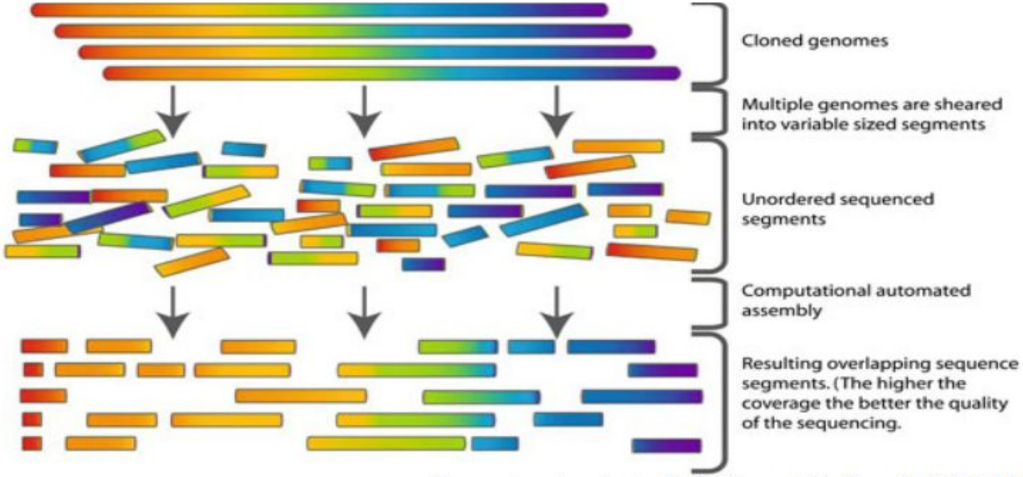
Short Read Alignment to a Reference Genome

Shamith Samarajiwa
MRC Cancer Unit
University of Cambridge

CRUK Bioinformatics Summer School 2021
22th July 2021

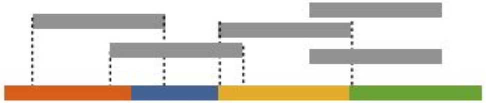
Genome Shotgun Sequencing and Assembly


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.

Commins J. et al, Biol Proced Online 11(1) 2015

Mapping to reference sequence

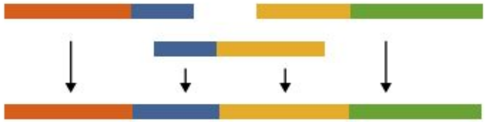
Recreate the genome with using prior knowledge as reference



Mapping is as good as reference used

De Novo assembly

Recreate the genome with no prior knowledge



Problem with repeated regions, high coverage and long

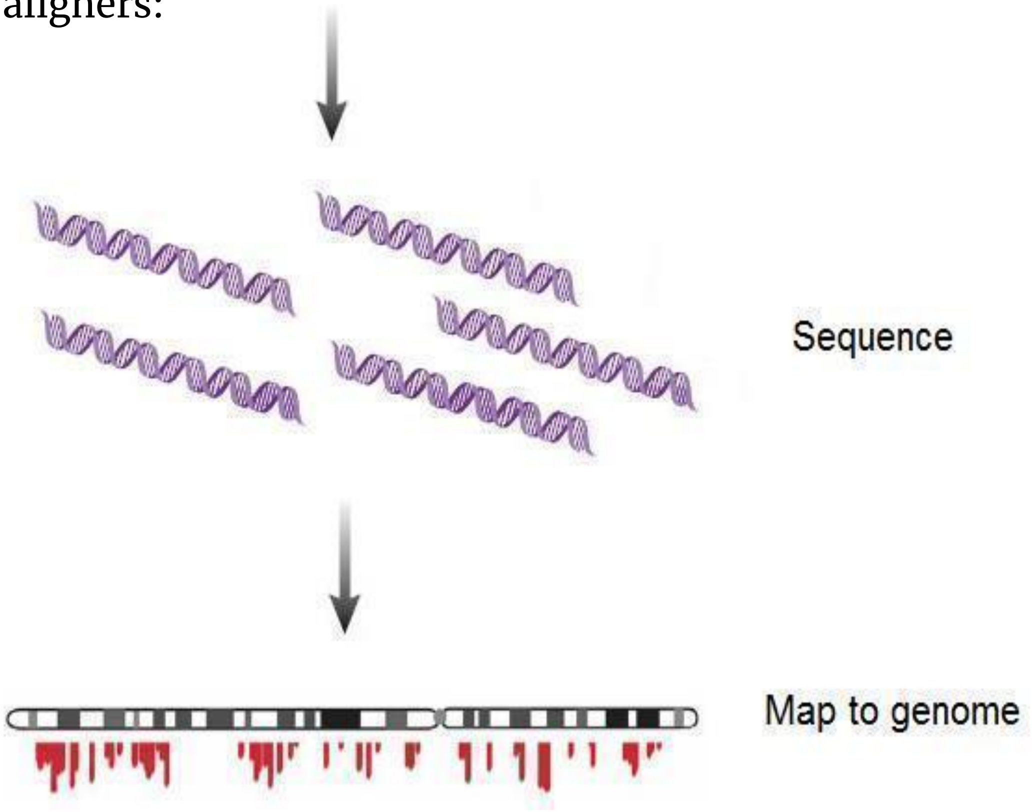
Aligning short-reads to a reference genome

