I – EXECUTIVE SUMMARY[1]

This synthesis, sketched during the course of the workshop proper, was developed in the months thereafter on the basis of written contributions provided by most participants under Laura Giuliano's coordination. Frédéric Briand reviewed and edited the entire Monograph. Céline Barrier was responsible for the physical production of the volume.

1. INTRODUCTION

Reports of infectious diseases affecting humans and marine organisms are more and more frequent. Whether these increases reflect better reporting or actual global trends is a subject of active research, which is receiving much attention worldwide, given the heightened human dependence on marine environments. Bacterial infections, in particular, figure among the emerging threats to human health, especially in heavily polluted coastal areas where they are associated with recreational and commercial uses of marine resources (Tamplin, 2001).

In addition evidence is mounting of an increased sensitivity of various marine organisms to infectious agents, leading to the occurrence of opportunistic pathogens, and/or amplification of resident infectious agents. Various research studies, for example, refer to ecologically and economically important species from the oceans, such as oysters and corals, which have been affected by large-scale epidemics. Increasing human pressure on marine ecosystems and ongoing climate change / warming are widely believed to further foster the spread of pathogens in the sea (Vezzulli *et al.*, 2016).

Environmental and climatic conditions usually play a major role in the distribution of indigenous pathogenic microorganisms in the marine environment and in their transmission to humans and animals, with less certainty in the case of *Vibrio cholerae* (Domman *et al.*, 2017). Beyond the (re)-emergence of indigenous pathogens (e.g. *Vibrio* spp.), the introduction of allochtonous pathogens by agricultural and urban runoff, ballast water or animal transfer, can be the cause of new infectious diseases with severity depending on the virulence, ecology and survival of the infectious agent.

Large-scale intensive aquaculture practices further contribute to the dramatic increase in severe disease outbreaks caused by a diverse range of pathogens, including parasites, viruses and bacteria (Soto-Rodriguez *et al.*, 2015; Sundberg *et al.*, 2016).

Today, 'omics approaches (metagenomics, metatranscriptomics, metaproteomics) provide great promise for a better understanding of marine microbial communities, including marine pathogens (Coyle *et al.*, this volume). By analysing the whole environmental DNA (eDNA), potential pathogens can be monitored for early detection and management. Novel next generation sequencing techniques allow the detection of DNA sequences even at very low concentrations, and through existing reference databases the sequences can be used to identify the presence of pathogenic microorganisms, their

[1] to be cited as :

Giuliano L., Dorman C., Bowler C., Sugiyama M., Vezzulli L., Czerucka D., Le Roux F., D'Auria G., Troussellier M. and F. Briand. 2017. Searching for bacterial pathogens in the Digital Ocean - Executive Summary. pp. 5 - 25 in CIESM Workshop Monograph n°49 [F. Briand, ed.] 158 p. CIESM Publisher, Monaco and Paris.

pathogenic potential, and mechanisms of evolution. Third generation, portable real-time sequencing devices are now available for genome sequencing of bacterial strains in the field (Bleidorn, 2016).

With billions of 'omics data already available in the public repositories (see sections below) most pathogenic microorganisms will be discovered and characterized in the future by the analysis of sequences, with an ongoing shift from molecular barcoding towards metagenomics and metatranscriptomics (Pallen, 2016). No less than 40 million novel genes were predicted from the recent Tara Oceans expedition alone (Suganawa *et al.,* 2015), and yet the molecular mechanisms of virulence of many environmental pathogens remain unknown. Even if recent metagenomic analyses have revealed that putative virulence genes are widespread in the ocean, drawing conclusions about the role of virulence genes in the absence of a model of pathogenesis would be highly premature. Caution in the interpretation of such data is strongly recommended.

In any case, all trends are leading to a "digital ocean era". The large amount of digital information (e.g. genetic sequences in digital format) requires the development of new analytical tools to transform the huge amount of data into biological knowledge. Environmental bioinformatics will provide new solutions for this vibrant and exciting field of research, allowing to scale-up from the analysis of the thousands of marine genomes to the millions of metagenomes in their environmental context (see more in Coyle *et al.,* this volume).

In opening the meeting, Drs Frédéric Briand and Laura Giuliano, respectively Director General and Scientific Director of CIESM, presented the overall background and objectives of the workshop to the participants (see list at the end of volume), emphasizing the urgent need for a more complete understanding of the emergence of bacterial pathogens outbreaks and the possible role of marine ecosystems as reservoirs. In summarizing the huge complexity of the marine environment, (e.g. with respect to scales, particles density, structure, etc.), they stressed the importance of tracking the phylo-geography of bacterial pathogens of humans and animals in marine areas.

The central question that framed the discussions was 'to what extent can the digital ocean teach us something about bacterial pathogens?' As reflected in Figure 1, which attempts to capture the many processes surrounding the analysis and mining of large data sets (oceanographic, environmental, biomedical, epidemiological, etc.), this is a complex, intricate question. In order to reveal meaningful patterns of potential marine pathogens (i.e. propagation, interaction with the host, effector delivery etc.), attention must be given to large gene expression data sets as depositories of module molecular markers of pathogenicity, and to the development of recent machine-learning algorithms that will help recognise pathogen-specific fingerprints. In fact, again and again the present volume will highlight the huge power of fast growing data sets as molecular epidemiological tools for reconstructing individual transmission events and for tracking the emergence and spread of resistant and/ or virulent clones.

To explore these questions, some fifteen experts of various geographic horizons and backgrounds (marine microbial ecology, pathogenicity and its genomic signature, virulence genes, integrated genomics and post-genomics approaches, computational molecular biology, metadata management) were invited by CIESM at the Oceanographic Institute in Paris, in late September 2017.

The present summary synthesizes the outcome of the discussions and exchanges conducted both during and in the immediate aftermath of this CIESM brainstorming workshop. Considering the fast-increasing number of datasets and platforms, and the diverse available tools for integrated analyses of metadata informing on pathogens and the marine environment, this chapter reflects the complexity of such metadata depositories, and the current difficulty to clearly identify pathogens by using the known associated (molecular) traits, in particular with respect to virulence which usually involves the interaction of different genes and gene regulators, in addition to possible host effects. The conclusions, while preliminary, are quite new in this field.

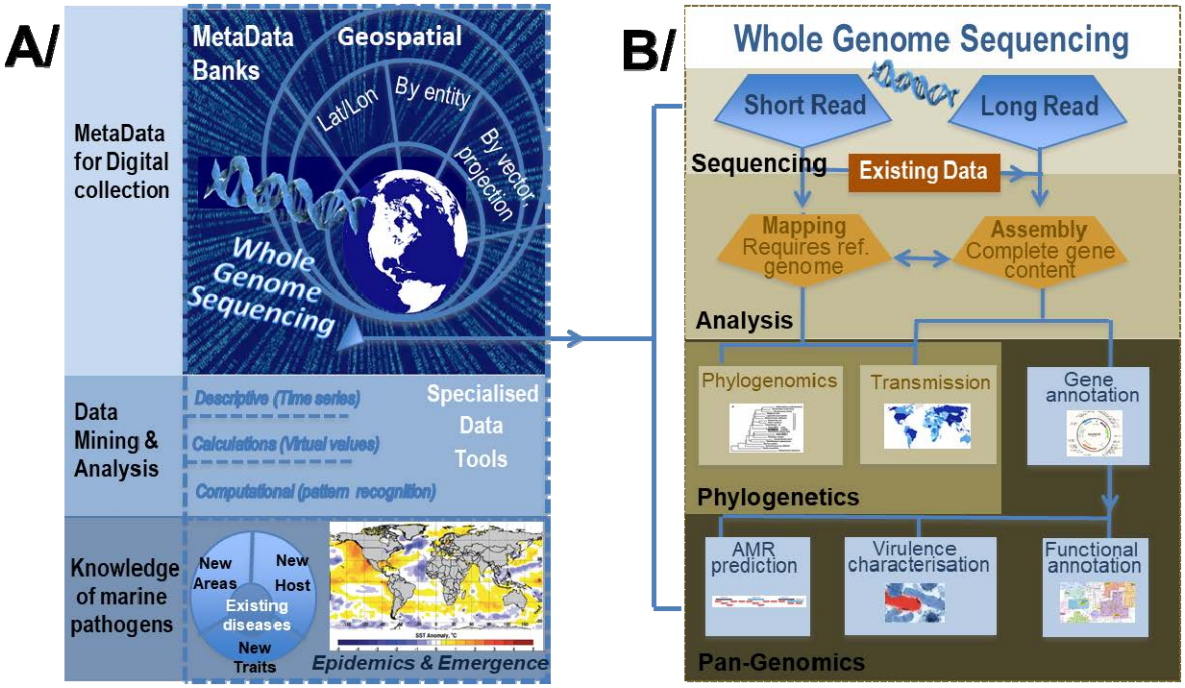


Figure 1: **Decrypting the Digital Ocean – Main Steps.**
A/ Processing information from Metadata banks to knowledge of the distribution, behaviour and epidemiology of marine pathogenic bacteria. **B/** *(*Adapted from Bayliss *et al.*, 2017) Zoom on the treatment of molecular data obtained from the whole genome sequencing of pathogens.

2. IDENTIFICATION OF USEFUL VARIABLES / MARKERS

In the mind of the public, bacteria can be divided into the 'good' and the 'bad', with the bad bacteria equating to pathogens. The potential for a microorganism to cause an infectious disease is influenced by several factors. A bacterium possessing genes that encode known pathogenic traits will have an increased probability to be pathogenic, but realising its potential for harm will also depend on environmental factors and on the physiological state of the host. The condition of the host, i.e. its ability to resist infection, may broaden the range of bacteria to which that host will be susceptible. In extreme cases, such as that of an immuno-compromised host, even normally benign ('good') bacteria may present a significant threat to health. For all these reasons, the development of rapid diagnostic methods with suitable sensitivity and specificity is very complex and requires multi-disciplinary, integrated efforts (Bruto *et al.*, this volume).

Pathogenic bacteria usually maintain their virulence genes under some form of control at the level of expression to avoid wasteful production of traits that impose a fitness cost on the organism if an advantage is not achieved. In general, two strategies can be observed at work in a bacterial population. The first involves a stereotypical response in which all, or almost all, the bacteria respond to a signal that a suitable host is present by activating their virulence genes. The second involves the operation of randomly-acting genetic switches that switch virulence genes on or off in a stochastic manner across the bacterial population. In this way, just a subset of the bacteria will be primed to infect if a host appears. In practice, both strategies can be detected in individual pathogens. These strategies imply that bacteria possess the machinery to interpret their environments and to respond accordingly. Are environmental sensing systems also virulence traits? It could be argued that without them virulence mechanisms would lack 'intelligence' and be much less effective. Which illustrates the importance of looking beyond classical virulence genes when making assessments of the pathogenic potential of bacteria (Dorman, 1994).

The genetic switches responsible for controlling the expression of bacterial virulence genes can be indistinguishable from those that control housekeeping genes, but there are exceptions that are frequently associated with genes involved in disease. These are the bi-phasic switches that govern gene expression stochastically through mechanisms that involve DNA segment inversions, DNA methylation and point mutations within poly-pyrimidine tracts (Dorman and Bogue, 2016). It is thought that this phase-variable method of gene expression allows pathogens to evade the host defences by presenting the immune system with novel antigens that it is not primed to recognise (Moxon *et al.,* 2006). Each of these systems has characteristic features that can be identified by interrogating bacterial genomes. For example, DNA inversion requires a site-specific recombination system that operates on inverted repeat sequences; DNA methylation by the Dam methylase requires GATC sequences that are associated with transcription signals and poly-pyrimidine tracts produce frame-shift mutations in open reading frames that alter the repertoire of surface proteins expressed by pathogens. All of these features are discoverable by genome sequence analysis (Chen *et al.,* 2014). Typically, the genes affected encode secreted proteins or cell surface components and will have protein secretion signals, which are also detectable bioinformatically.

Gene regulation using 'conventional' transcription factors is common among virulence genes that are subject to stereotypic control. These DNA binding proteins are divided into classes based on their domain structure and can be studied bioinformatically. Examples include the AraC-like proteins, the LysR-like proteins and members of the sensor-kinase/response regulator superfamily (Maddocks and Oynston, 2008; Yang *et al.,* 2011). Almost every major virulence gene control network contains examples of these transcription-controlling proteins. In many cases the binding sites of the proteins can also be detected bioinformatically. The proteins have signal reception domains that bind small molecules or metals; the response regulators are phosphorylated by the sensor kinases on conserved aspartic acid residues, making the proteins proficient for DNA binding (Dorman and Dorman, 2017).

DNA base composition can be a clue to the presence of virulence genes in the genome. Abnormal A+T base content may indicate that the genes in question have arrived by horizontal transfer and have been maintained because they added some advantageous features (possibly connected with pathogenicity) to the receptor organism. The link is quite reliable based on the studies done so far. While not every horizontally-acquired gene is a virulence gene, most virulence genes have been acquired horizontally (Gyles and Boerlin, 2014). These unusual DNA structural properties have implications for gene regulation, the choice of sigma factor to operate RNA polymerase and the likelihood that the genes will be targeted by transcription silencing nucleoid-associated proteins. Other clues that a portion of the genome has been acquired by lateral transfer include the presence nearby of phage attachment sites and/or genes (or pseudogenes) encoding integrases or transposases.

The expression of many virulence genes is controlled by small RNAs (sRNA) and the sRNA-mRNA interaction frequently requires a chaperone protein (Holmqvist *et al.,* 2016). Small RNAs can be discovered by whole genome analysis methods, as can potential RNA chaperone proteins (Barquist *et al.,* 2016). RNA control that acts *in cis* can also be predicted bioinformatically because it frequently relies on differential transcription termination that leads to alternative folding of the transcript.

The creation of physiological variety across a bacterial population is an excellent strategy for survival because it reduces the risk that all members of the population will be eliminated by a single catastrophe. The arrival of a powerful antibiotic in the midst of a population of susceptible bacteria is an example of such a catastrophe. However, if the action of the antibiotic is restricted to periods when the bacterium is growing, any non-growing organisms will escape death or inhibition. Persistence describes the ability of bacterial populations to bring forward metabolically inert members in a completely stochastic manner. These non-growing organisms are temporarily non-susceptible to antibiotic action and will be capable of carrying on the life of the population after the antibiotic is removed. Persistence mechanisms often involve a 'toxin/'anti-toxin' binary system in which the toxin component inhibits growth following the stochastic disappearance of the (usually-unstable) anti-toxin (Harms *et al.,* 2016). The presence of such systems can be an aid to virulence and may underlie recurrent infections that appear to be recalcitrant to antibiotic therapy. Persistence is distinct from antibiotic resistance, which involves mechanisms by which the bacterium becomes permanently

resistant through modification of the drug target, over-expression of the target, inactivation of the drug, or an ability to pump the drug out of the cell. Nevertheless, the detection of resistance genes in association with classical virulence traits may be indicative of enhanced pathogenic potential. Similarly, the presence of persistence systems could indicate an ability on the part of the bacterium to outwit host defences that rely on killing metabolically active bacteria (Maisonneuve and Gerdes, 2014).

Each infection is characterised by a minimum infectious dose. The fewer organisms needed to initiate a successful infection in a healthy host, the more virulent that pathogen will be. Do bacteria 'count' one another to determine if they have reached a threshold at which initiating an infection is worthwhile? Data from studies of bacterial cell-to-cell signalling suggest that they do (Papenfort and Bassler, 2016). Signal production and signal detection form the basis of quorum sensing. Here, small molecules are produced by individual bacteria in concentrations too low to elicit a response, but when pooled with the same molecules from other nearby bacteria a threshold is crossed that leads to the elaboration of a new behaviour, perhaps directly involved in the infection of a host. The genes that encode the enzymes of the pathways for expression of the signalling molecules and the genes that encode the receptors that detect those molecules can all contribute to the pathogenic potential of the bacterium.

3. PATHOGENIC TRAITS

Bacteria often engage on pathogenic behaviour in order to win new resources or to escape from an unfavourable environment. As to marine environments, depending on the environmental conditions, bacteria can move in a free individual manner or remain in the same place to form colony groups and colonize surfaces. As a group, bacteria can optimize growth and survival by the presence of different cell types that are able to perform specialized functions (i.e. better access to nutrients; better defence mechanisms for protection against unfavourable environmental conditions, etc.). Some bacteria can secrete polysaccharides to form biofilms which enhance adhesion, survival, and movement. The main pathogenic factors associated with such functions are surface polysaccharides (capsule, lipopolysaccharide, and glucan), S-layers, iron-binding systems, exotoxins and extracellular enzymes, secretion systems, fimbriae and other nonfilamentous adhesins, motility and flagella. An integrated compilation of bacterial pathogenic traits in *Aeromonas* spp. is provided by Tomas (2012).

Many virulence genes are switched on in response to a shortage of iron. Although this is the most abundant metal on earth, iron is largely unavailable to biology and is strongly sequestered by living organisms. Bacteria must therefore compete with their hosts to acquire it. Many successful bacteria manufacture iron-carrying molecules (siderophores) that have a higher affinity for iron than those of their customary hosts. They also have efficient transport systems to bring the iron-siderophore complexes into their cytoplasm where the metal can be released and used, for example to create active centres in proteins involved in electron transport in the respiratory chain (Palmer and Skaar, 2016). Iron acquisition genes can be found in bacterial chromosomes but they are also found on plasmids, including plasmids that harbour virulence genes. Detection of iron uptake genes does not in itself prove that a bacterium is a pathogen but their association with other disease-associated genes may reveal pathogenic potential.

According to the scientific literature, temperature increases in oceans can stimulate opportunistic pathogens and favour waterborne disease outbreaks, including sometimes the “reverse zoonosis”, namely the transmission of human pathogens to marine organisms. Among many examples figure various coral disease outbreaks (Ben-Haim *et al.,* 2003; Sutherland *et al.,* 2011), which have been also observed in Mediterranean waters (Rubio-Portillo *et al.,* 2014).

3.1 Genes / proteins with clinical relevance

Secretion systems are required for the export of virulence factors, including toxins (Mecsas and Strauss, 1996). Toxin genes can be found in isolation or as part of operons that include genes for toxin subunits or other virulence traits. Adhesins are essential for the colonisation of surfaces and for

assembling the bacteria into communities. Matrix-binding proteins allow bacteria to adhere tightly to the surfaces of host cells on the exterior or in the interior of the body (Chagnot *et al.*, 2012). Biofilm protects the bacterial community from the environment – aids colonisation of rocks in watercourses, on the seashore, the linings of pipes in the plumbing in buildings, the surface of teeth, etc. Biofilm glues microbes together and helps them resist shear forces or other forces that might wash them away. In some cases, flagellar motion provides bacteria with the capacity to avoid being trapped by the host cell physical defences (i.e., epithelial mucus in the gastrointestinal tract; see Czerucka and Peruani, this volume). Capsules mask bacterial surface antigens and help the bacteria to evade the host immune system. Capsules also defend the bacteria from desiccation and from thermal stress. Colanic acid is produced by many bacteria in response to low temperature and osmotic stress. The drying action associated with osmotic stress results in bacterial accumulation of compatible solutes that replaces lost water. All of this relies on a system for sensing and responding to osmotic stress. The list of known virulence factors of *V. parahaemolyticus* is provided below as an example (Table 1).

Table1: List of known V. parahaemolyticus virulence factors

	Name	Domain	Activity	Function
Toxins/adhesion	**TDH**	Thermostable directed hemolysin	Pore forming toxin	Cytotoxicity & enterotoxicity
	TRH	TDH related hemolysin	Pore forming toxin	Cytotoxicity & enterotoxicity
	MAM7	mce domain	Binds to fibronectin & phospho- lipid / atidic acid	Attachment to the host cell
T3SS1 effectors	VopQVP1680	Non conserved	Binds to V-ATPase	Autophagy induction
	VopSVP1686	Fic domain	AMPylates Rho family GTPases	Cytoskeleton disruption
	VPA450	Inositol 5-phosphate	Hydrolyzes PI(4,5)P2 to PIP4	Plasma & membrane disruption
T3SS2 effectors	VopCVPA1321	Rac & CDC42 deamination	Cytotoxic necrozing factor	Disturbs actin, bact. invasion
	VopTVPA1346	ADP/ribosyltransfase	ADP-ribosylates Ras	Unknown
	VopA/T	Acetyltransferase	Inhibits MAPK signaling	Imm-resp.suppress.
	VopVVPA1357	Non-conserved	Actin binding / bundling	Cytotoxicity & enterotoxicity
	VopLVPA1370	WH2 domain	Actin nucleation	Induction of actin stress fiber

3.2 Quorum sensing

What are the genes involved in the synthesis of the quorum-sensing molecule, the degradation of the quorum sensing molecule and the detection of such molecules? Are the molecules confined to intra-species signalling or can they participate in inter-species communication? What behaviour might be controlled by the signal? Some influence cell-to-cell transfer of DNA, others control the expression of biofilm material, while others regulate the expression of bioluminescence. Many bacteria have integrated their virulence genes into control networks that have a role in quorum sensing (e.g. *Pseudomonas*).

3.3 Whole metabolic pathways

The fewer complete metabolic pathways a microbe has, the more dependent it is on its host or other environments for survival. *Mycoplasmas* spp. have a very small genome and are obligate parasites of humans and other animals. *Mycobacterium tuberculosis* has many genome gaps that have occurred over the period it has lived in intimate association with cattle and humans. Virulence factors are often produced in branches of central metabolic pathways. For example, many siderophores contain a ring structure that is derived from chorismate, an intermediate in the same pathway that produces para-amino benzoic acid (pABA), folate and the aromatic amino acids.

3.4 Invasion factors

Invasion factors come in many forms but share the property of being able to get a normally non-phagocytic cell to phagocytose the microbe. Many systems inject effector proteins into the cytosol of the host cell to modify its cytoskeleton so that the surface envelops the microbe, bringing it inside. Pathogens can also use specialist proteins to escape from phagocytic vacuoles and to recruit host actin for locomotion across a cell and through the barriers that separate cells. They can weaken the structures that hold host epithelial cells together in tissues, facilitating the passage of the bacteria between those cells. Some pathogens (*Salmonella*) can subvert host defence cells such as macrophage, using them as a means to move between tissues in the host (Liss and Hensel, 2015). Bacteria also have the potential to induce cell death in macrophage, killing them and releasing the pathogens (Sridharan and Upton, 2014).

3.5 Competitive fitness

Competitive fitness is a very important concept in bacterial biology (von Bronk *et al.*, 2017). It compares the relative abilities of different bacteria to reproduce in a given environment. Any change to the genetic composition of the bacterium that enhances fitness may improve its virulence too, although there are often tradeoffs. For example, acquiring a plasmid that encodes a highly efficient iron uptake system may benefit competitive fitness in a low iron environment. In an anaerobic environment where iron, in its reduced Fe^{2+} state, can enter the cell without the aid of an uptake system, the burden of plasmid carriage may undermine the fitness of the bacterium. Fitness is enhanced by carrying just enough genetic capacity to survive and reproduce while ensuring that the expression of that genetic capacity is regulated very finely to ensure the optimal use of energy and resources. Sluggish responses to environmental change or wasteful operation of metabolic pathways can severely undermine fitness; so can inappropriate expression of virulence traits. Failing to avoid or evade host defences or other environmental hazards due to poor or inaccurate interpretation of the environment can prove deleterious to the bacterium. The list of virulence determinants for *Vibrio* species is provided here as an example (Table 2).

Table 2. Examples of virulence determinants for *Vibrio* species. *cheR*: Chemotaxis gene; Fla, Laf: Flagellar & lateral flagellar gene; GacA/s: two-component signal transduction system; PilA: Type IV pilin; PilD: Prepilin peptidase; CpsA: capsular polyssaccharide A; GbpA: GlcNacbinding protein A; MAM: Multivalent Adhesion molecule; Wza: exopolysaccharide genes; Orf1: insertion site of mini-Tn*5phoA* transposon ; OmpU: outer membrane protein; Apha/ OpaR: transcription regulator; T3SS1/2: type three secretion system 1 or 2; T6SS: Type six secretion system; *tdh :* thermostable direct hemolysin; *trh:* thermostable direct hemolysin *(tdh)*-related hemolysin; *hlyIII,;* VvhA, HlyA*:* hemolysin in different species; CTX: cholera toxin; *toxR*: toxin regulated system; *tcp*: toxin co-regulated pilus; RTX: exotoxin and virulence factors.

Category	Fitness factors	Species
Motility	cheR	*V. anguillarum*
Flagella	Fla, Laf	*V. parahaemolyticus*
Biofilm formation	GacA/S, pilA, pilD	*V. vulnificus*
Capsule	CpsA	*V. parahaemolyticus*
Adhesion	GbpA	*V. cholerae*
	Chitinase	*V. anguillarum*
	MAM	*V. parahaemolyticus*
	wza, orf1	*V. anguillarum*
	OmpU	*V. vulnificus, V. fisheri*
Quorum sensing	Apha	*V. parahaemolyticus*
	OpaR	*V. parahaemolyticus*
Secretion	T3SS1	*V. parahaemolyticus, V. cholera*
	T3SS2	*V. parahaemolyticus*
	T6SS	*V. parahaemolyticus*
Hemolysis	tdh	*V. parahaemolyticus*
	trh	*V. parahaemolyticus*
	hlyIII	*V. vulnificus*
	VvhA	*V. vulnificus*
	hlyA	*V. cholerae*
	HlyA	*V. tubiashii*
Toxicity	CTX, toxR, tcp	V. cholera
	RTX	V. vulnificus

3.6 Bacteriophages, integrons and mobile genetic elements

Mobile genetic elements of all kinds drive bacterial evolution over impressively short timescales (Koonin and Makarova, 2017; Wu *et al.*, 2015). They allow bacteria to sample a very wide range of genetic elements and to experiment with them. Incorporating useful ones permanently into the genome permits a bacterium to acquire sophisticated new behaviours in one step (for example, the ability to resist an antibiotic; the ability to utilise a complex carbon source that was previously beyond its physiological capacity). Many overt virulence systems such as cholera toxin (Waldor and Mekalanos, 1996) are delivered by bacteriophages. The process of lysogenic conversion' in which a bacteriophage interrupts one virulence gene physically while simultaneously delivering a new one creates variety among the virulence traits of a population of pathogens (see more in Dorman, this volume). Integrons are a means of building a repertoire of antibiotic resistance cassettes in the genome (Gillings, 2017). These collections can themselves become mobile, spreading the multi-drug-resistance phenotype through the microbial population. Plasmids harbour a wide variety of traits that are relevant to

virulence. They can be self-transmissible between bacterial cells or capable of being mobilised (Bañuelos-Vazquez *et al.,* 2017). Plasmids carry regions of DNA sequence homology with the bacterial chromosome, allowing them to become fused with the chromosome and to mediate its rearrangement and evolution, a behaviour that has important implications for the fitness of the bacterium and its ability to interact with the host (Humphrey *et al.,* 2012). Plasmid transfer rates are intimately linked to the formation of biofilm, which is itself linked to virulence. Plasmids often carry transposons and these in turn carry genes for resistance to antibiotics and heavy metals. Transposition is a key driver of genome evolution, not only causing genes to be knocked out but also causing previously silent genes to become active. Plasmids vary greatly in size: the so-called second chromosome in *Vibrio cholerae* is really a very big plasmid (Orlova *et al.,* 2017). This serves to illustrate the capacity of plasmids to absorb more and more genetic information. By being separate from the chromosome that houses the genes essential for the survival of the bacterium, plasmids can be jettisoned when selective pressure for their maintenance is absent. In practice, many large plasmids have evolved systems that kill/ inhibit bacteria that manage to lose the plasmid (Gerdes *et al.,* 2005). These toxin/anti-toxin systems have been co-opted by the bacteria and serve randomly to create persister cells in the population (see above).

3.7 CRISPR/*cas*

CRISPR/*cas* is a form of immunity that allows a bacterium to identify and destroy the DNA of an invader, such as a bacteriophage (Horvath and Barrangou, 2010). While many phages are temperate and do not kill the bacterium (at least immediately) there are virulent phages that will hijack the bacterial cell to reproduce themselves straightaway, leading to the death of the bacterium. Being able to destroy these viruses is very useful to the bacterium. Also useful is the ability to use restriction endonucleases to destroy foreign DNA and to use DNA methylation patterns to distinguish between self and non-self at the level of DNA (Loenen *et al.,* 2014). A failure to detect a working CRISPR/*cas* system in a pathogen may indicate that it is evolving at a high rate through the risky strategy of sampling a wide range of DNA from the environment.

3.8 Resistance

It is important to recall that perhaps 70% of antibiotics are made by bacteria and so the producers have to possess the means to survive their own products (Martin and Liras, 2012). These resistance genes can be exported via horizontal gene transfer throughout the microbial world (Yamashita *et al.,* 2014). Loss of competitive fitness associated with the carriage of a resistance gene that is currently not under selection represents a force that limits the process. Nowadays, widespread pollution of the environment with low concentrations of antimicrobials of all kinds is imposing just this type of selective pressure and is driving the spread of resistance genes. Pathogens are dangerous; antibiotic resistant pathogens are very dangerous indeed.

3.9 Epigenetic markers

These are associated with certain virulence genes and operons and epigenomic studies are likely to reveal many more (Chen *et al.,* 2014). For example, *agn43,* the gene for antigen 43, is an important autotransporter that is under dual control by Dam-mediated methylation and oxidative stress in *E. coli*. The *pap* operon that encodes Pap pili, important for pyelonephritis and kidney infection, is controlled phase-variably by Dam methylation and the leucine-responsive regulatory protein, Lrp (Hernday *et al.,* 2003). These genetic switches are characterised by alternate fully- and hemi-methylated states that are permissive or non-permissive for gene expression, creating cell-surface variety among bacteria in a genetically homogeneous population. The result is an increased probability of evading the host defences during infection.

3.10 Pathogenic lineages

To date, the known lineages containing human and/or animal pathogens are the Flavobacteria-bacterioides, the Spirochetes, the Chlamydia, the Cyanobacteria, the Proteobacteria and the Gram-positive bacteria (*i.e. Mycobacterium, Clostridia, Listeria, Rhodococcus and Streptococcus*). A large majority of known marine pathogens belong to the Gammaproteobacteria. Within these, the genus *Vibrio* alone contains 12 recognized human pathogens and many more animal pathogens. Other proteobacteria frequently associated to disease are *Aeromonas* and *Shewanella*.

However, diagnosis based on taxonomy is not sufficient to conclude about pathogenicity since the functional unit of pathogenesis is more often the strain or clone within a species thanks to recent acquisition of lateral gene transfer (LGT). There is an urgent need for accurate and rapid laboratory diagnostic methods leading to better control and treatment strategies, which shall include detection of specific virulence factors playing a role in the different kinds of infections (ex. adherence, contact-independent factors etc.)

4. THE PATHOGENIC ENVIRONMENT

Changes in host range, in pathogenic traits displayed in the same host, and the geographic distribution of a disease complex form three distinct sets of complementary and only slightly intersecting disease emergence scenarios. Together, these scenarios present the full picture and range of possible disease emergence dynamics (Engering *et al.*, 2013). A new era in medical science has dawned with the realization of the critical role of evolutionary and ecological factors, including both the microbial community structure and host health conditions in bacterial infectious diseases.

By providing for the first time a more comprehensive view of the "microbiome", the *Human Gut Microbiome* initiative showed the importance of selective pressures and community dynamics in shaping the microbiome (including the "extra-intestinal" one) in diseased humans and in gut pathogens. These breakthrough results have opened new scenarios whereby human 'disease susceptibility' could well exhibit geographical patterns depending on social, economic and ecological features (Ruth *et al.*, 2016).

4.1 Geographic location

Geographic location determines the probability of potential contact with host organisms. Coastal regions with dense human populations obviously represent areas with higher risks of disease transmission than other more pristine or open ocean sites. Ocean circulation further determines the connectivity of different sites with each other, such that Lagrangian transport distances are more relevant in the ocean than are actual geographic distances (see CIESM 2016). Thus one will find regions where transport times are short and extensive such as the Gulfstream in the North Atlantic, and areas where mixing is extremely limited such as in the large oceanic gyres. Further the biodiversity of different oceanic regions will determine quite distinct oceanic biogeographical provinces. As for all other oceanic microbes, the spatial distribution of pathogenic organisms is driven by wind, vertical transport, transport in ballast water, or rafting on marine litter such as microplastics (see CIESM 2014).

Mechanisms responsible for genome plasticity are found to specifically drive bacterial adaptive response and bacterial evolution in each host environment (see Dorman; Vezzulli *et al.*, both in this volume).

4.2 Physico chemical variables

These physical and chemical signals individually and collectively impose selective pressure on the bacterial metagenome, allowing some microbes to prosper while others decline. Knowledge of environmental structure and composition and of the genetic make-up of the bacterial population allows informed predictions to be made about which organisms are likely to inhabit particular niches and which are likely to be excluded. For example, a strict aerobe is likely to do poorly in a fully anoxic environment. When making predictions, one must keep in mind the ability of bacteria to become quiescent, entering a long-term dormant state, and in some cases to sporulate, before ruling out *a priori* the ability of bacteria to inhabit any part of the natural environment.

4.3 Biotic interactions

In addition to environmental (or abiotic) variables, the likelihood that microbes form myriads of associations between each other is receiving renewed interest. This has been explored in depth with the *Tara* oceans dataset by exploring global patterns of co-occurrence of organisms (Lima-Mendez *et al.*, 2015; Bowler, this volume). Moreover, a growing volume of research data supports the hypothesis that marine organisms may function as inter-epizootic reservoirs or vectors of pathogenic bacteria, and sometimes carry them over long distances (Troussellier *et al.*, this volume). A notable example is the association of *Vibrio cholerae* as a commensal microbe on marine copepods (Vezzulli *et al.*, this volume).

5. RELEVANT, RELIABLE, AVAILABLE DATABASES

The evolution of molecular biology protocols and sequencing technologies, from Sanger sequencing trough 2nd and 3rd generations parallel methods, allows us to collect a continuously growing amount of sequencing data which historically had a Moore's law doubling time of seven months, now estimated for 2nd generation Illumina sequencing at about twelve months (Goodwin *et al.*, 2016). Even within this more conservative view of a twelve months doubling time, we will face such a huge amount of data within a decade that the main problem will rely more on data transfer than on proper data processing (Muir *et al.*, 2016; Stephens *et al.*, 2015).

At this point in time where sequencing is no longer an economical or a technical issue, the bottleneck for many laboratories is in the analysis step. Yet shotgun and ribosomal genes amplicons data continue growing in volume at a vertiginous rhythm. Fortunately, the current tendency is to collect and to organize sequencing data so as to provide the users with powerful, open source tools such as the Metagenomics RAST Server (see below) to analyse and compare datasets from different environments.

5.1 Sequences / Nucleotides Databases

Publication of any genomics data requires submission of the corresponding DNA sequence data to one of the three main nucleotide archives joined under the “International Nucleotide Sequence Database Collaboration (INSDC)” initiative (http://www.insdc.org/), namely:

a/ The National Center for Biotechnology Information (NCBI, USA, https://www.ncbi.nlm.nih.gov/);

b/ The DNA Data Bank of Japan (DDBJ, Japan, http://www.ddbj.nig.ac.jp/);

c/ The European Nucleotide Archive (ENA), which is part of the European Molecular Biology Laboratory and European Bioinformatics Institute (EMBL-EBI, Europe, https://www.ebi.ac.uk/).

The content of each database is automatically mirrored at least every twenty-four hours. These databases have been increasing in size since their first appearance in the late 1980s and, given the advances in sequencing technologies, they have also been growing in complexity. Nowadays, in addition to the sequencing data, they also request the deposition of as many associated metadata as possible. This has required the development of standards that must be followed in order to deposit sequence information, which is crucial for allowing heterogeneous datasets to be compared with each other. The Genomics Standards Consortium (http://gensc.org/) plays a crucial role in ensuring the application of these recommendations. For marine metagenomics studies, the recommendations described in Ten Hoopen *et al*. (2015) are particularly relevant.

5.2 EMG (EBI Metagenomics Portal)

(https://www.ebi.ac.uk/metagenomics/)

The EBI Metagenomics service is an automated pipeline for the analysis and archiving of metagenomic data which aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of a sample. It is possible to freely browse all the public data in the repository, which includes *Tara* Oceans, OSD, etc. (Mitchell *et al.,* 2017) (more details in Bowler *et al.*; Villarroya *et al.,* both in this volume)

5.3 PANGAEA

(https://www.pangaea.de/)

The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. Each dataset can be identified, shared, published and cited by using a Digital Object Identifier (DOI).

5.4 OBIS

(http://www.iobis.org/)

OBIS – the Ocean Biogeographic Information System - is a global open-access data and information clearing-house on marine biodiversity for science, conservation and sustainable development. It emanates from the Census of Marine Life (2000 – 2010) and is pursued, with contributions of 500 institutions, under the aegis of the IOC International Oceanographic Data Exchange Programme (IODE).

5.5 MG-RAST

(http://metagenomics.anl.gov/)

MG-RAST is a web application server that allows the users to upload metagenomes for automated analysis and phylogenetic classification of sequence fragments and functional classification of samples.

5.6 Ribosomal sequences databases

One of the most important resources enabling the taxonomic description of microbial life is represented by databases and search tools providing the research community with aligned and annotated rRNA gene sequence data such as the Ribosomal Database Project (Cole and Tiedje, 2014 ;

Cole *et al.*, 2005) and the SILVA databases, jointly with the ARB project for phylogenetic tree reconstructions and representations, which includes small and large ribosomal gene subunits from the Bacteria, Archaea and Eukarya domains (Quast *et al.*, 2013; Yilmaz *et al*, 2014). More precisely, the RDP and SILVA projects began with less than 500 entries obtained in Charles Woese laboratory in 1991 and now contain more than three (RDP) and six (SILVA) million sequences in their most recent releases.

Interestingly, together with Bacteria/Archaea domains, Fungi have also been targeted for wide spectrum taxonomic characterization by the means of the Intergenic Transcribed Spacers (ITS) region (Deshpande *et al.*, 2016). Today, 18S rDNA sequences from eukaryotes are most comprehensively represented in the PR2 database (Guillou *et al.* 2013).

5.7 MAGE (*Microbial Annotation Genome and Analysis, Genoscope*)

A large amount of public genomic data concerning various bacterial species have been integrated in the MicroScope platform[1] to ease analysis by a common set of methods and parameters (e.g. the VibrioScope project benefits from dynamic and permanent updates of genome annotations). Twice a year, trainings are organised by the MAGE team for genome annotation and comparative genomics, RNAseq and metabolic network analysis.

6. STATISTICAL/ MATHEMATICAL TOOLS FOR DATA MINING

Data mining is the discipline of discovering patterns of various types of variables from databases (Aggarwal, 2016; Leskovec *et al.*, 2014; Zaki and Meira, 2014), and the techniques of *machine learning* plays a central role to achieve the task (Bishop, 2007; Murphy, 2012). That sector is vast and our meeting focused on recent advances of data mining and machine learning methods that can find potential associations of variables from biological data.

To find relevant variables from a dataset, *feature selection* is the representative approach in machine learning. Feature selection detects variables, or features, that are associated with the target variable from the set of all variables in a given dataset (Guyon and Elisseeff, 2003). The target variable can be binary (0 and 1 for cases and controls) in a case-control study or continuous in regression. The simplest method is *variable ranking*, where we compute the association score, for example, Pearson's correlation coefficient, to detect linear association or the mutual information for nonlinear association, between each variable and the target variable, followed by ranking the variables according to the association scores. Then one can find top-ranked highly associated variables from the ranking. This type of the two-step procedure, measuring association and making a ranking, is called a *filter method*. Although this approach can be easily applied for removing unnecessary variables from a dataset, redundant features might be selected as interactions between variables are not considered. For instance, if a dataset contains exactly the same variables that have the strong association with the target variable, both variables are selected.

The other two approaches for feature selection are a *wrapper method* and an *embedded method*, where the quality of each variable is assessed by the accuracy of a prediction model with respect to the target variable. A wrapper method repeats to construct a prediction model for each subset of variables; hence it is computationally too expensive in most cases. By contrast, in an embedded method, variables are automatically selected during the process of learning a prediction model from a dataset. A popular

[1] www.genoscope.cns.fr/agc/microscope/about/collabprojects.php?P_id=35

method is *lasso* (Tibshirani, 1996) that learns a linear prediction model, where a set of variables, which receive nonzero coefficients, is automatically selected in the learning process by regularizing the number of variables. The joint additive effect of selected variables maximizes the prediction accuracy of the model. The recent advances in *selective inference* (Taylor and Tibshirani, 2015) now enable us to assess the statistical significance of selected features in an embedded method.

To discover not single features or linear combinations of features but *patterns*, that is, combinations of features with multiplicative effects, *pattern mining* is the promising approach in feature selection (Aggarwal and Han, 2014; see more in Sugiyama, this volume), which was originally developed to find frequently co-purchased items in market basket analysis (Agrawal and Srikant, 1994) and increasingly used in a wide range of applications including bioinformatics (Zhang *et al.*, 2014). In particular, *significant pattern mining* (Llinares-López *et al.*, 2017; Terada *et al.*, 2013) has been recently developed, where one can find all variable combinations that are statistically significantly associated with the target variable while rigorously controlling the family-wise error rate (FWER). Moreover, one can apply pattern mining to detect sequential patterns (a sequence of variables) from time-series data and substructures of graphs from a graph-structured data.

In addition to feature selection, *dimension reduction* is often used to reduce the number of variables in the dataset, where variables are not directly selected but transformed into principal variables. Among a number of dimension reduction techniques, *t-SNE* is recently becoming a popular method and often used to visualize a multi-dimensional dataset (van der Maaten and Hinton, 2008).

Clustering is the standard approach to find groups of data points, which does not usually need the target variable. Given the number K of groups of data points, or *clusters*, the *K-means* algorithm divides data points into K disjoint clusters. Since the clustering quality obtained by the K-means algorithm is deteriorated by *outliers* in a dataset, one can perform *outlier detection* to remove such outliers before applying clustering, which can be also used as pre-processing for feature selection. State-of-the-art algorithms can efficiently detect outliers from a dataset using subsampling of data points (Sugiyama and Borgwardt, 2013).

By combining pattern mining and clustering, text mining can be achieved to analyse large amounts of text data. For example, one can automatically categorize text data such as scientific papers or news articles by finding frequently co-occurring words from them as keywords, followed by clustering them according to such keywords. Text mining is now actively used in the construction of biological databases (Wardeh *et al.*, 2015 and in this volume).

Recently, machine learning techniques are used not only for directly predicting the target variable but for designing experiments. *Bayesian optimization* offers the sequential experimental design strategy, which tells us which data point should be examined at the next experiment to efficiently maximize the prediction accuracy with respect to the target variable (Mockus, 2011). This technique has been already successfully used in various applications such as materials science (Ju *et al.*, 2017). Thus it might be interesting to use Bayesian optimization to build a biological database in an efficient manner.

References

Aggarwal C.C. and Han J. [editors] 2014. Frequent Pattern Mining. Springer.

Aggarwal C.C. 2016. Data Mining: The Textbook. Springer.

Agrawal R. and Srikant R. 1994. Fast algorithms for mining association rules. *In*: Proceedings of the 20th International Conference on Very Large Data Bases: 487–499.

Bañuelos-Vazquez L.A., Torres Tejerizo G., Brom S. 2017. Regulation of conjugative transfer of plasmids and integrative conjugative elements. *Plasmid* 91: 82-89.

Barquist L., Westermann A.J., Vogel J. 2016. Molecular phenotyping of infection-associated small non-coding RNAs. *Philos Trans R Soc Lond B Biol Sci* 371 (1707). pii: 20160081.

Ben-Haim Y., Zicherman-Keren M., Rosenberg E. 2003. Temperature-regulated bleaching and lysis of the Coral *Pocillopora damicornis* caused by the novel Pathogen *Vibrio coralliilyticus*. *Appl Environ Microbiol*. 69 (7): 4236–4242.

Bishop C.M. 2007. Pattern Recognition and Machine Learning. Springer.

Bleidorn C. 2016 Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. System. Biodivers. 14 (1): 1-8

Chagnot C., Listrat A., Astruc T., Desvaux M. 2012. Bacterial adhesion to animal tissues: protein determinants for recognition of extracellular matrix components. *Cell Microbiol* 14 (11):1687-1696.

Chen P., Jeannotte R., Weimer B.C. 2014. Exploring bacterial epigenomics in the next-generation sequencing era: a new approach for an emerging frontier. *Trends Microbiol* 22(5):292-300.

CIESM 2014. Marine litter in the Mediterranean and Black Seas. CIESM Workshop Monograph n°46 [F. Briand ed.], 180 p., CIESM Publisher, Monaco.

CIESM 2016. Marine connectivity – migration and larval dispersal. CIESM Workshop Monograph n°48 [F. Briand ed.], 172 p., CIESM Publisher, Monaco.

Cole J.R. and Tiedje J.M. 2014. History and impact of RDP: a legacy from Carl Woese to microbiology. *RNA Biol.* 11 (3): 239–43.

Cole J.R., Chai B., Farris R.J., Wang Q., Kulam S.A., McGarrell D.M., Garrity G.M. and Tiedje J.M. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* 33 (Database issue): D294–296.

Deshpande V., Wang Q., Greenfield P., Charleston M., Porras-Alfaro A., Kuske C.R., Cole J.R., Midgley D.J. and Tran-Dinh N. 2016. Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia* 108 (1): 1–5.

Domman D., Quilici, M.L., Dorman, M.J., Njamkepo, E., Mutreja, A., Mather, A.E., Delgado, G., Morales-Espinosa, R., Grimont, P.A.D., Lizárraga-Partida, M.L., Bouchier, C., Aanensen, D.M., Kuri-Morales, P., Tarr, C.L., Dougan, G., Parkhill, J., Campos, J., Cravioto, A., Weill, F.X., Thomson, N.R. 2017. Integrated view of *Vibrio cholerae* in the Americas. *Science* 358 (6364): 789-793.

Dorman C.J. 1994. Genetics of Bacterial Virulence. Blackwell.

Dorman C.J., Bogue M.M. 2016. The interplay between DNA topology and accessory factors in site-specific recombination in bacteria and their bacteriophages. *Science Progress* 99 (4): 420-437.

Dorman C.J., Dorman M.J. 2017. Control of virulence gene transcription by indirect readout in Vibrio cholerae and Salmonella enterica serovar Typhimurium. *Environ Microbiol* 19 (10): 3834-3845.

Engering A., Hogerwerf L., Slingenbergh J. 2013. Pathogen–host–environment interplay and disease emergence. *Emerging Microbes & Infections* 2, e5, doi:10.1038/emi.2013.5Gerdes K., Christensen S.K., Løbner-Olesen A. 2005. Prokaryotic toxin-antitoxin stress response loci. *Nat Rev Microbiol* 3 (5): 371-382.

Gillings M.R. 2017. Class 1 integrons as invasive species. *Curr Opin Microbiol* 38: 10-15.

Goodwin S., McPherson J. D., & McCombie W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17* (6), 333-351.

Guillou L., Bachar D., Audic S., Bass D., Berney C., Bittner L. and many others 2013. The Protist Ribosomal Reference database (PR2): a catalogue of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* (Database issue): D597-604.

Guyon I. and Elisseeff A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1156–1182.

Gyles C., Boerlin P. 2014. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Vet Pathol* 51 (2): 328-340.

Harms A., Maisonneuve E., Gerdes K. 2016. Mechanisms of bacterial persistence during stress and antibiotic exposure. *Science* 354 (6318) pii: aaf4268.

Hernday A.D., Braaten B.A., Low D.A. 2003. The mechanism by which DNA adenine methylase and PapI activate the pap epigenetic switch. *Mol Cell* 12 (4): 947-957.

Holmqvist E., Wright P.R., Li L., Bischler T., Barquist L., Reinhardt R., Backofen R., Vogel J. 2016. Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. *EMBO J* 35 (9): 991-1011.

Horvath P., Barrangou R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327 (5962): 167-170.

Humphrey B., Thomson N.R., Thomas C.M., Brooks K., Sanders M., Delsol A.A., Roe J.M., Bennett P.M., Enne V.I. 2012. Fitness of Escherichia coli strains carrying expressed and partially silent IncN and IncP1 plasmids. *BMC Microbiol* 4: 12:53.

Ju S., Shiga T., Feng L., Hou Z., Tsuda, K. and Shiomi, J. 2017. Designing nanostructures for phonon transport via bayesian optimization. *Phys. Rev. X*, 7 (021024): 1-10.

Koonin E.V., and Makarova, K.S. 2017. Mobile genetic elements and evolution of CRISPR-Cas systems: all the way there and back. *Genome Biol Evol* 9 (10): 2812-2825.

Leskovec J., Rajaraman, A., and Ullman, J. D. 2014. Mining of Massive Datasets. Cambridge University Press. 484 pp.

Lima-Mendez G., Faust K., Henry N., Decelle J., Colin S., Carcillo F., Chaffron S., Ignacio-Espinosa J.C., Roux S., Vincent F., Bittner L., Darzi Y. *et al.* 2015. Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* 348(6237): 1262073- (1-9).

Liss V., Hensel M. 2015. Take the tube: remodelling of the endosomal system by intracellular Salmonella enterica. *Cell Microbiol* 17 (5): 639-647.

Llinares-López F., Papaxanthos L., Bodenham D., Roqueiro D., and Borgwardt K. 2017. Genome-wide genetic heterogeneity discovery with categorical covariates. *Bioinformatics*, 33 (12): 1820–1828.

Loenen W.A., Dryden D.T., Raleigh E.A., Wilson G.G., Murray N.E. 2014. Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res* 42 (1): 3-19.

Maddocks S.E., Oyston P.C. 2008. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* 154: 3609-3623.

Maisonneuve E., Gerdes K. 2014. Molecular mechanisms underlying bacterial persisters. *Cell* 156 (3): 539-548.

Martín J.F., Liras P. 2012. Cascades and networks of regulatory genes that control antibiotic biosynthesis. *Subcell Biochem* 64: 115-138.

Mecsas J.J., Strauss E.J. 1996. Molecular mechanisms of bacterial virulence: type III secretion and pathogenicity islands. *Emerg Infect Dis* 2 (4): 270-88.

Mitchell A.L., Scheremetjew M., Denise H., Potter S., Tarkowska A., Qureshi M., Salazar G.A., Pesseat S., Boland M.A., Hunter F.M.I., ten Hoopen P., *et al.* 2017. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res*. doi: 10.1093/nar/gkx967

Mockus J. 2011. Bayesian Approach to Global Optimization: Theory and Applications. Springer.

Moxon R., Bayliss C., Hood D. 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.*, 40: 307-333.

Muir P., Li S., Lou S., Wang D., Spakowicz D.J., Salichos L., Zhang J., *et al.* 2016. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* 17 (March): 53.

Murphy K.P. 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

Olsen G.J., Overbeek R., Larsen N., Marsh T.L., McCaughey M.J., Maciukenas M.A., Kuan W.M., *et al.* 1992. The Ribosomal Database Project. *Nucleic Acids Res*. 20 Suppl (May): 2199–2200.

Orlova N., Gerding M., Ivashkiv O., Olinares P.D.B., Chait B.T., Waldor M.K., Jeruzalmi D. 2017. The replication initiator of the cholera pathogen's second chromosome shows structural similarity to plasmid initiators. *Nucleic Acids Res* 45 (7): 3724-3737.

Pallen M.J. 2016. Microbial bioinformatics 2020. *Microb Biotechnol*. 9 (5): 681-6.

Palmer L.D., Skaar E.P. 2016. Transition metals and virulence in bacteria. *Annu Rev Genet* 50: 67-91.

Papenfort K., Bassler B.L. 2016. Quorum sensing signal-response systems in Gram-negative bacteria. *Nat Rev Microbiol* 14 (9): 576-588.

Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., and Glockner F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 41 (Database issue): D590–596.

Rubio-Portillo, E., Yarza P., Peñalver, C., Ramos-Esplá, A.A., Antón, J. 2014. New insights into *Oculina patagonica* coral diseases and their associated *Vibrio* spp. Communities. ISME J. 8 (9): 1794–1807.

Ruth E., Le, Peterson D.A., Gordon J.I. 2016. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124 (4): 837-848.

Soto-Rodriguez S. A., Gomez-Gil B., Lozano-Olvera R., Betancourt-Lozano M., Morales-Covarrubias M. S. 2015. Field and experimental evidence of *Vibrio parahaemolyticus* as the causative agent of acute hepatopancreatic necrosis disease of cultured shrimp (*Litopenaeus vannamei*) in Northwestern Mexico. *Appl. Environ. Microbiol.* 81: 1689–1699.

Sridharan H., Upton J.W. 2014. Programmed necrosis in microbial pathogenesis. *Trends Microbiol* 22 (4): 199-207.

Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., Iyer R., Schatz M.C., Sinha S., and Robinson G.E. 2015. Big Data: Astronomical or Genomical? *PLoS Biol.* 13 (7): e1002195.

Sugiyama M. and Borgwardt K.M. 2013. Rapid distance-based outlier detection via sampling. *In:* Advances in Neural Information Processing Systems, 26: 467–475.

Sunagawa S., Coelho L.P., Chaffron S., Kultima J.R., Labadie K., Salazar G. Djahanschiri B. *et al.* 2015. Ocean plankton. Structure and function of the global ocean microbiome. *Science*. 348 (6237): 1261359-(1-9).

Sundberg L.-R., Ketola T., Laanto E., Kinnula H., Bamford J.K.H., Penttinen R., Mappes J. 2016. Intensive aquaculture selects for increased virulence and interference competition in bacteria. *Proc. Biol. Sci.* 283 (1826): 20153069.

Sutherland K.P., Shaban S., Joyner J.L., Porter J.W., Lipp E.K. 2011. Human pathogen shown to cause disease in the threatened eklhorn coral *Acropora palmata. PLoS One* 6 (8): e23468.

Tamplin M.L. 2001. Coastal vibrios: Identifying relationships between environmental condition and human disease. *Hum. Ecol. Risk Assess*. 7: 1437–1445.

Taylor J. and Tibshirani R. J. 2015. Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA*, 112 (25): 7629–7634.

Ten Hoopen P., Pesant S., Kottmann R., Kopf A., Bicak M., Claus S., Deneudt K., Borremans C., Thijsse P., Dekeyzer S., Schaap D.M., Bowler C., Glöckner F.O., Cochrane G. 2015. Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. *Stand. Genomic Sci.* 8: 10:20. doi: 10.1186/s40793-015-0001-5

Terada A., Okada-Hatakeyama M., Tsuda K., and Sese J., 2013. Statistical significance of combinatorial regulations. *Proc. Natl. Acad. Sci. USA*, 110 (32): 12996–13001.

Tibshirani R., 1996. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. (Series B, Statistical Methodology)*, 58 (1): 267–288.

Tomás J.M. 2012. The main *Aeromonas* pathogenic factors. *ISRN Microbiology*. vol. 2012, Article ID 256261, 22 pages doi:10.5402/2012/256261

van der Maaten, L. and Hinton, G. E., 2008. Visualizing data using t-sne. *J. Mach. Learn. Res*., 9: 2579–2605.

Vezzulli L., Grande C., Reid P.C., Hélaouët P., Edwards M., Höfle M.G., Brettar I., Colwell R.R., Pruzzo C. 2016. Climate influence on *Vibrio* and associated human diseases during the past half-century in the coastal North Atlantic. *Proc. Natl. Acad. Sci.* USA. 113 (34): E5062-71

von Bronk B., Schaffer S.A., Götz A., Opitz M. 2017. Effects of stochasticity and division of labor in toxin production on two-strain bacterial competition in Escherichia coli. *PLoS Biol* 15 (5): e2001457.

Waldor M.K., Mekalanos J.J. 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272 (5270): 1910-1914.

