Annotating Non-coding Regions of the Human Genome

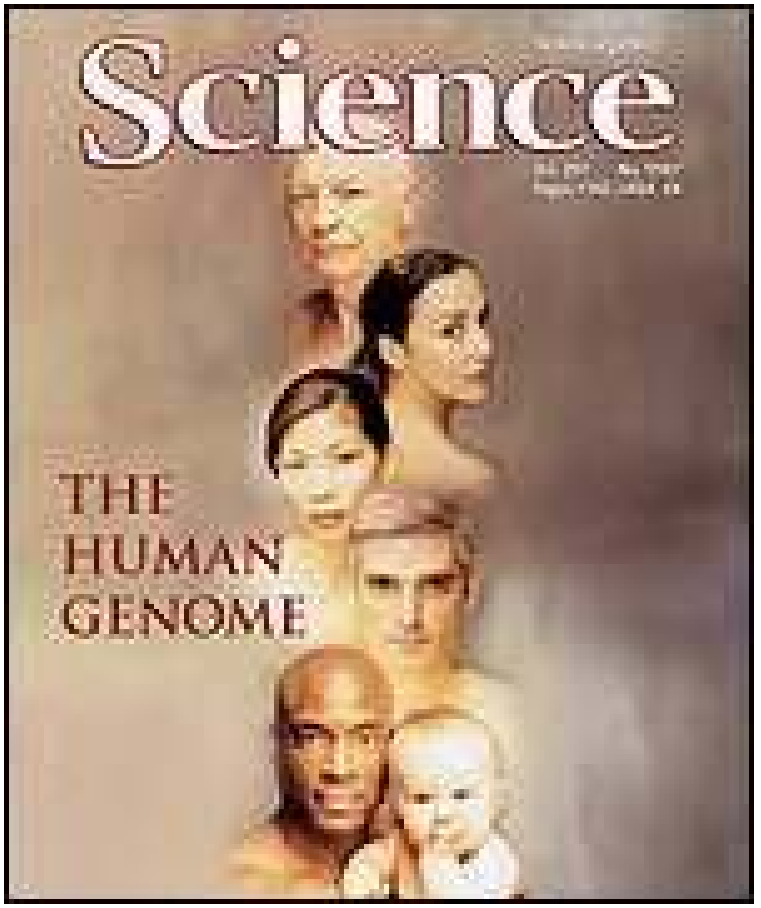
RECOMB Satellite Meeting
on Regulatory Genomics
Cambridge, MA
2008.10.31, 18:45-19:15

Mark B Gerstein
Yale (Comp. Bio. & Bioinformatics)

Slides from Lectures.GersteinLab.org

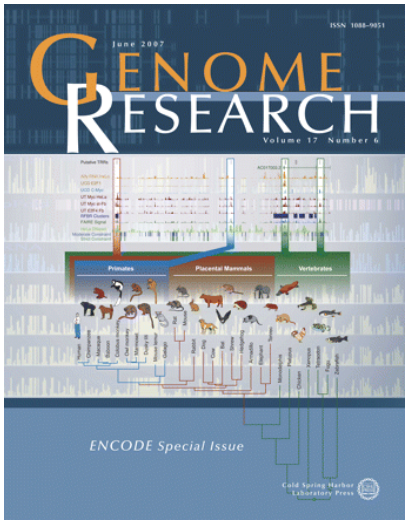
(Please read permissions statement.)
Paper references mostly from Papers.GersteinLab.org.
See streams.gerstein.info on photos & images

(Genome tech and Genome annotation talk, including:
Seq. Sim + PeakSeq ; MSB; DART TAR classification ;
TRE clustering + biplot, BoCaTFBS, sdcnvcorr [I:RECOMBSAT])



2001: Most of the genome is not coding (only ~1.2% exon). It consists of elements such as repeats, regulatory regions, non-coding RNAs, origins of replication, pseudogenes, segmental duplications....What do these elements do? How should they be annotated?

[IHGSC, ***Nature*** 409, 2001]
[Venter et al. ***Science*** 29, 2001]



2007 : Pilot results from ENCODE Consortium on decoding what the bases do

- 1% of Genome (30 Mb in 44 regions)
- Tiling Arrays to assay Transcription & Binding
- Multi-organism sequencing and alignment
- Careful Annotation
- Variation Data

[IHGSC, ***Nature*** 409, 2001]
[ENCODE Consortium, ***Nature*** 447, 2007]

How might we annotate a human text ?

Color is Function

Lines are Similarity

[B Hayes, Am. Sci. (Jul.- Aug. '06)]

The Semicolon Wars

Brian Hayes

IF YOU WANT TO BE a thorough-going world traveler, you need to learn 6,912 ways to say "Where is the toilet, please?" That's the number of languages known to be spoken by the peoples of planet Earth, according to Ethnologue.com.

If you want to be the complete polyglot programmer, you also have quite a challenge ahead of you, learning all the ways to say:

```
printf("hello, world\n");
```

(This one is in C.) A catalog maintained by Bill Kinnersley of the University of Kansas lists about 2,500 programming languages. Another survey, compiled by Diarmuid Piggott, puts the total even higher, at more than 8,500. And keep in mind that whereas human languages have had millennia to evolve and diversify, all the computer languages have sprung up in just 50 years. Even by the more-conservative standards of the Kinnersley count, that means we've been inventing one language a week, on average, ever since Fortran.

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity. All human languages are valuable; the

Every programmer knows there is one true programming language. A new one every week

a good-enough notation—for expressing an algorithm or defining a data structure.

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect, they'll concede, but it's built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will favor a different language. It's Lisp,

cide which end of a boiled egg to crack. This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the least-significant bit or the most-significant bit can go first. Which way is better? It hardly matters, although life would be easier if everyone made the same choice. But that's *not* what has happened, and so quite a lot of hardware and software is needed just to swap ends at boundaries between systems.

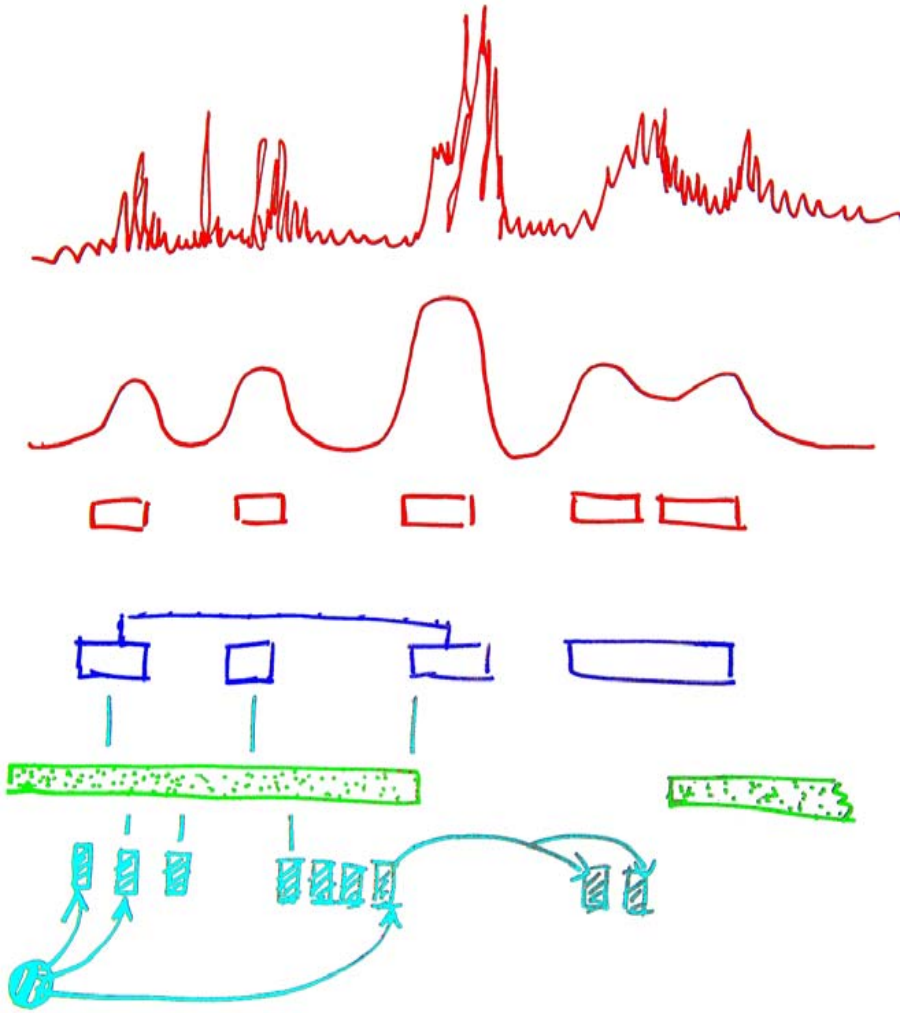
This modern echo of Swift's Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, "On holy wars and a plea for peace." The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

Another feud—largely forgotten, think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements have to be separated by semicolons. For example, in `x:=0; y:=x+1; z:=2` the semicolons tell the compiler where one statement ends and the next begins. C programs are also peppered with semi-

Overview of the Process of Annotation of non-coding Regions

- Basic Inputs
  1. Doing large-scale similarity comparison, looking for repeated or deleted regions
  2. Determining experimental signals for biochemical activity (e.g. transcription) across each base of genome
- **Results of Analyzing Similarity Comparison**
  1. Finding large repeated or deleted blocks (e.g. CNVs) as a function of degree of similarity
     1. within reference human genome
     2. within human population
     3. between related organisms (e.g. mouse)
  2. Finding smaller "exon-level" similarities (e.g. pseudogenes)

