CHIP-seq and ATAC-seq for Epigenomics Profiling and Functional Analysis

Yurii Chinenov, PhD
David Oliver, PhD

Hospital for Special Surgery
Genomics Center

Presentation Breakdown

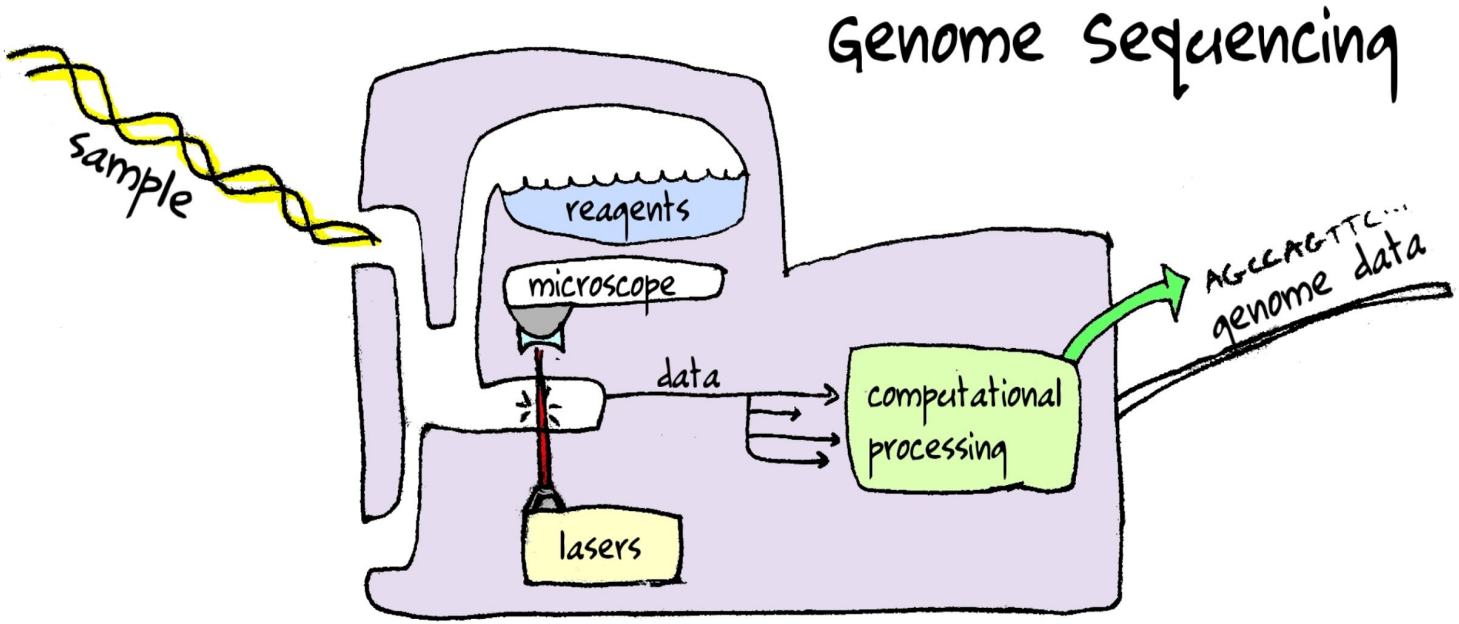
- **Part 1: Sequencing Technology**
  - Sequencing
  - ChIP-seq/ATAC-seq
- **Part 2: Pipelines and Downstream Analysis**
  - Sequencer to Data
  - Data to Meaning

Sequencing Technology

Sequencing Terminology

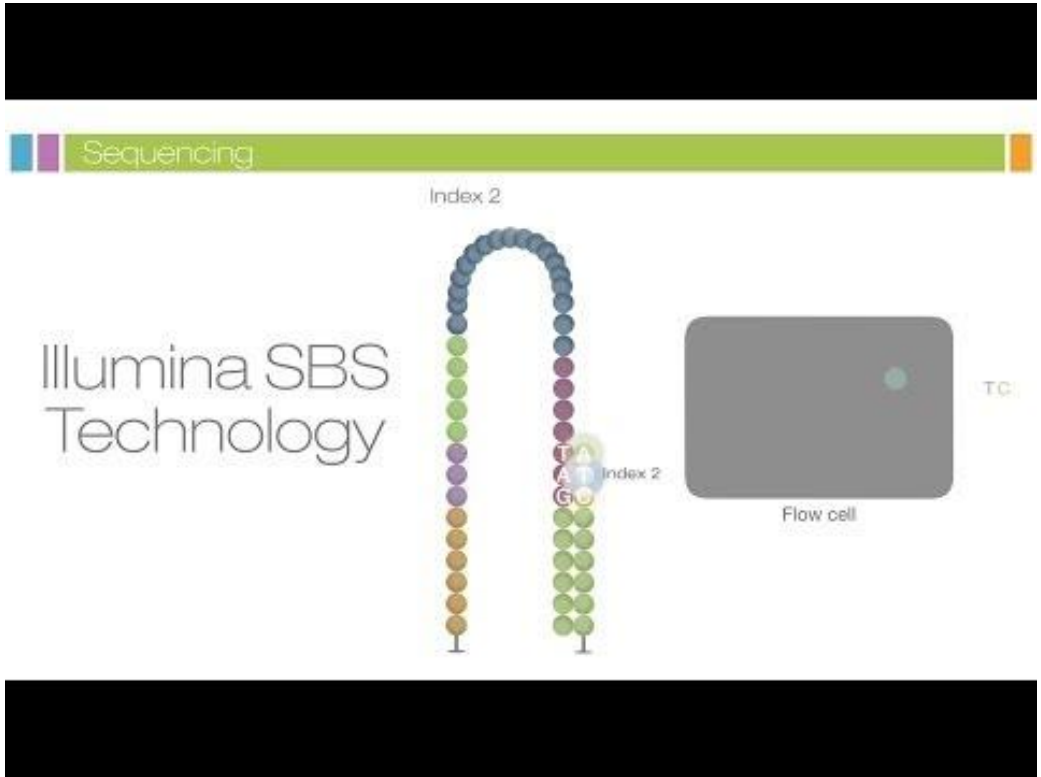
- **Sample** - The raw RNA or DNA to be sequenced
- **Library** - The **sample** after it has been prepared for sequencing (a selected portion of your sample)
- **Flow Cell** - The disposable portion of the sequencer where where the **library** is sequenced
- **Cluster** - A single molecule from the **library** amplified 1000 times to create on ‘spot’ on the **flow cell**
- **Reads** - The base calls for each **cluster**.
- **Quality** - The per-base probability that each base in the **read** was called correctly

Short Read Sequencing



https://blog.personalgenomes.org/2012/11/29/seeking-diversity/

Short Read Sequencing By Synthesis

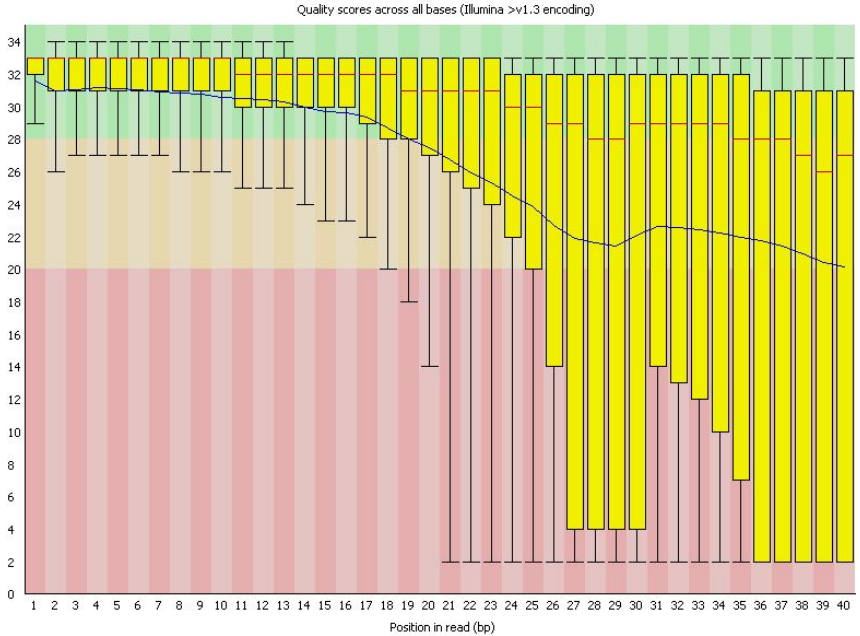


The Cost of Sequencing (Illumina)

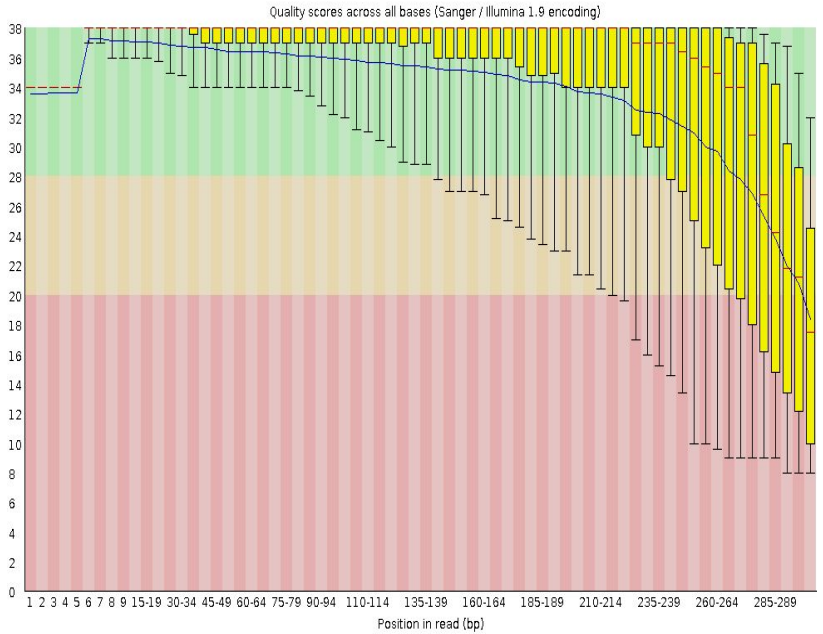
- Quality and length of sequencing has increased
- Sequencer chemistry and technology has improved
- Cost of sequencing has decreased dramatically

