Introduction to ChIP-seq analysis

Shamith Samarajiwa

Integrative Systems Biomedicine Group

MRC Cancer Unit

University of Cambridge

CRUK Bioinformatics Summer School

July 2015

UNIVERSITY OF CAMBRIDGE

Where to get help!

http://seqanswers.com

http://www.biostars.org

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

Read the posting guide before sending email!

Important!!!

- Good Experimental Design

- Optimize Conditions (Cells, Antibodies, Sonication etc.)

- **Biological Replicates (at least 3)**!!
  sample biological variation / improve signal to noise
  capture the desired effect size
  statistical power to test null hypothesis

- ChIP-seq controls – **KO, Input** or IgG

What is ChIP Sequencing?

- Combination of chromatin immunoprecipitation (ChIP) with ultra high-throughput massively parallel sequencing.

- Allows mapping of protein–DNA interactions ***in vivo*** on a genome scale.
- Enables mapping of transcription factors binding, RNA Pol II occupancy or Histone modification marks on a genome scale.

- The typical ChIP assay usually take 4–5 days, and require approx. 10^6~ 10^7 cells.

Origins of ChIP-seq technology

• Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., et al. “High-resolution profiling of histone methylations in the human genome.” *Cell 2007*

• Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. “Genome-wide mapping of *in vivo* protein-DNA interactions.” *Science* 316, 2007

• Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., et al. “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.” *Nature* 2007

• Robertson et al., "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." *Nat Methods.* 2007

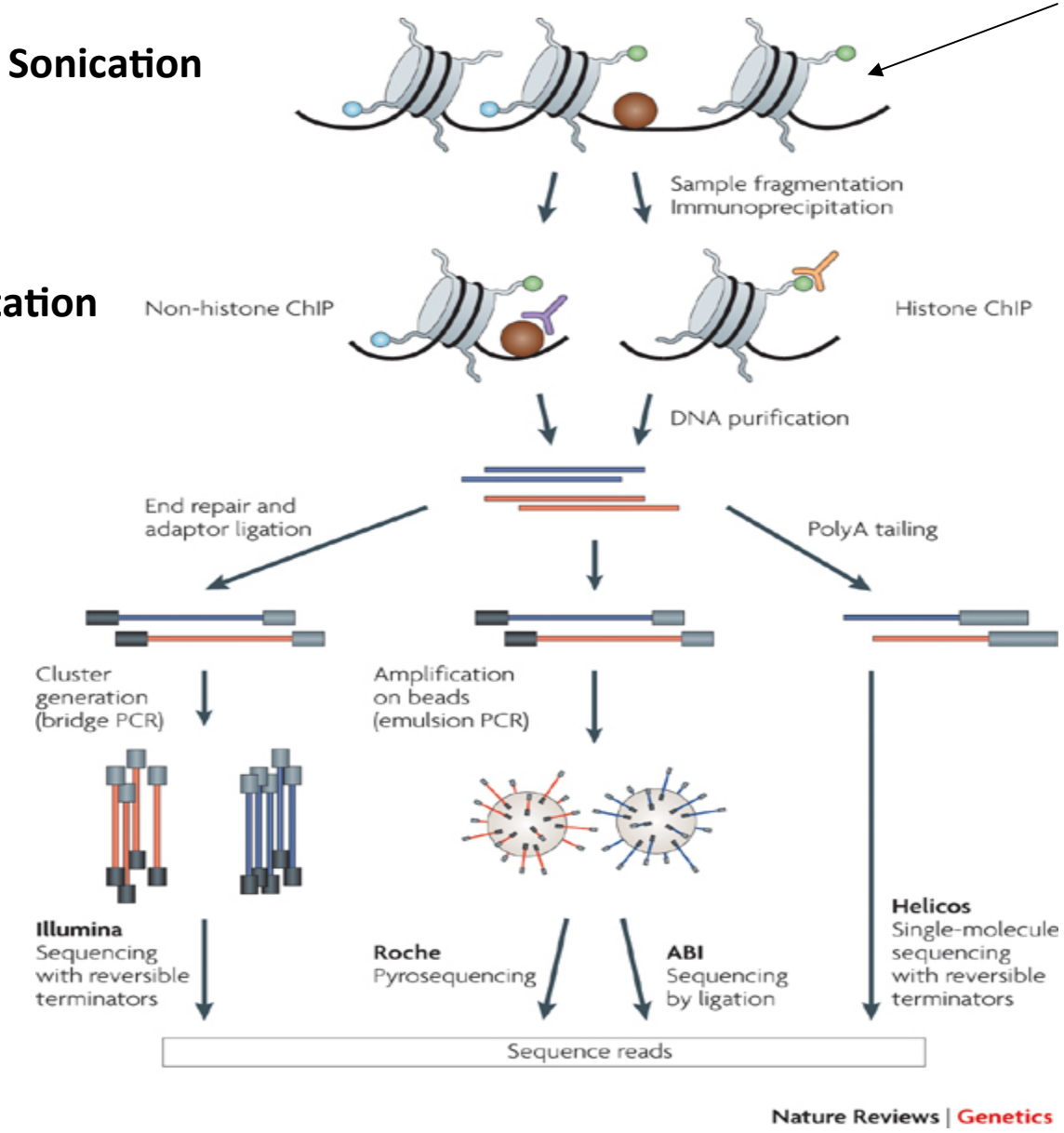
ChIP-seq methodology

Cross-link cells

Isolate genomic DNA

Sonication

Immuno-precipitation



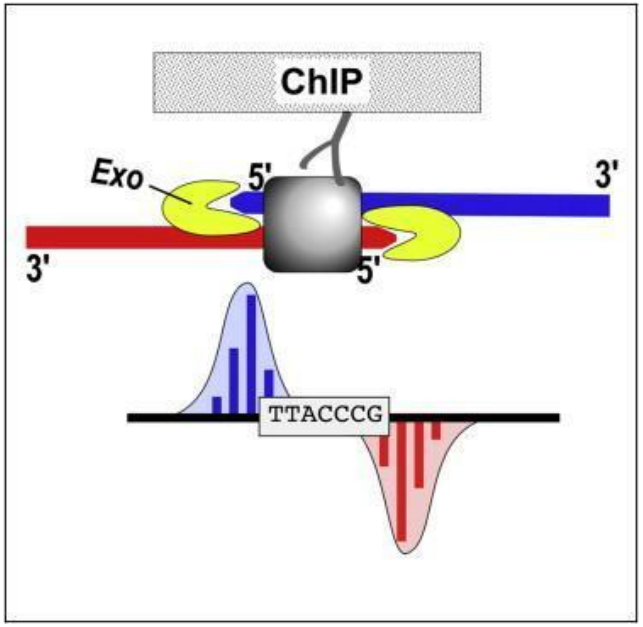
Park 2009 *Nat. Rev Genet.*

Advances in technologies for nucleic acid-protein interaction detection

• ChIP-chip : combines ChIP with microarray technology.

• ChIP-PET : ChIP with paired end tag sequencing

• ChIP-exo : ChIP-seq with exonuclease digestion



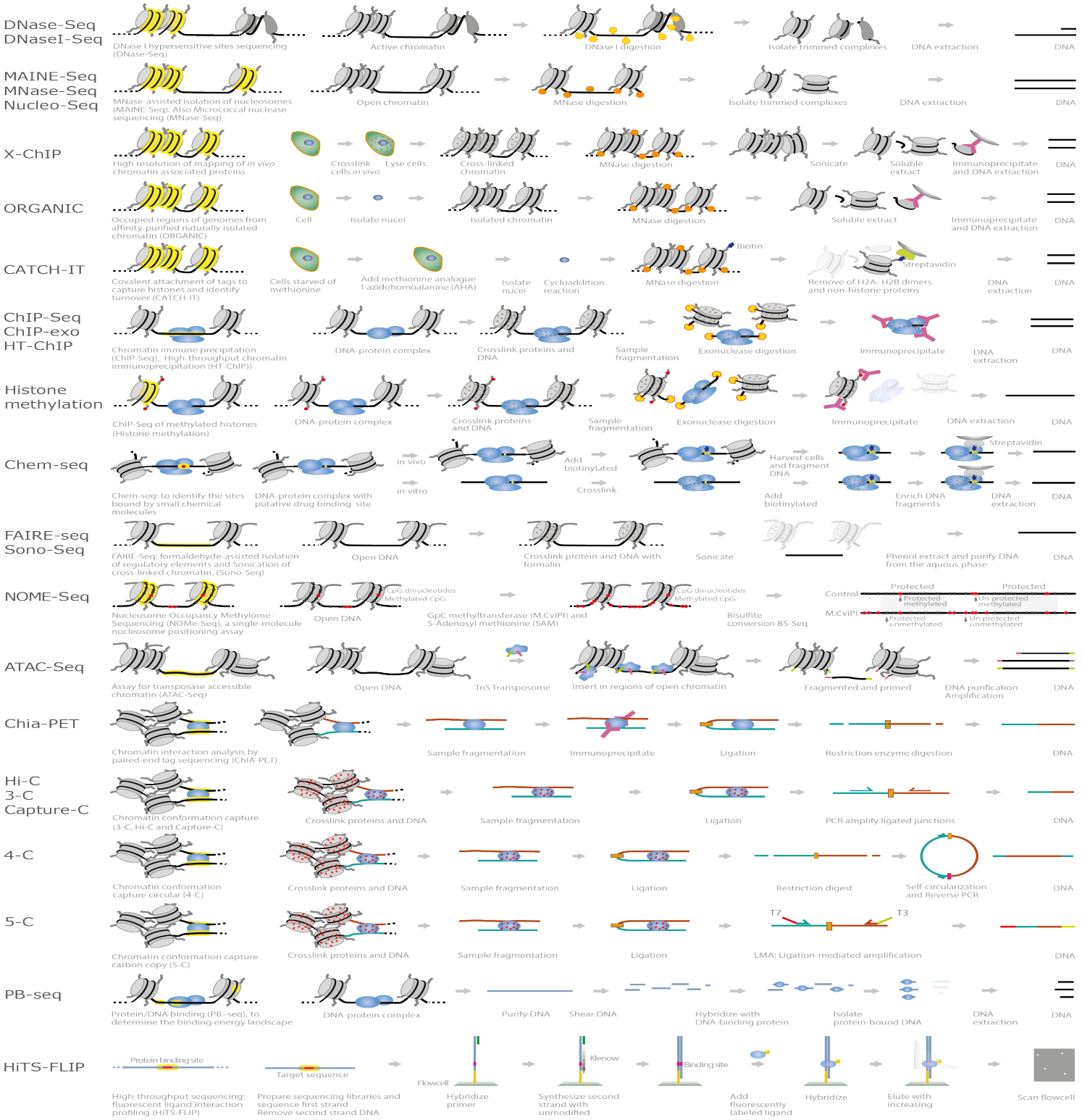
• CLIP-seq / HITS-CLIP : cross-linking immunoprecipitation high throughput sequencing

