Metagenomics 1

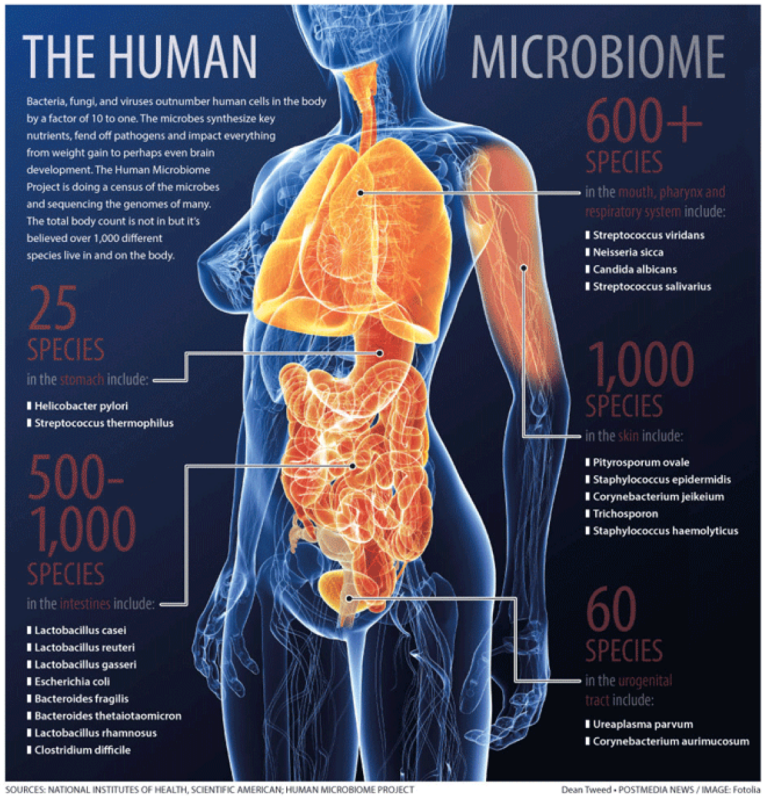
BCB 5250 Introduction to Bioinformatics II

Spring 2020

Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

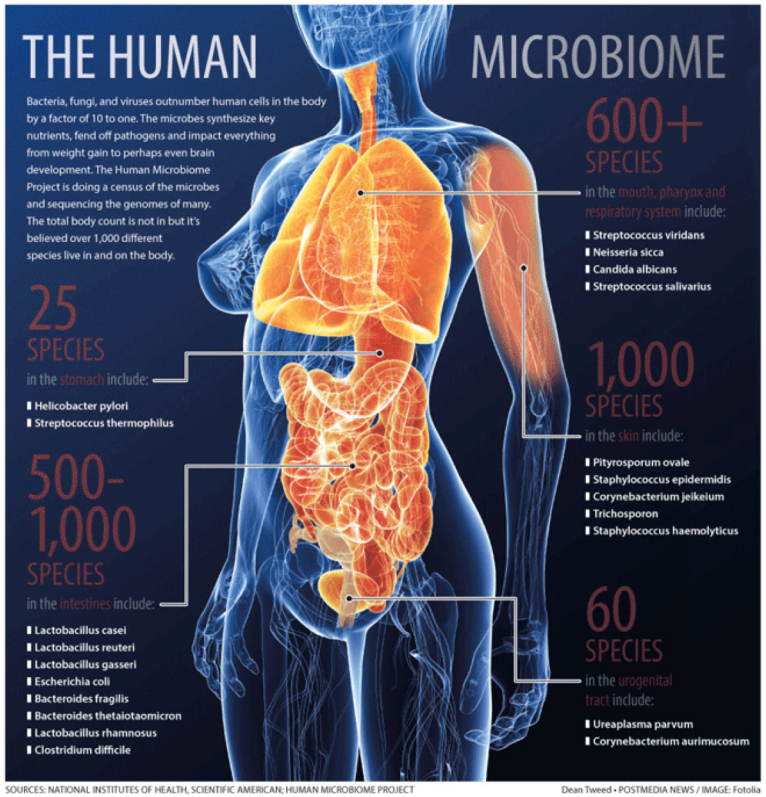
SAINT LOUIS
UNIVERSITY
EST. 1818

Human + Microbes Fact Sheet



- How many human cells in the human body?
  - 10^{13}
- Microbes in your body:
  - 10 times more cells than you.
  - 100 times more genes than you.
  - 1000 different species.

Human + Microbes Fact Sheet



- How many human cells?
  - 3.0×10^{13}
- How many bacterial cells in the human body?
  - 3.8×10^{13}

PLOS | BIOLOGY

ESSAY

Revised Estimates for the Number of Human and Bacteria Cells in the Body

Ron Sender[1], **Shai Fuchs**[2¤]*, **Ron Milo**[1]*

1 Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot, Israel,
2 Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

¤ Current address: Division of Endocrinology, The Hospital for Sick Children and Department of Pediatrics, The University of Toronto, Toronto, Canada
* shai.fuchs@sickkids.ca (SF); ron.milo@weizmann.ac.il (RM)

CrossMark
click for updates

OPEN ACCESS

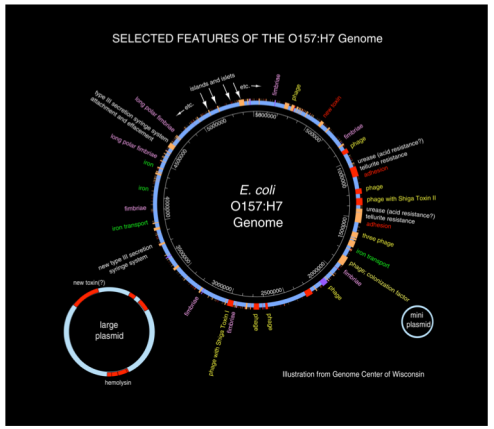
Citation: Sender R, Fuchs S, Milo R (2016) Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol 14(8): e1002533. doi:10.1371/journal.pbio.1002533

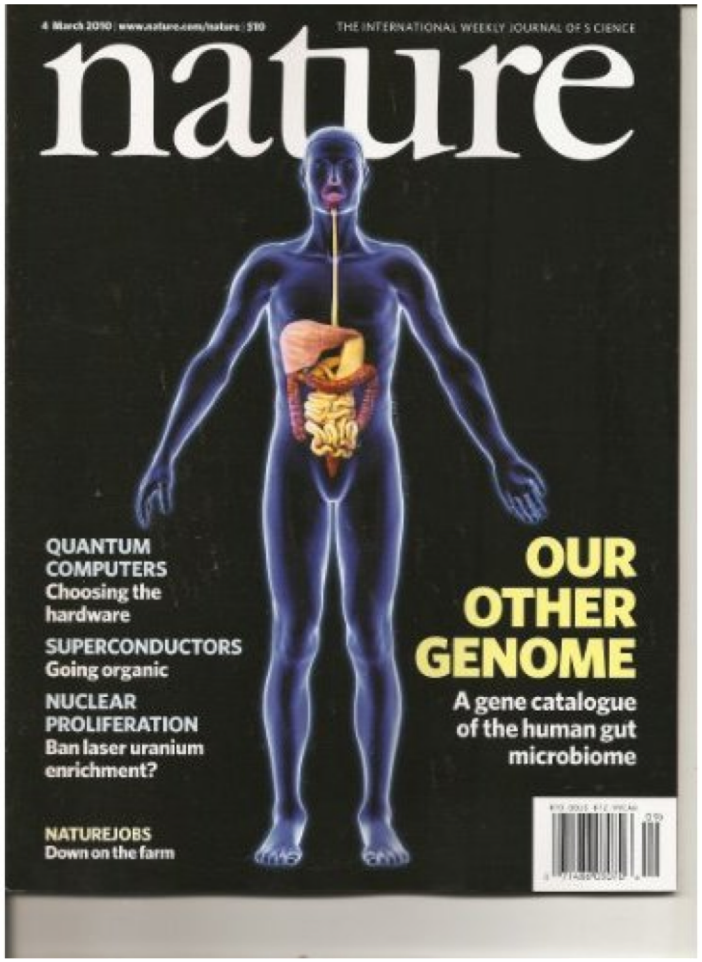
Published: August 19, 2016

Abstract

Reported values in the literature on the number of cells in the body differ by orders of magnitude and are very seldom supported by any measurements or calculations. Here, we integrate the most up-to-date information on the number of human and bacterial cells in the body. We estimate the total number of bacteria in the 70 kg "reference man" to be $3.8 \cdot 10^{13}$. For human cells, we identify the dominant role of the hematopoietic lineage to the total count (≈90%) and revise past estimates to $3.0 \cdot 10^{13}$ human cells. Our analysis also updates the widely-cited 10:1 ratio, showing that the number of bacteria in the body is actually of the same order as the number of human cells, and their total mass is about 0.2 kg.

