Introduction to ChIP-seq analysis

Shamith Samarajiwa

Integrative Systems Biomedicine Group

MRC Cancer Unit

University of Cambridge

Analysis of High-throughput sequencing data with BioConductor

1-3 June 2015

MRC Medical Research Council

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.

Origin 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

Science
AAAS

Genome-Wide Mapping of in Vivo Protein-DNA Interactions
David S. Johnson, *et al.*
Science **316**, 1497 (2007);
DOI: 10.1126/science.1141319

Genome-Wide Mapping of in Vivo Protein-DNA Interactions

David S. Johnson,[1]* Ali Mortazavi,[2]* Richard M. Myers,[1]† Barbara Wold[2,3]†

In vivo protein-DNA interactions connect each transcription factor with its direct targets to form a gene network scaffold. To map these protein-DNA interactions comprehensively across entire mammalian genomes, we developed a large-scale chromatin immunoprecipitation assay (ChIPSeq) based on direct ultrahigh-throughput DNA sequencing. This sequence census method was then used to map in vivo binding of the neuron-restrictive silencer factor (NRSF; also known as REST, for repressor element–1 silencing transcription factor) to 1946 locations in the human genome. The data display sharp resolution of binding position [±50 base pairs (bp)], which facilitated our finding motifs and allowed us to identify noncanonical NRSF-binding motifs. These ChIPSeq data also have high sensitivity and specificity [ROC (receiver operator characteristic) area ≥ 0.96] and statistical confidence ($P < 10^{-4}$), properties that were important for inferring new candidate interactions. These include key transcription factors in the gene network that regulates pancreatic islet cell development.

putational discovery of binding motifs feasible, this dictates the quality of regulatory site annotation relative to other gene anatomy landmarks, such as transcription start sites, enhancers, introns and exons, and conserved noncoding features (*2*). Finally, if high-quality protein-DNA interactome measurements can be performed routinely and at reasonable cost, it will open the way to detailed studies of interactome dynamics in response to specific signaling stimuli or genetic mutations. To address these issues, we turned to ultrahigh-throughput DNA sequencing to gain sampling power and applied size selection on immuno-enriched DNA to enhance positional resolution.

The ChIPSeq assay shown here differs from other large-scale ChIP methods such as ChIPArray, also called ChIPchip (*1*); ChIPSAGE (SACO) (*3*); or ChIPPet (*4*) in design, data produced, and cost. The design is simple (Fig.

Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing

Gordon Robertson[1], Martin Hirst[1], Matthew Bainbridge[1], Misha Bilenky[1], Yongjun Zhao[1], Thomas Zeng[1], Ghia Euskirchen[2], Bridget Bernier[1], Richard Varhol[1], Allen Delaney[1], Nina Thiessen[1], Obi L Griffith[1], Ann He[1], Marco Marra[1], Michael Snyder[2] & Steven Jones[1]

We developed a method, ChIP-sequencing (ChIP-seq), combining chromatin immunoprecipitation (ChIP) and massively parallel sequencing to identify mammalian DNA sequences bound by transcription factors *in vivo*. We used ChIP-seq to map STAT1 targets in interferon-γ (IFN-γ)–stimulated and unstimulated human HeLa S3 cells, and compared the method's performance to ChIP-PCR and to ChIP-chip for four chromosomes. By ChIP-seq, using 15.1 and 12.9 million uniquely mapped sequence reads, and an estimated false discovery rate of less than 0.001, we identified 41,582 and 11,004 putative STAT1-binding regions in stimulated and unstimulated cells, respectively. Of the 34 loci known to contain STAT1 interferon-responsive binding sites, ChIP-seq found 24 (71%). ChIP-seq targets were enriched in sequences similar to known STAT1 binding motifs. Comparisons with two ChIP-PCR data sets suggested that ChIP-seq sensitivity was between 70% and 92% and specificity was at least 95%.

single-end tags (SETs), which are simpler to prepare than PETs, may be effective for profiling mammalian protein-DNA interactions. Thus we appraised the 1G system as a platform for ChIP with tag sequencing.

As a test system, we selected the mammalian transcription factor STAT1, whose cellular biology is relatively well characterized, and whose use permits a comparison of unstimulated and stimulated cellular states[12–16]. In both resting and stimulated cells, STAT proteins shuttle continuously between cytoplasm and nucleus[12,13,15]. Signaling by several cytokines, growth factors and hormone receptors leads to activation of receptor-associated JAK family kinases that phosphorylate a substantial fraction of cytoplasmic STAT1 proteins[12,15,17–20]. Phosphorylated STAT1 forms specific homodimers, heterodimers and heterotrimers that bind DNA with high affinity, and thus accumulate in the nucleus. STAT1 complexes activate or repress transcription primarily by the homodimer binding to IFN-γ activation site (GAS) elements, but also to interferon-stimulated response elements (ISREs)[16,17]. The regulatory activity of STAT1

Resource

Cell

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski,[1,3] Suresh Cuddapah,[1,3] Kairong Cui,[1,3] Tae-Young Roh,[1,3] Dustin E. Schones,[1,3] Zhibin Wang,[1,3] Gang Wei,[1,3] Iouri Chepelev,[2] and Keji Zhao[1,*]

[1] Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA

[2] Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

[3] These authors contributed equally to this work and are listed alphabetically.

*Correspondence: zhaok@nhlbi.nih.gov

DOI 10.1016/j.cell.2007.05.009

SUMMARY

Histone modifications are implicated in influencing gene expression. We have generated high-resolution maps for the genome-wide distribution of 20 histone lysine and arginine methylations as well as histone variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF across the human genome using the Solexa 1G sequencing technology. Typical patterns of histone methylations exhibited at promoters, insulators, enhancers, and transcribed regions are identified. The monomethylations of H3K27, H3K9, H4K20, H3K79, and H2BK5 are all linked to gene activation, whereas trimethylations of H3K27, H3K9, and H3K79 are linked to repression. H2A.Z associates with functional regulatory elements, and CTCF marks boundaries of histone methylation domains. Chromosome banding patterns are correlated with unique patterns of histone modifications. Chromosome breakpoints detected in T cell cancers frequently reside in chromatin regions associated with H3K4 methylations. Our data provide new insights into the function of histone methylation and chromatin organization in genome function.

biological processes. Among the various modifications, histone methylations at lysine and arginine residues are relatively stable and are therefore considered potential marks for carrying the epigenetic information that is stable through cell divisions. Indeed, enzymes that catalyze the methylation reaction have been implicated in playing critical roles in development and pathological processes.

Remarkable progress has been made during the past few years in the characterization of histone modifications on a genome-wide scale. The main driving force has been the development and improvement of the "ChIP-on-chip" technique by combining chromatin immunoprecipitation (ChIP) and DNA-microarray analysis (chip). With almost complete coverage of the yeast genome on DNA microarrays, its histone modification patterns have been extensively studied. The general picture emerging from these studies is that promoter regions of active genes have reduced nucleosome occupancy and elevated histone acetylation (Bernstein et al., 2002, 2004; Lee et al., 2004; Liu et al., 2005; Pokholok et al., 2005; Sekinger et al., 2005; Yuan et al., 2005). High levels of H3K4me1, H3K4me2, and H3K4me3 are detected surrounding transcription start sites (TSSs), whereas H3K36me3 peaks near the 3′ end of genes.