• ATAC-seq : Assay for Transposon Accessible Chromatin

• Sono-seq : Sonication of cross linked chromatin sequencing.

• Hi-C: High throughput long distance chromatin interactions

DNA-Protein Interactions



Statistical aspects and best practices

Experimental guidelines:

- Landt *et al.*, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." *Genome Res. 2012.*
- Marinov *et al.,* "Large-scale quality analysis of published ChIP-seq data." 2014 *G3*
- Rozowsky *et al.,* "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls" *Nat Biotechnol.* 2009

Statistical aspects:

- Cairns *et al.*, "Statistical Aspects of ChIP-Seq Analysis." *Adv. in Stat Bioinf.*, 2013.
- Carroll TS *et al.*, "Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data." *Front Genet.* 2014
- Bailey *et al.*, "Practical guidelines for the comprehensive analysis of ChIP-seq data." *PLoS Comput Biol.* 2013.
- Sims et al., "Sequencing depth and coverage: key considerations in genomic analyses." *Nat. Rev. Genet*. 2014.

These guidelines address :

- Antibody validation
- Experimental replication
- Sequencing depth
- Data and metadata reporting
- Data quality assessment
- Replicates

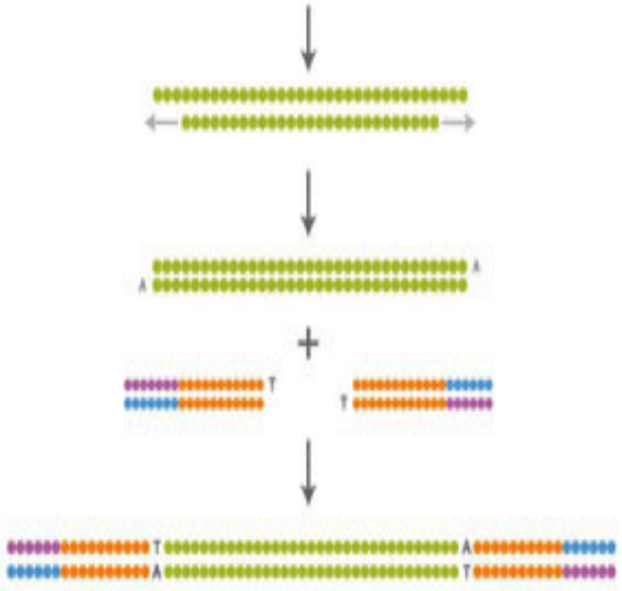
ChIP-seq workflow overview I

- ChIP-seq 'wet-lab processing and library preparation'

Sequencing and Data processing:

- Quality control of raw reads
- Mapping reads to a reference genome
- Remove artefacts and technical noise
- Visualization and Replicate comparison
- Binding site identification: peak calling and other methods
- Peak QC
- Identify replicated, high confidence binding sites (IDR and other methods).