From Genomics to Metagenomics





From Genomics to Metagenomics

- In Greek, Meta means “transcendent”.
- Genomics: the study of an organism's entire genome.
- Metagenomics: Culture-independent genomic analysis of a community of microorganisms from environmental sample.

What is Metagenomics?

- Metagenomics (Environmental Genomics, Ecogenomics or Community Genomics) is the study of genetic material recovered directly from environmental samples. (Wikipedia)

- Metagenomics is application of modern genomic techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species. (Chen & Pachter, 2005)

About Metagenomics

- Most microbial activities are carried out by complex communities of microorganisms.
- 99% of microbial species cannot be currently cultivated
  - But we can still see them under microscope and can retrieve their DNA
- Next-generation sequencing technology (NGS) allows a much deeper characterization of the structure of microbial communities using metagenomic approaches and cheap sequences.

Why Metagenomics?

Discovery of:

- novel natural products
- new antibiotica
- new molecules with new functions
- new enzymes and bioactive molecules
- what is a genome/species
- diversity of life
- interplay between human and microbes
- how do microbial communities work and how stable are they

Fascinating Facts about Gut Microbiome

- Our microbes make us more different from one another than our DNA.
  - While we're 99.9 percent identical to other people in terms of human DNA, we might only share 10 percent similarity with our fellow human beings in terms of our gut microbes. Each of us has about 20,000 human genes, but as many as 2-20 million microbial genes.
- Microbes in one part of the body are significantly different from microbes in other regions of the body.
  - A few feet of difference in the human body has more of an impact on your microbial biology than hundreds of miles on earth.

http://themindbodyshift.com/index.php/2015/02/27/how-gut-bacteria-affect-behavior-mood-and-disease/

Fascinating Facts about Gut Microbiome

- Microbes determine how medications and mosquitoes affect you.
  - Which microbes you have in your gut determine whether particular painkillers are toxic to your liver and whether other drugs will work for your heart condition. Microbes on your skin determine which chemicals are produced that either attract or repel mosquitoes to bite.
- Our lives depend on microbes. They help:
  - Digest food
  - Educate our immune system
  - Resist disease
  - Metabolize drugs

http://themindbodyshift.com/index.php/2015/02/27/how-gut-bacteria-affect-behavior-mood-and-disease/

Fascinating Facts about Gut Microbiome

- Our first microbial communities depend largely on how we're born.
  - Babies that are delivered through the birth canal have microbes similar to the vaginal community, while babies delivered by C-section have microbes that look like those of the skin. It turns out that natural birth may provide protective microbes that are critical to our health. This might help to explain why children born by Caesarean birth are more prone to have asthma, allergies and obesity—all conditions that have now been linked to microbes.

http://themindbodyshift.com/index.php/2015/02/27/how-gut-bacteria-affect-behavior-mood-and-disease/

Sample Preparation and DNA Extraction

- Design of study and sampling (sample size, timing, replicates)
- Avoid contamination
- Pre-treatment, e.g. filtering
- Lysation and DNA extraction, many methods available, different biases

Wooley et al., PLoS Comp Bio, 2010

Deciding Methods is Important

SCIENTIFIC REPORTS

OPEN

Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing

Received: 2 March 2017
Accepted: 15 June 2017
Published online: 31 July 2017

Michael Tessler[1,2], Johannes S. Neumann[3], Ebrahim Afshinnekoo[4,5,6], Michael Pineda[4,5], Rebecca Hersch[1], Luiz Felipe M. Velho[7,8], Bianca T. Segovia[7], Fabio A. Lansac-Toha[7], Michael Lemke[9], Rob DeSalle[1], Christopher E. Mason[4,5,10] & Mercer R. Brugler[1,11]

frontiers in Microbiology

TECHNOLOGY REPORT
published: 20 April 2016
doi: 10.3389/fmicb.2016.00459

CrossMark

Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics

Juan Jovel[1†], Jordan Patterson[1†], Weiwei Wang[1], Naomi Hotte[1], Sandra O'Keefe[1], Troy Mitchel[1], Troy Perry[1], Dina Kao[1], Andrew L. Mason[1], Karen L. Madsen[1] and Gane K.-S. Wong[1,2,3*]*

[1] Department of Medicine, University of Alberta, Edmonton, AB, Canada, [2] Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada, [3] BGI-Shenzhen, Shenzhen, China

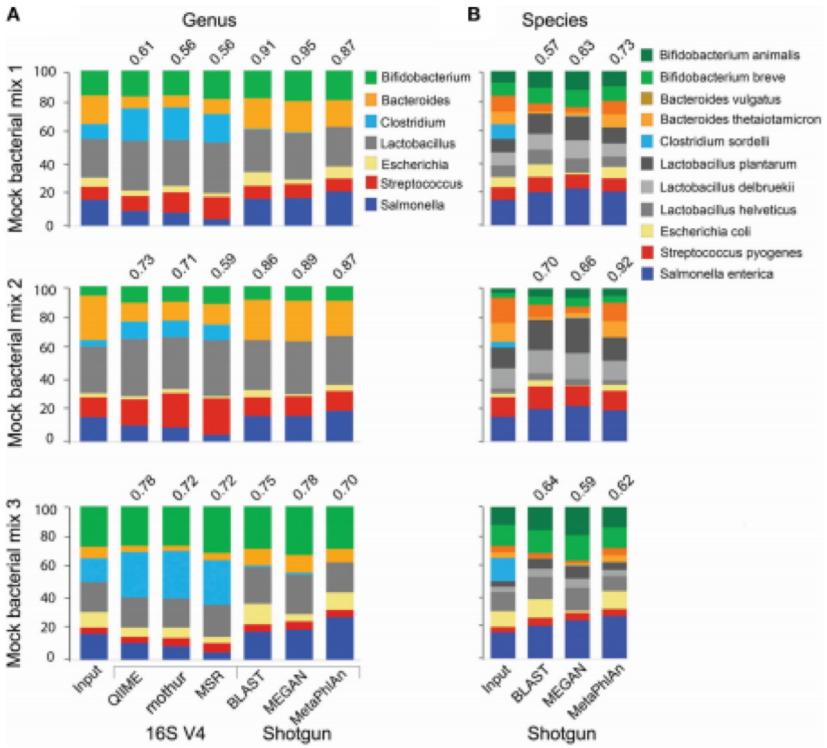


FIGURE 1 | Comparison of taxonomic analyses of a low complexity artificial microbial population using 16S amplicon or shotgun metagenomic approaches. Eleven bacterial species (representing 7 genera) were cultured under standard laboratory conditions. DNA was extracted using the FastDNA spin kit for feces (MPBio). 16S amplicon and shotgun metagenomics libraries were constructed using the NEXTflex 16S V4 Amplicon-Seq (BioO Scientific) and the Nextera XT (Illumina) kits, respectively. Libraries were paired-end sequenced on a MiSeq sequencer using a 500-cycle kit. For 16S libraries, sequences were trimmed with the "split_fastq_libraries.py" script from QIIME. Default parameters were used, with the exception that the quality threshold for trimming was raised to 30. PCR primer sequences were trimmed with in-house Perl scripts. Shotgun metagenomics libraries were trimmed with the fastqMcf tool, and a quality threshold of 15. The relative abundance of each species was determined with the software indicated at the bottom of the bar graph, using default parameters, at the genus **(A)** or species **(B)** levels. The Pearson correlation coefficient between the expected (Input) relative abundance and the classification performed by each program is indicated on top of the bar graph.

