Sequencing and applications

IN-BIOS5000/9000
Genome Sequencing Technologies, Assembly, Variant Calling and Statistical Genomics
17 October 2022

Torbjørn Rognes
Dept. of Informatics, UiO & Dept. of Microbiology, OUS
torognes@ifi.uio.no

UiO : University of Oslo

Oslo
University Hospital

Overview

- Sequencing technologies
- Important sequencing properties
- Developments in sequencing
- Paired-end reads & mate pair sequencing
- Overview of main applications
- Whole genome *de novo* sequencing and assembly
- Resequencing & variant calling
- Other applications: Metagenomics & RNA-Seq
- Challenges

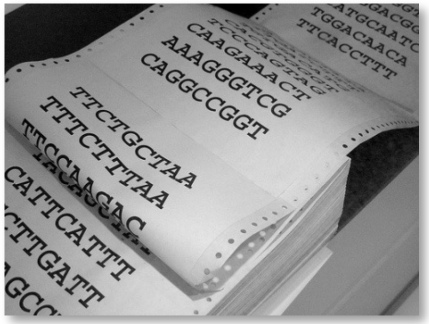
DNA sequencing

High-Throughput Sequencing (HTS), Deep sequencing, Next Generation Sequencing (NGS)

Obtain sample

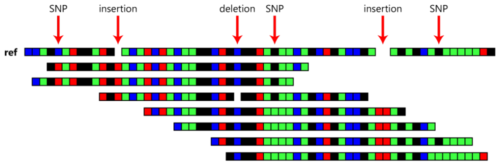
DNA extraction, preparation

Sequencing



Raw data

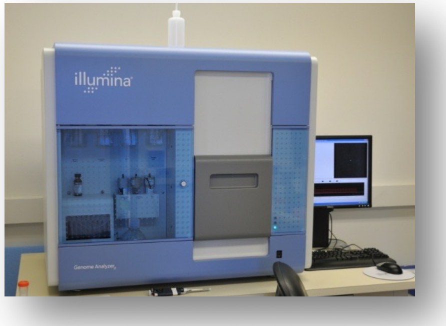
Sequence analysis



Interpretation

Illumina

- Sequencing by synthesis using fluorescence
- One fragment = one cluster = one read
- Read lengths up to 300bp, paired-end reads
- Dominant technology today
- Formerly known as Solexa
- NovaSeq 6000 specifications:
  - 6000 billion bases per run (2 days)
  - Up to 20 billion single reads or
    40 billion paired-end reads per run
  - Up to 2x250 bp
  - ~48 human genomes (40X) in 2 days



GA IIx

Sanger sequencing centre

PacBio

- Pacific Biosciences RS II and Sequel systems
- Long reads
- Single molecule (no PCR)
- Uses a "zero-mode waveguide (ZMW)"
- High error rate if not corrected
- Sequel II HiFi performance:
  - Average read length 30 000 – 100 000 bases
  - Throughput up to 500 GB per SMRT cell, raw reads
  - Up to 50 GB corrected reads (>Q20)

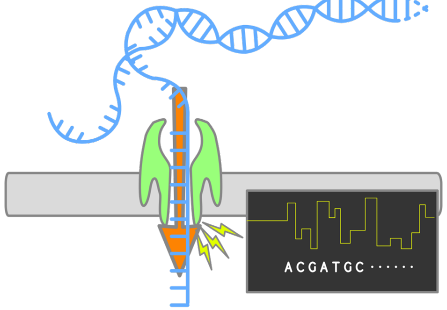
PacBio Sequel

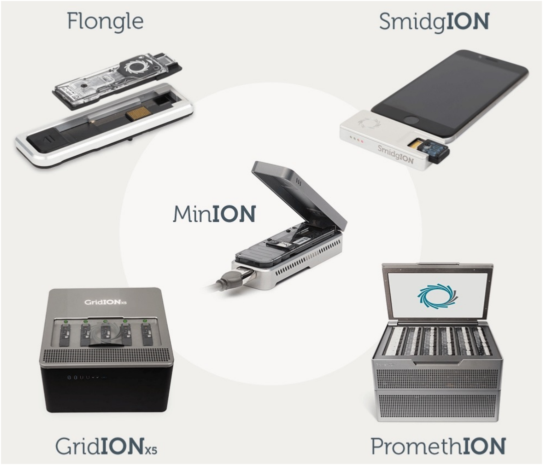
PacBio RSII

Oxford Nanopore

- Oxford Nanopore systems
  - MinION
  - PremethION
  - and others
- Various equipment from small portable to large high-capacity
- DNA is passing through a nanopore and voltage potential is measured
- Single molecule, no PCR
- Long reads
- High uncorrected error rate
- Varying capacity

Oxford Nanopore





Older sequencing technologies

Roche (454)

