Facultat de Ciències
Biològiques

Universitat
de València

Fundación para el Fomento de la
Investigación Sanitaria y Biomédica
de la Comunitat Valenciana

Área de Genómica
y Salud

OMICS APPROACHES TO STUDY THE ORAL MICROBIOME

Programa de Doctorado en Biotecnología

Doctorando

Pedro Belda Ferre

Director

Alejandro Mira Obrador

Valencia 2015

Back and front cover image: Anonymous ivory sculpture of a tooth from the 18th century, closed in the back cover, and opened in the front cover, representing a tooth infected by the “tooth-worm” and the pain and suffering it caused. This mythological figure was thought to cause dental caries, gum disease and toothache in general. Its origins come from the Sumerian civilization, where it was described in an ancient text as the cause of toothache.

Facultat de Ciències
Biològiques

VNIVERSITAT
DE VALÈNCIA

Fundación para el Fomento de la
Investigación Sanitaria y Biomédica
de la Comunitat Valenciana

Área de Genómica
y Salud

OMICS APPROACHES TO STUDY THE ORAL MICROBIOME

Programa de Doctorado en Biotecnología

Doctorando

Pedro Belda Ferre

Director

Alex Mira Obrador

Valencia 2015

Don ALEJANDRO MIRA OBRADOR, Doctor en Ciencias Biológicas e investigador del Área de Genómica y Salud de la Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana, INFORMA que esta memoria titulada "Omics approaches to study the oral microbiome" ha sido realizada por PEDRO BELDA FERRE bajo su dirección para optar al grado de Doctor Internacional por la Universidad de Valencia.

Y para que así conste, firma el presente certificado.

Valencia, a 11 de Mayo de 2015.

Alejandro Mira Obrador

Agradecimientos

Quiero agradecer en primer lugar a Alex Mira, el haberme dado la oportunidad de realizar contigo esta tesis doctoral. De ti he aprendido una infinidad de cosas, tanto a nivel científico como personal. Gracias de todo corazón. A Áurea, por haberse ofrecido de manera altruista a tomar las muestras para tantos y tantos trabajos incluso estando en el extranjero, y por las fructíferas discusiones de grupo en las que tanto hemos aprendido todos. A Arantxa, por brindarse siempre a revisar esta tesis y hacer que mejore con cada revisión. Gracias también al *"Equipo Cariátide"* y todos los que han ido pasando por él, sin vosotros esto hubiera sido imposible de hacer, y por supuesto mucho más aburrido.

No puedo dejar de agradecer a título personal a Luis David Alcaraz, Alfonso Benítez y James Williamson, por la estrecha colaboración que he tenido con ellos. Vosotros me habéis enseñado gran parte de lo aprendido en estos años. Por supuesto no hubiera aprendido tanto de vosotros sin las chelas, las Club Colombia ni las Friday Beers. ¡Gracias! Thanks!

A las salas de becarios 1 y 2, con vosotros he convivido durante estos años y habéis conseguido que ir al trabajo llegue a ser divertido. Gracias por vuestra ayuda en todos los momentos que la he necesitado y la habéis prestado hasta dejando vuestros asuntos de lado, eso no tiene precio. Ana Elena, Leo, Bea, Ana Dj, Ana D, Loles, Maria, Sandrine, Araceli, Anny, Pilar, Marta, Peris, Raúl, Jorge, Rodrigo, Marc y ese largo etcéra que formáis toda el Área de Genómica y Salud, habéis sido mi familia en este periodo, haciendo de estos años unos de los mejores en mi vida. ¡Gracias!

A Alex, Cheti, Monti, Pable, Roge y Sisard. Habéis sido mi válvula de escape desde hace ya tantos años, que ya uno ni se acuerda. Gracias por los cafés, acampadas, planes míticos a Pirineos, carreras, bicis, mudanzas, Maigmonas, homenajes varios, etc. Algún día saldrá el Mas de la Magra. A Marién, por tu incondicional amistad y por ser mi alma gemela. A Raúl, escudero de tantas y tantas batallas. A Loles, por abrirme las puertas de tu casa siempre que lo he necesitado.

A mi familia, por estar siempre a mi lado apoyándome en todas las decisiones que he tomado. Y por (al menos intentar) comprender esta aventura que ha sido hacer la tesis doctoral. Moltes gràcies Lipinoides!

Y por último, quiero agradecer a todas aquellas personas que han estado a mi lado durante esta tesis animándome a seguir adelante y compartiendo los mejores y peores momentos. Gracias por apoyarme de las mil y una manera en que lo habéis hecho.

¡¡¡GRACIAS A TODOS!!!

INDEX

INDEX

1

INTRODUCTION

1. INTRODUCTION

1.1. Insights into the human microbiome

Although unicellular organisms are thought to be one of the earliest forms of life, they were not discovered until 1677. A dutch draper merchant named Antony van Leeuwenhoek improved the existing microscope lenses, which at that time were only able to amplify up to 5 times the image. With his microscopes, some of them amplifying up to 275x, he discovered a new microscopic world. In his letters, he described the presence of small “animalcules” in human dental plaque samples in 1683 and was fascinated by the great variety of shapes and movements those “animalcules” showed. That was the first description of human-associated microbes. In the mid 1800s, Louis Pasteur discovered that bacteria present in wine and dairy products were the reason why they become sour (Pasteur 1857). He hypothesized that bacteria could also be the reason of human diseases, settling the “germ theory” of infectious diseases. He also started one of the most successful strategies in infectious disease prevention, vaccination. Pasteur found by chance that when a decaying culture of *Pastereulella multocida* was administered to chickens, it induced a mild version of chicken cholera. Those chickens could not be reinfected with a fresh culture of *P. multocida* afterwards, preventing them from suffering the virulent version of the disease. Robert Koch confirmed the germ theory in 1870, based on his experiments with anthrax, discovering its causal agent, *Bacillus anthracis*. This finding led him to formulate the “Koch's postulates”, used to establish the etiology of infectious diseases (Koch 1870). From this, a golden age came for the microbiology, leading to the identification of the etiological agents of many infectious diseases and thus, opening the possibility of halting epidemics by avoiding their spread. Etiological agent identification and vaccines development greatly contributed to the prevention of bacterial diseases such as diphteria, pertussis, tetanus, Q fever, typhus, cholera, etc. After the Second World War, antibiotics, discovered by Alexander Fleming, were produced industrially and applied to routine medicine, reducing the mortality of pneumonia, syphilis, tuberculosis, meningitis and other bacterial infectious diseases.

All those milestones achieved by microbiological sciences have affected the way we think about microorganisms. We tend to feel anxiety and fear when thinking about microbes, as if their only reason to exist was causing disease. But if we think about it, microbes have been living on Earth for at least 3.5 billion years, which is the estimated age of the oldest fossil of a living form (Woese & Gogarten 1999). All other kinds of life had to coexist in a world dominated by microbes. Nowadays, the number of bacterial cells in the world has been estimated to be around $5x10^{30}$ cells (Whitman et al. 1998), making bacteria the most common kind of organism on Earth (probably after virus[1]). Thus it seems unlikely that microbes have always caused disease. In contrast, specific mutualistic relationships must have been developed between humans and bacteria, and this seems to be the rule more than the exception. In the case of the human body, it is composed by around 10^{13} eukaryotic cells, but the fact is that they only account for 10% of the total number of cells, the rest being bacterial (Savage 1977). Our picture looks even more bacterial if we consider the number of bacterial genes, as they outnumber human ones by a hundred times (Gill et al. 2006). This overwhelming abundance of bacteria was first noted early in 1932 by Razumov, who observed that most bacteria he could view through the microscope, were not growing in pure culture (Razumov 1932). This fact was later termed as the "great plate count anomaly", showing for the first time the limitations of culture-based microbiological studies.

As a consequence of this close contact between bacteria and humans through time, symbiotic relationships have been developed, adapting to each other and co-evolving (Dubos et al. 1965, McFall-Ngai 2002). The emerging concept of humans as holobionts[2], is focusing research efforts towards the physiological benefits that microbiome[3] supply to their hosts. Some of the benefits recognized to the microbiota are the contribution to food digestion and nutrition (Wostmann 1981, Turnbaugh & Gordon 2009); the regulation of human metabolism (Qin et al. 2012, Ferrer et al. 2013); the maturation of the immune system (Lee & Mazmanian 2010, Chung et al. 2012); the prevention of colonization of host tissues by pathogens; the influence in xenobiotics processing and detoxification (Björkholm et al. 2009); and the epithelial development (Fons et al. 2000), among others. Thus disease may arise not only due to an overgrow of a single pathogen, but from the deterioration or loss of those beneficial effects proportioned by the whole microbiome under healthy conditions.

This fact changed the point of view of some infectious diseases. Given that a wide variety of microbes are found in healthy and diseased individuals, the etiological cause of those kinds of diseases cannot be attributed to any of the species isolated in diseased sites. Koch himself realized that soon after the publication of his postulates, which led to the

[1] The total number of bacteriophages has been estimated to outnumber in at least in an order of magnitude the number of bacterial cells (Williamson et al. 2008)

[2] Holobiont is a term coined by Lynn Margulis to denote any symbiotic association between individuals of different species for significant portions of their life history. All participants are bionts and the resulting organism is a holobiont.

[3] Microbial community that lives associated to a host, virtually inhabiting all body surfaces and cavities, either in a commensal, symbiotic or pathogenic relationship. The microbiome can be considered as part of the holobiont composed by both human being and its microbial community.

concept of “healthy carriers”. In addition, there are many diseases that cannot be attributed to single species (Peters et al. 2012). In these cases, the alteration of the microbial communities living in healthy sites (dysbiosis) can conduce to lose the healthy status and cause disease. Diseases that have been associated with an altered bacterial community include obesity (Turnbaugh & Gordon 2009, Turnbaugh et al. 2009, Ferrer et al. 2013), inflammatory bowel disease (Seksik 2010) , bacterial vaginosis (Ravel et al. 2013), dental caries (Marsh 2010), periodontitis (Kumar et al. 2006), etc, which are now being studied from a different point of view.

High-throughput molecular techniques were developed from late 90's, to overcome the limitation of the culturing step, allowing a better knowledge of those bacteria that are non retrievable by the standard culturing techniques. In consequence, the human microbiome has emerged as a complex bacterial community that virtually cover all surfaces and cavities of the human body, ranging from skin, digestive tract, reproductive and urinary system, respiratory system, eyes, etc. Even more surprising was the discovery of bacteria in parts of the body that until recently had always been considered sterile (Burcelin et al. 2013). For instance, bacteria have been found in placenta (Aagaard et al. 2014), human milk (Cabrera-Rubio et al. 2012), blood samples (Benítez-Páez et al. 2013) and atheroma plaque lesions (Kozarov et al. 2005, Koren et al. 2011). Thus, the assumed sterility of human tissues must be reconsidered, at least in some special circumstances.

Given the ubiquitous presence of bacteria even in the absence of disease, an important issue needs to be addressed, what is a healthy microbiome? The definition and discovery of what we can consider a healthy microbiota is an essential challenge for those diseases caused by a dysbiosis of the bacterial community, especially when they have slow development. Knowledge about the microbial community structure, their symbiotic and antagonistic relationships, as well as the metabolic functions of healthy-state microbiomes is fundamental for comparison with diseased-state microbiomes, allowing the understanding of the changes leading to the appearance of the pathology. Unfortunately, definition of what is a healthy state microbiota, is in many cases elusive, as some infectious diseases are only detected when clinical symptomatology is manifested. However, sub-clinical changes are happening before the onset of the first symptoms, which can be caused by microbial shifts towards a diseased community. This can lead to misconceptions about findings discovered by the comparison of sub-clinical diseased and clinically diseased microbiomes.

To confront this issue, there is a need to improve diagnosis of early sub-clinical stages of those diseases. This may be achieved with an increase in the number of healthy volunteers analyzed, so that sub-clinical diseased patients included by lack of diagnosis can be detected. Furthermore, long-term follow up studies of sub-clinical patients should confirm the disease outcome and also show which changes in the microbiome composition lead to the onset of the disease, or if changes in the microbiota are consequence of changes inherent to the disease. Today, the biggest efforts in that direction have been done by international consortia, such as the Human Microbiome Project[4] (HMP) and the META-HIT[5] consortium. META-HIT project has focused on the gut microbiome and its related diseases, such as obesity and irritable bowel disease. HMP has already analyzed using shot-gun metagenomics over 1500 samples from healthy subjects at 18 different body sites and three time points, comprising the largest human-associated bacterial gene catalog available today[6].

One of the most complex niches of the human microbiome is the oral cavity, given the huge diversity and the variety of sub-niches found on it. Among oral diseases, caries, gingivitis and [7]periodontitis are the most common ones caused by bacterial agents. They are caused by the formation and accumulation of dental plaque, which is a bacterial biofilm that grows over the clean tooth surface. Dental plaque is formed by bacterial and fungal cells, salivary glycoproteins, polysaccharides secreted by microbes and desquamated epithelial cells from gingival tissue (Mosby 2013). The supragingival dental plaque (supraGDP) accumulates over the tooth clean surface and it favors the growth of acidogenic and acidophilic bacteria. Subgingival dental plaque (subGDP) grows in the gingival sulcus, between the tooth and gingiva, is neutral or alkaline and is mainly composed of Gram negative bacteria. Caries is one of the most prevalent infectious diseases in the world and it has been estimated that 80% of the US population has suffered the disease(Petersen et al. 2005). Briefly, caries is produced when the pH over the tooth surface drops below a critical value, from which the minerals of the outermost layer of the teeth, the enamel, starts to dissolve. This pH decrease is mainly produced by fermentation of carbohydrates coming from the diet and glucoproteins from the host, and by the intake of acid beverages (sodas, carbonated drinks). Although caries has been traditionally associated to the presence of *Streptococcus mutans (Fitzgerald & Keyes 1960)* and *S. sobrinus* (Loesche 1986), there is a percentage of cases where those mutans streptococci (MS) have not been isolated from the lesion. Furthermore, some healthy individuals also carry in lower abundance *S. mutans* (Toi & Mogodiri 2000). Therefore, recent hypothesis have proposed that caries originates together with a dysbiotic state of the microbiota, where not only MS are responsible for acid production. In fact, the whole bacterial community ability to produce acids is now considered as the main cause. All

[4] http://commonfund.nih.gov/hmp/index

[5] http://www.metahit.eu/

[6] http://www.igs.umaryland.edu/doc/DACC_fin.pdf

[7] http://www.nidcr.nih.gov/DataStatistics/FindDataByTopic/DentalCaries/DentalCariesAdults20to64.htm#Table1

hypothesis proposed about the etiology of dental caries will be further discussed later on (see section 1.3.4 "Etiology of dental caries").

Through this thesis I will expand and discuss some of the microbiota concepts presented here, related to dental caries. The use of state-of-the-art high-throughput techniques, such as metagenomics (MTG), metatranscriptomics (MTT), metaproteomics (MTP), second-generation sequencing, bioinformatic analysis tools and statistical methods, have been applied to study the complex bacterial community of the supraGDP. Metagenomics works included in this thesis (Chapter 1 “The Oral Metagenome in Health and Disease” and Chapter 2 “Identifying the Healthy Oral Microbiome”) have intended to find out the differences in the microbial community composition between healthy and caries bearing individuals, together with the development of a newly described probiotic anti-caries strain, *Streptococcus dentisani* (Annex 1 “*Streptococcus dentisani* sp. nov. a new member of the Mitis group”), which is being tested for safety and industrial feasibility. Using metagenomics techniques and next generation sequencing (NGS), we proposed a new approach to identify virulence genes of pathogenic bacterial strains by comparing its genomes with a metagenome of a sample where similar non-pathogenic strains are usually found (Chapter 3 “Mining Virulence Genes Using Metagenomics”). Transcriptomic approaches have been applied to two different unsolved questions, the description of the bacterial colonization succession during the formation of the supraGDP under *in vivo* conditions, and the identification of the active bacteria after a carbohydrate-rich meal intake, when pH drops and enamel degradation takes place, trying to uncover those bacteria responsible for acidification (Chapter 4 “Microbiota Diversity and Gene Expression Dynamics in Human Oral Biofilms”). Metaproteomics has been applied to describe the protein composition of the supraGDP, under health and disease conditions, with the objective of finding putative biomarkers that will allow to differentiate healthy and diseased samples (Chapter 5 “The Human Oral Metaproteome reveals Potential Biomarkers for Caries Disease”).

1.2. Oral cavity

The oral cavity is the first entrance point to the digestive tract, whose main biological function is food selection and processing, before transit to the gastrointestinal tract. It is delimited by the lips, the cheeks, the palate, the tongue and floor of the mouth (Figure 1A). The palate is divided in hard palate, which separates the oral cavity from the nasal cavity, and the soft palate, separating oro- and naso-pharynx. The gingiva is a soft tissue that surrounds the base of the teeth and the maxillar and mandible bones. From the sockets of the alveolar bone, emerge a total of 20 deciduous teeth in children (2 incisors, 1 canine and 2 premolars per quadrant[8]) and 32 permanent teeth in adults (2 incisors, 1 canine, 2 premolars, 2 molars per quadrant, and in most individuals 1 wisdom tooth) (Figure 1B).

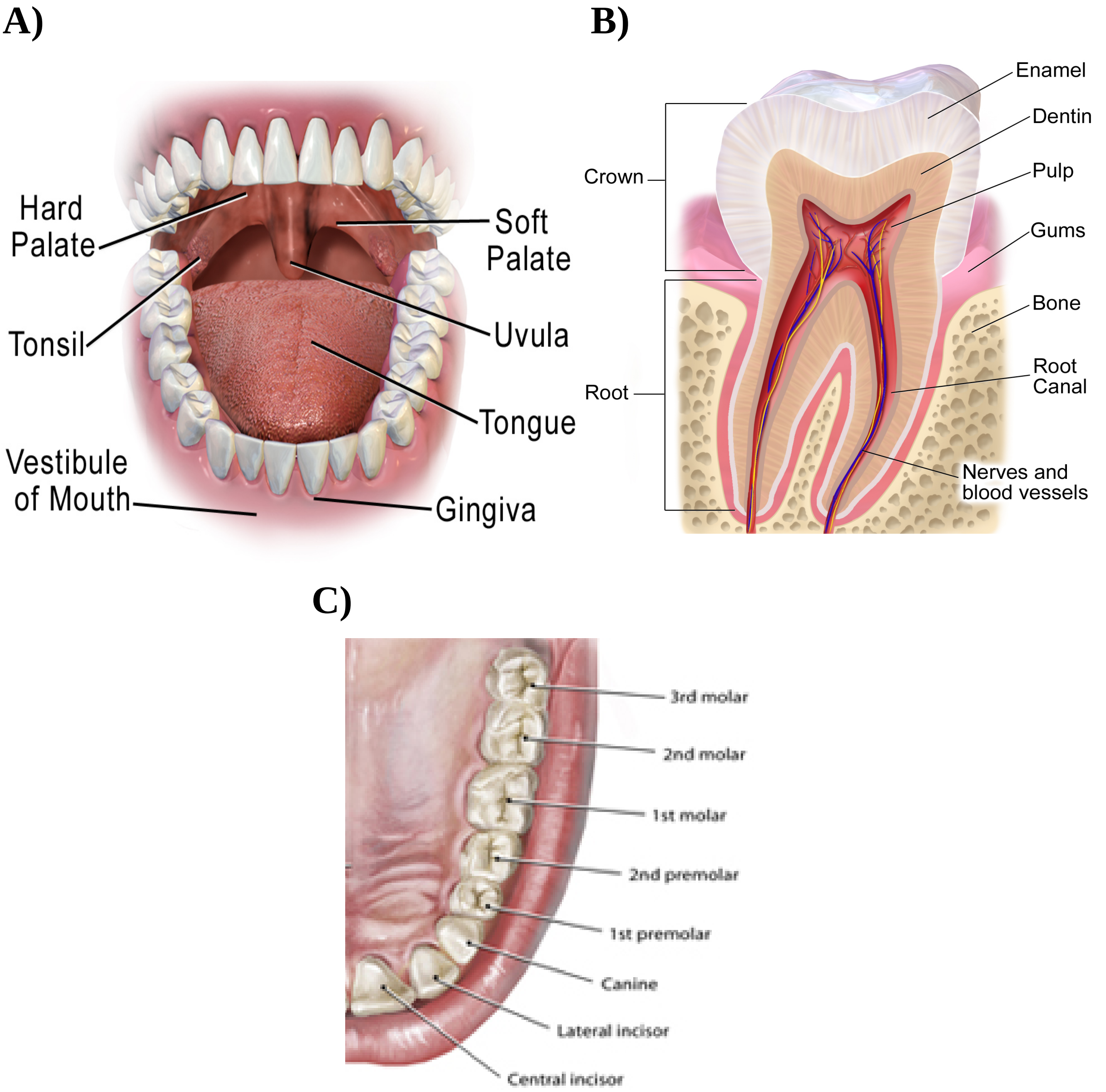


Figure 1. Oral cavity anterior view (A), teeth structure details (B) and detail of upper-right jaw (C). Adapted from (Blausen 2014)

All those mucosal structures are covered by three different types of epithelial tissue. In the case of the gingiva and hard palate, they are covered by keratinized, stratified and

[8] The teeth arcades are usually divided into 4 quadrants, starting from the patient's upper right and counting clockwise to the remaining quadrants. The first quadrant includes the patient's upper right incisor up to the upper right third molar.

squamous mucosa, in order to confer adequate resistance to masticatory forces. The tongue's side and upper surfaces are covered by filiform papillae, which has keratinized, stratified and squamous epithelium, whereas the epithelium between the papillae is non-keratinized, providing extensibility and flexibility. Soft palate, floor of the mouth, cheek, lower side of the tongue and the inside of the lips are all covered by non-keratinized, stratified and squamous epithelium, whose main function is lining and does not require special qualities.

Teeth are mineralized structures, anatomically divided in two parts, the crown and the root (Figure 1C). The crown is the visible part of teeth, and the root is inserted in the alveolar bone. Teeth are anchored to the alveolar bone at the root of the tooth, by the periodontal ligaments (cementum, alveolar bone and periodontal ligaments, comprise the supporting tissue, also known as periodontium). The upper part of the bone and the root of the tooth is covered with keratinized gingival tissue. The gingiva extends a short distance into the socket in the alveolar bone, creating a small depression (the gingival crevice) around the tooth, of no more than 2 mm when no gingival disease is present. The gingival tissue in the gingival crevice is called junctional epitelium (JE) and is not keratinized. It has an increased permeability, which facilitates the continuous flow of the gingival crevicular fluid (GCF), a serum-like fluid that baths the gingival crevice and exudates outside it.

The outer-most layer of the teeth crown is the enamel. It is 96% made of inorganic material, the rest being organic matter and water. The enamel is the hardest material in the human body and it is aimed to protect the tooth from the chewing weathering and from acids in the diet. The inorganic part of the enamel is mainly composed by hydroxyapatite (HAP) and fluoroapatite (when fluoride is added to drinking water, toothpastes or as diet supplement). The cementum, is the layer that covers the teeth root. It is composed by 45% of inorganic material (HAP), 33% organic material (collagen from the inserted periodontal ligaments) and 22% of water. Cementum's role is to serve as anchor to the periodontal ligaments, fixing the tooth to the alveolar bone.

Dentin is the substance between the enamel or cementum and the pulp chamber. Dentin is less mineralized than enamel (70%) and contains dentinal tubules, which are host to a matrix of collagenous proteins (20%) and cell processes of odontoblasts, the rest being water (for a better review on dentin composition and structure see (Goldberg et al. 2011)). Although degradation of dentine is faster than enamel given its lower mineral content, it also plays an important role in protection, enamel support and mitigating the pressure exerted on the crown by chewing. In fact, the 200 µm of dentine in contact with enamel is less mineralized, lacks dentinal tubules and is less hard than the rest of dentine, allowing a better dissipation of chewing pressure (Wang & Weiner 1997). Although it is not a vascularized tissue, dentinal tubules contain fluids and collagenous proteins. This particular structure of dentinal tissue can accelerate the progression of tooth decay, as acids can freely flow through them, but also due to the availability of proteins in the tubules that may allow faster bacterial growth. Those tubules are organized radially from the pulp chamber to the dentinoenamel junction (DEJ), reducing its inner diameter from 2.5 µm near the pulp up to 900 nm in the DEJ (Nanci 2008). Dentin is continuously formed throughout the whole life, therefore the increasing pressure inside the teeth bends the dentinal tubes in S-curves shape around the withdrawing pulp.

The pulp is the central part of the tooth and is the only vascularized and innervated part of the teeth. It is composed by connective tissue and hosts the cellular bodies of the dentinoblasts, whose cells processes extend through dentinal tissue, as well as fibroblasts, mesenchymal pluripotential cells, macrophages, granulocytes, mast and plasma cells. Nervous terminations in the pulp monitor dentinal damage, sensing different aggression signals, such as high masticatory pressure, dentinal caries or traumas.

Another key players in the oral cavity are the salivary glands' secretions, that are continuously being produced. Saliva is an aqueous fluid which has small amounts of other compounds diluted, such as mucus, glycoproteins, electrolytes, enzymes and antibacterial compounds (defensins, cathelicidins, lysozyme, Ig...) (Amerongen & Veerman 2002). The main functions of saliva are digestion of fat and starch, lubrication of mucossal surfaces and food for deglution, temperature and humidity regulation, defense against infections, buffering of pH variations and control of demineralization-remineralization balance of teeth. The submandibular and parotid salivary glands produce 90% of the daily saliva amount (750-1300 ml), but other glands also contribute to the production of saliva. All surfaces exposed to saliva, acquire a thin salivary layer over them of around 8-40 µm, termed the acquired enamel pellicle (AEP). Glucoproteins found in the AEP are used by microorganisms to adhere to the clean tooth surface, avoid their clearance by saliva swallowing and serve as an initial anchor for biofilm development (Jenkinson & Lamont 2005).

1.2.1. Oral microbial communities

Microbial life in the oral cavity has to face off the defense systems that the human body uses against microbial invasion. The most evident is the continuous clearance of free living microbes by the swallowing of saliva and mechanical forces exerted by chewing and tongue movements. In fact, bacteria found in the saliva are not formally considered oral inhabitants, as they are continuously being swallowed and have no time to grow and reproduce. Hence adhesion is a critical capacity that has been acquired by the oral microbiota to survive. Virtually all surfaces of the mouth are susceptible to colonization by microbial biofilms. All bacteria living in the oral cavity must possess the ability to adhere to solid surfaces coated with salivary pellicles (e.g. teeth surface), to desquamating epithelium or to bacteria that are already attached to a surface. Biofilms growing over mucosal sites do not reach the same thickness and complexity, except tongue dorsum, as in the case of dental plaque. This is due to the continuous desquamation process of the epithelium removing the outer-most layers, which facilitates the clearance of bacteria growing on top of the tissue. Furthermore, mucosal tissues have access to more immune components (mucosa-associated lymphoid tissue (Holmgren & Czerkinsky 2005)) than the teeth surface, which is inert and whose main immune defense is given through salivary and GCF antimicrobial components.

Given the wide variety of conditions that are present in the different parts of the mouth, microbiota composition changes in each of those micro-niches (Zaura et al. 2009, Segata et al. 2012, Simón-Soro et al. 2013a). In the next sections, I will describe the main characteristics of the different oral micro-niches and their bacterial composition as known today.

1.2.1.1. Saliva

Saliva is a key factor in the normal physiology of the oral cavity. It baths all the oral surfaces providing them with immune proteins for defense, calcium salts to help remineralization of teeth, pH-buffering salts, etc. But given the reduced time it stays in the mouth by continuous swallowing, there is not too much time for microbes to grow in it. The microbes found in saliva can be then considered in transit through the mouth. Salivary fluids collect all cells dislodging from any other surface, carrying with them the biofilm pieces that were attached to them (supraGDP, subGDP, mucosa-associated biofilms, etc). Because of those reasons, the salivary microbiome is considered to be a mixture of all the surfaces it is in contact with. Recent studies using 16S rRNA gene amplicon sequencing have stated that saliva's microbiome, specially when stimulated saliva is collected, is more closely related to the tongue's than to any other site in the mouth (Zaura et al. 2009, Segata et al. 2012, Simón-Soro et al. 2013a), as it is the shedding surface with a thicker biofilm and thus contributes with more bacteria to the composition of the salivary microbiome. This is of important consideration when looking for microbial effectors of the etiology of oral diseases occurring elsewhere than the tongue, as the potential bacterial or molecular agents may be masked by their relative dilution compared to the diseases' natural location (e.g. teeth surface for caries, gingival crevice for periodontitis...).

In addition, differences in the sampling procedure of saliva samples greatly affects the microbial composition of the sample. Saliva has been collected in several different ways, mainly unstimulated saliva by drooling, paraffin-stimulated saliva, active spitting, cotton swab, oral rinses with sterile solutions and paper tips), all of them recovering different bacterial communities, which are not equivalent. In an unpublished experiment conducted in our group, saliva samples were taken with those 6 methods from the same healthy individual, plus supraGDP samples. 16S rRNA gene amplicon sequencing showed that the bacterial communities recovered by saliva were highly variable, depending on the sampling procedure, which resulted in saliva samples from the same individual not clustering together (Figure 2). Saliva samples clustered in a different group from plaque samples, and reflected a higher variability than supraGDP samples. This indicates the high variability introduced by different saliva sampling methodologies, as there are many factors than cannot be controlled by the clinician (i.e. tongue movements to stimulate saliva production, temporal variation of the saliva production and composition, exposure to external stimuli that promote salivation, etc).

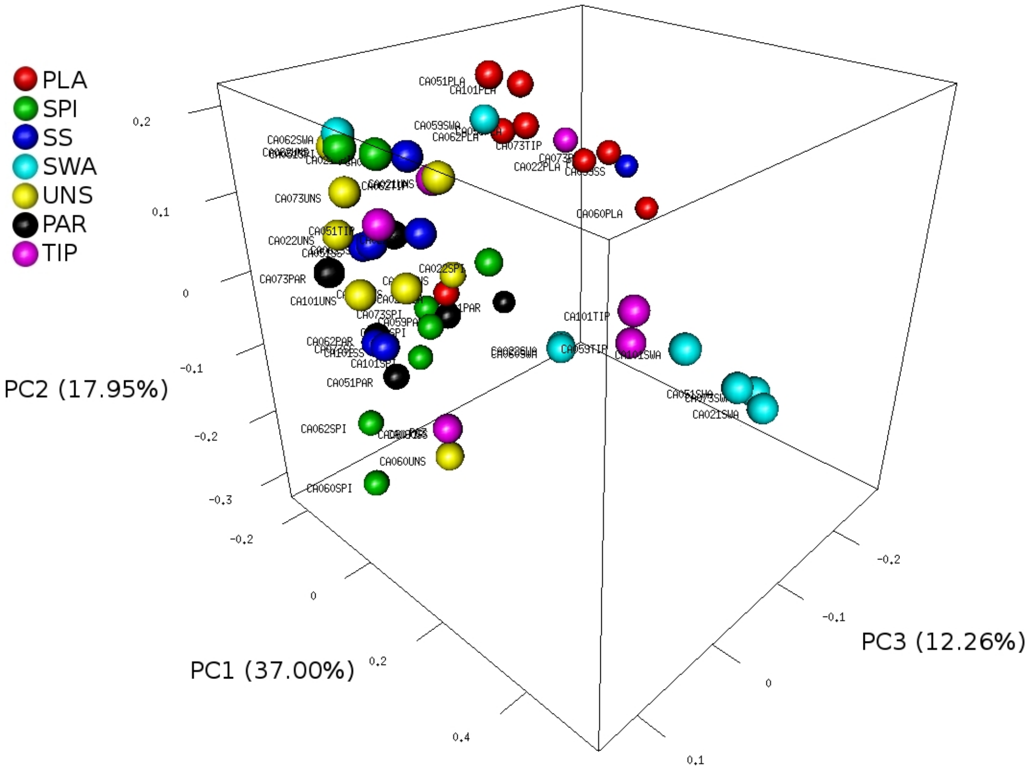


Figure 2. PCA analysis of 16S rRNA sequences amplified from supraGDP (PLA), spitted saliva (SPI), oral rinse with saline solution (SS), oral swab (SWA), unstimulated saliva (UNS), paraffin-stimulated saliva (PAR) and paper tips (TIP) (J Jorissën et al., unpublished).

Even though this highly variation in saliva composition, its microbiome has been widely used to associate its bacterial composition to health status (Streckfus & Bigler 2002). Saliva analysis is still of interest for screening of those diseases that would require a biopsy to reach the disease's site (i.e. cancer), or in epidemiological studies, where the vast number of volunteers needed would dramatically increase the need of specialized clinicians. In fact, the same occurs with the most commonly used body fluid, the blood, as it just recovers the outcome of many tissues. Saliva has been proposed to be a diagnostic fluid suitable for different systemic diseases. In the case of the dental plaque-derived oral diseases, caries, gingivitis and periodontitis, there have been many efforts to predict disease-risk by measuring pathogens related with them. For example, risk of suffering caries has been traditionally associated with the presence and abundance of MS *(S. mutans, S. sobrinus, S. cricetus* and *S. rattus)* and *Lactobacillus acidophilus* (Loesche 1986, van Houte 1994, Liljemark & Bloomquist 1996). Thus several diagnostic kits measuring those species levels in saliva have been developed (Dentocult SM (Orion, Finland), CRT (Vivadent, Liechtenstein), Cario Check SM (Sunstar, Japan) and Saliva-Check SM (GC, Japan)), but none of them have been successful in predicting appearance of caries. Therefore they are mainly used for educational purposes or for preventive reasons in association with other tests or information, as in the CAMBRA method (Steinberg 2009).

Regarding the microbial composition of saliva, its most abundant inhabitants are *Prevotella*, *Streptococcus*, *Veillonella* and *Neisseria* (Zaura et al. 2009, Yang et al. 2011, Segata et al. 2012, Simón-Soro et al. 2013a, Gomar-Vercher et al. 2014), As mentioned before, relative amounts of those genera varied depending on the saliva sampling method chosen. The total species number in saliva also varies greatly between different studies, ranging from 160 to 1400 species-level Operational Taxonomic Units (OTUs)[9].

1.2.1.2. Mucosa-associated microbial communities

Although oral mucosal surfaces are continuously shedding its outer-most layer of cells, they are usually covered by bacterial and fungal cells. There is still some controversy about their way of life, as it is not clear if they just attach to the surface or if they actually form biofilm structures (Dongari-Bagtzoglou 2008). A biofilm is considered as a well structured microbial community immersed in an extracellular polymeric substance (EPS), which has the capacity to adhere to a surface. In contrast to planktonic cells, biofilms are far too difficult to study, as it is still challenging to grow them under *in vitro* conditions, counting the number of cells, performing metabolic assays or examining with traditional microscopy

[9] Operational taxonomic unit, species distinction in microbiology. Typically using rDNA and a percent similarity threshold for classifying microbes within the same, or different, OTUs (Wooley et al. 2010).

techniques. In the case of oral mucosal biofilms, only some pathologic infections of *Candida albicans* have been proven to be biofilms *(Ganguly & Mitchell 2011)*, in association with other commensal species, and on the tongue dorsum. But given the difficulties to study mucosal biofilms and their higher complexity, it needs to be further investigated.

As mentioned before, tongue dorsum has certain particularities that favor the accumulation of microbial biofilms. Its surface morphology is full of fissures and grooves, enabling the retention of bacterial cells and food debris. This particular conditions facilitates the proliferation of anaerobic bacteria, whose respiration process yields volatile sulfur compounds (VSC) and aromatic compounds (indol and skatole), responsible for oral halitosis (De Boever & Loesche 1995, Roldán et al. 2003).

Among the microbial inhabitants found in the tongue, it is common to find bacterial inhabitants of other mouth niches, such as subgingival bacteria. Periodontal pathogens such as *Porphyromonas gingivalis*, *Prevotella intermedia*, *Aggregatibacter actinomycetemcomitans*, *Eikenella corrodens* and oral spirochetes are usually isolated from the tongue, and it has been proposed as a reservoir for bacterial recolonization after periodontal treatment (Roldán et al. 2003). Environmental conditions in the papillae crypts may resemble those in the subGDP, as low redox conditions and mucosal exudate similar to crevicular fluid may be present. On the other hand, tongue microbiota has access to salivary glycoproteins and its buffering effect, in contrast with subGDP. But probably the most determinant environmental condition is the shedding surface of the tongue mucosa, as its microbial composition resembles to the tonsils, throat and saliva (where all the shedded cells are suspended) (Segata et al. 2012). The most abundant inhabitants found in the tongue are *Streptococcus*, *Haemophilus*, *Prevotella*, *Veillonella*, *Moraxella*, *Fusobacterium* and *Actinomyces* (Huttenhower et al. 2012, Segata et al. 2012).

1.2.1.3. Dental plaque as a biofilm

Dental plaque is a complex microbial biofilm that adheres to clean surface of teeth. Teeth comprise the only body part of the human body that lacks a regulated system for shedding exposed surfaces. This makes the dental plaque a preferential location for complex biofilm development. Dental plaque is divided in two ecologically different parts, subGDP and supraGDP. SubGDP is formed in the gingival crevice, in-between the teeth and the gingiva covering it. Its accumulation is related with the appearance of gingivitis, gum bleeding and periodontits, causing gingiva recession and alveolar bone loss, destabilizing the tooth supporting tissues. This at the end can eventually cause the loss of dental pieces and other systemic complications (Koren et al. 2011, Aagaard et al. 2014), mainly related to the inflammatory response it triggers. SupraGDP develops over the clean surface of teeth, in the zone emerged from the gingiva. SupraGDP is associated with the development of dental caries, and it will be further treated later on this thesis.

Both SubGDP and SupraGDP are the most commonly studied biofilms. Its development start with the formation of a salivary pellicle over the clean surface of teeth, called acquired enamel pellicle (AEP). This pellicle is mainly formed by salivary glycoproteins (statherin, mucins, proline-rich proteins, IgA, cystatins, lysozime and lactoferrin), lipids and degradation products from dead human and bacterial cells (Al-Hashimi & Levine 1989). As bacteria come close to the teeth surface, weak van der Wall's forces attracts them to the AEP. Then specific bacterial proteins called adhesins recognize different epitopes of the proteins in the AEP, forming covalent bonds and binding tightly. Those initial colonizers are typically aerobic Gram-positive, such as *Streptococcus* and *Actinomyces*, but also aerobic Gram-negative, such as *Eikenella* and *Neisseria* (Li et al. 2004). They start to grow as microcolonies over the tooth. Secretion of polysaccharides enhances adhesion, creating a matrix around the initial biofilm that prevents bacteria from being detached. Furthermore, those cells lacking the ability to adhere to the salivary pellicle, are able to coadhere to at least another partner of the normal inhabitants of biofilms (Kolenbrander & London 1993). The biofilm is at the beginning an aerobic environment, with high redox potential and neutral pH, given the continuous access to oxygen and the buffering capacity of saliva.

As the biofilm continues growing, environmental conditions change inside it. Oxygen is rapidly depleted and CO_2 is generated, creating a microaerophilic or even anaerobic atmosphere in the inner layers. Redox potential is also reduced. Saliva finds it difficult to access to those layers, reducing its buffering effect. Thus pH is able to drop to lower values and stay at low pH for longer periods of time. Host cells and macromolecules are degraded and metabolic end products are secreted and retained in the biofilm matrix, increasing the range of nutrients available for exploiting. All those changes create environmental gradients in very short space, allowing the appearance of new microniches inside the biofilm matrix. When those new conditions appear, growth of late colonizers is enabled. They are typically Gram-negative rod-shaped bacteria, that require anaerobic or microaerophilic conditions to grow. *Fusobacterium nucleatum* is a critical player in scaffolding interactions between early and late colonizers (Kolenbrander et al. 2002), as it is able to co-adhere with many different species (Kolenbrander et al. 1989).

Although the process is quite well established under *in vitro* conditions, the wide variety of environmental conditions affecting *in vivo* biofilm formation have added extra difficulties to the study of this process. First, different exposure to mechanical forces throughout the teeth surfaces greatly affects the thickness, complexity and microbial composition of the biofilm that is formed at different sites (Simón-Soro et al. 2013a, Zaura et al. 2009). For instance, stagnation sites can clearly be seen in the preferential sites of biofilm accumulation, mainly gingival borders, crypts in the occlusal surface and interproximal surfaces. There is a gradient in the amount of dental plaque accumulated from the incisal region (low abundance) to the cervical region (higher amount of plaque). This differences in thickness of the biofilm, may impose different environmental conditions and thus the biofilm may not be homogeneous along this small surface. Another bias introduced by *in vivo* conditions is the difficulty to completely remove all bacteria from the dental surface, so the characterization of the first bacteria attaching to the AEP or those present in deep fissures may be difficult. Furthermore, it is impossible to add sequentially bacteria in order to know the colonization pattern of dental plaque. In fact, early and late colonizer bacteria will be attached, although they lack the proper conditions to grow under initial biofilm conditions.

Furthermore, early image studies have observed the presence of epithelial cells carrying bacterial cells in early dental plaque (Brecx et al. 1981). This may alter the sequential colonization order discovered under *in vitro* conditions, by allowing the attachment of complex microbial communities coming from mucosal surfaces biofilms (Tinanoff & Gross 1976, Tinanoff et al. 1976). Epithelial cells are commonly found in 2-day biofilm samples in close contact with bacterial components, but after 7 days of biofilm growth its abundance is reduced. This may point to an important nutrient source in dental plaque, apart from diet carbohydrates and glucosalivary proteins. Additionally, studies analyzing the structure of the dental plaque show that early colonizers, such as streptococci, are placed in the outer-most layers of the biofilm (Zijnge et al. 2010), where aerobic conditions are found, whereas late colonizers are mainly found in deep layers of the biofilm. This is in contrastwith *in vitro* biofilms models, which would suggest a structure where early colonizers are placed in the lower layers, and late colonizers in the upper ones.

The formation of oral biofilms is highly influenced by intercellular communication. Evidence of this communication comes from the specific synergistic interactions between certain groups of bacteria, such as the corn-cob structures formed by *Streptococcus* and *Candida* species (Zijnge et al. 2010). Anaerobic bacteria also tend to appear in close contact to oxygen consuming species, as in the case of *Porphyromonas gingivalis* associated to *Neisseria sp.* (Marsh et al. 2011). Antagonistic competence between different species is also present. Secretion of inhibitory compounds (acids, H_2O_2, bacteriocins, etc) to the biofilm matrix provide a competitive advantage against other microbes (Rogers et al. 1979), explaining why some species only appear on discrete clusters. Furthermore, cells inside the biofilm can coordinate their transcriptional activity, by secreting small molecules and sensing their levels in the environment (quorum sensing) (Kolenbrander et al. 2010). Another characteristic of the oral biofilms, is the close contact between different species cells, embedded in polysaccharide matrix. This close contact facilitates the exchange of genetic material among phylogenetically distant oral species, through horizontal gene transfer (HGT) (Mira 2008).

1.2.1.4. Subgingival dental plaque diversity

Physical conditions in the gingival crevice, where SubGDP is formed, are slightly different from other oral sites. SubGDP is confined to the gingival crevice and the access to salivary glycoproteins and food debris is reduced. Bacteria living on SubGDP mainly feed on the GCF continuously flowing through the non-keratinized epithelium, and on the cells being desquamated from gingival tissue. The lack of sugar fermentation and the GCF buffering effect prevent the establishment of acidic conditions. Mechanical forces exerted by chewing and tongue movements are minimal, reducing the erosion of the biofilm. Oxygen is scarce, as the continuous GCF flow outside the gingival crevice, prevents oxygenated saliva to enter, and oxygen is rapidly consumed. Thus, the basophilic and anaerobic conditions favor the growth of strict anaerobes or microaerophiles, such as spirochaetes, and basophilic species.

Another physical conditioning is the presence of two different kind of surfaces that are in close contact with the SubGDP, the cementum at the tooth side, and the non-keratinized epithelia in the other side. Different bacterial species have variable adhesion tropism towards either cementum or epithelium, creating a layer-organized biofilm. Additionally, as the GCF emanates from the epithelium, there is a gradient of nutrients from this side to the cementum, affecting the spatial availability of nutrients. Those conditions allow Gram negative and/or motile species to preferentially appear close to the epithelium, whereas Gram positive rods and cocci appear mainly close to the root surface (Listgarten 1976, 1994). Those particularities make SubGDP a highly diverse ecosystem.

Total bacterial diversity has been extensively studied using different techniques. As previously stated, most of the bacterial species in many niches are not yet cultivable. Scientific community has bypassed this limitation by applying molecular techniques to investigate the total bacterial diversity in a community. In SubGDP, studies using 16S amplicon cloning and sequencing[10] reported an estimated total diversity of 34-179 species-level phylotypes in healthy individuals (Paster et al. 2001, Aas et al. 2005, Bik et al. 2010). Recently, with the advent of NGS, high-throughput sequencing of 16S amplicons has increased the total diversity found in SubGDP up to 194-300 species-level phylotypes reaching up to 87% of the total diversity expected in this niche (Griffen et al. 2012, Huttenhower et al. 2012, Abusleme et al. 2013). The structure of the subGDP community is dominated by a few abundant species (*Streptococcus*, *Prevotella, Corynebacterium, Veillonella* and *Haemophilus*) and many other corresponding to other genera contributing to a lesser extent, making the so called long tail effect (Zaura et al. 2009). This high dominance in the bacterial community makes it difficult to fully describe the total diversity present, as the sampling effort needed is enormous.

Health problems potentially caused by subGDP are mainly related to inflammatory processes in the gingival tissue. This gingivitis can progress to periodontitis, where the tooth supporting tissues are infected and inflamed, ending with alveolar bone loss (Kawar et al. 2011). This inflammatory response is triggered by the prolonged accumulation of subGDP, and it can be induced by letting subGDP to accumulate during 21 days (Grant et al. 2010). The continued close contact between bacterial and epithelial host's cells, pose a challenge to the host and needs to fight against bacterial tissue invasion. The persistence of the subGDP, increased GCF flow, altered immune response and inflammation, lead to the destruction of the periodontium, which can eventually end in the loss of dental pieces[11].

Those bacteria associated with periodontitis lesions, have been termed periodontopathogens, and were grouped by complexes. The red complex, which is formed by *Tannerella forsythia*, *Treponema denticola* and *Porphyromonas gingivalis* (Socransky et al. 1998), has been found to be highly correlated with gingival pocket depth. The orange complex includes *Fusobacterium nucleatum/periodonticum, Prevotella intermedia, Prevotella nigrescens* and *Peptostreptococcus micros,* and it was also related to pocket depth. They have been accepted as the main ethiological agents of the disease, but some studies also correlate other Gram-negative bacteria (*Aggregatibacter actinomycetemcomitans, Fusobacterium*, *Prevotella*, *Campylobacter*, *Bacteroidetes, Sphorocytophaga, Synergistes, Negativicutes* and *Treponema*), Gram-positive (*Peptostreptococcus, Filifactor, Megasphaera, Staphylococcus aureus, Pseudoramibacter, Shuttleworthia, Mycoplasma* and *Mogibacterium*)

[10] 16S amplicon cloning and sequencing, is a molecular technique where the 16S rRNA gene is amplified using universal primers. Those amplicons are then cloned into *E. coli* and afterwards, sanger sequencing is done to obtain the sequence of each of the cloned amplicons.

[11] There are several variations of the etiological hypothesis presented here. For further reading, consult (Bartold & Van Dyke 2013, Rosier et al. 2014). Other hypotheses are also discussed later on applied to caries.

and *Archaea* (*Methanobrevibacter*) (Lepp et al. 2004, Kumar et al. 2006, Fritschi et al. 2008, Colombo et al. 2009, Griffen et al. 2012) . This polymicrobial origin of the disease, makes it difficult to characterize the specific etiological agent, as it may not be that a single bacterial species triggers the inflammatory response. Furthermore, the self-destructing inflammatory response against bacterial invasion, can be triggered by more than a single species.

In fact, some studies have proposed a change in the whole community together with the initiation of the disease (Kumar et al. 2006, Darveau 2010). As mentioned before, it is clear that periodontitis comes along with bacterial shifts in the subGDP. But nowadays it remains unclear the reasons for both bacterial dysbiosis and the origin of the disease, and even more important, what comes first.

Actual research lines are also focusing on the host factors that influence the appearance of the disease. Understanding the role of those host-specific factors in the onset of the disease, opens the door to explore of new therapeutic approaches, not only focused on the bacterial component, but also on the inflammatory response of the host.

1.2.1.5. Supragingival dental plaque.

The supraGDP is probably one of the better studied biofilms models to date. Since the discovery of bacteria inhabiting the supraGDP by Antony van Leeuwenhoek, a great deal of research has been done trying to characterize this particular microbial ecosystem. The metabolic activity of its microbiota is responsible for enamel degradation, which leads to caries lesions. SupraGDP grows over the visible teeth surfaces, i.e. those not covered by the gingival tissue. The main environmental constraint that bacteria have to face, is the mechanical force exerted by tongue movements, salivary flow, chewing, etc. This makes adhesion a critical capacity for surviving in the supraGDP. In fact, nearly all bacterial species studied have at least one adhesion partner or the capacity to adhere to the AEP (Kolenbrander et al. 2006), and can therefore adhere to the biofilm. The main nutrient supplies in supraGDP are the glycoproteins present in the saliva, food debris and cellular components from epithelium. The continuous access to saliva allows also a buffering effect, mainly through phosphate, bicarbonate and proteins present in it (Bardow et al. 2000). Buffering is crucial to prevent dental caries, as it will be later discussed. Access to oxygen is constant through oxygenated saliva, making the environmental conditions aerobic. Oxygen may become scarce if the biofilm is let grow and accumulate. Aerobic inhabitants can consume all the oxygen present, as the extracellular matrix prevents proper access of oxygen to deeper layers of the biofilm. This situation, together with bacterial associations between strict and facultative anaerobes, allows the growth of microaerophilic and anaerobic bacteria in this micro-niche (Kolenbrander et al. 2006).

Under healthy conditions, the supraGDP presents a high diversity. Different studies have predicted the total number of bacterial inhabitants in the mouth between 16 and 247 based on 16S rRNA cloning and sequencing (Aas et al. 2005, Bik et al. 2010), and 500-6888 based on 16S rRNA amplicon sequencing (Keijser et al. 2008, Zaura et al. 2009). This diversity is highly dominated by a reduced number of inhabitants (*Streptococcus, Neisseria, Veillonella, Rothia, Actinomyces, Corynebacterium* and *Haemophilus*), with the rest being present only at low abundance (*Leptotrichia, Campylobacter, TM7, Selenomonas, Kingella, Porphyromonas, Cardiobacterium, Gemella, Treponema, Aggregatibacter, Abiotrophia, Tannerella, Propionibacterium, Actinobacillus,* etc). This microbiota is able to ferment sugars, producing acidic compounds that can eventually cause caries lesions. A description of caries disease, its origin and etiology will be discussed in the next sections.

1.3. Caries. Definition and types

Dental caries can be defined as the result of the dissolution of the tooth mineral surface, caused by the acidic metabolic compounds produced by dental plaque's microbes, growing over the lesion (Fejerskov et al. 2008). Under neutral pH conditions, there is an equilibrium between enamel demineralization and remineralization, given the mineral saturation of the AEP that baths teeth surface. But when supraGDP microbiota metabolizes dietary fermentable sugars, producing acids as byproducts, the pH over the teeth surface is reduced. When the pH reaches the "critical pH" on the AEP, the equilibrium established between solid hydroxyapatite and its soluble ions, is displaced towards the soluble phase. The protons added remove phosphate and hydroxyl groups, converting them into water and HPO_4^{-2} and $H_2PO_4^{-1}$. Thus the equilibrium is unbalanced and hydroxyapatite is further dissolved until equilibrium is reestablished.

$$Ca_5(PO_4)_3OH \rightleftharpoons 5Ca^{2+} + 3PO_4^{3-} + OH^- \overset{4H+}{\rightharpoonup} 5Ca^{2+} + 3HPO_4^{-2} + H_2O \overset{3H+}{\rightharpoonup} 5Ca^{2+} + 3H_2PO_4^{-1}$$

When fermentable sugars are no longer available, buffering systems from saliva and dissolution of hydroxyapatite increase again the pH surpassing the critical pH. Then, the AEP becomes supersaturated on phosphate and calcium ions and they can precipitate and form new hydroxyapatite crystals, compensating the previous demineralization. This demineralization and remineralization is continuously happening under healthy conditions. Caries appears only when those cycles of demineralization and remineralization are unpaired and biased towards enamel dissolution, for instance when the biofilm is let grow and saliva buffering capacity is hindered (Figure 3).

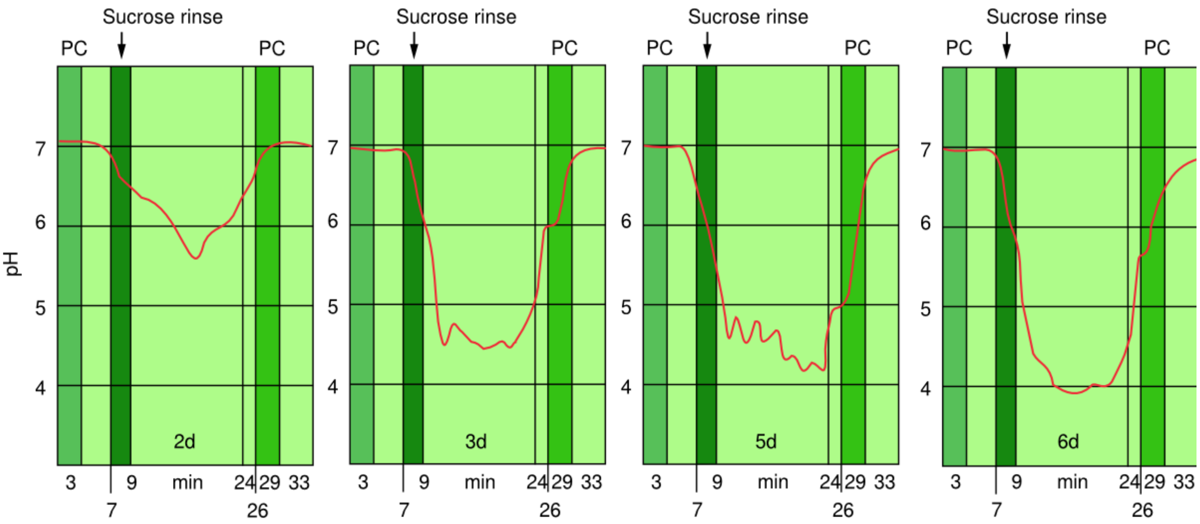


Figure 3. pH changes occurring in two, three, five and six-day-old interdental plaque of a 62 year-old volunteer, after rinsing during two minutes with a 10% sucrose solution. Paraffin chewing (PC) was given before and after the experiment. Adapted from (Imfeld & Lutz 1980).

It is important to differentiate dental caries from dental erosion, which also consists in the degradation of the teeth surface, although it is caused by factors other than bacterial acids (bruxism, bulimia, low salivary flow, abrasion, etc) (Imfeld 1996). Erosion has a different pattern of distribution compared to caries, depending on the source of enamel degradation. For instance, in the case of regurgitation and vomit, acid erosion caused by HCl is predominantly seen at the buccal and occlusal surfaces of premolars and molars in the mandible, as other sites are either less exposed to acid contact or are closer to a salivary ducts, suffering less erosion.

1.3.1. Caries as a multifactorial and chronic disease

Caries has to be considered a multi-factorial disease, where multiple factors influence the appearance and development of the pathology (Figure 4). None of the known factors influencing the disease has proved to be sufficient to predict the onset of caries. There are at least 4 fundamental factors necessary for the appearance of caries: cariogenic dental plaque accumulation; a tooth surface whose shape, disposition and composition make it susceptible of suffering caries; the presence of fermentable carbohydrates; and the co-ocurrence of all those factor along time, making caries a chronic disease.

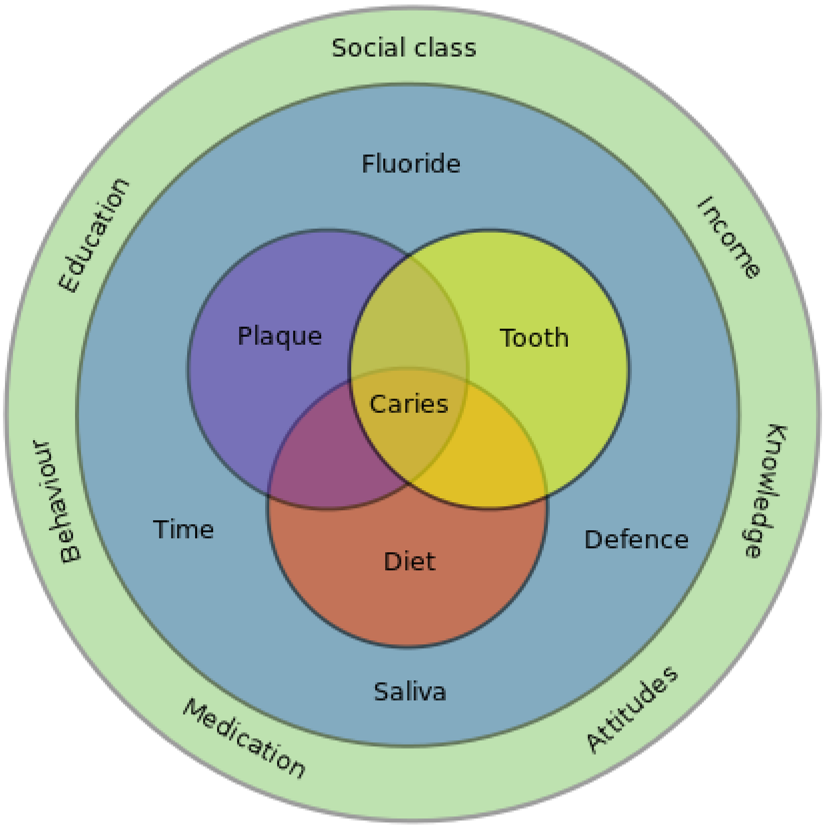


Figure 4. Keyes diagram representing the multifactorial origin of dental caries. Multiple factors are needed in confluence for caries appearance. Inner circles represent factors which play more important roles in caries onset, whereas outer circles reflect those conditions affecting the inner circles. Adapted from (ten Cate 2009).

Biofilm accumulation is a pre-requisite for caries to appear, so those spots where dental plaque is preferentially developed during long periods are more prone to caries development. Dental plaque prevents efficient buffering by saliva, retaining for longer periods the acids produced by the microbiota. For that reason, locations protected from mechanical forces (toothbrushing, chewing, tongue movements, etc) such as pits, grooves and fissures in occlusal surfaces, approximal surfaces and along the gingival margin, are the places where caries usually appears. Nevertheless, all teeth surfaces are potentially able to suffer dental caries, if biofilm is let to accumulate during long time periods, as usually seen for instance in orthodontic patients (Richter et al. 2011).

Another key factor for caries appearance is the pathogenicity potential of the established biofilm (Loesche 1986). The ability to produce acid in a sustained manner through time is the main pathogenic feature for caries. Acid is produced by the fermentation of simple carbohydrates, such as sucrose, from dietary sources. Thus, it has been shown that biofilms with an increased capacity to transport fermentable sugars inside the cells are more pathogenic. Phosphoenolpyruvate-phosphotransferase systems (PEP-PTS) are able to actively transport sugars even if they are scarce (Deutscher et al. 2006). The increased acid production reduces the pH, and the cariogenic biofilm must be able to survive this acidic environment for long time periods, and continue growing and metabolizing sugars[12], in order to consistently degrade the enamel. Lastly, both the production of extracellular (EPS) and intracellular (IPS) polysaccharides contribute to matrix biofilm formation and energy storage, respectively. EPS also help to retain localized the acids, so saliva has difficulties in buffering them. IPS are used under starvation periods and their metabolization helps to extend acid production when sugar from dietary sources is depleted.

Fermentable sugars intake is the main source of acid production in dental biofilms. Its metabolization produces acidic compounds as by-products that lower the pH, favoring conditions for the growth of acidouric and acidogenic bacteria. Higher sugar amounts and frequent intakes increase the risk to suffer tooth decay (Rugg-Gunn et al. 1984, Szpunar et al. 1995, Touger-Decker & van Loveren 2003, Moynihan & Petersen 2007). In developing countries where oral hygiene is deficient, water fluoridation programs are absent, and the consumption of mono- and disaccharide sugars is below 15-20kg per year, caries prevalence is low (Moynihan & Petersen 2007). Observational studies following caries prevalence and sugar availability, particularly during the World War II and beyond, detected a clear correlation between caries prevalence and sugar intake (Sognnaes 1948, Marthaler 1967). Probably the confirmation of the important role of sugars in caries appearance was based on the Vipeholm study, where a group of mentally disabled patients were frequently given high doses of sugar and candies to test whether sugar was causing or not caries (Gustafsson et al. 1954). The result was an increase of caries prevalence and loss of dental pieces in most of the patients receiving sugars. Additionally, *in vivo* experiments have measured the pH drop in dental biofilms due to different sugars and foods, showing that there is a clear acidification of dental biofilms (Rugg-Gunn et al. 1975). All the evidence available relating caries to carbohydrate consumption, indicate the key role of this factor into caries incidence.

Among the protection mechanisms that humans display against dental caries, one of the most important is saliva secretion. Salivary components are able to control and

[12] A bacteria or biofilm that produces acidic conditions is considered “acidogenic”. When it preferentially grow on acidic conditions, rather than neutral pH, it is considered “acidophile”. If it is able to produce acid even under acidic conditions, it is considered “aciduric”.

compensate the acidic demineralization caused by dental plaque's bacteria. Continuous salivary flow washes out bacterial cells that are not adhered to any surface. Antibacterial proteins present in the saliva (peroxidase, immunoglobulins, histatins, lactoferrin, lysozyme, mucins, agglutinin, cystatins, defensins, etc) (Wei et al. 2007), inhibit the excessive growth of microbes, reducing the amount of biofilm that is formed, although the extracellular matrix secreted by bacteria prevents the access of antimicrobial compounds inside the biofilm and thus reduces their killing efficiency. The formation of the AEP, provides a static liquid layer, which is not easily removed neither by toothbrushing nor by normal chewing and tongue movements. This makes the AEP a bacterial-free layer between the biofilm and the teeth surface, which acts as a diffusion barrier, preventing acids to reach the teeth surface. Its high concentration in calcium and phosphate saturates the pellicle, prevents the demineralization process, or even reverts it by depositing new mineral ions on the enamel surface. Eventually, mineral precipitation of the salivary fluid bathing the dental plaque can lead to calculus formation, creating an inorganic matrix that adheres firmly to the tooth surface and traps inside the bacteria present in the dental plaque (White 1997). Furthermore, as the AEP is the anchor to which initial colonizers are attached, different polymorphisms in the proteins adsorbed, may select a different set of bacteria being attached, determining a healthy or disease-prone microbial composition of the supraGDP. It is clear that saliva plays important roles in caries control, and its lack of production ends with severe dental caries, as seen in the Sjögren's syndrome (Ahmadi et al. 2013).

Fluoride has been included as an important environmental factor for caries. It was first noticed in 1923 by the observations from a dental practitioner, Frederick McKay, of mottled enamel in Colorado Springs' population. Inhabitants of this region suffering this enamel coloration were extremely resistant to tooth decay. Comparing water supplies from populations presenting those symptoms and lacking them, he discovered that fluoride salts present in drinking water, were the reason of this coloration and acid resistance of dental pieces (Peterson 1997). Since then, fluoride supplementation of water (Harding & O'Mullane 2013), toothpastes, table salt (Marthaler 2013) or milk (Bánóczy et al. 2013) has become the most effective public health intervention for caries prevention (Petersen & Lennon 2004, Jones et al. 2005). The low cost of fluoride and the posibility of supplementing community essential supplies, has a great impact in oral health in societies receiving regularly low doses of fluoride. Countries with high proportion of their population receiving fluoride have significantly decreased the prevalence of caries (Jones et al. 2005), and with no proved adverse effects due to its usage. Cancer has been proposed as a putative side-effect of fluoride supplementation in drinking water, given a study on lab animals, which found “equivocal” (uncertain) evidences of causing bone cancer in male rats (National Toxicology Program

1990). However, several studies have found no significative differences in osteosarcoma risk between fluoridated and no-fluoridated water regions (Kim et al. 2011, Levy & Leclerc 2012, Blakey et al. 2014). The discussion is open about the ethics of giving a preventive treatment to the whole population without their explicit informated consent.

Recently, other factors affecting caries susceptibility came from the genetic variability in humans. This became clear in the unethical Vipeholm study (Gustafsson et al. 1954). Most of the patients to whom sugars were frequently administered suffered caries through the study development, but around 20% of them, remained without caries. Vipeholm's researchers investigated the presence of caries in the families of those caries-free patients, and found that caries prevalence was significantly lower than that of the caries-prone patients' families (Böök & Grahnén 1953). They hypothesized that this was due to genetic heritable factors. Twins studies have been incredibly important to resolve this point (Townsend et al. 2003), as several caries indicators show high concordancy in monozygotic but not in dizygotic twins (Boraas et al. 1988). Several studies have shown increased caries risk in individuals with non-synonymous SNPs[13] in genes involved in taste pathways (Wendell et al. 2010), enamel formation genes (Patir et al. 2008), salivary components (Küchler et al. 2013) and immune system (Lehner et al. 1981, Acton et al. 1999, Hollox et al. 2008). Furthermore, environmental conditions may affect not only in a direct manner to caries experience, but also in an indirect way, by epigenetic changes in DNA, which may affect teeth development (Brook 2009, Chmurzynska 2010).

Other factors that influence caries appearance are from socio-economical nature. Education, social environment, attitudes, etc, have an impact on the quality of oral hygiene and number of visits to professional oral health care providers (Pine et al. 2004). Further efforts must be done in implementing educational programs directed to both adult parents and children at school, so that preventive measures can be applied correctly by most of the population.

1.3.2. Origins of caries.

Although caries is widely spread in modern societies, it has not been so common through the evolutionary history of hominids. Caries has been detected in numerous fossil records in hominids (Grine et al. 1990, Meng et al. 2011), but the frequency of caries highly increased coincidently with the technological revolution of agriculture development in the Neolithic (Richards 2002), which introduced into the normal diet fermentable carbohydrates,

[13] SNP stands for Single Nucleotide Polymorphism in a DNA sequence. If they are non-synonimous, it entails an aminoacid change at the protein level, and thus it can affect the protein's functionality.

posing a selective pressure towards the settlement of acidogenic and aciduric microbiota. Fossils from other hominids that lived before the Neolithic revolution were also found to suffer caries (Grine et al. 1990, Tillier et al. 1995, Aufderheide & Rodriguez-Martin 2011), although its prevalence has been always estimated to be smaller than in current times, around 3% (Grine et al. 1990), or than agricultural civilizations (Larsen et al. 1984, Lanfranco & Eggers 2010). It has not been until recently that a human hunter-gatherer settlement, has been described to have a high prevalence of caries (51.2%) (Humphrey et al. 2014). However, this particular population suffered from heavy tooth wear and was mainly fed by carbohydrate rich fruits, which may have favored the appearance of carious lesions together with erosive lesions.

The development of agriculture during the Neolithic introduced a high proportion of carbohydrates from grain and fruits from the incorporated crops. This increase in carbohydrate consumption posed an ecological challenge to the oral microbiota. This pressure reduced the biodiversity as seen in calculus, and favored the appearance of cariogenic species, such as *S. mutans, Veillonella sp.,* and periodontal pathogens such as *P. gingivalis, Treponema sp.* and *Tannerella sp*. (Adler et al. 2013, Warinner et al. 2014). The relatively short co-evolution history of acidophilic microbes can explain the high susceptibility of humans to caries, as the human genome has not been adapted yet to the microbiome changes in the oral cavity (Cordain et al. 2005). With the introduction of processed sugars and flour, putative cariogenic bacteria (such as mutans streptococci) became ubiquitous among human populations, and diversity was further reduced (Adler et al. 2013, Warinner et al. 2014). This reduction in the total diversity of the oral microbiota may have influenced the resilience of the normal microbiome against ecological stresses, and thus actual oral microbiome is more prone to dysbiosis.

Furthermore, the introduction of refined sugars since the industrial revolution in 1850 and the removal of taxes on sugar, further increased the selective pressure in the oral ecosystem. Therefore the presence and abundance of *S. mutans* and periodontopathogens (*Tannerella sp, P. gingivalis* and *Treponema sp.*) in today's population is higher than in preindustrialized ones, and the opposite is happening with health-associated microbes such as Ruminococcaceae. Those changes have facilitated the current widespread distribution of dental caries in industrialized societies (Adler et al. 2013, Warinner et al. 2014).

1.3.3. Caries development

Caries signs and symptoms are materialized in the demineralization of the mineral structure of the teeth. Initially, the disease curses asymptomatically at the macroscopic level, although some changes happen at the microscopic level. The enamel surface starts to dissolve, specially the intercrystalline spaces, which become wider increasing enamel's porosity (Holmen et al. 1985). Acids have then easier access to the subsurface enamel, whose fluoroapatite concentration is lower, making it more susceptible to acid demineralization. Additionally, the presence of proteins such as proline-rich proteins and statherins in the AEP, calcium and fluoride ions and buffering components of saliva, prevents surface enamel to continue dissolving (Aoba et al. 1984). Thus the lesion proceeds underneath a thin, and normally mineralized, layer of enamel. The maintenance of this external enamel layer, prevents bacterial colonization inside the enamel and the lesion is caused only by diffusion of acids produced by the dental plaque and dissolved enamel ions capacity to diffuse outside the enamel. When the enamel's pores become large and deep enough, they can be clinically detectable as white spots lesions, with air-drying at the beginning or even without it in later stages. The progression follows a perpendicular direction to the surface (Ekstrand et al. 1998), following the pores of the enamel's structure that are under the acidogenic biofilm, until demineralization reaches the dentin-enamel junction (Bjørndal et al. 1999) . This partially explain the typical double-funnel shape of some lesions, as the wider funnel found in dentin compared to the deeper radiographically visible lesion of the enamel, is caused by acids before bacterial invasion of dentin. Thus, when bacterial colonization of dentin happens, the mineralization degree is lower than healthy dentin, allowing a faster growth than in enamel.

As dentinal tissue has cellular components, it is able to react towards the acid diffusion through the enamel's pores, mainly by adding mineral deposits in the dentinal tubules, occluding them in order to avoid further damage and bacterial invasion (Stanley et al. 1983). When the dentin supporting the enamel becomes soft due to the loss of mineral, surface enamel above may break down and the barrier to the dental plaque is not present anymore. From this point, the disease's signs develop faster as the biofilm advances inside the tooth, reducing the diffusion gap and facilitating acid diffusion. Additionally, the biofilm is entering to previously demineralized enamel or dentin, whose hardness is smaller than sound enamel or dentin. It is important to remember that infected dentin, which appears softer to the practitioner, is not the front of the disease, as further demineralization is happening towards inner layer of dentin. Thus, there are three sequential fronts in the caries lesion; 1.- the infected dentin by biofilms, 2.- dentin with invaded dentinal tubules by isolated cells and 3.- sclerotic or reactionary dentin in response to acid diffusion (Bjørndal 1992).

When the demineralization front is close to the dental pulp, tertiary dentin is formed at the pulp-dentin junction, in an attempt to arrest the disease. Tertiary or reactionary dentin is formed when a traumatic stimuli is applied and reaches the pulp chamber. The odontoblasts-like cells in the dentin-pulp junction start the deposition of dentinal tissue towards the pulp chamber, and if the trauma is strong enough, the disposition of the dentinal tubules becomes irregular (Klinge 2001) in comparison with primary and secondary dentin. If the stimuli are prolonged in time, the pulp chamber can be completely obliterated, as the continued deposition of new tertiary dentin occupies the whole pulp chamber, preventing the advance of microorganisms.

1.3.4. Etiology of dental caries.

The etiology of dental caries has to be considered as multi-factorial, as it requires the conjunction of several factors for the disease onset (Figure 4). Undoubtedly, the conjunction of cariogenic plaque accumulation over time, together with the regular intake of fermentable sugars, at a teeth location prone to caries development, leads to caries onset. The role that microbial communities play in the disease etiology has been widely discussed and many theories have been proposed. In the following sections several hypotheses are described.

1.3.4.1. Non-specific plaque hypothesis

Willoughby D. Miller, a Koch's pupil, first observed that tooth decay was mainly caused by acidic demineralization. He isolated 23 bacterial strains, and tested them in search of acid and alkali production. He found that all of them were able to acidify the growth medium when sugar or starch was added, and to alkalinize it when meat or “albuminous substances” were added (Miller 1890). Thus, he concluded that caries could be caused by the combined effect of acid production of the whole microbiota and not by a single species. Furthermore, he pointed to caries as a two-step disease, where first the acid producing bacteria demineralize enamel and underlying dentin, softening them. Once the enamel is dissolved and softened dentin has been exposed, bacteria with an ability to degrade “albuminous substances” (i.e. dentinal proteins) will be more capable of degrading dentinal tissue.

Miller also demonstrated that when sugar or starch-paste was administered, the pH of saliva and plaque samples became acidic, and he could reproduce artificially caries. Additionally he observed that saliva samples incubated with albuminous substances, like meat, did not produce any lactic acid and pH was close to neutral.

Other authors also have observed that oral diseases are not the product of a unique species nor even a characteristic set of species, but in contrast it can be caused by different combinations of bacterial species (Theilade 1986). Thus, the classical mono-agent etiology as seen by Koch's postulates may not be supported by those observations. This opened new possibilities to establish the etiology of dental caries, as a non-specific multi-agent disease, where the whole metabolic outcome of the dental plaque's microbiota is the cause of caries.

The therapeutic consequences of this theory, are mainly based on deprivation of fermentable sugars and continuous removal of dental biofilm, as biofilm biomass must be correlated with acid production when sugars are present.

