Introduction to genome tiling microarray analysis

Biological motivations

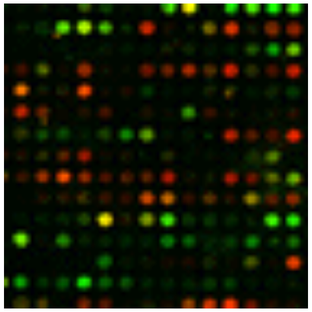
- There are many types of “events” happen at different locations on the genome. For example, protein bindings, epigenetic modifications (DNA methylation and histone modifications), copy number variations, etc.
- It is often of great interests to detect the genomic locations where a specific event happens, or quantify the events along the genome.
- The locations of these events provide explanations for many biological processes.

An example: transcription factor(TF) binding

- Transcription factors (TF): proteins that binds to specific DNA sequences and control the transcription from DNA to mRNA.
- There are many different types of TFs, each recognize different DNA sequences (motifs).
- The functions of the TFs are important for understanding gene regulatory mechanisms.
- The first step toward the understanding is to detect the TF binding sites (TFBS).

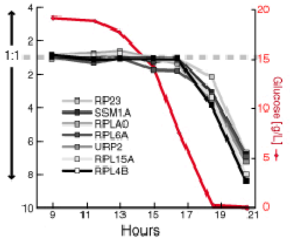
Traditional Method for Understanding Transcription Regulation

Gene expression microarray analysis



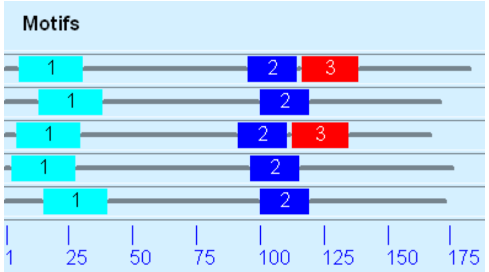
↓

Clustering genes by expression profile



↓

Search conserved sequence motifs in cluster promoters



Very challenging for mammalian genomes!

By Hongkai Ji at Hopkins

Use tiling array (ChIP-chip) to detect TFBS

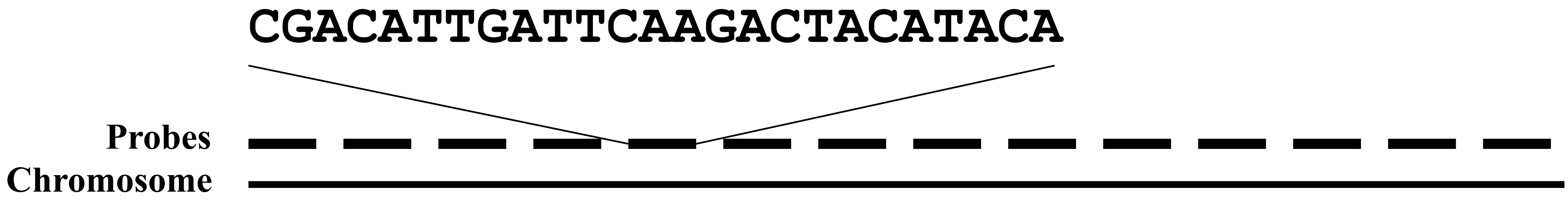
- Detect genome-wide *in vivo* location of TF and other DNA-binding proteins.
- Can learn the regulatory mechanism of a transcription factor or DNA-binding protein much better and faster.

Another example: DNA methylation and histone modification

- DNA methylation and histone modifications are chemical modification of DNA molecule.
- The strengths of such modifications varies along the genome, and they are related to gene regulatory and many diseases.
- The methylation or modification strengths can be measured using ChIP-chip or MeDIP-chip.

Tiling arrays

- The goal is to quantify the events of interests along the genomes, and/or detect the genomic coordinates for the events.
- Work the same as gene expression array (hybridization based), except that the probes are designed to tile up the genome at non-repeat regions.
- Data for probes in the location of interest often behave differently from backgrounds (e.g., bigger intensities).



Types of tiling arrays

- ChIP-chip: Chromatin ImmunoPrecipitation (ChIP) + tiling array (chip) for detecting transcription factor binding sites or measuring histone modification levels.
- MeDIP-chip: Methyl-DNA ImmunoPrecipitation (MeDIP) + tiling array (chip) for measuring DNA methylation level.
- ArrayCGH (Comparative Genomic Hybridization) for detecting copy number variations.

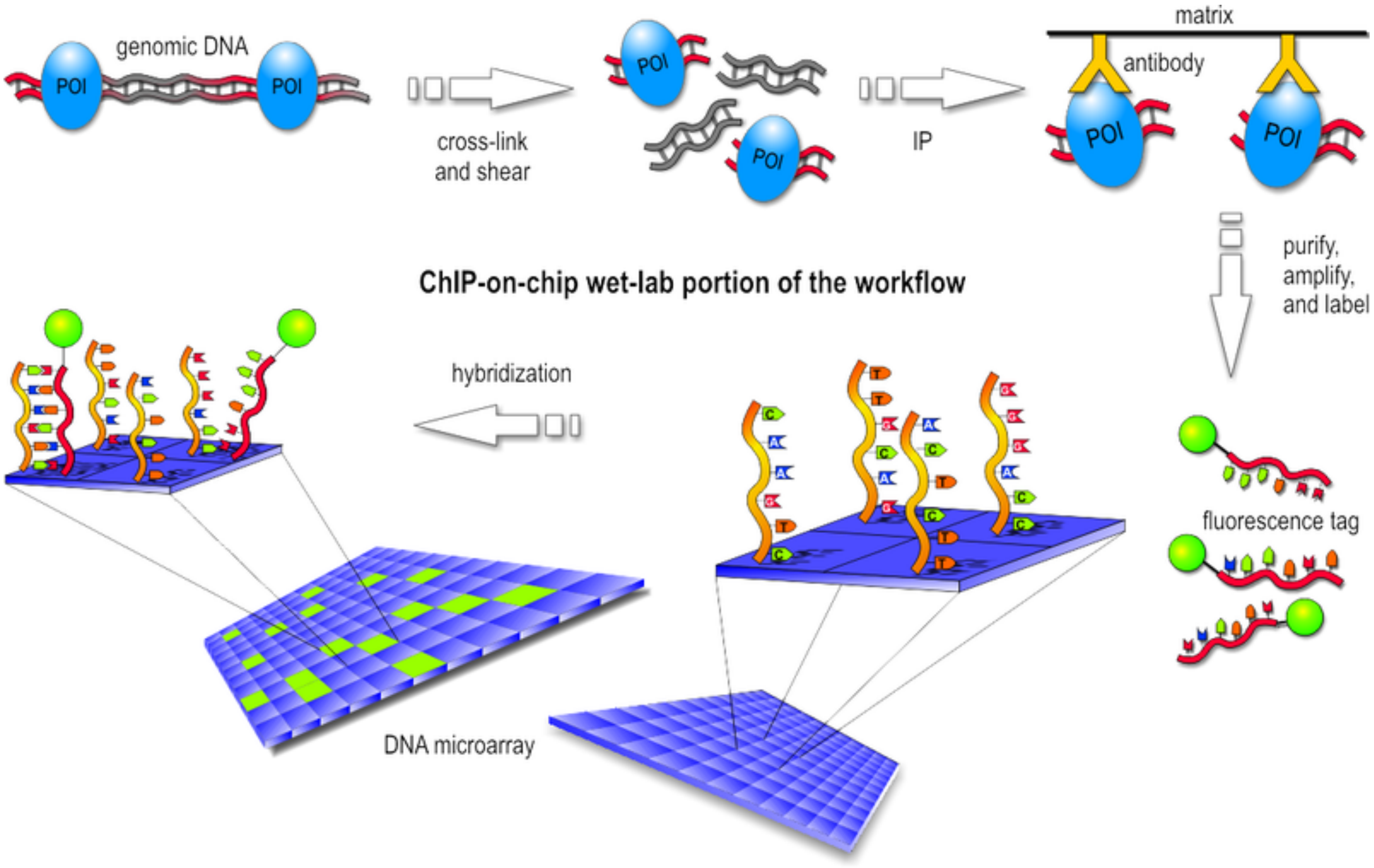
There's no major differences in array designs. Difference are the ways to prepare biological samples.