Quality of sequencing has increased

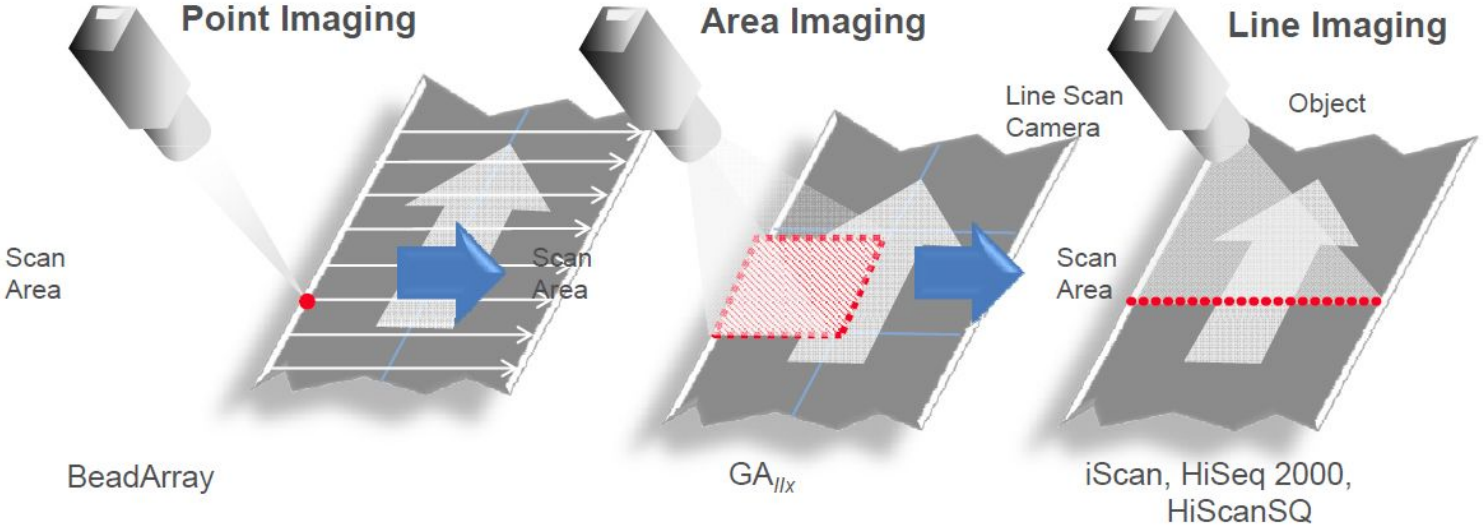
Circa GAIIx



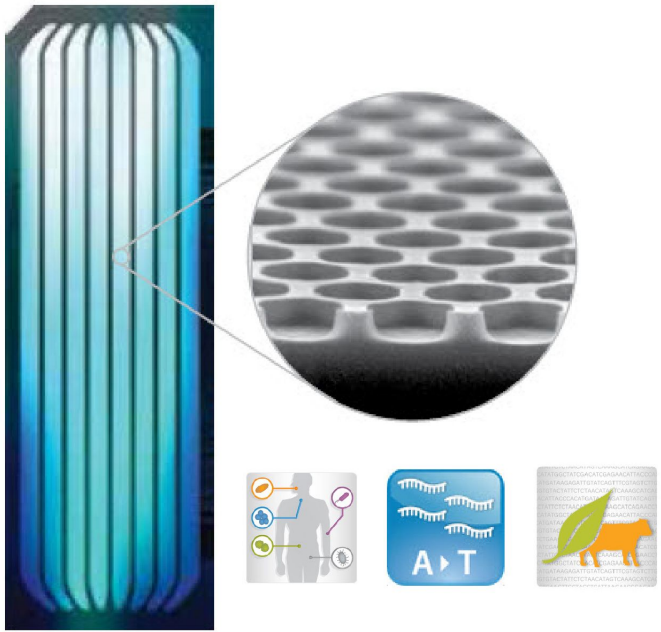
Circa MiSeq

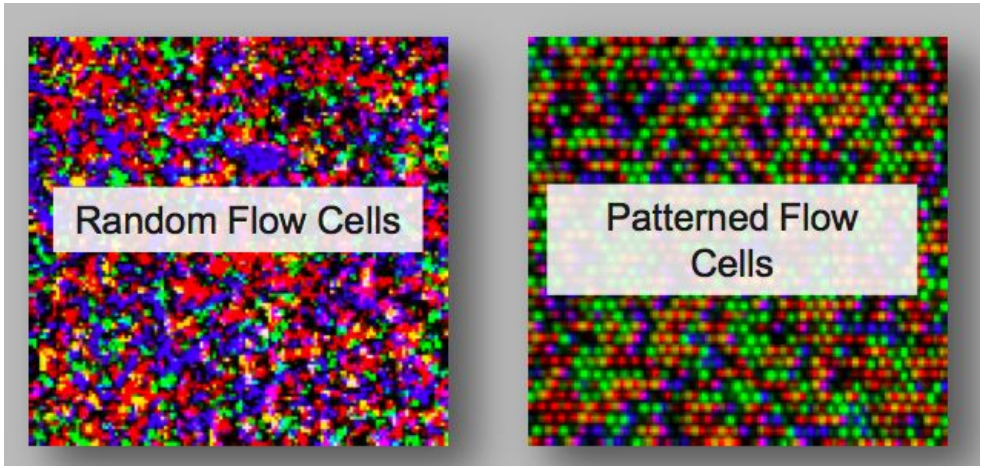


Sequencer technology has improved

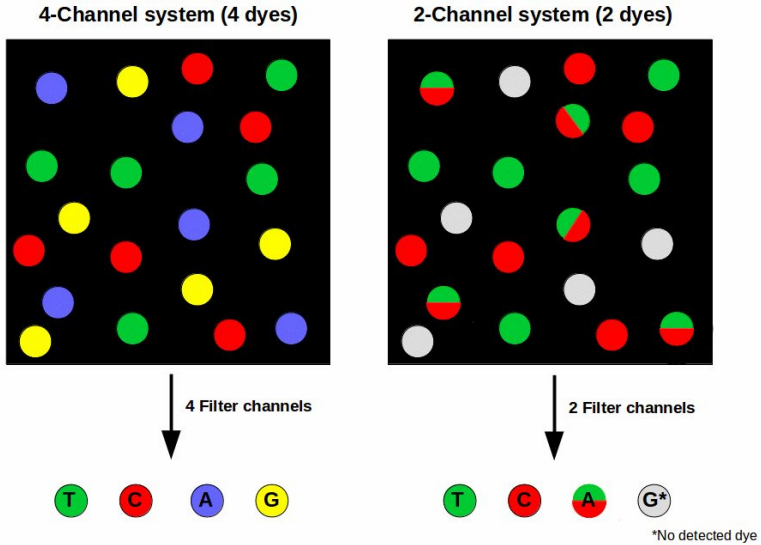


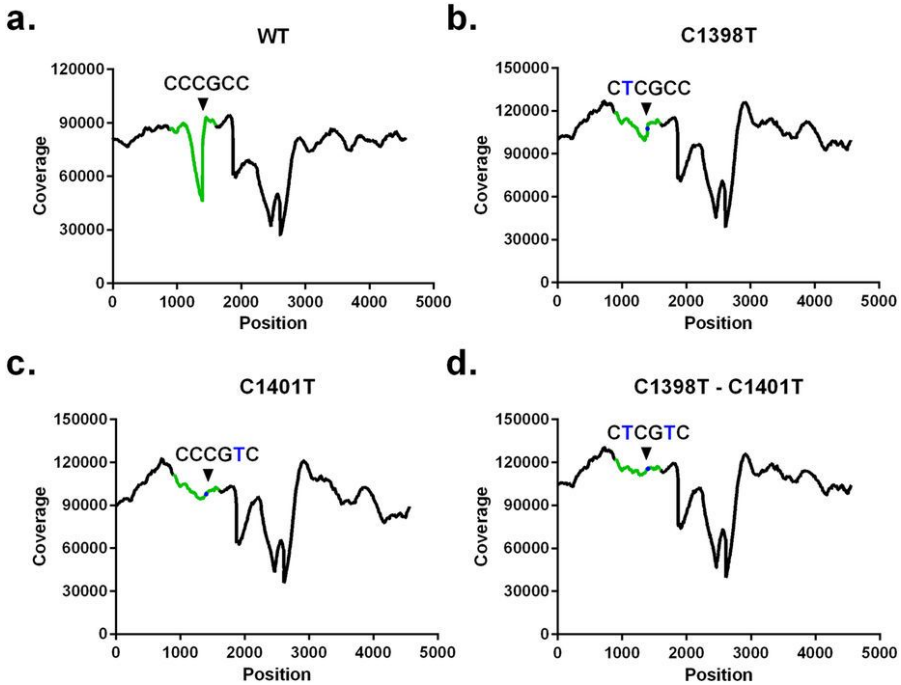
Sequencer technology has improved





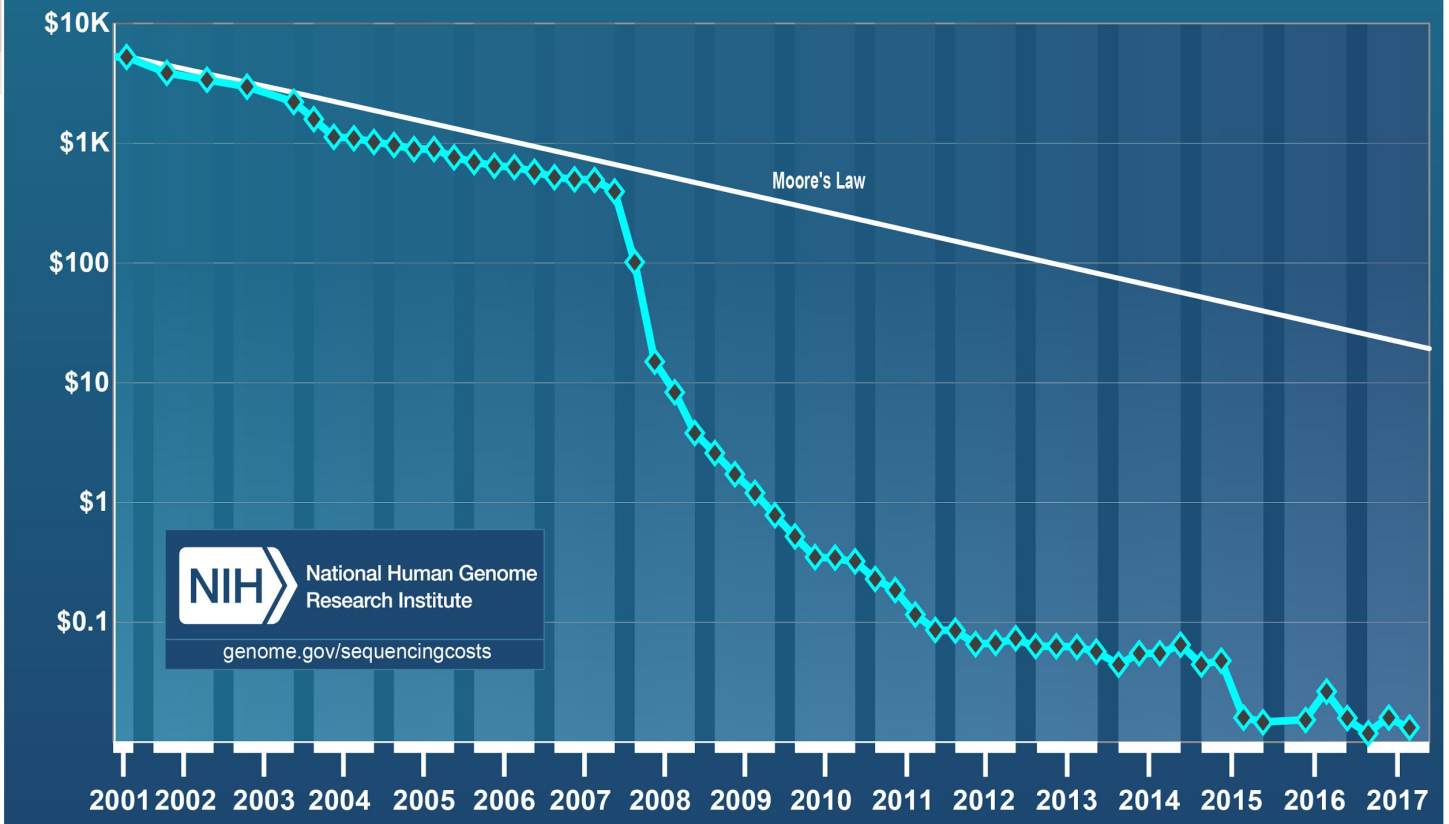
Sequencer Chemistry has improved

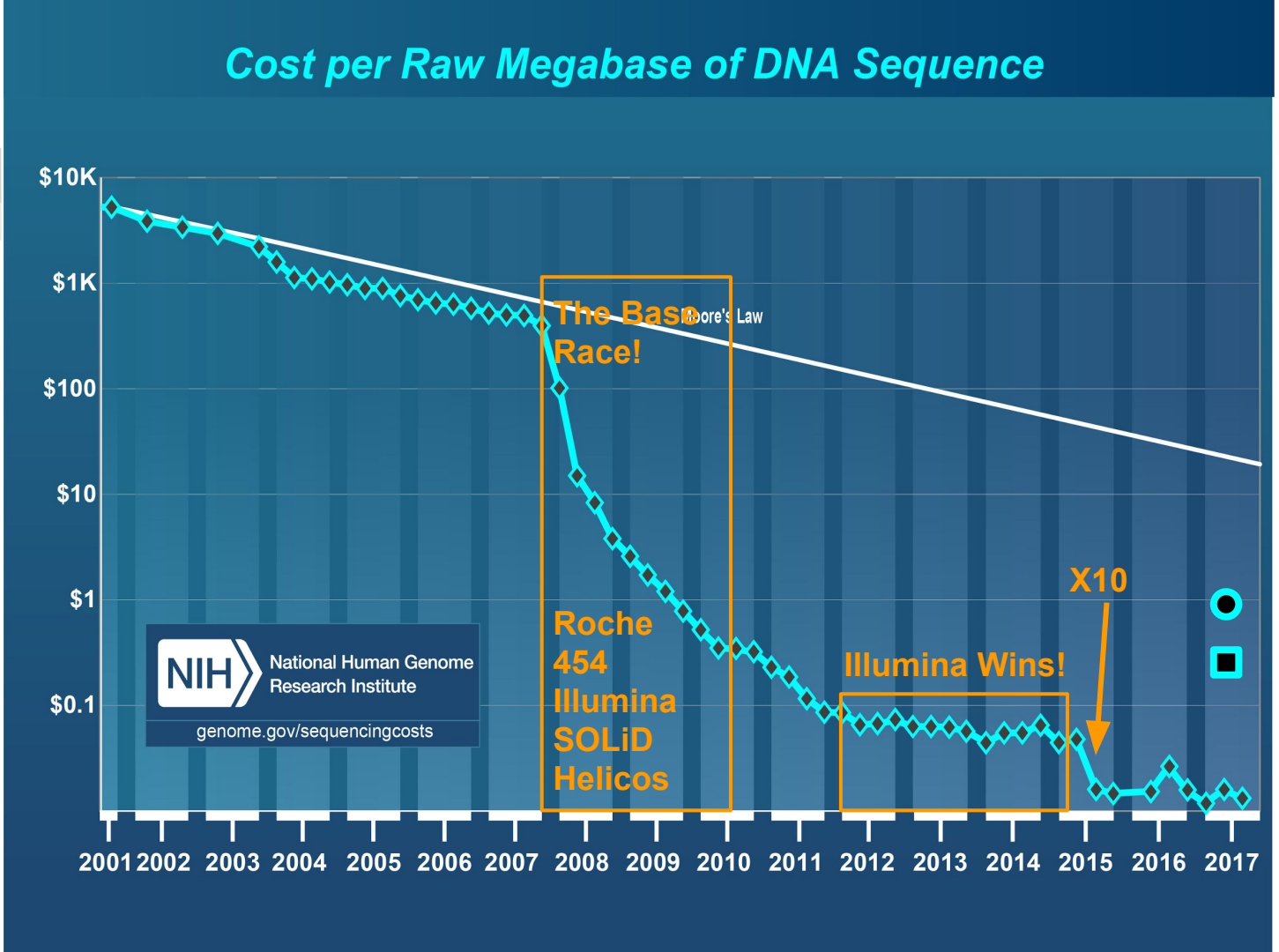




Hoecke et al., Scientific Reports 6, Article number: 26314 (2016)

Cost per Raw Megabase of DNA Sequence




Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
The Base Race!
Roche 454
Illumina SOLiD
Helicos
Illumina Wins!
X10
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Nanopore
PacBio

With Great Quality Comes Great Cost

Current Illumina Machines

Non-Illumina Sequencing for ChIP/ATAC

Sequencing Experiment Design

- Experimental Considerations
  - Special needs
  - Is an input control necessary
  - How many replicates?
- Library Preparation
  - How long should fragments be?
  - Paired or single end?
  - How long should my reads be?
- Sequencing Considerations
  - What platform should be used
  - How deep should I sequence?

Experimental Considerations: Special Needs

- Are there special adapters?
  - Non-standard adapters require special library prep and a sequencing facility that's willing to accommodate your needs.
- Is the sample non-"standard" material?
  - Too short/Too long (50 > length < 500)
  - Low Quality (e.g. ancient DNA)
  - GC rich/Highly repetitive
- How much starting material is there?
  - Is there enough material for the assay?

Experimental Considerations: Library Preparation

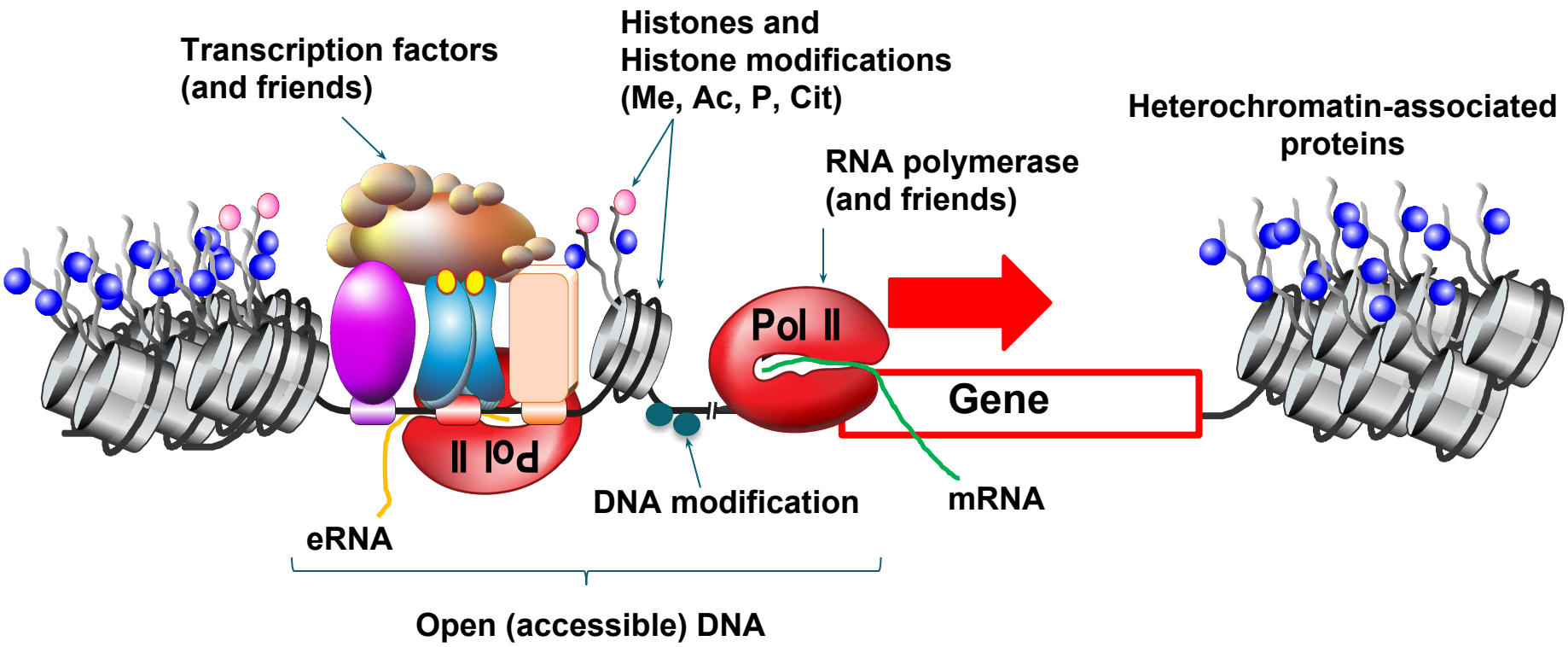
- How long should fragments be?
  - ChIP-seq -- 150-200bp
  - ATAC-seq -- 100-800bp
- Paired or single end?
  - ChIP-seq -- Single End
  - ATAC-seq -- Paired End
- How long should my reads be?
  - > 50bp

Experimental Considerations: Sequencing

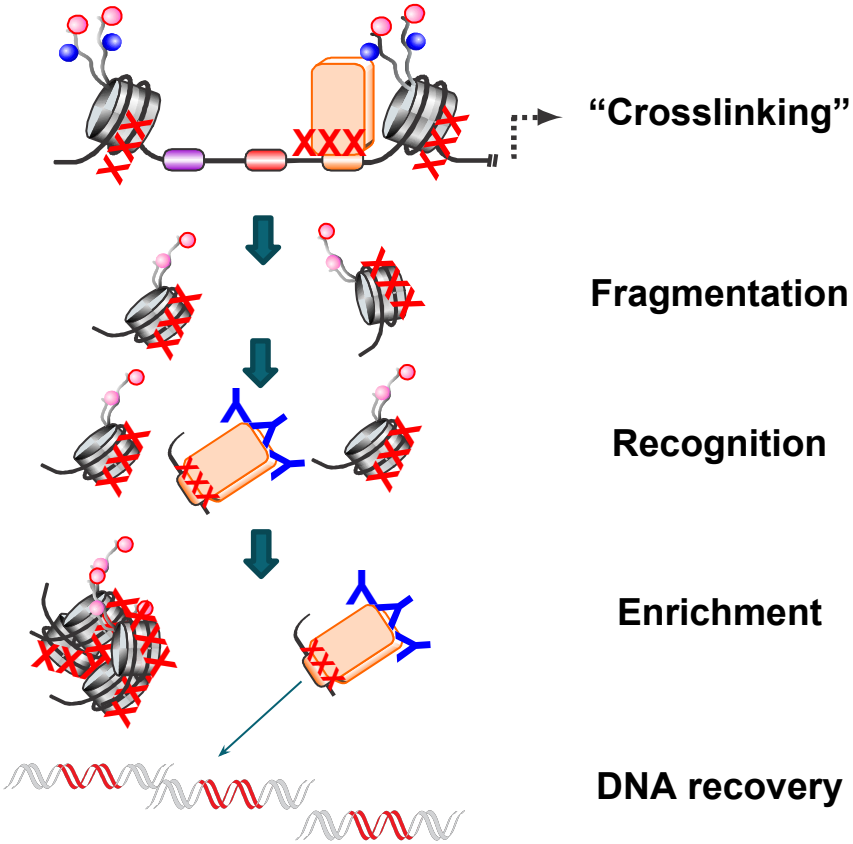
- What platform should I use?
  - For ChIP/ATAC-seq there is no need to use anything other than Illumina.
  - Other NGS based questions may require a specific platform
- How deep should I sequence?
  - ChIP - 15-25M single end reads
  - ATAC - 50-200M paired end reads
  - Use QC steps to determine if you need more.