1.3.4.2. Specific-plaque hypothesis

The specific-plaque theory of caries, states that among the whole bacterial community of the supraGDP, only a limited number of them are the responsible of the acid production, and thus, the causative agents of the disease or odontopathogens (Loesche 1986). This theory was firstly presented as observations pointed that even the most careful daily debridement of dental plaque was *per se* not enough to prevent caries, concluding that its appearance was not due to the whole bacterial community as stated by the non-specific plaque hypothesis. The first bacteria associated with tooth decay was *S. mutans*, which was isolated in 1924 by Kilian Clarke from initial caries lesions (Clarke 1924). He reproduced caries *ex-vivo*, immersing extracted dental pieces into a *S. mutans* culture, showing that it could even colonize dentinal tubules. In the early 60's, Fitzgerald and Keyes demonstrated for the first time that caries could be caused by a single organism, in albino hamsters with a normal non-cariogenic oral microbiota (Fitzgerald & Keyes 1960). They also showed its infectious potential, as albino hamsters (not having caries naturally), could develop cavities if caged together with golden hamsters inoculated with streptomycin-resistant *S. mutans*. Additional evidence pointing to *S. mutans* as the causative agent of caries was based on the reproduction of caries in animal germ-free models, showing that certain clinical isolates of *S. mutans* were able to cause dental caries in specimens fed with a cariogenic diet and strains of *S. mutans* (Ooshima et al. 1981). Other species have been proposed to cause caries, such as *Streptoccocus sobrinus, Lactobacillus casei* and *Actinomyces odontolyticus*, as they have also been isolated in some cavity lesions (Loesche 1986). All those species have been proved to be acidogenic and aciduric, making them potential candidates to be odontopathogens.

Caries prevention strategies proposed by this hypothesis, are focused in avoiding the colonization of dental plaque by odontopathogens. This may have been accomplished by active or passive immunization (Smith 2002), using antibacterial agents or bacteriophage therapy against odontopathogens (Loesche 1979, Delisle & Rostkowski 1993, Eckert et al. 2012).

1.3.4.3. Ecological plaque hypothesis

As none of the previous hypothesis seemed to be completely correct, Marsh proposed a new one combining aspects of both theories (Marsh 1994a). Under normal circumstances, dental plaque goes through ecological alterations that may affect the conditions inside the biofilm. For instance, the availability of fermentable sugars ends with acid production by acidogenic species. However, the high taxonomic and functional ecosystem's diversity, rapidly compensates these changes and thus, the biofilm resists light fluctuations in external factors, for instance by alkali production or acid consumption by other non-acidogenic species. The resilience of the community maintains an equilibrium between all bacterial species, maintaining the microbial homeostasis (Marsh 2003). Resuming, under normal conditions, numerous inter-microbial and host microbial interactions compensate each other and maintain the balance, but when the disturbance is strong enough, this equilibrium may be broken.

Although acidogenic bacteria can be detected in dental plaque samples from healthy individuals, they are usually in low abundance (Tanner et al. 2002). In the event of the appearance of an ecological pressure that favors its growth, for instance a pH decrease in the biofilm due to higher sugar intake or reduced salivary flow, those species that are more capable to adapt to the new environmental conditions, will overgrow and dominate the community (Figure 5). This has already been proved in the gut microbiota, as changes in the diet modified rapidly and reproducibly both the bacterial composition and the genes expressed to adapt to the ecological pressure of nutrient type availability (David et al. 2013). If the same is applied to dental plaque, it can be expected that those bacteria that are more capable of harvesting sugars, even in acidic conditions and surviving to sub-lethal pH, will rapidly displace the rest of microbial inhabitants, whose efficacy in harvesting sugar and survival in low pH is worse. This change consists typically in a dominance of acidogenic and acid-tolerant Gram-positives bacteria (e.g. MS and lactobacilli), to the detriment of acid-sensitive species associated with sound enamel.

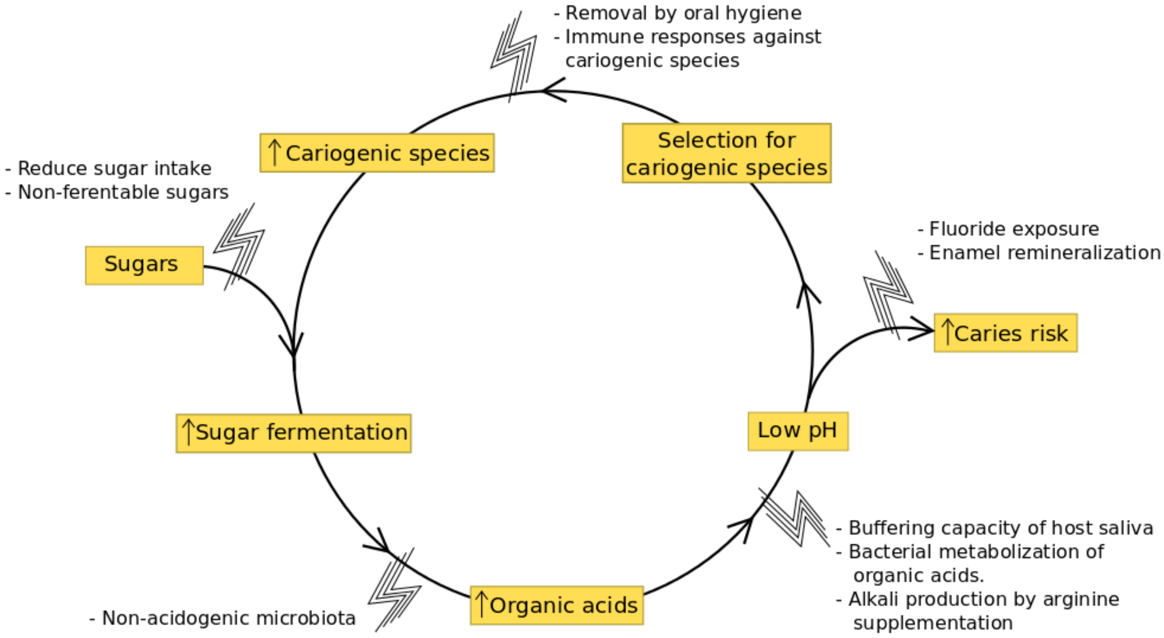

- Removal by oral hygiene
- Immune responses against cariogenic species
↑Cariogenic species
Selection for cariogenic species
- Reduce sugar intake
- Non-ferentable sugars
Sugars
- Fluoride exposure
- Enamel remineralization
↑Caries risk
↑Sugar fermentation
Low pH
- Buffering capacity of host saliva
- Bacterial metabolization of organic acids.
- Alkali production by arginine supplementation
↑Organic acids
- Non-acidogenic microbiota

Figure 5. Caries disease process as proposed by the ecological plaque hypothesis (Adapted from (Rosier et al. 2014)).

Kleinberg proposed that alkali production by arginine metabolism was a cornerstone for compensating acid production after carbohydrate fermentation (Wijeyeweera & Kleinberg 1989, Kleinberg 2002). He discarded mutans streptococci and lactobacilli as the sole causative agents of caries, and considered them as just another acid producing bacteria of the plaque ecosystem. After a series of elegant experiments, he identified a key element in the salivary components apart from bicarbonate, small arginine peptides, that were able to increase the pH of the plaque when metabolized by the microbiota (Kleinberg 1979). He then found that those acidogenic species, associated with caries, were only able to ferment glucose and produce a pH fall, but were unable to rise the pH afterwards, as they lacked arginolytic activity. Only those that had both the ability to ferment glucose and to degrade arginine where able to reproduce the fall and rise of pH (Wijeyeweera & Kleinberg 1989). Thus he proposed that the etiology of caries disease is an imbalance between arginolytic and non-arginolytic bacteria, caused by ecological changes in the plaque ecosystem.

This ecological change is proposed in this hypothesis as the motor of the dysbiosis of the resident microbiota, and thus the ultimately responsible factor for caries appearance. There are no etiological agents, as any bacteria with the capability of adapting to acidic environment may contribute to the disease. Derived from these assumptions, the therapeutic approaches proposed by this hypothesis are not only focused on antimicrobial, anti-adhesive or immunization strategies against putative pathogens, but also to the prevention of the selective pressures that favor the ecological shift (Figure 5). Those would include i) the restriction of fermentable sugars intake and the usage of non-fermentable sugars, ii) reshaping of oral microbiota in order to reduce the abundance of acidogenic bacteria and increasing those able to metabolize organic acids, iii) increase salivary flow and its buffering capacity, iv) the supplementation of tooth-paste with arginine (Kleinberg 1999, Acevedo et al. 2005, Liu et al. 2012), v) interference of acid production in dental plaque by fluoride and xylitol addition (Maehara et al. 2005), vi) the use of antimicrobial agents in dental care products, vii) stimulation of salivary flow after main meals (Marsh 2006), and viii) increasing remineralization periods by fluoride and calcium exposure or enamel remineralization agents (Kirkham et al. 2007).

1.3.4.4. Extended ecological plaque hypothesis

Traditional ecological plaque hypothesis considers caries as the consequence of cumulative processes of demineralization and remineralization, where the overall balance leads to net mineral loss. Thus, the lesions caused by a caries process may not be indicative of actual on-going lesion, as lesions can be arrested at any stage and have net mineral gain. The extended ecological plaque hypothesis proposed by Takahasi and Nyvad (Takahashi & Nyvad 2008), includes the clinical manifestations of caries lesions, as well as the detailed process by which the biofilm adapts to acid and selects for aciduric species (Figure 6).

This hypothesis proposes that under healthy conditions, dental plaque is a dynamic ecosystem, where non-mutans streptococci (non-MS) and *Actinomyces* are key responsible for the maintenance of the dynamic stability. They are able to ferment dietary sugars and thus reduce the pH on the tooth surface, even below the critical pH value of 5.5 (Sneath et al. 1986), although this drop in the pH can be easily buffered by homeostatic mechanisms in the plaque. In this dynamic stability stage, the enamel is smooth and shiny and dentine is shiny and hard. If the biofilm conditions favor a drop in the pH (by frequent supply of fermentable sugars or insufficient salivary flow to neutralize the acids produced), non-MS and *Actinomyces* may adapt its phenotype and increase its acidogenicity. The increased acid production, may select ‘low- pH’ species of non-MS and *Actinomyces*, shifting the microbiota into one even more acidogenic. Both phenotypic and genotypic changes in the plaque microbial community may unbalance the demineralization-remineralization equilibrium, towards a net mineral loss, initiating the lesion development. As the acidogenicity of the biofilm increases, aciduric bacteria such as MS and lactobacilli may increase its proportion in the community, and facilitate the lesion progression by maintaining an acidic environment that favors a net mineral loss. Throughout this process, the enamel becomes dull and rough, and dentin becomes dull and soft.

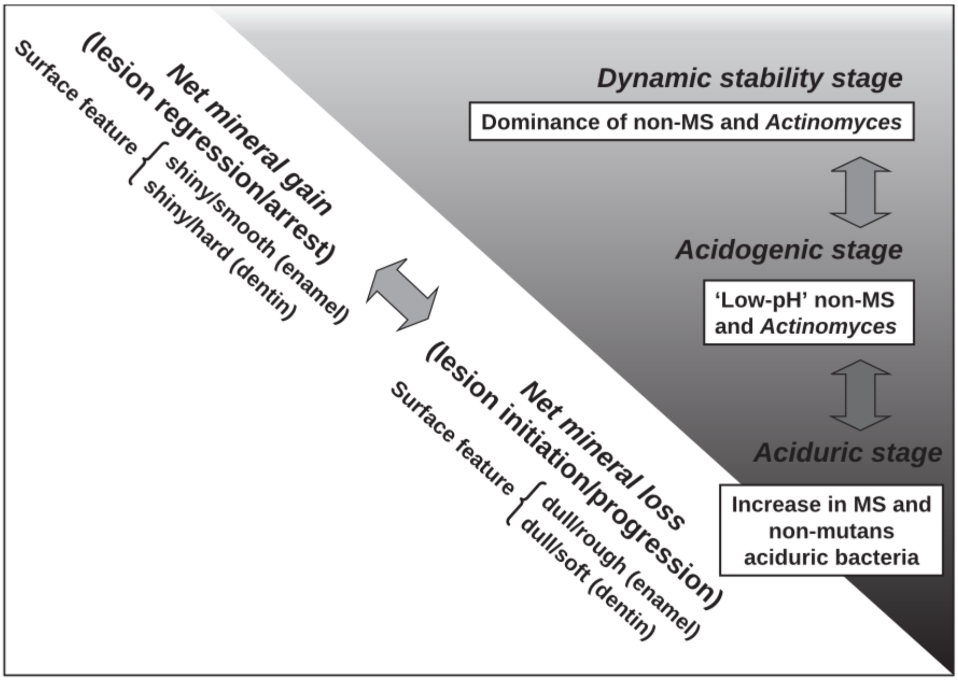


Figure 6. Extended caries ecological hypothesis, taking into account the relationship between the microbiota shifts in the dental biofilm and changes in clinical appearance of teeth lesions together with demineralization and remineralization processes (Takahashi & Nyvad 2008).

1.3.4.5. Tissue-dependent hypothesis of caries.

Although caries is considered as a pH-dependent disease, as low pH values are needed for demineralization of enamel's and dentin's hydroxyapatite, it is still not clear which are the changes in the microbial community as the disease enters different tooth tissues. Based on metagenomics observations, our group proposed a new hypothesis in which the disease can be divided into two different stages, enamel lesions and dentin lesions (Simón-Soro et al. 2013b). The first step would be caused by an acidogenic community, where adhesion, acid production by complex carbohydrate fermentation and acid stress resistance become key-values to survive in the acidic environment present in the enamel lesions. Once the biofilm overpass the enamel and reaches dentin, acid resistance is not so critical, as mineral content has already been dissolved and dentinal proteins are readily available. Other traits become more relevant to grow in those conditions, for instance those related with osmotic stress resistance, ability to degrade proteins, adhesion to collagen and fibronectin, and usage of human glycans. This would partially explain why caries can progress through dentinary tissue even with limited access to dietary sugars, particularly on interproximal lesions where occlusal forces prevent enamel from fracturing. During the enamel acidic demineralization, dentin is also demineralized by the acids flowing through the enamel pores (see section 1.3.3 “Caries development”), and thus, when the lesion reaches the dentinal tissue, its mineralization is lower than in sound dentin. The higher proportion of proteins in dentin (20% of collagenous proteins) than in enamel (Goldberg et al. 2011), implies an ecological advantage to those bacteria able to metabolize them more efficiently, and thus they become predominant. Protein degradation avoids pH to fall as much as in enamel lesions, and consequently, enamel retains its funnel shape and is not further acid-degraded at the dentinal side. Furthermore, the reduced mineral content of dentinal lesions, accelerates its progression once the enamel barrier has been compromised.

According to this hypothesis, therapeutic approaches leading with reduction of acid production by different strategies may be useful only in enamel lesions. But once the enamel barrier is broken, they may not be as effective. Further studies on bacterial isolates from dentin may show the critical players in dentin degradation and their collagenolytic contribution, and propose new strategies to avoid bacterial invasion of dentinal tissue.

1.3.4.6. Keystone-pathogen hypothesis

The keystone-pathogen hypothesis has been recently proposed in order to explain certain polymicrobial inflammatory diseases, such as periodontitis, intestinal inflammatory diseases or even colon cancer (Hajishengallis & Lambris 2012, Hajishengallis et al. 2012). It states that minor inhabitants of the normal healthy microbiota, are able to induce major changes into the microbial community, and cause a dysbiotic state that can lead to the onset of the disease. In the case of periodontitis, a plaque-derived inflammatory disease, *P. gingivalis* has been proposed as keystone-pathogen. Although it is well established that the subGDP is different under health and disease (Griffen et al. 2012, Abusleme et al. 2013), there is little knowledge about the factors that trigger this dysbiotic state. *P. gingivalis* has the ability to evade and subvert the immune system response, modifying the innate response and thus enabling an altered growth of the whole biofilm, which at the end causes inflammation and periodontic tissue destruction (Darveau 2009, 2010; Hajishengallis & Lambris 2011).

In caries, this hypothesis has not yet been proposed, as none of the usual inhabitants has been proved to induce the ecological changes in the microbiota or in the host's responses. But it could be perfectly plausible that a normal inhabitant of the microbiota could trigger changes in the structure and composition of the microbial community, promoting the development of an acidogenic biofilm and increasing the probabilities of suffering the disease. This seems quite difficult to prove, as long-term longitudinal studies would be needed to find potential keystone-pathogen candidates, and sub-clinical lesions must be detected in order to determine when the onset of the disease occurs.

1.3.5. Problems to approach the treatment of dental caries.

Although dental caries has been present in the human being since ancient times (Richards 2002, Meng et al. 2011, Wade et al. 2012, Adler et al. 2013), no effective cure has been found to date. Several reasons have hindered the discovery of an efficient, effective and preventive solution to caries onset. As it has been presented before, there are open discussions about the etiological causative agent of the disease, making it difficult to fight against an unknown pathogen. Caries is considered an infectious disease, but it does not strictly fulfill the classical Koch's postulates[14] (Koch 1870), which have been dominating infectious diseases etiology research since their establishment in 1870. Although Koch's postulates were reviewed to adapt them due to the advances in molecular techniques (Fredericks & Relman 1996), they are still not able to respond to the particularities of complex multifactorial diseases, such as dental caries (Russell 2009). As stated above, the main problem is that caries can not be linked to a single bacterial species, although several studies have proposed *S. mutans* to be "the" causing agent (Loesche 1986). Caries cannot be assigned only to the acid production of *S. mutans*, as other bacteria are also capable of producing acids in sufficient amounts to cause enamel demineralization (Russell 2009, Gross et al. 2012). Furthermore, it has been impossible to isolate any of the proposed candidates from every lesion studied. In contrast, potential candidates have been isolated in some healthy individuals (Thenisch et al. 2006). These misinterpretations of the disease have conducted some of the strategies to a blind alley, mainly focused in just one of the players in caries etiology and obviating the acidogenic potential of the whole dental plaque.

Probably, one of the reasons that have hampered finding an etiological agent is the long-term course of the disease. Caries is not a classic infection, where once the pathogen is established, it flourishes and causes the disease. Sustained periods of net mineral loss can be arrested and even reverted if proper hygiene and other interventions are accomplished.

[14] 1. The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy organisms.
2.The microorganism must be isolated from a diseased organism and grown in pure culture.
3.The cultured microorganism should cause disease when introduced into a healthy organism.
4.The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.

Observations made in different activity periods during the course of the disease (active demineralization or remineralization periods), can confound the microbial populations present at each period, as the environmental conditions in the biofilm are different. Then, it seems critical for the understanding of the disease a better classification of both clinically healthy and diseased individuals (from the point of view of the practitioner), depending on the stage and activity of the disease. Initial states of the disease may be crucial in understanding the origins of the disease, even before they can be clinically perceived.

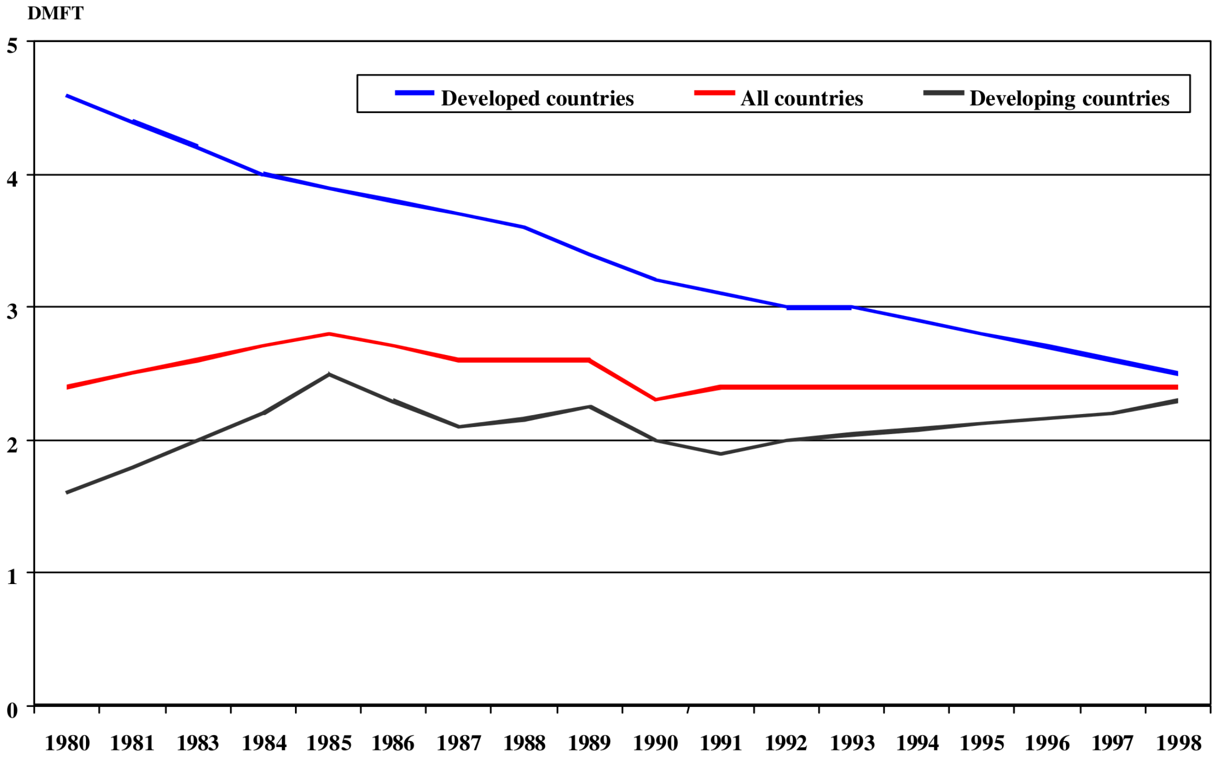


Figure 7. Evolution of caries prevalence in 12 year-old children from 1980 to 1998, in developed and developing countries as seen by the average number of decayed, missing and filled teeth (DMFT). Adapted from (Petersen 2003)

Another important issue that has hindered caries research, is the cultivation bias. Classical microbiological methods have relied on bacteria which are easily retrieved by culturing techniques. The oral cavity is one of the ecosystems with more cultivable members, but around 50% of the total diversity found in dental plaque still remains uncultivated (Paster et al. 2001, Wade 2002, Donachie et al. 2007, Marsh et al. 2011). Although there have been some efforts trying to develop new culture media to improve the diversity of species recovered by culturing (Nichols et al. 2010, Tian et al. 2010, Tanaka et al. 2014), this gives still a very limited picture of microbial diversity. Introduction of molecular techniques, such as cloning, denaturing gradient gel electrophoresis (DGGE), DNA and RNA microarrays and

16S rRNA sequencing, have highly improved the available knowledge of the taxonomic diversity, but they still have their own biases and limitations (Nyvad et al. 2013). In this thesis other techniques were applied to the knowledge of oral microbiology, which can potentially overcome some of those limitations. Metagenomics has been used to overcome both the culturing bias and the limitation of 16S rRNA gene sequencing to retrieve functional information (Alcaraz et al. 2012, Belda-Ferre et al. 2012) . Additionally, metagenomics has been proposed to be used in the detection of virulence genes of pathogens (Belda-Ferre et al. 2011). Metatranscriptomics was applied to know which microbes were transcriptionally active through the biofilm formation process and during the pH drop after a carbohydrate rich meal. Finally, metaproteomics was used in order to detect potential biomarkers of caries susceptibility, which could be used for caries risk assessment (Belda-Ferre et al., 2015, under review).

One of the most important limitations of caries research is the disease multifactorial origin. Many variables can affect the susceptibility of an individual to suffer caries throughout its life (see section 1.3.1 “Caries as a multifactorial and chronic disease”). This multi-factorial origin of the disease has made it difficult to eradicate caries, as all efforts performed up-to-date have focused only in a few of those factors. Those efforts have reduced significantly the severity of the disease in wealthy societies (Figure 7), in which the easy access to simple sugars in the diet rose the DMFT[15] levels alarmingly (Edelstein 2006). Partially successful strategies have reduced the severity of the disease, as it is the case of the implementation of fluoride in drinking water or toothpastes (Petersen & Lennon 2004), which increase the resistance of enamel to acidic dissolution either by incorporating to the enamel itself or, more probably, by increasing the remineralization capacity of saliva after acid challenges (Fejerskov 2004). Even recognized as one of the 10 most important advances in public health of the 20th century[16], some studies reduce the impact of this preventive effect, pointing to a disease slowdown rather than etiology counteract. Fluorides only delay the appearance of dentin caries, but the presence of white-spots lesions is not always prevented, pointing out that the origin of the disease (acid production) is still present (Groeneveld 1985). Other strategies have focused in eradicating one of the potential causal agents of caries, *S. mutans*, either by vaccination (Taubman & Smith 1974, Russell et al. 2004) or by replacement therapy with genetically modified strains of *S. mutans* (Hillman 2002, Hillman et al. 2007). Vaccines have been developed to target epitopes of *S. mutans*, trying to increase mucosal and salivary immunity against it (Taubman & Smith 1974, Russell & Wu 1991). But this approach has also its weakness, as caries is nowadays thought to be caused not only by *S. mutans* but by a complex and dysbiotic bacterial community (Kleinberg 2002, Marsh 2003, Belda-Ferre et al.

[15] DMFT: Decayed, missing and filled teeth.

[16] http://www.cdc.gov/mmwr/preview/mmwrhtml/00056796.htm

2012). Thus, if this hypothesis is true, the benefits of a caries vaccine only targeting *S. mutans* may not be completely effective, as other species may produce equal amounts of acidic compounds and demineralize enamel. Sugar intake is another important factor for caries onset and consumption reduction is key for caries prevention. Unfortunately, nowadays most of the processed food products contain substantial amounts of refined sugar as additives, which makes it difficult for a consumer to easily avoid sugar intake.

1.3.6. New techniques to confront dental caries.

As a consequence of all those factors influencing the onset of caries disease, there is a need to apply new approaches and research lines to fight against dental caries. In the last decade new high-throughput techniques have been introduced that allow to overcome the difficulties in culturing all the inhabitants of an ecosystem, by studying informative molecules such as DNA, RNA, proteins and metabolites (Zoetendal et al. 2008, Nyvad et al. 2013).

The analysis of nucleic acids has been revolutionized by the introduction of new sequencing techniques. Table 1 compiles a comparison of the main characteristics between different generations of sequencing technologies, which will be briefly described. Traditional or first generation sequencing (Sanger sequencing), based on dye-terminator chemistry (Sanger et al. 1977) has been superseded by second generation sequencing techniques, such as Illumina (Bennett 2004), 454 pyrosequencing (Ronaghi et al. 1996, Margulies et al. 2005), SOLiD (McKernan et al. 2009) or IonTorrent (Rothberg et al. 2011). Those techniques are based on previous library preparation step which implies a PCR amplification of the sample, with the associated biases this may introduce to the final results (Schwientek et al. 2011). Additionally, the length of the sequences obtained is typically shorter than Sanger technology. Nevertheless, the current sequence length of 454 pyrosequencing (700-800 bp) and its high number of reads (1 million per sequencing plate) has made it a technique of choice for metagenomic studies until third-generation sequencing techniques are fully developed. The extensive hands-on work needed for the library construction, the high cost of acquiring the sequencing machine, the need of special high quality reagents and the long time needed for each sequencing run, have brought about the emergence of the third generation sequencing technologies. Those techniques have the peculiarity of directly sequencing single molecules, without the need of a previous amplification step or complex library preparation protocols, reducing the bias of the PCR and library preparation, reducing the operative costs, and highly increasing the theoretical maximum sequence length that can be achieved. Examples of this sequencers are Pacific Biosciences (Eid et al. 2009), Oxford Nanopore Technologies (Clarke et al. 2009), Helicos Biosciences (Ozsolak et al. 2009) or IBM DNA transistor (Luan et al.

2012) . In the near future those technologies will truly democratize the sequencing techniques to every application, given its higher throughput and its lower cost per base sequenced than previous technologies and the miniaturization of sequencing machines. Both second and third generation technologies have still high margin for improving the actual technology in order to achieve longer and higher quality reads.

In order to overcome the above mentioned culturing techniques limitations, bacterial community diversity has been studied using molecular techniques, and at the beginning most of them were based on 16S rRNA gene sequencing (Weisburg et al. 1991). The time and cost reduction provided by this technique, allowed many research groups to conduct diversity studies on a wide range of microbial communities. Briefly, this methods rely on PCR amplification of the full 16S rDNA gene or a part of it, which are later analyzed either by DGGE (Muyzer 1999, Li et al. 2007, Tian et al. 2010), microarray hybridization (Wagner et al. 2007), cloning and sequencing (Becker et al. 2002, Kumar et al. 2006) or direct sequencing (Yang et al. 2011, Gomar-Vercher et al. 2014). This approach showed that many species remained uncultured and were not considered in traditional culture-based taxonomic studies (Pace 1997, Kroes et al. 1999, Sogin et al. 2006), involving a revolution in the understanding of microbial communities. Despite those milestones accomplished through 16S rRNA gene sequencing, a high number of drawbacks have limited the usefulness of this technique (V. Wintzingerode et al. 2006).

First, the sequence of the conserved target region of the 16S rRNA gene that is used to design primers for the PCR amplification, is not exactly identical among different bacterial species, and some phylogenetic groups systematically fail to be amplified (Sipos et al. 2007, Hong et al. 2009). The hypervariable regions that are used to identify bacteria, vary among members of the same species (Martínez-Murcia et al. 1999) or even between different ribosomal operons in a given genome (Acinas et al. 2004), which can lead to increased biodiversity estimates and misidentification of species. There is also a variation in the number of ribosomal operons among different genomes, which can lead to errors in the quantification of the number of cells present in the sample, although there are some methods to reduce its impact (Kembel et al. 2012). Furthermore, the sequence length of the reads obtained with current sequencing methods, impede the complete sequence of the 16S rRNA gene, which constrains the analysis to a particular region. Depending on the region selected, different taxonomic composition and accuracy in the assignment will be obtained from the same sample (Wang et al. 2007, Cruaud et al. 2014). The PCR amplification protocol used, including annealing temperature, extension time, initial template composition and DNA concentration, the total number of amplification cycles or the fidelity of the polymerase used,

Table 1. Comparison of the main characteristics of first, second and third generation sequencing technologies. Adapted from (Schadt et al. 2010) .

	First generation	Second generation	Third generation
Fundamental technology	Size-separation of specifically end- labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS[17]	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

[17] SBS: Sequence by synthesis

also inflicts considerable biases in the results obtained (Suzuki & Giovannoni 1996, Polz & Cavanaugh 1998, Sipos et al. 2007, Wu et al. 2010, Haas et al. 2011, Gonzalez et al. 2012). Additionally, the introduction of miscalled base-pairs during the sequencing process or the generation of chimeras during the amplification can artificially increase the number of OTUs found and give an inaccurate taxonomic composition (Gomez-Alvarez et al. 2009, Kunin et al. 2010). Furthermore, 16S rRNA gene profiling is limited in terms of the functional information it provides, as only taxonomic information of the bacteria present can be unraveled. This issue has been partially solved recently, with the development of a functional inference method from 16S rRNA gene data (Langille et al. 2013). Another major problem is that the depth of the taxonomic assignments obtained from current sequence length and binning methods, can only be made reliably at the genus level, making the inference more difficult. Additionally, pangenomic information from all species of a given environment needed for this method to be precise is far from being available for most niches, and in any case, the presence of specific clones of a given species may mark the difference between two similar samples. For all those reasons, the results obtained by this methodology must be interpreted with caution, and keep in mind all the possible biases in order to reduce their impact in the results.

Metagenomics was developed to study the genomes of a whole microbial community, in contrast to the single-marker gene approach of 16S rRNA gene sequencing. By a truly metagenomic approach, both taxonomic and functional composition of the studied sample can be obtained, as other functional genes apart from the 16S rRNA are obtained. Metagenomic studies expanded when capillary sequencers from the human genome sequencing initiatives, were available after completion of the human genome (Lander et al. 2001) (Venter et al. 2001). They were used initially for sequencing bacterial genomes[18], and afterwards for shot-gun sequencing of microbial communities. The first metagenomics approaches developed were performed through cloning fragmented DNA pieces from a given environment in long-insert vectors, such as fosmids or BACs, or in shorter insert hosts (Rondon et al. 2000, Breitbart et al. 2002, Venter et al. 2004, Gill et al. 2006) (Figure 8). This metagenomic approach has the potential of performing functional screenings to each of the clones from a given library, allowing the characterization of novel genes of biotechnological interest, if the gene of interest is able to be expressed in the host bacteria. The possibility of discovering long stretches of DNA in long-insert vectors enables the discovery of complete gene clusters or operons, leading to the discovery of complete metabolic pathways in a single genome,

[18] http://microbialgenomics.energy.gov/index.shtml

which provides more ecological information than the presence of all the genes in different genomes (Rodríguez-Valera 2004). However, the cloning step associated to this methodology, introduces some biases that can highly alter the taxonomic composition of the analyzed sample. For instance, long insert vectors, such as fosmids, typically contain sequences with higher GC content compared to direct sequencing metagenomics, and can preferentially clone viral or eukaryotic DNA at the expense of the dominant bacterial DNA (Temperton et al. 2009, Ghai et al. 2010). This is explained because the presence of genes coding for toxic compounds to the host, can impede the growth of those clones in the library.

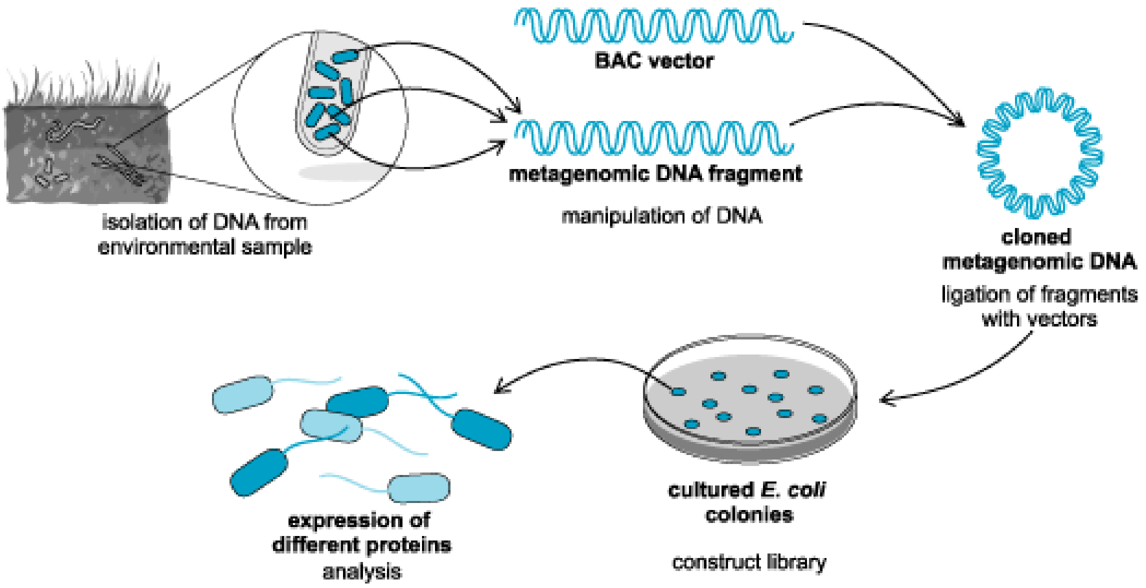

BAC vector
metagenomic DNA fragment
manipulation of DNA
isolation of DNA from environmental sample
cloned metagenomic DNA
ligation of fragments with vectors
cultured *E. coli* colonies
construct library
expression of different proteins
analysis

Figure 8. Typical shot-gun metagenomic approach protocol. First DNA is extracted from an environmental sample, which is fragmented to the appropriate size for the cloning vector used. Then, the metagenomic DNA fragments are ligated with the cloning vector and the resulting construction is then inserted into a competent *E. coli* strain. The resulting metagenomic library can be then tested in different ways, such as functional screenings, PCR amplification of interest genes or direct plasmid/fosmid sequencing.

In order to overcome the cloning bias associated to metagenomic clone libraries, direct sequencing has been proposed as an alternative approach. This methodology avoids both amplification and cloning steps of 16S rRNA gene and cloning metagenomic approaches for the characterization of a microbial community. However, it has the limitation of relying in the sequences available in public databases for the identification of genes and functions, which are mainly composed of cultivable microorganisms. This limitation can be clearly noted in the percentage of annotated sequences in different metagenomic studies, which usually ranges from 40 to 60% of the total reads (Belda-Ferre et al. 2012, Kriete & Eils 2013). Even with those drawbacks, direct sequencing metagenomics allows to identify putative differences in the microorganisms community of a given niche, as it reveals differences at the functional level that may not be apparent when analysing only the taxonomy as seen by the 16S rRNA gene. Furthermore, the posibility of discovering bacterial genomes without the need of previous isolation in pure culture, allows the discovery of a wide collection of microorganisms that are not adapted to laboratory conditions, better reflecting its ecology and the important traits that allow the microbe to survive in a particular environment. Future sequencing techniques should allow the complete sequencing of single cells (Marcy et al. 2007, Koren & Phillippy 2015, Raveh-Sadka et al. 2015), allowing the complete sequencing of the genomes in a particular niche, and thus enabling a better characterization of its microbial ecology.

Although metagenomics is able to inform us about the genetic potential of a given environment, it is not able to characterize which of those ORFs are actually real genes or just wrongly predicted ones. Furthermore, it is impossible to determine if they are being actively transcribed in a given moment, and thus playing a role in the adaptation process to given conditions. To illustrate this point, imagine the case where low-abundance members of the community are highly active. The metagenome would give an erroneous view of what is taking place in that community, as it would not detect its high activity. When analyzing RNA instead of DNA, actively transcribed genes can be revealed, obviating those bacteria that are present in the sample but not actively transcribing genes under given conditions (Gentile et al. 2006). Applying the Baas-Becking hypothesis, “everything is everywhere, but the environment selects”, microorganisms can be present in a given sample even if they do not live there, and could be detected by metagenomic approaches. However, if they are not adapted to that environment, or to the conditions present at that moment, their transcription activity will be low or absent. This gives the potential to metatranscriptomics approaches of knowing which members of the community are somehow active, and which genes are being transcribed under given circumstances, which can shed light about the microbial adaptation to changing conditions.

Current metatranscriptomics techniques are based on the construction of cDNA from RNA by reverse transcription, in order to be able to sequence them (Gosalbes et al. 2011). During this step, the RNA is processed using a set of enzymes to construct first a fragmented single strand of cDNA, and those fragments are joined using the hybridization to the parent RNA as backbone. Then the RNA is removed and finally a second cDNA strand is constructed. Through this process, the high temperatures that are applied can rapidly degrade the RNA, given its highly labile nature and the widespread presence of ribonucleases (RNases). This can lead to a biased representation of the transcribed genes at a given moment.

However, some of the newest sequencing technologies claim that direct RNA-sequencing could be performed without the reverse transcription step (Ozsolak et al. 2009). Additionally, it has to be kept in mind that over 90% of the transcribed genes in a given cell correspond to ribosomal genes, which allows the identification of those members of the community with active transcription without previous PCR amplification as in the case of DNA, but on the other side of the coin, this hinders the discovery of other functional genes being transcribed. For that purpose, before sequencing the cDNA, the ribosomal transcripts must be withdrawn from the sample, otherwise the sequencing depth that would be required to have enough genes transcripts reads could be unaffordable. This has been partially solved through the development of ribosomal transcripts subtraction methods that increase the proportion of mRNA in the RNA prior sequencing (Stewart et al. 2010, Yi et al. 2011, Giannoukos et al. 2012, Peano et al. 2013). The four main strategies developed for rRNA removal (1.- rRNA hybridization capture pulldown, 2.- degradation of processed RNA with a 5'-3' exonuclease, 3.- duplex-specific nuclease normalization and 4.- selective hexamers priming for cDNA construction), obviously can introduce extra biases to the data obtained, for instance the loss of the quantitative information on transcripts abundance (except in the case of hybridization capture approach), accidental mRNA removal, etc. Nevertheless, metatranscriptomics is a powerful approach to uncover the activity patterns of microbial communities.

Still, the existence of post-transcriptional regulation in bacteria makes the correlation between transcription and translation not completely direct (Nogueira & Springer 2000, Pradet-Balade et al. 2001). RNA translation is susceptible of regulation by the degradation of ribonucleases, riboswitches or interference RNA (iRNA), among others. Thus, the study of the molecules that are truly performing the activity of the cell, the proteins, is highly relevant and is usually termed metaproteomics (MTP) (Rodríguez-Valera 2004). Metaproteomics, together with metametabolomics (study of the metabolites of a community), are probably the techniques that more closely describe the functional activity of the community at a given time point. The metaproteome of a microbial community has been analyzed through the use of different techniques. Although the idea of analyzing proteins from microbial communities in response to changing environmental conditions comes from the early 90's (Ogunseitan 1993), metaproteomics has not been widely used until the development of efficient separation techniques and analytical tools. Mass spectrometry has been continuously being developed and improved, increasing the sensitivity and accuracy of the instruments. It basically consists in weighting molecules of interest through the measuring of its mass-to-charge ratio (m/z). First the analyte is ionized (most commonly used techniques in proteomics are MALDI[19] and ESI[20]) (Yamashita & Fenn 1984, Hillenkamp & Karas 1990), and is conducted to a mass

[19] MALDI: Matrix-Assisted Laser Desorption Ionization

[20] ESI: electrospray ionization

analyzer (time-of-flight, quadrupole, iontrap and orbitrap are examples of mass analyzers), where the mass-to-charge ratio is measured. Sequential MS steps can be applied to the same protein (tandem MS or MS/MS), by fragmenting the initial ion into smaller ions, that can potentially give the peptide sequence. Among the separation techniques needed for protein separation, the first ones relied on bi-dimensional polyacrylamide gel electrophoresis (2D-PAGE) fractionation, followed by the identification of relevant spots using MALDI-TOF mass spectrometry (Wilmes & Bond 2004). The throughput of this method is limited by the separation efficiency of the 2D-PAGE and the number of spots selected for MS identification. For that reason, HPLC[21] has been coupled in-line with MS in order to improve the separation resolution and to increase the throughput of the technique, as performed by Verbekmoes and collaborators in the first paper about the human gut metaproteome (Verberkmoes et al. 2009).

Metaproteomics, as metagenomics, has the limitation of relying on previously available sequences, in order to identify the obtained m/z values. This fact complicates the discovery of unknown proteins. Another issue that has been pointed, is that one cell is not translating at a given moment all the proteins coded in its genome. Thus the full set of proteins that can potentially be translated must be looked for under a wide range of environmental conditions. Thus, the reproducibility of those studies can be affected by uncontrolled variables.

To conclude, nowadays several tools are available to approach the study of microbial communities directly from their natural environment, obviating the need of culture implied in traditional techniques.

1.3.7. Omics approaches to study dental caries.

In this thesis, new high-throughput “omics” techniques have been applied to oral samples in order to shed light on critical aspects of dental caries. Metagenomics has been applied to overcome the culturing, the PCR and the cloning approaches of previous classical works, showing a less biased picture of members inhabiting the human dental plaque, both under health and disease conditions (Belda-Ferre et al. 2012). Previous studies based on either culturing or 16S rRNA gene sequencing, were limited by the biases imposed by those techniques, mainly the inability to culture most of the oral inhabitants (Paster et al. 2001, Wade 2002, Donachie et al. 2007, Marsh et al. 2011) or PCR biases due to over-amplification of some taxonomic groups depending on the PCR conditions used (Suzuki & Giovannoni 1996, Sipos et al. 2007, Hong et al. 2009, Morales & Holben 2009, Haas et al. 2011) as well

[21] HPLC: High Performance Liquid Chromatography

as to under-amplification of low-level members of the community (Gonzalez et al. 2012). In this thesis, metagenomics was used to compare differences in both taxonomic and functional composition between healthy and caries-bearing individuals, as well as within single carious lesions (Chapter 1 (Belda-Ferre et al. 2012) and 2 (Alcaraz et al. 2012)). Once taxonomic differences were identified, bacteria with potential antagonistic effects against cariogenic organisms were searched, by isolating in several culture media oral inhabitants that were specific to healthy individuals, which were later tested for inhibitory properties against mutans streptococci (Chapter 1 (Belda-Ferre et al. 2012)). Another metagenomic application presented in this thesis, is the comparison of pathogen genomes against healthy metagenomes of the same habitat where the pathogen was isolated, in order to find putative virulence genes (Chapter 3 (Belda-Ferre et al. 2011)).

Nowadays the understanding of dental caries is not compatible with a single-species etiology of the disease (Marsh 2003). In fact it is the acids produced by the whole bacterial community the responsible for the degradation of mineral components of teeth. For that reason, we applied metatranscriptomic approaches to distinguish the transcriptionally active dental plaque microbiota 30 minutes after a carbohydrate-rich meal, coincident with the pH drop (Chapter 4 (Benítez-Páez et al. 2014)). We also used metatranscriptomics to establish the biofilm formation process under *in vivo* conditions, as little is known about this process.

Metaproteomics was applied for two different purposes. First, to describe the protein content of human supraGDP for the first time, as previous metaproteomic studies had only focused on saliva (Jagtap et al. 2012), subGDP (Grant et al. 2010) or gingival crevicular fluid (Bostanci et al. 2010) protein composition. The second objective was to compare the protein composition of supraGDP between healthy and diseased volunteers, in order to find potential biomarkers susceptible of being included in a caries diagnostic kit (Chapter 5 (Belda-Ferre et al. 2015, under review)).

2

OBJECTIVES

OBJECTIVES

The study of human-associated microbial communities has experienced a boost since the beginning of the century, due to the introduction of new high-throughput and cost-effective techniques. Hence it has enabled to study the relationship of many diseases with the microbial community patients carry. Through the development of this thesis, some of those techniques have been applied to study the human dental plaque, in order to answer several biological questions regarding human dental caries. NGS and metagenomics have been used to overcome the limitations imposed by traditional culture-based, 16S-based or cloning techniques, with the purpose of deciphering the microbial composition under health and disease. The driving idea of the present thesis was to deepen the available knowledge on the total and transcriptionally active oral microbiome, under health and disease conditions, with the aim of deciphering the etiology of caries disease and proposing preventive and diagnostic strategies. This global objective was subdivided into the following specific goals:

- Taxonomic characterization of the microbial community present in the oral dental plaque of healthy and caries-bearing individuals, using open-ended techniques which are not biased by culturing, PCR amplification or cloning methodologies.
- Characterization of the genetic potential of the microbes inhabiting supraGDP. This should ascertain if the taxonomic differences between healthy and caries bearing individuals are also matched by functional differences.
- Detection of potentially caries-protective bacteria from healthy individuals, which could be susceptible of development as anti-caries probiotics.
- Propose putative virulence genes in pathogenic bacterial genomes, by comparing them with a metagenomic sample of healthy volunteers, obtained from the same environment where the pathogen was isolated.
- Compare the microbial community present in dental biofilms, as seen by metagenomics, with its transcriptionally active portion, in order to differentiate between the transient and active bacteria of this microbial niche.
- Characterization of the species being transcriptionally active during the pH drop which occurs on the teeth surface after a carbohydrate-rich meal, in order to discover the active players in the metabolization of sugars and acid production, and thus, potentially responsible for caries disease.
- Compare the oral microbiome composition as seen by metagenomics,

metatransciptomics and metaproteomics.

- Describe for the first time the human and bacterial protein catalog present in human supraGDP.
- Search for potential protein biomarkers of health and disease in dental caries for its use in a diagnostic test.

3

PUBLICATIONS

3.1

“The Oral Metagenome in Health and disease”

P Belda-Ferre, LD Alcaraz, R Cabrera-Rubio, H Romero, A Simón-Soro, M Pignatelli, A Mira

The ISME Journal (2012) 6, 46–56

The ISME Journal (2012) 6, 46–56
© 2012 International Society for Microbial Ecology All rights reserved 1751-7362/12
www.nature.com/ismej

ORIGINAL ARTICLE

The oral metagenome in health and disease

Pedro Belda-Ferre[1], Luis David Alcaraz[1], Raúl Cabrera-Rubio[1], Héctor Romero[2], Aurea Simón-Soro[1], Miguel Pignatelli[1] and Alex Mira[1]
[1]*Department of Genomics and Health, Center for Advanced Research in Public Health, Valencia, Spain and* [2]*Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias/C.U.R.E., Universidad de la República, Montevideo, Uruguay*

The oral cavity of humans is inhabited by hundreds of bacterial species and some of them have a key role in the development of oral diseases, mainly dental caries and periodontitis. We describe for the first time the metagenome of the human oral cavity under health and diseased conditions, with a focus on supragingival dental plaque and cavities. Direct pyrosequencing of eight samples with different oral-health status produced 1 Gbp of sequence without the biases imposed by PCR or cloning. These data show that cavities are not dominated by *Streptococcus mutans* (the species originally identified as the ethiological agent of dental caries) but are in fact a complex community formed by tens of bacterial species, in agreement with the view that caries is a polymicrobial disease. The analysis of the reads indicated that the oral cavity is functionally a different environment from the gut, with many functional categories enriched in one of the two environments and depleted in the other. Individuals who had never suffered from dental caries showed an over-representation of several functional categories, like genes for antimicrobial peptides and quorum sensing. In addition, they did not have *mutans* streptococci but displayed high recruitment of other species. Several isolates belonging to these dominant bacteria in healthy individuals were cultured and shown to inhibit the growth of cariogenic bacteria, suggesting the use of these commensal bacterial strains as probiotics to promote oral health and prevent dental caries.

The ISME Journal (2012) **6,** 46–56; doi:10.1038/ismej.2011.85; published online 30 June 2011
Subject Category: microbe–microbe and microbe–host interactions
Keywords: metagenomics; human microbiome; dental caries; *Streptococcus mutans*; pyrosequencing; probiotics

Introduction

The oral cavity of humans is inhabited by hundreds of bacterial species, most of which are commensal and required to keep equilibrium in the mouth ecosystem. However, some of them have a key role in the development of oral diseases, mainly dental caries and periodontal disease (Marsh, 2010). Oral diseases initiate with the growth of the dental plaque, a biofilm formed by the accumulation of bacteria in a timely manner together with the human salivary glycoproteins and polysaccharides secreted by the microbes (Marsh, 2006). The subgingival plaque, located within the neutral or alkaline subgingival sulcus, is typically inhabited by anaerobic Gram negatives and is responsible for the development of gingivitis and periodontitis. The supragingival dental plaque is formed on the teeth surfaces by acidogenic and acidophilic bacteria, which are responsible for dental caries. This is considered the most extended infectious disease in the world, affecting over 80% of the human population (Petersen, 2004). A poor oral health has also been related to the stomach ulcers, gastric cancer or cardiovascular disease, among others (Watabe *et al.*, 1998; Wu *et al.*, 2000). It is therefore surprising that no efficient strategies to combat oral diseases have been developed, despite their dramatic impact on human health. Some of the main reasons that oral pathogens have not been eradicated are related to the difficulty of studying the microbial communities inhabiting the oral cavity: First, the complexity of the ecosystem (several hundreds of species have been reported with multiple interaction levels) makes the potential pathogenical species difficult to target (Socransky *et al.*, 1998); second, not a single ethiological agent can be identified as in classical, Koch's postulates diseases. This has been clearly shown in periodontal disease, where at least three bacterial species that belong to very different taxonomic groups (the so-called 'red complex' of periodontal pathogens) are known to be involved in the illness (Darveau, 2010); and third, a large proportion of oral bacteria cannot be cultured (Paster *et al.*, 2001), and therefore traditional microbiological approaches give an incomplete picture of the natural communities inhabiting the

Correspondence: A Mira, Department of Genomics and Health, Center for Advanced Research in Public Health (CSISP), Avda. Cataluña 21, Valencia 46020, Spain.
E-mail: mira_ale@gva.es

Received 17 March 2011; revised 10 May 2011; accepted 12 May 2011; published online 30 June 2011

dental plaque. However, the development of metagenomic techniques and next-generation sequencing technology now allows the study of whole bacterial communities by analysing the total DNA pool from complex microbial samples.

Pioneering metagenomic studies in the human microbiome centred in the gut ecosystem, initially through a shot-gun approach, in which DNA was cloned in small-size plasmids followed by traditional Sanger sequencing method (Gill *et al.*, 2006; Kurokawa *et al.*, 2007), obtaining reads of about 800–1000-bp long. Recent approaches include the end sequencing of large-size fosmids (Vaishampayan *et al.*, 2010) and the use of Illumina sequencing technology to deliver vast amounts of small-size reads that could be later assembled (Qin *et al.*, 2010). Studies of the oral cavity microbiota, as well as other body habitats within the human microbiome such as the skin, the vagina or the respiratory tract, have mainly focused on the sequencing of PCR-amplified rRNA genes (Aas *et al.*, 2005; Grice *et al.*, 2008). These PCR-based studies have provided a substantial improvement of our knowledge of oral bacterial communities compared with past culture-based research, but the estimates of microbial diversity are hampered by biases in PCR amplification (de Lillo *et al.*, 2006), cloning bias (Ghai *et al.*, 2010) and when short pyrosequencing reads of the 16S rRNA gene were used, uncertainties in taxonomic assignment (Keijser *et al.*, 2008; Lazarevic *et al.*, 2009) and inflated diversity due to pyrosequencing errors (Quince *et al.*, 2009). Recently, the first study of the oral metagenome has been carried out by directly applying next-generation sequencing to a single sample from a healthy individual (Xie *et al.*, 2010), thus removing potential biases imposed by cloning and PCR. We have applied a similar approach to several samples varying in health status, directly sequencing the metagenomic DNA by 454 pyrosequencing, which has allowed us to compare the total genetic repertoire of the bacterial community under different health conditions.

Materials and methods

Sample collection

Supragingival dental plaque was obtained from 25 volunteers after signing an informed consent. The sampling procedure was approved by the Ethical Committee for Clinical Research from the DGSP-CSISP (Valencian Health Authority, Spain). The oral health status of each individual was evaluated by a dentist following recommendations and nomenclature from the Oral Health Surveys from the WHO, taking samples with sterile curettes. Plaque material from all teeth surfaces from each individual was pooled. In volunteers with active caries, the dental plaque samples were taken without touching cavities. In those cases, material from individual cavities was also extracted and kept separately. The volunteers were asked not to brush their teeth 24 h before the sampling. Information was obtained regarding oral hygiene, diet and signs of periodontal disease. DNA was extracted using the MasterPure Complete DNA and RNA Purification Kit (Epicentre Biotechnologies, Madison, WI, USA), following the manufacturer's instructions, adding a lysozyme treatment ($5\,\mathrm{mg\,ml^{-1}}$, at 37 °C for 30 min). For this study, eight samples were used for subsequent pyrosequencing, selected on the basis of homogeneity in their clinical features, including similar age, periodontal status, smoking habits and mucosal health. Supragingival dental plaque samples were taken from six individuals that were divided in three groups according to the number of caries they had suffered and that represented different degrees of oral health: two individuals had never developed caries in their lives (healthy controls), another two individuals had been regularly treated for caries in the past and had a low number of active caries at the moment of sampling (one and four cavities, respectively); and the last two individuals had a high number of active caries (8 and 15) and poor oral hygiene. In addition, samples from individual cavities were collected, and for two of them enough DNA for pyrosequencing was obtained: one at an intermediate stage and the other one at an advanced stage of caries development (dentin lesion), corresponding to teeth 1.6 and 4.6 following WHO nomenclature. The sequencing was performed at Macrogen Inc. (Seoul, South Korea) using the GS-FLX sequencer (Roche, Basel, Switzerland) with Titanium chemistry. After quality checking, average read length was 425 ± 117 bp. Sequences were deposited, and are publicly available in the MG-RAST server with the following accesions: 4447192.3, 4447102.3, 4447103.3, 4447101.3, 4447943.3, 4447903.3, 4447971.3 and 4447970.3.

Sequence analysis

Artificially replicated sequences (accounting for 1.2–4.54% of the raw reads) were removed from the data set using the '454 replicate filter' (Gomez-Alvarez *et al.*, 2009). The human sequences were identified by MegaBlast (Altschul *et al.*, 1990) against the human genome (e-value cutoff 1e−10) and were removed from the final data set. They accounted for 2.23–74.99% of the replicate-filtered reads (Supplementary Table 1). The metagenomic reads were mapped against 1117 sequenced reference genomes using the Nucmer and Promer v3.06 alignment algorithms, with the default parameters (Kurtz *et al.*, 2004). The nucleotide identity values of each read against its hit in the genome were used to generate frequency histograms. If the mode was 94% or higher the plot was considered to represent sequence identity against the same species (Konstantinidis and Tiedje, 2005). Stand-alone RPSBlast was used to align reads (translated into all six possible reading frames) to protein profiles (represented by position-specific scoring matrices). Queries were performed against the complete conserved domains database (Marchler-Bauer *et al.*, 2009) and against the COGs (Tatusov *et al.*, 2003) and Tigrfams (Selengut *et al.*, 2007) databases. Fractions of sequences assigned in each case are shown in Supplementary Table 2. TFams classification assignments were integrated into higher hierarchical levels, according to the Tigrfam classification scheme, in subroles and main roles. COGs assignments were also integrated into the higher level of COG's functional categories. In addition, samples were uploaded to the MGRAST server (Meyer *et al.*, 2008) and the functional assignment based on SEED subsystems was retrieved for the three hierarchical levels used: Subsystem, subsystem hierarchy 2 and subsystem hierarchy 1 (bottom up). In all cases, a table containing the counts of functional categories per sample was generated and used for subsequent analysis. All statistical analyses were conducted on *R* (2.6.2). Heat maps of taxonomic composition were generated using the gplots library of *R* (Warnes *et al.*, 2009) with relative frequencies per sample, as well as Euclidean distance, or normal medians. The relative rates of over-represented features present in the people without caries were estimated using a control of the false discovery rate, for testing the amount of false positive predictions (*q*-values) for a given *P*-value of significance, with the algorithm described by White *et al.* (2009).

Taxonomic assignment

16S rRNA sequences were extracted from the reads of each metagenome by similarity search using BLASTn (Altschul *et al.*, 1990) against the RDP database, with an e-value cutoff of 1e–10. Sequences <200 bp were removed. Phylogenetic assignment of the sequences was made using the RDP Classifier (Wang *et al.*, 2007), using an 80% confidence threshold. New operational taxonomic units were proposed if the reads were over 400 bp in length and had a nucleotide identity between 80–95% to known 16S sequences. Taxonomic assignments of all open reading frames were carried out based on a lowest common ancestor (LCA) algorithm (Alstrup *et al.*, 2004) with the characteristics described in the MEGAN software (Huson *et al.*, 2007). We implemented the algorithm in a multi-threaded command-line oriented in-house software in order to obtain faster analysis and simplify its integration in pipelines and downstream analysis. To obtain the LCA of each sequence, we carried out BLASTx homology searches against a custom database comprising the non-eukaryotic sequences of the NCBI's non-redundant database. For each query sequence (read), only hits with a bit score at least 90% of the best matches were considered in the LCA computation. We also made use of the script phymmBL (Brady and Salzberg, 2009) that combines the assignment of sequences both by homology and by nucleotide composition using hidden Markov Models. All the available complete and WGS genomes were retrieved from the human oral microbiome database (Chen *et al.*, 2010), as well as the RefSeq of NCBI containing all bacterial and archaea genomes (june 2010), and were used to build a local database to perform taxonomic model constructions and homology searches, using sequences larger than 200 bp to predict taxonomic affiliation. At this read length, phymmBL's performance at the class level has been estimated to be over 75%. All the taxonomic and functional results were parsed into a MySQL database for further analysis.

Results and discussion

The oral microbiome by pyrosequencing

Supragingival dental plaque samples were taken from six individuals that were divided in three groups according to the number of caries they had suffered and that represented different degrees of oral health: two individuals had never developed caries in their lives (healthy controls), another two individuals had been regularly treated for caries in the past and had a low number of active caries at the moment of sampling; and the last two individuals had a high number of active caries and poor oral hygiene. In addition, samples from individual cavities were collected, and for two of them enough DNA for pyrosequencing was obtained. A total of 1 Gbp of DNA sequence was obtained from the eight samples selected. The amount of human DNA in the metagenomes varied from 0.5–40% in supragingival dental plaque samples (Supplementary Table 1), thus the total size of the studied metagenome was reduced to 842 Mbp of sequence. We obtained an average read length of 425 ± 117 bp, which allowed a functional assignment in a significant fraction of the metagenome (Supplementary Table 2). In addition, assembly of those reads produced 1103 contigs larger than 5 Kb and 354 longer than 10 Kb. Success in the assembly of large contigs was dependent on sequencing effort. We obtained an average of 129.5 Mbp of filtered, high-quality sequences for each of the six oral samples. In the two cavity samples, around 70% of the reads corresponded to human DNA, and an average of 32.5 Mbp of filtered, high-quality reads were obtained.

Estimating diversity in the oral metagenome

We estimated microbial diversity in all samples by three different methods. First, we selected the reads matching 16S rRNA genes, assigning them to different taxonomic levels. A total of 4254 16S rRNA sequences were obtained (Supplementary Table 1), giving a similar picture of diversity to that obtained through 16S rRNA PCR-dependent procedures (Bik *et al.*, 2010), although the relative proportions of each taxonomic group were different (Figure 1). These

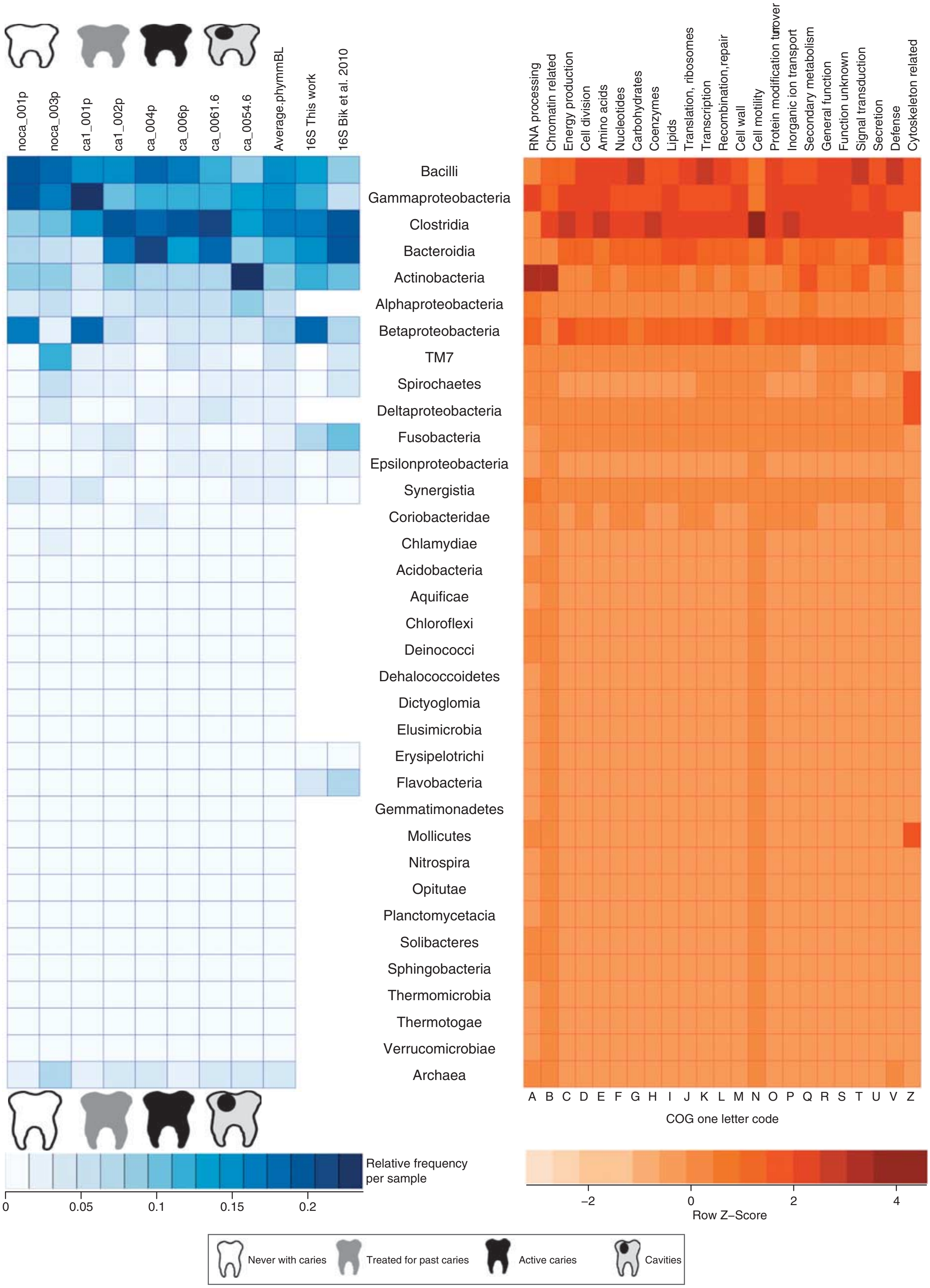


Figure 1 Bacterial diversity in the oral cavity. The graph on the left shows the relative frequency of different bacterial taxa, based on the assignment of the DNA reads by the PhymmBL software and by 16S rRNA reads extracted from the metagenome, and compared with the PCR results obtained by Bik *et al.* (2010). The graph on the right indicates the relative contribution of each taxonomic group to the coding potential of the ecosystem, based on the COGs functional classification system. It can be observed that the functional contribution is not equal among taxa.

16S rRNA reads identified 186 sequences representing novel operational taxonomic units previously undetected by PCR amplification (Supplementary Table 3). Rarefaction curves and different diversity indexes based on the rRNA sequences obtained from the metagenomic reads indicate an estimate of 73–120 genera for dental plaque samples (Supplementary Table 1 and Supplementary Figure 2). A second approach to estimate diversity was the use of a LCA algorithm to classify all reads giving a hit in public databases at the taxonomic level for which the assignment was unambiguous (Huson *et al.*, 2007). Over 1.5 million reads were assigned by this procedure, confirming the presence of bacterial groups detected by 16S rRNA genes, but suggesting that a wider range of taxonomic groups was present (Supplementary Figure 1). Finally, the recently developed phymmBL binning procedure (Brady and Salzberg, 2009) was used to taxonomically assign 1.94 million reads from our data set. The results agreed again with the taxonomic distribution described by the 16S rRNA and the LCA approaches, but with further implication of other bacterial taxa. The results from these three methods show that the relatively small numbers of 16S genes in directly sequenced metagenomes are enough to describe the main taxonomic groups present without cloning or PCR-based biases, although at the expense of lower sequence depth. Some of the taxa found at low proportions in our data set were also detected by large-scale 16S rRNA cloning studies (Paster *et al.*, 2001; Bik *et al.*, 2010) but others were not (Figure 1). This could be not only due to lower amplification efficiency of these bacteria by universal primers, but also due to the detection of false positive hits by the LCA and phymmBL approaches.

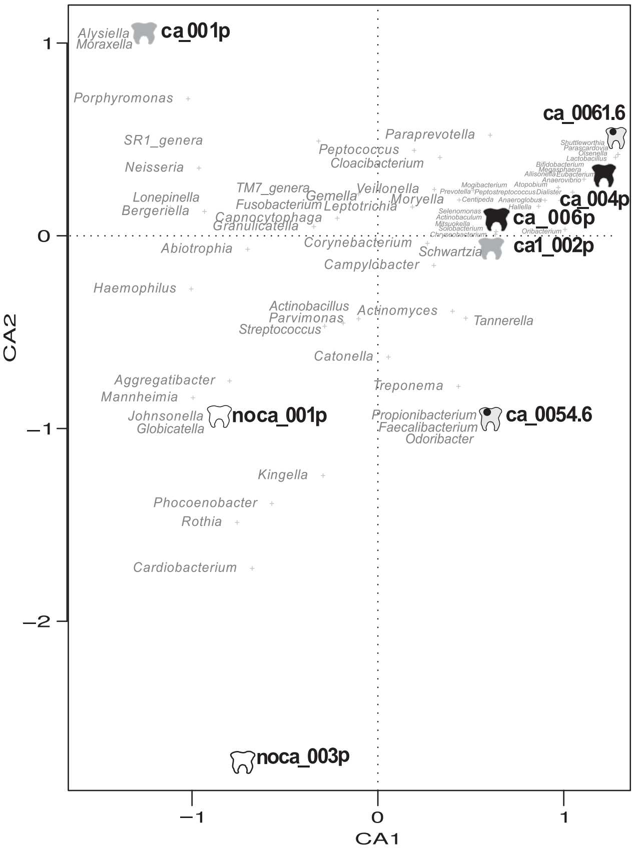


Figure 2 Correspondence analysis (CoA) of the bacterial diversity in oral samples based on 16S rRNA reads extracted from the metagenomes. The first axis successfully separates healthy from diseased individuals. The graph suggests bacterial genera which are potentially associated with absence of caries.

Despite the low number of samples examined, interesting differences in diversity can be seen between healthy and diseased individuals. All three methods showed a tendency for Bacilli and Gamma-Proteobacteria to be more common in healthy individuals, whereas typically anaerobic taxa like Clostridiales and Bacteroidetes are more frequent in diseased samples (Figure 1, Supplementary Figure 1). Bacilli are particularly depleted in the two samples from within cavities, and one of them showed a high proportion of Actinobacteria. Reads assigned to beta-Proteobacteria (mainly Neisseriales) and TM7 were at very low proportions in diseased samples, and studies based on a larger number of individuals should test whether their presence could be associated to healthy conditions. Correspondence analysis between the metagenomes based on the taxonomic assignation by 16S rRNA reads showed that samples with poor oral health tended to cluster together, whereas different consortia of bacteria can be found in healthy individuals (Figure 2). Some genera, like *Rothia* or *Aggregatibacter* appear to be specifically associated to healthy samples, in agreement with PCR-based studies that compared bacterial diversity in healthy controls and diseased volunteers (Aas *et al.*, 2005, 2008; Corby *et al.*, 2005). The metagenomic recruitments also showed *Aggregatibacter* as one of the prevalent species in individuals without caries (see below).

Sequence similarity searches against 18S rRNA databases revealed very few significant hits against eukaryotic species. No rRNA reads were identified from *Candida* or other fungi that are regular inhabitants of the oral cavity, indicating that although these organisms are frequently detected by PCR amplification (Ghannoum *et al.*, 2010), they are probably present at low proportions. In sample CA-04, significant hits to the rRNA ITS region of the protozoan *Trichomonas tenax* were found. *Trichomonas tenax* is found particularly in the oral cavity of patients with poor oral hygiene and advanced periodontal disease (Kleinberg, 2002), and it has been shown to be involved in bronchopulmonary infections.

An effective tool to quantify the presence of selected species in metagenomes is provided by sequence recruitments (Rodriguez-Valera *et al.*, 2009). Individual metagenomic reads that give a hit over a certain identity threshold against a reference bacterial genome are 'recruited' to plot a graph, which will vary in density depending on the abundance of that organism in the sample. If the average nucleotide identity displayed is above 94%, the recruitment is very likely made against reads of the same species (Konstantinidis and Tiedje, 2005). By comparing our metagenomes against the genomes of 1117 fully sequenced genomes available in databases, we were able to estimate the abundance of close relatives of these reference species in our samples (Supplementary Figure 3A). Interestingly, bacteria closely related to *Aggregatibacter* and *Streptococcus sanguis* were among the three with the highest level of recruitment in individuals without caries, in agreement with these species being more frequently amplified from the oral cavity of healthy individuals (Aas *et al.*, 2005; Corby *et al.*, 2005). On the other hand, *Streptococcus gordonii* and *Leptotrichia buccalis* were abundant in individuals with caries. Strains of *Veillonella parvula* were the most abundant in all individuals with caries and appeared to be common to all samples, but interestingly the recruitment plots show differences between strains (Supplementary Figure 4). For instance, the *Veillonella* present in the two healthy individuals shows a genomic island without recruitment, even at the protein level, between positions 2066–2094 Kb of the reference genome. Individuals with caries CA-04 and CA1-01 do contain this region, which includes CRISPR-associated genes, hypothetical proteins, a protein involved in DNA uptake and an amidophosphoribosyltransferase. This way, differences between strains of the same species can be identified which would pass unnoticed by 16S rRNA studies, and future work should identify whether those differential genes might be involved in pathogenesis. In addition, recruitment plots indicate that few taxa are normally dominant in each metagenome (Supplementary Figure 3B). This suggests that although bacterial diversity is indeed very large in the oral cavity, very few taxa account for most of the bacterial cells, and a big portion of the identified species are present at very low densities.

Functional diversity in the oral ecosystem

One of the powerful applications of LCA and phymmBL approaches is that each read with a significant hit can be assigned a taxonomic origin, and at the same time can also be related in many cases to a putative function. By relating taxonomy to function we have been able to predict what ecological or metabolic role each bacterial group can have. An example of this 'who can do what' approach can be seen in Figure 1 by using the COGs function classification system. It shows that categories are not equally distributed, and that some taxonomic groups are especially endowed for performing concrete functions. For example, a large portion of genes involved in defence mechanisms (that is, restriction endonucleases and drug efflux pumps) appear to be encoded by Bacilli. Other functions unequally distributed were cell motility genes in Clostridiales (mainly flagellar proteins) or signal transduction and carbohydrate metabolism in Bacilli (Figure 1, right). A more detailed functional analysis of the metagenome was performed using several systems for gene classification at different hierarchical levels. All pyrosequencing reads were compared against the conserved domains database, the Subsystems annotation environment (SEED) and the Tigrfams profiles (see Materials and methods section). Correspondence analysis (CoA) of the eight samples according to the functional assignment of the reads gave similar clustering patterns for the three function classification systems (Supplementary Figure 5). Samples from diseased individuals tended to cluster together, indicating that a similar set of functions were encoded in their metagenomes, and the two samples from individuals that had never suffered from caries, together with sample CA1-01 (with only one cavity at the moment of sampling), could be separated from the rest by the principal component. When the functional assignment of the oral microbiome was compared with that of the adult gut microbiome (Kurokawa *et al.*, 2007) a χ^2-test of independence revealed that the overall gut and oral functional roles depicted in the RAST subsystems are significantly different ($\chi^2_{(\mathrm{df}=158)} = 17\,057.42$, $P < 2.2\mathrm{e}{-16}$, $\phi = 0.123$), and this was supported also by clustering analysis where the oral samples clustered together (Figure 3), indicating that the gut and the mouth are two different ecosystems in terms of the relative frequencies of functions encoded in their metagenomes. It had previously been shown that the taxonomic diversity of the gut and oral ecosystems is clearly distinct (Bik *et al.*, 2010), despite the fact that clear examples of horizontal gene transfer have been shown between these two interconnected niches (Mira, 2007). Our data show large blocks of over-represented functions in the gut microbiome, while others appear over-represented in the oral samples (a detailed list of these functional categories is represented in Supplementary Figure 6). It is interesting to note that metabolic genes, like those involved in sugar uptake and assimilation, are enriched in gut bacteria together with adhesion proteins and prophage genes, whereas gene families related to oxidative and osmotic stress or iron scavenging are more frequent in the oral microbiome (Figure 3). Thus, the relative proportion of these functional categories provides important insights into the ecology of each ecosystem and the potential role of the corresponding microbiotas for human health.