A few examples of widely used short read aligners:

- BWA
- BWA-MEM2
- Bowtie2
- GEM

Splice Aware:

- **STAR**
- HISAT2
- TopHat2



(Splice Junction information from Genomic Annotation plus alignment to genome and transcriptome)

Annotations: GTF/GFF

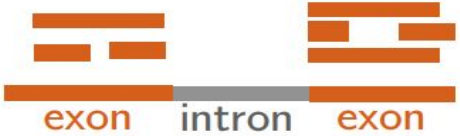
Resources:

GENCODE

RefSeq

Ensembl

GENCODE annotation is made by merging the manual gene annotation produced by the Ensembl-Havana team and the Ensembl-genebuild automated gene annotation.



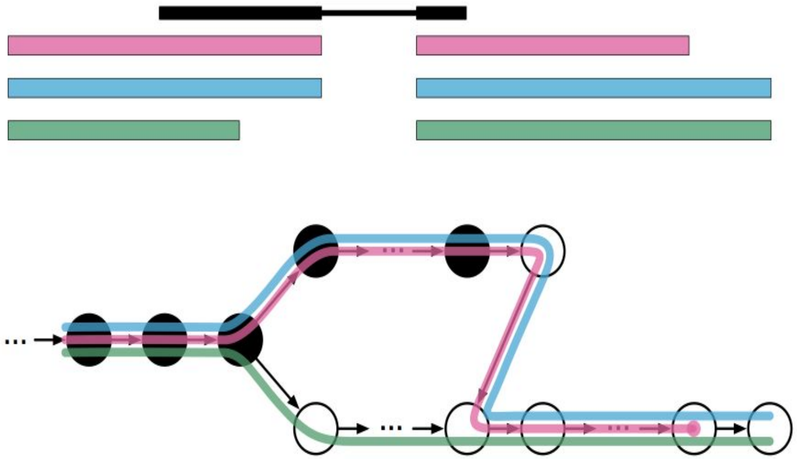
Gencode vs. Ensembl

- The gene annotation is the same in both files. The only exception is that the genes which are common to the human chromosome X and Y PAR regions can be found twice in the GENCODE GTF, while they are shown only for chromosome X in the Ensembl file.
- GENCODE GTF contains also APPRIS tags and the annotation are on the reference chromosomes only

Always make sure that annotations match the genome FASTA file (the same version & source)

Pseudo Aligners

- Used for RNA-seq quantification at a transcript level
    - Kallisto *(Bray et al., Nat. Biotech. 2016)*
    - Salmon *(Patro et al., Nat. Methods 2017)*
    - Sailfish
- Quantification estimates rather than base-to-base alignment
- Can model sequencing bias, eg. GC-bias, fragment length
- Fast, can handle multi-mapping
- Improved accuracy at transcript level