Available platforms for ChIP-chip

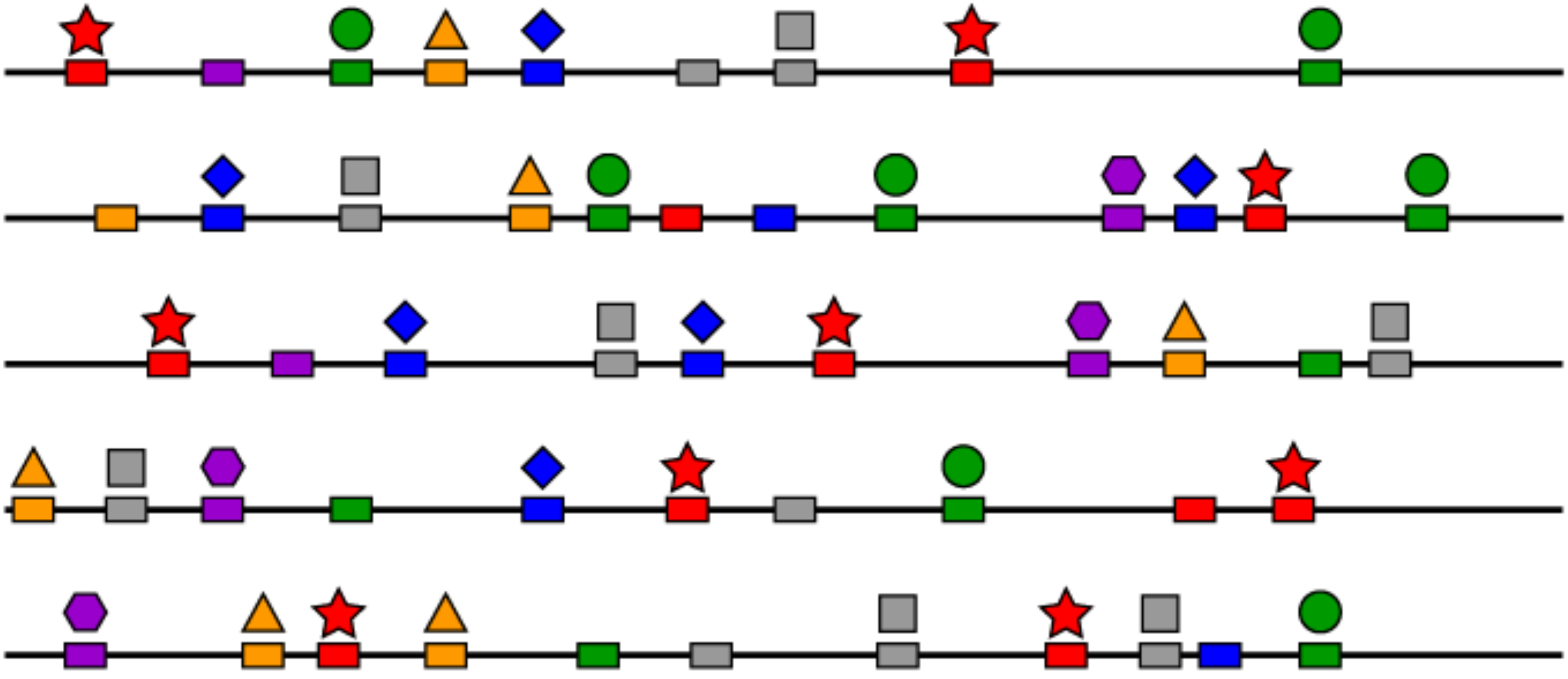
	# Arrays human genome	# Probes / Array	# Total Probes	Probe Length	Probe Resolution
Affymetrix	7	6M	42.0M	25mer	35 bp
Nimblegen	10	2.1M	21M	50mer	100 bp
Agilent	21	244K	5.1M	60mer	300 bp in genes; 500 bp in intergenic

By Xiaole Shirley Liu at Harvard

ChIP-chip procedures

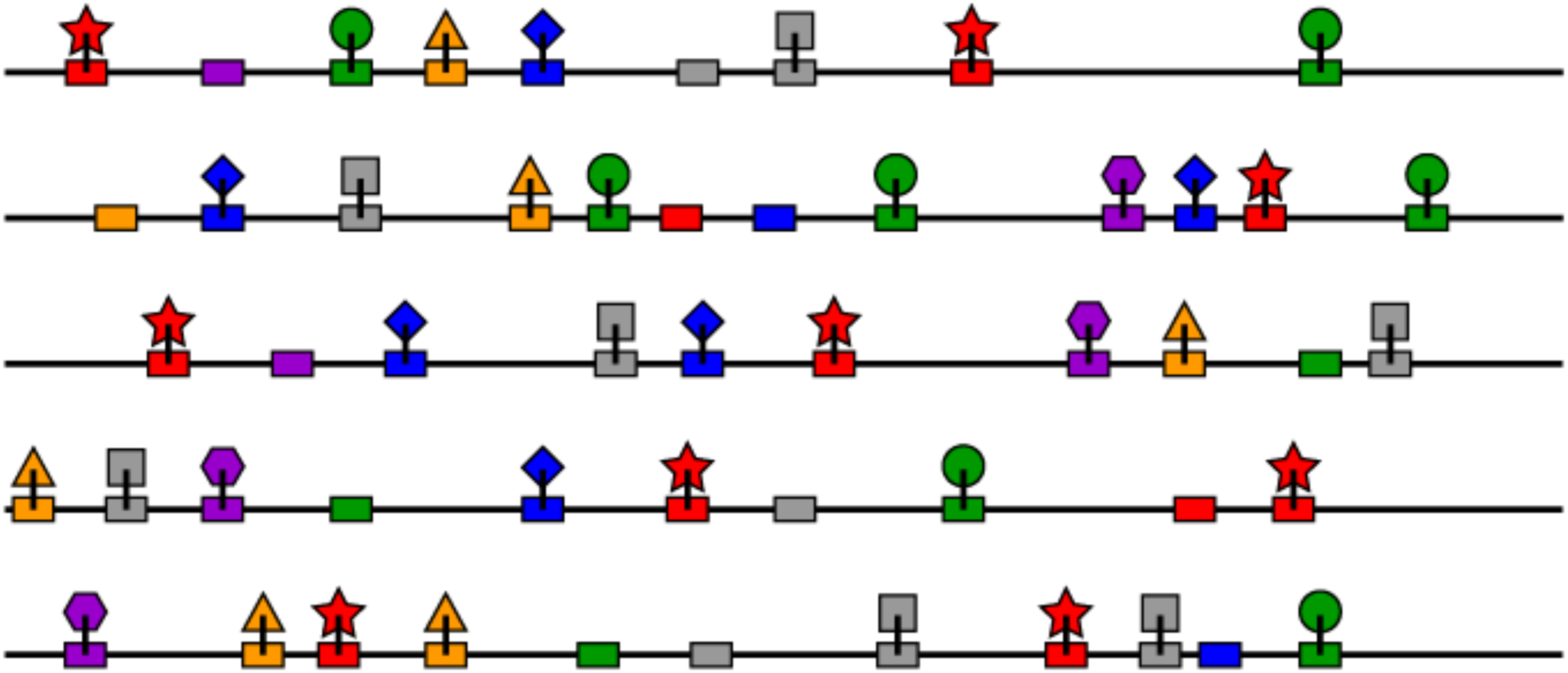


Chromatin ImmunoPrecipitation (ChIP)



