

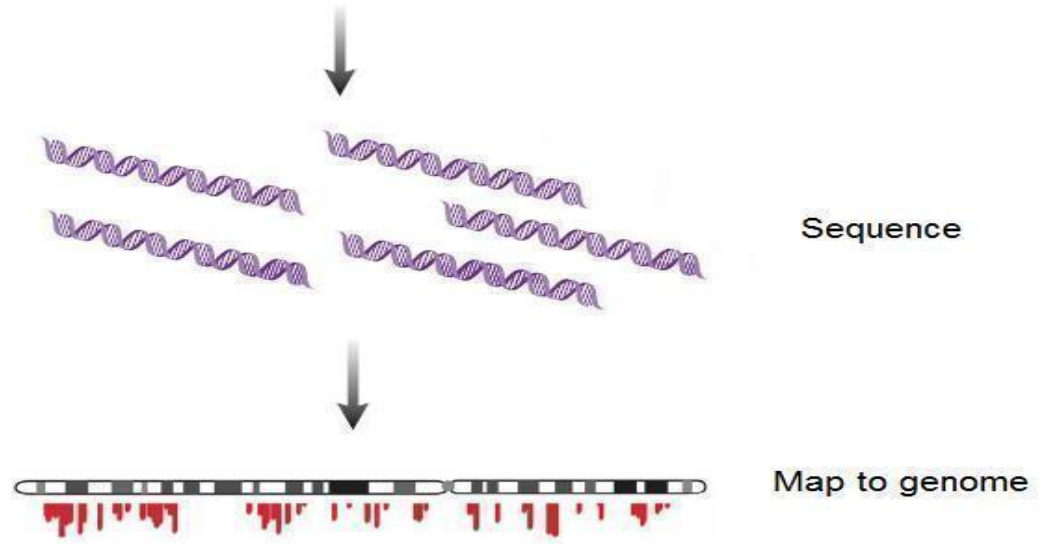
Short Read Alignment to a Reference Genome

Shamith Samarajiwa

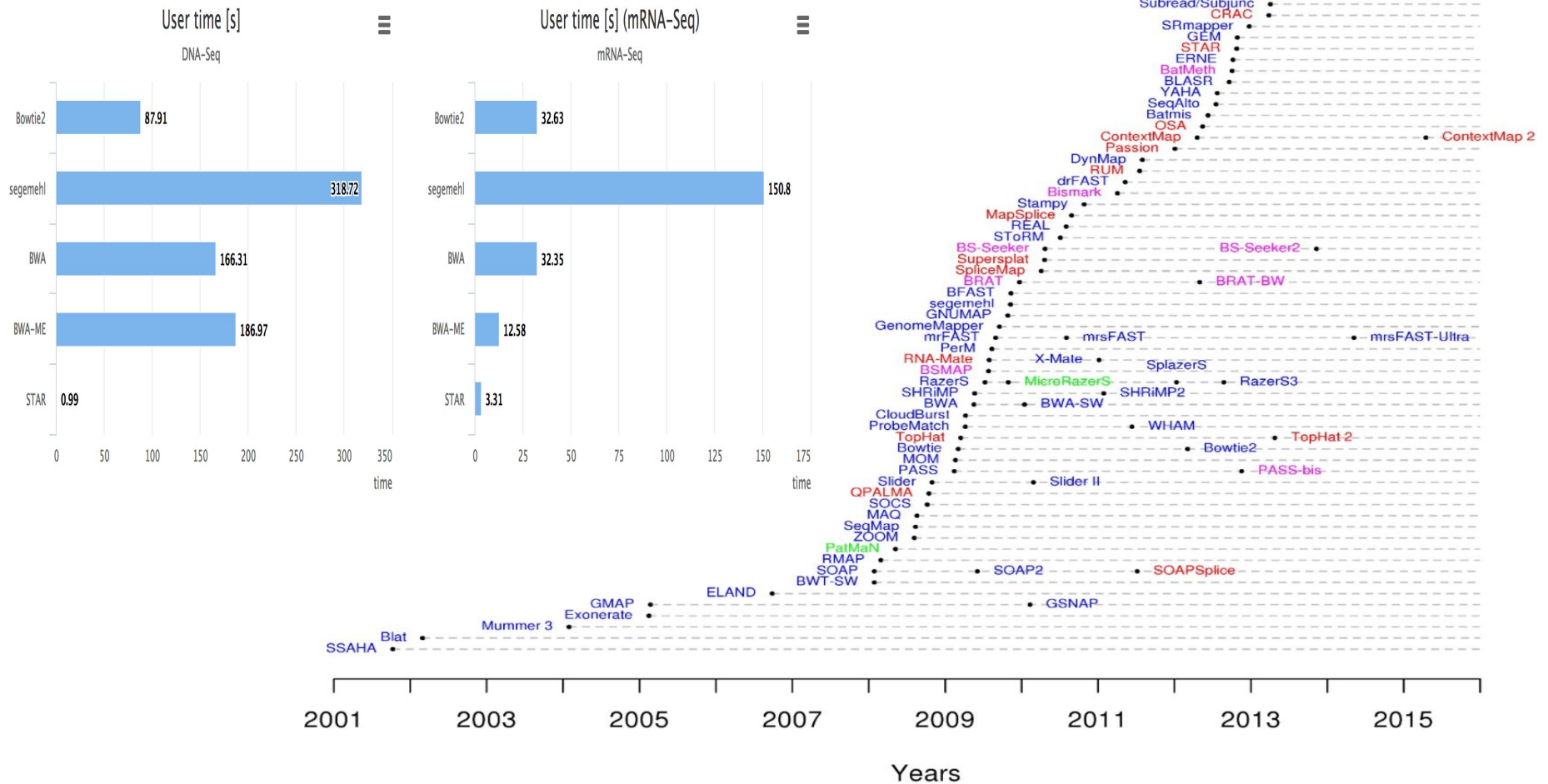
CRUK Summer School in Bioinformatics
Cambridge, September 2018