Illumina Genome Analysis System



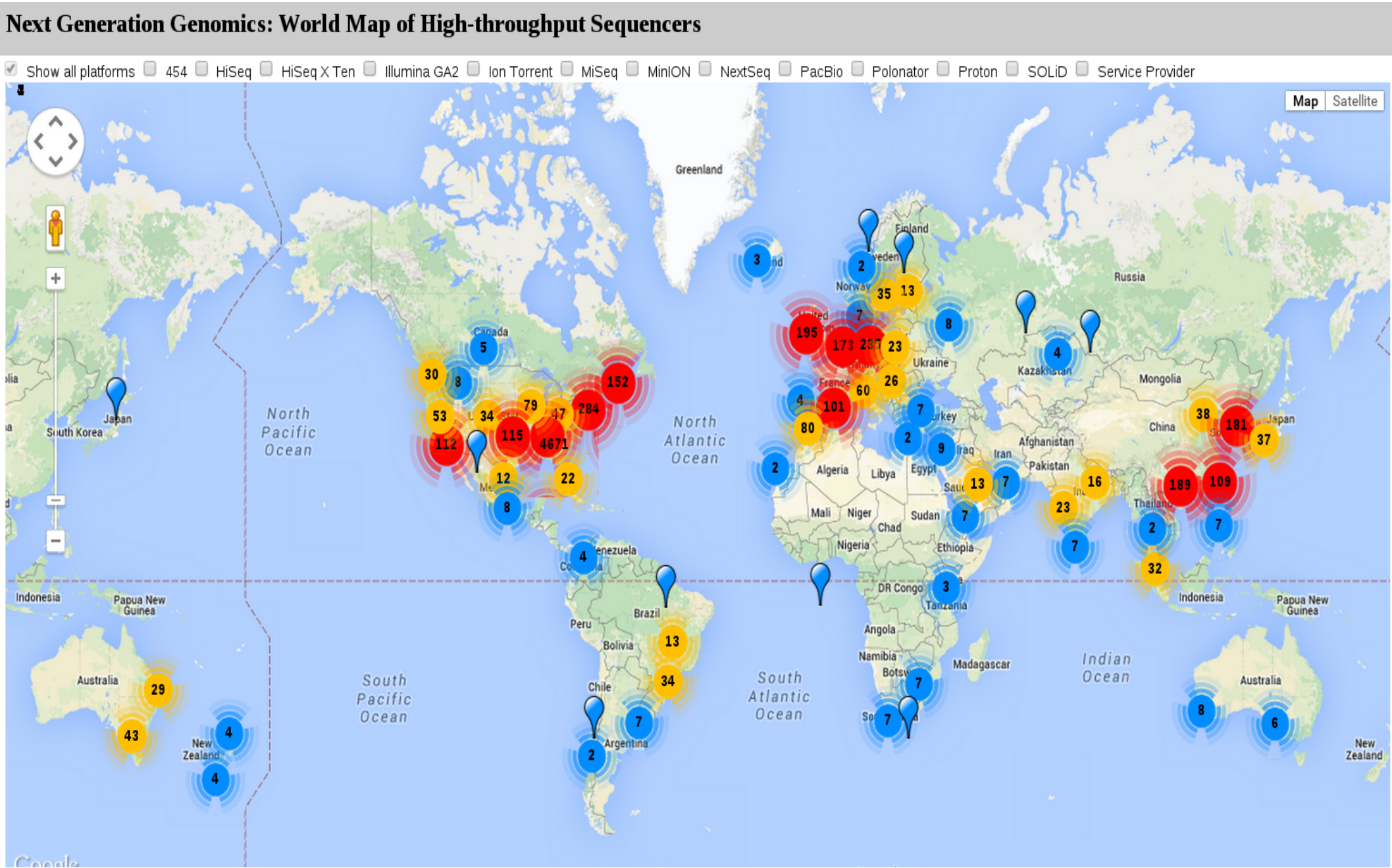
Library Preparation

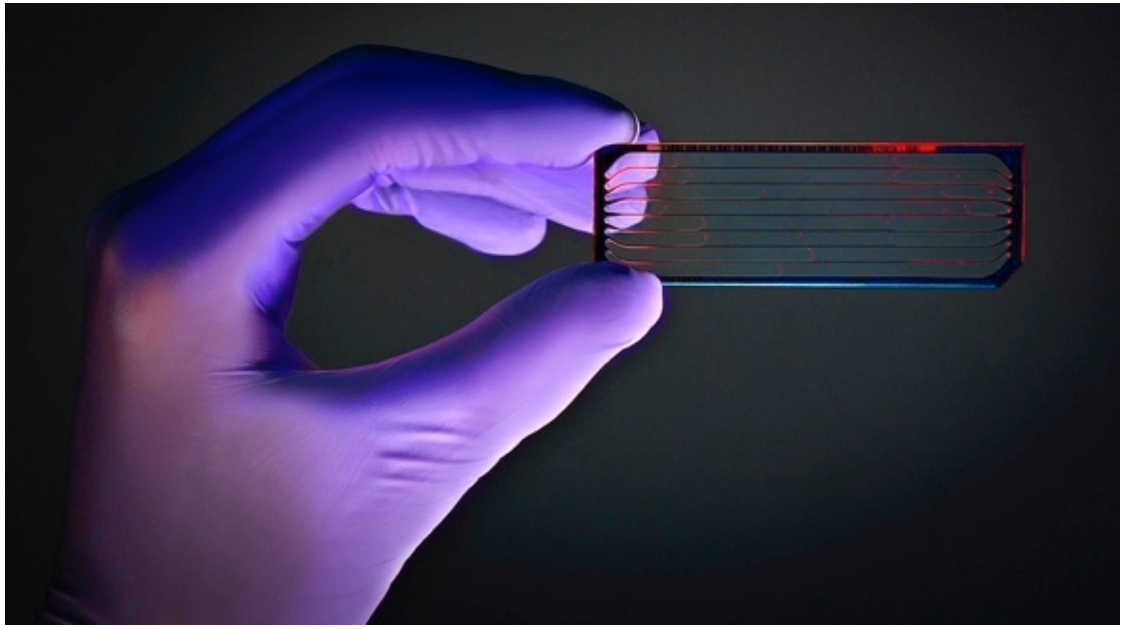
Cluster Generation

Sequencing by Synthesis

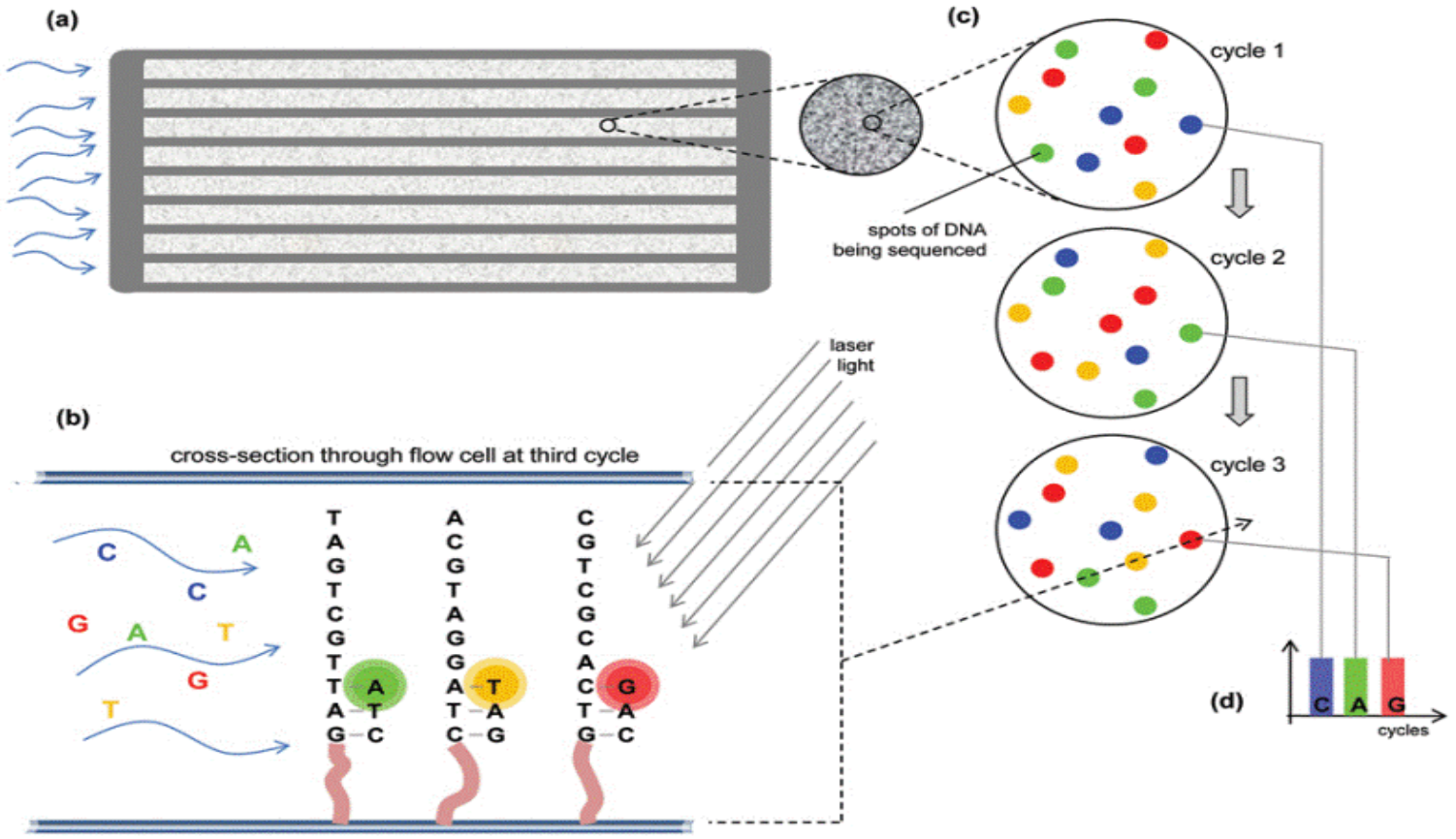
Sequencers

http://omicsmaps.com

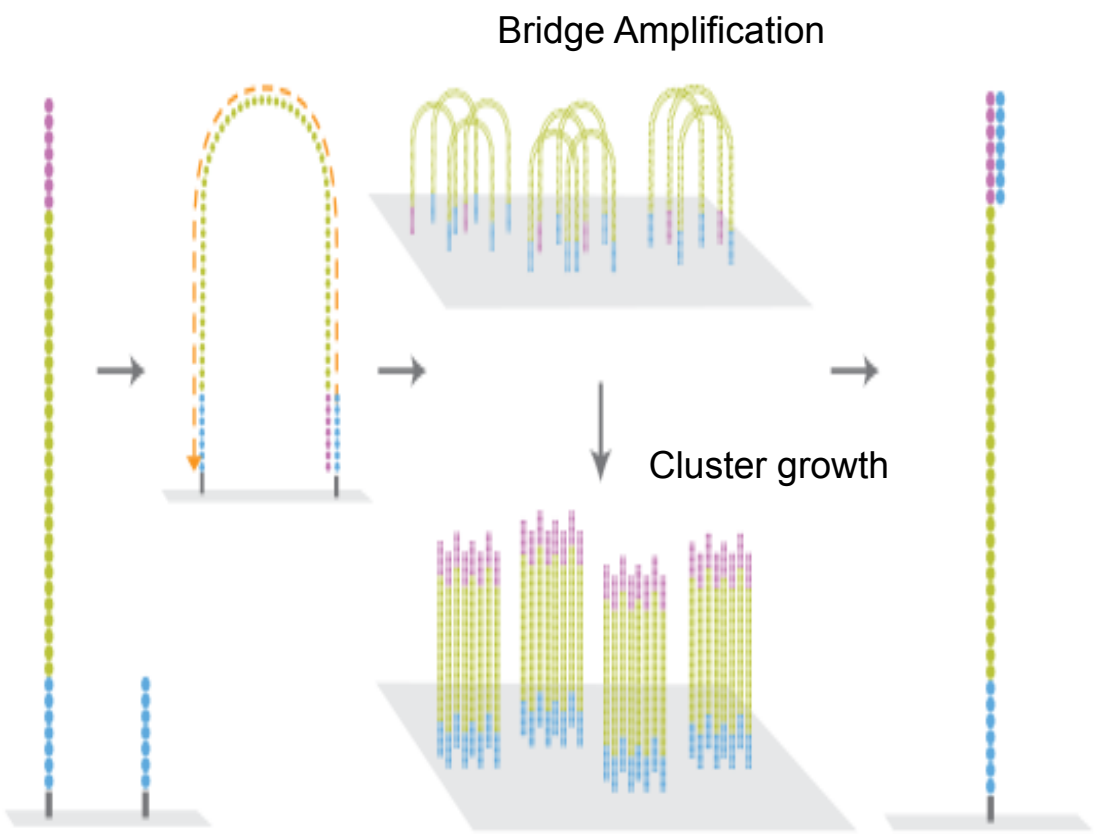


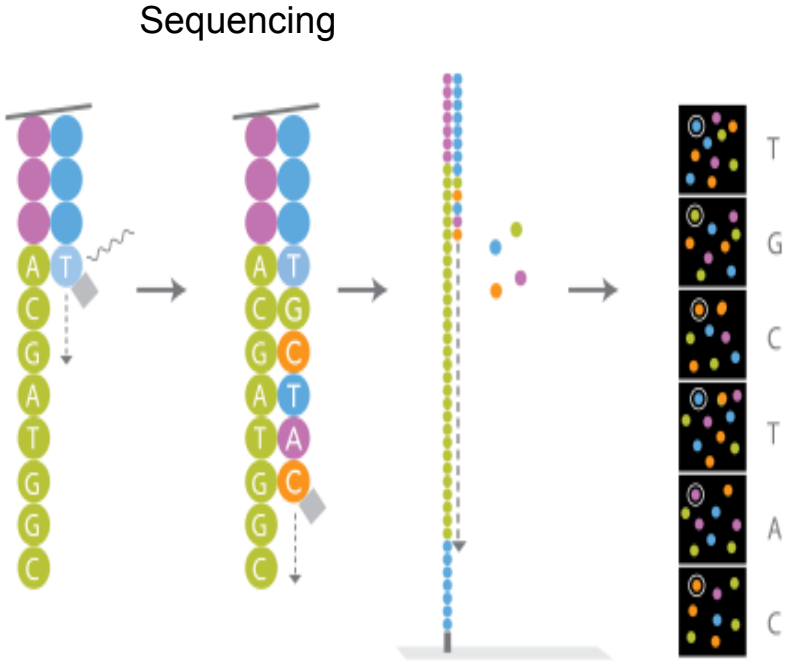


Illumina Flowcell



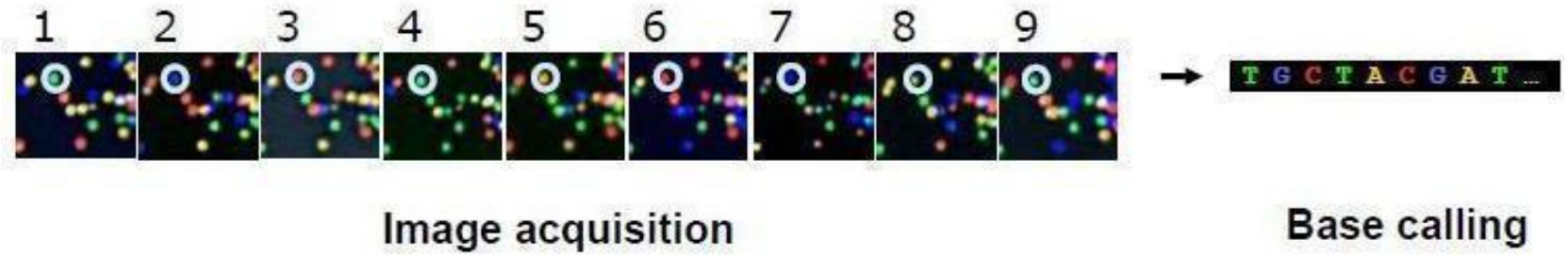
Sequencing





Incorporation of fluorescence,
reversibly terminated tagged nt

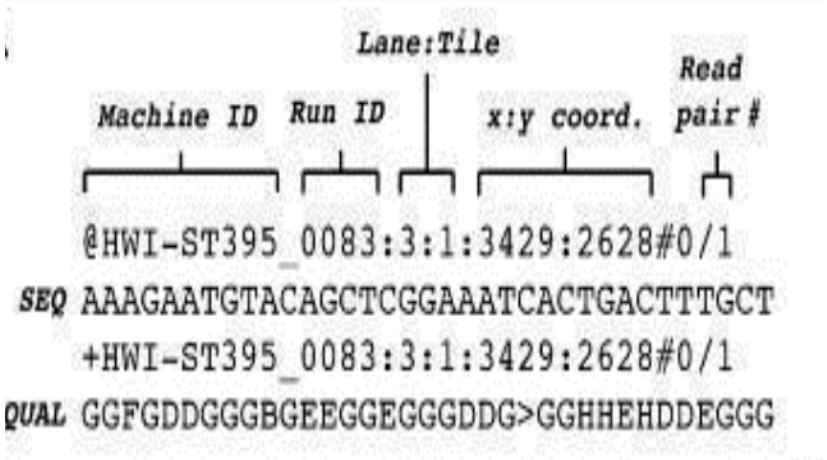
Sequencer Output



Quality Scores

Sequence Files

FASTQ formats



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.

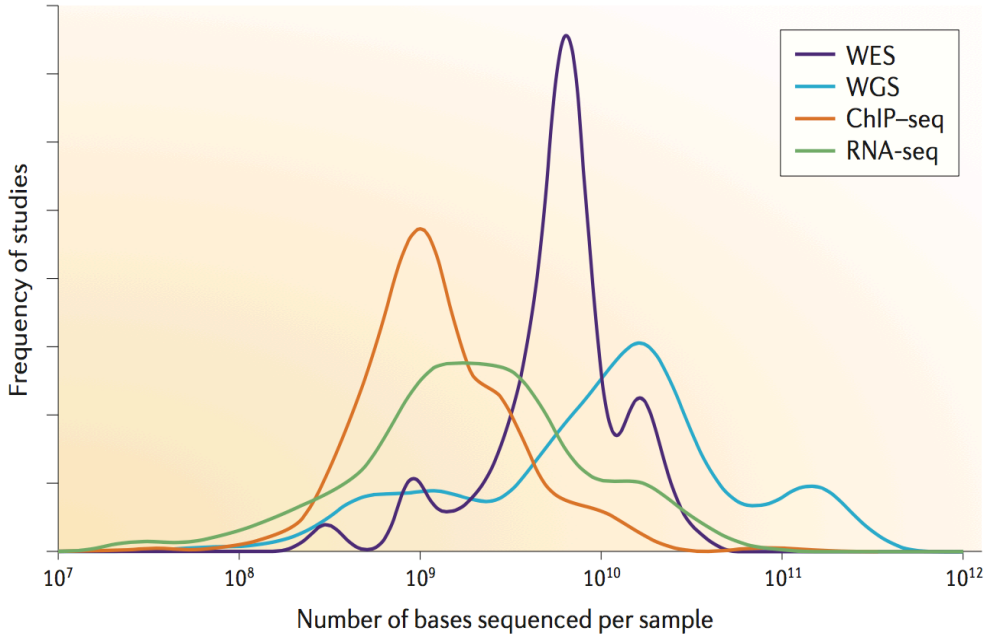
