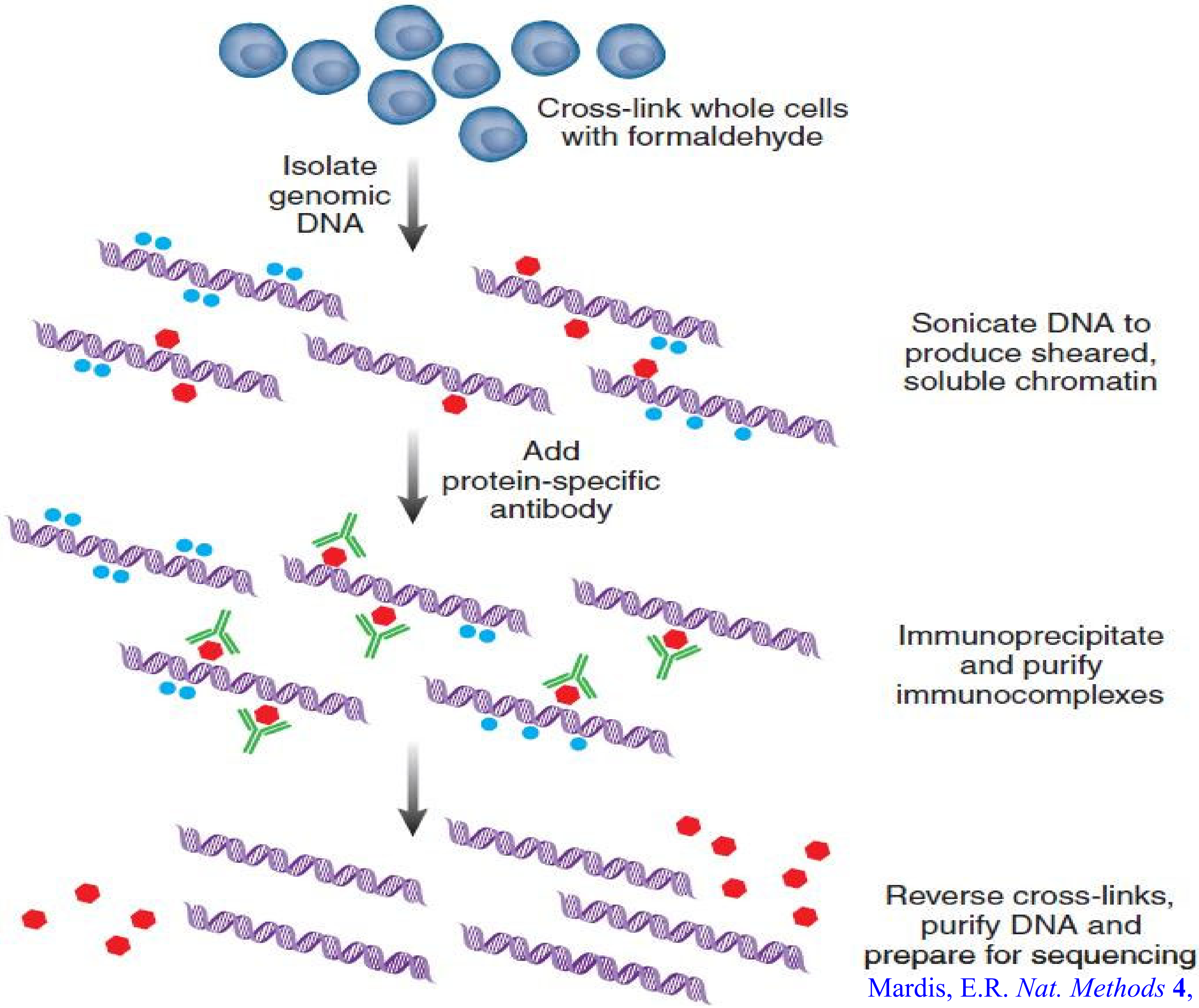
ChIP-seq

What is ChIP-Sequencing?

- ChIP-Sequencing is a new frontier technology to analyze protein interactions with DNA.
- ChIP-Seq
  - Combination of chromatin immunoprecipitation (ChIP) with ultra high-throughput massively parallel sequencing
  - Allow mapping of protein–DNA interactions in-vivo on a genome scale
  - TFs such as NRSF and STAT1
  - Histone modifications such as H3K4me1, H3K27ac

Workflow



Cross-link whole cells with formaldehyde

Isolate genomic DNA

Sonicate DNA to produce sheared, soluble chromatin

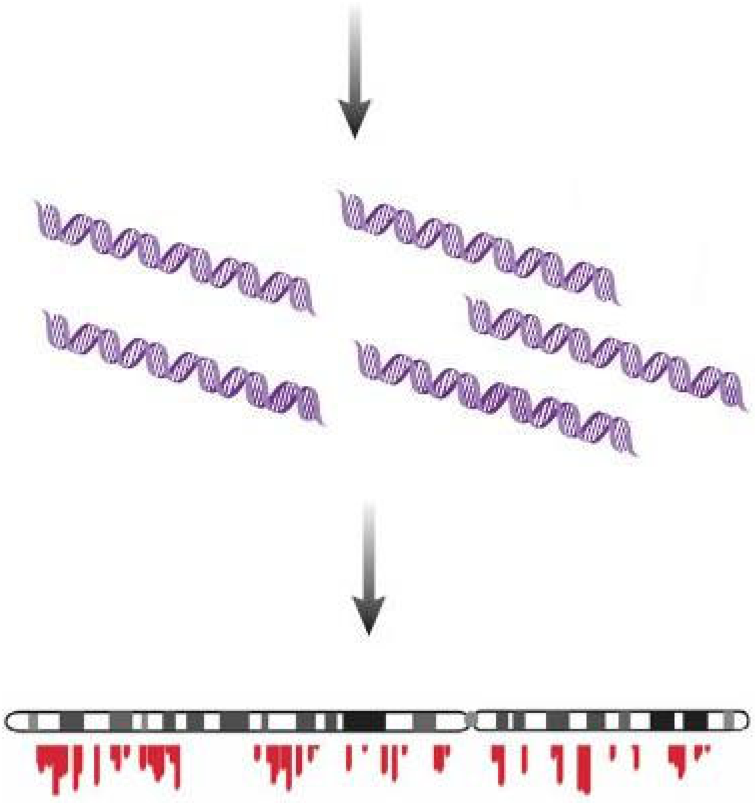
Add protein-specific antibody

Immunoprecipitate and purify immunocomplexes

Reverse cross-links, purify DNA and prepare for sequencing

Mardis, E.R. *Nat. Methods* **4**, 613-614 (2007)

Workflow



Sequence

Map to genome

Why ChIP-Sequencing?

- Previous microarray and ChIP-chip designs require knowing sequence of interest as a promoter, enhancer, or RNA-coding domain.
- ChIP-seq evaluates the entire genome in a single experiment.
- Much lower cost
- Less work
- Higher accuracy

Main advantages of ChIP-seq

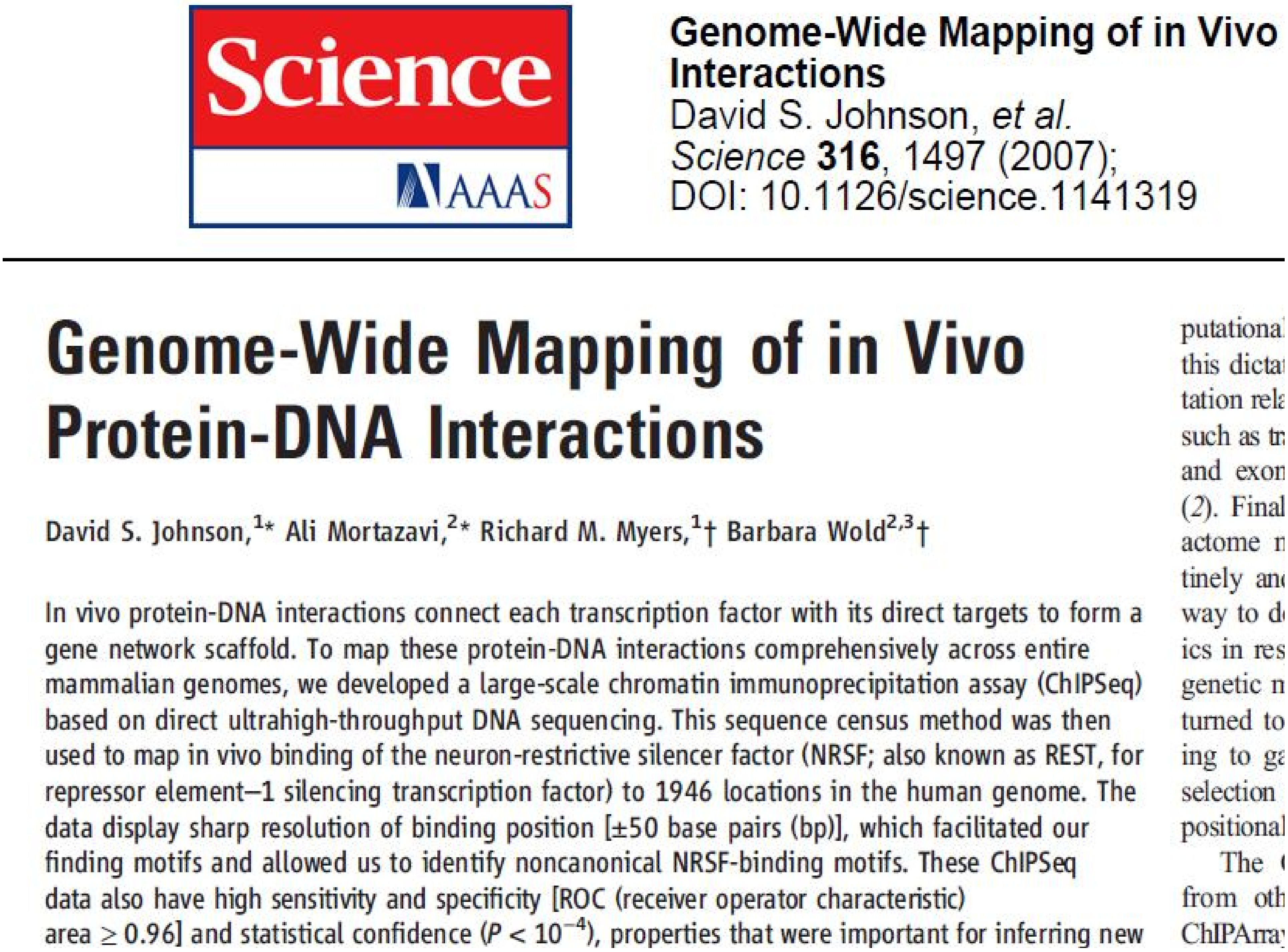
1. . Potential binding regions need not be specified prior to experiment

