Metagenomics & Bioinformatics

Thomas H.A. Haverkamp
@Thomieh
Norwegian Veterinary Institute
Outline

● A introduction to metagenomics
● The history of metagenomics
● Metagenomics and bioinformatics
● Metagenomics for surveillance.
Metagenomics definitions

**Metagenome**: the collective genome of all the microorganisms in an environment. ([Handelsman et al., 1998](#))

**Metagenomics**: the study of genetic material (DNA) recovered from an environmental sample.

**Microbiome**: all the microbes in a specific environmental niche.
Metagenomic diversity

Amplicon sequencing

Shotgun Sequencing

Nasir et al., 2012 – DOI: 10.1186/1471-2148-12-156
Metagenomics

Who is there?

Taxonomy

What are they doing?

Metabolism

The anaerobic ones are just sitting there, but the aerobic bacteria are doing jumping jacks, sit-ups, leg lifts....
Great plate count anomaly

< 1 % of cells can be cultured
> 99% uncultured
(1977)
Origin of Metagenomics – 1984

REPORTS

Analysis of Hydrothermal Vent-Associated Symbionts by Ribosomal RNA Sequences

DAVID A. STAHL¹, DAVID J. LANE¹, GARY J. OLSEN¹, NORMAN R. PACE¹
+ See all authors and affiliations

Science 27 Apr 1984:
Vol. 224, Issue 4647, pp. 409-411
DOI: 10.1126/science.224.4647.409

Analysis of 5S rRNA clones

Giant tube Worm (Riftia pachyptila) - image source: Wikipedia
Origin of Metagenomics – 1984

From the abstract of Stahl et al., 1984:

“5S rRNA's were extracted from symbiont-bearing tissues, separated into unique forms, and their nucleotide sequences determined and related to other 5S rRNA's in a phylogenetic tree analysis. The prokaryotic symbionts are related to one another and affiliated with the same narrow phylogenetic grouping as Escherichia coli and Pseudomonas aeruginosa.”

Comparison of environmental sequences with reference sequences !!!
Bacterial Evolution

CARL R. WOESE

Department of Microbiology, University of Illinois, Urbana, Illinois 61801

PERSPECTIVE .................................................................................................................... 222
A Fruitless Search and Its Consequences ........................................................................... 222
Three Ideas That Shape Our Concept of Bacterial Evolution ........................................... 224
  Procaryote-eucaryote dichotomy ..................................................................................... 224
  Oparin Ocean scenario ..................................................................................................... 224
  Darwin's warm little pond ............................................................................................... 226
MEASUREMENT OF BACTERIAL PHYLOGENETIC RELATIONSHIPS ............................... 226
16S rRNA sequencing – 1987

- 12 Bacterial phyla with cultured representatives
- > 500 bacterial species described

Genbank Nucleotide Archive 1986/1987

Woese C. Microbiological reviews 1987
16S rRNA sequencing – 2004

Database: Genbank 16S rRNA genes
- 21,466 cultured microbes
- 54,655 uncultured microbes
- 26 phyla with no cultured representatives.

How can we understand these uncultured microbes?
Shotgun metagenomics - 2004

March 2004
Acid mine drainage metagenome

April 2004
Sargasso Sea Metagenome
Sargasso sea metagenome

- 1.045 billion bases sequenced (SANGER)
- 1800 species detected (148 were not identified before)
- 1.2 million novel genes (Genbank 1-1-2004: 32 million genes)
- 700 different rhodopsin-like proteins discovered in a very nutrient poor environment
Proteorhodopsins in the oceans

- Light driven proton pump
- First discovered in heterotrophic bacteria (*Pelagibacter* spp.)
- Present in marine planktonic bacteria, archaea and eukaryotes (e.g. Dinoflagellates)
- Changed the understanding energy use in the oceans.
2004 >> Sequencing revolution

