Étude bioinformatique des génomes de Porphyromonas

Luis Alberto Acuña Amador

ANNÉE 2017

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Génétique, Génomique, Bioinformatique

Ecole doctorale Biologie Santé

présentée par

Luis Alberto ACUÑA AMADOR

Préparée à l'unité de recherche IGDR UMR 6290 CNRS - UR1
Institut de Génétique et Développement de Rennes
Composante universitaire : Sciences de la Vie et de l'Environnement

---

Étude bioinformatique des génomes de *Porphyromonas*

Thèse soutenue à Rennes
le 20 décembre 2017

devant le jury composé de :

Dr. Eric DUCHAUD
Directeur de recherche, VIM-UR0892, INRA, Jouy en Josas / *rapporteur*

Dr. Simonetta GRIBALDO
Directeur de Recherche, BMGE, Institut Pasteur, Paris / *rapporteur*

Dr. Monique ZAGOREC
Directeur de Recherche, UMR 1014 SECALIM, ONIRIS INRA, Nantes / *examinateur*

Dr. Hélène FALENTIN
Ingénieur de Recherche, UMR STLO, Agrocampus Ouest, Rennes / *examinateur*

Dr. Catherine ANDRÉ
Directeur de Recherche, UMR 6290 CNRS-UR1, IGDR, Rennes / *examinateur*

Dr. César RODRÍGUEZ SÁNCHEZ
Profesor, LIBA CIET, Facultad de Microbiología, Universidad de Costa Rica / *examinateur*

Dr. Frédérique BARLOY-HUBLER
Chargé de Recherche, UMR 6290 CNRS-UR1, IGDR, Rennes / *directeur de thèse*

Remerciements – Agradecimientos

Mes premiers mots de remerciements doivent aller à Fred. Tu m'as appris tant que dire un simple merci n'est pas suffisant. Malheureusement je ne connais pas d'autre mot. Tu m'as motivé, encouragé. Nous avons passé des moments difficiles, nous avons finalement réussi à nous en sortir. Merci, un grand merci. Je n'aurais pu mener cette bataille aussi loin sans toi. Merci.

Je veux remercier sincèrement les membres du jury pour le temps qu'ils vont consacrer à ce travail de thèse : Dr. Eric Duchaud et Dr. Simonetta Gribaldo qui ont accepté de juger ce travail en assumant le rôle de rapporteurs ; Dr. Monique Zagorec, Dr. Hélène Falentin, Dr. Catherine André et Dr. César Rodríguez qui ont accepté de participer à ce jury de thèse.

Un grand merci à Odile Tresse et à Mohamed Jebbar qui m'ont conseillé pendant mes comités de thèse. Un grand merci également à Denis Tagu, mon tuteur, qui m'a aidé à retrouver des bonnes conditions de travail pour ma dernière année. À Nathalie Théret également pour son rôle médiateur lors de cette période délicate, mes sincères remerciements.

Je voudrais également remercier l'équipe génétique du chien qui m'a accueilli pour la fin de ma thèse. Tout d'abord à Catherine, tu as eu toujours des mots encourageants, même avec ton planning surchargé. Tu m'as toujours fait sentir bien pour travailler dans ton équipe, merci beaucoup. Merci à Pascale pour tes conseils pour la rédaction de ce document, à Édouard pour notre dose de café même quand on avait pas le temps de descendre, à Maud, à Ronan et à Benoît pour vos encouragements et conseils, à Morgane et Fonzi pour les coupures dans la journée et les pensées positives toujours, à Christophe et à Thomas qui savent toujours nous faire rire et qui s'intéressent à notre bien-être, avec le sourire et des pensées positives. Merci à Sophie, toujours des bons conseils et des encouragements. Merci également aux doctorantes de l'équipe, courage pour quand ce sera votre tour, Solenne, Anaïs et Céline. Merci pour les soirées pâte feuilletée et brioche chez Céline, on a bien rigolé. Merci à Naoual pour tes encouragements et tes conseils. Merci à Nadine qui m'a accueilli dans sa pièce, merci de faire toujours au mieux ton travail et de nous aider avec le projet. Un grand merci à Annabelle, je crois que finalement j'ai réussi à apprendre mes couleurs, lol ; merci pour ton implication dans le projet microbiote, sans toi le projet ne serait pas où il est. Un grand merci à Aline, dès le début, tu t'es impliquée dans le projet, tu m'as aidé à mettre au point les

longues PCR qui ont terminé de valider nos constructions génomiques, et tout ça en plus de tout ce que tu fais dans l'équipe. Merci, merci beaucoup. Bon courage pour ta suite, tu mérites que de bonnes choses. Mes remerciements à l'équipe entière, et aux additionnels, Stéphane Dréano (avec tes carabes) et Stéphane Deschamps, pour ces cafés de 10h30, cet appel et vous voir descendre en meute pour passer 15-20 min à parler de tout et de rien vont me manquer. J'admire votre façon de passer des bons moments ensemble, de rigoler et de travailler dans une ambiance unique. Merci, travailler cette dernière année de ma thèse, c'était toujours un bonheur et c'est en grande partie grâce à vous.

Un grand merci à Reynald Gillet pour son accueil au sein de l'IGDR et pour son intervention cruciale pour apaiser les fortes tensions avec l'équipe que j'ai quittée. Merci à Géraldine (que je peux pas appeler GG), merci pour toujours m'aider avec tout ce que t'as pu, et tous les matins nous dire un « BONJOUR » qui nous fait sourire. Merci à Marion, toujours bienveillante et souriante. Merci à la gestion, principalement à Nadine et Isabelle, je n'ai jamais rencontré des personnes aussi efficaces et sympathiques que vous, un grand merci pour votre travail. Merci à Bénédicte, Sylvain, Philippe, toujours sympathiques, puis aux gens que je croise dans les couloirs et qui sont toujours souriants et agréables.

Merci à la famille Hubler, de m'avoir accueilli chez eux plusieurs samedis. Merci à Dominique, tant de bienveillance, les séances de yoga avec Fred et la relecture de ce manuscrit. Merci également à vous deux de m'avoir introduit au yoga, merci à Flora pour cette découverte.

Un grand merci également à tous les gens que j'ai croisé mes deux premières années à Bordeaux, profs, collègues de master et personnel de l'Inserm U853. Vous m'avez conforté dans mon envie de poursuivre en doctorat.

Mi más profundo agradecimento a la Oficina de Asuntos Internacionales y de Cooperación Externa y a la Universidad de Costa Rica que financiaron mis estudios doctorales. Gracias a todos los que fueron mis profesores en la Facultad de Microbiología, me siento muy orgulloso de haber aprendido de ustedes, gracias por transmitirme la curiosidad científica y las ganas de enseñar lo que sé, de aportar al país con mis conocimientos, de formar nuevos profesionales comprometidos con el desarrollo de Costa Rica. Gracias a Costa Rica, que supo invertir en su gente, en Salud y en Educación, en esos motores de desarrollo.

Espero que sepamos volver a ellos, cuidarlos y enriquecerlos; es lo que nos ha distinguido en el mundo y es lo que ha hecho de alguien como yo, sin importantes ingresos económicos, alguien que pudo salir a aprender y que quiere volver para mejorar el país, la vida de sus habitantes y seguir siendo ejemplo de país para el mundo. Gracias, gracias infinitas.

Gracias a César, Carlos, Evelyn, María del Mar, ñor Pablo, Martín, trabajar con ustedes siempre fue una dicha. Gracias a Diana, a Ricardo, a Diego, a Glori, a Nati, a Karencita Limón y a todos los demás con los que trabajé en el LIBA, ustedes hicieron de mis días allí de los más felices. Estoy ansioso de volver, trabajar y volverlos a ver.

Gracias a mis amigas de la U, Eva, Adri, Wendy, Tefa, Lucha, Calabacita, Tere y Llira; pasamos muy buenos tiempos. Gracias a mi amiga de toda la vida, gracias Pame por tus palabras *et tes encouragements*. Un abrazo, aunque no nos veamos tan seguido, sé que ahí están y ustedes saben que ahí estoy. Aprovecho para dar gracias a todos los profes y amigos del Franco, mi querido colegio, espero que algún día reciban a mi(s) hijo(a)(s).

Gracias a mi familia, en especial a doña Nena, mi Ayita, la matriarca. Vos formaste a tus hijas con el sudor de tu frente, con el mismo ahínco con el que luego ellas formaron a tus nietos, si estoy donde estoy es gracias a vos. ¡Qué fuerza y qué caracter! Espero tener un poco para enfrentar lo que venga. Gracias a Ma, otra fiera para sus hijos. Fuiste y seguirás siendo la mejor madre del mundo, la fuerza que me motiva todos los días para ser alguien de bien, alguien de quien te sintás orgullosa. No tengo palabras para agradecerte. Gracias a mis tías, María Elena, Sonia y Tina. Son un ejemplo de madres y de mujeres. En mi corazón no cabe el amor y el orgullo. A mis primos, Carlos Enrique, Ana Sofía, Juan Manuel, Rebecca, Laura y a mi gemela inteligente, Anamaría; gracias, cuiden siempre a sus madres. A papaito, gracias por tu apoyo. Gracias a Daniel, mi oso, a veces yo soy fuerte, a veces sos vos. Mi hermanito. Te quiero con todo mi corazón.

Gracias a Andrea y a su familia. Mi moo, gracias porque un sólo día no he estado sin vos. Horas de horas de teléfono y facetime. Tus berrinches, los míos, mi mal humor, el tuyo… Somos lampreas. Te amo, como nunca pensé que iba a querer a nadie, y como nunca pensé que alguien me iba a querer. Me has dado fuerzas y espero seguir el camino con vos, hasta que Dios quiera.

Gracias Dios mío por todo lo que me has dado, por cada prueba y por todas las alegrías y por todo lo que falta bueno y malo. Gracias.

a Ma, Ayita, Tina, Daniel y Andrea

Résumé

Les bactéries du phylum Bacteroidetes, classe Bacteroidia, sont parmi les plus importantes dans microbiotes gastrointestinaux des humains et d'autres mammifères. La bouche, entrée du tube digestif, est un environnement avec des sites anatomiques variés, auxquels s'associent des microbiotes de composition différente. L'union de la gencive et des dents, le sillon gingivo-dentaire ou sulcus, est un site de dépôt d'un biofilm complexe appelé plaque dentaire. Une bactérie de ce phylum, *Porphyromonas gingivalis*, est capable de perturber le système immunitaire humain et de produire un déséquilibre du biofilm oral également nommée dysbiose. Ceci déclenche la formation de la poche parodontale, un creusement pathologique du sulcus, et l'apparition de la parodontite. D'autres espèces du genre *Porphyromonas* sont également associées à la parodontite notamment chez les canidés.

Les populations de *P. gingivalis* sont panmictiques et la plasticité de leurs génomes importante. La bioinformatique peut aider à identifier les causes de la mosaïcité des génomes de cette bactérie, à étudier les facteurs de virulence au niveau du genre bactérien pour expliquer l'existence d'espèces pathogènes et d'autres commensales et à décrire la dysbiose liée à la parodontite.

La génomique comparative de *P. gingivalis* a démontré une corrélation entre le nombre de contigs dans les génomes draft de cette espèce et les répétitions génomiques, notamment des séquences d'insertion. Nous avons re-séquencé, re-assemblé et re-annoté trois souches de référence de cette bactérie qui avaient des génomes complets, en utilisant un séquençage en *long-read*. Nous avons mis en évidence des erreurs d'assemblage sur les trois génomes publiés, que nous avons corrigé. Une étude du pangénome de ces trois souches montre un génome *core* important. La plasticité de l'espèce serait donc plus dans l'organisation du génome que dans les différentes capacités de codage.

Une sous partie du génome *core*, dont les gènes ont un pourcentage d'identité nucléotidique plus faible que la plupart (*génome core variant*) est intéressante pour expliquer les différences phénotypiques de ces bactéries. Nous avons étudié la répartition d'un facteur de virulence, les fimbriae, structures d'adhésion, au sein du genre *Porphyromonas* et lié les *loci* à la phylogénie et au caractère pathogène des espèces.

Finalement, une description de la dysbiose qui a lieu lors d'une parodontite est faite par une analyse du microbiote de patients atteints de parodontite et d'individus sains. Les genres prépondérants lors des deux états sont mis en évidence.

Au cours de ces travaux, nous montrons l'importance de la biocuration et sa valeur ajoutée dans les travaux de génomique et bioinformatique en général. Seulement en faisant ce travail lent et lourd de biocuration, les réponses apportées aux questions biologiques seront pertinentes.

Abstract

Bacteria of Bacteroidetes phylum, Bacteroidia class, are amongst the more important in gastrointestimal microbiota, either human or from other mammals. The mouth, digestive tube entry, is an environment with varied anatomic sites, each having a particular microbiota with different composition. The union between gingiva and teeth, the gingival sulcus, is a site for biofilm (dental plaque) formation and accumulation. *Porphyromonas gingivalis*, a bacterium from this phylum, can modulate the inmune system and produce an oral biofilm desequilibrium called dysbiosis. This triggers the formation of a periodontal pocket, a pathological deepening of the gingival sulcus, and the emergence of periodontitis. Other *Porphyromonas* species are also associated to periodontitis, mainly in canids.

P. gingivalis populations are panmictic and their genomes are highly plastic. Bioinformatics can help to identify the causes of this genomic mosaicity, to study *Porphyromonas* virulence factors in order to explain why some species are pathogens and other are commensal, and to describe the dysbiosis linked to periodontitis.

P. gingivalis comparative genomics showed a correlation between the number of contigs in draft genomes and genomic repeats, mainly insertion sequences. We resequenced, reassembled and reannotated three reference strains of this bacterium that already had complete published genomes, using long-read sequencing. We showed that misassemblies were present in the three published genomes, and we corrected them. A pangenome study of the three strains showed that the core genome is preponderant. The species plasticity might be related more to the genome organization than to different coding capacities.

A subpart of th core genome, with genes having a nucleotidic identity percentage lower than the majority (*variable core genome*), is interesting for explaining the phenotypic differences of bacteria. We analysed the repertoire of a virulence factor, fimbriae, adhesion structures, in the *Porphyromonas* genus to link the loci to phylogeny and pathogenicity of its species.

Finally, we described the dysbiosis occuring with periodontitis, analysing gingival microbiota of patients having the illness and healthy individuals. Preponderant genera in both states are highlighted.

With this work, we demonstrate the importance of biocuration and its added value for genomic and bioinformatic studies in general. Only with this slow and arduous work, the answers to biological questions will be relevant.

Resumen

Las bacterias del *phylum* Bacteroidetes, clase Bacteroidia, son de las más importantes en los microbiotas gastrointestinales humanos y de otros mamíferos. La boca, entrada del tubo digestivo, posee sitios anatómicos diferentes a los cuales se asocian microbiotas de variable composición. La unión de la encía con los dientes, el surco gingival, es sitio de déposito de un *biofilm* complejo llamado placa dental. Una bacteria de este *phylum*, *Porphyromonas gingivalis*, es capaz de perturbar el sistema inmune y de producir un desequilibrio del *biofilm* oral denominado disbiosis. Esto desencadena la formación de una bolsa periodontal, la profundización patológica del surco gingival, y la aparición de la periodontitis. Otras especies del género *Porphyromonas* han sido asociadas a la periodontitis de otros mamíferos, en particular de los cánidos.

Las poblaciones de *P. gingivalis* son panmícticas y la plasticidad de sus genomas es importante. La bioinformática puede ayudar a identificar las causas de la mosaicidad de los genomas de esta bacteria, a estudiar sus factores de virulencia a nivel del género bacteriano para explicar la existencia de especies patógenas y de otras comensales, y a describir la disbiosis relacionada con la periodontitis.

Mediante nuestros trabajos de genómica comparada de *P. gingivalis*, demostramos una correlación entre el número de *contigs* en los genomas *draft* de esta especie y sus repeticiones genómicas, principalmente, secuencias de inserción. Secuenciamos, ensamblamos y anotamos de nuevo tres genomas de cepas de referencia de esta bacteria, usando una tecnología de secuenciación de *long-read*. Estas cepas ya poseían genomas completos. Mostramos que los tres genomas previamente publicados contenían errores de reconstrucción, que corregimos. El estudio del *pangenoma* de estas tres cepas muestra un porcentaje importante de genes en común, un genoma *core*. La plasticidad de la especie estaría relacionada más con la organización de su genoma que con el contenido de genes.

Una parte del genoma *core*, cuyos genes poseen un porcentaje de identidad nucleotídica más baja que la mayoría (*genoma core variante*) es interesante para explicar las diferencias fenotípicas de estas bacterias. Estudiamos la repartición de un factor de virulencia, las fimbrias, estructuras de adhesión, en el género *Porphyromonas* y relacionamos sus *loci* a la filogenia y a las características de patogenicidad de las especies de dicho género.

Finalmente, hicimos una descripción de la disbiosis que se lleva a cabo durante el desarrollo de una periodontitis, analizando los microbiotas gingivales de pacientes con periodontitis y de individuos sanos. Los géneros bacterianos preponderantes, en ambos estados, fueron descritos.

Mediante estos trabajos, mostramos la importancia de la biocuración y su valor agregado en los tabajos de genómica y de bioinformática en general. Solamente realizando este trabajo lento y arduo, las respuestas dadas a las preguntas biológicas serán pertinentes.

Table des matières

Table des illustrations

Abbréviations

^{32}P	Phosphore 32, isotope radioactif du phosphore
A	Adenine
ACPA	AntiCitrullinated Protein Antibodies ou autoanticorps anti-protéines citrullinées
ADN	Acide Désoxyribonucléique
AFLP	Amplified Fragment Length Polymorphism
ANI	Average Nucleotide Identity
ARN	Acide Ribonucléique
ARNm	ARN messager
ARNnc	ARN non-codants régulateurs
ARNr	ARN ribosomiques
ARNt	ARN de tranfert
ARPANET	Advanced Research Projects Agency Network
ASCII	American Standard Code for Information Interchange
ATCC	American Type Culture Collection
BAC	Bacterial Artificial Chromosome ou chromosomes bactériens artificiels
BBH	Best Bidirectional Hits ou meilleurs alignements réciproques
C	Cytosine
CDS	Coding DNA Sequence
CFB	Cytophaga-Flavobacter-Bacteroides
CM	Covariance Model ou modèle de covariance
COGs	Cluster of Orthologous Groups (of proteins)
CSC	Cancer Stem Cells ou Cellules Souches Cancéreuses
DDBJ	DNA Data Bank of Japan
ddNTP	didésoxyriboNucleotide Tri-Phosphate
dNTP	désoxyriboNucleotide Tri-Phosphate
DoE	Department of Energy des États-Unis
DUF	Domains of Unknown Functions
eDNA	ADN extracellulaire
eggNOG	evolutionary genealogy of genes : Non-supervised Orthologous Groups
ELF	Evil-Little Fellows
EMBL-EBI	European Molecular Biology Laboratory-European Bioinformatics Institute
emPCR	PCR en émulsion
EMT	Epithelial to Mesenchymal Transition ou Transition Épithélio-Mésenchimateuse
ENA	European Nucleotide Archive
ENIAC	Electronic Numerical Integrator And Computer
FIG	Fellowship for Interpretation of Genomes
FIGfams	Family of FIG
FORTRAN	FORmula TRANslator, un langage de programmation
G	Guanine
Gb	Gigabase ou 1 000 000 000 nt
GF	Germ-free
GGD	Genome-to-Genome Distance
HGT	Horizontal Gene Transfer ou transfert horizontal de gènes

HMM	Hidden Markov Models ou modèles de Markov cachés
HMP	Human Microbiome Project
HR	Holmes Ribgrass ou souche Ribgrass du Tobacco Mosaic Virus (TMV)
HTS	High-Throughput Sequencing, synonyme de NGS et de SGS
IBM	International Business Machines Corporation
ICE	Integrative and Conjugative Element
IHGSC	International Human Genome Sequencing Consortium
IMG ER	Integrated Microbial Genomes - Expert Review
INSDC	International Nucleotide Sequence Database Collaboration
kb	Kilobase ou 1 000 nt
LCB	Locally Collinear Blocks
Mb	Mégabase ou 1 000 000 nt
MG-RAST	MetaGenomics Rapid Annotation using Subsystems Technology
MGA	Multiple Genome Aligner
MLST	Multilocus Sequence Typing
multiMEMs	multiple Maximal Exact Matches
MUM	Maximal Unique Matchs
N	Nucléotide, représente une base quelconque
NAP	Nucleoid-Associated Proteins
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing, synonyme de HTS et de SGS
NIG	National Institute of Genetics (Japon)
NIH	National Institutes of Health des États-Unis
NLM	National Library of Medicine (États-Unis)
nt	nucléotides
OLC	Overlap-Layout-Consensus
ORF	Open Reading Frame
oriC	origine de réplication du chromosome
OSCC	Oral Squamous Cell Carcinoma ou carcinome à cellules squameuses gingivales
OTU	Operational Taxonomic Unit
PacBio	Pacific Biosciences
PAD	Peptidyl-Arginine Deiminase
pb	paires de bases
PC	Personal Computer
PCR	Polymerase Chain Reaction ou Réaction en Chaîne par Polymérase
PFGE	Pulsed Field Gel Electrophoresis
PGAAP	Prokaryotic Genome Automatic Annotation Pipeline
PWM	Positional Weight Matrix ou matrice de score-position
RAST	Rapid Annotation of microbial genomes using Subsystems Technology
RBS	Ribosome Binding Site
RFLP	Restriction Fragment Length Polymorphism
rrn	opéron des ARN ribosomiques
SGS	Second-Generation Sequencing, synonyme de HTS et de NGS
SMRT	Single-Molecule Real-Time
SOLiD	Sequencing of Oligonucleotids by Ligation and Detection
SPF	Specific-Pathogen-Free

SRA	Sequence Read Archive
T	Thymine
T9SS	Type 9 Secretion System ou système de sécrétion de type IX ou Por
ter	terminus de la réplication chromosomique
TGS	Troisième Génération de Séquençage
Tm	Melting Temperature ou température de fusion
TMV	Tobacco Mosaic Virus
U	Uracile
UNIX	UNiplexed Information and Computing Service, Unics qui sera après écrit Unix
VAMPS	Visualization and Analysis of Microbial Population Structure
WGA	Whole-Genome Amplification
WGS	Whole-Genome Shotgun actuellement aussi Whole-Genome Sequencing
ZMW	Zero-Mode Waveguide

Introduction

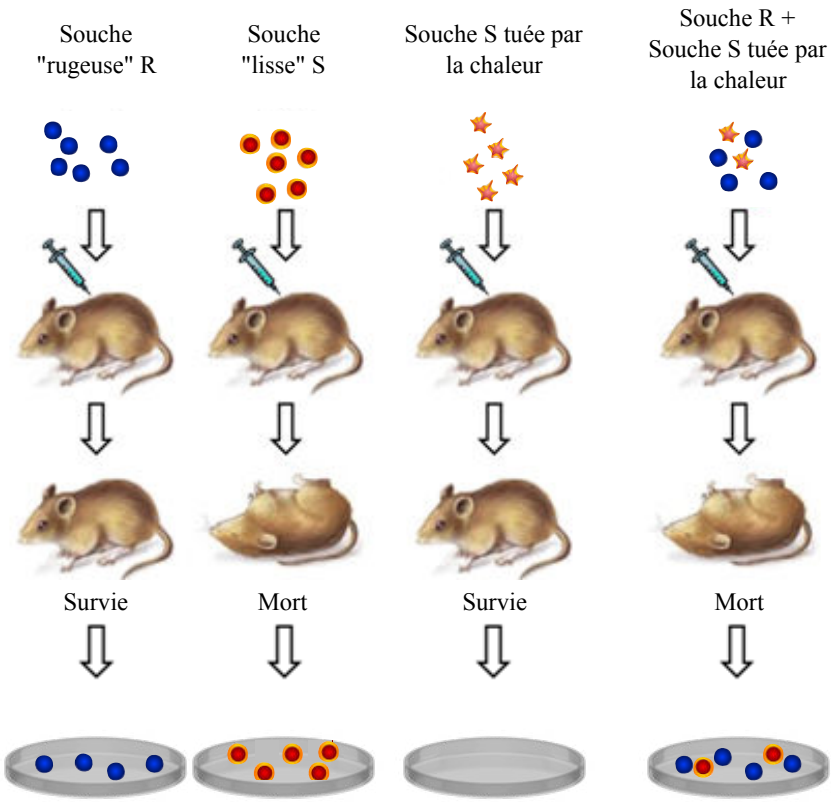


Figure 1. Expérience de Griffith. L'inoculation des souris avec des bactéries S virulentes tue les souris, tandis que les bactéries R avirulentes ne les tuent pas. Les bactéries S tuées par la chaleur ne sont pas létales, mais ces mêmes bactéries S tuées par la chaleur inoculées avec des bactéries R vivantes, sont létales pour la souris. Un "agent transformant" dans les bactéries S tuées par la chaleur est capable de transformer les bactéries R en S.

Adapté de http://istudy.pk/bacterial-transformation/ et https://commons.wikimedia.org/wiki/File:Griffith_experiment.svg
Téléchargées le 22 octobre 2017

1. De l'ADN aux annotations génomiques

1.1. Découverte de l'ADN comme support de l'information génétique

La première étude concernant la molécule d'ADN est généralement attribuée au chercheur suisse F. Miescher qui a réussi à isoler, en 1869, dans le noyau des cellules, une substance qu'il nomme nucléine, du latin *nucleus*, noyau (Wallis 1999). En 1889, son élève, R. Altmann isole les deux composants de la nucléine : des protéines et une substance acide qu'il désigne *acide nucléique* dont les quatre bases azotées (adénine, cytosine, thymine et guanine) seront décrites en 1896 par A. Kossel. Le désoxyribose, associé à l'acide nucléique est découvert en 1928 par P. Levene et J. Lunn.

En 1928, le médecin britannique F. Griffith décrit deux types de pneumocoques isolés de biopsies de poumon et d'expectorations de près de 300 cas de pneumonie (Griffith 1928). Les deux types sont appelés S pour *smooth* (lisse) et R pour *rough* (rugueux), pour l'aspect des colonies en culture pure. Le type S est virulent et tue les souris inoculées en quelques jours, tandis que les pneumocoques de type R sont avirulents. L'inoculation de souris avec le type R vivant et le type S tué par la chaleur entraîne la mort des souris et la récupération de colonies de type S **(Figure 1)**. Par cette expérience, Griffith montre qu'il survient une transformation des bactéries de type R en type S et que ce changement est stable et irréversible. Il existerait donc un "agent transformant" provenant des cellules mortes et conférant des propriétés héréditaires.

Plus d'une décennie après ces observations, O. Avery et collaborateurs caractérisent cet agent (Avery et al. 1944). Avec les méthodes de l'époque, ils démontrent que la fraction active, responsable de cette transformation, ne contient ni protéines, ni lipides, ni polysaccharides et correspond exclusivement à "une forme visqueuse et polymérisée d'acide désoxyribonucléique".

Malgré ces preuves, la communauté scientifique de l'époque est divisée. La simplicité de l'acide désoxyribonucléique (ADN) ne pourrait pas véhiculer une information complexe comme l'hérédité et les protéines, bien plus diverses et complexes, sembleraient de bien meilleures candidates. Peu à peu, plusieurs travaux s'accumulent pour démontrer que l'ADN est bien le porteur de l'information génétique. En 1950, un chimiste autrichien, E. Chargaff, prouve que cette molécule est constituée de quantités variables de quatre bases azotées : deux

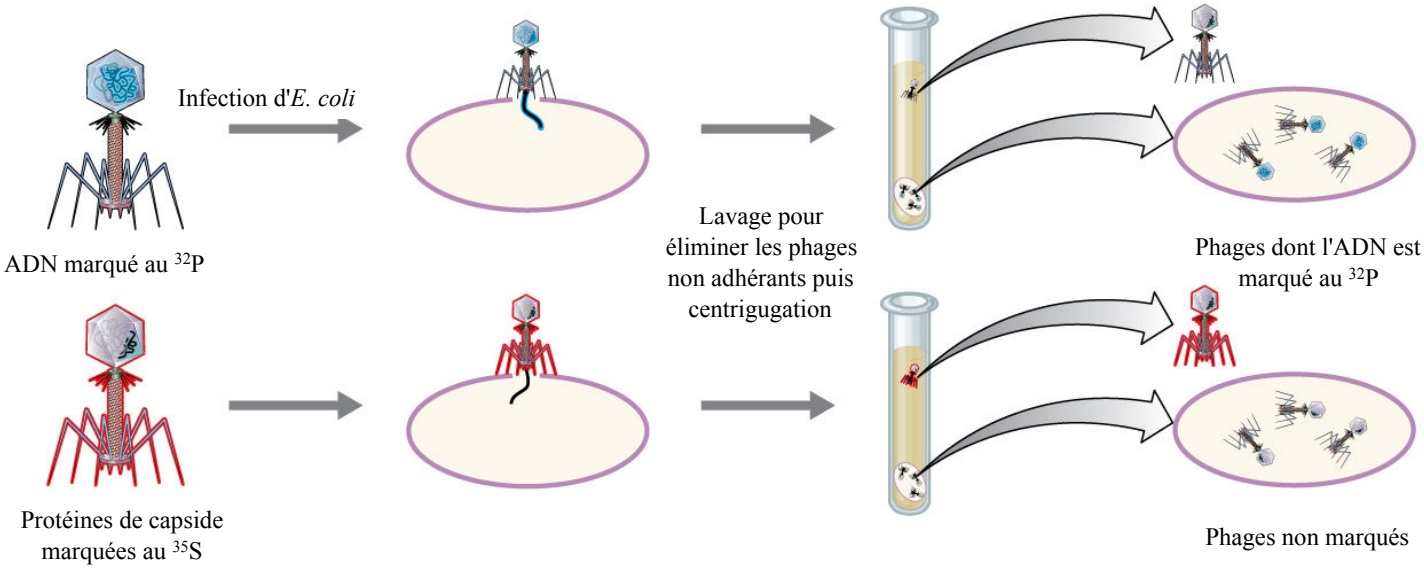


Figure 2. Expérience de Hershey & Chase. L'ADN du phage T2 est marqué au phosphore 32, une autre culture de phages est marquée au souffre 35 qui s'incorpore aux protéines de capside. Ces deux types de phages sont mis en contact avec une culture d'*Escherichia coli*. Après l'infection, les cellules sont lavées pour éliminer les phages non adhérés aux bactéries, puis les cultures sont centrifugées. Les bactéries infectées par les phages marqués au ^{35}S ne produisent pas des phages marqués, tandis que les cellules infectées par les phages marqués au ^{32}P, produisent des phages marqués au ^{32}P.

Adapté de https://commons.wikimedia.org/wiki/File:OSC_Microbio_10_01_HersheyChase.jpg
Téléchargée le 22 octobre 2017

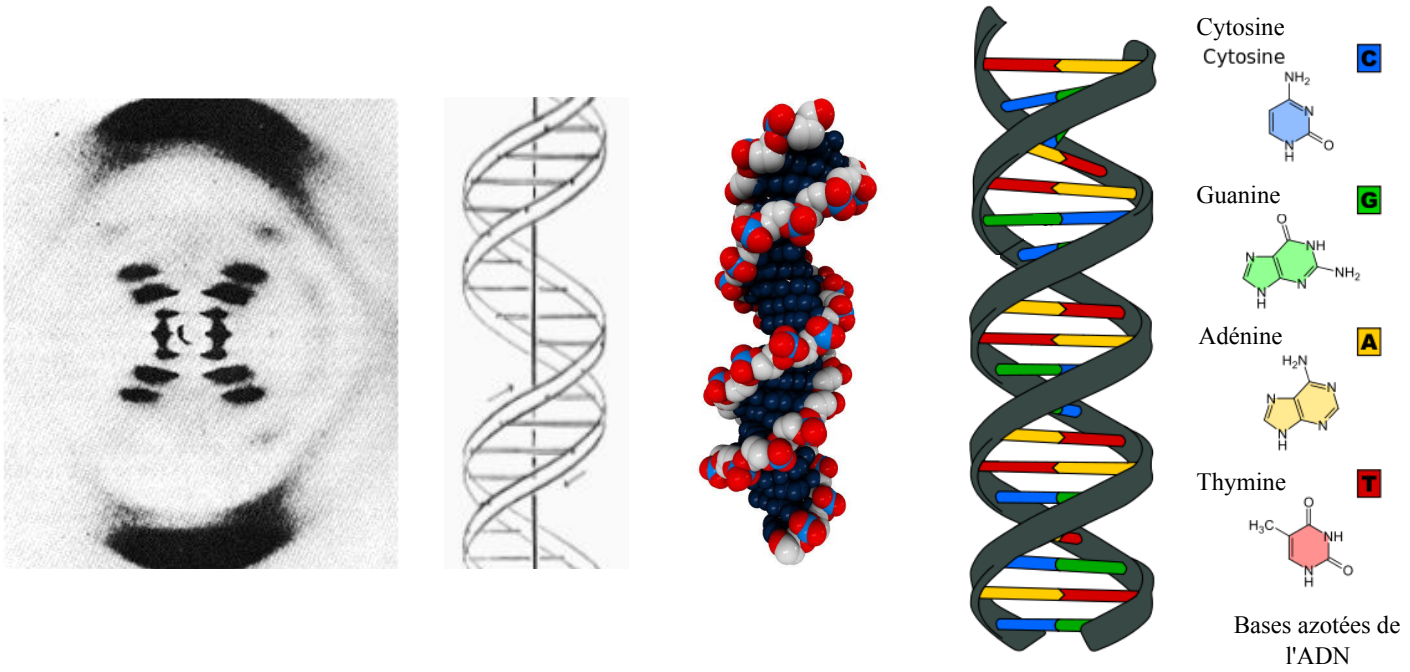


Figure 3. Structure de l'ADN. Photographie de la cristallographie aux rayon X de l'ADN, modèle schématique de l'ADN présenté par Watson et Crick dans leur publication de 1953 dans Nature, représentation des atomes dans une double hélice d'ADN et schéma en couleur de l'ADN, avec les différentes bases azotées représentées à droite.

Adaptés de https://knowledgeclass.blogspot.fr/2015/05/chemical-composition-of-dna.html ; Watson and Crick 1953a ; http://www.enzymlogic.com (pris de flickr.com) et https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg
Téléchargées le 22 octobre 2017

purines (adénine A et guanine G) et deux pyrimidines (cytosine C et thymine T) et que sa composition est caractéristique de chaque espèce. Il propose deux règles qui portent son nom (les règles de Chargaff) qui stipulent que dans l'ADN, la somme des purines égale la somme des pyrimidines et que les ratios entre les bases A-T ou C-G sont constants et égaux à un, et ceci dans toutes les espèces étudiées (Chargaff 1950).

Une des preuves les plus remarquable viendra des travaux d'A. Hershey et M. Chase en 1952 **(Figure 2)**. En marquant différentiellement l'ADN et les protéines du phage T2 puis en infectant une souche sensible d'*Escherichia coli*, ils démontrèrent que seul l'ADN est capable de pénétrer la bactérie, les protéines de capside restant à la surface de la bactérie. Cette internalisation de l'ADN seul suffit à produire de nouvelles particules phagiques, ce qui prouve que chez les virus, c'est bien la molécule d'ADN qui constitue le matériel héréditaire (Hershey and Chase 1952).

En 1953, sur la base des observations réalisées par R. Franklin et M. Wilkins en diffraction aux rayons X, J. Watson et F. Crick publient le fameux modèle d'ADN en double hélice **(Figure 3)**, constitué de deux brins antiparallèles de groupements phosphate et désoxyribose, appariés par des liaisons hydrogène entre les pyrimidines d'un brin et les purines de l'autre : A avec T et C avec G (Watson and Crick 1953a). Cet appariement correspond parfaitement aux règles de Chargaff et suggère le mécanisme de reproduction du matériel génétique par dissociation des deux brins, chacun servant de matrice à une nouvelle double hélice (Watson and Crick 1953b). En 1958, M. Meselson et F. Stahl confirment que la réplication de l'ADN est semi-conservative (Meselson and Stahl 1958) et la même année, l'équipe de A. Kornberg isole et caractérise l'ADN polymérase, enzyme responsable de cette réplication semi-conservative (Lehman et al. 1958).

En 1955, les expériences de H. Fraenkel-Conrat et R. Williams sur différents sérotypes du virus ARN de la mosaïque du tabac montrent que l'infection de la plante avec un virus hybride reconstitué à partir de protéines de TMV (*tobacco mosaic virus* commun) et l'ARN du HR (*Holmes ribgrass* aussi connu comme *Ribgrass strain of TMV*) résulte en lésions caractéristiques de HR. Les particules virales produites *in planta* sont identiques en composition d'acides aminés et sérologiquement au virus HR qui a fourni l'ARN. L'ARN est donc le déterminant génétique (Fraenkel-Conrat and Williams 1955). L'année suivante, les travaux de P. Zamecnik et collaborateurs décrivent que la synthèse des protéines a lieu dans des "particules microsomales constituées de nucléoprotéines", les ribosomes (Zamecnik et al.

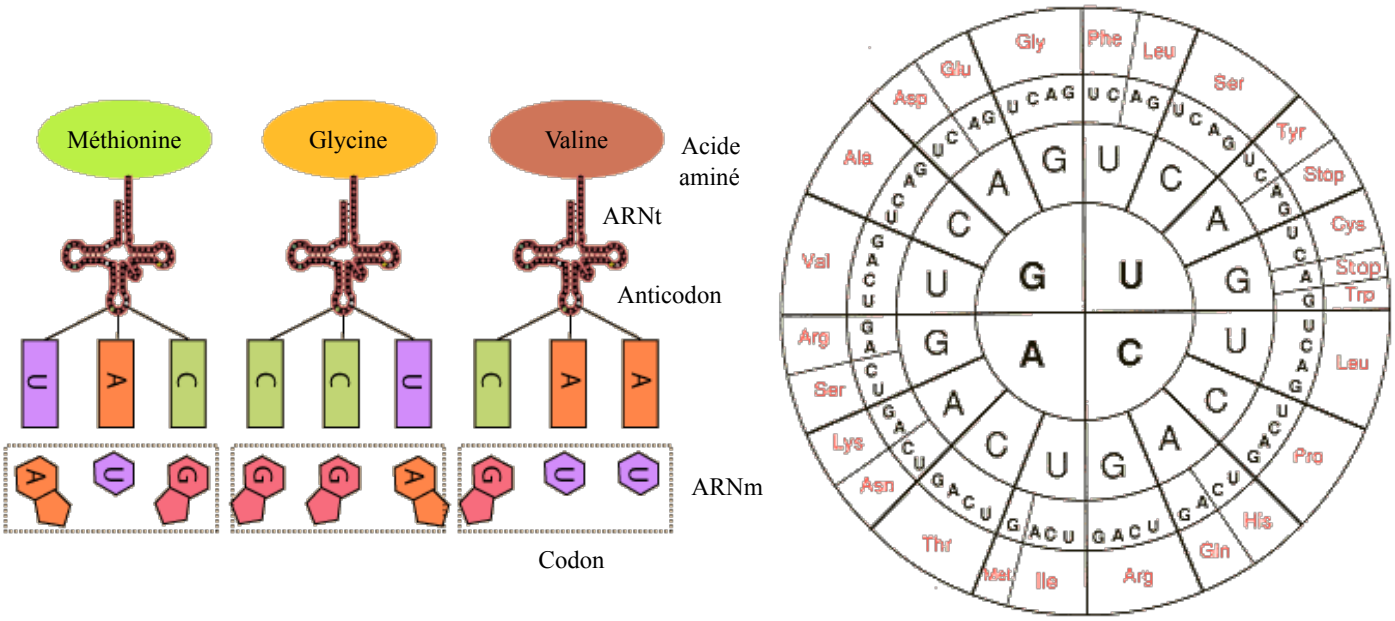


Figure 4. Le code génétique. La lecture de l'ARNm se fait par des triplets non-chevauchants appelés codons. À chaque codon, correspond un anticodon sur les ARNt qui transportent un seul acide aminé. À droite, une représentation circulaire du code génétique avec toutes les correspondances pour les 64 codons possibles. La lecture se fait de l'intérieur vers l'extérieur et la correspondance en acide aminés peut être lue en rouge (code de trois lettres pour chaque acide aminé).

Adaptés de http://hyperphysics.phy-astr.gsu.edu/Nave-html/Faithpathh/codelife2.html et https://www.flickr.com/photos/aleiex/1208438950/
Téléchargées le 22 octobre 2017

1956). Avec peu d'informations supplémentaires, F. Crick émet deux principes qu'il ne peut pas prouver, l'hypothèse de la séquence (*The Sequence Hypothesis*) et le dogme central (*The Central Dogma*). L'hypothèse de la séquence est que la spécificité d'un acide nucléique est portée seulement par la séquence de ses bases et que cette séquence est un code simple pour la séquence d'acides aminés d'une protéine donnée. Le dogme central est que la transmission de l'information pour la détermination de la séquence est unidirectionnelle avec seulement deux possibilités : d'un acide nucléique vers un autre acide nucléique ou d'un acide nucléique vers une protéine (Crick 1958).

La décennie suivante plusieurs travaux visent à décrypter le code des acides nucléiques pour prédire la séquence des acides aminés. En 1957, l'hypothèse du code génétique par triplets non-chevauchants est émise (Crick et al. 1957) et la synthèse *in vitro* d'un polypeptide de phénylalanine à partir d'un ARN-poly-U (Matthaei et al. 1962) ont servi à décoder le premier codon UUU qui code donc pour la phénylalanine. L'élucidation finale du code génétique par lecture non chevauchante et dégénérée de triplets d'ADN **(Figure 4)** fut le travail de plusieurs chercheurs (Khorana et al. 1966; Matthaei et al. 1962; Nirenberg and Leder 1964; Nirenberg et al. 1966; Ochoa 1964). Depuis, certaines de ces découvertes ont été nuancées avec les phénomènes d'ARN non codant, d'épissage alternatif ou bien encore de modifications épigénétiques. Cependant, encore aujourd'hui, presque 60 ans après ces découvertes, l'ADN reste une molécule centrale de la recherche en biologie.

Simultanément à ces découvertes sur la molécule d'ADN et sur les mécanismes associés, l'informatique émerge, avec l'apparition des premiers calculateurs puis des premiers ordinateurs ; la biologie et l'informatique, deux sciences que, combinées, permettront l'émergence d'une nouvelle discipline : la bio-informatique.

1.2. Emergence de la (bio)-informatique

D'après le dictionnaire Larousse, le terme *informatique* vient des mots information et automatique, c'est la "science du traitement automatique et relationnel de l'information considérée comme le support des connaissances et des communications". Elle correspond également à "l'ensemble des applications de cette science, mettant en œuvre des matériels (ordinateurs) et des logiciels". En anglais, l'informatique est appelée *computer science*, accentuant le lien de cette science avec les *computers*, calculateurs qui sont à l'origine des

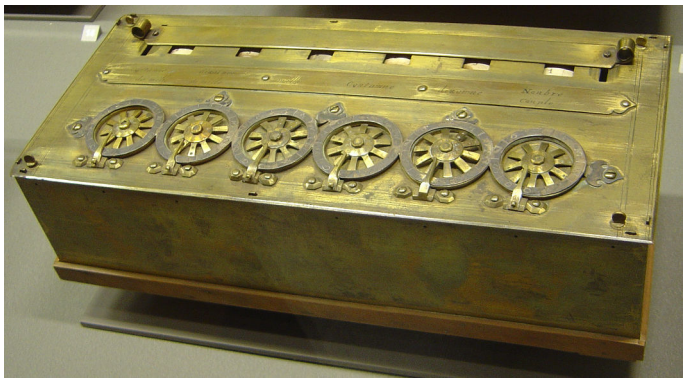


Figure 5. Horloge calculante de Schickard et la Pascaline. Deux des premières machines à calculer de l'histoire, à gauche : réplique du "Rechenuhr" de l'astronome allemand Schickard ; à droite, Pascaline exposée au Musée des Arts et des Métiers à Paris.

Extraits de https://commons.wikimedia.org/wiki/Category:Schickard%27s_calculating_machine#/media/File:Schickardmaschine.jpg et https://commons.wikimedia.org/wiki/Pascaline#/media/File:Arts_et_Metiers_Pascaline_dsc03869.jpg
Téléchargées le 22 octobre 2017

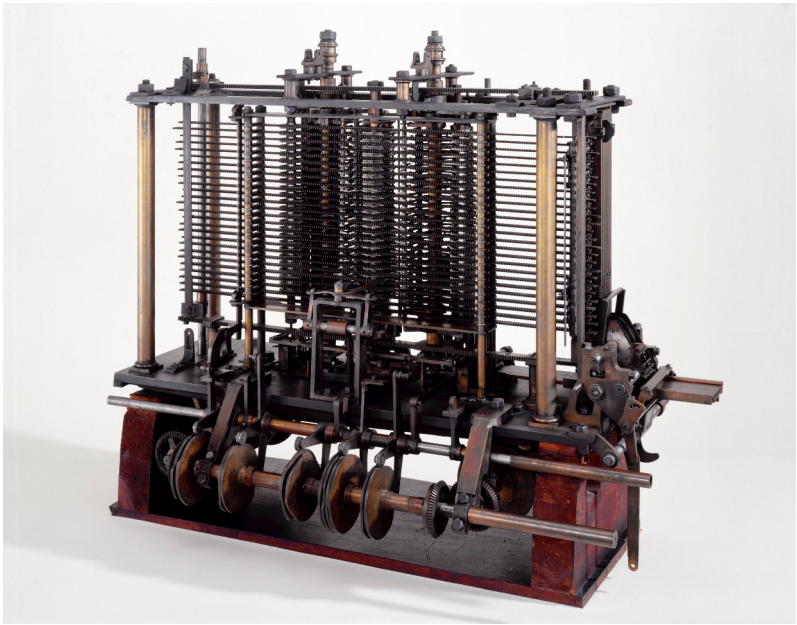


Figure 6. La machine analytique de Babbage. Une des premières machines analytiques construites à partir des notes de Babbage. Actuellement exposée au Science Museum à Londres. À droite, les cartes perforées utilisées pour programmer la machine.

Extraits de https://blog.sciencemuseum.org.uk/the-pride-and-passion-of-mr-babbage/ et http://history-computer.com/Babbage/AnalyticalEngine.html
Téléchargées le 22 octobre 2017

machines à calculer, puis des appareils électroniques capables d'exécuter des opérations arithmétiques et logiques appelées ordinateurs (Collen 1994).

1.2.1. Les premiers calculateurs et machines à calculer

L'informatique puisse ses fondements loin dans l'histoire. Ainsi, les calculateurs modernes dérivent d'un principe élaboré en Abyssinie et en Mésopotamie vers 2000 avant J.-C. à l'aide de pierres disposés sur le sol. De cette pratique dérive le nom moderne de calcul (*calculus* signifiant caillou en latin). Différentes appareils pour aider à calculer apparaissent, tables de calcul, abaques ou bouliers, simultanément chez plusieurs peuples (Étrusques, Grecs, Égyptiens, Indiens, Chinois…), puis la première règle à calculer, inventée par l'écossais J. Napier entre 1617 et 1620 et connue sous le nom de « Bâton de Neper » (Overton 2001).

Les premières machines à calculer mécaniques **(Figure 5)** dateraient de 1623 avec *l'horloge calculante* de l'astronome allemand W. Schickard et de 1642 pour la *Pascaline*, inventée par le philosophe et mathématicien français B. Pascal (Mainzer 2004; Overton 2001). Les premiers automatismes permettant d'exécuter des séquences d'opérations préenregistrées étaient sans rapport avec le calcul et concernaient d'abord les horloges puis les automates, très en vogue au XVIII^{ème} siècle. Au cours de ce siècle, apparaissent les cartes perforées, utilisées sur les métiers à tisser pour contrôler les mouvements des aiguilles, système mis au point par B. Bouchon en 1725, puis amélioré par J-M. Jacquard en 1800.

En 1833, le mathématicien britannique C. Babbage est le premier à énoncer le principe d'un ordinateur. Constatant que les tables de calculs comportaient beaucoup d'erreurs, il essaie de concevoir une machine qui exécuterait le travail sans fautes. Il utilise ces cartes perforées pour stocker des informations et fournir des instructions à sa *machine analytique* **(Figure 6)**. Cet appareil, qu'il ne pourra pas terminer, est l'ancêtre mécanique des ordinateurs. Il a une unité d'entrée qui reçoit les cartes perforées, une unité de traitement qui préfigurait les processeurs actuels, une unité de commande qui supervise les opérations et un magasin (*store*) où étaient rangées les données et les résultats intermédiaires et finaux (Collen and Kulikowski 2015; Mainzer 2004).

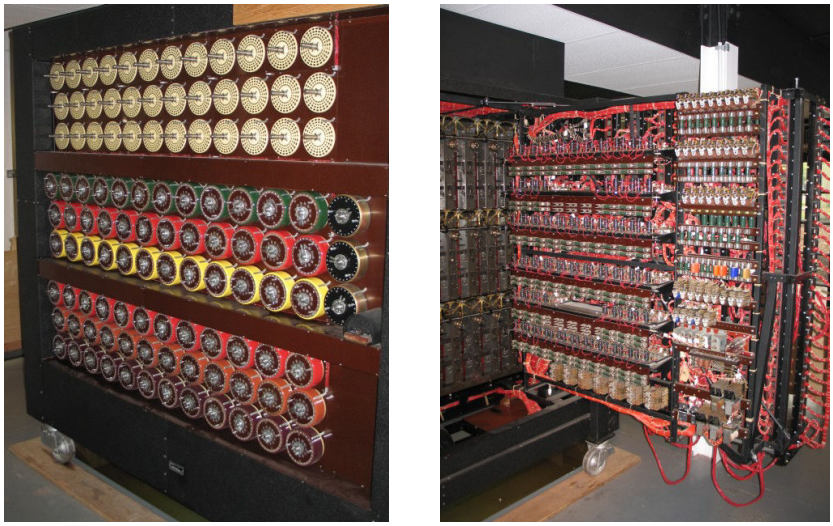


Figure 7. Réplique de Bombe. Réplique complète de la machine électro-mécanique utilisée par les britaniques pour décrypter les messages codées des Nazies pendant la Seconde Guerre Mondiale, conçue par Alan Turing à Bletchley Park. À gauche, vue de la machine fermée et à droite, ouverte.

Extraits de https://commons.wikimedia.org/wiki/Category:Bletchley_Park_Bombe#/media/File:A_Turing_Bombe,_Bletchley_Park_-_geograph.org.uk_-_1590986.jpg et ~590997.jpg
Téléchargées le 22 octobre 2017

Figure 8. L'ENIAC et l'IBM 650. Premier ordinateur de l'histoire, l'ENIAC fut utilisé pour le calcul de trajectoires des missiles par l'armée des États-Unis. L'IBM 650 peut être considéré comme le premier ordinateur commercial et plusieurs universités se sont équipées de cet ordinateur.

Extraits de https://commons.wikimedia.org/wiki/ENIAC#/media/File:Classic_shot_of_the_ENIAC.jpg et https://commons.wikimedia.org/wiki/Category:IBM_650#/media/File:IBM_650_at_Texas_A%26M.jpg
Téléchargées le 22 octobre 2017

En 1842, A. Lovelace, collaboratrice de C. Babbage, définit le principe des itérations successives dans l'exécution d'une opération. Elle crée ainsi le premier algorithme destiné à être exécuté par une machine, et devient la première *programmatrice* (Sammet 1969).

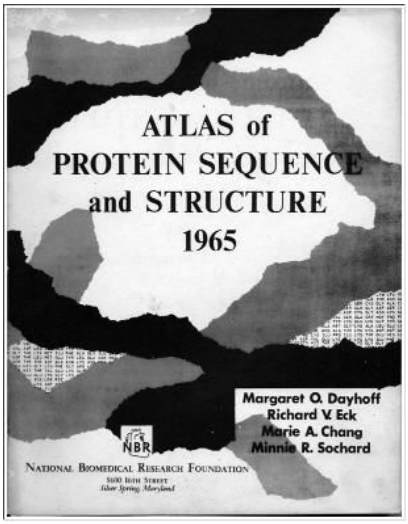
En 1896, H. Hollerith crée la société *Tabulation Machine Corporation* qui deviendra *International Business Machines Corporation* (IBM) en 1924. En 1937, G. Stibitz, ingénieur de Bell Labs, met au point le premier calculateur binaire électromécanique à partir de relais téléphoniques et faisant appel à la logique booléenne mise en place par G. Boole en 1854 (Collen and Kulikowski 2015). En 1937, A. Turing publia les principes des machines appelée actuellement *Machines de Turing*, capables de suivre de manière autonome les ordres codés par un algorithme (Turing 1937). Pendant la Seconde Guerre Mondiale, Turing et ses collaborateurs ont utilisé le premier calculateur électro-mécanique, Bombe **(Figure 7)**, pour décoder le code Enigma des Nazis, prouvant que les calculateurs pouvaient être utilisés pour bien plus que traiter des chiffres (Collen and Kulikowski 2015).

1.2.2. Premiers ordinateurs et émergence conjointe de la bioinformatique

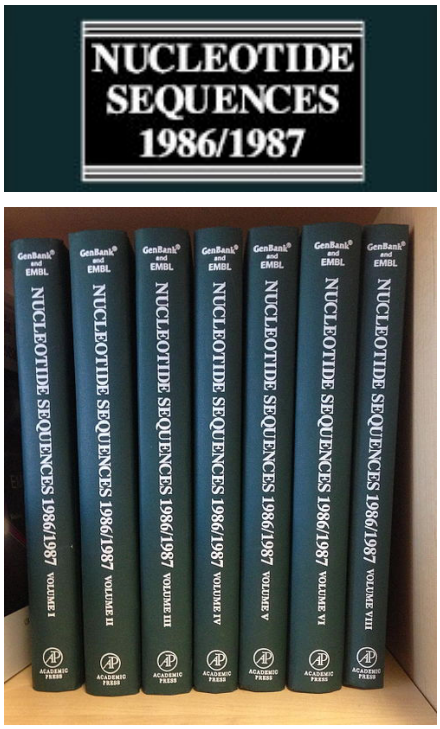
En 1943, P. Eckert et J. Mauchly, aux États-Unis, proposent la construction du premier véritable ordinateur de l'histoire, connu sous le nom de *Electronic Numerical Integrator and Computer* (ENIAC) pour aider l'armée à calculer les trajectoires des nouvelles armes et munitions produites pour la Seconde Guerre Mondiale **(Figure 8)**. Sa construction se termina en février 1946, six mois après la fin de la guerre (Polachek 1997).

Afin de communiquer avec la machine, les premiers langages de programmation, dits d'assemblage, sont utilisés pour représenter, avec des mots lisibles par l'humain, des instructions en langage machine (binaire). En 1955, la compagnie IBM, introduit l'IBM 650, le premier ordinateur *commercial* **(Figure 8)** et offre des "bourses d'études" (*educational grants*) aux universités pour qu'elles s'en équipent. Plusieurs universités ont installé l'ordinateur, principalement pour réaliser des calculs statistiques. Peu à peu des structures d'informatique ont été créées (Galler 1986). Entre 1954-1957, le langage FORTRAN (FORmula TRANslator), un des premiers langages procéduraux de haut niveau est développé par J. Backus et collaborateurs (Backus and Heising 1964).

Dans les années 1960, alors que le code génétique est déchiffré et que J. Monod, F. Jacob et A. Lwoff décrivent les mécanismes de la régulation génétique (Anonymous 1965),



50 séquences de protéines



Collection de 8 livres

NCBI ENA DDBJ DNA Data Bank of Japan

Year	Nucleotides in assembled/annotated sequences
September 2012	450 481 663 919
September 2013	670 004 320 378
September 2014	997 958 152 853
September 2015	1 401 669 271 501

Figure 9. Les bases de données de 1965 à nos jours. La croissance des bases de données est vertigineuse. Dès l'ouvrage de Dayhoff en 1965 à la fin des années 1980, les données étaient publiées en livres ; puis en CD-ROM au début des années 1990. Actuellement, la quantité de données double près de tous les 18 mois.

Adapté de http://blog.openhelix.eu/?p=1078 ; https://commons.wikimedia.org/wiki/File:NucleotideSequences_86_87.jpeg et Cochrane et al. (2016) 'The International Nucleotide Sequence Database Collaboration' *Nucleic Acids Res*, 44 (Database issue), D48-D50.
Téléchargées le 22 octobre 2017

les ordinateurs se dotent d'écran, de clavier, de mémoire virtuelle et de circuits imprimés (puces) leur permettant de réaliser des millions d'opérations par seconde, les faisant plus facilement utilisables. Plusieurs utilisateurs, des chercheurs de différents domaines pour la plupart, réalisent l'intérêt des ordinateurs dans les sciences de la vie.

En 1964, M. Dayhoff utilise des logiciels écrits en Fortran pour aider à déterminer la séquence d'acides aminés d'une protéine en utilisant l'information des peptides résultant de son hydrolyse. Elle fut une des premières scientifiques à démontrer concrètement l'utilité des ordinateurs en biologie et à rendre son code publique (Dayhoff 1965b). Cette même année, elle publie le premier recueil, sous forme de livre, de 50 protéines **(Figure 9)** : *Atlas of Protein Sequence and Structure*, ancêtre de nos banques de données modernes (Dayhoff 1965a). En 1967, W. Fitch et E. Margoliash établissent les premiers arbres phylogénétiques en comparant, par ordinateur, les séquences des protéines du cytochome C de nombreuses espèces animales (Fitch and Margoliash 1967).

En 1969, le système opératif *UNiplexed Information and Computing Service* ou *Unics* qui deviendra UNIX et l'*Advanced Research Projects Agency* Network (ARPANET), l'ancêtre de notre internet actuel, font leur apparition. À l'origine, ARPANET permettait de relier quatre universités américaines (Glowniak 1998). Pendant les dix années suivantes se développe l'email, l'Ethernet, l'Internet et les ordinateurs personnels (*Personal Computer* ou PC par des entreprises comme IBM et Apple) alors que la biologie progresse vers une vision moléculaire du vivant avec le développement du génie génétique et de la biologie moléculaire.

Cette nouvelle approche de l'étude du vivant va entraîner une accumulation sans précédent de données variées (gènes, génomes, protéines…), qu'il est nécessaire de stocker et d'analyser. Aux États-Unis, la *National Library of Medicine* (NLM) est très consultée dans les années 1980 via Internet. La création d'une subdivision de la NLM, le *National Center for Biotechnology Information* (NCBI) en 1988, regroupe plusieurs outils et bases de données, par exemple l'outil BLAST d'alignement de séquences (Altschul et al. 1990), le système de recherche et d'accès à toutes les données du NCBI via ENTREZ (littéralement *entrez*), les banques de données de séquences nucléiques GenBank crée en 1982, entre bien d'autres (K. A. Smith 2008). La *International Nucleotide Sequence Database Collaboration* (INSDC) est une initiative de coopération des bases de données nucléotidiques : GenBank du NCBI, *European Nucleotide Archive* (ENA) du *European Molecular Biology Laboratory-European*

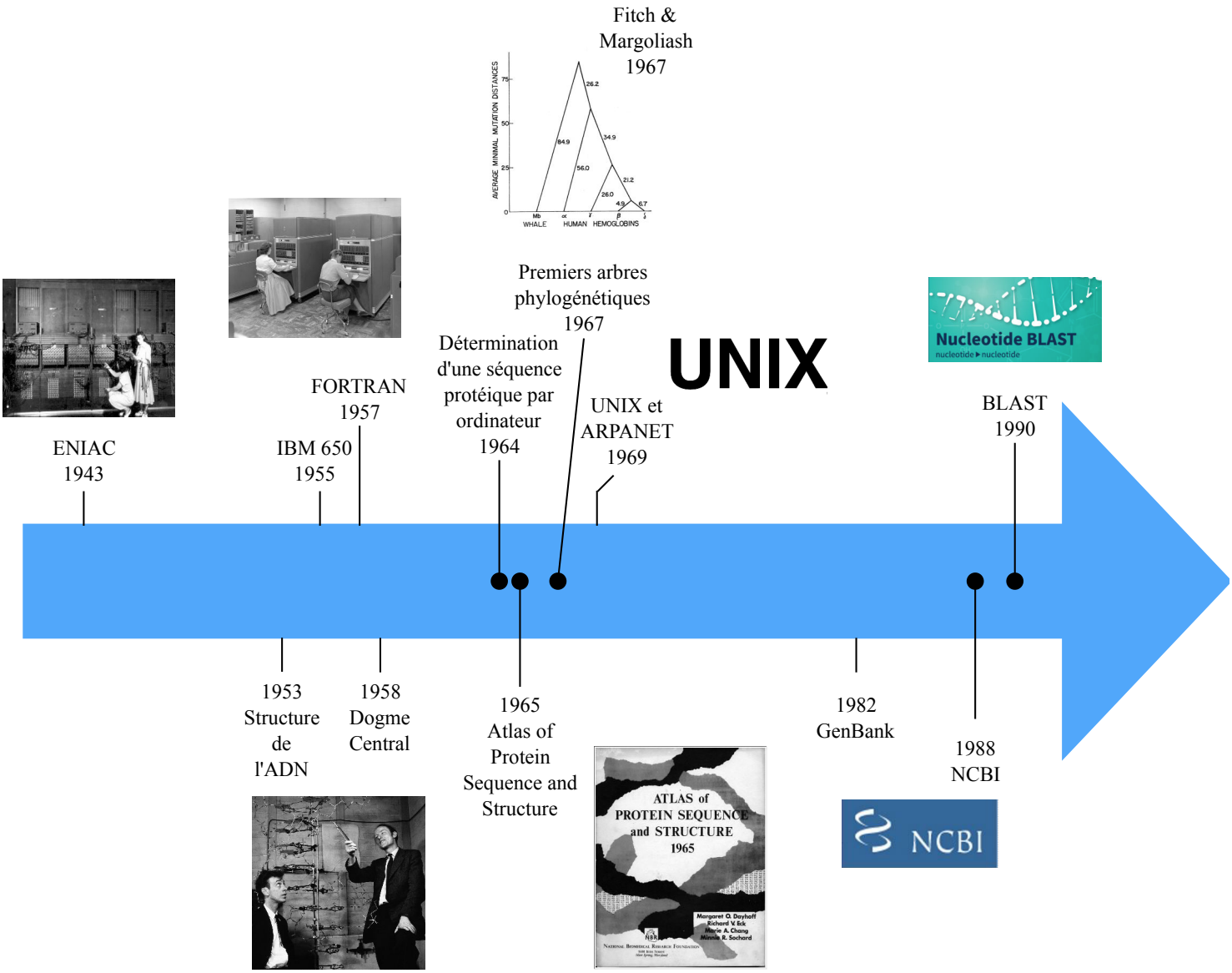


Figure 10. Évolution de l'informatique et apparition de la bioinformatique. Dès les années 1940 avec l'apparition des premiers ordinateurs aux années 1990, informatique et biologie ont progressé en parallèle au point qu'au milieu des années 1960, la distinction entre les deux disciplines est de plus en plus floue. De nos jours, nous parlons de bioinformatique et nous ne concevons plus la biologie sans la biologie *in silico*.

Photos extraites de : https://www.flickr.com/ (thekibster ENIAC ; Mikel Agirregabiria ibm650de1953) ; http://www.thehistoryblog.com/archives/date/2013/05/11 ; http://blog.openhelix.eu/?p=1078 ; https://commons.wikimedia.org/wiki/File:Unix_Logo.gif ; https://www.ncbi.nlm.nih.gov et Fitch & Margoliash 1967.
Téléchargées le 24 octobre 2017.

Bioinformatics Institute (EMBL-EBI) et le *DNA Data Bank of Japan* (DDBJ) du *National Institute of Genetics* (NIG). Les années récentes, ont connu une explosion de données, des bases de données, des méthodes pour les traiter **(Figure 9)**.

La "bioinformatique" marque l'apparition d'une nouvelle branche de la biologie, qualifiée d'*in silico*, discipline à part entière qui collecte et d'analyse des données biologiques complexes, en élaborant des outils et des stratégies adaptées pour les traiter et les comprendre **(Figure 10)**.

1.3. Génomes et génomique

D'après le dictionnaire Larousse, le terme *génome* correspond à "l'ensemble du matériel génétique, c'est-à-dire des molécules d'ADN, d'une cellule". Le terme *génome* est attribué au botaniste allemand H. Winkler, qui utilisa ce nom pour décrire l'ensemble des gènes (déterminants héréditaires) présents les chromosomes et spécifique à chaque espèce. Chaque organisme, virus, procaryote ou eucaryote, possède un génome composé d'ADN, à l'exception de certains virus à ARN. Le génome porte l'ensemble des informations génétiques d'un individu et sera intégralement ou partiellement transmis à sa descendance (Lederberg 2001).

L'organisation des génomes des cellules eucaryotes et procaryotes diffère. Chez les procaryotes, l'ADN est généralement haploïde, constitué d'un chromosome unique et circulaire, compacté en nucléoïde. Certaines ADN circulaires extra-chromosomiques nommées plasmides ou méga-plasmides sont parfois présents. Chez certains organismes bactériens, deux chromosomes circulaires distincts peuvent exister et une petite fraction des chromosomes bactériens (tels que ceux de *Streptomyces*, *Agrobacterium* et *Borrelia*) sont linéaires et possèdent des télomères, toutefois très différentes de ceux des chromosomes eucaryotes (Badrinarayanan et al. 2015). Chez les eucaryotes, le terme génome, évoque essentiellement le génome nucléaire diploïde qui comporte un nombre N de chromosomes qui varie selon les espèces (par exemple N=14 pour le parasite *Plasmodium falciparum*, responsable du paludisme, N=46 pour l'humain et N=308 pour le mûrier noir). Au génome nucléaire, s'ajoute un génome mitochondrial chez la quasi-totalité des eucaryotes et un génome chloroplastique chez les plantes et les algues photosynthétiques (Howe et al. 2003; Shi et al. 2017).

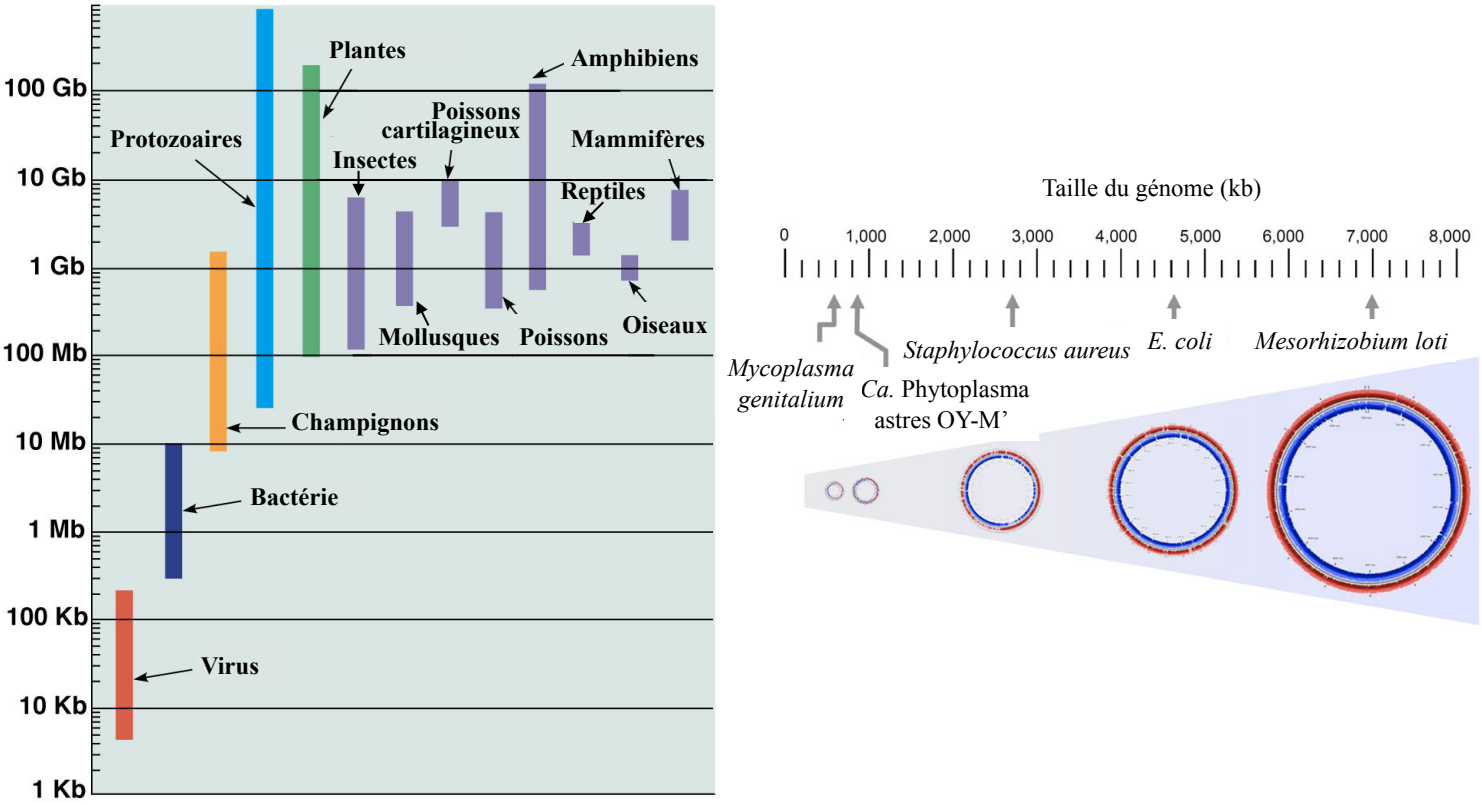


Figure 11. Tailles de génomes d'organismes procaryotes et eucaryotes. À gauche, rang des tailles des génomes par groupe d'organismes. À droite, quelques exemples de tailles de génomes bactériens.

Adapté de https://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-18/18_11.jpg et Oshima et al. (2013) 'Genomic and Evolutionary aspects of phytoplasmas' *Front Microbiol*, 4, 230.
Téléchargées le 22 octobre 2017

En génomique, la taille des génomes est mesurée en nucléotides (nt) et exprimé en paires de bases (pb) : kilobase (kb = 1 000 nt) ou mégabase (Mb = 1 000 000 nt). La biodiversité des tailles génomiques est remarquable, allant de petits génomes de milliers de bases pour les virus à des génomes gigantesques de 147.3 Gb chez une espèce de fougère, *Tmesipteris obliqua* (Hidalgo et al. 2017). Chez les bactéries, le rang est d'environ 150 kb chez des bactérie endosymbiotiques comme *Carsonella ruddii* (Riley et al. 2017) et *Hodgkinia cicadicola* (McCutcheon et al. 2009) à 12-13 Mb pour des bactéries des environnements complexes comme *Sorangium cellulosum* (Han et al. 2013) ou *Archangium gephyra* (Sharma and Subramanian 2017) **(Figure 11)**.

La *génomique*, est un terme proposé en 1986 par T. Roderick, désignant l'étude des génomes. Cette discipline de la bioinformatique, est à l'interface de la biologie et de l'informatique et intègre les connaissance de biologie cellulaire et moléculaire avec les techniques informatiques et mathématiques pour permettre une interprétation des séquences génomiques et l'élaboration de modèles et d'hypothèses fonctionnelles (Cole and Saint Girons 1994).

En fait, il serait plus correct de parler de génomiques au pluriel car cette discipline comporte différents volets, chacun faisant appel à des expertises différentes : *la génomique structurale* qui cherche à cartographier la structure des génomes, *la génomique fonctionnelle* qui explore la fonction des gènes et des protéines en intégrant des études *in silico* et *in vitro* comme la transcriptomique ou la protéomique, *la génomique comparative* qui confronte l'organisation et la répartition des gènes entre plusieurs organismes pour en tirer des hypothèses fonctionnelles ou phylogénétiques et *la génomique environnementale* avec notamment l'étude des métagénomes et des microbiomes qui rassemblent les études des biodiversités géniques d'environnements comme les océans, les sols ou les intestins sans culture ou isolement préalable des organismes présents.

Les domaines d'application de la génomique sont vastes mais sans chercher l'exhaustivité, des retombées en santé humaine, animale et végétale peuvent être citées, avec la compréhension des maladies d'origine génétique, l'étude des interactions entre gènes et environnement, les constructions et le transfert de gènes (thérapie génique), la production de molécules pharmaceutiques. La génomique constitue ainsi un pôle d'innovation important en biotechnologie, en agroalimentaire, en agronomie et en sciences environnementales.

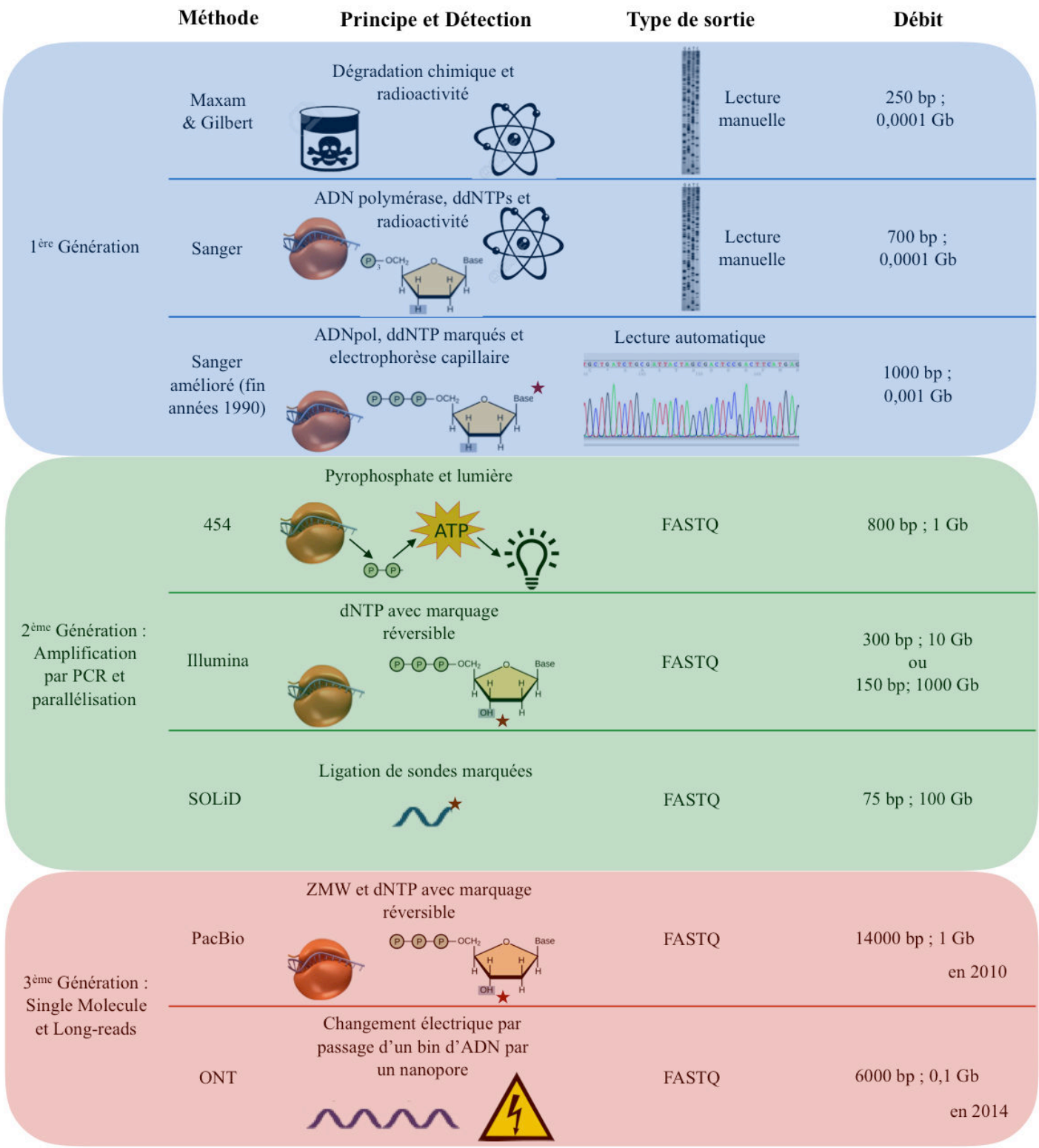


Figure 12. Schéma récapitulatif des méthodes de séquençage. Pour chaque génération, les principales méthodologies de séquençage sont présentées. Pour chacune, le principe et le mode de détection sont schématisés, avec le type de sortie et le débit. À partir de la deuxième génération, les sorties du séquenceur sont uniquement des fichiers informatiques avec les reads au format FASTQ. Le meilleur débit est présenté en dernier. Pour plus de détails et les sources, voir le texte.

Pictogrammes et images : http://emboj.embopress.org/content/17/17/5192.figures-only ; https://fr.dreamstime.com ; https://bio.libretexts.org

1.4. Séquençage de génomes

La base de toutes les branches de la génomique est le génome, il est donc nécessaire de connaître la séquence des nucléotides du génome pour pouvoir l'étudier. La détermination de l'ordre des nucléotides dans un génome se fait par un processus dit de séquençage **(Figure 12)**.

1.4.1. Les origines du séquençage

Le principal frein à l'étude de l'ADN a été la difficulté technique à le séquencer, contrairement au séquençage de protéines. Ce dernier débuta dès 1951-1955 avec la série de publications visant à déterminer de la séquence de l'insuline par F. Sanger (Sanger and Tuppy 1951b, 1951a; Sanger and Thompson 1953b, 1953a; Sanger et al. 1955). Avec la publication d'une méthode de séquençage protéique par P. Edman et G. Begg (Edman and Begg 1967) et la commercialisation quasi simultanée par la société Beckmann d'un appareil automatique, le séquençage des protéines s'est rapidement démocratisé.

Le premier problème de l'époque était le manque de disponibilité d'ADN sous forme moléculaire, les seules macromolécules de petite taille disponibles étant les virus. En se basant sur la méthode de dépurination acide décrite par Chargaff en 1952, V. Ling obtient en 1972 les premières séquences d'oligonucléotides à partir de phages Fd, F1 et PhiX (Ling 1972a, 1972b). En 1973, l'utilisation *in vitro* de systèmes enzymatiques puis du mécanisme de réplication permit la détermination de séquences de fragments des phages PhiX 174 (Galibert et al. 1974; Ziff et al. 1973) et f1 (Sanger et al. 1973).

Toujours sur le principe de la réplication, F. Sanger utilisa les propriétés de synthèse et d'hydrolyse du fragment de Klenow issu de l'hydrolyse de l'ADN polymérase I d'*Escherichia coli*. S'il manque un désoxyribonucleotide tri-phosphate (dNTP) dans le milieu, l'enzyme de Klenow hydrolyse l'ADN en cours de synthèse et ce, jusqu'à la position du prochain dNTP disponible. Sur ce principe, F. Sanger développa l'approche connue sous le nom de *méthode Plus-Moins* (*Plus-Minus method*) (Sanger and Coulson 1975) où après une courte synthèse en présence des quatre dNTP dont un marqué au ^{32}P, deux polymérisations supplémentaires sont effectuées : une dite *plus* où, dans quatre tubes, un seul dNTP est ajouté (provoquant l'arrêt de toutes les extensions à cette base) et une réaction dite *moins* où trois dNTPs sont utilisés (provoquant l'arrêt des séquences avant le nucléotide manquant). Les huit réactions sont

dénaturées, séparées en gel de polyacrylamide et les résultats lus sur autoradiographies. C'est avec cette méthode que la séquence complète du phage Phi X174 (5 375 nt) a été déterminée (Sanger et al. 1977b), faisant de ce virus, le premier génome ADN complètement déterminé.

1.4.2. Séquençage de première génération

F. Sanger améliora la technique *Plus-Moins* en remplaçant les dNTPs et l'activité exonucléase de la Klenow par des didésoxynucléotides (ddNTP), nucléotides identiques aux dérivés naturels ribonucléotides ou désoxyribonucléotides mais dépourvus de l'hydroxyle en position 3' (en plus du 2' des dNTPs). Ce groupement en 3' est requis pour la liaison phosphodiester qui relie les sucres et les phosphates du squelette de l'ADN, et donc l'incorporation d'un ddNTP dans une chaîne naissante d'ADN entraîne une terminaison prématurée de la synthèse. Les amorces utilisées pour la synthèse d'ADN sont marquées radioactivement pour rendre les fragments d'ADN polymérisés détectables lors de l'électrophorèse. L'utilisation d'un ddNTP à la fois, génère des chaînes de longueur variable par incorporation au hasard du ddNTP et permet de reconstituer une séquence d'environ 250 nt. Cette nouvelle méthode dites de *terminaison de chaîne* (Sanger et al. 1977a) et connue sous le nom de *séquençage Sanger*.

Quelques mois avant la publication de cette méthode, une équipe américaine publiait une méthode alternative où les synthèses enzymatiques étaient remplacées par des dégradations partielles d'un fragment d'ADN (marqué *in vitro* en position 5' par du ^{32}P) par des réactifs chimiques endommageant l'ADN préférentiellement sur les bases A, G, C ou T. La méthode utilise le diméthyl-sulfate qui coupe les purines (A/G), préférentiellement les A en condition acide, et l'hydrazine et la pipéridine qui rompent les pyrimidines (C/T), la coupure des T étant inhibée par l'addition de NaCl. Secondairement, les liaisons phosphate-sucre adjacentes sont cassées, libérant les ADN, tous marqués au ^{32}P en 5' et se terminant sur la base qui précède celle endommagée, identifiée selon le réactif chimique utilisé. Les positions des bases peuvent être lues sur gel de polyacrylamide (Maxam and Gilbert 1977). Cette méthode dite de *Maxam et Gilbert* a été pendant plusieurs années la méthode de référence. Elle s'appliquait particulièrement bien à l'ADN double brin alors que celle de Sanger nécessitait un ADN monobrin sur lequel s'hybridait une amorce. Or à l'exception de quelques phages, la très grande majorité des génomes sont double brin.

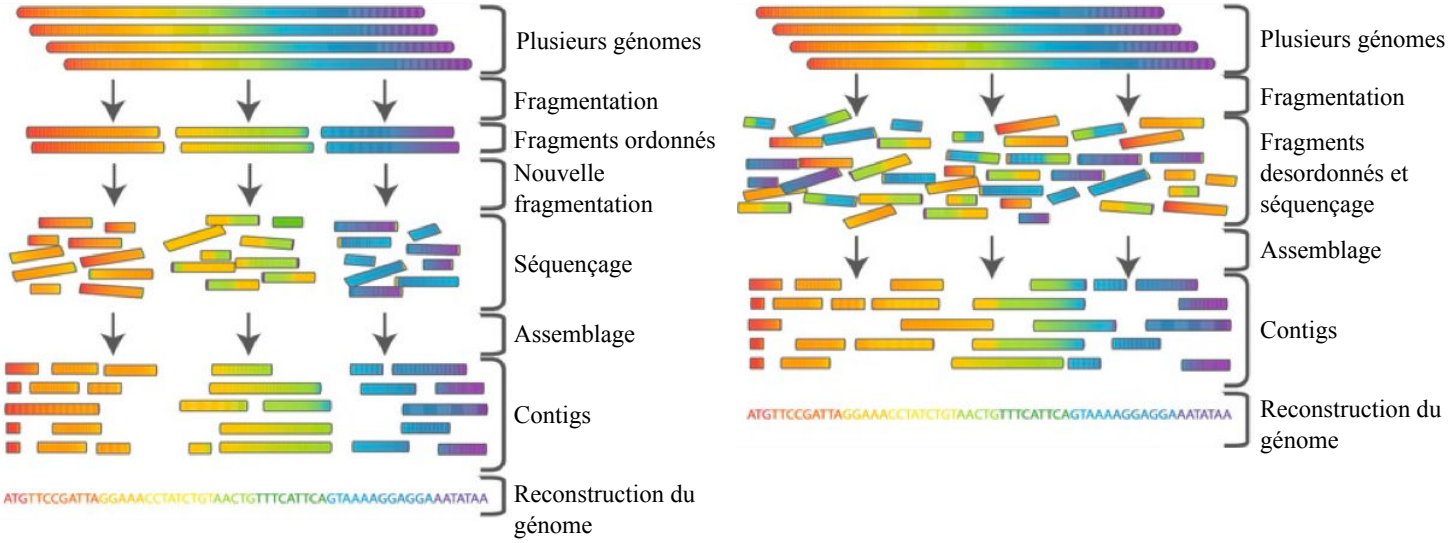


Figure 13. Stratégies de séquençage : ordonné et shotgun. À gauche, le séquençage dit ordonné avec une première fragmentation et ordonnancement des fragments. Ces grands fragments (ca 200 kb) sont à nouveau fragmentés puis séquencés. Après assemblage, les contigs peuvent être ordonnés par leur origine pour reconstruire le génome. À droite, le séquençage dit "Whole-genome shotgun", les génomes sont fragmentés puis séquencés et assemblés en contigs. Les chevauchements entre contigs sont utilisés pour reconstruire le génome.

Adapté de https://commons.wikimedia.org/wiki/File:Whole_genome_shotgun_sequencing_versus_Hierarchical_shotgun_sequencing.png
Téléchargées le 23 octobre 2017

Cependant, d'importantes innovations techniques ont finalement donné l'avantage à la méthode de Sanger. La première provient des travaux de J. Messing à la fin des années 1970, qui développa le clonage à partir du phage M13 (Messing et al. 1977), permettant de cloner n'importe quel fragment d'ADN double brin et de le récupérer sous forme monobrin. Le second grand changement fut le marquage radioactif des ddNTP (et non plus des amorces) puis l'utilisation de fluorochromes en remplacement de la radioactivité (Prober et al. 1987). Ces fluorochromes peuvent être excités par un laser et le signal peut être lu par un appareil optique ce qui permet de passer d'une lecture manuelle sur gel d'électrophorèse au traitement automatique des résultats. Ainsi, en 1986, le premier protocole de séquençage d'ADN partiellement automatisé est établi (L. M. Smith et al. 1986). La commercialisation du premier séquenceur automatique, l'ABI 370A d'Applied Biosystems, suit en 1987. Un autre progrès apparu en 1990 correspond à l'introduction de l'électrophorèse capillaire pour séparer les ADN polymérisés lors des réactions de séquençage ce qui permet d'augmenter le débit (Luckey et al. 1990). Toutefois, en dépit de ces innovations, la vitesse de séquençage de l'ADN reste encore modeste et les tailles des séquences lues limitées (500-1000 nt). Afin de pouvoir séquencer des génomes complets de plusieurs millions de bases, il est nécessaire de coupler le séquençage à des étapes de biologie moléculaire, qui permettent une fragmentation préalable de l'ADN puis son clonage afin de générer une banque génomique.

Une des premières stratégies pour séquencer un génome de taille importante fut l'utilisation d'un séquençage dit *ordonné*, *par ordonnancement hiérarchique* ou *clone-par-clone* **(Figure 13)**, consistant à créer des cartographies physiques des génomes (Kohara et al. 1987; H. O. Smith and Birnstiel 1976) permettant de classer des fragments génomiques clonés avant de les séquencer. La fragmentation peut être réalisé soit par des enzymes de restriction (digestion) soit par des ultrasons (sonication). Les grands fragments obtenus (ca 200 kb) sont insérés dans des vecteurs réplicatifs de type chromosomes bactériens artificiels (BAC). Une fois l'ordonnancement et le chevauchement de ces BAC réalisés, une banque dite *minimale*, assurant une couverture de 5 à 10 fois la longueur totale du génome étudié, est sélectionnée. Les fragments d'ADN de chacun des BAC retenus sont extraits puis fragmentés à leur tour (ca 5Kb) avant clonage dans des plasmides puis séquençage (Meyers et al. 2004). Cette approche présente l'avantage de faciliter l'assemblage post-séquençage grâce aux chevauchements des BAC et offre la possibilité de diviser le travail de séquençage chromosome par chromosome, entre plusieurs laboratoires. Toutefois, l'inconvénient majeur

Figure 14. Pages de couverture des publications du génome humain. À gauche, la première publication par le consortitum public dans Nature le 15 février 2001 et à droite, la publication du consortium privé du 16 février 2001 dans Science.

Extraits de http://www.nature.com/nature/journal/v409/n6822/index.html et http://science.sciencemag.org/content/291/5507/1304
Téléchargées le 23 octobre 2017

reste le clonage de certaines séquences notamment celle contenant des faibles complexités ou des répétitions.

La seconde procédure, connue sous le nom de *whole-genome shotgun* (WGS), néglige la phase préalable d'ordonnancement et ne nécessite pas de cartes physiques **(Figure 13)**. De nombreux fragments d'ADN, de tailles différentes sont générés au hasard, souvent de manière mécanique (par *shearing*) puis clonés et séquencés et le génome est *reconstruit* par un traitement bioinformatique post-séquençage d'assemblage. Cette méthode a été employée en 1995 pour séquencer le premier génome procaryote, celui d'une souche de laboratoire de *Haemophilus influenzae* (Fleischmann et al. 1995). Puis celui de la bactérie avec un des plus petits génomes connus, *Mycoplasma genitalium*, cette même année (Fraser et al. 1995). Cette méthode est plus rapide et moins coûteuse en temps et en argent que la précédente mais elle présente les mêmes limites puisque les séquences répétées de grande taille, fréquentes dans les génomes des mammifères et de plantes, ne peuvent être correctement assemblés.

Les objectifs de gain de temps et de réduction du coût, toujours plus ambitieux, ont été la force motrice du développement des techniques de séquençage d'ADN. Le projet phare étant le séquençage du génome humain car le volume d'ADN à séquencer (ca $3 \cdot 10^9$) était un vrai défi, nécessitant des améliorations expérimentales et bioinformatiques. Deux groupes concurrents ont relevé ce défi : le consortium international de séquençage du génome humain (*International Human Genome Sequencing Consortium* ou IHGSC), un groupement de laboratoires publiques coordonné à travers le monde, qui a utilisé l'approche *ordonnée* et Craig Venter qui a dirigé un groupe privé dont la stratégie a été le séquençage WGS et un protocole s'appuyant sur un puissant logiciel d'assemblage des séquences.

Le projet public, plus lent, a publié plusieurs versions préliminaires de haute qualité de parties du génome humain. La course se termina par une négociation afin de publier les résultats des génomes *drafts* simultanément même si l'équipe de C. Venter avait un net avantage. Le projet public publia dans Nature le 15 février 2001 (Lander et al. 2001; McPherson et al. 2001) et le projet privé dans Science le 16 février 2001 (Venter et al. 2001) **(Figure 14)**. Cependant, la séquence complète, de haute qualité ne fut délivrée qu'en octobre 2004 (International Human Genome Sequencing Consortium 2004).

La vitesse et le coût inférieur de la stratégie de *shotgun*, même pour un génome de grande taille et de grande complexité, ont conduit à l'adoption de cette méthodologie pour la

plupart des projets de séquençage ultérieurs, ce qui marqua le début de la fin des 30 ans d'hégémonie de la méthode Sanger.

1.4.3. Séquençage de deuxième génération

A partir de 2005, des améliorations technologiques ouvrent l'ère du séquençage haut débit (*high-throughput sequencing* ou HTS) plus couramment appelé NGS pour *next-generation sequencing* ou actuellement SGS pour *second-generation sequencing.* Plusieurs méthodes SGS ont été développées en même temps, par des sociétés privées concurrentes, à partir de chimies différentes. Toutefois, elles présentent en commun l'utilisation de la réaction en chaîne par polymérase (PCR) pour amplifier l'ADN avant son séquençage, éliminant l'étape de clonage, et le haut taux de parallélisation à l'origine de leur haut débit. Ce rendement permet de réduire considérablement le coût de séquençage au détriment de la taille des séquences lues, les *reads* (Dark 2013; Mardis 2008; Myllykangas et al. 2012).

Introduite en 2005, la première de ces méthodes est le *pyroséquençage* dit *454* de Roche (Margulies et al. 2005) qui adapte les travaux de P. Nyren (Nyren 1987) et de M. Ronaghi et collaborateurs (Ronaghi et al. 1996) sur l'utilisation des enzymes ATP-sulfurylase et luciférase pour séquencer l'ADN. Cette biochimie appelée *pyroséquençage* est la base du séquençage 454. Dans cette technique, l'ADN génomique est fragmenté mécaniquement par nébulisation, dénaturé (monocaténaire) et deux adaptateurs différents sont liés à chaque extrémité. L'un de ces adaptateurs est ensuite lié à une bille et la séquence est amplifiée par un processus appelé PCR en émulsion (emPCR), dans un mélange d'eau et d'huile permettant d'individualiser chaque bille comme autant de micro-réacteurs où plusieurs copies de l'ADN fixé sont produites. Les billes sont ensuite séparées sur une plaque composée de 100 000 puits de diamètre 30 µm ne pouvant accueillir qu'une bille chacun. C'est au sein de chacun de ces puits que va se réaliser la réaction de pyroséquençage, avec une amorce complémentaire au second adaptateur. Contrairement au séquençage Sanger, les nucléotides sont ajoutés de manière séquentielle (l'un après l'autre) et cyclique. Si le nucléotide ajouté correspond à celui attendu, il s'incorpore en libérant un pyrophosphate, transformé par l'ATP-sulfurylase en ATP et utilisé ensuite par la luciférase pour émettre un signal lumineux. Un traitement par une apyrase permet d'éliminer le surplus, puis le nucléotide suivant est incorporé et ainsi de suite. Le signal est enregistré par une caméra générant un pic dont la hauteur est fonction de l'intensité du signal lumineux, elle-même proportionnelle au nombre de nucléotides

incorporés en même temps. La séquence peut alors être déduite de la taille des pics obtenus. Avec cette technologie, le débit est considérable avec une capacité de presque 1 Gb soit 1 000 fois plus qu'en séquençage Sanger. Cependant, le taux d'erreur reste élevé notamment au niveau des homopolymères (répétitions d'un nucléotide), car l'intensité du signal n'est linéaire que pour huit nucléotides consécutifs (Margulies et al. 2005). Ainsi, en 2013, Roche annonce sa décision d'arrêter complètement le développement et le support des instruments utilisant cette technologie.