2. ChIP-seq data are likely to have less noise or artifacts.

Main disadvantages of ChIP-seq

1. The number of tags needed to characterize a protein binding or modifications is highly variable, while the number of sequence tags generated from a single sequencing run is relatively fixed.

2. ChIP-seq does not tell where the factor does not bind, may not comparable for different antibodies



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

putational
this dicta
tation rela
such as tr
and exon
(*2*). Final
actome n
tinely and
way to d
ics in res
genetic n
turned to
ing to ga
selection
positional
The
from oth
ChIPArray

Johnson *et al*, 2007

- ChIP-Seq technology is used to understand in vivo binding of the neuron-restrictive silencer factor (NRSF)
- Results are compared to known binding sites
  - ChIP-Seq signals are strongly agree with the existing knowledge
- Sharp resolution of binding position
- New noncanonical NRSF binding motifs are identified

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]

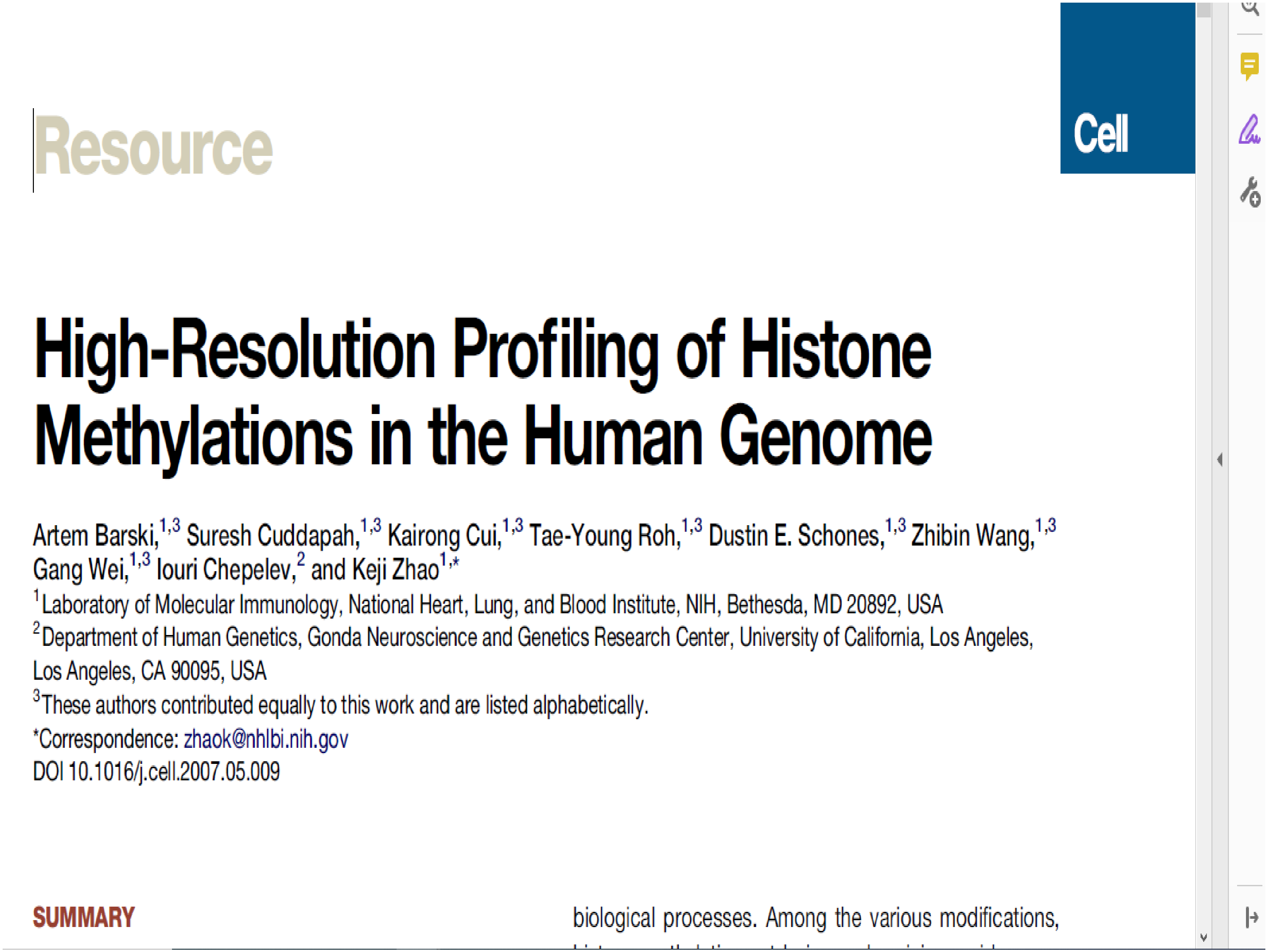
We developed a method, ChIP-sequencing (ChIP-seq), combining chromatin immunoprecipitation (ChIP) and massively parallel sequencing to identify mammalian DNA sequences bound by transcription factors *in vivo*. We used ChIP-seq to map STAT1 targets in interferon-γ (IFN-γ)–stimulated and unstimulated human 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 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 STAT-1 binding regions are identified in stimulated and unstimulated cells respectively.