Within the oral samples, individuals are clustered according to their health status (Figure 3). From an applied viewpoint, it is interesting that several functional categories are over-represented in samples from individuals without caries. Remarkable

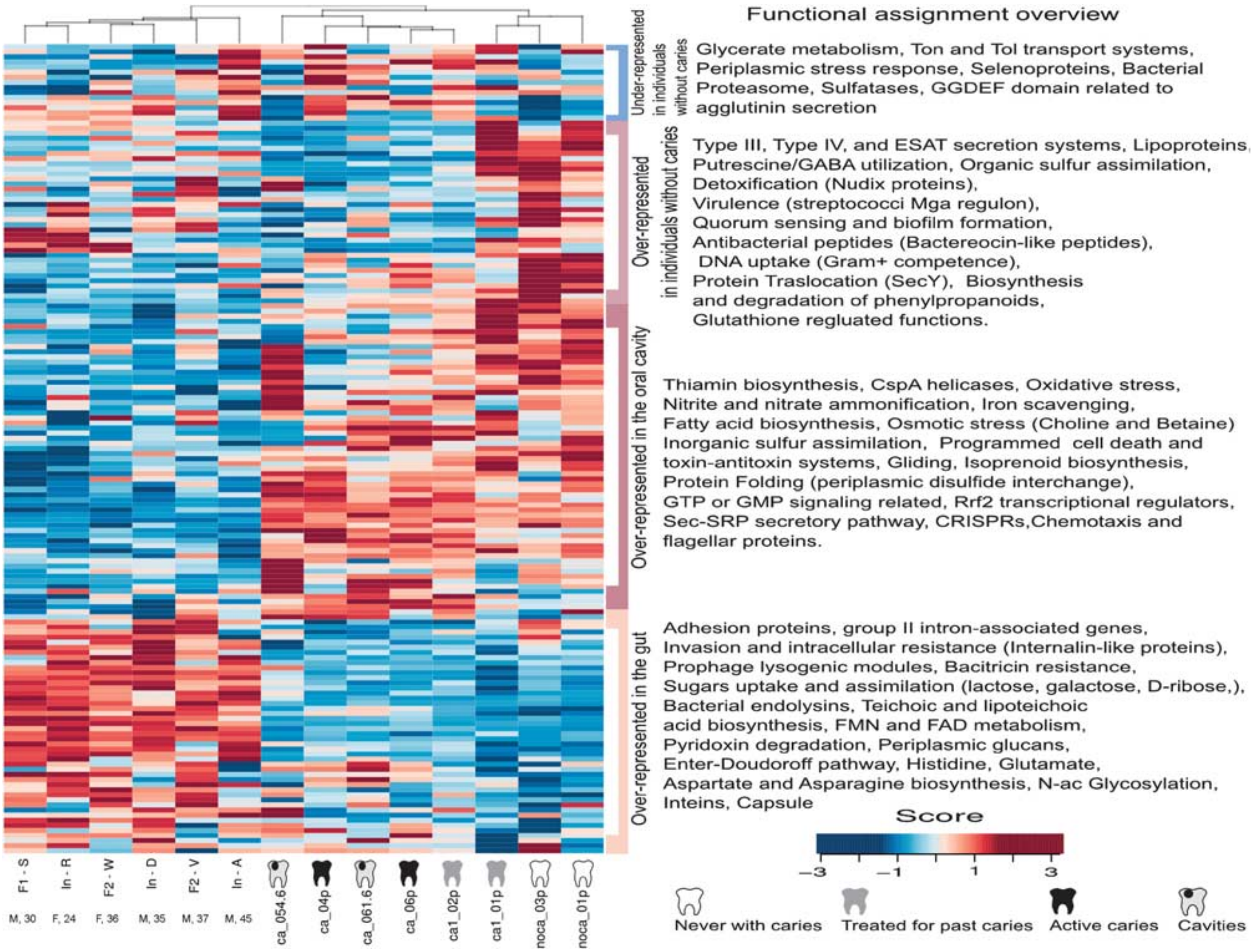


Figure 3 Functional profiles from oral and adult-gut metagenomic samples. Classification was based on Subsystem hierarchy 2 of MG-RAST. Counts were normalized to the total number of reads per sample and then normalized by function. Blue to red gradient indicates levels of under/over-representation. Large blocks of gene categories are over-represented in each of the two microbiotas, indicating that the gut and the oral cavity are two functionally distinct ecosystems. Within the oral microbiome, some functional roles are over-represented in individuals without caries. A full version of this figure indicating all 101 functional categories is included in Supplementary Figure 6. Sequences from the healthy adult-gut metagenomes were taken from Kurokawa *et al.* (2007). The age and sex of each individual are indicated below each label.

uprepresented genes in healthy individuals are involved in antibacterial peptides like bacteriocins (*P*-value = 2.95 e−7; *q*-value = 4.63 e−8), periplasmic stress response genes like *degS, degQ* (*P* = 2.46 e−46; *q* = 3.22 e−46), capsular and extracellular polysaccharides (*P* = 7.04 e−5; *q* = 8.5 e−6) and bacitracin stress response genes (*P* = 3.4 e−3; *q* = 3.24 e−4). Other functional categories were also over-represented but the difference was not statistically significant, like genes involved in quorum sensing and phospholipid metabolism. The higher presence of bacteriocin-related genes points at these bioactive compounds as promising potential anticaries agents. Some gene features over-represented in individuals with active caries are involved in mixed-acid fermentation (*P* = 2.85 e−260; *q* = 2.65 e−259) and DNA uptake and competence (*P* = 6.29 e−8; *q* = 1.13 e−8). Finally, it must be underlined that some over-represented genes in healthy individuals have an unknown function, and future studies should elucidate whether they are involved in the protection of the teeth against cariogenic conditions.

Cavities are complex ecosystems

We were able to extract sufficient DNA for 454 pyrosequencig in two samples from individual teeth, one at an intermediate stage and the other one at an advanced stage of caries development (dentin lesion). Given that mutans streptococci initially were considered to be the main ethiological agents of dental caries (Loesche, 1986), it is not surprising that most strategies against this disease have aimed at targeting *Streptococcus mutans*. These include the development of a vaccine using known surface antigens, passive immunization strategies that could neutralize the bacterium, the co-aggregation of *S. mutans* to probiotic strains or the use of specific inhibitors of *S. mutans* proteins, among others (Russell *et al.*, 2004). In addition, the

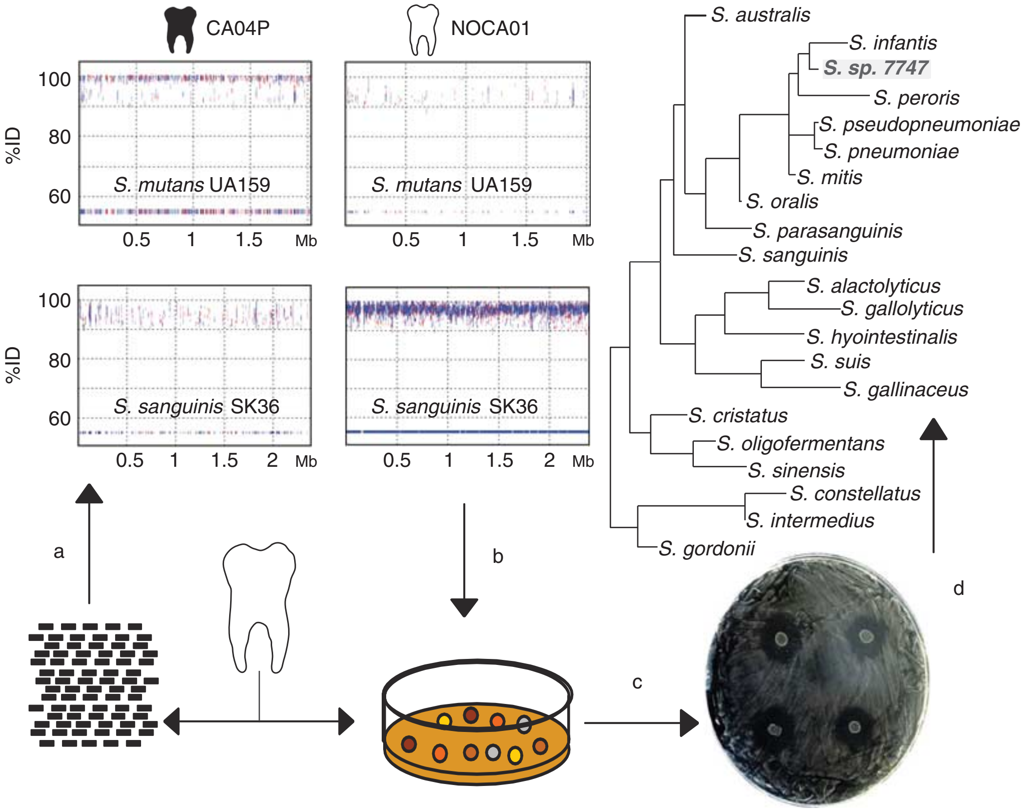


Figure 4 Searching of bacterial strains with a potential antagonistic effect against cariogenic bacteria. Metagenomic recruitment plots are used to detect the species (**a**), which are at low frequencies in individuals with caries but are among the most common in caries-free subjects. These species are then selected based on culture conditions and microscopic examination (**b**). The isolates are grown in solid media to provide an inhibition screening against caries-producing bacteria (**c**), selecting the strains that display inhibition rings (**d**), such as the *Streptococcus* strain 7747. Sequencing the genome of these inhibitory strains and comparing it against the metagenome of caried individuals must confirm that these strains are absent under diseased conditions.

presence of mutans streptococci in children is typically associated to caries risk in oral-health evaluation protocols (Ge *et al*., 2008). However, pioneering molecular-based studies of cavities have failed to amplify mutans streptococci by PCR or hybridization in a significant proportion of cavities, suggesting that other bacterial genera like *Lactobacillus*, *Actinomyces* or *Bifidobacterium* could be involved in the disease (Aas *et al*., 2008; Becker *et al*., 2002). Recent molecular work has confirmed this finding and expanded the list of potential cariogenic bacteria to other species like *Veillonella*, *Propionibacterium* and *Atopobium* (Aas *et al*., 2008), most of them are poorly studied bacteria. The metagenomes of cavities studied here showed an almost complete absence of *S. mutans*. However, they displayed a large taxonomic diversity, which are included among the most common genera, *Veillonella, Corynebacterium* or *Leptotrichia* (Supplementary Table 4). Some of these bacteria, particularly *Veillonella*, have been shown to be predominant at all stages of caries progression (Aas *et al*., 2008) and under high-glucose conditions, and appear to be implied in acid production (Bradshaw and Marsh, 1998). Interestingly, consortia between *Veillonella alcalescens* and *S. mutans* were shown to produce more acid than any one of these species separately (Noorda *et al*., 1988), suggesting that synergistic effects probably take place, as it has been demonstrated in other complex microbial communities. Thus, although these data are based on the metagenomes from only two cavities, they favour a nonspecific plaque hypothesis for the development of dental caries (Marsh, 1994; Kleinberg, 2002). Further work should elucidate the potential role these bacteria had other than mutans streptococci in the progression of caries, as well as their synergistic and antagonistic interactions. The forecoming improvements in the amount of DNA required for next-generation sequencing techniques will allow a metagenomic study of cavities at different stages of development, including initial, white-spot lessions. This is important because mutans streptococci could be instrumental at initial stages of caries, after which other species could colonize the niche. If caries is confirmed to be a polymicrobial disease, this should be taken into account for future therapeutic strategies. For instance, a potential solution for immunization strategies could pass through the selection of vaccine targets shared by different pathogens involved in the process of tooth decay (Mira *et al*., 2004; Mira, 2007).

Search for potential probiotics through metagenomics
The existence of a small proportion of the human adult population that has never suffered from dental caries has led some authors to suggest the presence of some bacterial species with a potential antagonistic effect against cariogenic bacteria (Corby *et al.*, 2005). Bacterial replacement of pathogenic strains by innocuous isolates obtained from healthy individuals has been successfully shown to prevent pharynx infections and is the basis for probioticts preventing infectious disease in the gut and other human niches (Tagg and Dierksen, 2003). Metagenomic recruitment of cariogenic bacteria against the oral microbiome of healthy individuals shows a complete absence of *S. mutans* and *S. sobrinus*. Interestingly, the lack of detection of the cariogenic bacteria is accompanied by an intense recruitment of other streptococci (mainly those related to *S. sanguis*) and *Neisseria*, which comprise the most abundant genera in these individuals (Supplementary Figure 3B). Given the possibility that isolates of these dominant genera could be involved in antagonistic interactions with cariogenic bacteria, fresh dental plaque samples from 10 healthy individuals (including those from which the metagenomic sequences were obtained) were collected and used for culturing under conditions optimal for the growth of neisserial and streptococcal species. After microscopic examination, diplococci and streptococci were selected, providing a collection of 249 isolates. Those that could be grown on the same culture medium as *S. mutans* and *S. sobrinus* were transferred to a loan culture of these cariogenic bacteria. This simple screening identified 16 strains that displayed inhibition rings (Figure 4). PCR amplification of the 16S rRNA gene identified most of them as streptococci, with a 96–99% sequence identity to *S. oralis*, *S. mitis* and *S. sanguis*. Thus, this metagenomic approach allowed us to quantify the most abundant bacteria and confirms the previously hypothesized presence of bacteria with a protective effect against cariogenic species. This effect appears to be direct (that is, inhibitory), but other indirect effects such as stimulation of the immune response or direct competition for the same substrate or niche cannot be ruled out. Future research on these isolates should aim at identifying the secreted compounds responsible for the inhibition of caries-producing bacteria, and metagenomic libraries of dental plaque DNA may prove useful in this respect (Seville *et al.*, 2009). Our own inhibition screenings performed on metagenomic fosmid libraries from dental plaque of healthy individuals against cariogenic bacteria suggest that antimicrobial peptides are among the products causing the inhibition. We propose the probiotic use of these anti-cariogenic bacteria or the utilization of the antibiotics they encode as promising new therapies against dental caries and other oral diseases (Devine and Marsh, 2009).

Conclusion

We have shown that the direct pyrosequencing of human samples is a feasible approach to study the human microbiome, which would obviate the biases imposed by cloning and PCR and that would provide a more complete view of human-related bacterial communities beyond their composition inferred from the 16S rRNA gene (Ghai *et al.*, 2010; Xie *et al.*, 2010). Even in samples with a large proportion of human DNA such as cavities, the large throughput of next-generation sequencing has provided enough sequences to gain insights into the microbiology of caries, suggesting that it is the outcome of a complex bacterial community. Despite the limited number of samples analyzed in this first study, important differences between healthy and diseased sites and individuals can be observed at the taxonomic and functional level, suggesting that the dental plaque of individuals that have never suffered from caries can be a genetic reservoir of new anticaries compounds and probiotics. Future population-based studies must evaluate whether the trends described in this study hold when higher sample sizes are used. We hope that these results stimulate further sequencing of the oral metagenome and metatranscriptome in the future as a tool to understand and combat the development of oral diseases.

Acknowledgements

This work was funded by the following projects from the Spanish MICINN: SAF2009-13032-C02-02 from the I+D program, BIO2008-03419-E from the EXPLORA program and MICROGEN CSD2009-00006 from the Consolider-Ingenio program. We thank Professor F Rodriguez-Valera and Professor Siv G Andersson for their advice and comments, and three anonymous referees for their constructive comments that significantly improved the manuscript.

References

Aas JA, Griffen AL, Dardis SR, Lee AM, Olsen I, Dewhirst FE *et al.* (2008). Bacteria of dental caries in primary and permanent teeth in children and young adults. *J Clin Microbiol* **46**: 1407–1417.

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.

Alstrup S, Gavoille C, Kaplan HRT. (2004). Nearest common ancestors: a survey and a new Algorithm for a distributed environment. *Theory Comp Syst* **37**: 441–456.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Becker MR, Paster BJ, Leys EJ, Moeschberger ML, Kenyon SG, Galvin JL *et al.* (2002). Molecular analysis of bacterial species associated with childhood caries. *J Clin Microbiol* **40**: 1001–1009.

Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF *et al.* (2010). Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J* **4**: 962–974.

Bradshaw DJ, Marsh PD. (1998). Analysis of pH-driven disruption of oral microbial communities *in vitro*. *Caries Res* **32**: 456–462.

Brady A, Salzberg SL. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6**: 673–676.

Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. (2010). The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database: J Biol Databases Curation* **2010**: baq013.

Corby PM, Lyons-Weiler J, Bretz WA, Hart TC, Aas JA, Boumenna T *et al.* (2005). Microbial risk indicators of early childhood caries. *J Clin Microbiol* **43**: 5753–5759.

Darveau RP. (2010). Periodontitis: a polymicrobial disruption of host homeostasis. Nature reviews. *Microbiology* **8**: 481–490.

de Lillo A, Ashley FP, Palmer RM, Munson MA, Kyriacou L, Weightman AJ *et al.* (2006). Novel subgingival bacterial phylotypes detected using multiple universal polymerase chain reaction primer sets. *Oral Microbiol Immunol* **21**: 61–68.

Devine DA, Marsh PD. (2009). Prospects for the development of probiotics and prebiotics for oral applications. *J Oral Microbiol* **1**: 1–11.

Ge Y, Caufield PW, Fisch GS, Li Y. (2008). *Streptococcus mutans* and *Streptococcus sanguinis* colonization correlated with caries experience in children. *Caries Res* **42**: 444–448.

Ghai R, Martin-Cuadrado AB, Molto AG, Heredia IG, Cabrera R, Martin J *et al.* (2010). Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J* **4**: 1154–1166.

Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A *et al.* (2010). Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathogens* **6**: e1000713.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS *et al.* (2006). Metagenomic analysis of the human distal gut microbiome. *Science (New York, NY)* **312**: 1355–1359.

Gomez-Alvarez V, Teal TK, Schmidt TM. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3**: 1314–1317.

Grice EA, Kong HH, Renaud G, Young AC, Bouffard GG, Blakesley RW *et al.* (2008). A diversity profile of the human skin microbiota. *Genome Res* **18**: 1043–1050.

Huson DH, Auch AF, Qi J, Schuster SC. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

Keijser BJF, Zaura E, Huse SM, van der Vossen JMBM, Schuren FHJ, Montijn RC *et al.* (2008). Pyrosequencing analysis of the oral microflora of healthy adults. *J Dental Res* **87**: 1016–1020.

Kleinberg I. (2002). A mixed-bacteria ecological approach to understanding the role of the oral bacteria in dental caries causation: an alternative to Streptococcus mutans and the specific-plaque hypothesis. *Crit Rev Oral Biol Med* **13**: 108–125.

Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**: 2567–2572.

Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A *et al.* (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**: 169–181.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C *et al.* (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osterås M *et al.* (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* **79**: 266–271.

Loesche WJ. (1986). Role of Streptococcus mutans in human dental decay. *Microbiol Rev* **50**: 353–380.

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH *et al.* (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**: D205–D210.

Marsh PD. (1994). Microbial ecology of dental plaque and its significance in health and disease. *Adv Dental Res* **8**: 263–271.

Marsh PD. (2006). Dental plaque as a biofilm and a microbial community—implications for health and disease. *BMC Oral Health* **6**(Suppl 1): S14.

Marsh PD. (2010). Microbiology of dental plaque biofilms and their role in oral health and caries. *Dental clin North Am* **54**: 441–454.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinfo* **9**: 386.

Mira A. (2007). Horizontal gene transfer in oral bacteria. In: Rogers AH (ed). *Oral Molecular Microbiology*. Horizon Scientific Press: Norfolk, UK, pp 65–85.

Mira A, Pushker R, Legault BA, Moreira D, Rodríguez-Valera F. (2004). Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics. *BMC Evol Biol* **4**: 50.

Noorda WD, Purdell-Lewis DJ, van Montfort AM, Weerkamp AH. (1988). Monobacterial and mixed bacterial plaques of Streptococcus mutans and Veillonella alcalescens in an artificial mouth: development, metabolism, and effect on human dental enamel. *Caries Res* **22**: 342–347.

Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA *et al.* (2001). Bacterial diversity in human subgingival plaque. *J Bacteriol* **183**: 3770–3783.

Petersen PE. (2004). [Continuous improvement of oral health in the 21st century: the approach of the WHO Global Oral Health Programme]. *Zhonghua kou qiang yi xue za zhi = Zhonghua kouqiang yixue zazhi = Chin J Stomatol* **39**: 441–444.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59–65.

Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM *et al.* (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.

Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F *et al.* (2009). Explaining microbial population genomics through phage predation. Nature reviews. *Microbiology* **7**: 828–836.

Russell MW, Childers NK, Michalek SM, Smith DJ, Taubman MA. (2004). A Caries Vaccine? The state of the science of immunization against dental caries. *Caries Res* **38**: 230–235.

Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC *et al.* (2007). TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* **35**: D260–D264.

Seville LA, Patterson AJ, Scott KP, Mullany P, Quail MA, Parkhill J *et al.* (2009). Distribution of tetracycline and erythromycin resistance genes among human oral and fecal metagenomic DNA. *Microbial Drug Resist (Larchmont, NY)* **15**: 159–166.

Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL. (1998). Microbial complexes in subgingival plaque. *J Clin Periodontol* **25**: 134–144.

Tagg JR, Dierksen KP. (2003). Bacterial replacement therapy: adapting 'germ warfare' to infection prevention. *Trends Biotechnol* **21**: 217–223.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV *et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinfo* **4**: 41.

Vaishampayan PA, Kuehl JV, Froula JL, Morgan JL, Ochman H, Francino MP. (2010). Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome Biol Evol* **2010**: 53–66.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T *et al.* (2009). gplots: Various R programming tools for plotting data. The Comprehensive R Archive Network. http://cran.r-project.org/package=gplots.

Watabe K, Nishi M, Miyake H, Hirata K. (1998). Lifestyle and gastric cancer: a case-control study. *Oncol Rep* **5**: 1191–1194.

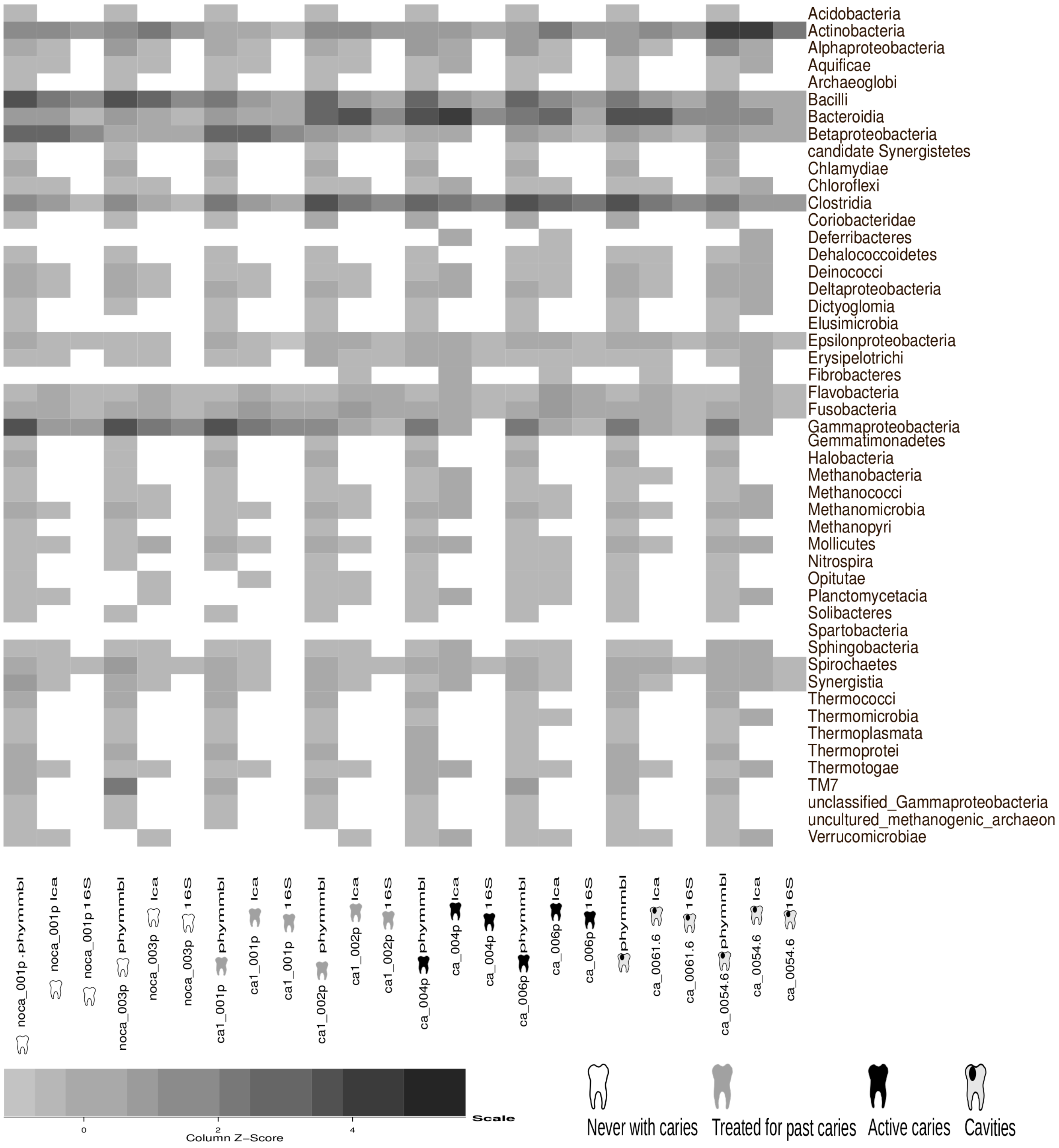
White JR, Nagarajan N, Pop M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples (CA Ouzounis, Ed.) *PLoS Comput Biol* **5**: e1000352.

Wu T, Trevisan M, Genco RJ, Dorn JP, Falkner KL, Sempos CT. (2000). Periodontal disease and risk of cerebrovascular disease: the first national health and nutrition examination survey and its follow-up study. *Arch Int Med* **160**: 2749–2755.

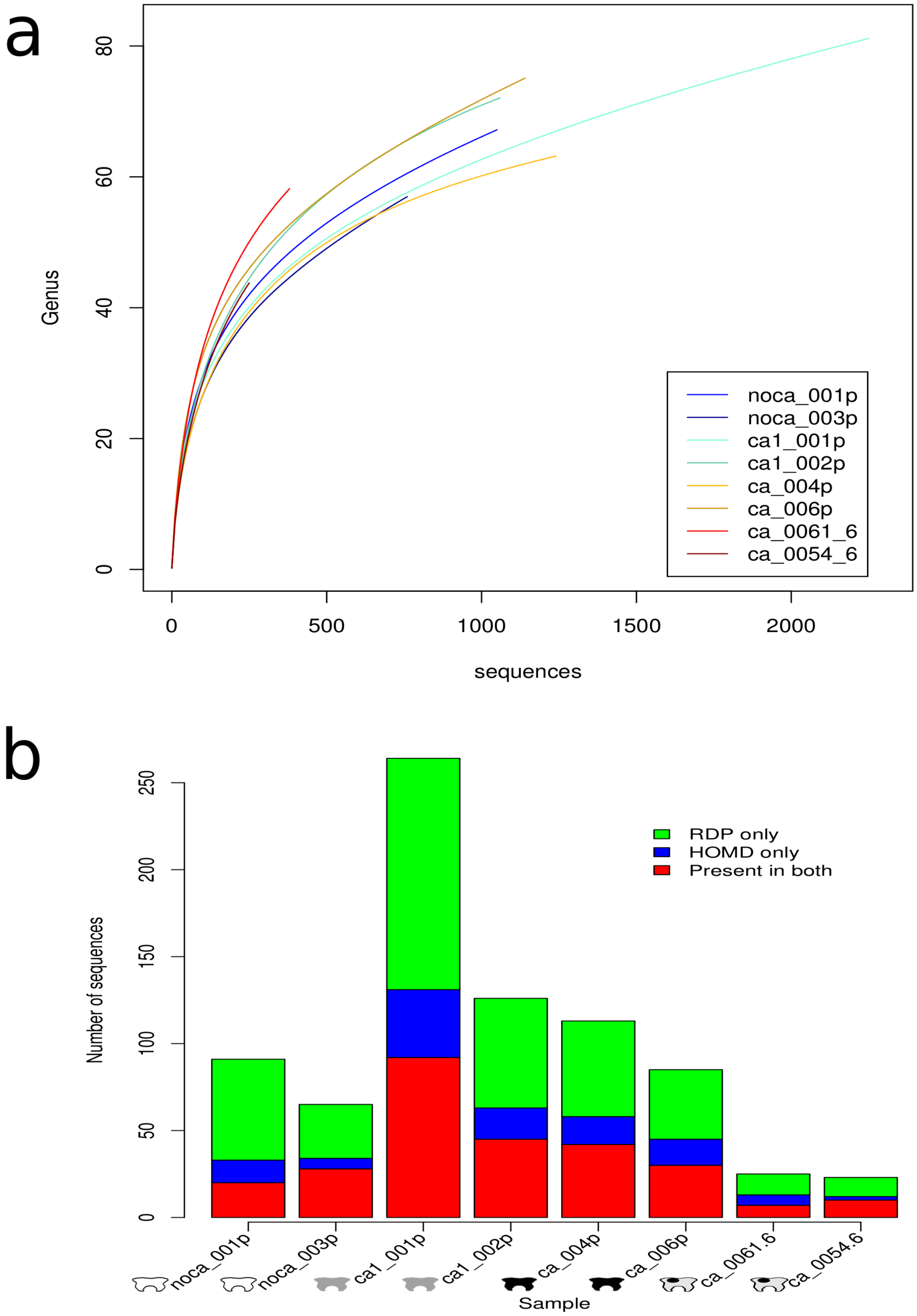
Xie G, Chain PSG, Lo C-C, Liu K-L, Gans J, Merritt J *et al.* (2010). Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Mol Oral Microbiol* **25**: 391–405.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

Belda-Ferre *et al.* Supplementary Fig 1

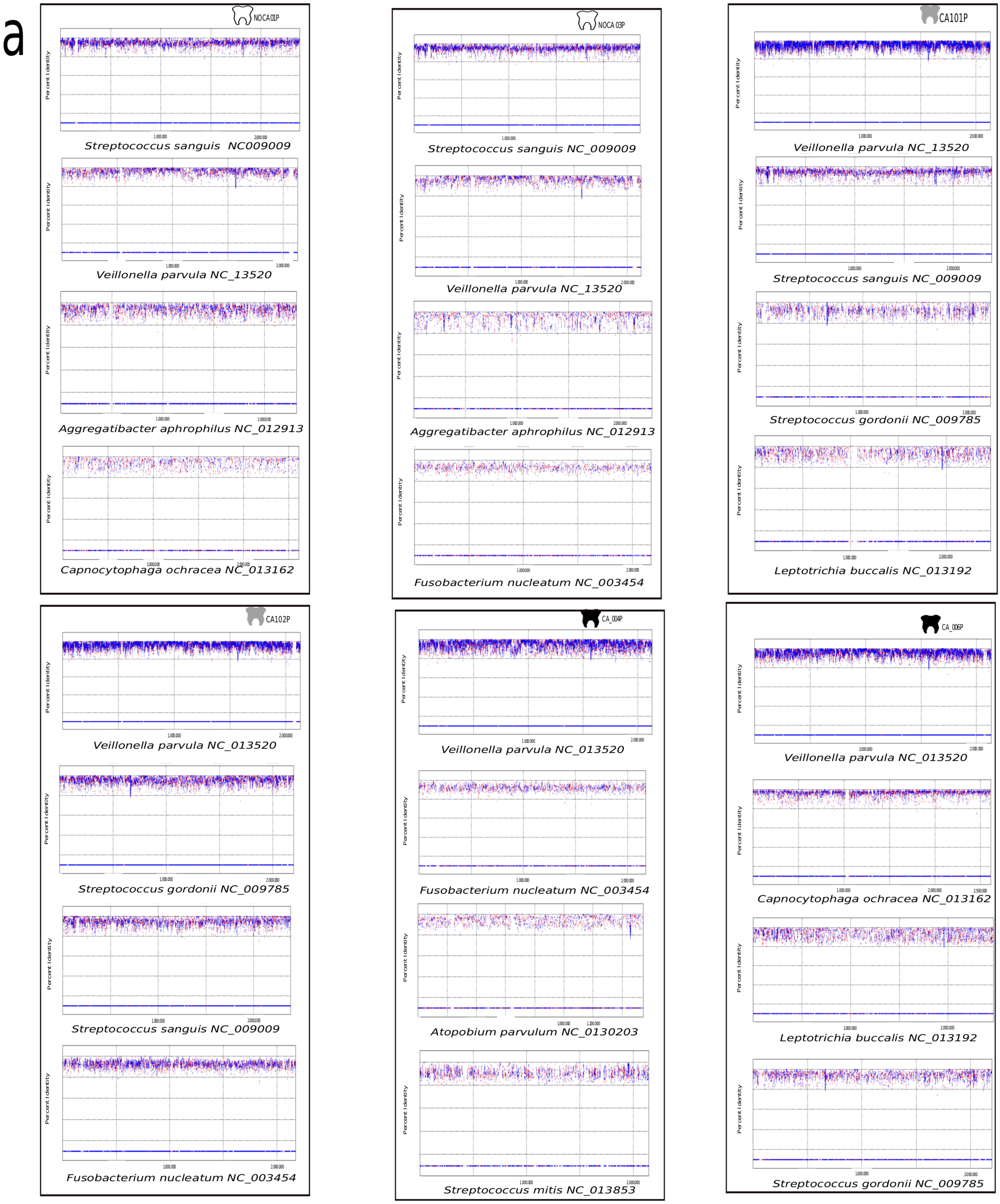


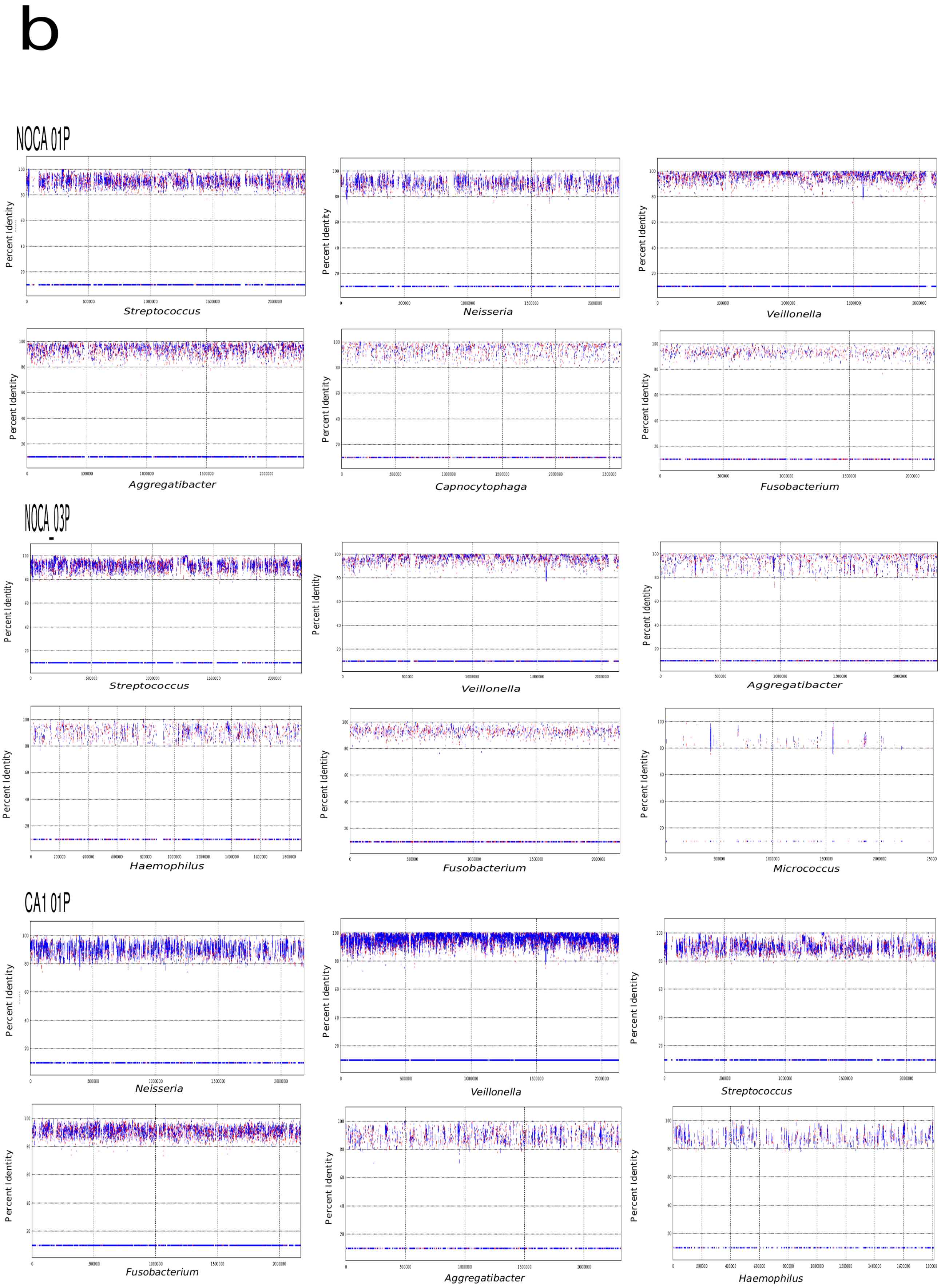
Supplementary Fig 1 | Bacterial diversity in the metagenome of the oral cavity by 16S rRNA, LCA and phymmBL approaches. Similar results are obtained for the most common microbial groups, but 16S rRNA reads detect a smaller fraction of taxa**.**

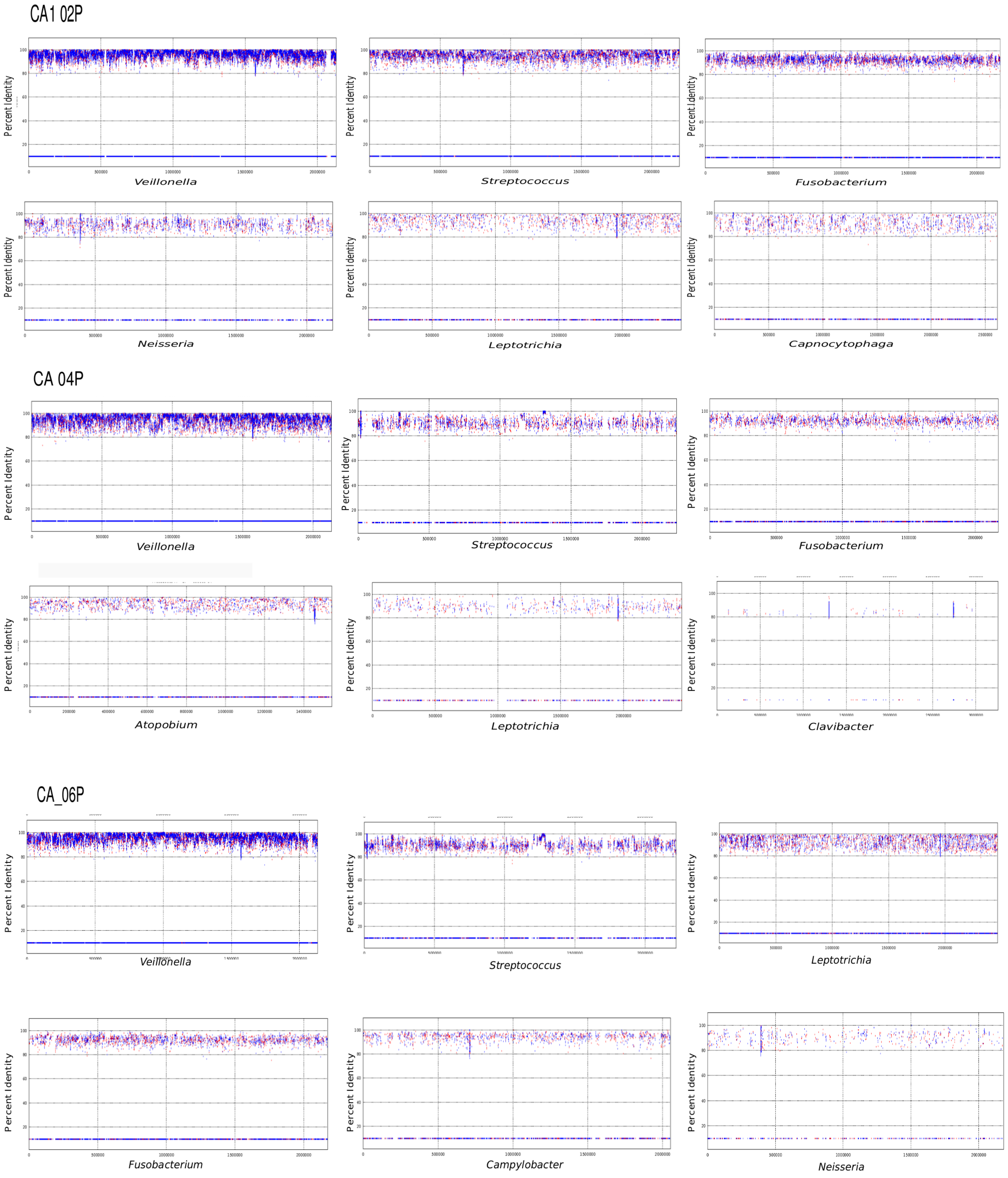


Supplementary Figure 2. (a) Rarefaction curves for the 8 metagenomic samples studied. Curves were estimated based on the 16S rRNA reads extracted from the metagenomes. No clear differences in diversity are observed between the health status of the samples. (b) New Operational Taxonomic Units (OTUs) in the oral cavity. Sequences corresponding to 16S rRNA genes were BLASTed against the Ribosomal Database Project (RDP) and the Human Oral Microbiome Database (HOMD) using reads which were over 400 bp and which aligned over 90% of the length. A read was considered to represent a new OTU if the sequence identity against its top hit was between 80-95%. The graphs show the number of new OTUs with hits in the RDP, the HOMD or both. A complete list is available from the authors upon request.

Belda-Ferre et al. Supplementary Fig.3

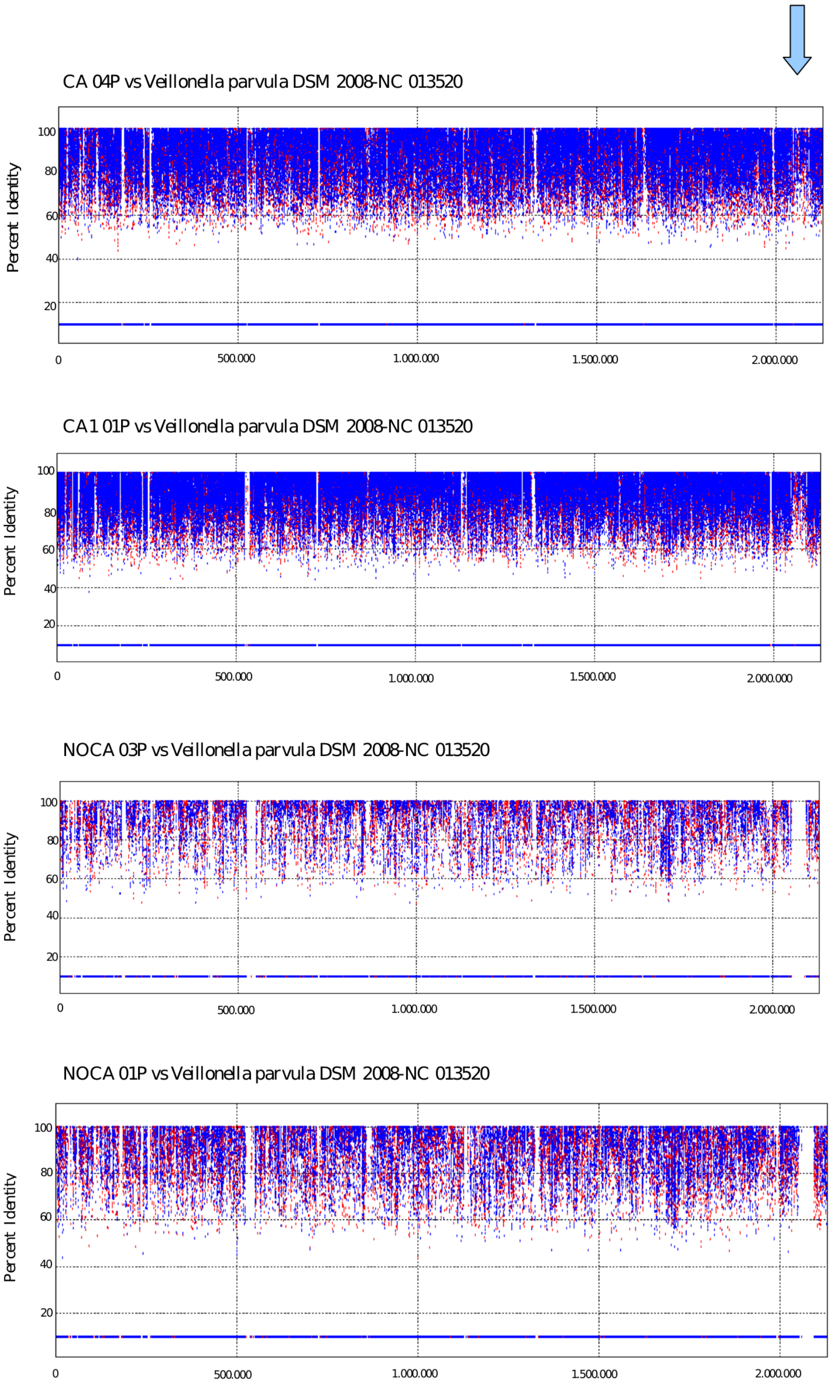



b
NOCA 01P
Percent Identity
100
80
60
40
20
0
500000
1000000
1500000
2000000
2500000
Streptococcus
Neisseria
Veillonella
Aggregatibacter
Capnocytophaga
Fusobacterium
NOCA 03P
Streptococcus
Veillonella
Aggregatibacter
Haemophilus
Fusobacterium
Micrococcus
CA1 01P
Neisseria
Veillonella
Streptococcus
Fusobacterium
Aggregatibacter
Haemophilus



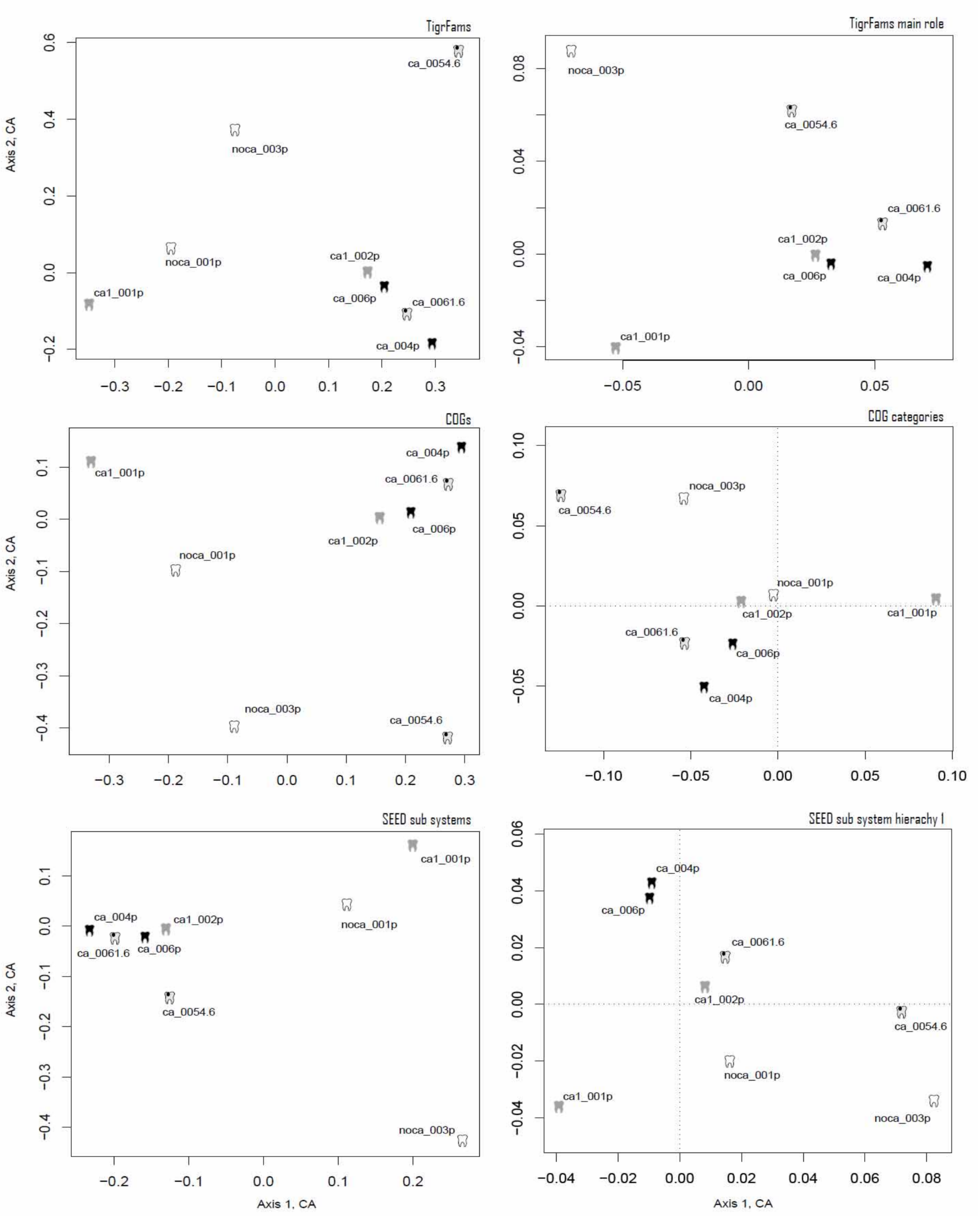
Supplementary Fig. 3 | Metagenomic recruitment of the 4 most common species per sample (**a**), and the 6 most common genera (**b**) against six metagenomes corresponding to supragingival dental plaque samples. A species was assumed to be present in the metagenome if the mode of the frequency distribution of percent identity was above 94%, following Konstantinidis and Tiedje[31]. A genus was assumed to be present in the metagenome if the mode of the frequency distribution of percent identity was above 90%. For genera recruitment plots, the graph corresponding to the most common species for each genus was selected as a representative

Belda-Ferre *et al.* Supplementary Fig 4



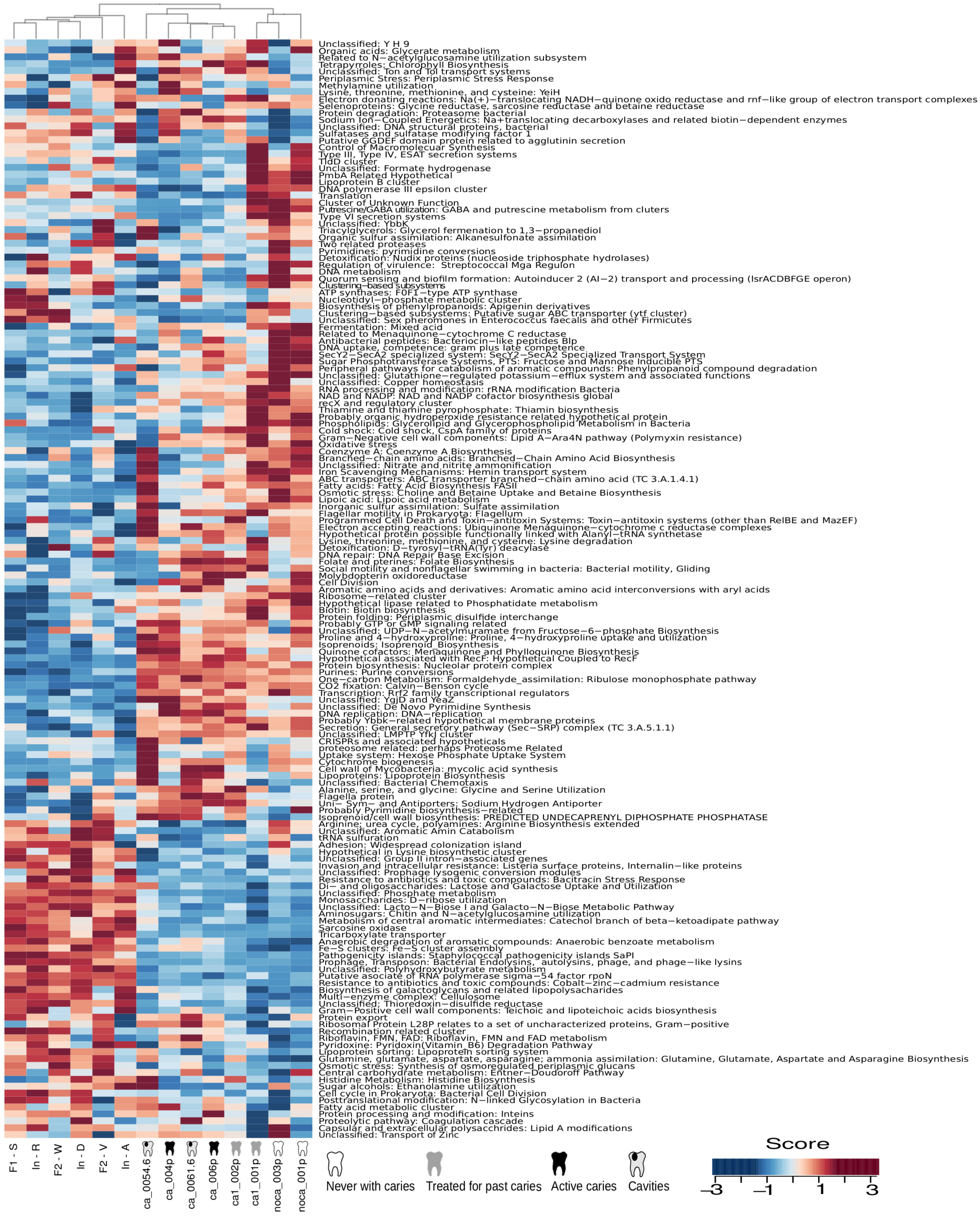
Supplementary Fig. 4 | Metagenomic recruitment of the species *Veillonella parvula* at the protein level in four supragingival dental plaque samples from the oral microbiome. Plots were done using Promer (see supplementary methods). The arrow indicates a genomic island that is absent in the two individuals without caries, showing differences in gene content between strains of this species inhabiting different individuals. Genes in the island were still present in samples CA_04P and CA1_01P under lower levels of coverage equivalent to those of samples NOCA_01P and NOCA_03P.

Belda-Ferre *et al.* Supplementary Fig 5



Supplementary Fig. 5 | Correspondence Analyses (CoAs) of the functional annotation of oral samples. This figure depicts the plot of the first 2 axes of a CoA performed on a table of frequencies of functional classes per sample. Six independent analyses were done over different annotations systems: TigrFam, COG and SEED Subsystems in the right column; and their respective integration into higher hierarchies in the left column: TigrFam main roles, COG categories, and Subsystem Hierarchy 1. Similar clustering of the samples were obtained with the different methods

Belda-Ferre *et al.*, Supplementary Fig. 6



Supplementary Fig. 6 | Functional profiles from oral and gut metagenomic samples. Classification was based on SEED Subsystem Hierarchy 2. Counts were first normalized to the total number of reads per sample and then normalized by function. Blue to red gradient indicates levels of under/ over-representation. Large blocks of gene categories are over-represented in each of the two microbiotas, indicating that the gut and the oral cavity are two functionally distinct ecosystems. Within the oral samples, some categories are over-represented in individuals without dental caries.

Supplementary Table 1. Features of oral samples and their metagenomes.

Age	Sex	Sample[1]	CAO's Index[2]	Number of reads	% Replic[3]	% Human DNA	Total Mbp	Contigs >5kbp	Largest contig	N50 Contig Size	16S reads[4]	Simpson Index[5]	Shannon Index[5]	Chao1 Index[6]	ACE Index[7]
23	Male	NOCA_01P	0	347927	1.2	40.59	77.54	13	12856	898	543	0.93	3.19	100 ± 24.6	94.6 ± 4.9
39	Male	NOCA_03P	0	330073	3.95	22.76	100.13	49	43857	1083	374	0.91	2.94	92 ± 28.4	83.7 ± 4.8
36	Male	CA1_01P	8 (1)	494659	3.93	2.23	203.71	657	46856	2230	1160	0.94	3.21	120 ± 24.8	120.4 ±5.8
29	Male	CA1_02P	6 (4)	315892	3.95	2.74	129.85	154	15919	1071	575	0.92	3.11	85.2 ± 9	89.9 ± 4.7
36	Fem	CA_04P	25 (15)	402049	4.54	11.54	142.37	181	19835	939	663	0.89	2.89	74.4 ± 9.9	73.5 ± 4.2
49	Male	CA_06P	11 (8)	354192	2.88	10.83	123.27	47	51033	872	615	0.95	3.38	129.2 ± 41	115.9 ±5.8
49	Male	CA_06_1.6	11 (8)	305820	3.25	66.97	37.52	0	3376	667	194	0.92	3.21	77 ± 13.3	77.1 ± 4.3
42	Male	CA_05_4.6	10 (7)	291162	3.19	74.99	27.67	2	29784	661	130	0.88	2.82	55.3 ± 8.3	66.3 ± 4.6

[1] Samples marked with "P" indicate suprangingival dental plaque samples. Samples with a number code indicate the tooth from which the cavity sample was taken, following the international WHO nomenclature

[2] Number of caried, absent and obstructed teeth (wisdom teeth were excluded). Number between brackets indicate the number of exposed caries

[3] Proportion of filtered artificial replicates during pyrosequencing

[4] Number of 16S rRNA sequences detected in the metagenome and assigned by the RDP classifier.

[5] Diversity indexes were calculated at the genus level based on 16S rRNA sequences extracted from the metagenomes

[6] Data indicate Chao1 richness index (number of expected genera in the sample) and its corresponding standard errors

[7] Data indicate ACE richness index (number of expected genera in the sample) and its corresponding standard errors

Supplementary Table 2. Level of funcional assignment of the metagenomic sequences.

ID	Health status	Total reads	cd (n) [a]	cd (%)	cog (n) [b]	cog (%)	Tfam (n) [c]	Tfam (%)	seed (n) [d]	seed(%)
NOCA_01P	□	204218	126729	62	108929	53	82457	40	111497	50
NOCA_03P	□	244881	116575	48	95327	39	74356	30	93391	38
CA1_01P	▒	464594	321997	69	280652	60	214050	46	271868	59
CA1_02P	▒	295072	182091	62	150966	51	118716	40	146161	55
CA_04P	■	339503	192003	57	161384	48	126281	37	158887	48
CA_06P	■	306740	182349	59	151524	49	119477	39	146032	47
CA_05_4.6	▣	70503	40999	58	31864	45	26245	37	29625	42
CA_06_1.6	▣	97722	54305	56	45440	46	35395	36	44552	46

(n): absolute count (%): percentage of the total reads in sample

[a] cdd : conserved domains of NCBI Conserved Domains Database

[b]cog: cluster of orthologous groups

[c]Tfam: Tigr Fams

[d]seed : Seed / MG-RAST sub systems

Supplementary Table 3 | Potential new Operational Taxonomic Units (OTUs) in the oral cavity. Sequences corresponding to 16S rRNA genes were extracted from the 8 metagenomes. BLASTN was performed against the Ribosomal Database Project and the Human Oral Microbiome Database (HOMD) using reads which were over 400 bp and which aligned over 90% of the length. Only OTUs with hits between 80-95% nucleotide identity were considered.

CLOSEST SPECIES	Number of Reads	%ID
uncultured bacterium	71	0.808-0.946
Acetobacter pasteurianus IFO 3283-01-42C	61	0.801-0.886
Streptococcus pyogenes MGAS10750	33	0.812-0.949
Mycoplasma arthritidis 158L3-1	23	0.801-0.864
Lactococcus lactis	20	0.801-0.945
Helicobacter mustelae 12198	18	0.802-0.86
Bacteroides vulgatus	17	0.803-0.898
Dickeya chrysanthemi	17	0.828-0.932
Actinobacillus pleuropneumoniae L20	16	0.808-0.929
Neisseria meningitidis 8013	16	0.808-0.943
Dechloromonas sp. HZ	16	0.812-0.896
Streptococcus sp.	15	0.805-0.949
Prevotella sp.	12	0.883-0.947
Francisella noatunensis subsp. noatunensis	12	0.803-0.915
Bacteroides helcogenes	12	0.832-0.935
Kinetoplastibacterium blastocrithidii	11	0.840-0.939
Veillonella parvula DSM 2008	11	0.815-0.903
Pseudomonas filiscindens	11	0.824-0.883
psychrophilic marine bacterium PS32	10	0.808-0.949
Bacillus cereus	9	0.824-0.896
Prevotella melaninogenica (T)	9	0.841-0.944
Parabacteroides goldsteinii	8	0.804-0.910
Bacteroides massiliensis	8	0.828-0.924
uncultured *Moraxellaceae bacterium*	8	0.834-0.948
Lactobacillus rhamnosus	8	0.811-0.907
Capnocytophaga sp.	7	0.911-0.946
Porphyromonas sp. oral clone EP003	7	0.857-0.923
uncultured *Veillonella sp.*	7	0.801-0.939
Bacillus pumilus	7	0.816-0.909
uncultured *Porphyromonas sp.*	6	0.917-0.929
Haemophilus parasuis SH0165	6	0.808-0.935
uncultured *Tsukamurella sp.*	6	0.810-0.923
Actinomyces naeslundii	6	0.841-0.948
Veillonella parvula	6	0.872-0.948
Fusobacterium nucleatum (T)	6	0.817-0.934
Neisseria sicca	6	0.912-0.949
Streptococcus sanguinis SK36	6	0.816-0.931
Actinomyces sp.	5	0.885-0.939
Bacteroides stercoris ATCC 43183	5	0.833-0.895
Prevotella melaninogenica	5	0.853-0.947
Fusobacterium sp. oral clone ASCF06	5	0.83-0.933
TM7 [G-1] sp.	5	0.839-0.947
uncultured *Actinomyces sp.*	5	0.817-0.923
uncultured *Prevotella sp.*	5	0.830-0.910
Parabacteroides distasonis	4	0.838-0.922
Actinomyces odontolyticus	4	0.844-0.941
uncultured *Corynebacterium sp.*	4	0.933-0.944
Campylobacter gracilis	4	0.803-0.914
uncultured *Neisseria sp.*	4	0.813-0.876
Moraxella catarrhalis	4	0.872-0.917
Burkholderia glathei (T)	4	0.855-0.865
Corynebacterium sp.	4	0.815-0.942
Terrahaemophilus aromaticivorans	4	0.929-0.942
Actinomyces sp. oral strain Hal-1065	3	0.836-0.844
Streptococcus mitis	3	0.816-0.948
Bacteroides tectus	3	0.871-0.921
Selenomonas sp.	3	0.916-0.931
Fusobacterium nucleatum ss. nucleatum	3	0.886-0.949
Aggregatibacter aphrophilus	3	0.863-0.897
Actinomyces sp. oral clone IP073	3	0.818-0.843
Leptotrichia sp. oral clone HE012	3	0.8-0.8897
Parabacteroides merdae	3	0.880-0.918
Bacteroides coprocola	3	0.867-0.886
Moraxella bovoculi (T)	3	0.871-0.934
Bacteroides thetaiotaomicron	3	0.841-0.909
uncultured candidate division TM7	2	0.882-0.889
Haemophilus parainfluenzae	2	0.939-0.946
Eikenella corrodens	2	0.867-0.876
uncultured *Capnocytophaga sp.*	2	0.932-0.944
Bacteroides intestinalis	2	0.805-0.904
Bifidobacterium adolescentis (T)	2	0.830-0.836
Bacteroides salyersiae	2	0.839-0.887
Prevotella denticola	2	0.910-0.948
Rehmannia glutinosa var. purpurea' phytoplasma	2	0.801-0.830
Moraxella sp.	2	0.916-0.933
Acetobacter pasteurianus IFO 3283-07	2	0.805-0.807
Streptococcus anginosus	2	0.876-0.944
uncultured *Abiotrophia sp.*	2	0.849-0.912
Actinomyces israelii	2	0.942-0.944
uncultured *Megasphaera sp.*	2	0.830-0.874
Xanthomonas translucens pv. poae	2	0.822-0.834
Terrahaemophilus sp.	2	0.914-0.915
Bacteroides caccae	2	0.919-0.937
Aggregatibacter sp.	2	0.874-0.934
Prevotella salivae (T)	2	0.830-0.929
Actinomyces sp. oral taxon 180	2	0.830-0.944
Selenomonas sputigena	2	0.817-0.854
Veillonella dispar	2	0.947-0.949
uncultured *Selenomonas sp.*	2	0.803-0.891
Actinomyces oris	2	0.880-0.889
Fusobacterium nucleatum ss. vincentii	2	0.896-0.942
Clostridium difficile 630	2	0.803-0.845
Pectobacterium atrosepticum	2	0.887-0.91
Prevotella denticola (T)	2	0.831-0.947
Selenomonas sp. oral clone GT010	2	0.931-0.939
Kingella oralis (T)	2	0.808-0.934
Parascardovia denticolens	2	0.834-0.910
Neisseria elongata (T)	2	0.807-0.903
Selenomonas infelix (T)	2	0.895-0.897
Rothia dentiocariosa	2	0.895-0.900
Actinomyces sp. oral clone IO076	2	0.907-0.918
Bacteroides uniformis	1	0.81
Capnocytophaga granulosa	1	0.8
Rothia sp. oral taxon 188	1	0.95
Prevotella sp. oral clone GI032	1	0.82
Haemophilus sp.	1	0.95
Actinomyces israelii (T)	1	0.95
uncultured *Gemella sp.*	1	0.81
uncultured *Pseudanabaena sp.*	1	0.9
Gemella sp. oral clone ASCE02	1	0.85
Prevotella tannerae	1	0.92
Bacteroides acidifaciens	1	0.92
Campylobacter concisus	1	0.85
Bergeyella sp.	1	0.94
Lactobacillus paracasei subsp. paracasei	1	0.82
Streptococcus sp. oral clone ASCG04	1	0.86
Streptococcus sp. oral clone ASCA03	1	0.83
Capnocytophaga sp. oral clone BR085	1	0.82
Lachnospiraceae [G-4] sp.	1	0.94
Streptococcus suis BM407	1	0.9
Neisseria gonorrhoeae	1	0.95
Micrococcus luteus NCTC 2665	1	0.95
Olsenella uli	1	0.94
Pseudomonas aeruginosa	1	0.86
Atopobium parvulum DSM 20469	1	0.85
Prevotella sp. oral clone BE073	1	0.8
Capnocytophaga sp. AHN9756	1	0.94
Prevotella intermedia	1	0.91
Leptotrichia sp. oral clone IK040	1	0.93
Aster yellows phytoplasma B	1	0.82
Kytococcus sedentarius	1	0.81
Kordia algicida (T)	1	0.9
Tannerella forsythensis	1	0.92
Rothia sp.	1	0.81
Xanthomonas axonopodis pv. citrumelo	1	0.81
Capnocytophaga sp. AHN9687	1	0.94
Neisseria pharyngis	1	0.92
Streptococcus gordonii	1	0.93
Eubacterium sp. oral clone DO016	1	0.91
Porphyromonas sp. oral clone DP023	1	0.95
Streptococcus sp. Culture clone SRC DSC22	1	0.89
Actinomyces oricola	1	0.93
Chryseobacterium sp. IMMIB L-1519	1	0.85
Xanthomonas axonopodis pv. syngonii	1	0.85
Streptococcus mitis B6	1	0.93
Prevotella oralis	1	0.81
Burkholderia sp. m35b	1	0.94
TM7 [G-3] sp.	1	0.87
Rothia dentocariosa (T)	1	0.92
Bacteroides-like sp. oral clone AU126	1	0.86
Neisseria sp.	1	0.86
Lautropia mirabilis	1	0.89
Fusobacterium nucleatum ss. animalis	1	0.89
Lachnospiraceae bacterium 'Oral Taxon 107'	1	0.9
uncultured *eubacterium*	1	0.92
uncultured candidate division *SR1 bacterium*	1	0.85
Clostridiales bacterium CD3:22	1	0.84
Eubacterium [XIVa] [G-1] saburreum	1	0.94
Treponema pectinovorum	1	0.86
Fusobacterium nucleatum ss. polymorphum	1	0.87
uncultured *gamma proteobacterium*	1	0.81
Bacteroides pyogenes	1	0.89
Prevotella sp. oral clone GU027	1	0.9
Leptotrichia sp.	1	0.8
Bacteroides fragilis	1	0.87
Neisseria flava	1	0.94
Solobacterium moorei	1	0.92
Capnocytophaga gingivalis	1	0.83
uncultured *Haemophilus sp.*	1	0.95
Kingella oralis	1	0.93
Fusobacterium sp. oral clone CZ006	1	0.93
Veillonellaceae bacterium oral taxon 155	1	0.83
Granulicatella sp. oral clone ASCG05	1	0.85
Leptotrichia sp. oral clone DR011	1	0.89

Supplementary Table 4 | Bacterial diversity in cavities from samples CA06-1.6 (a) and CA05-4.6 (b). Data show the number of contigs >500 bp giving a significant hit in the NR database (score >100).

a

BLAST hit against NRdb	Class	Contigs >500 bp
Veillonella parvula	Clostridia	166
Streptococcus pneumoniae	Bacilli	59
Streptococcus mitis	Bacilli	35
Capnocytophaga ochracea	Flavobacteria	20
Prevotella ruminicola	Bacteroidia	15
Porphyromonas gingivalis	Bacteroidia	10
Streptococcus gordonii	Bacilli	9
Corynebacterium aurimucosum	Actinobacteria	7
Corynebacterium efficiens	Actinobacteria	7
Fusobacterium nucleatum	Fusobacterium	7
Bacteroides fragilis	Bacteroides	6
Alistipes shahii	Bacteroides	5
Corynebacterium diphtheriae	Actinobacteria	5
Leptotrichia buccalis	Fusobacteria	5
Xylanimonas cellulosilytica	Actinobacteria	5
Aggregatibacter aphrophilus	γ-proteobacteria	4
Bacteroides thetaiotaomicron	Bacteroides	4
Corynebacterium glutamicum	Actinobacterium	4
Slackia heliotrinireducens	Actinobacteria	4
Bacteroides fragilis	Bacteroides	3
Campylobacter concisus	ε-proteobacteria	3
Corynebacterium kroppenstedtii	Actinobacteria	3
Streptococcus oralis	Bacilli	3
Bacteriophage Dp-1	Virus	2
Bacteroides vulgatus	Bacteroides	2
Beutenbergia cavernae	Actinobacteria	2
Campylobacter hominis	ε-proteobacteria	2
Conexibacter woesei	Actinobacteria	2
Corynebacterium jeikeium	Actinobacteria	2
Corynebacterium urealyticum	Actinobacteria	2
Kribbella flavida	Actinobacteria	2
Micrococcus luteus	Actinobacteria	2
Mycobacterium gilvum	Actinobacteria	2
Prevotella intermedia	Bacteroidia	2
Rhodopseudomonas palustris	α-proteobacteria	2

b

BLAST hit against NRdb	Class	Contigs >500 bp
Veillonella parvula	Clostridia	135
Xylanimonas cellulosilytica	Actinobacteria	46
Sanguibacter keddieii	Actinobacteria	30
Kineococcus radiotolerans	Actinobacteria	20
Porphyromonas gingivalis	Bacteroidia	19
Beutenbergia cavernae	Actinobacteria	17
Treponema denticola	Spirochaetes	17
Brachybacterium faecium	Actinobacteria	16
Kytococcus sedentarius	Actinobacteria	14
Actinomyces naeslundii	Actinobacteria	11
Kocuria rhizophila	Actinobacteria	10
Catenulispora acidiphila	Actinobacteria	9
Geodermatophilus obscurus	Actinobacteria	9
Micrococcus luteus	Actinobacteria	9
Nocardioides sp.	Actinobacteria	9
Thermomonospora curvata	Actinobacteria	9
Rothia mucilaginosa	Actinobacteria	8
Neisseria meningitidis	β-proteobacteria	7
Streptomyces coelicolor	Actinobacteria	6
Actinosynnema mirum	Actinobacteria	5
Arthrobacter chlorophenolicus	Actinobacteria	5
Clavibacter michiganensis	Actinobacteria	5
Corynebacterium efficiens	Actinobacteria	5
Nakamurella multipartita	Actinobacteria	5
Streptococcus sanguinis	Bacilli	5
Streptomyces avermitilis	Actinobacteria	5
Streptomyces griseus	Actinobacteria	5
Bifidobacterium longum	Actinobacteria	4
Nocardia farcinica	Actinobacteria	4
Propionibacterium acnes	Actinobacteria	4
Actinomyces oris	Actinobacteria	3
Bifidobacterium adolescentis	Actinobacteria	3
Gordonia bronchialis	Actinobacteria	3
Saccharopolyspora erythraea	Actinobacteria	3
Streptococcus gordonii	Bacilli	3
Streptomyces scabiei	Actinobacteria	3
Alistipes shahii	Bacteroides	2
Arthrobacter sp.	Actinobacteria	2
Corynebacterium aurimucosum	Actinobacteria	2
Corynebacterium diphtheriae	Actinobacteria	2
Corynebacterium urealyticum	Actinobacteria	2
Eggerthella lenta	Actinobacteria	2
Leifsonia xyli	Actinobacteria	2
Leptotrichia buccalis	Fusobacteria	2
Mycobacterium avium	Actinobacteria	2
Rhodococcus opacus	Actinobacteria	2
Stackebrandtia nassauensis	Actinobacteria	2
Streptosporangium roseum	Actinobacteria	2
Thermobifida fusca	Actinobacteria	2

3.2

“Identifying a healthy oral microbiome through metagenomics”

LD Alcaraz, P Belda-Ferre, R Cabrera-Rubio, H Romero, A Simón-Soro, M Pignatelli, A Mira

Clin Microbiol Infect 2012; 18 (Suppl. 4): 54–57

ORIGINAL ARTICLE 10.1111/j.1469-0691.2012.03857.x

Identifying a healthy oral microbiome through metagenomics

L. D. Alcaraz[1], P. Belda-Ferre[1], R. Cabrera-Rubio[1], H. Romero[2], Á. Simón-Soro[1], M. Pignatelli[1] *and* A. Mira[1]

1) *Department of Genomics and Health, Center for Advanced Research in Public Health, Avda. Cataluña, Valencia, Spain* and 2) *Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias/CURE, Universidad de la República, Montevideo, Uruguay*

Abstract

We present the results of an exploratory study of the bacterial communities from the human oral cavity showing the advantages of pyrosequencing complex samples. Over 1.6 million reads from the metagenomes of eight dental plaque samples were taxonomically assigned through a binning procedure. We performed clustering analysis to discern if there were associations between non-caries and caries conditions in the community composition. Our results show a given bacterial consortium associated with cariogenic and non-cariogenic conditions, in agreement with the existence of a healthy oral microbiome and giving support to the idea of dental caries being a polymicrobial disease. The data are coherent with those previously reported in the literature by 16S rRNA amplification, thus giving the chance to link gene functions with taxonomy in further studies involving larger sample numbers.