Aligning to a reference genome

- BWA
- Bowtie2
- STAR
- GEM
- Pseudo Aligners for RNA-seq quantification
 - Kallisto
 - Salmon
 - Sailfish



More than 90 Short Read Aligners



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

Table 4: Overall evaluation and comparison of multiple aligners.

Aligners	Computational speed			Overall evaluation	Memory usage		Accuracy			
	Speed with single thread	Speed with multithread	Key factor impacting speed (genome size or read count)		Key factor impacting memory (Genome size or read count)	Memory usage with multithread	Sensitivity	Precision	% of multimapped	%Corrected Multi-Mapped
Bowtie1	Fast	↑	Genome size	Low	Genome size	=	High	—	—	
BWA	Fast	↑	Both	Low	Genome size	=				
BOAT	Slow	↑↑	Genome size	Low	Read count	↑↑	High	—	—	Low
GASSST	—	↑	Genome size	High★★	Genome size	=	Low	High	—	
Gnumap	Slow	↓	Genome size	High★★	Genome size	=				
GenomeMapper	Slow	=	Genome size	Low▲	Genome size	=	High	—	—	
mrFAST	Slow	×	Genome size	High★★	Read count	×	High	—	—	
mrsFAST	—	×	Genome size	Low	Read count	×	High	—	—	
MAQ	—	×	Genome size	High★★	Read count	×				
NovoAlign [#]	—	/	Read count	Low▲	Genome size	/	High	High	Low	Low
PASS	—	↑	Genome size	Low▲	Genome size	↑	High	High	Low	Low
PerM [*]	Fast		Genome size	Low▲	Genome size	/	Ind: low	—	Low	
RazerS	Slow	×	Genome size	High★★	Read count	×	High	—	—	
RMAP	—	×	Genome size	High★	Genome size	×	Mis: low	High	Low	
SeqMap	—	×	Genome size	High★★★★	Read count	×	High	—	—	
SOAPv2	Fast	↑	Genome size	Low	Genome size	=	High	High	Low	
SHRiMAP2	Slow	↑	Genome size	High★★	Genome size	↑	High	Low	High	
Segemehl	—	↑	Both	High★★★★	Genome size	=	High	—		

PerM^{*} could adjust the threads automatically during running process.

Novoalign[#] could support multithread only for commercial version.

For computational speed, we defined the aligners which are extremely faster than others as fast, while we defined the ones which are extremely slower as slow.

For memory usage, we evaluated the aligners as follow: among the 5 even datasets, the maximum memory usage ≤4 G, low; the maximum memory usage ≥32 G, high★★★★.

