Infectious Disease 'Omics

Metagenomics and Microbiomes

Ernest Diez Benavente
LSHTM

ernest.diezbenavente@lshtm.ac.uk

Course outline

Infectious Disease 'Omics workshop

London School of Hygiene and Tropical Medicine

	September 17, 2018	September 18, 2018	September 19, 2018
0845 – 1045	Introduction to 'omic data and quality control	Phylo-genetics/-dynamics and populations	Microbiomes
	Coffee Break	Coffee Break	Coffee Break
1100 - 1230	Mapping and variant detection	Phylogenetics/dynamics and populations, GWAS	Transcriptomics RNA-seq Differential expression
	Lunch	Lunch	Lunch
1315 - 1515	Variant detection	Case study: Third generation sequencing and analysis (e.g. MinION)	Epigenetics, methylation and eQTLS
	Coffee Break	Coffee Break	Coffee Break
1530 – 1730	Assembly of genomes	GWAS	Advanced topics (e.g. host-pathogen studies)
	Participant feedback	Participant feedback	Participant feedback

What is metagenomics?

- ***In situ***, **culture-free** genomic characterization of the **taxonomic and functional profiles** of a **microbial community**.

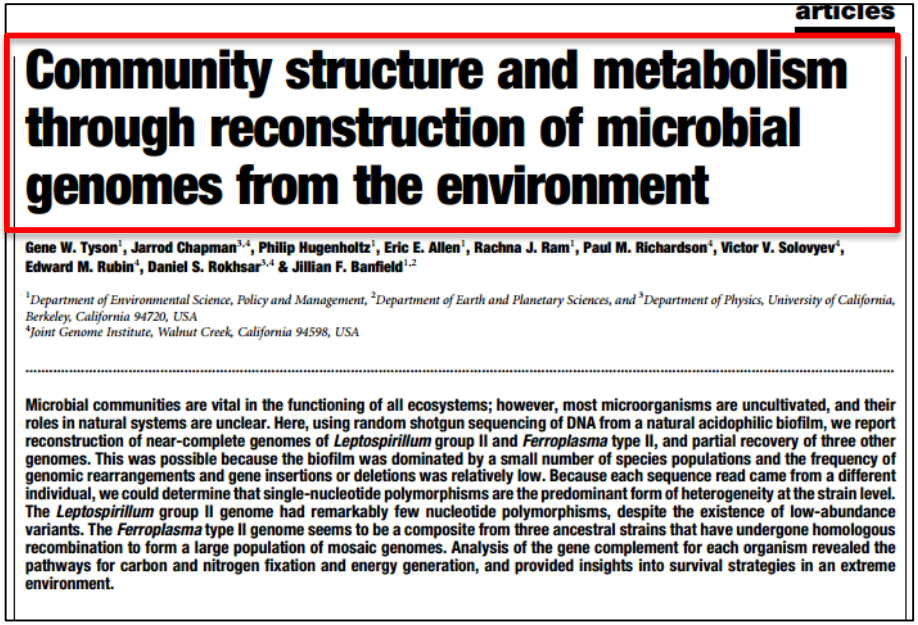
- Identifies and quantifies microbial taxa and/or genes, to know **"who" is there** and **what functions** can they perform.

- Generates **millions of reads**, typically more than a single parasite genomic project.

- Challenges:
  - Data management and analysis (the data **is large** and **diverse**).
  - **Experimental protocols and data-cleaning can produce bias** in the samples.

en.wikipedia.org

What can be achieved with metagenomics/microbiome analysis?



articles

Community structure and metabolism through reconstruction of microbial genomes from the environment

Gene W. Tyson[1], Jarrod Chapman[3,4], Philip Hugenholtz[1], Eric E. Allen[1], Rachna J. Ram[1], Paul M. Richardson[4], Victor V. Solovyev[4], Edward M. Rubin[4], Daniel S. Rokhsar[3,4] & Jillian F. Banfield[1,2]

[1]*Department of Environmental Science, Policy and Management,* [2]*Department of Earth and Planetary Sciences, and* [3]*Department of Physics, University of California, Berkeley, California 94720, USA*
[4]*Joint Genome Institute, Walnut Creek, California 94598, USA*

Microbial communities are vital in the functioning of all ecosystems; however, most microorganisms are uncultivated, and their roles in natural systems are unclear. Here, using random shotgun sequencing of DNA from a natural acidophilic biofilm, we report reconstruction of near-complete genomes of *Leptospirillum* group II and *Ferroplasma* type II, and partial recovery of three other genomes. This was possible because the biofilm was dominated by a small number of species populations and the frequency of genomic rearrangements and gene insertions or deletions was relatively low. Because each sequence read came from a different individual, we could determine that single-nucleotide polymorphisms are the predominant form of heterogeneity at the strain level. The *Leptospirillum* group II genome had remarkably few nucleotide polymorphisms, despite the existence of low-abundance variants. The *Ferroplasma* type II genome seems to be a composite from three ancestral strains that have undergone homologous recombination to form a large population of mosaic genomes. Analysis of the gene complement for each organism revealed the pathways for carbon and nitrogen fixation and energy generation, and provided insights into survival strategies in an extreme environment.

PNAS PNAS

Vaginal microbiome of reproductive-age women

Jacques Ravel[a,1], Pawel Gajer[a], Zaid Abdo[b], G. Maria Schneider[c], Sara S. K. Koenig[a], Stacey L. McCulle[a], Shara Karlebach[a], Reshma Gorle[a], Jennifer Russell[a], Carol O. Tacket[a], Rebecca M. Brotman[a], Catherine C. Davis[a], Kevin Ault[d], Ligia Peralta[e], and Larry J. Forney[c,1]

[a]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201; [b]Departments of Mathematics and Statistics and the Initiative for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844; [c]Department of Biological Sciences and the Initiative for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844; [d]Emory University School of Medicine, Atlanta, GA 30322; [e]Department of Pediatrics Adolescent and Young Adult Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; [f]Center for Vaccine Development, University of Maryland School of Medicine, Baltimore, MD 21201; and [g]The Procter & Gamble Company, Cincinnati, OH 45224

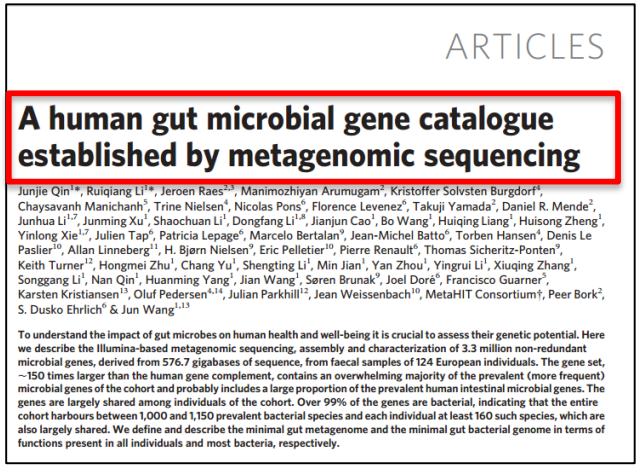
Edited by Jeffrey I. Gordon, Washington University School of Medicine, St. Louis, MO, and approved May 7, 2010 (received for review March 14, 2010)