Applications of Sequencing Technology: ChIP-seq/ATAC-seq

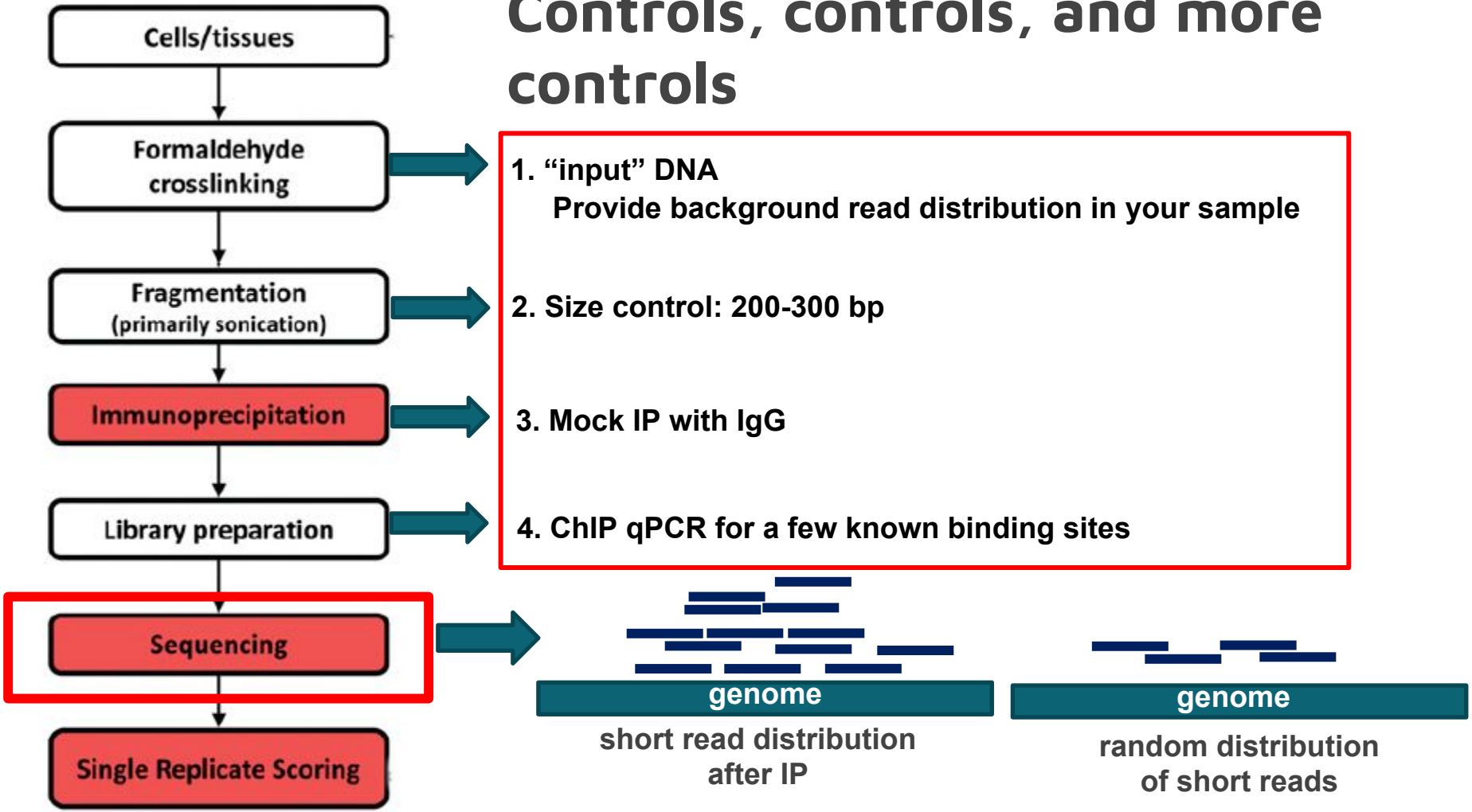
Identifying signals in chromatin



ChIP-seq: Identifying signals in chromatin

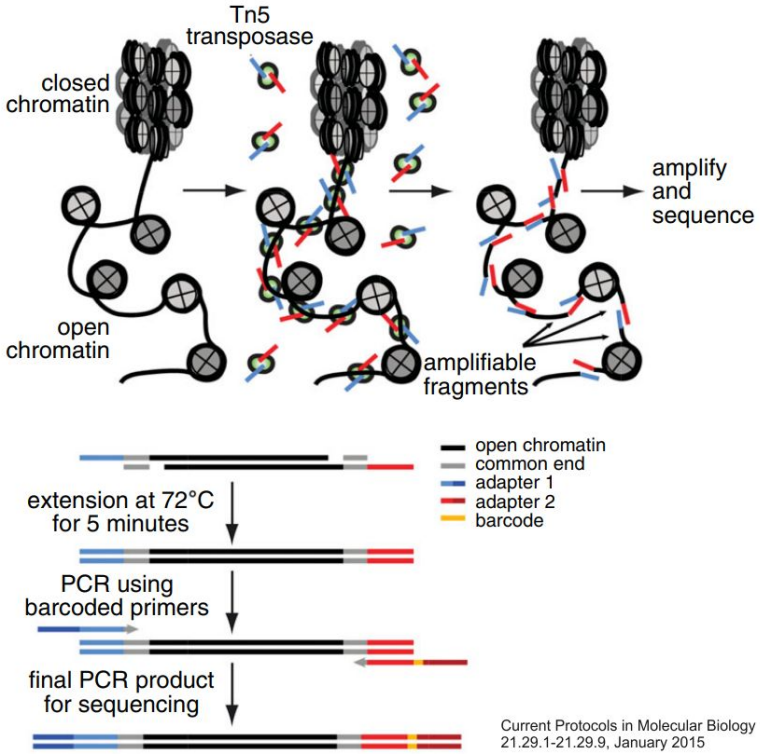


Controls, controls, and more controls



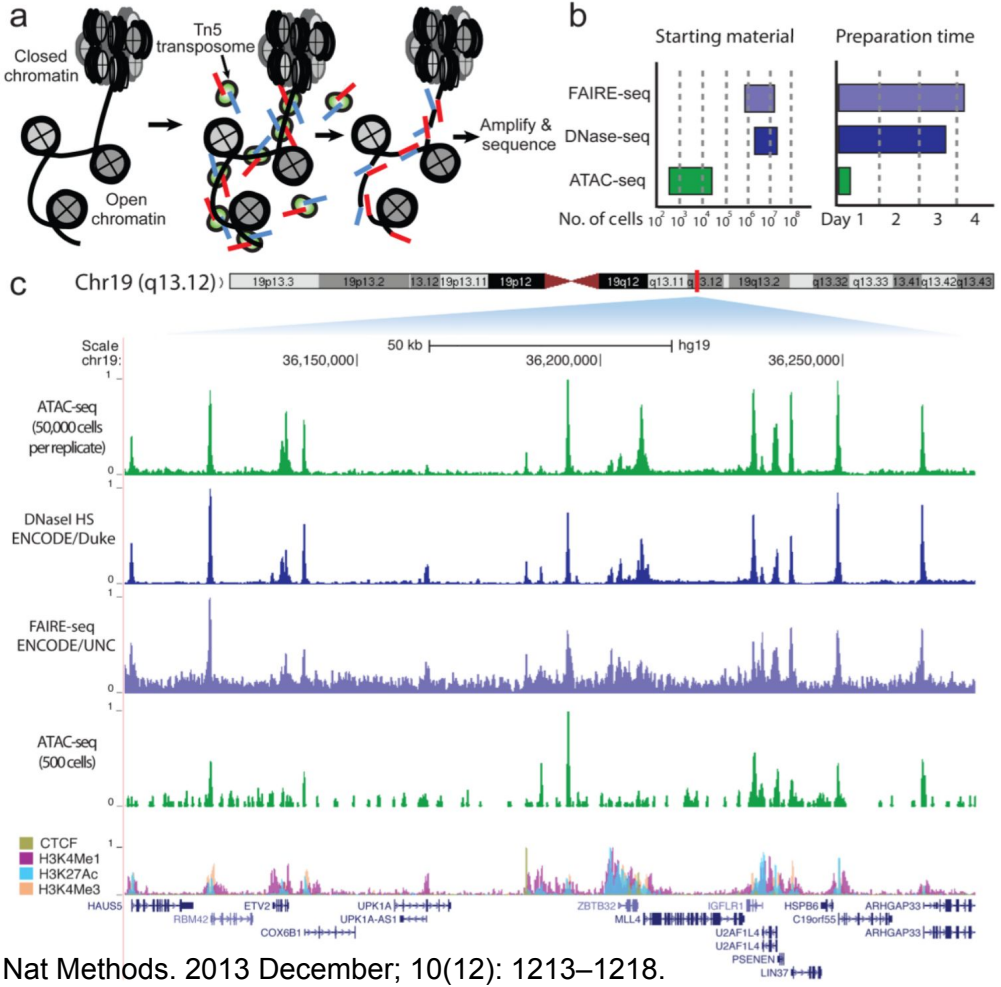
ATAC-seq

- Maps “open” chromatin genome-wide
- DNA Fragmentation / incorporation of sequencing adaptors in one step



Why ATAC-seq?

- Advantages:
  - More information
  - Less time
  - Less material
  - No sonication!
  - No need for “good quality antibody”
- Disadvantages
  - More Information

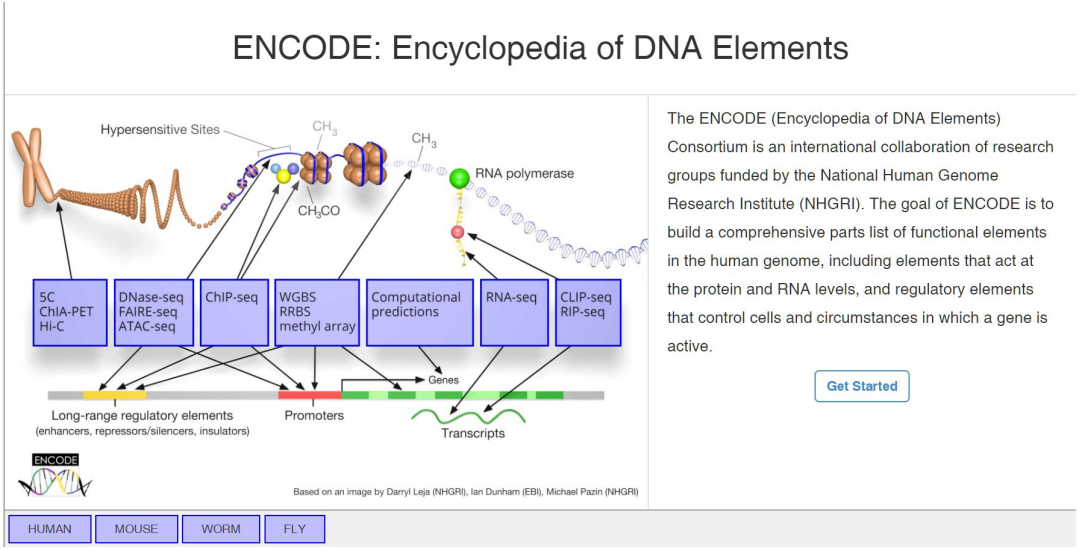


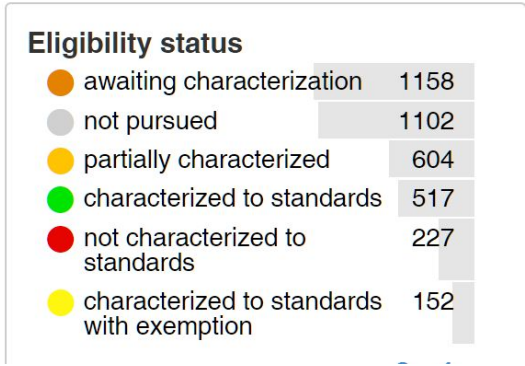
Nat Methods. 2013 December; 10(12): 1213–1218.

Caveats to ChIP/ATAC-seq

- False positive rate for peak detection is high
- Variation between peak detection methods is HIGH
- Association between binding and function is LOW
- The cost of the experiment is HIGHER than you plan
- Optimization requires MUCH MORE time than you plan
- The bioinformatics takes more time than you expect

Encode: Resource for Protocols, Workflows, and Data



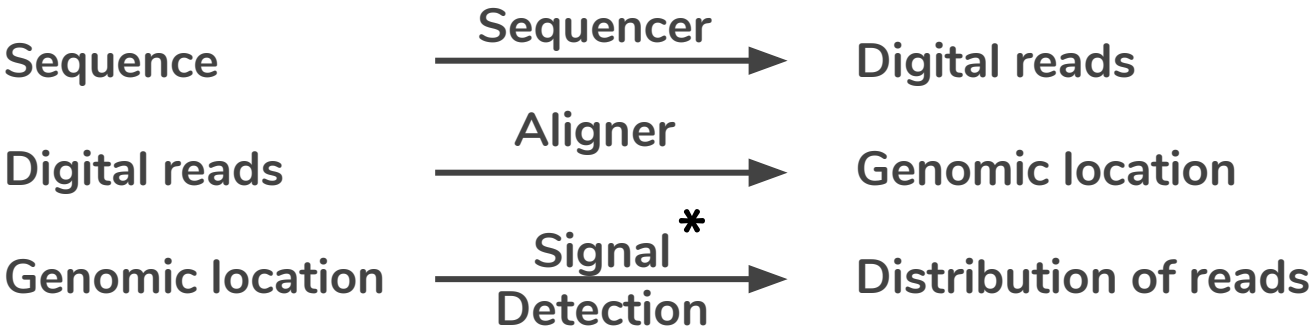


https://www.encodeproject.org/search/?type=AntibodyLot

End Part 1: Break

Sequencing Pipeline: Reads to Data

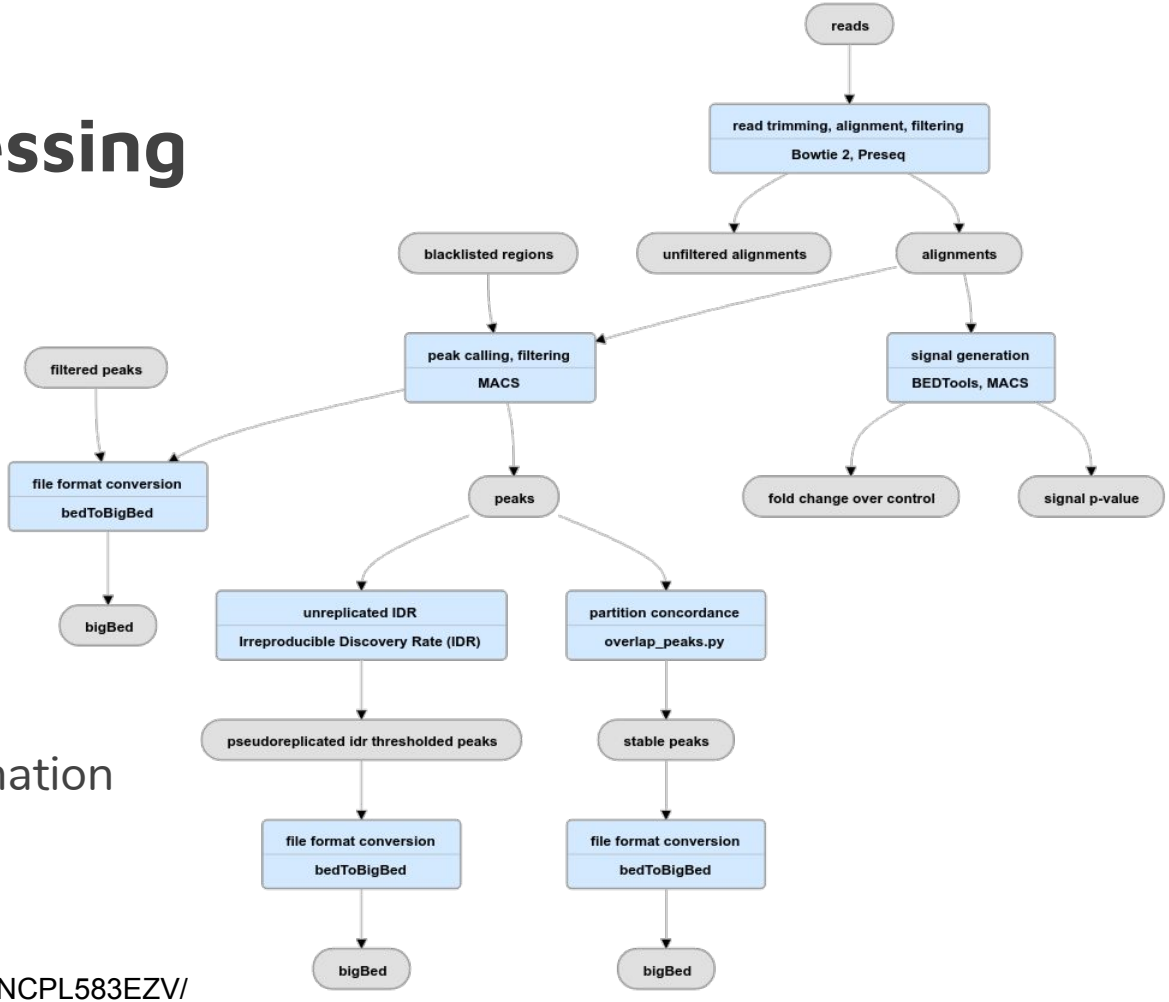
Downstream Terminology



*** Specifically, in ChIP/ATAC-seq this is called peak calling and the read distribution is called a peak.**

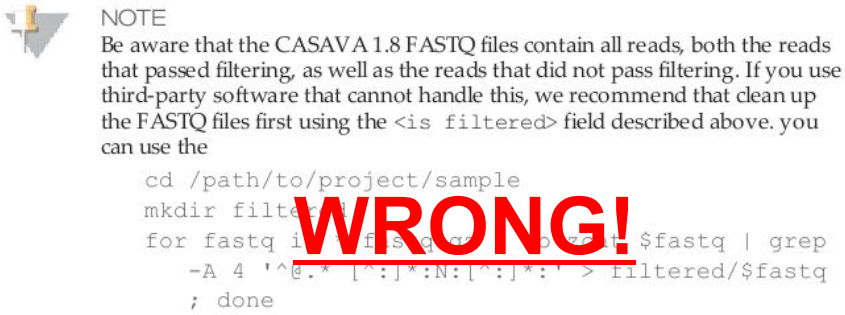
Read Processing

1. Adapter Trimming
2. Read QC
3. Read Alignment
4. Alignment QC
5. Peak Calling
6. Peak QC
7. Peak Confidence Estimation
8. Peak Visualization



https://www.encodeproject.org/pipelines/ENCPL583EZV/

Read Processing: Filtering



NOTE

Be aware that the CASAVA 1.8 FASTQ files contain all reads, both the reads that passed filtering, as well as the reads that did not pass filtering. If you use third-party software that cannot handle this, we recommend that clean up the FASTQ files first using the <is filtered> field described above. you can use the

```
cd /path/to/project/sample
mkdir filtered
for fastq in *fastq.gz ; do echo $fastq | grep
   -A 4 '^@.* [^:]*:N:[^:]*:' > filtered/$fastq
   ; done
```

WRONG!

1. grep:

```
$ cat input.fq | grep -A 3 '^@.* [^:]*:N:[^:]*:' | grep -v "^--$"
> output.fq
```

2. Fastq_illumina_filter
   a. http://cancan.cshl.edu/labmembers/gordon/fastq_illumina_filter/

Read Processing: Adapter Trimming

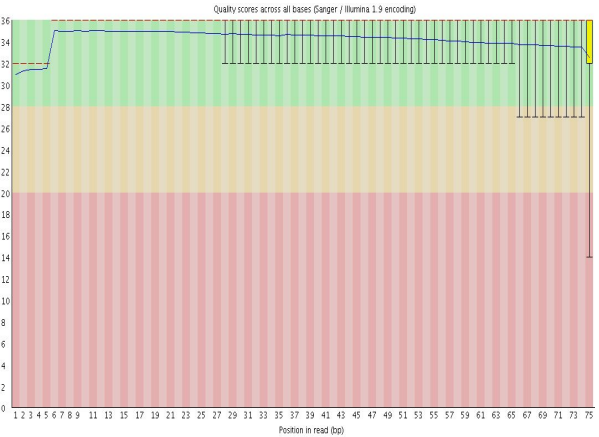
1. Cutadapt
    a. Auto-detect adapters
    b. Handles short adapters well
    c. Doesn't keep searching
2. Trimmomatic
    a. No auto-detection
    b. Keeps trimming as long as it finds adapter sequence
3. **Trim Galore! (recommended)**
    a. Wrapper for Cutadapt and FastQC
    b. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Read Processing: FastQC

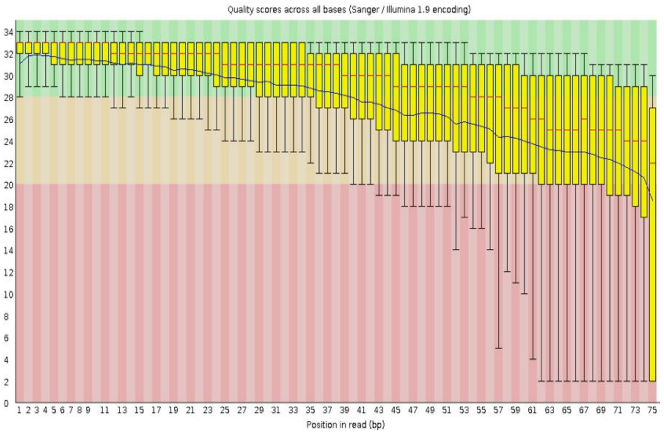
1. FastQC
    a. May be the only thing that (almost) everyone agrees on (https://www.biostars.org/p/204261/).