Line-4 encodes the quality scores (ASCII) for the sequence in Line 2.

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

The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. Nucleic Acids Res. 2010 Apr;38(6):1767-71.

Sequencing Depth and Coverage

Coverage: The average number of times that each nucleotide is expected to be sequenced given a certain number of reads of a given length and the assumption that the reads are randomly distributed across the genome.

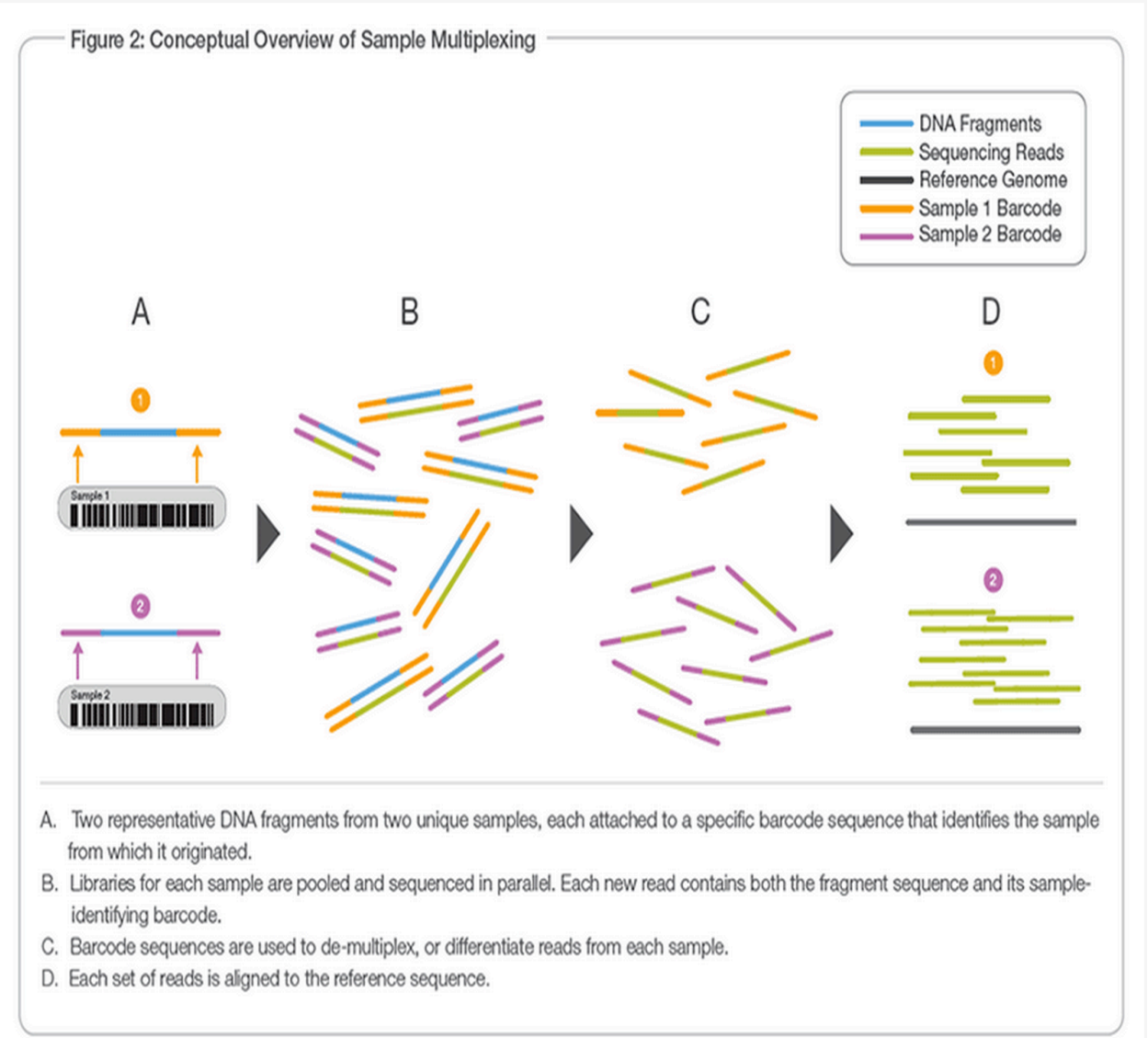


Depth: redundancy of coverage

Sims et al., 2014. Sequencing depth and coverage: key considerations in genomic analyses., Nat. Rev. Genet. 15(2), pp.121–132.

