Sequencing and applications

IN-BIOS5000/9000
Sequencing technologies, data analysis, and applications
28 October 2024

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

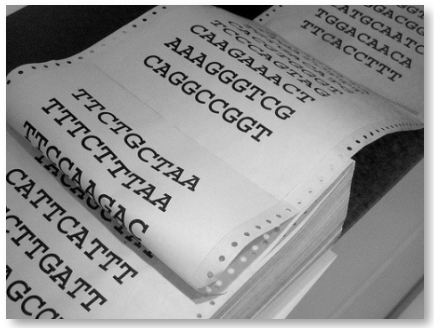
DNA sequencing

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

Obtain sample

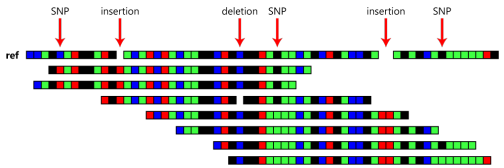
DNA extraction, preparation

Sequencing



Raw data

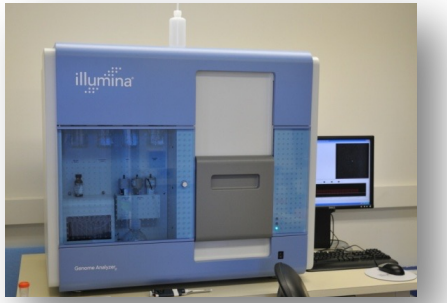
Sequence analysis



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

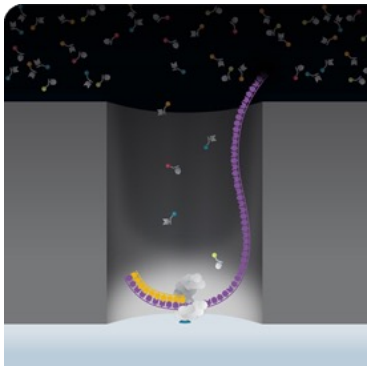
Sanger sequencing centre

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 up to 10-25 000 bases
  - Throughput up to 90 Gbp per run, 24h, Q33 (99,95% correct)

PacBio Sequel



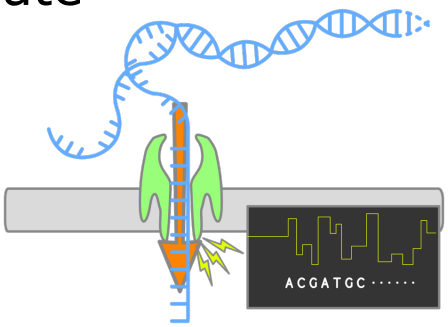
ZMW

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)