Significant progress has also been made in characterizing global levels of histone modifications in mammals. Several large-scale studies have revealed interesting insights into the complex relationship between gene expression and histone modifications. Generally, high levels of histone acetylation and H3K4 methylation are detected in promoter regions of active genes (Bernstein et al., 2005;

Vol 448|2 August 2007|doi:10.1038/nature06008 nature

ARTICLES

Genome-wide maps of chromatin state in pluripotent and lineage-committed cells

Tarjei S. Mikkelsen[1,2], Manching Ku[1,4], David B. Jaffe[1], Biju Issac[1,4], Erez Lieberman[1,2], Georgia Giannoukos[1], Pablo Alvarez[1], William Brockman[1], Tae-Kyung Kim[5], Richard P. Koche[1,2,4], William Lee[1], Eric Mendenhall[1,4], Aisling O'Donovan[4], Aviva Presser[1], Carsten Russ[1], Xiaohui Xie[1], Alexander Meissner[3], Marius Wernig[3], Rudolf Jaenisch[3], Chad Nusbaum[1], Eric S. Lander[1,3]* & Bradley E. Bernstein[1,4,6]*

We report the application of single-molecule-based sequencing technology for high-throughput profiling of histone modifications in mammalian cells. By obtaining over four billion bases of sequence from chromatin immunoprecipitated DNA, we generated genome-wide chromatin-state maps of mouse embryonic stem cells, neural progenitor cells and embryonic fibroblasts. We find that lysine 4 and lysine 27 trimethylation effectively discriminates genes that are expressed, poised for expression, or stably repressed, and therefore reflect cell state and lineage potential. Lysine 36 trimethylation marks primary coding and non-coding transcripts, facilitating gene annotation. Trimethylation of lysine 9 and lysine 20 is detected at satellite, telomeric and active long-terminal repeats, and can spread into proximal unique sequences. Lysine 4 and lysine 9 trimethylation marks imprinting control regions. Finally, we show that chromatin state can be read in an allele-specific manner by using single nucleotide polymorphisms. This study provides a framework for the application of comprehensive chromatin profiling towards characterization of diverse mammalian cell populations.