Overview of Annotation Process

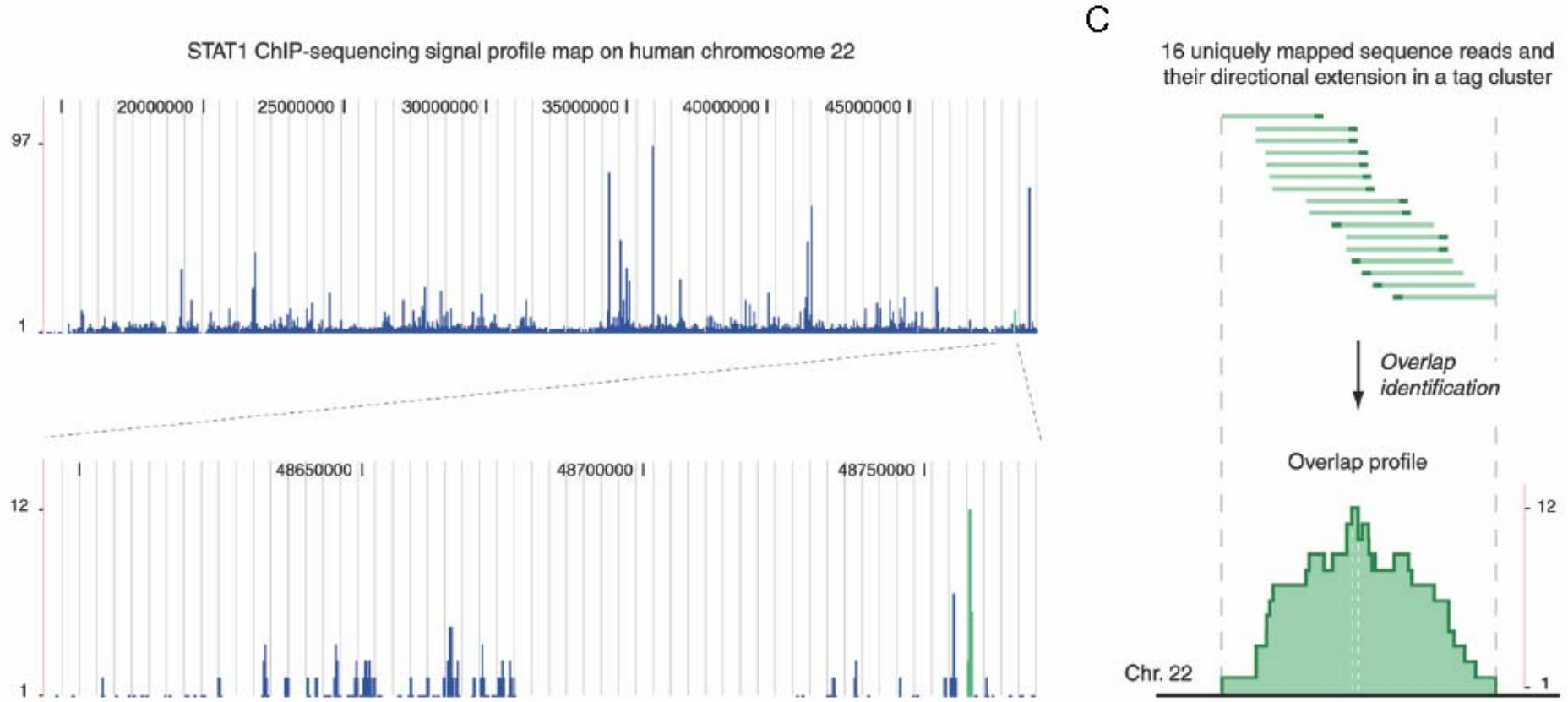


- **Determining experimental signals for biochemical activity (e.g. transcription) across each base of genome**
  - ◊ **Development of Sequence (and Array) Technology**
    - **Normalizing & Scoring Signal, Correcting Artifacts, Segmenting to create Small Annotation Blocks**
  - ◊ **Output of Production Pipelines and Surveying a Single Type of Annotation on a Large-scale**
    - **Clustering Small Blocks into Larger Ones, Surveying**
  - ◊ **Integrated Analysis Connecting Different Types of Annotation**
    - **Building networks and beyond**

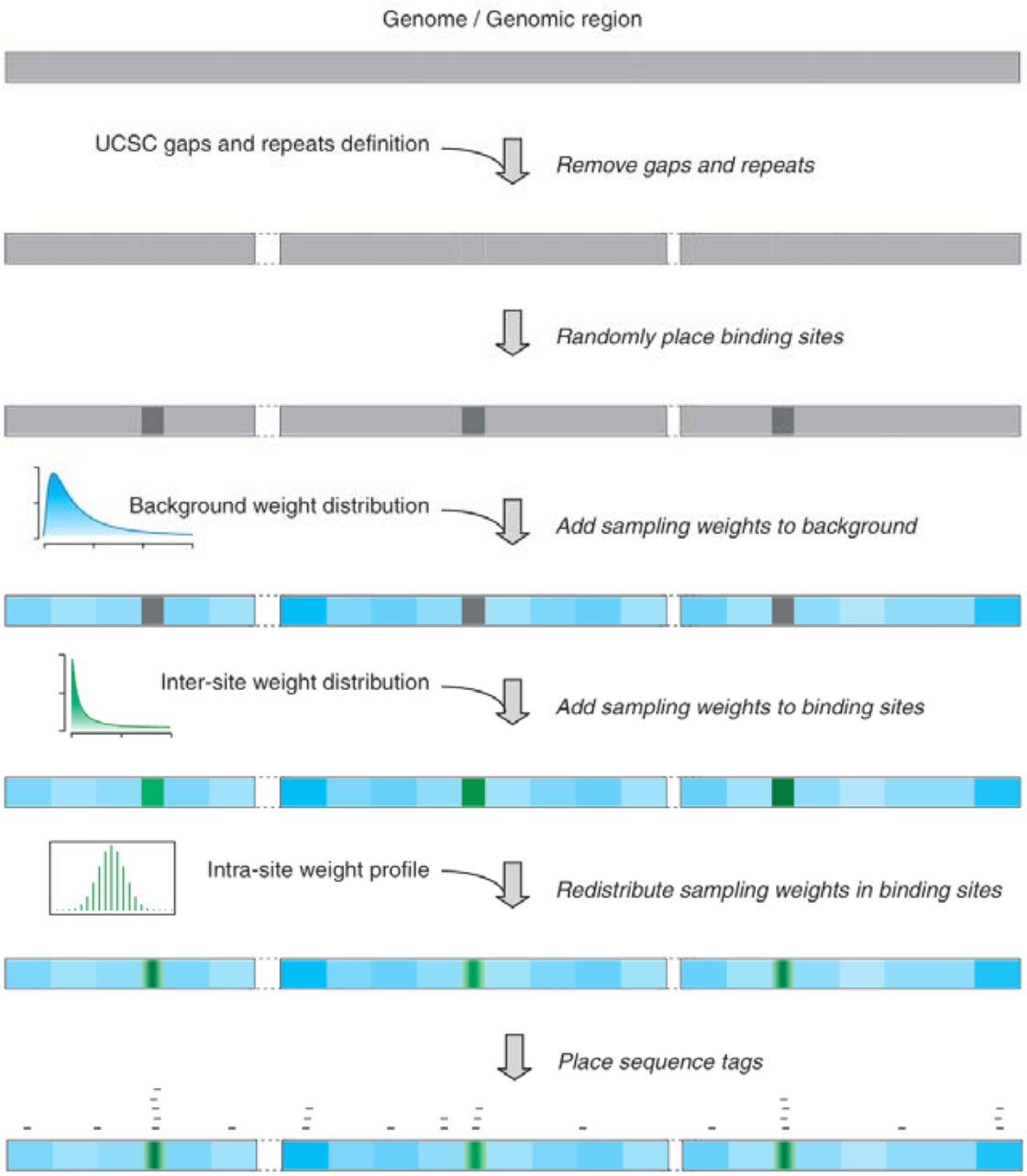
Signal Processing: Normalizing Signal and Finding Initial Annotation Blocks ("Hits")

pers. **photo**, see streams.gerstein.info

Representative Signal from Chip-Seq



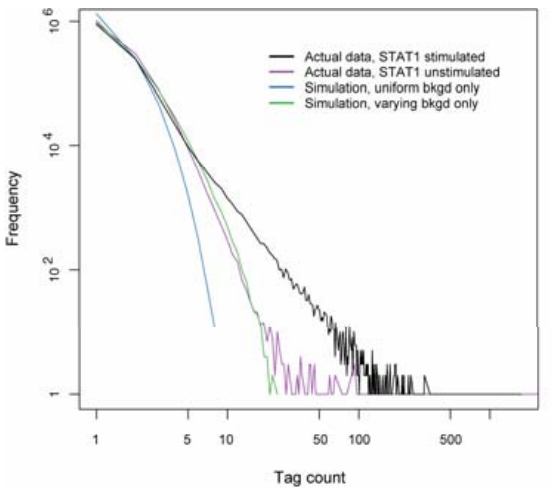
[Robertson et al., Nat. Meth. ('07); Zhang et al. PLOS Comp. Bio. (in revision, '08)]

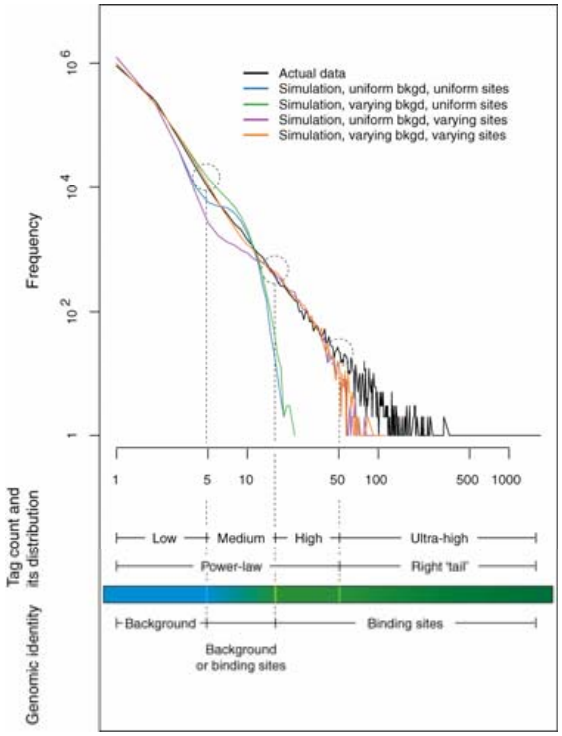


Correcting Chip-seq Signal by Simulating a Non-uniform Genomic Background

- We developed *in silico* ChIP sequencing, a computational method to simulate the experimental outcome.

[Zhang et al. PLoS Comp Bio. ('08)]





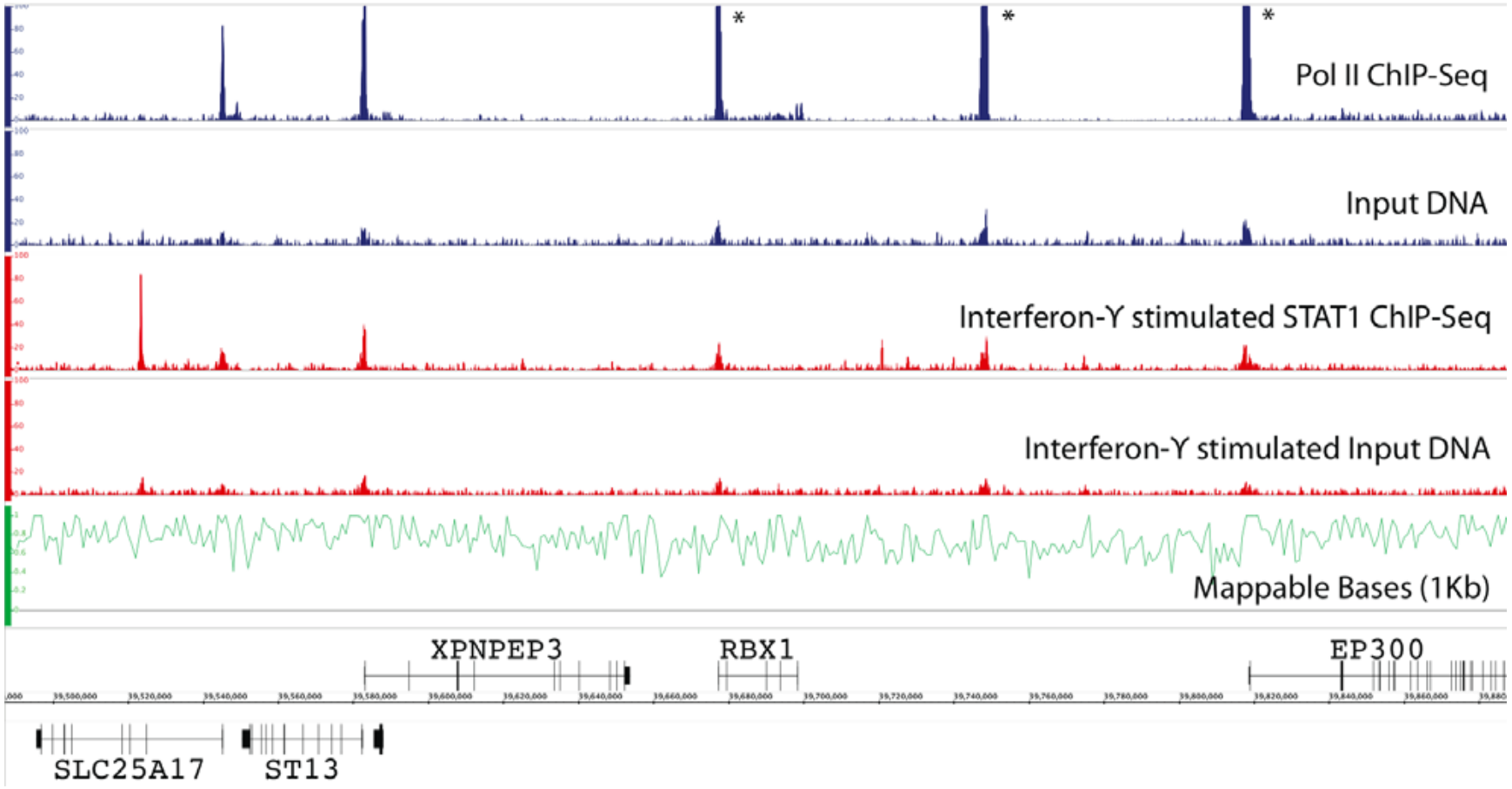
ChIP-sequencing simulation

- Contrary to the common belief, the background is mildly fluctuating and contains some ‘hot’ spots.
- Simple uniform background model does not count for all the variation in the background and thus leads to a serious underestimation of the background noise.

