METHODOLOGICAL PREPARATION AND CHARACTERIZATION OF **THE MICROBIAL ECOLOGY OF THE SKIN**

Marc Garcia i Garcerà

TESI DOCTORAL UPF / 2013

DIRECTOR DE LA TESI:
Dr. Francesc Calafell i Majó

Departament de Ciències Experimentals i de la Salut

UNIVERSITAT POMPEU FABRA

Als meus avis: Milagro i Ramón, Salvador i Rosa,

Als meus pares, Miracle i Salvador,

Al Jordi i Alba,

Als meus tios i ties, que en sou molts

A Enric,

I, en especial, a Francesc i Koldo,

Perquè malgrat tot l'esforç, gràcies a tots vosaltres, he arribat.

"Un sociólogo norteamericano dijo hace más de treinta años
que la propaganda era una formidable vendedora de sueños,
pero resulta que yo no quiero que me vendan sueños ajenos,
sino sencillamente que se cumplan los míos"

Mario Benedetti

"...But here's what gives me a hard-on:
I am a tiny, insignificant, ignorant lump of carbon.
I have one life, and it is short
And unimportant
But thanks to recent scientific advances
I get to live twice as long
As my great great great great uncleses and auntses.
Twice as long to live this life of mine
Twice as long to love this wife of mine
Twice as many years of friends and wine
Of sharing curries and getting shitty
With good-looking hippies
With fairies on their spines
And butterflies on their titties"

Tim Minchin - Storm

TABLE OF CONTENTS

AGRAÏMENTS

Abans que ningú, vull agrair-te a tu, lector desconegut, que t'hages atrevit a agafar aquesta tesi. Per que entengues tot allò que vull escriure i comprengues tot el camí que hem recorregut ella i jo per poder dir que és una realitat. Perquè no siga excessiva, sino accessible i entretinguda, i que sigues capaç d'arribar al final i desxifrar allò que no s'hi diu però que hi és entre línies.

Al Francesc, per tots els esforços posats en aquest treball. Perquè sense el seu suport incondicional, la seua empenta i la seua seguretat creient en el projecte i en mi, aquesta tesi ni aquest doctorat serien una realitat. Moltíssimes gràcies pel recolzament, la llibertat, i l'empenta que m'has oferit al llarg de tots aquests anys.

A l'Amparo, per oferir-me un lloc on desenvolupar la meua feina, per escoltar les meues hipòtesis i recolzar-me a tirar endavant, per tractar-me com un més del grup de genòmica del CSISP, sense cap tipus de condició, i tractar-me com un doctorand més, amb les reunions, discussions i anàlisi de resultats.

A Ana (D), perquè si algú m'ha servit de conciència científica, eixa eres tu. Perquè, quan la meua imaginació desbordava, tu has sabut plantar els meus peus a terra. Per les teues idees, discussions, cafès, desdejunis i passejos. Per les bones i les males històries, perquè al cap i a la fi, les amistats es forgen d'això i més.

A Koldo, perquè sense ell, el grup de metagenòmica de bioevo estaria més que coix. Perquè ha cregut en el projecte inclús quan la resta començavem a dubtar. Per dubtar de la convenció establerta i demostrar-me que de vegades hi ha que aprofundir més per trobar millors resultats. Per tots els moments que hem compartit i tot el recolzament en els pitjors moments. Eskerrik asko, Koldo.

A Leo, per fer-me estimar la filogenòmica com només ella pot estimar-la. Per fer-me voler ser millor científic. Per totes les hores davant arbres i aliniaments, escollint el millor model per explicar una història evolutiva que, a priori, no té cap sentit. Per la seua perseverància.

A Fernando, per oferir-me recolzament i projectes en els moments que més ho he necessitat. Per incloure'm al seu grup i fer-me sentir-ne part. Per recordar-me perquè faig ciència.

À Sandrine pour l'escalade, pour ton control parental et, bien, tu sais parfaitement pourquoi. Merci.

A Pedro, Raúl, Ana Elena, Ana (Dj), Peris, per les hores compartint despatx, riure, ciència i cafès.

Als membres del grup de genòmica del CSISP, per acollir-me i fer-me sentir com un més.

A Mireia, per tota l'ajuda, les hores al despatx, per allotjar-me a casa teua inclús sense estar-hi. Gràcies per tot, Mireia.

A l'Arcadi i el Tomàs, per haver cregut en mi fins i tot quan jo no ho tenia massa clar. Per haver-me oferit la vostra ajuda quan encara no era part de bioevo, per posar-me en contacte amb Francesc i oferir-me la possibilitat de ser i voler ser científic.

Al Carles Lalueza i el Sergi Civit, per tota la feina junts.

A l'Angel i el Txema, per la vostra paciència ensenyant-me informàtica

A la Núria Bonet i la Neus Solé, pel temps al davant la "poyata" amb les pells i els esputs i moltíssimes coses desagradables. Per fer-me passar el fàstic amb alegria.

A tot Bioevo. Als que ja no hi sou (Anna Ramírez, Valeria, Laura, Urko, Ixa,...) als que no he tingut el plaer de conèixer però formeu ara Bioevo, i als que hi formeu part encara (Javi, Marta, Belén, Elena, Gabriel,...). En sou molts, però a tots us dic: Moltíssimes gràcies.

To Cedric, for all the help with the multiple alignments, showing me why any evolutionary history can't be understood without a good alignment.

To Jeroen, for opening his doors at the VIB in brussels, for teaching me through long meetings that study the microbiome is not just 16S.

To the good friends I made in Brussels: Youssef, Sam, Ye, Falk, for all the hours in front of the computer and in front of the beers and chocolates.

A Marta, pels Brunchs, el pis, la companyia, i tot el suport durant la meua estada a Brussel·les. On està el meu Brunch, Marta?

A Marisa, per tot el que ens uneix, ens ha unit, i (espere) ens unirà. Perquè malgrat tot, has viscut amb mi la tesi que ara pareix que és una realitat. Gràcies.

A Paco, per creure que totes les malalties podien estar provocades per microbis.

A Enric, perquè no m'imagine una vida sense el meu amic. Perquè he passat amb ell més coses que amb ningú i m'ha vist en els meus millors i pitjors moments. I malgrat tota la distància, els problemes i els anys, ahí seguim, dia rere dia, any rere any. Amics com tu, Andrew, no n'hi ha.

A les meues amigues: Anaïs, Violeta, Laura, Lissi, Margarita, Lucía, perquè sabeu que sóc un desastre, i puc desapareixer molt de temps, però malgrat tot, ahí seguiu. Gràcies.

A Arantxa, Anna, Belén, Núria i Mònica. Perquè sense vosaltres la carrera haguera sigut un patiment, perquè tot i canviar d'universitat continueu sent les meues "amigues de la facultat". En especial a Arantxa, gràcies per assignatures i apunts *compartits.*

A Ana Ibars i Mercè Pamblanco, per ensenyar-me el que és la ciència i voler formar-ne part. A Rosa Blasco per inculcar-me l'interés per la biologia i fer-me decidir per ella.

A la meua cunyada, Alba, per tots els bons moments, per interessar-te per mi i voler ajudar-me fins i tot quan no podies.

Al meu germà, perquè sense tu aquesta vida seria molt més avorrida.

Als meus tios i ties, cosins i cosines, per preocupar-vos per mi, i facilitar-me les coses sempre que us ha sigut possible. Pels dinars, sopars, demostrant-me la importància de cuidar la familia.

Als meus "uelos": Perquè sempre vos heu preocupat per mi, i m'heu ajudat sempre que heu pogut. Pels macarrons, la coca de poma, les creïlles fregides i la paella, característics de cada casa que m'han fet creixer sa i feliç.

I per suposat, als meus pares. Perquè sense vosaltres ni aquesta ni ninguna de les meues experiències seria una realitat. Gràcies per fer-me creixer amb criteri, convicció i salut. Moltíssimes gràcies per estar ahí, quan cal i quan no.

ABSTRACT

The study of skin microbiota has always been focused from a clinical point of view. However, an ecological approach to the skin microbiota is impeded by different methodological limitations, including the high host/microbial DNA ratio or the low microbial content. In contrast with the burgeoning field of gut metagenomics, skin metagenomics has been hindered by the absence of an efficient methodology to work with skin microbial DNA. This thesis aims to settle the basis for further human skin microbiome studies from a systems approach. I have set four different important approaches for a ecological and systematic view of skin: 1) I have tested and compared a method to construct NGS libraries from trace amounts of DNA, allowing to work with very rare samples, and performing multiple functional experiments on the same sample; 2) I have defined a method to isolate the microbial DNA from a skin sample, to perform actual metagenomic studies. We have proved the utility of the method by constructing 2 metagenomic libraries from mouse skin biopsies; 3) I have tested the relationship between microbial diversity and unrelated phenotypes in mouse skin samples; 4) I have analyzed the host phenotypical spectra, characterized what it is metabolic health, and assessed the relationship between health and microbiota.

RESUM

L'estudi de la microbiota de la pell ha estat sempre enfocat cap a un punt de vista clínic. No obstant, una aproximació ecològica a la microbiota cutània és impossibilitada per multiples limitacions metodològiques, que inclouen la baixa ràtio de DNA hospedador/DNA comensal o la baixa quantitat absoluta de DNA microbià. En contrast al pròsper camp de la metagenòmica intestinal, la metagenòmica cutània ha estat obstaculitzada metodològicament, i per això, aquesta tesi intenta sentar les bases per al futur de l'estudi del microbioma cutàni i la aplicació d'una aproximació de sistemes. He aplicat quatre aproximacions importants per a la observació sistemàtica i ecològica de la pell: 1) He testat i comparat un mètode per construir llibreries de NGS a partir de traces de DNA, permetent treballar amb mostres de difícil obtenció i aplicar-hi multiples experiments funcionals a la mateixa mostra, sense haver d'usar tot el material en la seqüenciació; 2) He definit un mètode per a aïllar el DNA microbià d'una mostra de pell completa per tal de dur a terme una anàlisi metagenòmica de la mostra. Hem demostrat la seva utilitat construint dues llibreries independents a partir de pell de ratolí; 3) He analitzat la relació entre la diversitat microbiana i un fenotip aïllat de l'hospedador; 4) He analitzat l'espectre fenotípic de l'hoste des d'un punt de vista sistèmic, he caracteritzat l'estat de salut metabòlica i la seua relació amb la microbiota.

PROLOGUE

The skin is the human body's largest organ, colonized by a diverse milieu of microorganisms, which live in and cooperate with the host. Although most of the host-microbiome studies have been focused on the gastrointestinal tract, the skin ecosystem has been considered to be as variable as the latter, and to carry a huge intra-individual variability depending on topographical location.

Skin microbiota has an important role in the development of the innate and adaptive responses at systemic level. It is known to educate host immune system to discriminate between host, commensal and pathogen patterns and act in consequence. Thus, Skin microbiota may have an important role in chronic inflammatory and autoimmune diseases, and then it deserves to be studied carefully. However, to date the methodology to analyze the skin microbiota has limited their study.

This thesis sets the preliminary methodology to perform a systems approach to the skin microbial community from a metagenomic point of view, solving the primary limitations to study the different ecological niches of skin and their possible relationship with host health. This thesis focuses mostly on the methodological aspects of the skin metagenomics.

ABBREVIATIONS

Of the statistical methods used

CCA	Canonical Correspondence Analysis
FDR	False Discovery Rate
KL	Kullback-Leibler test
KW	Kruskal-Wallis
NMDS	non-metric multidimensional scaling
PCA	Principal Component Analysis
PCoA	Principal Coordinates Analysis

Of the bioinformatical algorithms, databases and methods

COG	Cluster of Orthologous Genes
HMM	Hidden Markov Model
KO	Kegg Orthologous Genes
LCA	Lowest Common Ancestor Algorithm

Of Immune-related molecules and cells

AMP	Anti-Microbial Peptide
APC	Antigen-Presenting Cell
CCL	CxC Chemokine Ligand
HLA	Human Leukocyte Antigen
ICAM-1	Intercellular Adhesion Molecule 1
IL	Interleukin
MHC	Major Histocompatibility Complex
MyD88	Myeloid differentiation primary response gene (88)
NF-kB	nuclear factor kappa-light-chain-enhancer of activated B cells
NOD	Nucleic-binding Oligomerization Domain
PAmP	Pathogen-Associated molecular pattern
TLRX	Toll-like Receptor X
TNFa	Tumor Necrosis Factor alpha

Of body parts, organs, and related traits

BMI	Body Mass Index
FAC	Fatty-Acid metabolism cluster
GI or GIT	Gastrointestinal Tract
GMC	Glucose Metabolism cluster
IC	Inflammatory-related Cluster

Of skin or related to

CE	Cornified Cell Envelope
EPU	Epidermal Proliferative Unit
FFA	Free Fatty-Acids
KX	Keratin type X

NHE1	Non-energy dependent sodium-proton exchange mechanism 1

Of microbial diversity and phylogeny

E. coli	*E. Escherichia coli*
H‘	Shannon Diversity Index - Relative
P. acnes	*Propionibacterium acnes*
S. aureus	*Staphylococcus aureus*
S. epidermidis	*Staphylococcus epidermidis*

Of other molecules and characteristics

16S rRNA	component of the 30S small subunit ribosomal RNA
CUB	Codon Usage Bias
DNA	deoxyribonucleic Acid
LPS	lipopolysaccharide
mRNA	messenger Ribonucleic Acid
RNA	Ribonucleic Acid
rRNA	ribosomal Ribonucleic Acid

Of Diseases or Disease-related

AD	Atopic Dermatitis
CD	Crohn’s Disease
IBD	Inflammatory Bowel’s Disease
ID	immunodepressed
IMID	Immune-Mediated Inflammatory Disease
SCID	Severe Combined Immunodeficiency
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes

Of Molecular methods

AF	Whole Genome Amplification-free method
Cq	quantification cut-off
emPCR	emulsion PCR
MDA	Multidisplacement Amplification
NG or NGS	Next Generation Sequencing
PCR	Polymerase-Chain Reaction
qPCR	quantitative PCR

Of projects

HMP	NIH Human Microbiome Project
MetaHIT	Metagenomics of the Human Intestinal Tract project

Introduction

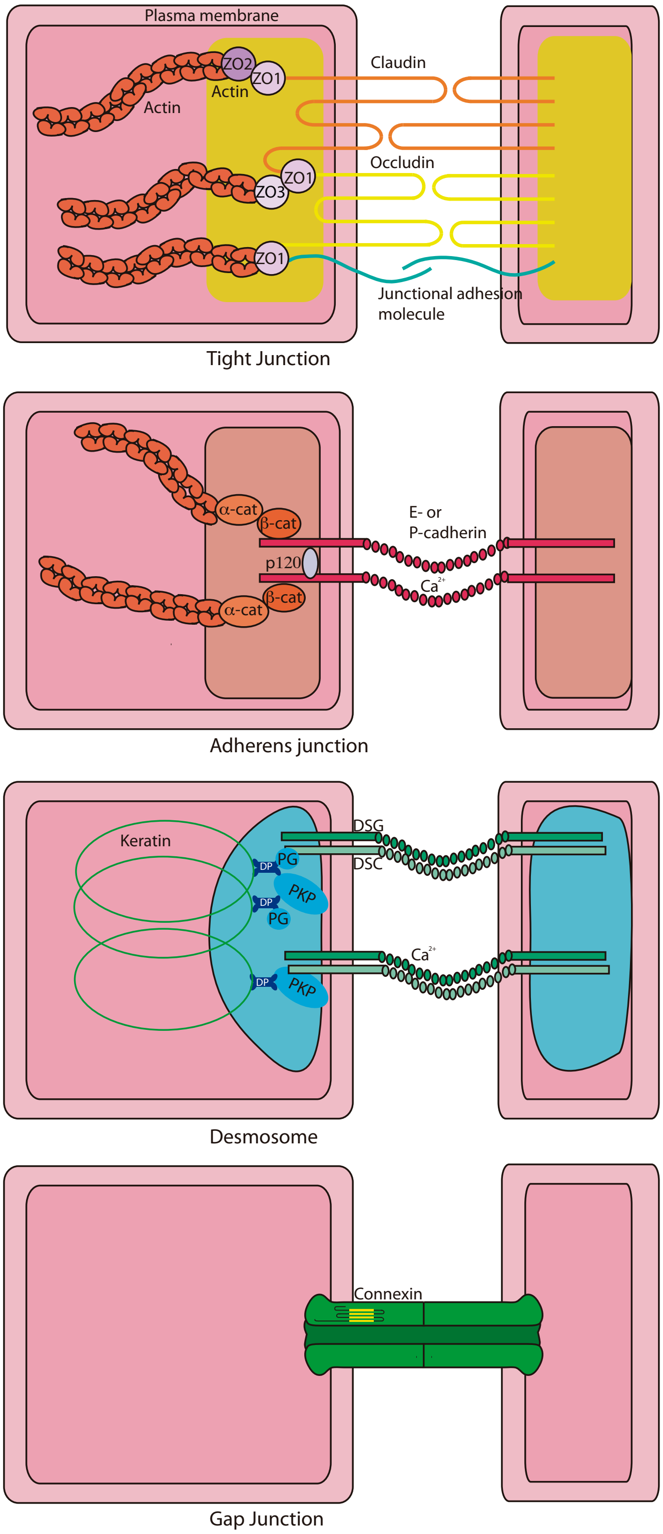
In this thesis, we will focus on the skin as an ecosystem, on their inhabitants, bacteria and other microorganisms, and on how to study them. We will assess the diversity of its residents and the relationships between them and their host. We will also analyze the limitations in their study and how to solve them.

Skin is the body's largest organ in most chordates, covering all the body. And being in contact with the external environment, it is colonized by a diverse pool of microorganisms, most of which are harmless or even beneficial to their host. Colonization is driven by the ecology of the skin surface, which is highly variable depending on topographical location, endogenous host factors and exogenous environmental factors. The cutaneous innate and adaptive immune responses can modulate the skin microbiota, but the microbiota also functions in educating the immune system. However, to completely understand the ecology of skin, considered all together as a complete ecosystem, we need to understand how the skin works as a tissue and how the different elements of the ecosystem work together to maintain homoeostasis.

The (not so) inorganic part. The biotope.

The skin as a external barrier.

Multicellular organisms, from Porifera to Mammalia, require tissue compartmentalization to allow internal cells to specialize and take control of some specific functions, maintaining the homoeostasis of the whole system. Epithelial cells are in contact with the outer space, and must cover the internal subsystems, preventing them to be physically or biologically injured, disrupting the system homoeostasis, provoking a loss of efficiency, even the total malfunction of the system, which would mean death(Marchiando et al. 2010). To establish defined boundaries, epithelial cells must cover the external surface and line internal compartments, and must also form barriers to prevent unrestricted exchange of materials. The critical nature of this barrier, and the consequences of its loss are demonstrated in burn wounds and wound infections, where further functionality and prognosis is directly related to the depth and area damaged(Peck 2011). The functions and the



Tight Junction

Adherens junction

Desmosome

Gap Junction

Figure 1. Intercellular junctions of the epidermis

Tight junctions form a belt at the apical side of keratinocytes, providing an additional barrier beneath the stratum corneum that controls fluid loss and protects against pathogens.

Adherens junctions coordinate the assembly and organization of the cortical actin cytoskeleton throughout the epidermis.

Desmosomes are a third class of intercellular adhesions; they link to the intracellular network of keratin intermediate filaments.

Gap junctions are unique in their ability to provide a direct connection between neighboring cells.

Adapted from Simpson et al. 2011

permeability of those barriers depend strictly on the level of interaction with the environment. While skin maintains an almost total isolation of the organism from the environment, preventing not only pathogen attacks but also physical and chemical assaults, other epithelia such as the gut, allow nutrients and other indispensable solutes to go through the barrier, keeping the income-outcome ratio in equilibrium(Stevens and Hume 1998). As the main role of epithelial cells is protection, their plasma membranes prevent most hydrophilic solutes from crossing the boundaries. But, more importantly, the paracellular pathway, where most pathogens may intrude, must also be sealed. This function is the responsibility of the apical junctional complexes(Stevenson et al. 1986; Anderson et al. 1988; Masahiko Itoh 1997). While tight (apical) junctions seal the paracellular pathway, the adherent junctions and desmosomes provide the strong bonds necessary to maintain cellular proximity (Figure 1) and allow apical junction complex to assemble(Farquhar and Palade 1963). External epithelia, such gut or skin, are considered absolute barriers, which prevent any communication from the exterior space, at least via the paracellular pathway. The apical plasma membrane will alter its permeability to allow the interchange of solutes, according to its alternative functions. Examples of those differences in permeability may be the mammalian small intestinal epithelia, which has low-resistance tight junctions, which allow short molecules to go through, increasing the nutrient intake, or the gallbladder epithelium, which must prevent concentrated bile acids from entering the circulation and forms tight junctions with high resistance.

Skin, in contrast, has different methods to maintain the internal tissues isolated. The epidermis is a thin layer of stratified squamous epithelium, located over a protein matrix called the basal lamina, which separates it from the underlying mesenchymally derived dermis. The epidermis is tough and resilient, and is able to withstand the physical and chemical traumas of each passing day(Fuchs and Raghavan 2002). As any other epithelia, it must keep harmful microorganisms out in order to guard against infection, but it must also retain body fluids to prevent dehydration.

Skin proliferation and differentiation

The epidermis is a self-renewing tissue: a single adult skin stem cell has sufficient proliferative capacity to produce enough new epidermis to cover the whole body(Rochat et al. 1994). As a self-renewing tissue, skin cells differentiate from a multipotent state layer(Watt and Hogan 2000), where cells tightly adhere to each other but allow themselves to separate during cell division and differentiation, to a totally irreversible, structurally tight and resilient state where cells are transformed into keratin sacks connected to each other by junctions with strong physical and chemical resistance. This final definitive state, called stratum corneum, forms a continuous sheet of protein-enriched cells (corneocytes) connected by corneodesmosomes and embedded in an intercellular matrix enriched in non-polar lipids and organized as lamellar lipid layers, which will serve as lipophilic barrier against dehydration(Proksch et al. 2008). The balance between growth and differentiation is crucial for the proper functionality of skin, and depends on a large number of variables, both intrinsic and environmental.

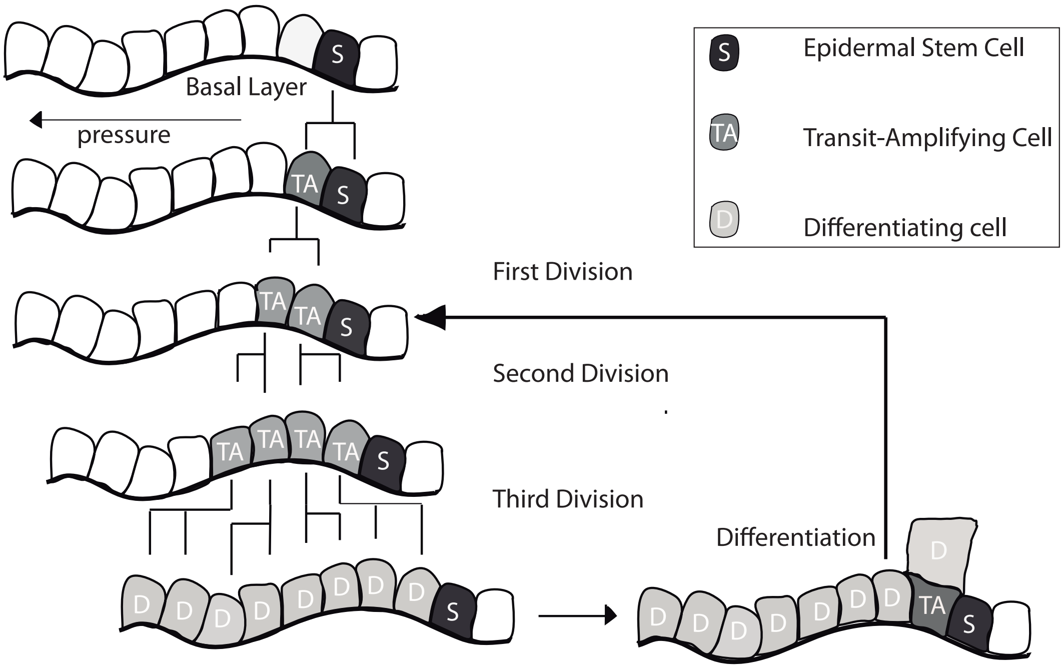


Figure 2. Keratinocyte differentiation.
According to Elias, a single stem-cell would construct a group of 10 tightly packed basal cells, that would start the differentiation and the travel through the different layers of the epidermis. Diagram according to Houben et al. 2007

However, to achieve the corneocyte state, skin stem cells must proliferate in a controlled way to avoid hyper proliferative disorders of the skin, such as squamous-cell carcinomas or psoriasis. The epidermis maintains a single inner (basal) layer of proliferative cells that adhere to an underlying basement membrane rich in extracellular matrix and growth factors. Basal cells express several characteristic markers, including keratins and transcription factors. Periodically, these cells withdraw from the cell cycle, commit to terminally differentiate, move outward, converted into definitive enucleated corneocytes, and are eventually shed from the skin surface(Fuchs 2008). Upon commitment to terminally differentiate, they go through three different stages, after the basal state: Spinous, granular, and corneocytic. However, major changes in transcription and morphology occur at the first-stage transition. Once the cell reaches the cell surface, the cell is totally differentiated, an enucleated cell skeleton that is packed with strings of keratin filaments encased by a gamma-glutamyl-epsilon-lysine cross-linked cornified envelope of proteins. After that, the extrusion of a lipid bilayer will wrap-seal the body surface preventing the transference of any polar molecule(de Guzman Strong et al. 2006).

One of the most intriguing questions about skin proliferation is where do stem cells locate in the epidermis. The constant renewal of the stratum corneum is highly important to the maintenance of the adequate skin barrier function and then the proliferation of stem cells in the basal layer is crucial to achieve the main skin function. Although researchers have known that stem cells exist in the basal layer, there is still discussion whether all cells within the basal layer have the multipotent capacity or only a small number of them exist in this layer(Houben et al. 2007; Fuchs 2009). For over 30 years the basal layer has been defined as a succession of groups of 10 tightly packed basal cells (in process of differentiation) surrounding a multipotent stem cell (Figure 2). These groups of cells have been named Epidermal Proliferative Units, consisting in one self-renewing stem cell per EPU producing the rest of the basal cells, so-called transit amplifying cells(Elias 2007). However, other studies based on lineage tracing in mice tail keratinocytes have shown that, although most labeled cells were lost within the following 3 months, some of them survived and clonally expanded in size over time. This behavior is against the EPU hypothesis, and supports the idea of the differential expression of the same cell type depending on the surrounding environment(Clayton et al. 2007). This hypothesis is also supported by the fact that in the skin of furry mammals, the epithelial stem cells reside in the bulge of the hair follicle. Stem-cell progeny exits the bulge and migrates upwards into the basal layer. The rate of proliferation and migration is highly related with the skin state, being accelerated during wound healing(Taylor et al. 2000; Oshima et al. 2001).

According to the latter hypothesis, the stem cells located in the basal layer (or in the bulge of the hair follicles in the case of mammals) would divide generating one new stem cell and one daughter cell which would start the differentiation process while migrating through the basal lamina. However, the differentiation process becomes phenotypically apparent when the differentiated cell modifies its transit, and is released from the basal lamina and starts its vertical transit until its release by desquamation(Houben et al. 2007).

The expression profile changes during the transit to desquamation to maintain the homoeostasis regulation. The basal layer, conformed by early differentiated cells going out from the hair follicle, receives the poorly understood trigger or triggers that specify its differentiation programs (Fuchs 1995). This new differentiation program alters mainly the adhesive properties of the keratinocyte, and also modifies the cell morphologically and biochemically(Houben et al. 2007). The same stem cells will produce other far more complex structures, such as sweet glands and hair follicles. However, I will just focus specifically on the epidermis and their keratinocytes.

Keratinocytes mainly gear their metabolic machinery towards the production of keratins(Sun et al. 1983), and then, most of the expression changes observed during the keratinocyte differentiation will affect those proteins(Roop et al. 1983). Keratins are classically divided in two classes (Figure 2), according to their behavior in electrophoretic mobility and the isoelectric point(Fuchs 1995). As discussed below, it appears that keratins are expressed as specific pairs on the suprabasal keratinocytes, and will form heterodimers. Heterodimers will form heteropolymers and ultimately heterofilaments(Coulombe and Wong 2004). Keratin filaments will interact with actin microfilaments and microtubules. And together they will form the cell skeleton of the final state of keratinocytes that will maintain skin homoeostasis. All keratins seem to be encoded by different genes and only one case of alternative splicing has been reported(Langbein et al. 2010). Keratin expression is highly regulated at transcription level(Blanpain and Fuchs 2009).Basal cells periodically execute the same termination process, going through the same morphological states. The transition from the basal to suprabasal state (or spinous layer) is a key first step in the expression modification of keratins. Keratins 5, 14, and 15 (K5, K14, and K15), mostly expressed in basal cells, disappear by a total down-regulation of their expression, being replaced by K1, K10 and K11. It seems that K5 and K14 mRNA expression in mouse is highly restricted to cells that maintain their proliferative capacity(Byrne et al. 1994; Alam et al. 2011). In suprabasal

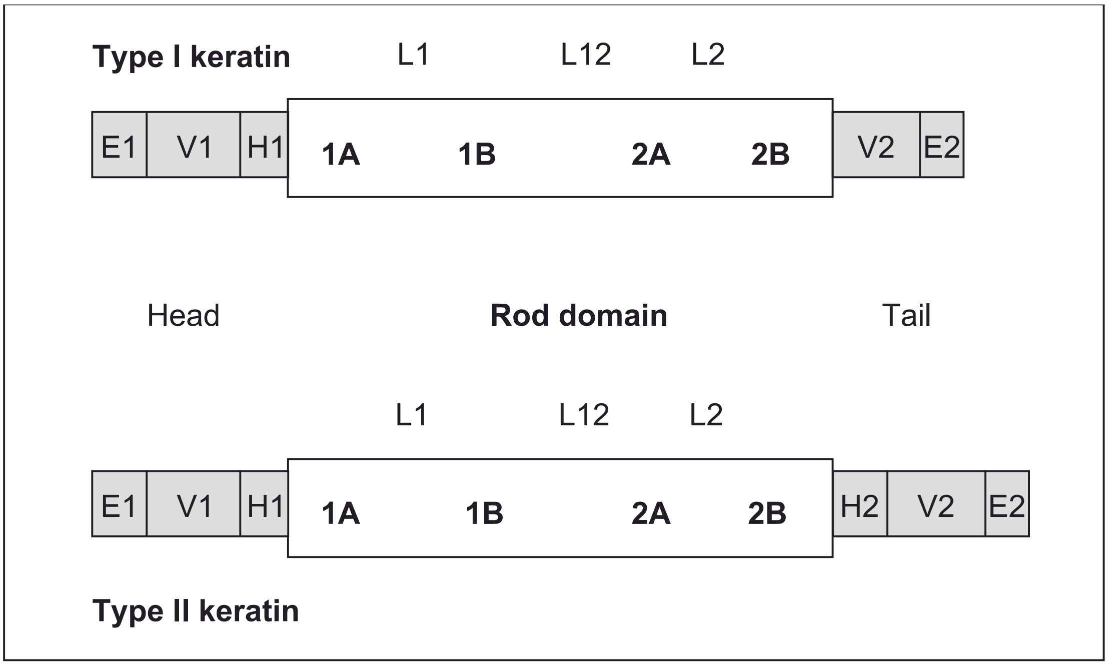


Figure 3. Structural organization of keratins.
Keratins contain an alpha helical rod domain which is important for polymerization. The rod domain is subdivided into four helical segments, 1A, 1B, 2A and 2B that are interrupted by three flexible non-helical linker domains L1, L12 and L2. The globular variable domains V1 and V2 vary widely in size and amino acid sequence among the individual keratins. Type II keratins contain in addition the homology subdomains H1 and H2 which are also thought to be important for polymerization. HIP and HTP are denoted by shaded boxes. Adapted from Arin and Mueller, 2007.

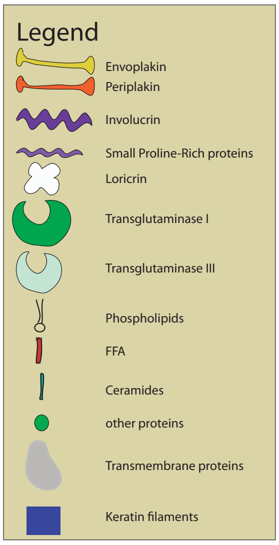
layers, some subsets of keratins are restricted to certain regions or physiological events, such as K9 in the palmo-plantar region, or K6 and K16 in wound healing, where hyper-proliferation is needed(Fuchs 1995). The over-expression of K6 and K16 has also been observed in hyper-proliferative disorders such as psoriasis(Leigh et al. 1995; Paladini and Coulombe 1998; Mommers et al. 2000).

Keratinocytes over the suprabasal (spinous) layer will start the expression of the enzymatic machinery to synthesize keratohyaline granules, which are basically composed by profilaggrin, a heteropolymer of 10-12 filaggrin molecules connected by link proteins and a final NH2-terminal filaggrin s-100(Matoltsy 1975). Filaggrin molecules, once released by proteolysis, will aid in the aggregation of all filaments mentioned above, will interact with the junctions to allow desquamation, and the terminal subunit of profilaggrin, the S-100 protein, will translocate to the nucleus, promoting the cell-cycle arrest and further cell death(Matoltsy 1975; Kuechle et al. 2000).

Cornification requires a massive activation of epidermal proteases, not only during the cleavage of profilaggrin. However, for most of them, the precise role remains elusive. Metalloproteases are mainly involved in the migration of other skin cells, such as Langerhans cells or melanoblasts, and also during wound healing. Serine- and cysteine- protease inhibitors are abundant and pivotal in skin, suggesting that both types of proteases are crucial in keratinocyte differentiation. But although the main role of each protease is not solved, proteases are known to be involved in three processes in skin differentiation: First, as I mentioned above, profilaggrin needs proteolytic processing before the release of filaggrin and S-100. Second, during the loss of the nucleus and other organelles, strong proteolysis eliminates their protein components. And third, during desquamation, corneodesmosomes, the junctional structures that mediate corneocyte cohesion, are cleaved through the enzymatic activity of, at least, two serine-proteases(Caubet et al. 2004).

Another important step through the final differentiation of the keratinocyte is the formation and assembly of the cornified cell envelope (CE), a 10nm-thick layered protein aggregate, located beneath the plasma membrane (Figure

4)(Kalinin et al. 2002). This CE will be triggered by elevated levels of Ca+2 (Figure 4), enhancing the assembly of two proteins, envoplakin and periplakin, forming heterotetramers which will form covalent bonds with a



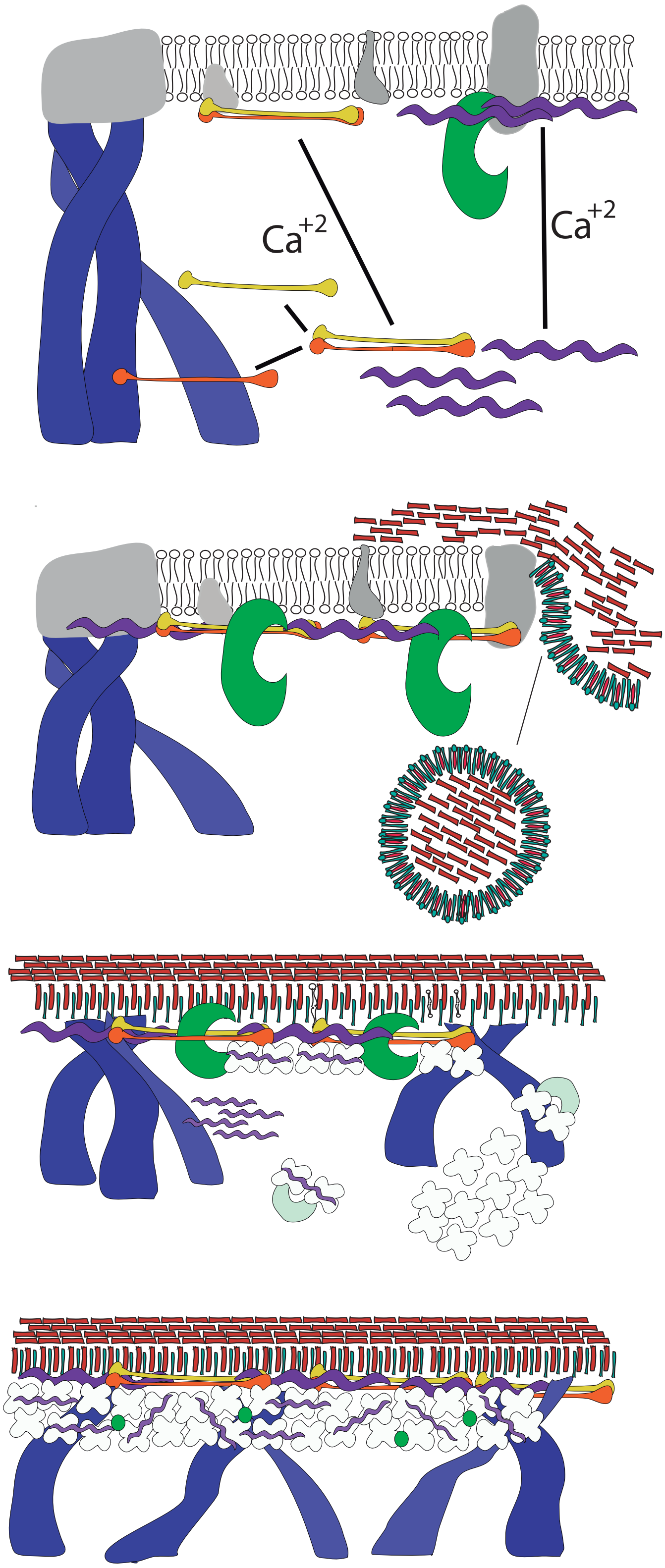


Figure 4. Model for CE assembly
See explanations in the text. Adapted from Kalinin et al. 2002

third protein unit, involucrin, which will serve as a connection with the plasma membrane through transglutaminase-1 membrane bond(Nemes et al. 1999). An initial structure based on tetramer units of periplakin and envoplakin will be formed at the cytosolic site of the plasma membrane, which will serve as a scaffold for the cornified envelope formation. At the same time, during the cornified envelope build, lipid precursors start forming; they will constitute the intracellular lipid domain, mainly formed by ceramides (60%), cholesterol (20%) and free fatty acids (FFA) (20%). A fusion of the surrounding membranes of the keratohyaline granulles with the plasma membrane also takes place, reducing the concentration of cholesterol in the plasma membrane to the complete replacement by hydroxyceramides(Bouwstra et al. 1998). Simultaneously, the filaggrin connected filaments will link with the cornified envelope through envoplakin, giving more resilience to the forming corneocyte against physical aggressions. The last component of the cornified envelope, loricrin, is assembled only when the intermediated filaggrin-CE complex is complete and the cornified envelope is assembled and attached to the ceramide-rich plasma membrane through involucrin. Loricrin is an insoluble protein expressed during the suprabasal state and located in the keratohyaline granules, which comprises 80% of the cornified envelope mass(Steinert and Marekov 1995; Kim et al. 2008; Kim and Leung 2012). Loricrin acts as a main reinforcement of the cornified envelope, together with a set of small, proline-rich proteins, being deposited onto the pre-existing scaffold, giving consistency and resistance. Alterations in the expression of the cornified envelope components are associated with inflammatory diseases such as atopic dermatitis (AD) or psoriasis(Kim and Leung 2012; Wolf et al. 2012).

Two more processes occur during cornification. First, the cornified keratinocytes, now called corneocytes, are embedded in an intercellular lipid matrix. This matrix, composed mainly of ceramides and cholesterol, forms at least two different lamellar domains with repeat distances of 6 and 13.2nm(Bouwstra et al. 1995). This structure has been called the sandwich model, where a narrow fluid lipid phase is interspersed between two broad crystalline lattices, basically formed by ceramides and cholesterol. This structure is constructed to repel polar substrates and water, improving the hydric isolation of the body and preventing dehydration(Grubauer et al. 1989).

Second, during the terminal differentiation of the keratinocyte, all organelles are destroyed and DNA is degraded, resulting in the death of the cell(Candi et al. 2005). However, in this case, the process does not follow any of the two pathways that lead to apoptosis, and the mechanism that leads to the corneocyte cell death is not considered apoptotic, despite the fact that the cell death is programmed from the moment the keratinocyte leaves the basal lamina and goes through differentiation(Takahashi et al. 2000). However, corneocytes are not phagocytosed, the cytoskeleton is not broken but reorganized, and the nucleus is destroyed through a different pathway, not involving the same caspase cascade(Nagata et al. 2003). The nuclei of corneocytes do not show chromatin condensation or DNA laddering(Lippens et al. 2005). The nucleus disappears completely, leaving no remnants in the corneocytes, while in apoptotic cells the nucleus is engulfed, the chromatin condensed and fragmented with the rest of the cell in the apoptotic bodies. DNA degradation in corneocytes is performed by DNase 2(Fischer et al. 2011), while apoptosis DNA degradation is executed by a caspase-activated DNase.

Once the corneocyte rises to the surface, all cell-cell junctions are cleaved in order to release the damaged cover, being replaced by a new, integer one. There are three different types of interactions between the cells in the skin: Hemidesmosomes, corneodesmosomes, and adherens junctions. Hemidesmosomes are exclusive of the basal keratinocytes and link to the basal lamina(Borradori and Sonnenberg 1999; Koster et al. 2004). Corneodesmosomes function as adhesive complexes integrating the keratin intermediate filaments between cells via cytoplasmic plaque proteins plakoglobin, desmoplakin and plakophilin. The extracellular proteins of the desmosomes, called desmoglein and desmocollin, are proteins of the family of cadherins. Adherens junctions serve as helpers of the desmosome complexes, by linking the actin cytoskeleton of the different cells, next to the desmosome surroundings(Ishiko et al. 2003).

During keratinocyte differentiation, desmosomes also suffer changes in their properties and abundance. When keratinocyte is located in the stratum corneum, desmosomes contain a novel glycoprotein, the corneodesmosin, which will link covalently to the cornified envelope.

When it is the turn of the mature corneocyte to be released, it starts the rupture of the cell-cell junctions by enzymatic proteolysis(Lin et al. 2012). As mentioned above, desquamation proteolysis is associated with serine proteases of the kallikrein family. Activation of proteolysis is shown to be sequential. The tryptic enzyme activates the inactive chymotryptic enzyme proform, in order to degrade the desmosome subunits and the adherens junctions. This serine proteolysis is accompanied with a slightly increased pH, which is related with an attenuated fatty acid content in the stratum corneum(Caubet et al. 2004). This increased pH level results in a rapid activation of serine proteases, which have an alkaline optimum pH for activity(Fluhr et al. 2001; Hachem et al. 2003; Gunathilake et al. 2009).

Once the cell-cell junctions are disconnected, the cell remainings will stay over the skin until a mechanical aggression eliminates them. During this time, skin remainings are a source of nutrients for an ensemble of microorganisms, from bacteria to microscopic arthropods. Microorganisms are widespread in all the environments on Earth. Given their ecological ubiquity it is not surprising to find many microbial taxa in close relationship with members of many eukaryotic taxa, even establishing associations of mutual benefit(Moya et al. 2008). However, to achieve that benefit, both sides must teach and train each other to allow coexistence.

The other inhabitants.

Defense mechanisms against biological attacks

Terrestrial chordates are continuously threatened with desiccation, injury, UV irradiation, and other physical and chemical insults. The structural to molecular mechanisms explained above comprise a major thrust of prior research in skin as a barrier, protecting the whole body against this kind of aggressions. However, the epidermis mediates a broad set of protective functions against microbe challenges(Elias 2007). Although we have focused the last paragraphs to explain the process of keratinocyte differentiation, given that it is the main source of defense, it is important to understand that skin hosts a vast diversity of host cells, being most of them part of the immune system.

Although the stratum corneum barrier is supposed to be impenetrable, some pathogens have developed systems to penetrate the physical barrier(Schaller et al. 2000; Monod et al. 2002). It is then not surprising that the most integrated protective systems of the epidermis are the permeability barrier and the antimicrobial defense(Aberg et al. 2007). One of the interesting facts about the permeability barrier is that, except in particular cases, the extracellular matrix is the main pathway through which bacterial pathogens breach the stratum corneum(Miller et al. 1988; Schaller et al. 2000; Rouse et al. 2005). Lamellar bilayers serve as an important, both chemical and physical, barrier. As noted above, intercellular lipids forming the lamellar bilayer contain free fatty acids and sphingosine, which exhibit themselves a potent activity against a wide range of pathogens. Several non-lipid antimicrobial proteins, including cathelicidin LL-37(Braff et al. 2005a; 2005b), and beta-defensin hBD2(Oren et al. 2003) are delivered to the lamellar bilayers by the stratum corneum, to the point that antimicrobial peptides appear to contribute to the supramolecular organization of the extracellular matrix.

Two more physical characteristics are crucial to the fight against pathogen aggressions. First, the acidic character of skin surface pH has been noted both in vivo and in vitro. The origin of the stratum corneum acidity comes mainly from two different origins: First, the complete hydrolysis of epidermal phospholipids and the free-fatty acid formation during the organelle extrusion, which are critical, not only for the structure of the lamellar bilayers but also to lower the pH of the surrounding environment by at least one unit. This hydrolysis of phospholipids into free fatty acids helps to form the acidic milieu, so hostile to invading pathogens(Mao-Qiang et al. 1995; Fluhr et al. 2001). The second origin of acidity is the non-energy dependent sodium-proton exchange mechanism, NHE1(Behne et al. 2002; Hachem et al. 2005). This transporter is ubiquitous in all outer nucleated layers, and acidifies the extracellular domain in the lower stratum corneum. The activity of NHE1 is reduced with age, with a resultant increase in pH, and then the risk of pathogen invasion is increased in aging skin(Choi et al. 2007).

The second physical characteristic is the low hydration of the stratum corneum(Warner et al. 1988). The water content of the stratum corneum drops drastically, creating dissecating conditions, which limit pathogen colonization. However, some bacteria seem to thrive under those conditions, for reasons that remain unknown. In contrast, more occlusive regions suffering super-hydration, such the axilla or the groin, are more prone to develop mild infections such as folliculitis by *Streptococcus sp.* and *Staphylococcus sp*. Moreover, filaggrin proteolysis is inhibited under over-hydration conditions, resulting not only in an increased surface pH, which is beneficial for opportunistic pathogens, but also in altered stratum corneum moisturizing. Filaggrin proteolysis generates as secondary products a wide range of metabolites collectively known as natural moisturizing factors. This reduction in the moisture point produced in over-hydration conditions could be a response to maintain skin homoeostasis, reducing the self-products generated to increase humidity(Harding and Scott n.d.).

The immune sentinels of the skin.

Being the skin the first interphase between the environment and the body, and then the first defense layer against environmental assaults, one may think that only structural/biochemical defenses are not sufficient to deal with such a broad range of possible external aggressions. We live in a hostile environment surrounded by microbial organisms, most of them potential pathogens. To survive in this kind of pathogenic broth, vertebrates have evolved an immune subsystem in their skins. As the largest body organ, the skin has a central role in the immune defense, and then, its mechanical defenses are reinforced by a versatile and robust system of immune surveillance(Kupper and Fuhlbrigge 2004). The crucial role of immune surveillance in maintaining system homoeostasis is evident from the strong relationship between cutaneous malignancies and infections when the immune function of the skin is limited, such in cases of acquired or provoked immunodepression(Lugo-Janer et al. 1991), or the relationship between the missregulation of the immune activity in a wide variety of inflammatory disorders, such as psoriasis or AD, and the increased susceptibility to infections(Wang et al. 2008; Ivanov et al. 2009).

As in a systemic point of view, we can divide the skin immune system into innate or acquired. Although the previous sections were focused on keratinocytes, skin is way more complex at cellular level, with a strong variety of cell types that interact with each other to maintain homoeostasis(Simpson et al. 2011). Immune surveillance is performed by resident immune cells, which function as sentinels against the possible invasion of potential pathogens. Nucleated keratinocytes, Langerhans cells and mature dendritic cells provide an early warning system and a first defense, producing a wide variety of molecules that act as a first innate defense and signal to attract the immune effectors(Nestle et al. 2009). Keratinocytes are often under-appreciated participants in cutaneous immune responses, but produce indeed huge variety of effectors against external assaults(Kupper et al. 1987; Kupper and Groves 1995; Wood et al. 1996).

Keratinocytes, similar to gut epithelial cells, can sense pathogens and respond to block the first steps of infections. In this line, eukaryotic cells sense microbial products in two different ways. Pathogens are recognized by epithelial cells through the molecular recognition of evolutionary conserved microbial components called Pathogen-associated Molecular patterns (PAmPs), which include the lipopolysaccharide, peptidoglycans, flagellin, or specific patterns of nucleic acids(Janeway 1989; Medzhitov 2009). Keratinocytes express two different PAmP recognition receptors: Toll-Like Receptors (TLRs) and Nucleic-binding Oligomerization Domain containing proteins (NOD). Toll-like receptors are a family of conserved membrane-spanning molecules that contain an ectodomain of leucine-rich repeats, a trans- membrane domain and a cytoplasmic domain known as the TIR (Toll/ IL-1 receptor) domain(Gilliet et al. 2008). These receptors lead to the activation of host cell signaling pathways and subsequent innate and adaptive immune responses. Keratinocytes are able to express a wide variety of TLRs, which specifically detect different PAmPs from different pathogen types. TLR recognition will send a signal that will activate two main pathways: First, the production of molecules which deal with the infection by themselves, such as cytokines and chemokines, which have a paracrine activity. And second, the translocation of the PAmP to the rough endoplasmic reticulum where it will be prepared to be presented in the surface. The keratinocyte is considered as a non-professional antigen-presenting cell(APC), because of its production of major-histocompatibility complex class II genes (MHC-II). MHC is one of the most studied regions of the mammalian regions (Figure 5); it is related to inflammation and immunity, located at the short arm of the chromosome 6, and comprises a cluster of duplicated and highly polymorphic genes related in self-recognition(Horton et al. 2004; Bentley et al. 2009). The presentation of the PAmP as foreign, through the MHC-II, will result in the recognition by professional APCs, such as dendritic and Langerhans cells, and will ultimately result in the activation of a

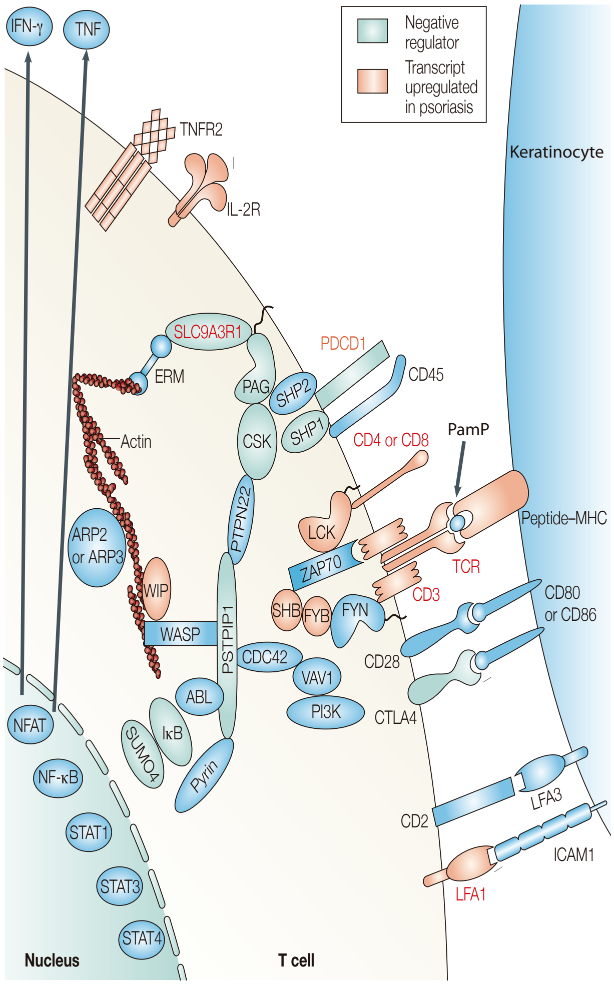


Figure 5. Molecular mechanisms involved in during the keratinocyte Antigen presentation. After formation of the immunological synapse between a T cell and a keratinocyte, the T-cell activation is regulated by a complex network of positive and negative (green) regulators. Proteins that are involved in the immunological synapse and have up-regulated expression.

Reproduced from Bowcock

T1-type immune response and the production of type-1 interferons(Lebre et al. 2006; Miller and Modlin 2007).

Once the keratinocyte presents the PAmP through MHC-II, it will be initially recognized by a specific type of dendritic cells: the Langerhans cells. Those cells will migrate to the nearer lymph node where they will present the antigen to their antigen-specific T cells. In vitro studies have shown that Langerhans cells process the antigens presented by keratinocytes for being presented to the effector T-cells(Seth et al. 2005) and naïve B-cells. Moreover, Langerhans cells preferentially induce the differentiation of T-helper cells. Langerhans cells, with the T-cells and the keratinocytes, will be the mediators of the immune response against the pathogen assaults.