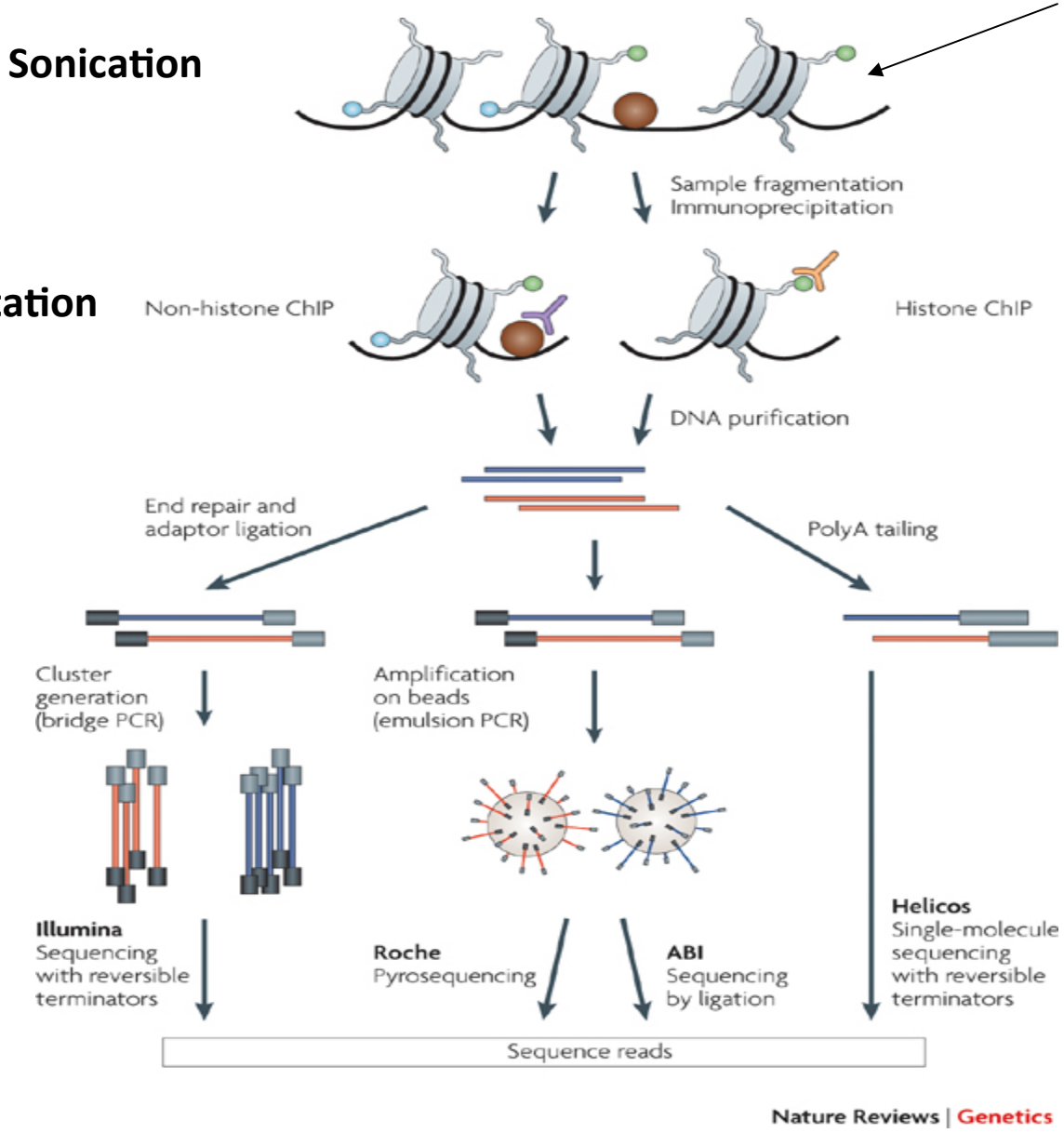
ChIP-seq methodology

Cross-link cells

Isolate genomic DNA

Sonication

Immuno-precipitation



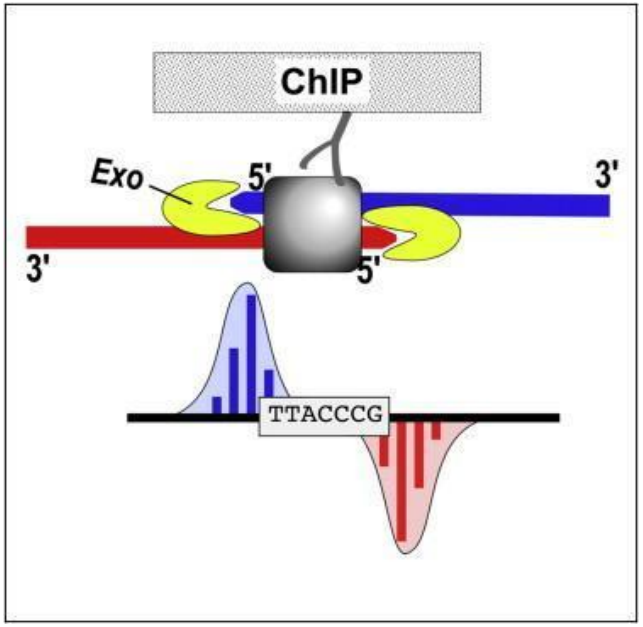
Park 2009 *Nat. Rev Genet.*

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



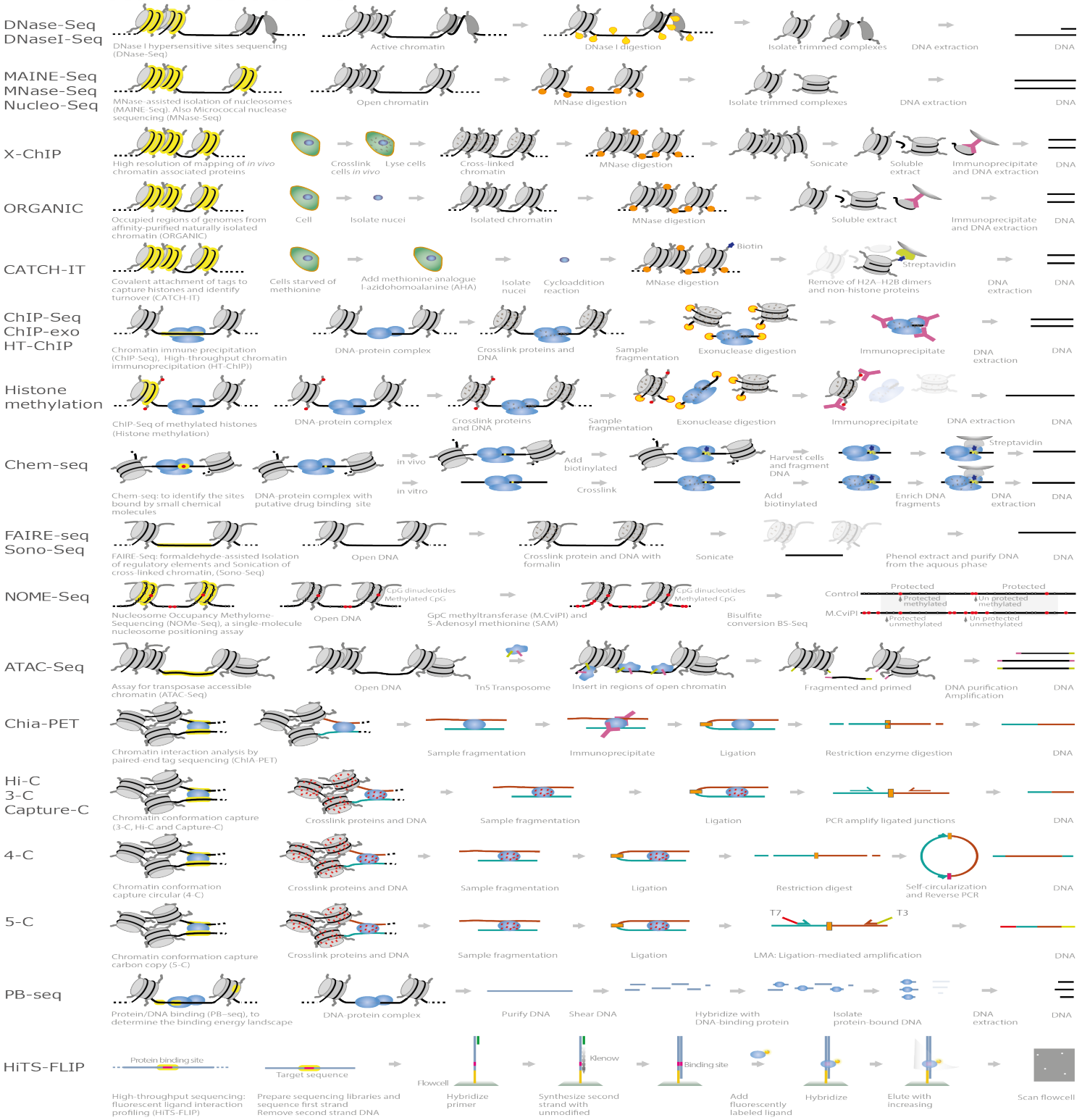
• CLIP-seq / HITS-CLIP : cross-linking immunoprecipitation high throughput sequencing

• 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.” *Advances in Statistical Bioinformatics*, 2013.
- Carroll TS *et al.*, “Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data.” *Front Genet.* 2014
- Bailey T *et al.*, "Practical guidelines for the comprehensive analysis of ChIP-seq data.” *PLoS Comput Biol.* 2013.

These guidelines address :