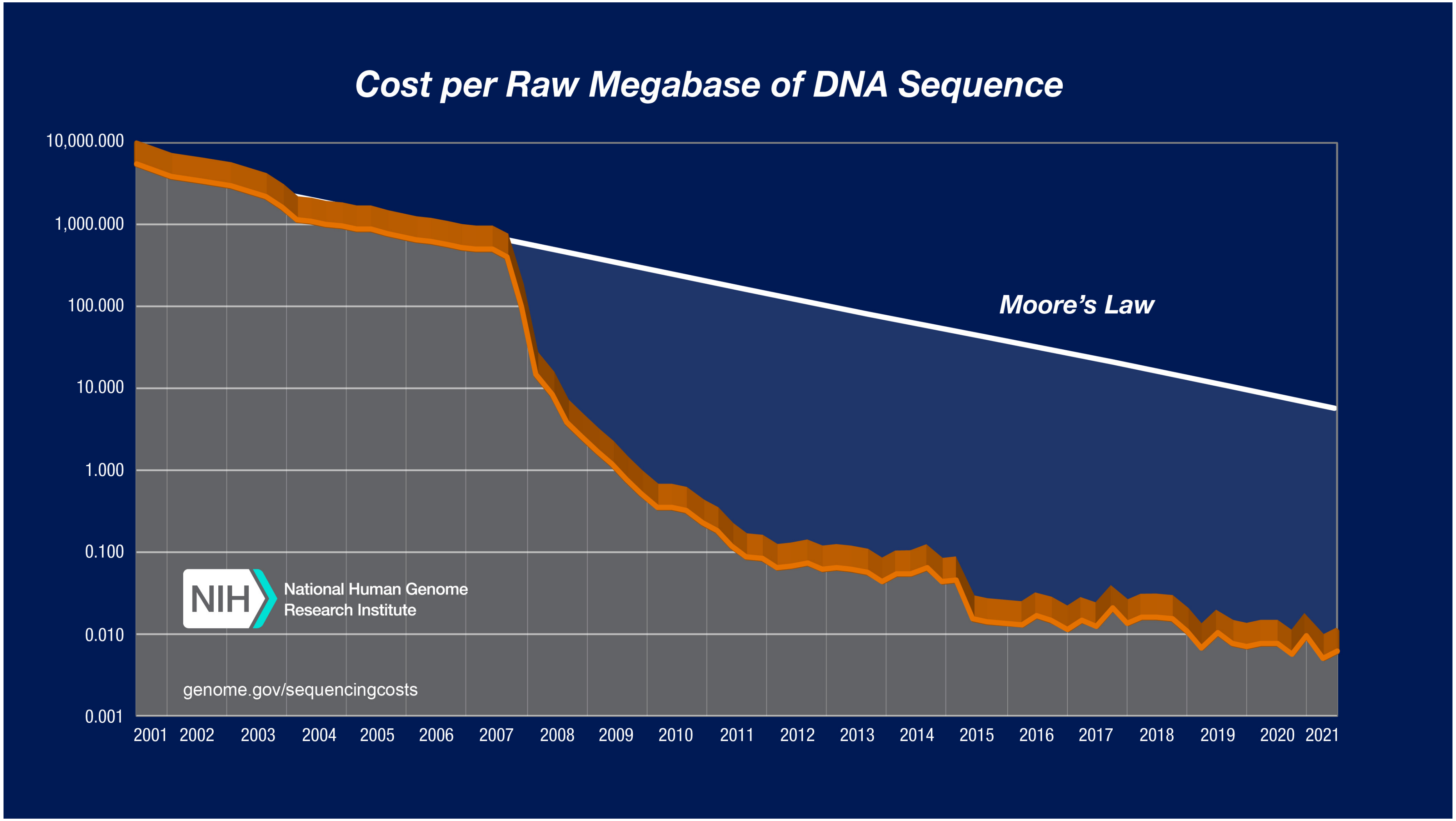
ABI (SOLiD)

