A Genomic View of Epigenetic and Transcription Regulation in Cancer

Wei Li, PhD
Division of Biostatistics
Dan L. Duncan Cancer Center
Department of Molecular and Cellular Biology
Baylor College of Medicine
http://lilab.openwetware.org/

Outline

- Introduction
- ChIP-chip with genome tiling arrays
- ChIP-seq with next-gen sequencing
- Estrogen Receptor and FoxA1 regulation in breast and prostate cancers

Genetics

- The science of heredity and variation in living organisms
- DNA is the molecular basis for inheritance

- The Human Genome Project
  - Determine the sequences of the 3 billion chemical base pairs that make up human DNA
  - 1991 - 2003 *(1953 -)*

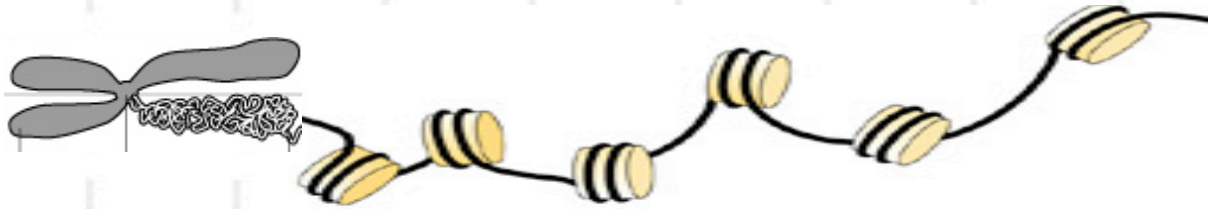
Epigenetics

- Heritable changes in gene function that occur without a change in the DNA sequence
- Stable, long-term alterations in the transcriptional potential of a cell that are not necessarily heritable

- NIH Epigenomics Roadmap
  - Develop comprehensive reference epigenome maps
  - 2008 - 2013

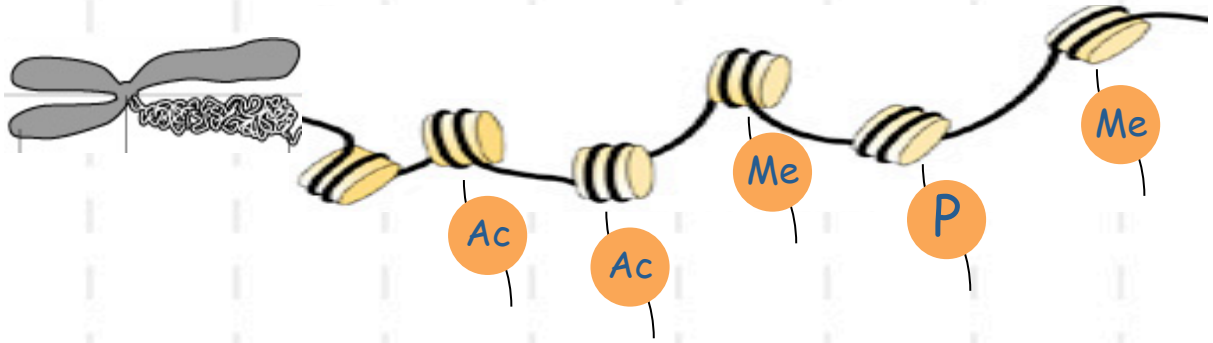
Nucleosome is the primary structural unit of chromatin

- DNA wrapped around histone proteins
- Nucleosomes control DNA accessibility



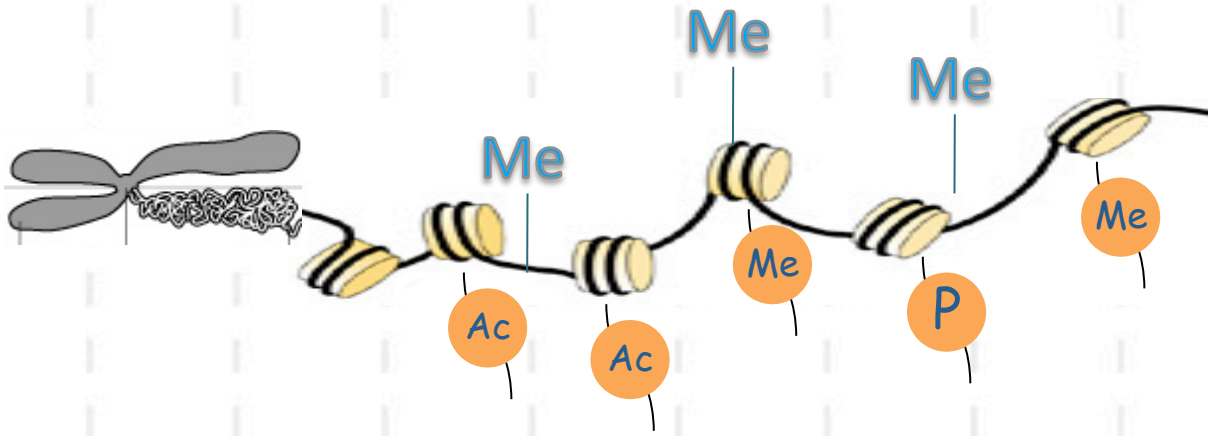
Histone Modifications can open/ close chromatin and influence gene expression

- Active histone marks: H3K4me1/2/3, H3K36me3
- Repressive histone marks: H3K9me3, H3K27me3
- ChIP-chip/seq: antibody against histone modification

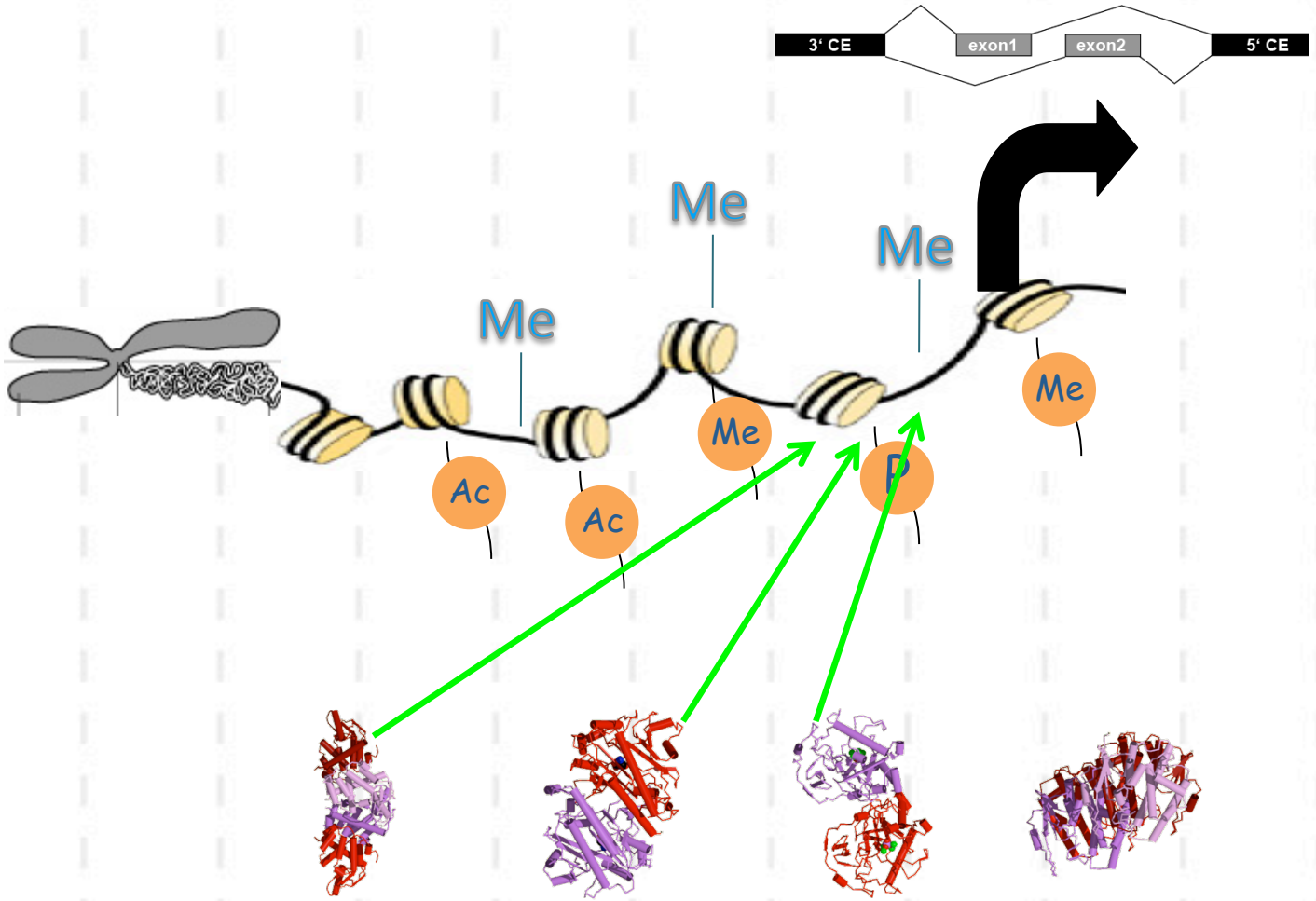


DNA methylation can close chromatin and repress gene expression

- longer term effect
- mDIP-chip/seq: antibody against $C^{M}G$



Epigenetic and Transcriptional Regulation



Alternative Splicing

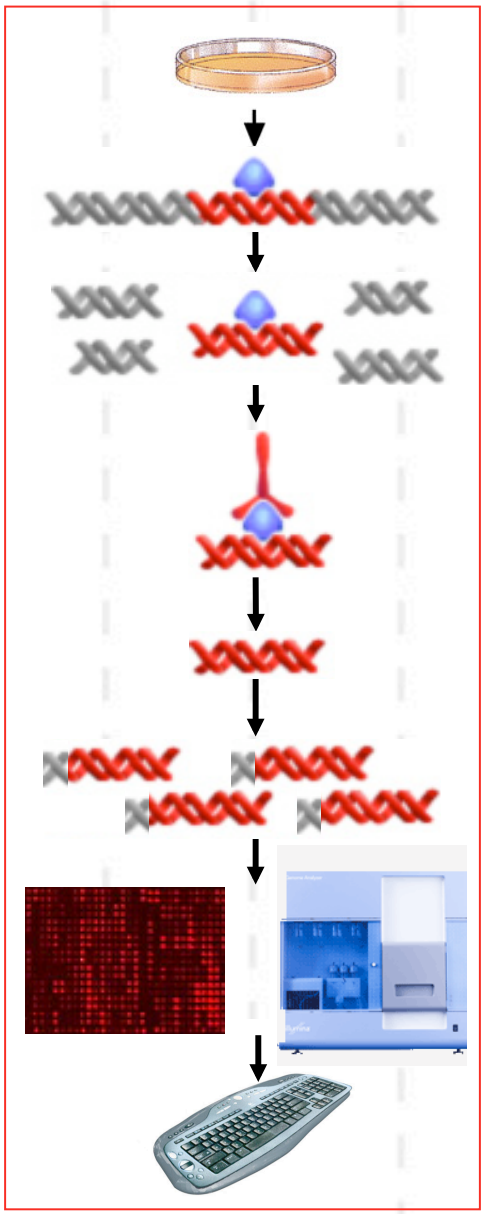
DNA Methylation

Nucleosome

Histone Modifications

Transcription Factors

ChIP on chip/seq



Cell treatment

Crosslinking

Sonication

Chromatin IP

Reverse croosslink
& DNA purification

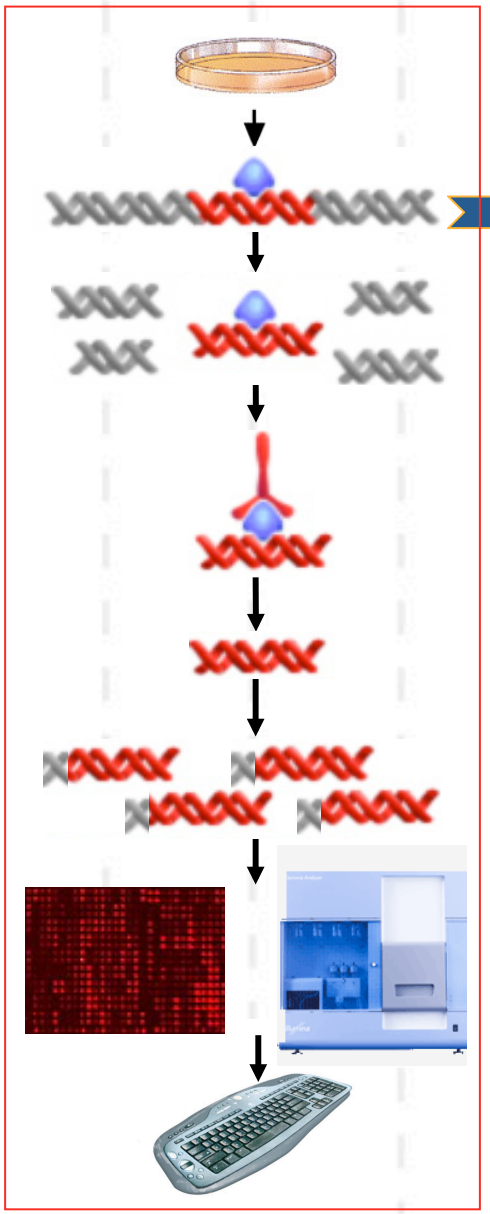
PCR amplification

Array hybridization

Array data analysis

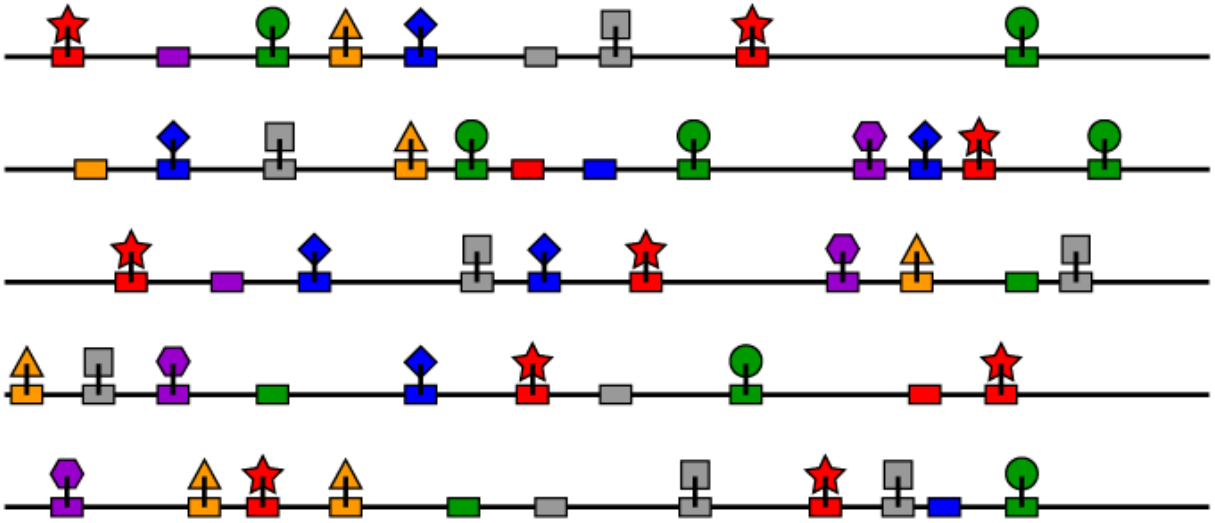
- ChIP-chip/seq
  - Chromatin ImmunoPrecipitation + microarray or sequencing
  - Genome-wide Location Analysis for TFs, histone modifications and DNA methylation (mDIP-chip/seq)

ChIP on chip/seq



TF-DNA Interaction

in vivo

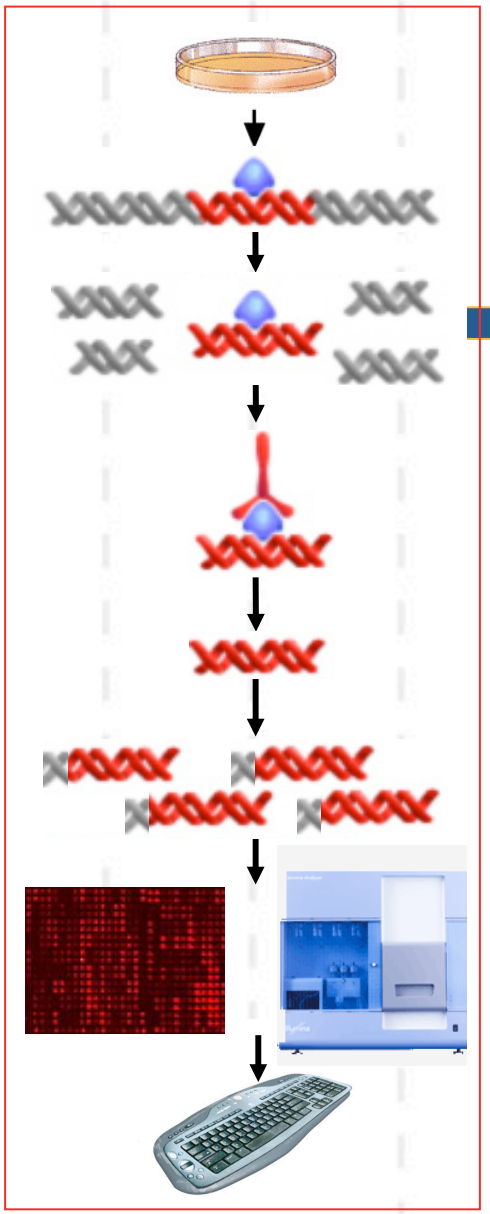


TF

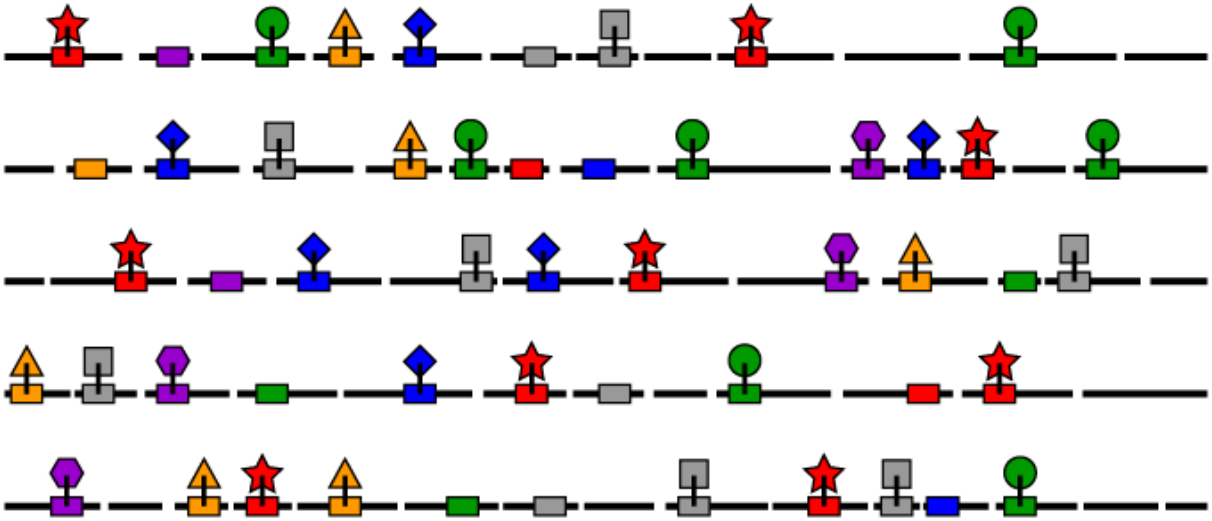
Binding Sites

some pics from Richard Bourgon

ChIP on chip/seq



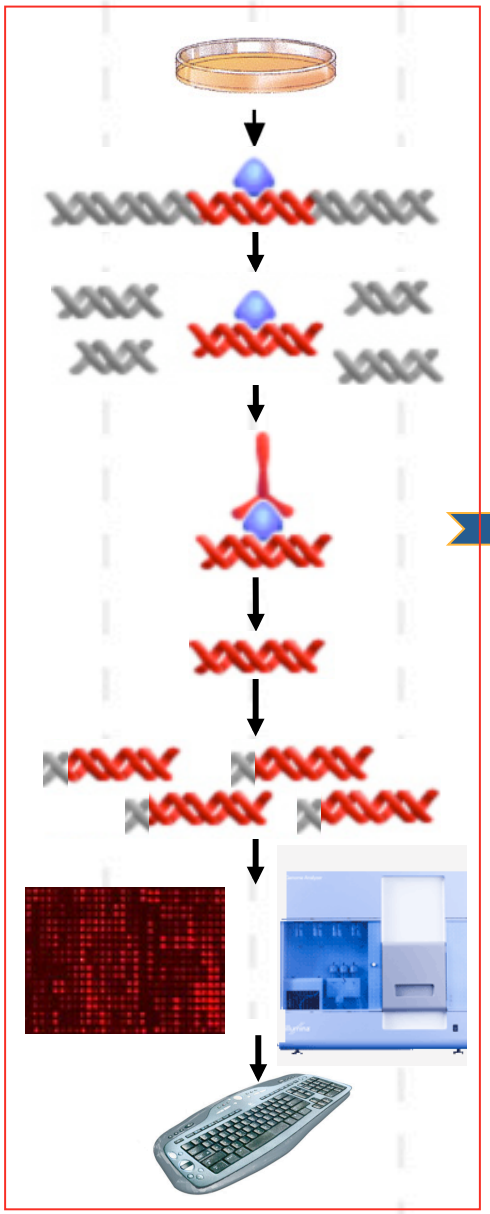
Sonication (~500bp)