- Our study demonstrates that both the genomic background of ChIP and binding sites are not uniform.
- Simulated distributions segments the actual distribution into four sections.

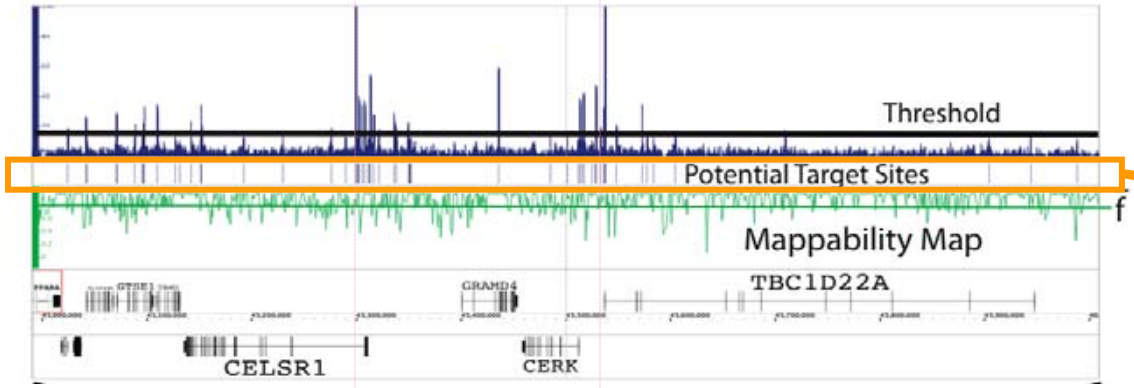
[Zhang et al. PLoS Comp Bio. ('08)]

ChIP-Seq vs Input DNA Control



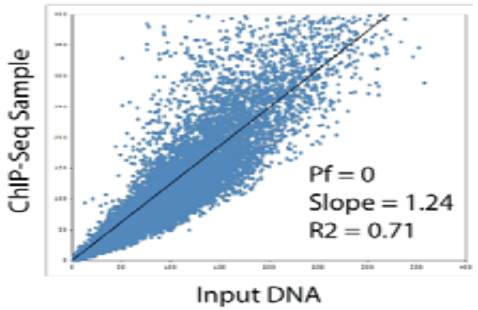
[Rozowsky et al. (submitted)]

PeakSeq: Scoring Relative to Controls



Filter for Potential Targets based on "Mappability" Simulation

Scale Input Relative to ChIP

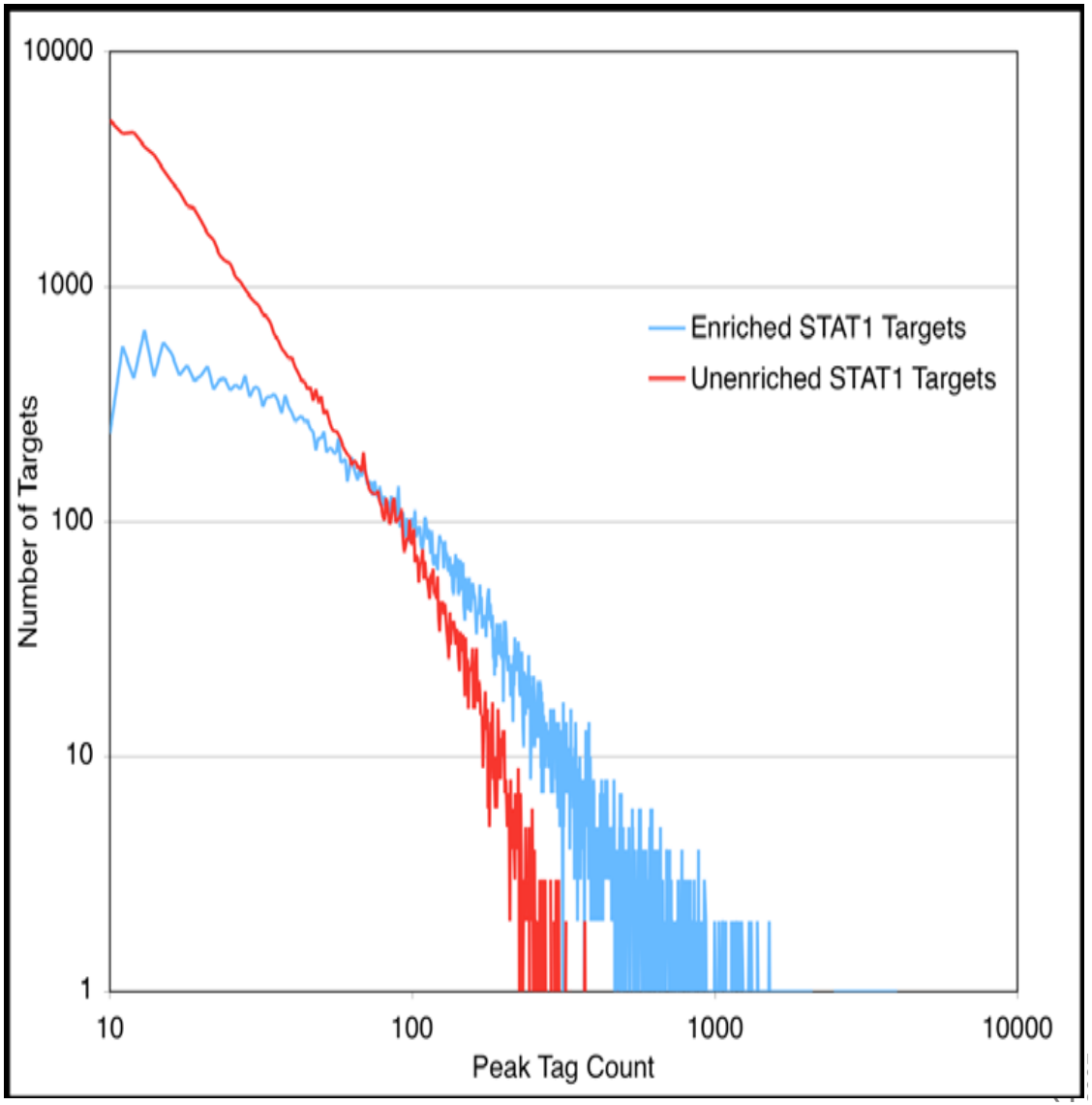


Score Relative to Bionomial Expectation

[Rozowsky et al. (submitted)]

Scored results consistent with simulation

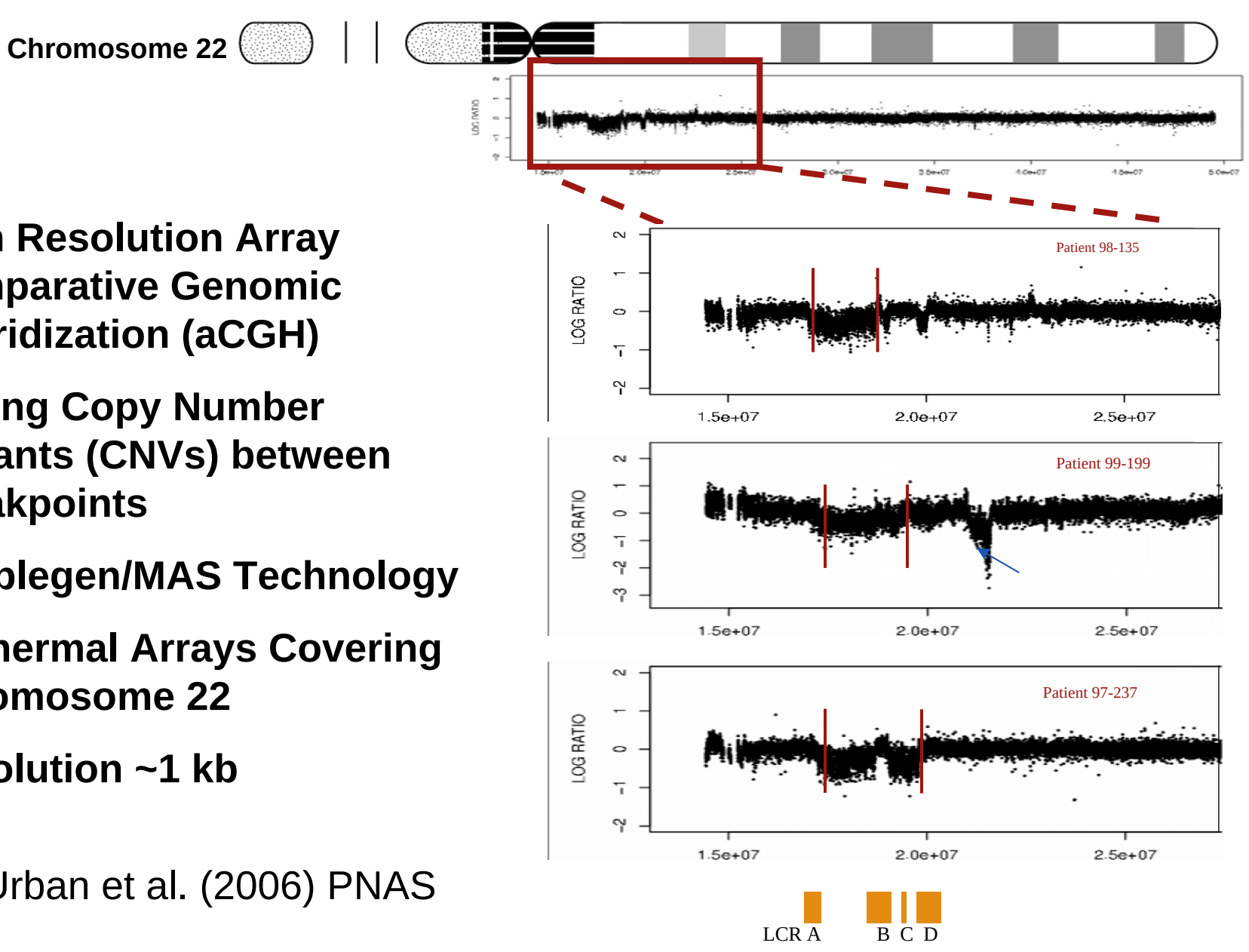
Actual peaks at tail of power-law graph



[Rozowsky et al. (submitted)]

Punctate Regions vs Broad Regions

Representative Signal from aCGH with CNVs & Breakpoints



High Resolution Array Comparative Genomic Hybridization (aCGH)