Ion Torrent

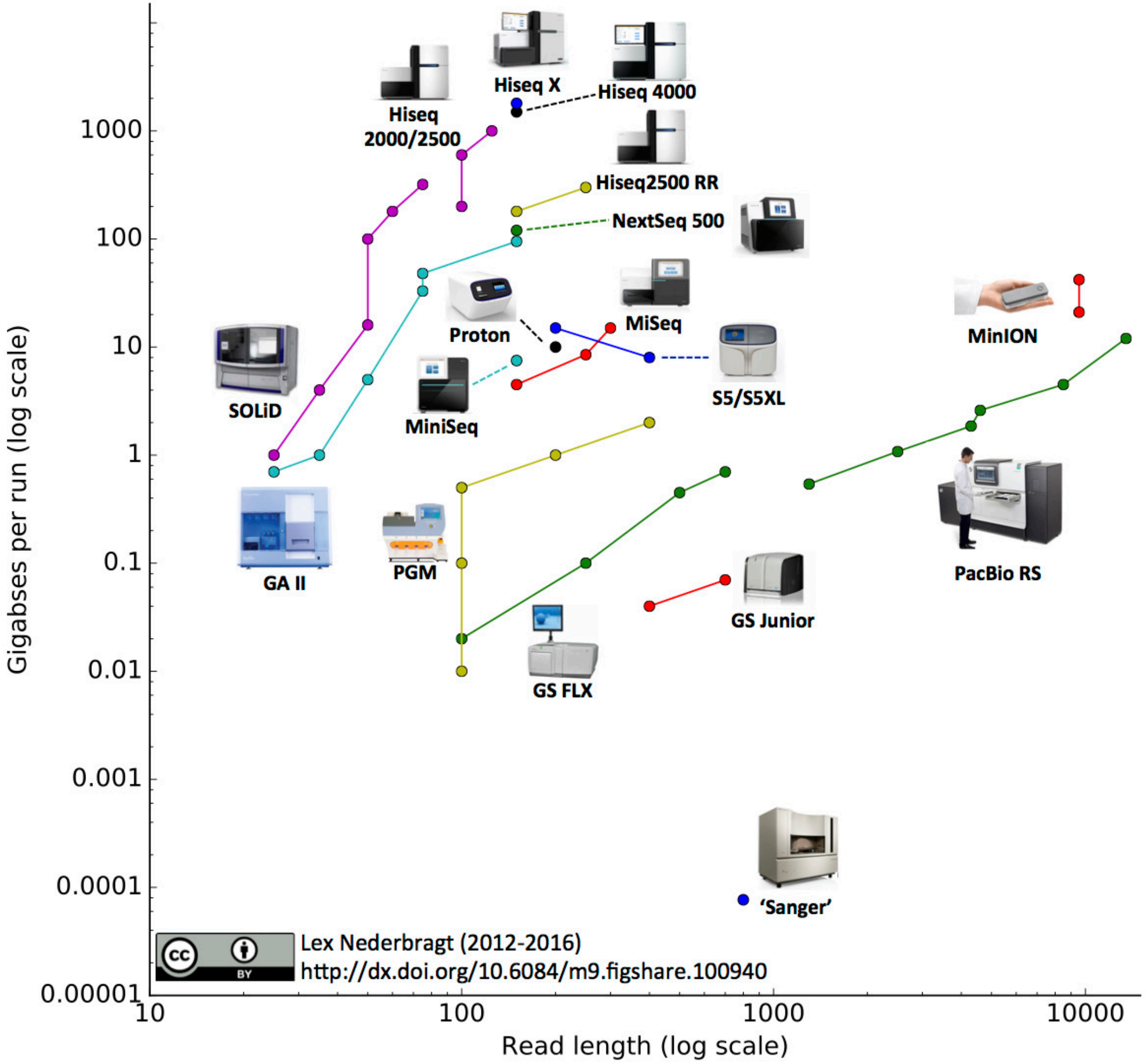
Important technology properties

- Cost
  - Per base
  - Investment
- Read length
- Speed / capacity (bases per day)
- Sequencing errors
  - Frequency
  - Profile (indels, substitutions)
  - Random or systematic?
- Paired-end support
- Single molecule or PCR-based
- Amount of lab work necessary
- Portability of equipment