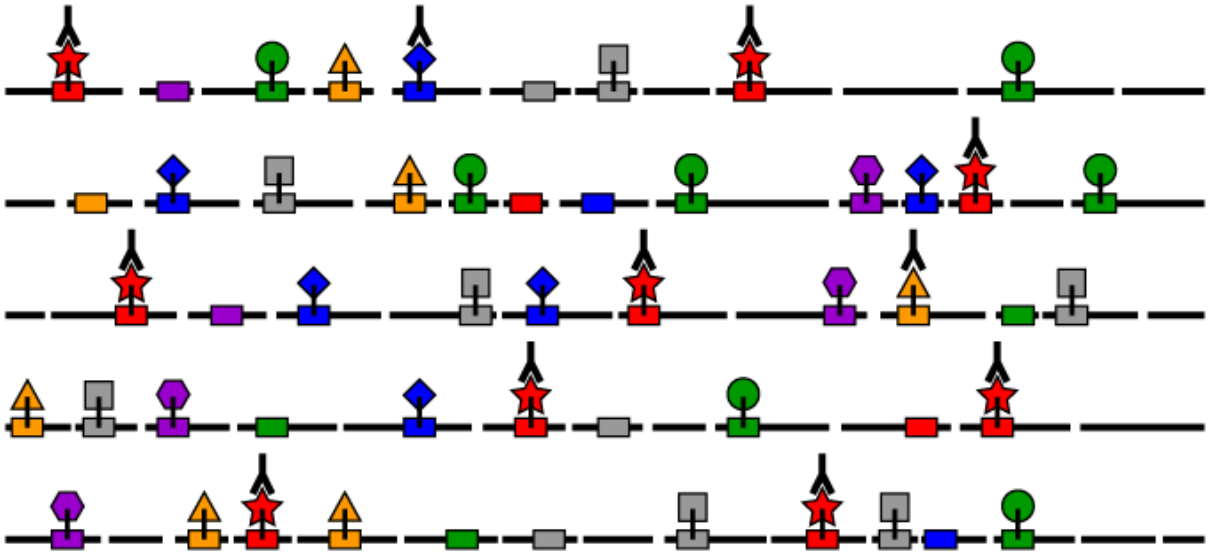
TF

Binding Sites

ChIP on chip/seq



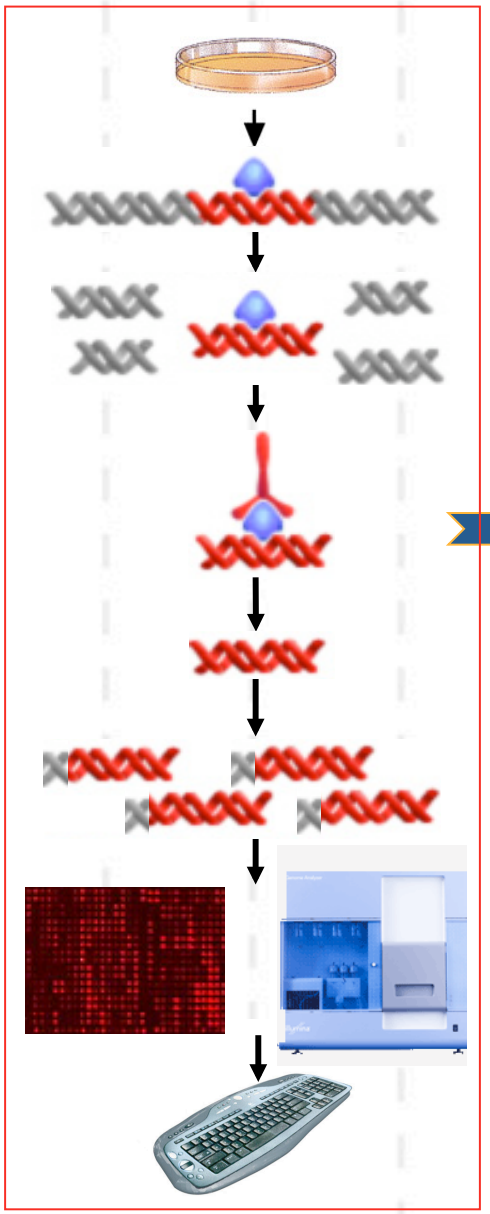
TF-specific Antibody



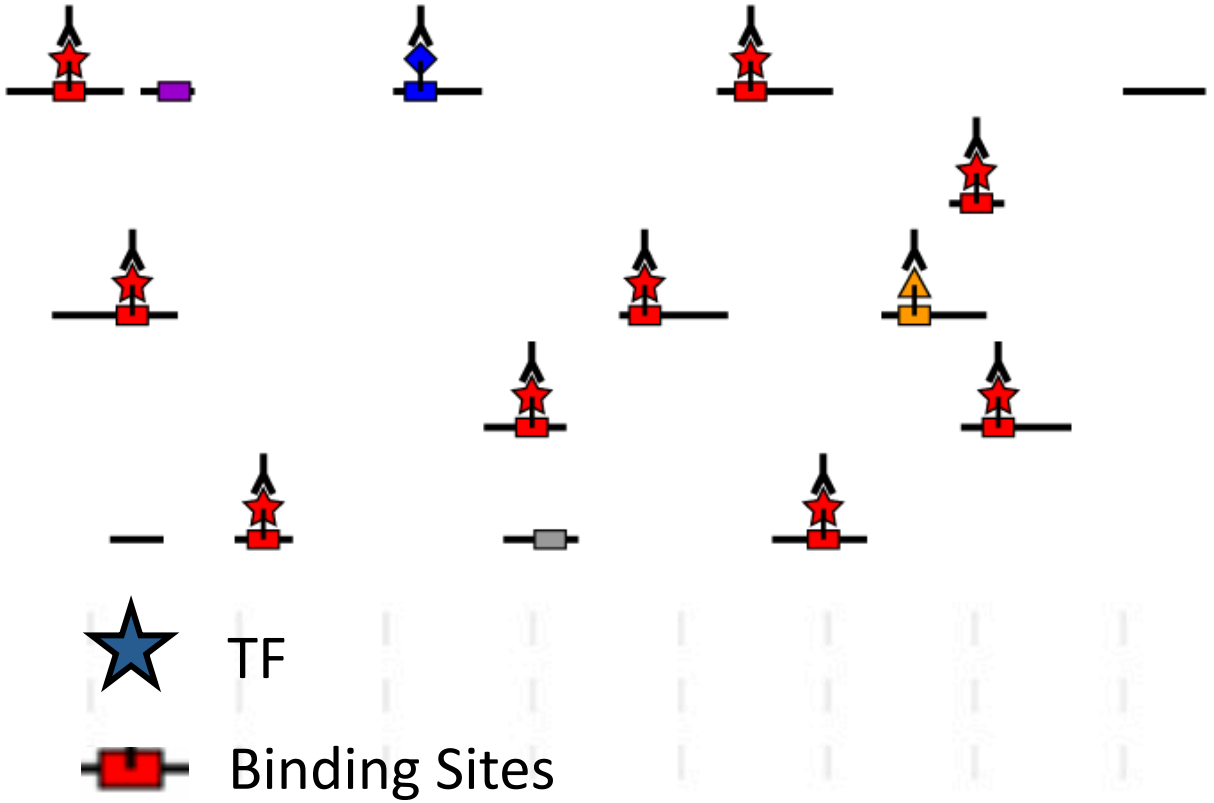
TF

Binding Sites

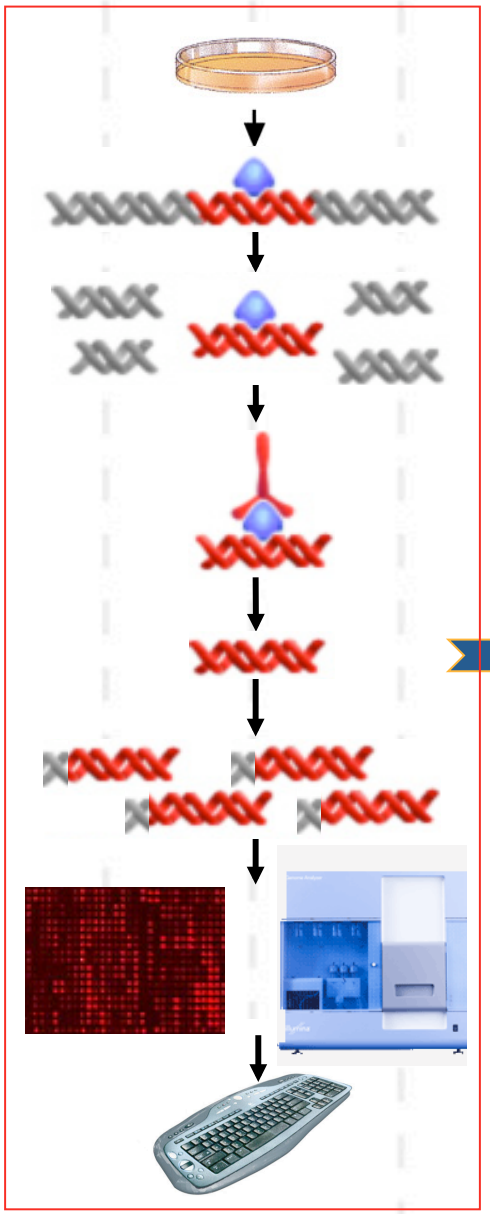
ChIP on chip/seq



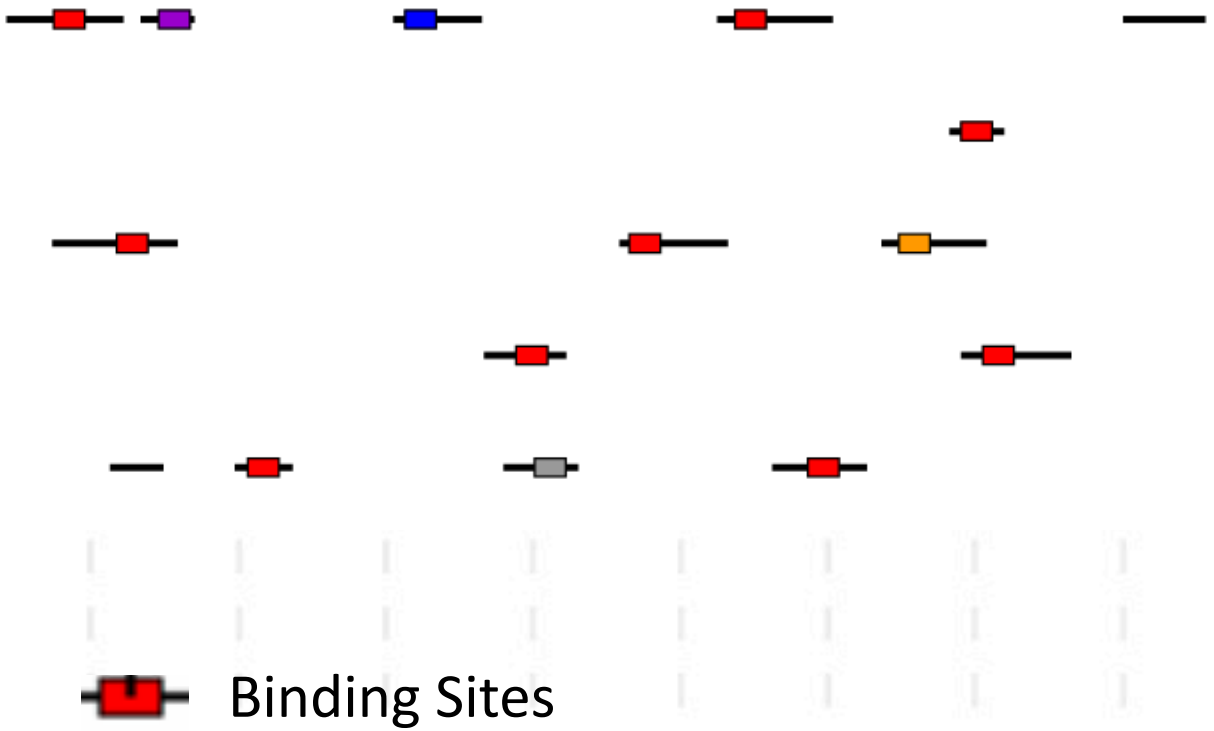
Immunoprecipitation



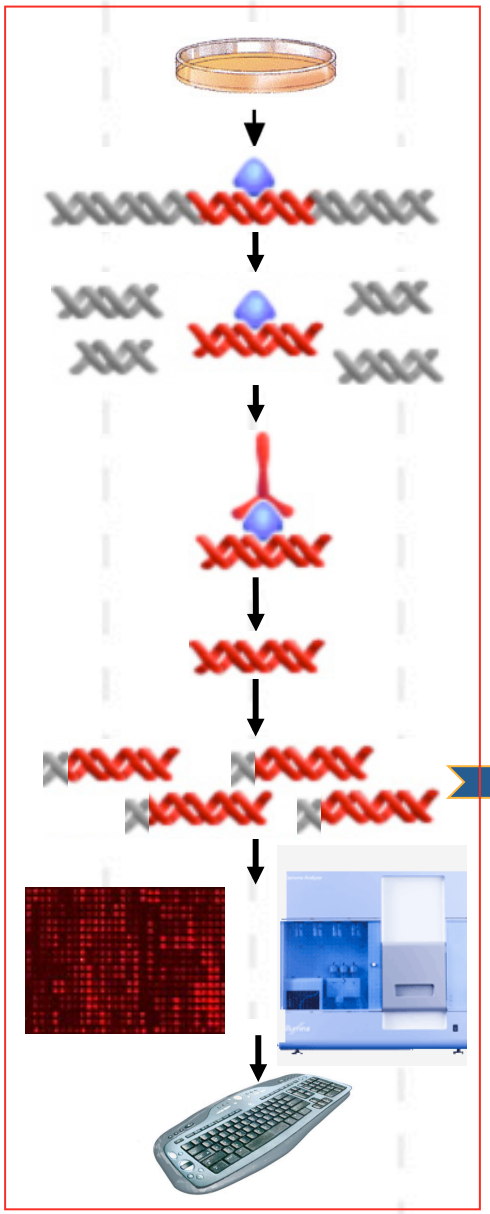
ChIP on chip/seq



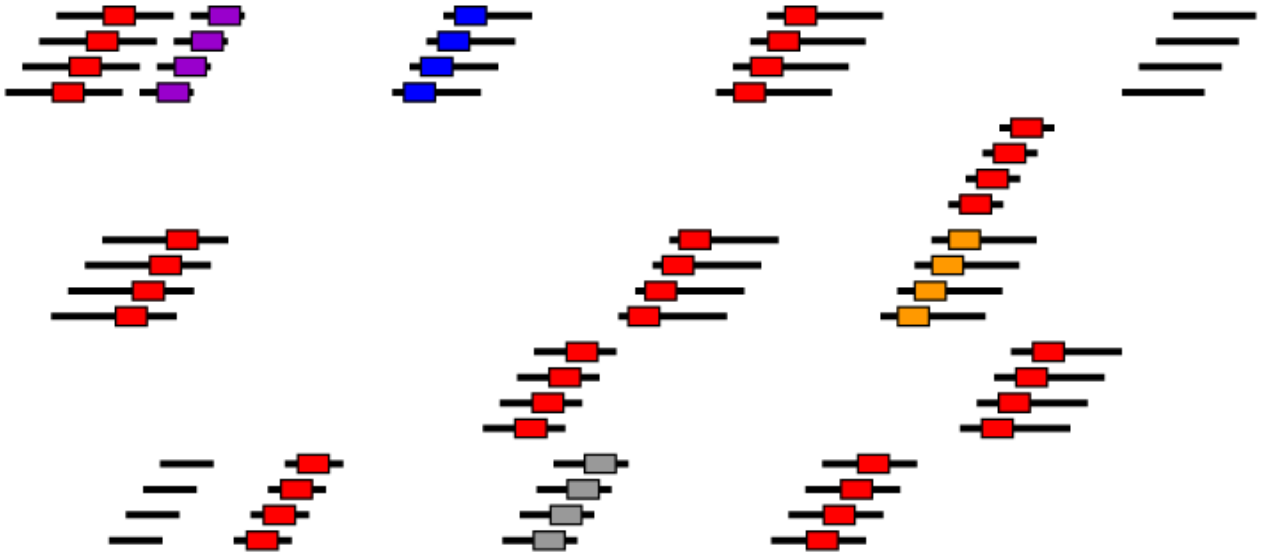
DNA Purification



ChIP on chip/seq

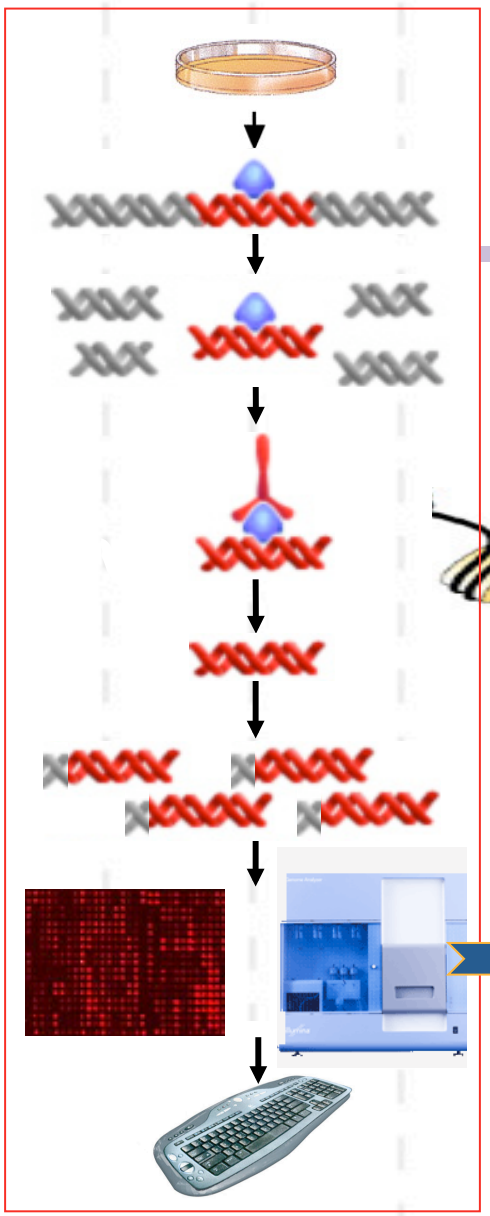


PCR Amplification

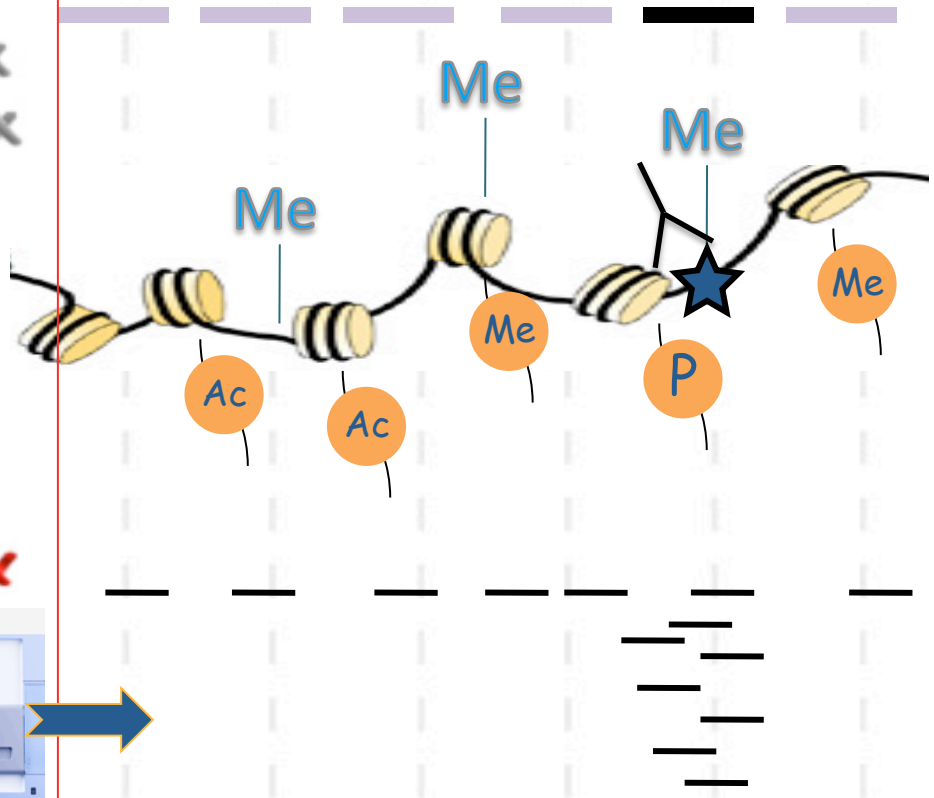


Binding Sites

ChIP on chip/seq



Unbiased Quantitative Read-out of the Entire Genome



Affymetrix

Tiling Array ($1.5k)

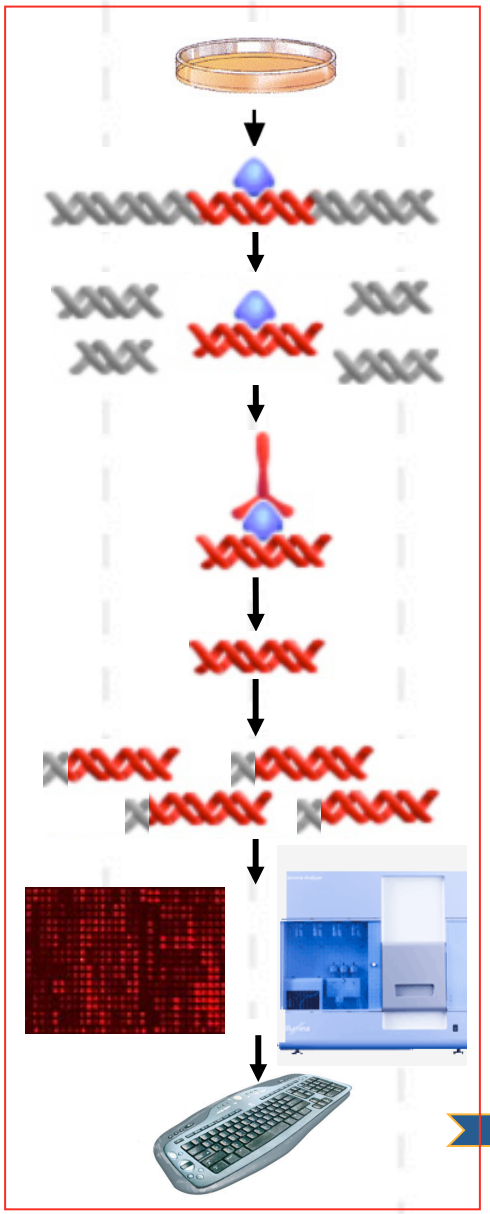
- 42 million probes (35 bp resolution)

Solexa/SOLiD

Sequencing ($2k)

- Gigabases of data from a single run

ChIP on chip/seq



Data Analysis

Probe / Read alignment

Peak Detection

Annotation

Visualization

Sequence Analysis

Hypothesis generation

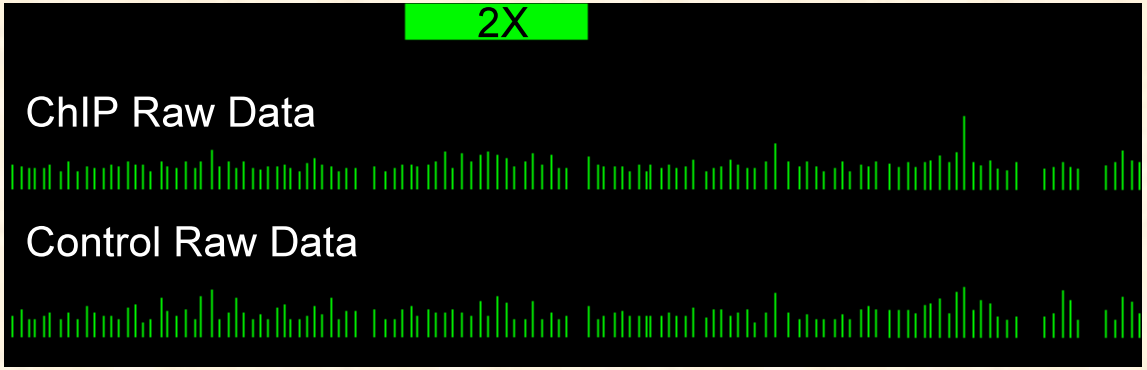
Outline

- Introduction
- ChIP-chip with genome tiling arrays
- ChIP-seq with next-gen sequencing
- Estrogen Receptor and FoxA1 regulation in breast and prostate cancers

Identify ChIP-enriched Region

- ChIP: endogenous or activated TF
- Controls: sonicated genomic Input DNA
- Often 3 ChIP, 3 Ctrl replicates needed

Why is it so Difficult to Analyze the Tiling Array Data?