Types of Metagenomics Studies

- 16S rRNA-based surveys
  - 16S rRNA sequencing
  - 16S rRNA microarrays
  - PhyloChip: the PhyloChip detects on average twice as many taxa as 16S rRNA gene sequencing
  - Reveal microbial diversity and abundance
- Shotgun metagenomic sequencing
  - Reveal gene content of a community and its metabolic potential
- Targeted metagenomics -- Function-driven metagenomics analysis
  - Transcriptomics, Proteomics, Cell sorting

Sequence-based Metagenomics

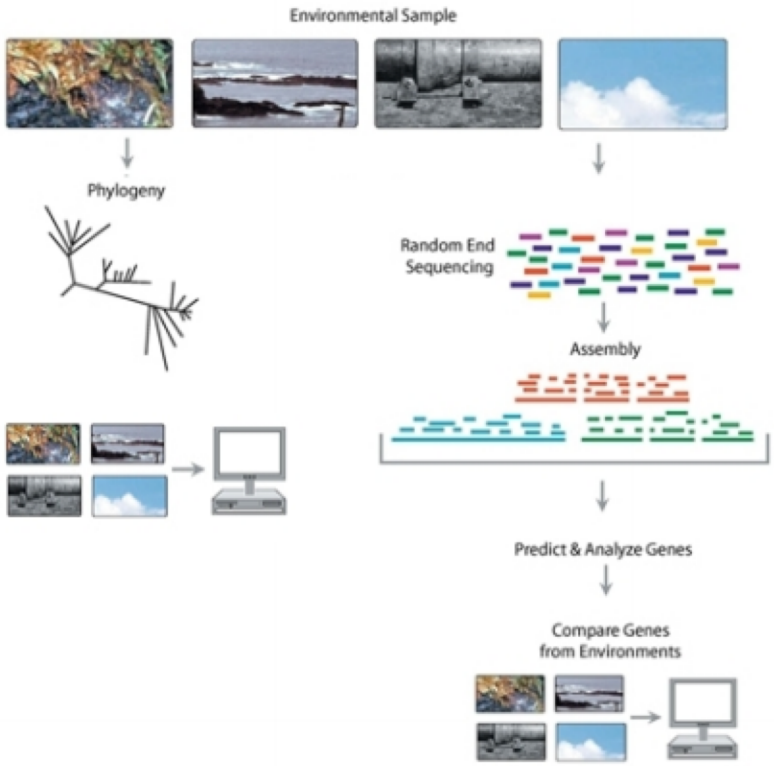
16S rDNA sequencing

Whole-genome sequencing

Sampling
DNA extraction

PCR+sequencing of 16S rDNA
Phylogeny analysis

Comparison



Sequencing

Assembly

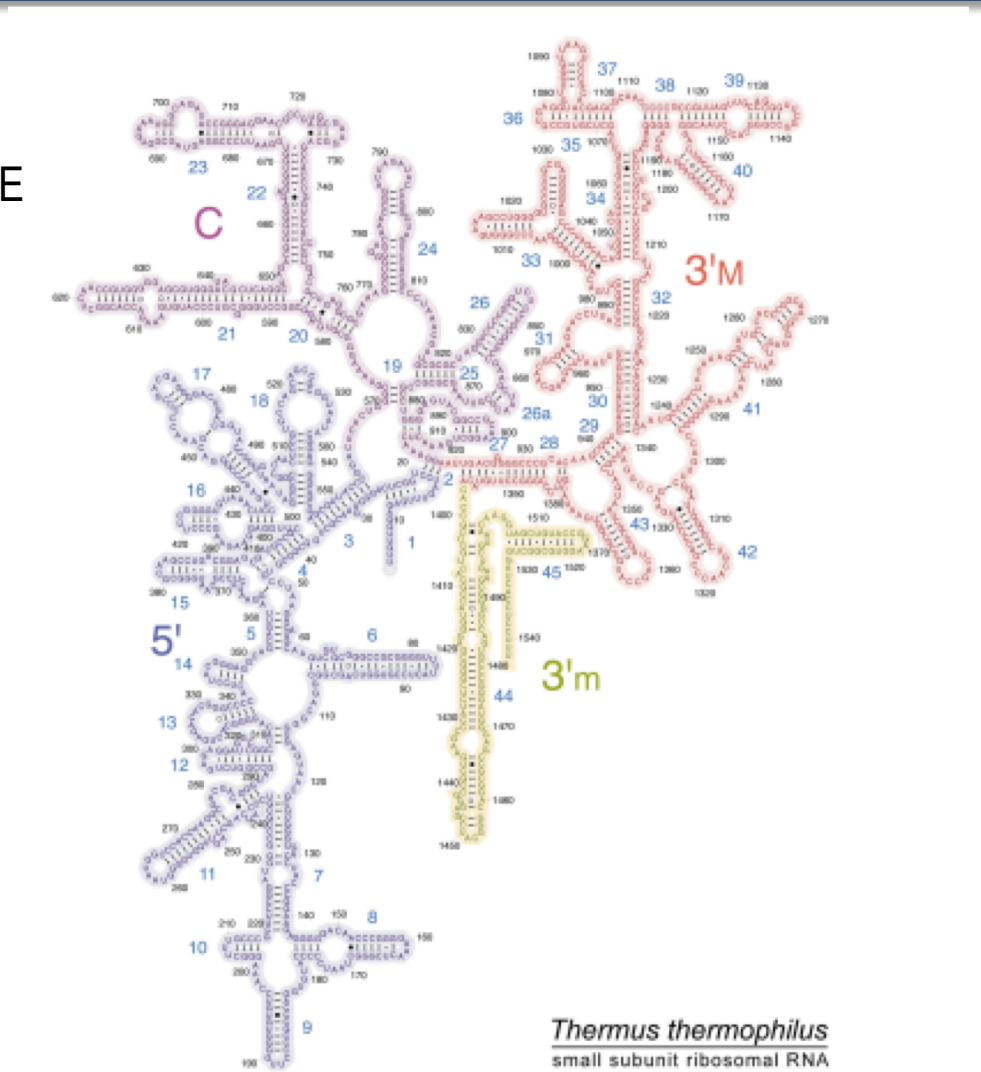
Gene finding
and annotation

Comparison

http://www.cbs.dtu.dk/courses/27626/

What is the 16S rRNA?