But keratinocytes are not only antigen presenting cells (Figure 6). As explained above, keratinocytes are an important source of cationic antimicrobial peptides, mainly ß-defensins and cathelicidins that actively fight pathogens. Antimicrobial peptide production is enhanced by different processes. During skin infections, keratinocytes are able to recognize the pathogen assault through PAmP recognition through TLR response, resulting on the activation of the expression of multiple immune signaling molecules, such as interleukins (IL-1,6,10,18 are some examples) and tumor necrosis factors(Albanesi et al. 2005). Those molecules, in particular IL-1, are pleotropic cytokines with a broad range of biological effects, including the activation of local T-lymphocytes and Dendritic cells and the promotion of B-lymphocyte maturation and clonal expansion(Wei et al. 2005; Ehrchen et al. 2010). The different TLR activation results in different immune associated responses, including chemokine and chemokine receptors. For instance, TLR3, 4, 5, and 9 activation by a wide range of PAmPs results in the activation of TNF-alpha, CXC chemokine ligand (CCL) 8, CCL2, and CCL20 expression. In contrast, TLR3 stimuli by poly-I:C induces the expression of type I Interferons and CCL27, but not other PAmPs. Upregulation of ICAM-1, HLA-DR, HLA-ABC, and CD40 was observed in response to flagellin and LPS. All those stimuli resulted in the phosphorilation of IKBa following the translocation of NF-kB to the nucleus, provoking the regulation of the immune production of IL-1a, IL-1b, IL-8, IL-9, and TNFa(Lebre et al. 2006). Keratinocytes are pro-

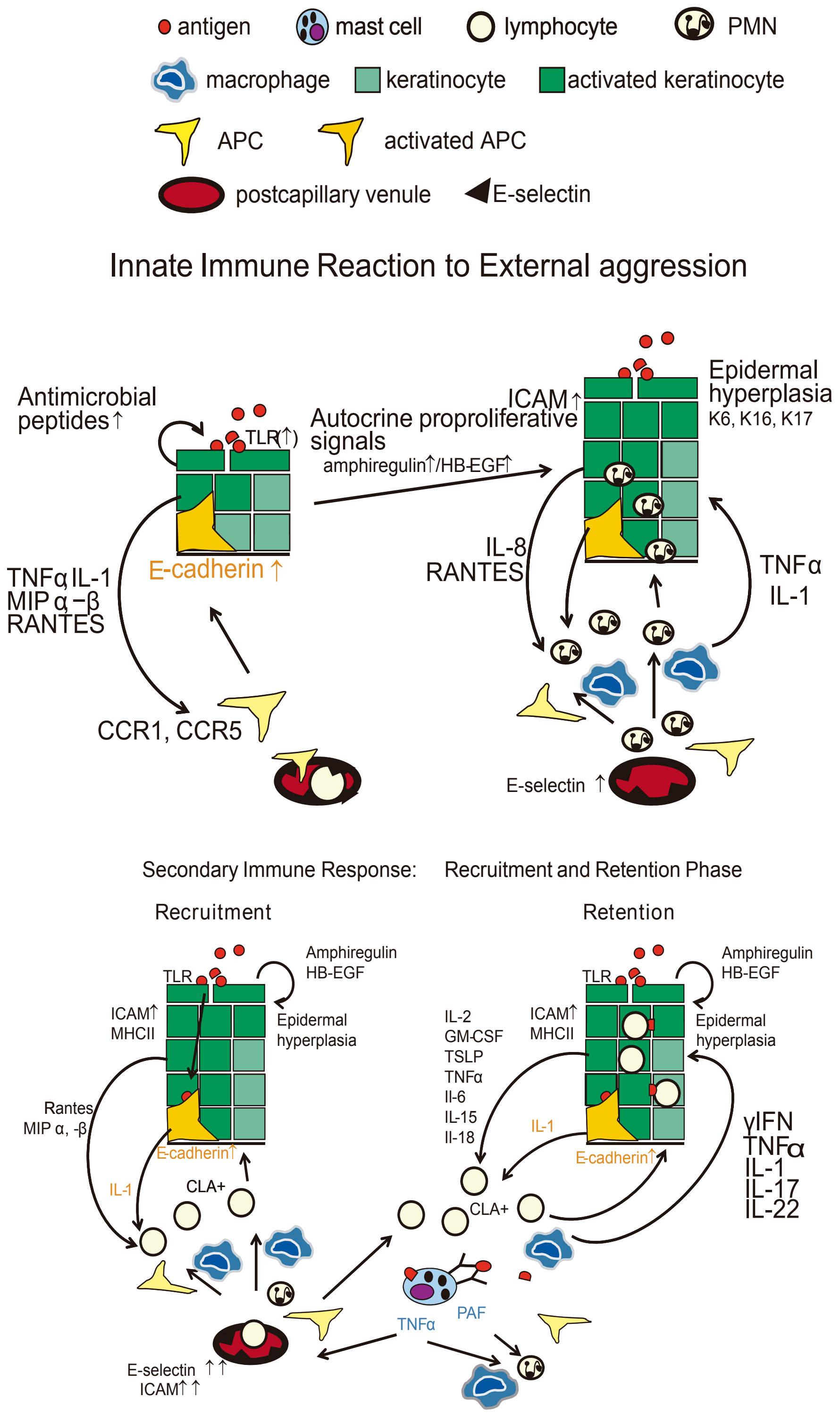


Figure 6. Simplified schematic representation of cutaneous immune cell traffic with emphasis on innate and adaptive keratinocyte immune function

In response to the antigen recognition by TLR, the keratinocyte releases chemokines that will attract other APC precursors into the epidermis, and will activate other keratinocytes through autocrine mechanism. APC and keratinocytes will secrete pro-proliferative factors that will induce epidermal hyperplasia. APC will attract inflammatory cells by secretion of chemokines. Inflammatory cells will adhere to kerationcytes through ICAM-1, and will induce the immune response against the possible pathogens. Adapted from Suter et al.2009 inflammatory effectors, strategically positioned at the outermost layer of the body to react to possible insults through the coordinated production of antimicrobial peptides, cytokines and chemokines, according to the type of infection(Nestle et al. 2009).

One of the most important pathways regulating the skin homoeostasis is the NF-kappaB pathway. NF-kB pathway receives its signal through TLR-PAmP recognition and TNF stimulation. Both stimuli activate a signaling cascade that results in the translocation of NF-kB to the nucleus and the regulation of inflammatory pathways(Wang et al. 2001; Sato et al. 2005; Omori et al. 2006). Alteration in any of the steps of the signaling cascade in keratinocytes spontaneously triggers a strong inflammatory response, establishing a crucial role for NF-kB signaling in keratinocytes in maintenance of the skin immune homoeostasis.The study of chronic inflammatory diseases, such as psoriasis or AD, has shown that disruption of the NF-kB pathway is also associated with the autoimmune response(Masters et al. 2009).

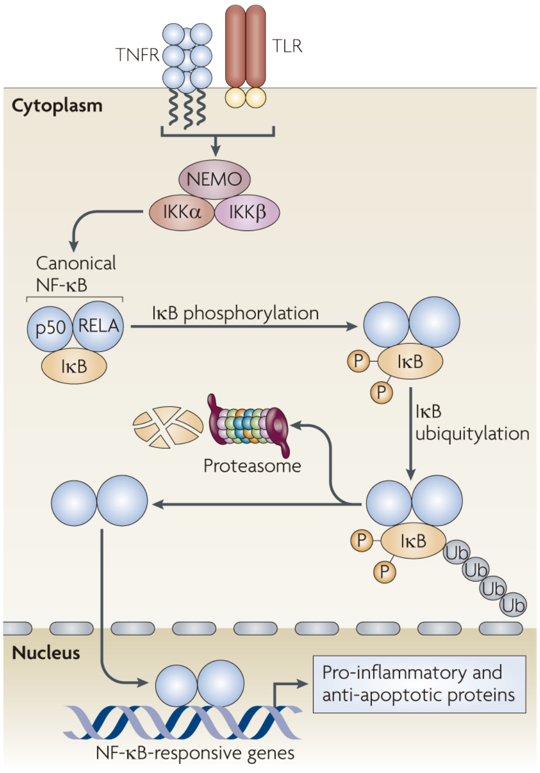


Figure 7. Schematic depiction of the canonical NF−κB. Nuclear factor-κB dimers are normally sequestered in the cytoplasm of resting cells by association with inhibitor of NF-κB (IκB) proteins. Pro-inflammatory signals stimulate receptors belonging to the tumor necrosis factor receptor or IL-1/TLR superfamilies, which activate the IκB kinase (IKK) complex. The IKK complex phosphorylates IκB proteins on specific serine residues, thereby triggering their polyubiquitylation (polyUb) and proteasome - dependent degradation.

Reproduced from Pasparakis et al. 2009

The maintenance of the skin homoeostasis depends on an extensive cross-talk between the keratinocytes and the rest of the epithelial cells(Omori et al.

2006). Alteration in NF-kB may disrupt this communication. However, the cellular and molecular mechanisms by which NF-kB inhibition disturbs the skin immune homoeostasis remain elusive, but it seems that impaired NF-kB signaling in keratinocytes results in a similar response to the one observed during wound healing and other hyperproliferative disorders, such different as squamous cell carcinoma or psoriasis(van Hogerlinden et al. 1999; Lind et al. 2004). It is probable that the NF-kB altered epidermis releases stress signals, that are interpreted by the rest of immune cells as skin lesions and injuries, eliciting a wound-healing response, basically defined by a marked hyperproliferation of apparently normal keratinocytes. Nevertheless, the fact that TNF-targeted therapies in psoriasis are still efficient(Mössner et al. 2008), leads to the question of whether something else is involved and triggers the injury signaling on those hyperproliferative disorders.

Our tenants, the inhabitants of the skin.

A historical introduction

As we explained in depth above, the skin has a strong arsenal against possible pathogen assaults. This fact can be translated to all the other epithelia. And still, the human microbiota consists of approximately 100 trillion (10^{15}) microorganisms, including bacteria, archaea and microeukaryotes(Hooper et al. 2002). So, the question that emerges, given this fact, is how do they manage to live in such an actively harmful environment. One may think that after a long co-existence and co-evolution between host cells and the microorganisms that attempt to colonize the body, a set of microorganisms consolidated into a skin commensal/ mutualistic microbiota. This microbiota is distributed in multiple niches, depending on the amount of nutrients and the physical properties that might result in the most suitable growth conditions for them(Moya et al. 2008). The improved understanding of the microbiota, mainly focused on the gut environment, indicates that the traditional dichotomy of commensal versus pathogen is clearly too simplistic. Rather, some bacteria, such as *Lactobacillus* or *Bifidobacterium*, are highly associated with good health, and a large number of studies in animal models supports t this benefit. Other microorganisms, such as E.coli, cause no harm to healthy hosts but are associated with, and have the capacity to exacerbate, disease, and can be viewed as opportunistic pathogens(Klijn et al. 2005; Barnich et al. 2007; Carvalho et al. 2009).

The study of microbial communities was first revolutionized by the study of the composition through the analysis of variability of a highly conserved genomic region in bacteria(Woese and Fox 1977). Since it was proposed as a universal phylogenetic marker of the bacterial domain, sequence analysis of the small subunit of the rRNA gene has become the gold standard for the assessment of microbial diversity within microbial samples(Hugenholtz and Pace 1996). Similarly, the study of host-associated microbiota was revolutionized by the use of 16S rRNA gene amplification and sequencing(Suau et al. 1999). All these methods started being associated to cloning-based approaches and Sanger sequencing(Schmidt et al. 1991). However, the beginning the "Omics" era, with the sequencing of the reference human genome in 2001, opened a new window of research, increasing our understanding of biological entities at all levels, from molecular to ecological(Venter et al. 2001). This milestone uncovered the necessity of a new set of technologies to produce massively-parallelized sequencing data, avoiding the cloning steps, and all the bioinformatical tools needed to manipulate, analyze and compare all this new information from a wide variety of genetic sources. Since this first genomic approach, substantive changes have occurred. High-throughput sequencing technologies, commercially appeared on 2004 (Margulies et al. 2005; Mardis 2008), claimed to reduce costs, avoiding the cloning bias and increasing the amount of data obtained. The technological race between the different high-throughput sequencing platforms ensued, and has continued with one main goal: The obtention of a technology inexpensive enough to allow any genetic sample to be sequenced, analyzed and understood(Wolinsky 2007). With the possibility to deeply sequence a sample without the need of performing cloning libraries, the times and costs to sequence and analyze the bacterial/viral composition of a sample just dropped drastically. More interestingly, this technology raises the possibility to characterize structures, functions, and dynamic relationships among the members of a microbial community. In microbial ecology, the disappearance of the need for culture has resulted in the emergence of a new discipline, called metagenomics(Venter et al. 2004; Remington et al. 2005). This discipline takes molecular ecology to a new level, allowing to assess the molecular machinery of a microbial community to survive and interact with the environment, and, in case this environment is a specific host, to assess the specific relationships among host and microbiota, allowing to answer specific questions such as how do microbes avoid the host defense mechanisms or which benefits the host receives for housing that precise microbial community and not any others.

It is accepted that, in utero, the human embryo is sterile, as is probably the case in most metazoans (Pflughoeft and Versalovic 2012). The first year of human life provides an attractive window of opportunity for microbes to colonize a new environment full of possibilities, with the almost sole defense of the innate immunity. During birth, the newborn will receive microbial communities from the mother's vaginal skin, and lactation will put it in contact with the milk community(Palmer et al. 2007; Eggesbø et al. 2011). Complexity will increase during infancy as the host contacts and interacts with other communities. As the community increases in complexity, the interactions and relationships with the host also increase(Hooper et al. 2002; Carvalho et al. 2012). One of the most striking findings that helped to define these complex interactions and relationships between host and microbiota was the role of gut microbial composition in energy harvest. Multiple studies relate the relationship between adiposity and the Firmicutes/Bacteroidetes ratio in mice and humans(Ley et al. 2005; Turnbaugh et al. 2009). Other types of interactions have been acknowledged. Interestingly, commensal microbiota from the gastrointestinal (GI) tract are known to produce metabolites, such as short-chain FFAs, by anaerobic fermentation, to provide substrates for oxidative metabolism in epithelial cells. The same FFAs provide protection against bacterial pathogens(Bengmark 1999).

Skin microbiota, the undiscovered continent

All these examples raise the idea that microbial communities are characterized by unparalleled complexity. Although our technological ability to characterize this complexity contributes to understand the ecological processes that drive microbe-host interactions, we are still far away from a complete understanding. Most studies have been focused on mucosal tissues because samples can be easily collected (feces, vaginal mucus, oral saliva), but the true interactions between host and microbiota have not been yet elucidated, due to the complexity to separate host from microbial DNA, making the direct sequencing expensive in terms of information and costs. This is especially true for skin. As described above, skin is a dry tissue, prepared to avoid all kinds of bacteria. Considered as an ecosystem, the skin is a 1.8m2 surface of diverse habitats with an abundance of folds, invaginations and specialized niches that support a wide range of microorganisms (Figure 8), according to diversity analysis. The physical and chemical characteristics of this ecosystem select for unique sets of microorganisms adapted to the niche they habitat. Skin is cool, acidic and desiccated, factors that, as we have explained above, protect against most microorganisms. Skin is the first barrier in contact with potential commensal microbiota and its main cells are key to the development of the immune system. And still the commensal microbiota associated with the skin is almost unknown(Kong and Segre 2011).

The study of the skin microbiota has a long tradition(Leyden et al. 1981). The involvement and relationship of bacteria in certain physiological traits, such as body odor, or the association between specific bacteria and certain diseases, are known for almost fifty years(Marples et al. 1974; Leyden et al. 1981). But these relationships have always been observed as direct metabolic links between one specific microbe and a certain host trait. The rest of the ecological variables, like the other commensals, or the host genetics are always underestimated or forgotten. The perception of the skin as an ecosystem, discreetly studied over the last 40 years(McBride et al. 1977; Fredricks 2001; Grice and Segre 2011), can improve our understanding of the relationships between host phenotypes and the causes or the consequences on the microbiota(Grice and Segre 2011). To further our understanding on health, disease and skin infection, it is important to interconnect our knowledge of the skin immunology, genomics and microbiology, and develop a better understanding of the ecosystem itself and the consequences in all its components when a phenotypic change occurs. It is important to remember that the same environmental factors that affect the

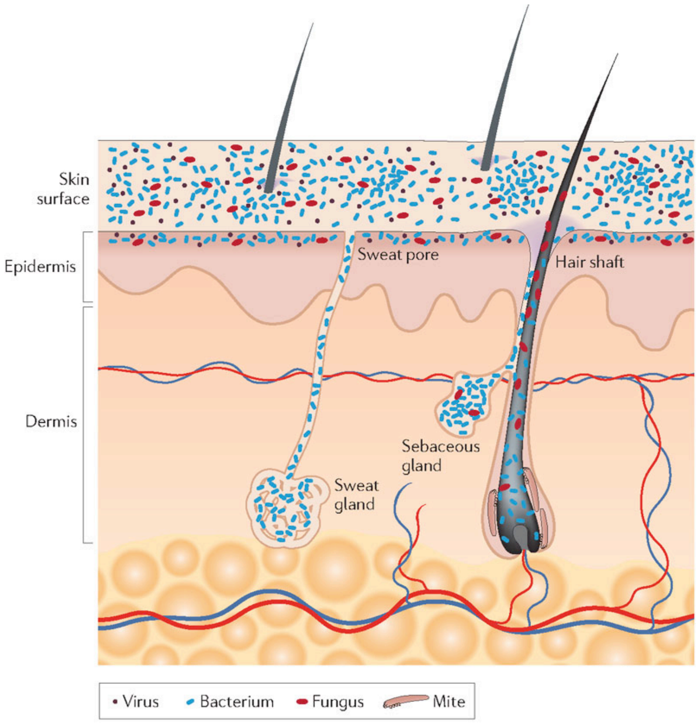


Figure 8. Schematic of skin histology viewed in cross-section with microorganisms and skin appendages. Microorganisms (viruses, bacteria and fungi) and mites cover the surface of the skin and reside deep in the hair and glands.

Reproduced from Grice et al. 2011

host, may also affect the microbiota, changing the composition or their metabolic activities. Occupation, clothing or antibiotic usage may modulate colonization by the skin microbiota(Grice and Segre 2011). Although we accumulate knowledge on environmental factors for other host-associated ecosystems, a similar assessment of skin microbiota in healthy individuals does not exist(Dethlefsen et al. 2008; Dethlefsen and Relman 2011). Other characteristics in human behavior, like the use of cosmetics, soaps and other hygienic products may also have their effect on the skin microbiota composition, and still their effects on it remain unclear. Quantitative culture studies demonstrate a strong variability associated with the range of light exposure, humidity and temperature on the surface of the skin, resulting in a wide range of ecological niches on the skin of a single individual(McBride et al. 1977), as different as a rainforest to a desert(Marples 1969). Only in the last five years, skin has become of ecological interest, and with the radiation of the field of metagenomics, skin microbiota is now starting to be understood(Grice et al. 2008).

The first approaches to characterize skin bacteria, and I stress in just bacteria, have revealed a much greater diversity than that showed by culture-based methods(Gao et al. 2007; Fierer et al. 2008; Grice et al. 2008; Costello et al. 2009; Grice et al. 2009), as expected by the results in other host-associated environments. As defined by 16S rRNA sequencing, most skin bacteria fall into four different phyla: Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria. Similar results are found in other mucosal surfaces, but the relative abundances of each is different for skin: while Actinobacteria and Proteobacteria are more abundant in skin(Grice et al. 2008; 2009), Firmicutes and Bacteroidetes are prevail in the oral cavity and in the GI tract(Gill et al. 2006; Belda-Ferre et al. 2011). And still, it seems a common feature of host-associated environments to present low diversity at high taxonomic levels (phylum), but high diversity at low levels (genus, species,...).

We can study bacterial diversity at three main levels: Intraindividual, temporal and interpersonal. One of the more important features in skin microbial diversity is the high dependence of diversity and body site. It seems that colonization of bacteria in skin is highly dependent on the physiology of the region, with specific species associated with it, depending on the level of humidity, secretion, or pH, suggesting that there are local selective pressures that allow only a few phylogenetically homogeneous groups, called phylotypes, to colonize certain body regions. Sebaceous sites, such as the forehead,the retroauricular crease or the back, are regions with low phylogenetic diversity (6-20 phylotypes), being dominated by *Propionibacterium ssp.*, confirming *Propionibacterium ssp*. as lipophylic residents of the pilosebaceous units(Grice et al. 2009). In contrast, *Corynebacterium ssp* is the most common bacteria of moist areas, maintaining Actinobacteria as the most prevalent phylum of those areas, leaving the prevalence of Proteobacteria for the dry areas, such as the forearm, knees, or back of the hand. Those dry areas are also the most diverse ones, harboring even more phylogenetic diversity than the gut or the oral cavity(Costello et al. 2009).

While Firmicutes are not a main phylum in skin, it is important to emphasize that in all previous studies, *Staphylococcus ssp*. are also prevalent in moist areas, consistent with culture data. Staphylococci occupy an aerobic niche on the skin and probably use urea from sweat as nitrogen source. Together with Corynebacteria, processing of apocrine sweat by staphylococci would result in the characteristic malodor associated with sweat in humans(Leyden et al. 1981).

Another characteristic of skin is the higher temporal variability compared with other host-associated environments. Being highly variable, skin also harbors temporal variability across body sites, being the most occlusive ones more consistent over time. In general, sites that harbor a greater diversity tend to be more associated with host-environment manipulation and interaction, and then they are less stable across time(Grice et al. 2009). And still, compared to the gut and oral cavity, the skin microbiota has the greatest variability over time(Costello et al. 2009).

In general, intrapersonal variation in microbial community diversity between equivalent skin sites is smaller than the interpersonal variation, according to 16S rRNA sequencing(Gao et al. 2007; Costello et al. 2009; Grice et al. 2009). While similar phylotype content is found between individuals, in equivalent skin regions, this common content is reduced. Gao et al. presented a proof of concept with only 6 individuals, showing that 68.1% of genera were uniquely associated to one subject(Gao et al. 2007). Fierer et al. presented similar results in hand microbiota, resulting in only 17% of species-level shared phylotypes between individuals, suggesting the possible use of bacterial diversity as a forensic tool(Fierer et al. 2008; 2010).

Comparison of skin microbial diversity among genders showed a greater diversity associated with female hands(Gao et al. 2007; Fierer et al. 2008). However, the reason why women harbor more diversity remains unclear. It is possible that this greater diversity is due to a combination of hygienic/ cosmetic differences with physiological factors.

Considering the situation of skin as the outer layer of the body, the high interpersonal variability can be framed as made up of two components: a low number of shared taxa, which are stable and helpful inhabitants of the skin, and are accompanied by a high proportion of transient species that fill the metabolic niches that remain unoccupied by the stable residents(Grice and Segre 2011). Factors driving variability are still unclear, but environmental, historical and host factors may be crucial to understand the reasons why some bacteria are present in one host and not the others.

Knowledge gaps

Although high throughput sequencing of metagenomic DNA is available, to date most studies performed are based on 16S rRNA amplification, which, indeed, provides a less biased description of skin microbiota than culture-based assays, but it misses some important information. First, the molecular approaches that are currently in use for skin are unable to distinguish between living bacteria and dead organisms(Kong and Segre 2011). This information may be important to survey the history of skin microbiota, but does not give important information, such as which bacteria are interacting with the host cells.

Second, as the skin is an exposed organ, 16S rRNA analysis does not discriminate between resident and transient bacteria. When one wants to assess the interactions of resident bacteria in a certain host phenotype, such as physiological characteristics or a complex disease, this question is crucial. Many common skin disorders are postulated to have an underlying microbial contribution. But this contribution does not satisfy Koch's postulates and then a new perspective is needed to understand how microbiota affects host cells, resulting in an altered host phenotype. Although classical studies relate a certain bacteria to a phenotype, like teenage acne and *Propionibacterium acnes*(Dessinioti and Katsambas 2010), molecular methods to characterize the microbiota associated with acne has been limited and did not relate acne to any novel association with bacteria, and the relationship with *P.acnes* was also limited (Bek-Thomsen et al. 2008).

Another case is AD, a chronic relapsing disorder that affects 15% of children population in Spain, and is associated with abnormal host-microbiota relationships. The fact that AD is more prevalent in industrialized and more hygienic environments raises the intriguing possibility that skin microbial fluctuations modulate the gene-environment interaction on the skin surface, resulting in episodic exacerbations of the disease(Kong et al.

2012). This hypothesis raises, then, the most important limitation of skin studies: If skin microbiota is related directly or indirectly to a certain disease by genomic interaction with the environment or the host, it is important to describe the gene content of the microbiota in a certain region affected by a certain disease. And still, this is yet unfeasible because of two main reasons. The first question is methodological. Skin microbiota occupies all possible niches, including sebaceous glands and follicles, even deep layers of the epidermis. To assess the bacteria in all those environments, deep biopsies are needed, but extracting the whole DNA from that sample, results in a very low ratio of bacterial/host DNA, which is translated in a very low amount of bacterial reads, which makes this methodology expensive and inefficient. The second reason is conceptual. The analysis of skin microbiota by any methodology (even in case we are able to assess the gene content of the skin microbiota) has to be considered as a single time point. Even when we observe a change in diversity(Gao et al. 2008) we need to asses whether that variation is a cause or a consequence of the phenotypic change, and discern when a diversity shift is spurious or related to that phenotype in a causative manner.

Filling the gaps

In this work we will try to cover, and solve, all the weak points in the metagenomic study of the skin. First, we will analyze the host metabolism and its effect on the microbiota. Since dataset comprising both host metabolic data and skin microbiome does not yet exist, we will test the methodology on the GI tract dataset from MetaHIT, which is the only project with enough information to carry this type of studies.

Second, we will assess the effect of the immune system, and its impairment, in the skin bacterial diversity. We have worked with SCID mice, which have a rare mutation in the gene *Prkdc*, which inactivates the V(D)J recombination, resulting in the inability for these mice to get mature B and T cells, which leads to an impaired acquired immunity. To assess the effect of the acquired immune system on the microbiota, we will compare the taxonomic diversity of SCID mice skin with a healthy dataset.

Third, we will try to solve the host contamination on skin metagenomic studies, by devising a method for microbial enrichment in skin biopsies. This method has been applied to the skin of healthy mice, and has been extensively validated.

Forth and last, we will try to solve the problem of the low yield of bacterial DNA obtained from skin, which has been one of the most widespread weak points of every study of skin microbiota, limiting skin microbiome studies only to 16S rRNA diversity.

Objectives

General Objectives

To devise a standard, reproducible, methodology to produce unbiased reliable metagenomic data from skin biopsy samples, allowing to analyze them according to the general ecosystems perspective.

Specific Objectives

1. Construct a library from trace amounts of DNA. Compare the methodology against the most widely used method when low yield is expected

2. To characterize the phylogenetic origin of a metagenomic read using alignment-free methodologies.

3. Devise and validate a method to isolate the microbial DNA from a skin biopsy, testing the microbial diversity bias and the functional variation of skin microbial community

4. To characterize the effect of unrelated genetic defects on the microbiota diversity

5. To assess the taxonomic and functional variation of the microbiota based on a systems perspective of health.

Results

Chapter 1. Direct sequencing from the minimal number of DNA molecules needed to fill a 454 picotiterplate.

Authors: Maria Dzunkova*, Marc Garcia-Garcerà*, Llúcia Martinez-Priego, Giuseppe D'Auria, Francesc Calafell[1], Andrés Moya[1].

* These authors contributed equally to the development of this work

Manuscript in preparation

ABSTRACT

Background. whole-genome amplification (WGA) is a fundamental approach for sequencing scarce DNA samples. However, WGA presents a GC-content bias, develops chimeras, suffers contamination enrichment, which add to the further sequencing analysis undesired noise and which compromise functional and/or taxonomic characterizations.

Hypothesis. As an alternative, whole-genome amplification-free (AF) methods can be the solution for all these problems. We explored the limits of the FLX titanium sequencing platform and compared a WGA protocol to an adaptation to the AF method proposed by Zheng et al. with a very low DNA yield (equivalent to 10.000 *E. coli* cells).

Results. The AF methodology resulted in an almost complete coverage of the reference genome, compared to a sparse, biased read distribution in WGA, with a very high amount of unassigned and unspecific DNA amplifications.

Conclusions. AF methods are an interesting alternative to Whole Genome Amplification, resolving most of its problems.

Introduction

The sequencing of the reference human genome in 2001 can be considered the start point of the "Omics" era, increasing our understanding of biological entities at all levels, from molecular to ecological(Venter et al. 2001). This milestone ushered in a new research phase that produced high throughput DNA and RNA sequencing data and all the bioinformatical tools needed to manipulate, analyze and compare a wide variety of genetic sources. Since the first genomic approach using cloning-based applications combined with the classical sequencing technology, substantive changes have occurred. High-throughput sequencing technologies, commercially appeared on 2004(Margulies et al. 2005), claimed to reduce costs, avoid the cloning bias and increase the amount of data obtained. The technological race between the different high-throughput sequencing platforms appeared, and has continued with one main goal: Obtaining a technology inexpensive enough to allow any genetic sample to be sequenced, analyzed and understood(Wolinsky 2007).

Still, next-generation sequencing technologies present some notorious limitations, such as the inverse correlation between the fragment size and the overall yield(Metzker 2009), which is crucial in data classification(Wommack et al. 2008), assembly(Pop and Salzberg 2008), and contamination detection; their high error rate(Metzker 2009) and their positive bias towards high GC rich regions. All these problems cause noise and complications in further data processing(Erlich et al. 2008), and have been partially solved increasing the DNA sequencing efficiency, enhancing by approximately 100,000-fold the sequencing output(Treangen and Salzberg 2012; Van Geystelen et al. 2013).

But the most limiting factor of next-generation sequencing platforms is the requirement of high amounts of starting genetic material to prepare libraries for sequencing (e.g. 1µg of starting material for a rapid library preparation method in 454 FLX+ technology). This limitation prevents working with most of the sample origins, including biopsies, which are rare and difficult to collect due to ethical limitations, or novel uncultivable microorganisms. To overcome this limitation, a number of methods have been developed to enrich genomic DNA without bias, with varying levels of success(Hosono et al. 2003; Albert et al. 2007; Hodges et al. 2007; Porreca et al. 2007; Garber 2008). Each one of them has been selectively used in different fields, according to the necessity, but any of them was bias free, leaving whole genome amplification as the option of choice.

Whole genome Amplification using the multi displacement amplification (MDA) reaction(Binga et al. 2008) is the unbiased random amplification using isothermal polymerase φ29 from *Bacillus subtilis* phage φ29(Vlcek and Paces 1986). This reaction was originally designed to be used in circular DNA templates, resulting in 10,000-fold amplification after a few hours of amplification. MDA has been successfully used in a wide variety of fields, from tumor genomics to bacterial genomics or single cell microbial genomics(Dean et al. 2002b; Paez et al. 2004; Raghunathan et al. 2005), in studies of selected gene loci.

However, MDA has been shown to have two main problems when it has been applied to whole genome sequencing approaches; first, this reaction may produce genomic bias, which may lead to errors in repetitive regions, or lack of genome coverage(Dean et al. 2001).This bias has been suggested to be caused by different inter-primer distances in the genome(Lage et al. 2003). Given that MDA uses random hexamers, hexamer hybridization and the distance among them may result in regions that are underrepresented or even not amplified at all. And second, the unspecificity of the random hexamers, and the amplification temperature of 30°C, make this reaction prone to amplify template-free hexamer concatenations, contaminant sequences and chimeric formations(Lasken and Stockwell 2007) which may be the process introducing more noise when MDA is used in de novo genome sequencing.

Still, although large amounts of starting genetic material are needed for 454 library preparation, only a few picograms are used in the emulsion PCR (emPCR) and the following pyrosequencing procedure(Meyer et al.

2008b). EmPCR involves mixing single-stranded library templates with DNA-capturing sepharose beads in an oil emulsion, expecting single sequence capture. However, empirical results have shown a strong multi sequence enrichment when the ratio DNA/sepharose bead exceeds the 0.35 limit, which may result in mixed signals(Margulies et al. 2005). Then successful quantification of the number of molecules may result critical to reduce the amount of starting material needed to sequence a sample. As a possibility quantitative PCR has been suggested as the solution to avoid the titration step to calculate the template/bead ratio, reducing the amount of starting material and allowing rare or limiting samples to be sequenced without MDA(Meyer et al. 2008b). Zheng et al. described an alternative method to prepare 454 libraries from samples with 1ng of total weight (Zheng et al. 2010), and quantify them through MGB taqman probes against 454 adaptors. These probes are suitable to quantify low amounts down to a few zeptograms, even below to the minimum amount needed for FLX sequencing(Huang et al. 2011). To assess the limits of the 454 sequencing platform, we adapted the method proposed by Zheng et al. to start with very limited samples. We constructed and sequenced a library out of 10,000 cultured Escherichia coli cells on a 1/8th Picotitter plate using FLX titanium+ technology. To test whether this method performed better than the MDA, we have compared our library without amplification, to a library constructed from the same amount of DNA but amplified using MDA.

Materials and Methods

Escherichia coli cell preparation

Methods are schematically presented in Figure 1. It is important to note that all the steps in the protocol were specially designed to prevent DNA loss, making each step where DNA yield can be reduced as effective as possible.

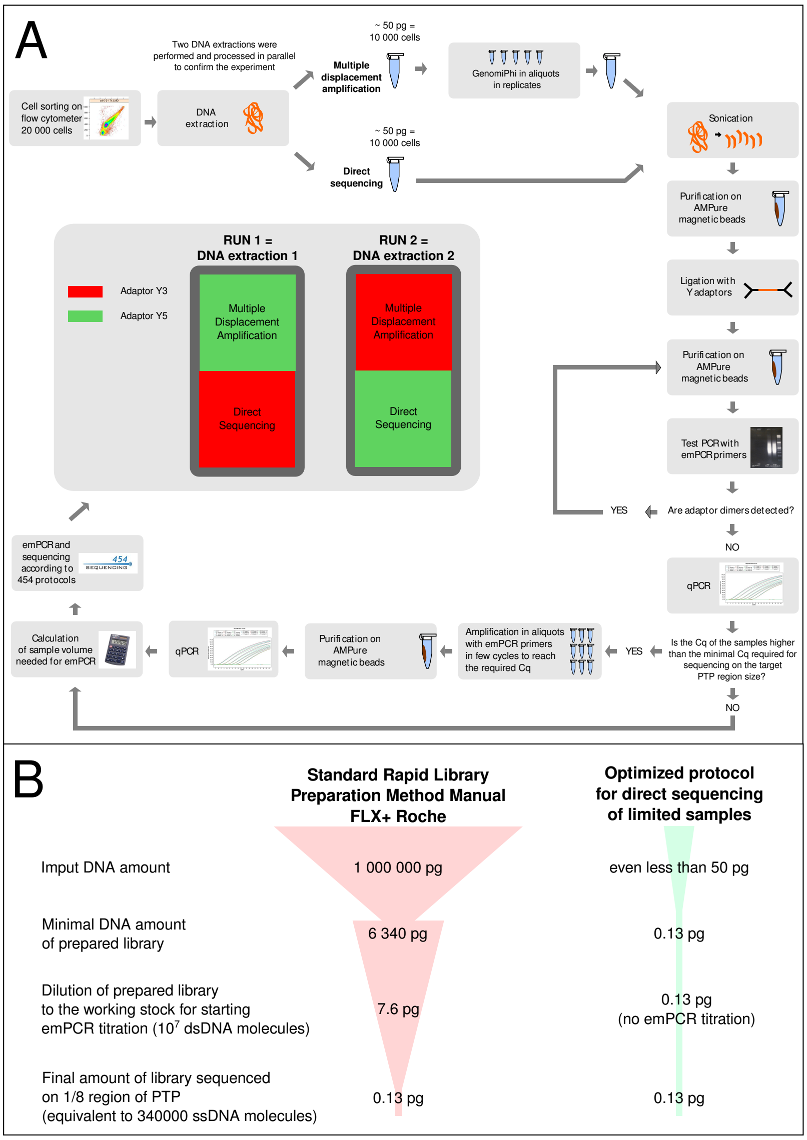

A
Cell sorting on flow cytometer 20 000 cells
DNA extraction
Two DNA extractions were performed and processed in parallel to confirm the experiment
~ 50 pg = 10 000 cells
Multiple displacement amplification
GenomiPhi in aliquots in replicates
~ 50 pg = 10 000 cells
Direct sequencing
Sonication
Purification on AMPure magnetic beads
Ligation with Y adaptors
Purification on AMPure magnetic beads
Test PCR with emPCR primers
Are adaptor dimers detected?
YES
NO
qPCR
Is the Cq of the samples higher than the minimal Cq required for sequencing on the target PTP region size?
YES
NO
Amplification in aliquots with emPCR primers in few cycles to reach the required Cq
Purification on AMPure magnetic beads
qPCR
Calculation of sample volume needed for emPCR
emPCR and sequencing according to 454 protocols
RUN 1 = DNA extraction 1
RUN 2 = DNA extraction 2
🟥 Adaptor Y3
🟩 Adaptor Y5
Multiple Displacement Amplification
Direct Sequencing
Multiple Displacement Amplification
Direct Sequencing
B
Standard Rapid Library Preparation Method Manual FLX+ Roche
Optimized protocol for direct sequencing of limited samples
Imput DNA amount
1 000 000 pg
even less than 50 pg
Minimal DNA amount of prepared library
6 340 pg
0.13 pg
Dilution of prepared library to the working stock for starting emPCR titration (10^7 dsDNA molecules)
7.6 pg
0.13 pg (no emPCR titration)
Final amount of library sequenced on 1/8 region of PTP (equivalent to 340000 ssDNA molecules)
0.13 pg
0.13 pg

Figure 1. Flowchart of the minimal library preparation protocol. Panel A: Schematic representation of the experimental work performed. The DNA from 20,000 cells was extracted and split in two aliquots, one further processed by the optimized library preparation protocol proposed (DS), and the other previously amplified by MDA (MDA) before preparing the library. Both library qualities were checked by test PCR with emPCR primers. Self-ligated adaptors were removed. The purified samples were quantified by qPCR. In case of obtaining a yield below the minimum, a short PCR against the adaptors was performed to increase the numbers.

Panel B: The comparison of standard Rapid Library Preparation Method Manual GS FLX+ Series.

Escherichia coli strain K12 was cultured overnight (O/N) in liquid Lysogeny Broth medium at 37°C. The culture was pelleted by centrifugation at 2000 rpm for 4 min at 4°C, and washed twice in cool physiological solution (NaCl 0.9%). Cells were immediately fixed adding formaldehyde 3.7%, and incubated O/N at 4°C. Fixed cells were washed twice to remove the remainings of formaldehyde and resuspended in physiological solution to reach optical density OD600 = 0.1.

Flow Cytometry separation of 10,000 cells.

We prepared 10,000 E. coli cells since we estimated that this number would provide enough staring genomic material to prepare a 1/8th plate for 454 platform. The expected DNA yield was 50pg, which was enough to sequence a genome with an expected 10X coverage.The diluted sample was stained with SYTO62 DNA staining (Invitrogen. Paisley PA4 9RF, UK) according to manufacturer's instructions, to distinguish the bacteria from cytometry noise that appears at the same size range. Flow cytometry sorting was performed using a MoFlo™ XDP cell sorter (Beckman-Coulter. Pasadena CA). Wavelength emission was set at 635nm, and absorption at 670, to detect signal from E.coli DNA stain. Gates were set using the side-scatter vs fluorescent signal to separate the cells. Sorted cells were placed in 1.5mL tubes containing physiologic solution to reach the number of 20,000 cells.

DNA extraction

We extracted DNA from 20,000 cells and then split the output DNA in two tubes in order to minimize any difference caused by random effects in the DNA extraction method. DNA was extracted according to the protocol of Ausubel et al. (Ausubel et al. 1992) in sterile conditions. All the chemicals used were previously sterilized by autoclave plus filtration through 0.2 µm-pore sterile filters. Briefly, sorted cells were lysed in lysozyme 10% (Applichem. Omaha. NE) in PBS for 30min at 37°C. SDS 0.5% (Applichem) and Proteinase K 0.13mg/mL (Applichem) were added to the cell lysis, which was incubated for 1 hour at 50°C. After the digestion, NaCl 0.6M and CTAB 1%(Applichem) were added. Solution was incubated at 65°C for 15 min. DNA was extracted with phenol:chloroform:isoamyl alcohol

(25:24:1) solution. DNA precipitation was performed with NH4-acetate (5M) and isopropanol. According to the low yield expected, 1µL of glycogen (20mg/mL) was added as a DNA carrier to help precipitation, and visualize the pellet. DNA was resuspended in 20µL nuclease-free water and divided in two aliquots to perform both protocols separately.

φ29 MDA-based whole genome amplification

MDA was performed with the commercial version of GenomiPhi V2 amplification kit(GE Healthcare Waukesha, WI). The reaction was performed according to manufacturer's instructions with incubation at 30ºC for 2.5 hours. In order to reduce amplification bias, we performed 5 replicates of the amplification using 2µL of extracted DNA per tube, and finally pooled after the reaction finished.

Shotgun 454 library preparation

After MDA amplification, both samples followed the same protocol. DNA shearing was set up previously with test DNA in a Raypa UCI-50 sonicator to obtain the correct fragment range. Sonicator water was set and maintained at 2ºC to avoid any side effect on the DNA. We chose this sonicator because it was prepared to work in closed tubes, avoiding possible side contamination. Query DNA was then sheared for 3 minutes at maximum intensity, obtaining a fragment distribution of 200-1000bp.

DNA fragments shorter than 400 bp were removed by Agencourt AMPure Beads XP (Beckman-Coulter), using the protocol proposed by Roche. According to the bead calibration, we added 1.2 volumes of AMPure beads to our sample (v/v). Fragment purification was performed in a magnetic particle concentrator, according to AMPure protocol.

454 adaptor ligation was performed according to the protocol proposed by Zheng et al.(Zheng et al. 2011). 12µL of DNA were added to a blunt-end mixture containing 1µL of dNTPs 25µM each (Fermentas. Thermo-Fischer. Waltham, MA), 2.5µL of ligation buffer (NEB. Ipswich, MA), 2.5µL of ATP (Agilent. Santa Clara, CA), 2µL of quick blunting enzyme mix (NEB) and 0.5µL of Klenow Fragment 3' 5' exo- (NEB). The mixture was incubated in a thermocycler (Eppendorf. Hamburg, Germany) for 15 min. at 12ºC

followed by 15 min. at 72ºC. After this incubation, the solution was ice-cooled.

Two Y adaptors were prepared with different multiplex identifiers (MIDs), according to Zheng et al. (Sigma-Aldrich. St. Louis, MO). 1µL of Y adaptors (100µM initial concentration) were added to the enzymatic mixture with the repaired DNA. 1µL of T4 DNA ligase (NEB) was also added to the mixture. The whole solution was incubated at 12ºC O/N. In order to test for possible adaptor bias, we performed this protocol twice for each DNA protocol with different MIDs. Self-ligated adaptors, which may produce short fragment sequencing bias in the following steps, were removed the day after by AMPure bead purification.

Library Quality Control

In order to prove adaptor ligation to the DNA and discard adaptor dimer presence after library purification, we prepared a test PCR using emPCR primers from Roche (emPCR-F 5'-CCAT-CTCATCCCTGCGTGTC-3', empCR-R: 5'-CCTATCCCCTGTGTGCCTTG-3', synthesized by Isogen Life-Science. De Meern, The Netherlands). 1µL of sample was tested for amplification using Go Taq Green polymerase Mix 2x (Promega. Fitchburg, WI) and 1µL of each emPCR primer, using the following conditions: an initial denaturation at 94ºC for 2 min was followed by 25 cycles of 94ºC for 2 sec, 60ºC for 60 sec, and 72ºC for 60 sec, and a final extension at 72ºC for 8 min. PCR product was visualized in a 1% agarose gel performed under standard conditions. If self-ligated adaptor was present, which could be observed as a band around the 100bp region, we repeated the purification until the band could not be longer detected (Supplementary Figure 1).

Quantitative PCR

Quantitative PCR was performed on a Roche LightCycler LC480 II, with MGB probe, according to Zheng et. al(Zheng et al. 2011). Each test was performed three times. For each reaction, the mix was prepared as follows: 10µL of Kapa Probe Fast Universal 2x qPCR master mix (Kapa Biosystems. Woburn, MA) were mixed with 1.4µL of each emPCR primers, with 1.2µL of MGB-probe 10µM and 1µL of sample. The reaction was finally adjusted to 20µL with nuclease-free water. To calculate the exact number of molecules, a standard curve was prepared using an amplification product of known length (202bp) and known concentration which contained the same adaptors used for 454 sequencing. Serial dilutions 1:10 of the standard were prepared and amplified with the same qPCR protocol. Quantitative PCR was performed using the following cycling protocol: a first denaturation step at 94°C for 10 minutes was followed by 40 cycles of 95°C for 30 sec, 60°C for 15 sec and 68°C for 1 min, allowing the longer reads to be extended.

Given that quantification was performed by MGB probe, we can calculate the exact number of molecules, independently of the fragment length. A standard curve was used to calculate the Cq vs log(Number of molecules) linear equation. We then used that equation to calculate the exact number of molecules per microliter in our samples.

Emulsion PCR and sequencing

MDA amplified DNA plus Whole Genome Amplification-free (AF) DNA were prepared using different MIDs, allowing them to be combined in a single PTP sequencing plate. Both samples were pooled equally according to the minimum number of molecules needed for one plate. The emulsion PCR was prepared using a Small Volume emPCR kit (Roche Applied-Science. Penzberg. Germany). According to this, the initial number of beads in this kit is 2.4 x106 beads per vial. To avoid mixed beads we calculated the sample volume to obtain a 15% enrichment, meaning the number of molecules to enrich 340,000 beads, according to Roche protocol. Then, a pool of 170,000 molecules from each sample was prepared and mixed with the beads. emPCR was performed according to manufacturer's instructions. Enriched beads were purified and used in 1/8th of PTP plate, and sequenced using a GS FLX Titanium Sequencing XLR70 Kit (Roche Applied Science).

Sequence processing

Sff files were processed according to the following pipeline. Fasta+Quality files were obtained and sequences were separated by MID, allowing up to three mismatches, using a customized perl script. Sequences were then checked for the presence of Y adaptors in the 3' end using the Blast algorithm implemented in Blast2GO(Altschul et al. 1990; Conesa et al. 2005) with an e-value below 10-3 . In case they were present, adaptor sequences were trimmed with the R package Biostrings v.2.11(Pages et al. 2012). Low complexity reads were removed from the analysis using a combinational script that included functions from the R packages ShortRead(Morgan et al. 2012),Biostrings and Entropy(Hausser and Strimmer 2012).

Genome Mapping and Data analysis

To test whether the coverage obtained with both sequencing approaches was similar, post processed reads were mapped to the genome of Escherichia coli K12 (gi:49175990) using SSAHA 2.5.4(Ning et al. 2001), using the following parameters; Word size = 13, minimum length for cross_match matching = 10, word size for cross_match matching = 10, number of k-mer matches required = 1.

Coverage was visualized and manipulated with the R packages Rsamtools(Li et al. 2009), ShortRead(Morgan et al. 2009) and Chipseq.

Coverage distribution differences between both methodologies were checked using Cramer von Mises test. We also tested for normality of the coverage. Given that coverage was not normally distributed, we tested differences between coverage distributions between both methods using the Kruskal-Wallis test. Given that the number of reads that matched to E. coli K12 was very different in both methods, we subsampled AF, 100 times and tested for differences between AF subsamples and GP.

Reads that did not match to E.coli were aligned using NCBI-blast against the "nr" database using the Megablast algorithm. Reads with e-value < 10-128 were considered to originate from theoretically contaminating bacterial species (TCB). TCB genomes were retrieved from the NCBI repository and used in SSAHA2 to calculate the percentage of coverage on those genomes.

We suggest that reads without a hit against NCBI, we proposed that they could originate from a concatenation of hexamers, based on the protocol from GenomiPhi, as GenomiPhi uses random hexamers to be used as primers for MDA. We divided the 'no-hit' reads in hexamers, and calculated the hexamer distribution among the reads, counting only the complete hexamers. We tested the hexamer distribution of our sequences against all TCB genomes, plus two random hexamer distributions, plus two Escherichia coli strain hexamer distributions, plus different strains of *Bacillus subtilis* strains (we called the union of these sets TCB+), since the enzyme used in the multi-displacement amplification was obtained from a *Bacillus subtilis* phage, which could be a source of contamination.

Hexamer distribution differences were tested using Cramer von Mises test, which judges the bona fide of the cumulative distribution function F*, compared to an empirical distribution. In this case, we tested our sequencing hexamer distribution among the different TCB+ genomes, which were used as the expected possible distributions. If 'no-hit' sequences were experimentally constructed by random hexamer concatenation, their distribution would be random and then we would expect a random cumulative distribution similar to a normal distribution. However, if our 'no-hit' reads would come from an unknown genome, they would fit in a gamma distribution which is common for all the known bacterial genomes so far. Cramer von Mises test was performed using the R package 'CvM2SL2Test' (Xiao and Xiao 2012).

To test whether relative abundances could pinpoint the origin of our unclassified reads, hexamer signatures were used to link unclassified reads to all TCB+ genomes. Canonical Correlation Analysis was performed with the hexamer relative abundances using the 'cca' function implemented in the R package 'vegan'(Oksanen et al. 2011). The association between the hexamer distribution of our sequencing projects and the rest of the TCB+ genome hexamer distributions was tested by grouping the distributions according to their origin in: E. coli genomes, B. subtilis genomes, GenomiPhi (GP), Amplification-Free (AF) and TCB genomes. ANOVA was performed using the correlation eigenvalues and the different theoretical clusters to assess similarities between the different groups.

Hierarchical clustering was also performed among the different groups. The robustness of each hierarchical grouping was measured with a bootstrap analysis with 1000 generations. Hierarchical clustering was performed using the R packages ‘hclust’ and ‘pvclust’(Suzuki and Shimodaira 2006).

Results

Library quantification and Sequencing results

DNA was extracted and processed according to its respective protocol. Libraries AF-Y3 and AF-Y5 were prepared and quantified, resulting in 414,443 and 41,043 ssDNA molecules respectively. Given that AF-Y5 had not enough molecules to be sequenced, we performed a 4 cycle emPCR primer amplification to further enrich the sample. This enrichment was performed separately in 20 tubes with 2μL of starting material to avoid possible PCR bias. The sample was re-quantified to test whether DNA amount was sufficient, obtaining 384,561 molecules. AF samples were pooled with their diluted GP counterparts. However, one sequencing process was unsuccessful, resulting in only 8,648 reads. This result could be caused by a poor library or a faulty emPCR enrichment. The second run resulted in an output of 66,512 sequences of average quality score 32, which was sufficient for the analysis.

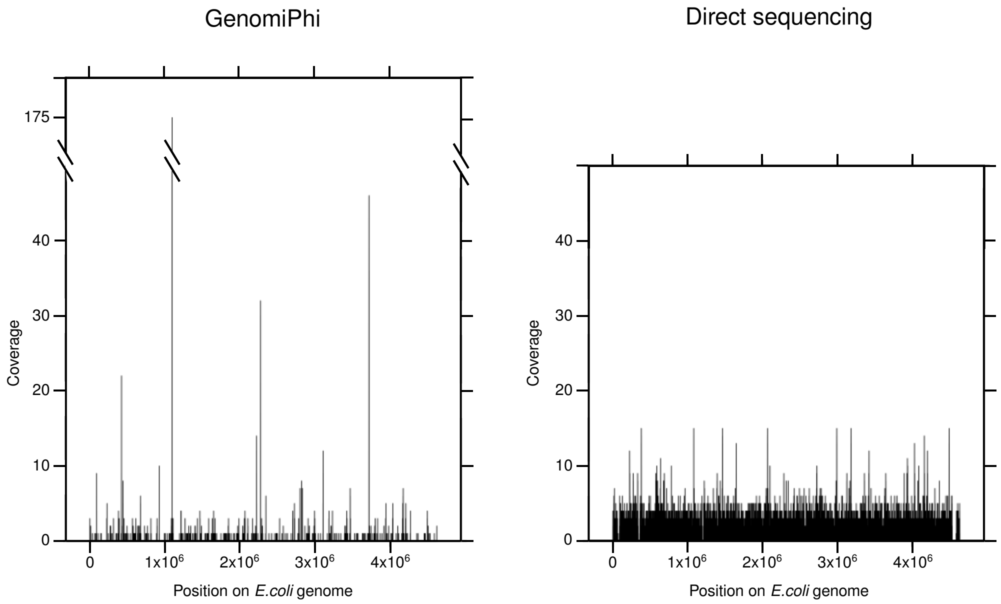


Figure 2: Distribution of coverage along E. coli genome.
Comparison of the genome coverage obtained by MDA and DS. The genome coverage of MDA reads was characterized by unequal distribution with many gaps and several areas with extremely high coverage (up to 175x), while the highest coverage obtained by DS was only 15x and it was better distributed along the whole genome.

Sequencing output was divided by adaptor, resulting in 24,654 and 50,506 reads with an average read length of 300 ± 151 bp. Expectedly, we obtained enough sequences to cover the whole E.coli genome in both methodologies. Sequences were filtered by complexity and trimmed to remove 3' adaptors and low quality ends, leaving 20,927 and 48,140 reads for further analyses. GC content was analyzed in both methodologies. We observed a significant decrease of the GC content in the GenomiPhi method compared to both the unamplified method (2 tailed t-test p-value = 0.0021) and the actual E.coli genome GC content (2 tailed t-test p-value = 0.037).