Multiplexing

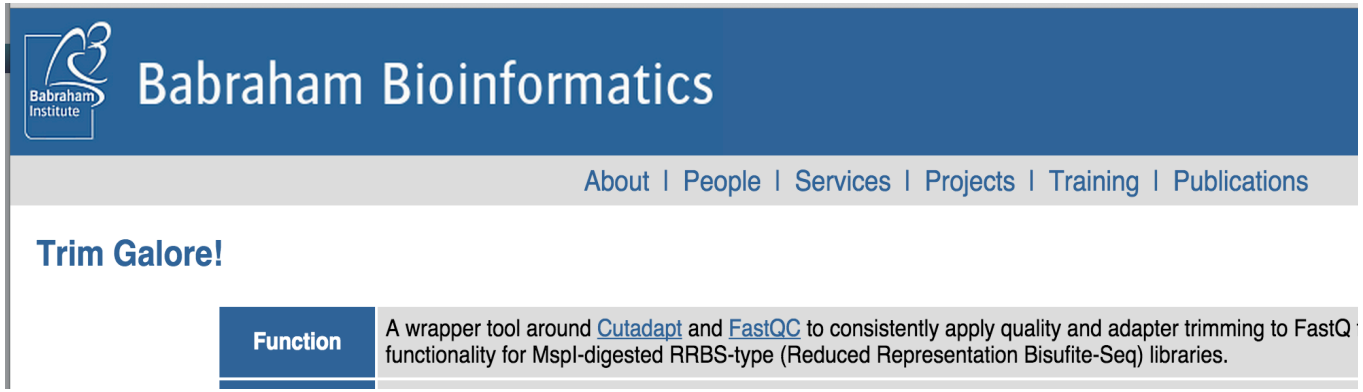
- Multiplexing gives the ability to sequence multiple samples at the same time.
- Useful when sequencing small genomes or specific genomic regions.
- Different barcode adaptors are ligated to different samples.
- Reads de-multiplexd after sequencing.



Quality control of short reads

- If samples were multiplexed on flow-cells, use barcodes to de-multiplex reads.
- Detect and trim adapters.
- Remove primers and other artefact sequences.
- Check for PCR duplicates.

Tools:



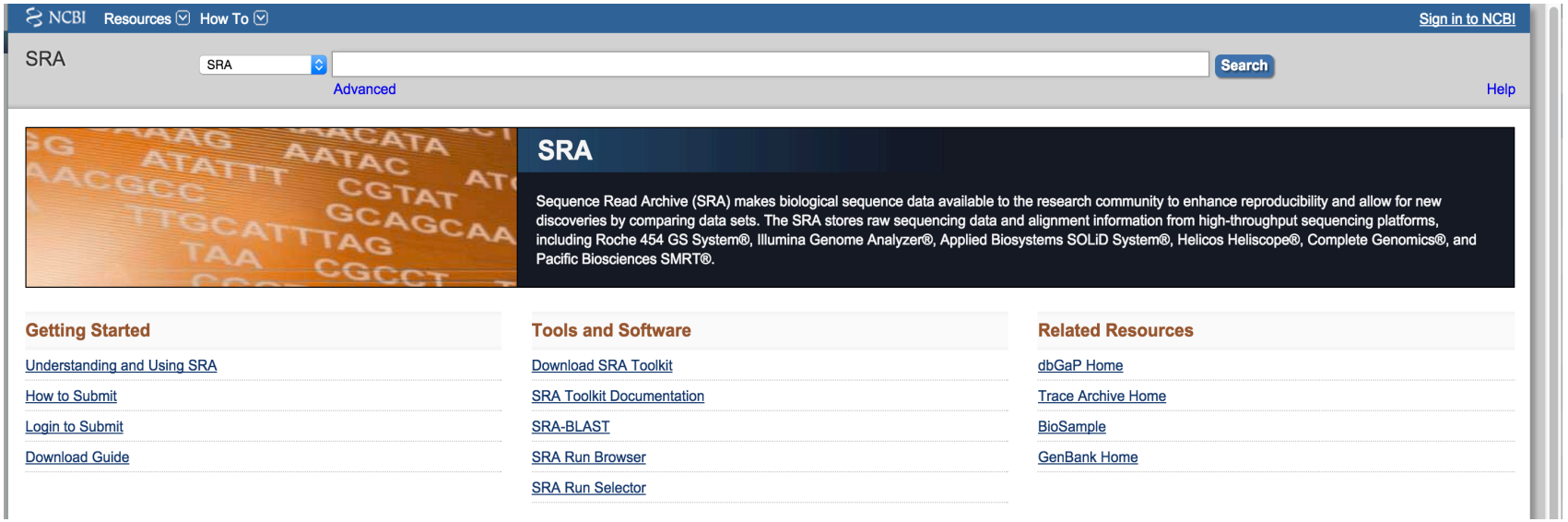
De-multiplexing: FASTX toolkit, QIIME, ea-utils

Artefact detection: **FASTQC,** NGSQC

Artefact removal: **CutAdapt, TrimGalore, ShortRead,** Useq, TagDust, FASTX toolkit, Trimmomatic

How to get external sequencing data via SRA tool-kit

- 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



Map to reference genome

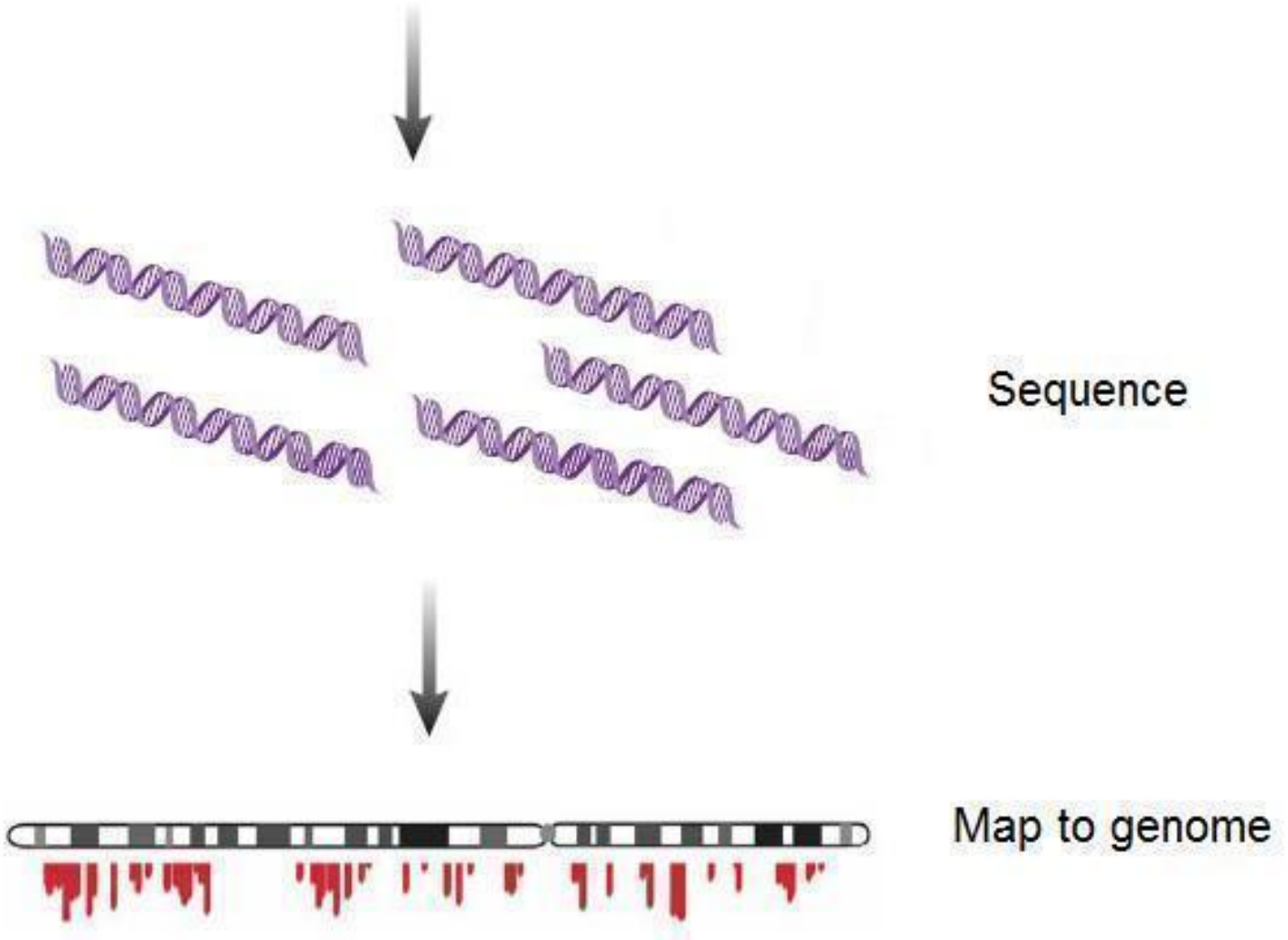


Table 4: Overall evaluation and comparison of multiple aligners.

Aligners

Aligners	Computational speed				Memory usage			Accuracy		
	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.