The cost of sequencing

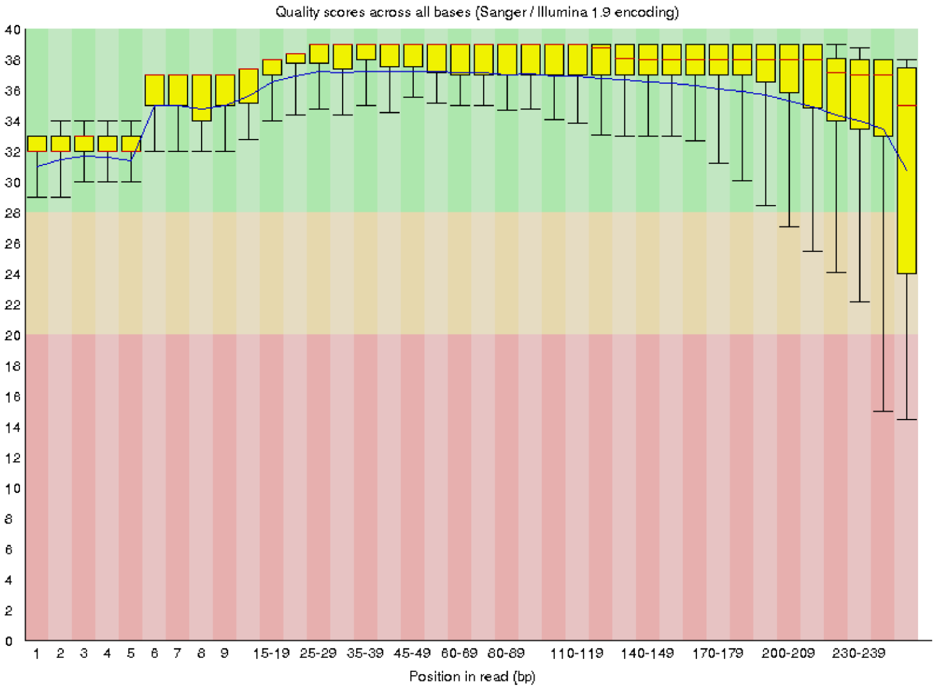


Sequencing technology development

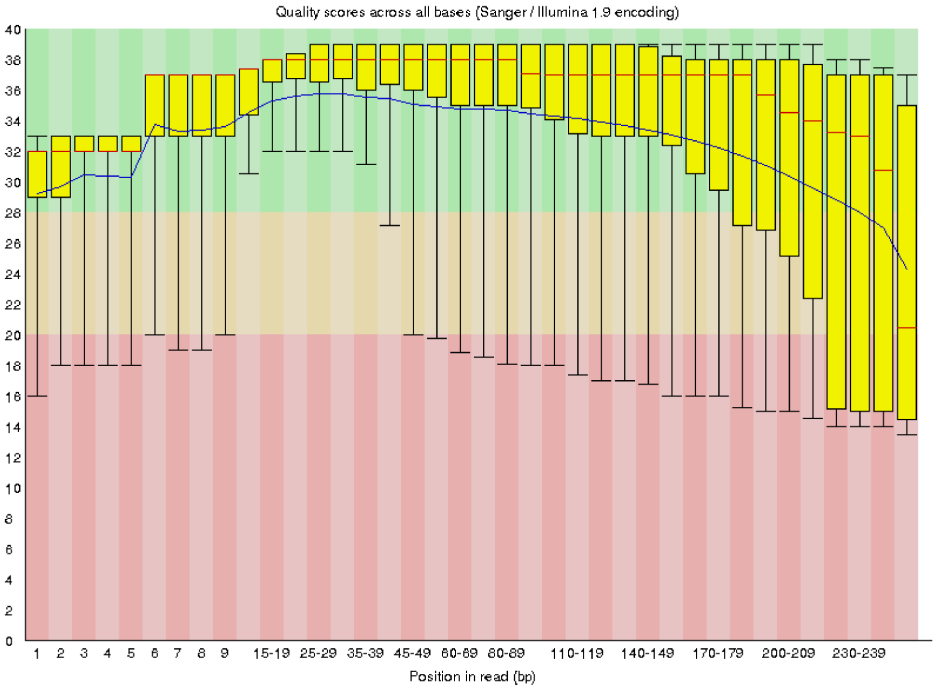


Source: Lex Nederbragt (2012-2016) https://doi.org/10.6084/m9.figshare.100940

Quality plots of Illumina MiSeq reads



Forward reads

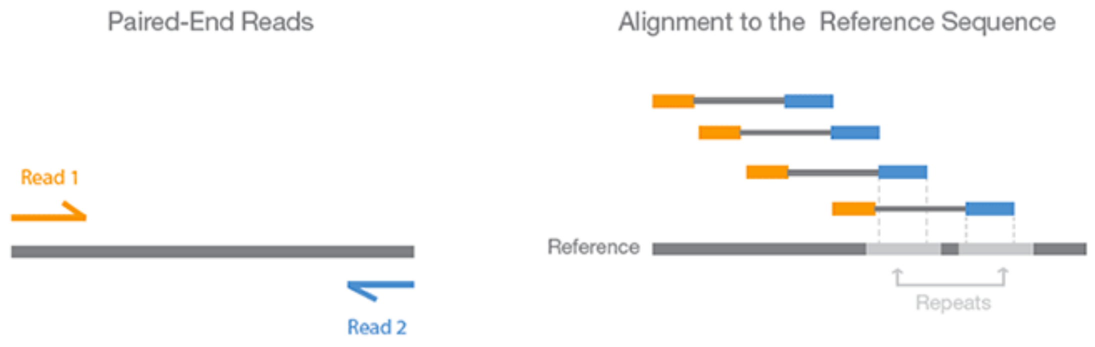


Reverse reads

Paired-end / mate pair sequencing

- Paired-end reads or mate pair reads are pairs of reads known to come from the same regions in the genome within a certain fixed distance
- Typically paired ends are a ~100-500bp apart, while mate pairs are ~2-10kb apart
- Performed by sequencing fragments from both ends
- Alleviates problems of short reads in repetitive genomic regions

Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Common HTS applications

***De novo* genome sequencing**	Determining the complete genome sequence of an organism for the first time
Whole genome re-sequencing and variant calling	Finding polymorphisms (SNPs) and discover mutations in an individual
Exome sequencing and variant calling	Sequencing only protein-coding regions of a genome from an individual to identify mutations or polymorphisms (SNPs)
Transcriptomics (RNA-seq)	Sequencing of expressed RNA (after reverse transcription to cDNA), (small RNA, mRNA or total RNA) to determine level
Chromatin immunoprecipitation-sequencing (ChIP-Seq) (ChIP-exo)	Mapping of genome-wide protein-DNA interactions
Methylation sequencing (Methyl-Seq)	Determining methylation patterns in the genome (epigenomics) (often on bisulfite-treated DNA)
Metagenomics	Sequencing the whole genomic DNA of multiple species (microorganisms) simultaneously from a certain environment
Metatranscriptomics	Sequencing RNA from multiple species (microorganisms) simultaneously
Amplicon sequencing	Sequencing of genomic regions selected and amplified by PCR, from multiple species simultaneously