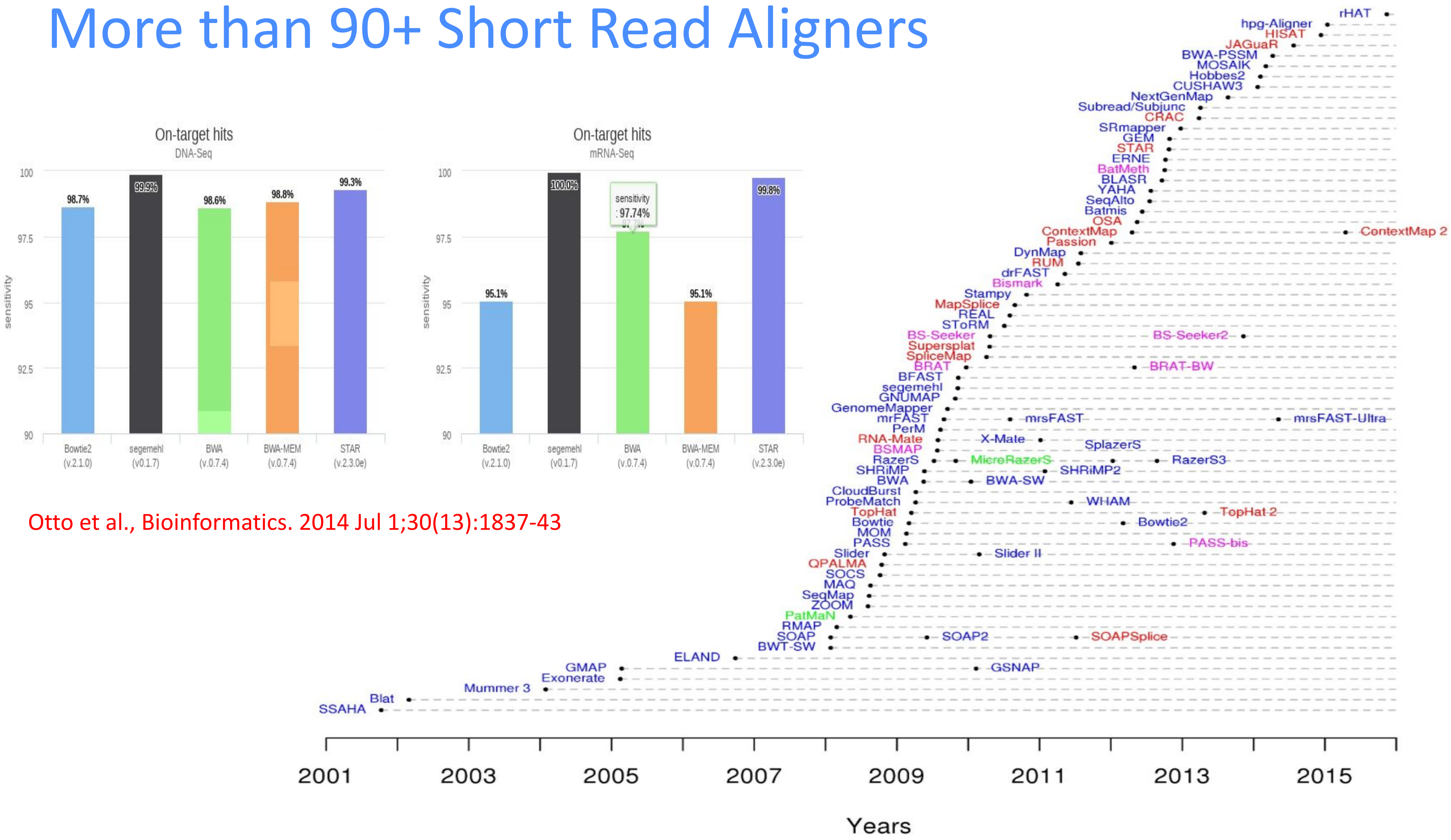


Evaluation and comparison of computational tools for RNA-seq isoform quantification

Chi Zhang, Baohong Zhang, Lih-Ling Lin & Shanrong Zhao

BMC Genomics **18**, Article number: 583 (2017) | Cite this article

More than 90+ Short Read Aligners



Otto et al., Bioinformatics. 2014 Jul 1;30(13):1837-43

https://www.ecseq.com/support/ngs/what-is-the-best-ngs-alignment-software

Features supported by the tools

	Bowtie	Bowtie2	BWA	SOAP2	MAQ	RMAP	GSNAP	FANGS	Novoalign	mrFAST	mrsFAST
Seed mm.	Up to 3		Any	Up to 2	Any	Any					
Non-seed mm.	QS	AS	Count	Count	QS	Count	Count	Count	QS	Count	Count
Var. seed len.	> 5		Any	> 28							
Mapping qual.		Yes	Yes		Yes				Yes		
Gapped align.		Yes	Yes	PE	PE		Yes	Yes	Yes	Yes	
Colorspace	Yes		Yes		Yes				Yes		
Splicing							Yes				
SNP tolerance							Yes				
Bisulphite reads						Yes	Yes		Yes	Yes	

PE: paired-end only, mm.: mismatches, QS: base quality score, count: total count of mismatches in the read, AS: alignment score, and empty cells mean not supported.

BWA

- Burrows-Wheeler Transform (BWT) algorithm with FM-index using suffix arrays.
- BWA can map low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms:
  - BWA-backtrack (Illumina sequence reads up to 100bp)
  - BWA-SW (more sensitive when alignment gaps are frequent)
  - BWA-MEM (maximum exact matches)
- BWA-SW and BWA-MEM can map longer sequences (70bp to Mbp) and share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate.
- BWA-MEM also has better performance than BWA-backtrack for 70-100 bp Illumina reads.
- Need to prepare a genome index
- BWA-MEM2 is significantly faster and a has a smaller memory footprint than BWA-MEM

Li and Durbin, 2009, Bioinformatics

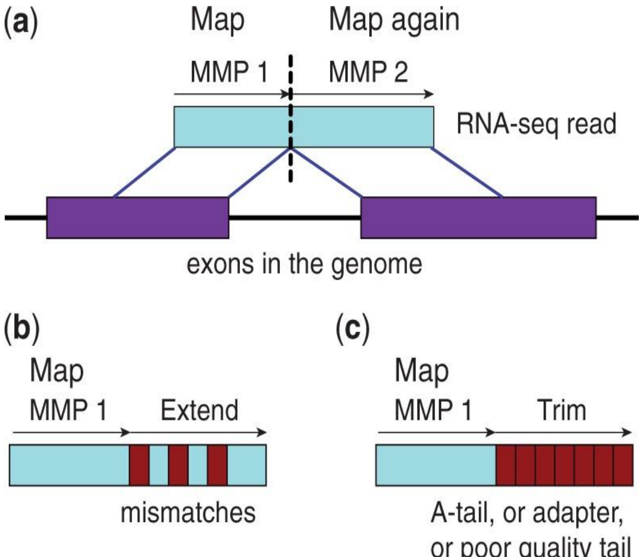
Bowtie2

- Bowtie2 is a Burrows-Wheeler Transform (BWT) aligner and handles reads longer than 50 nt.
- The transform is performed by sorting all rotations of the test and these acts as the index for the sequence. The aim is to find out from which part of the genome a the 'read' originates.
- Given a reference and a set of reads, this method reports at least one good local alignment for each read if one exists.
- Since genomes and sequencing datasets are usually large, dynamic programing proves to be inefficient and high-memory machines are required, with lots of secondary storage, etc.
- Need to prepare a genome index.