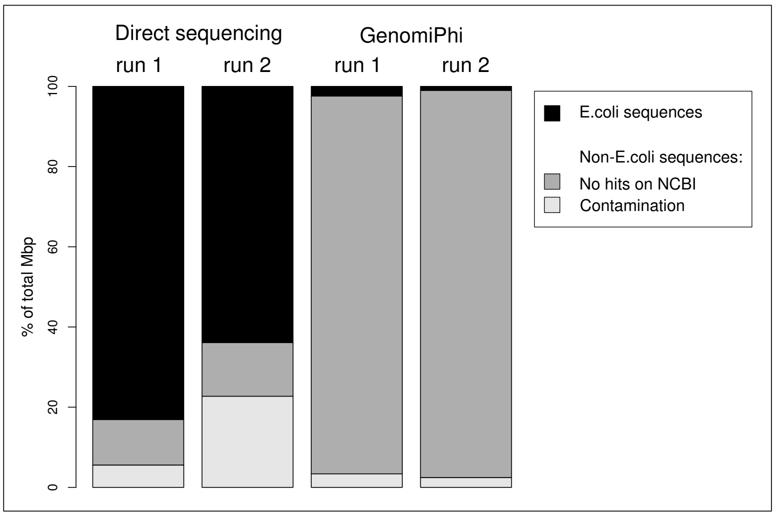


Figure 3. Results of Escherichia coli genome mapping and blast to NCBI database. Proportions (in %) of Mbp mapped by SSAHA2 to E. coli genome are showed for MDA and DS sequences

Escherichia coli genome mapping

Filtered and processed reads were mapped against the reference genome of E.coli. The unamplified method resulted in 80.59% of reads that matched with the genome (Figure 2). The average genome coverage was 5.8X. Coverage followed a normal distribution. In contrast, GenomiPhi method only presented 2.1% of reads matching the reference genome. Of them, coverage was poorly spread along the genome, with more than 50% of the genome uncovered, registering peaks of overrepresented regions up to a 175X coverage, and an average coverage of less than 3%. Distribution differences were tested using a one-way Kruskal-Wallis test, resulting in significative differences on the coverage distribution (p-value =

0.0017). Coverage was confirmed with MUMmer 3.0(Kurtz et al. 2004), supporting that the amplification-free method is more effective for genome sequencing projects in low yield DNA situations.

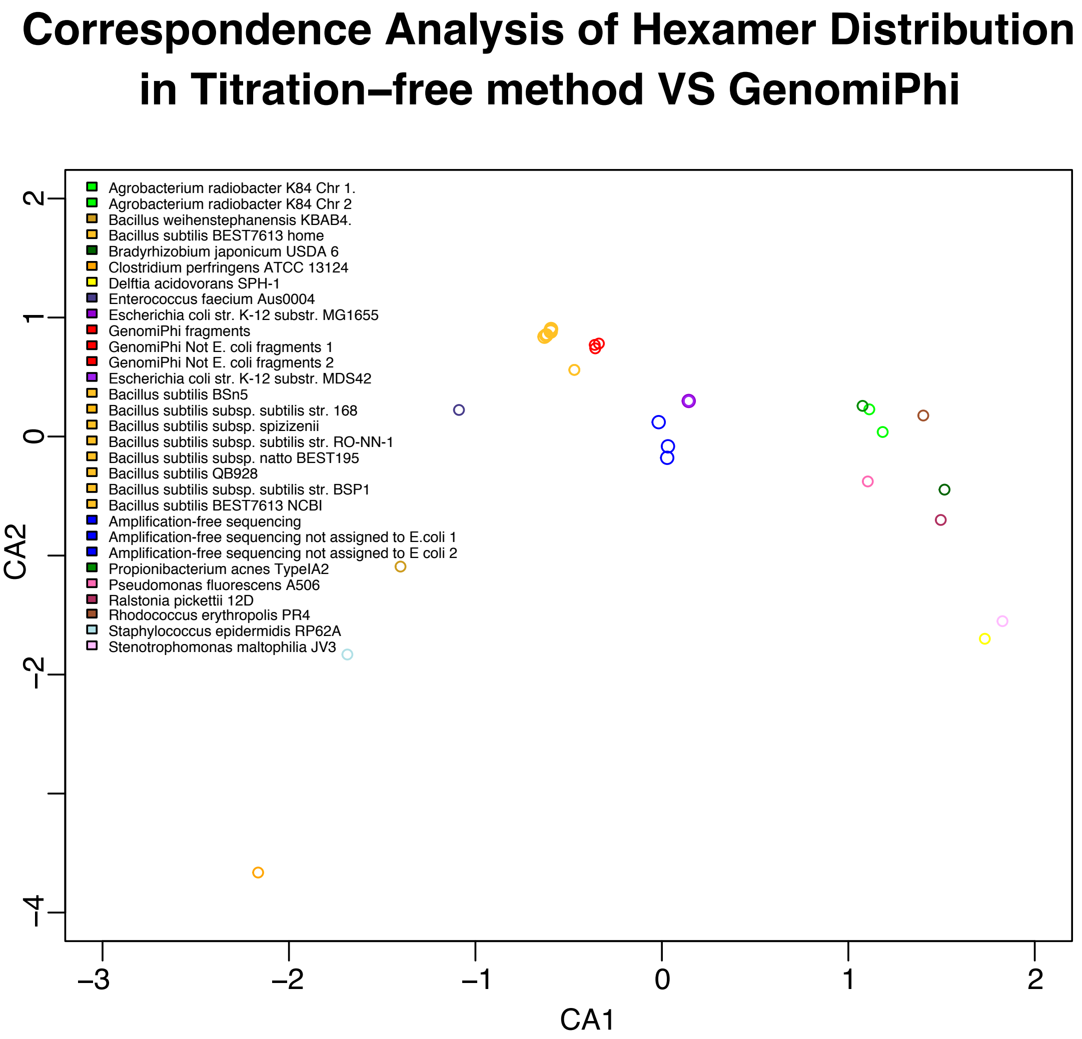


Figure 4. Correspondence analysis of the relative abundances of hexamers
6 nt-size relative abundances were considered as individual variables. hexamer space was used to construct the orthogonal vectors and calculate the distances between the different samples. three clear clusters can be observed. The one from GenomiPhi+*B.subtilis,* the *E.coli* + direct sequencing, and the *Agrobacterium +Propionibacterium+Psedomonas*

To test whether the reads not assigned to E.coli K12 by mapping originated indeed from E. coli, BLAST was performed against the 'nr' database. Unassigned GenomiPhi sequences only resulted in 2.16% of reads that matched some E. coli genomes with an e-value of 10-10. This result did not vary even rising the e-value cutoff to 10-4.

Analysis of unassigned reads

Once all the reads that matched to E.coli were assigned, we tried to assign the remaining reads to other genomes (Figure 3).1460 reads (6.98%) of the direct sequencing method and 1423 reads (2.96%) of the GenomiPhi method could be assigned to other species. Approximately one third of these sequences were identified as human: the rest came from bacteria. All these bacterial species were considered as TCB, and included mainly Proteobacteria. We selected those with a very low e-value (~ 0.0).

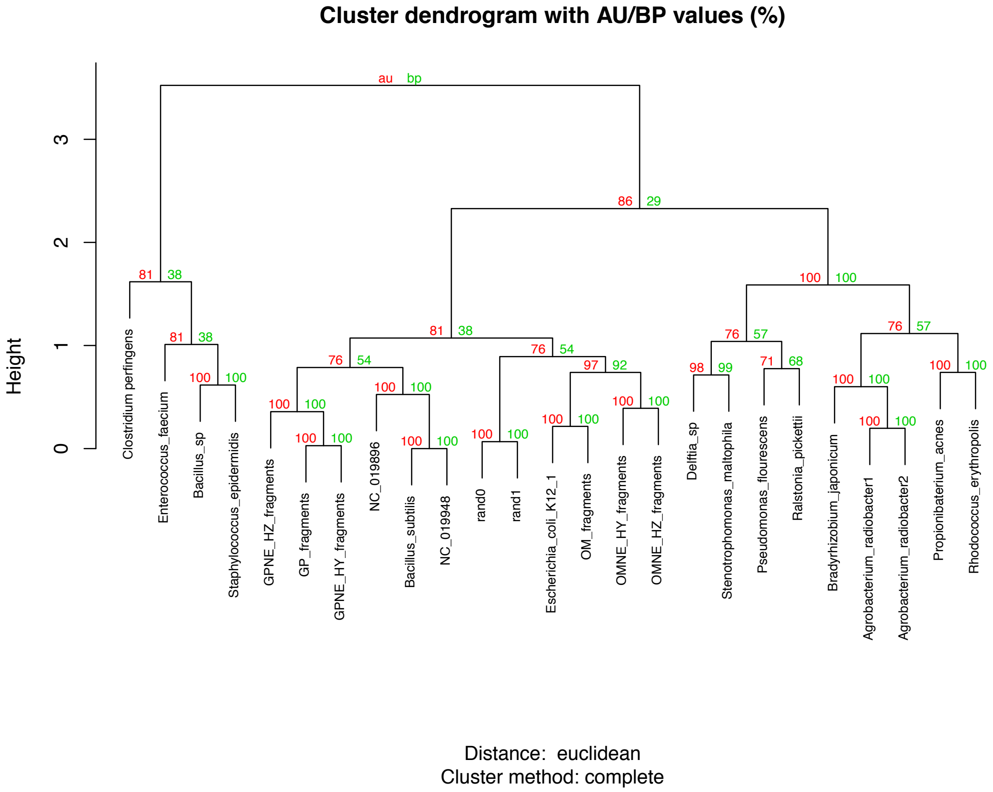


Figure 5. Hierarchical clustering of the hexamer distribution.
Comparison of the relative abundances of hexamers in the TCB+ genomes, and location of the different methods analyzed in this work. Bootstrap support (BP in green) and uncertainty decay (AU in red) show the statistical support on each node. Only BP and AU over 90 are considered statistically significant

TCB genomes were retrieved from the NCBI repository and were used in SSAHA2 to assess where the reads matched to. Of all reads assigned and mapped to TCB, 63.88% were in regions shared with E.coli and did match previously with that genome.

However, most of the reads obtained by the GP method (94.2%) were not assigned to any organism present in the 'nr' database. In the direct sequencing approach we also found a fraction (12.84%) that could not be assigned to any 'nr' represented species. This inability to assign all these reads is not related the length of the read or any other obvious feature. In fact, unclassified and classified reads followed the same read length distribution, as can be observed in Supplementary Figure 1.

In order to assess the origin of the unclassified reads, we tested the differences between AF and GP datasets. Unclassified reads in the GP dataset could have been artificially synthesized through hexamer concatenation. Hexamer frequency was obtained by screening each read using a 6-bp window size. As a null distribution, we constructed an artificial genome based on an average purine-pyrimidine ratio of 0.5, and calculated the hexamer distribution for that genome. We calculated, as alternative hypotheses, the hexamer distribution for all TCB+ genomes. Contrary to the expectation, reads coming from GP method did not display a random distribution and were totally different from the normal distribution of the artificial genome (Cramer von Mises test. p-value = 2.99x10-11). However, we did not observe differences between GP unknown reads and any of the rest of the hexamer distributions that came from the TCB genomes. We observed that any bacterial genome displayed a gamma distribution of hexamers, and GP-unknown were equivalent to them (Cramer von Mises test. P-values ranging from 0.16 to 0.51, depending on the genome). According to these results, we tested the hexamer relative abundance profile between our methodology outputs and of the different genomes' distributions, to try to assign the origin of those unclassified sequences. We also included in this analysis the species previously suggested to be common contaminants of sequencing projects using MDA(Woyke et al. 2011). Expectedly, reads coming from the amplification-free approach clustered with Escherichia coli K12 genomes (Figure 4). However, AF reads not assigned to Escherichia coli by any of the methods previously described, were also adjacent to the Escherichia coli distribution in CA space. Hierarchical clustering (Figure 5) with bootstrap reconciliation confirmed this result (Bootstrap support =

92%). The correspondence analysis showed a group that included *Bacillus subtilis* genomes and GP-unknown reads. This group was also confirmed by hierarchical clustering (Bootstrap support = 100%). To reduce disturbances produced by large variance spectra in hexamer abundances created by including very different genomes in the analysis, we eliminated all but *Bacillus subtilis* and Escherichia coli K12 genomes, obtaining similar results. However, AF-unassigned reads separated from the E. coli cluster in hierarchical clustering analysis, although the CA plot showed more proximity to E.coli genome distributions than to any other. All these results suggest that unranked GP reads might come from *Bacillus subtilis* DNA remainings present in the MDA enzyme buffer.

Discussion

In this work we demonstrate the possibilities and benefits of using a whole genome amplification-free method on samples with trace amounts of input DNA, such that proposed by Zheng and collaborators(Zheng et al. 2011). This method eliminates all the side effects produced by the deep amplification methods, commonly applied in samples with a very low yield. Zheng et al. worked with samples diluted to reduce the concentration to 1ng of total yield. In our case, we optimized all the steps on the library preparation protocol to work with very limited samples, limiting the DNA loss by skipping the steps where most of the DNA was lost, like during the fragmentation, and switching them to alternative methodologies that keep better the DNA(Rodrigue et al. 2010).

Current high throughput sequencing platforms require a minimum amount of DNA of 10-6 g. to prepare a successful library(Hutchison and Venter 2006). But several authors have suggested alternative protocols to construct a DNA library from a reduced amount of DNA and proved them by a successful sequencing(Meyer et al. 2008b; White et al. 2009; Zheng et al. 2010). They demonstrated that the true limiting factor in 454 sequencing is to obtain the number of enriched beads needed by the platform (e.g. 340,000 enriched beads for 1/8th PTP plate). If we round the fragment size to 700bp, the actual DNA amount needed would be less than 1% of the starting DNA required by Roche. This input requirement results in an insuperable bump for limited or rare samples, and implies to throw away a sample that, alternatively, could be used for more specific experiments afterwards(Belda-Ferre et al. 2011; Pérez-Cobas et al. 2012). To overcome this limitation, the most widely used method has been isothermal multi-displacement amplification. However, MDA presents a set of important problems previously reported by multiple authors. In this work we have compared our optimized whole-genome amplification free protocol to a MDA-based library construction and sequencing of a few Escherichia coli cells, showing that all of those problems may be avoided using the WGA-free method.

The main benefit of our method was the high and homogeneous coverage output assigned to E.coli. Other authors' results reported poor efficiency results assigning the sequencing output to the specific microorganism to be sequenced(Rodrigue et al. 2009; Woyke et al. 2011). Our results in MDA control method agreed with these observations. In contrast, our alternative behave way better in terms of absolute and relative numbers. An almost complete coverage of the genome, with only two major gaps can be considered a success when starting from as low as 385,000 initial sequences, even more when the coverage was very homogeneous. Amplification bias was also an important issue in MDA output: We find a very wide range of coverture, with regions from 0, in more than 60% of the genome to a maximum of 175X coverage. Supporting this problem, missing genomic proportions of more than 40% have been previously reported(Marcy et al. 2007). All these problems result in a very low sequencing efficiency, increasing the effort needed to finish a target genome sequence. Those low coverage rates and heterogeneous coverage distribution result in a frequent failure to close the target genome. Still they may lead to important information about novel gene discovery(Marcy et al. 2007; Podar et al. 2007), or complex gene regulatory systems(Dupont et al. 2011). But to achieve the complete genome information alternative and complementary methods are needed, more when the starting genetic material is very low. Our results show that the combination of MGB-probe based qPCR with 454 sequencing results in a strong improvement. The introduction of a qPCR step to calculate the volume needed to enrich the exact number of beads to be sequenced allows to reduce emPCR steps, imply a reduction on the initial amount of material needed and the cost. As Zheng et al. showed, this approach may result in DNA saving for further experiments, but also it will make the NGS methods available to very limited samples. Also, avoiding MDA will prevent the coverage bias, the lack of efficiency and all the other problems previously described, which are especially problematic when they are coming from reagent contaminants, increasing the efficiency, and the cost-benefit ratio, and making it possible to sequence samples of ever decreasing amounts of starting DNA.

As mentioned before we have shown that only a minimum DNA amount is necessary to construct a library. In our case, we obtained around 385,000 sequences after library preparation. If, in theory, all of them would be assigned to Escherichia coli, they would cover the whole genome by 10 fold. Our results showed an average coverage of more than 5-fold on almost 81% of the reference genome(Rasko et al. 2008). The minimum limit defined by the 454 specifications requires to use a very sensitive quantitative method. To date, the most sensitive methods for DNA quantification are Minor Groove Binding (MGB) and Locked Nucleic Acid (LNA) probes(Buh Gasparic et al. 2010).To take advantage of its significantly higher sensitivity, specificity and reproducibility with a shorter length, Zheng et al. chose and designed a Y-adaptor specific MGB probe(Kutyavin et al. 2000). Our results support that the quantification method proposed by Zheng, is sensitive enough to quantify below the femtogram level, hundred times below the minimum amount required to prepare a library for 1/8th PTP plate(Roche Diagnostics GmbH 2008). Roche standard method uses a Y-adaptor with a fluorophore, which is used to quantify the number of correctly ligated molecules. However, its sensitivity is much lower than MGB-probes, and then they should be skipped(Roche 2011). Roche standard adaptors interfere with MGB-probes by using similar absorption and emission wavelengths. To avoid so, two alternatives should work: The first one is to use a MGB label which does not interfere with Roche Y adaptors. The second, which is the one we have chosen, is to synthesize Y adaptors without fluorochrome(Zheng et al. 2011).

Other problems may be solved by WGA-free methods, and nonspecific sequence formation is one of those. MDA presents a widely reported problem of nonspecific product formation. This chimeric and side product is mainly created through displacement of the DNA strand, becoming available to prime on a second template and form chimeras(Lasken and Stockwell 2007). When the initial DNA amount is high enough (on the order of nanograms) this background amplification merely reduces the product yield, but when the initial DNA source has a really low yield, as it is the case in single-cell genomics, this process is specially harmful. A typical bacterial chromosome contains a few femtograms (10-15g) of DNA, which, logically, once sheared it would result in a very low number of molecules, and probably, not enough DNA to prepare a standard library. If MDA is used the low yield of source DNA may allow the missanealing of different strains during multi displacement amplification. This is specially harmful when there are different DNA sources. The commercially available MDA reagents have frequently been reported of being contaminated by unwanted DNA, and one of the most probable sources is the host species where the enzyme comes from, in this case *Bacillus subtillis* phage(Woyke et al. 2011). φ29 polymerase isolation requires source DNA elimination, which include DNA hydrolysis and filtration. However, short reads may be kept in the solution, affecting the MDA process through chimeric sequence formation. Several research groups reported unspecific product contamination using commercially available MDA kits(Bredel et al. 2005; Jiang et al. 2005; Le Caignec et al. 2006; Spits et al. 2006; Iwamoto et al. 2007). All of them concluded that contamination did not come from human DNA, but could come from the enzyme preparation process, including host bacteria *Bacillus subtillis*(Vlcek and Paces 1986). Our results support these conclusions by using the specific k-mer distribution to discriminate through DNA origin. This method was previously used to compare metagenomic sources, allowing to discriminate samples with potential contamination(Willner et al. 2009b). Theoretically, samples from the same genomic origin should present equivalent or similar hexamer content, depending on the level of exogenous DNA content. In this case, we have observed that all MDA unranked reads were adjacent to the *Bacillus subtilis* genome. Although DNA cleavage and hydrolysis would eliminate most of the *B. subtilis* DNA, short reads could be used during the MDA process to create chimeras through strand displacement, as mentioned above. Of course, this contamination source is eliminated avoiding the MDA.

As we mentioned before, any alternative source of DNA is really harmful to any genomic study but is more harmful in the MDA based ones. We know that the cell sorting process, even when thorough decontaminating procedures are used, can result in a new source of contamination. Despite the fact that a low volume surrounds a single cell sorted which reduces the contamination of extracellular DNA(Stepanauskas and Sieracki 2007), it is very difficult to completely avoid contamination during cell sorting(Zhang et al. 2006). This is specially true when MDA is used afterwards, because even short DNA remainings of previous equipment usages may be a DNA source for chimera construction. Sodium hypochlorite wash is the most popular method for cell sorter cleaning(Schmid et al. 1997). However, sodium hypochlorite wash produces basically depurination, which introduces weak points in the structure of the DNA, making strain break-up easier. The length of the fragments produced is inversely correlated with the depth of the wash (Garcia-Garcerà et al. 2011). But still, DNA remains may still be found on further samples, which may be also amplified by MDA. It is important to note that same TCB species were found in both GP and AF methodologies in low ratios, suggesting possible contamination through cytometry DNA traces which, even in the best case scenario, cannot be avoided. However, using a direct sequencing method contamination resulted in a 3% of the total sequence output, while in MDA method we do not know whether they were contributing to increase the number of unassigned reads.

Previous reports have shown that chimeric sequences may complicate the whole genome assembly process, but they do not seem to affect in our case(Stepanauskas and Sieracki 2007; Swan et al. 2011). Even when a scaffold was used, we did not find E.coli rearrangements. to be sure, we split the GP unassigned reads in substrings of 25 nt, and repeated the mapping with SSAHA2 and Blast (Data not shown). Still, our results were not modified, discarding possible chimeric E.coli reads which were not able to be assigned to the reference genome. All the E.coli reads found in GP (2.16%) dataset were successfully mapped to E.coli.

Concluding, Whole-Genome Amplification free methods are the only ones to obtain unbiased, reliable and replicable genetic information from any sample with a minimum amount of starting material. We suggest that these kind of methods should replace MDA in most Omic projects, including genome sequencing studies, given its sequencing efficiency and its lower cost-benefit ratio.

Chapter 2. A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin.

Authors: Marc Garcia-Garcerà, Koldo García-Etxebarria, Mireia Coscollà,Amparo Latorre, Francesc Calafell

Submitted mansucript

ABSTRACT

In the last decade, an extensive effort has been made to characterize the human microbiota, due to its clinical and economic interests. However, a metagenomic approach to the skin microbiota is impeded by the high proportion of host DNA that is recovered. In contrast with the burgeoning field of gut metagenomics, skin metagenomics has been hindered by the absence of an efficient method to avoid sequencing the host DNA. We present here a method for recovering microbial DNA from skin samples, based on a combination of molecular techniques. We have applied this method to mouse skin, and have validated it by standard, quantitative PCR and amplicon sequencing of 16S rRNA. The taxonomic diversity recovered was not altered by this new method, as proved by comparing the phylogenetic structure revealed by 16S rRNA sequencing in untreated vs. treated samples. As proof of concept, we also present the first two mouse skin metagenomes, which allowed discovering new taxa (not only prokaryotes but also viruses and eukaryotes) not reachable by 16S rRNA sequencing, as well as to characterize the skin microbiome functional landscape. Our method paves the way for the development of skin metagenomics, which will allow a much deeper knowledge of the skin microbiome and its relationship with the host, both in a healthy state and in relation to disease.

Introduction

Despite the great interest of skin as an ecosystem, the study of skin microbiome has been recurrently limited by the low host-commensal cell ratio and the high taxonomical divergence among skin sites(Grice and Segre 2011). The skin is the most external organ in the mammalian body. Its main role is to protect the internal tissues and interact with the external environment, collecting information, preventing loss of temperature and moisture, and defending the body against pathogenic agents (Bennett et al. 2008; Proksch et al. 2008). After a long co-existence between host skin cells and the microorganisms that attempt to colonize the body, a set of bacteria consolidated into a skin commensal/mutualistic microbiota. This microbiota is distributed in multiple niches, depending on the amount of nutrients and the physical properties that might result in the most suitable growth conditions for them(Moya et al. 2008). The effect of commensal-host interaction may lead to complex behaviors of the whole host system, which, beyond the skin, involves also the immune system(Nestle et al. 2009). Understanding how the whole system works may lead to crucial knowledge of skin physiology that may be highly relevant to public health and cosmetic pharmacology. Moreover, and under certain conditions (including the host genetic predisposition), the skin microbiota may lead to the disruption of the skin homeostasis, leading to complex skin diseases, such as atopic dermatitis, psoriasis or eczema. Then, the knowledge of these bacterial disease triggers is crucial to clinical dermatology (Dethlefsen et al. 2007).

Historically, the study of skin microbiota has been severely limited. Culture-based characterization has been shown to be restricted only to the species that grow rapidly under standard laboratory conditions, which are estimated at just ~1% of the species in the skin(Fredricks 2001). Although more complex culture media have been generated, the main solution has come from the culture-independent methods for studying the composition of the microbiota. The most successful relies in amplifying the phylogenetic informative 16S ribosomal RNA gene regions from the bacterial community (16S rRNA). This has been applied in most human body habitats, including the gut, oral cavity, or skin, among others(Gao et al. 2007; Grice et al. 2008; Costello et al. 2009). Specifically in skin, these pioneering studies have shown a high diversity in microbiota composition, which is highly dependent on the skin physical conditions(Grice et al. 2009), thus defining a large amount of possible niches. Variability in skin microbiota has been shown to be site-to-site more diverse than between the left- and right-hand sites of the same individual, or between the same sites in different individuals(Grice et al. 2009).

But second, and more interestingly, the most diverse regions have been shown to be at least as diverse as the gut microbiota. 16S rRNA amplification can identify the species present and their dynamics, but it provides limited or null information about bacterial gene composition, cell function and dynamics, or on microbe-microbe or microbe-host interactions. This limitation has been solved with shotgun high throughput sequencing, pioneered in soil and ocean metagenomic studies(Venter et al. 2004; Angly et al. 2006; López-Bueno et al. 2009), and applied to functional analyses of the gut bacterial communities(Gill et al. 2006). But, despite the success of the functional analysis of other human microbiomes(Rogers et al. 2005; Willner et al. 2009a; Belda-Ferre et al. 2011; Docktor et al. 2011; Gosalbes et al. 2011), this method cannot be applied to the skin microbiome, which occupies all upper layers on the skin, making deep sequencing inefficient and expensive due to the high amount of host DNA being sequenced. Recently a large dataset of human skin microbiome samples has been published under the Human Microbiome Project Consortium(Consortium 2012). However, those samples have been collected using non invasive superficial methodologies that do not access to the full microbiota, or specific subniches such as the eccrine sweat glands or the hair follicle, where a specific diversity may be found and where microbes may be more active(Grice et al. 2008; Grice and Segre 2011). Thus, although a wider perspective has been achieved by this methodology, a prokaryotic DNA enrichment is a crucial step for deep skin metagenomic analysis. For that reason, we propose a new approach that allows the prokaryotic DNA isolation making subsequent shotgun high throughput sequencing feasible and efficient. This protocol is based on the combination of enzymatic digestion of all bounds between skin cells, and the subsequent separation of cells by size without disrupting their integrity, avoiding the contamination by host DNA. As a proof of concept we present the metagenomic analysis of two murine deep-skin microbiomes. Given the long-time use of mouse as a model, and the evolutionary relationship between mice and other mammals, we think that murine validation of this method may extend to human skin samples, and allowing the retrieval of wider information than a less invasive sampling process.

Materials and Methods

Mouse skin sampling

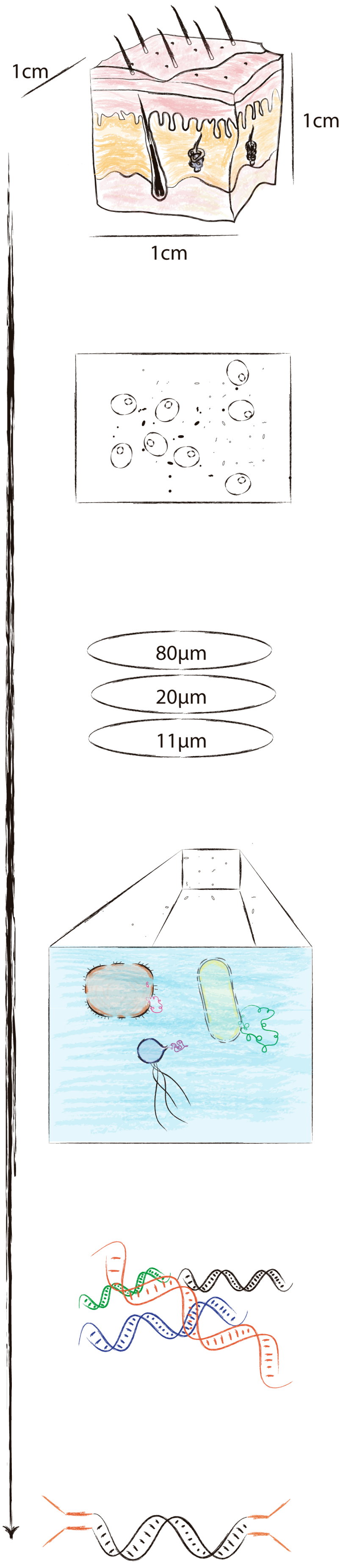
Eight healthy C57BL/6J mice of 8 weeks of age were included in this study. All steps of animal handling were carried out to minimize the stress of the animals while keeping them isolated to reduce possible microbiota share. All mice were euthanized through cervical dislocation, according to a local IRB-board (PRBB, IACU committee) approved protocol, and a region of 3x3 cm was excised from the dorsum-lumbar region, using a sterile blade, and frozen in liquid nitrogen to preserve the integrity of the skin. The samples were subsequently split using a 6 mm punch blade and stored at -80°C until further experiments. All mice followed the same feeding rate, and were born and housed under the same conditions in an animal core facility, accredited by AAALAC International. Of these individuals, six were used to test bacterial diversity (named B1 to B6 when used for bacterial enrichment and T1 to T6 for total DNA extraction) and two, namely Sample A and Sample B, for functional analysis.

A negative control was included in the analysis to test all possible materials and reagent contaminations.

Procedure

The proposed method and its validation is shown in Supplementary materials and methods. In brief, a graphical diagram of the procedure can be seen in Figure 1. A 6mm skin sample is digested using a buffered enzymatic solution for 3 times in constant shaking. The resulting solution was sequentially filtered using sterile nylon filters, eliminating all the host cells.

To avoid mitochondrial contamination (see supplementary materials for further information), mitochondria were eliminated through flow cytometry, leaving only cells with a genome size larger than 0.5Mb.The remaining cells were then digested and DNA was collected using a standard Ph-Chloroform method.



Skin samples were obtained using a 6mm punch blade, obtaining a 6x6x10mm region.

Sample was weighted and put into a 2mL Eppendorf tube

A combination of tripsin + dispase and collagenase was prepared in HBSS medium.

Skin was incubated in series of 30 minutes with the enzimatic solution

Resulting cell suspension was then filtered using sterile nylon filters of 80,20 and 11µm pore diameter.
Filters were washed using sterile HBSS

Bacteria were recollected in low binding tubes
Bacterial suspension was incubated with Lisozyme (10mg/mL) 30' and DNA was extracted using Ph-Chl standard method.
DNA was quantified by qPCR using 16S rRNA universal primers

Bacterial DNA was fragmented by sonication using a Diagenode® Bioruptor, in low-binding tubes

454 Libraries were prepared according to Zheng et al. (2010) and quantified using qPCR method proposed in the same experiment
Sequencing was performed using Titanium plus system

Figure 1. Schematic description of the proposed method.

The method was validated through specific amplification of bacterial and host gene, and 16S rRNA amplification sequencing to assess the amount of host and bacterial DNA, and the possible bacterial diversity bias associated to the method, compared to a standard DNA extraction without Microbial DNA enrichment.

Finally, the method was tested by performing a metagenomic analysis of two mouse skin samples, presenting for the first time a functional assessment of the deep skin microbiota.

Results

Microbial and host DNA detection by standard PCR in the proposed method

Skin samples were collected from genetically and environmental homogeneous mice to reduce as much as possible the environmental variability that could affect the diversity of the skin microbiota. Our method resulted in a successful isolation of around 5 ng of DNA (ranging from 3.2 ng, to, 8'7 ng). We tried to amplify a host gene, IRGA6, and the bacterial 16S rRNA to assess the ability of our protocol to substantially reduce the amount of host DNA. No band was observed in the IRGA6 amplification, but all of 16S rRNA amplification resulted in one size-specific band (Supplementary Figure 1). Control DNA extraction, as expected, resulted in a specific band in both 16S and IRGA6 genes. No visible bands were observed in any of the samples or extraction methods for the primate-specific NPIP gene used to monitor human contamination. Negative mock sample did not amplify for any of the three methods, suggesting no contamination was carried through the reagents.

Host DNA reduction test by qPCR

A qPCR experiment was performed using the same genes and the same samples, to measure the actual amount of host DNA in each method (Supplementary Figure 2). Cq values are shown in the supplementary table 1. The same trend in the standard amplification protocol was observed in quantitative PCR analyses. Although in most cases bacterial DNA yield reduction was observed in the method proposed due to the retention of bacteria by nylon filters, no statistical difference was observed (Chi-square p-value=0.089), while host DNA yield was significantly reduced beyond the limit of detection of the thermocycler (Cq > 35) in all cases (p-value<10-5). If we consider that the standard curve minimum was set at 10-4 ng for host DNA, with a Cq of 32, values of Cq over 35 suggest a complete elimination of host DNA from the sample. Mock sample resulted in lower scores, similar to the negative ones (16S amplification Cq = 33.08/33.46, IRGA6 amplification Cq = 37.27/36.89). We have observed that 16S amplification always result in an negative amplification over the 30th cycle. This result has been corroborated by other colleagues, and it suggested to be normal. According to this, 16S amplification of mock sample may be considered as negative.

Diversity bias assessment

Taking into account the bacterial DNA yield reduction, the following question was to assess whether the reduction was associated with a loss of diversity in our bacterial community associated to the skin. The bacterial 16S rRNA gene was amplified in all samples and extraction methods and sequenced using parallel-tagged 454 titanium technologies, resulting in 654,472 sequences with lengths ranging between 50 and 850 nucleotides. Of these, 645,474 reads (98.6%) passed all filters of length and quality and were used for further analyses. Pooled reads were split by sample resulting in 19 samples with reads, ranging from 1,000 to 22,000. Mock sample just retrieved 17 reads of low quality, and then it was eliminated of further analyses.

At the same time, pooled sequences were clustered using the greedy algorithm implemented in CD-HIT resulting in 8,696 and 1,275 OTUs at level 99% and 97%. A phylogenetic tree was constructed using the prior probabilities and seed tree provided by RaxML using the default configuration. As our alignment contained only 181 phylogenetically informative positions, which are not sufficient to classify the 1,275 phylotypes, we used the reference sequences from the NCBI to construct the tree. Taking into account the stringency of our assignment, the resulting tree can be considered representative and more realistic than the one resulting from using only the 181 informative positions. Supplementary Figure 3 shows the phylogenetic architecture of our community, which was predominantly represented by Proteobacteria, with a mean frequency of 79% and a standard deviation of 8% (supplementary Figure 4). Of these, the gammaproteobacteria were the major representative of this group with ~ 50% of the reads. But in terms of taxonomic diversity, we observed that most samples tended to carry more species belonging to the betaproteobacteria genera than the rest of

Proteobacteria classes (an average of 28% of betaproteobacteria, compared to a 5% of alphaproteobacteria, or 0.5% of the rest). The rest of OTUs were spread in several bacterial phyla, but most sequences were assigned to Firmicutes, Actinobacteria, and Bacteroidetes, the other three principal phyla previously defined (Grice et al. 2009).

To further test whether a diversity bias exists associated with the method, the taxonomic abundance was compared between the new and the standard protocols. Interestingly, the bacterial enrichment method yields a higher number of phylotypes compared to the standard method. The same trend was also observed by Shannon Diversity and Rarefaction curves(Supplementary table 2 and Supplementary figure 5). However, differences among groups are not statistically significant (Wilcoxon test. p-value = 0.327). Samples appear to be different, but they do not cluster by method. CCA shows no separation by sample or extraction method (Figure 2). On the contrary, NMDS analysis showed a complete scattering and admixture of the samples showing no aggregation by method (Supplementary figure 6). Taxon abundances were not statistically significantly different among samples (p-value=0.375) or among extraction methods (p-value=0.795).

Metagenomic library preparation and analysis proof of concept.

Two independent samples were extracted according to the method proposed. Sheared DNA was blunt-end repaired and ligated to the adaptors. Library quantification was performed by qPCR, and the correct amount was used for emPCRand sequencing according to the method adaptation from Zheng et al(Zheng et al. 2011), resulting in 60,488 (sample A, MG-RAST accession number 4496968.3) and 65,647 reads (sample B, 4496969.3).

As for taxonomical classification, in both samples, 95% of reads were assigned to bacteria (Figure 3), 1.92% to other commensals (fungi, arthropods), and 0.02% to host DNA. Thus, the enrichment protocol has resulted both in a total amount of DNA (~ 5 ng) and in a proportion of microbial DNA that make a metagenomic analysis feasible.

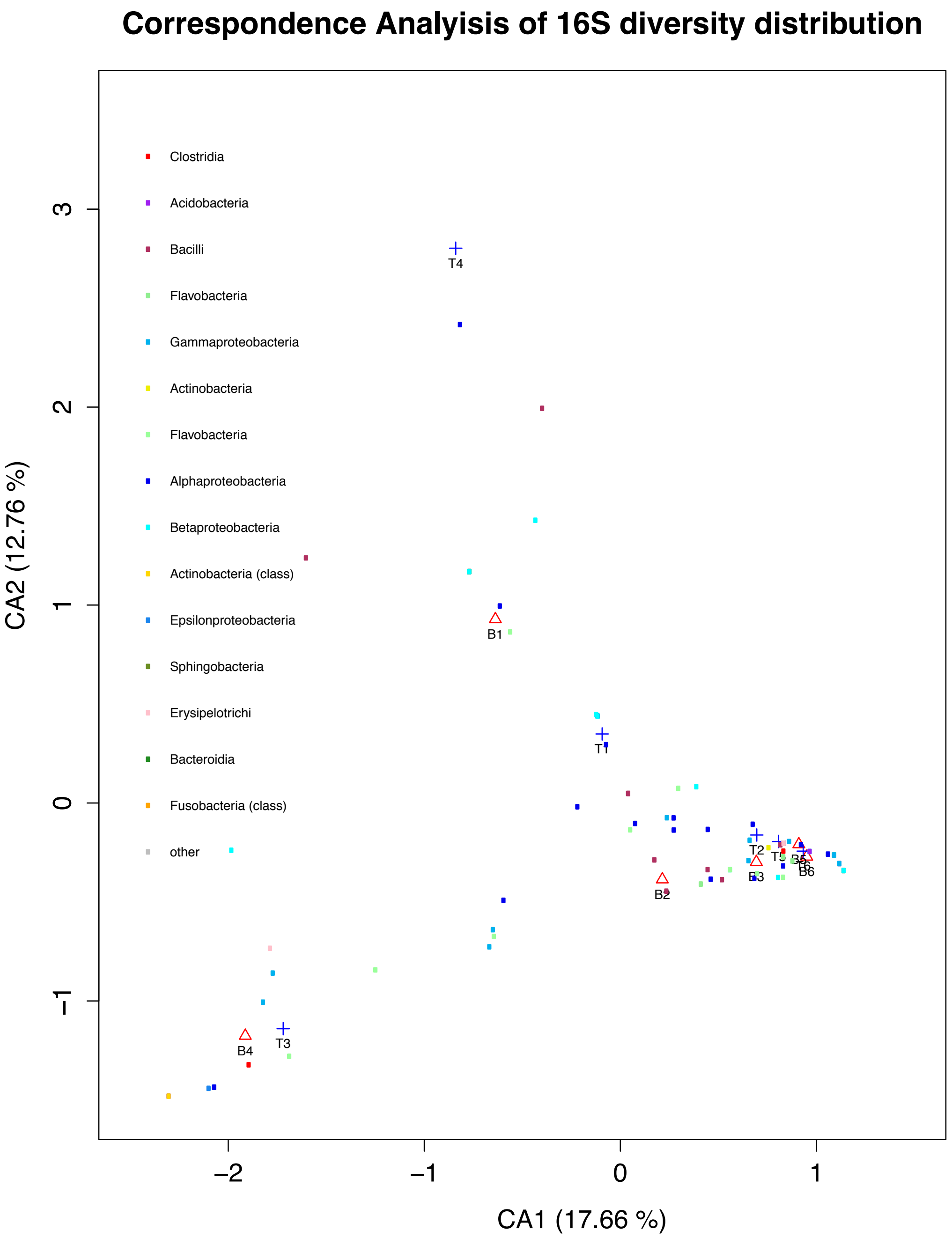


Figure 2. Canonical Correspondence Analysis (CCA) of the bacterial diversity in skin in Standard and Proposed methods.
Methods are named by the capital letter (B from Bacterial Enrichment extraction and T from Total DNA extraction)

Most of the bacterial reads belonged to the Proteobacteria phylum (85% of reads in sample A and 88% in sample B) whereas the presence of the rest of phyla was quite limited, except in Firmicutes. These proportions are similar to those observed with 16S rRNA sequencing (Supplementary Figure 4). In lower taxonomical levels a difference appears between samples:

Sample A seemed to be less diverse than sample B. At class level, although in both samples Gammaproteobacteria was the dominant class, the presence of other classes was quite limited in sample A, whereas in sample B the presence of other classes, such as Betaproteobacteria, was higher. The same pattern occurred in lower taxonomical levels. For example, at order level 74% of the reads of sample A were assigned to Pseudomonadales and only 48% in sample B. In this sample the presence of Burkholderiales and Enterobacteriales was higher (16% and 15% respectively) than in sample A (around 2% both). At family and genus

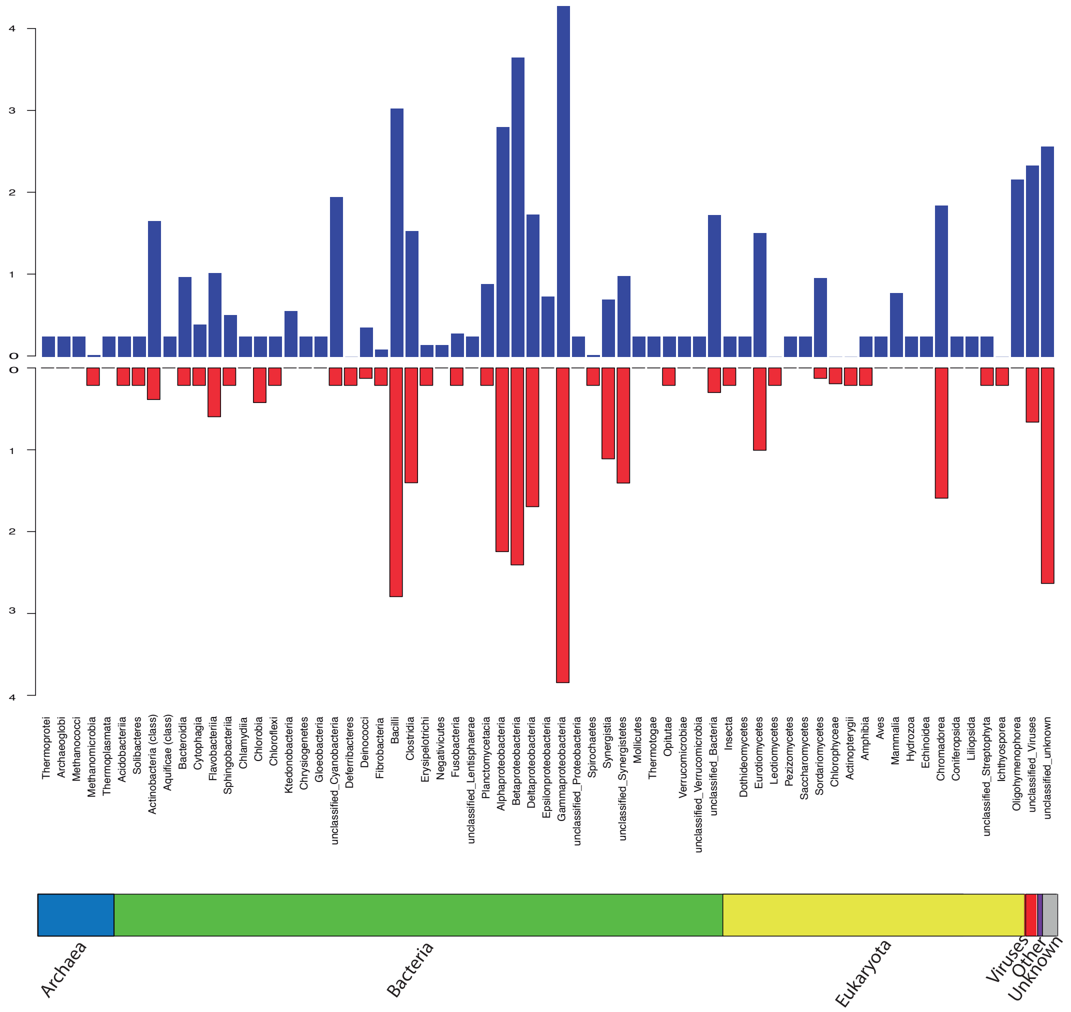


Figure 3. Comparison of the relative abundance of bacteria in skin metagenomic samples.
454 reads were split by taxonomy at class level. Relative abundance was log transformed to reduce the drastic differences among taxa assignation. Samples A (red) and B (blue) are faced by taxa, to facilitate the comparison.

levels, Moraxellaceae (72% in sample A and 48% in sample B) and *Acinetobacter* (71% and 45%), respectively, were the dominant taxa. These results agree with the 16S taxonomical distribution for both methodologies.

In addition, the metagenomic approach allowed discovering non-bacterial species. 1.8% of all the dataset was assigned to Eukarya, and of those, 56% was assigned to Fungi, 13% to Arthropoda, and less than 20% to Chordata. Fungi classes such as Dothideomycetes(Schoch et al. 2009), Eurotiomycetes(Ng et al. 2013), and Leotiomycetes, among others (Figure 3), have been observed in our dataset, and have previously been isolated from mammalian skin. Interestingly, a wide range of plant classes were retrieved.

Functional analysis was performed using HMM constructed with the multiple alignments of the functional categories from COG(Altenhoff and Dessimoz 2009) and eggNOG(Jensen et al. 2008). At COG 1 level, unknown function or general function prediction, both categories assigned to unknown function categories, were the most frequent function in both samples (53.3% and 27.2% each). Within the known functions, catabolism, as expected, was the most frequent, comprising mostly protein and carbohydrate degradation (11.93% and 29.85% respectively). Other categories such as nitrogen metabolism or respiration were also enriched, although to a lesser degree. Although in general, Sample B was more diverse than Sample A, similar trends were observed in both samples.

Following on the functional analysis, datasets were split taxonomically, to assess similarities and differences at functional level for each taxonomic subcluster. The functional distribution in both samples was different due to the different rate of reads assigned to the unknown function cluster, but at taxonomic subsystems the differences were more dramatic. Relative abundance patterns were different among taxonomic subsamples (Figure 4). Although we observe that the basic functional trends are present in all class clusters (including replication, protein and energy production), the most abundant function present in each sample and each taxonomic cluster was different. For instance, genes assigned to motility were enriched in the Bacteroidia cluster from sample B, but not in other class clusters of the same sample or in any of the taxonomic clusters from sample A. In general, the functional signal from sample B was broader and stronger than that from sample A, in agreement with the reduced diversity observed. Sample A was enriched in replication, recombination and repair genes, and this trend was shared by distant taxonomic groups, suggesting a possible environmental effect affecting the functionality and diversity of skin microbiota of mouse A. But differences were observed even in taxonomic groups with a wide range of functions observed, such as Alphaproteobacteria,. Taxonomic clusters were analyzed using

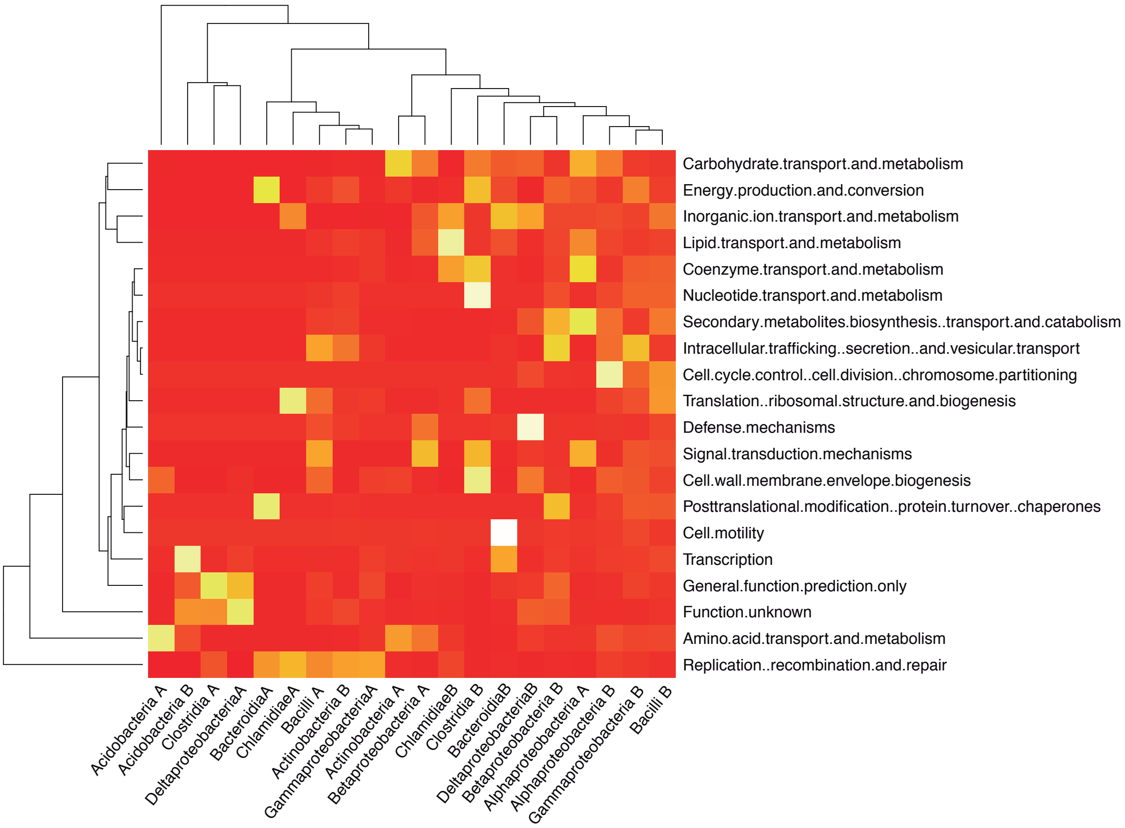


Figure 4. Functional profiles of Skin metagenomes, split by taxa.
Classification was made first by taxonomic assignation using PhymmBL, and then functional classification was based on EggNOG using HMM classification. Counts were normalized by sample and taxa using log transformation. Color gradient indicates the level of representation for that taxa. Hierarchical clustering was performed by function and taxa.

Correspondence Analysis and Non-Metric Multidimensional Scaling (NMDS) (Supplementary Figure 7). In both cases, they showed no association by taxon. While Sample B aggregated around the centroid, suggesting that functions were homogeneous at all taxonomic levels, sample A was more spread, separating more functions. We do not observe aggregation by sample either.

Discussion

Here we present a new method, based in cell and molecular biology techniques, to sharply reduce the amount of host DNA from a skin biopsy sample obtained by standard procedures. We have validated the new method in terms of isolation, efficiency and taxonomic bias inferred by qPCR results. With this protocol we have obtained 95% of bacterial DNA recovery compared to less than 1% expected by direct skin biopsy DNA extraction and sequencing (standard procedures). On average we obtained 5 ng of bacterial DNA, which is sufficient to prepare a sample using the protocol by Zheng et al.(Zheng et al. 2011). Although pool sequencing was performed to obtain equimolar quantities of each sample, we observed a bias on the resulting read count for each sample. Keeping in mind that the quantification of samples prior to the pooling was done by qPCR, the variance in sequence retrieval difference is likely to be due to stochastic variables of the sequencing procedure or differences in the efficiency of the fusion primers used for the experiment. Nevertheless, the taxonomic structure was maintained, even recovering more diversity than in standard procedures, including eukaryotic members of the microbial consortium associated to skin. The importance of detecting fungi or arthropoda is clear since they are also commensals, and may be also involved in the maintenance of the skin homoeostasis. It is important to note that we could detect those organisms even with the tight particle size restrictions imposed by our method. Our results are in agreement with the phylogenetic diversity observed in other skin microbiome 16S rRNA studies (Grice et al. 2008; Garcia-Garcerà et al. 2012), which implies that our method does not introduce any major taxonomic bias. Moreover, our study unveils that mouse skin carries less number of phylotypes, suggesting that is a low diversity ecosystem characterized by the prevalence of Proteobacteria, in comparison with other environments such as gut, where Firmicutes and Bacteroidetes are the most abundant, or human skin where the predominant phylum seems to be variable according to the skin niche analyzed(Grice et al. 2009; Blaser et al. 2013; Redel et al. 2013). Rarefaction curves (supplementary figure 8) show lower number of phylotypes per 1000 reads, with a lower slope than the ones observed in gut(Turnbaugh et al. 2010). These results agree with previous 16S-based works (Grice et al. 2008; Costello et al. 2009; Garcia-Garcerà et al. 2012), implying that the use of this method does not entail a bias and, in addition, new taxa may be discovered.