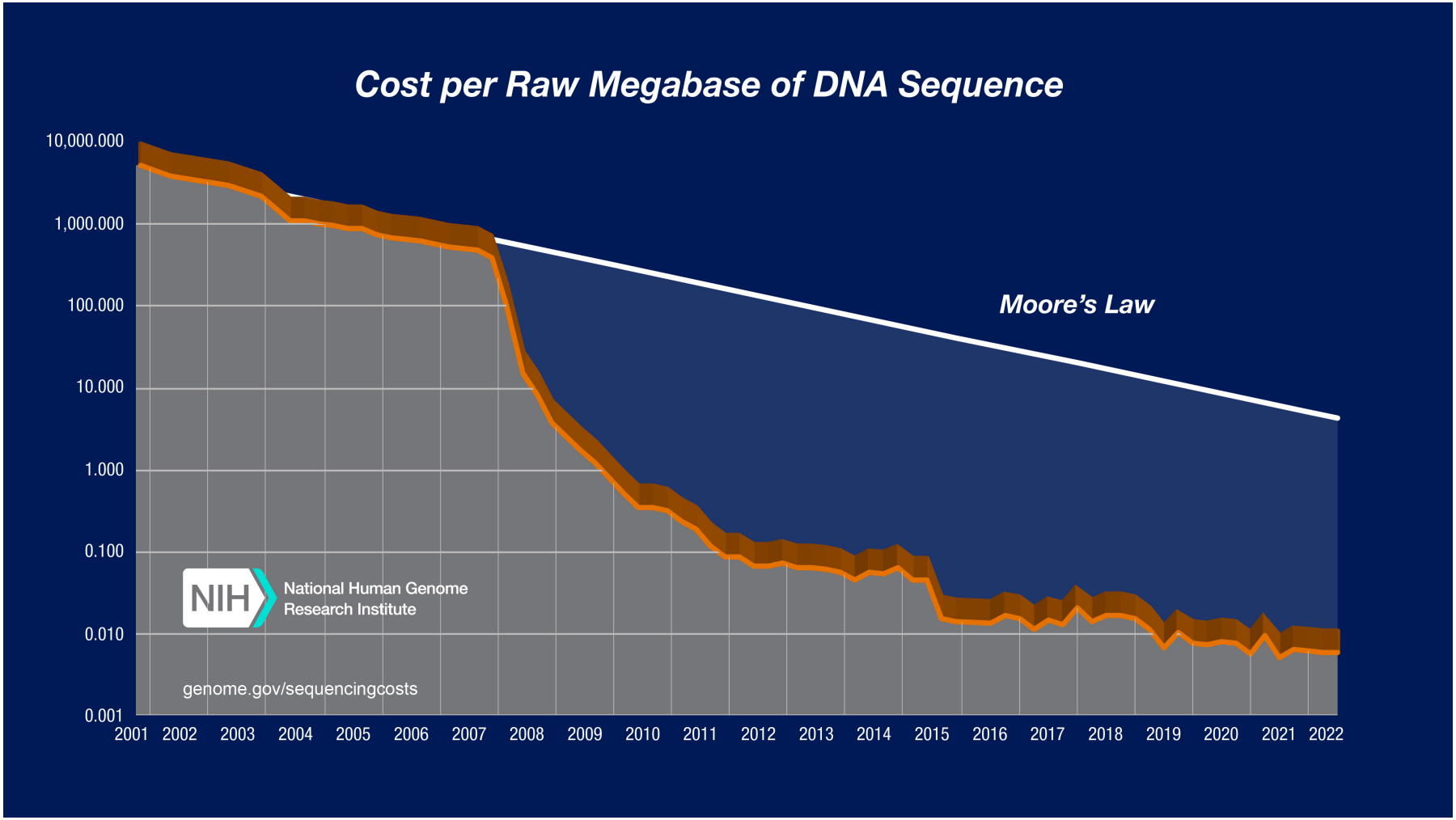
ABI (SOLiD)