Challenge	Expression Array (U133P2)	Tiling Array	
Massive data	0.5 M Probes	42 M Probes	*~80 X more*
Signal/noise ratio (present call)	~30%	~0.1-1%	*~60 X smaller*
Background	Total RNA (2% WG)	Whole Genome	*~50 X bigger*
Noisy probe values			

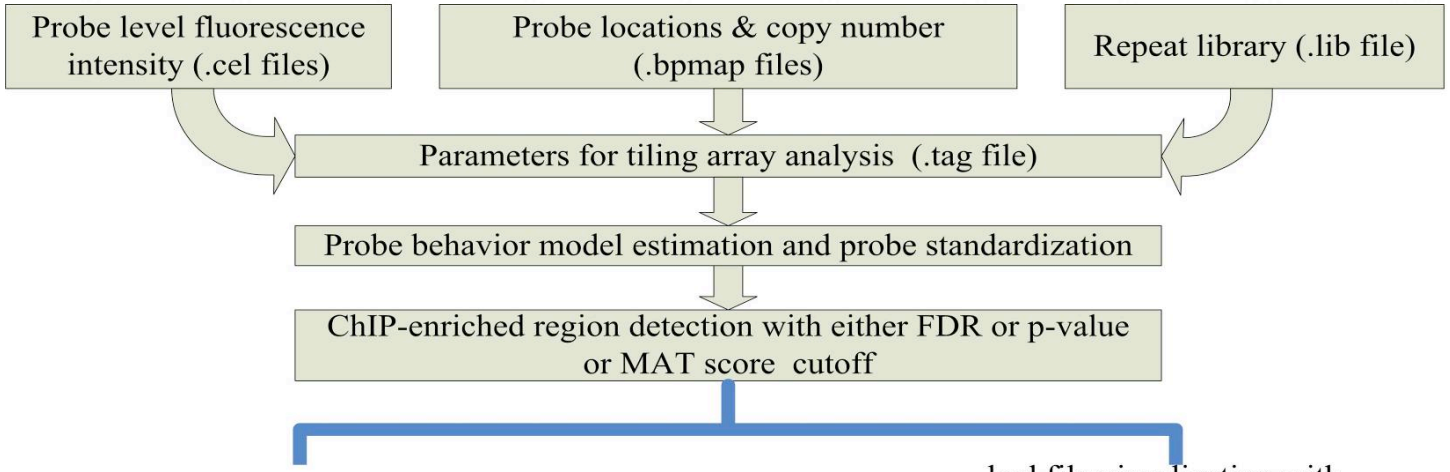
xMAN: eXtreme MApping of oligoNucleotides

- Mapping ~42 M probes to the human genome in less than 6 CPU hours
  - BLAST needs 20 CPU years
  - BLAT needs 55 CPU days (low sensitivity)
- xMAN probe mapping
  - ~13 k probes are not in the new genome version
  - Probe TCCCAGCACTTTGGGAGGCTGAGGC (chr1 91712) maps to 50,660 times in the genome
  - ~1.3 M probes have multiple genomic hits (internal spike-in)
- xMAN probe mapping and filtering significantly improves peak detection

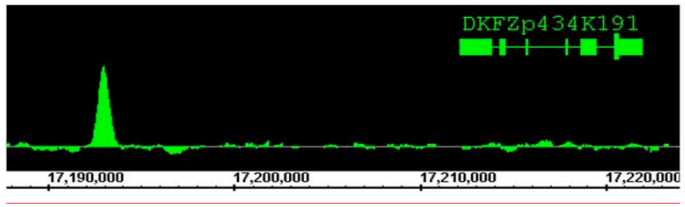
BMC Genomics (2008)

MAT: Model-based Analysis Of Tiling-array

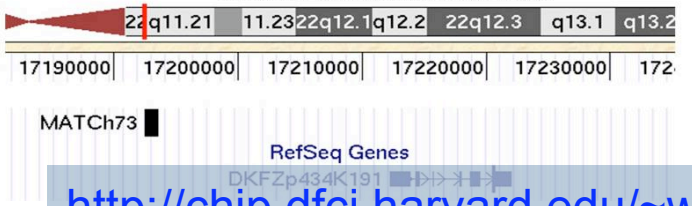
- Can work with single ChIP without replicates and controls
- Can Find individual failed sample
- More sensitive, specific and quantitative



.bar file visualization with IGB



.bed file visualization with UCSC Genome Browser



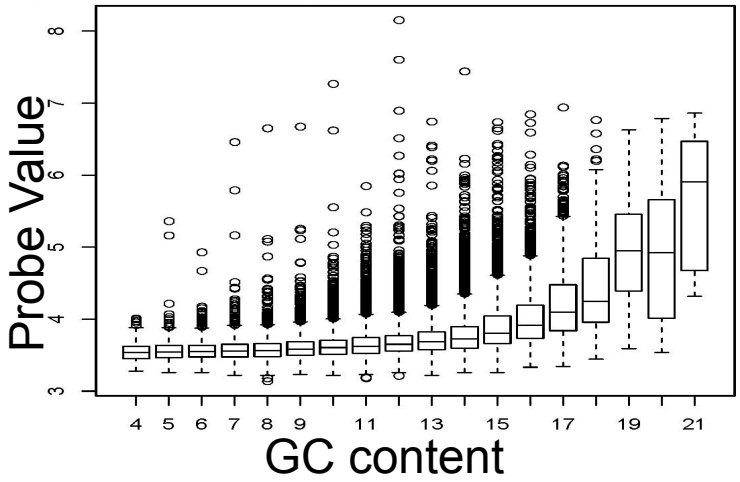
http://chip.dfci.harvard.edu/~wli/MAT/

Johnson et al, PNAS 2006

MAT: Intuitions

- Most of the probes on ChIP-chip measures genomic background
- Each Affy array has enough probes to estimate the effect of probe sequence on probe intensity
- Estimate probe behavior by borrowing similar probes on the same array

Probe sequence plays a big role in signal value



MAT: Probe Behavior Model

$$Log(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k=A,C,G} \beta_{jk} I_{ijk} + \sum_{l=A,C,G,T} \gamma_l n_{il}^2 + \delta Log(c_i) + \varepsilon_i$$

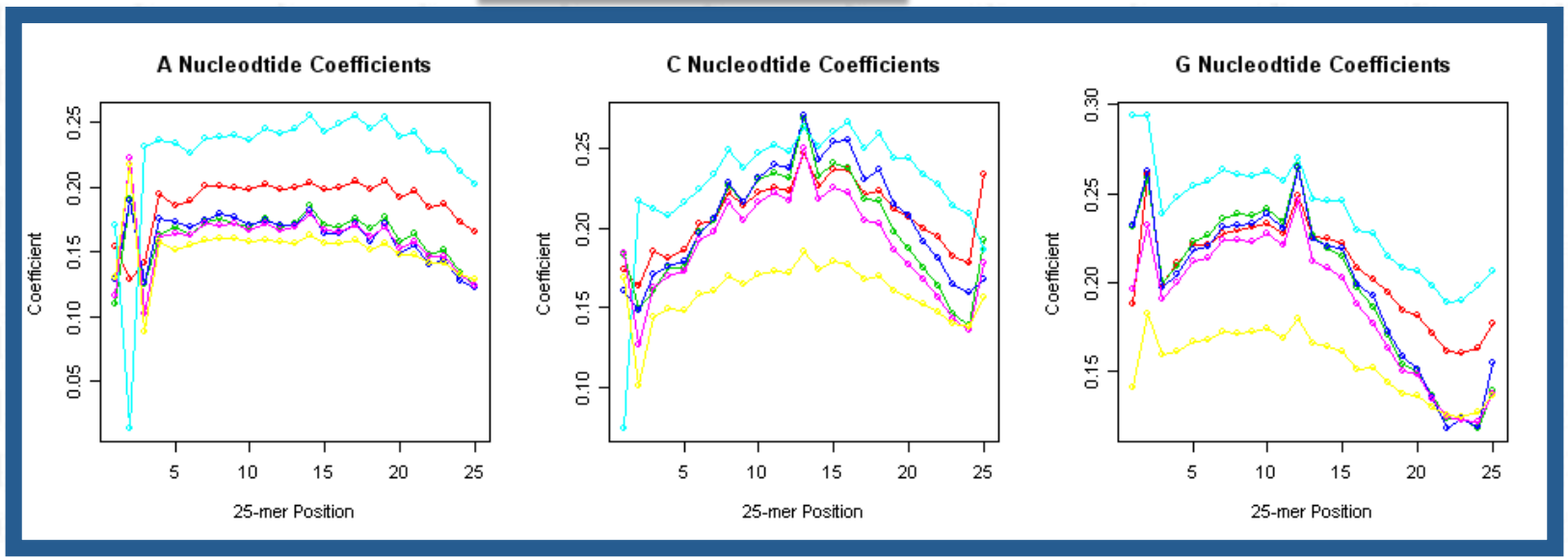
Probe Signal

Baseline on Number of Ts

Position Specific A,C,G Effect

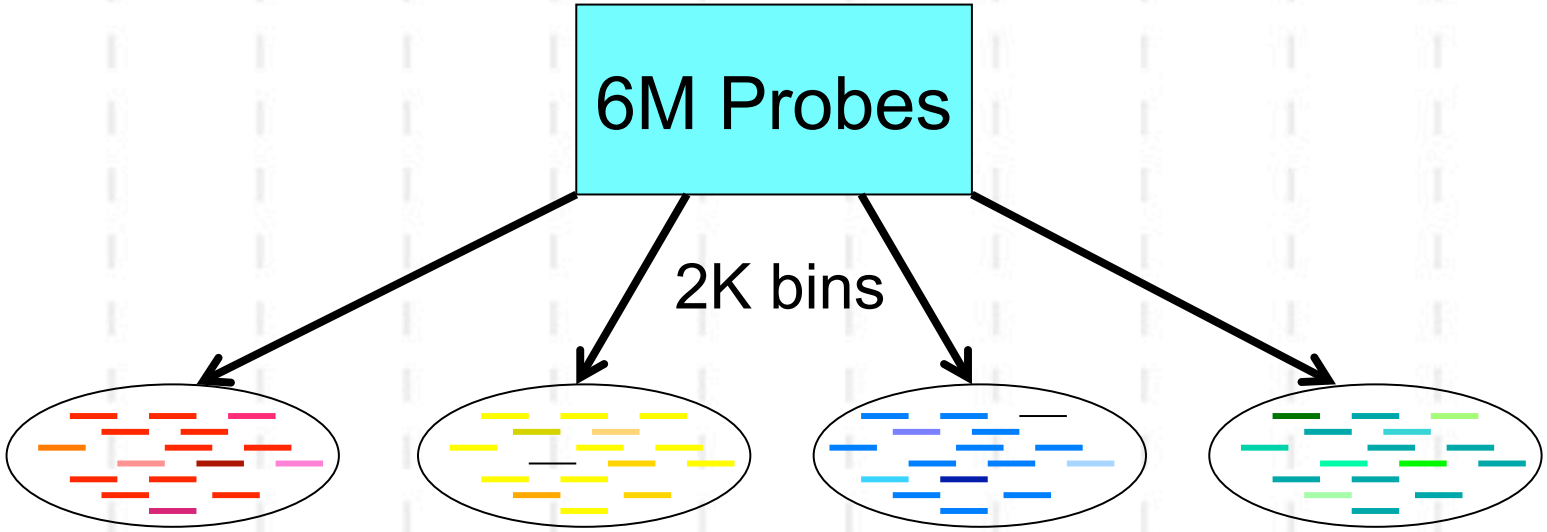
A,C,G,T Count Square

25mer Copy Number Along the Genome



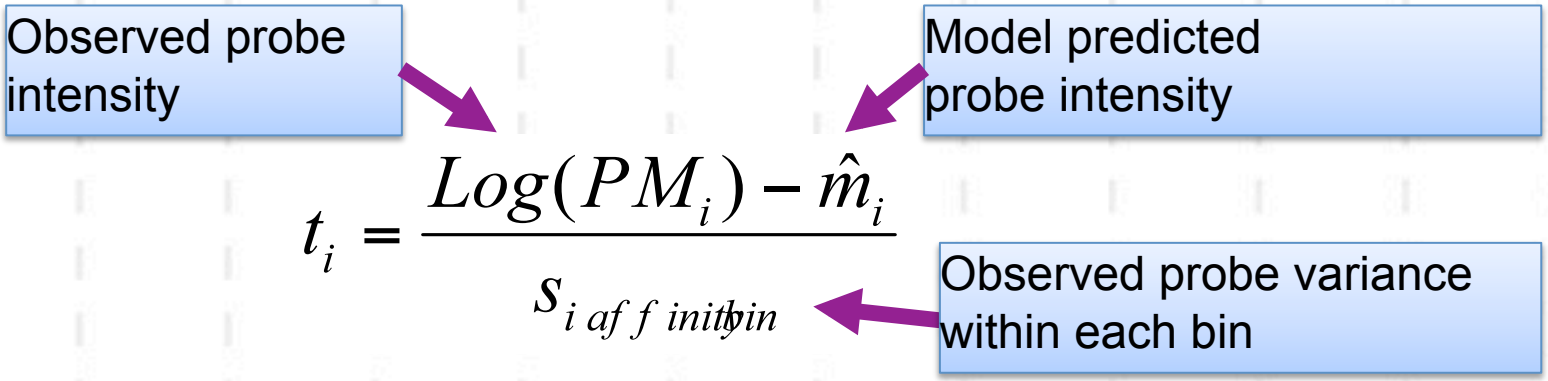
MAT: Probe Standardization

- Fit the probe model array by array
- Divide 6M array probes to bins (3k probes/bin)



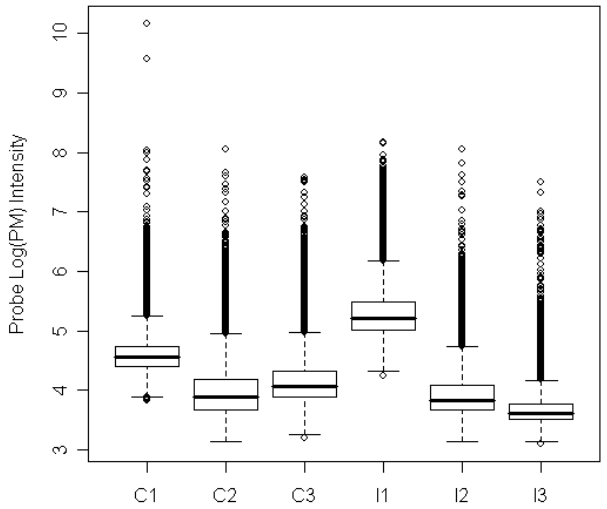
MAT: Probe Standardization

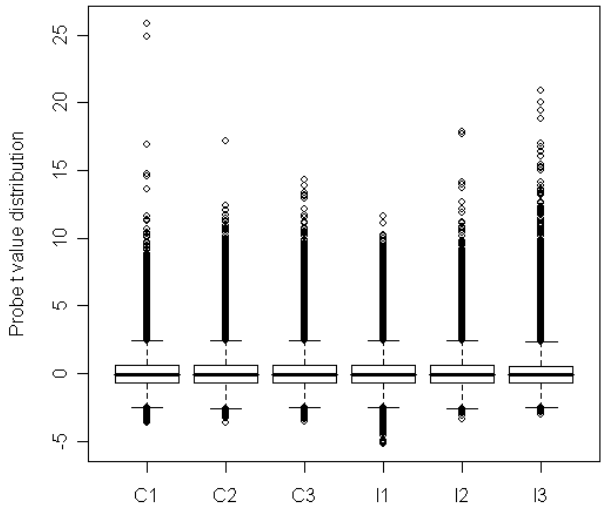
- Standardize probe behavior within each bin on a single array (Background-subtraction + normalization)



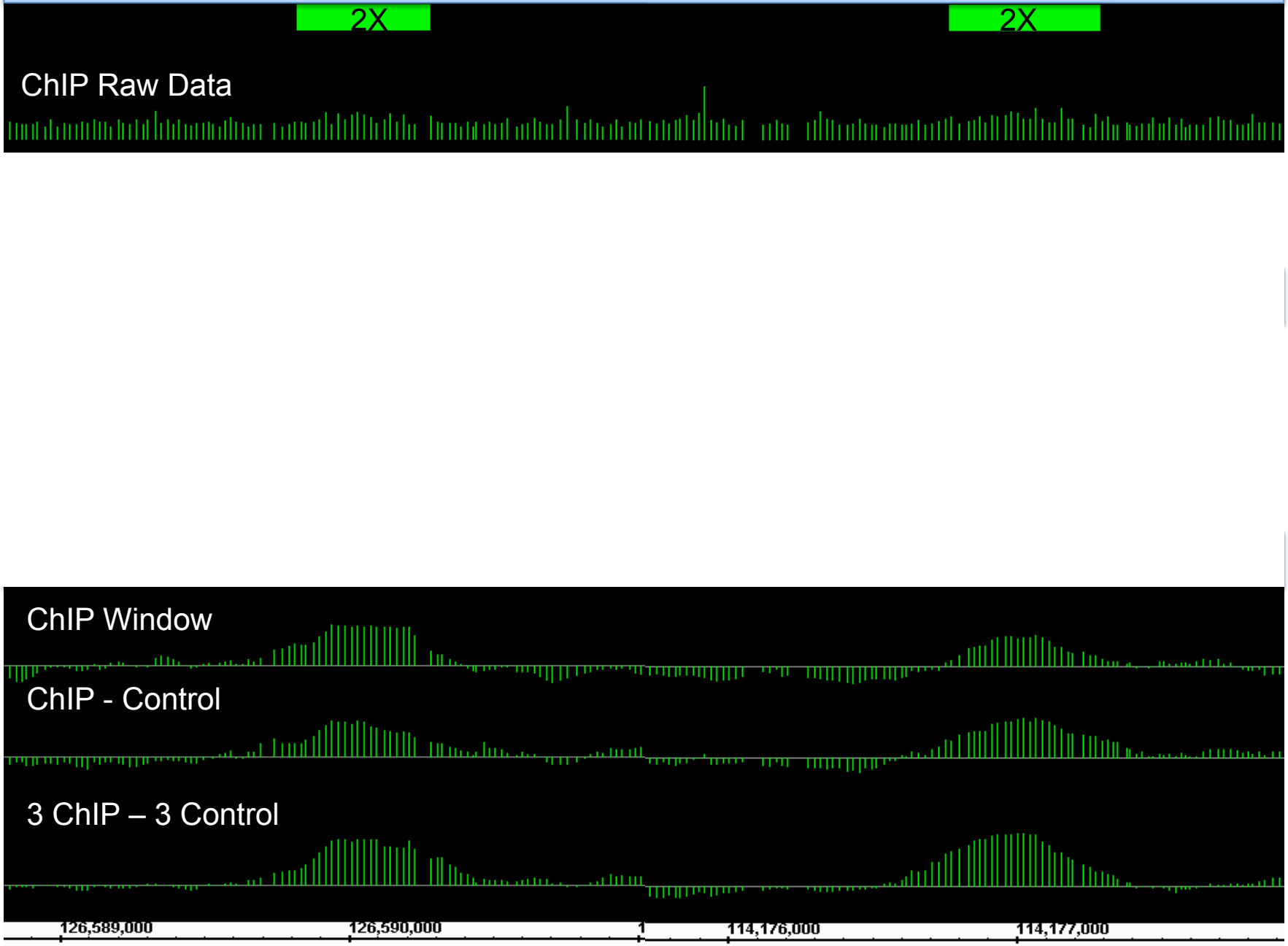
MAT: Probe Standardization

- After standardization: different probes across different samples are comparable