Alignment to Reference Genome

BWA & SAMtools example

make reference genome index:

```
bwa index -p hg19bwaidx -a bwtsw hg19.fa
```

align to hg19 reference:

```
bwa aln -t 4 hg19bwaidx sequence.fq.gz >  sequence.fq.sai
```

generate SAM file:

```
bwa samse hg19bwaidx sequence.fq.sai sequence.fq.gz > sequence.fq.sam
```

make BAM file:

```
samtools view –b sequence.fq.sam > sequence.fq.bam
```

sort:

```
samtools sort –o –O bam –T sorted sequence.fq.bam
```

index:

```
samtools index –b sequence.fq_sorted.bam sequence.fq_sorted.bai
```

BWA (Li & Durbin 2009)

SAMtools (Li et al., 2009)

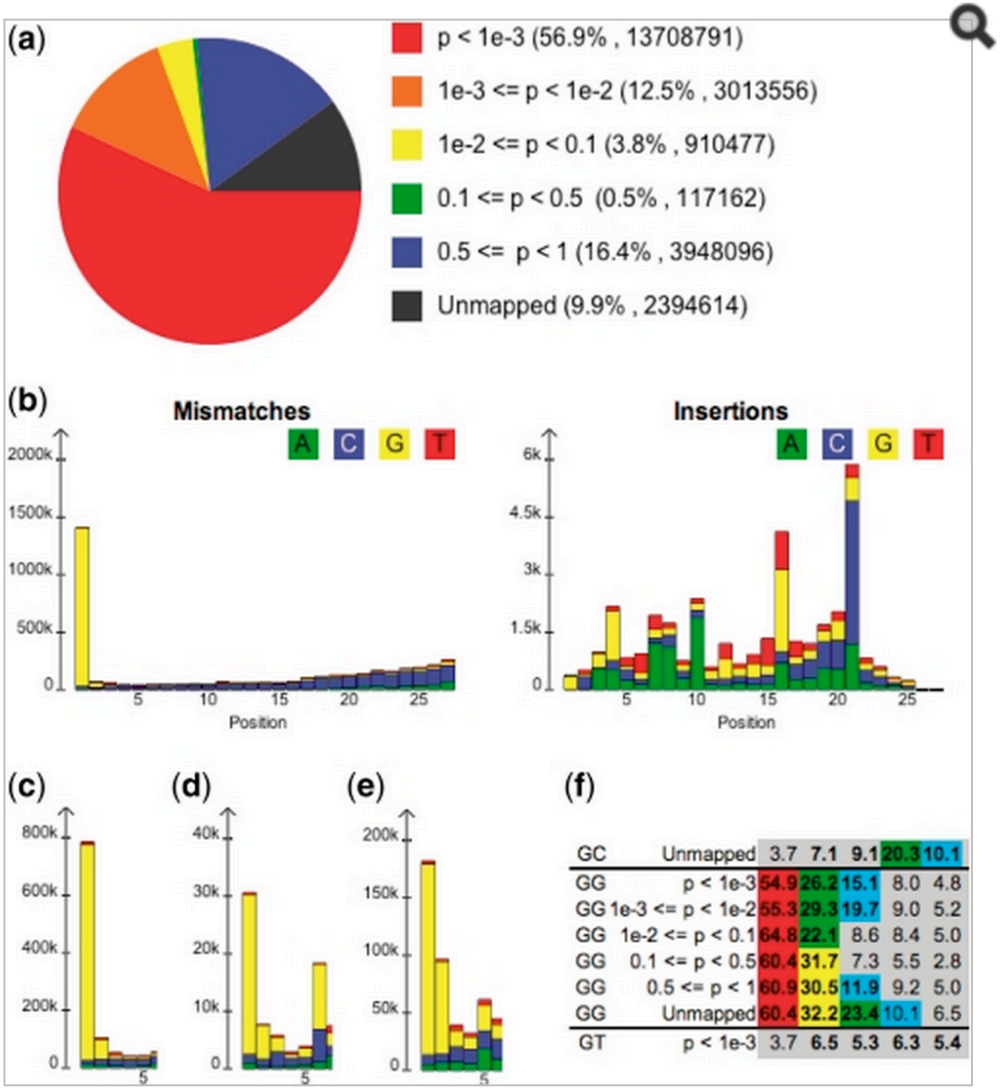
SAMstat for mapping QC

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

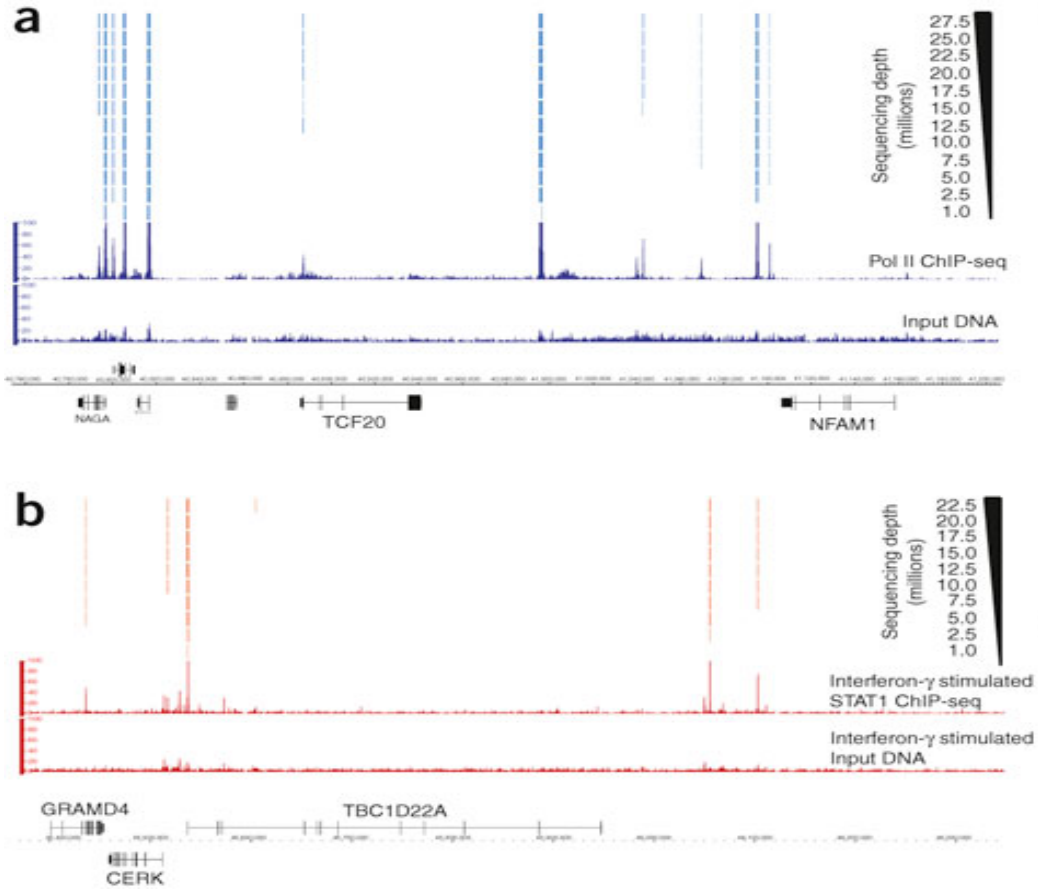
Overview of SAMstat output

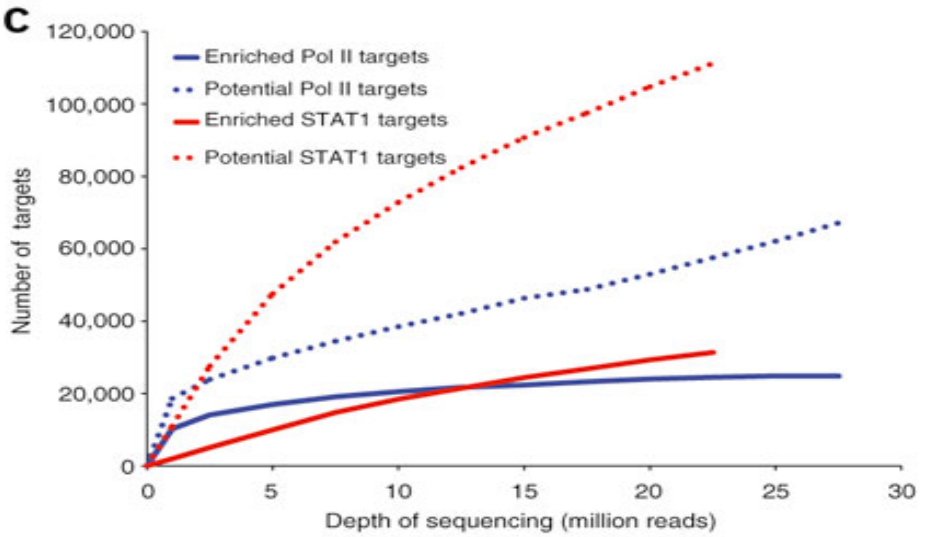
Reported statistics
Mapping rate[a]
Read length distribution
Nucleotide composition
Mean base quality at each read position
Overrepresented 10mers
Overrepresented dinucleotides along read
Mismatch, insertion and deletion profile[a]

[a]Only reported for SAM files.