Low▲ represents that the maximum memory usage will have an extreme increase with *H. sapiens* datasets (≥4 G).

×: without multithread function.

— represents medium level remark.

= means there is no obvious change.

Short Read Aligners

Download and install BWA

download

wget https://sourceforge.net/projects/bio-bwa/files/bwa-0.7.17.tar.bz2

extract

x extracts, v is verbose (details of what it is doing), f skips prompting for each individual file, and j tells it to unzip .bz2 files

tar -xvfj bwa-0.7.17.tar.bz2

cd bwa-0.7.17

make

Add BWA to your PATH by editing ~/.bashrc file (or .bash_profile or .profile file)

export PATH=\$PATH:/path/to/bwa-0.7.17

source ~/.bashrc

/path/to/ is a placeholder. Replace with real path to BWA on your machine

manual

man ./bwa.1

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 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 70-100bp Illumina reads.
- Need to prepare a [genome index](#)

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 text 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.

Features supported by the tools

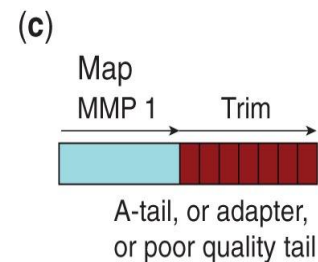
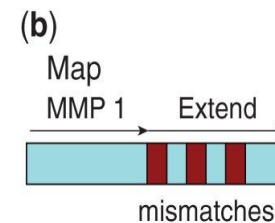
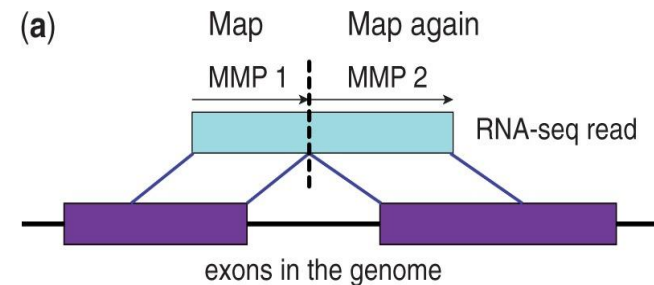
	Bowtie	Bowtie2	BWA	SOAP2	MAQ	RMAP	GSNAP	FANGS	Novoalign	mrFAST	mrsFAST
Seed mm.	Up to 3		Any	Up to 2	Any	Any					
Non-seed mm.	QS	AS	Count	Count	QS	Count	Count	Count	QS	Count	Count
Var. seed len.	> 5		Any	> 28							
Mapping qual.		Yes	Yes		Yes				Yes		
Gapped align.		Yes	Yes	PE	PE		Yes	Yes	Yes	Yes	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

STAR: Splicing Transcripts Alignment to a Reference

- Non-contiguous nature of transcripts, presence of splice-forms make short-read (36-200nt) 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 (a)
- Can detect chimeric transcripts
- Generate indices using a reference genome fasta, and

annotation gtf or gff from Ensembl/UCSC.



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 Pachter video](#)

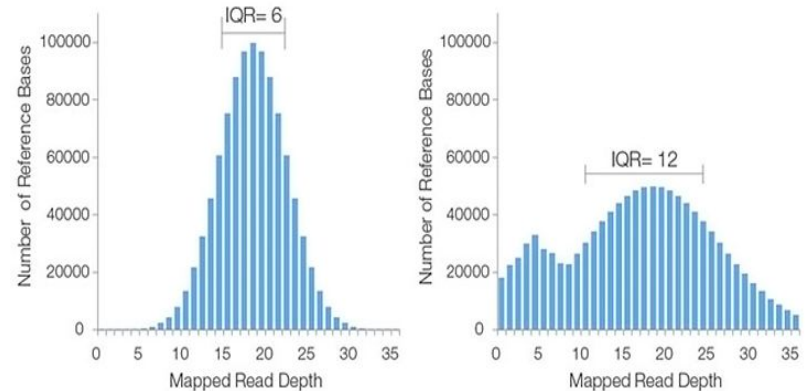
Coverage and Depth

Coverage: 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.