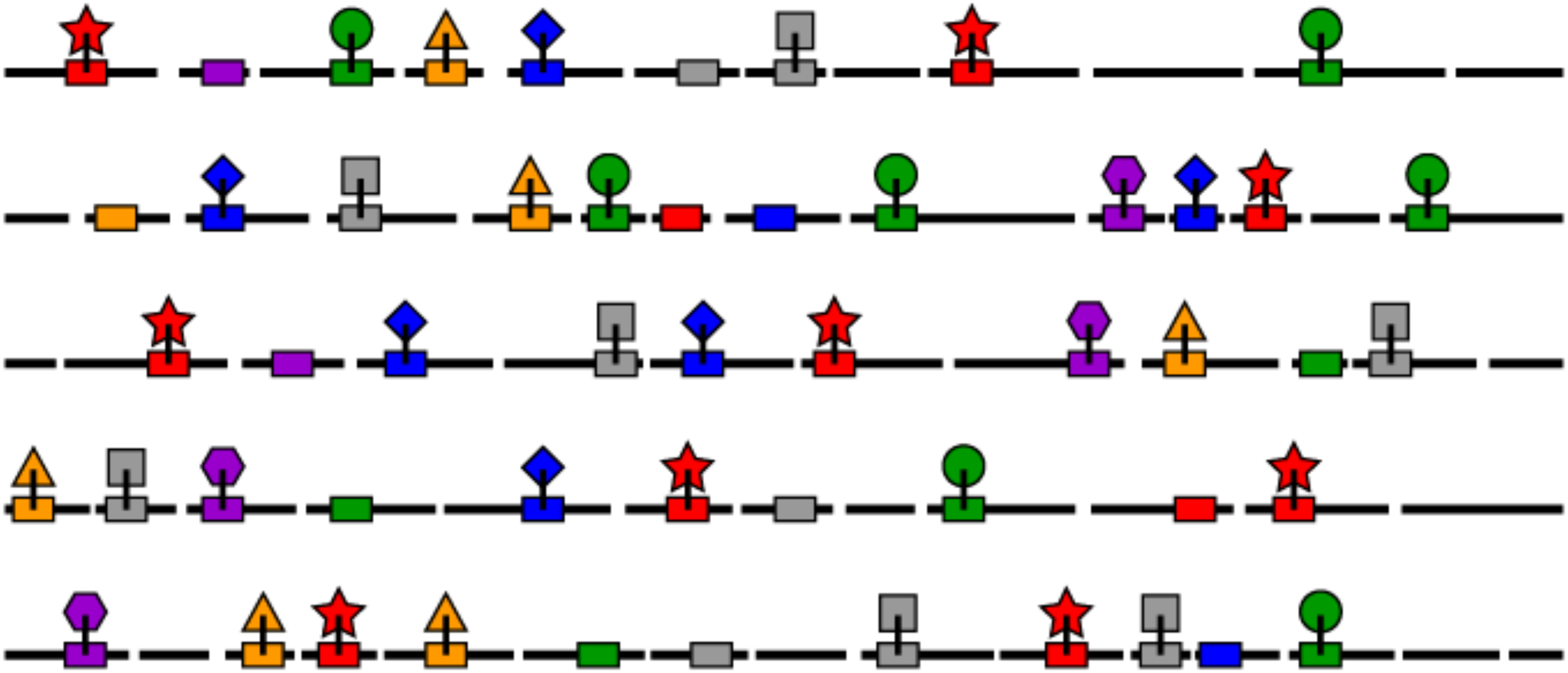
By Richard Bourgon at UC Berkley

TF/DNA Crosslinking *in vivo*



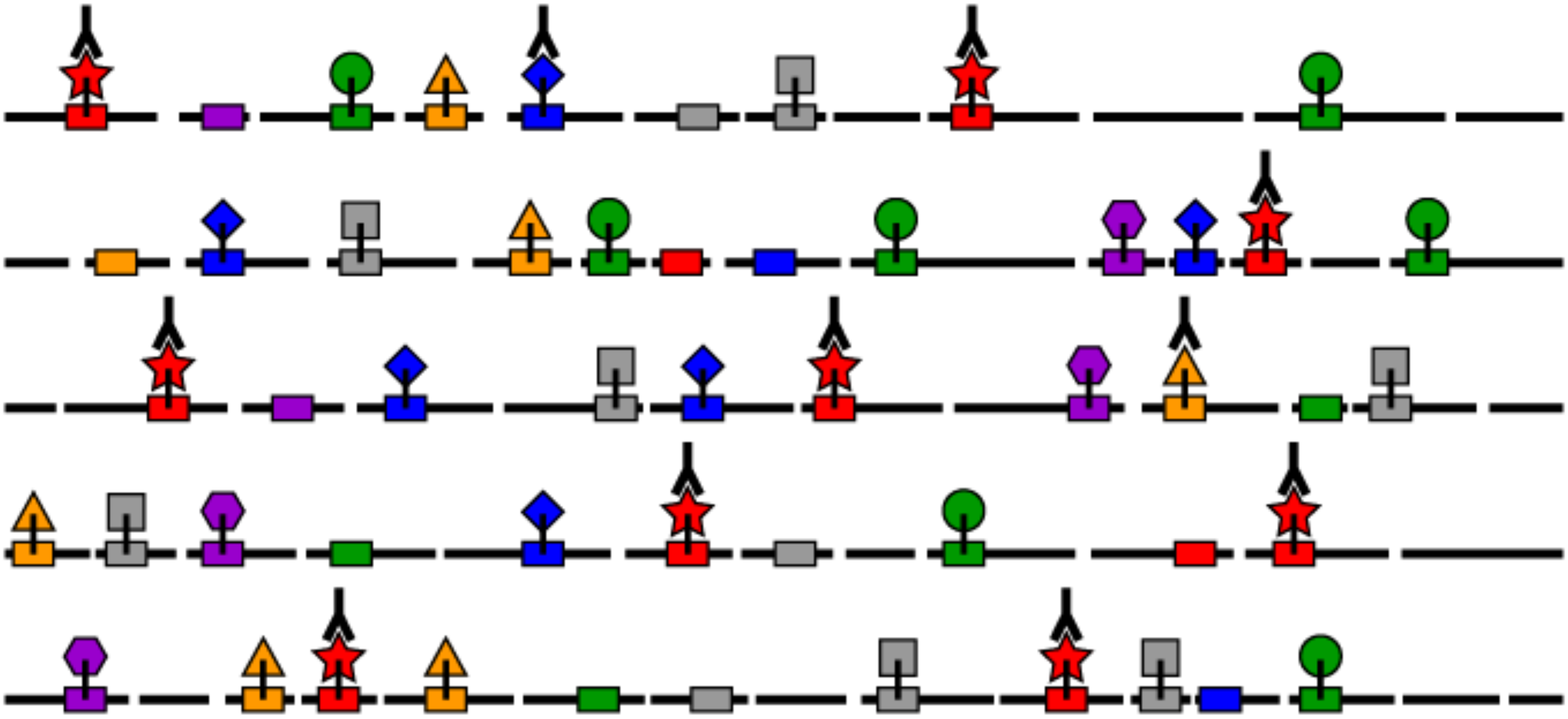
By Richard Bourgon at UC Berkley

Sonication (~500bp)



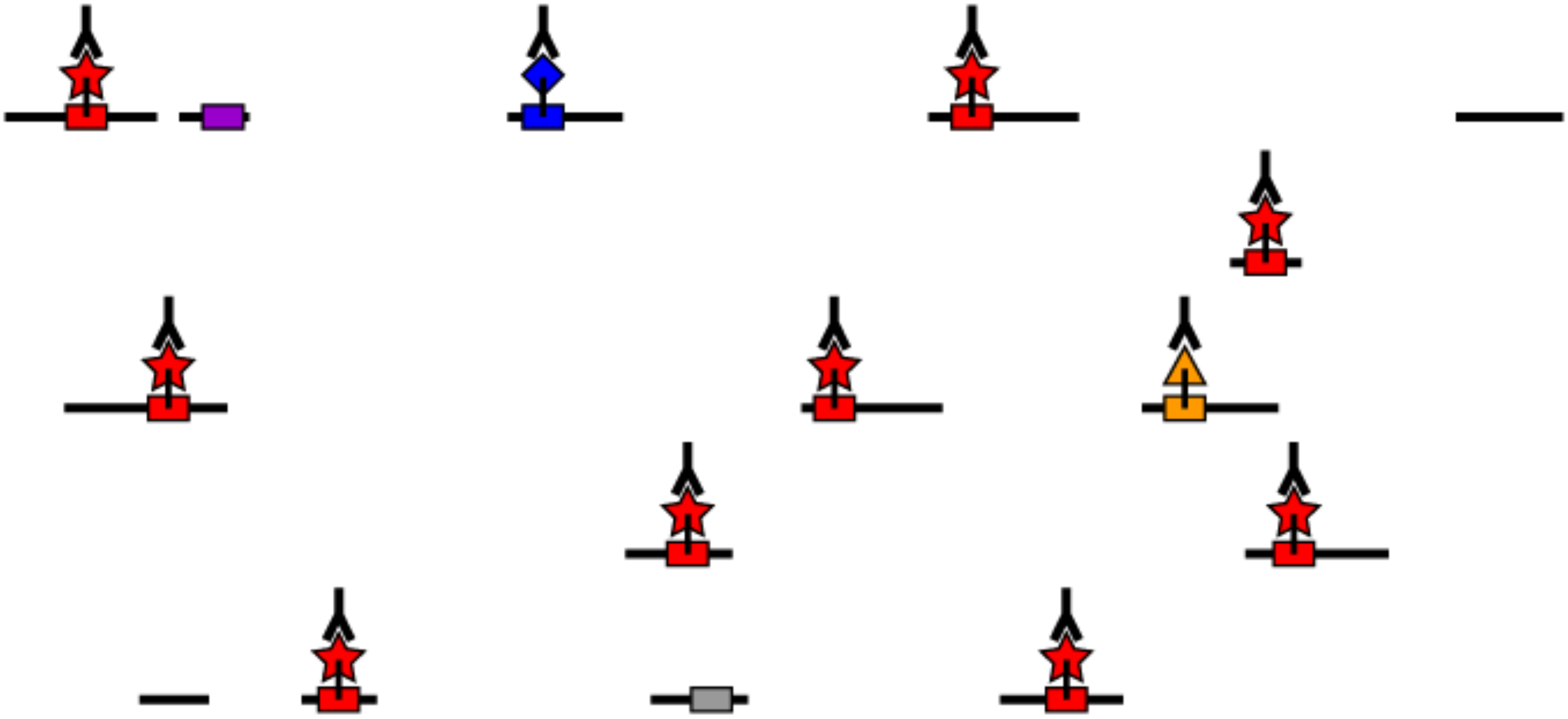
By Richard Bourgon at UC Berkley

TF-specific Antibody



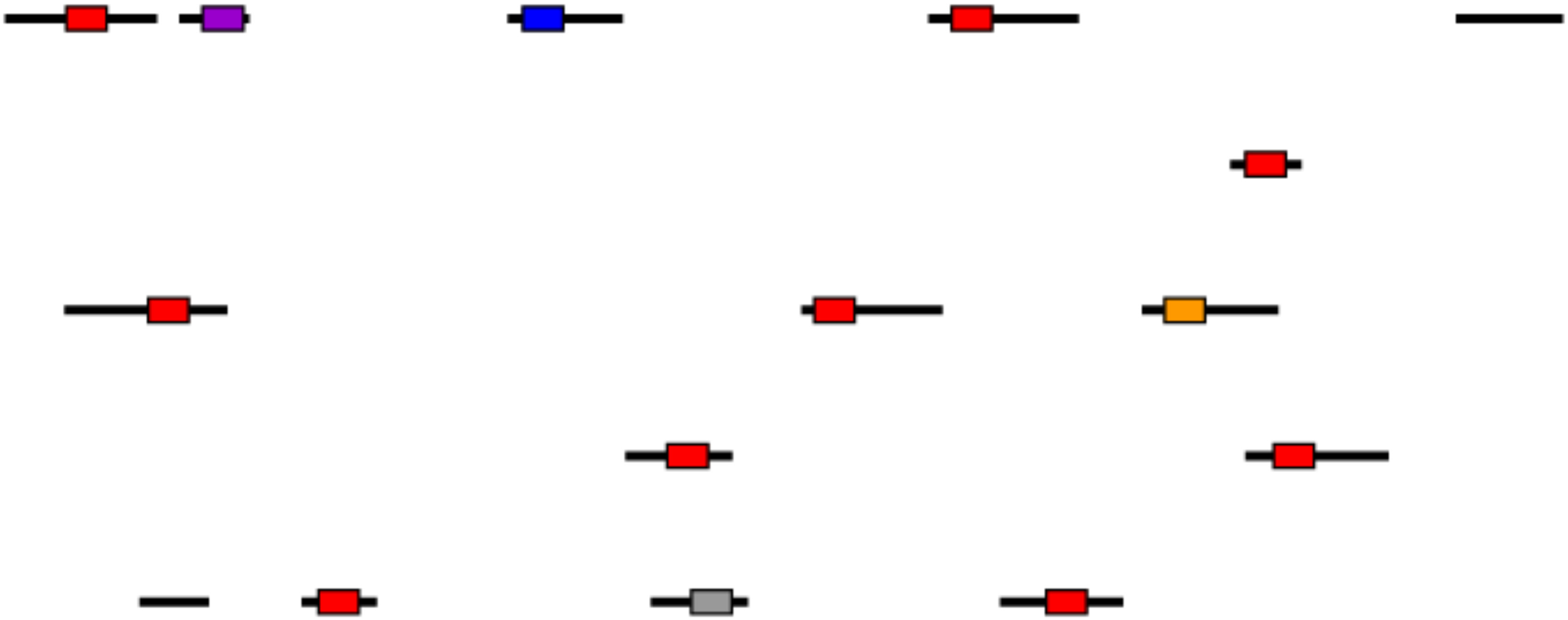
By Richard Bourgon at UC Berkley

Immunoprecipitation (IP)