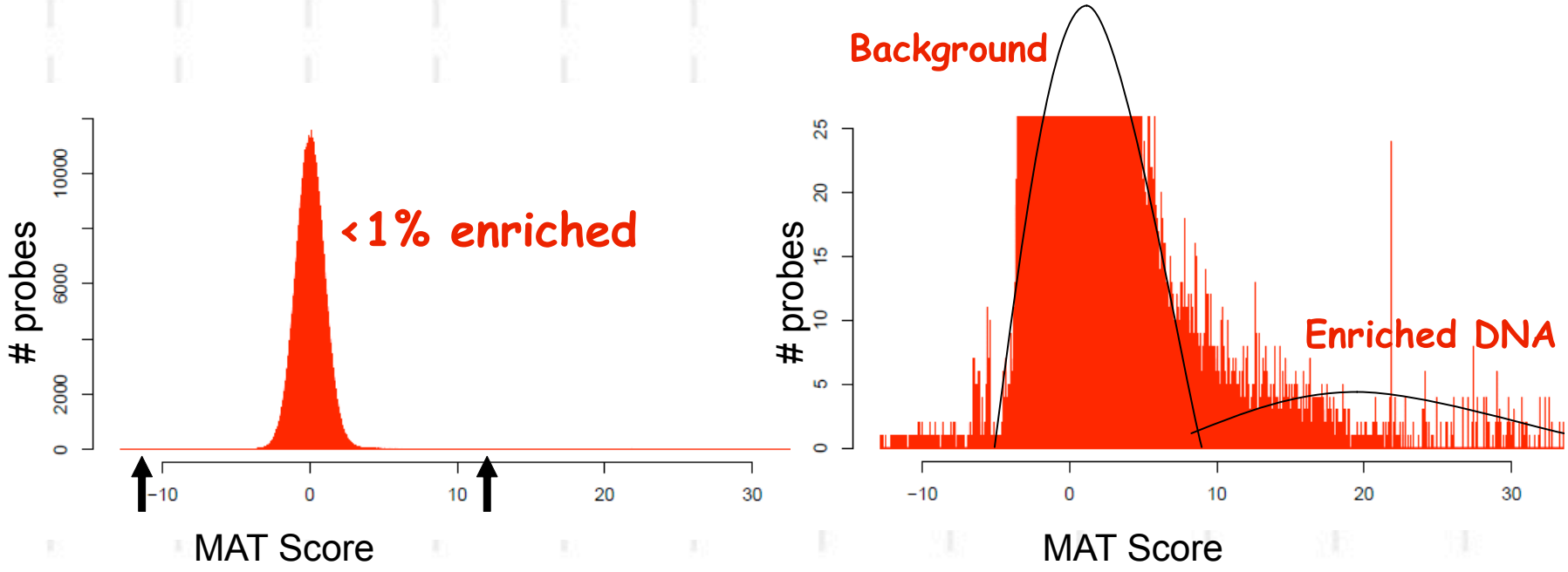




Raw probe values at two spike-in regions with concentration 2X

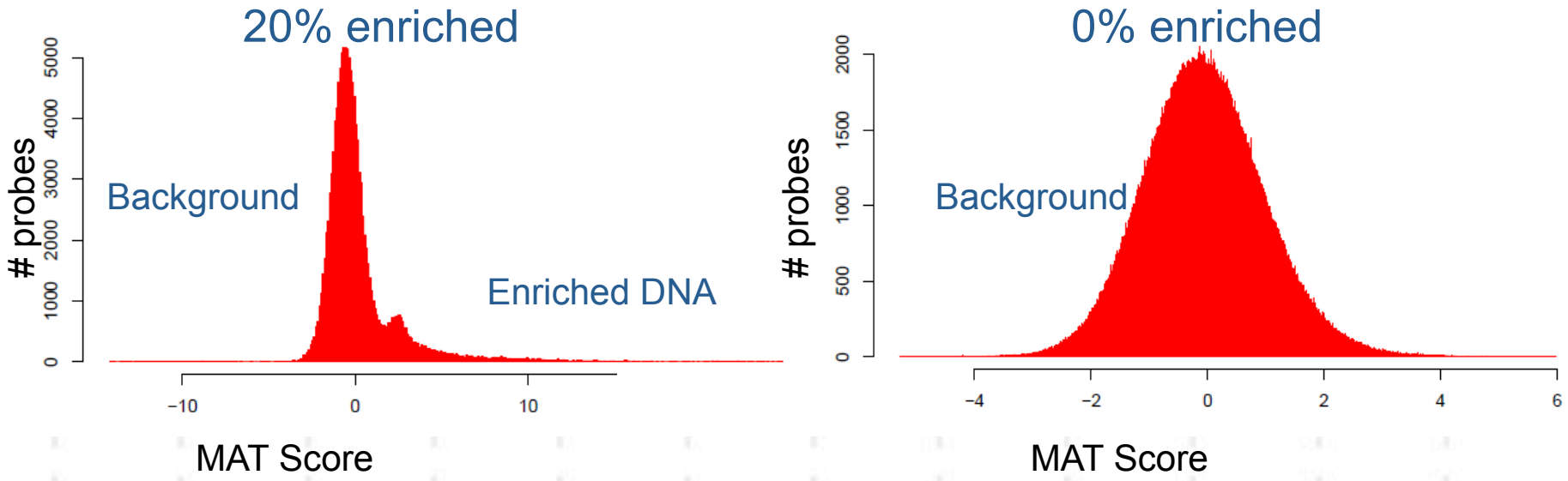


MAT: ChIP-region Detection

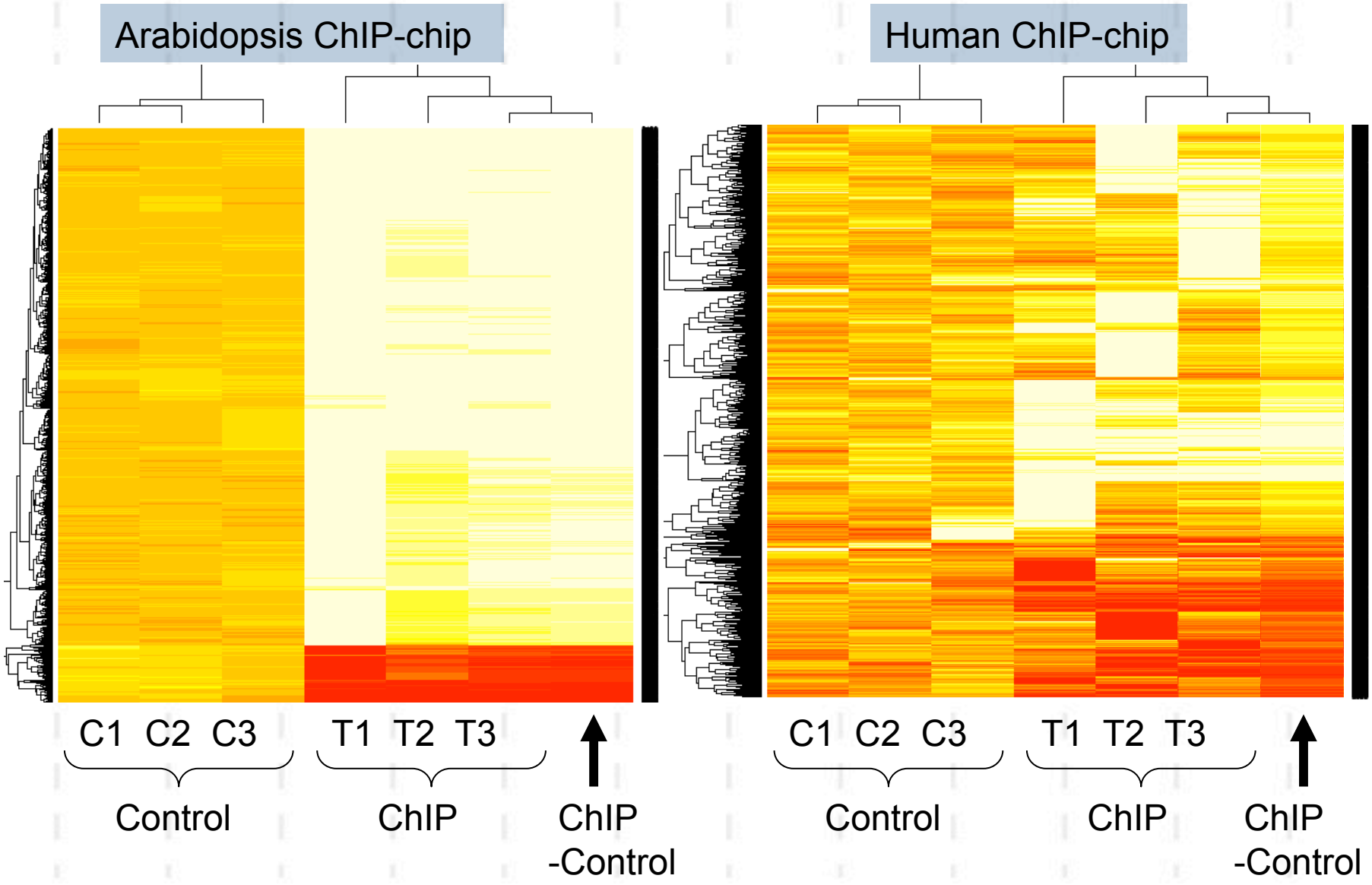


Regional FDR estimation
#negative_peaks / #positive_peaks

MAT: ChIP-region Detection

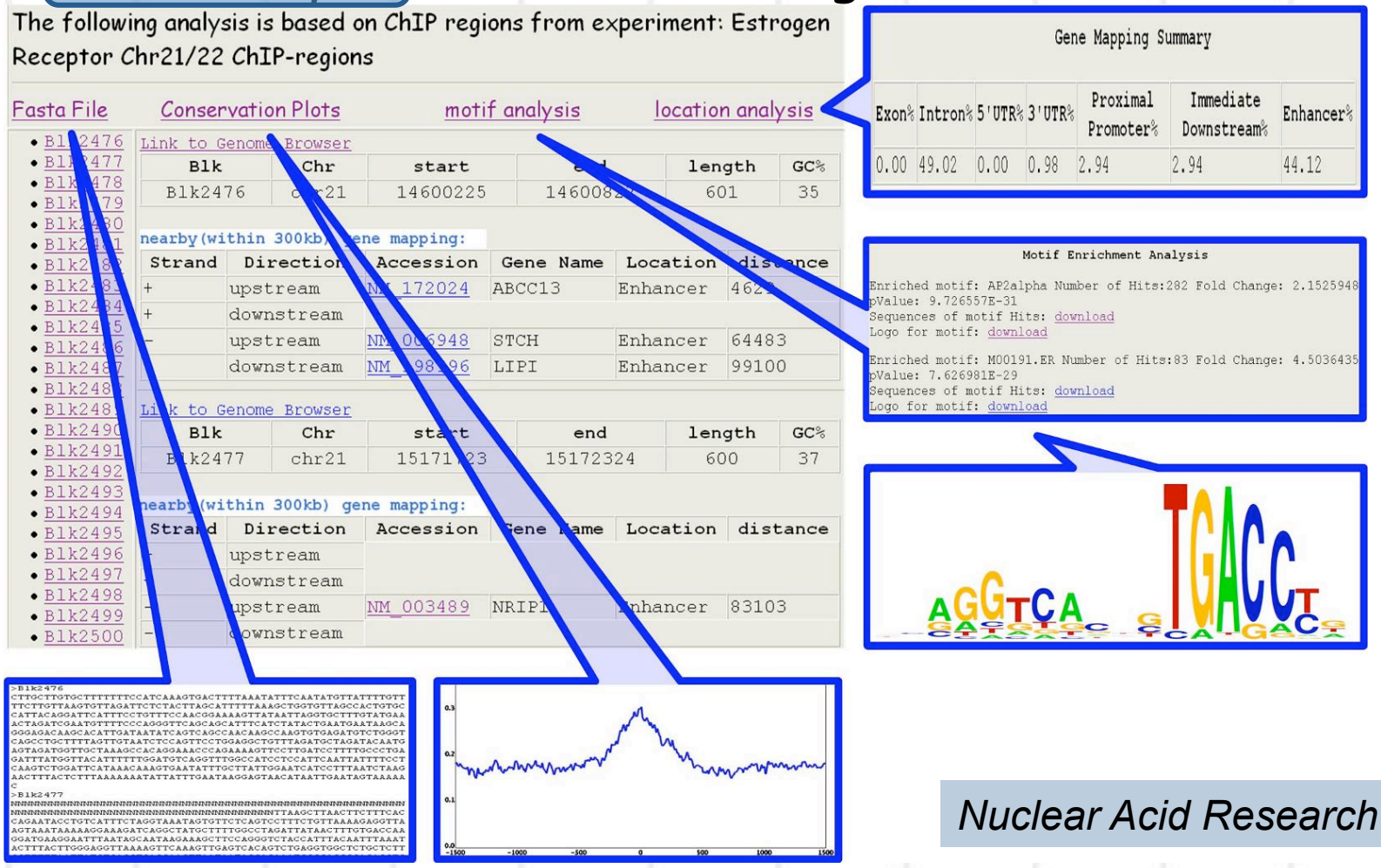


MAT: Reproducibility



CEAS: *Cis*-regulatory Element Annotation System

- Data Analysis Button for Biologists



Nuclear Acid Research (2006)

Mann-Whitney U-test for ChIP-region Detection

- Affy GTRANS/TAS, Cawley et al (*Cell* 2004):
  - Each probe: rank probes within [-500bp, +500bp] window, ~170 probes
  - Check whether sum of ChIP ranks is much smaller

	ctrl 1	ctrl 2	ChIP 1	ChIP 2
probe 1	1.71	2.23	3.02	2.25
probe 2	4.27	3.10	3.86	4.70
probe 3	4.06	3.67	4.03	4.74
probe 4	1.20	0.40	1.31	1.85
probe 5	4.29	3.95	4.56	4.76

	ctrl 1	ctrl 2	ChIP 1	ChIP 2
probe 1	17	15	13	14
probe 2	6	12	10	3
probe 3	7	11	8	2
probe 4	19	20	18	16
probe 5	5	9	4	1

Hidden Markov Model for ChIP-region Detection

- Li et al (Bioinformatics 2005), Carroll et all (Cell 2005)
- Collect and normalize data from as many labs and experiments as possible to estimate the behavior of each probe pair (PM-MM) ~ $N(\mu, \sigma^2)$
- Find regions with higher than baseline probes using HMM
  - Run HMM on ChIPs and Ctrls separately
  - Find high regions in ChIP that are low in Ctrl

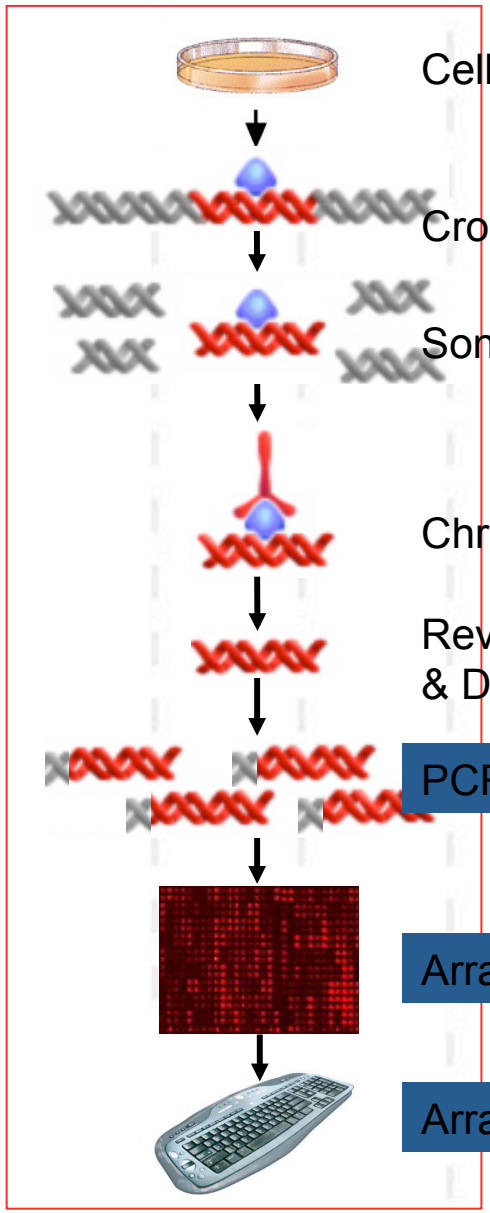
Empirical Bayes Shrinkage for ChIP-region Detection

- TileMap, Ji & Wong (Bioinformatics 2005)
- For each probe pair (either PM-MM or logPM):

$$t_i = \frac{\overline{X}_{i\,ChIP} - \overline{X}_{i\,Control}}{\sqrt{\hat{\sigma}^2_{i\,ChIP} / K_{ChIP} + \hat{\sigma}^2_{i\,Control} / K_{Control}}}$$

  - Variance adjusted by empirical Bayes based on observed variance and pooled variances from all the probes
- Then moving average of probe t's in a sliding window or HMM to identify ChIP-regions

ChIP on chip



Benchmark for ChIP-chip Target Detection

- ENCODE Spike-in experiment: both amplified and non-amplified

ChIP

~100 ENCODE clones, 0.25, 0.5, 1, 3,...,127X enrichment + total chromatin DNA

Control

total chromatin DNA

- Blind test
  - Samples hybridized to different tiling arrays.
  - Predictions made before the key was released

Genome Research (2008)