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

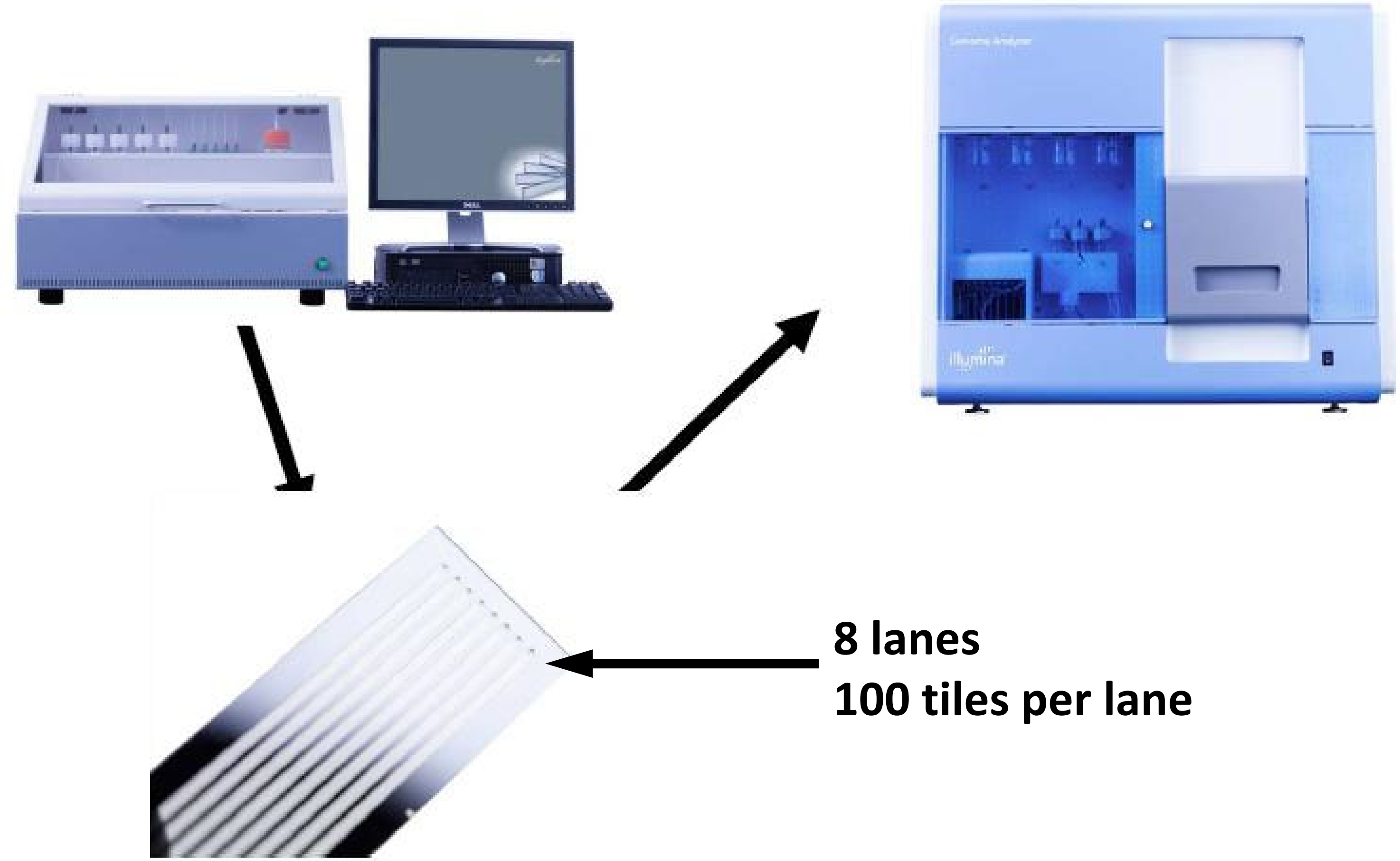
SUMMARY

biological processes. Among the various modifications,

Barski *et al*, 2007

- 21 histone ChIP-Seq together with Pol2 and CTCF ChIP-seq
- Typical patterns of histone methylations exhibited at promoters, insulators, enhancers, and transcribed regions are identified.
- H3K27me1, H3K9me1, H4K20me1, H3K79me1, and H2BK5me1 are all linked to gene activation
- H3K27me3, H3K9me3, and H3K79me3 are linked to repression.
- H2A.Z associates with functional regulatory elements, and CTCF marks boundaries of histone methylation domains.

Illumina Genome Analysis System



8 lanes
100 tiles per lane

Data output and processing

Image data output (tiff files) (around 2008)

100 tiles per lane, 8 lanes per flow cell, 36 cycles.

4 images (A,G,C,T) per tile per cycle = 115,200 images

Each tiff image is ~ 7 MB = 806,400 MB of data

1.6 TB for 70 nt reads,

3.2 TB for 70 nt Paired-end reads

llumina Pipeline:

- Firecrest (image analysis)Locates clusters and calculates intensity and noise
- Bustard (base calling)Deconvolutes signal and corrects for cross-talk, phasing
- GERALD –generation of recursive analyses linked by dependency
- ELAND –Efficient large-scale alignment of nucleotide databases

File formats

A brief note

Sequence formats

- FASTA
- FASTQ

Feature formats

- GFF
- GTF

Alignment formats

- *SAM*
- *BAM*

Formats: **FASTA**

```
>unique_sequence_ID My sequence is pretty cool
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAA
```

- Deceptively simple format (e.g. there is no standard)
- However in general:
    - Header line, starts with ‘>’
    - followed **directly** by an ID
    - … and an optional description (separated by a space)
- Files can be fairly large (whole genomes)
- Any residue type (DNA, RNA, protein), but simple alphabet

Formats: **FASTA**

E.g. a read

```
>unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAA
```

E.g. a chromosome

```
>Group10 gi|323388978|ref|NC_007079.3| Amel_4.5, whole genome shotgun
sequence
TAATTTATATATCTATTTTTTTTATTAAAAAATTTATATTTTTGTTAAAATTTTATTTGATTAGAAATAT
TTTTACTATTGTTCATTAATCGTTAATTAAAGATAGCACAGCACATGTAAGAATTCTAGGTCATGCGAAA
TTAAAAATTAAAAATATTCATATTTCTATAATAATTAAATTATTGTTTTAATTTAAGTAAAAAAATTTCT
AAGAAATCAAAAATTTGTTGTAATATTGAAACAAAATTTTGTTGTCTGCTTTTTATAGTAACTAATAAAT
ATTTAATAAAAAATTACTTTATTTAATATTTTATAATAAATCAAATTGTCCAATTTGAAATTTATTTTAT
CACTAAAAATATCTTTATTATAGTCAATATTTTTTGTTAGGTTTAAATAATTGTTAAAATTAGAAAATGA
TCGATATTTTCAAATAGTACGTTTAACTAATACTTAAGTGAAAGGTAAAGCGGTTATTTAAAATATTGAT
TTATAATATTCGTGACATAATATATTTATAAATAGATTATATATATATATATACATCAAAATATTATACG
AGAACTAGAAAATATTACAGATGCAAAATAAATTAAATTTTGTAAATGTTACAGAATTAAAAATCGAAGT
```

Formats: **FASTQ**

- **FASTQ – FASTA with quality**

```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAA
+
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4
```

- DNA sequence with quality metadata
- The header line, starts with '@',followed directly by an ID and an optional description (separated by a space)
- May be 'raw' data (straight from sequencing) or processed (trimmed)
- Variations: Sanger, Illumina, Solexa (Sanger is most common)
- Can hold 100's of millions of records
- **Files can be very large - 100's of GB apiece**

“Phred” quality (Q) scores

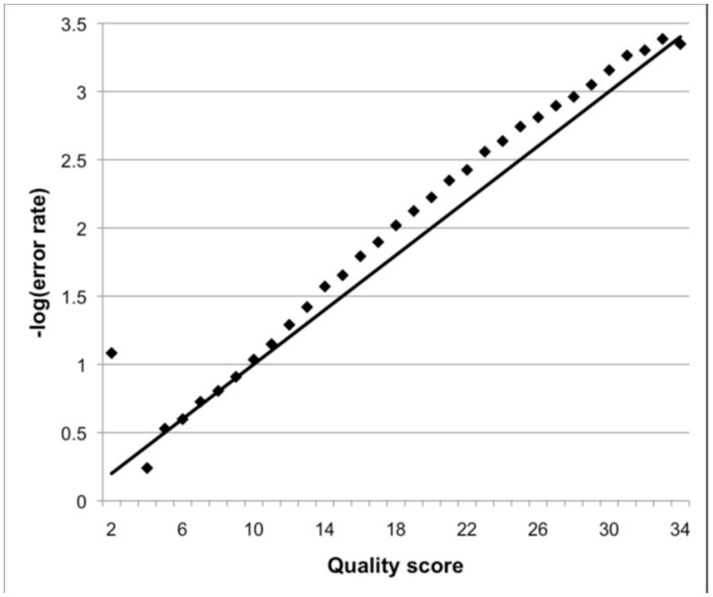
Historically developed for the phred program, an open source base caller for Sanger sequencing

$$Q = -10 * \log 10\ (P)$$

Where P is the probability that a base call is erroneous

Q score	Prob. of wrong call	Accuracy
10	1 in 10 (0.1)	90%
20	1 in 100 (0.01)	99%
30	1 in 1000 (0.001)	99.9%
40	1 in 10000 (0.0001)	99.99%