La deuxième méthode commercialisée est celle de Solexa, qui a changé ensuite son nom en Illumina. Dans cette méthodologie, une banque d'ADN double brin est généré par fractionnement aléatoire d'environ 200 pb puis des adaptateurs spécifiques sont ajoutés aux extrémités. Les fragments sont dénaturés et les ADN monocaténaires immobilisés à des endroits spécifiques de la cellule grâce à un des deux adaptateurs. Une *PCR en pont* (*bridge amplification*) est utilisée pour générer des groupes (*clusters*) de molécules identiques monocaténaires aux endroits spécifiques de la cellule. Le séquençage est dit *sequencing-by-synthesis* puisqu'il est fait au même temps que l'ADN polymérase incorpore des dNTP modifiés sur leur groupement hydroxyle en 3' et couplés chimiquement à un fluorophore différent pour chaque base. Une fois qu'un nucléotide est intégré, son hydroxyle modifié bloque l'ajout d'un autre nucléotide. Ensuite le milieu réactionnel est nettoyé pour supprimer les nucléotides non-incorporés et les fluorochromes des derniers nucléotides intégrés sont excités par un laser. Les émissions lumineuses résultantes sont détectées puis traduites en séquences. À la fin de chacun des cycles, une fois la mesure optique effectuée, le fluorochrome est clivé et un hydroxyle normal est restauré en 3', permettant à l'ADN polymérase d'intégrer le nucléotide suivant (Ju et al. 2006). Avec cette technologie, Illumina peut produire jusqu'à 600 Gb en 10 jours, soit un débit 50X supérieur au 454 de Roche, lui permettant de devenir le leader du marché (Heather and Chain 2016; van Dijk et al. 2014).

La troisième méthode a été le *Sequencing of Oligonucleotids by Ligation and Detection* (SOLiD) d'Applied Biosystems. La technologie repose sur une emPCR sur billes puis un séquençage, non pas par synthèse ou polymérisation comme précédemment, mais par ligation d'une sonde (un octamère) marquée par un fluorophore en 5' d'une amorce de séquençage universelle. Cette sonde est conçue pour s'hybrider spécifiquement à l'ADN sur ses deux premières positions et de manière non-spécifique pour le reste. Une fois hybridée dans la position correcte, elle est liée à l'amorce complémentaire à l'adaptateur. Une fois la détection optique effectuée, les trois derniers nucléotides de la sonde sont clivés avec le

fluorophore, en laissant une extrémité 5' disponible pour la ligature de la sonde suivante. Le séquençage se déroule en sondant deux bases et en sautant trois jusqu'à la fin du fragment. Le cycle est réinitialisé cinq fois en utilisant des amorces de tailles différentes de telle sorte qu'à la fin, chaque base a été interrogée deux fois (Valouev et al. 2008). Le rendement de cette technologie est de 3 Gb par run, mais le traitement informatique notamment le système de codage couleur rend l'analyse relativement complexe.

Finalement, la technologie IonTorrent de ThermoFisher utilise également l'emPCR pour générer les banques d'ADN matrice et les dNTP sont également incorporés de manière séquentielle. Cependant la méthode de détection diffère : lorsqu'un dNTP est incorporé, l'ADN polymérase libère un proton (ion H+) provoquant un changement de pH qui est détecté par une puce semi-conductrice (Rothberg et al. 2011). Ce fut la première technique qui n'utilisa plus l'optique pour lire la séquence d'ADN. Avec ce système, le débit est de jusqu'à 1 Gb en moins de 2 heures.

Une des améliorations de ces SGS est le séquençage appelé *paired-end* qui consiste à séquencer les deux extrémités d'un même fragment d'ADN matrice. Les deux *reads* sont indépendants, mais ils proviennent de la même molécule (ils sont appariés) et la taille du fragment qui les sépare, l'insert, peut être estimée (Risca and Greenleaf 2015).

Comme indiqué antérieurement, une des caractéristiques majeure de toutes ces technologies est l'amplification par PCR de l'ADN matrice préalable au séquençage. Cependant, des biais ou des erreurs peuvent être introduits lors de cette amplification (Kircher and Kelso 2010). D'où une troisième génération des techniques de séquençage qui tentera de remédier à ce problème.

1.4.4. Séquençage de troisième génération

La grande révolution des méthodes de troisième génération de séquençage (TGS) est la fin de l'utilisation de la PCR pour amplifier l'ADN matrice (techniques appelées *single-molecule*) et le séquençage de fragments de grande longueur (*long-read*).

La première méthode est l'HeliScope de Helicos. L'ADN est fragmenté (100 à 200 nt) et un adaptateur poly-A marqué avec un fluorochrome est lié à la matrice monocaténaire. Les fragments sont immobilisés par un oligomère de poly-T sur un support solide. Un laser excite

les fragments et une caméra détecte leur position. L'ajout séquentiel des dNTP marqués est détecté par la caméra, puis le fluorochome est clivé et le processus peut itérer jusqu'à la fin du fragment (Harris et al. 2008). Cet appareil n'est plus disponible à la vente et Helicos s'est converti en société de service de séquençage.

La technologie *Single-Molecule Real-Time* (SMRT) de Pacific Biosciences ou PacBio est le premier appareil capable de séquencer une molécule unique en temps réel. Il utilise une structure composé de cellules SMRT possédant chacune 75 000 nanostructures appelés Zero-Mode Waveguide (ZMW) et possédant chacune une ADN polymérase immobilisée au fond qui incorpore des dNTP couplés à un fluorophore et dont l'excitation laser entraîne son clivage, permettant un détection en permanence, en *temps réel* (Eid et al. 2009). Cette technologie permet d'obtenir des *reads* appelés *long-reads* de 5 à 20 Kb, ce qui est un progrès considérable par rapport aux techniques précédemment citées. Toutefois, un des problèmes majeurs de cette technologie est l'énorme taux d'erreur qui est proche des 15%, erreurs aléatoires qui peuvent être corrigées ensuite algorithmiquement (Rhoads and Au 2015).

La dernière technologie en date est celle d'Oxford Nanopore. La révolution de cette technologie est qu'elle n'utilise pas d'ADN polymérase et ne synthétise pas de brin complémentaire. La technique repose sur des pores protéiques (les nanopores) insérés dans une bicouche lipidique qui autorise le passage unidirectionnel d'un ADN simple brin. Ce passage produit un changement de courant électrique ou un potentiel qui diffère si le nanopore est obstrué par un A, un T, un G ou un C (Haque et al. 2013). Comparée aux techniques enzymatiques précédentes (polymérisation, ligation), le séquençage par nanopore nécessite moins de manipulations et un volume d'échantillon très faible. La technique est donc peu coûteuse mais a encore un fort taux d'imprécision. C'est toutefois une technologie en rapide évolution avec des améliorations constantes.

D'autres méthodologies comme la lecture directe des bases azotées par microscopie électronique (Bell et al. 2012) ont été proposées. Ces méthodes directes pour le séquençage sont parfois appelées de quatrième génération (Feng et al. 2015). Cependant l'introduction de nouvelles méthodes comme le *single-cell sequencing* (Gawad et al. 2016) qui pour le moment continue à dépendre d'une amplification (*Whole-Genome Amplification* ou WGA) et, surtout, le *in situ sequencing* (Nawy 2014) sont en train de changer, encore une fois, le paysage du séquençage. Même si pour le moment, ces méthodologies sont encore en développement et que quelque temps sera nécessaire pour leur mise en place et utilisation courante, les

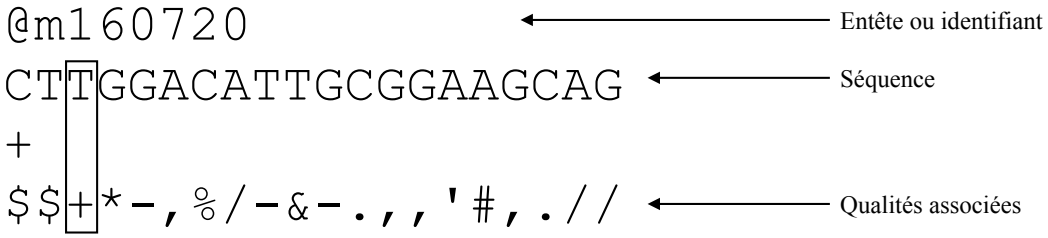


Figure 15. Format FastQ. Début d'un fichier au format fastq. L'information de chaque read est codée en quatre lignes, la première est un identifiant, la deuxième est la séquence des bases, la troisième commence toujours par un "+" et peut répéter ou non l'identifiant et la dernière ligne correspond aux qualités liées à chaque base, codées en Phred-33.

bénéfices pourraient être importants pour mieux décrire la diversité génomique dans une population de cellules (Gawad et al. 2016) et même le coupler à leur localisation spatiale dans un tissu ou biofilm (Ke et al. 2016; Nawy 2014).

1.5. Assemblage des reads de séquençage

L'assemblage est le processus de reconstruction du génome qui a été fragmenté pour être séquencé et qui doit être reconstitué pour permettre l'étude de sa structure et de son contenu. Pour bien comprendre cette étape, il faut connaitre les caractéristiques des reads et le concept d'alignement de séquences biologiques.

1.5.1. Reads de séquençage

Quelle que soit la technologie de séquençage utilisée, le signal détecté doit être interprété et converti en une des quatre bases des nucléotides de l'ADN. Ce procédé de conversion est appelé *base-calling* (Ewing et al. 1998). Ces données sont stockées dans un fichier de texte et à chaque base est associée une *qualité*, c'est-à-dire un nombre qui quantifie la confiance de l'exactitude du *base-calling*. Un *base-calling* ayant une probabilité d'être incorrect de 1/100 obtient par exemple une qualité de 20, noté Q20 (1/1000 une qualité de 30 noté Q30, etc) (Ewing and Green 1998). Pour simplifier l'association entre chaque base et sa qualité, cette dernière est encodée en un seul caractère. Généralement, le base-calling est réalisé par le séquenceur lui-même et produit des fichiers de sortie au format FastQ **(Figure 15)** combinant séquence et score (Cock et al. 2010). Le premier programme utilisant ce système d'assignation de scores à chaque base fut le programme Phred (Ewing et al. 1998; Ewing and Green 1998) et l'encodage de qualité le plus utilisé actuellement est le Phred-33 qui assigne à chaque valeur numérique un caractère ASCII commençant par le 33 (qui correspond au symbole !).

Avant de procéder à l'assemblage, la qualité des reads est évaluée afin d'éliminer ou de corriger les erreurs. La distribution des scores de qualité le long du read, sa taille et sa fréquence en oligonucléotides (k-mer) sont souvent utilisées ainsi que le taux de couverture, c'est à dire le nombre moyen de fois qu'un base a été lue (Yang et al. 2013). Par exemple, pour les SGS, notamment la technologie Illumina, la qualité des reads décroît aux extrémités

et les corrections vont donc essentiellement concerner une stratégie dite de *trimming* qui consiste à amputer ces extrémités afin d'améliorer la qualité des séquences (Del Fabbro et al. 2013). La deuxième stratégie consiste à corriger les erreurs des reads. Pour des reads dits courts (*short-reads*, <1 kb), deux stratégies sont utilisées : la méthode des k-mers et l'alignement multiple. Dans la première approche, les reads sont subdivisés en k-mers (sous-séquence ou *substring* de taille k) chevauchants et la distribution des k-mers est visualisée. Les k-mers rares sont identifiés et corrigés en faisant le plus petit nombre d'*éditions* (changement de bases) afin qu'il devienne un k-mer fréquent (Yang et al. 2013). La seconde approche de correction de short-reads correspond à l'utilisation d'alignements multiples des reads, qui sont alors corrigés (avec le plus petit nombre d'éditions) afin de se conformer au consensus de l'alignement.

Pour les reads dits longs (long-reads, >1 kb), il existe également deux stratégies : la correction dite *hybride* par des short-reads et l'auto-correction par alignement multiple. Le taux d'erreur des short-reads produit par les SGS étant largement moindre à celui des long-reads, ils peuvent servir pour corriger les long-reads. La correction consiste à aligner short-reads sur long reads, les *mapper*. Cet alignement (*mapping*) permet de corriger les *mismatch* (substitutions) et les *indels* (insertions et/ou délétions) (Au et al. 2012). La deuxième méthode n'utilise pas de références externes mais corrige les long-reads entre-eux par alignement multiple et production de consensus (Salmela et al. 2017).

1.5.2. Alignement de séquences

L'alignement de séquences est capital dans l'assemblage des génomes, que ce soit pour corriger les reads ou pour reconstruite le génome complet. Encore une fois, il existe plusieurs approches d'alignement, portés par différents algorithmes et programmes.

L'alignement de deux séquences (*pairwise alignment*) est un processus consistant à ajuster les bases (ou les acides aminés) de deux séquences biologiques afin d'atteindre le maximum d'identité entre elles. Cet ajustement intègre des bases identiques, des bases différentes appelés *mismatch* et des trous ou *gaps* correspondant soit à une insertion dans une des deux séquences, soit à une délétion dans l'autre (*indel*). L'alignement est exprimé en distance, qui peut être calculé à l'aide de deux métriques. Pour des séquences de même longueur, la distance de Hamming, qui correspond au nombre de substitutions minimum

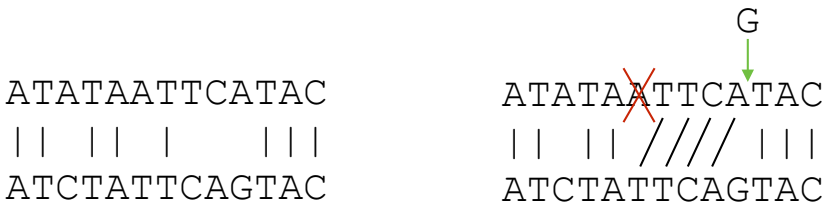


Figure 16. Comparaison des distances de Hamming et de Levenshtein. Le même exemple est utilisé dans les deux cas, les traits joignent des lettres identiques. À gauche avec la distance de Hamming, qui ne permet que des substitutions, nous observons 5 substitutions. À droite, avec la distance de Levenshtein qui permet des indels, nous observons 1 substitution, 1 délétion et 1 insertion.

Adapté de http://genoweb.univ-rennes1.fr/Serveur-GPO/outils/tutoriel/algo_distance.php

Figure 17. Comparaison entre un alignement global et un local. Les deux mêmes séquences sont alignées dans les deux cas, les traits joignent des lettres identiques. En haut, un alignement global "force" un alignement complet sur toute la séquence. En bas, un aligment local permet d'aligner uniquement des sous-parties de la séquence très similaires.

Adapté de http://rosalind.info/glossary/local-alignment/

nécessaire pour transformer une séquence en l'autre, est utilisée (Ristov 2016). Toutefois, comme il est assez rare que deux séquences biologiques soient de même longueur, une seconde métrique nommée distance de Levenshtein est utilisée (Yujian and Bo 2007). Cette dernière métrique permet également de calculer le nombre minimum d'opérations nécessaires pour transformer une chaîne de caractères en une autre, les opérations autorisées étant l'insertion, la suppression, la substitution d'un simple caractère ou la transposition de deux caractères adjacents **(Figure 16)**. À chacune de ces opérations unitaires est associé un coût prédéfini dans une matrice (pour les substitutions) ou par l'algorithme (pénalités d'ouverture de gap, de fermeture de gap, d'insertion…). L'ensemble est sommé et donne le score final de l'alignement.

Trois types d'alignement sont possibles : global, local et un type intermédiaire semi-global ou glocal (global-local). L'alignement global est adapté pour aligner deux séquences de taille similaire alors que l'approche locale permet d'identifier des similitudes régionales, courtes et éventuellement séparées de gaps **(Figure 17)**. Ces deux techniques ont leur algorithme optimal respectif : Needleman et Wunsch pour l'alignement global (Needleman and Wunsch 1970) et Smith et Waterman pour l'alignement local (T. F. Smith and Waterman 1981). Ces algorithmes sont très performants, mais lents ; leur vitesse étant inversement proportionnelle à la longueur des séquences. Afin de pouvoir aligner des séquences longues et de les comparer aux séquences présentes dans les bases de données biologiques, des approches heuristiques ont été adoptées pour l'alignement local : FASTP pour l'alignement de protéines (Lipman and Pearson 1985), FASTA pour l'alignement d'acides nucléiques (Pearson and Lipman 1988) et le très célèbre BLAST (Altschul et al. 1990). Ces algorithmes peuvent être considérés comme les premiers outils de mapping, les premiers *mappeurs* (Trapnell and Salzberg 2009). L'alignement hybride semi-global, qui ne pénalise pas les gaps au début et/ou la fin d'une ou des deux séquences est particulièrement adapté pour détecter les chevauchements et donc privilégié pour l'assemblage des reads de séquençage (Daily 2016).

1.5.3. Types et algorithmes d'assemblage de reads

L'hypothèse faite par tous les assembleurs est que deux séquences très similaires proviennent de la même position sur le génome (Nagarajan and Pop 2013). Deux stratégies d'assemblage existent pour reconstruire un génome : l'assemblage guidé par référence et l'assemblage *de novo*.

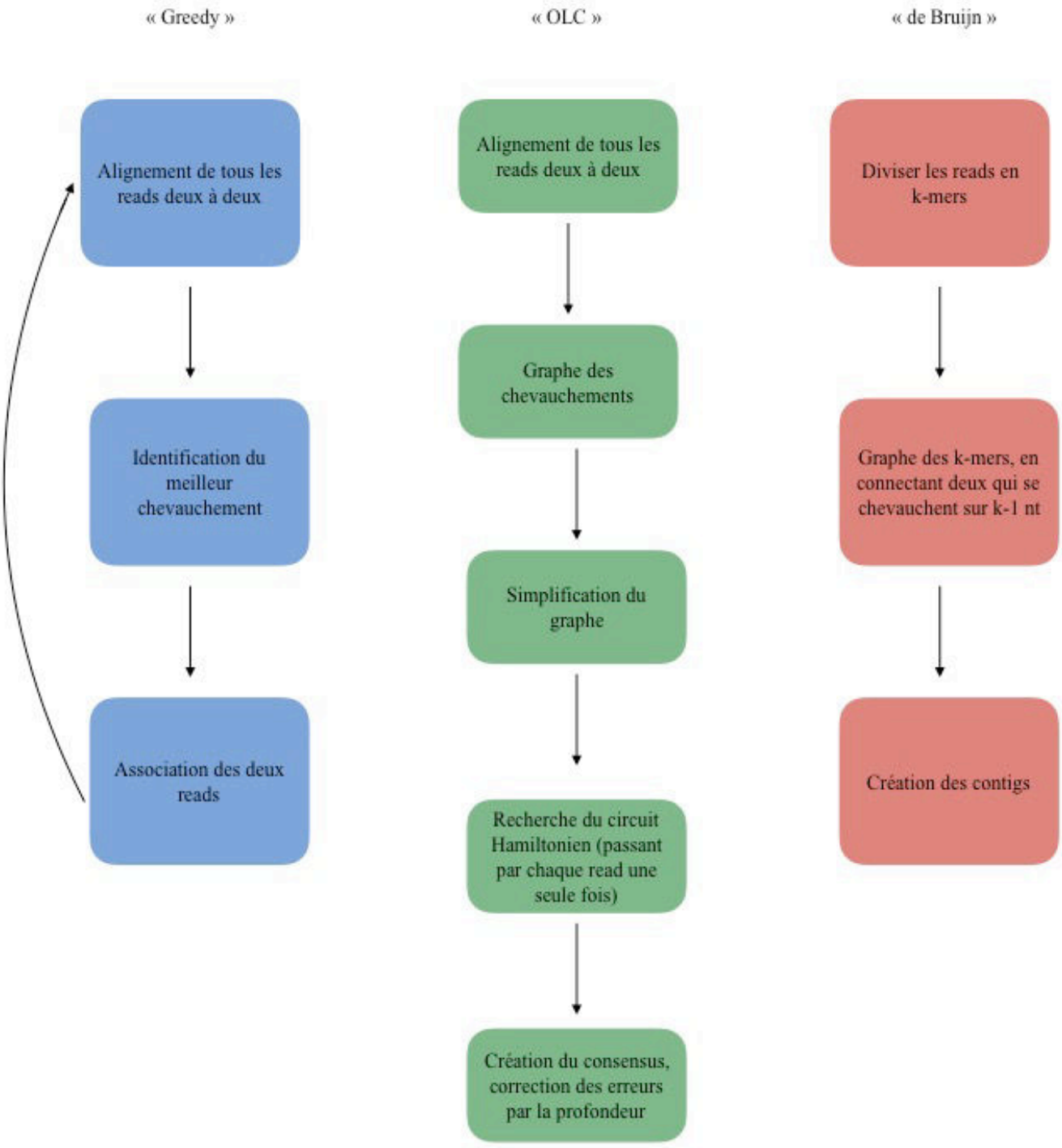


Figure 18. Étapes de l'assemblage *de novo* de génomes par trois méthodes courantes. L'approche Greedy est intuitive mais très lente pour des séquençages de profondeur importante du fait de sa nature itérative et du re-calcul des chevauchements des alignements à chaque fois. La méthode OLC fut la plus utilisée pour les séquençages Sanger (taille des reads compatible et profondeur adaptée), elle est gourmande en mémoire et lente comparée avec la dernière méthode. L'approche des graphes de de Bruijn est adaptée à des reads courts et à l'importante profondeur des séquençages de deuxième génération. Cependant, elle est très sensible aux répétitions génomiques de taille importante et produit des assemblages très fragmentés.

La première stratégie consiste à utiliser un génome de référence, prioritairement de la même espèce ou très proche du génome que l'on veut assembler. Ce type d'assemblage est simple et consiste à mapper les reads sur la référence puis à calculer une séquence consensus entre les reads mappés (Schneeberger et al. 2011). Cette stratégie n'est pas applicable que si un génome de référence existe et surtout si ces génomes proches sont faiblement réarrangés, connaissance indisponible *a priori*. Un autre problème est que certains génomes contiennent des erreurs d'assemblage et que leur utilisation comme référence entraîne fatalement leur propagation à d'autres génomes (Salzberg and Yorke 2005).

La seconde stratégie d'assemblage *de novo* n'utilise pas, comme son nom l'indique, de connaissances *a priori* (et notamment pas de génomes de référence), mais exploite la similitude entre reads pour trouver des chevauchements (*overlaps*) et les joindre, formant des séquences contiguës ou *contigs*. Plus un chevauchement est long, plus la confiance de l'association est importante. Pour augmenter la possibilité d'avoir des overlaps, la redondance de la lecture du génome, exprimé en taux de couverture, est utilisée. Trois types d'algorithmes sont utilisées pour joindre les reads : *Greedy*, *Overlap-Layout-Consensus* (OLC) et graphes de de Bruijn (Miller et al. 2010) **(Figure 18)**.

L'approche *Greedy* correspond à la méthode intuitive des premiers assembleurs développés et adopte une procédure itérative d'union de reads. Plus un chevauchement est long, plus la confiance de l'association des reads est forte, donc les reads avec le chevauchement le plus long sont associés. Puis itérativement des choix successifs pour identifier le read qui étend le mieux le contig résultant de l'étape précédente est associé, et ceci jusqu'à ce qu'il soit impossible d'associer plus de reads (Pop and Salzberg 2008). Parmi les assembleurs utilisant cette méthode, SSAKE (Warren et al. 2007), PE-Assembler (Ariyaratne and Sung 2011) ou encore GAPFiller (Nadalin et al. 2012), peuvent être cités comme exemples. Si la démarche *Greedy* est simple et permet d'obtenir rapidement des résultats satisfaisants, son efficacité nécessite une couverture élevée, rarement atteinte, ce qui conduit à la production de plus d'erreurs d'assemblage que celles observées avec les approches OLC ou de Bruijn (W. Zhang et al. 2011).

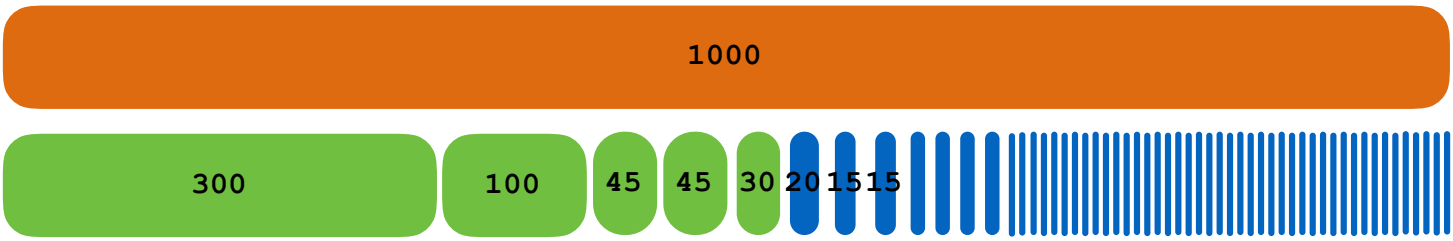
L'approche l'OLC comprend trois étapes, la première est l'*Overlap* qui aligne deux à deux tous les reads, un graphe de chevauchements est obtenu à la fin de cette étape. La seconde étape nommé *Layout* (*schéma*) analyse et simplifie ce graphe pour trouver le chemin qui parcourt chaque read une seule fois. Ce chemin est nommé circuit Hamiltonien, ou

Hamiltonian path. La dernière étape de *Consensus* fusionne les reads alignés en utilisant la profondeur du séquençage pour corriger les erreurs. Parmi les algorithmes d'assemblage utilisant la méthode OLC, Minimus (Sommer et al. 2007), Edena (Hernandez et al. 2008) ou encore SGA (Simpson and Durbin 2012) peuvent être cités. Les assembleurs OLC réalisent des assemblages de très bonne qualité mais nécessitent des reads assez longs (Pop 2009) ou une grande profondeur de couverture (Salzberg et al. 2012). Même si une variante appelée *String graph* plus performante et plus rapide (Myers 2005) a été développée, l'approche OLC a été progressivement abandonnée en raison de son incompatibilité avec les longueurs plutôt courtes des reads SGS. Toutefois, avec le développement des TGS, cette stratégie revient sur le devant de la scène.

Le troisième type d'algorithme utilise le principe des graphes de de Bruijn. Cette technique est la plus utilisée pour assembler des reads SGS. La première étape consiste à diviser les reads en k-mers chevauchants puis à construire un graphe de k-mers uniques et finalement d'utiliser le chemin de ce graphe pour extraire l'assemblage. L'avantage de cette approche, outre qu'elle correspond bien à des reads courts, est qu'aucun alignement deux à deux n'est nécessaire ce qui augmente la rapidité de résolution (Z. Li et al. 2012). L'inconvénient majeur est qu'en présence de répétitions et d'erreurs dans les données, l'approche est peu efficace. Une amélioration de la méthode a été proposé en 2001 (Pevzner et al. 2001) notamment en corrigeant les reads avant de construire le graphe mais ne résout pas le problème des répétitions. Un nombre important d'assembleurs sont basés sur cette méthode comme Velvet (Zerbino and Birney 2008), ALLPATHS (Butler et al. 2008) et SOAPdenovo (R. Li et al. 2010). Certains intègrent un traitement parallélisé pour accroître leur performance comme ABySS (Jackman et al. 2017) ou différentes valeurs de k comme IDBA (Peng et al. 2010).

1.5.4. Qualité de l'assemblage

Puisque différentes méthodes existent pour reconstruire *in silico* un génome fragmenté, il est important de pouvoir les comparer en évaluant la qualité de différents assemblages proposées. Le produit de l'assemblage est, en général, plusieurs séquences contiguës appelées *contigs* et l'évaluation des assembleurs consiste généralement à comparer ces contigs (Magoc et al. 2013; Salzberg et al. 2012). Les stratégies d'évaluations sont différentes selon qu'il existe ou non un génome de référence pour l'espèce séquencée. Sans



300 + 100 + 45 + 45 + 30 = 520 kb ; donc le N50 est 30 kb.

Figure 19. Définition et visualisation de la mesure "N50". Le N50 correspond à la taille du plus petit contig au dessus duquel 50% du génome est représenté.

Adapté de http://nekrut.github.io/BMMB554/post/topic13/
Visité le 23 octobre 2017

connaissance *a priori* (sans référence), l'assemblage sera évaluée sur un nombre minimum de contigs, de préférence de grande taille et comportant un minimum d'incertitude. En présence d'un génome de référence fiable, on évaluera plutôt le taux de couverture des contigs et leur agencement (*scaffolding*) (S. Bao et al. 2011).

Une méthode fréquemment évoquée dans les études pour évaluer la qualité des assemblages *de novo* est le N50. Cette mesure est définie comme la taille pour laquelle la longueur combinée de tous les contigs plus grands que cette valeur représente au moins 50 % de la somme des tailles de tous les contigs (Narzisi and Mishra 2011). Ainsi, 50 % de l'assemblage est composé de contigs d'une longueur ≥ N50 **(Figure 19)**. Plus un N50 est grand, meilleur serait l'assemblage. Cependant, cette métrique est imparfaite car elle dépend uniquement de la taille des contigs sans évaluer leur exactitude. Ainsi, un contig long mais erroné provoquera une inflation du N50.

Certains groupes ont essayé d'introduire la notion de NG50 qui permet de confronter différents assemblages du même génome en ayant recours à une taille attendue du génome au lieu de la taille totale de l'assemblage (Earl et al. 2011). Toutefois, comme le N50, cette métrique n'évalue pas les erreurs d'assemblage (*misassembly*). D'autres méthodes proposent de favoriser les assemblages avec le moins de contigs ou ceux avec la meilleure couverture mais finalement, aucune de ces approches n'est satisfaisante. En pratique, la qualité et la difficulté d'assemblage d'un génome dépendent moins de la stratégie employée que de ses caractéristiques intrinsèques, notamment le nombre de séquences répétées qu'il renferme, leurs tailles, leurs organisations et leurs localisations. Dans tous les cas de figure, la qualité d'un assemblage dépendra toujours directement de la qualité et de la longueur des reads et il sera extrêmement difficile voire impossible d'obtenir un assemblage de qualité si le séquençage contient un taux d'erreurs élevé ou trop de reads de petites tailles.

1.5.5. Niveau de finition ou complétude des génomes

Pour presque tous les projets de séquençage de type SGS (mais également TGS), la sortie des assembleurs est une liste de contigs qui constitue un assemblage partiel appelé *draft* (*brouillon*), généralement produit en quelques heures. Du fait de la fragmentation aléatoire du génome séquencé, l'ordre et l'orientation des contigs ne sont pas connus *a priori*. Ces assemblages draft sont de qualité inconnue et peuvent contenir des erreurs d'assemblage, des

séquences contaminantes comme des adaptateurs de séquençage et des séquences d'ADN d'autres organismes (L. Mallet et al. 2017; Mavromatis et al. 2012; Utturkar et al. 2017).

L'incapacité quasiment systématique d'obtenir un ensemble contigu unique reconstruisant le génome complet d'origine peut avoir plusieurs explications comme une couverture incomplète ou inégale de certaines régions du génome comme les répétitions, les homopolymères ou certaines zones riches en GC qui produisent des problèmes combinatoires (M. Kamada et al. 2014; Treangen and Salzberg 2011; Williams et al. 2013). À ceci s'ajoutent différentes erreurs possiblement introduites lors de la préparation des librairies par PCR (Kircher and Kelso 2010) et des erreurs de *base-calling*, principalement à la fin des reads pouvant produire des chemins sans issue dans les graphes de de Bruijn (des *tips*).

Pour compléter l'assemblage draft, une étape dite de finition (*finishing*) est nécessaire et consiste à (essayer de) ordonner les contigs et à combler les trous (*gaps*) qui les séparent. Cette étape longue et coûteuse nécessite souvent des connaissances *a priori* (Nagarajan et al. 2010). Souvent négligée, cette étape est cruciale pour obtenir un génome complet et de bonne qualité. Trois stratégies de finishing peuvent être suivies en utilisant un ou des génome(s) de référence(s) (E. Bao et al. 2014; Kolmogorov et al. 2014), des cartes génomiques (Madoui et al. 2016; Mariano et al. 2016) ou des long-reads (English et al. 2012; Madoui et al. 2015).

La première stratégie est très similaire à l'assemblage guidé par référence mais ce sont les contigs qui sont mappés pour être ordonnés et orientés. Le résultat est un *scaffold* (*échafaudage*) des contigs avec souvent des *assembly gaps* représentés par des séquences de N (E. Bao et al. 2014). Encore une fois, le problème se pose quand aucun génome de référence n'est disponible ou en présence de génomes fortement mosaïques car dans ce cas, soit le scaffolding ne sera pas possible, soit il pourrait être faux (Kolmogorov et al. 2014). De plus, même si le scaffolding est possible, il est préférable de le vérifier par des expériences complémentaires de *fermeture* des gaps (*gap closing*) qui utilisent soit le *primer-walking* (D. C. Richter et al. 2007), soit l'information des reads *paired-end* ou *mate-pair* (Dayarian et al. 2010). Le primer-walking consiste à concevoir des amorces aux extrémités des contigs et à amplifier puis séquencer les gaps itérativement jusqu'à leur fermeture (Sverdlov and Azhikina 2005). La deuxième stratégie correspond aux cartes physiques. Lorsqu'elles sont assez précises, elles peuvent permettre de détecter les éventuels erreurs d'assemblage (Mariano et al. 2016). Finalement, la dernière stratégie, les long-reads ou l'information de reads appariés avec

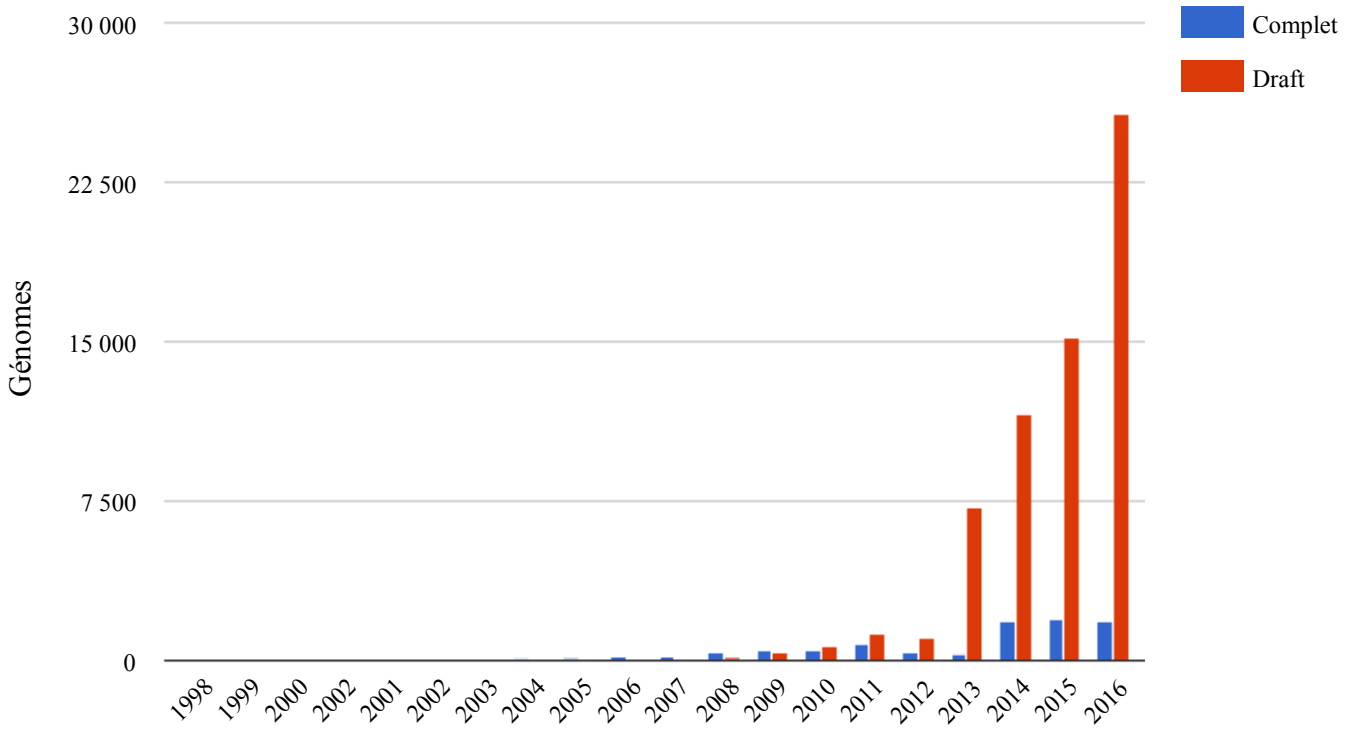


Figure 20. Statistiques des génomes draft et complets de Genomes OnLine Database (GOLD). Histogramme du nombre de génomes en permanent draft et complets par an. de 1998 à 2016.

Adapté de https://gold.jgi.doe.gov/statistics
Téléchargé le 23 octobre 2017

de longueurs d'inserts suffisantes peuvent être utilisés pour orienter et ordonner les contigs et fermer les gaps (Dayarian et al. 2010; English et al. 2012; Madoui et al. 2015).

Toutes ces étapes de finition impliquent un investissement économique supplémentaire, coûteux en temps et en ressources humaines. Certains auteurs estiment le coût du finishing à près de 95% du coût total du projet de séquençage et émettent l'hypothèse que de plus en plus de chercheurs considèrent que cette étape n'est pas rentable et préfèrent arrêter leurs efforts au stade de draft (Land et al. 2015) **(Figure 20)**.

1.6. Annotation du génome

L'annotation d'un génome correspond à l'identification des régions fonctionnelles biologiquement : les séquences codant des protéines (*Coding DNA Sequence* ou CDS) et les séquences transcrites mais non traduites comme les ARN ribosomiques (ARNr) qui constituent les ribosomes (avec des protéines ribosomiques), les ARN de tranfert (ARNt) qui sont des intermédiaires de la traduction qui permettent la lecture du code génétique et les petits ARN non-codants régulateurs (ARNnc, les riboswitch, etc). Afin de comprendre l'annotation d'un génome bactérien, quelques notions de l'organisation des gènes procaryotes sont nécessaires.

1.6.1. Gènes procaryotes

La transmission de l'information génétique se fait par l'intermédiaire de l'ADN dont le brin matrice sert de modèle pour créer soit un ARN directement fonctionnel (ARNr, ARNt et ARNnc), soit un ARN dit messager (ARNm). Cet ARNm permettra la synthèse protéique lors du mécanisme dit de traduction par les ribosomes, où chaque triplet de nucléotides (appelé codon), sera traduit en acide aminé. La protéine ou la chaîne polypeptidique sera synthétisée jusqu'au signal de terminaison de traduction.

Les gènes codant des protéines (ou *gènes codants*) possèdent des régions promotrices situées avant le codon d'initiation (*start codon*), généralement AUG (codant une méthionine) mais des start alternatifs existent, e.g. GUG (valine) et UUG (leucine). Les promoteurs sont responsables de l'initiation et de la régulation de la transcription de ces gènes qui se termine

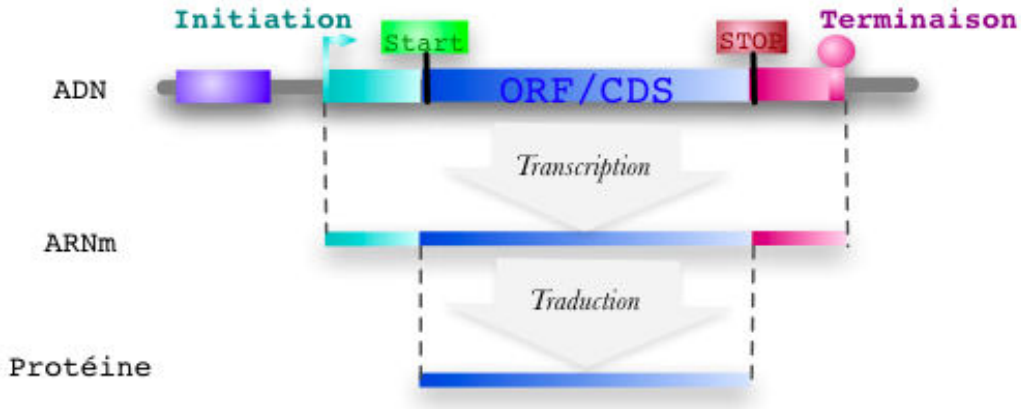


Figure 21. Structure des gènes codants des protéines chez les bactéries. Sur l'ADN, le rectangle mauve correspond aux régions régulatrices en amont du gène. Le gène est composé du bleu clair (le 5'UTR), le bleu (lORF ou CDS) et du rose (3'UTR). En général, l'annotation des gènes codants se limite à définir les CDS.

Extrait de http://sam.bioinfo.free.fr/animations_science/sejourdecouverte/tp/web.html
Téléchargé le 23 octobre 2017

sur un codon de terminaison (*stop codon* : UAA, UAG et UGA). Entre deux codons stop, un cadre ouvert de lecture (*Open Reading Frame* ou ORF) est défini. Le segment compris entre un codon start et un codon stop est appelé CDS et son identification est la détection de gènes, également dite *gene calling* ou *gene finding* (Shmatkov et al. 1999) **(Figure 21)**.

La détection de gènes chez les procaryotes est relativement plus simple que chez les eucaryotes. Les génomes procaryotes sont plus denses (ca 85% de la séquence est codante) et la structure des gènes est plus simple puisque ils ne possèdent pas d'introns et ne procèdent pas à l'épissage (Touchon and Rocha 2016). Certains gènes procaryotes, codés en opérons, peuvent être exprimés sous forme polycistronique, chacun avec ses codons start et stop, permettant leur traduction individuelle. En général, chaque opéron est précédé d'un opérateur composée d'un promoteur et d'un site de fixation du ribosome (*Ribosome Binding Site* ou RBS) également appelée séquence de Shine-Dalgarno (Shine and Dalgarno 1974).

Une fois les régions codantes positionnées (annotation syntaxique), le sens leur est donné en assignant des rôles biologiques (annotation fonctionnelle). Dans les faits, ces processus sont souvent simultanés. Une troisième étape d'annotation relationnelle peut ensuite venir compléter le tableau. Les définitions, caractéristiques et les stratégies de ces types d'annotation sont détaillées ci-après.

1.6.2. Annotation syntaxique ou structurale

L'objectif de l'annotation syntaxique est de définir les bornes des gènes (CDS et différents ARN), mais également des pseudogènes (gènes inactifs ou non-fonctionnels du génome) et séquences répétées et régulatrices (Tripp et al. 2015). Afin d'identifier ces régions d'intérêt, deux méthodes sont employées : les prédictions *ab initio* également appelées *de novo* (ou intrinsèques) et les méthodes par comparaison de séquences ou de motifs appelées extrinsèques. Ces méthodes sont complémentaires et souvent utilisées conjointement (Besemer and Borodovsky 2005). Dans les deux cas, les prédictions se font dans les six cadres de lecture de l'ADN (*reading frames*) qui correspondent à la division des séquences de nucléotides par triplets (consécutifs et non-chevauchants, sur les deux brins).

Les prédictions *ab initio* s'appliquent essentiellement aux CDS qui possèdent des caractéristiques précises : un codon start au début et un codon stop à la fin. La première étape consiste à détecter des ORF dans les six cadres de lecture. Puis les ORF de petite taille,

généralement de moins de 250 bp, et considérés comme des faux positifs sont éliminés (Hyatt et al. 2010). Pour les bactéries, la taille moyenne des protéines est d'environ 300 acides aminés (Brocchieri and Karlin 2005) ce qui correspond globalement à des CDS de *ca* 1 kb. Les prédictions peuvent ensuite être affinées en considérant le biais d'usage des codons propre à chaque espèce (Grantham et al. 1980). Un même acide aminé peut être codé par plusieurs codons synonymes (le code génétique est dégénéré), cependant, statistiquement, les fréquences d'usage des codons synonymes au sein d'une espèce ne sont pas homogènes. Ces préférences de certains codons par rapport aux autres introduit un biais qui peut être corrélé à l'abondance relative des ARNt (Bulmer 1987). Des modèles mathématiques peuvent être crées pour distinguer, grâce à ce biais d'usage des codons, les régions codantes des non-codantes (Besemer and Borodovsky 2005) et, encore une fois, éliminer des faux positifs.

Concernant les ARN directement fonctionnels sans traduction, leur structure n'est pas définie aussi bien que pour les CDS et donc les méthodes d'identification reposent sur l'identification de motifs conservés (une séquence de nucléotides présente dans plusieurs ARN de différentes espèces) et de prédiction de leur structure secondaire (le repliement de la chaîne monocaténaire d'ARN qui la stabilise et lui permet sa fonction). Deux modèles mathématiques sont utilisés pour identifier ses molécules : les modèles de Markov cachés (*Hidden Markov Models* ou HMM) pour les motifs conservés et le modèle de covariance qui prend en compte les contraintes de structure secondaire pour identifier ces ARN.

Comme déjà décrit, la distribution des nucléotides le long du génome et des gènes n'est pas aléatoire. Par exemple, l'alignement de différents gènes codant l'ARNr 16S permet d'identifier des régions très similaires et des régions beaucoup plus variables. Grâce aux observations faites sur des alignements multiples, le *patron* (*pattern*) des régions *constantes* peut être décrit avec une matrice de probabilité dite matrice de position pondérée (*matrice de score-position*, *positional weight matrix* ou PWM). À chaque position dans la séquence conservée, la base observée est associée à sa fréquence. Par exemple, le premier nucléotide de la région conservée est 7% des fois A, 10% C, 7% G et 76% T. Ainsi, toutes les positions du motif conservé sont décrites. Pour décrire le motif et l'enchaînement des nucléotides, la séquence peut être modélisée comme une chaîne de Markov, c'est à dire une suite d'états où l'état présent ne dépend que de l'état précédent. Elle est dite cachée parce que seul le résultat (ici le nucléotide) est observé, les mutations sous-jacentes (substitution, indel) pour arriver à cet état sont *cachées*. Ce modèle est donc un HMM. Pour décrire ces modèles statistiques, quatre informations sont nécessaires : i. l'alphabet (pour les acides nucléiques c'est les quatre

bases ; pour les protéines, les 20 acides aminés), ii. le nombre d'états du modèle (la taille du motif dans l'exemple cité), iii. les probabilités d'émission (la fréquence de chaque lettre de l'alphabet dans chaque état ; pour chaque position, la somme étant égale à un ou 100%) et iv. la probabilité de transition pour passer d'un état à un autre (ou de rester dans le même état) (Eddy 2004; J. Wu and Xie 2010).

Les HMM ne sont pas des bons modèles pour décrire la structure secondaire des ARN car ils ne peuvent pas prendre en compte les contraintes d'appariement des nucléotides à distance (Eddy 2004). Pour remédier à cela, un autre modèle dit *de covariance* (*Covariance Model* ou CM) a été développé et consiste à une généralisation des HMM qui est capable de prendre en compte la covariance due à l'appariement à distance des bases pour créer un repliement de l'ARN et sa structure secondaire. Ils permettent d'identifier les différentes combinaisons de conservation de la séquence primaire et de les corréler avec les autres régions conservées à d'autres positions dans la séquence (Eddy and Durbin 1994; Yao et al. 2006). Ces deux modèles sont utilisés pour détecter les ARN dans le génome bactérien (Kolbe and Eddy 2011; Lagesen et al. 2007; Laslett and Canback 2004), mais peuvent également être utiles pour l'annotation de CDS (Borodovsky et al. 1995).

Une fois les régions d'intérêt définies, les fonctions de chacun de ces gènes doit être assignée. Afin de comprendre les stratégies utilisées pour cela, certains concepts seront définis à continuation.

1.6.3. Homologie de séquences et annotation fonctionnelle

L'objectif de l'annotation fonctionnelle est de prédire les fonctions des produits des gènes identifiés lors de l'annotation syntaxique. De manière générale, l'annotation fonctionnelle peut avoir au moins deux sources d'information : la validation expérimentale et le transfert par homologie de séquence.

La validation expérimentale nécessite, comme son nom l'indique, des expériences menées *in vitro* ou *in vivo* permettant de découvrir biologiquement la fonction du gène. Ces travaux sollicitent plusieurs protocoles de biologie moléculaire comme des expériences de mutations par interruption de gènes (par des transposons, et bien d'autres techniques) et/ou de complémentation. Ces résultats d'études génétiques, font souvent l'objet de publications scientifiques qui peuvent servir de référence lors de l’annotation par des biocurateurs.

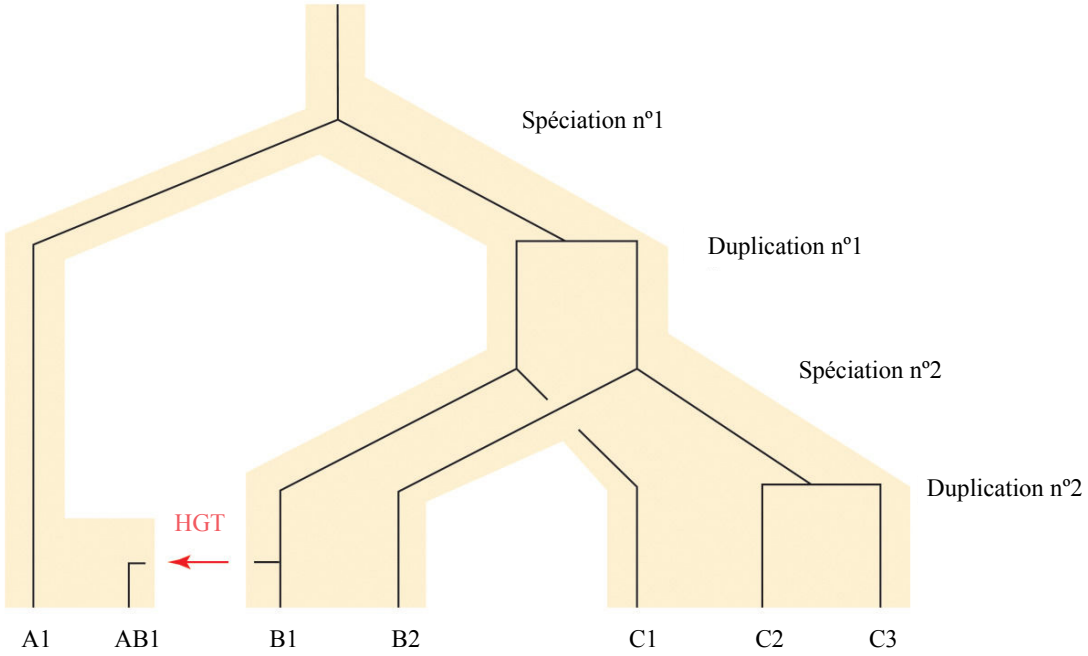


Figure 22. Homologie, Orthologie, Paralogie et Xénologie. Tous les gènes des espèces A, B et C sont homologues puisqu'ils proviennent tous d'un ancêtre commun. L'acquisition du gène AB1 est faite par transfert horizontal (HGT), ce gène est donc xénologue des autres. Les événements de spéciation sont à l'origine de gènes orthologues, ainsi B1 est orthologue de C1. Enfin, les phénomènes de duplication sont à l'origine de gènes paralogues, ainsi C2 et C3 sont paralogues entre eux.

Adapté de Fitch 2000.
Téléchargée le 23 octobre 2017

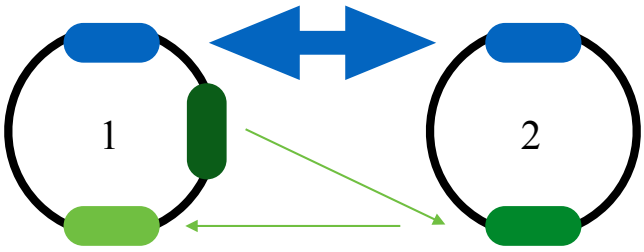


Figure 23. Best Bidirectional Hits (BBH) ou Reciprocal Best Hit. Deux situations sont décrites dans le schéma. Les gènes en vert, où le meilleur résultat du gène vert foncé sur le génome 1 correspond au gène vert du génome 2. Cependant, le meilleur alignement du gène vert du génome 2 est le gène vert clair du gènome 1. Ces gènes ne sont donc pas des BBH. Dans le cas des gènes en bleu, il s'agit bien de BBH.

Adapté de https://www.slideshare.net/melvinzhang/ortholog-assignment
Visité le 23 octobre 2017

L'autre stratégie, qui est de loin la plus utilisée pour assigner une fonction à un gène, est le transfert de fonction *in silico* par homologie de séquence. Le terme d'*homologie* a été utilisé, en anatomie, pour décrire un même organe ayant, chez des espèces différentes d'animaux, la même fonction. À cette définition s'ajoutera un critère d'évolution pour introduire dans la définition d'homologie l'origine ancestrale commune (Stormo 2009). Au concept d'homologie, s'oppose celui d'analogie, qui décrit des organes avec des fonctions similaires mais qui ne sont pas de relation évolutive. L'exemple typique étant les ailes d'insectes et celles de mammifères qui sont analogues, tandis que les ailes de mammifères sont homologues des pattes d'autres mammifères.

Cette relation d'homologie est précisée au niveau génétique par W. Fitch en utilisant les termes d'évolution divergente à partir d'un ancêtre commun pour caractériser des gènes homologues et d'évolution convergente à partir de gènes non apparentés pour des gènes analogues (Fitch 1970). L'homologie peut être divisée en trois types : l'*orthologie* qui correspond à des gènes divergents à la suite d'un événement de spéciation; la *paralogie* où les gènes divergent suite à un événement de duplication au sein de la même espèce et finalement, la *xénologie* où l'histoire évolutive du gène implique un transfert inter-espèce (*transfert horizontal*, *horizontal gene transfer* ou HGT) **(Figure 22)**. Il est recommandé d'utiliser ces définitions précises chaque fois que le type de relation est connu ou supposé, et de réserver le terme d'*homologie* quand la distinction n'est pas possible (Fitch 2000; Koonin 2005). La *conjecture de l'orthologie* (*ortholog conjecture*) établit que les gènes orthologues ont la *même* fonction, ou plutôt des fonctions équivalentes, dans les différents organismes. Cette hypothèse est centrale pour l'annotation fonctionnelle des génomes (Wolf and Koonin 2012).

Pour déterminer la relation d'orthologie entre deux séquences, une des premières méthodes utilisée est de réaliser des alignements de séquences dits *réciproques*. Cette technique, des *meilleurs alignements réciproques* (*Best Bidirectional Hits* ou BBH) est basée sur l'hypothèse que les séquences de gènes orthologues (et celles des protéines qu'ils codent) ont un pourcentage de similarité plus important entre elles qu'avec toute autre séquence de leur propre génome et toute autre séquence de l'autre génome (Wolf and Koonin 2012) **(Figure 23)**.

La grande quantité de génomes séquencés et leur mise à disposition dans des bases de données internationales, permettent de comparer les séquences à annoter avec un ensemble de séquences portant des informations fonctionnelles. Les bases de données peuvent être

généralistes comme GenBank, UniProtKB (C. H. Wu et al. 2006), Pfam (Finn et al. 2016; Sonnhammer et al. 1997) ou Rfam (Griffiths-Jones et al. 2003), certaines étant biocurées comme le sous-ensemble Swiss-Prot qui fait partie de UniProtKB (Boutet et al. 2016). D'autre part, les bases de données peuvent être dédiées : i) à une espèce particulière comme EcoCyc pour *Escherichia coli* K-12 (Keseler et al. 2017) ou SubtiWiki pour *Bacillus subtilis* (Michna et al. 2016) ; ii) à des groupes d'organismes proches comme CyanoBase pour les cyanobactéries ou RhizoBase pour les rhizobactéries (Fujisawa et al. 2014) ; iii) à des groupes de protéines comme The Histone Database pour les histones (Marino-Ramirez et al. 2011) ; ou iv) à des voies métaboliques particulières comme REPAIRtoire pour la réparation de l'ADN (Milanowska et al. 2011) ou plus générales comme MetaCyc pour les enzymes du métabolisme des petites molécules (Caspi et al. 2014) ou l'ensemble des voies métaboliques comme KEGG (Kanehisa et al. 2017).

La stratégie générale pour utiliser ces bases de données est l'utilisation d'algorithmes d'alignement local, principalement BLAST (Pearson 2013) et ses adaptations pour des séquences moins conservées comme PSI-BLAST (Altschul et al. 1997). La base de données est interrogée pour trouver les séquences donnant les meilleurs scores (pour chaque séquence requête ou *query*) et l'annotation de la séquence la plus proche sera transférée (Koestler et al. 2010; Sasson et al. 2006). Cette méthodologie est toutefois source d'erreur comme par exemple des transferts de fonction alors que seulement une partie de la séquence est conservé (Sasson et al. 2006), des erreurs d'annotation dans la base de données non biocurées qui seront alors propagées impunément (Jones et al. 2007; Schnoes et al. 2009), des seuils de pourcentage de similitude pour transférer des fonctions différentes selon le groupe de protéines (e.g. des protéines très conservées comme les ADN polymérases *vs* des protéines plus variantes comme les récepteurs de surface des bactéries pathogènes), la faible spécificité/sensibilité de la méthode d'alignement local par rapport à des alignements dits supervisés (qui utilisent des modèles spécifiques aux groupes de protéines, e.g. HMM ou CV pour des domaines fonctionnels) (Borodovsky et al. 1995) et le problème de détecter des vrais orthologues et non des paralogues avec des fonctions qui peuvent être assez divergentes (Kuzniar et al. 2008; Sasson et al. 2006).

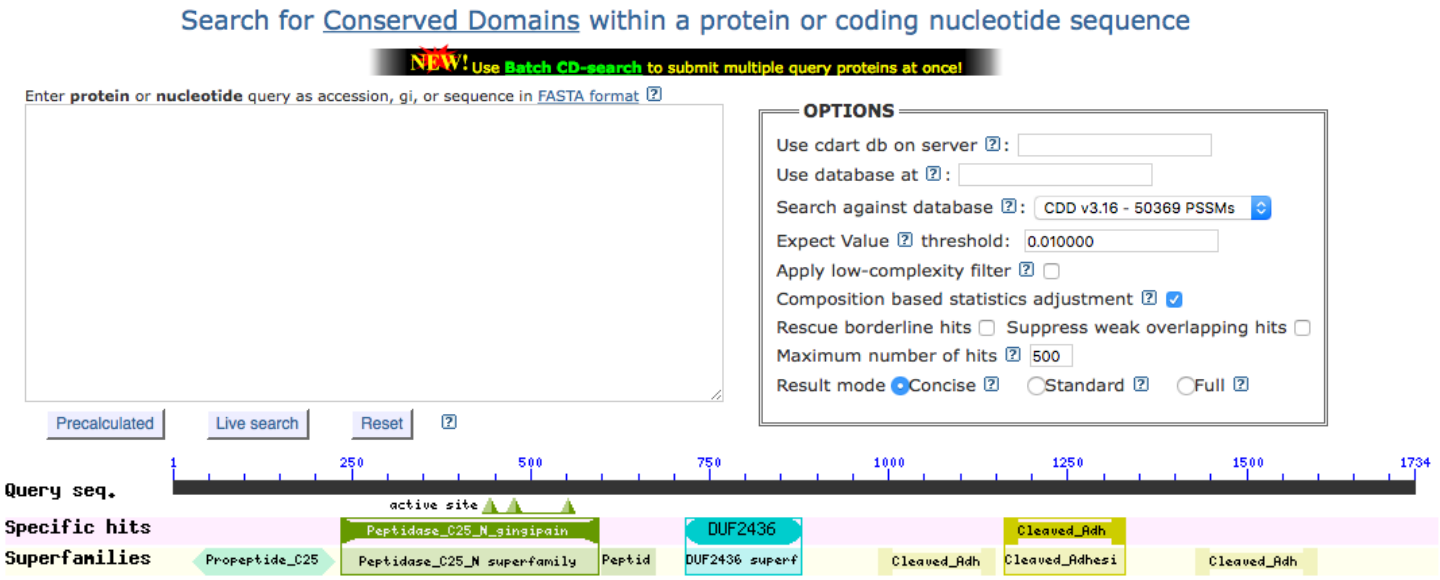
Une autre méthodologie, dite de *regroupement* (*clustering*) tente d'identifier de manière générale des groupes de gènes orthologues entre plusieurs génomes. La méthode des COGs (*Cluster of Orthologous Groups of proteins*) et sa base de données, reflètent la possibilité des relations d'orthologie un à un, un à plusieurs et plusieurs à plusieurs (Tatusov

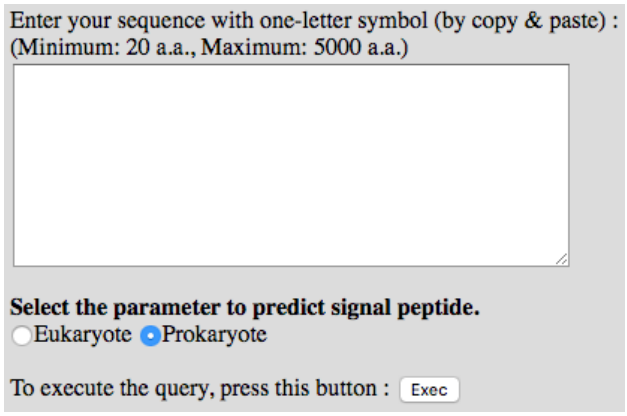
et al. 2000). Plusieurs méthodes utilisent cette notion de groupes de gènes orthologues qui permet d'étendre la réciprocité à l'ensemble des relations d'homologie, en utilisant les arbres phylogénétiques pour vérifier la concordance avec les arbres d'espèces et supprimer des paralogues dans les groupes en divisant les groupements (Altenhoff et al. 2016). La base de données eggNOG (*evolutionary genealogy of genes : Non-supervised Orthologous Groups*) peut être citée comme un exemple récent (Huerta-Cepas et al. 2017).

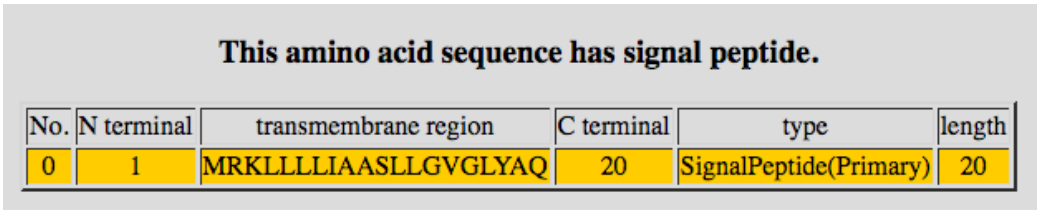
Une aide à l'identification des relations d'orthologie est l'identification de la *synténie*. L'introduction du terme de synténie est attribué au généticien britannique J. Renwick qui définit le terme comme la présence de deux *loci* sur le même chromosome (Renwick 1971). Cette définition est utile seulement pour des organismes avec plusieurs chromosomes, mais pas pour la vaste majorité des bactéries. Le sens premier du terme est élargi, pour désigner des *loci* d'organismes différents mais localisés dans une région chromosomique homologue (McCouch 2001; Passarge et al. 1999). L'ordre des gènes dans les génomes procaryotes n'est pas conservé à grande échelle (Mushegian and Koonin 1996) et le maintien de l'ordre de certains des gènes serait le résultat de sélection par épistasie (interaction des produits géniques) (Nei 2003) mais également par co-expression et co-régulation (Lemoine et al. 2007). L'ordre des gènes dans ces régions homologues, un cas particulier de synténie, où les orientations, proximité et position sont conservées est appelé colinéarité (H. Tang et al. 2008). Cette conservation de la synténie sert à vérifier les annotations existantes et les prédictions de fonction (Sridhar and Rafi 2007).

1.6.4. Annoter sans attribuer de fonction

Le haut débit de séquençage et les coûts plus faibles des projets de génomique s'accompagnent d'une croissance exponentielle des bases de données. Cependant, la quantité de données ne signifie pas nécessairement une connaissance plus approfondie des capacités de codage des génomes. Plus d'un tiers des CDS prédites sont annotées comme *protéines hypothétiques* ce qui veut dire que la communauté scientifique n'a pas d'autre information sur ces protéines que leur conservation de séquences. Ces protéines hypothétiques ne sont pas des artefacts et peuvent être présentes chez plusieurs organismes. Certaines ont été validées par des expériences de transcriptomique, protéomique ou lors d'études d'essentialité des gènes (Jaroszewski et al. 2009). La principale raison du manque de connaissance de ces CDS







This amino acid sequence has signal peptide.

No.	N terminal	transmembrane region	C terminal	type	length
0	1	MRKLLLLIAASLLGVGLYAQ	20	SignalPeptide(Primary)	20

Figure 24. Exemple de recherche de domaines protéiques. En haut, utilisant l'outil du NCBI Conserved Domain Search sur une protéine de 1734 acides aminés, cette séquence possède trois domaines : Peptidase_C25_N_gingipain, DUF2436 et Cleaved_Adh. Le premier correspond au domaine N-terminal des gingipaïnes (un sous-groupe de la famille des peptidases C25), le deuxième comme son nom l'indique est un domaine de fonction inconnue et le troisième est un domaine répété en tandem de régions d'hémagglutinine/ adhesine. En bas, utilisant l'outil SOSUI signal, une prédiction de signal peptide est faite. Ainsi, avec ces deux outils, nous pouvons savoir qu'il s'agit d'une peptidase qui n'est probablement pas cytoplasmique.

Sites utilisés : https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi et http://harrier.nagahama-i-bio.ac.jp/sosui/sosuisignal/sosuisignal_submit.html
Visités le 23 octobre 2017

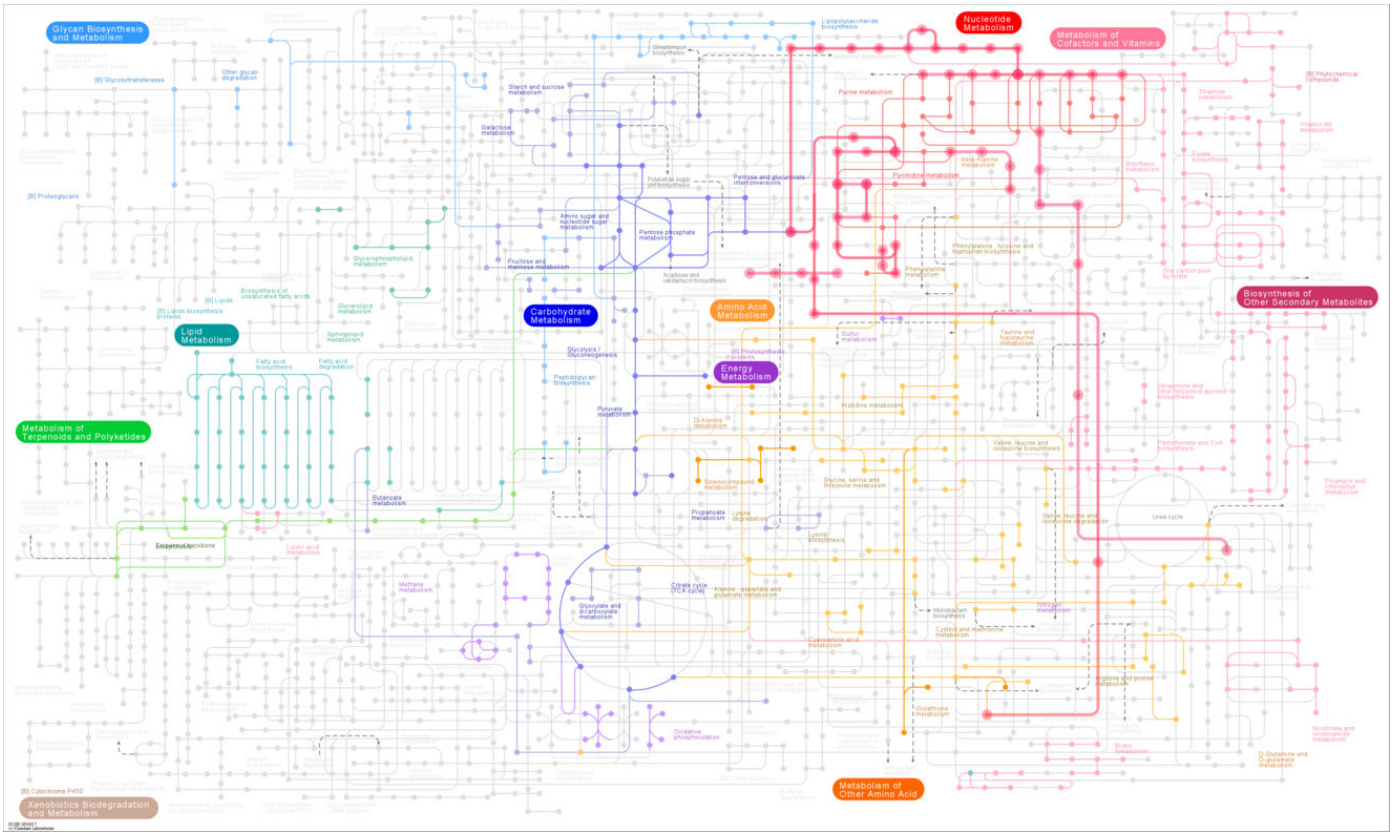
hypothétiques est qu'elles n'ont pas fait l'objet d'études traditionnelles de biochimie et biologie moléculaire (Roberts 2004).

Les protéines ont des domaines, qui correspondent à des unités évolutives, et qui peuvent être dupliqués et/ou recombinés. Les protéines de petite taille ont généralement un seul domaine et celles de taille plus importante peuvent avoir plusieurs domaines différents (Chothia et al. 2003). Ces domaines sont également des unités avec une structure et constituent les sous-unités fonctionnelles des protéines (C. Vogel et al. 2005) **(Figure 24)**.

Malheureusement, près d'un quart de tous les domaines protéiques décrits actuellement sont annotés comme des *domaines de fonction inconnue* (*Domains of Unknown Functions* ou DUF) (Bateman et al. 2010; Mudgal et al. 2015). Selon la base de données utilisée, des équivalents existent où la famille de domaines/protéines n'est pas caractérisée et possède uniquement un numéro, par exemple des *Fellowship for Interpretation of Genomes* (FIG) pour les FIGfams (Meyer et al. 2009) ou de COG (Galperin et al. 2015). Pour ces domaines DUFs, les biocurateurs sont incapables de leur associer toute information fonctionnelle en raison d'un manque dans la littérature scientifique du moment (Bateman et al. 2010). Ainsi, lors du processus d'annotation fonctionnelle, certaines CDS peuvent recevoir des annotations de protéines avec des domaines conservés de fonction inconnue. Cependant, la conservation évolutive et les expériences de mutation caractérisant les *gènes essentiels* d'un organisme suggèrent des rôles importants de certaines de ces CDS (Goodacre et al. 2013).

En plus de ces domaines DUFs, certains domaines portent des annotations fonctionnelles peu descriptives du processus biologique auquel participe la protéine et ne décrivent que des propriétés structurales ou de capacité de liaison à d'autres molécules (Bateman et al. 2010). Des domaines de type *nucleotide-binding P-loop motif*, *RNA-binding domain* ou *Helix-turn-Helix motif* peuvent être cités en exemple. Si seulement les DUFs et autres équivalents sont considérés comme non caractérisés, la sous-estimation des familles de domaines/protéines dans les bases de données peut être importante (Bateman et al. 2010). Ces annotations sont transférés par le biocurateur, mais restent peu informatives.

Pour les CDS recrutant des domaines de type DUF ou ne recrutant aucun domaine (*orphelines*), une désanonymation partielle est également possible en prédisant leurs localisations cellulaires. En effet, certains signaux de localisation, sont aisément identifiables dans les séquences d'acides aminés et peuvent être très utiles pour orienter l'annotation de séquences. La prédiction des domaines transmembranaires (hélices alpha), des signaux de



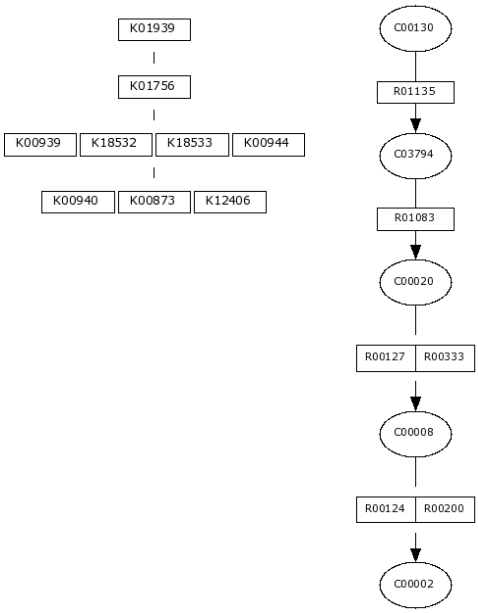


Figure 25. Exemples d'annotations relationnelles du KEGG. Toutes les voies métaboliques de *Porphyromonas gingivalis* sont schématisées, mises en évidence en rose foncé, les voies métaboliques des purines et pyrimidines. En bas, schéma générique de la biosynthèse des ribonucléotides d'adénine (ATP et ADP).

Site utilisé : http://www.genome.jp/kegg/pathway.html
Visité le 23 octobre 2017

sécrétion (signaux d'adressage N-terminaux de type signal peptide) ou encore des caractéristiques de protéines de la membrane externe (tonneaux bêta) peuvent être cités à titre d'exemple. Ces prédictions n'indiquent pas vraiment des fonctions, mais décrivent le lieu de leurs actions. Par exemple, les protéines membranaires, qui représentent près de 25% des gènes codants se retrouvent impliquées dans des processus comme le transport de molécules ou la signalisation entre cellules (Cuthbertson et al. 2005; Tusnady et al. 2005).

1.6.5. Annotation relationnelle

Ce dernier type d'annotation a pour but de donner une vision intégrée des relations entre les produits des gènes. Elle établit les voies métaboliques avec les enzymes et autres protéines qui interviennent et les réseaux de régulation. Une intégration de résultats expérimentaux, des descriptions mathématiques des processus biochimiques et la simulation de ceux-ci par ordinateur sont nécessaires. Le résultat est une modélisation des voies métaboliques qui est capable de décrire le processus (Endler et al. 2009). Un des premiers efforts de construction des réseaux de gènes cohérents capable de signaler des inconsistances dans l'annotation fonctionnelle afin de la compléter, sont les bases de données de la *Kyoto Encyclopedia of Genes and Genommes* ou KEGG (Ogata et al. 1999) **(Figure 25)**.

1.6.6. Annotation automatique et biocuration

Les premiers génomes séquencés ont dû être annotés manuellement, tant pour l'annotation syntaxique que pour l'annotation fonctionnelle. Avec l'accumulation des génomes séquencés, une grande quantité d'outils ont été développés. Pour l'annotation syntaxique des CDS, des outils comme GeneMark (Borodovsky and McIninch 1993) ou Glimmer (Salzberg et al. 1998) ont été créés tandis que des outils spécifiques ont également été développés pour les ARNr comme RNAmmer (Lagesen et al. 2007) ou les ARNt comme tRNAscan-SE (Lowe and Eddy 1997). Puis, des logiciels automatiques pour l'annotation fonctionnelle ont vu le jour comme par exemple GeneQuiz (Andrade et al. 1999). L'étape suivante consiste à unifier les deux types d'annotation dans une seule suite de logiciels d'annotation automatique qu'on nomme *pipeline*. Outre la facilité d'annoter avec une seule plateforme, la difficulté d'installer et de mettre à jour les logiciels individuels et les bases de données et de choisir les paramètres des logiciels sont les raisons du choix, surtout par de non

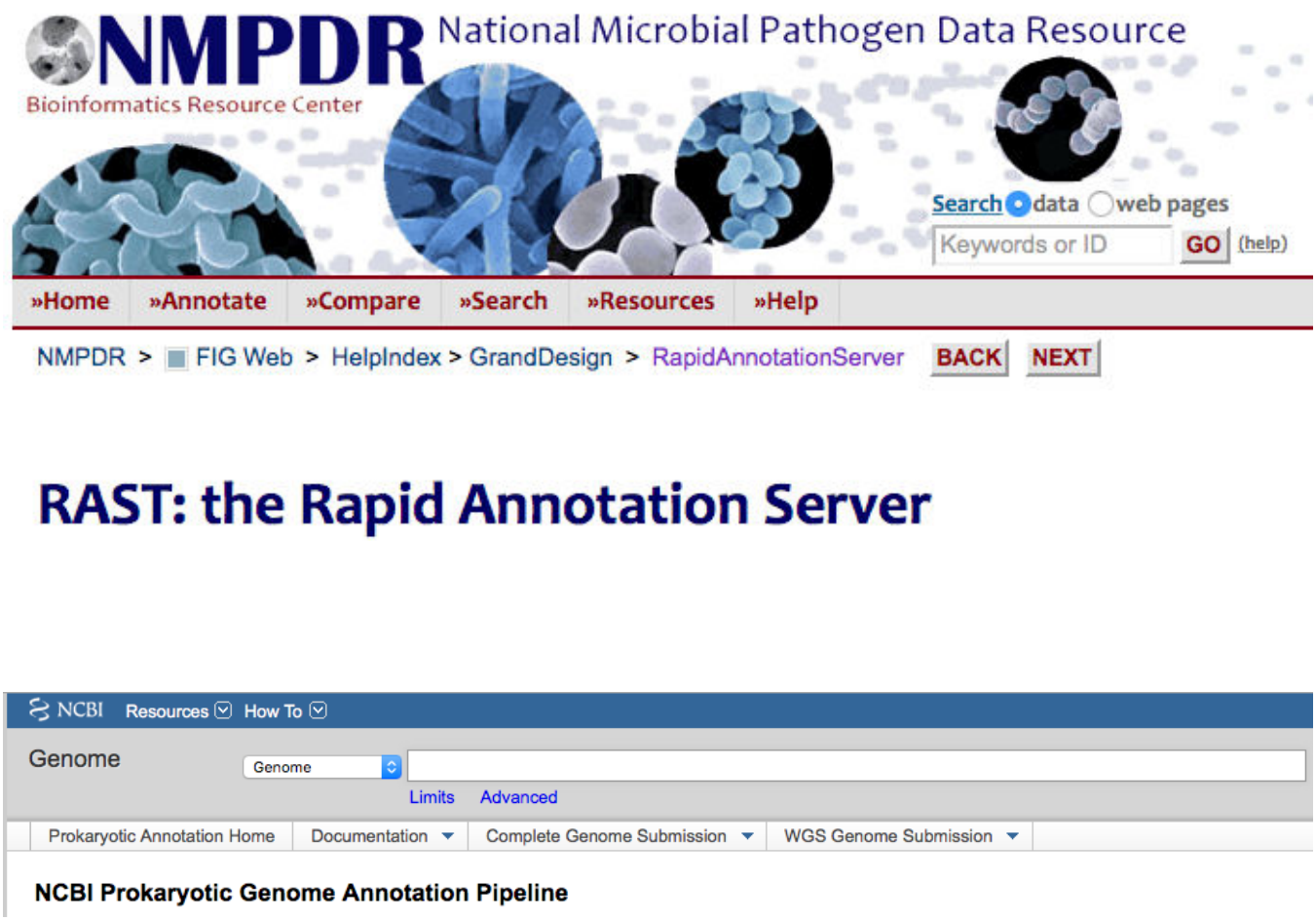


Figure 26. Exemples de pipelines d'annotation automatique en ligne. Du haut vers le bas : la pipeline RAST, la pipeline PGAAP et la plateforme MicoScope. Chaque site web a des procédures différentes pour annoter les génomes et nécessitent de niveaux d'expertise différents pour les utiliser.

Sites utilisés : http://www.nmpdr.org/FIG/wiki/view.cgi/FIG/RapidAnnotationServer ; https://www.ncbi.nlm.nih.gov/genome/annotation_prok/ et http://www.genoscope.cns.fr/agc/microscope/home/index.php
Visités le 23 octobre 2017

spécialistes, des pipelines d'annotation (Andrade et al. 1999). L'évolution de la bioinformatique au sein des structures de séquençage a fait que plusieurs plateformes d'annotation sont proposées actuellement. Ces pipelines peuvent être locales (installées sur l'ordinateur de l'utilisateur) mais la tendance est à un usage délocalisé, à travers un formulaire internet, permettant d'augmenter la force de calcul avec des clusters à distance.

Parmi les pipelines locales dédiés aux génomes bactériens, il est possible de citer Prokka (Seemann 2014). Ces plateformes permettent à l'utilisateur de connaître les logiciels utilisés pour chaque étape de l'annotation et de les paramétrer individuellement. Pour les pipelines via le web, deux tendances co-existent. Certains comme RAST (*Rapid Annotation of microbial genomes using Subsystems Technology*) (Aziz et al. 2008) qui permettent un paramétrage minimum et d'autres comme PGAAP (*Prokaryotic Genome Automatic Annotation Pipeline*) proposée par le NCBI qui sont complètement automatiques, sans aucune intervention du chercheur qui soumet sa séquence (Tatusova et al. 2016). De plus en plus de scientifiques utilisent ce type d'annotateurs automatiques qui sont, la plupart du temps, des boîtes noires (Papanicolaou 2016). Deux autres exemples remarquables de ce type de pipeline sont les plateformes du DoE (*Department of Energy* des États-Unis) appelée *Integrated Microbial Genomes - Expert Review* (IMG ER) (Markowitz et al. 2009) et la plateforme MicroScope du Génoscope (Vallenet et al. 2009) **(Figure 26)**. Des logiciels avec toujours plus de fonctionnalités sont développés au point de proposer assemblage et annotation en même temps (Liao et al. 2015).

Cette multitude de choix, nécessitant des niveaux très différents d'expertise et d'équipement de calcul local, peuvent faciliter amplement l'annotation, mais peuvent également poser des problèmes lors du choix et la comparaison des méthodes. Sur une même séquence génomique, deux plateformes d'annotation peuvent donner des résultats différents voire divergents (Stothard and Wishart 2006). Le but de la biocuration manuelle est de concilier ces annotations automatiques et de détecter, puis de corriger ou d'éliminer les annotations peu informatives, trop évasives ou totalement erronées (Richardson and Watson 2013). Ces erreurs peuvent être multiples et inclure des mauvais choix dans les codons d'initiation et/ou de terminaison, des inconsistances dans les fonctions attribuées à différentes CDS, par exemple dans deux souches annotées par des annotateurs distincts ou encore des gènes orthologues portant des noms ou des descriptions différentes de deux souches apparentées (Richardson and Watson 2013).

À ceci s'ajoutent au moins deux problèmes, la sur-annotation des annotateurs automatiques qui conduit à la production de faux-positifs appelées *Evil-Little Fellows* ou ELFs (Ochman 2002) et la production de faux négatifs correspondant aux gènes non annotés automatiquement, pour plusieurs raisons (H. X. Zhang et al. 2014). De plus, la qualité et la complétude de l'assemblage du génome compliquent l'annotation. En effet, les annotateurs automatiques ont des difficultés avec des assemblages fragmentés (draft) : des paralogues différents mais avec un pourcentage important de similitude peuvent être fusionnés en un seul gène (Indrischek et al. 2016), les gènes avec des domaines répétés peuvent être divisés en deux *loci* et des gènes peuvent tout simplement être absents de l'assemblage et se retrouver dans un gap (Denton et al. 2014).Tous ces possibles biais doivent être pris en compte pour l'étape suivante, la comparaison des génomes suivant la méthodologie dite de génomique comparative.

2. Génomique bactérienne

2.1. Taxonomie bactérienne : espèces et souches

La *taxonomie* est la science de la classification, identification et nomenclature des organismes d'après des critères définis. La *systématique* correspond à la taxonomie quand les critères utilisés sont les liens évolutifs entre les organismes. Dans les deux cas, trois termes sont à définir : i) la *classification* est la formation de groupes de bactéries en espèces, genres, familles, ordres, classes et phyla ; ii) l'identification est l'utilisation de la classification pour distinguer entre les organismes et assigner un isolat comme appartenant à une bactérie ; et iii) la *nomenclature* consiste à donner un nom à une espèce pour la définir et communiquer entre scientifiques (Baron 1996).

La notion d'espèce est définie comme un ensemble d'organismes partageant une même niche écologique, des caractères morphologiques comparables voire identiques et surtout une capacité à se reproduire avec une progéniture fertile (J. Mallet 1995). Chez les procaryotes, qui ne possèdent pas de reproduction sexuée, cette notion est difficile à définir. Chez ces organismes, la reproduction par fission binaire est dominante mais des espèces assez éloignées peuvent échanger du matériel génétique. La notion de souche correspond aux

descendants d'un seul isolement en culture pure, généralement d'une seule colonie. Ceci fait d'une souche un groupement de clones. En microbiologie, la souche est souvent *l'unité de travail* et la description de l'espèce comprend souvent une souche dite *type* (ou une référence) qui définit l'espèce. L'ensemble des souches considérées comme suffisamment proches de la souche type seront incluses dans cette espèce (Lapage et al. 1992).

Entre 1872-1875, le biologiste allemand F. Cohn fut un des premiers à proposer une classification des bactéries qui reposait sur des critères morphologiques auxquels s'ajoutent peu à peu des critères physiologiques (Fox and Stackebrandt 1988). Jusqu'au début des années 1960, la définition d'une espèce reposait sur une classification *phénétique*, utilisant un grand nombre de caractéristiques phénotypiques comme des caractères morphologiques, biochimiques ou environnementaux.

En 1957, P. Sneath s'inspirant des travaux de M. Adanson (1763) propose une taxonomie *numérique* qui code de manière binaire (1 pour présence, 0 pour l'absence) une centaine de caractères morphologiques, biochimiques et culturaux (Sneath 1957b, 1957a). Avec le développement des études génétiques puis génomiques, la place de l'évolution toujours plus importante dans la volonté de classer les organismes, des comités *ad hoc* ont été organisés pour définir les espèces bactériennes et décider sur les caractères à utiliser pour la taxonomie et la systématique. En 1987 et en 2002, les comités Wayne (International Committee on Systematic Bacteriology) réuni à l'Institut Pasteur à Paris (Wayne et al. 1987) et Stackebrandt (International Commettee on Systematic of Procaryotes) réuni à Gent en Belgique (Stackebrandt et al. 2002), ont défini des critères décrivant l'espèce bactérienne. Wayne et collaborateurs définissent *phylogénétiquement* les espèces comme l'ensemble des souches dont les ADN sont identiques à 70% ou plus, mesuré par des expériences de réassociation d'ADN (d'hybridation d'un hétéroduplex) avec une différence de température de fusion (*melting temperature* ou Tm) inférieure ou égale à 5ºC par rapport à la Tm de la molécule naturelle (Wayne et al. 1987). Une mesure plus facile à utiliser est le séquençage en entier du gène codant l'ARNr 16S. Sur les près de 1 500 nt du gène, une identité de 97% et plus peut se substituer à celle de 70% sur le génome complet (Stackebrandt and Goebel 1994).

Toutefois, ces comités ont encouragé le développement de méthodes moléculaires permettant une meilleure définition des espèces bactériennes. Ainsi, différents protocoles d'étude du polymorphisme nucléotidique des espèces/souches ont été réalisés à partir des profils de longueur de fragments de restriction (*Restriction Fragment Length Polymorphism*

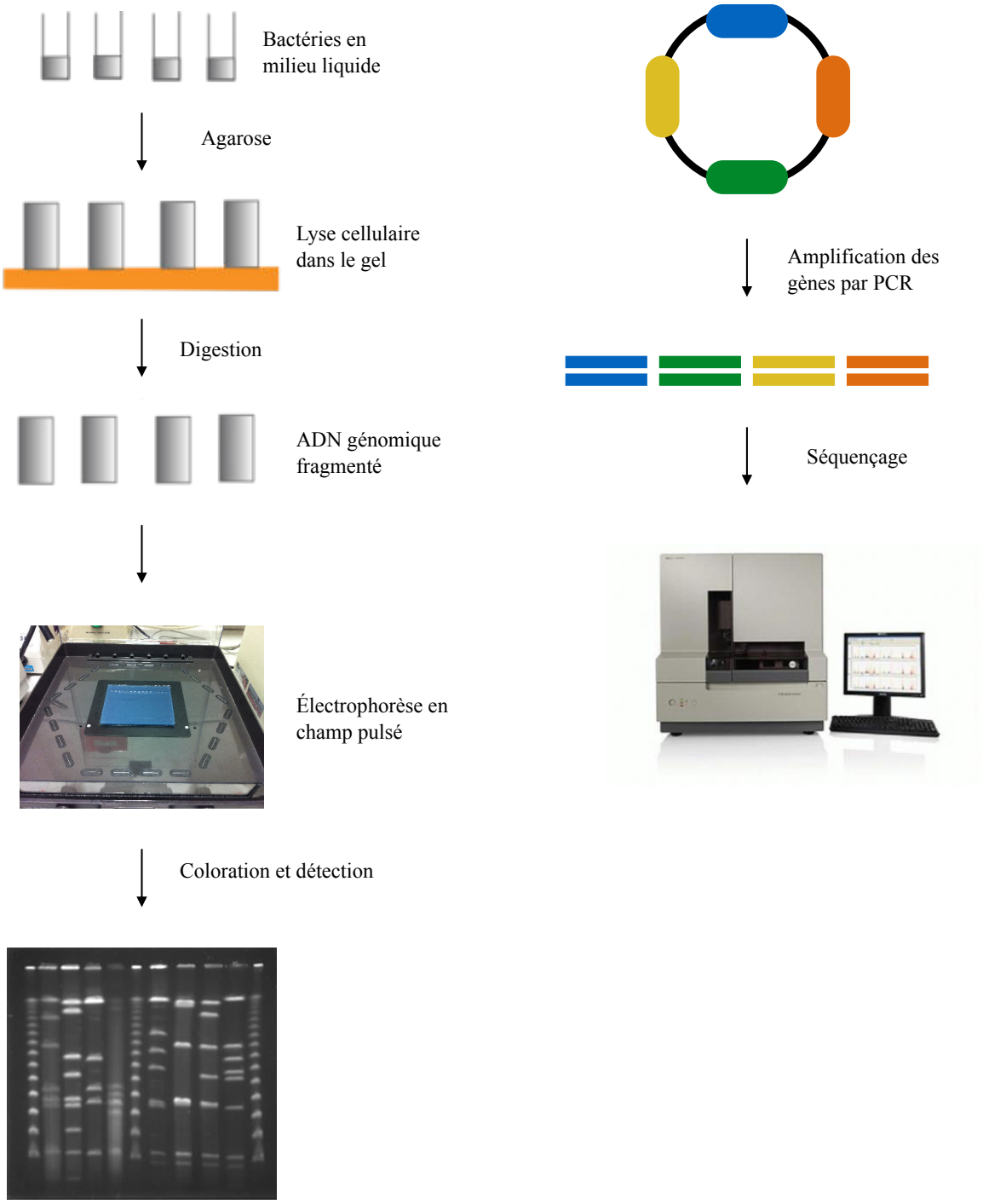


Figure 27. Exemples de techniques de typage. À gauche, le Pulsed Field Gel Electrophoresis (PFGE) de l'ADN génomique d'une bactérie digéré avec une enzyme de restriction. À droite, un schéma d'un Multilocus Séquence Typing (MLST) d'une bactérie, plusieurs gènes sont utilisés. Dans le premier cas, le pattern de bandes est comparé contre une base de données pour déterminer le type de la bactérie. Dans le deuxième cas, la combinaison des allèles détectés est utilisée pour déterminer le type.

Site utilisé : http://www.applied-maths.com/applications/pulsed-field-gel-electrophoresis-pfge-typing
Visité le 23 octobre 2017

ou RFLP) couplée à la *Pulsed Field Gel Electrophoresis* (PFGE) (Gicquel 1993; Klaassen et al. 2002; Muller et al. 1999), de séquençage multilocus (*Multilocus Sequence Typing* ou MLST) (Maiden 2006; Urwin and Maiden 2003), d'hybridations sur des puces à ADN génomique (Schena et al. 1998) ou d'études de polymorphisme de longueur de fragments de restriction (*Amplified Fragment Length Polymorphism* ou AFLP) (P. Vos et al. 1995) **(Figure 27)**. Cependant, l'avènement des méthodes de séquençage a permis d'ouvrir de nouvelles perspectives d'études *phylogénomiques* permettant d'augmenter considérablement la résolution de la classification bactérienne.

La notion d'espèce bactérienne fait encore débat et est parfois remplacée par celle de clusters bactériens plutôt que d'espèce (L. Tang and Liu 2012). Même dans cette vision de clusters, l'importance des transferts horizontaux est un obstacle qui a conduit à proposer une visualisation de l'évolution des bactéries sous la forme de réseaux de gènes transférés plutôt que d'arbres (Dagan 2011; Kunin et al. 2005). Ces deux visions des espèces bactériennes comme des clusters et/ou des réseaux de gènes dépendent en fait du groupe bactérien étudié. Pour les populations *clonales*, très homogènes, le regroupement en clusters est aisé et correspond à la définition canonique de l'espèce (Segerman 2012). Pour les populations *panmictiques*, une grande partie des transferts de gènes ont lieu entre bactéries partageant le même habitat (Popa et al. 2011), les clusters regroupent plutôt des organismes habitat-spécifique avec certains organismes centraux (Kunin et al. 2005).