RIBOSOME SECONDARY STRUCTURE

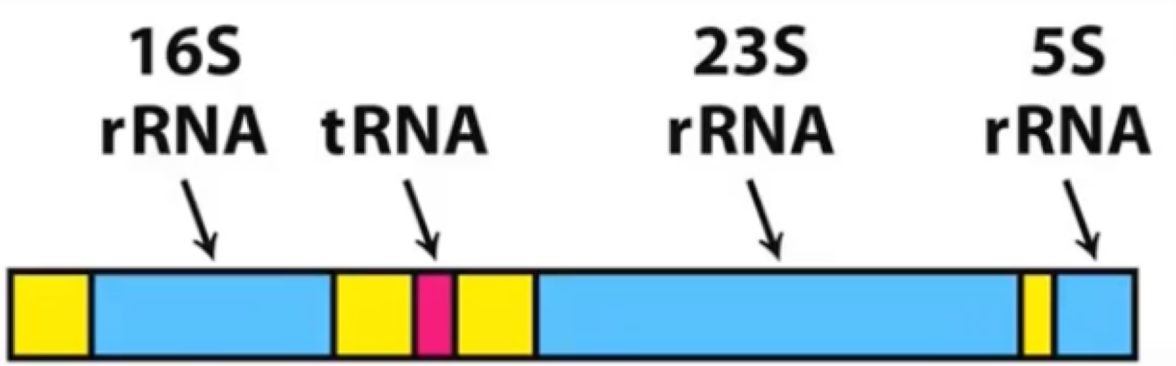


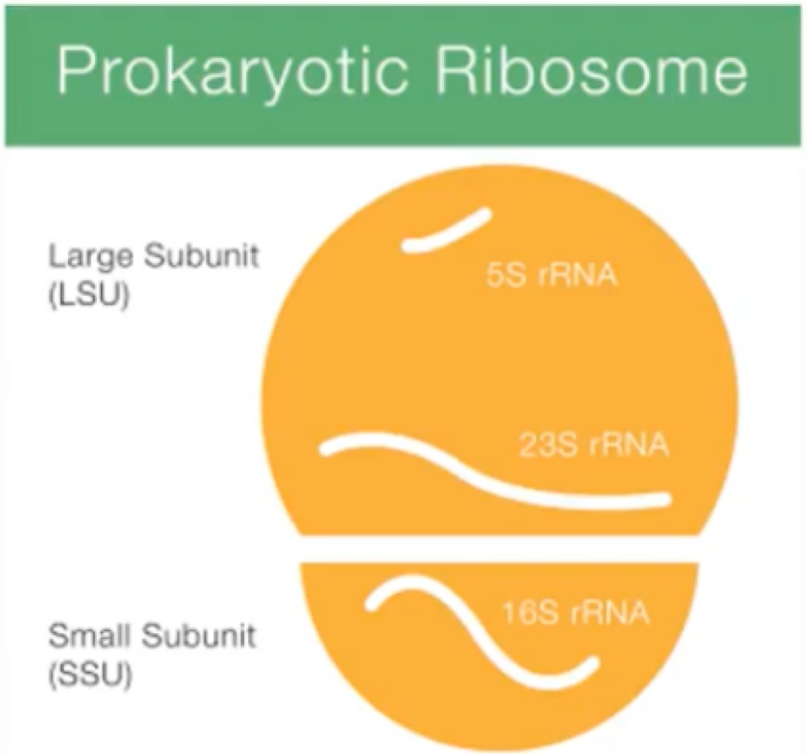
http://rna.ucsc.edu/rnacenter/ribosome_images.html

What is the 16S rRNA?

rRNA:

1. Universality
2. Activity in cellular functions
3. Extremely conserved structure and sequence





CD Genomics

What is the 16S rRNA?

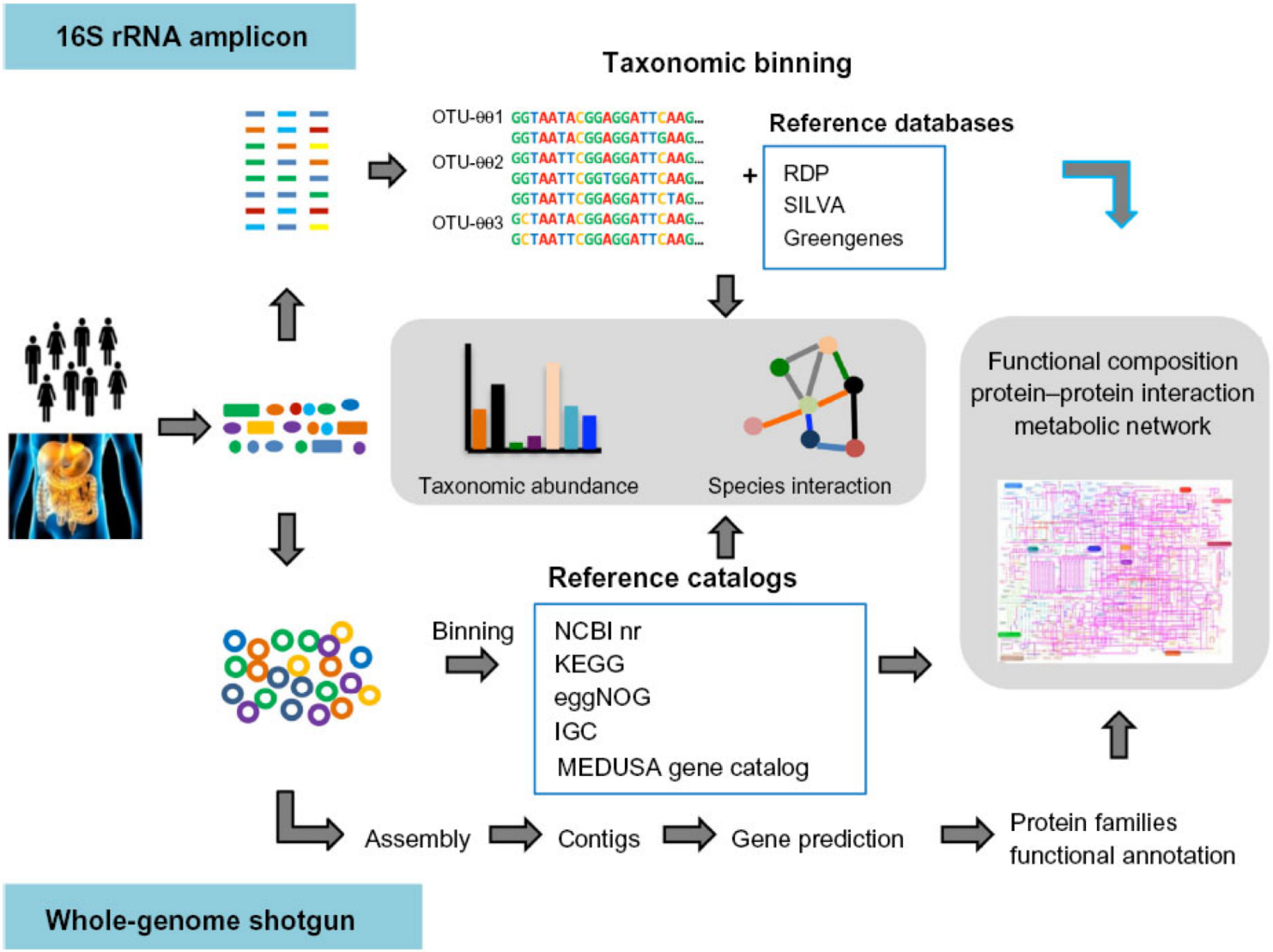
- This is a component of the small (30S) subunit or prokaryotic ribosomes (part of the RNA translate machinery).
- Ribosomes (and DNA that codes for them) have been mostly conserved over time
- Relatively short (1.5Kb), easy, fast and cheap to sequence (relative to other genes)
- It also stabilizes anti-codon pairing and the structure of the 23S rRNA, which is the other half of the 30S subunit.
- The term 16S refers to how it settles to when centrifuged (it's called a sedimentation rate, and it's measured in Svedberg (S) units).
  - The svedberg is actually a measure of time; it is defined as exactly 10^{-13} seconds (100 fs (femtosecond – 10^{-15})).