Whole genome *de novo* sequencing

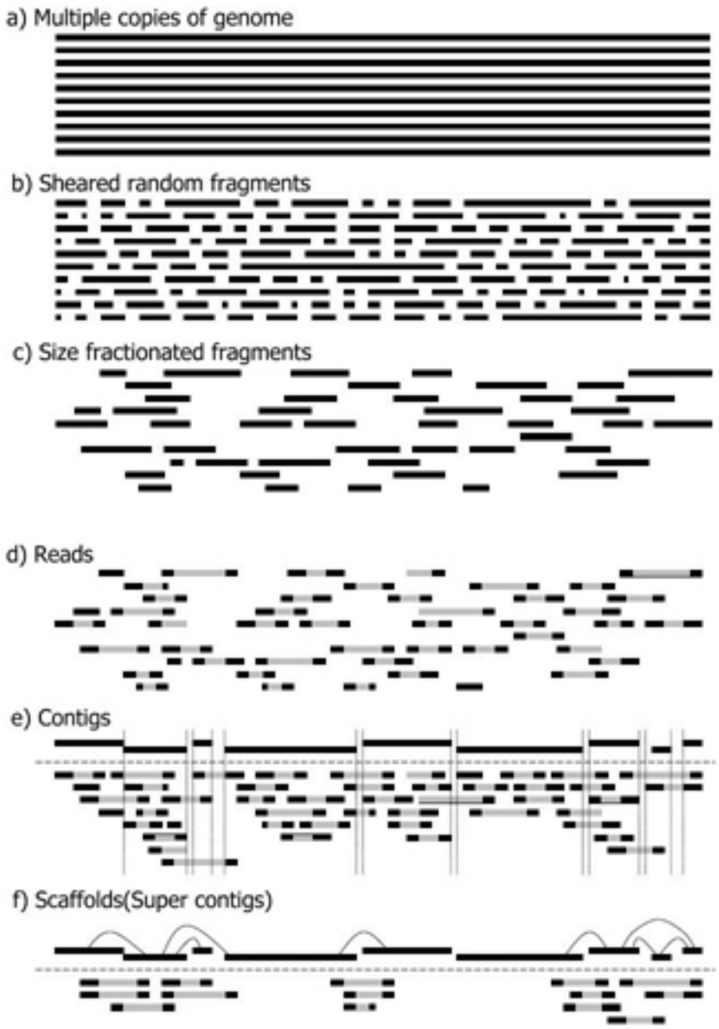
- Whole genome sequencing results in millions of small pieces of the full genome

- The challenge is to puzzle these together in the right order

- From reads to contigs, to scaffolds

- Genome sizes ranging from 2Mbp (bacteria) to 3Gbp (human) to 150Gbp (plant)

