Introduction to ChIP-seq analysis

Shamith Samarajiwa

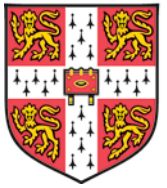
Computational Biology and Statistics group

University of Cambridge

Analysis of High-throughput sequencing data with BioConductor

4-6 September 2014

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE



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/Technical Replicates (at least 3) !!**
- ChIP-seq controls – KO, Input or IgG

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.

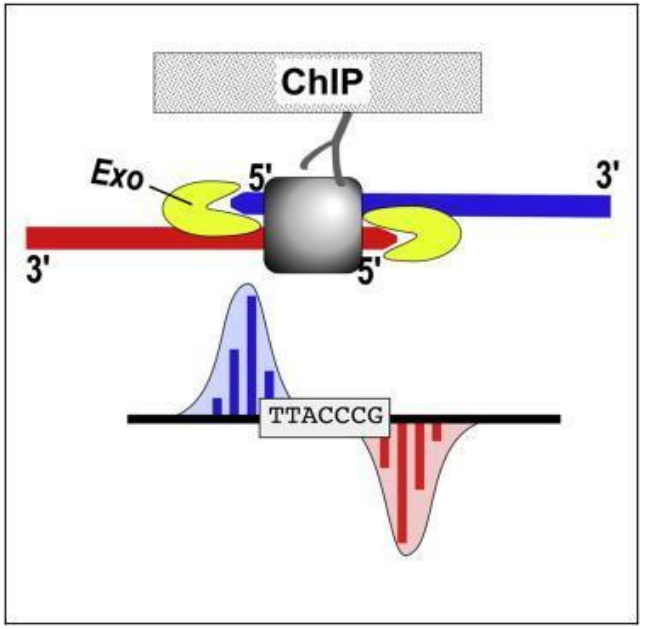
As a test system, we selected the mammalian transcription factor STAT1, whose cellular biology is relatively well characterized, and whose use permits a comparison of unstimulated and stimulated cellular states[12–16]. In both resting and stimulated cells, STAT proteins shuttle continuously between cytoplasm 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

Robertson *et al*, 2007

- ChIP-Seq technology was used to study genome-wide profiles of STAT1 DNA association.
- STAT1 targets in interferon-γ-stimulated and unstimulated human HeLA S3 cells are compared.
- The performance of ChIP-seq is compared to the alternative protein-DNA interaction methods of ChIP-PCR and ChIP-chip.
- 41,582 and 11,004 putative STAT1 binding regions are identified in stimulated and unstimulated cells respectively.

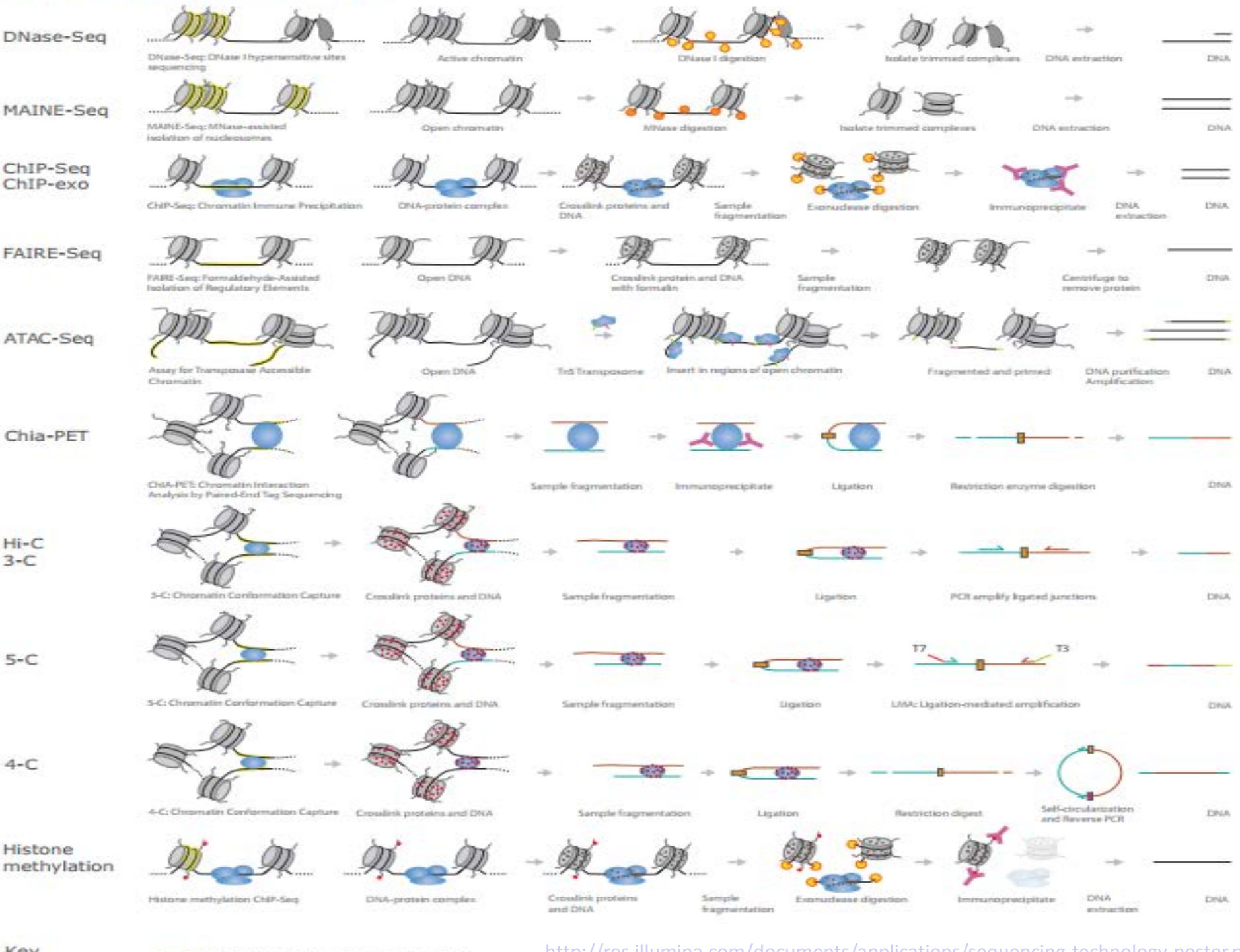
DNA-protein interaction technologies

- ChIP-chip : combines ChIP with microarray technology.
- ChIP-PET : ChIP with paired end tag sequencing
- ChIP-exo : ChIP-seq with exonuclease digestion



- Sono-seq : Sonication of cross linked chromatin sequencing.
- CLIP-seq (HITS-CLIP): cross-linking immuno-precipitation high throughput sequencing