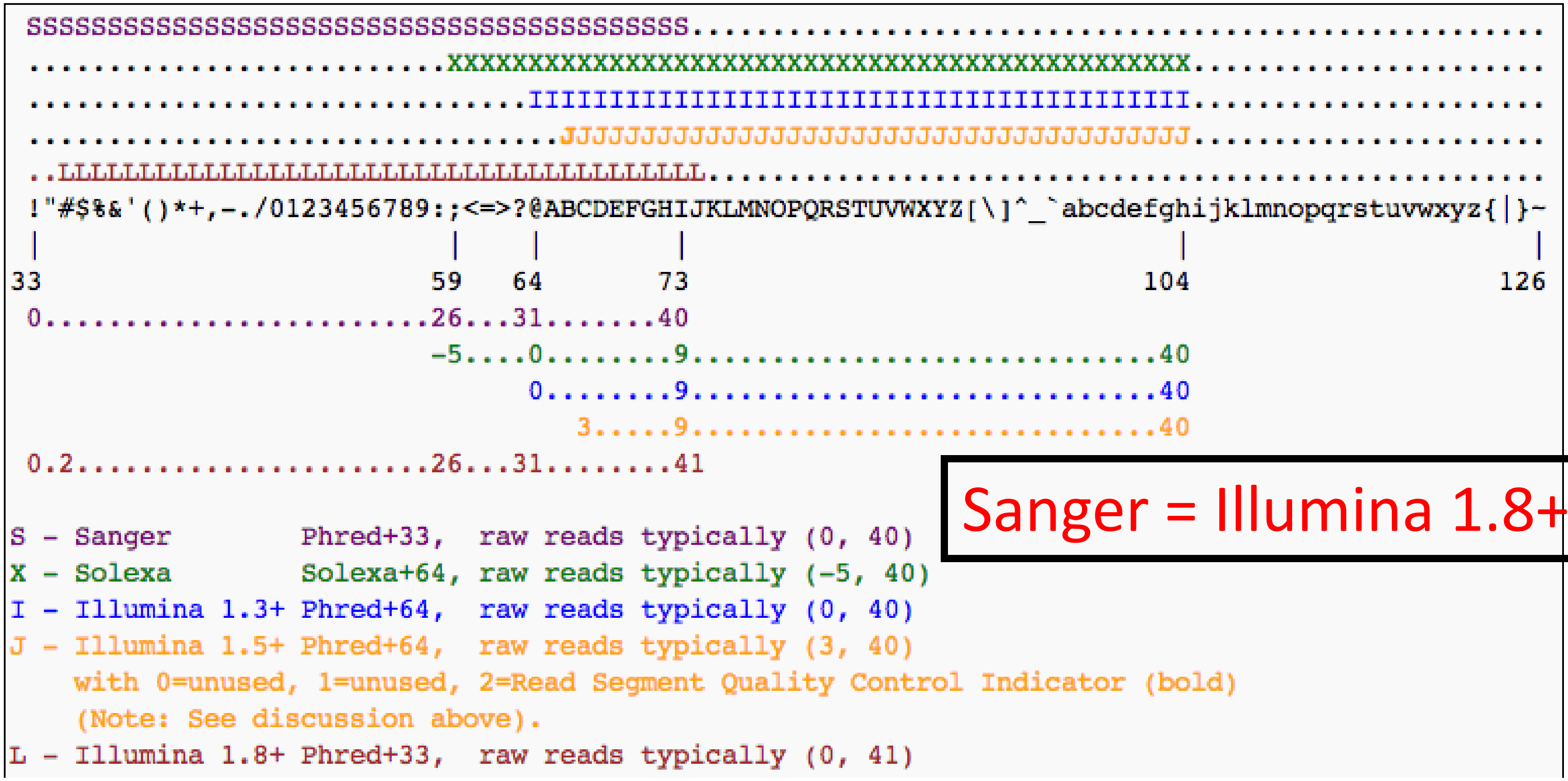
Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing
Akihiro Fujimoto, Hidewaki Nakagawa, Naoya Hosono, Kaoru Nakano, Tetsuo Abe, Keith A Boroevich, Masao Nagasaki, Rui Yamaguchi, Tetsuo Shibuya, Michiaki Kubo, Satoru Miyano, Yusuke Nakamura & Tatsuhiko Tsunoda
Nature Genetics, 2010



Formats: **FASTQ**

- **FASTQ – FASTA with quality**

```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAA
+unique_sequence_ID
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4
```



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
  ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9.............................40
  0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Sanger = Illumina 1.8+

http://en.wikipedia.org/wiki/FASTQ_format

Formats : **SAM**

- **SAM – Sequence Alignment/Map format**
  - SAM file format stores alignment information
- **Plain text**
- **Specification**: http://samtools.sourceforge.net/SAM1.pdf
- Contains quality information, meta data, alignment information, sequence etc.
- **Files can be very large:** Many 100’s of GB or more
- Normally converted into **BAM** to save space (and text format is mostly useless for downstream analyses)

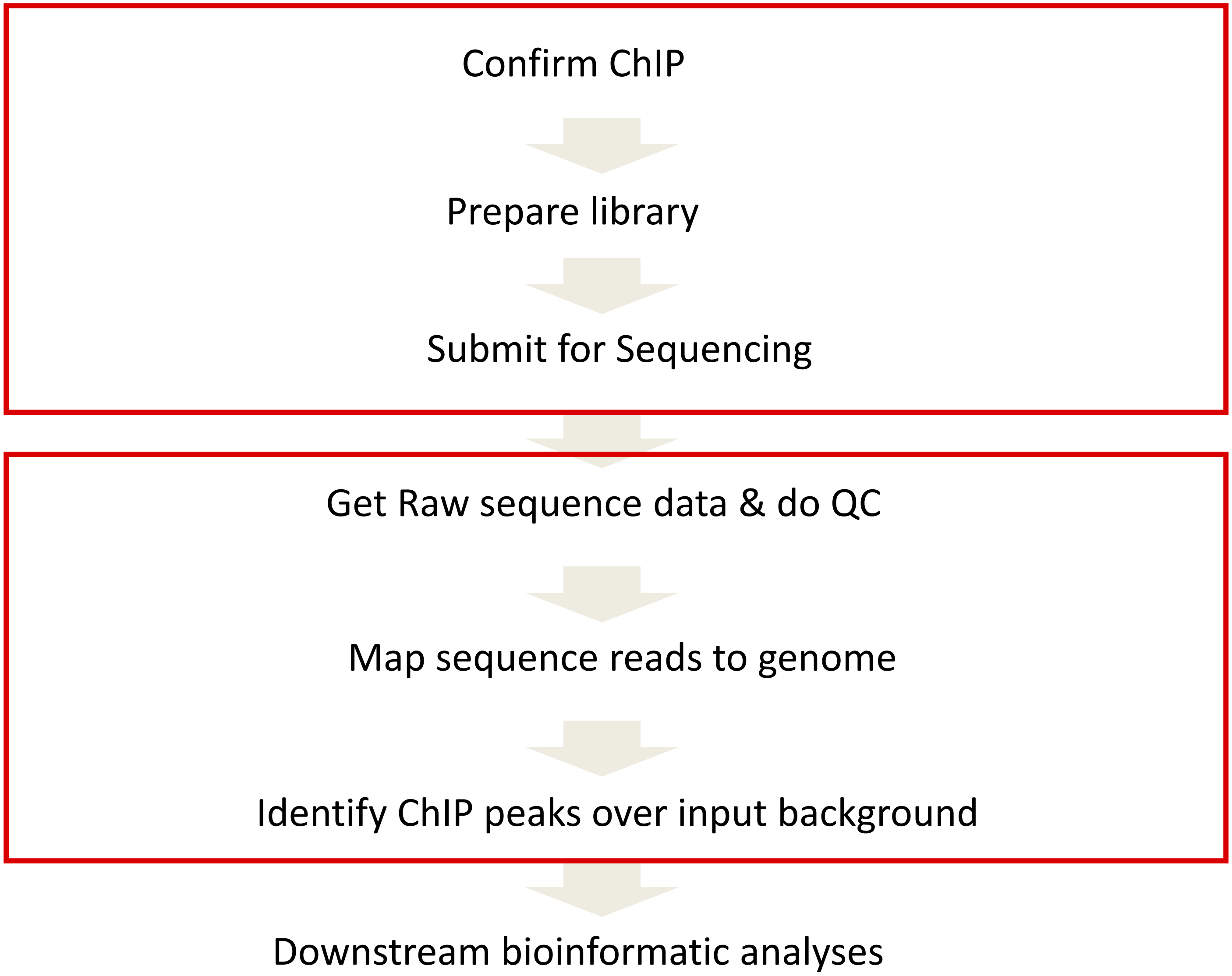
```
@HD       [format version]
@SQ       SN:chr_1 LN:12345678
@PG       [information about program that made this]
HWI-D00758:59:C7U2JANXX:1:1101:1398:2079     0     chr_1     130447256     255     1S9M   *     0
0     NAGCTCTTTA     #/<<BFBBFF NH:i:1  HI:i:1  AS:i:93 nM:i:2
```

Formats : **BAM**

- ✧ **BAM – BGZF compressed SAM format**
  - ✧ Compressed/binary version of SAM and is **not human readable.** Uses a specialized compression algorithm optimized for indexing and record retrieval (bgzip)
  - ✧ Makes the alignment information easily accessible to downstream applications (large genome file not necessary)
  - ✧ Unsorted, sorted by sequence name, **sorted by genome coordinates**
  - ✧ May be accompanied by an index file (.bai) (only if coordinate sorted)