General principle of WMS

Whole Metagenome Sequencing (Shotgun Metagenomics)

- Relatively new and powerful technique
- Sequences all the genomic material present in the environment
- Increases the resolution, and allows the discovery of archaea and viruses
- BUT is more expensive than 16s, produces a lot more data

Overview of bioinformatics methods

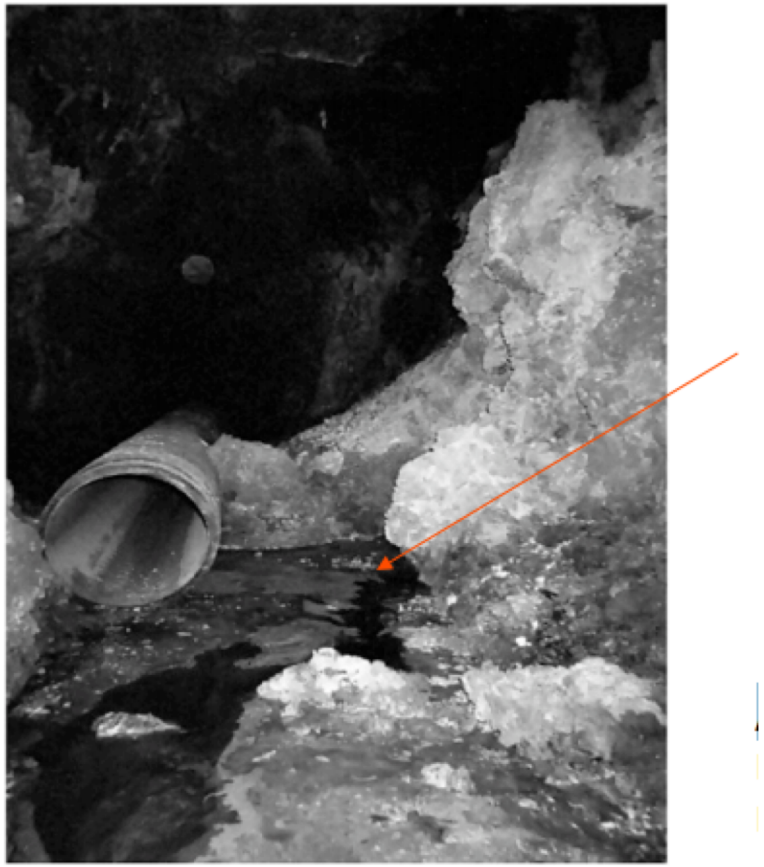


https://doi.org/10.2147/AGG.S57215

Metagenomics History

- The early stages of environmental ‘metagenomic’ studies targeted 16S rRNA (Phylogenic Tags) genes to obtain better picture of the species composing the community. Now WGS dominates metagenomics.
- The first well-analyzed WGS sequencing metagenomic study of microbial genomes from the environment focused on **acid mine drainage (AMD)** microbial biofilm by Jill Banfield and colleagues.

The Acid Mine Drainage (AMD) Project

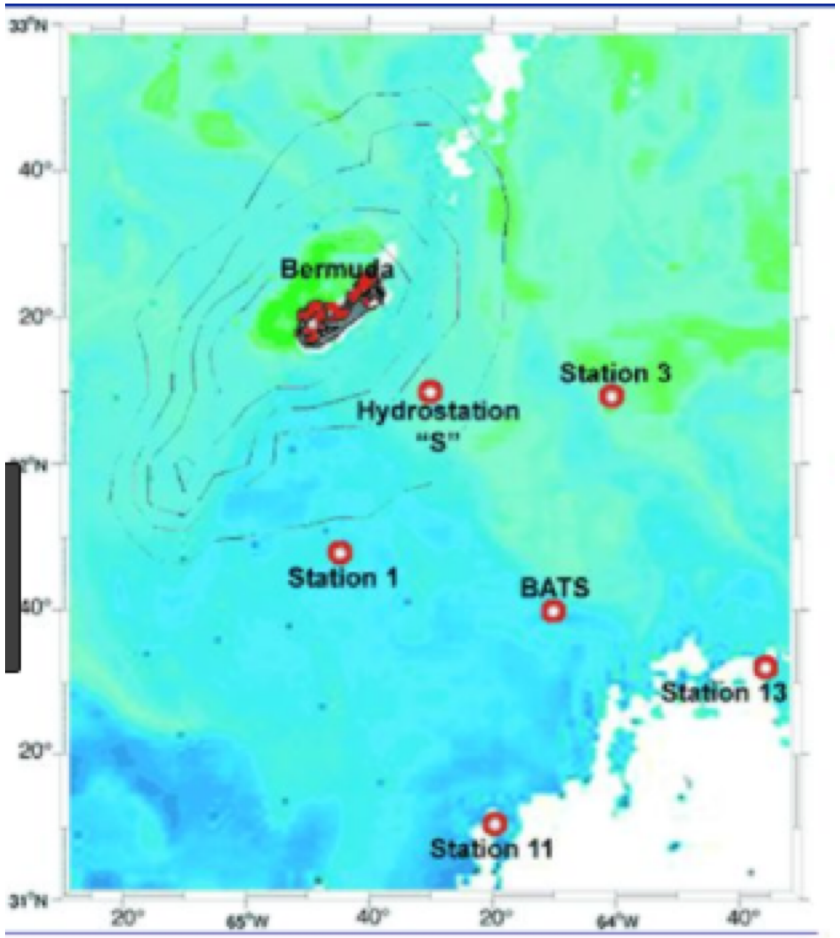


- Biofilms growing on the surface of flowing AMD in the five-way region of the Richmond mine at Iron Mountain, California, were sampled in March 2000
  - Acid is produced by oxidation of sulfide minerals that are exposed to air as a result of mining activity

Why AMD Biofilm?

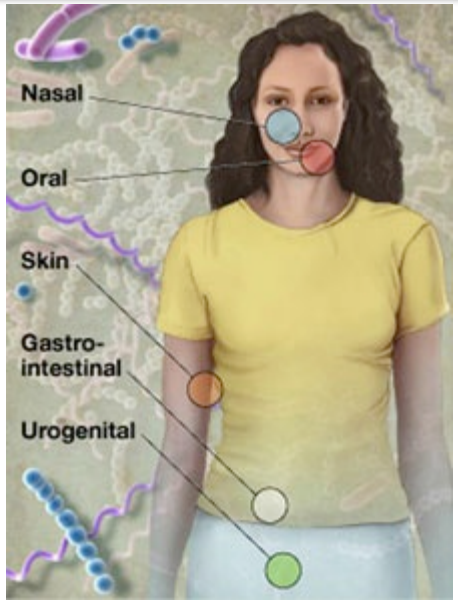
- Extreme acidic environment (self-contained ecosystem)
- Scientists are interested in the metabolic potential of such an environment: nitrogen fixation, sulfur oxidation, and iron oxidation
- To understand how the microbes tolerate the extremely acidic environment
- And it is a good pick -- low species complexity

The Sargasso Sea Project



- Science 2004, 304:66-74
- A pilot project of Venter's global ocean voyage (with the ultimate goal of finding solutions to energy problems)
- J. Craig Venter is not the first to sequence the genes of microbes from the ocean; but he is the first to do it on a large and ambitious scale
- Sargasso Sea takes its name from the sargassum seaweed that floats on its surface
- It is in the middle of the North Atlantic Ocean near Bermuda.
- Identified over 1.2 Million unknown genes

Human Microbiome Project (HMP)



NIH HUMAN MICROBIOME PROJECT

- From 242 healthy U.S
- More than 5,000 sam
- 3000 reference geno genomes are available

image courtesy of the NIH Common Fund

Human Microbiome Project (HMP)

- Humans have more bacterial cells (10^{14}) in habiting our body than our own cells (10^{13}), and 100 more bacterial genes.
  - But a recent paper revealed that the ratio is 1.3 (38 trillion bacterial cells vs. 30 trillion human cells from a reference man) : Sender, R., S. Fuchs, and R. Milo, Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol*, 2016. 14(8): p. e1002533.
- Consists of archaea, bacteria, and viruses.
- What are the composition and gene content of human microbiome?
- What are the differences of microbiome composition across individuals?
- What are the differences of microbiome composition across body parts?
- In Europe, the MetaHit project focused mainly on human gut microbiome.

Metagenomics of the human gut

Vol 464|4 March 2010|doi:10.1038/nature08821 nature

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.

Diagnostic Metagenomics

Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections

M. J. PALLEN*

Division of Microbiology and Infection, Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK

(*Received 21 November 2013; revised 16 January 2014; accepted 16 January 2014*)

SUMMARY

The term 'shotgun metagenomics' is applied to the direct sequencing of DNA extracted from a sample without culture or target-specific amplification or capture. In diagnostic metagenomics, this approach is applied to clinical samples in the hope of detecting and characterizing pathogens. Here, I provide a conceptual overview, before reviewing several recent promising proof-of-principle applications of metagenomics in virus discovery, analysis of outbreaks and detection of pathogens in contemporary and historical samples. I also evaluate future prospects for diagnostic metagenomics in the light of relentless improvements in sequencing technologies.