The means by which vaginal microbiomes help prevent urogenital diseases in women and maintain health are poorly understood. To gain insight into this, the vaginal bacterial communities of 396 asymptomatic North American women who represented four ethnic groups (white, black, Hispanic, and Asian) were sampled and the species composition characterized by pyrosequencing of barcoded 16S rRNA genes. The communities clustered into five groups: four were dominated by *Lactobacillus iners*, *L. crispatus*, *L. gasseri*, or *L. jensenii*, whereas the fifth had lower proportions of lactic acid bacteria and higher proportions of strictly anaerobic organisms, indicating that a potential key ecological function, the production of lactic acid, seems to be conserved in all communities. The proportions of each community group varied among the four ethnic groups, and these differences were statistically significant [$\chi^2(10) = 36.8$, $P < 0.0001$]. Moreover, the vaginal pH of women in different ethnic groups also differed and was higher in Hispanic (pH 5.0 ± 0.59) and black (pH 4.7 ± 1.04) women as compared with Asian (pH 4.4 ± 0.59) and white (pH 4.2 ± 0.3) women. Phylotypes with correlated relative abundances were found in all communities, and these patterns were associated with either high or low Nugent scores, which are used as a factor for the diagnosis of bacterial vaginosis. The inherent differences within and between women in different ethnic groups strongly argues for a more refined definition of the kinds of bacterial communities normally found

of samples have usually been analyzed, and the depth of sample analysis was not great.

In this study we sought to develop an in-depth and accurate understanding of the composition and ecology of the vagina microbial ecosystem in asymptomatic women using a high-throughput method based on pyrosequencing of barcoded 16S rRNA genes. The data obtained are an essential prerequisite for comprehending the role and ultimately the function of vaginal microbiota in reducing the risk of acquiring diseases and identifying factors that determine disease susceptibility. Specifically we sought to characterize the vaginal microbial communities in a cohort of 396 North American women equally representing four ethnic backgrounds (Asian, white, black, and Hispanic) and further address three aims. The first was to establish whether there were correlations between community composition and vaginal pH because these would be indicative of community performance. The second was to explore how the species composition of vaginal communities was reflected in Nugent scores (25), a diagnostic factor commonly used to identify women with bacterial vaginosis (26). Finally, the third aim was to identify patterns in the relative abundances of different species because these might reflect antagonistic or cooperative interspecies interactions.

Results and Discussion

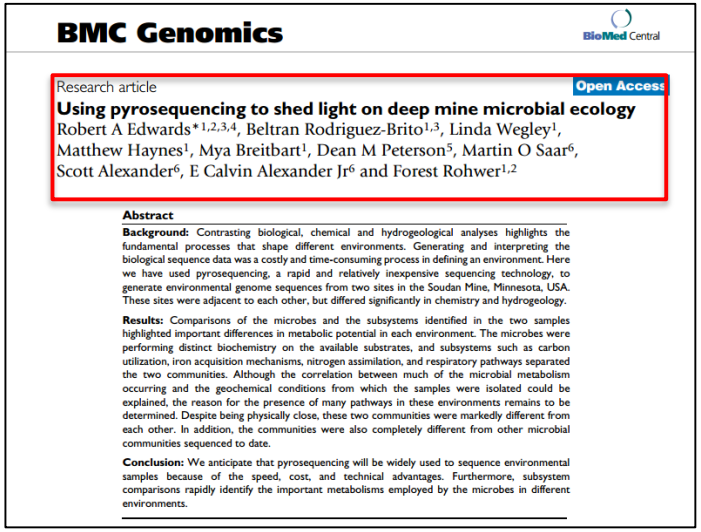


ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1*], Ruiqiang Li[1*], Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.



BMC Genomics

BioMed Central

Research article

Open Access

Using pyrosequencing to shed light on deep mine microbial ecology

Robert A Edwards*[1,2,3,4], Beltran Rodriguez-Brito[1,3], Linda Wegley[1], Matthew Haynes[1], Mya Breitbart[1], Dean M Peterson[5], Martin O Saar[6], Scott Alexander[6], E Calvin Alexander Jr[6] and Forest Rohwer[1,2]

