L3: Short Read Alignment to a Reference Genome

Shamith Samarajiwa

CRUK Autumn School in Bioinformatics
Cambridge, September 2017

UNIVERSITY OF CAMBRIDGE

MRC Cancer Unit

Where to get help!

SEQanswers the next generation sequencing community

http://seqanswers.com

http://www.biostars.org

Biostar BIOINFORMATICS EXPLAINED

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

http://www.bioconductor.org/help/mailing-list

Read the posting guide before sending email!

Overview

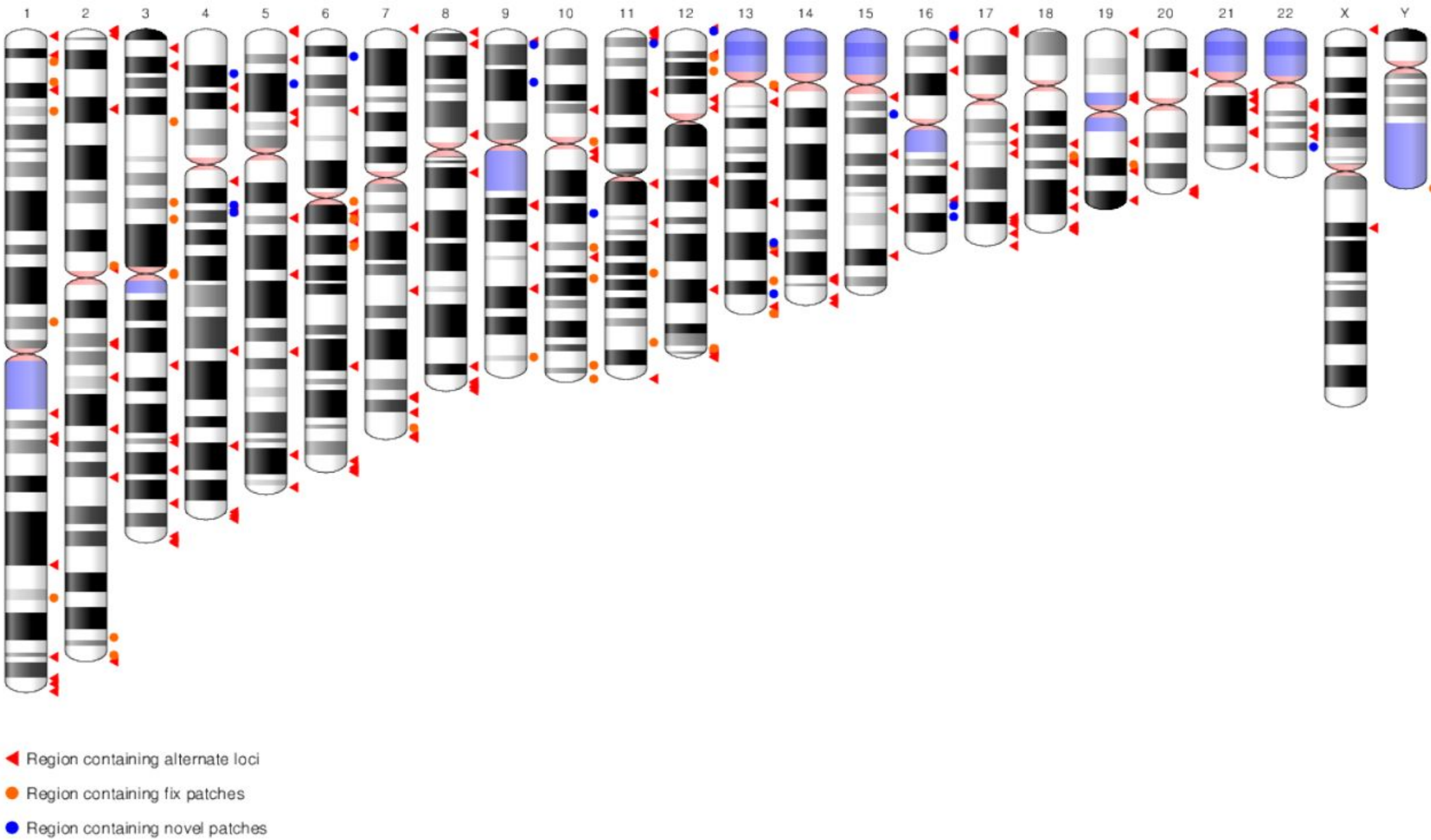
- Understand the difference between reference genome builds
- Introduction to Illumina sequencing
- Short read aligners
  - BWA
  - Bowtie
  - STAR
  - Other aligners
- Coverage and Depth
- Mappability
- Use of decoy and sponge databases
- Alignment Quality, SAMStat, Qualimap
- Samtools and Picard tools,
- Visualization of alignment data
- A very brief look at long reads, graph genome aligners and *de novo* genome assembly