Langmead and Salzberg, 2012, Nat. Methods

STAR

- Non-contiguous nature of transcripts, presence of splice-forms make short-read (36-200 nt) RNA-seq alignment to a genome challenging.
  - Reads contain mismatches, insertions and deletions caused by genomic variation and sequencing errors.
  - Mapping spliced sequence from non contiguous genomic regions.
  - Multi-mapping reads
- Two steps: Seed searching and clustering/stitching/scoring (find MMP -maximal mappable prefix using Suffix Arrays)
- Fast splice aware aligner, high memory (RAM) footprint
- Can detect chimeric transcripts
- Generate indices using a reference genome fasta, and annotation gtf or gff from Ensembl/UCSC.

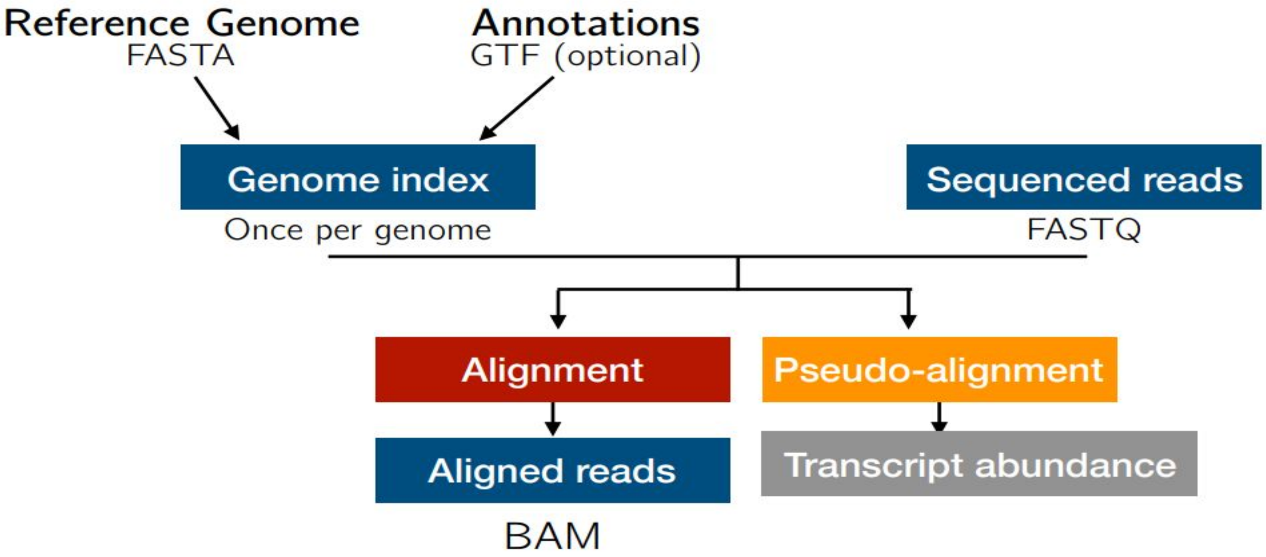


Dobin et al., 2013 Bioinformatics

Before you align checklist

- Do I need splice-aware aligner?
- Am I using right genome version? (hg38 - human, mm10 -mouse?)
- Do annotations match the reference genome?
- Read manual, select parameters, check default settings

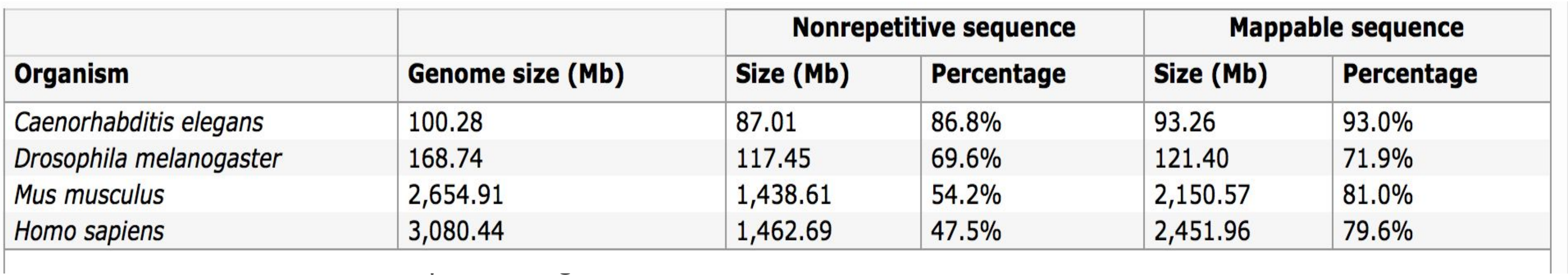
Standard alignment workflow



Some useful concepts in short read alignment

- Alignment Coverage and Depth
- Mappability
  - Alignability
  - Uniqueness
- Read Count Normalization
- File format specific tools: SAM/BAM files
  - SAM tools
  - Picard tools
- Mapping QC
  - SAMStat
- Visualization
  - IGV
- Downloading sequence data from repositories
  - SRA toolkit