Wardeh M., Risley C., McIntyre M. K., Setzkorn C., Baylis M. 2015. Database of host-pathogen and related species interactions, and their global distribution. *Sci. Data*, 2: 150049.

Wu Y., Aandahl R.Z., Tanaka M.M. 2015. Dynamics of bacterial insertion sequences: can transposition bursts help the elements persist? *BMC Evol Biol* ;15:288.

Yamashita A., Sekizuka T., Kuroda M. 2014. Characterization of antimicrobial resistance dissemination across plasmid communities classified by network analysis. *Pathogens* 3 (2): 356-376.

Yang J., Tauschek M., Robins-Browne R.M. 2011. Control of bacterial virulence by AraC-like regulators that respond to chemical signals. *Trends Microbiol* 19 (3): 128-135.

Yilmaz P., Parfrey L.W., Yarza P., Gerken J., Pruesse E., Quast C., Schweer T., Peplies J., Ludwig W., and Glockner F.O. 2014. The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks. *Nucleic Acids Res*. 42 (Database issue): D643–648.

Zaki M.J. and Meira Jr., W. 2014. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.

Zhang Q., Long Q., and Ott J. 2014. AprioriGWAS, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. *PLoS Comput. Biol.*, 10 (6): e1003627.

Diversity of metagenomics studies - dissecting microbiomes

Mária Džunková[1, §], Francesc Peris-Bondia[4,§], Andrés Moya[2, 3, 4], Giuseppe D'Auria[4, 5]

[1]-Australian Centre for Ecogenomics, University of Queensland, St Lucia QLD 4072, Australia
[2]- Integrative Systems Biology Institute, University of Valencia and CSIC, Valencia, Spain
[3]-Genomics and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencian Community (FISABIO), Valencia, Spain.
[4]-Sequencing and Bioinformatics Service, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencian Community (FISABIO), Valencia, Spain.
[5]- CIBER in Epidemiology and Public Health, Madrid, Spain
[§]- Equal contribution

Abstract

The advancements of sequencing methods have opened the way for the characterization of almost all microbial niches in the environment, passing from cloning-based studies to the second generation parallel short reads sequencing towards the third generation based on long reads direct sequencing, which probably will soon become a new standard or at least co-exist with the short reads sequencing methods. Behind the fever of DNA or cDNA sequencing for metagenomics or metatranscriptomics studies respectively, several additional methodologies focused on subsetting specific compartments of microbial environments appear. This growth and diversity of data collection are also favoured by the drop of sequencing costs and the rise of the output leading to a so-called "democratization" of DNA sequencing.

We provide here an overview of the sequencing approaches used in metagenomics studies focused on dissecting the complexity of microbial ecosystems.

Approaching microbiomes: High throughput sequencing

In the past, prior to the bloom of the DNA sequencing technologies, there was always a need for objective methods for bacterial taxonomic classification. At that time, microscopic observations, numerical taxonomy, chemotaxonomy and other biochemical techniques were being used for characterization of the microorganism studied. All these methods were based on the cultivation of the organisms. In the late '70s, probe based detection opened the door for cultivation-independent methods that showed the magnitude of microbial diversity (Amann *et al*., 1995; Rossello-Mora and

Amann, 2001). Thus, in the '80s, all kinds of environments started to be studied using available methodologies which, for the first time, did not require microbial cultures. The development of the DNA sequencing technologies represents a big step forward, which provided a real insight of microbial diversity including unculturable organisms.

When sequencing became cheaper and faster, knowledge on microbial diversity expanded and under the lemma of "metagenomics" researchers collected microbial diversity bits from almost all sites where microbial life blooms (Schmidt *et al.,* 1991; Handelsman *et al.,* 1998). The environment is now seen not only as a soup containing lifeforms of almost all taxonomic levels, but also as an important source of potential industrial applications which can be easily uncovered by sequencing technologies (Zhao and Bajic, 2015). The expansion of metagenomics could have its parallel in the invention of the radio telescope which captures radio waves coming from very remote planets, stars, nebulas and galaxies and so provides an insight into the universe which is now understood in a much deeper way than the one provided by optical instruments.

Shot-gun metagenomics and PCR-based surveys (mainly 16S rDNA gene amplicons), provided a new view of the tree of life. Nowadays we experience a fever for microbial life description along all possible gradients of temperature, pH, salinity, oxygen saturation levels, pressure, light or radiation exposition. Thousands of projects describe host-associated microbiomes. The studies of host-related microbiomes showed that the microbes do not follow a random distribution but respond to specific patterns according to micro-niches established in every sector (Trivedi, 2012; Grice and Segre, 2011; Lloyd-Price, Abu-Ali, and Huttenhower, 2016; Donaldson, Lee, and Mazmanian, 2016). In 2008, NIH started the Human Microbiome Project aiming to describe all microorganisms living with us. This approach revealed intimate mutualistic/symbiotic relationships between humans and microorganisms and showed how bacteria interact with humans during their life. Statistical correlations of specific bacterial communities with healthy status, different diseases and human age have been reported (Turnbaugh *et al.,* 2007; Huttenhower *et al.,* 2012).

The amplicon sequencing, metagenomics and metatranscriptomics are the most applied protocols in current microbial studies. In these approaches, DNA or RNA extracted directly from the environment is used as a representative of whole ecosystems. It is very important to keep in mind that the DNA is extracted from all bacterial cells in the environment, so their metabolic state (alive, death, spore form) is not considered. While RNA extraction provides a snapshot of the metabolic activities in the sample, it does not provide mRNA levels *per cell*. This means that the correlation of the expression levels with microbial counts is unfeasible. A recent review by Emerson *et al.* (2017) about the distinction between living and dead microbes in microbiome studies provides also a clear view about current methods and limitations in their applications.

In the '90s, prior to the metagenomic expansion, several reports already pointed out that the inactive bacterial fraction in aquatic environment can reach up to 90% of total bacterial counts (Porter *et al.,* 1995; Gasol *et al.,* 1995). These data have been further confirmed with the arrival of modern counting and sequencing technologies highlighting how taxonomic and functional metagenomic profiles are depicting the real active bacterial players (Romanowicz *et al.,* 2016; Gosalbes *et al.,* 2011; Tanca *et al.,* 2017).

Another important issue to consider is that microorganisms (including small eukaryotes, fungi, eubacteria, archaea, and viruses) do not inhabit uniquely the planktonic fraction, but are also associated to suspended matter, biofilms and other micro-niches and micro-physicochemical gradients. Although the anaerobiosis is generally considered a uniform selection parameter in the gut microbiome, one can observe a multitude of environments along the whole intestine where bacteria establish their own niche interacting passively or actively with the host (Pereira and Berry, 2017). However, the majority of published studies are based on the characterization of the stool samples for the simplicity of collection.

Fractioning microbiomes

The high level of specialization in current sampling methodologies and sequencing technologies provides a high resolution insight into environments. The diversity of a microbiome can be sub-grouped into many different fractions down to the level of single cells. The sub-grouping can be performed on the level of cell activity (active *vs.* inactive fraction), cell size (from eukaryotic cells to viruses) while cells belonging to different taxonomic and functional groups can be targeted too.

Flow cytometry for enrichment of microbiome fractions

One of the more powerful tools for sub-setting the microbiome is the flow cytometry which allows to select microbiomes fractions according to different parameters such as cell size, morphology, DNA/RNA content, fluorescence emitted by hybridization probes targeting genes (on DNA or RNA) or by immunoglobulin-based antibodies, etc. (Warnecke and Hugenholtz, 2007; D'Auria *et al.*, 2013; Džunková *et al.*, 2016; Simon-Soro *et al.*, 2015).

In previous work we reported the usage of the flow cytometry for selection and sorting of bacteria with high RNA concentration (Figure 1). This active subset of bacteria was taxonomically more diverse than the inactive fraction. These results revealed a core of active bacteria which is very underrepresented (almost invisible) in the routine DNA analysis of the whole bacterial community (Peris-Bondia *et al.*, 2011).

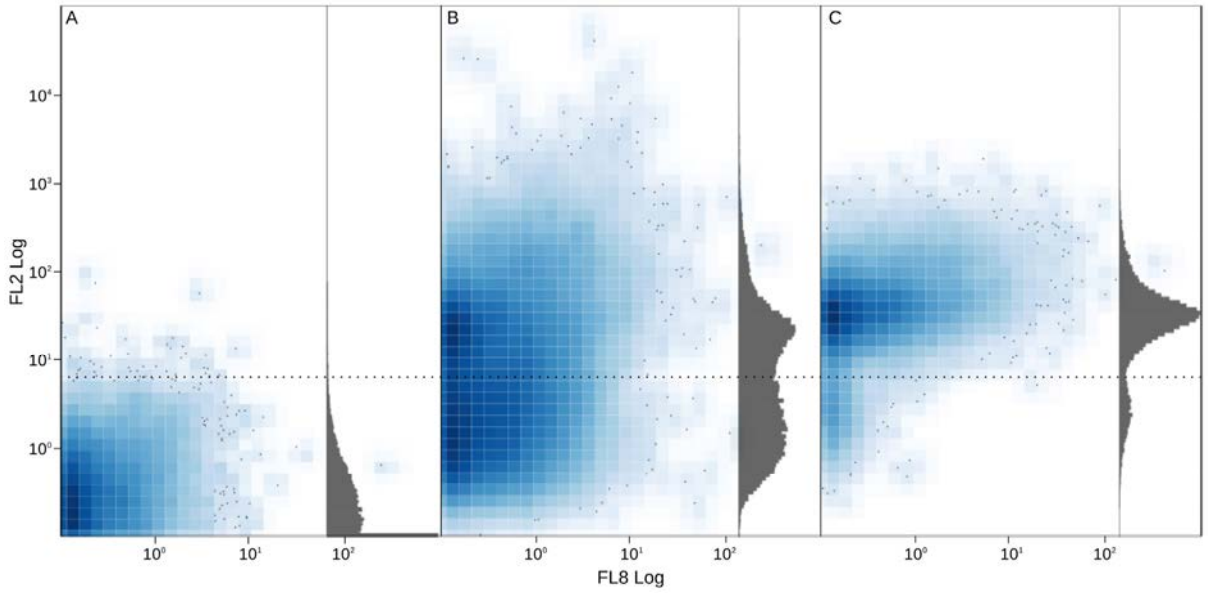


Figure 1. Distribution of events (cells) in bi-plots showing fluorescence levels captured on flow cytometer channels FL8 PMT and FL2 PMT. X and Y axis are defined in logarithmic arbitrary units. Gray histograms resume the distribution of events along FL2 PMT. Panel A shows unstained cells. Panel B shows cells stained with pyronin-Y targeting RNA; it is possible to observe two populations of unstained and stained cells respectively (the dotted black line separates the two populations). Panel C shows only pyronin-Y stained cells passing the threshold filter on FL2 PMT and sorted for DNA-based amplicon sequencing. Reproduced from Peris-Bondia *et al.* (2011).

In other words, the standard DNA based approach shows well known over-represented bacteria while directing sequencing efforts towards the active fraction gives importance to those species which are normally considered a minority (Peris-Bondia *et al.*, 2011).

In addition, we tried to investigate the interaction between microbiome and host by sorting bacteria opsonized by human IgA. The opsonized cells were labelled with fluorescent anti-human IgA and separated by flow cytometry which was followed by 16S gene amplification of the collected fraction. We were able to identify a core group of bacteria commonly present in samples from different individuals. Again, these bacteria are underrepresented when the samples are investigated by the standard methods targeting the whole faecal bacterial community (D'Auria *et al.*, 2013).

The targeted labelling of specific fraction of gut microbiota was also used to investigate the correlation between the commensal microbiota and the onset of *Clostridium difficile* infections (CDI). The work of Džunková *et al.* (2016) demonstrates that *C. difficile* infection is statistically more probable when the gut microbiota is found under a state of dysbiosis and when beneficial microbes such as *Lactobacillales* and members of *Clostridium* cluster IV appear to have reduced activity. The study showed that selective sequencing of specific bacterial fractions provides additional information about dysbiotic processes in the gut (Džunková *et al.*, 2016).

Accessing the virome

While the majority of sequencing efforts in metagenomic characterization provides information on prokaryotic/eukaryotic systems, a considerable viral fraction is also usually retrieved from sequence pools by standard bioinformatics methods (Paez-Espino *et al.*, 2016). It is worth to mention that phages are widely accepted to be orders of magnitude more abundant then the proper microbial fraction (Mizuno *et al.*, 2013). In order to improve sequencing efforts aiming to characterize natural viral fractions, several protocols have been developed. They are mostly based on PCR directing viral genes or on size/based sequential filtrations and density/based enrichment of viral particles prior to sequencing (Hjelms *et al.*, 2017; Goldsmith *et al.*, 2015).

The viral fraction is composed of very small particles containing a very reduced amount of DNA compared with surrounding eukaryotes/prokaryotes chromosomes. When viral particles are collected/ enriched for further DNA/RNA extraction for high throughput sequencing, all research efforts may be jeopardized by the presence of residual host or microbial DNA. Nonetheless, the flow cytometry represents a sensitive tool for selection of viral particles avoiding DNA contaminations from organisms of larger genome size.

In previous work we developed a protocol for viral particle enrichment after the standard filtration steps in which the flow cytometry represents an additional purification step. The selection of smaller viral particles with similar size and DNA content by FACS (fluorescence activated cell sorter) improved their genome assembly because reducing diversity and complexity facilitates the assembly. Although we proposed the assembly from the entire faecal virome, the FACS could be applied on different fractions visualized on the flow cytometry bi-plot charts allowing to capture higher diversity and consequently simplifying the assemblies (Figure 2) (Džunková, D'Auria, and Moya, 2015). In this work we optimized our sequencing efforts toward viral fraction. We showed that FACS coupled with the sequencing library preparation protocol omitting whole genome amplification helped to solve the difficulties reported in many virome sequencing projects mainly related to the reads assembly and annotation.

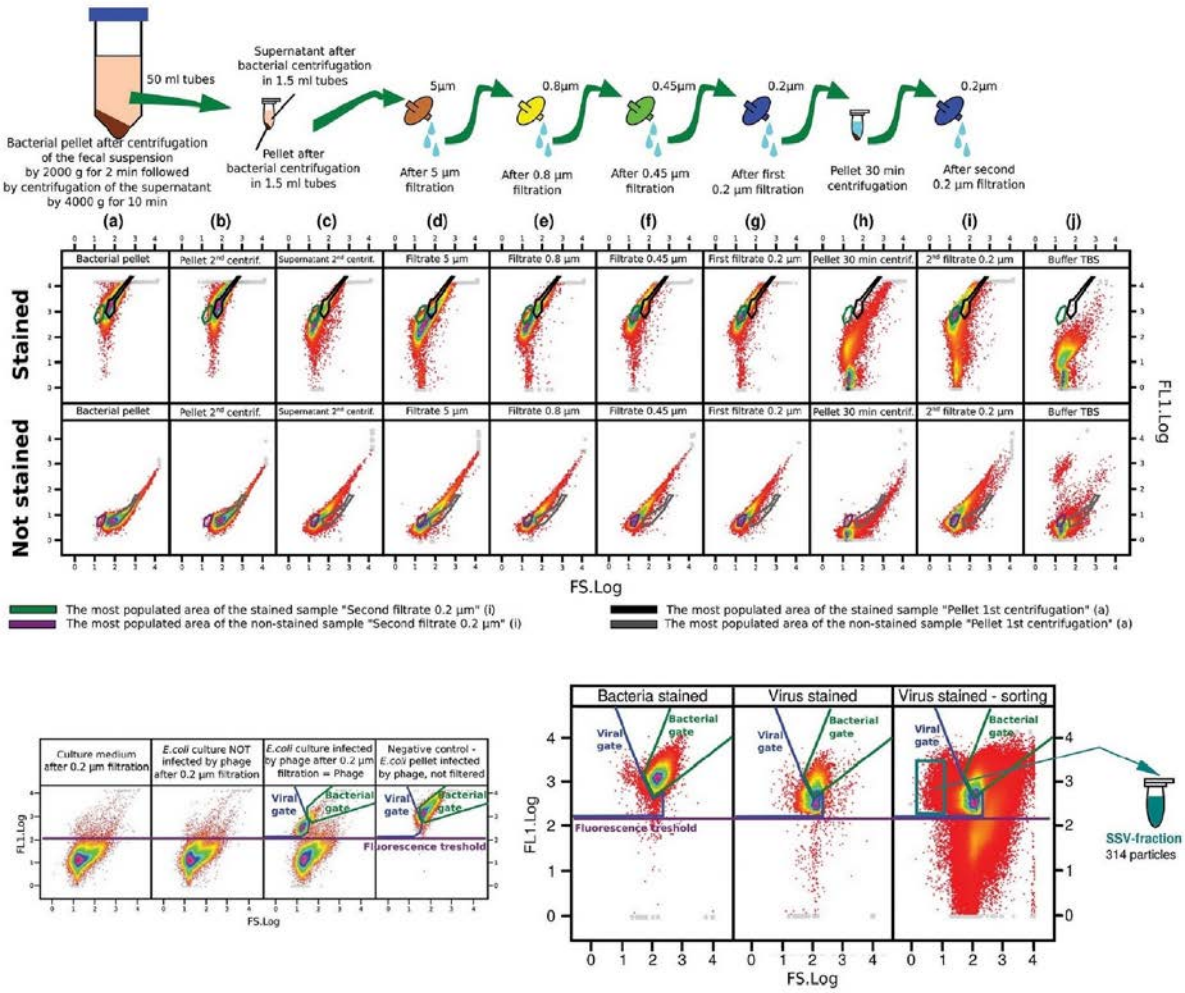


Figure 2. (A) Flow cytometry bi-plots of control samples across all steps of the faecal purification protocol. (B) FC bi-plots of Escherichia coli culture infected/not infected by phage M13KE. (C) FACS of the faecal sample. Reproduced from Džunková, D'Auria and Moya (2015).

Single-cell genomics

While metagenomics provides an integrated view on the genomic potential of the whole environment, it is overlooking the cell individuality and hindering the links between the cells and its function. If cultivation is not an option, the single-cell genomics is nowadays possible for single bacterial cells and even single viral particles which can be distributed one by one by a flow cytometer into well plates and their whole genome is amplified by a special type of DNA polymerase (Lasken, 2012; Martinez-Hernandez *et al.*, 2017; Rinke *et al*,. 2014).

In one of the first studies in which flow cytometry and single-cell genomics were used, the genome of a member of the under-represented unculturable TM7 phylum was successfully recovered from a soil sample (Podar *et al.*, 2007) and from saliva (Marcy *et al.*, 2007). Later on, the single-cell genomics allowed describing tens of novel bacterial taxonomical groups belonging to so called "microbial dark matter" (Rinke *et al.*, 2013).

One of the problems in single-cell genomics relays on the difficulties in obtaining enough DNA for sequencing. Very specialized facilities are a must in this discipline where the whole genome

amplification of the single cells must be performed in extremely sterile conditions, as any foreign DNA may be over-amplified and contaminate the sequencing output. The flow cytometer used for distribution of the single cells into well plates must work using sterilized buffers. The lysis buffer and the whole genome amplification reagents must be UV-radiated. Moreover, due to the high risk of bacterial contamination by air, the reagents should be distributed into well plates by robotic pipettors in closed UV-sterilized hoods, rather than by humans (Rinke *et al.*, 2014).

Still, while the standard metagenomic approaches report commonly only the most dominant taxa, the flow cytometry based sub-setting of the whole microbial community leads quite often to discoveries of novel taxa. Nowadays, the complete or scaffolds representations of genomes of unculturable organisms is being promoted with the aim to avoid the noise produced by dominant taxa (Andrei *et al.*, 2017).

* this chapter is to be cited as :

D'Auria G., Džunková M., Peris-Bondia F. and Moya A. 2017. Diversity of metagenomics studies - dissecting microbiomes. pp. 27 – 32 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

The Enhanced Infectious Disease Database (EID2) system of species interactions and location automatic detection and its applications

Maya Wardeh, Marie McIntyre and Matthew Baylis

Institute of Infection and Global Health, University of Liverpool, UK

Abstract

Communicable diseases continue to impose a great burden on public health and economies worldwide. It is well established that the ability of a pathogen to infect multiple hosts, particularly hosts in different taxonomic orders or wildlife, is a risk factor for emergence. Despite the significance of communicable diseases and the role multiple host species play in their transmission, amplification and 'spill-over' incidents, there have been few attempts to systematically collate information on all disease agents, their locations and interactions with established or potential hosts. We utilise data mining techniques and Big Data approaches to automatically populate the Enhanced Infectious Diseases Database (EID2). EID2 attempts to answer questions such as what are all the pathogens of a host, and what are all the hosts of a pathogen, what are all the countries where a pathogen was found, and what are all the pathogens found in a given country. The presented database provides a valuable resource for researchers of infectious agents, including bacteria. We provide examples of past and current utilisations of EID2 in the field of infectious diseases, before concluding with a few remarks on the challenges systems like EID2 often encounter.

Introduction

Communicable diseases continue to impose a great burden on public health and economies worldwide. The most recent assessment of the global burden of disease in human found that despite progress in compacting established infectious diseases, most notably malaria, hepatitis C, and HIV/AIDS, emerging infectious diseases such Ebola and Zika, as well as antimicrobial-resistant pathogens, still represent serious threats to life (GBD 2015 DALYs and HALE Collaborators). Wide-host range and vector-involvement have been linked to emergence of disease in new species (Jones *et al.*, 2008;

Morens *et al.,* 2004; Weiss and McMichael, 2004; Taylor *et al,.* 2001; Cleaveland *et al.,* 2001). In their most recent list of emerging diseases, the World Health Organisation (WHO) identified eight pathogens which are likely to cause sever epidemics, and further three diseases were also listed as serious and requiring action as soon as possible. These eleven pathogens are all zoonotic and/or involve arthropod vectors in their transmission[1].

Many emerging domesticated animal infections have also been linked to 'spill-overs' from other species including wildlife (Cleaveland *et al.,* 2001; Daszak *et al.,* 2000 ; Wiethoelter *et al.,* 2015). These diseases continue to inflict tremendous direct and indirect costs on the livestock sector, as a result of animal deaths, reduced productivity, and the cost of disease control. These costs are felt the most in low income countries, where animals are a significant source of income. Perhaps most importantly, communicable diseases impact negatively on animal welfare, and may limit or cause extreme fluctuations in the population size of wild animals (Tompkins *et al.,* 2001) and reduce the chances of survival of endangered or threatened species (Smith *et al.,* 2006). Emerging wild-life diseases have also been linked to 'spill-overs' from humans or domesticated animals, as well as resulting from human intervention, e.g. via host or parasite translocations (Daszak *et al.,* 2000). A similar pattern of emergence in wildlife has also been observed in the field of plant disease, where the introduction of new host or pathogen species has led to major outbreaks (Anderson *et al.,* 2004). One of the most famous examples is the potato late blight, caused by the oomycete *Phytophthora infestans*, which initially emerged as a disease of the cultivated potato when transported to Mexico from the south American Andes, and continued to emerge as an important disease of the world's seventh most valuable crop wherever potato was introduced to new countries and naïve host plants, until as recently as the 1990s. Emerging plant diseases have many considerable socio-economic effects, for instance, more than 600 million more people could be fed each year by halting the spread of fungal diseases in the world's five most important crops alone (Fisher *et al.,* 2012).

Bacterial agents have perhaps been less correlated with disease emergence in new host species (Taylor *et al.,* 2001; Cleaveland *et al.,* 2001); however, they still present a major threat (over 10% of human emerging diseases are bacterial), with *Neoehrlichia mikurensis* (2010) as one of the most recent none-opportunistic emerging bacteria. Additionally, a multitude of opportunistic bacteria continue to emerge and cause complications in the elderly, newly-born and the immune supressed (e.g. *Cronobacter*). Moreover, with the potential impact of antimicrobial-resistance in wildlife, as well as in domesticated animals, the identification of host ranges of bacterial species, pathogenic or not, have become of paramount importance.

Despite the significance of communicable diseases, and the role multiple host species play in their transmission, amplification and 'spill-over' incidents, there have been few attempts to systematically collate information on all disease agents, their locations and interactions with established or potential hosts. Examples include: most notably Taylor *et al.,* (2001) for human pathogens; Cleaveland *et al.,* (2001) for domestic mammal pathogens; and Stephens *et al.,* (2017) for pathogens of wild mammals. Each of these examples focused on a well-defined group of hosts. Their data were compiled manually from textbooks and the scientific literature. The final datasets in most cases are time-limited and do not reference infectious agents that have emerged since publication. However, with the advent of Big Data platforms and analytics, and the exponential increase in processing power, we can now develop automated systems to capture interactions between species. We present below The Enhanced Infectious Diseases Database (EID2) system, examples of past and current utilisations of the system in

[1] http://www.who.int/medicines/ebola-treatment/WHO-list-of-top-emerging-diseases/en/

the field of infectious diseases, before concluding with few remarks on the challenges that systems like EID2 often encounter.

The Enhanced Infectious Diseases Database (EID2) system

Recent years have seen a massive increase in open-access scientific output, both in terms of publications and genomic sequences. For instance, last year alone saw the publication of over 16% of the total number of papers indexed by PubMed, and approximately 20% of the total number of sequences uploaded to Genbank. The sheer volume, not to mention other complexities, of scientific output exceeds the ability of researchers, using traditional methods, to make effective use and assessment of all available findings. The Enhanced Infectious Diseases Database system (EID2) (Wardeh *et al.*, 2015) utilises data and text mining tools, with minimal expert input, in order to answer a range of questions such as: 1) What is the host-range of given pathogen/microbe[1]? 2) What are all the pathogens/microbes of a given host? 3) What are all the vector species of certain pathogen? And which hosts do they transmit this pathogen to? 4) What is the geographical range of an organism (host, pathogen or vector)?

In order to provide answers to these questions, the EID2 system comprises the following components:

1) Data repositories: EID2 maintains a number of complex data repositories and mapping dictionaries to facilitate interaction discovery and named entity recognition, including: 1) Organisms and their taxonomic lineage relationships (over 1 million organisms to date). 2) Alternative names (e.g. common names, common misspelling, breeds and acronyms), inclusion (AND) and exclusion (NOT) terms for the organisms. 3) Geographical names and hierarchies, including countries, administrative divisions, major cities and natural features. 4) Climate (e.g., temperature and rainfall) and demographic (human and livestock) data for the whole world.

2) Data acquisition layer: EID2 continually retrieves and classifies evidence from two sources: NCBI Nucleotide Sequences database; and PubMed (and soon to include Scopus as a third). Each piece of evidence is then linked to the organisms and geographical location. Sequences are often linked to one "cargo" organism which is either microbe (pathogen) or arthropod vector, one host organism and one location. Publications however are often linked to multiple organisms and locations. One powerful utilisation of EID2 is our ability to quickly extract and filter evidence based on the number of hosts/pathogens/vectors species or locations it mentions. This facilitates other processes of EID2, and enables us to conduct research in other avenues (such as transmission route discovery and co-infection interactions discovery).

3) Interactions discovery pipeline: EID2 extracts three types of interactions from its evidence bases: organism-organism interactions, organism-location interactions and organism-organism-location interactions. Wardeh *et al* (2015) provides detailed explanation of the process.

4) EID2 Portal: publically accessible at: https://eid2.liverpool.ac.uk/. The portal enables users to browse through EID2 data, look up interactions for one or more organisms, and produce tailored maps.

[1] We distinguish between microbes: organisms of the five classes: bacteria, fungi, helminth, protozoa and virus; and pathogens, which are species or strains of these classes which cause disease in at least one host species.

Table 1 – Species level interactions between five classes of microbes/pathogens and the different types of hosts (rows).

	Bacteria	**Fungi*[1]**	**Helminth**	**Protozoa**	**Virus*[2]**	**Total**
Algae	142	90	0	1	8	**241**
Fungi	112	146	1	1	32	**292**
Aquatic	312	96	101	32	14	**553**
Amphibian	58	3	93	12	12	**180**
Fish	401	83	989	54	89	**1616**
Reptiles	74	11	76	42	14	**217**
Aves	238	37	107	134	102	**618**
Rodents	187	23	110	67	205	**592**
Bats	65	4	13	16	94	**192**
Primates	78	9	246	132	131	**595**
Human	1219	466	203	82	262	**2232**
Carnivores	402	146	429	131	108	**1216**
Ruminants	501	142	340	202	152	**1337**
Other livestock*[3]	353	110	242	114	163	**982**
Other mammals	40	36	113	78	41	**308**
Cetacea	12	32	26	4	6	**80**
Arthropod vector	553	315	28	112	153	**1161**
Arthropod	430	215	57	71	73	**846**
Spermatophytes	1453	5112	211	14	907	**7697**
Tracheophytes	33	83	4	0	2	**122**
Bryophytes	55	45	0	0	11	**111**
Worms	252	28	10	10	4	**304**
Unique hosts	**4305**	**2018**	**4852**	**1479**	**3096**	**11288**
Unique microbes/pathogens	**3263**	**2691**	**5808**	**793**	**1848**	**14403**

[1] Includes oomycetes

[2] Includes viroid and prions.

[3] Comprises equids, cameladae, Suina and Leporidae (rabbits and hares).

To date, EID2 has extracted around 200,000 interactions between organisms (at any taxonomic level, including genus, species, and subspecies). Roughly 94% are between the five classes of pathogens/microbes, and 5% between arthropod vectors and hosts. Each of those interactions is linked to evidence (either publication or sequence meta-data) which can be filtered by date, or other values (such as type of scientific publication, affiliation of authors, location or name of journal etc.). EID2 has information on bacterial interactions with 8674 host species. The top 10 bacterial species in terms of number of hosts identified per bacterial species are: *Escherichia coli* (461, 10.71%), *Bacillus subtilis* (298, 5.64%), *Bacillus cereus* (281, 5.23%), *Wolbachia pipientis* (243, 5.65%), *Bacillus pumilus* (225, 5.23%), *Bacillus* megaterium (181, 4.21%), *Aeromonas hydrophila* (168, 3.90%), *Salmonella enterica* (146, 3.39%), *Agrobacterium tumefaciens* (145, 3.37%) and *Bacillus thuringiensis* (140, 3.25%). As Table 1 illustrates, EID2 data are sensitive to research trends and economic biases (e.g. human, food animals and crop centred). In order to correct for this we are working on enhancing our interaction discovery process, and on expanding our evidence base so that it contains research indexed and curated by multiple sources. Incorporating data from other sources, such as the OIE WAHIS database, which collects mandatory weekly/monthly reports on a cohort of domestic and wild animal infections from animal health organisations from around the world[1]. EID2 has also obtained the geographical ranges (~400,000 locations) of 180,000 organisms. On-the-fly maps of most can be produced using the portal.

Applications of the EID2 system

EID2 data and system have been utilised to provide insight into vector-borne diseases (Blagrove *et al.*, 2017), climate sensitivity of important human and animal diseases (McIntyre *et al.*, 2017), aquatic culture (Murray *et al.*, 2016) and relation between domestication time and risk of zoonotic infection (Morand *et al.*, 2014).

Network analysis was used in Morand *et al.* (2014) to visualise the overall interactions between humans and domestic animals and estimate which hosts are potential sources of parasites/pathogens for humans by investigating the network architecture and centrality. Ecological networks, in which nodes represent species, and links illustrate different interactions between those species, have been used to investigate a wide range of important phenomena such as food webs, predation, mutualism, and parasitism. In multi-host ecological networks, nodes are host species which are linked through sharing of microbes/pathogens. Two host species have a stronger connection when they share more microbes/pathogens. Sharing of microbes/pathogens may be a result of common evolutionary descent, or regular and prolonged interactions (such as through domestication or food-webs). This type of networks is an excellent example of both the power and challenges of Big Data, as on the one hand networks provide means to visualise and explore huge amount of data in an intuitive way; and on the other, the larger the network the more complex it is to analyse using traditional methods.

61,408 species level interactions between five classes of microbes and host species were recursively extracted from EID2 and transformed into multi-host networks depicting interactions between host species including all classes and each class individually. One especially relevant network metric, centrality, indicates the relative importance of nodes in the network. By implementing a centrality index based on a cohort of centrality measures (incorporating similarities between pathogen communities of connected hosts and weighted links measures), we can identify which hosts act as

[1] WAHIS, OIE World Animal Health Information System, (available at https://www.oie.int/wahis_2/public/wahid.php/Diseaseinformation/WI)

interspecies super-spreaders in a given network. Interspecies super-spreaders are particularly important to policy makers who develop disease management protocols, and to predict future disease emergence. Centrality can be estimated using several metrics. Here we implement a centrality index based on the most common metrics: degree, strength, closeness, eigenvector; and weighted networks metrics: Opsahl degree and Opsahl closeness (Opsahl *et al.,* 2010). Strength, closeness, Opsahl degree and Opsahl closeness were calculated twice: once using the number of shared organisms as weight metric, the second taking similarity between linked nodes as weight metric. The similarity was calculated using weighted Cohen's Kappa similarity coefficient between the two nodes (similarly to the method in Glass *et al*., (2015). Whereas strength, Opsahl degree (Opsahl *et al.,* 2010) and eigenvector centrality are more related to the pattern of direct co-sharing of microbe/pathogens with the rest of the hosts in the network, closeness captures aspects of indirect sharing through other species across the entire network.

Figure 1 displays the results of the centrality analysis, as well as a matrix representation of the networks aggregated by the host-type. The network of shared bacteria stands out in the number of shared microbes/pathogens. This could be because the underlying dataset includes commensal bacteria as well as pathogenic strains. However, only by keeping all the existing links would we be able to infer, at a later stage, the potential of pathogenic emergence. Figure 2 presents the bacterial network (nodes to the left are coloured by their type and to the right by their average centrality across all measures).

The E-I index was used to test how the different classes of microbes were shared between host groups (identified in Table 1). The index was calculated by dividing the number of ties internal to a group subtracted from that external to the other groups by the total number of ties. Hence, a positive E-I index indicates a tendency to share agents outside the group (i.e. extrovert) whereas a negative E-I index indicates a tendency to share agents within the group (i.e. introvert). As illustrated in Table 2, most host categories shared most of their pathogens with members from a different host category. Notable exceptions are spermatophytes (seed) plants (in fungi, helminths, virus and all -classes networks); fish (helminths and virus networks); amphibians (fungi networks); bats and non-human primates (helminths network); aves and arthropods (in protozoa network).

The EID2 system has also been utilised for the purposes of transmission route discovery. Single pathogens papers were extracted from EID2 evidence based and passed through an ensemble of classifiers trained with feature-reduced term-document matrices from a training set (built using EID2, and typed manually). Figure 3 presents some initial results (40,000+ publications processed for 600+ bacterial pathogens, over 100,000 publications for the working set of 3000+ pathogens).

Table 2 – E/I index across the six networks calculated for each of the different types of hosts (rows).

	Bacteria	**Fungi***	**Helminth**	**Protozoa**	**Virus***	**All classes**
algae	0.95	0.96	0	1.00	0.74	0.96
fungi	0.95	0.78	0	0	0.72	0.88
aquatic	0.88	0.97	0.72	0.22	0.48	0.87
amphibians	0.85	-0.97	0.51	0.56	0.26	0.75
fish	0.60	0.88	-0.25	0.38	-0.83	0.48
reptiles	0.95	0.95	0.90	0.76	0.84	0.91
aves	0.61	0.94	0.34	-0.44	0.07	0.20
rodents	0.84	0.73	0.49	0.81	0.72	0.78
bats	0.90	0.28	-0.22	0.77	0.57	0.66
non-human-primates	0.90	0.64	-0.05	0.64	0.16	0.72
primates with humans	0.90	0.83	0.10	0.64	0.25	0.74
human	1.00	1.00	1.00	1.00	1.00	1.00
carnivores	0.90	0.91	0.44	0.64	0.62	0.80
ruminants	0.92	0.95	0.24	0.80	0.75	0.87
other livestock*	0.98	0.98	0.90	0.96	0.95	0.98
other mammals	0.97	1.00	0.39	0.94	0.99	0.95
cetaceans	0.96	0.98	0.83	0.97	0.84	0.92
arthropod vector	0.25	0.67	0.54	0.72	0.64	0.42
arthropods	0.74	0.67	0.78	-0.68	0.20	0.72
spermatophytes	0.11	-0.71	-0.92	0.31	-0.95	-0.38
tracheophytes	1.00	0.98	1.00	0	1.00	0.98
bryophytes	0.99	0.99	0	0	1.00	0.99
worms	0.94	0.95	1.00	0.65	0.59	0.94

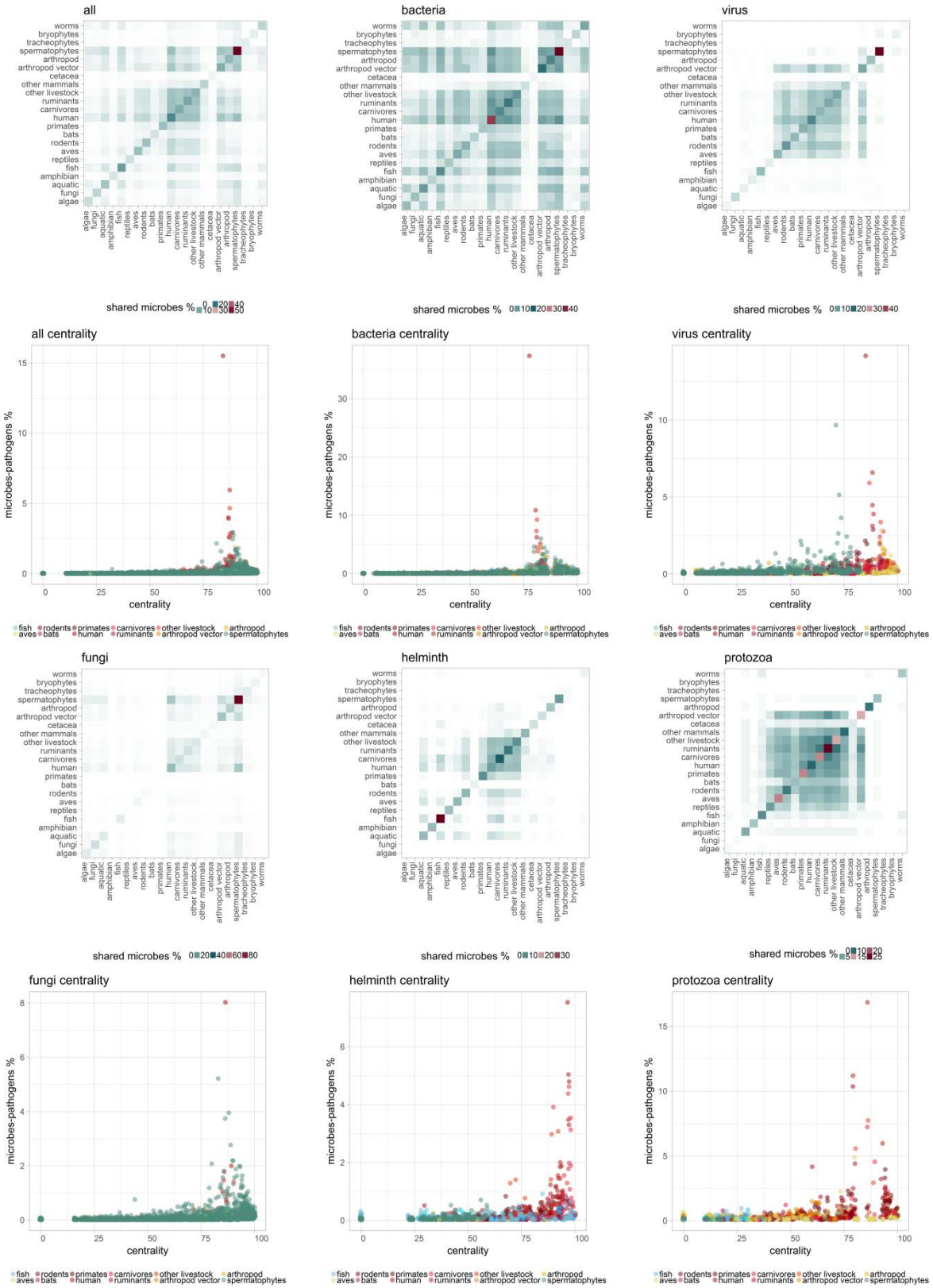


Figure 1 – Centrality and shared microbes/pathogens analysis results.

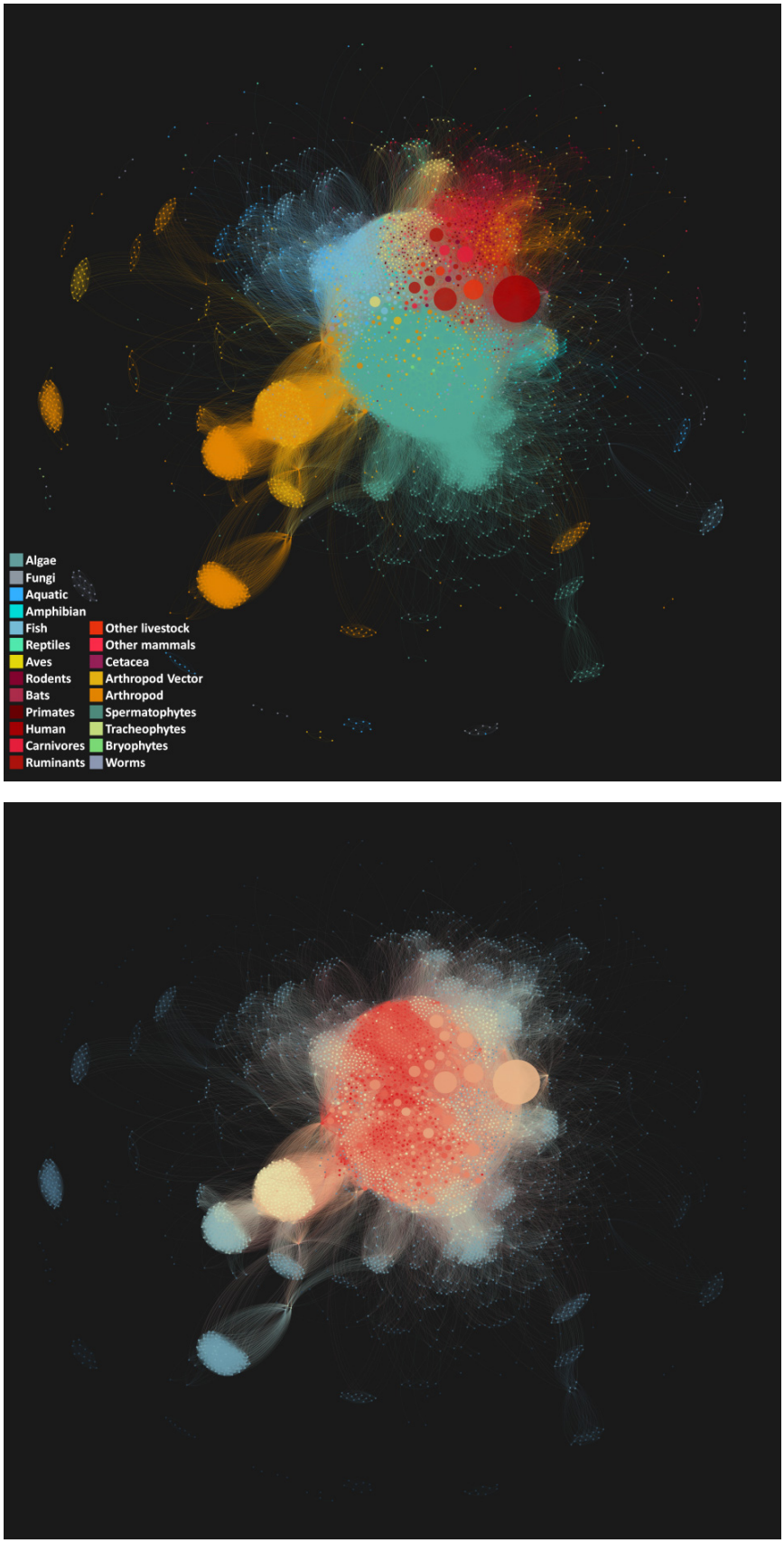


Figure 2 – Network of shared bacteria. Top (a): coloured by host type. Bottom (b): coloured by average centrality rank over the cohort of metrics (blue = low centrality, red = high centrality).

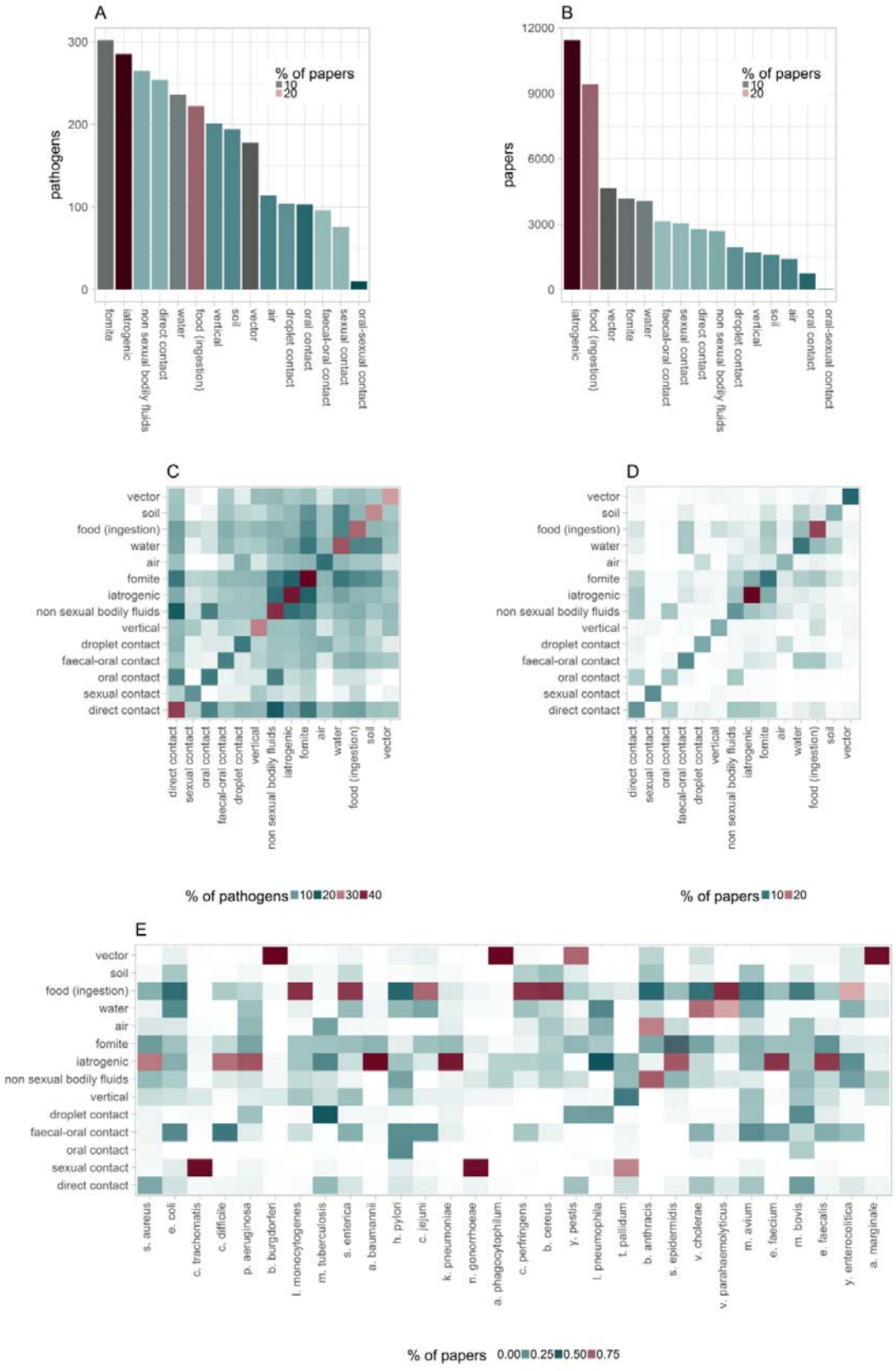


Figure 3 – detection pf transmission routes of bacteria. (a) Number of pathogens detected per route;(b) Number of papers per route; (c) co-detection of routes per pathogen; (d) co-mentions of route per paper; (e) routes detected (% of papers) of the top 35 bacteria in terms of papers linked to routes.

Challenges

Aside from the huge amount of scientific output which needs to be mined and assessed continually, the main challenges which systems like EID2 face in terms of Big Data are: 1) inconsistencies in the naming of organisms, particularly viruses where communities of researchers (e.g. human and plant virologist), tend to use same acronyms for completed different viruses (e.g. CMV is used both for cucumber mosaic virus and human cytomegalovirus); 2) trends in science, which often see an explosive number of research papers being published whenever there is an outbreak of a human disease, not often related to the disease at hand, whereas many other diseases remain understudied; 3) the multitude of taxonomies available and the continuous evolution of these taxonomies, which leads to information being missed due to inability to detect the full set of organisms mentioned in a paper; and 4) a reduction or lack of papers/sequences published on 'well known' interactions. For example if it is very well known that one given species is present everywhere in Asia or Europe, there are not many papers published with the exact locations of the said species, as such information is not noteworthy. Some of these challenges could be tackled with machine learning, intelligent coding and consolidating of resources; but others remain as open-ended, and as a result, correcting for sample biases is an important next step in the development of EID2 and its applications.

* this chapter is to be cited as :

Wardeh M., McIntyre M. and Baylis M. 2017. The Enhanced Infectious Disease Database (EID2) system of species interactions and location automatic detection and its applications. pp. 33 – 43 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

Sequencing the ocean

Chris Bowler[1], Frank-Oliver Glöckner[2] and Antonio Fernandez-Guerra[2]

[1] *Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Research University, 75005 Paris, France*
[2] *MPI for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany*

Abstract

Mankind has been exploring the ocean for thousands of years. Most recently the exploitation of DNA sequencing technology, developed largely for biomedical research, has generated comprehensive information about the diversity and distributions of oceanic microbes, and the genes they contain. These huge datasets permit the exploration of factors influencing biogeochemical processes, the evolution of life, and the transfer of carbon and energy to higher trophic levels. They also permit the study of how human activities are influencing the ocean, as well as the monitoring of human pathogens in order to prevent epidemics. Because ocean health is intimately connected with human health we can use genomics to understand the links between them.

1. Early history

There is nothing new about ocean exploration. For millennia mankind has been inspired by the lure of the far horizon and has sought to go where others had never been, most often in search of new trade routes and to map oceans, seas and coastlines so as to make the oceans safer for those who would follow. Sometimes, the objective of exploration was simply to be the first to discover something new, as was the case for the races to the North and South Poles. Scientific research was typically a welcome but only secondary pursuit, such as Joseph Banks' sojourn with Captain James Cook (1768–1771) on *HMS Endeavour* and Charles Darwin's voyage on *HMS Beagle* (1831-36) with Captain Robert FitzRoy. The first truly scientific expedition had to wait until 1872-1877, when Charles Wyville Thompson and John Murray aboard the *HMS Challenger* explored the world ocean to test the theory current at the time that the ocean below the sunlit upper layer was 'azoic,' devoid of life. They disproved the theory and in fact found thousands of spectacularly new planktonic organisms, subsequently introduced to the general public by Ernst Haeckel's fabulous art work. The voyage of the *Challenger* is now considered the first oceanographic expedition.

For over a century, oceanographic research has advanced apace with technology. In addition to measuring discrete parameters from water samples collected onboard ocean-going research vessels, arrays of autonomous floats and gliders can relay information directly from the ocean to laboratories, and remote sensing from space further allows us to monitor ocean processes from afar. Thanks to these technologies we today have a very good understanding of ocean processes from a physico-chemical perspective. On the other hand, the biology of the ocean has proved much more difficult to

constrain beyond measurements of chlorophyll, which can be easily measured from space, as a proxy for primary productivity (photosynthesis) in the surface of the ocean. Our inability to integrate biological processes into the functioning of the ocean is a serious issue, particularly concerning the microscopic planktonic organisms. These organisms collectively represent the majority of the biomass in the oceans, they drive many of the biogeochemical cycles of elements upon which the Earth system depends, and they support essentially all the food chains in the ocean. As one case in point, microscopic phytoplankton is responsible for half of the primary production on Earth, and they sustain the biological carbon pump, which can sequester carbon in the deep ocean for millennia.

2. Recent developments

At the same time that oceanographic exploration has progressed, advances in molecular biology, principally in biomedical research, have led to our ability to sequence DNA, the blueprint of life, at high-throughput and at very little cost. In the last 15 years, the cost of DNA sequencing has decreased by around one million-fold (Green *et al.*, 2017). The revolution in DNA sequencing now drives much of biological research. On the other hand, its impact on oceanographic research has been limited and it is far from becoming a mainstream tool integrated with other systems for ocean observation (Bowler *et al.*, 2009).

Two ocean expeditions have to date exploited DNA sequencing technology to explore marine plankton at global scale. The first was Craig Venter's Global Ocean Sampling (GOS) expedition on the *Sorcerer II* from 2004-2006 (Rusch *et al.*, 2007), the second was *Tara* Oceans from 2009-2013 (Karsenti *et al.*, 2011; Karsenti, 2015). Coming first, GOS was much more limited in scope, focusing only on bacteria in surface waters, essentially only in tropical and sub-tropical waters. Notwithstanding, thousands of research articles have now been published based on the GOS data set and several major discoveries have emerged, such as the widespread distribution of proteorhodopsin genes (de la Torre *et al.*, 2003). *Tara* Oceans had much greater ambitions, to characterize the entire plankton community, including viruses, protists and zooplankton in addition to bacteria, to sample down to 1,000 meters and to reach latitudes from the tropics to the poles. The massively parallel Illumina sequencing technology deployed by *Tara* Oceans was furthermore much more powerful than the Sanger-based technology used to sequence GOS samples several years earlier.

In 2015 the *Tara* Oceans consortium published five scientific papers in the journal Science presenting the initial wave of scientific results from the first six years of the project (Brum *et al.*, 2015; de Vargas *et al.*, 2015; Lima-Mendez *et al.*, 2015; Sunagawa *et al.*, 2015a; Villar *et al.*, 2015) The findings show the extraordinary diversity of plankton in the world's oceans, uncover many of the interactions between them, and reveal how plankton influence and are influenced by the environment. One of the papers (Sunagawa *et al.*, 2015a) describes an ocean microbial reference gene catalog containing 40 million genes from marine microbes (bacteria, viruses, Archaea and picoeukaryotes). Derived from samples of seawater collected from all over the world and at depths down to 1,000 meters the authors show that this publicly available DNA sequence dataset is more than 1,000 times larger than the GOS data set and is likely to have captured all of the abundant microbial genes present in the areas sampled.

The *Tara* Oceans project has captured the attention and imagination of both scientists and the general public, with TV, film and radio presentations, as well as thousands of articles being written in the international press (http://oceans.taraexpeditions.org/en/). Starting as a grass-roots initiative by a group of academic scientists, the research-enabled 36 meter schooner *Tara* spent almost four years circumnavigating the globe and going around the Arctic Circle. Overall, *Tara* Oceans sampled plankton at more than 210 sites and at multiple depth layers in all the major oceanic regions. The 35,000 samples collected from the expedition now form the basis for extensive processing, analysis, and data integration on land. One of the wettest experiments ever has generated more than seven terabasepairs of information, one of the largest contiguous sets of DNA sequence available to the scientific community (Alberti *et al.*, 2017; Karsenti, 2015; Pesant *et al.*, 2015; Sunagawa *et al.*, 2015b).