Sequencing Depth





d

Number of replicas	Number of targets	True positives	False positives	False negatives	Sensitivity	Positive predictive value
1	20,902	18,833	2,069	5,906	0.761	0.901
2	20,733	19,257	1,476	5,482	0.778	0.929
3	20,328	19,126	1,202	5,613	0.773	0.941

Rozowsky 2009

- More prominent peaks are identified with fewer reads, versus weaker peaks that require greater depth
- Number of putative target regions continues to increase significantly as a function of sequencing depth
- Narrow Peaks: 15-20 million reads, Broad Peaks: 20-40 million reads
- https://genohub.com/recommended-sequencing-coverage-by-application/

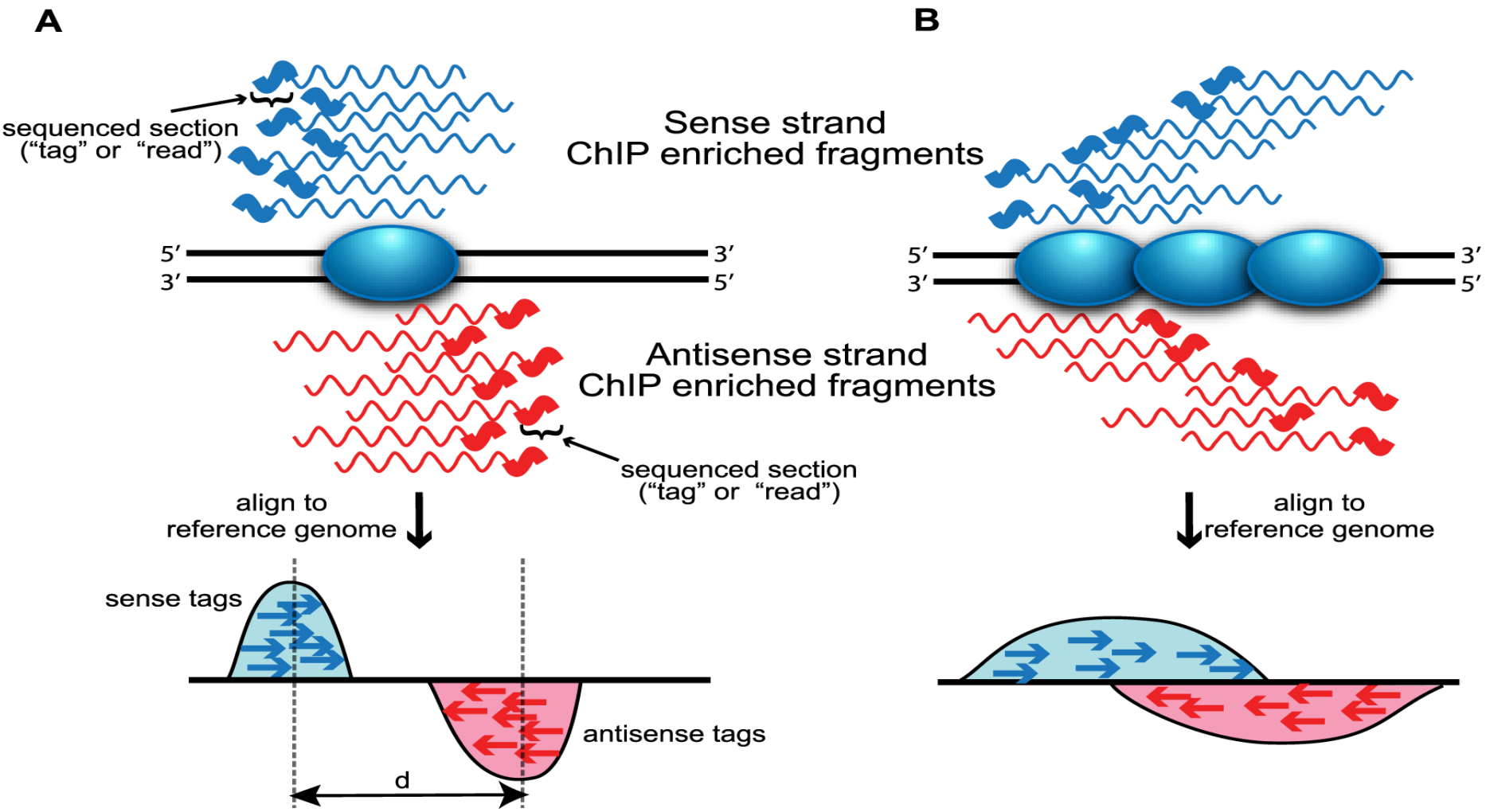
Mappability

		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%

*Calculated based on 30nt sequence tags
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.

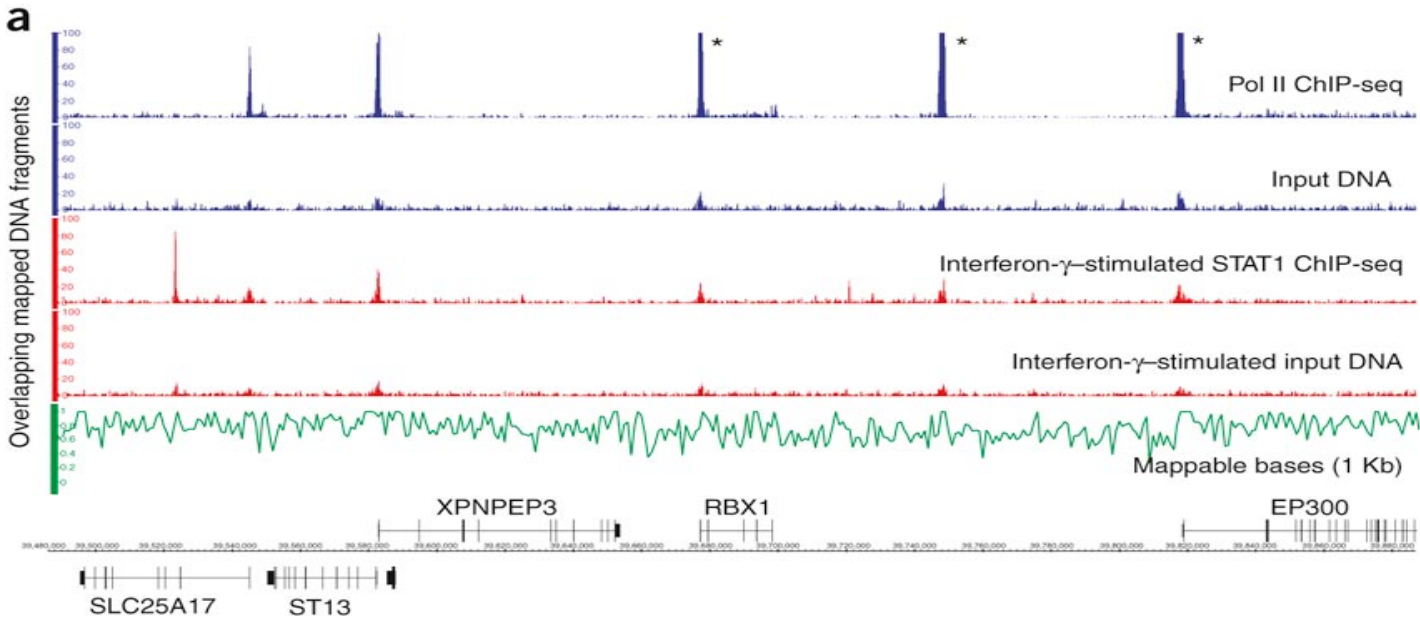
Strand dependent bimodality

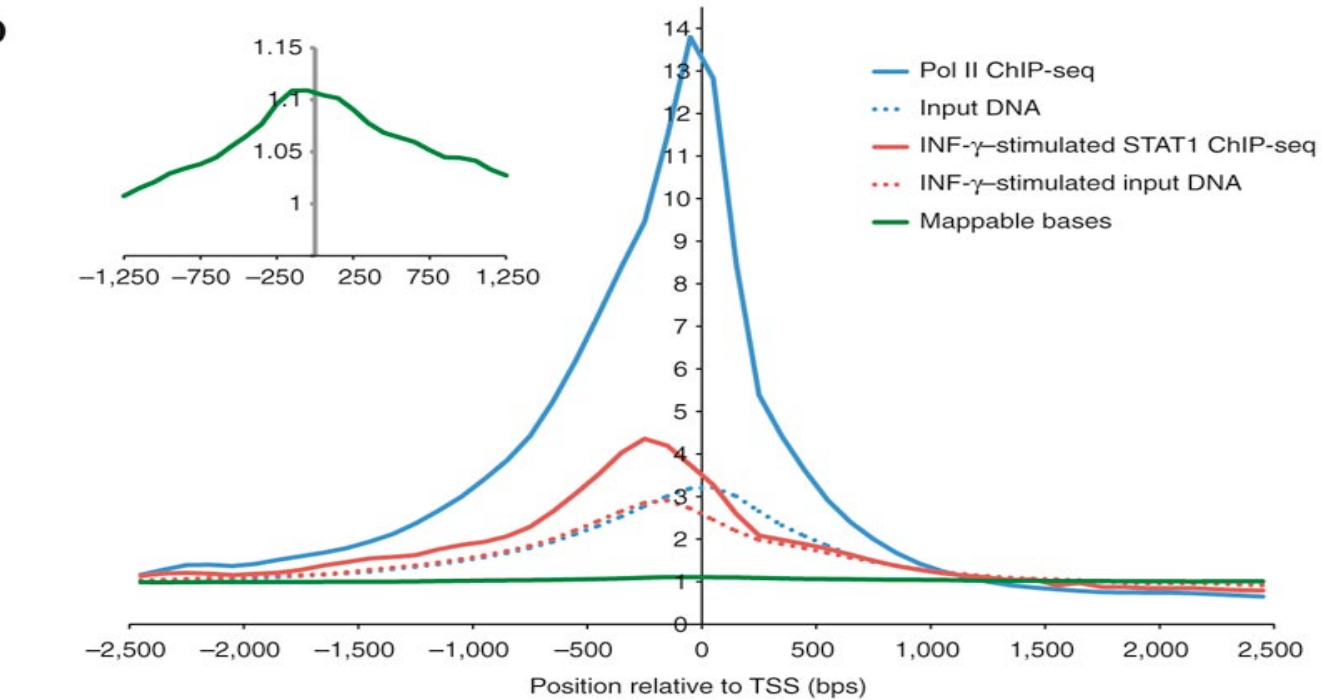


Wilbanks et al. 2010 PLOS One

Why we need control samples

- Open chromatin regions are more easily fragmented than closed regions.
- Uneven read distribution
- Repetitive sequences may appear to be enriched.
- Compare ChIPseq peak with same region in Input control.





Artefact removal 1

• After reads have been aligned to the reference genome, “blacklisted regions” are removed from **BAM** files before peak calling.