FastQC: Per-Base Quality Score

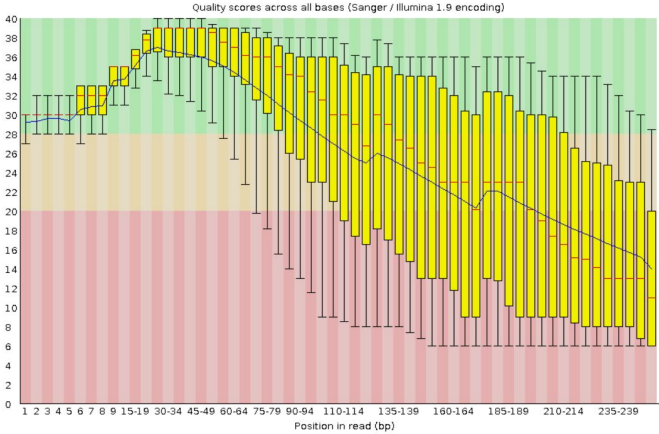
Good



Bad

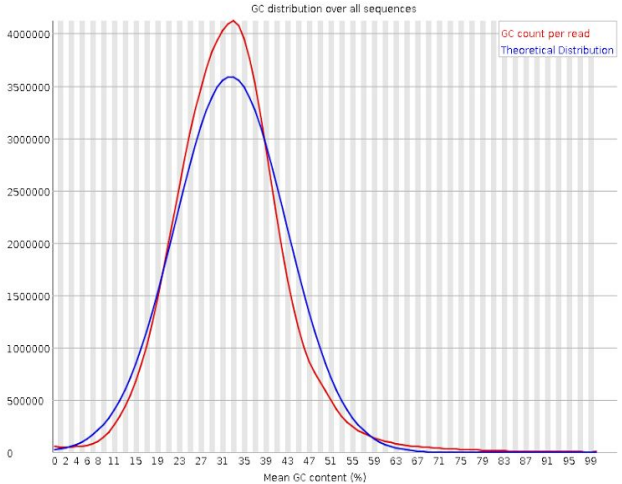


Ugly

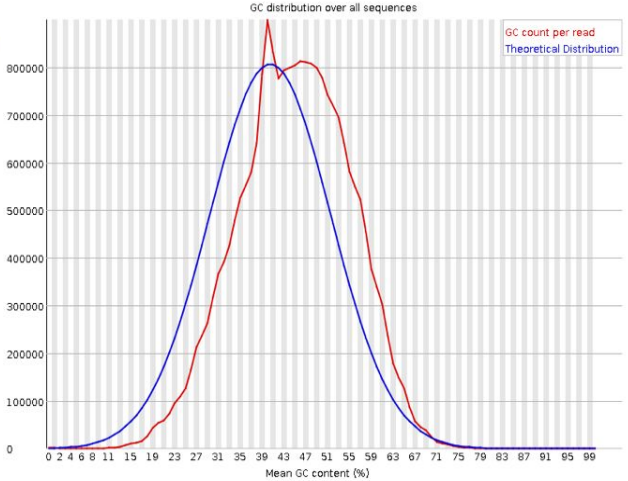


FastQC: Average GC Content

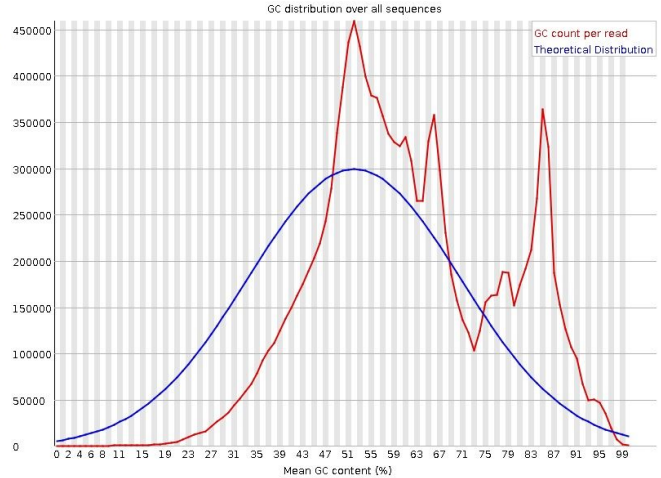
Good



Bad

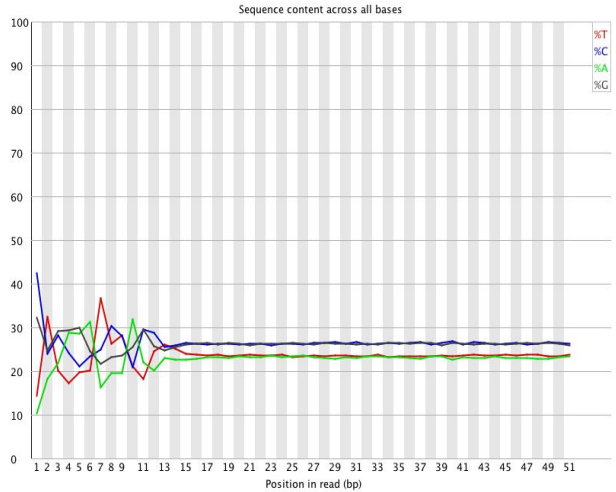


Ugly

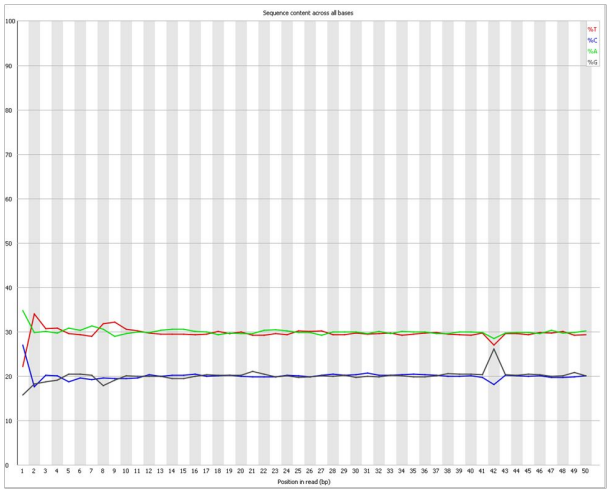


FastQC: Per-Base Nucleotide Content

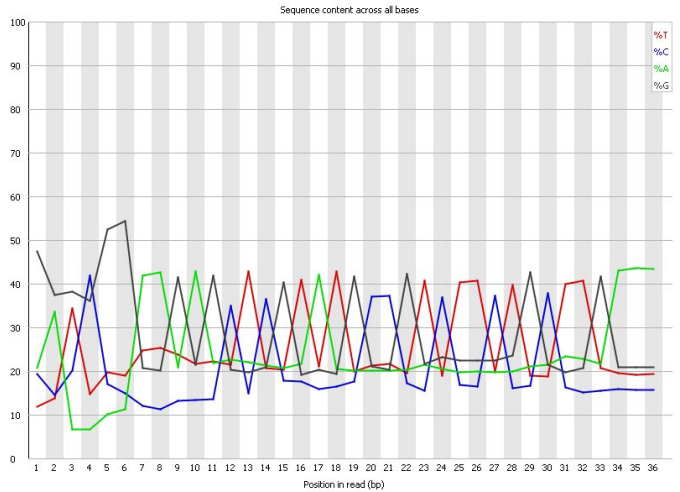
Good



Bad

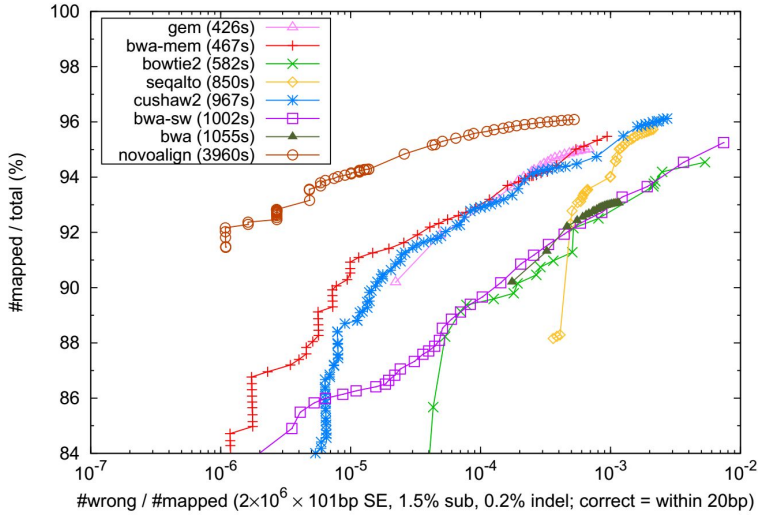


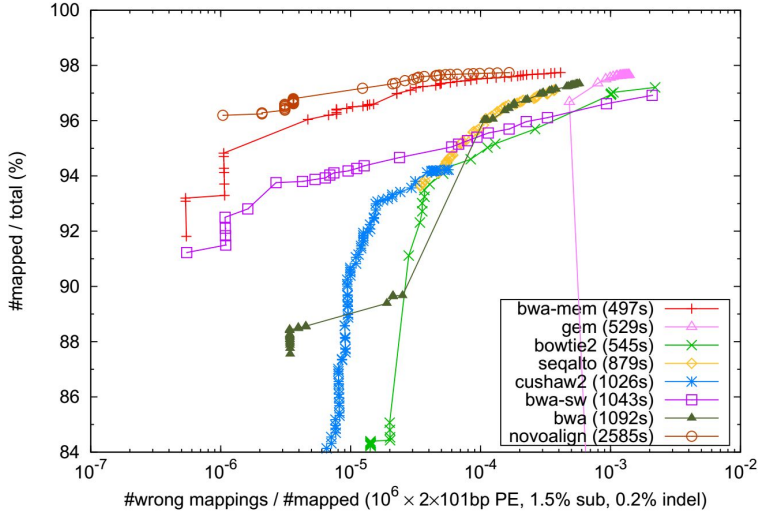
Ugly



Read Alignment

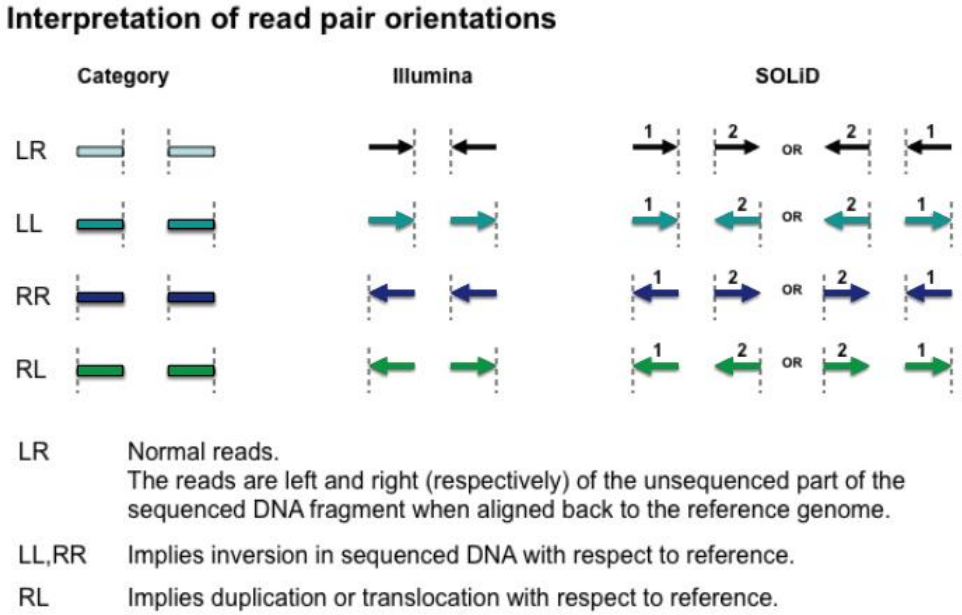
1. Bowtie2
    a. Most popular
2. BWA
    a. Sensitive but SLOW
3. ELAND (Illumina)
    a. Core's often use this
4. Novoalign
    a. Proprietary
5. STAR
    a. Careful!
    b. Need to prevent spliced alignment





Read Alignment: QC

```
## alignment summary
##   18200175 (100.00%) were paired; of these:
##     497844 (2.74%) aligned concordantly 0 times
##     11291185 (62.04%) aligned concordantly exactly 1 time
##     6411146 (35.23%) aligned concordantly >1 times
##     ----
##     497844 pairs aligned concordantly 0 times; of these:
##       139262 (27.97%) aligned discordantly 1 time
##     ----
##     358582 pairs aligned 0 times concordantly or discordantly; o
##       717164 mates make up the pairs; of these:
##         162541 (22.66%) aligned 0 times
##         174764 (24.37%) aligned exactly 1 time
##         379859 (52.97%) aligned >1 times
## 99.55% overall alignment rate
```



Alignments: SAM Format

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0, 2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0, 2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Filtering Alignments

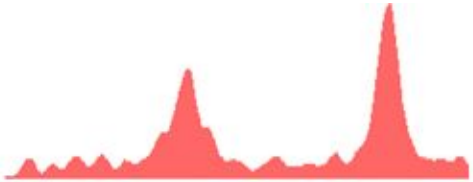
- Samtools
  - -f keeps all alignments with specific flag bits
  - -F excludes all alignments with specific flags
  - Decoding SAM flags:
    - https://broadinstitute.github.io/picard/explain-flags.html

Removing PCR Duplicates

- Samtools rmdup
  - Does not remove PCR duplicates for chromosomal translocation reads
- Picard Tools MarkDuplicates
  - Marks PCR duplicates and will remove them if requested
  - Removes translocation duplicates missed by samtools rmdup command.

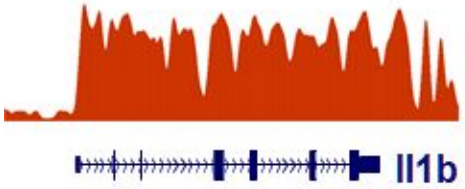
Peak calling: genomic read distribution

TF ChIP



Narrow Peaks

Histone Mod ChIP



Broad Peaks

ATAC

Mixed

Peak Callers

- MACS
  - ENCODE
  - Most widely used
  - Stable performance
- SPP
  - R implementation
- FSEQ
  - Kernel density based
  - ENCODE
  - Can be convinced to call peaks.

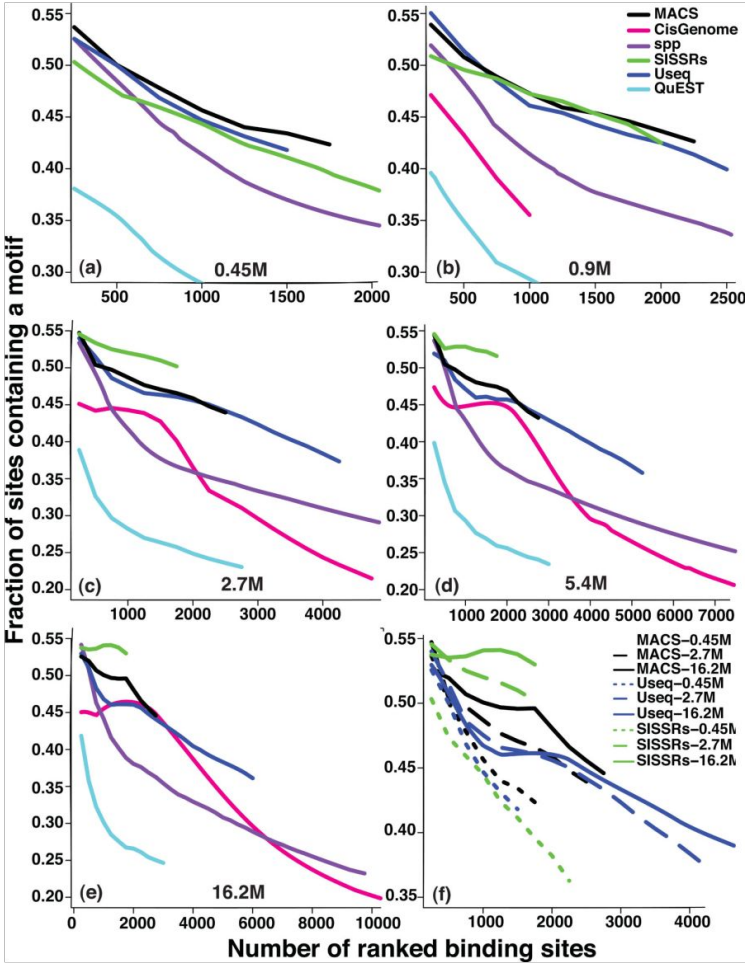
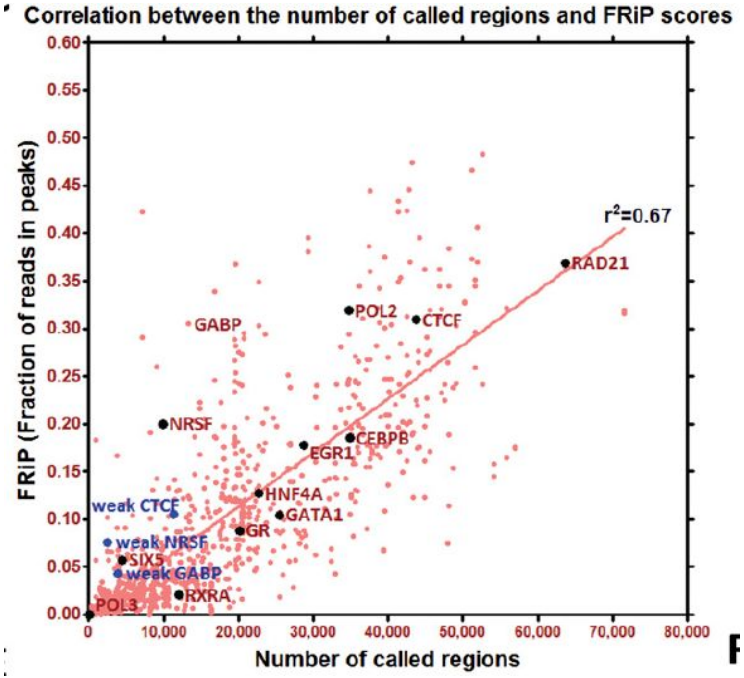


Figure 3. Quality of the Su(Hw) peaks
The fraction Su(Hw) peaks, identified by the indicated peak callers, that contains a Su(Hw) binding motif is plotted as a function of the number of top-ranked binding sites at the