The enormous 40 million gene set described (Figure 1) by Sunagawa *et al* (2015a) is shown to derive principally from around 35,000 species of marine bacteria. Around 80% of these sequences are novel, and most of the novelty derives from oceanic regions traditionally undersampled (such as the Southern Ocean), and from the twilight zone below the upper +/- 100 meters of the water column where sunlight can pass. The microbial communities are quite different from one location to another, and the authors found that temperature is the main factor determining their composition. In fact just by looking at the species content they could predict the temperature of the water with a high precision – quite a sophisticated thermometer! The implication of this finding is that temperature changes in a future ocean impacted by climate change are likely to affect the functioning of the whole marine ecosystem, impacting the food chain and the biogeochemical cycles that depend on this microbial world.

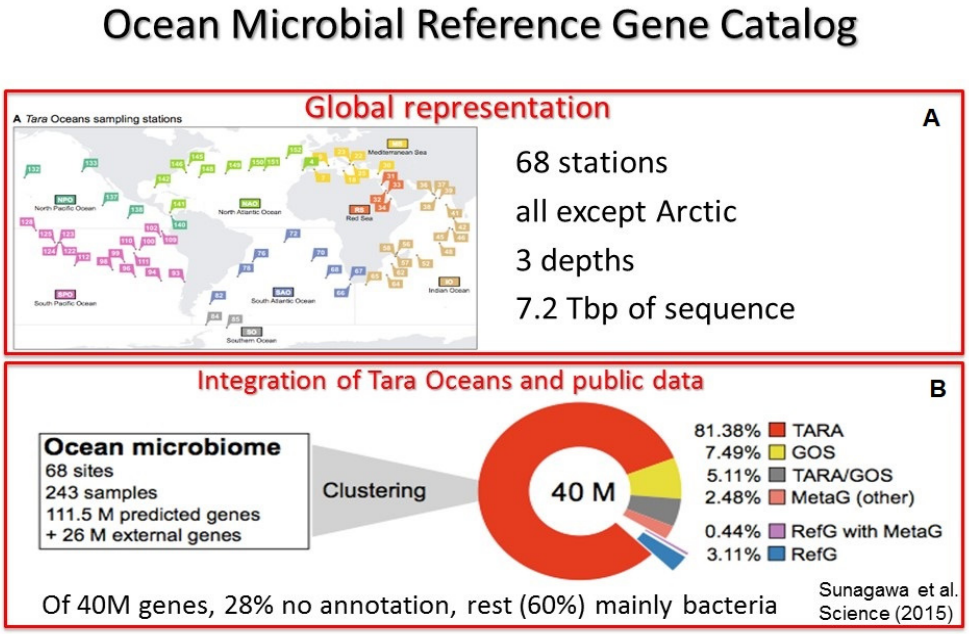


Figure 1: The *Tara* Oceans Microbial Reference Gene Catalogue. (A) Geographic distribution of 68 representative *Tara* Oceans sampling stations at which seawater samples and environmental data were collected from multiple depth layers. (B) The deep Illumina shotgun sequencing of 243 samples, followed by metagenomic assembly and gene prediction, resulted in the identification of >40 M reference genes. The combined clustering of genes identified in *Tara* Oceans samples with those obtained from public resources allowed annotation of genes according to the composition of each cluster. For example, a gene was labeled as: "TARA/GOS" if its original cluster contained sequences from both *Tara* Oceans and GOS samples. More than 81% of the genes were found only in samples collected by *Tara* Oceans. A breakdown of taxonomic annotations shows that the reference gene catalogue is mainly composed of bacterial genes. The catalogue is publicly available (www.ebi.ac.uk/services/tara-oceans-data) and represents 7.2 Tbp of data.

An interesting aspect of the work was to compare the ocean microbiome with the human gut microbiome, on the premise that both represent single contiguous ecosystems. The 40 million genes in the global ocean compare to ca. 10 million genes in the human gut microbiome, so the numbers are quite similar. Furthermore, in terms of abundance close to ¾ of ocean and gut genes are the same, in spite of them being very different ecosystems and containing different microbes. On the other hand, around 90% of these gut microbiome genes have putative functions assigned to them, which compares to less than 60% of the ocean microbiome. These striking differences reflect the priorities of funding biomedical research at the expense of environmental research, even though a healthy ecosystem promotes human health too. Another striking difference between the ocean and gut microbiomes is that whereas many gut microbiome genes are involved in signaling and cell-cell communication, the

primary signal from the ocean microbiome is one of nutrient transport, energy utilization and basic metabolism.

Life evolved in the ocean, and the complex consortium of organisms that make us human is not unlike the organisms that co-evolved to adapt to life in marine plankton ecosystems. The striking similarities between the two ecosystems are food for thought to ponder about our origins and to remember the importance of marine plankton ecosystems for generating and maintaining planet Earth habitable for us. In spite of our anthropocentric focus on biomedical research related to human healthcare, we should not forget the microbial life support systems that made our planet habitable for us in the first place and that continue to do so. In today's world of changing climate, the strong sensitivity of marine microbes to temperature should be a wake-up call to stop exploiting marine resources and treating the ocean as a garbage site for everything we want to get rid of.

Mankind is currently consuming annually a quantity of oil and gas fossil fuels that represent one million years of deposition by marine planktonic life. It is clear that the release of CO2 and other greenhouse gases from such a huge quantity of carbon reserves is having a huge impact on the ocean, from increasing temperatures and acidification, as well as deoxygenation (Figure 2).

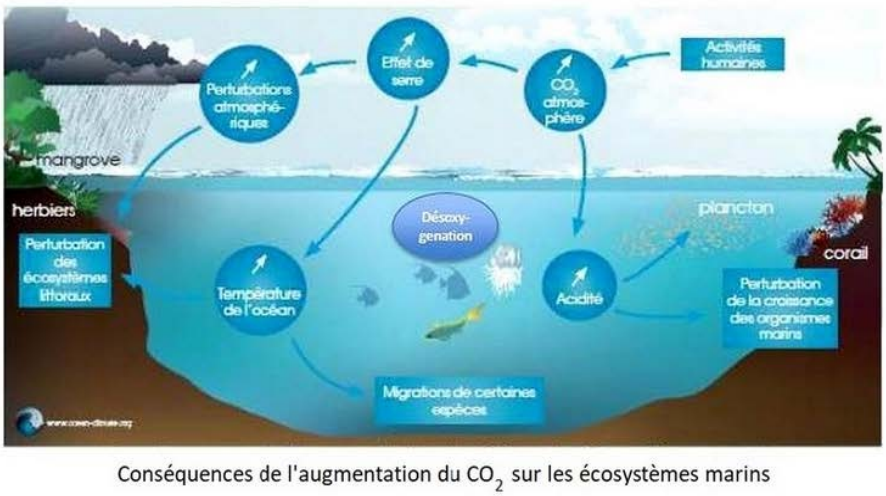


Figure 2: An ocean under pressure. Increased atmospheric temperatures and CO2 from anthropogenic activities are increasing ocean temperatures and leading to ocean acidification. These combined processes are also resulting in ocean deoxygenation. Collectively, these processes are affecting coastal and open ocean ecosystems in multiple ways. Figure from the Ocean and Climate Platform.

3. Human health

Contamination of the ocean with microplastics and other toxins are additional concerns because they can have detrimental effects on human health. Indeed, there are many direct links between ocean health and human health (Figure 3).

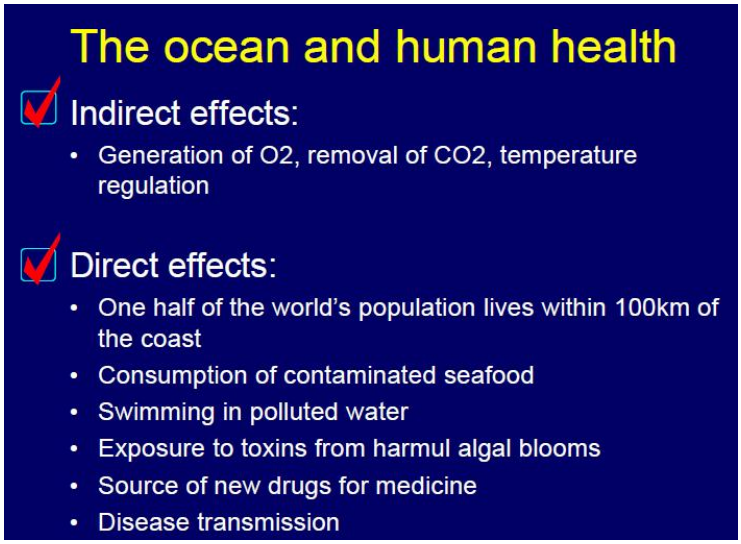


Figure 3: The health of the ocean is linked to human health in multiple direct and indirect ways.

One of the first direct connections between ocean and human health was the case of cholera, because it was found that the aetiological agent of the disease, *Vibrio cholerae*, is in fact a commensal of zooplankton in marine environments. Consequently, predictions about the risk of cholera outbreaks can now be made by monitoring plankton populations, and preventative measures can be taken to filter contaminated water to remove zooplankton using a simple sari cloth (Figure 4).

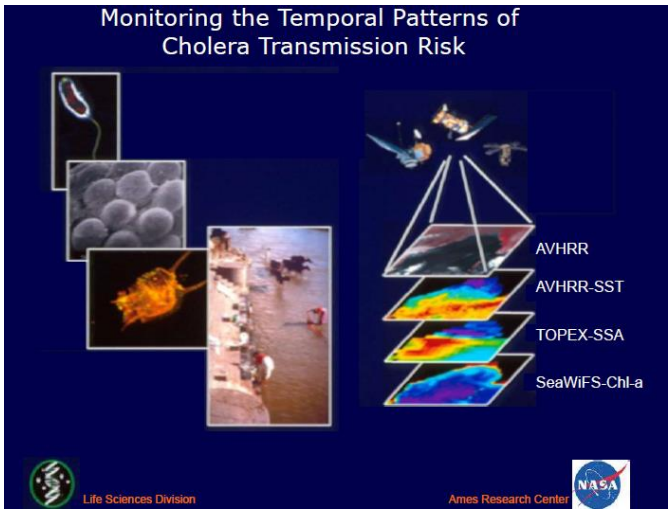


Figure 4: Monitoring the temporal patterns of cholera transmission risk. Satellite-based remote sensing is used to monitor phytoplankton blooms in the Bay of Bengal, which can increase incidence of cholera on land due to proliferation of copepod zooplankton, a commensal carrier of *Vibrio cholerae*. Figure courtesy of Rita Colwell.

Beyond this first example, efforts have been made to map the various types of human impact on marine ecosystems, including shipping, acidification, invasive species and temperature change, and global maps have been generated by Benjamin Halpern at the University of California Santa Barbara, collectively providing an ocean health index (Halpern *et al.*, 2012) (Figure 5).

The availability of such data allows human impacts to be analyzed with respect to plankton community structure and gene expression. Such analyses can reveal the incidence of potential pathogens in marine environments, as well as a myriad of other human impacts, such as the distribution of genes involved in xenobiotic detoxification or antibiotic resistance genes. For such studies the availability of genomics-enabled observatories at coastal sites will be essential.

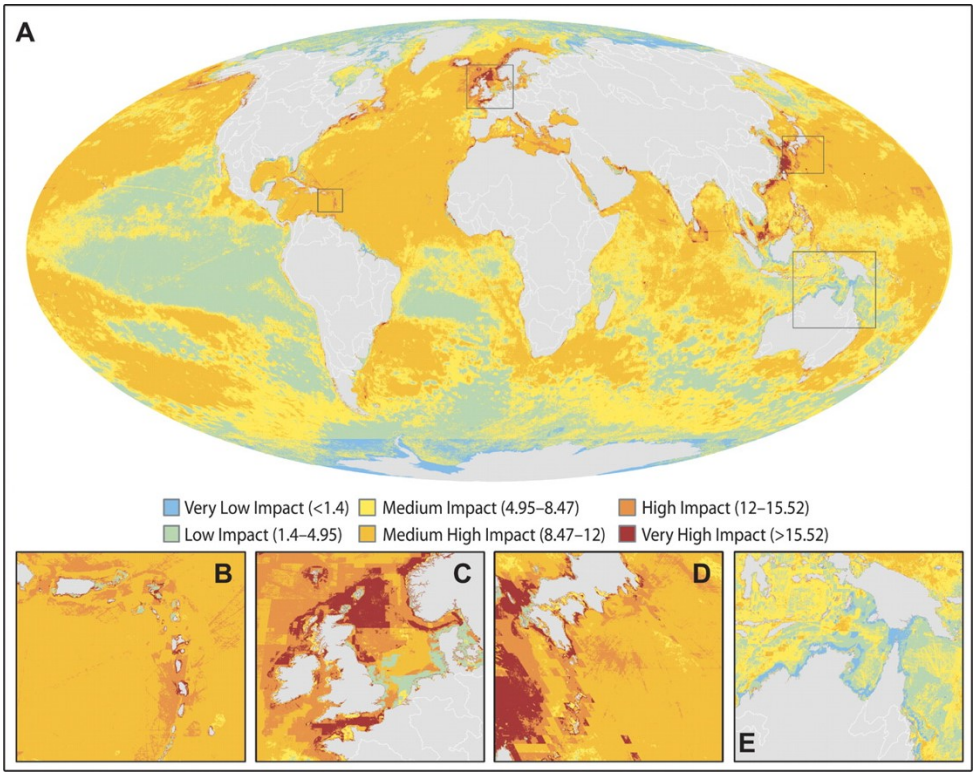


Figure 5: The impacts of human activities on the ocean. The map shows a compilation of the combined impacts of human activities on the ocean, derived from multiple parameters including increased temperatures and ocean acidification, maritime shipping, pollution, and incidence of invasive species. See Halpern *et al.* (2012) for further details.

The genomics data generated by the Ocean Sampling Day initiative (OSD; Kopf *et al.*, 2015) represents an ideal data set to perform such analyses because the majority of samples were collected at coastal sites (Figure 6). Comparison of the incidence of antibiotic resistance genes in OSD samples with respect to *Tara* Oceans samples indeed reveals the prevalence of such genes at coastal sites (Figure 7). This finding highlights that the impact of mankind on the ocean is truly pervasive – just as microplastics represent a purely man-made addition to ocean ecosystems since the middle of the 20th century, so the widespread use of antibiotics in modern medicine has led to the widespread occurrence of antibiotic resistance genes in marine microbes.

Databases containing DNA sequence information from potential pathogenic organisms can provide valuable references that can help to detect such microbes in omics data sets derived from the ocean, such as OSD and *Tara* Oceans. The Enhanced Infectious Disease Database (EID2) is a particularly useful source of sequences from human pathogenic bacteria (see Wardeh, this volume). Moving forward, the ability to generate Metagenome-Assembled Genomes (MAGs) and longer sequencing reads from new DNA sequencing technologies (Green *et al.*, 2017) will provide more powerful approaches because they will permit the identification of pathogenicity islands and plasmids involved in infection.

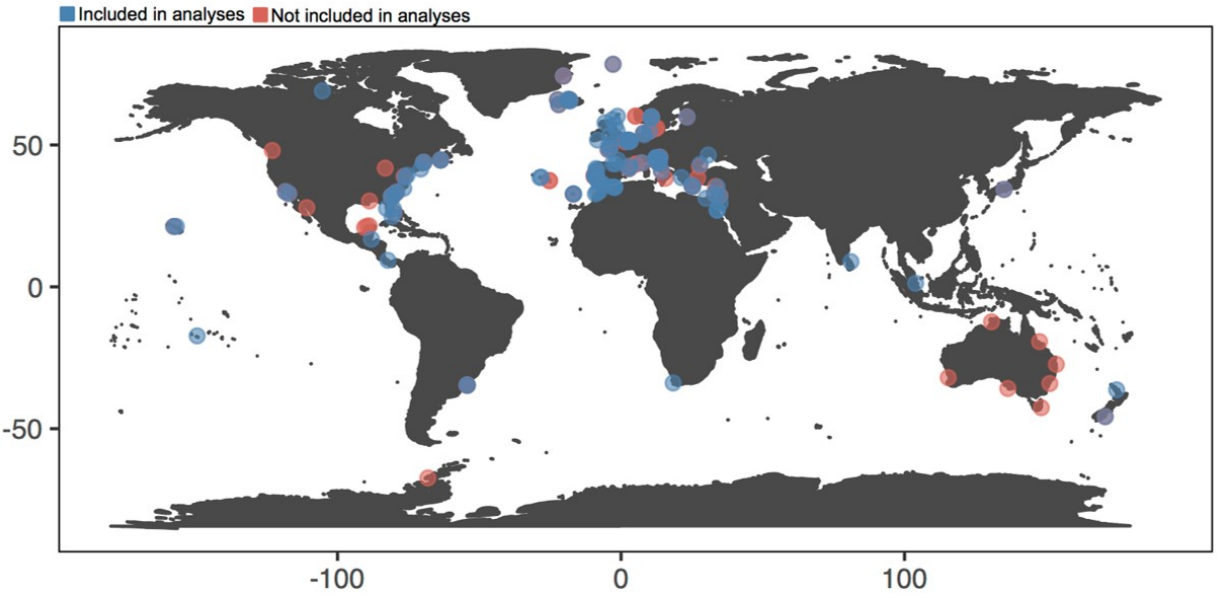


Figure 6: Distribution of samples collected on Ocean Sampling Day, June 21st 2014. Ninety percent of samples were collected from within 10 km of the shore, allowing human impacts to be examined in detail. See Kopf *et al.* (2015) for further details.

Projecting even further into the future, genomics is likely to become a standard component of ocean observation systems, and sequencing DNA in real time in the water column will add further possibilities for pathogen detection (Bowler *et al.*, 2009). Notwithstanding, while detection of potential human pathogens in the ocean can now be considered quite straightforward, it is also necessary to measure their pathogenic potential. This will require a better understanding of pathogenicity and the ecology of marine microbe ecosystems, indicating that additional aspects such as adhesion, toxin generation, invasion, colonization, and competitive fitness, will also need to be addressed (see Le Roux, this volume). Ultimately, the investments made in monitoring pathogens in the ocean will depend on the value we give to ocean health, which will be dependent in turn on the importance we give it with respect to human health. As we turn more to the ocean for food production, it will also be related to the development of a sustainable aquaculture industry.

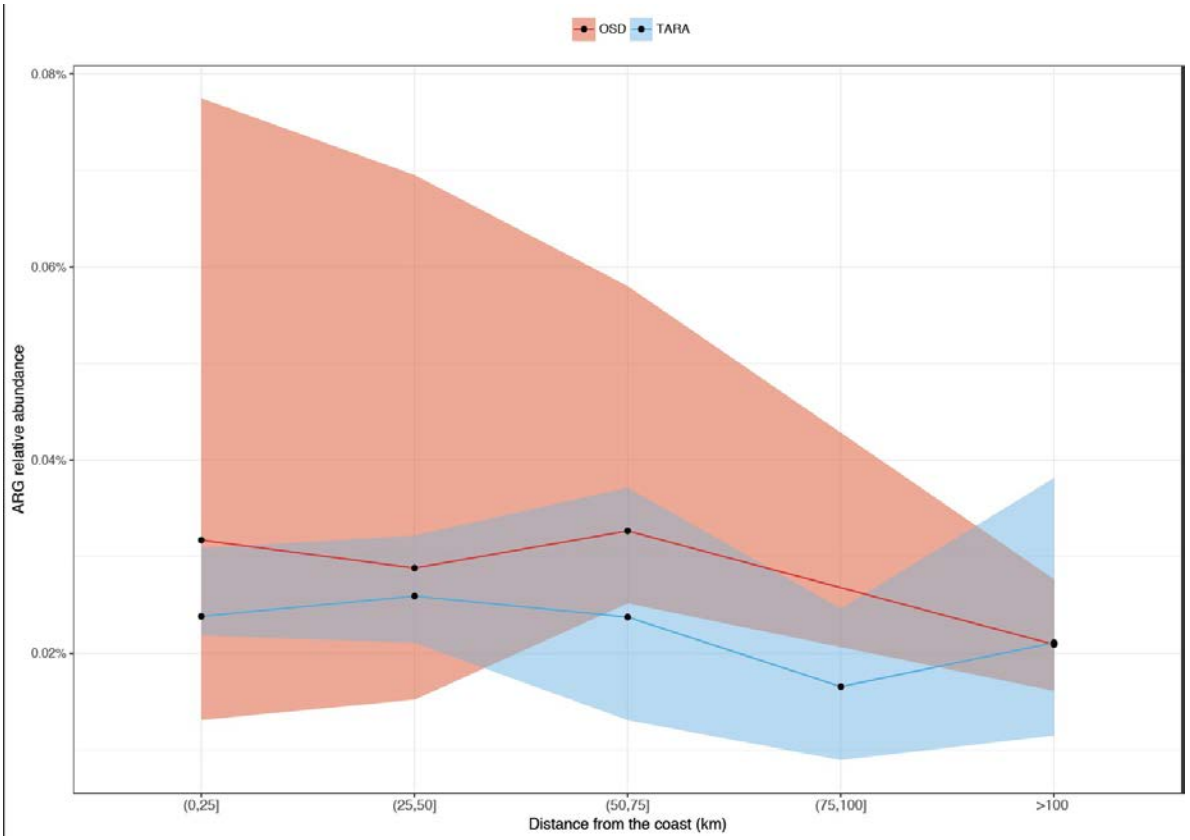


Figure 7: Relative abundance of antibiotic resistance genes (ARG) in *Tara* Oceans and OSD samples with respect to distance from the coast. The lines indicate mean values; the light red and light blue shaded areas represent the variations in abundance.

In conclusion, the incorporation of genomics into ocean observing platforms can be instrumental in better understanding the impacts of climate change on our ocean and in assessing the profound impacts of mankind on marine ecosystems. Political decisions need to reflect the urgency of understanding the current changes in the ocean and it is reassuring that policy makers at the United Nations Conference on Climate Change (COP21) and the Paris Treaty have recognized the importance of protecting the ocean, but words must now be replaced by concrete actions.

* this chapter is to be cited as :

Bowler C., Glockner F.O. and Fernandez-Guerra A. 2017. Sequencing the ocean. pp. 45 – 52 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

Finding Statistically Significant Patterns from Data

Mahito Sugiyama[1, 2]

[1] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan.*
[2] *JST PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan.*

Abstract: We introduce the technique of *significant pattern mining*, which enables us to find combinatorial *patterns*, such as feature sets, subgraphs, or sequences that are statistically significantly enriched in a class of samples. The technique offers the *p*-value of each pattern, which shows how much it is associated with classes, e.g., two groups of cases and controls in a case-control study. The resulting FWER (family-wise error rate) of the entire collection of patterns is rigorously controlled under the predetermined threshold α. Significant pattern mining has a wide range of potential applications for example in computational biology, statistical genetics, and healthcare.

Key words: Pattern mining, statistical significance, family-wise error rate, multiple testing corrections

1. Introduction

Pattern mining (Aggarwal and Han, 2014) is the process of finding interesting feature combinations, or *patterns*, from data and is a central topic in the area of data mining and big data analysis (Zaki and Meira Jr., 2016). Various types of patterns have been analyzed in applications: *itemsets* (Han *et al.*, 2000), combinations of binary features (variables), originally used in market basket analysis to find frequently copurchased items and recently used in other applications such as GWAS (Zhang *et al.*, 2014), *subgraphs* (Yan and Han, 2002), often used in drug discovery to detect commonly occurring substructures in a set of chemical compounds modeled as graphs (Takigawa and Mamitsuka, 2013), and sequences (Pei *et al.*, 2001), employed in such as DNA sequence analysis and customer behavior analysis.

As an extension of the traditional pattern mining problem, *significant* (*discriminative*) *pattern mining* is recently developed, which finds patterns enriched in one class relative to another class. For example, in a case-control study, one can find patterns that are enriched in the case group while not in the control group. The objective is to find *all* patterns that are statistically significantly associated with the class variable while correcting for multiple testing to ensure rigorous control of the *FWER* (*family-wise error rate*), that is, the probability to detect one or more false positive patterns. Significant pattern mining can provide *p*-values, which are indispensable in scientific fields such as biology and

medicine, and it has been actively studied and applied to various types of data, including itemsets (Llinares-López *et al.*, 2015; Papaxanthos *et al.*, 2016; Terada *et al.*, 2016; Terada *et al.*, 2013), association rules (Hämäläinen 2012; Webb 2007), subgraphs (Sugiyama *et al.*, 2015), and contiguous intervals (Llinares-López *et al.*, 2015; Llinares-López *et al.*, 2017).

In this paper, we give an overview of significant pattern mining. First, we review the current techniques of significant pattern mining in Section 54. Next, we introduce the problem setting of pattern mining and formulate it in Section 3. We extend to significant pattern mining in Section 4. To assess the statistical significance of each pattern, we use hypothesis testing, in particular, Fisher's exact test in Section 4.1, followed by discussing how to correct for multiple testing in Section 4.2. Then we briefly introduce the key technique of significant pattern mining, the *testability* of patterns, in Section 4.3. Finally, we conclude the paper in Section 5.

2. Review of Significant Pattern Mining

In the following, we summarize the current state-of-the-art of significant pattern mining techniques.
To control the FWER, the *testability* is widely used across methods (see Section 4.3 for a brief introduction), which was first introduced into pattern mining by Terada *et al.* (2013) to achieve significant pattern mining. The method is called LAMP and it is implemented in PLINK as LAMPLINK (Terada *et al.*, 2016) to use it in GWAS and is also parallelized (Yoshizoe *et al.*, 2015) for further efficiency. To achieve the optimal FWER control, Wesfall-Young permutation (Westfall and Young, 1993) is employed in addition to the testability in FastWY (Terada *et al.*, 2013) and it is further improved in terms of runtime and memory efficiency in the method Westfall-Young light (Llinares-López *et al.*, 2015), which is the current state-of-the-art. To include categorical covariates, which often exist in GWAS, FACS (Papaxanthos *et al.*, 2016) and LAMP-ELR (Terada *et al.*, 2016) were proposed.
All of the above methods can treat only itemsets, where each data point is a binary vector and each pattern is a set of (binary) features, except for Westfall-Young light. To apply significant pattern mining to other data structure, significant subgraph mining was proposed (Sugiyama *et al.*, 2015) and was integrated in Westall-Young light. Moreover, significant pattern mining was also applied to find contiguous intervals in (Llinares-López *et al.*, 2015) and categorical covariates are considered in Llinares-López *et al.* (2017). These methods have been applied to find all arbitrary sizes of contiguous intervals of single nucleotide polymorphisms (SNPs) in the genome that are jointly associated with the phenotype, and Llinares-López *et al.* (2015) have reported that 70% of significant intervals detected by the method in the *Arabidopsis thaliana* GWAS dataset are novel intervals that were not known before.

In addition to control the FWER by the above methods, recently Komiyama *et al.* (2017) proposed a method that can control the FDR.

3. A Primer of Pattern Mining

We begin with an example of market basket analysis, which is a basic scenario of pattern (itemset) mining. Let V be a set of items in a supermarket, for example, $V = \{\text{Milk, Egg, Bread, Potato}\}$, and each subset of V, like $\{\text{Milk}\}$ or $\{\text{Egg, Bread}\}$, is called a *pattern* or an *itemset*. Note that this set V can be massive in the standard situation as there are often tens of thousands of items in a supermarket. Each data point, or a *transaction*, x in a dataset D is a set of copurchased items by an individual customer. For example, if a customer id1 bought milk and egg, the corresponding data point $x_1 = \{\text{Milk, Egg}\}$, and a dataset D is a set of such data points for all customers (Table 1). It is equivalent if we use binary representation. An item corresponds to a binary variable and a data point becomes a binary vector $\boldsymbol{x} = (x^1, x^2, x^3, x^4)$, where four elements correspond to Milk, Egg, Bread, and Potato, respectively, and each element $x^j = 1$ if the corresponding item is purchased and

$x^j = 0$ otherwise. For example, the above data point $x_1 = \{\text{Milk, Egg}\}$ becomes $x_1 = (1,1,0,0)$ (see Table 1).

Table 1. Example of a dataset and the corresponding binary representation with four customers in market basket analysis.

Individuals	Copurchased items
x_1	Milk, Egg
x_2	Milk, Egg, Bread, Potato
x_3	Egg, Potato
x_4	Egg, Bread, Potato

Individuals	Milk	Egg	Bread	Potato
x_1	1	1	0	0
x_2	1	1	1	1
x_3	0	1	0	1
x_4	0	1	1	1

The importance of each pattern s, which is always a subset of V, is measured as the *support* of it, denoted by $\text{supp}(s)$. The support is defined as the number of customers who bought s, that is, $\text{"supp"}(x) = |\{x \in D \mid s \subseteq x\}|$, where $|A|$ denotes the number of elements in A. For example, in Table 1, $\text{supp}(\{\text{Egg}\}) = 4$ and $\text{supp}(\{\text{Egg, Bread}\}) = 2$. Here the goal of (frequent) pattern mining is to find all *frequent patterns* whose support is higher than σ, which is a threshold predetermined by the user.

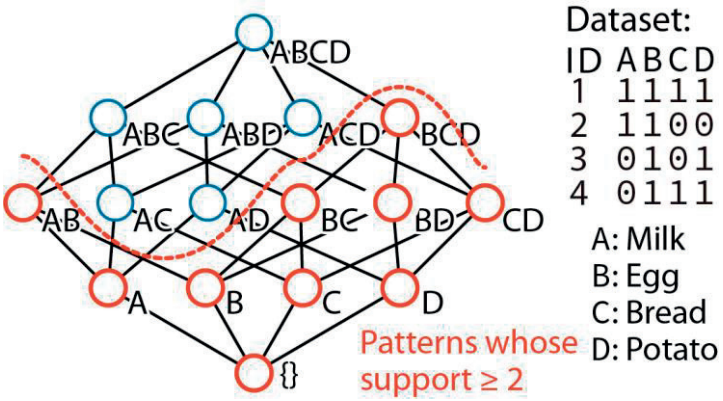
Apparently, pattern mining seems to be an easy task as we can find all frequent patterns by the following simple *generate-and-test strategy*: check each pattern one by one and output it if its support is larger than σ. However, the difficulty of pattern mining stems from the massiveness of the number of candidate patterns, which is $2^{|V|}$ as we need to consider all subsets of V. In Table 2, we summarize approximate time required to check all patterns with respect to changes in the number of items. As we can see, it is unfeasible to check all patterns even if the number of items is 70. This effect is often called *combinatorial explosion*. Since the number of items easily becomes tens of thousands in real-world situations, our simple generate-and-test strategy does not work at all.

The key to solve the above problem of combinatorial explosion and achieve frequent pattern mining is the *Apriori principle* (Agrawal, et al., 1993; Agrawal and Srikant, 1994). This is based on a simple observation as follows: For any pattern s, none of its super patterns, that is, patterns that include s can have a support larger than that of s. This means that if a pattern s is not frequent, its super patterns can never be frequent. In our example dataset in Table 1, $\text{supp}(\{\text{Milk, Egg}\}) = 2$, $\text{supp}(\{\text{Egg, Bread}\}) = 2$, $\text{supp}(\{\text{Egg, Potato}\}) = 3$, $\text{supp}(\{\text{Milk, Egg, Bread}\}) = 1$ $\text{supp}(\{\text{Milk, Egg, Potato}\}) = 1$, $\text{supp}(\{\text{Milk, Bread, Potato}\}) = 2$ $\text{supp}(\{\text{Milk, Egg, Bread, Potato}\}) = 1$ are always equal to or smaller than $\text{supp}(\{\text{Egg}\}) = 4$.

Table 2. Approximate time required to check all patterns.

The number $\|V\|$ of items	Approximate time required
10	0.00000057 sec.
20	0.00059 sec.
30	0.6 sec.
40	10.2 min.
50	174 hours.
70	7 million days
100	8 thousand billion days

This Apriori principle allows us to use a *branch-and-bound strategy* to find frequent patterns. We check from smaller patterns to larger patterns. In each step, if we encounter a pattern whose support is lower than the threshold, then we can *prune* all its super patterns without losing any frequent patterns (see Figure 1 below). This strategy can dramatically reduce the number of patterns to be checked and makes pattern mining possible.



The same formulation can be applied to other data structures. A representative example is *subgraph mining* (Inokuchi, *et al.*, 2000; Yan and Han, 2002), where each data point is a *graph*, and each pattern is a *subgraph* of it. Given a collection of graphs as a dataset, the goal of subgraph mining is to find all frequent subgraphs, where each subgraph is frequent if the number of graphs (data points) that includes the subgraph is higher than the predetermined threshold.

4. Significant Pattern Mining

Next, we consider the case in which a dataset D is divided into two groups C and $\overline{C}$, and always assume that $|C| \leq |\overline{C}|$. This setting is called *contrast pattern mining* (or emerging pattern mining,

discriminative pattern mining) (Dong and Bailey, 2013), and corresponds to supervised learning in the field of machine learning. For example, in a case-control study, C and $\overline{C}$ correspond to case and control groups, respectively. Then the goal is to find patterns that are more frequent in one class than in the other, that is, they are *associated with the class membership*. Significant pattern mining is a branch of contrast pattern mining in which we use statistical tests to quantify the discriminability of patterns as *p*-values.

4.1. Statistical Association Testing

Statistical association testing controls the false positive rate. A pattern is *false positive* if it is judged to be associated with class membership, while the truth is that it is independent of it. To examine the statistical association of supports of patterns, we use *contingency tables*. When we focus on a single pattern s, the contingency table with respect to two groups C and $\overline{C}$ and the occurrences/non-occurrences of s can be obtained in Table 3, where $\mathrm{supp}_C(s)$ is the support with respect to the group C, that is, $\mathrm{supp}_C(s) = |\{x \in C \mid s \subseteq x\}|$.

The statistical association can be tested by the *Fisher's exact test*, and if the *p*-value obtained by the test is smaller than the significance level α, which is predetermined by the user and often set to be $\alpha = 0.05$, it is called *statistically significant*. The probability of observing the value $\mathrm{supp}_C(s)$ is known to be obtained as $q = \binom{|C|}{\mathrm{supp}_C(s)}\binom{|\overline{C}|}{\mathrm{supp}_{\overline{C}}(s)} \Big/ \binom{|D|}{supp(s)}$ and its *p*-value is computed by summing up all probabilities of more extreme values than $\mathrm{supp}_C(s)$.

Table 3. Contingency table.

	Occurrences	Non-occurrences	Total
C	$\mathrm{supp}_C(s)$	$\lvert C\rvert - \mathrm{supp}_C(s)$	$\lvert C\rvert$
$\overline{C}$	$\mathrm{supp}_{\overline{C}}(s)$	$\lvert\overline{C}\rvert - \mathrm{supp}_{\overline{C}}(s)$	$\lvert\overline{C}\rvert$
D	$\mathrm{supp}(s)$	$\lvert D\rvert - \mathrm{supp}(s)$	$\lvert D\rvert$

4.2. Multiple Testing Correction

If we test only one pattern and its *p*-value is smaller than α, it is guaranteed that the false positive rate is controlled under α. However, in pattern mining, we have massive patterns and we need to test all of them to find all significant patterns. Here, if we test m patterns and output all significant patterns whose *p*-values are smaller than α, $m\alpha$ patterns could be false positives. Since the number m is massive in pattern mining, $m\alpha$ is also massive, hence a large number of false positives can be included in the output. For instance, if there are 100,000 items, i.e., $|V| = 100{,}000$, the number of patterns becomes 2^{100000}, thus under the setting $\alpha = 0.01$ the number of false positives is $2^{100000} \cdot 0.01 = 10^{30101}$, which is not acceptable in most applications.

To solve the problem and appropriately control the false positive rate, *multiple testing correction* is needed. In particular, one needs to control the *FWER* (*family-wise error rate*), the probability of one or

more false positive patterns occurring, using a new *p*-value threshold δ , which is always smaller than α . Let $\mathbf{FWER}(\delta)$ be the FWER under the threshold δ . Although the ultimate goal of multiple testing correction is to find $\delta_{\mathbf{opt}}$ satisfying $\mathbf{FWER}(\delta_{\mathbf{opt}}) = \alpha$, it is impossible to analytically obtain this ideal threshold $\delta_{\mathbf{opt}}$. Thus the objective of multiple testing correction is to efficiently find δ as large as possible while satisfying $\mathbf{FWER}(\delta) = \alpha$.

The most popular method of multiple testing correction is *Bonferroni correction* (Bonferroni, 1935; Bonferroni, 1936), which simply sets $\delta_{\mathbf{Bon}} = \alpha/m$, where m is the number of tests, or the number of patterns in our case. Then $\mathbf{FWER}(\delta_{\mathbf{Bon}}) < \alpha$ is always satisfied. However, it is well known that Bonferroni correction is too conservative, that is, $\delta_{\mathbf{Bon}}$ is too small and the resulting $\mathbf{FWER}(\delta_{\mathbf{Bon}})$ is also too small. It is even more extreme in our case as the number m of patterns is massive in pattern mining.

4.3. Testability

The key to solve multiple testing correction in pattern mining is the testability introduced by Tarone (1990), which allows for removing untestable patterns without changing the resulting FWER. Terada *et al.* (2013) were the first to use the testability in pattern mining in their method LAMP, which was the first to achieve rigorous FWER control.

For each pattern s , let us denote by $\psi(s)$ the *minimum achievable p-value of* s when we fix the marginal of the contingency table, that is, $|C|, |\overline{C}|$, and $|D|$, and let $s_1, s_2, s_3, \cdots$ be all patterns sorted in increasing order with respect to the minimum achievable *p*-values, that is, $\psi(s_1) \leq \psi(s_2) \leq \psi(s_3) \leq \cdots$ holds. Then, for the k th pattern s_k such that $k \cdot \psi(s_k) < \alpha$ and $(k+1) \cdot \psi(s_{k+1})$, patterns $s_1, s_2, \ldots, s_k$ are called *testable*. Tarone (1990) showed that $\mathbf{FWER}(\delta_{\mathbf{Tarone}}) < \alpha$ is always satisfied even if we ignore untestable patterns and test testable patterns with the threshold $\delta_{\mathbf{Tarone}} = \alpha/k$.

In Fisher's exact test, the minimum achievable *p*-value is obtained for the most extreme case. For example, if $0 \leq \mathrm{supp}(s) \leq |C|$, $\mathrm{supp}_C(s) = \mathrm{supp}(s)$ holds, which leads to $\psi(s) = \binom{|C|}{\mathrm{supp}(s)} \Big/ \binom{|D|}{\mathrm{supp}(s)}$.
From this equation, $\psi(s) \leq \psi(u)$ always holds if $s \subseteq u$, which means that the Apriori strategy of pattern mining can be directly used to find all testable patterns.
Finally, after finding all testable patterns, compute *p*-values for them using Fisher's exact test and output patterns whose *p*-value is smaller than α/k.

5. Conclusion

This paper introduces the current state-of-the-art of significant pattern mining techniques. Implementations of almost all methods are available online and can be used in applications. Since significant pattern mining can provide *p*-values and control the FWER (or the FDR), further collaborations with other scientific fields are likely to contribute interesting discoveries.

* this chapter is to be cited as :

Sugiyama M. 2017. Finding Statistically Significant Patterns from Data. pp. 53 – 58 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

The "explore and exploit" strategy of bacteria: from the gut to the ocean

Dorota Czerucka [1] and Fernando Peruani [2]

1. Centre Scientifique de Monaco, MC 98000.
2. Laboratoire J.A. Dieudonné (UMR7351 CNRS UNSA) Université Nice Sophia Antipolis, Nice.

Abstract: Scattered empirical data suggest that motility, often considered a biosignature, plays a key role in bacterial survival as well as in the bacterial infection processes. For enteric pathogenic bacteria such as *Helicobacter pylori*, *Salmonella* and *E. coli*, motility allows the pathogens to cross the mucus barrier located in the gastrointestinal tract and move toward the epithelial cells. For marine bacteria, motility is crucial to find colonization niches and follow nutrient patches adverted by currents. Data-driven mathematical models of bacterial motility, based on *in vitro* experimental data, can be used to elucidate and quantify bacterial exploring strategies, estimate the survival advantage conferred by motility, and to quantitatively demonstrate the correlation between bacterial motility and infection capacity.

Key words: flagella, chemotaxis, motility, quantitative biology, mathematical modelling

Introduction

In the 17th century, Antonie van Leeuwenhoek reported for the first time, using single lens microscopes, about the motion of peculiar microscopic objects, bacteria and protozoa. It was the vivid motion of these objects that initially convinced him that these hectic tiny objects were alive. Since then, motility became a biosignature. However, the observation of motion *per se* is not necessarily a manifestation of life. A century after the pioneering work of van Leeuwenhoek, around 1820, Robert Brown observed through his microscope the motion of tiny pollen grains. His initial guess was that these tiny grains were also alive as those reported by van Leeuwenhoek. His inquisitive mind pushed him to test this hypothesis by placing in water and alcohol grains of inorganic matter. To his surprise, inorganic grains also displayed motion.

It took another century to understand that the motion of those grains were of thermal origin. The experiments of Jean Baptiste Perrin and the theory of Albert Einstein provided the evidence and arguments that clarified the intriguing observation made by Brown. But then, under which

circumstances (if any) does motion become a life-related phenomenon? Nowadays, we know that the thermal motion of micrometer-size objects such as a bacterium (i.e. its diffusion coefficient) is in the range of 0.1 $\mu m^2/s$ at room temperature. However, direct measurements of bacterial motion (in the bulk) reveal that bacteria exhibit diffusion coefficients of the order of 10^4 $\mu m^2/s$, i.e. more than 4 orders of magnitude larger than the expected thermal diffusion coefficient. This quantitative difference is the signature of an active, energy-consuming process - the rotation of the flagellar bundle - and a clear biosignature. Interesting, the majority of bacterial species are known to be motile during at least a part of their life cycle (Fenchel *et al.,* 2002). Active motion (in contrast to thermal motion) confers a survival advantage to bacteria by letting bacteria explore space at a much faster pace than just by thermal diffusion. Some bacteria are also able use chemical cues to guide their motion and move toward favourable microenvironments (or away from unfavourable ones). The ability to direct their active motion along chemical gradients is known as chemotaxis. For intestinal pathogenic bacteria, the combination of motility and chemotaxis enables bacteria to detect and pursue nutrients and to reach their preferred niches for colonization (Stecher *et al.,* 2004).

Bacterial motility in the gut

As enormous surface area of mucosal cells in the gastrointestinal tract is potentially exposed to enteric microorganisms, the mucosal epithelium has developed different strategies to provide both a barrier to the commensal microorganisms and protection against potential viral, bacterial and eukaryotic pathogens. The organism's defence system against infection can be viewed as consisting of several "levels". The first one is the stratified mucus layer, which together with the glycolalix of the epithelial cells provides physical protection. Mucus is produced by goblet cells that interpolate between epithelial cells. The thickness of mucus layer as well as its composition varies between the stomach and the rectum (reviewed in McGuckin, 2011).

The major function of mucus is to limit bacterial contact with the epithelium. Enteric pathogens have developed a range of strategies to subvert and avoid this barrier. Pathogens can penetrate the mucus barrier physically through enzymatic degradation of the mucus as for instance in *Vibrio cholera*, *Yersinia enterocolitica*, *Clostridium perfingens* and *Shigella flexeneri*, or through flagella-mediated motility that propels bacteria in viscous environment, e.g. in *Escherichia coli* and *Salmonella typhimurium* (Macnab *et al.,* 1987).

Mechanism of motility and guidance: flagella and chemotaxis

The flagellum, together with the flagellar motor, provides self-propulsion to bacterium. In peritrichously flagellated bacteria, flagella are distributed over the cell body (e.g *Escherichia coli* and *Salmonella* sp.), while other bacteria display flagella at one pole (e.g *Helicobacter pylori* or *V parahaemolyticus*). Flagella generate bacterial movement via rotation of the filaments and most flagellar motors are reversible rotary machines, able to rotate both clockwise (CW) and counterclockwise (CCW). In case of "polar flagellated bacteria" the CW and CCW rotations of the flagellum correspond respectively to forwards and backwards swimming modes. In case of "petrichous flagellated bacteria" CCW spinning of the motor generates forces, which cause the individual filaments to sweep around the cell and form a single flagellar bundle propelling the bacterium forward in a "smooth" swimming motion. When the motor spins CW the propulsive flagellar bundle flies apart and moves individually, thus propelling bacterium in a "tumbly swimming" motion. "Petrichously flagellated bacteria" display a swimming pattern in which the "smooth" and "tumble" modes, of short duration (1ms), are alternately repeated (Macnab *et al.,* 1987).

In addition to motility, flagella have roles in other microbial process such as adherence to host cells, cell invasion, protein secretion, autoagglutination, and induction of pro-inflammatory response in host cells (Anderson *et al.,* 2009).

Structural studies revealed that the bacterial flagellum is comprised of three basic parts: the filament (helical propeller), the hook (universal joint), and the basal structure (rotary motor) (reviewed in Terashima *et al.,* 2008, Anderson *et al.,* 2009). About 50 gene products are involved in the construction of a functional flagellum. Since the flagellum is such a big organelle and its production and assembly requires a large commitment of resources, bacteria have developed a precise regulation system that controls flagellar construction. In *Salmonella* this control mechanism is well characterized: in brief the flagellum assembly from the inner structure base to the outer ones, beginning with basal body construction followed by hook assembly and finally filament formation. This assembly-coupled flagellar gene regulation is achieved by a cascade of flagellar gene operon called flagellar regulon. In *Salmonella* there are three classes of operons: early, middle and late. The master relator for the flagellar regulon (FlhDC) belongs to the early operon that induces the expression of middle operon. The middle operon contains gene encoding for basal body and hook and the regulator 28 (FliA) that control gene belonging to the late operon (filament and motor). In *Vibrio* the gene regulation for polar flagellar synthesis is more complex that for *Salmonella*. An inverse relationship exists between motility and pathogenicity in *Vibrio* spp. However it is undeniable that motility induced by the polar flagellum contributes to the virulence of pathogenic *Vibrio* through adhesion or biofilm formation regardless of the environment (reviewed in Zhu *et al.,* 2013).

Chemotaxis is the ability to sense and direct movement in response to chemical gradients. By using transmembrane chemoreceptors to measure chemical concentrations in their immediate vicinity and signal transduction system to subsequently process this information, bacteria can sense chemical gradients and tune motility accordingly. Many pathogenic bacteria use chemotaxis to find suitable colonization sites: chemotaxis can guide *Helicobacter pylori* to the mucus of the human stomach, *Vibrio cholerae* toward the intestinal mucosa. For chemotaxis, environmental gradients of attractants (amino acids, sugars and oligopeptides) and repellents (extreme pH, some metal ions, hydrophobic amino acids) are perceived by methyl-accepting chemotaxis proteins (MCPs). In the excitation phase, conformational changes caused by ligand binding to MCPs are conveyed to the cytoplasmic face of the membrane where they are recognized by an associated “transmitter” complex (CheA-CheW) (reviewed in Lux *et al.;* 2004). Thus, motility is regulated by a very complex network.

Bacterial motility and the ocean

Oceans contain a multitude of bacteria - approximately 10^8 microorganisms per liter - and the majority of these bacteria in culture are motile. The fraction of motile bacteria is low in coastal seawater (around 10%) and varies from 5% to 70% in the water column. Motility levels appear to be subject to large natural variability: the fraction of mobile bacteria is larger in summer than in winter and larger by day than by night. In sea water, motility is often favored when a high level of dissolved nutrients becomes available, particularly from point sources such as organic particles or individual phytoplankton in an algal bloom, which upon releasing nutrients creates a chemical gradient, confirming that chemotaxis is one of the fundamental functions of motility.

Both biological and physical factors determine where and when motility will be favored and what the optimal swimming speed is. Biological factors that determine motility include cell size, flagellar structure and presence or absence of protozoan predators. Physical factors include the density, characteristic of nutrient patches, the presence of fluid flows and turbulence. Importantly, the

dynamics of swimming is dictated by low-Reynolds-number physics. It is thus not surprising that marine bacteria exhibit strong phenotypic differences from enteric models like *E. coli* including higher swimming speeds, drastically different motility patterns and higher level of chemotactic performance. Chemotaxis is widespread among diverse taxa in the ocean. Many of the marine bacteria that have been isolated from seawater including *Vibrios*, *Silicibacter*, *Rosebacter*, *Pseudoalteromonas*, *Pseudomonas* etc, exhibit chemotaxis toward a wide range of attractants including amino acids, sugars, carboxylic acid, organic sulfur compounds, oxygen, nitrate, nitrite ammonium urea and phosphate and mucus secreted for example by corals. In terms of pathogenicity some Vibrio species such as *V cholera*, *V Vulnificus* and *V parahaemolyticus* have been described as human pathogens and are also pathogenic to fishes and other animals.

Relation between *Salmonella* motility and pathogenicity

In our laboratory we currently study the interaction between enteropathogenic bacteria and their host. We developed an *in vitro* model of human adenocarcinoma T84 cells that form polarized epithelium. This cellular model was used to follow the different steps of infection by *Eschericia coli* and *Salmonella typhimurium*: adhesion of bacteria to host cells wall and invasion, increase of monolayer permeability and finally pro-inflammatory response of the host cells *i.e* activation of NF-κB nuclear translocation, activation of MAP kinase resulting in IL-8 secretion (Czerucka *et al.,* 2001, Dahan *et al.,* 2002, Martins *et al.,* 2010).

Salmonella motility is crucial for the infectious process as strains cultivated with shaking (condition in which flagella are destroyed) display the pathogenic response in host cells. To investigate the relationship between motility and infection we performed video records of T84 cells exposed to ST (Pontier-Bres *et al.,* 2012). Real time computer tracking was assessed to follow bacterial motility. Using MTrackJ plugin we performed mathematical reconstitution of trajectories in x, y plane allowing us to determine the bacterial velocity (CLV) and linearity (LT). Data reported in this study show that in ST-infected cells, bacteria moved with a median CLV of 43.2 µm/sec ranging from 102.0 µm/sec to 1.2 µm/sec. Implication of flagella in ST motility was confirmed by the use of the mutated strain M913. These bacteria that were deleted in flagella (Stecher *et al.,* 2004) moved with a median CLV of 0.57 µm/sec that is significantly lower than the CLV of the wild type strain. Given that cell invasion is a main step in *Salmonella* pathogenesis, we have investigated invasion in conditions of human colonic T84 cells infected by the wild type strain SL1344 and the non-motile mutated strain 913. We showed that invasion of T84 cells by the M913 strain is significantly less efficient (around 100 times) than invasion by the wild type strain (Figure 1).

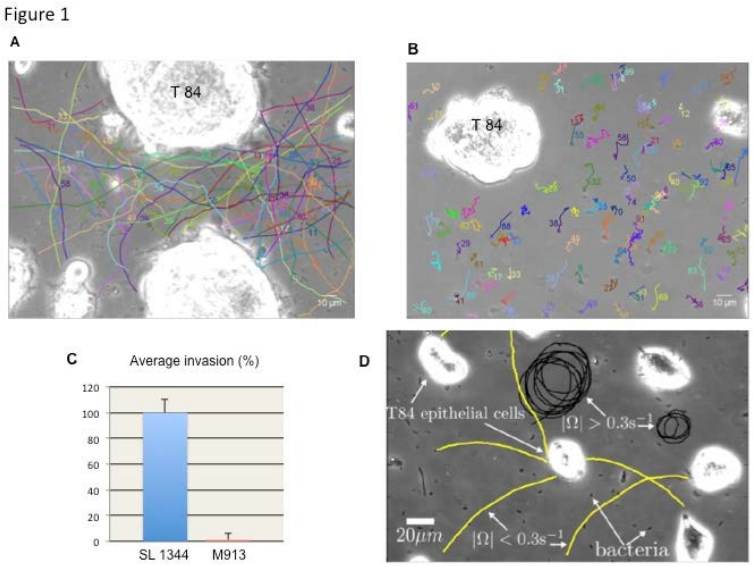


Figure 1 shows swimming trajectories of the wide type strain SL1344 (A), the mutated strain (Fla-) M913 (B) incubated with T84 cells. Records were performed 30 min post infection (PI), and the time between consecutive images was 0.1 second. Using MTrackJ software we determined the locations for each bacterium, and this information was then translated into coordinates (x, y) for each bacterial cell and the process was repeated in times series. The 2D trajectories of each bacterial cell were represented; different colours represent different trajectories. Each trajectory has it own number. Panel C presents the invasion of T84 cells infected one hour with the wild type strain SL1344 or with the mutant M913. Invasion was assessed by the gentamicin protection method. % of invasion was normalized versus SL1344 as 100%. Panel D displays bacterial trajectories of the same bacterial strain (SL1344), which illustrate the large variability of motility patterns: while same bacteria move in roughly straight trajectories, others perform smooth circular trajectories. Differences between these trajectories can be quantified by a series of motility parameter, and in particular the parameter Omega let us measure the curvature of these trajectories. For bacteria exhibiting large values of Omega, the time to find a host cell is 10 times longer than those with small values of Omega.

Quantitative approach of bacterial exploration strategies

Swimming, multiple-flagellated bacteria such as *Escherichia coli* and *Salmonella typhimurium* when swimming far away from surfaces exhibit swimming patterns characterized, as explained above, by the presence of roughly straight runs followed by sudden changes of direction (tumbles). This kind of motility patterns is known by the name of “run-and-tumble”. The characterization of the three-dimensional motion of bacteria requires the use of a special microscope equipped with a fast responding, automatized stage that allows keeping focus a given bacterium as it swims through the medium (Berg, 1972). However, most bacterial studies are performed with “regular” microscopes, which only allow observing and tracking the motion of individual bacteria for long periods of time - from tens of seconds to a minute - by focusing on the bottom surface of the chamber containing the bacteria. However, if the microscope is focused close to surface, we do not longer observe the above-mentioned run-and-tumble motion. Near surfaces, the swimming trajectories become smoother and tend to be circular (Frymier *et al.,* 1995). Moreover, the swimming of bacteria is strongly dominated by hydrodynamic interactions that lead to: i) an effective attraction towards the surface, ii) the above-mentioned circular motion of bacteria, and iii) a strong suppression of tumbling events (Otte *et al.*, 2017).

Knowing the statistical features of near-surface bacterial motion is key to characterize bacterial motility properties using regular microscope (i.e. without fast moving automatized stage) as well as to understand how bacteria explore surfaces. Note that most colonization niches are located on surfaces where bacteria anchor and biofilms grow. In order to obtain a quantitative characterization of near-surface swimming, we first tracked the motion of individual bacteria close to surfaces and performed a systematic study of the statistical features of the obtained swimming trajectories. Secondly, we showed that the experimental trajectories are consistent with a mathematical model of chiral active particles with active fluctuations with four parameters: the average speed v, the fluctuations of the

speed Dv, the angular velocity Ω - directly related to the curvature of the trajectories - and D_θ , which provides a measure of the fluctuation in the moving direction.

We developed a statistical method to extract each of these parameters from the experimental trajectories of individual bacteria. Our analysis (Otte *et al.*, 2017) revealed that our bacteria, besides of being alive and very motile, are strongly influenced by the thermal fluctuations first observed by Brown and typically neglected in bacterial motility analysis. We found that D_θ is of thermal origin, while Dv is of biological one. We used the mathematical model to express the diffusion coefficient (i.e. the exploring capacity) of bacteria swimming near-surfaces as function of these four parameters. We found that though each of these parameters varies in a small range from individual to individual, the diffusion coefficient of individual bacteria ranges over four orders of magnitude, i.e. in a range comparable to the one we discussed at the beginning between micrometer-size inanimate objects and swimming bacteria (in the bulk). We suspect that such giant inter-individual variability of the diffusion coefficient may provide a bacterial population with an important adaptation capacity to find colonization niches in different external conditions.