2.2. Dynamique des génomes bactériens

L'échange de matériel génétique, vu comme un échange sexuel, entre les bactéries est assez permissif. En effet, les populations bactériennes ne sont pas homogènes et peuvent ne pas être strictement clonales, mais panmictiques ou avoir des caractéristiques des deux types, en fonction du taux des transferts horizontaux et de leur écosystème (Falush et al. 2003; J. M. Smith et al. 1993). L'asexualité apparente des bactéries a conduit à surestimer l'importance de la *clonalité* bactérienne. Il existe des populations bactériennes clonales où tous les individus dérivent d'un génotype fondateur par accumulation de mutations (Spratt and Maiden 1999). Dans ce cas, comme les transferts horizontaux sont quasi inexistants, les hypothèses phylogénétiques inférées à partir de certains gènes de ménage peu mutables sont congruentes. Pour certaines bactéries, l'évolution est au contraire façonnée par transfert horizontal et la

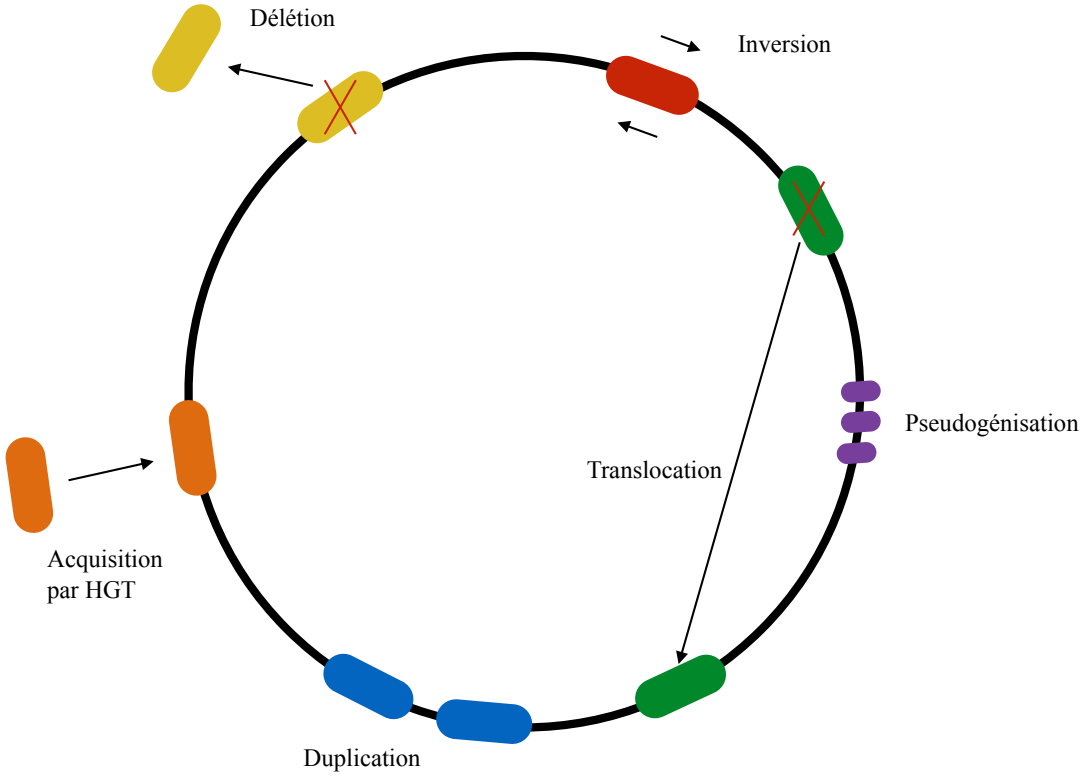


Figure 28. Quelques types de mécanismes de diversité dans une population bactérienne. Des mécanismes intra-chromosomiques produisent un gain (par duplication), une perte (par délation) ou un déplacement (par translocation ou inversion). Un mécanisme de gain inter-chromosomique est l'acquisition de gènes par Horizontal Gene Transfer (HGT). Finalement, la pseudogénisation peut être lente par accumulation de mutations, ou plus rapide par une translocation intra-chromosomique ou inter-chromosomique.

Adapté de : Abby & Daubin 2007.
Visité le 23 octobre 2017

recombinaison. Ces populations, dites *panmictiques*, peuvent survenir avec une fréquence de 200 à 300 fois supérieure aux mutations ponctuelles (Ashton and Caugant 2001; Takuno et al. 2012; M. Vos and Didelot 2009) à tel point qu'au final, chaque isolat représente un génotype unique et que l'histoire évolutive de ces populations est pratiquement impossible à reconstruire. En réalité, la majorité des espèces bactériennes ne suivent pas cette dichotomie clonale/panmictique mais ont une population clonale où des évènements de recombinaison permettent l'adaptation à des conditions environnementales. Ceci crée souvent des écotypes qui sont souvent des pathogènes et dont l'étude génomique peut être complexe. La dynamique des génomes bactériens est donc marqué par des réarrangements génomiques qui peuvent être intra- ou inter-chromosomiques **(Figure 28)**.

Les réarrangements *intra-chromosomiques* conduisant à un déplacement, une inversion, un gain ou une perte de certaines régions d'un génome par recombinaison interne entre séquences d'ADN répétées, notamment les opérons ribosomiques *rrn* (Helm et al. 2003). Les génomes peuvent également abriter de nombreux intégrons (Escudero et al. 2015), et d'autres éléments capables de se déplacer de façon autonome comme les transposons ou les séquences d'insertion (Siguier et al. 2015) ou non comme les MITEs (Delihas 2008). Par des systèmes de *copier-coller* ou de *couper-coller*, ces éléments sont des sources importantes de réarrangements intra-chromosomiques, de pseudogénisation et d'instabilités génomiques (Bennett 2004; Darmon and Leach 2014).

Les réarrangements *inter-chromosomiques* correspondent à des transferts horizontaux et à l'intégration d'ADN étranger dans les génomes selon différents mécanismes :

- La *transformation* est la capture et l'internalisation d'ADN exogène dans des bactéries dites compétentes. Cette compétence peut être acquise par des modifications nutritionnelles ou physiologiques (Claverys and Martin 2003; Johnston et al. 2014a). Dans les communautés microbiennes organisées en biofilm, la matrice externe contient de l'ADN extracellulaire (eDNA), majoritairement du à la lyse cellulaire intra- et inter- populations bactériennes (Allesen-Holm et al. 2006; Liu and Burne 2011) et disponible pour la transformation. L'ADN étranger peut également être transmis par des vésicules de membrane externe (Renelli et al. 2004). À cela s'ajoute la compétence naturelle de certaines bactéries, notamment des pathogènes de la sphère orale comme *Streptococcus pneumoniae* (Johnston et al. 2014b), *Neisseria meningitidis* (Hovland et al. 2017) et *Porphyromonas gingivalis* (Tribble et al. 2012).

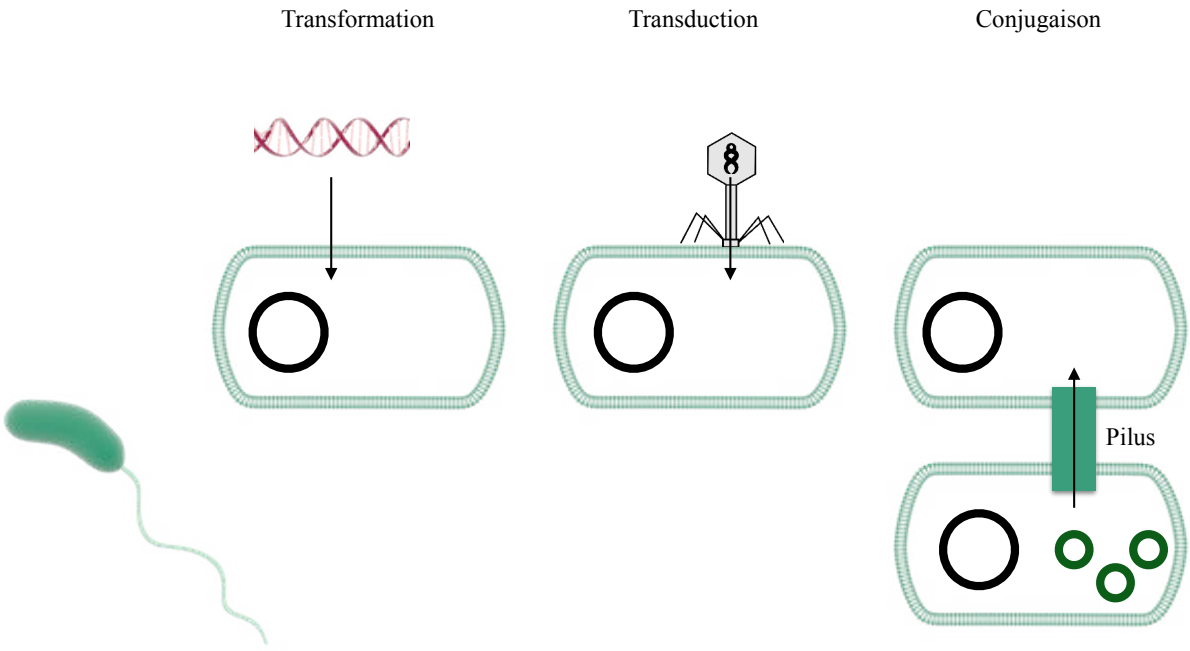


Figure 29. Mécanismes classiques de transfert horizontal de gènes chez les bactéries. La transformation est l'acquisition d'ADN exogène dans le milieu par la bactérie. La transduction est l'injection de l'ADN d'un bactériophage dans sa cellule hôte et la conjugaison est l'échange de matériel génétique (e.g. un plasmide) entre deux cellules (une donneuse et une receveuse) à l'aide d'un pilus.

Formes téléchargées de : http://www.somersault1824.com/science-illustrations/

- La *conjugaison* est la transmission d'ADN d'une cellule donneuse à une cellule receveuse (Arutyunov and Frost 2013). Ce transfert se fait par une structure spéciale appelée *pilus* et le processus est s'apparente le plus à un réel échange sexuel, car les *pili* mettent en contact direct les deux cellules. Les ADN échangés peuvent être d'origine plasmidique ou correspondre à des éléments intégrés dans le chromosome et transférables, comme les transposons conjugatifs (Salyers et al. 1995), également appelés éléments conjugatifs et intégratifs (*Integrative and Conjugative Elements* ou ICE) (C. M. Johnson and Grossman 2015).

- La *transduction* est l'échange d'ADN entre deux cellules par l'intermédiaire d'un virus (bactériophage ou phage). Les phages dits *tempérés* s'intègrent au chromosome de leur hôte établissant la *lysogénie* ou *cycle lysogénique*. Selon les conditions environnementales, le phage peut devenir *virulent* et entrer dans un *cycle lytique*. En encapsulant son génome, le phage peut transférer d'autres segments d'ADN de la cellule donneuse à la prochaine cellule infectée qui dévient la cellule receveuse. Ce mécanisme peut-être généralisé ou spécialisée selon la nature du phage et son rang d'hôte (Brussow et al. 2004; Canchaya et al. 2003) **(Figure 29)**.

2.3. La génomique comparative

La génomique comparative en bactériologie permet d'associer génotype et phénotype, de répondre à des questions d'évolution, de physiologie et de pathogénicité (Prentice 2004). Cependant, dans de nombreux cas, les résultats de ces études restent descriptifs, certaines modifications génomiques n'ont pas nécessairement de conséquences visibles ou faciles à évaluer phénotypiquement. À l'inverse, des différences phénotypiques peuvent ne pas être facilement reliées à des modifications génomiques. La génomique comparative peut faire intervenir de 2 à n génomes. Les comparaisons de génomes peuvent être faites à différents niveaux, chaque niveau donnant accès à des types différents d'information.

La partie du génome contrainte par l'évolution est surtout composée de gènes codants et les comparaisons génomiques sont souvent des approches *gène-centrées* (Koonin and Galperin 1997). Le séquençage et la comparaison de plusieurs génomes au sein d'une espèce (ou d'un genre bactérien) permet de définir le génome total ou *pangénome* (*pan*, tout en grec) (Tettelin et al. 2005; Tetz 2005). Le pangénome est composé du *core génome* contenant les

gènes présents chez tous les organismes comparés, du *génome accessoire* contenant les gènes présents au moins deux des souches ou espèces étudiés et du *génome unique* qui est spécifique d'une seule souche ou espèce (Medini et al. 2005; Tettelin et al. 2008). Le *pangénome ouvert* est celui où l'ajout d'un nouveau génome ajoute de nouveaux gènes**,** à l'opposé d'un *pangénome fermé* où sa taille n'augmente que très peu avec l'ajout de nouveaux génomes, par exemple pour des bactéries à expansion clonale (Rouli et al. 2015).

Le *core génome* inclut l'ensemble des gènes conservés qui codent les fonctions physiologiques fondamentales et les caractères phénotypiques caractéristiques d'une espèce. Certains de ces gènes sont jugés *essentiels* car leur délétion est létale dans les conditions expérimentales testées. L'essentialité d'un gène est néanmoins sujette à caution puisque certains gènes peuvent être individuellement essentiels, mais leur perte pourrait être compensée ou tolérée par d'autres mutations (Hutcherson et al. 2016; Smalley et al. 2003; Yang et al. 2016).

Les génomes *accessoires* et *uniques* sont ceux qui gouvernent la diversité observée des souches ou des espèces et sont généralement retrouvés dans les îlots génomiques. Ils codent, à de rares exceptions près, des fonctions supplémentaires conférant un avantage adaptatif. Chez les bactéries pathogènes, ces gènes peuvent être des déterminants de pathogénicité (toxines, protéines d'export ou de sécrétion) et d'interaction avec l'hôte (molécules d'adhésion, protéines de surface…) (Jackson et al. 2011).

En plus des analyses comparatives *gène-centrées*, les génomes peuvent être comparés selon une approche *structure-centrée*. À l'exception de quelques espèces bactériennes qui possèdent des chromosomes linéaires, comme *Agrobacterium tumefaciens* (Wood et al. 2001), *Borrelia burgdorferi* (Fraser et al. 1997), ou *Streptomyces coelicolor* (Bentley et al. 2002) ; la grande majorité des chromosomes bactériens sont circulaires et l'organisation des gènes le long de ce cercle est faite de façon à maximiser l'efficacité de la réplication du chromosome et de l'expression des gènes.

En 1963, le biochimiste britannique J. Cairns démontre la réplication bidirectionnelle du chromosome circulaire bactérien avec des fourches de réplication (*replication fork*) (Cairns 1963; Griffiths et al. 2000). Deux fourches de réplication se forment, la réplication commence à l'*origine de réplication du chromosome* (*oriC*) et se termine au *terminus* ou site *ter*, et progressent dans des directions opposées, avec synthèse continue d'un *brin précoce* ou *leading strand* à réplication directe, et discontinue d'un *brin retardé* ou *lagging strand*. Ce

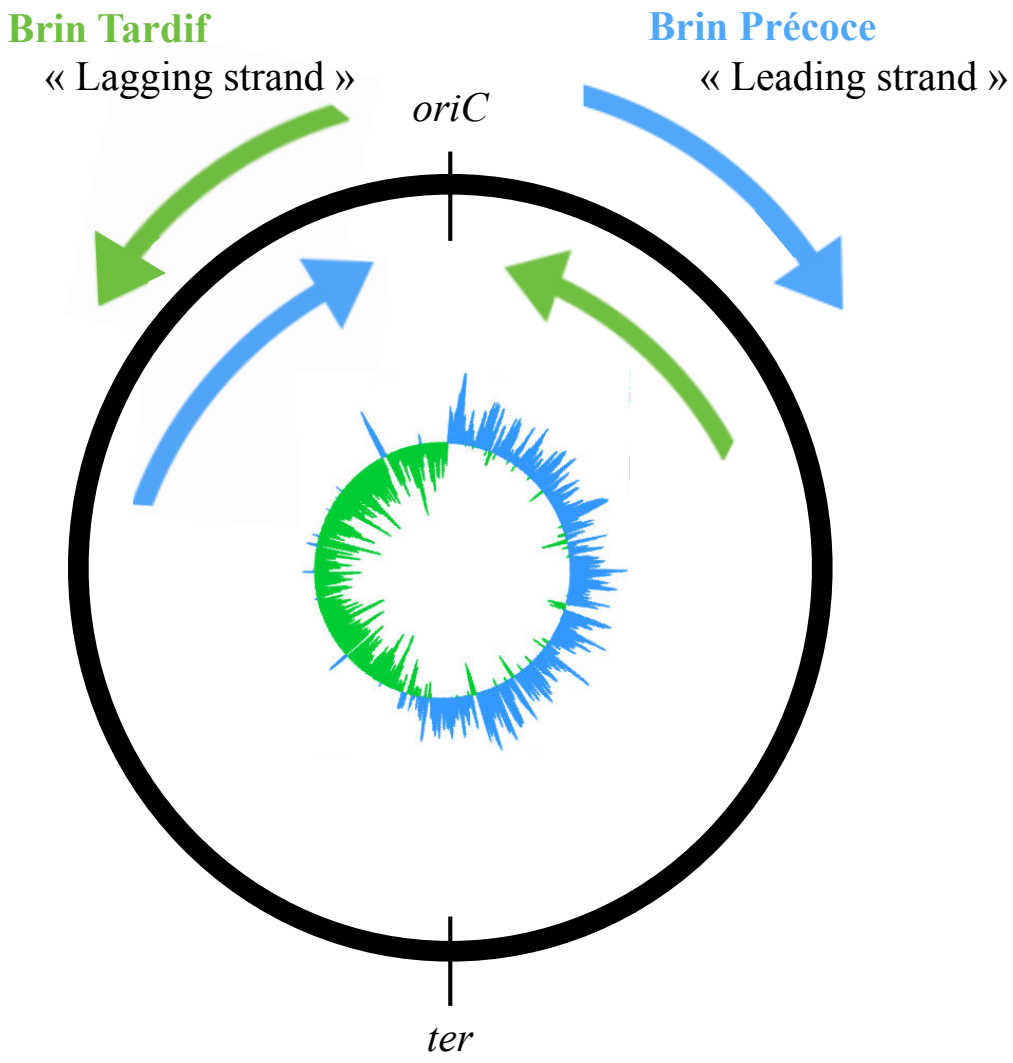


Figure 30. Structure du chromosome circulaire des bactéries. La forme de réplication implique la synthèse directe du brin précoce et discontinue du brin tardif. La réplication commence à l'*oriC* et se termine au *ter*. À l'intérieur du schéma est représenté le biais de CG le long du chromosome avec des valeurs positives pour le brin précoce riche en G et négatives pour le brin tardif. Les gènes fortement exprimés seront préférentiellement près de l'*oriC* et les gènes essentiels seront préférentiellement sur le brin précoce.

Adapté de : Abby & Daubin 2007, Rocha 2008 et Rocha & Danchin 2003.

mécanisme de réplication influence fortement la répartition des gènes sur le chromosome (Rocha 2008). La fréquence des mutations diffère selon les brins, entraînant des *biais* (*skews*) de composition remarquables entre les deux brins, notamment dans la distribution de G et de C (Kowalczuk et al. 2001; Lobry and Louarn 2003) **(Figure 30)**.

Chez certaines bactéries à croissance rapide, les gènes impliqués dans la croissance bactérienne sont proches d'*oriC* alors que ceux impliqués dans l'adaptation sont en revanche situés autour du site *ter* (Couturier and Rocha 2006), les gènes essentiels se trouvent préférentiellement sur le brin précoce où ADN et ARN polymérase progressent dans le même sens, évitant la collision frontale des deux enzymes et l'interruption prématuré de la transcription. À l'inverse, si un gène essentiel se trouve sur le brin tardif, la fréquente interruption de sa transcription entraînerait sa contre-sélection (Rocha and Danchin 2003). Pour cette raison, l'évolution des bactéries se produirait de préférence selon un gradient de plasticité génomique, centrée autour de certaines régions plus proches du *ter* que de l'*oriC*. La comparaison des structures de génomes bactériens proches permet d'identifier ces régions de plasticité génomique, définies comme des *loci* où la synténie est rompue et/ou des régions absentes/présentes, en raison de phénomène de réarrangements intra- ou inter-chromosomiques décrit précédemment (Abby and Daubin 2007; Daubin and Perriere 2003; Suyama and Bork 2001).

En raison de ces réarrangements, les génomes sont des entités mosaïques constitués de segments conservés, de régions plus ou moins partagées avec d'autres et de zones uniques, éventuellement réorganisées d'un génome à l'autre. Afin de pouvoir comparer à la fois la structure, l'organisation et le contenu génique de plusieurs génomes d'intérêt, la méthode la plus souvent utilisée, surtout en première instance, est l'alignement. Aligner de 2 à n séquences génomiques est une tâche plus difficile que l'alignement de séquences géniques courtes. Les algorithmes cités précédemment ne sont pas optimisés pour cette tâche, le temps de calcul étant trop long et la mémoire de l'ordinateur requise excessive (Delcher et al. 1999).

Pour les alignements génomiques, qui sollicitent donc des méthodes globales, deux stratégies existent. L'alignement itératif de paires de génomes (*iterative pairwise alignment*) où des alignements locaux de plusieurs sous-parties génomiques sont fusionnées de façon itérative et la recherche de points d'ancrage (*anchor-based multiple alignment),* particulièrement efficace pour ce type de séquences longues (Hohl et al. 2002).

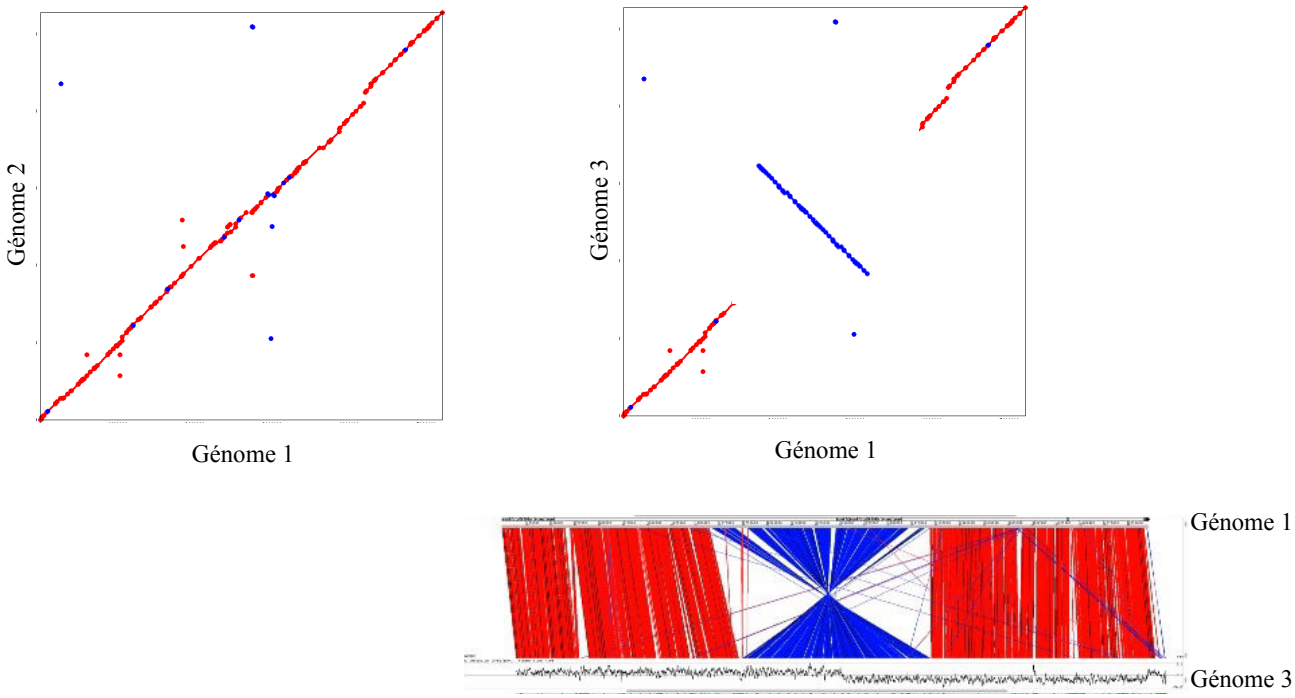


Figure 31. Alignement de génomes. L'alignement de génomes peut être fait par paires ou être multiple. En haut, des représentations de type dot plot (faites avec le logiciel MUMmer) de deux génomes colinéaires (1 et 2), et de deux génomes où l'un présente une inversion centrale par rapport à l'autre (en bleu, génomes 1 et 3). Pour ces deux génomes, une autre représentation (faite avec Artemis-ACT) est proposée en bas, l'inversion centrale peut également être observée.

En bas, alignement multiple de trois génomes (A, B et C), une grande quantité de réarrangements peuvent être observés (alignement fait avec progressiveMAUVE).

Figure d'Artemis-ACT adaptée de : https://fr.slideshare.net/leightonp/comparative-genomics-and-visualisation-part-1
Téléchargée le 24 octobre 2017

Un des premiers algorithmes utilisant l'ancrage est l'outil MUMmer, initialement conçu pour aligner rapidement une paire de génomes (entiers ou incomplets). Dans ce type d'alignements, la première étape consiste à décomposer les génomes à aligner, à l'aide d'un arbre des suffixes, pour trouver des *Maximal Unique Matchs* (MUM). Les MUMs correspondent à des sous-séquences *uniques* (non répétées sur les génomes) *communes* (sans mismatch) aux n génomes étudiés et *maximales* (ne pouvant pas être étendues). Les MUMs sont ensuite replacés sur chaque génome et servent d'ancrage pour comparer les gaps qui correspondent aux SNP, indels et régions polymorphes ou répétées (Delcher et al. 1999). La visualisation des résultats est proposée sous forme de *dot plot* (graphique de points) représentant en diagonale directe les séquences identiques et en diagonale inverse, les régions d'inversion (Gilbert 1990). Cette approche a été reprise dans l'outil Multiple Genome Aligner (MGA), en étendant le concept de MUM pour en définir des *maximal multiple exact matches* (multiMEMs), qui peuvent être uniques ou non (Hohl et al. 2002). MUMmer et MGA sont très utiles et rapides pour aligner des génomes proches, mais ne sont pas adaptés pour d'aligner des génomes éloignés et mosaïques. Le logiciel Mauve et son amélioration progressiveMauve (A. C. Darling et al. 2004; A. E. Darling et al. 2010) détectent et alignent des régions localement colinéaires (*Locally Collinear Blocks* ou LCB) en ancrant des multi-MUMs qui peuvent n'être commun qu'à un sous-ensemble des génomes comparés (A. E. Darling et al. 2010) **(Figure 31)**. D'autres outils destinés à l'alignement de génomes complets existent mais, en génomique microbienne, les trois précédemment cités sont les plus fréquemment utilisés. Les autres outils qui peuvent être cités sont Mugsy (Angiuoli and Salzberg 2011), GR-Aligner (Chu et al. 2009), GAME (Choi et al. 2005) ou encore LAGAN et Multi-LAGAN (Brudno et al. 2003).

En plus de ces alignements globaux de génomes permettant de comparer l'organisation des génomes entiers, différentes autres méthodes de comparaison ciblée existent. Certaines méthodes se concentrent sur les régions codantes. Ainsi, certains outils analysent les pangénomes et créent des clusters de gènes orthologues, visualisés généralement par des diagrammes de Venn montrant les proportions relatives de génomes core, accessoire et uniques. Certains outils comme GenoSets (Cain et al. 2012), VennPainter (Lin et al. 2016), BioVenn (Hulsen et al. 2008) ou VennDiagram (Chen and Boutros 2011) peuvent être cités. D'autres approches sont utilisées pour comparer l'organisation des gènes, c'est à dire leur synténie comme SyntenyTracker (Donthu et al. 2009) ; alors que certains logiciels permettent l'évaluation des liens évolutifs entre différents isolats, comme par exemple le calcul des ANI

(*average nucleotide identity*) (M. Richter and Rossello-Mora 2009) et autres comme le GGD (*Genome-to-Genome Distance*) (Meier-Kolthoff et al. 2014), les pourcentages et biais en k-mer ou en GC ou encore l'identification d'îlots génomiques ou de *hot spots* de recombinaison. La liste d'outils et d'approches permettant de comparer des génomes est longue, chacun permettant d'obtenir des informations qui aident à la compréhension des génomes.

3. Bacteroidetes et *Porphyromonas*

3.1. Phylum Bacteroidetes

3.1.1. Classification et caractéristiques générales

Dès le milieu des années 1980 et jusqu'au milieu des années 1990, les travaux du groupe de C. R. Woese ont été les premiers à mettre en évidence les liens phylogénétiques entre les bactéries anaérobies du groupe des *Bacteroides* et les bactéries aérobies des groupes des *Flavobacteria* et *Cytophaga* (Paster et al. 1985; Woese 1987). Ce groupe de bactéries à coloration Gram-négative (didermes) non sporulantes est hétérogène phénotypiquement mais phylogénétiquement cohérent. Il est alors classé au rang de phylum et appelé Cytophaga-Flavobacter-Bacteroides ou CFB (Gherna and Woese 1992; C. J. Smith et al. 2006). En 2010, avec la publication de la deuxième édition du *Bergey's Manual of Systematic Bacteriology*, ce phylum a été renommé Bacteroidetes (Krieg et al. 2010). En 2015, une nouvelle proposition pour inclure le rang "phylum" dans le code international de nomenclature des procaryotes propose de changer à nouveau son nom en Bacteroidaeota (Oren et al. 2015). Ces propositions n'ont pas encore été adoptées par la communauté scientifique, et donc le terme Bacteroidetes reste employé.

Bien que ce groupe soit phénotypiquement hétérogène, certaines caractéristiques distinctives du phylum sont remarquables, par exemple, concernant le mécanisme de la transcription des gènes. Ainsi, la séquence protéique de la sous-unité B de l'ADN gyrase contient une insertion de 4 acides aminés, unique au phylum (Gupta 2004; Gupta and Lorenzini 2007). Les membres de ce groupe ont également un facteur d'initiation de la transcription (facteur sigma de l'ARN polymérase) unique à ce phylum, capable de transcrire exclusivement des gènes de Bacteroidetes et non d'autres bactéries (Vingadassalom et al.

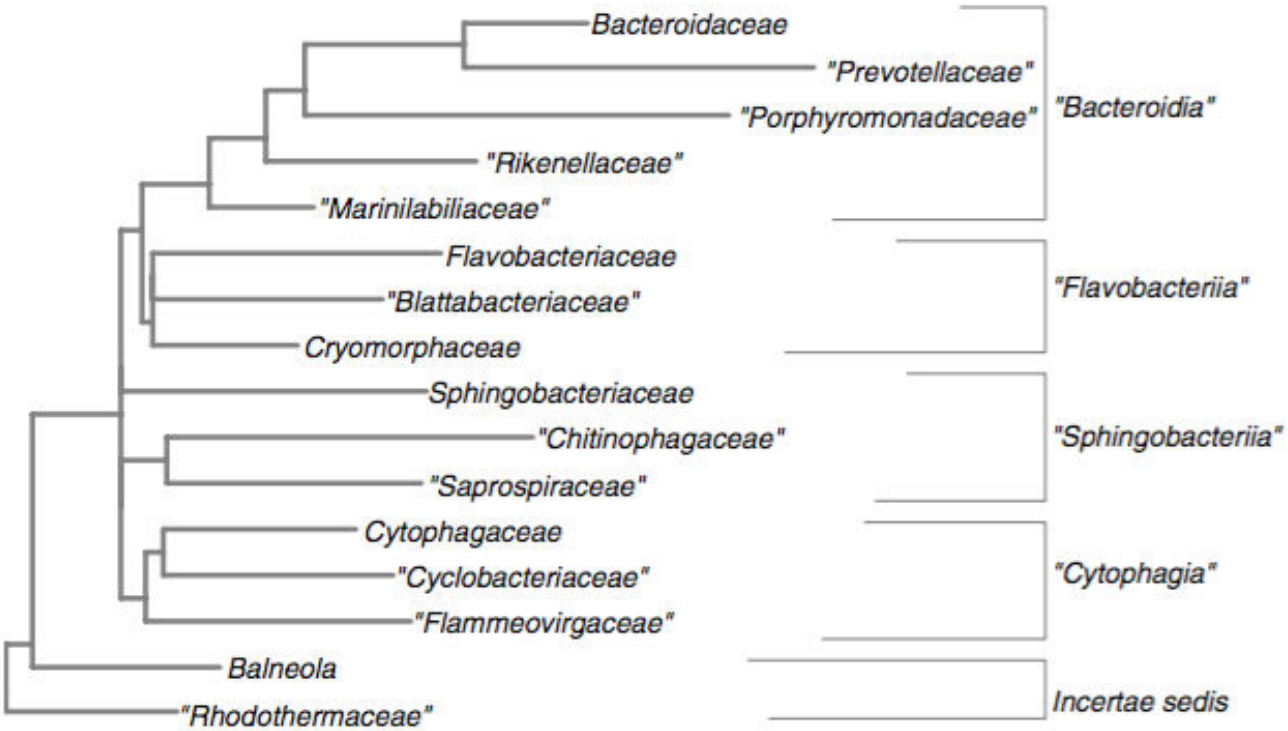


Figure 32. Phylum Bacteroidetes. D'après le Bergey's Manual of Systematic Bacteriology (2ème édition), le phylum Bacteroidetes se divise en quatre classes avec deux groupes *Incertae sedis*.

Extrait de : Kieg et al. 2010

2005). Le troisième élément commun est la faible proportion de gènes précédés de séquences de Shine-Dalgarno (Accetto and Avgustin 2011), la vaste majorité des transcrits de ces bactéries seraient sans séquence leader (sans 5' UTR) et capables de se lier directement au ribosome complet ou la traduction serait médiée par la protéine ribosomique S1 (Nakagawa et al. 2017). Une autre caractéristique métabolique exclusive de ce phylum est le système de glycosylation des protéines, qui s'effectue par une liaison O-mannose avec une rhamnose branchée (Coyne et al. 2013). Finalement, les bactéries du phylum Bacteroidetes possèdent un système de sécrétion unique dit de type IX (T9SS ou système de sécrétion Por) qui permet le relargage de protéines du périplasme vers l'extérieur (McBride and Zhu 2013). Les gènes du T9SS sont également nécessaires pour la mobilité de certaines espèces de Bacteroidetes par un mouvement dit glissant ou *gliding mobility* (Lasica et al. 2017).

Au sein des Bacteroidetes, on distingue au moins cinq sous-groupes, décrits dès les années 1990 : les cytophaga, les flavobacter, les bacteroides, les sphingobacter et les saprospira (Gherna and Woese 1992). Malgré ces observations, le phylum Bacteroidetes est divisé en quatre classes dans le Bergey's Manual : Bacteroidia, Flavobacteriia, Sphingobacteriia et Cytophagia auxquels s'ajoutent deux groupes qui n'ont pas pu être classés (Incertae sedis) (Krieg et al. 2010) **(Figure 32)**. Deux équipes de recherche ont aidé à clarifier la sous-division des Bacteroidetes. Dans un premier temps, Muñoz et al. séparent un sous-groupe des Sphingobacteria qu'ils nomment Chitinophagia (Muñoz et al. 2016), puis Hanke et al. séparent un second sous-goupe de cette dernière classe et donnent aux Saprospira le rang de classe (Hahnke et al. 2016). Ainsi, actuellement, ce phylum contient six classes : Bacteroidia, Flavobacteriia, Cytophagia, Sphingobacteria, Chitinophagia et Saprospiria ; les deux groupes Incertae sedis sont classés chacun dans un phylum à part (Hahnke et al. 2016). Une particularité des classes des Bacteroidetes est que chacune ne contient qu'un seul ordre et cette division n'est donc pas informative (Hahnke et al. 2016).

3.1.2. Écologie

Une des autres caractéristiques importantes des Bacteroidetes est leur distribution dans des écosystèmes variés. Ainsi, en plus de leur classification phylogénétique, leurs niches écologiques les séparent en groupes environnementaux ou associés au tube oro-gastrointestinal d'animaux. Les Bacteroidetes environnementaux correspondent principalement aux classes Flavobacteriia, Cytophagia et Sphyngobacteriia. Ces bactéries ont

Human Microbiome Project

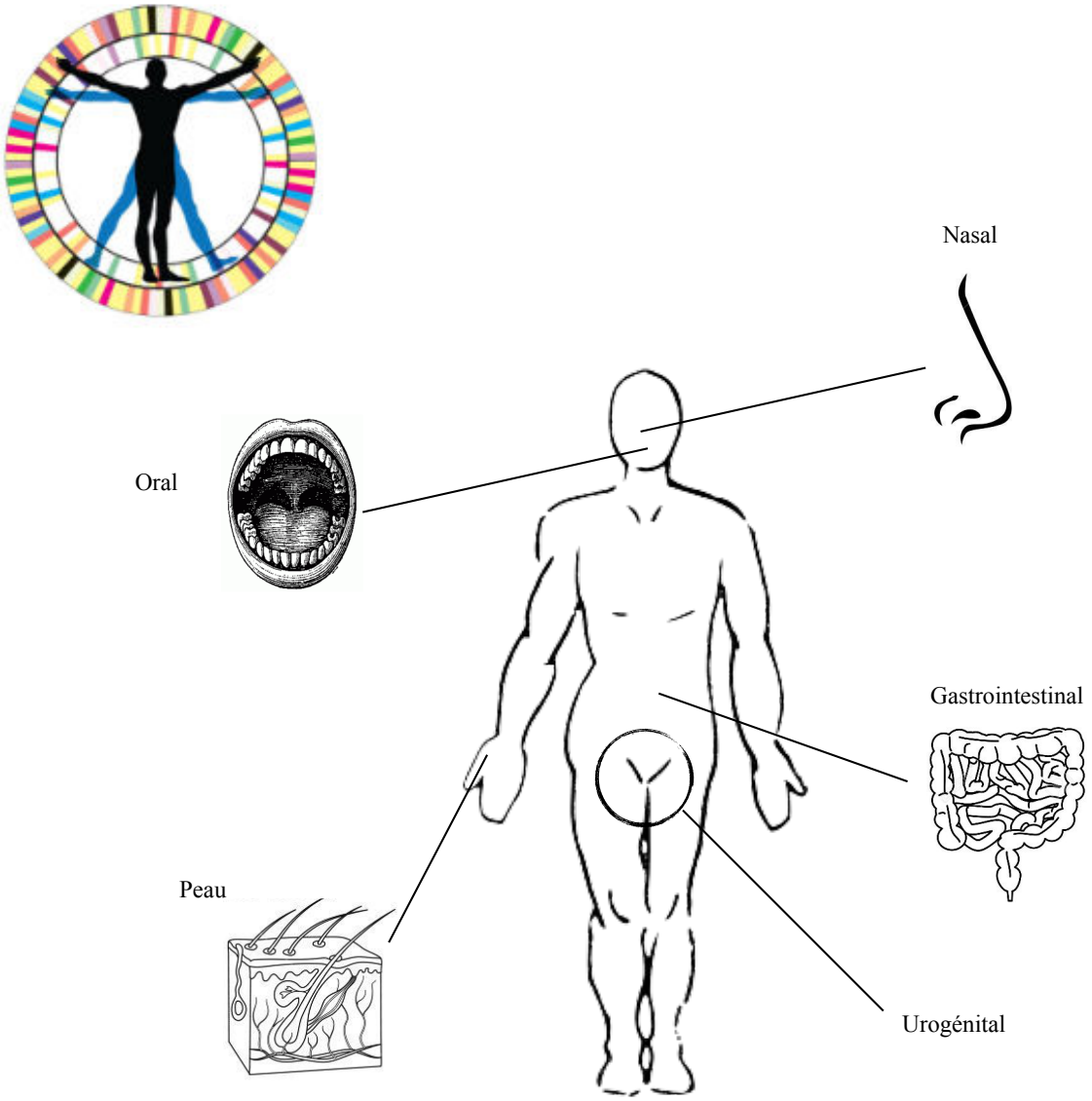


Figure 33. Sites anatomiques étudiés lors du Human Microbiome Project. Ce premier grand projet visant à répertorier les sites colonisés par les bactéries les a divisé en cinq types : peau, gastrointestinal, urogénital, nasal et oral. Seulement des personnes saines ont été recrutées pour cette étude. Des mesures de diversité pour chaque site et entre les différents sites des personnes ont été fournies.

Adapté de : https://commons.wikimedia.org/wiki/File:Anotomical_Position_Sketch.png ; http://outspokenarts.org/mouth-open-huge307x379/ ; https://www.123rf.com et https://www.shutterstock.com
Téléchargées le 24 octobre 2017

été isolées du sol, de compost, de plantes en décomposition, d'eau douce et d'échantillons d'eau de mer (Thomas et al. 2011). Il est estimé que les Bacteroidetes correspondent au troisième phylum en abondance dans les milieux marins (Fernandez-Gomez et al. 2013). Les Cytophagia sont des bactéries mobiles, retrouvées dans des habitats riches en matière organique comme le sol, les plantes en décomposition et les excréments d'herbivores mais aussi le sédiment des rivières, les lacs ou les estuaires (Reichenbach 2006). Les Sphingobacteriia ont été isolés d'habitats aqueux (eau douce et marine) vierges ou hautement contaminés par des pesticides, des hydrocarbures, etc (Balkwill et al. 2006). Finalement, les Flavobacteriia est le groupe le plus nombreux des Bacteroidetes en terme d'espèces. Leur gamme d'habitat est ample et comprend les sols, les environnements aqueux, certains aliments et produits laitiers, les animaux, les amibes et les plantes. Certains genres de cette classe sont spécifiques de certains écosystèmes. Par exemple, les espèces du genre *Flavobacterium* sont associées principalement aux poissons et les milieux aqueux alors que les *Riemerella* sont associées aux oiseaux et les *Capnocytophaga* aux cavités buccales des mammifères, notamment humaines et canines (Bernardet and Nakagawa 2006).

Concernant les Bacteroidetes associées au tube oro-gastrointestinal des animaux, ce sont principalement des bactéries de la classe Bacteroidia. Les Bacteroidetes correspondent à près de la moitié des bactéries détectées dans le colon (E. L. Johnson et al. 2017) et sont également présents dans la cavité orale, la gorge, l'œsophage et même l'estomac. De plus, ils ne sont pas restreints à l'humain mais colonisent également d'autres mammifères comme la souris, le chien, le porc et les ruminants. La relation avec l'hôte est mutualiste puisqu'elle bénéficie aux deux partenaires : les sources de nutriments pour les bactéries sont importantes et l'hôte bénéficie de la dégradation des polysaccharides complexes qui peuvent représenter une source additionnelle d'énergie de près du 10% journalier (Thomas et al. 2011).

3.1.3. Microbiotes et genres d'importance en santé humaine

Le *microbiote* des animaux comprend des procaryotes (bactéries et archées) et des eucaryotes (champignons et protozoires). La grande majorité correspond à des bactéries (E. L. Johnson et al. 2017). Le *Human Microbiome Project* (HMP) a répertorié les sites colonisés par les bactéries et les a classés en cinq aires : peau, nasal, oral, urogenital et gastrointestinal **(Figure 33)**. Les microbiotes avec la plus grande *alpha diversité* d'espèces, métrique qui prend en compte la richesse (le nombre d'espèces dans la communauté) et l'uniformité (une

mesure de l'équité des abondances des espèces), sont ceux de la bouche et des fèces (Human Microbiome Project Consortium 2012; Morgan et al. 2013). Cependant leur beta-diversité (la diversité entre les communautés du même site) est jugée faible (Human Microbiome Project Consortium 2012).

Le tube oro-gastrointestinal est dominé, au niveau du colon, par deux phyla, les Firmicutes et les Bacteroidetes (ca. 90%) et le genre le plus abondant est *Bacteroides* (Bull and Plummer 2014; Heinken et al. 2013). Un déséquilibre du microbiote, appelé *dysbiose*, est associé à plusieurs maladies comme le diabète, l'asthme, les allergies, l'autisme et le cancer (Lloyd-Price et al. 2016). Plusieurs études se sont concentrées sur les changements des microbiotes intestinaux et l'obésité, en observant les microbiotes des fèces et en résumant la richesse du microbiote à un ratio Bacteroidetes/Firmicutes (Greenhill 2015; Koliada et al. 2017; Lloyd-Price et al. 2016). De plus en plus d'auteurs estiment que parler de Bacteroidetes au sens large dans ces études est incorrecte et qu'il serait souhaitable de préciser que seule une cinquantaine d'espèces, réparties en sept genres et issues de la famille Bacteroidia sont présents dans ces échantillons fécaux. Caractériser, identifier et étudier plus précisément ces espèces permettrait de mieux comprendre les relations de certaines dysbioses avec plusieurs pathologies (Hoyles and McCartney 2009; Wexler and Goodman 2017). Une description des microbiotes plus fine et rigoureuse est nécessaire pour comprendre les interactions microbe-microbe et microbe-hôte, dans l'objectif de diagnostiquer des maladies (prédisposition, progression et/ou pronostic) et/ou de déterminer des interventions thérapeutiques (Morgan et al. 2013). La capacité de déterminer la virulence de certaines bactéries pathogènes du microbiote est également prometteuse (Hugon et al. 2015).

L'entrée du tube gastrointestinal est la bouche, seule partie de corps où des tissus durs (les dents) sont exposés à l'environnement. La bouche est divisée en vestibule entre les arcades alvéolo-dentaires (maxillaire et mandibulaire), l'intérieur des lèvres et des joues. La cavité orale est, à proprement parler, à l'intérieur des arcades et limitée par le palais, le plancher buccal sous la langue et l'isthme du gosier. Les arcades alvéolo-dentaires sont revêtues d'une muqueuse, la gencive. Les tissus de support des dents, appelés parodonte, sont l'os alvéolaire, la gencive, le ligament alvéolo-dentaire et le cément, tissu minéralisé qui recouvre la racine des dents (Lamont et al. 2014).

Le microbiote oral est organisé en biofilm, souvent appelé plaque dentaire, qui adhère aux surfaces de la cavité buccale alors que certains formes microbiennes, dites planctoniques,

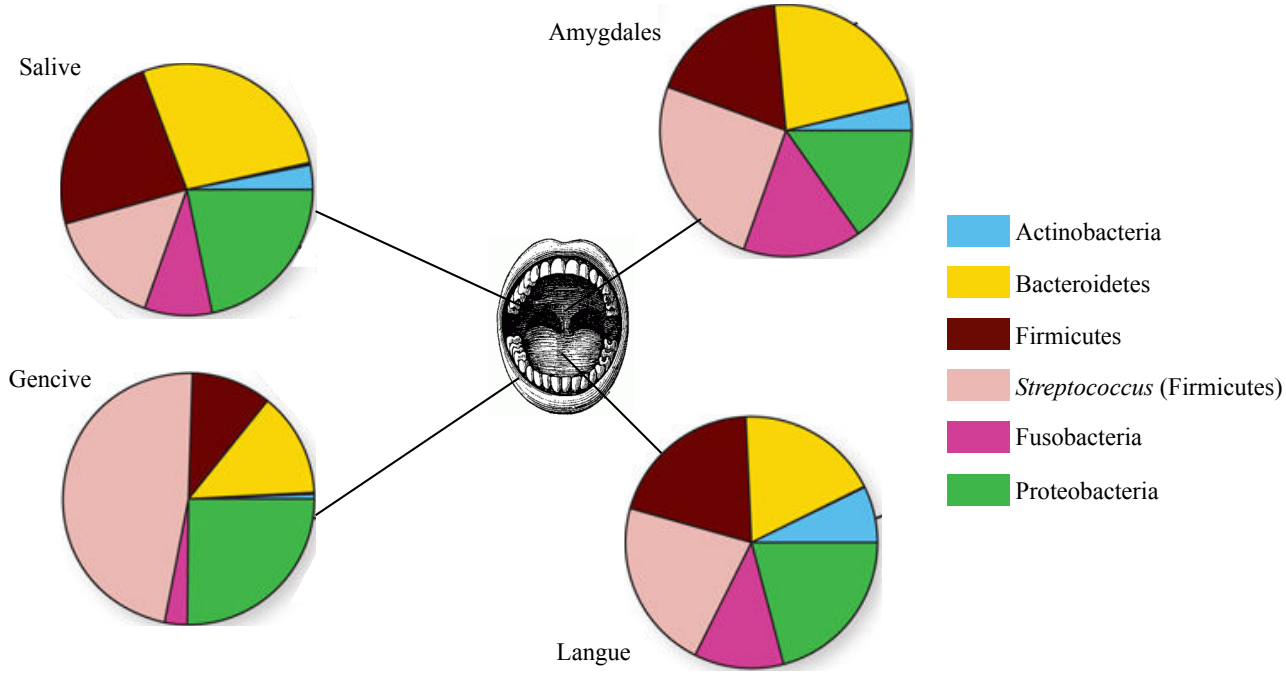


Figure 34. Composition du microbiote oral des différents sites d'une bouche saine. Les différents espaces de la bouche ont des microbiotes différents en composition au niveau du phylum bactérien. Cinq phyla représentent la quasi-totalité des bactéries présentes mais leur proportion dépend du site anatomique analysé.

Adapté de : http://outspokenarts.org/mouth-open-huge307x379/ et Grice & Segre (2012) 'The Human Microbiome: Our Second Genome' *Annu Rev Genomics Hum Genet*, 13, 151-70.
Téléchargées le 24 octobre 2017

sont suspendues dans la salive. Ce fluide est la source principale de nutriments et contient des glycoprotéines, des peptides et des acides aminés. Toutes les surfaces de la bouche sont colonisées par les microbes : la muqueuse des joues et de la gencive, les deux faces de la langue, le palais, et la plaque dentaire s'accumule sur la gencive et l'espace sous-gingival, entre les dents et l'épithélium (Lamont et al. 2014). Plus de 500 espèces bactériennes peuvent être détectées dans la bouche et appartiennent principalement aux phyla Bacteroidetes, Firmicutes, Fusobacteria, Actinobacteria, Proteobacteria et Spirochaetes (Reynolds-Campbell et al. 2017).

Le microbiote oral est acquis rapidement après la naissance, la principale source est le microbiote oral de la mère. Un processus de colonisation permanente est initié par des colonisateurs primaires dont l'adhérence et la croissance altèrent le micro-environnement, avec excrétion de produits métaboliques qui favorisent la croissance d'autres espèces bactériennes. Ce processus de succession d'étapes de colonisation augmente la diversité de la communauté et conduit à la formation d'un biofilm stable et complexe. Des facteurs externes comme la diète, le tabagisme, l'hygiène orale, les maladies systémiques, les changements hormonaux et les troubles physiques et psychiques peuvent induire des changements dans la composition du microbiote oral (Sampaio-Maia et al. 2016).

L'homéostasie des biofilms oraux et sa relation symbiotique avec l'hôte sont importantes pour la santé orale. Leur déséquilibre est l'élément déclencheur notamment de deux maladies buccales importantes : les caries et les maladies parodontales. Les bactéries commensales du microbiote oral ont des effets bénéfiques dans la physiologie, la nutrition et l'immunité de l'hôte. Les colonisateurs permanents forment une barrière aux microorganismes exogènes qui transitent et qui pourraient être pathogènes (Samaranayake and Matsubara 2017; Turnbaugh and Stintzi 2011).

Tous les sites de la bouche ont des microbiotes de compositions différentes (Takahashi 2015) **(Figure 34)**. Par exemple, au niveau du sillon gingivo-dentaire ou sulcus, qui correspond à l'espace d'union entre la marge de la gencive et la dent, un individu sain de maladies buccales aura un microbiote principalement composé de coques à Gram-positif du genre *Streptococcus*. Cependant, comme ce sillon a un potentiel redox inférieur au reste de la bouche, l'accumulation de plaque chez un individu atteint, réduit encore la quantité d'oxygène disponible, favorisant l'installation de bacilles à Gram-négatif, principalement bactéries anaérobies appartenant aux phyla Fusobacteria, Spirochaeta et Bacteroidetes. Dans ce dernier

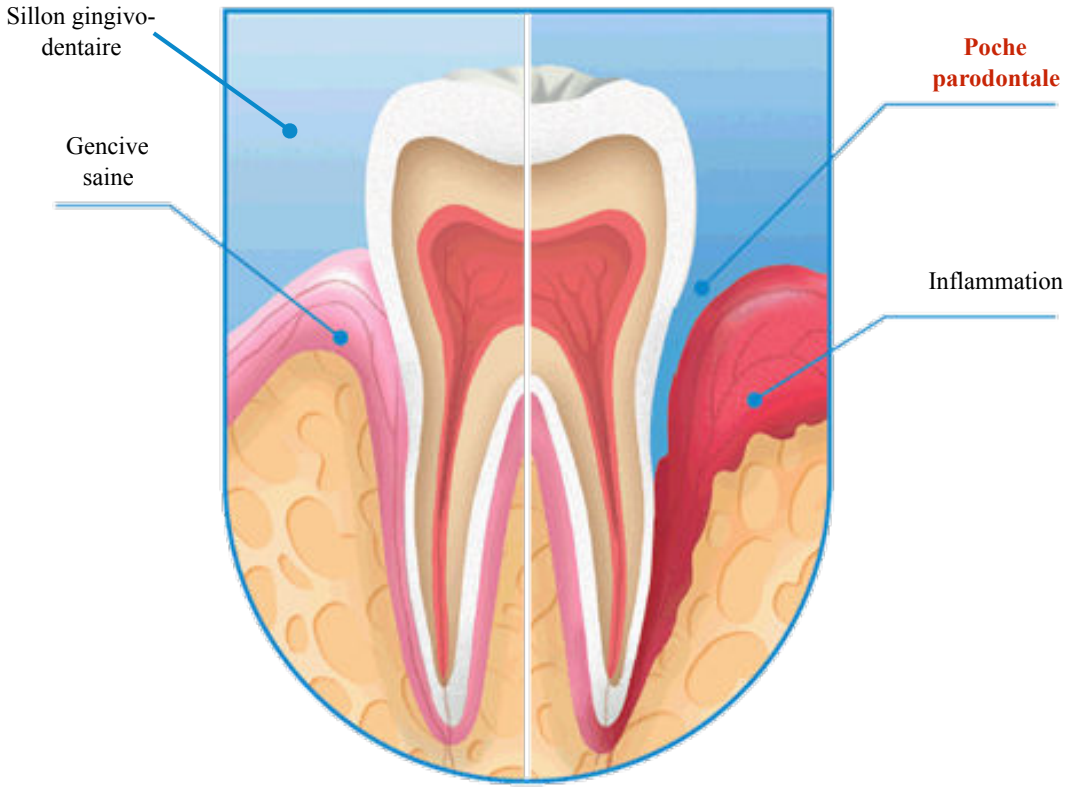


Figure 35. Sillon gingivo-dentaire et poche parodontale. Le sillon gingivo-dentaire est le site d'union entre la gencive et la dent. Dans une bouche saine, il est peut profond (1-2 mm) et colonisé principalement par des bactéries du genre *Streptococcus*. Lorsque le sillon s'inflamme et des bacilles à Gram négatif s'installent, il s'approfondit en créant une poche parodontale.

Adapté de : http://www.dr-philippe-bludzien-chirurgiens-dentistes.fr/parodontologie-suite.html et Costalonga & Herzberg (2014) 'The oral microbiome and the immunobiology of periodontal disease and caries' *Immunol Lett*, 162, 22-38.
Téléchargée le 24 octobre 2017

phylum, les genres plus communs sont *Prevotella*, *Tannerella* et *Porphyromonas*. Lorsque le sillon s'agrandit et devient de plus en plus profond, une poche parodontale est créée (Sampaio-Maia et al. 2016) **(Figure 35)**.

3.2. Genre *Porphyromonas*, microbiotes et pathologies associées

3.2.1. Classification du pouvoir pathogène des bactéries

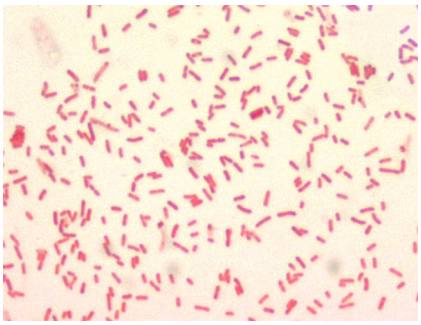
Les bactéries du microbiote sont en relation de mutualisme ou de commensalisme avec l'hôte. Elles sont adaptées et participent à des réseaux écologiques pour acquérir des nutriments. Ces bactéries *bénéfiques* sont capables d'empêcher les bactéries pathogènes de s'installer et de produire des effets négatifs pour l'hôte. Les organismes commensaux peuvent interagir directement ou indirectement avec les pathogènes. La compétition directe est faite soit en produisant des molécules qui inhibent la croissance des pathogènes comme les bactériocines, soit par un changement de l'environnement de l'hôte comme une baisse du pH ou encore la rétention de nutriments essentiels comme le fer ou certains acides aminés. Les bactéries commensales peuvent également produire certains métabolites spécifiques pouvant interférer avec la croissance des pathogènes ou l'action de leurs facteurs de virulence (e.g. le butyrate ou le fucose peuvent diminuer l'expression de facteurs de virulence de *Salmonella enterica* Enteritidis ou Typhimurium et de *Escherichia coli* entérohémorragique). La compétition peut également être indirecte puisque les bactéries commensales du microbiote stimulent le système immunitaire de l'hôte. Elles font partie des barrières muqueuses en occupant les récepteurs épithéliaux et stimulent la production de peptides antimicrobiens et l'immunoglobuline A par le système immunitaire de l'hôte (N. Kamada et al. 2013).

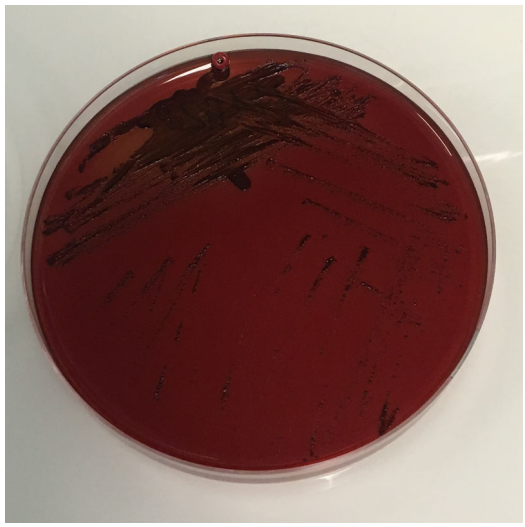
Les facteurs de virulence sont responsables des symptômes causés par les bactéries *pathogènes*. Les exemples classiques sont les toxines bactériennes mais le concept a été élargi pour englober également les facteurs responsables de l'évasion du système immunitaire de l'hôte et ou du dysfonctionnement des fonctions cellulaires et des organes. Par exemple, la capsule bactérienne, les sidérophores, les molécules d'attachement et de tropisme, les enzymes de dégradation de la matrice extracellulaire et des molécules de défense de l'hôte qui contribuent à la propagation de la bactérie ou encore les systèmes d'export et de translocation de protéines effectrices qui manipulent la fonction des cellules de l'hôte sont considérés des facteurs de virulence (Hornef 2015).

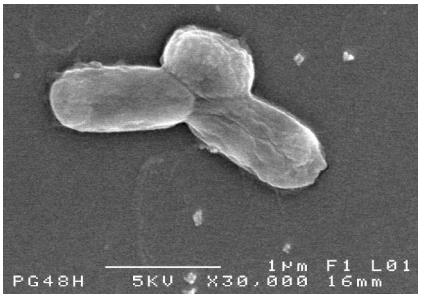
La vision des bactéries comme symbiotes ou pathogènes est simpliste et imprécise. Les progrès de la médecine et l'identification chez une partie de la population d'un certain état d'immunosuppression primaire ou acquise ont alerté sur l'existence d'un groupe de bactéries qui causent des morbidités et même la mortalité chez ces individus ; ce qui n'est pas le cas pour des individus sains. Ces bactéries sont appelées des *pathogènes opportunistes* (Hornef 2015). De même, les bactéries symbiontes sont en constante interaction dynamique avec le système immunitaire et les systèmes de réparation et de renouvellement des épithéliums. Ces interactions "éduquent" le système immunitaire et le préparent à l'interaction avec les pathogènes. Les bactéries spécifiques capables de développer le système immunitaire ou de produire une inflammation sont appelées *pathobionts*, ce qui décrit leur potentiel ambivalent de symbiont et de pathogène. À la différence des pathogènes opportunistes, les pathobionts produisent leurs effets sur l'hôte de manière indirecte via la stimulation du système immunitaire. Les pathobionts sont présents en absence de maladie, mais sous certaines conditions, ils sont capables de produire l'autoimmunité et de promouvoir donc le développement d'une maladie. Un cas de pathobiont cité dans la littérature est la bactérie *Porphyromonas gingivalis* (Hornef 2015).

3.2.2. Espèces du genre *Porphyromonas* et *Porphyromonas gingivalis*

En 1921, W. Oliver et W. Wherry décrivent des bactéries anaérobies de la cavité buccale qui poussent sur gélose au sang en formant des colonies noires et isolées (Oliver and Wherry 1921). Cette espèce appelée *Bacteroides melaninogenicus* comprend des souches asaccharolytiques regroupées dans la sous-espèce *B. melaninogenicus* subesp *asaccharolyticus*, reclassée en 1977 dans la nouvelle espèce *Bacteroides assacharolyticus* (Finegold and Barnes 1977). En 1980, A. Coykendall et collaborateurs démontrent l'existence de deux sous-groupes dans *Bacteroides assacharolyticus*, les souches isolées des cavités orales étant différentes de celles isolées d'autres sources. Ils proposent la création d'une espèce à part, *Bacteroides gingivalis*, pour regrouper les souches orales. La souche Slots 2561 est proposée comme souche type et déposée à l'*American Type Culture Collection* comme la souche ATCC 33277 (Coykendall et al. 1980). En 1988, H. Shah et M. Collins étudient les caractéristiques biochimiques très éloignées de *B. gingivalis* par rapport à l'espèce type du genre, *B. fragilis*, et proposent la création d'un nouveau genre bactérien, *Porphyromonas* dans lequel ils reclassent *B. gingivalis*, *B. asaccharolytica* et *B. endodontalis* (Shah and Collins







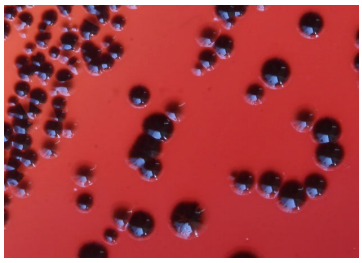


Figure 36. ***Porphyromonas gingivalis*** **: microscopie et colonies sur gélose.** À gauche, en haut, coloration de Gram ; en bas, microscopie électronique à balayage. À droite, bactérie en culture pure sur gélose au sang supplémenté pour anaérobies ; en bas, agrandissement pour apprécier mieux les colonies.

Adapté de : https://www.researchgate.net/figure/270107715_fig1_Figure-1-The-pigmentation-of-Porphyromonas-gingivalis-colonies-on-blood-agar
Téléchargée le 24 octobre 2017

1988). Le nom du genre vient de l'adjectif grec *porphyreos* (pourpre) et du nom grec *monas* (unité) ce qui veut donc dire cellule à porphyrine, du fait que ces bactéries accumulent de la protoporhyrine IX (protohème) ce qui donne aux colonies, sur gélose, leur coloration noire caractéristique **(Figure 36)**.

L'espèce type du genre est *Porphyromonas asaccharolytica* et, n'incluait à l'origine que deux autres espèces, *P. gingivalis* et *P. endodontalis* (Shah and Collins 1988). Actuellement, les espèces reconnues dans ce genre sont au nombre de 15 et correspondent à : *P. asaccharolytica* (isolée de plusieurs infections humaines), *P. bennonis* (isolée d'infections humaines), *P. cangingivalis* (isolée de poches parodontales de chiens sains ou atteints de parodontite), *P. canoris* (isolée de poches parodontales de chiens atteints de parodontite), *P. catoniae* (isolée de poches parodontales d'humains sains ou atteints de parodontite), *P. circumdentaria* (isolée d'infections de tissus mous et de gencives de félins), *P. crevioricanis* (isolée de poches parodontales et du fluide gingival de chiens atteints de parodontite ; synonyme de *P. cansulci*), *P. endodontalis* (isolée de plusieurs sites oraux humains), *P. gingivalis* (isolée de plusieurs sites oraux humains, principalement des poches parodontales), *P. gingivicanis* (isolée du fluide gingival de chiens ; synonyme de *P. canis*), *P. gulae* (isolée de la plaque sous-gingivale de plusieurs mammifères, principalement canidés, félins et primates), *P. levii* (isolée de plusieurs infections non orales de bovins), *P. macacae* (isolée de cavités orales de chiens, chats et singes, pathogènes important des infections par morsure d'animal ; un ancien synonyme de cette espèce était *P. salivosa*), *P. somerae* (isolée de plusieurs infection non orales) et *P. uenonis* (isolée de plusieurs infections polymicrobiennes intestinales) (Krieg et al. 2010; Sakamoto and Ohkuma 2013).

Les conclusions qui peuvent être faites de cette énumération sont : plus de deux tiers des espèces de *Porphyromonas* sont orales qui sont, elles mêmes, principalement isolées de chiens comme l'indique le nom des espèces (*P. cangingivalis*, *P. canoris*, *P. cansulci*, *P. crevioricanis*, *P. gingivicanis*), les autres espèces étant félines (*P. circumdentaria*), ou isolées de différentes sources canines, félines et primates (*P. macacae*) ou plus généralistes mais dominées par les canidés (*P. gulae*) et finalement trois espèces humaines (*P. catoniae*, *P. endodontalis* et *P. gingivalis*).

L'espèce la plus étudiée au sein de ce genre est *Porphyromonas gingivalis*. Il s'agit de la principale bactérie du genre liée aux pathologies humaines dont l'espèce la plus proche est

Porphyromonas gulae qui partage notamment certains facteurs de virulence (do Nascimento Silva et al. 2017; O'Flynn et al. 2015; Oishi et al. 2012).

3.2.3. Maladies systémiques, locales et lien avec l'oncogenèse

Porphyromonas gingivalis a été décrite comme une bactérie importante dans plusieurs maladies humaines. Son habitat naturel est la bouche, principalement les sites sous-gingivaux, mais elle peut être isolée, moins fréquemment certes, de la salive et des muqueuses orales. Ses facteurs de virulence incluent des peptidases comme les gingipaïnes, les collagenases et autres protéases, la capsule, les fimbriae, les hémagglutinines et son lipopolysaccharide (do Nascimento Silva et al. 2017).

La principale maladie associée à *Porphyromonas gingivalis* est la parodontite, une maladie locale au niveau de la gencive. En plus de ces manifestions locales, plusieurs études épidémiologiques montrent un lien entre cette bactérie et plusieurs maladies systémiques comme la polyarthrite rhumatoïde, des maladies cardiovasculaires, le diabète et différents types de cancers orodigestifs (Atanasova and Yilmaz 2014). Tant la parodontite comme le diabète sont des maladies chroniques fréquentes. Les individus atteints de diabète, surtout ceux qui ont une glycémie mal contrôlée, ont une parodontite sévère et généralisée. *P. gingivalis* est plus fréquente et en quantité plus importante chez les patients diabétiques comparé aux personnes non diabétiques (Aemaimanan et al. 2013). Concernant la polyarthrite rhumatoïde, il s'agit d'une maladie inflammatoire chronique caractérisée par la présence d'auto-anticorps anti-protéines citrullinées (*anticitrullinated protein antibodies* ou ACPA). *P. gingivalis* possède une peptidyl-arginine deiminase (PAD), enzyme qui produit des peptides citrullinés *in vivo*. Une association existe entre la présence d'anticorps anti-*P. gingivalis* et des ACPA (Hitchon et al. 2010). D'ailleurs l'ADN de la bactérie a été retrouvé dans les articulations des patients atteints de polyarthrite rhumatoïde (Totaro et al. 2013). Finalement, une association positive entre parodontite et maladie cardiovasculaire est connue. Des études *in vitro*, avec des animaux et des observations cliniques montrent que des bactéries paropathogènes, dont *P. gingivalis*, influencent des mécanismes majeurs dits proatherogéniques et peuvent donc jouer un rôle dans la pathogénèse de l'athérosclérose (Chistiakov et al. 2016).

Comme déjà dit, l'implication la plus forte de *P. gingivalis* dans une maladie concerne la parodontite. La parodontite est une maladie des tissus de soutien de la dent, le paradonte (la gencive, le cément, le ligament alvéolo-dentaire et l'os alvéolaire). Les maladies parodontales commencent par une gingivite, une inflammation localisée de la gencive qui est initialisée par la plaque dentaire. La parodontite chronique est le résultat d'une gingivite non traitée et progresse vers la perte de la gencive, des ligaments et une destruction de l'os alvéolaire. Ceci entraine l'apparition d'une poche parodontale au niveau du sillon gingivo-dentaire (ou sulcus), zone d'inflammation accompagnée d'une dysbiose correspondant à une diminution des bactéries symbiotiques au profit des paropathogènes (Kinane et al. 2017). *P. gingivalis* y est présente en faible quantité mais peut "orchestrer" une inflammation via l'activation du système du complément de l'hôte et ses interactions avec les bactéries du microbiote oral. L'infection de modèles animaux dits *gnotobiotiques* ou encore libres de microorganismes (*germ-free* ou GF) avec *P. gingivalis* ne produit pas de parodontite, mais son inoculation dans les modèles animaux libres de pathogènes (*specific-pathogen-free* ou SPF) produit des symptômes typiques de cette maladie. Cela veut dire que *P. gingivalis* n'est pas capable à elle seule de produire une parodontite, mais qu'elle est un des acteurs responsables de son apparition. De plus, son introduction dans les bouches saines, produit une augmentation de la charge bactérienne dans la gencive et change qualitativement le microbiote (Hajishengallis et al. 2011). À cause de ces caractéristiques, *P. gingivalis*, bactérie capable d'entraîner un changement du microbiote et de stabiliser ce microbiote dysbiotique associé à une maladie, est nommée *pathogène clé de voûte* (*keystone pathogen*) (Hajishengallis et al. 2012).

Malgré l'implication de la dysbiose pour l'initiation de la parodontite, la réponse inflammatoire de l'hôte est celle qui produit les lésions irréversibles du parodonte ce qui entraîne la perte des dents. Encore une fois, *P. gingivalis* est capable d'entretenir cette inflammation par plusieurs mécanismes d'évasion de la réponse immunitaire. Les gingipaïnes peuvent avoir une activité de catalyseur de l'activation du système de complément de l'hôte. En même temps, la bactérie active les leucocytes et augmente l'inflammation, mais réduit leur capacité de tuer les bactéries. Une serine phosphatase sécrétée, SerB, est capable d'inhiber la production d'interleukine 8 (IL-8) ce qui diminue le recrutement des neutrophiles dans la gencive. L'élimination spécifique de *P. gingivalis* est capable de retourner le microbiote dysbiotique à son état antérieur (Hajishengallis et al. 2012). *P. gingivalis* est donc un paropathogène clé de voûte mais également un pathobiont par sa manipulation du système immunitaire de l'hôte.

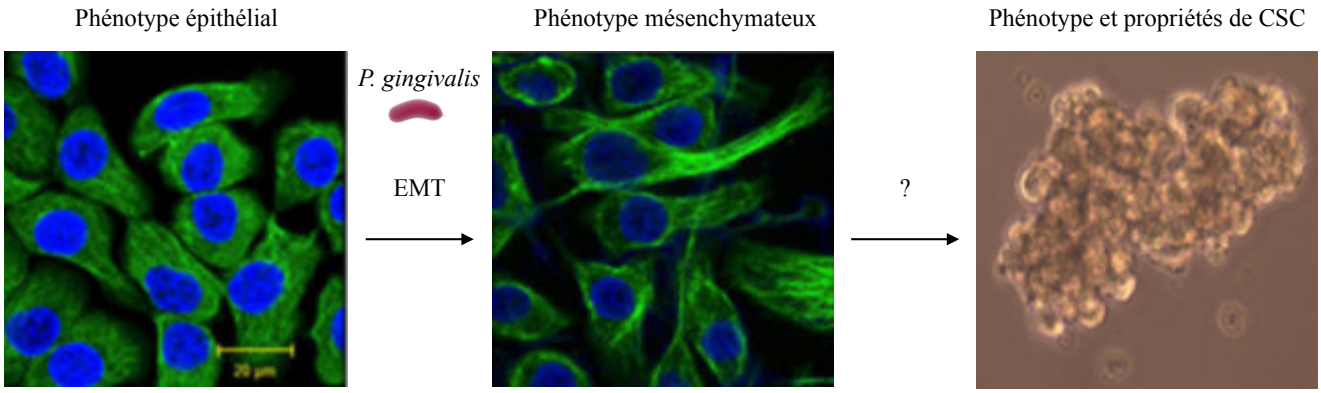


Figure 37. *Porphyromonas gingivalis*, EMT et apparition de CSC. *In vitro*, une co-culture avec *Porphyromonas gingivalis* produit un changement de phénotype des cellules épithéliales de gencive vers un phénotype mésenchymateux avec une diminution des marqueurs épithéliaux et une augmentation des mésenchymateux caractéristique d'une transition épithélio-mésenchymateuse (EMT). Cette EMT confère des propriétés de cellules souches cancéreuses (CSC) aux cellules exposées à la bactérie, par un mécanisme encore inconnu.

Adapté de : Ha et al. 2015 et Sztukowska et al. 2016
Téléchargée le 24 octobre 2017

Une autre caractéristique intéressante de cette bactérie, et de la maladie qu'elle est capable de produire, est son implication de plus en plus convaincante dans l'initiation et le développement du cancer. Un cancer oral nommé carcinome à cellules squameuses gingivales (*oral squamous cell carcinoma* ou OSCC) est parmi les dix premiers cancers les plus communs au monde et possède une présentation clinique similaire aux états avancés de la maladie parodontale : inflammation, saignements, poches parodontales profondes, destruction osseuse et mobilité des dents (Katz et al. 2011). Plusieurs études sur ce cancer identifient la présence de *P. gingivalis* : dans des biopsies de OSCC (Katz et al. 2011), comme stimulant *in vitro* la prolifération de cellules d'OSCC (Binder Gallimidi et al. 2015), ou comme inhibant l'apoptose des cellules via des cascades d'activation souvent citées en oncogènese (JAK1/STAT3 et PI3K/Akt) (Perera et al. 2016).

La *transition épithélio-mésenchimateuse* (EMT) est un phénomène qui conduit une cellule épithéliale polarisée à perdre ses caractéristiques épithéliales et à acquérir un phénotype de cellule mésenchymateuse. Il s'agit d'un processus normal lors du développement embryonnaire et de la cicatrisation, mais qui devient pathologique lors de la fibrose et du cancer. Ce changement de phénotype est caractérisé par un remaniement du cytosquelette, l'augmentation des capacités de migration et d'invasion, une résistance élevée à l'apoptose et l'expression de certains marqueurs spécifiques. Cette reprogrammation est faite par certains facteurs de transcription spécialisés comme Snail, Twist et Zeb (Lamouille et al. 2014; Voon et al. 2013). L'inflammation chronique est un facteur prédisposant à l'EMT et ce phénomène a déjà été décrit lors de la fibrose des gencives par certaines drogues (Ha et al. 2015; Sume et al. 2010). L'EMT est un processus désormais reconnu comme participant à l'initiation et à la progression tumorale et qui confère des propriétés de cellules souches cancéreuses aux cellules épithéliales différenciées. Les *cellules souches cancéreuses* (CSC) sont une sous-population minoritaire des cellules d'une tumeur qui a conservé ou acquis les propriétés d'auto-renouvellement et de division asymétrique des cellules souches (Singh 2013).

Au moins deux équipes montrent l'implication de *P. gingivalis* dans des processus d'EMT et d'apparition de CSC **(Figure 37)**. Une équipe de Corée du Sud a été la première à publier que la co-culture de cellules d'OSCC avec *P. gingivalis* produit une internalisation des bactéries dans les cellules et induit des changements morphologiques via une EMT, l'acquisition de capacités de migration et d'invasion, la résistance aux agents de chimiothérapie et des propriétés de CSC (Ha et al. 2015). Une collaboration d'équipes des États-Unis et du Royaume-Uni a confirmé ces résultats en utilisant des kératinocytes

inmortalisées dérivés de l'épithélium gingival et ont observé une augmentation du facteur de transcription ZEB1, entraînant une EMT et des capacités de migration des cellules. La bactérie est également capable d'augmenter l'expression de ZEB1 *in vivo* (Sztukowska et al. 2016).

Tous ces résultats sont similaires à ceux obtenus pour une autre bactérie, *Helicobacter pylori* et sa relation avec le cancer gastrique qui est actuellement reconnue par toute la communauté scientifique comme un acteur important de l'oncogenèse digestive. Au cours de mes études de Master à l'Université de Bordeaux entre 2012-2014, j'ai participé à la description de l'EMT induite par *H. pylori* sur des cellules épithéliales gastriques et à l'apparition de cellules avec des propriétés de CSC (Bessede et al. 2014), et également à la mise en évidence d'un mécanisme moléculaire qui conduit à l'induction de la dysplasie gastrique *in vivo* et l'acquisition de propriétés de CSC *in vitro* (Bessede et al. 2016). Ceci ouvre des possibilités énormes d'étudier *P. gingivalis* et des espèces très proches (comme *P. gulae*) dans le contexte de carcinogénèse et il semble pertinent de se poser la question : est-ce que *Porphyromonas gingivalis* est l'*Helicobacter pylori* de la bouche ?

Objectifs de cette thèse

Comme toute thèse scientifique, deux types d'objectifs co-existent dans les travaux qui seront présentés par la suite.

Tout d'abord, le premier concerne mes objectifs personnels de formation. Je suis microbiologiste, formé aux analyses biomédicales des agents infectieux, à la biochimie clinique, à l'hématologie et l'immuno-hématologie. Au Costa Rica, mon pays natal, cet éventail de connaissances sont proposées aux étudiants comme toutes les autres formations des sciences de la santé (médecine, pharmacie, odontologie), dans une formation à part, appelée "Microbiología y Química Clínica". C'est ma formation initiale. J'ai décidé de poursuivre mes études pour devenir chercheur en microbiologie. Après un master en Microbiologie-Immunologie à l'Université de Bordeaux, j'ai obtenu une bourse de l'Universidad de Costa Rica pour me former en bioinformatique lors d'un doctorat. Mon objectif principal était d'acquérir des compétences en bioinformatique. Je me suis rapidement

aperçu que ce domaine est très vaste et j'ai choisi de concentrer mes efforts en génomique et en biocuration. L'intérêt pour mon futur employeur était le même, et je retourne dans mon pays avec des connaissances que je n'aurais pas pu obtenir sur place et formé pour pouvoir développer des projets de recherche et participer à la formation des nouveaux microbiologistes.

D'un point de vue scientifique, les objectifs de cette thèse sont:

- Décrire le panorama des génomes bactériens disponibles dans les bases de données
- Expliquer les raisons de l'incomplétude de ces génomes, notamment au sein du groupe des Bacteroidia
- Identifier les causes de la plasticité génomique (ou mosaïcité) de *Porphyromonas gingivalis*
- Analyser les différences génomiques qui pourraient expliquer les différences de virulence des souches de *Porphyromonas gingivalis*, notamment leurs facteurs de virulence
- Décrire, pour un de ces facteurs de virulence, les fimbriae, la répartition au sein du genre *Porphyromonas* et faire le lien avec la phylogénie du clade
- Décrire la dysbiose liée à la parodontite en réutilisant les données publiques pour distinguer les microbiotes des sites sains par rapport à ceux de sites atteints
- Reconnaître l'importance de la biocuration et sa valeur ajoutée dans les travaux de génomique

Résultats

1. ARTICLE 1: Répétitions génomiques et erreurs d'assemblage : étude du re-séquençage en long read des souches de référence de *Porphyromonas gingivalis*

Genomic repeats and misassembly: a case study with long-read resequencing of *Porphyromonas gingivalis* reference strains

In press à BMC Genomics 2018

La capacité actuelle de séquençage et la volonté de description et de compréhension des microbiotes ont permis à plusieurs équipes à travers le monde de détailler les bactéries associées aux humains. Le mot *microbiote* est utilisé même dans des revues de vulgarisation scientifique et est associé à plusieurs états de santé. Les résultats attendus de ces études sont énormes, de la même manière que l'étaient ceux des travaux sur le séquençage du génome humain.

Pour comprendre les fonctions associées à certains microbiotes, il faut pouvoir extrapoler les potentialités métaboliques des organismes présents et pour le faire avec précision, il est nécessaire d'identifier le plus précisément possible les espèces (voire les souches) présentes au sein de la communauté étudiée et de connaître leurs génomes. Sans les séquences génomiques, les capacités de compréhension des interactions entre bactéries et hôte sont très limitées, incomplètes voire incorrectes.

La grande majorité des phyla présents dans les microbiotes humains et de mammifères sont les Bacteroidetes, Firmicutes, Actinobacteria et Proteobacteria. Se limitant au tube digestif, les Bacteroidetes et Firmicutes représentent près de 90% de toutes les bactéries présentes. Au niveau de la bouche, la répartition de phyla présents est moins déséquilibrée sauf en bactéries anaérobies qui sont présentes principalement au niveau du sillon gingivo-dentaire.

Le premier article présenté dans ce manuscrit concerne une étude génomique des Bacteroidetes, phylum important dans les microbiotes humains, tant en nombre (pourcentage d'individus présents) qu'en terme de capacités physiologiques. Dans un premier temps, nous avons fait un état des lieux descriptif des génomes disponibles dans les bases de données pour ce phylum et nous les avons comparé aux autres phyla. Ainsi, nous avons pu noter que les génomes des Bacteroidetes ne représentent que 2% de tous les génomes de bactéries disponibles, et ceci malgré leur prépondérance dans les microbiotes humains et dans toutes les niches écologiques présentées en introduction de ce document. De plus, à ce manque de génomes s'ajoute un très faible niveau de complétude de ceux-ci, puisque de tous les génomes disponibles, à peine 15% sont des génomes finis (complets), la vaste majorité correspondant à des génomes "en brouillon" (drafts) composé d'une collection de séquences contiguës (contigs), et rendant difficiles les études de génomique comparative.

La raison de cette accumulation de génomes drafts est principalement le résultat d'une contradiction dans la conception de nombreux projets de recherche qui, tentés par la facilité de séquencer à des prix toujours plus bas, se lance dans des séquençage massif de génomes microbiens sans expertise en génomique et/ou en bioinformatique et sans investir, en temps et en personnel, dans les étapes pourtant très importantes de finition et de biocuration des génomes. Ces étapes sont pourtant cruciales puisqu'elles valident, orientent et associent correctement les contigs générés automatiquement par les logiciels d'assemblage et permettent de pointer sur des expériences complémentaires à réaliser pour valider et confirmer la reconstruction du génome étudié.

La valeur des génomes draft est insuffisante parce que la certitude associée à leur assemblage est faible. De plus, leur utilité en génomique comparative reste réduite. Ils sont difficilement utilisables pour étudier l'architecture du génome, les gènes qui sont annotés peuvent l'être de façon incorrecte (par transfert automatique d'annotations erronées) et l'orféome proposé est souvent incomplet avec des gènes séparés, mal assemblés ou tout simplement non reconstruits et donc interprétés à tort comme absents du génome. Même si les drafts peuvent être utilisés pour analyser le transfert horizontal de gènes, la phylogénomique, l'évolution de la synténie du génome et les analyses de pangénome (tous les gènes présents dans des souches proches et qui pourrait expliquer des différences phénotypiques entre elles), les conclusions sont moins solides qu'avec des génomes complets.

Un autre résultat surprenant de cet article est la grande variabilité du nombre de contigs par génome draft que ce soit dans un même genre bactérien ou au sein d'une même espèce. Le nombre de variables qui peuvent impacter l'assemblage du génome est grand et inclut des considérations de laboratoire comme l'extraction, la purification et la fragmentation de l'ADN, sans parler des possibles contaminations. D'autres raisons permettant d'expliquer la variabilité des ces drafts proviennent des méthodes de séquençage, d'assemblage (logiciel utilisé, paramètres utilisés) et finalement du génome lui même (pourcentage de GC, taux de répétitions nucléiques, taux de paralogie).

Les métadonnées associées aux projets de séquençage des génomes disponibles sur la base de données du NCBI sont incomplètes et mal renseignées, les reads de ces projets sont très rarement mis à disposition. Nous nous sommes efforcés de faire une revue approfondie des métadonnées des génomes disponibles en draft, mais nous ne sommes pas en capacité de pointer sur une raison particulière pour expliquer la variation des génomes draft, mais nous pouvons éliminer un certain nombre de possibilités. Par exemple, l'analyse des souches de *Bacteroides fragilis* est assez démonstrative. La grande majorité des souches (99 sur 107) ont été séquencées avec une seule technologie (Illumina) et plus des deux tiers (69 sur 99) assemblés avec un seul logiciel (MaSuRCA). Pourtant, la variabilité du nombre de contigs est énorme, de 31 à 2566 selon les drafts. L'autre espèce avec un nombre suffisant de drafts pour être analysés est *Porphyromonas gingivalis* avec 17 génomes sur 20 séquencés en Illumina et assemblés avec le logiciel Velvet. Comme précédemment, la variabilité est remarquable, allant de 22 à 192 contigs par draft. Les hypothèses sur la technologie de séquençage et le logiciel d'assemblage semblent donc insuffisants pour expliquer ces variabilités et nous avons donc cherché l'impact des caractéristiques intrinsèques aux génomes.

La principale raison invoquée dans la littérature pour expliquer l'incapacité des logiciels d'assemblage à générer des génomes complets à partir des reads de séquençage est la présence de séquences génomiques répétées. Ces répétitions, quand elles sont de taille supérieure ou égale à la taille du read, créent des problèmes combinatoires qui se concluent généralement par un arrêt de l'assemblage et une interruption du contig en cours de reconstruction, provoquant des assemblages morcelés. Si ces répétitions sont souvent désignées dans la littérature comme responsables de la fragmentation des génomes, les démonstrations sont rares voire absentes. En effet, pour prouver cette hypothèse, il faut connaître le nombre de répétions du génome avant de l'avoir assemblé. Afin de prouver si ce sont bien ces répétitions qui sont responsables, nous nous sommes intéressés aux génomes

complets de différentes souches de la même espèce. Dans un premier temps, nous avons détecté le nombre de répétitions génomiques d'au moins 500 bp (avec un minimum d'identité à 95%) dans les génomes complets des Bacteroidia. Nous avons également observé leur disposition dans le génome et leur nombre de copies. En utilisant ces critères, nous avons pu montrer que *P. gingivalis* est l'espèce avec le plus de répétitions, en nombre et en fréquence et que ces répétitions se distribuent de manière homogène mais aléatoire le long des génomes. Nous avons donc choisi de poursuivre notre étude en nous focalisant plus en détail sur cette espèce microbienne.

Pour cela, nous avons utilisé une stratégie qui permet de simuler des reads de séquençage à partir de génome complet, en mimant les résultats obtenus lors d'un séquençage Illumina MiSeq de type paired-end et nous avons identifié les éléments génomiques contenus dans les interruptions d'assemblage (gaps) en mappant les contigs générés sur les séquences complètes de chaque génome. Les résultats cette expérience, décrit dans l'article 1, montrent une corrélation linéaire entre le nombre de contigs générés à partir d'un séquençage *in silico* et le nombre de répétitions génomiques. De même, nos résultats mettent en évidence que la mesure de la qualité d'un assemblage par la taille des contigs générés (N50) est une mesure qui peut être trompeuse et donc à éviter. Finalement, les gaps dans l'assemblage draft des reads artificiels nous a permis d'identifier, pour *P. gingivalis*, les séquences responsables des échecs des assembleurs, classés en Séquences d'Insertion (IS) et Miniature Inverted-repeat Transposable Element (MITE), des séquences intergéniques multicopies, des répétitions dans les régions codantes de gènes paralogues (motifs internes répétés), les opérons ribosomiques et certains îlots génomiques dupliqués. Les deux catégories les moins fréquentes, les opérons ribosomiques et les îlots génomiques sont celles de plus grande : environ 5.5 Kb pour chacun des quatre opérons ribosomiques de *P. gingivalis* et jusqu'à 145 Kb pour certains îlots génomiques comme CnTPg1.

Nous avons donc choisi de réaliser un séquençage de type *long-read* ou de troisième génération pour vérifier si cette technologie permet de résoudre le problème de l'assemblage des répétitions. Cette expérience, réalisée sur les trois souches de référence de *P. gingivalis*, montre que même en sélectionnant des fragments de 10Kb pour séquencer, la taille médiane des reads est plutôt proche de 6 Kb. Toutefois, cette taille permet de réduire considérablement le nombre de contigs obtenus lors de l'assemblage. En complément et afin d'obtenir des génomes complets, nous avons élaboré une nouvelle stratégie d'utilisation des reads par le biocurateur et permettant de limiter la validation par biologie moléculaire aux gaps de taille

supérieure à 6 Kb : les opérons ribosomiques pour les trois souches re-séquencées et un transposon conjugatif dupliqué pour l'une d'entre elles. Nous avons eu la surprise, en comparant les génomes que nous avons re-séquencés, re-assemblés et vérifiés et ceux d'origine, d'identifier les erreurs suivantes :

- Le mauvais assemblage d'origine pour le génome de la souche TDC60 avec trois des quatre opérons ribosomiques et plusieurs protéines avec des domaines répétés mal combinés.

- Une inversion centrale au niveau du génome de la souche W83 et une erreur sur la région CRISPR qui est finalement plus longue qu'initialement publiée (plus de spacers et de direct repeats).

- La reconstruction, pour la souche ATCC 33277, de deux transposons conjugatifs complets et orientés dans la même direction alors que la version d'origine contient une copie incomplète et orienté dans le sens contraire de la copie complète.

- Après la correction de ces erreurs d'assemblage, nous avons également entrepris une annotation biocurée manuellement, avec plusieurs améliorations notamment dans la description des ncRNA et des CDS. Nous avons corrigé la sur-annotation des génomes et avons désannonymé de très nombreux CDS initialement décrits de fonction inconnue ou hypothétique en prédisant des domaines fonctionnels et modélisant des localisations cellulaires. Ce travail a permis de réduire de plus de la moitié le pourcentage de CDS sans aucune description (de 22% initialement à 9%).

- L'article se termine par une analyse du pangénome, en comparant, non pas les séquences via des *best bidirectional hits* mais les annotations (notamment le nom des gènes) que nous avons rendues homogènes pour les trois souches et donc facilement comparables. Ce travail a permis de prouver que :

- Le compartiment de gènes uniques et accessoires est assez réduit.

- Près de 10% des gènes sont répétés en plus de deux copies dans chacune des trois souches.

- 85% des gènes sont dans le core génome, répertoire commun que nous avons proposé de diviser en un core constant (gènes à plus de 97% d'identité nucléique), un core

variant (moins de 97% d'identité nucléique), un core avec des pseudogènes dans au moins une des trois souches et finalement un core avec un phénomène de paralogie dans au moins une des trois souches. Encore une fois, la partie la plus importante de ces 4 core-génomes est le core constant qui représente 93% l'ensemble.

En résumé, l'article 1 démontre :

1. les problèmes liés à l'accumulation de génomes drafts
2. la sur-accumulation de ces génomes depuis l'arrivée des NGS
3. la corrélation entre le nombre de contigs générés et le nombre de répétitions génomiques
4. le risque de l'utilisation de la métrique N50
5. l'existence d'erreurs d'assemblage dans des génomes de référence, pourtant circularisés
6. l'intérêt du séquençage long-read
7. l'importance de la validation par PCR des longues répétitions
8. le gain, en génomique comparative, d'une annotation biocurée et homogénéisée

Finalement, nous montrons que la plasticité génomique de l'espèce *P. gingivalis* est plus liée à des remaniements par recombinaison intra-génomiques (mosaïcité) qu'à des acquisitions de gènes par transfert horizontaux (même s'ils existent). Ces résultats ouvrent des perspectives pour nos travaux futurs, notamment sur les gènes présents dans le génome core variable qui sont, en grande partie, impliqués comme des gènes de virulence. Une première étude sur les fimbriae, structures d'adhésion, est présentée comme deuxième article de ce document.

Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of *Porphyromonas gingivalis* reference strains

Luis Acuña-Amador,[1,2] Aline Primot,[1] Edouard Cadieu,[1] Alain Roulet,[3] and Frédérique Barloy-Hubler[1*]

1. Institut de Génétique et Développement de Rennes, CNRS, UMR6290, Université de Rennes 1, Rennes, France
2. Laboratorio de Investigación en Bacteriología Anaerobia, Centro de Investigación en Enfermedades Tropicales, Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica
3. GenoToul Genome & Transcriptome (GeT-PlaGe), INRA, US1426, Castanet-Tolosan, France

* Corresponding author: fhubler@univ-rennes1.fr

Authors' e-mail addresses: luis.acuna-amador@univ-rennes1.fr; aline.primot@univ-rennes1.fr; edouard.cadieu@univ-rennes1.fr; alain.roulet@toulouse.inra.fr; fhubler@univ-rennes1.fr

Abstract

Background: Without knowledge of their genomic sequences, it is impossible to make functional models of the bacteria that make up human and animal microbiota. Unfortunately, the vast majority of publicly available genomes are only working drafts, an incompleteness that causes numerous problems and constitutes a major obstacle to genotypic and phenotypic interpretation. In this work, we began with an example from the class Bacteroidia in the

phylum Bacteroidetes, which is preponderant among human orodigestive microbiota. We successfully identify the genetic loci responsible for assembly breaks and misassemblies and demonstrate the importance and usefulness of long-read sequencing and curated reannotation.

Results: We showed that the fragmentation in Bacteroidia draft genomes assembled from massively parallel sequencing linearly correlates with genomic repeats of the same or greater size than the reads. We also demonstrated that some of these repeats, especially the long ones, correspond to misassembled loci in three reference *Porphyromonas gingivalis* genomes marked as circularized (thus complete or finished). We prove that even at modest coverage (30X), long-read resequencing together with PCR contiguity verification (*rrn* operons and an integrative and conjugative element or ICE) can be used to identify and correct the wrongly combined or assembled regions. Finally, although time-consuming and labor-intensive, consistent manual biocuration of three *P. gingivalis* strains allowed us to compare and correct the existing genomic annotations, resulting in a more accurate interpretation of the genomic differences among these strains.