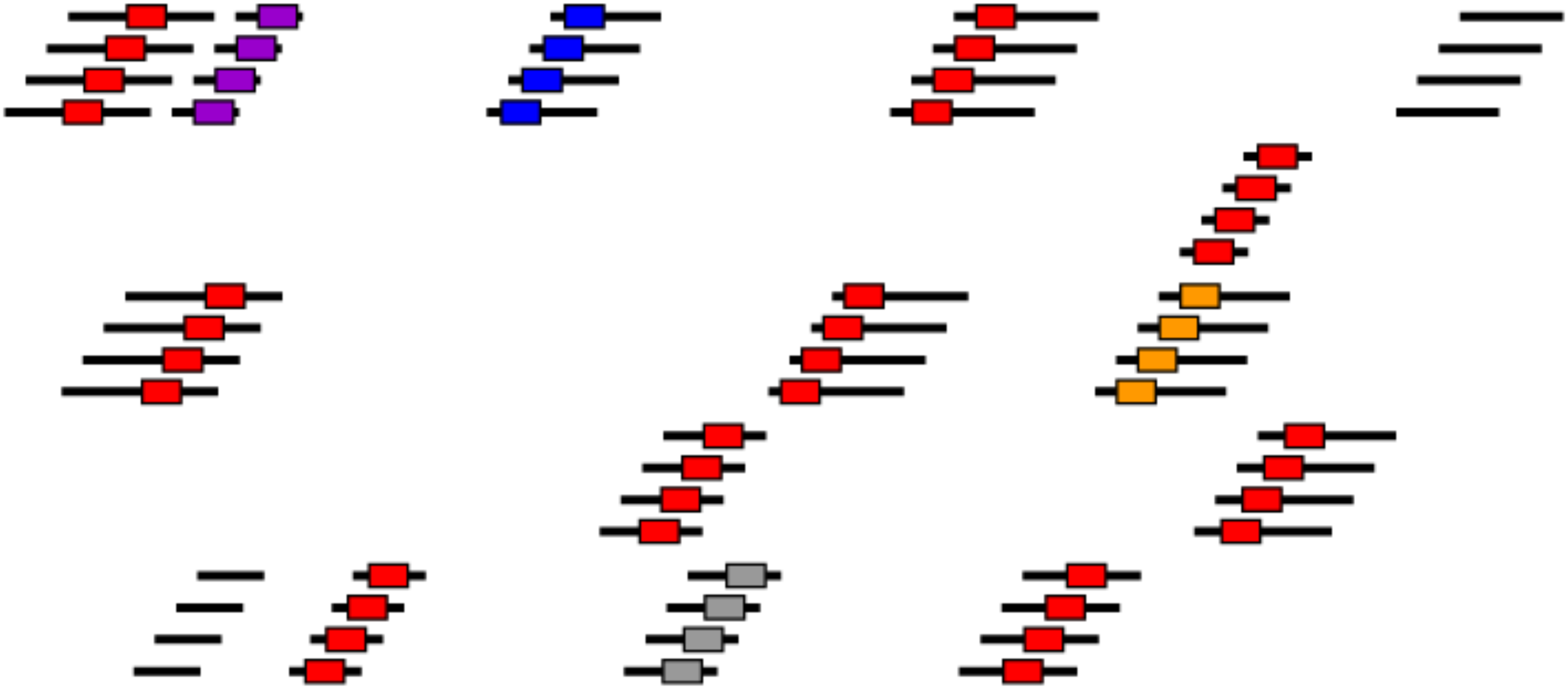
By Richard Bourgon at UC Berkley

Reverse Crosslink and DNA Purification



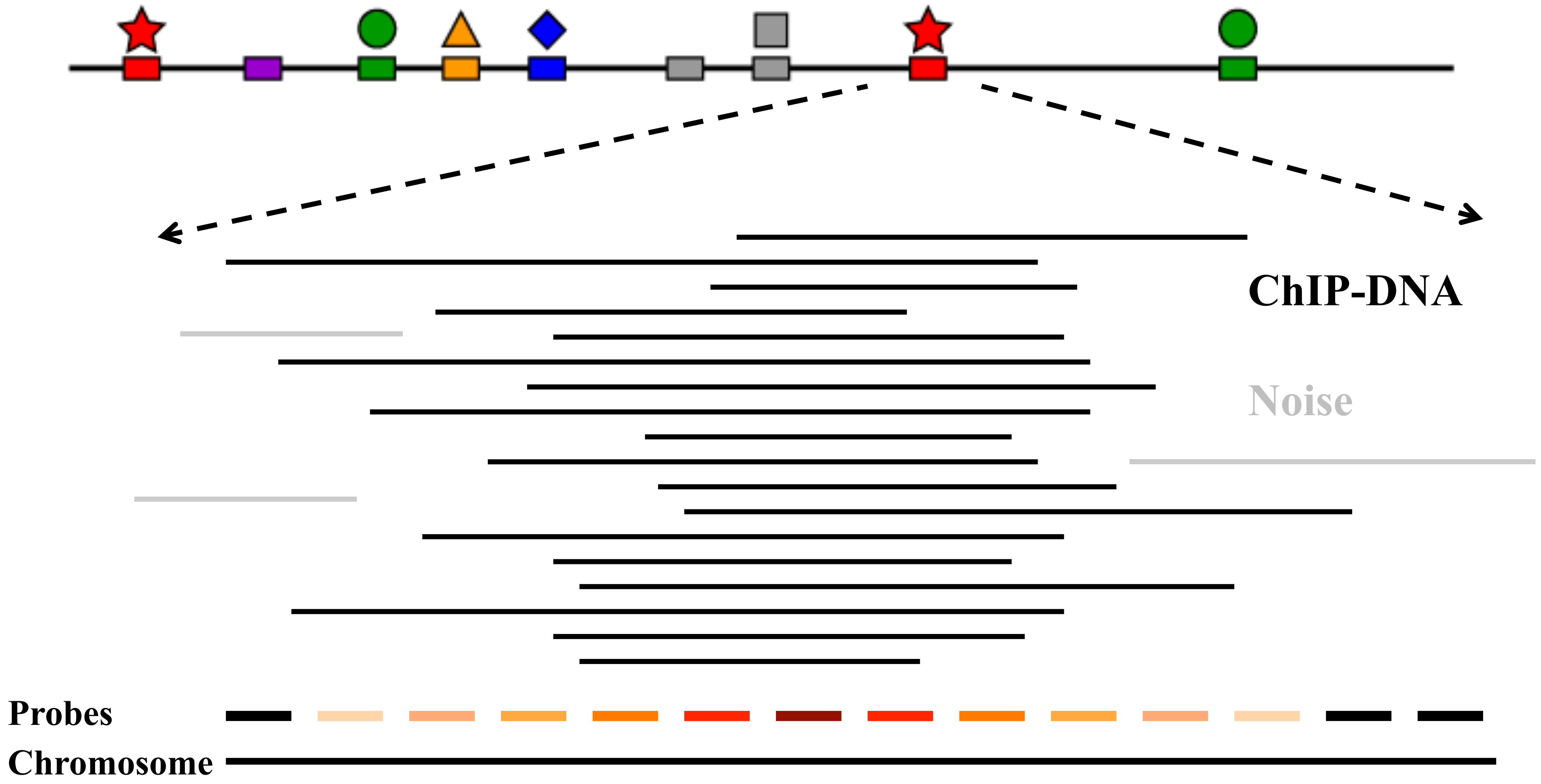
By Richard Bourgon at UC Berkley

Amplification



By Richard Bourgon at UC Berkley

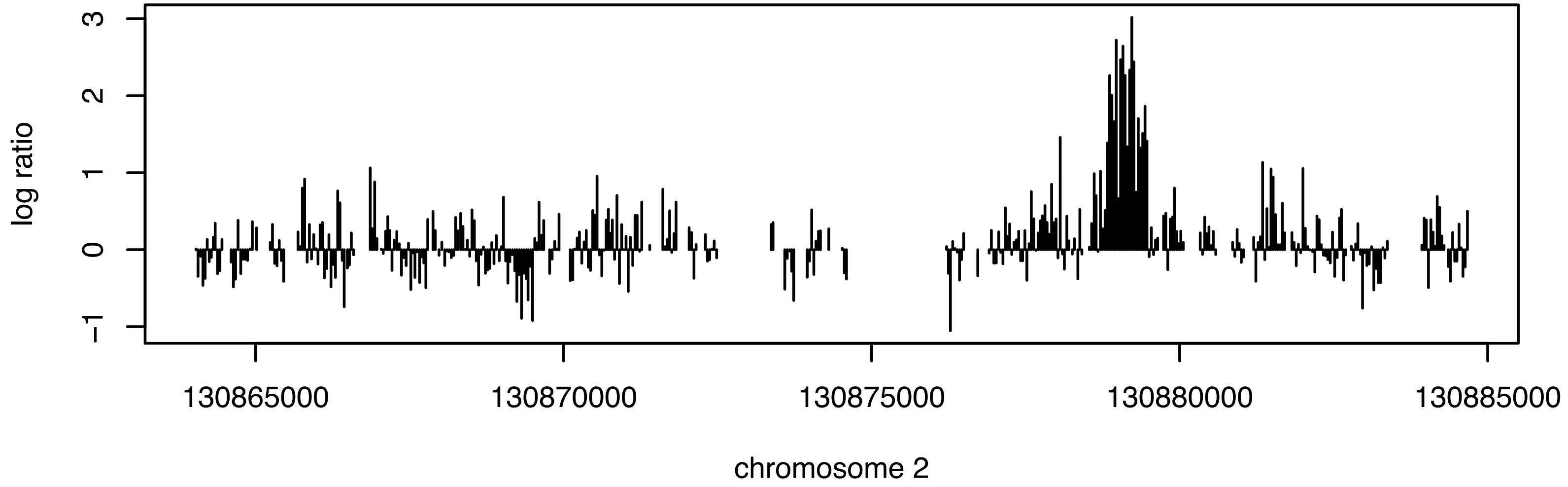
ChIP-chip Hybridization



Based on Xiaole Shirley Liu at Harvard

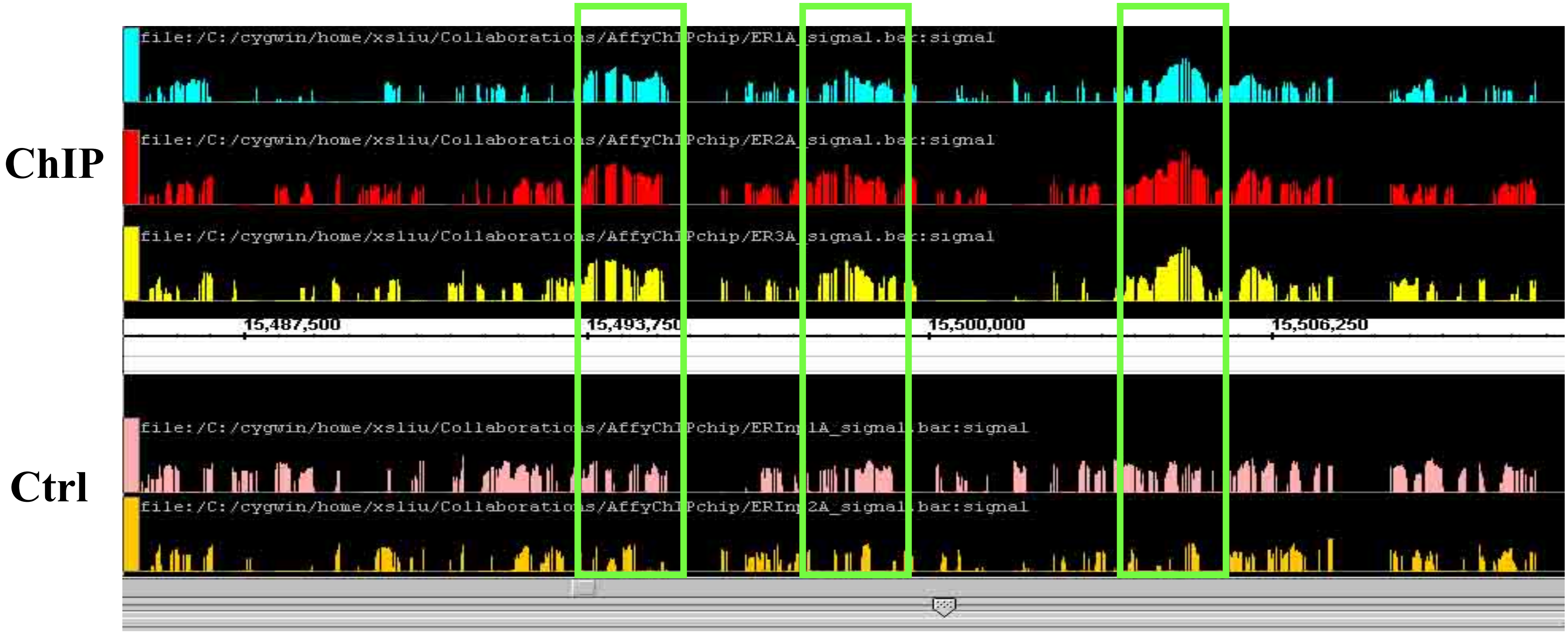
Data from ChIP-chip

- Can be thought as a file with millions of rows and three columns.
  - Each row is for a probe.
  - Columns are chromosome number, probe location on the genome, and signal (intensity values or log fold change).
- To visualize: plot the probe signals against probe locations.



Identify ChIP-enriched Region

- Controls: sonicated genomic input DNA (non-treated).
- Often 3 ChIP, 3 Ctrl replicates are needed


ChIP
file:/C:/cygwin/home/xsliu/Collaborations/AffyChIPchip/ER1A_signal.bar:signal
file:/C:/cygwin/home/xsliu/Collaborations/AffyChIPchip/ER2A_signal.bar:signal
file:/C:/cygwin/home/xsliu/Collaborations/AffyChIPchip/ER3A_signal.bar:signal
15,487,500
15,493,750
15,500,000
15,506,250
Ctrl