Calling Copy Number Variants (CNVs) between Breakpoints

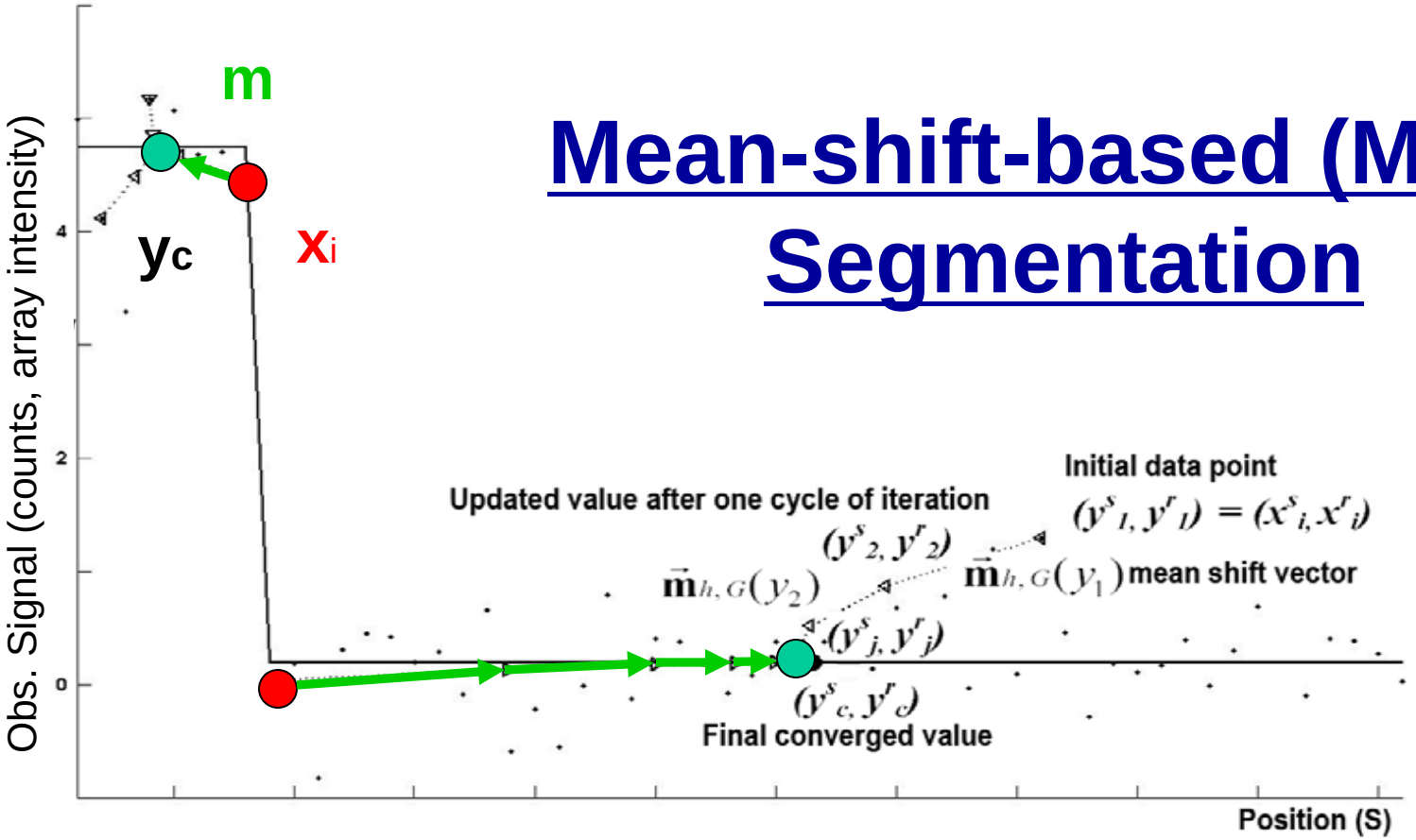
Nimblegen/MAS Technology

Isothermal Arrays Covering Chromosome 22

Resolution ~1 kb

Urban et al. (2006) PNAS

Mean-shift-based (MSB) Segmentation

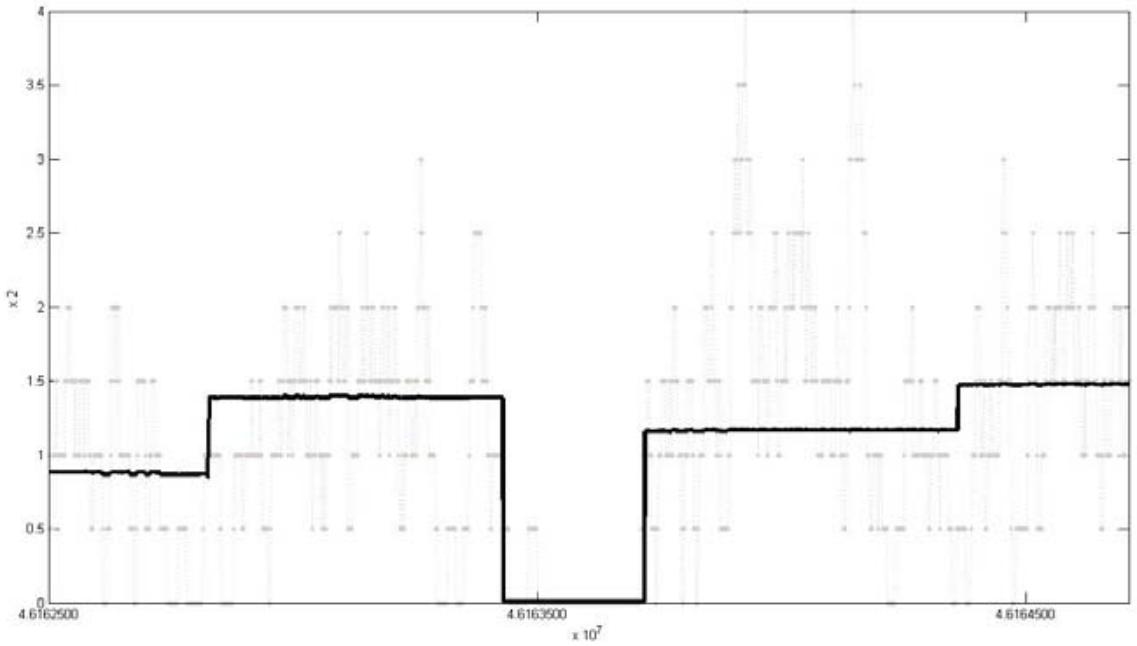


(x_i) Observed depth of coverage counts (or array signal) as samples from PDF
(m) Kernel-based approach to estimate local gradient of PDF
(y_c) Iteratively follow grad to determine local modes

Not Model-based (e.g. like HMM)
with global optimization, distr. assumption & parms. (e.g. num. of segments).
Achieves discontinuity-preserving smoothing

[Wang et al. Gen. Res (in press, '08)]

Representative Result Showing Segmentation Based on Depth of Coverage



NA11995 (seq. by Sanger, MAQ mapping)
chr 21 (46162500 to 46164711)

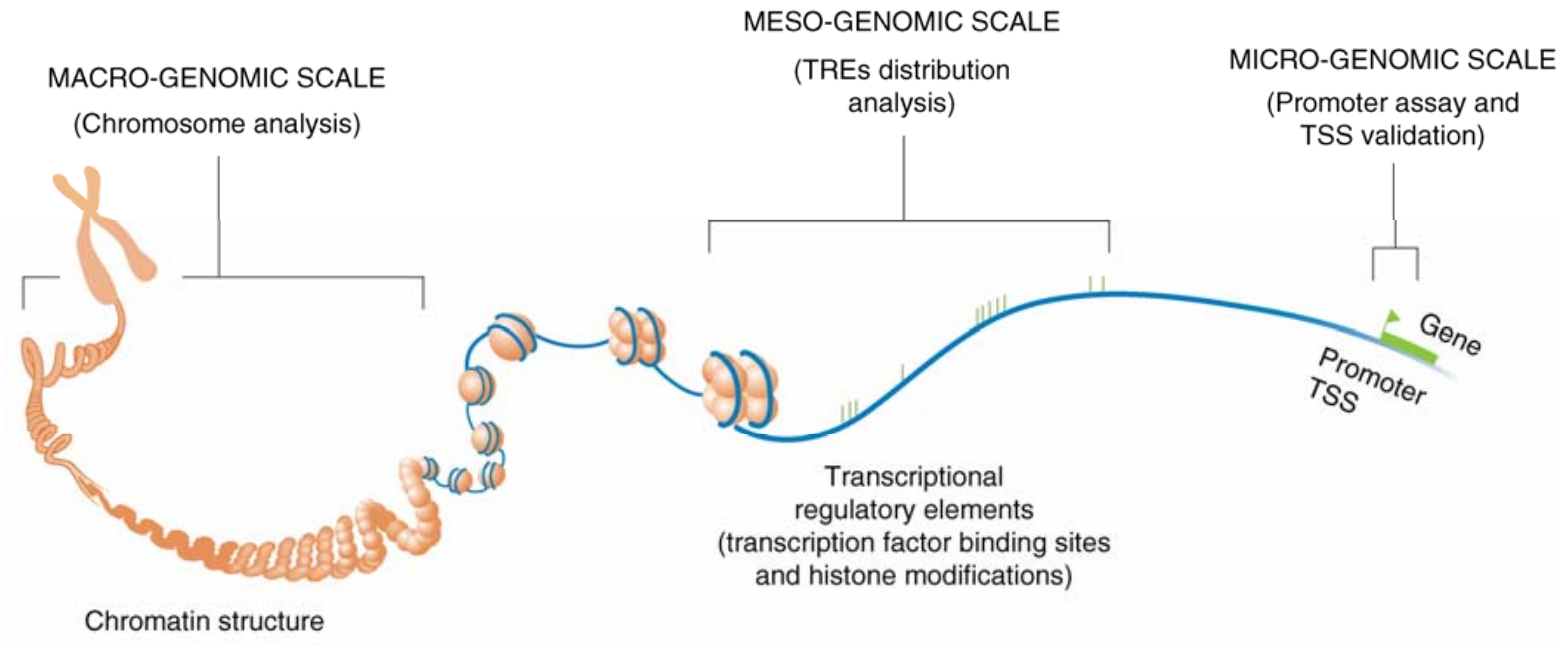
MSB is not model based so can be applied equally well to pseudo-signal from coverage depth as to CGH arrays ??

[Wang et al. Gen. Res (in press, '08)]

Annotating a single type of signal on a large-scale: Clustering and Characterizing Binding Sites (TREs)

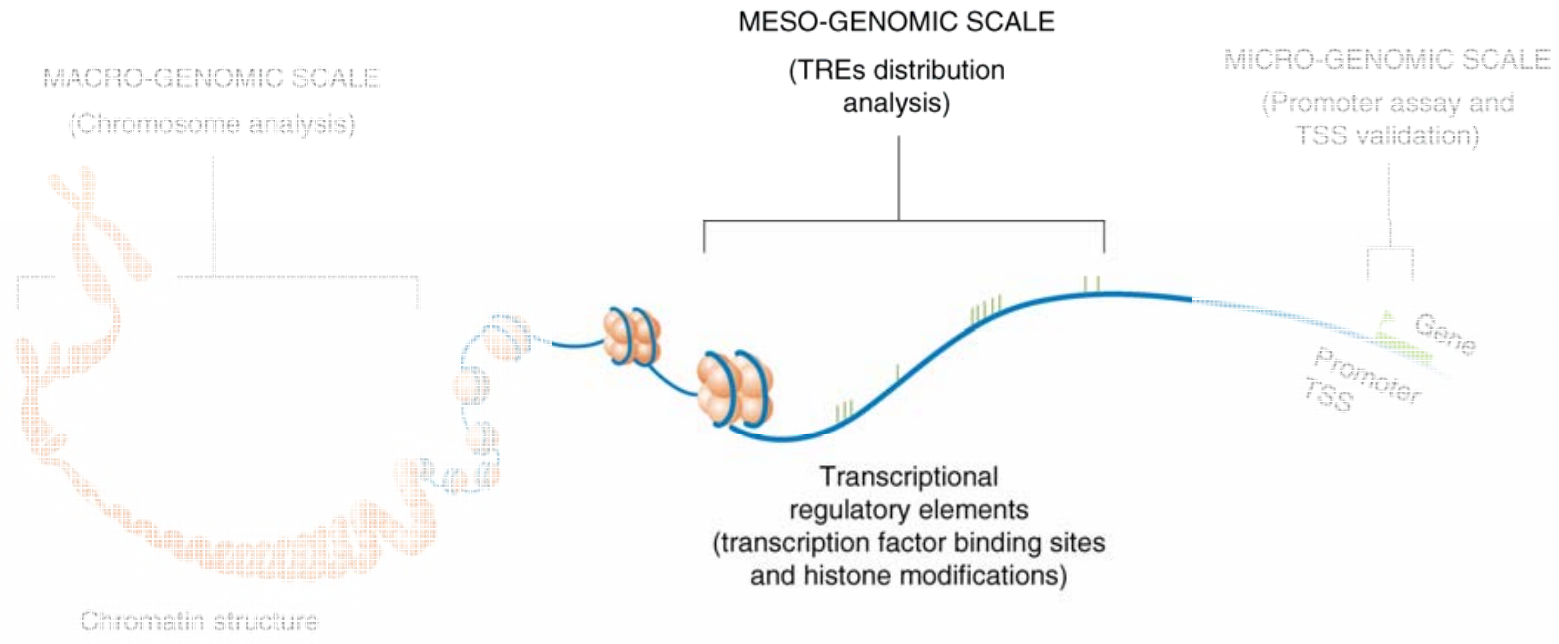
pers. **photo**, see streams.gerstein.info

Clustering Binding Sites at ~50kb resolution



[Zhang et al. (2007) Gen. Res.]