Conclusions: In this study, we demonstrate the usefulness and importance of long-read sequencing in verifying published genomes (even when complete) and generating assemblies for new bacterial strains/species with high genomic plasticity. We also show that when combined with biological validation processes and diligent biocurated annotation, this strategy helps reduce the propagation of errors in shared databases, thus limiting false conclusions based on incomplete or misleading information.

Keywords

Porphyromonas gingivalis, Bacteroidetes, long-read sequencing, misassembly, genomic repeats, annotation, biocuration, comparative genomics

Background

Pioneer studies such as the MetaHIT consortium [1] and the Human Microbiome Project [2] have used high-throughput sequencing techniques to produce a detailed catalog of human-associated bacterial taxa. The vast majority of prokaryotes identified in and on human beings belong to only four phyla: Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria [3]. To date, the best-described human microbial community is the gut microbiota [4], which is mostly (~90%) composed of members of the phyla Bacteroidetes and Firmicutes [5, 6]. Among these, the dominant classes are the strict anaerobes Bacteroidia and Clostridia, respectively [7, 8]. Although their relative proportions may vary [9, 10], Bacteroidetes make up approximately 50% of the gut microbiome [11].

The phylum Bacteroidetes (or "Bacteroidaeota" as recently proposed [12]), is highly diverse, and its phylogenetics has been well explored [13-15]. The Bacteroidia are Gram-negative chemoorganotrophic rod-shaped organisms. Either non-motile or moving by gliding, they have colonized several ecological niches, including soil, oceans, fresh water and the abovementioned gastrointestinal tract [8]. Their genomes can undergo massive reorganization, with extensive and frequent horizontal gene transfers (HGTs), and the sizes of their genomes correlate with their functional specialization [14]. The class Bacteroidia includes some commensal genera that can present as opportunistic pathogens, such as the intestinal *Bacteroides* and the oral *Prevotella*, *Porphyromonas,* and *Tannerella* [8, 14, 15].

Approximately 10 years ago, the technological breakthrough known as next-generation sequencing (NGS, now also called SGS for second-generation methods or massively parallel sequencing) exponentially lowered sequencing costs, making these techniques widely accessible [16]. In recent years, massively parallel sequencing has almost entirely been conducted using Illumina's MiSeq and HiSeq platforms [17, 18]. Prior to these developments, most whole-genome sequencing (WGS) projects were conducted on organisms

that were selected due to their relevance to medicine or biotechnology [19], resulting in a strongly biased portrait of microbial diversity [20]. Researchers such as those involved in the Genomic Encyclopedia of Bacteria and Archaea (GEBA) are currently attempting to compensate for this by sequencing at least one species from each known genus [19].

WGS generates primary information and a catalog of reference genomes. The associated biological information is typically stored in online databases and used for downstream purposes such as comparative genomics, transcriptomics, and proteomics [16, 21-23]. For genome assembly using massively parallel methods, computation time and memory efficiency have led to the use of algorithms based on de Bruijn graphs [24]. In this method, a genome is constructed using graphs, but if the assembly software encounters a genomic repeat that is equal to or longer than the read length, it either continues the assembly by guessing (which can create false joins) or breaks it, leaving repeat-induced gaps [25-27]. These assembly breaks are common because genomic repeats in bacteria account for 5 to 10% of the total genome. Their frequency is variable but not adaptively neutral since repeated DNA sequences are involved in protein-DNA interactions, bacterial immunity, specialization, speciation and transcriptional regulation [28, 29].

The expected outcome of WGS is the complete DNA sequence of a genome. An initial draft sequence is usually obtained in a matter of days and then completed through genome finishing, a cost-intensive process that may require months or even years [30]. For some bacteria, assembly finishing may be the most important step in genome sequencing, and there are at least three strategies for gap closure and assembly validation. The first of these strategies is reference-assisted gap closure, which consists of arranging contigs into a putative chromosome using information from a closely related complete genome [31, 32]. This assumes that the published reference sequences are accurate [33], biologically validated and closely phylogenetically related [34]. The second strategy involves long read (re)sequencing

using either mate-pair libraries [35] or long-read techniques such as PacBio [36] and Nanopore [37]. The third strategy is based on genome maps [38, 39]. These methodologies incur additional cost and require significant time investments. In fact, in a WGS project, assembly finishing comprises over 95% of the total cost and timeframe. Therefore, researchers often decide that finishing is not cost-effective, preferring to publish the draft genomes, which abound in the databases [40]. The accumulation of numerous draft genomes creates massive databases; unfortunately, the quality of the genomes is decreased since it is based on incomplete and/or incorrect genomic data [41, 42]. Indeed, draft genome assemblies are not sufficient for studying large-scale genome architecture [43], can result in incomplete or incorrectly annotated genes (e.g., partitioned annotation of ORFs), and may hinder evolutionary studies [44, 45] by leaving significant portions of the genome unclear or inaccurate [17]. Furthermore, some studies have demonstrated the limitations and difficulties associated with using drafts for studies involving HGT analysis, phylogenomics, the evolution of genome synteny, genome structural analysis, and pangenomic approaches [34, 42, 46, 47].

In this study, we focused on Bacteroidetes genomes because of the importance of this phylum in microbiota communities but also because this clade contains genomes that are mostly available as drafts. We attempted to identify the reasons for this incompleteness. We first explored the number of published genomes in Bacteria and determined the percentage of genomes released as drafts. We then studied species in which at least two strains have been sequenced, since this is required for comparative analysis of their repeats. We followed this with *in silico* sequencing to elucidate the links between assembly and repetitions. Finally, we narrowed in on the genomic diversity of an interesting model in this clade, *Porphyromonas gingivalis*. Using three principal reference strains, we combined long-read resequencing and *de novo* assembly with manually biocurated annotations. In this way, we produced correct and

consistent genome sequences that are valuable resources for future comparative genomic studies.

Methods

Analysis of the NCBI genome database

On 3 April 2017, we accessed all data through 2016 in the NCBI genome database (https://www.ncbi.nlm.nih.gov/genome/browse). For analysis purposes, we considered the "FCB Bacteroidetes/Chlorobi," "Terrabacteria Firmicutes," "Terrabacteria Actinobacteria," and "All Proteobacteria" subgroups to represent the phyla Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria, respectively. The levels "Complete" and "Chromosome" were considered finished or complete genomes, and "Scaffolds" and "Contigs" were considered draft or incomplete genomes. Analysis was performed using R (v3.3.2) [48] and in-house parsing scripts in Python (v2.7.10). We excluded all entries marked “Candidatus” and entries that did not include full species identification specifying both the genus and the species.

Analysis of Bacteroidetes genomes

Bacteroidetes genomes were categorized by class and sequencing status. The sequences of all copies of genes encoding 16S rRNA were extracted from the complete genomes. Alignment was performed using the MAFFT (v7.222) plug-in [49] in Geneious (v10.2.3) [50] set to the following parameters: automatic detection for the algorithm; a scoring matrix of 200 PAM/k = 2; a gap opening penalty of 1.53; and an offset value of 0.123. A consensus sequence was then created for each species using the default Geneious settings. Finally, a phylogenetic tree

was constructed using the PhyML plug-in [51]. The tree was based on an HKY85 substitution model [52] and features 100-bootstrap branch support; optimized topology, branch lengths, rates, and nearest-neighbor interchange (NNI); and a subtree pruning and regrafting (SPR) topology search. The tree was simplified to the class level according to the NCBI taxonomy. The isolation sources of all complete genomes were identified via the NCBI BioProject and BioSample databases. The sources were classified into three categories: environmental (soil, fresh or marine water, sludge, mud, plants and algae samples), animal (insects, mollusks, fish, birds, cattle, and domestic animals), and human (isolated from various body sites of healthy or sick individuals).

Complete Bacteroidia genomes and genomic repeats

Genomes of species of the class Bacteroidia were classified by sequencing status ("Complete" or "Draft") and by year of publication. The number of contigs in draft genomes was analyzed by genus. We retrieved the complete genomes (those having no assembly gaps) of all species identified as having only one chromosome in the NCBI database. If plasmids existed, only the chromosome was analyzed. Species with two or more chromosomes were excluded (Table S1). For further analysis, the complete genomes of species for which genomic sequences of two or more strains were available were studied; these included the four *Bacteroides* species *B. dorei, B. fragilis, B. ovatus,* and *B. thetaiotaomicron*, *Porphyromonas gingivalis*, and *Tannerella forsythia*. We studied the number of contigs present in draft assemblies for those species (Table 1). The genome-based similarity measure OrthoANI was used to assess intra- and inter-species relatedness [53].

Table 1. Information on the complete genomes used in this study

Species	Strain	Sequencing Technology	Assembler	Coverage	GenBank assembly accession
Bacteroides dorei	HS1_L_1_B_010	PacBio	Celera	306	GCA_000738045
Bacteroides dorei	HS1_L_3_B_079	PacBio	Celera	370	GCA_000738065
Bacteroides dorei	HS2_L_2_B_045b	PacBio	Celera	185	GCA_001274835
Bacteroides dorei	CL03T12C01	PacBio	SMRT Analysis	193	GCA_001640865
Bacteroides fragilis	NCTC 9343	Sanger	Phrap	10	GCA_000025985
Bacteroides fragilis	638R	Sanger	Phrap	9	GCA_000210835
Bacteroides fragilis	BE1	Illumina + Nanopore	SPAdes	68 + 8	GCA_001286525
Bacteroides fragilis	BOB25	454 + IonTorrent + Sanger	Newbler	29	GCA_000965785
Bacteroides fragilis	S14	Illumina	CLC-GW + SPAdes	73	GCA_001682215
Bacteroides fragilis	YCH46	Sanger	Pherd/Phrap	10	GCA_000009925
Bacteroides ovatus	ATCC 8483	PacBio + Illumina	HGAP + Celera	350	GCA_001314995
Bacteroides ovatus	V975	454 + Sanger	Newbler	23	GCA_900095495
Bacteroides thetaiotaomicron	7330	PacBio + Illumina	HGAP + Celera	395	GCA_001314975
Bacteroides thetaiotaomicron	VPI-5482	Sanger	Phrap	7	GCA_000011065
Porphyromonas gingivalis	381	454 + Sanger	Velvet + Newbler	50	GCA_001314265
Porphyromonas gingivalis	A7436	454 + Sanger	Velvet + Newbler	57	GCA_001263815
Porphyromonas gingivalis	A7A1-28	454 + Sanger	Velvet + Newbler	94	GCA_001444325
Porphyromonas gingivalis	AJW4	454 + Sanger	Velvet + Newbler	60	GCA_001274615
Porphyromonas gingivalis	ATCC 33277	Sanger	Pherd/Phrap	9.5	GCA_000010505
Porphyromonas gingivalis	TDC60	454 + Sanger	Newble + Phrap	9 + 7	GCA_000270225
Porphyromonas gingivalis	W83	Sanger	TIGR Assembler	8	GCA_000007585
Tannerella forsythia	3313	Sanger + 454 + Illumina	Newbler	21	GCA_001547875
Tannerella forsythia	92A2	Sanger	Celera	12	GCA_000238215
Tannerella forsythia	KS16	Sanger + 454 + Illumina	Newbler	23	GCA_001547855

For complete genomes, we used Repeatoire to identify genomic repetitions that were at least 95% similar to each other and longer than 500 base pairs (bp) [54], visualizing them with Circos (v0.69) [55]. In each genome, the repeat's initial location was fixed as the starting point for all of the links to the other positions of that repeat, and we color-coded the copy numbers of each repetition.

Simulated read assembly

For each genome, artificial reads were produced using ART software (v2.5.8) [56] set for paired-end reads of 250 nucleotides (nt) each, 500 nt insert size, and a coverage of 40; the built-in MiSeq simulation profile (v1) was used. For the *P. gingivalis* genomes, eleven *de novo* assemblers were tested using the default parameters: A5-miseq (v20160825) [57]; CodonCode Aligner (CCA v7.0.1); CLC Genomics Workbench (CLC-GW v8.5.1); fermi (v1.1) [58]; Geneious (v10.2.3) [50]; Minia (v2.0.3) [59]; MIRA (v4.0.2) [60]; PERGA (v0.5.03.02) [61]; SOAPdenovo (v2.04) [62]; SPAdes (v3.10.1) [63]; and Velvet (v1.2.10) [64]. For CCA and Geneious, reads were preassembled with PEAR (v0.9.10) [65]. Where required, KmerGenie (v1.7016) was used to determine the *k-mer* length parameter [66].

After the initial test step, we selected three software packages for assembly of all 24 genomes. The first was A5-miseq, in which we used the default parameters. For Geneious, reads were preassembled with PEAR and then assembled using the default Medium Sensitivity/Fast setting and generating consensus with a 50% strict threshold for calling bases. Finally, SPAdes was used with the default parameters but with the "careful" option and *k-mer* lengths of 21 to 127.

For each genome and assembly tool, we rejected contigs of less than 1 Kbp because they are not informative. We assessed the assemblies with QUAST (v4.5) [67]. The fidelity of each assembly was evaluated by mapping each set of contigs to the reference genome using Geneious mapper (Medium Sensitivity/Fast option); the unmapped contigs were rejected, and QUAST was then used again. The correlation between the number of repeats (copy number > 3) and each assembly was tested using the "ggplot2" and "nortest" packages in R. We identified the *P. gingivalis* reference genome segments that correspond to assembly gaps and classified these into five categories: genomic islands, *rrn* operons, coding sequences (CDS)

with repeated domains, intergenic sequences, and insertion sequences or miniature inverted-repeat transposable elements (IS/MITE).

Bacterial strain cultures and DNA extraction

We purchased the ATCC 33277 and W83 (also known as BAA-308) *P. gingivalis* strains from the American Type Culture Collection-LGC Standards (Manassas, VA, USA) in September 2006; a low (< 20) passage number was used. *P. gingivalis* TDC60 (also known as JCM 19600) was purchased from the Japan Collection of Microorganisms (Riken BioResource Center, Koyadai, Japan) in November 2015, and a low (< 10) passage number was used. All strains were cultured on Columbia European Pharmacopoeia agar plates (Conda, Madrid, Spain) supplemented with 5% (v/v) defibrinated horse blood (Eurobio, Courtaboeuf, France), 5 g/L yeast extract (Conda), 25 mg/L hemin (Sigma-Aldrich, Saint-Quentin Fallavier, France), and 10 mg/L menadione (Sigma-Aldrich). The cultures were incubated in an anaerobic chamber in a Whitley DG500 Workstation (Don Whitley Scientific, Shipley, UK) for 5 days at 37 °C in an atmosphere composed of 80% N_2, 10% H_2, and 10% CO_2.

For DNA extraction, each strain was cultured for 48 h at 37 ºC under the same atmospheric conditions described above. This was done in 50 mL BHI broth (BioMérieux, Marcy l'Etoile, France) enriched with 5 g/L yeast extract (Conda), 25 mg/L hemin (Sigma-Aldrich), and 10 mg menadione (Sigma-Aldrich). After harvesting, the cells were washed twice in Dulbecco's PBS (Dominique Dutscher, Brumath, France). A QIAamp DNA Mini Kit (QIAGEN, Courtaboeuf, France) was used for cell lysis and protein denaturation. The following steps were performed using standard methods: DNA precipitation with 5 mol/L NaCl (Sigma-Aldrich) and 0.7 volumes of cold isopropanol (VWR Chemicals, Fontenay-

sous-Bois, France), two washes in 70º ethanol (WWR), and resuspension in sterile Milli-Q ultrapure water (Merck, Darmstadt, Germany).

DNA library preparation and PacBio SMRT sequencing

Barcoded DNA library preparation and single molecule real-time (SMRT) sequencing were performed using the Genome et Transcriptome (GeT) GénoToul platform (Toulouse, France) according to the manufacturers' instructions. Quality control was performed at each step. DNA was purified using AMPure PB beads (Pacific Biosciences, Menlo Park, CA, USA), and the mass of dsDNA was verified using a Qubit fluorometer (Thermo Fisher Scientific, Villebon sur Yvette, France). The purity of the DNA was determined based on absorbance ratio using a Nanodrop spectrophotometer (Thermo Fisher Scientific). Sizing measurements were performed using a Fragment Analyzer (Advanced Analytical Technologies, Evry, France). In brief, each sample was diluted to 10 µg/mL and sheared on a Megaruptor (Diagenode, Seraing, Belgium). Using 5 µM of various barcoded adapters, a SMRTbell Barcoded Adapter Prep Kit (Pac Bio) was used to repair and ligate 150 ng of DNA fragments. After end-repair and ligation using a SMRTbell DNA Damage Repair Kit, the samples were pooled. To remove unligated DNA fragments, the library was treated with an exonuclease cocktail consisting of 1.81 U/µL Exo III and 0.18 U/µL Exo VII (PacBio).

Library selection in the 6-50 Kbp range was performed using BluePippin 0.75% agarose cassettes (Sage Science, Beverly, MA, USA). Primers were annealed to the size-selected SMRTbell with the full-length libraries. The primer-template complex was then bound to the P5 enzyme using a 10:1 ratio of polymerase:SMRTbell for 4 h at 30 °C. The magnetic bead loading step was conducted at 4 °C for 1 h. The complexes were placed into the PacBio RS II sequencer, which was configured to run continuously for 6 h at a sequencing

concentration of 80 pM. The sequencing results were validated using the NG6 integrated next-generation sequencing storage and processing environment [68].

Genome assembly and finishing strategy

We mapped the contigs obtained from the *in silico* reads to each reference genome. We identified and retrieved the sequences not covered by the contigs, which corresponded to the assembly gaps. Because these gaps are repeated regions, we clustered the sequences at 99% nucleotide identity and generated a consensus sequence. Thus, each consensus represents one type of repeated region. We used canu (v1.3) to correct, trim, and assemble the raw reads [69]. We mapped the corrected long reads to each consensus sequence that was previously identified; for each, we selected only reads overhanging at least 500 nt at both ends. The selected reads were *de novo* assembled using Geneious at 100% nucleotide identity. They were then used to scaffold the contigs generated by canu and to finish the assembly and reconstruct the genome organization.

To confirm our construction, PCR was used on repeats longer than the median trimmed/corrected read length (i.e., *rrn* operons and CTnPg1). The primers used are listed in Table 2. For the *rrn* operons, 22 additional *P. gingivalis* strains were tested; 16 of these were isolated from Colombian patients with periodontitis (UIBO421B, UIBO465, UIBO472, UIBO537B, UIBO655H4, UIBO695H2, UIBO710B, UIBO728B, UIBO728H3, UIBO742, UIBO760B, UIBO771H2, UIBO783, UIBO801H3, UIBO1047B, and UIBO1047H [70]), 4 were isolated from French patients with periodontitis (2J14, M71, MAJ and TN), and 2 strains (OMZ314 and OMZ409) were provided by Prof. J. Gmür of Zurich, Switzerland. For all PCR reactions, we used 50 ng genomic DNA, 1X Phusion GC buffer, 7% DMSO, 0.02 U/µL Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Fisher), and 200 nM of each primer (Eurogentec, Seraing, Belgium). PCR was performed under the following

conditions: an initial denaturation step at 98 ºC for 3 min; 30 cycles of 98 ºC for 20 s, 63 ºC for 30 s, and 72 ºC for 7 min 30 s; and a final elongation step at 72 ºC for 10 min. The CTnPg1 validation was performed under the same conditions as the *rrn* PCR except that no DMSO was used and the annealing temperature was 67 ºC.

Table 2. Primers used to validate the architecture of the *Porphyromonas gingivalis* genome

Primer Name	Primer sequence (5' to 3')	Tm (ºC)
rrn1F	TCCCCACCGGCAAAAACATC	68.0
rrn1R	GAGATGTCCGAAAGTCCATGTCAC	66.3
rrn2F	AGATAGCCAGTTTCGTTACGTCCG	67.3
rrn2R	TACAGCAACGGTTACTTCCGCG	68.6
rrn3F	CTATGGATATTCTGCGGTGTACGG	66.3
rrn3R	GTTGTAGGACAGCAACCTTTTGG	64.2
rrn4F	ACAAGTCAGAACATGGCCGAT	64.3
rrn4R	CAGGCACAAACCGCTTTACC	65.0
ctnpg1_5out1	GACGGAATTTGCGTGTTGATATAGT	64.3
ctnpg1_5out2	ATAAACGTGTGGCCGAAATAGATTC	65.3
ctnpg1_5in	CAATAGCGTTTGCATTACCTCATCT	65.3
ctnpg1_mid	ATCGGTGGAGATGTTCATACTACTG	63.9
ctnpg1_3in	GTATTTGCCCAATACTCTCTGAACG	64.9
ctnpg1_3out1	CGACAACATCGTATTTCTCTGTCAG	64.9
ctnpg1_3out2	CACCGAGATTCAAGGTTATGTGATG	66.9

Genome annotation and biocuration

Genomes were annotated using Prokka (v1.12-beta) [71], Genix online [72], and RASTtk (v1.3.0) [73]. The annotations for the gene starts/ends, gene names, gene product descriptions, gene status (gene, pseudogene by stop in-frame or frame-shift), EC numbers, and functional descriptions were all manually biocurated [74]. For this purpose, we performed NCBI BLAST searches [75] against non-redundant databases as well as domain searches using the Conserved Domains search tool [76]. The results were then compared to the precomputed annotations from MicroScope [77].

293 The presence of signal peptides was analyzed using the SignalP 4.1 server [78] and
294 SOSUIsignal [79]. Protein subcellular localization predictions were made using PSORTb
295 (v3.0) [80] and CELLO (v2.5) [81]. The outer membrane proteins predicted by these tools
296 were then confirmed using BOMP [82], and LipoP (v1.0) [83] and DOLOP [84] were used to
297 predict the lipoproteins.
298 Insertion sequence transposases were renamed as per the ISfinder [85] nomenclature.
299 Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) were identified using
300 CRISPRfinder [86], CRISPRDetect [87], CRISPI [88] webtools and the CRT plug-in [89]
301 with Geneious. We predicted genomic islands using IslandViewer (v4) [90], which has
302 precomputed predictions from SIGI-HMM [91], IslandPath [92], and IslandPick [93]. Island
303 prediction was also performed using EGID [94], which implements AlienHunter [95], SIGI-
304 HMM, IslandPath, INDeGenIUS [96], and PAI-IDA [97].
305 All genomic sequences and the reads used to produce them were deposited in
306 GenBank and in the NCBI BioProject database with links to the BioProject accession number
307 PRJNA393092.
308

309 ## Comparative genomics

310 Genome-level comparisons were made with progressiveMauve [98] using the default settings.
311 The same program was used to identify locally collinear blocks (LCBs) and single-nucleotide
312 polymorphisms (SNPs).
313 For each strain, we compared the available annotation in the NCBI database to our
314 manually biocurated annotations feature-by-feature: rRNA; tRNA; tmRNA; ncRNA;
315 regulatory; repeat region; and coding DNA sequence (CDS). The CDS were separated into
316 pseudogenes and “true” coding regions. For the latter category, the CDS/pseudogenes present
317 in both annotation versions were termed “common,” and those that were present only in our

new version were called “new.” We noted all instances of changes from a CDS to a pseudogene and vice versa, fusion (more than one interval becoming one interval) and separation (one interval becoming more than one interval), and changes in the coding strand. Finally, we also analyzed the changes in start/stop codons.

We concluded our study with a pangenomic analysis of the *P. gingivalis* strains ATCC 33277, TDC60 and W83. The CDS and pseudogenes were classified into six groups: multiple-copy genes (nucleoid-associated proteins, *tra* genes, *xer* genes, and transposases in ISPg); those that appeared in only one copy in all of the strains and were highly conserved (> 97% nucleotide identity); those present in only one copy in all of the strains but with sequence divergence (< 97% nucleotide identity); those present in all three strains but present in more than one copy in at least one strain and/or having pseudogenes; those present only in two strains and either copied or with pseudogenes; and those present in only one strain.

Results

The Bacteroidetes phylum is represented by few genomes, mostly incomplete and with variable numbers of contigs

Bacterial genomes represent approximately 85% of the available genomes in the NCBI genome database. Despite the fact that Proteobacteria represents a relatively minor constituent of human microbiota, it is the most sequenced bacterial phylum, even in comparison to the two most abundant phyla, Bacteroidetes and Firmicutes (Fig. S1a). Furthermore, the accumulation of incomplete draft bacterial genomes is notable, comprising 85-95% of all entries depending on the phylum (Fig. S1b).

Within the phylum Bacteroidetes, the classes Bacteroidia and Flavobacteriia constitute 90% of the listed genomes. Flavobacteriia is mostly associated with environmentally obtained

samples, particularly with aquatic species such as water and fish pathogens, and rarely (18% of complete genomes) with humans. In contrast, Bacteroidia isolation sources are human in more than 75% of cases (Fig. S2a). Although generally considered part of the normal microbiota, Bacteroidia are pathobionts that can become pathogens upon dysbiosis. Prior to the introduction of massively parallel sequencing, only four complete Bacteroidia genomes were published. Since 2007, this number has grown exponentially, mainly due to the large number of draft genomes that have accumulated to such an extent that today 10 times as many draft genomes as complete genomes are available (Fig. S2b).

In addition to the preponderance of draft genomes, the Bacteroidia draft genomes are extremely fragmented; half of them contain more than 75 contigs. The number of contigs per draft ranges widely from 2 in *Prevotella oryzae* DSM17970 to 4357 in *Bacteroides acidifaciens* 1a3B. Observing the number of contigs per draft in bacterial genera with at least five draft genomes, we noted that 50% of all entries are *Bacteroides* and that *Bacteroides*, *Parabacteroides,* and *Prevotella* all have at least one draft genome featuring more than 500 contigs. Based on their interquartile ranges (IQRs), the genera with less variation were found to be *Alistipes* and *Tannerella*; coincidentally, these were the genera with the smallest number of draft genomes. At 93 contigs per draft, *Bacteroides* and *Tannerella* have the highest median number of draft genomes (Fig. S3a). Even when only species having at least two complete genomes (*B. dorei*, *B. fragilis*, *B. ovatus*, *B. thetaiotaomicron*, *P. gingivalis*, and *T. forsythia*) are considered, this variability is still quite large. *B. fragilis* represents almost 20% of all Bacteroidia drafts and is the most variable in terms of the number of contigs per draft. *B. dorei* and *T. forsythia*, although the species with the smallest number of drafts, are the least variable. Finally, *P. gingivalis* has an intermediate profile (Fig. S3b).

Draft variability cannot be explained by sequencing and assembly methods

Several hypotheses could explain the observed variability in the features of the draft genomes. It could be linked to differences in the assembly strategies (algorithm types or specific assembly tools) and/or the massively parallel technologies used. To test these hypotheses, we collected all of the available information on sequencing and assembly methodology for all of the draft genomes of the six species studied here (Table S2). Of the 166 drafts, 144 were sequenced using Illumina technology, and only six assemblers were used for *de novo* assembly. Five assemblers used the De Bruijn graphs (Allpaths, Velvet, ABySS, CLC Genomics Workbench), covering 40% of the drafts. The final MaSuRCA assembler, which is based upon the de Bruijn and Overlap-Layout-Consensus (OLC) approaches, was used for the remaining 60% of the drafts. However, the overrepresentation of MaSuRCA is due to a single sequencing project of different strains of *B. fragilis* at the Institute for Genome Sciences (University of Maryland, Baltimore, MD, USA). Once again, the results (whether per assembler or per bacterial species) were very diverse. In *B. fragilis* projects using Illumina for which an assembler was specified (99 of 107 drafts), MaSuRCA generated 31 to 2566 contigs (n = 69), Allpaths produced 5-14 contigs (n = 10), SPAdes generated 73-343 (n = 8), ABySS produced 150-1290 (n = 6), CLC Genomics Workbench produced 5-156 (n = 5), and Velvet, which was only used once, generated 52 contigs. The 20 *P. gingivalis* draft genomes were sequenced using Illumina technology and then assembled 17 times with Velvet (generating 22-192 contigs) and once each with SPAdes (92 contigs), Celera (104 contigs), and SOAPdenovo (117 contigs). For all other species, the number of drafts with known assemblers is too small (< 10) to draw conclusions. At this point, it therefore seems that technological differences in sequencing and assembly cannot explain the extensive differences in the number of contigs per draft.

391 The hypothesis that intra-species diversity might be responsible for this variability
392 remains to be explored. Because this hypothesis involves comparative genomics, only
393 complete genomes could be used to explore its validity.
394

395 Complete *Porphyromonas gingivalis* genomes are the most diverse and
396 repeated genomes within Bacteroidia

397 By calculating the average nucleotide identity (ANI), we separated two groups of species. The
398 first group contained the *Bacteroides* genus, whereas the second group included *P. gingivalis*
399 and *T. forsythia* (Fig. 1a). ANI values offer a robust and sensitive way to measure the
400 evolutionary relationship of bacterial strains [99]. With the exception of the *P. gingivalis*
401 group (11 of 21 pairs), the ANI values of all the species are high and uniform (Fig. 1b). These
402 values suggest that only minor genomic differences exist within species and refute the
403 hypothesis that global intra-species diversity could explain the variability in the number of
404 contigs per draft genome. However, because small variations such as repeated genomic
405 regions cannot be detected by the ANI method and since repeats have been declared to be the
406 main cause of assembly gaps, we decided to evaluate the possible role of repeats in the
407 fragmentation of draft genomes.
408 We determined the number and locations of each repeat type. Because Illumina is the
409 industry standard and its MiSeq platform generates 2 x 250 nt paired reads, we set the low
410 threshold to 500 bp and looked for the repeats. The histogram in Fig. S4 shows the mean
411 number of repeats in each species (2 to 10 copies at 95% identity). Notably, although *B.*
412 *fragilis* displays the greatest variation in number of contigs per draft, it has an intermediate
413 number of repeats, and these have low copy numbers and occur a maximum of only six times.
414 *P. gingivalis* can once again be distinguished as having the lowest total number of repeats but
415 the highest number of copies for each repeat type, with a total of approximately 40 different

repeats copied more than 10 times (Fig. S4).

We used a Circos figure to visualize the genomic repeats with more than 3 copies in each complete genome. The plot illustrates the distribution of the repeated loci and their copy numbers in the form of a heat map, the color of which changes from blue to red as the count increases (Fig. 2). The plots show that all six available strains of *B. fragilis* possess few types of repeats and that repeats are not frequent. In contrast, all of the other *Bacteroides* species genomes (*B. dorei*, *B. ovatus*, and *B. thetaiotaomicron*) have more repeats that occur more often and display greater variability within the strains. Finally, we observed that the seven *P. gingivalis* strains had the highest genomic complexity and diversity with respect to repeat frequency.

The intra-species diversity of *P. gingivalis* and its large number of highly frequent repeats make this species an interesting model for analysis of the impact of genomic repeats on bacterial genome assembly, especially in Bacteroidetes. The *P. gingivalis* strains appear as three branches on the DNA-DNA distance tree. Two of these branches correspond to the previously mentioned closely related strains: ATCC 33277 with 381, and W83 with A7436 (Fig. 3). The ATCC 33277/381 branch contains the genomes with the most repeats, with some loci repeated more than 25 times, followed by the W83/A7634 branch. The remaining branch contains TDC60, AJW4, and A7A1-28, which display the lowest numbers and frequencies of repetition. With the exception of W83 (from Bonn, Germany) and TDC60 (from Tokyo, Japan), all of the cited strains were originally isolated in the USA. Despite this common origin, there is no apparent link between isolation populations and genomic repetition frequencies. To test whether the variation in genomic repeat counts affects assembly completion and since the real sequencing reads from the initial sequencing projects were not available, we produced *in silico* reads based on the complete published genome of the seven *P. gingivalis* strains and *in silico* simulated paired-end reads.

Simulated sequencing reveals a correlation between contig and repeat counts

To generate the artificial reads, we used ART software. This set of tools mimics the real sequencing process and permits simulation of the data that are produced by massively parallel methods; the simulated data display each technology's inherent empirical errors and profile qualities. For the seven complete *P. gingivalis* genomes, 11 assemblers were tested. The assembler spectrum was restricted to software that can treat Illumina reads but was otherwise chosen to be as wide as possible. Three of the assemblers are part of commercial software suites (Geneious, CLC Genomics Workbench, and CodonCode Aligner); the others are freely available for academic purposes. Since de Bruijn graphs are inescapable (they are used by A5-miseq, CLC Genomics Workbench, Geneious, Minia, MIRA, SOAPdenovo2, SPAdes, and Velvet), we made an effort to also include different assembly methods, including Overlap-Layout-Consensus (CodonCode Aligner), string graph (fermi), and greedy algorithm (PERGA).

We assessed each assembly's fragmentation and compared it to that of the others using QUAST's N50 parameter. For all *P. gingivalis* strains, Geneious produces the longest contigs and the fewest per assembly. SPAdes and A5miseq are just behind, with very similar performances. The other assemblers generate more fragmented assemblies, and some (Velvet and PERGA) do not even seem suitable for this bacterial group (Fig. 4a). Closer examination of the results of the three assemblers producing the highest N50 (Fig. 4b, upper panel) shows that the number of contigs obtained is below the median (90.5) of the published draft genomes assembled with real reads (Fig. S3b).

N50 can be misleading, however, because it provides no information on assembly accuracy; it can therefore give the impression that one tool is more algorithmically efficient

than another even when its results do not agree with the biological sequence. Since the simulated reads were created from the complete NCBI reference genomes, mapping the resulting contigs against the genomes should be biologically consistent. The unmapped contigs would therefore be rejected as believed to be incorrect, resulting in the graph shown in the lower panel of Fig. 4b. The plot demonstrates that despite their high N50 values, Geneious contigs, especially the longer ones, are not consistent with the real sequence. The plot line will therefore fall short of the expected genome size, resulting in drafts that are not at all complete. In contrast, A5miseq and SPAdes yield similar plots since they accurately reproduce the biological sequences, despite the fact that their resulting contigs are shorter. Evaluating assembly completeness solely based on the N50 metric is a bad idea, and it is preferable to obtain a draft genome that is correct even if it is more fragmented. Consequently, A5miseq and SPAdes produce the best results for the *P. gingivalis* genomes (modeled with MiSeq reads), and we confirmed this to be true for the other Bacteroidia species as well (Fig. S5a).

A direct positive linear correlation can be observed in the number of *P. gingivalis* genomic repeats (at least three copied loci) and the number of biologically sound contigs produced by A5miseq or SPAdes from the *in silico* simulated reads (Fig. 4c). This correlation is also observed for the other complete Bacteroidia genomes studied here (Fig. S5b). In the case of *P. gingivalis*, it is true even when all genomic repeats, including duplications, are included (Fig. S5c).

We can classify the functions of the elements that are annotated in the breaking points of the A5miseq and SPAdes *P. gingivalis* assemblies. These gaps coincide with the repeated regions in the genome, and two thirds of them correspond to copied insertion sequences (IS) or miniature inverted-repeat transposable elements (MITEs) with a maximum length of 1.1 Kb. In order of their occurrence, the other four categories are: intergenic unannotated regions;

repeats found in coding regions (such as gingipains) that have several paralogs and internal repeated motifs; *rrn* ribosomal RNA operons (present in four copies in *P. gingivalis*); and, the least frequent, genomic islands (Fig. 4d).

This study shows that to correctly assemble a genome from Illumina MiSeq reads of a few hundred nucleotides, it is helpful and even indispensable to have at least a good estimate of the quantity and frequency of genomic repeats. This knowledge permits the calculation of contig counts with little or no misassembly expected. It could also help in the choice of the best finishing strategy for the assembly. However, as seen here, this knowledge is often impossible to obtain in advance, since even among bacterial species the strains are very diverse.

Choosing *Porphyromonas gingivalis* strains for resequencing and reassembly

Taking into account our initial results and observations, we chose to resequence three *P. gingivalis* strains (ATCC 33277, TDC60, and W83) using long-read technology from PacBio. We chose these strains because they are commercially available, allowing other researchers to reproduce our results, and because of their geographic diversity (they appear in North America, Europe, and Asia). Moreover, they were the first *P. gingivalis* strains to be sequenced and are considered reference strains. In addition, ATCC 33277 and W83 are frequently used to study mutants, in functional analysis, in adherence/invasion tests involving different types of human cells, and for other purposes.

For each strain, the mean coverage was approximately 30X (29.7 to 30.4), and the median corrected/trimmed read length obtained from canu was 6.3 Kb. The reads were assembled with canu, producing 18, 28, and 11 contigs for ATCC 33277, TDC60, and W83, respectively, and the finishing strategy described in the Methods section enabled assembly

completion. All of the resulting complete assemblies differ from the initially published sequences. To validate the results, we performed PCR to confirm the organization of the large genomic multicopy regions. These included the *rrn* operons (Fig. S6a) and, for ATCC 33277, the duplication of the CTnPg1 genomic island and the orientation of the two copies (Fig. S6b-c). All of the RNA ribosomal operons from the ATCC 33277 and W83 strains were validated and found to correspond to the published structures. However, we corrected 3 of the 4 *rrn* in the TDC60 published genome, and this result confirms the observations made by Naito *et al.* in 2008 (see their Supplementary Fig. 3) [100]. The most surprising result was the reorganization of 2 *rrn* loci in W83 compared to the other two strains (Fig. S6a). To determine whether this reorganization occurs only in this strain, 22 additional strains were subjected to PCR verification; we found that it is in fact restricted to W83 (Fig. S6d). The general arrangement of *rrn* operons in *P. gingivalis* chromosomes therefore seems remarkably stable despite the varied positions of these operons relative to the origin of replication (*ori*C) due to the high genomic mosaicism of the species. Homologous recombination at the *rrn* extremities seems extremely rare and probably accidental since of approximately 30 strains only W83 is affected.

Differences from published genome sequences

With all three circular genomes confirmed, we proceeded to identify the ways in which these genomes differed from the previously published sequences for these strains. We used progressiveMauve, which identifies locally collinear blocks (LCBs) to compare the published genomes with our constructions. Insertions and deletions in the *de novo* assembly compared to the published sequence were noted (Fig. S7). The differences are reported in detail below.

In ATCC 33277, CTnPg1 is found to be completely duplicated, with both copies oriented in the same direction (Fig. S7). The other differences in this strain are three deletions

and one insertion (Fig. S8a). Since the ANI values showed a significant similarity of the ATCC 33277 and 381 strains and the 381 strain also has two complete CTnPg1 copies, we compared our ATCC 33277 genome reconstruction with the 381 genome [101]. We observed two collinear genomes with a length difference of only 1 kb and approximately 100 SNPs. The dissimilarities, which are minor, occur mainly in repeats and in the CRISPR1 region. As the complete 381 genome was assembled from 454 reads using Velvet and Newbler, there could be assembly errors. By mapping the Illumina HiSeq reads of the recent 381 sequencing [102] to our *de novo* ATCC 33277 construction, we positioned 99.5% of the reads without any gaps, suggesting that these American strains could be variants of the same strain.

TDC60 is the most changed, since, as discussed, 3 of the 4 *rrn* were incorrectly assembled during the initial construction. This means that in our new assembly, large sections of the genome are translocated (LCB2 and LCB5) and inverted (LCB3) (Fig. S7). The other differences are one insertion into a BrickBuilt 7 MITE [103] and four deletions (Fig S8b).

Finally, the new W83 strain has a central inversion (LCB2) (Fig. S7). The new sequence consists only of a 523-bp insertion, which corresponds to eight additional direct repeats and the same number of new spacers in the CRISPR2 region (Fig. S8c). The genome length differences and numbers of SNPs in the three strains are shown in detail in Fig. S7.

Annotation and manual biocuration of three selected *P. gingivalis* strains

We automatically *de novo* annotated the three *Porphyromonas gingivalis* strains using three pipelines, after which they were manually biocurated. This allowed us to standardize the structural (syntactic) annotation, which consists mostly of CDS start/stop codon positions, and the functional annotation, using a single ontology for gene names and functional descriptions. The number and nature of all 12 rRNA genes (in four operons), 53 tRNA genes, and 1

566 tmRNA gene are identical to those in the published genomes, although the positions of some
567 relative to *oriC* vary due to the reassembly. Seven of the riboswitches (5 cobalamin and 2
568 thiamine pyrophosphate riboswitches) were already annotated in all of the strains, and we
569 added 1 S-adenosyl methionine (SAM)-II long-loop riboswitch per strain. It is noteworthy
570 that all of these riboswitches were described by Hovik *et al*. [104]. For the small non-coding
571 RNA (ncRNA), only *rnpB* (bacterial RNAse P) is present in the previous annotation. We
572 positioned and annotated additional ncRNAs: one each *ctRNA* (antisense RNA),
573 Bacteroidales-1 RNA, and bacterial signal recognition particle (SRP RNA) for each strain.
574 We also added various group II catalytic introns: 5 in ATCC 33277; 4 in TDC60; and 3 in
575 W83. For all riboswitches and ncRNA, the synteny is conserved. Two group II catalytic
576 introns are present in *haeR* and in *punA*. The variation in numbers in this group is explained
577 by the paralogy in *traE* (a gene containing this ncRNA is present once in W83 and twice in
578 ATCC 33277 and TDC60) and by an additional ncRNA in the ATCC 33277 *traG* gene that is
579 absent from the other two strains.

580 For ATCC 33277, TDC60, and W83, we added 213, 199, and 188 peptide signals and
581 23, 19, and 21 mobile elements (genomic islands, transposons, and conjugative transposons),
582 respectively, to the original annotations. We further completed the annotation of intergenic
583 sequences with repeated elements: BrickBuilt/MITEs (complete or partial), dispersed
584 genomic repeats (> 500 bp with at least 95% identity), and CRISPR regions. We did not
585 annotate any short repeats (3 to 31 bp) or sequence-tagged sites.

586 Finally, the longest and probably the most important biocuration work involved the
587 CDS and pseudogenes, which we were able to dramatically improve. To facilitate traceability
588 in publications and databases, we kept the initial locus_tags when the DNA sequence was
589 unchanged or very similar (only having a new start and/or stop position). When the
590 differences were larger (e.g., changes in the coding strand, ORF fusion, or a new CDS), we

591 created new locus_tags marked *PGN_n*, *PGTDC60_n*, and *PG_n* for ATCC 33277, TDC60,
592 and W83, respectively. In the previous versions, the pseudogenes had coding sequences and
593 gene annotations, with pseudogene status annotated in the gene qualifier "note." These could
594 be classified into the categories "frameshifted," "incomplete," and "internal stop." Following
595 the NCBI Prokaryotic Genome Annotation guidelines
596 (https://www.ncbi.nlm.nih.gov/genbank/genomesubmit_annotation/), we annotated the
597 pseudogenes with the gene feature only (no CDS) and added an asterisk to the gene name
598 field so that pseudogenes could be easily distinguished from normal coding regions. The
599 gene_desc qualifier field describes the function of the pseudogenized gene, and the reason for
600 the pseudogenization ("fragment" or "frameshift") is listed in the note qualifier. Where they
601 are relevant, gene fusions and coding strand changes are also mentioned in the "note" field.
602 For the three strains, our revised annotation contains fewer CDS and pseudogenes than
603 are found in the original annotations (Table 3). These differences are essentially due to over-
604 annotation of the original genomes and to genome-to-genome annotation propagation caused
605 by the difficulty of distinguishing between small coding ORFs and Evil Little Fellows or
606 ELFs [105, 106]). These are regions that can accidentally produce ORFs that are not
607 biologically verified. To avoid erroneous elimination of real small ORFs, we identified ORFs
608 as ELFs only if one of the following conditions applied: if the ORFs were annotated in
609 intergenic repeats such as MITEs, which by definition do not code [107]; if the ORFs were
610 not conserved in all 30 available *P. gingivalis* complete and draft genomes but the DNA
611 sequences and synteny were conserved; and if the ORFs were small and possessed an
612 AMIGA status that was "wrong" [108]. We systematically predicted the cellular localizations
613 and/or the conserved functional domains of the small conserved ORFs. This sometimes
614 enabled us to improve the annotation of the hypothetical functions of small CDS into
615 "lipoprotein," "inner membrane," "periplasmic," "outer membrane," or "secreted." Finally,

the use of manual biocuration permitted us to correctly identify some coding sequences as pseudogenes and vice versa, to fuse two initial CDS/pseudogenes into one, and to divide a CDS/pseudogene to create two or more. In rare cases (n = 5), our re-annotation produced a coding strand change for the ORF involved (Table S4).

Table 3. Summary of CDS and pseudogenes (in parentheses) in the original and curated annotations

	Original annotation	Curated annotation	Common to both	Curated start/stop positions	Eliminated ELFs
ATCC 33277	2051 (96)	1856 (60)	1722 (22)	155 (8)	165 (46)
TDC60	2031 (89)	1794 (34)	1742 (13)	164 (8)	172 (64)
W83	2027 (112)	1801 (77)	1703 (45)	163 (14)	165 (53)

Comparison of the original annotations and our comprehensive syntactic re-annotations

To compare the original and curated annotations, we began by summarizing the common CDS and genes in the two versions as well as the number of curated start and stop positions and eliminated ELFs (Table 3). Additionally, for ATCC 33277, we added 55 new coding sequences; these usually corresponded to the re-establishment of a complete CTnPg1 copy or the reconstruction of complete ISPg CDS copies. For the TDC60 strain, we added 2 new pseudogenes (ISPg) and 6 new CDS (coding for 2 ISPg, an NAP, a peptidase, a rhodanase, and a PF07877 domain-containing protein). Finally, in the W83 strain, we *de novo* annotated 6 pseudogenes (5 ISpg and a PF07877 domain-containing protein) and 12 CDS (these encode 7 ISPg, 2 integrases, an NAP, a transcriptional regulator HxlR, and a virulence-related RhuM protein).

To compare the functional annotations in the two versions, we grouped the common features into five categories: nucleic acids and protein metabolism (replication, transcription,

translation, histones, proteases, etc.); metabolism and transport of ions and other macromolecules (transporters, porines, ion channels, ferritin, kinases, hydrolases, etc.); mobile elements (conjugation, competence, phages, etc.); proteins without assigned biological functions but with a predicted subcellular localization or possessing an identified functional domain (lipoproteins, inner/outer membrane, or proteins which contain GLPGLI or zinc finger domains, etc.); and proteins that are conserved in all available *P. gingivalis* genomes but to which we could not assign a function (DUFs, FIGs, and some COGs, all marked as hypothetical).

Our biocuration and re-annotation work on the coding sequences and pseudogenes shared by the two versions resulted mainly in the de-anonymization of hypothetical CDS/proteins of unknown function (Fig. 5). These constituted approximately 22% of the original annotation and now represent only approximately 9%. The newly described functions were placed into four other categories as follows: 6% were added to macromolecule/ion transport and metabolism, an additional 3% were placed in nucleic acids/protein metabolism, mobile elements increased by 2%, and 2% were added to the list of CDS with a predicted conserved domain and/or subcellular localization. Finally, we noted that for the new pseudogenes and coding sequences as well as all of the other changes (features, fusions, separations, etc.), the new annotations are particularly enriched in mobile elements (Table S4).

Pangenome analysis

All gene annotations (both true CDS and pseudogenes) from the three strains were classified into the five previously cited categories (Fig. 6). As shown in Fig. S9, with the exception of genes related to mobile elements, the absolute number of genes in each category is uniform. These transposases, integrases, phages, and conjugation genes are overrepresented in ATCC

33277, and underrepresented in TDC60. Even after manual biocuration, 16.8% of all genes remain poorly characterized; 7.7% have only subcellular localization predictions or functional domains, and 9.1% of the proteins conserved in *P. gingivalis* still have unknown functions.

The *P. gingivalis* ATCC 33277, TDC60 and W83 pangenome can be divided into four categories

The first class is genes that are present in at least 2 copies in all of the studied strains; these represent 10.1% of the total number of genes in ATCC 33277 and 7.5% of the total genes in the other two strains. Some, such as DNA helicase, transcriptional regulators of the helix-turn-helix (HTH) 17 family, and nucleoid-associated proteins (NAP), are involved in nucleic acid metabolism. The others are associated with horizontal gene transfer and include integrases, tetracycline-resistance elements, *tra* genes of conjugative transposons, and transposases encoded by several types of insertion sequences (Fig. S10a). Transposases represent more than a third of the multicopy genes and are highly pseudogenized, particularly in W83 (Fig. S10b). Their distribution in the strains is variable; ISPg8 (previously known as ISPg1) occurs the most frequently, followed by IS195 (formerly ISPg3). It is notable that W83 is the only strain rich in ISPg4. In ATCC 33277, we annotated a pseudogenized ISPg5 despite this gene's being described as absent by Califano *et al.* [109] (Fig. S10c).

The next category is the "dispensable/accessory genome," which consists of genes that are present in only two of the three strains (70 genes in ATCC 33277, 82 in TDC60, and 49 in W83). Genes with unknown function are more abundant, with 40% (38.6 to 44.9%) poorly characterized genes in the accessory genome vs. 17% in the core (Fig. 5). Many of these genes are in chromosomal regions that have been annotated as genomic islands.

The third gene class ("strain-specific or unique genes") consists of genes that are present in only one of the three strains. This class includes 3.3% of the genes in ATCC 33277,

688 2.9% of the genes in TDC60, and 4.7% of the genes in W83. Within this class, genes related
689 to mobile elements *per se* are rare (< 5%), and genes of unknown function represent
690 approximately 45% in ATCC 33277 and approximately 35% in TDC60 and W83. The
691 singularity of these "unique" genes is relative since BLAST analysis of *Porphyromonas*
692 genomes yields homologs in at least one other genome. This is often the case for ATCC
693 33277 genes in the 381 or HG66 strains and for W83 genes in the A7A1-28, A7536, and
694 AJW4 strains. Some of the unique genes present in the three strains have hits in *Tannerella*
695 *forsythia* genomes. Many of the unique genes are clustered in the genome.

696 The final gene class can be referred to as the "core" genome because it contains the
697 genes that are present in all three strains. This class includes more than 80% of all genes
698 (82.8% in ATCC 33277, 85.0% in TDC60, and 84.7% in W83) and can be further subdivided
699 into a constant core, a variable core, a paralogy core, and a pseudogenized core (Fig. S11a).
700 Due to paralogy, the total number of genes in the core genome varies, but it starts at 1522
701 genes.

702 The first subgroup of genes in the core genome is the "core genome *sensus stricto*" or
703 the "constant core genome," which contains genes with only one ortholog in each strain and
704 more than 97% nucleotide identity. The constant core represents the majority of the core
705 genes (n = 1424, 93% of the core); we classified these genes into the functional categories
706 described above (Fig. S11b). There are only four genes related to horizontal gene transfer, and
707 all four (*pir*, *yhaI*, *vrgG,* and *p12*) correspond to phage-related proteins. Since no phage
708 infecting *P. gingivalis* has been described [110], it would be interesting to study the
709 significance of this finding. We also observed four pseudogenes that are present in the three
710 strains: *COG4335,* a hypothetical protein with a conserved domain associated with DNA
711 alkylation activity [111]; *pspB,* which encodes alpha-ribazole-5'-phosphate phosphatase, an
712 enzyme involved in coenzyme B12 biosynthesis; *pyrD,* coding for dihydroorotate

dehydrogenase, which is involved in the *de novo* biosynthesis of pyrimidine; and finally *tetR,* which encodes a transcriptional regulator.

The second subgroup is the "variable core", consisting of genes that also only have one ortholog but display less than 97% nucleotide identity (ranging from 96.9 to 66.3%). This subgroup represents only 3% of the core or 47 variable genes (Fig. S11c).

The third subgroup is the "core with paralogy"; it contains genes with only one ortholog in a strain but with duplications in at least one of the strains. This subgroup represents only 2.5% of the core genes (n = 36). The genes in this subgroup are related to DNA transcription, modification and repair (*alkD, betI, btr,* and *ytxK*), and transport (*irtA, ndvA,* and *ydfJ*). There are also four genes of known function (*ctpA* protease, *epsJ* glycosyltranferase, *era* GTPase, and the *rhuM* virulence protein) and four hypothetical proteins (Fig. S11d).

The fourth and final subgroup is the "core with pseudogenization", which includes genes with only one ortholog in each strain but also at least one pseudogene in another strain. This group contains 36 genes, represents 2.5% of the core genes, and includes genes with unknown function as well as genes involved in adaptation and/or pathogenicity, for example, genes involved in transport, nucleic acid metabolism and cellular appendages (Table S5).

Discussion

Bacteroidetes dominate human intestinal bacterial communities [112], but these taxa are also associated with other animals and found in soil and aquatic environments. This suggests that they play an important role in biogeochemical processes [8]. *In silico* modeling allows us switch from describing microbial communities to actually predicting genotype-phenotype relationships and microbe-microbe and microbe-host interactions, but it requires genomic

737 reference sequences that are both accurate and exhaustive [113, 114]. The existing databases
738 contain thousands of bacterial genomes, but their quality is poor since nearly 90% of all
739 available bacterial genomes are merely drafts [40, 115]. Furthermore, the biodiversity of the
740 databases is biased because the data were primarily obtained from pathogenic or
741 biotechnologically/economically interesting bacteria [116-118]. As an example,
742 Bacteroidetes, despite their abundance among human microbiota, represent a minority of the
743 species in genome databases, with some overrepresented species such as *Bacteroides fragilis*.
744 Moreover, even when a reference genome is available for a bacterial species, this is not
745 sufficient to permit evaluation of the true biological diversity of its strains [119]. Of even
746 more concern, the number of drafts is actually underestimated since some genomes marked
747 "complete" or "finished" are fragmented and have not been fully annotated. For example, we
748 observed (Table S1) that two of nine *Porphyromonas gingivalis* genomes described as
749 complete were artificially circularized: HG66 [120] has an assembly gap of unknown length
750 (represented by 100 N), and JCVI SC001 [121] contains 282 assembly gaps, abnormalities
751 also noticed by Chen *et al.* [122].

752 Draft genomes are made up of a set of contigs of varying size and unknown order,
753 orientation, and quality. They may contain sequencing and assembly errors as well as absent,
754 fragmented, and/or frameshifted genes and other artifacts [44, 117, 123, 124], and these
755 irregularities can also occur in closed genomes, as demonstrated by this study. Worse, without
756 a complete genome it is difficult to identify possible contaminant sequences; in some draft
757 genomes, even human genome sequences can be observed [125-127]. Nevertheless, most
758 current sequencing project teams settle for drafts, limiting our knowledge of the genomic
759 structural organizations crucial for understanding bacterial evolution and adaptability [128].
760 Comparative genomics of two incomplete genomes can yield a description of most of the
761 protein-coding genes, but only complete genomic information will allow us to explore the

762 frequency and localization of repeated sequences, paralogy, synteny, and genomic
763 rearrangements. Incomplete data also hamper the interpretation of ecological models and
764 evolutionary reconstructions [116, 129].

765 Draft quality assessment is difficult, especially in the absence of a reference genome
766 [130]. It is usually evaluated using the N50 parameter [117, 131], with a higher value
767 indicating better quality [132]. However, our study shows that N50 is inadequate for judging
768 an assembly's biological value because it favors long contigs even when they are
769 misassembled. Dozens of assembly software packages exist, and comparative studies have
770 shown that their effectiveness depends on factors such as genome size, coverage, sequencing
771 technology, and the presence of atypical DNA (transposons, plasmids, and phages) [124, 133-
772 136]. Our analysis shows that this biological variability also influences assembly.

773 Using an *in silico* sequencing strategy to evaluate *de novo* assemblies obtained using
774 different software packages, we compared the number of genomic repetitions (*rrn* operons,
775 insertion sequences, adhesins, proteases, etc.) and the number of obtained contigs. Despite the
776 fact that repetitions are often cited as being responsible for assembly breaks [102, 123, 137,
777 138], as far as we know our study is the first to demonstrate a strong correlation between the
778 number of repeats and the number of contigs. These repeats "break" the assembly and are
779 quite often present as individual contigs or at the ends of contigs in draft genomes. However,
780 repeated elements are important for genome plasticity (rearrangements, duplications,
781 inversions), and even considering that their genome coverage is estimated at 6.9% for
782 prokaryotes [139], their apparent percentages vary widely even within the same species (4.8
783 to 6.7% for *P. gingivalis*, 1.5 to 2.4% for *B. fragilis*), and they are impossible to evaluate *ab*
784 *initio*. Genomic repeats are frequently transposases in insertion sequences [140, 141], and
785 they represent two-thirds of the breakpoints in *P. gingivalis* assemblies. Sequencing

786 technologies that generate reads shorter than the repeat length are not suitable for resolving
787 these assembly problems [123, 142, 143].

788 All of the molecular studies of the *P. gingivalis* genome published in the last 10 years
789 have shown that the genomes of *P. gingivalis* strains are highly variable [144-148]. Our
790 average nucleotide identity (ANI) calculations show, however, that as described by other
791 authors [144, 149], *P. gingivalis* is a single species with a genomic heterogeneity indicating a
792 non-clonal population. Furthermore, this variability is common to opportunistic pathogens
793 that are responsible for chronic colonization and infection [150, 151]. *P. gingivalis* is
794 therefore an interesting model for exploring the relationship between strain genomic diversity
795 and potential differences in pathogenicity and virulence. The potential of DNA recombination
796 in this diversity is facilitated by the natural competence of *P. gingivalis* [148]. This suggests
797 that the species represents a panmictic population [146] with high genomic mosaicity, as
798 confirmed in our study. The nutritional use of exogenous DNA as a carbon and energy source
799 certainly facilitates recombination [148].

800 In this study, we chose to resequence three *P. gingivalis* reference strains from
801 international collections. We demonstrated that even modest long-read coverage (~30X)
802 combined with biocurated assembly and some PCR contiguity validation could correct these
803 highly plastic genome assemblies. This result is supported by previous studies of other
804 bacterial species [38, 43, 152, 153].

805 Our work confirms the genomic diversity and plasticity of *P. gingivalis* but also shows
806 that the species includes clonal subpopulations of closely related strains. This is especially the
807 case for ATCC 33277 and 381 and to a lesser extent for the W83 and A7436 strains. We
808 confirmed the relatedness of ATCC 33277 and 381 mainly through the reconstruction and
809 reorientation of two full copies of the CTnPg1 conjugative transposon in the ATCC 33277
810 genome. These two strains are either cited as identical strains [102, 122, 154] or as variants

(ATCC 33277 being a natural streptomycin-resistant mutant [155, 156]). Our study shows the importance of submitting the sequencing reads to databases such as the NCBI SRA sequence read archive so that they can be reused for further analysis. This allows the scientific community to complete the study of genomes, adding value to the work of the initial researchers. Unfortunately, however, many reads are not publicly available, since it is not mandatory to upload them (even if it is highly recommended).

As previously stated, the main reason for sequencing several strains from a single species is comparison of their genomes with the goal of explaining phenotypic differences and understanding the evolutionary history and adaptation of the species. To do this, we compared the three resequenced *P. gingivalis* strains after performing a thorough manually biocurated annotation. Similar to Guo *et al.* [157], our biocuration strategy involves homogenizing transcription initiation sites, rigorously identifying frameshifts, internal stop codons, and intergenic low complexity repeats, and eliminating false CDS predictions; finally, if coding sequences only have hypothetical functions, we assign functions or predict the subcellular localizations of their gene products. Although manual biocuration is time-consuming and labor-intensive, it is essential for proper comparison [115, 158] and to avoid the false positives and negatives propagated by automatic annotation pipelines [159-161]. After this step, the mean genetic density was 85.6%, closer to the mean value of approximately 85% that has been described for prokaryotic genomes [162] than to the value of 87.5% obtained through automatic annotation. CDS boundaries were analyzed via comparative ortholog analysis, and we made corrections in the corresponding genes for the three strains. Biocuration resulted in changes in start codon use, with AUG (Met) used in 97.4% of cases for the three strains (an increase from 85%). UUG (Leu) is the next alternative codon at 2.5% (initially 9.5%), and GUG (Val) is used in 1.1% of cases (vs. the original 5.5%).

836 Our consistent annotation yields an accurate description of the pangenome. However,
837 the multiple-copy genes deserve a study of their own to analyze their content and to determine
838 how frequency differences are related to phenotypes. We identified 1522 constant core genes,
839 equivalent to 82.5-85.0% of all protein-coding genes. This is close to the 83% (1488)
840 estimation of Dashper *et al.* [102], the 1490 described as common by Naito *et al.* [100], and
841 the 1476 core genes identified by Brunner *et al.* [163] but very different from the 55% (*n* =
842 1037) estimated by Chen *et al.* [122]. Why are these values so different? There are at least
843 four possible reasons for this. First, the number of strains studied varies widely (23, 2, 8, and
844 19 in the studies of Dashper, Naito, Brunner, and Chen, respectively). This could explain the
845 smaller differences (0.7% to 3.8%) between our study and the first three cited studies, but not
846 the difference of greater than 30% between our results and those of Chen *et al.* Next, the
847 differences may result from the nature of the studied genomes. Unlike our study, which was
848 based only on complete and verified genomes, the Chen and Dashper groups based their
849 results mostly on draft genomes. As previously mentioned, incompleteness can falsely
850 indicate that some coding regions are absent and can artificially enrich unique strain-specific
851 coding sequences. This may have been the case in the Chen analysis and would explain the
852 small number of core genes that were detected. Another possible explanation for the
853 differences is the way in which we calculated the core genes. Chen compares all of the
854 automatically annotated CDS, without any evident biocuration, whereas Dashper used a more
855 biocurated annotation. Finally, in the description of the core genome, we included variable
856 genes as well as pseudogenized or duplicated genes that are functional orthologs. These have
857 all evolved differently, and due to low nucleotide identity in reciprocal best hits BLAST
858 analysis were previously wrongly described as unique or strain-specific genes. For example,
859 this is the case for the major fimbrillin gene *fimA,* which displays 66.3% nucleotide identity
860 and 60.1% protein identity in the three strains.

Comparison of the essential genes described by Klein *et al.* (*n* = 463) [164] and Hutcherson *et al.* (*n* = 281) [165] to those in our classification shows that the vast majority of essential genes are present in the constant core genome (96% and 98.5%, respectively). Six of the genes described as essential by Klein but not by Hutcherson were eliminated by our biocuration due to the presence of MITEs or because the observations showed a conserved nucleic acid sequence but not an ORF. The genes eliminated were short and close to the 5'- or 3'-UTRs of coding genes (10 to 250 nt). This polar effect could be caused by the transposon mutagenesis used in both of the abovementioned studies, with a change in one gene perturbing the transcription of adjacent genes [166]. Two essential genes (*PGN_0919* and *PGN_1215*) described by Klein but not by Hutcherson are specific to the ATCC 33277 strain in our study, but their presence can be observed in *P. gingivalis* strains 381 and HG66 and in other bacterial genera in various phyla, including *Parabacteroides* and *Prevotella* in Bacteroidetes, *Bacillus* in Firmicutes, and *Rhizobium* and *Vibrio* in Proteobacteria. This might indicate an exogenous origin, which would be consistent with their locations in or near genomic islands. This observation confirms the importance of the biological characterization of proteins with unknown functions and shows that such effort is vital for functional genomic interpretation and identification of proteins of interest [167]. The presence of homologs of strain-specific genes in other strains or species challenges the existence of ORFans, unique or orphan open reading frames [168]. Our re-assembly and re-annotation work produced two noteworthy and highly correlated improvements: fewer genes of unknown function and fewer ORFans in all three strains. The number of unique genes in ATCC 33277, TDC60, and W83 was initially 461, 415 and 382, respectively [100, 169]. This represents 17-22% of all protein-coding genes and was reduced to approximately 3% in our study, a value that is closer to the estimate of 6-7% unique genes obtained using experimental microarrays [145, 154]. Of the 382 unique genes described for TDC60, more than two-thirds were described as hypothetical

886 [169]. Naito *et al.* noted that more than 60% of the unique coding sequences had similar
887 sequences in other strains that did not fulfill their study's criteria (cut-off of > 60% alignment
888 length and > 90% identity) [100]; these could be allelic isoforms of the same gene [170].
889 Since unique genes might be involved in adaptive responses to environmental changes, it is
890 important to obtain accurate annotations. Our analysis of the three strains shows that many
891 loci initially described as unique correspond to regions of synteny that display nucleotide
892 sequence homology but coding loss in one or many strains. In some of these regions,
893 biocurated annotation identifies pseudogenes or non-coding repeated interspersed elements
894 such as MITEs. In others, it is more likely that the coding loss involves a conserved region
895 that expresses a promoter, a terminator, or even a common ncRNA but not a CDS. We only
896 included the regions that conserved their coding capacities in all 29 *P. gingivalis* strains
897 studied (Tables S1-S3) and that did not have new overlapping annotations.

898 It is interesting to note that the genes in the variant core that have been reported in the
899 literature are mostly associated with virulence factors coding fimbriae/pili [171],
900 hemagglutinins, surface proteins and transporters [149, 163], and *cas* genes, completing and
901 reconfirming the observations of Igboin *et al.* [145]. However, although we have predicted
902 their subcellular localizations, the products of some of these genes are still of unknown
903 function. Perhaps, as in the case of many membrane proteins, these loci encode proteins that
904 have new functions, are involved in *P. gingivalis* environmental interactions or confer
905 differential pathogenicity or virulence [170, 172], since virulence differences are probably
906 due to differing external envelope components and adhesion capacities [173].

907 In ATCC 33277, our reannotation positioned 23 mobile elements (MEs) in 14 regions
908 (2 of these were separated by less than 2 kb); this is in accordance with the 13 atypical regions
909 initially described [100]. In contrast, in W83 we only annotated 12 MEs in 11 regions, 10

fewer than in the initial annotation [174]. In TDC60, we again annotated 12 MEs in 11 regions, thus actually enriching the previous annotation, which only included 4 [169].

As a final remark, reference-guided genome assembly should be avoided for a bacterial phylum such as Bacteroidetes (and especially its *Porphyromonas* genus) that has high genomic plasticity and frequent repeats. The method is unreliable and is a source of errors due to the numerous genomic rearrangements. Long-read *de novo* assembly is clearly the strategy of choice for obtaining complete and accurate finished genomes. Even though the sequencing and assembly of complete genomes is expensive, time-consuming, and requires manual biocuration, it should be the goal for high-quality sequencing projects [175, 176]. In a bacterial species, having several consistently sequenced, assembled, annotated, and biocurated genomes is essential for comparative genomic studies, permitting the analysis of genomic plasticity and evolutionary mechanisms [177].

Conclusions

Current sequencing capacity is yielding more and more bacterial genomes at a continuously lower price, yet the vast majority of these projects release draft genomes. In agreement with previous observations, in this study we showed that assembly breaks are caused by genomic repeats that are equal to or longer in length than the sequencing reads. Nevertheless, these repeats encompass a vast variety of elements that are essential to genome organization, stability, and function. They are therefore inherent parts of genomes, yet most are not shown in draft genomes. When possible and according to the biological question to be answered, a complete finished genome should be the preferred aim of sequencing projects, and long-read sequencing makes this feasible. We have demonstrated that this technology allows the verification of bacterial genomes that have been sequenced, assembled and circularized using

massively parallel sequencing technologies. This method will also detect misassembly errors that are often associated with erroneous combinations of ribosomal operons or very long genomic islands. Finally, we have validated the importance of biocurating automatic annotations and have shown that a strategy based on comparative genomics is very powerful for improving both structural and functional annotations.

Figure titles and legends

Figure 1. Relatedness of complete Bacteroidia genomes for species having at least two different strains. a. Dendrogram of the inter-species relatedness calculated with the OrthoANI algorithm, clustered using UPGMA, and shown with the corresponding pairwise identity heatmap. **b.** Dendrogram of the intra-species relatedness, shown with the corresponding pairwise identity heatmap.

Figure 2. Genomic distribution of repeats (at least 3 copies) in each genome studied. Circos representations of each strain's chromosome, with *oriC* positioned at the first nucleotide of the *dnaA* gene. For each repeat, its first occurrence in the genome is the starting point of each line that links it to all of the other positions. As the copy number increases, the line colours range from light blue to red. The total number of repeats can be visualized as the number of intersections of the circular chromosome. Strains of the same species are grouped together and arranged in ascending order of repeat counts.

Figure 3. Genomic repeats in *Porphyromonas gingivalis* (*P. g.*) strains. From left to right, strain relatedness, genomic repeat distribution, and number of copies. The dendogram shows intra-species relatedness calculated with OrthoANI and clustered with UPGMA. The circular chromosome of each strain is presented using Circos, with *oriC* positioned at the top. For each repeat (at least 3 copies), its first occurrence in the genome is the starting point of the

lines that link it to all other positions. As the copy number increases, the lines go from light blue to red. On the right, the number of repeats by copy number. Since all repeats have at least 2 copies, the total number of repeats corresponds to the light blue bar.

Figure 4. A *de novo* genome assembly of *Porphyromonas gingivalis* artificial reads. a. Eleven programs were used for *de novo* assembly of the seven strains in study. The main cumulative lengths were calculated, and plotted here against the contig index. **b.** The three assemblers that produced the highest N50 were plotted in the same manner as in a. (upper panel), then the assembly was mapped to the reference and only the mapped contigs were plotted (lower panel). **c.** The number of contigs (A5-miseq and SPAdes) was plotted against the amount of repeats (with at least 3 copies). **d.** Identification of gaps: after assembly with A5-miseq or SPAdes, genomic regions not covered by contigs were extracted. The gaps were classified into five categories: genomic islands, ribosomal RNA (*rrn*) operons, coding sequences (CDS) with repeated domains, intergenic sequences, and insertion sequences or miniature inverted-repeat transposable element (IS/MITEs).

Figure 5. Functional comparison of the common coding sequences in two *Porphyromonas gingivalis* annotations. Comparison of **a.** an annotation available at the NCBI, and **b.** this study's manually biocurated annotation. For both, the common CDS were classified into five categories. Both pie charts reflect mean values.

Figure 6. Pangenome overview of ATCC 33277, TDC60, and W83 strains, focusing on accessory and unique genomes. The central triangle represents the core genome, which has at least 1522 genes (see text for details). Each corner is a *Porphyromonas gingivalis* (*P. g.*) strain, with a pie chart showing the unique genome's distribution of functions, with total and absolute counts shown. On each triangle side, stacked histograms show the accessory genome of the strains in the adjacent vertices. Total and absolute counts are shown, and the differences between strain numbers are due to paralogy.

Additional files: Figure and table titles and legends

Additional file 1. Supplementary Figure 1. NCBI genome database distribution of the main bacterial phyla associated with humans. a. Pie chart featuring the genomes present in the database, by phylum. **b.** Stacked bar chart of the incidence of the genomes belonging to the four main phyla associated with humans. Absolute counts are presented by phylum and classified as either finished/complete (having at least one chromosome and/or plasmid), or draft/incomplete (having multiple contigs or scaffolds).

Additional file 2. Supplementary Figure 2. Bacteroidetes genomes by class. a. On the left, a phylogenetic tree based on the 16S rRNA genes of complete genomes, grouped by class. A stacked bar chart then shows the number of genomes belonging to each Bacteroidetes class. The absolute genome counts are given, and classified as being either finished/complete or draft/incomplete (having multiple contigs or scaffolds). Pie charts on the right indicate the isolation sources for each genome: environmental (soil, fresh or marine water, and plants), animal (insects, molluscs, fish, birds, and mammals), or human (different body sites and health conditions). **b.** Stacked bar chart of Bacteroidia genomes grouped by status (complete or draft), presented by their publication year.

Additional file 3. Supplementary Figure 3. Bacteroidia draft genomes binned by genus and by species. a. Box plot of draft/incomplete Bacteroidia genomes grouped by genus. With the exception of *Tannerella* which has complete genomes, any genus with less than 10 draft genomes was classified as "other." The number of assemblies is presented above the plot, and the median is shown for each box. If a genus has drafts with more than 500 contigs/scaffolds, it is marked with ⊙: Bacteroides (n = 17, 557 to 4357 contigs); Parabacteroides (n = 2, 1471 and 1920 contigs); and Prevotella (n = 3, 553 to 3171 contigs). **b.** Box plot of

1007 draft/incomplete Bacteroidia genomes for which at least two complete genomes exist,
1008 grouped by species, as per **a**. The drafts which have over 500 contigs/scaffolds are
1009 *Bacteroides fragilis* (n = 7, 557 to 2566 contigs), *B. ovatus* (n = 1, 556 contigs), and *B.*
1010 *thetaiotaomicron* (n = 2, 1730 and 2372 contigs).

1011 **Additional file 4. Supplementary Figure 4. Genomic repeats by species.** Genomic repeats
1012 were identified for each genome, and the cumulative mean copy numbers and their standard
1013 deviations are presented. The light blue bar indicates the total number of repeats (at least 2
1014 copies).

1015 **Additional file 5. Supplementary Figure 5. *De novo* assembly of artificial reads of the**
1016 **studied Bacteroidia genomes. a.** QUAST graph (cumulative length versus config index) for
1017 each assembly of each strain. The left column shows all contigs (> 1 Kbp), while the right
1018 shows only the contigs that mapped to its reference. The dotted line represents the reference
1019 genome size. **b.** For all 24 genomes, the contig counts from A5-miseq and SPAdes were
1020 plotted against the repeat counts (with at least 3 copies). **c.** As b, but showing all seven *P.*
1021 *gingivalis* strains.

1022 **Additional file 6. Supplementary Figure 6. PCR validation of 3 *P. gingivalis* strain**
1023 **constructions. a.** Agarose gel electrophoresis (0.8% in 1X TBE buffer, stained with 1X
1024 GelRed) of PCR products for rrn operons. Primers used and verified strains are indicated by
1025 lane, and the primer names were simplified ("rrn" is not mentioned). DNA molecular-weight
1026 size markers were used in the first and last lanes. The first band is 9 Kb, and the second is 4
1027 Kb. **b.** Schematic representation (not to scale) of both copies ("a" and "b") of CTnPg1 from
1028 *P. gingivalis* ATCC 33277. The *de novo* assembly identified two complete copies with the
1029 same orientation. The published ATCC 33277 strain had two copies, but the second was
1030 partial and inverted when compared to the first. Size and orientation are presented for clarity,
1031 and the dotted line indicates the absent CTnPg1-b region in the published genome. Primers

1032 names were simplified ("ctnpg1_" is not mentioned). The ctnpg1_5in is red, the ctnpg1_3in is
1033 green, and all other primers are black. **c.** Agarose gel electrophoresis done as in **a.** for the
1034 primer combination indicated over each lane. In the order shown, the expected sizes are 7.5,
1035 3.5, 3.0, 3.5, 5.0, 6.5, and 10 Kb. **d.** Agarose gel electrophoresis as above for *rrn* operons in
1036 22 *P. gingivalis* strains. For PCR conditions, primer sequences, and strain origins, see
1037 Methods.

1038 **Additional file 7. Supplementary Figure 7. Whole-genome alignments of the three**
1039 **resequenced *P. gingivalis* strains.** Published and *de novo* assembled genome architectures
1040 are compared. Locally collinear blocks (LCBs) were detected using the progressiveMauve
1041 algorithm. Shown are translocations and inversions, insertions (green arrows), and deletions
1042 (thin red arrows), along with size differences and SNP counts. For ATCC 33277, the blue
1043 blocks represent CTnPg1 copies; in TDC60, they are the *rrn* operons. For W83, the *de novo*
1044 sequence "a" was assembled from the published genome's "b" and "c" sequences (see text for
1045 details).

1046 **Additional file 8. Supplementary Figure 8. Insertions and deletions in the *de novo P.***
1047 ***gingivalis* assemblies as compared to the published genomes.** In all cases, alignment zooms
1048 are presented for the corresponding insertions or deletions detailed in Fig. S8 and in the text.
1049 The upper genome is the *de novo* assembly and the bottom is the published one. **a.** The ATCC
1050 33277 strain has four indels. **b.** TDC has four indels, with an additional 22-bp deletion in an
1051 intergenic region (not depicted). **c.** W83 has a single insertion.

1052 **Additional file 9. Supplementary Figure 9. The three *P. gingivalis* strains binned by**
1053 **their CDS/pseudogene functions.** The coding sequences and pseudogenes were classified
1054 into five categories as shown, and the histogram is based on their absolute counts.

1055 **Additional file 10. Supplementary Figure 10. CDS/pseudogenes of all strains that have at**
1056 **least two copies in all three *P. gingivalis* genomes. a.** Horizontal histogram of absolute gene

counts binned by category. Three of these categories are related to nucleic acids (DNA helicases, regulators mainly containing the HTH 17 domain, and histones), while the remaining ones are related to mobile elements (integrases, transposases, tetracycline-resistance genes, and the *tra* conjugative transposons). **b.** Histogram of all transposases coded by genome and divided into CDS and pseudogenes. **c.** Heatmap table of transposase families separated into coding sequences and pseudogenes. Absolute numbers are presented, as the copy numbers grow, the cell color move from light blue to dark red.

Additional file 11. Supplementary Figure 11. Overview of the core genomes of *P. gingivalis* strains ATCC 33277, TDC60, and W83. a. Pie chart of genes present in all three strains grouped by categories: constant (more than 97% nucleotide identity, none has paralogs); variable (less than 97% nucleotide identity, none have paralogs); with paralogs (at least one strain has paralogs); and with pseudogenisation (at least one strain has a pseudogene, another a functional CDS, and none have paralogs). **b.** Constant core genes classified into five categories. **c.** Genes in variable core genome. The 47 genes are presented grouped by function. **d.** Core genes with paralogs. Gene names and products are listed, and the number of paralogs detailed by strain. To facilitate reading, cells were shaded when at least two paralogs exist. *, pseudogenes; †, a hypothetical gene clustered with genes from the BF0131 conjugative transposon.

Additional file 12. Table S1. Complete genomes not included in this study. When available, the species, strain, sequencing technology, assembler, coverage, Pubmed ID, release date, exclusion reason, and FTP link are presented for each genome. When no publication was found, genomes were marked as "unpublished," and the sequencing organism was identified. Key: BCoM, Baylor College of Medicine; BU, Bielefeld University; DOE-JGI, United States Department of Energy-Joint Genome Institute; FI, The Forsyth Institute; HMP, Human Microbiome Project; JCVI, J. Craig Venter Institute; LIAEPB, Leibniz Institute

for Agricultural Engineering Potsdam-Bornim; SI, Sanger Institute; and UoB, University of Bern.

Additional file 13. Table S2. The 166 draft genomes of the six Bacteroidia species studied here. Species, strain, sequencing technology, assembler, number of contigs, Pubmed ID, release date, and FTP links are presented. When no publication was found, genomes were marked as "unpublished" and the sequencing organism was identified. Key: *, no sequencing centre could be identified; BCoM, Baylor College of Medicine; BI, Broad Institute; DOE-JGI, United States Department of Energy-Joint Genome Institute; FMBA, Federal Medical-Biological Agency, Russia; HMP, Human Microbiome Project; IGS, Institute for Genome Science, University of Maryland; JCVI, J. Craig Venter Institute; TUD, Technical University of Denmark; UoS, University of Sheffield; and WU, Washington University.

Additional file 14. Table S3. CDS/pseudogene differences by strain. Comparison of the publically available annotations accessed via NCBI versus this study's annotation showed that coding sequences changed to pseudogenes, pseudogenes changed to CDS, features fused, single features split into two, and coding strands changed. See the Results for information on shared CDS and pseudogenes, eliminated CDS/genes, and all other features (tRNA, rRNA, ncRNA, tmRNA, riboswitches, mobile elements, signal peptides, and repeat regions).

Additional file 15. Table S4. Classification of additional CDS/pseudogenes. After a manual biocurated annotation, the changes were separated into five functional categories. The absolute counts of the new CDS/pseudogenes, CDS changed to pseudogenes or vice versa, feature fusions or splitting of single features into two, and coding strand changes are presented here, strain by strain. The results are emphasised via a heatmap that goes from lilac to burgundy. For re-annotation of the CDS/pseudogenes that are shared between the two versions, see Results and Fig. 4.

Additional file 16. Table S5. Core genes with pseudogenisation. Gene names and products are listed for all 36 genes that are present in this subset of core genes. An asterisk indicates the strains in which they are pseudogenised, and genes with pseudogenes in more than one strain are bold.

Declarations

Acknowledgements

The authors would like to thank Sandrine LeGall-David for technical advice and Martine Bonnaure-Mallet for general support (NuMeCan, Rennes, France).

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

All genomic details (including sequences and isolation information) are publicly available in the appropriate NCBI databases (https://www.ncbi.nlm.nih.gov). The resequenced strains from this study are commercially available from the ATCC (Catalog numbers ATCC® 33277™ and ATCC® BAA-308™ for *P. gingivalis* strains ATCC33277 and W83 respectively) or JCM (Catalog number 19600 for *P. gingivalis* strain TDC60). Raw sequencing reads and genomes were deposited to the NCBI SRA and Genome databases. The

BioProject accession number for this study is PRJNA393092.

Competing interests

The authors declare that they have no competing interests.

Funding

The PhD research of LA-A is funded by a scholarship from the Oficina de Asuntos Internacinales y Cooperación Externa (OAICE), Universidad de Costa Rica, San José, Costa Rica. OAICE had no role in the design of the study and collection, analysis, interpretation of data, nor in writing the manuscript.

Authors' contributions

FB-H and LA-A designed this project. They analyzed the databases, genomes, and *in silico* sequencing; did the genome assembly and finishing; annotated, biocurated, and did the comparative genomics; and wrote the manuscript including the tables and figures. LA-A performed the bacterial cultures and DNA extractions. AR prepared and sequenced the library. LA-A, AP, and EC verified and validated the genome. All authors read and approved the final manuscript.

References

1. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al*: **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature* 2010, **464**(7285):59-65.

2. Human Microbiome Project Consortium: **Structure, function and diversity of the healthy human microbiome**. *Nature* 2012, **486**(7402):207-214.
3. Hugon P, Dufour JC, Colson P, Fournier PE, Sallah K, Raoult D: **A comprehensive repertoire of prokaryotic species identified in human beings**. *The Lancet Infectious diseases* 2015, **15**(10):1211-1219.
4. Lloyd-Price J, Abu-Ali G, Huttenhower C: **The healthy human microbiome**. *Genome medicine* 2016, **8**(1):51.
5. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R: **Diversity, stability and resilience of the human gut microbiota**. *Nature* 2012, **489**(7415):220-230.
6. Robles Alonso V, Guarner F: **Linking the gut microbiota to human health**. *The British journal of nutrition* 2013, **109 Suppl 2**:S21-26.
7. Winter SE, Baumler AJ: **Why related bacterial species bloom simultaneously in the gut: principles underlying the 'Like will to like' concept**. *Cellular microbiology* 2014, **16**(2):179-184.
8. Hahnke RL, Meier-Kolthoff JP, Garcia-Lopez M, Mukherjee S, Huntemann M, Ivanova NN, Woyke T, Kyrpides NC, Klenk HP, Goker M: **Genome-Based Taxonomic Classification of Bacteroidetes**. *Frontiers in microbiology* 2016, **7**:2003.
9. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM *et al*: **Enterotypes of the human gut microbiome**. *Nature* 2011, **473**(7346):174-180.
10. Bull MJ, Plummer NT: **Part 1: The Human Gut Microbiome in Health and Disease**. *Integrative medicine (Encinitas, Calif)* 2014, **13**(6):17-22.
11. Johnson EL, Heaver SL, Walters WA, Ley RE: **Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes**. *Journal of molecular medicine (Berlin, Germany)* 2017, **95**(1):1-8.
12. Oren A, da Costa MS, Garrity GM, Rainey FA, Rossello-Mora R, Schink B, Sutcliffe I, Trujillo ME, Whitman WB: **Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes**. *Int J Syst Evol Microbiol* 2015, **65**(11):4284-4287.
13. Krieg NR, Ludwig W, Euzéby J, Whitman WB: **Phylum XIV. Bacteroidetes phyl. nov**. In: *Bergey's Manual® of Systematic Bacteriology: Volume Four The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes.* Edited by Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, Whitman WB. New York, NY: Springer New York; 2010: 25-469.
14. Thomas F, Hehemann JH, Rebuffet E, Czjzek M, Michel G: **Environmental and gut bacteroidetes: the food connection**. *Frontiers in microbiology* 2011, **2**:93.
15. Muñoz R, Rossello-Mora R, Amann R: **Revised phylogeny of Bacteroidetes and proposal of sixteen new taxa and two new combinations including Rhodothermaeota phyl. nov**. *Syst Appl Microbiol* 2016, **39**(5):281-296.
16. Smith DR: **Goodbye genome paper, hello genome report: the increasing popularity of 'genome announcements' and their impact on science**. *Briefings in functional genomics* 2017, **16**(3):156-162.
17. Reuter JA, Spacek DV, Snyder MP: **High-throughput sequencing technologies**. *Molecular cell* 2015, **58**(4):586-597.
18. Heather JM, Chain B: **The sequence of sequencers: The history of sequencing DNA**. *Genomics* 2016, **107**(1):1-8.

19. Kyrpides NC, Hugenholtz P, Eisen JA, Woyke T, Goker M, Parker CT, Amann R, Beck BJ, Chain PS, Chun J *et al*: **Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains**. *PLoS biology* 2014, **12**(8):e1001920.
20. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ *et al*: **A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea**. *Nature* 2009, **462**(7276):1056-1060.
21. Ventura M, Canchaya C, Fitzgerald GF, Gupta RS, van Sinderen D: **Genomics as a means to understand bacterial phylogeny and ecological adaptation: the case of bifidobacteria**. *Antonie van Leeuwenhoek* 2007, **91**(4):351-372.
22. Forde BM, O'Toole PW: **Next-generation sequencing technologies and their impact on microbial genomics**. *Briefings in functional genomics* 2013, **12**(5):440-453.
23. Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-generation sequencing technologies**. *Nature reviews Genetics* 2016, **17**(6):333-351.
24. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B *et al*: **Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph**. *Briefings in functional genomics* 2012, **11**(1):25-37.
25. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions**. *Nature reviews Genetics* 2011, **13**(1):36-46.
26. Williams D, Trimble WL, Shilts M, Meyer F, Ochman H: **Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes**. *BMC genomics* 2013, **14**:537.
27. Kamada M, Hase S, Sato K, Toyoda A, Fujiyama A, Sakakibara Y: **Whole genome complete resequencing of Bacillus subtilis natto by combining long reads with high-quality short reads**. *PloS one* 2014, **9**(10):e109999.
28. Shapiro JA, von Sternberg R: **Why repetitive DNA is essential to genome function**. *Biological reviews of the Cambridge Philosophical Society* 2005, **80**(2):227-250.
29. Avershina E, Rudi K: **Dominant short repeated sequences in bacterial genomes**. *Genomics* 2015, **105**(3):175-181.
30. Nagarajan N, Cook C, Di Bonaventura M, Ge H, Richards A, Bishop-Lilly KA, DeSalle R, Read TD, Pop M: **Finishing genomes with limited resources: lessons from an ensemble of microbial genomes**. *BMC genomics* 2010, **11**:242.
31. Bao E, Jiang T, Girke T: **AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references**. *Bioinformatics (Oxford, England)* 2014, **30**(12):i319-i328.
32. Kolmogorov M, Raney B, Paten B, Pham S: **Ragout-a reference-assisted assembly tool for bacterial genomes**. *Bioinformatics (Oxford, England)* 2014, **30**(12):i302-309.
33. Salzberg SL, Yorke JA: **Beware of mis-assembled genomes**. *Bioinformatics (Oxford, England)* 2005, **21**(24):4320-4321.
34. Vincent AT, Derome N, Boyle B, Culley AI, Charette SJ: **Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money**. *Journal of microbiological methods* 2016.
35. Dayarian A, Michael TP, Sengupta AM: **SOPRA: Scaffolding algorithm for paired reads via statistical optimization**. *BMC bioinformatics* 2010, **11**:345.
36. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC *et al*: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology**. *PloS one* 2012, **7**(11):e47768.

37. Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, Aury JM: **Genome assembly using Nanopore-guided long and error-free DNA reads**. *BMC genomics* 2015, **16**:327.
38. Mariano DC, Sousa Tde J, Pereira FL, Aburjaile F, Barh D, Rocha F, Pinto AC, Hassan SS, Saraiva TD, Dorella FA *et al*: **Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002**. *BMC genomics* 2016, **17**:315.
39. Madoui MA, Dossat C, d'Agata L, van Oeveren J, van der Vossen E, Aury JM: **MaGuS: a tool for quality assessment and scaffolding of genome assemblies with Whole Genome Profiling Data**. *BMC bioinformatics* 2016, **17**:115.
40. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T *et al*: **Insights from 20 years of bacterial genome sequencing**. *Functional & integrative genomics* 2015, **15**(2):141-161.
41. Howison M, Zapata F, Dunn CW: **Toward a statistically explicit understanding of de novo sequence assembly**. *Bioinformatics (Oxford, England)* 2013, **29**(23):2959-2963.
42. Barbosa EG, Aburjaile FF, Ramos RT, Carneiro AR, Le Loir Y, Baumbach J, Miyoshi A, Silva A, Azevedo V: **Value of a newly sequenced bacterial genome**. *World journal of biological chemistry* 2014, **5**(2):161-168.
43. Stepanov VG, Tirumalai MR, Montazari S, Checinska A, Venkateswaran K, Fox GE: **Bacillus pumilus SAFR-032 Genome Revisited: Sequence Update and Re-Annotation**. *PloS one* 2016, **11**(6):e0157331.
44. Klassen JL, Currie CR: **Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation**. *BMC genomics* 2012, **13**:14.
45. Fierst JL: **Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools**. *Frontiers in genetics* 2015, **6**:220.
46. Turroni F, van Sinderen D, Ventura M: **Genomics and ecological overview of the genus Bifidobacterium**. *International journal of food microbiology* 2011, **149**(1):37-44.
47. Periwal V, Scaria V: **Insights into structural variations and genome rearrangements in prokaryotic genomes**. *Bioinformatics (Oxford, England)* 2015, **31**(1):1-9.
48. R Core Team: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria; 2016.
49. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic acids research* 2002, **30**(14):3059-3066.
50. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C *et al*: **Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data**. *Bioinformatics (Oxford, England)* 2012, **28**(12):1647-1649.
51. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood**. *Systematic biology* 2003, **52**(5):696-704.
52. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA**. *Journal of molecular evolution* 1985, **22**(2):160-174.
53. Lee I, Kim YO, Park SC, Chun J: **OrthoANI: An improved algorithm and software for calculating average nucleotide identity**. *International journal of systematic and evolutionary microbiology* 2015.

54. Treangen TJ, Darling AE, Achaz G, Ragan MA, Messeguer X, Rocha EP: **A novel heuristic for local multiple alignment of interspersed DNA repeats**. *IEEE/ACM transactions on computational biology and bioinformatics* 2009, **6**(2):180-189.
55. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics**. *Genome research* 2009, **19**(9):1639-1645.
56. Huang W, Li L, Myers JR, Marth GT: **ART: a next-generation sequencing read simulator**. *Bioinformatics (Oxford, England)* 2012, **28**(4):593-594.
57. Coil D, Jospin G, Darling AE: **A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data**. *Bioinformatics (Oxford, England)* 2015, **31**(4):587-589.
58. Li H: **Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly**. *Bioinformatics (Oxford, England)* 2012, **28**(14):1838-1844.
59. Chikhi R, Rizk G: **Space-efficient and exact de Bruijn graph representation based on a Bloom filter**. *Algorithms for molecular biology : AMB* 2013, **8**(1):22.
60. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome research* 2004, **14**(6):1147-1159.
61. Zhu X, Leung HC, Chin FY, Yiu SM, Quan G, Liu B, Wang Y: **PERGA: a paired-end read guided de novo assembler for extending contigs using SVM and look ahead approach**. *PloS one* 2014, **9**(12):e114253.
62. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al*: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler**. *GigaScience* 2012, **1**(1):18.
63. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *Journal of computational biology : a journal of computational molecular cell biology* 2012, **19**(5):455-477.
64. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs**. *Genome research* 2008, **18**(5):821-829.
65. Zhang J, Kobert K, Flouri T, Stamatakis A: **PEAR: a fast and accurate Illumina Paired-End reAd mergeR**. *Bioinformatics (Oxford, England)* 2014, **30**(5):614-620.
66. Chikhi R, Medvedev P: **Informed and automated k-mer size selection for genome assembly**. *Bioinformatics (Oxford, England)* 2014, **30**(1):31-37.
67. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies**. *Bioinformatics (Oxford, England)* 2013, **29**(8):1072-1075.
68. Mariette J, Escudie F, Allias N, Salin G, Noirot C, Thomas S, Klopp C: **NG6: Integrated next generation sequencing storage and processing environment**. *BMC genomics* 2012, **13**:462.
69. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation**. *Genome research* 2017, **27**(5):722-736.
70. Perez-Chaparro PJ, Lafaurie GI, Gracieux P, Meuric V, Tamanai-Shacoori Z, Castellanos JE, Bonnaure-Mallet M: **Distribution of Porphyromonas gingivalis fimA genotypes in isolates from subgingival plaque and blood sample during bacteremia**. *Biomedica : revista del Instituto Nacional de Salud* 2009, **29**(2):298-306.
71. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics (Oxford, England)* 2014, **30**(14):2068-2069.