file:/C:/cygwin/home/xsliu/Collaborations/AffyChIPchip/ERInp1A_signal.bar:signal
file:/C:/cygwin/home/xsliu/Collaborations/AffyChIPchip/ERInp2A_signal.bar:signal

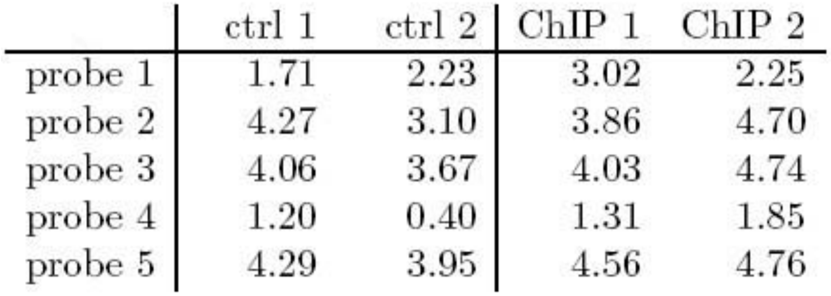
By Xiaole Shirley Liu at Harvard

ChIP-chip data analysis

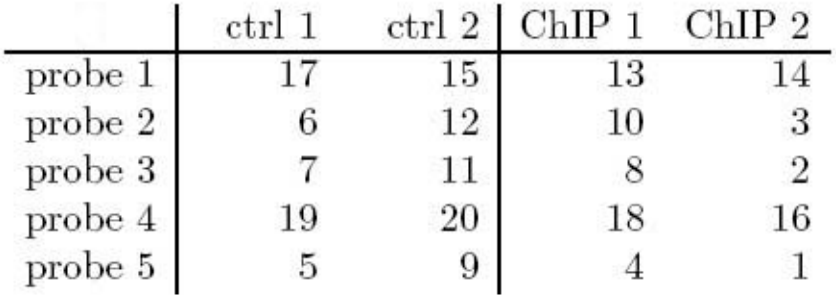
- Goal: detect locations of interests (e.g., binding sites, also called “peaks”) based on probe locations and signals.
- Normalization: remove technical artifacts.
- Detection for regions of interests:
  - Many different methods. Fundamentally data from neighboring probes need to be combined to make inference, because the regions of interests often overlap many probes.
  - Easiest method: moving average, then use an arbitrary cutoff.

Mann-Whitney U-test

- Affy TAS, Cawley et al (*Cell* 2004):
  - Each probe: rank probes signals within [-500bp, +500bp] window.
  - 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

By Xiaole Shirley Liu at Harvard

TileMap (Ji and Wong, Bioinformatics 2005)

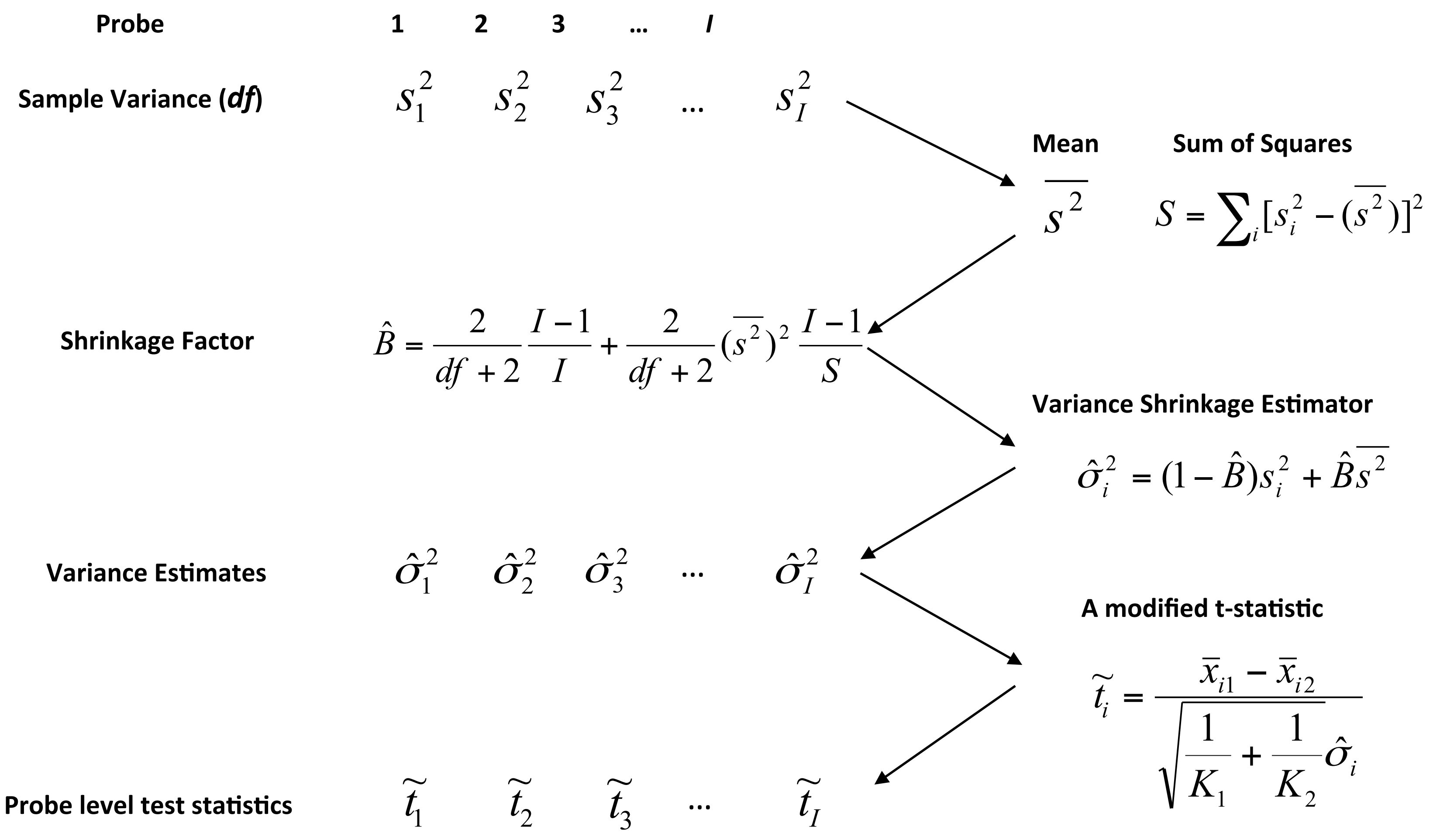
STEP 1:
Compute a test statistic for each probe to summarize probe level information