Conclusions

The bacterial motility machinery comprises two components: self-propulsion, involving the rotation of flagella, and the chemotactic system, which in turn interacts with the flagellar motor. These two components are subjected to a complex gene network with precise regulation that reflects the importance of motility for bacteria. Inspite of this, up to now there was not quantitative mechanistic explanation to relate bacterial motility with the infection capacity of bacteria and no quantitative estimate of the relevance of bacterial motility for bacterial survival. The difficulty was rooted to the fact that some bacterial species, pathogenic and not, are not motile. This observation suggests that the importance of motility is species specific, and arguably related to the natural environmental conditions of the bacteria. By deriving data-driven mathematical models for bacterial motility, based on *in vitro* experimental data, it is possible to elucidate and quantify the bacterial exploring strategies, which are species and environment specific. This is achieved by obtaining a set of specific motility parameters that provide an accurate quantitative description of the observed bacterial motility patterns. This allows estimating quantitatively the role played by motility in the infection process and bacterial survival, demonstrating the existence of a correlation between bacterial motility and infection capacity. The proposed approach, that combines *in vitro* experiments and mathematical modelling, can be applied to characterize the exploring capacity and role of motility in bacterial survival of marine bacteria. This will require acquisition of data at different geographic position as well as a detailed characterisation of the environmental conditions (temperature, pH, nutrient quality and spatial distribution, etc.). Such an approach may pave the way to estimate human impact on the environment and pathogenicity of marine bacteria.

* this chapter is to be cited as :

Czerucka D. and Peruani F. 2017. The “explore and exploit” strategy of bacteria: from the gut to the ocean. pp. 59 – 64 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

Tracking pathogens in the wild requires a plan

Maxime Bruto[1], Adèle James[1, 2], Damien Piel[1, 2], Sabine Chenivesse[1, 2], Yannick Labreuche[1, 2] and Frédérique Le Roux[1, 2]

[1]Sorbonne Universités, UPMC Paris 06, CNRS, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, F-29688, Roscoff Cedex, France.
[2]Ifremer, Unité Physiologie Fonctionnelle des Organismes Marins, ZI de la Pointe du Diable, CS 10070, F-29280 Plouzané, France.

Abstract

Climate change is correlated with a worldwide increase in reports of vibrio-associated diseases having ecosystem-wide impacts on humans and marine animals. In addition, the rapid growth of aquaculture has been the source of anthropogenic changes on a massive scale. Animals have been displaced from their natural environments, farmed at high densities and exposed to environmental stresses, including antibiotic treatment. A vast number of papers describes animal mortality outbreaks associated to vibrios but a large proportion of these works are based on i) moribund animals that may be infected secondary by vibrio (opportunistic colonization); ii) *Vibrio* strains identified taxonomically using rRNA genes that collapse the majority of genotype into only 2-3 taxa; iii) simplistic models of infection using only one or a few isolates. Hence it is often not clear whether vibrios isolated from diseased animals are the causative agent of the disease. Tracking pathogens in the wild requires a plan, *i.e.* unbiased sampling strategy, sufficient genomic coverage for microbial population structure, ecological realistic experimental infection model. This ultimately allows mapping of pathogenicity onto population structure to identify the functional unit of pathogenesis, a prerequisite to diagnose and monitor an infectious disease. Here we will describe why the interaction between oysters and vibrio is an up and coming model system to address new and original scientific questions concerning the dynamics of infectious diseases in the wild and its feedback on ecology and evolution of host-pathogen interactions.

Key words: ecology, evolution, vibrio, epidemiology

***Vibrio* is one of the best-described marine bacterial groups in evolutionary ecology**. Vibrios can be easily isolated and cultured, allowing multilocus or whole genome sequencing to obtain fine-scale genetic resolution among individual isolates (Fig.1A). In order to reduce the cost and time of the procedure as well as increasing the number and coverage of samples, developments of meta-barcoding sequencing using vibrio core genes are in progress (Cheng D. PhD thesis, MIT 2015). A series of studies (Shapiro and Polz, 2015) has shown that despite a high genetic diversity these bacteria are divided into phylogenetic groups sharing a lifestyle (planktonic or associated) and preferences for habitat (organic particles, phyto or zooplankton) (Hunt *et al.*, 2008) (Fig. 1A, B). Gene flux appears more frequent within these groups than between groups (Shapiro *et al.*, 2012) and the production of "public goods" governs social cohesion between strains (Cordero *et al.*, 2012a; 2012b). Thus, these phylogenetic groups satisfy the concept of population in ecology (and to some extend species unit) and provide a framework to investigate the functional unit of pathogenesis, *i.e.* a clone that emerges after a recent acquisition of virulence genes (Goudenege *et al.*, 2013), a population with virulence encoded by the core genome (Lemire *et al.*, 2014) or a consortium (Le Roux *et al.*, 2016).

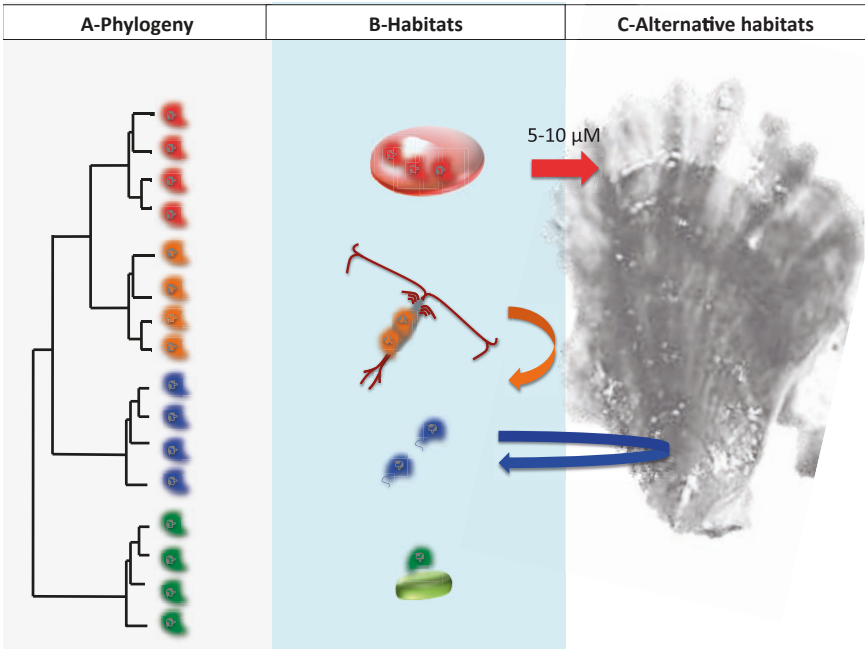


Figure 1: Ecological population of vibrios (adapted from Le Roux *et al.*, 2016). Despite extensive genetic diversity, vibrios cluster in phylogenetic clades (A) that show preference for habitat (B: blue: free living; red and green: particles associated; orange: host associated). To successfully colonize an alternative host such as oyster (C), cells must be taken up through the gills that act as a sieve (optimum size 5-10 µm) and be able to evade the immune system and out compete the host microbiota.

Investigating "vibrio virulence into the wild" requires an animal model of infection. Of the recent work aiming to improve the *in vivo* model, standardization of animal hatching seems to hold promising perspectives. In particular, we currently use specific pathogen-free (SPF) juvenile oysters (*Crassostrea gigas*) for experimental ecology and infection (Petton *et al.*, 2015). SPF oysters are descendants of a pool of genitors that are produced in hatcheries under highly controlled conditions so as to minimize the influence of genetic and environmental parameters that could affect the host sensitivity to the disease (Petton *et al.*, 2013a; 2013b). Among the infectious agents that have been associated to oyster diseases to date, the oyster herpes virus (OhSV) (Barbosa Solomieu *et al.*, 2015; Martenot *et al.*, 2011) is undetected in SPF oyster tissues. These animals are not axenic, as is the case with gnotobiotic animals developed to explore infectious processes without interference or influence

from unknown microbiota (Marques *et al.,* 2006); however the load of vibrios is low (<1 colony forming unit per mg tissue) and no mortalities are observed when maintaining these animals in the laboratory. SPF oysters can be placed in the environment where the onset of mortality and cumulative mortality rates can be monitored under natural conditions. Once disease is detected, live animals can be used in laboratory experiments to investigate the diversity and dynamics of microbes during disease progression. Moreover, these SPF oysters can be used for high-throughput experimental infections with hundreds of bacterial isolates (Goudenege *et al.,* 2015; Lemire *et al.,* 2014; Petton *et al.,* 2015). Hence standardized SPF oysters constitute an animal model to i) sample naturally colonizing vibrios from the environment, ii) allow natural progression of infection, and iii) determine virulence mechanisms across populations or strains of infecting vibrios.

On the oyster side, research can also benefit from the development of a wealth of genomic resources on *C. gigas* including linkage maps, transcriptomes and a whole genome sequence (Fabioux *et al.,* 2009; Hedgecock *et al.,* 2015; Sussarellu *et al.,* 2015; Zhang *et al.,* 2012) and microbiome metabarcode sequencing (Lokmer *et al.,* 2016; Lokmer and Wegner, 2015). Disease resistance encompasses a substantial genetic component that is responsive to selection (Degremont *et al.,* 2015; Wendling *et al.,* 2014; Wendling and Wegner, 2015) and the molecular diversity of oyster immune effectors is well-characterized (Bachere *et al.,* 2004; Schmitt *et al.,* 2010a; Schmitt *et al.,* 2010b). Such potential target genes can be manipulated by RNA interference (Fabioux *et al.,* 2009) thus offering population genomic resources to functionally characterize host-pathogen interactions.

Genetic strategies have been developed to delete genes, genomic regions or cure plasmids in numerous *Vibrio* species. Gene knock-out is essential for the formal demonstration of the predicted, or supposed, role of a gene/loci. However, this strategy is still limited to the strains in which available genetic tools can be used. Limitations can occur at several levels from the DNA delivery inside the cells, to the allelic exchange efficiency. DNA transformation (Gulig *et al.,* 2009; Marvig and Blokesch, 2010; Pollack-Berti *et al.,* 2010), be it natural or artificial, is either inoperative or inefficient in numerous vibrios and in many cases, exogenous DNA delivery relies on conjugation using *E. coli* donor strain (Simon *et al.,* 1983). The subsequent step of integration of the incoming DNA in the genome is in most cases achieved through the use of a non-replicative DNA molecule such as conditionally replicative R6K plasmid derivatives (Miller and Mekalanos, 1988). R6K replication is dependent on the binding of the *pir*-encoded (Kolter *et al.,* 1978). Plasmids carrying the R6K origin of replication ($oriV_{R6K}$) can only be replicated in E. coli strains expressing *pir* and behave as suicide vectors in *pir* $^{-}$ vibrio recipients. These plasmids have been successfully used to create mutants through gene disruption by insertion (Le Roux *et al.,* 2009) or transposon mutagenesis (Takemura *et al.,* 2017). Mutagenesis by allelic exchange (also named “Pop in, Pop out”) requires the use of a suicide vector carrying markers for counter selection of allelic exchange such as sacB (Ding et al., 2004) or ccdB (Le Roux *et al.,* 2007). These types of constructs have also been used to cure endogenous plasmid in vibrio (Bruto *et al.,* 2017). Alternatively, plasmid curing can also be performed by transferring a suicide vector carrying the origin of replication of the endogenous plasmid (mini-*oriV*) into the recipient, leading to incompatibility between the endogenous replicons and the mini-*oriV* (Le Roux *et al.,* 2011b). Finally, derivatives of plasmids originally isolated from *Vibrio* have been shown to be stably maintained by this host and are being used to express genes in trans (Le Roux *et al.,* 2011a). Finally although genetic strategies have been successfully applied to numerous *Vibrio* isolates, conjugation and mutagenesis efficiency can differ dramatically between closely related strains (Goudenege *et al.,* 2015). We thus strongly recommend testing the feasibility of genetic approaches prior to genome sequencing.

Analysis of natural infection dynamics, population genomics and molecular genetics has already provided important insights on oyster disease. First, virulence can coincide with population delineation (Bruto *et al.;* Le Roux *et al.,* 2011a). Using SPF oysters we showed that the proportion of strains capable of eliciting disease differs among populations of vibrios. Genome comparison within and between populations uncovers population specific genomic regions and some of them have been further demonstrated to contribute to virulence (Bruto *et al.,* 2017; Lemire *et al.,* 2014). Second, assessing the distribution of vibrios in animals supports the view that environmental dynamics is an important factor in colonization (Bruto *et al.,* 2017; Preheim *et al.,* 2011). Oyster-associated vibrios were analyzed in the context of a metapopulation framework, *i.e*., by considering potential overlap or differences in populations collected from spatially and temporally distinct habitats, which are connected by dispersal (Bruto *et al.,* 2017). This analysis revealed that several populations of *Vibrio* are preferentially associated with specific oyster tissues (Fig.1C). Among these, a population taxonomically assigned to *V. crassostreae* (Faury *et al.,* 2004) was found to be abundant in diseased animals, and its pathogenicity was correlated with the presence of a large mobilizable plasmid further demonstrated to be essential for killing but not colonizing oysters. Third, population diversity may increase the severity of pathogenesis. For example, *V. tasmaniensis* and *V. crassostreae* have been associated to diseased oysters and found to co-occur at the individual level (Bruto *et al., 2017*). Infection with *V. tasmaniensis* involves an intracellular phase in hemocytes and resistance to antimicrobial peptides, reactive oxygen species and copper (Duperthuy *et al.,* 2011; Vanhove *et al.,* 2015). Infection with *V. crassostreae* relies at least partially on distinct genes encoding for unknown functions (Bruto *et al.,* 2017; Lemire *et al.,* 2014). Hence oysters can be infected by species with different and potentially additive virulence mechanisms. Consistent with the hypothesis of a "shared weapons", experimental infections have demonstrated that some strains are moderately virulent when injected into animals individually and display heightened virulence in mixed experimental infections (Gay *et al.,* 2004; Lemire *et al.,* 2014).

The discovery that population can be the unit of *Vibrio* pathogenesis in oyster opens important perspectives to understand the evolution of these pathogens. First, the dynamics of vibrio species in the seawater column may have important consequences for when and where infection may be more likely to occur, i.e. population increases triggered by specific ecological conditions or decreases following the action of predators such as phages or grazers. Second, the observed population structure will yield insight into the evolutionary history of the pathogens and types of selection acting on pathogenicity determinant genes. Finally, oyster has been described as a reservoir for specific species of vibrio (Bruto *et al.;* Petton *et al.,* 2015), suggesting adaptations to living associated with oyster hosts (Wendling and Wegner, 2015). Holistic approaches taking into account the complexity of all levels of these interactions are needed in the future to deepen our understanding of disease processes and provide tools for an efficient management of disease in the wild.

Acknowledgements

This work has been supported by the ANR (project OPOPOP (13-ADAP-0007-01) and REVENGE (16-CE32-0008-01), region Bretagne and Ifremer (AJ and DP grants).

* this chapter is to be cited as :

Le Roux F., Bruto M., James A., Pied D., Chenivesse S. and Labreuche Y. 2017. Tracking pathogens in the wild requires a plan. pp. 65 – 68 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

Environmental stress and the control of gene expression in pathogenic bacteria

Charles J. Dorman

Department of Microbiology, Moyne Institute of Preventive Medicine, Trinity College Dublin, Ireland

Abstract

Bacterial pathogens of humans, animals and plants control the expression of their virulence genes using mechanisms that are environmentally responsive. This increases the probability that the genes will be expressed only when the bacterium experiences a combination of environmental signals that are characteristic of the host. The individual signals can be physical or chemical in nature and obtained from any environment, including a marine one. There are striking similarities between the control mechanisms used by pathogens that have an aquatic stage in their life cycles and those that are principally terrestrial. This allows us to make informed assessments about the regulatory strategies and mechanisms that are likely to be employed by bacterial pathogens in an oceanic setting.

Key words: DNA topology**,** Gene regulation**,** *Escherichia coli***,** *Salmonella enterica* serovar Typhimurium**,** *Vibrio cholera***,** Pathogenicity islands**,** DNA topoisomerases

Introduction

Marine bacteria face environmental challenges that are also encountered by human pathogens. These challenges include the need to adapt to an environment where nutrition is unpredictable and where competition for resources and habitable niches can be intense. The microbes must also survive physical changes to their surroundings, such as variations in pH and temperature, and chemical changes such as those leading to osmotic shock. Some marine bacteria, e.g. *Vibrio cholerae*, are human pathogens, making the link between the marine microbiome and human medicine an intimate one (Bruto *et al.*, 2017; Czerucka and Peruani, this volume).

Knowledge of virulence traits gained by studying model organisms that are relevant to human and veterinary medicine, as well as micro-organisms that infect plants, is of direct relevance to the advancement of our understanding of pathogens in aquatic environments. Lessons from bacteria such as *Salmonella enterica, Shigella flexneri* and *Escherichia coli* have been particularly valuable because they have illustrated the relationships between virulence genes and the core genome, the role of

horizontal gene transfer in the evolution of pathogens and the importance of environmental signals in controlling the expression of virulence traits. This information informs our views of what a pathogen might look like and of how it might behave. In this article I will focus on mechanisms of gene expression control in *Salmonella*, with a particular emphasis on the importance of DNA topology as a global influence on gene expression both in the core genome and in horizontally-acquired pathogenicity islands.

DNA supercoiling and gene expression

Our laboratory has investigated over many years the role of variable DNA topology, especially DNA supercoiling, as a regulatory principle in the control of bacterial gene expression (Dorman and Dorman, 2016; Travers and Muskhelishvili, 2005). Most bacteria of medical importance maintain their genomic DNA in a negatively supercoiled conformation, but the degree of supercoiling changes in response to environmental inputs (Cameron and Dorman, 2012; Dorman *et al.*, 2016). In the marine environment, extremophiles associated with deep ocean vents experience very high temperatures and many of these organisms maintain their DNA in a positively supercoiled state that tightens the DNA double helix (Lulchev and Klostermeier, 2014). In contrast, mesophilic bacteria such as *V. cholerae* have negatively supercoiled genomic DNA (Parsot and Mekalanos, 1992). Negative supercoils are introduced into DNA at a local level by the processes of transcription and DNA replication: the tracking of the polymerases causes local underwinding of the DNA in their wake and overwinding ahead. These under- and over-wound regions are equivalent topologically to negatively and positively supercoiled DNA, respectively (Liu and Wang. 1987). Specialist enzymes called topoisomerases eliminate these topological distortions, a process that is essential if the polymerases are to continue with their work (Ma *et al.*, 2013). The same topoisomerases operate on a genome-wide scale, maintaining the level of negative supercoiling in the genomic DNA within limits that are appropriate for the survival of the bacterium (Margolin *et al*., 1985; Sternglanz *et al*., 1981).

DNA topoisomerases

DNA gyrase is a topoisomerase that is unique to bacteria and it is responsible for removing positive supercoils, doing so through a mechanism that is identical to the one by which it introduces negative supercoils. It hydrolyses ATP to carry out the reaction, making the activity of gyrase sensitive to the ratio of ATP to ADP in the cell (Bates and Maxwell, 2007; Snoep *et al*., 2002). This creates an important link between DNA topology and the physiology of the bacterium: high metabolic fluxes are likely to promote higher negative supercoiling activity by gyrase, allowing it to neutralise positive supercoils and to convert relaxed DNA to a negatively supercoiled configuration. Topo I, an ATP-independent topoisomerase, uses the pent-up tension in negatively supercoiled DNA to drive its relaxation (Champoux, 2001). DNA relaxing activities are also found in topo III (ATP-independent) and topo IV (ATP-dependent) (Champoux, 2001).

Acid stress and DNA topology in *Salmonella*

We have studied the influence of exposure to acid pH on the superhelical state of the genomic DNA in the bacterium *Salmonella enterica* serovar Typhimurium (Cameron *et al*., 2011; Cameron and Dorman, 2012; Quinn *et al.*, 2014). This Gram-negative organism is a facultative intracellular pathogen that causes infection following ingestion of contaminated food or water (Boyle *et al*., 2007). It uses a set of specialist genes to promote invasion of the epithelial cells lining the small intestine and a second set to survive in the vacuole of macrophage. The macrophage attempts to kill the invader by acidifying the vacuole but the microbe exploits the drop in pH as a signal that governs the expression

of genes involved in acid stress survival and resistance to other killing measures initiated by the macrophage (see Fig. 1) (Chakraborty *et al*., 2015; Fass and Groisman, 2009; Yu *et al*., 2010).

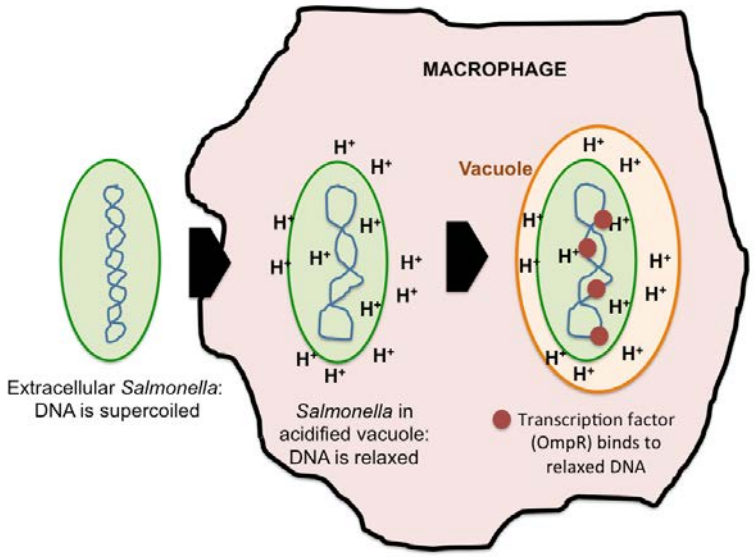


Figure 1. *Salmonella* adaptation to the macrophage acidified vacuole. In the extracellular state, the genomic DNA (blue) in the bacterium (green) is negatively supercoiled. In the intravacuolar environment (orange), *Salmonella* reacts to acidification of the niche by relaxing its DNA (actually due to acid-inhibition of the negative supercoiling activity of DNA gyrase). This produces a DNA template with an optimal topology for binding by the transcription factor, OmpR. This protein activates the expression of genes that promote the survival of *Salmonella* in the vacuole. The small maroon-coloured discs shown decorating the DNA in the intracellular *Salmonella* cell on the right represent OmpR.

The specialist virulence genes are grouped in the pathogenicity islands SPI1 and SPI2. Each encodes a distinct type III secretion system and a dedicated collection of effector proteins (Hensel, 2000; Rhen and Dorman, 2005; Shea *et al.,* 1996). Virulence gene activation occurs in response to thermal, osmotic and acid signals, as well as a decline in the concentration of magnesium ions (Fass and Groisman, 2009). A key problem to be solved concerns the relief of transcription repression that is imposed on the genes in SPI1 and SPI2 by the H-NS nucleoid-associated protein (Dorman, 2007). H-NS is a DNA binding protein with an ability to block transcription by polymerising along A+T-rich DNA or by bridging together segments of A+T-rich DNA in nucleoprotein complexes that silence transcription (van der Valk *et al.,* 2017). A multitude of mechanisms is used to reverse the silencing and many of these rely on other DNA binding proteins such as the NAP FIS and conventional transcription factors (Stoebel *et al.*, 2008). An important example of a transcription factor in the SPI1 and SPI2 regulatory circuit is OmpR as will be seen below.

We find that DNA in *Salmonella typhimurium* becomes progressively relaxed during the time spent in the macrophage vacuole (O Cróinín *et al.,* 2006). The same relaxing effect can be mimicked in the laboratory by treating the bacteria with the gyrase-inhibiting drug novobiocin, an agent that blocks the ATP-binding site in the GyrB subunit of the topoisomerase (Hardy and Cozzarelli, 2003; O Cróinín *et al.,* 2006). It was reasoned that the novobiocin effect might replicate that of an unfavourable ATP/ADP ratio in the bacterium as it adapts to the stressful conditions in the vacuole. However, subsequent analysis has shown that a previously reported decline in the cytoplasmic pH of *S.* Typhimurium when in the macrophage vacuole (Chakraborty *et al*., 2015; Choi and Groisman, 2016) was inhibiting the ATP-dependent DNA supercoiling activity of gyrase (Colgan A and Dorman CJ,

unpublished data). Experiments with purified gyrase *in vitro* showed a clear inhibitory effect of acid pH on its negative supercoiling activity (Colgan A. and Dorman C.J., unpublished data).

Gene regulation in the SPI2 island of *Salmonella*

What is the link between DNA relaxation and the control of gene expression that assists bacterial survival in the macrophage vacuole? A key step in the survival strategy concerns the expression of proteins involved in the assembly and operation of a type III secretion system that exports effector proteins to modify the vacuole in ways that benefit *Salmonella* (Hensel, 2000; Shea *et al.*, 1996). Many of these genes are clustered in the pathogenicity island SPI2, a region of A+T-rich DNA that has been acquired by horizontal gene transfer from an unknown source (Fig. 2) (Shea *et al.*, 1996). Transcription of the genes and operons in SPI2 is under complex control (Fass and Groisman, 2009). Part of the regulation is imposed by the SsrB DNA binding protein whose activity is controlled by the SpiR sensor-kinase in response to signals relevant to vacuole adaptation (Fig. 2) (Fass and Groisman, 2009). The *spiR* and *ssrB* genes are part of the SPI2 cluster and have been acquired by horizontal transfer (Fig. 3) (Hensel, 2000). The SsrB protein activates expression of its own gene and that of SpiR (Walthers *et al.*, 2011). In addition, the DNA binding protein OmpR is an important regulator of *ssrB* and *spiR* transcription (Fig. 2) (Feng *et al.*, 2003; 2004; Lee *et al.*, 2000; Quinn *et al.*, 2014). OmpR is encoded by the *ompRenvZ* (*ompB*) operon and its activity is controlled by the EnvZ sensor-kinase in response to acid stress (Bang *et al.*, 2002). In the related bacterium *Escherichia coli*, EnvZ transmits osmotic stress signals as well as pH ones to OmpR (Forst *et al.*, 1989; Hall and Silhavy, 1981; Stincone *et al.*, 2011).

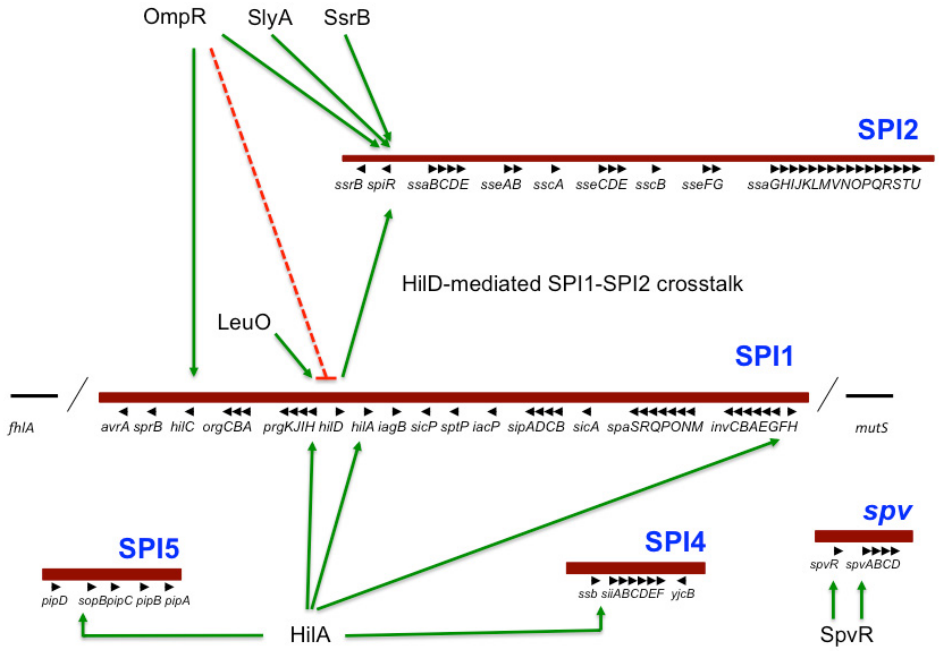


Figure 2. OmpR and the control of virulence gene transcription in *Salmonella.* Five major virulence loci are illustrated, four on the chromosome and one (*spv*) on a large plasmid. The OmpR wHTH DNA-binding protein regulates positively the master control genes in SPI2 (*ssrB spiR*) and the *hilC* regulatory gene in SPI1. It also represses the *hilD* regulatory gene in SPI1 that cross-regulates SPI2. Like OmpR, HilA, LeuO, SlyA and SsrB are all wHTH DNA-binding proteins that target A+T-rich DNA. The horizontal arrowheads at the virulence loci represent the individual open reading frames of the genes. Green arrows represent positive regulatory inputs; the dotted red line indicates a negative input by OmpR at *hilD*, which encodes an AraC-like DNA binding protein. The diagram is not to scale.

Integrated circuits: OmpR and SsrB control SPI2 transcription

Dual control by SsrB and OmpR is an example of the embedding of an imported set of genes into a pre-existing gene regulatory circuit in a way that exploits the regulatory inputs both of the imported regulator (SsrB) and of the one expressed by the core genome (OmpR). If OmpR evolved to control core genome transcription, how was it recruited by SPI2? Part of the solution seems to involve the very low DNA sequence requirements exhibited by OmpR in its binding sites. Essentially, this protein prefers to bind to A+T-rich DNA and that is a prominent feature of the SPI2 island.

Indirect readout: the importance of DNA shape in protein binding

Like SsrB, OmpR uses a winged helix-turn-helix (wHTH) motif to interact with both the major and the minor grooves of the DNA simultaneously (Martinez-Hackert and Stock, 1997). In A+T-rich DNA, the minor groove is narrower than in DNA found in the *Salmonella* core genome, possibly imposing a barrier to successful binding by OmpR (Rohs *et al.,* 2009). Modulation of the groove width occurs when the DNA twist is increased or decreased, providing a mechanism to influence the efficiency of OmpR binding (Dorman and Dorman, 2017).

DNA binding experiments performed *in vitro* with purified OmpR protein and circular DNA molecules containing high-affinity sites for OmpR binding show that this protein prefers DNA templates that are relaxed rather than negatively supercoiled (Cameron and Dorman, 2012). Whole genome analysis of OmpR binding to the *Salmonella* chromosome in living bacteria shows that the protein binds more avidly to its DNA targets when the DNA is relaxed compared to negatively supercoiled controls (Quinn *et al.,* 2014). Taken together, these results establish a role for variable DNA conformation in directing the binding pattern of the OmpR global regulator throughout the bacterial genome (Fig. 1). The negative impact of acid stress on DNA gyrase activity, leading to a loss of negative supercoiling in DNA, demonstrates a link between low pH stress (as experienced in the acidified vacuole of the macrophage), DNA relaxation, enhanced OmpR binding to DNA and the activation of transcription of the genes required by *Salmonella* to survive in the acidified vacuole (Fig.1).

The evolvability of simple regulatory circuits

One of the key features of the OmpR-DNA binding story is its relative simplicity and adaptability. By relying on an indirect readout mechanism that emphasises DNA shape rather than DNA sequence, wHTH proteins can be recruited to control the expression of newly-acquired genes if these genes can survive surveillance systems used by the bacterium to eliminate non-self DNA: restriction enzymes and CRISPR/*cas* (Fig. 3) (Dorman and Dorman, 2017). In Gram-negative bacteria the nucleoid-associated protein H-NS also plays a role in establishing genes that are acquired by horizontal transfer. This protein binds preferentially to A+T-rich DNA where it silences transcription through the creation of a chromatin-like nucleoprotein complex (Dillon *et al.,* 2010; Lucchini *et al.,* 2006). One of the tasks performed by OmpR involves acting as an anti-repressor to relieve the transcriptional silencing that is imposed by H-NS (Bang *et al.,* 2002).

Horizontal gene transfer is a key driver of bacterial evolution in all environments and is likely to be particularly potent in the marine setting due to the high concentrations there of bacteriophage (Fig. 3) (Hayes *et al.,* 2017). These bacterium-infecting viruses play a role in lateral gene transfer through transduction, the process in which bacterial genomic DNA is packaged in bacteriophage and then transmitted to their next bacterial hosts (Dorman, 2009). The scenario described above in which

foreign genes are silenced by H-NS and then activated when required by wHTH proteins in response to environmental signals accompanied by modifications to genomic DNA topology is not restricted to *Salmonella* and its close relatives. The master regulator of virulence gene expression in *Vibrio cholerae* is the wHTH DNA binding ToxR and this also controls the expression of A+T-rich genes that are subject to transcriptional silencing by the *V. cholerae* counterpart of H-NS, VicH (Cerdan *et al*., 2003; Kazi *et al*., 2016). The principal virulence factor used by this bacterium to cause cholera is the cholera toxin and its genes, *ctxAB*, are part of a bacteriophage called ctxφ (Waldor and Mekalanos, 1996). Individual *V. cholerae* cells can lose the ctxφ prophage from their genome, disarming them due to concomitant loss of the toxin genes. However, the *ctxAB* operon can be re-acquired if a new copy of the bacteriophage infects the cell (Waldor and Mekalanos, 1996). The virulence genes of *V. cholerae* also display sensitivity to changing DNA topology (Parsot and Mekalanos, 1992), giving them all of the components of the simple, environmentally responsive gene regulatory system described above for virulence genes in *Salmonella*.

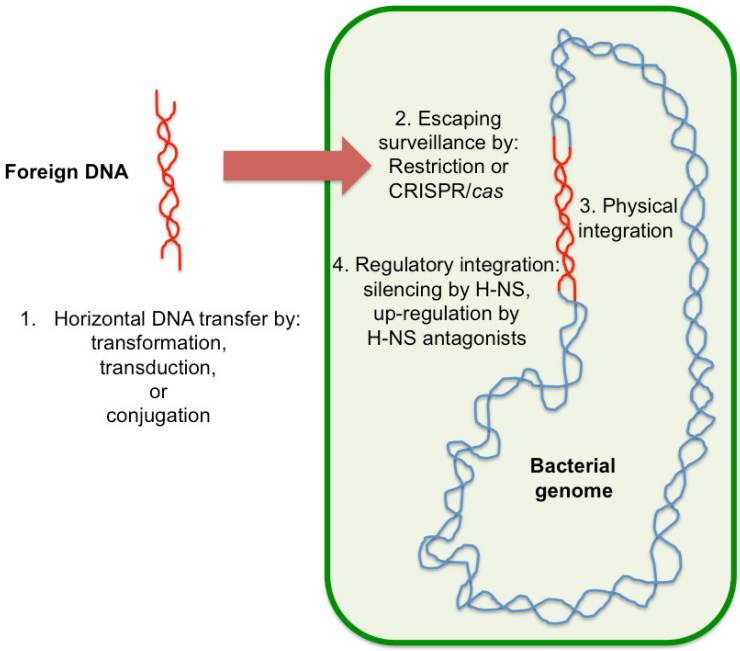


Figure 3. Acquisition of novel genetic characters (including virulence genes) by horizontal DNA transfer. Foreign DNA enters a new bacterial host (green) from the environment, with its transfer being mediated by transformation (naked DNA), transduction (bacteriophage-mediated) or conjugation (the new DNA is part of a self-transmissible plasmid). The foreign DNA (red) must escape destruction by cell defenses such as restriction endonucleases or CRISPR/*cas* systems and become integrated physically in the genome (blue). If the new DNA has an appropriate profile (A+T-rich and with intrinsic DNA curvature) its genes will be transcriptionally silenced by H-NS. Promiscuous DNA binding proteins, already present in the cell, may overcome this silencing, allowing the novel genes to be expressed in response to environmental signals relayed by those proteins.

DNA topology and global control of transcription

Variations in DNA supercoiling are not associated exclusively with adaptation to acid stress. They have also been detected in bacteria experiencing changes in temperature (Goldstein and Drlica, 1984), osmotic pressure (Higgins *et al*., 1988), carbon source (Balke and Gralla, 1987), growth phase (Conter *et al*., 1997), oxygen concentration (Cameron *et al.*, 2013) and hydrostatic pressure (Tang *et al.*,

1998). Bacteria with very simple genomes and few conventional transcription factors exploit variable DNA supercoiling as a component of their repertoire of gene regulators (Dorman, 2011; Zhang and Baseman, 2011). These observations support a model of gene regulation in which DNA topological variation in response to environmental stimuli provides a means simultaneously to affect the potential of multiple genes for expression, with targeted inputs from DNA binding proteins providing the specificity that determines whether or not the potential for expression is realised.

Information from model pathogens such as *Salmonella* will guide us in our search for bacteria with pathogenic potential in the oceans. The clues from the marine environment do not have to consist of complete, living, culturable organisms: their genomic DNA signals can be enough once we know what to look for.

Acknowledgement

Research in the author's laboratory is supported by a Principal Investigator Award (13/IA/1875) from Science Foundation Ireland.

* this chapter is to be cited as :

Dorman C.J. 2017. Environmental stress and the control of gene expression in pathogenic bacteria. pp. 69– 75 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

The management and utility of whole genome sequence data for infectious disease surveillance and intervention: from hospitals, to fish-farms, to the oceans

Nicola Coyle, Sion C. Bayliss, Harry A. Thorpe, Edward J Feil

The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, UK

The first surveys of molecular variation in natural populations using multi-locus ezyme electrophoresis (MLEE) in the 1960s revealed unexpectedly high levels of diversity, and helped frame fundamental debates about the relative roles of natural selection and genetic drift that have continued to the present day (Milkman, 1973). In subsequent years, progressive waves of molecular data, driven by technological advances, have informed on genome dynamics in increasingly fine detail. These quantum leaps have impacted most significantly on our understanding of microbial evolution, and in particular pathogenic bacteria that cause disease in man or commercially important animals. In addition to providing evidence on key evolutionary processes such as mutation and horizontal gene transfer, the establishment of global databases held on the internet have revolutionised molecular epidemiological disease surveillance.

The most recent wave of molecular data is perhaps the most significant of all. The current decade has seen the advent of next-generation sequencing platforms that have provided the means to assay the vast majority of single-nucleotide polymorphisms (SNPs) and gene content changes throughout entire genomes for large population samples. The era of "population genomics" has thus provided unprecedented resolution into microevolutionary changes over increasingly small temporal and spatial-scales, and vastly increased our understanding of the genome dynamics underpinning adaptation. An early motivation for the generation of whole-genome sequence (WGS) data, and still one of the key drivers is in the management of bacterial pathogens of public health importance. Proof-of-principle was first demonstrated by Harris *et al* (Harris *et al*., 2010), who sequenced 61 isolates of a specific lineage of hospital-acquired Methicillin Resistant *Staphylococcus aureus* (MRSA ST239). The study demonstrated how previous definitions of a single bacterial "strain" (based on methodologies such as Multilocus Sequence Typing or Pulsed-Field Gel Electrophoresis) masked finer-scaled variation at the genomic level that is highly spatially structured. For example, the lineage studied by Harris *et al*, namely ST239, was previously considered a single globally disseminated

strain, but the genome data revealed sub-lineages restricted to Europe, South America and Asia. Moreover, this spatial sub-structuring was apparent over the full range of geographical scales, from whole continents down to a single hospital.

The increased resolution afforded by WGS data thus provides unparalleled power to reconstruct individual transmission events, and to track the emergence and spread of resistant and/or virulent clones. However, the utility of these data goes far beyond ultra-high resolution bacterial typing. A further significant advantage of WGS is that it provides predictive data concerning clinically important phenotypes, notably antibiotic resistance and virulence, through database interrogation. For most major healthcare-associated pathogens, existing databases hold many key antibiotic resistance determinants, and *in silico* predictions of resistance profiles from WGS data consequently have high sensitivity and specificity. A number of dedicated databases and platforms have been developed to facilitate *in silico* predictions of resistance profiles, including ResFinder (Zankari *et al.,* 2012), Abricate (https://github.com/tseemann/abricate), CARD(McArthur *et al.* 2013), Mykrobe(Bradley *et al.*, 2015), ARIBA(Hunt *et al.,* 2017) and WGSA.net .

Virulence determination presents greater challenges, as this is a more complex phenotype commonly involving the interaction of many genes and gene regulators, in addition to possible host effects. However, significant analytical advances in Genome-Wide Association Studies (GWAS) have been made over recent years (Read and Massey, 2014; Farhat *et al.,* 2014). For example, programs such as SEER (Lees *et al.,* 2016), and Scoary (Brynildsrud *et al.,* 2016) can compare phenotypic data (for virulence or any other trait), and identify genes or SNPs that are statistically associated with these traits. Although the high level of population structure commonly found in bacterial populations can confound such analyses, these approaches will in time lead to a more complete understanding of the genetic basis of virulence, thus more comprehensive databases and increased sensitivity and specificity of *in silico* phenotypic predictions from WGS data.

The advent of WGS for bacteria thus simultaneously informs both on the evolutionary relatedness of an isolate (ie *what is it*?) and the phenotypic properties of an isolate (ie *what can it do?*). These two questions do not always inform on each other, a fact which can go a long way to explain the fundamental muddle at the heart of bacterial systematics (Turner and Feil, 2007; Feil, 2015). Moreover, the partition of bacterial genomes into a stable “core” and a dynamic “non-core” broadly reflects this distinction. Core genes are commonly defined as those universally present (or nearly so) within a given species, and are widely presumed to be primarily subject to stabilising selection, and mostly have an assigned function. The SNPs within these genes are used for phylogenetic reconstruction, and hence lineage definition. Non-core genes are variably present or absent, and are in fact typically very rare, and most have an unknown function or are assigned as “hypothetical proteins”. Whole-genome sequencing has revealed that, in many bacteria, individual strains frequently recruit non-core genes from a seemingly endless genetic reservoir in the environment. The total complement of genes observed across all strains (the so-called “pan-genome”) often numbers tens of thousands of genes, or an order of magnitude more than the number of genes present in any single genome. In contrast, the “core-genome”, which refers to the complement of genes present in all (or the vast majority) of sampled isolates, can be significantly smaller than the total number of genes in any given genome (Medini *et al.,* 2005; Page *et al.,* 2015). For example, a study of 328 *Klebsiella pneumoniae* isolates, each of which harbours 4-5,000 genes, revealed a pan-genome of 29,886 genes; only 1,888 (6.8%) of which were universally present (Holt *et al.,* 2015). Our unpublished analysis of 1,103 genomes of *Vibrio parahaemolytics* revealed a total of 60,526 genes, 97% of which were variably present or absent. Striking differences in gene content are even evident for single lineages at the sub-

species level which, until recently, would have been considered a single "strain". For example genome data for 228 isolates of the clinically important lineage *Escherichia coli* ST131 revealed a pan-genome of 11,401 genes, of which 2,722 (23.9%) were core (McNally *et al.*, 2016). Our unpublished analyses on individual clinical clones of V. Parahaemolyticus also reveal striking differences in gene content.

Although the vast majority of accessory genes are of unknown function, there is growing recognition that the acquisition of new genes through horizontal gene transfer (HGT) plays a central role in ecological adaptation (Vos *et al.*, 2015). The emergence and spread of antibiotic resistance, underpinned by the transfer of plasmids and other MGEs, is a pertinent example. The increasing availability of datasets containing thousands of isolates thus offers an unprecedented opportunity for describing the genetic basis of bacterial adaptation. However, the scale of these datasets presents serious logistic and conceptual challenges in terms of data management and analysis. Pioneering pan-genome analysis tools, such as PanOCT and PGAP relied on all-vs-all BLAST comparisons between protein sequences, and scaled approximately quadratically with the number of isolates (Fouts *et al.*, 2012; Zhao *et al.*, 2012). More recently, the Roary pipeline has rapidly gained in popularity for scalable, user-friendly, pan-genome characterisation (Page *et al.*, 2015). Roary uses a pre-clustering step based on CD-HIT (Fu *et al.*, 2012), meaning that it can analyse thousands of isolates relatively quickly using modest computing resources. The output from Roary can be visualised using online tools such as Phandango (http://www.biorxiv.org/content/early/2017/03/22/119545), which displays the presence / absence of accessory genes against a phylogenetic tree typically reconstructed based on SNPs in the core genome (Figure 1). In addition to protein-coding regions, variation in intergenic regions (IGRs) can also play a central role in adaptation (Thorpe *et al.*, 2017) and we have developed a pipeline that emulates Roary except that it focuses on these regions (http://www.biorxiv.org/content/early/2017/08/22/179515).

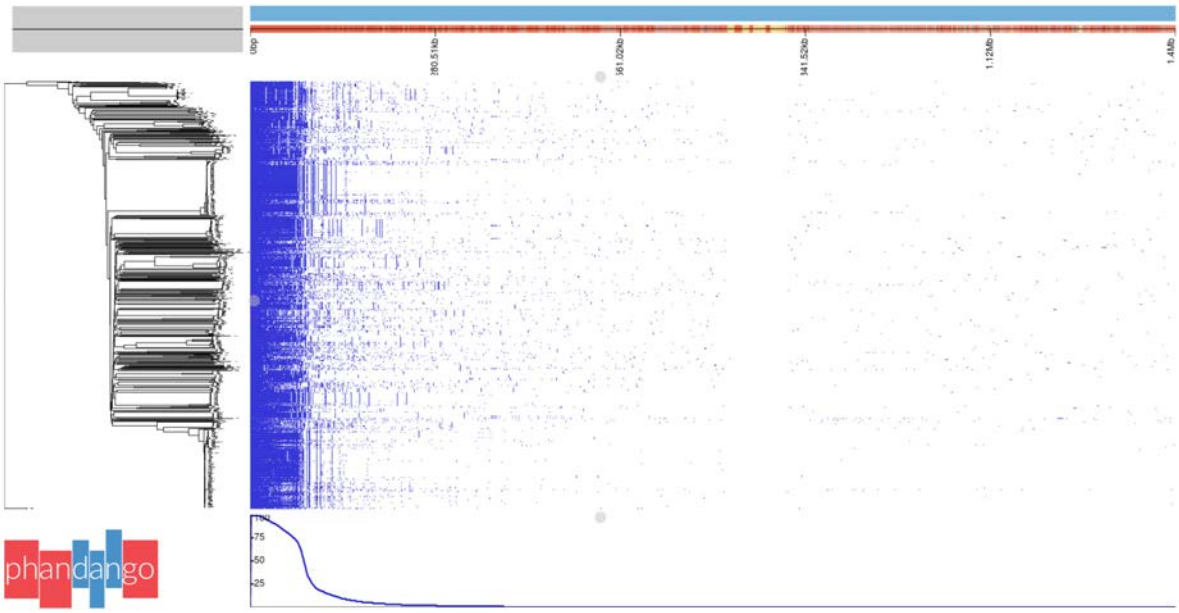


Figure 1: The pan-genome of 736 isolates of *Vibrio parahaemolyticus*. The tree on the left-hand side is constructed using the SNPs on the core genome. Blue dots indicate the presence of a gene. The plot at the bottom indicates the frequency of each gene in the sample. In total there are 70,125 genes, over 98% of which are non-core. The number of core genes (those present in all isolates) is only 1248. The figure was generated using Phandango (http://www.biorxiv.org/content/early/2017/03/22/119545).

An important caveat with pan-genome analyses is that the total number of genes observed is dependant upon the thresholds of protein (or nucleotide) identity used to define gene clusters. For example, as mentioned above, we found that 1,103 genomes of *Vibrio parahaemolytics* contained a total of 60,526 genes. However, this analysis was based on a threshold of 70% protein identity to define genes. Figure 1 illustrates the pan-genome of 736 of these isolates, but which contains more genes (70,125) than the total dataset. This is because the threshold protein identity for defining gene clusters has been made more conservative (95% identity). Currently there is no robust conceptual framework that can be used to inform a threshold level of identity that is most biologically and evolutionarily meaningful.

In addition to public health, the utility of WGS is now being deployed to manage pathogens of agriculture and aquaculture (Bayliss *et al.*, 2017). Aquaculture, which encompasses both finfish and shellfish, is the fastest growing food-producing sector, and plays a critical role in both food security and economic welfare in many developing nations, particularly in Asia. Since 2014, the majority of finfish consumed globally are farmed rather than caught from wild stocks. However, the rapid expansion of aquaculture, both in terms of scale but also diversification of target species, presents serious challenges with respect to sustainability, particularly in regard to infectious disease management. Fortunately, many of the analytical methods and platforms developed for WGS data of human pathogens can be applied, with minimal modification, to aquaculture pathogens (Bayliss *et al.*, 2017).

Whole genome sequencing of a large population sample of an aquaculture pathogen was first deployed as a molecular epidemiological tool by Brynilsdud *et al* (2014) who focused on *Renibacterium salmoninarum*, the causative agent of bacterial kidney disease (BKD) in salmonids. Similar to the previous study by Harris *et al* on MRSA, this study revealed a high level of previously undetected spatial structuring, and the utility of WGS data for tracking transmission over a large range of geographical scales. These data can be freely explored using the online interaction visualisation tool Microreact of the *R. salmoninarum* project URL: https://microreact.org/project/N1KdygLt (see Figure 2 for a screenshot). This tool provides the means to gauge the spatial structure in the context of the phylogenetic tree and associated metadata. The tree, map and metadata windows are mutually interactive, so that it is possible to choose a strain on the map, which is then highlighted on the tree (or *vice versa*), or to annotate the tree according to different metadata fields, including the presence / absence of specific resistance or virulence genes. Full details and instructions for the Microreact tool are available at microreact.org or described in Argimón *et al.* (2016).

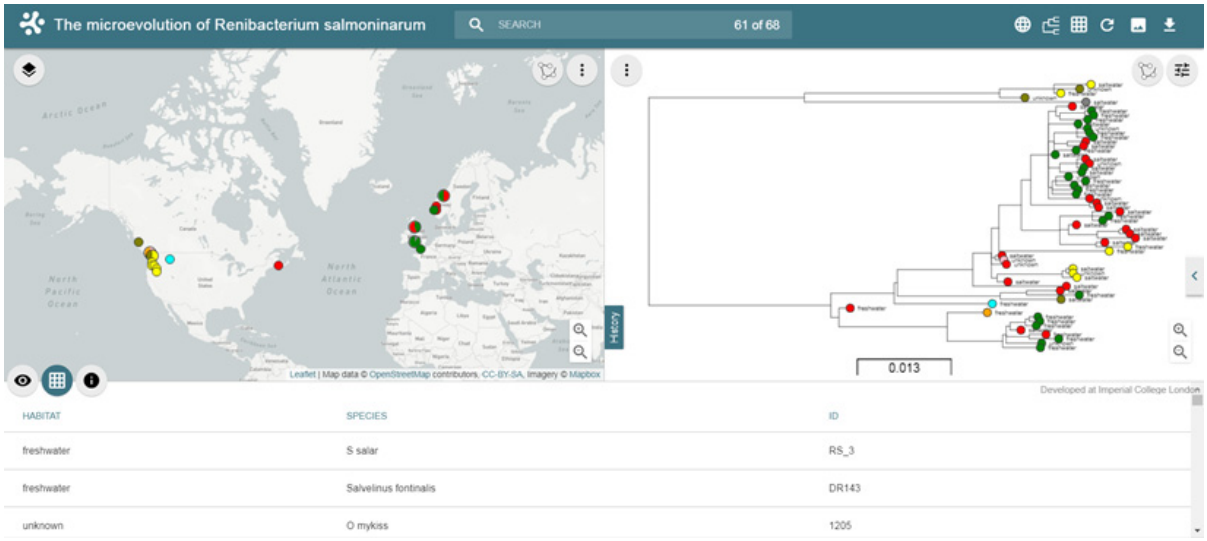


Figure 2: Screenshot of the R. salmoninarum data of Brynildsrud *et al.*, 2014 visualised using Microreact (Argimón *et al*. 2016). The project is free to explore at https://microreact.org/project/N1KdygLt.

At one extreme of geographical scale, the data indicated that *R. salmoninarum* was introduced into Europe from North America over the last 50 years, possibly via the trade in eggs and fry. At the other extreme, the authors showed that WGS can be used to detect outbreak variants at a very local level; that is, between neighboring farms. Moreover, the data informed on a basic biological property of the pathogen that has relevance for fish husbandry; that is, the degree to which the pathogen is adapted to a single host species or can freely switch between different species. The authors noted nearly indistinguishable isolates from the same geographical location, but from different host species, which suggests that host switching by this pathogen, can occur freely within the farm environment. This has implications for the sustainability of mixed farms, and also for assessing the risk of pathogen spillover into wild stocks.

The study by Brynilsdrud *et al.* (2014) touched on many questions that have been explored in more depth in subsequent studies on aquaculture pathogens. The global movement of pathogens via trade in eggs and fry is likely to be a key driver for pathogen dissemination on a global scale. Our subsequent unpublished WGS data on *R. salmoninarum* examines the origin of this pathogen in Chilean salmon farms, and points to at least three separate introductions from North America or Europe. In contrast, the study by Barnes *et al.* (2016) on *Yersinia ruckeri* (the causative agent of Enteric Red Mouth disease, ERM, in salmonids) demonstrated that the disease-causing strains in Australia and New Zealand are distinct from those in the Northern Hemisphere, and thus have not been imported but are instead local endemic strains.

The question of host adaptation is also being addressed in our unpublished study on the Gram-positive pathogen *Lactococcus garviae*. This is a genetically diverse environmental generalist that can colonise many different animal hosts and is an occasional opportunistic pathogen of humans. The pathogen has caused multiple outbreaks in rainbow trout farms in Southern Europe and the USA. Our WGS data show that these outbreaks have been caused by the independent emergence of lineages adapted to the rainbow trout host from environmental reservoirs, and comparisons of the outbreak clones reveal convergent gene acquisition that shed light on the genetic basis of this adaptation. A further unpublished study sequenced multiple isolates of *Aeromonas salmonicida* responsible for a prolonged outbreak of furunculosis in a large salmon farm in China. These data revealed how resistance to the antibiotics being used to treat the outbreak was conferred by the acquisition of novel plasmids from

environmental sources. Moreover, the data revealed evidence of sub-variants of the outbreak clone circulating within different workshops within the single farm.

A more complete understanding of the emergence of bacterial pathogens, or new resistant or virulence variants, relies critically on a greater focus on the environmental reservoirs in which these strains or genes first evolve. The marine environment is highly structured, particularly with respect to temperature and salinity, and our understanding of the phylogeography of bacterial pathogens of humans and animals in the marine environment is far from complete. However, there is strong evidence that variation in temperatures such as caused by El Niño events (Martinez-Urtaza *et al.,* 2016) or climate change, can have impacts on the frequency of infections caused by pathogenic *Vibrio* species. The summer of 2014 even witnessed cases of *V. parahaemolyticus* infection in sub-arctic regions due to an unprecedented heatwave across southern Scandinavia (Baker-Austin *et al.,* 2016). Our large *V. parahaemolyticus* dataset represents isolates recovered from a large geographical range and from multiple different sources, including human disease, infected finfish and shellfish from aquaculture settings, and the environment. We are currently mining these data for key virulence and resistance genes in order to identify those environmental compartments that might pose the greatest risk with respect to the emergence of new virulent variants.

In conclusion, whole-genome sequencing has revolutionised our ability to track the emergence and spread of bacterial pathogens and has greatly advanced our understanding of bacterial adaptation and genome dynamics. However, the utility of WGS data is critically reliant on the establishment of robust and community-oriented database structures for managing the data, and analysis tools for drawing biological inference. The frameworks developed for pathogens of public health importance can be readily adapted for aquatic pathogens of importance to aquaculture, but also to understanding the evolution and ecology of pathogens in the broader marine environment.

* this chapter is to be cited as :

Feil E.J., Coyle N., Bayliss S.C and Thorpe H.A. 2017. The management and utility of whole genome sequence data for infectious disease surveillance and intervention: from hospitals, to fish-farms, to the oceans. pp. 77 – 82 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

New potential NGS-based markers for detection of *Vibrionaceae* in marine environments

Aleksei Korzhenkov, Bogdan Efimenko, Stepan Toshchakov

Immanuel Kant Baltic Federal University

ABSTRACT

Extremely high rates of pathogen spread in marine environments require pathogen-specific and, at the same time, high-throughput and robust techniques for epidemiological monitoring. Due to the low abundance of prospective pathogens within microbial communities, low specificity and low target discrimination fuel the system, 16S-based methodologies of microbial community profiling do not seem to be applicable for monitoring of epidemiological risks. In this paper, we discuss current microbial community profiling methodologies in light of detecting pathogens in the ocean and provide a rationale for the development of new NGS-based markers for the detection of widely spread pathogens from the *Vibrionaceae* family. Since genes, chosen by our bioinformatic analysis are extremely abundant and amplified in genomes of pathogenic microflora, as opposed to non-pathogenic microorganisms, we expect this system to be more specific and sensitive in respect of pathogenic microflora, than standard 16S-based molecular techniques.