This protocol allows not only the taxonomically characterization of the skin microbiota but also to explore the functionality of that microbiota and the possible biochemical and molecular relationships among its individuals and between the microbiota and the host. To demonstrate it, two samples have been processed using our method, and have been fully sequenced to see differences in taxonomy and function, compared to the results obtained from 16S rRNA analysis; for the first time, a functional analysis of deep-skin microbiota can be produced. Besides the differences in the taxonomic distribution, one can also uncover previously unknown functional trends that occur on the skin ecosystem, which may have an important impact in health and cosmetic pharmaceutics.

The different functional databases used produce different results based on specific biases of their original datasets, and therefore we used different databases to characterize the functional annotation of the skin metagenomes. Although catabolism was the most abundant functional category in our datasets, other traits can be considered more interesting; categories such as respiration, nitrogen metabolism and other clusters associated with aerobiosis were enriched. Also aminoacid metabolism and protein degradation were important in our dataset. However, this may be related to taxonomic abundances and, then, it must be taken carefully. Nevertheless, all these trends are in agreement with the particularities of skin as an ecosystem, considering the continuous skin replacement and the direct contact with the atmosphere.

One of the most interesting questions that can be addressed using metagenomic data is the functional niche occupied by each taxonomic group, which has been called the “who does what” question. However, despite the fact this question can only be answered with transcription data, one may approximate the analysis to the “who can do what”, with functional analysis of metagenomic data. Even if our approach was just a proof of concept for the new methodology proposed, the functional analysis applied to taxonomic specific data showed interesting results. In some of the taxonomic clusters we observe different directions for both samples, suggesting that the main functional role for those samples was totally different in both ecosystems. These differences may be due to specific events that only occurred to that mouse, specific differences between both mice (metabolic, immune,...) or it was due to some bias produced by the procedure in that specific sample. Further sampling and research is needed to understand the skin in an ecosystem point of view.

Using this protocol, the level of knowledge of the skin microbiota may be brought to the level of the widely analyzed gut and oral microbiota, in which microbiota has been related to physiological and pathological changes in the host(Bäckhed et al. 2004; Bates et al. 2006; Benson et al. 2010). Gut microbiota has been related to immune system and gut differentiation(Butler et al. 2000; Mazmanian et al. 2005; Bates et al. 2006; Round and Mazmanian 2009) and chronic inflammatory diseases(Mazmanian et al. 2008; Nell et al. 2010; Docktor et al. 2011). With this method we open a new door to explore the skin microbiota and its possible implications in complex behaviors of skin, including skin complex diseases, as it has occurred with the gut and oral microbiomes.

However, one of the most striking points of this methodology is the facility to translate it to other host-microbiome scenarios. The most interesting case, and the easiest to apply, is the gut. It is interesting that none of the metagenomic studies performed in gut has been performed in biopsies, given the low bacterial/host DNA ratio in them. That is the reason why, up to date, all gut metagenomic studies have been performed in feces. The same problem occurs in the oral ecosystem where most studies have been performed in swabs and plaque extraction, where low host cell load exist. In fact the most recent paper from the Human Microbiome Project Consortium was performed using swabbing methods in 18 regions(Consortium 2012). But the bacteria in close relationship with the host cells, those in constant contact, are the most important in the progression and the maintenance of the ecosystem homoeostasis, and should be studied in more detail.

Although all the previous ideas are only suggestions, we think that the method can be a strong advance in the field of metagenomics and all its variants. Working with deep epithelial biopsies would shed light into the host-microbiota relationships in any conditions, pushing further the studies of host-microbiota interactions and its relationship in health and complex diseases.

Chapter 3. *Staphylococcus* prevails in the skin microbiota of long-term immunodeficient mice.

Authors: Marc Garcia-Garcerà, Mireia Coscollà, Koldo Garcia-Etxebarria, Juan Martín-Caballero, Fernando González-Candelas, Amparo Latorre, Francesc Calafell

Environ Microbiol. 2012 Aug;14(8):2087-98
PMID: 22524615

ABSTRACT

Host-commensal relationships in the skin are a complex system governed by variables related to the host, the bacteria, and the environment. A disruption of this system may lead to new steady states, which, in turn, may lead to disease. We have studied one such disruption by characterizing the skin microbiota in healthy and immunodepressed (ID) mice. A detailed anatomopathological study failed to reveal any difference between the skin of healthy and ID mice. We sequenced the 16S rDNA V1-V2 gene region to saturation in ten healthy and ten ID 8-week old mice, and found than all of the healthy and two of the ID mice had bacterial communities that were similar in composition to that of human skin, although, presumably because of the uniform raising conditions, less interindividual variation was found in mice. However, eight ID mice showed microbiota dominated by *Staphylococcus epidermidis*. Quantitative PCR amplification of 16S rDNA gene and of the *Staphylococcus*-specific *TstaG* region confirmed the previous results and indicated that the quantitative levels of *Staphylococcus* were similar in both groups while the total number of 16S copies was greater in the healthy mice. Thus, it is possible that, under long-term immunodeficiency, which removes the acquired but not the native immune system, *Staphylococcus epidermidis* may inhibit the growth of other bacteria but does not cause a pathogenic state.

Garcia-Garcera M, Coscolla M, Garcia-Etxebarria K, Martin-Caballero J, Gonzalez-Candelas F, Latorre A, et al. Staphylococcus prevails in the skin microbiota of long-term immunodeficient mice. Environ Microbiol. 2012 Aug;14(8):2087-2098.

Chapter 4. Meta-metagenomic analysis of gut microbiota and overall health status

Authors: Marc Garcia-Garcerà, Falk Hildebrand, Marie Joossens, Francesc Calafell, Jeroen Raes.

Work performed during the short stay in Brussels. Not meant for publication

ABSTRACT

The study of microbial ecosystems has shifted from a taxonomic point of view to a more systemic approach. The functional and compositional characteristics of host microbiota may be related to host phenotypic and metabolic traits, such as metabolite concentrations in blood or differences on the inflammatory conditions. Those variables may have an effect on the microbiota from a certain niche or vice versa.

We have performed a multidimensional analysis of a wide range of metabolic and inflammatory variables in relationship with the taxonomic and functional variation of the host GIT microbiota. We have defined a score based on those variables to define what can be called metabolic health.

Our results show a weak association between health score and microbiota at both functional and taxonomic level, which replicate previous analysis performed in more severe conditions like Crohn's disease. This association suggests a possible ecological succession from a more beneficial conditions, when the host is healthier, to the worse conditions, when the host suffers from a well-defined disease.

Introduction

The study of microbial ecosystems has shifted in the last decades from a taxonomic point of view to a more complex systemic approach. Given the ubiquity of microorganisms in Earth, the limit to study the microbial ecology of a given ecosystem was mostly technological. A high proportion of microorganisms cannot be cultured (between 80-99%) (Olsen et al. 1986) because of their specific resources usage. It is important to understand that the behavior in vivo may not be the same than in vitro, since they live in the presence of other microorganisms that interact with them, modifying their expression profile accordingly(Vieites et al. 2009). Eco-systems biology seeks to understand the complexity of all molecular processes from the different organisms in a given niche that contribute to the ecosystem homoeostasis. This systemic approach is possible by the appearance of new technologies that allow the retrieval of a vast variety of high-throughput data from a single sample or multiple correlated samples from the same time point(Raes and Bork 2008). For several decades, microbial ecology was limited to ribosomal RNA studies, characterizing the main participants on the picture and their main representation without, however, explaining their role. This fact has totally changed with high-throughput sequencing technologies(Mardis 2008). The ability of sequencing genetic material (both genomic and transcriptomic) directly from the sample takes the ecological knowledge from a phylogenetic point-of-view to a next level with a strong increase in complexity and information that has to be understood(Warnecke et al. 2007). The study of a system as a whole enables the modeling and further prediction of variation through the introduction of disturbances and external inputs(Bork 2005). A microbial ecosystem can be defined as the ecological system that includes all the biotic and abiotic elements that function and interact together in a certain geographical location(Zengler 2009) and a certain time point. The level of interaction will depend on the number of species, the nutrient sources and the geography of the habitat. In the case of a host-associated ecosystem, the role played by the host is akin to that of geography, and the system equilibrium will also depend on the host metabolism and the effects of the interaction between the host and their commensal microbiota(VerBerkmoes et al. 2009).

The human body surfaces harbor a great multitude of microbial communities that outnumbers our own cells by one order of magnitude(Gill et al. 2006). Of them, most inhabit the gastrointestinal tract. The 16S rRNA based phylogenetic classification of human microbial communities showed that among the 70 bacterial and 15 archeal lineages described to date, 22 bacterial lineages were found associated with human body, although most of the sequences belong to four main phyla: Actinobacteria, Firmicutes, Proteobacteria and Bacteroidetes(Costello et al. 2009). However, the relative abundances of each phyla vary depending on the niche studied. In the gut niche, the dominant phyla are Bacteroidetes and Firmicutes(Eckburg et al. 2005) compared to other ecosystems such as skin, where Proteobacteria and Actinobacteria prevail(Grice and Segre 2011). The gastrointestinal flora plays a fundamentally important role in health and disease, but the ecosystem remains incompletely characterized and its diversity poorly defined(Hooper et al. 2002; Sekirov et al. 2010). Growing evidence demonstrates that the normal gut microbiome contributes to the development of systemic alterations such as diet-induced obesity(Turnbaugh et al. 2006). Colonization of germ-free mice with the normal gut microbiota from conventionally housed mice leads to increase of fat mass, independently of the amount of food they were given(Bäckhed et al. 2004). This, and other facts, lead to the necessity to understand how gut microbiota affects the host in a systemic way. To do so, The MetaHIT (Metagenomics of the Human Intestinal Tract) project collected fecal specimens from 470 healthy individuals from Denmark and Spain(Members 2010), and has sequenced total DNA to catalogue the metagenomic gene pool to assess the complete metabolic picture of the gut microbiota. Additionally, they collected a wide range of host variables at different levels that may be used to construct a global schema of the host metabolism. We have characterized that metadata information to define a population structure based on healthiness to characterize whether how healthy random individuals are associated with which microbial community inhabits their gastrointestinal tract.

Materials and Methods

Data collection, and manipulation

Metagenomic and sample-associated metadata were collected from the human gut microbiome project "MetaHIT" database (www.metahit.eu). In total, 462 samples with an associated wide range of metadata information were collected from individuals of European ancestry in Spain and Denmark. Clinical information was classified into different categories: blood values, anthropometric variation, microbial diversity indices, vitals, and genetic information.

Our main concern was to assess the relationship between microbiota and health. Beyond the classical definition of health as the absence of disease, a number physiological parameters are known to be good predictors of future disease, often when a known threshold is exceeded. We first defined healthiness asa quantitative trait that summarizes such clinically relevant variables, and that can be used to grade individuals that would be usually considered as healthy. According to this definition we selected a subset of variables that were metabolically relevant to the definition of health. Thus, 35 out of 54 variables were chosen and tested for normality (SUPPL. TABLE 1). Variables with a bimodal distribution were considered to produce stratification, and then they were used to sub-classify the dataset according to population groups. The rest were normalized based on the original distribution, and centered to a normal distribution N(0,1) according to the clinical health thresholds.This normalization is crucial to make all variables comparable among them, reducing possible range size, and allowing multidimensional scaling without preponderance on any variable to the rest.

Defining healthiness, according to clinical variables.

First, all samples were categorized by the number of altered variables, based on the clinical thresholds. This classification was used as rank for further analyses. Correlation analysis was performed among variables, as a standard scale reduction method. Only variables with a correlation index below 0.7 were included on the analysis. For pairs of variables with a correlation index over that threshold, a mixed score was calculated as the mean of the sum of both values.

A non-metric multidimensional scaling (NMDS) approach was used for scale reduction. We calculated the Mahalanobis and Bray-Curtis distances between samples, considering each variable as a different dimension. The Bray-Curtis distance takes into account the dissimilarity between multidimensional points, while the Mahalanobis distance gauges the similarity of the dataset against a random distribution. The different samples were then ranked according to the dissimilarities of the points in the variable space. The ranking approach starts in a random configuration and finds the configuration that preserves the rank-order dissimilarities as closely as possible. We used a set of 200 initial random starting configurations, and retrieved the most likely final configuration. Distance matrices and NMDS were calculated using the R packages vegan and ecodist.

The first two or three components of the multidimensional scaling were considered for further classification of health. A new (2D or 3D, respectively) dimension-reduced space was defined. Two methods were used to classify the health level (the healthiness) of the different individuals. First, considering healthiness as a continuous variable, the healthiest group of individuals, defined as the group with less overall alteration in its clinical variables, was used, and the midpoint to all of them, i.e. the centroid, in the new dimension-reduced space. The distance from the centroid to all points was calculated, and considered as the healthiness score. And second, based on the number of altered variables, used to rank the health level before, we clustered individuals in groups. Given that the dimension-reduced space was continuous, our interest was to discretize the distribution, trying to achieve the less overlapping possible between groups.

Alternatively, metadata variables were clustered by its role in the systemic function. Functional subgroups were split and used to calculate a specific score for that function . In all cases, we applied Principal Component

Analysis (PCA), followed by the same scoring pipeline than in the whole metadata space analysis.

Taxonomic assignment of metagenomic reads

Taxonomic assignment was performed using the pipeline previously developed (Garcia-Garcera, M). An NCBI-NT database was filtered to remove all non-bacterial reads using a custom script based on the Lowest-Common Ancestor algorithm. Metagenomic reads, filtered by quality and length, were first clustered into phylotypes at 97% similarity and 90% coverage, and then aligned to the customized 'nt' database. Reads with taxonomic uncertainty, meaning the reads assigned to different taxonomic groups, were assigned as the lowest common taxonomic level, until family. Reads unassigned at family level were classified as 'unranked'.

Functional assignment of metagenomic reads

For each read, a putative aminoacid sequence, based on the six different frames, was aligned against the eggNOG database (v3.0), and KEGG database (release 58.0) using HMMER 3.0. Each protein was assigned to a KEGG orthologous group or a eggNOG orthologous group. Unclassified proteins were clustered at 70% and considered as novel protein families, but they were not used for further analyses.

Comparative analysis of healthy subgroups

A two-tailed Wilcoxon rank-sum test was used to identify the association between the metagenome profile and the health groups, and supported additionally with a multitest-corrected Kruskal-Wallis test . We applied the q-value method proposed in a previous study to estimate the false discovery rate (FDR)(Storey 2002). In our metadata-related metagenomic analysis, the statistical hypothesis tests were performed on a large number of features at class, family, genus, and species levels, and at KO and NOG profiles. Given that a FDR was obtained by the qvalue method, we estimated the power π for a given P-value threshold with the following equation:

Here, π0 is the proportion of null distribution P-values among all tested hypotheses; Ne is the number of P-values that were less than the P-value threshold; N is the total number of all tested hypotheses; FDRe is the estimated false discovery rate under the P-value threshold.

In contrast, at continuous level, a Spearman's correlation test was performed between the health score value and the relative amount of each taxonomic and functional category. A Spearman correlation test was also corrected using the same q-value method.

Results and Discussion

Multidimensional reduction, and classification of samples according to their health state.

Metadata was available for 356 out of 462 samples. All variables were tested for population stratification, by age and gender, resulting in two equivalent gender groups, with an approximately normal distribution of age, ranging between 30 and 70 years old. We did not observe any difference among those subpopulation distributions for any sample. However, two different subpopulations were observed on the 356 samples processed, based on the data distribution. Two different variables (Body Mass Index[BMI] and diabetes diagnosis) displayed a bimodal distribution, we subdivided the samples according to them (Figure 1). Three different datasets were prepared. First, a control group was constructed with all individuals without diabetes and a BMI below 30. Second, all individuals were included in the dataset, but divided by diabetes. We did not differentiate between T1D or T2D because, although they have different

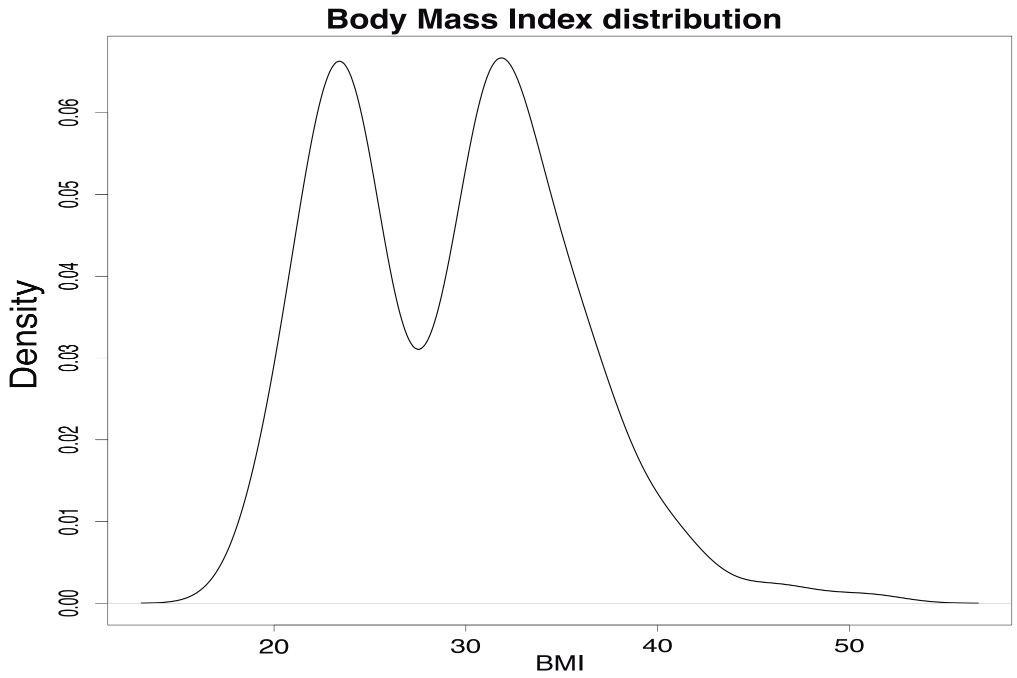


Figure 1. Bimodal distribution of BMI across the study population.

s o

BMI distribution shows a clear difference among two overlapping subpopulations. Subpopulation division was performed by the other clinical variables, resulting in two approximately normal populations.

triggering factors and molecular basis, at metabolic level they can be considered similar for our purposes. Finally, we defined a third dataset according to BMI. Given that BMI subpopulations strongly overlap , we proposed to divide the population in two subpopulations as subjects who had BMI < 25 or only BMI over the clinical threshold, and the population with one or more variables altered, including BMI. This division resulted in two approximately gamma distributions that were clearly differentiated and that could be normalized. Given that the second subpopulation had a wider range of alterations, only the first subgroup (only people with BMI below 25 or only BMI over the clinical threshold) was included to the study of health and microbiota.

Mahalanobis and Bray-Curtis dissimilarity distances were calculated and used to perform a multidimensional scale reduction using NMDS. Figure Xa shows a continuous distribution of samples according to the number of variables altered and the level of alteration. To remove possible variable neutralization, variance explained by the same rank level in the NMDS was taken as absolute value to construct the eigenvalue. Given that subject clusters were not readily apparent in the NMDS plot (Figure 2), we clustered the samples according to the number of variables they had beyond the clinical thresholds, and grouped allowing the minimum overlapping possible. Two possible solutions came up for grouping. Dividing the samples in groups based in equivalent alteration increase, the division by three groups was the best option, to minimize overlap. However, if we allowed an asymmetrical division to minimize overlap, we could separate the individuals with less than two clinical variables from the rest. This healthy group was almost completely isolated from the rest of the variables. These results were expected because in the three-way division, we observe a strong overlap between the two less healthy groups. Note that all this individuals were considered as control cases in all further studies performed by MetaHIT consortium.

Following the same procedure, we subdivided the variables according to their role in the systemic spectrum. We performed the same analysis for each subgroup of variables. Only fat-metabolism, glucose metabolism

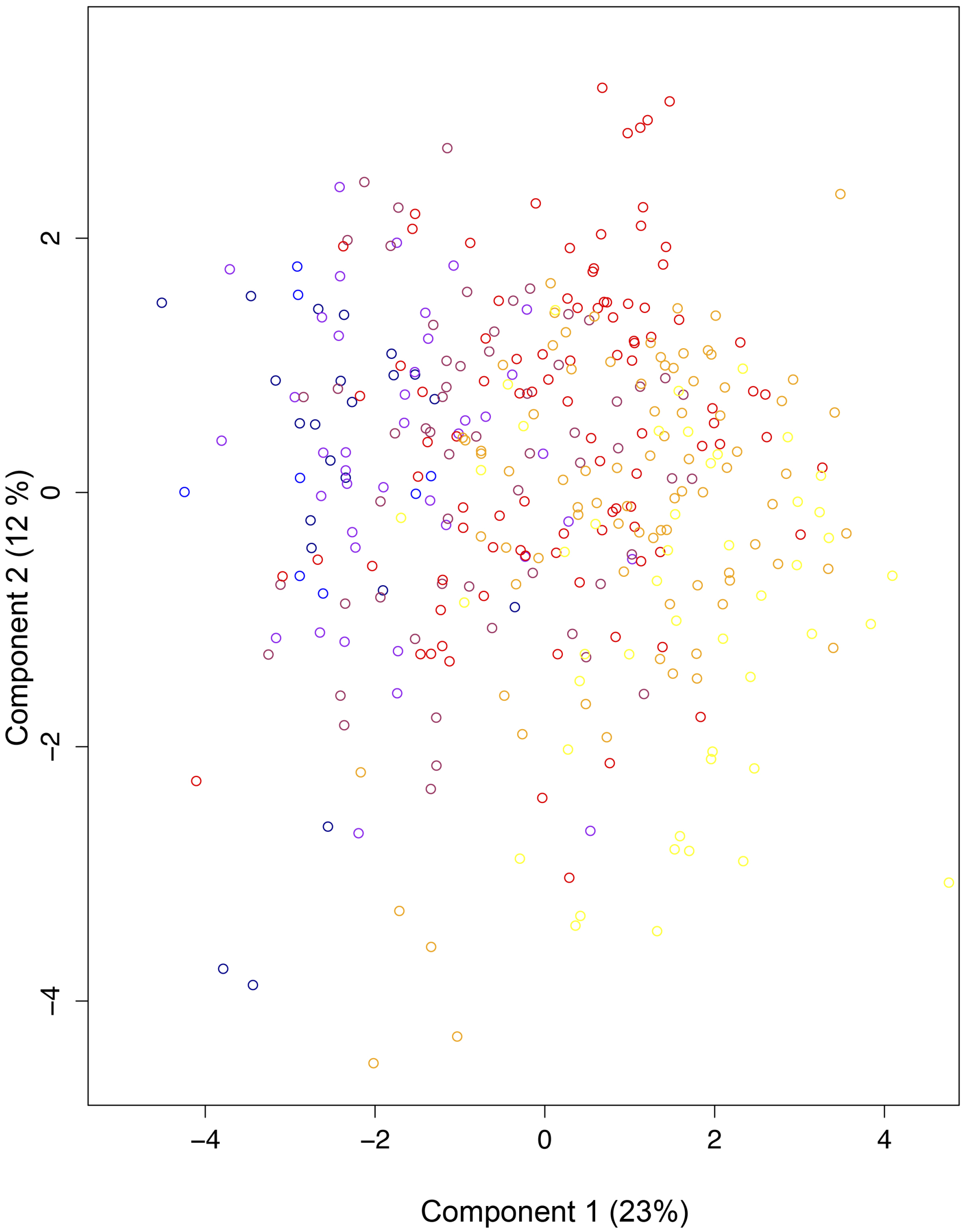


Figure 2. Non-metric Multidimensional Scaling of clinical variables
Mahalanobis distances was calculated for each sample, according to the normalized values of each variable. Ranks were constructed from those distances and the eigenvalues were calculated according to the rank distribution. Discreet classification was used to classify the different samples, from healthier (blue) to more altered (yellow).

and inflammation presented a gradient structure, and then were the only ones included for further analyses (Suppl. figures 1-3).

Taxonomic diversity and comparison

Based on the taxonomic assignation of reads, we calculated the alpha diversity for each sample using Shannon Index (H'). Comparison of the Shannon index between healthy groups showed a higher diversity associated with the healthier individuals compared to less healthy subjects. This association was concordant with the previous results of health clustering (Figure 3). Another interesting trend was that, together with the higher Shannon diversity associated with the healthiest, there was a lower standard deviation associated to the healthy groups. This lower standard deviation was also associated to a lower variability in species composition among individuals associated with the healthiest state.

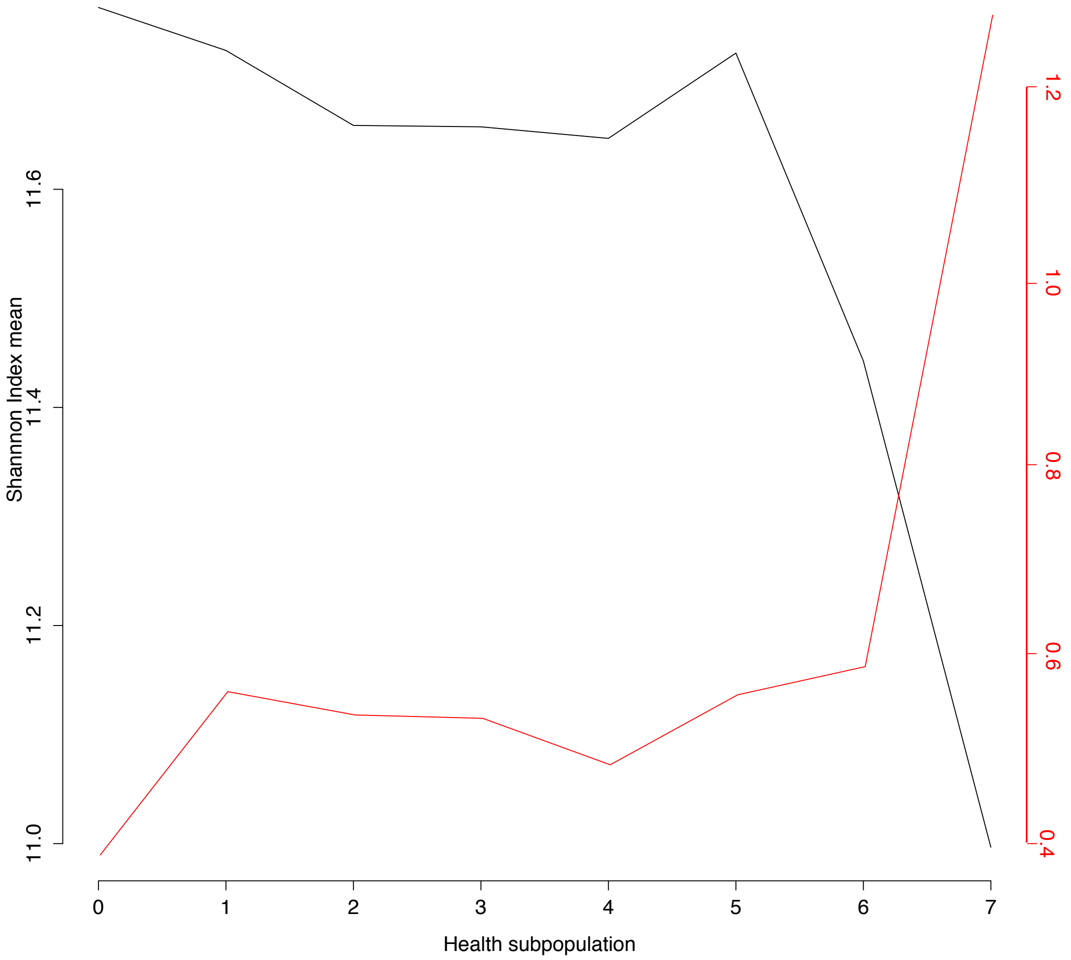


Figure 3. Comparison of the proposed health score with the alpha diversity
Discrete subpopulations were constructed by counting the number of variables over the clinical threshold. Mean (in black) and Standard deviation (in red) variation is shown.

Shannon diversity index distribution shows similar results than those observed in the health clustering analysis, where the healthier subpopulation presents a higher Shannon index, with less variation, followed by a plateau in the intermediate subgroups and a final group of a metabolically altered individuals with a really low Shannon diversity and a high variation between individuals. Similar results were obtained when comparing the alpha diversity with the health score obtained by NMDS. A negative correlation between the health score and the Shannon diversity was observed (p-value = 0.001091, Student's t-test).

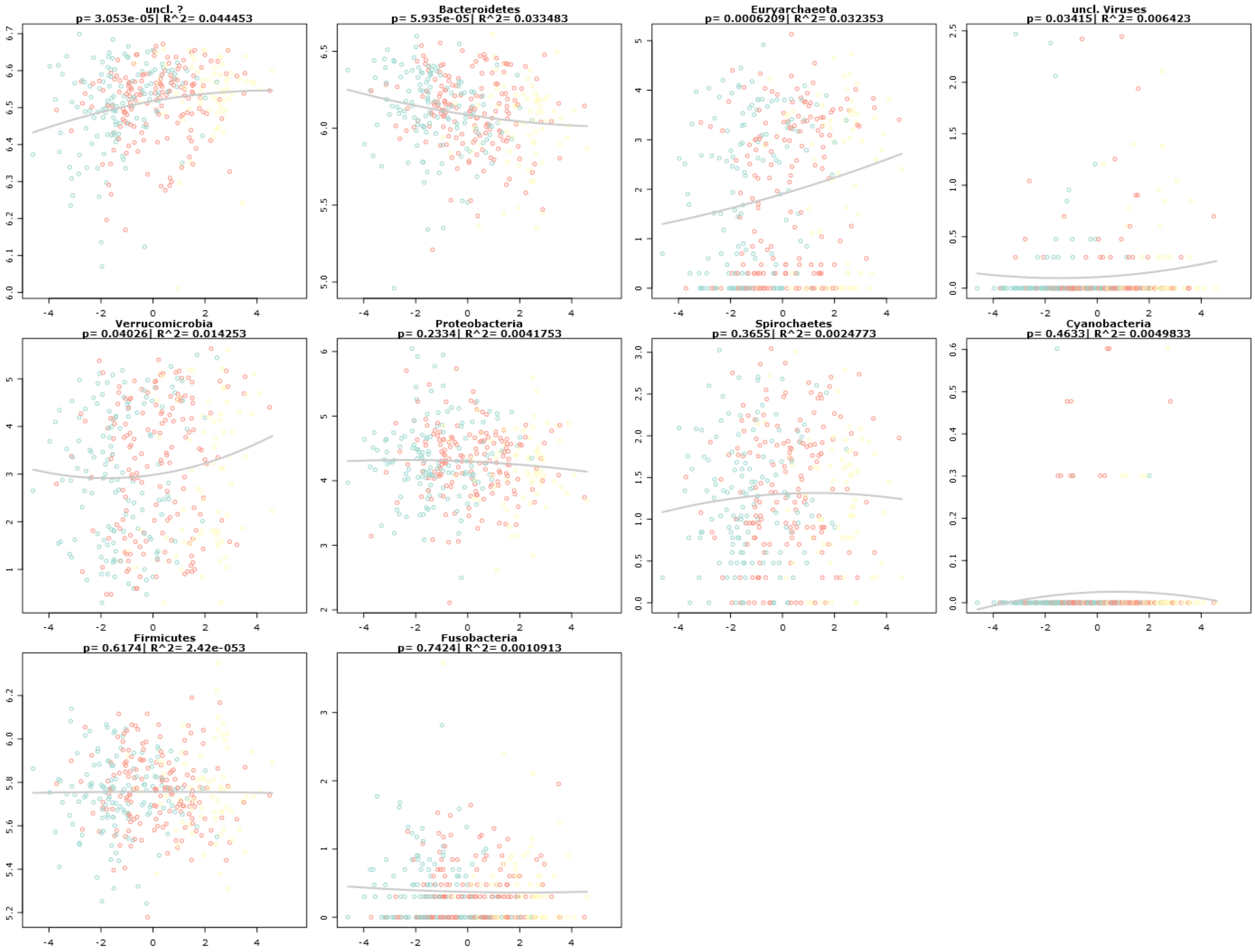


Figure 4. Spearman correlation between health score and phylum classification. Each subplot was constructed to characterize the possible correlation between one phylum group and the different variables according to the health score. Three subgroups were defined (see text for more information), going from green (healthier) to yellow (more altered). Health score was plot on the X axis, while the base 10 logarithm of the relative abundance of each taxonomic group was plotted in the Y axis.

The taxonomic comparison was performed at different levels, from phylum (Figure 4. See Supplementary tables 1-3 for full information, and real size figures in supplementary material) to species. One of the first trends observed was the correlation between the proportion of unclassified reads and the health score. This result was significant in both discreet (Kruskal-Wallis (KW) test with Pvalue=0.028 after multiple test correction) and continuous (Spearman correlation test with q-value = 8.165x10-5 after FDR correction) comparison at all taxonomic levels from phylum to genus. Comparison between healthy and diabetic individuals shows similar results, supporting the association between high diversity and health state.

A new classification of species composition has been recently defined with enterotypes, namely population clusters based on the species relative abundances in gut microbiota. This population stratification could be associated with different phenomena, including health or geographic origin. We tested the relationship between our health score and the relative abundance of each enterotypes. However, we did not find a clear association between the three different enterotypes defined by Arumugam et al. and health state in our samples (Arumugam et al. 2011) (KW test, p-value = 0.1375). This lack of association between the health state and the enterotype clustering was previously reported in the metagenomic-association study of Type-II diabetes(Qin et al. 2012) (Supplementary Figure 7). This lack of association with health, however, is interesting because the origin of this population differences, unrelated to geographic origin and health, remains unknown.

According to previous results, a decrease in the ratio Bacteroidetes/ Firmicutes was associated with obesity in humans(Ley et al. 2005). Considering obesity as an alteration of the health state, one may expect to observe a decrease in the relative amount of Bacteroidetes or an increase of the Firmicutes. Interestingly, although the level of Firmicutes does not seem to be related with our score method (KW test, q-value = 0.5855), the relative abundance of Bacteroidetes does decrease significantly (KW test, q-value = 0.0392), and supported by continuous univariate analysis (Spearman correlation test, p-value = 0.00323). Other significant phyla that were altered by our health clustering method in both discrete and continuous univariate analysis were Euriarchaeota (KW test q-value = 0.026) and Fusobacteria (Spearman Correlation test, p-value = 0.0014). This signal is maintained and amplified through higher levels of taxonomy. Class Bacteroidia correlated with healthiness (see supplementary figures 4-6 to see class, family and genus associations), according to our classification, following on the previous results. The association of *Bacteroides sp*. with health was also previously reported in comparative analysis of Crohn's disease and Inflammatory Bowel Disease (IBD)(Erickson et al. 2012). Previous comparative studies of taxa distribution in disease states such as IBD, Crohn's disease or Diabetes, were observed in our study. *Sutterella sp.* (KW test q-value = 0.008), *Butyvibrio sp*. (q-value = 0.04) or *Roseoburia sp.* (q-value = 0.047) (see table X and supplementary tables X,Y,Z for more information) were previously associated with healthy individuals(Walker et al. 2011; Erickson et al. 2012; Qin et al. 2012). However, although these taxa were significantly associated, even after multi-test correction, the correlation coefficient and regression slope in each case were very low. This weak correlation agrees with the fact that all individuals in our study were considered healthy and then, the differences among groups are not so strong. This observation can be considered to be supported by the fact that even all taxa with a p-value below 0.05, but not supported after FDR, are taxa previously suggested to have a main role in health in other studies. For instance, our study reports a significant association between the genus *Lactobacillus* and health. This result is supported by a high number of studies that demonstrate the role of *Lactobacillus* species as probiotic(Kleerebezem and Vaughan 2009).

Clinical information was split according to their systemic role in different clusters. We chose those the clusters that displayed a linear correlation with the Shannon diversity index. Those were the clusters of fatty acid metabolism, glucose metabolism and inflammation. These correlations agree with the fact that only Diabetes and BMI displayed a bimodal distribution, which are known to be associated with inflammatory and metabolic disorders, suggesting that maybe this population cannot be considered as a good healthy control group, although, they are, in fact, a good representation of the normal european population. Our results show a that health can be considered as a continuous characteristic that involves a high number of variables. We have observed that reduced alterations in those variables have an effect in the microbiota composition, at all levels. And more importantly, those alterations are also found, in a more drastic way, related to common diseases, such as Crohn's disease or IBD. Thereby, using the dataset we are analyzing here as a control group may mask true associations between microbiota and disease. According to our results, it is critical to be more strict when a control group is being chosen.

Fatty-acid metabolism Cluster (FAC) was also associated with higher relative proportions of unclassified phylotypes, supporting previous results. To assess whether the variables included in FAC were driving the association of unclassified phylotypes with health, they were removed from the general score and we repeated the analysis. Although the correlation was weaker without FAC variables, it was still present and significant, suggesting a true association between diversity and health. However, other phylotypes such as *Sutterella sp*., previously associated with health, were inversely correlated in FAC, and, once we eliminated the FAC variables from the general score, the signal was lost (Figure 5). This kind of situations where a phylotype is correlated with the general score but not with the metabolism-specific scores was common,highlighting the complex interactions between the microbiota and their host. Glucose Metabolism Cluster (GMC) showed different results. Higher scores in GMC, meaning more alterations in the glucose metabolism showed higher ratio of Proteobacteria/Bacteroidetes. Interestingly, we observe a totally different behavior between the FAC-related and the GMC-related taxonomy. Alterations in glucose and fat metabolism may be related to different diets, and then, those alterations could not be associated with host metabolism, resulting in a spurious correlation. Diet information, which was not available in this study, may add to the understanding of the interindividual differences in gut microbiota, although self-reported diet data is notably unreliable. The inflammation-related cluster of host metabolic variables (IC) was also correlated with H'. However, at all levels, the relationships observed were similar to the ones observed with FAC, with association of unclassified levels with low levels of inflammation. Also, higher levels of rare taxonomic groups like cyanobacteria were associated with a higher alteration of the inflammatory variables. In this case, we also observed associations of low Bacteroidetes and high Verrucomicrobia ratio with health.

We performed a bootstrap-like subsampling approach and posterior correlation test to further investigate associations between the different taxonomic groups and inflammatory panel. This analysis failed to find a correlation between the Bacteroidetes and Verrucomicrobia relative abundances and the inflammation level, but still showed association with the unclassified phylogroup. Bacteroidetes and Verrucomicrobia behave randomly in most of the samples except for a few outlier samples, based on the relative abundances of those taxa.The effect of those samples in the distribution was strong enough to make the correlation significant.. Removing just three individuals from the sample the correlation ceased to be significant, and then

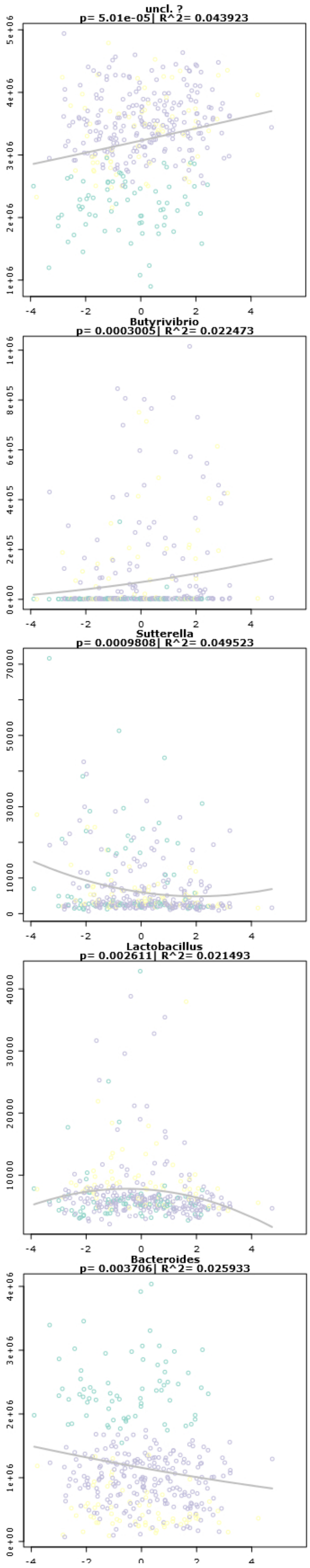


Figure 5. Correlation between fat-related health score clusters and genus.

The three health clusters (green as the healthiest and yellow as the more altered) obtained with the complete health score were used to discriminate the different samples. The new workspace was defined through multidimensional scaling of the fat metabolism-related variables. Spearman correlation test was performed between the fat-related score and the relative abundance of each taxa.

it must be considered as biologically irrelevant.

We finally considered the three scores (FAC,GMC and IC) as three dimensions of the health space. Given that those three scores were already ranked, the individual with the minimum average rank in the three dimensions was deemed the healthiest and used as a reference point. We calculated and Mahalanobis distance from that individual and all the other individuals i, resulting in a functional-based score. That score was tested for correlation with the same taxonomic groups analyzed before. At high taxonomic level (Phylum) we observed similar patterns than the observed separately, with higher *Bacteroides*, unclassified and Fusobacteria correlated, and Proteobacteria inversely correlated with health. Those correlations were maintained at lower taxonomic (genus) levels with association between less healthy individuals to potential pathogenic genera such as Escherichia or Streptococcus, and lower relative abundance of unclassified phylogroups in the case of more altered pathogenic states. This strong association of Proteobacteria and low health state may be result of the sampling approach performed by the MetaHIT consortium. New, alternative, and more stringent approaches, avoiding bimodal distributions for any of the variables included in the study would be needed to refine our knowledge of the relationships between the host and the microbiome associated.

Bacterial genomic functions and healthiness

We performed a correlation analysis using Spearman's Correlation and diversity using KL test against COG and KO databases to assess the functional association between health score and microbiota metabolism. Both discrete and continuous analyses showed a strong signal at COG level. Although most of the functional categories associated with our health score were not previously described, they were strongly associated, even after FDR. But, for the categories with known function, discrete analysis showed a strong association between health state and COG categories related basically to three subfunctions (Figure X): first, individuals with more metabolic variables altered, showed a reduction in the relative abundances of categories associated to catabolism.

Aminoacid and nitrogen metabolism were reduced. Some variables increased with the reduction of health, like pyrimidine synthesis, were previously associated with immune-mediated inflammatory diseases (IMID), such as psoriatic arthritis or rheumatoid arthritis. Enzymes related in those pathways have been used as treatment targets in IMID(Ash et al. 2012; Qi and Hua-Song 2013). Most of the functional gene categories that are elevated in low health individuals involving catabolism inhibition were associated with phosphonate metabolism; phosphonates are compounds that contain phosphorus-carbon bonds and act as enzymatic inhibitors in bacteria(Ford et al. 2010).

Second, the sporulation machinery was increased in individuals with more health variables altered. Multiple genes involved in the reduction of metabolism, capsid biosynthesis and sporulation were increased (p-values < 0.005 after FDR). Sporulation is a process mainly characteristic of Firmicutes and Actinobacteria, where spores act as a dormant form in case of reduced amount of nutrients or other stress events(Szponar et al. 2003; de Hoon et al. 2010). The association of sporulation with health may be related to the alteration of the normal ecosystem, with increased inflammation, resulting in a stressful condition for bacteria. However, taxonomically, we do not observe any modification of Firmicutes relative abundance across the different health subclusters. Further functional analyses, involving sequencing of bacterial mRNA would increase our comprehension of bacterial response to the increment of host alterations.

The third subfunction altered is the reduction of gene expression. Altered host states were associated with a reduction of gene expression in microbiota. In eukaryotes, genome-wide studies have shown that signal transduction pathways control a variety of downstream elements that allow a rapid change in the transcriptional landscape of a cell within minutes of exposure to stress(de Nadal et al. 2011), including down-regulation of catabolic and anabolic metabolism expression and increasing stress-dependent gene expression. Gene expression kinetics observed in response to stress is achieved by fine regulation of multiple steps of the mRNA biogenesis process. Although this is common to many stressors, the underlying mechanistic details of how such regulation is achieved are highly dependent on the particular stressor and organism. In bacteria, alteration of DNA topology to alter the gene expression panel is one of the most common(Franzmann et al. 2008). In fact, COG categories associated with stress, including heat-shock proteins, or two-component systems were present and correlated with the amount of alterations in host metabolism. Also, although catabolism was inhibited, a large amount of ribosomal proteins were increased in the low-health clusters and correlated with a less healthy score.

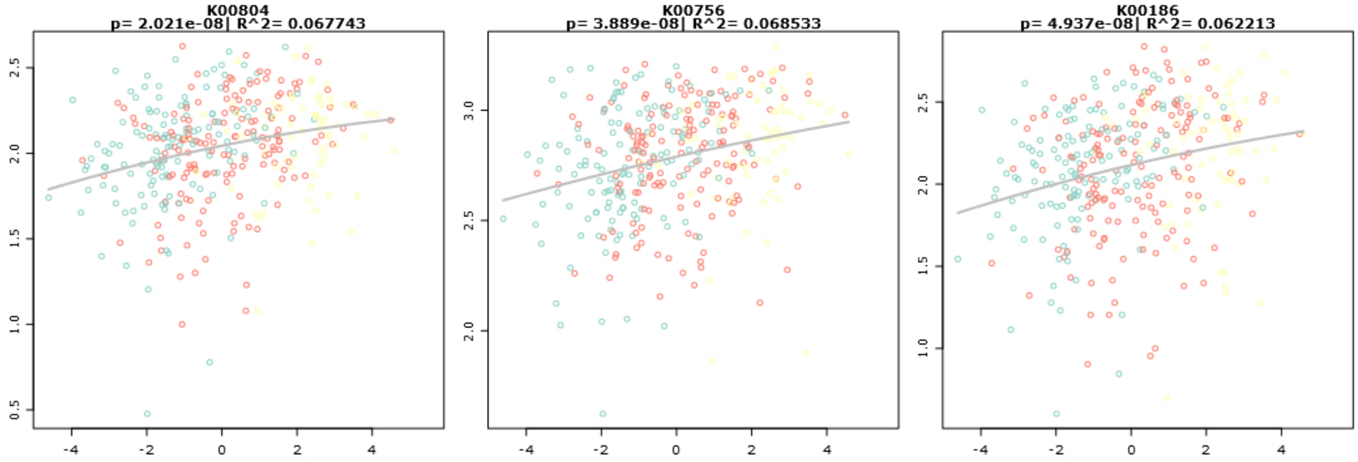


Figure 6. Significantly associated KO categories with health.
Health score and discreet subcategories (shown by colors) were tested for correlation using Spearman Correlation test. See Supplementary materials for further information.

Kegg Orthology (KO) comparison showed a lower but still strong signal at both discrete and continuous analysis (Figure 6). Although some of the categories are different, we observe similar results. First, phosphonate transport and metabolism was associated with low health clusters, as happened in the COG analysis. However, in the case of KO, protein-synthesis was correlated with high health clusters. As happened with COG, we observed a strong correlation of the health state and the primary, secondary metabolisms, at all levels, while dormancy, sporulation and replication were correlated with the reduction of health. Phages and prophages were elevated in low health levels. Mobile element movement and replication is highly correlated with stress conditions(Garcia-Russell et al. 2009). Interestingly, we observe an inverse correlation of health with virulence. Phages and prophages are transmitters of virulence and resistance factors. Given this striking result, we performed a bootstrap-like subsampling approach to test the significance of this result, but the correlation remained significant.

All these results can be considered as pieces in the construction of the whole picture of correlations between the host and microbiota metabolisms. Those relationships are based on covariance analysis in closely related subpopulations. Our significant observations are multiple, but weak, compared to those observed in other comparative analysis between highly different ecosystems. In case of disease-control comparisons, such as in diabetes or Crohn's disease, significance was much greater than in our study, mainly because clinical differences were also greater. But interestingly, our observations, although weaker, contain the same variation profiles than those observed in those diseases and follow the same trends observed in them, supporting that associations observed in diseases are already present just below the threshold that defines disease. This association between subclinical alterations and microbiota variation, at both taxonomic and functional level, supports the idea that health should not be dichotomized, but considered as a continuous trait that affects the system at all levels and all of their integrants. According to this idea of continuity in health and system disturbance, one should select groups clearly differentiated in terms of health to allow a meaningful difference, and avoid the noise produced by variable interaction.

We need to keep in mind that these results are just observational and from a single time point. Since it is known that the biocenosis affects the biotope and vice versa(Tringe et al. 2005), we ignore whether this correlation is related to causation and which would be the direction of causation. Since we lack properly collected diet, fitness, drug intake, and other variables that are important for health and that may modify the behavior of gut microbial communities, our results may be refined when we include more information at all levels, even at level of population structure. Our results, although significant and concordant with health-disease studies, may vary when using other population stratification methods, and then they must be taken carefully.

Moreover, since this study is based on one single time point, the result may change including more observations for the same individual. This is crucial, because ecosystems are not static but stationary and in equilibrium. The gut may be modified everyday just by changing the diet type or origin, introducing new taxa or new functions to the ecological equation, resolving it in a different manner than the previous observation. And still, the ecosystem will continue being stable and resulting on the best situation for both host and microbiota. The notion of a homeostatic but not immovable metagenome should always be taken in consideration to achieve a proper understanding of the host-microbiota system.

Discussion

Methodological improvements

from pure culture to skin metagenomics

In this thesis I have presented the preparatory work for the study of mammalian skin from an ecosystem point of view. To understand the skin from this perspective, one should consider all the characters represented in the picture, including the commensal microorganisms that live, interact and benefit our organism. To access that information, the genomic content of all the integrants of the ecosystem has to be considered. Still, some methodological gaps to study the skin as an ecosystem remain. First, although the skin microbiome has been shown to be highly diverse from a complete perspective, skin microbiota is limited in terms of population(Fredricks 2001; Grice et al. 2009). A practical implication is that a given skin region would yield limited amounts of bacterial DNA. Although the number of bacteria depends on the region, as does the relative taxonomic abundances, the numbers are still very low. Microbiology has traditionally been based on pure cultures grown in the laboratory, allowing researchers to increase their starting material (Riesenfeld et al. 2004). But most of the microbes cannot be cultured, because they have specific needs and signals from other microorganisms that co-exist with them, that create metabolites and other molecules that are essential for their growth. Traditionally, if a microorganism could not be cultured, its genomic content was not accessible to be sequenced, and then, its metabolic potential or its role in the ecosystem remained unknown. In the last decades, a strong effort has been put on the sequence of thousands of species from all three kingdoms (Bacteria, Archaea, Eukaryota), as well as a similar number of viral quasi-life forms. But also, one must remember that genomic information came from isolated organisms. The increasing knowledge of model organisms, such as Escherichia coli, has revealed a strong functional plasticity, that allows them to adapt and grow even in a pure culture(de Muinck et al. 2013; Dragosits et al. 2013). This plasticity, although allows it to be isolated, cultured and sequenced or manipulated, is not common. Only an estimated 5% of the bacterial species can to be cultured, and this percentage is reduced in other kingdoms or deep taxonomic groups(Tyson et al. 2004).

Two main milestones have been crucial to solve this limitation. The first one has been DNA cloning(Cohen et al. 1973; Nathans and Smith 1975). Molecular cloning allowed researchers to introduce fragments of known and unknown sources in a vector of known sequence to subsequently apply methods such as amplification or restriction digestion to characterize the sequence inserted on the vector. However, cloning is expensive in time and cost, and inefficient in retrieving both functional and phylogenetic data. This lack of efficiency fails to resolve the ecological puzzle of a given niche. This cultivation bottleneck in molecular cloning, although it provided a higher perspective than direct bacterial culture, skewed our view of microbial diversity and ecology.

But the second, and most important milestone, was DNA sequencing. Initially devised by Sanger and Coulson, DNA sequencing allowed to characterize the DNA molecule sequence, opening a very wide field of research(Sanger and Coulson 1975; Sanger et al. 1977). Specifically in ecology, the DNA sequencing methods allowed to start the characterization of a niche from a molecular point of view. The first attempt to directly sequence an environmental sample involved the cloning of genomic fragments and further sequencing using the classical method of dideoxynucleotide amplification termination, and posterior characterization. This type of studies showed a functional diversity far more diverse than previously expected. One example is the discovery of a previously unknown light-driven proton pump isolated from the SAR86 ribotype(Béjà et al. 2000). As already mentioned, cloning is expensive and limits the amount of information obtained. To characterize diversity, the solution was to use phylogenetic markers: genetic structures present in, up to date, all the organisms of a specific taxon with no or limited horizontal transference. 16S rRNA gene is one of the most widely used examples. Pioneered by by Carl Woese in the 1970s (Woese and Fox 1977), the comparison of 16S rRNA sequences has become a powerful tool to deduce the phylogenetic relationships between organisms, even at kingdom level(Böttger 1989). This tool has been applied in a really high number of environments, from deep-marine extreme environments(Bik et al. 2012) to the troposphere(Deleon-Rodriguez et al. 2013). A large range of host-associated environments have been studied, so far, including skin. Elizabeth A. Grice and collaborators first reported in 2008 the application of the cloning-based methodology to human and mouse skin, and revealed that Proteobacteria were the predominant phylum on skin(Grice et al. 2008); we replicated this result in healthy mice. But still, 16S rRNA only recovers the fraction of the phylogenetic diversity that is associated with the species amplified by the universal primers (Wang and Qian 2009; Kim et al. 2011) . Furthermore, it does not retrieve any functional information, further skewing the ecological information of the environment studied.