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

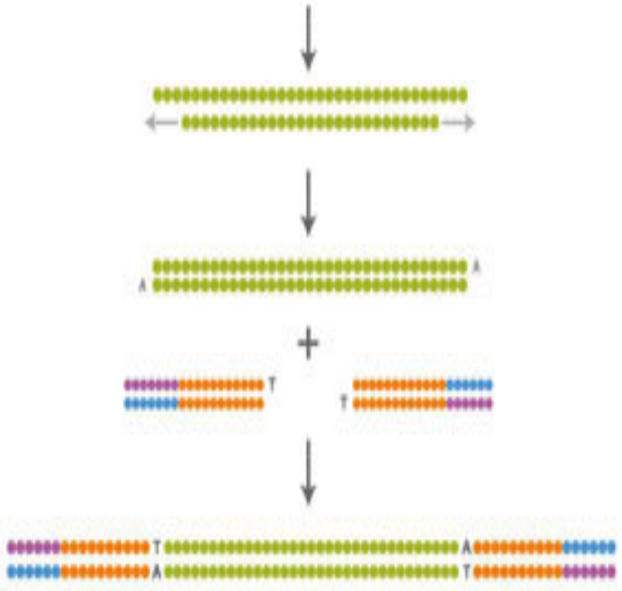
ChIP-seq workflow overview I

- 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



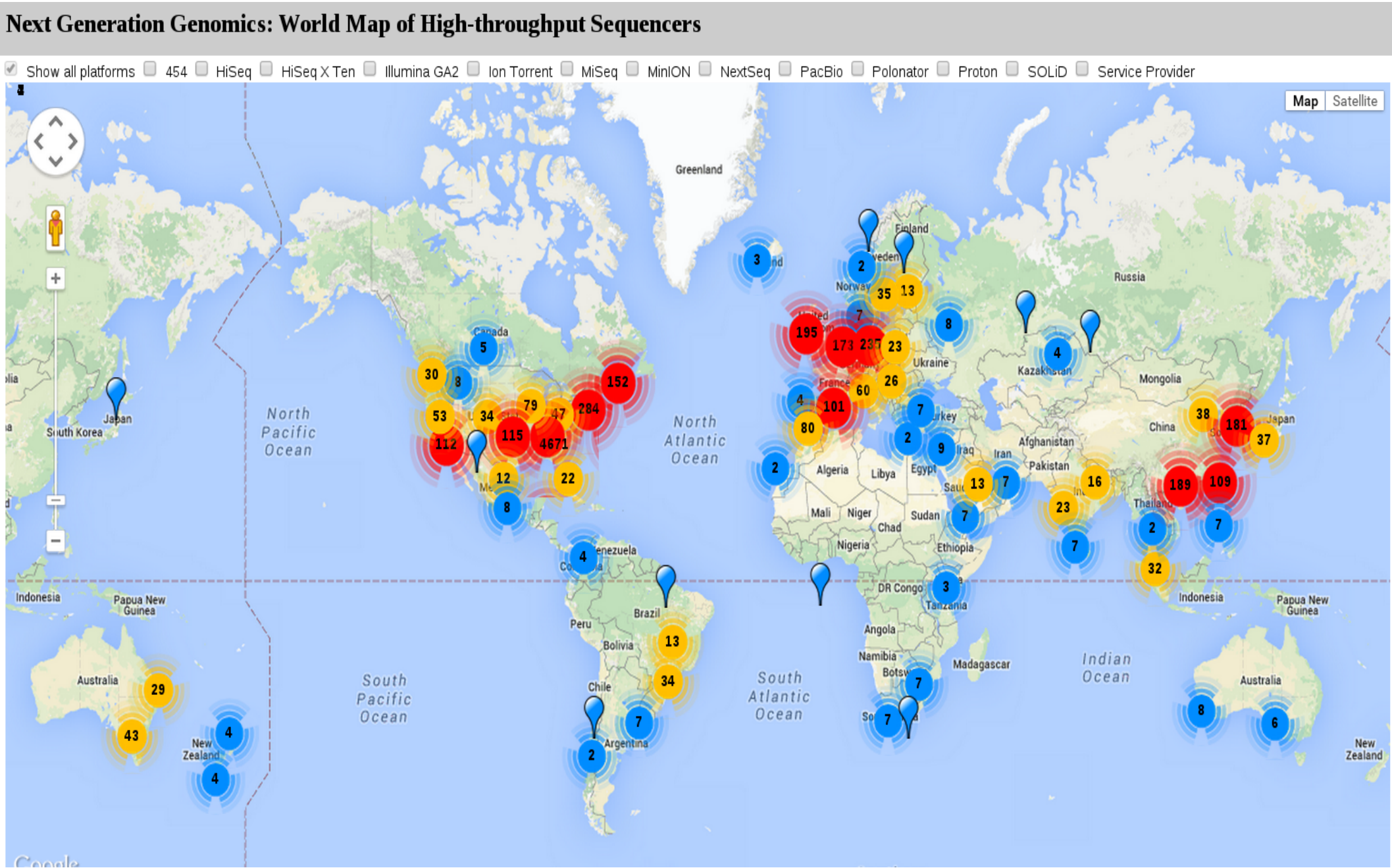
Library Preparation

Cluster Generation

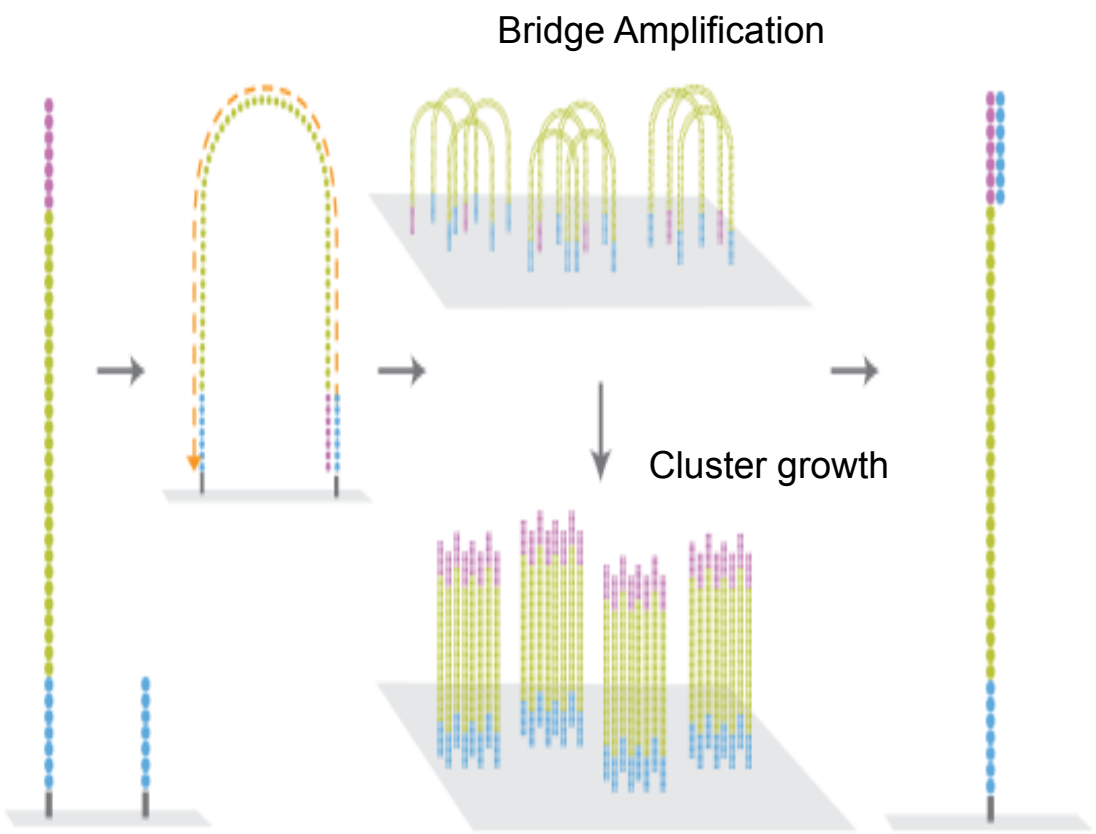
Sequencing by Synthesis

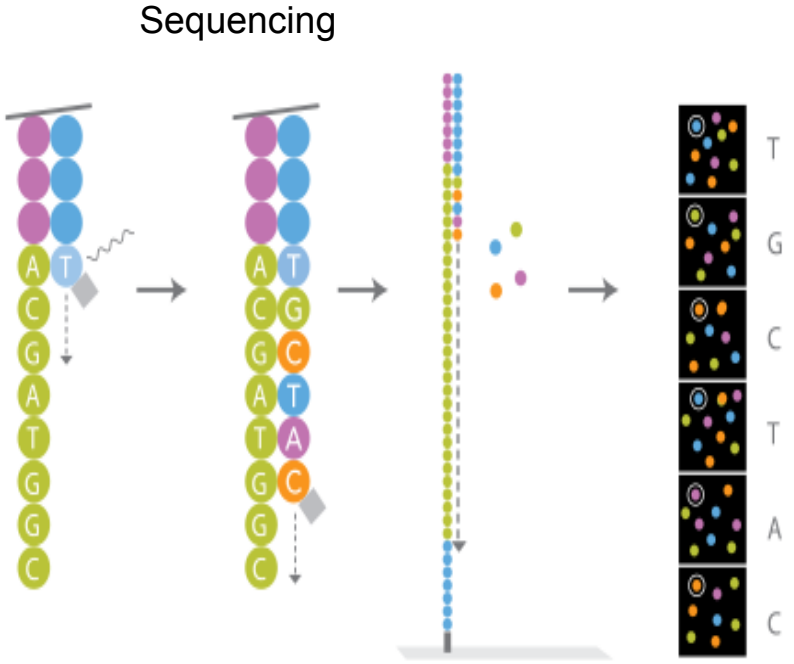
Sequencers

http://omicsmaps.com



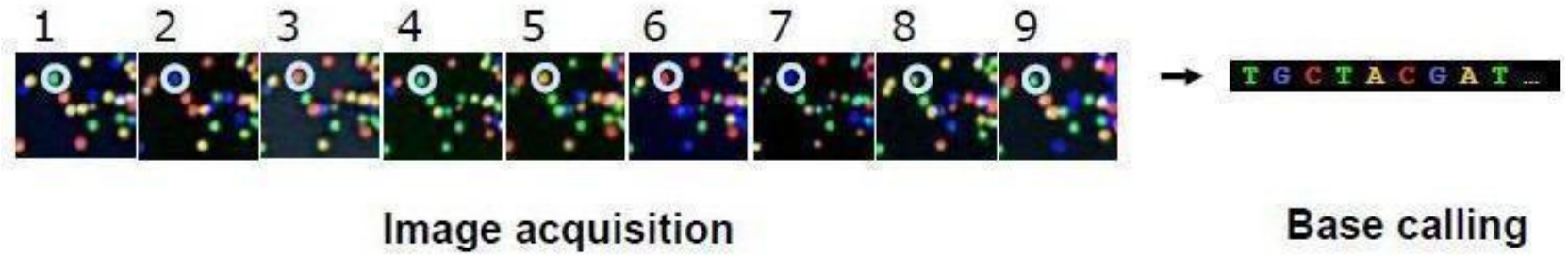
Sequencing





Incorporation of fluorescence,
reversibly terminated tagged nt

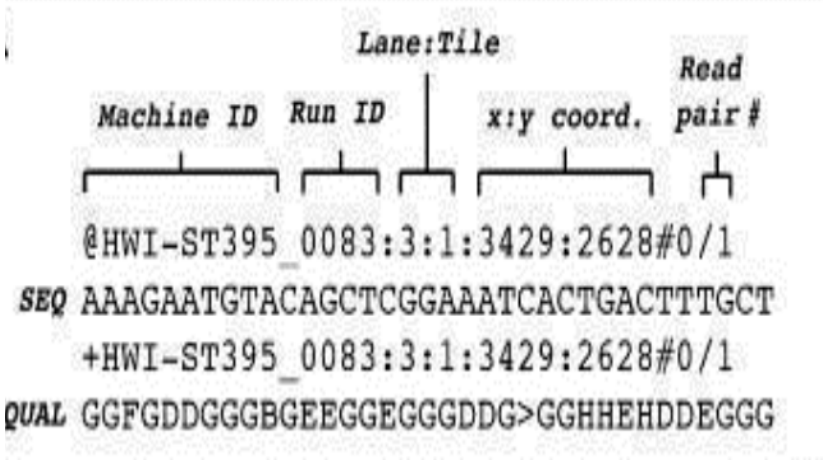
Sequencer Output



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.