INTRODUCTION

Dynamics of spread of pathogenic bacterial lineages in open marine and oceanic ecosystems is influenced by multiple factors, including sea currents, migrating fishes and mammals, maritime traffic and unusual weather activity. In several reported cases rates of epidemic spread can reach 5 000 km/year which significantly outweighs maximal rates of infection spread reported for terrestrial environments (McCallum *et al*., 2003). Therefore, constant monitoring of the level of pathogens in marine microbial communities is very important both for public health and food supply security.

Rapid development of next generation sequencing techniques in the first decade of this century led to the revolution in molecular microbial ecology. Scientific community got an instrument to study not culturable and minor fractions of microbial communities by high-throughput profiling of 16S rRNA gene or by shotgun metagenomic sequencing. Nevertheless, despite the significant progress in microbial profiling and detection techniques, the assessment of risks, associated with marine pathogenic microorganisms remains untrivial due to several factors:

> (i) Minor abundances of pathogenic microorganisms in comparison to major marine taxa (this volume) might require serious sequencing efforts to achieve significant coverage, which, it turns results in high costs, inappropriate to large-scale monitoring;

(ii) while being a powerful tool for the detection of pathogenic DNA in the environment, current microbial community profiling methodologies do not provide information regarding the viability of potential pathogens and therefore cannot be used as explicit method for the assessment of pathogenic risks in the environment;

(iii) one of the most simple and economically efficient methods of the analysis of marine microbial communities by 16S rRNA gene in most cases does not achieve a sufficient resolution to distinguish between pathogenic and non-pathogenic strains of the same species (Bruto *et al*., CIESM Chapter; Brooks JP *et al*., 2015; Ellegaard and Engel, 2016), therefore requiring additional experiments.

All these problems determine an increased demand in the development of molecular and microbiological tools for monitoring the level of pathogens in marine microbial communities, both in water column and sediments. In this paper we discuss current techniques, which were developed to overcome these issues and propose a rationale for the development of new NGS-based systems for the detection of pathogens in marine environments.

Problem I. Low abundance of pathogenic microorganisms

Enrichment of pathogenic DNA for shotgun metagenome sequencing

Pathogenic microorganisms are present in marine environments as "extreme minority" of the total microbial community and, therefore, can be classified as a rare biosphere (Troussellier *et al*., 2017; this volume). As a result, one of the largest barriers to the implementation of next generation sequencing in systematic pathogenic risk surveillance of marine environments is the low abundance of "pathogenic DNA" in the total DNA pool. Hence a vast majority of sequencing read, generated during shotgun NGS experiments, corresponds to non-pathogenic microbiota. This determines the high sequencing coverage needed for the reliable detection of pathogenic DNA and, therefore, high sequencing costs, inappropriate for large scale regular monitoring of marine environments.

There are two major strategies to overcome this issue. First, the "classical" one involves *in vivo* enrichment of pathogenic taxa using immunomagnetic separation or immunochromatography assays (Sakata *et al*., 2017). While this method works extremely well, it requires a prior knowledge of the target pathogen and very short sample handling time to keep microbial cells intact.

The second strategy implements enrichment on the DNA level and therefore excludes strict sample requirements. Total fragmented metagenomic DNA (or cDNA) is mixed with pre-designed oligonucleotide baits, which, in turn can be either biotinylated for or bound to the solid surface (on chip). Non-target DNA does not bind oligonucleotide baits and is eluted during washing steps. At the end enriched target DNA is eluted from the chip or magnetic beads in much higher concentration than found in the initial sample. This strategy is widely used for the separation of pathogen and host DNA, but reports concerning enrichment of specific taxa from microbial communities remain scarce. Nevertheless, Vezzulli *et al*. reported successful enrichment of pathogenic DNA of *Vibrio cholera* using RNA-baits on chip, confirming the presence of toxigenic *V. cholerae* O1 metagenomic DNA in an African river (Vezzulli *et al*., 2016). It should be also noted that cost of hybridization probes needed for the whole genome enrichment of pathogenic DNA remains high and currently may exceed the cost of sequencing. Therefore, these techniques can be rather applied to targeted experiments, but not to large-scale monitoring.

Detection of low abundant pathogenic DNA with species-specific 16S rRNA profiling

Universal 16S rRNA gene primer sets, commonly used for NGS microbial community profiling, amplify variable regions of 16S rRNA of many different taxa. Due to that fact, low abundant pathogenic microorganisms are quantified in very low levels, comparable to the level of contamination. Moreover, serious biases are observed; depending on 16S primer set, sequencing

protocol and analysis pipeline (Ceuppens *et al*., 2017). Better results are usually observed when genus-specific 16S primer sets are used (Pereira *et al*., 2017), but such approach requires optimization of separate PCR reaction for every minor taxa (genus), and therefore cannot be easily implemented for environmental monitoring.

Problem II. Pathogen viability

As discussed in detail by (Villaraya *et al*., 2017, this volume**)**, DNA-based microbial community profiling techniques do not provide information regarding metabolic status of microbes. Therefore, it is hard to assess the level of potential virulence of pathogen-related taxa identified in metagenomic experiments. Metatranscriptomic experiments utilizing rRNA itself instead of rRNA gene (DNA) show that there are many significant differences between transcriptionally active and total microbial communities of human gut (Peris-Bondia *et al*., 2011). Similar observations were made by Gentile and co-authors (2006) for Antarctic coastal waters: some taxa were very abundant, but obviously not active, as opposed to several underrepresented taxonomic groups, showing high transcriptional activity at RNA level. Furthermore, several metabolically active phylotypes were only detected in RNA libraries, indicating the presence of "active minority" in microbial communities. Recently published analysis of atmospheric communities showed that relative transcriptional activity of rare taxa was several times higher than it could be expected from total community rDNA analysis (Klein AM *et al*., 2016).

All these facts indicate that assessment of pathogenicity risks in the ocean should include complementary approaches, both metagenomic for the detection of "silent" pathogens and metatranscriptomic for the detection of low abundant, but potentially virulent taxa. In addition to metagenomic and metatranscriptomic tools, several techniques, based on fluorescence-based cell sorting, were developed to isolate active part of microbial community. These approaches are reviewed in detail in another chapter of this Monograph (Villaraya *et al*., this volume 2017).

Problem III. Method resolution.

Exact taxonomical identification of microorganisms is the key to the reliable detection of pathogenic risks. At the same time many microbial genera and even species include both extremely virulent and non-pathogenic microorganisms. That applies some limitations to the use of 16S-based community profiling for the pathogen monitoring. It especially concerns NGS-based techniques, where taxonomy is assigned on the basis of short 200-500 nt variable part of rRNA gene. Widely spread pathogens of the *Vibrionaceae* family can be a perfect example: even the whole rRNA gene phylogenetic tree does not show a good resolution of *Vibrio* clades (Figure 1).

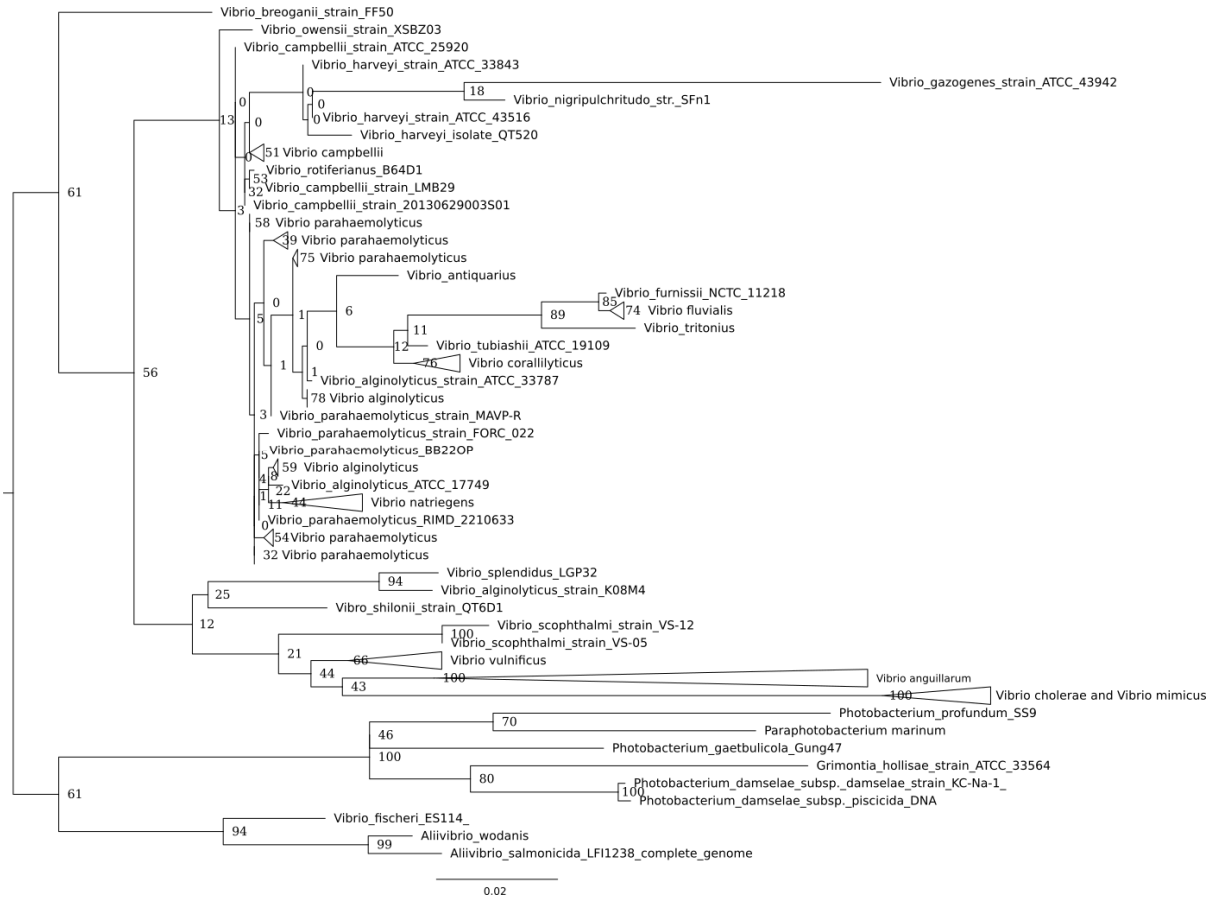


Figure 1. Figure 1. Maximum-Likelihood phylogenetic tree of Vibrionaceae representatives based on 16S rRNA gene. Tree was constructed in RAxML (Stamatakis A., 2014) using GTR substitution model, GAMMA+P-Invar model of rate heterogeneity, ML estimate of alpha-parameter, 100 bootstraps. Bootstrap values are presented next to the tree nodes. Scale bar shows 0.02 substitutions per site. Monophyletic groups of same or closely related species' strains were collapsed.

For that reason, nowadays only multi locus sequence typing (MLSA) method is widely used for the discrimination of *Vibrio* strains (Gabriel *et al*., 2014). The resolution power of this technique is very good and is comparable to whole genome sequencing, it involves Sanger sequencing step and is therefore quite expensive. At the same time, very little efforts were made to move the scope from 16S rRNA to other potential phylogenetic NGS markers. Probably, this might be linked to the low level of diversity of housekeeping prokaryotic genes, such as ribosomal proteins, RNA-polymerases, etc. Also a lack of existing curated taxonomic databases for functional genes with well-defined taxonomy and compatibility with NGS-based community profiling pipelines (such as SILVA database, see Quast *et al*., 2013) applies some challenges for further analysis.

Nevertheless, the *fur* gene encoding transcriptional factor involved in the regulation of ferric uptake was recently proposed as a phylogenetic marker for *Vibrio* (Machado and Gram, 2015, Machado *et al*., 2017). Despite this gene might alone be very good for the classification of *Vibrio* species, active horizontal gene transfer, well described for *Vibrio* (Metzger and Schulte, 2016), necessitates the development of additional phylogenetic markers. Below we provide a rationale for the development of new NGS-based markers for the detection of widely spread pathogens from the *Vibrionaceae* family and suggest two genes, which can be used both for rRNA and rDNA analysis.

SEARCH FOR NEW PHYLOGENETIC MARKERS FOR *VIBRIO* CLASSIFICATION

The ideal phylogenetic marker for use in NGS-based microbial community profiling should meet the following criteria:

a. gene should be present in one copy in each genome;
b. gene should contain both conservative (for annealing of universal primers) and variable (for species discrimination) regions;
c. length of target variable regions should not exceed maximal readlength of sequencing instrument.

To use our marker in rRNA metatranscriptomics profiling of active pathogens we also applied a fourth criterion:

d. gene should be associated with virulence, giving the opportunity to detect minor, but active pathogenic species.

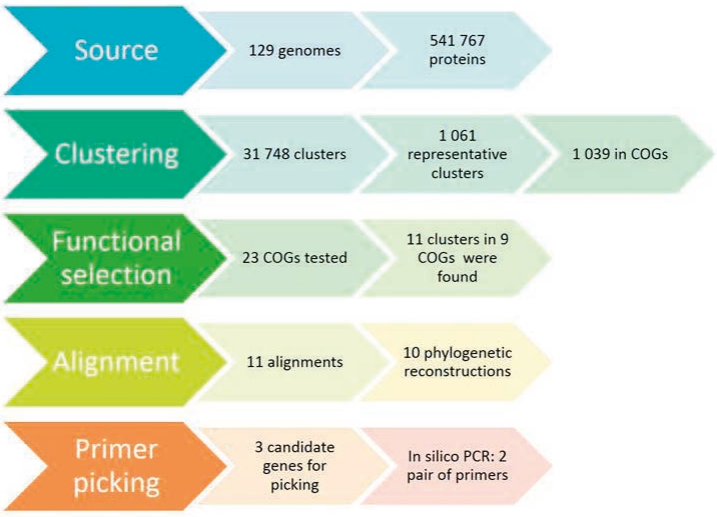


Figure 2. Schematic representation of a strategy for search of new NGS-based phylogenetic markers of *Vibrionaceae*.

As a primary dataset for the mining *Vibrionaceae* for new phylogenetic markers meeting the abovementioned criteria we used 129 complete *Vibrionaceae* genomes and *in silico* predicted proteomes were downloaded from NCBI RefSeq database (September 2017). The strategical scheme for the analysis is presented in Figure 2.

As a first step, in order to minimize potential laterally transferred genes and multicopy biases we picked only proteins, presented in one copy in each genome. To accomplish this, all 541 767 protein sequences were aligned 'all vs all' using diamond algorithm (Butchfink, 2015). Orthologous protein clusters were formed using OrthoMCL (Li *et al*., 2003). Protein clusters, containing more than one protein in any organism, as well as clusters, represented in less than 95% of genomes, were rejected. After that we obtained 1061 representative protein clusters, 1039 of which were assigned to function by blastp alignment algorithm (Altschul *et al*., 1997) against bacterial COG database (Galperin, 2014).

To get only proteins, associated with virulence and therefore highly expressed in pathogenic microorganisms, we used the list of 23 virulence-associated COGs (Table 1). 11 candidate proteins, corresponding to 9 virulence-associated COGs were identified by this approach. Nucleotide sequences of candidate genes, encoding identified proteins, were aligned using Muscle (Edgar, 2004) and used for phylogenetic reconstruction. Alignment quality and the presence of conservative and variable regions were assessed using JProfileGrid (Roca *et al*., 2011). Ten alignments showed sufficient quality and the presence of conservative and variable regions.

Phylogenetic reconstruction of candidate *Vibrionaceae* genes was conducted in RAxML (Stamatakis, 2014) using GTR substitution model, GAMMA+P-Invar model of rate heterogeneity, ML estimate of alpha-parameter 100 bootstraps were made. To assess the quality of the trees and therefore the usability of candidate genes for phylogenetic classification of *Vibrio* we compared its topology with *in silico* MLSA tree generated from the concatenated alignment of 20 ribosomal and conservative single copy genes found in selected *Vibrionaceae* genomes. Phylogenetic trees for PulG (Figure 3), ArsR (Figure 4) and TetR (data not shown) have the best match with concatenated single copy gene tree, so that three genes were selected for primer picking. These genes were arsR, metal responding transcriptional regulator involved in hemolysin expression in some Vibrio species (Saha and Chakrabarti, 2006), *pulG*, type II secretion system protein Involved in secretion of virulence factors (Gray *et al*, 2011) and tetR transcriptional regulator Involved in quorum-sensing (Pompeani *et al*., 2008). Visual inspection of these trees, in particular branch length and bootstrap values, confirmed that these genes showed good species resolution as well as length of terminal clades.

Primers were picked and initially tested with Geneous software. In silico PCR was conducted in EMBOSS primersearch (Rice, 2000) using Vibrionaceae complete genomes as reference. Despite all of these prospective primers contain some degeneracies (up to 6), optimization of PCR reaction as well as introduction of modified nucleotides (LNA), improving base pairing should yield a positive result. Finally, one pair of primers for ArsR gene with mean amplicon length of 215 bp and one pair of primers for PulG gene with mean amplicon length of 400 bp were found to anneal on to all of *Vibrionaceae* genomes analyzed.

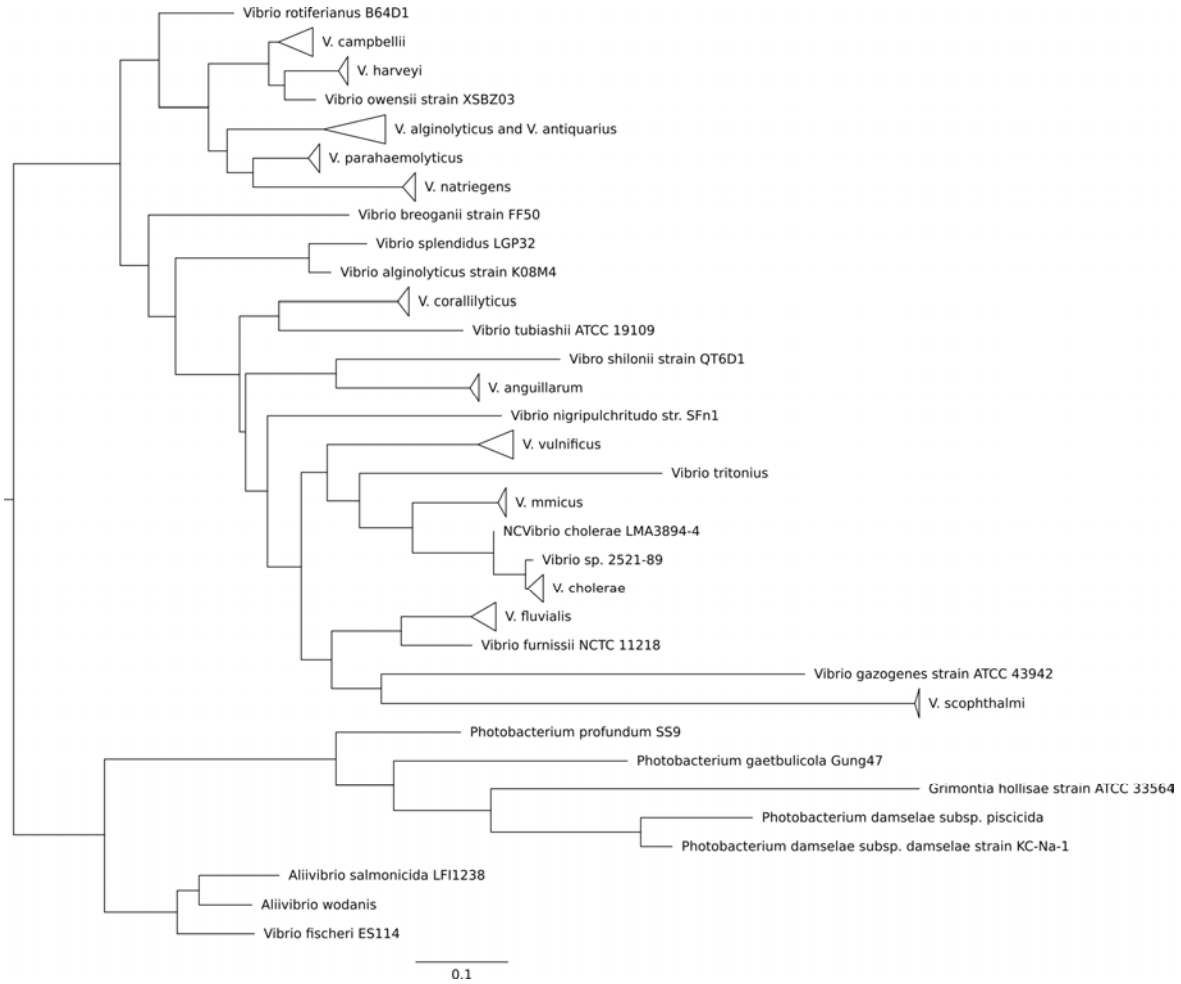


Figure 3. Maximum-Likelihood phylogenetic tree based on nucleotide sequences of *pulG* gene. Tree was constructed in RAxML (Stamatakis A., 2014) using GTR substitution model, GAMMA+P-Invar model of rate heterogeneity, ML estimate of alpha-parameter, 100 bootstraps (not shown). Scale bar shows 0.1 substitutions per site. Monophyletic groups of same or closely related species' strains were collapsed.

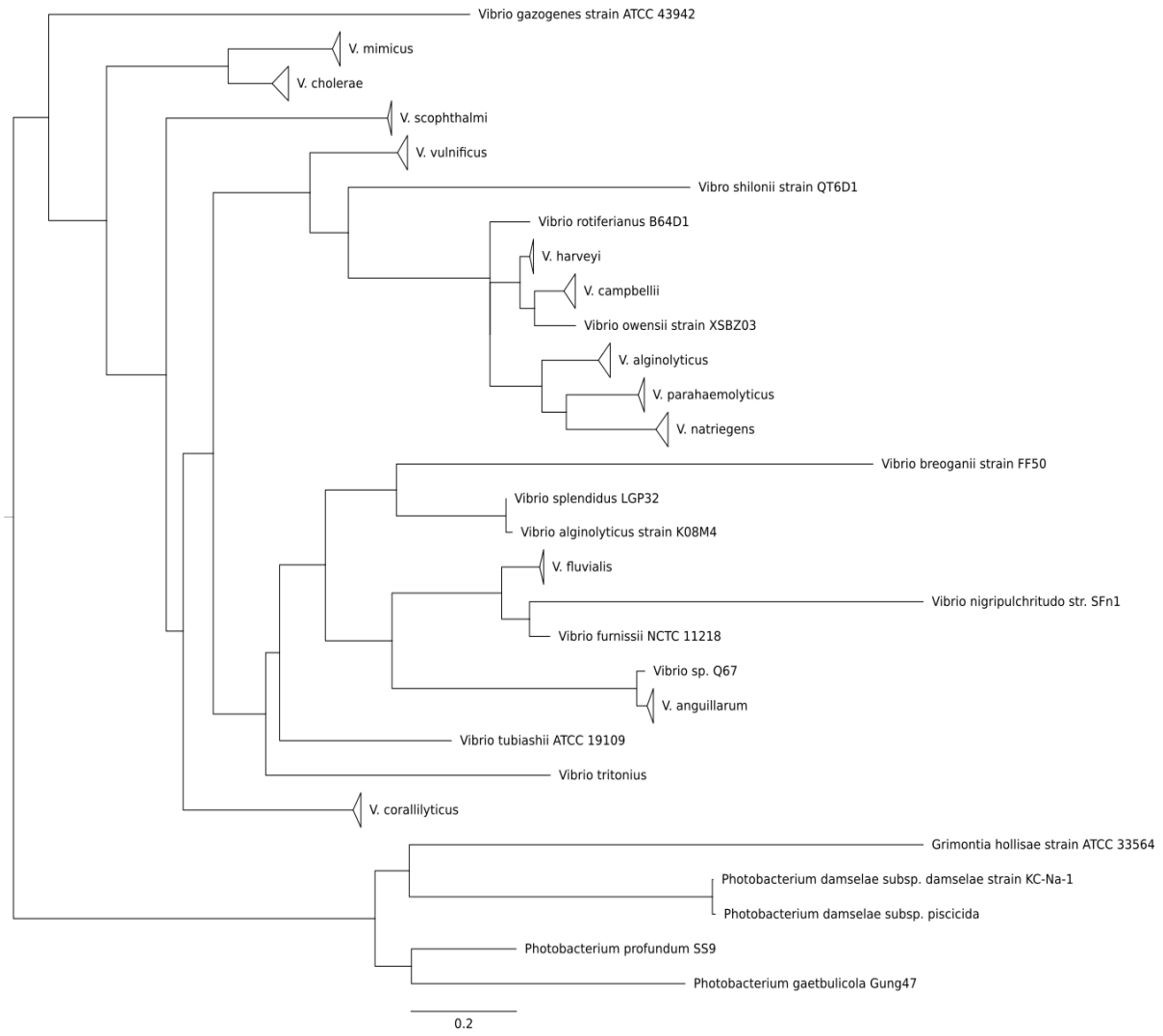


Figure 4. Maximum-Likelihood phylogenetic tree based on nucleotide sequences of *arsR* gene. Tree was constructed in RAxML (Stamatakis A., 2014) using GTR substitution model, GAMMA+P-Invar model of rate heterogeneity, ML estimate of alpha-parameter, 100 bootstraps (not shown). Scale bar shows 0.2 substitutions per site. Monophyletic groups of same or closely related species' strains were collapsed.

ADAPTATION TO ILLUMINA SEQUENCING

Proposed primer sequences can be used for introduction to any of currently developed 16S fusion primer system for the sequencing by Illumina technology. While Illumina kits propose the amplicon library preparation in to PCR steps (reference to illumine white paper for 16S), for the purpose of large-scale monitoring we would propose double barcoding system developed by Fadrosh *et al.* (2014) and co-authors since their strategy (i) allows to analyse several thousands of samples in one sequencing run and (ii) solves the problem of “sequence heterogeineity”, routinely observed during amplicon sequencing by Illumina technology.

CONCLUSION

Use of taxonomic markers other than 16S ribosomal RNA in next generation sequencing microbial community profiling techniques is largely underestimated. Here we propose a strategy for the development of NGS-based detection system of *Vibrionales* species, which can be used in large-scale

monitoring experiments due to dual-indexing sample barcoding technique. Despite the final conclusions regarding the usefulness of the system may be done only after sequencing and analysis, intermediary results of the *in silico* PCR look very promising.

[1] * this chapter is to be cited as :

Toshchakov S., Korzhenkov A. and Efimenko B. 2017. New potential NGS-based markers for detection of Vibrionaceae in marine environments. pp. 83 – 90 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

Investigating the macroecology of emerging *Vibrio* pathogens in the ocean using the Continuous Plankton Recorder technology

Luigi Vezzulli[a], William H Wilson[b], Carla Pruzzo[a]

[a] *Department of Earth, Environmental and Life Sciences (DISTAV), University of Genoa, Genoa, Italy*
[b] *Sir Alister Hardy Foundation for Ocean Science (SAHFOS), Plymouth, United Kingdom*

Summary

Vibrios still regarded by most marine microbiologists as the dominant culturable bacteria in the ocean represent an important cause of morbidity and mortality in humans and marine animals throughout the world. The 2010-2011 MCCIP Annual Report Card (www.mccip.org.uk/arc) that provides an up-to-date assessment of how climate change is affecting UK seas considered, for the first time, the potential future increases in marine vibrios as an emergent issue (also based on the unprecedented increase in the number of bathing infections associated with warm water *Vibrio* species in a number of northwest European countries, in recent years). This concern also applies on a global scale to most countries where human and non-human illnesses associated with these bacteria are increasing. However, until now there has been no observational or experimental evidence to support this view mainly due to a lack of historical data. Answering this global concern is pivotal to understanding, predicting and potentially managing climate change impacts on human and animal health associated with our oceans.

To address these issues we recently developed a novel approach based on the molecular analysis of formalin-fixed samples collected over the past half a century by the Continuous Plankton Recorder (CPR) survey. The CPR survey is one of the longest running marine biological monitoring programmes in the world, and provides a long-term archive of formalin preserved plankton samples. We exploited the well-established association between vibrios and plankton (that are considered to be the largest reservoir of these bacteria in nature) to assess a possible linkage between *Vibrio* occurrence in the sea and environmental variables over a decadal scale by applying a molecular and pyrosequencing analysis to the microbial community on historical CPR samples. We show here for the first time that vibrios, including the species *Vibrio cholera* (the causative agent of the diarrheal disease cholera) have increased in prevalence in the last 54 years in several coastal areas of the North Atlantic and North Sea and that this increase is correlated significantly with increasing sea surface temperature during the same period. Such increases are associated with an unprecedented occurrence of environmentally acquired Vibrio infections in the human population of northern Europe and the Atlantic coast of the United States in recent years.

As a follow up to these studies molecular methodologies, including a novel qPCR and Whole Genome Enrichment (WGE) NGS- based protocols, were developed and employed to detect and genotype *V. cholerae* DNA in historical CPR samples collected over more than 2000 km of coast in the Benguela

Current Large Marine Ecosystem (Southern Africa) which is known as an endemic area for cholera. Three samples, notably collected in coastal waters off the cities of Luanda in northern Angola and Cape Town and Port Elizabeth in South Africa, tested positive for *V. cholerae*. These findings are of particular significance and validate the concept of using the CPR technology in cholera studies.

Ultimately this is opening up a new avenue for depicting the large scale distribution of emerging pathogens in the ocean including the assessment and identification of the main factors which drive their occurrence and spread at a global scale.

Emergence of pathogenic Vibrio species in the marine environment

A growing number of human microbial infections have been associated with recreational and commercial uses of marine resources during the recent years. Whether these increases reflect better reporting or global trends is a subject of active research; given the heightened human dependence on marine environments for fisheries, aquaculture, waste disposal, and recreation, the potential for pathogen emergence from ocean ecosystems requires investigation. Among the surprising number of human pathogens that have been reported from marine environments, pathogenic *Vibrio* species are indigenous in aquatic systems and pose an emergent threat to human health. The primary route of human exposure to these microorganisms is through ingestion of contaminated seafood, but illness can also result from direct contact with seawater during recreational or occupational activities and from contact through aerosols containing toxins.

The genus *Vibrio* includes more than 30 species, at least 12 of which are pathogenic to humans, and/or have been associated with water- and food-borne infections (Oliver *et al.,* 2013). The infectious disease cholera caused by the pathogen *Vibrio cholerae* is responsible for the deaths of around 120,000 people worldwide. *Vibrio parahaemolyticus* and *Vibrio vulnificus* infections are associated with high morbidity and mortality throughout the world. Members from each of these species have become and continue to be formidable pathogens, especially *V. cholerae* O1 and *V. parahaemolyticus* serotype O3:K6, which are responsible for the only two existing bacterial pandemics. The greatest numbers of vibrio-associated illnesses in developed countries derive from the consumption of molluscs which concentrate particles present in surrounding waters. Recent studies also reported that non-foodborne *Vibrio* infections, mainly associated to the species *V. vulnificus*, *V. parahaemolyticus* and *non O1/O139 V. cholerae*, led to high morbidity and mortality especially in people with medical conditions predisposing to diseases.

Climate changes, leading to global and sea surface temperature rise in future years, are expected to increase human exposure to indigenous waterborne pathogens, for which growth opportunities increase (Vezzulli *et al.,* 2013). Global average temperatures have risen by nearly 0.8 ºC since the late 19th century, and have risen by approximately 0.2 ºC/decade over the past 25 years. Increased SST linked to El Niño events have been shown to pre-date increases in cholera incidence in both Asia and South America (Pascual et al., 2000). Similarly, climate and temperature anomalies promoted the expansion of the geographical and seasonal range of seafood-borne illnesses caused by the human pathogens *V. parahaemolyticus* and *V. vulnificus* (Martinez-Urtaza *et al.,* 2010). Evidence also links *Vibrio* infections to increasing mass mortality of marine life in the coastal marine environment (Vezzulli *et al.,* 2010a).

In Europe there has been an unprecedented increase in the number of bathing infections in recent years that have been associated with warm water *Vibrio* species (Baker-Austin *et al.,* 2013). Eutrophication of marine water as a direct consequence of human activities is also expected to increase the occurrence and abundance of these pathogens in the near future. However, despite the volume of indirect evidence, it is not clear whether vibrios, which are known to be thermodependant, are increasing within the complex and ecologically regulated bacterial communities in coastal marine waters. This is mainly due to a lack of historical data.

Ecology of Vibrios and epidemiology of their associated diseases

There is little understanding of the factors triggering *Vibrio* worldwide outbreaks and epidemics and of the presence and nature of a causal link between ongoing climate change and the registered spread of *Vibrio* illness. This is mainly due to a lack of historical data and of large sampling efforts.

In natural environments Vibrios are found attached to chitin, which is one of the most abundant polymers on earth and possibly the most abundant in the aquatic environment (Pruzzo *et al.*, 2008). In particular, chitin containing plankton, and especially copepods, represent one of the most important environmental reservoirs of these bacteria (Figure 1).

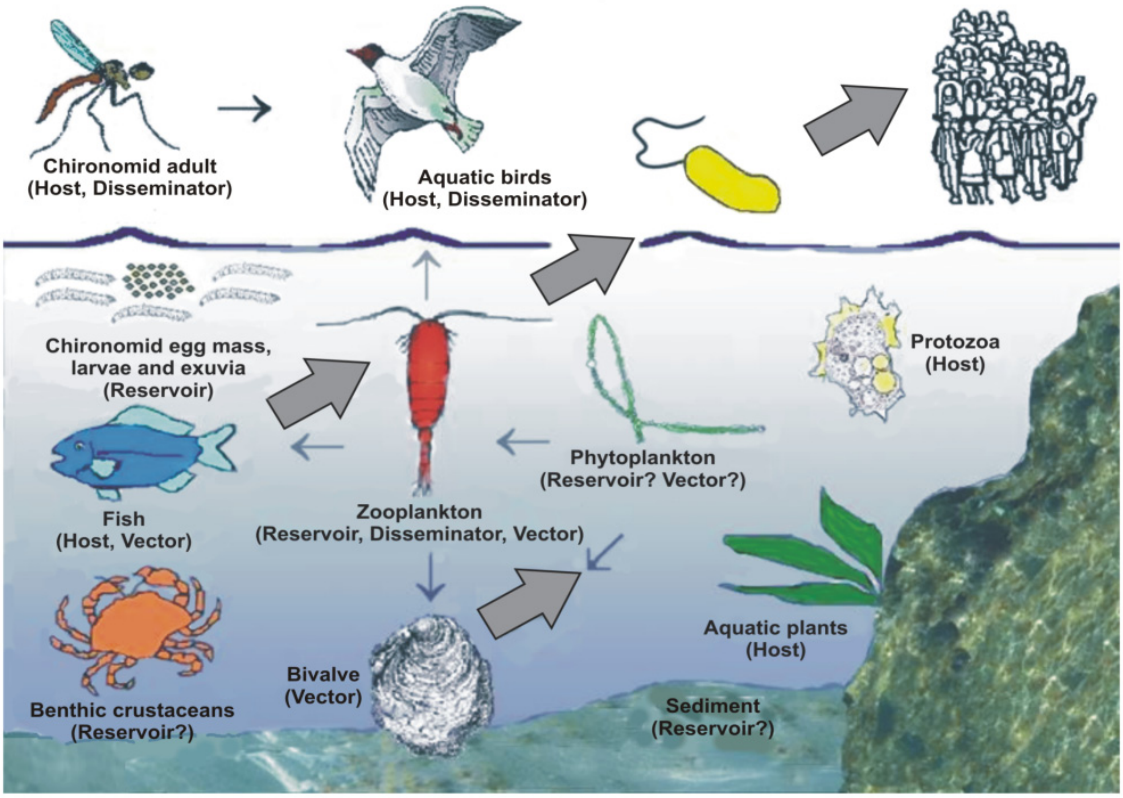


Figure 1. The variety of reported environmental reservoirs and hosts of *V. cholerae* in the aquatic environment. Question marks indicate only a putative role. Black arrows indicate documented interactions, e.g. trophic or incidental interactions. Heavy arrows indicate documented case of transmission from environmental reservoirs and/or hosts to humans (Adapted from Vezzulli *et al.*, 2010 b).

We know that in certain areas of the world the marine and brackish environment is the source of many epidemics, but little is known on the global occurrence and macroecology of vibrios in the aquatic environment, their link with large scale environmental and climatic process and with cases of clinical infections in human populations.

Global efforts to control vibrio epidemics, such as cholera, are focused on increasing water treatment and sanitation, public awareness of preventative and treatment measures, and access to health services that provide medicines and in particular oral rehydration salts and chlorine tablets. An important challenge is to improve poorly understood aspects of the epidemiology of the disease, especially those related to the origin of epidemics and the main factors that drive them. The role of large-scale hydroclimatic processes in propagating the disease during different seasons and at different spatial locations is also a crucial issue that is still poorly understood.

According to the "cholera paradigm" (Colwell, 1996), the epidemiology of the cholera disease is driven directly by aquatic reservoirs of *V. cholerae*. Studies conducted in Bangladesh have clearly demonstrated that the presence and spread of *V. cholerae* in the aquatic environment and the incidence of cholera outbreaks are strongly associated with the presence of zooplankton, which harbours the bacterium. Zooplankton abundance and distribution follow that of its food, the phytoplankton, whose growth in seawater is directly related to environmental variables such as the concentrations of nutrients and temperature. The latter are, subsequently, controlled by larger-scale climate variability (Lipp *et al.*, 2002). Improving knowledge of the occurrence, biogeography and ecology (e.g. deciphering the association between *Vibrio* and its plankton reservoirs and assessing global ecological factors which are driving vibrio re-emergence at the global scale) of vibrios in relation to macro-scale variability in marine ecosystems has considerable potential to help improve understanding of the epidemiology of their associated disease.

Continuous Plankton Recorder technology applied to *Vibrio* research: Rationale

Explorative and monitoring studies of the occurrence and ecology of vibrios in the field are very costly and time consuming. Generally they are based on the collection of environmental water samples from boats that must be directly analyzed or pre-treated (e.g. filtered and fixed) by local institutions. Sampling is a real bottleneck in this process as it can only be conducted in relatively small areas in punctual locations. All these constraints strongly limit field studies.

To overcome these problems we recently applied the Continuous Plankton Recorder (CPR) technology to *Vibrio* research (Vezzulli *et al.*, 2012, Vezzulli *et al.*, 2016). The CPR survey is one of the longest running marine biological monitoring programmes in the world and is operated by the Sir Alister Hardy Foundation for Ocean Science (SAHFOS) in Plymouth (UK) (https://www.sahfos.ac.uk/). The CPR is a high-speed plankton sampler designed to be towed from ships of opportunity over long distances (Reid et al., 2003) and enables the plankton and associated *Vibrio* bacteria to be sampled over thousands of miles (Figure 2).

Sampling takes place in the ocean surface layer (~7 meters) and plankton is collected on a band of silk (mesh-size 270 μm) that moves across the sampling aperture at a rate proportional to the speed of the towing ship. The CPR mesh width of 270 μm retains larger zooplankton with a high efficiency, but also collects small planktonic organisms such as nauplii, microzooplankton and phytoplankton. Upon return to the laboratory, the silk is removed from the device and divided into individual samples that are numbered along the route. Only odd samples are analysed, according to standard procedures. Both analysed and unanalysed samples are stored in plastic boxes in buffered formalin (usually comprising 4–10% buffered formalin) in the CPR archive in Plymouth (UK). Each sample represents 10 nautical miles of tow ($3m^3$ of filtered seawater) and captures a substantial fraction of the plankton associated *Vibrio* community, providing a valuable tool for macroecological studies of these bacteria in the aquatic environment.

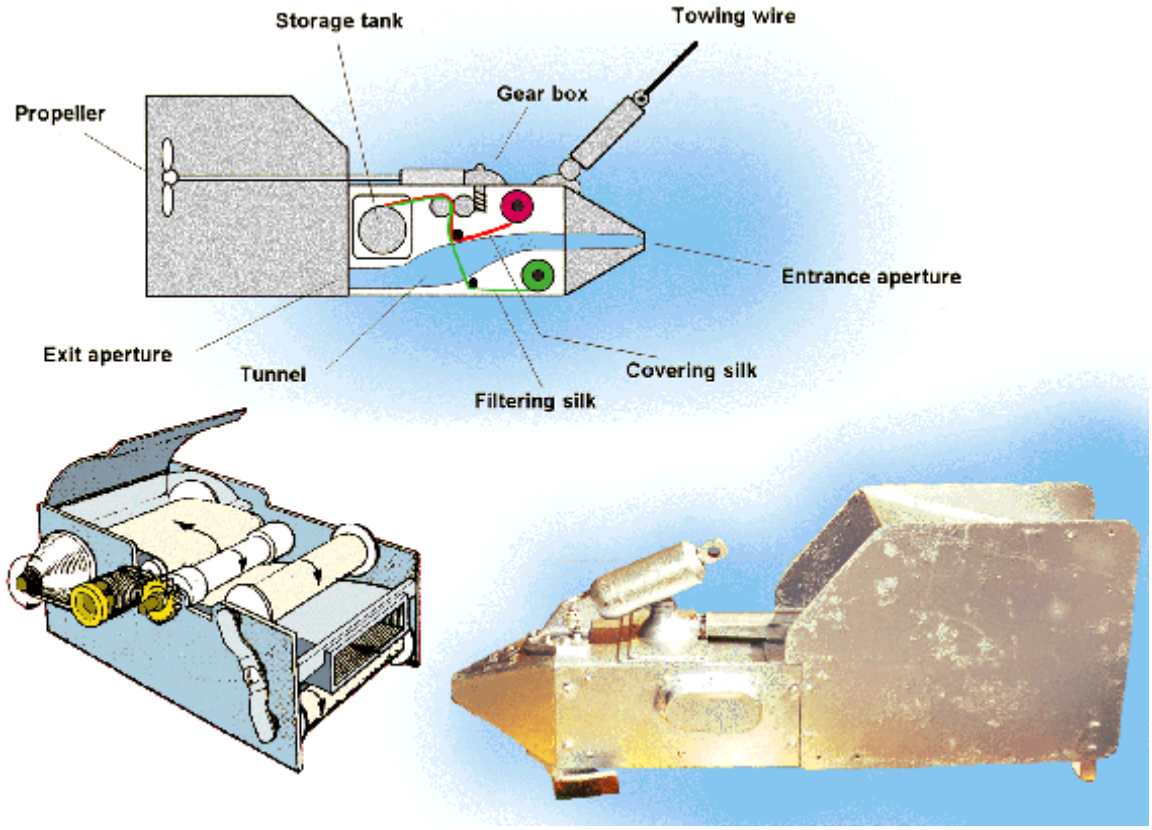


Figure 2. Continuous Plankton Recorder

Continuous Plankton Recorder technology applied to *Vibrio* research: long-term retrospective molecular studies of *Vibrio* populations

Archived CPR samples have been recently employed to retrospectively investigate long-term variations of *Vibrio* populations in the marine environment by applying molecular biology techniques. An extraction method was developed for the purification of DNA from historical formalin fixed CPR samples and an unbiased index of abundance for Vibrio quantification in the CPR samples, termed "Vibrio relative abundance index—VAI," was applied. This index measures the relative proportion of plankton-associated vibrios in comparison to the total number of associated bacterial cells. In particular, the ratio of *Vibrio spp*. cells to total bacterial cells is assessed by Real-Time PCR using genus-specific and universal primers, respectively, producing small amplicons of similar size (113 vs 98 bp) to avoid age- and formalin-induced bias. PCR protocol details for calculation of the index are given in Vezzulli *et al.* (2012).

VAI index was assessed on a large number of CPR samples collected in the temperate North Atlantic over the past half-century including the North Sea, western English Channel, Iberian coast, Iceland coast, Irish Sea, Newfoundland, Nova Scotia, and open North Atlantic (Vezzulli *et al.*, 2016) (Figure 3).

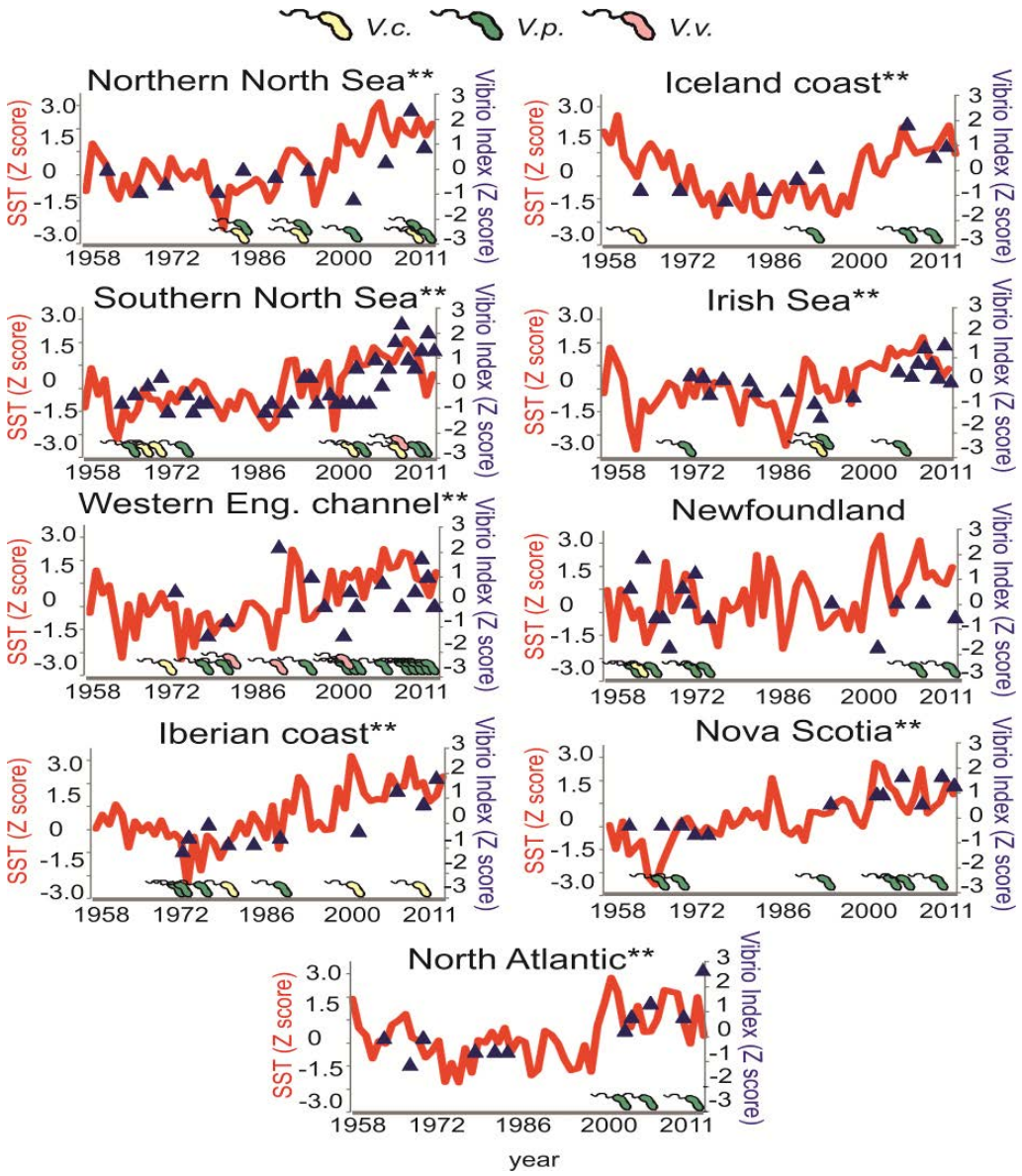


Figure 3.Multidecadal relationship between Vibrio prokaryote abundance and SST in the temperate North Atlantic and North Sea. Presence of the human pathogenic species *V. cholerae* (V.c.-yellow bacterial cell), *V. parahaemolyticus* (V.p.- green bacterial cell) and *V. vulnificus* (V.v-pink bacterial cell) is showed (from Vezzulli *et al.*, 2016).

The results showed that these bacteria have increased in prevalence in the last fifty years and that this increase is correlated significantly, during the same period, with warming sea surface temperature. Deep sequencing analysis applied on a subset of analyzed samples also provided evidence that vibrios, including the human pathogen *V. cholerae*, increased their dominance within the plankton associated bacterial community of coastal marine waters (Vezzulli *et al.*, 2012). Overall these findings provide support for the view that global warming is having a strong impact on the composition of marine bacterial communities with important implications for human and animal health into the future.

Continuous Plankton Recorder technology applied to *Vibrio* research: macroecology of *V. cholerae* in endemic areas

With a view to addressing some important questions regarding the ecology of *V. cholerae* in endemic areas such as the role of environmental factors in the origin, transmission and spreading of the cholera disease a species-specific qPCR assay for the diagnostic detection of this bacterium in CPR samples was recently developed (Vezzulli *et al.*, 2015). The method is based on the amplification of a small

amplicon of 206 bp of the gbpA gene of *V. cholerae* and was optimised for the analysis of formalin-fixed samples, such as historical CPR samples. In addition the method is highly specific for *V. cholerae* as it fails to amplify strains of closely-related *Vibrio* species including *V. mimicus* (Vezzulli *et al.*, 2015).

The protocol was tested on 18 samples collected by the southern African CPR Sister Survey in the Benguela Current Large Marine Ecosystem (BCLME) during its inaugural survey in September 2011 along the coasts of Angola, Namibia and South Africa which represent endemic areas for cholera (Figure 4).

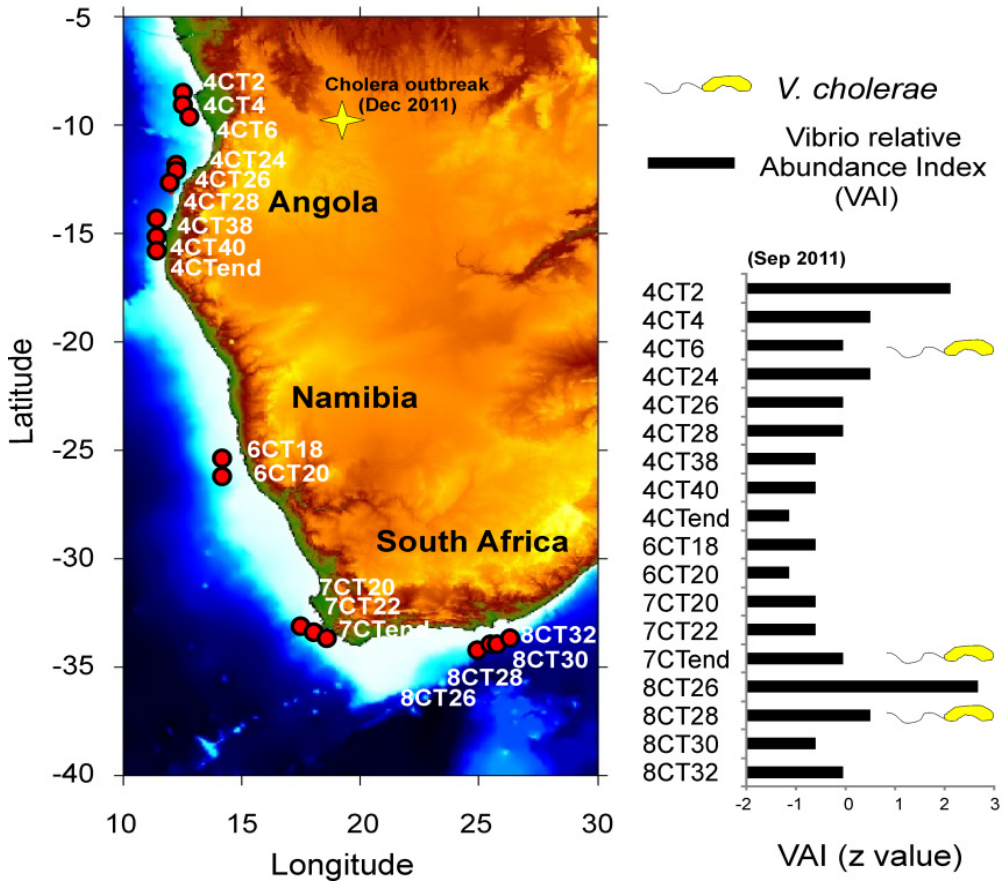


Figure 4. *V. cholerae* detection in CPR samples collected in cholera endemic areas of the Benguela Current Large Marine Ecosystem region (from Vezzulli *et al.*, 2015).

Three of the samples, notably collected in coastal waters off the cities of Luanda (northern Angola) and Cape Town and Port Elizabeth (South Africa), tested positive for *V. cholerae* (Vezzulli *et al.*, 2015). Incidentally, some two months after CPR sampling, numerous cases of the disease were reported in the district of Lucapa, about 800 km inland from Luanda. It is presently unknown whether the *V. cholerae* DNA recovered from these samples is from toxigenic genotypes. To investigate this further a whole-genome enrichment (WGE) and next-generation sequencing (NGS) approach was recently developed for direct genotyping and metagenomic analysis of low abundant *V. cholerae* DNA from complex environmental samples (Vezzulli et al., 2017). The protocol is based on the use of biotinylated RNA baits for target enrichment of *V. cholerae* metagenomic DNA via hybridization and was successfully applied on CPR samples (unpublished data).

These findings demonstrate the usefulness of the CPR in cholera studies with the potential to mark a significant breakthrough in the investigation of the ‘cholera paradigm’ (e.g. the role of environmental versus human factors in the origin of cholera outbreaks and epidemics) in cholera endemic areas.

Conclusions

CPR is a promising technology for the study of the macroecology of Vibrio species. Such a technology is of great potential interest considering the sampling coverage that can be achieved over large spatial (thousands of miles) and temporal (multidecadal period) scales. This holds true also at the local scale if we consider that one 10 nautical mile CPR sample represents multiple point samples, which is the approach traditionally adopted in acquiring environmental microbiological data. As such, using CPR samples is an improvement of the spatial sampling resolution by several orders of magnitude. Thanks to newly developed molecular techniques optimized for the analysis of *Vibrio* populations associated to formalin-fixed CPR samples (e.g. enabling detection, enumeration and typing of *Vibrio* species on CPR samples) a key element of the CPR methodology is its ability to obtain large amounts of information from remote areas at low sampling cost by using merchant ships of opportunity to tow the sampling machine for free. The costs involved, other than for purchase of equipment, are largely for subsequent laboratory analysis of the bacteria and plankton which can be undertaken thousands of miles away from areas of collection. Utility and performance of the method can be further improved by extracting DNA immediately after sampling and/or avoiding/improving fixation of the sample. Ultimately it is a procedure that could be possibly extended to other members of the picoplankton community or more in general, as long as new molecular biology and bioinformatic techniques will be available to contrast the bias of sample contamination, to the entire prokaryotic and viral communities of the ocean.

* this chapter is to be cited as :

Vezzulli L., Wilson W.H. and Pruzzo C. 2017. Investigating the macroecology of emerging Vibrio pathogens in the ocean using the Continuous Plankton Recorder technology. pp. 91 – 98 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

Diversity and distribution of potential human pathogenic bacteria in the seas: novel insights from exploration of NGS databases

Marc Troussellier[1], Jean-Christophe Auguet[1], Arthur Escalas[2]

[1]UMR 9190 MARBEC CNRS, Université de Montpellier, IRD Ifremer, France.
[2]UMR 7245 MCAM CNRS-MNHN, Cyanobactéries, Cyanotoxines et Environnement, Paris, France.