Ion Torrent

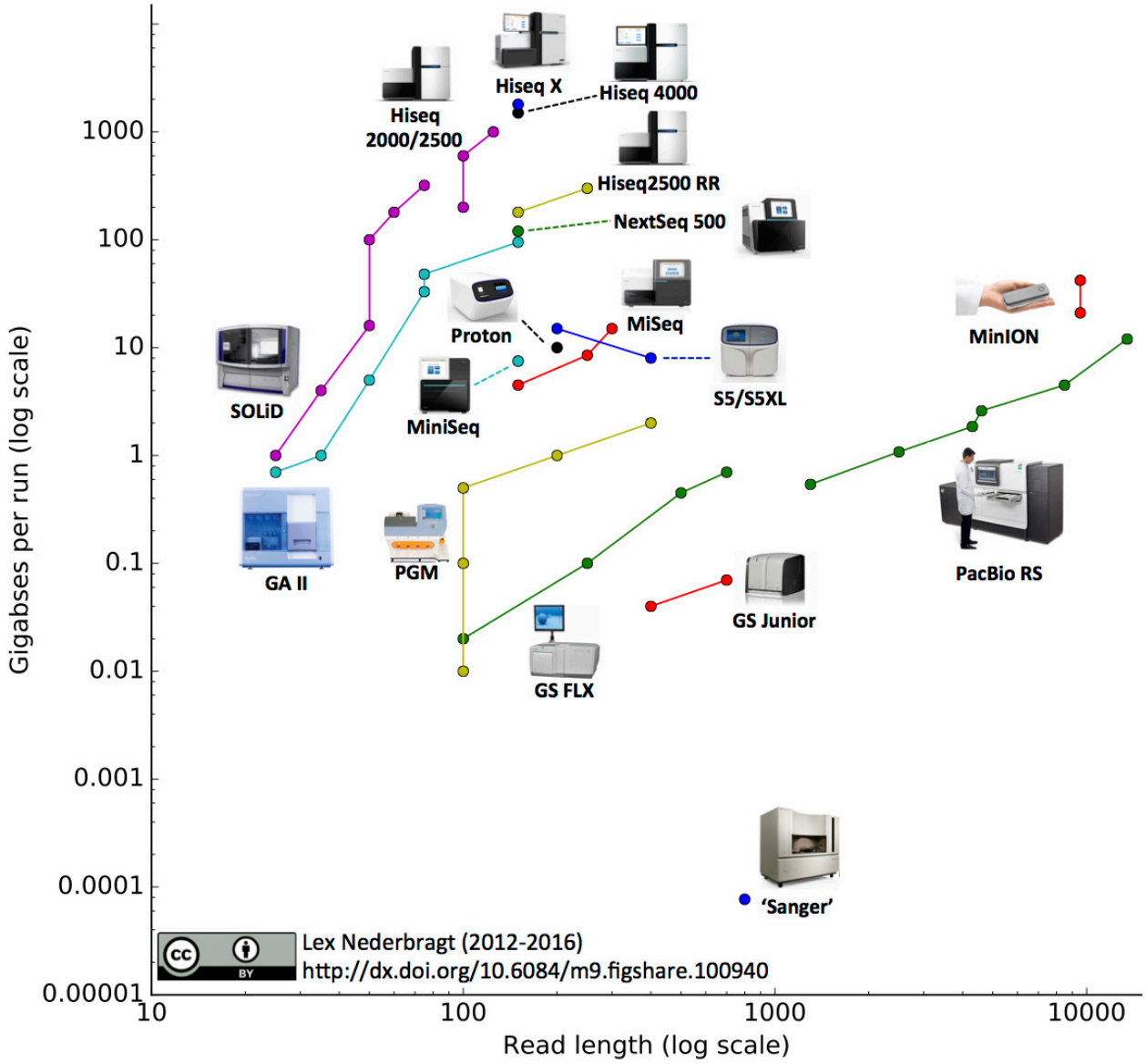
Important technology properties

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

The cost of sequencing