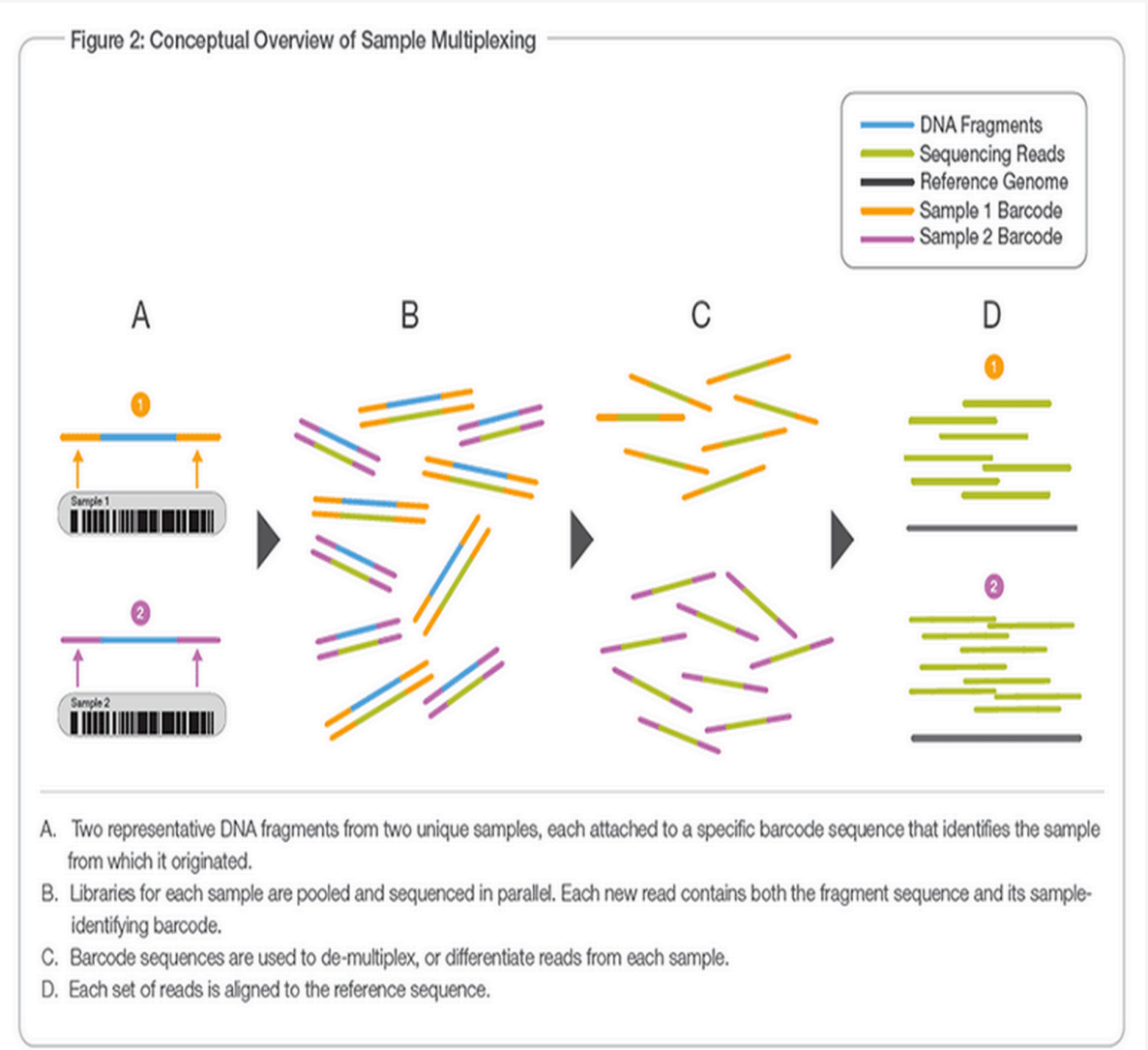
Line-4 encodes the quality scores (ASCII) for the sequence in Line 2.

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.

Multiplexing

- Number of reads per run continue to increase.
- Multiplexing gives the ability to sequence multiple samples at the same.
- Useful when sequencing small genomes or specific genomic regions.
- Different barcode adaptors are ligated to different samples



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

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 SRA format
- 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