Abstract

Few studies have been dedicated to the diversity and distribution of potential human pathogenic bacteria (PHPB) in offshore waters. The first reason is that the probability of human exposure to PHPB decreases with their concentration in seawater, which decreases with the distance to the coast. Thus the risk associated to PHPB is a priori considered as negligible in the open sea. The second main reason is that the methodologies used to search and count the PHPB within the marine bacterial communities' exhibit several limitations such as a low resolutive power, the ability to screen a limited number of PHPB species at a time, and/or were focused on the culturable fraction. Recent developments in next generation sequencing (NGS) have changed in a drastic way our view of microbial community ecology, allowing the detection of the less abundant members of bacterial communities and revealing that in most, if not all, microbial communities there is a large proportion of rare bacterial species that cannot be detected with previous methodologies. Here, we first consider whether PHPB can be detected in existing NGS databases obtained from different marine ecosystems exposed to contrasted anthropogenic pressures. The associated hypothesis is that PHPB relative occurrence and/or relative abundance followed a decreasing gradient from coastal areas with heavy anthropogenic pressures to open ocean waters far from PHPB sources. We also explore the hypothesis that marine macroorganisms can be large reservoirs and vectors of PHPB.

Keywords: Potential human pathogen bacteria, marine ecosystems, NGS databases, marine macro-organisms.

1- Introduction

The relationship between humans and the sea deeply changed during the last century as increasing numbers of human communities developed in coastal areas. Today, most of the world's largest cities

are located in coastal areas (Brown *et al.*, 2013). Not all these coastal cities are able to efficiently treat the huge amount of human wastes that are discharged in the sea. These sewage discharges constitute the main source of potential human pathogenic bacteria (PHPB) in coastal areas (*e.g.* Munn, 2005), which may subsequently affect people who use the sea for recreational activities or as a source of food.

Coastal seawater contamination is very well documented through scientific studies and/or survey of the PHPB that can be discharged in sensible areas such as aquaculture or recreational areas (*e.g.* Belkin and Colwell, 2005). Data on PHPB in offshore waters are less common. One of the reasons is that the probability of human exposure to PHPB decreases with their concentration in seawater, which decreases with the distance to the coast. Dilution of sewage and "mortality" of PHPB in seawater were considered as the causes of their disappearance (*e.g.* Mattioli *et al.*, 2017). Thus the risk (i.e. exposure x hazard) associated to PHPB is a priori considered as negligible in the open sea. The second reason is that the methodologies used to search and count the PHPB within the marine bacterial communities were only able to detect the most abundant PHPB cells and/or a limited number of PHPB species and/or only their culturable fraction.

The recent development of next generation sequencing (NGS) technologies has changed in a drastic way our view of microbial community ecology, allowing the detection of the less abundant members of bacterial communities and revealing that in most, if not all, microbial communities there was a large proportion of rare bacterial species that cannot be detected with previous methodologies. During the last decade, NGS have been extensively used to explore the diversity of marine microbial communities in different ocean sectors, including the "rare biosphere" (Box 1).

In the seas, one can hypothesize that bacterial taxa from allochtonous origin such as potential human pathogenic bacteria (PHPB) may belong to the rare biosphere. Among the different questions related to the ecology of the rare biosphere, the most frequently addressed were related to their persistence and wide distribution in seawater (e.g. Pedros-Alios, 2012; Lynch and Neufeld, 2015; Skopina *et al.*, 2016). We recently proposed that marine macroorganisms might sustain the diversity of seawater bacterial community and especially of their rare components (Troussellier *et al.*, 2017).The existing and future NGS databases offer the possibility to (i) explore the diversity and relative abundance of PHPB in different marine waters, (ii) validate or invalidate the absence of PHPB in offshore waters, and (iii) test whether marine macroorganisms can be reservoirs and vectors of PHPB.

Here, we first consider whether PHPB can be detected in existing NGS databases obtained from different marine ecosystems under contrasted anthropogenic pressures. The associated hypothesis is that PHPB occurrence and/or relative abundance follow a decreasing gradient from coastal areas with heavy anthropogenic pressures to open ocean waters far from PHPB sources. Then, we explore whether the recently proposed view of the interactions between marine microbes and macroorganisms, namely sustaining the rares (Troussellier *et al.*, 2017), which suggests that macroorganisms favor the maintenance of marine microbial diversity, may also be applied to PHPB. In other words, we test the hypothesis that marine macroorganisms can be reservoirs and vectors of PHPB.

2- Do potential human pathogen bacteria belong to the rare biosphere?

From a methodological point of view, there are at least two ways to estimate the proportion of a bacterial community that corresponds to PHPB:

The first approach, and the most classical one, is to compare the abundance of pathogenic bacteria to the total abundance of bacteria. However, this requires (i) to know which pathogenic species we want to search, and (ii) to enumerate in a separate way the different pathogenic bacteria through either specific culture media or molecular tools.

A second option, more global and exhaustive, relies on the use of NGS data to estimate the proportion of sequences that can be attributed to all known potential pathogenic bacteria OTUs.

2-1 Looking for PHPB with NGS

Recently, different authors have used NGS to detect potential pathogens and/or indicator bacteria in aquatic ecosystem with different objectives.

McLellan *et al.*, (2014) used NGS to discern between human (*e.g.* sewage) and other animal pathogen sources. This is an important point because of the explicit health risk posed by human pathogens and because the ways to remediate sewage contamination is different from the strategy to avoid animal waste contamination carried in surface runoff. Ahmed *et al.* (2015) showed that bacterial community and host-associated molecular marker analyses can be combined to identify potential sources of fecal pollution in an urban river.

Another main objective is to estimate the proportion that pathogens and/or bacterial indicators can form in bacterial communities. In fact, not all the NGS-based studies succeeded in detecting pathogens or fecal indicator bacteria. One of the reasons can be the level of sequencing depth (*i.e.* the number of sequences) retained to perform NGS data analysis, which has to be high enough to detect microorganisms present with a low or a very low abundance.

For instance, Vierheilig *et al.* (2015) were among the first authors to propose the use of NGS platform for water quality assessment. One of the main questions they addressed was: "*What is the potential of NGS methods for the detection of fecal pollution in environmental waters*?". The test sample set comprised water, sediment, soil, and fecal samples from a backwater study area influenced by the river Danube, as well as fecal samples from various zoo animals. However, as reported by the authors, the « low » sampling depth used did not allow them to detect populations with low abundances such as fecal indicators in surface waters.

Luna *et al.* (2016) successfully applied NGS to detect a large number of OTUs affiliated with different fecal indicator bacteria. In the most polluted area which they studied (Po delta), the relative abundance of traditional indicators such as *Enterobacteriaceae* accounted for 0.01 to 0.19% of the bacterial assemblage, while the genus *Enterococcus* represented between 0 (no sequences detected) to 0.01% of the community. The total number of OTUs belonging to these traditional fecal indicators recorded across all the stations of the Po area was 46 for *Enterobacteriaceae* and 4 for *Enterococcus.* Unfortunately, the authors did not report the sampling depth.

Brinkmeyer (2016) addressed explicitly the use of NGS to detect and estimate the frequency of potential pathogens in the ballast water of five ships. With a sampling depth of more than 48,000 sequences per sample, she was able to identify 60 different potential pathogens, 47 of which were human pathogens (78%), 8 from fishes (13%) and 5 from terrestrial plants (< 1%). The data provided also allowed estimating the relative frequency of each potential pathogen in the bacterial community. All the potential pathogens detected in each of the five ballast seawater exhibited relative frequency

lower than 1%. Taking the 0.1% value as a threshold for defining rare bacteria, a large proportion of detected potential pathogens (85-97%) can be considered as part of the rare biosphere (Figure 1).

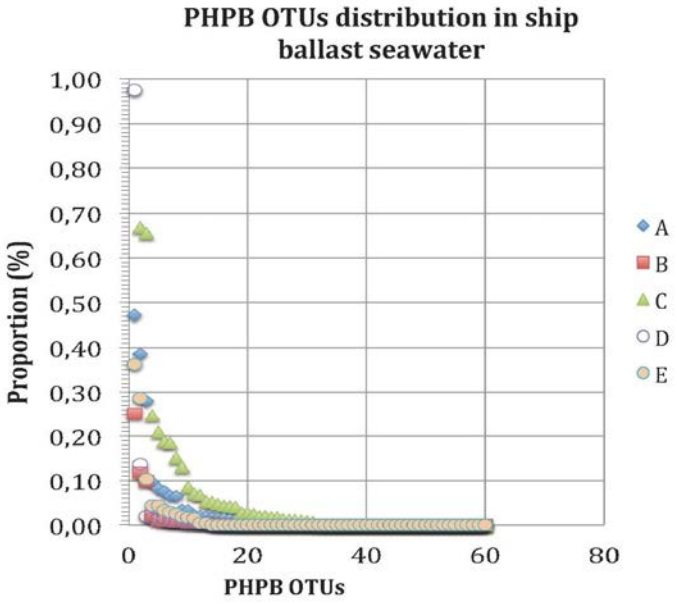


Figure 1. PHPB OTUs distribution in ballast seawater of five different ships (A-E) (From Brinkmeyer, 2016).

2-2 The opportunities offered by existing NGS databases

Many other NGS-based studies, while not explicitly dedicated to the study of PHPB, produced databases from where it is possible to extract interesting information on PHPB.

As a preliminary step to be able to extract specific information on the relative abundances of pathogenic bacteria, one needs to obtain an exhaustive list of these bacteria. Such a reference database has been published recently by Wardeh *et al.* (2015) who provided a list of hosts (humans, domestic animals, wild mammals, plants…) along with all the microbes associated with them through various types microbial-macrobial interactions, from pathogenic, mutualistic or commensal interactions.

With such a reference list in hand, we can begin to explore different kinds of NGS-based marine databases to identify the presence of PHPB in a systematic way. Here, we explored two types of marine environments, under contrasted anthropogenic pressure and where the frequencies of PHPB are expected to differ.

2-2-1 Coastal marine ecosystems exposed to anthropogenic pressure

Aravindraja *et al.* (2013) studied the diversity of bacterial communities associated with different coastal habitats (see Table 1). Samples were collected from Karandaku coastal area (India), which is impacted by multiple sewage discharges (Thinesh *et al.*, 2011; Sneha *et al.*, 2016). Their data allow the investigation of PHPB distribution in different coastal habitats.

Thanks to the sampling depth (137050 - 266103 sequences), these data show that PHPB OTUs may represent a large proportion of the total OTU richness (14-18%) and of the total number of bacterial

sequences (up to 34% of the community in the sediment sample) exposed to large anthropogenic pressure and show that investigating compartments of coastal waters could be promising.

Table 1. Richness of PHPB observed in different habitats of the Karandaku coastal area (adapted from Aravindraja *et al.*, 2013).

Samples	Total sequences	Total OTU richness	Total pathogen species sequences	Pathogen species richness	Proportion of pathogen sequences (%)	Proportion of pathogen OTUs (%)
seawater (S3)	137050	1263	6973	205	5.09	16.23
sediment (S1)	199439	1337	67649	234	33.92	17.50
rhizophere sediment (S2)	266103	1524	17259	215	6.49	14.11
seaweed (S4)	173150	1329	17848	252	10,31	18.96
seagrass (S5)	223953	1512	17233	242	7.69	16.01

Thus, as expected, in coastal areas exposed to large anthropogenic pressure, PHPB OTUs cannot be considered as members of the rare biosphere. Instead, they can represent a large proportion of bacterial community richness and abundance.

2-2-2 Open ocean ecosystems

A contrario, one may ask whether pathogenic bacteria can also be detected through NGS in offshore areas, which are *a priori* less affected by anthropogenic pressures.

This question could be usefully tackled by mining large NGS databases (e.g. ICoMM, SRA, MG-RAST), which contain many samples from open ocean sampling campaigns. The genomic data collected in global surveys such as the TARA expeditions or the Ocean Sampling Day (see Bowler, this volume) are particularly interesting in this regard.

Here, we analyzed miTAG 16S data from the metagenome of 139 samples referenced in the public *in-silico* database of the Tara Ocean consortium (Sunagawa *et al.*, 2015). Samples and environmental data were collected in 68 stations divided within mesopelagic and epipelagic waters explored by Tara Ocean expeditions, at four distinct depths: surface water layer (SRF), deep chlorophyll maximum layer (DCM), mesopelagic zone (MES) and subsurface epipelagic mixed layer (MIX). TARA sequences were screened using the list of potential human pathogens (n = 874) that can be extracted from the Wardeh *et al.* database (2015).

At least one human pathogen was detected in every one of the TARA samples we screened (n = 139). The richness of PHPB in these samples ranges from 2 to 31 OTUs. Overall, 10% of the pathogens from the reference list were observed (n = 87). The relative abundance of PHPB sequences ranged from 1 10^{-4} % to 8 10^{-3} %, which clearly suggests that on a global scale, PHPB belong to the rare biosphere in open ocean bacterial communities.

We also explored the TARA database to search for the most widespread PHPB OTUs, which belong to the core bacterial community (Figure 2). Eighteen PHPB taxa can be detected on the basis of these criteria in the whole TARA database (139 samples). Some of these taxa can be detected in > 50% of the TARA samples (*Acinetobacter johnsonii*, *Pseudomonas stutzeri*) and most of them (75%) occurred in more than 10% of TARA samples.

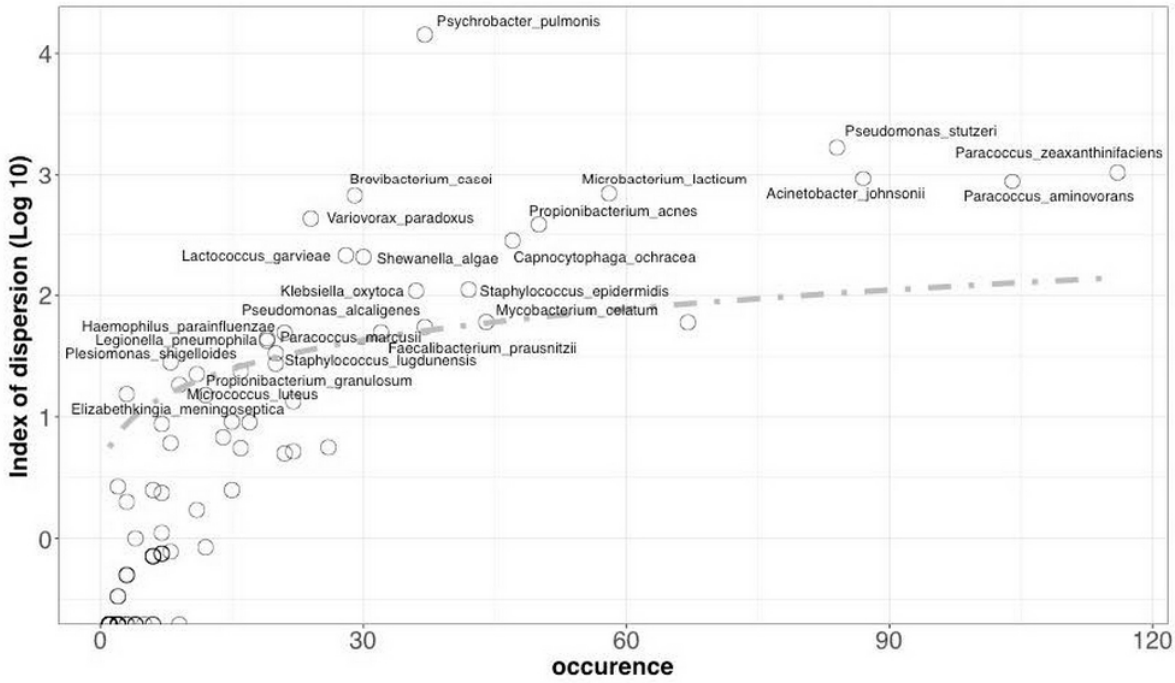


Figure 2. Occurrence of non-randomly distributed PHPB OTUs in the TARA database plotted against its dispersion index. The index of dispersion for each OTU was calculated as the ratio of the variance to the mean abundance multiplied by the occurrence. The line depicts the 2.5% confidence limit of the Chi^2 distribution: lineages falling below this line follow a Poisson distribution and are randomly dispersed in space.

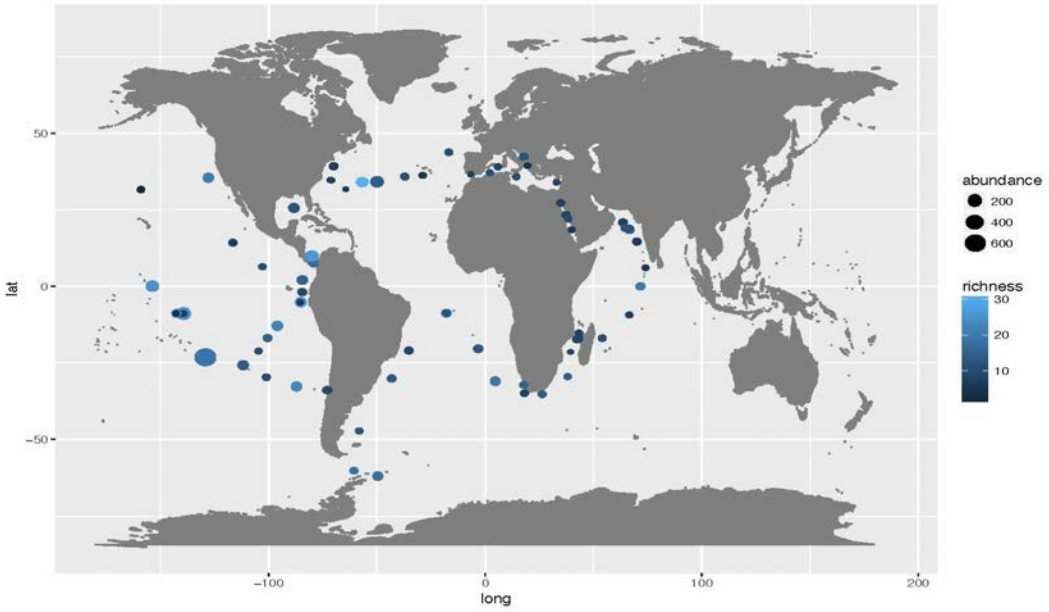


Figure 3. Geographic distribution of richness and abundance of sequences of the PHPB OTUs detected in the TARA database.

Figure 3 illustrates the geographical distribution of the richness and frequency of PHPB as observed through TARA sampling.

We did not observe any large-scale pattern, such as latitudinal or longitudinal patterns. However, when considering the correlations among PHPB relative abundances, at least two different patterns can be observed in the distribution of the OTUs (Figure 4).

A majority of the detected PHPB did not show any correlation of their spatial relative abundances with those of other species (*Elizabethkingia meningoseptica*, *Plesiomonas shigelloides*, *Acinetobacter johnsoni*, *Capnocytophaga ochracea*). Some of these species were already frequently detected/isolated in seawater or marine animals (*Acinetobacter johnsoni :* Venkateswaran *et al.*, 1991; Rohwer *et al.*, 2002; McInnes *et al.*, 2002; Farto *et al.*, 2006; Schulze *et al.*, 2006; Santiago-Vazquez *et al.*, 2007; Oh *et al.*, 2009; Kobayashi *et al.*, 2012; Benhamed *et al.*, 2014; De Santi *et al.*, 2016; Lee and Eom, 2016) or as contaminant originating from human fecal pollution in filter-feeders (*Plesiomonas shigelloides,* Herrera *et al.*, 2006) or, to our knowledge, never described in marine samples (*Elizabethkingia meningoseptica, Capnocytophaga ochracea).*

The most striking feature of this correlation matrix (Figure 4) was the existence of a cluster of species that exhibited a similar geographical distribution. The species that form this cluster are listed in Table 2.

Table 2. PHPB species showing significant and positive correlations of their abundance in the TARA database.

Species	**Occurrence (number of samples where the species was detected)**	**Relative occurrence (% of samples where the species was detected)**	**Pathology nature in humans**
Propionibacterium_acnes	50	36	acne, endocarditis, endopthalmitis, osteomyelitis
Staphylococcus_epidermidis	42	30	nosocomial infections
Brevibacterium_casei	29	21	pericarditis, brain infection, bacteremia
Staphylococcus_lugdunensis	20	15	endocarditis, osteomyelitis, abces
Legionella_pneumophila	19	14	pulmonary damage

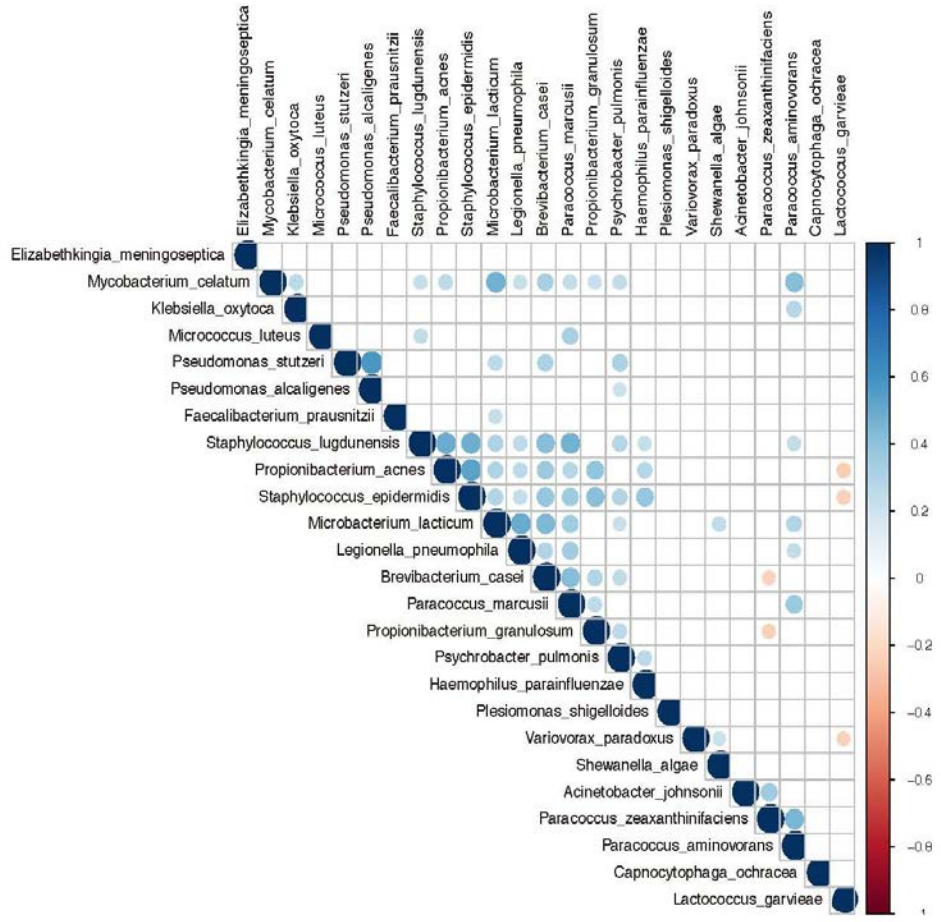


Figure 4. Non parametric correlation matrix among the abundances of the different PHPB OTUs detected in the TARA database. The scale for correlation intensity and sign is given on the axis part of the figure. Species are clustered following their correlation level (Spearman ρ). Only significant correlations ($P \leq 0;01\%$) are shown.

The positive, significant correlation of the spatial abundances of these species may lead to hypothesize that they could have the same origin and/or the same behavior in the sea.

It is interesting to note that the most frequently detected species (*Propionibacterium acnes*) is often observed in coastal seawater bacterial communities or associated with marine animals, while, to our knowledge, the second most occurring species (*Staphylococcus epidermidis*) had not been reported from marine samples. *Legionella pneumophila*, the species with the lowest occurrence is known to be able to survive in the marine environment (Heller *et al.*, 1998).

Only one species showed a negative correlation with other PHPBs: *Lactococcus garvieae* (versus *Propionibacterium acne, Staphylococcus epidermidis*), which is considered as an unusual human pathogen but an important one in aquaculture populations (Wang *et al.*, 2007). One explanation may

be that aquaculture locations generally take place far from the sewage discharges of large coastal cities.

3- Which anthropogenic and environmental factors can act on the PHPB distribution and survival in the seas?

The dispersal of free-living microbes across the ocean can result from different processes acting at different scales. Because of their very small size, microbial cells cannot disperse over large scale by their own means and they are passively transported from one place to another within water masses. Hence, as it is the case for other marine organisms, the dispersal capacity of bacteria first depends on hydrodynamic conditions (Mincer *et al.,* 2007; Hellweger *et al.,* 2014).

Other environmental or anthropogenic processes have been identified as factors that can act to transport PHPB far from their original source:

- Aeolian particle transport: PHPB have been detected on dust (*e.g.* Rosselli *et al.,* 2015). Dust or other particles from different continental areas can be transported over large distances (*e.g.* Kellog *et al.,* 2006).

- Growing international maritime traffic: Both volume and density of maritime traffic have shown a large increase during the last decades (UNCTAD, 2016; Halpern *et al.,* 2008). Cruise ships generate a number of waste streams, including sewage, blackwater and graywater and large amounts of associated PHPB (*e.g.* Copeland, 2008). Sewage can be discharged into the seawater only after it is treated and the distance of the ship is 4 nautical miles from the nearest land. But if the sewage is not treated it can be discharged 12 nautical miles away from the nearest land.

- Microplastics: Bacteria such as potential human pathogen *Vibrio spp*. (Zettler *et al.,* 2013; Kirstein *et al.,* 2016; Foulon *et al.,* 2016) and other Vibrionaceae (De Tender *et al.,* 2015) have been found colonizing microplastics particles in the marine environment which can travel long distances and form large patches in different oceanic provinces (*e.g.* Dohan and Maximenko, 2010).

Whatever the dispersion process, the large geographic distribution of PHPB raises the question of their persistence in marine waters. As for rare marine OTUs, being or becoming rare is a way to escape grazing and/or viral lysis (*e.g.* Yooseph *et al.,* 2010). In planktonic marine ecosystems many OTUs, especially the copiotroph ones, can use dormancy as a way to resist, at least temporarily, to adverse conditions such as nutrient deprivation (Lennon and Jones, 2011). More recently, we have shown that marine macroorganisms may have a very important role in sustaining the rare seawater bacterial OTUs (Troussellier *et al.,* 2017).

4- The potential role of marine macroorganisms as reservoir and dispersion vector of PHPB

For pathogens which are copiotrophs, marine animals may represent a favorable environment where to survive and, in the case of motile animals, a free and efficient dissemination vector. We propose that (i) host-microbes interactions benefit certain microbial populations, that are part of the rare biosphere (*i.e.,* opportunistic copiotrophic microorganisms) and that (ii) macroorganisms, especially the motile ones, may have a major role in the dispersal and geographic distribution of microbes.

4-1 Marine macroorganisms as reservoir of PHPB

Association between large plankton and bacteria, particularly PHPBs, is a well-documented story. Maugeri *et al.* (2004) showed higher occurrences of different culturable pathogens associated to large

plankton than in seawater. In the same study, *Vibrio vulnificus* was detected in 80% of the samples taken along the coastal area of the Strait of Messina using PCR based assay while culture-based method provided a lower number of positive detection. In the same coastal area *Helicobacter pylori* was only detected in seawater or plankton samples by PCR assay, but not detected by cultural techniques (Carbone *et al.*, 2005).

Species from the genus *Vibrio* were also observed in the microbiota of different macroorganisms. For instance, in Mediterranean waters, Gdoura *et al.* (2016) used real-time PCR-based assay to show that *Vibrio spp*. involved in human foodborne pathology in Tunisia were present in the vast majority of seawater and sediment samples as well as in different fish or clams species collected in Tunisian coastal areas.

Bogomolni *et al.* (2008) reported a large number of isolates recognized as human pathogens or potential human pathogens in the feces of marine mammals, sea birds and sharks (*Acinetobacter calcoaceticus-baumannii, Citrobacter braaki, C. freundii, Enterobacter cloacae, Leclercia adecarboxylata, Morganella morganii, Pseudomonas aeruginosa, Pseudomonas spp., Shewanella spp. and Stenotrophomonas maltophilia*).

Compared to seawater or sediment, higher marine organisms may also (i) constitute a favorable environment for the survival of potential bacterial pathogens (*e.g. V. parahaemolyticus* in clams, Karunasagar *et al.*, 1987), and (ii) act as a reservoir and vector of pathogens for other animals. For example marine fireworms (Sussman *et al.*, 2003), snails (Gignoux-Wolfsohn *et al.*, 2012) zooplankton (Certner *et al.*, 2017) may act as reservoir for coral disease-causing pathogens.

4-1-1 Occurrence of PHPB in corals versus seawater

We have performed a preliminary exploration of NGS databases (VAMPS: https://vamps.mbl.edu/) reporting on the composition of both bacterial communities associated with some marine macroorganisms and those of surrounding seawater.

Most available data were obtained for corals. Seven files of bacterial sequences associated to corals were checked for PHPB affiliated OTUs using the same reference database provided by Wardeh *et al. (*2015).

110 and 36 PHPB species were detected in coral bacterial communities (CBC) and in surrounding seawater bacterial communities (SWBC), respectively.

While no PHPB species were detected in all the CBC, some of them were more frequently observed. In table 3, we report the most occurring species in the CBC ($n \geq 4$, *i.e.* detected in more than 50% of CBC samples) and their occurrence in the surrounding SWBC. The profile of these PHPB species is given in box 2.

Coxiella burnetii, *Escherichia coli* and *Micrococcus luteus* were detected more frequently in CBC than in SWBC, while *Photobacterium damselae* and *Vibrio parahaemolyticus* were detected with the same frequency. *E. coli* and *M. luteus* have been reported to be capable of entering the dormant state (Box 2). While, the characteristics of *C. burnetii* cells indicate a high probability of survival in marine animals (Box 2), this is, to our best knowledge, the first time that this species is reported in coral samples. On the opposite, due to its intracellular replication mode, it appears logical that this bacteria had a low occurrence (Table 3) in SWBC.

The two recorded Vibrionaceae species (*P. damselae* and *V. parahaemolyticus*) can be considered as marine autochtonous bacteria, with the ability to support nutrient deprivation and/or low temperature and to regrowth when facing more favorable environmental conditions (Box 2).

Table 3. Most occurring PHPB species in CBC samples (n = 7).

Species	**CBC occurrence (number of CBC where the species was detected)**	**SWBC occurrence (number of SWBC where the species was detected)**	**Pathogenic characteristics**
Coxiella burnetii	4	1	Q fever,
Escherichia coli	5	2	Shiga toxin-producing *Escherichia coli* (STEC): hemorrhagic colitis, intestinal disease, cramps, abdominal pain, low grade fever
Micrococcus luteus	4	2	Endocarditis, bacteremia
Photobacterium damselae	5	5	Fish pasteurellosis, necrotizing fasciitis
Vibrio parahaemolyticus	5	4	Gastroenteritis, wound infection

These data reinforce the role of marine animals as a support of seawater bacterial community diversity, especially regarding their rare fraction (Troussellier *et al.*, 2017) including PHPB.

4-1-2 Relative abundance distributions of PHPB in marine animals *vs.* seawater

A step further we have compared the relative abundance distributions of PHPB in different marine animals and in surrounding seawater (MGRAST: http://metagenomics.anl.gov).

We report here two examples of the observed distributions of PHPB in coral and sponge bacterial communities and their surrounding seawater (Figure 5).

These two examples highlight that many PHPB taxa of SWBC increase in relative abundances in the animals (Figure 5A) including some PHPB taxa that were not detected in SWBC (Figure 5B).

4-2 Marine macroorganisms as a free-transport way for PHPB

Several species of marine macroorganisms are known to represent local vectors of pathogenic bacteria for surrounding species. In addition, Bogolmoni *et al.* (2008) suggested that marine mammals and birds in the Northwest Atlantic are reservoirs of potentially zoonotic pathogens, which they may transmit to beach visitors, fishermen and wildlife health personnel.

Recently, we estimated the spreading distance (SD) of ingested bacteria in different groups of marine macroorganisms (Figure 6). For the 16 fish species considered in this study, SD ranges between 2 and 190 km, corresponding to a dispersal potential from 200 to 200,000 times greater than the one microbial cells can reach by themselves (considering microbes swim between 15 and 100 $\mu m.s^{-1}$

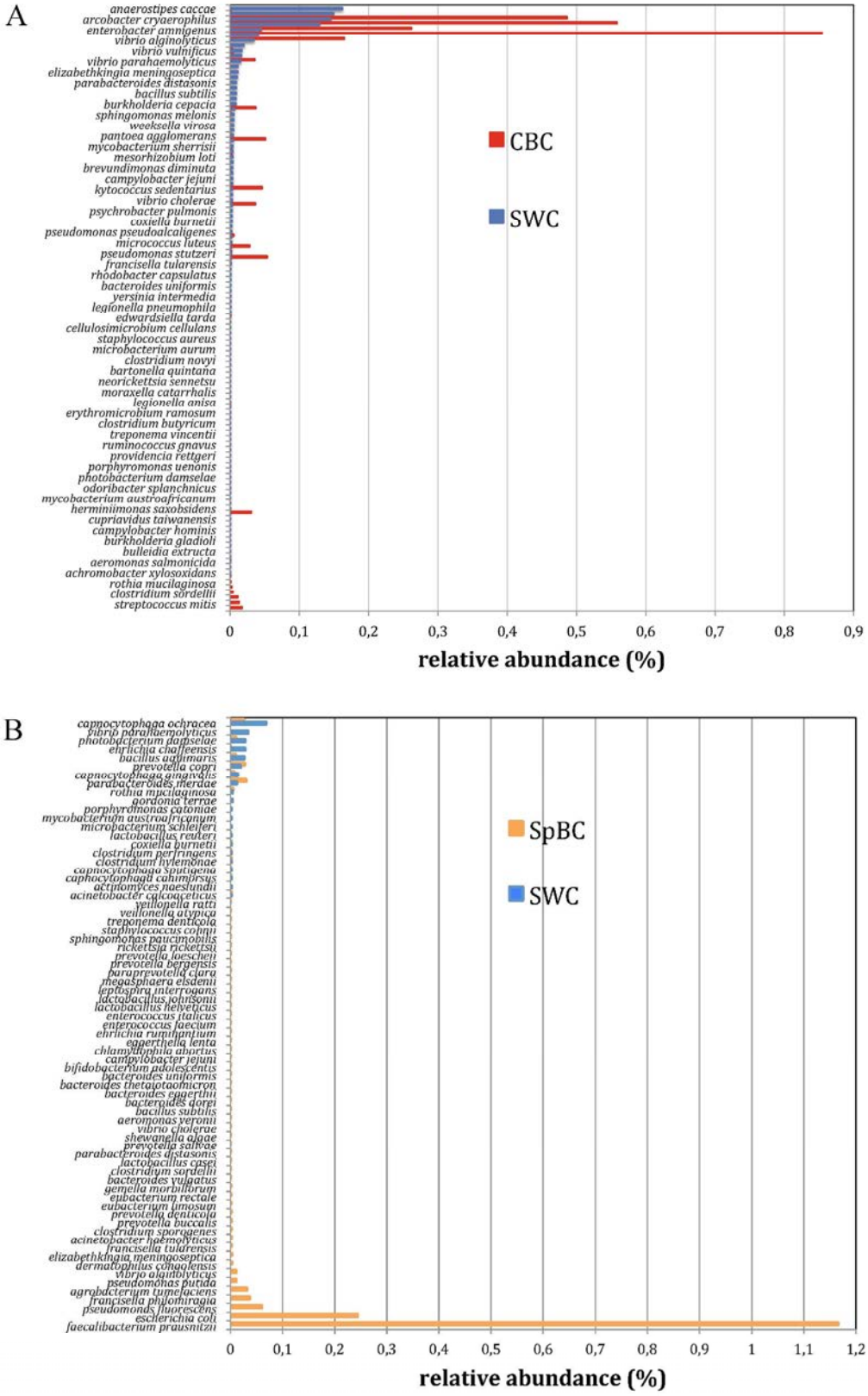


Figure 5. Abundance distribution of PHPB in the microbiota of two different benthic species and their surrounding seawater (SWBC). A: Coral bacterial community (CBC); B: Sponge bacterial community (SpBC).

(Wadhams and Armitage, 2004). Furthermore, mesopelagic fishes perform daily vertical migrations between the epipelagic layer where they forage and deeper layers where they digest (Radchenko, 2007). Such back and forth movements across habitats (pelagic-benthic, surface-deep layer), may allow marine microbes to cross chemo- or thermoclines which normally represent geographic dispersal barriers (Schaus and Vanni, 2000; Brenner and Krumme, 2007). Besides, fish predators may convey microbial communities over larger distances through predation and transfer them across food webs. For instance, marine mammals preying on fishes exhibit SSS equivalent or higher than the fastest fishes (*i.e.* 72 to 206 km.day^{-1}; (Watanabe *et al.*, 2011). Although the average GTT of most marine mammals is relatively short (2 hours to 2-3 days), considering their large intestine length (Carter *et al.*, 1999) their SD can be higher than 400 km. Thus, marine mammals exhibit a very high potential for the transport of their own gut microbes but also those of their prey.

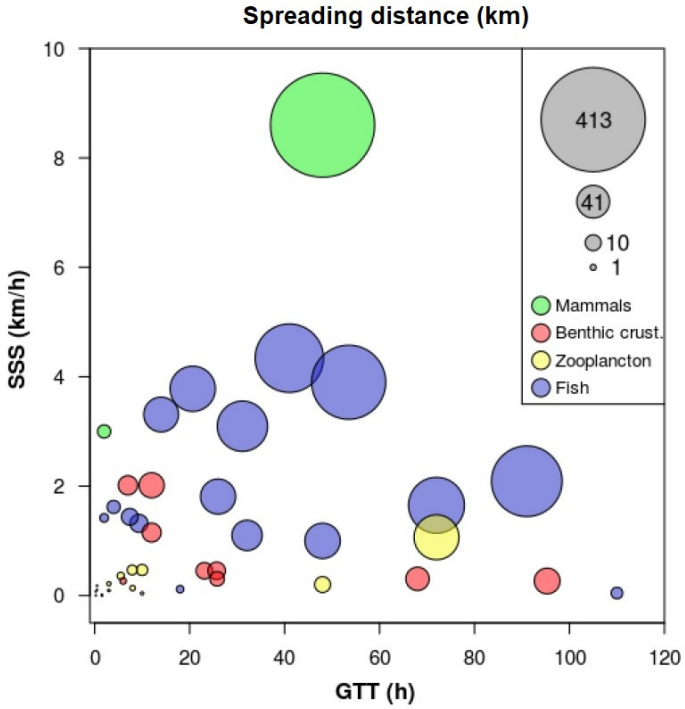


Figure 6. Spreading distance in different groups of marine macroorganisms. Spreading distance (SD in km) of bacterial populations is estimated as the product of gut transit time (GTT in h) and sustainable swimming speed (SSS in km h-1). The surface of the bubbles is proportional to SD. From Troussellier *et al.* (2017).

In conclusion, pelagic microbes that are ingested and travel inside macroorganisms' guts are "extracted" from the seawater matrix where viscosity strongly constrains their mobility and dispersal. Whatever the horizontal and vertical dispersion potential of motile macroorganisms, even a quick transit through their guts allows planktonic microbes to make "giant steps" compared to their own capacity. This process has the potential to seed local communities with organisms from another place, thus increasing the spatial distribution of rare microbes, including PHPB. Considering their dispersal potential and their gut transit time, macroorganisms are more likely to be involved in spatial dynamics at local (e.g. between coral reefs or habitats within a lagoon) or meso scales (e.g. between coastal lagoons or islands) than in large scale dispersal across water masses which depend more on sea currents (Amend *et al.*, 2013; Giovannoni and Nemergut, 2014).

5- How to improve our knowledge on PHPB distribution and associated risks?

This brief review of human pathogen occurrence in NGS database of marine bacterial communities brings us to raise several open questions:

- In depth exploration of existing metadata banks to detect potential pathogens. Can we imagine that each marine NGS database can be explored in an automatic way to detect (and monitor) potential pathogenic OTUs in marine ecosystems?

- Identification of marine environmental niches of different kinds of potential pathogens. Is it possible to define the PHPB environmental niches based on the ecological characteristics of the environments from where they are detected?

- Standardized high-throughput tools to monitor pathogens and/or genes involved in pathogenicity (DNAChip). Can we improve the existing DNAChip in order to detect in a standardized way the most frequently detected pathogens in the seas on the base of the list of PHPB observed in NGS marine databases?

- An old question without a satisfactory answer: from genes to infectivity and from dormant state to revival. How to detect in a simultaneous way the identity of a bacteria and the functionality of specific genes? Based on the list of detected PHPB in marine NGS database, do we need to explore both the ability to enter in dormancy and revival mechanisms of some (*e.g.* dominant) PHPB not studied before?

- How marine animals allow associated PHPB to increase their abundances? Are certain marine macroorganism species more favorable than others to PHPB?

- Risk associated to pathogens from the sea: the product of hazard and exposure? Could the relative abundance of PHPB in NGS database be used as a minimum proxy of a sanitary risk?

Box 1 The rare biosphere

Within the last decade, the increasing sequencing depth reached by Next Generation platforms has shed some light on the community assembly of microbes, unveiling a new compartment of these assemblages composed by very low abundant taxa or OTUs: the rare biosphere (Sogin *et al.,* 2006).

Rarity can be defined in several ways, including for instance local abundance, habitat specificity and geographical spread (Rabinowitz, 1981). Such concepts are routinely applied to plants and animals and can be easily transferred to microbial communities. Local abundance is the easiest and the most common index used to quantify species' rarity in microbial ecology. However, microbial rarity may also be expressed as a restriction to a low number of habitats, thereby reflecting habitat specificity (Barberán *et al.,* 2014) and geographic range (Tedersoo *et al.,* 2014).

Within the marine realm, rare OTUs are the rule not the exception. Since the pioneer study of Sogin *et al.,* (2006), an increasing number of studies have shown that the rare biosphere constitutes most of the microbial diversity over large spatial and temporal scales (Szabo *et al.,* 2007; Elshahed *et al.,* 2008; Youssef and Elshahed, 2009; Vergin *et al.,* 2013) . This rarity feature of microbial communities has been reported in most, if not all, marine systems (Youssef *et al.,* 2010; Campbell *et al.,* 2011; Lynch and Neufeld, 2015) as observed for other communities of microbial (*e.g.* in phytoplankton; Campbell *et al.,* 2011; Hugoni *et al.,* 2013; Lynch and Neufeld, 2015) or macrobial organims (*e.g.* fish, plants or trees; Mouillot *et al.,* 2013; ter Steege *et al.,*2013). Thus, rare microbes, understood here as those that have low abundance within communities, are more frequent than previously thought (Amend *et al.,* 2013).

Box 2 Profiles of PHPB occurring in coral bacterial communities from the VAMPS NGS database

Coxiella burnetii is the cause of Q fever, a zoonotic disease (e.g. Eldin *et al.,* 2017). Humans are exposed to *C. burnetii* via the waste products of domesticated terrestrial animals (cattle, sheep, goats…). It has recently been identified in several marine mammals species (e.g., Tryland *et al.,* 2014), in seabird ticks (e.g. Wilkinson *et al.,* 2014), probably in harmful microalga bacterial consortium (Maki *et al.,* 2004) and marine sediments (Urakawa *et al.,* 1999). It is an intracellular pathogen, replicating in eukaryotic cells. Sodium-proton exchangers and transporters for osmoprotectants are found in the *C. burnetii* genome, allowing this bacterium to confront osmotic and oxidative stresses.

Escherichia coli includes not only commensal but also pathogenic strains that cause a variety of human diseases—resulting in more than 2 million deaths each year (Jang *et al.,* 2017). Based on studies done in recent decades, the presence of environmental *Escherichia* is now well recognized. These environmental *E. coli* may be of animal-origin and have become adapted to their surrounding environments; or they may retain the characteristics of their ancestral linage, which originated from soil and sediment habitats (Jang *et al.,* 2017). In coastal marine areas, *E. coli* is used as a fecal contamination indicator, and is frequently detected in waters, biofilms, sediment and bivalves, especially in areas under the influence of sewage and/or runoff from land. Moreover, *E. coli* was commonly isolated in live and stranded birds or marine mammals (Bogomolni *et al.,* 2008). Marine animal pathologies caused by *E. coli* were already reported (Kim *et al.,* 2002). The loss of culturability of *E. coli* in seawater is well known, especially in surface waters where cells can be exposed to high level of solar radiations, but its viability and integrity can be retained for a long time period (Orruno *et al.,* 2017).

Micrococcus luteus appears more as an opportunistic as a systematic human pathogen. It has been isolated from marine fishes (e.g. Al Bulushi *et al.,* 2010), sponges (e.g. Bultel-Poncé *et al.,* 1998), biofilms (Kwon *et al.,* 2002). It has been described as salt-tolerant and able to enter in a reversible long-term dormant state (Greenblatt *et al.,* 2004, Keep *et al.,* 2006, Young *et al.,* 2010).

Photobacterium damselae was considered as a pathogen for a large number of marine animal species (wild or farmed) and as a causative agent of opportunistic infections in humans (e.g. Rivas *et al.*, 2013). It is considered as the pathogenic bacteria most frequently involved in fish diseases (Labella *et al.*, 2011) and as an emerging pathogen with increasing geographical distribution (Terceti *et al.*, 2016). Fish-virulent strains of *P. damselae* subsp. *damselae* are able to survive as culturable bacteria in seawater microcosms under starvation conditions, at 14 and 22 °C, for prolonged periods of time (Fouz *et al.*, 1998). The bacterium can shift between ambient seawater and hosts, through outer membrane proteins (OMPs) rapidly responding to salt concentration (Wu *et al.*, 2006).

Vibrio parahaemolyticus is highly successful in most marine systems. It is an important pathogen of animals ranging from microcrustaceans to humans and is a causative agent of seafood-associated poisoning (e.g. Ghenem *et al.*, 2017). The spread of vibrios may be linked to the increase in sea surface temperature (Vezzulli *et al.*, 2016). It is known to enter into the viable but nonculturable (VBNC) state as a response to injurious environmental conditions, such as low temperature and nutrient deprivation, and to regain their culturability several days after starvation (Mizunoe *et al.*, 2000).

* this chapter is to be cited as :

Troussellier M., Auguet J.C. and Escalas A. 2017. Diversity and distribution of potential human pathogenic bacteria in the seas: novel insights from exploration of NGS databases. pp. 99 – 114 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

Next generation sequencing-based approaches to characterize microbial pathogenic community and their potential relation to the Black Sea ecosystem status

Elena Stoica[1], Mariia Pavlovska[2], Evgen Dykyi[2], Konstantinos Kormas[3]

[1]*Nat. Inst. for Marine Research and Development "Grigore Antipa" (NIMRD), Constanta, Romania*
[2]*Ukrainian scientific center of Ecology of Sea (UkrSCES), Odesa, Ukraine*
[3]*Department of Ichthyology & Aquatic Environment, Faculty of Agricultural Sciences University of Thessaly, Greece*

Abstract

The Black Sea harbors unique microbiotas shaped by its specific hydrographic and hydro-chemical conditions. However the observational data on specific pathogenic bacteria are scarce. The monitoring programs in the Black Sea are mostly designed to determine the level of faecal bacteriological indicators in the recreational waters based on cultivation methods. This is the first study to assess the presence and diversity of the pathogenic bacterial community in the inshore and offshore Black Sea waters by means of an approach mining DNA next-generation-sequencing dataset. The present investigation reveals an unprecedented diversity of pathogenic bacteria dominated by *Actinobacteria, alpha-proteobacteria* and *gamma-proteobacteria*, which varied in relative abundance between species and regions. The bacterial OTUs obtained by NGS Illumina sequencing showed that high-throughput tools are needed to decipher the pathogen bacterial diversity of the Black Sea.

Keywords: *microbial community, marine pathogens, high-throughput sequencing, data mining, Black Sea*

Introduction

Marine microbial pathogenic community represents a mix of indigenous-pathogenic species and those externally introduced from contaminated sewage outfalls, nonpoint sources pollution, and river discharges (Stewart *et al*., 2008). Most pathogenic microbes are found in coastal marine habitats, both

as free-floating through the seawater column (planktonic) and attached to different surfaces (marine animals, phytoplankton, zooplankton, sediments and detritus). They can cause a broad spectrum of human and marine animal infectious diseases and thus successively affect marine ecosystem functioning and health status (Brettar *et al*., 2007; Lu *et al*., 2015). However, microbial contamination in coastal marine environments is a worldwide phenomenon not yet well understood due mostly to traditional culture and isolate methods used for pathogens detection.

The emerging field of metagenomics, the culture-independent sequence-based approach, has the potential to sidestep the major difficulties associated with culture-based methods (Thomas *et al*., 2012). The recent advent of the next-generation sequencing (NGS) technologies has revolutionized aquatic pathogens research, allowing simultaneous fast detection of all microorganisms in an environmental sample, both known and novel pathogenic microbes (Tan *et al*., 2012; Deshmukh *et al*., 2016). These developments support the current interest in using high-throughput DNA sequencing as promising methodology for monitoring marine ecosystems status on the basis of microbial indicators and/or pathogens (Bourlat *et al*., 2013; Bruno *et al*., 2016: Goodwin *et al*., 2017).

The Black Sea, the largest landlocked and vulnerable sea in the world, is suitable environment for many species of pathogenic and non-pathogenic bacteria. However, the current data available on the microbial pathogens in the Black Sea region are limited and mostly rely on traditional cultivation and molecular methods. This paper provides an estimate of next generation high-throughput Illumina-based sequencing approaches to assess microbial pathogenic community of estuarine, coastal and open Black Sea.

Detection technologies for microbial pathogens in the Black Sea

Quality assessment of the Black Sea marine environment is performed thru national level monitoring programs under the Bathing Water Directives 2006/7/EC, 76/160 EEC and the Shellfish Water Directive 2006/113/EC in most of the countries in the region (BSC, 2002). Intestinal enterococci and *Escherichia coli* for bathing waters, as well as total fecal coliforms, *Escherichia coli* and *Salmonella spp.* for designated shellfish growing areas are determined, on culture-based enumeration and detection of fecal indicator bacteria, to assess the microbiological pollution of surface seawater with the purpose of protecting human health and the Black Sea environment. However, monitoring for fecal indicator bacteria does not provide insights on the possible occurrence of other microbial contaminants (e.g. pathogens). Furthermore, such culture-based methods are prone to false-negative results that arise from the failure to resuscitate viable but non-cultivable cells (Deshmukh *et al*., 2016).

Additional data on microbial pathogens of most predominant fish species cultured in the Black Sea (e.g. sea bass, turbot) were obtained in recent years by using a combination of classical cultivation methods and traditional molecular techniques, including PCR technique and Sanger sequencing (Kayiş *et al*., 2009; Oztürk *et al*., 2014; Kayiş *et al*., 2017). However the high costs and protracted detection technology due to difficult growth or requirement for host constitute major problems in this combined approach.

High-throughput sequencing allows faster and more accurate species identification and decreases dependence on cultivation and morphological taxonomic expertise (Glockner *et al*., 2012). Although information derived from DNA NGS sequencing is not included yet in the current Black Sea status assessment programs, there is widespread recognition of the importance of this approach, and an

increasing number of Black Sea projects, such as MARCY and EMBLAS, generate sequence-based bacterial community diversity inventory.

Black Sea pathogenic bacterial diversity at different trophic conditions

During 2012 and 2016, high-throughput 16S drRNA gene amplicon sequencing was performed to analyze the bacterial assemblage composition of Black Sea estuarine, coastal and open stations characterized by different environmental conditions and human-derived pressures. As a result taxonomic composition of Black Sea prokaryotic communities was determined within the range of different hydro-chemical and hydro-physical conditions and distinct communities specific to different Black Sea areas were identified (Pavlovska *et al*., 2016).

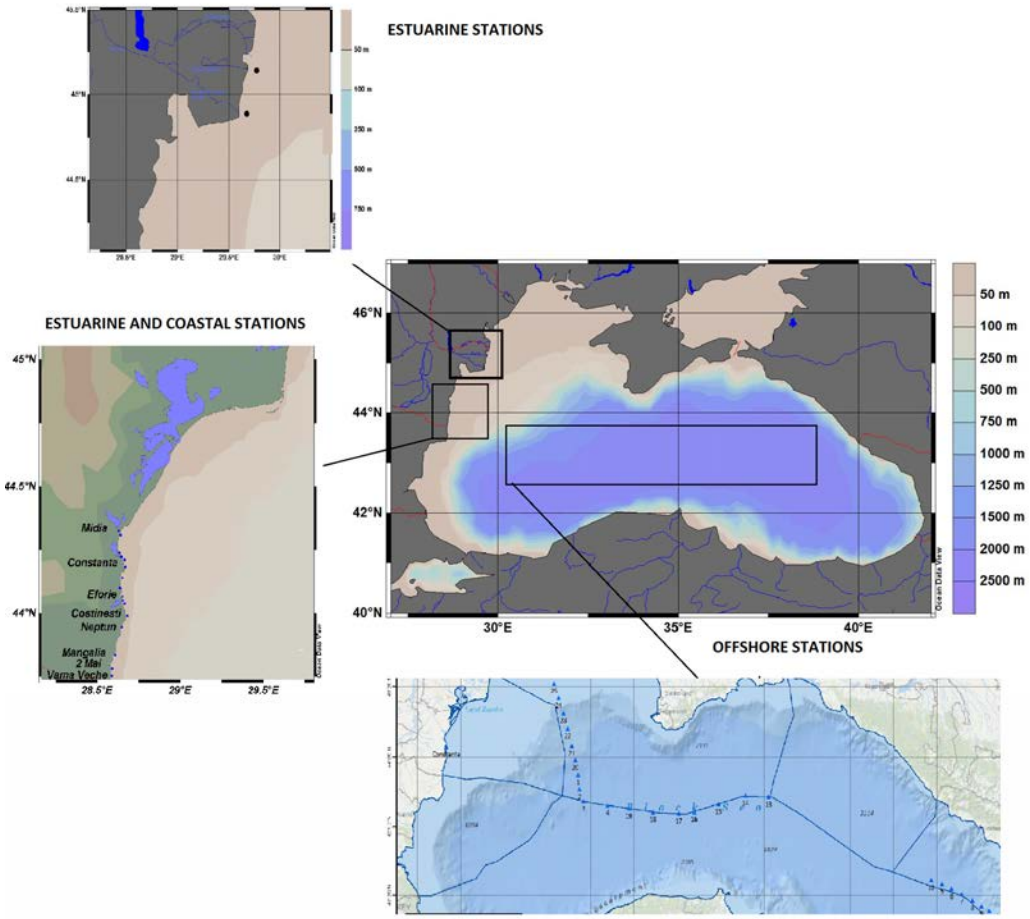


Figure 1 Location of sampling sites in three different trophic statuses regions of the Black Sea: a) estuarine, eutrophic (5.2 - 14.9 µg/l Chl-a): MARCY 1 (45.1443°N 29.7705°E), MARCY 2 (44.8768°N 29.61936°E), MARCY 3 (44.3250°N 28.6286°E), MARCY 4 (44.2178°N 28.6406°E), MARCY 5 (44.0998°N 28.6376°E); b) coastal, medium eutrophic (1.35 - 8.62 µg/l Chl-a): MARCY 6 (44.2376°N 28.6339°E), MARCY 7 (44.2194°N 28.6374°E), MARCY 8 (43.8183°N 28.5888°E), MARCY 9 (43.9453°N 28.6390°E), MARCY 10 (43.7537°N 28.5751°E); and c) open, oligotrophic (211-719 µg/l total nitrogen): EMBLAS 3 (41°39.895'N 41°35.513'E), EMBLAS 10 (41°45.675'N 41°42.728'E), EMBLAS 13 (41°53.742'N 41°37.225'E), EMBLAS 16 (42°11.750'N 41°25.433'E), EMBLAS 23 (41°47.976'N 41°28.871'E). The estuarine and coastal samples were collected during MARCY sampling campaign (April-September 2012) while the offshore waters were sampled during EMBLAS 2016 research cruise (May-June 2016).