Key words: metagenomics, high-throughput sequencing, diagnosis.

Toilet Waste from Long Distance Flights?

SCIENTIFIC REPORTS

OPEN

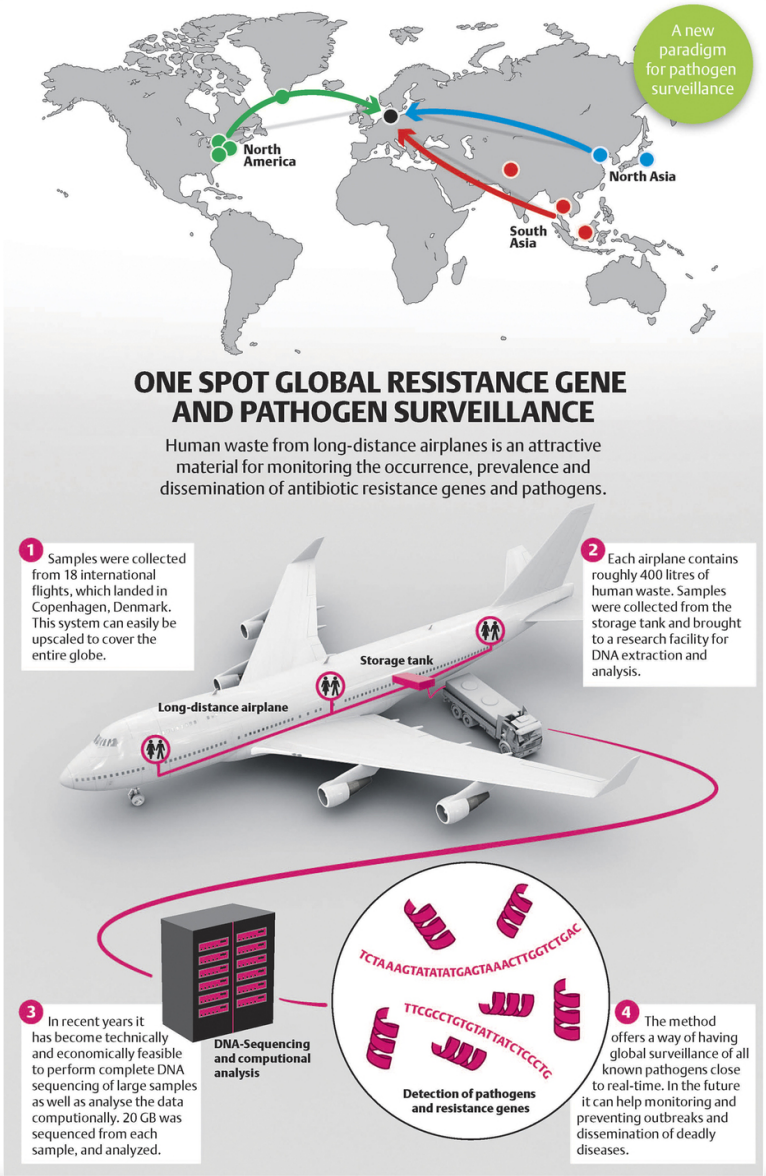
Received: 17 December 2014
Accepted: 17 April 2015
Published: 10 July 2015

Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance

Thomas Nordahl Petersen[1], Simon Rasmussen[1], Henrik Hasman[2], Christian Carøe[1], Jacob Bælum[1], Anna Charlotte Schultz[2], Lasse Bergmark[2], Christina A. Svendsen[2], Ole Lund[1], Thomas Sicheritz-Pontén[1] & Frank M. Aarestrup[2]

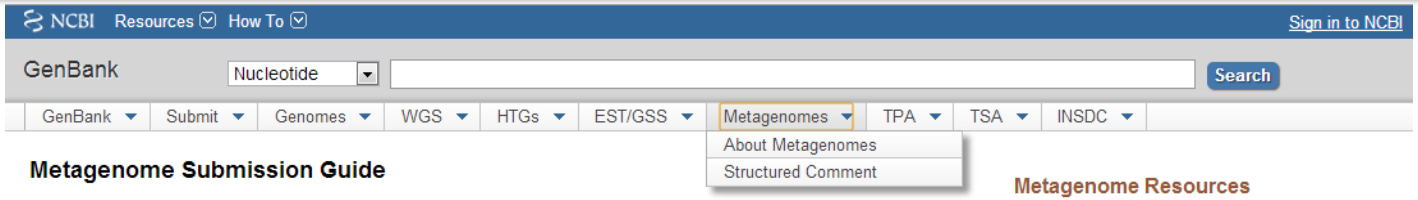
Human populations worldwide are increasingly confronted with infectious diseases and antimicrobial resistance spreading faster and appearing more frequently. Knowledge regarding their occurrence and worldwide transmission is important to control outbreaks and prevent epidemics. Here, we performed shotgun sequencing of toilet waste from 18 international airplanes arriving in

Toilet Waste from Long Distance Flights?

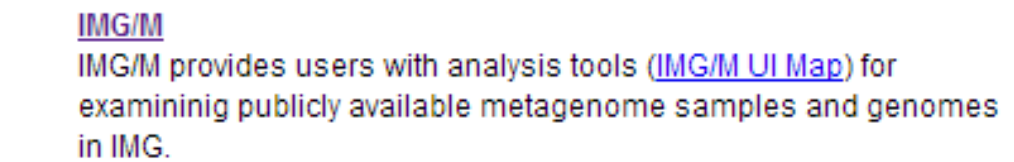


Metagenomic Database

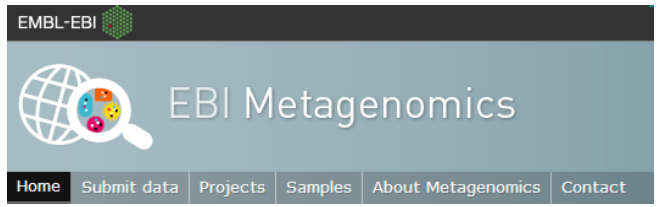
- NCBI



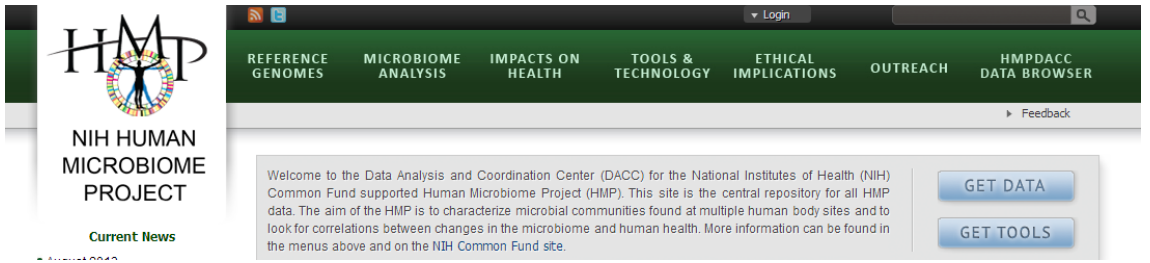
- IMG/M (JGI)