Affymetrix

Agilent

NimbleGen

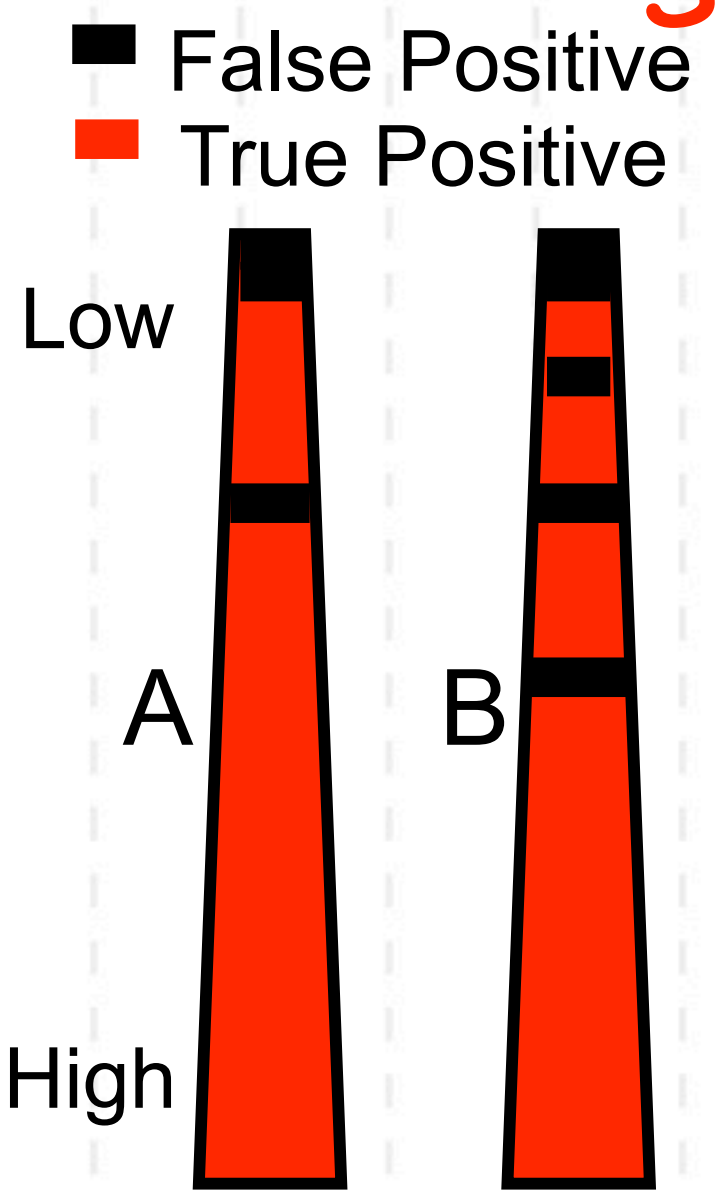
Others

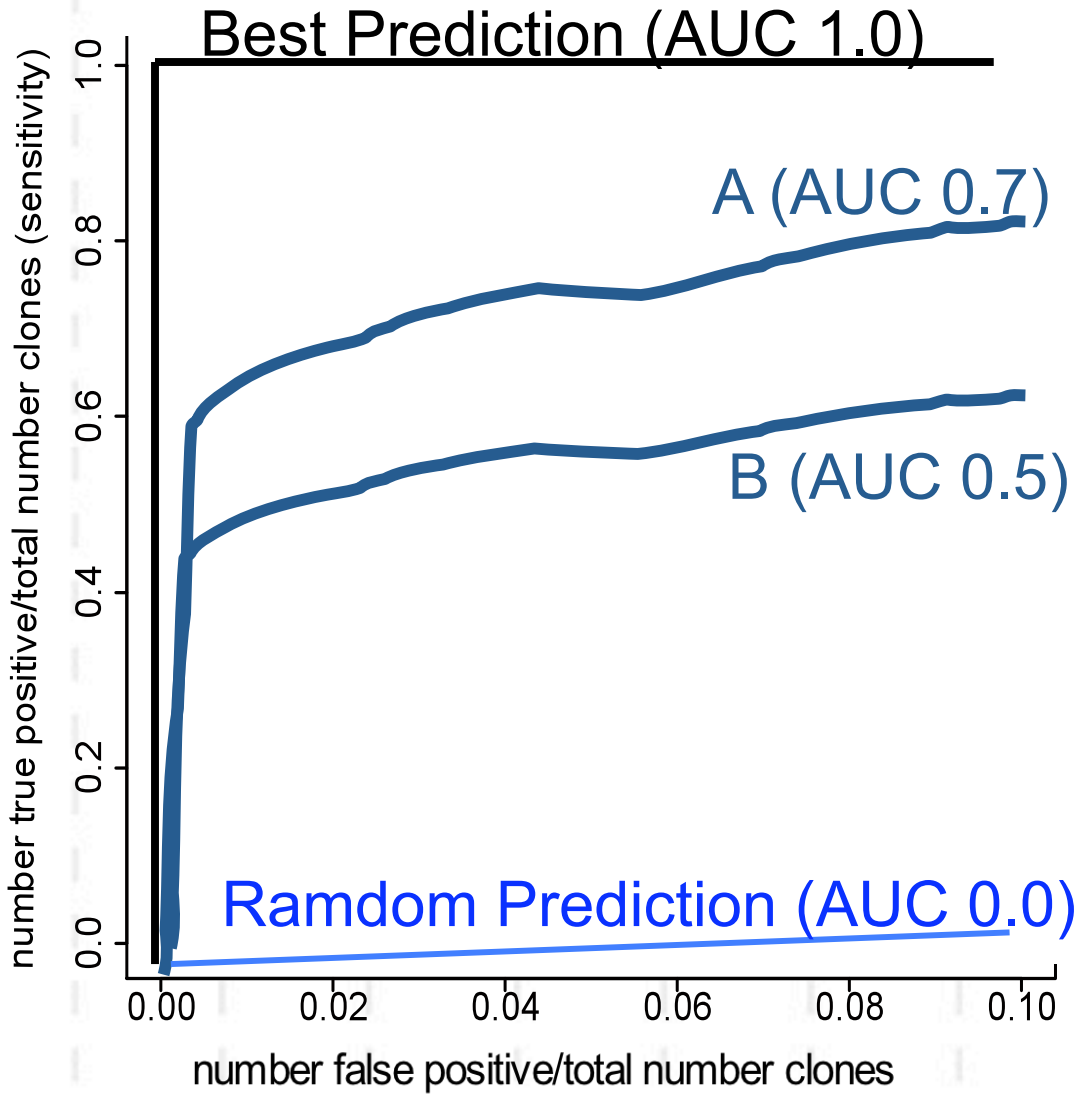
Coordinators

Sanger_Koch_PCRArray

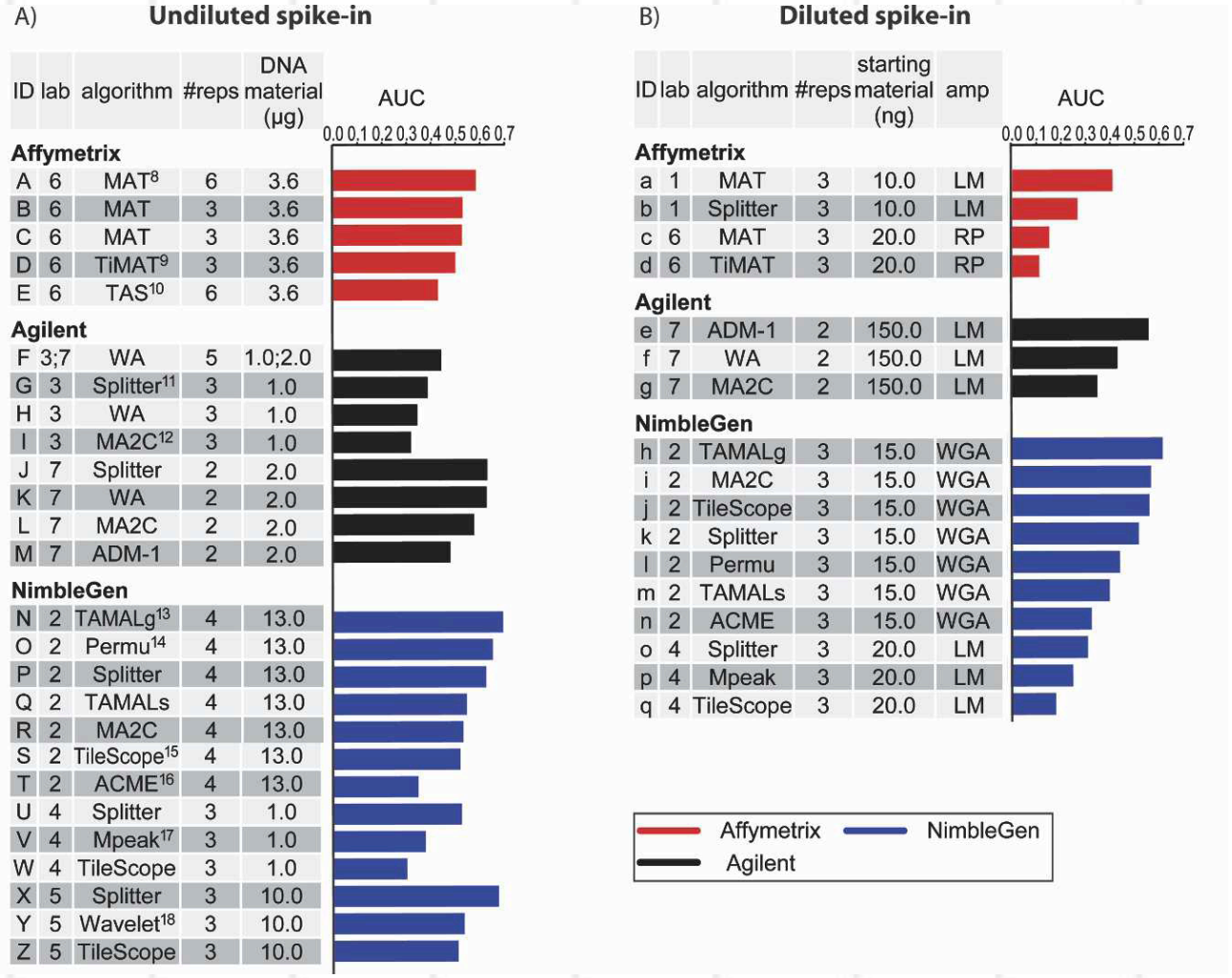
Singapore_Ruan_ChIP-PET

Sens/Spec


False Positive
True Positive
Low
A
B
High

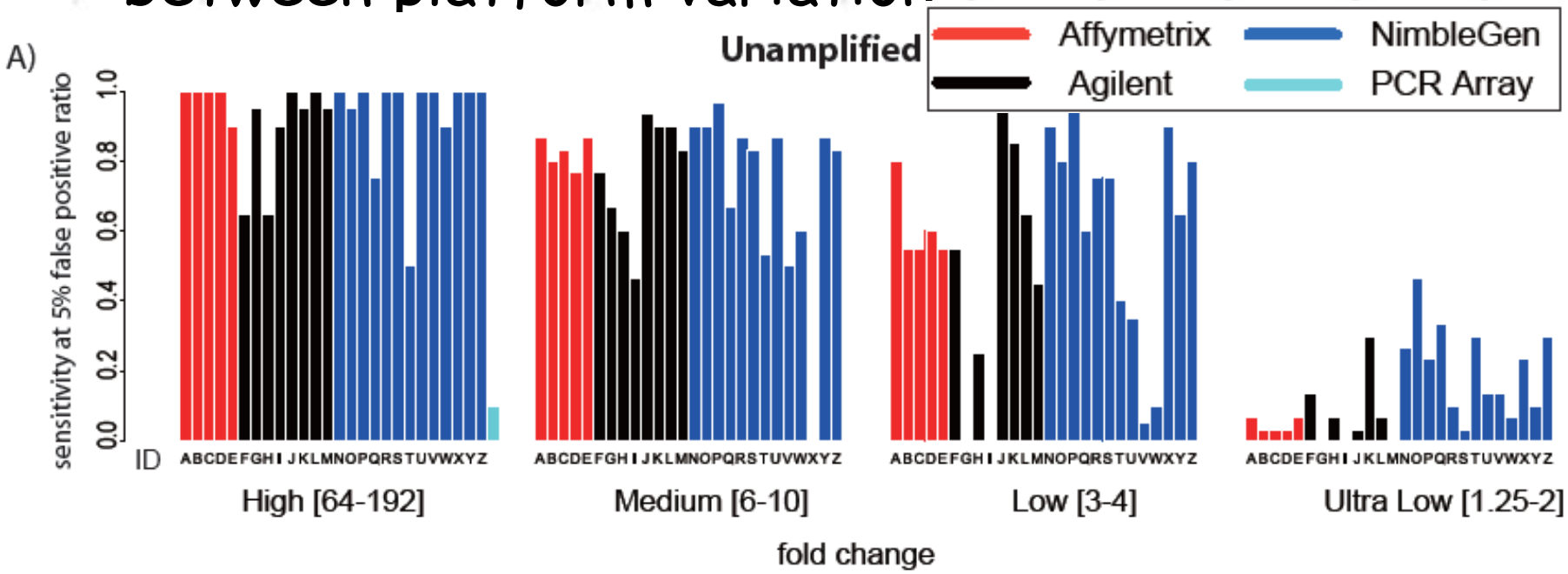

Best Prediction (AUC 1.0)
A (AUC 0.7)
B (AUC 0.5)
Ramdom Prediction (AUC 0.0)
number true positive/total number clones (sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
number false positive/total number clones

Sens/Spec

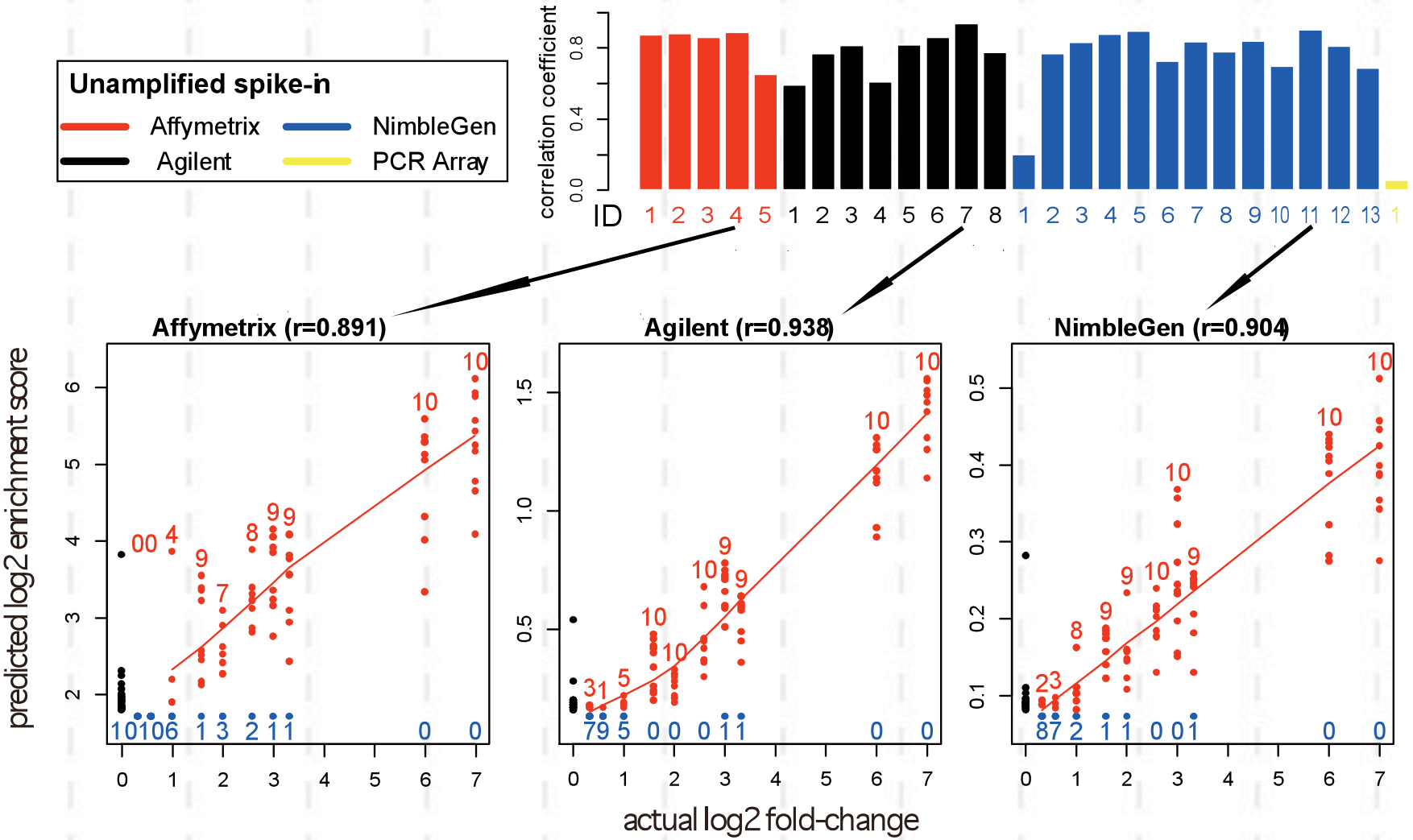

A)
Undiluted spike-in
ID lab algorithm #reps DNA material (µg)
AUC
0,0 0,1 0,2 0,3 0,4 0,5 0,6 0,7
Affymetrix
A 6 MAT[8] 6 3.6
B 6 MAT 3 3.6
C 6 MAT 3 3.6
D 6 TiMAT[9] 3 3.6
E 6 TAS[10] 6 3.6
Agilent
F 3;7 WA 5 1.0;2.0
G 3 Splitter[11] 3 1.0
H 3 WA 3 1.0
I 3 MA2C[12] 3 1.0
J 7 Splitter 2 2.0
K 7 WA 2 2.0
L 7 MA2C 2 2.0
M 7 ADM-1 2 2.0
NimbleGen
N 2 TAMALg[13] 4 13.0
O 2 Permu[14] 4 13.0
P 2 Splitter 4 13.0
Q 2 TAMALs 4 13.0
R 2 MA2C 4 13.0
S 2 TileScope[15] 4 13.0
T 2 ACME[16] 4 13.0
U 4 Splitter 3 1.0
V 4 Mpeak[17] 3 1.0
W 4 TileScope 3 1.0
X 5 Splitter 3 10.0
Y 5 Wavelet[18] 3 10.0
Z 5 TileScope 3 10.0
B)
Diluted spike-in
ID lab algorithm #reps starting material (ng) amp
AUC
0,0 0,1 0,2 0,3 0,4 0,5 0,6 0,7
Affymetrix
a 1 MAT 3 10.0 LM
b 1 Splitter 3 10.0 LM
c 6 MAT 3 20.0 RP
d 6 TiMAT 3 20.0 RP
Agilent
e 7 ADM-1 2 150.0 LM
f 7 WA 2 150.0 LM
g 7 MA2C 2 150.0 LM
NimbleGen
h 2 TAMALg 3 15.0 WGA
i 2 MA2C 3 15.0 WGA
j 2 TileScope 3 15.0 WGA
k 2 Splitter 3 15.0 WGA
l 2 Permu 3 15.0 WGA
m 2 TAMALs 3 15.0 WGA
n 2 ACME 3 15.0 WGA
o 4 Splitter 3 20.0 LM
p 4 Mpeak 3 20.0 LM
q 4 TileScope 3 20.0 LM
Affymetrix
NimbleGen
Agilent

Sensitivity at Different Fold Changes

- Commercial tiling arrays are comparable ≥ 3 fold, NimbleGen is more sensitive at low FC
- Between lab/analysis variation is bigger than between platform variation

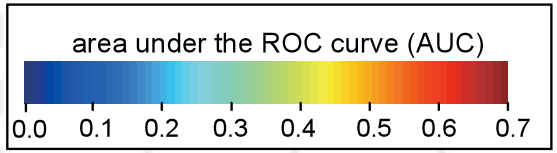


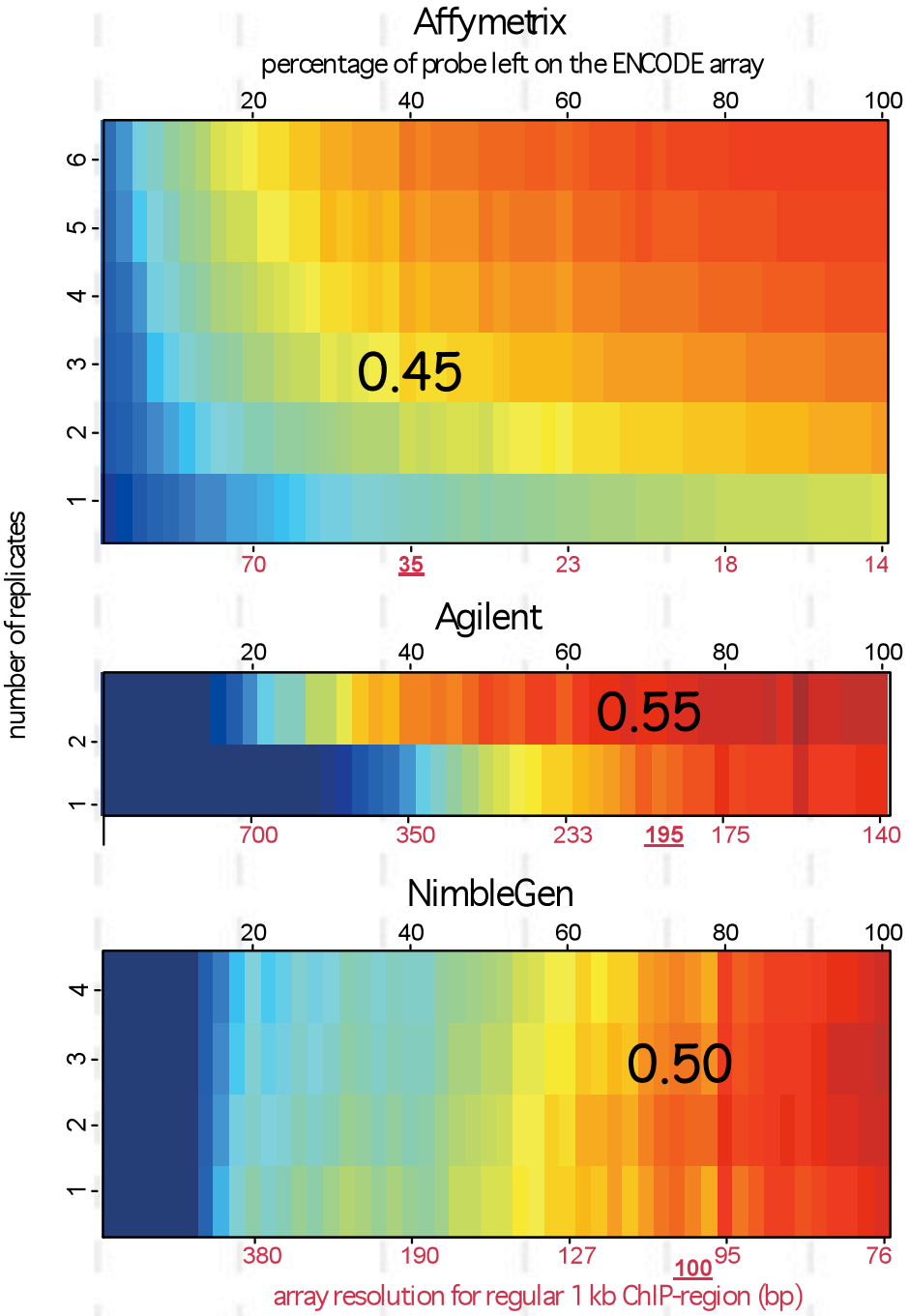
Quantitative Evaluation



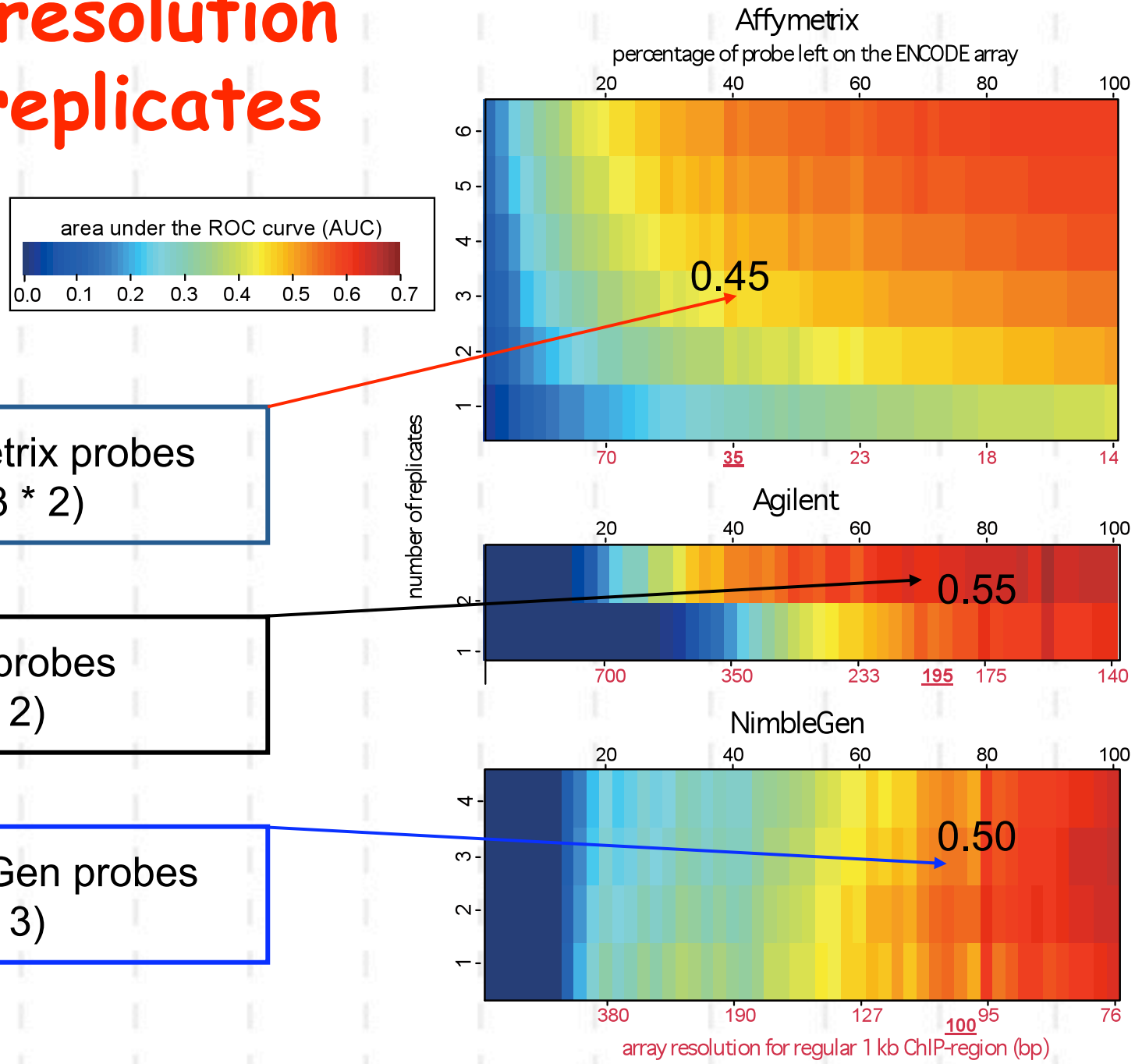
Tiling Resolution and Replicates

- Gradually deleting probes on the array
- Make predictions using different number of replicates