DNA-Protein Interactions



http://res.illumina.com/documents/applications/sequencing-technology-poster.pdf

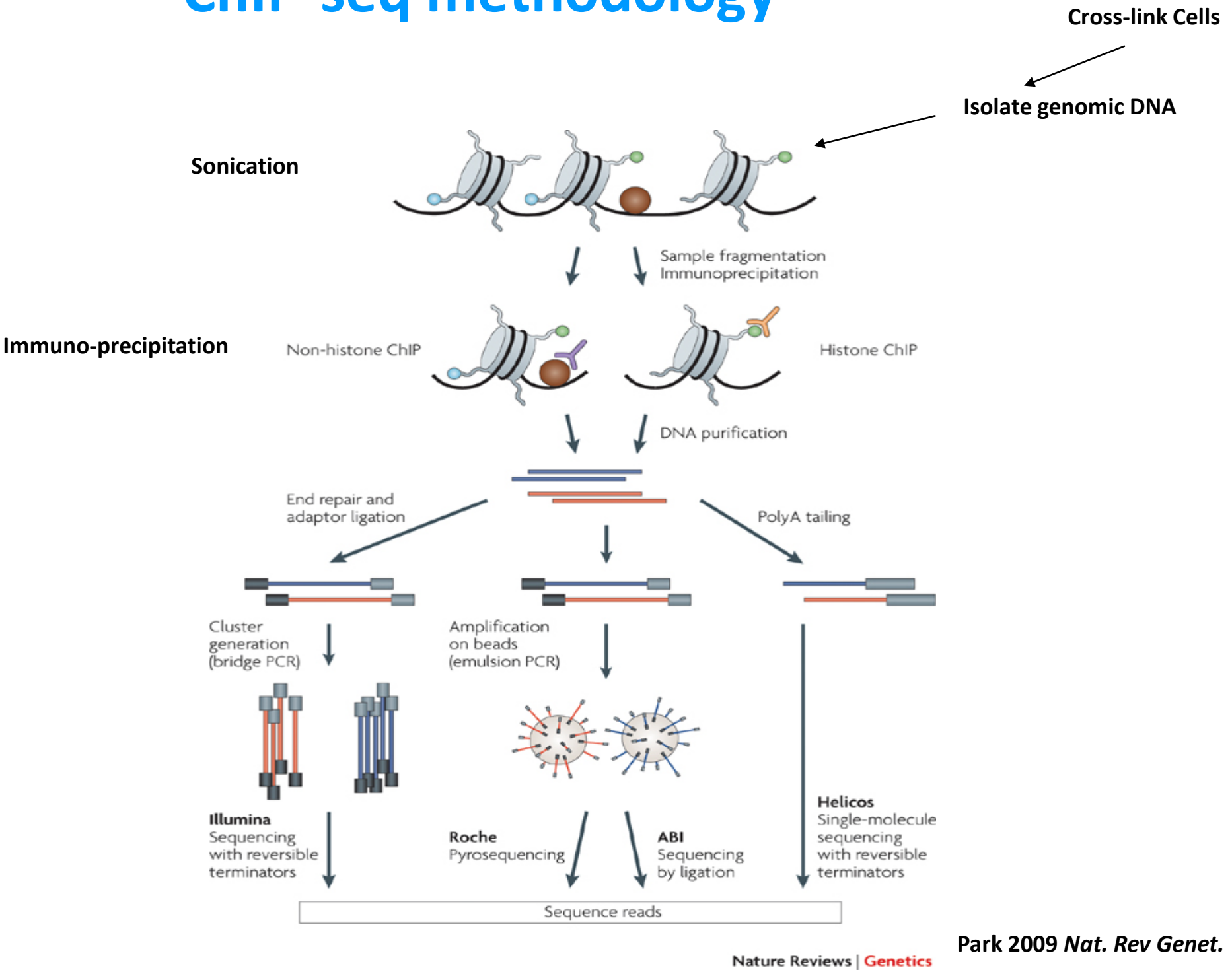
Jacques Retief, Illumina

What is ChIP-Sequencing?

ChIP-seq

- Combination of chromatin immuno-precipitation (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 or other chromatin associated protein binding (ex: PolII) or Histone modifications on the genome.
- The typical ChIP assay usually take 4–5 days, and require approx. 10^6~ 10^7 cells.

ChIP-seq methodology



Park 2009 *Nat. Rev Genet.*

ChIP-seq workflow overview 1

- ChIP-seq 'wet-lab processing and library preparation’
- Sequencing
- Quality control of raw reads
- Mapping reads to a reference genome
- Remove artifacts
- Visualization and Replicate comparison
- Binding site identification: peak calling and other methods
- Peak QC
- Annotation of peaks to genomic features
- Binding site distributions

ChIP-seq statistical aspects and best practices

Statistical Aspects:

Cairns *et al.*, “Statistical Aspects of ChIP-Seq Analysis.” In Advances in Statistical Bioinformatics, edited by Kim-Anh Do, Zhaohui S Qin, and Marina Vannucci, 138–169. Cambridge University Press, 2013.

Experimental guidelines:

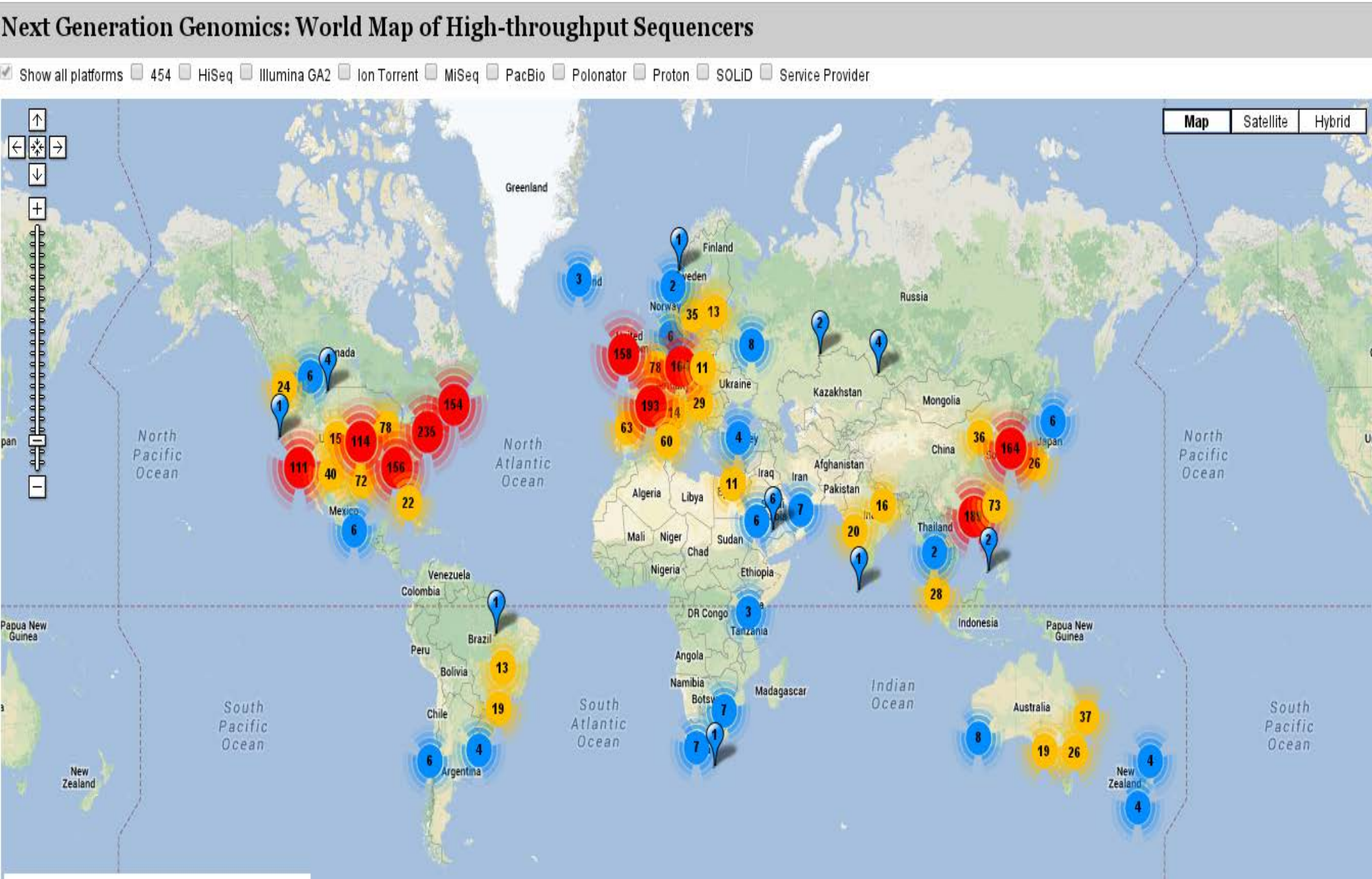
Landt *et al.*, “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.” 2012 *Genome Res.*

