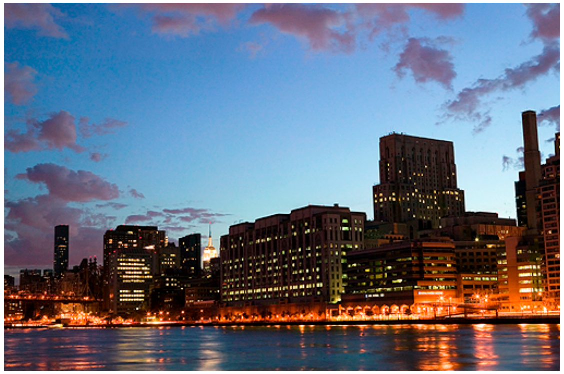
Weill Cornell Medical College

Institute for Computational Biomedicine



ChIP-seq and Hi-C

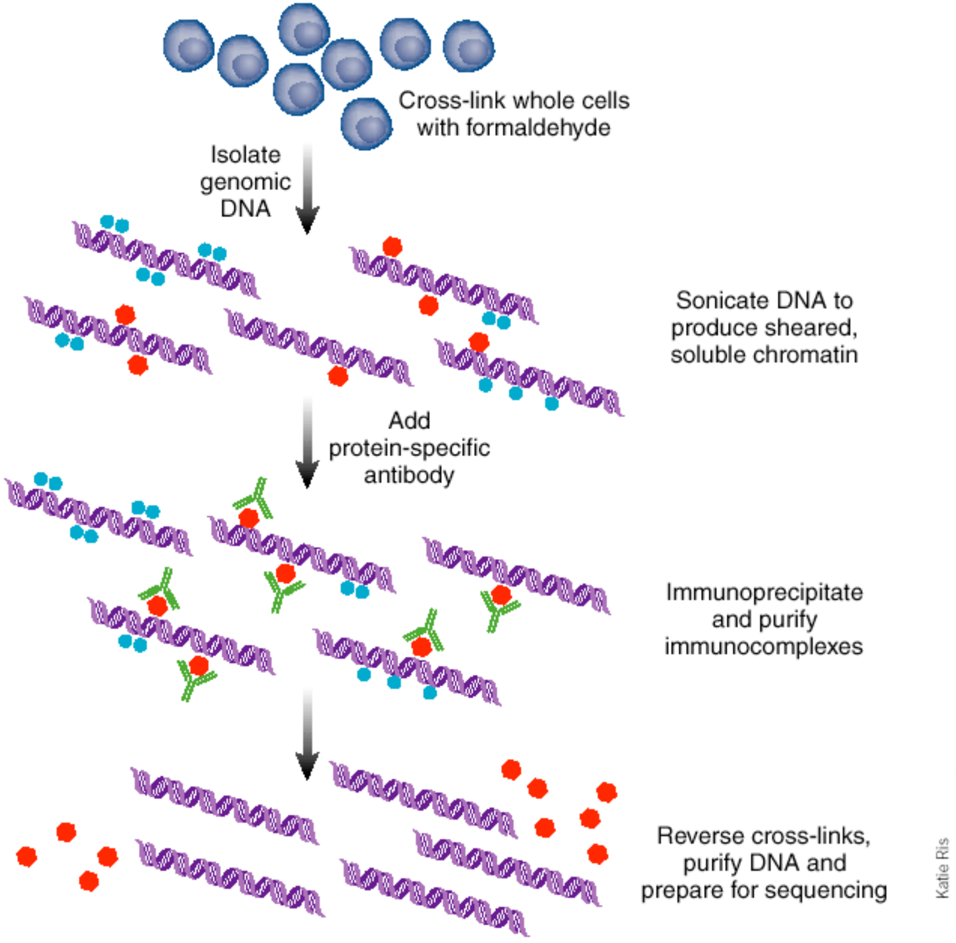
Olivier Elemento, PhD

Laboratory of Cancer Systems Biology

Plan

1. ChIP-seq
2. A few interesting ChIP-seq papers
3. Quality Control of ChIP-seq data
4. ChIP-seq Peak detection
5. Peak Analysis and Interpretation
6. Mapping chromatin interactions using Hi-C

ChIP-seq



Cross-link whole cells with formaldehyde

Isolate genomic DNA

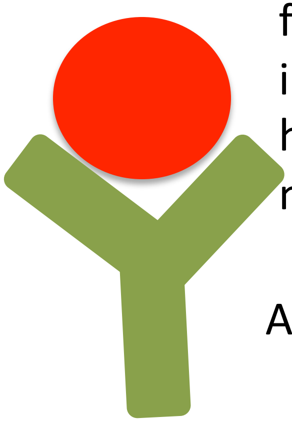
Sonicate DNA to produce sheared, soluble chromatin

Add protein-specific antibody

Immunoprecipitate and purify immunocomplexes

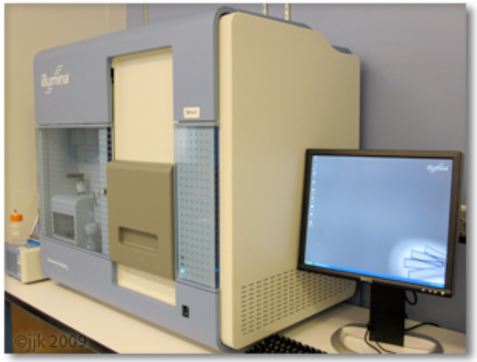
Reverse cross-links, purify DNA and prepare for sequencing

Katie Ris



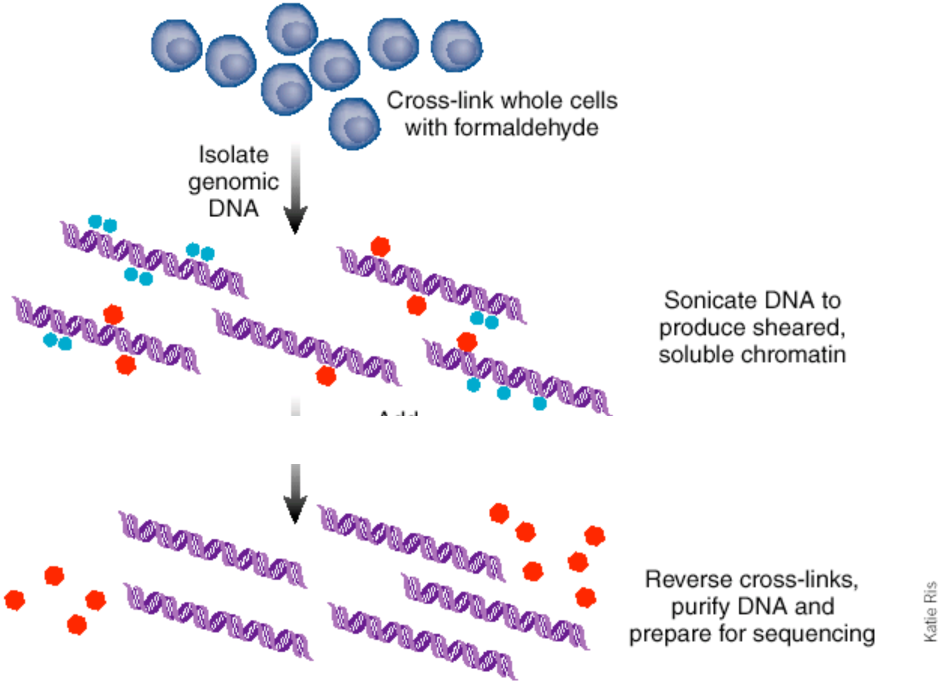
Transcription factor of interest (or histone modification)

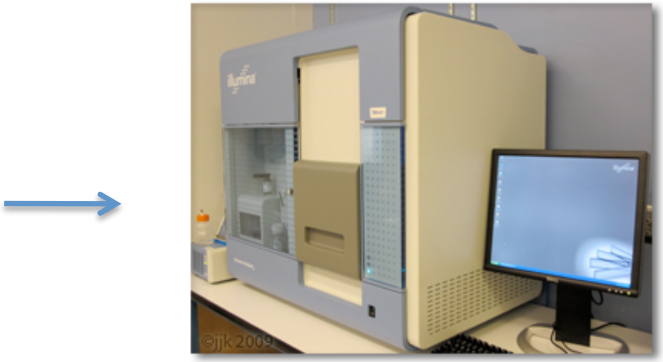
Antibody



Illumina

Control: input DNA


Cross-link whole cells with formaldehyde
Isolate genomic DNA
Sonicate DNA to produce sheared, soluble chromatin
Reverse cross-links, purify DNA and prepare for sequencing
Katie Ris



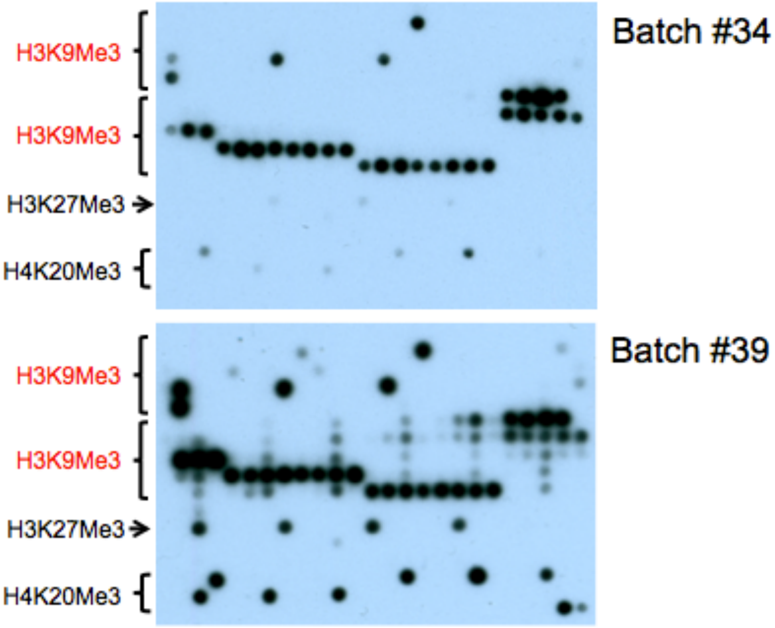
Illumina

Can use IgG as additional control

ChIP-seq methodology

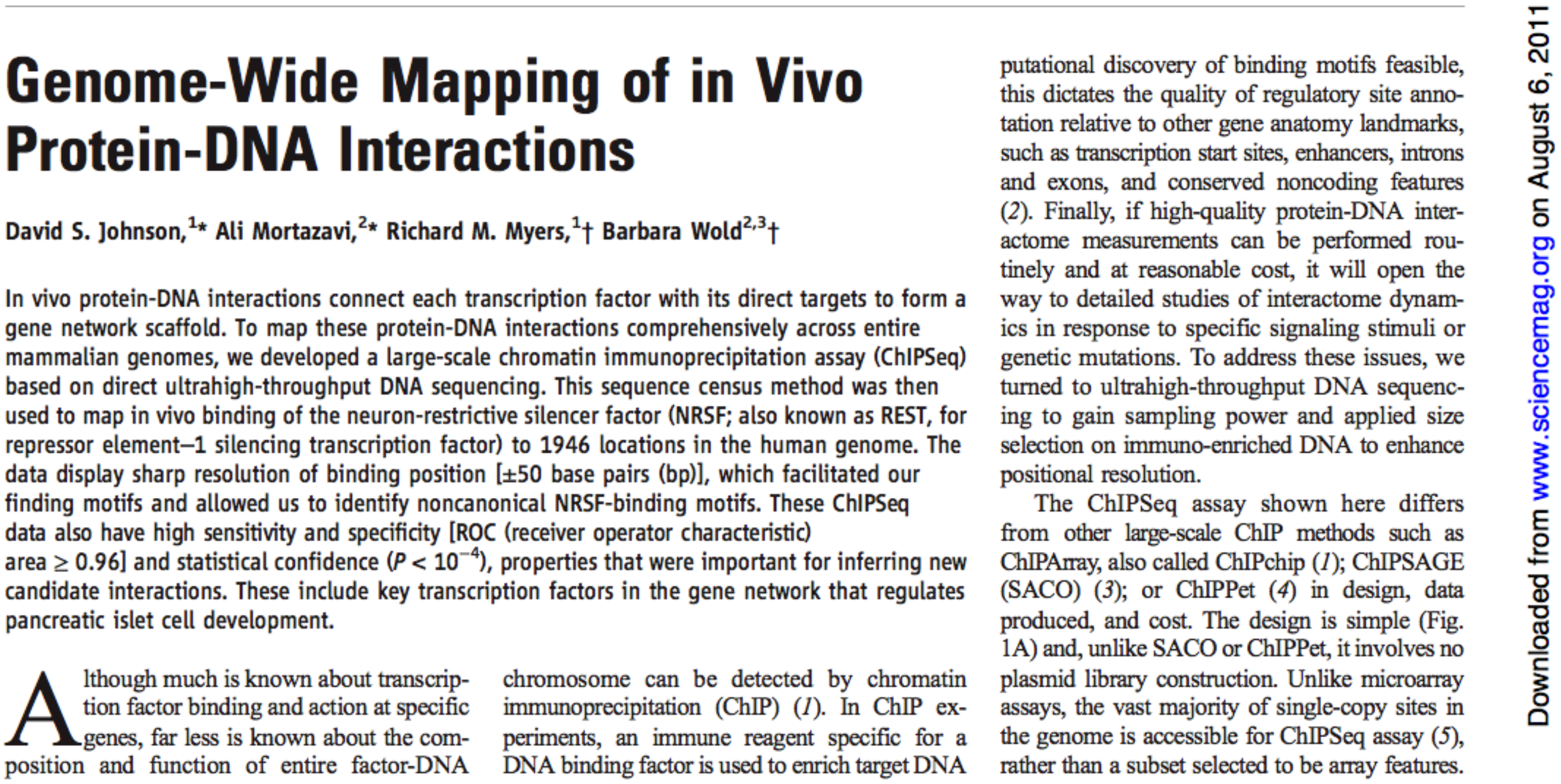
- Identify ChIP-grade antibody, determine specificity (Western, histone peptide array)
- Optimize conditions using single-locus ChIP-PCR (positive and negative controls)
- Sequence ChIP product using 1 Illumina lane per sample (no TruSeq ChIP-seq), single end
- Sequence input/IgG as control

Abcam H3K9Me3 rabbit polyclonal (ab8898)



Assessing the specificity of a commercial H3K9m3 antibody using histone peptide arrays, K. Bunting & B. Swed, WCMC

First ChIP-seq paper



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.

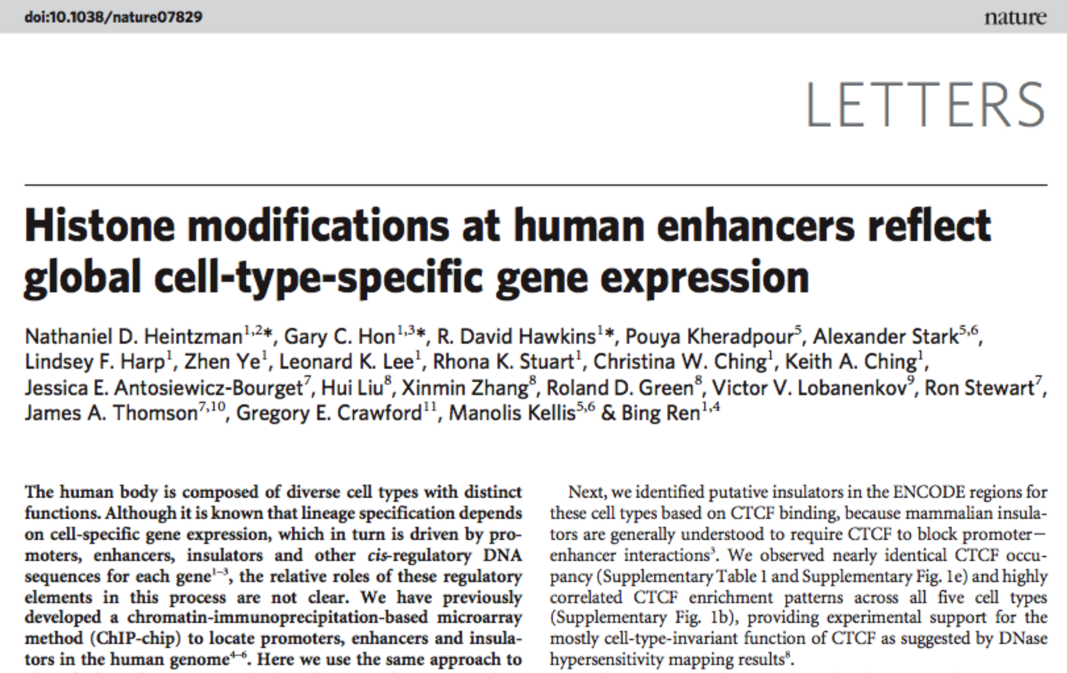
Although much is known about transcription factor binding and action at specific genes, far less is known about the composition and function of entire factor-DNA chromosome can be detected by chromatin immunoprecipitation (ChIP) (*1*). In ChIP experiments, an immune reagent specific for a DNA binding factor is used to enrich target DNA

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. 1A) and, unlike SACO or ChIPPet, it involves no plasmid library construction. Unlike microarray assays, the vast majority of single-copy sites in the genome is accessible for ChIPSeq assay (*5*), rather than a subset selected to be array features.

2007

Epigenetic modifications at enhancer regions



doi:10.1038/nature07829 nature

LETTERS

Histone modifications at human enhancers reflect global cell-type-specific gene expression

Nathaniel D. Heintzman[1,2]*, Gary C. Hon[1,3]*, R. David Hawkins[1]*, Pouya Kheradpour[5], Alexander Stark[5,6], Lindsey F. Harp[1], Zhen Ye[1], Leonard K. Lee[1], Rhona K. Stuart[1], Christina W. Ching[1], Keith A. Ching[1], Jessica E. Antosiewicz-Bourget[7], Hui Liu[8], Xinmin Zhang[8], Roland D. Green[8], Victor V. Lobanenkov[9], Ron Stewart[7], James A. Thomson[7,10], Gregory E. Crawford[11], Manolis Kellis[5,6] & Bing Ren[1,4]

The human body is composed of diverse cell types with distinct functions. Although it is known that lineage specification depends on cell-specific gene expression, which in turn is driven by promoters, enhancers, insulators and other *cis*-regulatory DNA sequences for each gene[1–3], the relative roles of these regulatory elements in this process are not clear. We have previously developed a chromatin-immunoprecipitation-based microarray method (ChIP-chip) to locate promoters, enhancers and insulators in the human genome[4–6]. Here we use the same approach to

Next, we identified putative insulators in the ENCODE regions for these cell types based on CTCF binding, because mammalian insulators are generally understood to require CTCF to block promoter–enhancer interactions[3]. We observed nearly identical CTCF occupancy (Supplementary Table 1 and Supplementary Fig. 1e) and highly correlated CTCF enrichment patterns across all five cell types (Supplementary Fig. 1b), providing experimental support for the mostly cell-type-invariant function of CTCF as suggested by DNase hypersensitivity mapping results[8].