72. Kremer FS, Eslabao MR, Dellagostin OA, Pinto LD: **Genix: a new online automated pipeline for bacterial genome annotation**. *FEMS microbiology letters* 2016, **363**(23).
73. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD *et al*: **RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes**. *Scientific reports* 2015, **5**:8365.
74. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA: **Genenames.org: the HGNC resources in 2015**. *Nucleic acids research* 2015, **43**(Database issue):D1079-1085.
75. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of molecular biology* 1990, **215**(3):403-410.
76. Marchler-Bauer A, Bryant SH: **CD-Search: protein domain annotations on the fly**. *Nucleic acids research* 2004, **32**(Web Server issue):W327-331.
77. Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, Le Fevre F, Longin C, Mornico D, Roche D *et al*: **MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data**. *Nucleic acids research* 2013, **41**(Database issue):D636-647.
78. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nature methods* 2011, **8**(10):785-786.
79. Gomi M, Sonoyama M, Mitaku S: **High performance system for signal peptide prediction: SOSUIsignal**. *Chem-bio informatics journal* 2004, **4**(4):142-147.
80. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ *et al*: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes**. *Bioinformatics (Oxford, England)* 2010, **26**(13):1608-1615.
81. Yu CS, Lin CJ, Hwang JK: **Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions**. *Protein science : a publication of the Protein Society* 2004, **13**(5):1402-1406.
82. Berven FS, Flikka K, Jensen HB, Eidhammer I: **BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria**. *Nucleic acids research* 2004, **32**(Web Server issue):W394-399.
83. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A: **Prediction of lipoprotein signal peptides in Gram-negative bacteria**. *Protein science : a publication of the Protein Society* 2003, **12**(8):1652-1662.
84. Babu MM, Priya ML, Selvan AT, Madera M, Gough J, Aravind L, Sankaran K: **A database of bacterial lipoproteins (DOLOP) with functional assignments to predicted lipoproteins**. *Journal of bacteriology* 2006, **188**(8):2761-2773.
85. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M: **ISfinder: the reference centre for bacterial insertion sequences**. *Nucleic acids research* 2006, **34**(Database issue):D32-36.
86. Grissa I, Vergnaud G, Pourcel C: **CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats**. *Nucleic acids research* 2007, **35**(Web Server issue):W52-57.
87. Biswas A, Staals RH, Morales SE, Fineran PC, Brown CM: **CRISPRDetect: A flexible algorithm to define CRISPR arrays**. *BMC genomics* 2016, **17**:356.
88. Rousseau C, Gonnet M, Le Romancer M, Nicolas J: **CRISPI: a CRISPR interactive database**. *Bioinformatics (Oxford, England)* 2009, **25**(24):3317-3318.

89. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P: **CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats**. *BMC bioinformatics* 2007, **8**:209.
90. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL, Brinkman FS: **IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets**. *Nucleic acids research* 2017.
91. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R: **Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models**. *BMC bioinformatics* 2006, **7**:142.
92. Hsiao W, Wan I, Jones SJ, Brinkman FS: **IslandPath: aiding detection of genomic islands in prokaryotes**. *Bioinformatics (Oxford, England)* 2003, **19**(3):418-420.
93. Langille MG, Hsiao WW, Brinkman FS: **Evaluation of genomic island predictors using a comparative genomics approach**. *BMC bioinformatics* 2008, **9**:329.
94. Che D, Hasan MS, Wang H, Fazekas J, Huang J, Liu Q: **EGID: an ensemble algorithm for improved genomic island detection in genomic sequences**. *Bioinformation* 2011, **7**(6):311-314.
95. Vernikos GS, Parkhill J: **Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands**. *Bioinformatics (Oxford, England)* 2006, **22**(18):2196-2203.
96. Shrivastava S, Reddy Ch V, Mande SS: **INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms**. *Journal of biosciences* 2010, **35**(3):351-364.
97. Tu Q, Ding D: **Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis**. *FEMS microbiology letters* 2003, **221**(2):269-275.
98. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement**. *PloS one* 2010, **5**(6):e11147.
99. Konstantinidis KT, Ramette A, Tiedje JM: **The bacterial species definition in the genomic era**. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2006, **361**(1475):1929-1940.
100. Naito M, Hirakawa H, Yamashita A, Ohara N, Shoji M, Yukitake H, Nakayama K, Toh H, Yoshimura F, Kuhara S *et al*: **Determination of the genome sequence of Porphyromonas gingivalis strain ATCC 33277 and genomic comparison with strain W83 revealed extensive genome rearrangements in P. gingivalis**. *DNA research : an international journal for rapid publication of reports on genes and genomes* 2008, **15**(4):215-225.
101. Chastain-Gross RP, Xie G, Belanger M, Kumar D, Whitlock JA, Liu L, Raines SM, Farmerie WG, Daligault HE, Han CS *et al*: **Genome Sequence of Porphyromonas gingivalis Strain 381**. *Genome Announc* 2017, **5**(2).
102. Dashper SG, Mitchell HL, Seers CA, Gladman SL, Seemann T, Bulach DM, Chandry PS, Cross KJ, Cleal SM, Reynolds EC: **Porphyromonas gingivalis Uses Specific Domain Rearrangements and Allelic Exchange to Generate Diversity in Surface Virulence Factors**. *Frontiers in microbiology* 2017, **8**:48.
103. Klein BA, Chen T, Scott JC, Koenigsberg AL, Duncan MJ, Hu LT: **Identification and characterization of a minisatellite contained within a novel miniature inverted-repeat transposable element (MITE) of Porphyromonas gingivalis**. *Mobile DNA* 2015, **6**:18.
104. Hovik H, Yu WH, Olsen I, Chen T: **Comprehensive transcriptome analysis of the periodontopathogenic bacterium Porphyromonas gingivalis W83**. *Journal of bacteriology* 2012, **194**(1):100-114.

105. Ochman H: **Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes**. *Trends in genetics : TIG* 2002, **18**(7):335-337.
106. Lawrence J: **When ELFs are ORFs, but don't act like them**. *Trends in genetics : TIG* 2003, **19**(3):131-132.
107. Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P, Yang G: **Miniature inverted-repeat transposable elements: discovery, distribution, and activity**. *Genome* 2013, **56**(9):475-486.
108. Bocs S, Danchin A, Medigue C: **Re-annotation of genome microbial coding-sequences: finding new genes and inaccurately annotated genes**. *BMC bioinformatics* 2002, **3**:5.
109. Califano JV, Kitten T, Lewis JP, Macrina FL, Fleischmann RD, Fraser CM, Duncan MJ, Dewhirst FE: **Characterization of Porphyromonas gingivalis insertion sequence-like element ISPg5**. *Infection and immunity* 2000, **68**(9):5247-5253.
110. Szafranski SP, Winkel A, Stiesch M: **The use of bacteriophages to biocontrol oral biofilms**. *Journal of biotechnology* 2017, **250**:29-44.
111. Lenhart JS, Schroeder JW, Walsh BW, Simmons LA: **DNA repair and genome maintenance in Bacillus subtilis**. *Microbiology and molecular biology reviews : MMBR* 2012, **76**(3):530-564.
112. Heinken A, Sahoo S, Fleming RM, Thiele I: **Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut**. *Gut microbes* 2013, **4**(1):28-40.
113. Garza DR, Van Verk MC, Huynen MA, Dutilh BE: **Bottom-up ecology of the human microbiome: from metagenomes to metabolomes**. *bioRxiv* 2016:060673.
114. Magnusdottir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jager C, Baginska J, Wilmes P *et al*: **Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota**. *Nature biotechnology* 2017, **35**(1):81-89.
115. Papanicolaou A: **The life cycle of a genome project: perspectives and guidelines inspired by insect genome projects**. *F1000Research* 2016, **5**:18.
116. Field D, Wilson G, van der Gast C: **How do we compare hundreds of bacterial genomes?** *Current opinion in microbiology* 2006, **9**(5):499-504.
117. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I: **RefSeq microbial genomes database: new representation and annotation strategy**. *Nucleic acids research* 2014, **42**(Database issue):D553-559.
118. Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Goker M, Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D *et al*: **1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life**. *Nature biotechnology* 2017, **35**(7):676-683.
119. Galperin MY, Koonin EV: **From complete genome sequence to 'complete' understanding?** *Trends in biotechnology* 2010, **28**(8):398-406.
120. Siddiqui H, Yoder-Himes DR, Mizgalska D, Nguyen KA, Potempa J, Olsen I: **Genome Sequence of Porphyromonas gingivalis Strain HG66 (DSM 28984)**. *Genome announcements* 2014, **2**(5).
121. McLean JS, Lombardo MJ, Ziegler MG, Novotny M, Yee-Greenbaum J, Badger JH, Tesler G, Nurk S, Lesin V, Brami D *et al*: **Genome of the pathogen Porphyromonas gingivalis recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform**. *Genome research* 2013, **23**(5):867-877.
122. Chen T, Siddiqui H, Olsen I: **In silico Comparison of 19 Porphyromonas gingivalis Strains in Genomics, Phylogenetics, Phylogenomics and Functional Genomics**. *Frontiers in cellular and infection microbiology* 2017, **7**:28.

123. Mavromatis K, Land ML, Brettin TS, Quest DJ, Copeland A, Clum A, Goodwin L, Woyke T, Lapidus A, Klenk HP *et al*: **The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation**. *PloS one* 2012, **7**(12):e48837.
124. Utturkar SM, Klingeman DM, Hurt RA, Jr., Brown SD: **A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies**. *Frontiers in microbiology* 2017, **8**:1272.
125. Fadeev E, De Pascale F, Vezzi A, Hubner S, Aharonovich D, Sher D: **Why Close a Bacterial Genome? The Plasmid of Alteromonas Macleodii HOT1A3 is a Vector for Inter-Specific Transfer of a Flexible Genomic Island**. *Frontiers in microbiology* 2016, **7**:248.
126. Kryukov K, Imanishi T: **Human Contamination in Public Genome Assemblies**. *PloS one* 2016, **11**(9):e0162424.
127. Mallet L, Bitard-Feildel T, Cerutti F, Chiapello H: **PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies**. *Bioinformatics (Oxford, England)* 2017.
128. Thomma B, Seidl MF, Shi-Kunne X, Cook DE, Bolton MD, van Kan JAL, Faino L: **Mind the gap; seven reasons to close fragmented genome assemblies**. *Fungal genetics and biology : FG & B* 2016, **90**:24-30.
129. Mardis E, McPherson J, Martienssen R, Wilson RK, McCombie WR: **What is finished, and why does it matter**. *Genome research* 2002, **12**(5):669-671.
130. Riba-Grognuz O, Keller L, Falquet L, Xenarios I, Wurm Y: **Visualization and quality assessment of de novo genome assemblies**. *Bioinformatics (Oxford, England)* 2011, **27**(24):3425-3426.
131. Baker M: **De novo genome assembly: what every biologist should know**. *Nat Meth* 2012, **9**(4):333-337.
132. Khiste N, Ilie L: **LASER: Large genome ASsembly EvaluatoR**. *BMC research notes* 2015, **8**:709.
133. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M *et al*: **Assemblathon 1: a competitive assessment of de novo short read assembly methods**. *Genome research* 2011, **21**(12):2224-2241.
134. Dias Z, Dias U, Setubal JC: **SIS: a program to generate draft genome sequence scaffolds for prokaryotes**. *BMC bioinformatics* 2012, **13**:96.
135. Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL: **GAGE-B: an evaluation of genome assemblers for bacterial organisms**. *Bioinformatics (Oxford, England)* 2013, **29**(14):1718-1725.
136. Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A: **Icarus: visualizer for de novo assembly evaluation**. *Bioinformatics (Oxford, England)* 2016, **32**(21):3321-3323.
137. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms**. *Nature biotechnology* 2012, **30**(5):434-439.
138. Hunt M, Newbold C, Berriman M, Otto TD: **A comprehensive evaluation of assembly scaffolding tools**. *Genome biology* 2014, **15**(3):R42.
139. Treangen TJ, Abraham AL, Touchon M, Rocha EP: **Genesis, effects and fates of repeats in prokaryotic genomes**. *FEMS microbiology reviews* 2009, **33**(3):539-571.
140. Touchon M, Rocha EP: **Causes of insertion sequences abundance in prokaryotic genomes**. *Molecular biology and evolution* 2007, **24**(4):969-981.
141. Newton IL, Bordenstein SR: **Correlations between bacterial ecology and mobile DNA**. *Current microbiology* 2011, **62**(1):198-208.

142. Schatz MC, Delcher AL, Salzberg SL: **Assembly of large genomes using second-generation sequencing**. *Genome research* 2010, **20**(9):1165-1173.
143. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW: **Extensive error in the number of genes inferred from draft genome assemblies**. *PLoS computational biology* 2014, **10**(12):e1003998.
144. Enersen M, Olsen I, van Winkelhoff AJ, Caugant DA: **Multilocus sequence typing of Porphyromonas gingivalis strains from different geographic origins**. *Journal of clinical microbiology* 2006, **44**(1):35-41.
145. Igboin CO, Griffen AL, Leys EJ: **Porphyromonas gingivalis strain diversity**. *Journal of clinical microbiology* 2009, **47**(10):3073-3081.
146. Enersen M: **Porphyromonas gingivalis: a clonal pathogen?: Diversities in housekeeping genes and the major fimbriae gene**. *Journal of oral microbiology* 2011, **3**.
147. Dolgilevich S, Rafferty B, Luchinskaya D, Kozarov E: **Genomic comparison of invasive and rare non-invasive strains reveals Porphyromonas gingivalis genetic polymorphisms**. *Journal of oral microbiology* 2011, **3**.
148. Tribble GD, Rigney TW, Dao DH, Wong CT, Kerr JE, Taylor BE, Pacha S, Kaplan HB: **Natural competence is a major mechanism for horizontal DNA transfer in the oral pathogen Porphyromonas gingivalis**. *mBio* 2012, **3**(1).
149. Tribble GD, Lamont GJ, Progulske-Fox A, Lamont RJ: **Conjugal transfer of chromosomal DNA contributes to genetic variation in the oral pathogen Porphyromonas gingivalis**. *Journal of bacteriology* 2007, **189**(17):6382-6388.
150. Feil EJ, Spratt BG: **Recombination and the population structures of bacterial pathogens**. *Annual review of microbiology* 2001, **55**:561-590.
151. Tibayrenc M, Ayala FJ: **How clonal are Neisseria species? The epidemic clonality model revisited**. *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**(29):8909-8913.
152. Riedel T, Bunk B, Thurmer A, Sproer C, Brzuszkiewicz E, Abt B, Gronow S, Liesegang H, Daniel R, Overmann J: **Genome Resequencing of the Virulent and Multidrug-Resistant Reference Strain Clostridium difficile 630**. *Genome announcements* 2015, **3**(2).
153. Malone KM, Farrell D, Stuber TP, Schubert OT, Aebersold R, Robbe-Austerman S, Gordon SV: **Updated Reference Genome Sequence and Annotation of Mycobacterium bovis AF2122/97**. *Genome announcements* 2017, **5**(14).
154. Chen T, Hosogi Y, Nishikawa K, Abbey K, Fleischmann RD, Walling J, Duncan MJ: **Comparative whole-genome analysis of virulent and avirulent strains of Porphyromonas gingivalis**. *Journal of bacteriology* 2004, **186**(16):5473-5479.
155. Slots J, Gibbons RJ: **Attachment of Bacteroides melaninogenicus subsp. asaccharolyticus to oral surfaces and its possible role in colonization of the mouth and of periodontal pockets**. *Infection and immunity* 1978, **19**(1):254-264.
156. Loos BG, Mayrand D, Genco RJ, Dickinson DP: **Genetic heterogeneity of Porphyromonas (Bacteroides) gingivalis by genomic DNA fingerprinting**. *Journal of dental research* 1990, **69**(8):1488-1493.
157. Guo FB, Xiong L, Teng JL, Yuen KY, Lau SK, Woo PC: **Re-annotation of protein-coding genes in 10 complete genomes of Neisseriaceae family by combining similarity-based and composition-based methods**. *DNA research : an international journal for rapid publication of reports on genes and genomes* 2013, **20**(3):273-286.
158. Zhang HX, Li SJ, Zhou HQ: **Evaluating the annotation of protein-coding genes in bacterial genomes: Chloroflexus aurantiacus strain J-10-fl and Natrinema sp J7-2 as case studies**. *Genetics and molecular research : GMR* 2014, **13**(4):10891-10897.

159. Richardson EJ, Watson M: **The automatic annotation of bacterial genomes**. *Briefings in bioinformatics* 2013, **14**(1):1-12.
160. Indrischek H, Wieseke N, Stadler PF, Prohaska SJ: **The paralog-to-contig assignment problem: high quality gene models from fragmented assemblies**. *Algorithms for molecular biology : AMB* 2016, **11**:1.
161. Vernikos G, Medini D, Riley DR, Tettelin H: **Ten years of pan-genome analyses**. *Current opinion in microbiology* 2015, **23**:148-154.
162. Touchon M, Rocha EP: **Coevolution of the Organization and Structure of Prokaryotic Genomes**. *Cold Spring Harbor perspectives in biology* 2016, **8**(1):a018168.
163. Brunner J, Wittink FR, Jonker MJ, de Jong M, Breit TM, Laine ML, de Soet JJ, Crielaard W: **The core genome of the anaerobic oral pathogenic bacterium Porphyromonas gingivalis**. *BMC microbiology* 2010, **10**:252.
164. Klein BA, Tenorio EL, Lazinski DW, Camilli A, Duncan MJ, Hu LT: **Identification of essential genes of the periodontal pathogen Porphyromonas gingivalis**. *BMC genomics* 2012, **13**:578.
165. Hutcherson JA, Gogeneni H, Yoder-Himes D, Hendrickson EL, Hackett M, Whiteley M, Lamont RJ, Scott DA: **Comparison of inherently essential genes of Porphyromonas gingivalis identified in two transposon-sequencing libraries**. *Molecular oral microbiology* 2016, **31**(4):354-364.
166. Korona R: **Gene dispensability**. *Current opinion in biotechnology* 2011, **22**(4):547-551.
167. Ijaq J, Chandrasekharan M, Poddar R, Bethi N, Sundararajan VS: **Annotation and curation of uncharacterized proteins- challenges**. *Frontiers in genetics* 2015, **6**:119.
168. Fischer D, Eisenberg D: **Finding families for genomic ORFans**. *Bioinformatics (Oxford, England)* 1999, **15**(9):759-762.
169. Watanabe T, Maruyama F, Nozawa T, Aoki A, Okano S, Shibata Y, Oshima K, Kurokawa K, Hattori M, Nakagawa I *et al*: **Complete genome sequence of the bacterium Porphyromonas gingivalis TDC60, which causes periodontal disease**. *Journal of bacteriology* 2011, **193**(16):4259-4260.
170. Tribble GD, Kerr JE, Wang BY: **Genetic diversity in the oral pathogen Porphyromonas gingivalis: molecular mechanisms and biological consequences**. *Future microbiology* 2013, **8**(5):607-620.
171. Kerr JE, Abramian JR, Dao DH, Rigney TW, Fritz J, Pham T, Gay I, Parthasarathy K, Wang BY, Zhang W *et al*: **Genetic exchange of fimbrial alleles exemplifies the adaptive virulence strategy of Porphyromonas gingivalis**. *PloS one* 2014, **9**(3):e91696.
172. Lamont RJ, Jenkinson HF: **Life below the gum line: pathogenic mechanisms of Porphyromonas gingivalis**. *Microbiology and molecular biology reviews : MMBR* 1998, **62**(4):1244-1263.
173. How KY, Song KP, Chan KG: **Porphyromonas gingivalis: An Overview of Periodontopathic Pathogen below the Gum Line**. *Frontiers in microbiology* 2016, **7**:53.
174. Nelson KE, Fleischmann RD, DeBoy RT, Paulsen IT, Fouts DE, Eisen JA, Daugherty SC, Dodson RJ, Durkin AS, Gwinn M *et al*: **Complete genome sequence of the oral pathogenic Bacterium porphyromonas gingivalis strain W83**. *Journal of bacteriology* 2003, **185**(18):5591-5601.
175. Koren S, Phillippy AM: **One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly**. *Current opinion in microbiology* 2015, **23**:110-120.

176. Teng JLL, Yeung ML, Chan E, Jia L, Lin CH, Huang Y, Tse H, Wong SSY, Sham PC, Lau SKP *et al*: **PacBio But Not Illumina Technology Can Achieve Fast, Accurate and Complete Closure of the High GC, Complex Burkholderia pseudomallei Two-Chromosome Genome**. *Frontiers in microbiology* 2017, **8**:1448.
177. Craddock T, Harwood CR, Hallinan J, Wipat A: **e-Science: relieving bottlenecks in large-scale genome analyses**. *Nature reviews Microbiology* 2008, **6**(12):948-954.

List of abbreviations

ANI	Average Nucleotide Identity
CDS	Coding DNA Sequence
COG	Cluster of Orthologous Groups (of proteins)
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTnPg	Conjugative Transposon of Porphyromonas gingivalis
DUF	Domain of Unknown Function
ELF	Evil-Little Fellows
FIG	Fellowship for Interpretation of Genomes
GEBA	Genomic Encyclopedia of Bacteria and Archaea
HGT	Horizontal Gene Transfer
ICE	Integrative and Conjugative Element
IQR	Interquartile Ranges
IS	Insertion Sequence
LCB	Locally Collinear Blocks
ME	Mobile Element
MITE	Miniature Inverted-repeat Transposable Elements
NAP	Nucleoid-Associated Proteins
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing, also SGS or massively parallel sequencing
NNI	Nearest-Neighbor Interchange
OLC	Overlap-Layout-Consensus
ORF	Open Reading Frame
rrn	ribosomal RNA operon
SGS	Second Generatation Sequencing, also NGS or massively parallel sequencing
SNP	Single Nucleotide Polymorphism
SPR	Subtree Pruning and Regrafting
WGS	Whole Genome Sequencing

Figure 1. Relatedness of complete Bacteroidia genomes for species having at least two different strains.

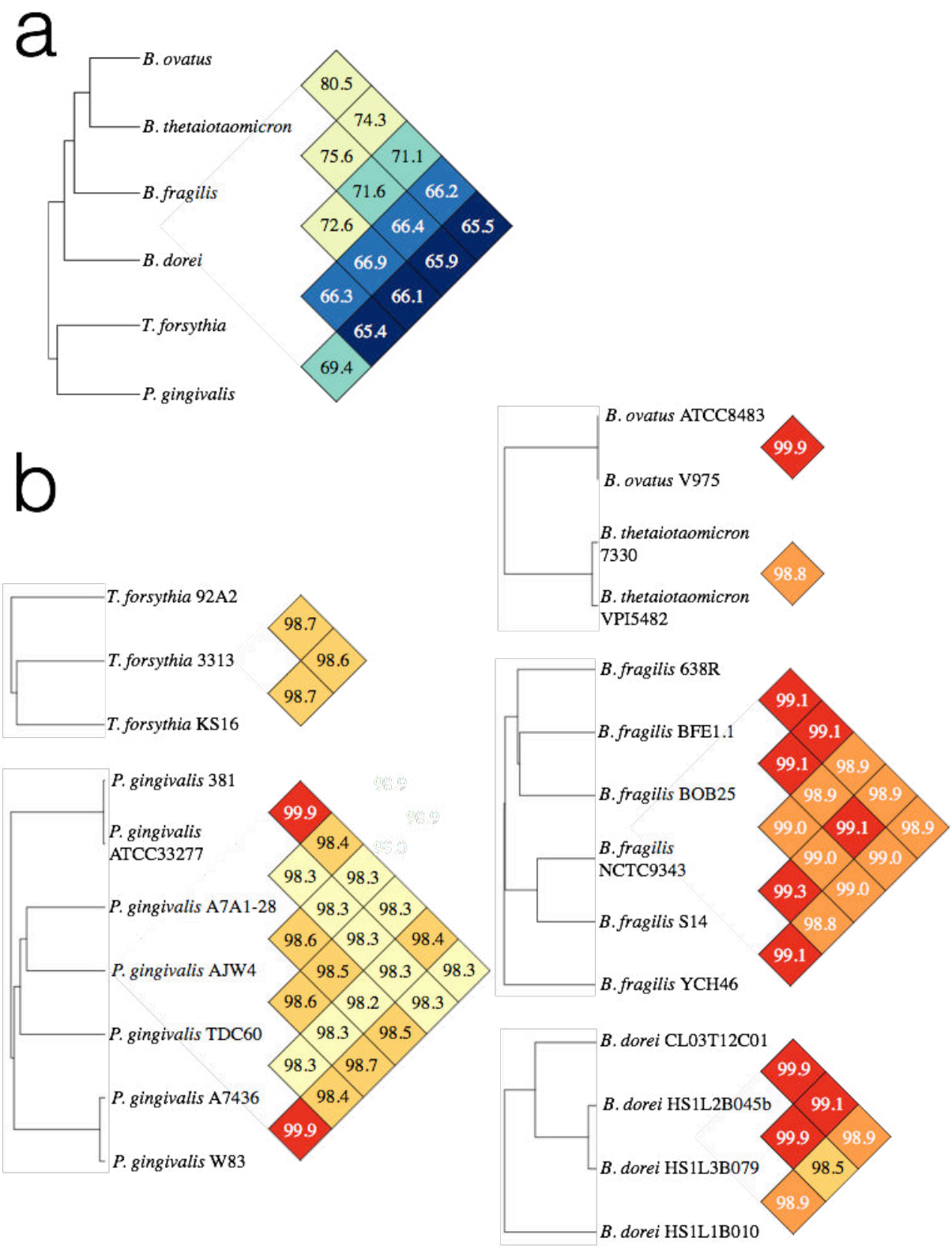


Figure 2. Genomic distribution of repeats (at least 3 copies) in each genome studied.

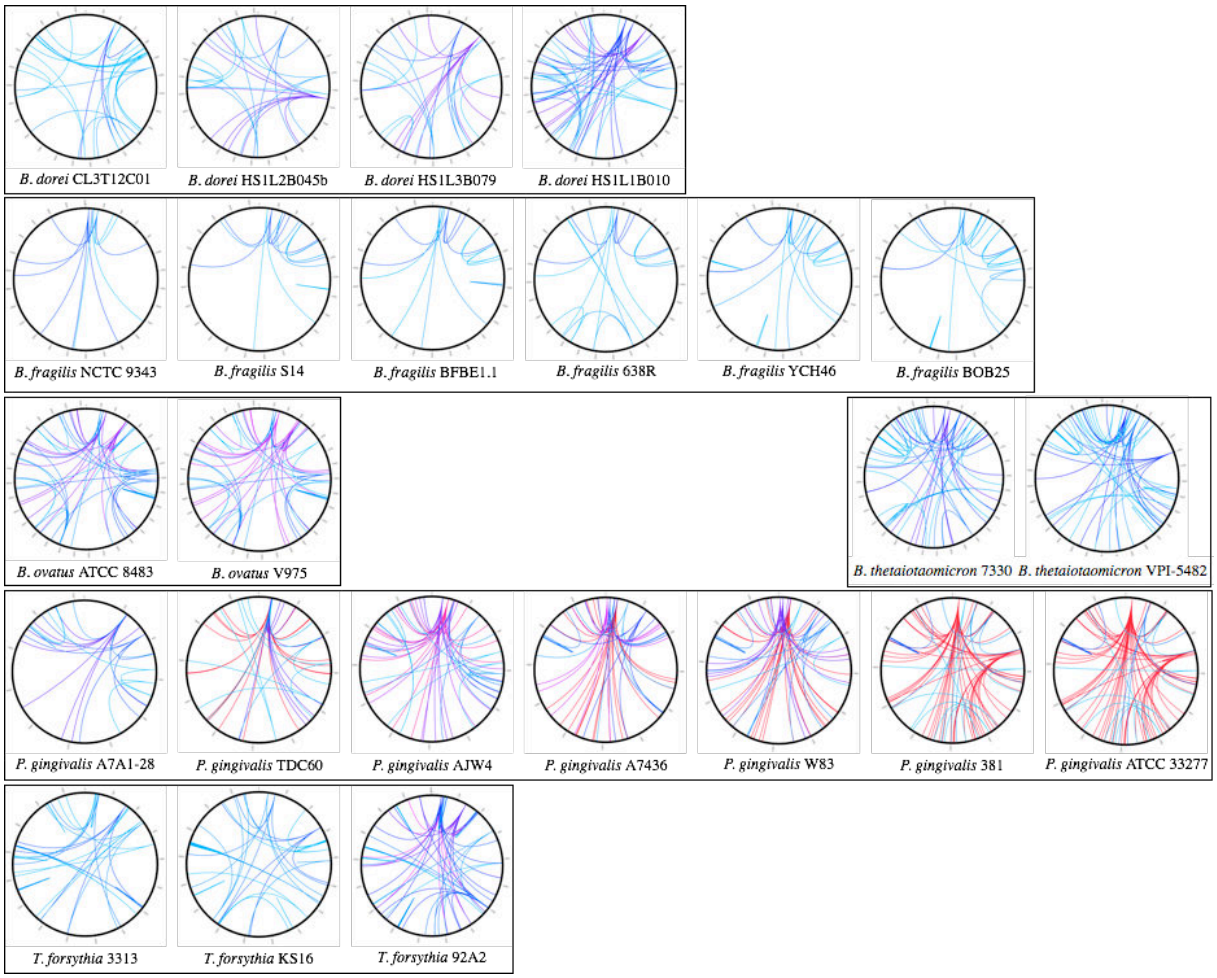


Figure 3. Genomic repeats in *Porphyromonas gingivalis* (P.g.) strains.

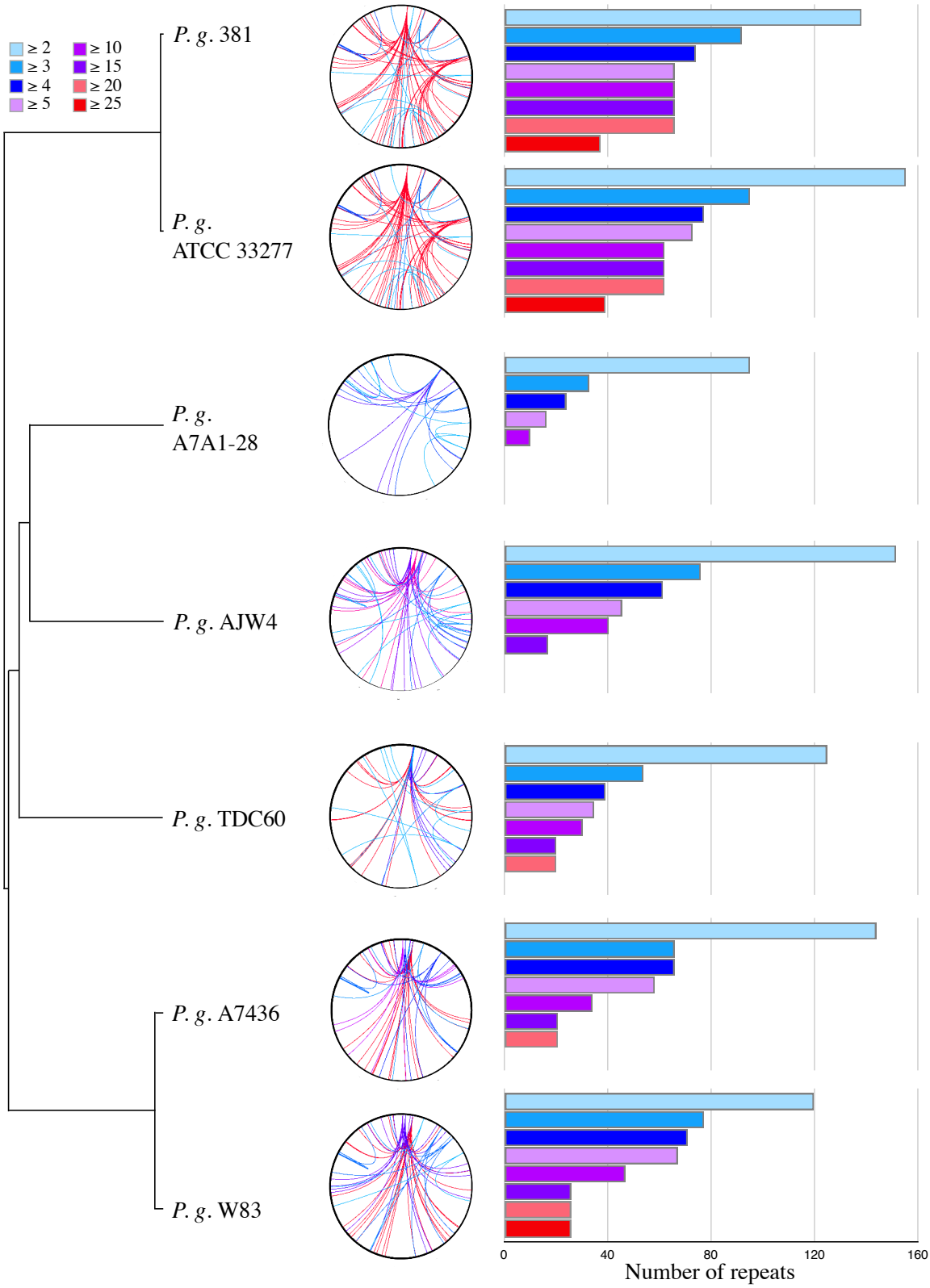


Figure 4. A *de novo* genome assembly of *Porphyromonas gingivalis* artificial reads.

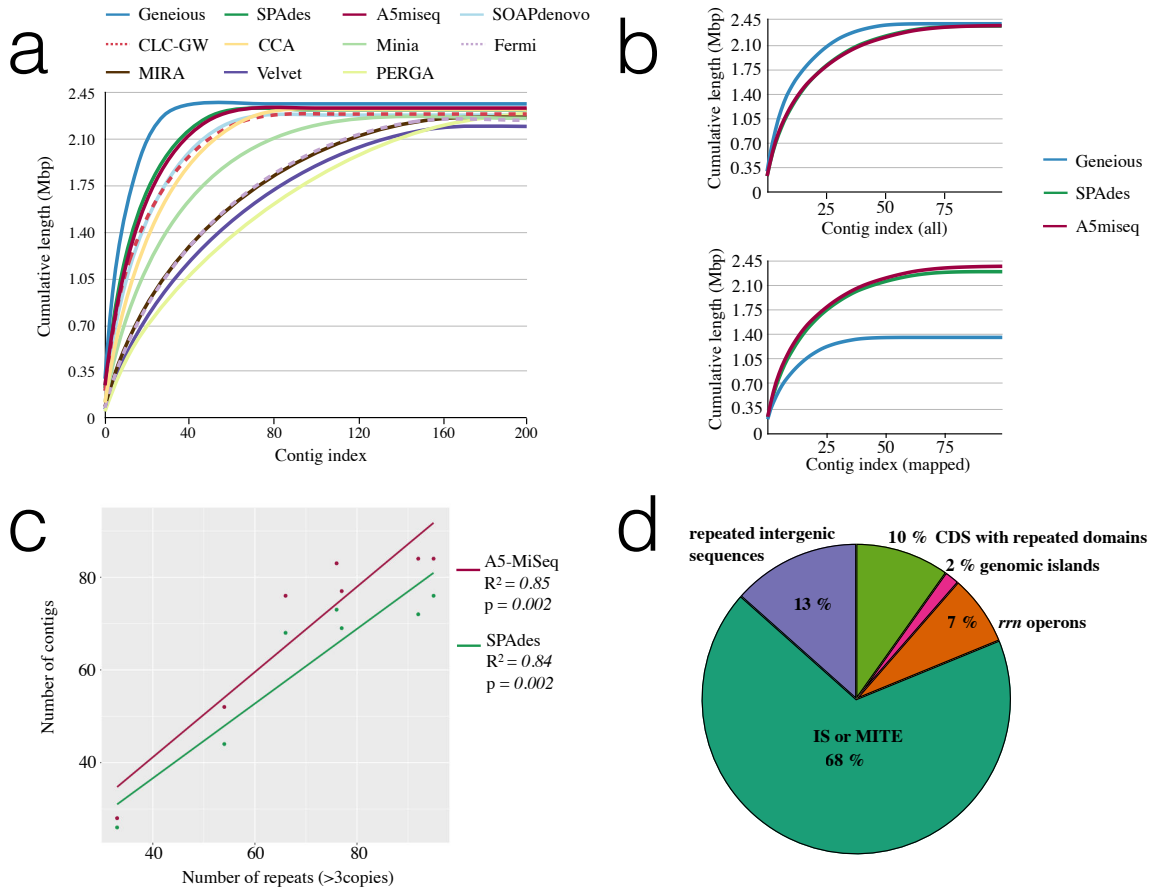


Figure 5. Functional comparison of the common coding sequences in two *Porphyromonas gingivalis* annotations.

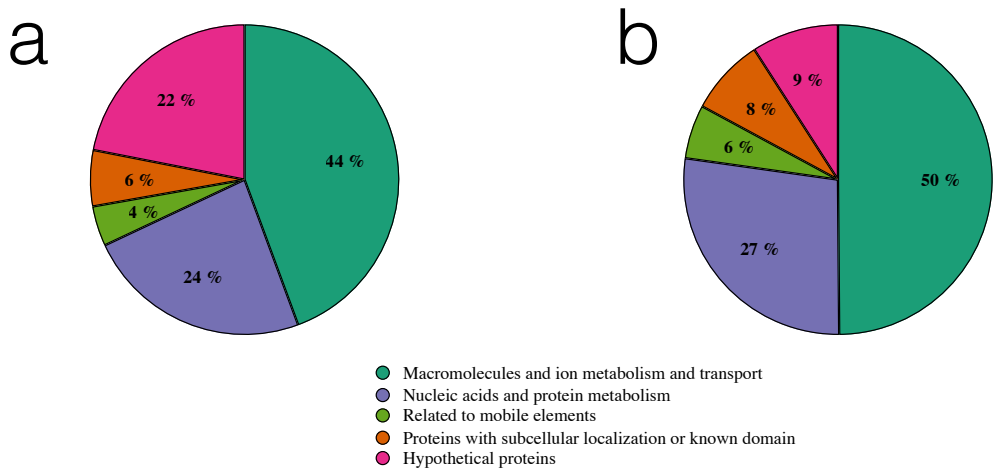
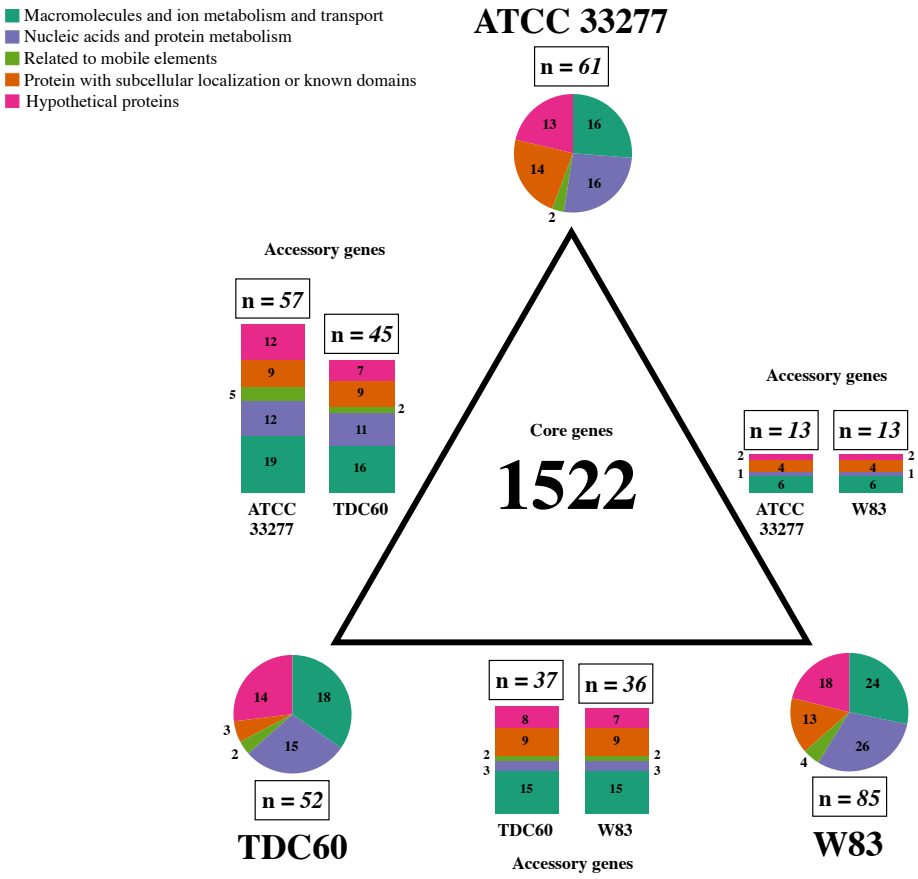
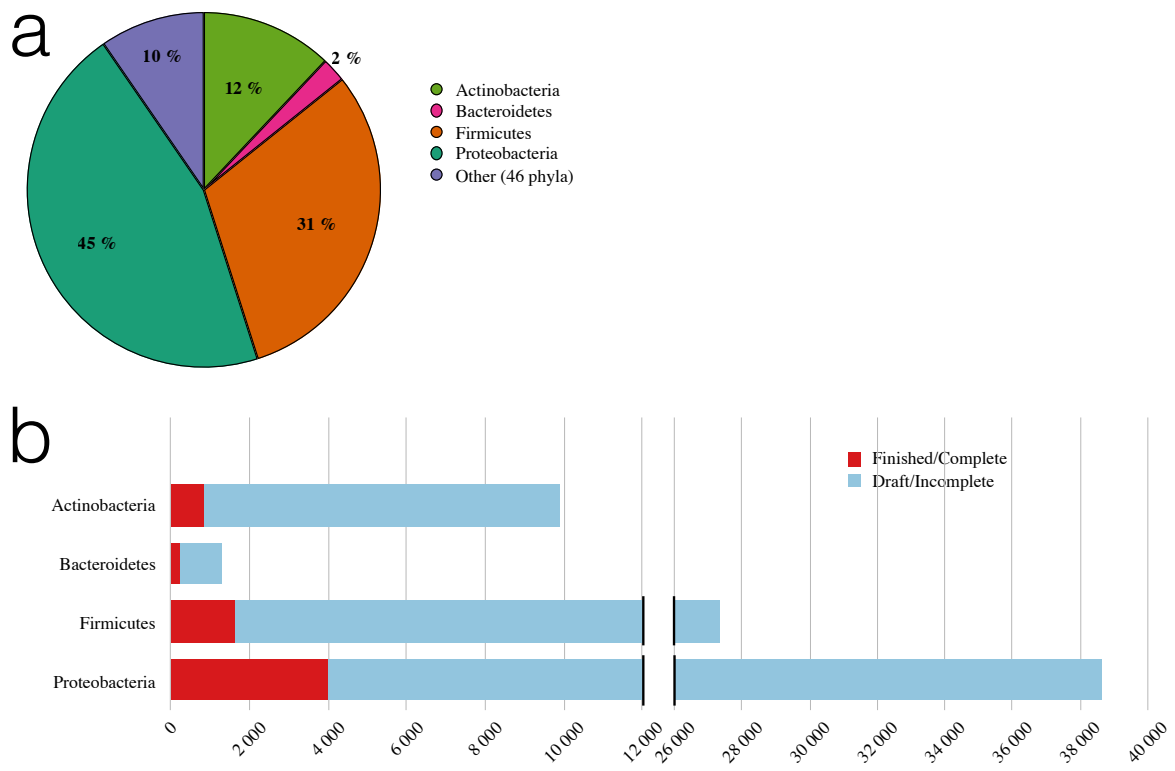


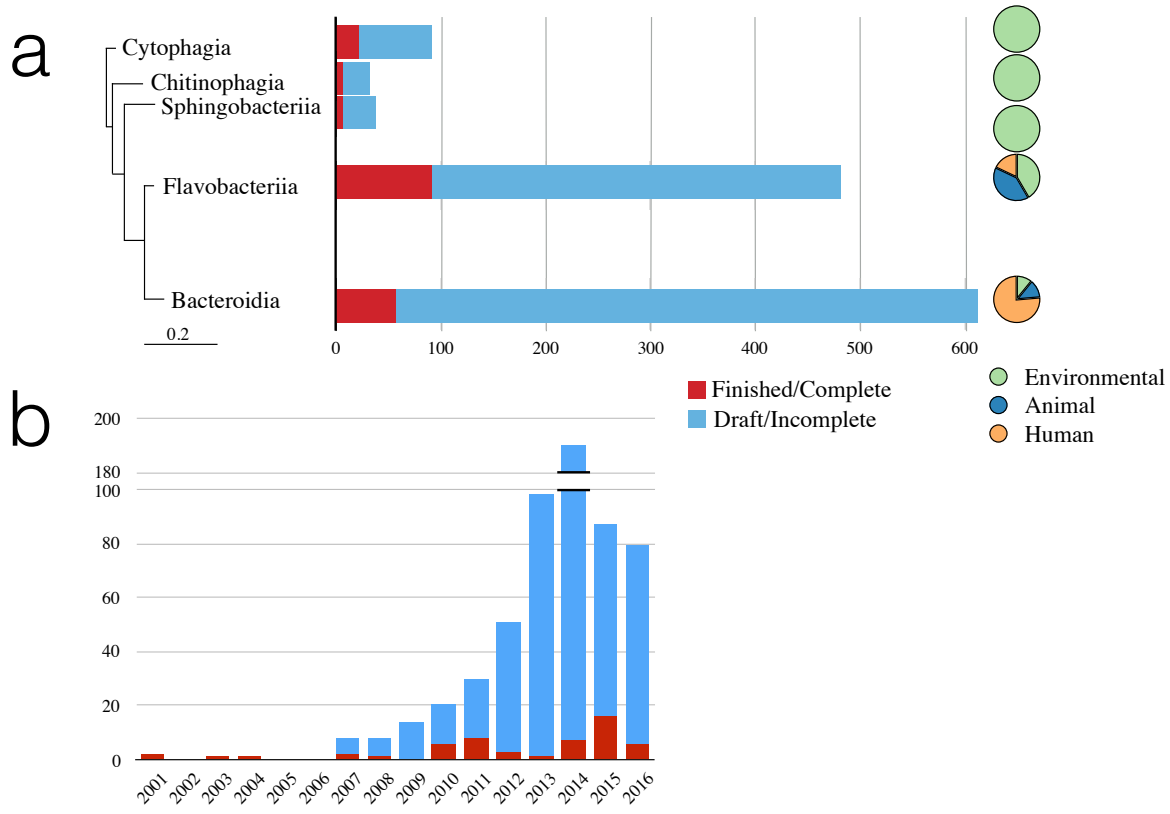
Figure 6. Pangenome overview of ATCC 33277, TDC60, and W83 strains, focusing on accessory and unique genomes.



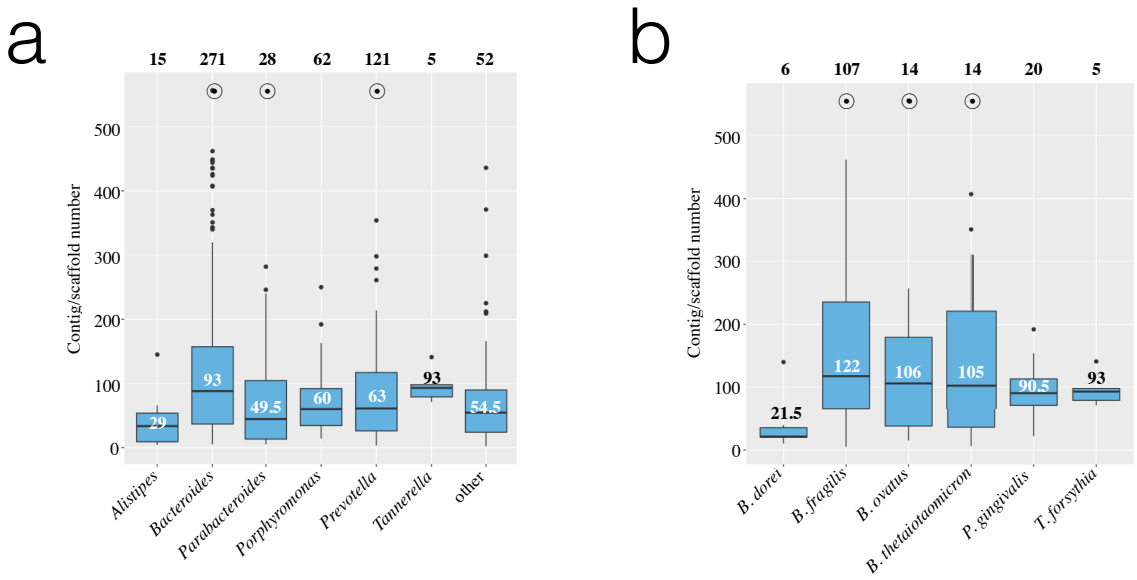
Additional file 1. Supplementary Figure 1. NCBI genome database distribution of the main bacterial phyla associated with humans.



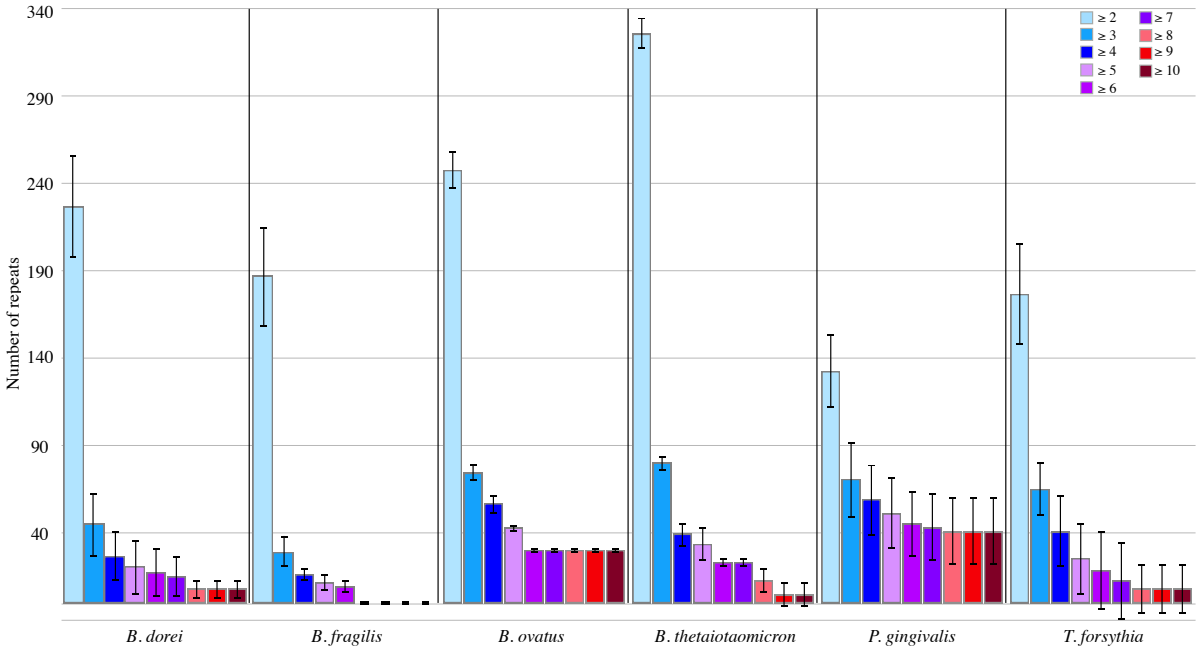
Additional file 2. Supplementary Figure 2. Bacteroidetes genomes by class.



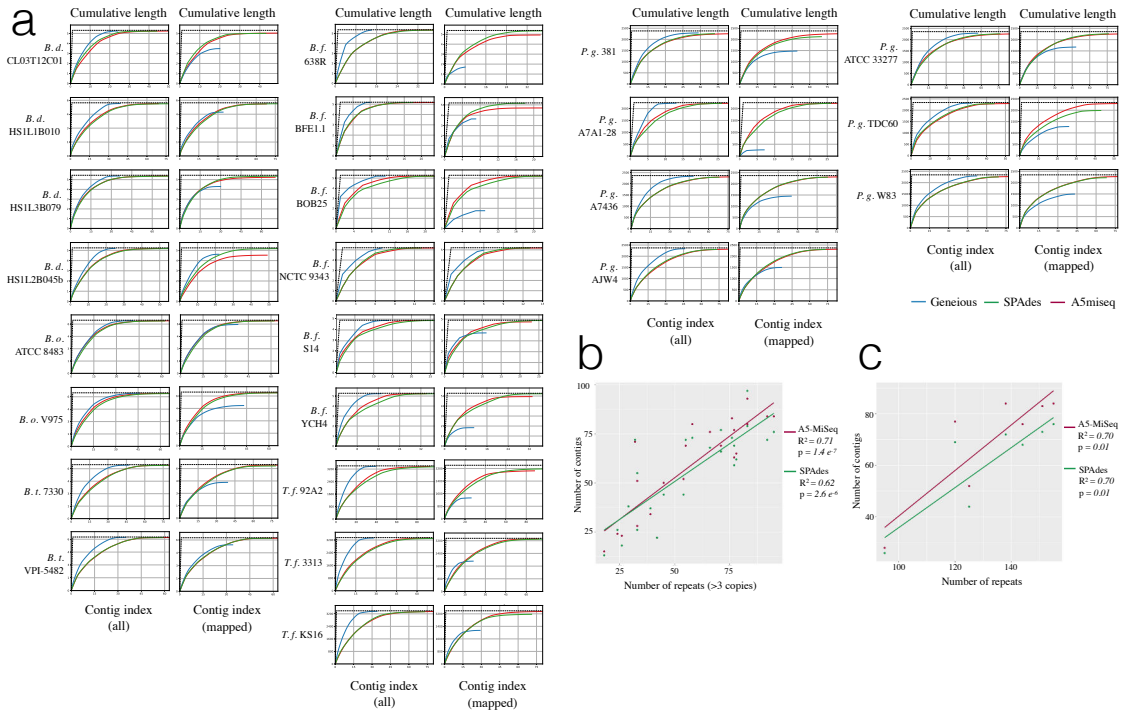
Additional file 3. Supplementary Figure 3. Bacteroidia draft genomes binned by genus and by species.



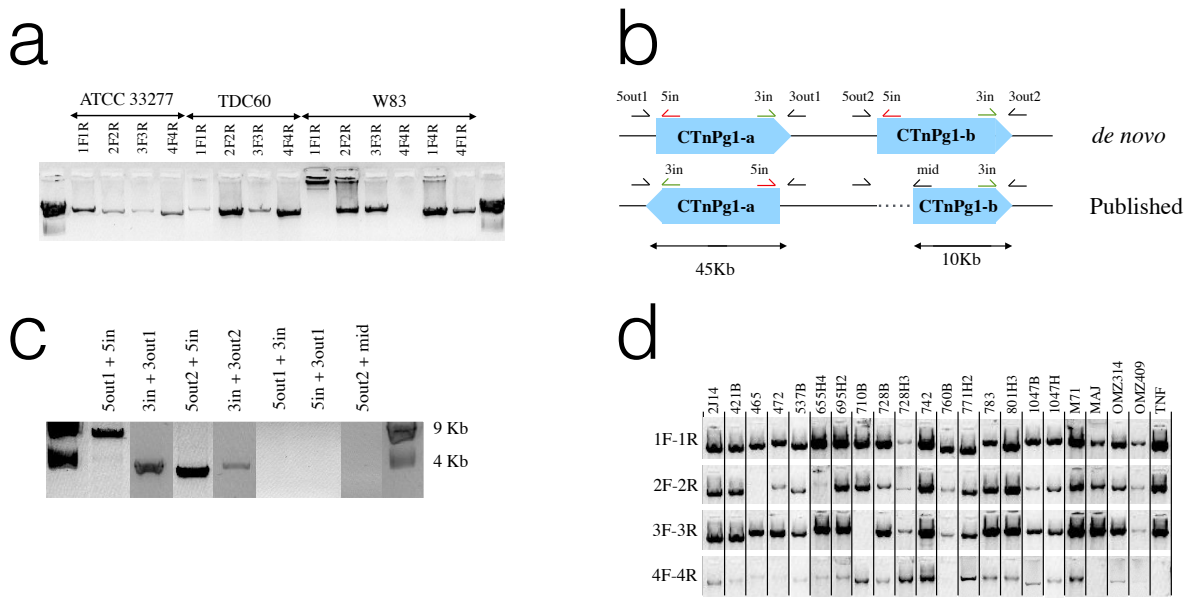
Additional file 4. Supplementary Figure 4. Genomic repeats by species.



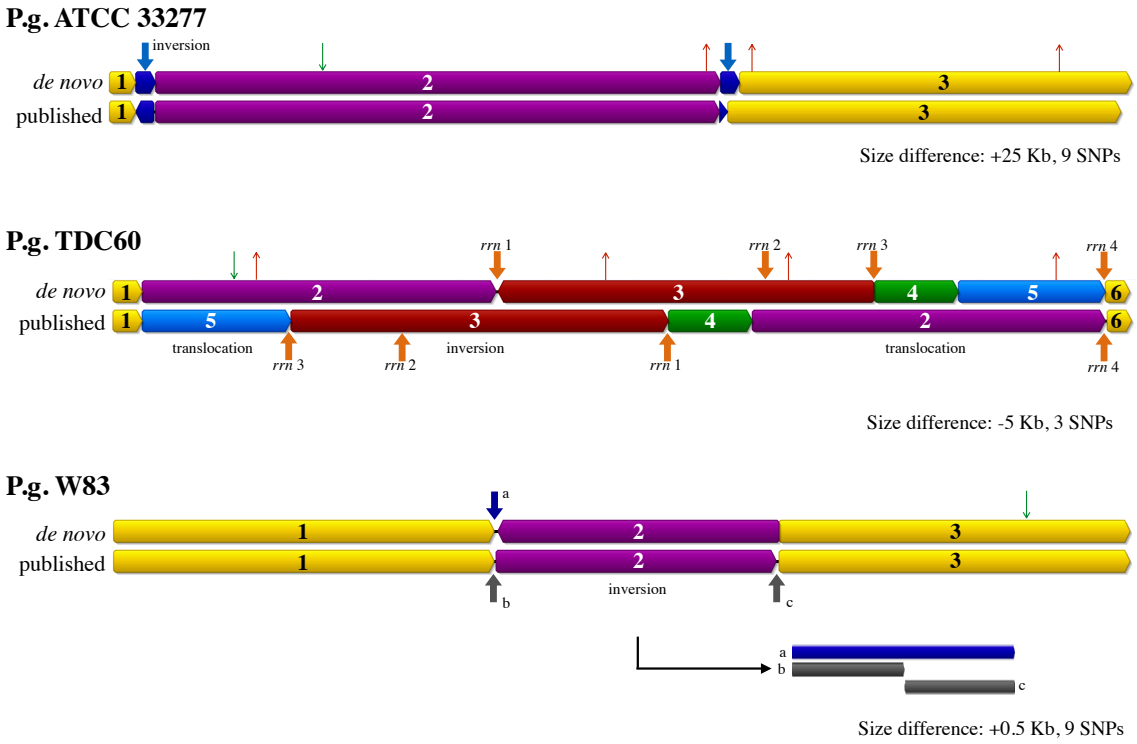
Additional file 5. Supplementary Figure 5. *De novo* assembly of artificial reads of the studied Bacteroidia genomes.



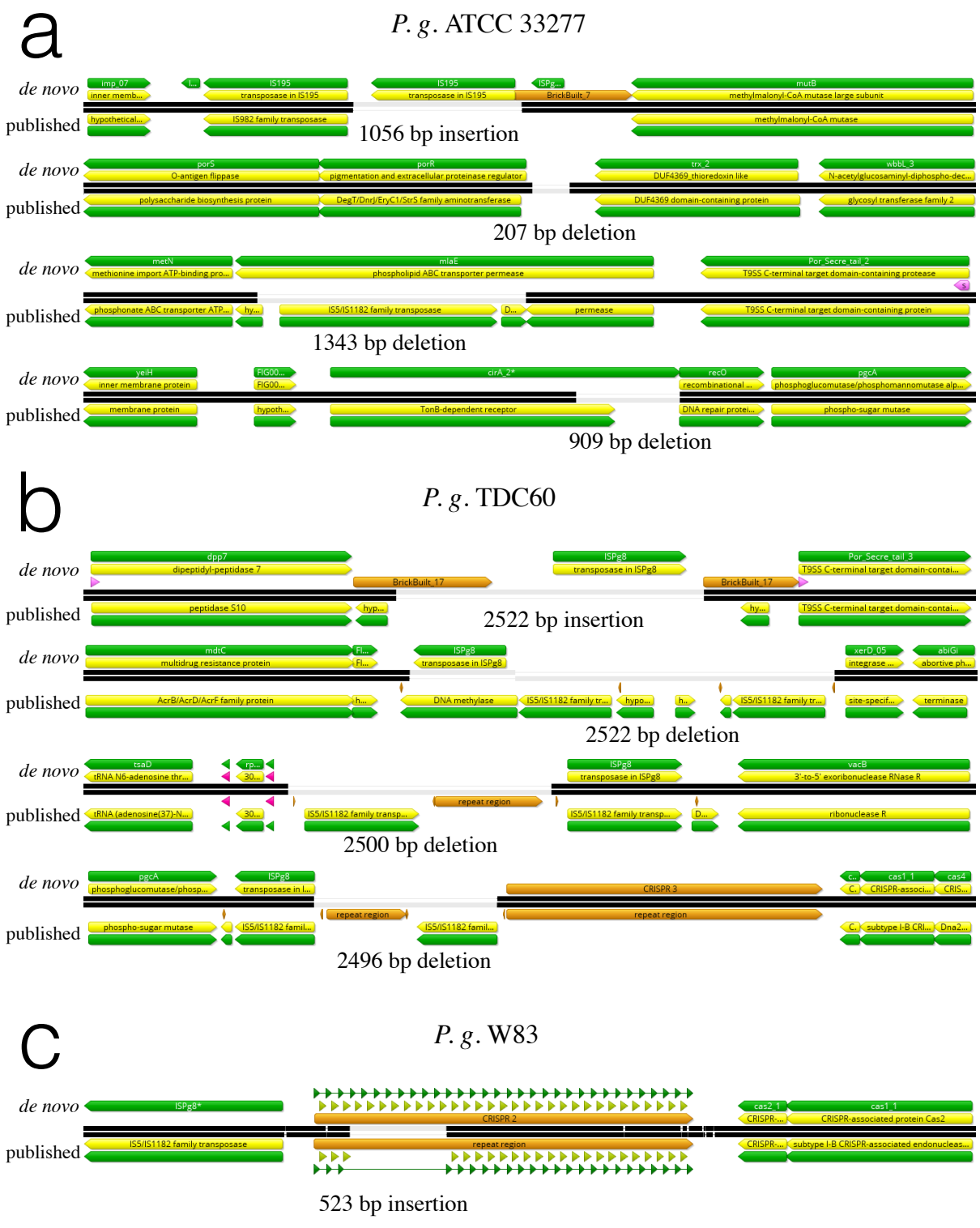
Additional file 6. Supplementary Figure 6. PCR validation of 3 P.g. strain constructions.



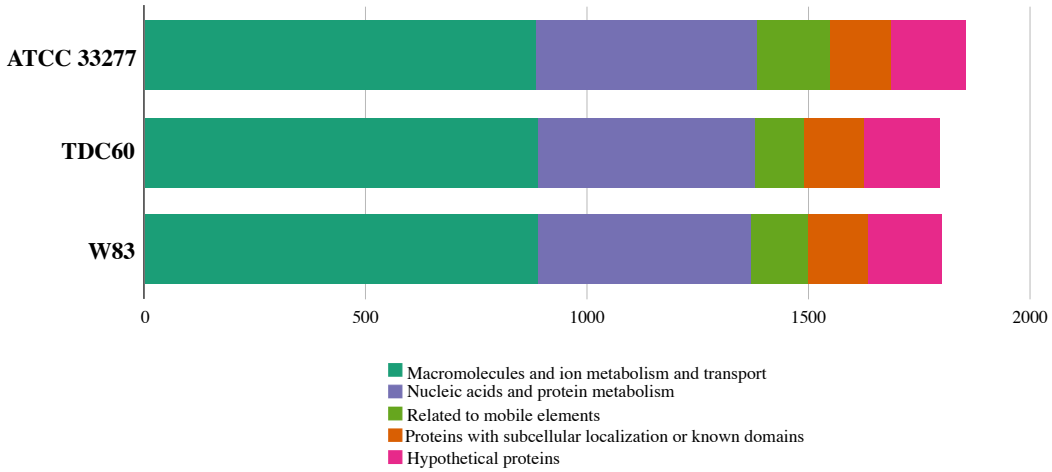
Additional file 7. Supplementary Figure 7. Whole-genome alignments of the three resequenced P.g. strains.



Additional file 8. Supplementary Figure 8. Insertions and deletions in the *de novo* P.g. assemblies as compared to the published genomes.

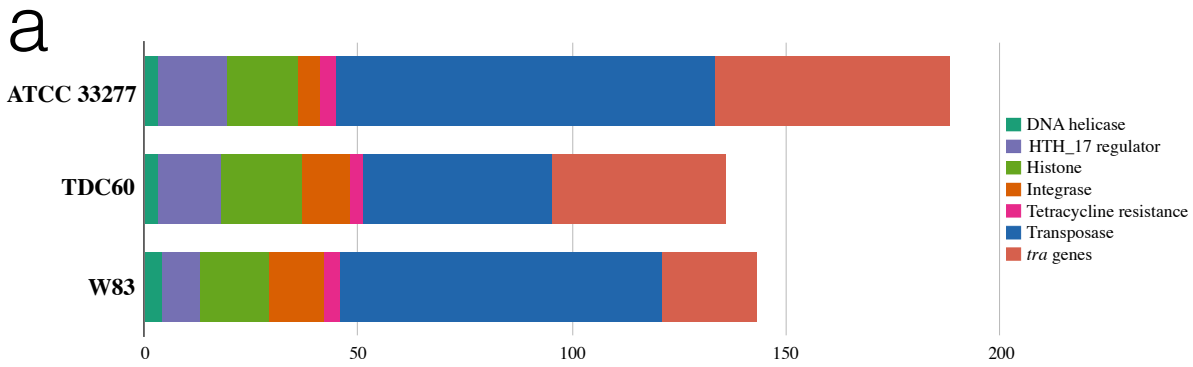


Additional file 9. Supplementary Figure 9. The three P.g. strains binned by their CDS/pseudogene functions.

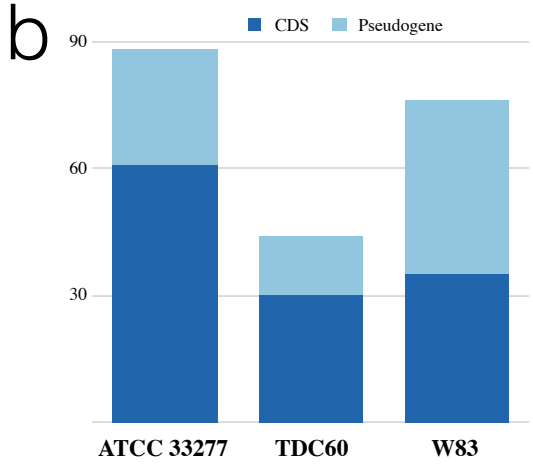


Additional file 10. Supplementary Figure 10. CDS/pseudogenes of all strains that have at least two copies in all three P.g. genomes.

a



b

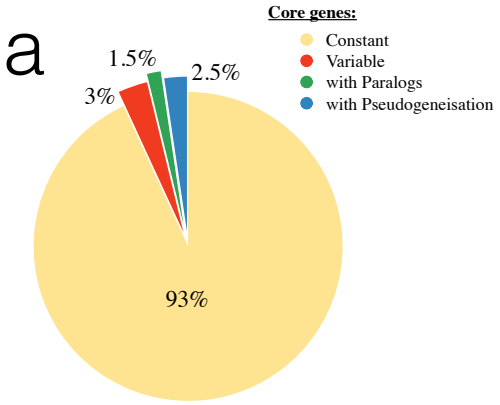


c

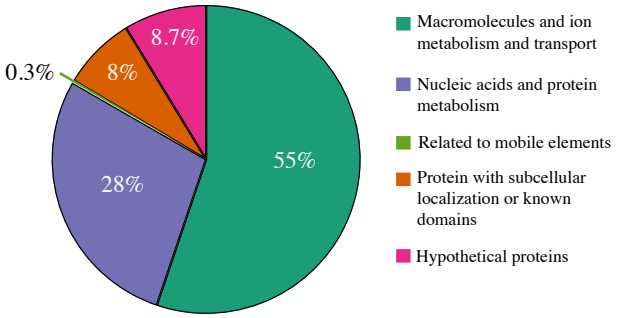
	ATCC 33277		TDC60		W83	
	CDS	pseudogene	CDS	pseudogene	CDS	pseudogene
ISPg8	32	10	16	5	11	22
ISPg2	3	4	1	2	5	3
IS195	24	4	10	1	5	3
ISPg4	0	1	0	1	10	1
ISPg5	0	5	0	0	1	11
ISPg6	0	3	1	2	1	0
ISPg7	0	0	1	0	1	0
IS256	0	0	0	2	0	1
IS4	0	0	1	1	1	0
ISL3	2	0	0	0	0	0

Additional file 11. Supplementary Figure 11. Overview of the core genomes of P.g. strains ATCC 33277, TDC60, and W83.

a



b



c

Function	Genes
CRISPR	*cas3*, *cas7*, *cmr2*, *cmr3*, *cmr5*, and *cmr6*
Fimbriae and pili	*fimA*, *fimC*, *fimD*, *fimE*, *mfa4;* and *ycf3_1*
Transport	*dtpT*, *ftsX_2*, *inlA*, *mepA_4*, *tonB_3*, and *wza*
Central metabolism	*capM*, *cysE*, *ghf*, *gppA*, *mfuc_2*, *mnmE*, *msuE*, *pheT_2*, *ragA*, *ragB*, *serB_2*, *topB3*, *wecG*, and *bmgA*
Type I restriction enzyme	*hsdR_1*
Haemagglutinin	*hagA*
Lysine gingipain protease	*kgp_4*
Unknown	*FIG00935998*, *FIG00936006*, *FIG00936766*, *FIG00936810*, *hypo_47*, *imp_35*, *imp_lipo_01*, *omp_lipo_1*, *omp_lipo_2*, *omp28*, and *yrrB_2*

d

gene	product	ATCC 33277	TDC60	W83
alkD	DNA alkylation repair enzyme	2	1	1
betI	transcriptional regulator	2	1	1
btr	transcriptional activator	2	1	2
ctpA	serine protease carboxy-terminal processing	2	1	1
DUF1896	hypothetical protein	2	2	1
epsJ	glycosyltransferase	2	2	1
era	GTPase	1 + 1*	1	1
FIG00935715	hypothetical protein	2	1	1
FIG00936125	hypothetical protein	2	1	1
hypo_01	hypothetical protein†	2	1	2
irtA	iron import ATP-binding/permease	2	1	1
ndvA	ABC transporter ATP-binding protein	2	1	1
rhuM	virulence protein RhuM family	2	1	2
ydfJ	exporter RND superfamily	2	1	1
ytxK	DNA methylase	2	2	1*

Additional file 12. Table S1. Complete genomes not included in this study.

Species	Strain	Sequencing Technology	Assembler	Coverage	Reference Pubmed ID	Release Date	Exclusion reason	FTP
Alistipes finegoldii	DSM 17242	454 + Illumina	Newbler	30	Unpublished (DOE-JGI)	2012	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/265/365/GCA_000265365.1_ASM26536v1
Alistipes shahii	WAL 8301	454	Unknown	22	Unpublished (SI)	2010	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/210/575/GCA_000210575.1_ASM21057v1
Bacteroides caccae	34	PacBio	HGAP + Celera	243	Unpublished (UoB)	2016	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/688/725/GCA_001688725.1_ASM168872v1
Bacteroides cellulosilyticus	WH2	PacBio + Illumina	HGAP + Celera	209	26430127	2015	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/318/345/GCA_001318345.1_ASM131834v1
Bacteroides coprosuis	DSM 18011	454 + Illumina	Newbler	30	21677890	2011	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/212/915/GCA_000212915.1_ASM21291v1
Bacteroides helcogenes	P 36-108	454 + Illumina	Newbler	30	21475586	2011	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/186/225/GCA_000186225.1_ASM18622v1
Bacteroides salanitronis	DSM 18170	454 + Illumina	Newbler	30	21677856	2011	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/190/575/GCA_000190575.1_ASM19057v1
Bacteroides vulgatus	ATCC 8482	Sanger	Phrap + PCAP	13	17576514	2007	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/012/825/GCA_000012825.1_ASM1282v1
Bacteroides vulgatus	mpk	PacBio	Celera	330	23071651	2015	Gap(s) in sequence	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/412/315/GCA_001412315.1_ASM141231v1
Bacteroides xylanisolvens	XB1A	454	Unknown	18	Unpublished (SI)	2010	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/210/075/GCA_000210075.1_ASM21007v1
Barnesiella viscericola	C46, DSM 18177	Unknown	ALLPATHS + Velvet + Phrap	Unknown	Unpublished (DOE-JGI)	2014	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/512/915/GCA_000512915.1_ASM51291v1
Draconibacterium orientale	FH5	Sanger + 454 + Illumina	Newbler	15	26796622	2014	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/962/635/GCA_000962635.1_ASM26263v1
Fermentimonas caenicola	ING2-E5B	Unknown	Unknown	Unknown	Unpublished (BU)	2014	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/953/535/GCA_000953535.1_E5B
Muricauda ruestringensis	M5T	PacBio + Illumina	HGAP	Unknown	25657285	2014	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/723/505/GCA_000723505.1_Muricauda_M5_Schon-Bonar-Fixed
Odoribacter splanchnicus	DSM 220712	454 + Illumina	Newbler	30	21677857	2011	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/190/535/GCA_000190535.1_ASM19053v1
Paludibacter propionicigenes	WB4	454 + Illumina	Newbler	30	21475585	2010	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/183/135/GCA_000183135.1_ASM18313v1
Parabacteroides distasonis	ATCC 8503	Sanger	Phrap + PCAP	13	17576514	2007	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/012/845/GCA_000012845.1_ASM1284v1
Petrimonas sulfuriphila	ING2-E5A	Unknown	Unknown	Unknown	Unpublished (LIAEPB)	2016	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/095/795/GCA_900095795.1_E5A
Porphyromonas asaccharolytica	DSM 20707	454 + Illumina	Newbler	30	Unpublished (DOE-JGI)	2011	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/212/375/GCA_000212375.1_ASM21237v1
Porphyromonas gingivalis	HG66	PacBio	SMRT Analysis	198	25291768	2014	Gap(s) in sequence	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/739/415/GCA_000739415.1_ASM73941v1
Porphyromonas gingivalis	JCVI SC001	Illumina	SPAdes	237	23564253	2013	Gap(s) in sequence	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/380/305/GCA_000380305.1_PgingivalisJCVISC001v1.0
Prevotella denticola	DSM 3688	454 + Illumina	Newbler	30	Unpublished (BCoM)	2012	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/242/335/GCA_000242335.1_ASM24233v3
Prevotella dentricola	F0289	454 + Illumina	Celera	43	Unpublished (JCVI)	2011	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/193/395/GCA_000193395.1_ASM19339v1
Prevotella enoeca	F0113	PacBio	HGAP	366	Unpublished (BCoM)	2015	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/444/445/GCA_001444445.1_ASM144444v1
Prevotella fusca	W1435	PacBio	HGAP	117	Unpublished (BCoM)	2015	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/262/015/GCA_001262015.1_ASM126201v1
Prevotella intermedia	17	Sanger	Celera	7	Unpublished (JCVI)	2012	More than 1 chromosome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/261/025/GCA_000261025.1_ASM26102v1
Prevotella intermedia	17-2	PacBio	HGAP	200	26294638	2015	More than 1 chromosome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/548/195/GCA_001548195.1_ASM154819v1
Prevotella intermedia	ATCC 25611	PacBio	HGAP + Celera	200	Unpublished (FI)	2017	More than 1 chromosome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/953/955/GCA_001953955.1_ASM195395v1
Prevotella intermedia	strain 17	PacBio	HGAP + Celera	200	Unpublished (FI)	2017	More than 1 chromosome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/953/935/GCA_001953935.1_ASM195393v1
Prevotella melaninogenica	ATCC 25845	454	Celera	14	Unpublished (HMP)	2010	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/144/805/GCA_000144805.1_ASM14480v1
Prevotella ruminicola	Bryant 23	Sanger	Unknown	8	20585943	2010	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/025/925/GCA_000025925.1_ASM2592v1
Prevotella scopos	W2052	Illumina	SPAdes	26	Unpublished (BCoM)	2016	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/683/355/GCA_001683355.1_ASM168335v1
Proteiniphilum saccharofermentans	M3/6	Unknown	Unknown	Unknown	Unpublished (BU)	2017	Only 1 complete genome	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/095/135/GCA_900095135.1_M3/6

Additional file 13. Table S2. The 166 draft genomes of the six Bacteroidia species studied here (1 of 4).

Species	Strain	Sequencing Technology	Assembler	Contigs	Reference Pubmed ID	Release Date	FTP
Bacteroides dorei	CAG:222	Unknown	Unknown	140	Unpublished (TUD)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/436/135/GCA_000436135.1_MGS222
Bacteroides dorei	5_1_36/D4	454	Newbler	10	Unpublished (HMP)	2009	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/158/335/GCA_000158335.2_Bact_dorei_5_1_36_D4_V2
Bacteroides dorei	CL02T00C15	Illumina	allpaths	22	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/035/GCA_000273035.1_Bact_dore_CL02T00C15_V1
Bacteroides dorei	CL02T12C06	Illumina	allpaths	21	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/055/GCA_000273055.1_Bact_dore_CL02T12C06_V1
Bacteroides dorei	CL03T12C01	Illumina	allpaths	20	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/075/GCA_000273075.1_Bact_dore_CL03T12C01_V1
Bacteroides dorei	DSM 17855	454	Newbler	49	Unpublished (BCcM)	2008	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/156/075/GCA_000156075.1_ASM15607v1
Bacteroides fragilis	CAG:47	Unknown	Unknown	118	Unpublished (TUD)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/434/095/GCA_000434095.1_MGS47
Bacteroides fragilis	CAG:558	Unknown	Unknown	175	Unpublished (TUD)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/432/495/GCA_000432495.1_MGS558
Bacteroides fragilis	1007-1-F #10	Illumina	MaSuRCA	83	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/685/GCA_000598685.2_ASM59868v2
Bacteroides fragilis	1007-1-F #3	Illumina	MaSuRCA	106	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/265/GCA_000599265.1_ASM59926v1
Bacteroides fragilis	1007-1-F #4	Illumina	MaSuRCA	167	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/545/GCA_000598545.2_ASM59854v2
Bacteroides fragilis	1007-1-F #5	Illumina	MaSuRCA	157	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/601/035/GCA_000601035.1_ASM60103v1
Bacteroides fragilis	1007-1-F #6	Illumina	MaSuRCA	87	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/601/095/GCA_000601095.1_ASM60109v1
Bacteroides fragilis	1007-1-F #7	Illumina	MaSuRCA	130	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/145/GCA_000599145.2_ASM59914v2
Bacteroides fragilis	1007-1-F #8	Illumina	MaSuRCA	315	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/265/GCA_000598265.1_ASM59826v1
Bacteroides fragilis	1007-1-F #9	Illumina	MaSuRCA	66	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/885/GCA_000598885.1_ASM59888v1
Bacteroides fragilis	1009-4-F #10	Illumina	MaSuRCA	63	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/705/GCA_000598705.1_ASM59870v1
Bacteroides fragilis	1009-4-F #7	Illumina	MaSuRCA	46	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/285/GCA_000599285.2_ASM59928v2
Bacteroides fragilis	14-106904-1	Illumina	SPAdes	73	Poster (Sydenham et al. 2015)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/816/225/GCA_001816225.1_ASM181622v1
Bacteroides fragilis	2-078382-3	Illumina	SPAdes	140	27348220	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/699/885/GCA_001699885.1_ASM169988v1
Bacteroides fragilis	2-F-2 #4	Illumina	MaSuRCA	213	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/825/GCA_000598825.1_ASM59882v1
Bacteroides fragilis	2-F-2 #5	Illumina	MaSuRCA	250	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/285/GCA_000598285.1_ASM59828v1
Bacteroides fragilis	2-F-2 #7	Illumina	MaSuRCA	363	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/145/GCA_000598145.1_ASM59814v1
Bacteroides fragilis	20656-2-1	Illumina	CLC-WB	68	27348220	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/699/875/GCA_001699875.1_ASM169987v1
Bacteroides fragilis	20793-3	Illumina	MaSuRCA	148	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/905/GCA_000598905.1_ASM59890v1
Bacteroides fragilis	20793-3	Unknown	Unknown	52	Unpublished*	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/699/855/GCA_001699855.1_ASM169985v1
Bacteroides fragilis	2d2A	Illumina	Unknown	1025	25717097	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/944/095/GCA_000944095.1_2d2A_assembly
Bacteroides fragilis	3-F-2 #6	Illumina	MaSuRCA	201	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/865/GCA_000598865.1_ASM59886v1
Bacteroides fragilis	320_BFRA	Illumina	Abyss	150	26230489	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/054/865/GCA_001054865.1_ASM105486v1
Bacteroides fragilis	321_BFRA	Illumina	Abyss	150	26230489	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/056/335/GCA_001056335.1_ASM105633v1
Bacteroides fragilis	322_BFRA	Illumina	Abyss	193	26230489	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/054/895/GCA_001054895.1_ASM105489v1
Bacteroides fragilis	3397 N2	Illumina	MaSuRCA	93	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/565/GCA_000598565.1_ASM59856v1
Bacteroides fragilis	3397 N3	Illumina	MaSuRCA	102	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/925/GCA_000598925.1_ASM59892v1
Bacteroides fragilis	3397 T10	Illumina	MaSuRCA	2566	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/405/GCA_000598405.1_ASM59840v1
Bacteroides fragilis	3397 T14	Illumina	MaSuRCA	94	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/165/GCA_000599165.2_ASM59916v2
Bacteroides fragilis	34-F-2 #13	Illumina	MaSuRCA	426	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/425/GCA_000598425.1_ASM59842v1
Bacteroides fragilis	3719 A10	Illumina	MaSuRCA	117	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/845/GCA_000598845.1_ASM59884v1
Bacteroides fragilis	3719 T6	Illumina	MaSuRCA	64	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/725/GCA_000598725.1_ASM59872v1
Bacteroides fragilis	3725 D9 ii	Illumina	MaSuRCA	91	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/699/685/GCA_000699685.1_ASM69968v1
Bacteroides fragilis	3725 D9(v)	Illumina	MaSuRCA	75	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/585/GCA_000598585.1_ASM59858v1
Bacteroides fragilis	3774 T13	Illumina	MaSuRCA	340	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/305/GCA_000598305.1_ASM59830v1
Bacteroides fragilis	3783N1-2	Illumina	MaSuRCA	449	Unpublished (JGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/325/GCA_000598325.1_ASM59832v1

1

Additional file 13. Table S2. Continuation (2 of 4).

Species	Strain	Sequencing Technology	Assembler	Contigs	Reference Pubmed ID	Release Date	FTP
Bacteroides fragilis	3783N1-6	Illumina	MaSuRCA	57	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/590/065/GCA_000590065.2_ASM59006v2
Bacteroides fragilis	3783N1-8	Illumina	MaSuRCA	85	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/605/GCA_000598605.1_ASM59860v1
Bacteroides fragilis	3783N2-1	Illumina	MaSuRCA	436	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/345/GCA_000598345.1_ASM59834v1
Bacteroides fragilis	3976T7	Illumina	MaSuRCA	447	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/165/GCA_000598165.1_ASM59816v1
Bacteroides fragilis	3976T8	Illumina	MaSuRCA	73	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/185/GCA_000599185.2_ASM59918v2
Bacteroides fragilis	3986 N(B)19	Illumina	MaSuRCA	792	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/445/GCA_000598445.1_ASM59844v1
Bacteroides fragilis	3986 N(B)22	Illumina	MaSuRCA	31	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/945/GCA_000598945.1_ASM59894v1
Bacteroides fragilis	3986 N3	Illumina	MaSuRCA	36	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/601/115/GCA_000601115.1_ASM60111v1
Bacteroides fragilis	3986 T(B)13	Illumina	MaSuRCA	44	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/965/GCA_000598965.1_ASM59896v1
Bacteroides fragilis	3986 T(B)9	Illumina	MaSuRCA	211	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/465/GCA_000598465.1_ASM59846v1
Bacteroides fragilis	3986T(B)10	Illumina	MaSuRCA	313	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/185/GCA_000598185.2_ASM59818v2
Bacteroides fragilis	3988 T1	Illumina	MaSuRCA	462	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/205/GCA_000598205.1_ASM59820v1
Bacteroides fragilis	3988T(B)14	Illumina	MaSuRCA	408	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/365/GCA_000598365.1_ASM59836v1
Bacteroides fragilis	3996 N(B) 6	Illumina	MaSuRCA	298	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/225/GCA_000598225.1_ASM59822v1
Bacteroides fragilis	3998 T(B)1-4	Illumina	MaSuRCA	424	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/385/GCA_000598385.1_ASM59838v1
Bacteroides fragilis	3998T(B)3	Illumina	MaSuRCA	444	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/485/GCA_000598485.1_ASM59848v1
Bacteroides fragilis	3_1_12	454	Unknown	33	Unpublished (HMP)	2009	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/157/015/GCA_000157015.1_ASM15701v1
Bacteroides fragilis	4g8B	Illumina	Unknown	1402	25717097	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/373/095/GCA_001373095.1_4g8B_assembly
Bacteroides fragilis	86-5443-2-2	Illumina	CLC-WB	156	27348220	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/699/885/GCA_001699885.1_ASM169988v1
Bacteroides fragilis	885_B5RA	Illumina	Abyss	435	26230489	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/058/755/GCA_001058755.1_ASM105875v1
Bacteroides fragilis	894_B5RA	Illumina	Abyss	370	26230489	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/058/775/GCA_001058775.1_ASM105877v1
Bacteroides fragilis	915_B5RA	Illumina	Abyss	1290	26230489	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/077/245/GCA_001077245.1_ASM107724v1
Bacteroides fragilis	A7 (UDC12-2)	Illumina	MaSuRCA	87	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/985/GCA_000598985.1_ASM59898v1
Bacteroides fragilis	ATCC 25285	Illumina	CLC-WB	58	28592550	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/997/325/GCA_001997325.1_ASM199732v1
Bacteroides fragilis	B1 (UDC16-1)	Illumina	MaSuRCA	311	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/625/GCA_000598625.1_ASM59862v1
Bacteroides fragilis	B9	Illumina	CLC-WB	5	27246231	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/695/355/GCA_001695355.1_ASM169535v1
Bacteroides fragilis	CL03T00C08	Illumina	allpaths	8	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/095/GCA_000273095.1_Bact_frag_CL03T00C08_V1
Bacteroides fragilis	CL03T12C07	Illumina	allpaths	7	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/115/GCA_000273115.1_Bact_frag_CL03T12C07_V1
Bacteroides fragilis	CL05T00C42	Illumina	allpaths	12	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/269/525/GCA_000269525.1_PB_Bact_frag_CL05T00C42_V1
Bacteroides fragilis	CL05T00C42	Illumina	allpaths	5	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/765/GCA_000273765.1_Bact_frag_CL05T00C42_V1
Bacteroides fragilis	CL05T12C13	Illumina	allpaths	13	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/135/GCA_000273135.1_Bact_frag_CL05T12C13_V1
Bacteroides fragilis	CL07T00C01	Illumina	allpaths	9	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/263/115/GCA_000263115.1_Bact_frag_CL07T00C01_V1
Bacteroides fragilis	CL07T12C05	Illumina	allpaths	11	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/155/GCA_000273155.1_Bact_frag_CL07T12C05_V1
Bacteroides fragilis	DCMOUB0017B	Illumina	SPAdes	232	Poster (Sydenham et al. 2015)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/710/375/GCA_000710375.2_DCMOUB0017B2.0
Bacteroides fragilis	DCMOUB0018B	Illumina	SPAdes	268	Poster (Sydenham et al. 2015)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/724/665/GCA_000724665.2_DCMOUB0018B2.0
Bacteroides fragilis	DCMOUB0042B	Illumina	SPAdes	197	Poster (Sydenham et al. 2015)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/724/795/GCA_000724795.1_DCMOUB0042B1.0
Bacteroides fragilis	DCMOUB0067B	Illumina	SPAdes	343	Poster (Sydenham et al. 2015)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/724/805/GCA_000724805.1_DCMOUB0067B_1.0
Bacteroides fragilis	DCMOUB0085B	Illumina	SPAdes	157	Poster (Sydenham et al. 2015)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/724/815/GCA_000724815.1_DCMOUB0085B_1.0
Bacteroides fragilis	DCMSKEBY00018	Illumina	SPAdes	213	Poster (Sydenham et al. 2015)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/710/365/GCA_000710365.2_DCMSKEBY00018B2.0
Bacteroides fragilis	DS-166	Illumina	MaSuRCA	124	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/245/GCA_000598245.1_ASM59824v1
Bacteroides fragilis	DS-208	Illumina	MaSuRCA	271	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/505/GCA_000598505.1_ASM59850v1
Bacteroides fragilis	DS-71	Illumina	MaSuRCA	308	Unpublished (IGS)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/085/GCA_000599085.1_ASM59908v1

2

Additional file 13. Table S2. Continuation (3 of 4).

Species	Strain	Sequencing Technology	Assembler	Contigs	Reference Pubmed ID	Release Date	FTP
Bacteroides fragilis	Ds-233	Illumina	MaSuRCA	557	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/805/GCA_000598805.1_ASM59880v1
Bacteroides fragilis	HMW 610	Illumina	allpaths	9	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/297/695/GCA_000297695.1_Bact_frag_HMW_610_V1
Bacteroides fragilis	HMW 615	Illumina	allpaths	14	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/297/735/GCA_000297735.1_Bact_frag_HMW_615_V1
Bacteroides fragilis	HMW 616	Illumina	allpaths	9	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/297/755/GCA_000297755.1_Bact_frag_HMW_616_V1
Bacteroides fragilis	II 545	Illumina	MaSuRCA	76	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/785/GCA_000598785.2_ASM59878v2
Bacteroides fragilis	J-143-4	Illumina	MaSuRCA	270	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/525/GCA_000598525.1_ASM59852v1
Bacteroides fragilis	J38-1	Illumina	MaSuRCA	120	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/645/GCA_000598645.1_ASM59864v1
Bacteroides fragilis	JCM 11017	IonTorrent	Newbler	98	19934255	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/613/425/GCA_000613425.1_ASM61342v1
Bacteroides fragilis	JIM10	454 + IonTorrent	Newbler	81	Unpublished (FMBA)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/692/695/GCA_001692695.1_ASM169269v1
Bacteroides fragilis	KLE1758	Illumina	Velvet	52	Unpublished (WU)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/580/095/GCA_001580095.1_ASM158009v1
Bacteroides fragilis	Korea 419	Illumina	MaSuRCA	246	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/205/GCA_000599205.1_ASM59920v1
Bacteroides fragilis	O:21	Illumina	CLC-WB	14	27246231	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/693/695/GCA_001693695.1_ASM169369v1
Bacteroides fragilis	S13 L11	Illumina	MaSuRCA	790	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/105/GCA_000599105.1_ASM59910v1
Bacteroides fragilis	S23 R14	Illumina	MaSuRCA	263	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/865/GCA_000598865.1_ASM59886v1
Bacteroides fragilis	S23L17	Illumina	MaSuRCA	133	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/601/055/GCA_000601055.1_ASM60105v1
Bacteroides fragilis	S23L24	Illumina	MaSuRCA	122	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/305/GCA_000599305.1_ASM59930v1
Bacteroides fragilis	S24L15	Illumina	MaSuRCA	147	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/005/GCA_000599005.1_ASM59900v1
Bacteroides fragilis	S24L26	Illumina	MaSuRCA	72	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/745/GCA_000598745.1_ASM59874v1
Bacteroides fragilis	S24L34	Illumina	MaSuRCA	65	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/325/GCA_000599325.1_ASM59932v1
Bacteroides fragilis	S38L11	Illumina	MaSuRCA	294	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/125/GCA_000599125.1_ASM59912v1
Bacteroides fragilis	S38L12	Illumina	MaSuRCA	90	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/345/GCA_000599345.1_ASM59934v1
Bacteroides fragilis	S38L5	Illumina	MaSuRCA	105	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/025/GCA_000599025.1_ASM59902v1
Bacteroides fragilis	S38L3	Illumina	MaSuRCA	49	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/598/765/GCA_000598765.2_ASM59876v2
Bacteroides fragilis	S38L5	Illumina	MaSuRCA	83	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/365/GCA_000599365.1_ASM59936v1
Bacteroides fragilis	S6L3	Illumina	MaSuRCA	119	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/225/GCA_000599225.1_ASM59922v1
Bacteroides fragilis	S6L8	Illumina	MaSuRCA	100	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/385/GCA_000599385.1_ASM59938v1
Bacteroides fragilis	S6R5	Illumina	MaSuRCA	80	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/045/GCA_000599045.1_ASM59904v1
Bacteroides fragilis	S6R6	Illumina	MaSuRCA	84	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/599/245/GCA_000599245.1_ASM59924v1
Bacteroides fragilis	S6R8	Illumina	MaSuRCA	133	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/601/075/GCA_000601075.2_ASM60107v2
Bacteroides ovatus	CAG:22	Unknown	Unknown	257	Unpublished (TUD)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/436/435/GCA_000436435.1_MGS22
Bacteroides ovatus	2789STDY5834943	Illumina	Unknown	57	27144353	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/405/735/GCA_001405735.1_14207_7_66
Bacteroides ovatus	3725 D1 iv	Illumina	MaSuRCA	181	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/699/725/GCA_000699725.1_ASM69972v1
Bacteroides ovatus	3725 D9 iii	Illumina	MaSuRCA	556	Unpublished (JGI)	2014	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/699/665/GCA_000699665.1_ASM69966v1
Bacteroides ovatus	3_8_47FAA	454	Newbler	25	Unpublished (HMP)	2011	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/218/325/GCA_000218325.1_Bact_ovat_3_8_47FAA_V1
Bacteroides ovatus	ATCC 8483	454 + Sanger	Newbler + PCAP	32	Unpublished (HMP)	2007	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/154/125/GCA_000154125.1_ASM15412v1
Bacteroides ovatus	CL02T12C04	Illumina	allpaths	15	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/195/GCA_000273195.1_Bact_ovat_CL02T12C04_V1
Bacteroides ovatus	CL03T12C18	Illumina	allpaths	19	Unpublished (HMP)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/273/215/GCA_000273215.1_Bact_ovat_CL03T12C18_V1
Bacteroides ovatus	CL09T03C03	Illumina	Velvet	193	26768961	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/535/615/GCA_001535615.1_ASM153561v1
Bacteroides ovatus	DSM 1896	Unknown	Unknown	69	Unpublished (DOE-JGI)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/107/475/GCA_900107475.1_IMG-taxon_2693426857_annotated_assembly
Bacteroides ovatus	KLE1656	Illumina	Velvet	87	Unpublished (WU)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/578/575/GCA_001578575.1_ASM157857v1
Bacteroides ovatus	NLAE-zl-C500	Unknown	Unknown	125	Unpublished (DOE-JGI)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/102/645/GCA_900102645.1_IMG-taxon_2654588143_annotated_assembly
Bacteroides ovatus	NLAE-zl-C57	Unknown	Unknown	174	Unpublished (DOE-JGI)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/100/465/GCA_900100465.1_IMG-taxon_2654588146_annotated_assembly

3

Additional file 13. Table S2. Continuation (4 of 4).