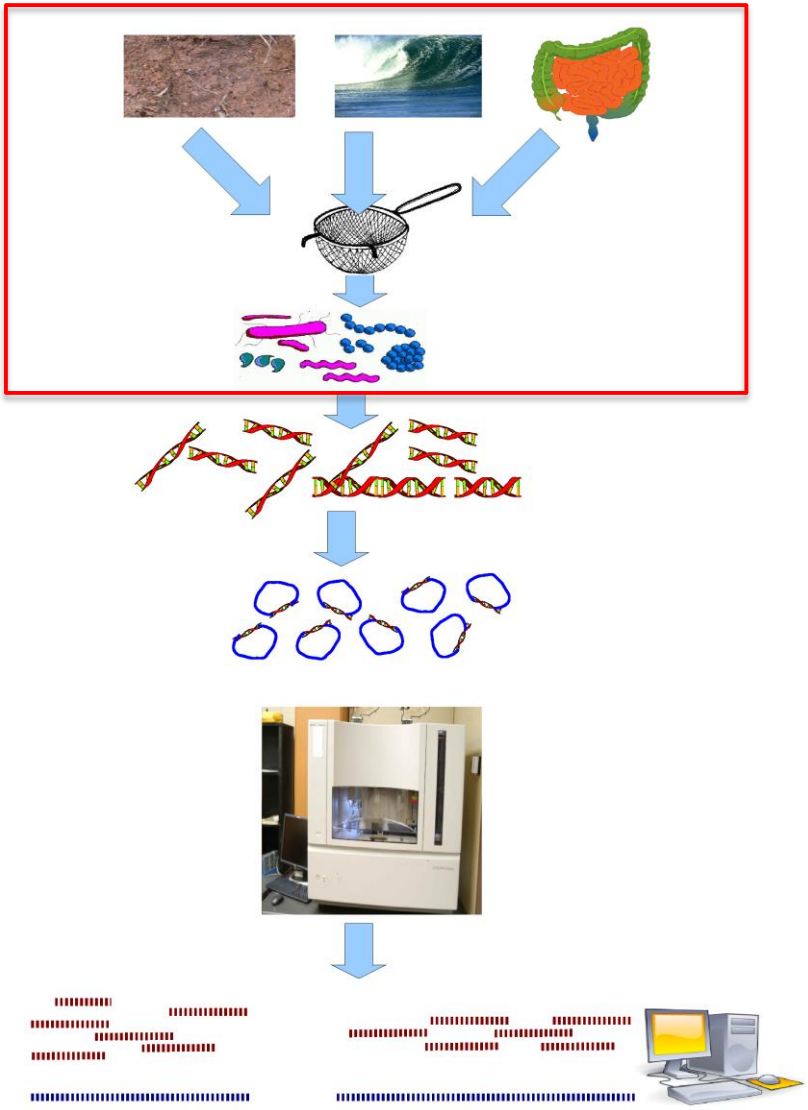
Abstract

Background: Contrasting biological, chemical and hydrogeological analyses highlights the fundamental processes that shape different environments. Generating and interpreting the biological sequence data was a costly and time-consuming process in defining an environment. Here we have used pyrosequencing, a rapid and relatively inexpensive sequencing technology, to generate environmental genome sequences from two sites in the Soudan Mine, Minnesota, USA. These sites were adjacent to each other, but differed significantly in chemistry and hydrogeology.

Results: Comparisons of the microbes and the subsystems identified in the two samples highlighted important differences in metabolic potential in each environment. The microbes were performing distinct biochemistry on the available substrates, and subsystems such as carbon utilization, iron acquisition mechanisms, nitrogen assimilation, and respiratory pathways separated the two communities. Although the correlation between much of the microbial metabolism occurring and the geochemical conditions from which the samples were isolated could be explained, the reason for the presence of many pathways in these environments remains to be determined. Despite being physically close, these two communities were markedly different from each other. In addition, the communities were also completely different from other microbial communities sequenced to date.

Conclusion: We anticipate that pyrosequencing will be widely used to sequence environmental samples because of the speed, cost, and technical advantages. Furthermore, subsystem comparisons rapidly identify the important metabolisms employed by the microbes in different environments.

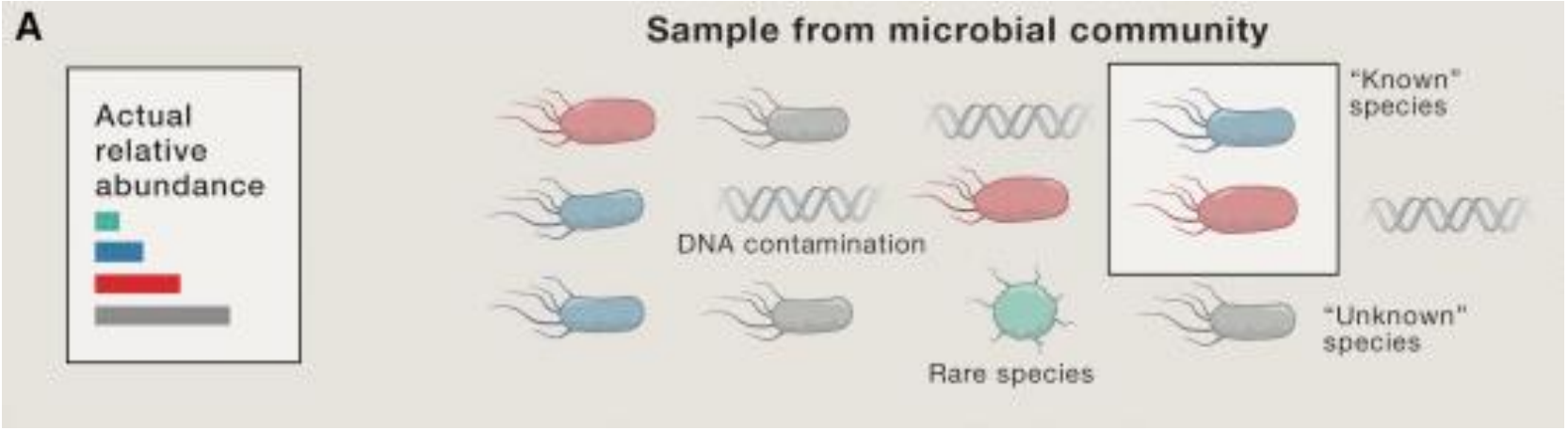
The metagenomics process



The metagenomics process: **Sample from microbial community**

- Collection of samples from a microbial community (water from the ocean, faecal sample from a patient, environmental sample, soil, etc.)

- Who is there and what functions they perform?

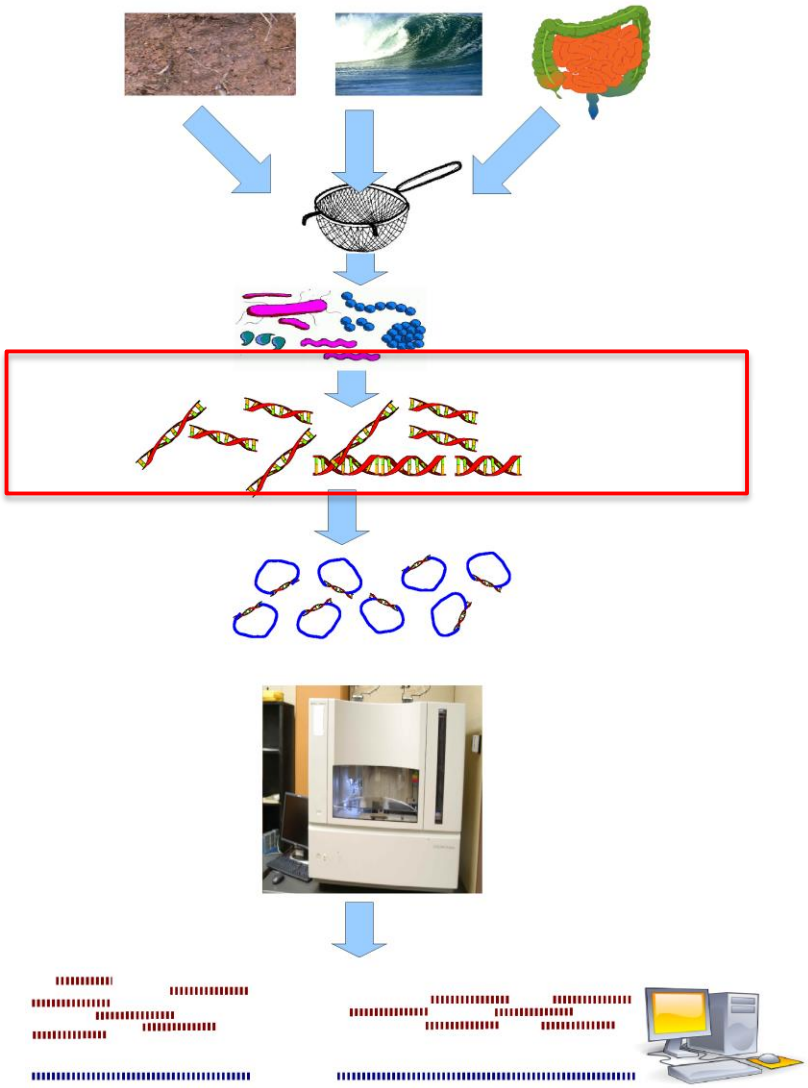


Main Challenges

- Unknown species can dominate microbial communities (not detected by reference-based methods).