- ✧ **Files are typically very large:** ~ 1/5 of SAM, but still very large

ChIP-seq Workflow



Confirm ChIP

Prepare library

Submit for Sequencing

Get Raw sequence data & do QC

Map sequence reads to genome

Identify ChIP peaks over input background

Downstream bioinformatic analyses

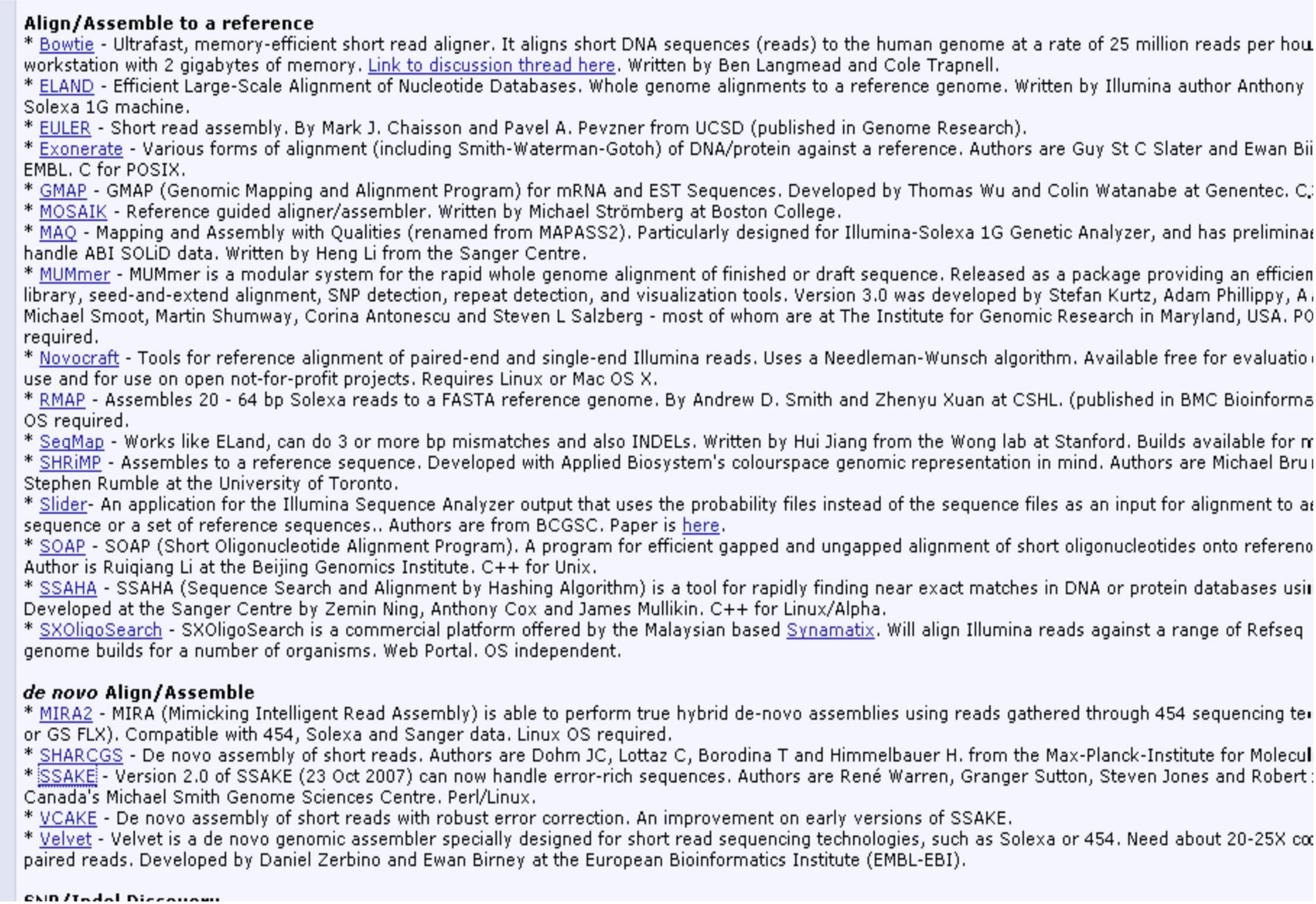
Adaptor sequence removal and read filtration

Illumina FASTQ file generation pipelines include an adapter trimming option for the removal of adapter sequences from the 3’ ends of reads

The adapters contain the sequencing primer binding sites, the index sequences, and the sites that allow library fragments to attach to the flow cell lawn.

Trimmomatic,
https://academic.oup.com/bioinformatics/article/30/15/2114/2390096

Cutadapt
https://cutadapt.readthedocs.io/en/stable/



Align/Assemble to a reference

* Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hou
workstation with 2 gigabytes of memory. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell.
* ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony
Solexa 1G machine.
* EULER - Short read assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research).
* Exonerate - Various forms of alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Bi
EMBL. C for POSIX.
* GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C,
* MOSAIK - Reference guided aligner/assembler. Written by Michael Strömberg at Boston College.
* MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has prelimina
handle ABI SOLiD data. Written by Heng Li from the Sanger Centre.
* MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficien
library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, A
Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. PO
required.
* Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Available free for evaluatio
use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* RMAP - Assembles 20 - 64 bp Solexa reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinforma
OS required.
* SeqMap - Works like ELand, can do 3 or more bp mismatches and also INDELs. Written by Hui Jiang from the Wong lab at Stanford. Builds available for m
* SHRiMP - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Bru
Stephen Rumble at the University of Toronto.
* Slider- An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a
sequence or a set of reference sequences.. Authors are from BCGSC. Paper is here.
* SOAP - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto referenc
Author is Ruiqiang Li at the Beijing Genomics Institute. C++ for Unix.
* SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases usi
Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
* SXOligoSearch - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq
genome builds for a number of organisms. Web Portal. OS independent.

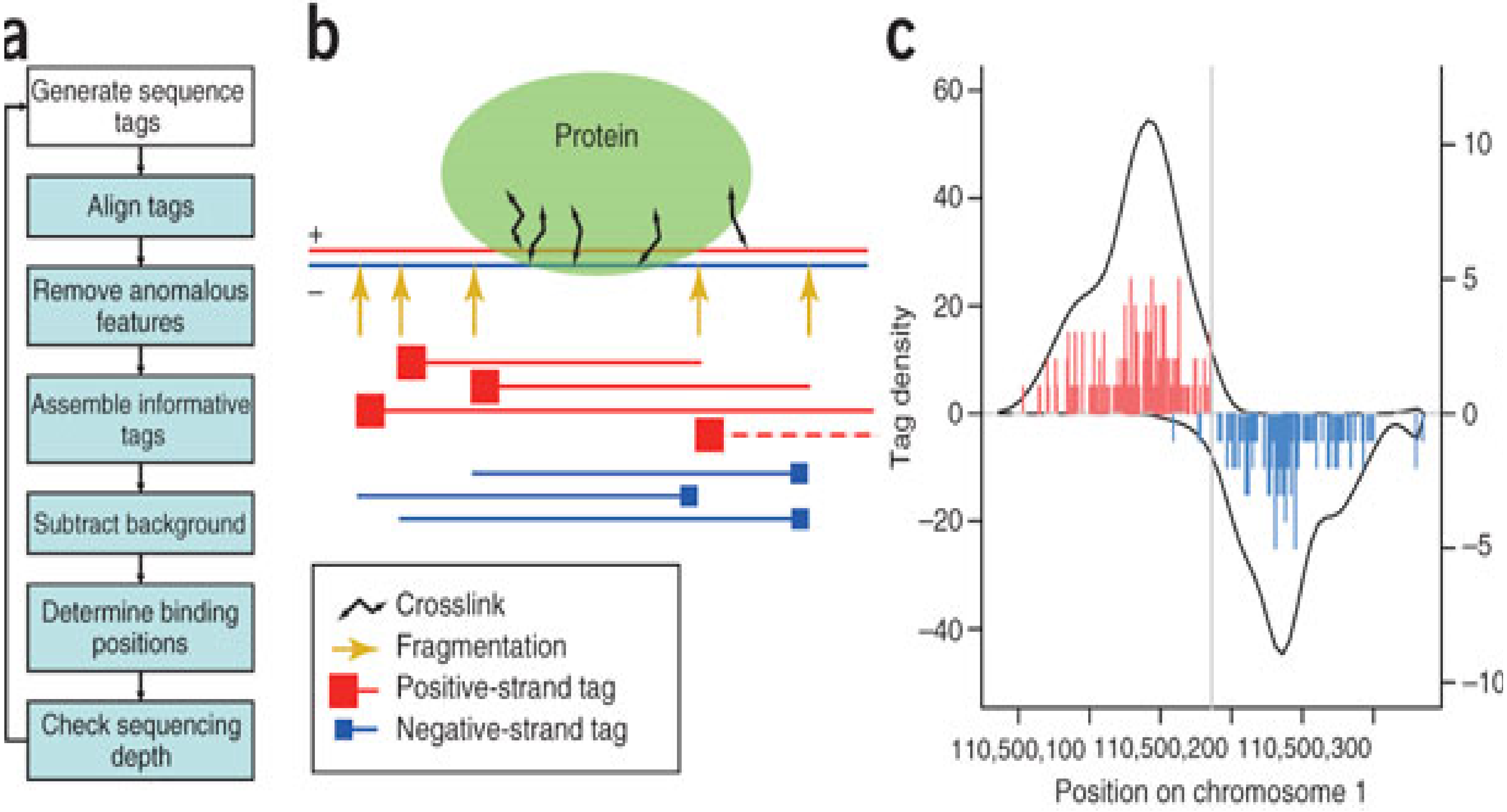
***de novo* Align/Assemble**

* MIRA2 - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing te
or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
* SHARCGS - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecul
* SSAKE - Version 2.0 of SSAKE (23 Oct 2007) can now handle error-rich sequences. Authors are René Warren, Granger Sutton, Steven Jones and Robert
Canada's Michael Smith Genome Sciences Centre. Perl/Linux.
* VCAKE - De novo assembly of short reads with robust error correction. An improvement on early versions of SSAKE.
* Velvet - Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X co
paired reads. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI).