Tiling resolution and replicates

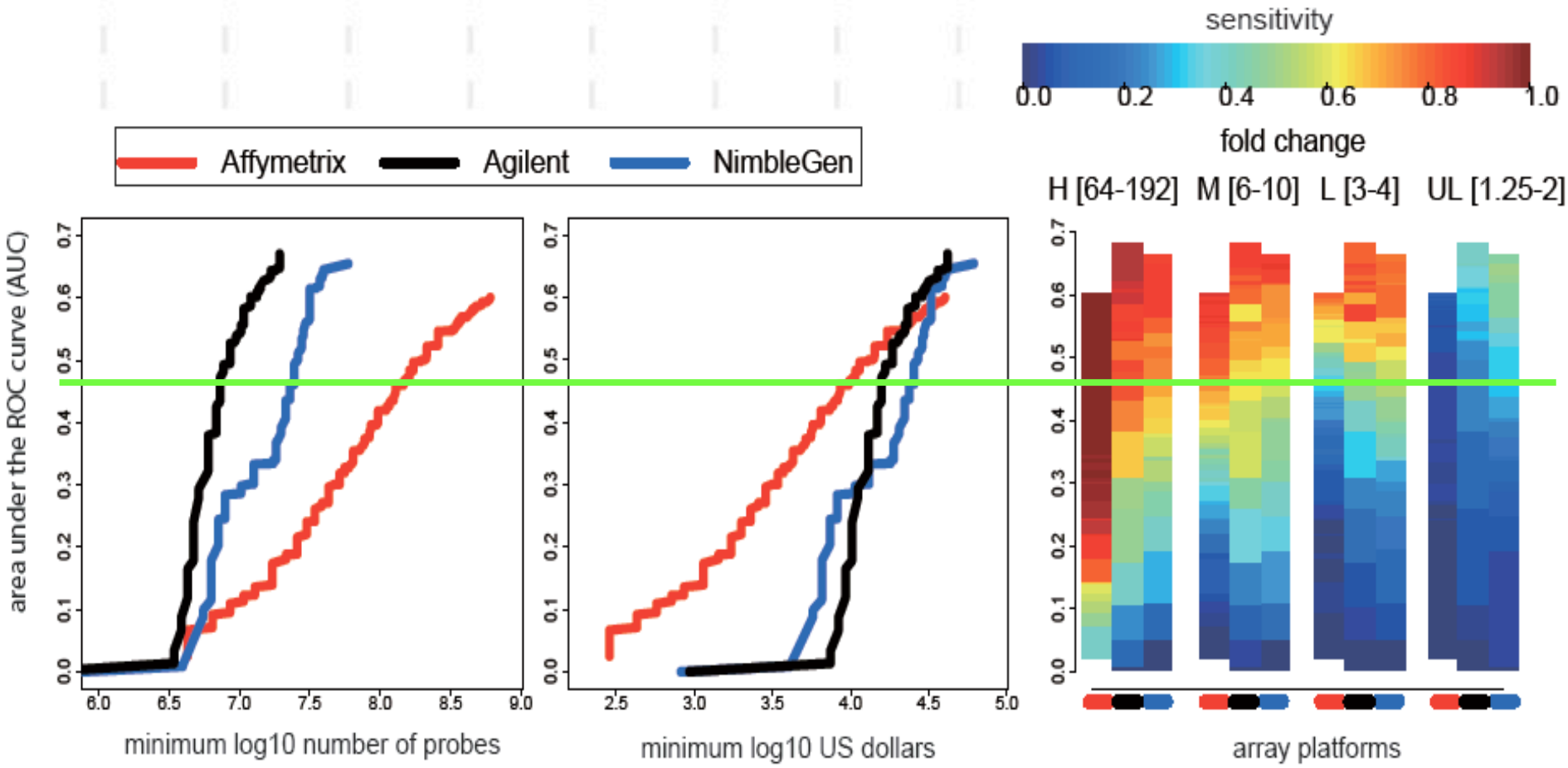


258M Affymetrix probes
(1.5G / 35 * 3 * 2)

15M Agilent probes
(1.5G / 195 * 2)

45M NimbleGen probes
(1.5G / 100 * 3)

Translate into Cost vs Performance



Outline

- Introduction
- ChIP-chip with genome tiling arrays
- ChIP-seq with next-gen sequencing
- Estrogen Receptor and FoxA1 regulation in breast and prostate cancers

Conventional sequencing

- Can sequence up to 1,000 bp, and per-base 'raw' accuracies as high as 99.999%. In the context of high-throughput shotgun genomic sequencing, Sanger sequencing costs on the order of $0.50 per kilobase.

Jay Shendure & Hanlee Ji, Nature Biotechnology 26, 1135 - 1145 (2008)

Next-generation DNA sequencing technologies

Table 1. Second-generation DNA sequencing technologies

▲ Figures and tables index Next table ▸

	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length	References
454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	Indel	250 bp	14,20
Solexa	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	Subst.	36 bp	17,22
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	Subst.	35 bp	13,26
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	Subst.	13 bp	13,20
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	Del	30 bp	18,30

The pace with which the field is moving makes it likely that estimates for costs and read-lengths will be quickly outdated. Vendors including Roche Applied Science, Illumina, and Applied Biosystems have major upgrade releases currently in progress. Estimated costs-per-megabase are approximate and inclusive only of reagents. Read-lengths are for single tags. Subst., substitutions; indel, insertions or deletions; del, deletions.

▲ Figures and tables index Next table ▸

Jay Shendure & Hanlee Ji, Nature Biotechnology 26, 1135 - 1145 (2008)

Applications of next-generation sequencing

Table 2. Applications of next-generation sequencing

Previous table | Figures and tables index | Next table

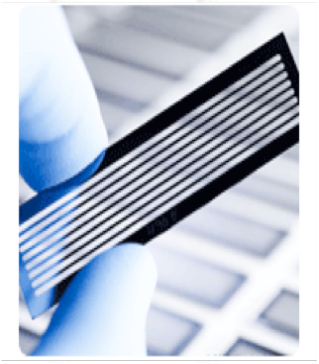
Category	Examples of applications	Refs
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes	44
Reduced representation sequencing	Large-scale polymorphism discovery	45
Targeted genomic resequencing	Targeted polymorphism and mutation discovery	46,47,48,49,50,51,52
Paired end sequencing	Discovery of inherited and acquired structural variation	53,54
Metagenomic sequencing	Discovery of infectious and commensal flora	55
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	56,57,58,59,60,61,62,63
Small RNA sequencing	microRNA profiling	64
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA	60,65,66
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions	67,68,68,69,70
Nuclease fragmentation and sequencing	Nucleosome positioning	69
Molecular barcoding	Multiplex sequencing of samples from multipleindividuals	61,71

Previous table | Figures and tables index | Next table

Flexible for all DNA/RNA applications

Jay Shendure & Hanlee Ji, Nature Biotechnology 26, 1135 - 1145 (2008)

When Sequencing Met Microarrays



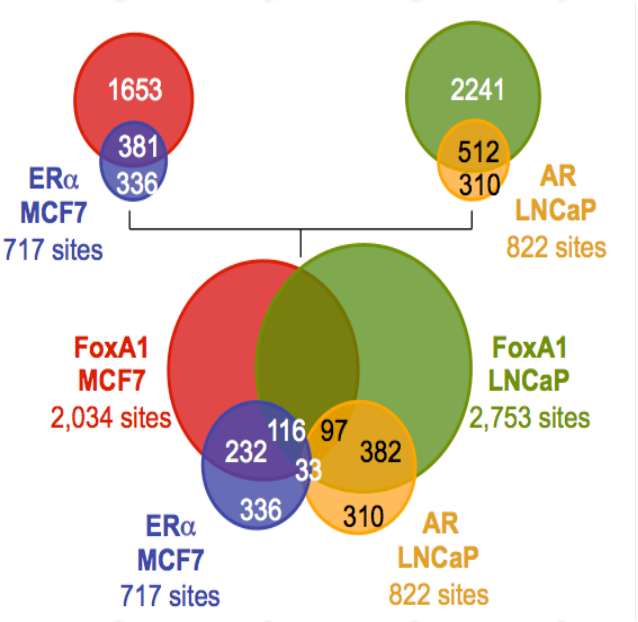
VS.

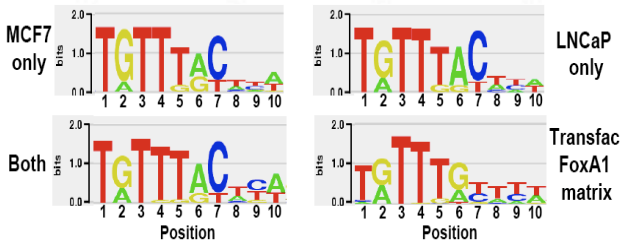
"ChIP-seq offers important advantages over ChIP-chip, including lower cost, minimal hands-on processing and a requirement for fewer replicate experiments as well as less input material."

Elaine R Mardis

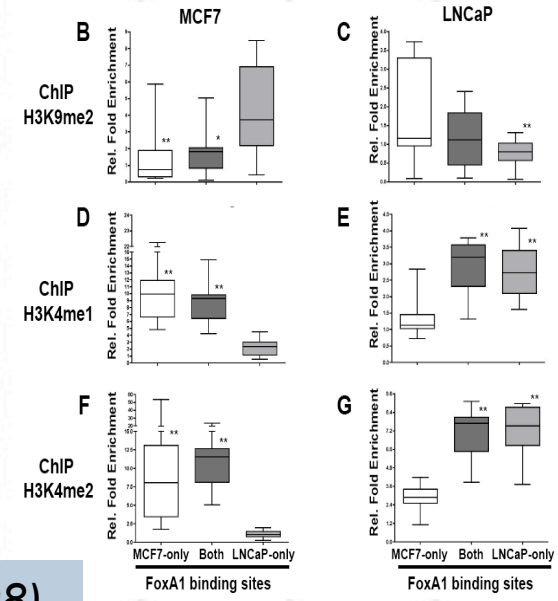
Nature Methods - **4**, 613 - 614 (2007)

FoxA1 Regulation





- FoxA1 ⟷ ER in breast cancer
- FoxA1 ⟷ AR in prostate cancer
- Distinct sets of targets in breast and prostate cancer, mostly at intergenic enhancers



Cell (2005); Nature Genetics (2006); Cell (2008)

ChIP-seq Preliminary Data Analysis

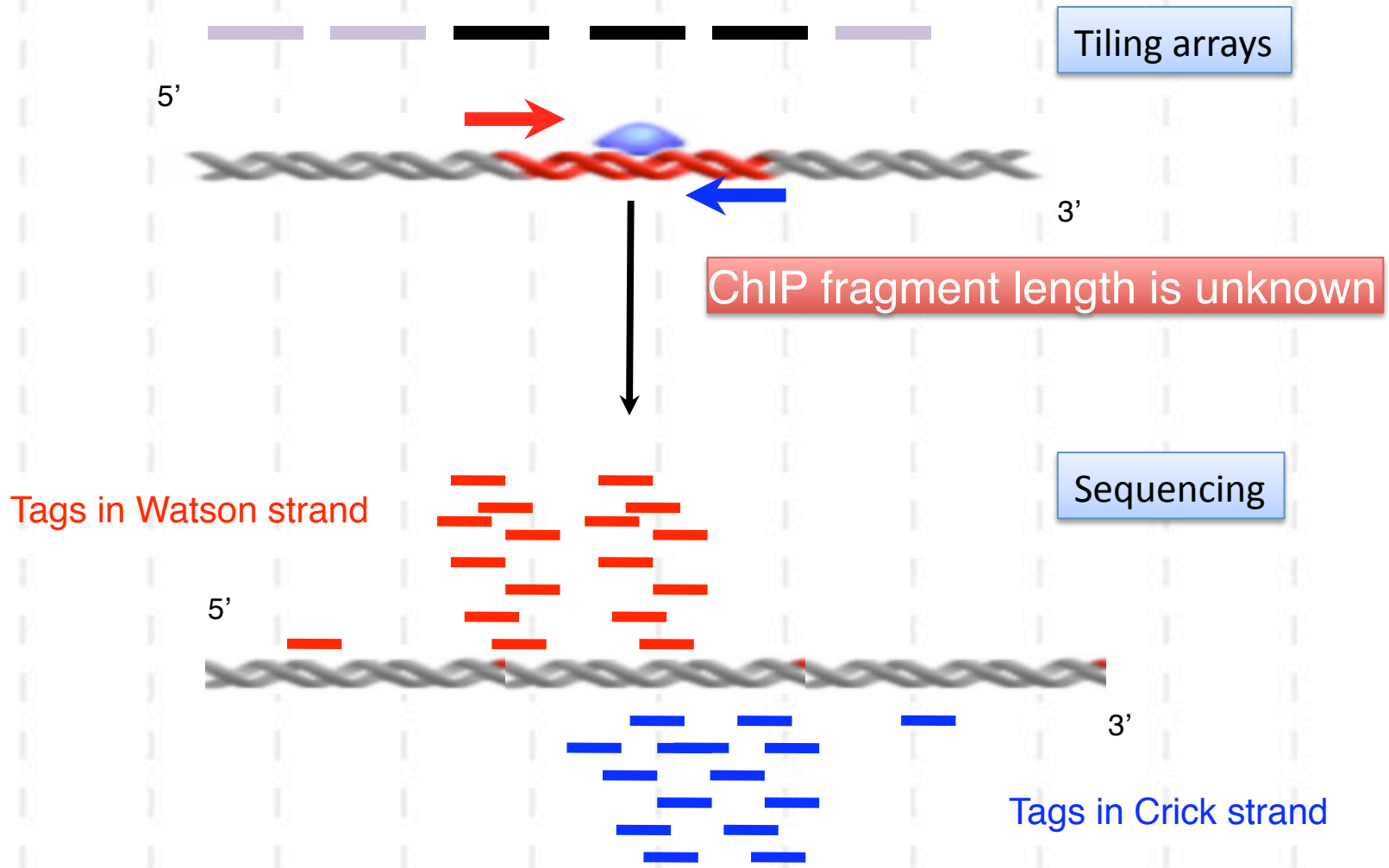
Read Alignment

- ELAND
- SOAP
- MAQ
- RMAP
- SeqMap
- …

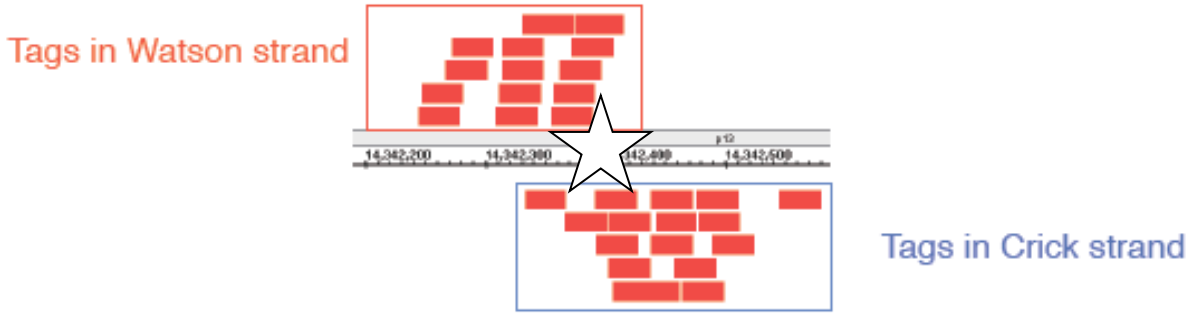
Peak Detection

- MACS
- FindPeaks
- CHiPSeq
- SISSRs
- QuEST
- …

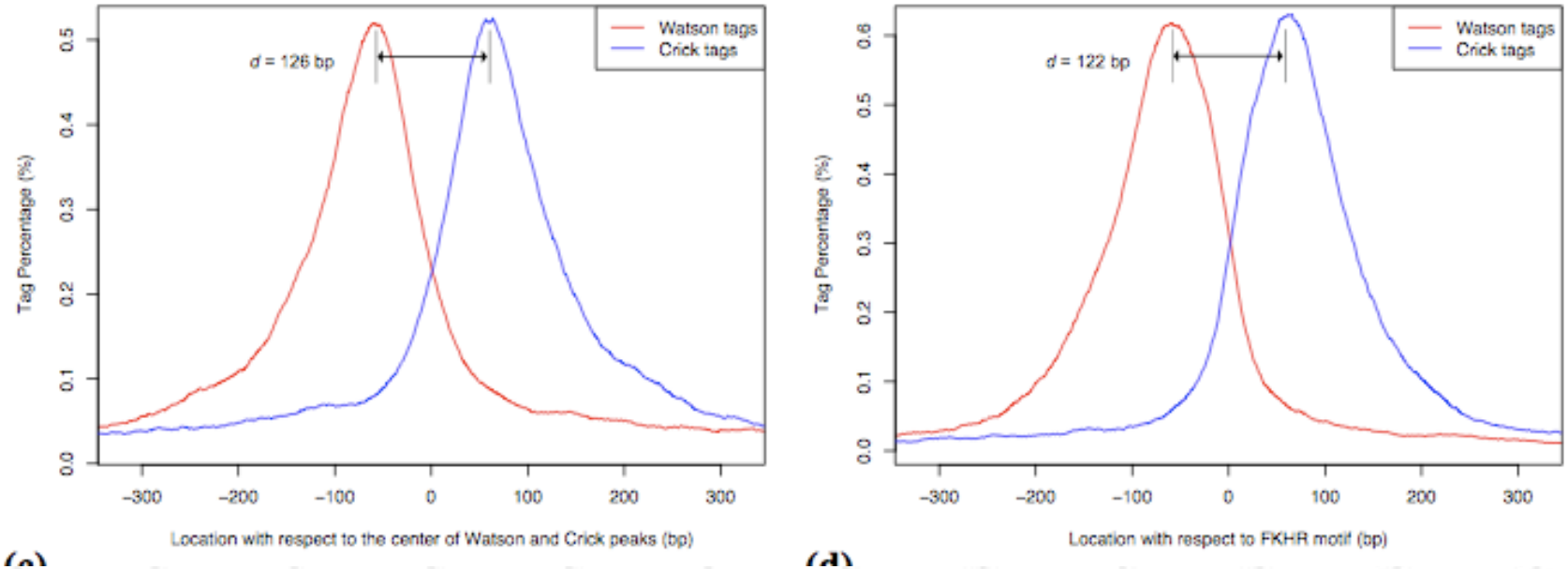
ChIP-Seq


Tiling arrays
5’
3’
ChIP fragment length is unknown
Sequencing
Tags in Watson strand
5’
3’
Tags in Crick strand

Tag Shift Size Model



- Model shift size from the most confident peaks



Peak Calls

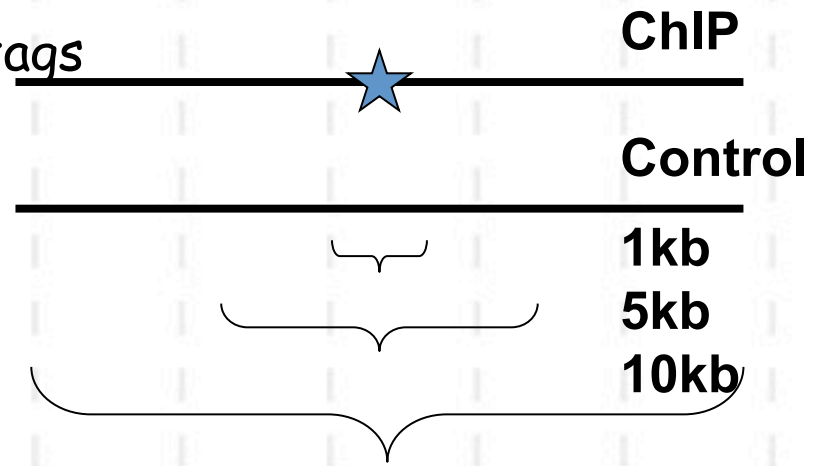
- Tag distribution along the genome ~ Poisson distribution

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

λ = expected number of occurrences of an event
k = observed number of occurrences of an event

- ChIP-Seq show local biases in the genome
  - biases from sequencing, mapping, chromatin structure and genome copy number variations.
  - 300bp controls have to few tags
  - But can look further

Dynamic λ_{local} = max(λ_{BG}, [λ_{1k},] λ_{5k}, λ_{10k})



FoxA1 ChIP-seq

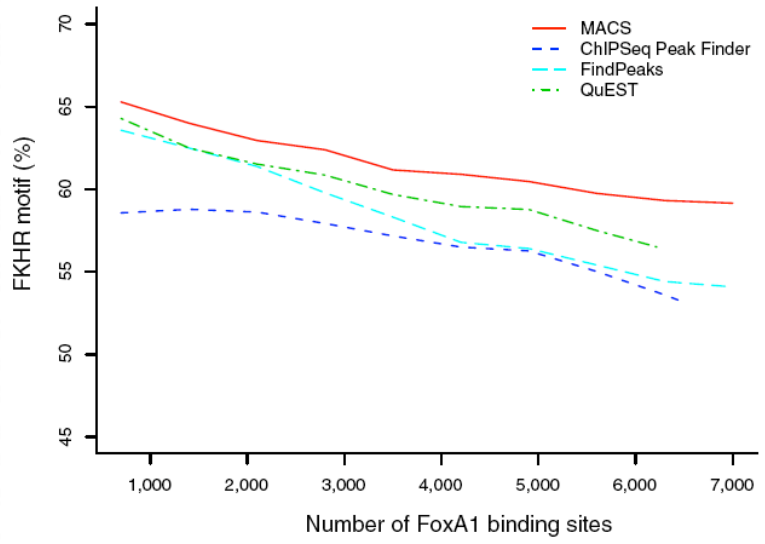
MACS
ChIPSeq Peak Finder
FindPeaks
QuEST