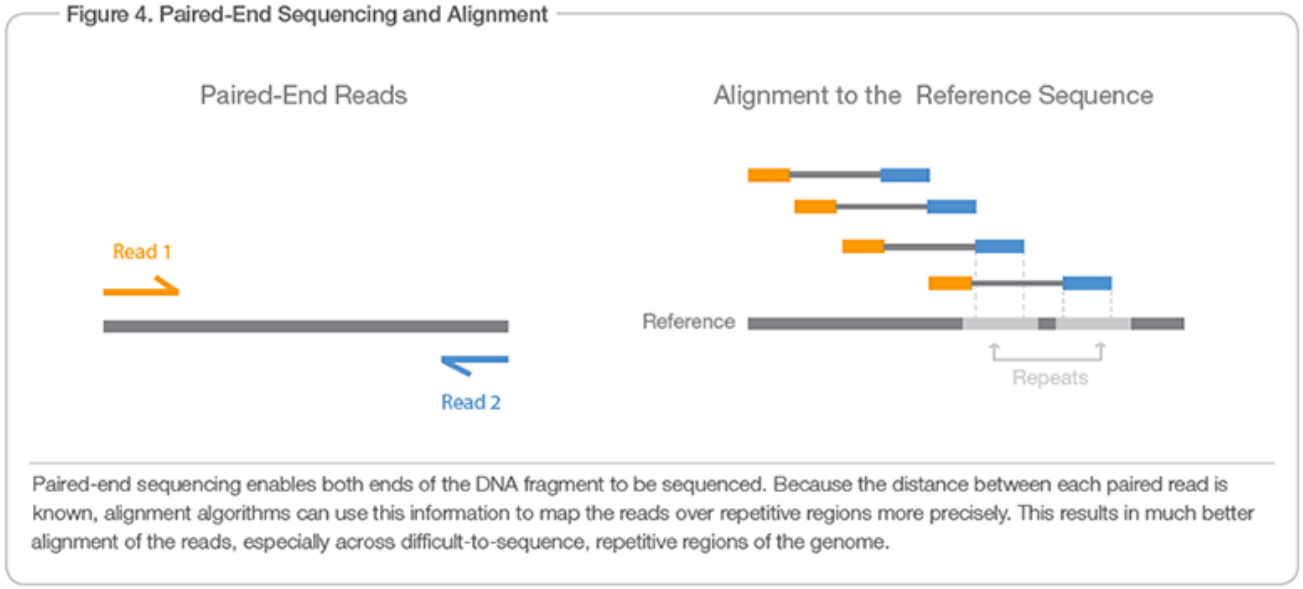
Sequencing technology development



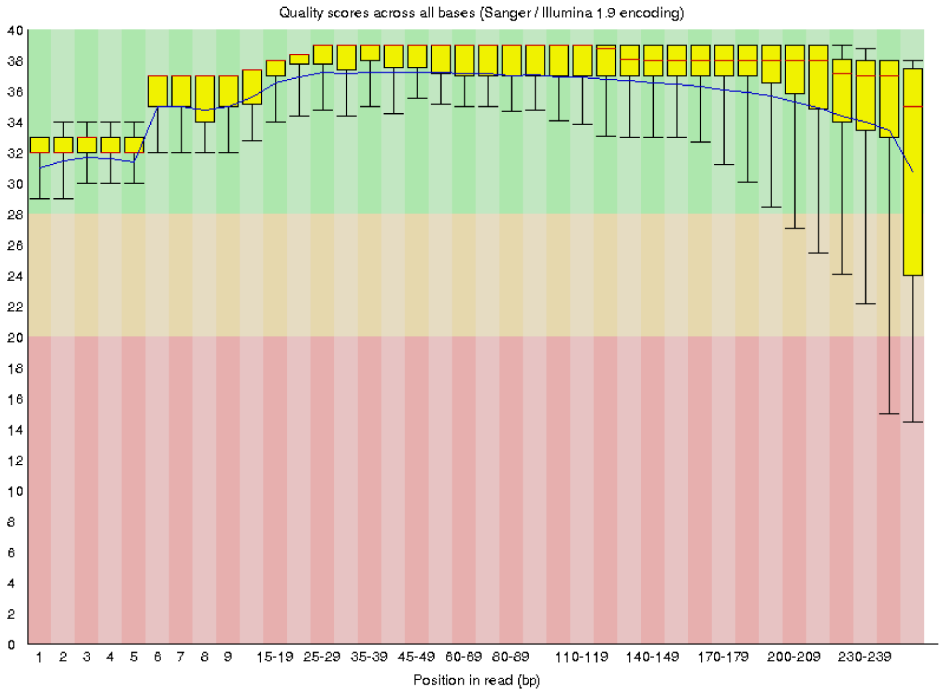
Source: Lex Nederbragt (2012-2016) https://doi.org/10.6084/m9.figshare.100940