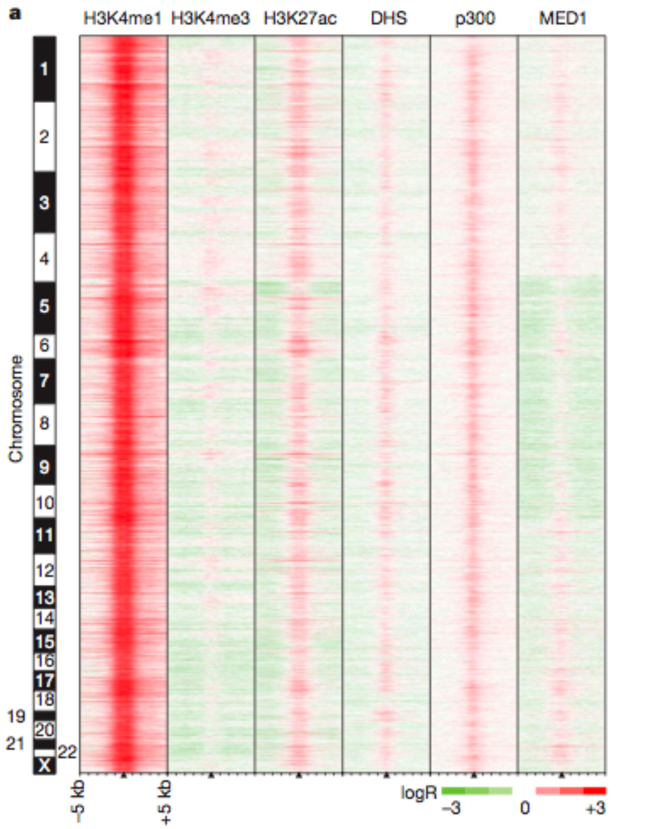
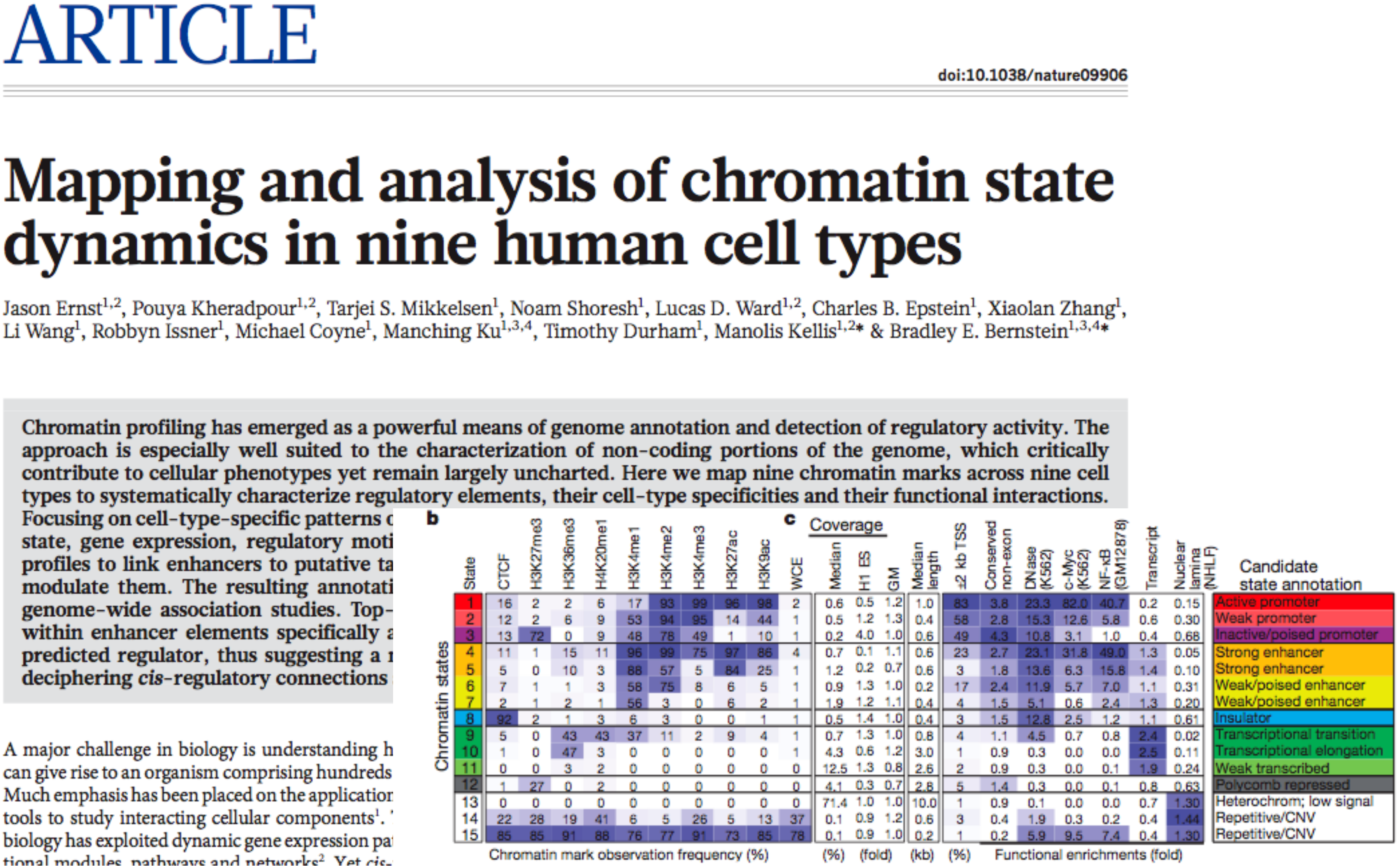


Figure 2 | Genome-wide enhancer predictions in human cells. a, We predict 36,589 enhancers in HeLa cells on the basis of chromatin signatures for H3K4me1 and H3K4me3 as determined by ChIP-chip using genome-wide tiling microarrays and condensed enhancer microarrays (see Supplementary Information). Enhancer predictions are located at the centre of 10-kb

Chromatin states



ARTICLE

doi:10.1038/nature09906

Mapping and analysis of chromatin state dynamics in nine human cell types

Jason Ernst[1,2], Pouya Kheradpour[1,2], Tarjei S. Mikkelsen[1], Noam Shoresh[1], Lucas D. Ward[1,2], Charles B. Epstein[1], Xiaolan Zhang[1], Li Wang[1], Robbyn Issner[1], Michael Coyne[1], Manching Ku[1,3,4], Timothy Durham[1], Manolis Kellis[1,2]* & Bradley E. Bernstein[1,3,4]*

**Chromatin profiling has emerged as a powerful means of genome annotation and detection of regulatory activity. The approach is especially well suited to the characterization of non-coding portions of the genome, which critically contribute to cellular phenotypes yet remain largely uncharted. Here we map nine chromatin marks across nine cell types to systematically characterize regulatory elements, their cell-type specificities and their functional interactions.
Focusing on cell-type-specific patterns o
state, gene expression, regulatory moti
profiles to link enhancers to putative ta
modulate them. The resulting annotati
genome-wide association studies. Top-
within enhancer elements specifically a
predicted regulator, thus suggesting a r
deciphering *cis*-regulatory connections**

A major challenge in biology is understanding h
can give rise to an organism comprising hundreds
Much emphasis has been placed on the application
tools to study interacting cellular components[1]. '
biology has exploited dynamic gene expression pat
tional modules, pathways and networks[2]. Yet *cis*-

b

State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
1	16	2	2	6	17	93	99	96	98	2
2	12	2	6	9	53	94	95	14	44	1
3	13	72	0	9	48	78	49	1	10	1
4	11	1	15	11	96	99	75	97	86	4
5	5	0	10	3	88	57	5	84	25	1
6	7	1	1	3	58	75	8	6	5	1
7	2	1	2	1	56	3	0	6	2	1
8	92	2	1	3	6	3	0	0	1	1
9	5	0	43	43	37	11	2	9	4	1
10	1	0	47	3	0	0	0	0	0	1
11	0	0	3	2	0	0	0	0	0	0
12	1	27	0	2	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	22	28	19	41	6	5	26	5	13	37
15	85	85	91	88	76	77	91	73	85	78

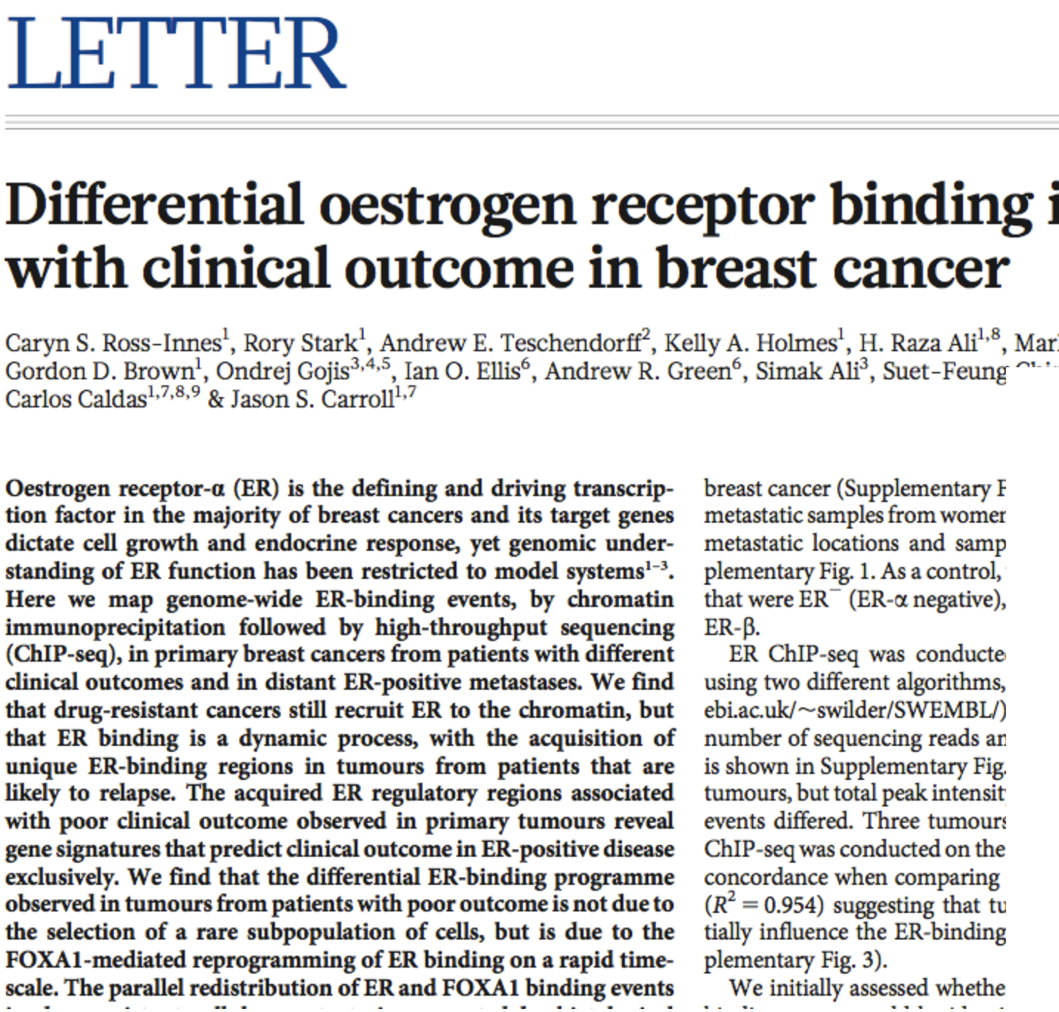
Chromatin states

Chromatin mark observation frequency (%)

c

State	Coverage: Median (%)	Coverage: H1 ES (fold)	Coverage: GM (fold)	Median length (kb)	±2 kb TSS (%)	Conserved non-exon	DNase (K562)	c-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear lamina (NHLF)	Candidate state annotation
1	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active promoter
2	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak promoter
3	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised promoter
4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb repressed
13	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV

Functional enrichments (fold)



LETTER

doi:10.1038/nature10730

Differential oestrogen receptor binding is associated with clinical outcome in breast cancer

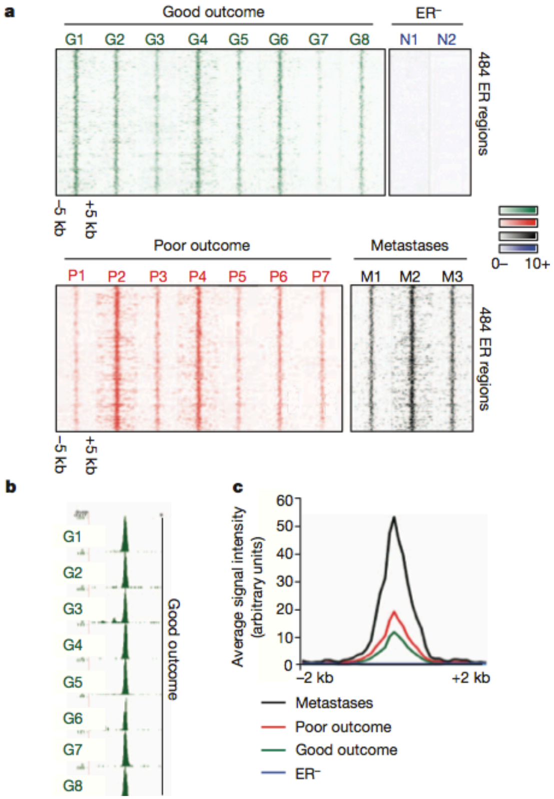
Caryn S. Ross-Innes[1], Rory Stark[1], Andrew E. Teschendorff[2], Kelly A. Holmes[1], H. Raza Ali[1,8], Mark J. Dunning[1],
Gordon D. Brown[1], Ondrej Gojis[3,4,5], Ian O. Ellis[6], Andrew R. Green[6], Simak Ali[3], Suet-Feung
Carlos Caldas[1,7,8,9] & Jason S. Carroll[1,7]

Oestrogen receptor-α (ER) is the defining and driving transcription factor in the majority of breast cancers and its target genes dictate cell growth and endocrine response, yet genomic understanding of ER function has been restricted to model systems[1–3]. Here we map genome-wide ER-binding events, by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), in primary breast cancers from patients with different clinical outcomes and in distant ER-positive metastases. We find that drug-resistant cancers still recruit ER to the chromatin, but that ER binding is a dynamic process, with the acquisition of unique ER-binding regions in tumours from patients that are likely to relapse. The acquired ER regulatory regions associated with poor clinical outcome observed in primary tumours reveal gene signatures that predict clinical outcome in ER-positive disease exclusively. We find that the differential ER-binding programme observed in tumours from patients with poor outcome is not due to the selection of a rare subpopulation of cells, but is due to the FOXA1-mediated reprogramming of ER binding on a rapid timescale. The parallel redistribution of ER and FOXA1 binding events

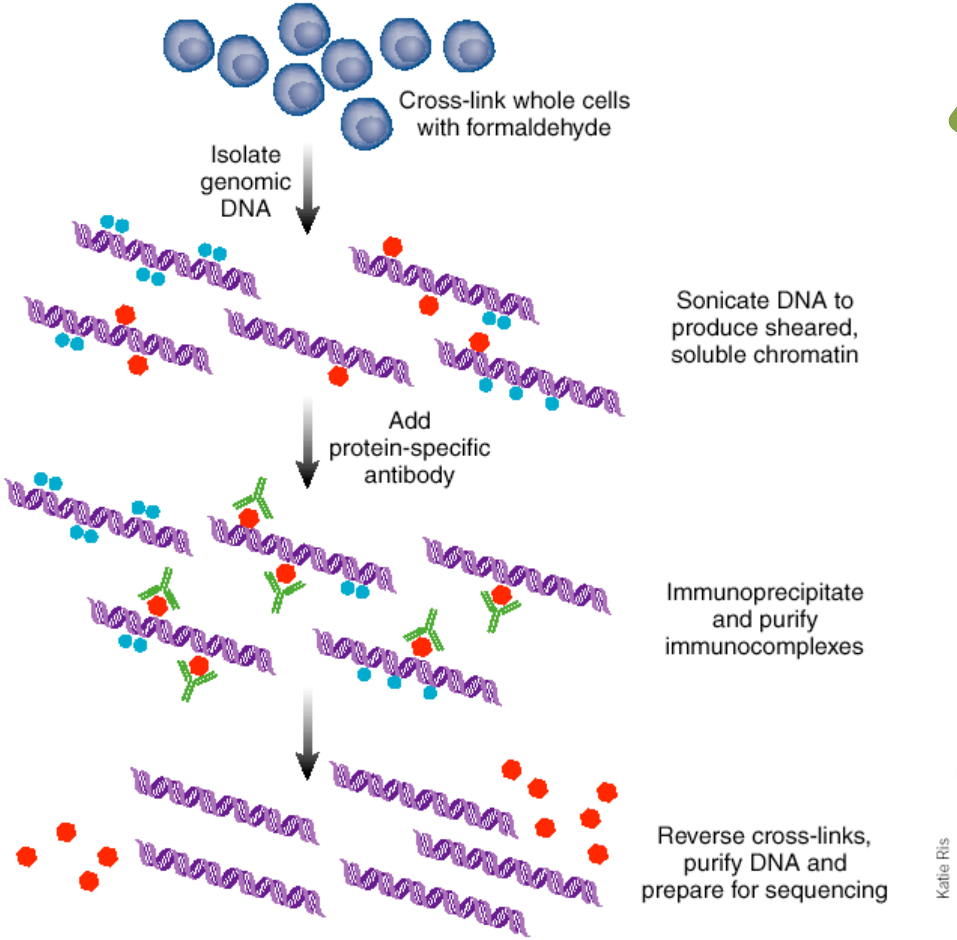
breast cancer (Supplementary F
metastatic samples from womer
metastatic locations and samp
plementary Fig. 1. As a control,
that were ER$^-$ (ER-α negative),
ER-β.

ER ChIP-seq was conducte
using two different algorithms,
ebi.ac.uk/~swilder/SWEMBL/)
number of sequencing reads ar
is shown in Supplementary Fig.
tumours, but total peak intensit
events differed. Three tumours
ChIP-seq was conducted on the
concordance when comparing
($R^2 = 0.954$) suggesting that tu
tially influence the ER-binding
plementary Fig. 3).

We initially assessed whethe



ChIP-seq



Cross-link whole cells with formaldehyde

Isolate genomic DNA

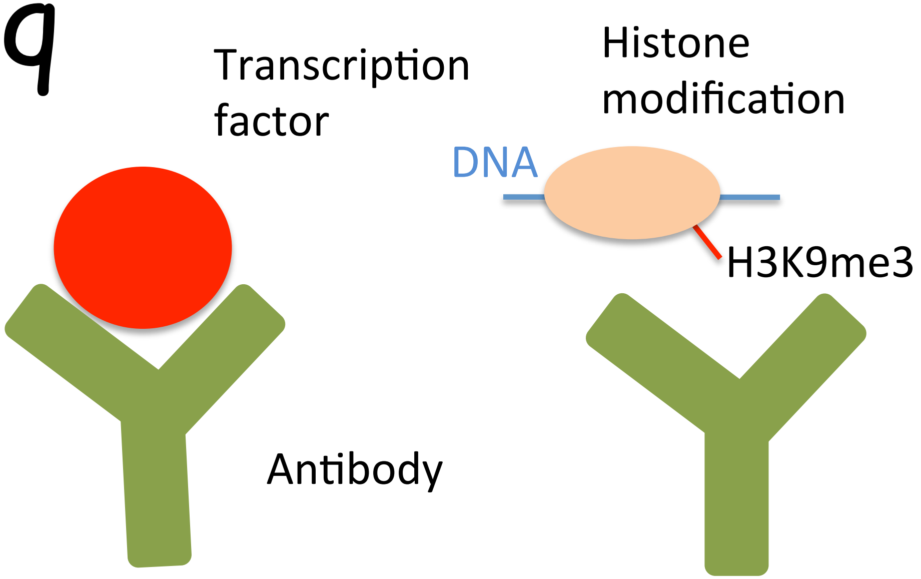
Sonicate DNA to produce sheared, soluble chromatin