These guidelines address :

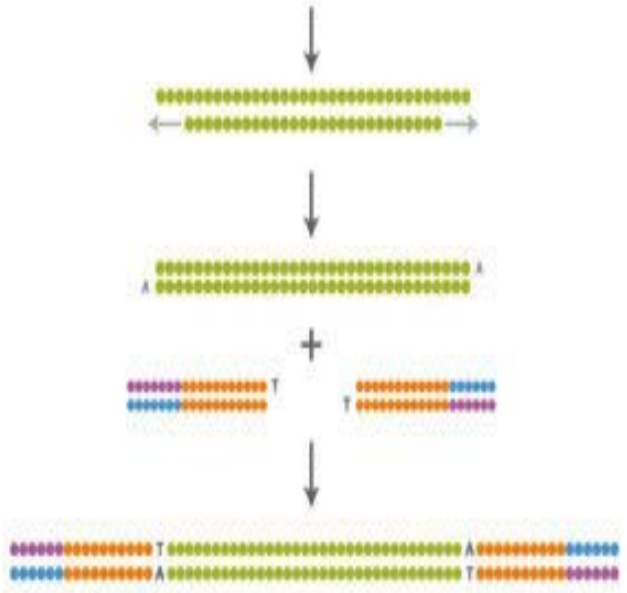
- antibody validation,
- experimental replication,
- sequencing depth,
- data and metadata reporting,
- and data quality assessment.

Sequencers

http://omicsmap.com



Illumina Genome Analysis System



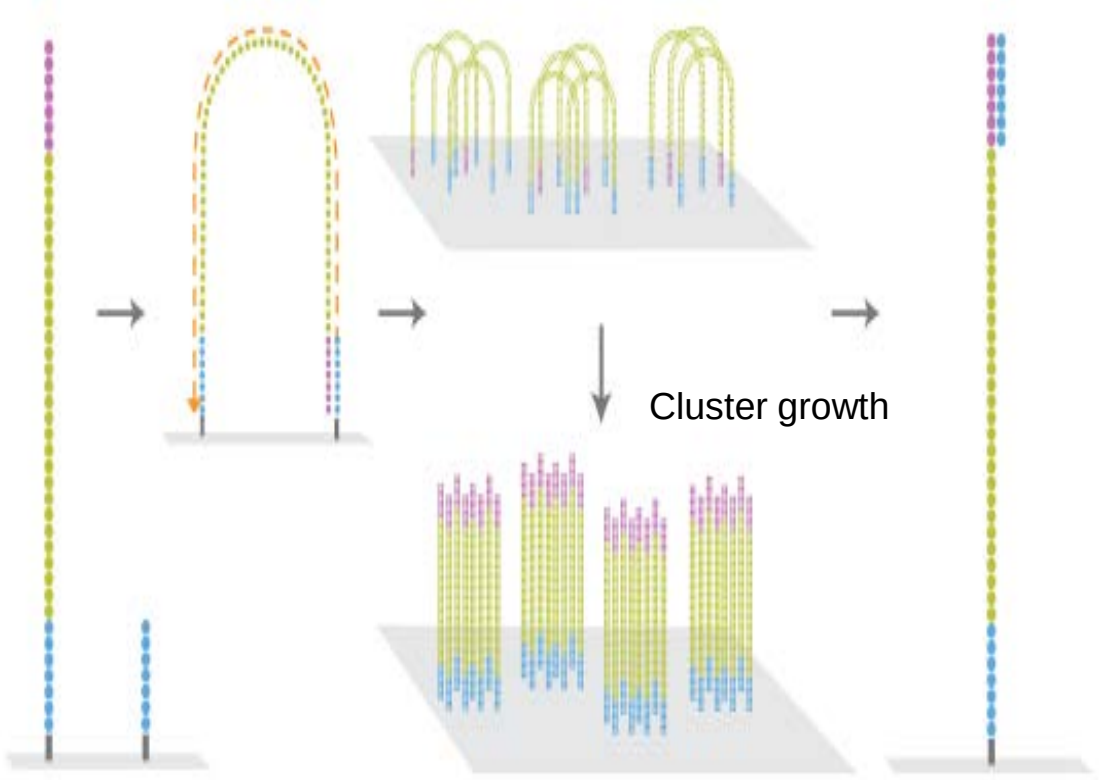
Library Preparation

Cluster Generation

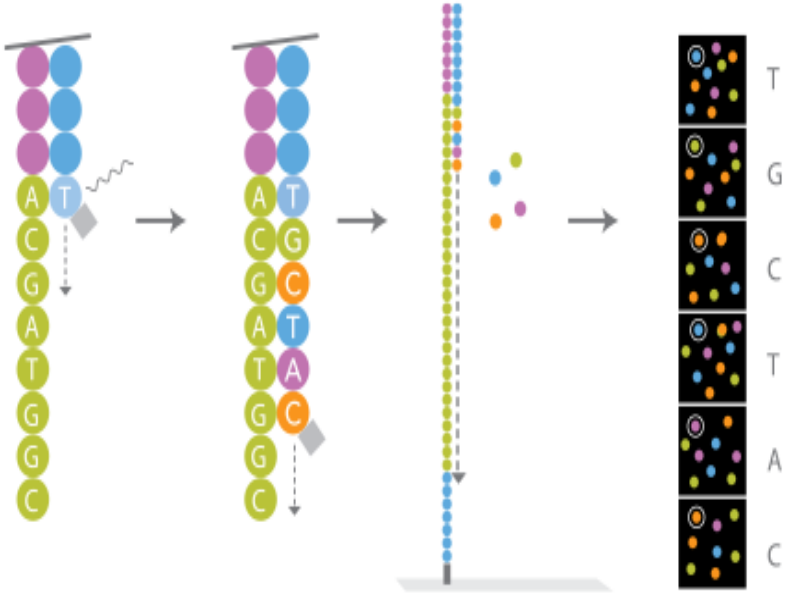
Sequencing by Synthesis

Sequencing

Bridge Amplification

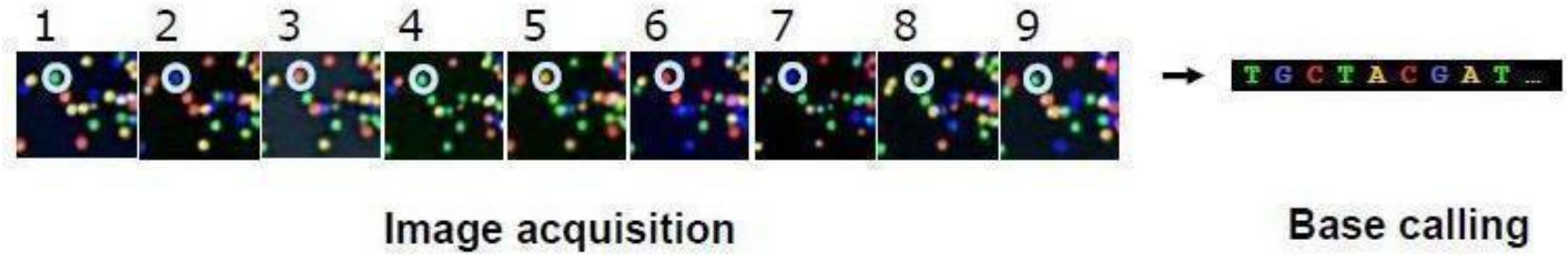


Sequencing



Incorporation of fluorescence,
reversibly terminated tagged nt

Sequencer Output



Quality Scores

Sequence Files

FASTQ formats

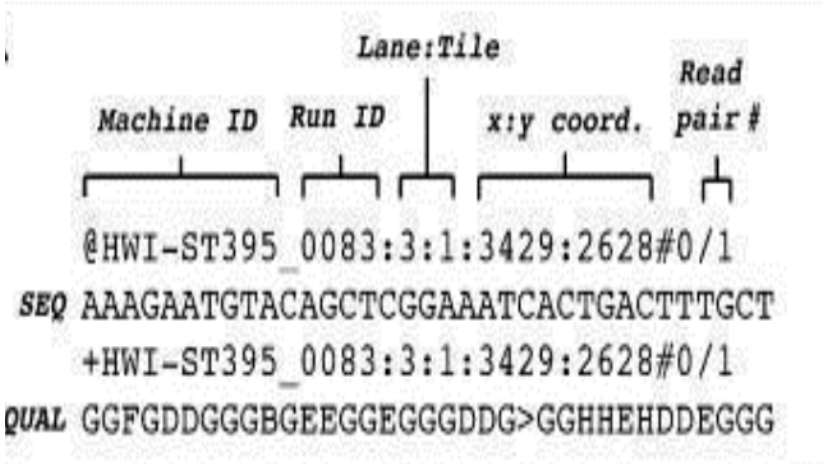
A FASTQ file normally uses four lines per sequence.