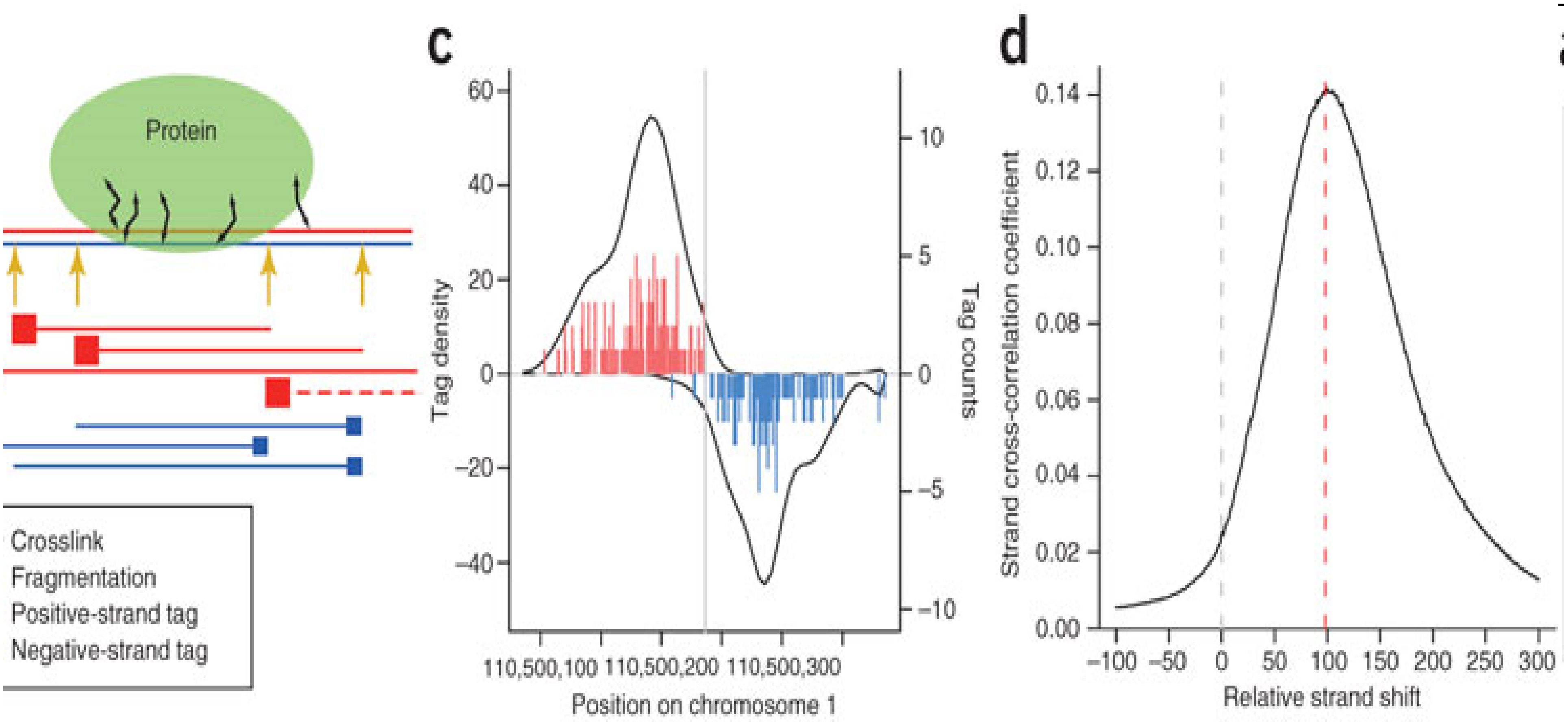
SNP/Indel Discovery

Read distribution around protein binding positions

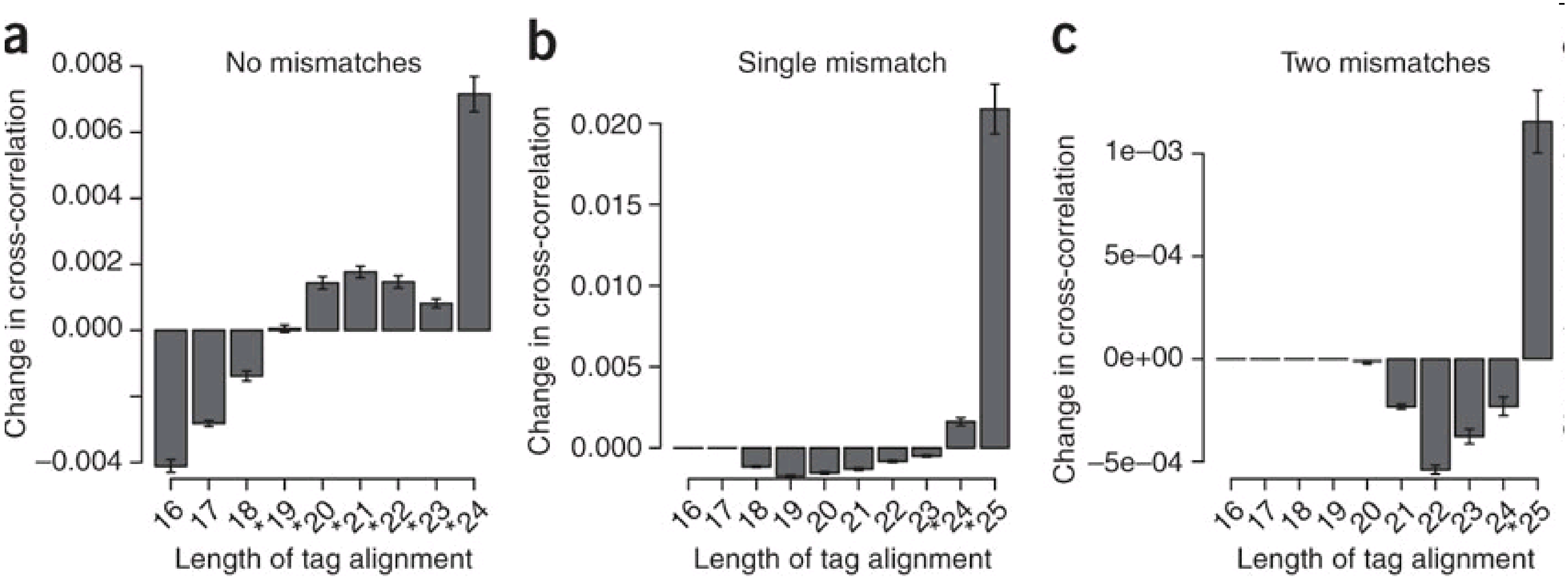
Nat Biotechnol. 2008; 26(12): 1351–1359



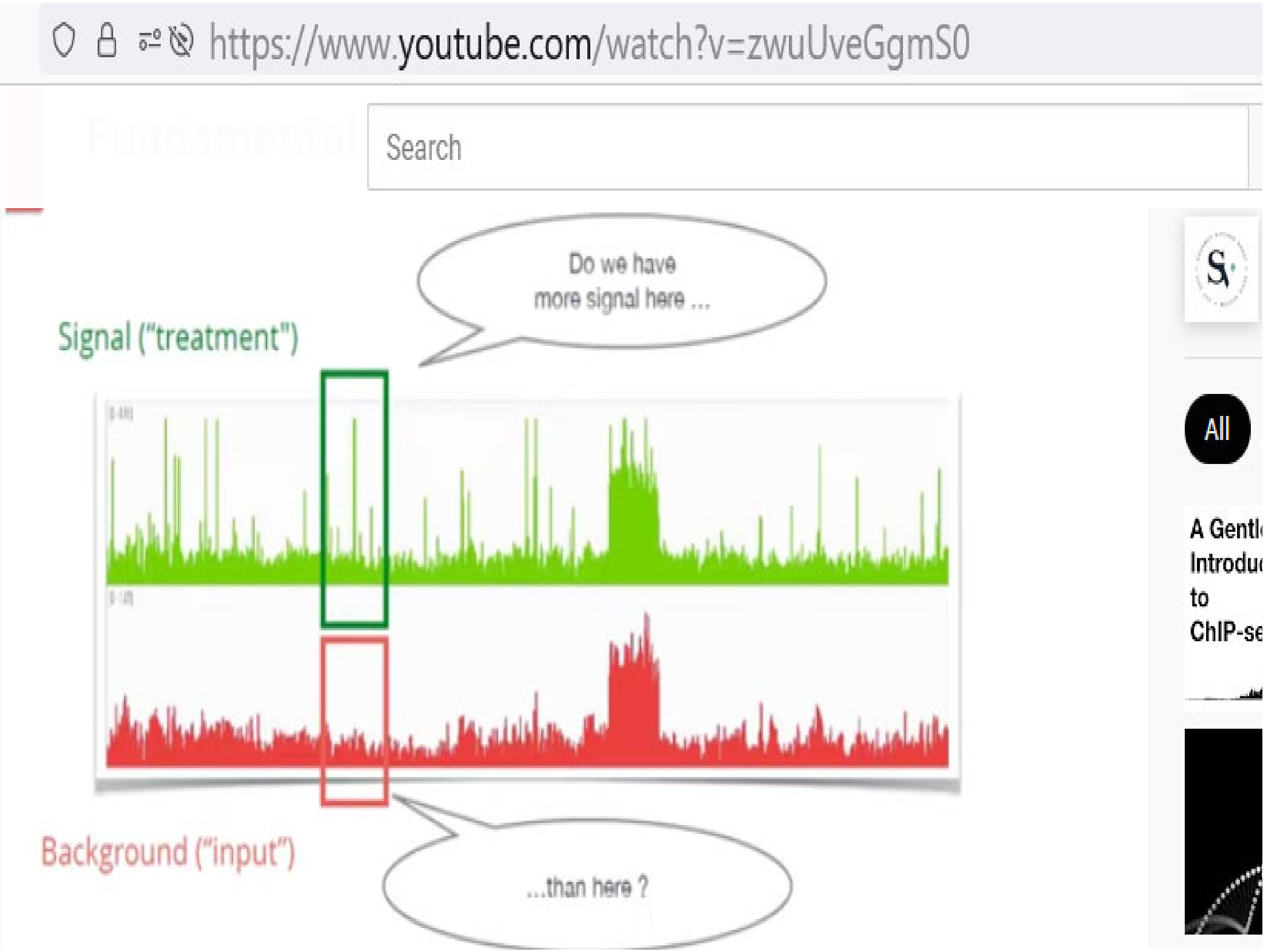
Read extension or shifting



Which reads to use?