Add protein-specific antibody

Immunoprecipitate and purify immunocomplexes

Reverse cross-links, purify DNA and prepare for sequencing

Katie Ris



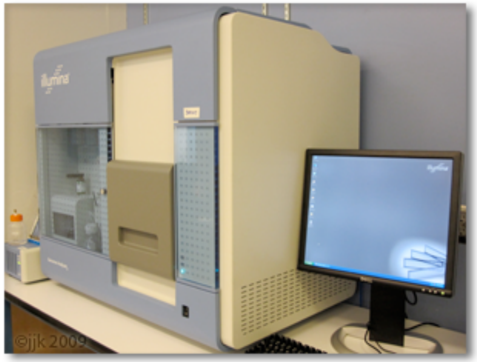
Transcription factor

Histone modification

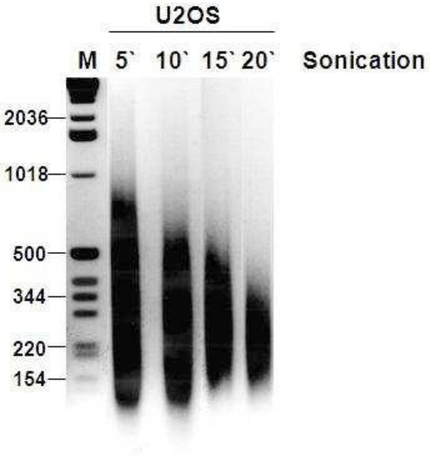
DNA

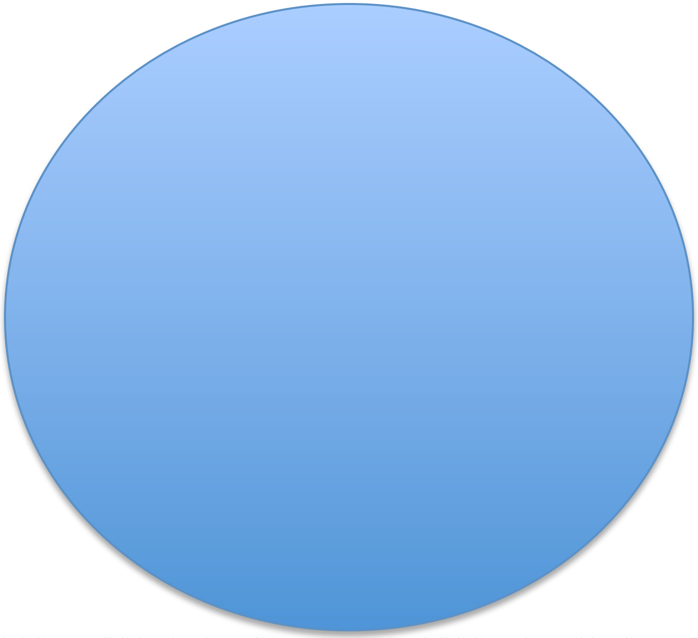
H3K9me3

Antibody



Illumina

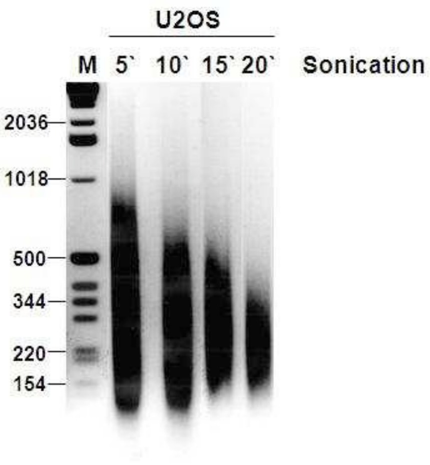


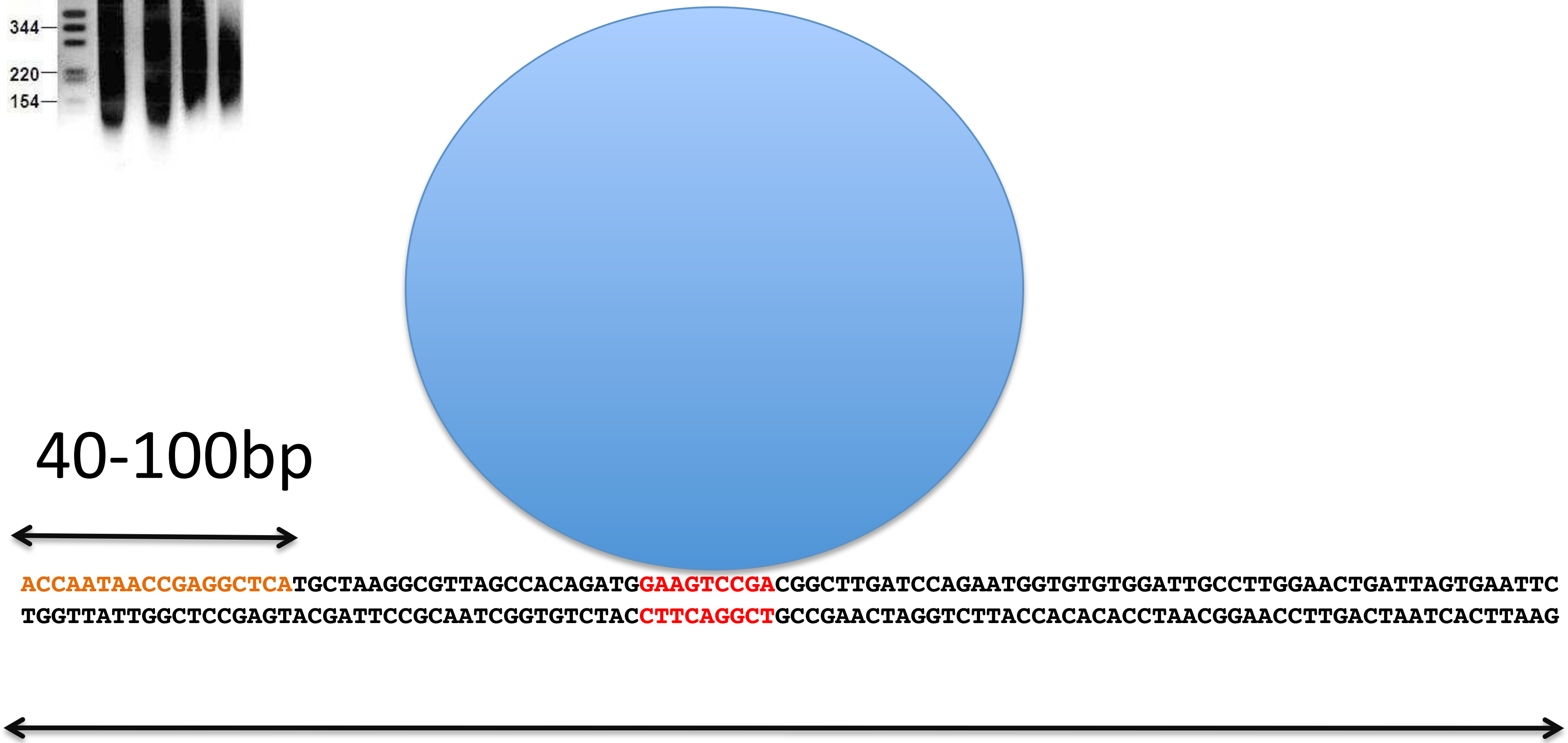


ACCAATAACCGAGGCTCATGCTAAGGCGTTAGCCACAGATGGAAGTCCGACGGCTTGATCCAGAATGGTGTGTGGATTGCCTTGGAACTGATTAGTGAATTC
TGGTTATTGGCTCCGAGTACGATTCCGCAATCGGTGTCTACCTTCAGGCTGCCGAACTAGGTCTTACCACACACCTAACGGAACCTTGACTAATCACTTAAG

↔

Average length ~ 170bp

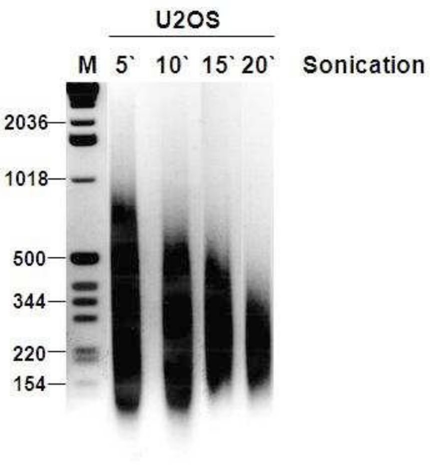


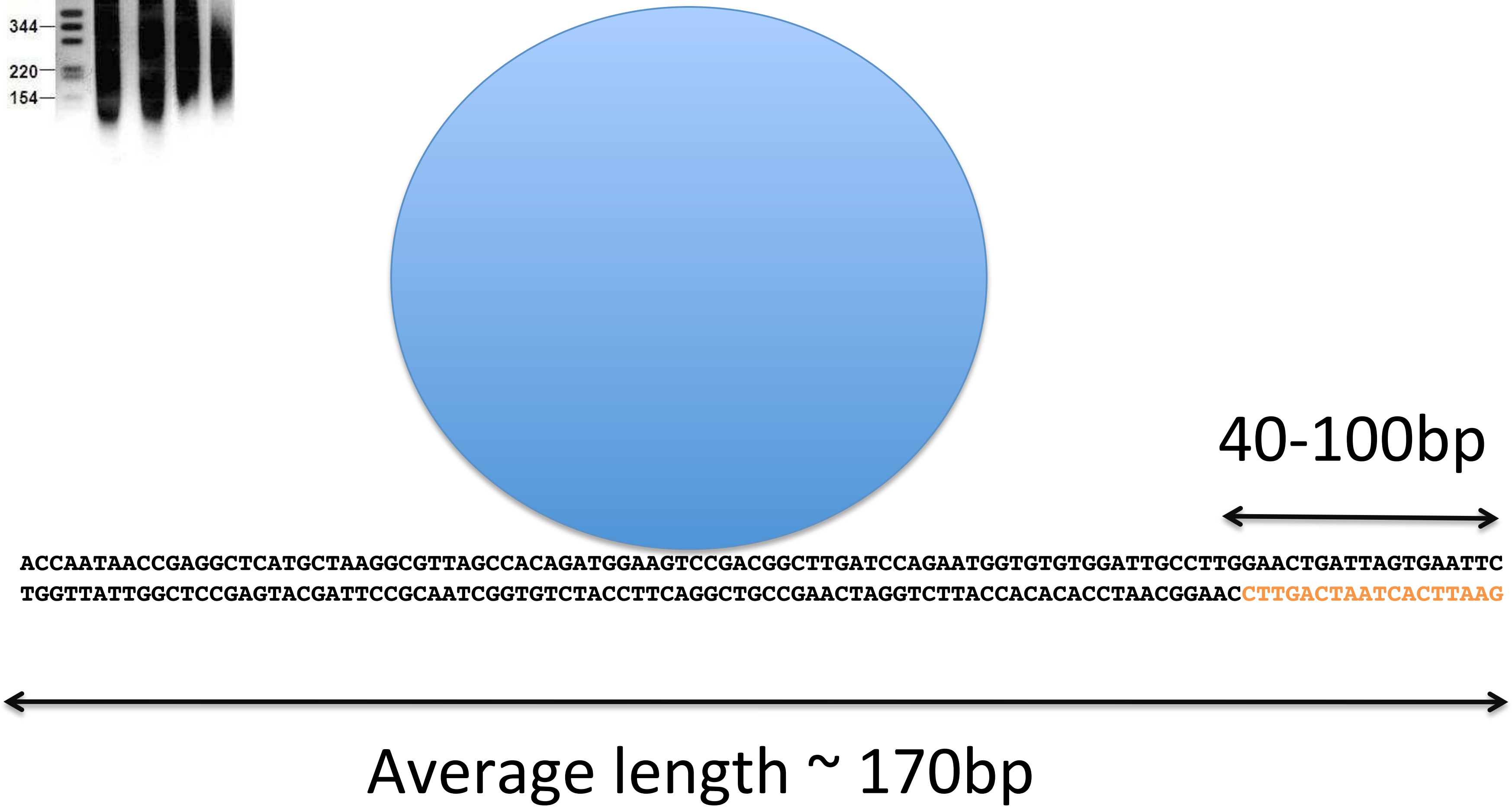


40-100bp

```
ACCAATAACCGAGGCTCATGCTAAGGCGTTAGCCACAGATGGAAGTCCGACGGCTTGATCCAGAATGGTGTGTGGATTGCCTTGGAACTGATTAGTGAATTC
TGGTTATTGGCTCCGAGTACGATTCCGCAATCGGTGTCTACCTTCAGGCTGCCGAACTAGGTCTTACCACACACCTAACGGAACCTTGACTAATCACTTAAG
```

Average length ~ 170bp





40-100bp

ACCAATAACCGAGGCTCATGCTAAGGCGTTAGCCACAGATGGAAGTCCGACGGCTTGATCCAGAATGGTGTGTGGATTGCCTTGGAACTGATTAGTGAATTC
TGGTTATTGGCTCCGAGTACGATTCCGCAATCGGTGTCTACCTTCAGGCTGCCGAACTAGGTCTTACCACACACCTAACGGAACCTTGACTAATCACTTAAG

Average length ~ 170bp

BWA tutorial (for aligning single end reads to genome)

- Get genome, e.g., from UCSC
  - http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz

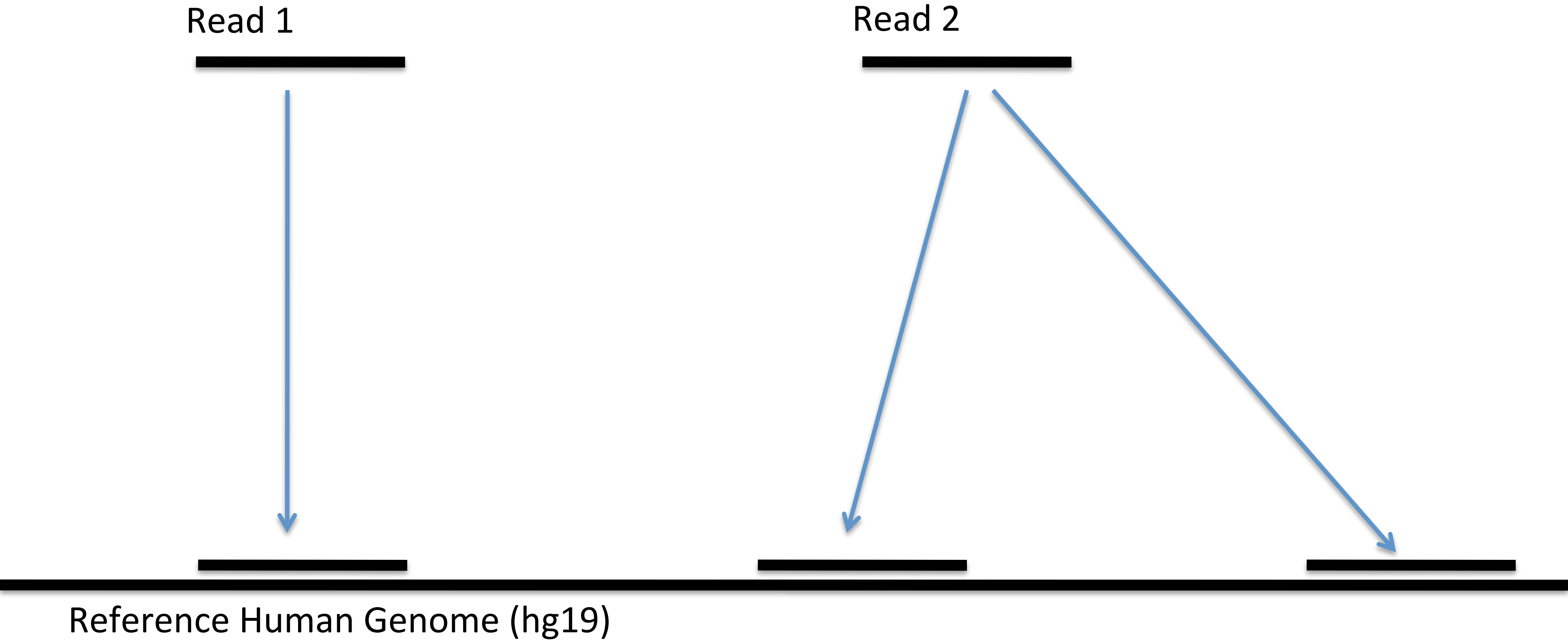
- Combine into 1 file
  - tar zvfx chromFa.tar.gz
  - cat *.fa > wg.fa

- Indexing the genome
  - bwa index -p hg19bwaidx -a bwtsw wg.fa

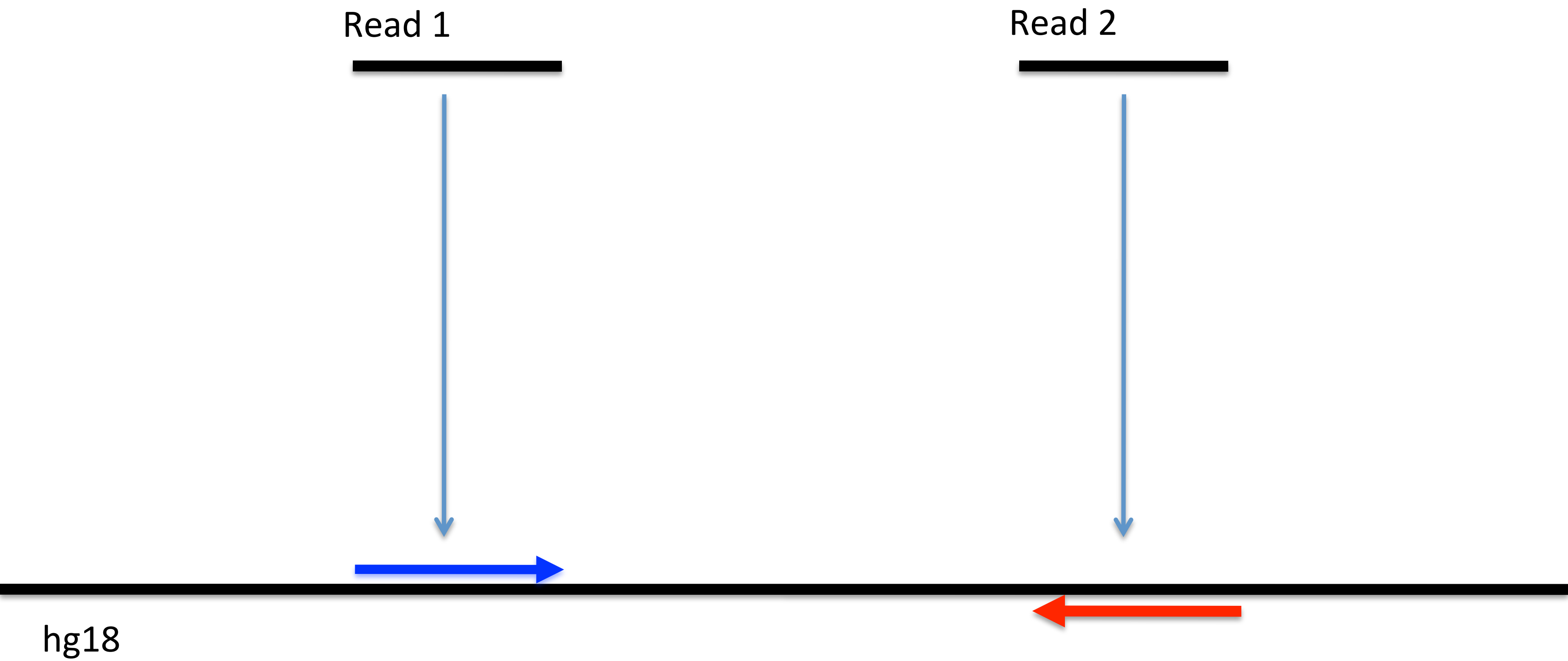
- Align ChIP reads to reference genome
  - bwa aln -t 4 hg19bwaidx s_3_sequence.txt.gz > s_3_sequence.txt.bwa
- Convert to SAM format
  - bwa samse hg19bwaidx s_3_sequence.txt.bwa s_3_sequence.txt.gz > s_3_sequence.txt.sam

- Align input reads to same reference genome
  - bwa aln -t 4 hg19bwaidx s_4_sequence.txt.gz > s_4_sequence.txt.bwa
- Convert to SAM format
  - bwa samse hg19bwaidx s_4_sequence.txt.bwa s_4_sequence.txt.gz > s_4_sequence.txt.sam