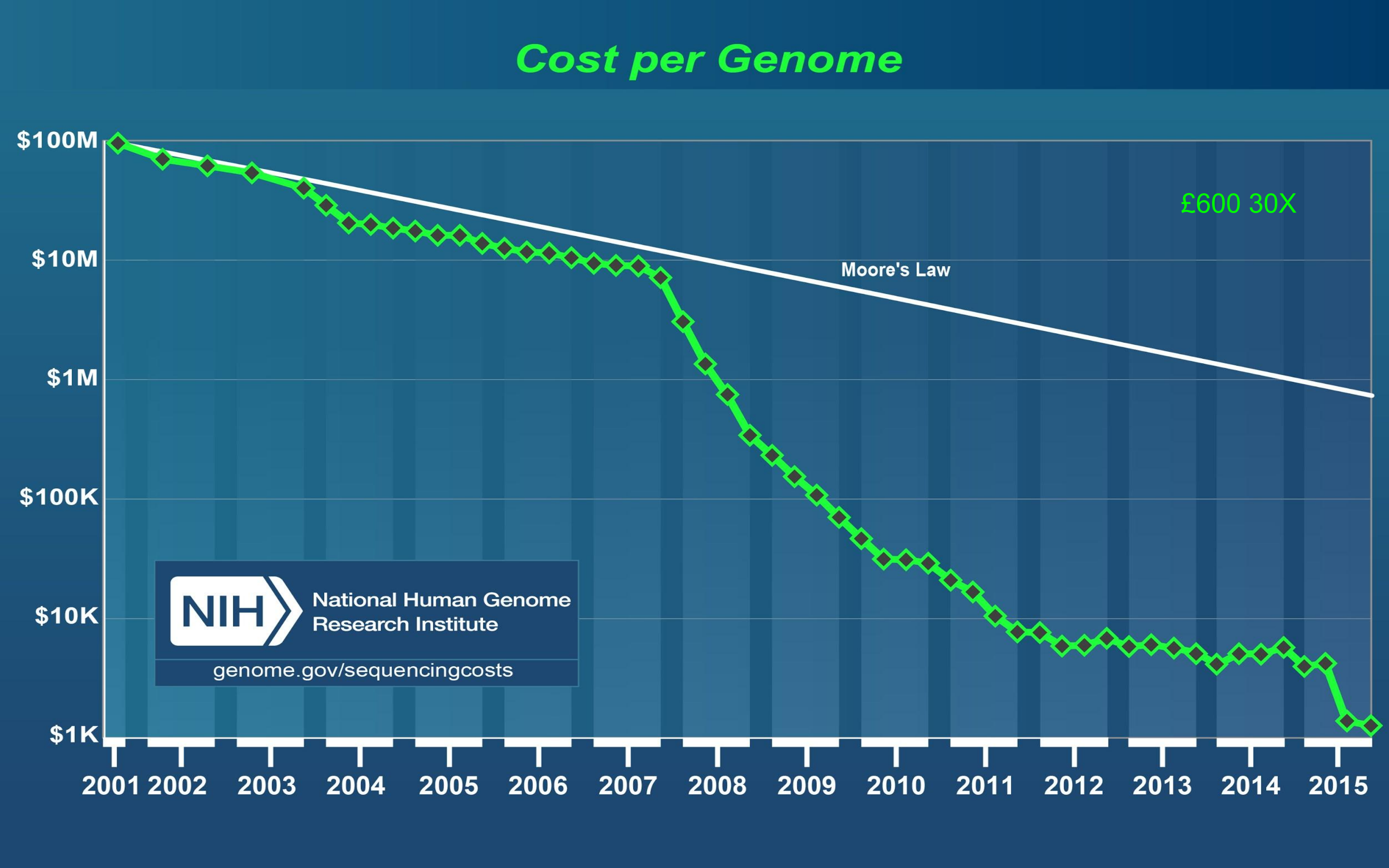
Reference Genomes

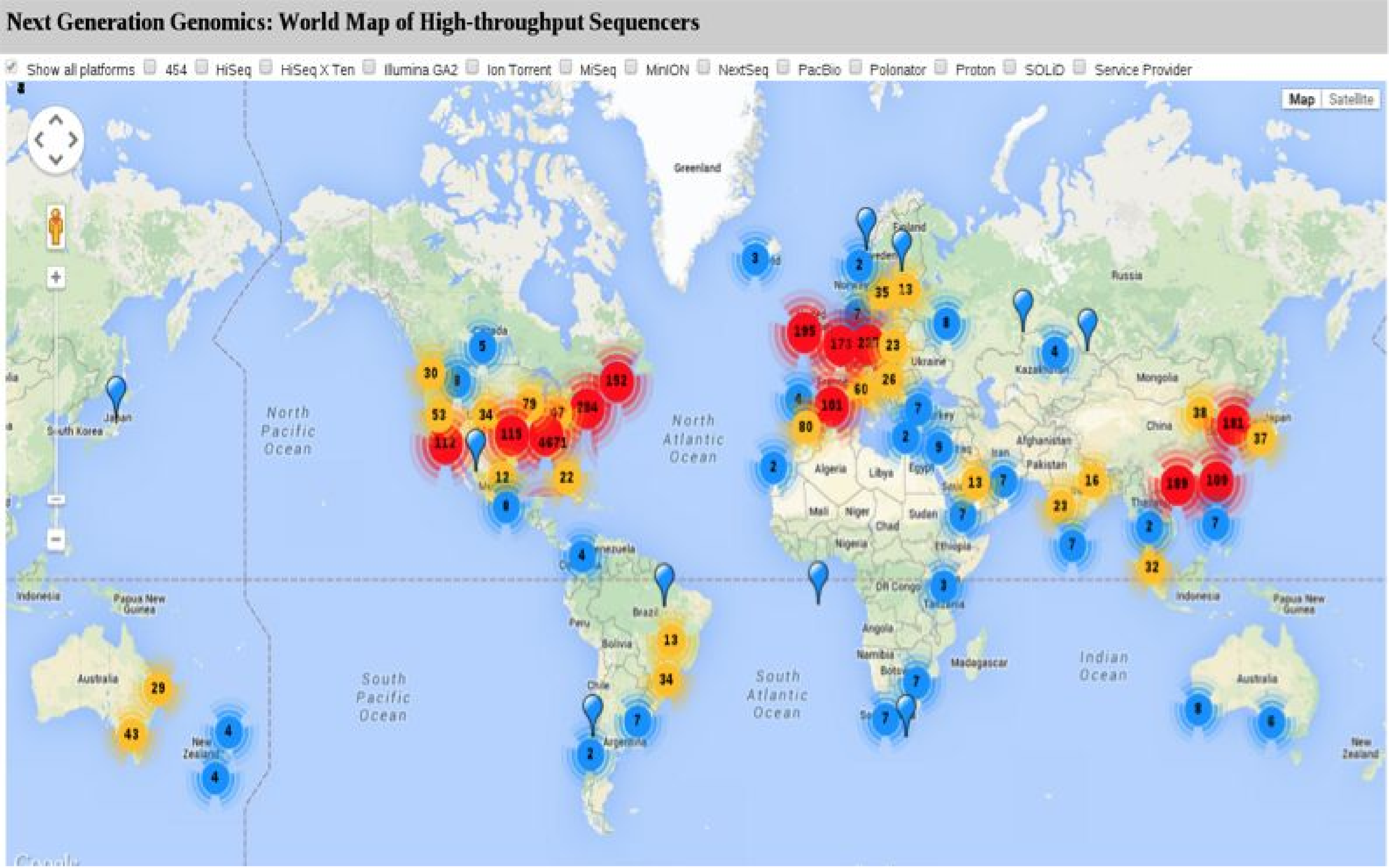
- A haploid representation of a species genome.
- The human genome is a haploid mosaic derived from 13 volunteer donors from Buffalo, NY. USA.
- In regions where there is known large scale population variation, sets of alternate loci (178 in GRCh38) are assembled alongside the reference locus.
- The current build has around 500 gaps, whereas the first version had ~150,000 gaps.

Genome Reference Consortium: https://www.ncbi.nlm.nih.gov/grc

GRCh 38


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Region containing alternate loci
Region containing fix patches
Region containing novel patches


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
£600 30X
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015



Next Generation Genomics: World Map of High-throughput Sequencers
Show all platforms
454
HiSeq
HiSeq X Ten
Illumina GA2
Ion Torrent
MiSeq
MinION
NextSeq
PacBio
Polonator
Proton
SOLiD
Service Provider
Map
Satellite
Greenland
Russia
Canada
Mongolia
China
Japan
South Korea
Kazakhstan
Ukraine
Afghanistan
Pakistan
Iran
Egypt
Algeria
Libya
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
DR Congo
Tanzania
Angola
Namibia
Madagascar
Venezuela
Brazil
Peru
Bolivia
Chile
Argentina
Indonesia
Papua New Guinea
Australia
New Zealand
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean

Illumina sequencers

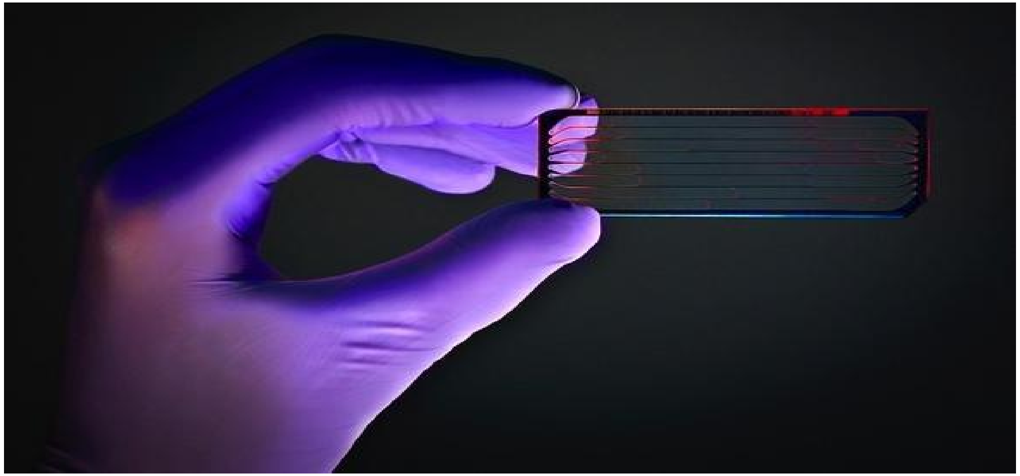



	NextSeq*	HiSeq 4000*	NovaSeq 5000††*	NovaSeq 6000††*
Output Range	20-120 Gb	125-1500 Gb	167-2000 Gb	167-6000 Gb
Run Time	11-29 hr	<1-3.5 days	TBA	19-40 hr
Reads per Run	130-400 million	2.5-5 billion	1.4-6.6 billion	1.4-20 billion
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp
Samples per Run†	8-24	160-320	96-192	96-192
Relative Price per Sample†	Lower Cost	Lower Cost	Lower Cost	Lower Cost
Relative Instrument Price†	Higher Cost	Higher Cost	Higher Cost	Higher Cost

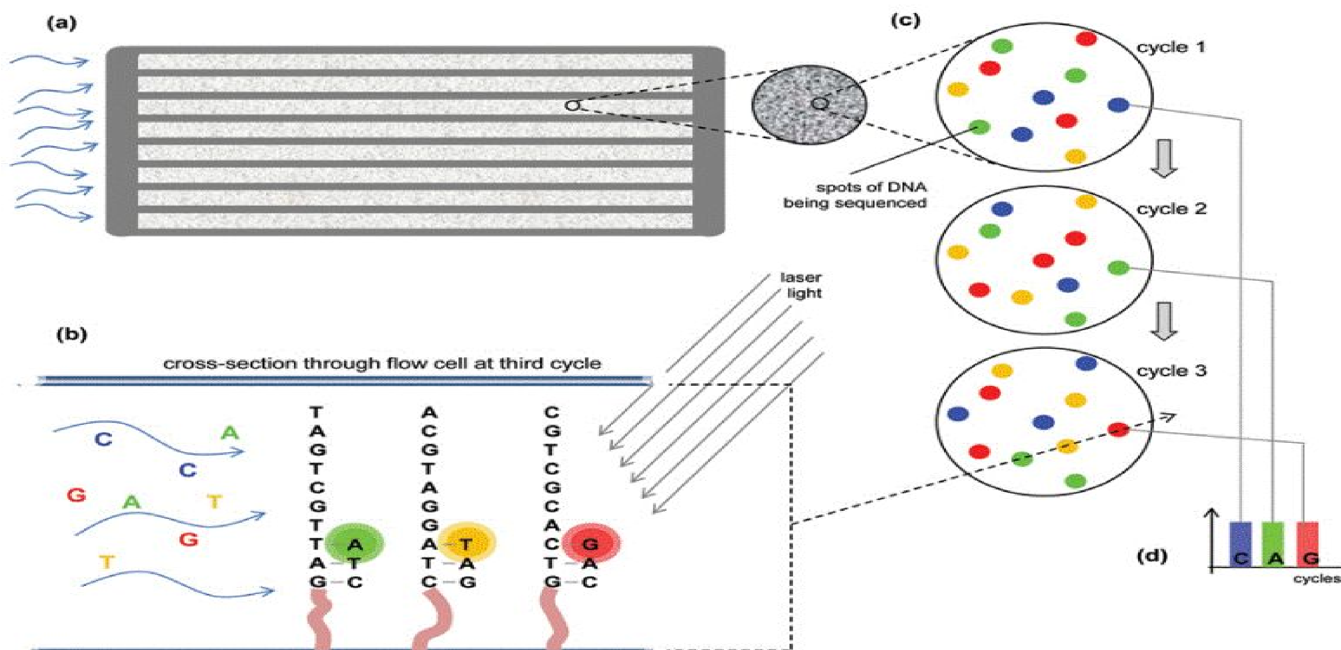
Illumina Genome Analyzer

Illumina sequencing technology

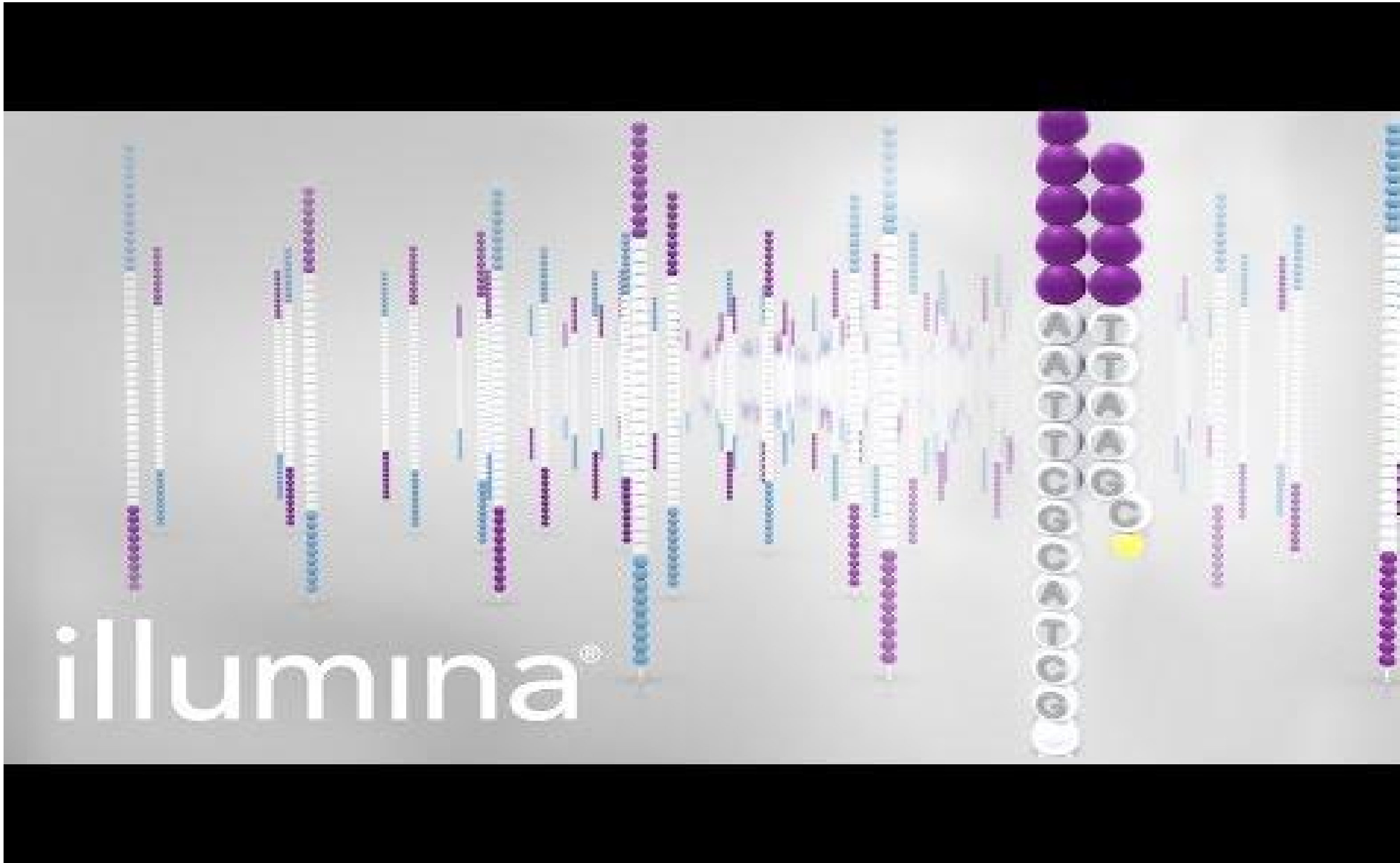
- Illumina sequencing is based on the Solexa technology developed by **Shankar Balasubramanian** and **David Klenerman** (1998) at the University of Cambridge.
- Multiple steps in “Sequencing by synthesis” (explained in next slide)
  - Library Preparation
  - Bridge amplification and Cluster generation
  - Sequencing using reversible terminators
  - Image acquisition and Fastq generation
  - *Alignment and data analysis*