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 values 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.

ChIP-seq workflow overview 2

- **Extracting biological meaning: Downstream analysis methods**

  Motif identification and Motif Enrichment Analysis

    Feature overlap analysis

    Functional enrichment analysis : Ontologies, Gene Sets, Pathways

    Differential peak analysis

    Integration with transcriptomic data : Identify true targets

    Deciphering Regulomes & Network Biology applications

Quality Control

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

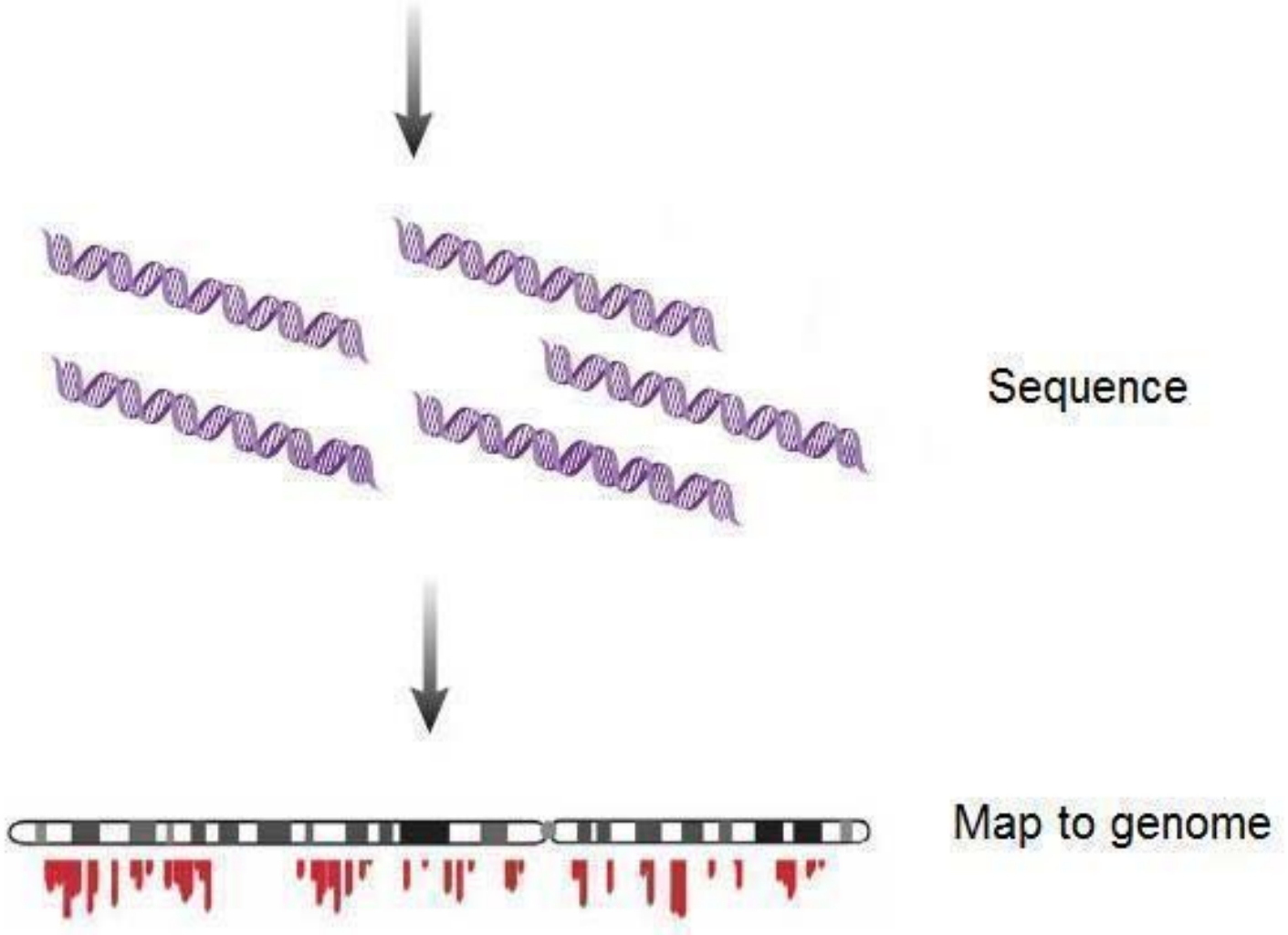
Tools:

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

Artefact detection : **FASTQC,** NGSQC

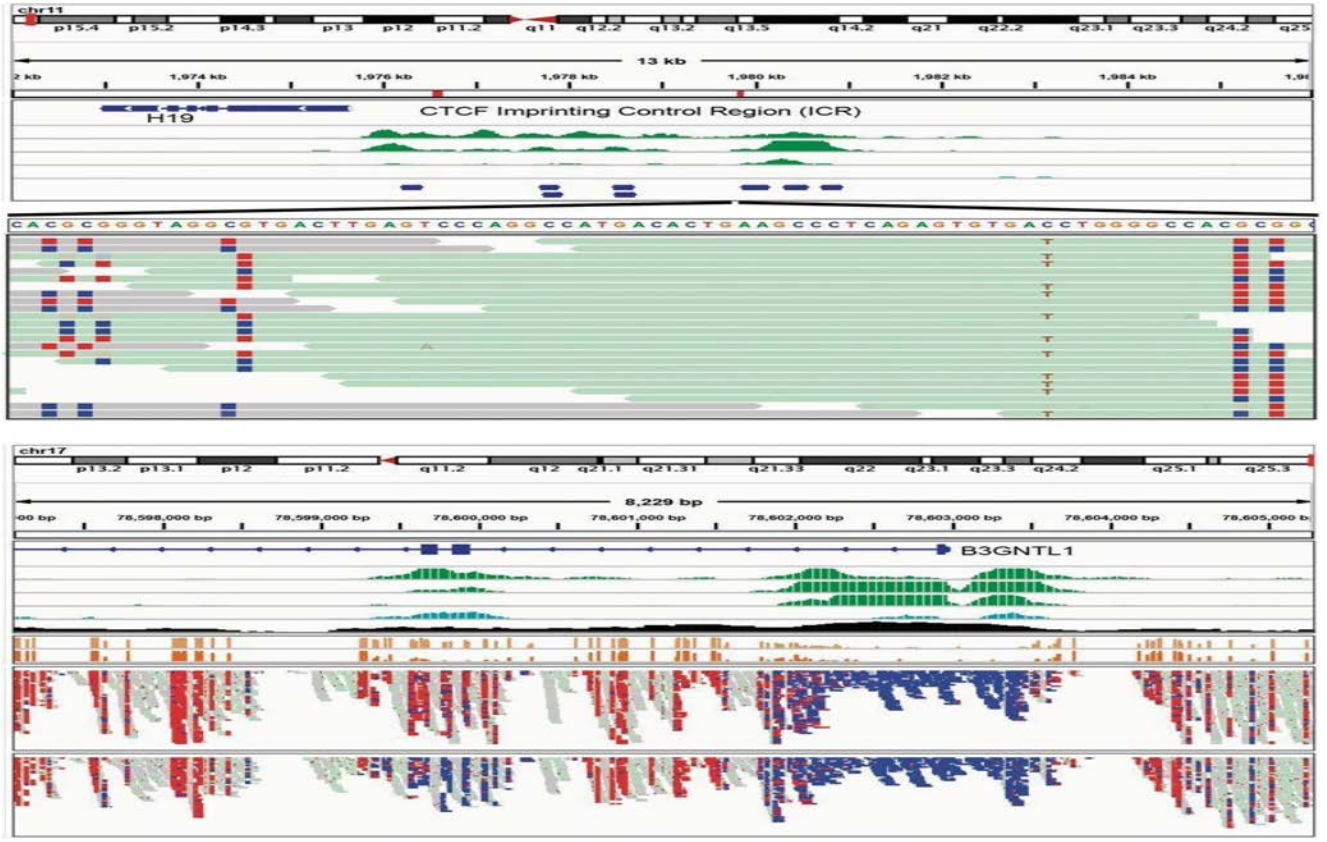
Artefact removal : **shortread (BioC),** Useq, TagDust, **CutAdapt, Trim Galore,** FASTX toolkit