Summary QC Reports with MultiQC

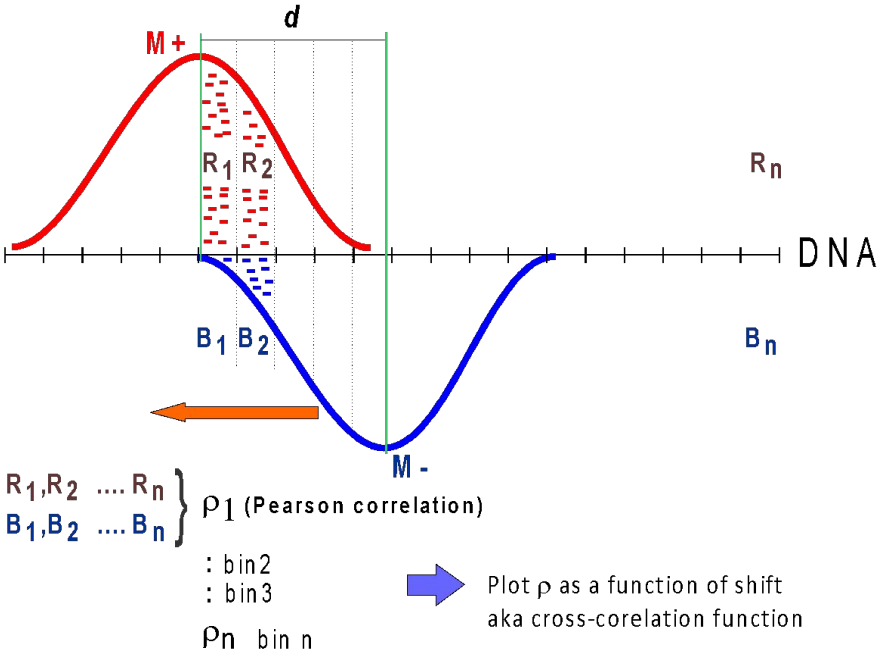
MultiQC Example

Quality Control: FRiP



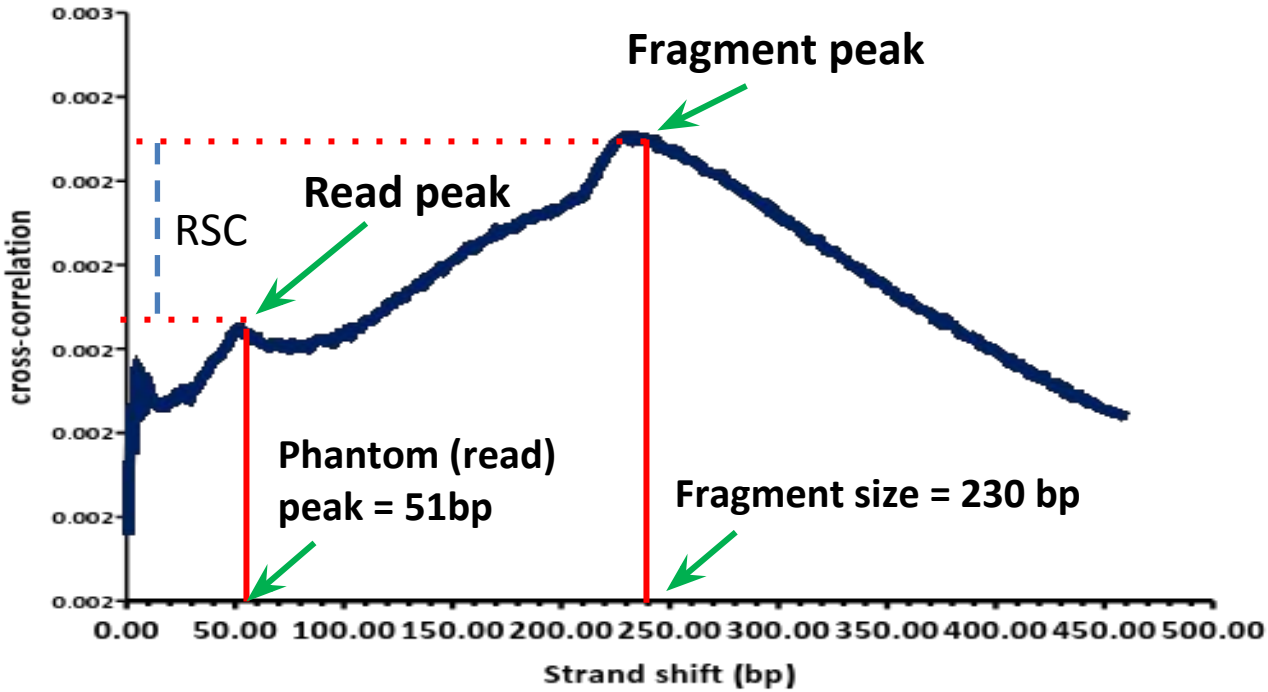
- FRiP = Reads in Peaks / Total Mapped Reads
- FRiP < 1% may be a cause for concern
- Correlated with peak number

Quality Control: Cross Correlation Analysis



- When the signal from the positive strand (red) matches the signal from the negative strand (blue) the correlation is maximized

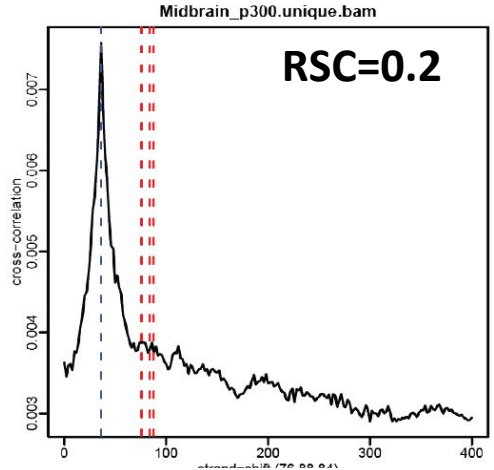
Cross Correlation Analysis Example



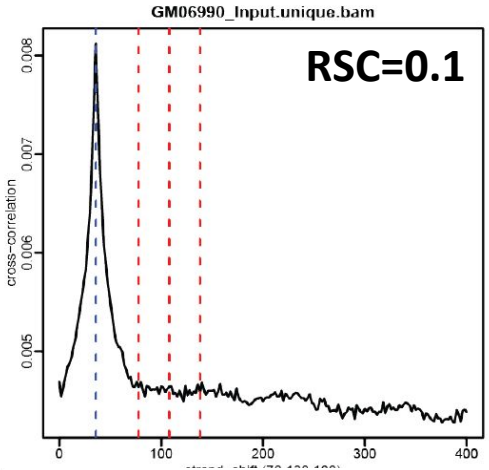
Relative strand correlation (RSC) = Fragment Peak/Read Peak = 1.617 (should be >0.8)

Cross Correlation analysis currently in repository

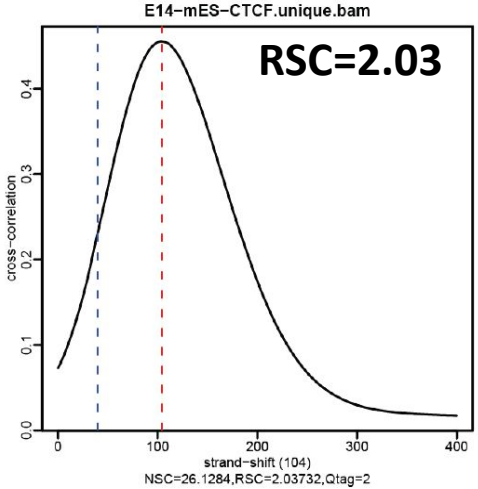
Bad datasets



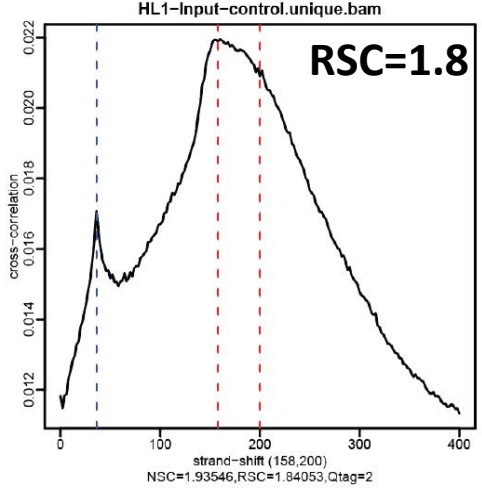
F



Good datasets



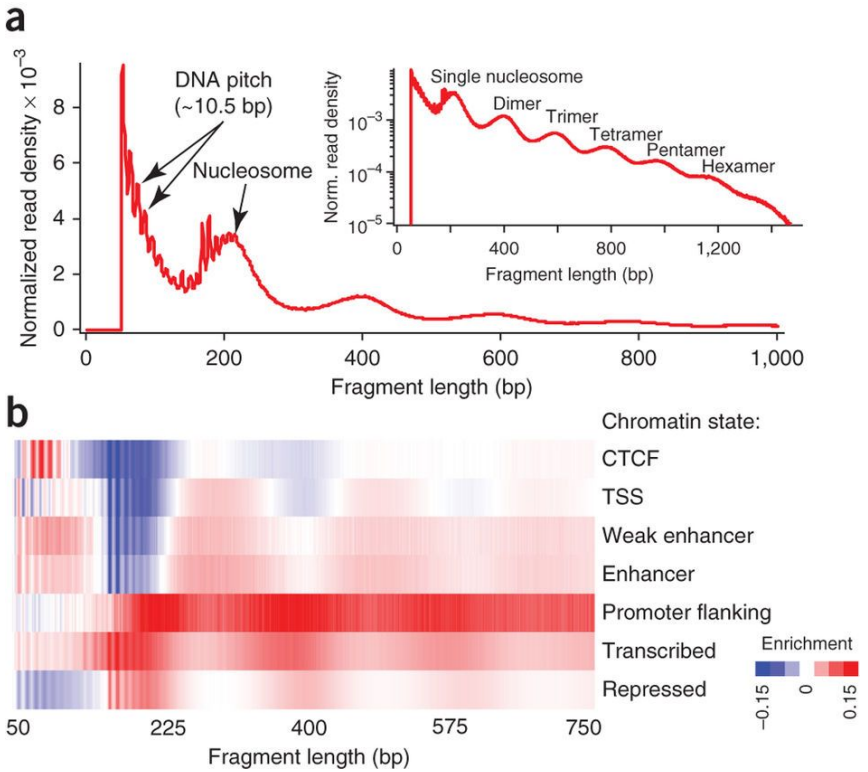
J



G3 (Bethesda). 2014 4(2):209-23.

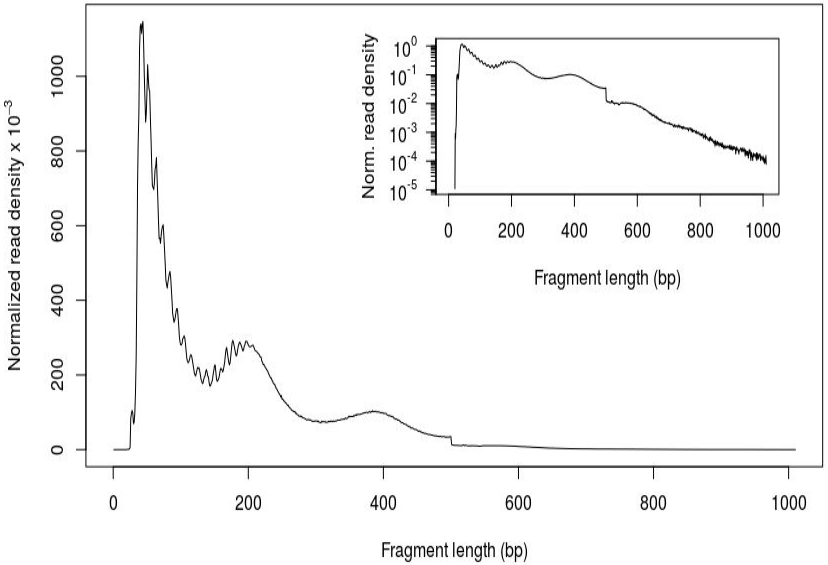
ATAC-seq Fragment Size Distribution

- Looking for distinct nucleosome profile.

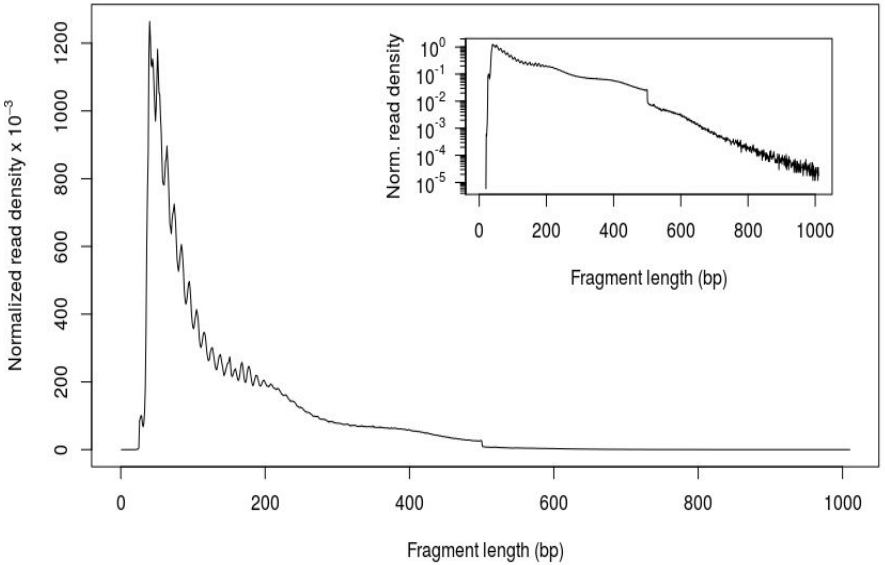


ATAC-seq Fragment Size Distribution

Good

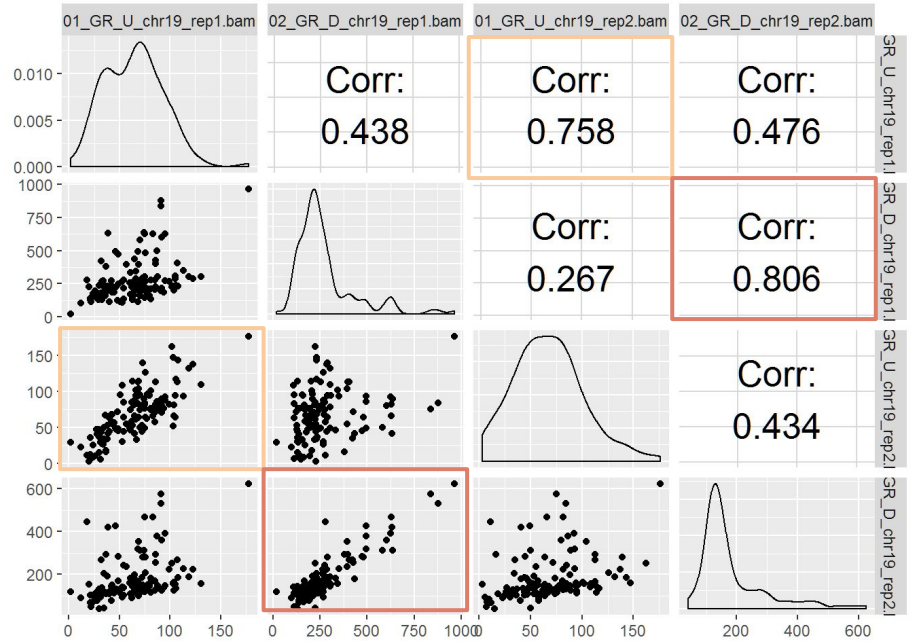


Poor



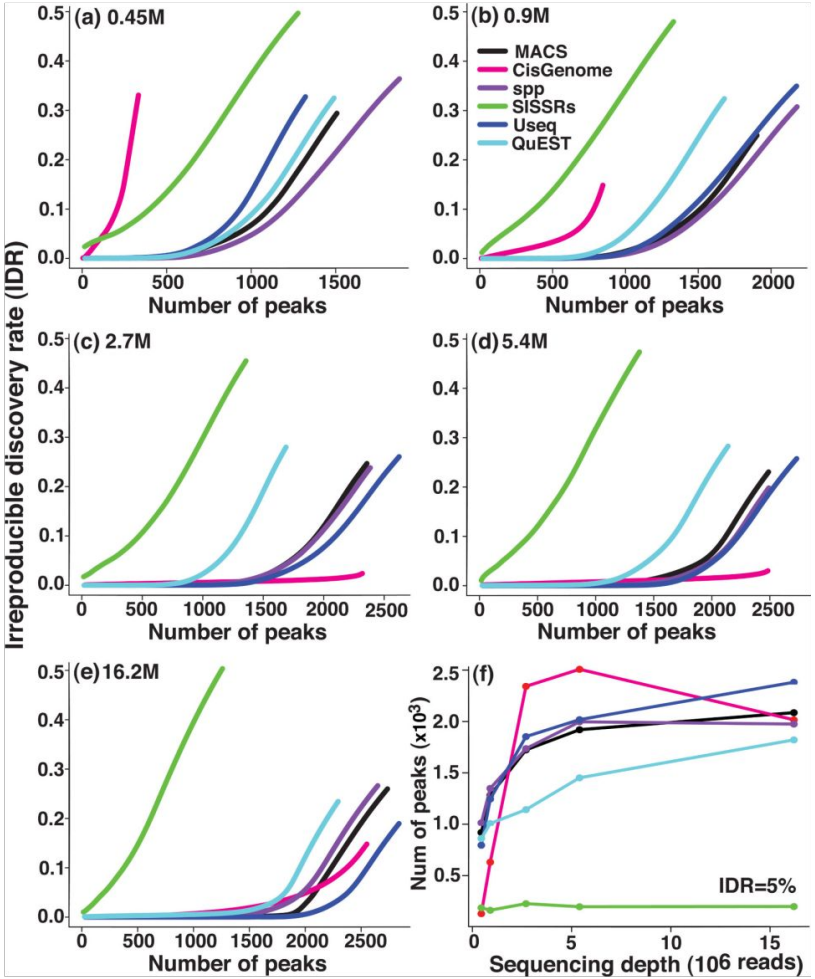
Correlation Between Read in Peaks

1. Union of peaks between replicates
2. Count reads in peaks
3. Calculate correlation



Irreproducible discovery rate (IDR)