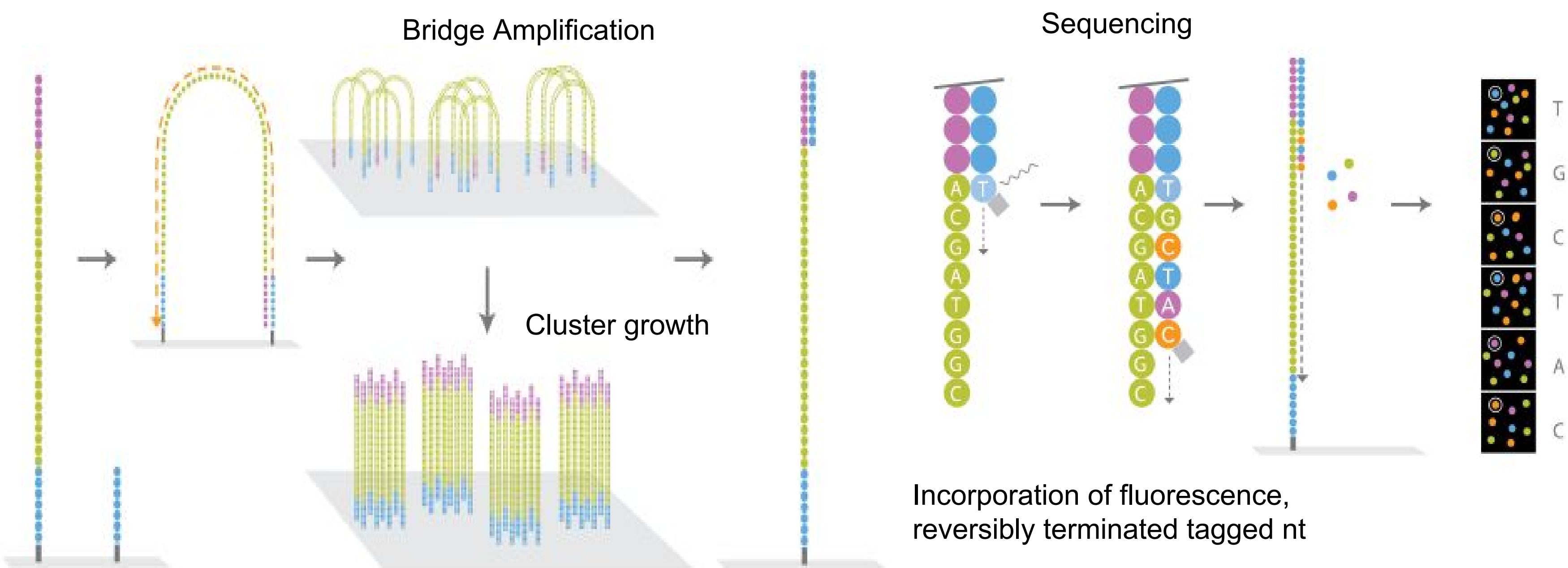
Illumina Flowcell



Sequencing By Synthesis technology



Illumina Sequencing

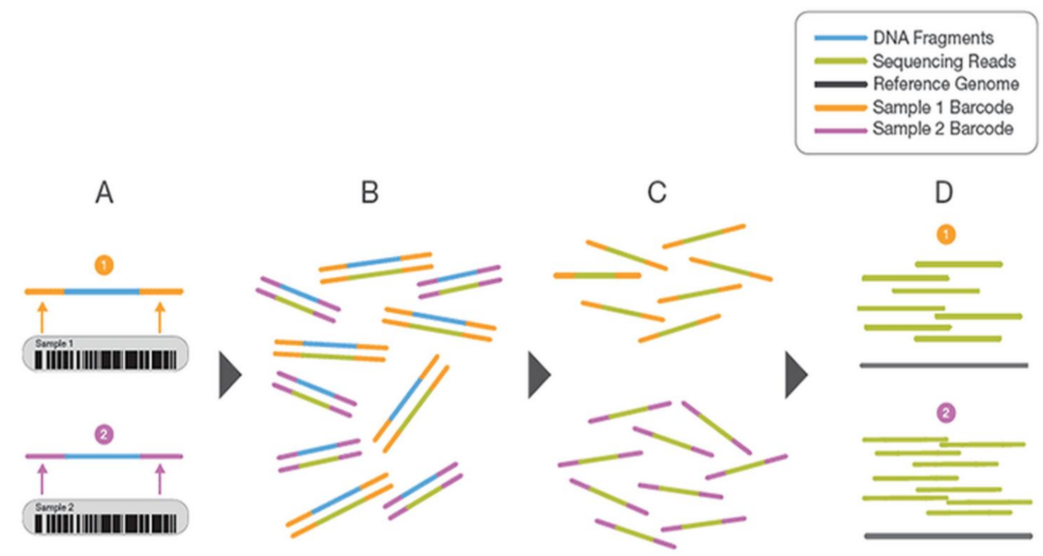


Incorporation of fluorescence,
reversibly terminated tagged nt

Multiplexing

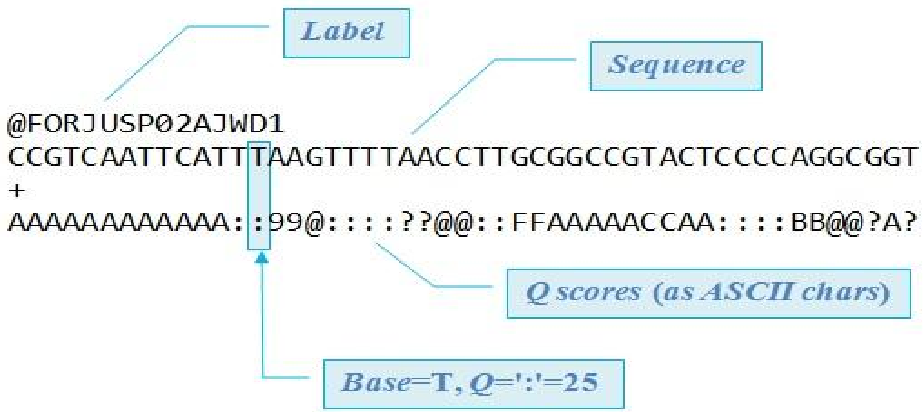
- Multiplexing gives the ability to sequence multiple samples at the same time.
- Useful when sequencing small genomes or specific genomic regions.
- Different barcode adaptors are ligated to different samples.
- Reads de-multiplexd after sequencing.