Map to reference genome



Visualizing binding sites and Replicate Comparisons

IGV



Peak shapes

- TF and Epigenetic marks have different peaks shapes.
- The same TF may have different peak shapes reflecting different biological properties of binding.
- Replicates should have similar binding patterns.

Artefact removal 1

- After reads have been aligned “blacklisted regions” are removed before peak calling.

Blacklisted regions attempt to identify regions of the reference genome which are troublesome (repetitive elements or other anomalies) for high throughput sequencing aligners.

These regions confuse peak callers and result in spurious signal.

- Alignability provides a measure of how often the sequence found at the particular location will align within the whole genome.
- Uniqueness is a direct measure of sequence uniqueness throughout the reference genome.

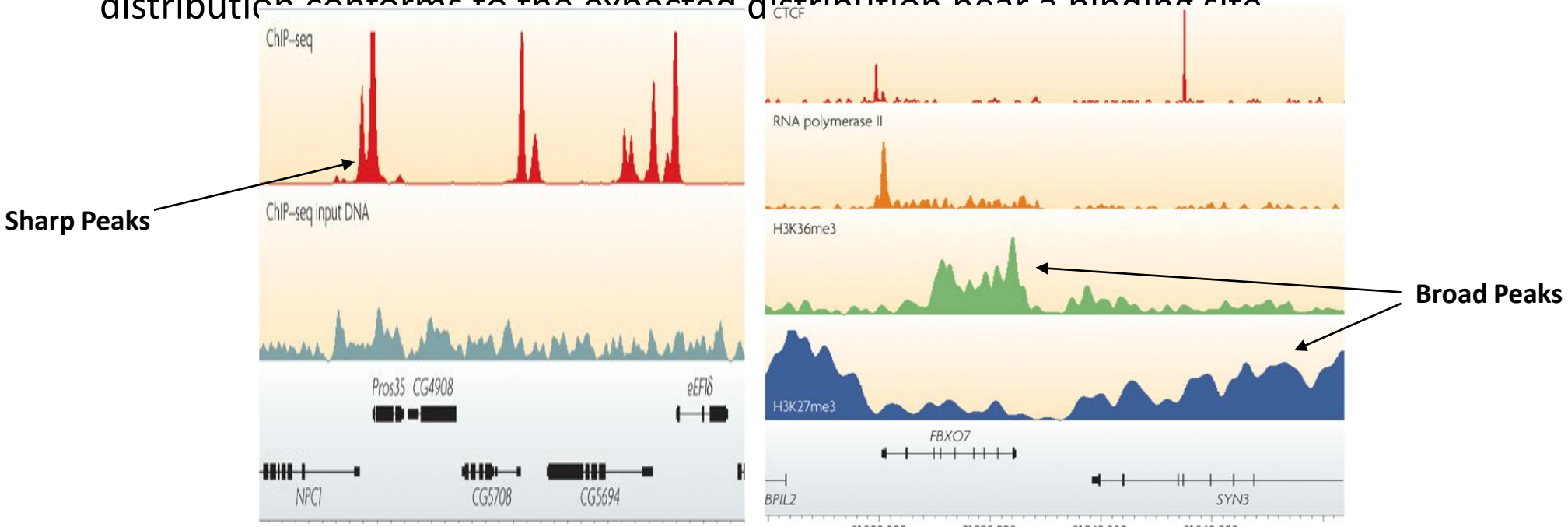
Artefact removal 2

- 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 next gen sequencing experiments independent of cell line and type of experiment.
- There were 80 open chromatin tracks (DNase and FAIRE datasets) 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. Some of these regions overlap pathological repeat elements such as satellite, centromeric and telomeric repeats. However, simple mappability based filters do not account for most of these regions. Hence, it is recommended to use this blacklist alongside mappability filters.
- The *DAC Blacklisted Regions* track was generated for the ENCODE project. The *Duke Excluded Regions* track displays genomic regions for which mapped sequence tags were filtered out before signal generation and peak calling for Open Chromatin: DNaseI HS and FAIRE tracks. This track contains problematic regions for short sequence tag signal detection (such as satellites and rRNA genes). The *Duke Excluded Regions* track was generated for the ENCODE project.

DATA PROCESSING

Peak-calling

- Identify TF binding locations (map a set of reads to a set of genomic intervals)
- Two main strategies:
  - Count-based - De
    fine regions. Count the number of reads falling into each region. When a region contains a statistically signifi
    cant number of reads, call that region as 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 distributio[illegible] d[illegible]



Peak Callers

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

Sharp TF peaks:

MACS v1.4.2 & MACS v2: model based analysis for ChIP-seq:

Zhang *et al.*, 2008, PMID 18798982

Feng et al., "Using MACS to Identify Peaks from ChIP-Seq Data" 2011, PMID: 2163394

BayesPeak (BioC) : A Bayesian peak caller

Cairns *et al.*, 2011, PMID 21245054

Jmosaics (Bioc): Joint analysis of multiple ChIP-seq datasets

Zeng et al., 2013, PMID: 23844871

SPP :

Kharchenko et al., 2008 PMID:19029915

T-PIC :

Hover et al., 2011 PMID:21226895

Diffuse chromatin modification peaks:

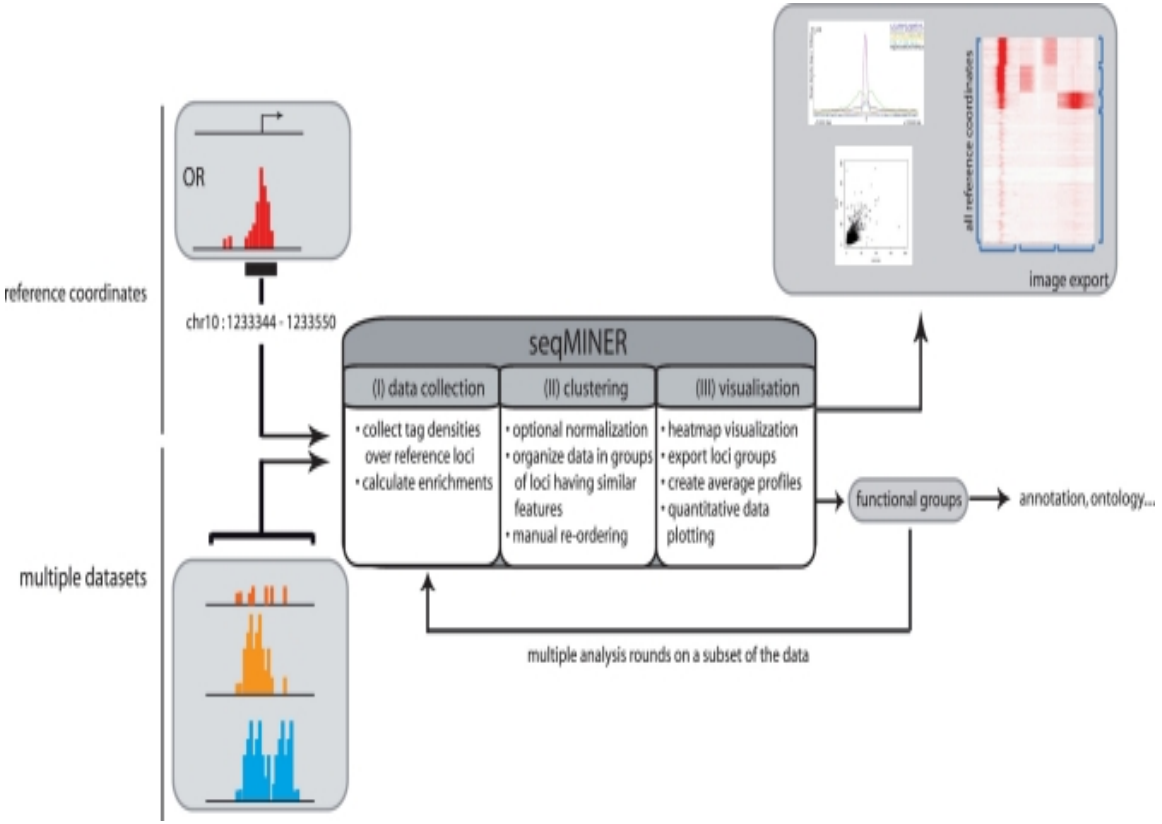
RSEG

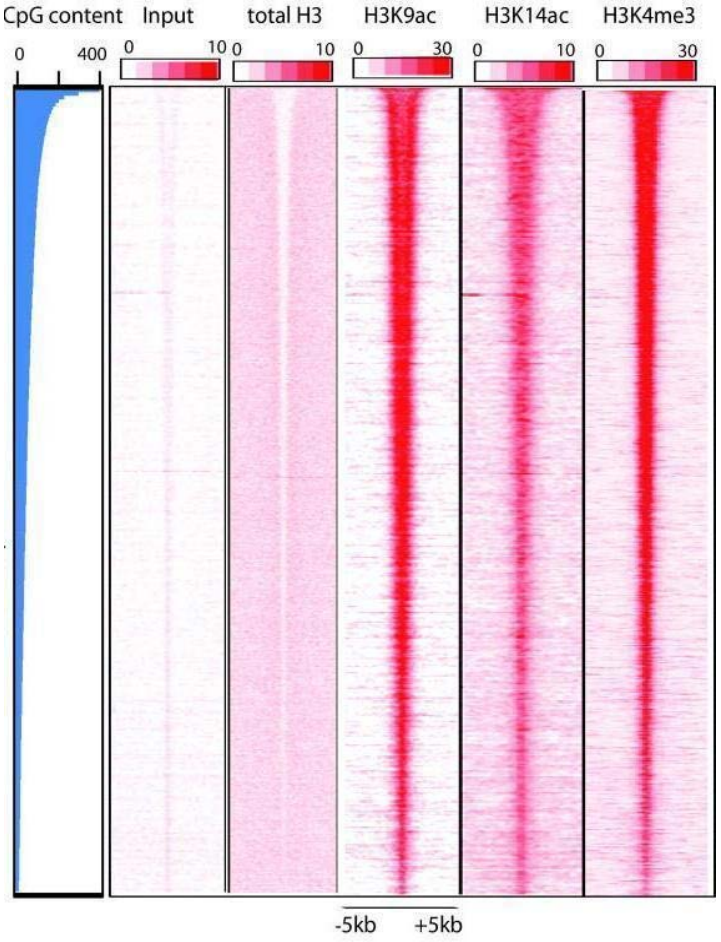
"Evaluation of algorithm performance in ChIP-seq peak detection." Wilbanks EG, Facciotti MT. PLoS One. 2010 Jul 8;5(7):e11471.

seqMINER

- Enables qualitative comparisons between a reference set of genomic positions and multiple ChIP-seq data-sets.
- Useful for comparing and visualizing replicates or conditions.

Ye *et al.*, 2011 Nucleic Acids Res. PMID: 21177645





Peak annotation 1

- **ChIPpeakAnno (BioC)** (Zhu *et al.*, 2010, BMC Bioinformatics)
- map peaks to nearest feature (TSS, gene, exon, miRNA or custom features)
  - extract peak sequences
  - find peaks with bidirectional promoters
  - obtain enriched gene ontology
  - map different annotation and gene identifiers to peaks
- can use biomaRt package to get annotation.
- IRanges, GenomicFeatures, GO.db, BSgenomes, multtest
- converts BED and GFF data formats to RangedData before calling peak annotate function.

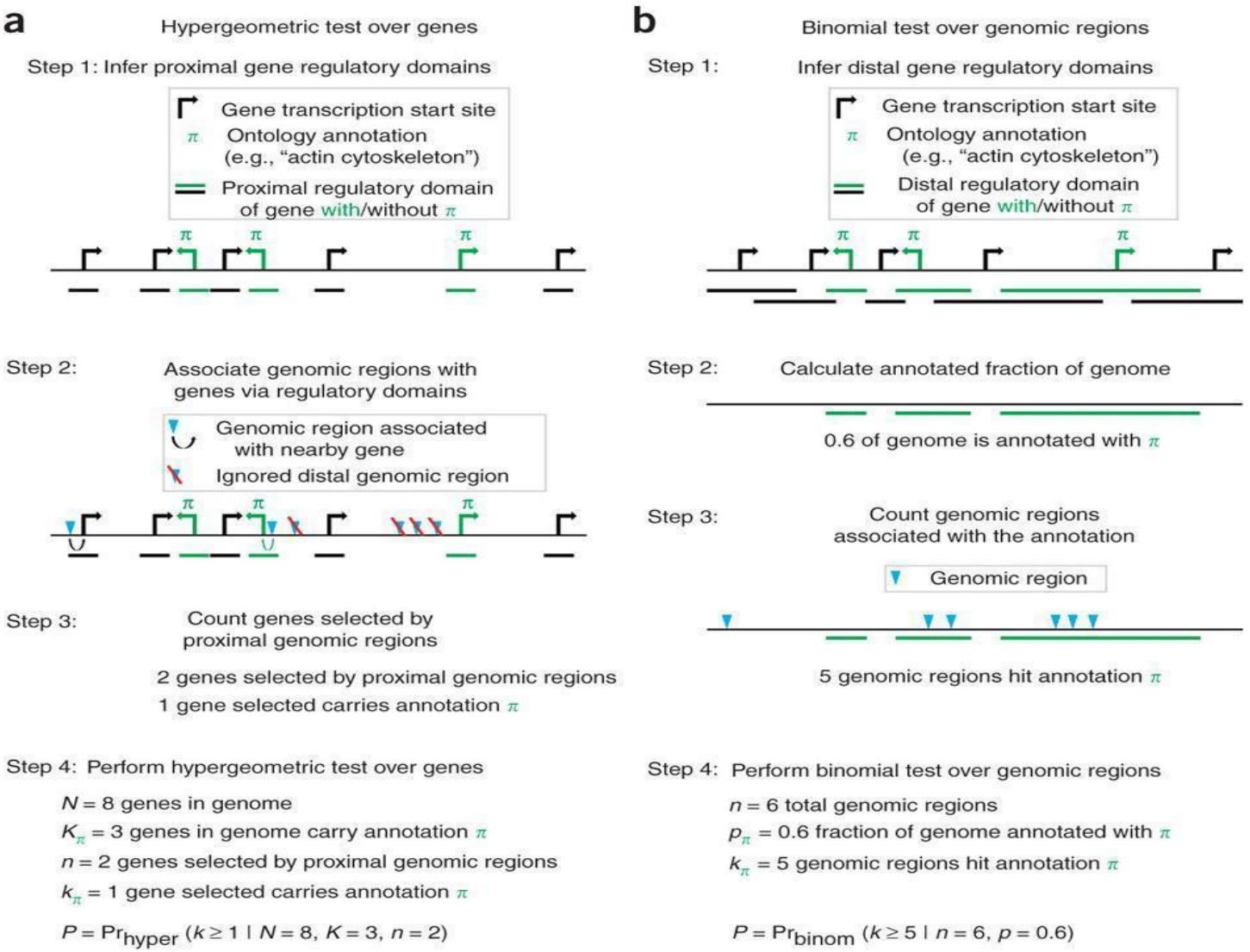
Peak annotation 2

PeakAnalyzer (Salmon-divon *et al.*, 2010, BMC Bioinformatics)