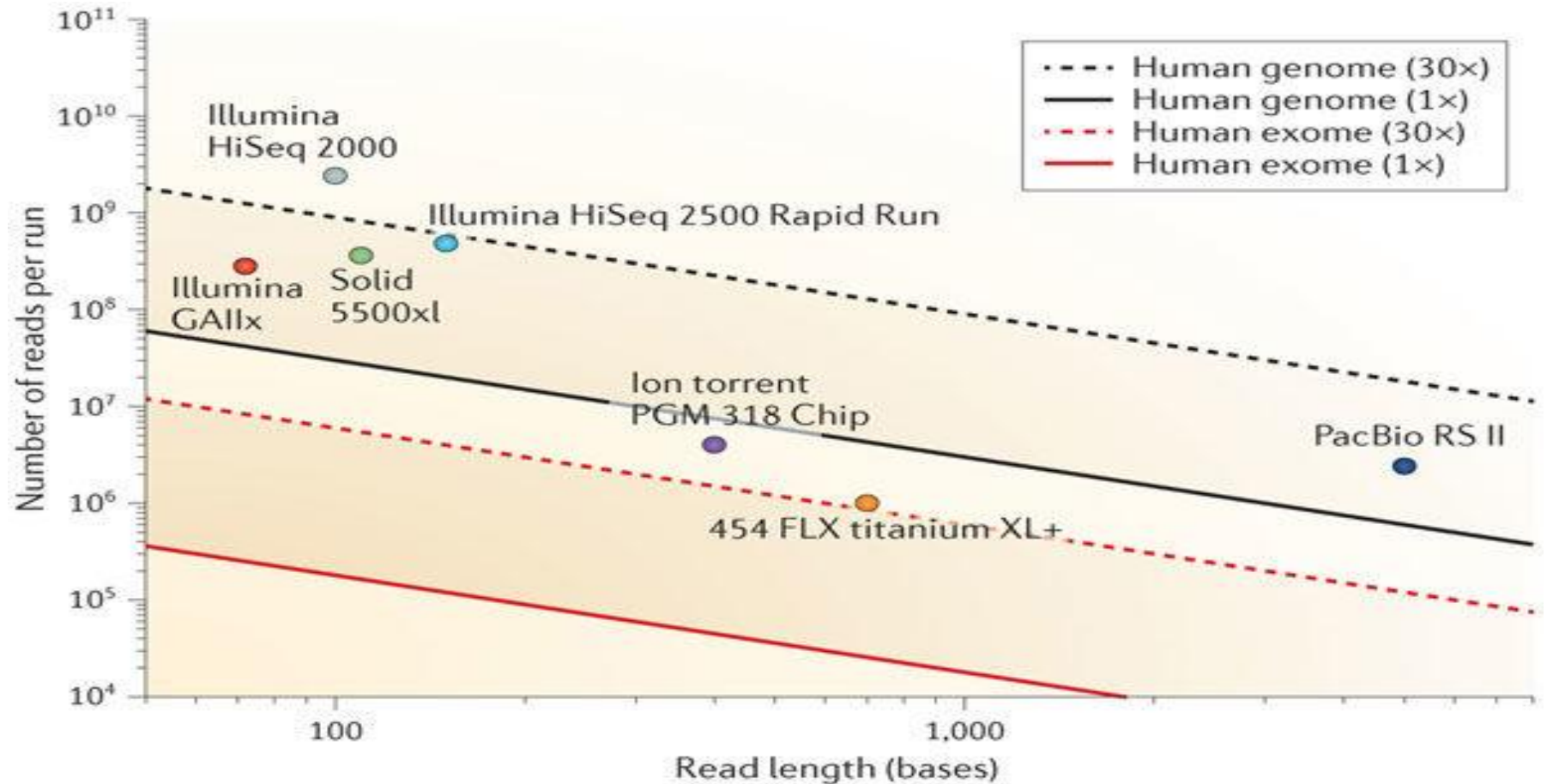
Depth: redundancy of coverage or the total number of bases sequenced and aligned at a given reference position.

Increased depth of coverage rescues inadequacies of sequencing methods.

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



Lander-Waterman model of Coverage



Mappability

Organism	Genome size (Mb)	Nonrepetitive sequence		Mappable sequence	
		Size (Mb)	Percentage	Size (Mb)	Percentage
<i>Caenorhabditis elegans</i>	100.28	87.01	86.8%	93.26	93.0%
<i>Drosophila melanogaster</i>	168.74	117.45	69.6%	121.40	71.9%
<i>Mus musculus</i>	2,654.91	1,438.61	54.2%	2,150.57	81.0%
<i>Homo sapiens</i>	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.