- DNA from host (i.e. human gut) or laboratory can contaminate.

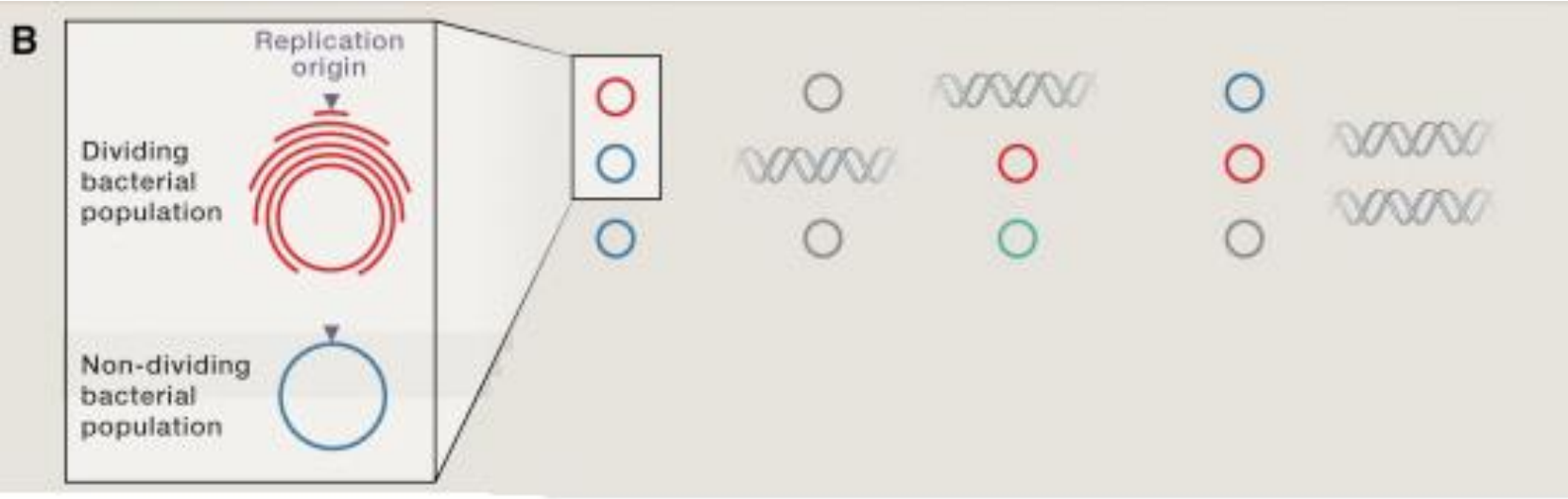
The metagenomics process



The metagenomics process: **DNA extraction**

- Extraction of the DNA from the organisms that are found in the bacterial sample.

- Discard all the rest (proteins, membranes, organelles)



Main Challenges

- Extraction efficiency varies between taxa and depends on the protocol used.

- Dividing (more active) bacteria have a higher and less even coverage than non-dividing bacteria.

The metagenomics process: **DNA fragmentation**

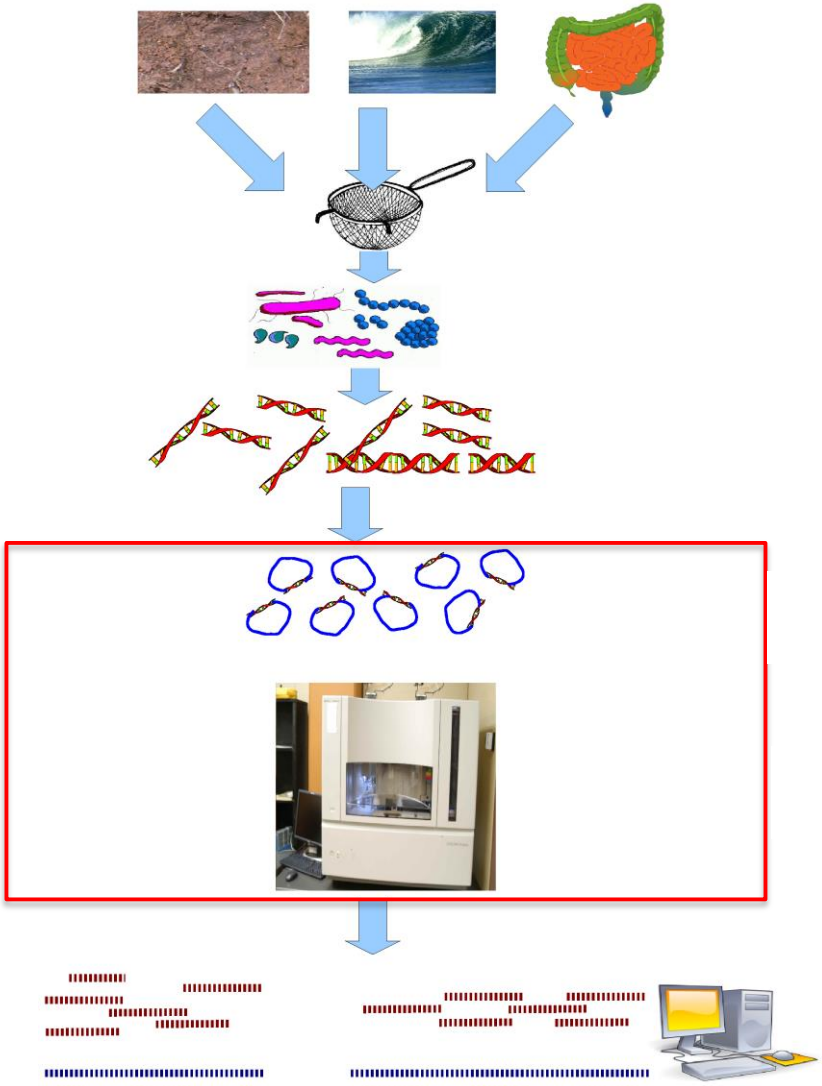
- Extracted DNA is fragmented by mechanical or enzymatic methods.

Main Challenges

- This fragmentation occurs at breakpoints that are not evenly distributed as they occur more often in certain di-nucleotides.