2017 Illumina Novaseq: 6 Terabyte of data / run
Metagenomic applications

- Ecology
- Gut Microbe Characterization
- Infectious disease diagnosis
- Biofuel
- Environmental remediation
- Biotechnology
- Agriculture
- Single cell metagenomics
Metagenomic methods

- Species detection and abundance (classification)
- Metabolic potential of a microbial community (pathways)
- Extraction of microbial genomes (assembly)
- Comparative analysis of within species diversity (Single nucleotide variants)
- Detection of specific functions (AMR, Toxins, other compounds of interest)
- Detection of mobile elements (phages, plasmids)
- Detection of Growth Dynamics

Bioinformatics is essential for all these methods
Metagenomic assembly

Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen

Robert D. Stewart¹, Marc D. Auffret², Amanda Warr¹, Andrew H. Wiser³, Maximilian O. Press³, Kyle W. Langford³, Ivan Liachko³, Timothy J. Snelling⁴, Richard J. Dewhurst², Alan W. Walker⁴, Rainer Roehe² & Mick Watson¹

Mining for biomass-degrading enzymes
Metagenomic assembly

913 draft genomes:
- 850 “normal”
- 63 Hi-C assisted

Hi-C-based proximity-guided assembly

913 draft genomes

Metagenome Amplified Genomes (MAGs)

Check for:

- Completeness ($\geq 80\%$)
- Contamination ($\leq 10\%$)

$\rightarrow$ Use single copy genes

Stewart et al., Nat Commun. 2018 Feb 28; 9(1) 870
Carbohydrate-active enzymes

Classification of metagenome enzymes

Fig. 3 Distribution of the maximum percentage identity of the RUG proteins against five public databases for six classes of carbohydrate-active enzymes.
GH glycoside hydrolase, GT glycosyl transferase, PL polysaccharide lyases, CE carbohydrate esterases, AA auxiliary activities, CB carbohydrate binding

Stewart et al., Nat Commun. 2018 Feb 28; 9(1) 870
Metagenomic classification

Classification of metagenome reads against:
7 custom databases

→ draft genomes improve classification results.

Stewart et al., Nat Commun. 2018 Feb 28; 9(1) 870
A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4

Importance: Identification of the bacterium responsible for an outbreak can aid in disease management. However, traditional culture-based diagnosis can be difficult, particularly if no specific diagnostic test is available for an outbreak strain.

Metagenome sequences assembled into Environmental gene tags (EGTs)

Loman et al., 2013 Jama (DOI: 10.1001/jama.2013.3231)
Metagenomics for outbreaks

Figure 2. Recovery of Sequences From the Outbreak Strain From the Outbreak Metagenome Through Iterative Filtering

EGT filtering to remove Non-pathogen sequences

Loman et al., 2013 Jama (DOI: 10.1001/jama.2013.3231)
Metagenomics for outbreaks

Figure 3. Reconstruction of the *Escherichia coli* O104:H4 Outbreak Strain Genome

Reconstruction of the *E. coli* outbreak strain genome Using EGTs

Loman et al., 2013 Jama (DOI: 10.1001/jama.2013.3231)
Metagenomic pitfalls

- **Classification**: → Database dependent!
  - What is not in the database, will not be detected !!!
  - Overclassification (e.g. finding more strains than there actually are)
- **Metabolic reconstructions**: → Again, database dependent!
  - Finding new pathways possible, with detailed manual curation
- **Metagenomic assembly**
  - To be successful, it depends on the organism of interest, its diversity, and the diversity of your sample !!!
  - Uneven coverage in actively growing communities.
  - For phages /plasmids, it is difficult to identify the host.
Metagenomic pitfalls

- AMR / Toxin gene detection, very dependent on prevalence in ecosystem
  - Deep sequencing needed for reliable estimates
- Comparing metagenomic samples.
  - Relative abundances mask a lot of the actual differences
  - Specialized normalization often needed (Cells counts)
- In host-microbiome samples, the host genome can be a significant fraction of the data. (10 - 70%)
Acknowledgements

Metagenomics: The Next Culture-Independent Game Changer