Decoy and Sponge databases

- **The decoy** contains human sequences missing from the hg19 reference, mitochondrial sequences and viral sequences integrated into the human genome. [blog article on decoys](#)
- **The sponge** contains ribosomal and mitochondrial sequences, non-centromeric Huref sequences absent in GRCh38 (hg38), centromeric models etc (Miga et al., 2015).
- These mop up ambiguous sequences, resulting in more accurate and faster alignment.

Nucleic Acids Research

[Nucleic Acids Res.](#) 2015 Nov 16; 43(20): e133.

Published online 2015 Jul 10. doi: [10.1093/nar/gkv671](https://doi.org/10.1093/nar/gkv671)

PMCID: PMC4787761

Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments

[Karen H. Miga](#),* [Christopher Eisenhart](#), and [W. James Kent](#)

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

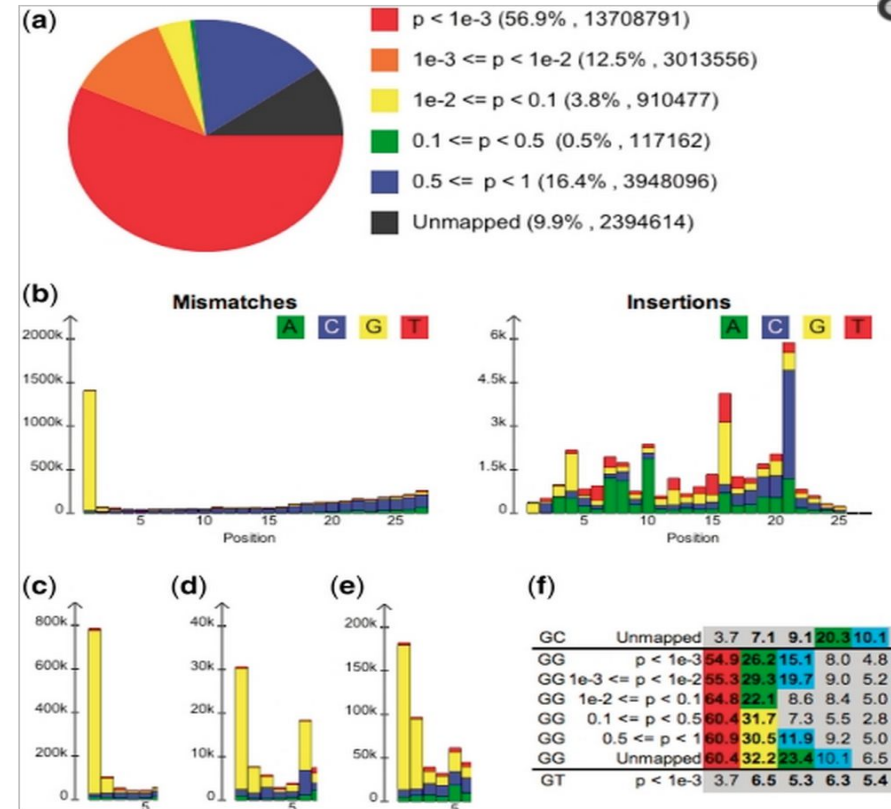
SAMStat for mapping QC

- 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, causing 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

^aOnly reported for SAM files.



Lassmann et al., 2011, Bioinformatics.