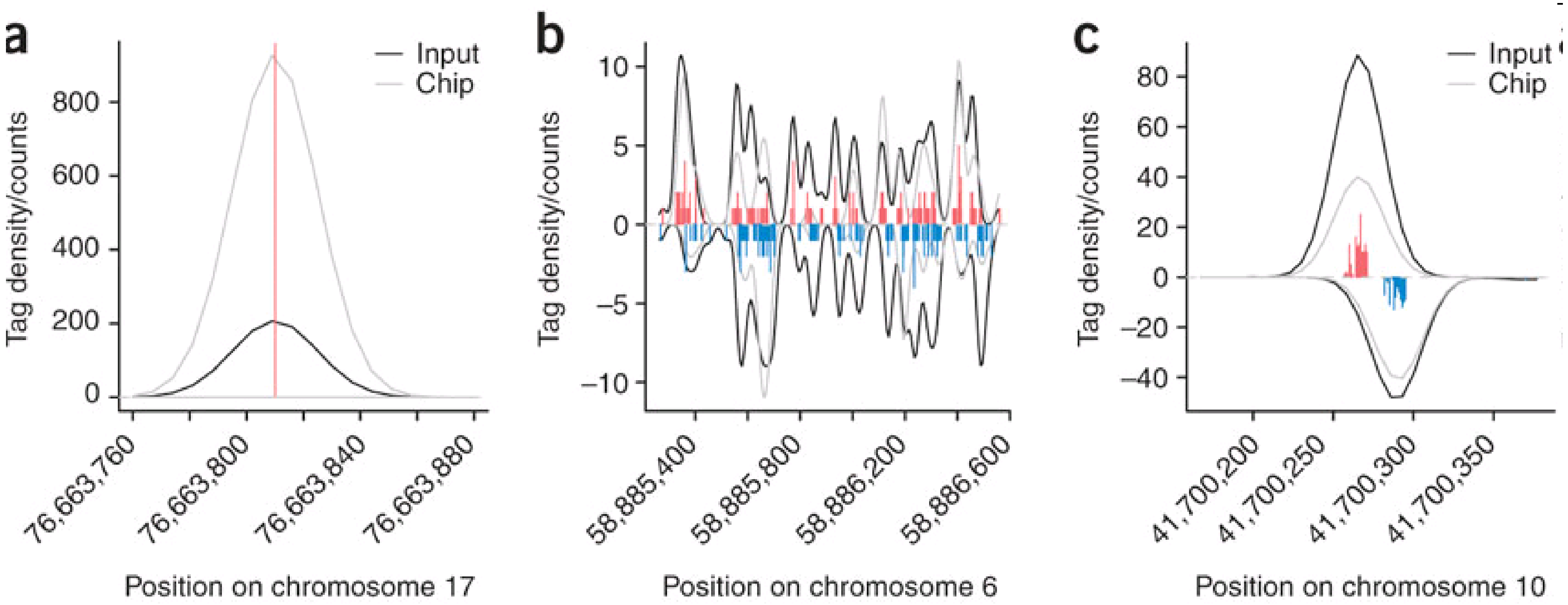


The plots show the change in strand mean cross-correlation profile when this class of tags is considered together with the base class of perfectly aligned tags (25 bp, no mismatches).



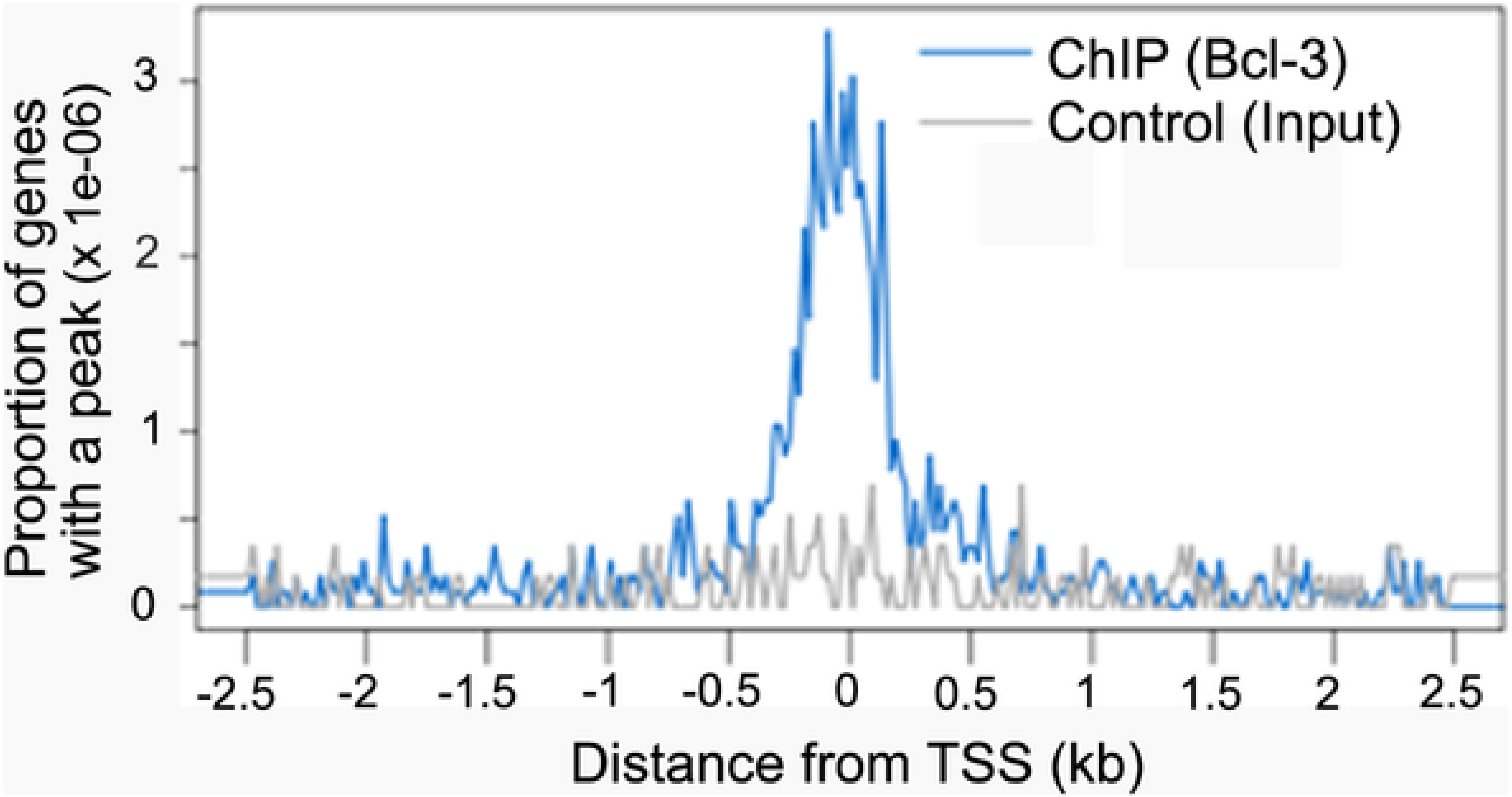
https://www.youtube.com/watch?v=zwuUveGgmS0