Johnson et al., Science 2007

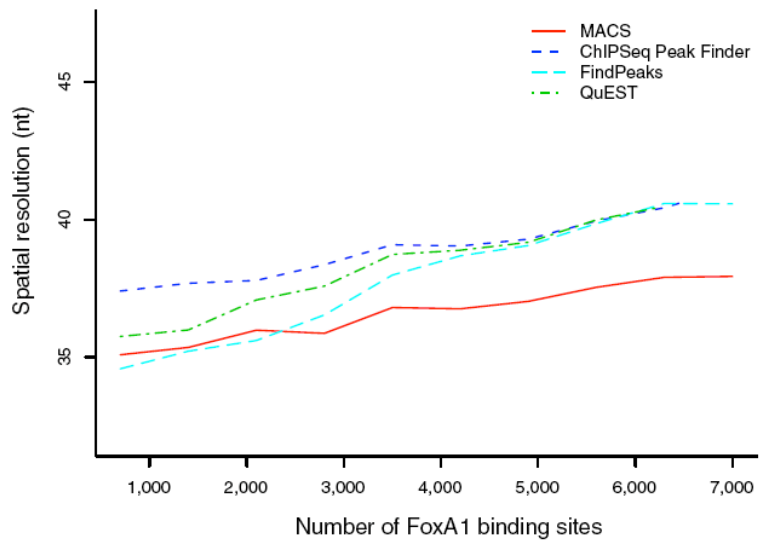
Unpublished, Genome Canada

Valouev et al., Nat Methods 2008

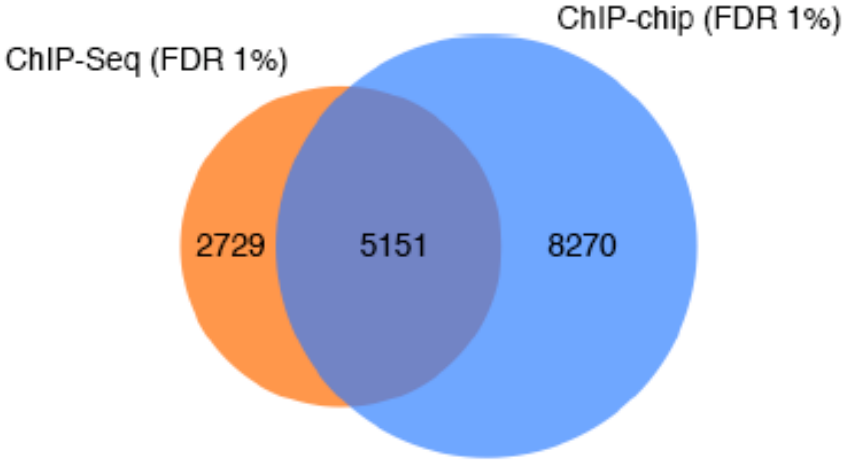
(c)

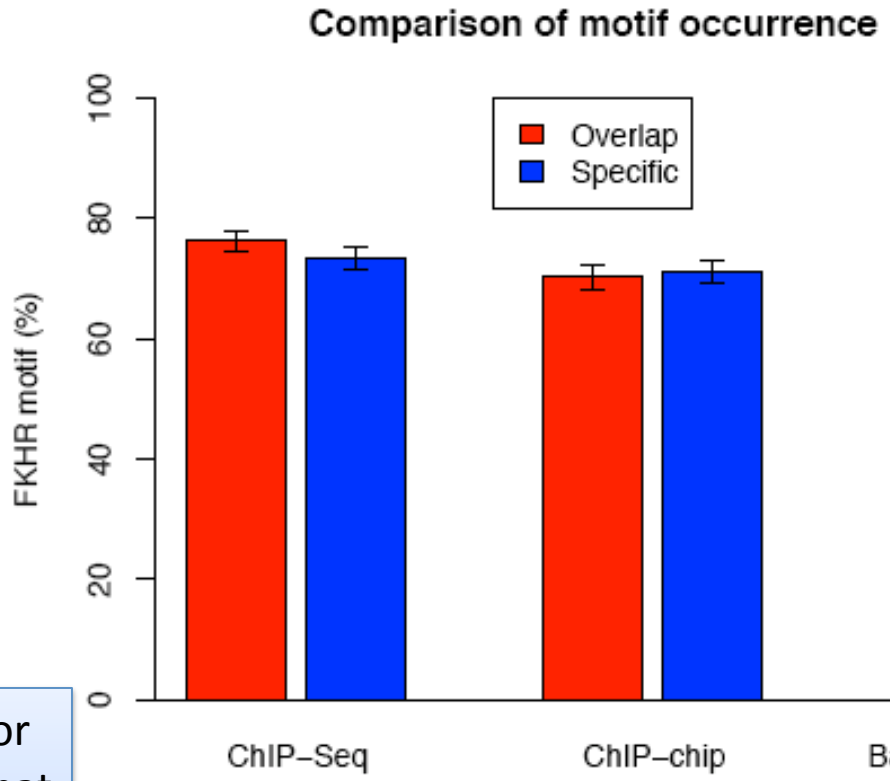


(e)



ChIP-seq and ChIP-chip Comparison (FoxA1)





- The FKHR motif occurrence from ChIP-chip or ChIP-Seq specific peaks is comparable with that from the overlapping peaks.
- This suggests that most of the platform-specific peaks are genuine binding sites.

ChIP-Seq Maybe Ineffective at Mapping Inactive Histone Marks

- Close chromatin are harder to sonicate, so resulting fragments are larger. ChIP-seq library construction biases shorter fragments, so may reduce the detectable enrichment of inactive marks
- Mikkelsen et al, Nat 2007, differentiation (chromatin closing) weakens ChIP-seq efficiency of inactive marks but not active marks

More Differentiated

		ES	MEF	NP
active	H3K4me3	19524/8.9M	16738/11.3M	17432/6.5M
inactive	H3K27me3	4652/6.5M	6548/11.4M	2215/7.9M
inactive	H3K9me3	1789/4.2M	991/3.7M	446/3.9M

#peaks

#tags

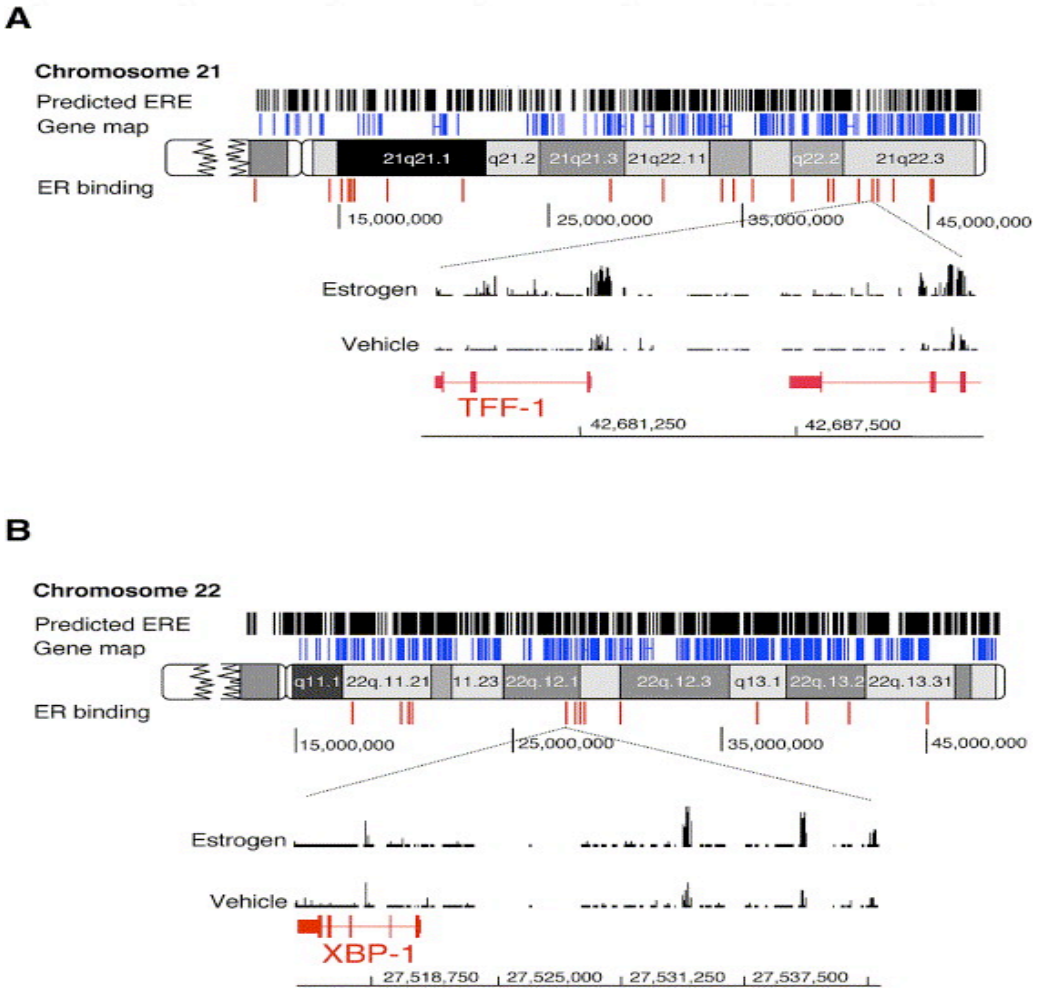
Outline

- Introduction
- ChIP-chip with genome tiling arrays
- ChIP-seq with next-gen sequencing
- Estrogen Receptor and FoxA1 regulation in breast and prostate cancers
  - FoxA1 translates epigenetic signatures into lineage-specific transcription.Cell 132 (2008) 958-970
  - Genome-wide analysis of estrogen receptor binding sites. Nat. Genet. 38 (2006) 1289-1297.
  - Chromosome-Wide Mapping of Estrogen Receptor Binding Reveals Long-Range Regulation Requiring the Forkhead Protein FoxA1. Cell 122 (2005) 33-43.

Nuclear Hormone Receptor Superfamily

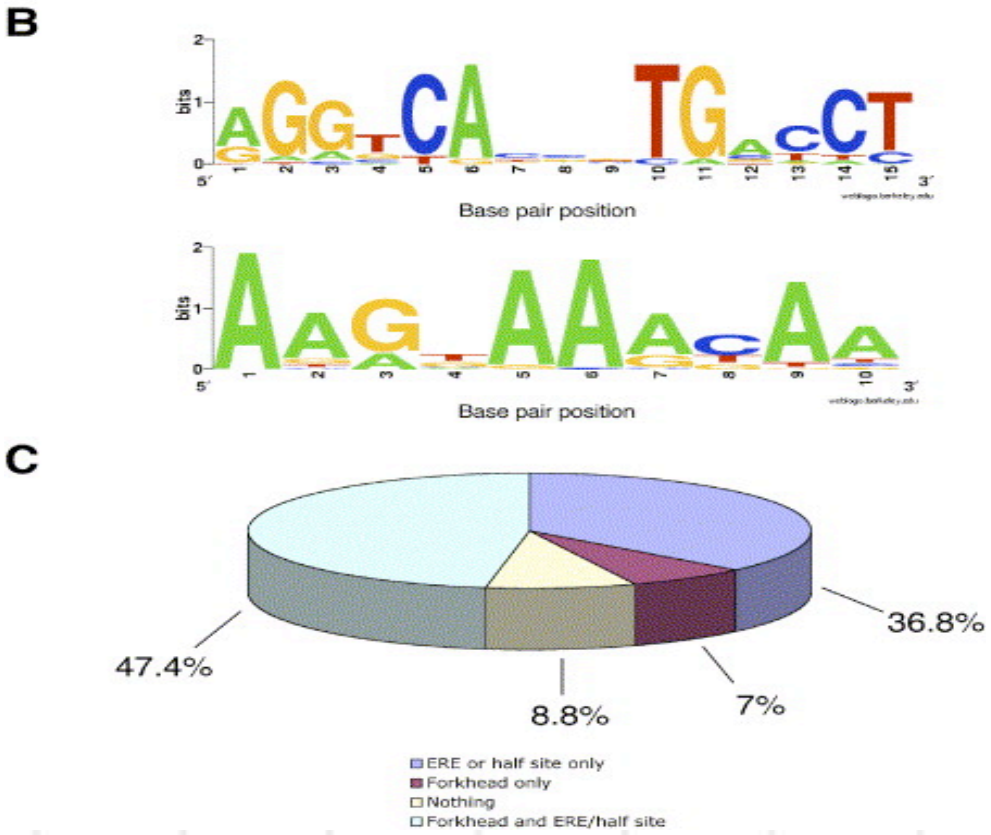
- The largest class of eukaryotic transcription factors
- Type I receptors (steroid) : progestins (PR), **estrogens (ER), androgens (AR)**, glucocorticoids (GR) and mineralocorticoids (MR), PPARγ
- Type II receptors (non-steroid): thyroid hormone (TR), vitamin D (VDR), 9-*cis* (RXRs) and all-*trans* retinoic acid (RARs)
- Orphan receptors: cognate ligands are unknown

ChIP-chip identified ER binding sites in chromosome 21 and 22

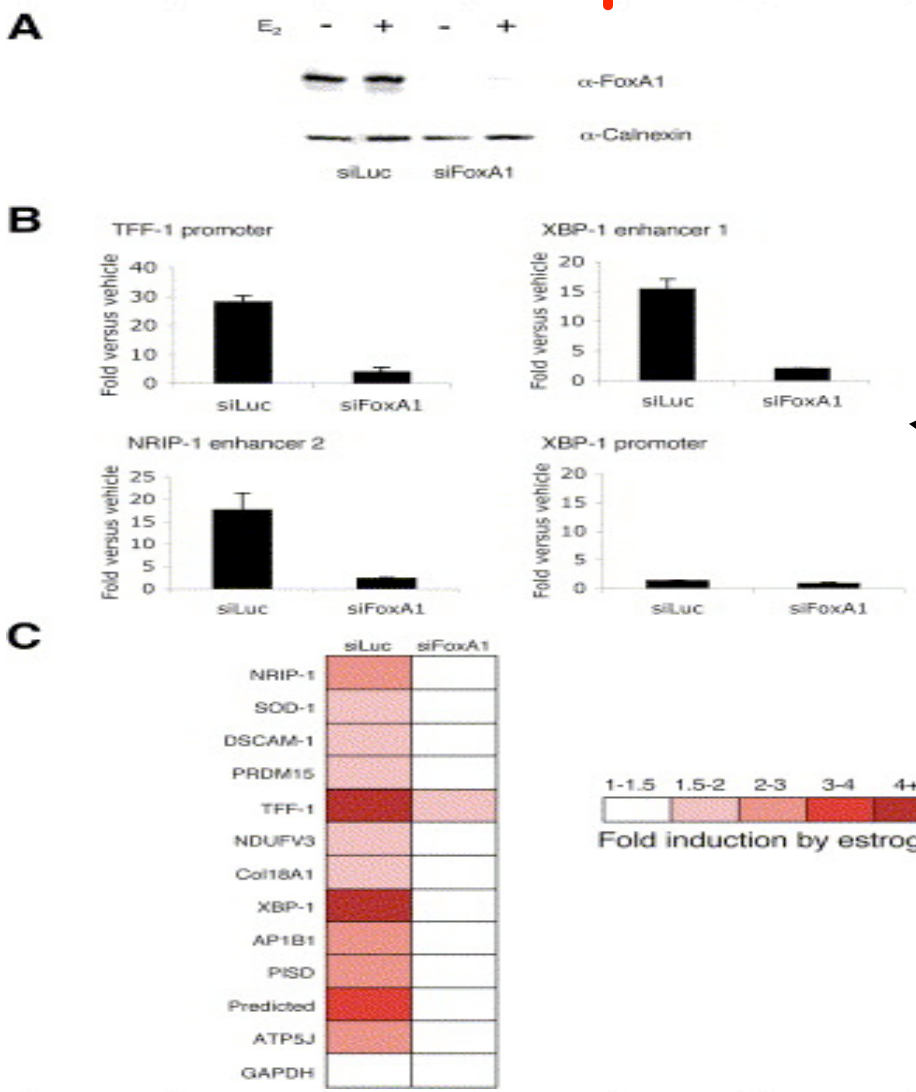


Carroll et al. Cell 2005

De novo motif discovery found FoxA1



Specific targeted knockdown of FoxA1 and the effects on estrogen-mediated transcription

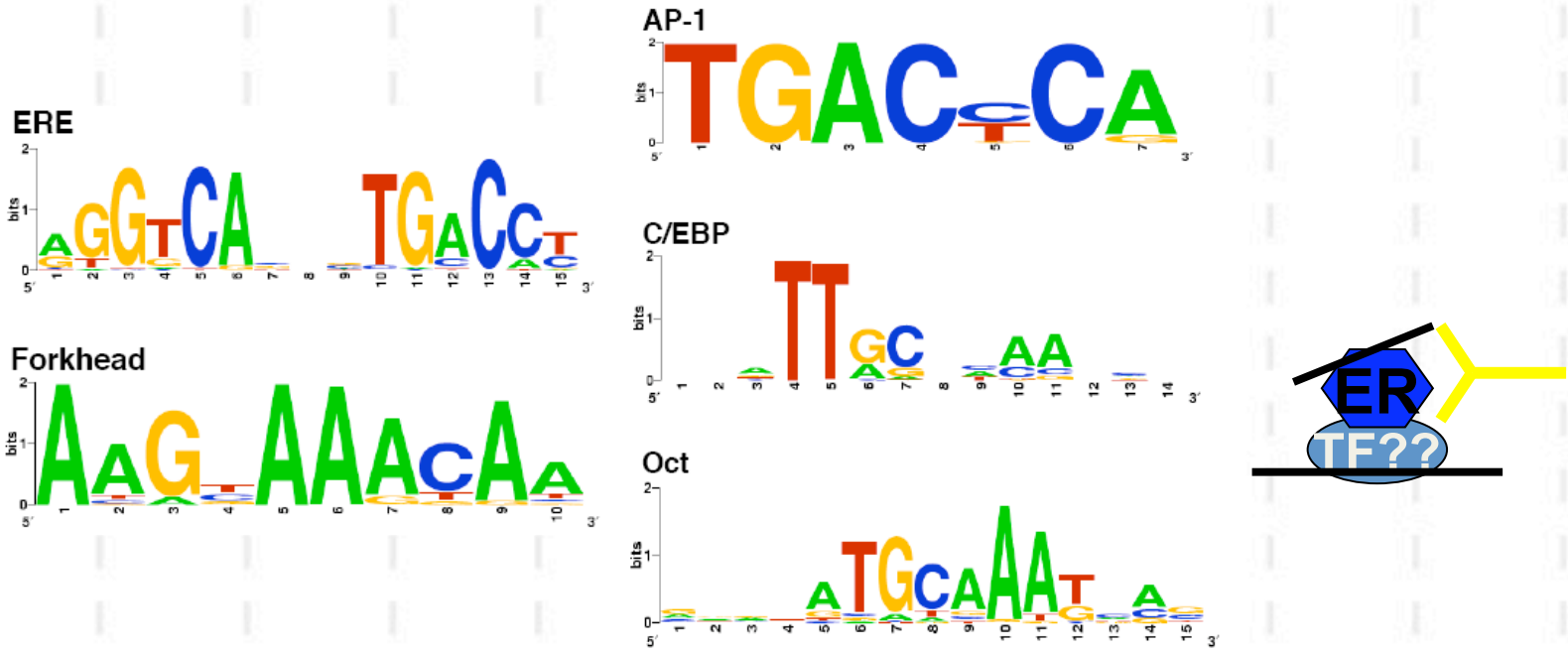


ER-binding after FoxA1 knockdown

Changes in mRNA levels (estrogen mediated vs. control fold change) after FoxA1 knockdown

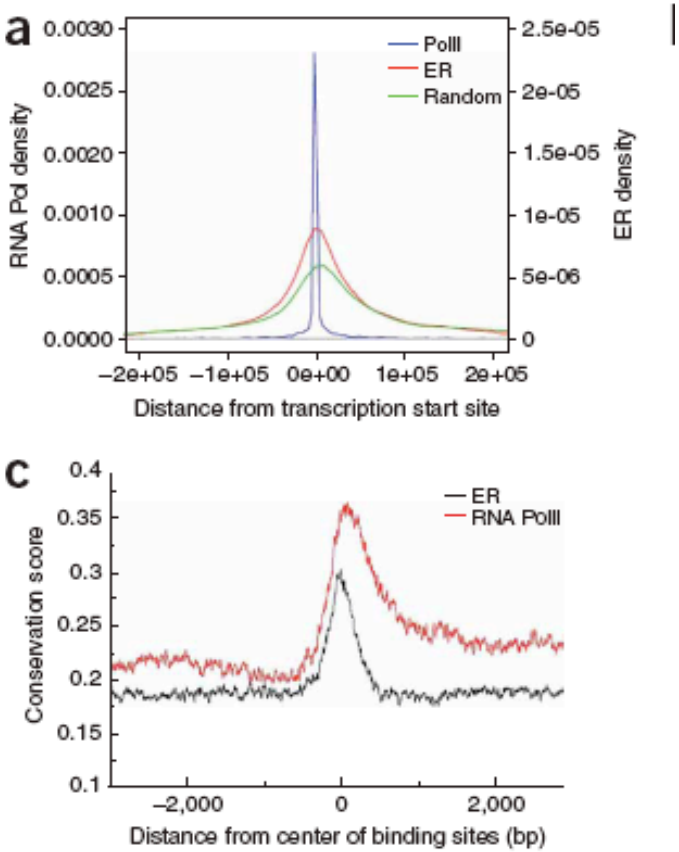
Estrogen Receptor (ER) Cistrome in Breast Cancer

- ~7000 unique ER binding sites at 5% FDR
- Sequence pattern finding from all ER ChIP-regions
  - Both de novo (MDscan) and TRANSFAC mapping