Additionally, in depth exploration of the obtained NGS 16S-rDNA sequences was performed to identify all possible pathogenic strains in the Black Sea bacterial assemblages. To mine the two available datasets (MARCY and EMBLAS) for the presence of pathogens, five offshore stations were combined with both five coastal and five estuarine (Danube) stations, at which water samples were collected (Figure 1). As a result, 50 taxa harboring human, animal and plant pathogenic bacterial species were identified in the joint dataset (Table 1).

Table 1 Occurrence of NGS sequences of the potentially pathogenic bacterial OTUs detected in 54 Black Sea water samples collected in 2014 (MARCY project) and 2016 (EMBLAS project): estuarine (n=12), coastal (n=13) and offshore (n=29).

Species	**Number of positive samples**	**Relative abundance (ratio of positive samples)**	**Associated diseases**	**Reference**
Aeromonas salmonicida	1	0,03	furunculosis in fish	Fernandez-Alvarez *et al*., 2016
Aeromonas spp.	11	0,44	intestinal and extra-intestinal infections in humans and animals	Parker *et al*., 2011
Aeromonas veronii	2	0,08	human pathogen (wound infections, diarrhea, septicemia)	Thomsen *et al*., 2001
Aeromonas veronii bv. sobria	1	0,04	human pathogen	Thomsen *et al*., 2001
Aeromonas veronii bv. veronii	8	0,32	fish pathogen	Cai *et al*., 2012
Borrelia japonica	3	0,10	arthritis in mammals	Kaneda *et al*., 1998
Brevundimonas vesicularis	6	0,21	bacteremia in humans	Shang *et al*., 2011
Brucella sp.	29	1,00	brucellosis in humans and animals	Paulsen *et al*., 2002
Burkholderia spp.	29	0,54*	cystic fibrosis-related pathogen	Van Pelt *et al*., 1999
Corynebacterium bovis	3	0,12	animal pathogen	Burr *et al*., 2012
Coxiella burnetii	5	0,17	Q fever in humans and animals	Marrie *et al*., 1990
Endozoicomonas elysicola	10	0,34	epithelicystic in fish	Katharios *et al*., 2015

Enterococcus cecorum	7	0,24	avian pathogen	Jung *et al*., 2017
Enterovibrio spp.	6	0,24	potential fish pathogen	Austin *et al*., 2016
Erythrobacter vulgaris	15	0,6	potential pathogen of marine invertebrates	Ivanova *et al*.,2005
Escherichia coli	8	0,28	gastroenteritis, urinary tract infections, neonatal meningitis, hemorrhagic colitis, and Crohn's disease	Chaudhuri *et al*., 2012
Escherichia shigella spp.	11	0.44	intestinal disease	Ud-Din *et al*., 2014
Fischerella spp.	1	0,04	toxic to fish, associated with the development of other pathogens	Wright *et al*., 2006
Granulicatella elegans	6	0,21	infective endocarditis	Ohara-Nemoto *et al*., 2005
Haemophilus parainfluenzae	7	0,24	pathogen in respiratory tract of humans	Middleton *et al*., 2003
Helicobacter rappini	6	0,21	human and animal pathogen (enteric and systemic diseases)	Fox, 2002
Legionella pneumophila	2	0,08	respiratory illness (Legionnaires' pneumonia)	Kawano *et al*., 2007
Legionella spp.	29	1,00	legionellosis	Diederen *et al*., 2008
Legionella taurinensis	1	0,04	potential human pathogen (Legionnaires′ disease)	Lo Presti *et al*., 1999
Leptospira spp.	5	0,17	leptospirosis in humans and different animal species	Plank *et al*., 2000
Leptotrichia goodfellowii	11	0,38	endocarditis in humans	Matias *et al*., 2016
Leucobacter spp.	17	0,68	invertebrates and fish diseases	Hodgkin *et al*., 2013
Phytophthora infestans	3	0,12	late blight plant pathogen (late blight disease)	Haas *et al*., 2009
Mycobacterium interjectum	11	0,44	mycobacterial human pathogen (pulmonary diseases)	Stella *et al*., 2017
Mycobacterium poriferae	29	1,00	human and fish pathogen (pulmonary diseases)	Ballester *et al*., 2011
Mycoplasma phocidae	3	0,10	marine mammals pathogen (respiratory infections)	Ailing *et al*., 2011

Nocardioides aquiterrae	1	0,04	potential plant pathogen	Yoon et la., 2004
Pirellula spp.	29	1,00	budding bacteria	Kerger et la., 1988
Pseudomonas mendocina	5	0,2	human pathogen (infective endocarditis)	Aragon *et al*., 1992
Pseudomonas stutzeri	37	0,68*	skin and other infections in humans	Jorge *et al*., 2006
Rickettsia spp.	40	0,74*	human, animal and plant pathogen	Lydyard *et al*., 2009
Shewanella spp.	2	0,08	bacteremia in humans	Sharma *et al*., 2010
Spirochaeta spp.	33	0,61*	human and animal pathogen	Lindboe, 2001
Staphylococcus pasteuri	28	0,97	nosocomial infections in humans	Saving *et al*., 2009
Staphylococcus spp.	3	0,12	potential human pathogen	Gill *et al*., 2005
Streptococcus anginosus	2	0,07	brain and liver abscesses	Yilmaz, 2012
Streptococcus salivarius	1	0,03	potential pathogen	Chen, 1996
Streptococcus sanguinis	25	0,86	bacterial endocarditis	Paik *et al*., 2005
Streptomyces spp.	30	0,55	plant pathogen	Pánková *et al*., 2012
Vbrio aestuarianus	3	0,10	invertebrates pathogen	Garnier *et al*., 2008
Vibrio harveyi	12	0,48	fish and invertebrates pathogen	Austin *et al*., 2006
Vibrio ichthyoenteri	2	0,08	fish pathogen	Kim *et al*., 2014
Vibrio kanaloae	2	0,08	invertebrates pathogen	Thompson *et al*., 2003
Vibrio rotiferianus	15	0,52	fish pathogen	Austin *et al*., 2016
Vibrio shilonii	5	0,2	coral pathogen	Kushmaro *et al*., 2001
Vibrio vulnificus	29	0,57*	wound infection, septicemia, gastrointestinal disease	Jones *et al*., 2009

***Present in all datasets and calculated based on offshore, coastal and estuarine data.*

Among the major bacterial phyla containing pathogenic species, *Actinobacteria* and *alpha-proteobacteria* were the most common in estuarine waters, whereas *gamma-proteobacteria* dominated in the coastal Black Sea. *Mycobacterium* and *Erythrobacter*, containing pathogenic species, formed the most abundant genera of microbial assemblage from the Danube estuarine waters (Figure 2). The high abundance of these phyla known to harbor pathogenic species in the eutrophic Black Sea region (5.21 – 14.9 μg/l range of chlorophyll *a*) may be an indication that these microbes have specific

nutritional or environmental preference. Previous studies have shown the influence of environmental gradients on the abundance and distribution of *Mycobacterium* in coastal estuaries. For example Jacobs *et al*. (2009) found a strong nutrient response, with all nitrogen components and turbidity measurements correlating positively with *Mycobacterium spp.* abundance. Moreover the genus *Erythrobacter* is known as a very relevant component of the marine planktonic assemblages, becoming in some cases one of the most dominant groups in eutrophic coastal environments (Frette *et al*., 2004). *Vibrio* and *Aeromonas*, formed a diverse and dynamic genera of pathogenic microbial assemblages from the Romanian coastal waters. Two gamma-proteobacterial pathogenic strains related to *Aeromonas* (*Aeromonas spp.* and *Aeromonas veronii bv. veronii*) were found to constitute a relatively large percentage of the total bacterial rRNA contribution (Figure 2). Aeromonads are known to cause severe disease both in humans and fish. *Aeromonas veronii bv. veronii* are predominantly pathogenic to fish and other cold-blooded species, and the mesophilic strains of *Aeromonas spp.* are emerging as important pathogens in humans, causing a variety of intestinal and extra-intestinal infections. Many *Vibrio* species are also important zoonotic bacterial agents, causing disease in fish and shellfish and death among domestic marine life (Vezzuli, this volume).

Occurrence of pathogens in offshore Black Sea

NGS sequences matching pathogenic species were surprisingly found in the offshore waters collected during the EMBLAS project in late May - early June 2016. A total of 29 taxa harboring potentially pathogenic species were identified at different water column depths ranging from the surface to the anoxic H_2S zone of the Black Sea (Figure 3). The potentially pathogenic bacterial species identified by this study in the offshore Black Sea waters have been reported elsewhere in various hosts (Table 1) and were exposed to a variety of abiotic (e.g. temperature, sunlight, nutrients and oxygen) and biotic challenges that are specific for the unique Black Sea ecosystems (BSC 2002, 2008). The pathogenic bacterial communities observed in the offshore waters were dominated by *Streptomyces spp.*, *Brucella sp.*, and *Pseudomonas stutzeri* (Figure 2). These microorganisms had previously reported in the various inshore regions of the world seas, mostly in water samples taken closest to the coast. No information is available on the occurrence of pathogenic bacteria in open marine waters.

Conclusions

The high-throughput pyrosequencing analysis offers an opportunity to study in depth the diversity of microbial pathogens in Black Sea waters. Mining of existing next-generation-sequencing datasets revealed unprecedented diversity of pathogenic bacteria in both inshore (estuarine and coastal) and offshore waters of the Black Sea. The composition of Black Sea pathogenic microbiota varied generally as a function of water trophic status. The survival and/or proliferation of microbial pathogens could be influenced by different environmental factors (e.g. temperature, salinity, nutrients, light and oxygen) specific for each studied site. Our study demonstrated the potential application of DNA high-throughput sequencing combined with data mining as a tool in routine monitoring of the Black Sea microbial water quality.

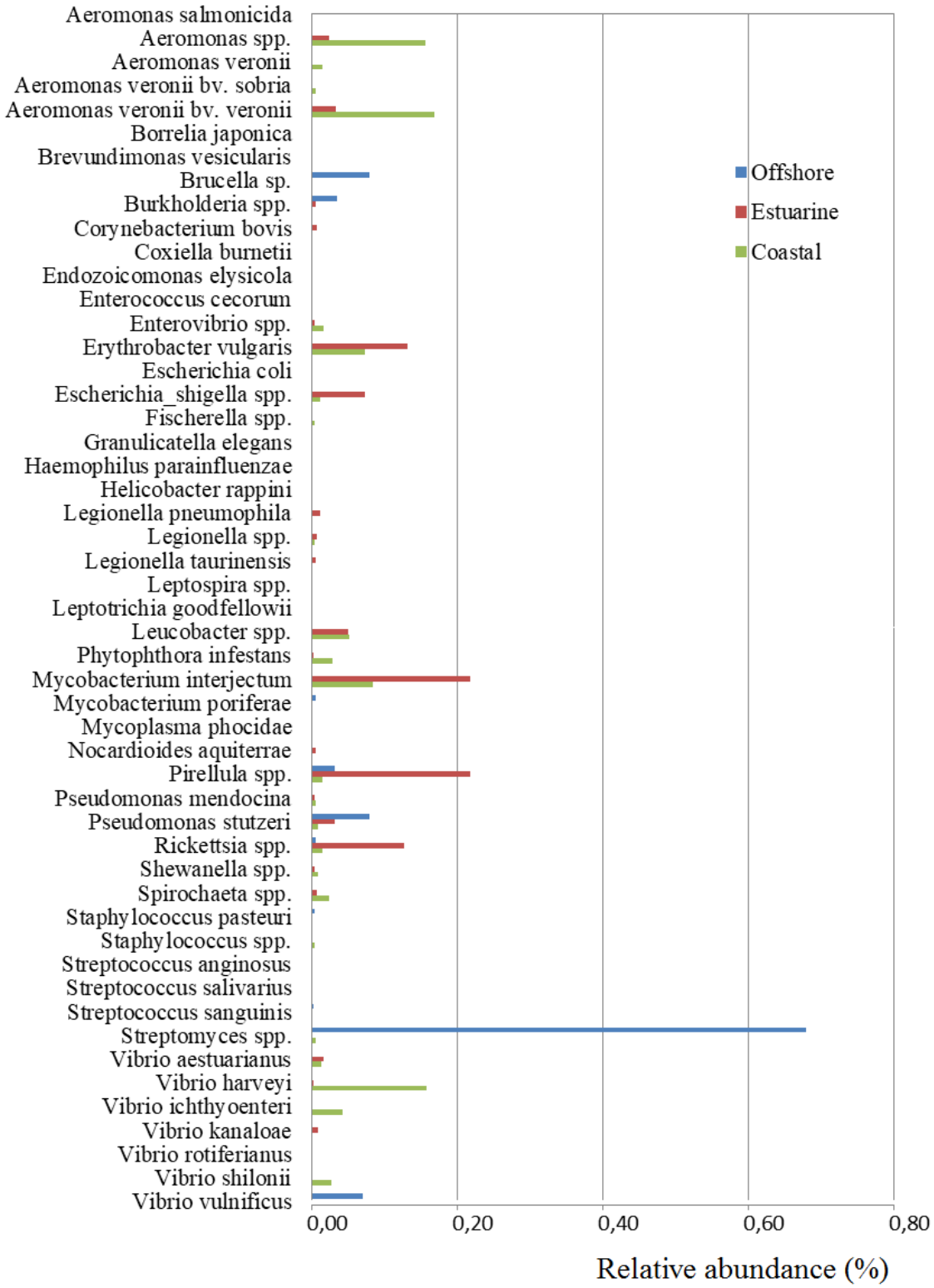


Figure 2. Relative abundance of pathogenic bacteria in the Black Sea. Bar charts indicate the relative abundance (the proportion of stations with the species present) of pathogenic bacterial species in estuarine, coastal and waters offshore. The chart is based on data collected within the MARCY and EMBLAS projects.

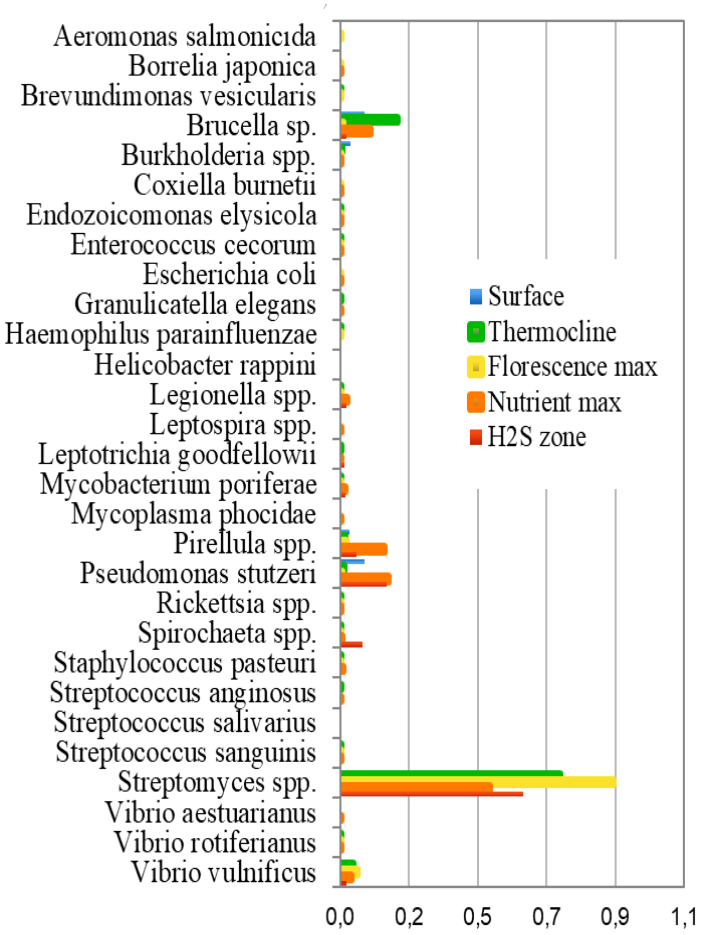


Figure 3. Relative abundance of potentially pathogenic bacteria in the offshore waters of the Black Sea. Bar charts indicate the relative abundance (the proportion of stations with the species present) of sequences matching pathogenic species found with 454 sequenced S-D-Bact-0341-b-S-17 /S-D-Bact-0785-a-A-21 universal bacterial V3-V4 amplicons at various depths: surface (0 – 0.5m), thermocline (5-25 m), fluorescence maximum (17-55 m), nutrient maximum (54 – 122 m) and anoxic (H_2S) zone (99 – 1000 m). The chart is based on the data collected in 2016 in the course of EMBLAS project for the selected five offshore stations.

Acknowledgements: This study was supported by UEFISCDI (MARCY BS-ERA.NET 019/BS 7-050-2011 and COFUND-ERA4CS-CoCliME project) and EU/UNDP (EMBLAS EU/UNDP Black Sea Project).

* this chapter is to be cited as :

Stoica E., Pavlovska M., Dykyi E and Kormas K. 2017. Next generation sequencing-based approaches to characterize microbial pathogenic community and their potential relation to the Black Sea ecosystem status. pp. 115 – 123 In CIESM Monograph 49 [F. Briand ed.] Searching for Bacterial Pathogens in the Digital Ocean, 158 p., CIESM Publisher, Monaco and Paris.

BIBLIOGRAPHIC REFERENCES

Aggarwal C. C. and Han, J. 2014. Frequent Pattern Mining. Springer.

Agrawal R. and Srikant R. 1994. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, 487-499.

Agrawal R., Imieliński T. and Swami A. 1993. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

Ahmed W., Staley C., Sadowsky M.J., Gyawali P., Sidhu J.P.S., Palmer A., ... and Toze S. 2015. Toolbox approaches using molecular markers and 16S rRNA gene amplicon data sets for identification of fecal pollution in surface water. *Applied and environmental microbiology*, 81 (20): 7067-7077.

Al Bulushi I.M., Poole S.E., Barlow R., Deeth H.C. and Dykes G.A. 2010. Speciation of Gram-positive bacteria in fresh and ambient-stored sub-tropical marine fish. *Inter. J. of food microbiology*, 138(1): 32-38.

Alberti A., Poulain J., Engelen S., Labadie K., Romac S., Ferrera I.and Cruaud, C. 2017. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific data*, 4: sdata 201793.

Altschul S. F., Madden T. L., Schäffer A. A., Zhang J., Zhang Z., Miller W.and Lipman D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25 (17): 3389-3402.

Amann R.I., Ludwig W. and Schleifer K.H. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev*. 59: 143–169.

Amend A. S., Oliver T. A., Amaral Zettler L. A., Boetius A., Fuhrman J. A., Horner Devine M. C., and Zinger, L. 2013. Macroecological patterns of marine bacteria on a global scale. *Journal of Biogeography*, 40 (4): 800-811.

Anderson J.K., Smith T.G. and T.R. Hoover. 2009. Sense and sensibility: flagellum-mediated gene regulation. *Trends in Microbiol*., 18: 30-37.

Anderson P.K., Cunningham A.A., Patel N.G., Morales F.J., Epstein P.R. and Daszak P. 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends Ecol. Evol*., 19 (10): 535-44.

Andrei A.Ş., Baricz A., Robeson M.S., Păuşan M.R., Tămaş T., Chiriac C., Szekeres E., Barbu-Tudoran L., Levei E.A., Coman C., Podar M., Banciu H.L. 2017. Hypersaline sapropels act as hotspots for microbial dark matter. *Sci. Rep*., 7: 6150.

Aragone M.R., Maurizi L.O., Clara J.L., Navarro E. and Ascione A. 1992. *Pseudomonas mendocina*, an environmental bacterium isolated from a patient with human infective endocarditis. *Journal of clinical microbiology*, 30: 1583-1584.

Aravindraja C., Viszwapriya D. and Pandian S.K. 2013. Ultradeep 16S rRNA sequencing analysis of geographically similar but diverse unexplored marine samples reveal varied bacterial community composition. *Plos one*, 8 (10): e76724.

Argimón S. et al., 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial genomics*, 2 (11), p.e000093.

Austin B. and Austin D.A. Bacterial Fish Pathogens. Disease of Farmed and Wild Fish. Springer International Publishing Switzerland 2016.

Austin B. and Zhang X.H. 2006. *Vibrio harveyi*: a significant pathogen of marine vertebrates and invertebrates. *Lett. Appl. Microbiol*., 43 (2): 119-24.

Bachere E., Gueguen Y., Gonzalez M., de Lorgeril J., Garnier J. and Romestand B. 2004. Insights into the anti-microbial defense of marine invertebrates: the penaeid shrimps and the oyster *Crassostrea gigas*. *Immunological reviews*, 198: 149-168.

Baker-Austin C. et al. 2016. Heat Wave-Associated *Vibriosis*, Sweden and Finland, 2014. *Emerging infectious diseases*, 22(7): 1216–1220.

Baker-Austin C., Trinanes J. A., Taylor N. G., Hartnell R., Siitonen A. and Martinez-Urtaza J. 2013. Emerging *Vibrio* risk at high latitudes in response to ocean warming. *Nature Climate Change*, 3(1): 73-77.

Balke V.L. and Gralla J.D. 1987. Changes in the linking number of supercoiled DNA accompany growth transitions in *Escherichia coli*. *J. Bacteriol.,* 169(10): 4499-4506

Bang I.S., Audia J.P., Park Y.K., and Foster J.W. 2002. Autoinduction of the ompR response regulator by acid shock and control of the *Salmonella enterica* acid tolerance response. *Mol. Microbiol*., 44(5): 1235-1250

Barberán A., Casamayor E.O. and Fierer N. 2014. The microbial contribution to macroecology. *Frontiers in Microbiology*, 5: 203.

Barbosa Solomieu V., Renault T. and Travers M.A. 2015. Mass mortality in bivalves and the intricate case of the Pacific oyster, *Crassostrea gigas*. *J. Invertebr. Pathol*., 131: 2-10.

Barnes A.C. et al., 2016. Whole genome analysis of *Yersinia ruckeri* isolated over 27 years in Australia and New Zealand reveals geographical endemism over multiple lineages and recent evolution under host selection. *Microbial genomics*, 2 (11), p.e000095.

Bates A.D. and Maxwell A. 2007. Energy coupling in type II topoisomerases: why do they hydrolyze ATP? *Biochemistry*, 46: 7929–7941

Bayliss S.C. et al., 2017. The promise of whole genome pathogen sequencing for the molecular epidemiology of emerging aquaculture pathogens. *Frontiers in microbiology*, 8: 121.

Belkin S. and Colwell R.R. (Eds.) 2005. Oceans and health: pathogens in the marine environment. Springer, New York.

Benhamed S., Guardiola F.A., Mars M. and Esteban M.Á. 2014. Pathogen bacteria adhesion to skin mucus of fishes. *Veterinary microbiology*, 171(1): 1-12.

Berg H.C. and D.A. Brown. 1972. Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking. *Nature*, 239: 500-504.

Bjarnason J., Southward C.M., Surette M.G. 2003. Genomic profiling of iron-responsive genes in *Salmonella enterica* serovar typhimurium by high-throughput screening of a random promoter library. *J Bacteriol*., 185 (16): 4973-82.

Black Sea Commission 2002. State of the Environment of the Black Sea: pressures and trends, 1996-2000. www.blacksea-commission.org/_publ-SOE2002-eng.asp, Istanbul, Turkey.

Black Sea Commission 2008. State of the Environment of the Black Sea (2001 - 2006/7). Edited by Temel Oguz. Publications of the Commission on the Protection of the Black Sea Against Pollution (BSC) 2008-3, Istanbul, Turkey, 448 pp.

Blagrove M.S.C, Caminade C., Waldmann E., Sutton E.R., Wardeh M. and Baylis M. 2017. Co-occurrence of viruses and mosquitoes at the vectors optimal climate range: An underestimated risk to temperate regions? *PLoS Negl. Trop. Dis.*, 15: 11(6):e0005604.

Bogomolni A.L., Gast R.J., Ellis J.C., Dennett M., Pugliares K.R., Lentell B.J. and Moore, M.J. 2008. Victims or vectors: a survey of marine vertebrate zoonoses from coastal waters of the Northwest Atlantic. *Diseases of aquatic organisms*, 81 (1): 13-38.

Bonferroni C. E. 1935. Il calcolo delle assicurazioni su gruppi di teste. Studi in Onore del Professore Salvatore Ortu Carboni, 13-60.

Bonferroni C. E. 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8: 3-62.

Bourlat S.J., Borja A., Gilbert J.A., Taylor M.I., Davies N., Weisberg S.B., Griffith J.F., Lettieri T., Field D., Benzie J., Glockner F.O., Rodriguez-Ezpeleta N., Faith D.P., Bean T.P. and Obst M. 2013. Genomics in marine monitoring: New opportunities for assessing marine health status. *Marine Pollution Bulletin*, 74 (1): 19–31.

Bowler C., Karl D. M. and Colwell R. R. 2009. Microbial oceanography in a sea of opportunity. *Nature*, 459: 180-184.

Boyle E.C., Bishop J.L., Grassl G.A. and Finlay B.B. 2007. Salmonella: from pathogenesis to therapeutics. *J. Bacteriol*., 189 (5): 1489-1495.

Bradley P. et al., 2015. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nature communications*, 6, p.10063.

Brenner M. and Krumme U. 2007. Tidal migration and patterns in feeding of the four-eyed fish *Anableps anableps L.* in a north Brazilian mangrove. *J. Fish Biol*., 70: 406–427.

Brettar I., Guzman C.A., Höfle M.G. In: CIESM Workshop Monographs No. 31: Marine Sciences and Public Health. Geneva: CIESM; 2007. Human pathogens in the marine environment - an ecological perspective. pp. 59–68.

Brinkmeyer R. 2016. Diversity of bacteria in ships ballast water as revealed by next generation DNA sequencing. *Marine Pollution Bulletin*, 107 (1): 277-285.

Brooks J.P., Edwards D.J., Harwich M.D., Rivera M.C., Fettweis J.M., Serrano M.G., Reris R.A., Sheth N.U., Huang B., Girerd P., Strauss J.F. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC microbiology*, 15 (1): 66.

Brown S., Nicholls R.J., Woodroffe C.D., Hanson S., Hinkel J., Kebede A.S., ... and Vafeidis A.T. 2013. Sea-level rise impacts and responses: a global perspective. In Coastal hazards (pp. 117-149). Springer Netherlands.

Brum J. R., Ignacio-Espinoza J. C., Roux S., Doulcier G., Acinas S. G., Alberti A.and Gorsky, G. 2015. Patterns and ecological drivers of ocean viral communities. *Science*, 348 (6237): 1261498.

Bruno A., Sandionigi A., Galimberti A., Siani E., Labra M., Cocuzza C. et al. 2017. One step forwards for the routine use of high-throughput DNA sequencing in environmental monitoring. An efficient and standardizable method to maximize the detection of environmental bacteria. *Microbiology Open*, 6:e00421. https://doi.org/10.1002/mbo3.421.

Bruto M., James A., Petton B., Labreuche Y., Chenivesse S., Alunno-Bruscia M., Polz M.F. and Le Roux F. 2017. *Vibrio crassostreae*, a benign oyster colonizer turned into a pathogen after plasmid acquisition. *Isme J.*, 11: 1043-1052.

Brynildsrud O. *et al.* 2014. Microevolution of *Renibacterium salmoninarum*: evidence for intercontinental dissemination associated with fish movements. *Isme J.*, 8 (4): 746–756.

Brynildsrud O. *et al.* 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome biology*, 17(1): 238.

Buchfink Benjamin Chao Xie and Daniel H. Huson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12 (1): 59-60.

Bultel-Poncé V., Debitus C., Berge J.P., Cerceau C. and Guyot M. 1998. Metabolites from the sponge-associated bacterium *Micrococcus luteus*. *Journal of marine biotechnology*, 6: 233-236.

Burr H.N., Wolf F.R. and Lipman N.S. 2012. *Corynebacterium bovis*: Epizootiologic features and environmental contamination in an enzootically infected rodent room. *Journal of the American Association for Laboratory Animal Science : JAALAS*. 51 (2): 189-198.

Cai S.H., Wu Z.H., Jian J.C., Lu Y.S. and Tang J.F. 2012. Characterization of pathogenic *Aeromonas veronii bv. veronii* associated with ulcerative syndrome from chinese longsnout catfish (*Leiocassis longirostris*, Günther). *Brazilian Journal of Microbiology*, 43 (1): 382-388.

Cameron A.D. and Dorman C.J. 2012. A fundamental regulatory mechanism operating through OmpR and DNA topology controls expression of Salmonella pathogenicity islands SPI1 and SPI2. *Plos Genet*., 8 (3): e1002615

Cameron A.D., Kröger C., Quinn H.J., Scally I.K., Daly A.J., Kary S.C. and Dorman C.J. 2013. Transmission of an oxygen availability signal at the *Salmonella enterica serovar Typhimurium* fis promoter. *Plos one*, 8 (12): e84382

Cameron A.D., Stoebel D.M. and Dorman C.J. 2011. DNA supercoiling is differentially regulated by environmental factors and FIS in *Escherichia coli* and *Salmonella enterica. Mol. Microbiol*., 80 (1): 85-101.

Campbell B.J., Yu L., Heidelberg J.F. and Kirchman D.L. 2011. Activity of abundant and rare bacteria in a coastal ocean. *PNAS*, 108 (31): 12776-12781.

Carbone M., Maugeri T.L., Gugliandolo C., La Camera E., Biondo C. and Fera M.T. 2005. Occurrence of *Helicobacter pylori* DNA in the coastal environment of southern Italy (Strait of Messina). *Journal of Applied microbiology*, 98 (3): 768-774.

Carter S.K., Fernando C.W., Cooper A.B., Cordeiro-Duarte A. 1999. Consumption rate, food preferences and transit time of captive giant otters *Pteronura brasiliensis*: implications for the study of wild populations. *Aquat. Mamm*. 25: 79–90.

Cerdan R., Bloch V., Yang Y., Bertin P., Dumas C., Rimsky S., Kochoyan M., and Arold S.T. 2003. Crystal structure of the N-terminal dimerisation domain of VicH, the H-NS-like protein of *Vibrio cholerae*. *J. Mol. Biol*., 334 (2): 179-185.

Certner R.H., Dwyer A.M., Patterson M.R. and Vollmer S.V. 2017. Zooplankton as a potential vector for white band disease transmission in the endangered coral, *Acropora cervicornis*. *Peer J.*, 5: e3502.

Ceuppens S., De Coninck D., Bottledoorn N., Van Nieuwerburgh F., Uyttendaele M. 2017. Microbial community profiling of fresh basil and pitfalls in taxonomic assignment of enterobacterial pathogenic species based upon 16S rRNA amplicon sequencing. *Int. J. Food Microbiol*., 257: 148-156.

Chakraborty S., Mizusaki H., and Kenney L.J. 2015. A FRET-based DNA biosensor tracks OmpR-dependent acidification of Salmonella during macrophage infection. *Plos Biol*., 13 (4): e1002116.

Champoux J.J. 2001. DNA topoisomerases: structure, function, and mechanism. *Annu. Rev. Biochem*., 70: 369-413.

Chaudhuri R.R. and Henderson I.R. 2012. The evolution of the *Escherichia coli* phylogeny. Infection, genetics and evolution. *Journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 12: 214-26.

Chen Y.Y. 1996. *Streptococcus salivarius* urease: genetic and biochemical characterization and expression in a dental plaque streptococcus. *Infection and Immunity*, 64(2): 585-592.

Choi J. and Groisman E.A. 2016. Acidic pH sensing in the bacterial cytoplasm is required for Salmonella virulence. *Mol. Microbiol*., 101: 1024-1038

Cleaveland S., Laurenson M.K. and Taylor L.H. 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 356 (1411): 991-999.

Coale K.H., Johnson K.S., Fitzwater S.E., Gordon R.M., Tanner S. …and Kudela R. 1996. A massive phytoplankton bloom induced by an ecosystem-scale iron fertilization experiment in the equatorial Pacific Ocean. *Nature*, 383 (6600): 495 – 501.

Colwell RR. 1996. Global climate and infectious disease: the cholera paradigm. *Science*, 274: 2031-2025.

Conter A., Menchon C., and Gutierrez C. 1997. Role of DNA supercoiling and RpoS sigma factor in the osmotic and growth phase-dependent induction of the gene osmE of *Escherichia coli* K12. *J. Mol. Biol*., 273 (1): 75-83

Copeland C. 2008. Cruise ship pollution: background, laws and regulations, and key issues. CRS Report for Congress, Congressional Research Service, 26 p.

Cordero O.X., Ventouras L.A., Delong E.F. and Polz M.F. 2012a. Public good dynamics drives evolution of iron acquisition strategies in natural bacterioplankton populations. *PNAS*, 109: 20059-20064.

Cordero O.X., Wildschutte H., Kirkup B., Proehl S., Ngo L., Hussain F., Le Roux F., Mincer T. and Polz M.F. 2012b. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science*, 337: 1228-1231.

Czerucka D., Dahan S., Mograbi B., Rossi B., and P. Rampal. 2001. Implication of mitogen-activated protein kinase in T84 cell responses to EPEC infection. *Inf. Immun*., 69: 1298-1305.

D'Auria G., Peris-Bondia F., Džunková M., Mira A., Collado M.C., Latorre A. & Moya A. 2013. Active and secreted IgA-coated bacterial fractions from the human gut reveal an under-represented microbiota core. *Sci. Rep.*, 3: 3515.

Dahan S., Busuttil V., Imbert V., Peyron J-F, Rampal P., and D. Czerucka. 2002. Enterohemorrhagic *Escherichia coli* infection induces interleukin-8 production via activation of mitogen-activated proteine kinases and the transcription factors NF-kB and AP-1 in T84 cells. *Inf. Immun*., 70: 2304-2310.

Daszak P., Cunningham A.M. and Hyatt A.D. 2000. Emerging infectious diseases of wildlife-- threats to biodiversity and human health. *Science*, 287 (5459): 1756.

de la Torre J. R., Christianson L. M., Béjà O., Suzuki M. T., Karl D. M., Heidelberg J. and DeLong E. F. 2003. Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc. Natl. Acad. Sci., USA100,* 12830-12835.

De Santi C., Altermark B., de Pascale D., and Willassen N.P. 2016. Bioprospecting around Arctic islands: Marine bacteria as rich source of biocatalysts. *Journal of basic microbiology*, 56 (3): 238-253.

De Vargas C., Audic S., Henry N., Decelle J., Mahé F., Logares R.and Carmichael M. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348 (6237): 1261605.

Degremont L., Garcia C. and Allen S.K., Jr. 2015. Genetic improvement for disease resistance in oysters: A review. *J. Invertebr. Pathol*., 131: 226-241.

Deshmukh R. A., Joshi K., Bhand S. and Roy U. 2016. Recent developments in detection and enumeration of waterborne bacteria: a retrospective minireview. *Microbiology Open*, 5 (6): 901-922.

Diederen B.M.W. 2008. Legionella spp. and Legionnaires' disease. *J. Infect*., 56: 1–12.

Dillon S.C, Cameron A.D, Hokamp K., Lucchini S., Hinton J.C. and Dorman C.J. 2010. Genome-wide analysis of the H-NS and Sfh regulatory networks in *Salmonella Typhimurium* identifies a plasmid-encoded transcription silencing mechanism. *Mol. Microbiol*., 76 (5): 1250-1265

Ding Y., Davis B.M. and Waldor M.K. 2004. Hfq is essential for *Vibrio cholerae* virulence and downregulates sigma expression. *Mol Microbiol*, 53: 345-354.

Dohan K. and Maximenko N. 2010. Monitoring ocean currents with satellite sensors. *Oceanography*, 23 (4): 94-103.

Donaldson G.P., Lee S.M. and Mazmanian S.K. 2016. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.*, 14: 20–32.

Dong G. and Bailey J. 2013. Contrast Data Mining: Concepts, Algorithms, and Applications. CRC Press.

Dorman C.J. 2007. H-NS, the genome sentinel. *Nat. Rev. Microbiol*., 5 (2): 156-161.

Dorman C.J. 2009. Regulatory integration of horizontally-transferred genes in bacteria. *Front. Biosci.* (Landmark Ed), 14: 4103-4112.

Dorman C.J. 2011. Regulation of transcription by DNA supercoiling in *Mycoplasma genitalium*: global control in the smallest known self-replicating genome. *Mol. Microbiol.*, 81 (2): 302-304.

Dorman C.J. and Dorman M.J. 2016. DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression. Biophys. Rev., 8(3): 209-220.

Dorman C.J. and Dorman M.J. 2017. Control of virulence gene transcription by indirect readout in *Vibrio cholerae* and *Salmonella enterica serovar Typhimurium*. *Environ. Microbiol*., doi: 10.1111/1462-2920.13838.

Dorman C.J., Colgan A. and Dorman M.J. 2016. Bacterial pathogen gene regulation: a DNA-structure-centred view of a protein-dominated domain. *Clin. Sci. (Lond.),* 130 (14): 1165-1177.

Dumontet et al. 2000. Prevalence and diversity of *Aeromonas* and *Vibrio spp*. in coastal waters of Southern Italy. *Comparative immunology, microbiology and infectious diseases*, 23 (1): 53-72.

Duperthuy M., Schmitt P., Garzón E., Caro A., Rosa R. D., Le Roux F….and Kieffer-Jaquinod S. 2011. Use of OmpU porins for attachment and invasion of *Crassostrea gigas* immune cells by the oyster pathogen *Vibrio splendidus*. *PNAS*, 108 (7): 2993-2998.

Džunková M., D'Auria G. and Moya A. 2015. Direct sequencing of human gut virome fractions obtained by flow cytometry. *Front. Microbiol*., 6: 955.

Džunková M., Moya A., Vazquez-Castellanos J.F., Artacho A., Chen X., Kelly C. & D'Auria G. 2016. Active and secretory IgA-coated bacterial fractions elucidate dysbiosis in Clostridium difficile infection. *mSphere*, 1.

Eberl L. and Vandamme P. 2016. Members of the genus Burkholderia: good and bad guys. *Food Research*, doi: 10.12688/f1000research.1.

Edgar R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32 (5): 1792-1797.

Eldin C., Mélenotte C., Mediannikov O., Ghigo E., Million M., Edouard S. ... and Raoult D. 2017. From Q fever to *Coxiella burnetii* infection: a paradigm change. *Clinical microbiology reviews*, 30 (1): 115-190.

Ellegaard K.M., Engel P. 2016. Beyond 16S rRNA community profiling: intra-species diversity in the gut microbiota. *Front. Microbiol*., 7: 1475

Elshahed M.S., Youssef N.H., Spain A.M., Sheik C., Najar F.Z., Sukharnikov L.O. ... and Krumholz, L.R. 2008. Novelty and uniqueness patterns of rare members of the soil biosphere. *Applied and Environmental Microbiology*, 74(17): 5422-5428.

Emerson J.B., Adams R.I., Roman C.M.B., Brooks B., Coil D.A. …and Rothschild L.J. 2017. Schrödinger's microbes: Tools for distinguishing the living from the dead in microbial ecosystems. *Microbiome*, 5: 86.

Fabioux C., Corporeau C., Quillien V., Favrel P. and Huvet A. 2009. In vivo RNA interference in oyster-vasa silencing inhibits germ cell development. *FEBS J.*, 276: 2566-2573.

Fadrosh D. W., Ma B., Gajer P., Sengamalay N., Ott S., Brotman R. M. and Ravel J. 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, 2(1): 6.

Farhat M.R. *et al.* 2014. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. *Genome medicine*, 6 (11): 101.

Farto R., Guisande J.A., Armada S.P., Prado S. and Nieto T.P. 2006. An improved and rapid biochemical identification of indigenous aerobic culturable bacteria associated with Galician oyster production. *Journal of Shellfish Research*, 25 (3): 1059-1065.

Fass E., and Groisman E.A. 2009. Control of Salmonella pathogenicity island-2 gene expression. *Curr. Opin. Microbiol*., 12 (2): 199-204.

Faury N., Saulnier D., Thompson F.L., Gay M., Swings J. and Le Roux F. 2004. *Vibrio crassostreae sp. nov.,* isolated from the haemolymph of oysters (*Crassostrea gigas*). *Int. J. Syst. Evol. Microbiol.,* 54: 2137-2140.

Feil E.J. 2015. Toward a synthesis of genotypic typing and phenotypic inference in the genomics era. *Future microbiology*, 10 (12): 1897–1899.

Fenchel T. 2002. Microbial behaviour in a heterogeneous world. *Science*, 296: 1068-1071.

Feng X., Oropeza R., and Kenney L.J. 2003. Dual regulation by phospho-OmpR of ssrA/B gene expression in Salmonella pathogenicity island 2. *Mol. Microbiol.,* 48 (4): 1131-1143.

Feng X., Walthers D., Oropeza R., and Kenney L.J. 2004. The response regulator SsrB activates transcription and binds to a region overlapping OmpR binding sites at Salmonella pathogenicity island 2. *Mol. Microbiol.,* 54(3): 823-835.

Fisher M.C., Henk D.A., Briggs C.J., Brownstein J.S., Madoff L.C, McCraw S.L. and Gurr S.J. 2012. Emerging fungal threats to animal, plant and ecosystem health. *Nature*, 484 (7393): 10.1038/nature10947.

Forst S., Delgado J., and Inouye M. 1989. Phosphorylation of OmpR by the osmosensor EnvZ modulates expression of the ompF and ompC genes in *Escherichia coli. Proc. Natl. Acad. Sci. USA,* 86 (16): 6052-6056.

Fouts D.E. *et al.* 2012. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic acids research*, 40 (22): 172.

Fouz B., Toranzo A.E., Marco-Noales E., and Amaro C. 1998. Survival of fish-virulent strains of *Photobacterium damselae subsp. damselae* in seawater under starvation conditions. *FEMS Microbiology Letters*, 168 (2): 181-186.

Frette K., Johnsen K., Jørgensen N.O.G., Nybroe O. and Kroer N. 2004. Functional characteristics of culturable bacterioplankton from marine and estuarine environments. *Int. Microbiol*., 7: 219–227.

Frymier P.D., Ford R.M., and H.C. Berg. 1995. Three-dimensional tracking of motile bacteria near a solid planar surface. *Proc. Natl. Acad. Sci. USA*, 92: 6195-6199.

Fu L. *et al.* 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* , 28(23), pp.3150–3152.

Gabriel M.W., Matsui G.Y., Friedman R., Lovell C.R. 2014. Optimization of multilocus sequence analysis for identification of species in the genus *Vibrio*. *Applied and environmental microbiology.,* 80 (17): 5359-65.

Galperin M. Y., Makarova K. S., Wolf Y. I.and Koonin E. V. 2014. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic acids research*, 43(D1): D261-D269.

Garnier M., Labreuche Y., Nicolas J.L.2008. Molecular and phenotypic characterization of *Vibrio aestuarianus subsp. francensis subsp. nov*., a pathogen of the oyster *Crassostrea gigas*. *Syst. Appl. Microbiol.,* 31: 358–365.

Gasol J.M., Giorgio P.A. del, Massana R. & Duarte C.M. 1995. Active versus inactive bacteria: size-dependence in a coastal marine plankton community. *Mar. Ecol. Prog. Ser.*, 128: 91–97.

Gay M., Renault T., Pons A.M. and Le Roux F. 2004. Two *Vibrio splendidus* related strains collaborate to kill *Crassostrea gigas*: taxonomy and host alterations. *Dis. Aquat. Org.,* 62: 65-74.

GBD 2015 DALYs and HALE Collaborators. 2016. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet, 388(10053):1603-1658.

Gdoura M., Sellami H, Nasfi H., Trabelsi R., Mansour S., Attia T., Nsaibia S., Vallaeys T., Gdoura R., Siala M. 2016. Molecular detection of the three major pathogenic vibrio species from seafood products and sediments in Tunisia using real-time PCR. *Journal of Food Protection*, 79 (12): 2086-2094.

Gentile G., Giuliano L., D'Auria G., Smedile F., Azzaro M., De Domenico M., Yakimov M.M. 2006. Study of bacterial communities in Antarctic coastal waters by a combination of 16S rRNA and 16S rDNA sequencing. *Environmental Microbiology*. 8 (12): 2150-61.

Ghenem L., Elhadi N., Alzahrani F. and Nishibuchi M. 2017. *Vibrio parahaemolyticus*: A review on distribution, pathogenesis, virulence determinants and epidemiology. *Saudi Journal of Medicine and Medical Sciences*, 5 (2): 93.

Gignoux-Wolfsohn S.A., Marks C.J. and Vollmer S.V. 2012. White Band Disease transmission in the threatened coral, *Acropora cervicornis*. *Scientific reports*, 2: 804.

Gill S.R., Founts D. E., ARcher G. L. and 25 others. 2005. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *Journal of bacteriology*, 187 (7): 2426-38.

Giovannoni S. and Nemergut, D. 2014. Microbes ride the current. *Science*, 345: 1246–1247.

Glass K., Ott E., Losert W. and Girvan M. 2015. Implications of Functional Similarity for Gene Regulatory Interactions. *Royal Society Interface*, 9 (72): 1625-1636.

Glockner F. O., Stal L. J., Sandaa R.A., Gasol J.M., O'Gara F., Hernandez F., Labrenz M., Stoica E., Varela M. M., Bordalo A. and Pitta P. 2012. Marine microbial diversity and its role in ecosystem functioning and environmental change, Marine Board Position Paper 17, Calewaert, J.B and McDonough N (Eds.). Publisher: European Scientific Foundation (ESF), Ostend, Belgium, pp. 81, ISBN, 978- 2-918428-71-8.

Goldsmith D.B., Parsons R.J., Beyene D., Salamon P. & Breitbart M. 2015. Deep sequencing of the viral phoH gene reveals temporal variation, depth-specific composition, and persistent dominance of the same viral phoH genes in the Sargasso Sea. *PeerJ*., 3: e997.

Goldstein E., and Drlica K. 1984. Regulation of bacterial DNA supercoiling: plasmid linking numbers vary with growth temperature. *Proc. Natl. Acad. Sci., USA*, 81 (13): 4046-4050.

Goodwin K.D., Thompson L.R., Duarte B., Kahlke T., Thompson A.R., Marques J.C. and Caçador I. 2017. DNA Sequencing as a tool to monitor marine ecological status. *Front. Mar. Sci.*, 4:107. doi: 10.3389/fmars.2017.00107.

Gosalbes M.J., Durban A., Pignatelli M., Abellan J.J., Jimenez-Hernandez N., Perez-Cobas A.E., Latorre A. and Moya A. 2011. Metatranscriptomic approach to analyze the functional human gut microbiota. *Plos one*, 6: e17447.

Goudenege D., Labreuche Y., Krin E., Ansquer D., Mangenot S., Calteau A., Medigue C., Mazel D., Polz M.F. and Le Roux F. 2013. Comparative genomics of pathogenic lineages of *Vibrio nigripulchritudo* identifies virulence-associated traits. *Isme J.*, 7: 1985-1996.

Goudenège D., Travers M. A., Lemire A., Petton B., Haffner P., Labreuche Y. …and Nicolas J. L. 2015. A single regulatory gene is sufficient to alter *Vibrio aestuarianus* pathogenicity in oysters. *Environmental microbiology*, 17 (11): 4189-4199.

Gray M.D., Bagdasarian M., Hol W.G., Sandkvist M. 2011. In vivo cross linking of EpsG to EpsL suggests a role for EpsL as an ATPase pseudopilin coupling protein in the Type II secretion system of *Vibrio cholerae*. *Molecular microbiology*, 79 (3):786-98.

Green E. D., Rubin E. M. and Olson M. V.20170 The future of DNA sequencing. *Nature*, 550: 179-181.

Greenblatt C.L., Baum J., Klein B.Y., Nachshon S., Koltunov V. and Cano R. J. 2004. *Micrococcus luteus*-survival in amber. *Microbial ecology*, 48 (1): 120-127.

Grice E.A. and Segre J.A. 2011. The skin microbiome. *Nat. Rev. Microbiol*. 9: 244–253.

Gulig P.A., Tucker M.S., Thiaville P.C., Joseph J.L. and Brown R.N. 2009. USER friendly cloning coupled with chitin-based natural transformation enables rapid mutagenesis of *Vibrio vulnificus*. *Appl. Environ. Microbiol*., 75: 4936-4949.

Haas B.J., Kamoun S., Zody M.C., Jiang R.H.Y., Handsaker R.E., Cano L.M. et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*, 461 (7262): 393-398.

Hall M.N. and Silhavy T.J. 1981. The ompB locus and the regulation of the major outer membrane porin proteins of *Escherichia coli* K12. *J. Mol. Biol.*, 146 (1): 23-43.

Halpern B. S., Longo C., Hardy D., McLeod K. L., Samhouri J. F., Katona S. K.and Rosenberg A. A. 2012. An index to assess the health and benefits of the global ocean. *Nature*, 488 (7413): 615-620.

Halpern B.S. *et al.* 2008. A global map of human impact on marine ecosystems. *Science*, 319 (5865): 948-952.

Hämäläinen W. 2012. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. *Knowledge and Information Systems*, 32 (2): 383-414.

Hamasaki K., Taniguchi A., Tada Y., Kaneko R. and Miki T. 2016. Active populations of rare microbes in oceanic environments as revealed by bromodeoxyuridine incorporation and 454 tag sequencing. *Gene*, 576 (2): 650-656.

Han J., Pei J. and Yin Y. 2000. Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 1-12.

Handelsman J., Rondon M.R., Brady S.F., Clardy J. & Goodman R.M. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol*., 5: R245–249.

Hardy C.D. and Cozzarelli N.R. 2003. Alteration of *Escherichia coli* topoisomerase IV to novobiocin resistance. *Antimicrob. Agents Chemother*., 47: 941-947.

Harris S.R. *et al.* 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, 327 (5964): 469–474.

Harvell C. D., Kim K., Burkholder J. M., Colwell R. R., Epstein P. R., Grimes D. J. …and Porter J. W. 1999. Emerging marine diseases - climate links and anthropogenic factors. *Science*, 285(5433): 1505-1510.

Harvell C.D., Mitchell C.E., Ward J.R., Altizer S., Dobson A.P., Ostfeld R.S. and Samuel M.D. 2002. Climate warming and disease risks for terrestrial and marine biota. *Science*, 296: 2158-2162.

Hayes S., Mahony J., Nauta A. and van Sinderen D. 2017. Metagenomic approaches to assess bacteriophages in various environmental niches. *Viruses*, 9 (6): E127.

Hedgecock D., Shin G., Gracey A. Y., Van Den Berg D. and Samanta M. P. 2015. Second-generation linkage maps for the pacific oyster *Crassostrea gigas* reveal errors in assembly of genome scaffolds. *G3: Genes, Genomes, Genetics*, 5 (10): 2007-2019.

Heller R., Höller C., Süssmuth R. and Gundermann K.O. 1998. Effect of salt concentration and temperature on survival of *Legionella pneumophila*. *Letters in Applied Microbiology*, 26(1): 64-68.

Hensel M. 2000. Salmonella pathogenicity island 2. *Mol. Microbiol.*, 36(5): 1015-1023

Herrera F.C., Santos J.A., Otero A. and García-López M.L. 2006. Occurrence of *Plesiomonas shigelloides* in displayed portions of saltwater fish determined by a PCR assay based on the hugA gene. *International journal of food microbiology*, 108 (2): 233-238.

Higgins C.F., Dorman C.J., Stirling D.A., Waddell L., Booth I.R., May G. and Bremer E. 1988. A physiological role for DNA supercoiling in the osmotic regulation of gene expression in *S. typhimurium* and *E. coli*. *Cell*., 52 (4): 569-584.

Hjelms M.H., Hellmer M., Fernandez-Cassi X., Timoneda N., Lukjancenko O. …and Schultz A.C. 2017. Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing. *Plos one*, 12: e0170199.

Hodgkin J., Félix M.A., Clark L.C., Stroud D., Gravato-Nobre M.J. 2013. Two Leucobacter strains exert complementary virulence on *Caenorhabditis* including death by worm-star formation. *Current Biology*, 23 (21): 2156-2161.

Holt K.E. *et al*. 2015. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *PNAS*, 112 (27): 3574–81.

Hugoni M., Taib N., Debroas D., Domaizon I., Dufournel I.J., Bronner G. ... and Galand P.E. 2013. Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *PNAS*, 110 (15): 6004-6009.

Hunt D.E., David L.A., Gevers D., Preheim S.P., Alm E.J. and Polz M.F. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science*, 320: 1081-1085.

Hunt M., Mather A. E., Sánchez-Busó L., Page A. J., Parkhill J., Keane J. A. and Harris S. R. 2017. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *bioRxiv*, 118000.

Huttenhower C., Gevers D., Knight R., Abubucker S., Badger J.H. …and White O. 2012. Structure, function and diversity of the healthy human microbiome. *Nature*, 486: 207–214.

Inokuchi A., Washio T. and Motoda H. 2000. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In: Zighed, D.A., Komorowski, J. and Żytkow, J., editors, Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg; 13-23.

Ivanova E.P., Bowman J.P., Lysenko A.M., Zhukova N.V., Gorshkova N.M., Kuznetsova T.A., Kalinovskaya N.I., Shevchenko L.S. and Mikhail V.V. 2005. *Erythrobacter vulgaris sp. nov*., a novel organism isolated from the marine invertebrates. *Systematic and Applied Microbiology*, 28 (2): 123-130.

Jacobs J., Rhodes M., Sturgis B. and Wood B. 2009. Influence of environmental gradients on the abundance and distribution of *Mycobacterium spp*. in a coastal lagoon estuary. *Applied and Environmental Microbiology*, 75 (23): 7378-7384.

Jang J., Hur H. G., Sadowsky M.J., Byappanahalli M.N., Yan T., and Ishii S. 2017. Environmental *Escherichia coli*: Ecology and public health implications A Review. *Journal of Applied Microbiology*, 123: 570–581.

Jones K.E., Patel N.G., Levy M.A., Storeygard A., Balk D. and Gittleman J.L. 2008. Global trends in emerging infectious diseases. *Nature*, 451 (7181): 990-993.

Jones M. K., Oliver J. D. 2009. *Vibrio vulnificus*: disease and pathogenesis. *Infect. Immun*., 77: 1723–1733.

Jorge L., Bennasar A., Bosch R., Garcia-Valdes E., Palleroni N. 2006. Biology of *Pseudomonas stutzeri*. *Microbiology and Molecular Biology Reviews*, 70 (2): 510–547.

Karsenti E. The making of Tara Oceans: funding blue skies research for our Blue Planet.2015. *Mol. Syst. Biol.*, 11: 811.

Karsenti E., Acinas S. G., Bork P., Bowler C., De Vargas C., Raes J.and Follows M. 2011. A holistic approach to marine eco-systems biology. *Plos biology*, 9 (10): e1001177.

Karunasagar I., Karunasagar I., Venugopal M.N. and Nagesha C.N. 1987. Survival of *Vibrio parahaemolyticus* in estuarine and sea water and in association with clams. *Systematic and applied microbiology*, 9 (3): 316-319.

Kawano K., Okada M., Kura F., Amemura-Maekawa J. and Watanabe H. 2007. Largest outbreak of legionellosis associated with spa baths: comparison of diagnostic tests. *Kansenshogaku Zasshi*, 2: 173-182.

Kayis S., Capkin E., Balta F. and Altinok I. 2009. Bacteria in rainbow trout (*Oncorhynchus mykiss*) in the southern Black Sea Region of Turkey - A survey. *Isr. J. Aquacult. Bamid*., 61: 339–344.