Keywords: Binning, dental caries, metagenomics, probiotic, pyrosequencing
Original Submission: 23 November 2011; **Accepted:** 20 March 2012
Editors: A. Moya, Rafael Cantón, and D. Raoult
Clin Microbiol Infect 2012; **18** (Suppl. 4): 54–57

Corresponding author: A. Mira, Department of Genomics and Health, Center for Advanced Research in Public Health, Avda. Cataluña 21, 46020 Valencia, Spain
E-mail: mira_ale@gva.es

Introduction

Unlike most infectious diseases where a single causing agent can be found responsible for the infection, oral diseases appear to be the outcome of multiple microorganisms. In periodontitis, for instance, at least three bacterial organisms have been found to be directly associated with the development of the disease [1]. Similarly, the complexity of the microbial community in the oral cavity has hampered the identification of a single aetiological agent for dental caries. It has been demonstrated that *Streptococcus sobrinus* and above all *S. mutans* are acidogenic and play an important role in caries initiation [2]. However, the use of molecular techniques like PCR amplification and cloning of the 16S rRNA gene have revealed that a high proportion of samples from cavities do not contain *mutans* streptococci, whereas other acid-producing bacteria are present [3]. These include *Lactobacillus*, *Actinomyces* or *Bifidobacterium*. Recent molecular work has confirmed these results and expanded the list of potential cariogenic species to *Veillonella*, *Propionibacterium* and *Atopobium*, among others [4], most of which are poorly characterized species.

Dental caries, microbiome and pyrosequencing

Dental caries is probably better understood as a polymicrobial disease [5] where the interaction and synergistic effect of multiple species should be taken into account for future strategies of diagnosis, prevention and treatment. Given that a large portion of oral bacteria cannot be cultured by current laboratory techniques, the introduction of molecular approaches has provided a significant improvement in our understanding of oral microbiota. However, PCR amplification and cloning still have significant biases that do not allow microbial diversity to be fully studied, as many species or DNA segments cannot be detected. Thus, a metagenomic approach by which the total DNA from a microbial community is obtained obviating the need for culture or PCR amplification has been proposed as a promising strategy to study the full genetic pool of the human microbiome in health and disease [6]. In addition, the extraordinary increase in sequencing output and the reduction of the associated cost provided by next generation sequencing has been applied to the study of gut microbiota, providing a more complete picture of human-associated bacterial communities [7].

We used a 454 GLX Titanium pyrosequencing approach to obtain over 800 Mbp of DNA sequence from supragingival dental plaque samples from eight individuals who varied in oral health status. Two of them (healthy controls, with no caries) were volunteers who had never suffered from dental caries in their lives and another four samples were from individuals with one, four, eight and 15 cavities at the moment of sampling. In addition, two samples were taken from individual cavities in order to give a first glimpse of the diversity at these diseased sites. Over 2 million pyrosequencing reads of 425 bp average length were analysed by phymmBL [8], a binning method that combines the assignment of sequences by homology and by nucleotide composition using hidden Markov models, thus allowing taxonomic binning and prediction for each single read. All the available complete whole genome sequencing as well as reference genomes for the Human Microbiome Project and the Human Oral Database were used to build a local database to predict taxonomic affiliation. Filtering the reads under 200 bp, we managed to taxonomically assign over 1.6 million reads to the 1150 genomes analysed, with an estimated accuracy at the class level over 75% [8].

When a correspondence analysis was performed with the assigned reads, samples with bad oral health tended to cluster together (Fig. 1). As can be observed in the figure, the principal component separated the two healthy samples and the sample from the individual with a single cavity from the other five samples with dental plaque of individuals with more than four cavities and from the two samples within cavities. Whereas the dental plaque samples from individuals of bad oral health clustered tightly at the positive values of the main axis, the three samples from healthy individuals occupied different positions at the secondary axis. Taken together, the results show hints of a specific microbiota associated with the presence of dental caries, and there appear to be several combinations of bacteria under good oral health, a finding that should be confirmed with larger

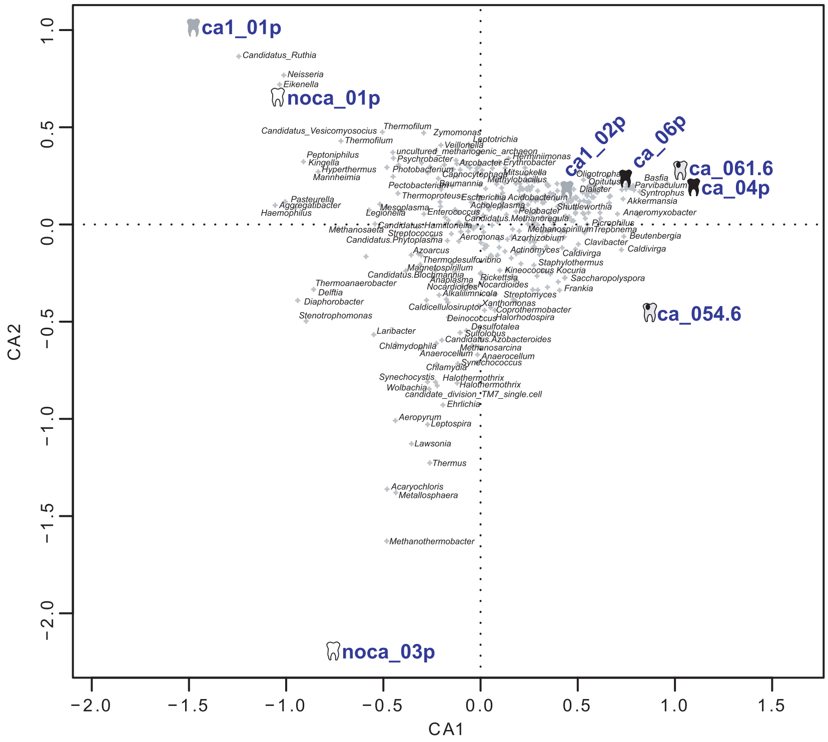


FIG. 1. Correspondence analysis of the bacterial diversity in eight oral samples based on the taxonomic assignment of 1.6 million pyrosequencing reads by the binning PhymmBL approach. The first axis successfully separates healthy from diseased individuals. Around the healthy samples some bacterial genera are suggested to be potentially associated with absence of caries. The samples are represented with symbols according to health status: individuals that have never suffered from dental caries are marked with white teeth symbols (samples noca-01p and noca-03p); individuals with one cavity (sample ca1-01p) and four cavities (sample ca1-02p) are marked with grey teeth symbols; individuals with eight and 15 cavities (samples ca-06p and ca-04p, respectively) are marked with black teeth symbols; samples from individual cavities are marked with a black spot within a white tooth and correspond to teeth 1.6 (sample ca-06_1.6) and 4.6 (sample ca-05_4.6), following WHO nomenclature.

sampling of healthy volunteers. The bacterial genera which are uniquely associated with the absence of caries can be observed in the figure, and include strains with highest similarity to *Neisseria*, *Cardiobacterium*, *Rothia*, *Kingella*, *Aggregatibacter* or *Mannheimia*. Some of these bacteria are poorly known, and include for instance members of the TM7 phylum, a bacterial group which does not have a single member cultured in the laboratory but which appears to be widely present in the oral cavity [9]. Thus, we propose that efforts towards improving culturing media for oral bacteria would be highly recommended to better understand the potential beneficial role of these microbes. In the caries-associated genera we can find *Dialister*, *Oligotropha*, *Basfia*, *Parvibaculum*, *Syntrophus* or *Treponema*, among others. Interestingly the genus *Streptococcus*, which includes the *mutans* streptococci traditionally associated with caries, is not associated to a health status in the correspondence analysis, probably reflecting the preventive nature of some other species from this genus and giving new insights into the contribution of other species to the diseased status.

Our results agree with those obtained by sequencing of PCR amplified 16S rRNA genes, where a number of studies have found a specific set of bacterial genera associated with non-caries oral conditions [10], including some of the health-related bacteria identified in our work. The finding that a given microbial community may be linked to non-caries status supports the idea of using health-associated bacteria as probiotics to prevent oral diseases [11]. The use of health-promoting bacteria has been successfully applied in pharynx infections by inoculating with bacteriocin-producing commensal strains isolated from healthy individuals and is the basis for replacement therapies to prevent infectious disease in the gut and the oral cavity.

Conclusions

The data presented here suggest that a more holistic view of the probiotic approach against dental caries may be needed, as the microbial contribution to good or bad oral health is probably related to bacterial consortia rather than to individual species. It is well established that dental caries is a multifactorial disease where diet, teeth shape and composition, saliva pH or the immune system among others play a role in the tendency to develop the disease. The development of metagenomics and next generation sequencing techniques now allows the contribution of different microbial consortia to oral diseases to be investigated. This is a challenging period in which experimental work should be designed to determine whether a combination of microbial species could be successfully transplanted to the oral cavity and form a stable biofilm that prevents cavities and contributes to oral health. Additionally, the microbial consortia associated with cariogenic conditions may also have important consequences for designing preventive strategies against dental caries. Promising results in passive and active immunization against antigens from specific oral pathogens have been obtained in the last decade [12]. However, if oral diseases are polymicrobial, single-species-based immunization strategies may be limited. We have previously proposed the use of surface antigens shared among several oral pathogens as more efficient targets for the design of vaccines against oral diseases [13]. Metagenomic approaches like the one presented here will help to determine the species against which these efforts should be directed.

Acknowledgements

This work was funded by the following projects from the Spanish MICINN: SAF2009-13032-C02-02 from the I+D program, BIO2008-03419-E from the EXPLORA program and MICROGEN CSD2009-00006 from the Consolider-Ingenio program, and a CONACyT fellowship to LDA.

Transparency Declaration

The authors do not have any potential conflicts of interest to declare.

References

1. Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL. Microbial complexes in subgingival plaque. *J Clin Periodontol* 1998; 25: 134–144.
2. Loesche WJ. Role of Streptococcus mutans in human dental decay. *Microbiol Rev* 1986; 50: 353–380.
3. Corby PM, Lyons-Weiler J, Bretz WA et al. Microbial risk indicators of early childhood caries. *J Clin Microbiol* 2005; 43: 5753–5759.
4. Aas JA, Griffen AL, Dardis SR et al. Bacteria of dental caries in primary and permanent teeth in children and young adults. *J Clin Microbiol* 2008; 46: 1407–1417.
5. Kleinberg I. A mixed-bacteria ecological approach to understanding the role of the oral bacteria in dental caries causation: an alternative to *Streptococcus mutans* and the specific-plaque hypothesis. *Crit Rev Oral Biol Med* 2002; 13: 108–125.
6. Mullany P, Hunter S, Allan E. Metagenomics of dental biofilms. *Adv Appl Microbiol* 2008; 64: 125–136.
7. Qin J, Li R, Raes J et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010; 464: 59–65.
8. Brady A, Salzberg SL. Phymm , Phymm BL. metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 2009; 6: 673–676.

9. Paster BJ, Boches SK, Galvin JL et al. Bacterial diversity in human subgingival plaque. *J Bacteriol* 2001; 183: 3770–3783.
10. Zaura E, Keijser BJF, Huse SM, Crielaard W. Defining the healthy 'core microbiome' of oral microbial communities. *BMC Microbiol* 2009; 9: 259.
11. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R et al. The oral metagenome in health and disease. *ISME J* 2012; 6: 46–56.
12. Abiko Y. Passive immunization against dental caries and periodontal disease: development of recombinant and human monoclonal antibodies. *Crit Rev Oral Biol Med* 2000; 11: 140–158.
13. Mira A. Horizontal gene transfer in oral bacteria. In: Rogers AH, ed. *Oral molecular microbiology*. Caister Academic Press: Norfolk, UK, 2007; 65–85.

3.3

“Mining Virulence Genes Using Metagenomics”

P Belda-Ferre, R Cabrera-Rubio, A Moya, A Mira

PloS one, 2011, 6,10,e24975

Mining Virulence Genes Using Metagenomics

Pedro Belda-Ferre[1], **Raúl Cabrera-Rubio**[1], **Andrés Moya**[1,2], **Alex Mira**[1]*

[1] Joint Unit of Research in Genomics and Health, Centre for Public Health Research-Cavanilles Institute for Biodiversity and Evolutionary Biology, University of Valencia, Valencia, Spain, [2] Centro de Investigación Biomédica en Red especializado en Epidemiología y Salud Pública, Madrid, Spain

Abstract

When a bacterial genome is compared to the metagenome of an environment it inhabits, most genes recruit at high sequence identity. In free-living bacteria (for instance marine bacteria compared against the ocean metagenome) certain genomic regions are totally absent in recruitment plots, representing therefore genes unique to individual bacterial isolates. We show that these Metagenomic Islands (MIs) are also visible in bacteria living in human hosts when their genomes are compared to sequences from the human microbiome, despite the compartmentalized structure of human-related environments such as the gut. From an applied point of view, MIs of human pathogens (e.g. those identified in enterohaemorragic *Escherichia coli* against the gut metagenome or in pathogenic *Neisseria meningitidis* against the oral metagenome) include virulence genes that appear to be absent in related strains or species present in the microbiome of healthy individuals. We propose that this strategy (i.e. recruitment analysis of pathogenic bacteria against the metagenome of healthy subjects) can be used to detect pathogenicity regions in species where the genes involved in virulence are poorly characterized. Using this approach, we detect well-known pathogenicity islands and identify new potential virulence genes in several human pathogens.

Citation: Belda-Ferre P, Cabrera-Rubio R, Moya A, Mira A (2011) Mining Virulence Genes Using Metagenomics. PLoS ONE 6(10): e24975. doi:10.1371/journal.pone.0024975

Editor: Jack Anthony Gilbert, Argonne National Laboratory, United States of America

Received March 16, 2011; **Accepted** August 25, 2011; **Published** October 19, 2011

Funding: This work was funded by the following projects from the Spanish Ministerio De Ciencia E Innovacion: SAF2009-13032-C02-01 and SAF2009-13032-C02-02 from the I+D program, BIO2008-03419-E from the EXPLORA program and MICROGEN CSD2009-00006 from the Consolider-Ingenio program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mira_ale@gva.es

Introduction

Identifying virulence genes experimentally is one of the cornerstones of bacterial pathogenesis research. Experimental approaches typically include cloning of genes potentially involved in pathogenesis into a laboratory strain, transposon mutagenesis to generate a collection of mutants, or detection of genes essential for survival in the host by *in vivo* expression technology [1]. The completion of bacterial genomes now allows to directly detecting genes that could be involved in pathogenicity: when both pathogenic and commensal strains of the same species are sequenced, the genes unique to the pathogen can easily be located [2]. The selection of potential candidate genes is more refined as the number of non-pathogenic strains for comparison increases. Thus, an ideal comparison would be provided by a pathogenic strain and a whole population of related, avirulent strains inhabiting the human body of a healthy individual. The advent of metagenomics and its application to the study of the human microbiome [3,4] now provides a unique opportunity to perform these comparisons, as the total gene pool from whole microbial populations can be compared against the genome of individual pathogenic strains.

A fast way to make these comparisons is achieved by metagenomic recruitments [5]. Individual metagenomic reads that give a hit over a certain identity threshold against a reference bacterial genome are "recruited" to plot a graph which will vary in density depending on the abundance of that organism in the sample. Interestingly, it has frequently been found that recruitments of marine bacteria against all marine metagenomes available identified several "islands" of extremely limited or absent coverage, even for species which were dominant in the sample [6]. These "Metagenomic Islands" have also been found in other free-living environments [7] and represent segments of the genome which are highly variable or specific to the reference strain. Assuming that virulent strains are absent from healthy individuals, metagenomic recruitments of pathogenic strains of bacteria whose commensal counterparts are typically found in the human microbiome should reveal MIs at the regions where virulence genes are located. To test this possibility we have compared the genomes of several human pathogens against available gut metagenomes and against several oral metagenomes obtained by ourselves and other groups through direct pyrosequencing from oral cavity samples.

Results

Human-associated bacteria display Metagenomic Islands (MI)

Similarly to free-living bacteria, when metagenomic recruitments are made between the genomes of gut-associated bacteria against the human gut metagenome, regions with low or absent recruitment are clearly visible (Figure 1A). This shows that gut inhabitants also have genomic regions that appear to be unique to individual strains. In free-living habitats like aquatic environments, intraspecific genomic diversification has been proposed as a strategy to exploit different microniches [8], and this would partly account for differences in gene content among strains from the same species. In the gut and other host-related environments, the

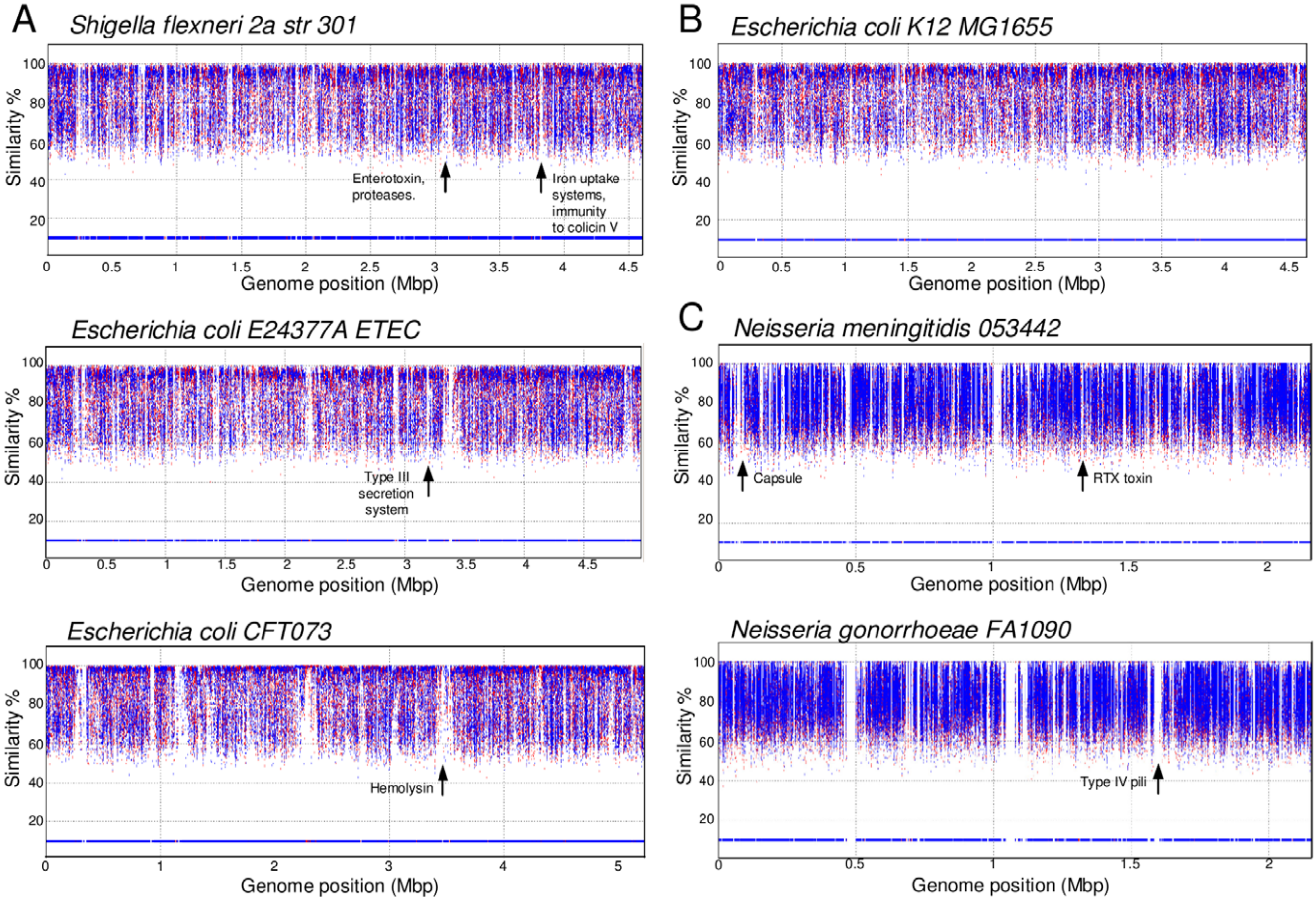


Figure 1. Comparing healthy microbiomes against bacterial genomes. Metagenomic recruitment of (A) the gut metagenome against the three enteric human pathogens *S. flexneri. E. coli* ETEC and *E. coli* CFT073; (B) the gut metagenome against the avirulent *E. coli* K12 laboratory strain; and (C) the dental plaque metagenome against twoe pathogenic neisserial species. Some relevant pathogenicity islands are indicated (for a full list of MIs gene content see Table S1). The few islands detected in the commensal *E. coli* K12 strain correspond to mobile genetic elements, mainly phage genes, as well as a few outer membrane genes.
doi:10.1371/journal.pone.0024975.g001

confinement of bacteria to a given host individual, with limited microbial exchange through the faecal-oral route, imposes a compartmentalized structure to these niches, which may contribute to the observed differences in genomic content between strains. However, the appearance of MIs could also be showing genes specific to virulent strains because recruitment plots for commensal bacteria displayed a higher coverage along the genome and a limited presence of MIs (Fig. 1B), which were limited to mobile genetic elements, mainly phage genes (48.8% of the total) and several outer membrane proteins (14.6% of the total). Thus, in order to determine whether regions of absent recruitment identify genes involved in pathogenicity, we performed a systematic description of gene content in MIs from human pathogens for which pathogenicity islands and virulence genes are well characterized.

MIs identify virulence genes

As shown in Figure 1A, recruitments of human pathogens against the gut metagenome of healthy individuals show MIs which correspond to virulence genes. The gene content of all MIs identified in pathogenic *Shigella* and *Escherichia* strains against the gut metagenome (Table S1) also reveals, as expected, the presence of mobile elements like IS elements and phage genes. In fact, prophages appear to be quite unique to individual strains and represent an important portion of MIs. This may reflect viral infection specificity for individual strains and also high divergence rates for genes which are among the fastest evolving in microbial genomes [6]. But apart from mobile elements, a large proportion of MIs was formed by genes shown experimentally to be involved in pathogenesis and other well-known virulence factors. These include fimbrial proteins, toxins, type I, II and III secretion systems, cell invasion proteins and various antigens, among others (Table S1).

A similar pattern was found when pathogenic Neisserial and Streptococcal species were compared against the oral metagenome of healthy individuals (Figure 1C). The oral microbiome is known to be rich in commensal Neisserial and Streptococcal species [9] and therefore the MIs, apart from containing mobile genetic elements, included many genes involved in pathogenesis such as well-characterized toxins, antigens, hemolysins and adhesins (Table S1). In addition to experimentally demonstrated virulence factors, the islands include many ORFs of unknown function, some of which could also be involved in pathogenesis and should therefore be characterized. An example is given by a 5.7 Kb MI in *Streptococcus pneumoniae* R6, where only hypothetical proteins are annotated (Table S1). Refined sequence similarity searches show that the second half of the island contains genes with homology to the *fmt*A protein family, which modulates antibiotic resistance in

Staphylococcus aureus [13] and adds to other antibiotic resistance genes found in other islands.

When known pathogenicity islands from seven well characterized pathogens were compared to the MIs identified in the present study, most virulence genes were detected (Table 1). Some of the genes which have been shown to play a role in virulence are not detected in MIs because they are involved in several vital cellular functions other than pathogenicity and therefore they are also found in non-virulent strains. These include iron uptake systems or genes involved in adherence to the host. However, most genes directly participating in virulence like toxins, immune evasion systems or proteins involved in cell invasion were readily identified.

Given that many virulence genes are coded in extrachromosomal elements, the same approach was followed for well-characterized bacterial plasmids of enteric bacteria. Despite the promiscuous nature of many extrachromosomal replicons, most plasmids genes from pathogenic strains of *E. coli* showed an intense coverage (Figure 2), showing that these are frequent among natural populations of commensal enteric bacteria. However, clear islands were also identified. Examination of gene content in plasmids' MIs indicated that virulence genes were again absent from the recruitments (Table S2), whereas genes involved in replication, conjugation and other basic plasmid functions were well represented in the gut metagenome. Thus, metagenomic recruitments can prove useful to detect virulence plasmids and to determine which regions from an uncharacterized plasmid may be involved in pathogenicity.

Detection of new virulence genes in Streptococci

We have applied the proposed method in two streptococcal species which vary in their degree of study and pathogenicity.

Table 1. Detection of virulence genes in Metagenomic Islands (MIs).

SPECIES	FUNCTION	VIRULENCE GENES	NUMBER OF GENES IN MI
Neisseria meningitidis FAM18	Adherence	27	13
	Immune evasion	11	11
	Invasion	6	5
	Iron uptake systems	14	7
	IgA protease	1	1
	Toxin	2	2
Neisseria gonorrhoeae FA 1090	Adherence	24	12
	Immune evasion	0	
	Invasion	13	12
	Iron uptake systems	12	6
	IgA protease	1	1
	Toxin	1	0
Shigella flexneri 2a str. 301 chromosome	Host immune evasion	3	3
	Iron uptake systems	20	6
	Protease	2	1
	Secretion system	7	7
	Toxin	2	2
Shigella flexneri 2a str. 301 plasmid	Protease	2	1
	Secretion system	52	51
	Others	4	2
Escherichia coli CFT073	Adherence	45	8
	Autotransporter	4	2
	Iron uptake systems	33	7
	Toxins	4	2
Escherichia coli O157:H7 str. Sakai	Adherence	17	2
	Autotransporter	1	0
	Iron uptake systems	7	0
	LEE encoded TTSS effectors	6	6
	Non-LEE encoded TTSS effectors	5	5
	Secretion system	34	32
	Toxins	4	4
E. coli O157:H7 str. Sakai plasmid O157	Adherence	1	1
	Autotransporter	1	1
	Toxin	4	3

Experimentally characterized virulence genes were obtained from the Virulence Factors Database [14].
doi:10.1371/journal.pone.0024975.t001

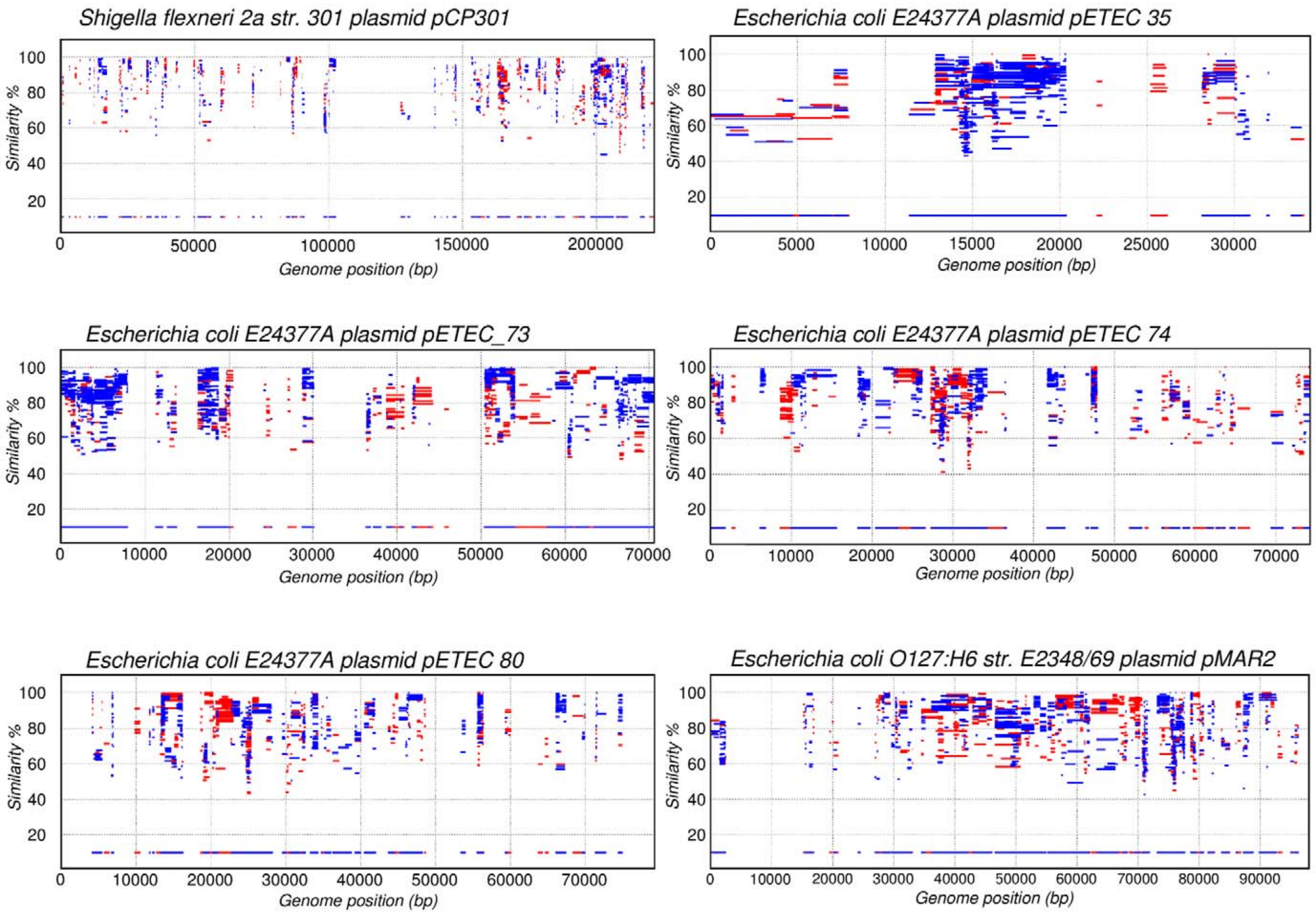


Figure 2. Detecting virulence plasmids by metagenomics. Protein recruitment plots obtained by comparing the healthy gut metagenome against plasmids of pathogenic *Shigella flexneri* and *E. coli* strains. The islands identify known virulence genes whereas genes involved in plasmid housekeeping functions display high recruitment (see Table S2).
doi:10.1371/journal.pone.0024975.g002

S. sanguinis is a normal, commensal inhabitant of the human mouth but after surgeries it can enter the bloodstream and cause endocarditis. When doing recruitment plots of its genome against the oral metagenome of healthy individuals, several MIs were identified (Table S1). Two of them, containing ORFs coding for a platelet binding glycoprotein and adhesion proteins had already been described as virulence factors [10], showing that the method detects known virulence genes. In addition, an 8.9 Kb metagenomic island contains a hemolysin gene which could have a role in red-blood cell lysis and another 6 Kb island includes a gene with high similarity to a precursor of surface antigens (see Table S1). *S. pneumoniae* is a major human pathogen, and different strains are involved in many types of infection ranging from pneumonia to otitis media, meningitis, endocarditis or bacteremia. We have compared the genome of strain R6 against the healthy oral metagenome and have also found new genes potentially involved in virulence (Table S1). These include several *lic* genes, which have been shown to be involved in adherence and nasopharyngeal colonization in animal models [11], and an exoribonuclease from the VacB family, which has been shown to be involved in virulence in enteric bacteria [12]. Genes with significant homology to an immunoglobulin protease, a type IV prepilin peptidase and several cell wall anchor proteins were also within different islands. In several occasions, the islands were located at positions where no genes had been annotated (Table S1). BlastX searches within these regions, however indicated homology to several proteins, like ORFs with repetitive domains and significant homology to hydrolases, fatty acid metabolism proteins and LPXTG-motif cell wall anchor domains (position 556,709 in the R6 genome), suggesting that virulence factors could be present in these islands where functional genes could have passed unnoticed in annotation procedures. Thus, we propose the above genes as potentially involved in pathogenicity and suggest that similar analyses can be helpful in other human or animal pathogens.

Conclusion

Although virulence is a complex trait determined by a large amount of genes and subject to intricate regulation (4% of *Salmonella typhimurium* has been shown to be involved in pathogenesis [14]) the simple procedure described here readily pinpoints most virulence genes unique to pathogenic strains. The procedure can be applied to other non-human pathogens. For instance, the genomes of fish pathogens can be compared against the metagenomes of marine samples where pathogenic strains of the same species are expected to be at low frequencies, and the healthy rumen metagenome could be used for comparison against pathogens of domestic farm animals. The presence of close relatives of plant pathogens in soil may also proof useful to determine genes involved in plant infection by comparing their genomes against the soil metagenome. As shown here, the recruitment of genomes from pathogens against the metagenome of healthy individuals containing commensal strains of the same species may prove extremely useful to select genes potentially involved in virulence and can be specially fruitful in species for which genes involved in pathogenicity are poorly characterized. The method has the advantage of using already-available sequenced metagenomes, whose number is rapidly increasing with the advent of more efficient and inexpensive sequencing techniques. In the future, this approach can be taken one step further when the metagenomes of diseased individuals are also available: In those cases, it will be possible to check whether the potential virulence determinants which are absent in the metagenomes of healthy individuals appear to be present in individuals with diseases such as Crohn's or ulcerative colitis. Obviously, once the approach proposed in this manuscript has narrowed down the list of potential candidates, such pathogenic capability must be tested experimentally. We anticipate that the use of MIs may reduce the number of genes to be cloned or mutated, therefore facilitating the basic process of characterizing virulence factors.

Methods

Oral DNA samples

A 22 year old caucasian healthy female participated in the study after signing informed consent. Sampling procedure was approved by the Ethical Committee for Clinical Research from the Center for Public Health Research (CEIC-DGSP/CSISP). The subject had not received any antibiotic treatment in the previous 2 months, had never suffered from caries and had no symptoms of gingivitis or gum bleeding at the sampling time. The volunteer was asked not to brush her teeth 24 hours before sample collection and not to eat in the prior 2 hours. Supragingival dental plaque was taken from all surfaces of all teeth with sterile toothpicks and pooled into a single sample. DNA was extracted using the AquaPure DNA extraction kit (BIORAD) following the manufacturer instructions and stored at −20°C. DNA concentration was measured with NanoDrop (Thermo Scientific), giving a 497.17 ng/µl concentration and a 260/280 ratio of 1.81 before pyrosequencing.

Gut metagenomes

The sequences of the gut metagenome used were retrieved from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genbank/wgs/), and were composed of 15 healthy individuals, two subjects from Gill et al. (2006) [3], and 13 subjects from Kurokawa et al. (2007) [4], together accounting for a total of 804 Mbp of high-quality reads obtained by shot-gun and subsequent Sanger sequencing.

Oral metagenomes

Oral DNA samples were sequenced using the GS-FLX pyrosequencer with the Titanium chemistry at Macrogen Inc, South Korea. A total of 175,401 reads were obtained, with an average length of 448 bp, adding to a total of 78.6 Mbp. Artificially replicated sequences that systematically appear in 454 data [15] were removed from the dataset using the "454 Replicate Filter" (http://microbiomes.msu.edu/replicates/). For that purpose, sequences that clustered together by CD-HIT and had their first 10 positions exactly identical, were replaced by the longest sequence in each cluster. Those spurious replicates accounted for 4.31% from the total. Additionally, reads corresponding to contaminating human DNA were also removed from the dataset by Megablast [16] against the human genome, with a E-value threshold of 1e-10, giving a final set of filtered sequences of 167,793 reads. These filtered reads have been deposited in the metagenome database from MG-RAST, with Accession Number 4447098.3. In addition to the metagenome we obtained, 113,312 pyrosequences from Xie et al. (2010) [17], accounting for 45.12 Mbp (MG-RAST ID:4446622.3) of sequence from a human dental plaque sample, were also added to the oral dataset, which contained 120.1 Mbp of high-quality sequence.

Reference genomes

The genomes of pathogenic bacteria used in this study were retrieved from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The annotation of those genomes were also downloaded in the. ptt format, in order to search for the function of genes within metagenomic islands. A list of the pathogen reference strains used with their genomes' accession IDs is supplied in Table S3.

Detection of Metagenomic Islands through Recruitment plots

The metagenomic reads were mapped against the sequenced reference genomes using the Nucmer and Promer v3.06 alignment algorithms, with the default parameters [18]. To visualize data, results were plotted using Mummerplot, adding the coverage option, which plots all matches in one dimension, so areas of no recruitment can be readily detected. Using the coordinates files, we considered Metagenomic Islands (MI) as those genomic regions spanning one or more genes which gave no significant hits when mapping against the metagenomic reads at the protein level. For each MI, the genes annotated in that region were identified from the corresponding ptt files.

Virulence genes

Information about virulence factors was obtained from the Virulence Factors Database (http://www.mgc.ac.cn/VFs/main.htm) [10] and from the Pathogenicity Island Database (http://www.gem.re.kr/paidb/) [19]. All virulence genes from the selected genomes were searched for within the detected MIs, thus the proportion of virulence genes contained inside the MIs could be quantified.

Comparisons against different strains, species or genera

An important aspect of metagenome recruitments is to know against which bacterial strains in the metagenome are we comparing the reference genome. It has been shown that the average nucleotide identity (ANI) between orthologous genes of different strains within the same species is on average above 94% [20]. In fact, a threshold of 95% has been proposed as a substitute for the classical DNA-DNA hibridization assays for taxonomical assignments of new species [21]. Values of ANI between 90–95% are typically found for homologous genes in genome comparisons between different species of the same genus [20,22–24]. Other thresholds for higher taxonomic levels are more difficult to establish, although they have been estimated for several two-way comparisons. For instance, the ANI for orthologous genes between *Escherichia* and *Salmonella* appears to be around 80% [25]. Thus, when comparing the reference genomes against the gut and oral metagenomes, a frequency histogram was made with the nucleotide identity obtained for each sequence (Figure S1). When the mode of the histogram was above 94%, the metagenomic recruitment was considered to correspond to strains of the same species. This is the case for the recruitment of *E. coli* O157:H7 against the healthy gut metagenome (Fig. S1A), where this pathogen is probably compared against the gut population of *E.*

coli commensal strains. When *Neisseria meningitidis* is compared against the healthy oral metagenome, the peak in the frequency histogram is located at 91% nucleotide identity (Fig. S1B), indicating that this species is absent in the metagenome and therefore the recruitment is done against other Neisserial species which are common commensal inhabitants from the dental plaque [26]. A third case is provided by the recruitment of the typhi serovar of *Salmonella enterica*, a species which is absent from the healthy human gut, providing a peak in the frequency histogram at 81% nucleotide identity (Fig. S1C). Thus, the plot in this case shows the recruitment of *S. typhi* genes against a different species, primarily against the *E. coli* gut population. Even in the latter case, MIs are readily seen. However, the islands identified will represent not only genes unique to that strain within *Salmonella enterica*, but also genes present in all *Salmonella* species but absent in *E. coli*, making the comparison less specific. The most informative recruitments are therefore those made between a pathogenic strain against the non-pathogenic strains from the same species.

Supporting Information

Figure S1 Nucleotide recruitment plots obtained from comparing the intestinal metagenome against the genome of *Escherichia coli O157:H7 Sakai str.* (*A*) and *Salmonella enterica subsp. enterica serovar Typhi str. CT18* (***C***), and from the oral metagenome against *Neisseria meningitidis FAM18* (***B***). Graphs on the right show frequency histograms of the similarity values of the reads mapped to a given genome. The black line marks the 94% standard threshold for mean identity values for strains from the same species. Taking this line as a threshold, the recruitments are performed against bacteria from the same species as the reference genome (A), different species from the same genus (B) or against a different genus (C).
(TIF)

Table S1 Full gene content of Metagenomic Islands detected by recruitment of selected pathogenic species of pathogenic bacteria against the gut and oral metagenomes.
(PDF)

Table S2 Full gene content of Metagenomic Islands detected by recruitment of selected virulence plasmids from enteric bacteria against the gut metagenome.
(PDF)

Table S3 List of strains analyzed in the manuscript and their corresponding NCBI accession numbers.
(PDF)

Acknowledgments

We thank Prof. F. Rodriguez-Valera for his advice and comments.

Author Contributions

Conceived and designed the experiments: A. Mira PB-F RC-R. Performed the experiments: PB-F RC-R. Analyzed the data: PB-F RC-R A. Mira. Contributed reagents/materials/analysis tools: A. Mira A. Moya. Wrote the paper: A. Mira.

References

1. Tan MW, Rahme LG, Sternberg JA, Tompkins RG, Ausubel FM (1999) *Pseudomonas aeruginosa* killing of *Caenorhabditis elegans* used to identify *P. aeruginosa* virulence factors. Proceedings of the National Academy of Sciences USA 96: 2408–13.
2. Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409: 529–33.
3. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355–9.
4. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Research 14: 169–81.
5. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311: 1768–70.
6. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, et al. (2009) Explaining microbial population genomics through phage predation. Nature Reviews Microbiology 7: 828–36.
7. Pasić L, Rodriguez-Mueller B, Martin-Cuadrado A-B, Mira A, Rohwer F, Rodriguez-Valera F (2009) Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. BMC Genomics 10: 570.
8. Cuadros-Orellana S, Martin-Cuadrado A-B, Legault B, D'Auria G, Zhaxybayeva O, et al. (2007) Genomic plasticity in prokaryotes: the case of the square haloarchaeon. The ISME Journal 1: 235–45.
9. Zaura E, Keijser BJF, Huse SM, Crielaard W (2009) Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiology 9: 259.
10. Yang J, Chen L, Sun L, Yu J, Jin Q (2008) VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Research 36: D539–42.
11. Zhang JR, Idanpaan-Heikkila I, Fischer W, Tuomanen EI (1999) Pneumococcal licD2 gene is involved in phosphorylcholine metabolism. Molecular Microbiology 31: 1477–88.
12. Cheng ZF, Zuo Y, Li Z, Rudd KE, Deutscher MP (1998) The vacB gene required for virulence in *Shigella flexneri* and *Escherichia coli* encodes the exoribonuclease RNase R. Journal of Biological Chemistry 273: 14077–80.
13. Fan X, Liu Y, Smith D, Konermann L, Siu KW, Golemi-Kotra D (2007) Diversity of penicillin-binding proteins. Resistance factor FmtA of *Staphylococcus aureus*. Journal of Biological Chemistry 282: 35143–52.
14. Chaudhuri RR, Peters SE, Pleasance SJ, Northen H, Willers C, et al. (2009) Comprehensive identification of *Salmonella enterica* serovar typhimurium genes required for infection of BALB/c mice. PLoS Pathogens 5: e1000529.
15. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. The ISME Journal 3: 1314–7.
16. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. Journal of Computational Biology 7: 203–14.
17. Xie G, Chain PS, Lo CC, Liu KL, Gans J, et al. (2010) Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. Molecular Oral Microbiology 25: 391–405.
18. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biology 5: R12.
19. Yoon SH, Park Y-K, Lee S, Choi D, Oh TK, et al. (2007) Towards pathogenomics: a web-based resource for pathogenicity islands. Nucleic Acids Research 35: D395–400.
20. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences USA 102: 2567–72.
21. Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences USA 106: 19126–31.
22. Mira A, Ochman H (2002) Gene location and bacterial sequence divergence. Molecular Biology and Evolution 19: 1350–8.
23. Ivars-Martinez E, Martin-Cuadrado A-B, D'Auria G, Mira A, Ferriera S, et al. (2008) Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. The ISME Journal 2: 1194–212.
24. Haley BJ, Grim CJ, Hasan NA, Choi S-Y, Chun J, et al. (2010) Comparative genomic analysis reveals evidence of two novel Vibrio species closely related to *V. cholerae*. BMC Microbiology 10: 154.
25. Goris J, Konstantinidis KT, Klappenbach Ja, Coenye T, Vandamme P, Tiedje JM (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology 57: 81–91.
26. Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, et al. (2010) Bacterial diversity in the oral cavity of 10 healthy individuals. The ISME Journal 4: 962–74.

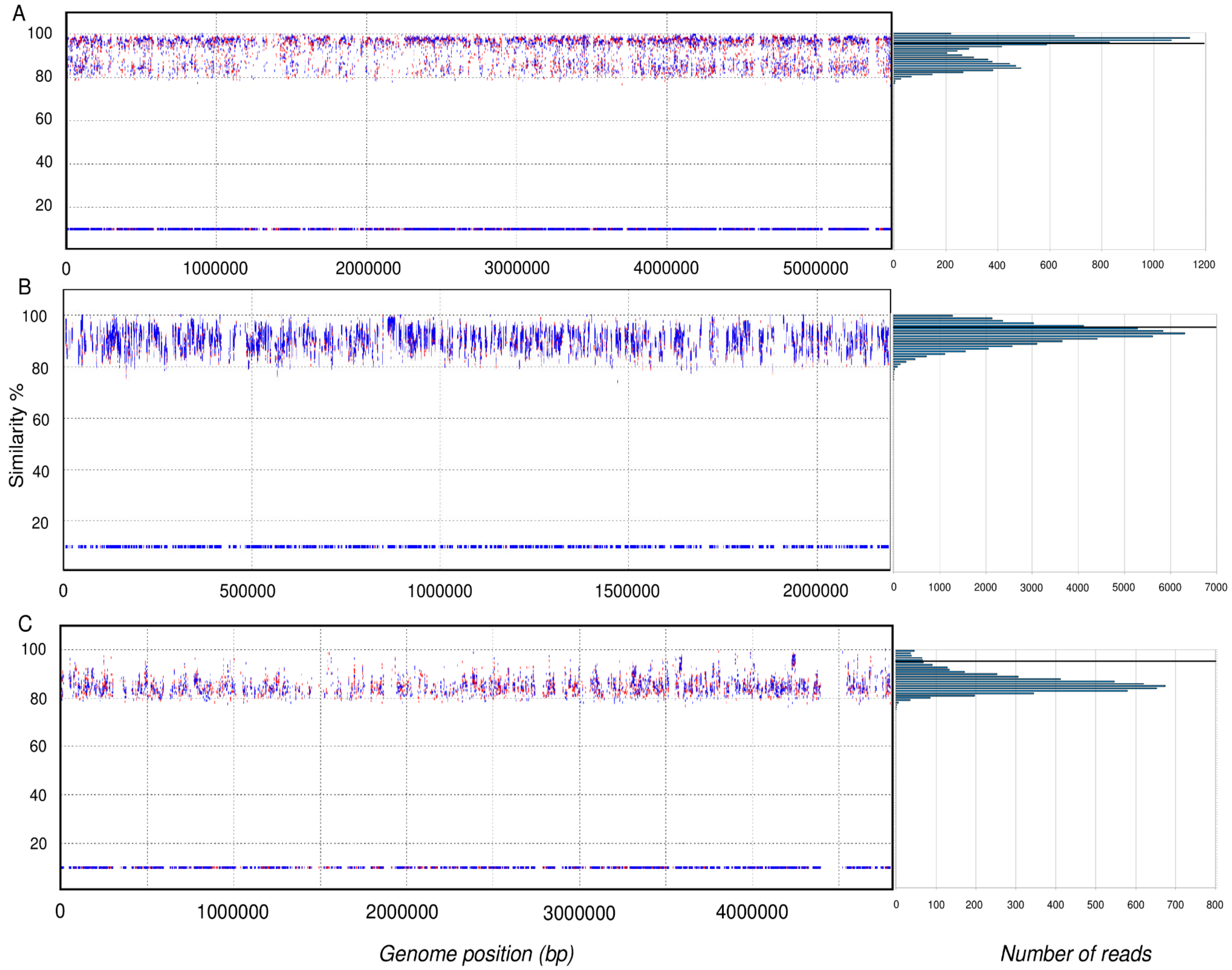


Figure S1. Nucleotide recruitment plots obtained from comparing the intestinal metagenome against the genome of Escherichia coli O157:H7 Sakai str. (A) and Salmonella enterica subsp. enterica serovar Typhi str. CT18 (C), and from the oral metagenome against Neisseria meningitidis FAM18 (B). Graphs on the right show frequency histograms of the similarity values of the reads mapped to a given genome. The black line marks the 94% standard threshold for mean identity values for strains from the same species. Taking this line as a threshold, the recruitments are performed against bacteria from the same species as the reference genome (A), different species from the same genus (B) or against a different genus (C).

Supplementary Table 1.
Full gene content of Metagenomic Islands detected by recruitment of selected pathogenic species of pathogenic bacteria against the gut and oral metagenomes.

Escherichia coli E24377A ETEC vs gut metagenome

Start-End (bp)	Length (kbp)	Number of ORFs	Main features (in brackets number of genes)
19539-22862	3.3	2	hypothetical protein(2)
240344-244042	3.7	3	hypothetical protein (2), ImpA domain-containing protein (1)
295666- 306687	11.02	7	hypothetical protein (5), prophage CP4-57 regulatory protein (1), SNF2 family helicase (1)
315072- 319208	4.13	2	hypothetical protein(1), UvrD family helicase (1)
608338- 614564	6.2	3	hypothetical protein (2), YD repeat-containing protein (1); *
698289- 702722	4.4	6	hypothetical protein (5), DnaJ domain-containing protein (1)
836672- 843690	7.01	6	hypothetical protein (2), multidrug efflux protein, IS66 family transposase, IS66 family orf2, IS66 family orf1 (*)
1000659-1013491	12.8	13	phage integrase family site specific recombinase, phage regulatory protein, putative replication gene B protein, C4-type zinc finger DksA/TraR family protein, replication gene A protein, hypothetical protein (5), PBSX family phage portal protein, hypothetical protein, IS21 family transposase, IS21 family transposition helper protein
1063848- 1083251	19.4	20	tail fiber assembly protein, putative phage tail fiber protein, hypothetical protein (2), baseplate assembly protein J, phage protein, phage baseplate assembly protein V, bacteriophage Mu P protein, bacteriophage Mu transposase MuA, putative repressor protein, aliphatic sulfonate ABC transporter periplasmic substrate-binding protein, NAD(P)H-dependent FMN reductase, fimbrial protein (3), periplasmic pilus chaperone family protein, outer membrane usher protein fimD, putative fimbrial protein, putativi pili assembly chaperone
1124464- 1128783	4.3	4	putative lipoprotein, hypothetical protein, group 4 capsule (G4C) polysaccharide; lipoprotein YmcC, putative inner membrane protein
1399972-1448135	48.7	50	phage N-6-adenine-methyltransferase, hypothetical protein (32), IS66 family orfs (3), putative recombinase, transport protein TonB, acyl-CoA thioester hydrolase, intracellular septation protein A, outer membrane protein W, phage integrase recombinase, exonuclease family protein, Rha family phage regulatory protein, DNA-binding transcriptional regulator DicC, crossover junction endodeoxyribonuclease RusA-like protein, phage antitermination protein Q, protein kinase domain-containing protein, DNA methylase, lambda phage portal protein
1626864- 1635094	8.2	4	hypothetical protein (2), type VI secretion system Vgr family protein, protein rhsD, truncation.
1755818- 1761438	5.6	7	hypothetical protein, IS66 family transposase (2), IS66 family orf2 (2), IS66 family orf1 (2)
2172994-2227046	54.05	91	hypothetical protein (38), tail fiber family protein, major tail sheath protein, phage tail protein I, baseplate assembly protein J, baseplate assembly protein W, phage baseplate assembly protein V, phage major capsid protein E, bacteriophage lambda head decoration protein D, family peptidase, IS21 (2), phage terminase large subunit (GpA), phage antitermination protein Q, lipoprotein (2), DnaB family helicase, transcriptional repressor DicA, phage integrase, invasion, shikimate transporter, AMP nucleosidase, transcriptional regulator Cbl, nitrogen assimilation transcriptional regulator, nicotinate dimethylbenzimidazole phosphoribosyltransferase, cobalamin synthase, adenosylcobinamide kinase/adenosylcobinamide-phosphate guanylyltransferase, cobalamin biosynthesis, regulatory protein PocR, propanediol diffusion facilitator, propanediol utilization protein PduA and B, propanediol dehydratase (18), gyrase inhibitor, D-alanyl-D-alanine carboxypeptidase, exonuclease I, YeeE/YedE, amino acid permease
2283962- 2292498	8.5	7	hypothetical protein (4), VI polysaccharide biosynthesis protein vipB/tviC, VI polysaccharide biosynthesis protein VipA/tviB, glycosyl transferase group 2 family protein (2)
2907653- 2949345	41.69	28	resolvase family site-specific recombinase, parB family protein, repB plasmid partitioning protein, N4/N6-methyltransferase family protein, putative type I restriction-modification system, S subunit, HsdR family type I site-specific deoxyribonuclease, hypothetical protein (14), relaxase/mobilization nuclease domain-containing protein, putative lipoprotein (2), IS21 family transposase (2), IS66 family (2)
3020396- 3032923	12.5	7	hypothetical protein (4), IS3; transposase orfA, IS3; transposase orfB, putative phage integrase family protein
3184041-3202475	18.4	18	hypothetical protein (6), transcriptional regulatory protein (C terminal), TPR repeat-containing protein, transcriptional regulator, LuxR family transcriptional regulator, type III secretion apparatus lipoprotein EprK, type III secretion apparatus protein EprH, FlhB/HrpN/YscU/SpaS family protein, type III secretion apparatus proteins EpaR/Q/O2, surface presentation of antigens protein SpaP, type III secretion apparatus protein (truncation).
3328251-3341175	12.9	8	hypothetical protein (3), SNF2 family helicase, AAA family ATPase, S8A family peptidase, DNA methylase, type III restriction enzyme, res subunit
3373202-3397803	24.6	21	phage integrase family protein, helicase/Zfx/Zfy transcription activation region domain-containing protein, hypothetical protein (8), Ig family protein, UvrD family helicase, IS66 family orf1, IS66 family orf2, IS66 family transposase, IS21 family (2), DnaB family helicase, chromosome partitioning protein, site-specific recombinase, phage integrase family protein.

3406613-3414091	7.4	7	IS66 transposase, IS66 family orf2, IS66 family orf1, DNA-binding protein H-NS-like protein, hypothetical protein, prophage CP4-57 regulatory protein, hypothetical protein
3519564-3521299	1.7	2	periplasmic pilus chaperone family protein, hypothetical protein
3575976- 3581426	5.4	4	putative pilus biogenesis initiator protein, hypothetical protein (2), putative CS1 type fimbrial major subunit.
3624219- 3629648	5.4	4	fimbrial protein, pili assembly chaperone protein, fimbrial usher family protein, putative fimbrial protein.
3942979- 3946457	3.5	2	hypothetical protein (2).
4078565-4083334	4.8	2	hypothetical protein, RHS domain-containing protein
4263383-4268559	5.2		Not annotated
4360765-4365797	5.03		Not annotated
4854251- 4858358	4.1	8	iron-dicitrate transporter ATP-binding subunit, iron-dicitrate transporter subunit FecD, iron-dicitrate transporter permease subunit, iron-dicitrate transporter substrate-binding subunit, fec operon regulator FecR, FecI.
4887773- 4898030	10.25	7	type III restriction enzyme, res subunit, N4/N6-methyltransferase family protein, hypothetical protein (2)

Escherichia coli **CFT073 vs gut metagenome**

Start-End (bp)	**Length (kbp)**	**Number of ORFs**	**Main features (in brackets number of genes)**
274627-279569	4.9	3	hypothetical protein (3)
292801- 298303	5.5	5	putative oligogalacturonide transporter, putative exopolygalacturonate lyase, hypothetical protein (3)
315930-320291	4.3	1	ShlA/HecA/FhaA exofamily protein
331863-347819	15.6	13	hypothetical protein (10), putative cytoplasmic membrane export protein, putative membrane spanning export protein, RTX family exoprotein A gene.
908887-942273	33.4	48	integrase, hypothetical protein (22), putative regulator for prophage, DNA adenine methylase, prophage terminase, ATPase subunit, putative capsid scaffolding protein, major capsid protein terminase, endonuclease subunit, putative capsid completion protein, phage tail protein secretory protein, Fels-2 prophage lysozyme, putative regulatory protein, phage tail protein (4), Phage baseplate assembly protein (3), variable tail fibre protein, major tail sheath protein, putative tail fiber protein of prophage (2), putative regulator of late gene expression, prophage P2 Ogr protein.
1127548-1135398	7.8	6	P4 family integrase, hypothetical protein (4), prophage CP4-57 regulatory protein alpA
1142842-1163920	21.07	24	hypothetical protein (16), 3-oxoacyl-(acyl carrier protein) synthase II, 3-ketoacyl-(acyl-carrier-protein) reductase, hypothetical protein, 3-oxoacyl-(acyl carrier protein) synthase I, acyl carrier protein, putative acyl carrier protein, putative phospholipid biosynthesis acyltransferase, putative O-methyltransferase
1169911- 1173103	3.2	4	hypothetical protein (2), putative transposase, phospho-2-dehydro-3-deoxyheptonate aldolase
1175693-1182738	7.04	5	MchB protein, MchC protein, MchD protein, microcin H47 secretion protein (2)
1186790- 1194173	7.4	10	putative F1C and S fimbrial switch regulatory protein (2), F1C fimbrial subunit precursor (5), F1C periplasmic chaperone, F1C fimbrial usher
1202703- 1207331	4.6	1	ABC transporter ATP-binding protein
1218885- 1224353	5.5	6	hypothetical protein (5), antigen 43 precursor
1350647-1353648	3	4	putative DNA packaging protein of prophage (terminase large subunit), putative DNA packaging protein of prophage, putative capsid protein of prophage (2, one truncated)
1759666-1763220	3.5	1-2	hypothetical protein, (zinc protease pqqL, truncated)
2748603- 2758995	10.2	4	yapH-like protein, hypothetical protein, Type 1 fimbriae regulatory protein fimB (2)
2988732- 2992446	3.7		Not annotated
3234417- 3239912	5.5	4	hypothetical protein (4)
3416671-3420519	3.8	3	hypothetical protein, hemolysin C, hemolysin A (truncated) *Region 3406225-3450866 corresponds to well characterized Pathogenicity Island I (Alpha-hemolysin, P-fimbriae, aerobactin)*
3432331-3436862	4.5	4	PapJ protein, PapD protein, PapC protein, PapH protein.
3452943-3456117	3.2	4	hypothetical protein (4)
3464575-3472925	8.3	9	IucC protein, IucB protein, IucA protein, shiF protein, hypothetical protein (5)
3520088-3527942	7.8	5	hypothetical protein (4), putative glycerol-3-phosphate cytidyltransferase.

Salmonella enterica subsp. enterica serovar Typhi str. CT18 vs gut metagenome

Start-End (bp)	**Length (kbp)**	**Number of ORFs**	**Main features (in brackets number of genes)**
14807-37104	22.3	20	Fimbrial proteins (6), hypothetical protein (9), transcriptional regulator (3), sulfatase (1), chitinase (1)
302188-360345	58.1	51	hypothetical protein (35), fimbrial proteins (7), outer membrane adhesin (1), lipoprotein (2) Rhs-family protein (3), transcriptional regulator (2), ClpB-like protein (1) *Region 302092-360757 established as Pathogenicity Island 6. Function : safA-D and tcsA-R chaperone-usher fimbrial operons*
378000-393000	15	16	Fimbrial proteins (5), hypothetical protein (5), transmembrane regulator (2), transcriptional regulator (1), lipoprotein (1), outer membrane protein (1).
1008756 -1053000	43	65	Putative bacteriophage proteins (33), hypothetical protein (10), putative secreted protein (2),putative DNA-binding protein (3), putative replication protein (1), putative prophage terminase small (1) and large (1) subunit, putative prophage membrane protein (2), putative prophage antitermination protein (1), putative

			methyltransferase (1), putative lipoprotein (1), bacteriophage recombination protein (1), excisionase (1), exonuclease (1), FtsZ inhibitor protein (1), host-nuclease inhibitor protein (1), putative damage-inducible protein (1), DNA invertase (1), DNA methylase (1), integrase (1).
1085173-1094111	8.9	9	Transposase for insertion sequence element is200, hypothetical protein (4), cell invasion protein (2), putative secreted peptidase, histidine kinase. *Region 1085068-1092563 established as PAI5. Function : Effector proteins for SPI-1 and SPI-2 (SopB, SigD, PipB)*
1465324-1486831	21.5	18	hydrogenase 1 maturation protease (1), hydrogenase isoenzyme formation protein (1), hydrogenase-1 operon protein HyaE2 (1), hydrogenase-1 operon protein HyaF2 (1), hypothetical protein (2), membrane transport protein (1), Ni/Fe-hydrogenase 1 b-type cytochrome subunit HyaC2 (1), putative alcohol dehydrogenase (1), putative aminotransferase (1), putative ATP/GTP-binding protein (1), putative isomerase (1), putative multidrug efflux protein (1), putative regulatory protein (2), putative secreted hydrolase (1), putative transport protein (1), uptake hydrogenase small subunit (1).
1625104-1650621	25.5	32	putative outer membrane secretory protein (1), putative pathogenicity island 2 secreted effector protein (7), putative pathogenicity island lipoprotein (1), putative pathogenicity island protein (9), putative secretion system protein (2), two-component response regulator (1), two-component sensor kinase (1), putative type III secretion protein (4), Type III secretion system chaperone protein (1), secretion system apparatus protein SsaU (1), secretion system apparatus protein SsaV (1), type III secretion system ATPase (1), type III secretion system protein (2). *Region 1624920-1666524 established as PAI2. Function : Type III secretion system, required for systemic infection and intracellular pathogenesis. Insertion site : tRNA-val*
1768224-1791992	23.8	33	Hypothetical protein (13), lysozyme inhibitor, outer membrane invasion protein, putative ABC transport ATP-binding (2), putative bacteriophage protein, putative cold shock protein, putative cytochrome, putative heat shock protein, putative inner membrane transport protein (2), putative lipoprotein (3), putative outer membrane virulence protein, substrate-binding transport protein, putative toxin-like protein, putative virulence protein, toxin subunit (2), transposase for IS200.
1818448- 1930625	112.2	131	23S rRNA methyltransferase A, alanine racemase, carboxy-terminal protease, cell division inhibitor MinD, cell division topological specificity factor MinE, cold shock-like protein CspC, D-amino acid dehydrogenase small subunit, exonuclease VIII, fatty acid metabolism regulator, FtsZ inhibitor protein, heat shock protein HtpX, host cell-killing modulation protein, hypothetical protein (58), L-serine deaminase 1, L,D-carboxypeptidase A, long-chain-fatty-acid--CoA ligase, mannose-specific PTS system protein IID, membrane-bound lytic murein transglycosylase E, para-aminobenzoate synthase component I, penicillin-binding protein, phosphotransferase enzyme II C component,
			potassium/proton antiporter, PTS system mannose-specific IIAB component, putative acetyltransferase, putative bacteriophage protein (31), bacteriophage tail protein (2), putative Cro repressor, DNA-binding protein, putative endolysin, putative hydrolase, lipoprotein, putative regulator, replication protein, solute/DNA competence effector, transposase, ribonuclease D, RsmF, septum formation inhibitor, serine/threonine protein phosphatase 1, sodium/proton antiporter, SpoVR family protein, transcriptional regulator KdgR.
2742887-2800000	57.2	42	Large repetitive protein, putative type I secretion protein (2), putative secretion protein ATP-binding protein, 4-aminobutyrate aminotransferase, DNA binding protein nucleoid-associated, DNA-binding transcriptional regulator CsiR, gamma-aminobutyrate transporter, hydroxyglutarate oxidase, hypothetical protein (12), major tail tube protein, outer membrane receptor FepA, putative ABC transporter protein, putative bacteriophage late gene regulator (2), putative bacteriophage major tail sheath protein, putative bacteriophage (2), putative ferric enterochelin esterase, putative transcriptional regulator (2), two-component system sensor kinase, putative type I secretion protein, succinate-semialdehyde dehydrogenase I, transcriptional regulator, virulence protein. *Region 2743495-2759190 established as PAI9. Function : Type I secretory apparatus, including large RTX-like protein*
2862867-2900000	37.1	38	acyl carrier protein, AraC family transcription regulator (4), ATP synthase SpaL, cell adherance/invasion protein (5), chaperone (associated with virulence), hypothetical protein (5), invasion protein regulator, pathogenicity 1 island effector protein (8), secretory protein (associated with virulence) (6), serine/threonine-specific protein phosphatase 2, surface presentation of antigens protein SpaO/SpaP/SpaS(associated with type III secretion and virulence), tyrosine phosphatase (associated with virulence) (4). *Region 2858736-2900586 established as PAI-1. Function : Type III secretion system, invasion into epithelial cells, apoptosis (InvA, OrgA, SptP, SipA, SipB, SipC, SipD, SopE, prgH). Insertion site : fhlA/mutS*
3042054-3059937	17.9	17	endonuclease fragment, fimbrial chaperone protein, fimbrial protein, hypothetical protein (10), integrase, outer membrane fimbrial usher protein, outer membrane protein (associated with virulence), plasmid maintenance.
3133624-3139922	6.3	12	Hypothetical protein (10), bacteriocin immunity protein. *Region 3132530-3139414 established as PAI8. Function : Two bacteriocin pseudogenes, genes conferring immunity to the bacteriocins. Insertion site : tRNA-phe*
3515395-3549044	33.6	47	bacteriophage integrase, capsid portal protein, DNA adenine methylase, DNA-invertase, endonuclease, hypothetical protein (15), lipoprotein (2), major capsid protein, major tail sheath protein, major tail tube protein, phage baseplate assembly protein (2), phage tail protein, putative capsid completion protein, putative capsid

			scaffolding protein, putative lysozyme, phage tail protein (5), putative positive regulator of late gene transcription (2), putative regulatory protein, regulatory protein cII, repressor protein, secretory protein, terminase ATPase subunit, terminase endonuclease subunit, transposase, variable tail fibre protein.
3889742-3897176	7.4	3	Hypothetical protein (2), putative DNA binding protein. Region 3883613-3900553 established as PAI3. Function: Invasion, survival in monocytes, Mg2+ uptake. Insertion site : tRNA-pro.
4321946-4346916	25	9	hypothetical protein (2), large repetitive protein (2), putative integral membrane protein, putative type-1 secretion protein (3), single-stranded DNA-binding protein. *Region 4322993-4346383 established as PAI-4. Function: Type I secretion system, putative toxin secretion, apoptosis, required for intracellular survival in macrophages, genes weakly similar to RTX-like toxins.*
4402991-4543069	140	153	AraC family transcription regulator, bacteriophage integrase, capsid portal protein, DNA adenine methylase, DNA helicase, DNA polymerase V subunit UmuC, DNA topoisomerase III,GerE family regulatory protein, hypothetical protein (86), IS1 (2), integrase (fragment), invasion-associated secreted protein, tail sheath protein, major tail tube protein, nonspecific acid phosphatase precursor, nucleotide-binding protein, phage baseplate assembly (2), phage integrase (2), phage regulatory protein (2), phage protein (3), PilN lipoprotein, pilus assembly protein, prepilin (4), putative acetyltransferase, capsid completion protein, putative capsid protein, DNA helicase, putative exonuclease, putative lipoprotein, lysozyme, major capsid protein, methyltransferase, phage baseplate assembly protein, phage tail protein (7), phage terminase, pilus assembly protein, positive regulator of late gene transcription (2), regulatory protein, secretion protein, shufflon-specific DNA recombinase, single-stranded DNA-binding protein, terminase subunit, transcriptional regulatory protein, VI polysaccharide biosynthesis protein (2), VI polysaccharide protein (10). *Region 4409511-4543148 established as PAI7. Function : Vi exopolysaccharide, SopE prophage and a type IVB pilus operon. Insertion site : tRNA-phe*

Escherichia coli O157:H7 str Sakai vs gut metagenome

Start-End (bp)	**Length (kbp)**	**Number of ORFs**	**Main features (in brackets number of genes)**
18284-24143	5.9	7	hypothetical protein (4), putative outer membrane usher protein precursor, putative fimbrial protein (2)
154305-161555	7.3	7	putative fimbrial related protein (7)
226821-231340	4.5	1	putative phosphatase
311962-315433	3.5	6	hypothetical protein (6)
579634-605089	25.5	5	hypothetical protein (2), putative outer membrane transport protein, putative ATP-binding component of a transport system, putative membrane fusion protein of a transport system
616665-620312	3.6	1	RhsD core protein with extension
659088-663506	4.4	2	bacteriophage N4 adsorption protein Nfr (2)
665897-673709	7.8	4	Rhs core protein (2), hypothetical protein (2)
756972-761392	4.4	5	hypothetical protein (3), putative enzyme of polynucleotide modification, putative tRNA ligase
808998-813939	4.9	2	RhsC core protein with extension, hypothetical protein
824993-827911	2.9	3	hypothetical protein, putative chaperone, putative outer membrane protein
896902-901800	4.9	6	NinG protein, serine/threonin protein phosphatase, putative outer membrane protein, antitermination protein, hypothetical membrane protein, hypothetical protein
925931-929688	3.8	4	Hypothetical protein (4)
1130621-1134072	3.5	5	hypothetical protein (3), homolog of Salmonella FimH protein, putative fimbrial-like protein.
1161155-1165787	4.6	5	putative integrase, hypothetical protein (3), putative division inhibition protein.
1180375-1182095	1.7	1	hypothetical protein
1185506-1189603	4.1	8	hypothetical protein (7), putative holin protein.
1189827-1196688	6.9	7	Phage related protein (6), hypothetical protein.
1266910-1268457	1.5	3	Hypothetical protein, Shiga toxin 2 subunit A and B
1275281-1285769	10.5	11	hypothetical protein (7), phage related protein (4)
1287067-1306278	19.2	17	hypothetical protein (12), putative outer membrane protein (3), putative tail tip fiber protein, MokW protein.
1321735-1324060	2.3	1	hypothetical protein
1337169-1361748	24.6	18	FidL-like protein, hemagglutinin related protein (2),
1365418-1378122	12.7	8	hypothetical protein (4), putative integrase, putative membrane protein, transposase, putative regulatory protein
1403704-1409766	6.1	9	Hypothetical protein (8), TerW protein
1412093-1415215	3.1	4	putative tellurium resistance protein (4)
1420952-1424909	4	6	hypothetical protein (6)
1544512-1549614	5.1	11	hypothetical protein (9), putative phage related protein (2)
1561435-1575771	14.3	15	Putative phage related protein (12), hypothetical prots. (3)
1579575-1585557	6	3	putative secreted effector protein, hypothetical protein (2)
1601708-1608895	7.2	11	Putative phage related protein (7), hypothetical protein (4)
1766113-1769281	3.2	3	putative fimbrial minor pilin protein precursor (2), putative colonization factor.
1774074-1789217	15.1	19	hypothetical protein (8), putative holin protein, putative endolysin, antirepressor protein, endopeptidase, lipoprotein Rz1 precursor, putative Dnase, putative phage related protein (5)

1790629-1794373	3.7	3	Phage related protein (3)
1800640-1803498	2.9	3	hypothetical protein (3)
1803739-1811589	7.9	12	Integrase, hypothetical protein (11)
1922707-1925718	3	5	hypothetical protein (2), restriction alleviation and modification enhancement protein, recombinase recT, exonuclease VIII RecE
1934311-1938964	4.7	4	hypothetical protein (3), putative methyltransferase.
1940951-1943871	2.9	3	Hypothetical protein.
1944553-1956524	12	17	putative endolysin, putative antirepressor protein, putative endopeptidase, hypothetical protein (7), putative Dnase, putative phage related protein (6),
1957922-1962025	4.1	3	Phage related protein (3)
2042654-2050534	7.9	3	hypothetical protein, VgrE protein, RhsE core protein
2095238-2098792	3.6	1	hypothetical protein
2116542-2122543	6	2	putative ATP-binding component of a transport system and adhesin protein, hypothetical protein
2158654-2165353	6.7	8	hypothetical protein (4), phage related protein (4)
2176148-2179103	3	3	Phage related protein (3)
2184405-2190561	6.2	6	putative endolysin, putative holin protein, putative transcriptional regulator, hypothetical protein (3)
2218918-2222988	4.1	1	putative tail length tape measure protein
2224586-2241664	17.1	24	Hypothetical protein (12), endopeptidase, lipoprotein Rz1 precursor, putative antirepressor protein, putative antitermination protein, putative Dnase, putative endolysin, putative head-tail adaptor, putative holin protein, putative major head protein/prohead protease, putative portal protein, putative terminase large subunit, putative terminase small subunit.
2484600-2487761	3.2	2	putative transport protein, hypothetical protein.
2668068-2670857	2.8	3	Hypothetical protein (2), EspF-like protein
2673153-2688564	15.4	17	Host specificity protein, hypothetical protein, phage related protein (15)
2694791-2698923	4.1	5	Hypothetical protein (4), antiterminator.
2714118-2718695	4.6	1	Putative factor
2780835-2787527	6.7	6	putative glycosyl transferase (3), perosamine synthetase, O antigen flippase, O antigen polymerase
2901449-2905616	4.2	5	Putative phage related protein (5)
2924228-2926965	2.7	4	Shiga toxin I precursor (2), antitermination protein, hypothetical protein.
3075378-3079251	3.9	4	putative antibiotic resistance protein, putative transcriptional regulator, hypothetical protein, glycerophosphodiester phosphodiesterase
3478307-3480077	1.8	2	hypothetical protein (2)
3492963-3498985	6	6	hypothetical protein (4), putative site specific recombinase, putative DNA binding protein
3505135-3511008	5.9	4	Hypothetical protein (4)
3709749-3737163	27.4	35	Hypothetical protein (15), putative invasion protein, putative sensory transducer, putative transcriptional regulator, tyoe III secretion system protein (17)
3866707-3872713	6	7	putative adherence factor (2), transposase (2), hypothetical protein (3)
4159436-4164182	4.7	0	Not annotated
4365737-4369236	3.5	3	Hypothetical protein (3)
4585467-4605785	20.3	25	CesT protein, Esc protein, Esp protein (4), gamma intimin, hypothetical protein (14), translocated intimin receptor, type III secretion system protein(3)
4606989-4609848	2.9	4	Type III secretion system protein(3), hypothetical protein
4610632-4614888	4.3	4	type III secretion system protein (4), hypothetical prot. (2)
4615973-4624447	8.5	14	Type III secretion system protein(4), hypothetical protein (8), Ler protein, EspG protein
4686222-4693486	7.3	7	Hypothetical protein (7)
4928827-4931803	3	1	RhsH core protein with extension
4975790-4979751	4	0	Not annotated
5017602-5022018	4.4	0	Not annotated
5028745-5031187	2.4	1	regulator of acetyl CoA synthetase
5057479-5075494	18	22	Hypothetical protein (11), phage related protein (11).
5359996-5367261	7.3	5	hypothetical protein (4), putative membrane protein.
5382504-5391510	9	4	putative RNA helicase, putative DNA helicase, hypothetical protein (2)
5416303-5421927	5.6	3	putative invasin, hypothetical protein (2)

Shigella flexneri 2a str. 301 vs gut metagenome

Start-End (bp)	**Length (kbp)**	**Number of ORFs**	**Main features (in brackets number of genes)**
214885-219996	5.1	0	Not annotated
229909-244882	15	19	Hypothetical protein (2), IS2 transposase InsD, insertion sequence 2 OrfA protein, IS911 (3), outer membrane usher protein, periplasmic chaperone of fimbral assembly machinery, coat protein, cytoplasmic protein (5), putative DNA stabilization protein, packaging glycoprotein, putative scaffolding protein, putative terminase large subunit.
262348-267750	5.4	9	Hypothetical protein (3), insertion element IS2 transposase InsD, insertion sequence 2 OrfA protein, IS1 (2), IS600 (2).
278203-287655	9.5	8	Hypothetical prots. (2), IS1 ORF (2), Rhs-family protein (4).
311516-332269	20.8	26	Hypothetical protein (7), integrase, IS600 ORF (8), IS629 ORF (3), putative bactoprenol glucosyl transferase, putative flippase, putative glucosyl tranferase II, putative phage integrase (3), putative phage tail fibre protein.