Reads can map to multiple locations/chromosomes



Reads map to one strand or the other



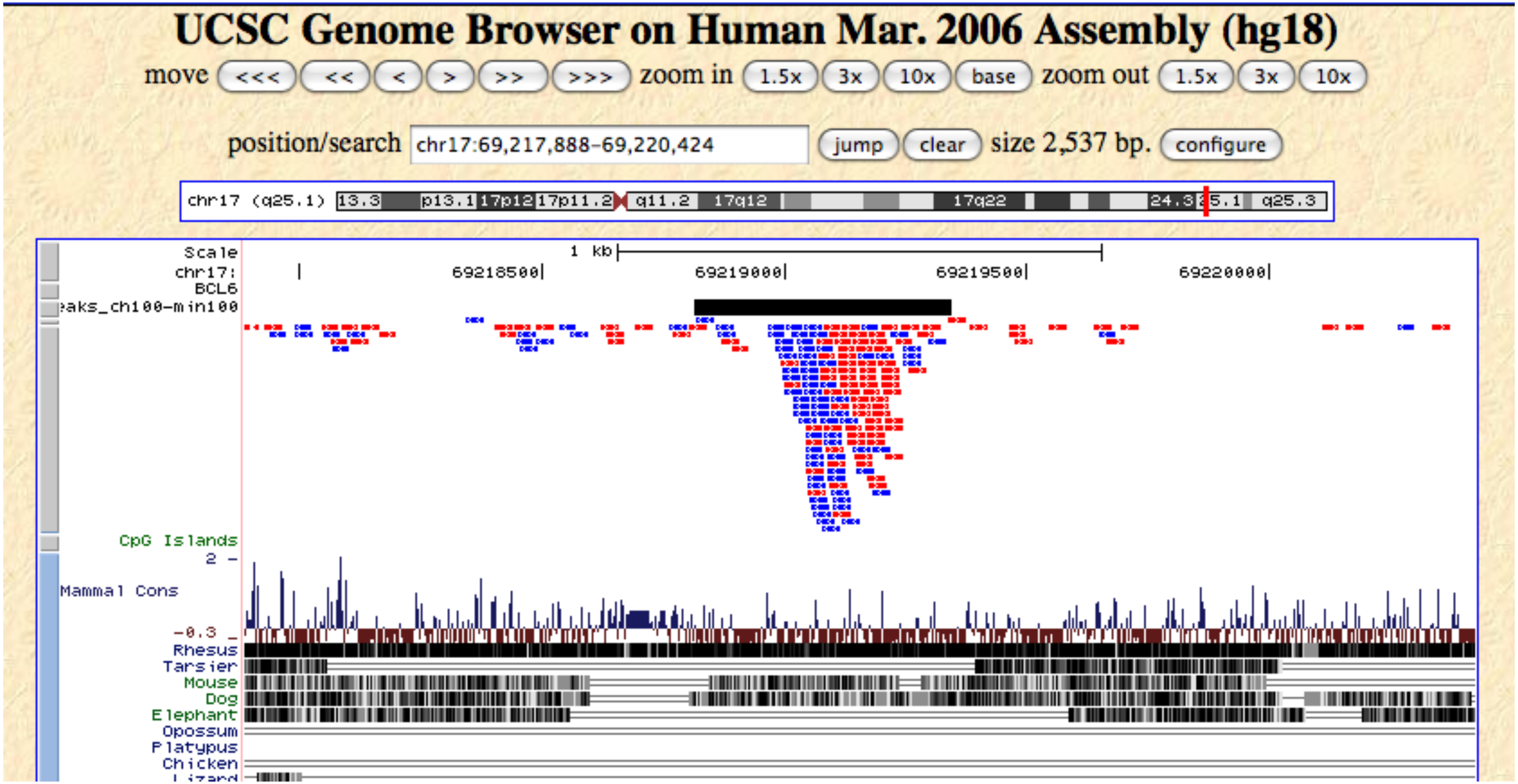
SAM format

```
DH1608P1_0130:6:1103:10579:166379#TTAGGC      16     chr1    1249828 37     51M     *      0      0
GGGCGTGACTCTGATCTCAGGCATCGTCTCCGCCGCGCTCCCGGACCCGCG     eb`XXYbZdadee^ceV]X][ccTcc^ebeece
eeeWbeeeeeeeceeaee     XX:Z:NM_017871,32      NM:i:0  MD:Z:51

DH1608P1_0130:6:1102:3415:150915#TTAGGC 16     chr1    1249828 37     51M     *      0      0
GGGCGGGACTCTGATCTCAGGCATCGTCTCCGCCGCGCTCCCGGACCCGCG     BBBBBBBBBBBac]bbbceedaeddeZceeea_ba_\_eee
eeeedaeeee     XX:Z:NM_017871,32      NM:i:1  MD:Z:5T45

DH1608P1_0130:6:1102:13118:62644#TTAGGC 16     chr1    1249828 37     51M     *      0      0
GGGCGTGCCTCGGATCTCAGGCATCGTCTCCGCCGCGCTCCCGGACCCGCG     BBBBBBBBBBBBBBBBBBBBB`XTbSa`cffegdggeccbe
effdeggggg     XX:Z:NM_017871,32      NM:i:2  MD:Z:7A3T39

DH1608P1_0130:6:1203:3012:157120#TTAGGC 16     chr1    1249826 25     51M     *      0      0
AAGGCCGTGACTCTGATCTCAGCCCTCGTCTCCGCCGCGCTCCCGGACCCG     BBBBBBBB^`QWZZ]UXYSZSTFRU]Z__SO[adcc[acdV
\`Y]YWY][_     XX:Z:NM_017871,34      NM:i:3  MD:Z:4G17G1A26

DH1608P1_0130:6:2206:4445:12756#TTAGGC  16     chr1    1246336 25     1M3487N50M     *      0      0
CCAAAGGGTGTGACTCTGATCTCGGGCATCGTCTCCGCCGCGCTCCCGGAC     BBBBBBBBBBBBBBBBBBBBBBBB`YdddYdc\
cacaNddddcdddaeeee     XX:Z:NM_017871,37      NM:i:3  MD:Z:2C5C14A27

DH1608P1_0130:6:2203:7903:43788#TTAGGC  16     chr1    1246336 37     1M3487N50M     *      0      0
CCCAAGGGCGTGACTCTGATCTCAGGCATCGTCTCCGCCGCGCTCCCGGAC     adbe[fbcbccb_cb^cb^^c^edgegggggdf
ggefffgfbfggggegeg     XX:Z:NM_017871,37      NM:i:0  MD:Z:51
```

CIGAR string, eg 5M3487N46M = 5bp-long block, 3487 insert, 46bp-long block

MD tag, e.g, MD:Z:4T46 = 5 matches, 1 mismatch (T in read), 46 matches

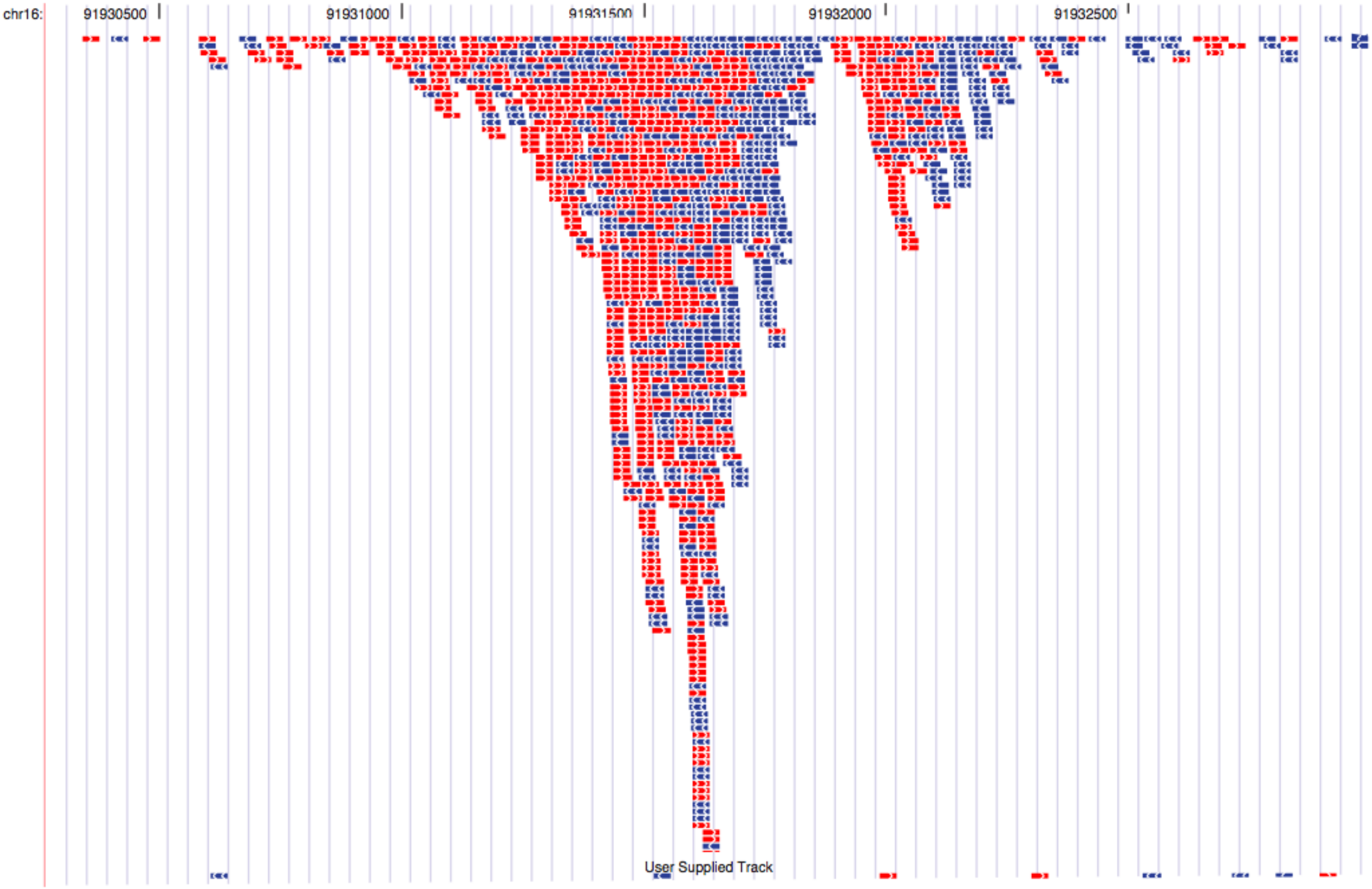
XT tag, e.g. XT:A:U = unique mapper; XT:A:R = more than 1 high-scoring matches



UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)

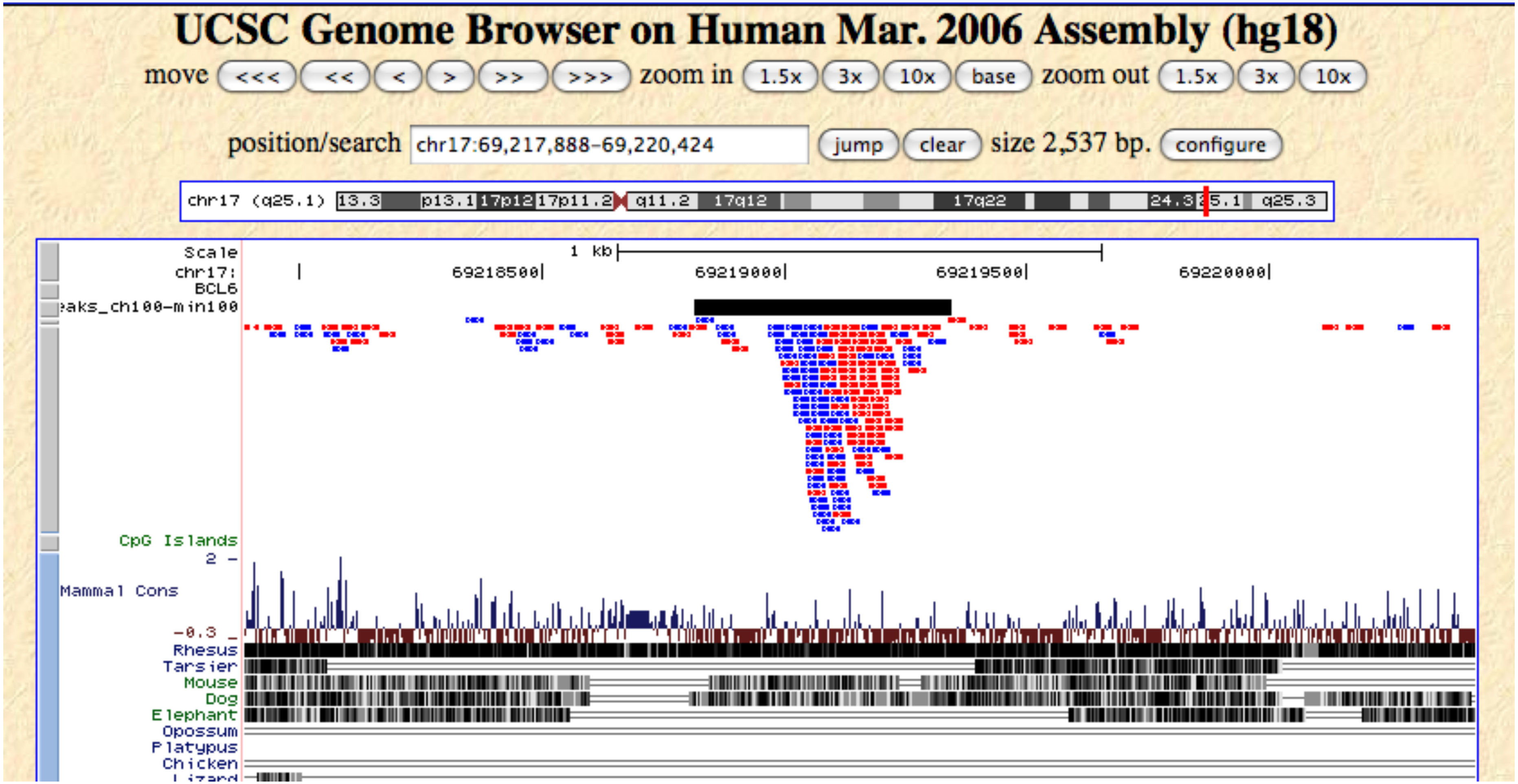
position/search chr17:69,217,888-69,220,424

size 2,537 bp.


chr16:
91930500
91931000
91931500
91932000
91932500
User Supplied Track

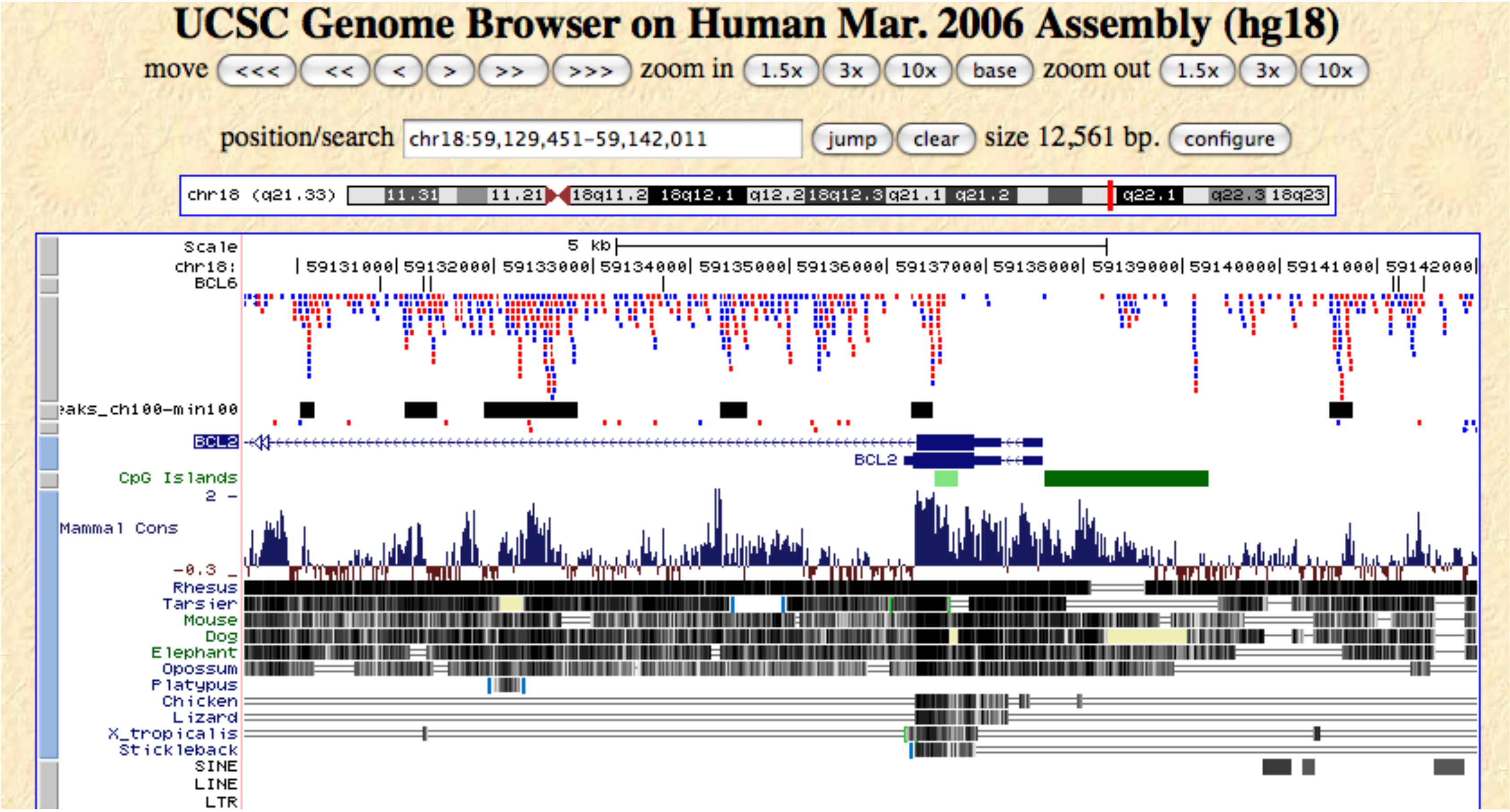
ChIP-seq peak calling

A nice peak



UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr17:69,217,888-69,220,424 jump clear size 2,537 bp. configure
chr17 (q25.1) 13.3 p13.1 17p12 17p11.2 q11.2 17q12 17q22 24.3 25.1 q25.3
Scale
1 kb
chr17:
69218500
69219000
69219500
69220000
BCL6
eaks_ch100-min100
CpG Islands
2 -
Mammal Cons
-0.3 _
Rhesus
Tarsier
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard

Not all peaks are nice

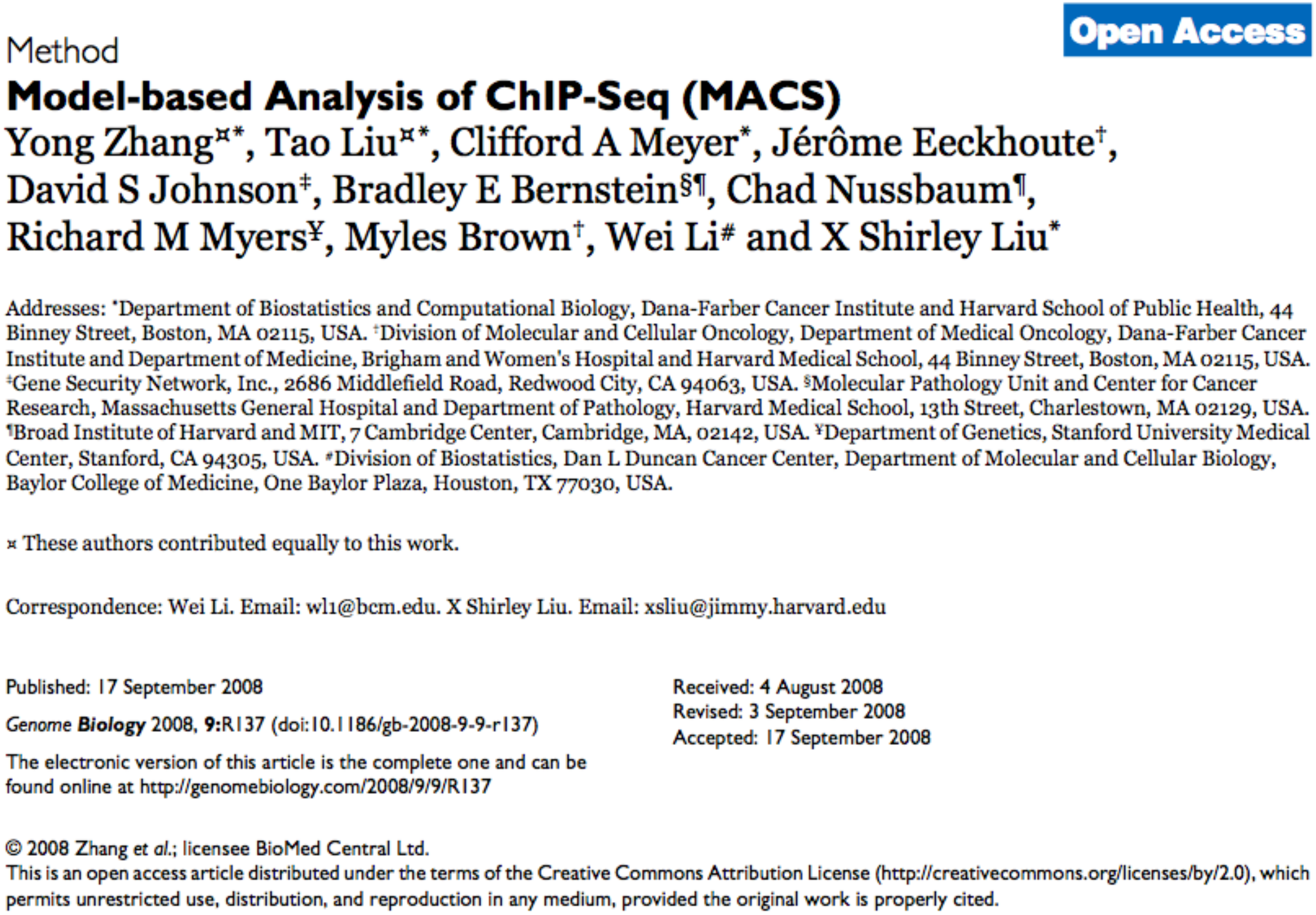


UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)

position/search chr18:59,129,451-59,142,011

size 12,561 bp.