Kayiş Ş., Er A., Kangel P. and Kurtoğlu İ.Z. 2017. Bacterial pathogens and health problems of *Acipenser gueldenstaedtii* and *Acipenser baerii* sturgeons reared in the eastern Black Sea region of Turkey, *Iranian Journal of Veterinary Research*, vol.18, pp.18-24, 2017.

Kazi M.I., Conrado A.R., Mey A.R., Payne S.M. and Davies B.W. 2016. ToxR antagonizes H-NS regulation of horizontally acquired genes to drive host colonization. *Plos Pathog*., 12 (4): e1005570.

Keep N.H., Ward J.M., Cohen-Gonsaud M. and Henderson B. 2006. Wake up! *Peptidoglycan lysis* and bacterial non-growth states. *Trends in microbiology*, 14 (6): 271-276.

Kellogg C.A., Griffin D.W. 2006. Aerobiology and the global transport of desert dust. *Trends Ecol. Evol.*, 21: 638-644.

Kerger B.D., Mancuso C.A., Nichols P.D., White D.C., Langworthy T., Sittig M., Schlesner H.and Hirsch P. 1988. The budding bacteria, *Pirellula* and *Planctomyces*, with atypical 16S rRNA and absence of peptidoglycan, show eubacterial phospholipids and uniquely high proportions of long chain beta-hydroxy fatty acids in the lipopolysaccharide lipid A. *Arch. Microbiol.*, 149: 255–260.

Kim J.H., Lee J.K., Yoo H.S., Shin N.R., Shin N.S., Lee K.H. and Kim D.Y. 2002. Endocarditis associated with *Escherichia coli* in a sea lion (*Zalophus californianus*). *Journal of veterinary diagnostic investigation*, 14 (3): 260-262.

Kim M.S., Cho J.Y. and Choi H.S. 2014. Identification of *Vibrio harveyi*, *Vibrio ichthyoenteri*, and *Photobacterium damselae* isolated from olive flounder *Paralichthys olivaceus* in Korea by multiplex PCR developed using the rpoB gene. *Fish Sci*., 80: 333–39.

Klein A.M., Bohannan B.J., Jaffe D.A., Levin D.A., Green J.L. 2016. Molecular evidence for metabolically active bacteria in the atmosphere. *Frontiers in microbiology*., 7.

Kobayashi F., Maki T. and Nakamura Y. 2012. Biodegradation of phenol in seawater using bacteria isolated from the intestinal contents of marine creatures. *International biodeterioration and biodegradation*, 69: 113-118.

Kolter R., Inuzuka M., and Helinski D.R. 1978. Trans-complementation-dependent replication of a low molecular weight origin fragment from plasmid R6K. *Cell*., 15: 1199-1208.

Komiyama J., Ishihata M., Arimura H., Nishibayashi T. and Minato S. 2017. Statistical Emerging Pattern Mining with Multiple Testing Correction. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 897-906.

Kopf A., Bicak M.and Kottmann R.2015. The ocean sampling day consortium. *GigaScience*, 4: 27. PubMed Abstract, Full Text, 1(2).

Kushmaro A., Banin E., Loya Y., Stackebrandt E. and E. Rosenberg. 2001. *Vibrio shiloi sp. nov.*, the causative agent of bleaching of the coral *Oculina patagonica*. *Int. J. Syst. Evol. Microbiol.*, 51 (4): 1383-1388.

Kwon K.K., Lee H.S., Jung S.Y., Yim J.H., Lee J.H. and Lee H.K. 2002. Isolation and identification of biofilm-forming marine bacteria on glass surfaces in Dae-Ho Dike, Korea. *Journal of Microbiology Seoul*, 40 (4): 260-266.

Labella A., Berbel C., Manchado M., Castro D. and Borrego J.J. 2011. *Photobacterium damselae subsp. damselae*, an emerging pathogen affecting new cultured marine fish species in southern Spain.1 35-152 In Recent Advances in Fish Farms. InTech.

Lafferty K.D. and Kuris A.M. 1993. Mass mortality of Abalone *Haliotis cracherodii* on the California Channel-Islands: tests of epidemiologic hypotheses. *Mar. Ecol. Prog. Ser.*, 96: 239–248.

Lasken R.S. 2012. Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.*, 10: 631–640.

Le Roux F., Binesse J., Saulnier D., and Mazel D. 2007. Construction of a *Vibrio splendidus* mutant lacking the metalloprotease gene vsm by use of a novel counterselectable suicide vector. *Appl. Environ. Microbiol.*, 73: 777-784.

Le Roux F., Davis B.M. and Waldor M.K. 2011a. Conserved small RNAs govern replication and incompatibility of a diverse new plasmid family from marine bacteria. *Nucleic Acids Res*., 39: 1004-1013.

Le Roux F., Labreuche Y., Davis B.M., Iqbal N., Mangenot S., Goarant C., Mazel D. and Waldor M.K. 2011b. Virulence of an emerging pathogenic lineage of *Vibrio nigripulchritudo* is dependent on two plasmids. *Environ. Microbiol.*, 13: 296-306.

Le Roux F., Wegner K. M., Baker-Austin C., Vezzulli L., Osorio C. R. Amaro, C. ... and Mazel D. 2015. The emergence of *Vibrio pathogens* in Europe: ecology, evolution, and pathogenesis (Paris, 11–12th March 2015). *Front. Microbiol*., 6: 830.

Le Roux F., Wegner K.M. and Polz M.F. 2016. Oysters and vibrios as a model for disease dynamics in wild animals. *Trends Microbiol*., 24 (7): 568-580.

Le Roux F., Zouine M., Chakroun N., Binesse J., Saulnier D., Bouchier C. ... and Buchrieser C. 2009. Genome sequence of *Vibrio splendidus*: an abundant planctonic marine species with a large genotypic diversity. *Environmental microbiology*, 11 (8): 1959-1970.

Lee A.K., Detweiler C.S., and Falkow S. 2000. OmpR regulates the two-component system SsrA-SsrB in Salmonella pathogenicity island 2. *J. Bacteriol*., 182 (3): 771-781.

Lee S.Y. and Eom Y.B. 2016. Analysis of microbial composition associated with freshwater and seawater. *Biomedical Science Letters*, 22 (4): 150-159.

Lee Y.J., van Nostrand J.D., Tu Q., Lu Z., Cheng L., Yuan T., Deng Y., Carter M.Q., He Z., Wu L., Yang F., Xu J., Zhou J.. 2013. The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities. *Isme J.*, 7 (10): 1974-84.

Lees J.A. *et al.* 2016. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature communications*, 7: 12797.

Lemire A., Goudenege D., Versigny T., Petton B., Calteau A., Labreuche Y. and Le Roux F. 2014. Populations, not clones, are the unit of vibrio pathogenesis in naturally infected oysters. *Isme J.*, 9: 1523-1531.

Lennon J.T. and Jones S.E. 2011. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology*, 9 (2): 119-130.

Lessios H.A., Robertson D.R. and Cubit J.D. 1984. Spread of *Diadema antillarum* mass mortality through the Caribbean. *Science*, 226: 335–337.

Li L., Stoeckert C. J. and Roos, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13 (9): 2178-2189.

Lima-Mendez G., Faust K., Henry N., Decelle J., Colin S., Carcillo F. and Bittner, L. 2015. Determinants of community structure in the global plankton interactome. *Science*, 348 (6237): 1262073.

Lindboe C. 2001. The prevalence of human intestinal spirochetosis in Norway. *Animal Health Research Reviews*, 2 (1): 117-120.

Lipp E.K., Huq A., Colwell R.R. 2002. Effects of global climate on infectious disease: the cholera model. *Clin. Microbiol. Rev.*, 15: 757-770.

Liu L.F. and Wang J.C. 1987. Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci.*, USA 84: 7024-7027.

Llinares-López F., Grimm D. G., Bodenham D. A., Gieraths U., Sugiyama M., Rowan B. and Borgwardt K. M. 2015. Genome-Wide Detection of Intervals of Genetic Heterogeneity Associated with Complex Traits. *Bioinformatics*, 31 (12): i240-i249.

Llinares-López F., Papaxanthos L., Bodenham D. A., Roqueiro D., Investigators COPDGene and Borgwardt K. M. 2017. Genome-wide genetic heterogeneity discovery with categorical covariates. *Bioinformatics*, 33 (12): 1820–1828.

Llinares-López F., Sugiyama M., Papaxanthos L. and Borgwardt K. M. 2015. Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing. Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 725-734.

Lloyd-Price J., Abu-Ali G. & Huttenhower C. 2016. The healthy human microbiome. *Genome Med.*, 8: 51.

Lo Presti F., Riffard S., Meugnier H., Reyrolle M., Lasne Y., Grimont P.A.D., Grimont F., Vandenesch F., Etienne J., Fleurette J. and Freney J. 1999. *Legionella taurinensis sp. nov.*, a new species antigenically similar to *Legionella spiritensis. Int. J. Syst. Bacteriol.*, 49: 397–403.

Lokmer A. and M. Wegner K. 2015. Hemolymph microbiome of Pacific oysters in response to temperature, temperature stress and infection. *Isme J.*, 9: 670-682.

Lokmer A., Kuenzel S., Baines J.F. and Wegner K.M. 2016. The role of tissue-specific microbiota in initial establishment success of Pacific oysters. *Environ. Microbiol.*, 18: 970-987.

Lu Y., Wang R., Zhang Y., Su H., Wang P., Jenkins A., Ferrier R.C., Bailey M. and G. Squire . 2015. Ecosystem health towards sustainability. Ecosystem Health and Sustainability, 1 (1): 2. http://dx.doi.org/10.1890/EHS14-0013.1.

Lucchini S., Rowley G., Goldberg M.D., Hurd D., Harrison M. and Hinton JC. 2006. H-NS mediates the silencing of laterally acquired genes in bacteria. *Plos Pathog*., 2 (8): e81.

Lulchev P. and Klostermeier D. 2014. Reverse gyrase: recent advances and current mechanistic understanding of positive DNA supercoiling. *Nucleic Acids Res*., 42 (13): 8200-8213.

Luna G.M., Quero G.M. and Perini L. 2016. Next generation sequencing reveals distinct fecal pollution signatures in aquatic sediments across gradients of anthropogenic influence. *Advances in Oceanography and Limnology*, 7 (2).

Lux R. and W. Shi. 2004 Chemotaxis-guided movements in bacteria. *Crit. Rev. Oral. Biol. Med*., 15: 207-220.

Lydyard P., Cole M., Holton J., Irving W., Porakishvili N., Venkatesan P. and Ward K. Case Studies in Infectious Disease. Garland Science Textbook, 2009.

Lynch M.D. and Neufeld J.D. 2015. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, 13 (4): 217-229.

Ma J., Bai L. and Wang M.D. 2013. Transcription under torsion. *Science*, 340:1580–1583.

Machado H., Cardoso J., Giubergia S., Rapacki K., Gram L. 2017. FurIOS: a web-based tool for identification of Vibrionaceae species using the fur gene. *Frontiers in microbiology*, 8.

Machado H., Gram L. 2015. The fur gene as a new phylogenetic marker for Vibrionaceae species identification. *Applied and environmental microbiology*, 81 (8): 2745-52.

MacInnes J., Robertson P.A.W. and Austin B. 2002. A comparison of the bacterial microflora between coastal sites in Qingdao, PR China and Loch Fyne, Scotland. *Journal of Ocean University of Qingdao*, 1 (2): 148-152.

MacLellan S.L. and Eren A.M. 2014. Discovering new indicators of fecal pollution. *Trends in microbiology*, 22 (12): 697-706.

Macnab R.M. 1987. Motility and chemotaxis. pp. 732-759. In Escherichia coli and *Salmonella typhimurium*. J. Ingrham, K.B. Low, B. Magasanik, M. Schaechter, H.E. Umbarger, and F.C. Neidhardt, eds. American Society for Microbiology, Washington, DC.

Maki T., Yoshinaga I., Katanozaka N., and Imai I. 2004. Phylogenetic analysis of intracellular bacteria of a harmful marine microalga, *Heterocapsa circularisquama* (Dinophyceae). *Aquatic microbial ecology*, 36 (2): 123-135.

Marcy Y., Ouverney C., Bik E.M., Losekann T., Ivanova N., Martin H.G., Szeto E., Platt D., Hugenholtz P., Relman D.A. & Quake S.R. 2007. Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci. U.S.A*, 104: 11889–11894.

Margolin P., Zumstein L., Sternglanz R., Wang J.C. 1985. The *Escherichia coli* supX locus is topA, the structural gene for DNA topoisomerase I. *Proc. Natl. Acad. Sci., USA*, 82: 5437-5441.

Marques A., Ollevier F., Verstraete W., Sorgeloos P. and Bossier P. 2006. Gnotobiotically grown aquatic animals: opportunities to investigate host-microbe interactions. *J. Appl. Microbiol*., 100: 903-918.

Marrie J. Thomas. 1990. “Q fever - a review”. *Canadian Veterinary Journal*, 31 (8): 555-563.

Martenot C., Oden E., Travaille E., Malas J.P. and Houssin M. 2011. Detection of different variants of Ostreid herpesvirus 1 in the Pacific oyster *Crassostrea gigas* between 2008 and 2010. *Virus Res*., 160: 25-31.

Martínez-Hackert E. and Stock A.M. 1997. The DNA-binding domain of OmpR: crystal structures of a winged helix transcription factor. *Structure*, 5 (1): 109-124.

Martinez-Hernandez F., Fornas O., Lluesma Gomez M., Bolduc B., Cruz Pena M.J. and Martinez-Garcia M. 2017. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun*., 8: 15892.

Martinez-Urtaza J., Bowers J.C., Trinanes J., DePaola A. 2010. Climate anomalies and the increasing risk of *Vibrio parahaemolyticus* and *Vibrio vulnificus* illnesses. *Food Res. Int*., 43: 1780-1790.

Martinez-Urtaza, J. *et al.* 2016. Is El Niño a long-distance corridor for waterborne disease? *Nature microbiology*, 1: 16018.

Martins F., Dalmasso G., Arantes R. M. E., Doye A., Lemichez E., Lagadec P., Imbert V., Peyron J.F., Rampal P., Nicoli J.R. and D. Czerucka. 2010. Interaction of *Saccharomyces boulardii* with *Salmonella enterica serovar Typhimurium* protects mice and modifies T84 cell response to the infection. *Plos one*, 5: e8925.

Marvig R.L. and Blokesch M. 2010. Natural transformation of *Vibrio cholerae* as a tool-optimizing the procedure. *BMC Microbiol*., 10: 155.

Mattioli M.C., Sassoubre L.M., Russell T.L. and Boehm A.B. 2017. Decay of sewage-sourced microbial source tracking markers and fecal indicator bacteria in marine waters. *Water research*, 108: 106-114.

Maugeri T.L., Carbone M., Fera M.T. and Gugliandolo C. 2006. Detection and differentiation of *Vibrio vulnificus* in seawater and plankton of a coastal zone of the Mediterranean Sea. *Research in microbiology*, 156 (2): 194-200.

Maugeri T.L., Carbone M., Fera M.T., Irrera G.P. and Gugliandolo C. 2004. Distribution of potentially pathogenic bacteria as free living and plankton associated in a marine coastal zone. *Journal of applied microbiology*, 97 (2): 354-361.

McArthur, A.G. et al., 2013. The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, 57 (7): 3348–3357.

McCallum Harvell and Dobson. 2003. Rates of spread of marine pathogens. *Ecology Letters*, 6 (12): 1062-1067.

MCCIP 2010. Marine Climate Change Impacts Annual Report Card 2010-2011. (Eds Baxter JM, Buckley PJ, and Wallace, CJ) Summary Report, MCCIP, Lowestoft, 12pp.

McGuckin M.A., Liden S.K., Sutton P. and T.H. Florin. 2011. Mucin dynamics and enteric pathogens. *Nature*, 9: 265-277.

McIntyre K. M., Setzkorn C., Hepworth P.J., Morand S., Morse A.P. and Baylis M. 2017. Systematic assessment of the climate sensitivity of important human and domestic animal pathogens in Europe. *Sci. Rep*., 7 (1):7134. doi: 10.1038/s41598-017-06948-9.

McNally A. *et al.* 2016. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *Plos genetics*, 12 (9): p.e1006280.

Medini D. *et al.* 2005. The microbial pan-genome. Current opinion in genetics & development, 15(6), pp.589–594.

Metzger D.C., Schulte P.M. 2016. Epigenomics in marine fishes. *Marine genomics*, 30: 43-54.

Miethke M., Marahiel M.A. 2007. Siderophore-based iron acquisition and pathogen control. *Microbiol Mol Biol Rev*., 71 (3): 413-51.

Milkman R. 1973. Electrophoretic variation in *Escherichia coli* from natural sources. *Science*, 182 (4116): 1024–1026.

Miller V.L. and Mekalanos J.J. 1988. A novel suicide vector and its use in construction of insertion mutations: osmoregulation of outer membrane proteins and virulence determinants in *Vibrio cholerae* requires toxR. *Journal of Bacteriology*, 170: 2575-2583.

Mizuno C.M., Rodriguez-Valera F., Kimes N.E. and Ghai R. 2013. Expanding the marine virosphere using metagenomics. *Plos Genet*., 9: e1003987.

Mizunoe Y., Wai S.N., Ishikawa T., Takade A., and Yoshida S.I. 2000. Resuscitation of viable but nonculturable cells of *Vibrio parahaemolyticus* induced at low temperature under starvation. *FEMS Microbiology Letters*, 186 (1): 115-120.

Morand S., McIntyre K. M. and Baylis M. 2014. Domesticated animals and human infectious diseases of zoonotic origins: domestication time matters. *Infect. Genet. Evol*., 24:76-81.

Morens D.M., Folkers G.K. and Fauci A.S. 2004. The challenge of emerging and re-emerging infectious diseases. *Nature*, 430 (6996): 242-249.

Mouillot D., Bellwood D.R., Baraloto C., Chave J., Galzin R., Harmelin-Vivien M. *et al.* 2013. Rare species support vulnerable functions in high-diversity ecosystems. *Plos Biol*., 11: e1001569.

Munn C.B. 2005. Pathogens in the sea: an overview. Oceans and Health: *Pathogens in the Marine Environment*, 1-28.

Murray A. G., Wardeh M. and McIntyre, K. M. 2016. Using the H-index to assess disease priorities for salmon aquaculture. *Prev. Vet. Med.*, 126: 199-207.

O Cróinín T., Carroll R.K., Kelly A. and Dorman C.J. 2006. Roles for DNA supercoiling and the Fis protein in modulating expression of virulence genes during intracellular growth of *Salmonella enterica serovar Typhimurium*. *Mol. Microbiol.*, 62 (3): 869-882.

Oh J.Y., Jeong Y.W., Joo H.S., Chong W.S., Lee J.C., Tamang M.D. ... and Park J.C. 2009. Distribution of genomic species and antimicrobial susceptibility in Acinetobacters isolated from Gangjin Bay, Korea. *Journal of Bacteriology and Virology*, 39 (4): 247-256.

O'Higgins T., A. Farmer, G. Daskalov, S. Knudsen, and L. Mee. 2014. Achieving good environmental status in the Black Sea: scale mismatches in environmental management. *Ecology and Society*, 19 (3): 54.

Oliver J.D., Pruzzo C., Vezzulli L., Kaper J.B. 2013. "Vibrio Species," Food Microbiology: Fundamentals and Frontiers, 4th Ed. M. P. Doyle and R. L. Buchanan Eds. ASM Press, Washington, D.C. doi:10.1128/9781555818463.ch16.

Opsahl T., Agneessens A. and Skvoretz J. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32 (3): 245-251.

Orruño M., Kaberdin V.R. and Arana I. 2017. Survival strategies of Escherichia coli and *Vibrio spp.*: contribution of the viable but nonculturable phenotype to their stress-resistance and persistence in adverse environments. *World Journal of Microbiology and Biotechnology*, 33 (3): 45.

Oztürk R.C. and Altinok I. 2014. Bacterial and viral fish diseases in Turkey. *Turk. J. Fish Aquat. Sc.*, 14: 275–297.

Paez-Espino D., Eloe-Fadrosh E.A., Pavlopoulos G.A., Thomas A.D., Huntemann M., Mikhailova N., Rubin E., Ivanova N.N. & Kyrpides N.C. 2016. Uncovering Earth's virome. *Nature*, 536: 425–430.

Page A.J. *et al.* 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31 (22): 3691–3693.

Pánková I., Krejzar V., Sedlák P. and Sedláková V. 2012. The Occurrence of Plant. Pathogenic Streptomyces spp. in Potato-growing Regions in Central Europe. *Am. J. Pot. Res*., 89: 207–215.

Papaxanthos L., Llinares-Lopez F., Bodenham D. and Borgwardt K. M. 2016. Finding significant combinations of features in the presence of categorical covariates. *Advances in Neural Information Processing Systems*, 2271-2279.

Parker J.L., Shaw J.G. 2011. *Aeromonas spp*. clinical microbiology and disease. *Journal of Infection*, 62 (2): 109-118.

Parsot C., and Mekalanos J.J. 1992. Structural analysis of the acfA and acfD genes of *Vibrio cholerae*: effects of DNA topology and transcriptional activators on expression. *J. Bacteriol*., 174 (16): 5211-5218.

Pascual M., Rodó X., Ellner S.P., Colwell R.R., Bouma M.J. 2000. Cholera Dynamics and El Niño-Southern Oscillation. *Science*, 289: 1766-1769.

Pavlovska M., Stoica E., E. Dykyi, D. Zasko, V. Ilyinsky 2016. Chap. II.6 - Bacterioplankton, in: EMBLAS-II National Pilot Monitoring Studies (NPMS) and Joint Open Sea Surveys in Georgia, Russian Federation and Ukraine (JOSS) Scientific Report, Editors: J. Slobodnik, B. Alexandrov, V. Komorin, A. Mikaelyan, A. Guchmanidze, M. Arabidze, A. Korshenko.

Pedrós-Alió C. 2012. The rare bacterial biosphere. *Annual review of marine science*, 4: 449-466.

Pei J., Han J., Mortazavi-Asl B., Pinto H., Chen Q., Dayal U. and Hsu M. C. 2001. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. *Proceedings of the 17th International Conference on Data Engineering*, 215-224.

Pereira F.C. and Berry D. 2017. Microbial nutrient niches in the gut. *Environ. Microbiol*. 19: 1366–1378.

Pereira R.P., Peplies J., Brettar I., Höfle M.G. 2017. Development of a genus-specific next generation sequencing approach for sensitive and quantitative determination of the Legionella microbiome in freshwater systems. *BMC Microbiol*, 17(1): 79.

Peris-Bondia F., Latorre A., Artacho A., Moya A. and D'Auria G. 2011. The active human gut microbiota differs from the total microbiota. *Plos one*. 6: e22448.

Peris-Bondia F., Latorre A., Artacho A., Moya A., D'Auria G. 2011. The active human gut microbiota differs from the total microbiota. *PloS one*, 6 (7): e22448.

Pesant S., Not F., Picheral M., Kandels-Lewis S., Le Bescot N., Gorsky G., Iudicone D., Karsenti E., Speich S., Troublé R., Dimier C., Searson S. 2015. Tara Oceans Consortium Coordinators. Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data*, 2: 150023.

Petton B., Bruto M., James A., Labreuche Y., Alunno-Brusci M. and Le Roux F. 2015. *Crassostrea gigas* mortality in France: the usual suspect, a herpes virus, may not be the killer in this polymicrobial opportunistic disease. *Front. Microbiol.*, 6: 686.

Petton B., Pernet F., Robert R. and Boudry P. 2013a. Temperature influence on pathogen transmission and subsequent mortalities in juvenile Pacific oysters *Crassostrea gigas*. *Aquacult. Environ. Interact*., 3: 257-273.

Petton B., Pernet F., Robert R. and Boudry P. 2013b. Temperature influence on pathogen transmission and subsequent mortalities in juvenile Pacific oysters *Crassostrea gigas*. *Aquacult. Environ. Interact.*, 3: 257-273.

Plank R. and Dean D. 2000. Overview of the epidemiology, microbiology, and pathogenesis of *Leptospira spp*. in humans, *In Microbes and Infection*, 2(10): 1265-1276.

Podar M., Abulencia C.B., Walcher M., Hutchison D., Zengler K., Garcia J.A., Holland T., Cotton D., Hauser L. & Keller M. 2007. Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol*. 73: 3205–3214.

Pollack-Berti A., Wollenberg M.S. and Ruby E.G. 2010. Natural transformation of *Vibrio fischeri* requires tfoX and tfoY. *Environ. Microbiol*., 12: 2302-2311.

Pompeani A.J., Irgon J.J., Berger M.F., Bulyk M.L., Wingreen N.S., Bassler B.L. 2008. The *Vibrio harveyi* master quorum sensing regulator, LuxR, a TetR type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters. *Molecular microbiology*, 70 (1): 76-88.

Pontier-Bres R., Prodon F., Munro P., Rampal P., Lemichez E., Peyron J-F, and D. Czerucka. 2012. Modification of *Salmonella typhimurium* motility by the probiotic yeast strain *Saccharomyces boulardii*. *Plos one*, 7(3): e33796.

Porter J., Diaper J., Edwards C. and Pickup R. 1995. Direct measurements of natural planktonic bacterial community viability by flow cytometry. *Appl. Environ. Microbiol*., 61: 2783–2786.

Preheim S.P., Boucher Y., Wildschutte H., David L.A., Veneziano D., Alm E.J. and Polz M.F. 2011. Metapopulation structure of Vibrionaceae among coastal marine invertebrates. *Environ. Microbiol*., 13: 265-275.

Pruzzo C., Vezzulli L., Colwell R.R. 2008. Global impact of *Vibrio cholerae* interactions with chitin. *Environ. Microbiol*., 10 (6): 1400–1410.

Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., Glöckner F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*., 41(Database issue): D590-6.

Quinn H.J., Cameron A.D. and Dorman C.J. 2014. Bacterial regulon evolution: distinct responses and roles for the identical OmpR proteins of *Salmonella typhimurium* and *Escherichia coli* in the acid stress response. *Plos Genet*., 10 (3): e1004215.

Rabinowitz D. 1981. Seven forms of rarity. *The biological aspects of rare plants conservation*, 205-217.

Radchenko V.I. 2007. Mesopelagic fish community supplies biological pump. *Raffles Bull. Zool. Suppl.*, 14: 265–271.

Read T.D. and Massey R.C. 2014. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome medicine*, 6 (11): 109.

Reid P.C., Colebrook J.M., Matthews J.B.L., Aiken J., Continuous Plankton Recorder Team 2003. The Continuous Plankton Recorder: concepts and history, from Plankton Indicator to undulating recorders. *Prog. Oceanogr*., 58: 117-173.

Rhen M. and Dorman C.J. 2005. Hierarchical gene regulators adapt *Salmonella enterica* to its host milieus. *Int. J. Med. Microbiol*., 294: 487-502.

Rice P., Longden I. and Bleasby A. 2000. "EMBOSS: the European molecular biology open software suite", 276-277.

Rinke C., Lee J., Nath N., Goudeau D., Thompson B., Poulton N., Dmitrieff E., Malmstrom R., Stepanauskas R. & Woyke T. 2014. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat. Protoc*., 9: 1038–1048.

Rinke C., Schwientek P., Sczyrba A., Ivanova N.N., Anderson I.J. …and Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499: 431–437.

Rivas A.J., Lemos M.L. and Osorio C.R. 2013. *Photobacterium damselae subsp. damselae*, a bacterium pathogenic for marine animals and humans. *Frontiers in microbiology*, 4: 283.

Roca Alberto I. and Aaron C. Abajian. 2011. Improvements to the JProfileGrid software for visualizing very large multiple sequence alignments. *The FASEB Journal*, 25 (1): 924-6.

Rohs R., West S.M., Sosinsky A., Liu P., Mann R.S. and Honig B. 2009. The role of DNA shape in protein-DNA recognition. *Nature*, 461 (7268): 1248-1253.

Rohwer F., Seguritan V., Azam F., Knowlton N. 2002. Diversity and distribution of coral-associated bacteria. *Mar. Ecol. Prog. Ser.*, 243: 1-10.

Romanowicz K.J., Freedman Z.B., Upchurch R.A., Argiroff W.A. and Zak D.R. 2016. Active microorganisms in forest soils differ from the total community yet are shaped by the same environmental factors: the influence of pH and soil moisture. *FEMS Microbiol. Ecol*., 92(10).

Rossello-Mora R. and Amann R. 2001. The species concept for prokaryotes. FEMS Microbiol. Rev., 25: 39–67.

Rusch D. B., Halpern A. L., Sutton G., Heidelberg K. B., Williamson S., Yooseph S.and Beeson K. 2007. The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology*, 5 (3): e77.

Saha R.P., Chakrabarti P. 2006. Molecular modeling and characterization of *Vibrio cholerae* transcription regulator HlyU. *BMC structural biology*, 6 (1): 24.

Sakata J., Yonekita T., Kawatsu K. 2017. Development of a rapid immunochromatographic assay to detect contamination of raw oysters with enteropathogenic *Vibrio parahaemolyticus. International Journal of Food Microbiology*, 264: 16-24.

Salter S.J., Cox M.J., Turek E.M., Calus S.T., Cookson W.O., Moffatt M.F., Turner P., Parkhill J., Loman N.J., Walker A.W. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*., 12:87.

Santiago-Vázquez L.Z., Brück T.B., Brück W.M., Duque-Alarcon A.P., McCarthy P.J. and Kerr R.G. 2007. The diversity of the bacterial communities associated with the azooxanthellate hexacoral *Cirrhipathes lutkeni*. *The ISME journal*, 1 (7): 654-659.

Schaus M.H. and Vanni M.J. 2000. Effects of gizzard shad on phytoplankton and nutrient dynamics: role of sediment feeding and fish size. *Ecology*, 81: 1701–1719.

Schmidt T.M., DeLong E.F. and Pace N.R. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol*., 173: 4371–4378.

Schmitt P., Gueguen Y., Desmarais E., Bachere E. and de Lorgeril J. 2010a. Molecular diversity of antimicrobial effectors in the oyster *Crassostrea gigas*. *BMC Evol. Biol.,* 10: 23.

Schmitt P., Wilmes M., Pugniere M., Aumelas A., Bachere E., Sahl H.G., Schneider T. and Destoumieux-Garzon D. 2010b. Insight into invertebrate defensin mechanism of action: oyster defensins inhibit peptidoglycan biosynthesis by binding to lipid II. *J. Biol. Chem.,* 285: 29208-29216.

Schulze A.D., Alabi A.O., Tattersall-Sheldrake A.R. and Miller K.M. 2006. Bacterial diversity in a marine hatchery: balance between pathogenic and potentially probiotic bacterial strains. *Aquaculture*, 256 (1): 50-73.

Shapiro B.J. and Polz M.F. 2015. Microbial Speciation. *Cold Spring Harb Perspect. Biol*., 7: a018143.

Shapiro B.J., Friedman J., Cordero O.X., Preheim S.P., Timberlake S.C., Szabo G., Polz M.F. and Alm E.J. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336: 48-51.

Sharma K.K., Kalawat U. 2010. Emerging infections: Shewanella - a series of five cases. *J. Lab. Physicians*, (2): 61–65.

Shea J.E., Hensel M., Gleeson C. and Holden D.W. 1996. Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*. *Proc. Natl. Acad. Sci., USA*, 93 (6): 2593-2597.

Sievers F., Wilm A., Dineen D., Gibson T. J., Karplus K., Li, W …and Thompson J. D. 2011. Fast, scalable generation of high quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7 (1): 539.

Simon R., Priefer U.B. and Puhler A. 1983. A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in Gram negative bacteria. *Bio/Technology,* 1: 784-791.

Simon-Soro A., D'Auria G., Collado M.C., Džunková M., Culshaw S. and Mira A. 2015. Revealing microbial recognition by specific antibodies. *BMC Microbiol*., 15: 132.

Skopina M.Y., Vasileva A.A., Pershina E.V. and Pinevich A.V. 2016. Diversity at low abundance: The phenomenon of the rare bacterial biosphere. *Microbiology*, 85 (3): 272-282.

Smith K.F, Sax D.F. and Laffetry K.D. 2006. Evidence for the role of infectious disease in species extinction and endangerment. *J. Cons. Biol*., 20: 1349-1357.

Sneha K.G., Anas A., Jayalakshmy K.V., Jasmin C., VipinDas P.V., Pai S.S. ... and Nair S. 2016. Distribution of multiple antibiotic resistant *Vibrio spp* across Palk Bay. *Regional Studies in Marine Science*, 3: 242-250.

Snoep J.L., van der Weijden C.C., Andersen H.W., Westerhoff H.V. and Jensen P.R. 2002. DNA supercoiling in *Escherichia coli* is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase I and DNA gyrase. *Eur. J. Biochem.,* 269: 1662–1669

Sogin M.L., Morrison H.G., Huber J.A., Welch D.M., Huse S.M., Neal P.R. ... and Herndl G.J. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *PNAS*, 103 (32): 12115-12120.

Sotello D, Hata DJ, Reza M, Satyanarayana R, Arunthari V, Bosch W. 2017. Disseminated *Mycobacterium interjectum* infection with bacteremia, hepatic and pulmonary involvement associated with a long-term catheter infection. *Case Reports in Infectious Diseases*, 2017: 6958204.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30 (9): 1312-1313.

Stecher B., Hapfelmeier S., Müller C., Kremer M., Stallmach T and H. Wolf-Dietrich. 2004. Flagella and chemotaxis are required for efficient induction of *Salmonella enterica Serovar Typhimurium colitis* in mice. *Infect. Immun*., 72: 4138-4150.

Stephens P.R., Pappalardo P., Huang S., Byers J.E., Farrell M.J., Gehman A., Ghai R.R., Haas S.E., Han B., Park A.W., Schmidt J.P., Altizer S., Ezenwa V.O. and Nunn C.L. 2017. Global mammal parasite database version 2.0. *Ecology*, 98 (5): 1476.

Sternglanz R., DiNardo S., Voelkel K.A., Nishimura Y., Hirota Y., Becherer K., Zumstein L., Wang J.C. 1981. Mutations in the gene coding for *Escherichia coli* DNA topoisomerase I affect transcription and transposition. *Proc. Natl. Acad. Sci., USA*, 78: 2747-2751.

Stewart J.R., Gast R.J., Fujioka R.S., Solo-Gabriele H.M., Meschke J.S., Amaral-Zettler L.A., Del Castillo E., Polz M.F., Collier T.K., Strom M.S., Sinigalliano C.D., Moeller P.D.R., Holland A.F. 2008. The coastal environment and human health: microbial indicators, pathogens, sentinels and reservoirs. *Environ Health*, 7: 1476–1069X.

Stincone A., Daudi N., Rahman A.S., Antczak P., Henderson I., Cole J., Johnson M.D., Lund P. and Falciani F. 2011. A systems biology approach sheds new light on *Escherichia coli* acid resistance. *Nucleic Acids Res.*, 39 (17): 7512-7528.

Stocker R. and J.R. Seymour. 2012. Ecology and physics of bacterial chemotaxis in the ocean. *MMBR*, 76: 792-812.

Stoebel D.M., Free A. and Dorman C.J. 2008. Anti-silencing: overcoming H-NS-mediated repression of transcription in Gram-negative enteric bacteria. *Microbiology*, 154: 2533-2545.

Sugiyama M., Llinares-López F., Kasenburg N. and Borgwardt K. M. 2015. Significant Subgraph Mining with Multiple Testing Correction. Proceedings of 2015 SIAM International Conference on Data Mining, 37-45.

Sunagawa S., Coelho L. P., Chaffron S., Kultima J. R., Labadie K., Salazar G. and Cornejo-Castillo F. M. 2015a. Structure and function of the global ocean microbiome. *Science*, 348 (6237): 1261359.

Sunagawa S., Karsenti E., Bowler C. and Bork P.2015b. Computational eco-systems biology in Tara Oceans: translating data into knowledge. *Mol. Syst. Biol.*, 11: 809.

Sunagawa S.L.P., Coelho S., Chaffron J.R., Kultima K., Labadie G., Salazar B., Djahanschiri et al., 2015. Structure and function of the global ocean microbiome. Science, 348(6237): 1261359–1261359.

Sussarellu R., Huvet A., Lapegue S., Quillen V., Lelong C., Cornette F., Jensen L.F., Bierne N. and Boudry P. 2015. Additive transcriptomic variation associated with reproductive traits suggest local adaptation in a recently settled population of the Pacific oyster, *Crassostrea gigas*. *BMC Genomics*, 16: 808.

Sussman M., Loya Y., Fine M. and Rosenberg E. 2003. The marine fireworm *Hermodice carunculata* is a winter reservoir and spring summer vector for the coral bleaching pathogen *Vibrio shiloi*. *Environmental Microbiology*, 5 (4): 250-255.

Sussman M., Loya Y., Fine M. and Rosenberg E. 2003. The marine fireworm *Hermodice carunculata* is a winter reservoir and spring summer vector for the coral bleaching pathogen *Vibrio shiloi*. *Environmental Microbiology*, 5 (4): 250-255.

Szabó K.É., Itor P.O., Bertilsson S., Tranvik L. and Eiler A. 2007. Importance of rare and abundant populations for the structure and functional potential of freshwater bacterial communities. *Aquatic microbial ecology*, 47 (1): 1-10.

Takemura A.F., Corzett C.H., Hussain F., Arevalo P., Datta M., Yu X., Le Roux F. and Polz M.F. 2017. Natural resource landscapes of a marine bacterium reveal distinct fitness-determining genes across the genome. *Environ. Microbiol.*, 19: 2422-2433.

Takigawa I. and Mamitsuka H. 2013. Graph mining: procedure, application to drug discovery and recent advances. *Drug Discovery Today*, 18(1–2): 50-57.

Tan B., Ng C., Nshimyimana J.P., Loh L.L., Gin K.Y.H. and Thompson J.R. 2015. Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Front. Microbiol.*, 6: 1027. doi: 10.3389/fmicb.2015.01027.

Tanca A., Abbondio M., Palomba A., Fraumene C., Manghina V., Cucca F., Fiorillo E. and Uzzau S. 2017. Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome*. 5: 79.

Tang G.Q., Tanaka N. and Kunugi S. 1998. In vitro increases in plasmid DNA supercoiling by hydrostatic pressure. *Biochim. Biophys. Acta.*, 1443 (3): 364-368.

Tarone R. E. 1990. A modified Bonferroni method for discrete data. *Biometrics*, 46 (2): 515-522.

Taylor L.H., Latham S.M. and Woolhouse M.E.J. 2001. Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 356 (1411): 983-989.

Tedersoo L., Bahram M., Põlme S., Kõljalg U., Yorou N. S., Wijesundera R., ... and Smith M.E. 2014. Global diversity and geography of soil fungi. *Science*, 346 (6213): 1256688.

Ter Steege H., Pitman N. C., Sabatier D., Baraloto C., Salomão R. P., Guevara J. E., ... and Monteagudo A. 2013. Hyperdominance in the Amazonian tree flora. *Science*, 342 (6156): 1243092.

Terada A., duVerle D. and Tsuda K. 2016. Significant Pattern Mining with Confounding Variables. *Advances in Knowledge Discovery and Data Mining*, 277-289.

Terada A., Okada-Hatakeyama M., Tsuda K. and Sese J. 2013. Statistical significance of combinatorial regulations. *Proc. Natl. Acad. Sci. USA*, 110 (32): 12996-13001.

Terada A., Tsuda K. and Sese J. 2013. Fast Westfall-Young permutation procedure for combinatorial regulation discovery. 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 153-158.

Terada A., Yamada R., Tsuda K. and Sese J. 2016. LAMPLINK: detection of statistically significant SNP combinations from GWAS data. *Bioinformatics*, 32 (22): 3513-3515.

Terashima H., Kojima S., and M. Homma. 2008. Flagellar motility in bacteria: structure and function of flagellar motor. *Int Rev Cell Mol Biol*, 270 : 39-85.

Terceti M.S., Ogut H. and Osorio C.R. 2016. *Photobacterium damselae subsp. damselae*, an emerging fish pathogen in the black sea: evidence of a multiclonal origin. *Applied and environmental microbiology*, 82 (13): 3736-3745.

Thinesh T., Mathews G. and Edward J.K. 2011. Coral disease prevalence in the Palk Bay, Southeastern India–with special emphasis to black band. *Indian J. Geo. Mar. Sci*., 40: 813-820.

Thomas T., Gilbert J., Meyer F. 2012. Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp*., 2 (1): 3. doi:10.1186/2042-5783-2-3.

Thompson F. L., Li Y., Gomez-Gil B., Thompson C. C., Hoste B., Vandemeulebroecke K. et al. 2003. *Vibrio neptunius sp. nov*., *Vibrio brasiliensis sp. nov*. and *Vibrio xuii sp. nov*., isolated from the marine aquaculture environment (bivalves, fish, rotifers and shrimps). *Int. J. Syst. Evol. Microbiol*., 53: 245–252.

Thomsen R.N., Kristiansen M.M. 2001. Three cases of bacteraemia caused by *Aeromonas veronii biovar sobria*. *Scandinavian Journal of Infectious Diseases*, 33: 718-719.

Thorpe H. A., Bayliss S. C., Hurst L. D., and Feil E. J. 2017. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*, 206(1): 363-376.

Tompkins D.M, Arneberg P., Begon M.E., Cattadori I.M., Greenman J.V., Heesterbeek, J.A.P., Hudson, P.J., Newborn D., Pugliese A., Rizzoli A.P., and Rosa R. 2001. Parasites and host population dynamics pp. 45-62. In the ecology of wildlife diseases [Hudson, P.J. & A.P Dobson eds]. Oxford University Press.

Touati D. 2000. Iron and oxidative stress in bacteria. *Arch Biochem Biophys*., 373 (1): 1-6.

Travers A. and Muskhelishvili G. 2005. DNA supercoiling - a global transcriptional regulator for enterobacterial growth? *Nat. Rev. Microbiol*., 3 (2): 156-169.

Trivedi B. 2012. Microbiome: The surface brigade. *Nature*, 492: S60–61.

Troussellier M., Escalas A., Bouvier T. and Mouillot D. 2017. Sustaining rare marine microorganisms: macroorganisms as repositories and dispersal agents of microbial diversity. Front. *Microbiol*., 8: 947.

Troxell B., Fink R.C., Porwollik S., McClelland M., Hassan H.M. 2011. The Fur regulon in anaerobically grown *Salmonella enterica sv. typhimurium*: identification of new Fur targets. *BMC Microbiol*., 11: 236.

Troxell B., Hassan H.M. 2013. Transcriptional regulation by Ferric Uptake Regulator (Fur) in pathogenic bacteria. *Front. Cell. Infect. Microbiol.*, 3:59.

Tryland M., Nesbakken T., Robertson L., Grahek Ogden D. and Lunestad B.T. 2014. Human pathogens in marine mammal meat a northern perspective. *Zoonoses and public health*, 61 (6): 377-394.

Turnbaugh P.J., Ley R.E., Hamady M., Fraser-Liggett C.M., Knight R. & Gordon J.I. 2007. The human microbiome project. *Nature*, 449: 804–810.

Turner K.M.E. and Feil, E.J. 2007. The secret life of the multilocus sequence type. *International journal of antimicrobial agents*, 29 (2): 129–135.

Ud-Din A, Wahid S. 2014. Relationship among *Shigella spp*. and enteroinvasive *Escherichia coli* (EIEC) and their differentiation. *Brazilian Journal of Microbiology*, 45 (4): 1131-1138.

UNCTAD 2016. Review of maritime transport 2016. United Nations Publications, xii + 104 p.

Urakawa H., Kita-Tsukamoto K. and Ohwada K. 1999. Microbial diversity in marine sediments from Sagami Bay and Tokyo Bay, Japan, as determined by 16S rRNA gene analysis. *Microbiology*, 145 (11): 3305-3315.

van der Valk R.A., Vreede J., Qin L., Moolenaar G.F., Hofmann A., Goosen N. and Dame R.T. 2017. Mechanism of environmentally driven conformational changes that modulate H-NS DNA bridging activity. *Elife*, 6: e27369

Van Pelt C., Verduin C.M., Goessens W.H.F. *et al*. 1999. Identification of *Burkholderia spp*. in the Clinical Microbiology Laboratory: Comparison of Conventional and Molecular Methods. *Journal of Clinical Microbiology*, 37 (7): 2158-2164.

Vanhove A. S., Rubio T. P., Nguyen A. N., Lemire A., Roche D., Nicod J. ... and Le Roux F. 2016. Copper homeostasis at the host vibrio interface: lessons from intracellular vibrio transcriptomics. *Environmental microbiology*, 18 (3): 875-888.

Venkateswaran K., Iwabuchi T., Matsui Y., Toki H., Hamada E. and Tanaka H. 1991. Distribution and biodegradation potential of oil-degrading bacteria in North Eastern Japanese coastal waters. *FEMS Microbiology Letters*, 86 (2): 113-121.

Venter J.C., Remington K., Heidelberg J.F., Halpern A.L., Rusch D., Eisen J.A. ... and Fouts D.E. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304 (5667): 66-74.

Vergin K.L., Done B., Carlson C.A., and Giovannoni, S.J. 2013. Spatiotemporal distributions of rare bacterioplankton populations indicate adaptive strategies in the oligotrophic ocean. *Aquatic Microbial Ecology*, 71 (1): 1-13.

Vezzulli L., Brettar I., Pezzati E., Reid P.C., Colwell R.R., Höfle M.G., Pruzzo C. 2012. Long-term effects of ocean warming on the prokaryotic community: evidence from the vibrios. *ISME J*., 6: 21-30.

Vezzulli L., Colwell R.R., Pruzzo C. 2013. Ocean warming and spread of pathogenic vibrios in the aquatic environment. *Microbial Ecology*, 65 (4): 817-825.

Vezzulli L., Grande C., Reid P.C., Hélaouët P., Edwards M., Höfle M.G., Brettar I., Colwell R.R., Pruzzo C. 2016. Climate influence on Vibrio and associated human diseases during the past half-century in the coastal North Atlantic. *PNAS*, 201609156.

Vezzulli L., Grande C., Tassistro G., Brettar I., Höfle M.G., Pereira R.P.A., Mushi D., Pallavicini A., Vassallo P., Pruzzo C. 2017. Whole-genome enrichment provides deep insights into *Vibrio cholerae* metagenome from an African river. *Microbial Ecology*, 73 (3): 734-738.

Vezzulli L., Previati M., Pruzzo C., Marchese A., Bourne D.G., Cerrano C. 2010a. Vibrio infections triggering mass mortality events in a warming Mediterranean Sea. *Environ. Microbiol.*, 12: 2007-2019.

Vezzulli L., Pruzzo C., Huq A., Colwell R.R. 2010b. Environmental reservoirs of *Vibrio cholerae* and their role in cholera. *Environmental Microbiology Reports*, 2: 27–33.

Vezzulli L., Stauder M., Grande C., Pezzati E., Verheye H.M., Owens N.J.P., Pruzzo C. 2015. gbpA as a novel qPCR target for the species-specific detection of *Vibrio cholera* O1, O139, non-O1/non-O139 in Environmental, Stool, and Historical Continuous Plankton Recorder Samples. Plos One,10(4): e0123983. doi:10.1371/journal.pone.0123983.

Vierheilig *et al.* 2015. Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons in microbial water quality monitoring. *Water Science and Technology*, 72 (11): 1962-1972.

Villar E., Farrant G. K., Follows M., Garczarek L., Speich S., Audic S.and Casotti R. 2015. Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science*, 348 (6237): 1261447.

Vos M. *et al.* 2015. Rates of lateral gene transfer in prokaryotes: high but why? *Trends in microbiology*, 23 (10): 598–605.

Wadhams G.H. and Armitage J.P. 2004. Making sense of it all: bacterial chemotaxis. *Nat. Rev. Mol. Cell. Biol.,* 5: 1024–1037.

Waldor M.K. and Mekalanos J.J. 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, 272 (5270): 1910-1914.

Walthers D., Li Y., Liu Y., Anand G., Yan J. and Kenney L.J. 2011. *Salmonella enterica* response regulator SsrB relieves H-NS silencing by displacing H-NS bound in polymerization mode and directly activates transcription. *J. Biol. Chem.*, 286 (3):1895-902.

Wang C.Y.C., Shie H.S., Chen S.C., Huang J.P., Hsieh I.C., Wen M.S.... and Wu D. 2007. Lactococcus garvieae infections in humans: possible association with aquaculture outbreaks. *International journal of clinical practice*, 61 (1): 68-73.

Wardeh M., Risely C., McIntyre K.M., Setzkorn C. and Baylis M. 2015. Database of host-pathogen and related species interactions, and their global distribution. *Sci. Data.,* 2: 150049. DOI:10.1038/sdata.2015.49.

Wardeh M., Risley C., McIntyre M.K., Setzkorn C. and Baylis M. 2015. Database of host-pathogen and related species interactions, and their global distribution. *Scientific data*, 2: 150049.

Warnecke F. and Hugenholtz P. 2007. Building on basic metagenomics with complementary technologies. *Genome Biol.*, 8: 231.

Watanabe YY., Sato K., Watanuki Y., Takahashi A., Mitani Y., Amano M. *et al.* 2011. Scaling of swim speed in breath-hold divers. *J. Anim. Ecol*., 80: 57–68.

Webb G. I. 2007. Discovering Significant Patterns. *Machine Learning*, 68 (1): 1-33.

Weiss R.A. and McMichael A.J. 2004. Social and environmental risk factors in the emergence of infectious diseases. *Nature Medicine Suppliment*, 10 (12): 570-576.

Wendling C.C. and Wegner K.M. 2015. Adaptation to enemy shifts: rapid resistance evolution to local *Vibrio spp*. in invasive Pacific oysters. *Proc. Biol. Sci*., 282: 20142244.

Wendling C.C., Batista F.M. and Wegner K.M. 2014. Persistence, seasonal dynamics and pathogenic potential of Vibrio communities from pacific oyster hemolymph. *Plos one*, 9: e94256.

Westfall P. H. and Young S. S. 1993. Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons.

Wiethoelter A.K., Beltrán-Alcrudo D., Kock R., and Siobhan M.M. 2015. Global trends in infectious diseases at the wildlife–livestock interface. *PNAS,* 112 (31): 9662-9667.

Wilkinson D.A., Dietrich M., Lebarbenchon C., Jaeger A., Le Rouzic C., Bastien M.... and Dellagi K. 2014. Massive infection of seabird ticks with bacterial species related to *Coxiella burnetii*. *Applied and environmental microbiology*, 80 (11): 3327-3333.

Wright A.D., Papendorf O., König G.M. and Oberemm A. 2006. Effects of cyanobacterium *Fischerella ambigua* isolates and cell free culture media on zebrafish (*Danio rerio*) embryo development. *Chemosphere*, 65 (4): 604-608.

Wu L., Lin X., Wang F., Ye D., Xiao X., Wang S. and Peng X. 2006. OmpW and OmpV are required for NaCl regulation in *Photobacterium damsela*. *Journal of proteome research*, 5 (9): 2250-2257.

Yan X. and Han J. 2002. gSpan: Graph-based substructure pattern mining. Proceedings of 2002 IEEE International Conference on Data Mining, 721-724.

Yong-Jin Lee, Joy D van Nostrand, Qichao Tu, Zhenmei Lu, Lei Cheng … and Jizhong Zhou. 2013. The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities. *The ISME Journal*, 7: 1974-1984.

Yoon J.H., Kim I.G., Kang K.H., Oh T.K.. and Park Y.H. 2004. *Nocardioides aquiterrae sp. nov*., isolated from groundwater in Korea. *Int. J. Syst. Evol. Microbiol.*, 54 (Pt 1): 71–5.

Yooseph S., Nealson K.H., Rusch D.B., McCrow J.P., Dupont C.L., Kim M., et al., 2010. Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, 468: 60–66.

Yoshizoe K., Terada A. and Tsuda K. 2015. Redesigning pattern mining algorithms for supercomputers. arXiv:1510.07787,

Young M., Artsatbanov V., Beller H. R., Chandra G., Chater K.F., Dover L.G.... and Lapidus A. 2010. Genome sequence of the Fleming strain of Micrococcus luteus, a simple free-living actinobacterium. *Journal of bacteriology*, 192 (3): 841-860.

Youssef N.H. and Elshahed M. S. 2009. Diversity rankings among bacterial lineages in soil. *The ISME journal*, 3 (3): 305.

Yu X.J., McGourty K., Liu M., Unsworth K.E. and Holden D.W. 2010. pH sensing by intracellular Salmonella induces effector translocation. *Science*, 328 (5981): 1040-1043.

Zaki M. J. and Meira Jr., W. 2016. Data Mining And Analysis. Cambridge.

Zankari, E. *et al.* 2012. Identification of acquired antimicrobial resistance genes. *The Journal of antimicrobial chemotherapy*, 67 (11): 2640–2644.

Zhang G., Fang X., Guo X., Li L., Luo R., Xu F. ... and Xiong Z. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, 490 (7418): 49-54.

Zhang Q., Long Q. and Ott J. 2014. AprioriGWAS, a New Pattern Mining Strategy for Detecting Genetic Variants Associated with Disease through Interaction Effects. *Plos Computational Biology*, 10 (6): e1003627.

Zhang W. and Baseman J.B. 2011. Transcriptional regulation of MG_149, an osmoinducible lipoprotein gene from Mycoplasma genitalium. *Mol. Microbiol.*, 81 (2): 327-339.

Zhao F. & Bajic V.B. 2015. The value and significance of metagenomics of marine environments. Preface. *Genomics Proteomics Bioinformatics*, 13: 271–274.

Zhao Y. *et al.* 2012. PGAP: pan-genomes analysis pipeline. *Bioinformatics*, 28 (3): 416-418.

Zhu S., Kojima S. and M. Homma. 2013. Structure, gene regulation and environmental response of flagella in Vibrio. *Frontiers in Microbiology*, 4: 1-9.

LIST OF PARTICIPANTS

Chris Bowler
Inst. Biologie / Ecole Normale Supérieure
Paris, France
cbowler@biologie.ens.fr

Frédéric Briand
Director General, CIESM
fbriand@ciesm.org

Dorota Czerucka
Centre Scientifique de Monaco
Monaco
dczerucka@centrescientifique.mc

Giuseppe D'Auria
Foundation FISABIO
Valencia, Spain
dauria_giu@gva.es

Charles Dorman
Trinity College
Dublin, Ireland
cjdorman@tcd.ie

Edward Feil
Univ. of Bath
Bath, UK
E.Feil@bath.ac.uk

Laura Giuliano
Director of Science, CIESM
lgiuliano@ciesm.org

Frédérique Le Roux
Station Biologique
Roscoff, France
fleroux@sb-roscoff.fr

Elena Stoica
National Inst. for Marine Research
Constanta, Romania
estoica@alpha.rmri.ro

Mahito Sugiyama
National Inst. of Informatics
Tokyo, Japan
mahito@nii.ac.jp

Stepan Toshchakov Emm. Kant Baltic Federal Univ.
Kaliningrad, Russia
stepan.toshchakov@gmail.com

Marc Troussellier UMR MARBEC
Univ. Montpellier, France
troussel@univ-montp2.fr

Maya Wardeh Inst. of Infection & Global Health
Univ. of Liverpool, UK
Maya.Wardeh@liverpool.ac.uk

Experts invited for inputs on modeling tools :

Fernando Peruani Univ. Nice Sophia–Antipolis
France
peruani@unice.fr

Stephane Robin AgroParisTech/ INRA
Paris, France
robin@agroparistech.fr

Excused

Frank Oliver Glöckner Max Planck Inst. & Jacobs Univ.
Bremen, Germany
fog@mpi-bremen.de

Luigi Vezzulli Dpt. of Earth, Env. & Life Sc.
University of Genoa, Italy
luigi.vezzuli@unige.it