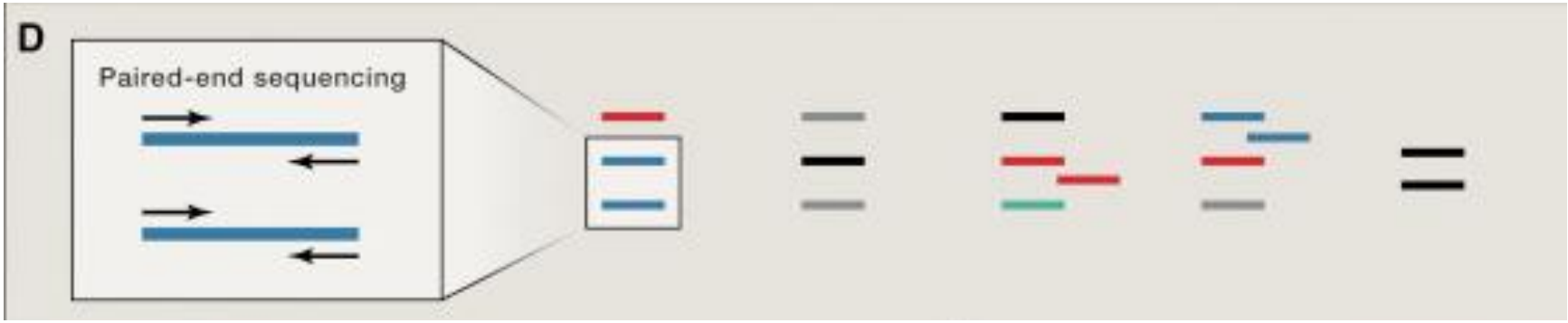
- Some sequences are more likely to be breakpoints than others.

The metagenomics process



The metagenomics process: **Preparation of the library and sequence**

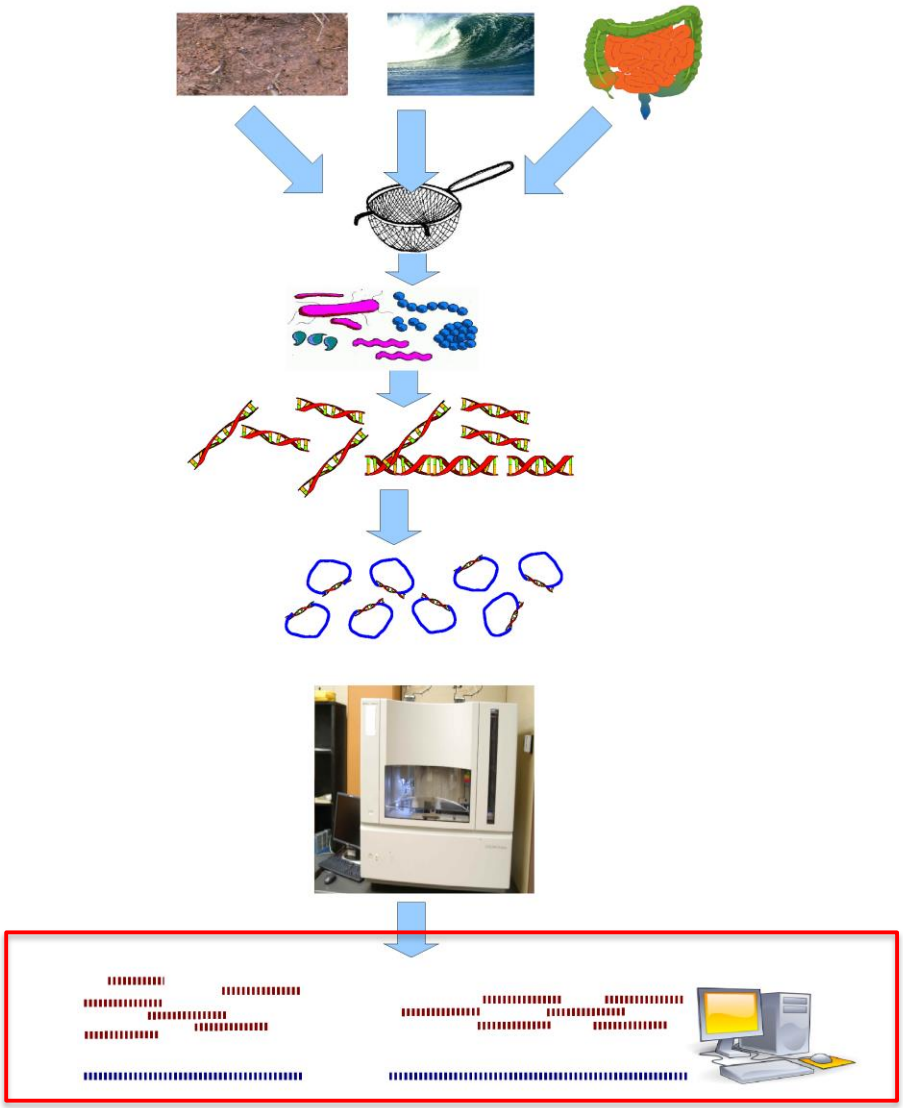
- The libraries are prepared from the fragmented DNA and sequenced.
- This sequencing can be performed using different technologies (pyrosequencing 454, Illumina paired or single sequencing, PacBio SMRT Cell sequencing)



Main Challenges

- Library preparation protocols affect estimates of community composition.
- The different sequencing technologies have different read lengths and error rates.
- DNA fragments with high or low GC% content can be under represented.

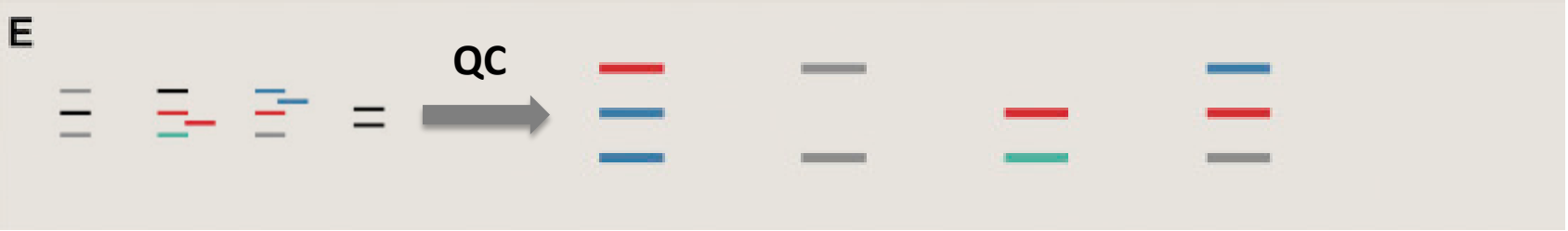
The metagenomics process



The metagenomics process: **Quality control**

- Bioinformatics tools are then performed on the reads to remove the low quality data and obtain a robust dataset.

- The step will: **eliminate duplicate reads**, **trim the read-tails** (having lower quality normally), **remove reads** that clearly come **from a contamination source** and **low quality reads** are filtered out.