MACS



Open Access

Method

Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nussbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

Addresses: [*]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, 44 Binney Street, Boston, MA 02115, USA. [†]Division of Molecular and Cellular Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute and Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA. [‡]Gene Security Network, Inc., 2686 Middlefield Road, Redwood City, CA 94063, USA. [§]Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital and Department of Pathology, Harvard Medical School, 13th Street, Charlestown, MA 02129, USA. [¶]Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142, USA. [¥]Department of Genetics, Stanford University Medical Center, Stanford, CA 94305, USA. [#]Division of Biostatistics, Dan L Duncan Cancer Center, Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

[¤] These authors contributed equally to this work.

Correspondence: Wei Li. Email: wl1@bcm.edu. X Shirley Liu. Email: xsliu@jimmy.harvard.edu

Published: 17 September 2008

Genome ***Biology*** 2008, **9:**R137 (doi:10.1186/gb-2008-9-9-r137)

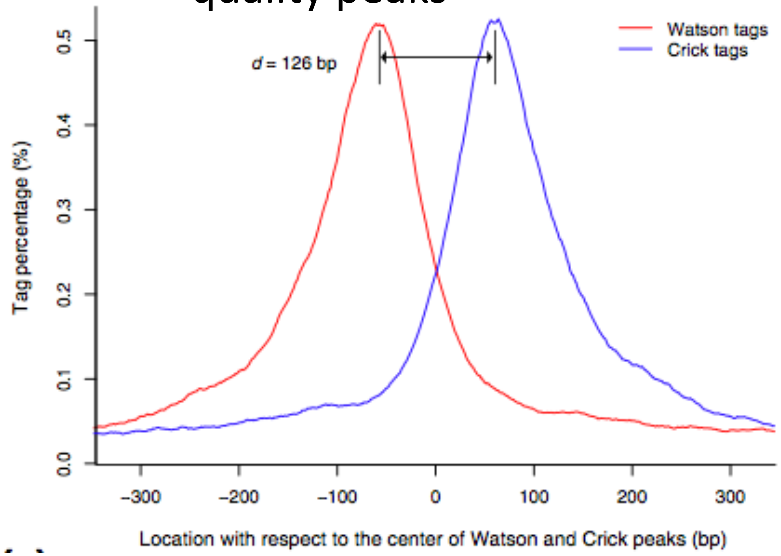
The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/9/R137

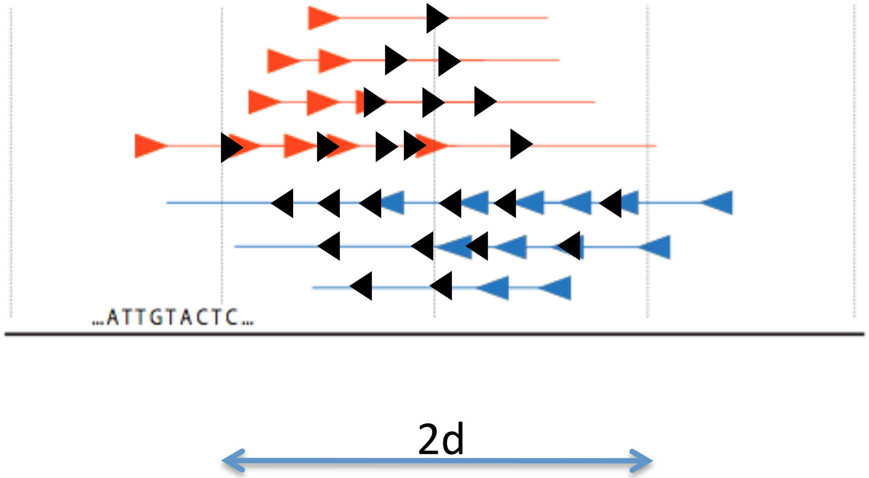
Received: 4 August 2008
Revised: 3 September 2008
Accepted: 17 September 2008

MACS

(a)

Estimate d based on high quality peaks





The Poisson distribution

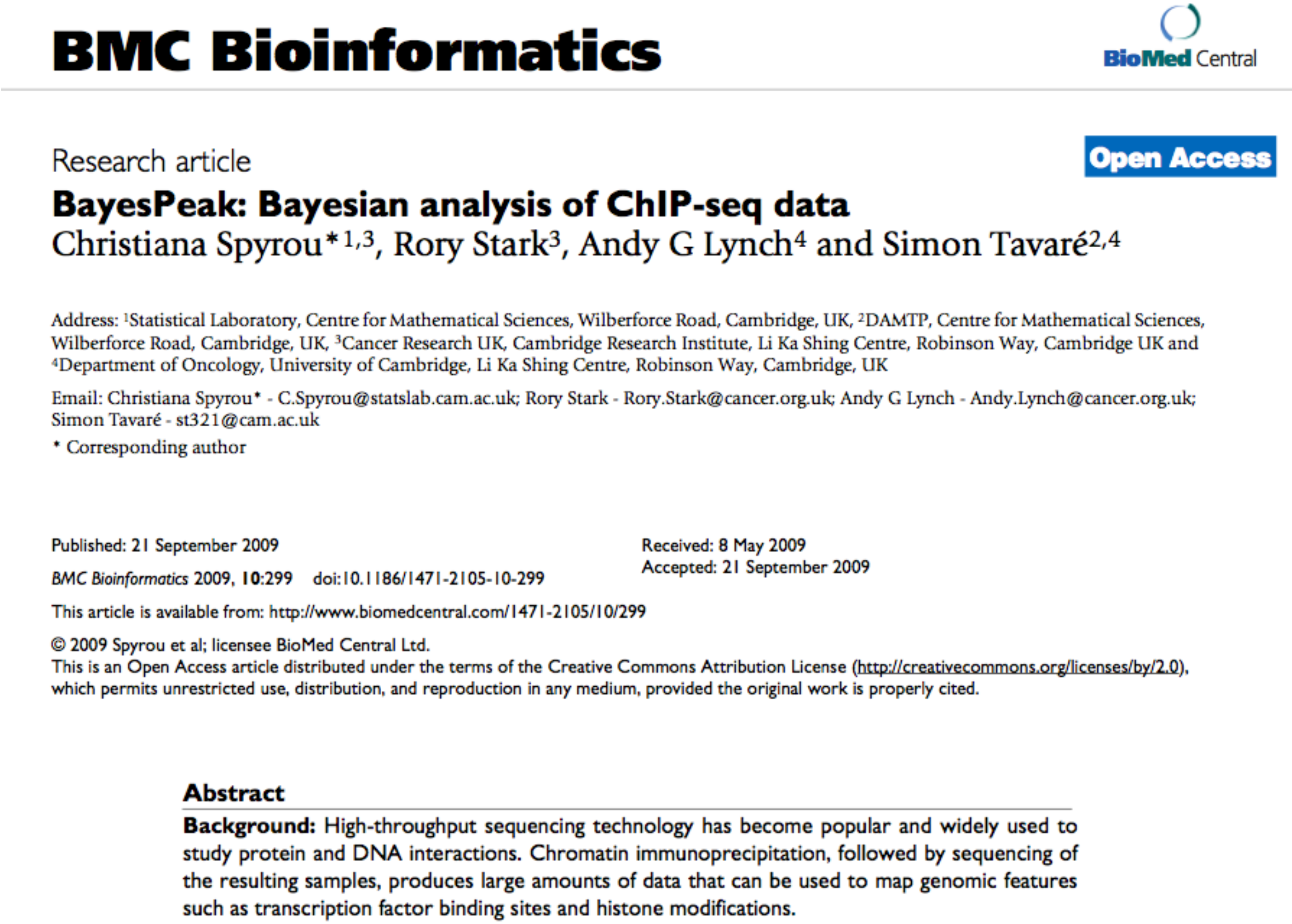
$$P(X \ge x) = 1 - \sum_{0}^{x-1} \frac{\lambda^x e^{-\lambda}}{x!}$$

λ=expected # of reads within an interval of 2*d* bp

$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$$

# in R P(X>=5|λ=0.001) is 1-sum(dpois(0:4, 0.001))

BayesPeak



BMC Bioinformatics

BioMed Central

Research article

Open Access

BayesPeak: Bayesian analysis of ChIP-seq data

Christiana Spyrou*[1,3], Rory Stark[3], Andy G Lynch[4] and Simon Tavaré[2,4]

Address: [1]Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, UK, [2]DAMTP, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, UK, [3]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge UK and [4]Department of Oncology, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, UK

Email: Christiana Spyrou* - C.Spyrou@statslab.cam.ac.uk; Rory Stark - Rory.Stark@cancer.org.uk; Andy G Lynch - Andy.Lynch@cancer.org.uk; Simon Tavaré - st321@cam.ac.uk

* Corresponding author

Published: 21 September 2009

BMC Bioinformatics 2009, **10**:299 doi:10.1186/1471-2105-10-299

Received: 8 May 2009
Accepted: 21 September 2009

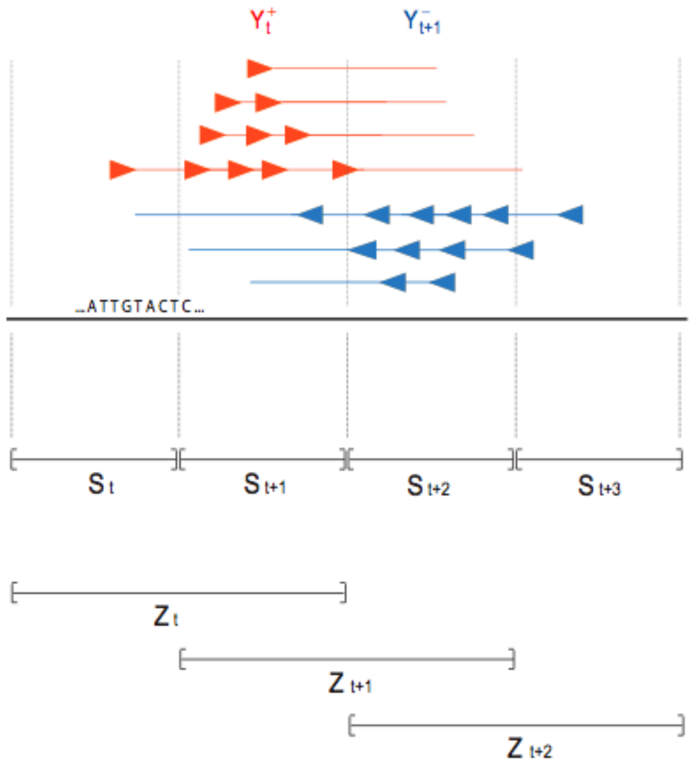
This article is available from: http://www.biomedcentral.com/1471-2105/10/299

Abstract

Background: High-throughput sequencing technology has become popular and widely used to study protein and DNA interactions. Chromatin immunoprecipitation, followed by sequencing of the resulting samples, produces large amounts of data that can be used to map genomic features such as transcription factor binding sites and histone modifications.

BayesPeak (Bayesian Hidden Markov Models)

Observed variable



Hidden states

$$Z_t = \begin{cases} 0 & \text{if} \quad (S_t, S_{t+1}) = (0,0) \\ 1 & \text{if} \quad (S_t, S_{t+1}) = (0,1) \\ 2 & \text{if} \quad (S_t, S_{t+1}) = (1,0) \\ 3 & \text{if} \quad (S_t, S_{t+1}) = (1,1) \end{cases}$$

The emission distributions of the model are

$$\begin{aligned} Y_t^+, Y_{t+1}^- \mid Z_t = 0 \sim &\quad \text{Poisson}(\lambda_0 \gamma^{w_t}) \\ Y_t^+, Y_{t+1}^- \mid Z_t = 1,2,3 \sim &\quad \text{Poisson}((\lambda_0 + \lambda_1)\gamma^{w_t}) \\ \lambda_0 \sim &\quad \Gamma(\alpha_0, \beta_0) \\ \lambda_1 \sim &\quad \Gamma(\alpha_1, \beta_1) \end{aligned}$$

Parameters estimated using Bayesian treatment

Other peak finders

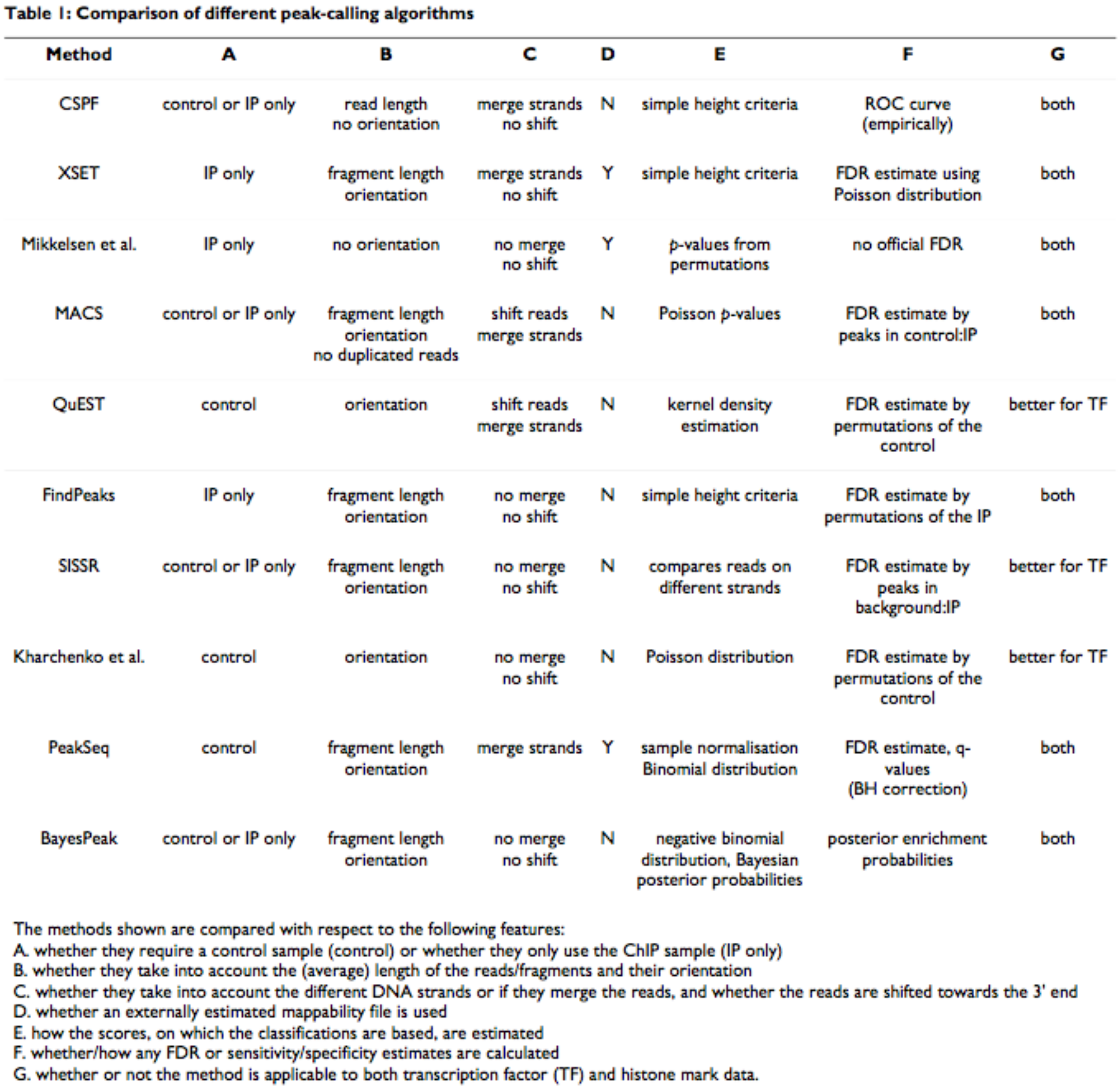


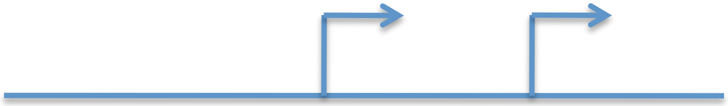
Table 1: Comparison of different peak-calling algorithms

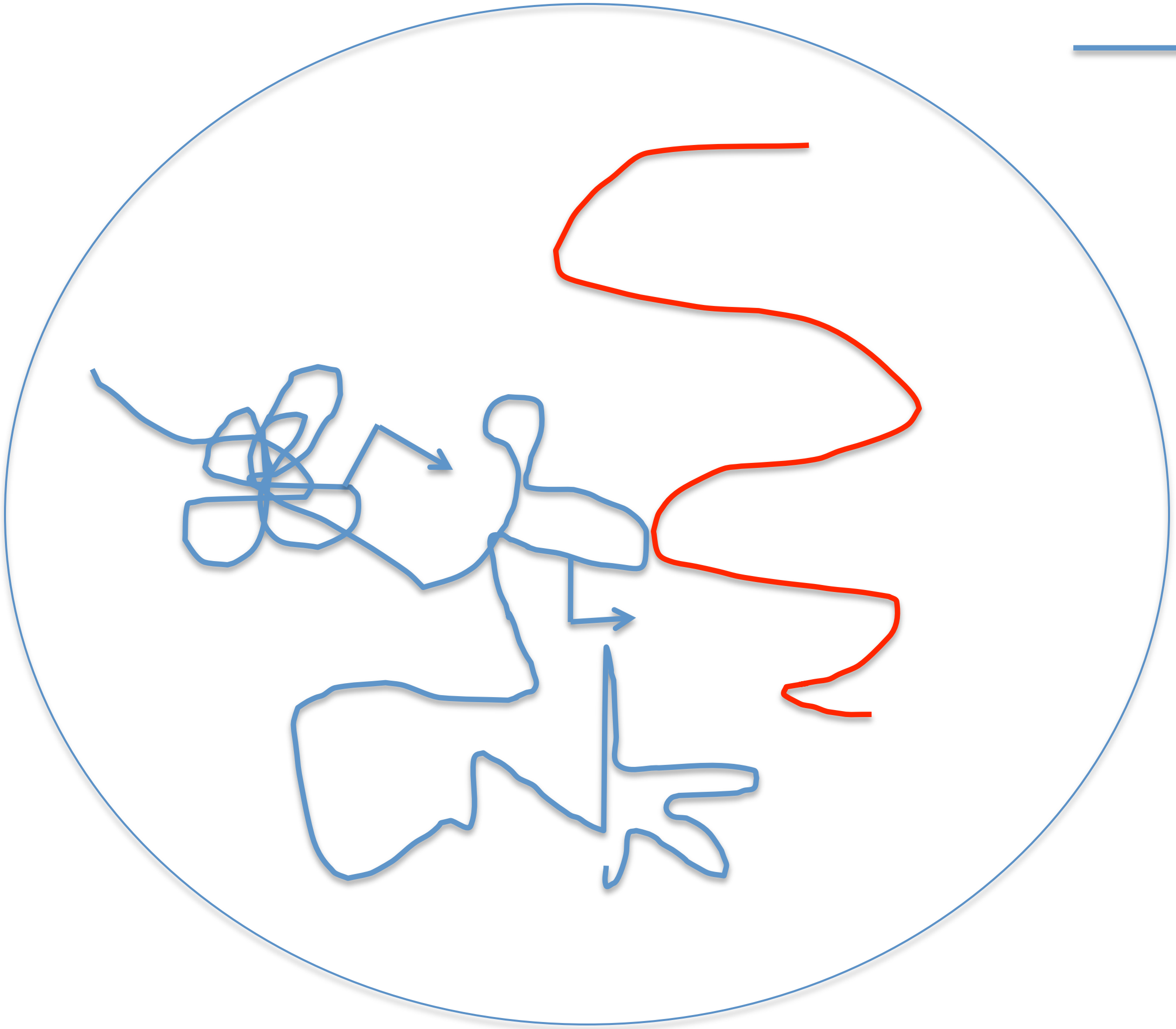
Method	A	B	C	D	E	F	G
CSPF	control or IP only	read length no orientation	merge strands no shift	N	simple height criteria	ROC curve (empirically)	both
XSET	IP only	fragment length orientation	merge strands no shift	Y	simple height criteria	FDR estimate using Poisson distribution	both
Mikkelsen et al.	IP only	no orientation	no merge no shift	Y	*p*-values from permutations	no official FDR	both
MACS	control or IP only	fragment length orientation no duplicated reads	shift reads merge strands	N	Poisson *p*-values	FDR estimate by peaks in control:IP	both
QuEST	control	orientation	shift reads merge strands	N	kernel density estimation	FDR estimate by permutations of the control	better for TF
FindPeaks	IP only	fragment length orientation	no merge no shift	N	simple height criteria	FDR estimate by permutations of the IP	both
SISSR	control or IP only	fragment length orientation	no merge no shift	N	compares reads on different strands	FDR estimate by peaks in background:IP	better for TF
Kharchenko et al.	control	orientation	no merge no shift	N	Poisson distribution	FDR estimate by permutations of the control	better for TF
PeakSeq	control	fragment length orientation	merge strands	Y	sample normalisation Binomial distribution	FDR estimate, q-values (BH correction)	both
BayesPeak	control or IP only	fragment length orientation	no merge no shift	N	negative binomial distribution, Bayesian posterior probabilities	posterior enrichment probabilities	both

The methods shown are compared with respect to the following features:
A. whether they require a control sample (control) or whether they only use the ChIP sample (IP only)
B. whether they take into account the (average) length of the reads/fragments and their orientation
C. whether they take into account the different DNA strands or if they merge the reads, and whether the reads are shifted towards the 3' end
D. whether an externally estimated mappability file is used
E. how the scores, on which the classifications are based, are estimated
F. whether/how any FDR or sensitivity/specificity estimates are calculated
G. whether or not the method is applicable to both transcription factor (TF) and histone mark data.