↓

STEP 2:
Combine probe level test statistics of neighboring probes to help infer binding regions

By Hongkai Ji at Hopkins

Probe level test statistic: empirical Bayes approach



Probe **1** **2** **3** **…** ***I***

Sample Variance (*df*) s_1^2 s_2^2 s_3^2 … s_I^2

Mean

$$\overline{s^2}$$

Sum of Squares

$$S = \sum_i [s_i^2 - (\overline{s^2})]^2$$

Shrinkage Factor

$$\hat{B} = \frac{2}{df+2}\frac{I-1}{I} + \frac{2}{df+2}(\overline{s^2})^2\frac{I-1}{S}$$

Variance Shrinkage Estimator

$$\hat{\sigma}_i^2 = (1-\hat{B})s_i^2 + \hat{B}\overline{s^2}$$

Variance Estimates $\hat{\sigma}_1^2$ $\hat{\sigma}_2^2$ $\hat{\sigma}_3^2$ … $\hat{\sigma}_I^2$

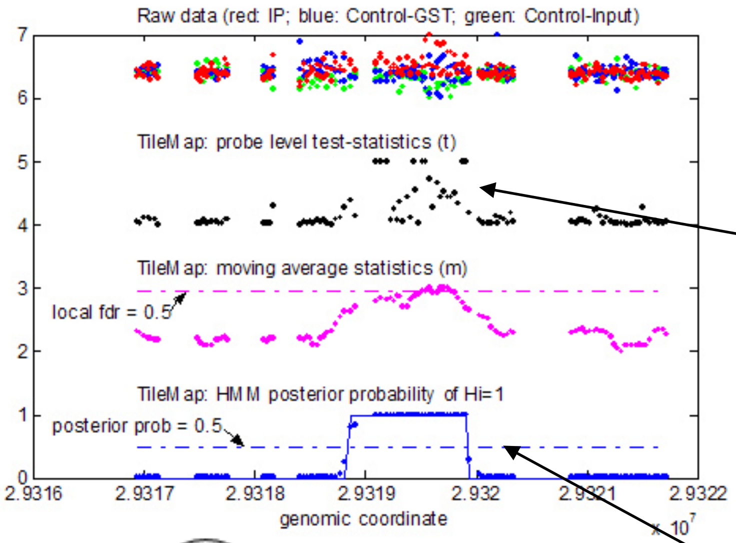
A modified t-statistic

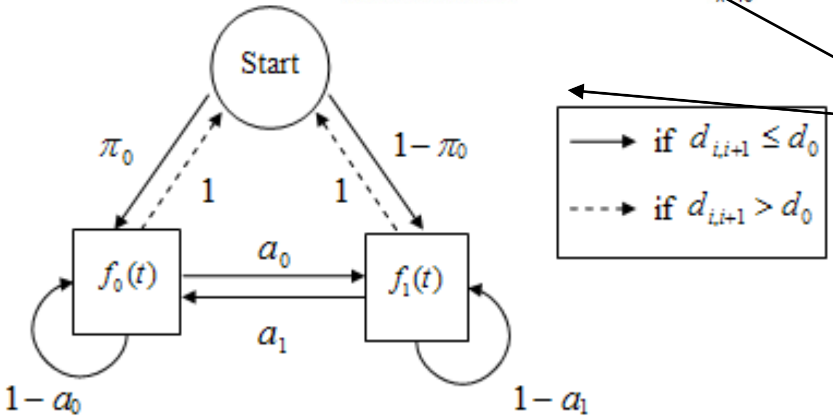
$$\tilde{t}_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{\sqrt{\frac{1}{K_1} + \frac{1}{K_2}}\hat{\sigma}_i}$$

Probe level test statistics $\tilde{t}_1$ $\tilde{t}_2$ $\tilde{t}_3$ … $\tilde{t}_I$

By Hongkai Ji at Hopkins

Combining neighboring probes





TileMap (MA)

1. Compute the probe level test statistic *t* for each probe;
2. Compute a moving average statistic to measure enrichment;
3. Estimate FDR.

TileMap (HMM)

1. Compute the probe level test statistic *t* for each probe;
2. Estimate the distribution of *t* under H_0 and H_1;
3. Model *t* by a Hidden Markov Model, and decode the HMM.

By Hongkai Ji at Hopkins

TileMap summary

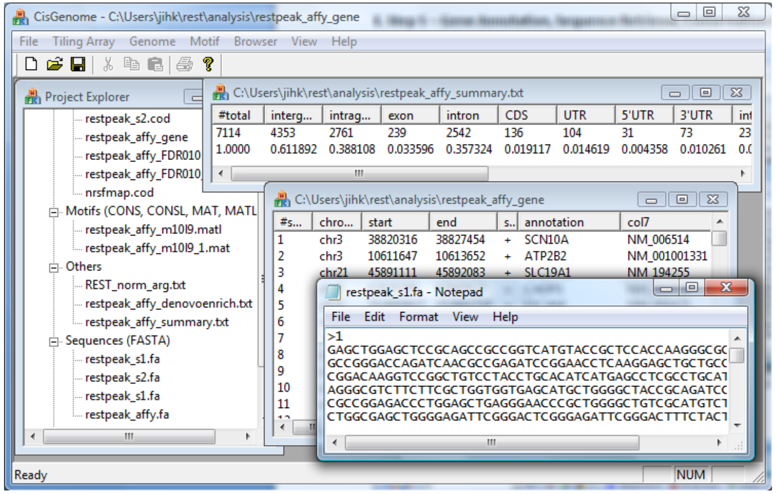
- Now a part of a software suite CisGenome.
- Windows based GUI.
- Command line version available for Mac and Linux.
- Freely available from:

  `http://www.biostat.jhsph.edu/~hji/cisgenome/`

CisGenome

(Ji H. et al. Nature Biotechnology, 2008)

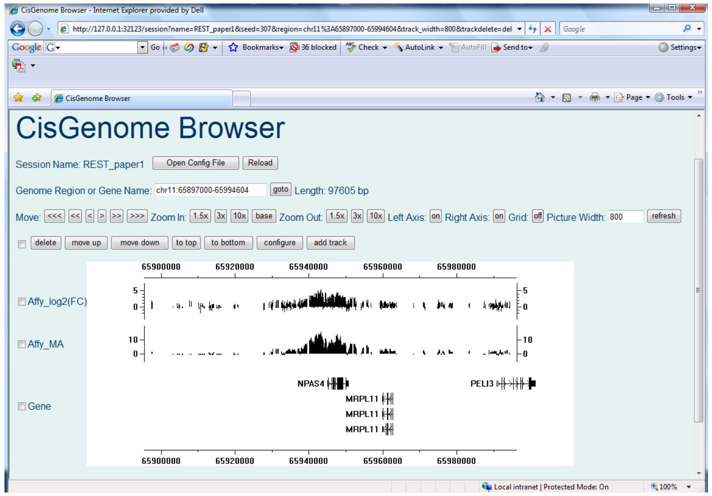
Graphic User Interface



#total	interg...	intrag...	exon	intron	CDS	UTR	5'UTR	3'UTR	int
7114	4353	2761	239	2542	136	104	31	73	23
1.0000	0.611892	0.388108	0.033596	0.357324	0.019117	0.014619	0.004358	0.010261	0.(

#s...	chro...	start	end	s..	annotation	col7
1	chr3	38820316	38827454	+	SCN10A	NM_006514
2	chr3	10611647	10613652	+	ATP2B2	NM_001001331
3	chr21	45891111	45892083	+	SLC19A1	NM 194255
4						
5						
6						
7						
8						
9						
10						
11						

```
>1
GAGCTGGAGCTCCGCAGCCGCCGGTCATGTACCGCTCCACCAAGGGCGC
GCCGGGACCAGATCAACGCCGAGATCCGGAACCTCAAGGAGCTGCTGCC
CGGACAAGGTCCGGCTGTCCTACCTGCACATCATGAGCCTCGCCTGCAT
AGGGCGTCTTCTTCGCTGGTGGTGAGCATGCTGGGGCTACCGCAGATCC
CGCCGGAGACCCTGGAGCTGAGGGAACCCGCTGGGGCTGTCGCATGTCT
CTGGCGAGCTGGGGAGATTCGGGACTCGGGAGATTCGGGACTTTCTACT
```



http://127.0.0.1:32123/session?name=REST_paper1&seed=307&region=chr11%3A65897000-65994604&track_width=800&trackdelete=del

CisGenome Browser

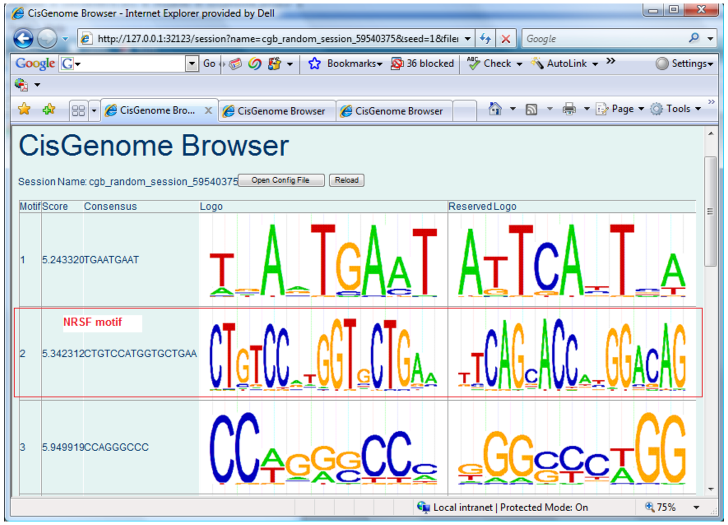
Session Name: REST_paper1

Genome Region or Gene Name: chr11:65897000-65994604 Length: 97605 bp

Move: Zoom In: Zoom Out: Left Axis: on Right Axis: on Grid: off Picture Width: 800