- A set of high-performance utilities for the automated processing of experimentally-derived peak regions and annotation of genomic loci.
- Consists of PeakSplitter and PeakAnnotator.
- Biologist' friendly tool.
- Get latest genome annotation files from Ensembl (gtf format) or UCSC (BED format).
- Map to either nearest downstream gene, TSS or user defined annotation.
- Determine overlap between peak sets.
- Split peaks to sub-peaks. May be useful for *de novo* motif analysis.

Functional Enrichment Analysis

GREAT: Genomic Regions Enrichment of Annotations Tool



McLean C.Y. et al., "GREAT improves functional interpretation of *cis*-regulatory regions". *Nat.Biotechnol.* **28**(5):495-501, 2010.

Motif detection

Don't scan a sequence with a motif and expect all sites identified to be biologically active. Random matches will swamp the biologically relevant matches! This is a well known problem in motif searching, amusingly called the **"Futility Theorem"** of motif finding. -Wasserman WW, Sandelin A. Nat Rev Genet 2004;5:276-87.

1. PWM based sequence scanning or word search methods. These methods uses prior information about TF binding sites and therefore can only be used to detect known Transcription Factor Binding Sites (TFBS).

2. *De novo* motif identification –Pattern discovery methods:

- **Word based** – occurrence of each 'word' of nucleotides of a certain length is counted and compared to a background distribution.

- **Probabilistic**- seek the most over-represented pattern using algorithmic approaches like Gibbs sampling and Expectation maximization. These iteratively evolve an initial random pattern until a more specific one is found.

Use *de novo* motif calling and alignment to build your own PWMs! Biostrings & Motiv packages have PFM to PWM conversion methods.

BioConductor motif analysis packages

- rGADEM -motif discovery
- MotifRG -motif discovery
- MotIV -map motif to known TFBS, visualize logos
- motifStack -plot sequence logos
- MotifDb -motif database
- TFBSTools – R interface to the JASPAR database
- PWMenrich -motif enrichment analysis

Position Weight Matrices

a

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Source binding sites

b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

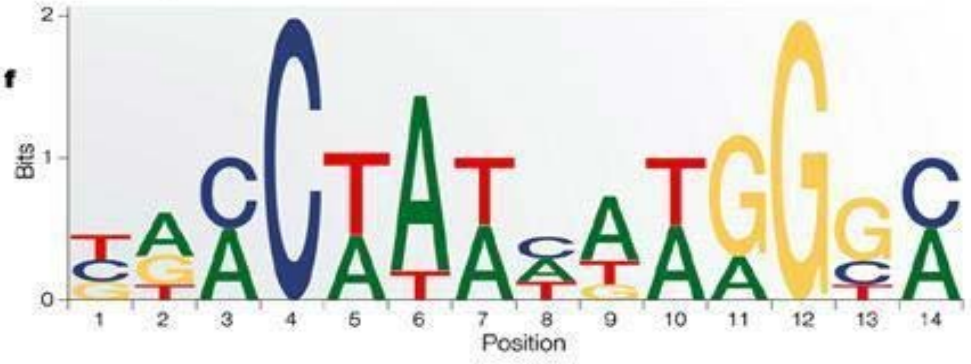
A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

e Site scoring

0.45	−0.66	0.79	1.68	0.45	−0.66	0.79	0.45	−0.66	0.79	0.00	1.68	−0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

Σ = 5.23, 76% of maximum

f



Nature Reviews | Genetics

PWM conversion:

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

TFBS PWM/PFM sources

TRANSFAC public	Matys et al., 2006	Multiple species	v7.0 2005, Not been updated for a while!
TRANSFAC professional	Matys et al., 2006	Multiple species	v2013.4
JASPAR 2014	Mathelier et al., 2014	Multiple species	(656)
ORegAnno		Multiple species	Curated collection from different sources.
hPDI	Xie et al., 2010	Human	(437)
SwissRegulon	Pachkov et al., 2010	mammalian	(190)
HOMER	Heinz et al., 2010	Human	(1865)
UniPROBE	Newburger & Bulyk, 2009	Multiple species	
Dimers	Jonawski et al., 2013	Human	(603) predicted dimers
FactorBook	Wang et al., 2012	Human	(79) ENCODE ChIP-seq motifs
SCPD, YetFasco		Yeast	
Elemento, Redfly FlyFactorSurvey,Tiffin		Drosophila	

Motif Enrichment Analysis

MEA identifies over- and under-represented known motifs in a set of genes.

The regulatory proteins whose DNA binding motifs are enriched in a set of regulatory sequences are candidate transcription regulators of that gene set.

Identifying co-regulated gene sets is difficult. Use Ontologies, pathways, GSEA etc.

Picking the right background model will determine the success of the motif enrichment analysis:

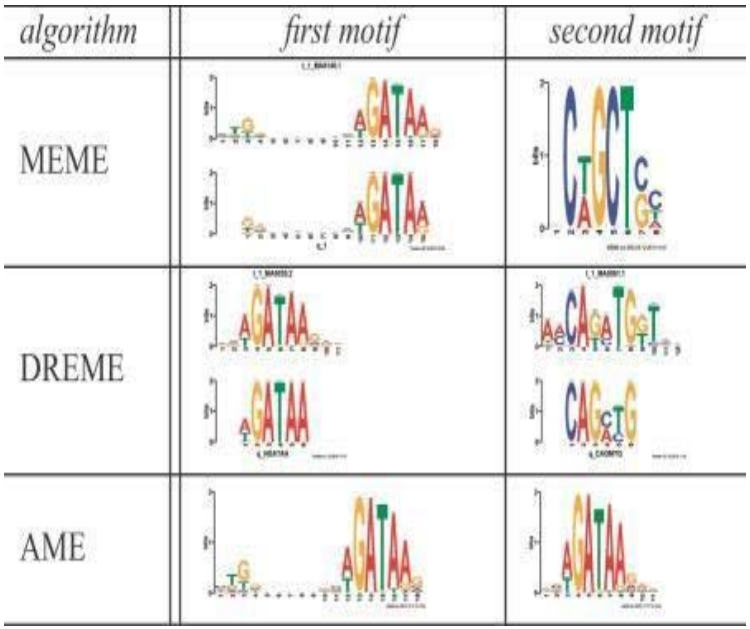
- All core-promoters from protein coding or non-coding genes etc,
- Higher order Markov model based backgrounds,
- A sequence set similar in nucleotide composition, length and number to the test set,
- Open chromatin regions,
- A shuffled test sequence set.