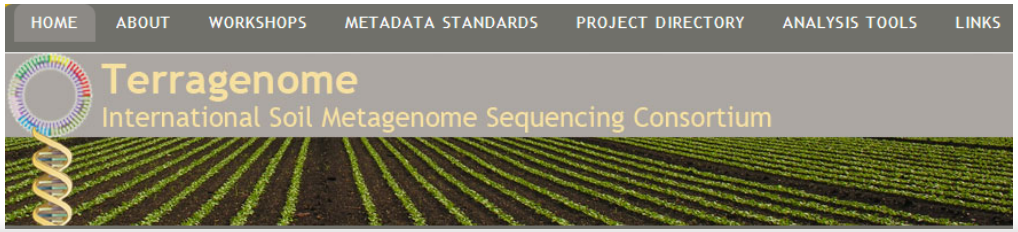
- EBI Metagenomics



- MG-RAST
- Kbase
- Human

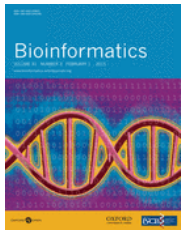


- Soil

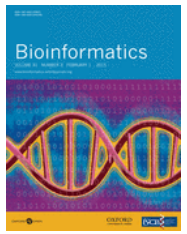


Metagenomic Analysis

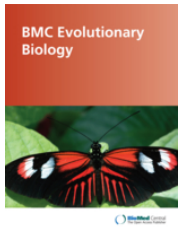
- **Taxonomic classification (reference based)**
  - similarity-based (MEGAN, Carma, Pathoscope, **Sigma**)
  - composition-based (PhyloPythia, TACOA , Phymm)
  - conserved marker genes (MetaPhlAn, Kraken)
  - web analysis (IMG/MER, MG-RAST, EBI Metagenomics)



- ***De novo* assembly (reference free)**
  - widely used: MetaVelvet, SoapDenovo, IDBA
  - newborns : RayMeta, **Omega**



- **Specific analysis**
  - identify rRNA sequences (rRNASelector)
  - functional analysis of protein coding sequence (InterPro, **Unifam**)



References

Read below metagenomics paper (do not need to return your review report)

- Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation, 2, 3. http://doi.org/10.1186/2042-5783-2-3

Additional Slides

What genomes are in your metagenomic sample?

- Metagenomics analysis with references

BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 170–177
doi:10.1093/bioinformatics/btu641

Genome analysis

Advance Access publication September 29, 2014

Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance

Tae-Hyuk Ahn, Juanjuan Chai and Chongle Pan*

Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Metagenomic sequencing of clinical samples provides a promising technique for direct pathogen detection and characterization in biosurveillance. Taxonomic analysis at the strain level can be used to resolve serotypes of a pathogen in biosurveillance. Sigma was developed for strain-level identification and quantification of pathogens using their reference genomes based on metagenomic analysis.

Results: Sigma provides not only accurate strain-level inferences, but also three unique capabilities: (i) Sigma quantifies the statistical uncertainty of its inferences, which includes hypothesis testing of identified genomes and confidence interval estimation of their relative abun-

sequences for pathophysiology analysis. In the future, metagenomics could be used to sequence fecal samples of patients in the early stage of such an outbreak, potentially allowing more rapid and unambiguous identification of the pathogen.

It is a computational challenge to achieve sensitive and specific identification of pathogens in a complex metagenome background. Many algorithms have been developed to infer the taxonomic composition of a microbial community using metagenomic sequencing data. The algorithms generally used the following three approaches. (i) The *k*-mer statistics approach compares *k*-mer frequency profiles of metagenomic reads with those of or-

2011 German *E. coli* Outbreak.

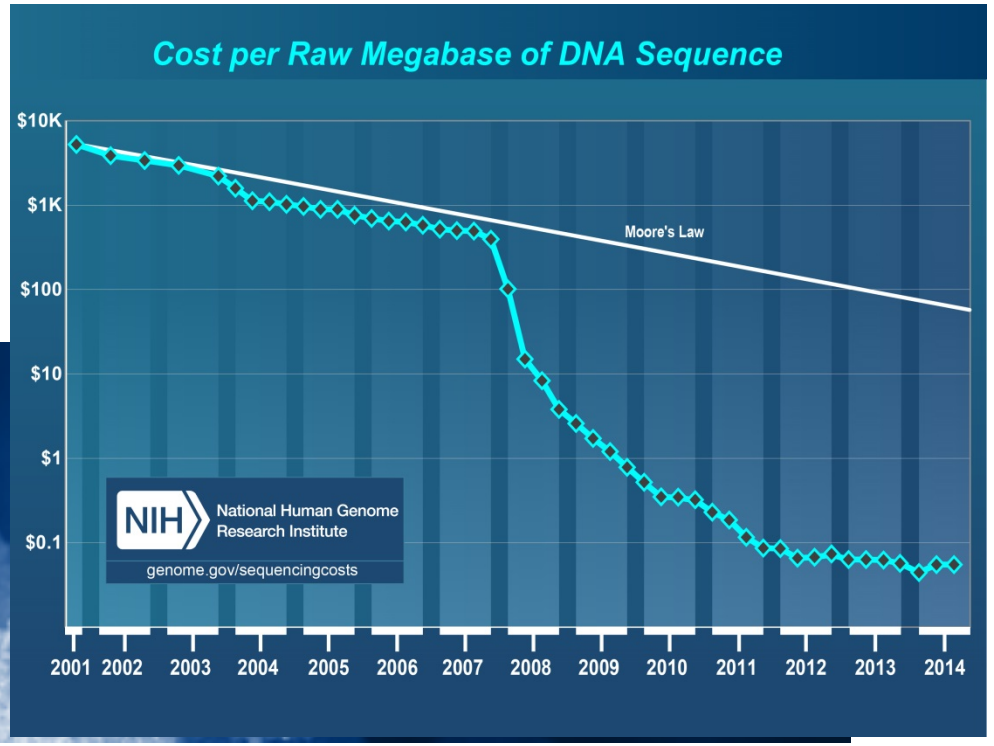
In May through June 2011, a novel strain of *Escherichia coli* O104:H4 bacteria caused a serious outbreak of foodborne illness.

In all, 3,950 people were affected and 53 died.

2014 Ebola Virus Disease

Metagenomics and Biosurveillance

- Conventional methods to identify pathogens: laboratory culture, immunoassays, and genotyping



- NGS metagenomic analysis can be a promising new method for biosurveillance.

Diagnostic Metagenomics

Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections

M. J. PALLEN*

Division of Microbiology and Infection, Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK

(*Received 21 November 2013; revised 16 January 2014; accepted 16 January 2014*)

SUMMARY

The term 'shotgun metagenomics' is applied to the direct sequencing of DNA extracted from a sample without culture or target-specific amplification or capture. In diagnostic metagenomics, this approach is applied to clinical samples in the hope of detecting and characterizing pathogens. Here, I provide a conceptual overview, before reviewing several recent promising proof-of-principle applications of metagenomics in virus discovery, analysis of outbreaks and detection of pathogens in contemporary and historical samples. I also evaluate future prospects for diagnostic metagenomics in the light of relentless improvements in sequencing technologies.

Key words: metagenomics, high-throughput sequencing, diagnosis.

Metagenomic data analysis for Biosurveillance

Requires:

- High-quality reference genomes
- Ability to resolve different strains of the same species
- Statistical confidence measurements