Mappability



		Nonrepetitive sequence		Mappable sequence	
Organism	**Genome size (Mb)**	**Size (Mb)**	**Percentage**	**Size (Mb)**	**Percentage**
Caenorhabditis elegans	100.28	87.01	86.8%	93.26	93.0%
Drosophila melanogaster	168.74	117.45	69.6%	121.40	71.9%
Mus musculus	2,654.91	1,438.61	54.2%	2,150.57	81.0%
Homo sapiens	3,080.44	1,462.69	47.5%	2,451.96	79.6%

Rozowsky, (2009)

- Not all of the genome is ‘available’ for mapping when reads are aligned to the unmasked genome.

- **Alignability:** This provide a measure of how often the sequence found at the particular location will align within the whole genome.

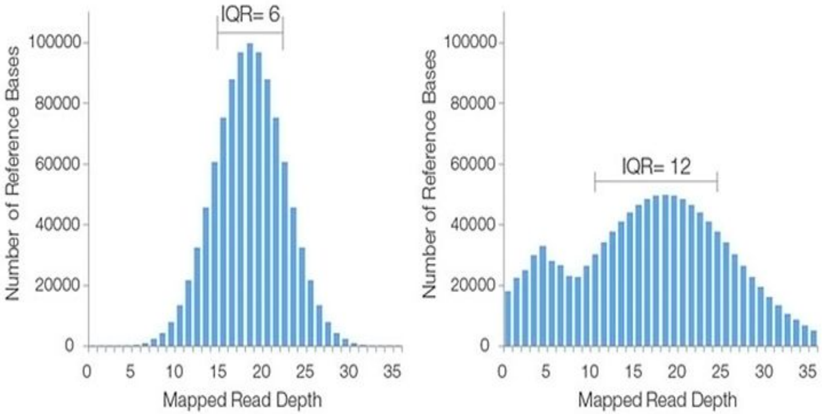
- **Uniqueness:** This is a direct measure of sequence uniqueness throughout the reference genome.

Coverage and Depth

Coverage: The average number of reads of a given length that align to or ‘cover’ known reference bases with the assumption that the reads are randomly distributed across the genome.

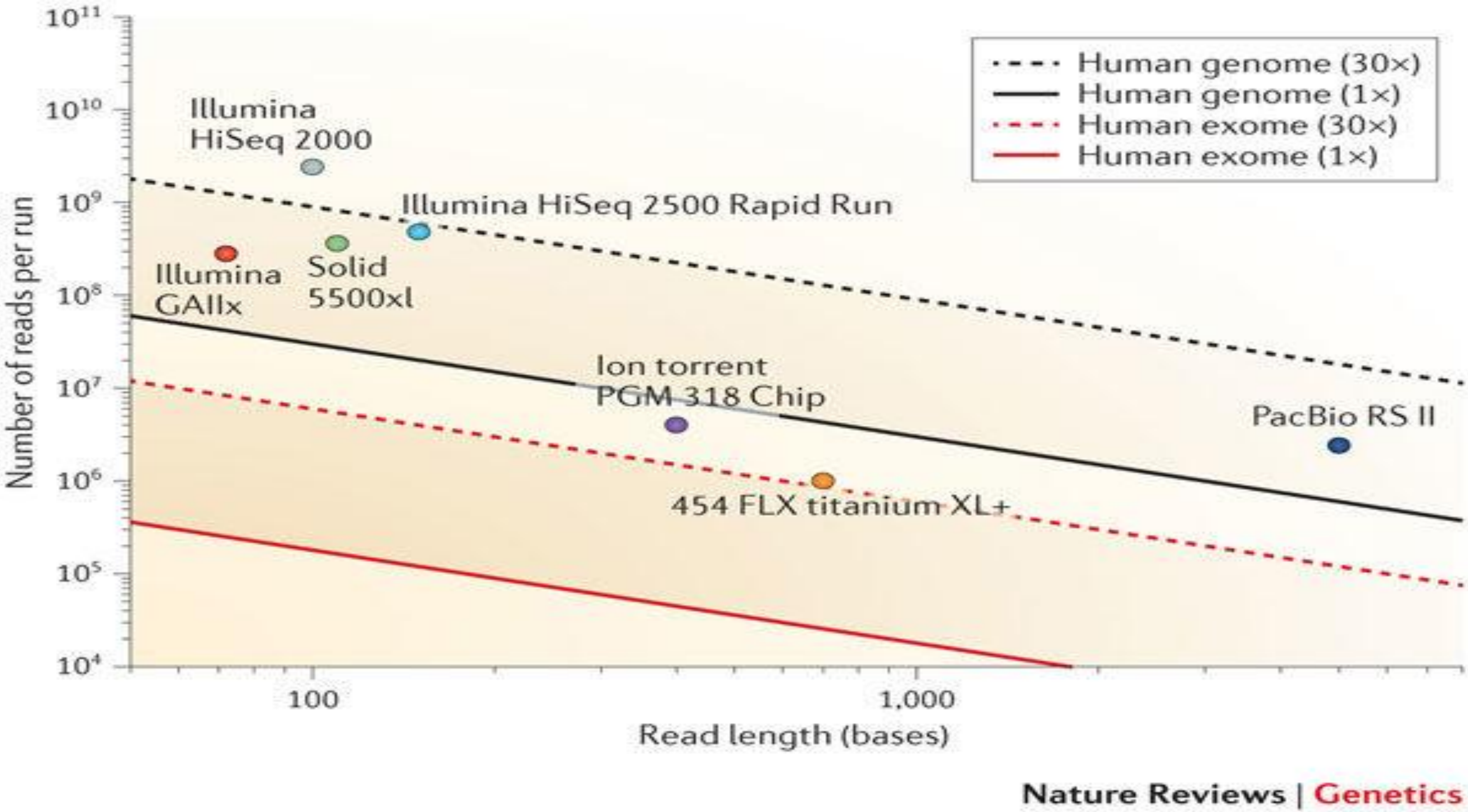
Depth: redundancy of coverage or the total number of bases sequenced and aligned at a given reference position. Increased depth of coverage rescues inadequacies of sequencing methods.