MEME-Chip

- http://meme.nbcr.net

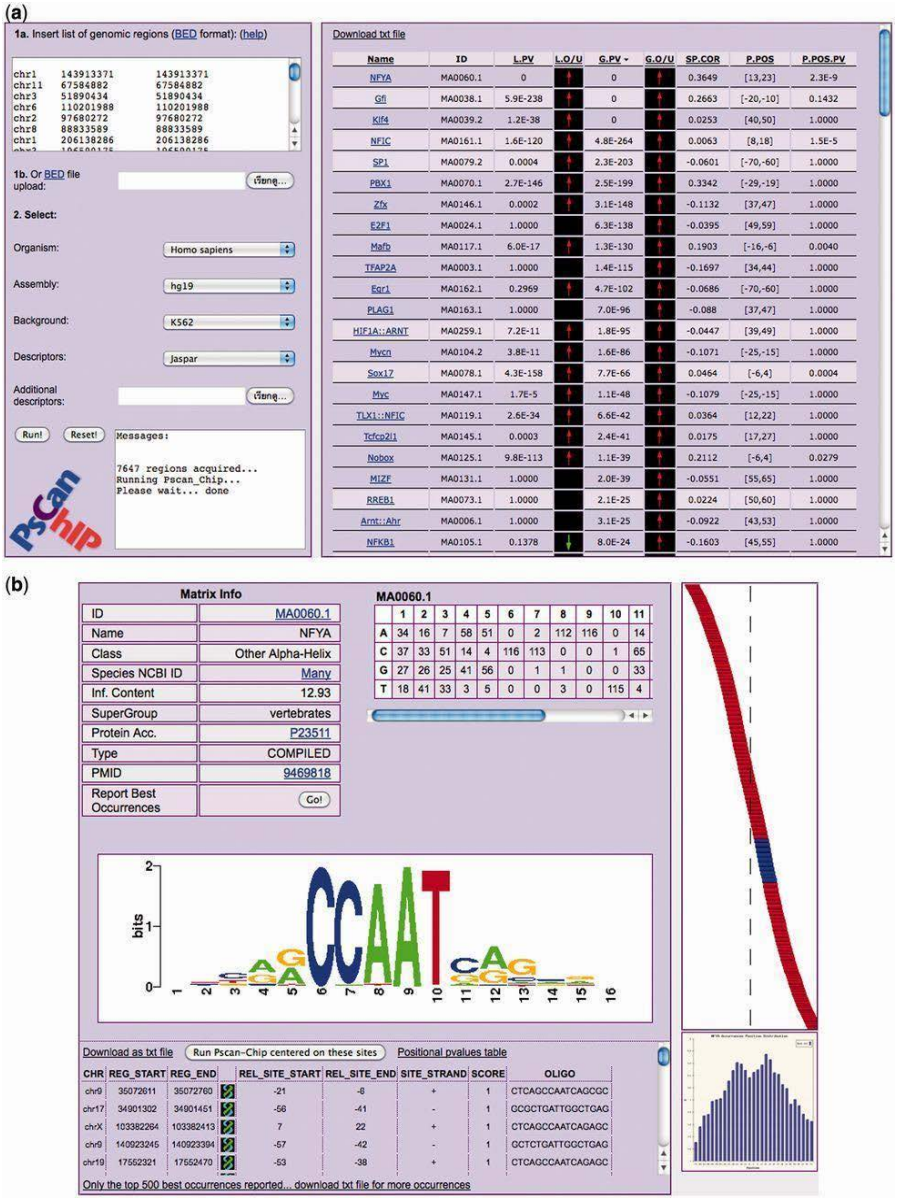
"MEME-ChIP: motif analysis of large DNA datasets." Machanick and Bailey, 2011 Bioinformatics

- Given a set of genomic regions, it performs
  - *ab initio* motif discovery -novel TF binding sites (MEME, DREME)
  - motif enrichment analysis -known TF enrichment (Centrimo/AME)
  - motif visualization (MAST and AMA)
  - binding affinity analysis
  - motif identification -compare to known motifs (TOMTOM)
- Uses two algorithms for motif discovery:
- MEME -expectation maximization (EM) to discover probabilistic models of DNA-binding by single TFs or TF complexes.
- DREME -simpler, non-probabilistic model (regular expressions) to describe the short binding motifs.



Pscan-Chip

- http://159.149.160.51/pscan_chip_dev/
- Motif enrichment analysis using PWM databases and user defined background models.
- Optimized for ChIP-seq.
- Ranked lists of enriched motifs.
- Sequence logo's and motif enrichment distribution plots.



"PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments." Zambelli *et al.*, 2013 *Nucleic Acids Res.*

HOMER (Hypergeometric Optimization of Motif EnRichment) v4

- http://homer.salk.edu/homer/index.html
- Large number of (Perl & C++) tools for ChIP-seq analysis.
- Provides both *de novo* and PWM scanning based motif identification and enrichment analysis.
- User can specify custom background. (Randomly selected, GC or CGI matched backgrounds.)
- Uses a collection of ChIP-seq derived PWMs or user can specify PWM.
- Peak annotation, GO enrichment analysis, Extract peak sequences, Visualization.

Meta-Motif Analyzers

http://131.174.198.125/bioinfo/gimmemotifs/

GimmeMotifs: a *de novo* motif prediction pipeline, especially suited for ChIP-seq datasets. It incorporates several existing motif prediction algorithms in an ensemble method to predict motifs and clusters these motifs using the weighted information content (WIC) similarity scoring metric.

BioProspector http://motif.stanford.edu/distributions/bioprospector/

GADEM http://www.niehs.nih.gov/research/resources/software/gadem/index.cfm

Improbizer http://users.soe.ucsc.edu/~kent/

MDmodule (included in the MotifRegressor Package) http://www.math.umass.edu/~conlon/mr.html

MEME http://meme.sdsc.edu/

MoAn http://moan.binf.ku.dk/

MotifSampler http://homes.esat.kuleuven.be/~sistawww/bioi/thijs/download.html

Trawler http://ani.embl.de/trawler/

Weeder http://159.149.160.51/modtools/

Network biology applications: Integrating transcriptomic data with ChIP-seq

- Not all TF binding sites are transcriptionally active. The collection of transcriptionally active targets of a TF is its regulome.
- Regulomes can be used to “explain” the phenotype under consideration and understand aspects of biological systems.
- Regulomes in combination with pathway and network modelling approaches can then be used decipher the networks underlying phenotypes.
- These networks provide information on connectivity, information flow, and regulatory, signaling and other interactions between cellular components.

- GeneNetworkBuilder, Bionet

Rcade

Rcade: R-based analysis of ChIP-seq And Differential Expression

- Rcade is a Bioconductor package developed by **Cairns *et al.***, that utilizes **Bayesian** methods to integrates ChIP-seq TF binding, with a transcriptomic Differential Expression (DE) analysis.
- The method is read-based and independent of peak-calling, thus avoids problems associated with peak-calling methods.
- A key application of Rcade is in inferring the direct targets of a transcription factor (TF).
- These targets should exhibit TF binding activity, and their expression levels should change in response to a perturbation of the TF.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Rcade

- Rcade integrates posterior probabilities of binding (determined via the baySeq package) with those of differential expression (determined via the limma package).

$$B = \log(\frac{PP}{1 - PP})$$

- Rcade uses a fully Bayesian modelling approach. In particular, it uses log-odds values (a measure of probability), or B-values, in both its input and output. The log-odds value is related to the posterior probability (PP) of an event, as per the formula above.
- Priors need to be defined.
- A number of output files are generated by Rcade. Usually, the file of interest is “DEandChIP.csv”, which contains a list of genes most likely to have both DE and ChIP signals ranked by their B-value.
- More on Rcade @ the practical!

DiffBind

BioConductor package by **Stark *et al.,*** for identifying sites that are differentially bound between two sample groups.

It includes functions to support the processing of peak sets, including overlapping and merging peak sets, counting sequencing reads overlapping intervals in peak sets, and identifying statistically significantly differentially bound sites based on evidence of binding affinity (measured by differences in read densities).

More on DiffBind @ the practical!