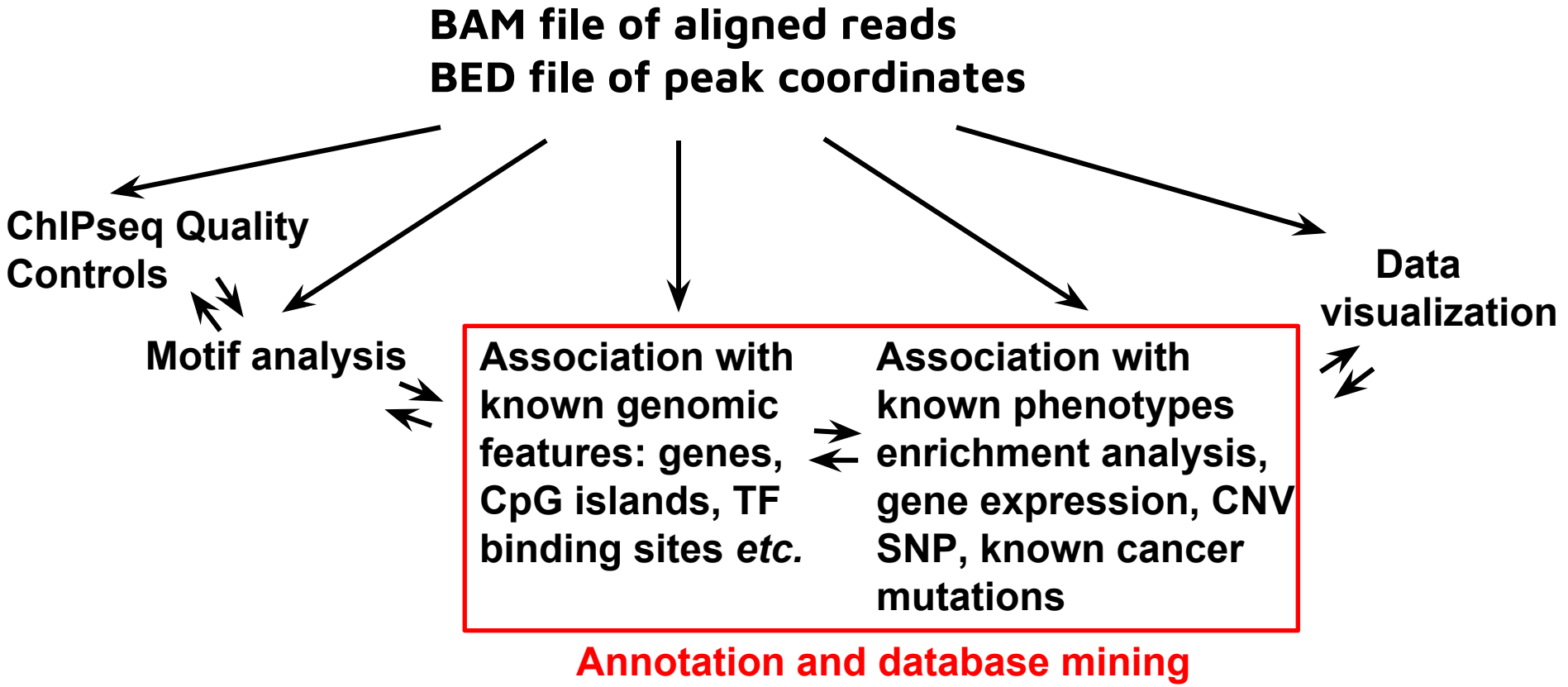
1. Peak reproducibility between replicates.
2. ENCODE uses it
3. Slightly better than Fisher's and Stouffer's methods.
4. Honestly, use one of these but the IDR program is finicky and the method is not defined for more than 2 replicates



End Part 2: Break

Functional Analysis: Data to Meaning

After peak calling:



DNA Motif discovery

Supervised

Compares peak sequences to known binding sites databases such as **JASPAR** (http://jaspar.genereg.net/) and **TRANSFAC**

- **Can only find "known" sites**
- **Relies on the quality of databases**
- **Can work with a small number of sequences**
- **Biased**

(Trust, but verify – there are mistakes and circular annotations)

Unsupervised

Doesn't care about binding sites.
Treats a collection of sequences as a text that might contain repeating words.

- **Slow**
- **Noisy**
- **Need large number of sequences**
- **Need to interpret motifs (which means comparing to databases)**
- **unbiased**

De novo Motif discovery

The MEME Suite

http://alternate.meme-suite.org/

http://xxmotif.genzentrum.lmu.de/

http://rsat.sb-roscoff.fr/index.html

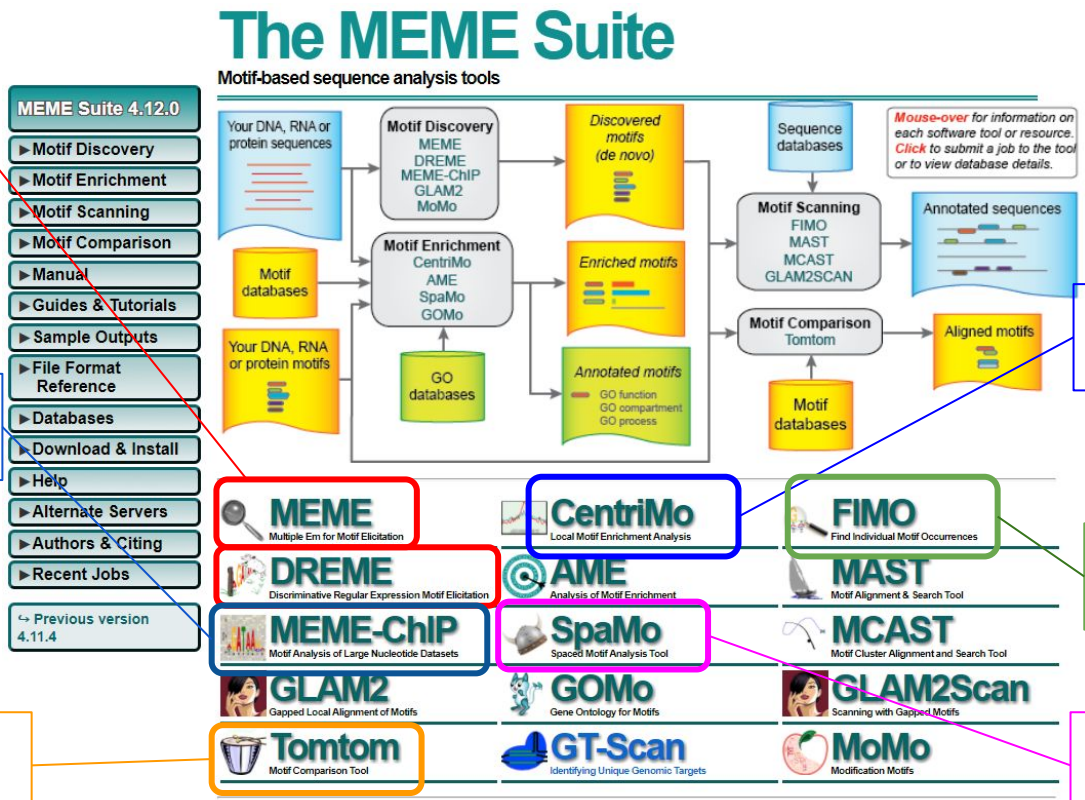
HOMER

http://homer.ucsd.edu/homer/index.html

De novo motif search overrepresentation with Gibbs sampling

Performs all analyses at once

Comparing identified motifs to known motifs



Motif positional enrichment

Tabulating motif occurrence

Co-occurrence of motifs

De novo motif prediction/overrepresentation
Typical MEME output

MEME-ChIP
Motif Analysis of Large Nucleotide Datasets

If you use MEME-ChIP in your research, please cite the following paper:
Philip Machanick and Timothy L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets", *Bioinformatics*, 2712, 1696-1697, 2011. [full text]

MOTIFS | PROGRAMS | INPUT FILES | PROGRAM INFORMATION | SUMMARY IN TSV FORMAT NEW! | MOTIFS IN MEME TEXT FORMAT NEW!

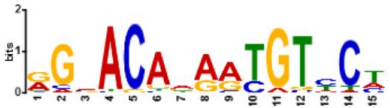
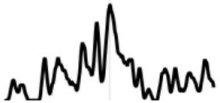
MOTIFS

The significant motifs (E-value ≤ 0.05) found by the programs MEME, DREME and CentriMo; clustered by similarity and ordered by E-value.

Expand All Clusters Collapse All Clusters

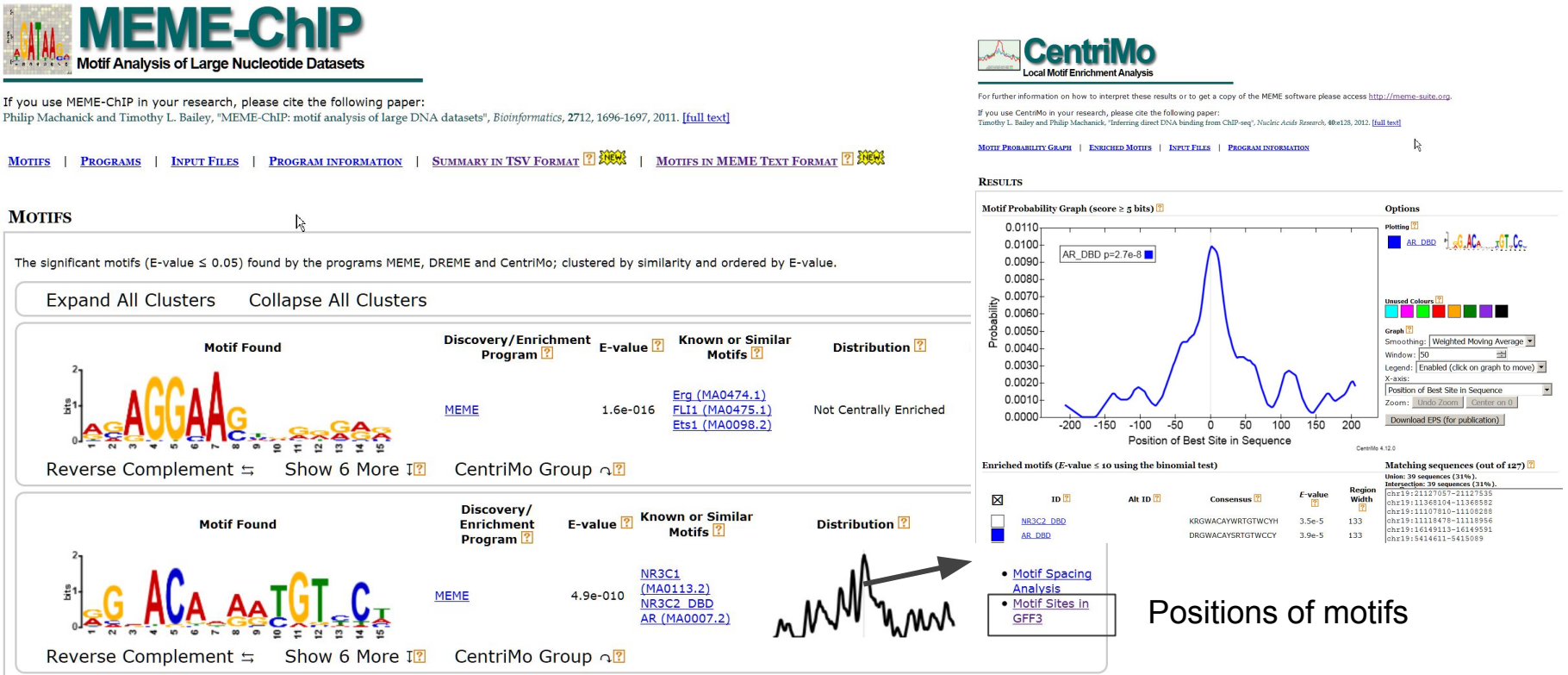
Motif Found	Discovery/Enrichment Program	E-value	Known or Similar Motifs	Distribution	SpaMo & FIMO
	MEME	1.6e-016	Erg (MA0474.1) FLI1 (MA0475.1) Ets1 (MA0098.2)	Not Centrally Enriched	• Motif Spacing Analysis • Motif Sites in GFF3

Reverse Complement ⇆ Show 6 More ↧ CentriMo Group ↷

Motif Found	Discovery/ Enrichment Program	E-value	Known or Similar Motifs	Distribution	SpaMo & FIMO
	MEME	4.9e-010	NR3C1 (MA0113.2) NR3C2_DBD AR (MA0007.2)		• Motif Spacing Analysis • Motif Sites in GFF3

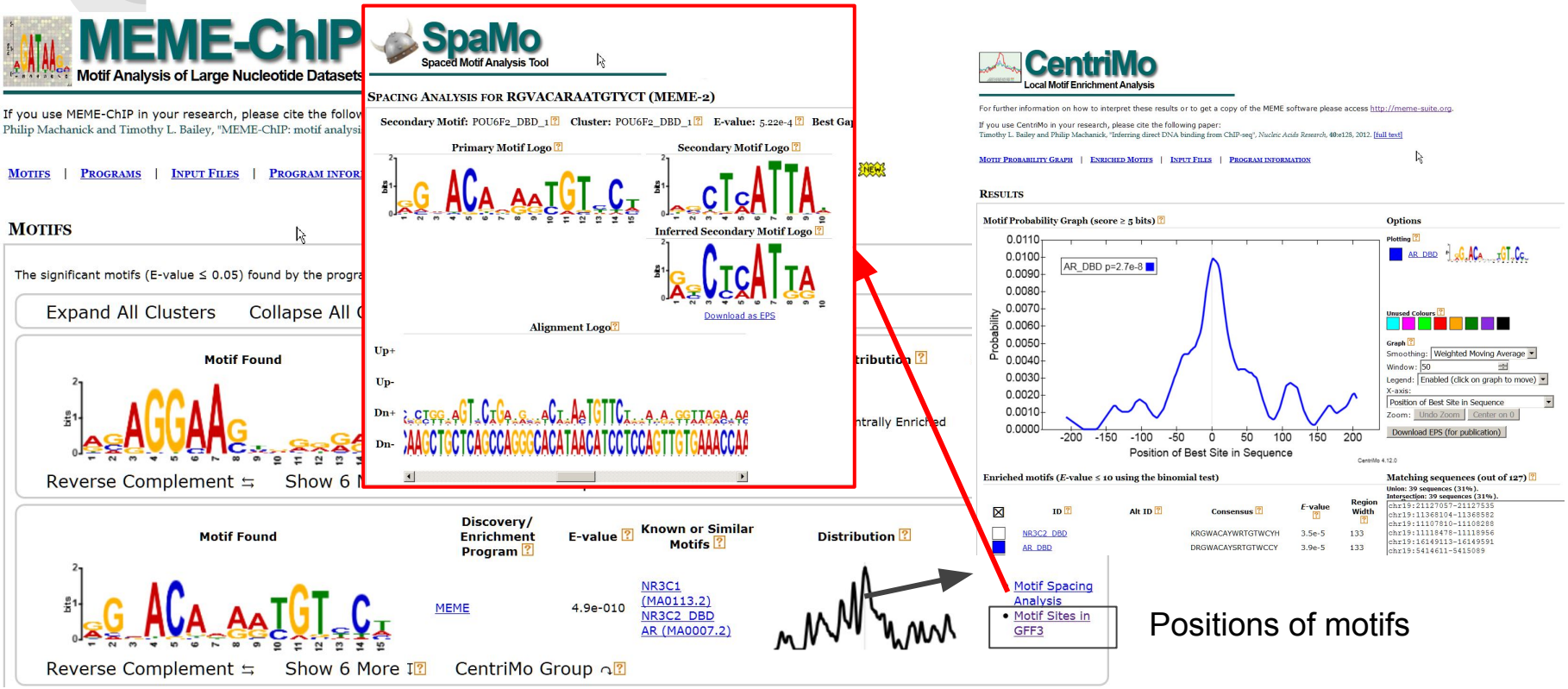
Reverse Complement ⇆ Show 6 More ↧ CentriMo Group ↷

De novo motif prediction/overrepresentation
Typical MEME output



Positions of motifs

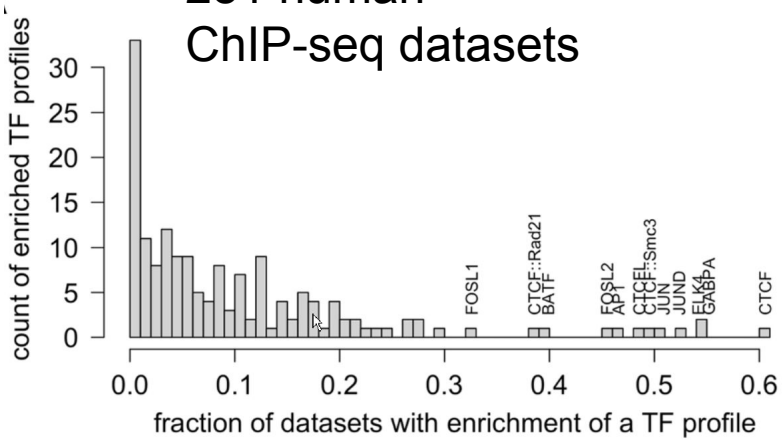
De novo motif prediction/overrepresentation
Typical MEME output

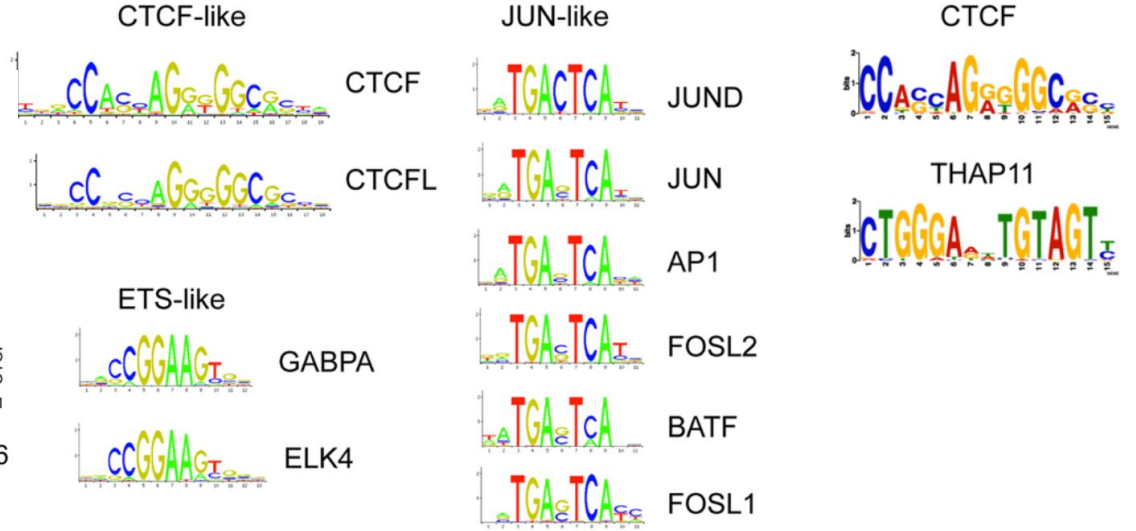


Positions of motifs

My program detected motif X in my peaks. It is a "real" binding site, right?