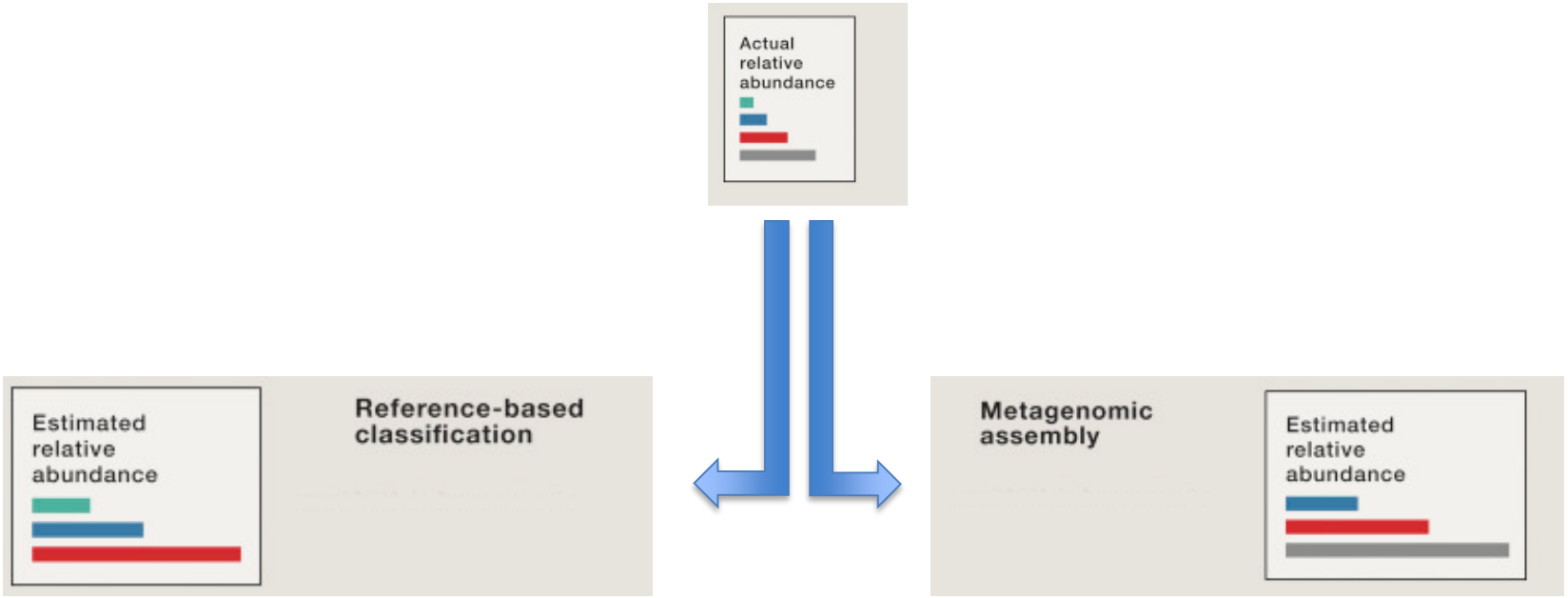
Main Challenges

- The scale of the data.

- What is a contaminant? Is a bacteria a contaminant or an unexpected part of the microbial community?

The metagenomics process: **Data analysis**

- To elucidate the composition of the microbial community, the resultant reads (high quality ones) are either **compared to a reference database** or ***de novo* assembled**.

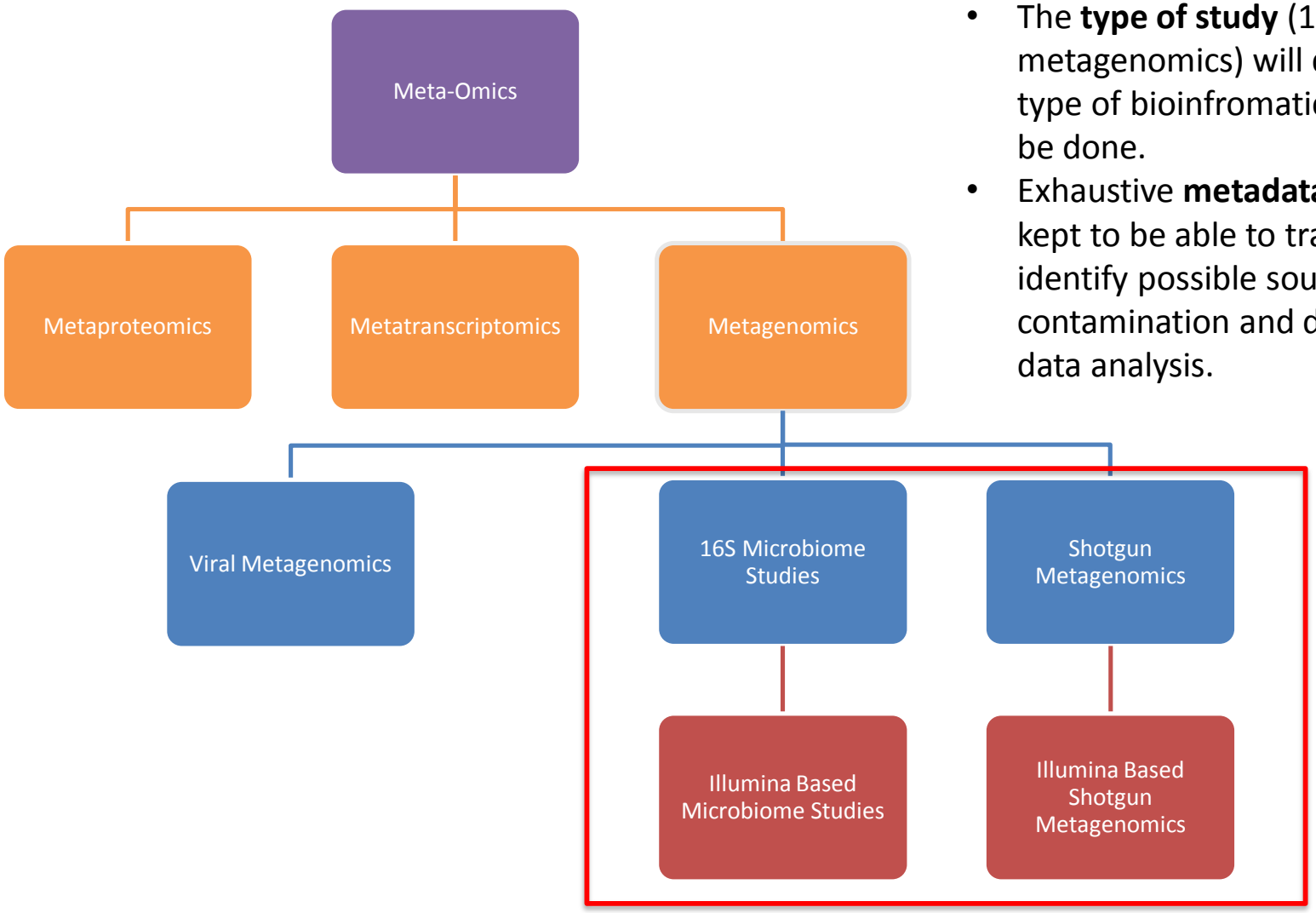


Main Challenges

- Reference-based approaches cannot account for unknown species and overestimates the known ones.