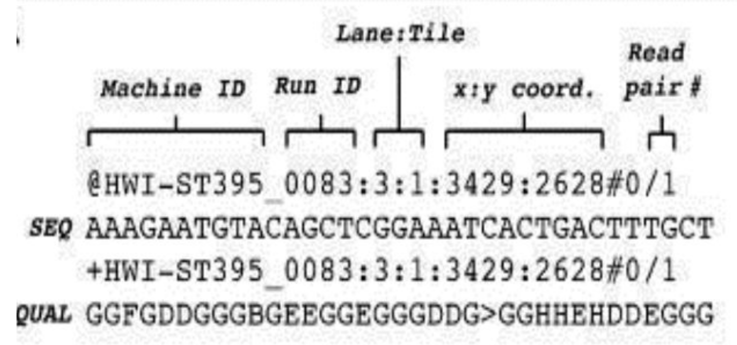
Figure 2: Conceptual Overview of Sample Multiplexing



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

FASTQ format





A FASTQ file normally uses four lines per sequence.

Line-1 begins with a **'@'** character and is followed by a sequence identifier and an optional description.

Line-2 is the raw sequence letters.

Line-3 begins with a **'+'** character and is optionally followed by the same sequence identifier again.

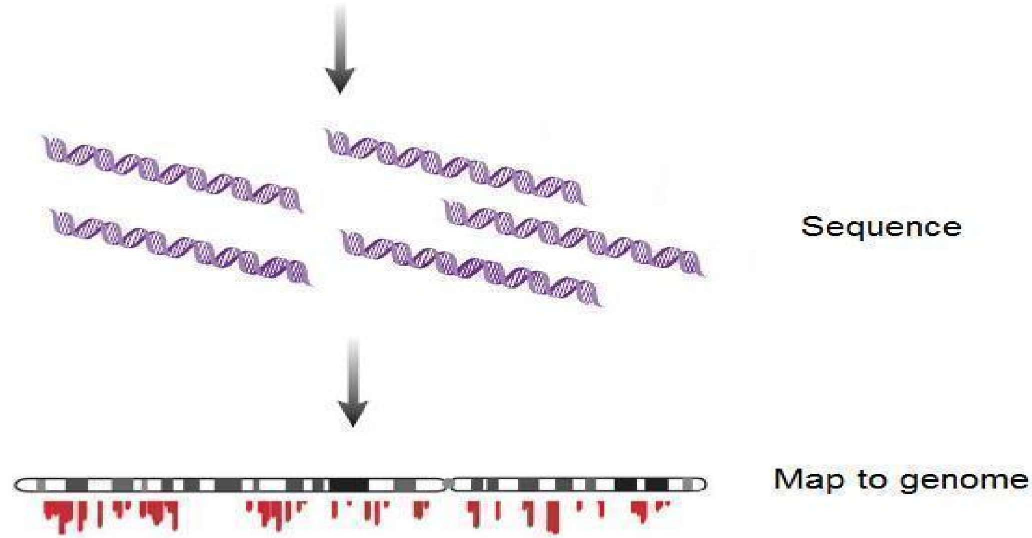
Line-4 encodes the quality scores (ASCII) for the sequence in Line 2.

Historically there are a number of different FASTQ formats. These include the Sanger Format, Illumina/Solexa 1.0, Illumina 1.3, 1.5, 1.8 and 1.9

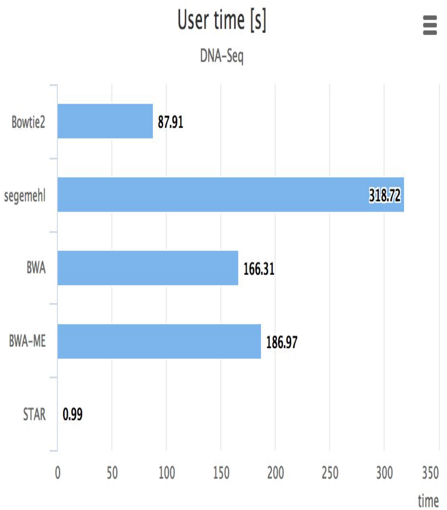
Cock et al., Nucleic Acids Res. 2010 Apr;38(6):1767-71.

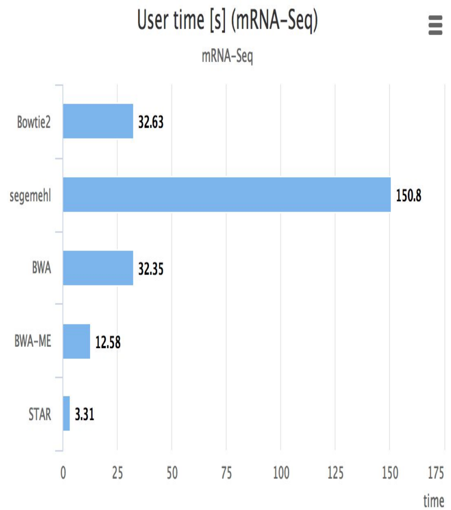
Aligning to a reference genome

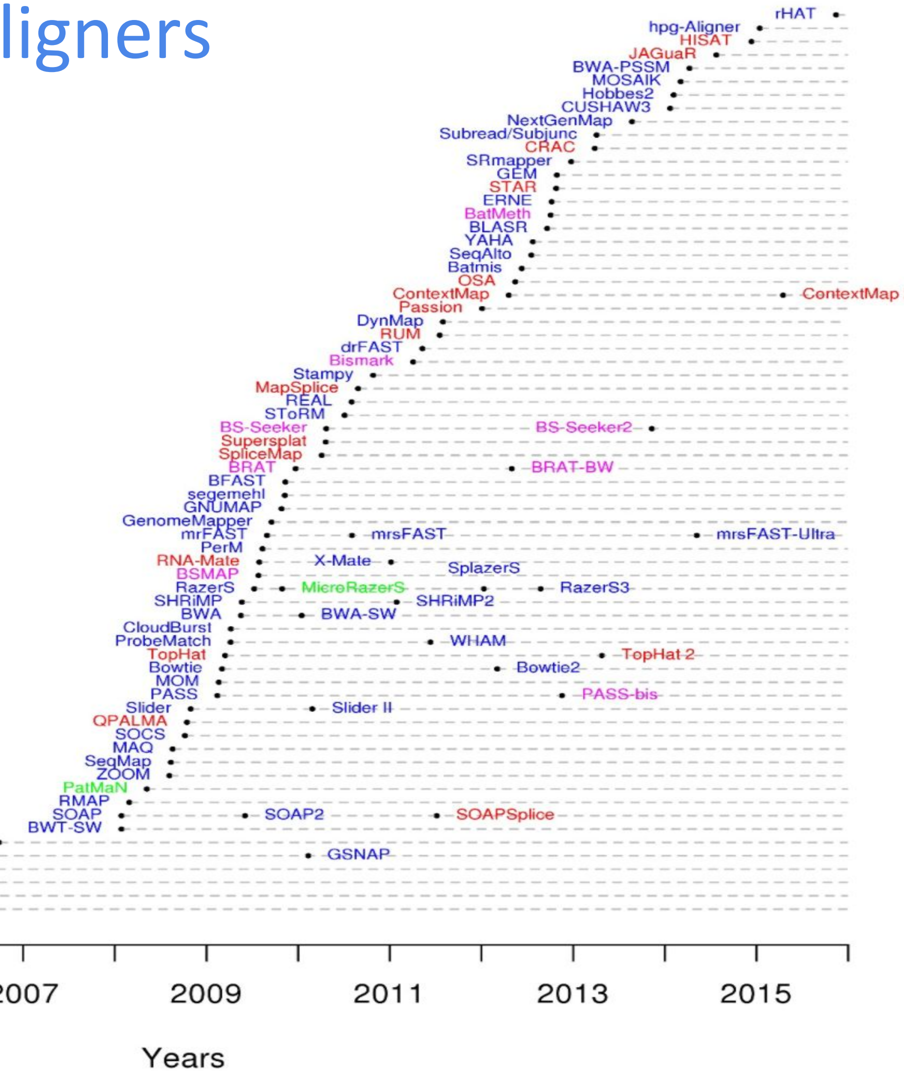
- BWA
- Bowtie2
- STAR
- GEM
- Pseudoaligners for RNA-seq quantification
  - Kallisto
  - Salmon
  - Sailfish