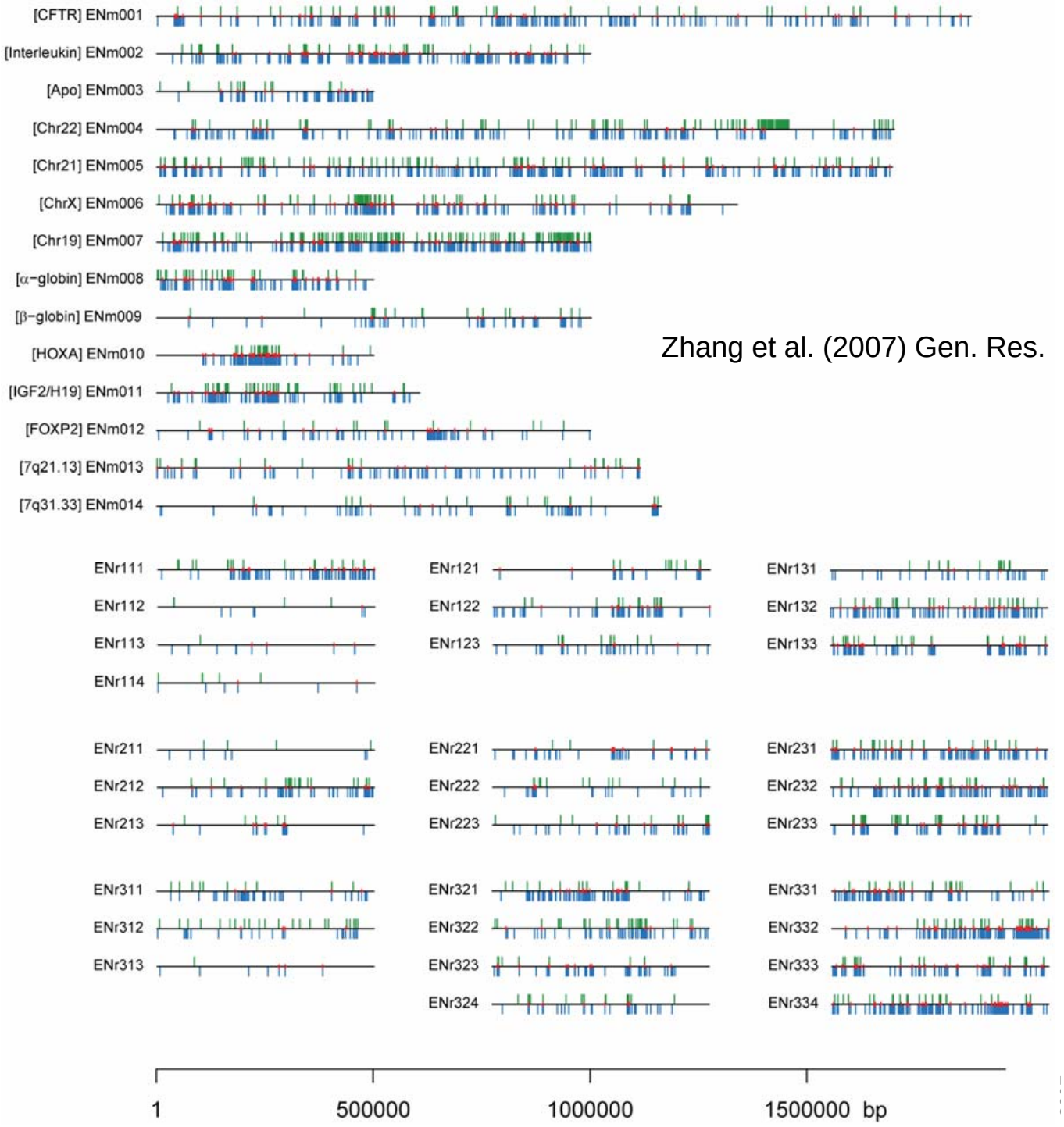
Clustering Binding Sites at ~50kb resolution



[Zhang et al. (2007) Gen. Res.]

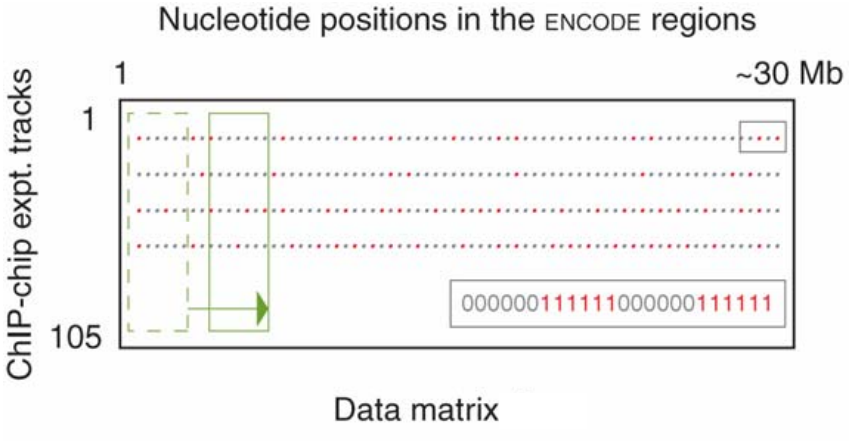
Landscape of ENCODE Transcriptional Regulatory Elements

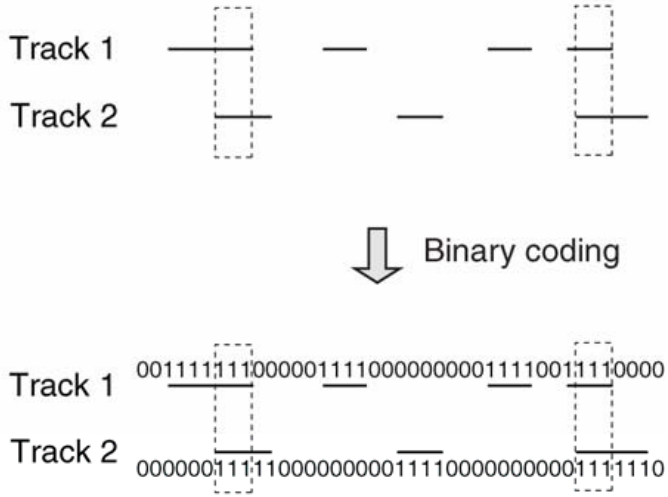
- Analyzed 105 lists of transcriptional regulatory elements in the encode regions
- 29 transcription factors, 9 cell lines, 2 time points
  - ◊ RNA Pol2
  - ◊ Histone modifications such as Ac & Me
  - ◊ Core promoters
  - ◊ Promoter proximal elements
  - ◊ Others such as enhancers, silencers, insulators, & response elements

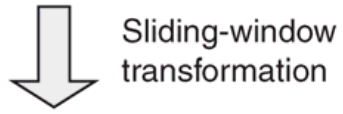


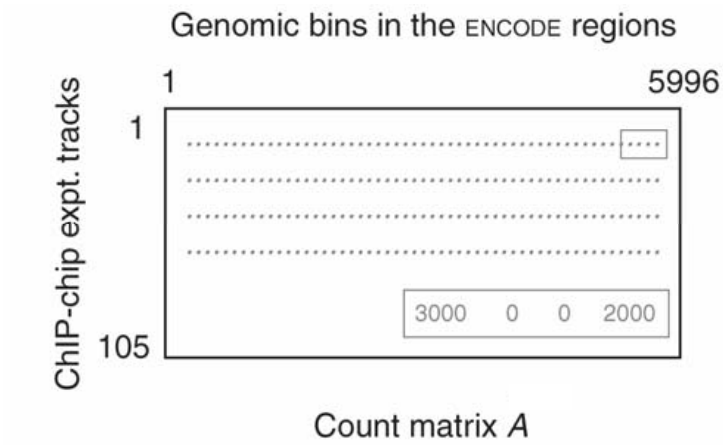
Zhang et al. (2007) Gen. Res.

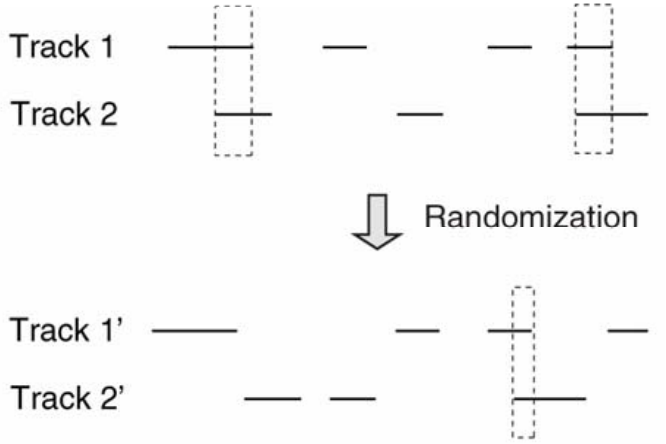
Collect Total Hits for Each Factor in ~6000 Bins of 10 to 100 kb and Compare to Random Control







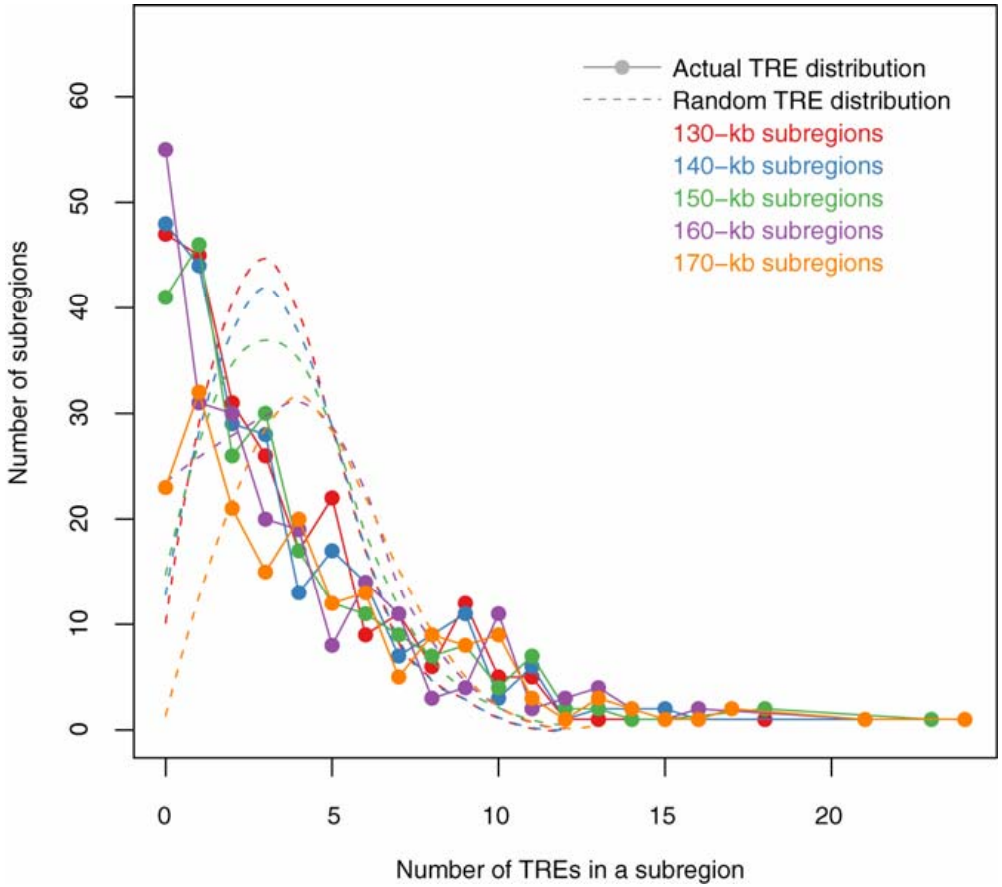




Zhang et al. (2007) Gen. Res.