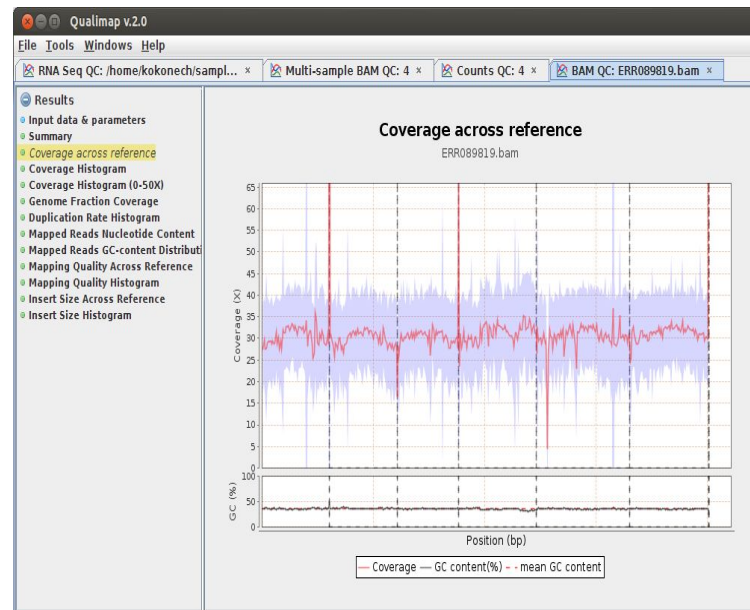
Qualimap

Qualimap provides both a GUI and a command-line interface to facilitate the quality control of alignment sequencing data and feature counts.