Species	Strain	Sequencing Technology	Assembler	Contigs	Reference Pubmed ID	Release Date	FTP
Bacteroides ovatus	SD CMC 3f	454	Newbler	156	Unpublished (JCVI)	2010	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/178/275/GCA_000178275.1_ASM17827v1
Bacteroides thetaiotaomicron	CAG:40	Unknown	Unknown	142	Unpublished (TUD)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/437/995/GCA_000437995.1_MGS40
Bacteroides thetaiotaomicron	3731	Unknown	Unknown	175	Unpublished (WU)	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/049/535/GCA_001049535.1_3731
Bacteroides thetaiotaomicron	7330	Unknown	Unknown	407	Unpublished (WU)	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/049/555/GCA_001049555.1_7330
Bacteroides thetaiotaomicron	14-1069-2	Illumina	SPAdes	102	Poster (Sydenham et al. 2015)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/816/245/GCA_001816245.1_ASM181624v1
Bacteroides thetaiotaomicron	19_BTHE	Illumina	Abyss	351	26230489	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/055/755/GCA_001055755.1_ASM105575v1
Bacteroides thetaiotaomicron	2789STDY5668873	Illumina	Unknown	31	27144353	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/405/255/GCA_001405255.1_13414_6_57
Bacteroides thetaiotaomicron	2789STDY5834846	Illumina	Unknown	63	27144354	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/405/095/GCA_001405095.1_13470_2_65
Bacteroides thetaiotaomicron	2789STDY5834899	Illumina	Unknown	52	27144355	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/405/895/GCA_001405895.1_14207_7_22
Bacteroides thetaiotaomicron	2789STDY5834945	Illumina	Unknown	28	27144356	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/406/075/GCA_001406075.1_14207_7_68
Bacteroides thetaiotaomicron	2e6A	Unknown	Unknown	1730	Unpublished (WU)	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/373/135/GCA_001373135.1_2e6A_assembly
Bacteroides thetaiotaomicron	3a5B	Unknown	Unknown	2372	Unpublished (WU)	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/937/835/GCA_000937835.1_3a5B_assembly
Bacteroides thetaiotaomicron	KLE1254	Illumina	Velvet	108	Unpublished (WU)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/578/565/GCA_001578565.1_ASM157885v1
Bacteroides thetaiotaomicron	KPPR-3	Unknown	Unknown	38	Unpublished (DOE-JGI)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/109/385/GCA_900109385.1_IMG-taxon_2593339226_annotated_assembly
Bacteroides thetaiotaomicron	dnLKV9	Illumina	allpaths	6	Unpublished (BI)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/403/155/GCA_000403155.2_Bact_thet_dnLKV9_V1
Porphyromonas gingivalis	11A	Illumina	Velvet	89	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/295/GCA_900157295.1_11A
Porphyromonas gingivalis	13_1	Illumina	Velvet	68	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/285/GCA_900157285.1_13-1
Porphyromonas gingivalis	3A1	Illumina	Velvet	56	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/325/GCA_900157325.1_3A1
Porphyromonas gingivalis	3_3	Illumina	Velvet	72	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/215/GCA_900157215.1_Strain_3-3
Porphyromonas gingivalis	7BTOBR	Illumina	Velvet	72	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/225/GCA_900157225.1_7BTOBR
Porphyromonas gingivalis	84_3	Illumina	Velvet	50	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/275/GCA_900157275.1_84-3
Porphyromonas gingivalis	A7A1_28	Illumina	Velvet	22	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/265/GCA_900157265.1_A7A1-28
Porphyromonas gingivalis	AFR5B1	Illumina	Velvet	88	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/345/GCA_900157345.1_AFR-5B1
Porphyromonas gingivalis	ATCC 49417	Illumina	Velvet	77	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/255/GCA_900157255.1_ATCC49417
Porphyromonas gingivalis	Ando	Illumina	Velvet	112	26543123	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/297/745/GCA_001297745.1_ASM129774v1
Porphyromonas gingivalis	F0185	Illumina	Velvet	113	Unpublished (HMP)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/467/955/GCA_000467955.1_ASM46795v1
Porphyromonas gingivalis	F0566	Illumina	Velvet	192	Unpublished (HMP)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/467/975/GCA_000467975.1_ASM46797v1
Porphyromonas gingivalis	F0568	Illumina	Velvet	154	Unpublished (HMP)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/467/795/GCA_000467795.1_ASM46779v1
Porphyromonas gingivalis	F0569	Illumina	Velvet	111	Unpublished (HMP)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/467/815/GCA_000467815.1_ASM46781v1
Porphyromonas gingivalis	F0570	Illumina	Velvet	117	Unpublished (HMP)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/467/835/GCA_000467835.1_ASM46783v1
Porphyromonas gingivalis	MP4-504	Illumina	SPAdes	92	27056232	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/553/795/GCA_001553795.1_ASM155379v1
Porphyromonas gingivalis	SJD2	Illumina	SOAPdenovo	117	24385574	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/503/975/GCA_000503975.1_SJD2
Porphyromonas gingivalis	SU60	Illumina	Velvet	53	28184216	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/157/235/GCA_900157235.1_YH522
Porphyromonas gingivalis	W4087	Illumina	Velvet	114	Unpublished (HMP)	2013	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/467/995/GCA_000467995.1_ASM46799v1
Porphyromonas gingivalis	W50	Illumina	Celera	104	Unpublished (JCVI)	2012	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/271/945/GCA_000271945.1_PgingivalisW50v1.0
Tannerella forsythia	9610	Illumina	SPAdes	79	Unpublished (WU)	2017	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/938/785/GCA_001938785.1_ASM193878v1
Tannerella forsythia	ATCC 43037	Illumina	SPAdes	141	26067981	2015	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/006/485/GCA_001006485.1_ASM100648v1
Tannerella forsythia	UB20	Illumina	Unknown	93	Unpublished (UoS)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/096/735/GCA_900096735.1_TFUB20
Tannerella forsythia	UB22	Illumina	Unknown	98	Unpublished (UoS)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/096/715/GCA_900096715.1_TFUB22
Tannerella forsythia	UB4	Illumina	Unknown	71	Unpublished (UoS)	2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/096/725/GCA_900096725.1_TFUB4

Additional file 14. Table S3. CDS/pseudogene differences by strain.

Strain	CDS to pseudogene	Pseudogene to CDS	Fusion	Separation	Strand change
ATCC 33277	18	13	3 CDS -> 1 pseudogene 2 pseudogene -> 1 pseudogene 8 CDS -> 4 pseudogene 2 CDS + 1 pseudogene -> 1 pseudogene 4 CDS + 4 pseudogene -> 4 pseudogene 1 CDS + 1 pseudogene -> 1 CDS 2 CDS -> 1 CDS	1 pseudogene -> 2 CDS	1 CDS -> 1 CDS 1 pseudogene -> 1 CDS
TDC60	15	8	8 CDS -> 4 pseudogene 2 pseudogene -> 1 CDS	1 pseudogene -> 2 CDS	1 pseudogene -> 1 CDS
W83	16	7	16 CDS -> 8 pseudogene 2 pseudogene + 1 CDS -> 1 pseudogene 1 pseudogene + 1 CDS -> 1 pseudogene	—	2 CDS -> 2 CDS

Additional file 15. Table S4. Classification of additional CDS/pseudogenes.

	Macromolecules and ion transport and metabolism			Nucleic Acids & protein metabolism			Related to mobile elements			Proteins with subcellular localization or known domains			Hypothetical proteins		
Strain	ATCC 33277	TDC	W83	ATCC 33277	TDC	W83	ATCC 33277	TDC	W83	ATCC 33277	TDC	W83	ATCC 33277	TDC	W83
New pseudogenes	0	0	0	0	0	0	8	2	5	1	0	1	0	0	0
New CDS	5	1	1	12	2	2	31	2	9	1	1	0	6	0	0
CDS to pseudogenes	3	1	2	2	5	2	13	8	9	0	1	1	0	0	2
Pseudogenes to CDS	3	3	4	6	1	1	3	3	2	0	1	0	1	0	0
Fusions	7	2	5	0	0	1	5	2	2	1	1	1	0	0	1
Separations	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0
Coding Strand Change	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1
Sub-Total	18	7	12	22	10	6	60	17	27	4	4	4	8	1	4

Additional file 16. Table S5. Core genes with pseudogenisation.

gene	product	ATCC 33277	TDC60	W83
araC	arabinose operon regulatory protein			*
bepA_4	**Beta-barrel assembly-enhancing protease**	*	*	
cas1_2	CRISPR-associated protein Cas1	*		
cirA_2	TonB-dependent receptor putative	*		
clcA	H(+)/Cl(-) exchange transporter			*
cobQ	cobyric acid synthase			*
cshA_1	ATP-dependent RNA helicase		*	
drsE	Coenzyme A disulfide reductase	*		
DUF3078_1	hypothetical protein	*		
feuA	periplasmic binding protein			*
fib3	fibronectin type III domain protein			*
FIG00936174	hypothetical protein	*		
fimB	fimbriae length regulator	*		
fimS	two-component sensor histidine kinase			*
fucP	L-fucose-proton symporter		*	
glkA	**glucokinase**	*		*
GSCFA	hypothetical protein			*
haeS	two-component sensor histidine kinase	*		
hipB_1	transcriptional regulator		*	
lipo_03	**lipoprotein FIG00936185**	*	*	
mepM	Murein DD-endopeptidase	*		
mfa1	fimbrillin protein			*
mfa5	fimbrillin associated protein			*
nahK	N-acetylhexosamine 1-kinase			*
natA_1	ABC transporter ATP-binding protein			*
nrfA	cytochrome c nitrite reductase catalytic subunit	*		
nrfB	cytochrome c biogenesis protein			*
ogt	Methylated-DNA-protein-cysteine methyltransferase	*		
omp_09	outer membrane protein immunoreactive 23 kDa antigen			*
PAD_porph_4	peptidylarginine deiminase		*	
PF07877	PF07877 family protein			*
pfeA_1	ferric enterobactin receptor			*
sprA	gliding motility protein			*
wapA_2	tRNA(Glu)-specific nuclease			*
ydaF_1	ribosomal N-acetyltransferase		*	
yqhD	**Fe-dependent oxidoreductase alcohol dehydrogenase**		*	*
Total		**13**	**8**	**19**

2. ARTICLE 2: L'évolution des *Porphyromonas* racontée par leurs loci de fimbriae

The evolution of *Porphyromonas* as narrated by their fimbriae gene loci

Soumis à Microbial Genomics

Le deuxième article présenté dans ce document est un *short paper* soumis à la revue Microbial Genomics. Il s'agit d'une communication courte avec un format précis : 2000 mots et 8 tables/figures maximum.

Suite à notre description sommaire du génome *core* de trois souches de *Porphyromonas gingivalis*, nous nous sommes aperçus que certains des gènes présents dans cette catégorie peuvent avoir des pourcentages d'identité nucléotidique plus faibles que la moyenne. Nous avons séparé les gènes qui ont moins de 97% d'identité et les avons classés dans ce que nous proposons d'appeler un *génome core variant*. Dans cette catégorie, nous avons observé une grande proportion de gènes codant des protéines dites *facteurs de virulence*. Dans les définitions plus récentes des facteurs de virulence, en plus des protéines effectrices classiques comme les toxines bactériennes, nous pouvons désormais lister toute molécule ou structure capable de moduler la réponse du système immunitaire de l'hôte et de permettre la permanence des bactéries en contact avec les cellules de l'hôte. En effet, toutes les structures qui permettent aux bactéries de nous coloniser sont nécessaires pour qu'elles puissent nous rendre malades.

Comme pour les autres types de bactéries, les fimbriae sont les adhésines les plus connues de *P. gingivalis*, même si cette bactérie possède également un nombre considérable d'hémagglutinines. De nombreuses équipes ont étudié la diversité des fimbriae intra-espèce et des études épidémiologiques et *in vitro* montrent un certain lien entre le type de fimbriae de la souche et la virulence de celle-ci. Cependant, les résultats sont contradictoires alors que plusieurs types de fimbriae peuvent être détectés *in vivo* dans une même poche parodontale d'un patient atteint de parodontite.

Dans la littérature scientifique il existe un vide d'information sur la diversité des fimbirae sur les autres espèces de *Porphyromonas*, que se soient des espèces isolées de cavités orales saines ou atteintes de paradontite, humaines ou d'animaux de compagnie, notamment les chiens et les chats, ou d'espèces non-orales.

Dans l'article présenté ici, nous avons fait un travail de biocuration pour annoter les gènes codant des fimbriae dans les génomes de *Porphyromonas*. Nous avons identifié les domaines protéiques conservés et prédit la localisation cellulaire des protéines afin de générer un répertoire permettant de modéliser les événements de gain ou perte de ces loci dans le genre *Porphyromonas*. Nous avons collecté également toutes les données associés aux génomes disponibles dans les bases de données publiques, notamment l'hôte, le site d'isolement et la présence ou non de maladie lors de l'isolement. Notre objectif étant de corréler ces métadonnées aux histoires évolutives du répertoire des gènes codant des fimbriae et à la phylogénie des espèces.

MICROBIAL GENOMICS

Short Paper template

The evolution of *Porphyromonas* as narrated by their fimbriae gene loci

Luis Acuña-Amador[1,2] and Frédérique Barloy-Hubler[1].

1. Institut de Génétique et Développement de Rennes, CNRS, UMR6290, Université de Rennes 1, Rennes, France.
2. Laboratorio de Investigación en Bacteriología Anaerobia, Centro de Investiación en Enfermedades Tropicales, Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica.

ABSTRACT

Fimbriae are bacterial adhesins that allow attachment to surfaces or cells. As such, they are often regarded as bacterial virulence factors, for example in *Porphyromonas gingivalis* and *P. gulae*, two oral bacteria associated with human and canine periodontal disease. These fimbriae have been widely described in *P. gingivalis*, but not or rarely in other *Porphyromonas* species. To fill this gap, we have collected all available genomes of 17 species of *Porphyromonas* and their associated metadata. We traced evolutionary relationships of these species using 16S rDNA and phylogenomic analysis. Then we identified and compared the fimbrillin genetic repertoire and organization in each of the 54 strains studied. We show a remarkable genetic proximity of *P. gulae*, *P. gingivalis* and *P. macacae* that possess the same fimbriae and host range. Our study establish that the *Porphyromonas* share only an ancestral fimbrillin locus and that all the remaining loci were acquired during evolution and adaptation of each species. We also demonstrate that non-oral *Porphyromonas*, described as little or non-pathogenic have more fimbrilin loci than the so-called more "virulent" species. We conclude that even if fimbriae are structures necessary for cell attachment and colonization, their primary role might be environmental adaptation.

DATA SUMMARY

All genomes and associated metadata used in this article are publically available in NCBI, ENA, PATRIC and JGI databases. Accession numbers are listed (Supplementary Table 1). Fibrillin protein sequences are available at NCBI protein database.

I/We confirm all supporting data, code and protocols have been provided within the article or through supplementary data files ☑

IMPACT STATEMENT

The *Porphyromonas* genus includes species isolated from healthy or sick oral cavities and skin lesions of humans, dogs, cats and nonhuman primates. These bacteria live in biofilms, attached to epithelial cells. *Porphyromonas* fimbriae are responsible for adhesion to other bacteria and host tissues, and have been exclusively described in *P. gingivalis* and *P. gulae*. This paper seeks to provide new information on all *Porphyromonas* fimbriae biodiversity. These bacteria are often associated with local pathologies as well as systemic diseases and cancers, and the interest of this study is substantial since fimbriae are regarded as major virulence factors. A better understanding of their evolutionary history will help to evaluate their possible role in host-specificity and pathogenicity. This study shows that fimbriae genes appear to be acquired during *Porphyromonas* evolution and that the number of loci does not correlate with virulence or host specificity. We propose that fimbrillin are proteins of environmental adaptation rather than virulence factors.

INTRODUCTION

The *Porphyromonas* belongs to the Bacteroidetes phylum and corresponds to anaerobic bacteria which are part of mammal microbiota. Two thirds of the species are of oral origin, primaraly isolated from dogs and humans (1-3), many being implicated in periodontitis development (4-6).

P. gingivalis is the most studied species and its principal virulence factors are the proteolytic enzymes called gingipains and its fimbriae (6). Fimbriae confer the capacity to adhere and invade gingival epithelial cells (7-9). *P. gingivalis* has three types of fimbriae: major fimbriae (polymers of FimA encoded by *fimA*) (10), minor fimbriae (polymers of Mfa1 encoded by *mfa1*) (11) and a recently described third type (encoded by the gene *pgn*_1808 in *P. gingivalis* ATCC 33277) (12). Studies showed that *P. gulae* fimbriae are similar to those of *P. gingivalis* (13, 14). Nevertheless, little is known about these attachment proteins in other species of the genus.

In this study, we analysed *in silico* all the available genomes to describe all fimbriae genes in *Porphyromonas*, to investigate the evolution of these loci and to correlate them to isolation site and/or host and pathogenicity.

METHODS

Porphyromonas genomes and associated metadata (isolation site/host) were retrieved from NCBI, ENA, PATRIC and JGI databases. Fimbrillin/Fimbriae protein sequences were retrieved from NCBI protein database and used to extract genomic loci in all genomes. For each gene, proteic domains and subcellular localisation were predicted. Genomes, 16S rRNA genes and fimbriae loci were used to generate trees and analyse fimbrillin gene repertoire and diversity (Supplementary Fig. 1).

RESULTS AND DISCUSSION

Porphyromonas genomic sequences have related data such as isolation source, host, site or organ, and sometimes associated disease or symptoms. By compiling several information sources, we synthetized and compare these information for the 17 species studied (**Supplementary Table 2**). Two groups are observed: i) non-oral species, isolated from digestive, genital or skin areas in contexts of infectious disease (abscesses, necrosis) and ii) oral species isolated from healthy sites (saliva, gingiva) or gingival lesions (gingivitis or periodontitis). Genomes are from a low number of studies and few hosts, but the observed diversity seems sufficient to explore if there is a relationship between fimbriae repertoires and the host, isolation body site and/or associated illness.

Conventional taxonomic analysis using the 16S rRNA gene phylogenetic tree **(Fig. 1)** demonstrated that all strains of each species are homogeneously clustered within separate branches. These results showed a separate branching of the non-oral species. Near the root, *P. bennonis*, *P. levii* and *P. somerae*, described as opportunistic pathogens in clinical infections (15-17) whereas the remaining species in this group, *P. asaccharolytica* and *P. uenonis* form a clade and clustered with three oral moderately virulent species. This is interesting since *P. asaccharolytica* is sometimes detected, in small proportions, in oral cavities of healthy patients (18). This analysis confirmed the close phylogenetic relationship between *P. gingivalis* and *P. gulae* (19).

Observing the distance tree based on the average nucleotide identity (ANIb and ANIm, **Fig. 2**), some of the results are the same including the robust clustering of *P. gingivalis* and *P. gulae*. However, using this metric, we can observe that the non-oral species cluster together. There is also a better correlation between the groups and the observed pathogenicity level, even if this information is poorly documented and should be used with caution. Finally, there is a good correlation with the host, as human and animal species separate, with the remarkable exception of the human-associated *P. gingivalis* that, as mentioned, cluster with the canine-associated *P. gulae*.

Predicted fimbrial proteins carry at least one of four different domains **(Fig. 3)**, all initially described in *P. gingivalis*: FimA (pfam06321), associated fimbrillin-C (pfam15495), Mfa-like-1 (pfam13149) and Mfa2 (pfam08842). These domains are mainly found in the Bacteroidetes with very few exceptions. We find them associated in operonic *loci* with recurring domains such as DUF3575 or OmpA_C found in ß-barrel proteins; or DUF5119 discovered in *Bacterioides ovatus* Fim4B protein, an anchoring subunit of the fimbriae complex (20). Other associations are less expected, like the domain Nuc (endonuclease), the BACON (Bacteroidetes-Associated Carbohydrate-binding Often N-terminal) or the von Willebrand factor, yet found associated with pilin in other bacteria (21).

To establish the most exhaustive repertoire of *Porphyromonas* fimbriae, we searched all genomes for these domains and retrieved all loci. We observed that more than 95% of genes identified in our study were annotated "of unknown function" and that a very large number of *loci* are located, complete or incomplete, at the end of contigs. We finally identify 35 non-redundant *loci* of fimbriae genes in the 17 species of *Porphyromonas* studied, that correspond to a mosaicity of genes **(Fig. 3)**. The different combinations observed concerned the position of the domains in proteins (N-terminal, C-terminal or central), the order of the different genes in the operons and the cellular localisation

predicted *in silico* from the amino acids sequences **(Fig. 3)**. All locus_tag are presented for each loci in **Supplementary Table 1**.

The fimC domain is primarily located in the C-terminal region of proteins (96%). The mfa2 domain is mainly central (85 %) but sometimes found in N-terminal whereas the fimA domain is preferentially found in N-terminal (76 %) but some central or C-terminal positions are observed. The mfa1 domain is found equally central and N-terminal **(Fig. 3)**. All these variations create an important biodiversity of fimbrillin proteins in the *Porphyromonas* genus.

If we extend this study of biodiversity to the genomic loci and draw the most likely evolutionary scenario from the matrix of presence/absence of these loci **(Fig. 4a)**, we notice that the *Porphyromonas* genus has one single ancestral fimbrial gene, namely *fimC* **(Fig. 4b)**. This gene encodes a protein recently studied in *P. gingivalis* as a novel detergent and heat labile fimbrillin constitutivly expressed (12). This is the only fimbrillin locus detected in non-virulent species *P. pasteri* and *P. catoniae* as well as in strains with no identified species (named "COT" in this study) whereas *P. somerae* seems to have lost this locus (or never acquired it). Then, *Porphyromonas* fimbrillin story seems to have mostly evolved though loci acquisitions (**Fig. 4b**), gains which are specific of species or group of species. Some loci are acquired early in evolution: locus 1 is present in 10 out of 17 species and locus 2 was acquired by only 4 oral species. The only locus that seems to be gained twice independently is locus 12 in *P. cangingivalis* and in *P. bennonis* (**Fig. 4b**). Interestingly, *P. asaccharolytica* and *P. uenonis*, which can be found in many human sites such as the cervix, ear, intestine, genitalia and mouth (1) are the species that has acquired greater number of loci. For both species, strains isolated from serious infections (ulcers or empyema) have additional loci, which may be favourable environments for loci transfers.

Since the majority of genomes are available as drafts, the absence of loci was verified by mapping the reads of the strain lacking the locus on the strain having it. When no reads cover these loci, they could therefore be judged absent (**Fig. 5a**). However, in the vast majority of cases, the reads are not available and therefore doubt persists on the absence of those loci. In addition, some genes in the loci are pseudogenised by mutation or insertion of transposases (**Fig. 5b**). Finally, the fragmented nature of the draft genomes is responsible for the separation of some of these loci into several contigs (**Fig. 5c**). These results show the importance of curation and caution to analyse loci repertories and diversity.

In general, some loci are also observed in Bacteroidetes from several environments (**Table 1**), while other are restricted to one *Porphyromonas* species. This could be a result either from a species specialisation or an incomplete genomic data in the Bacteroidetes phylum.

CONCLUSION

Porphyromonas fimbriae diversity is remarkable quantitatively and qualitatively. Some loci are ancestral while most seems to be acquiered through horizontal gene transfer, however restricted to the Bacteroidetes phylum. This repertoire, probably extended if other Bacteroidetes genus were studied, seems to correlate with the multitude of body sites that these bacteria could colonise. Habitat adaptation may drive fimbriae diversity and its relationship to virulence needs to be substantiated.

AUTHOR STATEMENTS

Funding information

The PhD research of LA-A is founded by a scholarship from the Oficina de Asuntos Internacionales y Cooperación Externa, Universidad de Costa Rica, San José, Costa Rica.

Acknowledgements

Not applicable

Ethical statement

No ethics or consent approval was requiered for the research in this study.

Conflicts of interest

The authors declare that they have no competing interests.

REFERENCES

1. Krieg NR, Ludwig W, Euzéby J, Whitman WB. Phylum XIV. Bacteroidetes phyl. nov. In: Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, et al., editors. Bergey's Manual® of Systematic Bacteriology: Volume Four The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes. New York, NY: Springer New York; 2010. p. 25-469.
2. Coil DA, Alexiev A, Wallis C, O'Flynn C, Deusch O, Davis I, et al. Draft genome sequences of 26 porphyromonas strains isolated from the canine oral microbiome. Genome announcements. 2015;3(2).
3. Davis IJ, Wallis C, Deusch O, Colyer A, Milella L, Loman N, et al. A cross-sectional survey of bacterial species in plaque from client owned dogs with healthy gingiva, gingivitis or mild periodontitis. PloS one. 2013;8(12):e83158.
4. O'Flynn C, Deusch O, Darling AE, Eisen JA, Wallis C, Davis IJ, et al. Comparative Genomics of the Genus Porphyromonas Identifies Adaptations for Heme Synthesis within the Prevalent Canine Oral Species Porphyromonas cangingivalis. Genome biology and evolution. 2015;7(12):3397-413.
5. do Nascimento Silva A, de Avila ED, Nakano V, Avila-Campos MJ. Pathogenicity and genetic profile of oral Porphyromonas species from canine periodontitis. Archives of oral biology. 2017;83:20-4.
6. Hajishengallis G. Porphyromonas gingivalis-host interactions: open war or intelligent guerilla tactics? Microbes and infection. 2009;11(6-7):637-45.
7. Moreno S, Contreras A. Functional differences of Porphyromonas gingivalis Fimbriae in determining periodontal disease pathogenesis: a literature review. Colombia medica (Cali, Colombia). 2013;44(1):48-56.

8. Njoroge T, Genco RJ, Sojar HT, Hamada N, Genco CA. A role for fimbriae in Porphyromonas gingivalis invasion of oral epithelial cells. Infection and immunity. 1997;65(5):1980-4.
9. Umemoto T, Hamada N. Characterization of biologically active cell surface components of a periodontal pathogen. The roles of major and minor fimbriae of Porphyromonas gingivalis. Journal of periodontology. 2003;74(1):119-22.
10. Yoshimura F, Takahashi K, Nodasaka Y, Suzuki T. Purification and characterization of a novel type of fimbriae from the oral anaerobe Bacteroides gingivalis. Journal of bacteriology. 1984;160(3):949-57.
11. Hamada N, Sojar HT, Cho MI, Genco RJ. Isolation and characterization of a minor fimbria from Porphyromonas gingivalis. Infection and immunity. 1996;64(11):4788-94.
12. Nagano K, Hasegawa Y, Yoshida Y, Yoshimura F. Novel fimbrilin PGN_1808 in Porphyromonas gingivalis. PloS one. 2017;12(3):e0173541.
13. Hamada N, Takahashi Y, Watanabe K, Kumada H, Oishi Y, Umemoto T. Molecular and antigenic similarities of the fimbrial major components between Porphyromonas gulae and P. gingivalis. Veterinary microbiology. 2008;128(1-2):108-17.
14. Nomura R, Shirai M, Kato Y, Murakami M, Nakano K, Hirai N, et al. Diversity of fimbrillin among Porphyromonas gulae clinical isolates from Japanese dogs. The Journal of veterinary medical science. 2012;74(7):885-91.
15. Summanen PH, Durmaz B, Vaisanen ML, Liu C, Molitoris D, Eerola E, et al. Porphyromonas somerae sp. nov., a pathogen isolated from humans and distinct from porphyromonas levii. Journal of clinical microbiology. 2005;43(9):4455-9.
16. Summanen PH, Lawson PA, Finegold SM. Porphyromonas bennonis sp. nov., isolated from human clinical specimens. International journal of systematic and evolutionary microbiology. 2009;59(Pt 7):1727-32.
17. Sweeney M, Watts J, Portis E, Lucas M, Nutsch R, Meeuwse D, et al. Identification of Porphyromonas levii isolated from clinical cases of bovine interdigital necrobacillosis by 16S rRNA sequencing. Veterinary therapeutics : research in applied veterinary medicine. 2009;10(4):E1-10.
18. Ziouani S, Klouche KN, Benyelles I, Hoceini A, Aissaoui N, Nas F, et al. Oral microflora of supragingival and subgingival biofilms in Algerian healthy adults. African Journal of Microbiology Research. 2015;9:1548-57.
19. Fournier D, Mouton C, Lapierre P, Kato T, Okuda K, Menard C. Porphyromonas gulae sp. nov., an anaerobic, gram-negative coccobacillus from the gingival sulcus of various animal hosts. International journal of systematic and evolutionary microbiology. 2001;51(Pt 3):1179-89.
20. Xu Q, Shoji M, Shibata S, Naito M, Sato K, Elsliger MA, et al. A Distinct Type of Pilus from the Human Microbiome. Cell. 2016;165(3):690-703.
21. Okura M, Osaki M, Fittipaldi N, Gottschalk M, Sekizaki T, Takamatsu D. The minor pilin subunit Sgp2 is necessary for assembly of the pilus encoded by the srtG cluster of Streptococcus suis. Journal of bacteriology. 2011;193(4):822-31.

FIGURES AND TABLES

Figure 1. Full length 16S rDNA *Porphyromonas* phylogenetic tree. In blue, species of oral origin and in orange, non-oral species.

Figure 2. Whole-genome ANI (ANIm and AMIb mean) *Porphyromonas* phylogenomic tree. Species are highlighted in blue or orange (oral or non-oral origin), their isolation host and pathogenicity level are depicted : in green, low virulence species and in red, species associated with pathology.

Figure 3. Fimbriae loci in *Porphyromonas* spp. Loci 1 to 3 are the most studied, mainly in *P. gingivalis*. For each gene, their proteic domains and the predicted subcellular localisation are presented.

Figure 4. Distribution of fimbriae loci in *Porphyromonas* spp. a. Loci matrix with absence in gray; presence in all strains in dark blue and in some strains in light blue. **b.** Derived tree with species specific loci in black; non exclusive *Porphyromonas* spp. in red (*= loci present in only some strains).

Figure 5. Exemples of fimbriae loci diversity. For different strains: **(a)** confirmed absence of a locus, **(b)** pseudogenisation and **(c)** splited loci.

Table 1. Non *Porphyromonas* spp. fimbriae loci repartition.

Supplementary Figure 1. Methods overview. Schematic view of data sources and methods used for analysis. Data types in blue, output in orange, figures in green and tables in purple.

Supplementary Table 1. Strains, locus_tag prefixes and numbers for each fimbiae locus identified in this study.

Supplementary Table 2. Metadata associated to all genomes used in this study.

Figure 1. Full length 16S rDNA *Porphyromonas* phylogenetic tree.

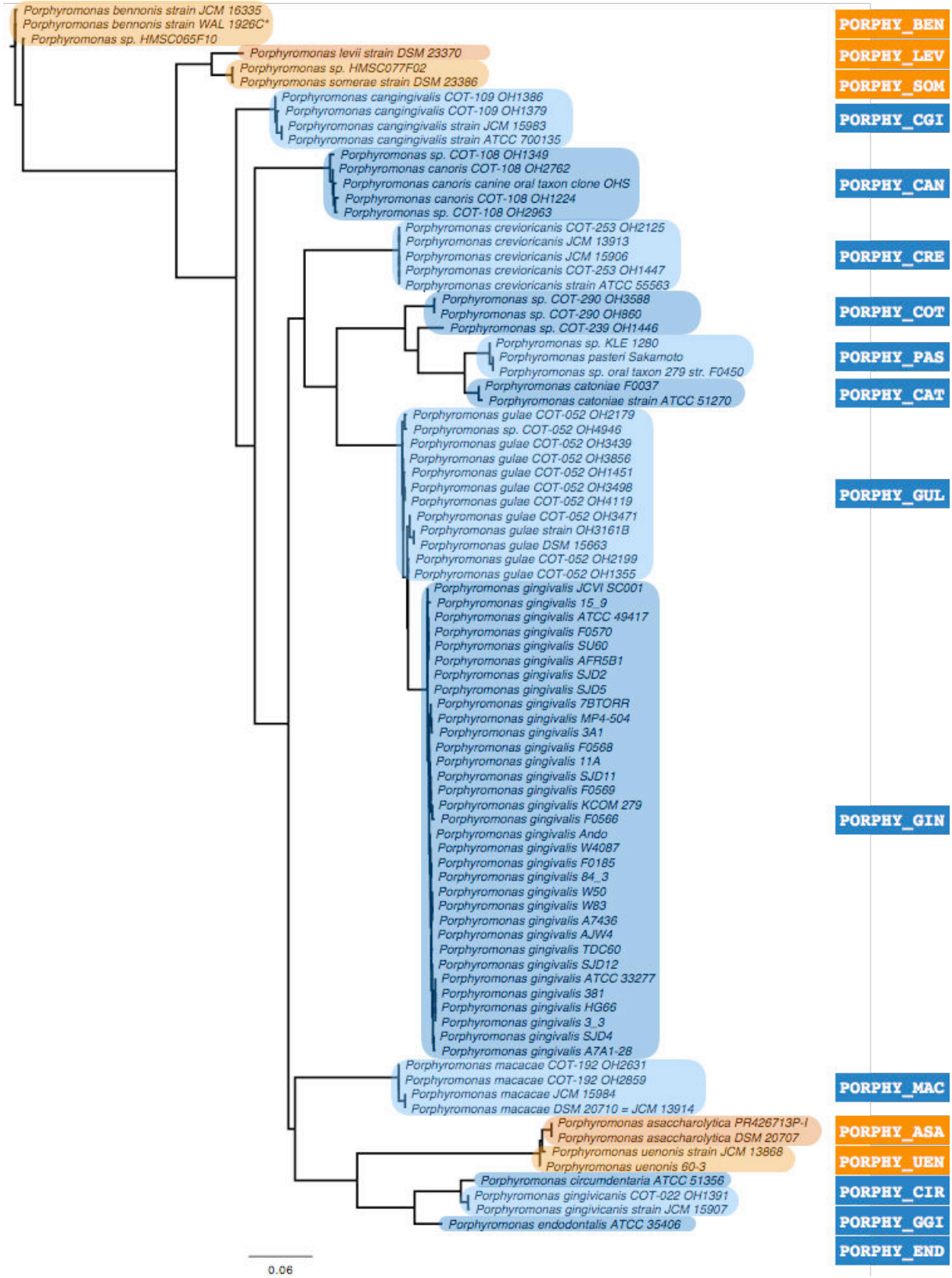


Figure 2. Whole-genome ANI (ANIm and AMIb mean) *Porphyromonas* phylogenomic tree.

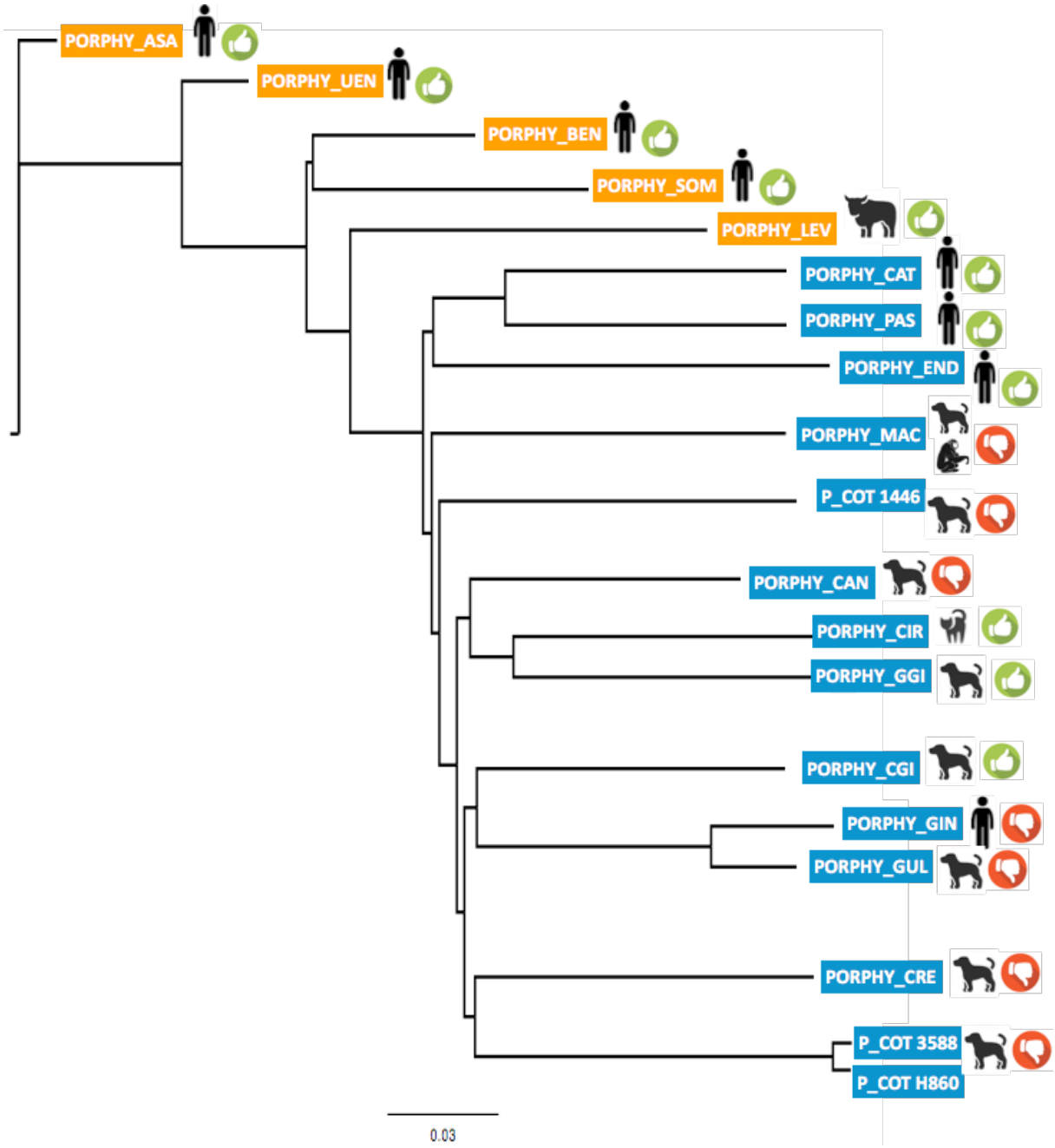


Figure 3. Fimbriae loci in *Porphyromonas* spp.

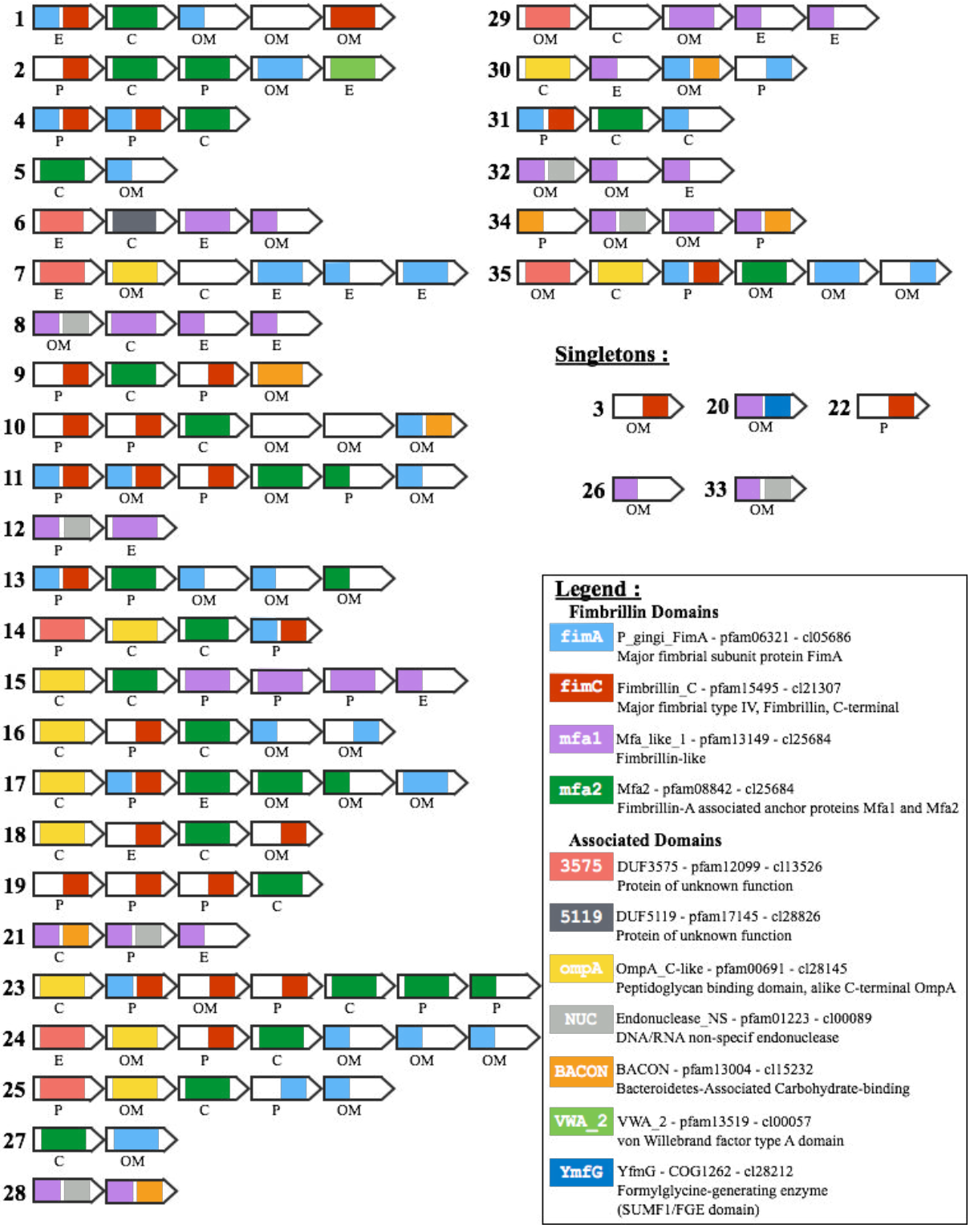


Figure 4. Distribution of fimbriae loci in *Porphyromonas* spp.

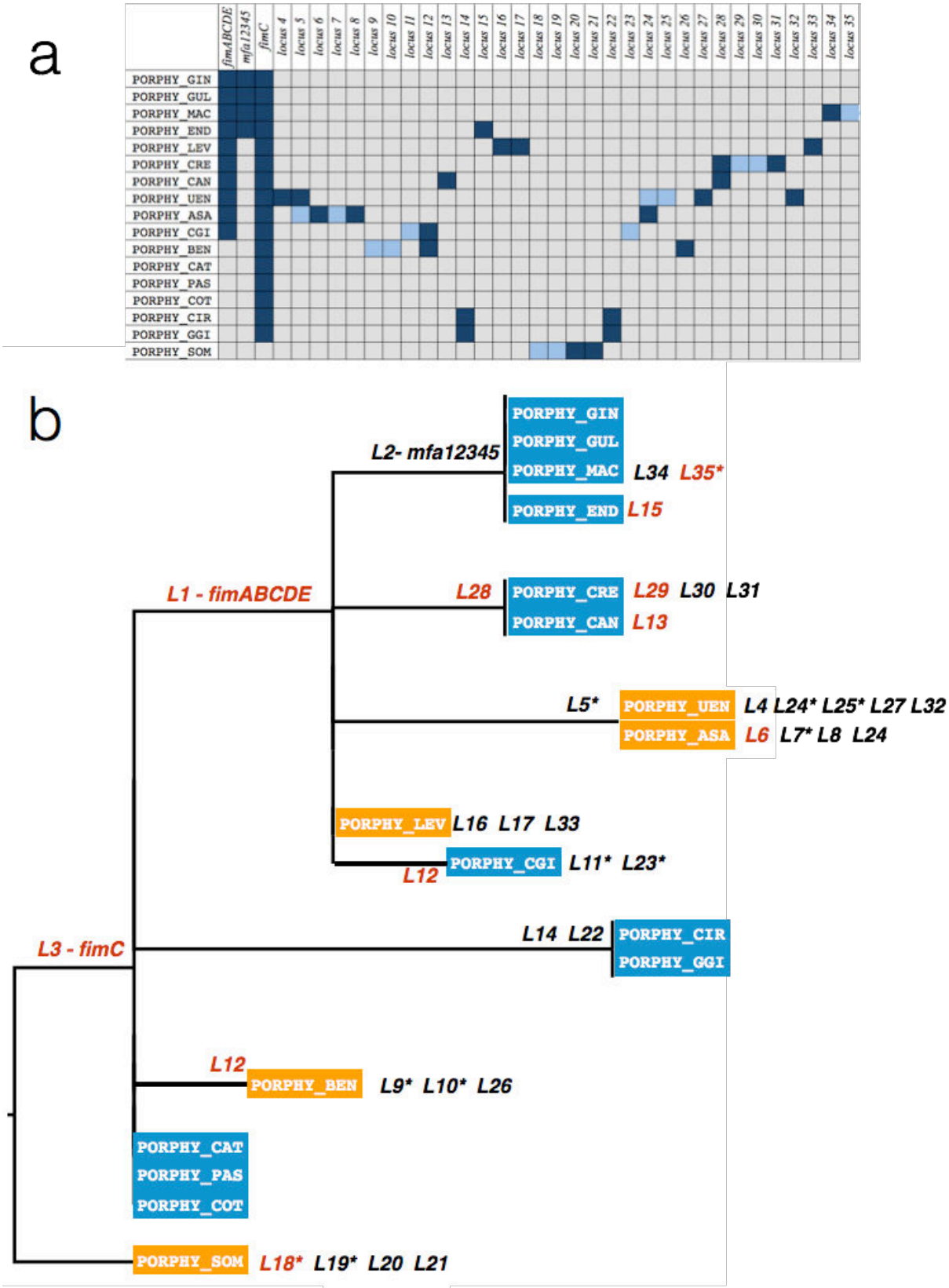
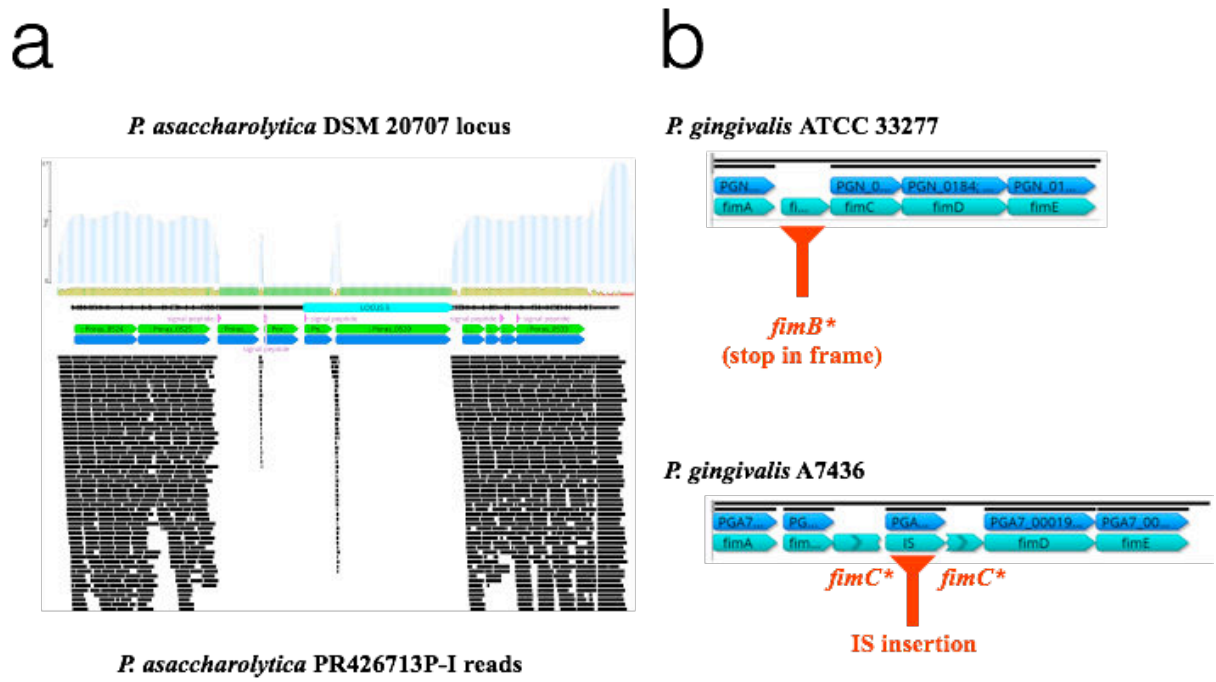


Figure 5. Exemples of fimbriae loci diversity.



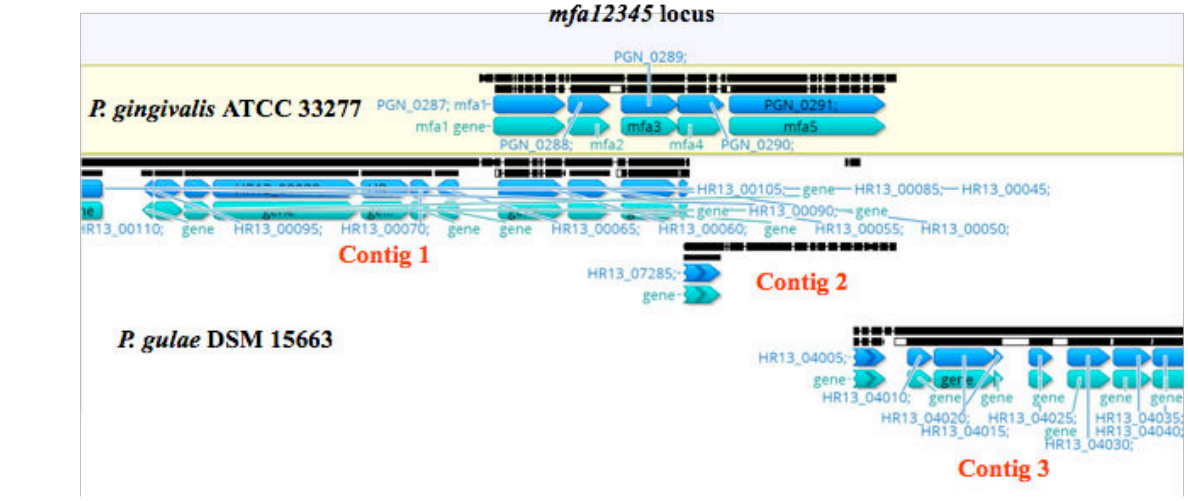
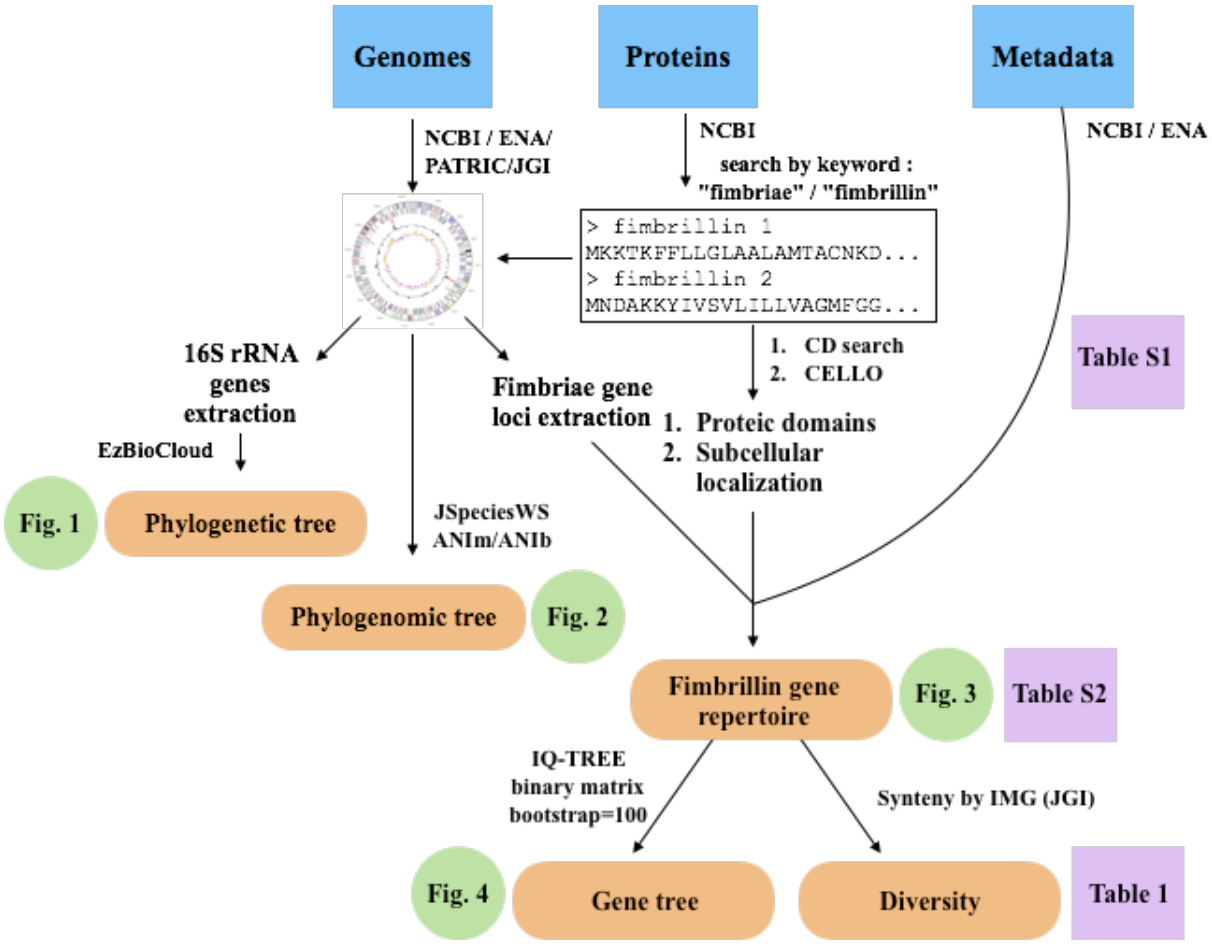


Table 1. Non *Porphyromonas* spp. fimbriae loci repartition.

	fimABCDE	*mfa12345*	*fimC*	*locus 6*	*locus 12*	*locus 13*	*locus 15*	*locus 18*	*locus 28*	*locus 29*	*locus 35*	Phylum
Alloprevotella rava F0323 (Human oral cavity)					oral							Bacteroidetes
Bacteroidales bacterium KA00251 (HMP, medical vagina isolate)			skin, blood/fluid									Bacteroidetes
Bacteroides eggerthii DSM 20697 (Human, Large intestine, Human feces)				gut								Bacteroidetes
Bacteroides fluxus YIT 12057 (Human feces)								gut				Bacteroidetes
Bacteroides graminisolvens (Rice-straw residue collected from cattle farms)					environmental							Bacteroidetes
Bacteroides intestinalis 341, DSM 17393 (Human feces collected in Japan)				gut								Bacteroidetes
Bacteroides neonati MS4 preterm neonate stool											gut	Bacteroidetes
Bacteroides ovatus (mammal digestive system)					gut							Bacteroidetes
Bacteroides plebeius M12, DSM 17135 Human feces collected in Japan									gut			Bacteroidetes
Bacteroides pyogenes DSM 20611 (pig skin)	skin, blood/fluid											Bacteroidetes
Bacteroides reticulotermitis JCM 10512 (Gut of the subterranean termite)				gut								Bacteroidetes
Bacteroides sp. 3_1_19 (Human)								Not available				Bacteroidetes
Bacteroides sp. 4_3_47FAA (Human fecal sample)								gut				Bacteroidetes
Bacteroides stercoris ATCC 43183 Human feces from Texas									gut			Bacteroidetes
Bacteroides vulgatus 274-1D4 (Gastrointestinal tract of mus musculus)				gut								Bacteroidetes
Bacteroides vulgatus CL09T03C04 (Human)								Not available				Bacteroidetes
Bacteroidetes bacterium oral taxon 272 F0290 (Host, Human intestinal microflora)							gut					Bacteroidetes
Dysgonomonas capnocytophagoides DSM 22835 Human cutaneous abscess											skin, blood/fluid	Bacteroidetes
Dysgonomonas gadei ATCC BAA-286 infected gall bladder of 68-yr old male											skin, blood/fluid	Bacteroidetes
Dysgonomonas mossii DSM 22836 Human abdominal drainage											skin, blood/fluid	Bacteroidetes
Elizabethkingia (Spinal fluid from premature infant, Massachusetts)					skin, blood/fluid							Bacteroidetes
Elizabethkingia anophelis PW2809 neonatal meningitis									skin, blood/fluid			Bacteroidetes
Odoribacter laneus (Human feces)					gut							Bacteroidetes
Parabacteroides sp. 2_1_7 (Human fecal sample)								gut				Bacteroidetes
Phocaeicola abscessus CCUG 55929 (human brain abscess)							skin, blood/fluid					Bacteroidetes
Porphyromonas loveana (oral)	oral	oral	oral									Bacteroidetes
Propionibacterium acidifaciens F0233 (99% identity to Porphyromonas, oral)	oral	oral	oral			oral	oral					Bacteroidetes
Proteiniphilum sp. EBM-41 (Anaerobic digester at 37C)	environmental											Bacteroidetes
Sanguibacteroides justesenii OUH 334697 Blood Cultures from Two Different Patients	skin, blood/fluid											Bacteroidetes
several species of Bacteroides, Parabacteroides, Alistipes and Prevotella										Not available		Bacteroidetes
Sphingobacterium (Host-associated)					gut							Bacteroidetes

oral gut environmental skin, blood/fluid Not available

Supplementary Figure 1. Methods overview.



Supplementary Table 1. Strains, locus_tag prefixes and numbers for each fimbiae locus identified in this study.

	Accession number	locus_tag prefix	LOCUS 1	LOCUS 2	LOCUS 3	LOCUS 4	LOCUS 5	LOCUS 6	LOCUS 7	LOCUS 8
P. asaccharolytica PR426713P-I	AENO00000000	HMPREF9294_	1598 to 1602	-	0374	-	-	*reassembled* 0468 to 0197	-	1066 to 1069
P. asaccharolytica DSM 20707	NC_015501	Poras_	0283 to 0287	-	1299	-	0528 to 0529	0825 to 0829	0857 to 0862	1019 to 1023
P. bennonis DSM 23058/JCM 16335	AQWR00000000	B088DRAFT_	-	-	01263	-	-	-	-	-
P. sp. HMSC065F10 (*bennonis-like*)	LTXH00000000	HMPREF2890_	-	-	*reannotated between 02250-02260*	-	-	-	-	-
P. canginginvalis COT-109 OH1379	JQJF00000000	HQ34_	05550 to 05570	-	02210	-	-	-	-	-
P. canginginvalis COT-109 OH1386	JQJD00000000	HQ35_	07010 to 07030	-	07500	-	-	-	-	-
P. canginginvalis JCM 15983	BAKR00000000	JCM15983DRAFT_	00062 to 00066	-	01684	-	-	-	-	-
P. canginginvalis ATCC 700135	FUWL01000000	SAMN02745205_	00049 to 00053	-	00414	-	-	-	-	-
P. canoris COT-108 OH1224	JQZX00000000	HQ29_	05110 to 05130	-	00650	-	-	-	-	-
P. sp. COT-108 OH1349 (*canoris-like*)	JRAH00000000	JT26_	03010 to 03030	-	09370	-	-	-	-	-
P. sp. COT-108 OH2963 (*canoris-like*)	JRAP00000000	HQ39_	03990 to 04010	-	00925	-	-	-	-	-
P. canoris COT-108 OH2762	JQZV00000000	HQ43_	07935 to 07955	-	09055	-	-	-	-	-
P. catoniae F0037	AMEQ00000000	HMPREF9134_	-	-	01065	-	-	-	-	-
P. catoniae ATCC 51270	JDFF00000000	HMPREF0636_	-	-	1008	-	-	-	-	-
P. circumdentaria ATCC 51356	FUXE00000000	SAMN02745171_	-	-	01593	-	-	-	-	-
P. crevioricanis ATCC 55563	FUXH00000000	SAMN02745203_	01135 to 01139	-	01515	-	-	-	-	-
P. crevioricanis COT-253 OH1447	JQJC00000000	HQ38_	08900 to 08920	-	00070	-	-	-	-	-
P. crevioricanis COT-253 OH2125	JQJB00000000	HQ45_	07450 to 07470	-	07910	-	-	-	-	-
P. crevioricanis JCM 13913	BAOV00000000	PORCAN_	1404 to 1408	-	38	-	-	-	-	-
P. crevioricanis JCM 15906	BAOU00000000	PORCRE_	36 to 41	-	924	-	-	-	-	-
P. endodontalis ATCC 35406	ACNN00000000	POREN0001_	0820 to 0824	0387 to 0392	0653	-	-	-	-	-
P. gingivalis 11A	FUFE00000000	PGIN_11A_	00437 to 00441	01632 to 01639	01261	-	-	-	-	-
P. gingivalis 15_9	FUGF00000000	PGIN_15-9_	01588 to 01592	00356 to 00360	01143	-	-	-	-	-
P. gingivalis 3_3	FUFB00000000	PGIN_3-3_	00383 to 00387	00672 to 00677	01611	-	-	-	-	-
P. gingivalis 381	CP012889	PGF_	00001760 to 00001800	00002830 to 00002880	00017830	-	-	-	-	-
P. gingivalis 3A1	FUFC00000000	PGIN_3A1_	00849 to 00853	00160 to 00164	01536	-	-	-	-	-
P. gingivalis 7BTORR	FUFD00000000	PGIN_7BTORR_	00745 to 00749	00469 to 00473	01660	-	-	-	-	-
P. gingivalis 84_3	FUFG00000000	PGIN_84-3_	01611 to 01615	01269 to 01273	01869	-	-	-	-	-
P. gingivalis A7436	CP011995	PGA7_	00019920 to 00019970	00001690 to 00001730	00017440	-	-	-	-	-
P. gingivalis A7A1-28	CP013131	PGS_	00001480 to 00001520	00002700 to 00002730	00016750	-	-	-	-	-
P. gingivalis AFR5B1	FUFJ00000000	PGIN_AFR-5B1_	01272 to 01276	01541 to 01545	00930	-	-	-	-	-
P. gingivalis AJW4	CP011996	PGJ_	00019870 to 00019910	00001690 to 00001730	00016930	-	-	-	-	-
P. gingivalis Ando	BCBV00000000	PGANDO_	0824 to 0828	1057 to 1061	1295	-	-	-	-	-
P. gingivalis ATCC 49417	FUFH00000000	PGIN_ATCC49417_	00097 to 00101	fragment	00365	-	-	-	-	-
P. gingivalis F0185	AWVC00000000	HMPREF1988_	01642 to 01646	00787 to 00791	00474	-	-	-	-	-
P. gingivalis F0566	AWVD00000000	HMPREF1989_	00946 to 00950	00305 to 00309	01686	-	-	-	-	-
P. gingivalis F0568	AWUU00000000	HMPREF1553_	02150 to 02154	00622 to 00627	02200	-	-	-	-	-
P. gingivalis F0569	AWUV00000000	HMPREF1554_	00379 to 00383	01958 to 01962	01931	-	-	-	-	-
P. gingivalis F0570	AWUW00000000	HMPREF1555_	01695 to 01699	00638 to 00642	00703	-	-	-	-	-
P. gingivalis HG66	CP007756	EG14_	02885 to 02905	04705 to 04725	00390	-	-	-	-	-
P. gingivalis JCVI SC001	CM001843	A343_	2068 to 2072	2224 to 2228	0931	-	-	-	-	-

	Accession number	locus_tag prefix	LOCUS 1	LOCUS 2	LOCUS 3	LOCUS 4	LOCUS 5	LOCUS 6	LOCUS 7	LOCUS 8
P. gingivalis KCOM 279	NHRU00000000	CBG53_	06025 to 06045	07595 to 07615	04755	-	-	-	-	-
P. gingivalis MP4-504	LOEL00000000	AT291_	05660 to 05680	05230 to 05250	01865	-	-	-	-	-
P. gingivalis SJD11	ASYO00000000	SJDPG11_	04110 to 04130	00755 to 00775	03950	-	-	-	-	-
P. gingivalis SJD12	ASYP00000000	SJDPG12_	04120 to 04140	01275 to 01295	04100	-	-	-	-	-
P. gingivalis SJD2	ASYL00000000	SJDPG2_	04650 to 04670	01265 to 01285	06325	-	-	-	-	-
P. gingivalis SJD4	ASYM00000000	SJDPG4_	02490 to 02510	02875 to 02895	07395	-	-	-	-	-
P. gingivalis SJD5	ASYN00000000	SJDPG5_	01730 to 01750	01215 to 01235	07225	-	-	-	-	-
P. gingivalis SU60/YH522	FUFI00000000	PGIN_YH522_	01746 to 01750	00929 to 00933	00088	-	-	-	-	-
P. gingivalis W4087	AWVE00000000	HMPREF1990_	01524 to 01528	01936 to 01939	01936	-	-	-	-	-
P. gingivalis W50	AJZS00000000	HMPREF1322_	0047 to 0051	1695 to 1700	0808	-	-	-	-	-
P. gingivalis W83	NC_002950	PG_	2132 to 2136	0177 to 0183	1881	-	-	-	-	-
P. gingivalis ATCC 33277	NC_010729	PGN_	0180 to 0185	0287 to 0291	1808	-	-	-	-	-
P. gingivalis TDC60	NC_015571	PGTDC60_	1286 to 1290	0450 to 0454	0141	-	-	-	-	-
P. gingivicanis JCM 15907	BAKX00000000	JCM15907DRAFT_	-	-	00453	-	-	-	-	-
P. gingivicanis COT-022 OH1391	JQZW00000000	HQ36_	-	-	03345	-	-	-	-	-
P. gulae COT-052 OH2857	JRFD01000000	HQ46_	07985 to 08005	00320 to 00335 (*IL)	04080	-	-	-	-	-
P. gulae COT-052 OH1355	JRAG00000000	HQ42_	08975 to 08995	07470 to 07485 (*IL)	08260	-	-	-	-	-
P. gulae COT-052 OH1451	JRAI00000000	HR08_	02475 to 02495	01640 to 01655 (*IL)	10540	-	-	-	-	-
P. gulae COT-052 OH2179	JRAJ00000000	HR09_	03160 to 03195	06390 to 06410	01945	-	-	-	-	-
P. gulae COT-052 OH2199	JRAE00000000	HR10_	02880 to 02900	fragment	02360	-	-	-	-	-
P. sp. COT-052 OH4946 *(gulae-like)*	JQZY01000000	HQ50_	00395 to 00415	fragment	02910	-	-	-	-	-
P. gulae COT-052 OH3439	JRAK00000000	HR15_	05525 to 05545	05835 to 05865	03625	-	-	-	-	-
P. gulae COT-052 OH3471	JRAQ00000000	HQ40_	01405 to 01425	01225 to 01240 (*IL)	10345	-	-	-	-	-
P. gulae COT-052 OH3498	JRAF01000000	HR16_	08440 to 08465	00005 to 00015 (*IL)	01335	-	-	-	-	-
P. gulae COT-052 OH3856	JRAT00000000	HQ49_	08570 to 08610	09020 to 09035 (*IL)	05395	-	-	-	-	-
P. gulae COT-052 OH4119	JRAL00000000	HR17_	00340 to 00360	08855 to 08870 (*IL)	03500	-	-	-	-	-
P. gulae DSM 15663	ARJN00000000	F452DRAFT_	00993 to 00999	locus 1 : 01161 to 01164 locus 2 : 02016 to 02019	00309	-	-	-	-	-
P. gulae OH3161B	JQJE00000000	HR13_	08085 to 08095	00045 to 00060 (*IL)	00630	-	-	-	-	-
P. levii DSM 23370	ARBX00000000	A3GGDRAFT_	00263 to 00267	-	01421	-	-	-	-	-
P. macacae COT-192 OH2631	JRFB01000000	HR11_	04230 to 04250	04875 to 04895	03565	-	-	-	-	-
P. macacae COT-192 OH2859	JRFA01000000	HQ47_	05755 to 05775	-	07890	-	-	-	-	-
P. macacae DSM 20710/JCM 13914	BAKQ01000000	A3GIDRAFT_	01152 to 01156	00500 to 00504	00647	-	-	-	-	-
P. macacae JCM 15984	BAKS01000000	JCM15984DRAFT_	00147 to 00152	00180 to 00186	02252	-	-	-	-	-
P. sp. KLE 1280 *(pasteri-like)*	JNOS01000000	HMPREF1121_	-	-	01658	-	-	-	-	-
P. sp. 279 str. F0450 *(pasteri-like)*	ALKJ01000000	HMPREF1323_	-	-	1150	-	-	-	-	-
P. somerae DSM 23386	AQVC00000000	A3GKDRAFT_	-	-	-	-	-	-	-	-
P. sp. HMSC077F02 *(somerae-like)*	LTSX01000000	HMPREF3027_	-	-	-	-	-	-	-	-
P. uenonis DSM 23387	BAJM01000000	L215_RS0	09465 to 107560	-	09265	100935 to 100945	107770 to 107790	-	-	-
P. uenonis 60-3	ACLR01000000	PORUE0001_	0550 to 0554	-	01094	1280 to 1282	1280 to 1282	-	-	-
P. sp. COT-239 OH1446	JRAO01000000	HQ37_	-	-	02910	-	-	-	-	-
P. sp. COT-290 OH860	JRAR01000000	HQ41_	-	-	08885	-	-	-	-	-
P. sp. COT-290 OH3588	JRFC01000000	HQ48_	-	-	05560	-	-	-	-	-

	Accession number	locus_tag prefix	**LOCUS 9**	**LOCUS 10**	**LOCUS 11**	**LOCUS 12**	**LOCUS 13**	**LOCUS 14**	**LOCUS 15**	**LOCUS 16**	**LOCUS 17**
P. asaccharolytica PR426713P-I	AENO00000000	HMPREF9294_	-	-	-	-	-	-	-	-	-
P. asaccharolytica DSM 20707	NC_015501	Poras_	-	-	-	-	-	-	-	-	-
P. bennonis DSM 23058/JCM 16335	AQWR00000000	B088DRAFT_	01059 to 01062	01449 to 01454	-	07520-*orf*	-	-	-	-	-
P. sp. HMSC065F10 (*bennonis-like*)	LTXH00000000	HMPREF2890_	-	-	-	00515 to 00516	-	-	-	-	-
P. cangingivalis COT-109 OH1379	JQJF00000000	HQ34_	-	-	03060 to 03095	06575-*orf*	-	-	-	-	-
P. cangingivalis COT-109 OH1386	JQJD00000000	HQ35_	-	-	-	09885 to 09890	-	-	-	-	-
P. cangingivalis JCM 15983	BAKR00000000	JCM15983DRAFT_	-	-	00526 to 00532	00714 to 00715	-	-	-	-	-
P. cangingivalis ATCC 700135	FUWL01000000	SAMN02745205_	-	-	00571 to 00576	00937 to 00938	-	-	-	-	-
P. canoris COT-108 OH1224	JQZX00000000	HQ29_	-	-	-	-	05110 to 05130	-	-	-	-
P. sp. COT-108 OH1349 (*canoris-like*)	JRAH00000000	JT26_	-	-	-	-	03010 to 03030	-	-	-	-
P. sp. COT-108 OH2963 (*canoris-like*)	JRAP00000000	HQ39_	-	-	-	-	03990 to 04010	-	-	-	-
P. canoris COT-108 OH2762	JQZV00000000	HQ43_	-	-	-	-	07950 to 07940	-	-	-	-
P. catoniae F0037	AMEQ00000000	HMPREF9134_	-	-	-	-	-	-	-	-	-
P. catoniae ATCC 51270	JDFF00000000	HMPREF0636_	-	-	-	-	-	-	-	-	-
P. circumdentaria ATCC 51356	FUXE00000000	SAMN02745171_	-	-	-	-	-	00578 to 00582	-	-	-
P. crevioricanis ATCC 55563	FUXH00000000	SAMN02745203_	-	-	-	-	-	-	-	-	-
P. crevioricanis COT-253 OH1447	JQJC00000000	HQ38_	-	-	-	-	-	-	-	-	-
P. crevioricanis COT-253 OH2125	JQJB00000000	HQ45_	-	-	-	-	-	-	-	-	-
P. crevioricanis JCM 13913	BAOV00000000	PORCAN_	-	-	-	-	-	-	-	-	-
P. crevioricanis JCM 15906	BAOU00000000	PORCRE_	-	-	-	-	-	-	-	-	-
P. endodontalis ATCC 35406	ACNN00000000	POREN0001_	-	-	-	-	-	-	0625 to 0632	-	-
P. gingivalis 11A	FUFE00000000	PGIN_11A_	-	-	-	-	-	-	-	-	-
P. gingivalis 15_9	FUGF00000000	PGIN_15-9_	-	-	-	-	-	-	-	-	-
P. gingivalis 3_3	FUFB00000000	PGIN_3-3_	-	-	-	-	-	-	-	-	-
P. gingivalis 381	CP012889	PGF_	-	-	-	-	-	-	-	-	-
P. gingivalis 3A1	FUFC00000000	PGIN_3A1_	-	-	-	-	-	-	-	-	-
P. gingivalis 7BTORR	FUFD00000000	PGIN_7BTORR_	-	-	-	-	-	-	-	-	-
P. gingivalis 84_3	FUFG00000000	PGIN_84-3_	-	-	-	-	-	-	-	-	-
P. gingivalis A7436	CP011995	PGA7_	-	-	-	-	-	-	-	-	-
P. gingivalis A7A1-28	CP013131	PGS_	-	-	-	-	-	-	-	-	-
P. gingivalis AFR5B1	FUFJ00000000	PGIN_AFR-5B1_	-	-	-	-	-	-	-	-	-
P. gingivalis AJW4	CP011996	PGJ_	-	-	-	-	-	-	-	-	-
P. gingivalis Ando	BCBV00000000	PGANDO_	-	-	-	-	-	-	-	-	-
P. gingivalis ATCC 49417	FUFH00000000	PGIN_ATCC49417_	-	-	-	-	-	-	-	-	-
P. gingivalis F0185	AWVC00000000	HMPREF1988_	-	-	-	-	-	-	-	-	-
P. gingivalis F0566	AWVD00000000	HMPREF1989_	-	-	-	-	-	-	-	-	-
P. gingivalis F0568	AWUU00000000	HMPREF1553_	-	-	-	-	-	-	-	-	-
P. gingivalis F0569	AWUV00000000	HMPREF1554_	-	-	-	-	-	-	-	-	-
P. gingivalis F0570	AWUW00000000	HMPREF1555_	-	-	-	-	-	-	-	-	-
P. gingivalis HG66	CP007756	EG14_	-	-	-	-	-	-	-	-	-
P. gingivalis JCVI SC001	CM001843	A343_	-	-	-	-	-	-	-	-	-

	Accession number	locus_tag prefix	LOCUS 9	LOCUS 10	LOCUS 11	LOCUS 12	LOCUS 13	LOCUS 14	LOCUS 15	LOCUS 16	LOCUS 17
P. gingivalis KCOM 279	NHRU00000000	CBG53_	-	-	-	-	-	-	-	-	-
P. gingivalis MP4-504	LOEL00000000	AT291_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD11	ASYO00000000	SJDPG11_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD12	ASYP00000000	SJDPG12_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD2	ASYL00000000	SJDPG2_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD4	ASYM00000000	SJDPG4_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD5	ASYN00000000	SJDPG5_	-	-	-	-	-	-	-	-	-
P. gingivalis SU60/YH522	FUFI00000000	PGIN_YH522_	-	-	-	-	-	-	-	-	-
P. gingivalis W4087	AWVE00000000	HMPREF1990_	-	-	-	-	-	-	-	-	-
P. gingivalis W50	AJZS00000000	HMPREF1322_	-	-	-	-	-	-	-	-	-
P. gingivalis W83	NC_002950	PG_	-	-	-	-	-	-	-	-	-
P. gingivalis ATCC 33277	NC_010729	PGN_	-	-	-	-	-	-	-	-	-
P. gingivalis TDC60	NC_015571	PGTDC60_	-	-	-	-	-	-	-	-	-
P. gingivicanis JCM 15907	BAKX00000000	JCM15907DRAFT_	-	-	-	-	-	00570 to 00573	-	-	-
P. gingivicanis COT-022 OH1391	JQZW00000000	HQ36_	-	-	-	-	-	01490 to 01475	-	-	-
P. gulae COT-052 OH2857	JRFD01000000	HQ46_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH1355	JRAG00000000	HQ42_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH1451	JRAI00000000	HR08_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH2179	JRAJ00000000	HR09_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH2199	JRAE00000000	HR10_	-	-	-	-	-	-	-	-	-
P. sp. COT-052 OH4946 *(gulae-like)*	JQZY01000000	HQ50_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3439	JRAK00000000	HR15_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3471	JRAQ00000000	HQ40_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3498	JRAF01000000	HR16_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3856	JRAT00000000	HQ49_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH4119	JRAL00000000	HR17_	-	-	-	-	-	-	-	-	-
P. gulae DSM 15663	ARJN00000000	F452DRAFT_	-	-	-	-	-	-	-	-	-
P. gulae OH3161B	JQJE00000000	HR13_	-	-	-	-	-	-	-	-	-
P. levii DSM 23370	ARBX00000000	A3GGDRAFT_	-	-	-	-	-	-	-	01447 to 01452	02005 to 02010
P. macacae COT-192 OH2631	JRFB01000000	HR11_	-	-	-	-	-	-	-	-	-
P. macacae COT-192 OH2859	JRFA01000000	HQ47_	-	-	-	-	-	-	-	-	-
P. macacae DSM 20710/JCM 13914	BAKQ01000000	A3GIDRAFT_	-	-	-	-	-	-	-	-	-
P. macacae JCM 15984	BAKS01000000	JCM15984DRAFT_	-	-	-	-	-	-	-	-	-
P. sp. KLE 1280 *(pasteri-like)*	JNOS01000000	HMPREF1121_	-	-	-	-	-	-	-	-	-
P. sp. 279 str. F0450 *(pasteri-like)*	ALKJ01000000	HMPREF1323_	-	-	-	-	-	-	-	-	-
P. somerae DSM 23386	AQVC00000000	A3GKDRAFT_	-	-	-	-	-	-	-	-	-
P. sp. HMSC077F02 *(somerae-like)*	LTSX01000000	HMPREF3027_	-	-	-	-	-	-	-	-	-
P. uenonis DSM 23387	BAJM01000000	L215_RS0	-	-	-	-	-	-	-	-	-
P. uenonis 60-3	ACLR01000000	PORUE0001_	-	-	-	-	-	-	-	-	-
P. sp. COT-239 OH1446	JRAO01000000	HQ37_	-	-	-	-	-	-	-	-	-
P. sp. COT-290 OH860	JRAR01000000	HQ41_	-	-	-	-	-	-	-	-	-
P. sp. COT-290 OH3588	JRFC01000000	HQ48_	-	-	-	-	-	-	-	-	-

	Accession number	locus_tag prefix	**LOCUS 18**	**LOCUS 19**	**LOCUS 20**	**LOCUS 21**	**LOCUS 22**	**LOCUS 23**	**LOCUS 24**	**LOCUS 25**	**LOCUS 26**
P. asaccharolytica PR426713P-I	AENO00000000	HMPREF9294_	-	-	-	-	-	-	1596 to 1602	-	-
P. asaccharolytica DSM 20707	NC_015501	Poras_	-	-	-	-	-	-	0283 to 0289	-	-
P. bennonis DSM 23058/JCM 16335	AQWR00000000	B088DRAFT_	-	-	-	-	-	-	-	-	00223
P. sp. HMSC065F10 (*bennonis-like*)	LTXH00000000	HMPREF2890_	-	-	-	-	-	-	-	-	06690
P. cangingivalis COT-109 OH1379	JQJF00000000	HQ34_	-	-	-	-	-	03060 to 03095	-	-	-
P. cangingivalis COT-109 OH1386	JQJD00000000	HQ35_	-	-	-	-	-	-	-	-	-
P. cangingivalis JCM 15983	BAKR00000000	JCM15983DRAFT_	-	-	-	-	-	00525 to 00532	-	-	-
P. cangingivalis ATCC 700135	FUWL01000000	SAMN02745205_	-	-	-	-	-	00571 to 00577	-	-	-
P. canoris COT-108 OH1224	JQZX00000000	HQ29_	-	-	-	-	-	-	-	-	-
P. sp. COT-108 OH1349 (c*anoris-like*)	JRAH00000000	JT26_	-	-	-	-	-	-	-	-	-
P. sp. COT-108 OH2963 (c*anoris-like*)	JRAP00000000	HQ39_	-	-	-	-	-	-	-	-	-
P. canoris COT-108 OH2762	JQZV00000000	HQ43_	-	-	-	-	-	-	-	-	-
P. catoniae F0037	AMEQ00000000	HMPREF9134_	-	-	-	-	-	-	-	-	-
P. catoniae ATCC 51270	JDFF00000000	HMPREF0636_	-	-	-	-	-	-	-	-	-
P. circumdentaria ATCC 51356	FUXE00000000	SAMN02745171_	-	-	-	-	00600	-	-	-	-
P. crevioricanis ATCC 55563	FUXH00000000	SAMN02745203_	-	-	-	-	-	-	-	-	-
P. crevioricanis COT-253 OH1447	JQJC00000000	HQ38_	-	-	-	-	-	-	-	-	-
P. crevioricanis COT-253 OH2125	JQJB00000000	HQ45_	-	-	-	-	-	-	-	-	-
P. crevioricanis JCM 13913	BAOV00000000	PORCAN_	-	-	-	-	-	-	-	-	-
P. crevioricanis JCM 15906	BAOU00000000	PORCRE_	-	-	-	-	-	-	-	-	-
P. endodontalis ATCC 35406	ACNN00000000	POREN0001_	-	-	-	-	-	-	-	-	-
P. gingivalis 11A	FUFE00000000	PGIN_11A_	-	-	-	-	-	-	-	-	-
P. gingivalis 15_9	FUGF00000000	PGIN_15-9_	-	-	-	-	-	-	-	-	-
P. gingivalis 3_3	FUFB00000000	PGIN_3-3_	-	-	-	-	-	-	-	-	-
P. gingivalis 381	CP012889	PGF_	-	-	-	-	-	-	-	-	-
P. gingivalis 3A1	FUFC00000000	PGIN_3A1_	-	-	-	-	-	-	-	-	-
P. gingivalis 7BTORR	FUFD00000000	PGIN_7BTORR_	-	-	-	-	-	-	-	-	-
P. gingivalis 84_3	FUFG00000000	PGIN_84-3_	-	-	-	-	-	-	-	-	-
P. gingivalis A7436	CP011995	PGA7_	-	-	-	-	-	-	-	-	-
P. gingivalis A7A1-28	CP013131	PGS_	-	-	-	-	-	-	-	-	-
P. gingivalis AFR5B1	FUFJ00000000	PGIN_AFR-5B1_	-	-	-	-	-	-	-	-	-
P. gingivalis AJW4	CP011996	PGJ_	-	-	-	-	-	-	-	-	-
P. gingivalis Ando	BCBV00000000	PGANDO_	-	-	-	-	-	-	-	-	-
P. gingivalis ATCC 49417	FUFH00000000	PGIN_ATCC49417_	-	-	-	-	-	-	-	-	-
P. gingivalis F0185	AWVC00000000	HMPREF1988_	-	-	-	-	-	-	-	-	-
P. gingivalis F0566	AWVD00000000	HMPREF1989_	-	-	-	-	-	-	-	-	-
P. gingivalis F0568	AWUU00000000	HMPREF1553_	-	-	-	-	-	-	-	-	-
P. gingivalis F0569	AWUV00000000	HMPREF1554_	-	-	-	-	-	-	-	-	-
P. gingivalis F0570	AWUW00000000	HMPREF1555_	-	-	-	-	-	-	-	-	-
P. gingivalis HG66	CP007756	EG14_	-	-	-	-	-	-	-	-	-
P. gingivalis JCVI SC001	CM001843	A343_	-	-	-	-	-	-	-	-	-

	Accession number	locus_tag prefix	LOCUS 18	LOCUS 19	LOCUS 20	LOCUS 21	LOCUS 22	LOCUS 23	LOCUS 24	LOCUS 25	LOCUS 26
P. gingivalis KCOM 279	NHRU00000000	CBG53_	-	-	-	-	-	-	-	-	-
P. gingivalis MP4-504	LOEL00000000	AT291_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD11	ASYO00000000	SJDPG11_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD12	ASYP00000000	SJDPG12_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD2	ASYL00000000	SJDPG2_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD4	ASYM00000000	SJDPG4_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD5	ASYN00000000	SJDPG5_	-	-	-	-	-	-	-	-	-
P. gingivalis SU60/YH522	FUFI00000000	PGIN_YH522_	-	-	-	-	-	-	-	-	-
P. gingivalis W4087	AWVE00000000	HMPREF1990_	-	-	-	-	-	-	-	-	-
P. gingivalis W50	AJZS00000000	HMPREF1322_	-	-	-	-	-	-	-	-	-
P. gingivalis W83	NC_002950	PG_	-	-	-	-	-	-	-	-	-
P. gingivalis ATCC 33277	NC_010729	PGN_	-	-	-	-	-	-	-	-	-
P. gingivalis TDC60	NC_015571	PGTDC60_	-	-	-	-	-	-	-	-	-
P. gingivicanis JCM 15907	BAKX00000000	JCM15907DRAFT_	-	-	-	-	00572	-	-	-	-
P. gingivicanis COT-022 OH1391	JQZW00000000	HQ36_	-	-	-	-	08780	-	-	-	-
P. gulae COT-052 OH2857	JRFD01000000	HQ46_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH1355	JRAG00000000	HQ42_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH1451	JRAI00000000	HR08_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH2179	JRAJ00000000	HR09_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH2199	JRAE00000000	HR10_	-	-	-	-	-	-	-	-	-
P. sp. COT-052 OH4946 *(gulae-like)*	JQZY01000000	HQ50_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3439	JRAK00000000	HR15_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3471	JRAQ00000000	HQ40_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3498	JRAF01000000	HR16_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3856	JRAT00000000	HQ49_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH4119	JRAL00000000	HR17_	-	-	-	-	-	-	-	-	-
P. gulae DSM 15663	ARJN00000000	F452DRAFT_	-	-	-	-	-	-	-	-	-
P. gulae OH3161B	JQJE00000000	HR13_	-	-	-	-	-	-	-	-	-
P. levii DSM 23370	ARBX00000000	A3GGDRAFT_	-	-	-	-	-	-	-	-	-
P. macacae COT-192 OH2631	JRFB01000000	HR11_	-	-	-	-	-	-	-	-	-
P. macacae COT-192 OH2859	JRFA01000000	HQ47_	-	-	-	-	-	-	-	-	-
P. macacae DSM 20710/JCM 13914	BAKQ01000000	A3GIDRAFT_	-	-	-	-	-	-	-	-	-
P. macacae JCM 15984	BAKS01000000	JCM15984DRAFT_	-	-	-	-	-	-	-	-	-
P. sp. KLE 1280 *(pasteri-like)*	JNOS01000000	HMPREF1121_	-	-	-	-	-	-	-	-	-
P. sp. 279 str. F0450 *(pasteri-like)*	ALKJ01000000	HMPREF1323_	-	-	-	-	-	-	-	-	-
P. somerae DSM 23386	AQVC00000000	A3GKDRAFT_	101600 to 101620	101765 to 101780	101230	11025 to 11030	-	-	-	-	-
P. sp. HMSC077F02 *(somerae-like)*	LTSX01000000	HMPREF3027_	-	-	01845	06535 to 06545	-	-	-	-	-
P. uenonis DSM 23387	BAJM01000000	L215_RS0	-	-	-	-	-	-	102995 to 103025	104960 to 104980	-
P. uenonis 60-3	ACLR01000000	PORUE0001_	-	-	-	-	-	-	-	-	-
P. sp. COT-239 OH1446	JRAO01000000	HQ37_	-	-	-	-	-	-	-	-	-
P. sp. COT-290 OH860	JRAR01000000	HQ41_	-	-	-	-	-	-	-	-	-
P. sp. COT-290 OH3588	JRFC01000000	HQ48_	-	-	-	-	-	-	-	-	-

	Accession number	locus_tag prefix	LOCUS 27	LOCUS 28	LOCUS 29	LOCUS 30	LOCUS 31	LOCUS 32	LOCUS 33	LOCUS 34	LOCUS 35
P. asaccharolytica PR426713P-I	AENO00000000	HMPREF9294_	-	-	-	-	-	-	-	-	-
P. asaccharolytica DSM 20707	NC_015501	Poras_	-	-	-	-	-	-	-	-	-
P. bennonis DSM 23058/JCM 16335	AQWR00000000	B088DRAFT_	-	-	-	-	-	-	-	-	-
P. sp. HMSC065F10 (*bennonis-like*)	LTXH00000000	HMPREF2890_	-	-	-	-	-	-	-	-	-
P. cangingivalis COT-109 OH1379	JQJF00000000	HQ34_	-	-	-	-	-	-	-	-	-
P. cangingivalis COT-109 OH1386	JQJD00000000	HQ35_	-	-	-	-	-	-	-	-	-
P. cangingivalis JCM 15983	BAKR00000000	JCM15983DRAFT_	-	-	-	-	-	-	-	-	-
P. cangingivalis ATCC 700135	FUWL01000000	SAMN02745205_	-	-	-	-	-	-	-	-	-
P. canoris COT-108 OH1224	JQZX00000000	HQ29_	-	*orf*-02220	-	-	-	-	-	-	-
P. sp. COT-108 OH1349 (c*anoris-like*)	JRAH00000000	JT26_	-	*orf*-05010	-	-	-	-	-	-	-
P. sp. COT-108 OH2963 (c*anoris-like*)	JRAP00000000	HQ39_	-	*orf*-02320	-	-	-	-	-	-	-
P. canoris COT-108 OH2762	JQZV00000000	HQ43_	-	*orf*-03215	-	-	-	-	-	-	-
P. catoniae F0037	AMEQ00000000	HMPREF9134_	-	-	-	-	-	-	-	-	-
P. catoniae ATCC 51270	JDFF00000000	HMPREF0636_	-	-	-	-	-	-	-	-	-
P. circumdentaria ATCC 51356	FUXE00000000	SAMN02745171_	-	-	-	-	-	-	-	-	-
P. crevioricanis ATCC 55563	FUXH00000000	SAMN02745203_	-	00140 to 00141	-	-	00297 to 00299	-	-	-	-
P. crevioricanis COT-253 OH1447	JQJC00000000	HQ38_	-	04780 to 04785	03905 to 03915	00720 to 00735	07595 to 07610	-	-	-	-
P. crevioricanis COT-253 OH2125	JQJB00000000	HQ45_	-	02515 to 02520	-	-	06265 to 06275	-	-	-	-
P. crevioricanis JCM 13913	BAOV00000000	PORCAN_	-	298 to 299	-	-	176 to 178	-	-	-	-
P. crevioricanis JCM 15906	BAOU00000000	PORCRE_	-	2060 to 2061	-	-	1398 to 1400	-	-	-	-
P. endodontalis ATCC 35406	ACNN00000000	POREN0001_	-	-	-	-	-	-	-	-	-
P. gingivalis 11A	FUFE00000000	PGIN_11A_	-	-	-	-	-	-	-	-	-
P. gingivalis 15_9	FUGF00000000	PGIN_15-9_	-	-	-	-	-	-	-	-	-
P. gingivalis 3_3	FUFB00000000	PGIN_3-3_	-	-	-	-	-	-	-	-	-
P. gingivalis 381	CP012889	PGF_	-	-	-	-	-	-	-	-	-
P. gingivalis 3A1	FUFC00000000	PGIN_3A1_	-	-	-	-	-	-	-	-	-
P. gingivalis 7BTORR	FUFD00000000	PGIN_7BTORR_	-	-	-	-	-	-	-	-	-
P. gingivalis 84_3	FUFG00000000	PGIN_84-3_	-	-	-	-	-	-	-	-	-
P. gingivalis A7436	CP011995	PGA7_	-	-	-	-	-	-	-	-	-
P. gingivalis A7A1-28	CP013131	PGS_	-	-	-	-	-	-	-	-	-
P. gingivalis AFR5B1	FUFJ00000000	PGIN_AFR-5B1_	-	-	-	-	-	-	-	-	-
P. gingivalis AJW4	CP011996	PGJ_	-	-	-	-	-	-	-	-	-
P. gingivalis Ando	BCBV00000000	PGANDO_	-	-	-	-	-	-	-	-	-
P. gingivalis ATCC 49417	FUFH00000000	PGIN_ATCC49417_	-	-	-	-	-	-	-	-	-
P. gingivalis F0185	AWVC00000000	HMPREF1988_	-	-	-	-	-	-	-	-	-
P. gingivalis F0566	AWVD00000000	HMPREF1989_	-	-	-	-	-	-	-	-	-
P. gingivalis F0568	AWUU00000000	HMPREF1553_	-	-	-	-	-	-	-	-	-
P. gingivalis F0569	AWUV00000000	HMPREF1554_	-	-	-	-	-	-	-	-	-
P. gingivalis F0570	AWUW00000000	HMPREF1555_	-	-	-	-	-	-	-	-	-
P. gingivalis HG66	CP007756	EG14_	-	-	-	-	-	-	-	-	-
P. gingivalis JCVI SC001	CM001843	A343_	-	-	-	-	-	-	-	-	-

	Accession number	locus_tag prefix	LOCUS 27	LOCUS 28	LOCUS 29	LOCUS 30	LOCUS 31	LOCUS 32	LOCUS 33	LOCUS 34	LOCUS 35
P. gingivalis KCOM 279	NHRU00000000	CBG53_	-	-	-	-	-	-	-	-	-
P. gingivalis MP4-504	LOEL00000000	AT291_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD11	ASYO00000000	SJDPG11_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD12	ASYP00000000	SJDPG12_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD2	ASYL00000000	SJDPG2_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD4	ASYM00000000	SJDPG4_	-	-	-	-	-	-	-	-	-
P. gingivalis SJD5	ASYN00000000	SJDPG5_	-	-	-	-	-	-	-	-	-
P. gingivalis SU60/YH522	FUFI00000000	PGIN_YH522_	-	-	-	-	-	-	-	-	-
P. gingivalis W4087	AWVE00000000	HMPREF1990_	-	-	-	-	-	-	-	-	-
P. gingivalis W50	AJZS00000000	HMPREF1322_	-	-	-	-	-	-	-	-	-
P. gingivalis W83	NC_002950	PG_	-	-	-	-	-	-	-	-	-
P. gingivalis ATCC 33277	NC_010729	PGN_	-	-	-	-	-	-	-	-	-
P. gingivalis TDC60	NC_015571	PGTDC60_	-	-	-	-	-	-	-	-	-
P. gingivicanis JCM 15907	BAKX00000000	JCM15907DRAFT_	-	-	-	-	-	-	-	-	-
P. gingivicanis COT-022 OH1391	JQZW00000000	HQ36_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH2857	JRFD01000000	HQ46_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH1355	JRAG00000000	HQ42_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH1451	JRAI00000000	HR08_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH2179	JRAJ00000000	HR09_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH2199	JRAE00000000	HR10_	-	-	-	-	-	-	-	-	-
P. sp. COT-052 OH4946 *(gulae-like)*	JQZY01000000	HQ50_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3439	JRAK00000000	HR15_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3471	JRAQ00000000	HQ40_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3498	JRAF01000000	HR16_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH3856	JRAT00000000	HQ49_	-	-	-	-	-	-	-	-	-
P. gulae COT-052 OH4119	JRAL00000000	HR17_	-	-	-	-	-	-	-	-	-
P. gulae DSM 15663	ARJN00000000	F452DRAFT_	-	-	-	-	-	-	-	-	-
P. gulae OH3161B	JQJE00000000	HR13_	-	-	-	-	-	-	-	-	-
P. levii DSM 23370	ARBX00000000	A3GGDRAFT_	-	-	-	-	-	-	00131	-	-
P. macacae COT-192 OH2631	JRFB01000000	HR11_	-	-	-	-	-	-	-	01165 to 01175 + *2 orfs*	-
P. macacae COT-192 OH2859	JRFA01000000	HQ47_	-	-	-	-	-	-	-	00710 + orf to 00720+2orfs	-
P. macacae DSM 20710/JCM 13914	BAKQ01000000	A3GIDRAFT_	-	-	-	-	-	-	-	01647 to 01650	-
P. macacae JCM 15984	BAKS01000000	JCM15984DRAFT_	-	-	-	-	-	-	-	01339 to 01345	01188 to 01193
P. sp. KLE 1280 *(pasteri-like)*	JNOS01000000	HMPREF1121_	-	-	-	-	-	-	-	-	-
P. sp. 279 str. F0450 *(pasteri-like)*	ALKJ01000000	HMPREF1323_	-	-	-	-	-	-	-	-	-
P. somerae DSM 23386	AQVC00000000	A3GKDRAFT_	-	-	-	-	-	-	-	-	-
P. sp. HMSC077F02 *(somerae-like)*	LTSX01000000	HMPREF3027_	-	-	-	-	-	-	-	-	-
P. uenonis DSM 23387	BAJM01000000	L215_RS0	108160 to 108170	-	-	-	-	09720 to 101135	-	-	-
P. uenonis 60-3	ACLR01000000	PORUE0001_	1566 to 1569	-	-	-	-	0299 to 0301	-	-	-
P. sp. COT-239 OH1446	JRAO01000000	HQ37_	-	-	-	-	-	-	-	-	-
P. sp. COT-290 OH860	JRAR01000000	HQ41_	-	-	-	-	-	-	-	-	-
P. sp. COT-290 OH3588	JRFC01000000	HQ48_	-	-	-	-	-	-	-	-	-

Supplementary Table 2. Metadata associated to all genomes used in this study.