281 human
ChIP-seq datasets





… may be?

Summit enrichment?
Site complexity?
Correlation with peak scores?
Distance conservation relative to other sites

Genome Biol. 2014 Jul 29;15(7):412.

Motif overrepresentation – data interpretation

Co-occurrence of motifs does not mean co-function

Worsley Hunt R, Wasserman WW. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. Genome Biol. 2014 Jul 29;15(7):412. PMID: 25070602;

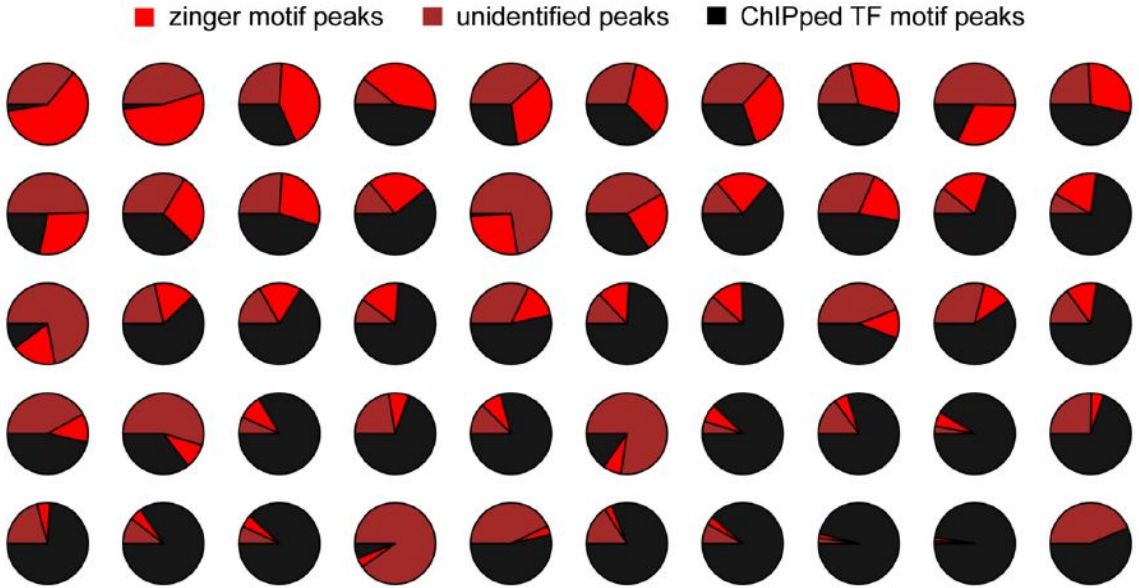


Figure 3 The fraction of zinger motif peaks and ChIPped TF motif peaks varies across ChIP-seq datasets. The pie charts present a random selection of 50 datasets for multiple TFs and cell-lines with zinger motifs present (>1% zinger). The charts are ordered by greatest zinger motif peak enrichment to the least. Black is the portion of peaks with the ChIPped TF's motif, red is the portion of zinger motif peaks, and brown is the remaining portion of peaks that do not contain either the ChIPped TF nor zinger motifs.

My program claims that the site enriched in my peak is a binding site for factor X. Is it?

JASPAR 2018

	ID	Name	Species	Class	Family	Logo
	MA0039.1	Klf4	Mus musculus	C2H2 zinc finger factors	Three-zinc finger Kruppel-related factors	
	MA0039.2	Klf4	Mus musculus	C2H2 zinc finger factors	Three-zinc finger Kruppel-related factors	
	MA0039.3	KLF4	Homo sapiens	C2H2 zinc finger factors	Three-zinc finger Kruppel-related factors	

Copy CSV

My program claims that the site enriched in my peak is a binding site for factor X. Is it?

JASPAR 2018

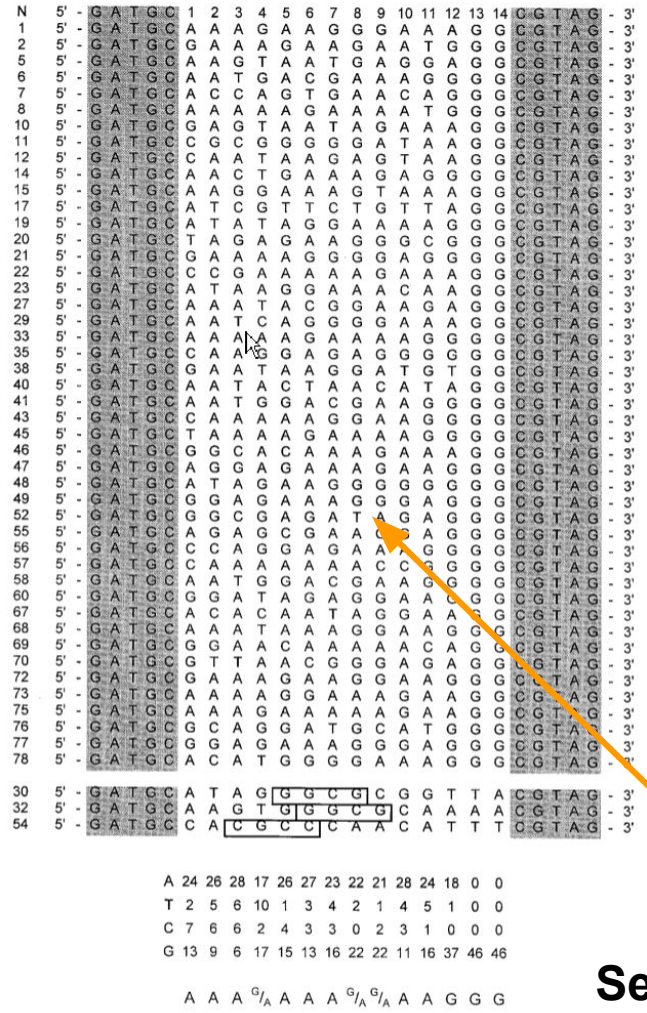
ID	Name	Family	Logo
MA0039.1	Klf4	Three-zinc finger Kruppel-related factors	
MA0039.2	Klf4	Three-zinc finger Kruppel-related factors	
MA0039.3	KLF	Three-zinc finger Kruppel-related factors	

Copy CSV

Profile summary Add

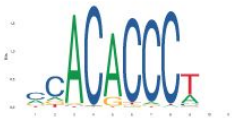
Name:	Klf4
Matrix ID:	MA0039.1
Class:	C2H2 zinc finger factors
Family:	Three-zinc finger Kruppel-related factors
Collection:	CORE
Taxon:	Vertebrates
Species:	Mus musculus
Data Type:	SELEX
Validation:	9443972
Uniprot ID:	Q60793

My program claims that the site enriched in my peak is a binding site for factor X. Is it?



Profile summary	Add
Name:	Klf4
Matrix ID:	MA0039.1
Class:	C2H2 zinc finger factors
Family:	Three-zinc finger Kruppel-related factors
Collection:	CORE
Taxon:	Vertebrates
Species:	Mus musculus
Data Type:	SELEX
Validation:	9443972
Uniprot ID:	Q60793

JASPAR 2018

ily	Logo
ee-zinc er Kruppel-ted factors	
ee-zinc er Kruppel-ted factors	
ee-zinc er Kruppel-ted factors	

Selected sites were not aligned!

My program claims that the site enriched in my peak is a binding site for factor X. Is it?

Circular References

Experimental binding site for a gene family member...

↓

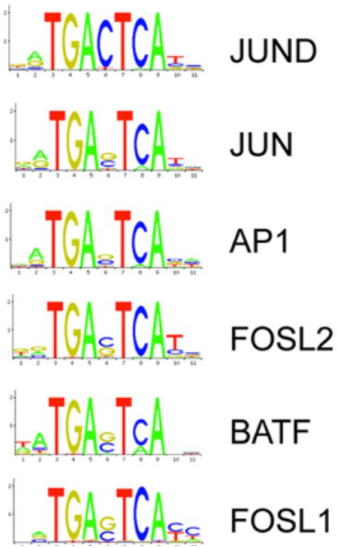
is cited in a review on entire family with a comma-separated list of family members

↓

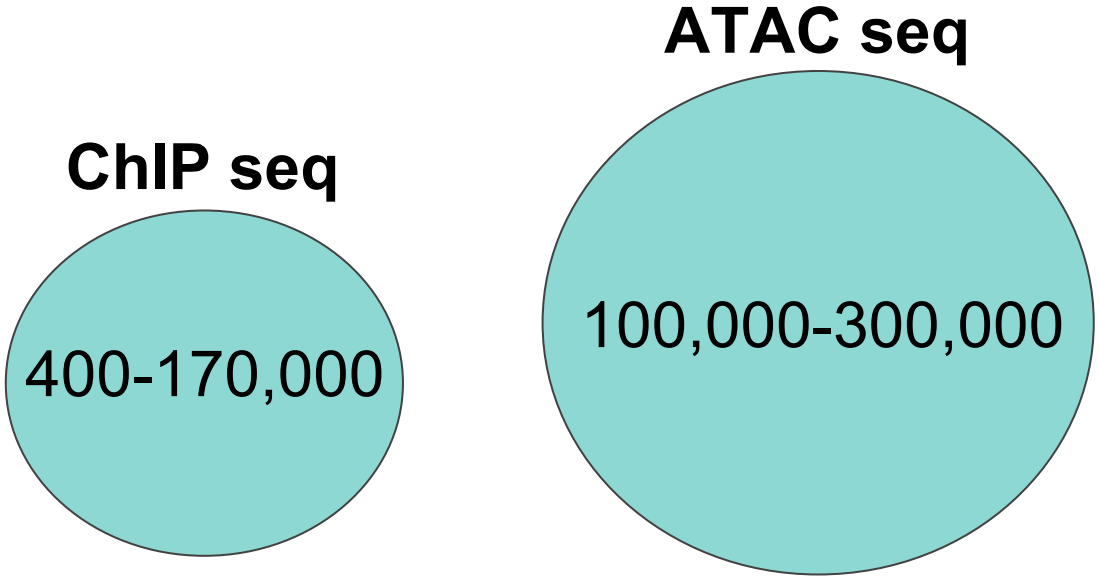
which creates new entry in a database for a new site for all other family member

Gene family members with similar sites

JUN-like



Peak annotation



Peak annotation

1. Descriptive: where binding events occur relative to known genomic features (genes, TSS, transcripts, CpG islands, binding sites of other transcription factors, any other “locations” of your choosing)

   Easy – overlay your peak coordinates with those of known genome features

2. Predictive : How does a binding event near feature X affects the function of this feature or any other phenotype of interest?

   Difficult

Predictive Peak annotation

- **Noise and low-affinity interactions**
- **Correlation between binding events and transcriptional response is often lacking <u>in a point expression study</u>.**

  " On average, **14.7%** of genes bound by a factor were differentially expressed following the knockdown of that factor, suggesting that most interactions between TF and chromatin do not result in measurable changes in gene expression levels of putative target genes. "

  Cussanovich et all. (2014) PlosGenetics 10(3): e1004226

- **Correlation between the number of peaks per gene and expression does exist**
- **Correlation between specific histone modifications and expression exists, but is it quantitative?**

Peak Annotation

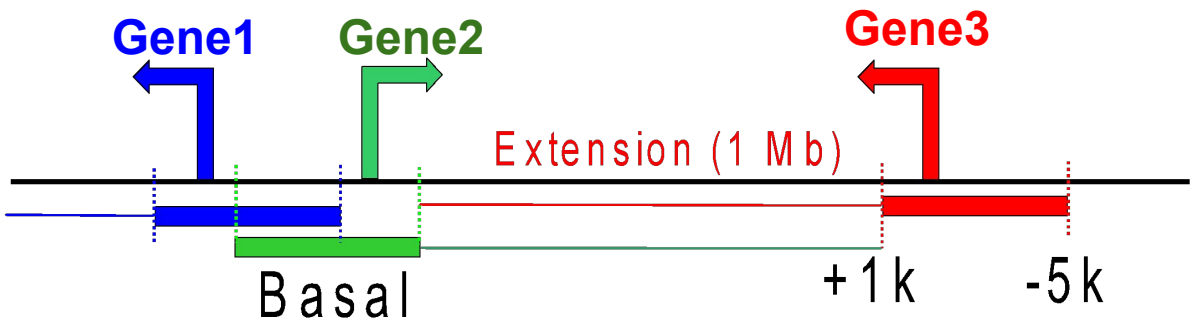
- “Naïve” TSS proximity – all peaks within X kb of a TSS

- “Augmented” TSS proximity – all peaks within X kb of TSS and
  - with a binding site of a factor in question
  - ranked by peak scores or p-values
  - overlapping other transcription factors binding sites
  - overlapping open chromatin
  - overlapping peaks of your target from different cells.
- All peak in a “regulatory region” belong to all genes in a regulatory region.
  - experimental topological domain
  - educated guess

- Distance-dependent stratification.

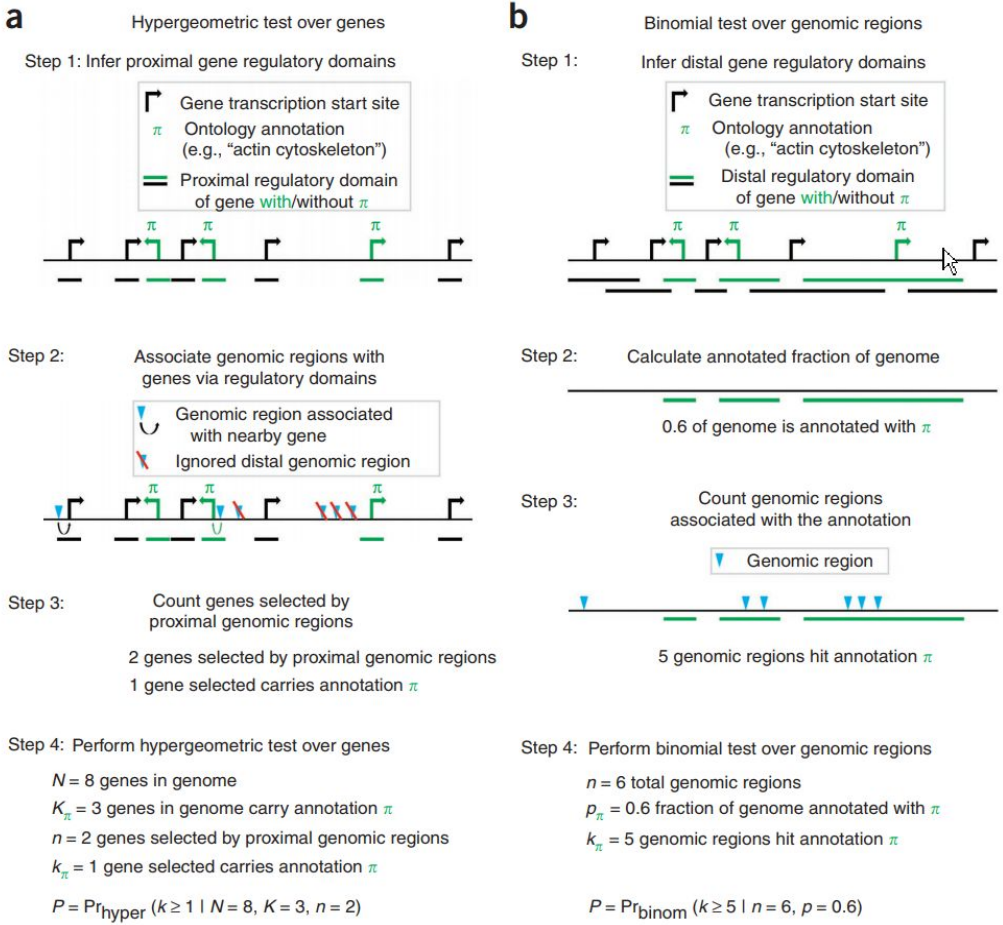
GREAT: Genomic Regions Enrichment of Annotations Tool

User-defined “Regulatory regions”



Uses experimentally-defined “regulatory regions”, when known

Peak Annotation: GREAT



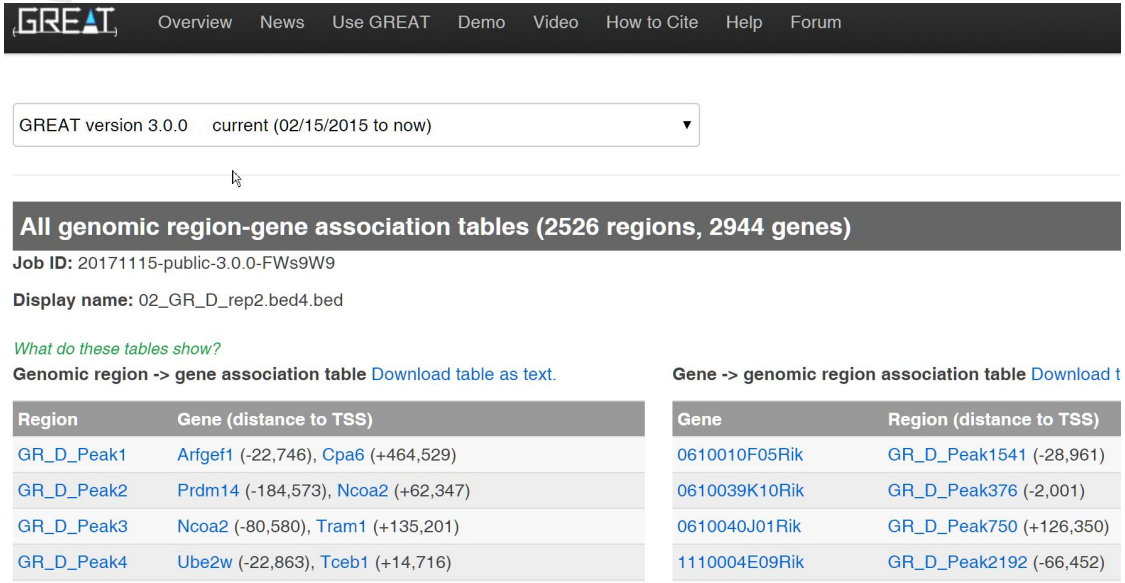
Nature Biotechnology 28, 495–501 (2010)

Ontology annotation of Regulatory regions

Calculate annotated fraction of genome

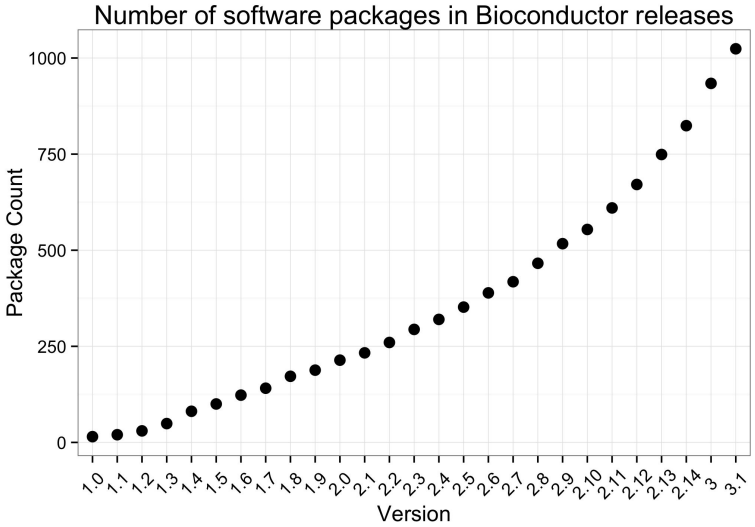
Counts peaks that hit "annotated" portion of genome

Peak Annotation: GREAT



GeneOntology
Mouse Phenotype (MGI)
Human Phenotype (OMIM)
Disease Ontology
MSigDb (Broad Institute)
PANTHER (UCLA)
InterPro (EMBL)
TreeFAM(EMBL)

Bioconductor: a repository of R tools for the analysis of genomic data.