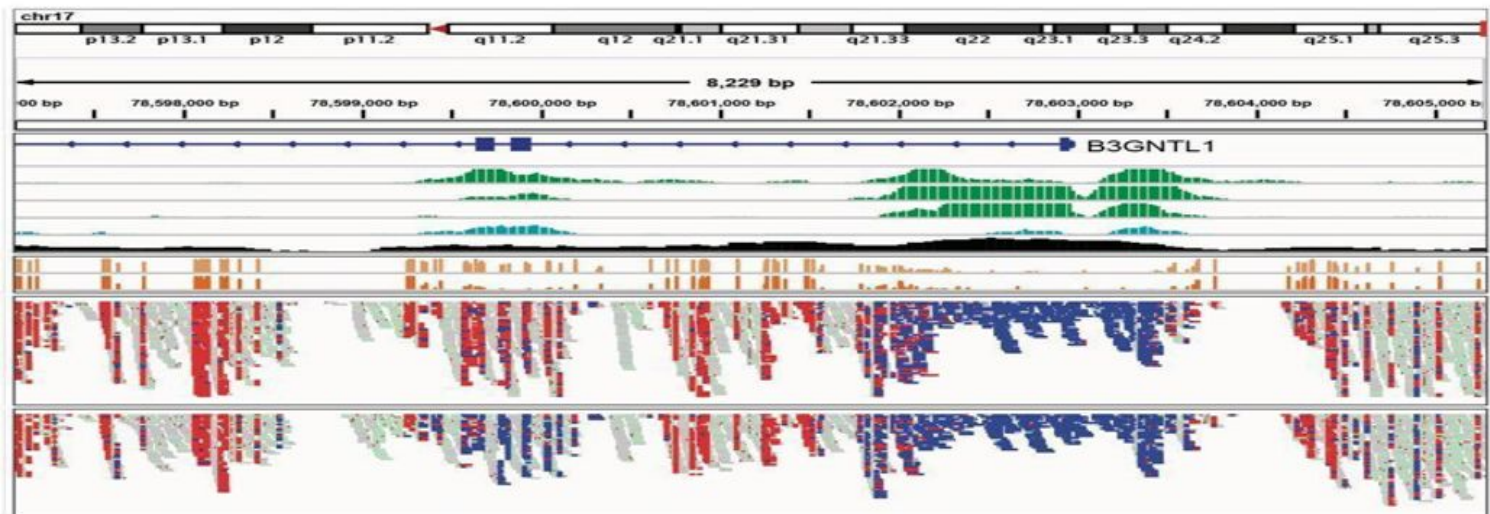
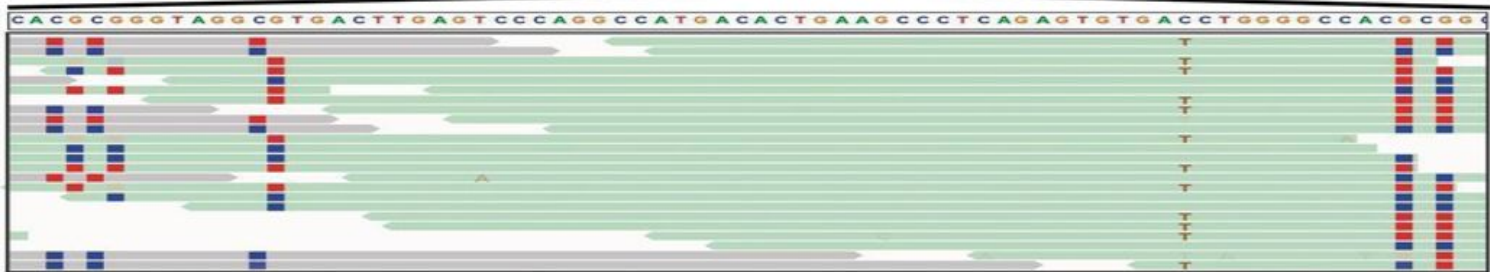
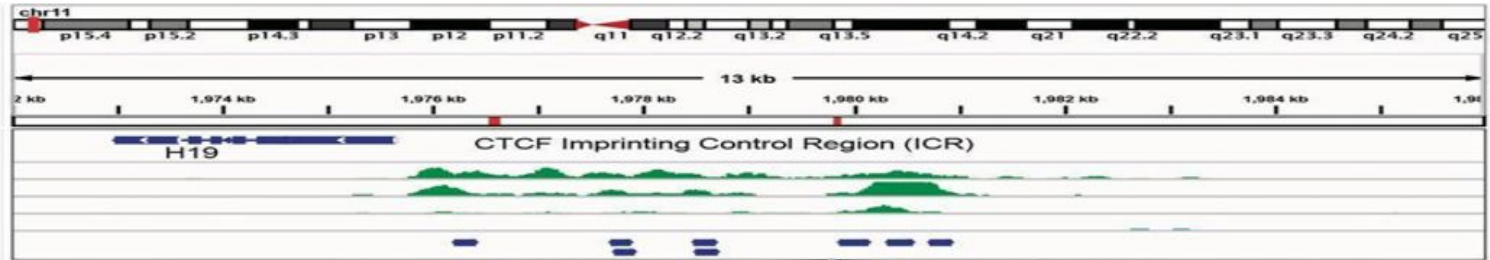
Supported types of experiments include:

- Whole-genome sequencing
- Whole-exome sequencing
- RNA-seq (special mode available)
- ChIP-seq

<http://qualimap.bioinfo.cipf.es/>

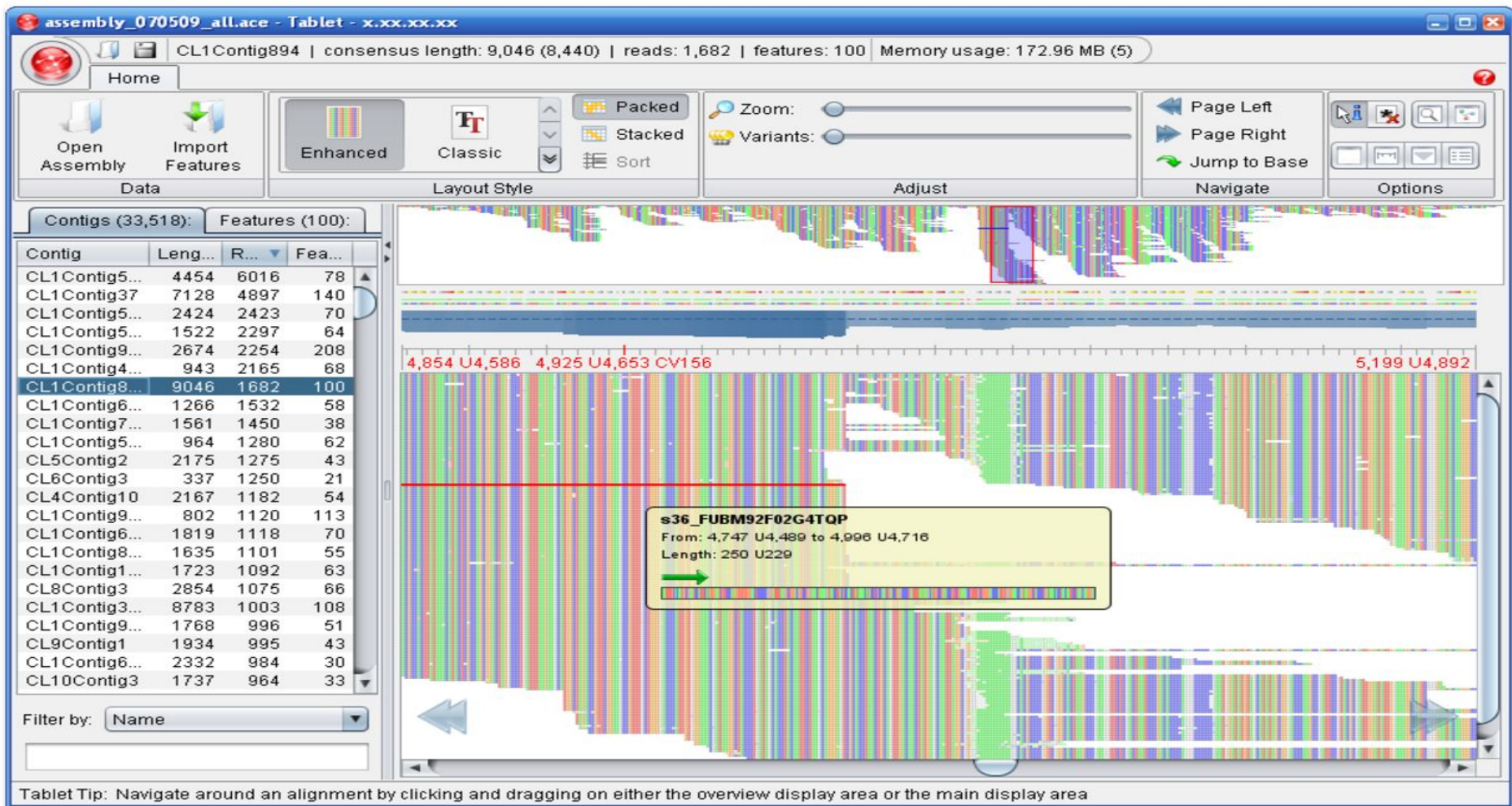


Visualization: IGV



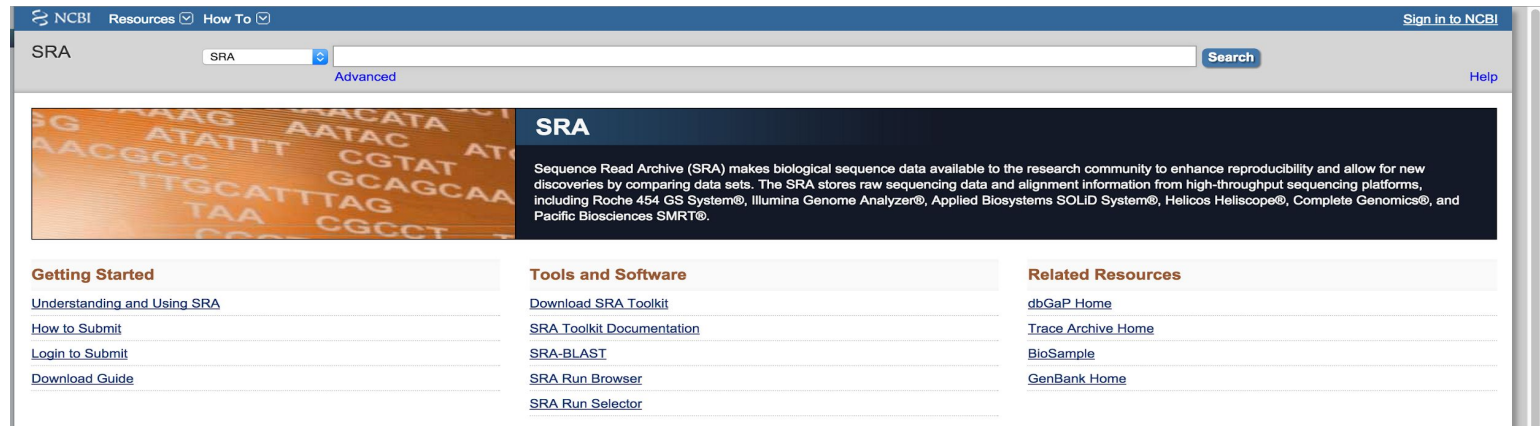
Integrated
Genome
Viewer (IGV)

Visualization: Tablet



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



The screenshot shows the NCBI SRA (Sequence Read Archive) website. At the top, there is a navigation bar with 'NCBI', 'Resources', and 'How To' links, along with a 'Sign in to NCBI' button. Below this is a search bar with 'SRA' entered and a 'Search' button. The main content area features a large banner with a background image of DNA sequence data and the text: '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 System®, Illumina Genome Analyzer®, Applied Biosystems SOLID System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.'

Below the banner, there are three columns of links:

- 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 (Oxford Nanopore, 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.