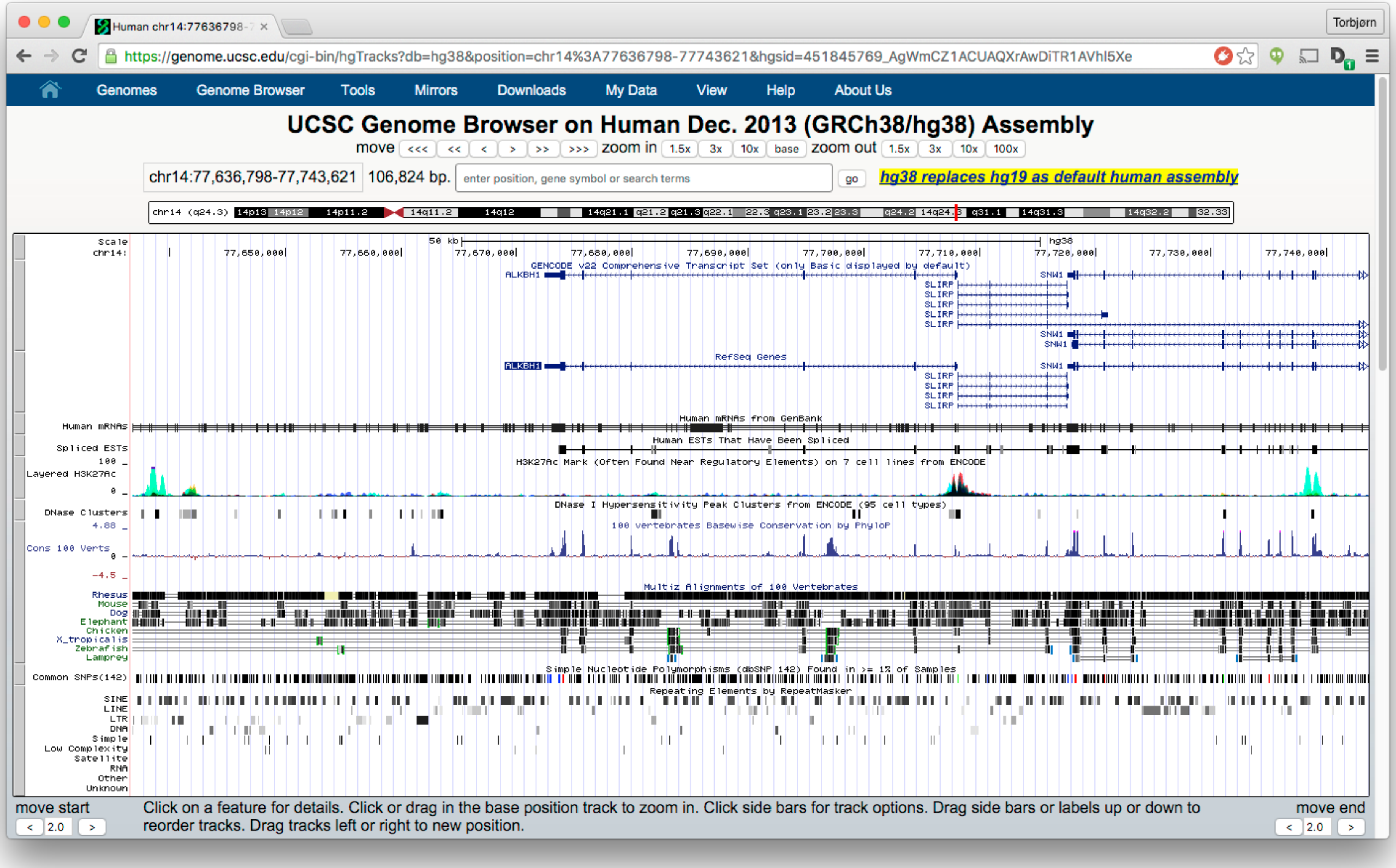
- Read size from 100 bp to 100 000 bp


a) Multiple copies of genome
b) Sheared random fragments
c) Size fractionated fragments
d) Reads
e) Contigs
f) Scaffolds(Super contigs)

Problematic issues

- Sequencing errors
  - Introduces false sequences into the assembly
  - May be alleviated by higher coverage / larger sequencing depth, or by error detection and correction
- Repeats
  - Genomes often contain many almost identical repeated sequences
  - Repeats longer than the read length makes it impossible to determine the exact location of the read
  - May cause compression or misassemblies
  - May be alleviated by longer reads or paired-end/mate pair reads
- Heterozygosity
  - Diploid organisms (e.g Humans) actually have two “genomes”, not one. Chromosome pairs 1-22 for all, plus XX or XY. One set of chromosomes from our mother and one from our father.
  - The two are mostly identical, but there are some differences
  - Causes “bubbles” in the assembly

Genome browsers



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr14:77,636,798-77,743,621 106,824 bp.

hg38 replaces hg19 as default human assembly

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Human chr14:77636798-7
Torbjørn
https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&position=chr14%3A77636798-77743621&hgsid=451845769_AgWmCZ1ACUAQXrAwDiTR1AVhI5Xe
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x 100x
enter position, gene symbol or search terms
go
chr14 (q24.3)
Scale
chr14:
50 kb
hg38
77,650,000
77,660,000
77,670,000
77,680,000
77,690,000
77,700,000
77,710,000
77,720,000
77,730,000
77,740,000
GENCODE v22 Comprehensive Transcript Set (only Basic displayed by default)
ALKBH1
SLIRP
SNW1
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(142)
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
move start
move end
< 2.0 >

Source: genome.ucsc.edu

Mapping reads to a reference genome

Goal: Identify positions in the genome that are most similar to the sequence reads

Input data:

- 10-1000 million reads, each 30-300bp
- Sequencing errors (typ. ~1% error rate)

Reference genome:

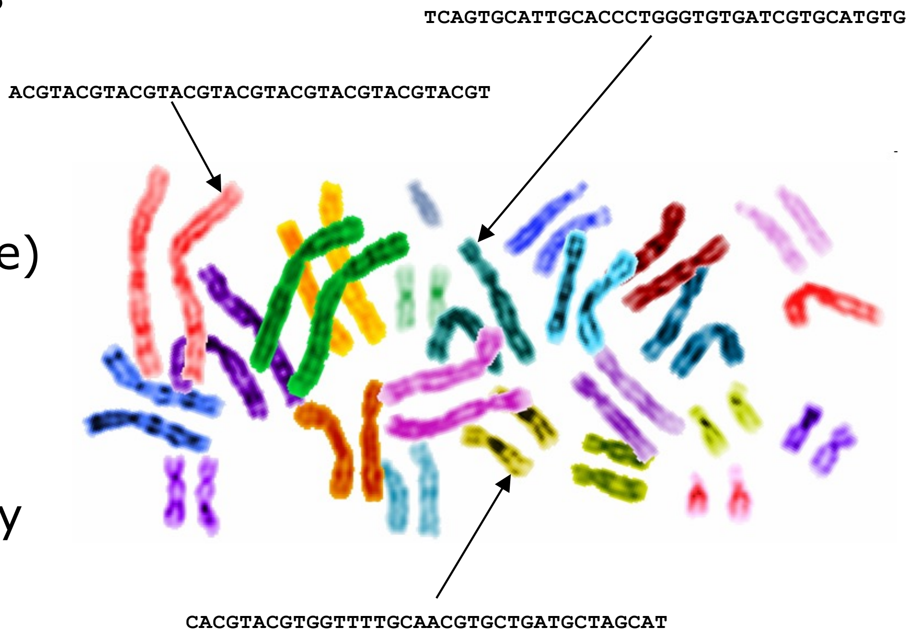
- E.g. human genome, 3 Gbp
- Some genome variation, heterozygocity

Output:

- 0, 1, or more potential genomic locations for each read
- Mapping quality assignment

Requirements:

- Sensitivity, specificity, speed, compactness



TCAGTGCATTGCACCCTGGGTGTGATCGTGCATGTG

ACGTACGTACGTACGTACGTACGTACGTACGTACGT

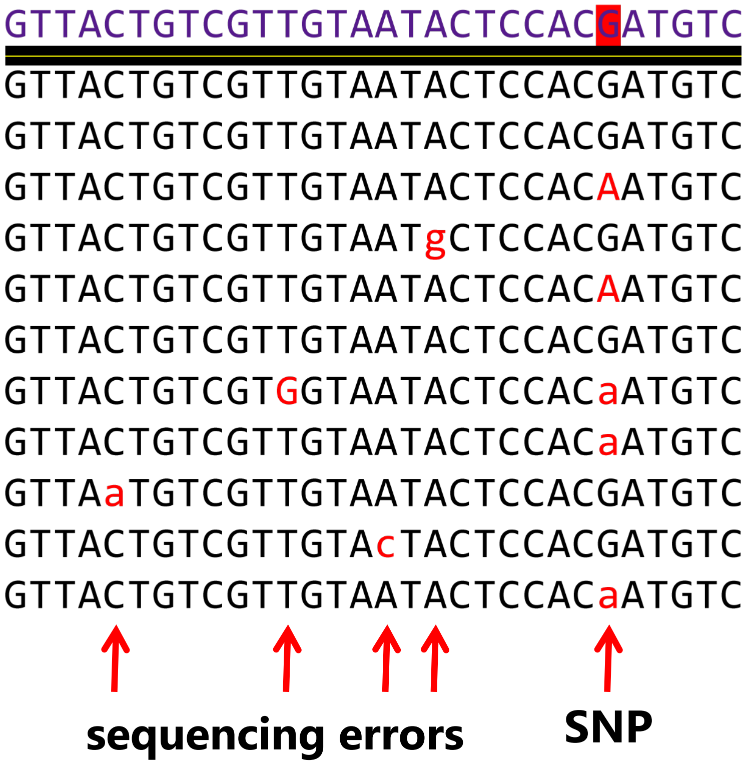
CACGTACGTGGTTTTGCAACGTGCTGATGCTAGCAT

Variation discovery by resequencing

- Variants may be called after mapping reads to a reference genome
- High coverage required, that is, the average number of times each base is sequenced (typically 40-100X)

- Natural variation discovery
- Mutation detection

- Single Nucleotide Polymorphisms (SNPs) and variants (SNVs)
- Small insertions & deletions (indels)
- Copy Number Variation (CNV)
- Large inversions, translocations etc



```
GTTACTGTCGTTGTAATACTCCACGATGTC

GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACAATGTC
GTTACTGTCGTTGTAATgCTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACAATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTGGTAATACTCCACaATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC
GTTAaTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAcTACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC
```

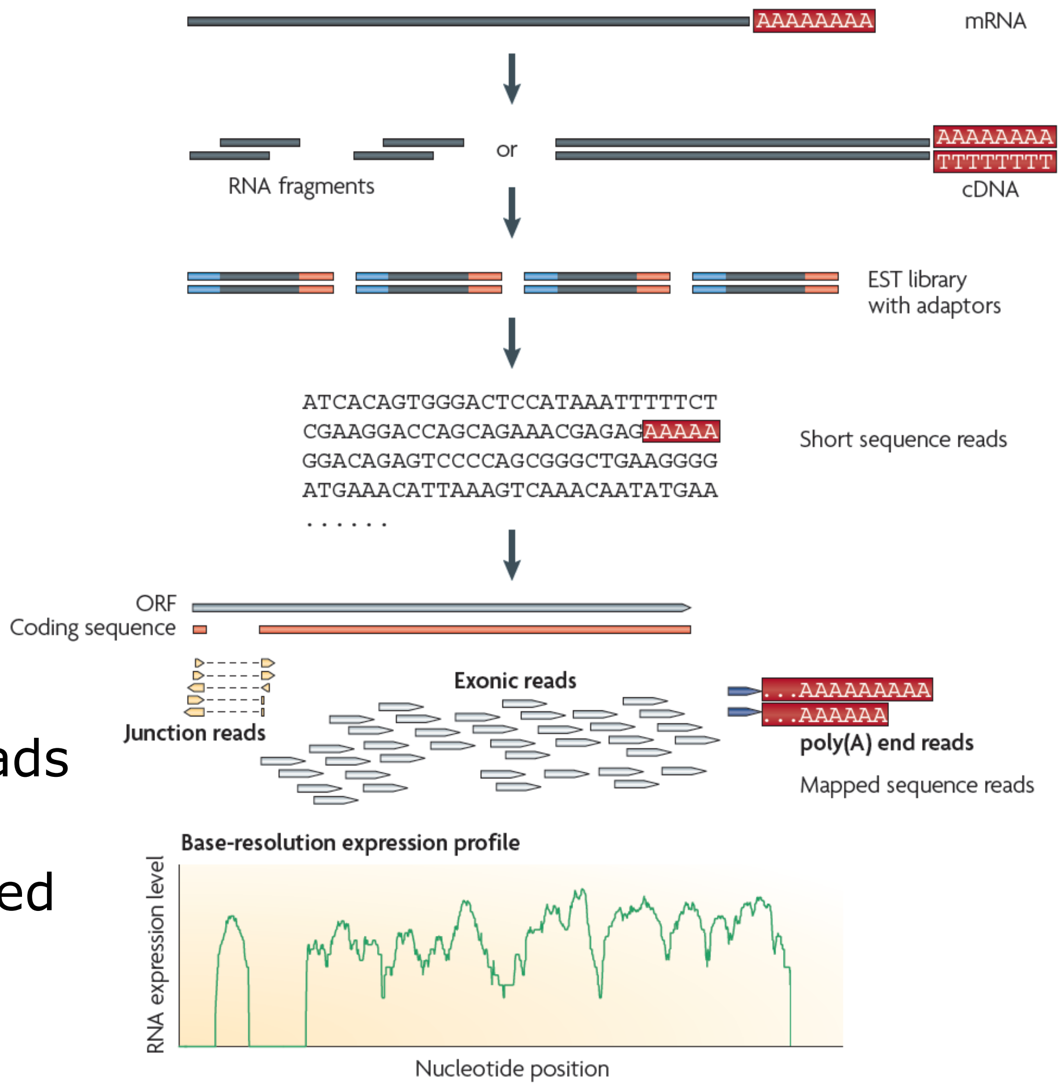
Gene expression (RNA-Seq)