Paired-end / mate pair sequencing

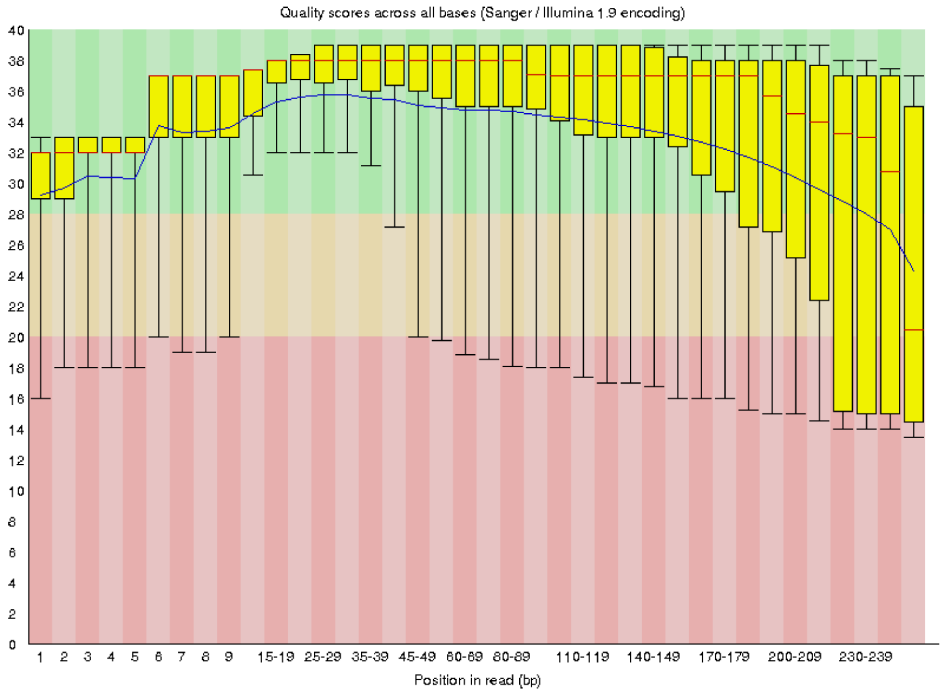
- Paired-end reads or mate pair reads are pairs of reads known to come from the same regions in the genome within a certain fixed distance
- Typically paired ends are a ~100-500bp apart, while mate pairs are ~2-10kb apart
- Performed by sequencing fragments from both ends
- Alleviates problems of short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