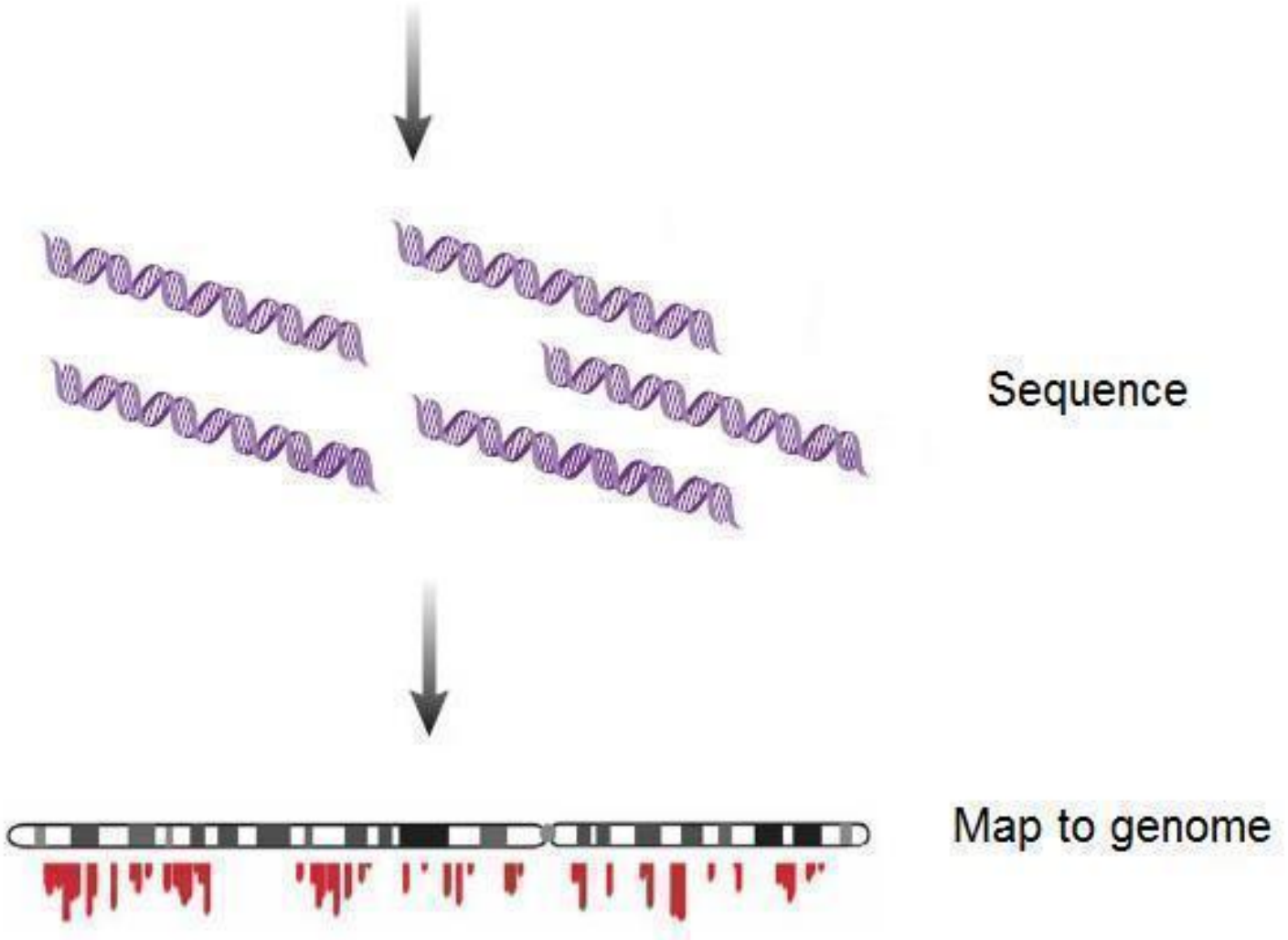


Table 4: Overall evaluation and comparison of multiple aligners.