Examples of good (left) and poor (right) sequencing coverage histograms



Sims et al., 2014, Nat. Rev. Genet.

Lander-Waterman model of Coverage



Nature Reviews | Genetics

Coverage = Read Length * Number of reads / haploid Genome length

Normalised Counts

- Do not use RPKM (Reads Per Kilobase Million) and FPKM (Fragments Per Kilobase Million) to express normalised counts in ChIP-seq (or RNA-seq).
- CPM (Counts Per Million) and TPM (Transcripts Per MIllion) is the less biased way of normalising read counts.
- When calculating TPM, the only difference from RPKM is that you normalize for gene/transcript length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound.

RPKM vs TPM

Lior Pachtor video

Processing SAM / BAM files

- **SAMtools** provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.
    - **import**: SAM-to-BAM conversion
    - **view**: BAM-to-SAM conversion and sub alignment retrieval
    - **sort**: sorting alignment
    - **merge**: merging multiple sorted alignments
    - **index**: indexing sorted alignment
    - **faidx**: FASTA indexing and subsequence retrieval
    - **tview**: text alignment viewer
    - **pileup**: generating position-based output and consensus/indel calling
- **RSamTools** package in *Bioconductor* allows similar functionality in R.

Picard tools

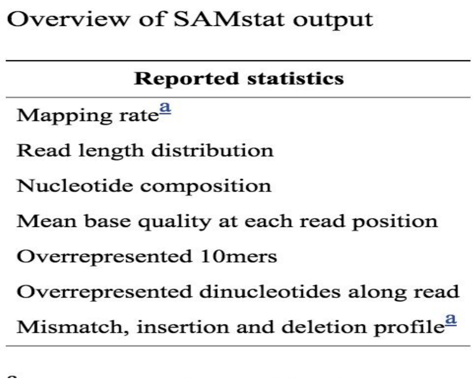
- Picard is a collection of Java-based command-line utilities that manipulate sequencing data and formats such as SAM/BAM/CRAM and VCF. It has a Java API (SAM-JDK) for creating new programs that read and write SAM files.

- The *mark duplicate* function is particularly useful.

Picard tools

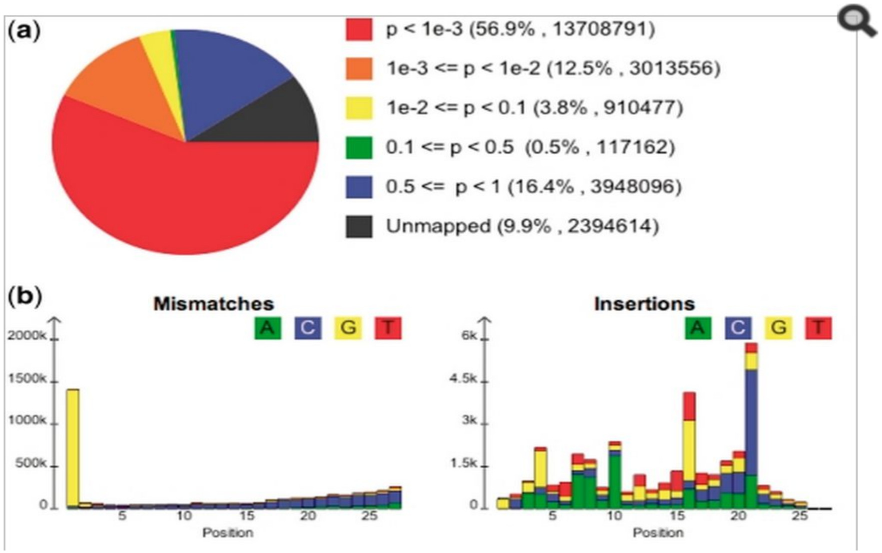
SAMStat for mapping QC

- SAMstat is a C program that plots nucleotide overrepresentation and other statistics in mapped and unmapped reads and helps understand the relationship between potential protocol biases and poor mapping.
- It reports statistics for unmapped, poorly and accurately *mapped reads* separately.
- This allows for identification of a variety of problems, such as remaining linker and adaptor sequences that cause poor mapping.