This limitation pushed, as explained in chapter whatever, the development and the appearance of a new set of technological platforms, that popularly received the name of “new generation sequencing (NGS) technologies”. Commercially appeared in 2004(Margulies et al. 2005), they solved all the cloning-associated problems of genomic studies, reducing the costs of data production with a substantial increase of the output. Thanks to NGS platforms, researchers leapfrogged the cloning and culture steps, pushing the ecological perspective of microbiology to a systemic level.

The multiple NGS platforms available to date present a wide range of sequence length outputs, starting from an average of 35bp to around 700bp, increasing their read length since their appearance(Metzker 2009). This read length has been shown to be inversely correlated with the amount of data retrieved. Depending on the project and the data needed, one must choose the platform that better fits their needs. Environmental studies samples present genetic material from different origins, and then one must solve which fragment belongs to which microorganism(Board 2008). Obtaining an unbiased view of the phylogenetic composition and functional diversity within a microbial community is one central objective of metagenomic analysis, and to characterize function and origin, DNA length is an important issue. To perform this characterization, alignment-based methodologies are the most widely used. But to differentiate between two closely related origins, or characterize the possible origin of a never characterized species, the read length has to be chosen carefully, if one wants to characterize a metagenome (Wommack et al. 2008). For that reason we have selected the 454 FLX Titanium technology (Roche Applied-Science), which produces reads of around 450bp, large enough to better resolve the metagenomic characterization of an environmental sample.

On the low microbial DNA yield in skin

Still all high throughput sequencing platforms without exception have a crucial limitation for our ecological study of skin, which is the required starting amount of DNA material(Hutchison and Venter 2006). As we suggested above, bacterial populations in skin are very low, which can be translated to a very low amount of bacterial DNA. Specifically, 454 technology requires 10-6 g of initial input to successfully construct a library, according to manufacturer's instructions, way over the concentrations we have been obtaining (not more than 5ng of total bacterial DNA, or 200 times less than required). However, the main limitation for NGS sequencing is to know exactly how many DNA fragments have been successfully ligated to their specific adaptors during library preparation, and then, it is crucial to perform a successful method to quantify that number. As we mentioned in chapters one and two, alternative strategies to construct a successful library without following Roche's protocol are possible (Meyer et al. 2008b; White et al. 2009; Zheng et al. 2010). In the 454 platform, the minimum amount of DNA material depends on the number of required enriched beads. To achieve that number, one must find a tight balance between the input DNA and the number of initial beads, to achieve a single DNA molecule per enriched bead. Beads enriched with multiple DNA fragments result in mixed signals during the sequencing process, which makes impossible to read the sequence, as happens in Sanger sequencing. On the opposite side, a low DNA/bead ratio results in an inadequate amount of beads, loosing the advantage of the full sequencing capacity. Zheng et collaborators demonstrated that, empirically, the input-DNA-to-bead ratio followed a Poisson distribution(Zheng et al. 2010) following the equation:

$$f(k,\lambda) = \lambda ke-\lambda/k!$$

where k is the number of molecules per capture bead, and λ was the input DNA-to-bead ratio. According to this distribution, the authors calculated the probability to have 0 or 1 molecules per bead, and the respective enrichment fraction to minimize the number of mixed beads. According to this result, it is important to know the exact number of molecules per sample, to correctly calculate the volume to use on the enrichment(Meyer et al. 2008b).

Here we have adopted and adapted the method by Zheng et al. to construct a library starting with the initial amount of bacterial DNA expected. Other alternatives include to previously amplify the sample using a whole genome amplification method, such as linker adaptor shotgun library(Breitbart et al. 2002), or Multi-displacement amplification(MDA)(Dean et al. 2002a). However, both methods have been previously reported to present multiple associated problems, including methodological associated bias, both at GC content(Bredel et al. 2005) and the relative representation of species and genomic regions(Kim and Bae 2011). It is clear that any method that includes a amplification step will result in a certain amount of bias, depending on the number of cycles, the relative differences of representations, and also a set of stochastic events, including what DNA fragments are initially nearer to the polymerase molecules(Kanagawa 2003; Aird et al. 2011; Pinto and Raskin 2012). In chapter one we demonstrate the possibilities and benefits of skipping any amplification process by using a very sensitive method to characterize the DNA amount before sequencing. In this case we have compared it with the most widely spread method: multi displacement amplification (MDA)(Raghunathan et al. 2005). In this work we compared the adaptation of Zheng's method and MDA, starting with 10,000 Escherichia coli K12 cells.

Using an amplification-free method has the main benefit of retrieving a better coverage output, compared to the control method. This is specially crucial when our sample DNA comes from multiple origins. Although most of the literature supporting this fact has been tested on 16S rRNA, their conclusions may be applied to any multi-template sources(Kanagawa

2003). Even when a sample has the same primer targets, the GC ratio is not the same on the different templates, and then the annealing and extension temperatures for each are different, conferring them different efficiencies to amplify. This fact can be extended when primers are different, as happens with MDA, where starting primers are a random distribution of hexamers, resulting in more abrupt differences. Our results have shown coverage differences of one to fifty in the different genomic regions, while more than 50% of the genome sequenced was totally uncovered. In contrast, the coverage output retrieved by the amplification-free method was homogeneous, with aleatory differences along the genome length. It is important to highlight that one of our replicates retrieved a very low number of sequences after the DNA extraction, shearing and adaptor ligation. In this case, we considered this result as an opportunity to assess the possibility to use a PCR step when the DNA source was below the minimum DNA amount. So we performed a 4-cycle PCR using primers against the adaptors. We did not observe any bias associated with that amplification. However, it is also true that four is a very small number of cycles, which do not seem to modify the average ratio of sources. Still, I would consider a failure to obtain less than 200,000 DNA fragments (meaning 110Mb of sequence) out of an initial DNA amount of 44Gb, needing to repeat the DNA extraction and improving each step to reduce the DNA loss for further experiments.

One interesting result from this work is the assessment of the origin or origins of an environmental DNA sample. In our case, what started as the determination of whether MDA unassigned hits originated from a concatenation of random hexamers, resulted in the ability to assign the taxonomic origin of those reads as phi29 polymerase production contamination. Bacteriophage phi29 is a Podoviridae-related phage that infects Bacillus *sp.*, mainly *Bacillus subtilis*(Vlcek and Paces 1986). This phage is able to catalyze a multi-stranded isothermal polymerization(Esteban et al. 1993) at very high accuracy. The enzyme that performs this reaction has been adopted as the best approach to perform whole genome amplification on very rare and limited samples, such is the case of single-cell approaches(Lasken 2007). MDA approaches have been very useful leading to important information about novel gene discovery(Marcy et al. 2007; Podar et al. 2007), or complex gene regulatory systems(Dupont et al. 2011). Still, our observations, in agreement with other authors' work, suggest a series of problems associated with MDA, described in detail in chapter one, which means that we would not use this method unless there is not any other option(Binga et al. 2008).

One of the critical limitations of MDA, as happens with any other universal primer amplification process, is the nonspecific product formation. As seen before, different sources have to be considered. But it is important to highlight that the most important trait to have into account is the ratio between the specific DNA to be amplified and the rest of the contaminants. Here we have been working with really minimal amounts of DNA, that is few tens of femtograms. Here, even when it is just a minimum, the possible contaminants, such as the ones described in chapter one, will be used as template for the isothermal amplification carried by phi29 polymerase.

On the alignment-free methods and the origin assessment

One of the side applications worth to mention in this discussion was the use of k-mer distribution to characterize the genomic origin of the set of unclassified sequencing reads. Intuitively, one accepts that a genome is not a random polymer of nucleobase bond pentoses. This non-randomness comes related to two main factors: First, the underlying genome structure is associated with the set of functions encoded in the sequence (in the form of genes or operons). Obviously, to perform a function, the substructure (gene or operon) must be translated into a protein or a set of proteins, that need to have a concrete amino acid sequence (or types of) in certain positions to be able to perform the specific function it is carrying. In this case, the structure comes from the function. This functionally associated structure is the characteristic used by the sequence-based algorithms to find the function, the phylogenetic origin, or their relationship with other species' peptides and genes(Rocha 2008).

Second, any genome has gone through an evolutionary history of selective pressures from an initial universal common ancestor to the present(Theobald 2010), adopting a (sort of) common genetic code, which also confers structure, evolving through reproductive events and adjusting their subpopulations in a certain niche according to selective pressures, gaining and losing information, fixing traits such as codon usage or amino acid limitations on that species in a certain time point(Gouy and Gautier 1982). In this case, the structure comes from the evolutionary history and the adaptation to the environment. This combination of environmental adaptation and common evolutionary history has suggested that more information is available rather than just sequence homology, which opens a new perspective to use alignment-independent sequence traits to characterize and compare a set of sequences.

The iterative functions for scale independent representation of biological sequences were proposed over twenty years ago (Jeffrey 1990). But they were first replaced by algorithms from computational statistics, such as stochastic modeling or hidden Markov model assignment for hypothesis testing and parameter estimation when comparing a set of two or more sequences. Still, these methods carry a bias associated with their string interpretation(Vinga and Almeida 2003). As happens with the other sequence-based algorithms, they fail when they face molecule fluidity. Fluidity is a molecular trait based on the idea that a proper folding of the protein is associated with a proper function. It seems that even in the possibility of recombination and shuffling, the function can be maintained, if, after the protein folding, the active sites are still available to the substrate(Zhang et al. 2002). This plasticity breaks the linearity needed for pairwise sequence comparison, making impossible to characterize function or homology in evolutionary related sequences. Assuming function conservation, local pairwise alignment algorithms emerged to solve this plasticity limitation, querying a sequence against a database of known templates(Altschul et al. 1990). Still, these methods are highly dependent on the size of the database, the knowledge of reference homologous sequences, and the dissimilarity matrix used to compare the different positions and calculate the score(Dayhoff et al. 1978).

In this work we found that MDA used contaminant DNA from different species. However, when we tried to assign the sequencing reads using standard alignment methods, we failed to characterize their phylogenetic origin. A large number of papers show that MDA reagents may carry DNA from the microorganism where the enzyme comes from, in this case a *Bacillus subtillis* phage φ29 (Bredel et al. 2005; Jiang et al. 2005; Le Caignec et al. 2006; Spits et al. 2006; Iwamoto et al. 2007; Woyke et al. 2011). We know that φ29 polymerase isolation requires source DNA elimination, but sometimes this process is inefficient, and short fragments may be kept in the solution, affecting the MDA process through chimeric sequence formation. If the fragments are short enough in origin (10-11 nt. long), the alignment-based algorithms will fail to characterize the origin of a chimeric sequence because, although they will be able to detect similar words in the database, they will not be able to progress through the elongation step of the dynamic programming algorithm(Phillips 2006). To overcome this problem, a number of algorithms based on the search of characteristic motifs have been developed. Some of them, including the one proposed by us in chapter one, are focused to project the high-dimensional genetic space into a low-dimensional numeric space and measure the similarity between different datasets(Blaisdell 1986; Almeida and Vinga 2006). Although different k-mer sizes have been used so far(Trifonov and Rabadan 2010), we decided to analyze the 6 nt. long motif distributions, since the priming substrate of MDA is a set of random hexamers. However, this methodology may be interesting for other motif sizes. An important case are the triplets, because of their implication, as codons, in the translation process from DNA to protein. One interesting observation has been the codon usage bias (CUB). CUB has been explained by two main factors: First, common codons in nature may speed up translation, and then may be selected for, while rare codon usage would slow down translation(Pan et al. 1998; Hershberg and Petrov 2008). Second, sequence traits such as GC content may also affect synonymous substitutions which may result in a pressure against the usage of certain codons(Chen et al. 2004). All these pressures may result in specific motif differences, which would be sufficient to discriminate among two different genomes by their relative codon frequency distribution(Belalov and Lukashev 2013). This CUB can also be observed by our 6-mer approach, where we observe significant differences that separate genomes according to their k-mer distribution.

In our case, we have used a multidimensional scaling approach to characterize and magnify the differences between a set of different genomes to test whether the hexamer distribution of our experimentally produced data was similar enough to some of them. Our results show a strong association between MDA datasets and *Bacillus subtilis*, the host bacterium of the phi29 phage. Interestingly, strong evidence suggests CUB coevolution between bacterial hosts and their phages(Brüssow et al. 2004; Carbone 2008; Lucks et al. 2008).

According to the extremely low rate of correctly assigned reads of our MDA dataset, my main concern would be to determine where the problems are in MDA, given that it is the approach of election when DNA amount is a limiting factor for further analyses(Hosono et al. 2003; Jiang et al. 2005; Lasken 2007; Rodrigue et al. 2009). Our results suggest that, in a very low-yield sample, MDA is not a method of choice, given that even reagent contaminants from the enzyme production process are in a concentration high enough to result in more than 80% of the further sequencing output.

Our results show that, in our case, where bacterial DNA yield in skin is so low and DNA loss may occur during the DNA extraction process, a whole metagenomic DNA enrichment procedure should solve our yield limitation. However, those methodologies carry a high number of biases and limitations, magnified when the yield falls below a minimum limit. According to our results, whole-(meta)genome amplification free methods are the only ones to provide unbiased, reliable and replicable genomic information from any sample with a minimum amount of starting material. I suggest, in agreement with my collaborators, that the kind of methodologies presented here should replace MDA in “Omics” projects, given their sequencing efficiency and lower cost-benefit ratio.

On the analysis of the skin microbial communities

Given this low microbial DNA yield in skin, researchers have been struggling to achieve the limits required by NG sequencing. Over the past five years, two main gigantic projects have been working to characterize the human microbiota: the metagenomics of the human intestinal tract (metaHIT) project(Members 2010) and the NIH Human Microbiome Project (HMP)(NIH HMP Working Group et al. 2009). Those projects have focused to increase the knowledge of the human genetic and physiological diversity, and its relationship with the microbiota and its effect on them, that reside with the human beings in a commensal/symbiotic relationship. Of both, only the HMP has considered skin. In fact, the knowledge about the skin microbiota has been really limited during the past years. Even when the HMP was producing an immense amount of data from other body regions, starting to leave the descriptive research for the new functional, hypothesis testing study of human microbiota, the skin microbiota was still a mystery in terms of function, co-occurrence, and interaction. During the progress of the HMP project (Weinstock, George. Personal communications. St. Louis, Vancouver, Geneva, Paris 2008-11) the inefficiency of the accepted methodologies and the enormous difficulties to obtain enough DNA from the skin microbiota have become apparent. These results point to the necessity of using alternative techniques, like those produced in our laboratory.

When the reference genome catalogue was constructed, less than half of the discovered skin-associated genomes were sequenced, focusing basically in three main genera: *Staphylococcus*, *Propionibacterium* and *Corynebacterium*(Human Microbiome Jumpstart Reference Strains Consortium et al. 2010). These three genera are the most representative taxa in skin, but they are far from being alone, and then, more information was expected to be produced in this first phase. During the past five years, the only knowledge about skin microbiota has been quite limited to its wide niche-specific phylogenetic diversity and its relationship with the physiological characteristics of those niches, in health(Gao et al. 2007; Fierer et al. 2008; Grice et al. 2008; Costello et al. 2009; Grice et al. 2009;

Fierer et al. 2010; Lemon et al. 2010; Blaser et al. 2013) and disease(Gao et al. 2008; Grice et al. 2010; Fahlén et al. 2012; Kong et al. 2012; Redel et al. 2013). But the functional profile of the skin microbiota has remained unknown during the past decade. Only recently, the HMP consortium has revealed a large dataset of functional and phylogenetic profiles of all human body sites, including four different skin regions(Consortium 2012). Still, to my understanding, the information retrieved by HMP is methodologically biased and incorrect. The sampling method chosen was based on ethical considerations, but not on real skin knowledge. It is known that microorganisms are ubiquitous in skin and its appendages. However, the sampling method chosen was based on the evidence that superficial swabbing was a good methodology to obtain a good microbial/host DNA ratio(Gao et al. 2010), and that there was the possibility to observe a similar diversity than with complete depth biopsies(Grice et al. 2008). But the different microorganisms have specific niches, like the hair follicle, in the case of mites, or certain fungi, like *Malasezzia sp.* And even accepting the constant skin layer replacement, as has been widely explained during the introduction of this thesis, the microbial cells being characterized through superficial sampling are either functionally different than the ones present in their niche, unviable, niche specific, or transient. The fact that skin is always in contact with the external environment and other individuals, may imply some superficial microbiota sharing, as has been observed by some researchers(Meadow et al. 2013). Still, as it occurs with other niche-specific microbiota(Turnbaugh et al. 2009; Arumugam et al. 2011), one should expect a skin microbiome core of resident bacteria, different from the total skin microbiome, located in deeper epidermic layers, being able to metabolize the side products from the skin, and maintaining the skin homoeostasis in constant interaction with the host cells(Grice and Segre 2012). Then, swabbing is not a good methodology and skin biopsies should be carried when someone wants to characterize the skin microbiota from a functional point of view. However, whole skin-deep biopsies contain a great amount of host DNA. Even when the host-microbial cell rate is very low, the differences in the order of magnitude make the direct sequencing inefficient in terms of cost and output.

Considering the experimental costs as an important variable in our research is as important as the proper experimental design. That is the main reason why we have designed a protocol to discard the host cells. Although the protocol that we have designed is more demanding in time, it yields a larger amount of DNA for a smaller cost. Also, the use of methods that retrieve the unbiased diversity in terms of molecular function and phylogenetic diversity should be a priority because they result in a better understanding of the skin (and any) ecosystem and permit further experimental designs to be more accurate to the actual homeostatic state.

One of the most interesting points of our approach is its ability to retrieve not only bacteria, but representatives of all three kingdoms present in the skin microbiota, and also the external group of the viruses. Until now, most of the publications have been focused on the bacterial kingdom, because of its implications in pathogenesis and immune system stimulation(Bek-Thomsen et al. 2008; Gao et al. 2008; Fitz-Gibbon et al. 2013). However, this classical point of view of a pathogen causing a disease does not fit with a broader, systems biology perspective where the microbial interaction with other microorganisms and the host may result in a complex host tissue behavior that may be interpreted as disease. In this sense, some results are emerging. This pathogenic complex behavior has been observed in the disease called colony collapse disorder, a bee (Apis mellifera) phenomenon in which all worker individuals of a colony drastically die. This phenomenon has been associated with a combination of multiple virus interaction which resulted in an abrupt death of the adult individuals of the colony(Cox-Foster et al. 2007). This is not an isolated case and also occurs in skin. Recently, Marinelli and colleagues reported that *Propionibacterium acnes* bacteriophage populations show a strikingly low genetic diversity, compared to other phage populations(Marinelli et al. 2012). This limited diversity pointed to unique evolutionary constraints imposed by the lipid-rich anaerobic environment in which *P. acnes* reside.

Interestingly, only two rare subpopulations of pathogenic *P.acnes* were resistant to the effect of *P.acnes* phages. Host variable modifications, such as lipid content, could result in ecological pressures to enrich a local region with transient *P. acnes* subspecies, which could activate the inflammatory host panel in a single occupied follicle, producing the complex physiological disorder called acne. On the same line, the authors of this work proposed to experimentally increase the diversity of *P. acnes* bacteriophages, as an ecological treatment for acne. These, and other examples have been reported recently, pushed the microbial ecology community of skin, to find a method to characterize the whole skin microbiota(Raes and Bork 2008; Grice and Segre 2011), and, in our case, to produce a methodology that allowed us to obtain a complete, ecosystem-based, picture of the skin.

The methodology presented here has been demonstrated to be efficient to retrieve complete, unbiased, microbial diversity in skin. Although most of the previous studies of human skin bacterial diversity retrieved a predominance of Actinobacteria, mouse skin resulted to be completely different in terms of taxa distribution. In this case, and in agreement of previous results, mouse skin was reported to be Proteobacteria-rich, with a strong fecal Firmicutes contamination(Grice et al. 2008; Scharschmidt et al. 2009). This contamination is commonly obtained through direct contact of the mice with the sawdust bedding, and cannot be controlled. The results presented in this thesis are just a descriptive approximation on the functional and taxonomic variability of skin microbiota, and then we cannot confirm if there is a role for this contamination, and if so, which is the role of this fecal contamination and other transient microorganisms in the skin ecosystem. Further specific analyses are needed to enlighten the specific role of them in the holistic perspective of the skin. And still, our work opens a new perspective for skin ecological studies because it allows to include archaea, fungi, viruses, and other microeukaryotes in the equation. The importance of detecting non-bacterial species is clear since they are also commensals and they may be important to maintain the skin homoeostasis. However, to date, we do ignore what they do in and to the skin(Consortium 2012), and, if they do so, how they interact with us. Even in the case that they could be merely commensals, important questions can be addressed. As happens with bacteria, other microorganisms (fungi,eukaryotic viruses, archaea other microeukaryotes) that reside in our skin need specific metabolic pathways to digest the side products of skin regeneration, while avoiding the immune system. Their methods to perform this molecular mimicry and hide from the host immune system are unknown(Elde and Malik 2009; Zhang et al. 2011). Commensal (not bacterial) skin microorganisms have been associated with complex diseases, such as asthma, atopic dermatitis or rosacea(Burgess et al. 2012; O'Reilly et al. 2012), and then they must be put on the complex ecosystems equation, when we want to characterize and compare the ecological differences between different skin states like skin complex diseases(Valdimarsson et al. 2009).

The methodology proposed in chapter two not only solves the phylogenetic bias produced. It also opens the access to the whole functional panel of the skin microbiome. In our case, we have observed that catabolic metabolism was predominant in our dataset. This trait agrees with the picture of the saprophytic microbiota that resides in the skin. Although our functional analysis was limited to bacteria, it could be extended to the other phylogenetic groups of skin commensals. However, since most of functional databases have been centered in bacteria, and then they annotations are highly biased towards them. In the case of skin, the bias is higher since there are not many datasets to compare with. Even in the case of the HMP, just a limited number of samples retrieved enough DNA to perform functional analyses, and only in one of both areas sampled(Consortium 2012). This limitation in data is associated with the limited amount of annotation obtained in our study (around 50%). Other studies working with the same information have retrieved even lower annotation rates, suggesting an even greater functional diversity associated with those samples, but not accessible due to the lack of references to perform a proper annotation (data not shown). This lack of proper annotation may be also translated to a phylogenetic level, and maybe the fraction of unassigned reads may be related to not-previously sequenced taxa. These biases should be reduced when the knowledge of skin microbiota increases(Wu et al. 2009). The HMP functional analysis of skin and our further applications of the method presented(data not shown) are examples of the effort of understanding the skin microbiota in a systemic perspective and the will of reducing the strong knowledge bias of skin microbiota.

However, one of the most striking points of this methodology is the facility to translate it to other host-microbiome scenarios. A similar situation than the skin can be applied to the gastrointestinal tract (GIT). The GIT microbiota has always been observed from a transient point of view. Given how easy is to collect stool, most of the studies carried on the GIT are based on that type of samples(Turnbaugh and Gordon 2009). The same problem occurs in the oral ecosystem, where most studies have been performed in swabs and plaque extraction, with low host cell load. In fact, the most recent paper from the Human Microbiome Project Consortium was performed using swabbing methods in 18 regions(Consortium 2012). But the bacteria in close relationship with the host cells, those in constant contact, are the most important in the progression and the maintenance of the ecosystem homoeostasis, and should be studied in more detail. Based on the enterotype classification, most of the beneficial microbes in the GIT are mucine-degraders, which implies that most of them are in close relationship with the mucosa and the gut epithelia(Wells et al. 2011). Functional differences may result in incorrect assumptions about how the gut microbiota interacts with the host cells, and how this relationship results beneficial for the host. Given that GIT and oral mucosa are epithelia, the methodology proposed may be used in both tissue samples.

Besides, based on the simplicity of each of the steps, the methodology can be adjusted according to the type of tissue that is being sampled, and not only in humans. The methodology may be adapted to any host-microbe interaction approach when it is important to observe both sides of the interaction. Until now, tissue manipulation only included a whole DNA/RNA/protein extraction and further bioinformatic separation. The method we have proposed, allows not only an unbiased separation of host and microbe cells, but also to work with both sides of the story from the same sample. Depending on the type of sampled tissue, one could adjust the enzymatic disruption to obtain a complete cell suspension, allowing to apply the filtration and cytometry steps according to its needs. Although all the previous ideas are only suggestions, we think that the method can be a strong advance in the field of metagenomics and all its variants. Working with deep epithelial biopsies would shed light into the host-microbiota relationships in any conditions, pushing further the studies of host-microbiota interactions and its relationship in health and complex diseases. Multiple studies have compared the microbial communities between healthy individuals and those with inflammatory bowel disease, retrieving significant differences between both environmental scenarios(Frank et al. 2007). Some of these differences have been shown to be associated with differences in TLR signaling; for instance, a lack of MyD88 signaling in some mouse strains significantly changes the composition of the gut microbiota, predisposing to disease(Larsson et al. 2012). This example shows the close relationship between host gut cells and microbial composition, but does not show any metagenomic functional profile, due to the inability to separate host cells from the mucosa-associated microbiota. Performing this kind of separation would allow to discriminate host and microbiota transcriptions, integrating the complete information to find the functional differences between both scenarios(Raes et al. 2011) and analyze the possible environmental variables that are affecting as cause for the ecosystems change (in this example, the disease), or are a consequence of that change.

Similar approaches could be applied to skin, to assess whether similar immune profiles occur on that tissue. Up to date, however, there are a very few examples relating relationships between skin cells and microbiota. The limitation in the ecological study of the skin results clear comparing the relative abundance of literature of the different human niches(Leydcn et al. 1987; Blauvelt 2001; Chiller et al. 2001; Cogen et al. 2008). While most of the literature focused on skin is based on classical methods of pathogen-disease relationship and other microbe-related physiological host alterations(Blauvelt 2001; Jenson et al. 2001; Casas et al.

2012), a high number of studies have been performed in GIT and oral mucosa relating shifts in microbiota with disease(Willing et al. 2010; Belda-Ferre et al. 2011; Docktor et al. 2011). The work presented in chapter three of this thesis shows an example of bacterial shift given an unrelated genetic factor, under a highly controlled environment. The skin ecological variability depends of a wide number of variables, as happens with any other ecosystem. All those variables may be considered dimensions in the ecological space and their variation may result in a strong physiological change in both the skin and the microbiota. To analyze a certain variable, one must isolate it and reduce the possible systemic variability to discern the true involvement of that variable on the whole system. Our research in chapter four concludes that, in the case of the relationship between health and microbiota, it is important to homogenize as much as possible the rest of the variables.

On the analysis of host variables and their relationship with microbiota

Although the reader may consider that the results presented in chapter four are unrelated to skin, skin and the GIT are not that different from a systems perspective. The reason why we chose gut as a target for this work was based on data availability. As we explained before, the only skin whole (meta)genome shotgun dataset available comes from the HMP and it is quite limited compared to other human-related niches(Consortium 2012). However, HMP host information is not available to date, which makes impossible to perform an analysis of meta-data like the one we have performed in GIT datasets. In contrast, MetaHIT consortium made available all the meta-information of their sampled individuals for us, allowing us to perform this kind of analysis, and assess whether health was truly related somehow to the microbiota(Members 2010).

As occurs with population stratification in population genetics, we can observe a dilution of the microbiota-host relationship signal in disease by ecological homogeneity in our subgroups(Wigginton et al. 2005). Our work in population stratification has shown an important deviation of the phylogenetic diversity depending on the health index defined in our study. We observed a strong correlation between the alpha diversity and the health state of our individuals. This result was also observed in mice and human when compared lean versus obese individuals(Turnbaugh et al. 2006; 2009) and diabetes(Qin et al. 2012). Our analysis, in agreement with the studies of Turnbaugh and Qin, suggests that gut microbiota may be strongly affected differently by two fronts: The fat mass percentage(Ley et al. 2008; Hildebrandt et al. 2009) and the host immune system(Tlaskalová-Hogenová et al. 2004; Round and Mazmanian 2009). Although the cited literature compared highly different environments, our approach has reported consistent results, even with closely related subpopulations. The relationship between host health and microbiota shows similar patterns than those observed in macro-ecological successions in different contamination states(Elliott et al. 2010). Relative abundances of the different taxonomic groups are inversely correlated one to each other as the health score increases, but they do not correlate with the enterotype classification proposed by Arumugam, Raes and collaborators(Arumugam et al. 2011), and contrary to the expectations, they can be considered as a continuous distribution(Jeffery et al. 2012). Still discreet enterotype classification has been observed in other subpopulations and species, and then they might be taken into consideration(Moeller et al. 2012; Hildebrand et al. 2013). The fact that our health classification is not associated with the enterotype classification may be related with the methodology used to construct our health score, which may behave better as a continuous distribution, or maybe, the enterotype classification correlates better with other variables not considered in this analysis, such as diet or geographic origin. This discreet-continuous dichotomy of the gut microbiota, also extended to other host-associated microbiomes, is now an important topic in host-associated ecology. Obviously, it is in human nature to classify and categorize, because it helps us to understand complex situations like the one that we are studying. However, labeling and categorizing also leads us to simplified perspectives that may not be accurate at all. The necessity of classifying our discoveries has led us to a critical point in microbiology. NG sequencing projects have produced a large amount of 16S rRNA sequences previously unknown, that need to be classified. Still, when we analyze the differences between all the novel sequences retrieved, they follow a continuous distribution, which makes hard to classify them. Researchers define thresholds on something that does not follow a discreet classification, and does not even follow the classical definition of taxon(Rocha 2008; Treangen and Rocha 2011). The same problem exists in the ecological definition of the host-tissue microbiota(Jeffery et al. 2012). In agreement to previous works in gut microbiota, more than 90% of gut bacteria are members of Bacteroidetes and Firmicutes, but the relative proportions of those follow a continuous gradient within our populations. Those distributions are associated with our health continuous classification.

Our observations suggest that Bacteroidetes predominance is correlated with low fat and low local inflammation, as was first presented by Peter J. Turnbaugh and Ruth E. Ley, on their respective works(Ley et al. 2005; Turnbaugh et al. 2009). Together, our observations also report a lack of alteration in the Firmicutes ratio. We do not observe, however, any significance with Actinobacteria, although these authors suggest a that this phylum behaves opposite to Bacteroidetes. These conclusions have been supported by a large number of publications in different datasets suggesting that a high Bacteroidetes ratio is an important trend associated with a low low inflammatory environment(Turnbaugh et al. 2006; Sekirov et al. 2010). The association of a low Bacteroidetes/ Actinobacteria ratio seems to be more linked to the sub-clinic inflammation level than to the diet itself(Tremaroli and Bäckhed 2012). Similar observations have been made in chronic inflammatory gut diseases such as Inflammatory Bowel's Disease (IBD) or Crohn's disease(CD). The subjacent reasons for that inflammation are still unknown but they are suggested to imply a complex interaction between host, microbial and environmental factors not yet resolved(Carvalho et al. 2009; Docktor et al. 2011; Greenblum et al. 2012). Other inflammatory diseases have been suggested to be associated by a host-microbial mimicry that generates auto-inflammation(Valdimarsson et al. 2009). However, it seems clear that the immune system recognizes some potential pathogenic factors through innate immune receptors, keeping the intestinal mucosa in a state of physiological inflammation, with continuous production of tissue repair factors, antimicrobial proteins and immunoglobulins, mostly IgA, that selectively provide a protection against any pathogenic assault(Honda and Littman 2012). When those unknown factors appear, the host immune system increases its response against the commensal flora, as any other pathogens, inducing pro-inflammatory signals, unbalancing the inflammation profile to a subclinical level, resulting in damage against the host's own cells(Sansonetti 2004). In agreement, Peterson and collaborators showed that IgA suppression was associated with an increase in the pro-inflammatory signaling against commensal microbiota (Peterson et al. 2007). According to these results, polyclonal

IgA would distinguish between potential pathogenic microbes from the members of the normal flora, and their suppression would be a sufficient signal to produce a strong inflammatory response which would finally affect the host's own cells. These and other results show the critical role of sustaining the homoeostasis state of the ecosystem (in this case the gut) and how the response to the commensal microbiota may result in the same etiological damage observed in complex inflammatory processes such as Crohn's disease(Joossens et al. 2011).

Some conclusions obtained from this gut analysis can be translated to our skin study in immunodeficient mice. In our case, we constructed two environmentally equivalent, but genetically different, mouse models conferring one of them a limited lymphocyte population(Ito et al. 2002). As occurs with diet in GIT, since the skin is in continuous interaction with the external environment, one should expect an important phylogenetic variability of skin microbiota associated with differences in that environment. That is why we reduced that variability generating homogeneous conditions for all the mice. In that highly controlled environment, we observed an important interindividual variability, and a strong shift related to a difference on the relative abundances of immune sentinels on the skin(Streilein 1983; Bennett et al. 2008).

It is interesting the strong association between the immune system alteration and the prevalence of *Staphylococcus* species. *Staphylococcus* species seems to be associated with skin since early stages of the *Tetrapoda* evolution(SUN et al. 2012), although their representation is limited and not predominant in normal skin(Grice et al. 2009). However, this taxonomic group seems to benefit from many alterations of the homoeostasis. *Staphylococcus* species are the earlier colonizers of skin during the first days after birth, when the acquired immune system is not properly developed(Capone et al. 2011). Given the limited exposure to antigens in utero and the well described defects in adaptive immunity of newborns, the control and interaction with the potential commensal microbiota depends on the innate immune system(Levy 2007). The innate immune system can instruct the adaptive immune system to tolerate certain microbes but not others, depending on their infection capacity and potential(Janeway and Medzhitov 2002). The innate immune system is highly conserved across the vertebrate phylogeny, being selected to recognize conserved products of microbial metabolism of facultative and opportunistic pathogens and defend against them, without reacting to the host. However, as we have seen before, the innate immune system is highly connected to the inflammatory machinery, through TLR activation, which may result harmful to the host in a non-selective way(Masters et al. 2009). This rapid colonization by Staphylococci is reduced during the first year after birth. Their predominance declines according to the development of the acquired immune system and the recognition of *Staphylococcus sp*. as a potential pathogen. However, we ignore whether this representation decrease is caused generically through a complete recognition of the *Staphylococcus* genus or, on the contrary, *Staphylococcus epidermidis* is reduced in absolute numbers as a side effect of the acquired recognition of *Staphylococcus aureus.*

Still, the newborn maintains a predominance of Firmicutes, which have a pivotal role in denying access to potentially infectious bacteria and contribute to the establishment of cutaneous homoeostasis(Capone et al. 2011). Although this prevalence of Firmicutes is lost during adulthood, it is interesting that any rupture of the homoeostasis results in a new prevalence of *Staphylococcus sp*. on skin. Our observations, in agreement with other authors' results, suggest that adaptive immune system is required to maintain the homoeostasis in skin. When we depleted the number of lymphoid cells through the knock-out of the Prkdc^{scid} gene, the relative abundance of *Staphylococcus ssp*. increased. This homeostatic disruption can be understood, according to the known capacity of *Staphylococcus* ssp. to avoid the innate immune system(Peschel et al. 2001; Kies et al. 2003; Jin et al. 2004; Cheung et al. 2010; Lo et al. 2011). However, we have observed similar reactions in other immune system impairments and over-inflammatory panels. The relationship between *Staphylococcus ssp.* infections and atopic dermatitis (AD) patients is known for more than 20 years(Ogawa et al. 1994; Higaki et al. 1999). AD is a chronic, relapsing, intensely inflammatory skin disorder with a strong prevalence in industrialized countries, and unknown etiology(Kong et al. 2012), but considered as a complex disease probably caused by a combination of genetic and environmental factors(Cramer et al. 2010). Although *Staphylococcus* colonization has long been associated with AD, their relationship does not seem to be causative for this disease(Kong et al. 2012). Still, microbial communities seem to change as a consequence of the barrier dysfunction and the defects on the innate and adaptive immunity(Scharschmidt et al. 2009). Kong and collaborators showed that the microbial community changed dramatically through the different phases of the disease. As we observed in our meta-data analysis, the authors observed a strong fall of the alpha diversity during the AD flares, characterized by a strong overgrowth and colonization of *Staphylococcus* species, including *S. aureus* and *S.epidermidis*, followed by a strong diversity recovery after the flare, increasing the diversity even over the baseline.

Another non-homeostatic situations where the immune system seems to be altered is psoriasis(Lebwohl 2003). Psoriasis, a chronic inflammatory condition of the skin, is present in about 2% of the world population. It appears to be the result of complex etiological factors, both genetic and environmental, and it is characterized by hyperkeratosis, hyperproliferation of keratinocytes, infiltration of skin by immune cells, and angiogenesis. Although all previous examples showed a decrease in the ecological diversity, psoriatic skin displayed a greater diversity in the skin lesions, compared to normal skin. This result has been replicated twice by independent groups, confirming its validity(Gao et al. 2008; Fahlén et al. 2012). Interestingly, psoriatic skin could be considered a more harmful environment for the commensal microbiota, since many AMPs are highly up-regulated in psoriatic lesions and most of them have not only antimicrobial activities but also immune-modulatory functions, promoting a stronger, sometimes nonspecific, immune response(Braff et al. 2005b; Dombrowski and Schauber 2012; Morizane and Gallo 2012). However, this alteration in the immune system is also related with an overrepresentation of Firmicutes(Gao et al. 2008). According to the information presented before, Gao and collaborators showed a reduction in the genus *Propionibacterium sp*. arguing that this underrepresentation could be due to the presence of immunomodulatory constituents on the genome of *Propionibacterium*, that could be playing a role in signaling human keratinocytes. The authors suggest that the reduction on the relative abundance of *P.acnes* eliminates, somehow, the protection produced by this species, allowing non pathogenic *Staphylococcus ssp*. and Streptococcus sp. to colonize the psoriatic skin. However, alternatively, Lo et al. showed that *Staphylococcus aureus* pathogenic strains, can use *P.acnes* to improve their infection of skin(Lo et al. 2011). Maybe, in presence of an impaired skin situation, the commensal Firmicutes species display an anti-Staphylococci activity that affects also *Propionibacterium* sp. to reduce the possibilities of skin infection by the opportunistic *Staphylococcus aureus* strains. The inhibitory effect of *St. epidermidis* against other *Staphylococcus* species has been previously explained in chapter three(Otto et al. 2001; Vuong et al. 2003).

However, this relationship between *Staphylococcus ssp*. and barrier disruption does not seem to be associated with a structural impairment. Supporting the relationship between the immune system disruption and *Staphylococcus* colonization, Scharschmidt reported that, under a structural barrier disruption, a strong fall in the alpha diversity was observed, and it was not associated with an increase of the Firmicutes abundance, but Proteobacteria(Scharschmidt et al. 2009). Matriptase knock-out and hypomorphic mice(Line et al. 2008) displayed hyperproliferative and retention ichthyosis with impaired desquamation, hypotrichosis with brittle, thin, uneven, and sparse hair, and tooth defects. All those structural defects, which could be associated with opportunistic infections of *Staphylococcus ssp*., did not display significant changes on the relative abundance of this genus. Similarly to our observations in unhealthy individuals in the GIT, we observe that any impairment on the homoeostasis of the skin barrier is associated with a strong diversity reduction, improving the colonization of certain commensal taxa. Our results in skin diversity have shown that *Staphylococcus epidermidis* fitness increases in a low adaptive immune system situation, reducing the absolute abundances of the rest of the taxa, benefitting from the available resources on the impaired skin. This observation cannot be contrasted against any other examples of ecological alteration of skin, since our work is the only one which included quantitative data. Still, recent publications show that *Staphylococcus sp*. could take advantage of other commensal bacteria to infect the skin, reducing their fitness and then their absolute abundance(Lo et al. 2011). Further quantitative and qualitative analyses could improve our knowledge of the relationship of the different commensal bacteria and their fitness under homeostatic and non-homeostatic situations.

As we have also seen in the GIT, skin can be considered as a multidimensional system with three main types of variables, host, microbial, and environmental, which interact and lead it to a transient system equilibrium. This equilibrium state will be maintained while all variables are constant. Previous studies have shown that a microbial shift may lead to altered inflammatory states and impaired healing during diabetic wound progression(Scharschmidt et al. 2009; Grice et al. 2010). The microbial shift may also be a consequence of the altered inflammatory state of the diabetic wound region. As recently observed by Redel et al. (Redel et al. 2013), the relative and absolute abundances of Staphylococcus species were highly increased in diabetic wounds, associated with an increase in diversity, measured by 16S rRNA sequencing. Interestingly, the work presented here showed a reduction of the absolute abundances of the non-*Staphylococcus* taxa in the immunodeficient mice, compared to the healthy individuals. If we observe the different all the works discussed here, including ours, we can see that, in situations of structural impairment, diversity tends to increase, while when the immune system is altered, the relative abundances of *Staphylococcus sp*. increase. Maybe the differences observed in diabetic wounds are a combination of both situations.

The previous discussion has mainly focused on phylogenetic diversity. Under the neutral theory in community ecology, one expects that under equivalent host and environmental conditions, even when a different taxon composition is observed, the functional profile for that niche will be the same(Hubbell 2005). An example of this functional equivalence can be observed in the HMP results(Consortium 2012). Our functional analysis and correlation with health shows significant changes associated with the health score proposed, which agree with the results presented in case-control studies of GIT diseases, such as CD or IBD(Joossens et al. 2011; Li et al. 2012). However, as happened with the taxonomic conclusions, we cannot translate our results to skin, since we do not have any information about the functional variation in altered skin ecosystems. Most of the work performed in skin has been performed in terms of phylogenetic diversity, but the differences in the relative functional abundances remain unknown. Therefore, the main goal of our methodology is the possibility to characterize the functional profile of a skin metagenome. We have analyzed two different skin samples using the available databases, retrieving homogeneity at functional level for both samples, in agreement with Hubbell's hypothesis(Jensen et al. 2008; Altenhoff and Dessimoz 2009; Powell et al. 2012). Although we observe differences between both samples, those differences agree with the taxonomic abundance reduction in one of both samples. Simulation analysis performed to characterize the equivalence in functional rarefaction showed similar trends in function acquisition (data not shown), supporting that the differences between both samples are due to methodological bias, not by actual differences between both samples. According to Hubbell's hypothesis one should expect no association between function and taxonomic category. Our observations agree with that hypothesis.

Final remarks and future work

Still, I must say that our work characterizing the skin as a ecosystem is not complete. Much work remains to be done. We have set a methodology to characterize the skin microbiome from a taxonomic and functional point of view. We have defined the metabolic potential of the mouse skin microbiota, which we know is totally different than human skin microbiome(Grice et al. 2008; 2009). But we do not know the actual functionality of each phylogenetic group in the mouse and human skin ecosystems. Metatranscriptomic analysis of skin microbiota is accessible through the methodology presented in this thesis, since it is possible to maintain the transcriptomic information of the sample while performing the whole procedure(Chomczynski and Sacchi 2006). The knowledge of 'who does what' in microbial ecology of host-associated populations is crucial to understand the relationship of the microbiota in the complex systems behavior like is disease. Given the strong phylogenetic variability observed in equivalent ecosystems, and the strong functional similarities(Pérez-Cobas et al. 2012), the possibility to observe differences associated (which do not mean causative) with an environmental state, may be related to functional changes, both at broad and detailed level(Tremaroli and Bäckhed 2012).

Our future work plan involves the functional and phylogenetic analysis of human skin in health and diseased condition, and characterize the differences between both ecological states, to discern whether any variation exist between both.

I also miss, and I think I can include the rest of my colleagues, the characterization of the metabolic interactions between the host and the skin microbiota(Musso et al. 2011; Grice and Segre 2012; Loper et al. 2012). The transcriptomic analysis of the interactions between host and microbes has always been observed from a clinical perspective(Bhatty et al. 2012). However, as we have suggested through this whole work, complex inflammatory diseases may be more difficult to interpret, but the answer to understand how they work and how do we fight them may be in the ecological perspective. The research community interested on this perspective is growing. I hope the work presented here helps future research works to disentangle the ecological complexity of the skin, and its implication on health.

Conclusions

1. Although commercial platforms require high amounts of DNA to construct a NGS library, the actual limiting factor is to obtain the minimum number of single-molecule enriched beads. Using quantitative PCR to achieve this number results in a better approximation than whole-genome amplification procedures.

2. Metagenomic DNA sources can be taxonomically assigned using alignment-free methods. Based on sequence characteristics, such as CUB, G-C content or k-mer distribution, one may differentiate the taxonomic origin of a read and discriminate genuine source DNA from contaminations

3. The new method proposed allows to discriminate host and microbial DNA, without applying any functional or phylogenetic bias. This method allows to perform actual metagenomic analyses from skin biopsy samples.

4. The anatomically normal skin of immunodefficient mice is gradually colonized by *Staphylococcus sp. Staphylococcus* colonization is highly related to the immune system impairment, but not to structural alterations.

5. Microbial communities may change through remote unrelated events that are not related to pathogenic events.

6. The microbial-host relationships are not static. The skin homoeostasis is a transient equilibrium that may get disrupted by multiple different alterations. To study those alterations one must consider host-microbial relationships as being in continuous

adaptation, and then study them accordingly

7. Metabolic health should be considered as a continuous variable. Microbial community variation gradually adapts to the homeostatic state of the ecosystem. As can be observed at macro ecological level, host-microbe ecosystem presents levels of species and functional succession, adapted to the host health state.

8. According to the previous conclusion, it is hard to discern between cause and consequence of microbial relationship with disease. To study the causative effects of microbiota in health and disease, one should isolate the different effects and possible causes and then combine them to fully understand the complex relationships between host and disease.

References

Aberg, KM, Man, M-Q, Gallo, RL, *et al.* (2007). Co-Regulation and Interdependence of the Mammalian Epidermal Permeability and Antimicrobial Barriers. *J Invest Dermatol* 128: 917–25.

Aird, D, Ross, MG, Chen, W-S, *et al.* (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12: R18.

Alam, H, Sehgal, L, Kundu, ST, *et al.* (2011). Novel function of keratins 5 and 14 in proliferation and differentiation of stratified epithelial cells. *Molecular Biology of the Cell* 22: 4068–78.

Albanesi, C, Scarponi, C, Giustizieri, M, *et al.* (2005). Keratinocytes in Inflammatory Skin Diseases. *CDTIA* 4: 329–34.

Albert, TJ, Molla, MN, Muzny, DM, *et al.* (2007). Direct selection of human genomic loci by microarray hybridization. *Nat Meth* 4: 903–5.

Almeida, JS, Vinga, S (2006). Computing distribution of scale independent motifs in biological sequences. *Algorithms Mol Biol* 1: 18.

Alstrup, S, Gavoille, C, Kaplan, H (2004). Nearest common ancestors: A survey and a new algorithm for a distributed environment. *Theory of Computing Systems*.

Altenhoff, AM, Dessimoz, C (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* 5: e1000262.

Altschul, SF, Gish, W, Miller, W, *et al.* (1990). Basic local alignment search tool. *J Mol Biol* 215: 403–10.

Amann, RI, Ludwig, W, Schleifer, KH (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59: 143–69.

Anderson, JM, Stevenson, BR, Jesaitis, LA, *et al.* (1988). Characterization of ZO-1, a protein component of the tight junction from mouse liver and Madin-Darby canine kidney cells. *The Journal of Cell Biology* 106: 1141–9.

Angly, FE, Felts, B, Breitbart, M, *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol* 4: e368.

Arumugam, M, Raes, J, Pelletier, E, *et al.* (2011). Enterotypes of the human gut microbiome. *Nature* 473: 174–80.

Ash, Z, Gaujoux-Viala, C, Gossec, L, *et al.* (2012). A systematic literature review of drug therapies for the treatment of psoriatic arthritis: current evidence and meta-analysis informing the EULAR recommendations for the management of psoriatic arthritis. *Ann Rheum Dis* 71: 319–26.

Ausubel, FM, Brent, R, Kingston, RE, *et al.* (1992). Current Protocols in Molecular Biology 2.1.1–2.4.5.

Baker, GC, Smith, JJ, Cowan, DA (2003). Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods* 55: 541–55.

Barnich, N, Carvalho, FA, Glasser, A-L, *et al.* (2007). CEACAM6 acts as a receptor for adherent-invasive E. coli, supporting ileal mucosa colonization in Crohn disease. *J Clin Invest* 117: 1566–74.

Bates, JM, Mittge, E, Kuhlman, J, *et al.* (2006). Distinct signals from the microbiota promote different aspects of zebrafish gut differentiation. *Dev Biol* 297: 374–86.

Bäckhed, F, Ding, H, Wang, T, *et al.* (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proc Natl Acad Sci USA* 101: 15718–23.

Bäckhed, F, Ley, RE, Sonnenburg, JL, *et al.* (2005). Host-bacterial mutualism in the human intestine. *Science* 307: 1915–20.

Behne, MJ, Meyer, JW, Hanson, KM, *et al.* (2002). NHE1 regulates the stratum corneum permeability barrier homeostasis. Microenvironment acidification assessed with fluorescence lifetime imaging. *J Biol Chem* 277: 47399–406.

Bek-Thomsen, M, Lomholt, HB, Kilian, M (2008). Acne is not associated with yet-uncultured bacteria. *Journal of Clinical Microbiology* 46: 3355–60.

Bekpen, C, Hunn, JP, Rohde, C, *et al.* (2005). The interferon-inducible p47 (IRG) GTPases in vertebrates: loss of the cell autonomous resistance mechanism in the human lineage. *Genome Biol* 6: R92.

Bekpen, C, Marques-Bonet, T, Alkan, C, *et al.* (2009). Death and Resurrection of the Human IRGM Gene. *PLoS Genet* 5: e1000403.

Belalov, IS, Lukashev, AN (2013). Causes and implications of codon usage bias in RNA viruses. *PLoS ONE* 8: e56642.

Belda-Ferre, P, Alcaraz, LD, Cabrera-Rubio, RUL, *et al.* (2011). The oral metagenome in health and disease. *ISME J* 1–11.

Bengmark, S (1999). Gut microenvironment and immune function. *Curr Opin Clin Nutr Metab Care* 2: 83–5.

Benítez-Páez, A, Álvarez, M, Belda-Ferre, P, *et al.* (2013). Detection of Transient Bacteraemia following Dental Extractions by 16S rDNA Pyrosequencing: A Pilot Study. *PLoS ONE* 8: e57782.

Bennett, MF, Robinson, MK, Baron, ED, *et al.* (2008). Skin immune systems and inflammation: protector of the skin or promoter of aging? *J Invest Dermatol* 13: 15–9.

Benson, AK, Kelly, SA, Legge, R, *et al.* (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci USA* 107: 18933–8.

Bentley, G, Higuchi, R, Hoglund, B, *et al.* (2009). High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens* 74: 393–403.

Béjà, O, Aravind, L, Koonin, EV, *et al.* (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289: 1902–6.

Bhatty, M, Fan, R, Muir, WM, *et al.* (2012). Transcriptomic analysis of peritoneal cells in a mouse model of sepsis: confirmatory and novel results in early and late sepsis. *BMC Genomics* 13: 509.

Bik, HM, Sung, W, De Ley, P, *et al.* (2012). Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Mol Ecol* 21: 1048–59.

Binga, EK, Lasken, RS, Neufeld, JD (2008). Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J* 2: 233–41.

Blaisdell, BE (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci USA* 83: 5155–9.

Blanpain, C, Fuchs, E (2009). Epidermal homeostasis: a balancing act of stem cells in the skin. *Nat Rev Mol Cell Biol* 10: 207–17.

Blaser, MJ, Dominguez-Bello, MG, Contreras, M, *et al.* (2013). Distinct cutaneous bacterial assemblages in a sampling of South American Amerindians and US residents. *ISME J* 7: 85–95.

Blauvelt, A (2001). Skin diseases associated with human herpesvirus 6, 7, and 8 infection. *J Invest Dermatol* 6: 197–202.

Board, E (2008). Sequencing the microbial soup. *Nat Struct Mol Biol* 15: 115.

Bork, P (2005). Is there biological research beyond Systems Biology? A comparative analysis of terms. *Mol Syst Biol* 1: E1–E2.

Borradori, L, Sonnenberg, A (1999). Structure and Function of Hemidesmosomes: More Than Simple Adhesion Complexes. *J Invest Dermatol* 112: 411–8.

Bosma, GC, Custer, RP, Bosma, MJ (1983). A severe combined immunodeficiency mutation in the mouse. *Nature* 301: 527–30.

Bouwstra, JA, Gooris, GS, Bras, W, *et al.* (1995). Lipid organization in pig stratum corneum. *J Lipid Res* 36: 685–95.

Bouwstra, JA, Gooris, GS, Dubbelaar, FE, *et al.* (1998). Role of ceramide 1 in the molecular organization of the stratum corneum lipids. *J Lipid Res* 39: 186–96.

Böttger, EC (1989). Rapid determination of bacterial ribosomal RNA sequences by direct sequencing of enzymatically amplified DNA. *FEMS Microbiol Lett* 53: 171–6.

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

Braff, MH, Di Nardo, A, Gallo, RL (2005a). Keratinocytes store the antimicrobial peptide cathelicidin in lamellar bodies. *J Invest Dermatol* 124: 394–400.

Braff, MH, Zaiou, M, Fierer, J, *et al.* (2005b). Keratinocyte production of cathelicidin provides direct activity against bacterial skin pathogens. *Infect Immun* 73: 6771–81.