Non-random distribution of TREs

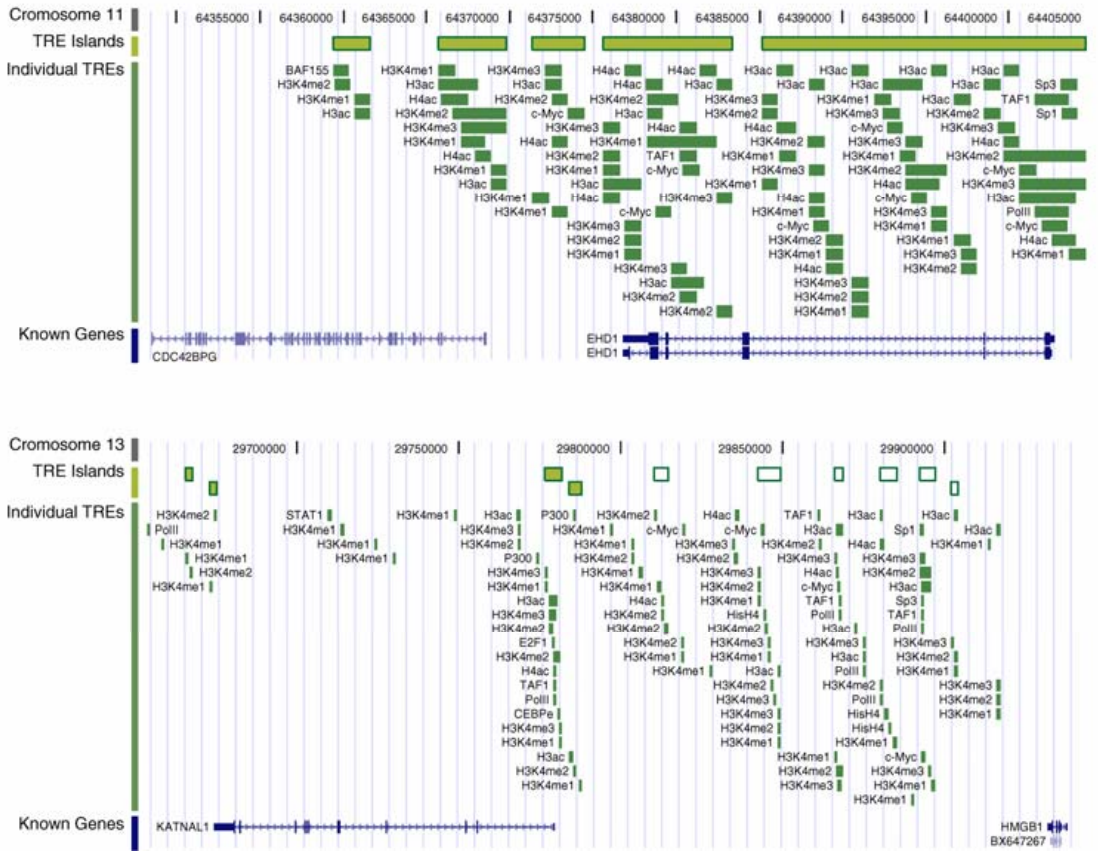
- TREs are not evenly distributed throughout the encode regions ($P < 2.2\times10^{-16}$).
- The actual TRE distribution is power-law.
- The null distribution is 'Poissonesque.'
- Many genomic subregions with extreme numbers of TREs.



Zhang et al. (2007) Gen. Res.

Local TRE enrichment and depletion: Annotation of Desserts and Forests

- Hundreds of TRE ‘forests’ and ‘deserts’ are identified in ENCODE regions.
- The entirety of *ehd1* on chromosome 11 is covered by TRE islands.
- Some of islands are located in the intergenic regions in the genome.



dart.gersteinlab.org/encode/tr/

Zhang et al. (2007) Gen. Res.

Biplot to Show Overall Relationship of TFs and Genomic Bins

TFs: a, b, c...

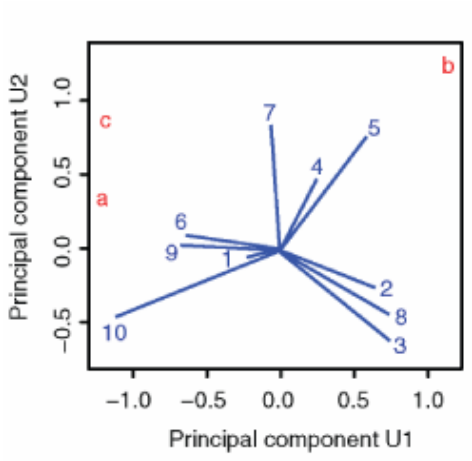
50kb Genomic Bins: 1,2,3...

	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

$A = USV^T$

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
2	0.70	1.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
3	0.69	1.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
4	0.77	0.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
5	0.54	0.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
6	0.99	0.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
7	0.95	0.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
10	0.97	0.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

AA^T

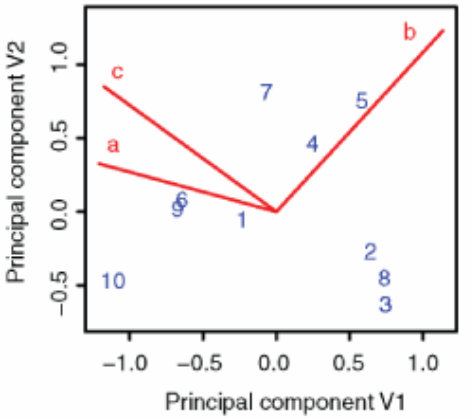


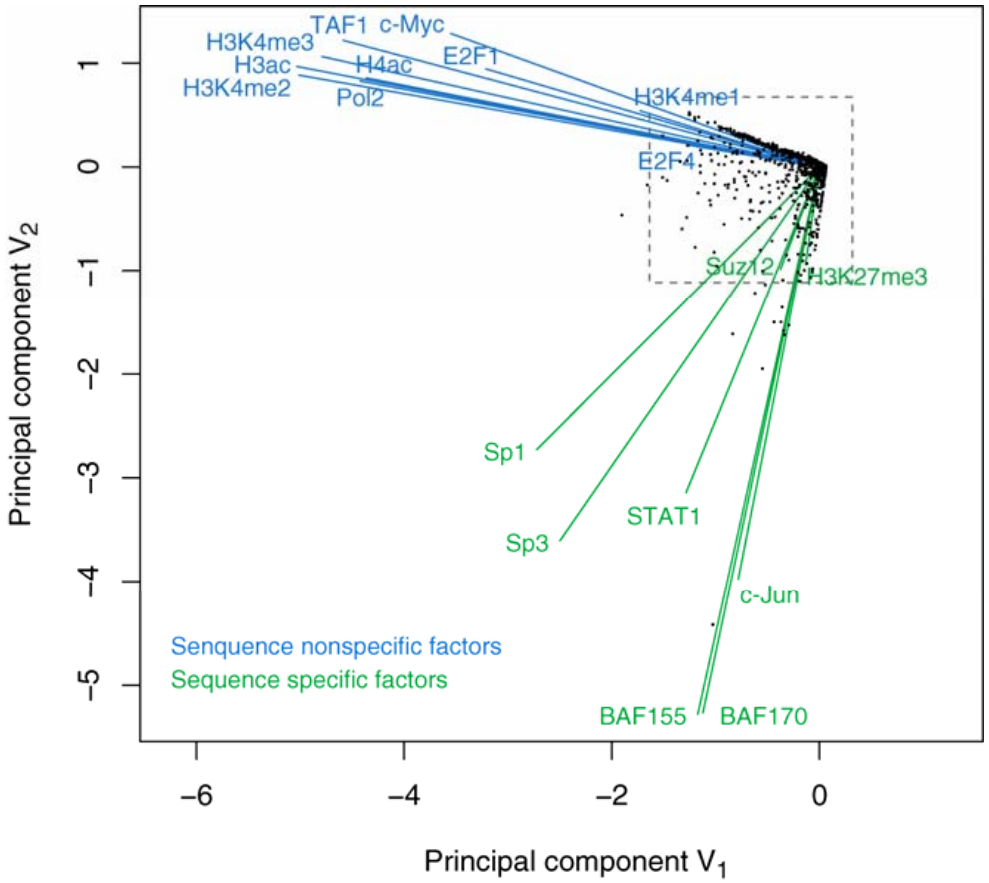
A^T

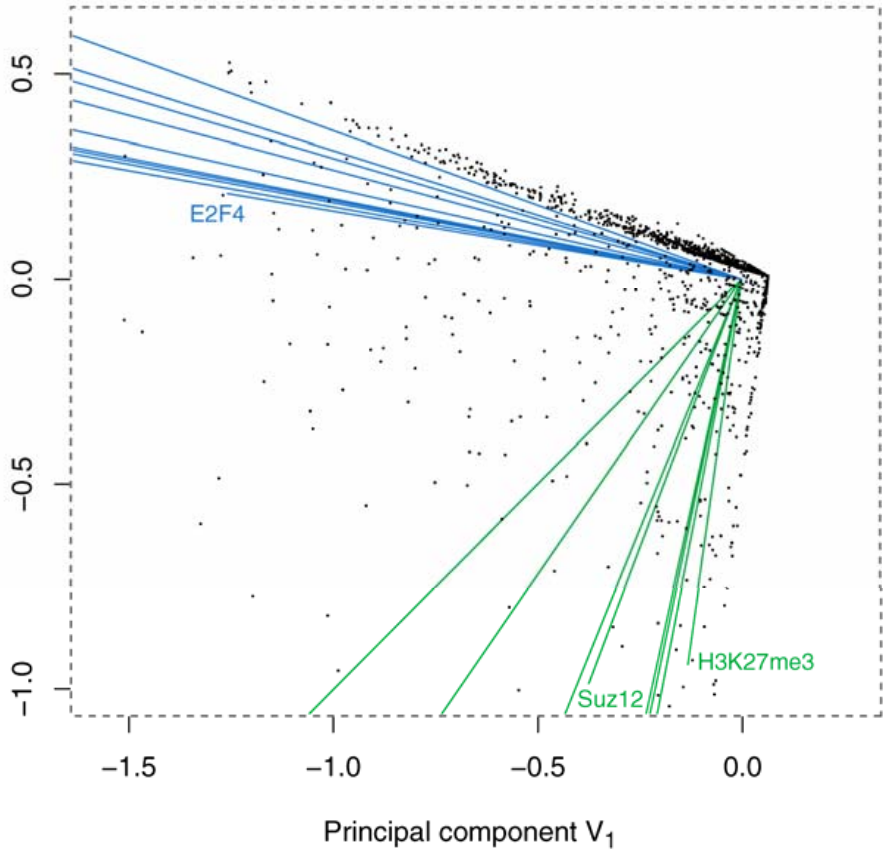
	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