Current version is 3.6 >1400 packages
http://bioconductor.org/packages/release/BiocViews.html#Software

Is updated twice a year

http://bioconductor.org/

R environment

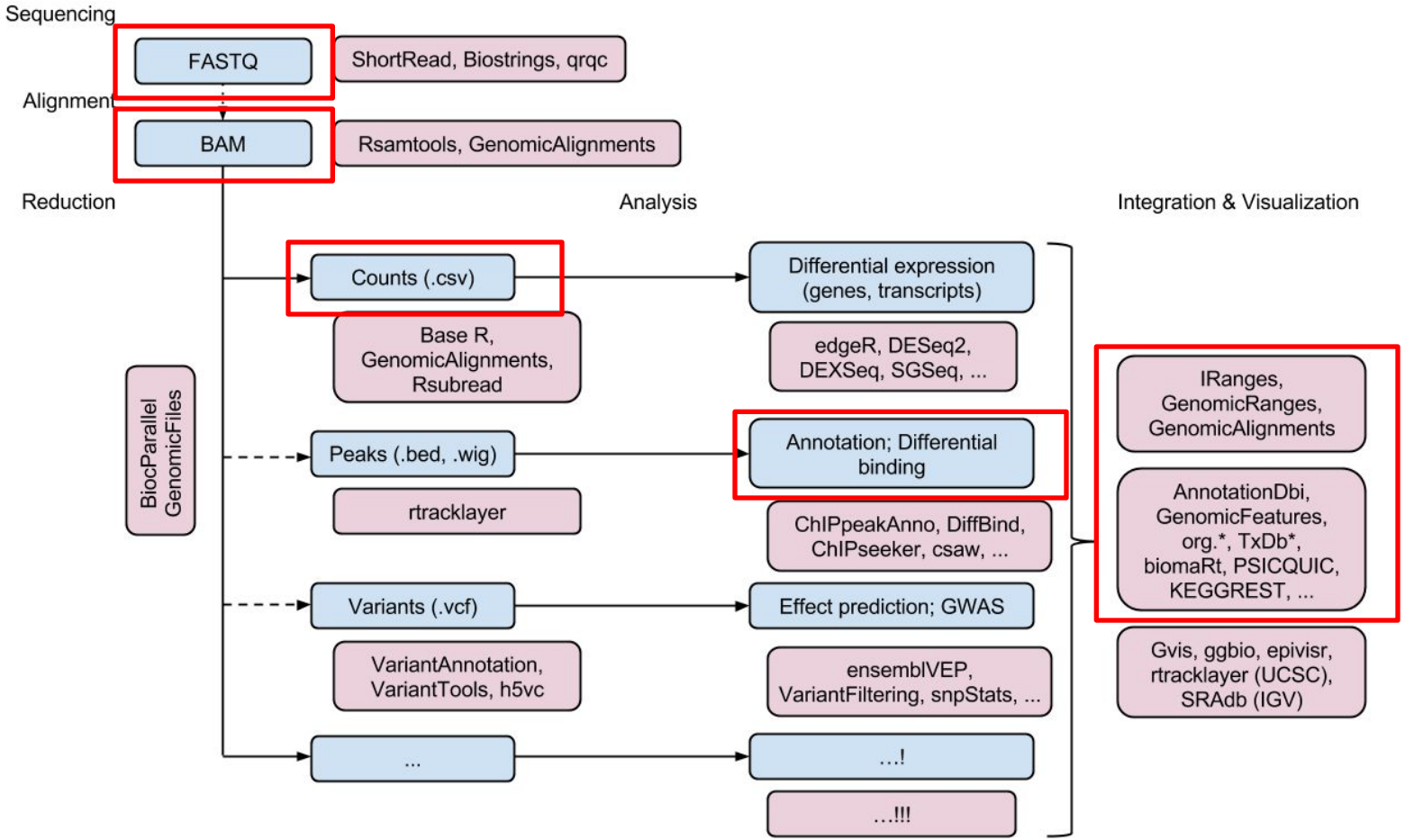
“Compatibility” with ~7000 packages in CRAN

Consistent data structure and access methods

Large user base and effective support

“Good” documentation (but not for all packages)

Bioconductor – (very) general overview



Explore Bioconductor: BiocViews

https://bioconductor.org/packages/release/BiocViews.html#___Software

All Packages

Bioconductor version 3.6 (Release)

Autocomplete biocViews search:

- Software (1476)
  - AssayDomain (573)
  - BiologicalQuestion (559)
  - Infrastructure (322)
  - ResearchField (413)
  - StatisticalMethod (487)
  - Technology (933)
  - WorkflowStep (774)
- AnnotationData (909)
  - ChipManufacturer (388)
  - ChipName (195)
  - CustomArray (2)
  - CustomDBSchema (3)
  - FunctionalAnnotation (14)
  - Organism (595)
  - PackageType (642)
  - SequenceAnnotation (1)

Packages found under Software:

Show All entries

Package	Maintainer
a4	Tobias Verbeke, Willem Ligtenberg
a4Base	Tobias Verbeke, Willem Ligtenberg
a4Classif	Tobias Verbeke, Willem Ligtenberg
a4Core	Tobias Verbeke, Willem Ligtenberg
a4Preproc	Tobias Verbeke, Willem Ligtenberg
a4Reporting	Tobias Verbeke, Willem Ligtenberg
ABAEnrichment	Steffi Grote

Annotation packages in Bioconductor

TxDb family (eg: TxDb.Hsapiens.UCSC.hg19.knownGene**) –**
Transcriptome coordinates for the known genes (introns, exons, UTRs)

BSgenome family (eg: BSgenome.Hsapiens.UCSC.hg19) - complete genome sequence

orgdb family (eg: org.Hs.eg.db) gene based gene ID mapper, plus location information

biomaRt – ENSEMBL-centric annotation package for both model and non-model organisms.

ChIP-Seq/Atac-Seq - related packages

ChIPQC - quality controls for Chip-seq experiments

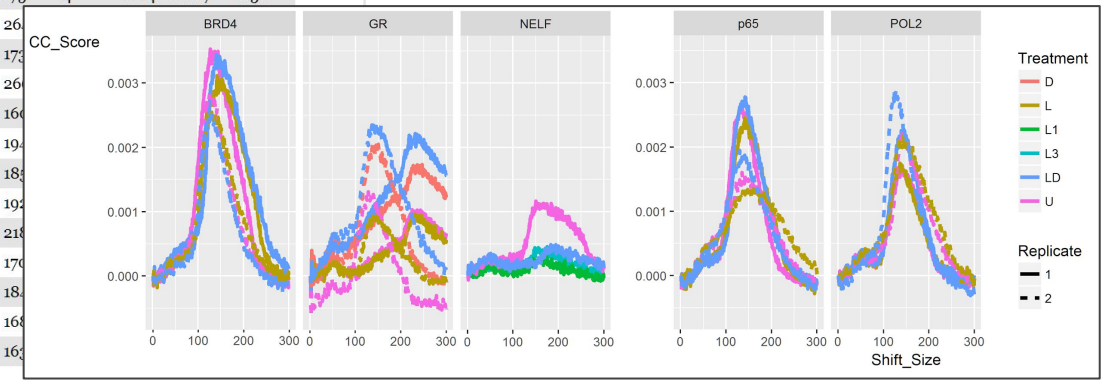
QC Summary

Table 1. Summary of ChIP-seq filtering and quality metrics.

ID	Tissue	Factor	Condition	Replicate	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%	RiBL%
BRD4_L_rep1		BRD4		1	3340300	0	51	175	10	0.99	3.2	0.32
BRD4_L_rep2		BRD4		2	3723935	0	51	160	13	0.93	1.7	0.68
BRD4_LD_rep1		BRD4		1	3290166	0	51	178	8	1	3.9	0.29
BRD4_LD_rep2		BRD4		2	3390004	0	51	142	11	0.83	0.55	0.75
BRD4_U_rep1		BRD4		1	3249236	0	51	163	12	1.1	2.3	0.3
BRD4_U_rep2		BRD4		2	3685080	0	51	142	9.2	0.92	0.65	0.74
GR_D_rep1		GR		1	6219988	0	51	267	3	0.77	1	0.55
GR_D_rep3		GR		2	2814984	0	51	172	5.1	0.82	1.2	0.53
GR_L_rep1		GR		1	5861427	0	51	267	4.7	0.71	0.12	0.53
GR_L_rep3		GR		2	2648221	0	51	175	4	0.64	0.067	0.58
GR_LD_rep1		GR		1	5870854	0	51	26				
GR_LD_rep3		GR		2	3669465	0	51	17				
GR_U_rep1		GR		1	6221929	0	51	26				
GR_U_rep3		GR		2	3065535	0	51	16				
Nelf1_LPS0h		NELF		1	2612013	0	49	19				
Nelf1_LPS1h		NELF		1	2605162	0	49	18				
Nelf1_LPS3h		NELF		1	2876635	0	49	19				
Nelf3_LD_1h		NELF		1	2257041	0	49	21				
p65_L_rep1		p65		1	3546168	0	51	17				
p65_L_rep2		p65		2	4770660	0	51	18				
p65_LD_rep1		p65		1	3546954	0	51	16				
p65_LD_rep2		p65		2	3847697	0	51	16				

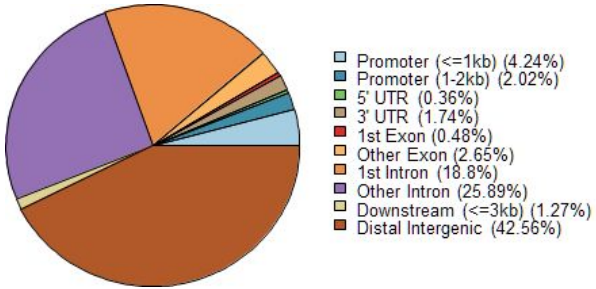
Summary table
Number of Reads
% Duplicates
Read Length
Fragment Length
Relative Cross Coverage
Frip%

Calculates Cross-Correlation but slow
Provides IRD filtering (very conservative)



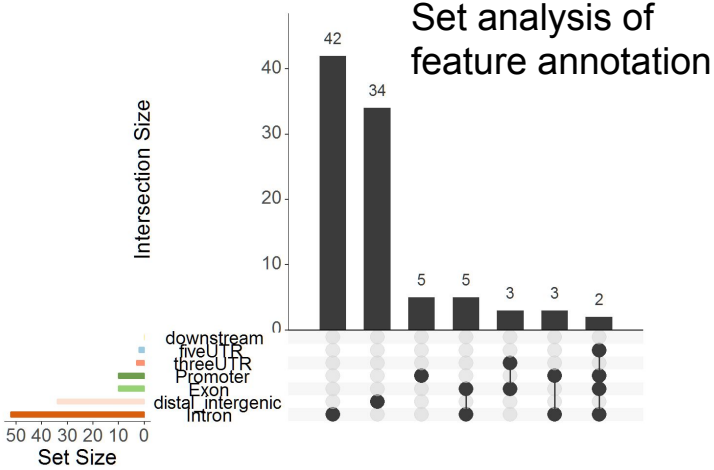
ChIPSeeker , ChIPpeakAnno peak summarization and annotation

Annotation relative to genomic features

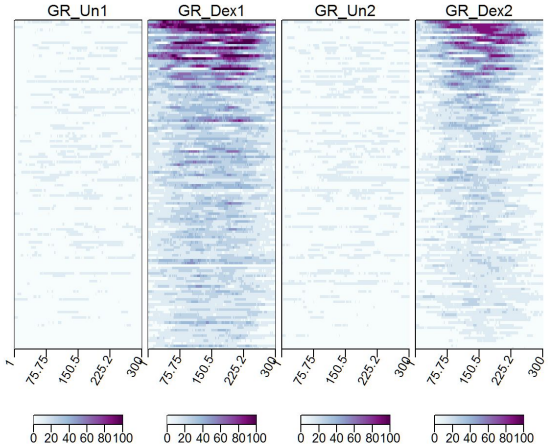


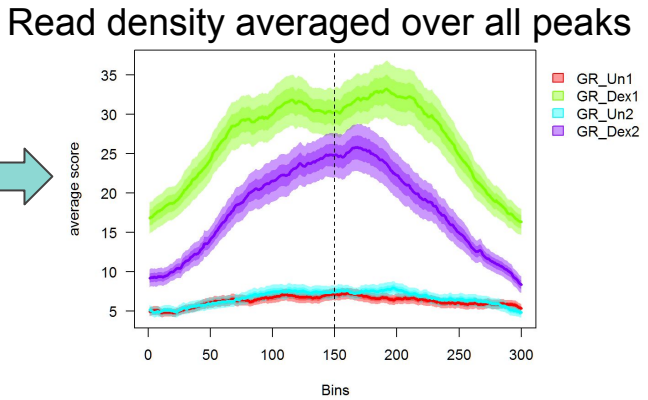
Heatmap of read densities in each peak

Set analysis of feature annotation

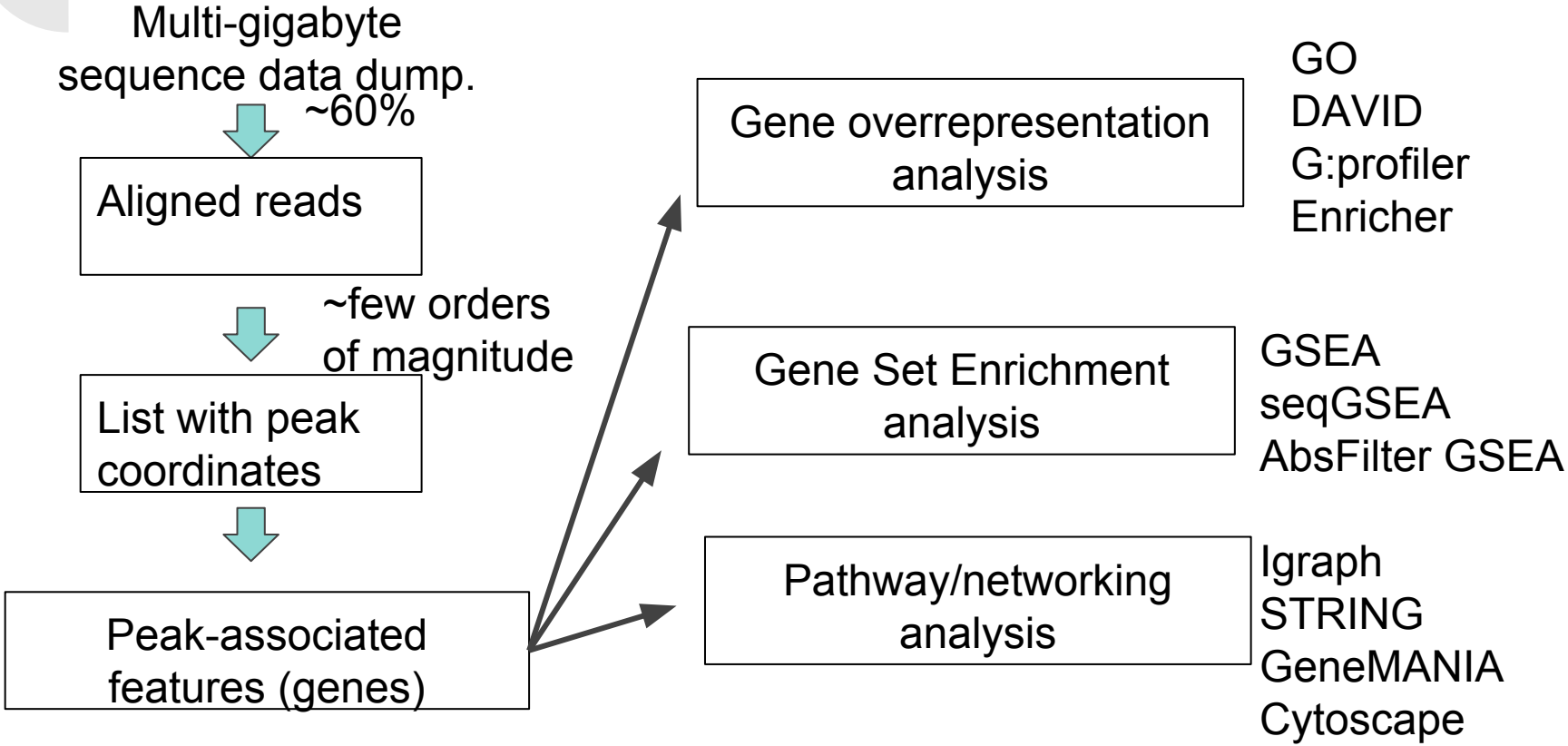


Genomation - peak summarization and annotation





From peak annotation to gene list - data reduction and summarization



Acknowledgements:

Hospital for Special Surgery

Funding:
The Tow Foundation
American Heart Association
NIH