	HOST	ISOLATION	SOURCE	HEALTH	DISEASE
PORPHY_ASA					
Porphyromonas asaccharolytica PR426713P-I		Vagina	-	+	+
Porphyromonas asaccharolytica DSM 20707		Lung/Pleura	Soft tissue infection	+	+
PORPHY_BEN					
Porphyromonas bennonis DSM 23058		Skin/Shoulder	Soft tissue infection	+	+
Porphyromonas sp. HMSC065F10		Skin/Shoulder	Soft tissue infection	+	+
PORPHY_CGI					
Porphyromonas cangingivalis COT-109 OH1379		Oral	Gingivitis	+++++	++
Porphyromonas cangingivalis COT-109 OH1386		Oral	Gingivitis	+++++	++
Porphyromonas cangingivalis JCM 15983		Oral	Periodontal pocket	+++++	++
Porphyromonas cangingivalis ATCC 700135		Oral	Periodontal pocket	+++++	++
PORPHY_CAN					
Porphyromonas canoris COT-108 OH1224		Oral	Periodontitis 1	++	++
Porphyromonas sp. COT-108 OH1349		Oral	Gingivitis	++	++
Porphyromonas sp. COT-108 OH2963		Oral	Periodontitis 1	++	++
Porphyromonas canoris COT-108 OH2762		Oral	Gingivitis	++	++
PORPHY_CAT					
Porphyromonas catoniae F0037		Oral	Gingival crevice	+++++	-
Porphyromonas catoniae ATCC 51270		Oral	Gingival crevice	+++++	-
PORPHY_CIR					
Porphyromonas circumdentaria ATCC 51356		Oral	Soft tissue infection	+	+
PORPHY_CRE					
Porphyromonas crevioricanis ATCC 55563		Oral	Periodontal pocket	-	+++
Porphyromonas crevioricanis COT-253 OH1447		Oral	Periodontitis 1	-	+++
Porphyromonas crevioricanis COT-253 OH2125		Oral	Gingivitis	-	+++
Porphyromonas crevioricanis JCM 13913		Oral	Gingivitis	-	+++
Porphyromonas crevioricanis JCM 15906		Oral	Periodontal pocket	-	+++
PORPHY_END					
Porphyromonas endodontalis ATCC 35406		Oral	Infected root canal	+	+
PORPHY_GIN					
Porphyromonas gingivalis 11A		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis 15_9		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis 3_3		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis 381		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis 3A1		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis 7BTORR		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis 84_3		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis A7436		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis A7A1_28		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis AFR5B1		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis AJW4		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis Ando		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis ATCC 49417		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis F0185		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis F0566		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis F0568		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis F0569		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis F0570		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis HG66		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis ICVI_SC001		Oral	Periodontitis	+++	+++++

	HOST	ISOLATION	SOURCE	HEALTH	DISEASE
Porphyromonas gingivalis KCOM 279		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis MP4-504		Oral	Periodontitis	+++	
Porphyromonas gingivalis SJD11		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis SJD12		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis SJD2		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis SJD4		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis SJD5		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis SU60		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis W4087		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis W50		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis W83		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis ATCC 33277		Oral	Periodontitis	+++	+++++
Porphyromonas gingivalis TDC60		Oral	Periodontitis	+++	+++++
PORPHY_GGI					
Porphyromonas gingivicanis JCM 15907		Oral	Gingival crevicular fluid	++	-
Porphyromonas gingivicanis COT-022 OH1391		Oral	Gingival crevicular fluid	++	-
PORPHY_GUL					
Porphyromonas gulae COT-052 OH1355		Oral	Gingivitis	+++	+++++
Porphyromonas gulae COT-052 OH1451		Oral	Periodontitis	+++	+++++
Porphyromonas gulae COT-052 OH2179		Oral	Periodontitis	+++	+++++
Porphyromonas gulae COT-052 OH2199		Oral	Periodontitis	+++	+++++
Porphyromonas sp. COT-052 OH4946		Oral	Periodontitis	+++	+++++
Porphyromonas gulae COT-052 OH3439		Oral	Periodontitis	+++	+++++
Porphyromonas gulae COT-052 OH3471		Oral	Health	+++	+++++
Porphyromonas gulae COT-052 OH3498		Oral	Periodontitis	+++	+++++
Porphyromonas gulae COT-052 OH3856		Oral	Periodontitis	+++	+++++
Porphyromonas gulae COT-052 OH4119		Oral	Periodontitis	+++	+++++
Porphyromonas gulae DSM 15663		Oral	Periodontitis	+++	+++++
Porphyromonas gulae OH3161B		Oral	Periodontitis	+++	+++++
PORPHY_LEV					
Porphyromonas levii DSM 23370		Gut	Rumen	+	+
PORPHY_MAC					
Porphyromonas macacae COT-192 OH2631		Oral	Periodontitis	+++	+++++
Porphyromonas macacae COT-192 OH2859		Oral	Periodontitis	+++	+++++
P. macacae DSM 20710 = JCM 13914		Oral	Periodontitis	+++	+++++
Porphyromonas macacae JCM 15984		Oral	Periodontitis	+++	+++++
PORPHY_PAS					
Porphyromonas sp. KLE 1280		Oral	Saliva	+++++	-
Porphyromonas sp. oral taxon 279 str. F0450		Oral	Saliva	+++++	-
PORPHY_SOM					
Porphyromonas somerae DSM 23386		Skin/bone	Soft tissue infection	+	+
Porphyromonas sp. HMSC077F02		Vagina	-	+	+
PORPHY_UEN					
Porphyromonas uenonis DSM 23387		Oral	Soft tissue infection	+	+
Porphyromonas uenonis 60-3		Vagina	-	+	+
***no species* PORPHY_COT**					
Porphyromonas sp. COT-239 OH1446		Oral	Periodontitis		
Porphyromonas sp. COT-290 OH860		Oral	Gingivitis		
Porphyromonas sp. COT-290 OH3588		Oral	-		

3. ARTICLE 3: Signature de la dysbiose bactérienne lors de la parodontite

Signature of Microbial Dysbiosis in Periodontitis

Publié dans Applied and Envionmental Microbiology

Le dernier article présenté dans ce document de thèse a été publié dans la revue Applied and Environmental Microbiology en mai 2017. Il s'agit d'un article de recherche qui utilise des données propres à l'équipe rennaise dans laquelle j'ai commencé mon doctorat et surtout des données publiques et disponibles dans des bases de données internationales comme le *Sequence Read Archive* (SRA) du NCBI, du serveur d'études métagénomiques comme *MetaGenomics Rapid Annotation using Subsystems Technology* (MG-RAST) de l'University of Chicago et de l'Argonne National Laboratory, et du *Human Microbiome Project* (HMP) du *National Institutes of Health* (NIH) des États-Unis.

L'objectif général de l'article est d'utiliser ces données hétérogènes pour décrire une communauté caractéristique de poches parodontales lors d'une parodontite chronique et de la différencier de la communauté bactérienne du sulcus gingivo-dentaire de la bouche saine. La description des communautés bactériennes, qui par abus de langage sont appelées microbiotes, est faite en dénombrant le nombre de bactéries présentes et en assignant chacune à un *Operational Taxonomic Unit* (OTU). Idéalement, chaque OTU devrait correspondre à une espèce. Cependant le niveau de résolution du marqueur choisi est inférieur à ce niveau taxonomique et la description la plus précise obtenue, dans le meilleur des cas, est le genre bactérien. Ces marqueurs sont traditionnellement des régions variables du gène codant l'ARN ribosomique 16S. Ce gène d'environ 1500 bp est trop grand pour être séquencé entièrement en SGS mais rempli les critères d'un bon *code-barre d'ADN* (ou *DNA barcode*) puisqu'il

contient des séquences variables encadrées par deux zones très conservées permettant son amplification par PCR.

L'hétérogénéité des données utilisées dans cet article vient des différentes méthodes utilisés pour obtenir les reads de séquençage : le protocole d'extraction d'ADN (qui pourrait ne pas être assez fort pour lyser toutes les cellules bactériennes et introduire un biais du fait de la libération de l'ADN des cellules faciles à lyser uniquement), des conditions d'amplification de la région choisie (production de chimères, amplification non spécifique), des amorces utilisées, de la région choisie (régions hypervariables de V1 à V9) et de l'analyse bioinformatique utilisée pour identifier les OTU. À cela s'ajoutent des considérations cliniques de définition des patients sains non atteints de parodontite, des patients atteints de parodontite et de collecte de l'échantillon qui sera ensuite utilisé pour extraire l'ADN de la communauté bactérienne.

Dans cette étude, l'hypothèse est que les différences entre les microbiotes des sillons sains et ceux issus de poches parodontales avec parodontite chronique sont plus importantes que celles dues aux variations techniques et qu'il est donc possible de déterminer une communauté caractéristique de sites avec parodontite, une "*signature de dysbiose*". Le nombre total d'échantillons de sites atteints est de près de 200 et de sites sains de plus du double avec près de 425 échantillons qui incluent des sulci sains et des échantillons contrôle externe : caries, plaque supragingivale et de la partie médiane du vagin. Nous avons utilisé une seule suite d'analyse pour définir les communautés, un pipeline nommé *Visualization and Analysis of Microbial Population Structure* (VAMPS) du Josephine Bay Paul Center et Marine Biological Laboratory de l'University of Chicago.

La première description de chaque échantillon correspond à la mesure de l'alpha-diversité (la diversité interne à l'échantillon : son nombre d'espèces et le poids relatif de chacune dans le total). Les échantillons issus de poches parodontales ont une alpha-diversité plus faible que ceux issus de sulci sains. La dysbiose produirait donc une communauté moins diverse, plus réduite en nombre d'espèces et/ou plus homogène. La deuxième mesure classique de la diversité correspond à la beta-diversité qui est une mesure de la différence de composition entre deux échantillons, plusieurs indices sont utilisés pour l'évaluer et dans l'article nous utilisons l'indice de Bray-Curtis. La formation de groupes basés sur cet indice (*clustering*) résulte en cinq groupes. En observant le statut des échantillons classés dans chacun des groupes, nous pouvons constater deux clusters qui regroupent, en majorité, des

échantillons de salive d'une part et de caries et vagin d'autre part. Ces deux groupes sont considérés comme les contrôles externes et leur regroupement entre eux était attendu et garantit la pertinence de l'analyse. Les deux clusters suivants regroupent majoritairement des échantillons de sites sains (supra et sub-gingivaux) et seulement un petit pourcentage des échantillons de parodontite (4% et 6% du total des échantillons de parodontite sont regroupés dans les clusters 1 et 2 respectivement). Le dernier cluster, le nº 5, regroupe 90% des échantillons de parodontite, ce qui correspond à près des deux tiers des échantillons classés dans ce groupe, le reste étant des poches parodontales peu profondes (*shallow*), des outgroup (caries et salive) et surtout des échantillons sains (sub ou supragingivaux) qui représentent près de un quart des échantillons du cluster 5.

Ces résultats de clustering malgré les différences techniques confortent notre hypothèse. Les microbiotes de sites avec parodontite sont plus similaires entre eux et peuvent être différentiés des sites sains, cela malgré les différentes méthodes de préparation des échantillons et de production des reads de séquençage. Le peu d'échantillons "*avec parodontite*" classés dans les groupes sains pourraient être des échantillons de sites en rémission. Le cas contraire, les échantillons sains dans le cluster "*à parodontite*" pourraient être des sites à risque élevé de développement de parodontite. Ces deux hypothèses restent à vérifier et nécessitent des études longitudinales de suivi de patients sains qui développent des parodontites et également des patients sous traitement et en rémission de leur parodontite.

La description macroscopique étant faite, l'étape suivante est la description des communautés bactériennes. Pour cela, une analyse au niveau du genre bactérien a été faite. Les genres présents dans 95% des échantillons de sulci sains et les genres présents dans 95% des échantillons de poches parodontales ont été identifiés. Les genres communs aux deux états ne sont pas utilisés pour la suite (e.g. *Fusobacterium*). Avec cette définition, les genres associés à un état sain sont *Veillonella*, *Neisseria*, *Rothia*, *Corynebacterium* et *Actinomyces* ; tandis que les genres associés à la maladie parodontale sont *Eubacterium*, *Campylobacter*, *Treponema* et *Tannerella*. La réduction de la diversité du microbiote à un ratio entre le nombre de reads des genres associées à la maladie sur ceux des genres associés aux sulci sains est capable de distinguer les échantillons sains des atteints de parodontite et même de pointer les échantillons sains qui sont regroupés avec les échantillons atteints et les échantillons avec parodontite qui forment un cluster avec les échantillons de sulci sains.

Ce ratio a été validé avec des échantillons d'une étude externe aux données utilisées précédemment. Une corrélation entre le statut de l'échantillon et le ratio a été observé ce qui conforte l'utilisation du ratio pour décrire la dysbiose caractéristique des maladies parodontales.

La description du microbiote au niveau de l'espèce n'est pas satisfaisante. Un genre bactérien peut avoir des espèces symbiontes et des espèces pathogènes. Puis, un genre peut être associé à d'autres pathologies que celle étudié particulièrement dans l'étude qui l'associe à l'état de santé. Les deux possibilités sont vraies dans cette étude. *Corynebacterium* est un genre divers avec plus de la moitié de ses espèces considérées comme non pathogènes, mais certaines sont hautement pathogènes. *Corynebacterium diphteriae* (agent de la diphtérie), *C. pseudotuberculosis* et *C. ulcerans* sont des espèces pathogènes reconnues. Si dans un microbiote une de ces espèces est détectée, le pronostique du patient est très différent de s'il s'agit d'une autre espèce non pathogène. Également, les espèces du genre *Neisseria* peuvent être pathogènes, il est à noter particulièrement *Neisseria meningitidis* et *N. gonorrhoeae* qui sont toutes les deux fortement pathogènes. Leur détection dans un sulcus gingival, même si le sulcus est sain de parodontite, peut signifier que la bactérie colonise la bouche de la personne et cela peut signifier une colonisation ultérieure de la gorge, étape précédant une méningite bactérienne (*N. meningitidis*), une infection locale ou pharyngite gonococcique (*N. gonorrhoeae*) ou diphtérie (*C. diphteriae*). Il est donc nécessaire de développer des stratégies de séquençage à haut débit de gènes complets de l'ARN 16 afin d'espérer identifier les bactéries jusqu'au niveau de l’espèce.

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

ENVIRONMENTAL MICROBIOLOGY

Signature of Microbial Dysbiosis in Periodontitis

Vincent Meuric,[a,b] Sandrine Le Gall-David,[b] Emile Boyer,[a,b] Luis Acuña-Amador,[c] Bénédicte Martin,[b] Shao Bing Fong,[b] Frederique Barloy-Hubler,[c] Martine Bonnaure-Mallet[a,b]

CHU Rennes, Pôle Odontologie, Rennes, France[a]; Université de Rennes 1, EA 1254, INSERM 1241, Equipe de Microbiologie, Rennes, France[b]; CNRS, UMR 6290, IGDR, Rennes, France[c]

ABSTRACT Periodontitis is driven by disproportionate host inflammatory immune responses induced by an imbalance in the composition of oral bacteria; this instigates microbial dysbiosis, along with failed resolution of the chronic destructive inflammation. The objectives of this study were to identify microbial signatures for health and chronic periodontitis at the genus level and to propose a model of dysbiosis, including the calculation of bacterial ratios. Published sequencing data obtained from several different studies (196 subgingival samples from patients with chronic periodontitis and 422 subgingival samples from healthy subjects) were pooled and subjected to a new microbiota analysis using the same Visualization and Analysis of Microbial Population Structures (VAMPS) pipeline, to identify microbiota specific to health and disease. Microbiota were visualized using CoNet and Cytoscape. Dysbiosis ratios, defined as the percentage of genera associated with disease relative to the percentage of genera associated with health, were calculated to distinguish disease from health. Correlations between the proposed dysbiosis ratio and the periodontal pocket depth were tested with a different set of data obtained from a recent study, to confirm the relevance of the ratio as a potential indicator of dysbiosis. Beta diversity showed significant clustering of periodontitis-associated microbiota, at the genus level, according to the clinical status and independent of the methods used. Specific genera (*Veillonella*, *Neisseria*, *Rothia*, *Corynebacterium*, and *Actinomyces*) were highly prevalent (>95%) in health, while other genera (*Eubacterium*, *Campylobacter*, *Treponema*, and *Tannerella*) were associated with chronic periodontitis. The calculation of dysbiosis ratios based on the relative abundance of the genera found in health versus periodontitis was tested. Nonperiodontitis samples were significantly identifiable by low ratios, compared to chronic periodontitis samples. When applied to a subgingival sample set with well-defined clinical data, the method showed a strong correlation between the dysbiosis ratio, as well as a simplified ratio (*Porphyromonas*, *Treponema*, and *Tannerella* to *Rothia* and *Corynebacterium*), and pocket depth. Microbial analysis of chronic periodontitis can be correlated with the pocket depth through specific signatures for microbial dysbiosis.

IMPORTANCE Defining microbiota typical of oral health or chronic periodontitis is difficult. The evaluation of periodontal disease is currently based on probing of the periodontal pocket. However, the status of pockets "on the mend" or sulci at risk of periodontitis cannot be addressed solely through pocket depth measurements or current microbiological tests available for practitioners. Thus, a more specific microbiological measure of dysbiosis could help in future diagnoses of periodontitis. In this work, data from different studies were pooled, to improve the accuracy of the results. However, analysis of multiple species from different studies intensified the bacterial network and complicated the search for reproducible microbial signatures. Despite the use of different methods in each study, investigation of the microbiota at the genus level showed that some genera were prevalent (up to 95% of the sam-

Received 23 February 2017 **Accepted** 2 May 2017

Accepted manuscript posted online 5 May 2017

Citation Meuric V, Le Gall-David S, Boyer E, Acuña-Amador L, Martin B, Fong SB, Barloy-Hubler F, Bonnaure-Mallet M. 2017. Signature of microbial dysbiosis in periodontitis. Appl Environ Microbiol 83:e00462-17. https://doi.org/10.1128/AEM.00462-17.

Editor Andrew J. McBain, University of Manchester

Address correspondence to Vincent Meuric, vincent.meuric@univ-rennes1.fr.

F.B.-H. and M.B.-M. contributed equally to this work.

ples) in health or disease, allowing the calculation of bacterial ratios (i.e., dysbiosis ratios). The correlation between the proposed ratios and the periodontal pocket depth was tested, which confirmed the link between dysbiosis ratios and the severity of the disease. The results of this work are promising, but longitudinal studies will be required to improve the ratios and to define the microbial signatures of the disease, which will allow monitoring of periodontal pocket recovery and, conceivably, determination of the potential risk of periodontitis among healthy patients.

KEYWORDS chronic periodontitis, health, microbiota, dysbiosis ratio

Chronic periodontitis (CP) is a type of chronic inflammation characterized by alveolar bone loss, with intermittent periods of remission and relapse. CP is currently considered an infection, mainly due to increases in bacteria in the sulcus, leading to the formation of a periodontal pocket (for review, see references 1 and 2). The major pathogen linked to CP is *Porphyromonas gingivalis*, with bacterial partners such as *Treponema denticola* and *Tannerella forsythia*. These three bacteria have been considered the major pathogenic "red complex" since 1998 (3). However, recent advances from metagenomic studies have developed a new model of periodontal disease pathogenesis. CP does not result from individual pathogens but rather from polymicrobial synergy and dysbiosis (4) associated with a dysregulated immune response inducing inflammation-mediated tissue damage (5). Host genetic components have also been implicated in CP, with multiple genes contributing cumulatively to the host's overall disease risk (or protection) through effects on the host immune response and the microbiome (6). Since the Human Microbiome Project (HMP) (7), microbiota have been analyzed based on partial sequencing of the 16S rRNA gene, with different numbers of healthy and CP samples. However, comparisons between studies are difficult because of the differences in the methods used (i.e., clinical examination and diagnosis of periodontitis and oral health, sample collection protocols, DNA extraction protocols, and analysis of hypervariable regions of the 16S rRNA gene). Because there is growing interest in the human microbiome, despite the difficulties mentioned earlier, the use of independent studies to look for "signal in the noise" should proceed as suggested previously (8), through reanalysis of all data with the same protocol. The difference between periodontal health-associated and disease-associated microbiota should be larger than the technical variations of the different studies, which would enable the identification of microbial signatures using next-generation sequencing (NGS) technologies. The first objective of this study was to explore the disease-associated changes in the subgingival microbiota at the genus level, using a unique Visualization and Analysis of Microbial Population Structure (VAMPS) pipeline (9), for beta diversity (Bray-Curtis dissimilarity) with a large number of samples (from 6 different studies) and to confirm that the microbiota identified did not cluster according to the methods used (primer or study type). Subgingival microbiota from patients with diagnosed chronic periodontitis (196 samples) and from healthy subjects (422 samples), as well as external control samples (from dentine caries, supragingival plaque, and the midvagina), were included. The second objective was to determine a dysbiosis ratio of bacteria that could predict health or disease severity from the subgingival samples and finally to test the ratio with an independent cohort of patients with well-described periodontal pocket measurements.

RESULTS

Microbial community structure analysis. Using a matrix correlation analysis, the possible clustering of microbiota according to the nature of the primers used, the site of sampling, or the study investigated was explored. Despite various studies, the analyzed data clustered into five groups according to the clinical status (healthy or CP) or the sampling site, as shown by the three-dimensional (3D) principal-coordinate analysis (PCoA) plots (Fig. 1). Healthy subgingival samples were primarily spread into two main clusters; control samples were clearly separated into two other clusters, corresponding to saliva and dentine caries/vagina, while the majority of CP samples

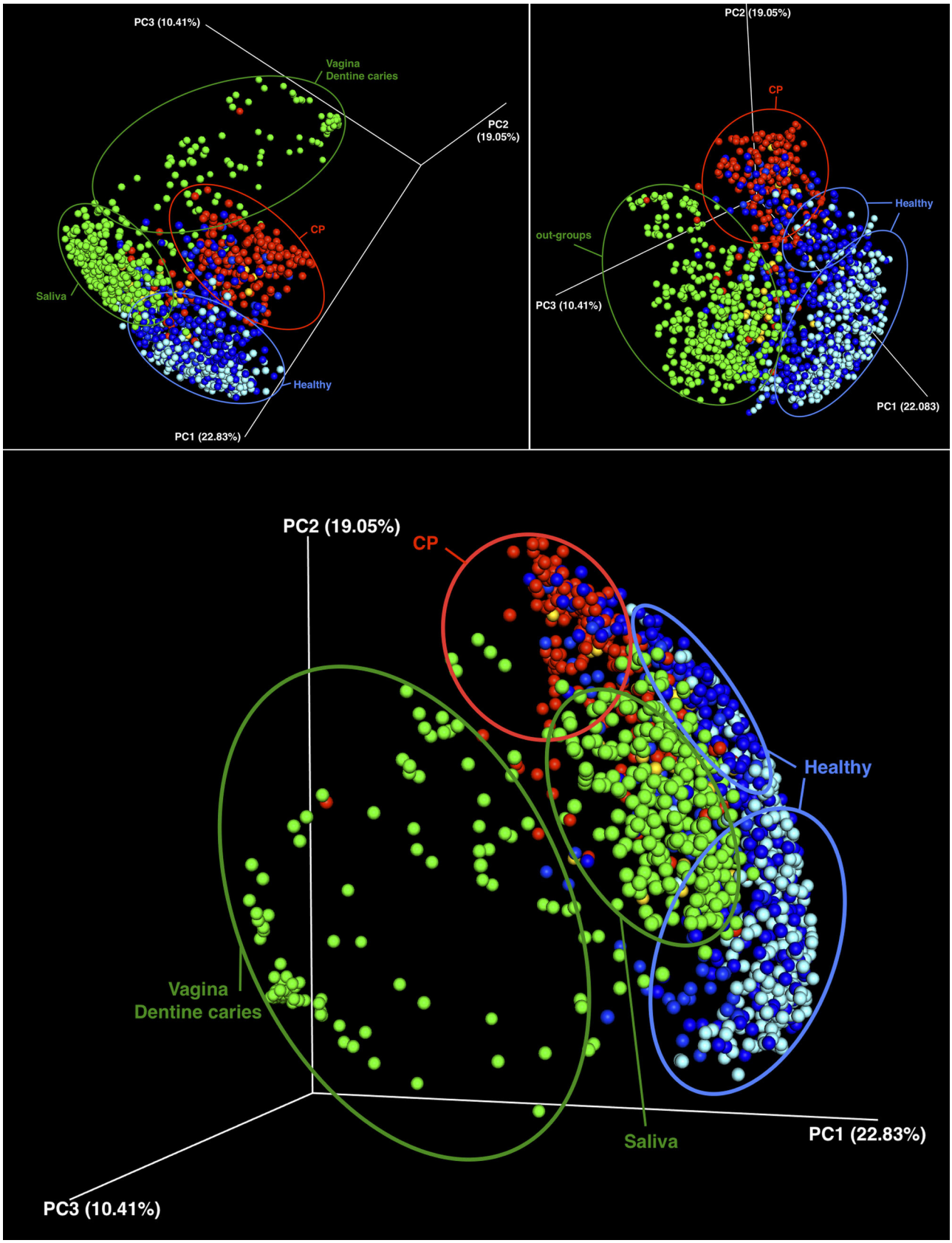


FIG 1 Different views of 3D PCoA plots illustrating the beta diversity of bacterial populations as a function of sampling site and diagnosis. Light blue, supragingival samples; dark blue, healthy subgingival samples; green, outgroups (saliva, midvagina, and dentine caries samples); red, CP samples. Percentages represent percent explained variance.

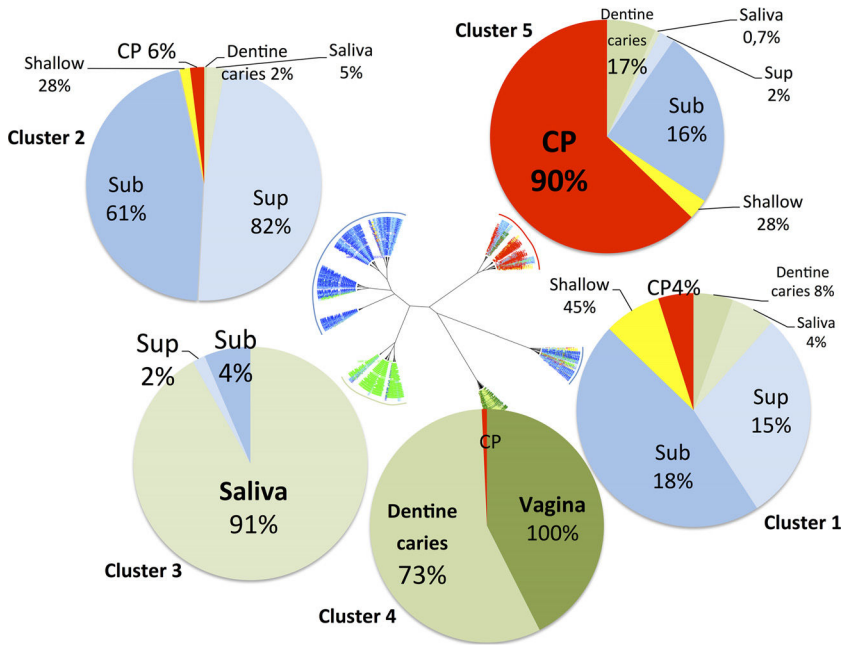


FIG 2 Unrooted tree displaying genus Bray-Curtis beta-diversity clustering of microbiota and pie charts related to the sample origins within each cluster. The tree was prepared using Figtree 1.4.2 software. The distribution of microbiota in each cluster is represented by pie charts, with different colors representing different sampling sites (supragingival [Sup] in light blue, saliva in light green, dentine caries in medium green, and midvagina in dark green) and diagnosis for subgingival (Sub) samples (healthy in dark blue, shallow in yellow, and CP in red). Percentages correspond to the number of samples from a specific sampling site in a given cluster relative to the total number of samples from the same sampling site.

were found in a fifth cluster. Two-dimensional (2D) beta-diversity analysis showed the precise distribution of the samples in the five clusters (Fig. 2). The search for an association between clusters and primers and/or study type showed that the fourth cluster was associated with V3V4 16S rRNA primers (correlation $r = 0.537$; $P < 0.001$) and with the study by Kianoush et al. (10), which used those specific primers (correlation $r = 0.608$; $P < 0.001$). No other correlations with primers were found. The two healthy clusters (clusters 1 and 2) were characterized by subgingival and supragingival samples in similar proportions (Fig. 2, light blue and dark blue sections). Focusing on healthy subgingival samples, the main difference between the two healthy clusters (clusters 1 and 2) was in the distribution of samples from the HMP study and from the other studies in the clusters; 225/323 samples from the HMP study were clustered in healthy cluster 2, while healthy cluster 1 was richer in samples derived from the other studies (44/99 samples). Cluster 3 was characterized by saliva, as 91% of the saliva samples (258/284 samples) were grouped in this cluster (correlation $r = 0.892$; $P < 0.001$) (Fig. 2). Cluster 4 was characterized by dentine caries samples (73% [80/110 samples]; correlation $r = 0.603$; $P < 0.001$) and midvagina samples (100% [60/60 samples]; correlation $r = 0.638$; $P < 0.001$). Finally, cluster 5 contained 90% of the CP samples (176/196 samples; correlation $r = 0.708$; $P < 0.001$).

It is interesting to note that 10% of the CP samples were found in the two healthy clusters (19/196 samples) and contained similar microbiota (analyzed on the basis of beta diversity), at the genus level, as dentine caries samples and/or midvagina samples (1/196 samples). Conversely, 16% of the healthy subgingival samples (69/422 samples) and 17% of the dentine caries samples (19/110 samples) were found in cluster 5.

Microbiota richness and alpha diversity in subgingival samples. A cluster comparison showed that the sampling depth (number of reads sequenced) was greater in healthy subgingival clusters 1 and 2 than in cluster 5. Nevertheless, no significant difference between healthy subgingival samples and CP samples of cluster 5 was found (Fig. 3). The observed richness (S) was lower in the CP samples of cluster 5 than in the

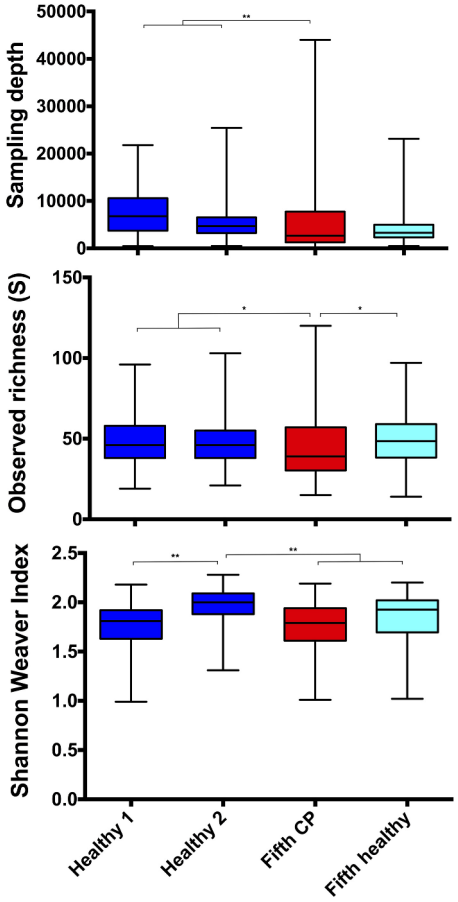


FIG 3 Alpha diversity index values. Comparisons of microbiota sampling depth, observed richness (number of different taxa per sample), and diversity (Shannon-Weaver index) in subgingival samples of healthy clusters 1 and 2 (dark blue) and either CP samples (red) or healthy subgingival samples (light blue) of cluster 5 were performed. *, $P < 0.05$; **, $P < 0.01$.

samples of healthy clusters 1 and 2 and the healthy subgingival samples of cluster 5 (Fig. 3). However, the Shannon-Weaver diversity index values showed that the diversity of healthy cluster 2 was significantly greater than the diversity of healthy cluster 1 and that of all samples from cluster 5, which were similar.

Patterns of microbial communities in subgingival samples (genus level). Genera that were present in at least 95% of all healthy subgingival samples or 95% of the CP samples from cluster 5 are presented in Fig. 4A and B, respectively. Results showed that healthy subgingival samples were dominated by 8 major genera, i.e., *Fusobacterium*, *Actinomyces*, *Streptococcus*, *Neisseria*, *Capnocytophaga*, *Prevotella*, *Corynebacterium*, and *Rothia*, and 6 minor genera, i.e., *Leptotrichia*, *Veillonella*, *Porphyromonas*, *Granulicatella*, *Kingella*, and *Gemella*. Associations were found between *Fusobacterium* and *Prevotella*, *Actinomyces*, and *Rothia* and between *Leptotrichia* and *Porphyromonas*. Common genera found in CP samples were less abundant, with 4 major genera, *Treponema*, *Porphyromonas*, *Prevotella*, and *Fusobacterium*, followed by *Streptococcus*, *Eubacterium*, *Tannerella*, and *Campylobacter* genera. Only one association was found, between *Eubacterium* and *Treponema*, while *Fusobacterium* and *Treponema* presented a negative correlation.

Calculation of dysbiosis ratios of bacteria. The dysbiosis ratios of the genera found mainly in CP samples (*Eubacteria*, *Campylobacter*, *Treponema*, and *Tannerella*) to the genera found mainly in healthy samples (*Veillonella*, *Neisseria*, *Rothia*, *Corynebacterium*, and *Actinomyces*) were significantly different among the samples according to their diagnosis. The dysbiosis ratios for healthy subgingival samples (from the HMP, $n = 323$,

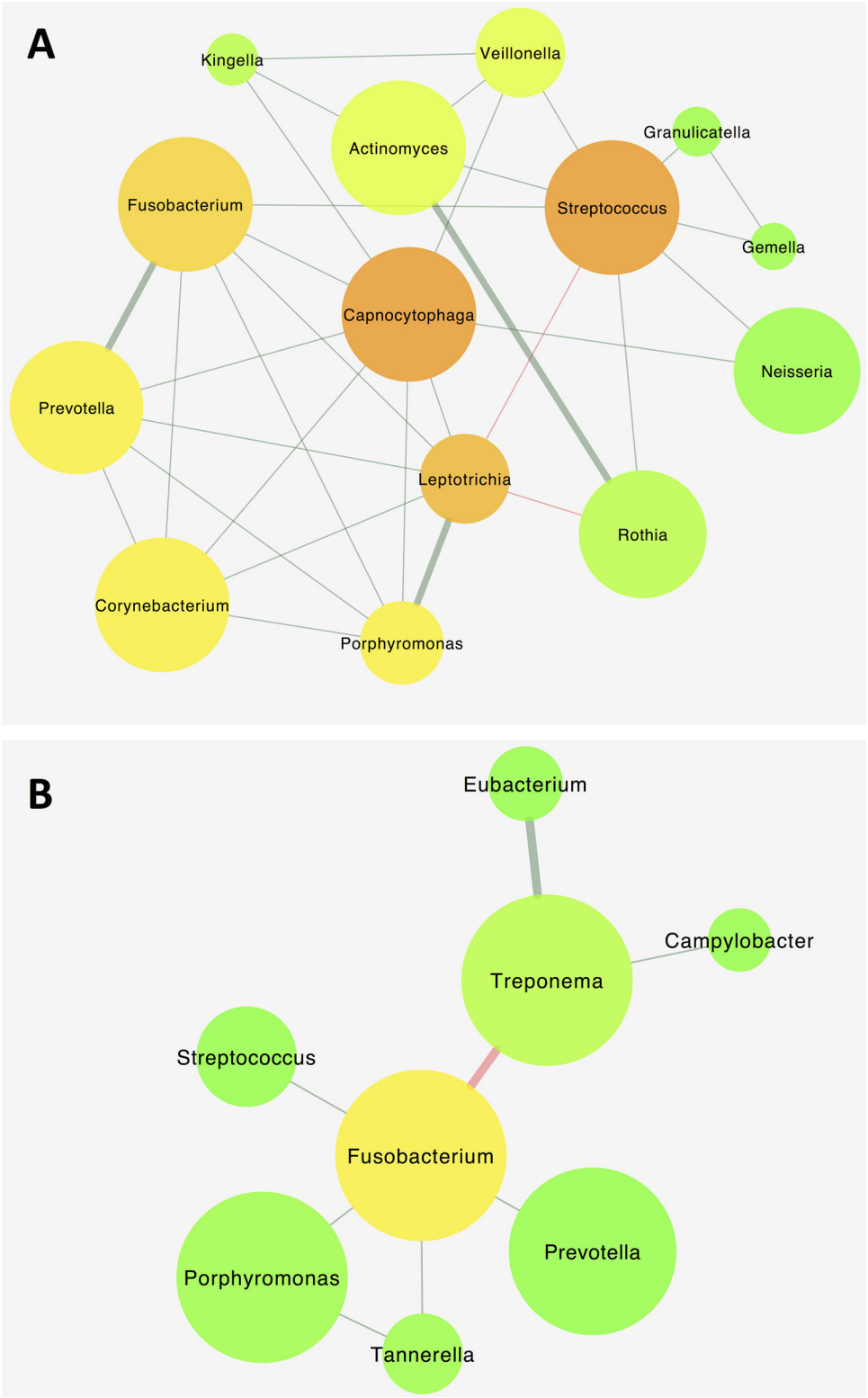


FIG 4 Patterns of subgingival microbial communities. (A) Patterns of genera present in at least 95% of all healthy subgingival samples. (B) Patterns of genera present in at least 95% of all CP samples from cluster 5. Edges represent 1 (thin line) or 2 or 3 (thick line) significant correlations between genera (green, positive; red, negative). Node colors represent the numbers of partners, ranging from 1 (green) to 7 (dark orange). Node sizes represent the abundance of each taxon.

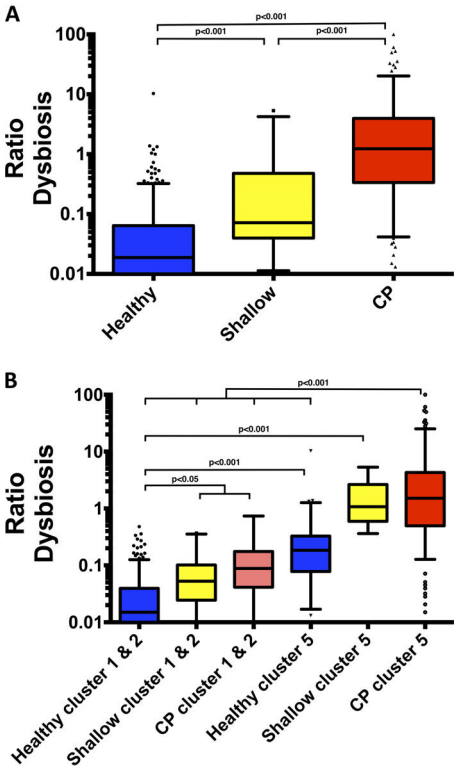


FIG 5 Subgingival dysbiosis ratios of *Eubacterium*, *Campylobacter*, *Treponema*, and *Tannerella* to *Veillonella*, *Neisseria*, *Rothia*, *Corynebacterium*, and *Actinomyces*. (A) Comparisons between healthy, shallow, and CP samples from all clusters. (B) Comparisons between clusters 1 and 2 and cluster 5 for healthy, shallow, and CP samples.

ratio of 0.016; from the other studies, $n = 99$, ratio of 0.021) yielded a median ratio of 0.018; the samples from shallow sites had a ratio of 0.071, and the CP samples had a ratio of 1.229 ($P < 0.001$) (Fig. 5A).

Although different clustering was achieved through beta-diversity analysis, no significant difference in the ratios of clusters 1 and 2 according to the clinical status (healthy, shallow, or CP) was found. Pooling of samples according to clinical status was performed, and the resulting ratios were compared to the ratios for cluster 5, as shown in Fig. 5B.

The dysbiosis ratio found for CP samples from cluster 5 (ratio of 1.510) was significantly greater than the ratios for the majority of samples from clusters 1 and 2 (healthy subgingival samples, ratio of 0.015; shallow samples, ratio of 0.052; CP samples, ratio of 0.088) and was also significantly greater than the ratio for healthy subgingival samples (ratio of 0.184) from cluster 5 ($P < 0.001$). In clusters 1 and 2, the dysbiosis ratios for CP samples were similar to the ratios for shallow sites. These two groups were significantly different from the healthy subgingival samples ($P < 0.05$) in the same cluster.

Healthy subgingival samples ($n = 69$) belonging to cluster 5 exhibited a dysbiosis ratio (ratio of 0.184) significantly different from that for the other healthy subgingival samples (ratio of 0.015) and also from that for the majority of the CP samples (cluster 5) (ratio of 1.510). These results confirmed the possible difference of these healthy subgingival microbiota ($P < 0.001$) from those of healthy clusters 1 and 2. Their ratio was not significantly different from the CP sample ratios in clusters 1 and 2, which could be considered "on the mend."

Validation of the dysbiosis ratio. A different data set, from Bizzarro et al. (11) and containing well-described samples (pocket depths of 2 to 8 mm), was used as an external control to confirm the relevance of the bacterial dysbiosis ratio. The dysbiosis

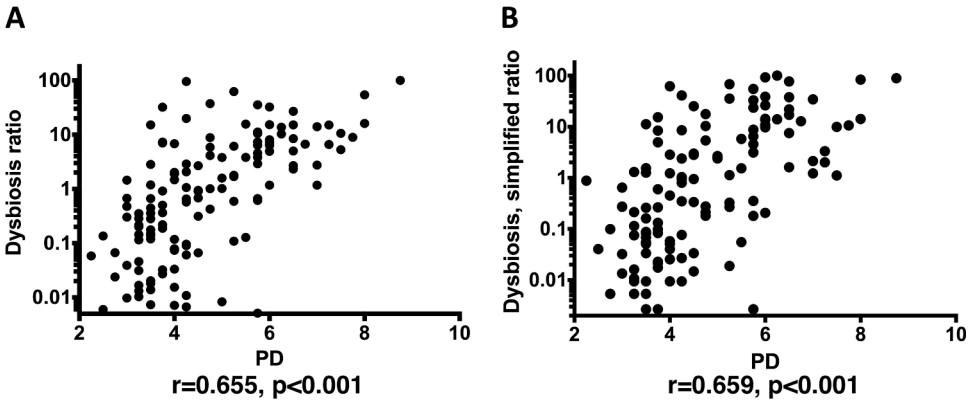


FIG 6 Correlation between pocket depth (PD) and dysbiosis. Samples from Bizzarro et al. (11) were analyzed by VAMPS, followed by calculation of the dysbiosis ratios. (A) Ratios of *Eubacterium*, *Campylobacter*, *Treponema*, and *Tannerella* to *Veillonella*, *Neisseria*, *Rothia*, *Corynebacterium*, and *Actinomyces*. (B) Simplified ratios of *Porphyromonas*, *Treponema*, and *Tannerella* to *Rothia* and *Corynebacterium*.

ratio at the genus level was correlated with the periodontal pocket depth ($r = 0.655$; $P < 0.001$) (Fig. 6A). These results, based on data for 37 patients (147 samples collected at different times, with different procedures for periodontal treatment), confirmed the link between dysbiosis and the depth of the periodontal pocket. The simplified ratio of *Porphyromonas*, *Treponema*, and *Tannerella* to *Rothia* and *Corynebacterium* showed a similar correlation ($r = 0.659$; $P < 0.001$) (Fig. 6B).

DISCUSSION

Many studies have been published since the Human Microbiome Project in 2009, increasing the volume of microbiota data available for the research community. However, comparisons between studies are challenging, at least at the species level, because of the use of different methods. This issue is a real limitation to understanding disease, as is the small number of samples in each study. Additionally, findings are more complicated for healthy subgingival samples, which usually represent less than one-half of the samples included in the studies (12, 13). This work is a taxon-based analysis, at the genus level, of sequence reads from several studies. Studying a large number of samples minimized individual variations and overcame technical variations by increasing the effective sample size. Such an analysis had already been proposed in a recent study of the microbiota in obesity (8). Studies with described healthy samples (sulci of ≤3 mm) and CP samples (pocket depths of ≥5 mm) and available raw sequence data in data banks were chosen. Data from the HMP resources (two different pairs of primers used) were added to increase the number of healthy subgingival samples with available microbiota data from 99 samples to 422 samples. The different microbiota clustered either by sampling site, such as the outgroups used as controls for this study (saliva samples in cluster 3 and dentine caries and vagina samples, which are both rich in *Lactobacillus*, in cluster 4), or by clinical status, such as subgingival samples (healthy samples in clusters 1 and 2 and CP samples in cluster 5). CP sites either can show greater microbial diversity and observed richness, compared with healthy subgingival sites (14, 15), or can present no significant difference in microbial diversity, as reported for health versus periodontitis (16). Thus, the large number of samples surpasses the technical variations, at least at the genus level, with the primers used in the different studies, and the difference between periodontal health and disease is larger than the technical variations, as described by Kirst et al. (16). No difference between the healthy subgingival and supragingival samples was found when beta-diversity analysis was performed at the genus level, as described previously (17). Ninety percent of the CP samples were found in cluster 5. To define cluster 5 as a "periodontitis cluster" by beta diversity was appealing. However, cluster 5 also contained healthy subgingival samples, indicating that further investigations are necessary to understand and to develop prediction markers for chronic periodontitis.

A core community (genera present in at least 50% of the samples) is usually identified in publications and provides a basis for disease diagnosis, prevention, and therapeutic targets (18, 19). However, the variability of genera expands as the sample size increases, thus limiting its use for establishing an easy microbiological marker for dysbiosis. In this work, genera present at higher prevalence in at least 95% of the samples were used to determine the genera implicated in health or in favor of the disease. The genera used to calculate the dysbiosis ratio in favor of periodontitis were *Treponema*, *Campylobacter*, *Eubacterium*, and *Tannerella*. These genera were identified at high levels and high prevalence in CP samples, compared to healthy samples. The genera include well-identified species (*Tannerella forsythia*, *Treponema denticola*, *C. rectus*, and *E. nodatum*) that are strongly associated with disease (3, 20–22). It should be noted that some species, such as the newly cultivated *Tannerella* clone BU063 (23, 24), which is thought to be health associated, are also found in active periodontal sites (25); therefore, this species is still controversial. Despite a significant difference in the abundance of the *Porphyromonas* genus (including *P. gingivalis*, which is strongly associated with periodontitis) between healthy samples (3.35%) and CP samples (13%), the genus was excluded in the first dysbiosis ratio because of its similar prevalence rates. Because the lowest abundance value among genera accounting for the CP calculations was that for *Campylobacter* (1.9%), this value was chosen as a cutoff value to minimize the number of genera used for the health calculations, i.e., *Rothia*, *Corynebacterium*, *Actinomyces*, *Veillonella*, and *Neisseria*. *Capnocytophaga* and *Leptotrichia* were not included because of their high rates of prevalence in CP samples (more than 90%; data not shown). Species belonging to the genus *Rothia* have been repeatedly described as being members of oral communities associated with periodontal health (26–31) or at least as being more predominant in health (28). In the same way, *Corynebacterium* appeared to be more associated with healthy subgingival biofilm (32, 33). Moreover, *Rothia* and *Corynebacterium* were among the bacteria that showed the greatest increases after periodontal treatment (34), while a study suggested that *Corynebacterium* might be considered a putative periodontal protector (35). *Veillonella* and *Actinomyces* have been negatively correlated with clinical markers in CP (36), and *Neisseria* was found in inactive sites (25). The calculated dysbiosis ratios distinguish clearly healthy subgingival samples from CP samples.

Shallow samples were divided into two groups, which can be easily explained based on the origin of the samples (healthy subgingival sites in mouths presenting chronic periodontitis). Two-thirds of the samples had low ratios (clusters 1 and 2) and could be considered microbiologically healthy. The remaining one-third of the samples (cluster 5) presented high ratios, certainly due to contamination of the sampling sites by bacteria from surrounding CP sites, and could be considered at risk of periodontitis. Thus, shallow samples may represent an intermediate stage in disease development, as proposed by Griffen et al. (14).

Healthy subgingival samples were divided into three groups. Two of the groups (clusters 1 and 2) presented the same low ratios and described an absence of dysbiosis. The third group had a higher dysbiosis ratio, similar to those for shallow sites and CP samples from clusters 1 and 2 but significantly lower than those for CP or shallow samples from cluster 5. Because healthy patients from the HMP were defined as patients with pockets depths of <4 mm, some of them could have explained this high-ratio group; however, healthy patients from other studies (19/99 subjects) were also included in this group. This result is similar to those of Zhou et al., in which a few healthy subjects with indicators of disease, such as increases in *Treponema*, were detected (37). Therefore, patients who presented with relatively high ratios could be considered at risk of periodontitis.

Conversely, a few CP samples with deep periodontal pockets (i.e., ≥ 5 mm) had low dysbiosis ratios. An hypothesis of appropriate host responses (such as a stronger immune response and/or better hygiene) could explain this discrepancy between the dysbiosis ratios and the diagnoses; these patients might be microbiologically on the mend, as revealed by both the clustering and the dysbiosis ratios. Another hypothesis

is a sampling issue between the top and base of the periodontal pocket (to be discussed later). To study the hypothesis of microbiota on the mend and the calculated dysbiosis ratios, a recent study presenting follow-up findings after treatment, with well-defined depths of periodontal pockets, was performed (11). The study was conducted with a different set of primers (for V5V7) and allowed testing of the dysbiosis ratios at the genus level with a new set of primers that had not been used to determine the ratios. Consequently, this comparative analysis can be considered a validation experiment for the ratios. A strong correlation between the dysbiosis ratios and the pocket depths was observed, thus highlighting the value of calculating the dysbiosis ratio (using the selected genera of our study) as a microbial signature to evaluate the microbiota of chronic periodontitis.

A major concern at the beginning of this work was the capacity to identify species with multiple data sets. However, the V1V2 and V5V7 primers used in three studies were not suitable for species identification. At the genus level, as reported by Bizzarro et al. (11), the proposed dysbiosis ratio (even as a simplified dysbiosis ratio) is a good microbial signature calculated using the online VAMPS software. Indeed, *Rothia* and *Corynebacterium* were the major healthy genera found and, although *Porphyromonas* was found in both health and disease, its abundance increased significantly in disease (from 3.34% to 13%). The result was interesting because it was found to be similar to the precedent ratio (correlation with pocket depth, $r = 0.659$; $P < 0.001$). However, the simplified ratio needed more adjustment, because 43 of 196 CP samples presented neither of the two healthy genera and a value of 0.1% was attributed for the calculation (see Materials and Methods ["Calculation of dysbiosis ratios of bacteria" section]).

Finally, using ratios, some data points still showed discrepancies in predicting the periodontal status. The variability in microbial composition and spatial distribution could explain these results. Deep periodontal pockets in CP patients may present gradients of oxygen tension, pH, and nutrients, as well as host defense factors, from the base of the pocket to the top (opening). This may explain why some genera (*Porphyromonas* and *Treponema*) are typically found at the base of the pocket (38, 39). However, the sampling could induce bias even after careful removal of the supragingival plaque. Healthy genera may be found predominantly at the top (opening) of the pocket, with the genera more closely associated with CP being located at the base of the pocket. Indeed, while the architecture of the periodontal pocket has not yet been clearly studied with the use of NGS analysis, the importance of the biogeography of the microbiome on the micron scale was clearly shown recently (40).

In conclusion, this study aimed to define ratios of bacteria as microbial signatures, after the analysis of publicly available raw data from different studies, independent of the technical methods used to generate the data. These ratios allowed the differentiation of healthy and diseased microbiota in the majority of samples. Standardized protocols for sampling and complete metadata in public data banks are necessary to study dysbiosis in oral health and to improve the proposed dysbiosis ratios. The addition of specific perioprotectors and potential specific pathogens to the dysbiosis calculations could also be promising. Longitudinal studies are necessary to predict exact pockets that are microbiologically on the mend or sulci with a risk of periodontitis.

MATERIALS AND METHODS

Microbiome data sets for comparison. Read sequences from healthy and CP subgingival samples from five different studies, i.e., those by Abusleme et al. (12), Kirst et al. (16), Griffen et al. (14) (shallow site samples also included), Zhou et al. (41), and Camelo-Castillo et al. (13), were retrieved from either the NCBI Sequence Read Archive (SRA) or the metagenomics (MG)-RAST server (Table 1). Twenty-four samples from patients with chronic periodontitis who were recruited between June 2010 and September 2011 at the University Hospital (Rennes, France), which were analyzed using V3V4 primers, were added (E. Boyer, S. Le Gall-David, Y. Deugnier, M. Bonnaure-Mallet, and V. Meuric, unpublished data). Each data set was manually imported into VAMPS (https://vamps.mbl.edu), while numerous healthy subgingival samples were added from the HMP (two different subgingival data sets, using V1V3 and V3V5 primers, available in VAMPS [9]). Three mouth control microbiota data sets from the HMP (saliva and supragingival, both V1V3 and V3V5 regions, available in VAMPS) and dentine caries data from the study by

TABLE 1 Subgingival microbiota samples used in this study

Authors	Accession no.	No. of subgingival microbiota samples				16S rRNA gene regions
			CP			
		Health	Shallow[a]	Diagnosis	Follow up after treatment	
Abusleme et al. (12)	GenBank SRA SRA051864	10		44		V1V2
Kirst et al. (16)	GenBank BioProject PRJNA269205	25		25		V1V3
Griffen et al. (14)	GenBank SRA SRP009299	29	29	29		V1V2 and V4
Camelo-Castillo et al. (13)	MG-RAST 12161	22		60		V1V3
Zhou et al. (37)	GenBank SRA SRA062091	13		18		V1V3
Boyer et al. (unpublished)	In progress[b]			24		V3V4
HMP (7)		119				V1V3
HMP (7)		204				V3V5
Bizzarro et al. (11)	GenBank BioProject PRJNA289294			37[c]	110[c]	V5V7

[a]Sites defined as healthy in patients with periodontitis (14).
[b]Data are available on VAMPS (data set designated Y_Hemoparo).
[c]Data on the CP microbiota from patients with follow-up findings after treatment were used to confirmed the dysbiosis ratio hypothesis (11).

Kianoush et al. (10) (V3V4 regions; BioProject accession no. PRJEB5178) were used. One midvagina microbiome data set from the HMP (V1V3 region, available in VAMPS) was used as an external mouth control. Finally, the data set from the study by Bizzarro et al. (11), containing well-described sample pocket depths (from 2 to 8 mm), was used to independently challenge the relevance of the dysbiosis ratio of bacteria involved in periodontitis.

Ecology diversity and taxonomic identifications. Reads from the different data sets were analyzed with VAMPS, using default parameters for taxonomic assignment to the genus level through the Global Alignment for Sequence Taxonomy (GAST) process and using the Ribosomal Database Project (RDP) classification to produce the best taxonomic assignment for each read. Reads identified as *Archaea*, *Eukarya*, or organelle and unknown reads were excluded from further analysis. The frequency of each taxonomic assignment in the data set was reported as a percentage (number of reads with the taxonomic assignment relative to the total number of reads in the data set). Alpha diversity, as the observed richness, and Shannon-Weaver index values were determined from the raw data sets. Differences between microbiota structures (beta diversity) were assessed using a 2D PCoA tree based on the Bray-Curtis distances, through VAMPS. Samples were divided into five clusters (clusters 1 to 5); visualizations were performed using Figtree software (version 1.4.2), and 3D PCoA plots were generated using Emperor software. Relative abundances were studied when the average abundance was above 1% in at least one sample. Assessments of significant patterns of microbial cooccurrence or mutual exclusion at the genus level were performed using Cytoscape 3.2.1 (42) and the CoNet plugin (43). Only genera found in the great majority (at least 95%) of the healthy subgingival samples or the CP samples (from cluster 5) are represented.

Calculation of dysbiosis ratios of bacteria. To measure the dysbiosis, a first ratio, based on the relative abundance of genera highly prevalent (>95%) in CP samples (*Eubacterium*, *Campylobacter*, *Treponema*, and *Tannerella*) versus genera highly prevalent (>95%) in healthy microbiota (*Veillonella*, *Neisseria*, *Rothia*, *Corynebacterium*, and *Actinomyces*), was calculated. The ratios were normalized between samples using GraphPad Prism 6 software before comparisons. A second simplified ratio of *Porphyromonas*, *Treponema*, and *Tannerella* versus *Rothia* and *Corynebacterium* was also tested. When no specific genus was detected, a value of 0.1% was attributed (because "not detected" does not mean "absence").

Statistical analysis. Normality tests for data distribution were performed. Data were studied with the Spearman correlation test for correlations of biological origins, primers used, published sample origins, and microbiota clusters. Observed richness (number of taxa per sample), Shannon-Weaver index values, and dysbiosis ratios of the genera found in disease to the genera found in health were analyzed with a Kruskal-Wallis test (nonparametric analysis of variance). Tests were carried out using GraphPad Prism 6 software and were considered significant with P values of <0.05. The significant patterns of microbial cooccurrence and mutual exclusion were analyzed as described by Faust et al. (43); a compilation of statistical analyses (Spearman and Pearson correlations and Bray-Curtis and Kullback-Leibler dissimilarity measures) was used, with a threshold set at 0.5. The data matrix was randomized by 100 row-wise permutations. The P values were adjusted with the Benjamini-Hochberg false discovery rate (FDR) correction for the number of tests, retaining only P values of <0.05. Finally, the ratios of genera and pocket depths were controlled for normality, followed by the Spearman correlation test.

ACKNOWLEDGMENTS

V.M., F.B.-H., and M.B.-M. conceived and designed the research, V.M. performed the sampling, S.L.G.-D. performed the molecular biological analyses, V.M., S.L.G.-D., E.B., L.A.-A., B.M., S.B.F., and M.B.-M. performed the bioinformatic and statistical analyses and wrote the manuscript, and M.B.-M. supervised the project.

We declare no competing financial interests.

REFERENCES

1. Teles R, Teles F, Frias-Lopez J, Paster B, Haffajee A. 2013. Lessons learned and unlearned in periodontal microbiology. Periodontol 2000 62: 95–162. https://doi.org/10.1111/prd.12010.
2. Kilian M, Chapple IL, Hannig M, Marsh PD, Meuric V, Pedersen AM, Tonetti MS, Wade WG, Zaura E. 2016. The oral microbiome: an update for oral healthcare professionals. Br Dent J 221:657–666. https://doi.org/10.1038/sj.bdj.2016.865.
3. Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL, Jr. 1998. Microbial complexes in subgingival plaque. J Clin Periodontol 25:134–144. https://doi.org/10.1111/j.1600-051X.1998.tb02419.x.
4. Hajishengallis G. 2015. Periodontitis: from microbial immune subversion to systemic inflammation. Nat Rev Immunol 15:30–44. https://doi.org/10.1038/nri3785.
5. Hajishengallis G, Darveau RP, Curtis MA. 2012. The keystone-pathogen hypothesis. Nat Rev Microbiol 10:717–725. https://doi.org/10.1038/nrmicro2873.
6. Hajishengallis G, Moutsopoulos NM, Hajishengallis E, Chavakis T. 2016. Immune and regulatory functions of neutrophils in inflammatory bone loss. Semin Immunol 28:146–158. https://doi.org/10.1016/j.smim.2016.02.002.
7. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M. 2009. The NIH Human Microbiome Project. Genome Res 19: 2317–2323. https://doi.org/10.1101/gr.096651.109.
8. Sze MA, Schloss PD. 2016. Looking for a signal in the noise: revisiting obesity and the microbiome. mBio 7:e01018-16. https://doi.org/10.1128/mBio.01018-16.
9. Huse SM, Mark Welch DB, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML. 2014. VAMPS: a website for visualization and analysis of microbial population structures. BMC Bioinformatics 15:41. https://doi.org/10.1186/1471-2105-15-41.
10. Kianoush N, Adler CJ, Nguyen KA, Browne GV, Simonian M, Hunter N. 2014. Bacterial profile of dentine caries and the impact of pH on bacterial population diversity. PLoS One 9:e92940. https://doi.org/10.1371/journal.pone.0092940.
11. Bizzarro S, Laine ML, Buijs MJ, Brandt BW, Crielaard W, Loos BG, Zaura E. 2016. Microbial profiles at baseline and not the use of antibiotics determine the clinical outcome of the treatment of chronic periodontitis. Sci Rep 6:20205. https://doi.org/10.1038/srep20205.
12. Abusleme L, Dupuy AK, Dutzan N, Silva N, Burleson JA, Strausbaugh LD, Gamonal J, Diaz PI. 2013. The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. ISME J 7:1016–1025. https://doi.org/10.1038/ismej.2012.174.
13. Camelo-Castillo AJ, Mira A, Pico A, Nibali L, Henderson B, Donos N, Tomas I. 2015. Subgingival microbiota in health compared to periodontitis and the influence of smoking. Front Microbiol 6:119. https://doi.org/10.3389/fmicb.2015.00119.
14. Griffen AL, Beall CJ, Campbell JH, Firestone ND, Kumar PS, Yang ZK, Podar M, Leys EJ. 2012. Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. ISME J 6:1176–1185. https://doi.org/10.1038/ismej.2011.191.
15. Socransky SS, Haffajee AD, Smith C, Dibart S. 1991. Relation of counts of microbial species to clinical status at the sampled site. J Clin Periodontol 18:766–775. https://doi.org/10.1111/j.1600-051X.1991.tb00070.x.
16. Kirst ME, Li EC, Alfant B, Chi YY, Walker C, Magnusson I, Wang GP. 2015. Dysbiosis and alterations in predicted functions of the subgingival microbiome in chronic periodontitis. Appl Environ Microbiol 81:783–793. https://doi.org/10.1128/AEM.02712-14.
17. Ning J, Beiko RG. 2015. Phylogenetic approaches to microbial community classification. Microbiome 3:47. https://doi.org/10.1186/s40168-015-0114-5.
18. Shade A, Handelsman J. 2012. Beyond the Venn diagram: the hunt for a core microbiome. Environ Microbiol 14:4–12. https://doi.org/10.1111/j.1462-2920.2011.02585.x.
19. Jalanka-Tuovinen J, Salonen A, Nikkila J, Immonen O, Kekkonen R, Lahti L, Palva A, de Vos WM. 2011. Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. PLoS One 6:e23035. https://doi.org/10.1371/journal.pone.0023035.
20. Laine ML, Moustakis V, Koumakis L, Potamias G, Loos BG. 2013. Modeling susceptibility to periodontitis. J Dent Res 92:45–50. https://doi.org/10.1177/0022034512465435.
21. Byrne SJ, Dashper SG, Darby IB, Adams GG, Hoffmann B, Reynolds EC. 2009. Progression of chronic periodontitis can be predicted by the levels of *Porphyromonas gingivalis* and *Treponema denticola* in subgingival plaque. Oral Microbiol Immunol 24:469–477. https://doi.org/10.1111/j.1399-302X.2009.00544.x.
22. Haffajee AD, Teles RP, Socransky SS. 2006. Association of *Eubacterium nodatum* and *Treponema denticola* with human periodontitis lesions. Oral Microbiol Immunol 21:269–282. https://doi.org/10.1111/j.1399-302X.2006.00287.x.
23. Leys EJ, Lyons SR, Moeschberger ML, Rumpf RW, Griffen AL. 2002. Association of *Bacteroides forsythus* and a novel *Bacteroides* phylotype with periodontitis. J Clin Microbiol 40:821–825. https://doi.org/10.1128/JCM.40.3.821-825.2002.
24. Vartoukian SR, Moazzez RV, Paster BJ, Dewhirst FE, Wade WG. 2016. First cultivation of health-associated *Tannerella* sp. HOT-286 (BU063). J Dent Res 95:1308–1313. https://doi.org/10.1177/0022034516651078.
25. Yost S, Duran-Pinedo AE, Teles R, Krishnan K, Frias-Lopez J. 2015. Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. Genome Med 7:27. https://doi.org/10.1186/s13073-015-0153-3.
26. Moore LV, Moore WE, Cato EP, Smibert RM, Burmeister JA, Best AM, Ranney RR. 1987. Bacteriology of human gingivitis. J Dent Res 66: 989–995. https://doi.org/10.1177/00220345870660052401.
27. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the normal bacterial flora of the oral cavity. J Clin Microbiol 43:5721–5732. https://doi.org/10.1128/JCM.43.11.5721-5732.2005.
28. Colombo AP, Boches SK, Cotton SL, Goodson JM, Kent R, Haffajee AD, Socransky SS, Hasturk H, Van Dyke TE, Dewhirst F, Paster BJ. 2009. Comparisons of subgingival microbial profiles of refractory periodontitis, severe periodontitis, and periodontal health using the human oral microbe identification microarray. J Periodontol 80:1421–1432. https://doi.org/10.1902/jop.2009.090185.
29. Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF, Nelson KE, Gill SR, Fraser-Liggett CM, Relman DA. 2010. Bacterial diversity in the oral cavity of 10 healthy individuals. ISME J 4:962–974. https://doi.org/10.1038/ismej.2010.30.
30. Heuer W, Stiesch M, Abraham WR. 2011. Microbial diversity of supra- and subgingival biofilms on freshly colonized titanium implant abutments in the human mouth. Eur J Clin Microbiol Infect Dis 30:193–200. https://doi.org/10.1007/s10096-010-1068-y.
31. Moutsopoulos NM, Chalmers NI, Barb JJ, Abusleme L, Greenwell-Wild T, Dutzan N, Paster BJ, Munson PJ, Fine DH, Uzel G, Holland SM. 2015. Subgingival microbial communities in leukocyte adhesion deficiency and their relationship with local immunopathology. PLoS Pathog 11: e1004698. https://doi.org/10.1371/journal.ppat.1004698.
32. Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA, Sahasrabudhe A, Dewhirst FE. 2001. Bacterial diversity in human subgingival plaque. J Bacteriol 183:3770–3783. https://doi.org/10.1128/JB.183.12.3770-3783.2001.
33. Ling Z, Liu X, Luo Y, Yuan L, Nelson KE, Wang Y, Xiang C, Li L. 2013. Pyrosequencing analysis of the human microbiota of healthy Chinese undergraduates. BMC Genomics 14:390. https://doi.org/10.1186/1471-2164-14-390.
34. Laksmana T, Kittichotirat W, Huang Y, Chen W, Jorgensen M, Bumgarner R, Chen C. 2012. Metagenomic analysis of subgingival microbiota following non-surgical periodontal therapy: a pilot study. Open Dent J 6:255–261. https://doi.org/10.2174/1874210601206010255.
35. Zorina OA, Petrukhina NB, Basova AA, Shibaeva AV, Trubnikova EV, Shevelev AB. 2014. Identification of key markers of normal and pathogenic microbiota determining health of periodontium by NGS-sequencing 16S-rDNA libraries of periodontal swabs. Stomatologiia (Mosk) 93:25–31. https://doi.org/10.17116/stomat201493625-31. (In Russian.)
36. Teles R, Sakellari D, Teles F, Konstantinidis A, Kent R, Socransky S, Haffajee A. 2010. Relationships among gingival crevicular fluid biomarkers, clinical parameters of periodontal disease, and the subgingival

microbiota. J Periodontol 81:89–98. https://doi.org/10.1902/jop.2009.090397.
37. Zhou Y, Mihindukulasuriya KA, Gao H, La Rosa PS, Wylie KM, Martin JC, Kota K, Shannon WD, Mitreva M, Sodergren E, Weinstock GM. 2014. Exploration of bacterial community classes in major human habitats. Genome Biol 15:R66. https://doi.org/10.1186/gb-2014-15-5-r66.
38. Zijnge V, van Leeuwen MB, Degener JE, Abbas F, Thurnheer T, Gmur R, Harmsen HJ. 2010. Oral biofilm architecture on natural teeth. PLoS One 5:e9321. https://doi.org/10.1371/journal.pone.0009321.
39. Kigure T, Saito A, Seida K, Yamada S, Ishihara K, Okuda K. 1995. Distribution of *Porphyromonas gingivalis* and *Treponema denticola* in human subgingival plaque at different periodontal pocket depths examined by immunohistochemical methods. J Periodont Res 30:332–341. https://doi.org/10.1111/j.1600-0765.1995.tb01284.x.
40. Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. 2016. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci U S A 113:E791–E800. https://doi.org/10.1073/pnas.1522149113.
41. Zhou M, Rong R, Munro D, Zhu C, Gao X, Zhang Q, Dong Q. 2013. Investigation of the effect of type 2 diabetes mellitus on subgingival plaque microbiota by high-throughput 16S rDNA pyrosequencing. PLoS One 8:e61516. https://doi.org/10.1371/journal.pone.0061516.
42. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303.
43. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. 2012. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol 8:e1002606. https://doi.org/10.1371/journal.pcbi.1002606.

Discussion et perspectives

Les bactéries sont des microorganismes unicellulaires dont l'organisation de la cellule est traditionnellement décrite comme simple, avec un contenu cellulaire sans compartiments spécialisés ni membranes internes, ni noyau (d'où le nom de procaryotes). Ainsi, la réplication du génome bactérien, sa transcription et même sa traduction ont lieu dans le cytoplasme et de manière simultanée. En réalité, depuis quelques années, nous découvrons que sous cette apparente simplicité se cachent des mécanismes et des organisations complexes avec notamment la présence de micro-compartiments protéiques qui sont des compartiments subcellulaires spécialisés permettant de colocaliser certaines enzymes pour catalyser des réactions, mais également pour protéger certaines protéines sensibles et pour séquestrer des intermédiaires toxiques pour la cellule (Sutter et al. 2017; Yeates et al. 2010; Yeates et al. 2013). De plus, le génome bactérien, généralement considéré comme composé d'un seul chromosome circulaire, peut être complexe par la présence de chromosomes additionnels, parfois linéaires, de mégaplasmides ou de plasmides. Ce génome est organisé sous forme d'un nucléoide, composé d'ADN et de protéines de type histones nommées NAPs (*Nucleoid-Associated Proteins*) (Feijoo-Siota et al. 2017). De plus, alors que les eucaryotes pluricellulaires utilisent différentes cellules spécialisés pour fonctionner, la cellule bactérienne doit être capable de tout faire pour survivre dans un environnement souvent très changeant et pouvoir capter des stimuli, y réagir, métaboliser des molécules pour se nourrir, produire ses éléments structuraux et s'adapter aux carences et au stress. Nous découvrons finalement que les organismes bactériens sont des machines bien plus complexes que leur apparente simplicité ne laisse imaginer.

Les bactéries sont des formes de vie les plus anciennes, les plus importantes et les plus diverses de notre planète. Il n'existe d'environnement sur Terre qui soit libre de vie microbienne. Les microbes sont retrouvés dans la troposphère, la stratosphère, sur les matériaux envoyés par les humains dans l'espace (Horneck et al. 2010; Pawar et al. 2012). Les microbes sont également présents dans les océans les plus profonds ainsi que plusieurs kilomètres sous la surface de la terre (Danovaro et al. 2016; Fredrickson and Onstott 1996). Certaines bactéries ont une capacité à survivre pendant des millions d'années dans un état de dormance sous forme de spores (Lennon and Jones 2011). Au final, plus de 50 % de la biomasse de la planète correspond aux procaryotes alors que ces cellules représentent en moyenne un millième du volume des cellules eucaryote (Whitman et al. 1998). Cette proportion atteint les 90 % dans les océans (Sogin et al. 2006). La plupart de la biodiversité

terrestre est et a toujours été microbienne, malgré de nombreuses différences dans la manière de la calculer.

En plus de ces valeurs quantitatives de biomasse et de diversité taxonomique, l'extraordinaire diversité des capacités microbiennes est à l'origine des cycles biogéochimiques qui assurent le maintien de la vie sur notre planète. Elles ont un rôle capital dans les cycles biogéochimiques des éléments clés de la vie sur Terre comme le carbone, l'azote, l'hydrogène et l'oxygène. Elles participent au recyclage de ces éléments et à leur fixation dans les molécules du vivant (Rousk and Bengtson 2014).

Enfin, il existe une multiplicité des symbioses entre bactéries et êtres multicellulaires. Ces relations symbiotiques comprennent des endosymbioses (à l'intérieur des cellules) et des ectosymbioses (à l'extérieur des cellules) (Moya et al. 2008). Ces symbioses proviennent de la co-évolution des entités biologiques impliquées et de leurs génomes (Herre et al. 1999). Les bactéries associées aux mammifères, par exemple, sont responsables de la dégradation, dans le tube digestif, de plusieurs types de polysaccharides ce qui les rend utilisables par les cellules eucaryotes. Elles sont également impliquées dans le développement du système immunitaire et sont même citées dans les livres d'immunologie comme faisant partie intégrale des barrières physiques et naturelles du système immunitaire.

La découverte de la présence d'une *flore microbienne* chez les mammifères a été faite par Antony van Leeuwenhoek en 1683, mais les recherches sur les microbes qui peuplent les humains et les animaux se sont poursuivies à un rythme très lent jusqu'à l'implication de ces microbes comme causants des maladies. Plus tard, leur contribution à notre santé a été mise en évidence. Le fait que les bactéries indigènes exercent un effet protecteur en empêchant la colonisation par des agents pathogènes exogènes a été démontrée vers la fin des années 1960 (Abrams and Bishop 1966; van der Waaij et al. 1971). La plupart des informations concernant le rôle des microbes dans le développement des mammifères ont été obtenues par des études comparatives impliquant une *microflore* normale et des animaux dits *germ-free* artificiellement colonisés par des espèces microbiennes choisies, études qui se sont développées dans les années 1970 (Coates 1975).

Dans ces environnements symbiotiques, comme dans tous les endroits qu'elles colonisent, les bactéries vivent en communautés complexes. En effet, les cultures pures de bactéries cultivées en laboratoire ne sont pas naturelles, et la structuration des communautés microbiennes en biofilms est la forme la plus fréquente dans la nature, plus fréquente que la

forme planctonique (libres dans un milieu généralement liquide). Un biofilm est une agglomération hétérogène et structurée de bactéries, et parfois d'autres microorganismes autour d'une matrice extracellulaire composée de polymères. La structure en 3D des biofilms et leur polarité (attaché à une surface généralement) génèrent des gradients internes de pH, d'oxygène ou encore de nutriments. Les structures internes des biofilms (pores et différentes couches de composition variée) permettent des relations de synergie qui permettent aux individus de s'adapter à différents stress comme la température, la pression et même aux molécules toxiques comme les antibiotiques (Jefferson 2004). Ces biofilms sont donc des structures où résident des coopérations essentielles, comparables à celles d'un organisme pluricellulaire.

Ces communautés sont nommées *microbiote*, terme qui englobe tous les microorganismes : bactéries, archées, virus, champignons filamenteux, levures et protozoaires présents dans un environnement ou un échantillon. Force est de constater que dans les études actuellement publiées, ce terme est utilisé pour décrire exclusivement la communauté bactérienne. Par abus de langage, les études parlent souvent de microbiote alors qu'elles décrivent des "*bactériotes*". La première apparition du terme *microbiote* est assez ancienne. A. Goetz cite ce terme dans un brevet déposé en 1945 et publié en 1950 (Goetz 1950) puis dans le contexte de la santé humaine, dans sa forme orale avec les travaux de S. Socransky et collaborateurs en 1963 (Socransky et al. 1963) et en 1966 pour la description du microbiote digestif par R. Dubos (Dubos 1966). L'intérêt récent pour l'étude des microbiotes animaux n'est lié qu'à deux phénomènes : 1) la prise de conscience de l'importance de ces communautés microbiennes dans les maladies, le développement cellulaire, la nutrition, le comportement ou encore le bien-être et 2) le développement de technologies qui nous permettent d'identifier les microbes présents. Cet intérêt s'est accompagné d'une tendance à la mode de "*faire du microbiote*" par des laboratoires et équipes sans réelle expertise en microbiologie ou en bioanalyse des données écologiques et bioinformatiques récoltées. Comme le titre un article de juillet 2016 : "*Suddenly everyone is a microbiota specialist*" (Boers et al. 2016), les articles s'accumulent avec un nombre affolant de catalogues de bactéries, présentes ou absentes dans telles ou telles conditions. Le microbiote est analysé partout, sur notre corps mais aussi sur les objets de la vie courante (claviers, téléphones…), nos lieux de vie (maison, trottoirs, sols…), nos animaux de compagnie, nos enfants, nos anciens … mais finalement, qu'est ce que ces études cherchent à prouver ? Quelle est l'hypothèse de travail, la question biologique ? Et que dire des plans expérimentaux ? Des

cohortes de quelques dizaines, au mieux quelques centaines d'individus pour établir des résultats relayés par la presse avec une telle couverture médiatique qu'elles deviennent des "*vérités*". Que penser aussi des expériences menés sur des animaux (rongeurs ou insectes) et extrapolées aux humains ? La liste des questionnements est longue sur la fiabilité de ces expériences : temps d'expérimentation souvent court, métadonnées souvent incomplètes ou inconsistantes, comparaisons entre situations (malade/sain, obèse/maigre) faites sur des individus différents, souvent non apparentés génétiquement, sur la base de diagnostiques mal documentés et des critères de diagnostique hétérogènes. Ces manquements en standardisation, en contrôle et en documentation entraîne des incohérences lorsque de la comparaison des études similaires. Comment pouvons nous comparer des études qui n'utilisent pas les mêmes protocoles de prélèvement, d'extraction d'ADN, de *barcoding*, d'analyses et de vérification ?

Oui, la diversité des microbiotes est importante et le nombre de cellules procaryotes cohabitant dans nos organismes et nos environnements est supérieure à celui de nos cellules eucaryotes mais cette diversité est la question et non la réponse. Comme l'indique vraiment bien l'article d'A. Shade de janvier 2017 (Shade 2017), la mesure de la diversité microbienne d'un environnement devrait servir de point de départ pour élucider des mécanismes écologiques plutôt que d'être publié comme une réponse. De nombreuses communautés microbiennes ont d'innombrables membres et leur diversité n'est ni bonne, ni mauvaise. Elle est mesurée à l'aide d'indices statistiques qui doivent nous permettre d'élaborer des hypothèses pour tester des mécanismes écologiques dynamiques. De plus, il n'existe aucune valeur absolue de la diversité, son calcul dépend de méthodes et de ces méthodes, les interprétations. Les études sur le microbiote ne peuvent et ne devraient pas être réalisés sans hypothèses biologiques, sans plans expérimentaux, ni expériences de validation finales au risque de ne produire qu'un nième catalogue de bactéries présentes dans un nième environnement.

Parmi les hypothèses biologiques qui nous ont intéressés dans cette thèse, il y a le problème des bactéries à effet préjudiciable. Certaines bactéries sont pathogènes pour l'humain et les animaux, et les maladies infectieuses sont toujours d'importance capitale pour la santé alors que les laboratoires de recherche en microbiologie tendent à disparaître du paysage international de la recherche. Pourtant, la découverte de nouvelles bactéries capables de nous rendre malades est d'actualité, soit en profitant d'une diminution des capacités de notre système immunitaire (pathogènes opportunistes), soit en le manipulant (pathobionts). Même si les limites entre les définitions sont floues, les bactéries qui sont capables de

produire des pathologies sont fréquentes et affectent une grande partie de la population humaine. De plus, des maladies souvent déclarées comme non infectieuses, comme le diabètes, les maladies cardiovasculaires ou le cancer, peuvent être modulées ou causées par des microorganismes. Nous sommes à un point de l'histoire où nous avons de plus en plus de connaissances sur la capacité des microorganismes à causer des pathologies. Le lien avéré entre les ulcères peptiques (gastro-duodénales) et *Helicobacter pylori* est une des premières preuves que les agents infectieux, notamment les bactéries, sont capables de provoquer des maladies considérées comme chroniques et non infectieuses. Cette preuve de concept qui a révolutionné la microbiologie actuelle a été récompensée par le prix Nobel de Physiologie et Médecine de 2005 attribué à B. Marshall et R. Warren (Zetterstrom 2006). Le lien avec le cancer gastrique et l'oncoprotéine CagA (la première oncoprotéine bactérienne décrite) d'*Helicobacter pylori* a été une découverte majeure et corrèle, pour la première fois, une bactérie et le cancer, même si le lien entre certains virus et le cancer étaient déjà connus.

Pour comprendre les microbiotes et les conséquences de certaines dysbioses, il est donc indispensable de connaitre et de comprendre la biologie des bactéries concernées, leurs métabolismes et leurs capacités à s'adapter et/ou modifier leur environnement. Que ce soit dans des conditions de santé, de maladie ou les étapes intermédiaires de prédisposition ou de rémission, sans recherche fondamentale ou avec uniquement la vision étriquée de quelques "*modèles*" bactériens comme *Escherichia coli* ou *Bacillus subtilis*, nos connaissances sont incomplètes et donc imprécises et non optimales. Certes, pour comprendre nos microbiotes, nous avons besoin de connaître leur composition quantitativement et qualitativement et les changements de ces compositions. Cependant, que l'espèce bactérienne soit majoritaire dans un microbiote ne veut pas forcément dire qu'elle est celle qui a le plus d'impact dans la communauté. Un exemple de ce type de bactérie, qui même en faible proportion est capable d'avoir un impact important est *Porphyromonas gingivalis*. Cette espèce, décrite comme clé de voûte des maladies parodontales, est capable, à des faibles concentrations dans les sulci gingivo-dentaires, de modifier, en modulant le système immunitaire de l'hôte, le biofilm sous-gingival. Cette modification entraîne une perte de l'homéostasie et une dysbiose qui, à son tour, stimule et amplifie l'inflammation (Hajishengallis et al. 2012).

Dans ce contexte de microbiologie moderne, qui place les bactéries non plus dans des tubes à essai ou des boites de Petri, mais dans leurs environnements naturels, la génomique bactérienne est sans nulle doute la discipline qui peut apporter des réponses. Cette discipline de la biologie qui combine expertise en microbiologie et en bioinformatique répond à

plusieurs questions, notamment sur les capacités de *pathogénicité* des bactéries mais également leurs potentialités métaboliques (Leao et al. 2017), l'adaptation des organismes microbiens à leurs hôtes eucaryotes (Hersemann et al. 2017; Lee et al. 2017), la compréhension des facteurs conditionnant les co-occurences microbiennes (Orata et al. 2015; X. Zhang et al. 2017), les apparitions et maintenances des résistance aux antiobiotiques ou autres drogues/métaux (Jeukens et al. 2017) ou encore la plasticité des génomes, l'acquisition ou la perte de fonctions et l'adaptation aux environnements (Cao et al. 2016; MacArthur et al. 2017). Ceci ne sont que des exemples car finalement, la génomique comparative permet d'adresser aux génotypes n'importe quelle question d'ordre fonctionnelle ou évolutive et d'en faire des hypothèses phénotypiques à vérifier par la suite.

La génomique est, par définition, l'étude des génomes entiers, de leurs architectures et de leurs contenus géniques. La matière première de cette discipline est une (ou plusieurs) séquences nucléotidiques ordonnées, obtenues après séquençage, reconstruites par assemblage des reads et annotées automatique et/ou manuellement.

Nous avons expliqué l'évolution des technologies de séquençage en introduction de ce document, de même que les stratégies pour assembler les génomes et nous avons pu démontrer que de nombreux groupes bactériens présentent de graves lacunes d'information, notamment les phyla bactériens du microbiote humain comme les Bacteroidetes. De même, nous avons pointé la faible et/ou douteuse qualité des génomes incomplets en draft et le retentissement sur les résultats d'études de génomique comparative. De plus, nous montrons par le premier article présenté que même les génomes dits *complets* parce que circularisés, peuvent contenir des erreurs d'assemblage, comme des inversions, des recombinaisons ou encore des loci répétés incomplètement reconstruits. Cette démonstration nous interroge sur les valeurs scientifiques des informations génomiques en microbiologie.

Le séquençage des premiers génomes microbiens, à la fin des années 1970, était publié dans des revues scientifiques à haut facteur d'impact et qui ont encore de nos jours une grande renommée scientifique, comme Cell, Science et Nature. De nos jours, publier un génome bactérien est banal et peu attractif pour les revues scientifiques, à tel point que certains chercheurs n'associent plus de publication à leurs travaux de séquençage et préfèrent juste les soumettre dans une base de données internationale comme le NCBI ou l'ENA. Si le génome est publié, il l'est souvent sous forme de *genome announcement*, *report* ou *note*, correspondant à une communication courte voire très courte (500-1500 mots, une page A4) dans des

journaux à faible facteur d'impact (D. R. Smith 2017). Les conséquences sont que d'abord, la publication passe la plupart du temps inaperçue et l'information fournie est souvent trop succincte pour susciter l'intérêt de la communauté internationale. De plus, comme nous l'avons dit à plusieurs reprises dans ce document, l'immense majorité des génomes, de 80% à 90% selon le phylum bactérien, sont des drafts. Nous avons un paradoxe d'accumulation massif de données de séquences alors que notre compréhension des génomes ne fait pas réellement de saut significatif si ce n'est qu'en terme d'évolution. Nous séquençons parce que nous pouvons séquencer, parce ce n'est pas cher et parce que nous voulons toujours plus de donnés. Comme le traduisent très correctement A. Vincent et collaborateurs en juillet 2017, ce coté *très abordable* de générer de l'information génomique a donné lieu à un faux sentiment de simplicité qui contredit le fait que de nombreux chercheurs considèrent encore les technologies de séquençage comme des boîtes noires (Vincent et al. 2017).

Étant donné que le coût du séquençage est devenu moins prohibitif, de nombreux laboratoires, dans le monde entier, sont maintenant en mesure de mener leurs propres projets de séquençage et même de maintenir leurs propres séquenceurs. Toutefois, cette nouvelle accessibilité a conduit beaucoup de non-spécialistes à utiliser les NGS sans connaissance préalable et, par conséquent, à utiliser la technologie de façon non optimale. C'est notamment le cas dans le domaine de la microbiologie où la taille relativement faible des génomes microbiens peut conduire à l'impression que le séquençage de ces génomes est simple. La réalité est que, sans une stratégie adéquate de séquençage et les expertises et moyens nécessaires en génomique, bioinformatique et biocuration, les chercheurs sont susceptibles d'être déçus et frustrés de ne pas être en mesure de générer des données de qualité en raison de la mauvaise planification, d'un manque de ressources ou d'attente irréaliste. Il est grand temps de repenser les projets de génomique microbienne, en privilégiant comme il se doit en biologie, l'importante étape préalable qui consiste à formuler une hypothèse qui va au-delà de la séquence et de développer une approche expérimentale appropriée. Pour utiliser optimalement le financement de la recherche, nous devons tout d'abord déterminer quel problème scientifique nous voulons résoudre, puis quel ensemble de données sera les plus utile pour répondre à cette question. C'est cette question qui doit conditionner les choix de la technologie de séquençage (longueur des lectures, type de lectures, profondeur de séquençage, taux d'erreur) et non le prix ou la proximité/disponibilité d'un appareil ou de plateformes. Une fois le meilleur séquenceur sélectionné, il faut penser à la création des bibliothèques (type de fragmentation de l'ADN, multiplexage, contrôles de qualité). Après le

séquençage, il faut prévoir le temps et les moyens nécessaires à l'analyse des données, étape indispensable pour obtenir des résultats de qualité, mais souvent jugée trop fastidieuse et/ou coûteuse pour être faite.

Ainsi, la qualité de ces données semble une considération mineure dans les objectifs des projets de séquençage. La question biologique adressée aux données n'est pas claire. Des exemples sont multiples comme ces projets colossaux de séquençage massif de la biodiversité des sols (T. M. Vogel et al. 2009), des océans (Tully et al. 2017) ou encore des environnements extrêmes de tous types qui sont très coûteux, mais dont le bénéfice scientifique reste discutable. Des projets pour séquencer et avoir des génomes de toutes les branches des bactéries existent (Kyrpides et al. 2014), cependant la valeur ajoutée est faible, si ce n'est que pour compléter l'inventaire de nouveaux phyla. De plus, les génomes sont assemblés en draft et des efforts de finition ne semblent pas à l'ordre du jour.

Nous vivons dans une époque où l'accumulation d'information est plus rapide que notre capacité ou notre volonté à l'analyser, l'interpréter et lui donner du sens. Pourtant, sans cette analyse, l'information reste inutile. C'est exactement ce qui c'est passé avec le séquençage du génome humain car même si l'obtention de la séquence en elle même a beaucoup de valeur et a été un challenge, au final, c'est la détection de gènes et la compréhension de leurs fonctions, de leurs interactions, de leur organisation et de leur régulation qui est important et pourtant, ces travaux sont toujours en cours depuis plus de 15 ans. Pour les génomes bactériens le problème est similaire. L'annotation est faite de manière automatique avec peu ou pas d'intervention humaine. La multitude de pipelines d'annotation automatiques encourage cette pratique et même le NCBI, qui est un des poids lourds des bases de données génomiques, propose son propre pipeline et/ou encourage les chercheurs à soumettre leur séquence sans aucune annotation. Les études se basant sur ces génomes annotés automatiquement contiennent de nombreuses erreurs comme un taux important de faux positifs ou d'annotation erronée, que ce soit structurellement (comme les bornes des CDS) ou fonctionnellement (description des fonctions).