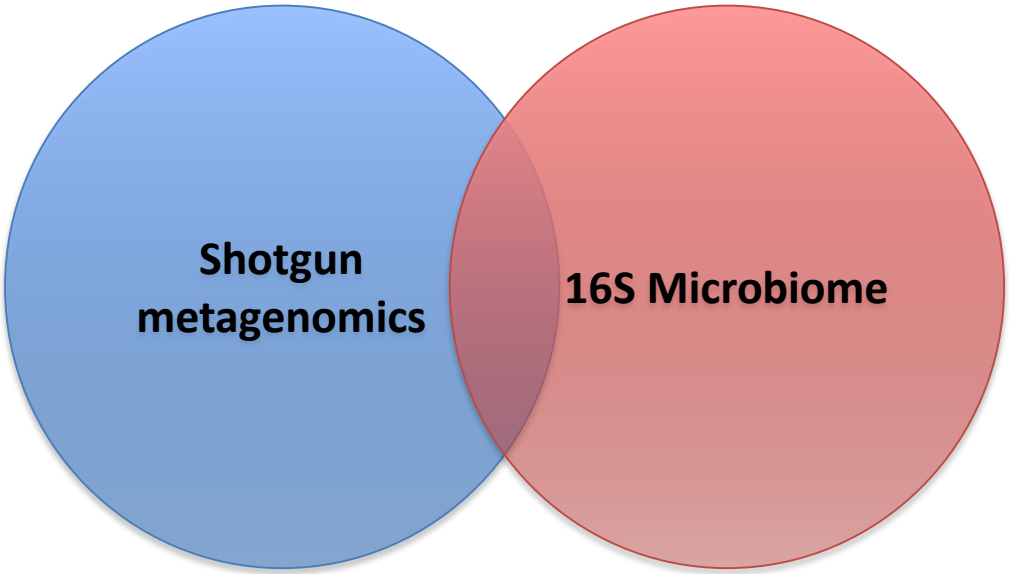
- The metagenomics assembly approach may not detect rare species and overestimates the most abundant species.

Types of metagenomics studies



- The **type of study** (16S, Shotgun metagenomics) will define the type of bioinfromatic analysis to be done.
- Exhaustive **metadata** should be kept to be able to track and identify possible sources of contamination and design the data analysis.

Types of study



- Less *a priori* knowledge before processing sample: we get the sample and we sequence it.
- Analysis more complex due to diversity and size of the data.
- High degree of QC steps needed.

- Amplification step.
- Needs more knowledge *a priori* about the community (selection of primers depending on community).
- Analysis of results and QC, only a small region is sequenced (easy to spot obvious contaminants).
- Are we capturing enough variation (strain variability)?

Sequencing methods

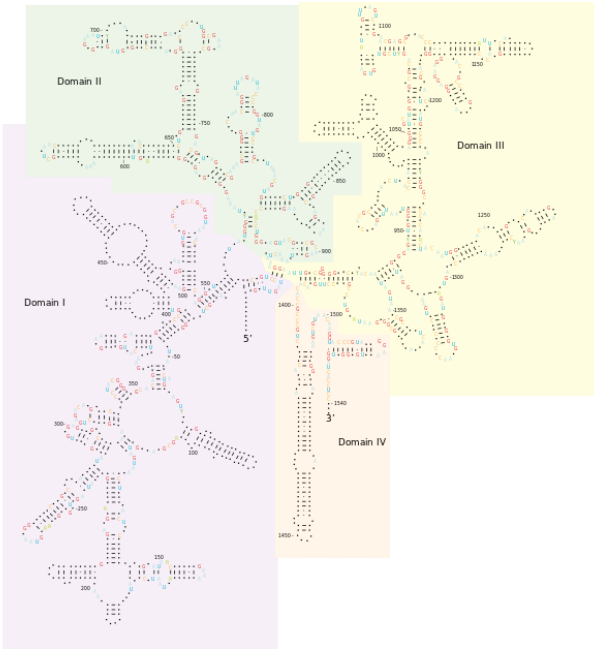
Illumina
MiSeq/HiSeq

PacBio ?

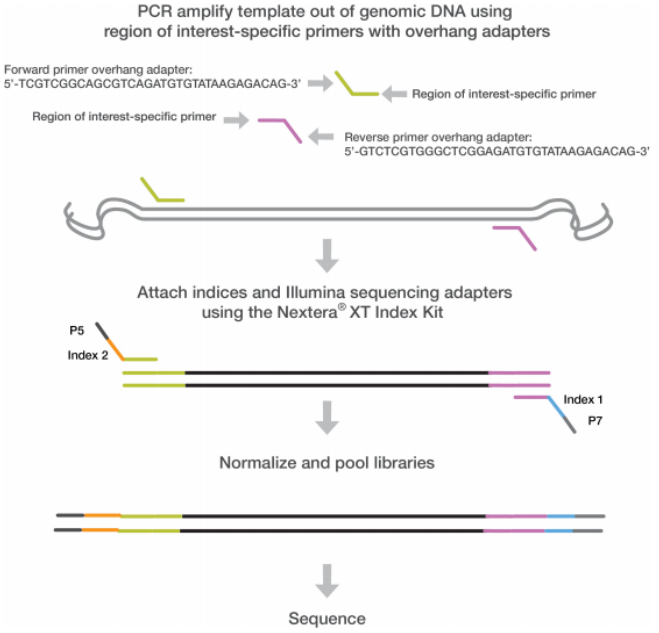
16S Microbiome analysis

- **The microbiome** is the genome collection of the microbial flora harboured by a human host in a specific tissue.
- The **gene** codes for a prokaryotic **16S ribosomal RNA**
  - Has a structural role as a scaffold defining the positions of the ribosomal proteins.
  - Highly conserved between bacteria and archaea.
  - Split in regions that primers can target (V1-V9)