Core Data Analysis Programs

CisGenome Browser



http://127.0.0.1:32123/session?name=cgb_random_session_59540375&seed=1&file

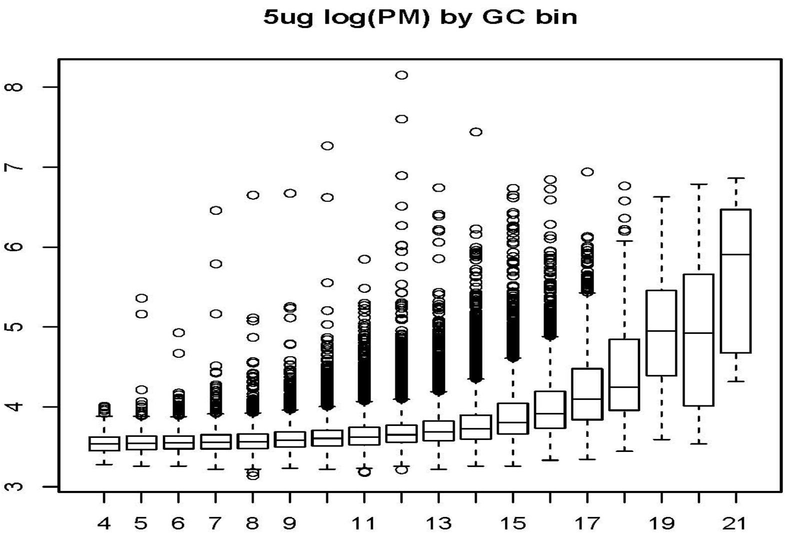
CisGenome Browser

Session Name: cgb_random_session_59540375

Motif	Score	Consensus	Logo	Reserved Logo
1	5.243320	TGAATGAAT		
2	5.342312	CTGTCCATGGTGCTGAA		
3	5.949919	CCAGGGCCC		

MAT: Model-based Analysis of Tiling arrays (Johnson W.E. et al. *PNAS*, 2006)

- Estimate probe behavior by checking other probes with similar sequence on the same array
- Probe sequence plays a big role in signal value.
- Most of the probes in ChIP-chip measures non-specific hybridization.



By Xiaole Shirley Liu at Harvard

Probe Behavior Model

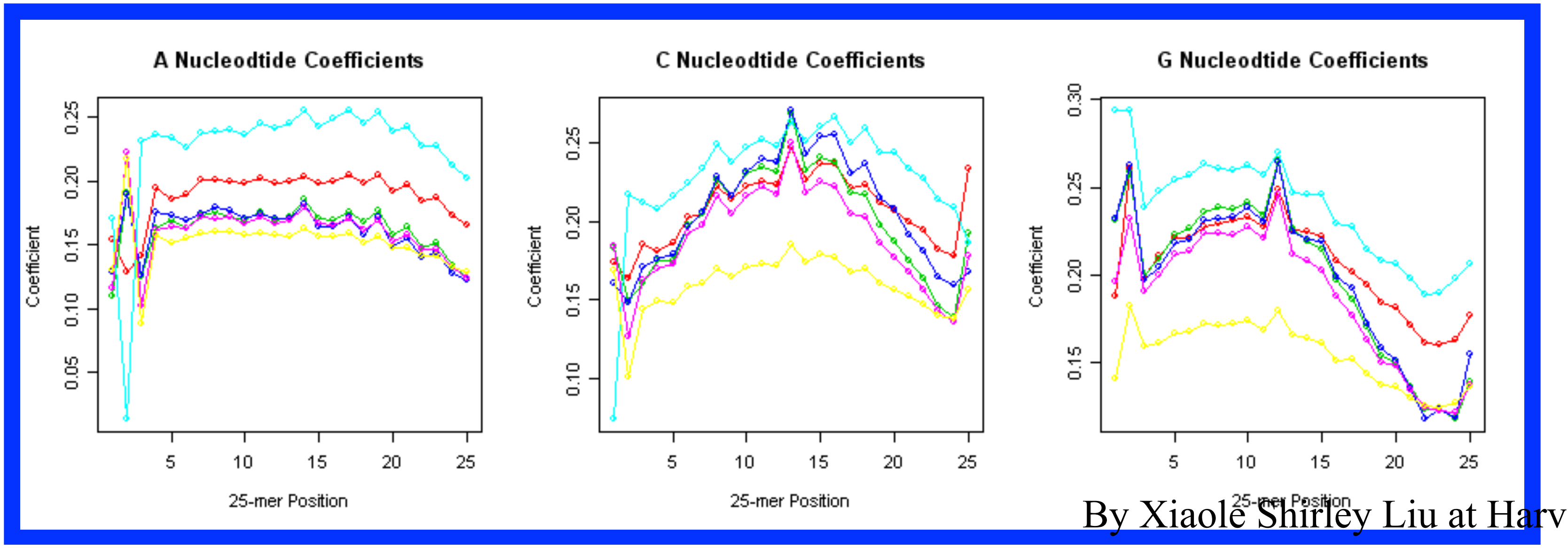
$$Log(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k=A,C,G} \beta_{jk} I_{ijk} + \sum_{l=A,C,G,T} \gamma_l n_{il}^2 + \delta Log(c_i) + \varepsilon_i$$

Baseline on number of Ts

A,C,G at each position of the 25mer

A,C,G,T Count Square

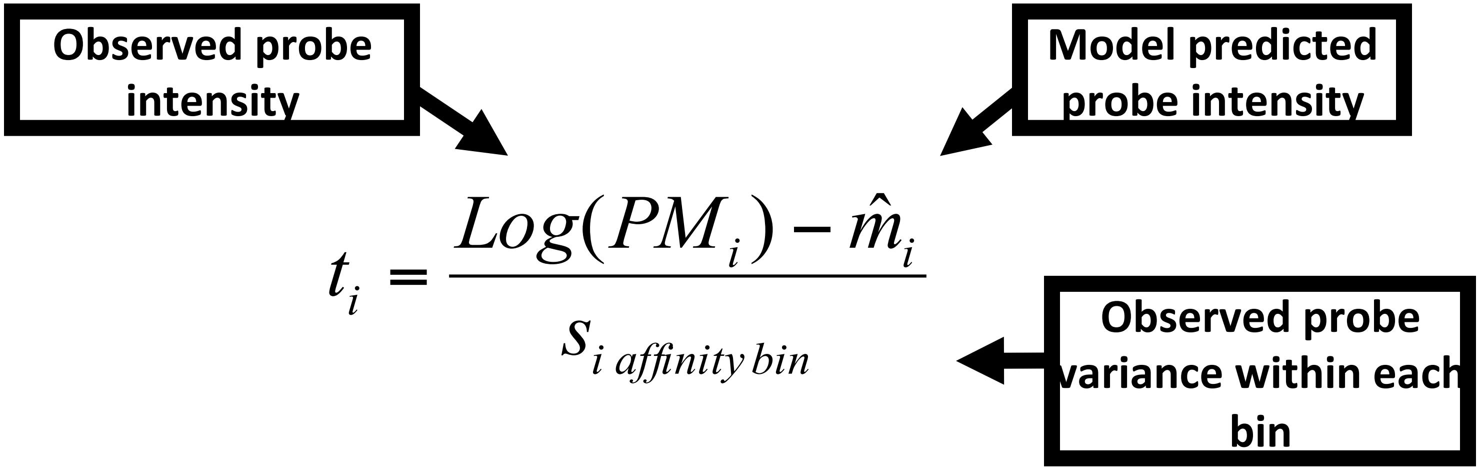
25mer Copy Number along the Genome



By Xiaole Shirley Liu at Harvard

Probe Standardization

- Fit the probe model array by array
- Divide array probes to bins (3k probes/bin)
- Background-subtraction and standardization (normalization) on a single array;



Observed probe intensity

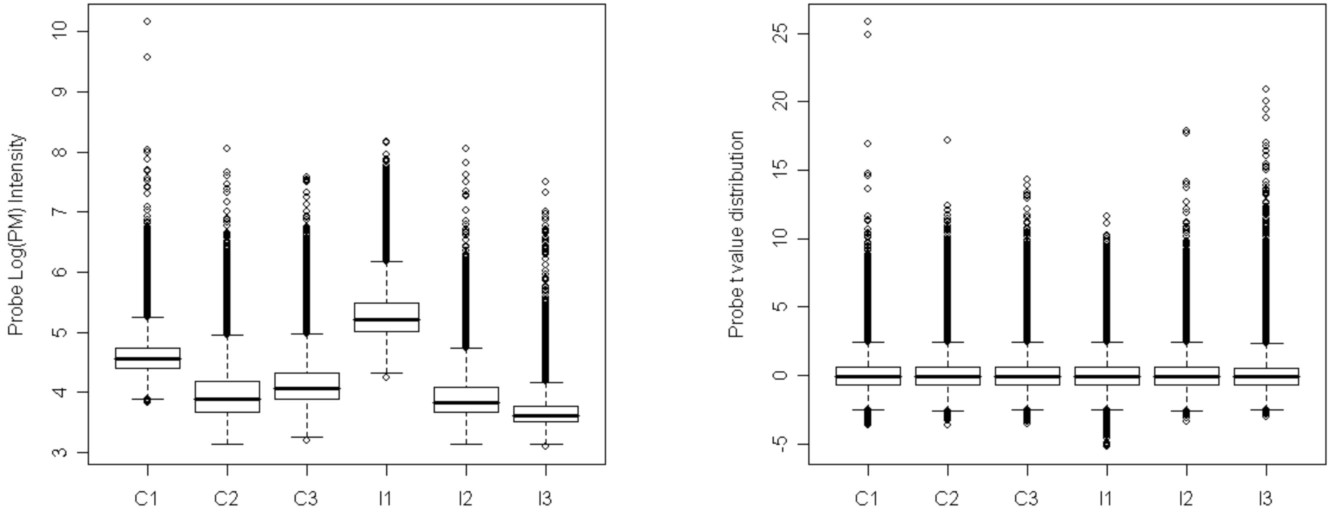
Model predicted probe intensity

$$t_i = \frac{Log(PM_i) - \hat{m}_i}{s_{i\ affinity\ bin}}$$

Observed probe variance within each bin

By Xiaole Shirley Liu at Harvard

- Probe signals before and after standardization



By Xiaole Shirley Liu at Harvard

Binding region detection

- Window-based MATscore
  - ChIP without Ctrl

$$MAT(region) = TM(t's\ in\ region)\sqrt{n_{probe}}$$

  - TM: trimmed mean
  - Multiple ChIP with multiple Ctrl

$$MAT(region) = \frac{TM(t's\ in\ ChIP) - TM(t's\ in\ Input)}{\sigma_{Input}}\sqrt{n_{probe}}$$