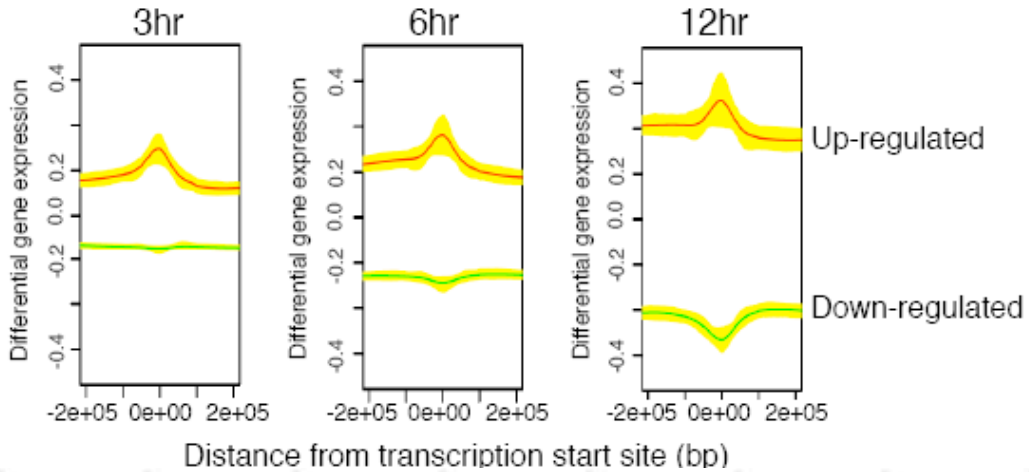
Carroll et al, Nat Gen 2006

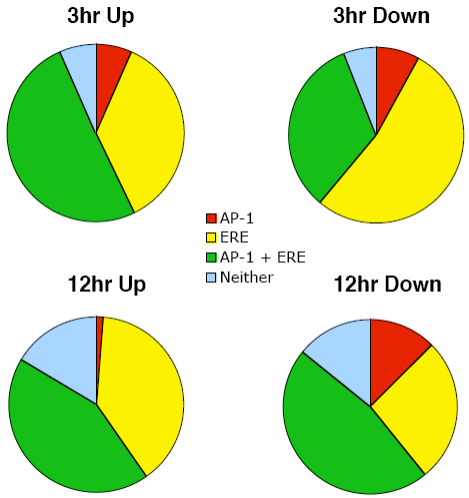
Estrogen Receptor (ER) Cistrome in Breast Cancer



ER Regulation Mechanism

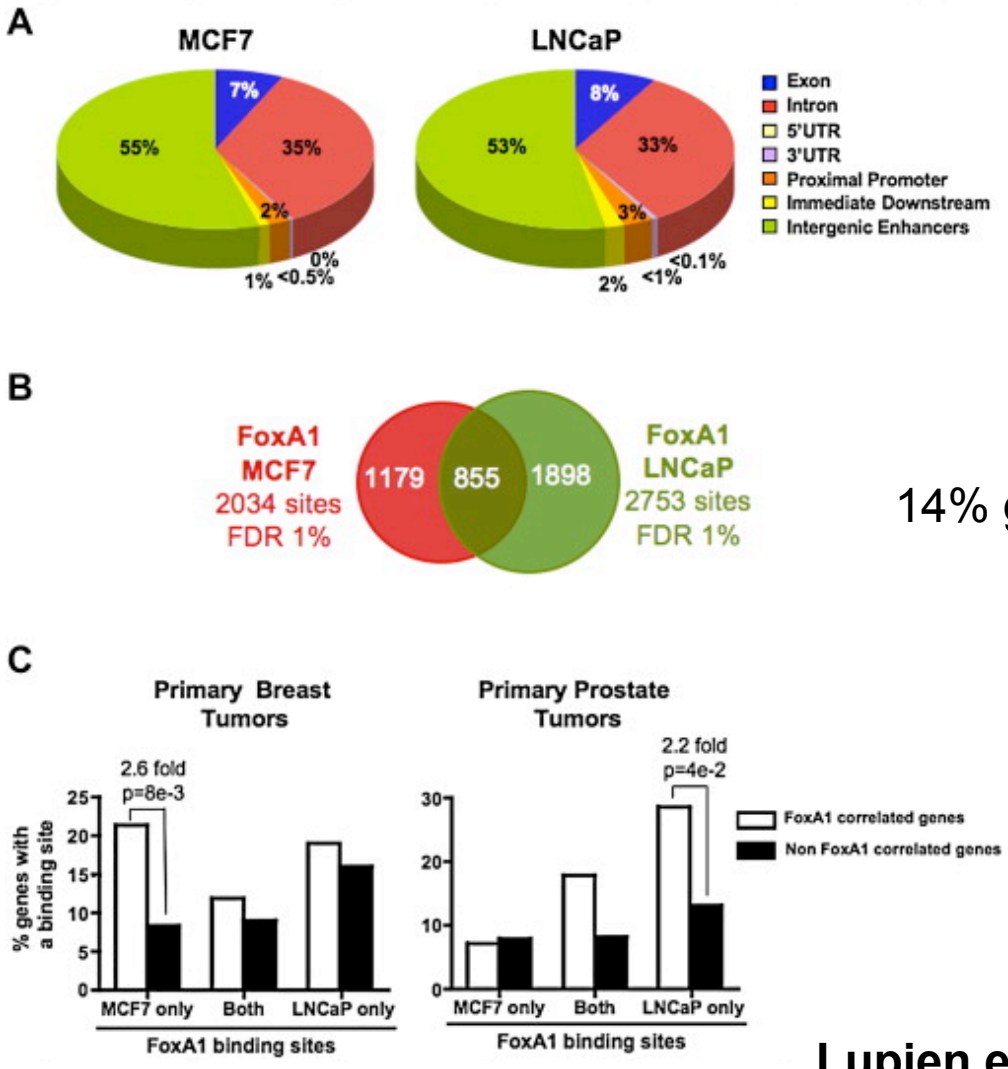
- 3hr, ER binding regulates mostly up genes
- 12 hr, ER binding regulates both up and down genes





Down through AP-1 and NRIP
siRNA NRIP abolish down regulation

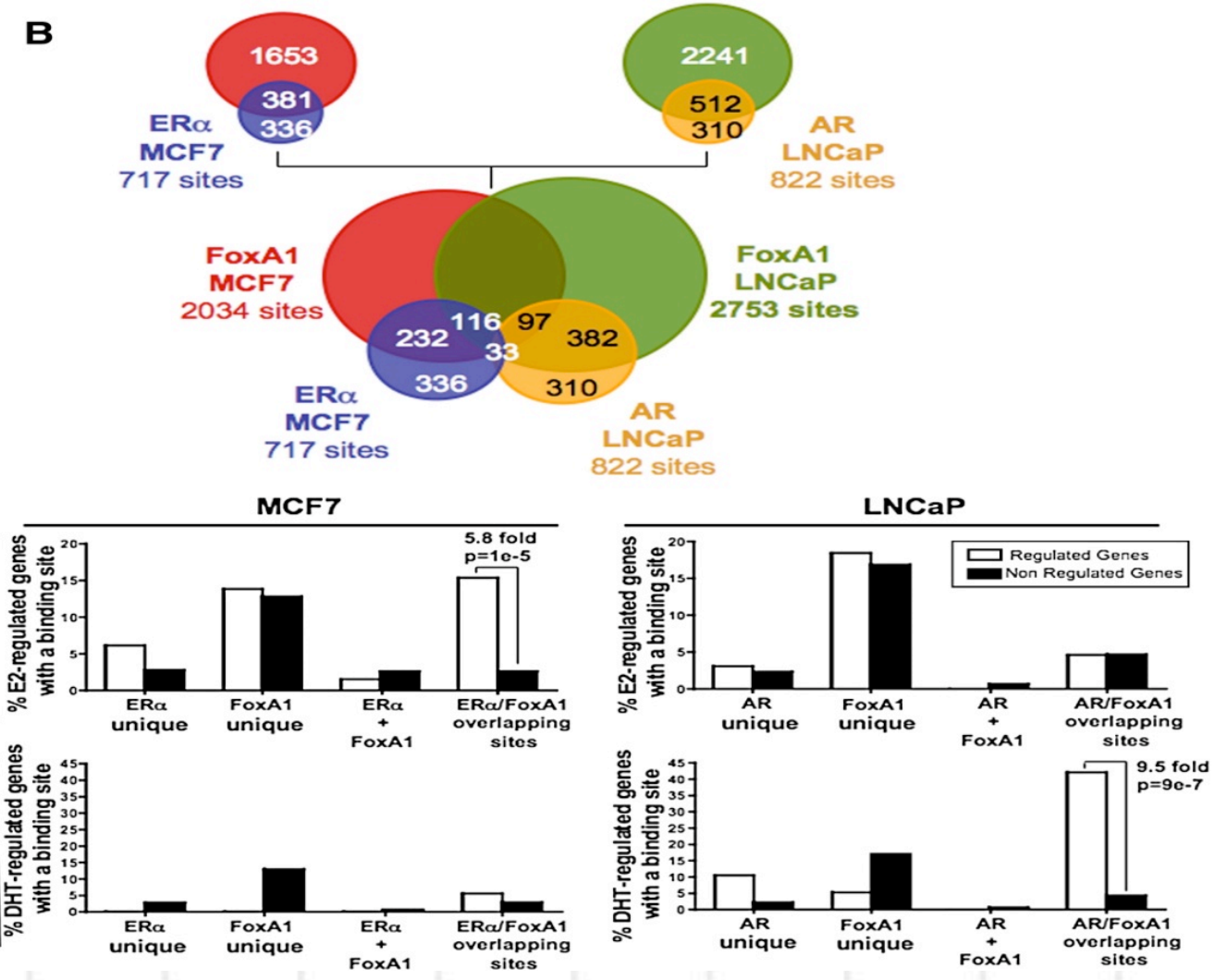
Cell Type-Specific Recruitment of FoxA1 Correlates with Differential Gene Expression Patterns



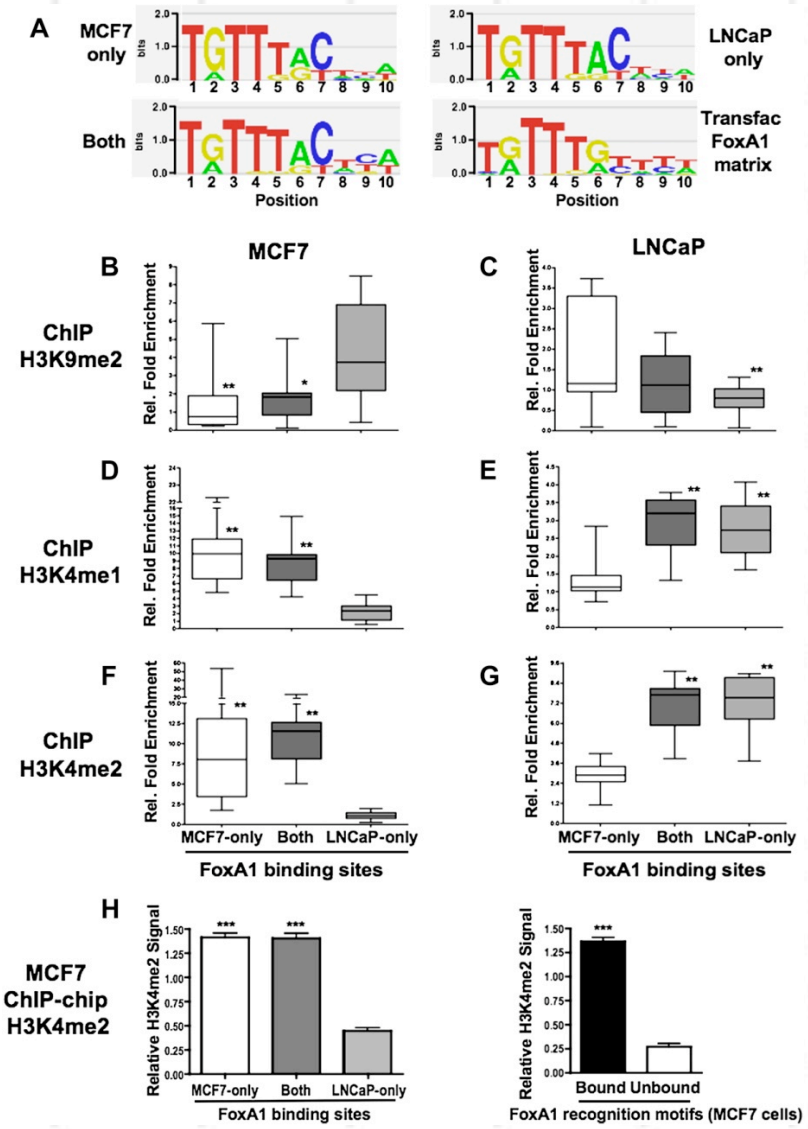
14% genome

Lupien et al. Cell 2008

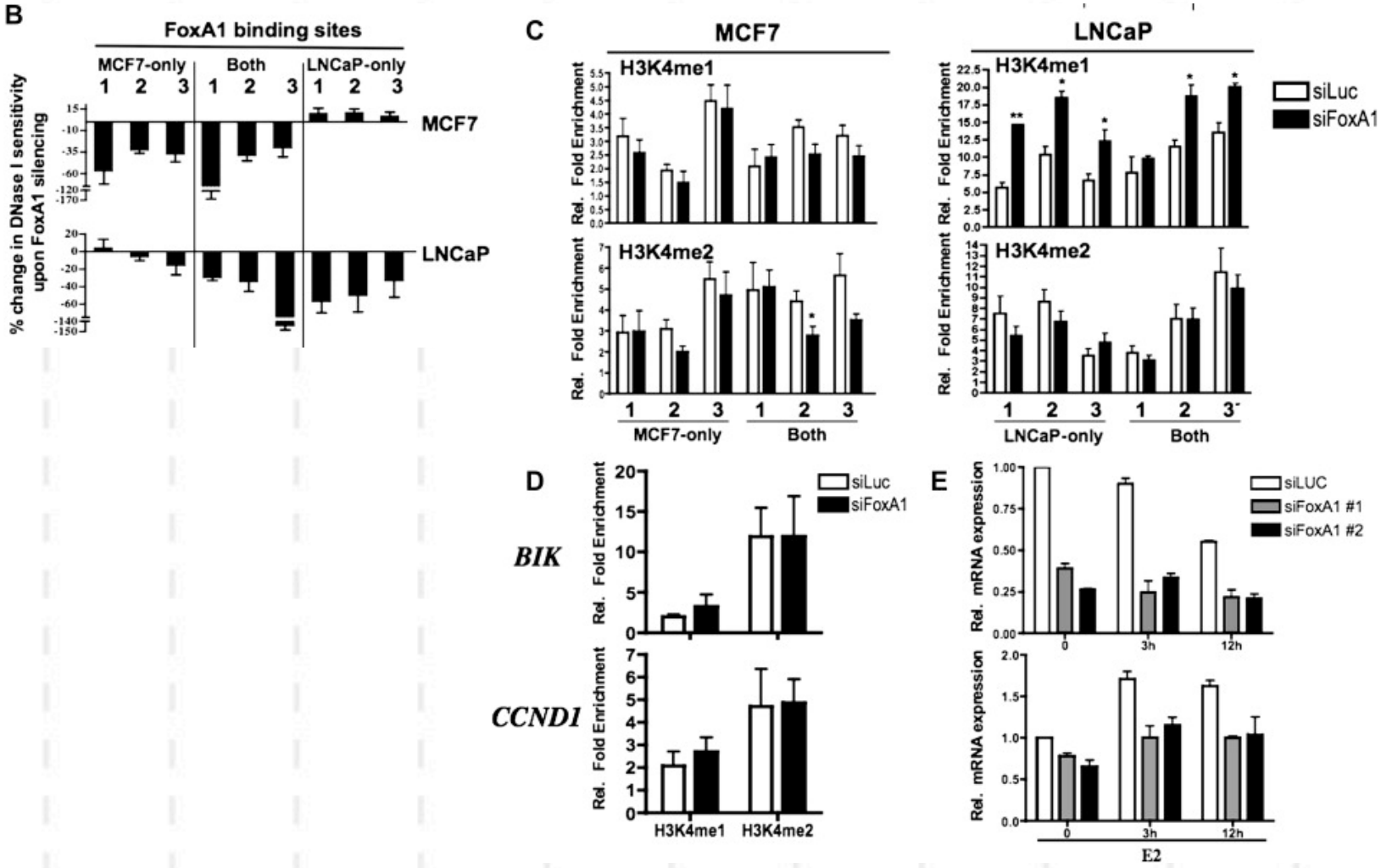
FoxA1 Cell Type-Specific Binding Sites Also Recruit Nuclear Receptors ERa or AR and Correlate with Regulation of Sex Steroid Signaling in Breast and Prostate Cancer Cells



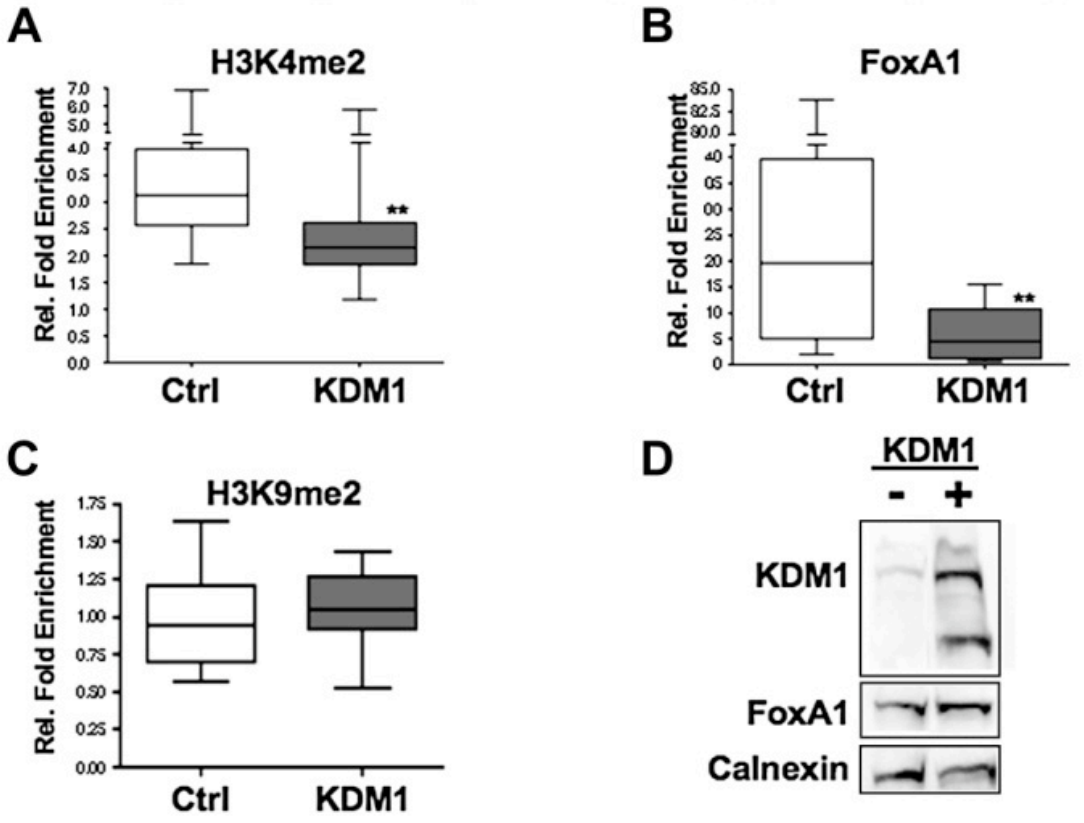
How is FoxA1 able to bind to distinct regions in the MCF7 and LNCaP cells?



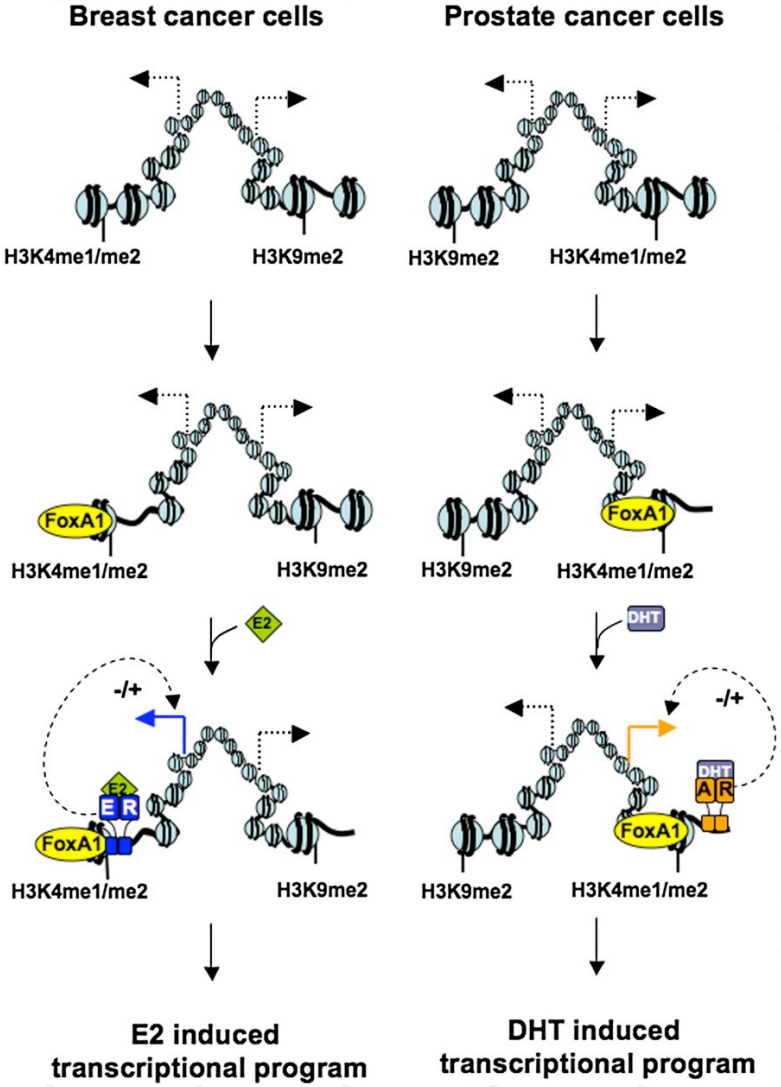
FoxA1 Silencing Decreases Chromatin Accessibility of Enhancers but Not H3K4 Methylation Levels



Reduction of H3K4 Methylation Impairs Cell Type-Specific FoxA1 Recruitment



FoxA1 translates epigenetic signatures into enhancer-driven cell type-specific transcription



Outline

- Introduction
- ChIP-chip with genome tiling arrays
- ChIP-seq with next-gen sequencing
- Estrogen Receptor and FoxA1 regulation in breast and prostate cancers

Acknowledgements

- X. Shirley Liu
- Myles Brown
- Hongkai Ji
- Qianben Wang