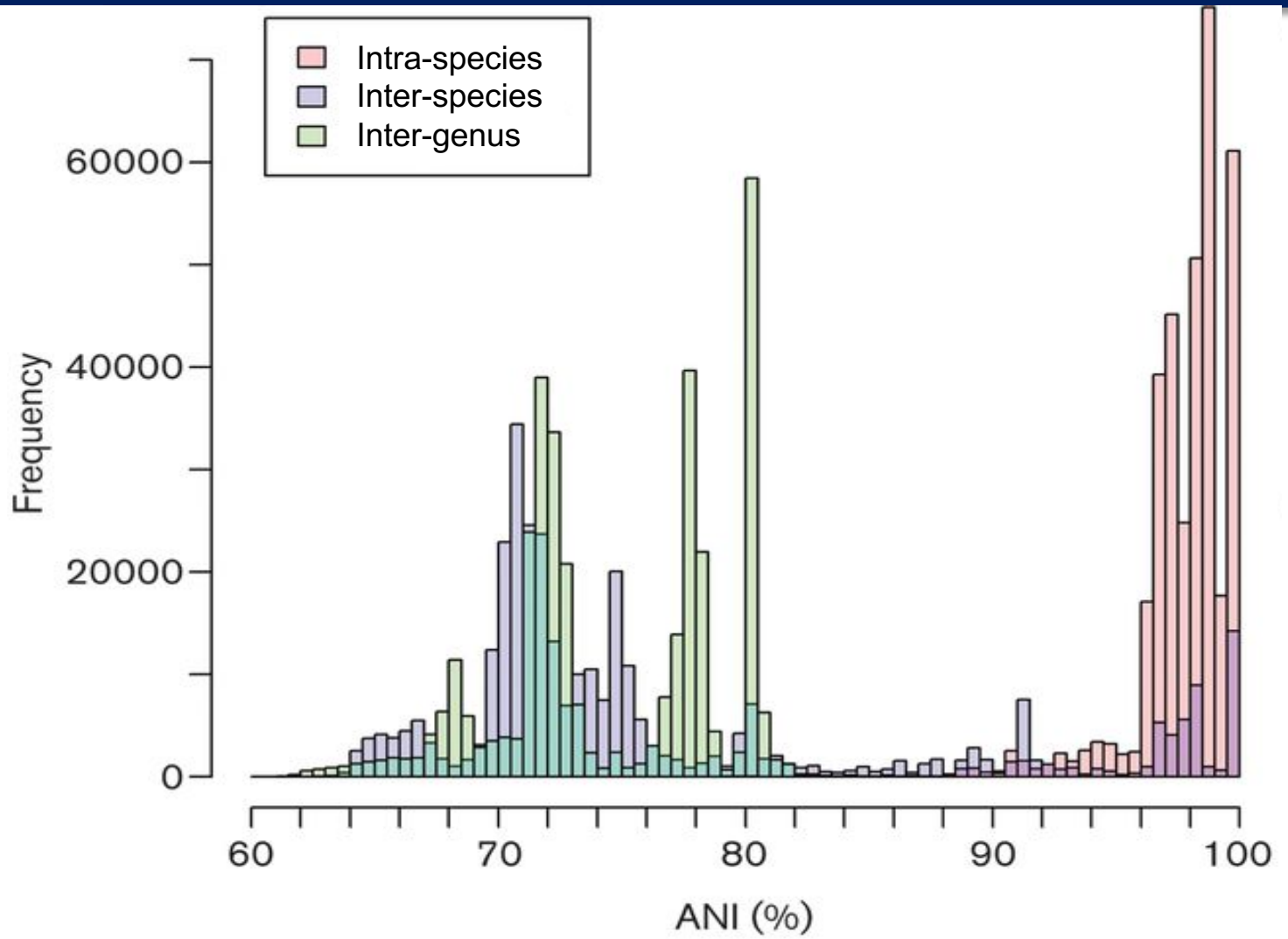
Most existing taxonomic classification methods:

- **species level or higher**

Current existing strain-level supporting methods:

- **slow, poor performance, no statistical confidence**

Strain Level Resolution Complexity



M. Kim et. al., "Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes", *International Journal of Systematic and Evolutionary Microbiology* (2014), 64, 346–351

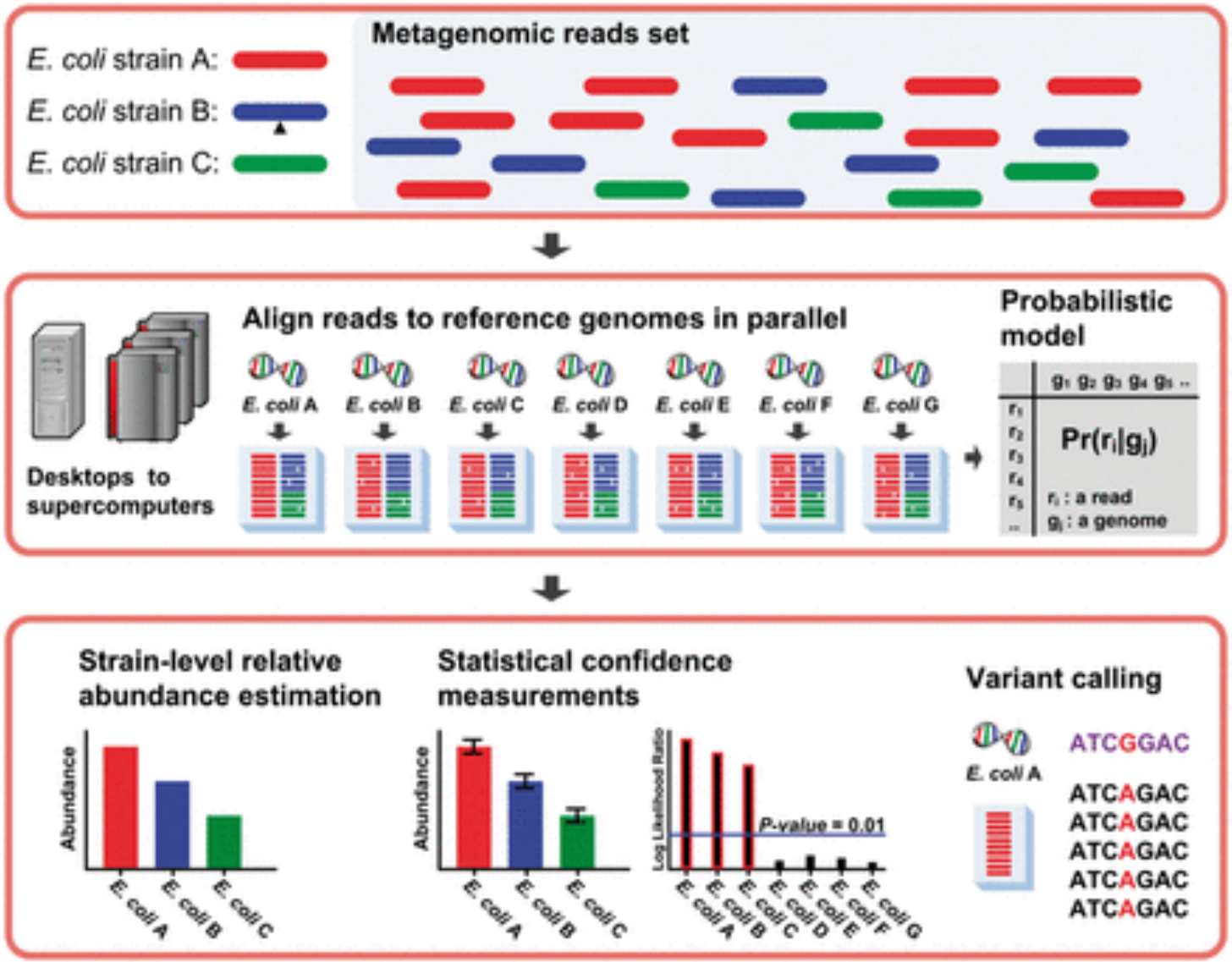
Sigma

- We present **Sigma** (http://sigma.omicsbio.org)

  A novel sequence similarity-based approach for Strain-level identification of genomes from metagenomic analysis for biosurveillance and taxonomic profiling

- **Significance and Impact**:
  - Unique capability of identifying the correct strain of a pathogen in a complex metagenomic background from many closely related candidates in the reference genome database.
  - Using a top open-science supercomputer, Titan, pathogenic bacteria strains can be identified in an hour from the 100 million human microbiome sequences.

Sigma


E. coli strain A:
E. coli strain B:
E. coli strain C:
Metagenomic reads set
Desktops to supercomputers
Align reads to reference genomes in parallel
E. coli A
E. coli B
E. coli C
E. coli D
E. coli E
E. coli F
E. coli G
Probabilistic model
Pr(r_i|g_j)
r_i : a read
g_j : a genome
Strain-level relative abundance estimation
Abundance
Statistical confidence measurements
Log Likelihood Ratio
P-value = 0.01
Variant calling
ATCGGAC
ATCAGAC
ATCAGAC
ATCAGAC
ATCAGAC
ATCAGAC