More than 90 Short Read Aligners







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

Short Read Aligners

Table 4: Overall evaluation and comparison of multiple aligners.

	Computational speed			Memory usage			Accuracy			
Aligners	**Speed with single thread**	**Speed with multithread**	**Key factor impacting speed (genome size or read count)**	**Overall evaluation**	**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 s 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.

Download and install BWA

download
wget https://sourceforge.net/projects/bio-bwa/files/bwa-0.7.16a.tar.bz2
extract
tar -xvfj bwa-0.7.16a.tar.bz2
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
cd bwa-0.7.16a
Make
Add BWA to your PATH by editing ~/.bashrc file (or .bash_profile or .profile file)
export PATH=$PATH:/path/to/bwa-0.7.16a
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 1Mbp) 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

Li and Durbin, 2009, Bioinformatics

Bowtie2

- Bowtie2 handles reads longer than 50 nt.
- Given a reference and a set of reads, this method reports at least one good local alignment for each read if one exists.
- Indexing: 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.
- Uses Burrows-Wheeler Transform (BWT)
- 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.
- Need to prepare a genome index.

Features supported by the tools

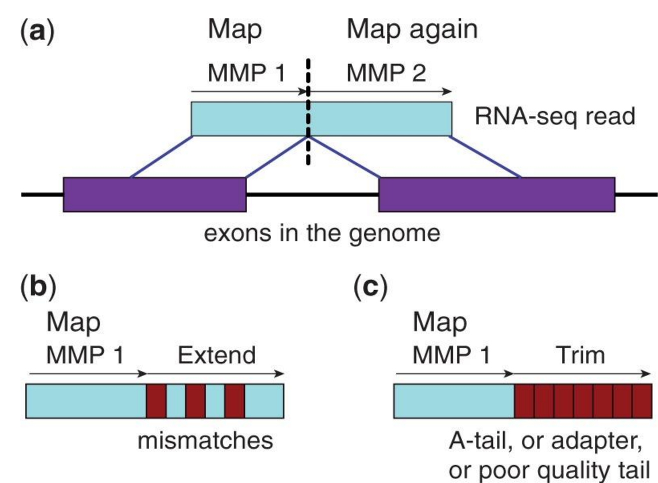
	Bowtie	Bowtie2	BWA	SOAP2	MAQ	RMAP	GSNAP	FANGS	Novoalign	mrFAST	mrsFAST
Seed mm.	Up to 3		Any	Up to 2	Any	Any					
Non-seed mm.	QS	AS	Count	Count	QS	Count	Count	Count	QS	Count	Count
Var. seed len.	> 5		Any	> 28							
Mapping qual.		Yes	Yes		Yes				Yes		
Gapped align.		Yes	Yes	PE	PE		Yes	Yes	Yes	Yes	
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
- Can detect chimeric transcripts
- Generate indices using a reference genome fasta, and

annotation gtf or gff from Ensembl/UCSC genome browsers



Dobin et al., 2013 Bioinformatics

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

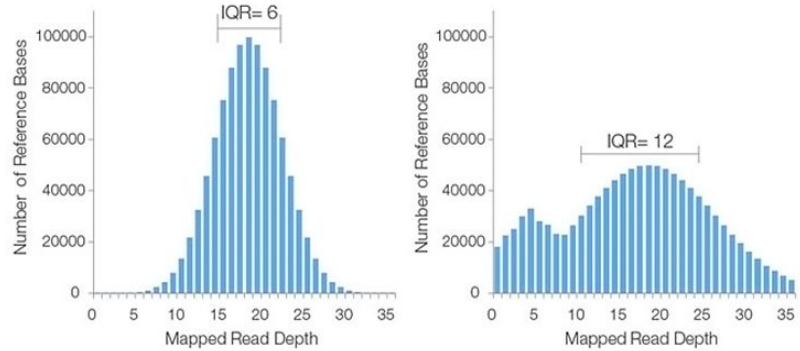
Coverage and Depth

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.

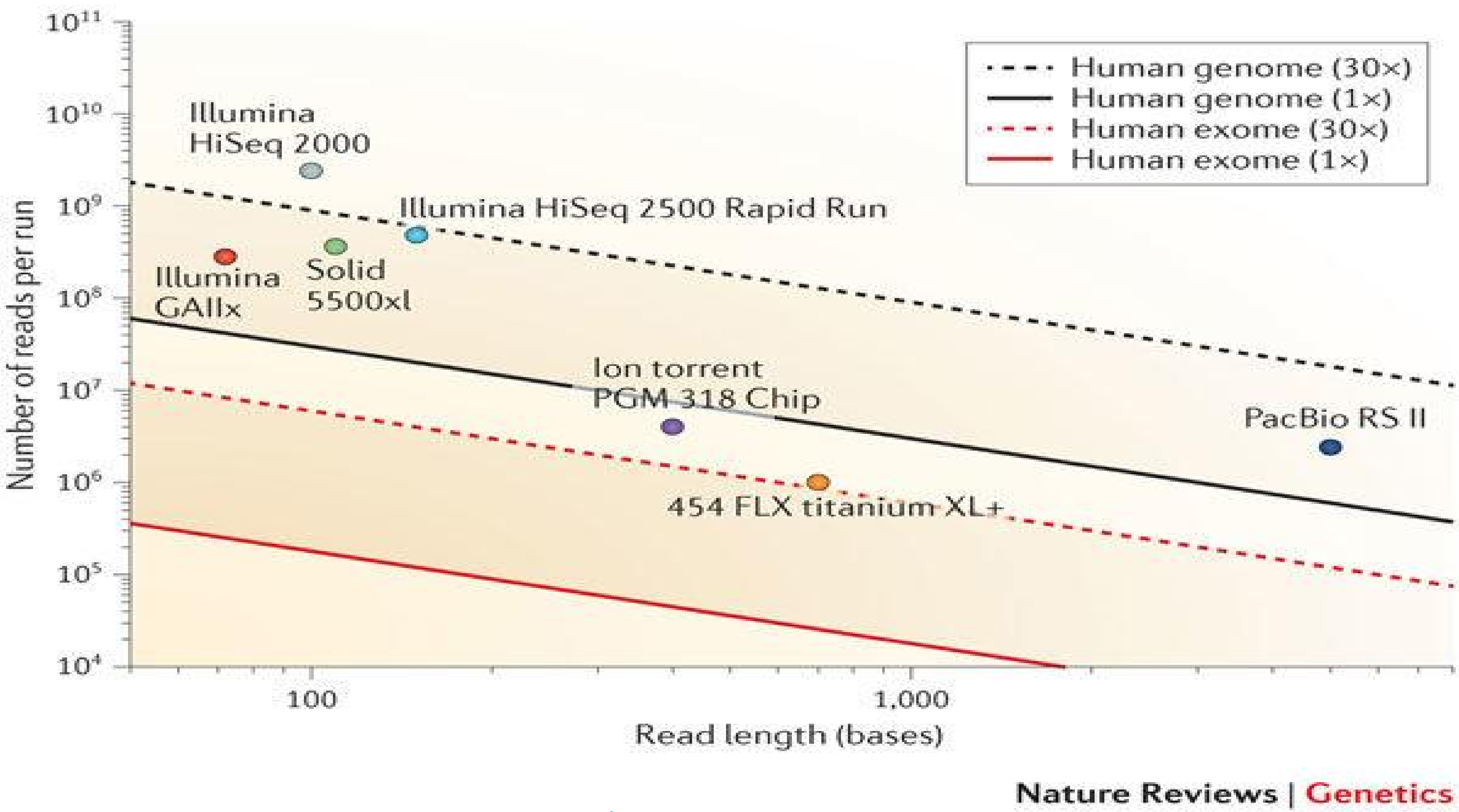
Increased depth of coverage rescues inadequacies of sequencing methods.