Bredel, M, Bredel, C, Juric, D, *et al.* (2005). Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA. *J Mol Diagn* 7: 171–82.

Breitbart, M, Salamon, P, Andresen, B, *et al.* (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 99: 14250–5.

Brüssow, H, Canchaya, C, Hardt, W-D (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* 68: 560–602–tableofcontents.

Buh Gasparic, M, Tengs, T, La Paz, JL, *et al.* (2010). Comparison of nine different real-time PCR chemistries for qualitative and quantitative applications in GMO detection. *Anal Bioanal Chem* 396: 2023–9.

Burgess, STG, Greer, A, Frew, D, *et al.* (2012). Transcriptomic analysis of circulating leukocytes reveals novel aspects of the host systemic inflammatory response to sheep scab mites. *PLoS ONE* 7: e42778.

Bustin, SA, Benes, V, Garson, JA, *et al.* (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*.

Butler, JE, Sun, J, Weber, P, *et al.* (2000). Antibody repertoire development in fetal and newborn piglets, III. Colonization of the gastrointestinal tract selectively diversifies the preimmune repertoire in mucosal lymphoid tissues. *Immunology* 100: 119–30.

Byrne, C, Tainsky, M, Fuchs, E (1994). Programming gene expression in developing epidermis. *Development* 120: 2369–83.

Candi, E, Schmidt, R, Melino, G (2005). The cornified envelope: a model of cell death in the skin. *Nat Rev Mol Cell Biol* 6: 328–40.

Capella-Gutiérrez, S, Silla-Martínez, JM, Gabaldón, T (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–3.

Capone, KA, Dowd, SE, Stamatas, GN, *et al.* (2011). Diversity of the human skin microbiome early in life. *J Invest Dermatol* 131: 2026–32.

Caporaso, JG, Kuczynski, J, Stombaugh, J, *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* 7: 335–6.

Carbone, A (2008). Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J Mol Evol* 66: 210–23.

Carvalho, FA, Aitken, JD, Vijay-Kumar, M, *et al.* (2012). Toll-Like Receptor–Gut Microbiota Interactions: Perturb at Your Own Risk! *Annu Rev Physiol* 74: 177–98.

Carvalho, FA, Barnich, N, Sivignon, A, *et al.* (2009). Crohn's disease adherent-invasive Escherichia coli colonize and induce strong gut inflammation in transgenic mice expressing human CEACAM. *J Exp Med* 206: 2179–89.

Casas, C, Paul, C, Lahfa, M, *et al.* (2012). Quantification of Demodex folliculorum by PCR in rosacea and its relationship to skin innate immune activation. *Experimental Dermatology* 21: 906–10.

Castillo, M, Martín-Orúe, SM, Manzanilla, EG, *et al.* (2006). Quantification of total bacteria, enterobacteria and lactobacilli populations in pig digesta by real-time PCR. *Vet Microbiol* 114: 165–70.

Caubet, C, Jonca, N, Brattsand, M, *et al.* (2004). Degradation of corneodesmosome proteins by two serine proteases of the kallikrein family, SCTE/KLK5/hK5 and SCCE/KLK7/hK7. *J Invest Dermatol* 122: 1235–44.

Chen, SL, Lee, W, Hottes, AK, *et al.* (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA* 101: 3480–5.

Cheung, GYC, Rigby, K, Wang, R, *et al.* (2010). Staphylococcus epidermidis strategies to avoid killing by human neutrophils. *PLoS Pathog* 6.

Chiller, K, Selkin, BA, Murakawa, GJ (2001). Skin microflora and bacterial infections of the skin. *J Invest Dermatol* 6: 170–4.

Choi, E-H, Man, M-Q, Xu, P, *et al.* (2007). Stratum corneum acidification is impaired in moderately aged human and murine skin. *J Invest Dermatol* 127: 2847–56.

Chomczynski, P, Sacchi, N (2006). The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. *Nat Protoc* 1: 581–5.

Clayton, E, Doupé, DP, Klein, AM, *et al.* (2007). A single type of progenitor cell maintains normal epidermis. *Nature* 446: 185–9.

Cogen, A, Nizet, V, Gallo, R (2008). Skin microbiota: a source of disease or defence? *British Journal of Dermatology* 158: 442–55.

Cohen, SN, Chang, AC, Boyer, HW, *et al.* (1973). Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci USA* 70: 3240–4.

Cole, JR, Chai, B, Farris, RJ, *et al.* (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Research* 35: D169–72.

Cole, JR, Wang, Q, Cardenas, E, *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* 37: D141–5.

Conesa, A, G o tz, S, Garc i a-G o mez, JMM, *et al.* (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)* 21: 3674–6.

Consortium, THMP (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486: 207–14.

Costello, EK, Lauber, CL, Hamady, M, *et al.* (2009). Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science* 326: 1694–7.

Coulombe, PA, Wong, P (2004). Cytoplasmic intermediate filaments revealed as dynamic and multipurpose scaffolds. *Nature cell biology* 6: 699–706.

Cox, MP, Peterson, DA, Biggs, PJ (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11: 485.

Cox-Foster, DL, Conlan, S, Holmes, EC, *et al.* (2007). A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* 318: 283–7.

Cramer, C, Link, E, Horster, M, *et al.* (2010). Elder siblings enhance the effect of filaggrin mutations on childhood eczema: Results from the 2 birth cohort studies LISAplus and GINIplus. *Journal of Allergy and Clinical Immunology* 125: 1254–5.

Dayhoff, MO, Schwartz, RM, Orcutt, BC (1978). A model of evolutionary change in proteins. *In Atlas of protein sequence and structure*.

de Guzman Strong, C, Wertz, PW, Wang, C, *et al.* (2006). Lipid defect underlies selective skin barrier impairment of an epidermal-specific deletion of Gata-3. *The Journal of Cell Biology* 661–70.

de Hoon, MJL, Eichenberger, P, Vitkup, D (2010). Hierarchical evolution of the bacterial sporulation network. *Curr Biol* 20: R735–45.

de Muinck, EJ, Lagesen, K, Afset, JE, *et al.* (2013). Comparisons of infant Escherichia coli isolates link genomic profiles with adaptation to the ecological niche. *BMC Genomics* 14: 81.

de Nadal, E, Ammerer, G, Posas, F (2011). Controlling gene expression in response to stress. *Nature Rev Genet* 12: 833–45.

Dean, FB, Hosono, S, Fang, L, *et al.* (2002a). Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* 99: 5261–6.

Dean, FB, Nelson, JR, Giesler, TL, *et al.* (2001). Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. *Genome Research* 11: 1095–9.

Dean, M, Carrington, M, O'Brien, SJ (2002b). Balanced polymorphism selected by genetic versus infectious human disease. *Annu Rev Genomics Hum Genet* 3: 263–92.

Dekio, I, Sakamoto, M, Hayashi, H, *et al.* (2007). Characterization of skin microbiota in patients with atopic dermatitis and in normal subjects using 16S rRNA gene-based comprehensive analysis 56: 1675–83.

Deleon-Rodriguez, N, Lathem, TL, Rodriguez-R, LM, *et al.* (2013). Microbiome of the upper troposphere: Species composition and prevalence, effects of tropical storms, and atmospheric implications. *Proceedings of the National Academy of Sciences* 110: 2575–80.

Dessinioti, C, Katsambas, AD (2010). The role of Propionibacterium acnes in acne pathogenesis: facts and controversies. *Clin Dermatol* 28: 2–7.

Dethlefsen, L, Huse, S, Sogin, ML, *et al.* (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* 6: e280.

Dethlefsen, L, McFall-Ngai, M, Relman, DA (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 449: 811–8.

Dethlefsen, L, Relman, DA (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc Natl Acad Sci USA* 108 Suppl 1: 4554–61.

Docktor, MJ, Paster, BJ, Abramowicz, S, *et al.* (2011). Alterations in diversity of the oral microbiome in pediatric inflammatory bowel disease. *Inflamm Bowel Dis*.

Dombrowski, Y, Schauber, J (2012). Cathelicidin LL-37: a defense molecule with a potential role in psoriasis pathogenesis. *Experimental Dermatology* 21: 327–30.

Dragosits, M, Mozhayskiy, V, Quinones-Soto, S, *et al.* (2013). Evolutionary potential, cross-stress behavior and the genetic basis of acquired stress resistance in Escherichia coli. *Mol Syst Biol* 9: 643.

Dupont, CL, Rusch, DB, Yooseph, S, *et al.* (2011). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6: 1186–99.

Eckburg, PB, Bik, EM, Bernstein, CN, *et al.* (2005). Diversity of the human intestinal microbial flora. *Science* 308: 1635–8.

Eggesbø, M, Moen, B, Peddada, S, *et al.* (2011). Development of gut microbiota in infants not exposed to medical interventions. *APMIS* 119: 17–35.

Ehrchen, JM, Roebrock, K, Foell, D, *et al.* (2010). Keratinocytes Determine Th1 Immunity during Early Experimental Leishmaniasis. *PLoS Pathog* 6: e1000871.

Elde, NC, Malik, HS (2009). The evolutionary conundrum of pathogen mimicry. *Nat Rev Microbiol* 7: 787–97.

Elias, PM (2007). The skin barrier as an innate immune element. *Semin Immunopathol* 29: 3–14.

Elliott, DR, Scholes, JD, Thornton, SF, *et al.* (2010). Dynamic changes in microbial community structure and function in phenol-degrading microcosms inoculated with cells from a contaminated aquifer. *FEMS Microbiology Ecology* 71: 247–59.

Erickson, AR, Cantarel, BL, Lamendella, R, *et al.* (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE* 7: e49138.

Erlich, Y, Mitra, PP, delaBastide, M, *et al.* (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Meth* 5: 679–82.

Esteban, JA, Salas, M, Blanco, L (1993). Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem* 268: 2719–26.

Fahlén, A, Engstrand, L, Baker, BS, *et al.* (2012). Comparison of bacterial microbiota in skin biopsies from normal and psoriatic skin. *Arch Dermatol Res* 304: 15–22.

Farquhar, MG, Palade, GE (1963). Junctional complexes in various epithelia. *The Journal of Cell Biology* 17: 375–412.

Fierer, N, Hamady, M, Lauber, CL, *et al.* (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci USA* 105: 17994–9.

Fierer, N, Lauber, CL, Zhou, N, *et al.* (2010). From the Cover: Forensic identification using skin bacterial communities. *Proc Natl Acad Sci USA* 107: 6477–81.

Fischer, H, Scherz, J, Szabo, S, *et al.* (2011). DNase 2 is the main DNA-degrading enzyme of the stratum corneum. *PLoS ONE* 6: e17581.

Fitz-Gibbon, S, Tomida, S, Chiu, B-H, *et al.* (2013). Propionibacterium acnes Strain Populations in the Human Skin Microbiome Associated with Acne. *J Invest Dermatol*.

Fluhr, JW, Kao, J, Jain, M, *et al.* (2001). Generation of free fatty acids from phospholipids regulates stratum corneum acidification and integrity. *J Invest Dermatol* 117: 44–51.

Ford, JL, Kaakoush, NO, Mendz, GL (2010). Phosphonate metabolism in Helicobacter pylori. *Antonie Van Leeuwenhoek* 97: 51–60.

Frank, DN, St Amand, AL, Feldman, RA, *et al.* (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* 104: 13780–5.

Franzmann, TM, Menhorn, P, Walter, S, *et al.* (2008). Activation of the chaperone Hsp26 is controlled by the rearrangement of its thermosensor domain. *Mol Cell* 29: 207–16.

Fredricks, DN (2001). Microbial ecology of human skin in health and disease. *J Invest Dermatol* 6: 167–9.

Fuchs, E (1995). Keratins and the skin. *Annu Rev Cell Dev Biol* 11: 123–53.

Fuchs, E (2008). Skin stem cells: rising to the surface. *The Journal of Cell Biology* 180: 273–84.

Fuchs, E (2009). Finding One's Niche in the Skin. *Cell Stem Cell* 4: 499–502.

Fuchs, E, Raghavan, S (2002). Getting under the skin of epidermal morphogenesis. *Nature Rev Genet* 3: 199–209.

Galdbart, J-O, Allignet, J, Tung, H-S, *et al.* (2000). Screening for Staphylococcus epidermidis markers discriminating between skin-flora strains and those responsible for infections of joint prostheses. *J Infect Dis* 182: 351–5.

Gao, Z, Perez-Perez, GI, Chen, Y, *et al.* (2010). Quantitation of major human cutaneous bacterial and fungal populations. *Journal of Clinical Microbiology* 48: 3575–81.

Gao, Z, Tseng, C-H, Pei, Z, *et al.* (2007). Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci USA* 104: 2927–32.

Gao, Z, Tseng, C-H, Strober, BE, *et al.* (2008). Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. *PLoS ONE* 3: e2719.

Garber, K (2008). Fixing the front end. *Nat Biotechnol* 26: 1101–4.

Garcia-Garcerà, M, Coscollá, M, Garcia-Etxebarria, K, *et al.* (2012). Staphylococcus prevails in the skin microbiota of long-term immunodeficient mice. *Environmental Microbiology*.

Garcia-Garcerà, M, Gigli, E, Sanchez-Quinto, F, *et al.* (2011). Fragmentation of Contaminant and Endogenous DNA in Ancient Samples Determined by Shotgun Sequencing; Prospects for Human Palaeogenomics. *PLoS ONE* 6: e24161.

Garcia-Russell, N, Elrod, B, Dominguez, K (2009). Stress-induced prophage DNA replication in Salmonella enterica serovar Typhimurium. *Infect Genet Evol* 9: 889–95.

Gill, SR, Pop, M, Deboy, RT, *et al.* (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312: 1355–9.

Gilliet, M, Cao, W, Liu, Y-J (2008). Plasmacytoid dendritic cells: sensing nucleic acids in viral infection and autoimmune diseases. *Nature Rev Immunol* 8: 594–606.

Gomez-Alvarez, V, Teal, TK, Schmidt, TM (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3: 1314–7.

Gosalbes, MJ, Durbán, A, Pignatelli, M, *et al.* (2011). Metatranscriptomic Approach to Analyze the Functional Human Gut Microbiota. *PLoS ONE* 6: e17447.

Gotelli, NJ, Colwell, RK (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology letters* 4: 379–91.

Gouy, M, Gautier, C (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research* 10: 7055–74.

Greenblum, S, Turnbaugh, PJ, Borenstein, E (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci USA* 109: 594–9.

Grice, EA, Kong, HH, Conlan, S, *et al.* (2009). Topographical and temporal diversity of the human skin microbiome. *Science* 324: 1190–2.

Grice, EA, Kong, HH, Renaud, G, *et al.* (2008). A diversity profile of the human skin microbiota. *Genome Research* 18: 1043–50.

Grice, EA, Segre, JA (2011). The skin microbiome. *Nat Rev Microbiol* 9: 244–53.

Grice, EA, Segre, JA (2012). Interaction of the microbiome with the innate immune response in chronic wounds. *Adv Exp Med Biol* 946: 55–68.

Grice, EA, Snitkin, ES, Yockey, LJ, *et al.* (2010). Longitudinal shift in diabetic wound microbiota correlates with prolonged skin defense response. *Proc Natl Acad Sci USA* 107: 14799–804.

Grubauer, G, Feingold, KR, Harris, RM, *et al.* (1989). Lipid content and lipid type as determinants of the epidermal permeability barrier. *J Lipid Res* 30: 89–96.

Gunathilake, R, Schurer, NY, Shoo, BA, *et al.* (2009). pH-regulated mechanisms account for pigment-type differences in epidermal barrier function. *J Invest Dermatol* 129: 1719–29.

Haas, BJ, Gevers, D, Earl, AM, *et al.* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* 21: 494–504.

Hachem, J-P, Behne, MJ, Aronchik, I, *et al.* (2005). Extracellular pH Controls NHE1 expression in epidermis and keratinocytes: implications for barrier repair. *J Invest Dermatol* 125: 790–7.

Hachem, J-P, Crumrine, D, Fluhr, J, *et al.* (2003). pH directly regulates epidermal permeability barrier homeostasis, and stratum corneum integrity/cohesion. *J Invest Dermatol* 121: 345–53.

Harding, CR, Scott, IR (n.d.). *Skin Moisturization*. Marcel Dekker: New York.

Hausser, J, Strimmer, K (2012). *entropy: Entropy and Mutual Information Estimation*.

Hershberg, R, Petrov, DA (2008). Selection on codon bias. *Annu Rev Genet* 42: 287–99.

Higaki, S, Morohashi, M, Yamagishi, T, *et al.* (1999). Comparative study of staphylococci from the skin of atopic dermatitis patients and from healthy subjects. *Int J Dermatol* 38: 265–9.

Higuchi, R, Fockler, C, Dollinger, G, *et al.* (1993). Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Nat Biotechnol* 11: 1026–30.

Hildebrand, F, Nguyen, ATL, Brinkman, B, *et al.* (2013). Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol* 14: R4.

Hildebrandt, MA, Hoffmann, C, Sherrill Mix, SA, *et al.* (2009). High-Fat Diet Determines the Composition of the Murine Gut Microbiome Independently of Obesity. *Gastroenterology* 137: 1716–1724.e2.

Hodges, E, Xuan, Z, Balija, V, *et al.* (2007). Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 39: 1522–7.

Honda, K, Littman, DR (2012). The Microbiome in Infectious Disease and Inflammation. *Annu Rev Immunol* 30: 759–95.

Hooper, LV, Midtvedt, T, Gordon, JI (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu Rev Nutr* 22: 283–307.

Horton, R, Wilming, L, Rand, V, *et al.* (2004). Gene map of the extended human MHC. *Nature Rev Genet* 5: 889–99.

Hosono, S, Faruqi, AF, Dean, FB, *et al.* (2003). Unbiased Whole-Genome Amplification Directly From Clinical Samples. *Genome Research* 13: 954–64.

Houben, E, De Paepe, K, Rogiers, V (2007). A keratinocyte's course of life. *Skin Pharmacol Physiol* 20: 122–32.

Huang, J, Zheng, Z, Andersson, AF, *et al.* (2011). Rapid screening of complex DNA samples by single-molecule amplification and sequencing. *PLoS ONE* 6: e19723–.

Hubbell, SP (2005). Neutral theory in community ecology and the hypothesis of functional equivalence. *Functional ecology* 19: 166–72.

Huerta-Cepas, J, Dopazo, J, Gabaldón, T (2010). ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* 11: 24.

Hugenholtz, P, Pace, NR (1996). Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends Biotechnol* 14: 190–7.

Human Microbiome Jumpstart Reference Strains Consortium, Nelson, KE, Weinstock, GM, *et al.* (2010). A catalog of reference genomes from the human microbiome. *Science* 328: 994–9.

Hutchison, CA, Venter, JC (2006). Single-cell genomics. *Nat Biotechnol* 24: 657–8.

Ikaha, R, Gentleman, R (1996). A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5: 299–314.

Ishiko, A, Matsunaga, Y, Masunaga, T, *et al.* (2003). Immunomolecular mapping of adherens junction and desmosomal components in normal human epidermis. *Experimental Dermatology* 12: 747–54.

Ito, M, Hiramatsu, H, Kobayashi, K, *et al.* (2002). NOD/SCID/gamma(c)(null) mouse: an excellent recipient mouse model for engraftment of human cells. *Blood* 100: 3175–82.

Ivanov, II, Atarashi, K, Manel, N, *et al.* (2009). Induction of Intestinal Th17 Cells by Segmented Filamentous Bacteria. *Cell* 1–14.

Iwamoto, K, Bundo, M, Ueda, J, *et al.* (2007). Detection of Chromosomal Structural Alterations in Single Cells by SNP Arrays: A Systematic Survey of Amplification Bias and Optimized Workflow. *PLoS ONE* 2: e1306–.

Janeway, CA (1989). Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb Symp Quant Biol* 54 Pt 1: 1–13.

Janeway, CA, Medzhitov, R (2002). Innate immune recognition. *Annu Rev Immunol* 20: 197–216.

Jean-Baptiste, N, Benjamin, DK, Cohen-Wolkowiez, M, *et al.* (2011). Coagulase-negative staphylococcal infections in the neonatal intensive care unit. *Infect Control Hosp Epidemiol* 32: 679–86.

Jeffery, IB, Claesson, MJ, O'Toole, PW, *et al.* (2012). Categorization of the gut microbiota: enterotypes or gradients? *Nat Rev Microbiol* 10: 591–2.

Jeffrey, HJ (1990). Chaos game representation of gene structure. *Nucleic Acids Research* 18: 2163–70.

Jensen, LJ, Julien, P, Kuhn, M, *et al.* (2008). eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research* 36: D250–4.

Jenson, AB, Geyer, S, Sundberg, JP, *et al.* (2001). Human papillomavirus and skin cancer. *J Invest Dermatol* 6: 203–6.

Jiang, Z, Zhang, X, Deka, R, *et al.* (2005). Genome amplification of single sperm using multiple displacement amplification. *Nucleic Acids Research* 33: e91.

Jin, T, Bokarewa, M, Foster, T, *et al.* (2004). Staphylococcus aureus resists human defensins by production of staphylokinase, a novel bacterial evasion mechanism. *J Immunol* 172: 1169–76.

Johnson, ME, Viggiano, L, Bailey, JA, *et al.* (2001). Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413: 514–9.

Joossens, M, Huys, G, Cnockaert, M, *et al.* (2011). Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* 60: 631–7.

Kalinin, AE, Kajava, AV, Steinert, PM (2002). Epithelial barrier function: assembly and structural features of the cornified cell envelope. *Bioessays* 24: 789–800.

Kanagawa, T (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* 96: 317–23.

Keane, TM, Goodstadt, L, Danecek, P, *et al.* (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477: 289–94.

Kies, S, Vuong, C, Hille, M, *et al.* (2003). Control of antimicrobial peptide synthesis by the agr quorum sensing system in Staphylococcus epidermidis: activity of the lantibiotic epidermin is regulated at the level of precursor peptide processing. *Peptides* 24: 329–38.

Kim, BE, Leung, DY (2012). Epidermal barrier in atopic dermatitis. *Allergy Asthma Immunol Res* 4: 12–6.

Kim, BE, Leung, DYM, Boguniewicz, M, *et al.* (2008). Loricrin and involucrin expression is down-regulated by Th2 cytokines through STAT-6. *Clin Immunol* 126: 332–7.

Kim, K-H, Bae, J-W (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Applied and Environmental Microbiology* 77: 7663–8.

Kim, M, Morrison, M, Yu, Z (2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods* 84: 81–7.

Kleerebezem, M, Vaughan, EE (2009). Probiotic and gut lactobacilli and bifidobacteria: molecular approaches to study diversity and activity. *Annu Rev Microbiol* 63: 269–90.

Klijn, A, Mercenier, A, Arigoni, F (2005). Lessons from the genomes of bifidobacteria. *FEMS Microbiology Reviews* 29: 491–509.

Kolbe, DL, Eddy, SR (2009). Local RNA structure alignment with incomplete sequence. *Bioinformatics* 25: 1236–43.

Kong, HH, Oh, J, Deming, C, *et al.* (2012). Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Research*.

Kong, HH, Segre, JA (2011). Skin Microbiome: Looking Back to Move Forward. *J Invest Dermatol* 132: 933–9.

Koster, J, Borradori, L, Sonnenberg, A (2004). Hemidesmosomes: Molecular Organization and Their Importance for Cell Adhesion and Disease. *In*: *Handbook of Experimental Pharmacology*, Handbook of Experimental Pharmacology. Springer Berlin Heidelberg: Berlin, Heidelberg, 243–80.

Köckritz-Blickwede, von, M, Rohde, M, Oehmcke, S, *et al.* (2008). Immunological mechanisms underlying the genetic predisposition to severe Staphylococcus aureus infection in the mouse model. *Am J Pathol* 173: 1657–68.

Kuechle, MK, Presland, RB, Lewis, SP, *et al.* (2000). Inducible expression of filaggrin increases keratinocyte susceptibility to apoptotic cell death. *Cell Death Differ* 7: 566–73.

Kupper, TS, Chua, AO, Flood, P, *et al.* (1987). Interleukin 1 gene expression in cultured human keratinocytes is augmented by ultraviolet irradiation. *J Clin Invest* 80: 430–6.

Kupper, TS, Fuhlbrigge, RC (2004). Immune surveillance in the skin: mechanisms and clinical consequences. *Nature Rev Immunol* 4: 211–22.

Kupper, TS, Groves, RW (1995). The interleukin-1 axis and cutaneous inflammation. *J Invest Dermatol* 105: 62S–66S.

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

Kutyavin, IV, Afonina, IA, Mills, A, *et al.* (2000). 3'-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Research* 28: 655–61.

Lage, JM, Leamon, JH, Pejovic, T, *et al.* (2003). Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome Research* 13: 294–307.

Lai, Y, Cogen, AL, Radek, KA, *et al.* (2010). Activation of TLR2 by a small molecule produced by Staphylococcus epidermidis increases antimicrobial defense against bacterial skin infections. *J Invest Dermatol* 130: 2211–21.

Lande, R, Gregorio, J, Facchinetti, V, *et al.* (2007). Plasmacytoid dendritic cells sense self-DNA coupled with antimicrobial peptide. *Nature* 449: 564–9.

Langbein, L, Eckhart, L, Rogers, MA, *et al.* (2010). Against the rules: human keratin K80: two functional alternative splice variants, K80 and K80.1, with special cellular localization in a wide range of epithelia. *J Biol Chem* 285: 36909–21.

Larsson, E, Tremaroli, V, Lee, YS, *et al.* (2012). Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut* 61: 1124–31.

Lartillot, N, Philippe, H (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21: 1095–109.

Lasken, RS (2007). Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr Opin Microbiol* 10: 510–6.

Lasken, RS, Stockwell, TB (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC biotechnology* 7: 19–9.

Le Caignec, C, Spits, C, Sermon, K, *et al.* (2006). Single-cell chromosomal imbalances detection by array CGH. *Nucleic Acids Research* 34: e68.

Lebre, MC, Van Der Aar, A, Van Baarsen, L (2006). Human keratinocytes express functional Toll-like receptor 3, 4, 5, and 9. *Dermatology*.

Lebwohl, M (2003). Psoriasis. *Lancet* 361: 1197–204.

Leigh, IM, Navsaria, H, Purkis, PE, *et al.* (1995). Keratins (K16 and K17) as markers of keratinocyte hyperproliferation in psoriasis in vivo and in vitro. *Br J Dermatol* 133: 501–11.

Lemon, KP, Klepac-Ceraj, V, Schiffer, HK, *et al.* (2010). Comparative analyses of the bacterial microbiota of the human nostril and oropharynx. *MBio* 1.

Levy, O (2007). Innate immunity of the newborn: basic mechanisms and clinical correlates. *Nature Rev Immunol* 7: 379–90.

Ley, RE, Bäckhed, F, Turnbaugh, P, *et al.* (2005). Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* 102: 11070–5.

Ley, RE, Hamady, M, Lozupone, C, *et al.* (2008). Evolution of mammals and their gut microbes. *Science* 320: 1647–51.

Ley, RE, Peterson, DA, Gordon, JI (2006). Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. *Cell* 124: 837–48.

Leydcn, JJ, McGinley, KJ, Nordstrom, KM, *et al.* (1987). Skin microflora. *J Invest Dermatol* 88: 65–72.

Leyden, JJ, McGinley, KJ, Hölzle, E, *et al.* (1981). The microbiology of the human axilla and its relationship to axillary odor. *J Invest Dermatol* 77: 413–6.

Li, E, Hamm, CM, Gulati, AS, *et al.* (2012). Inflammatory bowel diseases phenotype, C. difficile and NOD2 genotype are associated with shifts in human ileum associated microbial composition. *PLoS ONE* 7: e26284.

Li, H, Handsaker, B, Wysoker, A, *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–9.

Li, M, Cha, DJ, Lai, Y, *et al.* (2007). The antimicrobial peptide-sensing system aps of Staphylococcus aureus. *Mol Microbiol* 66: 1136–47.

Li, W, Godzik, A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–9.

Lilliefors, H (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*.

Lin, T-K, Crumrine, D, Ackerman, LD, *et al.* (2012). Cellular changes that accompany shedding of human corneocytes. *J Invest Dermatol* 132: 2430–9.

Lind, MH, Rozell, B, Wallin, RPA, *et al.* (2004). Tumor necrosis factor receptor 1-mediated signaling is required for skin cancer development induced by NF-kappaB inhibition. *Proc Natl Acad Sci USA* 101: 4972–7.

Line, JE, Svetoch, EA, Eruslanov, BV, *et al.* (2008). Isolation and purification of enterocin E-760 with broad antimicrobial activity against gram-positive and gram-negative bacteria. *Antimicrob Agents Chemother* 52: 1094–100.

Lippens, S, Denecker, G, Ovaere, P, *et al.* (2005). Death penalty for keratinocytes: apoptosis versus cornification. *Cell Death Differ* 12 Suppl 2: 1497–508.

Lo, C-W, Lai, Y-K, Liu, Y-T, *et al.* (2011). Staphylococcus aureus hijacks a skin commensal to intensify its virulence: immunization targeting β-hemolysin and CAMP factor. *J Invest Dermatol* 131: 401–9.

Loper, JE, Hassan, KA, Mavrodi, DV, *et al.* (2012). Comparative genomics of plant-associated Pseudomonas spp.: insights into diversity and inheritance of traits involved in multitrophic interactions. *PLoS Genet* 8.

López-Bueno, A, Tamames, J, Velázquez, D, *et al.* (2009). High diversity of the viral community from an Antarctic lake. *Science* 326: 858–61.

Lucks, JB, Nelson, DR, Kudla, GR, *et al.* (2008). Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol* 4: e1000001.

Lugo-Janer, G, Sánchez, JL, Santiago-Delpin, E (1991). Prevalence and clinical spectrum of skin diseases in kidney transplant recipients. *J Am Acad Dermatol* 24: 410–4.

Madera, M, Gough, J (2002). A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Research* 30: 4321–8.

Maidak, BL, Cole, JR, Lilburn, TG, *et al.* (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Research* 29: 173–4.

Mao-Qiang, M, Feingold, KR, Jain, M, *et al.* (1995). Extracellular processing of phospholipids is required for permeability barrier homeostasis. *J Lipid Res*.

Marchiando, AM, Graham, WV, Turner, JR (2010). Epithelial Barriers in Homeostasis and Disease. *Annu Rev Pathol Mech Dis* 5: 119–44.

Marcy, Y, Ouverney, C, Bik, EM, *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 USA* 104: 11889–94.

Mardis, ER (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9: 387–402.

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

Marinelli, LJ, Fitz-Gibbon, S, Hayes, C, *et al.* (2012). Propionibacterium acnes bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *MBio* 3.

Marples, MJ (1969). Life on the human skin. *Sci Am* 220: 108–15.

Marples, RR, Leyden, JJ, Stewart, RN, *et al.* (1974). The skin microflora in acne vulgaris. *J Invest Dermatol* 62: 37–41.

Martineau, F, Picard, FJ, Ke, D, *et al.* (2001). Development of a PCR assay for identification of staphylococci at genus and species levels. *Journal of Clinical Microbiology* 39: 2541–7.

Masahiko Itoh, ANSMST (1997). Involvement of ZO-1 in Cadherin-based Cell Adhesion through Its Direct Binding to a Catenin and Actin Filaments. *The Journal of Cell Biology* 138: 181.

Masters, SL, Simon, A, Aksentijevich, I, *et al.* (2009). Horror autoinflammaticus: the molecular pathophysiology of autoinflammatory disease (*). *Annu Rev Immunol* 27: 621–68.

Matoltsy, AG (1975). Desmosomes, filaments, and keratohyaline granules: their role in the stabilization and keratinization of the epidermis. *J Invest Dermatol* 65: 127–42.

Mazmanian, SK, Liu, CH, Tzianabos, AO, *et al.* (2005). An Immunomodulatory Molecule of Symbiotic Bacteria Directs Maturation of the Host Immune System. *Cell* 122: 107–18.

Mazmanian, SK, Round, JL, Kasper, DL (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature* 453: 620–5.

McBride, ME, Duncan, WC, Knox, JM (1977). The environment and the microbial ecology of human skin. *Applied and Environmental Microbiology* 33: 603–8.

Meadow, JF, Bateman, AC, Herkert, KM, *et al.* (2013). Significant changes in the skin microbiome mediated by the sport of roller derby. *PeerJ*.

Medzhitov, R (2009). Approaching the Asymptote: 20 Years Later. *Immunity* 30: 766–75.

Mehta, P, Goyal, S, Long, T, *et al.* (2009). Information processing and signal integration in bacterial quorum sensing. *Mol Syst Biol* 5: 325.

Members, MCA (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464: 59–65.

Metzker, ML (2009). Sequencing technologies — the next generation. *Nature Rev Genet* 11: 31–46.

Meyer, F, Paarmann, D, D'Souza, M, *et al.* (2008a). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386.

Meyer, M, Briggs, AW, Maricic, T, *et al.* (2008b). From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Research* 36: e5.

Miller, LS, Modlin, RL (2007). Human Keratinocyte Toll-like Receptors Promote Distinct Immune Responses. *J Invest Dermatol* 127: 262–3.

Miller, SJ, Aly, R, Shinefeld, HR, *et al.* (1988). In vitro and in vivo antistaphylococcal activity of human stratum corneum lipids. *Arch Dermatol* 124: 209–15.

Moeller, AH, Degnan, PH, Pusey, AE, *et al.* (2012). Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nat Commun* 3: 1179.

Mommers, JM, van Rossum, MMM, van Erp, PEP, *et al.* (2000). Changes in keratin 6 and keratin 10 (co-)expression in lesional and symptomless skin of spreading psoriasis. *Dermatology* 201: 15.

Monod, M, Capoccia, S, Léchenne, B, *et al.* (2002). Secreted proteases from pathogenic fungi. *International Journal of Medical Microbiology* 292: 405–19.

Morgan, M, Anders, S, Lawrence, M, *et al.* (2009). ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 25: 2607–8.

Morgan, XC, Segata, N, Huttenhower, C (2012). Biodiversity and functional genomics in the human microbiome. *Trends Genet* 1–8.

Morizane, S, Gallo, RL (2012). Antimicrobial peptides in the pathogenesis of psoriasis. *The Journal of Dermatology* 39: 225–30.

Morot-Bizot, SC, Talon, R, Leroy, S (2004). Development of a multiplex PCR for the identification of Staphylococcus genus and four staphylococcal species isolated from food. *Journal of Applied Microbiology* 97: 1087–94.

Mouse Genome Sequencing Consortium, Waterston, RH, Lindblad-Toh, K, *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–62.

Moya, A, Peretó, J, Gil, R, *et al.* (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nature Rev Genet* 9: 218–29.

Mössner, R, Schön, MP, Reich, K (2008). Tumor necrosis factor antagonists in the therapy of psoriasis. *Clin Dermatol* 26: 486–502.

Musso, G, Gambino, R, Cassader, M (2011). Interactions Between Gut Microbiota and Host Metabolism Predisposing to Obesity and Diabetes. *Annu Rev Med* 62: 361–80.

Nagata, S, Nagase, H, Kawane, K, *et al.* (2003). Degradation of chromosomal DNA during apoptosis. *Cell Death & …*.

Nathans, D, Smith, HO (1975). Restriction endonucleases in the analysis and restructuring of dna molecules. *Annu Rev Biochem* 44: 273–93.

Nawrocki, EP, Kolbe, DL, Eddy, SR (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25: 1335–7.

Nell, S, Suerbaum, S, Josenhans, C (2010). The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models. *Nat Rev Microbiol* 8: 564–77.

Nemes, Z, Marekov, LN, Fésüs, L, *et al.* (1999). A novel function for transglutaminase 1: attachment of long-chain omega-hydroxyceramides to involucrin by ester bond formation. *Proc Natl Acad Sci USA* 96: 8402–7.

Nestle, FO, Meglio, PD, Qin, J-Z, *et al.* (2009). Skin immune sentinels in health and disease. *Nature Rev Immunol* 9: 679–91.

Ng, KP, Yew, SM, Chan, CL, *et al.* (2013). Draft Genome Sequence of Herpotrichiellaceae sp. UM 238 Isolated from Human Skin Scraping. *Genome Announc* 1.

NIH HMP Working Group, Peterson, J, Garges, S, *et al.* (2009). The NIH Human Microbiome Project. *Genome Research* 19: 2317–23.

Ning, Z, Cox, AJ, Mullikin, JC (2001). SSAHA: a fast search method for large DNA databases. *Genome Research* 11: 1725–9.

Ogawa, T, Katsuoka, K, Kawano, K (1994). Comparative study of staphylococcal flora on the skin surface of atopic dermatitis patients and healthy subjects. *The Journal of Dermatology* 21: 453–60.

Oksanen, J, Blanchet, FG, Kindt, R, *et al.* (2011). Vegan: community ecology package. *R package version 117-10*.

Olsen, GJ, Lane, DJ, Giovannoni, SJ, *et al.* (1986). Microbial Ecology and Evolution: A Ribosomal RNA Approach. *Annu Rev Microbiol* 40: 337–65.

Omori, E, Matsumoto, K, Sanjo, H, *et al.* (2006). TAK1 is a master regulator of epidermal homeostasis involving skin inflammation and apoptosis. *J Biol Chem* 281: 19610–7.

Oren, A, Ganz, T, Liu, L, *et al.* (2003). In human epidermis, beta-defensin 2 is packaged in lamellar bodies. *Exp Mol Pathol* 74: 180–2.

Oshima, H, Rochat, A, Kedzia, C, *et al.* (2001). Morphogenesis and renewal of hair follicles from adult multipotent stem cells. *Cell* 104: 233–45.

Otto, M (2009). Staphylococcus epidermidis — the “accidental” pathogen. *Nat Rev Microbiol* 7: 555–67.

Otto, M, Echner, H, Voelter, W, *et al.* (2001). Pheromone cross-inhibition between Staphylococcus aureus and Staphylococcus epidermidis. *Infect Immun* 69: 1957–60.

O'Reilly, N, Bergin, D, Reeves, EP, *et al.* (2012). Demodex-associated bacterial proteins induce neutrophil activation. *British Journal of Dermatology* 166: 753–60.

Paez, JG, Lin, M, Beroukhim, R, *et al.* (2004). Genome coverage and sequence fidelity of φ29 polymerase–based multiple strand displacement whole genome amplification. *Nucleic Acids Research* 32: e71.

Pages, H, Aboyoun, P, Gentleman, R, *et al.* (2012). Biostrings: String objects representing biological sequences, and matching algorithms. *R package*.

Paladini, RD, Coulombe, PA (1998). Directed expression of keratin 16 to the progenitor basal cells of transgenic mouse skin delays skin maturation. *The Journal of Cell Biology* 142: 1035–51.

Palmer, C, Bik, EM, DiGiulio, DB, *et al.* (2007). Development of the Human Infant Intestinal Microbiota. *PLoS Biol* 5: e177.

Pan, A, Dutta, C, Das, J (1998). Codon usage in highly expressed genes of Haemophillus influenzae and Mycobacterium tuberculosis: translational selection versus mutational bias. *Gene* 215: 405–13.

Peck, MD (2011). Epidemiology of burns throughout the world. Part I: Distribution and risk factors. *Burns* 37: 1087–100.

Peschel, A, Jack, RW, Otto, M, *et al.* (2001). Staphylococcus aureus resistance to human defensins and evasion of neutrophil killing via the novel virulence factor MprF is based on modification of membrane lipids with l-lysine. *J Exp Med* 193: 1067–76.

Peschel, A, Otto, M, Jack, RW, *et al.* (1999). Inactivation of the dlt operon in Staphylococcus aureus confers sensitivity to defensins, protegrins, and other antimicrobial peptides. *J Biol Chem* 274: 8405–10.

Peterson, DA, Mcnulty, NP, Guruge, JL, *et al.* (2007). IgA response to symbiotic bacteria as a mediator of gut homeostasis. *Cell Host Microbe* 2: 328–39.

Pérez-Cobas, AE, Gosalbes, MJ, Friedrichs, A, *et al.* (2012). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*.

Pflughoeft, KJ, Versalovic, J (2012). Human Microbiome in Health and Disease. *Annu Rev Pathol Mech Dis* 7: 99–122.

Phillips, AJ (2006). Homology assessment and molecular sequence alignment. *J Biomed Inform* 39: 18–33.

Pinto, AJ, Raskin, L (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS ONE* 7: e43093.

Podar, M, Abulencia, CB, Walcher, M, *et al.* (2007). Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Applied and Environmental Microbiology* 73: 3205–14.

Pop, M, Salzberg, SL (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet* 24: 142–9.

Porreca, GJ, Zhang, K, Li, JB, *et al.* (2007). Multiplex amplification of large sets of human exons. *Nat Meth* 4: 931–6.

Posada, D (2008). jModelTest: Phylogenetic Model Averaging. *Mol Biol Evol* 25: 1253–6.

Powell, S, Szklarczyk, D, Trachana, K, *et al.* (2012). eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research* 40: D284–9.

Proksch, E, Brandner, JM, Jensen, J-M (2008). The skin: an indispensable barrier. *Experimental Dermatology* 17: 1063–72.

Qi, R, Hua-Song, Z (2013). Leflunomide inhibits the apoptosis of human embryonic lung fibroblasts infected by human cytomegalovirus. *Eur J Med Res* 18: 3.

Qin, J, Li, Y, Cai, Z, *et al.* (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490: 55–60.

Raes, J, Bork, P (2008). Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol* 6: 693–9.

Raes, J, Letunic, I, Yamada, T, *et al.* (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol* 7: 473.

Raghunathan, A, Ferguson, HR, Bornarth, CJ, *et al.* (2005). Genomic DNA amplification from a single bacterium. *Applied and Environmental Microbiology* 71: 3342–7.

Rasko, DA, Rosovitz, MJ, Myers, GSA, *et al.* (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates.

Redel, H, Gao, Z, Li, H, *et al.* (2013). Quantitation and composition of cutaneous microbiota in diabetic and nondiabetic men. *J Infect Dis* 207: 1105–14.

Remington, KA, Heidelberg, K, Venter, JC (2005). Taking metagenomic studies in context. *Trends Microbiol* 13: 404.

Riesenfeld, CS, Schloss, PD, Handelsman, J (2004). Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38: 525–52.

Rocha, EPC (2008). The organization of the bacterial genome. *Annu Rev Genet* 42: 211–33.

Rochat, A, Kobayashi, K, Barrandon, Y (1994). Location of stem cells of human hair follicles by clonal analysis. *Cell* 76: 1063–73.

Roche Diagnostics GmbH (2008). *GS FLX Titanium emPCR Method Manual*. Roche Diagnostics.

Rodrigue, L, Lavoie, MC (1996). Comparison of the proportions of oral bacterial species in BALB/c mice from different suppliers. *Lab Anim* 30: 108–13.

Rodrigue, S, Malmstrom, RR, Berlin, AM, *et al.* (2009). Whole genome amplification and de novo assembly of single bacterial cells. *PLoS ONE* 4: e6864.

Rodrigue, S, Materna, AC, Timberlake, SC, *et al.* (2010). Unlocking short read sequencing for metagenomics. *PLoS ONE* 5: e11840.

Rodríguez, F, Oliver, JL, Marín, A, *et al.* (1990). The general stochastic model of nucleotide substitution. *J Theor Biol* 142: 485–501.

Rogers, GB, Carroll, MP, Serisier, DJ, *et al.* (2005). Bacterial activity in cystic fibrosis lung infections. *Respir Res* 6: 49.

Roop, DR, Hawley-Nelson, P, Cheng, CK, *et al.* (1983). Keratin gene expression in mouse epidermis and cultured epidermal cells. *Proc Natl Acad Sci USA* 80: 716–20.

Rosenthal, M, Goldberg, D, Aiello, A, *et al.* (2011). Skin microbiota: Microbial community structure and its potential association with health and disease. *Infect Genet Evol* 11: 839–48.

Roth, RR, James, WD (1988). Microbial ecology of the skin. *Annu Rev Microbiol* 42: 441–64.

Round, JL, Mazmanian, SK (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature Rev Immunol* 9: 313–23.

Rouse, MS, Rotger, M, Piper, KE, *et al.* (2005). In vitro and in vivo evaluations of the activities of lauric acid monoester formulations against Staphylococcus aureus.

Sandiford, S, Upton, M (2012). Identification, Characterization, and Recombinant Expression of Epidermicin NI01, a Novel Unmodified Bacteriocin Produced by Staphylococcus epidermidis That Displays Potent Activity against Staphylococci. *Antimicrob Agents Chemother* 56: 1539–47.

Sanger, F, Coulson, AR (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94: 441–8.

Sanger, F, Nicklen, S, Coulson, AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74: 5463–7.

Sansonetti, PJ (2004). War and peace at mucosal surfaces. *Nature Rev Immunol* 4: 953–64.

Sato, S, Sanjo, H, Takeda, K, *et al.* (2005). Essential function for the kinase TAK1 in innate and adaptive immune responses. *Nat Immunol* 6: 1087–95.

Schaller, M, Schackert, C, Korting, HC, *et al.* (2000). Invasion of Candida albicans correlates with expression of secreted aspartic proteinases during experimental infection of human epidermis. *J Invest Dermatol* 114: 712–7.

Scharschmidt, TC, List, K, Grice, EA, *et al.* (2009). Matriptase-Deficient Mice Exhibit Ichthyotic Skin with a Selective Shift in Skin Microbiota. *J Invest Dermatol* 129: 2435–42.

Schloss, PD, Gevers, D, Westcott, SL (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6: e27310.

Schmid, I, Nicholson, JK, Giorgi, JV, *et al.* (1997). Biosafety guidelines for sorting of unfixed cells. *Cytometry* 28: 99–117.

Schmidt, TM, Delong, EF, Pace, NR (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol* 173: 4371–8.

Schmieder, R, Edwards, R (2011). Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS ONE* 6: e17288.

Schoch, CL, Crous, PW, Groenewald, JZ, *et al.* (2009). A class-wide phylogenetic assessment of Dothideomycetes. *Stud Mycol* 64: 1–15S10.

Sekirov, I, Russell, SL, Antunes, LCM, *et al.* (2010). Gut microbiota in health and disease. *Physiol Rev* 90: 859–904.

Seth, RB, Sun, L, Ea, C-K, *et al.* (2005). Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. *Cell* 122: 669–82.

Simpson, CL, Patel, DM, Green, KJ (2011). Deconstructing the skin: cytoarchitectural determinants of epidermal morphogenesis. *Nat Rev Mol Cell Biol* 12: 565–80.

Spits, C, Le Caignec, C, De Rycke, M, *et al.* (2006). Optimization and evaluation of single-cell whole-genome multiple displacement amplification. *Hum Mutat* 27: 496–503.

Stamatakis, A, Ludwig, T, Meier, H (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics (Oxford, England)* 21: 456–63.

Steinert, PM, Marekov, LN (1995). The proteins elafin, filaggrin, keratin intermediate filaments, loricrin, and small proline-rich proteins 1 and 2 are isodipeptide cross-linked components of the human epidermal cornified cell envelope. *J Biol Chem* 270: 17702–11.

Stepanauskas, R, Sieracki, ME (2007). Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* 104: 9052–7.

Stevens, CE, Hume, ID (1998). Contributions of microbes in vertebrate gastrointestinal tract to production and conservation of nutrients. *Physiol Rev* 78: 393–427.

Stevenson, BR, Siliciano, JD, Mooseker, MS, *et al.* (1986). Identification of ZO-1: a high molecular weight polypeptide associated with the tight junction (zonula occludens) in a variety of epithelia. *The Journal of Cell Biology* 103: 755–66.

Stiller, M, Knapp, M, Stenzel, U, *et al.* (2009). Direct multiplex sequencing (DMPS)--a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Research* 19: 1843–8.

Storey, JD (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64: 479–98.

Streilein, JW (1983). Skin-associated lymphoid tissues (SALT): origins and functions. *J Invest Dermatol* 80 Suppl: 12s–16s.

Suau, A, Bonnet, R, Sutren, M, *et al.* (1999). Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and Environmental Microbiology* 65: 4799–807.

Sun, C, Mathur, P, Dupuis, J, *et al.* (2006). Peptidoglycan recognition proteins Pglyrp3 and Pglyrp4 are encoded from the epidermal differentiation complex and are candidate genes for the Psors4 locus on chromosome 1q21. *Hum Genet* 119: 113–25.

Sun, T-T, Eichner, R, Nelson, WC, *et al.* (1983). Keratin classes: molecular markers for different types of epithelial differentiation. *J Invest Dermatol* 81: 109s–15s.

SUN, Y, LI, Q, LI, Z, *et al.* (2012). Molecular cloning, expression, purification, and functional characterization of palustrin-2CE, an antimicrobial peptide of Rana chensinensis. *Biosci Biotechnol Biochem* 76: 157–62.

Suzuki, R, Shimodaira, H (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics (Oxford, England)* 22: 1540–2.

Swan, BK, Martinez-Garcia, M, Preston, CM, *et al.* (2011). Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. *Science* 333: 1296–300.

Szponar, B, Pawlik, KJ, Gamian, A, *et al.* (2003). Protein fraction of barley spent grain as a new simple medium for growth and sporulation of soil actinobacteria. *Biotechnol Lett* 25: 1717–21.

Takahashi, H, Aoki, N, Nakamura, S, *et al.* (2000). Cornified cell envelope formation is distinct from apoptosis in epidermal keratinocytes. *J Dermatol Sci* 23: 161–9.

Taylor, G, Lehrer, MS, Jensen, PJ, *et al.* (2000). Involvement of Follicular Stem Cells in Forming Not Only the Follicle but Also the Epidermis. *Cell* 102: 451–61.

Theobald, DL (2010). A formal test of the theory of universal common ancestry. *Nature* 465: 219–22.

Thomas, T, Gilbert, J, Meyer, F (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* 2: 3.

Tlaskalová-Hogenová, H, Stepánková, R, Hudcovic, T, *et al.* (2004). Commensal bacteria (normal microflora), mucosal immunity and chronic inflammatory and autoimmune diseases. *Immunol Lett* 93: 97–108.

Treangen, TJ, Rocha, EPC (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7: e1001284.

Treangen, TJ, Salzberg, SL (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Rev Genet* 13: 36–46.

Tremaroli, V, Bäckhed, F (2012). Functional interactions between the gut microbiota and host metabolism. *Nature* 489: 242–9.

Trifonov, V, Rabadan, R (2010). Frequency analysis techniques for identification of viral genetic data. *MBio* 1.

Tringe, SG, Mering, von, C, Kobayashi, A, *et al.* (2005). Comparative Metagenomics of Microbial Communities. *Science* 308: 554–7.

Turnbaugh, PJ, Gordon, JI (2009). The core gut microbiome, energy balance and obesity. *J Physiol (Lond)* 587: 4153–8.

Turnbaugh, PJ, Hamady, M, Yatsunenko, T, *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* 457: 480–4.

Turnbaugh, PJ, Ley, RE, Hamady, M, *et al.* (2007). The human microbiome project. *Nature* 449: 804–10.

Turnbaugh, PJ, Ley, RE, Mahowald, MA, *et al.* (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444: 1027–31.

Turnbaugh, PJ, Quince, C, Faith, JJ, *et al.* (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proceedings of the National Academy of Sciences* 107: 7503–8.

Tyson, GW, Chapman, J, Hugenholtz, P, *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428: 37–43.

Uçkay, I, Pittet, D, Vaudaux, P, *et al.* (2009). Foreign body infections due to Staphylococcus epidermidis. *Ann Med* 41: 109–19.

Urban, DL, Goslee, SC (2007). The ecodist Package for Dissimilarity-based Analysis of Ecological Data. *Journal of Statistical Software*.

Valdimarsson, H, Thorleifsdottir, RH, Sigurdardottir, SL, *et al.* (2009). Psoriasis--as an autoimmune disease caused by molecular mimicry. *Trends Immunol* 30: 494–501.

Van Geystelen, A, Decorte, R, Larmuseau, MH (2013). AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. *BMC Genomics* 14: 101.

van Hogerlinden, M, Rozell, BL, Ahrlund-Richter, L, *et al.* (1999). Squamous cell carcinomas and increased apoptosis in skin with inhibited Rel/nuclear factor-kappaB signaling. *Cancer Res* 59: 3299–303.

Venter, JC, Adams, MD, Myers, EW, *et al.* (2001). The sequence of the human genome. *Science* 291: 1304–51.

Venter, JC, Remington, K, Heidelberg, JF, *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66–74.

VerBerkmoes, NC, Denef, VJ, Hettich, RL, *et al.* (2009). Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* 7: 196–205.

Vieites, JM, Guazzaroni, MA-E, Beloqui, A, *et al.* (2009). Metagenomics approaches in systems microbiology. *FEMS Microbiology Reviews* 33: 236–55.

Vinga, S, Almeida, JS (2003). Alignment-free sequence comparison-a review. *Bioinformatics* 19: 513–23.

Vlcek, C, Paces, V (1986). Nucleotide sequence of the late region of Bacillus phage phi 29 completes the 19,285-bp sequence of phi 29 genome. Comparison with the homologous sequence of phage PZA. *Gene* 46: 215–25.

Vuong, C, Gerke, C, Somerville, GA, *et al.* (2003). Quorum-sensing control of biofilm factors in Staphylococcus epidermidis. *J Infect Dis* 188: 706–18.

Walker, AW, Sanderson, JD, Churcher, C, *et al.* (2011). High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. *BMC Microbiol* 11: 7.