Mapping chromatin interactions using the Hi-C technique

The human genome is not linear





Terminology:

Intra-chromosomal interaction

Inter-chromosomal interaction

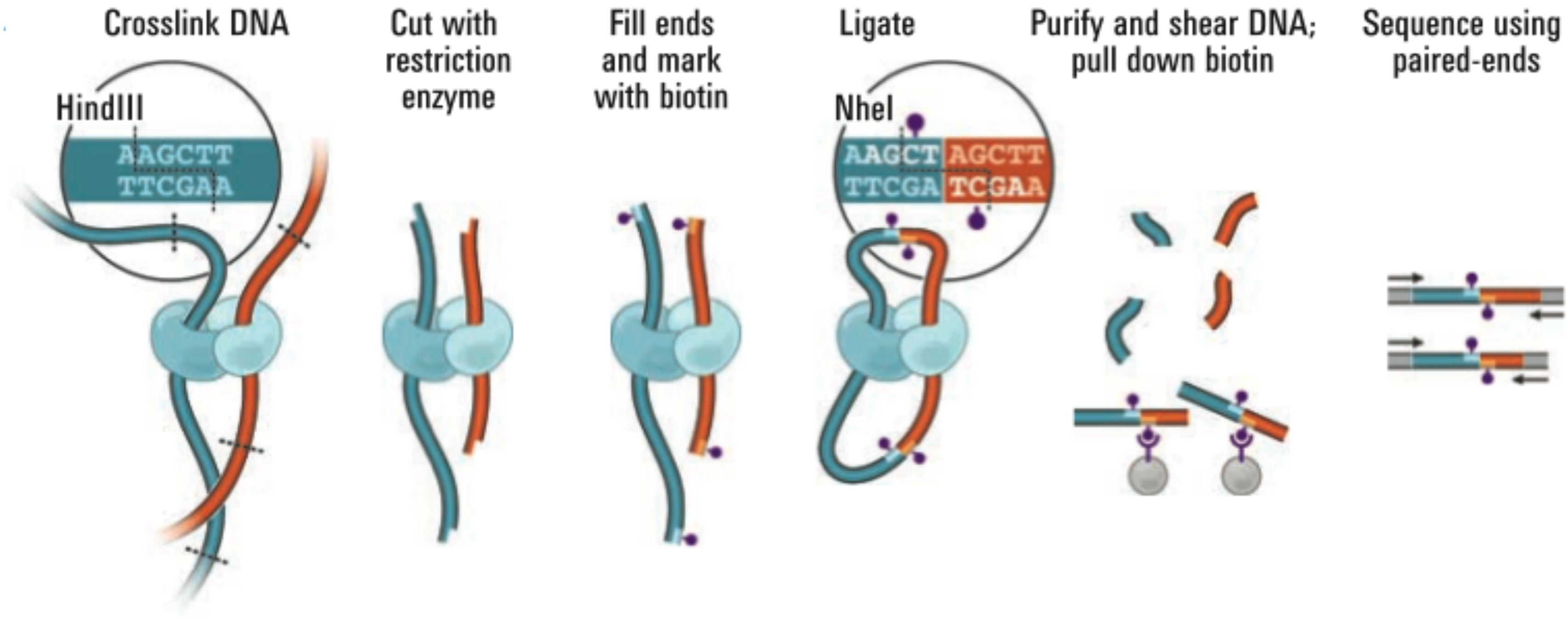
Interaction hub

How dynamic is the three dimensional architecture of the human genome ?

How does 3D localization impact gene expression ?

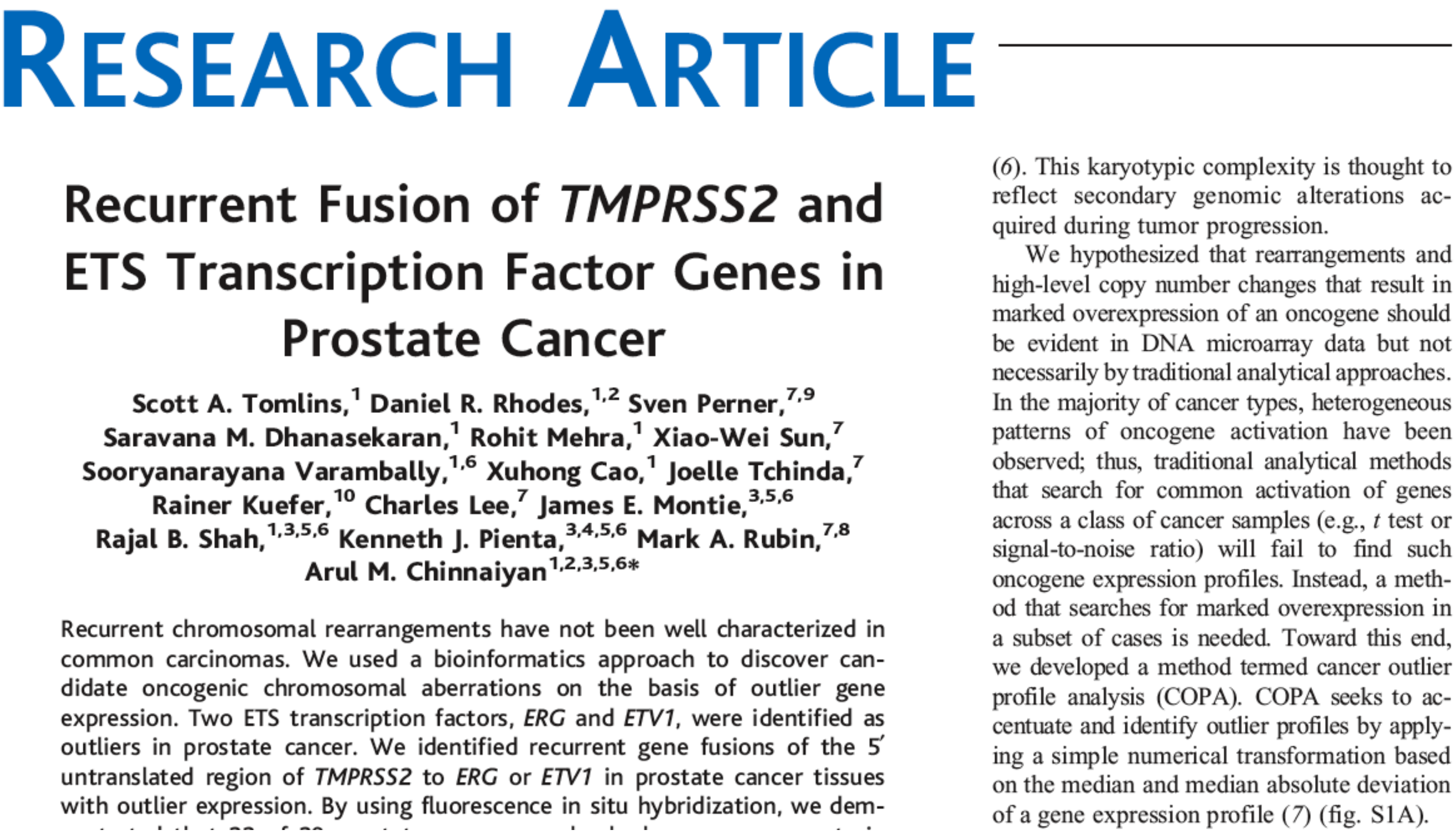
Is the 3D genome different in normal cells and cancer cells ?

Genome-scale mapping of chromatin interactions using HiC


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Lieberman-Aiden et al, 2009

Can a single (oncogenic) transcription factor induce global changes in chromatin structure ?



Research Article

Recurrent Fusion of *TMPRSS2* and ETS Transcription Factor Genes in Prostate Cancer

Scott A. Tomlins,[1] Daniel R. Rhodes,[1,2] Sven Perner,[7,9] Saravana M. Dhanasekaran,[1] Rohit Mehra,[1] Xiao-Wei Sun,[7] Sooryanarayana Varambally,[1,6] Xuhong Cao,[1] Joelle Tchinda,[7] Rainer Kuefer,[10] Charles Lee,[7] James E. Montie,[3,5,6] Rajal B. Shah,[1,3,5,6] Kenneth J. Pienta,[3,4,5,6] Mark A. Rubin,[7,8] Arul M. Chinnaiyan[1,2,3,5,6]*

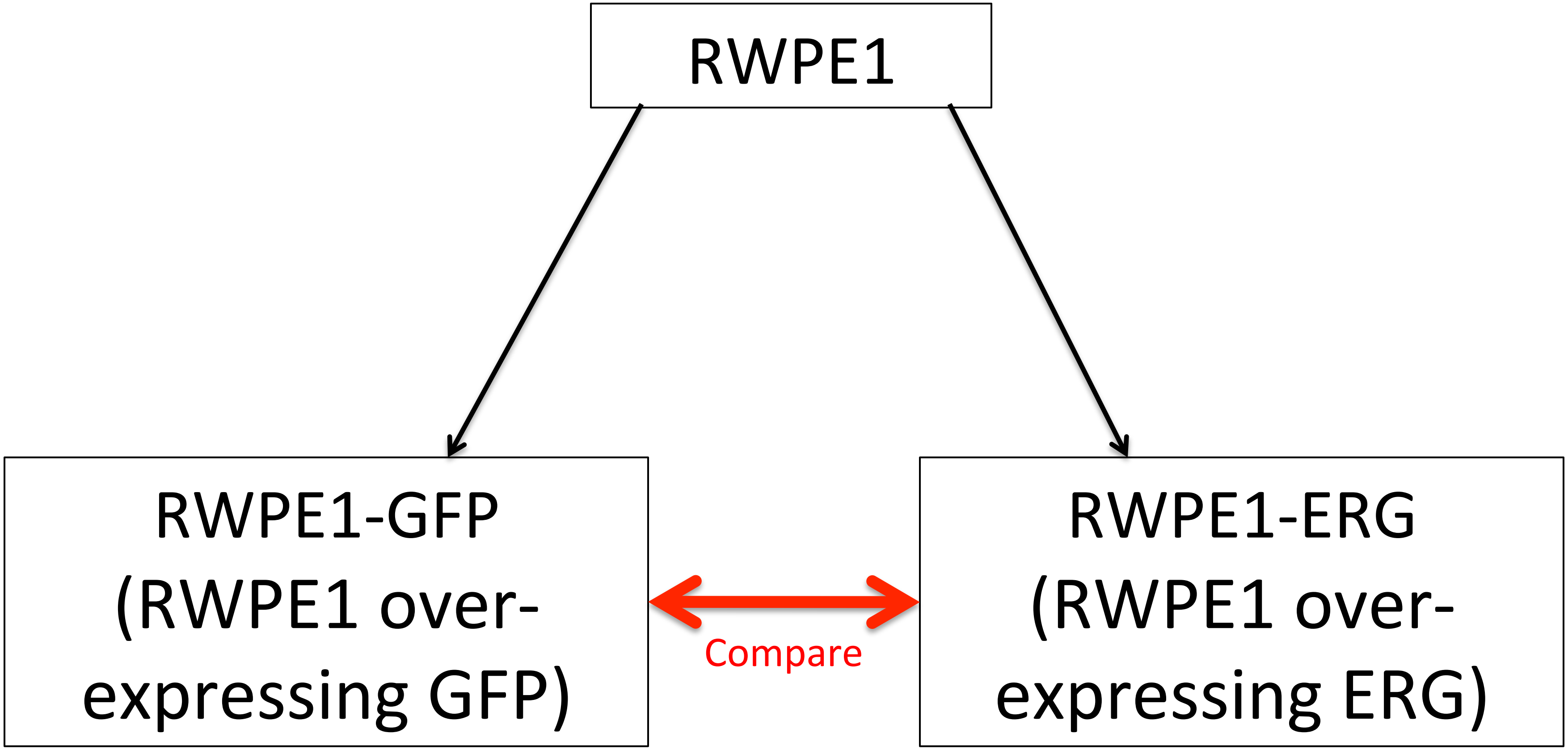
Recurrent chromosomal rearrangements have not been well characterized in common carcinomas. We used a bioinformatics approach to discover candidate oncogenic chromosomal aberrations on the basis of outlier gene expression. Two ETS transcription factors, *ERG* and *ETV1*, were identified as outliers in prostate cancer. We identified recurrent gene fusions of the 5′ untranslated region of *TMPRSS2* to *ERG* or *ETV1* in prostate cancer tissues with outlier expression. By using fluorescence in situ hybridization, we dem-

(*6*). This karyotypic complexity is thought to reflect secondary genomic alterations acquired during tumor progression.

We hypothesized that rearrangements and high-level copy number changes that result in marked overexpression of an oncogene should be evident in DNA microarray data but not necessarily by traditional analytical approaches. In the majority of cancer types, heterogeneous patterns of oncogene activation have been observed; thus, traditional analytical methods that search for common activation of genes across a class of cancer samples (e.g., *t* test or signal-to-noise ratio) will fail to find such oncogene expression profiles. Instead, a method that searches for marked overexpression in a subset of cases is needed. Toward this end, we developed a method termed cancer outlier profile analysis (COPA). COPA seeks to accentuate and identify outlier profiles by applying a simple numerical transformation based on the median and median absolute deviation of a gene expression profile (*7*) (fig. S1A).

ERG is fused to TMPRSS2 and over-expressed in > 50% of prostate tumors

Prostate epithelial lesion cell line (RWPE1)



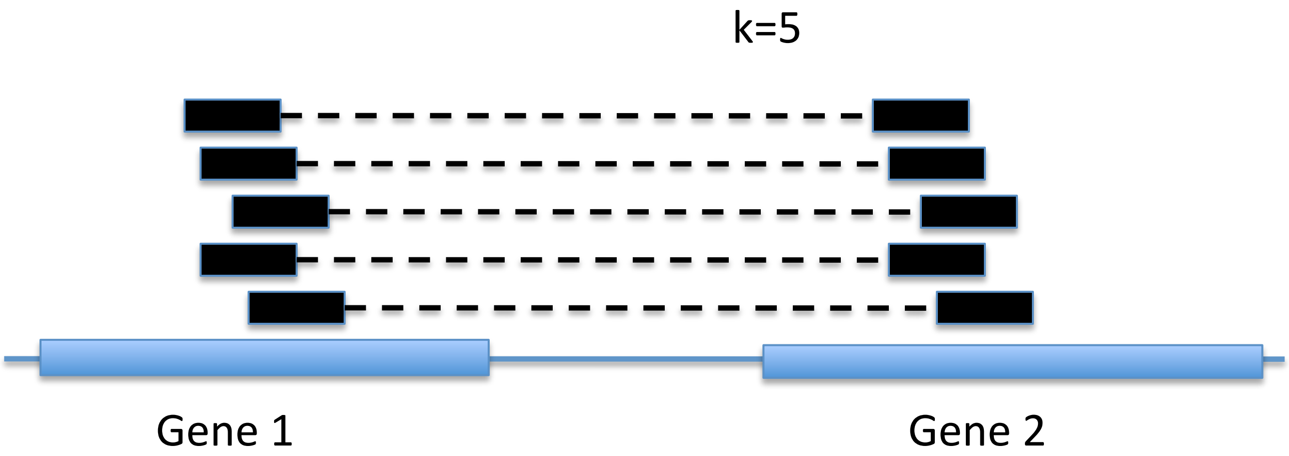
RWPE1

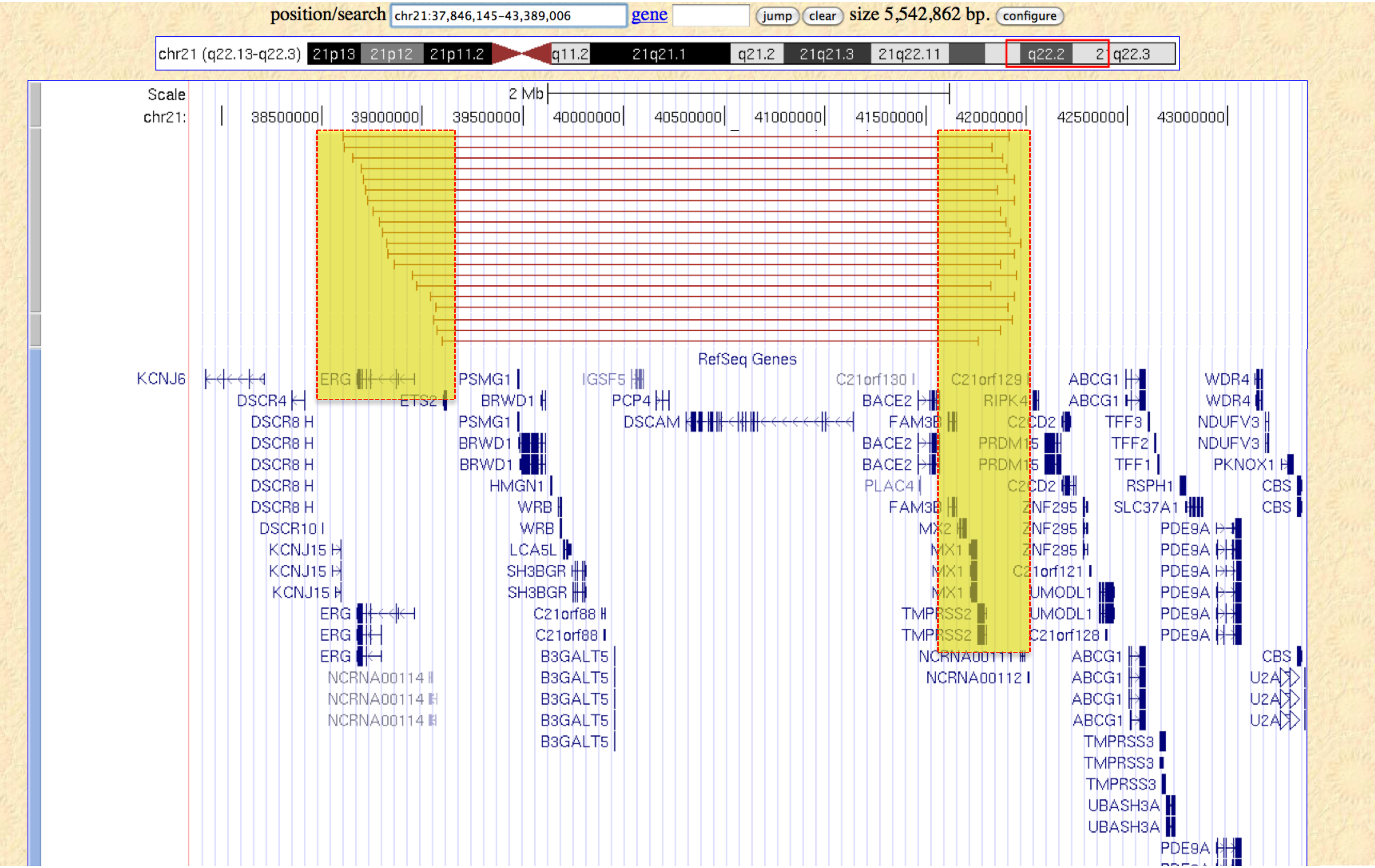
RWPE1-GFP (RWPE1 over-expressing GFP)