374550-380512	6	5	Hypothetical protein (4), IS1 ORF.
511825-512589	0.8	1	phosphopantetheinyltransferase component of enterobactin synthase multienzyme complex
618107-624102	6	6	Hypothetical protein (2), Rhs-family protein (2), IS1 (2).
699193-750156 (low coverage region)	51	60	Capsid protein small subunit, endopeptidase, head-tail preconnector gp5 (3), host specificity protein, hypothetical protein (3), insertion element IS2, insertion sequence 2 OrfA protein, invasion plasmid antigen, IS600 ORF (7), IS629 ORF (3), IS911 ORF (4), major capsid protein, minor tail protein, putative bacteriophage protein (19), putative Q protein, putative replication protein DnaC, putative S protein, putative tail assembly protein, putative tail attachment protein, putative tail component, putative tail component of prophage CP-933K (7), putative tail length tape measure protein precursor.
898345-921546	23.2	28	DNA-binding transcriptional regulator DicC, hypothetical protein (6), insertion element IS2 transposase InsD, insertion sequence 2 OrfA protein, IS2 ORF2, IS600 ORF (6), IS911 ORF (2), ISSfl4 ORF (3), putative bacteriophage protein (5), putative exodeoxyribonuclease VIII of prophage CP-933R, transcriptional repressor DicA.
984277-991700	7.4	9	Fimbrial protein, hypothetical protein, IS1 ORF (3), putative fimbrial-like protein (2), putative outer membrane protein, putativi pili assembly chaperone.
1034215-1038587	4.4	4	Hypothetical protein (3), putative regulator.
1090857-1103061	12.2	15	Hypothetical protein (3), insertion element IS2 transposase InsD, insertion sequence 2 OrfA protein, IS600 ORF (2), IS629 ORF (4), IS911 ORF (4).
1181034-1188414	7.4	9	Hypothetical protein (4), putative head maturation protease of prophage CP-933C, putative head portal protein, putative head-tail adaptor, putative holin protein of prophage CP-933C, putative terminase of prophage CP-933C.
1390300-1395946	5.6	8	Hypothetical protein, insertion element IS2 transposase InsD, insertion sequence 2 OrfA protein, IS600 ORF (2), putative bacteriophage protein (2), putative replication protein.
1399435-1405421	6	8	putative bacteriophage protein (4), IS600 ORF (2), hypothetical protein (2)
1420617-1425133	4.5	5	IS ORF (3), hypothetical protein, invasion plasmid antigen.
1626770-1634439	7.7	9	Hypothetical prot. (5), IS911 ORF (2), putative integrase (2)
1917457-1941923	24.4	25	Phage /mobile elements island Host specificity protein, invasion plasmid antigen, IS1 ORF (2), IS600 ORF (2), ISSfl2 ORF, minor tail protein (8), putative crossover junction endodeoxyribonuclease, putative DNA-packaging protein, putative membrane protein precursor, putative Q antiterminator encoded by prophage CP-933P, putative serine protease, putative tail component of prophage CP-933K (6), putative tail length tape measure protein precursor.
2044876-2053192	8.3	7	Hypothetical protein (3), IS1 ORF2, IS600 ORF2, iso-IS10R ORF, putative tail protein.
2107750-2115293	7.5	9	LPS SYNTHESIS ISLAND dTDP-rhamnosyl transferase (2), glycosyl transferase, glycosyl translocase, hypothetical protein (3), O-antigen polymerase, polysaccharide biosynthesis protein.
2227913-2236010	8.1	9	DNA-damage-inducible protein, hypothetical protein (3), IS1 ORF (2), iso-IS10R ORF, putative tail fiber assembly protein, putative tail fiber protein.
2552420-2557999	5.6	6	Hypothetical protein (6). Between genes of peptydoglycan synthesis.
2589066-2598121	9.1	7	Hydrogenase 4 Fe-S subunit, hydrogenase 4 membrane subunit, hydrogenase 4 subunit D, hydrogenase 4 membrane subunit, hydrogenase 4 subunit F, large subunit of hydrogenase 3 (2).
2684580-2696870	12.3	15	DNA-invertase, hypothetical protein (4), insertion element IS2 transposase InsD, insertion sequence 2 OrfA protein, invasion plasmid antigen, IS600 ORF (2), putative bacteriophage protein, putative tail component of prophage CP-933K, putative tail fiber assembly protein, putative tail fiber protein (2).
2754519-2761442	6.9	5	Integrase, IS1 ORF (2), IS3 ORF (2).
2831678-2836482	4.8	5	IS600 ORF (2), putative DNA-binding protein, putative phage transposase, putative regulatory protein.
2947609-2957653	10	11	Hypothetical protein (5), IS3 ORF (2), IS911 ORF (3), integrase.
3069582-3070249	0.7	2	Shet1A, Shet1B *Region 3042550-3097662 corresponds to well established PAI1. It shows good recruitment with small islands, which correspond exactly to regions where pathogenic genes are annotated*
3099253-3107662	8.4	11	Hypothetical protein (4), IS1 (4), IS3 (2), IS10R ORF.
3277113-3284686	7.6	7	Insertion sequence 2 OrfA protein, IS1 ORF (2), IS2 ORF, putative chaperone, putative fimbrial protein, putative IS2
3464484-3475679 (low recruitment)	11.2	10	DNA-binding transcriptional regulator FrlR, fructoselysine 3-epimerase, fructoselysine 6-kinase, fructoselysine-6-P-deglycase, hypothetical protein (2), nitrite reductase (NAD(P)H) subunit, nitrite reductase small subunit, nitrite reductase, NirC protein, siroheme synthase.
3592829-3613524	20.7	20	Hypothetical protein (11), insertion element IS2 transposase InsD, insertion sequence 2 OrfA protein, IS1 ORF (2), putative ATP-binding component of a transport system, putative IS1 encoded protein (2), putative outer membrane pore protein, putative periplasmic binding transport protein.
3812410-3836346	23.9	22	ColV-immunity protein, hypothetical protein (3), insertion element IS2 transposase InsD (2), insertion sequence 2 OrfA protein (2), IS1 ORF (3), IS629 ORF (2), lysine:N6-hydroxylase, putative ferric siderophore receptor, putative fimbrial protein, putative long polar fimbriae, putative membrane transport protein, serine protease, siderophore biosynthesis protein (3). *Region 3806404-3835200 established as PAI2. Function : Iron uptake systems, immunity to colicin V (IucA, IucB, IucC, IucD, IutA, aerobactin). Insertion site : tRNA-selC*

4047837-4053555	5.7	1	Protoporphyrinogen oxidase, 5 kb with nothing anotated.
4185519-4191708	6.2	1	Glutamate racemase, 6 kb with nothing anotated.
4226640-4232128	5.5	0	Not annotated
4410160-4414498	4.3	8	Hypothetical protein (4), IS1 ORF (2), oxidoreductase, pyrBI operon leader peptide.
4530286-4535960	5.7	9	30S ribosomal protein S18, 50S ribosomal protein L9, endoribonuclease SymE, hypothetical protein (3), IS600 ORF (2), ISEhe3 orfB.

Streptococcus sanguinis SK36 vs oral metagenome

Start-End (bp)	**Length (kbp)**	**Number of ORFs**	**Main features (in brackets number of genes)**
16589-22064	5.5	0	Not annotated
143437-169792	26.4	23	ATPase with chaperone activity ATP-binding subunit putative, Conserved hypothetical protein (9), Cro-like transcriptional repressor XRE family putative, Hypothetical protein (11), Uncharacterized protein
392676- 398388	5.7	4	Conserved hypothetical protein (4)
708764-716005	7.2	9	Putative ABC-type multidrug/protein/lipid transport system (pediocin PA-1 exporter) ATPase and permease components, putative Arsenical resistance operon transcription repressor (ArsR), putative FmtA-like protein, Hypothetical protein (2), putative Integral membrane protein, putative metal-dependent membrane protease, putative protease, putative transposase.
807557-813010	5.4	2	Platelet-binding glycoprotein, putative glycosyltransferase.
1114985-1123882	8.9	3	putative calcium binding hemolysin-like protein, putative hemolysin exporter ATPase component, multidrug resistance efflux pump/hemolysin secretion transmembrane protein.
1168936-1179877	10.9	9	Putative beta-glucosides PTS (2), beta-N-acetylhexosaminidase, conserved hypothetical protein, dihydrolipoamide acetyl transferase E2 component, glycosyl hydrolase family 1, putative Na+-driven multidrug efflux pump, phosphotransferase system, cellobiose-specific component IIA, tautomerase.
1306316-1310922	4.6	4	Conserved hypothetical protein (4)
1631616-1639447	7.8	5	Putative sortase-like protein, putative Surface protein, putative, imA fimbrial subunit-like protein, Heme utilization/adhesion exoprotein, Hypothetical protein
1742320-1747543	5.2	8	Conserved uncharacterized protein, hypothetical protein (6), putative transcriptional regulator GntR family (repressor of trehalose operon).
1796947-1806223	9.3	7	Conserved hypothetical protein (4), hypothetical protein, putative modification methylase, putative very short patch repair endonuclease.
1880884-1888335	7.5	5	Conserved hypothetical protein (4), putative acetyltransferase.
1975960-1981997	6	2	Putative cell surface SD repeat antigen precursor, conserved hypothetical protein.
2083756-2092917	9.1	10	Putative acetyltransferase, putative carbohydrate isomerase AraD/FucA family, putative carbohydrate kinase FGGY family, putative conserved hypothetical protein (5), putative glyoxylate reductase NADH-dependent, putative phosphotransferase system (PTS) galactitol-specific IIC component.
2172522-2178059	5.5	5	Hypothetical protein (4), putative 2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase.

Streptococcus pneumoniae R6 vs oral metagenome

Start-end (bp)	**Length (kbps)**	**Number of ORFs**	**Main features: genes (in brackets, number of genes)**
14358-19982	5.5	0	Not annotated
29729-30521	0.8	2	degenerate transposase (orf1), hypothetical protein
81093-83757	3.7	1	cell wall surface anchor family protein
113103-118087	5	5	transporter, truncation (2), hypothetical protein (3)
118343-126749	8.4	6	Hypothetical proteins
127111-130450	3.3	10	hypothetical proteins
140019-149379	9.3	8	degenerate transposase (3), glycosyltransferase involved in exopolysaccharide (EPS) synthesis, glycosyl transferase family protein, ABC transporter ATP-binding protein, hypothetical protein UDP-glucose dehydrogenase
218390-221322	3	4	DEOR-type transcriptional regulator, hypothetical protein, PTS system, IIA and B components
275026-281685	6.6	8	6-phospho-beta-glucosidase, hypothetical protein, cellobiose phosphotransferase system IIB, C and A components, BigG family transcription antiterminator, hypothetical protein (2)
296522-300559	4	5	preprotein translocase, YajC subunit, hypothetical protein, GalR family transcription regulator, hypothetical protein (2)
314511-318935	4.4	4	hypothetical proteins
323723-325132	1.4	3	hypothetical proteins
349704-351435	1.7	3	DNA alkylation repair enzyme, truncated, hypothetical protein, choline binding protein G, truncated, choline binding protein G
354576-360056	5.5	7	PTS system, mannitol-specific IIBC components, transcriptional regulator, mannitol-specific enzyme IIA component, mannitol-1-phosphate 5-dehydrogenase, hypothetical protein (2), trigger factor
415208-425429	10.2	12	hypothetical protein (3), ROK family protein, PTS system, cellobiose-specific IIC component, hypothetical protein, PTS system, lactose-specific IIA component, 6-P-beta-galactosidase phosphotransferase system sugar-specific EII

			component
451863-454743	2.9	3	type I restriction-modification system S subunit (2), integrase/recombinase
454834-458104	3.3	2	type I restriction-modification system, M and R subunits
473077-474055	1	2	hypothetical proteins
474868-476332	1.5	2	hypothetical proteins
496875-503646	6.8	9	degenerate transposase (2), hypothetical protein (5), BglG family transcriptional antiterminator, PTS system, beta-glucosides-specific IIABC components
539862-542483	2.6	3	tributyrin esterase, Serine/alanine adding enzyme, beta-lactam resistance factor
556171-558361	2.2	0	Not annotated
591533-593553	2	2	Zinc metalloprotease
611637-614204	2.6	3	hypothetical proteins
618769-622463	3.7	4	hypothetical protein (2), HesA/MoeB/ThiF family protein ABC transporter ATP-binding protein - unknown substrate
622971-624168	1.2	2	hypothetical protein, degenerate transposase
625795-629397	3.6	6	putative ribonuclease BN, cytochrome c-type biogenesis protein CcdAhypothetical protein, hypothetical protein (3), ABC transporter ATP-binding protein
651680-653048	1.4	2	degenerate transposase, hypothetical protein
698647-700072	1.4	2	hypothetical proteins
719596-721865	2.3	3	transposases
733869-735940	2	3	hypothetical protein (2), hemolysin-related protein
756396-757933	1.5	3	ABC transporter substrate-binding protein - oligopeptide transport, internal deletion, hypothetical protein (2)
804421-808057	3.5	6	hypothetical protein (5), degenerate transposase (orf1)
820142-821585	1.4	3	degenerate transposase (2), spermidine synthase
838923-841643	2.7	3	hypothetical protein (4), tranposase (orf1 and 2)
852353-854751	2.4	3	hypothetical protein (2), ABC transporter ATP-binding protein
886374-888345	2	5	degenerative transposase (2), hypothetical proteins
893052-899548	6.5	4	pneumococcal histidine triad protein D and E precursor, hypothetical protein, pneumococcal histidine triad protein E precursor, truncation
903628-904825	1.2	2	GtrA family protein, hypothetical protein
918685-927808	9.1	10	hypothetical protein, iron compound-binding protein, iron-compound ABC transporter, permease protein (2), iron-compound ABC transporter, ATP-binding protein
936137-941289	5.1	7	hypothetical protein (2), Tn5252 ORFs (4), putative positive transcriptional regulator MutR
941758-955165	13.4	9	UDP-N-acetyl-D-mannosaminuronic acid dehydrogenase, hypothetical proteins (3), nikkomycin biosynthesis protein, carboxylase, ABC transporter membrane-spanning permease - macrolide efflux
971190-973769	2.6	4	degenerate transposase (3)
1052960-1054225	1.3	2	hypothetical proteins
1056785-1058377	1.6	1	pneumococcal histidine triad protein A precursor
1066637-1068010	1.4	1	6-phospho-beta-galactosidase
1069773-1071118	1.3	2	PTS system, lactose-specific IIBC and A component
1087626-1090825	3.2	2	hypothetical proteins
1099892-1102232	2.3	1	type II restriction endonuclease, putative
1152329-1155285	3	2	polysaccharide biosynthesis protein, putative, required for phosphorylcholine incorporation in teichoic and lipoteichoic acids, licD protein, carbamoyl phosphate synthase large subunit
1184169-1187331	3.2	3	N-acetylneuraminate lyase subunit, truncation, cytidine deaminase, hypothetical protein (2), ABC transporter ATP-binding protein
1189233-1195071	5.8	6	ABC transporter membrane-spanning permease - oligopeptide transport (3), hypothetical protein, N-acetylmannosamine-6-phosphate 2-epimerase, degenerate transposase (2)
1195546-1205849	10.3	10	hypothetical protein (5), ABC transporter ATP-binding protein, drug efflux ABC transporter, ATP-binding/permease protein, prolyl oligopeptidase family protein
1268214-1270759	2.5	2	choline binding protein
1279559-1292573	13	11	Protease, Type II restriction endonuclease (3), hypothetical protein (3), ABC transporter ATP-binding protein (3), hypothetical proteins, degenerate transposase (2)
1330104-1333603	3.5	5	cell wall surface anchor family protein, hypothetical proteins, degenerate transposase (2)
1381557-1384182	2.4	1	hypothetical protein
1414003-1416052	2	1	1,4-beta-N-acetylmuramidase
1460011-1462499	2.5	3	hypothetical proteins
1510861-1515990	5.1	5	ABC transporter permease (2), ABC transporter substrate-binding protein, hypothetical protein, sialidase A precursor (neuraminidase A)
1529462-1533414	4	5	ABC transporter ATP-binding protein, hypothetical prot. (4)
1590612-1592989	2.3	3	sucrose-6-phosphate hydrolase, putative, ABC transporter membrane-spanning permease - sugar transport
1602308-1604450	2.1	3	transcriptional regulator, hypothetical protein
1611600-1613961	2.3	3	catabolite control protein, hypothetical protein, Mg2+ transporter
1682796-1684507	1.7	1	sugar ABC transporter, permease protein (2)

1691699-1697659	6	0	Not annotated
1713561-1722800	9.2	13	Pneumolysin, hypothetical proteins (5), degenerate transposase (orf1/2), N-acetylmuramoyl-L-alanine amidase
1729774-1741445	11.7	13	hypothetical proteins (6), transcriptional activator, bacteriocin formation protein, putative, toxin secretion ABC transporter, ATP-binding/permease protein, subtilisin-like serine protease
1771172-1775517	4.4	4	hypothetical protein (2), ABC transporter ATP-binding protein - unknown substrate, transcriptional regulator PlcR, putative
1788552-1795751	7.2	0	Not annotated BlastX indicates similarity to a Competence-specific global transcription modulator
1802154-1807851	5.7	7	Transposase (2), nicotinate-nucleotide pyrophosphorylase, hypothetical proteins (3), Beta-glucosidase
1808352-1811677	3.3	3	PTS system, IIC, B and A components
1851852-1856497	4.3	0	Not annotated
1871050-1877198	6.1	7	tranposase (orf2), hypothetical protein, response regulator sensor histidine kinase PnpS, phosphate ABC transporter phosphate-binding protein (4)
1877695-1879688	2		phosphate transporter PhoUm, truncated IS1380-Spn1 transposase, transcriptional regulator
1893316-1895607	2.3	3	hypothetical protein (2)
1917722-1920516	2.8	4	transketolase, C-terminal subunit, putative transketolase n-terminal section, PTS system ascorbate-specific transporter subunit IIC
1920916-1923966	3	5	PTS system, IIB component, putative, hypothetical prot. (2)
1924635-1929904	5.3	6	hypothetical protein, 50S ribosomal protein L32 and L33, choline binding protein PcpA, degenerate transposase (orf1), tranposase (orf2), hypothetical protein
1953474-1965988	12.5	13	fucolectin-related protein, hypothetical protein, PTS system, IIA, B, C and D components, fucose pathway protein, function unknown, L-fuculose phosphate aldolase, fucose kinase, fucose operon repressor, putative
1974387-1978096	3.7	2	hypothetical proteins
1987348-1992052	4.7	6	choline binding protein A, hypothetical protein, histidine kinase, response regulator

Neisseria_meningitidis FAM18 vs oral metagenome

46181-52657	6.5	2	Non-annotated region (5 kb), hypothetical protein, putative DNA transport competence protein.
55856-59447	3.6	5	LPS ISLAND. Putative inner membrane transport protein, dTDP-4-dehydrorhamnose 3,5-epimerase, glucose-1-phosphate thymidylyltransferase, dTDP-glucose 4,6-dehydratase, UDP-glucose 4-epimerase.
59826-68935	9.1	7	Hypothetical protein, alpha-2,9-polysialyltransferase, N-acetylneuraminic acid synthetase, acylneuraminate cytidylyltransferase, N-acetylglucosamine-6-phosphate 2-epimerase, capsule polysaccharide export outer membrane and inner membrane protein (2)
73146-82686	9.5	6	capsule polysaccharide modification protein (2), dTDP-D-glucose 4,6-dehydratase, dTDP-glucose 4,6-dehydratase, glucose-1-phosphate thymidylyltransferase, hypothetical protein.
89795-96284	6.5	8	hypothetical protein (3), IS1016 transposase partial CDS, putative inner membrane protein (2), putative protein export protein (2).
207663-210169	2.5	3	hypothetical protein, putative integral membrane protein, class II pilin PilE.
213814-216632	2.8	1	RNA polymerase sigma factor solo coge 400pb, 2388pb sin hits adicionales
236567-239209	2.6	2	Putative inner membrane protein (2)
278434-288427	10	12	hypothetical protein (6), putative inner membrane protein (2), putative invertase/transposase, putative periplasmic protein, putative rotamase, TspB protein.
374923-381021	6.1	2	Hypothetical protein, pilus-associated protein.
454574-485141	30.6	23	Hemagglutination island. Hypothetical protein (14), N-acetyl-gamma-glutamyl-phosphate reductase, putative hemagglutinin (2), putative hemagglutinin/hemolysin-related protein (2), putative hemolysin activator, lipoprotein, periplasmic protein, secretion protein.
551138-561195	10	4	Putative peptidase, putative transposase, putative outer membrane protein, iron-regulated protein FrpC.
620616-625974	5.4	8	Adhesin Maf (2), hypothetical protein (5), ribonuclease inhibitor barstar.
669848-675838	6	1	IgA1 protease
810425-817429	7	11	Hypothetical protein (8), putative integral membrane protein, putative periplasmic protein (2).
847060-877132	30	35	Hypothetical protein (23), putative D-lactate dehydrogenase-related protein, putative lipoprotein, putative Phage integrase, putative phage related protein (8), transcriptional regulator.
883641-891278	7.6	8	HlyD family secretion protein, hypothetical protein (2), ParA protein, putative ABC transporter, putative integral membrane protein (2), putative ParB protein.
899469-904123	4.5	3	putative acyl-CoA hydrolase, pseudogene (opacity protein), putative transposase for IS1655.
1038354-1046140	7.8	8	Phage island. Hypothetical protein (5), putative host-nuclease inhibitor protein, phage tail fibre protein, transposase for IS1655.
1299450-1302814	3.4	1	Putative type III restriction/modification system enzyme
1327713-1338626	10.9	3	putative cytolysin secretion ABC transporter, insertion element IS1016 transposase, putative RTX iron-regulated frpc protein outer membrane.
1401490-1405236	3.7	2	Pseudogene (opacity protein), hypothetical protein.
1480022-1489392	9.4	2	Lactoferrin binding protein

1599194-1605603	6.4	2	Hypothetical protein, putative DNA transport competence protein.
1692689-1700188	4.5	10	Hypothetical protein (5), putative cell-surface protein, putative integral membrane protein (2), putative periplasmic type I secretion system protein (2).
1741953-1754180	12.2	11	Hypothetical protein (5), Opa1800 outer membrane protein precursor, putative integral membrane protein (3), putative transposase, TspB protein.
1833155-1844849	11.7	17	Adhesin Maf (2), hypothetical protein (14), putative lipoprotein.
1890038-1902422	12.4	10	Hypothetical protein (5), putative integral membrane protein (2), putative invertase/transposase, putative replication initiation factor, TspB protein.
1905347-1910613	5.3	5	Hypothetical protein (3), pseudogene (outer membrane protein), putative DNA transport competence protein.
1965871-1971037	5.2	1	Putative outer membrane peptidase
1990115-1996307	6.2	2	Hypothetical protein, Ig-Aspecific serine endopeptidase.
2040719-2048085	7.4	3	Hypothetical protein, putative pilin, putative DNA transport competence protein.
2133941-2143373	9.4	15	Adhesin (2), hypothetical protein (12), mafb protein (fragment).

Supplementary Table 3. List of strains analyzed and their corresponding accession numbers.

STRAIN	GENOMIC ELEMENT	ACCESSION NUMBER
Shigella flexneri 2a str 301	Chromosome	NC_004337.1
	Plasmid pCP301	NC_004851.1
Escherichia coli E243777A ETEC	Chromosome	NC_009801.1
	Plasmid pETEC_80	NC_009786.1
	Plasmid pETEC_74	NC_009790.1
	Plasmid pETEC_73	NC_009788.1
	Plasmid pETEC_35	NC_009787.1
Escherichia coli CFT073	Chromosome	NC_004431.1
Escherichia coli K12 MG1655	Chromosome	NC_000913.2
Escherichia coli O127:H6 str. E2348/69	Plasmid pMAR2	NC_011603.1
Escherichia coli O157:H7 Sakai str.	Chromosome	BA000007.2
Salmonella enterica subsp. enterica serovar Typhi str. CT18	Chromosome	NC_003198.1
Neisseria meningitidis 053442	Chromosome	NC_010120.1
Neisseria meningitidis FAM18	Chromosome	NC_008767.1
Neisseria gonorrhoeae FA1090	Chromosome	NC_002946.2
Streptococcus sanguinis SK36	Chromosome	NC_009009.1
Streptococcus pneumoniae R6	Chromosome	NC_003098.1

3.4

“Microbiota diversity and gene expression dynamics in human oral biofilms”

A Benítez-Páez*, P Belda-Ferre*, A Simón-Soro, A Mira

BMC Genomics 2014, 15:311

BMC Genomics

RESEARCH ARTICLE Open Access

Microbiota diversity and gene expression dynamics in human oral biofilms

Alfonso Benítez-Páez[1,2*†], Pedro Belda-Ferre[1†], Aurea Simón-Soro[1] and Alex Mira[1*]

Abstract

Background: Micro-organisms inhabiting teeth surfaces grow on biofilms where a specific and complex succession of bacteria has been described by co-aggregation tests and DNA-based studies. Although the composition of oral biofilms is well established, the active portion of the bacterial community and the patterns of gene expression *in vivo* have not been studied.

Results: Using RNA-sequencing technologies, we present the first metatranscriptomic study of human dental plaque, performed by two different approaches: (1) A short-reads, high-coverage approach by Illumina sequencing to characterize the gene activity repertoire of the microbial community during biofilm development; (2) A long-reads, lower-coverage approach by pyrosequencing to determine the taxonomic identity of the active microbiome before and after a meal ingestion. The high-coverage approach allowed us to analyze over 398 million reads, revealing that microbial communities are individual-specific and no bacterial species was detected as key player at any time during biofilm formation. We could identify some gene expression patterns characteristic for early and mature oral biofilms. The transcriptomic profile of several adhesion genes was confirmed through qPCR by measuring expression of fimbriae-associated genes. In addition to the specific set of gene functions overexpressed in early and mature oral biofilms, as detected through the short-reads dataset, the long-reads approach detected specific changes when comparing the metatranscriptome of the same individual before and after a meal, which can narrow down the list of organisms responsible for acid production and therefore potentially involved in dental caries.

Conclusions: The bacteria changing activity during biofilm formation and after meal ingestion were person-specific. Interestingly, some individuals showed extreme homeostasis with virtually no changes in the active bacterial population after food ingestion, suggesting the presence of a microbial community which could be associated to dental health.

Keywords: Dental plaque, Metatranscriptomics, Biofilm formation, Human microbiome, RT-qPCR, RNAseq

Background

The study of microbial communities from environment- and human-derived samples through Next Generation Sequencing (NGS) methods has revealed a vast complexity in those ecological niches where hundreds or thousands of microbial species co-inhabit and functionally interact. One of these complex communities is that found in the human oral dental plaque (hereinafter, human oral biofilm). Although some studies, using NGS methods and 16S rRNA-based analysis, estimate that microbial diversity of the oral cavity is composed by thousands of species [1], more recent data have limited these estimates to a few hundreds [2-4]. Contrary to Koch's postulates, dental caries is not considered etiologically the outcome of a single-agent but is associated to an unbalance of microbial species that synergistically cause enamel demineralization by their acidogenic activity [5,6]. Thus, characterizing the composition of whole bacterial communities that actively engage in biofilm formation and sugar fermentation after the ingestion of food is vital for understanding community dynamics under health and disease conditions [7].

* Correspondence: abenitez@cidbio.org; mira_ale@gva.es
†Equal contributors
[1]Oral Microbiome Group – Department of Health and Genomics, Center for Advanced Research in Public Health (CSISP-FISABIO), Avda. Catalunya 21, 46020 Valencia, Spain
[2]Bioinformatics Analysis Group – GABi. Centro de Investigación y Desarrollo en Biotecnología (CIDBIO), Bogotá, D.C 111221, Colombia

Although the set of species present in the human oral biofilm is almost fully depicted, new efforts have to be conducted to establish microbial agonistic or antagonistic associations, to distinguish actively-growing bacteria from inactive or transient species, as well as to outline the role of individual species during biofilm formation on tooth surfaces. The co-aggregation detected to occur between streptococci and *Actinomyces* species has been proposed to be a major promoter of human oral biofilm formation [8]. Like most biofilms, the dental plaque is built in a continued process characterized by succession of different bacterial species, each one with relevant roles in every step of biofilm construction [9]. Formation of the oral biofilm could be dissected in three major stages, namely: i) attachment; ii) colonization; and iii) biofilm development [10]. However, species participating of the entire process are traditionally characterized as "early" and "late" colonizers, where early colonizers would be responsible of the two first stages [9]. Among early colonizers the viridans streptococci group is considered as a cornerstone of the oral biofilm puzzle given its ability to bind saliva proteins through Antigens I and II. In this manner, streptococci species become the first colonizers able to bind tooth surfaces and promoting arrival of secondary colonizers by intergeneric coaggregation (reviewed in [9]). *Actinomyces naeslundii* is one of the secondary colonizers and a well known coaggregation partner of streptococci [8,11]. *Fusobacterium nucleatum* is considered a key player given its capability to coaggregate both with early and late colonizers of the oral biofilm [12], the latter group characterized by species belonging to Bacteroidetes and Spirochaetes [6,9]. It is noteworthy to highlight that intergeneric coaggregation not only contributes to bacterial growth and colonization [13], but it is thought to facilitate the genetic and metabolic exchange among species, and even to create the adequate environment for arrival of some obligate anaerobic bacteria [10]. Therefore, any disruption in the development of the oral biofilm caused by impairing of early colonizers tooth attachment or inability to recruit other key players during biofilm formation, would affect the entire process avoiding presence of pathogens responsible for periodontal disease or caries [7].

Although few attempts to link specific gene expression profiles in oral bacteria with the establishment and maturation of oral biofilm have been done [14], further studies are needed to understand global gene dynamics and intracellular signalling which are the basis for cell-to-cell communication among oral bacteria and to promote biofilm formation on tooth surfaces. There are important limitations to study gene expression from *in vivo* oral samples, including RNA instability and amounts of sampling material, but a sequencing approach of total cDNA from an *in vitro* oral biofilm model has recently been performed [15].

Because gene transcripts typically occupy a small fraction of total bacterial RNA, even after mRNA-enriching protocols, a massive coverage is normally required to quantify gene expression by RNAseq technologies. On the other hand, high-coverage sequencing technologies are normally coupled to short read lengths, which jeopardize accurate taxonomic assignment of the sequences. The latter can be achieved through the use of longer reads, at the expense of a lower coverage of mRNA transcripts. In the present manuscript we present the first metatranscriptome analysis of *in vivo* human oral biofilm samples through two approaches: A short read-length, high coverage Illumina® approach to study oral biofilm formation through time, and a long read-length, lower coverage pyrosequencing strategy to study changes in community composition before and after a meal. For the first approach, a total of 16 samples of supragingival plaque from 4 healthy individuals were collected at four different time points (6, 12, 24 and 48 hours after a professional ultrasound cleaning) to disclose the microbiota and gene expression dynamics during oral biofilm formation. For the second experiment, the metatranscriptome of dental plaque from five individuals was studied 30 minutes before and after a controlled meal, in order to characterize the potential shifts in the active bacterial community when dietary nutrients are available for growth.

Results & discussion

Metagenome vs metatranscriptome

A preliminary test was performed by direct pyrosequencing of DNA (sequence data obtained from [3]) and cDNA from a 24-hour dental plaque sample from the same individual (NoCa1). The results show a very different pattern of bacterial genera in the metagenome and the metatranscriptome (Figure 1). *Actinomyces*, *Corynebacterium* and *Neisseria* were the three most abundant genera in the RNA-based community whereas *Veillonella*, *Streptococcus* and *Leptotrichia* were the most commonly found in the total DNA-based metagenome. In addition, a long tail of low-proportion genera is observed in the metagenome but absent in the metatranscriptome, suggesting they could correspond to transient or inactive bacteria. This shows the importance of obtaining both kind of data to understand the composition and dynamics of human-associated microbial populations.

Low-coverage, long-reads approach

Active communities before and after a meal

A total of 213,419 pyrosequencing reads were obtained after quality filtering [16-18]. An average of 38.9% corresponded to SSU rRNA sequences, 59.9% to LSU rRNA and 1.2% to other sequences, including mRNA. Taxonomic assignments based on 16S and 23S rDNA

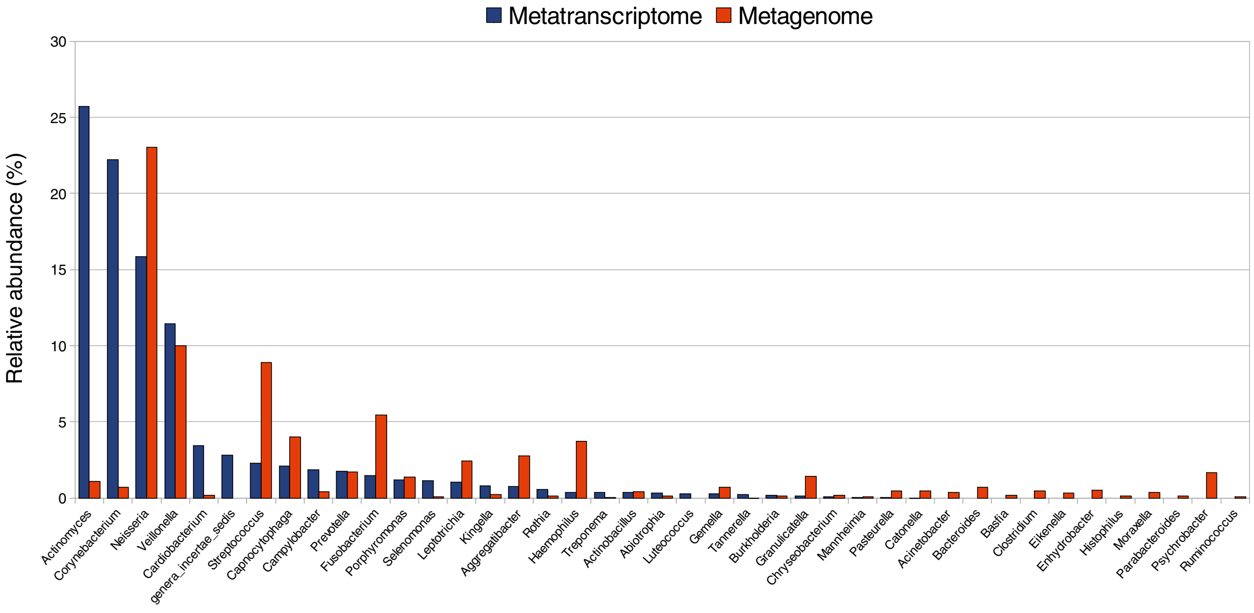


Figure 1 Total (DNA-based) and active (RNA-based) microbiota composition in the human oral biofilm from individual NoCa-01. Microbial diversity is inferred from 16S rDNA and 16S cDNA taxonomic assignment, respectively, from reads obtained by direct pyrosequencing of the DNA and cDNA of a 24 h dental plaque sample. Relative abundance of bacterial genera from metagenomic data differs from that obtained from metatranscriptomic data. The former would correspond to the total bacterial composition in the sample whereas the latter would represent the "active microbiota" as inferred by their presence in samples' RNA.

sequences gave similar results (Figure 2A and Additional file 1: Figure S1). A different bacterial composition was found for each individual. In some cases, over 80% of active bacteria corresponded to only three genera (for instance *Actinomyces*, *Corynebacterium* and *Rothia* for individuals NoCa1 and Ca2) whereas other individuals did not show any dominant genera in their active microbial community (Figure 2A, Additional file 2: Table S3). Some individuals were very resilient to changes after the meal (e.g. individual NoCa1), whereas others had more apparent changes in the proportions of some bacteria, but no specific pattern was common to all individuals (Additional file 3: Figure S2). Thus, the changes in active bacteria after a meal were not universal and depended on the original microbial population associated to each human host.

Actinomyces was the only genus found at a proportion over 10% in all samples and was found to be significantly more abundant in healthy individuals (Figure 2B) according to a high-dimensional class comparison test [19]. *Actinomyces* is an early colonizer of the oral biofilm, usually present in the first layers of dental plaque in contact with enamel [11]. It is able to increase local pH by producing ammonia through the degradation of arginine, lysilarginine and urea [20]. As a consequence, individual *Actinomyces* species could represent biomarkers of healthy biofilms with a protective role against acidogenic bacteria [21]. On the other hand, late colonizers being strictly anaerobes like *Porphyromonas*, *Fusobacterium*, *Capnocytophaga*, *Tannerella* and *Leptotrichia* were found significantly more abundant in oral biofilm from caries-bearing individuals (Figure 2B). This should be further studied given the observed over-representation of species belonging to the red complex of periodontal disease [22].

Although the number of assigned mRNA transcripts detected by this approach is quite small, we could identify some genes being expressed in all samples such as those encoding ribosomal proteins and basic housekeeping machinery like elongation factors and cell division proteins. Interestingly, we found expression of multiple sugar transport systems and central metabolism genes such as glyceraldehyde 3-phosphate dehydrogenase, L-lactate dehydrogenase, citrate synthase, enolase, and malate dehydrogenase.

High-coverage, short-reads approach

Microbial activity during oral biofilm formation

Although the high-coverage approach (~25 million Illumina reads per sample) was mainly aimed at determining gene expression patterns during biofilm formation, an attempt was first made to utilize the reads mapping to rRNA genes to characterize microbiota composition at different times of dental plaque formation. Potential errors in taxonomic assignment were minimized by 1) assigning reads at the family and genus taxonomic level only;

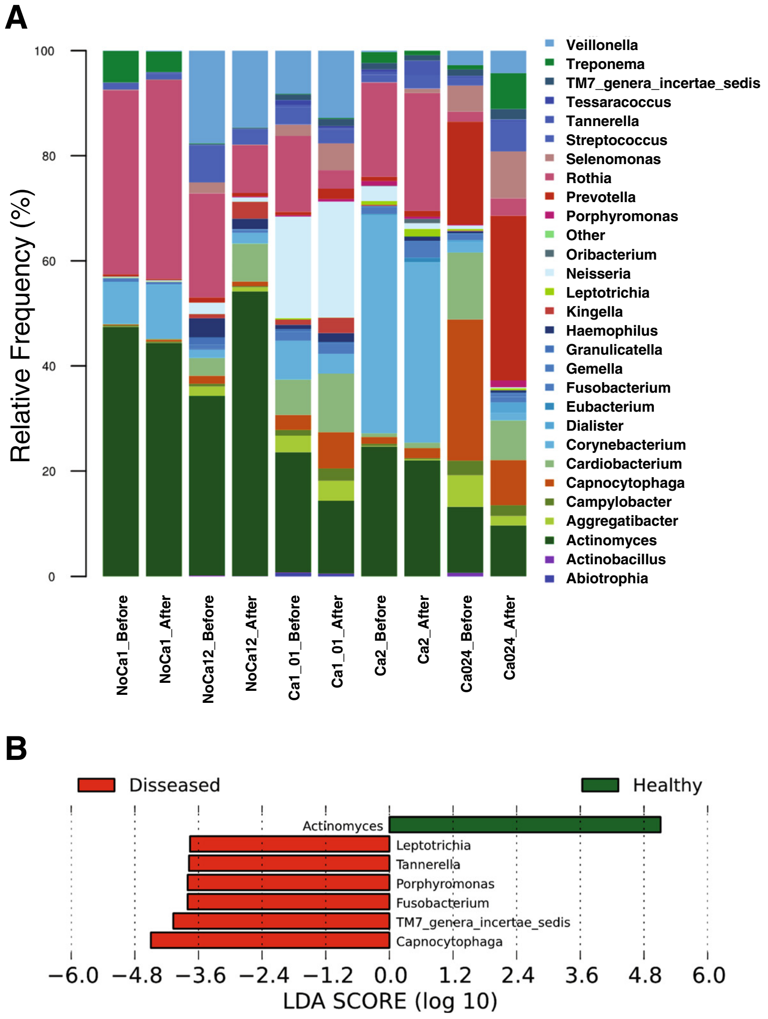


Figure 2 Meal-uptake-dependent active microbiota and association with health and disease. A – Graphical representation of the genera distribution according to 16S rRNA assignation of metatranscriptomic reads (obtained by pyrosequencng of total cDNA) based on RDP classifier. Relative frequency for most predominant active genera is shown in dental plaque samples obtained before and after a carbohydrate-rich meal. **B** – Health- and disease-associated genera in the above metatranscriptome, as inferred from Linear Discrimination Analysis (LDA) performed for dimensional class comparisons. The data were generated by using LEfSe test, available in the Galaxy Web Server toolkit. NoCa = individuals with no caries; Ca = individuals with caries.

2) selecting matches against 16S rRNA database of 100% sequence identity; and 3) eliminating hits to conserved regions of the 16S rRNA gene, keeping only hypervariable, informative regions.

When we tried to compare samples according to maturation time we found that samples from early biofilm (6 and 12 h) showed significantly ($p \leq 0.0118$) less genera than samples from mature biofilm (24 and 48 h). In average, 171 genera were found at a frequency above 0.01%, accounting for 99.4% of the global diversity. The 40 most predominant genera found in all analyzed samples are depicted in Figure 3. These 40 genera account for 68% of gene expression according to their frequency in the metatranscriptomic reads. The heat-map in Figure 3 shows the genus-level clustering according to frequency within each sample. Among predominant genera we could observe *Streptococcus* (found at relative abundances between 12 to 19% in different samples) and *Actinomyces* (in a range of 3-12%), both being well known partners for coaggregation [8,9]. Interestingly, *Actinomyces* showed higher frequencies in early biofilm samples, in agreement with its known role as early colonizer. In addition to *Streptococcus* and *Actinomyces,* other frequent genera were the Actinobacteria *Rothia, Angustibacter,* and *Kineococcus*; the Proteobacteria *Neisseria*, *Kingella* and *Alysiella*; the Firmicutes *Gemella, Paenibacillus* and *Veillonella,* the latter also reported as coaggregation partner with *Streptococcus* [23]; and finally *Capnocytophaga* and *Fusobacterium.* When we tried to discern a specific pattern of microbial organisms associated with different times of biofilm formation it was

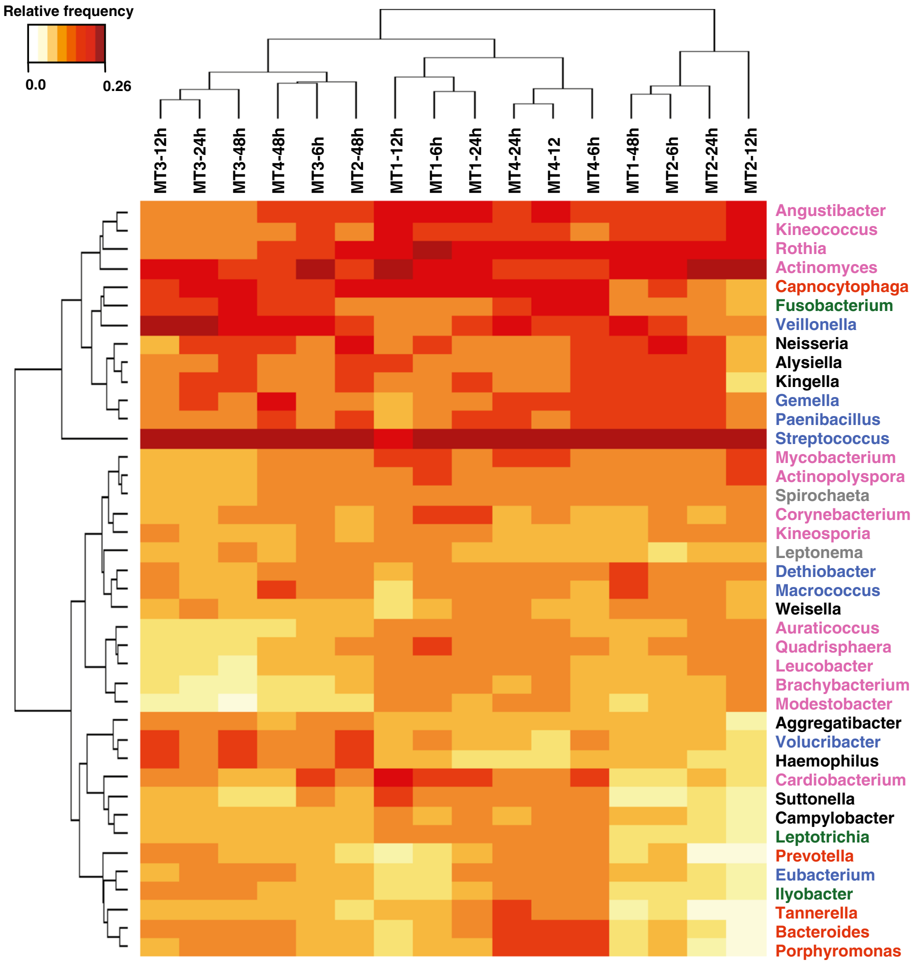


Figure 3 Clustering of oral biofilm samples according to the estimated microbial diversity. Log2 transformed frequencies of top 40 genera found in the dental plaque metatranscriptome through time are shown. Metatranscriptome samples were identified with MT prefix followed by time point (6, 12, 24, or 48 hours of biofilm formation). Number after dot refers to donor code. Bacterial genera labels are highlighted in colour according to phylum: Actinobacteria (pink), Bacteroidetes (Orange), Fusobacteria (Green), Firmicutes (Blue), Proteobacteria (Black), and Spirochaetes (Grey).

observed that samples predominantly clustered according to the donor they were extracted. Consequently, we could detect no clear association between bacterial composition and biofilm development stage. These results would globally fit within the concept that individual-specific microbial communities are a consequence of host-bacterial co-evolution to maintain host health [24,25]. Consequently, the host-specific microbiota could be considered as a genetic fingerprint almost unique for every person ([25,26] and references therein), and even preserved throughout the years in a very stable fashion [27].

Microbial interactions during oral biofilm formation

Although no association was found between specific bacteria and biofilm stage, we observed certain correlation patterns between different microbial groups which were reproducible in different patients. In order to detect monotonic functions associated to genera frequency fluctuation through time, we calculated the Spearman's rho parameter (ρ) for the top 40 more predominant genera listed in Figure 3 for all patients. Thus, we could obtain a map of significant positive and negative correlations that can indicate either pairwise interactions between genera or adaptation to similar environmental conditions (Figure 4A). Interestingly, most genera belonging to the same phylum showed positive correlations. In this way, Actinobacteria members ($\rho = 0.7346$ in average) appeared to show the same growth pattern during biofilm formation as well as Fusobacteria and Bacteroidetes ($\rho = 0.7833$ and 0.7450 on average, respectively). In contrast, genera assigned to Proteobacteria and Firmicutes showed lower correlation values ($\rho < 0.34$) because some species within these groups had different patterns of occurrence. Globally, several genera seem to have a negative

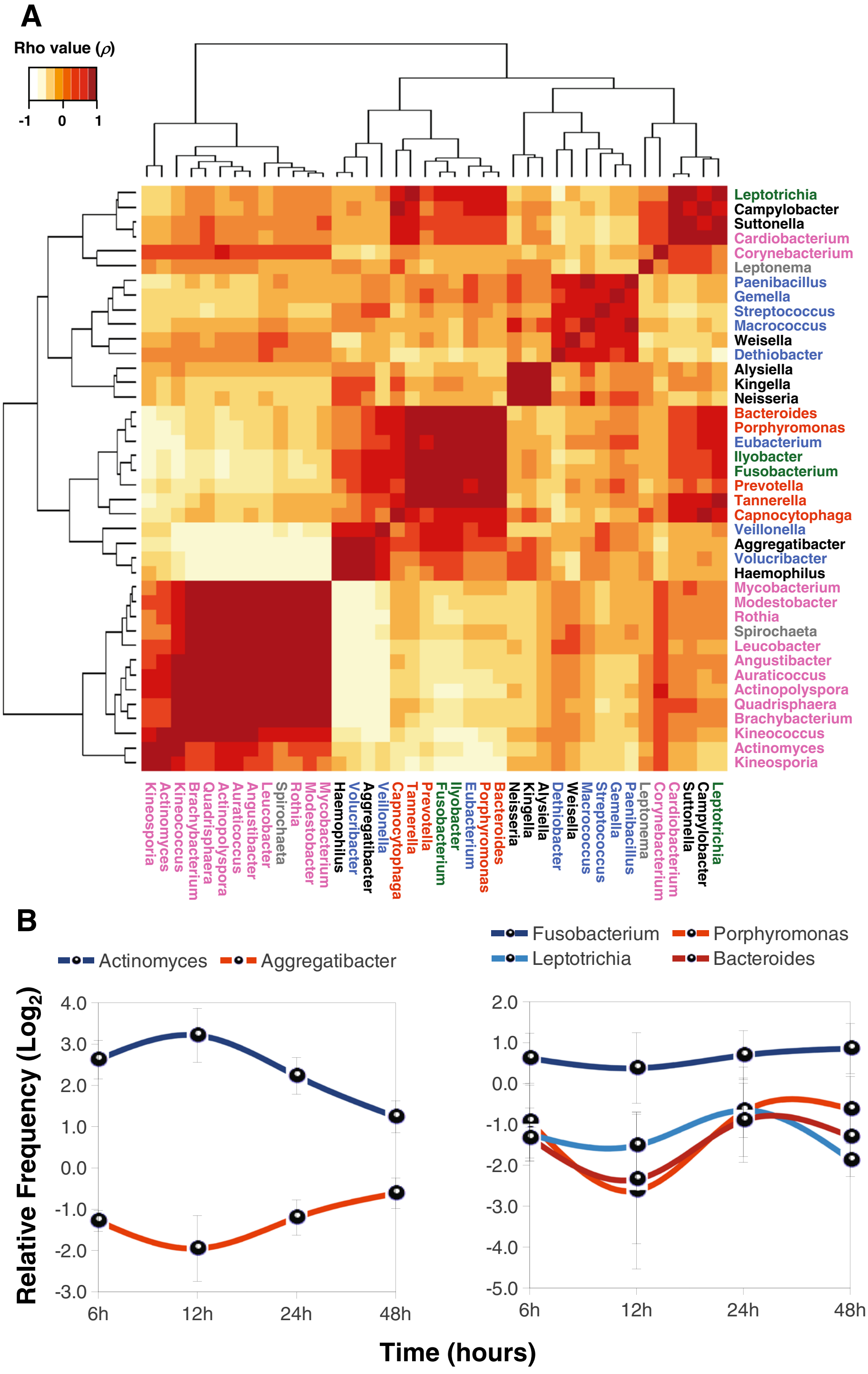


Figure 4 Positive and negative interactions in the oral biofilm active microbiota. A – Relative abundance of bacterial genera (based on 16S assignment, see Methods) and fluctuation through time were studied in a pairwise manner calculating the Spearman's rho parameter (see Methods). Genera labels are highlighted in color according to phylum: Actinobacteria (pink), Bacteroidetes (Orange), Fusobacteria (Green), Firmicutes (Blue), Proteobacteria (Black), and Spirochaetes (Grey). **B** – Detail of the Spearman correlations found in the analysis. The left panel shows a strong negative (ρ ~ - 0.75) correlation between *Actinomyces* and *Aggregatibacter* through time. In the right panel, a positive correlations is presented among late colonizers such as *Leptotrichia*, *Fusobacterium*, *Porphyromonas,* and *Bacteroides* (ρ ~ 0.80).

correlation with Actinobacteria, particularly *Veillonella* (Firmicutes) *Volucribacter* (Proteobacteria), *Haemophilus* (Proteobacteria), and *Aggregatibacter* (Proteobacteria), the latter showing strong negative correlations against 11 out of 15 different genera of Actinobacteria detected (Figure 4A). An example is given in Figure 4B (left panel), which shows the distribution of *Actinomyces* versus *Aggregatibacter* (ρ = -0.7581) during biofilm evolution. A quasi-mirror distribution is observed between these two genera suggesting that arrival and/or growth of *Actinomyces* could be outcompeted by *Aggregatibacter* presence in biofilm. In contrast, a multiple positive correlation is exemplified by the distribution of *Fusobacterium, Bacteroides, Porphyromonas* and *Leptotrichia* (Figure 4B, right panel) in full agreement with the coaggregation partners established for *Fusobacterium* and the classical view of species succession during oral biofilm formation and the establishment of late colonizers [9,12]. Indeed, Fusobacteria species seem to have the same distribution pattern than Bacteroidetes given the multiple significant positive correlations observed among their genera (Figure 4A). The presence of potential periodontal pathogens [28] in these multiple correlation patterns could be indicative of their synergy for arrival to the oral biofilm and probably for development of periodontal disease.

In vivo gene expression and functional analysis of the oral biofilm

For the functional analysis of gene expression patterns during oral biofilm formation the KEGG functional classification [29] was used. We detected expression of 19,519 genes (ORFs with aligned reads) on average per sample and the set of KO categories represented was 2,266, indicating that ~ 12% of expressed genes were functionally annotated. A distinguishable clustering according to biofilm stage was not fully depicted, although a pairwise clustering among early and mature biofilm samples was observed (data not shown). Some molecular pathways were differentially expressed between early and mature biofilm (i.e. Ribosome; Purine Metabolism; and Glycolysis). A comparison was performed between early (6 h and 12 h) and mature (24 h and 48 h) biofilm samples, determining the False Discovery Rate (FDR) with q-value ≤ 0.05. We detected a set of 271 KO categories differentially expressed (35 over-expressed in early biofilm and 236 in mature biofilm).

Over-expression in early biofilm

Over-expression of KO categories in early biofilm was predominantly grouped in genes involved in the metabolism of Carbohydrates, Energy, Amino Acids, Cofactor/Vitamins, and Xenobiotic Degradation. Translation functions were also overrepresented because several ribosome proteins showed higher expression during early biofilm as well as the Elongation Factors Tu (EF-Tu, K02358) and G (EF-G, K02355). Central role of Translation in early steps of oral biofilm formation was also evidenced by over-expression of K00566 category corresponding to the MnmA tRNA-modifying protein. MnmA is an evolutionarily conserved enzyme and incorporates the posttranscriptional modification s^2U at the wobble position of several tRNAs [30] which reads A/G ending codons during translation. The tRNA modifications are largely associated to control the fine-tuning of protein synthesis, thus improving ribosome accuracy [31]. As a consequence, they appear to be involved in controlling a wide range of bacterial phenotypic traits including biofilm formation by multi-drug resistant human pathogens [32] and could be important for oral biofilm formation as well. Our results are in agreement with previous reports where amino acid metabolism is critical for growth of early colonizers such as *Streptococcus gordonii,* which needs coaggregation to stabilize expression of genes involved in amino acid synthesis and membrane transporters [14].

Over-expression in mature biofilm

On the other hand, genes over-expressed at the late biofilm stage had a more variable functional profile. New functional categories over-expressed in late oral biofilms included the ABC transporters, the Cell motility represented by the orthology groups K03407 associated to bacterial chemotaxis, K02676 and K02390 involved in pilus and flagella assembly, respectively, and finally genes involved in Base Excision Repair, Mismatch Repair, and Homologous Recombination systems; and some putative transposases. The group of tRNA-modifying enzymes was also present in late oral biofilm with a larger set of genes over-expressed such as *mnmC, yfiC, cmoA, tadA, queE, trmK*, and the hydrouridine synthase gene *dusC.* Consequently, this molecular pathway seems to be involved in controlling the expression of proteins along the full oral biofilm process. Strikingly, a set of genes involved in competition between bacterial species also showed over-expression during this late stage of oral biofilm formation. The *comFC, comFA, comGB, comGC,* and *comGA* orthology groups are members of the Type II secretion system of which other members appear to be annotated as Competence-related DNA transformation transporters. These genes have been reported to be involved in quorum sensing response to produce mutacins; these are non-lantibiotic bacteriocins able to induce lysis and consequently DNA liberation in related species possibly supporting DNA horizontal transfer [33]. Likewise, *tfoX* orthologues were also found to be over-expressed in late oral biofilm, thus strengthening the idea of natural genetic transformation [34,35] occurring among close species in the mature oral biofilm. Globally, over-expression of these competence-related genes, permitting DNA transformations *in vivo*, could support the specific low ratio between functional diversity of genes and operational taxonomic units detected in supragingival plaque, thus indicating high functional redundancy and microbial population homogenization [36]. Other important functional categories over-expressed in late oral biofilm included those involved in Environmental Information Processing and membrane transporters, such as those belonging to the Phosphotransferase System (PTS) as well as MFS membrane receptors specialized in the importing/exporting of small molecules. Both major families of membrane transporters were found to be preferentially expressed from *Actinomyces* species indicating a high level of metabolic exchange between this genus and its environment. However, subfamilies such a *salX*-like ABC transporters associated to bacteriocin export and defense were detected to be predominantly active in streptococci species. Over-expression of some KEGG orthology categories belonging to Two-Component family of proteins indicate an active role of cells in perceiving external signals of nutrient availability in the environment. An over-expression was found of the PTS-Ntr-EIIA enzyme and the GlnB protein, both involved in nitrogen regulation, and the sigma factor 54 of the RNA polymerase involved in expression of genes for nitrogen metabolism. Therefore, processes related to nitrogen uptake/metabolism appear to be very relevant in the mature stage of oral biofilm probably indicating that nitrogen is a limiting factor for oral biofilm progression. In recent studies of cDNA massive sequencing from an *in vitro* five-species oral biofilm microbial community, similar results were obtained in terms of over-represented functions in mature biofilms [15]. Finally, *luxS* homologue in *Neisseria* spp. was found to be significantly over-expressed in early biofilm. The *luxS* genes are responsible of Autoinducer-2 (AI-2) synthesis, a molecule considered as a major interspecies signal for cell-cell communication [9,10]. Evidence for AI-2 role to control biofilm formation was previously observed when a *luxS* null strain of *S. gordonii* was unable to form a mixed-species biofilm with *P. gingivalis* [37].

Gene expression by qPCR

We selected adhesion genes involved in cell-to-matrix and cell-to-cell interactions to corroborate by qPCR the expression pattern inferred from Illumina sequencing. Among the adhesins we could find several molecules such as SspA and SspB proteins from *S. gordonii* and homologues present in several species of *Streptococcus*, all being very relevant for attachment to tooth surfaces [9]. Type 1 and Type 2 Fimbriae molecules found in *Actinomyces* species are described to mediate coaggregation with streptococci species. The surface expression of before mentioned proteins is sortase A (SrtA) dependent [9,38,39], catalyzing a peptidic linking to the cell wall [40] and promoting interactions with the extracellular molecules from bacterial counterparts or host tissues in the case of pathogens [39]. Regarding the role of *Actinomyces* sp. Type 2 Fimbriae in coaggregation with *Streptococcus* sp. [8,41], we studied the expression pattern of its gene (*fimA*) together with the *A. naeslundii* Fimbriae-Associated protein gene (*srtA*) and some other adhesins from *Streptococcus gordonii*. The Illumina- and qPCR-derived expression patterns during the oral biofilm formation for *fimA* and *srtA* homologue from *A. naeslundii* are showed in Figure 5A. qPCR data showed that these two genes had similar expression patterns, thus suggesting a co-expression pattern and quite probably dependent on their clustered localization in the *Actinomyces naeslundii* chromosome. We found that expression patterns were similar at all time points of oral biofilm formation from all patients with high correlation coefficients. In addition to the high degree of correlation between Illumina and qPCR expression data, all expression patterns of these genes present a common feature, namely a high level of expression at very early stage of oral biofilm formation and then decaying in a slight or noticeable manner. Once we showed gene expression dynamics of adhesins from *A. naeslundii* and *S. gordonii* during oral biofilm formation, we observed our data is in agreement with other *in vivo* analysis performed by immunodetection of surface molecules by fluorescence labelling and co-localization [8]. SspA protein is involved in attachment of *Streptococcus gordonii* to the tooth surface by recognizing salivary agglutinins and it also mediates interaction with *Actinomyces* sp. [42,43]. The expression profile from *sspA* from *S. gordonii* is presented in Figure 5B (top panel). The correlation coefficient between qPCR and Illumina data is ~ 0.91. Its expression pattern is similar to that found for Type 2 Fimbriae genes from *A. naeslundii*, which is in agreement with the role of streptococci as first colonizers and the requirement of SspA for attaching to the tooth surface and promote arrival of other early colonizers such as *Actinomyces* sp. [9,44]. Finally, we studied the expression patterns of *cshA* and *cshB* genes coding for two cell surface antigens in *S. gordonii* that increase hydrophobicity of cell surface and mediate interactions with *A. naeslundii* and human fibronectin [45,46]. Expression profiles for *cshA* and *cshB* obtained by qPCR (Figure 5B) are fairly similar to that observed for *sspA* gene and correlation with Illumina sequencing data is notable, at least for *cshA*. Expression patterns of different adhesion proteins analyzed showed a similar pattern, with higher expression in very early stages of oral biofilm. In addition, most of the adhesion genes studied here showed an increased level of expression at

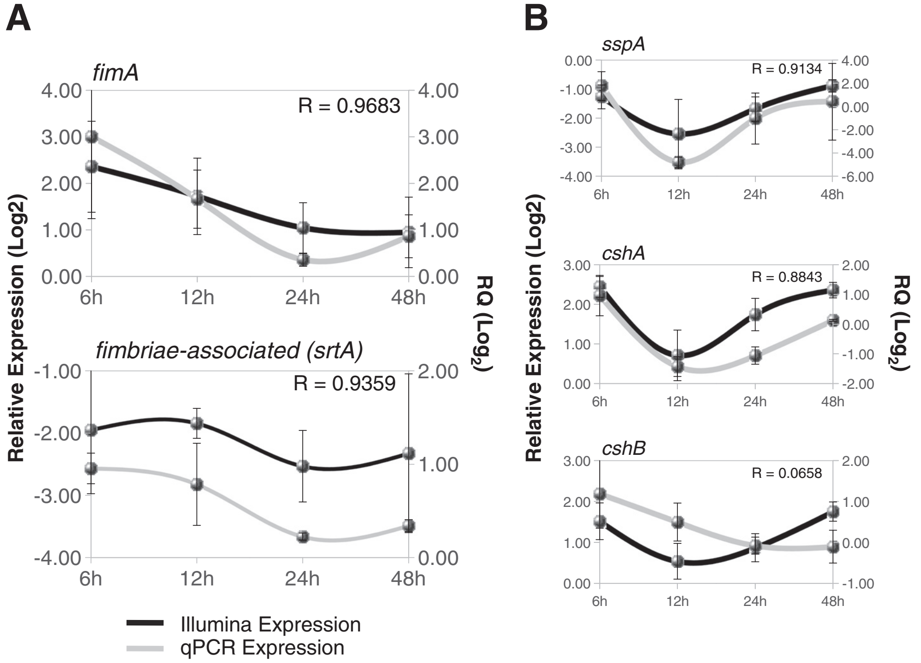


Figure 5 Gene expression comparison between Illumina sequencing and qPCR during biofilm formation. A - Genes associated to Type 2 Fimbriae assembly in *Actinomyces naeslundii* were analyzed and Pearson correlations calculated from expression values obtained by these two approaches. **B** – Adhesion genes from *Streptococcus gordonii* were also analyzed and their expression pattern was compared.

the end point of study (48 h). We hypothesize that such level of expression would reflect the last stage of the biofilm cycle where biofilm detachment occurs, thus releasing bacterial cells to colonize new niches [10,47].

Conclusions

Our study shows for the first time the microbial diversity and gene expression dynamics in the complex oral microbial community *in vivo*. We could follow oral biofilm formation and determine proportions of active microbiota through time, including before and after a carbohydrate-rich meal, when the process of acid production, responsible of enamel demineralization, takes place. We present a large set of correlations among bacterial groups and genera being in agreement with biological and classical interactions reported to be central for biofilm installation and development [8,48,49]. In the functional exploration of genes expressed during human oral biofilm formation, we present a quantitative analysis, further supported by results obtained by qPCR, demonstrating several functional categories of prevalence at different oral biofilm stages. Among them we showed that translation machinery is predominantly expressed in early biofilm stages whereas more specialized genes are required in mature biofilm. Some genes involved in competence, and reported to be involved in quorum sensing response and functionally related to mutacin production and DNA uptake, were over-expressed in late biofilm supporting the intricate level of cell-to-cell interactions in mature biofilm and suggesting strong competition for colonization. More than 70% of the genetic information compiled from this oral metatranscriptome has no functional assignment; therefore, further efforts must be conducted for classification and characterization of genes and their involvement in biofilm development and/or cell-to-cell communication. From an applied point of view, the identification of active bacterial species after food uptake can be considered a first step to narrow down the list of potential etiological agents of dental caries from the large set of micro-organisms found in the metagenome of dental plaque and cavities [3]. The striking homeostasis found in one of the individuals who had never suffered from dental caries, and where virtually no changes were found in the active microbiota before and after a meal, could indicate that the microbiota of some individuals is not affected by food ingestion, potentially reducing the risk of acidic pH and promoting dental health.

Methods

Sample collection and RNA processing

The sampling procedure was approved by the Ethical Committee for Clinical Research from the DGSP-CSISP (Valencian Health Authority, Spain) and all donors signed an informed consent. The oral health status of each individual was evaluated before sampling and following recommendations and nomenclature from the

WHO. Donors were 20-30 years of age, had all 28 teeth present (excluding third molars) had not suffered from any systemic disease and had not taken systemic antimicrobials in the previous 6 months. Dental plaque samples were taken with autoclaved spoon excavators from vestibular and lingual surfaces of teeth excluding a 1 mm region on the edges.

For the biofilm formation experiments, 16 supragingival dental plaque samples were obtained from 4 caries-free volunteers (DMFT = 0 [decayed, Missing, Filling Teeth], OHI = 1 [oral Hygiene Index]. GI = 1 [gingival Index]). The volunteers were subjected to professional teeth ultrasound cleaning. Oral biofilm (supragingival plaque) from all teeth surfaces was pooled and collected from each volunteer at 6, 12, 24, and 48 h of biofilm formation. After every sampling a professional brushing was performed to reset biofilm formation for next sampling. Total RNA was extracted using the MasterPure^{TM} RNA Purification Kit (Epicentre®). Samples were collected and processed for elimination of 5S rRNA and tRNAs through ion exchange chromatography with KCl gradient in Nucleobond AX 20 columns (Macherey-Nagel). Pre- and post-processed RNA were loaded in RNA chip (Agilent Technologies) and analyzed for integrity using Agilent Bioanalyzer 2100 (Agilent Technologies). The first strand of cDNA from processed RNA was synthesized using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). For this aim, two cDNA reactions were prepared for each RNA sample and modifying some manufacturer's instructions to obtain best performance during synthesis. Specifically, each cDNA reaction had 100U of Multiscribe Reverse transcriptase and synthesis was completed at 48°C during 210 min using 5-10 ug of RNA as template. Doubled stranded cDNA (ds-cDNA) was achieved in 100 uL of reaction containing 35U *E.coli* DNA Polymerase I (New England Biolabs), 5U *E. coli* DNA Ligase (New England Biolabs), 5U RNase H (Epicentre®), 300 uM dNTP's, and two reactions of first strand cDNA synthesis. The ds-cDNA synthesis was initiated by incubation during 150 min at 16°C and completed by adding 4.5U of T4 DNA Polymerase (New England Biolabs), and 1X BSA (New England Biolabs) followed by incubation during additional 30 min at 16°C. Purified ds-cDNAs were obtained using High Pure PCR Product Purification Kit (Roche®) and sent to GATC Biotech AG (Konstanz, Germany) for parallel single-end sequencing using HiSeq2000 system (Illumina®).

The five donors for the before/after meal transcriptome were asked not to brush their teeth for 16 hours. Three of them had active caries at the moment of sampling (Decayed Teeth = 3, OHI = 1, GI = 1) and the other two had no history of dental caries (DMFT = 0, OHI = 1, GI = 1). None of the donors had periodontal disease. All donors ingested the same meal, whose nutritional characteristics are indicated in Additional file 4: Table S1. Supragingival dental plaque was obtained from the right maxillary and left mandibular quadrant free teeth surfaces for the sample 30 minutes before eating and from the left maxillary and right mandibular quadrant 30 minutes after food intake, without touching caries lesions, if they were present. Opposite quadrants were sampled because, when analysing PCR-amplified 16S rRNA from cDNA, an equivalence in terms of taxonomic composition was found when sampling opposite mandibular and maxillary quadrants (Additional file 5: Figure S3). The obtained total ds-cDNA (as described above) was purified and enriched in fragments longer than 400 bp using AMPure beads (Agencourt). Those long cDNA fragments were sequenced using 454 GS-FLX technology with titanium chemistry (Roche).

Taxonomic assignment and correlations

Filtering and trimming of original data set was assisted by Galaxy Web Server [16-18], filtering by quality using the sliding window method (window size 25 with a minimum quality in the window of 20), and sequences shorter than 200 bp were removed. For the high-coverage biofilm samples, microbial diversity was established by taxonomic assignment using reads matching 16S rRNA sequences. For this aim we constructed a RDP-based (Release 10, Update 29) database containing almost 10,000 reference sequences of 16S rDNA annotated according to NCBI taxonomy [50]. This reference database was processed to filter out the conserved regions of 16S rDNA genes using Hidden Markov Models [51]. Then, using MegaBlast v2.2.21 algorithm [52] and selecting alignments for 48 nt in length and 100% identity we could assign taxonomy at genus level using only hypervariable regions of 16S rDNA sequences, thus determining predominant microbiota. Heat maps of taxonomic composition were generated using the gplots library of R [53], frequencies were $\log_2$ transformed and clustered with Euclidean distance. In the case of samples before/after a meal, microbial diversity was established using the 16S and 23S rRNA gene. 16S and 23S sequences were binned using META-RNA 1.0 [54]. 16S sequences were assigned using the online RDP assigner [50]. 23S sequences were assigned using the SILVA database and SINA assigner [55,56]. All statistical analyses were conducted on R v2.15. Non-parametric Spearman rank correlation was calculated among top 40 most frequent genera to associate frequency fluctuations during biofilm formation between genera. Then, Spearman's (ρ) coefficient and *t*-test significance was calculated for pairs of genera from all patients, using a Bonferroni correction for multiple comparisons.

Functional analysis

Based on predominant microbiota present at all states of biofilm formation, available complete and WGS genomes were retrieved from the RefSeq and the Human Oral Microbiome databases [57,58]. More than 80 genomes of oral related microorganisms were downloaded and used to build a local database with almost 300,000 coding sequences. More than 800 small predicted ORFs (100-400 nt) were removed, being 98-100% identical to different regions of 16S or 23S rDNAs [59]. The remaining set of ORFs were then submitted to the KEGG Automatic Annotation Server [29] for KEGG Orthology (KO) assignment. Using MegaBlast v2.2.21 algorithm [52] with e-value cutoff 1e-08 and selecting alignments longer than 60% of read with >80% of identity, we assigned KO numbers and PATH categories to the BRITE functional hierarchy [29]. Negative binomial distribution contained in DESeq [60] bioconductor v2.10 package (default parameters) was employed for differential expression analysis. KO over-representation was determined by comparison between early (6-12 h) and late (24-48 h) biofilm samples with q values ≤ 0.05. Counting of reads per gene and genome were normalized against genus frequency and size dataset and then transformed in $\log_2$ for comparison with qPCR expression data.

Quantitative PCR

Primers for qPCR were designed submitting the respective ORF sequences from *S. gordonii* and *A. naeslundii* to the Primer3Plus webserver [61] (Additional file 6: Table S2). Gene amplification was performed using LightCycler® 480 System (Roche), SYBR Green I Master (Roche), and a small aliquot from the respective sample sequenced by Illumina. The Cp values were calculated from three replicates using the LightCycler® 480 SW software v1.5 (Roche). Expression was normalized against 16S rRNA expression from *S. gordonii* and *A. naeslundii*, respectively, and referred to expression seen for every gene at 6h for all patients in average using the ΔΔCt method.

Data access

All sequence data derived from 454 pyrosequencing of cDNA from samples after/before meal experiments, microbial diversity associated to dental quadrants, and Illumina HiSeq2000 sequencing of cDNA from oral biofilm are stored in the MG-RAST server to be publicly available by accessing to the "Oral Metatranscriptome" project, id 935 (http://metagenomics.anl.gov/linkin.cgi?project=935). Sequence data is also available at the European Nucleotide Archive (ENA-EBML) with provisional accession number ERP003984.

Additional files

Additional file 1: Figure S1. Bacterial genera composition according to 23S rDNA. The taxonomic assignation was based on SINA analysis against reference samples from the SILVA database. Bars show the relative frequency for most predominant genera in metatranscriptomic samples obtained before and after a carbohydrate-rich meal.

Additional file 2: Table S3. Shannon Diversity Indexes for samples from the low-coverage approach.