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
```

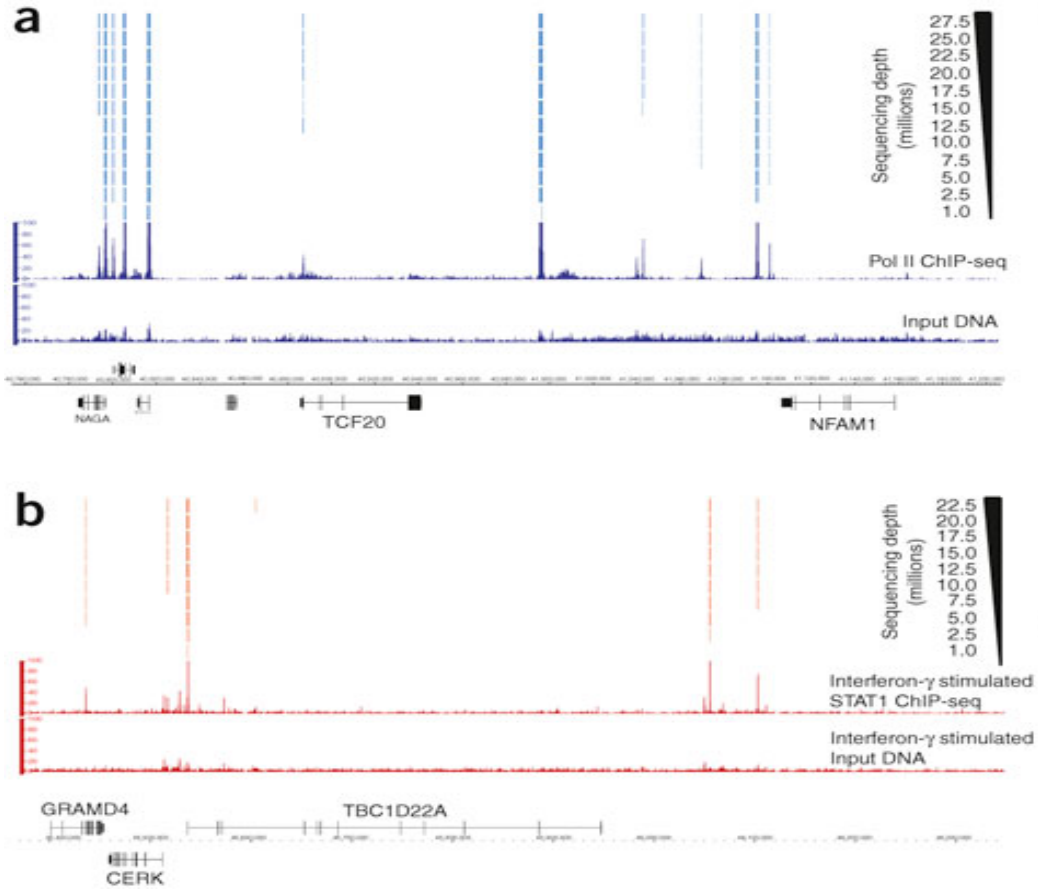
index:

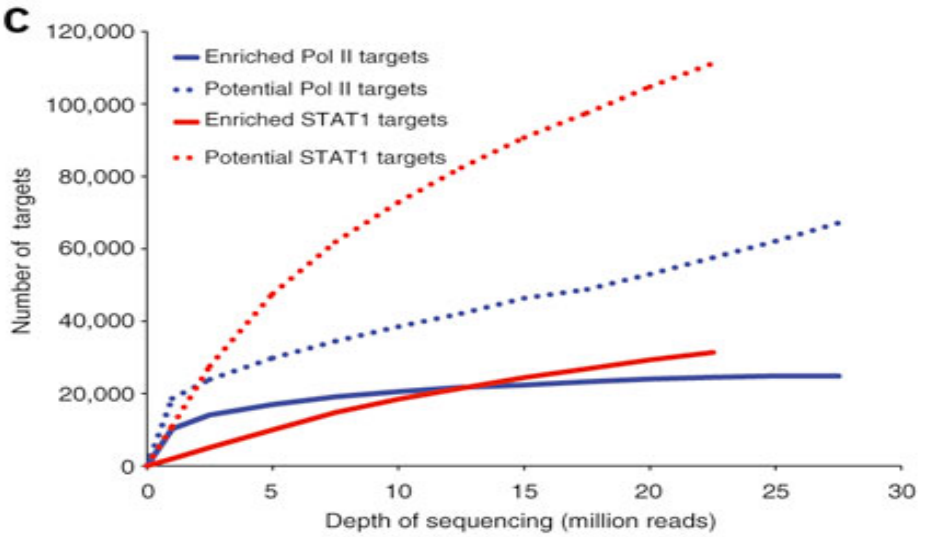
```
samtools index –b sequence.fq_sorted.bam sequence.fq_sorted.bai
```

BWA (Li & Durbin 2009)

SAMtools (Li et al., 2009)

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/

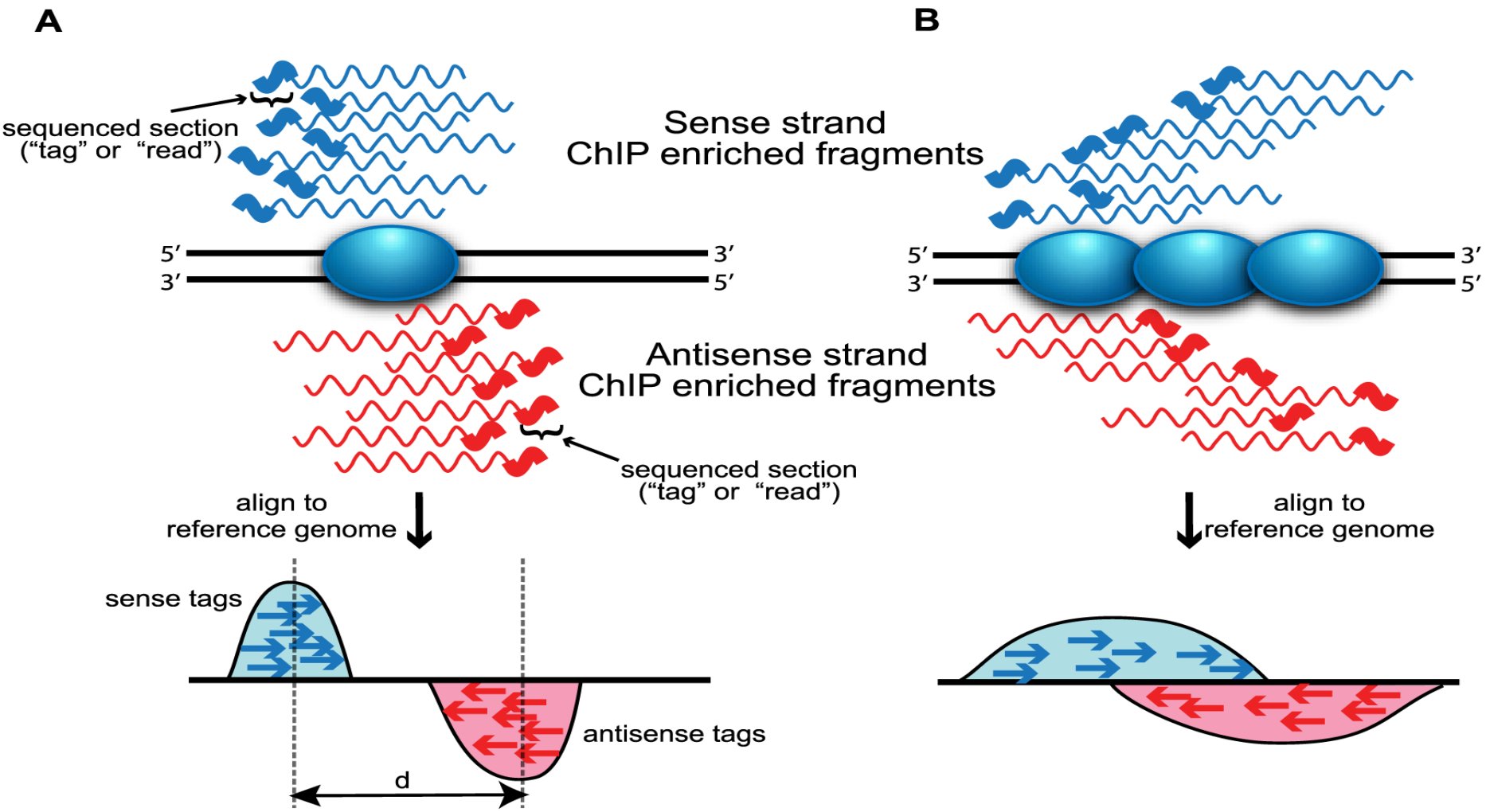
Mappability

		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.

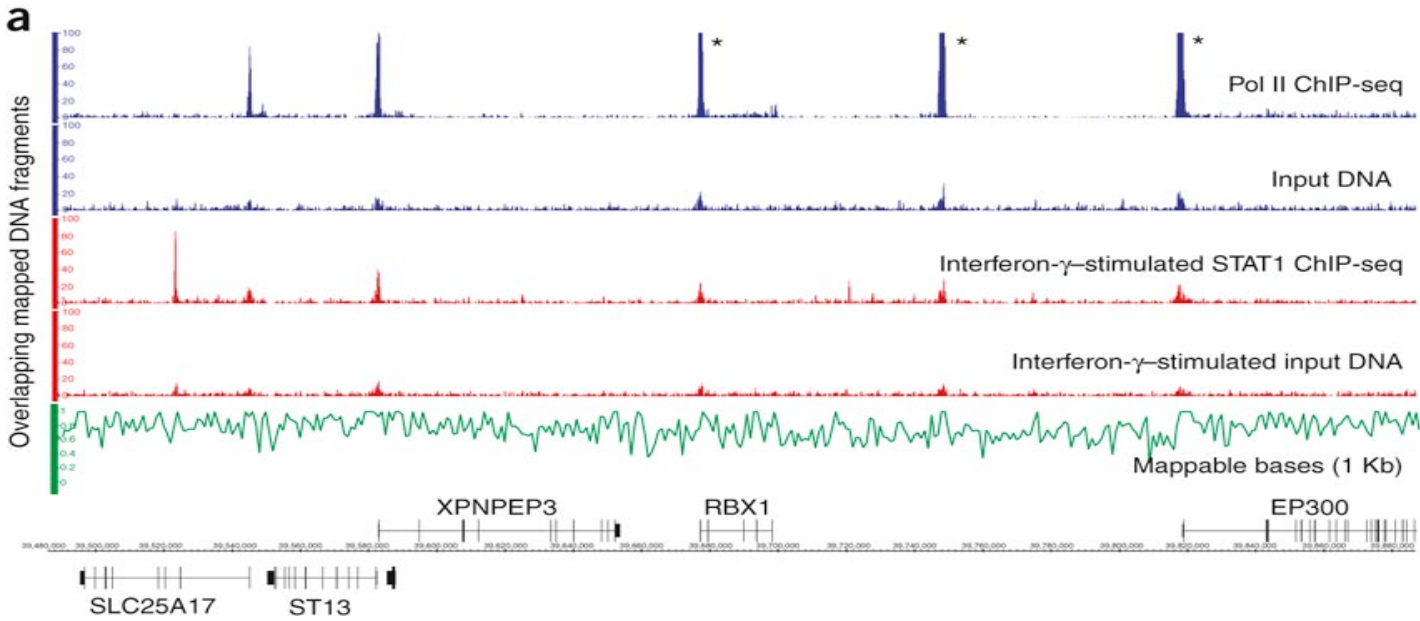