Compare

RWPE1-ERG (RWPE1 over-expressing ERG)

Quantifying proximity using HiC reads

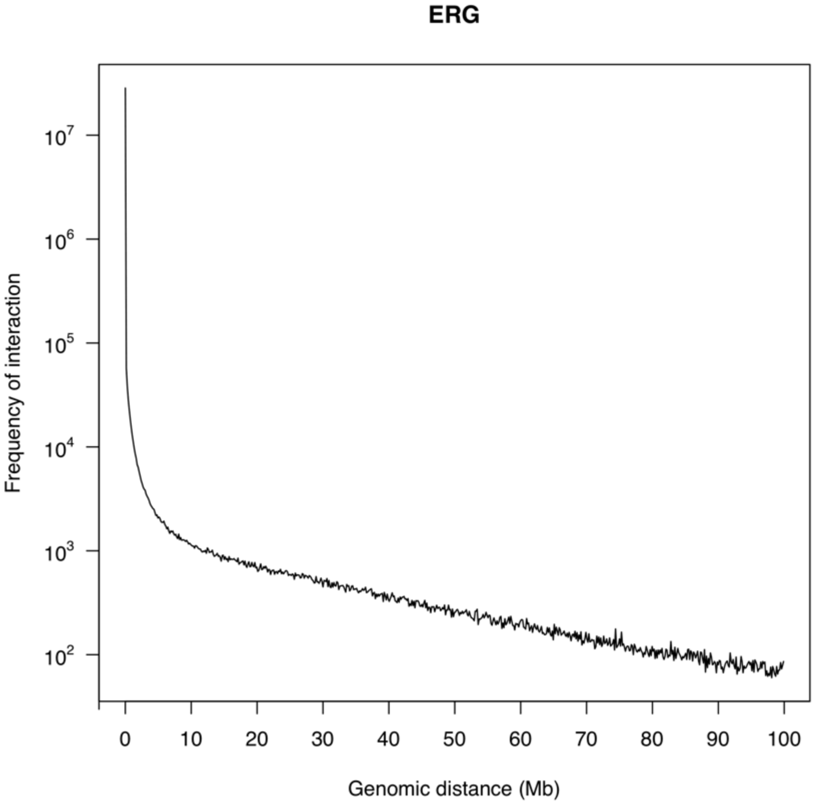


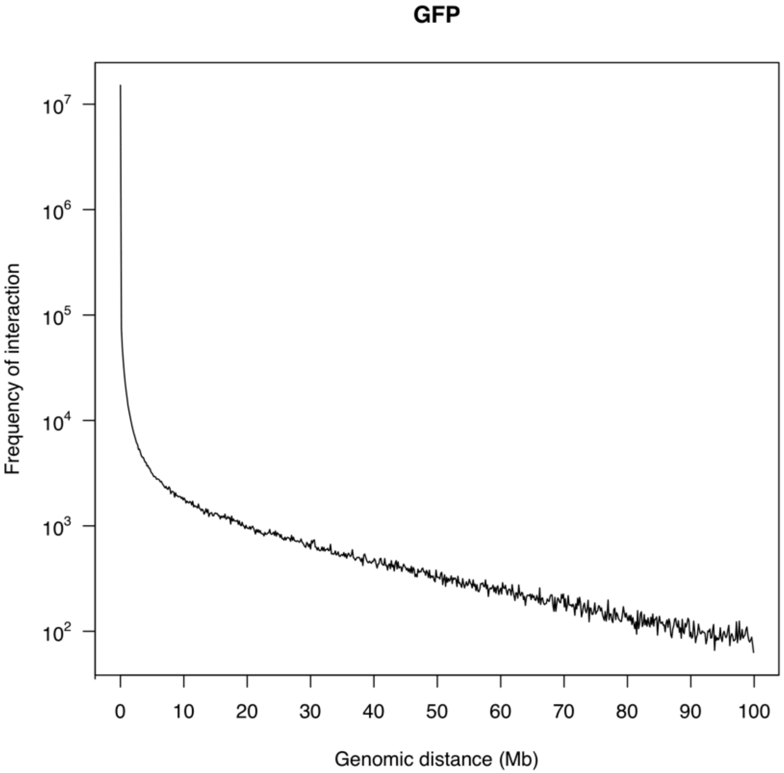


position/search chr21:37,846,145-43,389,006

size 5,542,862 bp.

Interaction frequency decreases with genomic distance





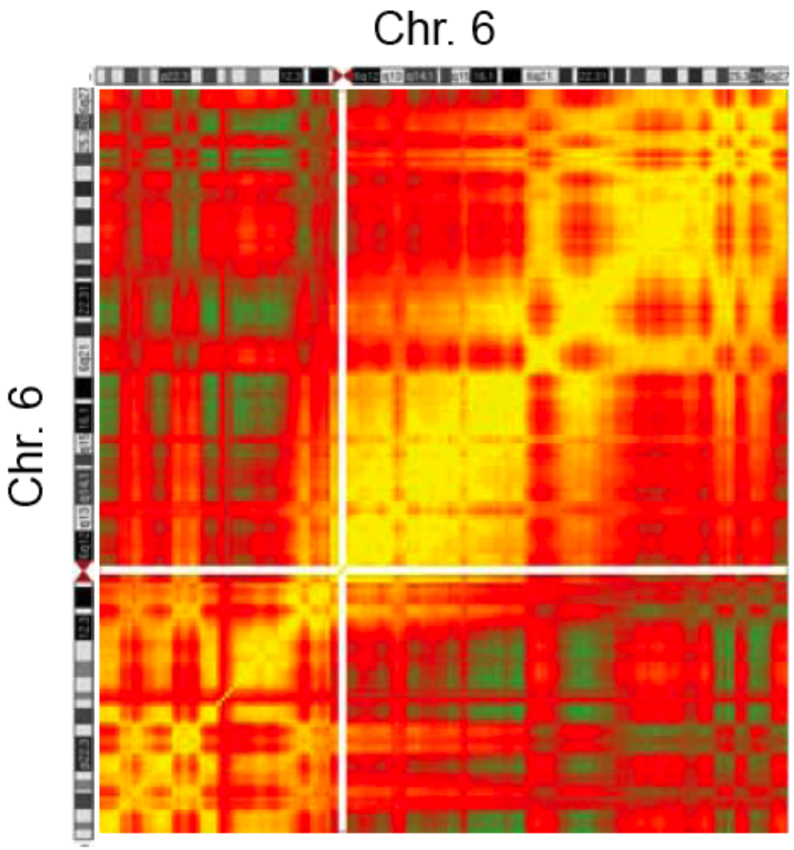
$$P_{\text{interaction}}(s) \sim s^{-1}$$

(500Kb-1Mb range)

$$P_{\text{interaction}}(s) \sim s^{-1}$$

(500Kb-1Mb range)

Contact maps show domains of interactions



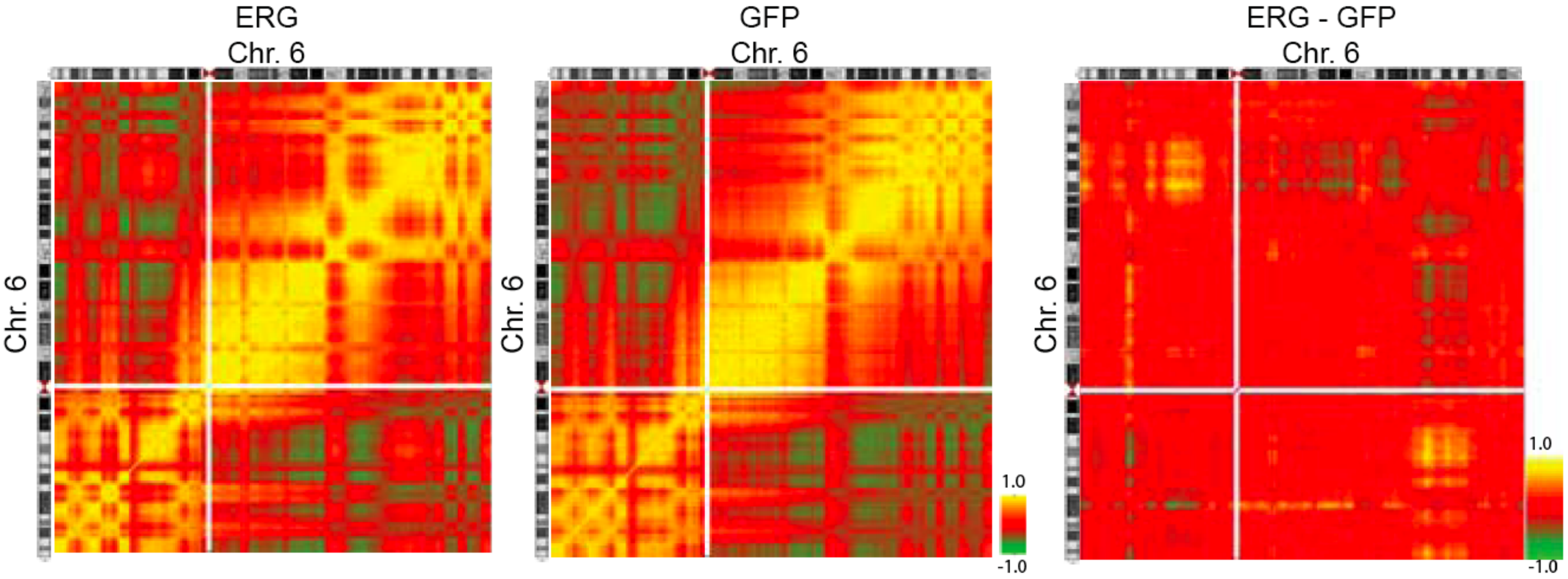
🟩 Less interaction than expected

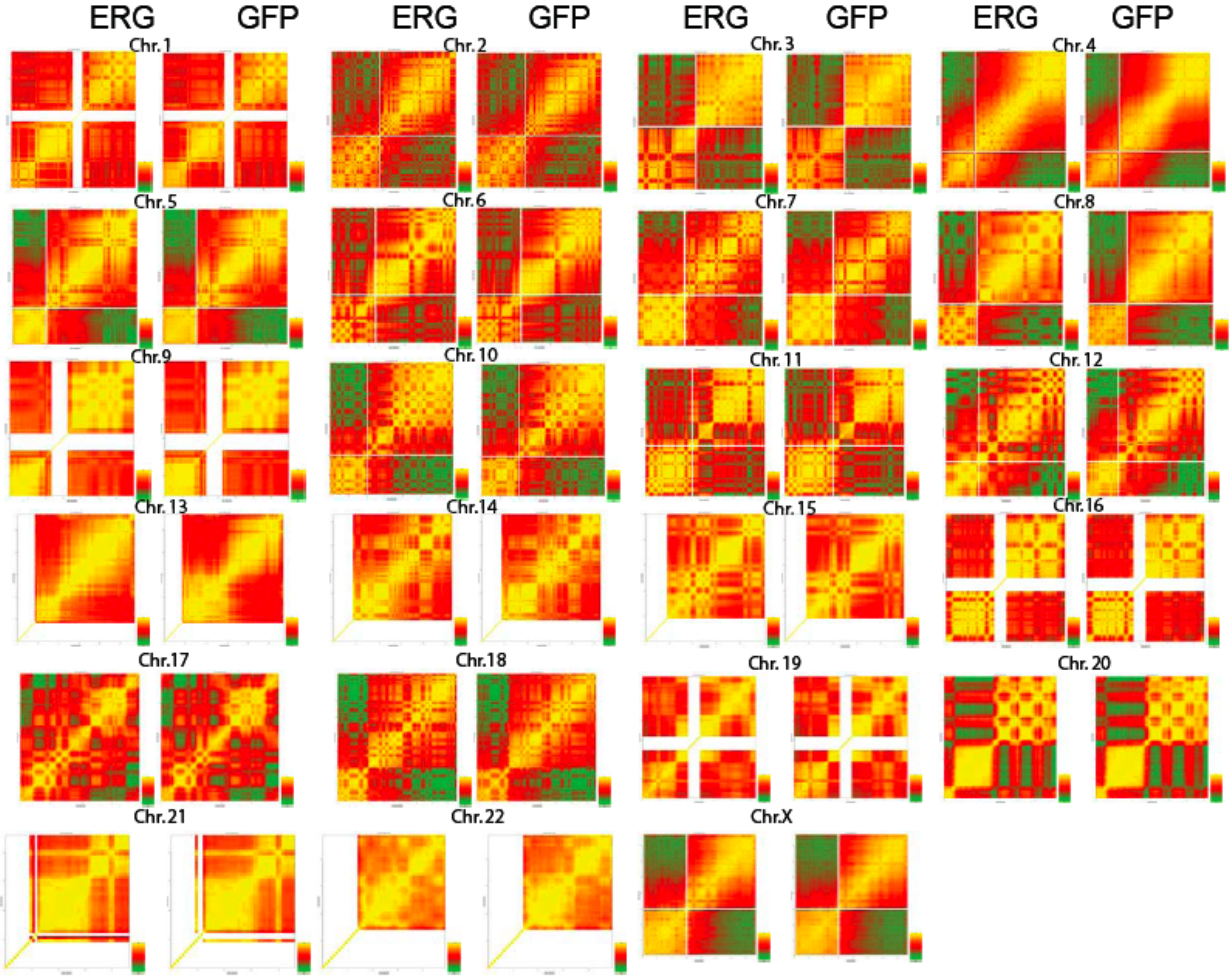
🟥 Expected

🟨 More interaction than expected

RWPE1-ERG

There are differences in interactions between ERG and GFP cells

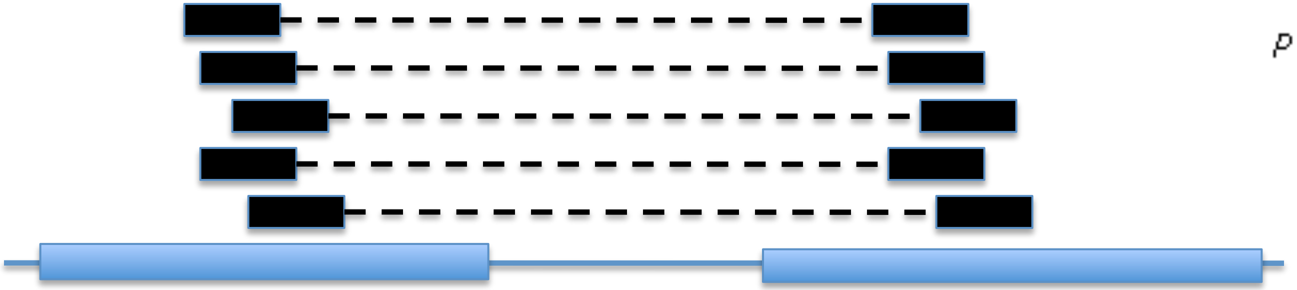



ERG
GFP
ERG
GFP
ERG
GFP
ERG
GFP
Chr.1
Chr.2
Chr.3
Chr.4
Chr.5
Chr.6
Chr.7
Chr.8
Chr.9
Chr.10
Chr.11
Chr.12
Chr.13
Chr.14
Chr.15
Chr.16
Chr.17
Chr.18
Chr.19
Chr.20
Chr.21
Chr.22
Chr.X