Alors que les bases de données encouragent fortement la mise à disposition des reads utilisés pour produire la séquence génomique, cela reste à la discrétion des équipes de recherche. Le résultat est que la grande majorité des génomes n'ont pas de reads associés disponibles et donc la seule façon de vérifier la construction et/ou l'annotation est de refaire un séquençage. Pourtant, l'utilisation des reads par une équipe externe ne devrait pas produire

de peur ou d'appréhension, c'est le principe même de la science ouverte (*Open Science*) qui vise à rendre la recherche, les données et leur diffusion accessibles à tous les niveaux de la société scientifique et universitaire. Cette approche est boudée par de nombreux laboratoires qui produisent des embargos sur leurs résultats et leurs matériels biologiques par crainte, ou dans l'objectif d'établir une certaine suprématie scientifique et/ou économique via un protectionnisme excessif. Pourtant, seule la possibilité de réutilisation des données (et des matériaux du vivant) par différents projets de recherche et la possibilité d'un examen supplémentaire et/ou complémentaire des résultats scientifiques favorise la transparence des interprétations et l'utilisation maximisée des données. La raison invoquée fréquemment pour ce tabou est la pression faite sur les laboratoires pour déposer des brevets alors que les études montrent que seule une faible proportion des brevets déposés par la recherche publique génèrent des revenus et que la majorité de ces derniers demeure inexploitée. De plus, les méthodes changent et s'améliorent, de nouvelles technologies apparaissent, de nouvelles stratégies permettant d'utiliser les données d'origine ou de nouvelles hypothèses biologiques voient le jour. En science, nous sommes confrontés à la vérification par des expériences complémentaires, et même quand nous avons fait de notre mieux, nous ne sommes pas à l'abri d'avoir publié des interprétations qui, au regard de nouveaux développements sont finalement inexacts ou d'avoir commis des erreurs, surtout dans le cadre d'analyse de gros volume de données (comme en génomique). Par ailleurs, même dans les rares cas où les données sont disponibles, de nombreuses informations (source précise d'isolement, année, méthodes utilisés…), qui devraient figurer dans les métadonnées et/ou dans l'article scientifique associée, sont absentes ou succinctes. De plus, dans quelques cas, ces informations sont contradictoires. Par exemple, le génome de *Porphyromonas gingivalis* 381 contient, dans l'entête du fichier genbank, la sous-partie "*Genome-Assembly-Data*" qui décrit l'assemblage de la souche comme étant fait de manière guidée sur le génome AE015924.1 qui correspond à la souche W83. Or dans le *genome announcement*, rien ne laisse penser à un assemblage sur référence, mais fait penser à un assemblage *de novo*.

Créer du sens dans ces données, les vérifier et les homogénéiser pour avoir des informations de qualité est crucial et correspond au processus dit de « biocuration » de l'annotation structurale et fonctionnelle. Dans le contexte de l'augmentation des volumes de données et d'accumulation des erreurs, ce travail hautement nécessaire est très long, méthodique et nécessite une grande rigueur.

Les articles présentés dans ce document sont des exemples de ce travail de biocuration, pour annoter des génomes et pour identifier des loci d'intérêt dans un genre bactérien. Le travail des bioinformaticiens n'est pas seulement de création de nouvelles données mais également de traitement des données existantes. Dans plusieurs cas, selon la question biologique, les données sont déjà disponibles et des meilleures règles communautaires pour la mise à disposition des données associées aux données brutes (les métadonnées) permettrait leur réutilisation. Avec les connaissances et l'expérience, un biocurateur peut prendre en considération le meilleur choix d'annotation, peut fournir un niveau de cohérence et enrichir les connaissances biologiques. La *curation* des données est une étape essentielle dans la gestion et l'analyse de données génomiques. Le biocurateur ne fait pas que capturer, trier, annoter et représenter les données, il recherche également les contradictions et les manquements. Ce travail nécessite également de lire et de trier les connaissances relevantes issues de la recherche en sciences biologiques et biomédicales ce qui, compte tenu de la croissance fulgurante de la littérature scientifique, reste très consommateur en temps. Pourtant, en dépit de ces difficultés et du temps nécessaire, c'est une démarche *gagnante* comme le montre ce travail de thèse.

La bioinformatique est une discipline qui a une longue trajectoire et les questions auxquelles elle apporte des réponses sont multiples. Dans une époque où les données massives appelées *big data* sont omniprésentes, la biologie et la microbiologie ne peuvent que s'enrichir de cette discipline. Cependant, malgré des algorithmes performants, le traitement de ces données ne peut pas être uniquement automatique. C'est le biocurateur qui donne du sens à l'analyse des données, c'est celui qui identifie les vides ou les erreurs dans les informations et qui planifie les expériences pour valider et compléter ces connaissances.

Pour moi et pour ma carrière future, l'acquisition de l'expérience dans le domaine de la bioinformatique est cruciale. La microbiologie moderne ne peut plus vraiment se passer de la bioinformatique et pour répondre de manière pertinente aux questions posées, le futur chercheur doit être capable de comprendre, d'utiliser et de comparer différents algorithmes ou logiciels qui composent la boîte à outils de la bioanalyse. La réflexion sur toutes les étapes de l'analyse est primordiale, et comme toujours en biologie et en sciences, la collecte des données est capitale. L'*Open Science* nous permet de réutiliser les données des autres et que d'autres utilisent nos données, ce qui finit par enrichir la communauté scientifique.

C'est dans ce contexte que j'espère poursuivre mes travaux en tant que futur chercheur. Le travail collaboratif est la meilleure façon d'avancer. Dans une époque où les chercheurs sont confrontés à des budgets de plus en plus réduits même dans les "pays riches", la recherche doit être faite par des groupes pluridisciplinaires où les domaines de compétences sont complémentaires.

À mon retour au Costa Rica, j'intégrerai une équipe de quatre chercheurs en bactériologie des anaérobies. Mon poste est équivalent à un Maître de Conférences des Universités à l'Universidad de Costa Rica, la plus grande université publique d'Amérique Centrale. Mon temps sera donc distribué entre des tâches administratives, mes enseignements et ma recherche. Le principal cours auquel je participerai correspond à celui de *Bacteriología General* (Bactériologie Générale) qui comprend des cours magistraux (environ 12 h par enseignant) et des travaux pratiques (environ 100 h par enseignant) pendant le deuxième semestre. L'encadrement d'étudiants qui préparent leur *Trabajo Final de Graduación* (Thèse d'Exercice) est également prévu. C'est souvent dans le cadre de ces encadrements que nous développons nos activités de recherche.

Les directions futures des travaux que nous envisageons de poursuivre sont multiples. La durée d'une thèse est finalement courte quand le thésard doit apprendre son métier. Au début de ma thèse, nous envisagions de séquencer une vingtaine de souches de *Porphyromonas gingivalis*. Nous avons les données de séquençage de 25 nouvelles souches de cette bactérie et nous envisageons de les assembler et annoter en suivant la méthodologie que nous avons développée et présentée dans le premier article de ce document. Ceci permettrait de tripler le nombre de génomes complets disponibles pour l'espèce bactérienne et de réaliser de réelles études du pangénome de l'espèce avec des souches caractérisées phénotypiquement sur lesquelles nous pouvons réaliser des expériences de laboratoire pour valider nos analyses *in silico*.

D'autre part, nous voulons étendre l'analyse des génomes au genre *Porphyromonas*. Pour cela, une première étape d'achat des souches type des collections internationales est en cours. Nous envisageons également de contacter les auteurs de plusieurs publications de génomes *draft* pour collaborer avec eux pour la production de génomes complets des souches qu'ils possèdent. Les modalités de ce dernier point sont encore à définir et nous attendons mon retour au Costa Rica pour commencer l'étape de collecte du matériel biologique. J'ai répondu

à un appel à projets lancé par l'Universidad de Costa Rica afin d'avoir un financement pour ce projet.

Finalement, je participerais aux projets dans lesquels je suis déjà implique au sein de l'équipe génétique du chien de l'IGDR. Dans un premier temps, nous cherchons à décrire les microbiotes oraux de chiens sains en utilisant une stratégie d'amplification de gène codant l'ARNr 16S en entier ce qui nous permettra de mesurer la diversité en bactéries et archées des échantillons collectés. Cette collecte est en cours, et nous mettons actuellement au point les techniques d'extraction de l'ADN communautaire et l'amplification du marqueur en entier. Cette collecte s'accompagne d'un renseignement de plusieurs paramètres environnementaux qui pourraient être liés aux différences attendues entre les différentes cavités orales des chiens.

Conclusion et apports de la thèse

Ce travail de thèse présente, dans un premier temps, des travaux préalables à la génomique : le séquençage, l'assemblage et l'annotation. Pour chacune de ces étapes nous avons choisi consciencieusement la meilleure stratégie et nous avons expliqué notre choix. Dans ce contexte, nous décrivons le paysage des génomes bactériens au milieu de la deuxième décennie du XXI^{ème} siècle, avec une immensité de génomes de qualité douteuse et de faible valeur pour de nombreuses études de génomique structurale, fonctionnelle ou comparative. Nous démontrons également que les assemblages complets ou finis doivent être vérifiés et mis à jour le cas échéant.

Nous nous sommes intéressés à une bactérie anaerobie de la bouche, *Porphyromonas gingivalis*. Cette bactérie est un pathobiont, capable de produire des changements profonds dans le microbiote amenant à une dysbiose et à une inflammation chronique. Au cours de mon parcours, j'ai eu l'opportunité de travailler avec des bactéries similaires, capables de produire des maladies lors d'une dysbiose intestinale comme le pathogène opportuniste *Clostridium difficile* ou de produire des dysplasies et éventuellement le cancer gastrique comme *Helicobacter pylori*. Toutes les trois, sont des modèles intéressants de bactéries avec des capacités variables de pathogénicité. La microbiologie ne doit plus voir tout noir (les mauvaises bactéries) ou tout blanc (les bonnes bactéries), c'est en essayant de comprendre toutes les "*nuances de gris*" que nous pourrons élucider plus précisément comme les microbes influencent nos vies. Dans ce contexte, la génomique est capable d'apporter des réponses puisque toutes les capacités de la bactérie sont codées dans son génome.

Ce n'est que suite à l'obtention des génomes complets vérifiés que nos travaux de "vraie génomique" ont pu commencer à être. Tout d'abord en génomique fonctionnelle, avec une annotation homogène et biocurée, processus long et parfois fastidieux, mais qui apporte un réel gain en connaissance des fonctions codées par le génome. Puis une analyse de génomique comparative du pangénome des trois souches re-séquencées, re-assemblées er re-annotées. Cette analyse, bien que préliminaire, montre bien que la plasticité génomique de l'espèce est plus dans l'organisation du génome (mosaïcité) que dans les différentes capacités de codage. Ainsi, même si certains îlots issus de transferts horizontaux sont identifiables, ils restent assez bien distribués dans les souches et génèrent peu de différence quantitative. Il est intéressant d'observer cette quantité impressionnante de séquences d'insertion avec au final peu de fonctions nouvelles apportés.

Ces travaux nous ont donné l'idée d'étudier des gènes qui sont communs aux trois souches mais qui ont une séquence avec un pourcentage d'identité nucléotidique faible (que nous proposons de classer dans un génome core dit variable). Plusieurs gènes codant pour des protéines dites *facteurs de virulence* sont classés dans ce cluster. Nous avons choisi d'étudier en première intention les fimbriae, structures d'adhésion, au sein du genre *Porphyromonas*. Nous avons observé, encore une fois, des annotations très disparates des gènes qui codent pour ces protéines et que dans une grande majorité des cas, les CDS étaient annotées de fonction inconnue. En observant cet exemple, nous constatons que le problème de l'annotation de mauvaise qualité est répandue et que toute analyse de génomique comparative doit être précédée de génomique fonctionnelle et d'un travail intense de biocuration. Seulement avec ces prérequis, les travaux de génomique comparative peuvent répondre efficacement aux questions des microbiologistes expérimentaux et fournir des hypothèses solides à tester en laboratoire. Au final le gain en temps est énorme, même si le projet est plus lent à démarrer.

Enfin, cette thèse présente des travaux de génomique environnementale en analysant des données principalement publiques de microbiotes oraux de sulci sains et de sites atteints de parodontite. Encore une fois le travail de recueil et traitement des données a été capital pour avoir des résultats qui répondaient à notre hypothèse de travail.

Au final ce travail de thèse est une preuve que la biocuration est la meilleure stratégie pour arriver à avoir des réponses à nos questions biologiques. Elle démontre que la vérification et la validation des résultats de bioinformatique est un travail nécessaire, que les machines les plus performantes, leurs algorithmes et logiciels, ne sont pas capables de donner du sens aux phénomènes biologiques sans l'intervention humaine. Il est donc nécessaire de se former, et cela en permanence, aux nouvelles technologies, nouvelles méthodologies. Nous devons arrêter d'être des utilisateurs passifs de la bioinformatique et nous devons comprendre ce que nous faisons, les questions posées et les contrôles à utiliser pour avoir des résultats de qualité. Nous devons devenir des chercheurs en bioinformatique et rester ouverts à nous former tout au long de notre carrière.

Références bibliographiques

Abby, S. and Daubin, V. (2007), 'Comparative genomics and the evolution of prokaryotes', *Trends Microbiol,* 15 (3), 135-41.

Abrams, G. D. and Bishop, J. E. (1966), 'Effect of the normal microbial flora on the resistance of the small intestine to infection', *J Bacteriol,* 92 (6), 1604-8.

Accetto, T. and Avgustin, G. (2011), 'Inability of Prevotella bryantii to Form a Functional Shine-Dalgarno Interaction Reflects Unique Evolution of Ribosome Binding Sites in Bacteroidetes', *PLOS ONE,* 6 (8), e22914.

Aemaimanan, P., Amimanan, P., and Taweechaisupapong, S. (2013), 'Quantification of key periodontal pathogens in insulin-dependent type 2 diabetic and non-diabetic patients with generalized chronic periodontitis', *Anaerobe,* 22, 64-8.

Allesen-Holm, M., et al. (2006), 'A characterization of DNA release in Pseudomonas aeruginosa cultures and biofilms', *Mol Microbiol,* 59 (4), 1114-28.

Altenhoff, A. M., et al. (2016), 'Standardized benchmarking in the quest for orthologs', *Nat Methods,* 13 (5), 425-30.

Altschul, S. F., et al. (1990), 'Basic local alignment search tool', *J Mol Biol,* 215 (3), 403-10.

Altschul, S. F., et al. (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res,* 25 (17), 3389-402.

Andrade, M. A., et al. (1999), 'Automated genome sequence analysis and annotation', *Bioinformatics,* 15 (5), 391-412.

Angiuoli, S. V. and Salzberg, S. L. (2011), 'Mugsy: fast multiple alignment of closely related whole genomes', *Bioinformatics,* 27 (3), 334-42.

Anonymous (1965), 'Nobel Prize for Medicine', *The Lancet,* 286 (7417), 836 - 37.

Ariyaratne, P. N. and Sung, W. K. (2011), 'PE-Assembler: de novo assembler using short paired-end reads', *Bioinformatics,* 27 (2), 167-74.

Arutyunov, D. and Frost, L. S. (2013), 'F conjugation: back to the beginning', *Plasmid,* 70 (1), 18-32.

Ashton, F. E. and Caugant, D. A. (2001), 'The panmictic nature of Neisseria meningitidis serogroup B during a period of endemic disease in Canada', *Can J Microbiol,* 47 (4), 283-9.

Atanasova, K. R. and Yilmaz, O. (2014), 'Looking in the Porphyromonas gingivalis cabinet of curiosities: the microbium, the host and cancer association', *Mol Oral Microbiol,* 29 (2), 55-66.

Au, K. F., et al. (2012), 'Improving PacBio long read accuracy by short read alignment', *PLoS One,* 7 (10), e46679.

Avery, O. T., Macleod, C. M., and McCarty, M. (1944), 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III.', *J Exp Med,* 79 (2), 137-58.

Aziz, R. K., et al. (2008), 'The RAST Server: rapid annotations using subsystems technology', *BMC Genomics,* 9, 75.

Backus, J. W. and Heising, W. P. (1964), 'Fortran', *IEEE Transactions on Electronic Computers,* EC-13 (4), 382-85.

Badrinarayanan, A., Le, T. B., and Laub, M. T. (2015), 'Bacterial chromosome organization and segregation', *Annu Rev Cell Dev Biol,* 31, 171-99.

Balkwill, David L., Fredrickson, J. K., and Romine, M. F. (2006), 'Sphingomonas and Related Genera', in Martin Dworkin, et al. (eds.), *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass* (New York, NY: Springer New York), 605-29.

Bao, E., Jiang, T., and Girke, T. (2014), 'AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references', *Bioinformatics,* 30 (12), i319-i28.

Bao, S., et al. (2011), 'Evaluation of next-generation sequencing software in mapping and assembly', *J Hum Genet,* 56 (6), 406-14.

Baron, E. J. (1996), 'Classification', in E. J. Baron (ed.), *Medical Microbiology* (4th edition edn.; Galveston, Texas: University of Texas Medical Branch at Galveston).
Bateman, A., Coggill, P., and Finn, R. D. (2010), 'DUFs: families in search of function', *Acta Crystallogr Sect F Struct Biol Cryst Commun,* 66 (Pt 10), 1148-52.
Bell, D. C., et al. (2012), 'DNA base identification by electron microscopy', *Microsc Microanal,* 18 (5), 1049-53.
Bennett, P. M. (2004), 'Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement', *Methods Mol Biol,* 266, 71-113.
Bentley, S. D., et al. (2002), 'Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2)', *Nature,* 417 (6885), 141-7.
Bernardet, Jean-François and Nakagawa, Yasuyoshi (2006), 'An Introduction to the Family Flavobacteriaceae', in Martin Dworkin, et al. (eds.), *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass* (New York, NY: Springer New York), 455-80.
Besemer, J. and Borodovsky, M. (2005), 'GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses', *Nucleic Acids Res,* 33 (Web Server issue), W451-4.
Bessede, E., et al. (2014), 'Helicobacter pylori generates cells with cancer stem cell properties via epithelial-mesenchymal transition-like changes', *Oncogene,* 33 (32), 4123-31.
Bessede, E., et al. (2016), 'Deletion of IQGAP1 promotes Helicobacter pylori-induced gastric dysplasia in mice and acquisition of cancer stem cell properties in vitro', *Oncotarget,* 7 (49), 80688-99.
Binder Gallimidi, A., et al. (2015), 'Periodontal pathogens Porphyromonas gingivalis and Fusobacterium nucleatum promote tumor progression in an oral-specific chemical carcinogenesis model', *Oncotarget,* 6 (26), 22613-23.
Boers, S. A., Jansen, R., and Hays, J. P. (2016), 'Suddenly everyone is a microbiota specialist', *Clin Microbiol Infect,* 22 (7), 581-2.
Borodovsky, M. and McIninch, J. (1993), 'GENMARK: Parallel gene recognition for both DNA strands', *Computers & Chemistry,* 17 (2), 123-33.
Borodovsky, M., et al. (1995), 'Detection of new genes in a bacterial genome using Markov models for three gene classes', *Nucleic Acids Res,* 23 (17), 3554-62.
Boutet, E., et al. (2016), 'UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View', *Methods Mol Biol,* 1374, 23-54.
Brocchieri, L. and Karlin, S. (2005), 'Protein length in eukaryotic and prokaryotic proteomes', *Nucleic Acids Res,* 33 (10), 3390-400.
Brudno, M., et al. (2003), 'LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA', *Genome Res,* 13 (4), 721-31.
Brussow, H., Canchaya, C., and Hardt, W. D. (2004), 'Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion', *Microbiol Mol Biol Rev,* 68 (3), 560-602, table of contents.
Bull, M. J. and Plummer, N. T. (2014), 'Part 1: The Human Gut Microbiome in Health and Disease', *Integr Med (Encinitas),* 13 (6), 17-22.
Bulmer, M. (1987), 'Coevolution of codon usage and transfer RNA abundance', *Nature,* 325 (6106), 728-30.
Butler, J., et al. (2008), 'ALLPATHS: de novo assembly of whole-genome shotgun microreads', *Genome Res,* 18 (5), 810-20.
Cain, A. A., Kosara, R., and Gibas, C. J. (2012), 'GenoSets: visual analytic methods for comparative genomics', *PLoS One,* 7 (10), e46401.
Cairns, J. (1963), 'The bacterial chromosome and its manner of replication as seen by autoradiography', *J Mol Biol,* 6, 208-13.
Canchaya, C., et al. (2003), 'Phage as agents of lateral gene transfer', *Curr Opin Microbiol,* 6 (4), 417-24.
Cao, D. M., et al. (2016), 'Comparative Genomics of H. pylori and Non-Pylori Helicobacter Species to Identify New Regions Associated with Its Pathogenicity and Adaptability', *Biomed Res Int,* 2016, 6106029.

Caspi, R., et al. (2014), 'The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases', *Nucleic Acids Res,* 42 (Database issue), D459-71.
Chargaff, E. (1950), 'Chemical specificity of nucleic acids and mechanism of their enzymatic degradation', *Experientia,* 6 (6), 201-9.
Chen, H. and Boutros, P. C. (2011), 'VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R', *BMC Bioinformatics,* 12, 35.
Chistiakov, D. A., Orekhov, A. N., and Bobryshev, Y. V. (2016), 'Links between atherosclerotic and periodontal disease', *Exp Mol Pathol,* 100 (1), 220-35.
Choi, J. H., Cho, H. G., and Kim, S. (2005), 'GAME: a simple and efficient whole genome alignment method using maximal exact match filtering', *Comput Biol Chem,* 29 (3), 244-53.
Chothia, C., et al. (2003), 'Evolution of the protein repertoire', *Science,* 300 (5626), 1701-3.
Chu, T. C., et al. (2009), 'GR-Aligner: an algorithm for aligning pairwise genomic sequences containing rearrangement events', *Bioinformatics,* 25 (17), 2188-93.
Claverys, J. P. and Martin, B. (2003), 'Bacterial "competence" genes: signatures of active transformation, or only remnants?', *Trends Microbiol,* 11 (4), 161-5.
Coates, M. E. (1975), 'Gnotobiotic animals in research: their uses and limitations', *Lab Anim,* 9 (4), 275-82.
Cock, P. J., et al. (2010), 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Res,* 38 (6), 1767-71.
Cole, S. T. and Saint Girons, I. (1994), 'Bacterial genomics', *FEMS Microbiol Rev,* 14 (2), 139-60.
Collen, M. F. (1994), 'The origins of informatics', *J Am Med Inform Assoc,* 1 (2), 91-107.
Collen, M. F. and Kulikowski, C. A. (2015), 'The Development of Digital Computers', in Morris F. Collen and Marion J. Ball (eds.), *The History of Medical Informatics in the United States* (London: Springer London), 3-73.
Couturier, E. and Rocha, E. P. (2006), 'Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes', *Mol Microbiol,* 59 (5), 1506-18.
Coykendall, A. L., Kaczmarek, F. S., and Slots, J. (1980), 'Genetic Heterogeneity in Bacteroides asaccharolyticus (Holdeman and Moore 1970) Finegold and Barnes 1977 (Approved Lists, 1980) and Proposal of Bacteroides gingivalis sp. nov. and Bacteroides macacae (Slots and Genco) comb. nov', *Int J Syst Bacteriol,* 30 (3), 559-64.
Coyne, M. J., et al. (2013), 'Phylum-wide general protein O-glycosylation system of the Bacteroidetes', *Mol Microbiol,* 88 (4), 772-83.
Crick, F. H. (1958), 'On protein synthesis', *Symp Soc Exp Biol,* 12, 138-63.
Crick, F. H., Griffith, J. S., and Orgel, L. E. (1957), 'Codes without commas', *Proc Natl Acad Sci U S A,* 43 (5), 416-21.
Cuthbertson, J. M., Doyle, D. A., and Sansom, M. S. (2005), 'Transmembrane helix prediction: a comparative evaluation and analysis', *Protein Eng Des Sel,* 18 (6), 295-308.
Dagan, T. (2011), 'Phylogenomic networks', *Trends Microbiol,* 19 (10), 483-91.
Daily, J. (2016), 'Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments', *BMC Bioinformatics,* 17, 81.
Danovaro, R., et al. (2016), 'Macroecological drivers of archaea and bacteria in benthic deep-sea ecosystems', *Sci Adv,* 2 (4), e1500961.
Dark, M. J. (2013), 'Whole-genome sequencing in bacteriology: state of the art', *Infect Drug Resist,* 6, 115-23.
Darling, A. C., et al. (2004), 'Mauve: multiple alignment of conserved genomic sequence with rearrangements', *Genome Res,* 14 (7), 1394-403.
Darling, A. E., Mau, B., and Perna, N. T. (2010), 'progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement', *PLoS One,* 5 (6), e11147.

Darmon, E. and Leach, D. R. (2014), 'Bacterial genome instability', *Microbiol Mol Biol Rev,* 78 (1), 1-39.
Daubin, V. and Perriere, G. (2003), 'G+C3 structuring along the genome: a common feature in prokaryotes', *Mol Biol Evol,* 20 (4), 471-83.
Dayarian, A., Michael, T. P., and Sengupta, A. M. (2010), 'SOPRA: Scaffolding algorithm for paired reads via statistical optimization', *BMC Bioinformatics,* 11, 345.
Dayhoff, M. O. (1965a), *Atlas of protein sequence and structure* (Silver Spring Md.: National Biomedical Research Foundation).
--- (1965b), 'Computer aids to protein sequence determination', *J Theor Biol,* 8 (1), 97-112.
Del Fabbro, C., et al. (2013), 'An extensive evaluation of read trimming effects on Illumina NGS data analysis', *PLoS One,* 8 (12), e85024.
Delcher, A. L., et al. (1999), 'Alignment of whole genomes', *Nucleic Acids Res,* 27 (11), 2369-76.
Delihas, N. (2008), 'Small mobile sequences in bacteria display diverse structure/function motifs', *Mol Microbiol,* 67 (3), 475-81.
Denton, J. F., et al. (2014), 'Extensive error in the number of genes inferred from draft genome assemblies', *PLoS Comput Biol,* 10 (12), e1003998.
do Nascimento Silva, A., et al. (2017), 'Pathogenicity and genetic profile of oral Porphyromonas species from canine periodontitis', *Arch Oral Biol,* 83, 20-24.
Donthu, R., Lewin, H. A., and Larkin, D. M. (2009), 'SyntenyTracker: a tool for defining homologous synteny blocks using radiation hybrid maps and whole-genome sequence', *BMC Res Notes,* 2, 148.
Dubos, R. (1966), 'The microbiota of the gastrointestinal tract', *Gastroenterology,* 51 (5), 868-74.
Earl, D., et al. (2011), 'Assemblathon 1: a competitive assessment of de novo short read assembly methods', *Genome Res,* 21 (12), 2224-41.
Eddy, S. R. (2004), 'What is a hidden Markov model?', *Nat Biotechnol,* 22 (10), 1315-6.
Eddy, S. R. and Durbin, R. (1994), 'RNA sequence analysis using covariance models', *Nucleic Acids Res,* 22 (11), 2079-88.
Edman, P. and Begg, G. (1967), 'A protein sequenator', *Eur J Biochem,* 1 (1), 80-91.
Eid, J., et al. (2009), 'Real-time DNA sequencing from single polymerase molecules', *Science,* 323 (5910), 133-8.
Endler, L., et al. (2009), 'Designing and encoding models for synthetic biology', *J R Soc Interface,* 6 Suppl 4, S405-17.
English, A. C., et al. (2012), 'Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology', *PLoS One,* 7 (11), e47768.
Escudero, J. A., et al. (2015), 'The Integron: Adaptation On Demand', *Microbiol Spectr,* 3 (2), Mdna3-0019-2014.
Ewing, B. and Green, P. (1998), 'Base-calling of automated sequencer traces using phred. II. Error probabilities', *Genome Res,* 8 (3), 186-94.
Ewing, B., et al. (1998), 'Base-calling of automated sequencer traces using phred. I. Accuracy assessment', *Genome Res,* 8 (3), 175-85.
Falush, D., Stephens, M., and Pritchard, J. K. (2003), 'Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies', *Genetics,* 164 (4), 1567-87.
Feijoo-Siota, L., et al. (2017), 'Considerations on bacterial nucleoids', *Appl Microbiol Biotechnol,* 101 (14), 5591-602.
Feng, Y., et al. (2015), 'Nanopore-based fourth-generation DNA sequencing technology', *Genomics Proteomics Bioinformatics,* 13 (1), 4-16.
Fernandez-Gomez, B., et al. (2013), 'Ecology of marine Bacteroidetes: a comparative genomics approach', *Isme j,* 7 (5), 1026-37.
Finegold, S. M. and Barnes, E. M. (1977), 'Report of the ICSB taxonomic subcommittee on gram-negative anaerobic rods. Proposal that the saccharolytic and asaccharolytic strains at present classified in the species Bacteroides melaninogenicus (Oliver and

Wherry) be reclassified in two species as Bacteroides malaninogenicus and Bacteroides asaccharolyticus', *Int J Syst Bacteriol,* 27, 388-91.
Finn, R. D., et al. (2016), 'The Pfam protein families database: towards a more sustainable future', *Nucleic Acids Res,* 44 (D1), D279-85.
Fitch, W. M. (1970), 'Distinguishing homologous from analogous proteins', *Syst Zool,* 19 (2), 99-113.
--- (2000), 'Homology a personal view on some of the problems', *Trends Genet,* 16 (5), 227-31.
Fitch, W. M. and Margoliash, E. (1967), 'Construction of phylogenetic trees', *Science,* 155 (3760), 279-84.
Fleischmann, R. D., et al. (1995), 'Whole-genome random sequencing and assembly of Haemophilus influenzae Rd', *Science,* 269 (5223), 496-512.
Fox, G. E. and Stackebrandt, E. (1988), 'The Application of 16S rRNA Cataloguing and 5S rRNA Sequencing in Bacterial Systematics', in R. R. Colwell and R. Grigorova (eds.), *Methods in Microbiology* (19: Academic Press), 405-58.
Fraenkel-Conrat, H. and Williams, R. C. (1955), 'Reconstitution of active tobacco mosaic virus from its inactive protein and nucleic acid components', *Proc Natl Acad Sci U S A,* 41 (10), 690-8.
Fraser, C. M., et al. (1995), 'The minimal gene complement of Mycoplasma genitalium', *Science,* 270 (5235), 397-403.
Fraser, C. M., et al. (1997), 'Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi', *Nature,* 390 (6660), 580-6.
Fredrickson, J. K. and Onstott, T.C. (1996), 'Microbes Deep inside the Earth', *Scientific American*.
Fujisawa, T., et al. (2014), 'CyanoBase and RhizoBase: databases of manually curated annotations for cyanobacterial and rhizobial genomes', *Nucleic Acids Res,* 42 (Database issue), D666-70.
Galibert, F., Sedat, J., and Ziff, E. (1974), 'Direct determination of DNA nucleotide sequences: structure of a fragment of bacteriophage phiX172 DNA', *J Mol Biol,* 87 (3), 377-407.
Galler, B. A. (1986), 'The IBM 650 and the Universities', *Annals of the History of Computing,* 8 (1), 36-38.
Galperin, M. Y., et al. (2015), 'Expanded microbial genome coverage and improved protein family annotation in the COG database', *Nucleic Acids Res,* 43 (Database issue), D261-9.
Gawad, C., Koh, W., and Quake, S. R. (2016), 'Single-cell genome sequencing: current state of the science', *Nat Rev Genet,* 17 (3), 175-88.
Gherna, R. and Woese, C. R. (1992), 'A partial phylogenetic analysis of the "flavobacter-bacteroides" phylum: basis for taxonomic restructuring', *Syst Appl Microbiol,* 15 (4), 513-21.
Gicquel, B. (1993), 'RFLP typing of the Mycobacterium tuberculosis bacilli', *Tuber Lung Dis,* 74 (4), 223-4.
Gilbert, D. G. (1990), 'Dot plot sequence comparisons on Macintosh computers', *Comput Appl Biosci,* 6 (2), 117.
Glowniak, J. (1998), 'History, structure, and function of the Internet', *Semin Nucl Med,* 28 (2), 135-44.
Goetz, A. (1950), 'Method for producing a microbicidal composition of matter', in United States Patents (ed.), *US2521713 A* (United States of America).
Goodacre, N. F., Gerloff, D. L., and Uetz, P. (2013), 'Protein domains of unknown function are essential in bacteria', *MBio,* 5 (1), e00744-13.
Grantham, R., et al. (1980), 'Codon catalog usage and the genome hypothesis', *Nucleic Acids Res,* 8 (1), r49-r62.
Greenhill, C. (2015), 'Gut microbiota: Firmicutes and Bacteroidetes involved in insulin resistance by mediating levels of glucagon-like peptide 1', *Nat Rev Endocrinol,* 11 (5), 254.

Griffith, F. (1928), 'The Significance of Pneumococcal Types', *J Hyg (Lond),* 27 (2), 113-59.
Griffiths, A. J. F., et al. (2000), 'The Stucture and Replication of DNA', *An introduction to genetic analysis* (7th edn.; New York: W. H. Freeman).
Griffiths-Jones, S., et al. (2003), 'Rfam: an RNA family database', *Nucleic Acids Res,* 31 (1), 439-41.
Gupta, R. S. (2004), 'The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes', *Crit Rev Microbiol,* 30 (2), 123-43.
Gupta, R. S. and Lorenzini, E. (2007), 'Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the Bacteroidetes and Chlorobi species', *BMC Evol Biol,* 7, 71.
Ha, N. H., et al. (2015), 'Prolonged and repetitive exposure to Porphyromonas gingivalis increases aggressiveness of oral cancer cells by promoting acquisition of cancer stem cell properties', *Tumour Biol,* 36 (12), 9947-60.
Hahnke, R. L., et al. (2016), 'Genome-Based Taxonomic Classification of Bacteroidetes', *Front Microbiol,* 7, 2003.
Hajishengallis, G., Darveau, R. P., and Curtis, M. A. (2012), 'The keystone-pathogen hypothesis', *Nat Rev Microbiol,* 10 (10), 717-25.
Hajishengallis, G., et al. (2011), 'Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement', *Cell Host Microbe,* 10 (5), 497-506.
Han, K., et al. (2013), 'Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu', *Sci Rep,* 3, 2101.
Haque, F., et al. (2013), 'Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA', *Nano Today,* 8 (1), 56-74.
Harris, T. D., et al. (2008), 'Single-molecule DNA sequencing of a viral genome', *Science,* 320 (5872), 106-9.
Heather, J. M. and Chain, B. (2016), 'The sequence of sequencers: The history of sequencing DNA', *Genomics,* 107 (1), 1-8.
Heinken, A., et al. (2013), 'Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut', *Gut Microbes,* 4 (1), 28-40.
Helm, R. A., et al. (2003), 'Genomic rearrangements at rrn operons in Salmonella', *Genetics,* 165 (3), 951-9.
Hernandez, D., et al. (2008), 'De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer', *Genome Res,* 18 (5), 802-9.
Herre, E. A., et al. (1999), 'The evolution of mutualisms: exploring the paths between conflict and cooperation', *Trends Ecol Evol,* 14 (2), 49-53.
Hersemann, L., et al. (2017), 'Comparative genomics of host adaptive traits in Xanthomonas translucens pv. graminis', *BMC Genomics,* 18 (1), 35.
Hershey, A. D. and Chase, M. (1952), 'Independent functions of viral protein and nucleic acid in growth of bacteriophage', *J Gen Physiol,* 36 (1), 39-56.
Hidalgo, O., et al. (2017), 'Is There an Upper Limit to Genome Size?', *Trends Plant Sci,* 22 (7), 567-73.
Hitchon, C. A., et al. (2010), 'Antibodies to porphyromonas gingivalis are associated with anticitrullinated protein antibodies in patients with rheumatoid arthritis and their relatives', *J Rheumatol,* 37 (6), 1105-12.
Hohl, M., Kurtz, S., and Ohlebusch, E. (2002), 'Efficient multiple genome alignment', *Bioinformatics,* 18 Suppl 1, S312-20.
Horneck, G., Klaus, D. M., and Mancinelli, R. L. (2010), 'Space microbiology', *Microbiol Mol Biol Rev,* 74 (1), 121-56.
Hornef, M. (2015), 'Pathogens, Commensal Symbionts, and Pathobionts: Discovery and Functional Effects on the Host', *Ilar j,* 56 (2), 159-62.
Hovland, E., et al. (2017), 'DprA from Neisseria meningitidis: properties and role in natural competence for transformation', *Microbiology,* 163 (7), 1016-29.
Howe, C. J., et al. (2003), 'Evolution of the chloroplast genome', *Philos Trans R Soc Lond B Biol Sci,* 358 (1429), 99-106; discussion 06-7.

Hoyles, L. and McCartney, A. L. (2009), 'What do we mean when we refer to Bacteroidetes populations in the human gastrointestinal microbiota?', *FEMS Microbiol Lett,* 299 (2), 175-83.
Huerta-Cepas, J., et al. (2017), 'Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper', *Mol Biol Evol,* 34 (8), 2115-22.
Hugon, P., et al. (2015), 'A comprehensive repertoire of prokaryotic species identified in human beings', *Lancet Infect Dis,* 15 (10), 1211-9.
Hulsen, T., de Vlieg, J., and Alkema, W. (2008), 'BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams', *BMC Genomics,* 9, 488.
Human Microbiome Project Consortium (2012), 'Structure, function and diversity of the healthy human microbiome', *Nature,* 486 (7402), 207-14.
Hutcherson, J. A., et al. (2016), 'Comparison of inherently essential genes of Porphyromonas gingivalis identified in two transposon-sequencing libraries', *Mol Oral Microbiol,* 31 (4), 354-64.
Hyatt, D., et al. (2010), 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics,* 11, 119.
Indrischek, H., et al. (2016), 'The paralog-to-contig assignment problem: high quality gene models from fragmented assemblies', *Algorithms Mol Biol,* 11, 1.
International Human Genome Sequencing Consortium (2004), 'Finishing the euchromatic sequence of the human genome', *Nature,* 431 (7011), 931-45.
Jackman, S. D., et al. (2017), 'ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter', *Genome Res,* 27 (5), 768-77.
Jackson, R. W., et al. (2011), 'The influence of the accessory genome on bacterial pathogen evolution', *Mob Genet Elements,* 1 (1), 55-65.
Jaroszewski, L., et al. (2009), 'Exploration of uncharted regions of the protein universe', *PLoS Biol,* 7 (9), e1000205.
Jefferson, K. K. (2004), 'What drives bacteria to produce a biofilm?', *FEMS Microbiol Lett,* 236 (2), 163-73.
Jeukens, J., et al. (2017), 'Comparative genomics of a drug-resistant Pseudomonas aeruginosa panel and the challenges of antimicrobial resistance prediction from genomes', *FEMS Microbiol Lett,* 364 (18).
Johnson, C. M. and Grossman, A. D. (2015), 'Integrative and Conjugative Elements (ICEs): What They Do and How They Work', *Annu Rev Genet,* 49, 577-601.
Johnson, E. L., et al. (2017), 'Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes', *J Mol Med (Berl),* 95 (1), 1-8.
Johnston, C., et al. (2014a), 'Bacterial transformation: distribution, shared mechanisms and divergent control', *Nat Rev Microbiol,* 12 (3), 181-96.
Johnston, C., et al. (2014b), 'Streptococcus pneumoniae, le transformiste', *Trends Microbiol,* 22 (3), 113-9.
Jones, C. E., Brown, A. L., and Baumann, U. (2007), 'Estimating the annotation error rate of curated GO database sequence annotations', *BMC Bioinformatics,* 8, 170.
Ju, J., et al. (2006), 'Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators', *Proc Natl Acad Sci U S A,* 103 (52), 19635-40.
Kamada, M., et al. (2014), 'Whole genome complete resequencing of Bacillus subtilis natto by combining long reads with high-quality short reads', *PLoS One,* 9 (10), e109999.
Kamada, N., et al. (2013), 'Control of pathogens and pathobionts by the gut microbiota', *Nat Immunol,* 14 (7), 685-90.
Kanehisa, M., et al. (2017), 'KEGG: new perspectives on genomes, pathways, diseases and drugs', *Nucleic Acids Res,* 45 (D1), D353-d61.
Katz, J., et al. (2011), 'Presence of Porphyromonas gingivalis in gingival squamous cell carcinoma', *Int J Oral Sci,* 3 (4), 209-15.
Ke, R., et al. (2016), 'Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences', *Hum Mutat,* 37 (12), 1363-67.

Keseler, I. M., et al. (2017), 'The EcoCyc database: reflecting new knowledge about Escherichia coli K-12', *Nucleic Acids Res,* 45 (D1), D543-d50.
Khorana, H. G., et al. (1966), 'Polynucleotide synthesis and the genetic code', *Cold Spring Harb Symp Quant Biol,* 31, 39-49.
Kinane, D. F., Stathopoulou, P. G., and Papapanou, P. N. (2017), 'Periodontal diseases', *Nat Rev Dis Primers,* 3, 17038.
Kircher, M. and Kelso, J. (2010), 'High-throughput DNA sequencing--concepts and limitations', *Bioessays,* 32 (6), 524-36.
Klaassen, C. H., van Haren, H. A., and Horrevorts, A. M. (2002), 'Molecular fingerprinting of Clostridium difficile isolates: pulsed-field gel electrophoresis versus amplified fragment length polymorphism', *J Clin Microbiol,* 40 (1), 101-4.
Koestler, T., von Haeseler, A., and Ebersberger, I. (2010), 'FACT: functional annotation transfer between proteins with similar feature architectures', *BMC Bioinformatics,* 11, 417.
Kohara, Y., Akiyama, K., and Isono, K. (1987), 'The physical map of the whole E. coli chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library', *Cell,* 50 (3), 495-508.
Kolbe, D. L. and Eddy, S. R. (2011), 'Fast filtering for RNA homology search', *Bioinformatics,* 27 (22), 3102-9.
Koliada, A., et al. (2017), 'Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population', *BMC Microbiol,* 17 (1), 120.
Kolmogorov, M., et al. (2014), 'Ragout-a reference-assisted assembly tool for bacterial genomes', *Bioinformatics,* 30 (12), i302-9.
Koonin, E. V. (2005), 'Orthologs, paralogs, and evolutionary genomics', *Annu Rev Genet,* 39, 309-38.
Koonin, E. V. and Galperin, M. Y. (1997), 'Prokaryotic genomes: the emerging paradigm of genome-based microbiology', *Curr Opin Genet Dev,* 7 (6), 757-63.
Kowalczuk, M., et al. (2001), 'DNA asymmetry and the replicational mutational pressure', *J Appl Genet,* 42 (4), 553-77.
Krieg, Noel R., et al. (2010), 'Phylum XIV. Bacteroidetes phyl. nov', in Noel R. Krieg, et al. (eds.), *Bergey's Manual® of Systematic Bacteriology: Volume Four The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes* (New York, NY: Springer New York), 25-469.
Kunin, V., et al. (2005), 'The net of life: reconstructing the microbial phylogenetic network', *Genome Res,* 15 (7), 954-9.
Kuzniar, A., et al. (2008), 'The quest for orthologs: finding the corresponding gene across genomes', *Trends Genet,* 24 (11), 539-51.
Kyrpides, N. C., et al. (2014), 'Genomic Encyclopedia of Type Strains, Phase I: The one thousand microbial genomes (KMG-I) project', *Stand Genomic Sci,* 9 (3), 1278-84.
Lagesen, K., et al. (2007), 'RNAmmer: consistent and rapid annotation of ribosomal RNA genes', *Nucleic Acids Res,* 35 (9), 3100-8.
Lamont, R. J., Hajishengallis, G. N., and Jenkinson, H. F. (2014), *Oral Microbiology and Immunology, Second Edition* (American Society of Microbiology).
Lamouille, S., Xu, J., and Derynck, R. (2014), 'Molecular mechanisms of epithelial-mesenchymal transition', *Nat Rev Mol Cell Biol,* 15 (3), 178-96.
Land, M., et al. (2015), 'Insights from 20 years of bacterial genome sequencing', *Funct Integr Genomics,* 15 (2), 141-61.
Lander, E. S., et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature,* 409 (6822), 860-921.
Lapage, S. P., et al. (1992), 'Appendix 10, Infrasubspecific Subdivisions.', in S. P. Lapage, et al. (eds.), *International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision* (Washington (DC): ASM Press

International Union of Microbiological Societies.).
Lasica, A. M., et al. (2017), 'The Type IX Secretion System (T9SS): Highlights and Recent Insights into Its Structure and Function', *Front Cell Infect Microbiol,* 7, 215.
Laslett, D. and Canback, B. (2004), 'ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences', *Nucleic Acids Res,* 32 (1), 11-6.
Leao, T., et al. (2017), 'Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea', *Proc Natl Acad Sci U S A,* 114 (12), 3198-203.
Lederberg, J. (2001), "Ome Sweet 'Omics-- A Genealogical Treasury of Words', *The Scientist*.
Lee, J. Y., et al. (2017), 'Comparative genomics of Lactobacillus salivarius strains focusing on their host adaptation', *Microbiol Res,* 205, 48-58.
Lehman, I. R., et al. (1958), 'Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli', *J Biol Chem,* 233 (1), 163-70.
Lemoine, F., Lespinet, O., and Labedan, B. (2007), 'Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data', *BMC Evol Biol,* 7, 237.
Lennon, J. T. and Jones, S. E. (2011), 'Microbial seed banks: the ecological and evolutionary implications of dormancy', *Nat Rev Microbiol,* 9 (2), 119-30.
Li, R., et al. (2010), 'De novo assembly of human genomes with massively parallel short read sequencing', *Genome Res,* 20 (2), 265-72.
Li, Z., et al. (2012), 'Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph', *Brief Funct Genomics,* 11 (1), 25-37.
Liao, Y. C., et al. (2015), 'MyPro: A seamless pipeline for automated prokaryotic genome assembly and annotation', *J Microbiol Methods,* 113, 72-4.
Lin, G., et al. (2016), 'VennPainter: A Tool for the Comparison and Identification of Candidate Genes Based on Venn Diagrams', *PLoS One,* 11 (4), e0154315.
Ling, V. (1972a), 'Fractionation and sequences of the large pyrimidine oligonucleotides from bacteriophage fd DNA', *J Mol Biol,* 64 (1), 87-102.
--- (1972b), 'Pyrimidine sequences from the DNA of bacteriophages fd, fl, and phi X174', *Proc Natl Acad Sci U S A,* 69 (3), 742-6.
Lipman, D. J. and Pearson, W. R. (1985), 'Rapid and sensitive protein similarity searches', *Science,* 227 (4693), 1435-41.
Liu, Y. and Burne, R. A. (2011), 'The major autolysin of Streptococcus gordonii is subject to complex regulation and modulates stress tolerance, biofilm formation, and extracellular-DNA release', *J Bacteriol,* 193 (11), 2826-37.
Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016), 'The healthy human microbiome', *Genome Med,* 8 (1), 51.
Lobry, J. R. and Louarn, J. M. (2003), 'Polarisation of prokaryotic chromosomes', *Curr Opin Microbiol,* 6 (2), 101-8.
Lowe, T. M. and Eddy, S. R. (1997), 'tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence', *Nucleic Acids Res,* 25 (5), 955-64.
Luckey, J. A., et al. (1990), 'High speed DNA sequencing by capillary electrophoresis', *Nucleic Acids Res,* 18 (15), 4417-21.
MacArthur, I., et al. (2017), 'Comparative Genomics of Rhodococcus equi Virulence Plasmids Indicates Host-Driven Evolution of the vap Pathogenicity Island', *Genome Biol Evol,* 9 (5), 1241-47.
Madoui, M. A., et al. (2016), 'MaGuS: a tool for quality assessment and scaffolding of genome assemblies with Whole Genome Profiling Data', *BMC Bioinformatics,* 17, 115.
Madoui, M. A., et al. (2015), 'Genome assembly using Nanopore-guided long and error-free DNA reads', *BMC Genomics,* 16, 327.
Magoc, T., et al. (2013), 'GAGE-B: an evaluation of genome assemblers for bacterial organisms', *Bioinformatics,* 29 (14), 1718-25.

Maiden, M. C. (2006), 'Multilocus sequence typing of bacteria', *Annu Rev Microbiol,* 60, 561-88.
Mainzer, K. (2004), 'Complex Systems and the Evolution of Computability', *Thinking in Complexity: The Computational Dynamics of Matter, Mind, and Mankind* (Berlin, Heidelberg: Springer Berlin Heidelberg), 187-239.
Mallet, J. (1995), 'A species definition for the modern synthesis', *Trends Ecol Evol,* 10 (7), 294-9.
Mallet, L., et al. (2017), 'PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies', *Bioinformatics*.
Mardis, E. R. (2008), 'Next-generation DNA sequencing methods', *Annu Rev Genomics Hum Genet,* 9, 387-402.
Margulies, M., et al. (2005), 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature,* 437 (7057), 376-80.
Mariano, D. C., et al. (2016), 'Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002', *BMC Genomics,* 17, 315.
Marino-Ramirez, L., et al. (2011), 'The Histone Database: an integrated resource for histones and histone fold-containing proteins', *Database (Oxford),* 2011, bar048.
Markowitz, V. M., et al. (2009), 'IMG ER: a system for microbial genome annotation expert review and curation', *Bioinformatics,* 25 (17), 2271-8.
Matthaei, J. H., et al. (1962), 'Characteristics and composition of RNA coding units', *Proc Natl Acad Sci U S A,* 48, 666-77.
Mavromatis, K., et al. (2012), 'The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation', *PLoS One,* 7 (12), e48837.
Maxam, A. M. and Gilbert, W. (1977), 'A new method for sequencing DNA', *Proc Natl Acad Sci U S A,* 74 (2), 560-4.
McBride, M. J. and Zhu, Y. (2013), 'Gliding motility and Por secretion system genes are widespread among members of the phylum bacteroidetes', *J Bacteriol,* 195 (2), 270-8.
McCouch, S. R. (2001), 'Genomics and synteny', *Plant Physiol,* 125 (1), 152-5.
McCutcheon, J. P., McDonald, B. R., and Moran, N. A. (2009), 'Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont', *PLoS Genet,* 5 (7), e1000565.
McPherson, J. D., et al. (2001), 'A physical map of the human genome', *Nature,* 409 (6822), 934-41.
Medini, D., et al. (2005), 'The microbial pan-genome', *Curr Opin Genet Dev,* 15 (6), 589-94.
Meier-Kolthoff, J. P., Klenk, H. P., and Goker, M. (2014), 'Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age', *Int J Syst Evol Microbiol,* 64 (Pt 2), 352-6.
Meselson, M. and Stahl, F. W. (1958), 'The replication of DNA', *Cold Spring Harb Symp Quant Biol,* 23, 9-12.
Messing, J., et al. (1977), 'Filamentous coliphage M13 as a cloning vehicle: insertion of a HindII fragment of the lac regulatory region in M13 replicative form in vitro', *Proc Natl Acad Sci U S A,* 74 (9), 3642-6.
Meyer, F., Overbeek, R., and Rodriguez, A. (2009), 'FIGfams: yet another set of protein families', *Nucleic Acids Res,* 37 (20), 6643-54.
Meyers, B. C., Scalabrin, S., and Morgante, M. (2004), 'Mapping and sequencing complex genomes: let's get physical!', *Nat Rev Genet,* 5 (8), 578-88.
Michna, R. H., et al. (2016), 'SubtiWiki 2.0--an integrated database for the model organism Bacillus subtilis', *Nucleic Acids Res,* 44 (D1), D654-62.
Milanowska, K., et al. (2011), 'REPAIRtoire--a database of DNA repair pathways', *Nucleic Acids Res,* 39 (Database issue), D788-92.
Miller, J. R., Koren, S., and Sutton, G. (2010), 'Assembly algorithms for next-generation sequencing data', *Genomics,* 95 (6), 315-27.

Morgan, X. C., Segata, N., and Huttenhower, C. (2013), 'Biodiversity and functional genomics in the human microbiome', *Trends Genet,* 29 (1), 51-8.
Moya, A., et al. (2008), 'Learning how to live together: genomic insights into prokaryote-animal symbioses', *Nat Rev Genet,* 9 (3), 218-29.
Mudgal, R., et al. (2015), 'De-DUFing the DUFs: Deciphering distant evolutionary relationships of Domains of Unknown Function using sensitive homology detection methods', *Biol Direct,* 10, 38.
Muller, F. M., et al. (1999), 'Standardized molecular typing', *Mycoses,* 42 Suppl 2, 69-72.
Muñoz, R., Rossello-Mora, R., and Amann, R. (2016), 'Revised phylogeny of Bacteroidetes and proposal of sixteen new taxa and two new combinations including Rhodothermaeota phyl. nov', *Syst Appl Microbiol,* 39 (5), 281-96.
Mushegian, A. R. and Koonin, E. V. (1996), 'Gene order is not conserved in bacterial evolution', *Trends Genet,* 12 (8), 289-90.
Myers, E. W. (2005), 'The fragment assembly string graph', *Bioinformatics,* 21 Suppl 2, ii79-85.
Myllykangas, Samuel, Buenrostro, Jason, and Ji, Hanlee P. (2012), 'Overview of Sequencing Technology Platforms', in Naiara Rodríguez-Ezpeleta, Michael Hackenberg, and Ana M. Aransay (eds.), *Bioinformatics for High Throughput Sequencing* (New York, NY: Springer New York), 11-25.
Nadalin, F., Vezzi, F., and Policriti, A. (2012), 'GapFiller: a de novo assembly approach to fill the gap within paired reads', *BMC Bioinformatics,* 13 Suppl 14, S8.
Nagarajan, N. and Pop, M. (2013), 'Sequence assembly demystified', *Nat Rev Genet,* 14 (3), 157-67.
Nagarajan, N., et al. (2010), 'Finishing genomes with limited resources: lessons from an ensemble of microbial genomes', *BMC Genomics,* 11, 242.
Nakagawa, S., Niimura, Y., and Gojobori, T. (2017), 'Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine-Dalgarno sequence in prokaryotes', *Nucleic Acids Res,* 45 (7), 3922-31.
Narzisi, G. and Mishra, B. (2011), 'Comparing de novo genome assembly: the long and short of it', *PLoS One,* 6 (4), e19175.
Nawy, T. (2014), 'In situ sequencing', *Nat Methods,* 11 (1), 29.
Needleman, S. B. and Wunsch, C. D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *J Mol Biol,* 48 (3), 443-53.
Nei, M. (2003), 'Genome evolution: let's stick together', *Heredity (Edinb),* 90 (6), 411-2.
Nirenberg, M. and Leder, P. (1964), 'RNA codewords and protein synthesis. The effect of trinucleotides upon the binding of sRNA to ribosomes', *Science,* 145 (3639), 1399-407.
Nirenberg, M., et al. (1966), 'The RNA code and protein synthesis', *Cold Spring Harb Symp Quant Biol,* 31, 11-24.
Nyren, P. (1987), 'Enzymatic method for continuous monitoring of DNA polymerase activity', *Anal Biochem,* 167 (2), 235-8.
O'Flynn, C., et al. (2015), 'Comparative Genomics of the Genus Porphyromonas Identifies Adaptations for Heme Synthesis within the Prevalent Canine Oral Species Porphyromonas cangingivalis', *Genome Biol Evol,* 7 (12), 3397-413.
Ochman, H. (2002), 'Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes', *Trends Genet,* 18 (7), 335-7.
Ochoa, S. (1964), 'The chemical basis of heredity--The genetic code', *Bull N Y Acad Med,* 40, 387-411.
Ogata, H., et al. (1999), 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Res,* 27 (1), 29-34.
Oishi, Y., et al. (2012), 'Purification and characterization of a novel secondary fimbrial protein from Porphyromonas gulae', *J Oral Microbiol,* 4.
Oliver, W. W. and Wherry, W. B. (1921), 'Notes on Some Bacterial Parasites of the Human Mucous Membranes', *J Infect Dis,* 28 (4), 341-44.

Orata, F. D., et al. (2015), 'The Dynamics of Genetic Interactions between Vibrio metoecus and Vibrio cholerae, Two Close Relatives Co-Occurring in the Environment', *Genome Biol Evol,* 7 (10), 2941-54.
Oren, A., et al. (2015), 'Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes', *Int J Syst Evol Microbiol,* 65 (11), 4284-87.
Overton, M. L. (2001), *Numerical computing with IEEE floating point arithmetic* (Society for Industrial and Applied Mathematics) 104.
Papanicolaou, A. (2016), 'The life cycle of a genome project: perspectives and guidelines inspired by insect genome projects', *F1000Res,* 5, 18.
Passarge, E., Horsthemke, B., and Farber, R. A. (1999), 'Incorrect use of the term synteny', *Nat Genet,* 23 (4), 387.
Paster, B. J., et al. (1985), 'A Phylogenetic Grouping of the Bacteroides, Cytophagas, and Certain Flavobacteria', *Syst Appl Microbiol,* 6 (1), 34-42.
Pawar, S. P., et al. (2012), 'Genome sequence of Janibacter hoylei MTCC8307, isolated from the stratospheric air', *J Bacteriol,* 194 (23), 6629-30.
Pearson, W. R. (2013), 'An introduction to sequence similarity ("homology") searching', *Curr Protoc Bioinformatics,* Chapter 3, Unit3.1.
Pearson, W. R. and Lipman, D. J. (1988), 'Improved tools for biological sequence comparison', *Proc Natl Acad Sci U S A,* 85 (8), 2444-8.
Peng, Y., et al. (2010), 'IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler', in Bonnie Berger (ed.), *Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010. Proceedings* (Berlin, Heidelberg: Springer Berlin Heidelberg), 426-40.
Perera, M., et al. (2016), 'Emerging role of bacteria in oral carcinogenesis: a review with special reference to perio-pathogenic bacteria', *J Oral Microbiol,* 8, 32762.
Pevzner, P. A., Tang, H., and Waterman, M. S. (2001), 'An Eulerian path approach to DNA fragment assembly', *Proc Natl Acad Sci U S A,* 98 (17), 9748-53.
Polachek, H. (1997), 'Before the ENIAC', *IEEE Ann. Hist. Comput.,* 19 (2), 25-30.
Pop, M. (2009), 'Genome assembly reborn: recent computational challenges', *Brief Bioinform,* 10 (4), 354-66.
Pop, M. and Salzberg, S. L. (2008), 'Bioinformatics challenges of new sequencing technology', *Trends Genet,* 24 (3), 142-9.
Popa, O., et al. (2011), 'Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes', *Genome Res,* 21 (4), 599-609.
Prentice, M. B. (2004), 'Bacterial comparative genomics', *Genome Biol,* 5 (8), 338.
Prober, J. M., et al. (1987), 'A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides', *Science,* 238 (4825), 336-41.
Reichenbach, Hans (2006), 'The Order Cytophagales', in Martin Dworkin, et al. (eds.), *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass* (New York, NY: Springer New York), 549-90.
Renelli, M., et al. (2004), 'DNA-containing membrane vesicles of Pseudomonas aeruginosa PAO1 and their genetic transformation potential', *Microbiology,* 150 (Pt 7), 2161-9.
Renwick, J. H. (1971), 'The mapping of human chromosomes', *Annu Rev Genet,* 5, 81-120.
Reynolds-Campbell, G., Nicholson, A., and Thoms-Rodriguez, C. A. (2017), 'Oral Bacterial Infections: Diagnosis and Management', *Dent Clin North Am,* 61 (2), 305-18.
Rhoads, A. and Au, K. F. (2015), 'PacBio Sequencing and Its Applications', *Genomics Proteomics Bioinformatics,* 13 (5), 278-89.
Richardson, E. J. and Watson, M. (2013), 'The automatic annotation of bacterial genomes', *Brief Bioinform,* 14 (1), 1-12.
Richter, D. C., Schuster, S. C., and Huson, D. H. (2007), 'OSLay: optimal syntenic layout of unfinished assemblies', *Bioinformatics,* 23 (13), 1573-9.
Richter, M. and Rossello-Mora, R. (2009), 'Shifting the genomic gold standard for the prokaryotic species definition', *Proc Natl Acad Sci U S A,* 106 (45), 19126-31.

Riley, A. B., Kim, D., and Hansen, A. K. (2017), 'Genome Sequence of "Candidatus Carsonella ruddii" Strain BC, a Nutritional Endosymbiont of Bactericera cockerelli', *Genome Announc,* 5 (17).
Risca, V. I. and Greenleaf, W. J. (2015), 'Beyond the Linear Genome: Paired-End Sequencing as a Biophysical Tool', *Trends Cell Biol,* 25 (12), 716-9.
Ristov, S. (2016), 'A Fast and Simple Pattern Matching with Hamming Distance on Large Alphabets', *J Comput Biol,* 23 (11), 874-76.
Roberts, R. J. (2004), 'Identifying protein function--a call for community action', *PLoS Biol,* 2 (3), E42.
Rocha, E. P. (2008), 'The organization of the bacterial genome', *Annu Rev Genet,* 42, 211-33.
Rocha, E. P. and Danchin, A. (2003), 'Essentiality, not expressiveness, drives gene-strand bias in bacteria', *Nat Genet,* 34 (4), 377-8.
Ronaghi, M., et al. (1996), 'Real-time DNA sequencing using detection of pyrophosphate release', *Anal Biochem,* 242 (1), 84-9.
Rothberg, J. M., et al. (2011), 'An integrated semiconductor device enabling non-optical genome sequencing', *Nature,* 475 (7356), 348-52.
Rouli, L., et al. (2015), 'The bacterial pangenome as a new tool for analysing pathogenic bacteria', *New Microbes New Infect,* 7, 72-85.
Rousk, J. and Bengtson, P. (2014), 'Microbial regulation of global biogeochemical cycles', *Front Microbiol,* 5, 103.
Sakamoto, M. and Ohkuma, M. (2013), 'Porphyromonas crevioricanis is an earlier heterotypic synonym of Porphyromonas cansulci and has priority', *Int J Syst Evol Microbiol,* 63 (Pt 2), 454-7.
Salmela, L., et al. (2017), 'Accurate self-correction of errors in long reads using de Bruijn graphs', *Bioinformatics,* 33 (6), 799-806.
Salyers, A. A., et al. (1995), 'Conjugative transposons: an unusual and diverse set of integrated gene transfer elements', *Microbiol Rev,* 59 (4), 579-90.
Salzberg, S. L. and Yorke, J. A. (2005), 'Beware of mis-assembled genomes', *Bioinformatics,* 21 (24), 4320-1.
Salzberg, S. L., et al. (1998), 'Microbial gene identification using interpolated Markov models', *Nucleic Acids Res,* 26 (2), 544-8.
Salzberg, S. L., et al. (2012), 'GAGE: A critical evaluation of genome assemblies and assembly algorithms', *Genome Res,* 22 (3), 557-67.
Samaranayake, L. and Matsubara, V. H. (2017), 'Normal Oral Flora and the Oral Ecosystem', *Dent Clin North Am,* 61 (2), 199-215.
Sammet, J. E. (1969), *Programming Languages: History and Fundamentals* (Englewood Cliffs, N.J.: Prentice-Hall).
Sampaio-Maia, B., et al. (2016), 'The Oral Microbiome in Health and Its Implication in Oral and Systemic Diseases', *Adv Appl Microbiol,* 97, 171-210.
Sanger, F. and Tuppy, H. (1951a), 'The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates', *Biochem J,* 49 (4), 463-81.
--- (1951b), 'The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates', *Biochem J,* 49 (4), 481-90.
Sanger, F. and Thompson, E. O. (1953a), 'The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates', *Biochem J,* 53 (3), 366-74.
--- (1953b), 'The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates', *Biochem J,* 53 (3), 353-66.
Sanger, F. and Coulson, A. R. (1975), 'A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase', *J Mol Biol,* 94 (3), 441-8.
Sanger, F., Thompson, E. O., and Kitai, R. (1955), 'The amide groups of insulin', *Biochem J,* 59 (3), 509-18.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977a), 'DNA sequencing with chain-terminating inhibitors', *Proc Natl Acad Sci U S A,* 74 (12), 5463-7.
Sanger, F., et al. (1973), 'Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA', *Proc Natl Acad Sci U S A,* 70 (4), 1209-13.
Sanger, F., et al. (1977b), 'Nucleotide sequence of bacteriophage phi X174 DNA', *Nature,* 265 (5596), 687-95.
Sasson, O., Kaplan, N., and Linial, M. (2006), 'Functional annotation prediction: all for one and one for all', *Protein Sci,* 15 (6), 1557-62.
Schena, M., et al. (1998), 'Microarrays: biotechnology's discovery platform for functional genomics', *Trends Biotechnol,* 16 (7), 301-6.
Schneeberger, K., et al. (2011), 'Reference-guided assembly of four diverse Arabidopsis thaliana genomes', *Proc Natl Acad Sci U S A,* 108 (25), 10249-54.
Schnoes, A. M., et al. (2009), 'Annotation error in public databases: misannotation of molecular function in enzyme superfamilies', *PLoS Comput Biol,* 5 (12), e1000605.
Seemann, T. (2014), 'Prokka: rapid prokaryotic genome annotation', *Bioinformatics,* 30 (14), 2068-9.
Segerman, B. (2012), 'The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories', *Front Cell Infect Microbiol,* 2, 116.
Shade, A. (2017), 'Diversity is the question, not the answer', *Isme j,* 11 (1), 1-6.
Shah, H. N. and Collins, M. D. (1988), 'Proposal for Reclassification of Bacteroides asaccharolyticus, Bacteroides gingivalis, and Bacteroides endodontalis in a New Genus, Porphyromonas', *Int J Syst Bacteriol,* 38 (1), 128-31.
Sharma, G. and Subramanian, S. (2017), 'Unravelling the Complete Genome of Archangium gephyra DSM 2261T and Evolutionary Insights into Myxobacterial Chitinases', *Genome Biol Evol,* 9 (5), 1304-11.
Shi, H., Xing, Y., and Mao, X. (2017), 'The little brown bat nuclear genome contains an entire mitochondrial genome: Real or artifact?', *Gene,* 629, 64-67.
Shine, J. and Dalgarno, L. (1974), 'The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites', *Proc Natl Acad Sci U S A,* 71 (4), 1342-6.
Shmatkov, A. M., et al. (1999), 'Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes', *Bioinformatics,* 15 (11), 874-86.
Siguier, P., et al. (2015), 'Everyman's Guide to Bacterial Insertion Sequences', *Microbiol Spectr,* 3 (2), Mdna3-0030-2014.
Simpson, J. T. and Durbin, R. (2012), 'Efficient de novo assembly of large genomes using compressed data structures', *Genome Res,* 22 (3), 549-56.
Singh, S. R. (2013), 'Gastric cancer stem cells: a novel therapeutic target', *Cancer Lett,* 338 (1), 110-9.
Smalley, D. J., Whiteley, M., and Conway, T. (2003), 'In search of the minimal Escherichia coli genome', *Trends Microbiol,* 11 (1), 6-8.
Smith, C. J., Rocha, E. R., and Paster, B. J. (2006), 'The Medically Important Bacteroides spp. in Health and Disease', in Martin Dworkin, et al. (eds.), *The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass* (New York, NY: Springer New York), 381-427.
Smith, D. R. (2017), 'Goodbye genome paper, hello genome report: the increasing popularity of 'genome announcements' and their impact on science', *Brief Funct Genomics,* 16 (3), 156-62.
Smith, H. O. and Birnstiel, M. L. (1976), 'A simple method for DNA restriction site mapping', *Nucleic Acids Res,* 3 (9), 2387-98.
Smith, J. M., et al. (1993), 'How clonal are bacteria?', *Proc Natl Acad Sci U S A,* 90 (10), 4384-8.
Smith, K. A. (2008), 'Laws, leaders, and legends of the modern National Library of Medicine', *J Med Libr Assoc,* 96 (2), 121-33.

Smith, L. M., et al. (1986), 'Fluorescence detection in automated DNA sequence analysis', *Nature,* 321 (6071), 674-9.
Smith, T. F. and Waterman, M. S. (1981), 'Identification of common molecular subsequences', *J Mol Biol,* 147 (1), 195-7.
Sneath, P. H. (1957a), 'Some thoughts on bacterial classification', *J Gen Microbiol,* 17 (1), 184-200.
--- (1957b), 'The application of computers to taxonomy', *J Gen Microbiol,* 17 (1), 201-26.
Socransky, S. S., et al. (1963), 'The microbiota of the gingival crevice area of man. I. Total microscopic and viable counts and counts of specific organisms', *Arch Oral Biol,* 8, 275-80.
Sogin, M. L., et al. (2006), 'Microbial diversity in the deep sea and the underexplored "rare biosphere"', *Proc Natl Acad Sci U S A,* 103 (32), 12115-20.
Sommer, D. D., et al. (2007), 'Minimus: a fast, lightweight genome assembler', *BMC Bioinformatics,* 8, 64.
Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997), 'Pfam: a comprehensive database of protein domain families based on seed alignments', *Proteins,* 28 (3), 405-20.
Spratt, B. G. and Maiden, M. C. (1999), 'Bacterial population genetics, evolution and epidemiology', *Philos Trans R Soc Lond B Biol Sci,* 354 (1384), 701-10.
Sridhar, J. and Rafi, Z. A. (2007), 'Small RNA identification in Enterobacteriaceae using synteny and genomic backbone retention', *Omics,* 11 (1), 74-99.
Stackebrandt, E. and Goebel, B. M. (1994), 'Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology', *Int J Syst Evol Microbiol,* 44 (4), 846-49.
Stackebrandt, E., et al. (2002), 'Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology', *Int J Syst Evol Microbiol,* 52 (Pt 3), 1043-7.
Stormo, G. D. (2009), 'An introduction to sequence similarity ("homology") searching', *Curr Protoc Bioinformatics,* Chapter 3, Unit 3.1 3.1.1-7.
Stothard, P. and Wishart, D. S. (2006), 'Automated bacterial genome analysis and annotation', *Curr Opin Microbiol,* 9 (5), 505-10.
Sume, S. S., et al. (2010), 'Epithelial to mesenchymal transition in gingival overgrowth', *Am J Pathol,* 177 (1), 208-18.
Sutter, M., et al. (2017), 'Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell', *Science,* 356 (6344), 1293-97.
Suyama, M. and Bork, P. (2001), 'Evolution of prokaryotic gene order: genome rearrangements in closely related species', *Trends Genet,* 17 (1), 10-3.
Sverdlov, E. and Azhikina, T. (2005), 'Primer Walking', *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd.).
Sztukowska, M. N., et al. (2016), 'Porphyromonas gingivalis initiates a mesenchymal-like transition through ZEB1 in gingival epithelial cells', *Cell Microbiol,* 18 (6), 844-58.
Takahashi, N. (2015), 'Oral Microbiome Metabolism: From "Who Are They?" to "What Are They Doing?"', *J Dent Res,* 94 (12), 1628-37.
Takuno, S., et al. (2012), 'Population genomics in bacteria: a case study of Staphylococcus aureus', *Mol Biol Evol,* 29 (2), 797-809.
Tang, H., et al. (2008), 'Synteny and collinearity in plant genomes', *Science,* 320 (5875), 486-8.
Tang, L. and Liu, S. L. (2012), 'The 3Cs provide a novel concept of bacterial species: messages from the genome as illustrated by Salmonella', *Antonie Van Leeuwenhoek,* 101 (1), 67-72.
Tatusov, R. L., et al. (2000), 'The COG database: a tool for genome-scale analysis of protein functions and evolution', *Nucleic Acids Res,* 28 (1), 33-6.
Tatusova, T., et al. (2016), 'NCBI prokaryotic genome annotation pipeline', *Nucleic Acids Res,* 44 (14), 6614-24.
Tettelin, H., et al. (2008), 'Comparative genomics: the bacterial pan-genome', *Curr Opin Microbiol,* 11 (5), 472-7.

Tettelin, H., et al. (2005), 'Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"', *Proc Natl Acad Sci U S A,* 102 (39), 13950-5.
Tetz, V. V. (2005), 'The pangenome concept: a unifying view of genetic information', *Med Sci Monit,* 11 (7), Hy24-9.
Thomas, F., et al. (2011), 'Environmental and gut bacteroidetes: the food connection', *Front Microbiol,* 2, 93.
Totaro, M. C., et al. (2013), 'Porphyromonas gingivalis and the pathogenesis of rheumatoid arthritis: analysis of various compartments including the synovial tissue', *Arthritis Res Ther,* 15 (3), R66.
Touchon, M. and Rocha, E. P. (2016), 'Coevolution of the Organization and Structure of Prokaryotic Genomes', *Cold Spring Harb Perspect Biol,* 8 (1), a018168.
Trapnell, C. and Salzberg, S. L. (2009), 'How to map billions of short reads onto genomes', *Nat Biotechnol,* 27 (5), 455-7.
Treangen, T. J. and Salzberg, S. L. (2011), 'Repetitive DNA and next-generation sequencing: computational challenges and solutions', *Nat Rev Genet,* 13 (1), 36-46.
Tribble, G. D., et al. (2012), 'Natural competence is a major mechanism for horizontal DNA transfer in the oral pathogen Porphyromonas gingivalis', *MBio,* 3 (1).
Tripp, H. J., et al. (2015), 'Toward a standard in structural genome annotation for prokaryotes', *Stand Genomic Sci,* 10, 45.
Tully, B. J., et al. (2017), '290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology', *PeerJ,* 5, e3558.
Turing, A. M. (1937), 'On Computable Numbers, with an Application to the Entscheidungsproblem', *P Lond Math Soc,* s2-42 (1), 230-65.
Turnbaugh, P. J. and Stintzi, A. (2011), 'Human health and disease in a microbial world', *Front Microbiol,* 2, 190.
Tusnady, G. E., Dosztanyi, Z., and Simon, I. (2005), 'PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank', *Nucleic Acids Res,* 33 (Database issue), D275-8.
Urwin, R. and Maiden, M. C. (2003), 'Multi-locus sequence typing: a tool for global epidemiology', *Trends Microbiol,* 11 (10), 479-87.
Utturkar, S. M., et al. (2017), 'A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies', *Front Microbiol,* 8, 1272.
Vallenet, D., et al. (2009), 'MicroScope: a platform for microbial genome annotation and comparative genomics', *Database (Oxford),* 2009, bap021.
Valouev, A., et al. (2008), 'A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning', *Genome Res,* 18 (7), 1051-63.
van der Waaij, D., Berghuis-de Vries, J. M., and Lekkerkerk-van der Wees, J. E. C. (1971), 'Colonization resistance of the digestive tract in conventional and antibiotic-treated mice', *J Hyg (Lond),* 69 (3), 405-11.
van Dijk, E. L., et al. (2014), 'Ten years of next-generation sequencing technology', *Trends Genet,* 30 (9), 418-26.
Venter, J. C., et al. (2001), 'The sequence of the human genome', *Science,* 291 (5507), 1304-51.
Vincent, A. T., et al. (2017), 'Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money', *J Microbiol Methods,* 138, 60-71.
Vingadassalom, D., et al. (2005), 'An unusual primary sigma factor in the Bacteroidetes phylum', *Mol Microbiol,* 56 (4), 888-902.
Vogel, C., Teichmann, S. A., and Pereira-Leal, J. (2005), 'The relationship between domain duplication and recombination', *J Mol Biol,* 346 (1), 355-65.
Vogel, T. M., et al. (2009), 'TerraGenome: a consortium for the sequencing of a soil metagenome', *Nat Rev Micro,* 7 (4), 252-52.
Voon, D. C., et al. (2013), 'EMT-induced stemness and tumorigenicity are fueled by the EGFR/Ras pathway', *PLoS One,* 8 (8), e70427.

Vos, M. and Didelot, X. (2009), 'A comparison of homologous recombination rates in bacteria and archaea', *Isme j,* 3 (2), 199-208.
Vos, P., et al. (1995), 'AFLP: a new technique for DNA fingerprinting', *Nucleic Acids Res,* 23 (21), 4407-14.
Wallis, G. (1999), 'DNA discovery', *Science,* 285 (5429), 837.
Warren, R. L., et al. (2007), 'Assembling millions of short DNA sequences using SSAKE', *Bioinformatics,* 23 (4), 500-1.
Watson, J. D. and Crick, F. H. (1953a), 'Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid', *Nature,* 171 (4356), 737-8.
--- (1953b), 'Genetical implications of the structure of deoxyribonucleic acid', *Nature,* 171 (4361), 964-7.
Wayne, L. G., et al. (1987), 'Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics', *Int J Syst Evol Microbiol,* 37 (4), 463-64.
Wexler, A. G. and Goodman, A. L. (2017), 'An insider's perspective: Bacteroides as a window into the microbiome', *Nat Microbiol,* 2, 17026.
Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998), 'Prokaryotes: the unseen majority', *Proc Natl Acad Sci U S A,* 95 (12), 6578-83.
Williams, D., et al. (2013), 'Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes', *BMC Genomics,* 14, 537.
Woese, C. R. (1987), 'Bacterial evolution', *Microbiol Rev,* 51 (2), 221-71.
Wolf, Y. I. and Koonin, E. V. (2012), 'A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes', *Genome Biol Evol,* 4 (12), 1286-94.
Wood, D. W., et al. (2001), 'The genome of the natural genetic engineer Agrobacterium tumefaciens C58', *Science,* 294 (5550), 2317-23.
Wu, C. H., et al. (2006), 'The Universal Protein Resource (UniProt): an expanding universe of protein information', *Nucleic Acids Res,* 34 (Database issue), D187-91.
Wu, J. and Xie, J. (2010), 'Hidden Markov model and its applications in motif findings', *Methods Mol Biol,* 620, 405-16.
Yang, X., Chockalingam, S. P., and Aluru, S. (2013), 'A survey of error-correction methods for next-generation sequencing', *Brief Bioinform,* 14 (1), 56-66.
Yang, X., et al. (2016), 'Analysis of pan-genome to identify the core genes and essential genes of Brucella spp', *Mol Genet Genomics,* 291 (2), 905-12.
Yao, Z., Weinberg, Z., and Ruzzo, W. L. (2006), 'CMfinder--a covariance model based RNA motif finding algorithm', *Bioinformatics,* 22 (4), 445-52.
Yeates, T. O., Crowley, C. S., and Tanaka, S. (2010), 'Bacterial microcompartment organelles: protein shell structure and evolution', *Annu Rev Biophys,* 39, 185-205.
Yeates, T. O., Jorda, J., and Bobik, T. A. (2013), 'The shells of BMC-type microcompartment organelles in bacteria', *J Mol Microbiol Biotechnol,* 23 (4-5), 290-9.
Yujian, L. and Bo, L. (2007), 'A normalized Levenshtein distance metric', *IEEE Trans Pattern Anal Mach Intell,* 29 (6), 1091-5.
Zamecnik, P. C., et al. (1956), 'Mechanism of incorporation of labeled amino acids into protein', *J Cell Physiol Suppl,* 47 (Suppl 1), 81-101.
Zerbino, D. R. and Birney, E. (2008), 'Velvet: algorithms for de novo short read assembly using de Bruijn graphs', *Genome Res,* 18 (5), 821-9.
Zetterstrom, R. (2006), 'The Nobel Prize in 2005 for the discovery of Helicobacter pylori: implications for child health', *Acta Paediatr,* 95 (1), 3-5.
Zhang, H. X., Li, S. J., and Zhou, H. Q. (2014), 'Evaluating the annotation of protein-coding genes in bacterial genomes: Chloroflexus aurantiacus strain J-10-fl and Natrinema sp J7-2 as case studies', *Genet Mol Res,* 13 (4), 10891-7.
Zhang, W., et al. (2011), 'A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies', *PLoS One,* 6 (3), e17915.
Zhang, X., et al. (2017), 'Comparative Genomics Unravels the Functional Roles of Co-occurring Acidophilic Bacteria in Bioleaching Heaps', *Front Microbiol,* 8, 790.
Ziff, E. B., Sedat, J. W., and Galibert, F. (1973), 'Determination of the nucleotide sequence of a fragment of bacteriophage phiX 174 DNA', *Nat New Biol,* 241 (106), 34-7.

Annexes

1. Poster et abstract présentés au 2nd International Conference on Porphyromonas gingivalis and Related Species in Oral and Systemic Diseases ; 23-25 juin 2015 à Londres, Royaume-Uni

Premier auteur et participation au congrès.

Programme disponible : http://www.dentistry.qmul.ac.uk/conferences/pglondon2015/

Poster and Selected Oral Presentation Abstracts

P-1

Study of subgingival microbiome in subjects with periodontal disease and HFE-related hereditary hemochromatosis

Luis A. Acuña Amador (1), Vincent Meuric, Sandrine Le Gall David (1), Emile Boyer (1), Martine Bonnaure-Mallet (1) and Frédérique Barloy-Hubler (2).

1. EA 1254, Université Rennes 1, Rennes, France

2. CNRS-UMR 6290, IGDR, Université de Rennes 1, Rennes, France.

Periodontitis is a chronic inflammatory disease related to bacterial infection. The disease involves the destruction of supporting periodontal tissue, ultimately resulting in tooth loss. Association with systemic diseases such as diabetes, cardiovascular diseases, rheumatoid arthritis or cancer has been found. In a recent study, we show that caucasian patients with HFE-related hereditary hemochromatosis (HFE-HH) have a highly prevalence of periodontitis. HFE-HH is an autosomal recessive disorder characterized by iron overload potentially damaging organs such as the liver, heart and pancreas.

To understand why the periodontal disease develops preferably in HFE-HH patients, we examined the subgingival bacterial biodiversity in HFE-HH periodontitis patients by sequencing V3-V4 regions of the 16S rRNA genes. Cleaned reads were taxonomically assigned using the RDP-classifier. Bacterial communities with independent healthy patients (no HFE-HH, nor periodontitis) as well as independent chronic periodontitis patients (no HFE-HH) from 5 published studies have been used as controls.

A total of 24 periodontal pocket microbial community samples were analyzed, yielded a combined total of 510,238 sequences across all samples (median count of 22,944, max. 30,863, min. 2040). The observed richness, alpha and beta diversity (Shannon and Simpson index) tend to significantly increase with the periodontal disease gravity, evaluate through the depth of subgingival pockets (4 and 5 mm = moderate, 6 mm= intermediate and >7 mm = severe). While interpersonal variability is usually observed, HFE-HH patients could be clustered into 5 microbial profiles only.

All controls and HFE-HH samples contains the same 5 major phylum (abundance >1%): Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria, Proteobacteria and Spirochaetes.

However, the distribution of these five phyla varied amongst health and periodontitis samples and we observed a significant reduction in the abundance of Actinobacteria in HFE-HH subgingival community (15 fold compared to healthy and 3 fold compared to periodontitis without HFE-HH samples). The disappearance of Actinobacteria population in hemochromatosis-related periodontitis is accompanied by an absence of TM7. TM7 is an obligate epibiont living in symbiosis with Actinomyces. This candidate division TM7 is described in oral healthy cavity at 1% of the whole oral bacterial community and, found increased in abundance in non-HFE periodontitis pockets. This change is accompanied by a significant increase in Bacteroidetes and Fusobacteria if compared with healthy and non-HFE samples whereas Proteobacteria population remained stable. At this stage of our analysis, our study shows that periodontitis is caused by ecological disturbances in subgingival communities and is also dependent with other chronic disease such as hemochromatosis.

Study of subgingival microbiome in subjects with periodontal disease and HFE-related hereditary hemochromatosis

Luis A Acuña-Amador [1], Vincent Meuric [1], Sandrine Le Gall-David [1], Émile Boyer [1], Martine Bonnaure-Mallet [1], Frédérique Barloy-Hubler [2]

1. EA 1254, Université Rennes 1 and CHU Rennes, France
2. CNRS-UMR 6290, IGDR, Université Rennes 1, France

Société Beaujolaise d'Odonto-Stomatologie · Université de Rennes 1 · CNRS · CHU Rennes

PgLONDON2015
23RD - 25TH JUNE 2015

INTRODUCTION

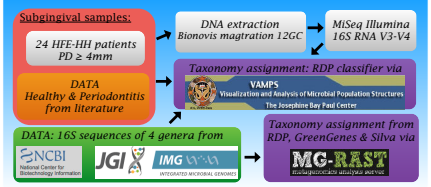
The subgingival biofilm contains between 10 to 16 phyla. A high proportion of Actinobacteria (A) and Firmicutes (F) is found in healthy biofilm while Bacteroidetes (B), Synergistetes (Sy) and Spirochetes (Sp) are the main phyla in periodontitis. NGS analysis confirms the association of three periopathogens, P. gingivalis (B), T. denticola (Sp) and T. forsythia (B) called the "red complex" and the newly described F. alocis (F).

In a recent study, we show that French patients with HFE-related hereditary hemochromatosis (HFE-HH) have a high prevalence of severe periodontitis. HFE-HH is an autosomal recessive disorder characterized by an inappropriate high absorption of iron in the gastrointestinal mucosa, potentially damaging organs (e.g. liver, heart and pancreas).

Our hypothesis is that the high iron burden in HFE-HH patients modify the oral microbiome to favor pathogenic species like P. gingivalis (B), leading to severe periodontitis.

In order to understand the links between periodontitis and hemochromatosis, we examined the subgingival bacterial biodiversity at the phylum level in HFE-HH periodontitis patients as well as in health and periodontitis. Then, we focused on Bacteroidetes and more precisely on P. gingivalis, testing 2 taxonomy assignment web tools with a dataset of known sequences publicly available at the NCBI and IMG websites.

METHODS



RESULTS

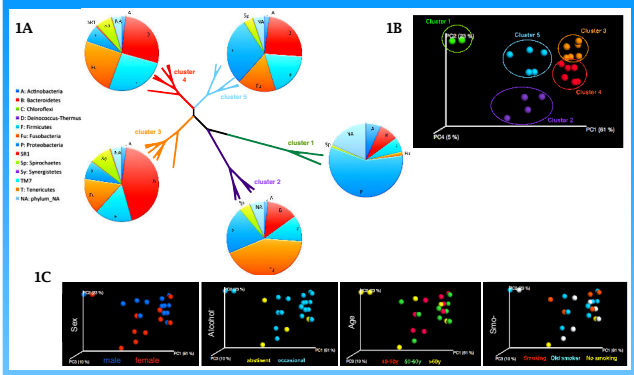


Figure 1 : *Unrooted tree displaying phylum Bray-Curtis beta diversity among patients with periodontitis associated to hemochromatosis* ***(fig. 1A)****. Principal coordinates analysis (PCoA) on Bray-Curtis distance matrices* ***(fig. 1B and 1C)*** *.*

24 HFE-HH patients were grouped in 5 clusters with intra-individual variation ***(fig. 1A and 1B)****.* ***Cluster 3****,* ***Cluster 4*** *and* ***Cluster 5*** *showed variable but significant proportion of Bacteroidetes (B, red) while* ***Cluster 2****, an enrichment of Fusobacteria (Fu, orange).* ***Cluster 1*** *(only 2 patients) shows an increase of Proteobacteria (P), Actinobacteria (A) and unclassified phyla (NA). Neither sex, alcohol consumption, smoking nor age explain the different microbiota observed* ***(fig. 1C)****.*
In order to determine if the HFE-HH condition had an impact on oral microbiota, we amplified the dataset to include both healthy and periodontitis (non-HFE-HH) samples ***(fig. 2)****.*

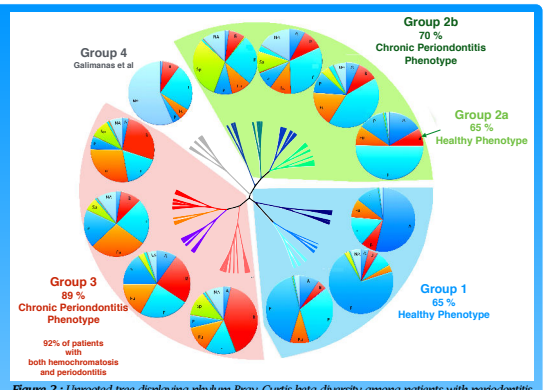


Figure 2 : *Unrooted tree displaying phylum Bray-Curtis beta diversity among patients with periodontitis associated to hemochromatosis (HP), healthy sub-gingival biofilm (H) and chronic periodontitis (CP).*

All H and CP data was taken from literature (5 studies, 242 samples). ***Group 1*** *consists mainly of healthy phenotypes and show microbiota with a high proportion of Proteobacteria (P) and Actinobacteria (A), while an increase of Bacteroidetes (B), Fusobacteria (Fu) and Spirochaetes (Sp) is found in the "chronic periodontitis"* ***Group 2b*** *and* ***Group 3****.* ***Group 2a*** *appears as an intermediate status with a "sick" microbiota but a "healthy" phenotype.* ***Group 3*** *includes almost all of the patients with hemochromatosis and may correspond to a "worsened" periodontitis, it showed a correlation between periodontitis/HFE-HH and the Bacteroidetes. Then, we needed to better define this phylum to describe shifts in genera/species within this taxonomic level. We chose 4 main genera to test if 16S/V3-V4 and VAMPS/MG-RAST are efficient in taxonomy assignment in this phylum* ***(fig. 3)****.*

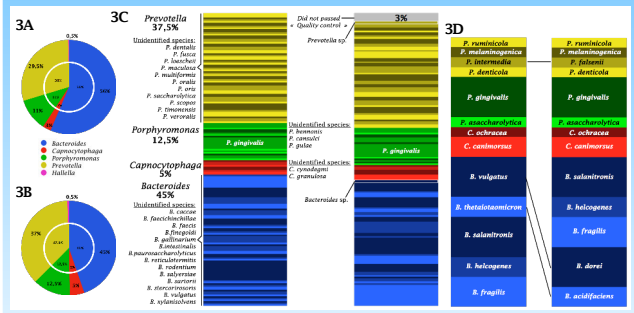


Figure 3 : *Distribution of taxonomy assignment using VAMPS (outer circle) vs. known taxonomy as annotated from databases (inner circle) in 387 whole 16S genes* ***(fig. 3A)*** *or in the 181 unique V3-V4 sequences* ***(fig. 3B)****. Those latter sequences (left bar) were identified with MG-RAST (right)* ***(fig. 3C)****. A simplification of data is presented : 27 V3-V4 sequences* ***(fig. 3D)****.*

VAMPS can robustly assign taxonomy when either the whole 16S rRNA or the V3-V4 portion of this gene are used. Only one sequence from the Prevotella genus is mistakenly identified as Hallella (both genera are from the same family) ***(fig. 3A and 3B)****. As VAMPS is limited to the genus level, we used another tool, MG-RAST, to assign the species level. Many species are not identified and some, although well described, were misidentified* ***(fig. 3C and 3D)****. This is due to either a wrong labeling in public databases or an incorrect identification by the web tool.*
For further analysis, P. gingivalis (the "keystone" periopathogen, well identified by both web tools) was chosen ***(fig. 4)****.*

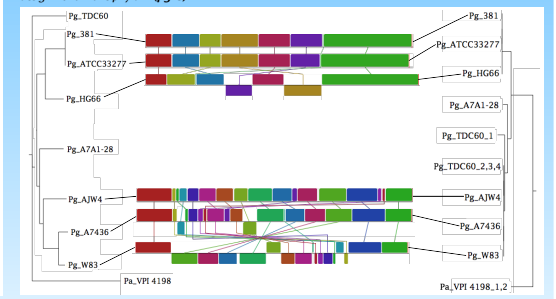


Figure 4 : *Phylogenetic tree of P. gingivalis whole genome (left) vs V3-V4 sequences (right) are shown with, in the middle, a MAUVE representation of recombination in selected genomes.*

A focus on 8 P. gingivalis and 1 P. asaccharolytica complete genomes was made. This analysis points out that homology in the V3-V4 16S gene sequence (as well as in the whole gene, data not shown) can result in discordant phylogeny when compared to whole genomes and that even if strains are evolutionarily close, they could be the result of different recombination history.
Description at the species level seems insufficient as our results of comparative genomics (ongoing) show important genomic variability that should be correlated to different periodontitis phenotypes.

CONCLUSIONS

This study highlights different microbial variations in periodontitis at the higher level in bacterial taxonomy: the phylum. These variations, regardless of the hemochromatosis status, show that the dysbiosis observed is not bimodal (sick or healthy) but rather progressive, resulting in shifts between phenotypic observations (presence or absence of gingival pockets) and the observed microbial communities. Various orotypes (defined as oral microbiome type by homology to enterotype) evolution paths may exist to explain the presence of healthy and periodontitis phenotypes in the same cluster. In order to study these variations, additional experiments must be conduced to evaluate the relationship between the periodicity of the disease and iron oral availability with microbial dysbiosis at the individual mouth scale during successive acute or remission phases.

Moreover, the future challenge will be the conception of more reliable tools and the identification of markers to study microbial communities in order to exploit microbiota data at the species and strain levels. Our current understanding may benefit from whole genome information and models taking into account recombination and other major sources of variation and evolution (as HGT) that might be useful to predict disease severity or outcome.

REFERENCES

1. **Abusleme, L. et al.** *The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation.* ***ISME J (2013)***
2. **Galimanas, V. et al.** *Bacterial community composition of chronic periodontitis and novel oral sampling sites for detecting disease indicators.* ***Microbiome (2014)***
3. **Griffen, A. L. et al.** *Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing.* ***ISME J (2012)***
4. **Huse, SM. et al.** *VAMPS: a website for visualization and analysis of microbial population structures.* ***BMC Bioinformatics (2014)***
5. **Kirst, M. E. et al.** *Dysbiosis and alterations in predicted functions of the subgingival microbiome in chronic periodontis.* ***Appl. Environ. Microbiol. (2015)***
6. **Meyer, F. et al.** *The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.* ***BMC Bioinformatics (2008)***
7. **Zhou, M. et al.** *Investigation of the effect of type 2 diabetes mellitus on subgingival plaque microbiota by high-throughput 16S rDNA pyrosequencing.* ***PLoS ONE (2013)***

2. Présentation orale au 4th Microbiome R&D and Business Collaboration Forum: Europe ; 3-4 avril 2017 à Amsterdam, Pays-Bas

Co-auteur dans les travaux présentés. Collaboration dans l'analyse des banques de reads V3-V4 du gène ARNr 16S pour décrire le microbiote.