A^TA





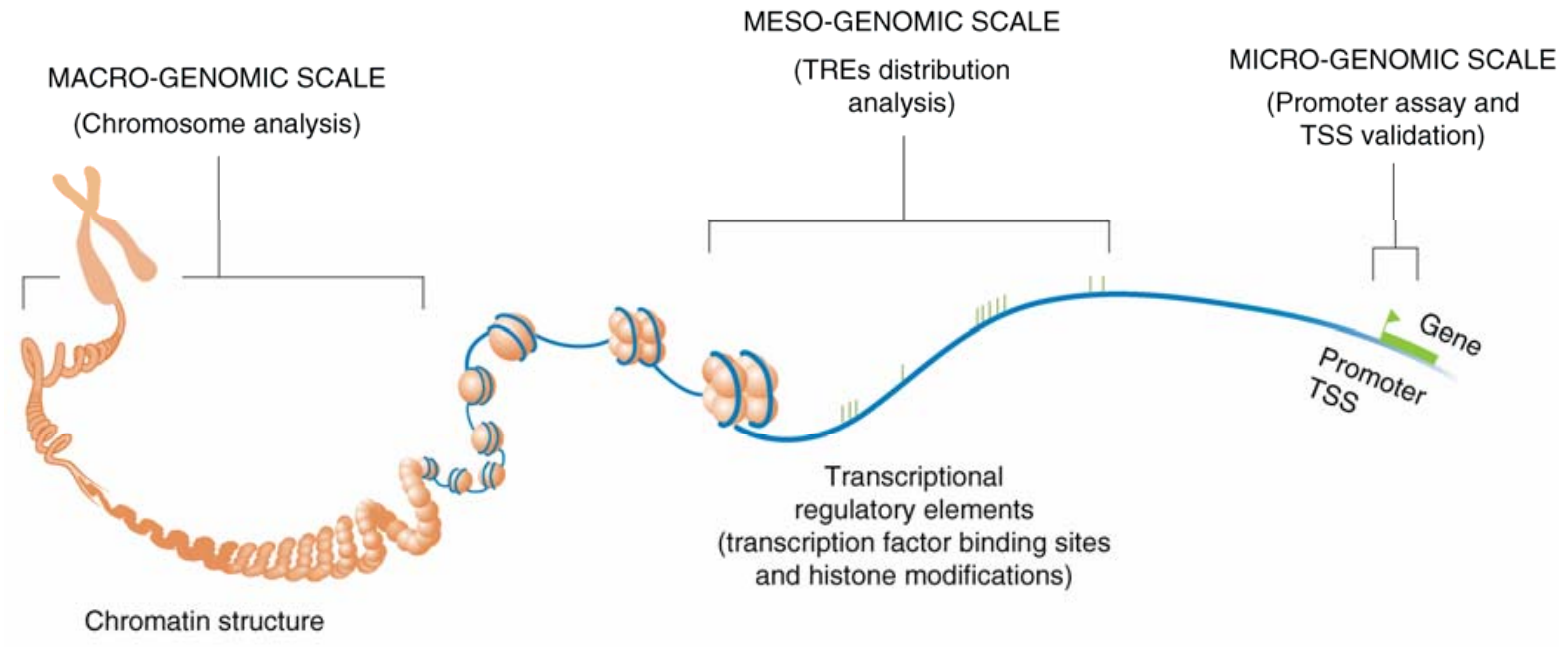


Results of Biplot

- Biplot groups TFs into sequence-specific and sequence-nonspecific clusters.
  - ◊ c-Myc may behave more like a sequence-nonspecific TF.
  - ◊ H3K27me3 functions in a transcriptional regulatory process in a rather sequence-specific manner.
- Genomic Bins are associated with different TFs and in this fashion each bin is "annotated" by closest TF cluster

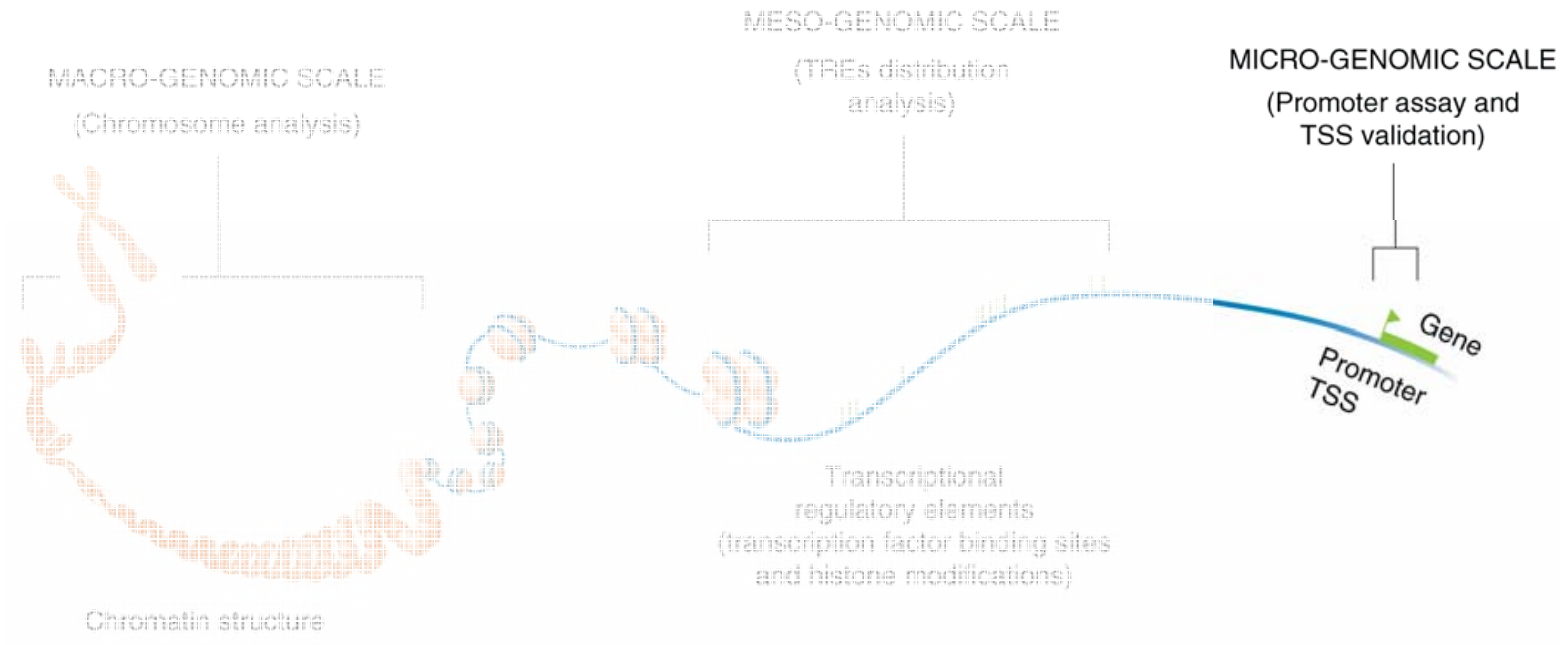
Zhang et al. (2007) Gen. Res.

TRE analysis on the micro-genomic scale



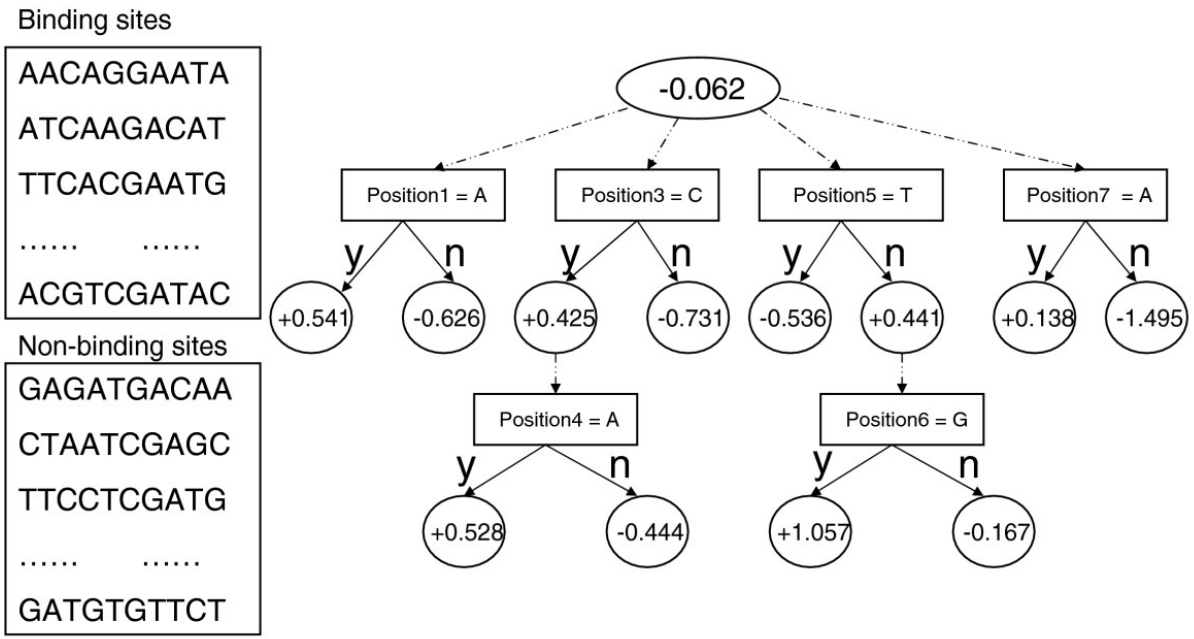
[Zhang et al. (2007) Gen. Res.]

TRE analysis on the micro-genomic scale



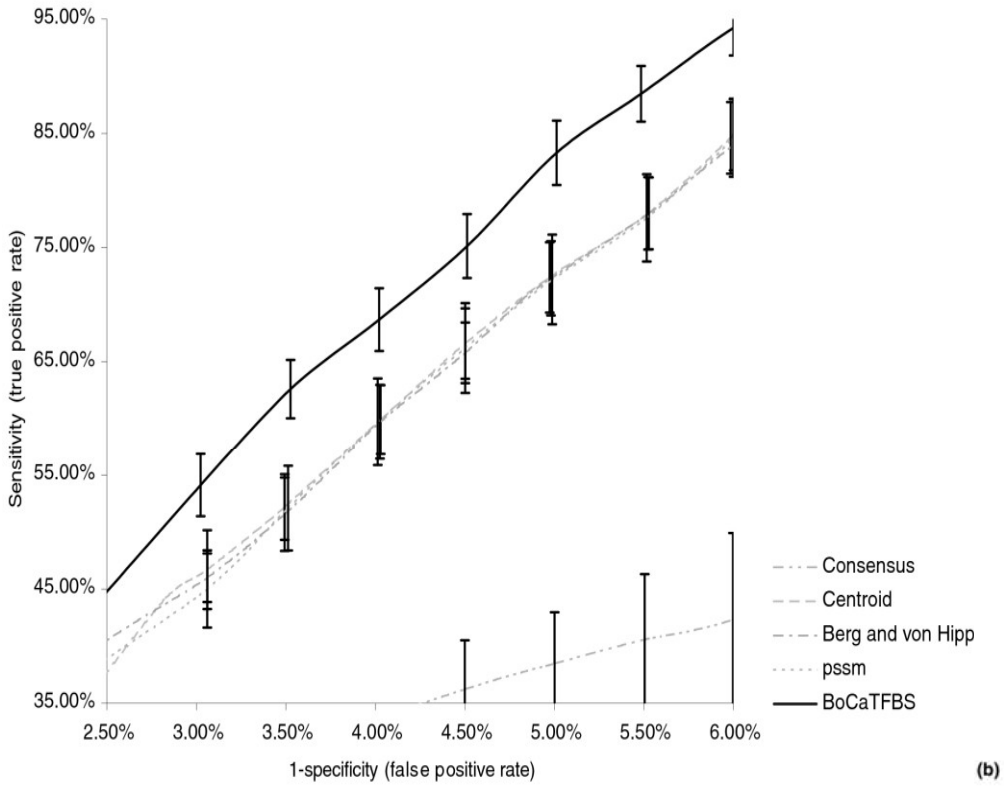
- Traditional motif learners (e.g. consensus sequences, profile methods, and HMMs) only use positive information
- ChIP-chip & Chip-seq give vast amount of negative information (regions not bound)
- Explicitly use this in constructing classifier that **refines** known positive motif seeds
- Use sequence of **Alternating Decision Trees (ADTboost),** which allow explicit inter-positional correlations between nucleotide positions