Blacklisted regions are genomic regions with anomalous, unstructured, high signal or read counts in NGS experiments, independent of cell type or experiment.

• The blacklisted regions typically appear uniquely mappable, so simple mappability filters do not remove them. These regions are often found at repetitive regions (Centromeres, Telomeres, Satellite repeats) and are troublesome for high throughput sequencing aligners and when computing genome wide correlations.

• These regions also confuse peak callers and result in spurious signal.

Artefact removal 2

- The *DAC Blacklisted Regions* aim to identify a comprehensive set of regions in the human genome that have anomalous, unstructured, high signal/read counts in NGS experiments, independent of cell line and type of experiment.

80 open chromatin tracks (DNase and FAIRE data-sets) and 20 ChIP-seq input/control tracks spanning ~60 human tissue types/cell lines in total used to identify these regions with signal artefacts. These regions tend to have a very high ratio of multi-mapping to uniquely mapping reads and high variance in mappability. The *DAC Blacklisted Regions* track was generated for the ENCODE project.

- The *Duke Excluded Regions* contains problematic regions for short sequence tag signal detection (such as satellites and rRNA genes).

- *Grey Lists* represent regions of high artefact signals that are specific to your cell-type or sample, and can be tuned depending on the stringency required.

Artefact removal 3

Resources:

Where to get Blacklist BED file:

- http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability

How they were generated:

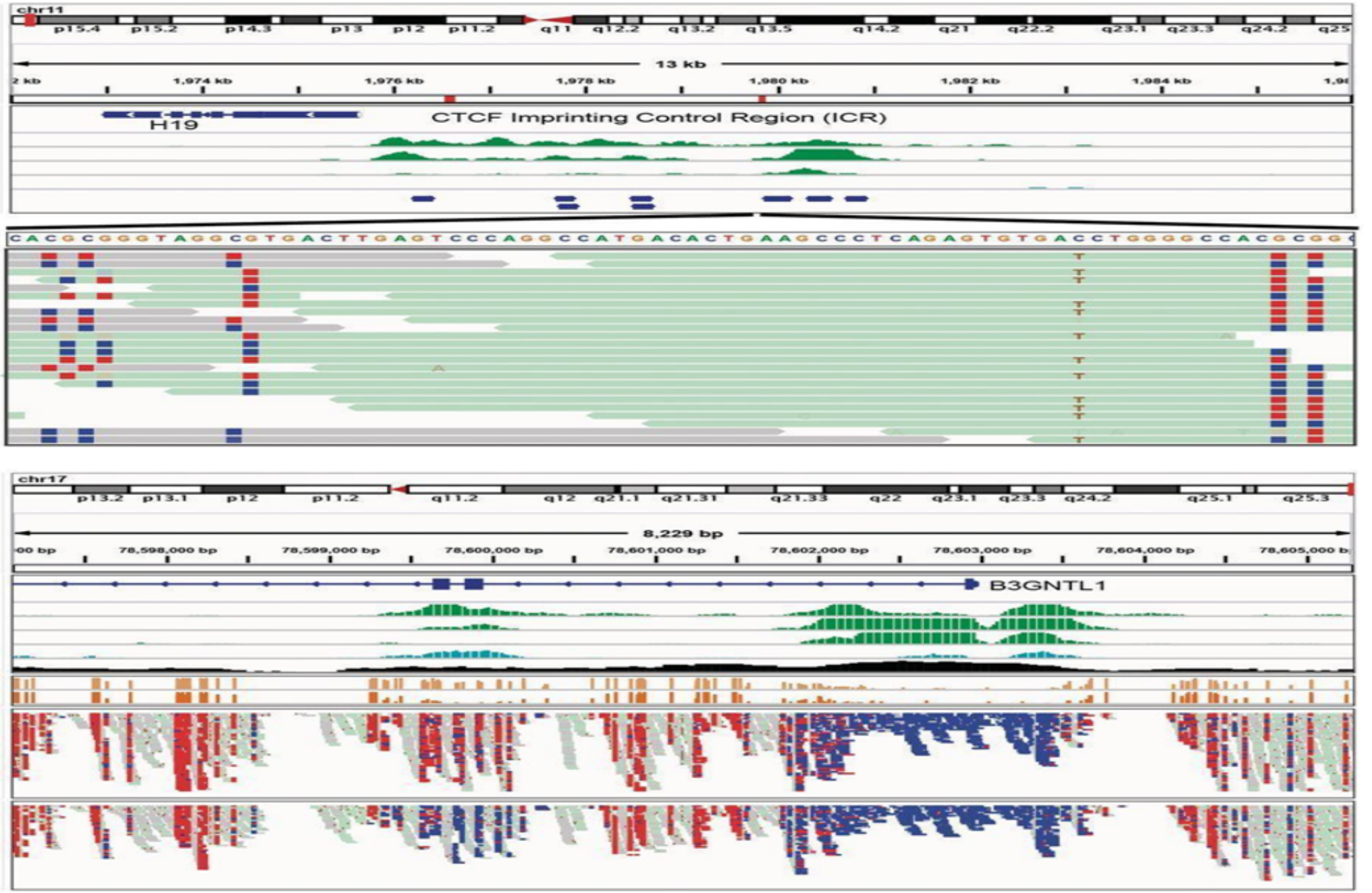
- http://www.broadinstitute.org/~anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf

ChIPseq Quality control :

- Carroll *et al.,* "Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data." *Front Genet.* 2014

- **GreyListChIP**
- **ChIPQC**

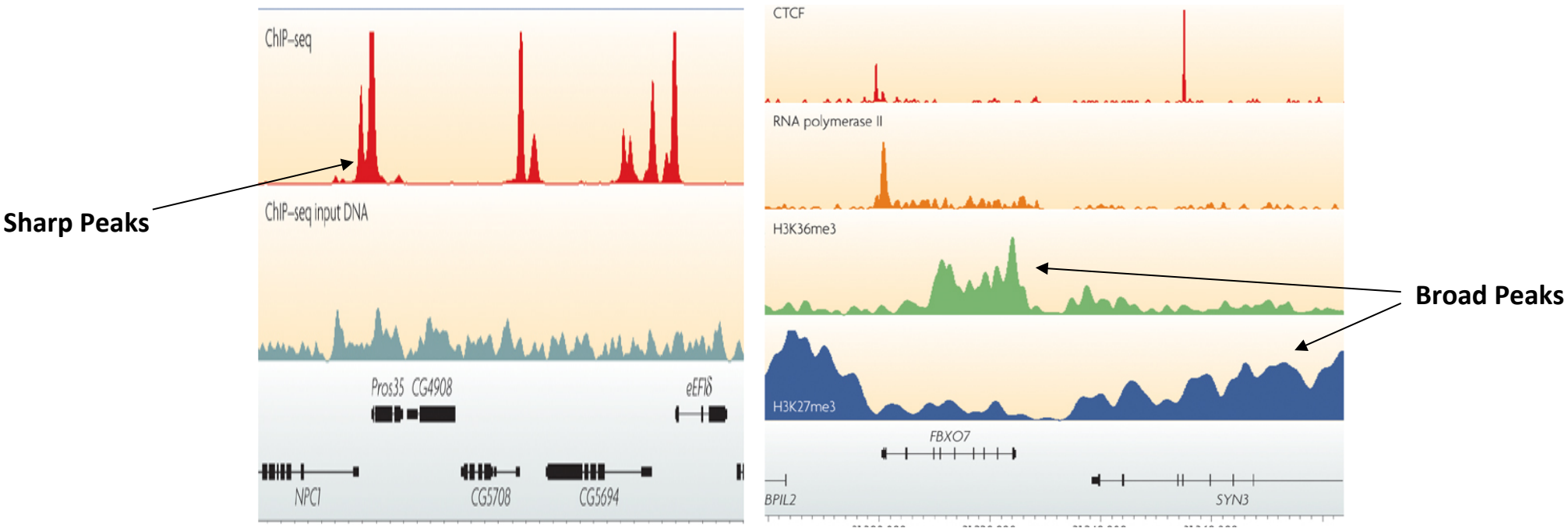
Visualizing binding sites and replicates



Integrated Genome Viewer (IGV)

Peak Calling

- Identifies TF binding sites
- Count based - Define regions. Count the number of reads falling into each region. When a region contains a statistically significant number of reads, call that region a peak.
- Shape based - Consider individual candidate binding sites. Model the spatial distribution of reads in surrounding regions, and call a peak when the read distribution conforms to the expected distribution near a binding site.



Peak Shapes

- Different ChIPseq applications have different peaks shapes.
- TF and regulatory element binding epigenetic marks are narrow, while histone modifications marking transcribed or repressed domains are broad.
- Most peak callers have been designed to find narrow peaks.
- The same TF may have different peak shapes reflecting different biological properties of binding.
- Replicates should have similar binding patterns.
- "Evaluation of algorithm performance in ChIP-seq peak detection." Wilbanks EG, Facciotti MT. PLoS One. 2010 Jul 8;5(7):e11471.

Peak Callers

- There are dozens of peaks callers. Some are good, others bad, none perfect!

Sharp TF & regulatory element associated epigenetic mark peaks:

- MACS v1.4.2 & MACS v2: model based analysis for ChIP-seq (Zhang *et al.*, 2008; Feng et al., 2011)
- **BayesPeak**: A Bayesian peak caller (Cairns *et al.*, 2011)
- **Jmosaics**: Joint analysis of multiple ChIP-seq datasets (Zeng et al., 2013)
- SPP (Kharchenko et al., 2008)
- T-PIC (Hover et al.)

Diffuse chromatin modification peaks:

- RSEG, SICER

ChIP-seq Quality Control