Additional file 3: Figure S2. Bacterial relative abundances between samples obtained before and after a meal. Positive values (expressed as $\log_2$ ratios) are colored in green and indicate a higher abundance of a given genus in the sample before the meal; negative values (also expressed as $\log_2$ ratios), colored in red, indicate a higher abundance in the after-meal sample.

Additional file 4: Table S1. Number of reads analyzed for taxonomy assignment from the low-coverage approach.

Additional file 5: Figure S3. Bacterial diversity analysis of the 24 h human oral biofilm according to dental quadrants. Bacterial composition was estimated by pyrosequencing of the 16S rRNA gene obtained by PCR amplification of cDNA. Diversity at the family taxonomic level (Actinobacteria as Phylum) was determined in biofilm samples coming from four dental quadrants of a unique donor. Pie charts for every quadrant show relative frequency for most predominant bacterial families. Rarefaction curves for each quadrant display a similar diversity for all samples and bacterial composition piecharts indicate slight differences at the frequency of some families like Neisseriaceae being less frequent in upper quadrants.

Additional file 6: Table S2. Sequence information for oligonucleotides used in the qPCR approach.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors participated in the study design. ASS carried out the sampling. ABP and PBF performed the molecular biology studies and high-throughput data analysis. ABP, PBF and AM worked in manuscript preparation. All authors read and approved the final manuscript.

Acknowledgments

Authors thank nutritionist Mercedes Mora-Ruiz for advice in designing the carbohydrate-rich diet for the before/after meal experiment. Authors also thank Anny Camelo for assistance in qPCR procedures and Bernadent Dental Clinic (Valencia, Spain) for facilities during sample collection. This study was funded by projects MICROGEN CSD2009-00006 and 2012-40007 from Spanish MECO.

Received: 22 September 2013 Accepted: 10 April 2014
Published: 27 April 2014

References

1. Keijser BJ, Zaura E, Huse SM, van der Vossen JM, Schuren FH, Montijn RC, ten Cate JM, Crielaard W: **Pyrosequencing analysis of the oral microflora of healthy adults.** *J Dent Res* 2008, **87**(11):1016–1020.
2. Ahn J, Yang L, Paster BJ, Ganly I, Morris L, Pei Z, Hayes RB: **Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison.** *PLoS One* 2011, **6**(7):e22788.
3. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simon-Soro A, Pignatelli M, Mira A: **The oral metagenome in health and disease.** *ISME J* 2012, **6**(1):46–56.
4. Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF, Nelson KE, Gill SR, Fraser-Liggett CM, Relman DA: **Bacterial diversity in the oral cavity of 10 healthy individuals.** *ISME J* 2010, **4**(8):962–974.
5. Marsh PD: **Dental plaque as a biofilm and a microbial community - implications for health and disease.** *BMC Oral Health* 2006, **6**(Suppl 1):S14.
6. Wilson M: **The oral cavity and its indigenous microbiota.** In *Microbial inhabitants of humans.* Edited by Wilson M. New York: Cambridge University Press; 2005:318–374.
7. Jenkinson HF, Lamont RJ: **Oral microbial communities in sickness and in health.** *Trends Microbiol* 2005, **13**(12):589–595.
8. Palmer RJ Jr, Gordon SM, Cisar JO, Kolenbrander PE: **Coaggregation-mediated interactions of streptococci and actinomyces detected in initial human dental plaque.** *J Bacteriol* 2003, **185**(11):3400–3409.

9. Kolenbrander PE, Andersen RN, Blehert DS, Egland PG, Foster JS, Palmer RJ Jr: **Communication among oral bacteria.** *Microbiol Mol Biol Rev* 2002, **66**(3):486–505.
10. Hojo K, Nagaoka S, Ohshima T, Maeda N: **Bacterial interactions in dental biofilm development.** *J Dent Res* 2009, **88**(11):982–990.
11. Dige I, Raarup MK, Nyengaard JR, Kilian M, Nyvad B: **Actinomyces naeslundii in initial dental biofilm formation.** *Microbiology* 2009, **155**(Pt 7):2116–2126.
12. Kolenbrander PE, Andersen RN, Moore LV: **Coaggregation of Fusobacterium nucleatum, Selenomonas flueggei, Selenomonas infelix, Selenomonas noxia, and Selenomonas sputigena with strains from 11 genera of oral bacteria.** *Infect Immun* 1989, **57**(10):3194–3203.
13. Periasamy S, Kolenbrander PE: **Aggregatibacter actinomycetemcomitans builds mutualistic biofilm communities with Fusobacterium nucleatum and Veillonella species in saliva.** *Infect Immun* 2009, **77**(9):3542–3551.
14. Jakubovics NS, Gill SR, Iobst SE, Vickerman MM, Kolenbrander PE: **Regulation of gene expression in a mixed-genus community: stabilized arginine biosynthesis in Streptococcus gordonii by coaggregation with Actinomyces naeslundii.** *J Bacteriol* 2008, **190**(10):3646–3657.
15. Frias-Lopez J, Duran-Pinedo A: **Effect of periodontal pathogens on the metatranscriptome of a healthy multispecies biofilm model.** *J Bacteriol* 2012, **194**(8):2082–2095.
16. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: a web-based genome analysis tool for experimentalists.** *Curr Protoc Mol Biol* 2010, **89:**19.10.1–19.10.21.
17. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res* 2005, **15**(10):1451–1455.
18. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**(8):R86.
19. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C: **Metagenomic biomarker discovery and explanation.** *Genome Biol* 2011, **12**(6):R60.
20. Liu Y, Hu T, Zhang J, Zhou X: **Characterization of the Actinomyces naeslundii ureolysis and its role in bacterial aciduricity and capacity to modulate pH homeostasis.** *Microbiol Res* 2006, **161**(4):304–310.
21. Takahashi N, Nyvad B: **The role of bacteria in the caries process: ecological perspectives.** *J Dent Res* 2011, **90**(3):294–303.
22. Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL Jr: **Microbial complexes in subgingival plaque.** *J Clin Periodontol* 1998, **25**(2):134–144.
23. Chalmers NI, Palmer RJ Jr, Kolenbrander PE, Cisar JO: **Characterization of a Streptococcus sp.-Veillonella sp. community micromanipulated from dental plaque.** *J Bacteriol* 2008, **190**(24):8145–8154.
24. Chung H, Pamp SJ, Hill JA, Surana NK, Edelman SM, Troy EB, Reading NC, Villablanca EJ, Wang S, Mora JR, Umesaki Y, Mathis D, Benoist C, Relman DA, Kasper DL: **Gut immune maturation depends on colonization with a host-specific microbiota.** *Cell* 2012, **149**(7):1578–1593.
25. Dethlefsen L, McFall-Ngai M, Relman DA: **An ecological and evolutionary perspective on human-microbe mutualism and disease.** *Nature* 2007, **449**(7164):811–818.
26. Filoche S, Wong L, Sissons CH: **Oral biofilms: emerging concepts in microbial ecology.** *J Dent Res* 2010, **89**(1):8–18.
27. Rajilic-Stojanovic M, Heilig HG, Tims S, Zoetendal EG, de Vos WM: **Long-term monitoring of the human intestinal microbiota composition.** *Environ Microbiol* 2012, **15:**1146–1159.
28. Abiko Y, Sato T, Mayanagi G, Takahashi N: **Profiling of subgingival plaque biofilm microflora from periodontally healthy subjects and from subjects with periodontitis using quantitative real-time PCR.** *J Periodontal Res* 2010, **45**(3):389–395.
29. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W182–W185.
30. Björk GR, Hagervall TG: **Transfer RNA modification.** In *EcoSal—Escherichia coli and Salmonella: cellular and molecular biology*. Edited by Böck RCI JB, Neidhardt FC, Nyström T, Rudd KE, Squires CL. Washington, D.C: ASM Press; 2005.
31. Grosjean H: **Fine tuning of RNA functions by modification and editing.** In *Topics in Current Genetics*. Edited by Hohmann S. New York: Springer Verlag; 2005.
32. Shin JH, Lee HW, Kim SM, Kim J: **Proteomic analysis of Acinetobacter baumannii in biofilm and planktonic growth mode.** *J Microbiol* 2009, **47**(6):728–735.
33. Kreth J, Merritt J, Shi W, Qi F: **Co-ordinated bacteriocin production and competence development: a possible mechanism for taking up DNA from neighbouring species.** *Mol Microbiol* 2005, **57**(2):392–404.
34. Bhattacharjee MK, Fine DH, Figurski DH: **tfoX (sxy)-dependent transformation of Aggregatibacter (Actinobacillus) actinomycetemcomitans.** *Gene* 2007, **399**(1):53–64.
35. Pollack-Berti A, Wollenberg MS, Ruby EG: **Natural transformation of Vibrio fischeri requires tfoX and tfoY.** *Environ Microbiol* 2010, **12**(8):2302–2311.
36. Human Microbiome Project Consortium: **A framework for human microbiome research.** *Nature* 2012, **486**(7402):215–221.
37. McNab R, Ford SK, El-Sabaeny A, Barbieri B, Cook GS, Lamont RJ: **LuxS-based signaling in Streptococcus gordonii: autoinducer 2 controls carbohydrate metabolism and biofilm formation with Porphyromonas gingivalis.** *J Bacteriol* 2003, **185**(1):274–284.
38. Nobbs AH, Vajna RM, Johnson JR, Zhang Y, Erlandsen SL, Oli MW, Kreth J, Brady LJ, Herzberg MC: **Consequences of a sortase A mutation in Streptococcus gordonii.** *Microbiology* 2007, **153**(Pt 12):4088–4097.
39. Ton-That H, Marraffini LA, Schneewind O: **Protein sorting to the cell wall envelope of Gram-positive bacteria.** *Biochim Biophys Acta* 2004, **1694**(1–3):269–278.
40. Ton-That H, Mazmanian SK, Faull KF, Schneewind O: **Anchoring of surface proteins to the cell wall of Staphylococcus aureus. Sortase catalyzed in vitro transpeptidation reaction using LPXTG peptide and NH(2)-Gly(3) substrates.** *J Biol Chem* 2000, **275**(13):9876–9881.
41. Mishra A, Wu C, Yang J, Cisar JO, Das A, Ton-That H: **The Actinomyces oris type 2 fimbrial shaft FimA mediates co-aggregation with oral streptococci, adherence to red blood cells and biofilm development.** *Mol Microbiol* 2010, **77:**841–854.
42. Jakubovics NS, Kerrigan SW, Nobbs AH, Stromberg N, van Dolleweerd CJ, Cox DM, Kelly CG, Jenkinson HF: **Functions of cell surface-anchored antigen I/II family and Hsa polypeptides in interactions of Streptococcus gordonii with host receptors.** *Infect Immun* 2005, **73**(10):6629–6638.
43. Jakubovics NS, Stromberg N, van Dolleweerd CJ, Kelly CG, Jenkinson HF: **Differential binding specificities of oral streptococcal antigen I/II family adhesins for human or bacterial ligands.** *Mol Microbiol* 2005, **55**(5):1591–1605.
44. Kolenbrander PE, Palmer RJ Jr, Periasamy S, Jakubovics NS: **Oral multispecies biofilm development and the key role of cell-cell distance.** *Nat Rev Microbiol* 2010, **8**(7):471–480.
45. McNab R, Holmes AR, Clarke JM, Tannock GW, Jenkinson HF: **Cell surface polypeptide CshA mediates binding of Streptococcus gordonii to other oral bacteria and to immobilized fibronectin.** *Infect Immun* 1996, **64**(10):4204–4210.
46. McNab R, Jenkinson HF, Loach DM, Tannock GW: **Cell-surface-associated polypeptides CshA and CshB of high molecular mass are colonization determinants in the oral bacterium Streptococcus gordonii.** *Mol Microbiol* 1994, **14**(4):743–754.
47. Stratul S, Didilescu A, Hanganu C, Greabu M, Totan A, Spinu T, Onisei D, Rusu D, Jentsch H, Sculean A: **On the molecular basis of biofilm formation. Oral biofilms and systemic infections.** *TMJ* 2008, 58:118–123.
48. Loozen G, Ozcelik O, Boon N, De Mol A, Schoen C, Quirynen M, Teughels W: **Inter-bacterial correlations in subgingival biofilms: a large-scale survey.** *J Clin Periodontol* 2014, **41**(1):1–10.
49. Ammann TW, Belibasakis GN, Thurnheer T: **Impact of early colonizers on in vitro subgingival biofilm formation.** *PLoS One* 2013, **8**(12):e83090.
50. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM: **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis.** *Nucleic Acids Res* 2009, **37**(Database issue):D141–D145.
51. Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH: **V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences.** *J Microbiol Methods* 2010, 83(2):250–253.
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403–410.
53. Warnes G, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B: **gplots: Various R programming tools for plotting data**. In *The Comprehensive R Archive Network*; 2009.
54. Huang Y, Gilna P, Li W: **Identification of ribosomal RNA genes in metagenomic fragments.** *Bioinformatics* 2009, **25**(10):1338–1340.

55. Pruesse E, Peplies J, Glockner FO: **SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes.** *Bioinformatics* 2012, **28**(14):1823–1829.
56. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.** *Nucleic Acids Res* 2007, **35**(21):7188–7196.
57. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE: **The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information.** *Database (Oxford)* 2010, **2010**:baq013.
58. Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.** *Nucleic Acids Res* 2012, **40**(Database issue):D130–D135.
59. Tripp HJ, Hewson I, Boyarsky S, Stuart JM, Zehr JP: **Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies.** *Nucleic Acids Res* 2011, **39**(20):8792–8802.
60. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**(10):R106.
61. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA: **Primer3Plus, an enhanced web interface to Primer3.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W71–W74.

doi:10.1186/1471-2164-15-311
Cite this article as: Benítez-Páez *et al.*: Microbiota diversity and gene expression dynamics in human oral biofilms. *BMC Genomics* 2014 **15**:311.

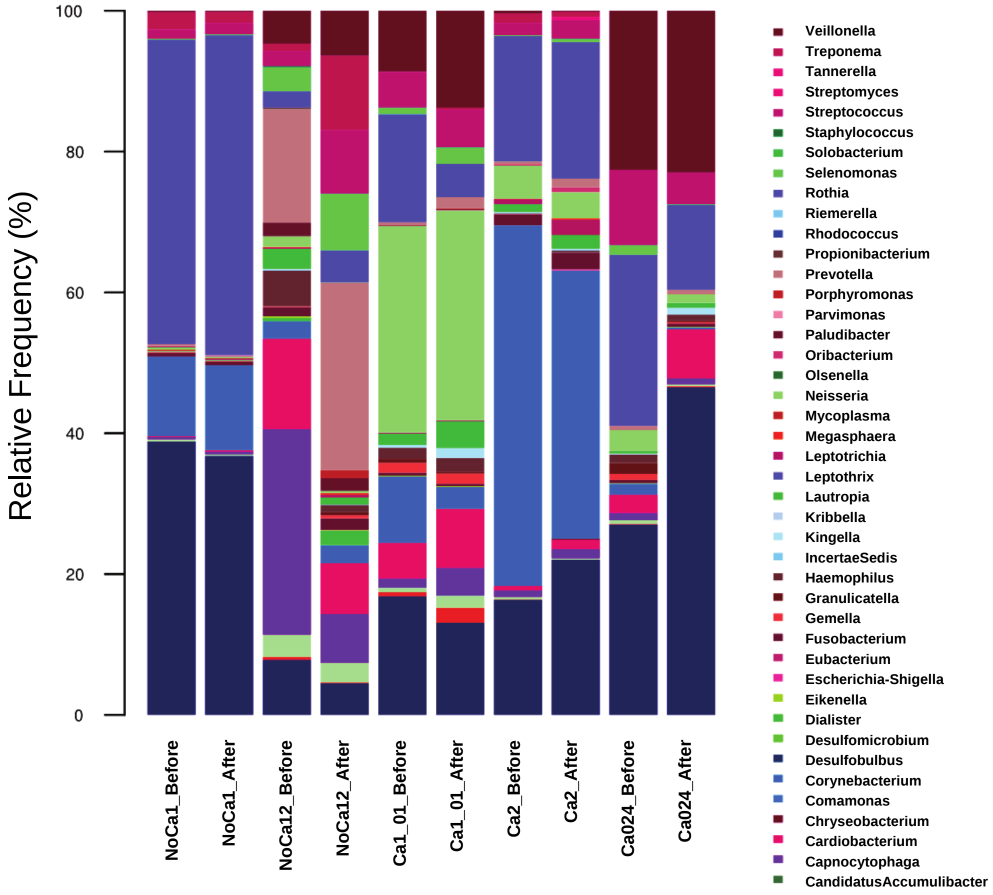


Figure S1. Bacterial genera composition according to 23S rDNA. The taxonomic assignation was based on SINA analysis againstreference samples from the SILVA database. Bars show the relativefrequency for most predominant genera in metatranscriptomic samplesobtained before and after a carbohydrate-rich meal.

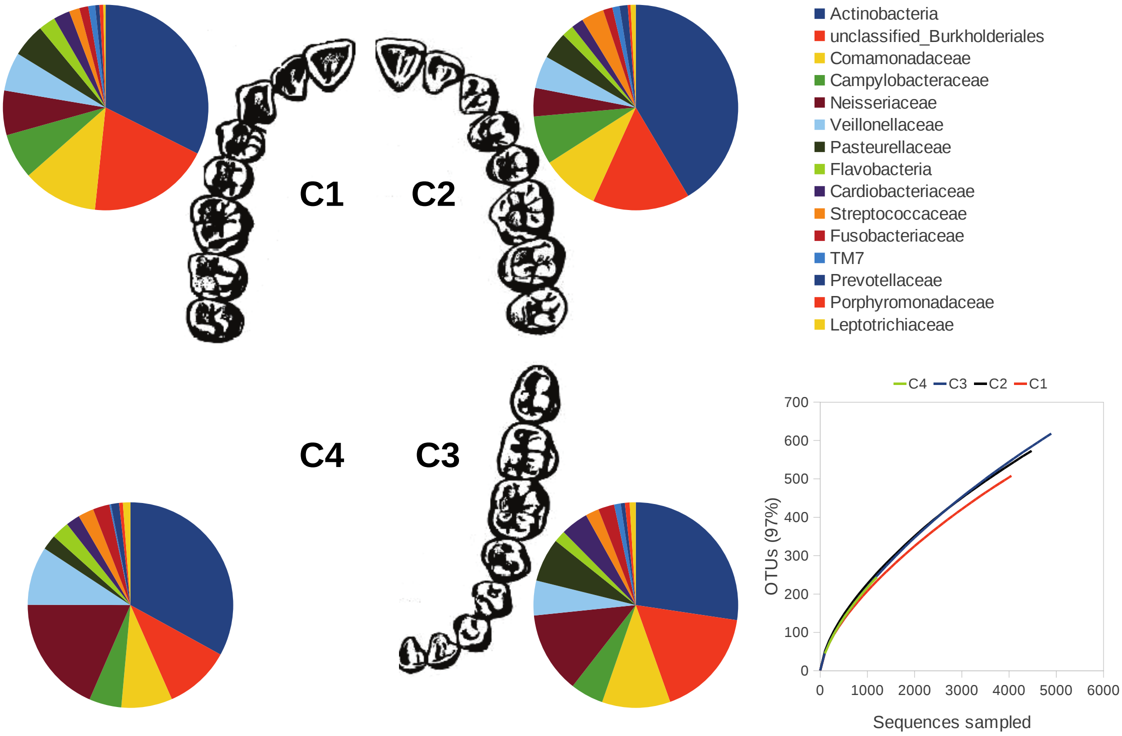


Figure S3. Bacterial diversity analysis of the 24 h human oral biofilm according to dental quadrants. Bacterial composition was estimated by pyrosequencing of the 16S rRNA gene obtained by PCR amplification of cDNA. Diversity at the family taxonomic level (Actinobacteria as Phylum) was determined in biofilm samples coming from four dental quadrants of a unique donor. Pie charts for every quadrant show relative frequency for most predominant bacterial families. Rarefaction curves for each quadrant display a similar diversity for all samples and bacterial composition piecharts indicate slight differences at the frequency of some families like Neisseriaceae being less frequent in upper quadrants.

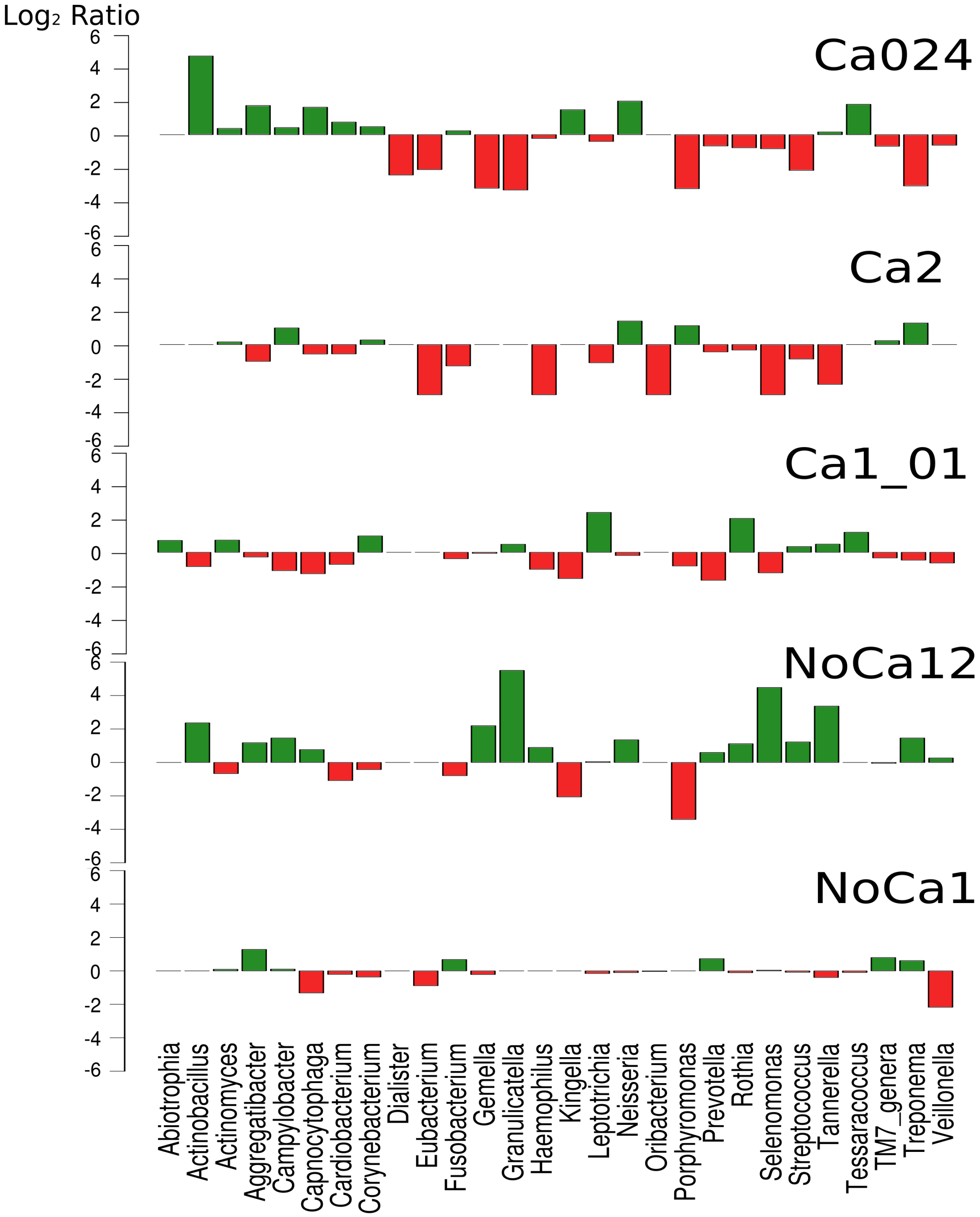


Figure S2. Bacterial relative abundances between samples obtained before and after a meal. Positive values (expressed as log2 ratios) are colored in green and indicate a higher abundance of a given genus in the sample before the meal; negative values (also expressed as log2 ratios), colored in red, indicate a higher abundance in the after-meal sample.

Table S1. Number of reads analyzed for taxonomy assignment from the low-coverage approach.

Sequence number	16S	23S	Other (mRNA)	Total sequences
NoCa1 Before	9769	12706	596	23071
NoCa1 After	23497	34589	1348	59434
NoCa12 Before	6626	11576	112	18314
NoCa12 After	6779	10117	179	17075
Ca1_01 Before	16500	26825	386	43711
Ca1_01 After	6695	9479	181	16355
Ca2 Before	5501	9238	186	14925
Ca2 After	561	1051	10	1622
Ca024 Before	2539	3647	66	6252
Ca024 After	5092	7398	170	12660

Table S2. Sequence information for oligonucleotides used in the qPCR approach.

Gene (species)	Name	Sequence	Gene Position	Tm (C°)	Amplicon Size (bp)
cshA	CshA-F	TTC CAT TCC CAG CTG ATT CGA CT	4361	62.9	100
(*S. gordonii*)	CshA-R	ACC TTA CCG TCT GCG TCC AC	4460	62.5	100
cshB	CshB-F	TCC GGC TAG CTT TGT GGA TGC	2487	63.2	116
(*S.gordonii*)	CshB-R	TCA CTT GGC CGG TAT TTG GAT C	2602	62.2	116
fimA	FimA-F	GAC GGC CAG TGG ATC TAC GA	556	62.5	126
(*A. naeslundii*)	FimA-R	GCT CAC CGG GAA CTT GAT GAG	681	63.2	126
srtA	Fimbriae-F	CGT CGA GGT CTT CGG AGA GG	477	64.6	114
(*A. naeslundii*)	Fimbriae-R	ACC AGG GTG AGC AGG TCC TT	590	62.5	114
sspA	SspA-F	CTT GGT ATG GTG CAG GGG CTA	2261	63.2	103
(*S. gordonii*)	SspA-R	TGA GGC ATT TCC GCT ACA GGC	2363	63.2	103
sspB	SspB-F	CGA CCG GAC ATT GGT TGC TAA AC	2811	64.6	118
(*S. gordonii*)	SspB-R	GCC AGT TGG AAG CGG ATC TAC	2928	63.2	118
16S rRNA	16S_strp-F	GGG GAT AAC TAT TGG AAA CGA TAG C	147	64.1	115
(*S. gordonii*)	16S_strp-R	ACT AGC TAA TAC AAC GCA GGT CCA T	261	64.1	115
16S rRNA	16S_acti-F	GAG TAA CAC GTG AGT AAC CTG CC	99	64.6	150
(*A. naeslundii*)	16S_acti-R	GAT AGG CCG CGA GCC CAT C	248	63.6	150

3.5

“The Human Oral Metaproteome reveals Potential Biomarkers for Caries Disease”

P Belda-Ferre*, J Williamson*, A Simón-Soro, A Artacho, ON Jensen, A Mira

Under second review in PROTEOMICS.

The Human Oral Metaproteome reveals Potential Biomarkers for Caries Disease

Pedro Belda-Ferre[1]*, James Williamson[2]*, Áurea Simon-Soro[1], Alejandro Artacho[1], Ole N. Jensen[2], Alex Mira[1‡]

[1] FISABIO Foundation, Center for Advanced Research in Public Health, Valencia, Spain

[2] Dep. of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark

* Equal contributors

‡ Corresponding autor: Alex Mira, Centro Superior de Investigación en Salud Publica, Avda. Cataluña 21, 46020 Valencia (Spain). Tel. +34 961 925 925, e-mail: mira_ale@gva.es

Abbreviations:

- NGS: Next generation sequencing
- HMP: Human Microbiome Project
- rRNA: Ribosomal RNA
- HILIC: Hydrophilic interaction liquid chromatography
- MTG/MTT/MTP: Metagenome/Metatranscriptome/Metaproteome

Keywords: Biomarkers, dental caries, dental plaque, microbiota, pH buffering.

Abstract

Tooth decay is considered the most prevalent human disease worldwide. We present the first metaproteomic study of the oral biofilm, using different mass spectrometry approaches that have allowed us to quantify individual peptides in healthy and caries-bearing individuals. A total of 7771 bacterial and 853 human proteins were identified in 17 individuals, which provide the first available protein repertoire of human dental plaque. *Actinomyces* and *Coryneybacterium* represent a large proportion of the protein activity followed by *Rothia* and *Streptococcus*. Those four genera account for 60-90% of total diversity. Healthy individuals appeared to have significantly higher amounts of L-lactate dehydrogenase and the arginine deiminase system, both implicated in pH buffering. Other proteins found to be at significantly higher levels in healthy individuals were involved in exopolysaccharide synthesis, iron metabolism and immune response. We applied multivariate analysis in order to find the minimum set of proteins that better allows discrimination of healthy and caries-affected dental plaque samples, detecting seven bacterial and five human protein functions that allow determining the health status of the studied individuals with an estimated specificity and sensitivity over 96%. We propose that future validation of these potential biomarkers in larger sample size studies may serve to develop diagnostic tests of caries risk that could be used in tooth decay prevention.

1. Introduction

The study of microbial populations associated with different niches in the human body, known as the human microbiome, has received large interest in recent years because of its important contribution to health and disease [1] . Beneficial aspects of human associated microbiota include modulation of the immune system, advantageous metabolic properties and prevention against infections, among others [2]. In the oral cavity, bacteria are responsible for some of the most prevalent diseases worldwide, including dental caries and periodontal disease [3]. A description of the microbial species present on the teeth surfaces, the tongue and the gingival tissues was first achieved by laboratory culture and later by molecular methods including the cloning of the 16S gene, DGGE or the use of hybridization chips, showing distinct microbial composition at each of these sites [4,5]. The application of Next Generation Sequencing (NGS) has enabled a complete description of the taxonomic composition of different oral niches [6–9]. This includes a large sequencing effort by the HMP describing seven mouth sites in a large population of healthy individuals which will serve as a reference against which to compare the potential shifts in composition associated to oral diseases [1,10].

A first attempt to relate microbial composition and function to disease has been performed by a metagenomic approach in which dental plaque from healthy individuals and from patients with dental caries was directly pyrosequenced without PCR or cloning steps, giving a picture of the total genetic reservoir of the bacterial populations [11]. However, the total bacterial genetic makeup includes genes from inactive or transient species, as well as large quantities of genes which are not expressed. Meta-transcriptomic approaches have begun to be applied in in vitro oral models [12] and dental plaque samples [13,14], but are limited by the short RNA half-life, the need to amplify and enrich the small mRNA fraction and the fraction of transcripts which are not translated. Recent advances in mass-spectrometry and high-throughput analysis of proteomic data [15] now offer the possibility to study the final output of microbial populations which is directly having an effect on oral ecology and health. In the current manuscript, we present the first metaproteomic study of the oral biofilm, in which total protein composition of dental plaque in healthy and caries-bearing individuals is analyzed by applying different mass spectrometry approaches that have allowed us to detect and quantify individual proteins and compare their levels between disease groups.

The metaproteomic data presented here will serve to give a clearer picture of microbial ecology, dynamics and activity in the oral cavity, as well as the interaction of bacterial activity with the host immune response. From an applied point of view, the identification of molecules at significantly different levels in healthy or caries-bearing individuals may serve as potential biomarkers for health and disease with diagnostic purposes. The metaproteomic strategy presented here has the advantage of looking at the output of the microbial activity at the disease site (i.e. the tooth surface), removing noise introduced by non-active microorganisms, transient or which are present in other parts of the oral cavity. Associated confounding effects are especially relevant when saliva is analyzed, as bacteria from tongue and mucosal surfaces are included in the samples [16], and these organisms are unrelated to the disease. An additional advantage of metaproteomic data is that not only microbial proteins would be detected but also those from the host. These human components present in the dental plaque, including immunoglobulins, antimicrobial peptides or salivary proteins adhering to the tooth, will have a vital role in bacterial modulation and adhesion [17], as well as their metabolic output, and their contribution to disease prevention must not be ignored.

The present study describes the metaproteome of 17 human samples of dental plaque with the aim of characterizing the most common proteins in this oral ecosystem, as well as their inter-individual variability. This exploratory analysis will serve to better understand the biology of oral microbiota and its contribution to dental homeostasis. The studied individuals are clinically well characterized in two distinct groups which are homogenous except for the absence of dental caries history or the presence of a similar number of active caries lesions at the moment of sampling. Thus, we have compared these two groups with the aim of identifying potential biomarkers for health or disease, as well as proteins with a potential therapeutic use against dental caries.

2. Materials and Methods

Sample collection

All donors signed a written informed consent and the sampling procedure was approved by the Ethical Committee for Clinical Research from the DGSP-CSISP (Valencian Health Authority, Spain). Oral health status was evaluated individually following recommendations and nomenclature from the Oral Health Surveys of the World Health Organization [18]. A full dental examination was performed on volunteers aged 19-39 who had not been treated with antibiotics or antifungals in the previous 6 months and had all 28 teeth present. Seventeen donors were selected for sampling on the basis of falling into two distinct groups according to caries health: (a) Individuals who had not previously suffered from dental caries (healthy) and with no active caries at the moment of sampling, with Decayed-Missing-Filled index (DMF)= 0, Oral Health index (OHI)=2, and Gingival Health index (GI)= 1; (b) Individuals with at least 3 active caries (diseased patients), as evaluated by their colour and texture [19], with OHI=1, GI=1 and variable number of previous restorations. None of the donors had periodontal disease. Full clinical information is included in Supplementary Table 1. Dental plaque was collected 24 hours after tooth brushing, pooling plaque from all palatine and lingual teeth surfaces in two different days (biological replicates), using autoclaved spoon excavators as previously described [20]. The pooling allowed to have enough material for MS and to get a representative sample from all teeth, given that microbial composition changes depending on sampling site [20]. Sampling from interproximal and occlusal surfaces were not included because they contain plaque of highly variable formation times [21] and bacterial composition and gene expression changes considerably during biofilm formation [14]. The second reason to sample free surfaces is related to biomarker discovery: the increased salivary flow across lingual and palatine surfaces [22] increases the chance for detecting disease biomarkers of human origin. In addition, the increased salivary flow reduces caries incidence at these sites. Given that potential biomarkers must be informative even at low caries prevalence sites, those identified at free surfaces are expected to be the most robust. Samples were kept in 350 µl of a sterile buffer solution (20mM HEPES, 1mM Benzamidine HCl, 5mM PMSF and 100mM EDTA, pH=7.4) and stored at -80ºC.

Disruption, solubilisation and digestion

Samples were defrosted on ice before usage. Biofilm suspension was pelleted at 5000xg, rinsed with PBS and resuspended in 400 µl of 1% sodium deoxycholate (SDC) and 20 mM triethylammonium bicarbonate (TEAB) pH 8.5. Each sample was sonicated using a tip probe sonicator (analogue cell disruptor, Branson, Germany) for 10 cycles of 15 sec and incubated 10 min at 80ºC. Protein content was quantified using the ProStain™ kit (Active Motif, Rixensart, Belgium), yielding on average 234 µg per sample. Reduction, alkylation and trypsin digestion was done using centrifuge ultrafiltration filters, modifying the method previously described [23]. Briefly, 30 µg of protein per sample were diluted up to 500 µl of 20mM TEAB, 0.5% SDC, 50mM DTT and reduced at 56ºC for 45 min. Alkylation was done by adding 55mM IAA, 20mM TEAB and 0.5% SDC for 20 min in the dark at room temperature. Trypsin digestion was done with 1:100 proportion of trypsin (Novozymes, Denmark) at 37ºC overnight in wet chamber. Peptides were recovered by centrifugation at 14000xg for 20 minutes. SDC was removed using the phase-transfer method [24].

Study Design

The present work consisted of two main parts (Figure 1). First, the discovery study aimed to generate a qualitative assessment of the oral metaproteome using a pool of samples from healthy individuals (white tooth) and another from patients with dental caries (black tooth). After pooling samples, they were prefractionated by HILIC chromatography, followed by LC-MS/MS analysis. The second comparative study was a quantitative assessment of individual samples in order to compare the metaproteomes of healthy and diseased individuals. In this case, each sample was directly analysed by LC-MS/MS, without the prefractionation step, allowing the comparison between healthy and caries-bearing volunteers.

HILIC Chromatography

For the in depth characterization of the dental plaque proteome, pooled samples from all individuals were separated by HILIC chromatography prior to on-line RP LC-MS/MS analysis. An in-house packed HILIC capillary column (300µm I.D x 15 cm, TSK amide 80, (Tosoh Bioscience, Belgium)) was coupled to an Agilent 1200 HPLC system with an integrated fraction collector. The gradient was run at 6µL/min. Solvent A was 90% ACN and 0.1% TFA, Solvent B was 0.1%TFA. The gradient was 5-40% B over 26min, 40-100% B over 2min and 100% for 5 min. One minute fractions were collected during the gradient elution.

LC-MS/MS Data Dependant Acquisition

HILIC fractions were resuspended in 5 or 10µL of 0.1% TFA and injected onto an Easy-LC chromatography system (Thermo Fisher Scientific, Germany). Samples were loaded using intelligent flow control onto home packed trap columns consisting 100µm ID x 2.5cm, 5µm, C18 (Reprosil, Dr. Maisch, Germany). Trapped peptides were then separated on an analytical column of 75µm ID x 15cm, 3µm, C18 (Reprosil, Dr. Maisch, Germany) during the gradient elution. The analytical solvents were A: 0.1% FA and B: 95% ACN, 0.1% FA. Analytical gradients were 0-34% B over 90mins, 34-100%B over 2 mins, 100%B for 10 mins, 100-0%B over 1min. Spectra were acquired in an LTQ-Orbitrap XL instrument (Thermo Fisher Scientific, Germany) operating in data dependant acquisition (DDA) mode. The instrument was set to acquire MS spectra to a resolution of 30000 in the Orbitrap and the top 7 most intense precursor ions per scan were selected for fragmentation in the LTQ at a normalised collision energy (NCE) of 35. Ion trapping times and gain control in the orbitrap and LTQ were set to ensure at least 10 points over each chromatographic peak.

For acquisition of individual patient samples, the chromatographic set up was the same as above. However the instrument MS used was an LTQ-Orbitrap Velos (Thermo Fisher Scientific, Germany). The instrument was set to aquire MS scans to a resolution of 60000 in the Orbitap and the top 20 most intense precursors per scan were selected for CID fragmentation in the LTQ (NCE 35). Gain control and ion trapping times were again set to ensure sufficient points over the typical chromatographic peak for label free quantitation.

Database construction and data analysis

Genomic annotated sequences were downloaded from the Human Oral Microbiome Database [25]. After an initial search of pilot metaproteomic data using this HOMD data, unmatched spectra were searched against non-matching metagenomes obtained from [11]. Metagenomic sequences that were present in the MTP data were taxonomically annotated and available genomes from those taxa in the sequenced genomes of the HMP were added to the HOMD database. In order to reduce redundancy of the database, all protein entries were clustered at 95% similarity using CD-HIT algorithm. A total of 3,805,985 sequences were used in the database.

RAW files were submitted to Proteome Discoverer software (Thermo Fisher Scientific, Bremen, Germany) for spectrum selection, search submission (via MASCOT), quantitation via area calculation of peptide precursors and scoring with Mascot Percolator (Matrix Science, UK). MASCOT search parameters were as follows. Enzyme: Trypsin/P, 2 missed cleavages, fixed modifications: Carbamidomethyl (C), variable modifications: Oxidation (M), peptide tol: 10ppm, MS/MS tol: 0.6Da, Peptide charge: 2, 3 and 4+, instrument: ESI-TRAP.

Statistical analysis is described in detail as Supplementary Methods.

3. Results and Discussion

3.1 Discovery Proteomics

The discovery proteomics dataset was interrogated in order to establish a general picture of the protein content of dental plaque. It should be noted that the amount of proteins were estimated through their peak area. This provides a better estimate of abundance than using only the number of proteins identified in a particular category. It was designed to maximize the detection of proteins, and allowed the identification of 7771 bacterial and 853 human proteins, which provide the first available protein repertoire of human dental plaque (Table 1). The dataset consists primarily of bacterial proteins although some of the most abundant proteins are of human origin (Supplementary Table 2). Twenty-one significant hits were obtained against fungal databases, corresponding to a wide variety of species (Supplementary Table 2). Although the low number of hits suggests that the proportion of fungi in the dental plaque is modest compared to the bacterial component, it has to be borne in mind that the extraction procedure was not optimized for fungal cells.

As expected, a large proportion of the identified human proteins are either known secreted proteins and/or involved in salivary secretion. Furthermore, a number of GO and KEGG terms indicative of proteins involved in response to bacterial pathogens are significantly enriched in our dataset (Supplementary Table 3).

To improve the functional annotation of bacterial proteins HMMER2 [26] was used against the TIGRFAMs (9.0 release) database of prokaryotic functional models [27], successfully annotating 58% of the identified proteins.

From the TIGR roles, as expected, the majority of bacterial proteins identified are involved in central metabolic and housekeeping processes such as energy metabolism and protein synthesis (Supplementary Figure 1). However, our analysis was deep enough to detect low abundance proteins, such as those involved in regulatory functions and signal transduction which each account for approximately 1% of the total bacterial protein content. Within the “Energy Metabolism” role, the dominant sub-role was “Glycolysis/Gluconeogenesis” with only a relatively small abundance of electron transport processes represented.

In addition to functional annotations we have assessed the abundance of the Bacterial Genera present in the metaproteome according to the abundance of proteins uniquely identified to each Genera. As shown in Supplementary Figure 2, a high dominance is seen in the bacterial composition of the samples analyzed. Indeed, the 6 most abundant genera (*Actinomyces*, *Streptococcus*, *Corynebacterium*, *Rothia*, *Leptotrichia* and *Veillonella*) were 1000 times more abundant than the least abundant (*Enterococcus*, *Turicella*, *Massilia*, *Rhodococcus*, *Eikenella*, *Succinatimonas* and *Pasteurella*).

3.2 Different omics to study dental plaque

In two cases, the present metaproteome data could be compared with previous metagenome [11] and metatranscriptome [14] data from the same individuals, from 24 h oral biofilms (Supplementary Figure 3). Although *Actinomyces* and *Corynebacterium* have low abundance in the metagenome (MTG), together they represent a large proportion of the RNA (from the metatranscriptome MTT) and protein synthesis (from the metaproteome, MTP) carried out in the population. The genus *Rothia* is also at high abundance in the proteome compared to genome and transcriptome. The higher abundance of *Actinomyces* in MTT and MTP may reflect that, although previous studies have stated that this genus is an early colonizer due to its ability to adhere to saliva pellicule on the teeth surface [28], its activity is maximal when a mature biofilm is formed (24 h after toothbrushing). This correlates with the establishment of anaerobic conditions at the base layer of the biofilm in close contact with teeth, where *Actinomyces* have mainly been found [29]. This could reflect that, although in a mature biofilm, *Actinomyces* cell counts are still low, its activity is remarkably high. Another plausible explanation is the known bias of second-generation sequencing methods against high GC templates [30], which could artificially reduce the proportion of high GC bacteria such as *Actinomyces*. On the other hand, in the MTP, there is a reduction of *Streptococcus*, *Neisseria*, *Veillonella*, *Capnocytophaga*, *Fusobacterium* and *Aggregatibacter*, probably displaced by the higher *Actinomyces* activity.

The differences in terms of bacterial taxonomic composition between the three methodological approaches could reflect the different biases each technique involves, or true biologically meaningful differences.

Figure 2 shows bacterial composition of individual samples from healthy and diseased volunteers. A high *Actinomyces* and *Corynebacterium* dominance is noted in all samples, as well as a high proportion of *Rothia* and *Streptococcus*. Those four genera account for 60-90% of total diversity. No significant differences in taxonomic composition were found at the genus level between healthy and diseased individuals. The high dominance seen in the MTP contrasts with other DNA-based studies [4,5,11,20], where *Actinomyces* or *Corynebacterium* have never been described as major components of supragingival dental plaque. In previous MTT studies [14] *Actinomyces* and *Corynebacterium* have also been described as major active components in 24-hour biofilms. This could be reflecting an activation of those genera in the mature biofilm that is only detectable when looking at functional molecules, such as RNA and proteins.

3.3 Comparative approach

Bacterial protein composition

Over 2300 bacterial proteins were identified by the quantitative approach considering all samples, with nearly 1000 of them uniquely identified with at least 2 peptides. Proteins found in at least 3 samples from each group were initially considered for statistical analysis. At a first view, functional composition is similar in all individuals when using the TIGR annotation nomenclature (Supplementary Figure 4). However, clear differences arise when looking at the deepest classification level of the TIGRFAM system, as samples clustered by caries status (Figure 3A). Differences between healthy and caries bearing individuals can be grouped in three main clusters of functions (Figure 3A). The first block includes functions that do not have a differential distribution pattern among the two groups. The second block is composed by functions over-represented in healthy individuals and the third one contains functions which are more abundant in diseased patients. This finding suggests that there is a differential pattern of protein abundance between healthy and caries bearing individuals, suggesting the presence of functions potentially protecting from tooth decay, as well as others potentially providing disease biomarkers that could be used to detect caries-prone patients before appearance of the disease.

Caries is caused by the acidic demineralization of teeth enamel. Acids produced as a consequence of the fermentation of sugars are one of the sources of enamel degrading acids [2,3,8]. In this study, healthy individuals appeared to have significantly higher amounts of L-lactate dehydrogenase (LDH) (p-val= 1.23x10-4). In *A. actinomycetemcomitans* an unusual LDH has been found that is expressed when grown with lactate [31]. Removal of lactate from media could reverse the drop in pH caused by glucose fermentation, increasing pyruvate levels. However, conventional L- lactate dehydrogenases from oral streptococci reduce NAD and generate lactate, which would decrease the pH. Given that the taxonomic assignment of the LDH identified in the current work could not be established, the reason for its higher frequency in caries-free individuals is unclear. A higher abundance of sugar transporters in caries-bearing individuals was observed, including PTS systems and ABC transporters such as those of the CPR0540 family (TIGR 03850), involved in the uptake of disaccharides (p=0.011), suggesting higher sugar intake rates. Another significant difference is the increased abundance in the caries group of N-acetylglucosamine-6-phosphate deacetylase (p-val=5.93x10-3), an enzyme that degrades this amino-sugar present in a wide variety of macromolecules coming from saliva and gingival crevicular fluid (GCF) [32].

Several proteins involved in exopolysaccharide synthesis were found to be at significantly higher levels in healthy individuals, including the glucose-1-phosphate thymidylyltransferase (p=0.0006) and a 1,4-alpha-glucan branching enzyme (p=0.0005), both of which have been found to be important for biofilm formation [33–35].

The ornithine carbamoyltransferase was over-represented in healthy volunteers. This enzyme is one of the three proteins involved in the arginine deiminase system. This system has been proposed to be an acid-protection system in dental biofilm for bacteria sensitive to very low pH and also to be the responsible for the pH increase to neutral values after meal ingestion [36,38]. In fact, clinical trials have shown that providing arginine in oral care products reduce caries formation [37,39].

Human protein composition

Among the proteins found in the supragingival dental plaque, 127 human proteins were detected and quantified. Apart from housekeeping functions like ribosomal proteins or histones, the most common categories corresponded to secreted proteins (complement system, antibacterial peptides, mucins, immunoglobulins and microglobulin), to others involved in the development of keratinized epithelia (hornerin, keratin, galectin-7, suprabasin and filaggrin) and iron metabolism (haptoglobin, ceruloplasmin, lipocalin-2, hemoglobin). Twenty-nine proteins were found to be differentially expressed (p-value < 0.05) in healthy and diseased individuals (Figure 3B). Some of the most relevant are discussed below and a full list can be seen in Supplementary Table 4.

Keratinized epithelium in supragingival dental plaque

Although samples were collected only from teeth surfaces, leaving a 1mm edge from gums, high amounts of epithelial related proteins (keratin, cell junctions, desmosomes) were found. Oral epithelia are continuously shed from the superficial layer, mainly composed by differentiated keratinocytes. This epithelial desquamation removes mucosal biofilm, which can eventually be attached to supragingival dental plaque. This can be the origin of microbial colonizers in supragingival dental plaque. Further studies should determine whether epithelial cells may serve as vehicles for microbial transportation into the biofilm [40] and also serve as nutrient source for the growing biofilm.

Immune host response proteins

Unlike other oral surfaces, which are protected by the epithelial innate immune system, supragingival dental plaque is only subjected to the mucosal adaptive immune system. In addition, secreted exoplysaccharides can impede the access of immunitary molecules such as secretory antibodies to the inside of the biofilm, impeding a proper immune surveillance. However, a large repertoire of immune proteins was detected in the biofilm metaproteome. The main origins of those components of the immune system are salivary fluids and GCF. Immunoglobulins are mainly secreted through the salivary glands, which produce high amounts of IgA [41]. High amounts of heavy (α and μ types) and light chains (λ), J chain and the secreted portion of the polymeric Ig receptor (pIgr) are found on the studied samples, which suggests that Ig are able to reach the dental plaque, exerting a selective pressure on biofilm formation [42]. Alpha and µ heavy chain were over-expressed in diseased volunteers, whereas the pIgR was more abundant in healthy individuals. Healthy individuals have been shown to contain higher amounts of secreted Ig in saliva [43]. Cellular immunity was also detected in our samples, mainly in form of proteins typically present in leucocytes' granules or salivary secretions (azurocidin, cysteine-rich secretory protein 3, cathelicidin, lysozime, neutrophil defensin 1, cathepsin G, coronin 1A and BPI) or its surface (Integrin alpha-M), proteins involved in phagocytosis, motility and cytoskeleton modification (profilin-1, hASC-3, RAC-2 and CEACAM1) and major histocompatibility system (alfa- and beta-2 macroglobulin). This suggests that leukocytes can have an active role in dental plaque, suggesting a migration of immune cells, probably from the gingival crevice. However, some of the proteins listed can also be secreted by the salivary glands (e.g. lysozyme, BPI and defensins). In healthy volunteers there is a statistically significant over-expression of azurocidin, complement component 3 (C3), pIgr, RAC-2 and hASC-3, whereas in diseased patients only Ig alpha and mu chains were over-represented, pointing to a wider variety of defence weapons in healthy individuals.

Fe metabolism

Iron is typically a limiting factor in bacterial growth [44]. Reduced availability of this nutrient in the oral cavity makes capture of iron a continuous battle between bacterial and human cells. Both secrete siderophores to capture iron molecules from the environment. Human Fe-chelating proteins found in the dental plaque metaproteome comprise hemoglobin beta-subunit, neutrophil gelatinase-associated lipocalin (NGAL), haptoglobin and ceruloplasmin. NGAL was the only one over-represented in healthy individuals. NGAL has a bacteriostatic effect by chelating bacterial siderophores [45]. This protein is secreted together with MMP-9 (also over-represented in healthy volunteers) linked by a disulfide-bond. Ceruloplasmin and haptoglobin were significantly over-represented (p-val=4.15x10-2 and 3.39x10-2) respectively) in caries-bearing individuals. Ceruloplasmin has a ferroxidase activity, which in blood is used to facilitate ferric iron (Fe+3) carriage by lactoferrin [46], and has also been described to reduce recruitment of polymorphonucleted neutrophiles and superoxide levels. This could end in a reduction of the cellular immune response and making soluble iron (Fe+3) more available. Haptoglobin is a protein binding free hemoglobin (Hb), preventing oxidative tissue damage and exerting an antimicrobial effect by reducing Hb availability.

Proteases and protease inhibitors

Salivary secretions contain proteins with proteolityc action, as well as their counterpart inhibitors to regulate their activity. In the present study we detected proteases (MMP-9, antileukoproteinase, cathepsin) and proteases inhibitors (cystatins, transgelin, alpha-2-macroglobulin). MMP-9 is a proteinase that degrades gelatin and collagen. It has been implicated in caries progression through dentinal tissue [47,48], but in this study, MMP-9 appears over-represented in caries-free individuals (p-value=9.76x10-6)). Inside dentinal tissue MMP-9 and cathepsin have been described to hydrolise collagen and gelatine. Cathepsins are released by lysosomes, degrading proteins for antigen presentation, degrading host's extracellular matrix, or digesting engulfed pathogens in macrophages, among others. Proteinase inhibitors may play also an important role in the regulation of this proteolytic activity leaking from macrophages. In addition, collagen has been found to be a binding site for oral pathogenic microorganisms, facilitating their adhesion to the biofilm [49]. Thus, differences in the concentration of that protein may be related to the tendency of developing a pathogenic biofilm.

3.4 Network analysis

One of the advantages of a metaproteomic approach is the possibility to analyse both microbial and human proteins together and recent advances in network analysis now allow the study of these potential interactions [50]. We have applied Bayesian Network Analysis to individual data from the caries and non-caries individuals, identifying a pattern of correlations between all the identified proteins that best explain data variability (Supplementary Figure 5). A total of 186 positive and negative interactions were found with correlation values over 0.85 among healthy individuals, whereas the caries group contained 165 correlations. Interestingly, several interactions in the network appear to be different between healthy and diseased individuals. For instance, cell division protein FtsZ appears to be differentially connected in the healthy and diseased networks (Figure 4). In healthy volunteers this protein, implicated in the Z-ring formation prior to cell division, shows negative correlations with human proteins cystatin-SN and neutrophil defensin 1. Twe propose that those two proteins may play a role in inhibiting bacterial growth, and thus affecting the presence of this division protein when they are expressed in high levels. Positive correlations were also found with other human and bacterial proteins. On the other hand, the diseased samples' network of FtsZ displays positive correlations on the fructose-biphosphate aldolase A (similar to the healthy network) and the CD59 glycoprotein, a protein that is able to inhibit cell lysis by the complement membrane attack system. Other differences in human-bacterial interactions can be observed in Supplementary Figure 5

3.5 Biomarker discovery

Previous DNA-based studies have detected differences in dental plaque samples between caries-affected and healthy individuals, in terms of both presence of multiple oral microorganisms [51–53] and functional composition [11]. Another study combined microbial DNA-based approaches in dental plaque with proteomic analysis of stimulated saliva samples to predict caries risk [54]. Previous work from our group found significant microbial differences between dental plaque and saliva [16], making the latter not the ideal sample for biomarker discovery given that differences between saliva and plaque composition may dilute true biologically meaningful caries biomarkers. Here we present a metaproteomic approach to find potential biomarkers to diagnose healthy or diseased tendency in dental plaque. A list of 53 bacterial and 29 human differentially abundant proteins was obtained by comparing the two groups through t-test statistics (Supplementary Table 4). Principal Component Analyses based on those selected proteins were very efficient at discriminating between healthy and diseased individuals (Figure 5). However, this large number of proteins would hamper the development of feasible diagnostic kits. Thus, we applied multivariate analysis implemented in GALGO R-package [55] in order to find the minimum set of proteins that allows discrimination of healthy and caries-affected dental plaque samples. This approach detected six TIGR bacterial roles and four human proteins that allow determination of the health status of dental plaque samples, with an estimated specificity and sensitivity over 96% (Figure 5). Those six roles appear to be biologically relevant, as they reflect different conditions associated to health and disease. For instance, the glucose PTS system, the copper-containing nitrite reductase and the stress response protein CspD were selected as biomarkers, being more abundant in diseased samples. Healthy associated biomarkers included L-lactate dehydrogenase, succinate-CoA ligase and succinate dehydrogenase. When applying GALGO to human proteins, 4 proteins were selected as biomarkers, namely cystatin-A, transglutaminase CRAa, hemoglobin beta subunit and protein S100-A9.

4. Concluding remarks

Current diagnostic tests for caries risk are based on genetic analysis, on microbial components or on individual caries-associated salivary compounds. However, caries risk diagnostic approaches based on genetic polymorphisms [56] are difficult to replicate on different human populations and are subject to a high degree of false positives due to the counteracting effect of compensatory mutations and epigenetic factors. Commercially available microbial tests based on single species like mutans streptococci or lactobacilli counts are extremely limited due to the polymicrobial nature of dental caries [11,20]. We believe that the measurements of multiple components and the interactions among those salivary and/or plaque constituents and functions will be more informative and sensitive than individual-compound tests. For instance, a patient may present normal values for a given compound or metabolic reaction but have out-of-range values for another. Here we propose a set of six bacterial and four human proteins that are able to differentiate healthy and caries-bearing individuals. Further large-scale longitudinal studies should validate their predictive value as indicators of future disease onset. We propose that diagnostic tests of caries risk open the possibility to design oral care products specifically adapted to the test outcome (personalized medicine) that could be used as a tooth decay preventive treatment.

5. Acknowledgements

This study was funded by projects MICROGEN CSD2009-00006 and 2012-40007 from the Spanish Ministry of Economy and Competitiveness; by Villum Fonden, Center for Analytical Biosciences (61310-0110-13761), and by Center for Epigenetics (62310-0112-023541).

6. Conflict of interest

The authors have declared no conflict of interest

7. References

[1] HMP, A framework for human microbiome research. *Nature* 2012, 486, 215–21.

[2] Wilson, M., in:, *Microb. Inhabitants Humans*, Cambridge University Press, 2005, pp. 318–374.

[3] Wade, W.G., The oral microbiome in health and disease. *Pharmacol. Res.* 2013, 69, 137–43.

[4] Aas, J.A., Paster, B.J., Stokes, L.N., Olsen, I., Dewhirst, F.E., Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* 2005, 43, 5721–32.

[5] Bik, E.M., Long, C.D., Armitage, G.C., Loomer, P., et al., Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J.* 2010, 4, 962–74.

[6] Zaura, E., Keijser, B.J.F., Huse, S.M., Crielaard, W., Defining the healthy “core microbiome” of oral microbial communities. *BMC Microbiol.* 2009, 9, 259.

[7] Lazarevic, V., Whiteson, K., Hernandez, D., François, P., Schrenzel, J., Study of inter- and intra-individual variations in the salivary microbiota. *BMC Genomics* 2010, 11, 523.

[8] Nyvad, B., Crielaard, W., Mira, a, Takahashi, N., Beighton, D., Dental caries from a molecular microbiological perspective. *Caries Res.* 2013, 47, 89–102.

[9] Keijser, B.J.F., Zaura, E., Huse, S.M., van der Vossen, J.M.B.M., et al., Pyrosequencing analysis of the oral microflora of healthy adults. *J. Dent. Res.* 2008, 87, 1016–20.

[10] Segata, N., Haake, S.K., Mannon, P., Lemon, K.P., et al., Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* 2012, 13, R42.

[11] Belda-Ferre, P., Alcaraz, L.D., Cabrera-Rubio, R., Romero, H., et al., The oral metagenome in health and disease. *ISME J.* 2012, 6, 46–56.

[12] Frias-Lopez, J., Duran-Pinedo, A., Effect of periodontal pathogens on the metatranscriptome of a healthy multispecies biofilm model. *J. Bacteriol.* 2012, 194, 2082–95.

[13] Duran-Pinedo, A.E., Chen, T., Teles, R., Starr, J.R., et al., Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. *ISME J.* 2014, 8, 1659–1672.

[14] Benítez-Páez, A., Belda-Ferre, P., Simón-Soro, A., Mira, A., Alfonso Benítez-Páez, Pedro Belda-Ferre, Áurea Simón-Soro, A.M., Microbiota diversity and gene expression dynamics in human oral biofilms. *BMC Genomics* 2014, 15, 311.

[15] Kolmeder, C.A., de Vos, W.M., Metaproteomics of our microbiome - developing insight in function and activity in man and model systems. *J. Proteomics* 2014, 97, 3–16.

[16] Simón-Soro, A., Tomás, I., Cabrera-Rubio, R., Catalan, M.D., et al., Microbial geography of the oral cavity. *J. Dent. Res.* 2013, 92, 616–21.

[17] Leone, C.W., Oppenheim, F.G., Physical and chemical aspects of saliva as indicators of risk for dental caries in humans. *J. Dent. Educ.* 2001, 65, 1054–62.

[18] World Health Organization, *Oral Health Surveys, Basic Methods*, Geneva 1997.

[19] Nyvad, B., Machiulskiene, V., Baelum, V., Construct and Predictive Validity of Clinical Caries Diagnostic Criteria Assessing Lesion Activity. *J. Dent. Res.* 2003, 82, 117–122.

[20] Simón-Soro, Á., Belda-Ferre, P., Cabrera-Rubio, R., Alcaraz, L.D., Mira, a., A Tissue-Dependent Hypothesis of Dental Caries. *Caries Res.* 2013, 47, 591–600.

[21] Marsh, P., Martin, M., Lewis, M., Williams, D., in:, *Oral Microbiol. 5th Ed.*, Churchill Livingstone, Edimburg, UK 2009, pp. 74–102.

[22] Lecomte, P., Dawes, C., The influence of salivary flow rate on diffusion of potassium chloride from artificial plaque at different sites in the mouth. *J. Dent. Res.* 1987, 66, 1614–8.

[23] Manza, L.L., Stamer, S.L., Ham, A.-J.L., Codreanu, S.G., Liebler, D.C., Sample preparation and digestion for proteomic analyses using spin filters. *Proteomics* 2005, 5, 1742–5.

[24] Masuda, T., Tomita, M., Ishihama, Y., Phase Transfer Surfactant-Aided Trypsin Digestion for Membrane Proteome Analysis research articles 2008, 731–740.

[25] Chen, T., Yu, W.-H., Izard, J., Baranova, O. V, et al., The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford).* 2010, 2010, baq013.

[26] Finn, R.D., Clements, J., Eddy, S.R., HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 2011, 39, W29–37.

[27] Haft, D.H., Selengut, J.D., White, O., The TIGRFAMs database of protein families. *Nucleic Acids Res.* 2003, 31, 371–373.

[28] Li, J., Helmerhorst, E.J., Leone, C.W., Troxler, R.F., et al., Identification of early microbial colonizers in human dental biofilm. *J. Appl. Microbiol.* 2004, 97, 1311–8.

[29] Zijnge, V., Van Leeuwen, M.B.M., Degener, J.E., Abbas, F., et al., Oral biofilm architecture on natural teeth. *PLoS One* 2010, 5, e9321.

[30] Schwientek, P., Szczepanowski, R., Rückert, C., Stoye, J., Pühler, A., Sequencing of high G+C microbial genomes using the ultrafast pyrosequencing technology. *J. Biotechnol.* 2011, 155, 68–77.

[31] Brown, S. a, Whiteley, M., Characterization of the L-lactate dehydrogenase from Aggregatibacter actinomycetemcomitans. *PLoS One* 2009, 4, e7864.

[32] Moye, Z.D., Burne, R.A., Zeng, L., Uptake and Metabolism of N-Acetylglucosamine and Glucosamine by Streptococcus mutans. *Appl. Environ. Microbiol.* 2014, 80, 5053–5067.

[33] Seibold, G.M., Breitinger, K.J., Kempkes, R., Both, L., et al., The glgB-encoded glycogen branching enzyme is essential for glycogen accumulation in Corynebacterium glutamicum. *Microbiology* 2011, 157, 3243–51.

[34] Huang, T.P., Somers, E.B., Wong, A.C.L., Differential biofilm formation and motility associated with lipopolysaccharide/exopolysaccharide-coupled biosynthetic genes in Stenotrophomonas maltophilia. *J. Bacteriol.* 2006, 188, 3116–3120.

[35] Paes Leme, A.F., Koo, H., Bellato, C.M., Bedi, G., Cury, J.A., The role of sucrose in cariogenic dental biofilm formation--new insight. *J. Dent. Res.* 2006, 85, 878–87.

[36] Casiano-colon, A., Marquis, R.E., Role of the arginine deiminase system in protecting oral bacteria and an enzymatic basis for acid tolerance. *Appl. Environ. Microbiol.* 1988, 54, 1318–1324.

[37] Acevedo, A.M., Machado, C., Rivera, L.E., Wolff, M., Kleinberg, I., The inhibitory effect of an arginine bicarbonate/calcium carbonate CaviStat-containing dentifrice on the development of dental caries in Venezuelan school children. *J. Clin. Dent.* 2005, 16, 63–70.

[38] Liu, Y.-L., Nascimento, M., Burne, R. a, Progress toward understanding the contribution of alkali generation in dental biofilms to inhibition of dental caries. *Int. J. Oral Sci.* 2012, 4, 135–40.

[39] Santarpia, R.P., Lavender, S., Gittins, E., Vandeven, M., et al., A 12-week clinical study assessing the clinical effects on plaque metabolism of a dentifrice containing 1.5% arginine, an insoluble calcium compound and 1,450 ppm fluoride. *Am. J. Dent.* 2014, 27, 100–5.

[40] Brecx, M., Rönström, A., Theilade, J., Attström, R., Early formation of dental plaque on plastic films. 2. Electron microscopic observations. *J. Periodontal Res.* 1981, 16, 213–27.

[41] Van Nieuw Amerongen, a, Bolscher, J.G.M., Veerman, E.C.I., Salivary proteins: protective and diagnostic value in cariology? *Caries Res.* 2004, 38, 247–53.

[42] Mestecky, J., Russell, M.W., Elson, C.O., Intestinal IgA: novel views on its function in

the defence of the largest mucosal surface. *Gut* 1999, 44, 2–5.

[43] Fidalgo, T.K. da S., Freitas-Fernandes, L.B., Ammari, M., Mattos, C.T., et al., The relationship between unspecific s-IgA and dental caries: A systematic review and meta-analysis. *J. Dent.* 2014, 42, 1372–1381.

[44] Wandersman, C., Delepelaire, P., Bacterial iron sources: from siderophores to hemophores. *Annu. Rev. Microbiol.* 2004, 58, 611–47.

[45] Goetz, D.H., Holmes, M. a, Borregaard, N., Bluhm, M.E., et al., The neutrophil lipocalin NGAL is a bacteriostatic agent that interferes with siderophore-mediated iron acquisition. *Mol. Cell* 2002, 10, 1033–43.

[46] Segelmark, M., Persson, B., Hellmark, T., Wieslander, J., Binding and inhibition of myeloperoxidase (MPO): a major function of ceruloplasmin? *Clin. Exp. Immunol.* 1997, 108, 167–74.

[47] Vidal, C., Tjäderhane, L., Scaffa, P., Tersariol, I., et al., Abundance of MMPs and cysteine cathepsins in caries-affected dentin. *J. Dent. Res.* 2014, 269–274.

[48] Tjäderhane, L., Larjava, H., Sorsa, T., Uitto, V.J., et al., The activation and function of host matrix metalloproteinases in dentin matrix breakdown in caries lesions. *J. Dent. Res.* 1998, 77, 1622–1629.

[49] Jenkinson, H.F., Lamont, R.J., Oral microbial communities in sickness and in health. *Trends Microbiol.* 2005, 13, 589–95.

[50] Su, C., Andrew, A., Karagas, M.R., Borsuk, M.E., Using Bayesian networks to discover relations between genes, environment, and disease. *BioData Min.* 2013, 6, 6.

[51] Becker, M.R., Paster, B.J., Leys, E.J., Moeschberger, M.L., et al., Molecular analysis of bacterial species associated with childhood caries. *J. Clin. Microbiol.* 2002, 40, 1001–9.

[52] Corby, P.M., Lyons-Weiler, J., Bretz, W.A., Hart, T.C., et al., Microbial risk indicators of early childhood caries. *J. Clin. Microbiol.* 2005, 43, 5753–5759.

[53] Li, Y., Ge, Y., Saxena, D., Caufield, P.W., Genetic profiling of the oral microbiota associated with severe early-childhood caries. *J. Clin. Microbiol.* 2007, 45, 81–7.

[54] Hart, T.C., Corby, P.M., Hauskrecht, M., Hee Ryu, O., et al., Identification of microbial and proteomic biomarkers in early childhood caries. *Int. J. Dent.* 2011, 2011, 196721.

[55] Trevino, V., Falciani, F., GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics* 2006, 22, 1154–6.

[56] Shaffer, J.R., Feingold, E., Wang, X., Lee, M., et al., GWAS of dental caries patterns in the permanent dentition. *J. Dent. Res.* 2013, 92, 38–44.

Table 1. Identified proteins groups in the human oral biofilm[1].

	Discovery approach	Comparative approach
	Bacterial	
Proteins (% 2 unique peptides)	7771 (48%)	2137 (46%)
Distinct proteins* (% 2 unique peptides)	1482 (48%)	363 (31%)
Genera (% 2 unique peptides)	134 (50%)	107 (34%)
	Human	
Proteins (% 2 unique peptides)	853 (61%)	228 (61%)
Distinct proteins* (% 2 unique peptides)	397 (56%)	35 (55%)

[1] Numbers between brackets indicate the proportion of the indicated number of proteins that were identified by two unique hits to the same peptide (high-stringency criterion).

* Refers to a protein group containing only 1 protein entry

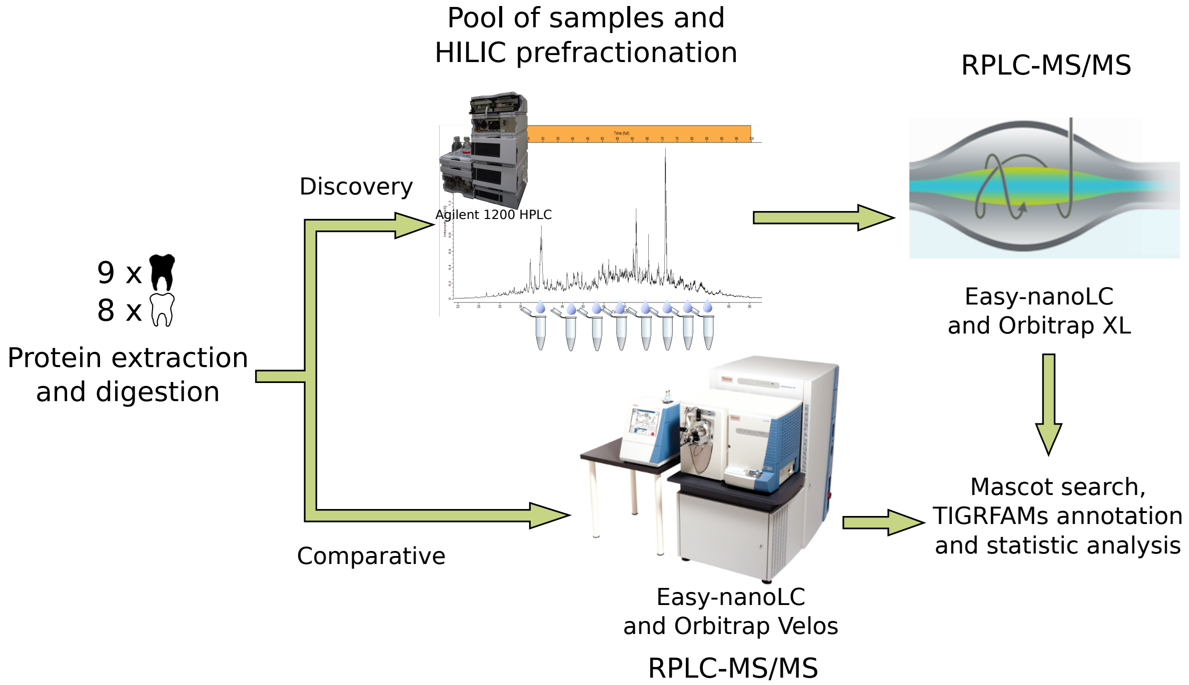


Figure 1. Workflow showing the two approaches of this study. In the Discovery approach, all samples were mixed in equal amounts after protein extraction and trypsin digestion, followed by HILIC prefractionation and subsequent RPLC-MS/MS analysis of each fraction. In the Comparative approach, all samples were analyzed independently by RPLC-MS/MS without prefractionation. Mascot server was used to identify peptides in the sample and TIGRFam annotation was performed upon the bacterial proteins found. White and black tooth icons represent samples from caries-free and caries-bearing individuals, respectively.

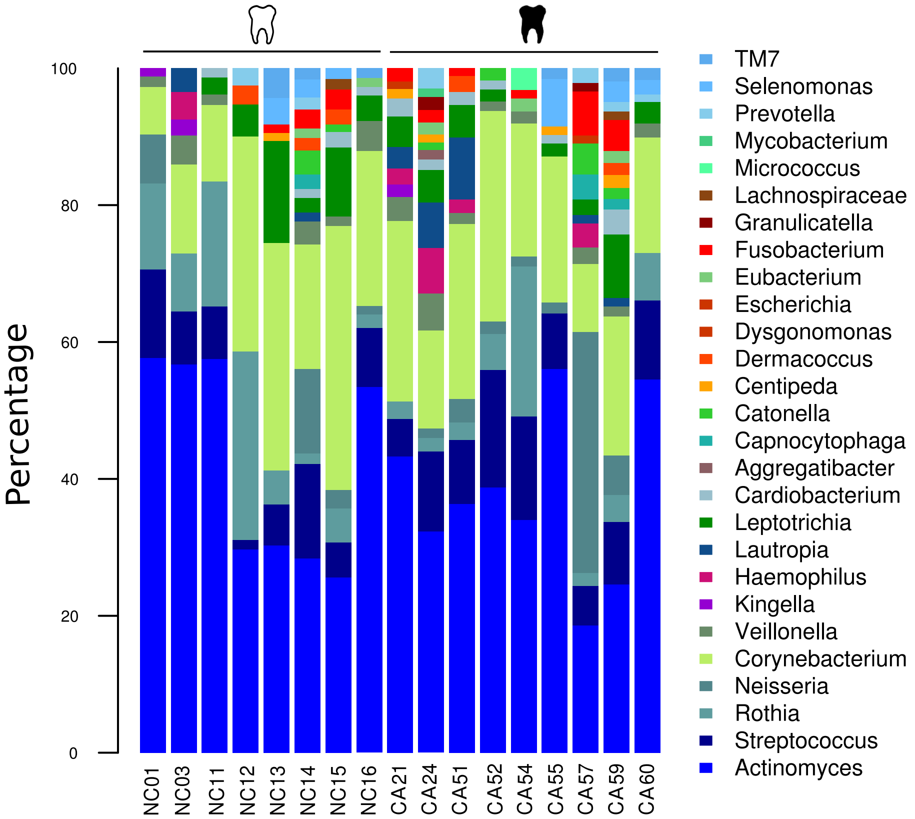


Figure 2. Taxonomic composition of supragingival dental plaque samples according to their metaproteomic profile in healthy (white tooth symbol) and diseased (black tooth symbol).