Wang, C, Deng, L, Hong, M, *et al.* (2001). TAK1 is a ubiquitin-dependent kinase of MKK and IKK. *Nature* 412: 346–51.

Wang, M, Karlsson, C, Olsson, C, *et al.* (2008). Reduced diversity in the early fecal microbiota of infants with atopic eczema. *J Allergy Clin Immunol* 121: 129–34.

Wang, Q, Garrity, GM, Tiedje, JM, *et al.* (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73: 5261–7.

Wang, Y, Qian, P-Y (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE* 4: e7401.

Warnecke, F, Luginbühl, P, Ivanova, N, *et al.* (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450: 560–5.

Warner, RR, Myers, MC, Taylor, DA (1988). Electron probe analysis of human skin: determination of the water concentration profile. *J Invest Dermatol* 90: 218–24.

Watt, FM, Hogan, BL (2000). Out of Eden: stem cells and their niches. *Science* 287: 1427–30.

Wei, C-C, Chen, W-Y, Wang, Y-C, *et al.* (2005). Detection of IL-20 and its receptors on psoriatic skin. *Clin Immunol* 117: 65–72.

Wells, JM, Rossi, O, Meijerink, M, *et al.* (2011). Epithelial crosstalk at the microbiota-mucosal interface. *Proc Natl Acad Sci USA* 108 Suppl 1: 4607–14.

White, RA, Blainey, PC, Fan, HC, *et al.* (2009). Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics* 10: 116.

Wigginton, JE, Cutler, DJ, Abecasis, GR (2005). A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 76: 887–93.

Willing, BP, Dicksved, J, Halfvarson, J, *et al.* (2010). A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* 139: 1844–1854.e1.

Willner, D, Furlan, M, Haynes, M, *et al.* (2009a). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 4: e7370.

Willner, D, Thurber, RV, Rohwer, F (2009b). Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental Microbiology* 11: 1752–66.

Woese, CR, Fox, GE (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* 74: 5088–90.

Wolf, R, Orion, E, Ruocco, E, *et al.* (2012). Abnormal epidermal barrier in the pathogenesis of psoriasis. *Clin Dermatol* 30: 323–8.

Wolinsky, H (2007). The thousand-dollar genome. Genetic brinkmanship or personalized medicine? *EMBO Rep* 8: 900–3.

Wommack, KE, Bhavsar, J, Ravel, J (2008). Metagenomics: read length matters. *Applied and Environmental Microbiology* 74: 1453–63.

Wood, LC, Elias, PM, Calhoun, C, *et al.* (1996). Barrier Disruption Stimulates Interleukin-1alpha Expression and Release from a Pre-Formed Pool in Murine Epidermis. *J Invest Dermatol* 106: 397–403.

Woyke, T, Sczyrba, A, Lee, J, *et al.* (2011). Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. *PLoS ONE* 6: e26161–.

Wu, D, Hugenholtz, P, Mavromatis, K, *et al.* (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462: 1056–60.

Xiao, Y, Xiao, MY (2012). Package "CvM2SL2Test."

Zengler, K (2009). Central role of the cell in microbial ecology. *Microbiol Mol Biol Rev* 73: 712–29.

Zhang, E, Tanaka, T, Tajima, M, *et al.* (2011). Characterization of the skin fungal microbiota in patients with atopic dermatitis and in healthy subjects. *Microbiol Immunol* 55: 625–32.

Zhang, K, Martiny, AC, Reppas, NB, *et al.* (2006). Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* 24: 680–6.

Zhang, Y-X, Perry, K, Vinci, VA, *et al.* (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 415: 644–6.

Zheng, Z, Advani, A, Melefors, Ö, *et al.* (2010). Titration-free massively parallel pyrosequencing using trace amounts of starting material. *Nucleic Acids Research* 38: e137–7.

Zheng, Z, Advani, A, Melefors, Ö, *et al.* (2011). Titration-free 454 sequencing using Y adapters. *Nat Protoc* 6: 1367–76.

Annexes

Annex 1. Supplementary materials and methods.

Chapter 2. A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin.

Proposed protocol details

Skin enzymatic digestion and bacterial first separation

One skin portion for each individual, 6 mm wide and 5-10 mm deep was incubated in 1.5 mL of buffered enzymatic solution containing sterile HBSS 1X(Invitrogen, Paisley PA4 9RF, UK), collagenase (1 mg/mL) (Sigma-Aldrich. St. Louis, MO), dispase II(1 mg/mL) (Roche Applied Science. Penzberg, Germany), and trypsin 0.025% (Sigma-Aldrich) for 30 min at 37°C constantly shaking. We included a mock "empty" sample which followed the same protocol than the rest of samples. The cell saturated solution was then neutralized with BSA 3%-EDTA and stored on ice, and the remaining skin was re-incubated with a new fresh aliquot of enzymatic solution for 30 more minutes. This digestion was repeated until no remaining skin was observed. Three digestions were needed to obtain the total homogenization of the skin sample.

Sterilized nylon filters of 80, 20 and 11 µm (Millipore,Billerica, MA) were pre-incubated in tween20 (Sigma-Aldrich) 0.1% for 10 min to facilitate the filtration, avoiding the possible adherence of bacteria to the filters, and mounted in a sterile support (swinnex, Millipore). The remaining tween20 was then washed away with BSA 3% in PBS (Sigma Aldrich). The cell suspension was filtered using the filters in a decreasing order, to sequentially reduce the amount of eukaryotic cells in the suspension. To avoid the attachment of the prokaryotic cells to the filters, 10 mL of fresh HBSS was filtered to drag and collect the maximum amount possible of prokaryotes. Cells were washed and concentrated by centrifugation in 500 µL of nuclease-free PBS and quantified by spectrophotometry.

Bacterial-mitochondrial separation by flow cytometry

Eukaryotic organelle contamination is an important issue to have in mind in bacterial diversity studies since they may be amplified by universal 16S rRNA primers(Benítez-Páez et al. 2013). Given that the original tissue was a complete slice of skin, and skin is rich in mitochondria, which may be released during the filtration, a cytometric separation was performed using a FACSAria II SORP cell sorter (Becton Dickinson, San Jose, CA), as follows: the remaining cells were centrifuged for 3 min at 12K rpm, treated with 25 µM of DNAse I (NEB. Hitchin, Herts. UK) stock solution (1mg/mL) for 20 min., to remove any possible released DNA, and resuspended in 1 mL of ethanol 70%, leaving the cells at 4°C overnight to fix. Ethanol was washed away three times with PBS. Propidium Iodide (PI) staining solution was prepared as follows: 50 µL of PI stock solution (1 mg/mL) were diluted together with 50 µL of RNAse A (Qiagen, Valencia. CA) stock solution (both 1 mg/mL) in PBS to a final volume of 1 mL. Cells were resuspended with staining solution and incubated at 4°C overnight. The cell sorter was prepared according to the type of cells we were trying to separate. Flow cytometry data were collected on the cell sorter using a 85 µm nozzle setup. A single laser assay was carried out using a 488-nm blue laser. Forward (FSC) and Side scatter (SSC) were collected on logarithmic scale to cover wide range of particle size detection (<1 to 10 µm) and SSC and PI thresholds were set at 300 and 200 respectively to exclude electronic noise but neither low FSC nor low PI emitting particles. PI fluorescence was measured through a bandpass (BP) 605/40 filter. Acquisition was stopped when 15,000 gated events were collected in the PI vs. autofluorescence dot plot. Autofluorescence was detected in the green channel through 525/30 BP filter. Gates for cell sorting were established on low, medium and bright PI relative fluorescence intensity accordingly to the size and DNA content by three different test controls.

First, sterile inert beads (Sigma Aldrich) of 1, 2.5 and 10 µm of diameter were used to locate the specific size window for bacteria and mitochondria. A size-specific gate was defined. Two different gates were defined for the mitochondria (1-2.5 µm) and bacteria (1-10 µm) sizes. Given that beads do not have fluorescence emission, PI background was set using the same beads. Second, Escherichia coli cells from a pure culture were used as bacterial standard to assess size and genomic content. Finally, bacteria from feces, which are almost free of eukaryotic cells(Arumugam et al. 2011) were used as a standard complex community to take into account possible deviations on the size range (see below for more information on the control preparation). Controls were sorted sequentially, and all variables (Forward scatter, side scatter, fluorescence range, autofluorescence) were set for further analyses. Isolated cells were then sorted using the previously set thresholds, removing all out-of-threshold events. Given that the genome size of mouse mitochondria is 16 kb, and accounting for ~ 25X polyploidy, genomic size threshold was set at 1 Mb. All particles containing more than 0.5 Mb of DNA and with a size of 1-10 µm were recovered. Also everything below the size range of 1µm was recovered.

The remaining cell suspension was centrifuged for 3 min at 8K rpm, to precipitate the cells, and washed with PBS 1X (Ambion. Life Technologies. Austin TX), to eliminate possible eukaryotic debris. The remaining cells were then incubated in a preparation of buffer solution (PBS 1X) and lysozyme (10 mg/mL) for 30 min at 37°C, and subsequently cooled on ice. DNA extraction was performed using a phenol-chloroform standard method. The exact details for DNA extraction are explained below.

Extracted DNA was eluted in 50 µL of nuclease-free Tris-EDTA buffer (TE), and the resulting suspension was quantified using Nanodrop.

As an alternative control protocol, frozen skin samples of similar weight were homogenized in PBS using a mechanical homogenizer (IKA® Ultraturrax. Thermo Fischer), and followed by the standard protocol. Whole genomic DNA was extracted using the same protocol. All the following steps were carried out in parallel for both extraction methods.

Method validation

We controlled for presence of bacterial, host, and human DNA in our enriched samples. To verify the presence or absence of bacterial DNA, a standard PCR amplification of the 16S rRNA gene was performed from the purified genomic DNA using the universal primers 8F (5'-AGAGTTTGATCCTGGCTCAG-3') and 355R (5'-CTGCTGCCTCCCGTAGGAGT-3') (Baker et al. 2003). For each 50 µL reaction, conditions were as follows: 5.0 µL of 10X buffer with MgCl2 (Roche Applied Science), 2 µL of dNTP mix (10 mM each; Roche Applied Science), 1.5 µL of each primer (20 µM; IDT), 5 µL of DMSO, 1.5ng of bacterial genomic DNA, and 1 U of FastStart High Fidelity Taq Polymerase (Roche Applied Science). Thermocycling conditions were the following: initial denaturation at 95°C for 2 min, followed by 30-32 cycles of a 30-sec 95°C denaturation, 30-sec annealing at 55°C, 1-min elongation at 72°C, and a final extension of 8 min at 72°C. PCR products were visualized on an 1% agarose gel. All of them should contain just one band corresponding to 350 bp. Negative and positive control PCR reactions were performed with each set of amplifications and in all cases did produce the expected result.

Host DNA was detected by means of the rodent-specific IRGA6 gene (Bekpen et al. 2005; 2009), which was PCR-amplified with primers 269F (5'-GAGGCATTGGGAATGAAGAA-3') and 668R (5'-GCAATGCCATTCTCCCTAAA-3') following the same conditions than for 16S rRNA gene amplification. Experimental contamination (which may alter the bacterial composition by introducing exogenous prokaryotes) was measured by detecting human DNA; to that effect, we amplified the NPIP gene(Johnson et al. 2001), which is specific for the primate lineage and is not shared with rodents. This combination of IRGA6 and NPIP genes allows us to discriminate between host and experimental contamination.

To assess the actual ratio between bacterial and mouse DNA isolated, a qPCR experiment was performed. First, the 16S rRNA gene V1-V2 region was amplified using primers 63F (5'-GCAGGCCTAACACATGCAAGTC-3') and 355R (5J CTGCTGCCTCCCGTAGGAGT-3')(Castillo et al. 2006; Grice et al. 2008). The possible host DNA was also quantified using the IRGA6 gene primers 151F(5' -AGAGCACACCGAGGGCTATTC-3') and 257R (5'-GAACAGCTGACCCATGACTTCA5'). To perform an absolute quantification and assess the efficiency of the PCR, a standard curve was constructed by amplifying serial dilutions of known quantities of E. coli DNA and mouse DNA for each one. The qPCR experiment was performed on a LightCycler 480 (Roche Applied Science) using optical grade 385-well plates. Each 10 µL reaction included 5 µL SybrGreen Master Mix (Roche Applied Science), 1 µL each primer (10 µM), 1 µL water, and 2 µL of DNA. For each DNA sample, three replicates were performed. The cycling conditions used were as follows: initial denaturation at 95°C for 4 min, followed by of 40 cycles of 10 s 95° denaturation and 60-sec 60°C elongation. The standard curve equations were: Cq =-0.43x + 8.43; for E. coli and Cq=-0.46x+13.54 for IRGA6, both with R^2=1 and an efficiency of 2.

According to the amount of DNA obtained after the protocol, the following nested PCR protocol was performed: for each sample we amplified 16S rRNA genes using 8F and 1510R (5'-CGGTTACCTTGTTACGACTT-3') covering the nine variable regions without any phylogenetic bias(Schloss et al. 2011). The mix conditions were the same than in the previous amplification, and the cycling conditions were as follows: a first denaturation step of 2 min at 95°C, followed by 25 cycles of 30 s at 95°C, 30 sat 55°C and 90 s at 72°C, followed by a final extension step of 8 min at 72°C. This first amplification was followed with a semi-nested PCR using primers 8F and 355R, modified according to Costello et al. (Costello et al. 2009). The forward primer (5'-CGTATCGCCTCCCTCGCGCCATCAG-XXXXXXXXXX-AGAGTTTGATYMTGGCTCAG -3') contained the 454 Life Sciences Titanium primer A sequence, followed by a 10 nt error correcting barcode, and the degenerated primer 8F. The reverse primer (5'-CTATGCGCCTTGCCAGCCCGCTCAGTGCTGCCTCC-CGTAGGAGT -3') contained the 454 Life Sciences primer B sequence, the broadly conserved bacterial primer 355R, and a two-base linker sequence ('TC'). All forward and reverse primers were digitally tested using GeneRunner. PCR conditions were modified from the previous amplification step reducing the elongation time to 30 s, and the annealing temperature to 52°C to allow maximum adherence. Amplification was repeated three times per sample. Replicate amplicons were pooled and visualized on 1.0% agarose gels using Red Safe DNA gel stain in 0.5X TBE. Amplicons were cleaned using the NucleoFast 96-well plates according to manufacturer's instructions.

Amplicon quantification, pooling, and pyrosequencing.

Amplicon PCR concentrations were determined using a Nanodrop Spectophotometer (Thermo Fisher, e). Cleaned amplicons were pooled in equimolar ratios into a single tube to a final amount of 700 ng of DNA. Pyrosequencing was carried out using primer A on a 454 Life Sciences Genome Sequencer FLX instrument (Roche Applied Science), to an expected final amount of 5,000 sequences per sample.

Mouse metagenomic library preparation and sequencing.

Two independent mouse skin samples were processed to assess the proper ability of this method to perform efficient cheaper metagenomic libraries. Extracted DNA was placed in low-binding tubes and sonicated with a Bioruptor®(Diagenode, Denville, NJ), according to manufacturer's instructions to obtain a 400-800 bp fragment range. As the amount of DNA was relatively low, fragment distribution was verified using an E. coli genomic DNA extraction, processed following the same conditions. Fragmented DNA was then purified using NucleoFast 96-well plates according to manufacturer's instructions, and resuspended on 50µL of TE. Subsequent purification, blunt-end repair, adapter ligation, amplicon quantification and pyrosequencing library generation were carried out following a previously published protocol (Zheng et al. 2011) to avoid a titration step and allowing to work with just 1ng of fragmented DNA. qPCR was performed, as in Zheng Z.*Science* 326: 1694–7.

Coulombe, PA, Wong, P (2004). Cytoplasmic intermediate filaments revealed as dynamic and multipurpose scaffolds. *Nature cell biology* 6: 699–706.

Cox, MP, Peterson, DA, Biggs, PJ (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. et al. (Zheng et al. 2011) to quantify the exact number of molecules in our samples. Further steps were performed according to the standard Roche protocol.

Amplification analysis

Amplicons were processed and analyzed following the procedure described previously (Garcia-Garcerà et al. 2012). First, only sequences between 200 and 400 nt with a quality score >25, containing unambiguous characters, and that did not contain an uncorrectable barcode, were used. The remaining sequences were assigned to samples by examining the 10-nt bar-code. Pyrosequencing noise and chimera sequences were filtered using AmpliconNoise, implemented in qiime(Caporaso et al. 2010) and Chimera sequencing(Haas et al. 2011). Similar sequences were clustered into operational taxonomic units (OTUs, also called phylotypes) using CD-HIT(Li and Godzik 2006) with a minimum coverage of 99% and a minimum identity of 98%. For each OTU, the longest sequence was chosen as representative and was aligned using BLAST(Altschul et al. 1990) against the "nt" database and against a local, self-formatted RDP database (Maidak et al. 2001; Cole et al. 2007; 2009). Only hits with a coverage over 99% and an E-value below 1E-10 were kept. Homopolymeric regions were corrected using the reference sequence (GI) of the top result. In case of taxon ambiguity, the read was assigned to the lowest common ancestor in the NCBI reference taxonomy. OTUs defined as environmental undefined sequence, unknown, or assigned to a level below family were removed from the analysis. A maximum likelihood (ML) phylogenetic tree was constructed using RaxML (Stamatakis et al. 2005) using 100 bootstrap generations and the default configuration. Setting the resulting probabilities as prior probabilities and the maximum likelihood tree as a seed tree, we calculated the posterior probabilities and the most plausible tree shape using phyloBayes(Lartillot and Philippe 2004) with the default configuration.

Shannon, Simpson (1-D) diversity indexes and Chao richness index were calculated for both samples, using the vegan R-package(Oksanen et al. 2011). Multidimensional methods, Correspondence analysis (CoA), Canonical Correspondence Analysis (CCA), principal components analysis (PCA) and Non-metric multidimensional scaling (NMDS) were performed using the vegan and the ecodist R-packages(Urban and Goslee 2007).

Metagenomic sequences post-processing and analysis.

Shotgun sequences were processed as follows. First, reads from the host were eliminated using Deconseq(Schmieder and Edwards 2011). Filtered datasets were then uploaded to MG-RAST(Thomas et al. 2012) in FastQ format and were deposited in the MG-RAST database(Meyer et al. 2008a) with the accession numbers **4496968.3** and **4496969.3**. Uploaded reads were filtered using the quality control (QC) pipeline implemented by MG-RAST(Cox et al. 2010). Reads with ambiguities, low complexity, or short size were eliminated from the analysis. Artificial duplicated sequences were filtered(Gomez-Alvarez et al. 2009).

Taxonomical and functional abundances were calculated with MG-RAST using data normalization and the best representative method. Original files were split in taxonomic specific files to assess functional differences among taxa using phymmBL(Brady and Salzberg 2009). At the same time, COG and eggNOG alignments were downloaded and specific Hidden Markov Models (HMM) were constructed for each alignment using HMMer 3.0(Madera and Gough 2002). Taxonomy-specific reads were aligned using the HMMSearch from HMMer. Results were joined in a single file for each taxon-specific file, and parsed using a customized script written in Perl. Functional assignation was performed according to the highest likelihood obtained. Reads with likelihood below 103 were discarded and considered as "unclassified". Comparisons and NMDS were performed using the package ecodist.

Standard Extraction method.

One skin portion for each individual, 6mm wide and 5-10mm deep was placed in a flat well 2mL tube. Tissue was homogenized with a mechanical homogenizer IKA Ultraturrax (Thermo Scientific, Waltham, MA). 300µL of Tissue lysis buffer (100mM Tris-HCl pH 8.0, 1mM EDTA pH 7.6, 100mM NaCl, SDS 10%) and 50µL of Proteinase K (Roche Applied Science. Penzberg. Germany) were added to the homogenized tissue, and mix was incubated for 30 min at 56°C. Homogenized tissue was ice cooled, and 300µL of Lysozyme (10mg/mL, Sigma-Aldrich, Munich, Germany) buffer in nuclease-free PBS (Ambion, Paisley, UK) was added. Mix was incubated for 30 min at 37°C. 600µL of Phenol:Chlorophorm:Isoamilalcohol (25:24:1) were added, mixed, and spinned at maximum speed for 3 minutes to separate the organic and the aquose phase. Aquose phase was placed in a new tube. Phenol-chlorophorm separation was repeated twice, last time using only chlorophorm. Around 400µL of aquose phase were recovered. 800µL of absolute ethanol and 40µL of ammonium acetate (5M) were added to the aquose phase. The solution was mixed and incubated at -80°C for 2h. DNA was precipitated and washed with ethanol 70%, air dried, and resuspended on 50µL of nuclease-free distilled water.

Standard Feces microbiota isolation and PI staining

10g of feces were collected from a healthy volunteer on a sterile collection 10mL tube with 15mL of PBS. Feces were homogenized and centrifuged for 8 minutes at 4K rpm. Bacteria precipitated as a whitey layer over the fecal debris. Bacteria was then resuspended, using a pipette, and the whole volume was recollected, and split in 10 independent eppendorf tubes. Bacteria were centrifuged for 3' at 8K rpm and resuspended in Ethanol 70% in one unique eppendorf tube. Bacteria were resuspended and incubated O/N at 4°C to fix. Ethanol was washed away three times with PBS. 100µL of Propidium Iodide (PI) staining solution with RNAse was used to stain the bacteria O/N in black eppendorf tubes to keep the light away. This bacterial suspension was used afterwards as bacterial DNA staining and size control.

Annex 2. Supplementary tables and figures.

Chapter 1. Direct sequencing from the minimal number of DNA molecules needed to fill a 454 picotiterplate.

A

Mean read length

		Total processed reads	E. coli mapped reads	Unassigned reads
Multiple displacement amplification	Run 1	277.00	267.60	276.10
	Run 2	159.10	237.20	158.20
Direct sequencing	Run 1	230.00	240.20	177.60
	Run 2	139.60	143.60	122.10

Median read length

		Total processed reads	E. coli mapped reads	Unassigned reads
Multiple displacement amplification	Run 1	245.00	225.00	244.00
	Run 2	131.00	169.00	131.00
Direct sequencing	Run 1	203.00	214.00	159.00
	Run 2	119.00	123.00	108.00

GC content

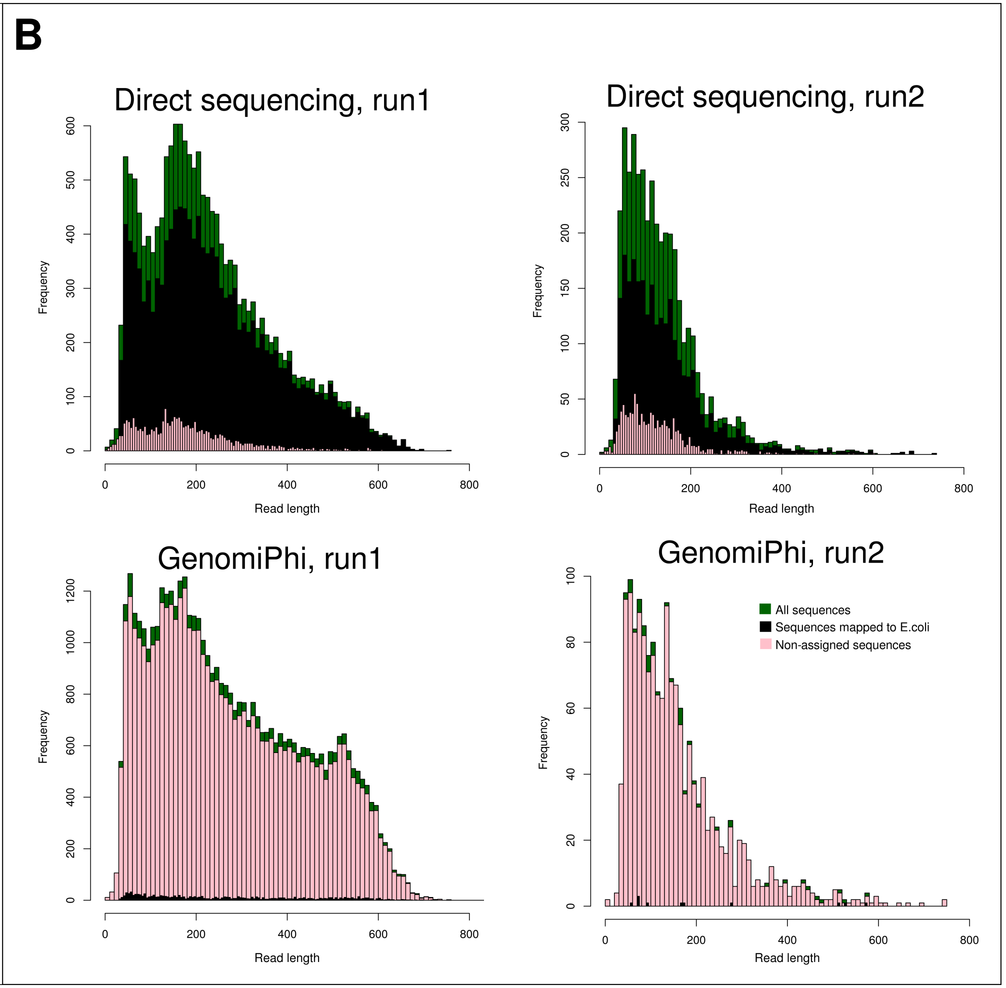
		Total processed reads	E. coli mapped reads	Unassigned reads
Multiple displacement amplification	Run 1	46.05	48.71	46.02
	Run 2	46.15	44.96	46.19
Direct sequencing	Run 1	48.94	48.95	49.57
	Run 2	48.54	48.89	49.36

Read quality

		Total processed reads	E. coli mapped reads	Unassigned reads
Multiple displacement amplification	Run 1	35.92	35.45	35.94
	Run 2	30.84	30.69	30.84
Direct sequencing	Run 1	36.18	36.21	35.90
	Run 2	31.12	31.10	31.19

Read complexity

		Total processed reads	E. coli mapped reads	Unassigned reads
Multiple displacement amplification	Run 1	3.85	3.83	3.85
	Run 2	3.73	3.77	3.73
Direct sequencing	Run 1	3.83	3.85	3.72
	Run 2	3.70	3.73	3.62



Supplementary Figure 1
Panel A: Comparison of read length, read complexity, sequence quality and GC content among total reads, E. coli mapped reads and unassigned reads of both approaches MDA and DS and the both runs. The results in the table indicate that run 2 was performing worse than run 1, but it did not influence the data analysis. The difference between MDA and DS datasets can be observed here only by lower GC content in MDA sample.
Panel B: Distribution of reads length. Distribution of length of total processed reads, reads mapped to E. coli and unclassified reads is compared between both runs of MDA and DS. No differences were found.

SUPPLEMENTARY TABLES AND GRAPHICS

Supplementary Table 1. qPCR values for bacterial 16S rRNA and the murine IRGA6 gene.

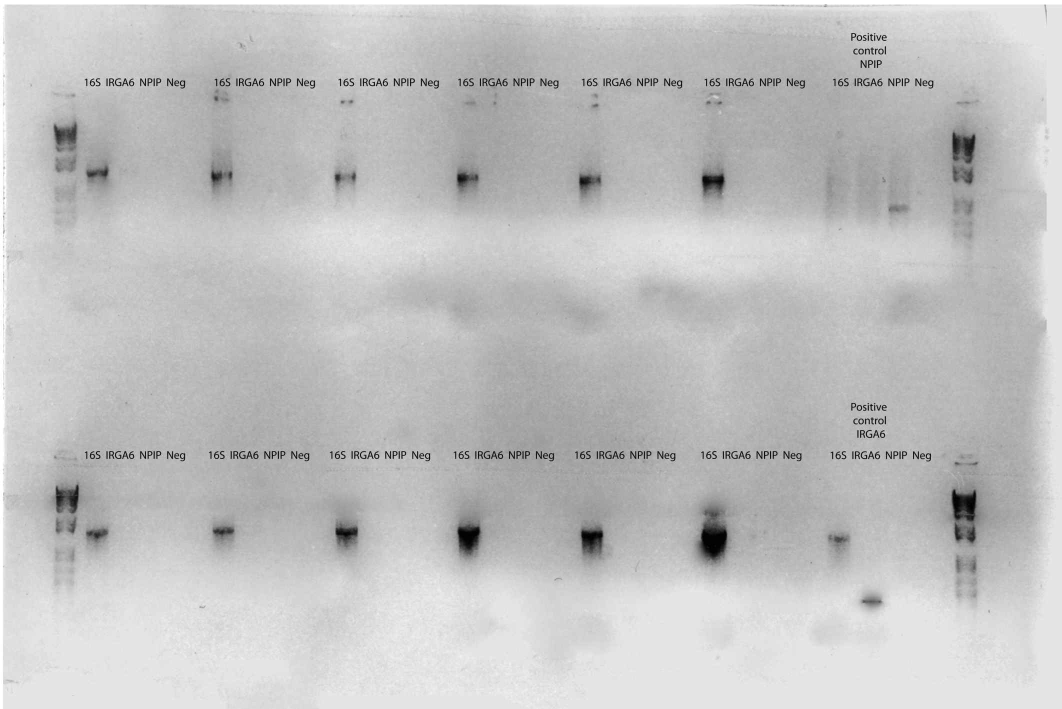
Sample	16S rRNA B	16S rRNA T	IRGA6 B	IRGA6 T
1	26.57	23.97	35.57	28.64
2	27.54	23.21	37.56	21.79
3	18.03	23.44	36.81	21.54
4	23.99	23.17	35.96	26.04
5	25.66	23.51	38.68	23.11
6	27.03	23.34	35.72	21.84
mock	33.08	33.46	37.27	36.89
negative	32.17		37.92	

B, samples enriched for bacteria; T, total extraction samples.

Supplementary Table 2. Diversity measurements of 16S rRNA sequences.

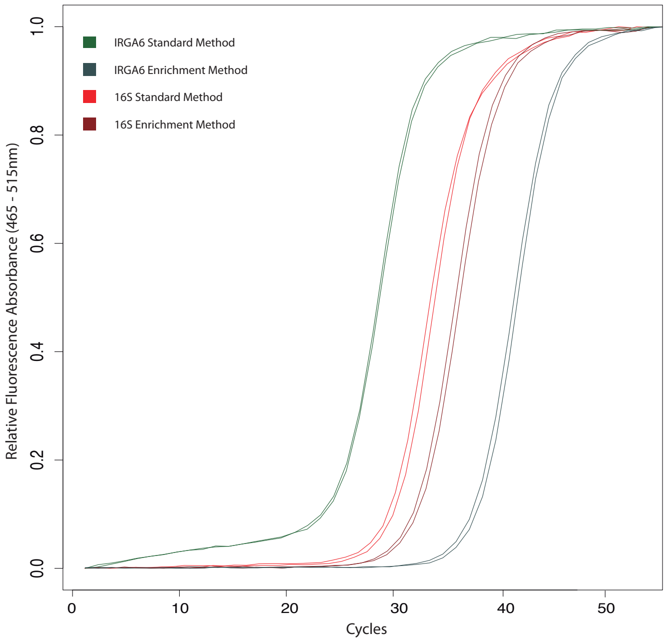
	N	Shannon	Chao1	SE.Chao1	ACE	SE.ACE
B1	33	1.77	40.00	13.15	44.26	3.31
T1	39	2.25	43.00	4.93	49.49	3.77
B2	43	2.04	54.00	11.18	55.69	3.53
T2	49	1.94	66.50	16.08	65.98	4.07
B3	34	1.76	35.25	2.19	37.92	3.05
T3	31	1.46	31.75	1.62	33.48	2.83
B4	27	1.11	30.00	4.80	32.64	2.69
T4	18	0.56	21.00	11.66	21.43	2.33
B5	22	1.36	22.60	1.77	23.43	2.27
T5	15	1.70	15.50	3.74	17.94	1.62
B6	9	1.76	9.50	3.74	11.94	1.51
T6	8	1.15	8.00	NaN	8.64	1.31

Diversity indexes (Shannon diversity index, Chao1 Richness and ACE) were calculated for each sample given a family-based abundance table. B, bacterial enrichment samples; T, total extraction samples. SE, Standard Error; ACE, Abundance-base Coverage Estimator. Undefined (NaN) values appear when all rare taxa are only assigned as singletons.



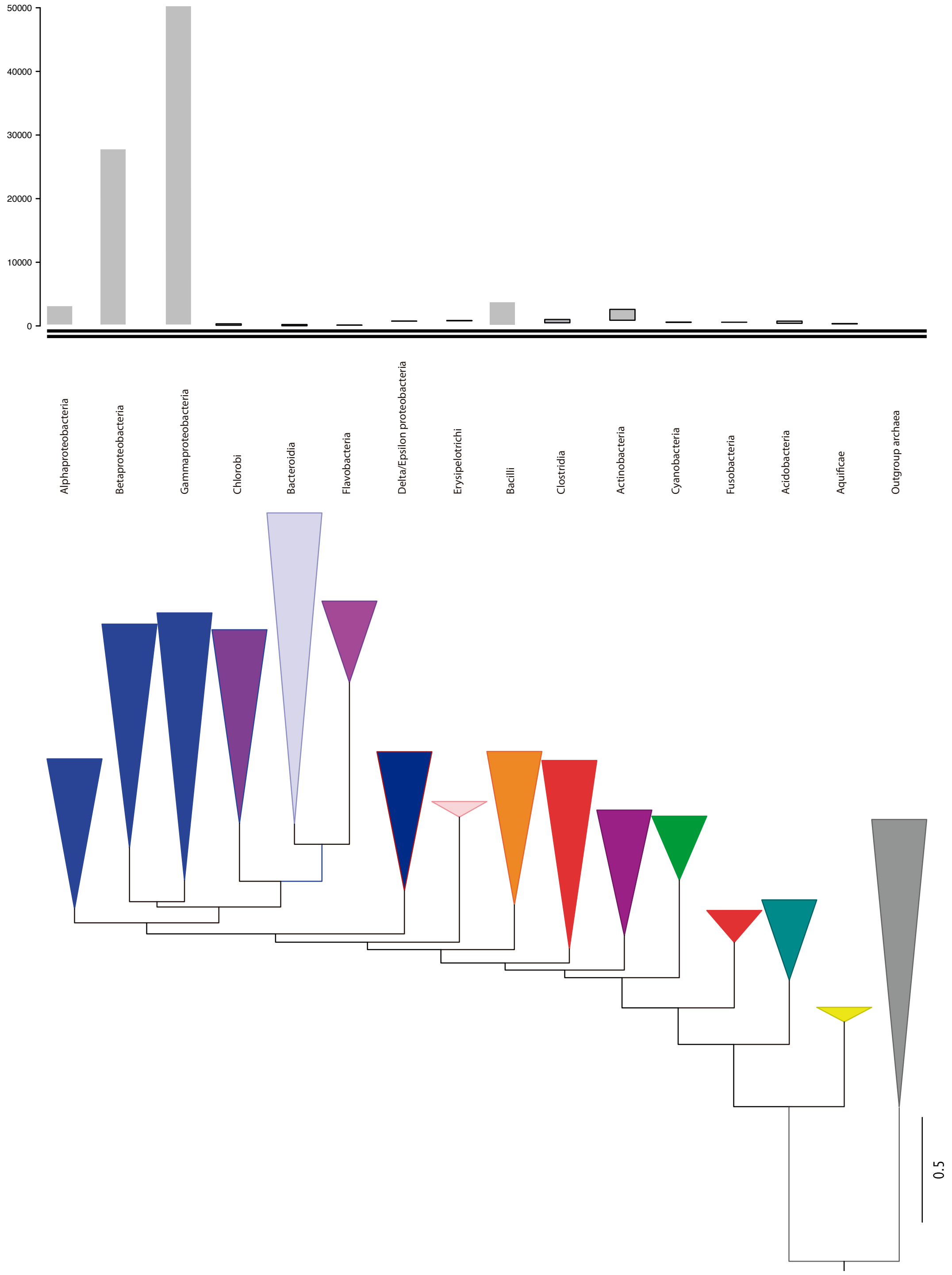
Supplementary Figure 1: Standard Amplification of Host, Bacterial and Contaminant DNA.

Gel visualization of the bacterial (16S), host (IRGA6) and human contaminant (NPIP) gene standard amplification in 12 independent samples, plus a host and a human contaminant controls. Each amplification was performed independently with its own negative control for 16S. Host and contaminant controls were tested independently with their own negative control.



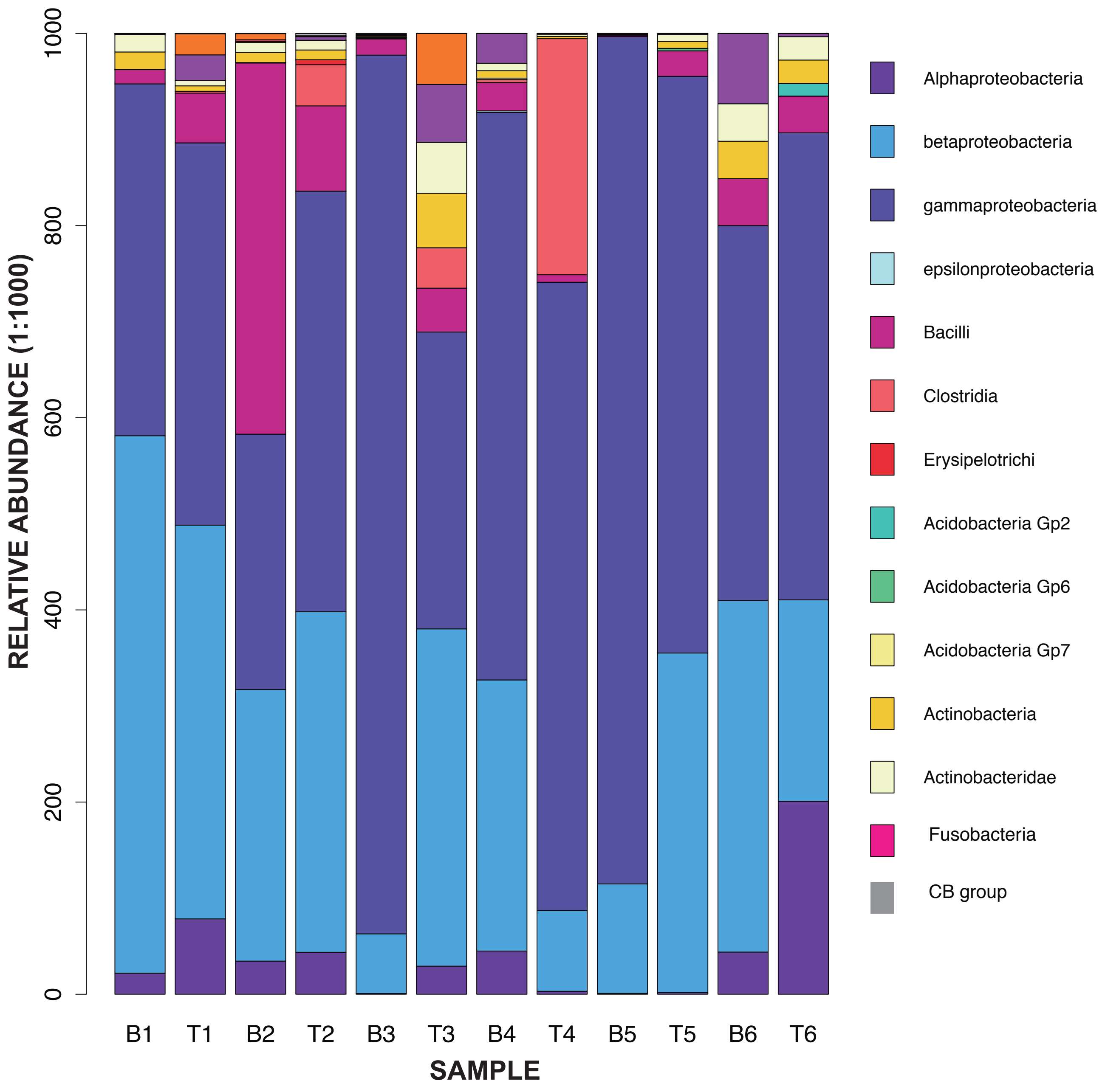
Supplementary Figure 02. Quantification of host and bacterial DNA by qPCR.

Amplification curves of 16S rRNA (red) and IRGA6 (green) genes were tested and compared among standard and proposed method. Two equivalent samples were processed to make comparable the amplification curves. Differences on Cq can be interpreted as differences in the amount of DNA on that sample for a given DNA type.



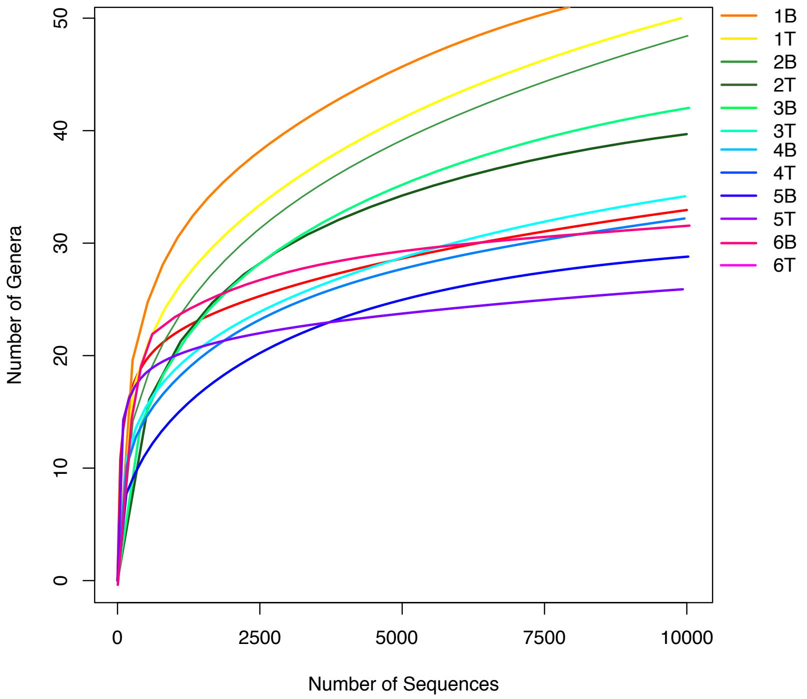
Suppl. 03. Phylogenetic tree reconstruction by RAxML.

Reference phylogenetic tree obtained by Maximum Likelihood analysis of the alignment of 16S reference sequences obtained from RDP and selected by similarity, using CD-HIT. Triangle height indicates phylogenetic diversity within each group. On the upper hand panel, relative abundances of each taxon.



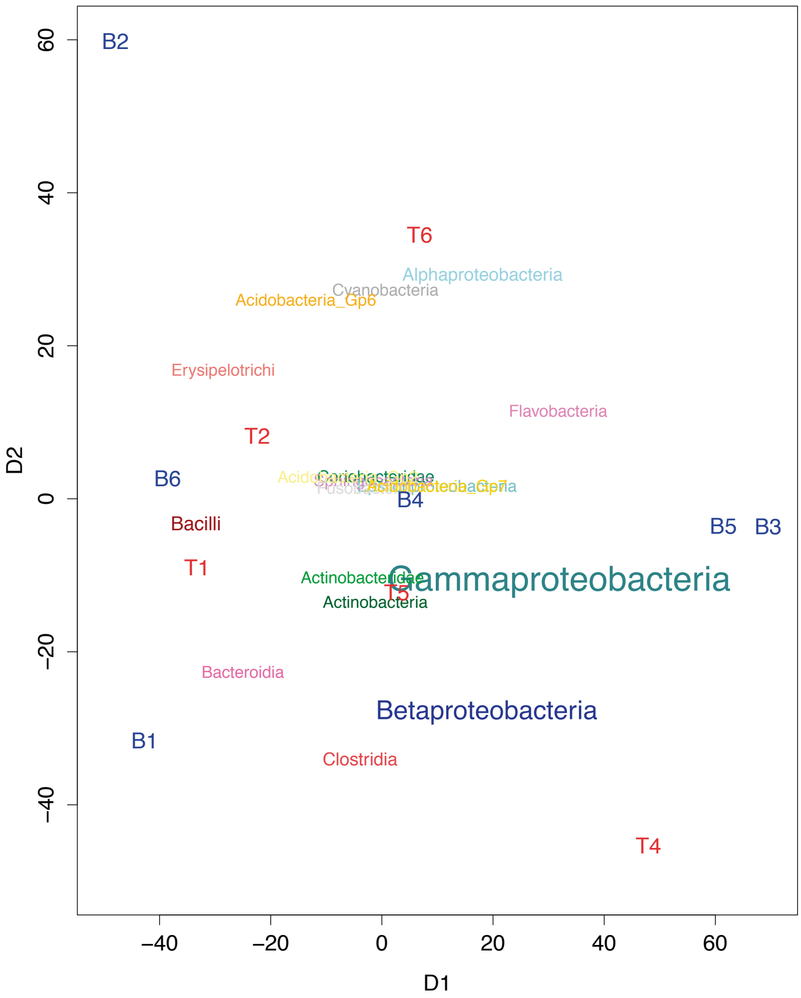
Supplementary Figure 04. Comparison of the relative abundance of bacteria in skin samples using both methodologies of extraction.

16S DNA sequences were assigned to the genera level. Only genera with more than three sequences assigned were used for the analysis. Then whole information was clustered to class level. Methods are named by the capital letter (B from Bacterial Enrichment extraction and T from Total DNA extraction)



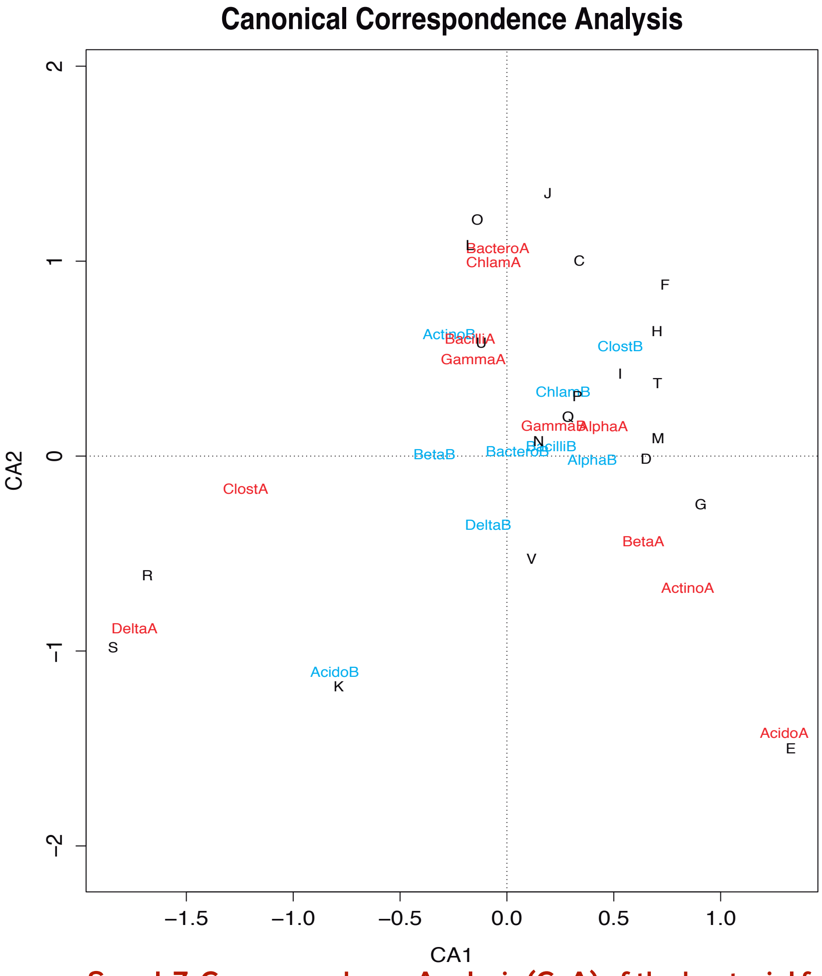
Supplementary Figure 05. Phylogenetic diversity rarefaction curves for the different samples and methods.

Rarefaction curves based on phylogenetic clusters at 97% similarity. Curves were normalized to 10,000 sequences according to the expected slope.



Supplementary Figure 06. Non metric Multidimensional Scaling Analysis of 16S rDNA diversity in standard and bacterial Enrichment methods.

First, NMDS was constructed for Sample diversity, using a Manhattan distance matrix and a 20X replicas. The resulting rank matrix was reduced to two dimensions, which were used to construct the graph. Colors (blue and red) separate both methods following the same pattern on the paper. NMDS matrix was constructed for taxa diversity using the same distance algorithm. Both matrices were normalized one to each other to be comparable. Taxa diversity was plotted by name. Letter size was associated to the relative mean abundance of each taxon. Color was selected by class, using a phylum-based chart.

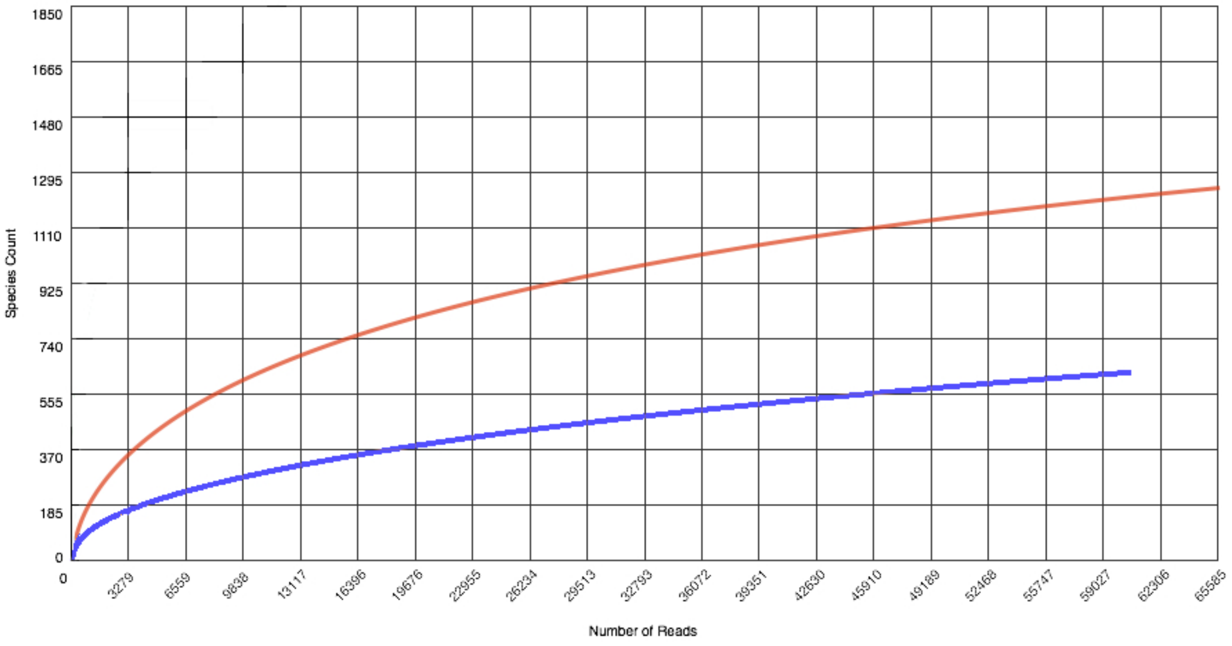


[J]	Translation, ribosomal structure and biogenesis
[A]	RNA processing and modification
[K]	Transcription
[L]	Replication, recombination and repair
[B]	Chromatin structure and dynamics
[D]	Cell cycle control, cell division, chromosome partitioning
[Y]	Nuclear structure
[V]	Defense mechanisms
[T]	Signal transduction mechanisms
[M]	Cell wall/membrane/envelope biogenesis
[N]	Cell motility
[Z]	Cytoskeleton
[W]	Extracellular structures
[U]	Intracellular trafficking, secretion, and vesicular transport
[O]	Posttranslational modification, protein turnover, chaperones
[C]	Energy production and conversion
[G]	Carbohydrate transport and metabolism
[E]	Amino acid transport and metabolism
[F]	Nucleotide transport and metabolism
[H]	Coenzyme transport and metabolism
[I]	Lipid transport and metabolism
[P]	Inorganic ion transport and metabolism
[Q]	Secondary metabolites biosynthesis, transport and catabolism
[R]	General function prediction only

[Acido]	Acidobacteria
[Actino]	Actinobacteria
[Alpha]	Alphaproteobacteria
[Bactero]	Bacteroidia
[Bacill]	Bacilli
[Beta]	Betaproteobacteria
[Chlam]	Chlamidiae
[Clost]	Clostridia
[Delta]	Deltaproteobacteria
[Gamma]	Gammaproteobacteria

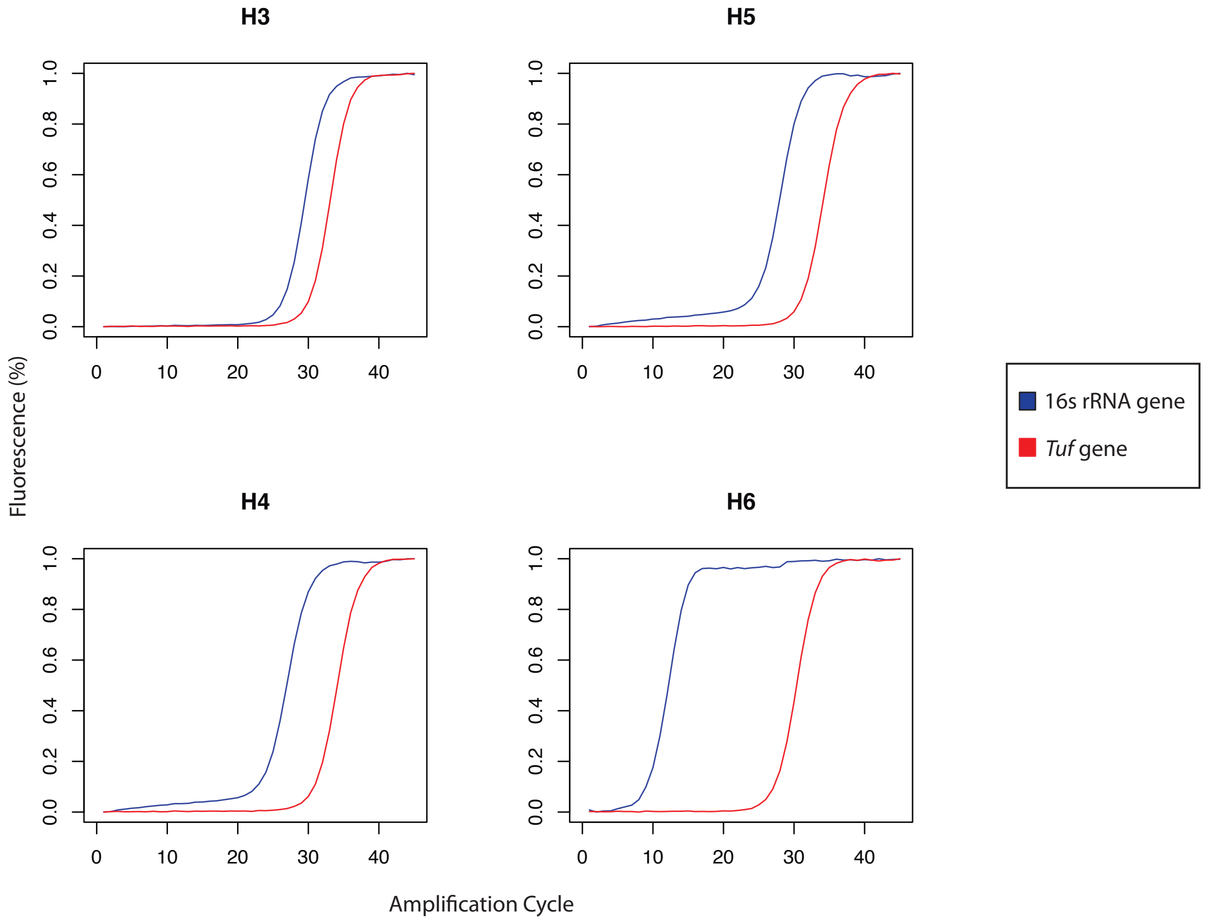
Suppl. 7. Correspondence Analysis (CoA) of the bacterial function in skin samples, separated by taxa.