Three major types of background anomalies

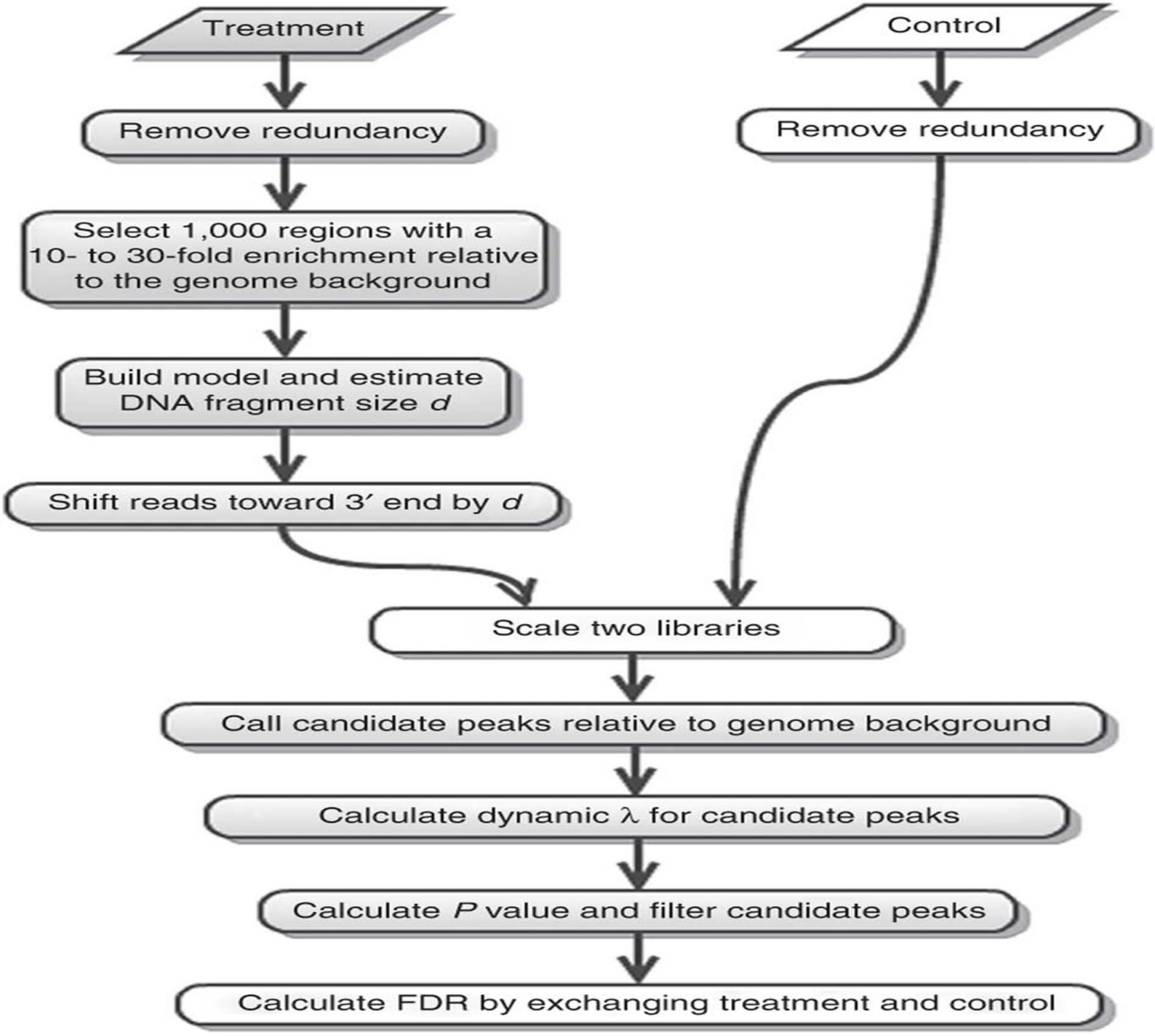


(**a**) Singular positions with extremely high tag count. (**b**) Larger, nonuniform regions of increased background tag density. (**c**) Background tag density patterns resembling true protein-binding positions. Each plot shows density of tags from ChIP and input samples. The tag histograms give combined tag counts.

Scale input and ChIP samples



MACS and MACS2



Treatment

Control

Remove redundancy

Remove redundancy

Select 1,000 regions with a 10- to 30-fold enrichment relative to the genome background

Build model and estimate DNA fragment size *d*

Shift reads toward 3′ end by *d*

Scale two libraries

Call candidate peaks relative to genome background

Calculate dynamic λ for candidate peaks

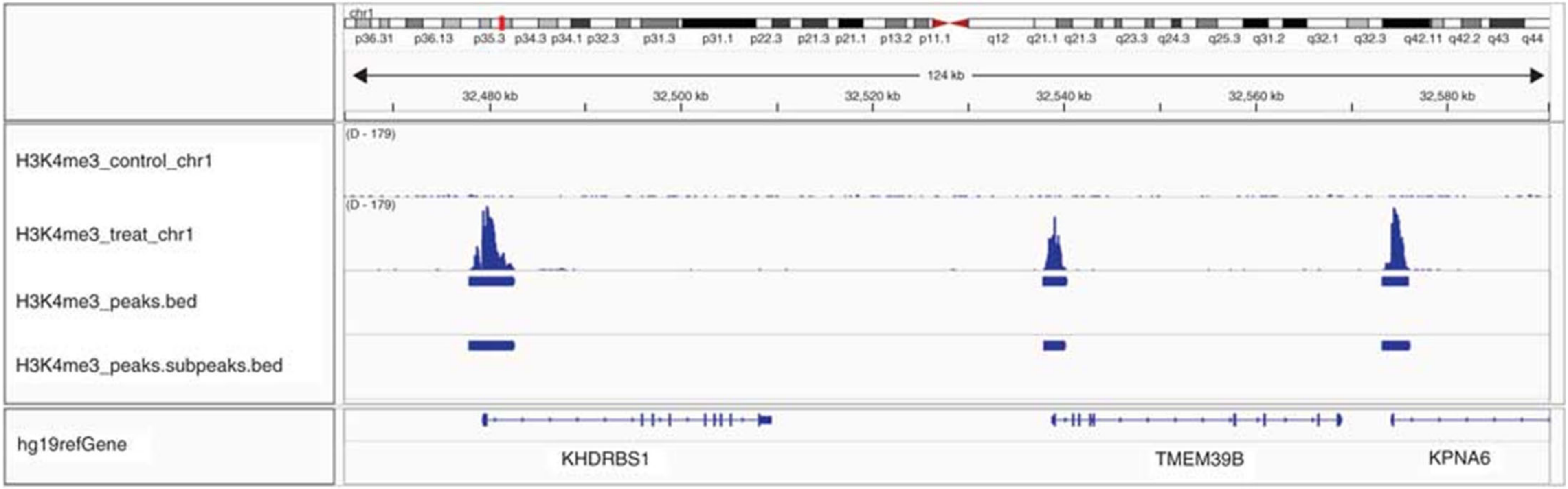
Calculate *P* value and filter candidate peaks

Calculate FDR by exchanging treatment and control

Integrative Genomics Viewer (IGV)

The **Integrative Genomics Viewer (IGV)** is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data. It supports flexible integration of all the common types of genomic data and metadata, investigator-generated or publicly available, loaded from local or cloud sources

IGV visualization of MACS results using the University of Washington H3K4me3 data set.


chr1
p36.31 p36.13 p35.3 p34.3 p34.1 p32.3 p31.3 p31.1 p22.3 p21.3 p21.1 p13.2 p11.1 q12 q21.1 q21.3 q23.3 q24.3 q25.3 q31.2 q32.1 q32.3 q42.11 q42.2 q43 q44
124 kb
32,480 kb
32,500 kb
32,520 kb
32,540 kb
32,560 kb
32,580 kb
H3K4me3_control_chr1
(0 - 179)
H3K4me3_treat_chr1
(0 - 179)
H3K4me3_peaks.bed
H3K4me3_peaks.subpeaks.bed
hg19refGene
KHDRBS1
TMEM39B
KPNA6

Summary Steps

- Confirm ChIP worked by qPCR before sending the ChIP samples to the sequencing center.
- Have a control sample
- Clean reads and mappability
- Non-redundant fraction of reads
- Read extension/shifting
- Scale the input and the ChIP sample
- MACS2

Summary Tools

- Trimmomatic or Cutadapt to remove adaptor sequences and filter low quality reads
- Bowtie 2 to map reads to references
- MACS2 to define peaks

references

- https://www.youtube.com/watch?v=zwuUveGgmS0
- https://galaxyproject.org/learn/
- https://rockefelleruniversity.github.io/RU_ChIPseq/

1^{st} In-class question

- What will you do if you can sequence a human genome with $10?

submit your answer at webcourse no later than 6pm today.