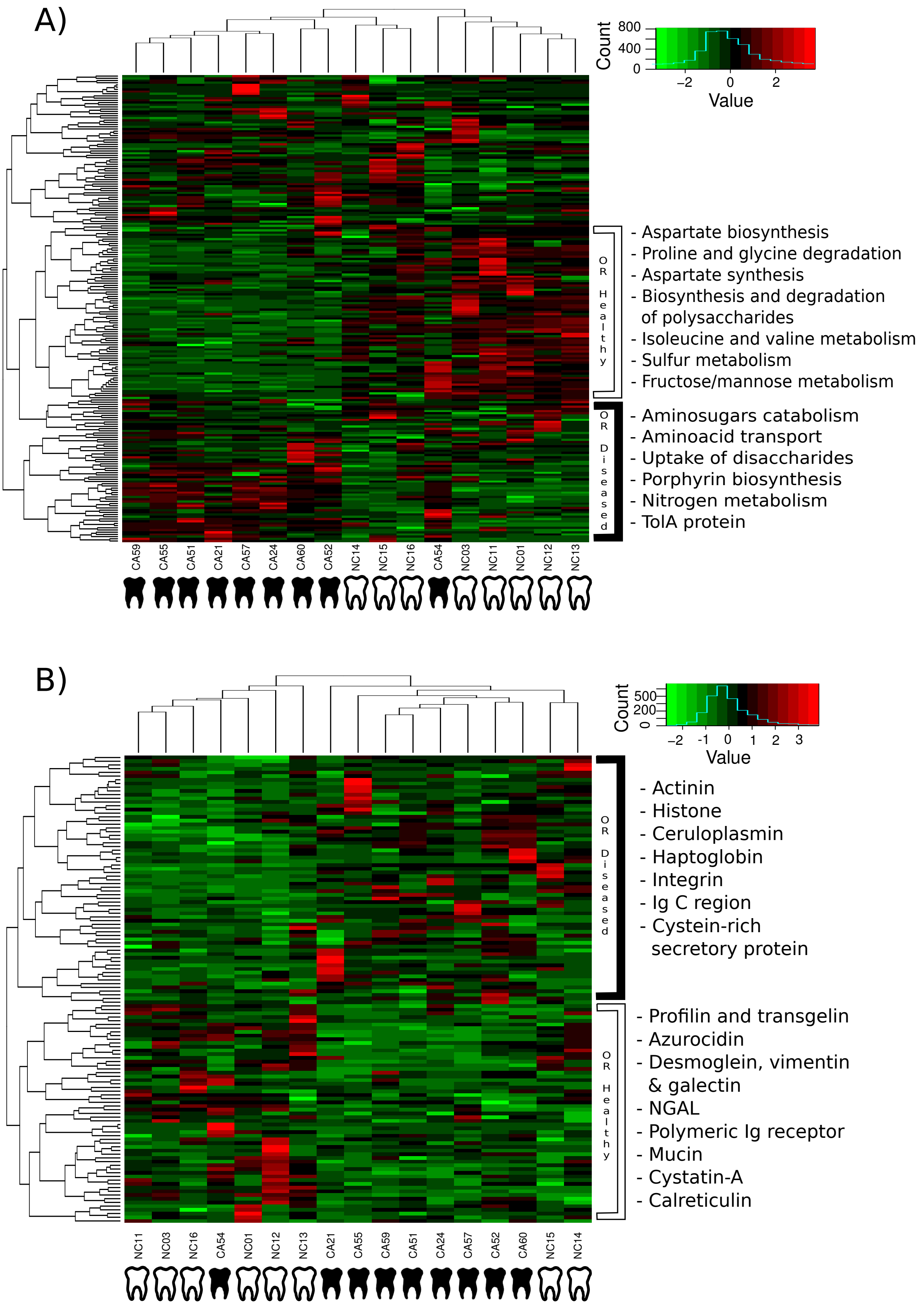


Figure 3. Two way hierarchical clustering based on the bacterial TIGRfam annotation (A) and human proteins (B), showing many proteins differentially represented between healthy and diseased samples. Selected over-represented functions are displayed on the right of the heatmap. Most samples cluster according to the caries status (white/black teeth icons represent healthy and diseased individuals), indicating that caries-free and caries-bearing individuals have different protein profiles. The full list of over-represented bacterial and human proteins can be found in Supplementary Table 4A and 4B, respectively.

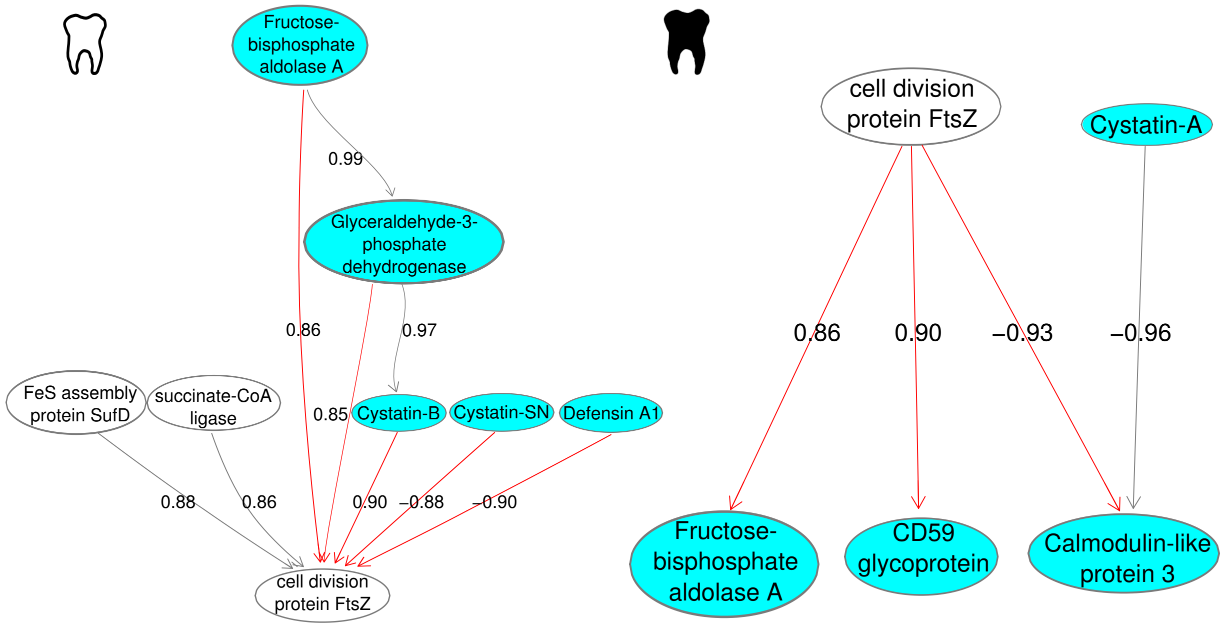

Fructose-bisphosphate aldolase A
0.99
Glyceraldehyde-3-phosphate dehydrogenase
0.86
0.97
FeS assembly protein SufD
succinate-CoA ligase
0.85
Cystatin-B
Cystatin-SN
Defensin A1
0.88
0.86
0.90
−0.88
−0.90
cell division protein FtsZ
cell division protein FtsZ
Cystatin-A
0.86
0.90
−0.93
−0.96
Fructose-bisphosphate aldolase A
CD59 glycoprotein
Calmodulin-like protein 3

Figure 4. Markov blanket networks for the bacterial cell division protein FtsZ. Different correlation networks are found for FtsZ for healthy (white tooth symbol) and diseased (black tooth) individuals. Bacterial proteins are in white, whereas human proteins are in blue. Red arrows connect bacterial and human proteins, and grey arrows connect either bacterial proteins or human proteins. Numbers in the arrows reflect correlation values. The full correlation network can be found in Supplementary Figure 5.

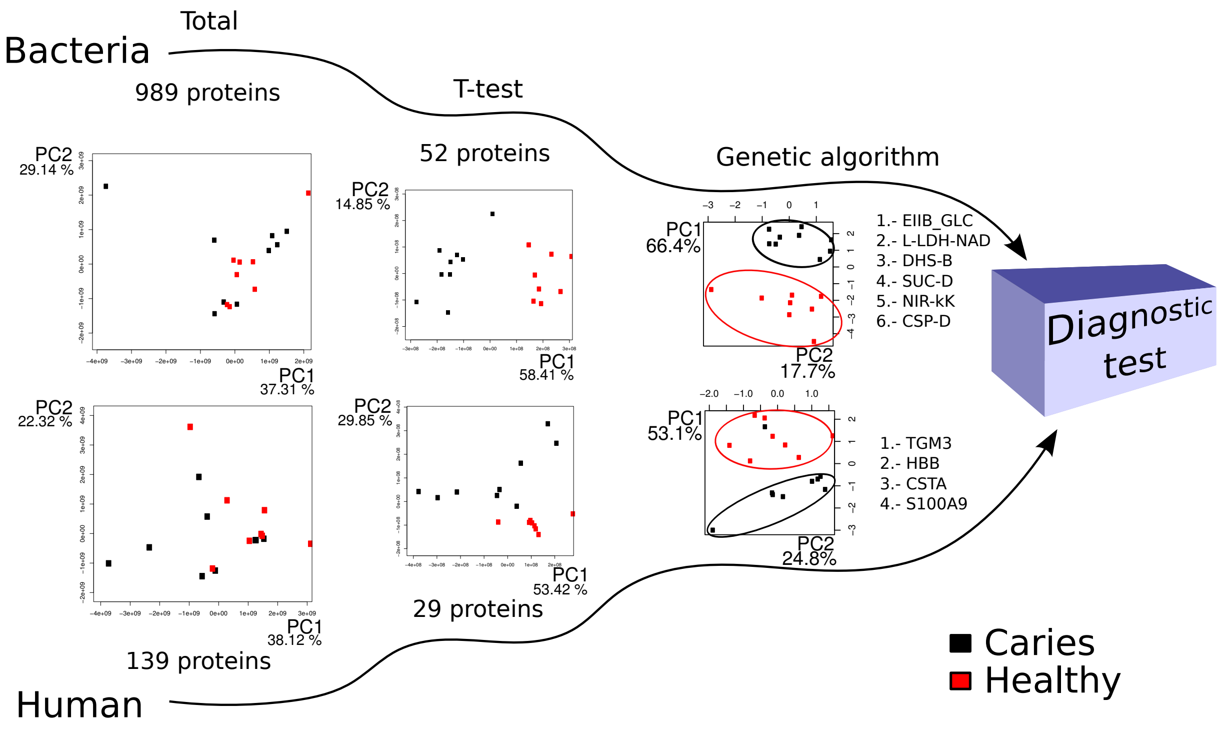
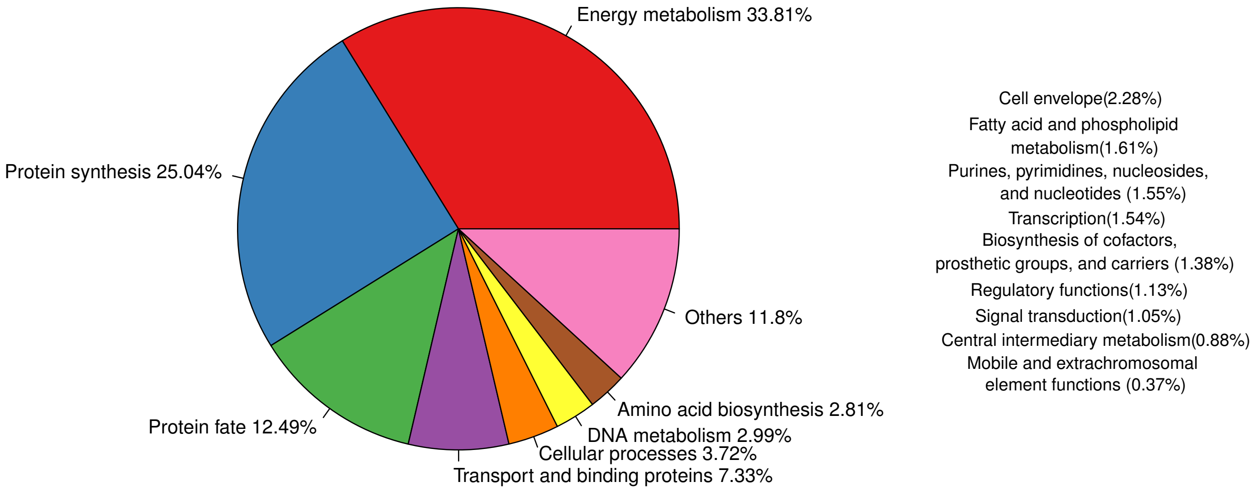
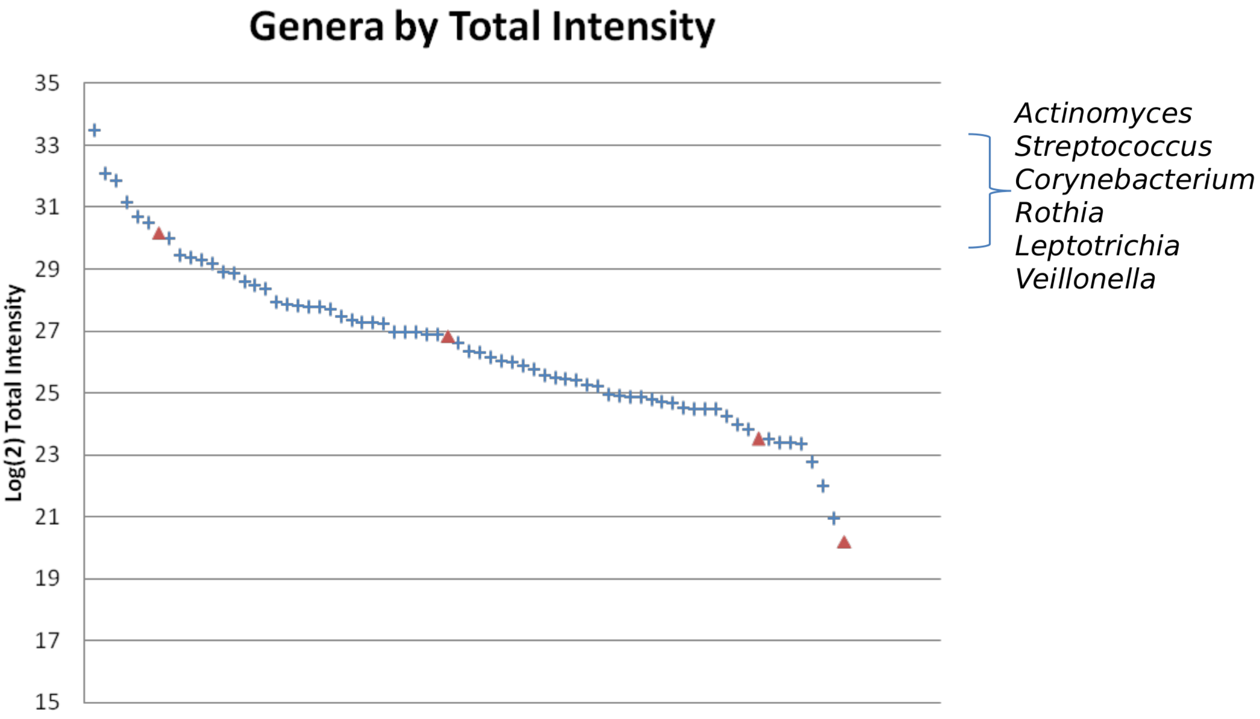


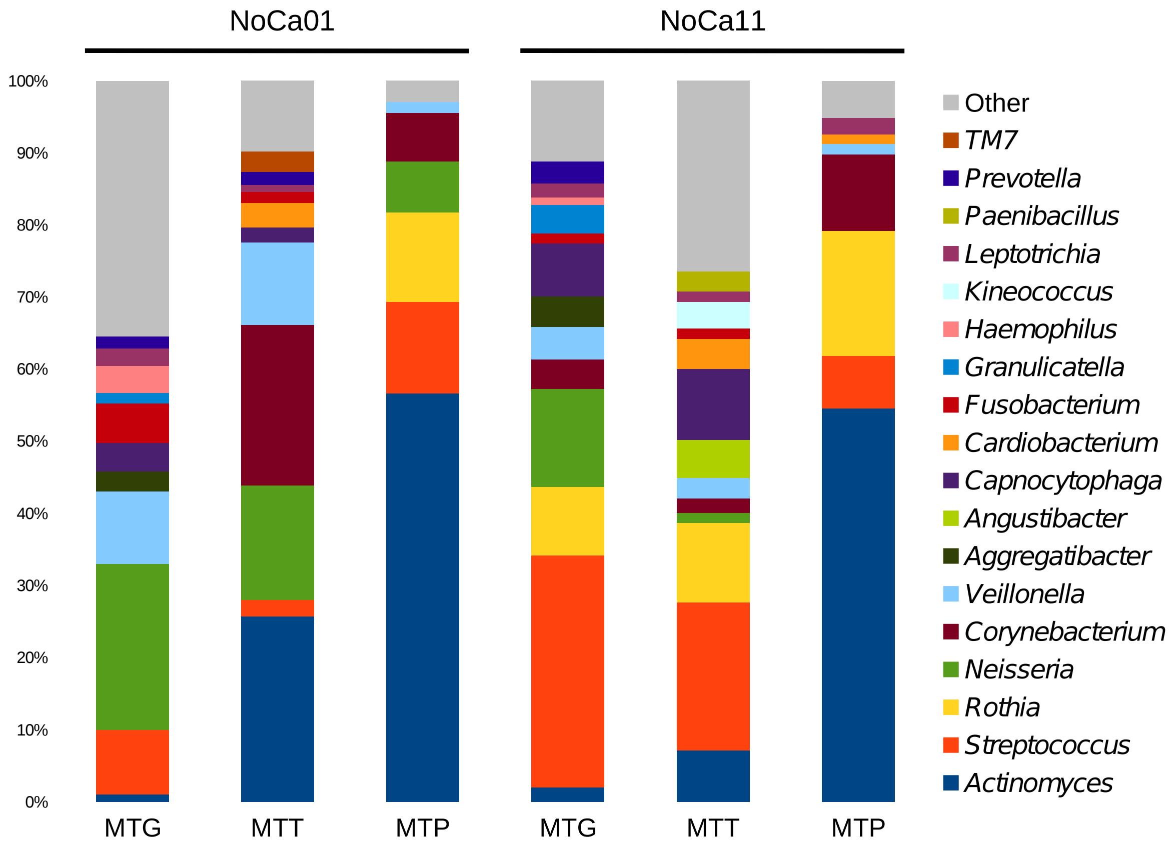
Figure 5. Metaproteomic approach for the identification of potential biomarkers for caries risk assesment. The plots show PCA analyses using bacterial (upper) and human (lower) proteins. Separation between healthy and diseased individuals is improved when using differentially present proteins, selected by a t-test analysis. GALGO analysis further provides a list of 6 bacterial and 4 human biomarkers that still achieve efficient separation and that provide a list of potential biomarkers that will need to be validated in larger, longitudinal cohorts. Ellipses are drawn for clarity to highlight the separate clustering of healthy and diseased individuals.



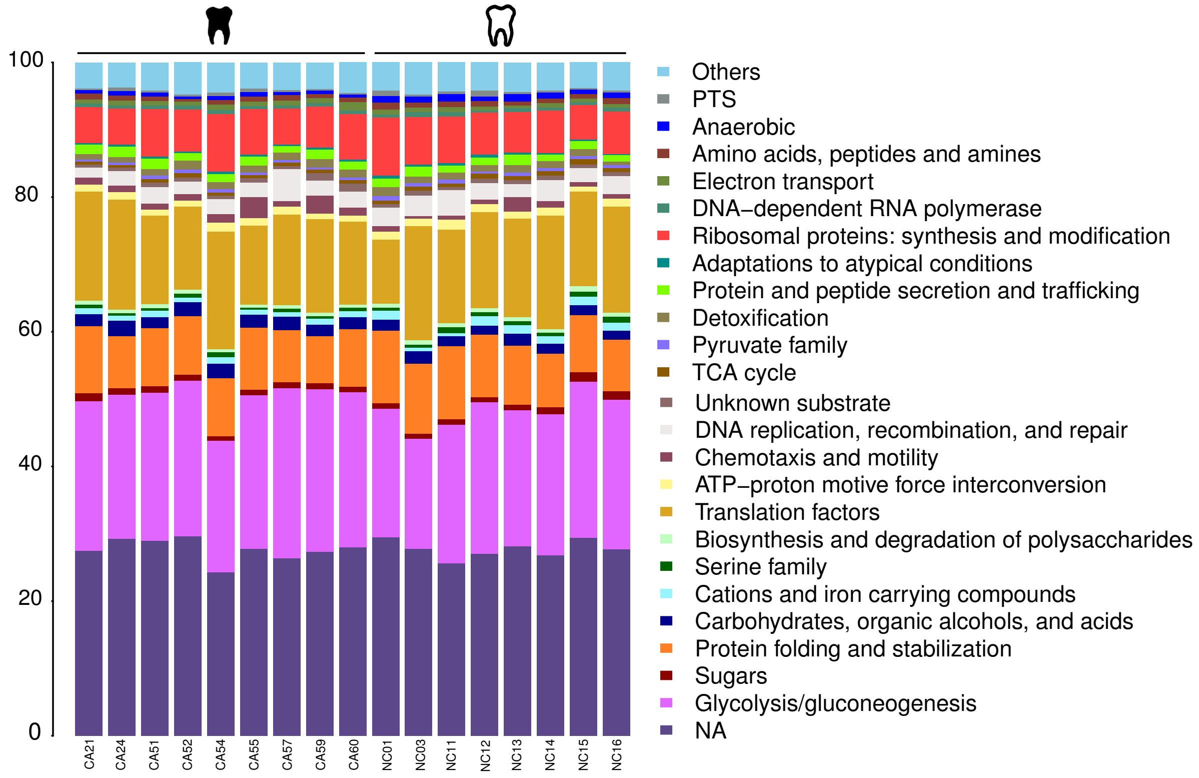
Supplementary Figure 1. TIGR Roles found in the discovery approach, quantified by the total peak area of each peptide. Roles accounting for less than 2.5% of the total bacterial proteins are displayed in the right smaller pie.



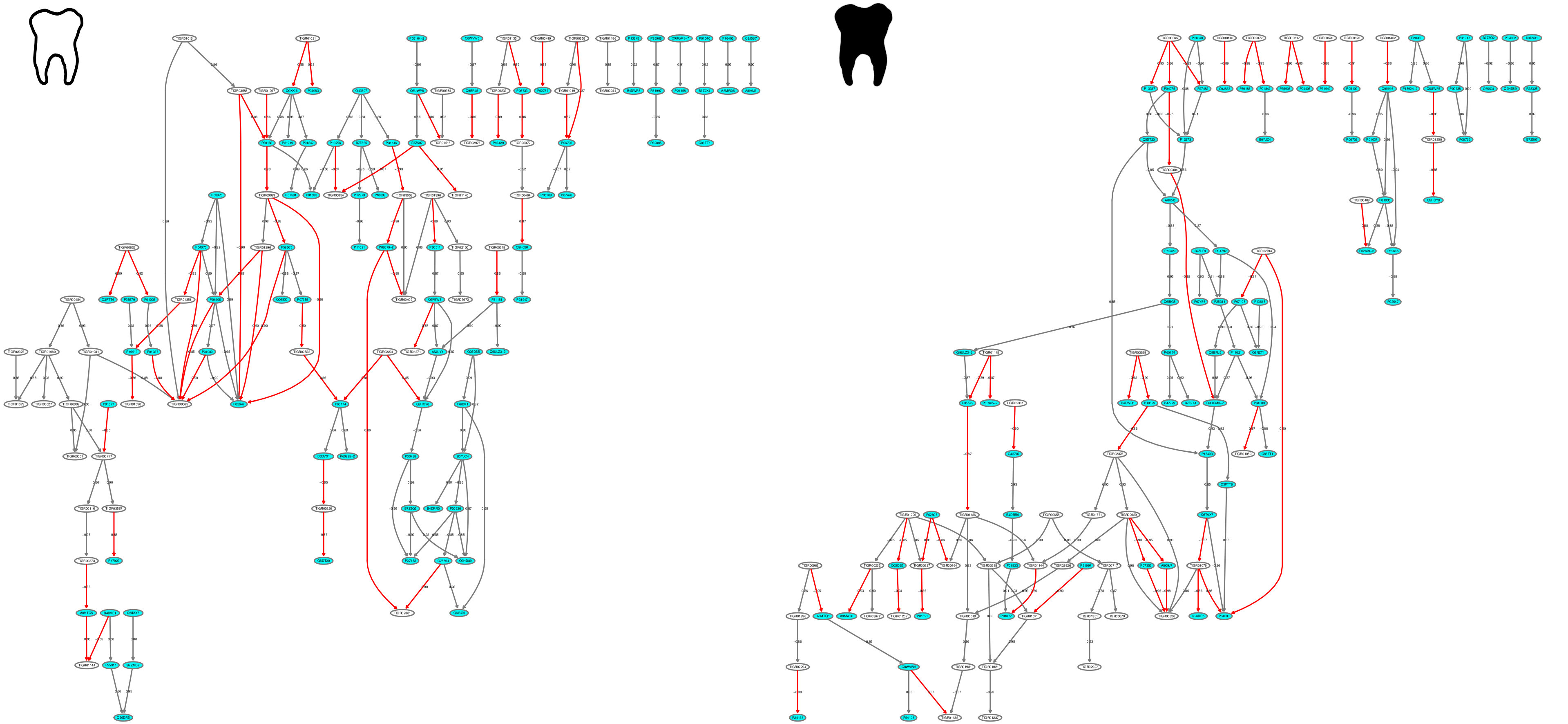
Supplementary Figure 2. Dynamic range plot of protein abundance for each genus. Red triangles mark an order of magnitude difference in abundance. Proteins from genera in the top bracket are therefore approximately 1000-fold more abundant than those at the bottom of the plot.



Supplementary Figure 3. Taxonomic composition of dental plaque from individuals NoCa01 and NoCa11, using metagenomics (MTG), metatranscriptomics (MTT) and metaproteomics (MTP) approaches.



Supplementary Figure 4. Bacterial functional profiles of human dental plaque using TIGR roles. Those roles accounting for less than 0.3% are grouped in “Others”.



Supplementary Figure 5. Bayesian network showing interactions between bacterial (white circles) and human proteins (blue circles), for healthy (white tooth icon) and diseased (black tooth) individuals. Red arrows connect bacterial and human proteins, and grey arrows connect either bacterial proteins or human proteins. Numbers in the arrows indicate correlation values. Numbers within the circles correspond to bacterial TIGRfam and human UniProt functional categories.

Supplementary Table 1. Patients clinical data

Sample	Age	Sex	OHI[a]	Gum status[b]	Decayed teeth	Missing teeth	Filled teeth
Ca021	25	Female	0	1	8	0	8
Ca024	34	Male	0	0	3	0	5
Ca051	37	Male	1	1	3	0	4
Ca052	32	Female	0	0	5	0	11
Ca054	25	Female	0	0	3	0	0
Ca055	23	Female	0	0	4	0	4
Ca057	26	Male	0	0	5	0	6
Ca059	33	Male	1	1	4	0	0
Ca060	27	Male	3	3	9	0	12
NoCa01	25	Male	0	1	0	0	0
NoCa03	39	Female	0	0	0	0	0
NoCa11	27	Female	0	0	0	0	0
NoCa12	24	Male	0	1	0	0	0
NoCa13	26	Male	0	1	0	0	0
NoCa14	37	Male	2	1	0	0	0
NoCa15	19	Male	3	2	0	0	0
NoCa16	36	Female	0	0	0	0	0

a) Oral Hygiene Index: 0.- No visible plaque. 1.- No visible plaque, but adheres to probe. 2.- Low to moderate thickness plaque. 3.- Thick and abundant plaque

b)Gum status: 0.- Healthy. 1.- Mild inflammation. 2.- Redness and/or induced bleeding. 3.- Spontaneous bleeding

4

GENERAL DISCUSSION

GENERAL DISCUSSION

Microbes and humans have been co-evolving through thousands of years, interacting and adapting to each other (McFall-Ngai 2002). In consequence, the human body is inhabited by highly diverse microbial communities, which outnumber in an order of magnitude the number of human cells (Savage 1977). This huge number of microbial cells is perfectly adapted and, under normal circumstances, does not cause disease. The oral cavity plays an important role to the selection of this community, as it is the first entrance point of nutrients and microbes to the gastrointestinal tract (GIT), the largest microbial community of the human body. However, the oral cavity's microbiome has the peculiarity of being prone to cause diseases if no preventive measures are applied such as teeth brushing. Cavities and gingivitis/periodontitis usually happen if dental biofilms are let to grow (Grant et al. 2010). Although those diseases were present in hominids since ancient times (Grine et al. 1990, Tillier et al. 1995, Aufderheide & Rodriguez-Martin 2011, Meng et al. 2011, Wade et al. 2012), their prevalence remained low. The higher intake of carbohydrates thanks to the introduction of agriculture during the Neolithic, and later on the consumption increase of refined sugars during the Industrial Revolution, posed important ecological pressures to the oral microbiome. Those ecological changes reduced the microbial diversity of the oral microbiome, reducing its resistance to the colonization of new unadapted microbes which became odontopathogens (Adler et al. 2013). Those changes lead to an increased prevalence of caries (Richards 2002).

Despite this long history of dental caries in humans, it is still one of the most prevalent infectious diseases affecting human beings (Petersen 2003). During the last century, significant efforts were done to understand the origin of the disease, different ways to combat its incidence and tools for caries diagnosis. But the limitations of the traditional microbiological techniques used, and difficulties due to the long-term and multifactorial origin of the disease, have led the focus to wrong directions and no effective cure or preventive measure has been developed. In fact, the new molecular techniques introduced since the 90's decade, have revolutionized the understanding of caries disease and new hypotheses of its etiology have been proposed, in the light of new high-throughput techniques (Figure 9). In this thesis, a compendium of 5 scientific works will be presented, which tackle open biological questions concerning dental caries disease. Different biological levels (DNA, RNA and proteins) were investigated to get a wide picture of the dental plaque's ecosystem, with genomics, metagenomics, metatransciptomics and metaproteomics approaches, under health and disease states. Furthermore, those still-open biological questions have been faced off applying some of the front-end high-throughput technologies available, such as 454 pyrosequencing, Illumina sequencing and nanoLC-MS/MS.

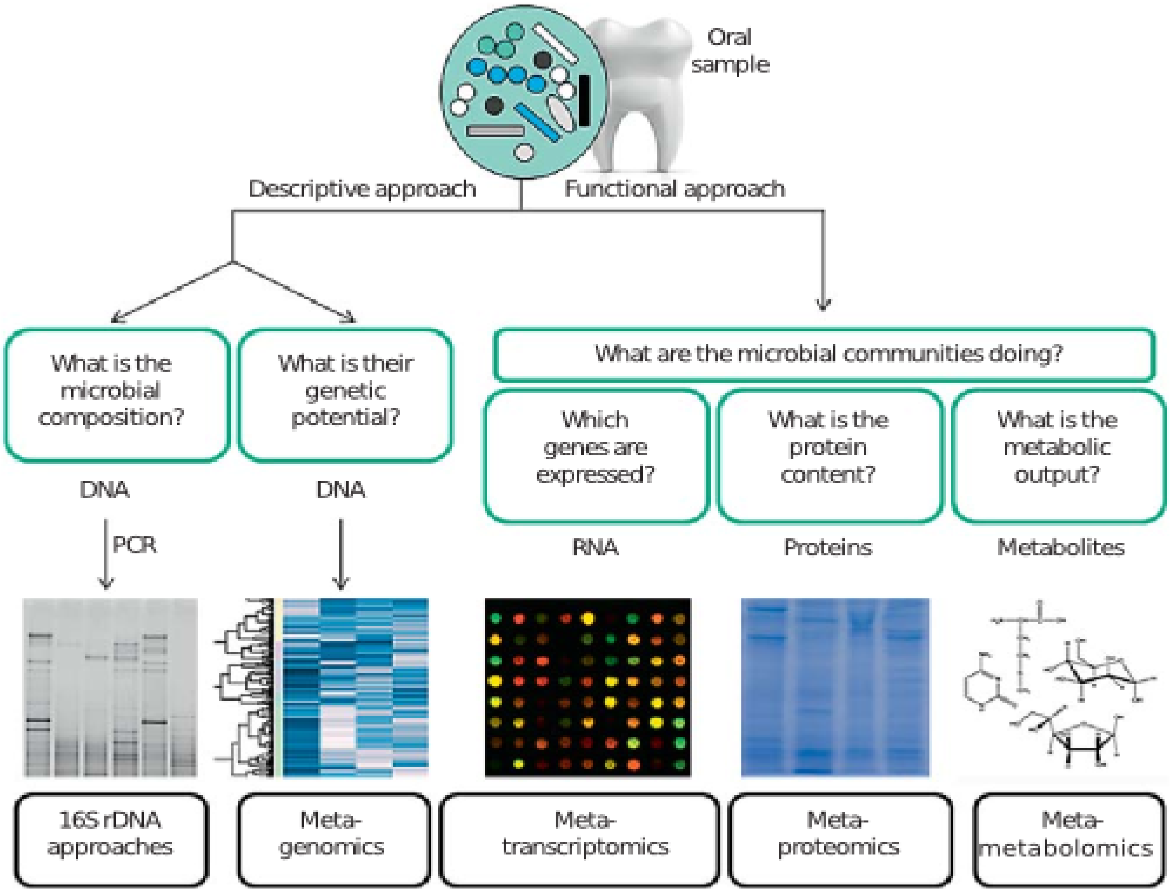

Oral sample
Descriptive approach
Functional approach
What is the microbial composition?
What is their genetic potential?
What are the microbial communities doing?
Which genes are expressed?
What is the protein content?
What is the metabolic output?
DNA
DNA
PCR
RNA
Proteins
Metabolites
16S rDNA approaches
Meta-genomics
Meta-transcriptomics
Meta-proteomics
Meta-metabolomics

Figure 9. Schematic representation of different “omics” approaches available for the study of oral communities. Different essential microbial molecules offer distinct information about the microbial community under study. When DNA is analyzed, based on 16S rRNA gene approaches, the taxonomic microbial composition of the sample can be described, but in order to obtain information about the total functional potential, it is necessary to directly sequence the whole DNA (metagenomics). RNA can be sequenced (metatranscriptome) to obtain which genes of the total genetic potential of the community are being transcribed under given conditions and which microbes are transcriptionally active. Proteins can be analyzed using different mass spectrometry techniques (metaproteome), disclosing information about the functional molecules which develop most of the functions of the cell. Furthermore, metabolites can be measured along (metametabolomics) to detect the activity of proteins. Adapted from (Nyvad et al. 2013).

Culture-based techniques were the first methods applied to dental caries, which led to the non-specific (Miller 1890) and specific plaque hypothesis (Clarke 1924, Fitzgerald & Keyes 1960) of caries etiology. Recent molecular works, based on 16S rRNA gene sequencing, uncovered a huge microbial diversity in the supraGDP (Aas et al. 2005, 2008; Keijser et al. 2008, Zaura et al. 2009, Bik et al. 2010), which has hindered the targeting of potential specific pathogens, compared with classical infectious diseases. Furthermore, the impossibility of culturing most of those inhabitants, and the limited information obtained from 16S rRNA sequence about the lifestyle, metabolic potential and quantitative composition of the community, impedes proper interpretation of the ecosystem ecology. With the aim of overcoming the limitations of the techniques previously used in order to better understand the ecology of the supraGDP ecosystem, we performed a metagenomic analysis in which 6 dental plaque samples from individuals with different caries status were pyrosequenced (Alcaraz et al. 2012, Belda-Ferre et al. 2012). Two samples were obtained from healthy individuals, two from patients who had less than 4 active caries (1 and 4 active lesions), and two samples from patients with a high number of active lesions (8 and 15 lesions). Additionally, enough DNA could be extracted from individual intermediate and advanced dentin lesions to be sequenced. This was the first metagenomic study comparing dental plaque samples from healthy and caries-bearing individuals, and also the first study in sequencing the microbiota present inside dentin cavities.

This methodology allowed the discovery of 186 potentially new 16S rRNA gene sequences previously undetected by PCR amplification methods, revealing a wide undiscovered diversity both in dental plaque and carious lesions. This could be due to lack of amplification efficiency of under-represented taxa and to lack of universal primers in PCR-based studies (Gonzalez et al. 2012). Furthermore, the availability of shot-gun sequences allowed the usage of two binning methods, LCA an phymmBL (Alstrup S., Gavoille C., Kaplan H. 2004, Brady & Salzberg 2009), which allowed to assign up to 75% of all reads at the class level. This in-depth analysis of the whole dental plaque's microbiome lead to important conclusions for the understanding of the caries process.

First, the differences found in previous studies (Aas et al. 2008, Crielaard et al. 2011), in terms of dental plaque samples' taxonomic composition between healthy and caries-bearing individuals, are also present when analyzed by metagenomics approaches. For instance, healthy individuals were more prone to carry Bacilli and Gammaproteobacteria in their microbiota, whereas Clostridia and Bacteroidia were more frequent in diseased samples. This reflects that dental health status is coupled with taxonomic shifts in the microbial composition. This finding, supports the ecological plaque hypothesis of caries (Marsh 1994a, Kleinberg 2002) (see section 1.3.4.3 “Ecological plaque hypothesis” and 1.3.4.4 “Extended ecological plaque hypothesis”), as potential pathogens, such as *S. mutans*, were not abundant or even undetectable in either diseased individuals or dentinal caries lesions. *S. mutans* could only be detected in a white-spot lesion sample analyzed later on (Simón-Soro et al. 2013b). However, this could be due to the relatively low sequencing coverage obtained through 454 sequencing, which hinders the detection of low-abundance microbes, such as mutans streptococci, which have been detected at levels lower than 1% in caries lesions by PCR approaches (Simón-Soro & Mira 2014). Other studies using different techniques, systematically detect *S. mutans* in caries-prone individuals and correlate their abundance to caries susceptibility (Thenisch et al. 2006). From our data, differences in abundance among the whole community suggests that caries is linked with changes in the microbiota composition, rather than by the role of a single species. This is in contrast with the specific-plaque hypothesis, where one or few members of the microbiota (traditionally *S. mutans*) are responsible of the acid production that degrades enamel.

Apart from differences at the genus or species-level detected in ours and other studies (Corby et al. 2005, Aas et al. 2008, Crielaard et al. 2011), variations at the strain level could be spotted by analyzing metagenomic recruitment plots[22]. For instance, *Veillonella parvula* strains presented a metagenomic island (MI) in healthy individuals and diseased patients with less than 3 active caries (Supplementary Figure 4 in Chapter 3.1). Typically, MI contain hypervariable genes, which are under strong selective pressures, such as recognition sites for bacteriophage infection (Cuadros-Orellana et al. 2007, Rodriguez-Valera et al. 2009) or for the immune system (Lan & Reeves 2000). The coexistence of different clones of the same species enables subniche specialization for better exploitation of resources and the ability to evade the attacks of bacteriophages and the immune response, by varying the cell envelop constantly. Those differences at the strain level remind us about the limitations of studies focused in just a single gene, such as 16S rRNA gene, as they remain undisclosed by those techniques. Further improvements in sequencing technologies may enable the complete sequencing of whole genomes when performing direct sequencing metagenomics, so that clonal diversity within a given sample can be spotted, and thus the real biodiversity of an ecosystem can be completely understood. Furthermore, taxonomic and functional binning methods may improve when a wider database of well curated pangenomic information is available for comparison, as hypervariable genes under selective pressure may be properly identified.

Analyzing the metagenomes of the two dentin lesions, we found a wide diversity of microbes. This, together with the fact that *S. mutans* could only be detected in a white-spot lesion that was analyzed later, suggests that the role of this microbe, at least in those lesions where enamel integrity has already been compromised, may not be essential for dentin

[22] Recruitment plots are graphs obtained by displaying the coordinates and similarity values of metagenomic sequences within a reference genome.

degradation. In contrast, it seems more plausible that caries is a tissue-dependent disease, where acidogenic species degrade the mineral content of both enamel and dentin at the beginning of the lesion. Once the enamel integrity has been compromised, and microbes are able to enter the dentinal tissue with a low level of mineralization, the availability of proteins as the main nutrient source, involves an advantage for those bacteria which are better at degrading the proteic content of dentin (see sections 1.3.3 "Caries development" and 1.3.4.5 "Tissue-dependent hypothesis of caries"). Although once the enamel barrier is compromised there is a need for clinical intervention to restore it, new strategies may arise to prevent further dentin degradation. The lower mineral content of dentin allows demineralization with a less acidic pH. In addition, human matrix metallo-proteinases (MMPs) produced by odontoblasts get activated by acidic pH and play an important role in dentin degradationby its collagenolytic activity (Tjäderhane et al. 1998, Vidal et al. 2014), which is the most common protein in dentinary tissue. In addition, bacterial-encoded collagenases have been described in dentin carious lesions and probably also contribute to tissue degradation (Simón-Soro et al. 2013b), even under close-to-neutral pH. Furthermore, some bacteria such as *Fusobacterium nucleatum* are able to recruit MMPs to their cell surface and take advantage to enhance their tissue invasive potential (Gendron et al. 2004). It seems that the control of those two dentin degrading agents may represent a promising research line for discovering new therapies against dentin caries.

The analysis of the functions encoded in the metagenomic sequences allowed to assign both taxonomical and functional affiliations to each of the obtained reads. With this information, we were able to describe which microbes were able to perform each of the functions found in the supraGDP samples. For instance, cell motility functions were mainly encoded in sequences belonging to the Clostridia class. The availability of this kind of information could be useful to determine which are the potential acid producers in dental plaque, in order to direct against them preventive measures to combat caries. Alternatively, depicting the oral inhabitants capable of alkali production may point to potential probiotics that reduce caries risk. The functional information obtained showed that healthy and diseased individuals carry different microbiotas, not only at the taxonomic level, but also in terms of distinct genetic repertoire. This finding suggests that it is important who is present in a bacterial community, as "not every microorganism can perform any function". The implications come from efforts in shaping the microbiome towards a healthy community. Strategies based in the implementation of a single strain, might not be stable through time, as the selective pressure of bacteriophages and the immune system will purge it (Rodriguez-Valera et al. 2009). The implementation of more than one strain of the probiotic species could help in the settlement of the newcomers and be more stable in the long-run, as it has been shown in multi-species probiotic strategies in *Clostridium difficle* infections (Petrof et al. 2013). Another strategy that is showing promising results is the supplementation with prebiotic compounds in order to modify the biofilm environment for the selection of a non-cariogenic microbiota, such as arginine supplementation (Liu et al. 2012, Koopman et al. 2014, Santarpia et al. 2014).

One of the limitations of direct sequencing metagenomics approaches is that the sample is destructed while processing it for sequencing. Other metagenomic approaches, such as cloning procedures, allow to access the cloned material for subsequent analysis or functional experiments. As we were interested in the possible differences between healthy and diseased samples, we partially overcame this limitation by taking samples with similar characteristics of the same individuals. Our goal was to detect as many as dental plaque inhabitants associated to healthy individuals, in order to search for potential probiotics for caries prevention. Based on the taxonomic differences found in the metagenomes between healthy and diseased individuals, we focused primarily on *Streptococcus* and *Neisseria* species. We obtained a total of 192 isolates from healthy individuals (Cabrera-Rubio 2014), which were screened for inhibitory properties against *S. mutans* and *S. sobrinus*, two microbes traditionally associated with caries risk (Thenisch et al. 2006). Four of those were selected for patent protection (Mira 2010), and two of them, isolates 7746 and 7747 CECT, were described as a new streptococcal species, *Streptococcus dentisani* (Camelo-Castillo et al. 2014). This work shows how metagenomics can be instrumental in the identification of potential new probiotic species with therapeutic potential. A key issue for developing successful probiotics is that they can perform its beneficial function at the site where the disease takes place (the teeth surface in the case of dental caries). However, most oral probiotics are bifidobacteria or lactobacilli isolated from human or animal fecal material, which have been selected because they are considered as safe by food safety agencies. Thus, they are not likely to colonize enamel surfaces and in vitro studies show that these gut bacteria are not good oral probiotics (Pham et al. 2011). The work presented in this thesis suggests that metagenomics can provide a feasible methodology to search for natural colonizers of dental plaque with better chances of becoming oral health promoting probiotics.

The finding of *S. dentisani*, based on the results obtained by metagenomics approaches, has taught us that the use of different techniques has synergistic potential. Although metagenomics has the advantage of recovering both culturable and unculturable microbes, it is not able to use the analyzed samples for later uses. For that reason, the development of new culturing techniques that closely resemble the environmental conditions *in vivo* is of vital importance for developing new applications using the yet uncultured majority of the microbial world (Lagier et al. 2012). Recently, a group of researchers has used an special device for simulating soil conditions for culturing microbes of this environment (Nichols et al. 2010), cultivating up to 50% of the diversity found in soil samples. Using this technology, a new antibiotic compound against Gram-positive was discovered from a previously uncultured soil inhabitant (*Eleftheria terrae*), which binds to the lipids II and III and inhibiting their incorporation to the cell wall, a previously unknown mechanism of action (Ling et al. 2015). Those findings reflects the potential of metagenomics approaches to discover new species, which may contain functions susceptible of future valuable applications, and the need of improvement of current culturing techniques, in order to be able to explore the new applications hidden in the “uncultured majority”. Similar approaches can be applied to isolate more microbial species that are not cultivable nowadays and thus increase our knowledge of oral diversity.

In the second paper of the present thesis, a metagenomics-based method to detect putative virulence genes in pathogenic strains is proposed (Belda-Ferre et al. 2011). When a reference strain bacterial genome is compared to a metagenome of a sample coming from the same environment where it was isolated, some genomic regions are not recovered in the metagenome. Those regions, termed as metagenomic islands (MI), typically include highly variable genes, such as those coding for exposed cell wall proteins, which are under high selective pressures due to the attack of either bacteriophages or the immune system in human-associated niches. In the case of pathogenic bacteria, where the pathogenic potential is provided by a gene or a group of genes (for instance a toxin), those are usually absent in non-pathogenic strains, making the comparison between pathogenic and non-pathogenic strains a strategy for detecting those virulence genes. However, the wide variety of strains present within a single environment, implies a big number of genes that are not shared between two strains, the so-called “dispensable pangenome” (D’Auria et al. 2010). Thus there is a need of comparing multiple strains between them in order to find potential pathogenic genes (Ho Sui et al. 2009, D’Auria et al. 2010, Hilker et al. 2014). The limitation of this approach is still the need for isolating enough strains of the same species which would be biased towards those strains better adapted to grow under laboratory conditions. Alternatively, metagenomics offers the possibility of comparing a bunch of non-pathogenic strains of a given species living in the same niche against the pathogenic strain of interest in a single comparison using recruitment plots. This facilitates the fast detection of virulence genes when an outbreak is detected and preventive epidemiological measures are to be taken. For instance, this method was tested with the *E. coli* O104:H4 strain that caused an outbreak in Germany in 2011. It was rapidly sequenced by several international groups, allowing the scientific community to analyze its genome. By comparing the genome of both the chromosome and the plasmid present in the *E.*

coli O104:H4 TY2482 strain, against stool metagenomes of healthy volunteers made available by the HMP, all virulence related genes described in this strain were found in the MIs. For instance, virulence determinants of the TY2482 genome, such as the shiga toxin encoding phage, the mercury resistance cluster, beta-lactamases, adhesins (*pig*,), several toxin-antitoxin systems, the aerobactin siderophore, Serine Protease AutoTransportes of Enterobacteriaceae (SPATEs, such as *pic*), and the antigen 43 (mediates cell-to-cell interaction in *E. coli* biofilms), do not recruit metagenomic sequences from stool metagenomes of healthy individuals. This shows that metagenomics is a highly versatile tool that can be used in clinical applications, such as outbreak investigations. The discovery of virulence genes in such cases may enhance treatment success of the affected patients, adapting it to the pathogenic potential of the causative bacteria.

Two of the most important factors on which caries appearance depends on, are biofilm formation and acid production through carbohydrate fermentation by supraGDP inhabitants. Those two determinants were studied using two different metatranscriptomics approaches under *in vivo* conditions (Benítez-Páez et al. 2014). Biofilm formation process has been traditionally studied using *in vitro* models (Kolenbrander et al. 2006), or using close-end technologies, such as checkerboard DNA-DNA hybridization arrays (Li et al. 2004, Teles et al. 2012) but the findings of those studies still remain to be proved on *in vivo* studies using open-ended technologies. Furthermore, little is known about the functional activity of those colonizers at different biofilm formation times. Biofilm is thought to be developed in a sequential manner, where there is a clear distinction between initial colonizers, late colonizers and the bridging capacity of *Fusobacterium sp.* between those two groups (Kolenbrander et al. 1989). This succession process and the coaggregation patterns are represented in Figure 10, where initial colonizers are those able to attach to the AEP over the clean surface of teeth, which at the same time are used by other inhabitants as anchor to attach to the biofilm. However, this biofilm model proposed by Kolenbrander and collaborators, was based on the coaggregation patterns observed in different oral species, tested in a pair-wise manner under *in vitro* conditions. Although some studies confirm higher abundances of early colonizers during the initial stages of oral biofilm formation (Li et al. 2004), there is scarce knowledge about the colonization process *in vivo*. For instance, it remains to be confirmed which of those microorganisms able to adhere are actually active at a given biofilm formation time.

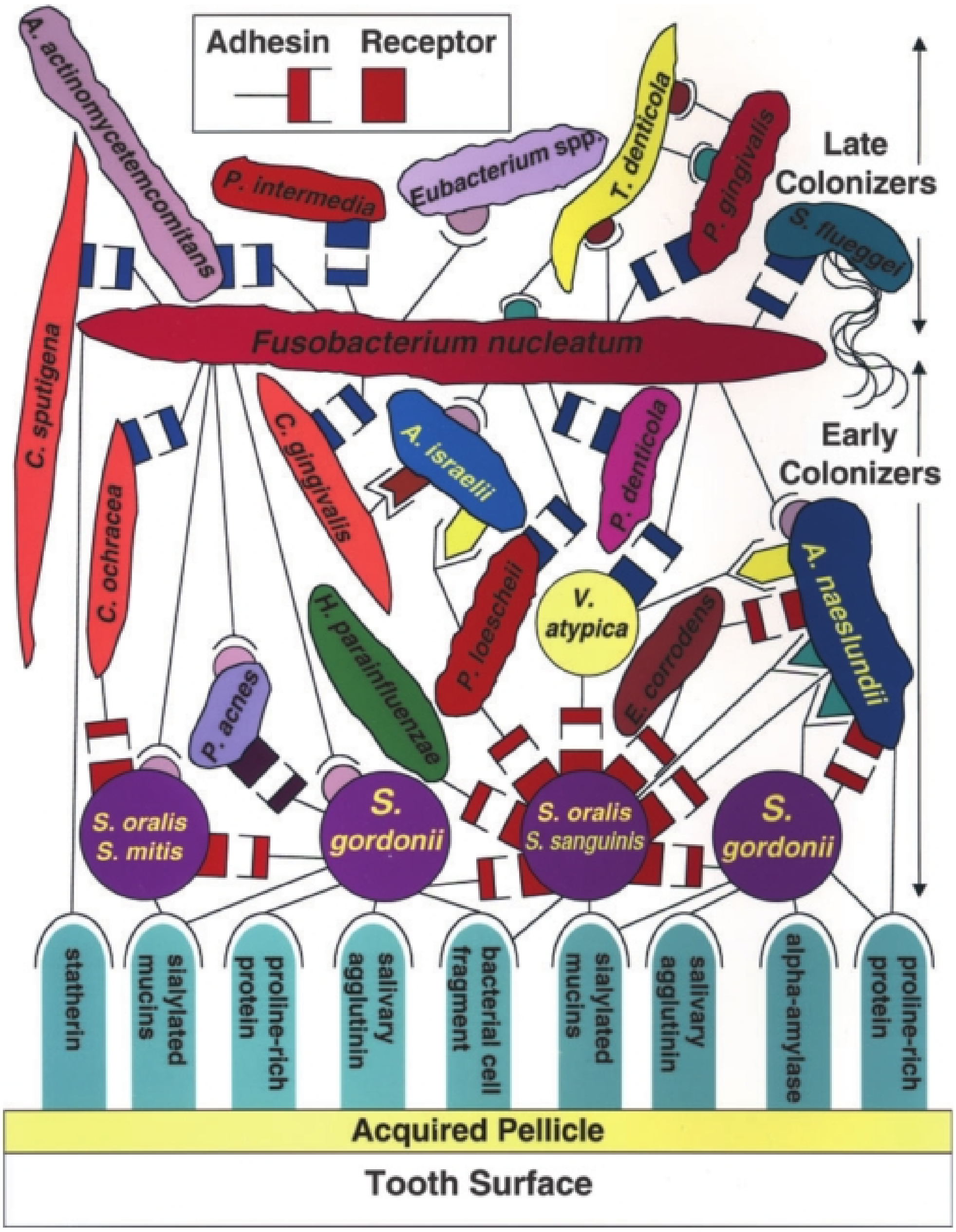


Figure 10. Spatiotemporal model of oral bacterial colonization, showing recognition of salivary pellicle receptors by early colonizing bacteria, and coaggregations between early colonizers, fusobacteria and late colonizers of the tooth surface. Each coaggregation depicted is known to occur in a pairwise test. Collectively, these interactions are proposed to represent development of dental plaque. Adapted from (Kolenbrander et al. 2006).

The microbial source available for colonizing teeth surfaces is the saliva, as it is continuously bathing all oral surfaces. However, it is known that its bacterial composition significantly differs from teeth microbial community. Therefore, differences in this microbial composition can be given either by a selective attachment process of salivary microbes, which leaves out of the biofilm those microbes that are not able to attach to the supraGDP, or by different growing capabilities inside supraGDP or in salivary suspension. Additionally, the gene expression pattern present at each stage of biofilm formation may reveal important information about the key functions involved in biofilm formation, which could be used for interfering the process. In this work, we aimed to decipher which of the bacteria able to attach to the biofilm were transcriptionally active at different biofilm formation times (6, 12, 24 and 48 hours) after a professional cleaning intervention, and which genes are being transcribed throughout the biofilm formation process. Along the time sampled, we were able to detect that many genera were active at all stages of biofilm formation, with some variations in the transcriptional activity. Those changes found presented positive and negative correlations between oral inhabitants, probably reflecting interspecies antagonism and mutualism, or different adaptation to the initial or mature oral biofilm. This could reflect that although all the inhabitants of the biofilm are present at any given time point, its activity is mostly favored when the conditions are optimal for their development. The high active diversity found in the early stages of biofilm formation sampled, suggests that attachment to the teeth surface may not happen through an ecological succession, or if it does, it occurs in a shorter time period than previously thought. In contrast, bacterial cells forming aggregates present in the salivary fluids or adhered to epithelial cells could be jointly attached to the teeth (Cabello Yeves 2014). Those who have the best conditions for growing, will present the higher transcriptional activity, as we were able to detect. For instance, the initial biofilm is not thick enough for the development of anaerobic microniches. Therefore, strict anaerobes and microaerophiles inhabitants of the supraGDP will increase their activity only when the thickness of the biofilm and its oxygen content decreases, allowing the continued anaerobic conditions required for their growth. For instance, we found negative correlations between the early colonizer *Actinomyces sp.* and the late colonizers *Aggregatibacter, Veillonella, Volucribacter and Haemophilus,* reflecting the preferential growth conditions of each of those microbes. On the other hand we found positive correlations between *Fusobacterium* and other microbes, such as *Bacteroides*, *Leptotrichia* and *Bacteroides*, in concordance with previous coaggregation studies (Kolenbrander et al. 1989, 2002). This reflects the bridging role previously described of *Fusobacterium* between early and late colonizers. Apart from the mere coaggregation, future studies must determine if there is any metabolic supplementation between coaggregating species, and which are the mutual benefits obtained by keeping aggregated. This could lead to finer strategies for biofilm control and microbiota reshaping towards a healthier and non-cariogenic community.

When the mRNA transcripts were analyzed, we found differential expression between early (6 and 12 hours) and late (24 and 48 hours) biofilm samples. In the early biofilm samples, we found 35 KEGG Orthology (KO) categories over-expressed, which reflect the importance of niche exploitation and the absence of competence at the initial moments of biofilm development. Among those KO there were genes involved in Carbohydrates metabolism, Energy, Amino Acids, Cofactor and Vitamins, Xenobiotic Degradation and translation (ribosomal proteins and Elogation Factor Tu and G). Those categories are mainly related with central metabolism of bacteria, and therefore used for fast replication. On the other hand, late biofilm samples showed more variable functional profiles, with over-expressed functions dealing with Cell Motility (chemotaxis, pilus and flagella assembly), ABC transporters, DNA repair systems and Homologous Recombination systems. Those categories, together with those of Type II secretion systems for sensing and producing mutacins and Competence related functions, reflect the importance of cell to cell communication and environment sensing, in order to produce and secrete molecules inhibiting the growth of potential competitors. Once the biofilm is fully established, the access of nutrients to the inside of the biofilm is reduced, and thus the competence for the scarce nutrients is increased. Cell motility and chemotaxis suggest that the motility inside mature biofilms is also important as a response to eventual disaggregation from the biofilm structure.

The other biological question that was confronted using metatranscriptomic approaches, was to determine which are the biofilm inhabitants that are more trascriptionally active during the pH drop typically observed after a carbohydrate-rich meal. The objective was to identify the number of potential acid-producing or acid-tolerant members of the biofilm, which may be considered as likely players in caries etiology. For that reason, we used 454 pyrosequencing for sequencing cDNA transcripts, which yielded over 200.000 reads, of which 98.8% were from SSU[23] and LSU[24] ribosomal genes transcripts. The high proportion of ribosomal transcripts allowed the description of the active members of the community without the need of amplification steps. Bacterial composition as obtained by both SSU and LSU ribosomal genes were highly similar. The main observation found was the high dominance observed in the active community: In some of the samples, over 80% of the transcripts corresponded to *Actinomyces*, *Corynebacterium* and *Rothia*. This contrasts with the typical pattern found in the metagenome, where the total abundance of those 3 genera is much lower, emphasizing the need of applying multiple techniques in order to have a more complete view of ecosystem functioning. Although no specific pattern could be detected to change before and after the meal, we observed that some individuals had a strong resilience to changes in the active community, whereas in others, the changes were more pronounced. This could be due to a faster pH increase in those individuals than in others, or a better counter-

[23] SSU: ribosomal small subunit

[24] LSU: ribosomal large subunit

response of the members of the community to the acidification process. For instance, caries-free individuals have been found to secrete higher amounts arginine and lysine in saliva (Van Wuyckhuyse et al. 1995), which may help reduce the pH drop occurring after carbohydrate consumption. For that reason, in future studies the samples should be taken not based on the time since the end of the meal but based on the pH curve produced by the biofilm, in order to retrieve the changes happening during the peak acid production, as it may happen at variable times depending on several factors.

When healthy and diseased samples were compared using a high-dimensional class comparison test (Segata et al. 2011), we found that *Actinomyces* was the only genera associated to healthy individuals. Members of the *Actinomyces* genus have been described to be pH-rising members of oral biofilms (Wijeyeweera & Kleinberg 1989, Liu et al. 2006), given their ureolytic activity. The higher transcriptional activity of *Actinomyces* can be the reason of a lower tendency to suffer enamel demineralization, as they are able to produce both the pH drop and rise, in contrast with other oral inhabitants. However, the extremely low number of mRNA transcripts sequenced did not allow to search for ureolytic genes and this extent could not be demonstrated. Caries-associated genera were *Leptotrichia*, *TM7*, *Tannerella*, *Porphyromonas*, *Fusobacterium* and *Capnocytophaga*.

One of the limitations of direct cDNA sequencing, is the low proportion of mRNA transcripts obtained. In consequence, only a small fraction of the highly transcribed genes can be detected. At the time of experiment design, rRNA substraction methods were not properly developed and rRNA removal was not consistent, depending on the sample analyzed (He et al. 2010). However, the most interesting part of this kind of experiments is the discovery of the actively transcribed genes, as this will show the transcriptional adaptation that the whole community presents under certain environmental changes. When the response to the acidic environment caused by carbohydrate fermentation will be fully understood, new interventions could be designed to interfere with the pH drop which ultimately initiates caries lesions.

For comparison purposes with our previous metagenomic research, we sequenced cDNA from dental plaque of the same individuals than the ones taken in (Belda-Ferre et al. 2012). Interestingly, we found that the transcriptionally active oral microbiota as seen by cDNA sequencing, highly differs from the total community seen by the metagenomic approach. For instance, minor members of the community as seen in the MTG, such as *Actinomyces*, *Corynebacterium*, *Cardiobacterium* and TM7 genera, had high transcriptional activity in the MTT. On the other hand, *Neisseria*, *Streptococcus*, *Capnocytophaga*, *Fusobacterium*, *Aggregatibacter* and *Haemophilus*, were more abundant in the MTG than in the MTT. Interestingly we found minor members of the community only in the MTG, such as *Acinetobacter*, *Bacteroides*, *Bastia*, *Clostridium*, *Eikenella*, *Enhydrobacter* and *Psychrobacter*, which reflect their low activity in the biofilm at the sampling time. Those genera may not play important roles in the oral environment at least under the conditions present during the sampling procedure, and may just be transient bacteria coming from food or other external sources. The advantage of knowing the transcribed genes is that spurious and not active members of the community can be discarded and not taken in consideration in the analysis, given the high lability of RNA molecules, which are rapidly degraded.

The last work included in the present thesis is the first metaproteomic study performed in human supraGDP samples. The main objective of this work was the discovery of all the proteins traduced by the dental plaque microbial community in mature biofilm samples, and determine whether the functional profiles of healthy and caries-bearing individuals differed and to what extent. Those differences could potentially be used as biomarkers of health and disease, which may be used for the implementation of preventive measures before the disease onset.

As previously stated in this discussion, MTG approaches show the whole functional potential of the microbial community, but is not able to distinguish which of those functions are relevant to face off certain environmental conditions. MTT has the ability to detect those genes that are actively being transcribed at a given moment, but the existence of post-transcriptional modifications and regulation makes the correspondence between transcriptional and functional activity not direct (Nogueira & Springer 2000, Pradet-Balade et al. 2001). As we could observe in our datasets, MTG, MTT and MTP analysis of similar samples provided different taxonomic compositions. The most surprising observation was that *Actinomyces*, which appeared as a minor member of the community as seen in the MTG approach, had much higher transcriptional activity in the MTT and an even higher abundance of proteins, as seen in the MTP. This reflects the complementarity of those techniques, as they show different aspects of a microbial community. For instance, if supraGDP was only analyzed using just one technique, one would think that the importance of *Actinomyces* in the oral cavity could be of minor interest, given its low abundance. However, its high transcriptional and proteic activity using MTT and MTP suggests a more important role in mature supraGDP. However, one needs to take into account the potential biases introduced by each methodological approach, as the results might be affected by them. For instance, the sequencing bias against high-GC content sequences, may influence the abundance of high-GC microorganisms in the MTG and the MTT, compared to the MTP. On the other hand, the low dynamic range available in the actual mass-spectrometers, impedes the discovery of low abundance proteins, hindering the discovery of proteins from minor inhabitants of the oral cavity.

The MTP paper presented in this thesis, is the first proteomics approach characterizing the human dental plaque. One of the objectives was to describe as many as possible proteins present in the supraGDP. To accomplish that, we performed a discovery experiment in which we applied a HILIC prefractionation of trypsin-digested proteins pooled from 17 samples, followed by nanoLC-MS/MS. With this approach, we could detect 7771 bacterial and 853 human proteins, whose distribution denoted a high dominance, with abundance differences of three orders of magnitude, being *Actinomyces* the most abundant genus. The presence of human proteins in the supraGDP in high amounts reflected their importance for biofilm development and control, being most of them of mucosal or salivary origin.

In order to detect the differences under health and disease conditions, we accomplished a quantitative experiment, aiming at describing and quantifying the proteins found in healthy and caries-bearing individuals supraGDP samples. Even with the lack of a prefractionation step of the samples, we achieved the detection of over 1000 proteins per sample on average. Interestingly, several features affecting the pH in the biofilm were differentially present in caries-free and caries-bearing individuals. Higher levels of L-lactate dehydrogenase were found in healthy individuals. An unusual LDH carried by *A. actinomycetemcomitans* converts lactate into pyruvate, and thus removes from the biofilm one of the acids with a bigger contribution to biofilm acidification. However, conventional LDH produce lactic acid, which decreases the pH. As taxonomic assignment of LDH found in this work could not be accomplished, the reason for the higher abundance in caries-free individuals is unclear. Healthy individuals also presented higher amounts of ornithine carbamoyltransferase, which participates in the arginine deiminase system, one of the acid protection systems exhibited by oral microorganism that is now being exploited with arginine supplemented toothpastes (Koopman et al. 2014, Santarpia et al. 2014). This higher abundance of ornithine carbamoyltransferase could reflect an adaptation of the microbiota to the higher arginine and lysine salivary content of healthy individuals (Van Wuyckhuyse et al. 1995). On the other hand, caries-affected individuals had higher amounts of PTS system and ABC disaccharides transporters, reflecting a higher sugar intake potential. They also presented higher amounts of N-acetylglucosamine-6-phosphate deacetylase, which is able to degrade this aminosugar present in macromolecules from saliva and GCF, allowing acid production after dietary sugars have been depleted from the biofilm (Moye et al. 2014). The differences found point towards important traits influencing the onset of caries disease. Using those microbial functions may potentially be used to detect early stages of the disease before they are clinically visible and apply corrective measures to prevent caries lesions.

In this quantitative approach, a total of 127 human proteins were detected and quantified, which were much more abundant than bacterial ones. This high concentration of human proteins, difficulted the detection of bacterial proteins, as actual mass-spectrometers' dynamic ranges are not able to include such big concentration differences. However, we could adjust the concentration and amount of sample loaded to obtain enough signal for both human and bacterial proteins. In the future, it would be advisable to separate human proteins prior to MS/MS analysis, in order to allow wider dynamic ranges for bacterial proteins and thus detecting a higher number of those.

The human proteins found in the quantitative approach showing significant differences included keratinized epithelium, host immune response proteins, Fe metabolism and proteases or their inhibitors, among others. The finding of proteins typically found in the keratinized epithelium suggested that desquamated mucosal cells can be attached to the growing biofilm (Brecx et al. 1981). Epithelial cells have been proposed as the transporters for bacterial enamel colonization, as they have adhered microbial cells that benefit from this adhesion to the teeth and from nutrient availability from the cell detritus (Tinanoff & Gross 1976, Tinanoff et al. 1976). Additional in vivo studies must confirm the role of epithelial cells in the development of supraGDP, as they could be susceptible of intervention strategies to control not only oral biofilms, but also other biofilm-associated diseases such as catheter and biomedical implants infections that might have similar initiation procedures.

Unlike other oral surfaces, which are protected by the epithelial innate immune system, supragingival dental plaque is only subject to the mucosal adaptive immune system. But secreted exoplysaccharide can impede the access of those molecules to the inside of the biofilm, impeding a proper immune surveillance. However, several proteins related to the human immune response were detected. It is still to be deciphered how the immune system interacts with biofilms and how influences its formation, but the results of this chapter suggest some of the inter-talks taking place between the host and the biofilm. For instance, high amounts of immunoglobulins were detected, which intriguingly were more abundant in diseased individuals. This is in congruence with a systematic review where IgA levels were found to be higher in caries-active subjects (Fidalgo et al. 2014). Although Igs have aggregating properties preventing bacterial adhesion to the teeth surface, bacteria have different strategies to avoid it. For instance, some bacteria can deplete IgA carbohydrates and use them as a nutrient source, and after the IgA has been deglycosylated, it becomes more susceptible to proteolitic degradation, loosing their activity (Frandsen 1994). Furthermore, some studies suggest that SIgA contribute to biofilm formation at least in the gut environment (Randal Bollinger et al. 2003), which can lead also to an increased biofilm growth rate. On the other hand, higher amounts of cellular immunity proteins (azurocidin, C3, pIgr, RAC-2 and hASC-3) were found in healthy volunteers. The importance of those immune responses in the protection against caries disease must be further explored, in order to develop new strategies that increase the immune response against biofilm development, different to current vaccination approaches directed towards *S. mutans*.

Taking into account both human and bacterial proteins abundance variation between healthy and caries-bearing individuals, we hypothesized that those differences could be used to diagnose more efficiently individuals with a tendency to develop caries disease. With this aim, we looked after those proteins presenting significant abundance differences by applying a T-test, which detected 53 bacterial and 29 human proteins. However, the high number of proteins detected is not suited for its implementation into a feasible diagnostic kit. Therefore, a multivariate analysis was applied in order to detect the minimum number of proteins that, when measured together in supraGDP samples, could discriminate healthy from caries-affected supraGDP samples. This analysis selected a set of six bacterial and four human proteins, that allowed to determine health status with an estimated specificity and sensitivity of over 96%. Those proteins could potentially serve as biomarkers and be measured routinely in dental care clinics, for better prospective treatments. Although those biomarkers found in this study are theoretically valid for disease diagnosis, they will need to be confirmed for use in caries risk assessment previous to the disease onset, in order to develop preventive treatments. Further large-scale long-term follow-up studies need to be performed to confirm the validity of those biomarkers or propose new ones that complement them, that could be used to detect caries risk in healthy patients. The health impact of this kind of measures could alleviate the burden of oral diseases derived from dental caries, which are highly prevalent in today's societies.

5

CONCLUSIONS

CONCLUSIONS

1.- A true metagenomics approach through direct pyrosequencing of supraGDP allows a direct microbial abundance quantification without the limitations of culture, cloning and PCR amplification steps imposed by previous methodologies.

2.- Binning analysis of metagenomic reads allow to identify the taxonomic affiliation of a large proportion of all reads obtained, not only 16S rRNA gene sequences, representing a wider taxonomic landscape than traditional 16S-based molecular approaches. Additionally, each read can be both taxonomically and functionally annotated, which enables to know the genetic potential of each taxonomic group.

3.- Caries-bearing individuals harbor a taxonomically and functionally different microbiota from healthy controls, although this tendency should be confirmed using larger sample sizes. Diseased volunteers presented higher amounts of Clostridiales and Bacteroidetes, whereas healthy individuals showed higher abundances of Bacilli and Gamma-Proteobacteria, particularly *Aggregatibacter*, *Rothia* and *Kingella*.

4.- Functional composition of caries-bearing individuals seems to be more conserved than in healthy individuals, suggesting that more diverse functional potential represents a higher resilience to caries onset.

5.- Cavities are diverse ecosystems, not dominated by *Streptococcus mutans*, suggesting a polymicrobial etiology. Therefore, strategies only focusing in combating *S. mutans* are likely to be ineffective.

6.- Healthy microbiotas are a natural reservoir of potential therapeutic probiotics and antibacterial compounds. *Streptococcus dentisani*, isolated from healthy individuals, represents a clear example, which is being developed as an anticaries probiotic.

7.- Metagenomes from healthy individuals can be used to detect putative virulence genes from pathogens, by comparing their genome sequence against a healthy MTG obtained from the same environment.

8.- Transcriptional variations before and after carbohydrate intake are highly variable and mainly depend on individual's microbial composition. Some individuals present a highly stable transcriptional activity, which may reflect a high resilience to environmental changes such as acidification.

9.- Proteic content of 24 hours supraGDP is highly dominated by *Actinomyces*, *Corynebacterium*, *Rothia* and *Streptococcus* proteins, representing a higher dominance than previous MTG and MTT studies showed.

10.- Human salivary and epithelial-related proteins are present in high abundance in the supraGDP, presumably exerting defense functions dealing with immune response, Fe metabolism, proteases and their inhibitors.

11.- Proteic content differences between healthy and diseased individuals may allow the development of tests which could be useful for caries-risk assessment at pre-clinical stages of the disease.

12.- The different ‘omic’ approaches presented in this thesis are complementary to each other and necessary to obtain a complete picture of the composition, functional potential and activity of the microbiome.

6

SHORT SPANISH VERSION

APROXIMACIONES "-ÓMICAS" PARA EL ESTUDIO DEL MICROBIOMA ORAL

Introducción

La visión clásica de los microorganismos como agentes infecciosos ha propiciado que se consideren como meros agentes causales de enfermedades. Sin embargo, la larga historia de coexistencia entre microorganismos y los seres humanos, ha hecho que ambos se hayan adaptado mutuamente y coevolucionen (Dubos et al. 1965, McFall-Ngai 2002). De hecho, en el cuerpo humano habitan 10^{14} células bacterianas, un orden de magnitud más que células humanas (Savage 1977). Entre los beneficios que la microbiota aporta, están su contribución a digerir alimentos y a aportar nutrientes, regulación del metabolismo, maduración del sistema inmune, etc. Por ello, hoy en día las enfermedades de origen bacteriano no solo se consideran aquellas en las que un único agente causal produce una infección o produce toxinas, sino también aquellas enfermedades causadas por la pérdida de sus efectos beneficiosos. Entre las enfermedades que transcurren con una alteración de la microbiota están la obesidad (Turnbaugh & Gordon 2009, Turnbaugh et al. 2009, Ferrer et al. 2013), enfermedad inflamatoria intestinal (Seksik 2010), vaginosis bacteriana (Ravel et al. 2013), caries dental (Marsh 2010), periodontitis (Kumar et al. 2006), etc, y por tanto su abordaje terapéutico no puede realizarse de una manera clásica, como por ejemplo mediante el uso de antibióticos o vacunas dirigidas al agente causal.

Por ello, el estudio de la microbiota en su conjunto y no limitado a ciertas especies patógenas, es fundamental para entender este nuevo tipo de enfermedades. Las técnicas tradicionales basadas en cultivo, tienen la gran limitación de no ser capaces de aislar a la gran mayoría de microorganismos, tal y como ya observó en 1932 Razumov, quien definió la "gran anomalía del recuento en placa" el hecho de que al cultivar en una placa petri una muestra, tan solo crecían un pequeño número de las bacterias observadas al microscopio (Razumov 1932). El desarrollo a finales de los años 90 de técnicas moleculares, permitieron salvar parcialmente las limitaciones del cultivo. Así se observó que prácticamente todas las superficies y cavidades del cuerpo humano están colonizadas por bacterias. Se han encontrado bacterias incluso partes del cuerpo que se creían estériles, tales como el hígado, leche materna, placas de ateroma, placenta, tejido adiposo o en la sangre bajo ciertas condiciones (Burcelin et al. 2013).

Debido a esta ubicuidad de las bacterias incluso en ausencia de enfermedad, es necesario poder distinguir entre un microbioma sano y uno enfermo, para así poder diagnosticar aquellas enfermedades que transcurren junto con una disbiosis de la comunidad bacteriana, incluso antes de que los síntomas clínicos aparezcan. Grandes consorcios internacionales están llevando a cabo estudios de la microbiota de un gran número de individuos sanos, como por ejemplo el Proyecto del Microbioma Humano estadounidense (HMP 2012) o el Proyecto META-HIT europeo (Qin et al. 2010).