Quality plots of Illumina MiSeq reads



Forward reads



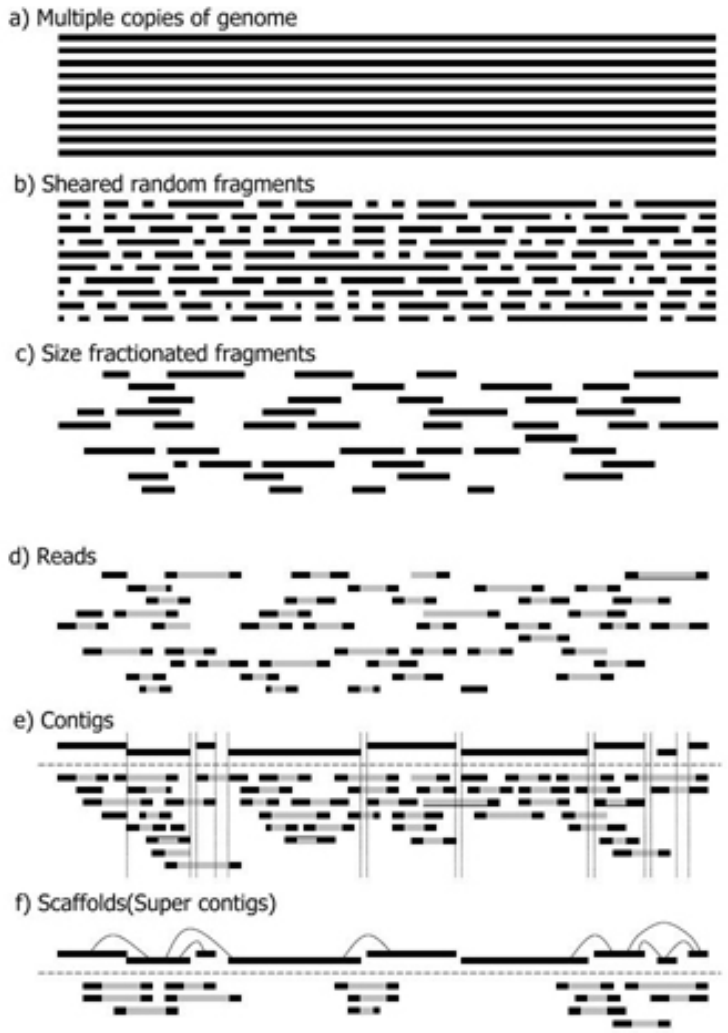
Reverse reads

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



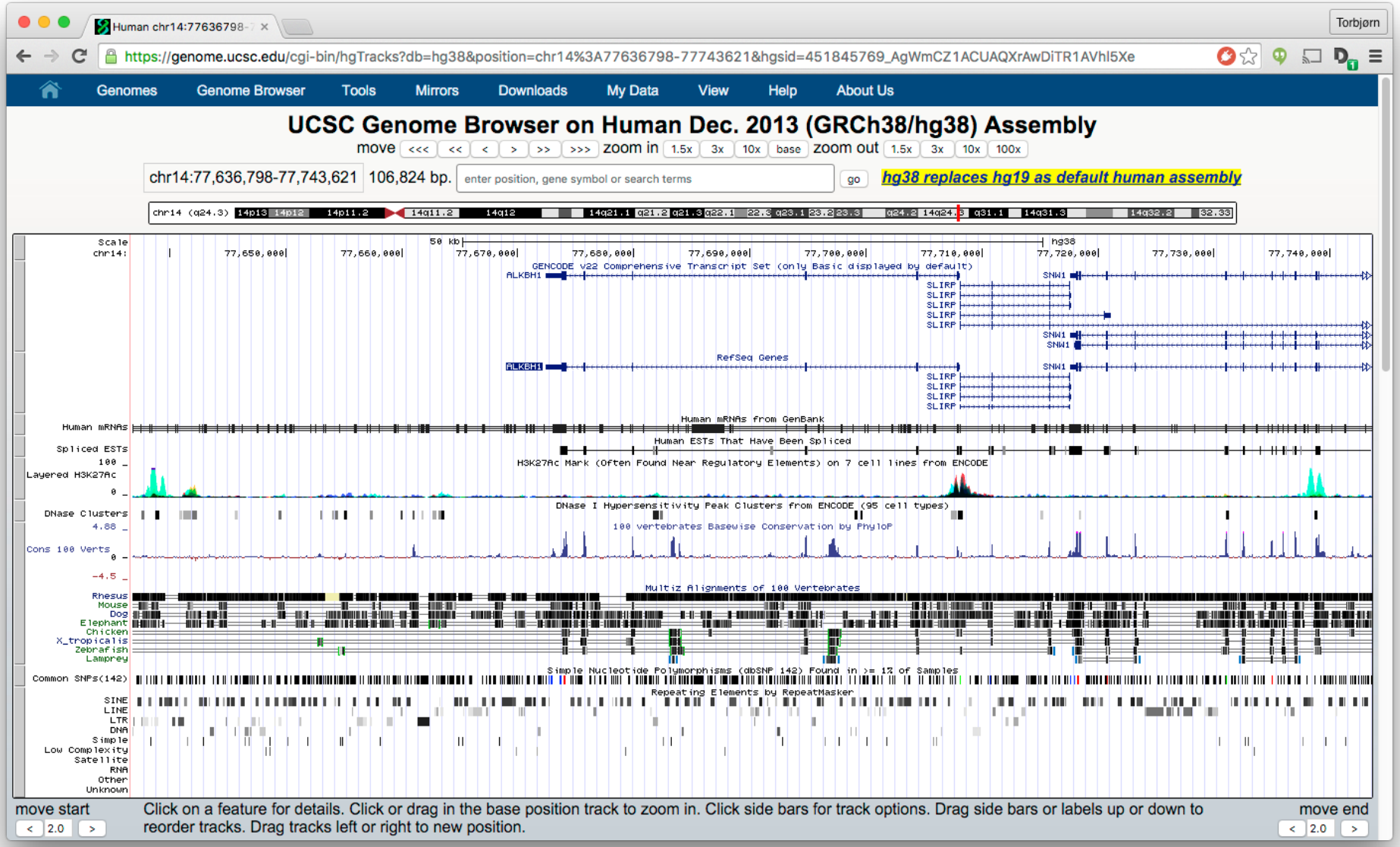
Problematic issues

- 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

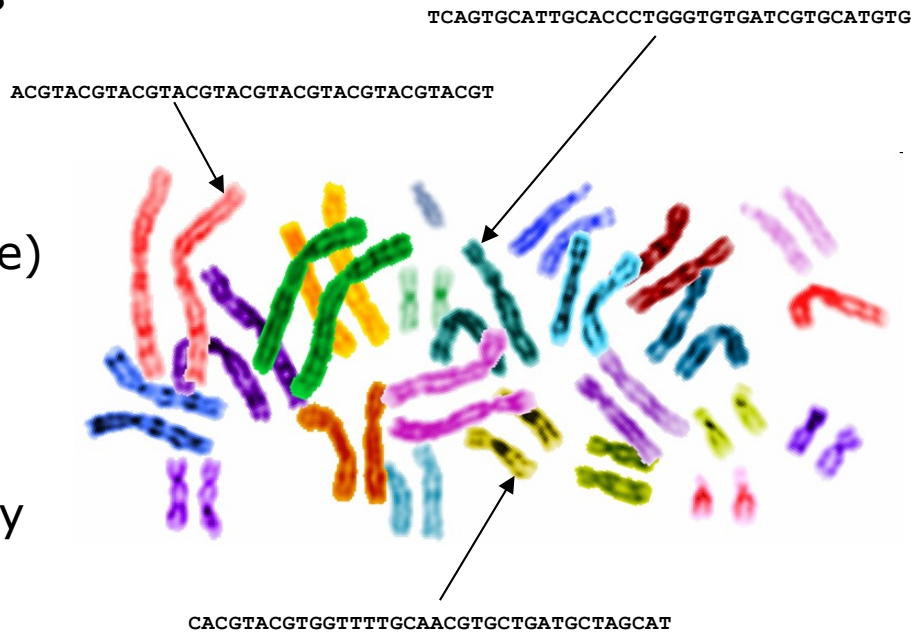
Genome browsers