Fisher exact tests to quantify difference in interactions between GFP and ERG

GFP

k=5



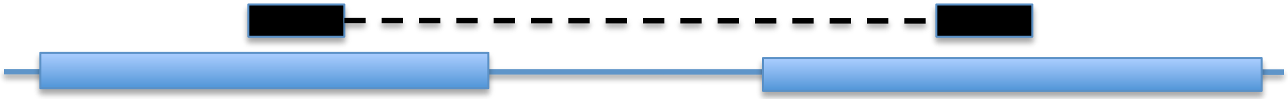
Gene 1

Gene 2

$$P(X=k)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

ERG

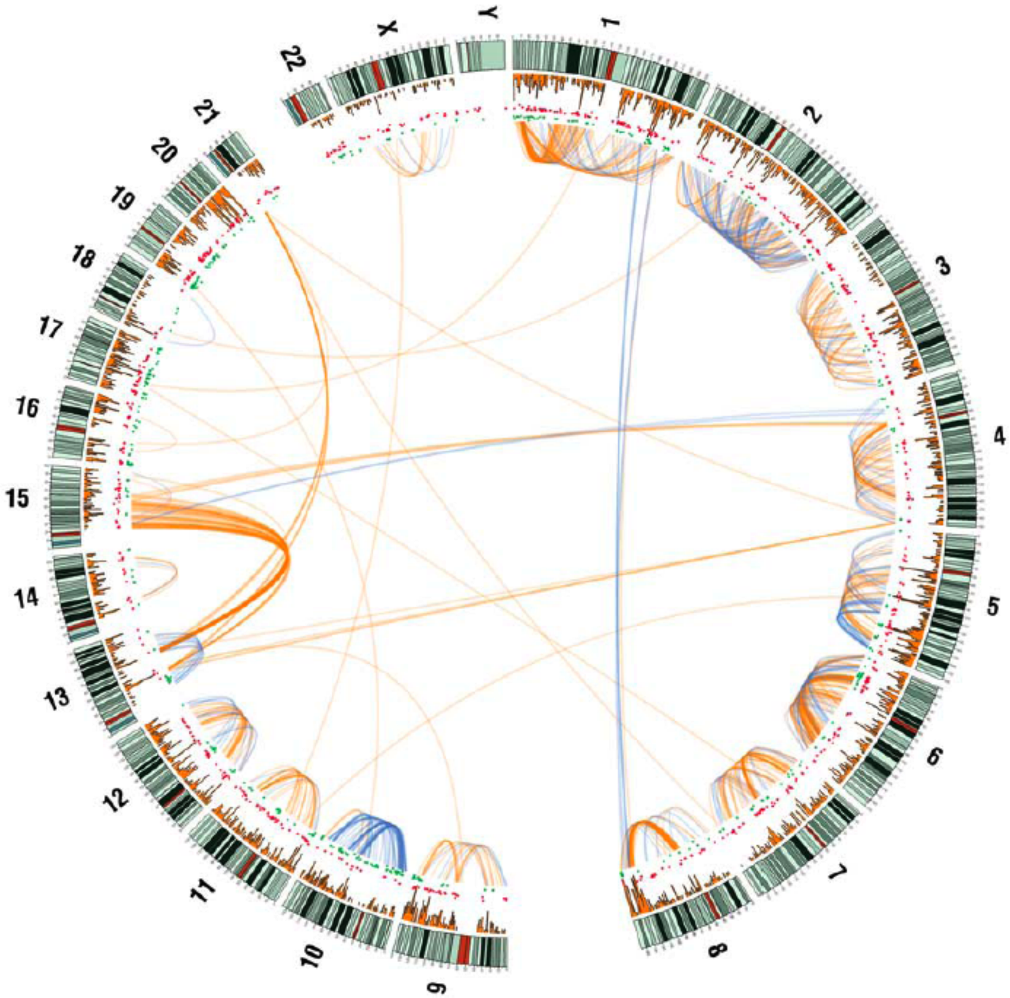
k=1



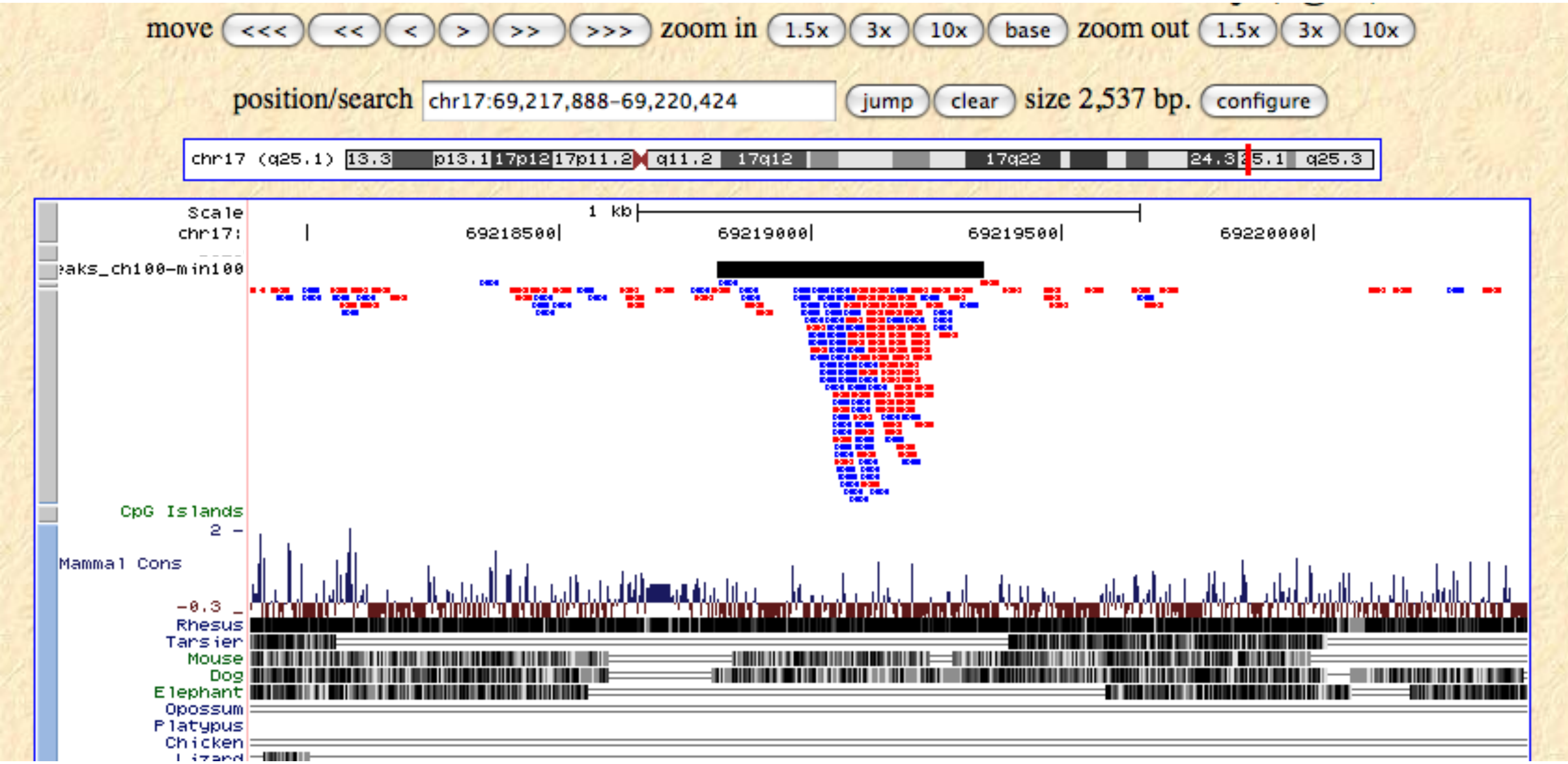
Gene 1

Gene 2

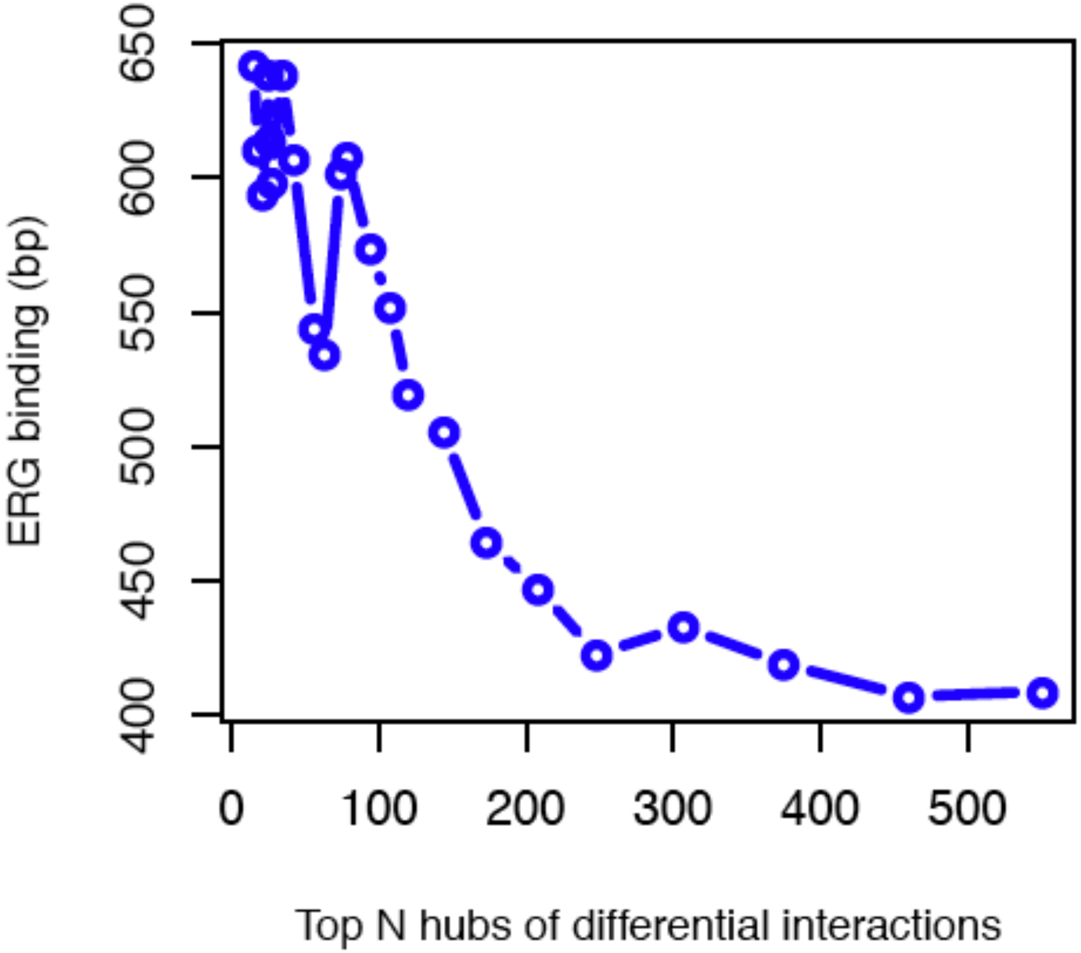
A graphical representation of all gains or losses of interaction in ERG cells vs GFP



ChIP-seq: ERG binds to >6,000 location in RWPE1-ERG cells


move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr17:69,217,888-69,220,424 jump clear size 2,537 bp. configure
chr17 (q25.1) 13.3 p13.1 17p12 17p11.2 q11.2 17q12 17q22 24.3 25.1 q25.3
Scale
1 kb
chr17:
69218500
69219000
69219500
69220000
eaks_ch100-min100
CpG Islands
2 -
Mammal Cons
-0.3 _
Rhesus
Tarsier
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard

Loci/genes that lose or gain many interactions are more likely to be bound by ERG

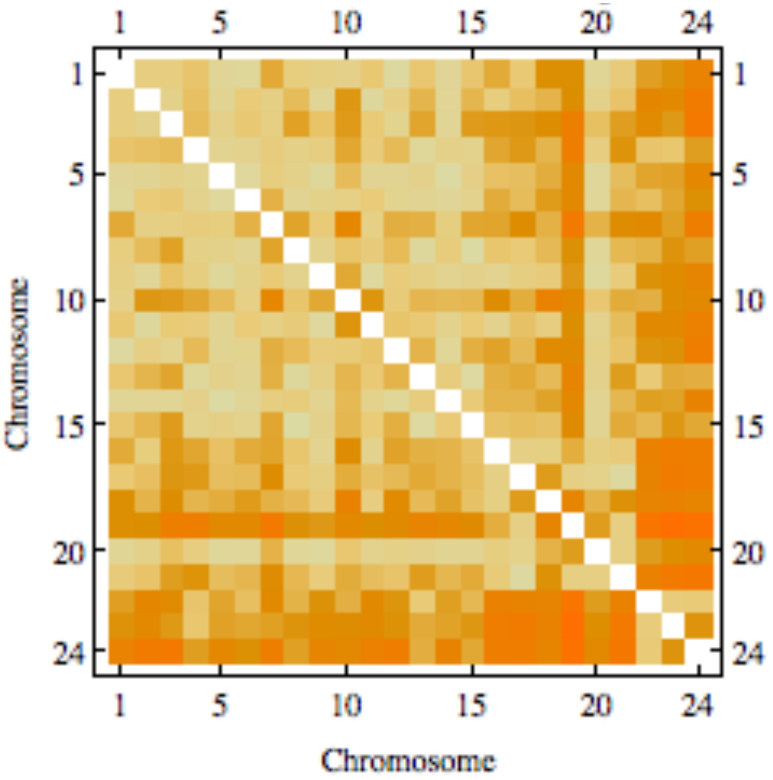


3D genome models

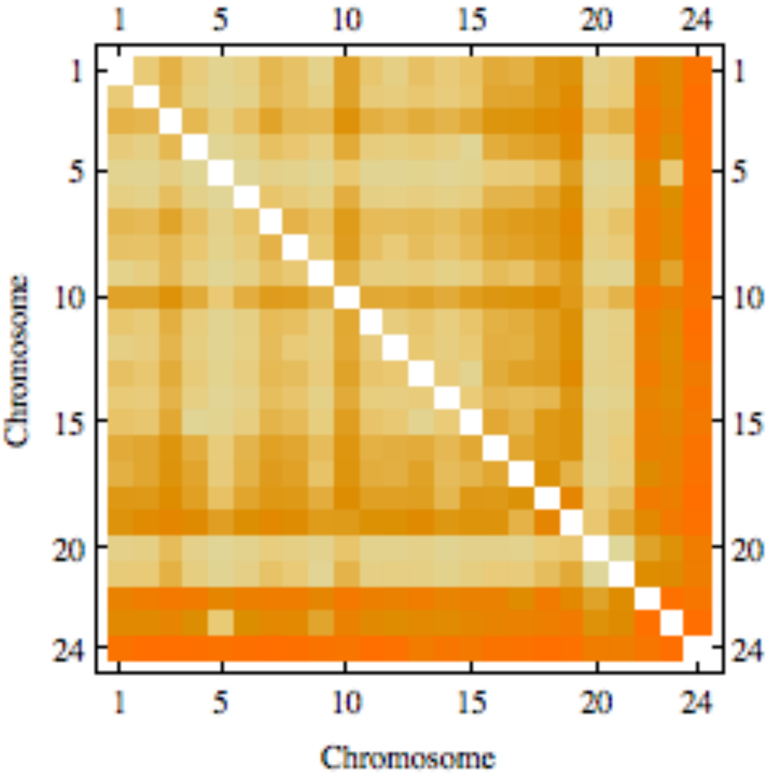
- Bin the genome into 1Mb blocks
- D = 1 / k where k = number of reads "connecting" two blocks
- Start with random 3D topologies, use gradient descent algorithm to make 3D models that best recapitulate the HiC distances

3D genome models

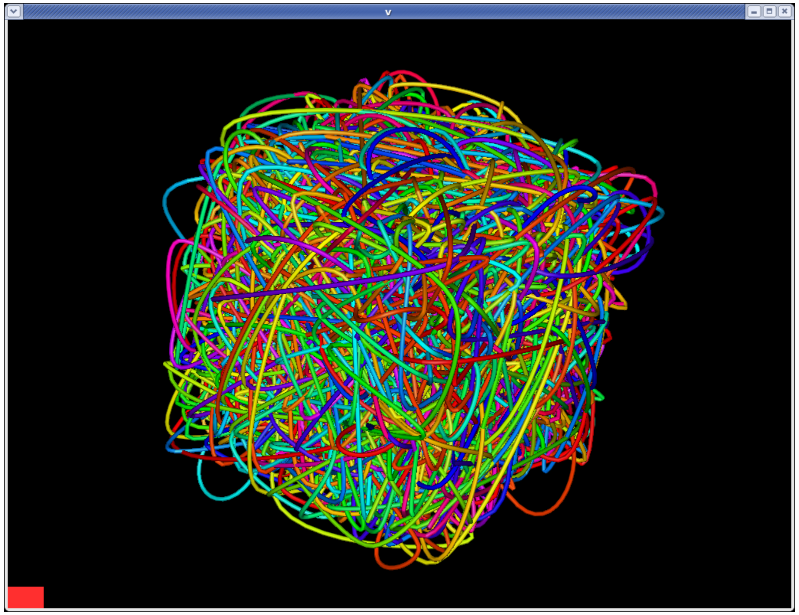
HiC distances



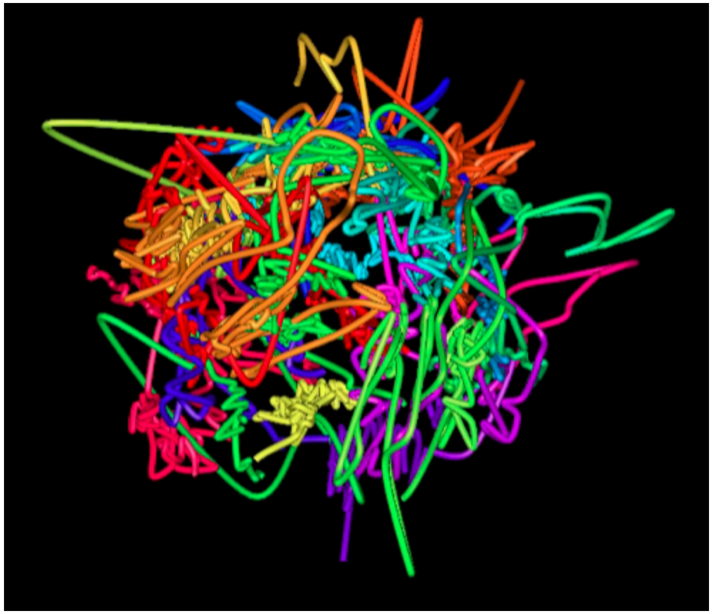
3D model

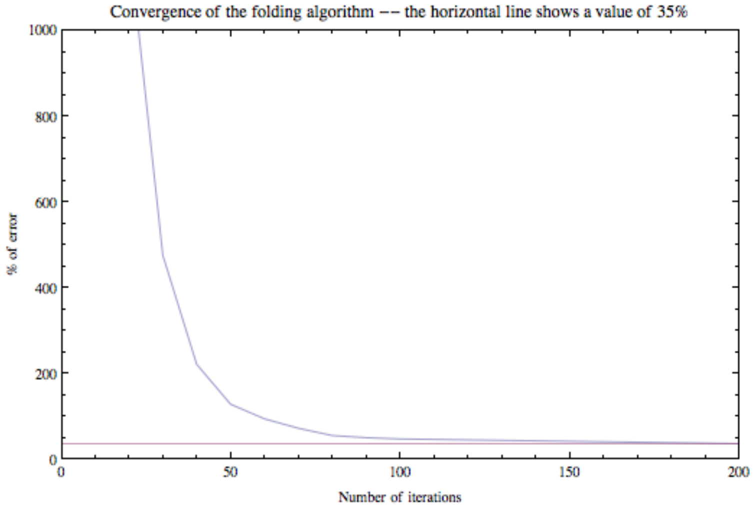


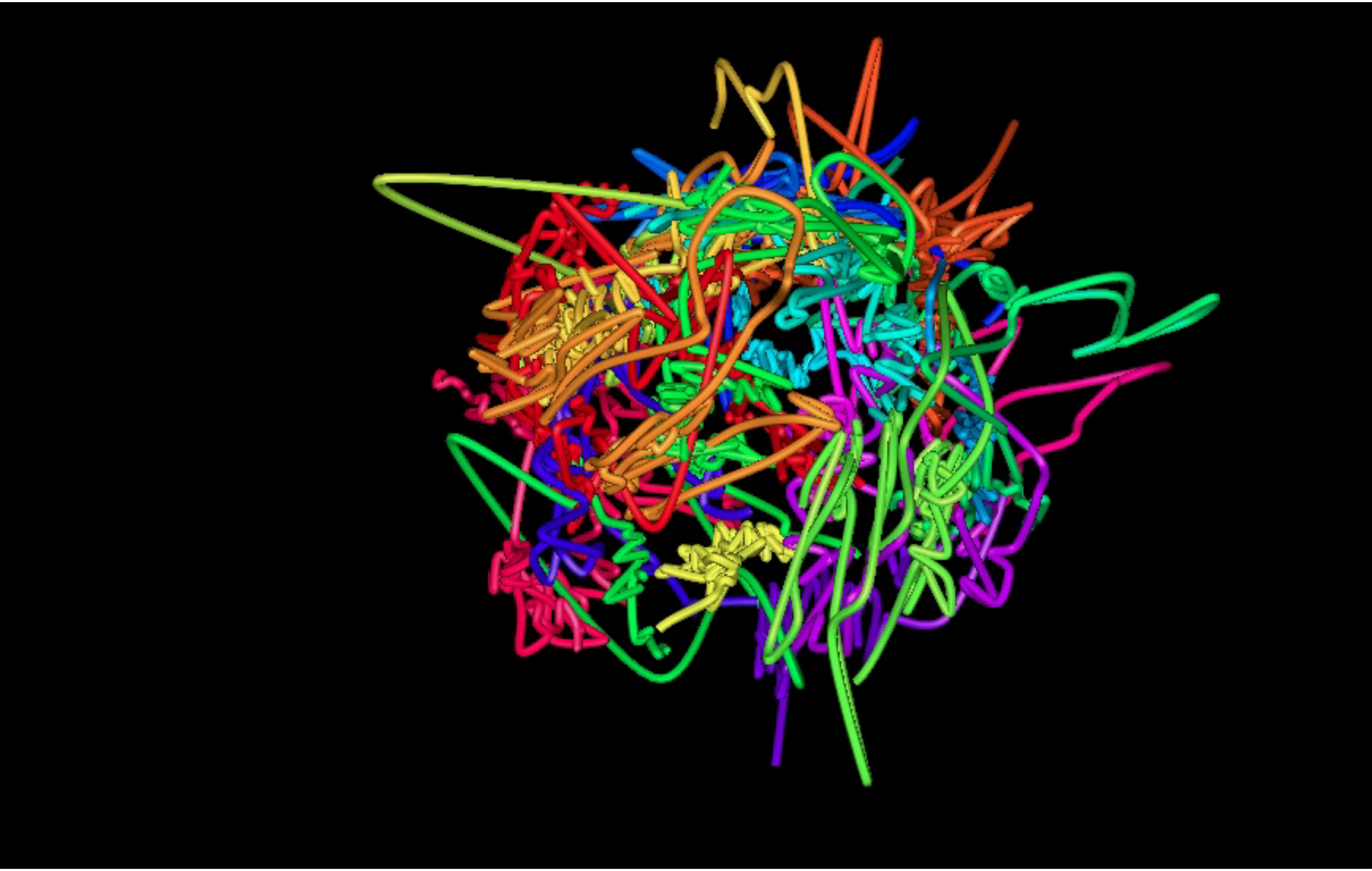
Initial random model



Final model




Convergence of the folding algorithm –– the horizontal line shows a value of 35%
% of error
1000
800
600
400
200
0
0
50
100
150
200
Number of iterations



How does ERG over-expression change chromosome conformation ?