- Gene expression analysis
- Transcriptomics
- Replaces microarrays

- mRNAs
- Small RNAs (miRNA, piRNA…)

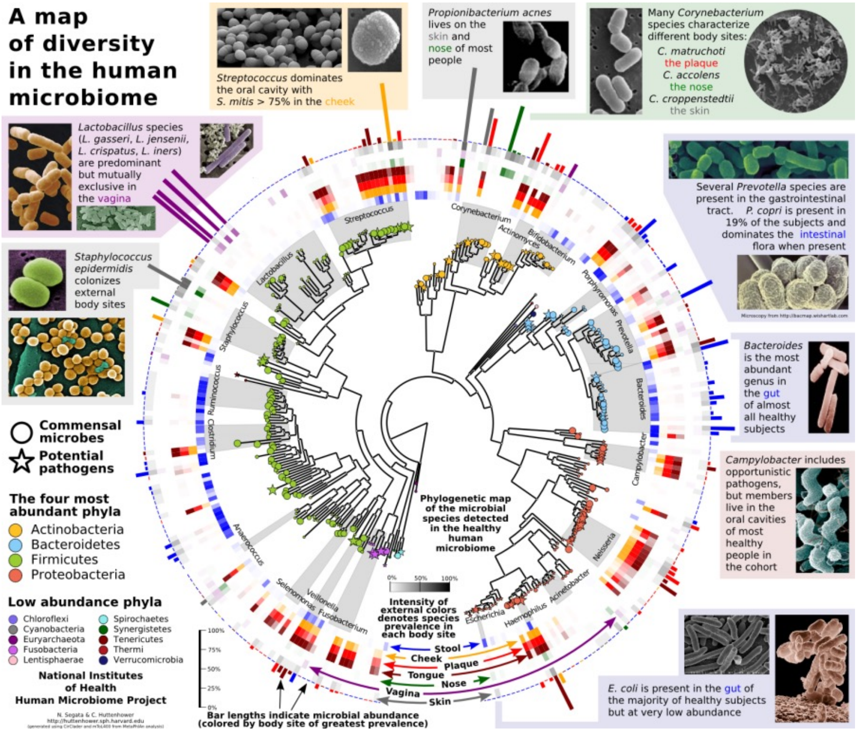
- Splice variants

- Counts the number of reads for each RNA
- Statistical analysis required for interpretation



Metagenomics/metatranscriptomics

- Samples contains collection of DNA/RNA from many microorganisms present in some niche - a microbial community
- Sequences all the DNA at once
- Sources: Soil, ocean, mine, human body, the built environment, ...
- Ecological diversity studies
- Clinical studies (e.g. human gut)
- Big data: Many hundred million sequences



A map of diversity in the human microbiome

Streptococcus dominates the oral cavity with *S. mitis* > 75% in the cheek

Propionibacterium acnes lives on the skin and nose of most people

Many *Corynebacterium* species characterize different body sites: *C. matruchoti* the plaque *C. accolens* the nose *C. croppenstedtii* the skin

Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina

Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the intestinal flora when present

Microscopy from http://bacmap.wishartlab.com

Staphylococcus epidermidis colonizes external body sites

Bacteroides is the most abundant genus in the gut of almost all healthy subjects

Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort

E. coli is present in the gut of the majority of healthy subjects but at very low abundance

Phylogenetic map of the microbial species detected in the healthy human microbiome

Intensity of external colors denotes species prevalence in each body site

Bar lengths indicate microbial abundance (colored by body site of greatest prevalence)

National Institutes of Health Human Microbiome Project

N. Segata & C. Huttenhower
http://huttenhower.sph.harvard.edu
(generated using CirCladder and mToL400 from MetaPhlAn analysis)

TARA
OCEANS

earth
microbiomeproject

HMP

Human Microbiome Project

Challenges

- Cost of actual sequencing is decreasing, but what about the cost of analysis?
- Lack of competent people for bioinformatics analysis
- Large storage needs due to the amounts of data generated. Terabytes of data.
- Compute intensive analysis (read mapping, assembly, etc)
- Security and privacy issues related to sensitive human data