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
- Mapping quality assignment

Requirements:

- High accuracy, high speed, little memory

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 30-100X)

- Natural variation discovery
- Mutation detection

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

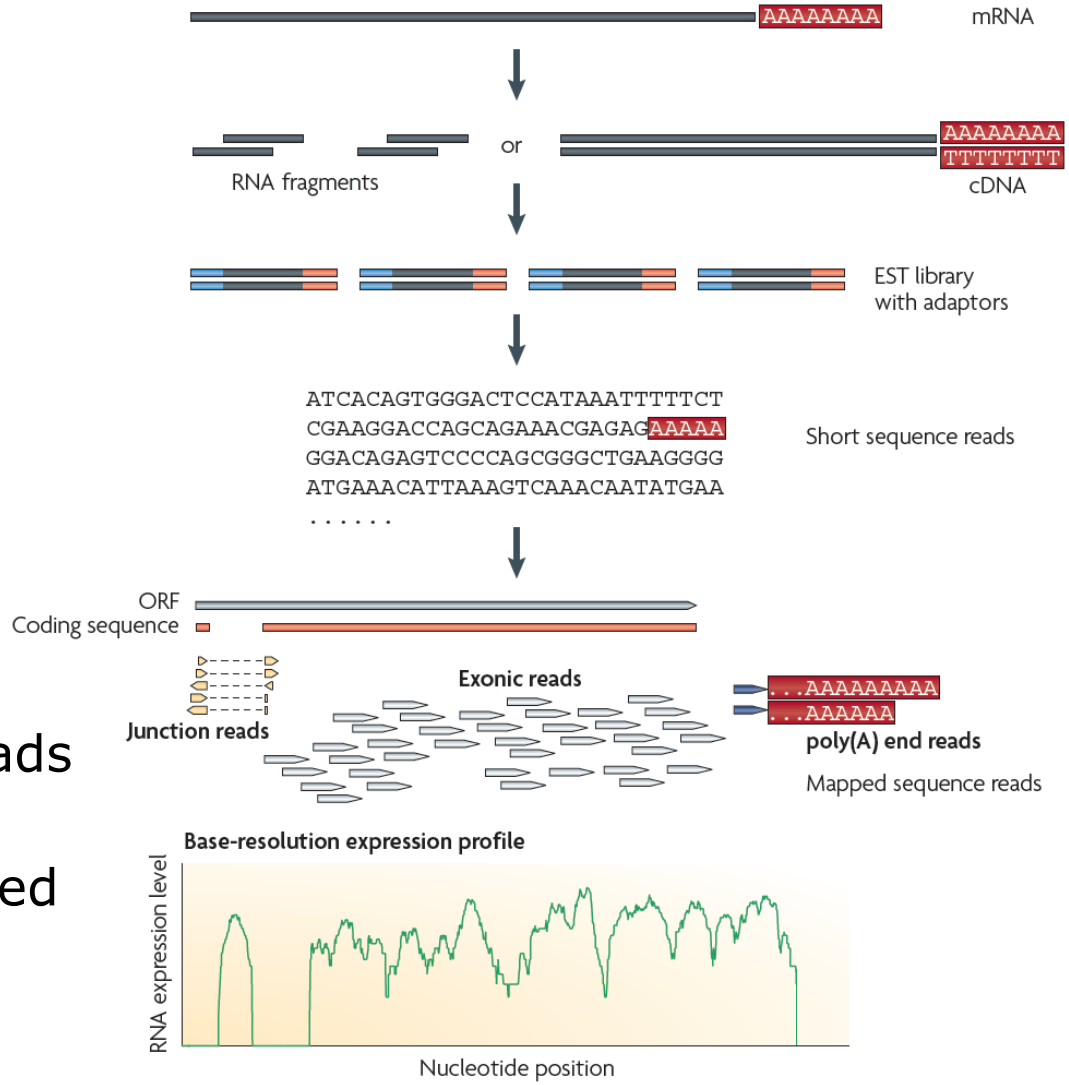
Gene expression (RNA-Seq)