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



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

Lander-Waterman model of Coverage



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

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.

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. PMCID: PMC4787761
Published online 2015 Jul 10. doi: 10.1093/nar/gkv671

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

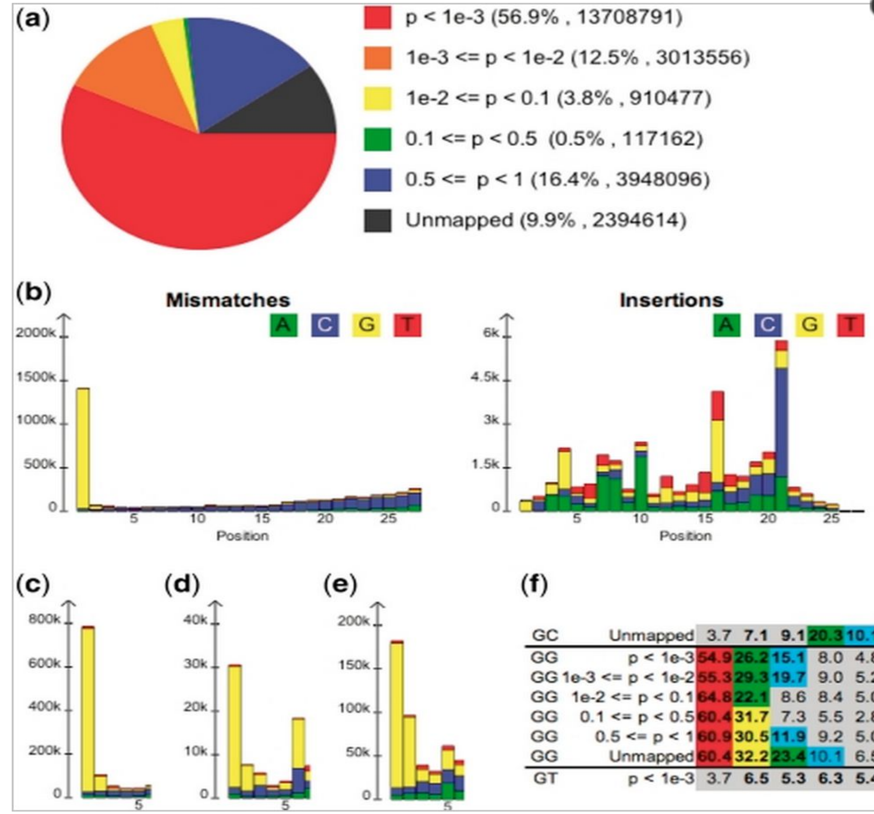
SAMStat for mapping QC

- 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]

[a]Only 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.

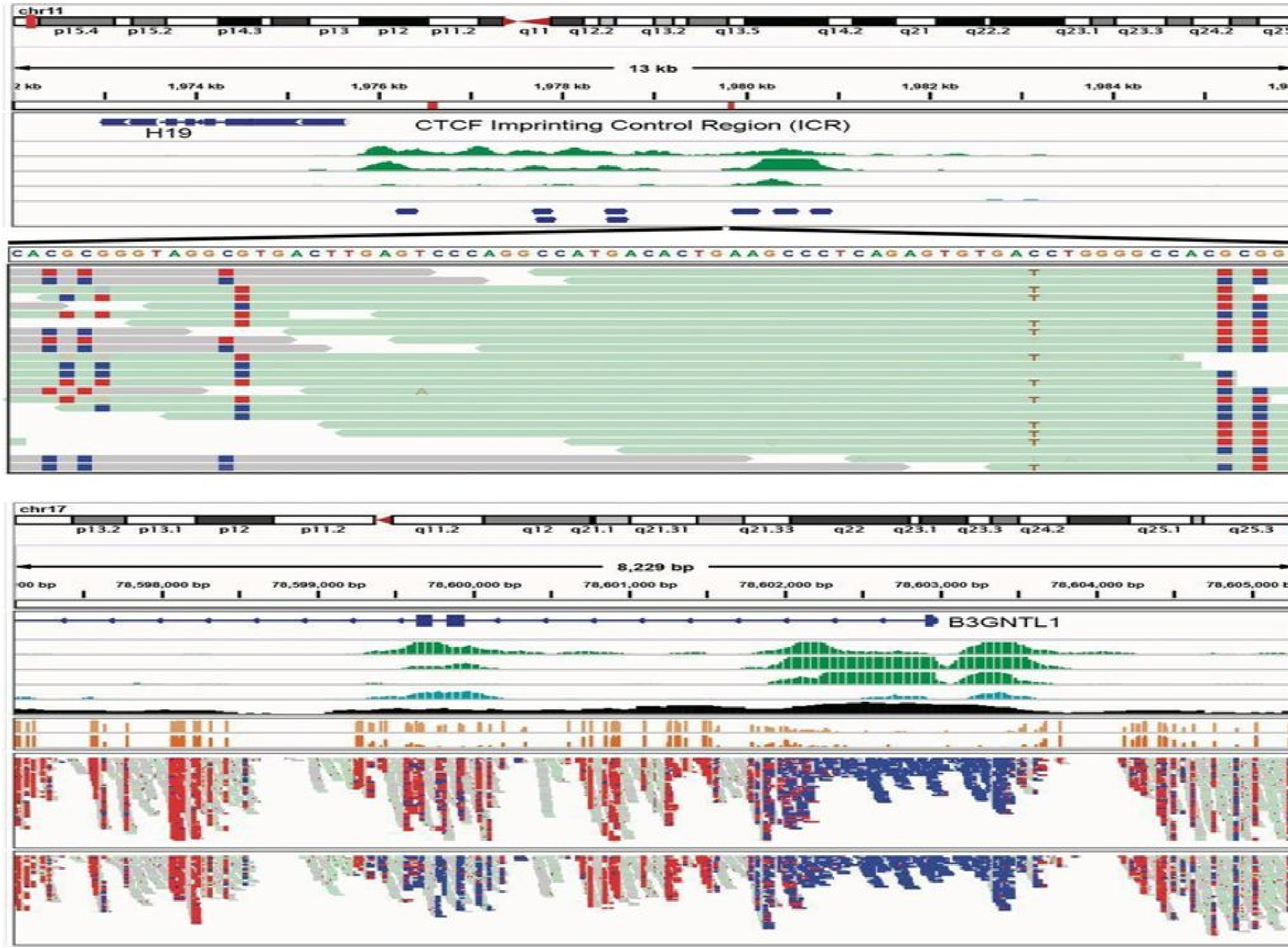
Supported types of experiments include:

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

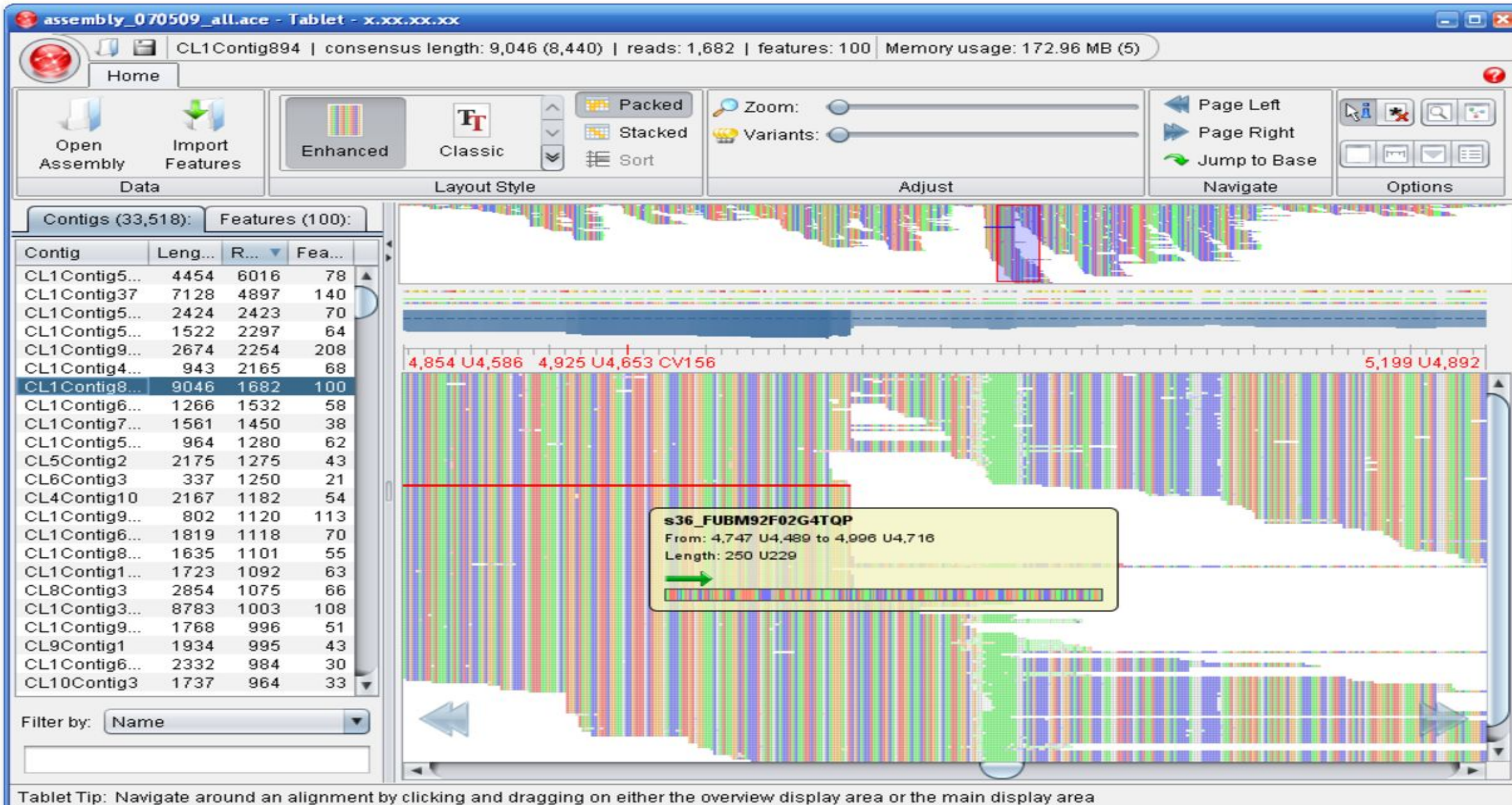
http://qualimap.bioinfo.cipf.es/

Visualizing binding sites and replicates

Integrated Genome Viewer (IGV)

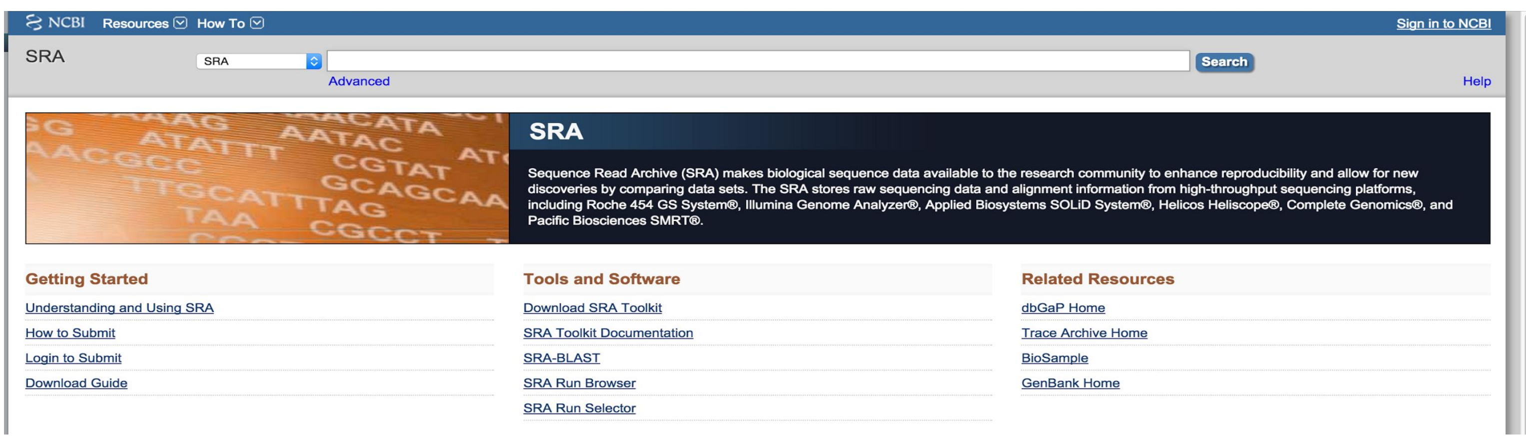


Visualization: Tablet



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 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 Nonopore, 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.