Cavidad oral

Uno de los nichos del microbioma humano más complejo es la cavidad oral, la cual alberga una gran variedad de subnichos y en consecuencia una enorme diversidad microbiana. Está delimitada por los labios, las mejillas, el paladar (dividido en paladar duro y blando), la lengua y el suelo de la boca.

Los dientes son estructuras mineralizadas, que están anatómicamente divididas en dos partes, la corona, la parte visible del diente, y la raíz, que es la parte que se inserta en los alveolos dentales. El diente está compuesto por 3 capas diferenciadas, esmalte (cemento en las raíces), dentina y pulpa. El esmalte es la capa más externa de la corona y es el material más duro del cuerpo humano. Está compuesto en un 96% de material inorgánico, principalmente hidroxiapatita. El cemento es la capa más externa de la raíz y su contenido inorgánico es menor (45%), ya que en él se insertan los ligamentos periodontales, haciendo que un 33% sea material orgánico y un 22% de agua.

La dentina tiene un 70% de hidroxiapatita y contiene túbulos dentinarios que albergan una matriz de proteínas colágenas y prolongaciones celulares de los odontoblastos (Goldberg et al. 2011). Aunque no está vascularizada, debido a su estructura tubular y contenido celular hace que sea un tejido vivo, capaz de reaccionar frente a estímulos externos. La pulpa es la parte central del diente, y es la única parte del diente que está vascularizada e inervada. Está compuesta principalmente de tejido conectivo y de los cuerpos celulares de los odontoblastos, los cuales se prolongan a través de los túbulos dentinarios.

Las glándulas salivares juegan un importante papel en la cavidad oral, que continuamente secretan saliva. La saliva es un fluido acuoso con una pequeña proporción de compuestos disueltos, como moco, glicoproteínas, electrolitos, enzimas y compuestos antibacterianos (Amerongen & Veerman 2002). Sus principales funciones son participar en la digestión de grasas y almidón, lubricación de las superficies mucosas y comida para facilitar su deglución, regulación de temperatura y humedad, defensa frente a infecciones, tampón variaciones de pH y control del balance de mineralización-desmineralización del diente. La saliva, al entrar en contacto con las diferentes superficies de la boca, forma una fina capa de unos 8-40 µm llamada película adquirida. Las glicoproteínas presentes en esta película son utilizadas por las bacterias como anclaje para poder adherirse a la superficie del diente, evitando ser arrastradas por la saliva y ser deglutidas (Jenkinson & Lamont 2005).

Comunidades microbianas orales

Todas las bacterias que viven en la cavidad oral han de ser capaces de adherirse a alguna superficie, para así evitar ser arrastradas junto con la saliva hacia el estómago. Por ello, prácticamente todas las superficies de la boca son susceptibles de ser colonizadas por biofilms bacterianos. La microbiota presente en la saliva, no es considerada como habitante oral, sin embargo su composición es fruto de la descamación de los biofilms de diversas partes de la boca. Además, debida a la gran variedad de condiciones presentes, la composición microbiana varía entre los diferentes micronichos orales (Zaura et al. 2009, Segata et al. 2012, Simón-Soro et al. 2013a).

Entre las enfermedades orales, caries, gingivitis y periodontitis son las más comunes causadas por microorganismos. Estas enfermedades requieren la formación y acumulación de placa dental, que es un biofilm bacteriano que crece sobre la superficie del diente. Está formado por células bacterianas, fúngicas y epiteliales descamadas, así como glicoproteínas salivares y polisacáridos y proteínas secretadas por microorganismos (Tinanoff & Gross 1976, Mosby 2013). La placa supragingival se acumula sobre la superficie visible del diente y favorece el crecimiento de bacterias acidogénicas y acidófilas, y es la causante de la caries dental. La placa subgingival se acumula en el surco subgingival, ente la encía y el diente, es un ambiente neutro o ligeramente alcalino y está principalmente compuesto de bacterias Gram negativas.

La placa subgingival se caracteriza por desarrollarse en un ambiente típicamente anaeróbico, con un pH neutro o alcalino, y cuya principal fuente de nutrientes es el líquido crevicular y las células descamadas epiteliales. La acumulación de placa subgingival provoca inflamación en las encías, que puede progresar a periodontitis, inflamándose los tejidos de soporte del diente, perdida de hueso alveolar y potencialmente, la pérdida de la pieza dental (Kawar et al. 2011). Las especies bacterianas asociadas tradicionalmente a gingivitis y periodontitis son las que componen el complejo rojo, *Tannerella forsythia*, *Treponema denticola* y *Porphyromonas gingivalis (Socransky et al. 1998)*. Sin embargo, estudios más recientes también sugieren la implicación de un mayor número de especies en esta patología (Kumar et al. 2006, Fritschi et al. 2008, Colombo et al. 2009, Griffen et al. 2012). Esto hace que el origen de la gingivitis/periodontitis sea polimicrobiano.

La placa supragingival crece sobre las superficies del diente no cubiertas por encía. Las bacterias que viven en este nicho, han de ser capaces de soportar las fuerzas mecánicas de los movimientos de la lengua, saliva, mejillas y masticación, mediante su adhesión. Prácticamente todos los habitantes de la placa supragingival son capaces de adherirse a la película adquirida o a otro de los miembros de la placa (Kolenbrander et al. 2006). Aunque en principio es un ambiente aeróbico, si el biofilm adquiere cierto grosor, el oxígeno puede ser consumido por las bacterias aeróbicas, creando zonas de microaerofilia dentro del biofilm, favoreciéndose el crecimiento de anaerobios. Las principales fuentes de nutrientes son las glicoproteínas salivares, restos de comida y los componentes celulares desprendidos del epitelio. La actividad metábolica de este biofilm es responsable de la acidificación que causa la desmineralización del esmalte, iniciándose las lesiones de caries.

Caries dental

La caries dental es la patología derivada de la disolución de la superficie orgánica y mineral del diente, causada esta última por los ácidos producidos por los microorganismos de la placa dental supragingival (Fejerskov et al. 2008). En condiciones de pH neutro, hay un equilibrio entre mineralización y desmineralización del esmalte, ya que la película adquirida sobre éste está saturada de hidroxiapatita. Cuando el pH baja fruto del metabolismo microbiano de azúcares fermentables, la solubilidad de la hidroxiapatita aumenta y se favorece la desmineralización del esmalte. Una vez los azúcares dejan de estar disponibles, la capacidad de tampón de la saliva hace aumentar de nuevo el pH, reestableciéndose de nuevo el equilibrio. Además, la película adquirida, sobresaturada de iones fosfato y calcio, favorece su precipitación y la remineralización del esmalte. La caries dental se inicia cuando estos ciclos de desmineralización-remineralización se descompensan y aumenta la desmineralización, dándose una pérdida neta de mineral.

Gracias a todo el conocimiento acumulado, diferentes teorías se han ido planteando sobre la contribución microbiana a la caries dental. Una de las primeras fue la hipótesis de la placa no específica planteada por Miller (Miller 1890). En ella sugería que los ácidos producidos por el conjunto de las bacterias presentes en la placa dental al proporcionarles azúcar o almidón eran los responsables de la degradación del esmalte, y no eran fruto de una única especie. También propuso que la caries era una enfermedad con dos etapas, la primera caracterizada por la disolución ácida del esmalte, y la segunda caracterizada por la descomposición de las "sustancias albuminosas" de la dentina, una vez esta queda expuesta. Más tarde, se propuso la hipótesis específica de la placa (Loesche 1986), coincidiendo con el descubrimiento de *Streptococcus mutans*. Fitzgerald y Keyes demostraron que esta especie acidogénica era capaz de producir lesiones de caries por sí sola en hamsters albinos (Fitzgerald & Keyes 1960). Otras especies propuestas como patógenas siguiendo esta hipótesis específica son *Streptococcus sobrinus*, *Lactobacillus casei* y *Actinomyces odontolyticus* (Loesche 1986). La hipótesis de la placa ecológica propuesta por Marsh (Marsh 1994b), plantea que la caries tiene lugar cuando una fuente de estrés para el ecosistema de la placa dental, como por ejemplo un mayor aporte de azúcares en la dieta, produce un cambio ambiental en el nicho, una bajada de pH, fruto de la fermentación de éstos. Esta bajada de pH, además de aumentar el riesgo de caries, favorece un cambio en la comunidad bacteriana en la placa dental, debido a la selección de especies acidófilas y acidogénicas. Esto termina por retroalimentar una mayor producción de ácido y por tanto aumentar el riesgo de caries. Por tanto, se considera que la caries acontece junto con una disbiosis de la microbiota, causante última de la caries.

Problemas para estudiar la caries dental

Aunque la caries dental ha estado presente en el ser humano durante miles de años, no se ha podido encontrar niguna cura efectiva. Entre los múltiples motivos que dificultan el estudio de esta patología se encuentran la falta de consenso sobre la etiología de la enfermedad, a la vista de las múltiples teorías que se han ido planteando a lo largo del tiempo (Rosier et al. 2014). El hecho de ser una enfermedad polimicrobiana, no cumplir los postulados de Koch clásicos (Koch 1870), tener un largo período de desarrollo hasta que las lesiones son detectables, ha dificultado su abordaje. Esto unido al gran número factores que influyen en la aparición de la enfermedad, ha hecho que los esfuerzos realizados no hayan sido capaces de encontrar un tratamiento para evitar la caries dental.

Por otro lado, las técnicas clásicas de microbiología basadas en el cultivo de cepas ha dificultado el conocimiento exhaustivo del microbioma oral. Aún siendo éste uno de los ecosistemas con más miembros cultivables, tan solo ha sido posible cultivar alrededor del 50% de todos sus habitantes (Paster et al. 2001, Wade 2002, Donachie et al. 2007, Marsh et al. 2011). Esto ha proporcionado una visión sesgada de este ecosistema, estudiándose únicamente aquellas especies cultivables. La introducción de técnicas moleculares como la clonación, DGGE, microarrays de DNA y RNA y secuenciación del gen 16S rRNA, han mejorado el conocimiento acerca de la diversidad taxonómica en la placa dental, pero presentan un gran número de sesgos y limitaciones (Nyvad et al. 2013). Las técnicas clásicas moleculares no permiten conocer por ejemplo el conjunto de funciones que una comunidad bacteriana puede llevar a cabo, los genes que se están transcribiendo en determinadas circunstancias o el contenido proteico en su conjunto.

Nuevas técnicas para estudiar la caries dental

Como consecuencia de estos factores que influyen en la aparición de caries, se necesitan nuevas herramientas que permitan superar las limitaciones existentes. En la última década se han introducido técnicas de alto rendimiento que permiten superar las limitaciones del cultivo, mediante el estudio de moléculas informativas de las comunidades microbianas, como el DNA, RNA, proteínas y metabolitos (Zoetendal et al. 2008, Nyvad et al. 2013).

El análisis de los ácidos nucleicos ha vivido una revolución con la introducción de las técnicas de secuenciación de segunda generación, que han permitido aumentar la cantidad de bases secuenciadas por carrera. A la vez se ha disminuido el tiempo necesario y el coste de obtener la misma cantidad de secuencias. Esto ha permitido el uso de la secuenciación directa tanto de DNA y RNA, obviando la necesidad de cultivo y otras limitaciones asociadas a las técnicas moleculares clásicas como el DGGE, amplificación y secuenciación del gen 16S rRNA, etc.

La metagenómica se desarrolló para poder estudiar los genomas presentes en el conjunto de una comunidad bacteriana, y así poder obtener mayor información que con las técnicas basadas en un solo gen marcador filogenético (16S rRNA). De esta manera se puede obtener información sobre la composición taxonómica de la comunidad analizada, así como del total de funciones que potencialmente pueden ser llevadas a cabo. La metatranscriptómica, mediante el análisis de las moléculas de RNA, da a conocer qué especies y genes están transcripcionalmente activos en un momento determinado, lo cual permite conocer las adaptaciones del conjunto de la comunidad bacteriana ante diferentes condiciones ambientales. Por último, la metaproteómica permite conocer las proteínas presentes en la comunidad, reflejando de forma más cercana la actividad funcional que se lleva a cabo.

Objetivos

El estudio de las comunidades microbianas asociadas al cuerpo humano han experimentado un gran auge desde el inicio del siglo XXI, debido a la introducción de nuevas técnicas de alto rendimiento y eficiencia. Por lo tanto ha permitido estudiar la relación existente entre muchas enfermedades y las comunidades microbianas que poseen los pacientes. A lo largo de esta tesis, algunas de estas nuevas técnicas se han aplicado al estudio de la microbiota de la placa dental humana, para responder a múltiples cuestiones biológicas relativas a la caries dental. Técnicas de secuenciación de alto rendimiento y de metagenómica se han utilizado para salvar las limitaciones inherentes a las técnicas tradicionales basadas en el cultivo de cepas, en un único gen como el 16S rRNA o técnicas de clonación, con el objetivo de descifrar la composición microbiana en estados de salud y enfermedad. El hilo conductor de esta tesis ha sido profundizar en el conocimiento disponible sobre el microbioma en su conjunto y en la porción transcripcionalmente activa, bajo diferentes estados de salud y enfermedad, con el objetivo profundizar en la etiología de la caries dental y proponer estrategias de prevención y diagnóstico. Este objetivo global se subdivide en los siguientes objetivos específicos:

- Caracterización taxonómica de la comunidad microbiana presente en la placa dental de individuos sanos y con caries, mediante el uso de técnicas no limitadas de antemano ni sesgadas por metodologías como el cultivo, la amplificación mediante PCR o la clonación.
- Caracterización del potencial genético de los microorganismos habitantes de la placa dental supragingival. Esto debe determinar si las diferencias taxonómicas entre individuos sanos y con caries también se corresponden con diferencias funcionales.
- Detección de bacterias potencialmente protectoras frente a la caries procedentes de individuos sanos, que puedan ser susceptibles de desarrollo como probióticos anticaries.
- Proponer posibles genes de virulencia en los genomas de bacterias patógenas, mediante su comparación con muestras metagenómicas de voluntarios sanos, obtenidas del mismo ecosistema de donde se aisló al patógeno en cuestión.
- Comparar la comunidad microbiana encontrada mediante metagenómica en los biofilms dentales con la fracción transcripcionalmente activa, para poder diferenciar entre las bacterias de paso de las activas en este nicho.
- Caracterizar las especies transcripcionalmente activas durante la bajada de pH que tiene lugar sobre la superficie del diente después de la ingesta de comidas ricas en

carbohidratos, para descubrir las bacterias activas durante la metabolización de azúcares y producción de ácido, potencialmente responsables de la caries dental.

- Comparar la composición del microbioma oral mediante técnicas metagenómicas, metatranscriptómicas y metaproteómicas.
- Describir por primera vez el catálogo de proteínas humanas y bacterianas presentes en la placa dental supragingival.
- Buscar posibles biomarcadores de salud y enfermedad en caries dental que puedan ser potencialmente utilizados en un test diagnóstico.

Resultados y conclusiones

En el capítulo 3.1 de la presente tesis, investigamos las diferencias a nivel taxonómico y funcional de muestras de placa dental de voluntarios sanos, con caries y muestras obtenidas de lesiones avanzadas de caries, mediante la secuenciación directa del DNA extraído (Alcaraz et al. 2012, Belda-Ferre et al. 2012). El uso de la metagenómica nos permitió observar que la composición taxonómica, aún siendo similar a otros estudios previos basados en técnicas diferentes, presentaban ciertas diferencias, como un mayor número de géneros minoritarios no detectados por técnicas de PCR del gen 16S rRNA. Al comparar las muestras de individuos sanos con muestras de pacientes con caries, observamos diferencias en la composición taxonómica, corroborando que la caries dental está asociada a un cambio en la microbiota a nivel taxonómico, de acuerdo con la hipótesis ecológica de la placa. Además, al disponer de secuencias procedentes del conjunto de genomas presentes en las muestras, pudimos observar que las diferencias no se debían únicamente a la presencia de diferentes géneros o especies, sino también a diferentes cepas de la misma especie.

Tras observar que las personas sanas eran portadoras de cepas de *Streptococcus* diferentes a las presentes en personas con caries, aislamos 192 cepas. De éstas, se seleccionaron varios aislados debido a su potencial como probiótico protector frente a caries dental, dos de las cuales se describieron como la nueva especie *Streptococcus dentisani* (Camelo-Castillo et al. 2014), y en la actualidad están siendo desarrolladas para su comercialización. Este capítulo muestra por tanto cómo la metagenómica puede ayudar a identificar microorganismos con potencial probiótico.

En cuanto a la composición funcional, se pudo comprobar que la microbiota intestinal y de la placa dental, no solo se diferencian a nivel taxonómico, sino que también el repertorio funcional de la microbiota de estos dos ecosistemas es diferente. Al poder asignar a cada lectura de DNA tanto una afiliación taxonómica como funcional, se pudo relacionar qué funciones eran llevadas a cabo por cada grupo taxonómico. Se encontraron diferencias entre sujetos sanos y con caries, que incluían funciones sobrerrepresentadas en sujetos sanos, en concreto funciones relacionadas con péptidos antibacterianos, genes de respuesta a estrés periplásmico y polisacáridos extracelulares. En los individuos con caries activas, se encontró que estaban sobrerrepresentadas las funciones relacionadas con fermentación ácida, incorporación de DNA y competencia.

Por último, este trabajo fue el primero en secuenciar de forma directa DNA bacteriano procedente de lesiones de caries. Esto permitió observar que son ecosistemas complejos en el que habita una gran diversidad de especies, y en el que *S. mutans* se muestra prácticamente ausente. En estudios posteriores se pudo confirmar que *S. mutans* estaba presente en lesiones de mancha blanca, aunque su proporción siempre era menor al 1% (Simón-Soro et al. 2014). Estos resultados sugieren que la caries está causada por un conjunto de bacterias diferentes a las encontradas en personas sanas, aunque no existe una única combinación de bacterias cariogénicas.

En el capítulo 3.2 se propuso un método basado en la metagenómica para poder detectar posibles genes de virulencia presentes en cepas patógenas (Belda-Ferre et al. 2011). Éste consiste en comparar el genoma de la cepa patógena en cuestión frente a un metagenoma de una persona sana, obtenido del mismo lugar donde se aisló la cepa. Así las regiones del genoma que no están presentes en el metagenoma, llamadas islas metagenómicas, suelen contener genes hipervariables como los que codifican proteínas expuestas de la pared celular, sometidas a presión selectiva debido al ataque de fagos o del sistema inmune. En el caso de bacterias patógenas, cuando el potencial patógeno es debido a un gen o grupo de genes (p.ej. toxinas), estos suelen estar ausentes en las cepas no patógenas, haciendo que la comparación entre cepas patógenas y no patógenas sea una estrategia para la búsqueda de estos genes de virulencia. Sin embargo, el gran número de genes no compartidos por cepas de la misma especie, el “pangenoma accesorio” (D’Auria et al. 2010), hace necesaria la comparación de múltiples cepas para poder encontrar genes de patogenicidad potenciales (Ho Sui et al. 2009, D’Auria et al. 2010, Hilker et al. 2014). Esta estrategia está limitada por la necesidad de aislar suficientes cepas, lo que podría estar sesgado hacia aquellas cepas mejor adaptadas al crecimiento en condiciones de laboratorio. Como alternativa, la metagenómica permite comparar el conjunto de cepas presentes en un nicho obviando el paso de cultivo frente a la cepa patogénica en cuestión mediante los gráficos de reclutamiento. Esto facilita la detección de candidatos a genes de virulencia cuando se detecta un brote infeccioso y se requieren aplicar medidas epidemiológicas preventivas, como el caso del brote por *E. coli* O104:H4 enterohemorrágica (Qin et al. 2011, Ahmed et al. 2012). Este capítulo prueba que el método es fiable al identificar claramente genes de virulencia conocidos, y propone nuevos candidatos en bacterias orales e intestinales.

En el capítulo 3.3 se abordaron dos de los procesos de los que depende la aparición de caries, la formación de biofilm y la producción de ácido por la fermentación de carbohidratos, mediante una aproximación de metatranscriptómica (Benítez-Páez et al. 2014). La formación del biofilm de la placa dental se ha estudiado tradicionalmente mediante el uso de modelos *in vitro* (Kolenbrander et al. 2006) o con técnicas de microarrays de DNA (Li et al. 2004, Teles et al. 2012). Sin embargo, no se conoce mucho en condiciones *in vivo*, ni de la actividad funcional a lo largo de este proceso. Este trabajo se planteó para intentar conocer qué bacterias de las que se adhieren al diente son activas así como el patrón de expresión de éstas en la placa dental supragingival durante diferentes etapas de su formación tras una limpieza dental profesional (6, 12, 24 y 48 horas). Así se pudo observar una gran diversidad de especies activas en todas las etapas analizadas, incluyendo correlaciones en la actividad de varios habitantes orales, tanto positivas (mutualistas) como negativas (potencialmente antagonistas). Esto sugiere que la adhesión al biofilm puede no ocurrir mediante una sucesión ecológica, o que ésta tiene lugar en períodos de tiempo más cortos que los propuestos hasta la fecha. Una posible explicación es que las células bacterianas que forman agregados presentes en la saliva o adheridos a células epiteliales, se adhieren a la vez a la superficie del diente (Cabello Yeves 2014). En función de las condiciones ambientales del biofilm, aquellos que estén mejor adaptados a ellas mostrarán mayor actividad transcripcional.

En cuanto a los tránscritos de mRNA, encontramos diferentes patrones de expresión entre las muestras de biofilm tempranas (6 y 12 horas) y las tardías (24 y 48 horas). Entre las 35 categorías sobre-expresadas en las muestras tempranas se encuentran genes relacionados con funciones housekeeping y de aprovechamiento de recursos, que sugieren una reducida competencia en los momentos iniciales de formación del biofilm. Sin embargo, en las muestras tardías se encuentran sobre-expresadas funciones de motilidad celular, transportadores ABC, sistemas de reparación de DNA y de recombinación homóloga, reflejando la importancia de la comunicación entre células y la detección de señales del ambiente, lo que puede indicar una mayor competencia por los nutrientes limitados en el biofilm maduro.

La otra cuestión abordada en el capítulo 3.3 fue determinar qué habitantes de la placa dental supragingival eran transcripcionalmente más activos durante la bajada de pH que sucede tras la ingesta de una comida rica en carbohidratos, para así poder identificar los posibles miembros productores de ácido o ácido tolerantes, y por tanto, posibles contribuyentes a la etiología de la caries. Uno de los resultados observados fue una gran dominancia a nivel transcripcional de los géneros *Actinomyces*, *Corynebacterium* y *Rothia*, a los que correspondía más del 80% de los tránscritos. Usando otras aproximaciones no basadas en RNA su contribución porcentual al total de la comunidad microbiana es mucho menor (capítulo 3.1), lo cual demuestra la necesidad de aplicar diferentes técnicas para poder tener una mejor visión global de cómo funciona un ecosistema. En cuanto a los cambios observados entre las muestras de antes y después de la ingesta de la comida rica en carbohidratos, no se observó un patrón específico. Sin embargo se detectó en algunos individuos una mayor resiliencia en la comunidad activa frente a los cambios de pH, mientras que en otros los cambios fueron muy pronunciados. Además se observó una mayor actividad transcripcional de *Actinomyces* en las personas sanas, que podría ser un factor protector frente a la acidificación del biofilm, aunque no se pudo determinar a qué especie en concreto se correspondía.

En el último capítulo de esta tesis (Capítulo 3.4), se realizó el primer estudio metaproteómico aplicado a la placa dental supragingival humana. Los objetivos principales eran describir el conjunto de proteínas traducidas en la placa dental, así como indagar en las diferencias entre personas sanas y con caries, para proponer dichas diferencias como posibles marcadores de salud y enfermedad. En la primera aproximación, se utilizó un prefraccionamiento HILIC de las proteínas digeridas con tripsina obtenidas de un total de 17 muestras combinadas, seguido de nanoLC-MS/MS. Con esta aproximación conseguimos detectar un total de 7771 proteínas bacterianas y 853 humanas, observándose una gran dominancia con diferencias de hasta 3 órdenes de magnitud entre las proteínas identificadas a mayor y menor concentración.

En cuanto a las diferencias entre muestras en estado de salud y enfermedad, procedimos a cuantificar de manera individual las proteínas presentes en las muestras de placa dental supragingival de 17 voluntarios (9 con caries y 8 sin caries). En esta aproximación se pudieron detectar más de 1000 proteínas por muestra de media. Entre las proteínas con mayor abundancia en personas sanas que en pacientes con caries, se encuentran enzimas relacionadas con mecanismos que afectan al pH del biofilm (L-lactato deshidrogenasa y ornitina carbamoiltransferasa). En el caso de sujetos con caries, se observó una mayor abundancia de proteínas relacionadas con sistemas PTS, transportadores de disacáridos y N-acetilglucosamina-6-fosfato desacetilasa, sugiriendo mejores capacidades para explotar los azúcares disponibles y su fermentación. En cuanto a las 127 proteínas humanas encontradas y cuantificadas, también se encontraron diferencias en abundancia entre personas sanas y enfermas, incluyendo proteínas relacionadas con epitelio queratinizado, proteínas de respuesta inmune del huésped, metabolismo de hierro y proteasas e inhibidores de proteasas entre otros.

Teniendo en cuenta las diferencias observadas, se intentó buscar aquellas proteínas que permitieran diferenciar a personas sanas de las enfermas. En principio, mediante la aplicación de una prueba univariante de T, se encontraron 53 proteínas bacterianas y 29 humanas, lo cual es muy prometedor, pero probablemente inviable para su aplicación en un test diagnóstico comercial. Por ello se aplicó un análisis multivariante (Trevino & Falciani 2006) para poder detectar el mínimo número de proteínas que fueran capaces de diferenciar si una muestra procede de un individuo con o sin caries de la forma más fiable posible. Así se obtuvo un conjunto de 6 proteínas bacterianas y 4 humanas, capaces de diferenciar entre sanos y enfermos con una especificidad y sensitividad superior al 96%. Futuros estudios deberán confirmar la validez de estas proteínas para diagnosticar de forma temprana el riesgo de padecer caries, con el fin de poder establecer medidas preventivas antes de la aparición de signos clínicos de esta enfermedad.

7

BIBLIOGRAPHY

BIBLIOGRAPHY

Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. 2014. The placenta harbors a unique microbiome. *Sci. Transl. Med.* 6(237):237ra65

Aas JA, Griffen AL, Dardis SR, Lee AM, Olsen I, et al. 2008. Bacteria of dental caries in primary and permanent teeth in children and young adults. *J. Clin. Microbiol.* 46(4):1407–17

Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* 43(11):5721–32

Abusleme L, Dupuy AK, Dutzan N, Silva N, Burleson J a, et al. 2013. The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *ISME J.* 7(5):1016–25

Acevedo AM, Machado C, Rivera LE, Wolff M, Kleinberg I. 2005. The inhibitory effect of an arginine bicarbonate/calcium carbonate CaviStat-containing dentifrice on the development of dental caries in Venezuelan school children. *J. Clin. Dent.* 16(3):63–70

Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J. Bacteriol.* 186(9):2629–35

Acton RT, Dasanayake AP, Harrison RA, Li Y, Roseman JM, et al. 1999. Associations of MHC genes with levels of caries-inducing organisms and caries severity in African-American women. *Hum. Immunol.* 60(10):984–89

Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, et al. 2013. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat. Genet.* 45(4):450–55, 455e1

Ahmadi E, Fallahi S, Alaeddini M, Hasani Tabatabaei M. 2013. Severe dental caries as the first presenting clinical feature in primary Sjögren’s syndrome. *Casp. J. Intern. Med.* 4(3):731–34

Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, et al. 2012. Genomic comparison of Escherichia coli O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2. *PLoS One.* 7(11):e48228

Al-Hashimi I, Levine MJ. 1989. Characterization of in vivo salivary-derived enamel pellicle. *Arch. Oral Biol.* 34(4):289–95

Alcaraz LD, Belda-Ferre P, Cabrera-Rubio R, Romero H, Simón-Soro Á, et al. 2012. Identifying a healthy oral microbiome through metagenomics. *Clin. Microbiol. Infect.* 18:54–57

Alstrup S., Gavoille C., Kaplan H. RT. 2004. Nearest Common Ancestors: A Survey and a New Algorithm for a Distributed Environment. *Theory Comput. Syst.* 37:441–56

Amerongen AVN, Veerman ECI. 2002. Saliva--the defender of the oral cavity. *Oral Dis.* 8(1):12–22

Aoba T, Moreno EC, Hay DI. 1984. Inhibition of apatite crystal growth by the amino-terminal segment of human salivary acidic proline-rich proteins. *Calcif. Tissue Int.* 36(1):651–58

Aufderheide AC, Rodriguez-Martin C. 2011. *The Cambridge Encyclopedia of Human Paleopathology*

Bánóczy J, Rugg-Gunn A, Woodward M. 2013. Milk fluoridation for the prevention of dental caries. *Acta Med. Acad.* 42(2):156–67

Bardow A, Moe D, Nyvad B, Nauntofte B. 2000. The buffer capacity and buffer systems of human whole saliva measured without loss of CO2. *Arch. Oral Biol.* 45(1):1–12

Bartold PM, Van Dyke TE. 2013. Periodontitis: a host-mediated disruption of microbial homeostasis. Unlearning learned concepts. *Periodontol. 2000.* 62(1):203–17

Becker MR, Paster BJ, Leys EJ, Moeschberger ML, Kenyon SG, et al. 2002. Molecular analysis of bacterial species associated with childhood caries. *J. Clin. Microbiol.* 40(3):1001–9

Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simón-Soro A, et al. 2012. The oral metagenome in health and disease. *ISME J.* 6(1):46–56

Belda-Ferre P, Cabrera-Rubio R, Moya A, Mira A. 2011. Mining virulence genes using metagenomics. *PLoS One.* 6(10):e24975

Benítez-Páez A, Álvarez M, Belda-Ferre P, Rubido S, Mira A, Tomás I. 2013. Detection of Transient Bacteraemia following Dental Extractions by 16S rDNA Pyrosequencing: A Pilot Study. *PLoS One.* 8(3):e57782

Benítez-Páez A, Belda-Ferre P, Simón-Soro A, Mira A. 2014. Microbiota diversity and gene expression dynamics in human oral biofilms. *BMC Genomics.* 15(1):311

Bennett S. 2004. Solexa Ltd. *Pharmacogenomics.* 5(4):433–38

Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, et al. 2010. Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J.* 4(8):962–74

Björkholm B, Bok CM, Lundin A, Rafter J, Hibberd ML, Pettersson S. 2009. Intestinal microbiota regulate xenobiotic metabolism in the liver. *PLoS One.* 4(9):e6958

Bjørndal L. 1992. Carieslæsionens tidlige udvikling i emalje og pulpa-dentinorganet. *Tandlægebladet.* 96(11):469–72

Bjørndal L, Darvann T, Lussi A. 1999. A computerized analysis of the relation between the occlusal enamel caries lesion and the demineralized dentin. *Eur. J. Oral Sci.* 107:176–82

Blakey K, Feltbower RG, Parslow RC, James PW, Gómez Pozo B, et al. 2014. Is fluoride a risk factor for bone cancer? Small area analysis of osteosarcoma and Ewing sarcoma diagnosed among 0-49-year-olds in Great Britain, 1980-2005. *Int. J. Epidemiol.* 43(1):224–34

Blausen. 2014. Blausen gallery 2014. *Wikiversity J. Med.*

Böök J, Grahnén H. 1953. Clinical and genetical studies of dental caries. II. Parents and sibs of adult highly resistant (caries-free) propositi. *Odontol Rev.* Jan;4(1):1–53

Boraas JC, Messer LB, Till MJ. 1988. A genetic contribution to dental caries, occlusion, and morphology as demonstrated by twins reared apart. *J. Dent. Res.* 67(9):1150–55

Bostanci N, Heywood W, Mills K, Parkar M, Nibali L, Donos N. 2010. Application of label-free absolute quantitative proteomics in human gingival crevicular fluid by LC/MS E (gingival exudatome). *J. Proteome Res.* 9(5):2191–99

Brady A, Salzberg SL. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods.* 6(9):673–76

Brecx M, Rönström A, Theilade J, Attström R. 1981. Early formation of dental plaque on plastic films. 2. Electron microscopic observations. *J. Periodontal Res.* 16(2):213–27

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. 2002. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* 99(22):14250–55

Brook a H. 2009. Multilevel complex interactions between genetic, epigenetic and environmental factors in the aetiology of anomalies of dental development. *Arch. Oral Biol.* 54 Suppl 1:S3–17

Burcelin R, Serino M, Chabo C, Garidou L, Pomié C, et al. 2013. Metagenome and metabolism: the tissue microbiota hypothesis. *Diabetes, Obes. Metab.* 15(s3):61–70

Cabello Yeves PJ. 2014. *Arquitectura microbiana de la saliva mediante aproximaciones de citometría de flujo y pirosecuenciación.* Universidad Politécnica de Valencia

Cabrera-Rubio R. 2014. *Análisis Taxonómico y Funcional del Microbioma Humano mediante Aproximaciones Clásicas Moleculares y Metagenómicas*. Universitat de Valencia

Cabrera-Rubio R, Collado MC, Laitinen K, Salminen S, Isolauri E, Mira A. 2012. The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery 1 – 4. . 544–51

Camelo-Castillo A, Benítez-Páez A, Belda-Ferre P, Cabrera-Rubio R, Mira A. 2014. Streptococcus dentisani sp. nov., a novel member of the mitis group. *Int. J. Syst. Evol. Microbiol.* 64(Pt 1):60–65

Chmurzynska A. 2010. Fetal programming: link between early nutrition, DNA methylation, and complex diseases. *Nutr. Rev.* 68(2):87–98

Chung H, Pamp SJ, Hill J a, Surana NK, Edelman SM, et al. 2012. Gut immune maturation depends on colonization with a host-specific microbiota. *Cell.* 149(7):1578–93

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4(4):265–70

Clarke JK. 1924. On the bacterial factor in the aetiology of dental caries. *Br J Exp Pathol.* 5(3):141–47

Colombo AP V., Boches SK, Cotton† SL, Goodson JM, Kent R, et al. 2009. Comparisons of Subgingival Microbial Profiles of Refractory Periodontitis, Severe Periodontitis and Periodontal Health using the Human Oral Microbe Identification Microarray (HOMIM). *J. Periodontol.* 80(9):1421–32

Corby PM, Lyons-Weiler J, Bretz WA, Hart TC, Aas JA, et al. 2005. Microbial risk indicators of early childhood caries. *J. Clin. Microbiol.* 43(11):5753–59

Cordain L, Eaton SB, Sebastian A, Mann N, Lindeberg S, et al. 2005. Origins and evolution of the Western diet: health implications for the 21st century. *Am J Clin Nutr.* 81(2):341–54

Crielaard W, Zaura E, Schuller A a, Huse SM, Montijn RC, Keijser BJF. 2011. Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. *BMC Med. Genomics.* 4(1):22

Cruaud P, Vigneron A, Lucchetti-Miganeh C, Ciron PE, Godfroy A, Cambon-Bonavita M-A. 2014. Influence of DNA extraction method, 16S rRNA targeted hypervariable regions, and sample origin on microbial diversity detected by 454 pyrosequencing in marine chemosynthetic ecosystems. *Appl. Environ. Microbiol.* 80(15):4626–39

Cuadros-Orellana S, Martin-Cuadrado A-B, Legault B, D'Auria G, Zhaxybayeva O, et al. 2007. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J.* 1(3):235–45

D'Auria G, Jiménez-Hernández N, Peris-Bondia F, Moya A, Latorre A. 2010. Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics.* 11:181

Darveau RP. 2009. The oral microbial consortium's interaction with the periodontal innate defense system. *DNA Cell Biol.* 28(8):389–95

Darveau RP. 2010. Periodontitis: a polymicrobial disruption of host homeostasis. *Nat. Rev. Microbiol.* 8(7):481–90

David L a., Maurice CF, Carmody RN, Gootenberg DB, Button JE, et al. 2013. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*

De Boever E, Loesche W. 1995. Assessing the contribution of anaerobic microflora of the tongue to oral malodor. *J. Am. Dent. Assoc.* 126(10):1384–93

Delisle AL, Rostkowski CA. 1993. Lytic bacteriophages ofStreptococcus mutans. *Curr. Microbiol.* 27(3):163–67

Deutscher J, Francke C, Postma PW. 2006. How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiol. Mol. Biol. Rev.* 70(4):939–1031

Donachie SP, Foster JS, Brown M V. 2007. Culture clash: challenging the dogma of microbial diversity. *ISME J.* 1(2):97–99

Dongari-Bagtzoglou A. 2008. Mucosal biofilms: challenges and future directions. *Expert Rev. Anti. Infect. Ther.* 6(2):141–44

Dubos R, Schaedler RW, Costello R, Hoet P. 1965. Indigenous, normal, and autochtonous flora of the gastrointestinal tract. *J. Exp. Med.* 122(1):67–77

Eckert R, Sullivan R, Shi W. 2012. Targeted antimicrobial treatment to re-establish a healthy microbial flora for long-term protection. *Adv. Dent. Res.* 24(2):94–97

Edelstein BL. 2006. The dental caries pandemic and disparities problem. *BMC Oral Health.* 6 Suppl 1:S2

Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science.* 323(5910):133–38

Ekstrand KR, Ricketts DN, Kidd E a. 1998. Do occlusal carious lesions spread laterally at the enamel-dentin junction? A histolopathological study. *Clin. Oral Investig.* 2(1):15–20

Fejerskov O. 2004. Changing paradigms in concepts on dental caries: consequences for oral health care. *Caries Res.* 38(3):182–91

Fejerskov O, Kidd E, Nyvad B, Baelum V. 2008. *Dental Caries The Disease and Its Clinical Management*. Oxford: Blackwell Munksgaard

Ferrer M, Ruiz A, Lanza F, Haange S-B, Oberbach A, et al. 2013. Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environ. Microbiol.* 15(1):211–26

Fidalgo TK da S, Freitas-Fernandes LB, Ammari M, Mattos CT, de Souza IPR, Maia LC. 2014. The relationship between unspecific s-IgA and dental caries: A systematic review and meta-analysis. *J. Dent.* 42(11):1372–81

Fitzgerald RJ, Keyes PH. 1960. Demonstration of the etiologic role of streptococci in experimental caries in the hamster. *J. Am. Dent. Assoc.* 61:9–19

Fons M, Gomez A, Karjalainen T. 2000. Mechanisms of Colonisation and Colonisation Resistance of the Digestive Tract. *Microb. Ecol. Health Dis.* 2:240–46

Frandsen E V. 1994. Carbohydrate depletion of immunoglobulin A1 by oral species of gram-positive rods. *Oral Microbiol. Immunol.* 9(6):352–58

Fredericks DN, Relman DA. 1996. Sequence-based identification of microbial pathogens: a reconsideration of Koch’s postulates. *Clin. Microbiol. Rev.* 9(1):18–33

Fritschi BZ, Albert-Kiszely A, Persson GR. 2008. Staphylococcus aureus and Other Bacteria in Untreated Periodontitis. *J. Dent. Res.* 87(6):589–93

Ganguly S, Mitchell AP. 2011. Mucosal biofilms of Candida albicans. *Curr. Opin. Microbiol.*

14(4):380–85

Gendron R, Plamondon P, Grenier D. 2004. Binding of Pro-Matrix Metalloproteinase 9 by Fusobacterium nucleatum subsp . nucleatum as a Mechanism To Promote the Invasion of a Reconstituted Basement Membrane Binding of Pro-Matrix Metalloproteinase 9 by Fusobacterium nucleatum subsp . nucleatum as a. *Infect. Immun.* 72(10):6160–63

Gentile G, Giuliano L, D'Auria G, Smedile F, Azzaro M, et al. 2006. Study of bacterial communities in Antarctic coastal waters by a combination of 16S rRNA and 16S rDNA sequencing. *Environ. Microbiol.* 8(12):2150–61

Ghai R, Martin-Cuadrado A-B, Molto AG, Heredia IG, Cabrera R, et al. 2010. Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.* 4(9):1154–66

Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, et al. 2012. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* 13(3):R23

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. 2006. Metagenomic analysis of the human distal gut microbiome. *Science.* 312(5778):1355–59

Goldberg M, Kulkarni AB, Young M, Boskey A. 2011. Dentin: structure, composition and mineralization. *Front. Biosci. (Elite Ed).* 3:711–35

Gomar-Vercher S, Cabrera-Rubio R, Mira a, Montiel-Company JM, Almerich-Silla JM. 2014. Relationship of children's salivary microbiota with their caries status: a pyrosequencing study. *Clin. Oral Investig.*

Gomez-Alvarez V, Teal TK, Schmidt TM. 2009. Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* 3(11):1314–17

Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A. 2012. Amplification by PCR Artificially Reduces the Proportion of the Rare Biosphere in Microbial Communities. *PLoS One.* 7(1):e29973

Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, Jiménez-Hernández N, et al. 2011. Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS One.* 6(3):e17447

Grant MM, Creese AJ, Barr G, Ling MR, Scott AE, et al. 2010. Proteomic Analysis of a Noninvasive Human Model of Acute Inflammation and Its Resolution : The Twenty-one Day Gingivitis Model research articles. *J. Proteome Res.* 4732–44

Griffen AL, Beall CJ, Campbell JH, Firestone ND, Kumar PS, et al. 2012. Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* 6(6):1176–85

Grine FE, Gwinnett AJ, Oaks JH. 1990. Early hominid dental pathology: Interproximal caries in 1.5 million-year-old Paranthropus robustus from Swartkrans. *Arch. Oral Biol.* 35(5):381–86

Groeneveld A. 1985. Longitudinal study of prevalence of enamel lesions in a fluoridated and non-fluoridated area. *Community Dent. Oral Epidemiol.* 13(3):159–63

Gross EL, Beall CJ, Kutsch SR, Firestone ND, Leys EJ, Griffen AL. 2012. Beyond Streptococcus mutans: Dental Caries Onset Linked to Multiple Species by 16S rRNA Community Analysis. *PLoS One.* 7(10):e47722

Gustafsson BE, Quensel CE, Lanke LS, Lundqvist C, Grahnen H, et al. 1954. The Vipeholm dental caries study; the effect of different levels of carbohydrate intake on caries activity in 436 individuals observed for five years. *Acta Odontol. Scand.* 11(3-4):232–64

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward D V, et al. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 494–504

Hajishengallis G, Darveau RP, Curtis MA. 2012. The keystone-pathogen hypothesis. *Nat. Rev. Microbiol.* 10(10):717–25

Hajishengallis G, Lambris JD. 2011. Microbial manipulation of receptor crosstalk in innate immunity. *Nat. Rev. Immunol.* 11(3):187–200

Hajishengallis G, Lambris JD. 2012. Complement and dysbiosis in periodontal disease. *Immunobiology.* 217(11):1111–16

Harding MA, O'Mullane DM. 2013. Water fluoridation and oral health. *Acta Med. Acad.* 42(2):131–39

He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, et al. 2010. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods.* 7(10):807–12

Hilker R, Munder A, Klockgether J, Losada PM, Chouvarine P, et al. 2014. Interclonal gradient of virulence in the Pseudomonas aeruginosa pangenome from disease and environment. *Environ. Microbiol.*

Hillenkamp F, Karas M. 1990. Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. *Methods Enzymol.* 193:280–95

Hillman JD. 2002. Genetically modified Streptococcus mutans for the prevention of dental caries. *Antonie Van Leeuwenhoek.* 82(1-4):361–66

Hillman JD, Mo J, McDonell E, Cvitkovitch D, Hillman CH. 2007. Modification of an effector strain for replacement therapy of dental caries to enable clinical safety trials. *J. Appl. Microbiol.* 102(5):1209–19

HMP. 2012. A framework for human microbiome research. *Nature.* 486(7402):215–21

Ho Sui SJ, Fedynak A, Hsiao WWL, Langille MGI, Brinkman FSL. 2009. The association of virulence factors with genomic islands. *PLoS One.* 4(12):e8094

Hollox EJ, Barber JCK, Brookes AJ, Armour J a L. 2008. Defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1. *Genome Res.* 18(11):1686–97

Holmen L, Thylstrup A, Øgaard B, Kragh F. 1985. A Scanning Electron Microscopic Study of Progressive Stages of Enamel Caries in vivo. *Caries Res.* 19(4):355–67

Holmgren J, Czerkinsky C. 2005. Mucosal immunity and vaccines. *Nat. Med.* 11(4 Suppl):S45–53

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3(12):1365–73

Humphrey LT, De Groote I, Morales J, Barton N, Collcutt S, et al. 2014. Earliest evidence for caries and exploitation of starchy plant foods in Pleistocene hunter-gatherers from Morocco. *Proc. Natl. Acad. Sci. U. S. A.* 111(3):954–59

Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, et al. 2012. Structure, function and diversity of the healthy human microbiome. *Nature.* 486(7402):207–14

Imfeld T. 1996. Dental erosion. Definition, classification and links. *Eur. J. Oral Sci.* 104(2 (Pt 2)):151–55

Imfeld T, Lutz F. 1980. Intraplaque acid formation assessed in vivo in children and young adults. *Pediatr Dent*

Jagtap P, McGowan T, Bandhakavi S, Tu ZJ, Seymour S, et al. 2012. Deep metaproteomic analysis of human salivary supernatant. *Proteomics.* 12(7):992–1001

Jenkinson HF, Lamont RJ. 2005. Oral microbial communities in sickness and in health. *Trends Microbiol.* 13(12):589–95

Jones S, Burt BA, Petersen PE, Lennon MA. 2005. The effective use of fluorides in public health. *Bull. World Health Organ.* 83(9):670–76

Kawar N, Gajendrareddy PK, Hart TC, Nouneh R, Maniar N, Alrayyes S. 2011. Periodontal disease for the primary care physician. *Dis. Mon.* 57(4):174–83

Keijser BJF, Zaura E, Huse SM, van der Vossen JMBM, Schuren FHJ, et al. 2008. Pyrosequencing analysis of the oral microflora of healthy adults. *J. Dent. Res.* 87(11):1016–20

Kembel SW, Wu M, Eisen J a, Green JL. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.* 8(10):e1002743

Kim FM, Hayes C, Williams PL, Whitford GM, Joshipura KJ, et al. 2011. An assessment of bone fluoride and osteosarcoma. *J. Dent. Res.* 90(10):1171–76

Kirkham J, Firth A, Vernals D, Boden N, Robinson C, et al. 2007. Self-assembling Peptide Scaffolds Promote Enamel Remineralization. *J. Dent. Res.* 86(5):426–30

Kleinberg I. 1979. Etiology of dental caries. *J. Can. Dent. Assoc.* 45(12):661–68

Kleinberg I. 1999. A new saliva-based anticaries composition. *Dent. Today.* 18(2):98–103

Kleinberg I. 2002. A mixed-bacteria ecological approach to understanding the role of the oral bacteria in dental caries causation: an alternative to Streptococcus mutans and the specific-plaque hypothesis. *Crit. Rev. Oral Biol. Med.* 13(2):108–25

Klinge RF. 2001. Further Observations on Tertiary Dentin in Human Deciduous Teeth. *Adv. Dent. Res.* 15(1):76–79

Koch R. 1870. Die Ätiologie der Milzbrand-Krankheit , begründet auf die Entwicklungsgeschichte des Bacillus Anthracis . 1). *Cohns Beiträge zur Biol. der Pflanz.* 2(2):5–27

Kolenbrander PE, Andersen RN, Blehert DS, Egland PG, Foster JS, Palmer RJ. 2002. Communication among Oral Bacteria. *Microbiol. Mol. Biol. Rev.* 66(3):486–505

Kolenbrander PE, Andersen RN, Moore L V. 1989. Coaggregation of Fusobacterium nucleatum , Selenomonas flueggei , Selenomonas infelix , Selenomonas noxia , and Selenomonas sputigena with Strains from 11 Genera of Oral Bacteria

Kolenbrander PE, London J. 1993. Adhere Today , Here Tomorrow : Oral Bacterial Adherence. *J. Bacteriol.* 175(11):3247–52

Kolenbrander PE, Palmer RJ, Periasamy S, Jakubovics NS. 2010. Oral multispecies biofilm development and the key role of cell-cell distance. *Nat. Rev. Microbiol.* 8(7):471–80

Kolenbrander PE, Palmer RJ, Rickard AH, Jakubovics NS, Chalmers NI, Diaz PI. 2006. Bacterial interactions and successions during plaque development. *Periodontol. 2000.* 42(5):47–79

Koopman JE, Röling WFM, Buijs MJ, Sissons CH, Ten Cate JM, et al. 2014. Stability and Resilience of Oral Microcosms Toward Acidification and Candida Outgrowth by Arginine Supplementation. *Microb. Ecol.*

Koren O, Spor A, Felin J, Fåk F, Stombaugh J, et al. 2011. Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc. Natl. Acad. Sci. U. S. A.* 108 Suppl :4592–98

Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* 23:110–20

Kozarov E V, Dorn BR, Shelburne CE, Dunn W a, Progulske-Fox A. 2005. Human atherosclerotic plaque contains viable invasive Actinobacillus actinomycetemcomitans and Porphyromonas gingivalis. *Arterioscler. Thromb. Vasc. Biol.* 25(3):e17–18

Kriete A, Eils R. 2013. *Computational Systems Biology: From Molecular Mechanisms to Disease*, Vol. 26. Academic Press

Kroes I, Lepp PW, Relman DA. 1999. Bacterial diversity within the human subgingival crevice. *Proc. Natl. Acad. Sci. U. S. A.* 96(25):14547–52

Küchler EC, Deeley K, Ho B, Linkowski S, Meyer C, et al. 2013. Genetic mapping of high caries experience on human chromosome 13. *BMC Med. Genet.* 14:116

Kumar PS, Leys EJ, Bryk JM, Martinez FJ, Moeschberger ML, Griffen AL. 2006. Changes in periodontal health status are associated with bacterial community shifts as assessed by quantitative 16S cloning and sequencing. *J. Clin. Microbiol.* 44(10):3665–73

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* 12(1):118–23

Lagier J-C, Armougom F, Million M, Hugon P, Pagnier I, et al. 2012. Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin. Microbiol. Infect.* 18(12):1185–93

Lan R, Reeves PR. 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol.* 8(9):396–401

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature.* 409(6822):860–921

Lanfranco LP, Eggers S. 2010. The usefulness of caries frequency, depth, and location in determining cariogenicity and past subsistence: a test on early and later agriculturalists from the Peruvian coast. *Am. J. Phys. Anthropol.* 143(1):75–91

Langille M, Zaneveld J, Caporaso JG, McDonald D, Knights D, et al. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31(9):814–21

Larsen CS, Shavit R, Griffin MC, Anthropolag D. 1984. Dental Caries Evidence for Dietary Change : An Archaeological Context. *Adv. Dent. Anthropol.* 179–202

Lee YK, Mazmanian SK. 2010. Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science.* 330(6012):1768–73

Lehner T, Lamb JR, Welsh KL, Batchelor RJ. 1981. Association between HLA-DR antigens and helper cell activity in the control of dental caries. *Nature.* 292(5825):770–72

Lepp PW, Brinig MM, Ouverney CC, Palm K, Armitage GC, Relman DA. 2004. Methanogenic Archaea and human periodontal disease. *Proc. Natl. Acad. Sci. U. S. A.* 101(16):6176–81

Levy M, Leclerc B-S. 2012. Fluoride in drinking water and osteosarcoma incidence rates in the continental United States among children and adolescents. *Cancer Epidemiol.* 36(2):e83–88

Li J, Helmerhorst EJ, Leone CW, Troxler RF, Yaskell T, et al. 2004. Identification of early microbial colonizers in human dental biofilm. *J. Appl. Microbiol.* 97(6):1311–18

Li Y, Ge Y, Saxena D, Caufield PW. 2007. Genetic profiling of the oral microbiota associated with severe early-childhood caries. *J. Clin. Microbiol.* 45(1):81–87

Liljemark WF, Bloomquist C. 1996. Human oral microbial ecology and dental caries and periodontal diseases. *Crit. Rev. Oral Biol. Med.* 7(2):180–98

Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, et al. 2015. A new antibiotic kills pathogens without detectable resistance. *Nature*

Listgarten MA. 1976. Structure of the microbial flora associated with periodontal health and disease in man. A light and electron microscopic study. *J. Periodontol.* 47(1):1–18

Listgarten MA. 1994. The structure of dental plaque. *Periodontol. 2000.* 5:52–65

Liu Y, Yaling L, Hu T, Tao H, Zhang J, et al. 2006. Characterization of the Actinomyces naeslundii ureolysis and its role in bacterial aciduricity and capacity to modulate pH homeostasis. *Microbiol. Res.* 161(4):304–10

Liu Y-L, Nascimento M, Burne R a. 2012. Progress toward understanding the contribution of alkali generation in dental biofilms to inhibition of dental caries. *Int. J. Oral Sci.* 4(3):135–40

Loesche WJ. 1979. Clinical and microbiological aspects of chemotherapeutic agents used according to the specific plaque hypothesis. *J. Dent. Res.* 58(12):2404–12

Loesche WJ. 1986. Role of Streptococcus mutans in human dental decay. *Microbiol. Rev.* 50(4):353–80

Luan B, Wang D, Zhou R, Harrer S, Peng H, Stolovitzky G. 2012. Dynamics of DNA translocation in a solid-state nanopore immersed in aqueous glycerol. *Nanotechnology.* 23(45):455102

Maehara H, Iwami Y, Mayanagi H, Takahashi N. 2005. Synergistic inhibition by combination of fluoride and xylitol on glycolysis by mutans streptococci and its biochemical mechanism. *Caries Res.* 39(6):521–28

Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, et al. 2007. Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci. U. S. A.* 104(29):11889–94

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 437(7057):376–80

Marsh PD. 1994a. Microbial ecology of dental plaque and its significance in health and disease. *Adv. Dent. Res.* 8(2):263–71

Marsh PD. 1994b. Microbial ecology of dental plaque and its significance in health and disease. *Adv. Dent. Res.* 8(2):263–71

Marsh PD. 2003. Are dental diseases examples of ecological catastrophes? *Microbiology.* 149(2):279–94

Marsh PD. 2006. Dental plaque as a biofilm and a microbial community - implications for health and disease. *BMC Oral Health.* 6 Suppl 1:S14

Marsh PD. 2010. Microbiology of dental plaque biofilms and their role in oral health and caries. *Dent. Clin. North Am.* 54(3):441–54

Marsh PD, Moter A, Devine DA. 2011. Dental plaque biofilms: communities, conflict and control. *Periodontol. 2000.* 55(1):16–35

Marthaler TM. 1967. Epidemiological and clinical dental findings in relation to intake of carbohydrates. *Caries Res.* 1(3):222–38

Marthaler TM. 2013. Salt fluoridation and oral health. *Acta Med. Acad.* 42(2):140–55

Martínez-Murcia AJ, Antón AI, Rodríguez-Valera F. 1999. Patterns of sequence variation in two regions of the 16S rRNA multigene family of Escherichia coli. *Int. J. Syst. Bacteriol.* 49 Pt 2:601–10

McFall-Ngai MJ. 2002. Unseen forces: the influence of bacteria on animal development. *Dev. Biol.* 242(1):1–14

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19(9):1527–41

Meng Y, Zhang H-Q, Pan F, He Z-D, Shao J-L, Ding Y. 2011. Prevalence of dental caries and tooth wear in a Neolithic population (6700-5600 years BP) from northern China. *Arch. Oral Biol.* 56(11):1424–35

Miller WD. 1890. *Microorganism of the Human Mouth. The Local and General Diseases Which Are Caused by Them.* Philadelphia

Mira A. 2008. Horizontal Gene Transfer in Oral Bacteria. In *Molecular Oral Microbiology*, ed. AH Rogers, pp. 65–81. Norfolk, UK: Caister Academic Press

Mira A. 2010. Anticaries compositions and probiotics/prebiotics

Morales SE, Holben WE. 2009. Empirical Testing of 16S rRNA Gene PCR Primer Pairs Reveals Variance in Target Specificity and Efficacy Not Suggested by In Silico Analysis. *Appl. Environ. Microbiol.* 75(9):2677–83

Mosby. 2013. *Mosby's Dental Dictionary.* Elsevier

Moye ZD, Burne RA, Zeng L. 2014. Uptake and Metabolism of N-Acetylglucosamine and Glucosamine by Streptococcus mutans. *Appl. Environ. Microbiol.* 80(16):5053–67

Moynihan P, Petersen PE. 2007. Diet, nutrition and the prevention of dental diseases. *Public Health Nutr.* 7(1a):201–26

Muyzer G. 1999. DGGE/TGGE a method for identifying genes from natural ecosystems. *Curr. Opin. Microbiol.* 2(3):317–22

Nanci A. 2008. *Ten Cate's Oral Histology: Development, Structure, and Function*. Elsevier Health Sciences

National Toxicology Program. 1990. NTP Toxicology and Carcinogenesis Studies of Sodium Fluoride (CAS No. 7681-49-4)in F344/N Rats and B6C3F1 Mice (Drinking Water Studies). *Natl. Toxicol. Program Tech. Rep. Ser.* 393:1–448

Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, et al. 2010. Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. *Appl. Environ. Microbiol.* 76(8):2445–50

Nogueira T, Springer M. 2000. Post-transcriptional control by global regulators of gene expression in bacteria. *Curr. Opin. Microbiol.* 3(2):154–58

Nyvad B, Crielaard W, Mira a, Takahashi N, Beighton D. 2013. Dental caries from a molecular microbiological perspective. *Caries Res.* 47(2):89–102

Ogunseitan OA. 1993. Direct extraction of proteins from environmental samples. *J. Microbiol. Methods.* 17(4):273–81

Ooshima T, Sobue S, Hamada S, Kotani S. 1981. Susceptibility of rats, hamsters, and mice to carious infection by Streptococcus mutans serotype c and d organisms. *J. Dent. Res.* 60(4):855–59

Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, et al. 2009. Direct RNA sequencing. *Nature.* 461(7265):814–18

Pace NR. 1997. A Molecular View of Microbial Diversity and the Biosphere. *Science (80-.).* 276(5313):734–40

Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, et al. 2001. Bacterial diversity in human subgingival plaque. *J. Bacteriol.* 183(12):3770–83

Pasteur L. 1857. Mémoire sur la fermentation appelée lactique (Extrait par l’auteur). *Comptes rendus des seances l’Academie des Sci.* 45:913–16

Patir A, Seymen F, Yildirim M, Deeley K, Cooper ME, et al. 2008. Enamel formation genes are associated with high caries experience in Turkish children. *Caries Res.* 42(5):394–400

Peano C, Pietrelli A, Consolandi C, Rossi E, Petiti L, et al. 2013. An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb. Inform. Exp.* 3(1):1

Peters BM, Jabra-Rizk MA, O’May GA, Costerton JW, Shirtliff ME. 2012. Polymicrobial interactions: impact on pathogenesis and human disease. *Clin. Microbiol. Rev.* 25(1):193–213

Petersen PE. 2003. The World Oral Health Report 2003

Petersen PE, Bourgeois D, Ogawa H, Estupinan-Day S, Ndiaye C. 2005. The global burden of oral diseases and risks to oral health. *Bull. World Health Organ.* 83(9):661–69

Petersen PE, Lennon MA. 2004. Effective use of fluorides for the prevention of dental caries in the 21st century: the WHO approach. *Community Dent. Oral Epidemiol.* 32(5):319–21

Peterson J. 1997. Solving the mystery of the Colorado Brown Stain. *J. Hist. Dent.* 45(2):57–61

Petrof EO, Gloor GB, Vanner SJ, Weese SJ, Carter D, et al. 2013. Stool substitute transplant therapy for the eradication of Clostridium difficile infection: “RePOOPulating” the gut. *Microbiome.* 1(1):3

Pham LC, Hoogenkamp MA, Exterkate RAM, Terefework Z, de Soet JJ, et al. 2011. Effects of Lactobacillus rhamnosus GG on saliva-derived microcosms. *Arch. Oral Biol.* 56(2):136–47

Pine CM, Adair PM, Nicoll AD, Burnside G, Petersen PE, et al. 2004. International comparisons of health inequalities in childhood dental caries. *Community Dent. Health.* 21(1 Suppl):121–30

Polz MF, Cavanaugh CM. 1998. Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* 64(10):3724–30

Pradet-Balade B, Boulmé F, Beug H, Müllner EW, Garcia-Sanz JA. 2001. Translation control: bridging the gap between genomics and proteomics? *Trends Biochem. Sci.* 26(4):225–29

Qin J, Cui Y, Zhao X, Rohde H, Liang T, et al. 2011. Identification of the Shiga toxin-producing Escherichia coli O104:H4 strain responsible for a food poisoning outbreak in Germany by PCR. *J. Clin. Microbiol.* 49(9):3439–40

Qin J, Li Y, Cai Z, Li S, Zhu J, et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature.* 490(7418):55–60

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 464(7285):59–65

Randal Bollinger R, Everett M Lou, Palestrant D, Love SD, Lin SS, Parker W. 2003. Human secretory immunoglobulin A may contribute to biofilm formation in the gut. *Immunology.* 109(4):580–87

Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, et al. 2015. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *Elife.* 4:1–25

Ravel J, Brotman RM, Gajer P, Ma B, Nandy M, et al. 2013. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. *Microbiome.* 1(1):29

Razumov A. 1932. The direct method of calculation of bacteria in water: comparison with the Koch method. *Mikrobiologija.* 1:131–46

Richards MP. 2002. A brief review of the archaeological evidence for Palaeolithic and Neolithic subsistence. *Eur. J. Clin. Nutr.* 56(12):16 p following 1262

Richter AE, Arruda AO, Peters MC, Sohn W. 2011. Incidence of caries lesions among patients treated with comprehensive orthodontics. *Am. J. Orthod. Dentofacial Orthop.* 139(5):657–64

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, et al. 2009. Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* 7(11):828–36

Rodríguez-Valera F. 2004. Environmental genomics, the big picture?

Rogers AH, Hoeven JSVANDER, Mikxe FHM. 1979. Effect of Bacteriocin Production by Streptococcus mutans on the Plaque of Gnotobiotic Rats. . 23(3):571–76

Roldán S, Herrera D, Sanz M. 2003. Biofilms and the tongue: therapeutical approaches for the control of halitosis. *Clin. Oral Investig.* 7(4):189–97

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* 242(1):84–89

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, et al. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66(6):2541–47

Rosier BT, Jager M De, Zaura E, Krom BP. 2014. Historical and contemporary hypotheses on the development of oral diseases : are we there yet ? *Front. Cell. Infect. Microbiol.* 4(92):1–11

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 475(7356):348–52

Rugg-Gunn AJ, Edgar WM, Geddes DA, Jenkins GN. 1975. The effect of different meal patterns upon plaque pH in human subjects. *Br. Dent. J.* 139(9):351–56

Rugg-Gunn AJ, Hackett AF, Appleton DR, Jenkins GN, Eastoe JE. 1984. Relationship between dietary habits and caries increment assessed over two years in 405 English adolescent school children. *Arch. Oral Biol.* 29(12):983–92

Russell MW, Childers NK, Michalek SM, Smith DJ, Taubman MA. 2004. A Caries Vaccine? The state of the science of immunization against dental caries. *Caries Res.* 38(3):230–35

Russell MW, Wu HY. 1991. Distribution, persistence, and recall of serum and salivary antibody responses to peroral immunization with protein antigen I/II of Streptococcus mutans coupled to the cholera toxin B subunit. *Infect. Immun.* 59(11):4061–70

Russell RR. 2009. Changing concepts in caries microbiology. *Am. J. Dent.* 22(5):304–10

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74(12):5463–67

Santarpia RP, Lavender S, Gittins E, Vandeven M, Cummins D, Sullivan R. 2014. A 12-week clinical study assessing the clinical effects on plaque metabolism of a dentifrice containing 1.5% arginine, an insoluble calcium compound and 1,450 ppm fluoride. *Am. J. Dent.* 27(2):100–105

Savage DC. 1977. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.* 31(70):107–33

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Hum. Mol. Genet.* 19(2):227–40

Schwientek P, Szczepanowski R, Rückert C, Stoye J, Pühler A. 2011. Sequencing of high G+C microbial genomes using the ultrafast pyrosequencing technology. *J. Biotechnol.* 155(1):68–77

Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, et al. 2012. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* 13(6):R42

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, et al. 2011. Metagenomic biomarker discovery and explanation. *Genome Biol.* 12(6):R60

Seksik P. 2010. [Gut microbiota and IBD]. *Gastroentérologie Clin. Biol.* 34 Suppl 1:S44–51

Simón-Soro A, Guillen-Navarro M, Mira A. 2014. Metatranscriptomics reveals overall active bacterial composition in caries lesions. *J. Oral Microbiol.* 6:25443

Simón-Soro A, Mira A. 2014. Solving the etiology of dental caries. *Trends Microbiol.*

Simón-Soro A, Tomás I, Cabrera-Rubio R, Catalan MD, Nyvad B, Mira A. 2013a. Microbial geography of the oral cavity. *J. Dent. Res.* 92(7):616–21

Simón-Soro Á, Belda-Ferre P, Cabrera-Rubio R, Alcaraz LD, Mira a. 2013b. A Tissue-Dependent Hypothesis of Dental Caries. *Caries Res.* 47(6):591–600

Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M. 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* 60(2):341–50

Smith DJ. 2002. DENTAL CARIES VACCINES: PROSPECTS AND CONCERNS. *Crit. Rev. Oral Biol. Med.* 13(4):335–49

Sneath P, Mair N, Sharpe M, Holt J, eds. 1986. Oral streptococci. In *Bergey's Manual of Systematic Bacteriology. Baltimore*, pp. 1054–63. Baltimore: Williams & Wilkins. Vol 2 ed.

Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL. 1998. Microbial complexes in subgingival plaque. *J. Clin. Periodontol.* 25(2):134–44

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. 2006. Microbial diversity in the deep sea and the underexplored "' rare biosphere .'" . (30):

Sognnaes RF. 1948. Analysis of wartime reduction of dental caries in European children; with special regard to observations in Norway. *Am. J. Dis. Child.* 75(6):792–821

Stanley HR, Pereira JC, Spiegel E, Broom C, Schultz M. 1983. The detection and prevalence of reactive and physiologic sclerotic dentin, reparative dentin and dead tracts beneath various types of dental lesions according to tooth surface and age. *J. Oral Pathol.* 12(4):257–89

Steinberg S. 2009. Adding caries diagnosis to caries risk assessment: the next step in caries management by risk assessment (CAMBRA). *Compend. Contin. Educ. Dent.* 30(8):522, 524–26, 528 passim

Stewart FJ, Ottesen E a, DeLong EF. 2010. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.* 4(7):896–907

Streckfus CF, Bigler LR. 2002. Saliva as a diagnostic fluid. *Oral Dis.* 8(2):69–76

Suzuki MT, Giovannoni SJ. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* 62(2):625–30

Szpunar SM, Eklund SA, Burt BA. 1995. Sugar consumption and caries risk in schoolchildren with low caries experience. *Community Dent. Oral Epidemiol.* 23(3):142–46

Takahashi N, Nyvad B. 2008. Caries ecology revisited: microbial dynamics and the caries process. *Caries Res.* 42(6):409–18

Tanaka T, Kawasaki K, Daimon S, Kitagawa W, Yamamoto K, et al. 2014. A hidden pitfall in agar media preparation undermines cultivability of microorganisms. *Appl. Environ. Microbiol.* 80(24):7659–66

Tanner a. CR, Milgrom PM, Kent R, Mokeem S a., Page RC, et al. 2002. The Microbiota of Young

Children from Tooth and Tongue Samples. *J. Dent. Res.* 81(1):53–57

Taubman MA, Smith DJ. 1974. Effects of local immunization with Streptococcus mutans on induction of salivary immunoglobulin A antibody and experimental dental caries in rats.*Infect. Immun.* 9(6):1079–91

Teles FR, Teles RP, Uzel NG, Song XQ, Torresyap G, et al. 2012. Early microbial succession in redeveloping dental biofilms in periodontal health and disease.*J. Periodontal Res.* 47(1):95–104

Temperton B, Field D, Oliver A, Tiwari B, Mühling M, et al. 2009. Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing.*ISME J.* 3(7):792–96

Ten Cate JMB. 2009. The need for antibacterial approaches to improve caries control.*Adv. Dent. Res.* 21(1):8–12

Theilade E. 1986. The non-specific theory in microbial etiology of inflammatory periodontal diseases. *J. Clin. Periodontol.* 13(10):905–11

Thenisch NL, Bachmann LM, Imfeld T, Leisebach Minder T, Steurer J. 2006. Are mutans streptococci detected in preschool children a reliable predictive factor for dental caries risk? A systematic review. *Caries Res.* 40(5):366–74

Tian Y, He X, Torralba M, Yooseph S, Nelson KE, et al. 2010. Using DGGE profiling to develop a novel culture medium suitable for oral microbial communities.*Mol. Oral Microbiol.* 25(5):357–67

Tillier A, Arensburg B, Rak Y, Vandermeersch B. 1995. Middle Palaeolithic dental caries: new evidence from Kebara (Mount Carmel, Israel). *J. Hum. Evol.* 29(2):189–92

Tinanoff N, Gross A. 1976. Epithelial cells associated with the development of dental plaque.*J. Dent. Res.* 55(4):580–83

Tinanoff N, Gross A, Brady JM. 1976. Development of plaque on enamel. Parallel investigations.*J. Periodontal Res.* 11(4):197–209

Tjäderhane L, Larjava H, Sorsa T, Uitto VJ, Larmas M, Salo T. 1998. The activation and function of host matrix metalloproteinases in dentin matrix breakdown in caries lesions.*J. Dent. Res.* 77(8):1622–29

Toi CS, Mogodiri R. 2000. Mutans streptococci and lactobacilli on healthy and carious teeth in the same mouth of children with and without dental caries

Touger-Decker R, van Loveren C. 2003. Sugars and dental caries.*Am. J. Clin. Nutr.* 78(4):881S – 892S

Townsend GC, Richards L, Hughes T, Pinkerton S, Schwerdt W. 2003. The value of twins in dental research. . (2):82–88

Trevino V, Falciani F. 2006. GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics.* 22(9):1154–56

Turnbaugh PJ, Gordon JI. 2009. The core gut microbiome, energy balance and obesity.*J. Physiol.* 587(Pt 17):4153–58

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. 2009. A core gut microbiome in obese and lean twins. *Nature.* 457(7228):480–84

V. Wintzingerode F, Göbel UB, Stackebrandt E. 2006. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis.*FEMS Microbiol. Rev.* 21(3):213–29

Van Houte J. 1994. Role of micro-organisms in caries etiology.*J. Dent. Res.* 73(3):672–81

Van Wuyckhuyse BC, Perinpanayagam HE, Bevacqua D, Raubertas RF, Billings RJ, et al. 1995. Association of free arginine and lysine concentrations in human parotid saliva with caries experience. *J. Dent. Res.* 74(2):686–90

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science*. 291(5507):1304–51

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 304(5667):66–74

Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, et al. 2009. Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* 3(2):179–89

Vidal C, Tjäderhane L, Scaffa P, Tersariol I, Pashley D, et al. 2014. *Abundance of MMPs and cysteine cathepsins in caries-affected dentin*. Journal of dental research. http://www.ncbi.nlm.nih.gov/pubmed/24356440

Wade AD, Hurnanen J, Lawson B, Tampieri D, Nelson AJ. 2012. Early dental intervention in the Redpath Ptolemaic Theban Male. *Int. J. Paleopathol.*

Wade W. 2002. Unculturable bacteria--the uncharacterized organisms that cause oral infections. *J. R. Soc. Med.* 95(2):81–83

Wagner M, Smidt H, Loy A, Zhou J. 2007. Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microb. Ecol.* 53(3):498–506

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73(16):5261–67

Wang R., Weiner S. 1997. Strain–structure relations in human teeth using Moiré fringes. *J. Biomech.* 31(2):135–41

Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, et al. 2014. Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* (February):

Wei G-X, Campagna AN, Bobek LA. 2007. Factors affecting antimicrobial activity of MUC7 12-mer, a human salivary mucin-derived peptide. *Ann. Clin. Microbiol. Antimicrob.* 6:14

Weisburg WG, Barns SM, Pelletier DA, Lane DJ. 1991. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* 173(2):697–703

Wendell S, Wang X, Brown M, Cooper ME, DeSensi RS, et al. 2010. Taste genes associated with dental caries. *J. Dent. Res.* 89(11):1198–1202

White DJ. 1997. Dental calculus: recent insights into occurrence, formation, prevention, removal and oral health effects of supragingival and subgingival deposits. *Eur. J. Oral Sci.* 105(5):508–22

Whitman WB, Coleman DC, Wiebe WJ. 1998. Perspective Prokaryotes : The unseen majority. *Proc. Natl. Acad. Sci.* 95(June):6578–83

Wijeyeweera RL, Kleinberg I. 1989. Arginolytic and ureolytic activities of pure cultures of human oral bacteria and their effects on the pH response of salivary sediment and dental plaque in vitro. *Arch. Oral Biol.* 34(1):43–53

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, et al. 2008. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One*. 3(1):e1456

Wilmes P, Bond PL. 2004. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* 6(9):911–20

Woese CR, Gogarten PJ. 1999. When did eukaryotic cells (cells with nuclei and other internal organelles) first evolve? What do we know about how they evolved from earlier life-forms? *Sci. Am.*, . http://www.scientificamerican.com/article/when-did-eukaryotic-cells/?print=true

Wooley JC, Godzik A, Friedberg I. 2010. A primer on metagenomics. *PLoS Comput. Biol.* 6(2):e1000667

Wostmann BS. 1981. The germfree animal in nutritional studies. *Annu. Rev. Nutr.* 1:257–79

Wu J-Y, Jiang X-T, Jiang Y-X, Lu S-Y, Zou F, Zhou H-W. 2010. Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiol.* 10:255

Yamashita M, Fenn JB. 1984. Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.* 88(20):4451–59

Yang F, Zeng X, Ning K, Liu K-L, Lo C-C, et al. 2011. Saliva microbiomes distinguish caries-active from healthy human populations. *ISME J.*

Yi H, Cho Y-J, Won S, Lee J-E, Jin Yu H, et al. 2011. Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res.* 39(20):e140

Zaura E, Keijser BJF, Huse SM, Crielaard W. 2009. Defining the healthy “core microbiome” of oral microbial communities. *BMC Microbiol.* 9:259

Zijnge V, Van Leeuwen MBM, Degener JE, Abbas F, Thurnheer T, et al. 2010. Oral biofilm architecture on natural teeth. *PLoS One.* 5(2):e9321

Zoetendal EG, Rajilic-Stojanovic M, de Vos WM. 2008. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut.* 57(11):1605–15