Structure

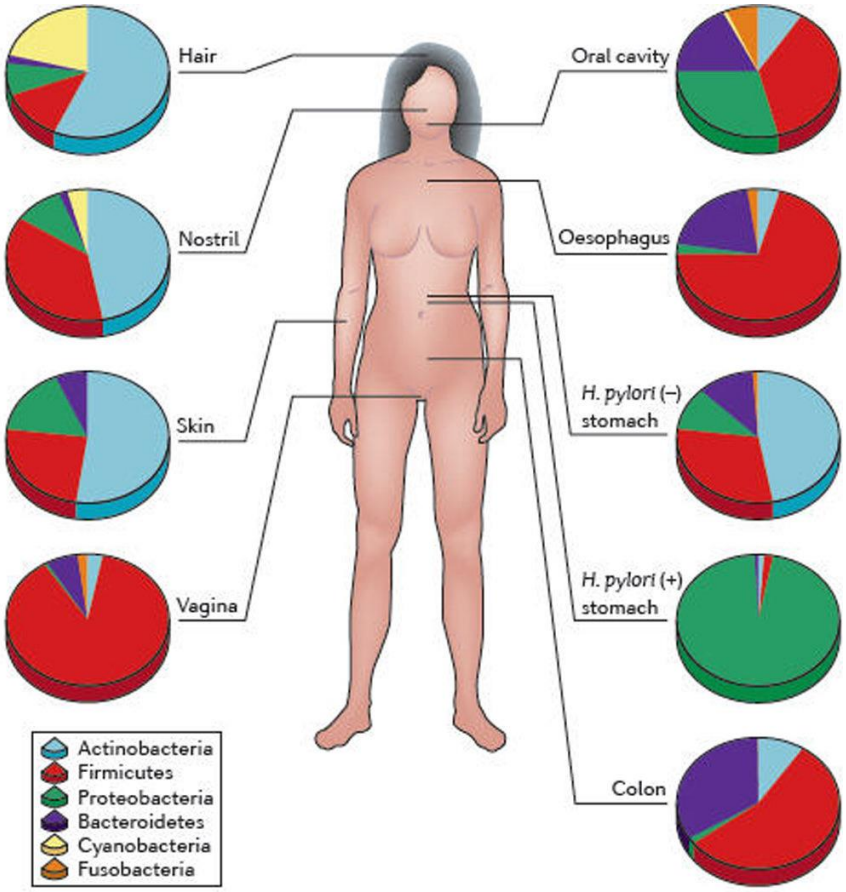


Amplification

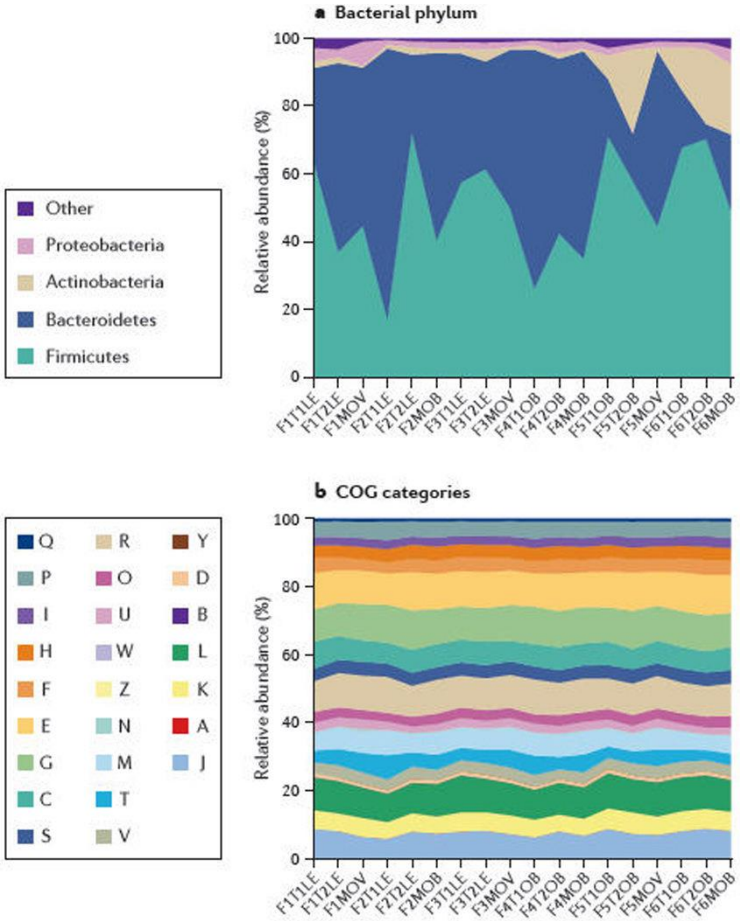


The human microbiome diversity

Differences between physiological sites



Relatively dynamic at phylum and species level, but stable in terms of functions within the community.



Software and databases available for Metagenomics

- **Shotgun Metagenomics**
  - Qiime, (pronounced *Xiime*) is one of the most used metagenomics tool.
  - **Kraken**, k-mer base approach to classify metagenomics sequences.
  - Orione, Galaxy based platform to study metagenomics and microbiome datasets.
  - PandaSeq, to assemble paired end read and correct errors.

- **16S Microbiome analysis**
  - **mothur**, k-mer base approach to classify metagenomics sequences.
  - Qiime, it can also be used for Microbiome data.
  - Ribosomal Database Project (RDP), includes online data analysis and aligned and annotated Bacterial and Archaeal small-subunit 16S rRNA sequences

- **Visualization tools**
  - **KronaTools**, visualization tool that allows intuitive exploration of relative abundances within the complex hierarchies of metagenomics classifications.
  - **R**, coding language.
  - Elviz, website tool for metagenomics data visualization

References

- Bahl, M. I., Bergström, A., & Licht, T. R. (2012). Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *FEMS Microbiology Letters*, *329*(2), 193 LP-197. JOUR. Retrieved from http://femsle.oxfordjournals.org/content/329/2/193.abstract
- Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., … Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, *7*, 57. JOUR. http://doi.org/10.1186/1471-2164-7-57
- Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., … Wong, G. K.-S. (2016). Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*, *7*, 459. JOUR. http://doi.org/10.3389/fmicb.2016.00459
- Nayfach, S., & Pollard, K. S. (2016). Toward Accurate and Quantitative Comparative Metagenomics. *Cell*, *166*(5), 1103–1116. http://doi.org/10.1016/j.cell.2016.08.007
- Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., … Iliopoulos, I. (2015). Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. *Bioinformatics and Biology Insights*, *9*, 75–88. JOUR. http://doi.org/10.4137/BBI.S12462
- Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., … Forney, L. J. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences* , *108*(Supplement 1), 4680–4687. JOUR. Retrieved from http://www.pnas.org/content/108/Supplement_1/4680.abstract
- Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, *5*, 209. JOUR. http://doi.org/10.3389/fpls.2014.00209
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., … Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43. JOUR. Retrieved from http://www.nature.com/nature/journal/v428/n6978/suppinfo/nature02340_S1.html
- http://www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/shotgun-metagenomic-sequencing.html