Samples are represented by color (Sample A in red, sample B in blue), and functions by letter, using the standard eggNOG function category code.



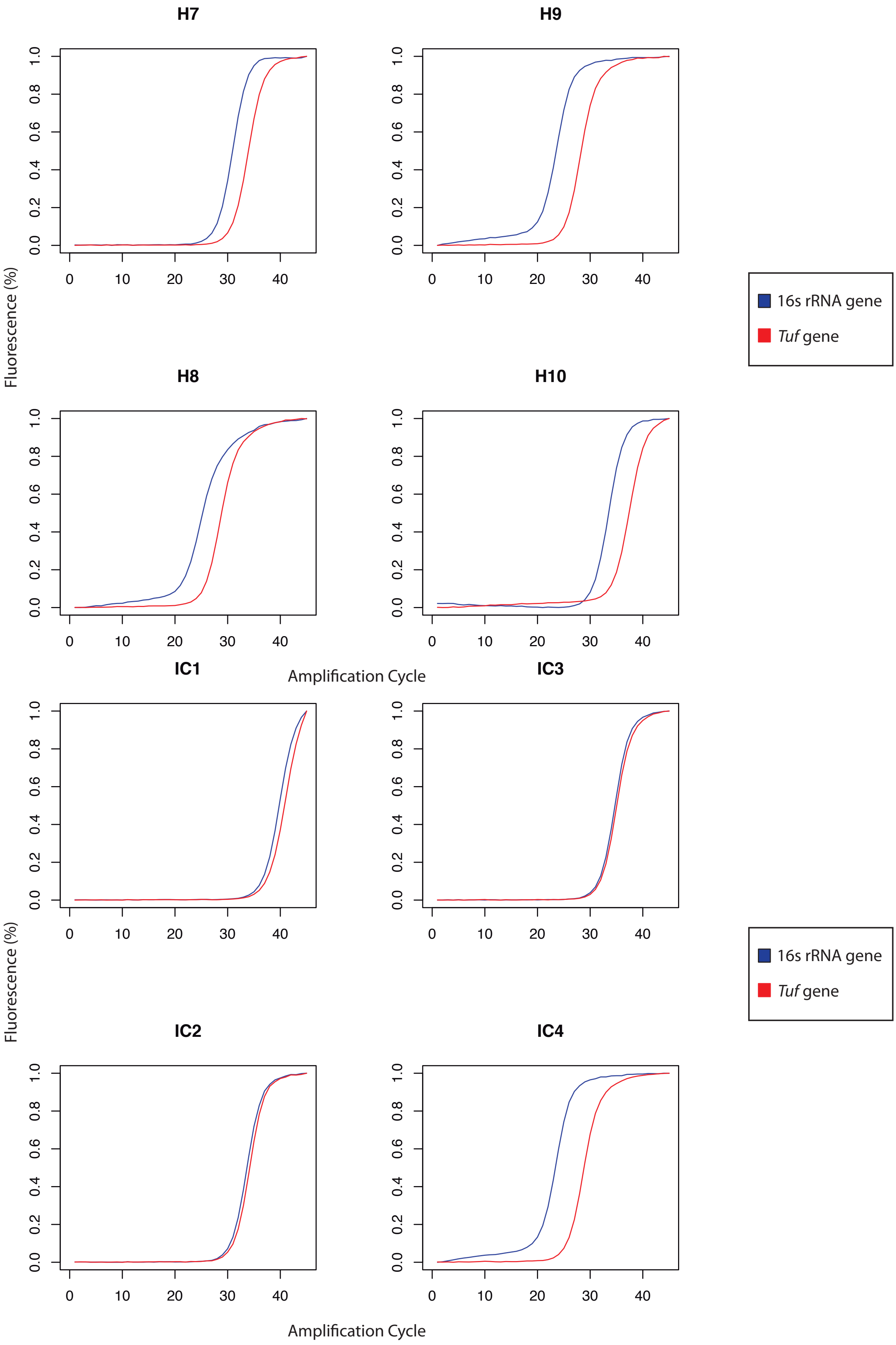
Supplementary figure 08. Phylogenetic diversity rarefaction curves for metagenomic samples.

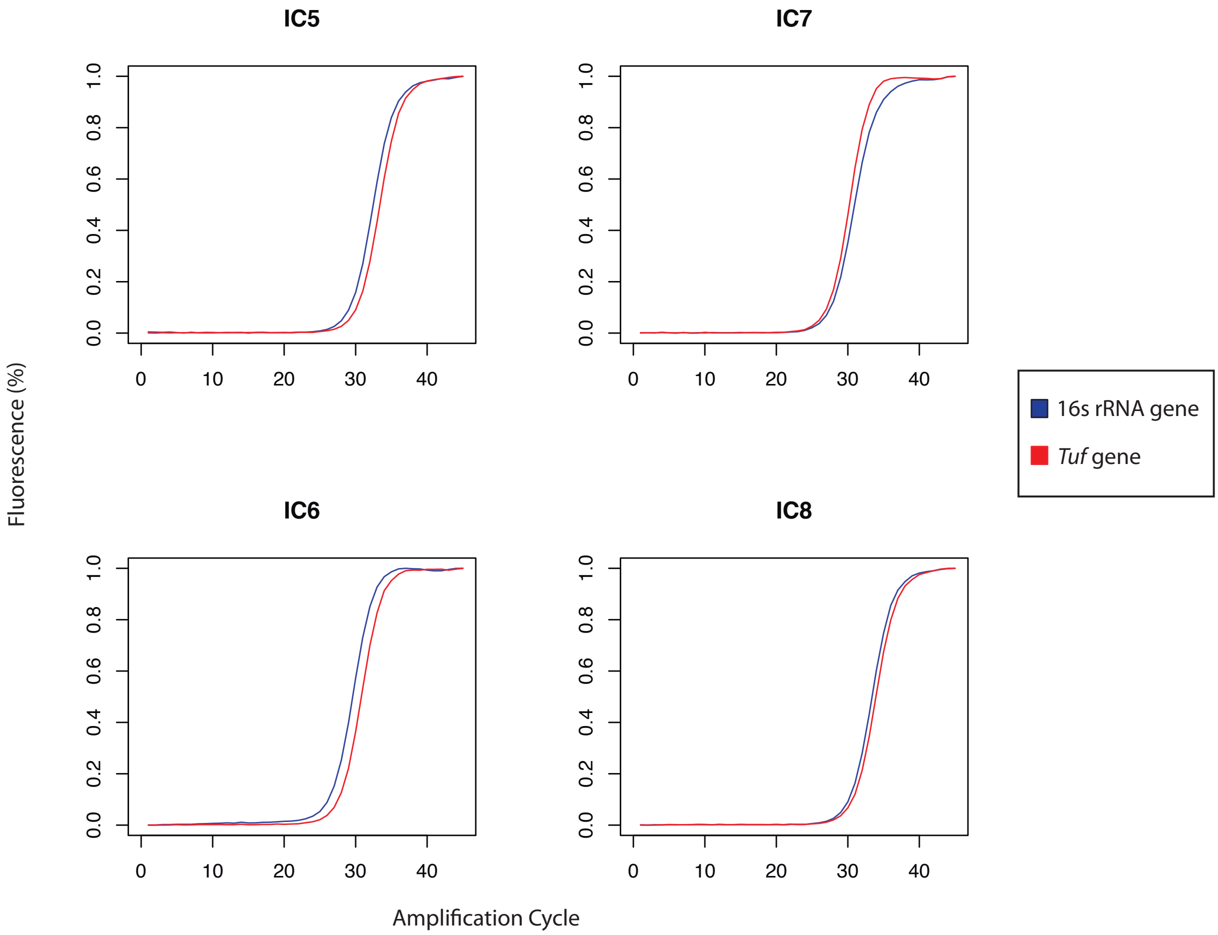
Rarefaction curves were constructed by random sampling of reads with taxonomic assignment, using the taxonomic information previously obtained.

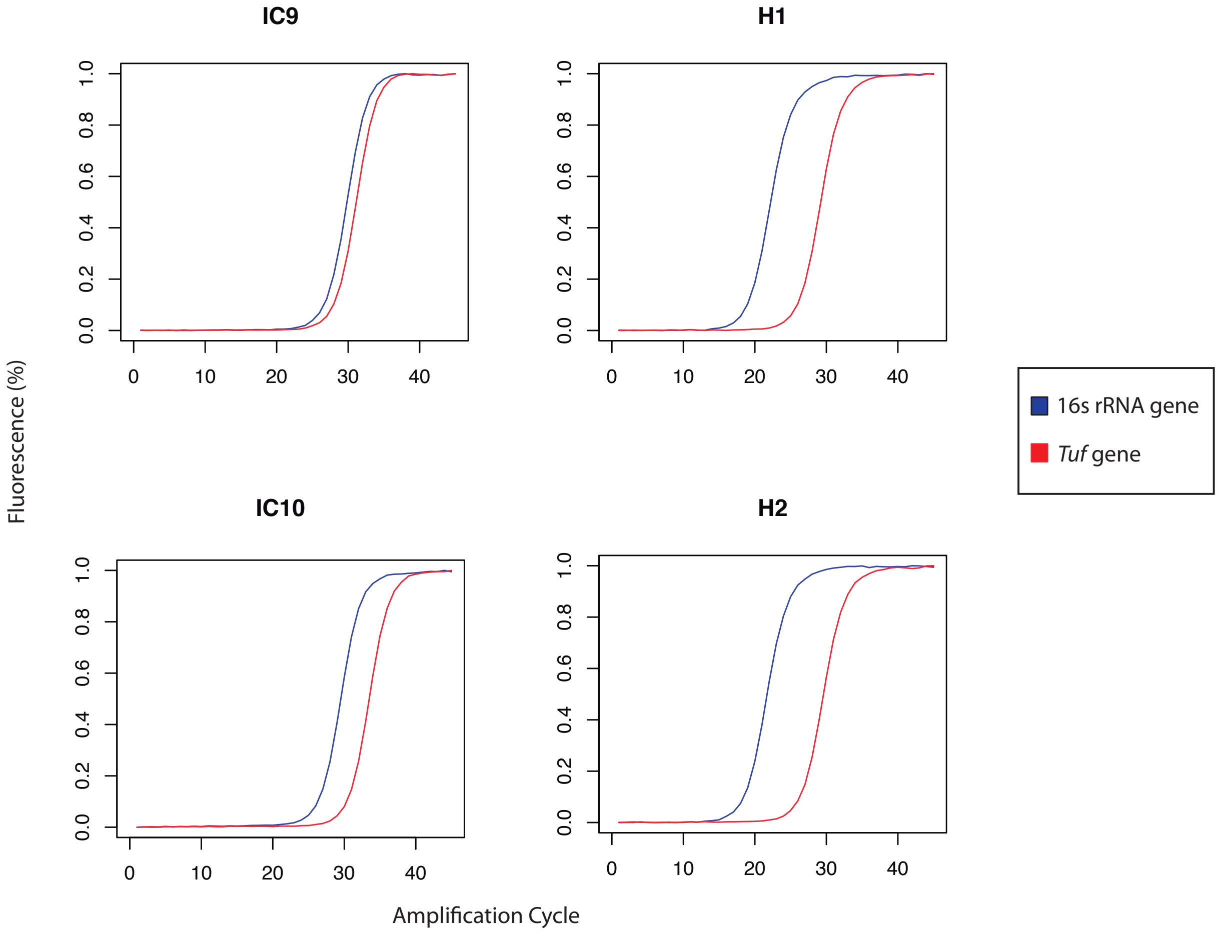


Supplementary figure 1. Comparison of 16s rRNA and tuf genes

Amplification curves for 16s rRNA (red) and tuf (blue) genes were compared to assess the relative amount of Staphylococci ratio in the sample. To allow comparison between both curves both 16s rRNA and tuf associated curves were normalized with the standard and the negative controls. Amplification curves are separated in different plots by sample (plots 1-10H and 1-10ID).


H7
H9
H8
H10
IC1
IC3
IC2
IC4
Fluorescence (%)
0.0
0.2
0.4
0.6
0.8
1.0
Amplification Cycle
0
10
20
30
40
16s rRNA gene
Tuf gene

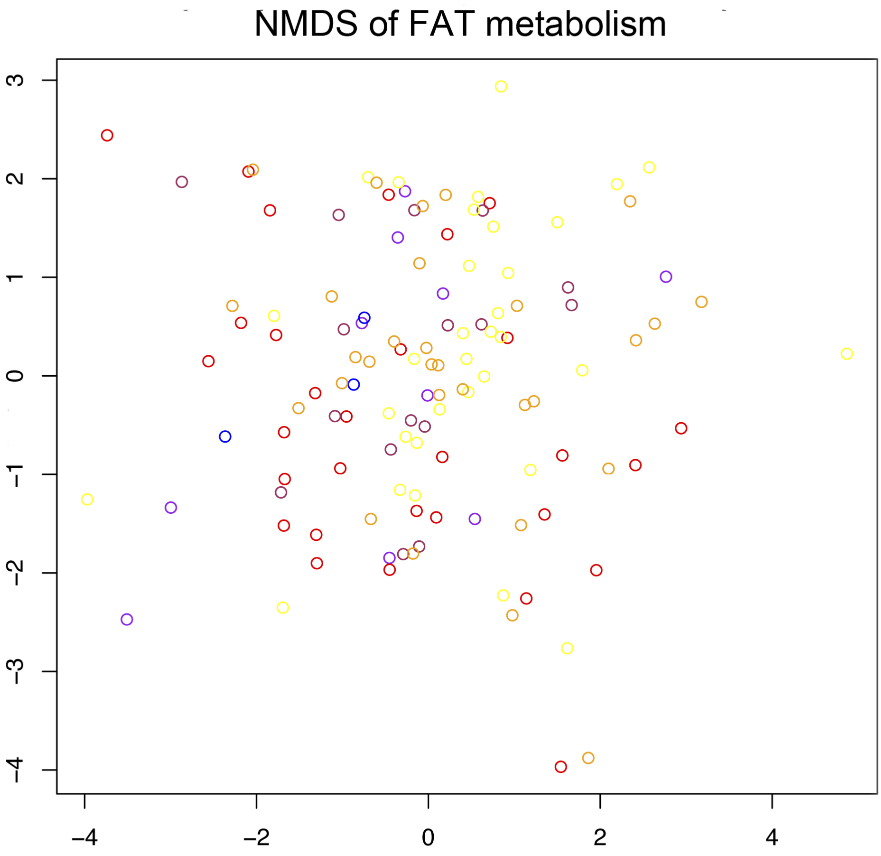

IC5
IC7
IC6
IC8
Fluorescence (%)
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Amplification Cycle
🟦 16s rRNA gene
🟥 *Tuf* gene

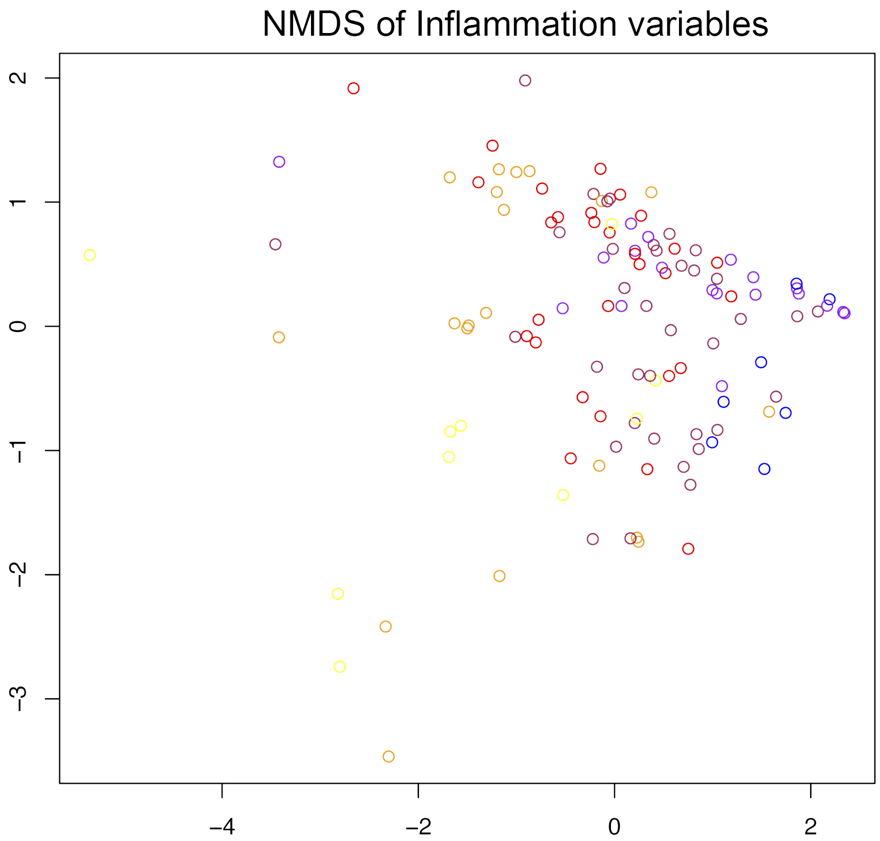

IC9
H1
IC10
H2
Fluorescence (%)
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Amplification Cycle
🟦 16s rRNA gene
🟥 *Tuf* gene

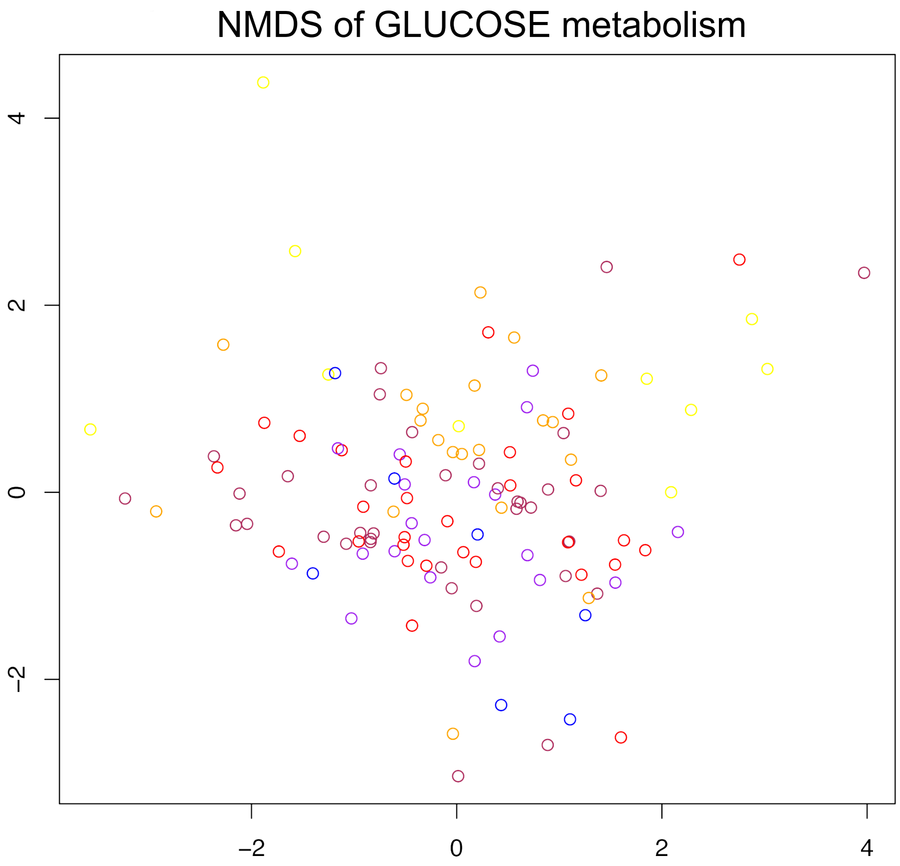
Chapter 4. Meta-metagenomic analysis of gut microbiota and overall health status .

ID	health threshold	category
cholesterol	5	MET
glucose	7	MET
Body Mass Index	25	MET
systolic blood pressure	120	MET
diastolic blood pressure	80	MET
TNF-alpha	32,5	INF
IL-6	12,1	INF
C-reactive Protein	10	INF
IL-1Ra	400	INF
basophiles	0,003	INF
eosinophiles	0,5	INF
lymphocytes	3,9	*INF*
neutrophiles	5,4	INF
mononuclear cells	0,8	INF
free fatty acids	NA	NA
Alanin transaminase	56	MET
Aspartate transaminase	38	MET
leptin	10	MET
fat-percentage	NA	NA
hba1c	5,5	MET
hemoglobin	17	MET
hdl cholesterol	2,2	MET
ldl cholesterol	3	MET
triglycerides	1,7	MET
insulin	90	MET
gender	NA	NA
diabetes	NA	NA
white blood cell count	9	INF
platelets	350	NA
shannon diversity index	NA	NA
serum Amyloid a	325	MET
Adiponectin	10	MET
asting Induced Adipocyte Fact	NA	MET
CPEP	NA	NA

Supplementary Table 1. Host-related variables included on the study.
Table shows the clinical threshold, that separates normality from alteration, in the second column. Units depend on the variable. Information retrieved by the clinical analysis department of the Ghant General Hospital. Variables are divided in two groups; metabolism and inflammation.


NMDS of FAT metabolism
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4


NMDS of Inflammation variables
2
1
0
−1
−2
−3
−4
−2
0
2



Supplementary Figures 1-3. NMDS of variable subcategories.
Rank was constructed from the subcategory-assigned variables. Correlation between the subcategory score and the global health state was performed to assign the colors.

PHYLUM

Rank	Feature name	p-value	q-value	Rho	percentage
1	?	8,165E-06	8,17E-05	0,232	58,706
2	Bacteroidetes	1,942E-05	9,71E-05	-0,223	27,112
3	Euryarchaeota	7,700E-03	0,0258	0,140	0,065
4	Fusobacteria	0,048	0,1205	-0,104	0,063
5	Proteobacteria	0,065	0,1315	-0,097	0,978
6	Actinobacteria	0,092	0,1537	-0,089	0,564
7	Viruses	0,169	0,2303	0,072	0,000
8	Verrucomicrobia	0,184	0,2303	0,070	0,485
9	Firmicutes	0,534	0,5938	-0,033	12,027
10	Cyanobacteria	0,984	0,9846	0,001	2,25E-06

CLASS

Rank	Feature name	p-value	q-value	Rho	Percentage
1	?	8,19E-06	0,0001	0,2323	58,7049
2	Bacteroidia	1,92E-05	0,0001	-0,2228	27,1121
3	Bacilli	2,52E-05	0,0001	-0,2197	0,2433
4	Betaproteobacteria	0,0020	0,0086	-0,1617	0,1857
5	Methanobacteria	0,0059	0,0202	0,1443	0,0651
6	Fusobacteria(class)	0,0507	0,1438	-0,1026	0,0631
7	Viruses	0,0620	0,1506	0,0981	0,0001
8	Actinobacteria (class	0,0899	0,1911	-0,0891	0,5645
9	Sphingobacteria	0,1740	0,2987	0,0715	0,0007
10	Verrucomicrobiae	0,1757	0,2987	0,0712	0,4854

FAMILY

Rank	Feature name	p-value	q-value	Rho	Percentage
1	?	8,42E-06	0,0004	0,2320	58,7061
2	Lactobacillaceae	0,0001	0,0024	-0,2031	0,1311
3	Sutterellaceae	0,0003	0,0041	-0,1910	0,1216
4	Streptococcaceae	0,0014	0,0162	-0,1675	0,1088
5	Methanobacteriacea	0,0072	0,0678	0,1409	0,0650
6	Bacteroidaceae	0,0141	0,1105	-0,1288	21,1314
7	Leuconostocaceae	0,0241	0,1619	-0,1184	0,0001
8	Actinomycetaceae	0,0357	0,1955	0,1103	0,0013
9	Ruminococcaceae	0,0379	0,1955	-0,1090	2,4850
10	Prevotellaceae	0,0429	0,1955	-0,1063	3,0594

GENUS

Rank	Feature name	p-value	q-value	Rho	Percentage
1	?	8,40E-06	0,0008	0,2321	58,7054
2	Lactobacillus	7,56E-05	0,0036	-0,2066	0,1310
3	Sutterella	0,0003	0,0081	-0,1911	0,1216
4	Streptococcus	0,0012	0,0289	-0,1692	0,1088
5	Butyrivibrio	0,0026	0,0404	0,1579	1,2871

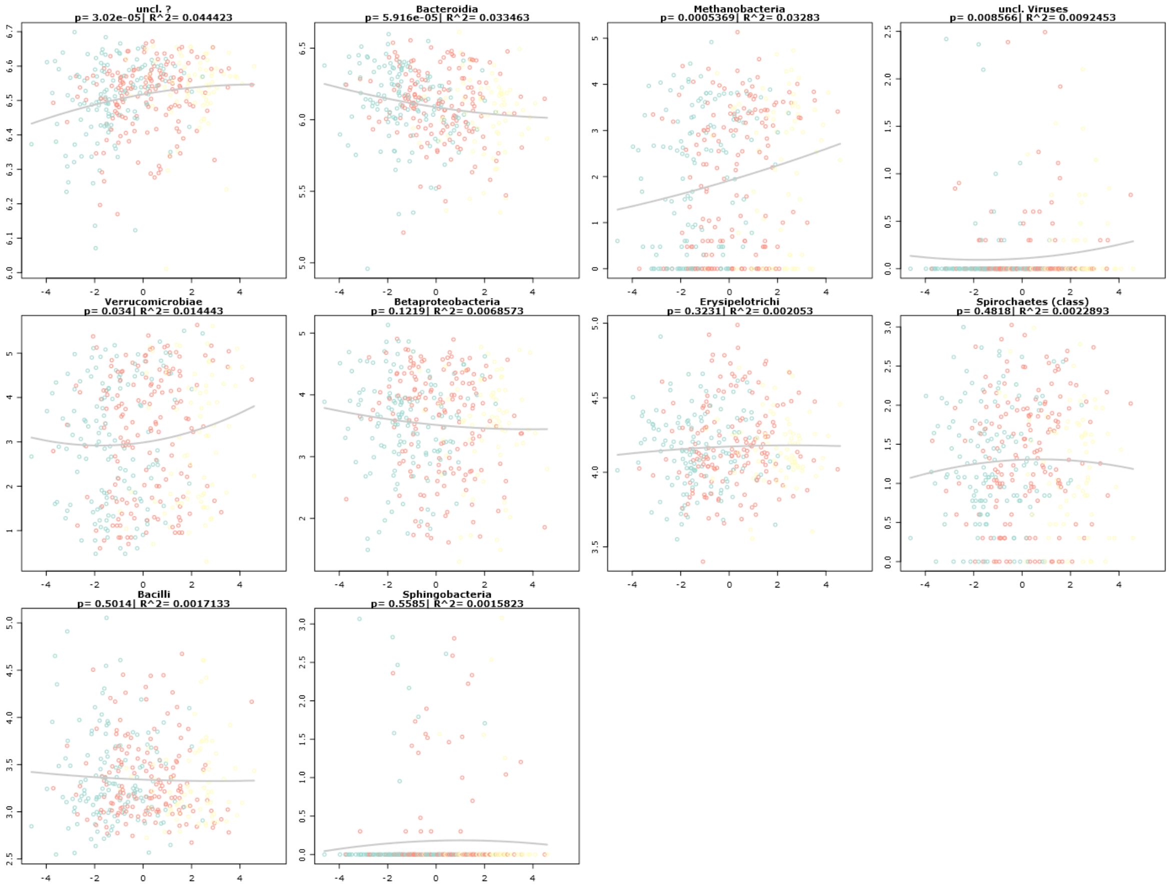
Rank	Feature name	p-value	q-value	Rho	Percentage
6	Oribacterium	0,0026	0,0404	0,1579	0,0047
7	Solobacterium	0,0030	0,0407	-0,1553	0,0195
8	Roseburia	0,0047	0,0552	-0,1482	1,4543
9	Acidaminococcus	0,0054	0,0561	-0,1459	0,0628
10	Subdoligranulum	0,0060	0,0561	0,1442	0,0624
11	Methanobrevibacter	0,0071	0,0609	0,1411	0,0651
12	Bacteroides	0,0139	0,1091	-0,1290	21,1313
13	Myoviridae	0,0198	0,1355	0,1223	9,24E-05
14	Desulfovibrio	0,0202	0,1355	0,1219	0,0435
15	Eggerthella	0,0450	0,2406	0,1053	0,0062
16	Prevotella	0,0462	0,2406	-0,1047	3,0599
17	Ruminococcus	0,0489	0,2406	-0,1035	0,9291
18	Blautia	0,0495	0,2406	-0,1032	0,0789
19	Bilophila	0,0511	0,2406	-0,1025	0,1378
20	Oxalobacter	0,0548	0,2406	0,1009	0,0006

SPECIES

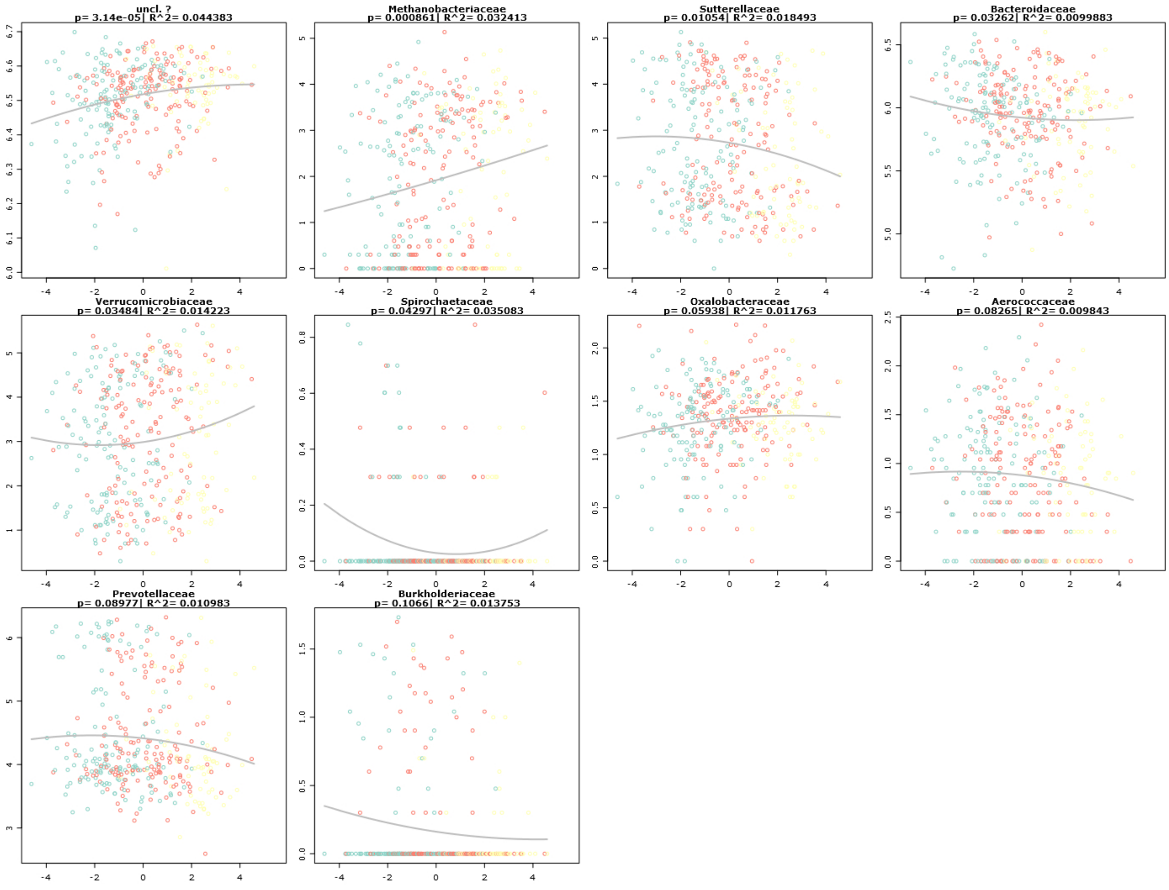
Rank	Feature name	p-value	q-value	Rho	Percentage
1	Bifidobacterium bifidu	8,65E-06	0,0013	-0,2317	0,0614
2	?	8,69E-06	0,0013	0,2317	58,7047
3	Ruminococcus torque	3,14E-05	0,0032	-0,2172	0,0769
4	Lactobacillus sakei	0,0002	0,0115	0,1980	2,16E-05
5	Sutterella wadsworthe	0,0003	0,0163	-0,1905	0,1218
6	Lactobacillus reuteri	0,0004	0,0181	-0,1844	0,0518
7	Bifidobacterium dentiu	0,0004	0,0181	-0,1837	0,0607
8	Roseburia inulinivorans	0,0005	0,0181	-0,1827	0,6145
9	Lachnospiraceae bac	0,0006	0,0182	-0,1796	0,0714
10	Streptococcus salivari	0,0006	0,0182	-0,1795	0,0408
11	Clostridium bolteae	0,0007	0,0182	-0,1782	0,1438
12	Streptococcus parasa	0,0014	0,0364	-0,1669	0,0355
13	Desulfovibrio sp, 3_1_s	0,0016	0,0364	0,1657	0,0036
14	Mobiluncus mulieris	0,0019	0,0410	0,1628	0,0007
15	Oribacterium sp, oral t	0,0024	0,0495	0,1587	0,0046
16	Butyrivibrio crossotus	0,0026	0,0497	0,1577	1,2871
17	Lactobacillus oris	0,0032	0,0567	0,1544	0,0002
18	Solobacterium moorei	0,0034	0,0567	-0,1537	0,0195
19	Streptococcus angino	0,0038	0,0603	-0,1518	0,0005
20	Lactobacillus fermentu	0,0041	0,0606	-0,1506	0,0292
21	Acidaminococcus sp,	0,0042	0,0606	-0,1501	0,0534
22	Blautia hansenii	0,0048	0,0664	-0,1478	0,0702

Supplementary tables 2-6. Spearman correlation between health score and taxonomic category.
Color rows show significantly associated taxonomic categories after FDR. In pink, it is shown the unknown category.

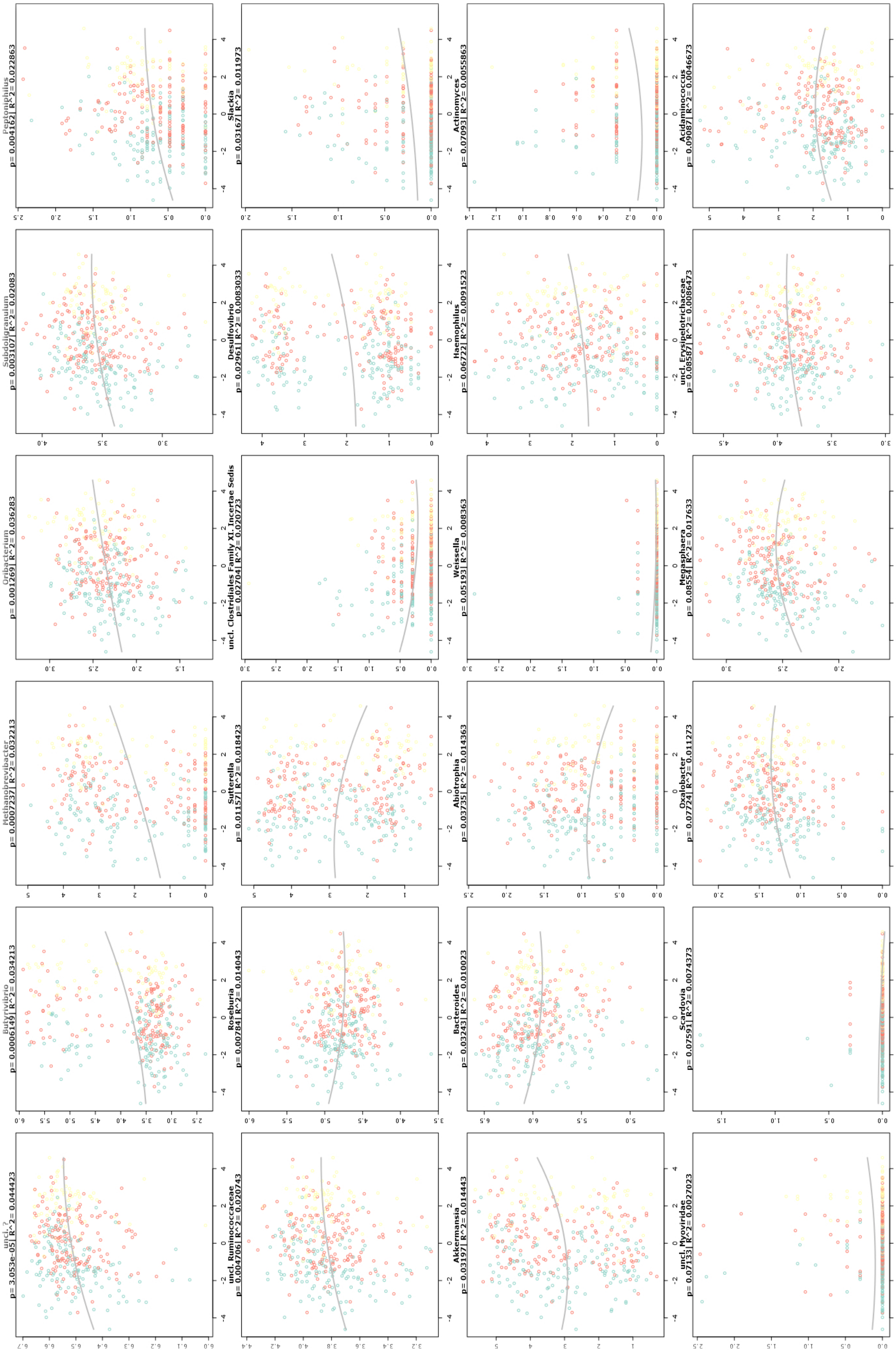
CLASS



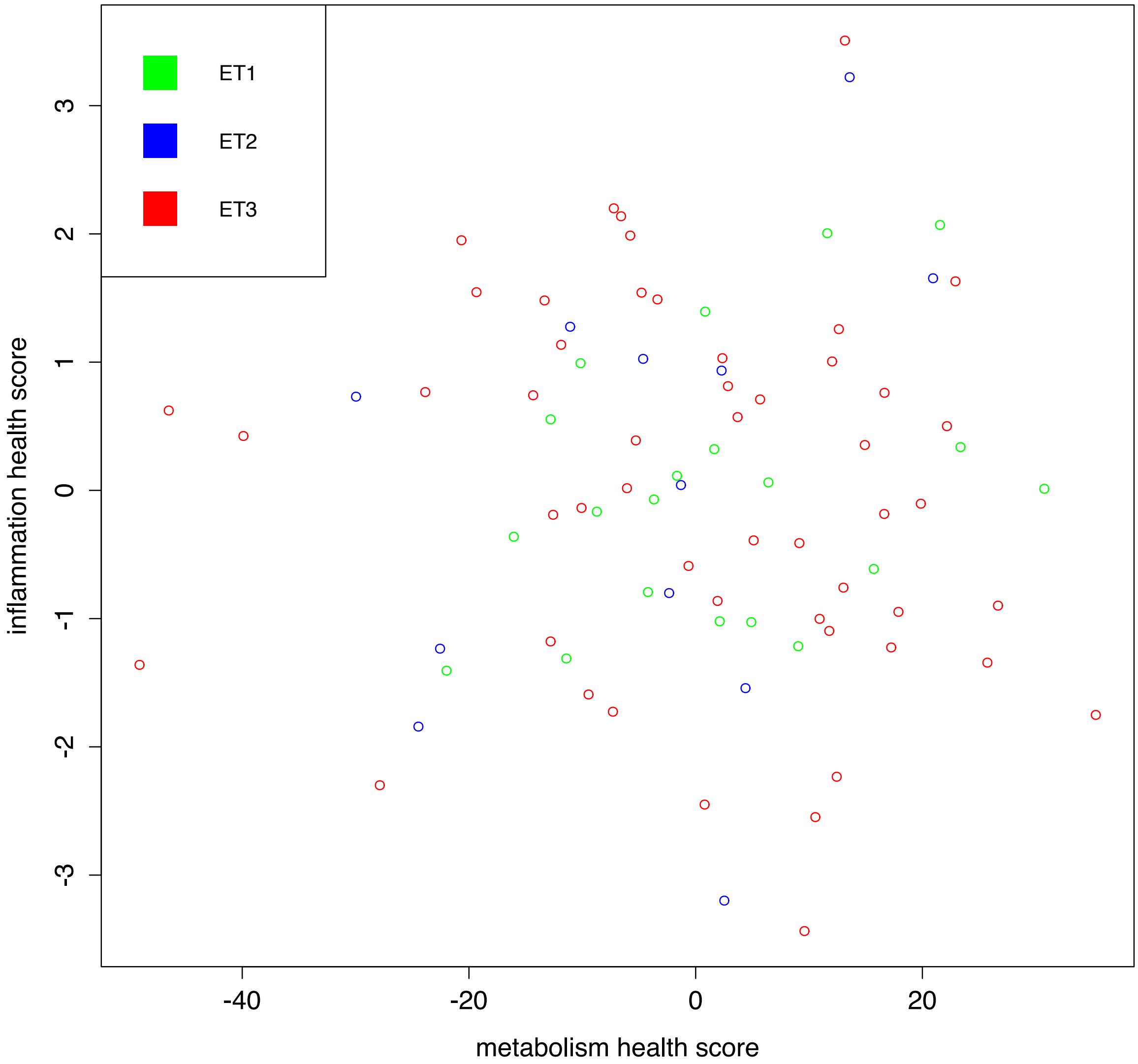
FAMILY



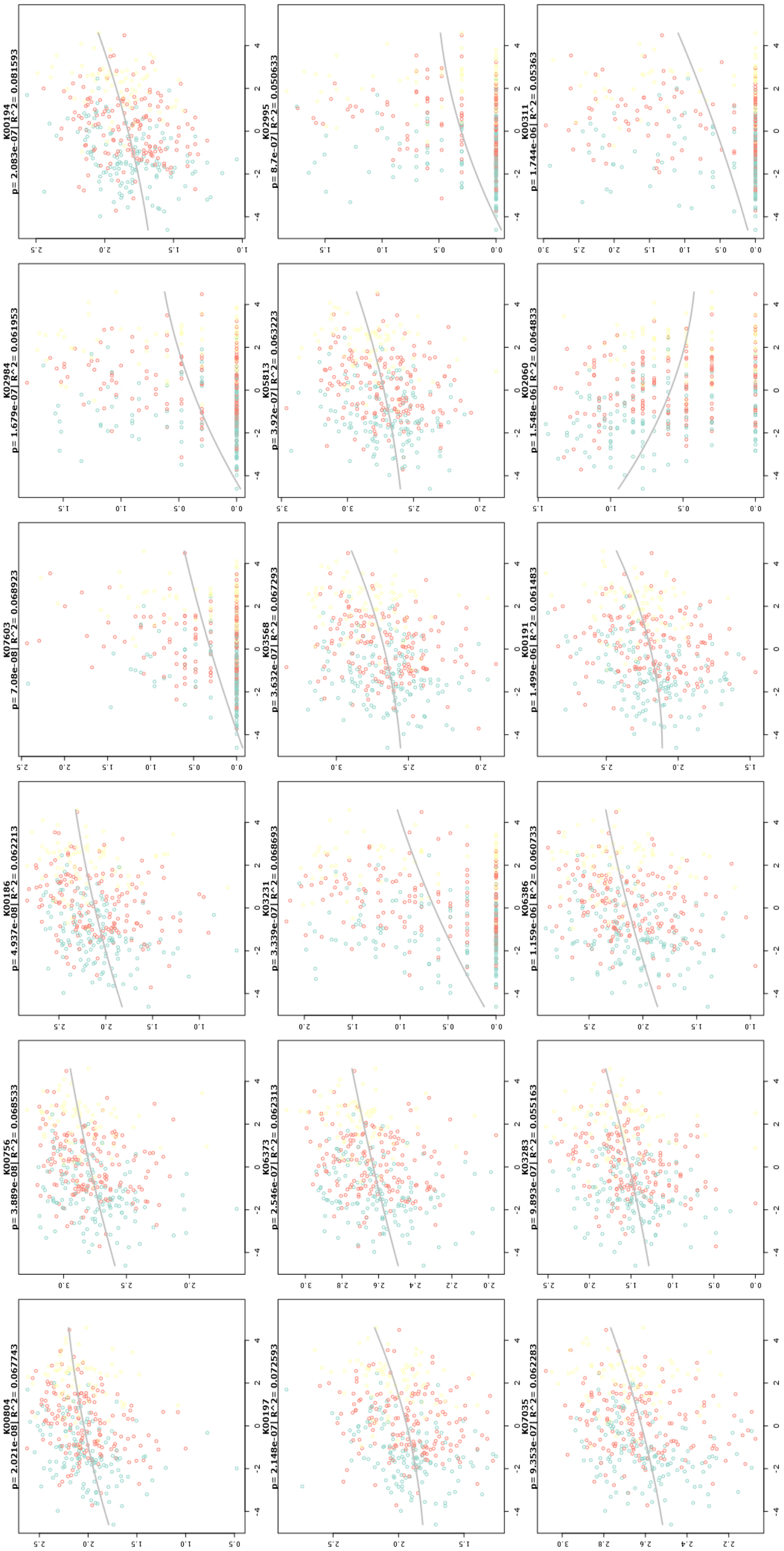
GENUS

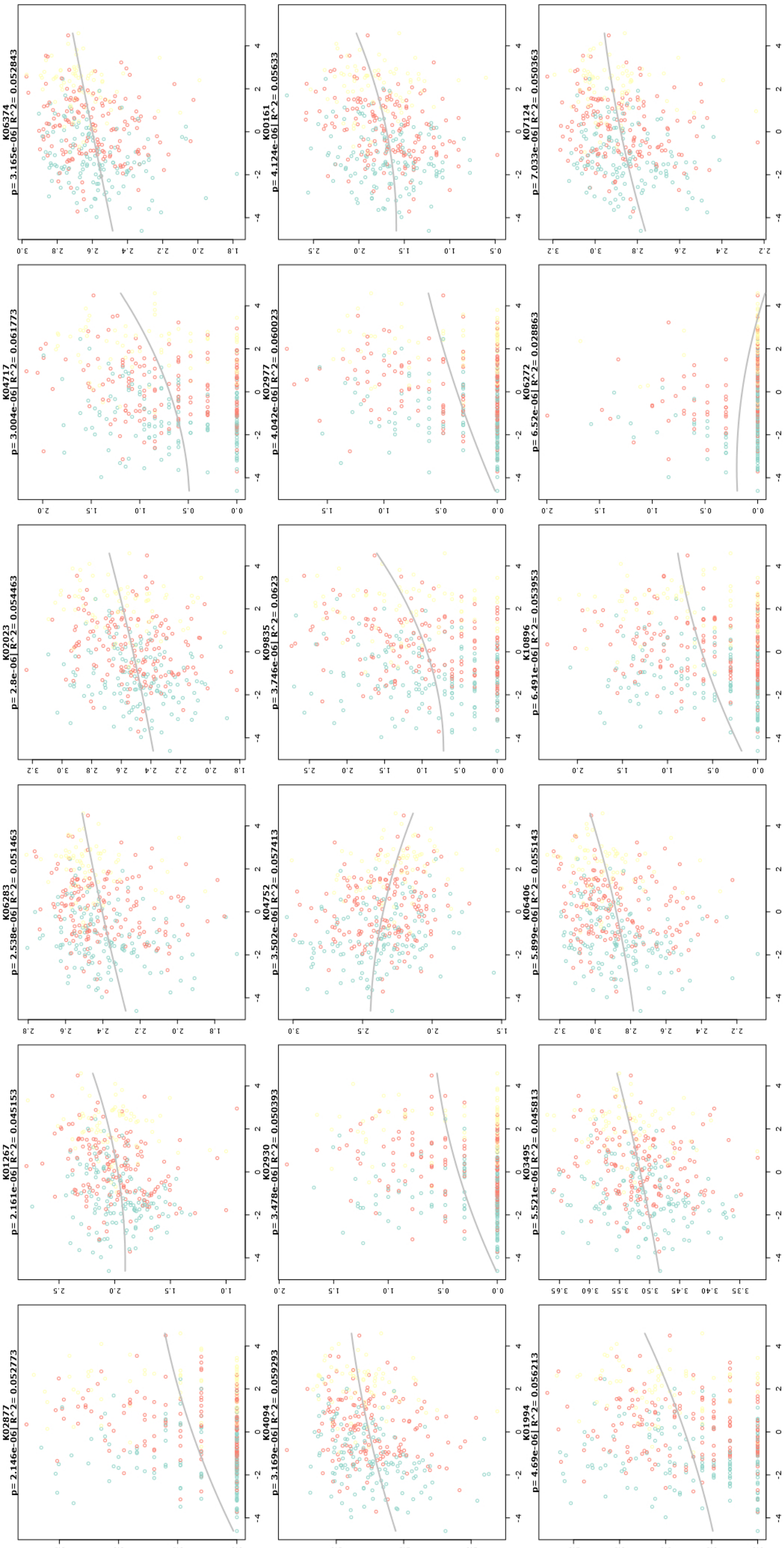


Supplementary figures 4-6. Spearman correlation of health state with taxonomic category.
See CD content for real size figures



Supplementary Figure 7. Relationship between enterotypes and health score
Health score was split by inflammation and metabolism to assess possible relationship with the enterotype classification in the low BMI subpopulation.
Colors are associated with the three enterotypes presented by Arumugam et al. 2011.


K00194
p= 2.083e-07| R^2= 0.081593
K02995
p= 8.7e-07| R^2= 0.050633
K00311
p= 1.744e-06| R^2= 0.05363
K02984
p= 1.679e-07| R^2= 0.061953
K05813
p= 3.92e-07| R^2= 0.063223
K02060
p= 1.548e-06| R^2= 0.064833
K07603
p= 7.08e-08| R^2= 0.068923
K03568
p= 3.632e-07| R^2= 0.067293
K00191
p= 1.499e-06| R^2= 0.061483
K00186
p= 4.937e-08| R^2= 0.062213
K03231
p= 3.339e-07| R^2= 0.068693
K06386
p= 1.159e-06| R^2= 0.060733
K00756
p= 3.889e-08| R^2= 0.068533
K06373
p= 2.546e-07| R^2= 0.062313
K03283
p= 9.893e-07| R^2= 0.055163
K00804
p= 2.021e-08| R^2= 0.067743
K00197
p= 2.148e-07| R^2= 0.072593
K07035
p= 9.353e-07| R^2= 0.062283



Supplementary Figure 8. Spearman Correlation of Kegg Othologous categories and health score