- Gene expression analysis
- Transcriptomics
- Replaces microarrays

- mRNAs
- Small RNAs (miRNA, piRNA…)

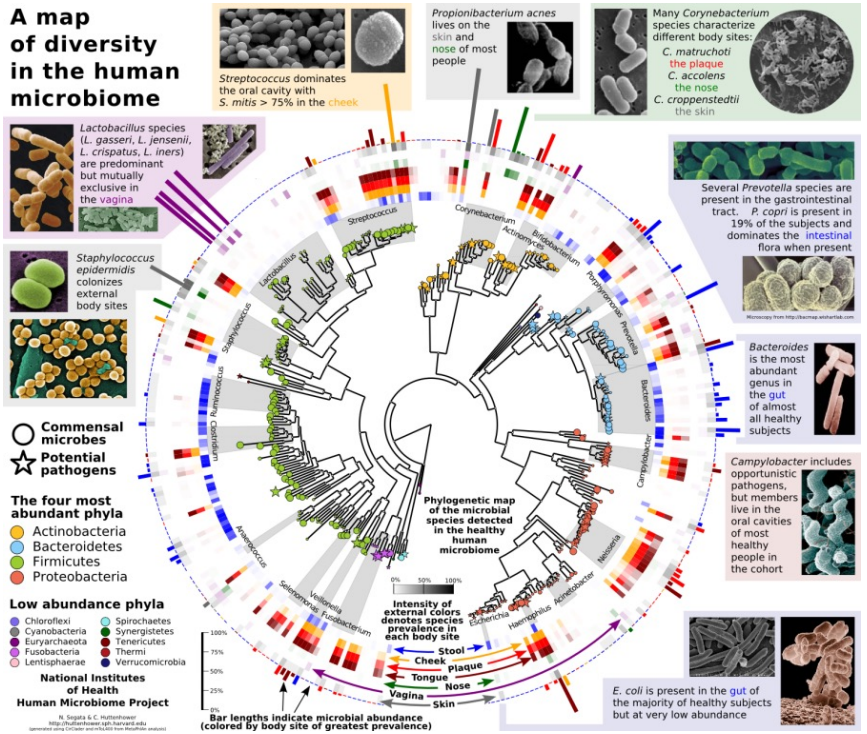
- Splice variants

- 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



TARA OCEANS

earth microbiome project

HMP

Human Microbiome Project

Challenges

- 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