Overview of SAMstat output

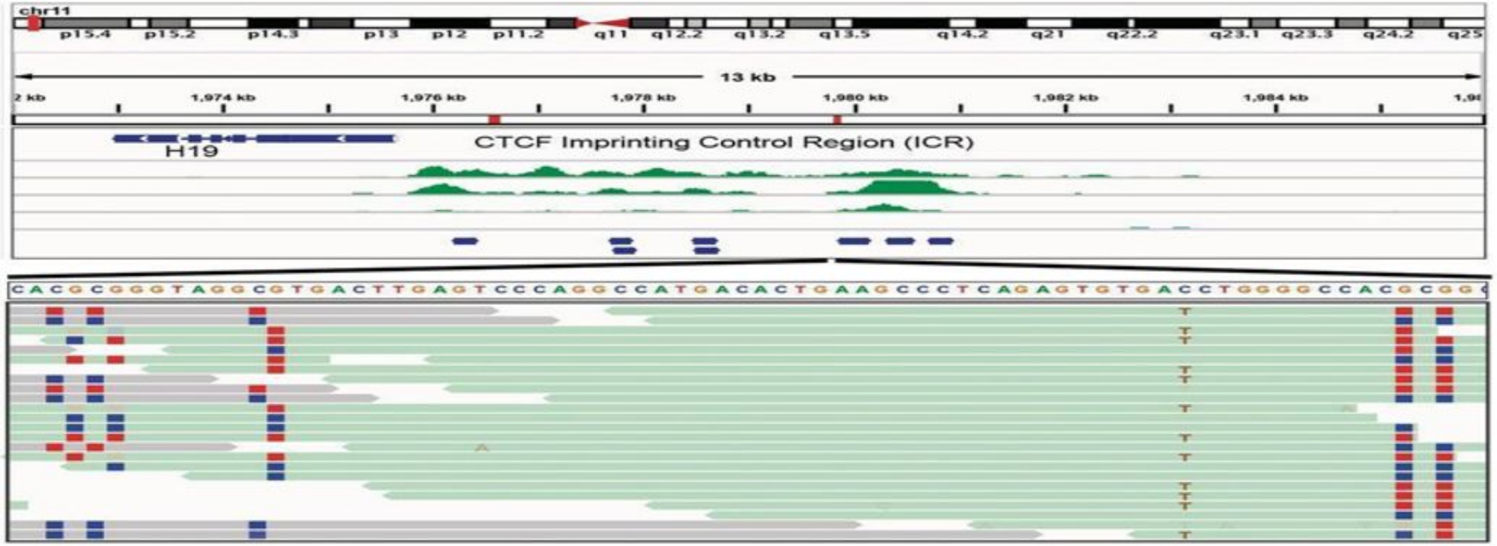
Reported statistics
Mapping rate[a]
Read length distribution
Nucleotide composition
Mean base quality at each read position
Overrepresented 10mers
Overrepresented dinucleotides along read
Mismatch, insertion and deletion profile[a]

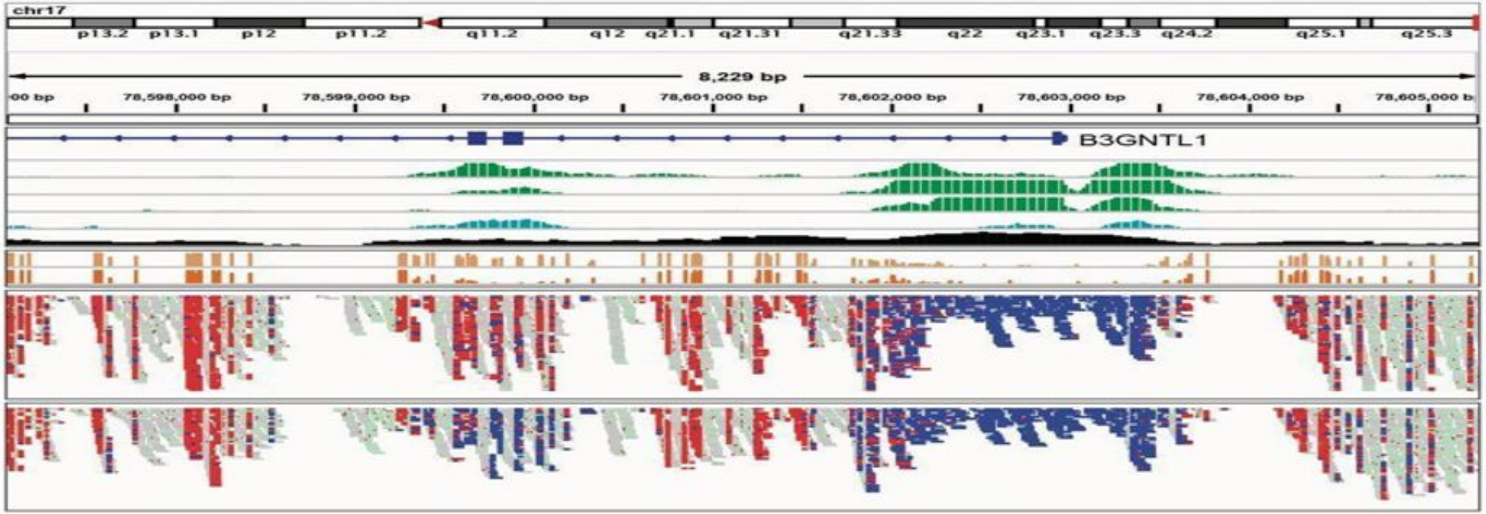


Lassmann et al., 2011, Bioinformatics.