  - More probes, higher t values in ChIP, less variance (fluctuation) → more confident

By Xiaole Shirley Liu at Harvard

To use MAT

- Create a text configuration file (config.txt):

```
[data]
BpmapFolder = /home/bst/student/hwu/Project/Ji/MVHMM/DREAM/rawdata/
CelFolder = /home/bst/student/hwu/Project/Ji/MVHMM/DREAM/rawdata
GenomeGrp =Hs
Group = 111000
[bpmap]
1=Hs_PromPR_v02-3_NCBIv36.bpmap
[cel]
1=IP1.CEL IP2.CEL IP3.CEL CT1.CEL CT2.CEL CT3.CEL
[intensity analysis]
BandWidth =     300
MaxGap =     300
MinProbe  =     10
[interval analysis]
Pvalue = 1e-3
```

- Then run "`MAT config.txt`" at command line.

MAT summary

- Open source, written in python at http://chip.dfci.harvard.edu/~wli/MAT/
- Installation could be tricky.
- Good computational performance.
- Can work with single ChIP, multiple ChIP, and multiple ChIP with controls with increasing accuracy.

Bioconductor packages for analyzing ChIP-chip data

- Most of the ChIP-chip analysis are done using MAT or CisGenome, so there are relatively fewer R packages.
- Useful ones:
  - rMAT: R implementation of MAT model. Works for Affy ChIP-chip.
  - Ringo (R Investigation of NimbleGen Oligoarrays): works for NimbleGen two-color tiling arrrays.
  - Starr: an extension of Ringo, works for Affymetrix arrays.
  - ChIPpeakAnno: annotation of peaks, e.g., find closeby genes, GO terms, DNA sequences, etc.

rMAT

- R implementation of MAT.
- Works for Affymetrix arrays only.
- Needs bpmap file (factory provided file to probe annotations), and raw data file in CEL format.

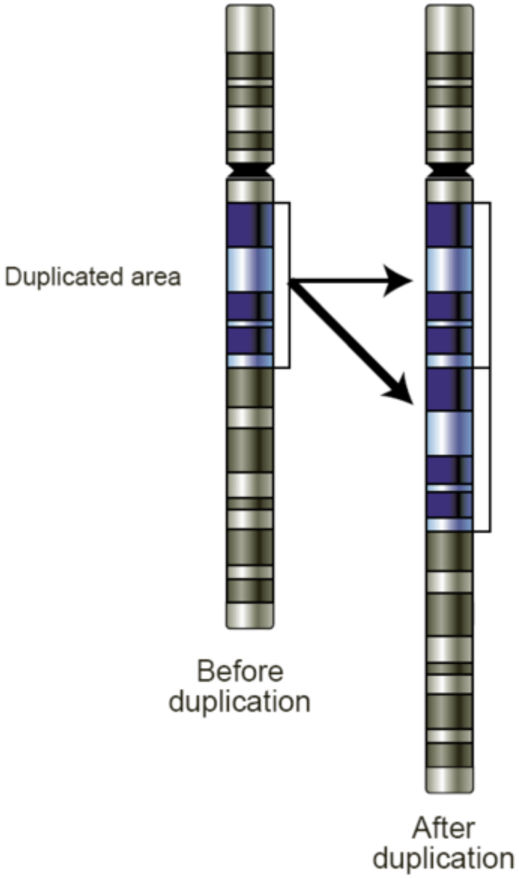
```
library(rMAT)
bpmapFile=“Hs.bpmap”
### read in data
arrayFile=c(“IP.CEL”, “Control.CEL”)
rawdata=BPMAPCelParser(bpmapFile, arrayFile, groupName="Sc")

## normalization - run MAT model
normdata=NormalizeProbes(rawdata,method="MAT”)

## compute MAT scores and find peaks
RD=computeMATScore(normdata,cName=“Control”, dMax=600)
Enrich=callEnrichedRegions(RD, dMax=600, dMerge=300,
   nProbesMin=8, method="score", threshold=2)
```

Copy number variation arrays

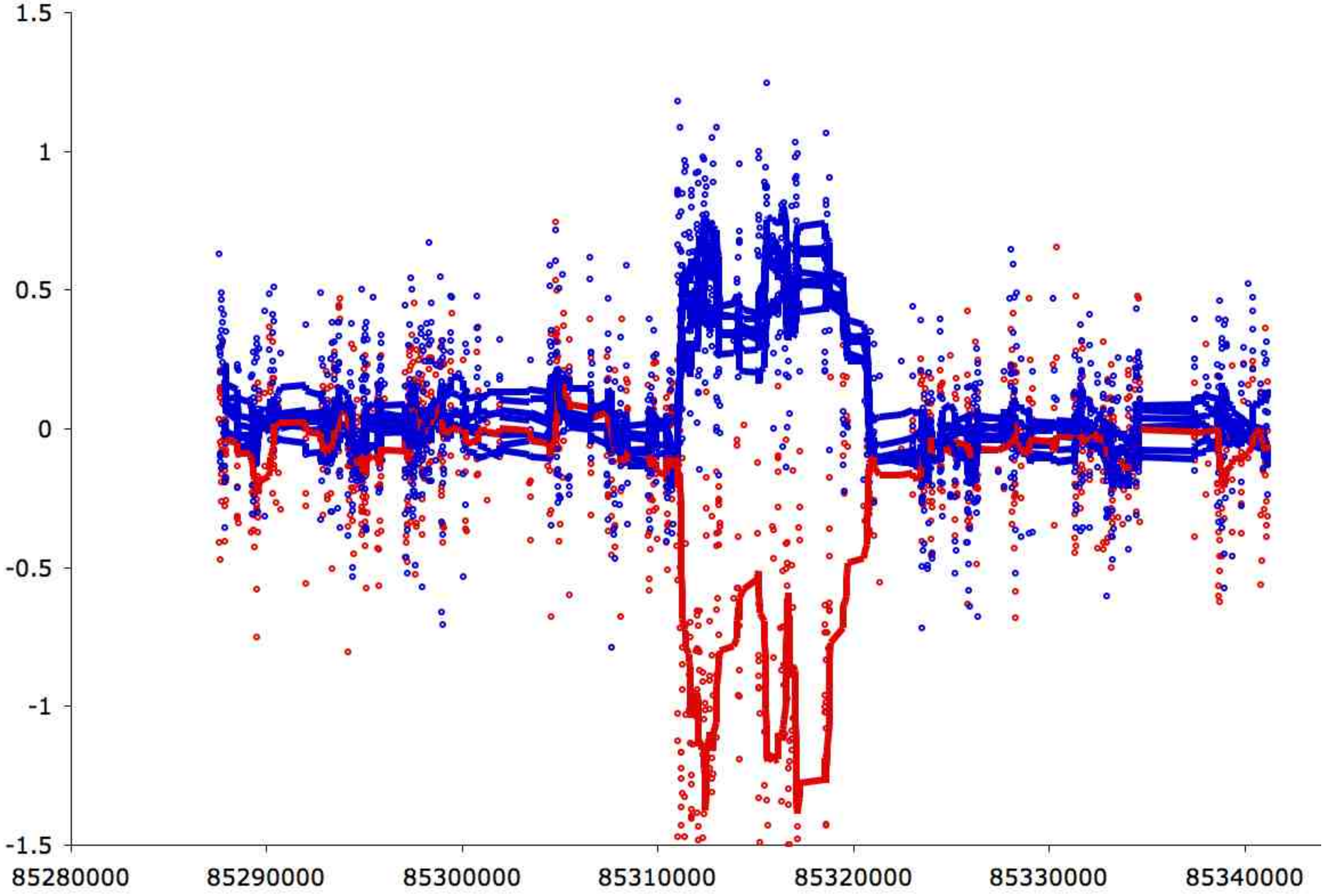
- CNV: phenomenon that sections of DNA have abnormal number of copies (deviate from 2).
- Can be detected by SNP arrays (for one sample) or arrayCGH (comparing two samples case vs. control).



Data from CNV arrays

- Data format are the same, e.g., probe locations and signal, but characteristics are different:
  - No peaks, but long, flat-topped "plateaus".
  - Heights of the plateaus are discrete, corresponding to different number of copies (integers: 1, 2, 3, …)

Example data from arrayCGH



Analysis of CNV arrays

- Methods are different from ChIP-chip, but still smoothing based to combine neighboring probe information, for example, Hidden Markov Model.

A list of CNV array software

- Affymetrix:
  - APT: uses a hidden Markov model
  - R package VanillaICE: HMM base. R. Scharpf *et al.* (2008) AOAS
  - R package DNAcopy: Circular Binary Segmentation. Olshen *et al.* (2004) Biostatistics
- Illumina:
  - QuantiSNP: S. Colella *et al.* (2007), NAR
  - PennCNV: K. Wang *et al.* (2008), NAR

Review

- Tiling arrays are DNA microarrays for detecting locational modifications of genome.
- Probes tile up a part of whole genome.
- Still hybridization based (DNA segments stick to probes), same as gene expression arrays.
- Data need to be visualized along genome.
- Location of interests shows some patterns: peaks for TFBS, or plateau for CNV.
- Need to combine data from neighboring probes to make calls.