Strand dependent bimodality

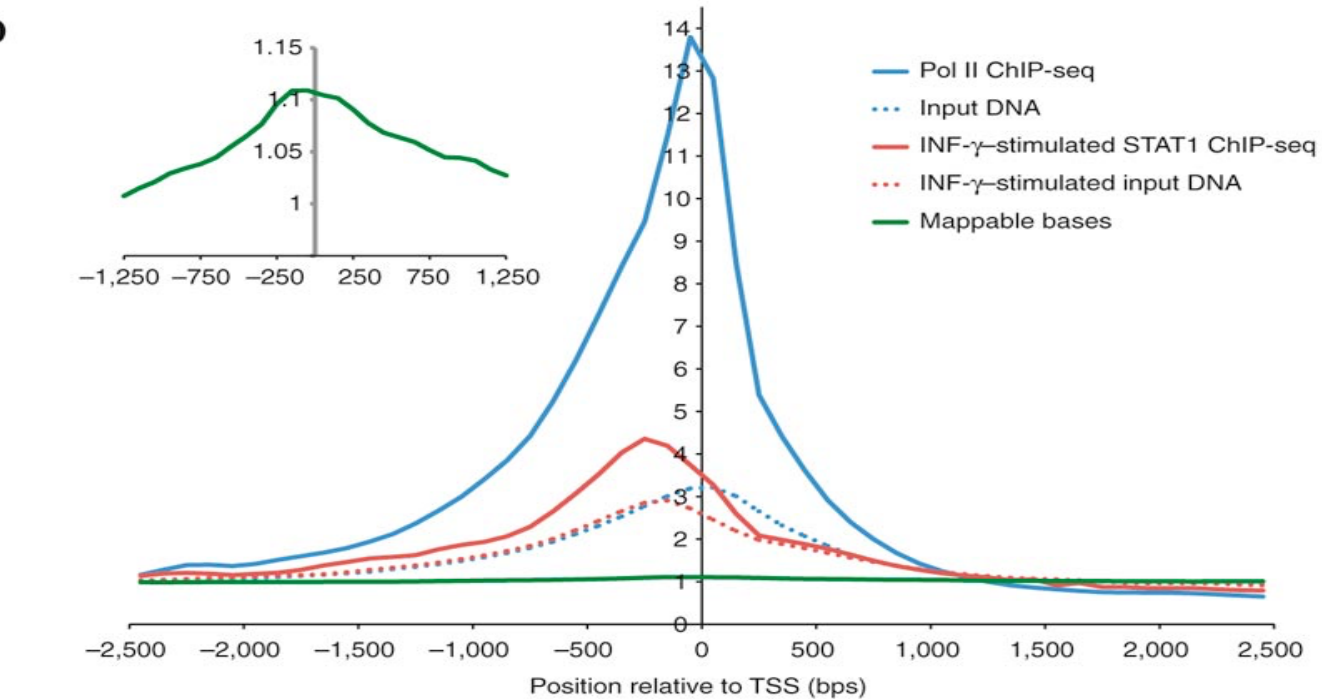


Wilbanks et al. 2010 PLOS One

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:

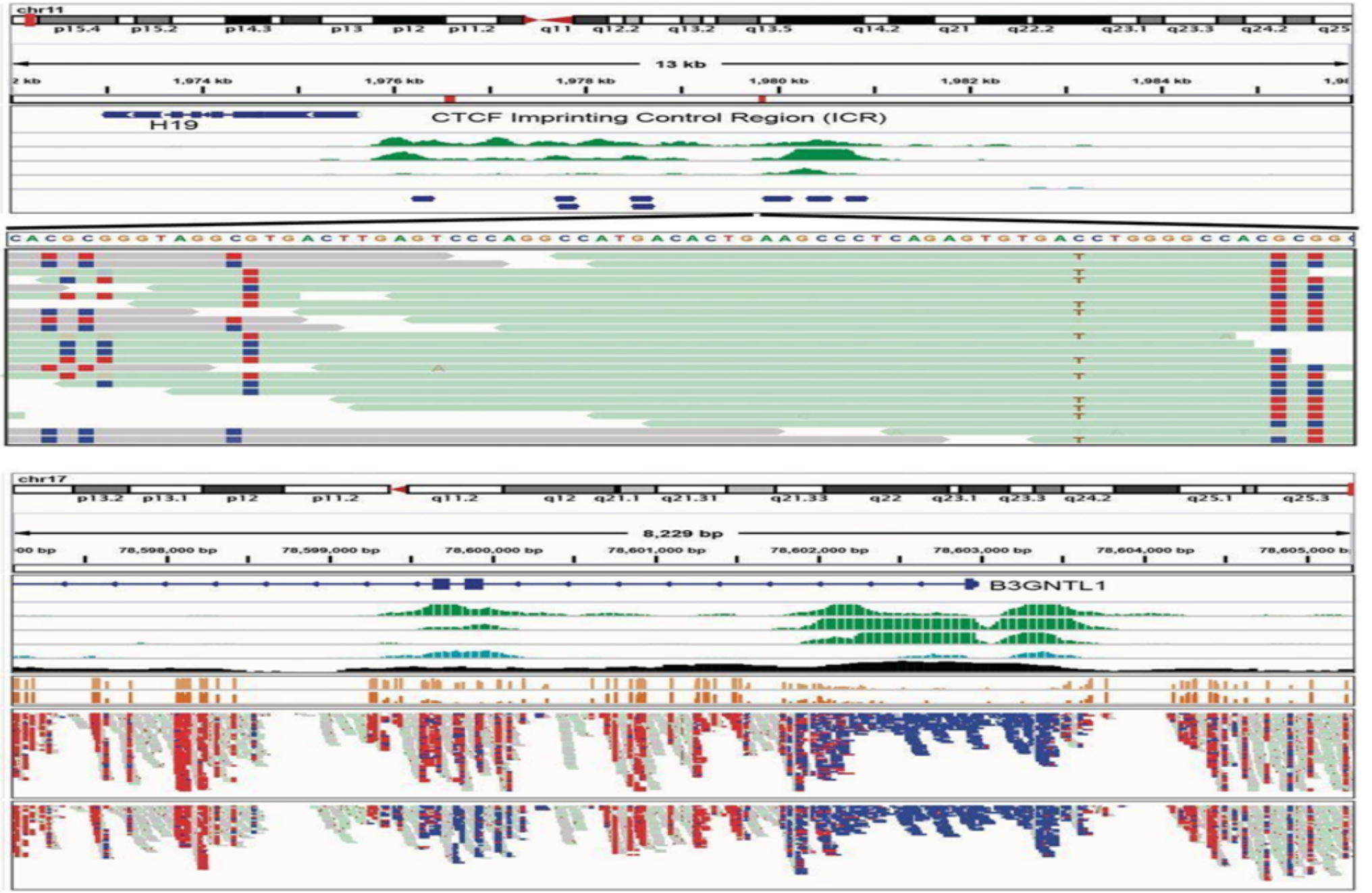
- http://www.broadinstitute.org/~anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf

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**

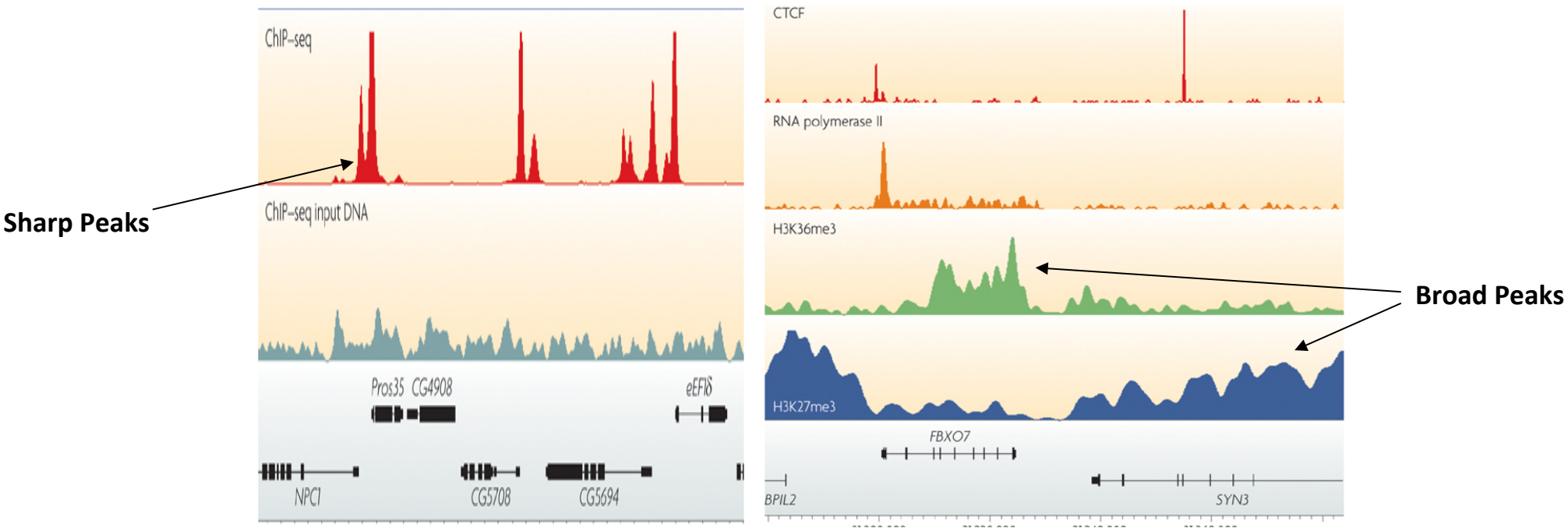
Visualizing binding sites and replicates



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

ChIPseq Quality Control