Visualization with IGV

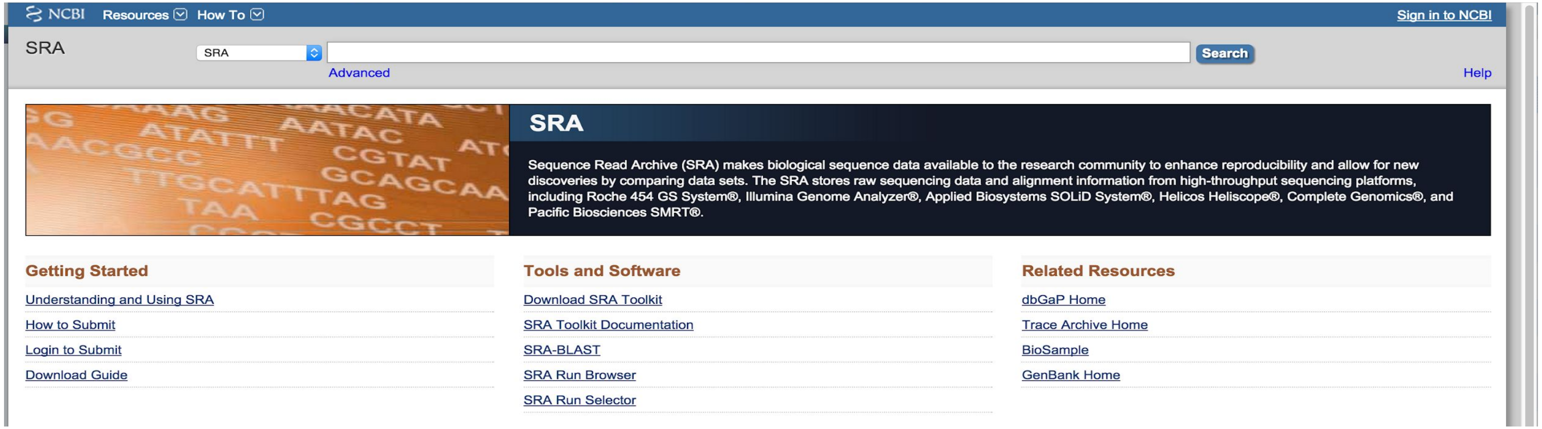
Integrated Genome Viewer (IGV)





How to get external sequencing data via SRA toolkit

- Extract data sets from the **Sequence Read Archive** or **dbGAP** (NCBI)
- These repositories store sequencing data in the SRA format
- Prefetch: fetch fastq data
- Fastq-dump: Convert SRA data into fastq format
- sam-dump: Convert SRA data to SAM format
- sra-stat: Generate statistics about SRA data (quality distribution, etc.)
- vdb-validate: Validate the integrity of downloaded SRA data



SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

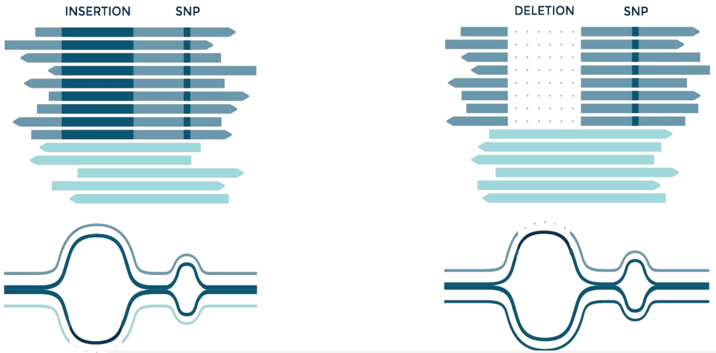
Getting Started
- Understanding and Using SRA
- How to Submit
- Login to Submit
- Download Guide

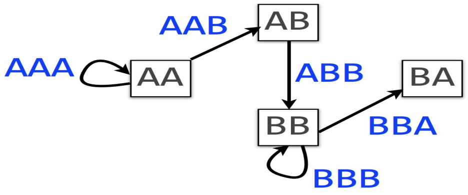
Tools and Software
- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources
- dbGaP Home
- Trace Archive Home
- BioSample
- GenBank Home

The Future

- Graph based reference genomes and aligners are beginning to make an appearance and will eventually replace linear genome representations.
- Long read sequencing technologies are becoming more robust (Oxford Nanopore Technologies, Pacific Bioscience, Illumina and others)
- *De novo* assembly of genomes (usually using De Bruijn graph methods for species without reference genomes) is an alternative to mapping.


INSERTION
SNP
DELETION
SNP


AAB
AB
AAA
AA
ABB
BA
BB
BBA
BBB