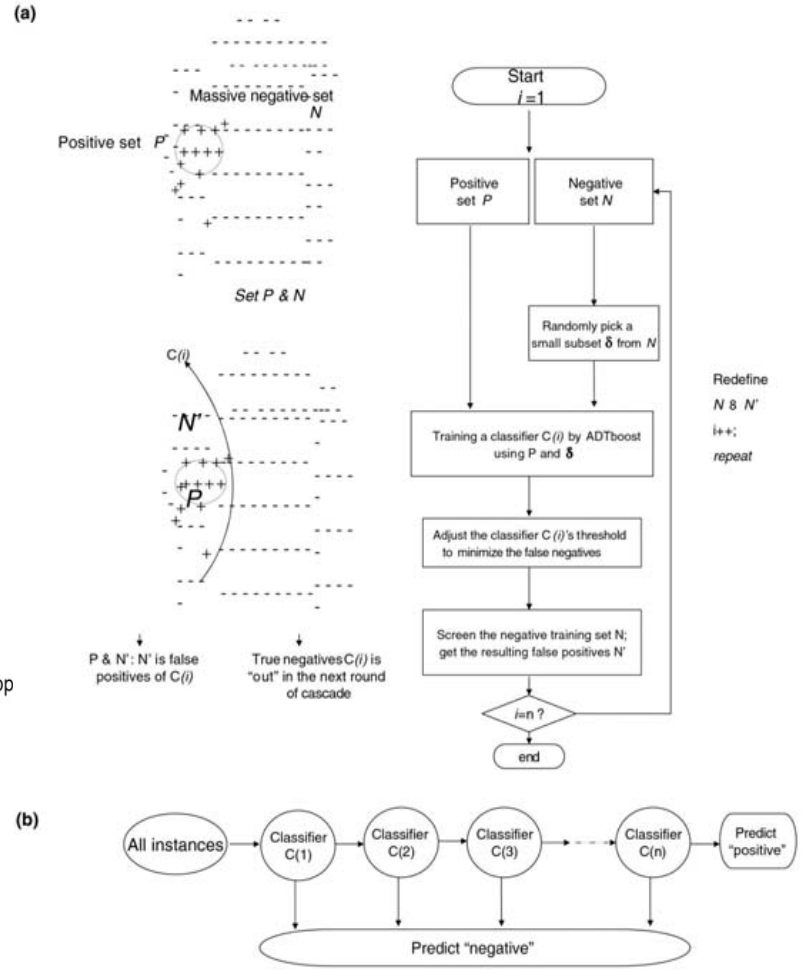
Using Binding Site Regions Found by ChIP-chip to refine motifs: BoCaTFBS



[Wang et al., GenomeBiology ('06)]

Good performance compared to traditional motif-finders but large negative set requires training and detection cascade for efficiency and balance

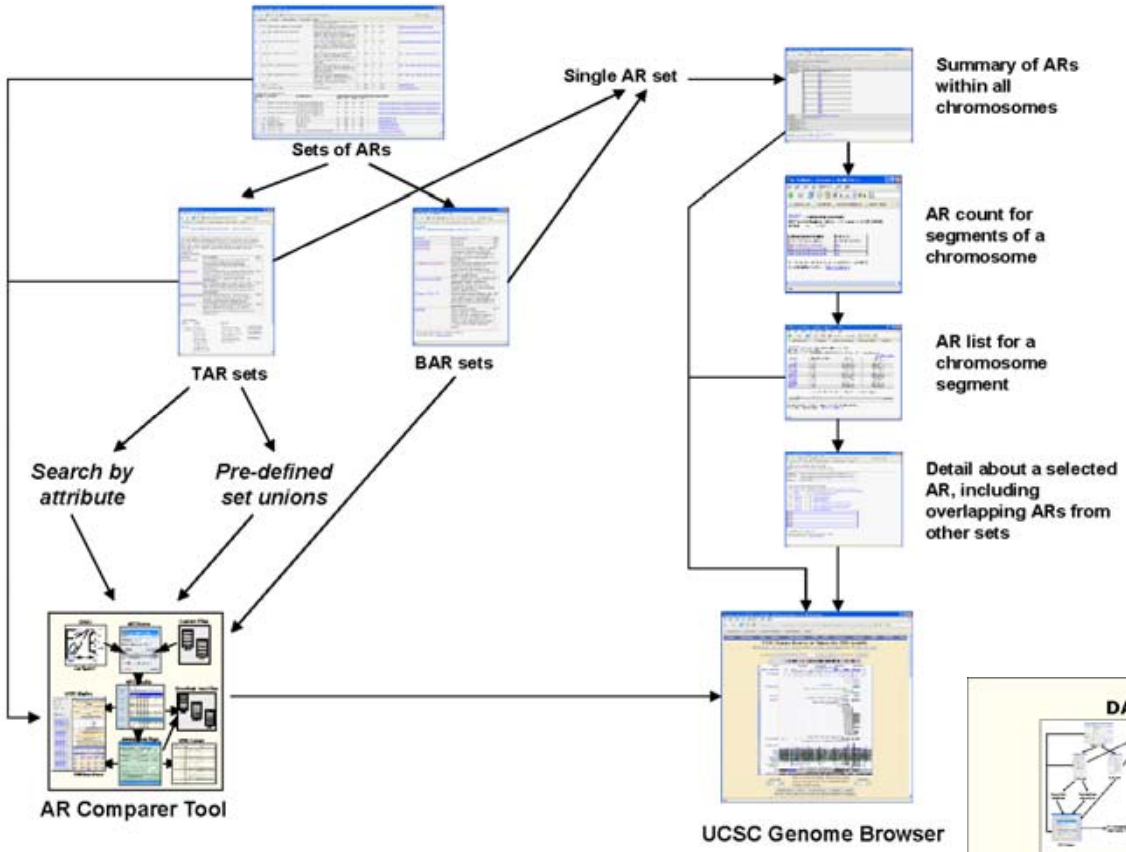


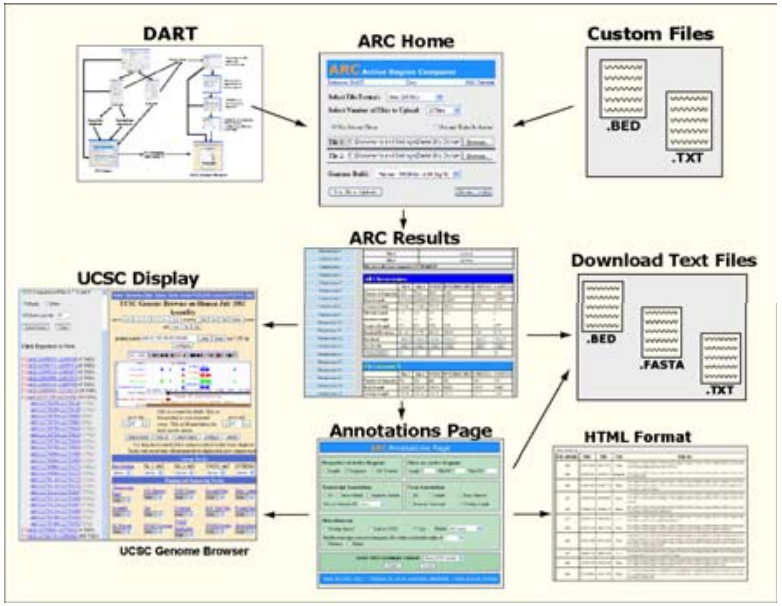


[Wang et al., GenomeBiology ('06)]

Annotating a single type of signal on a large-scale: Clustering and Classifying Un-annotated Transcription (TARs)

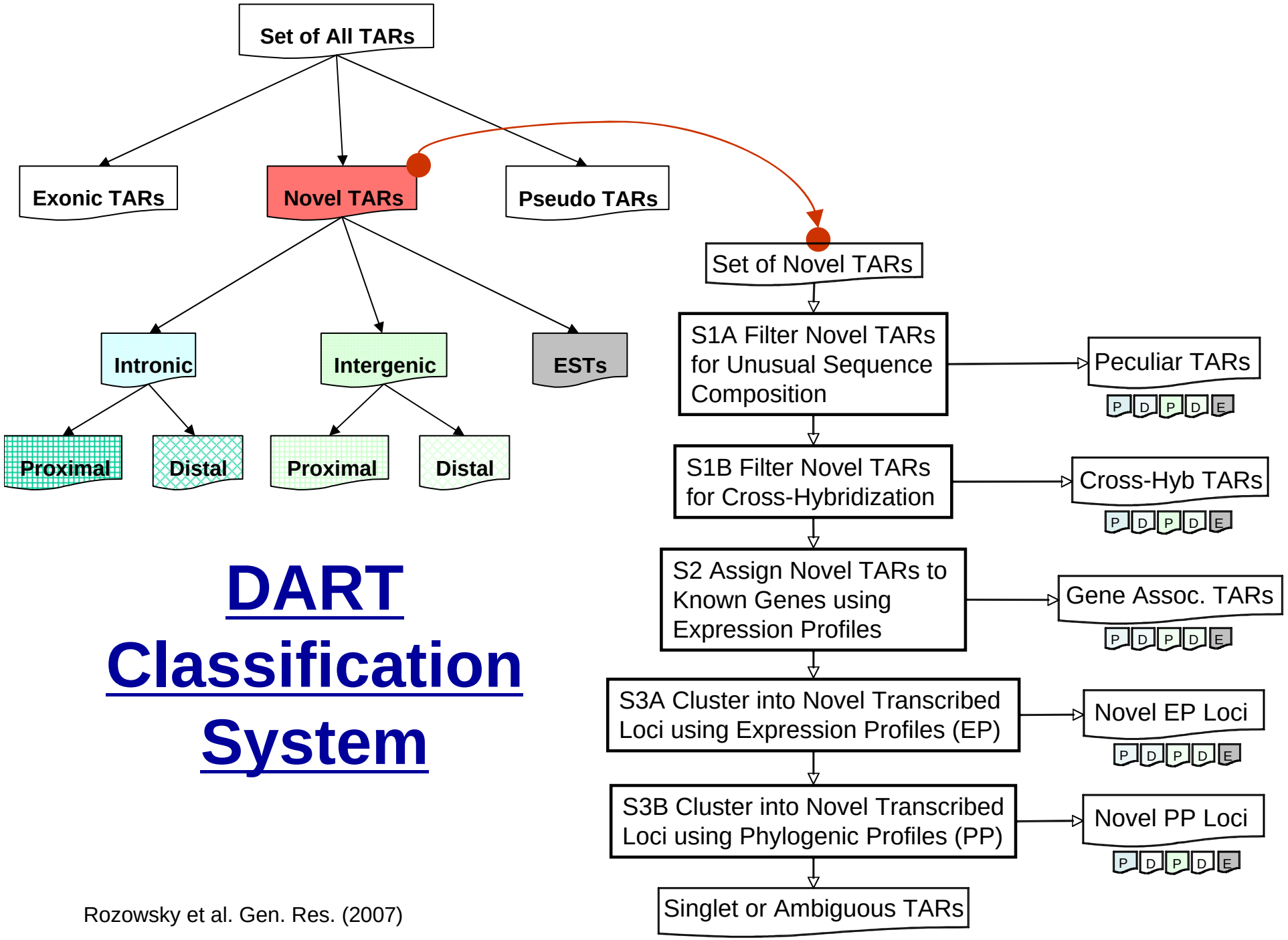
pers. **photo**, see streams.gerstein.info





DART: Database & Tools
- Interfaces with UCSC
- Tools use Ensembl API

Rozowsky et al. Genome Research (2007)

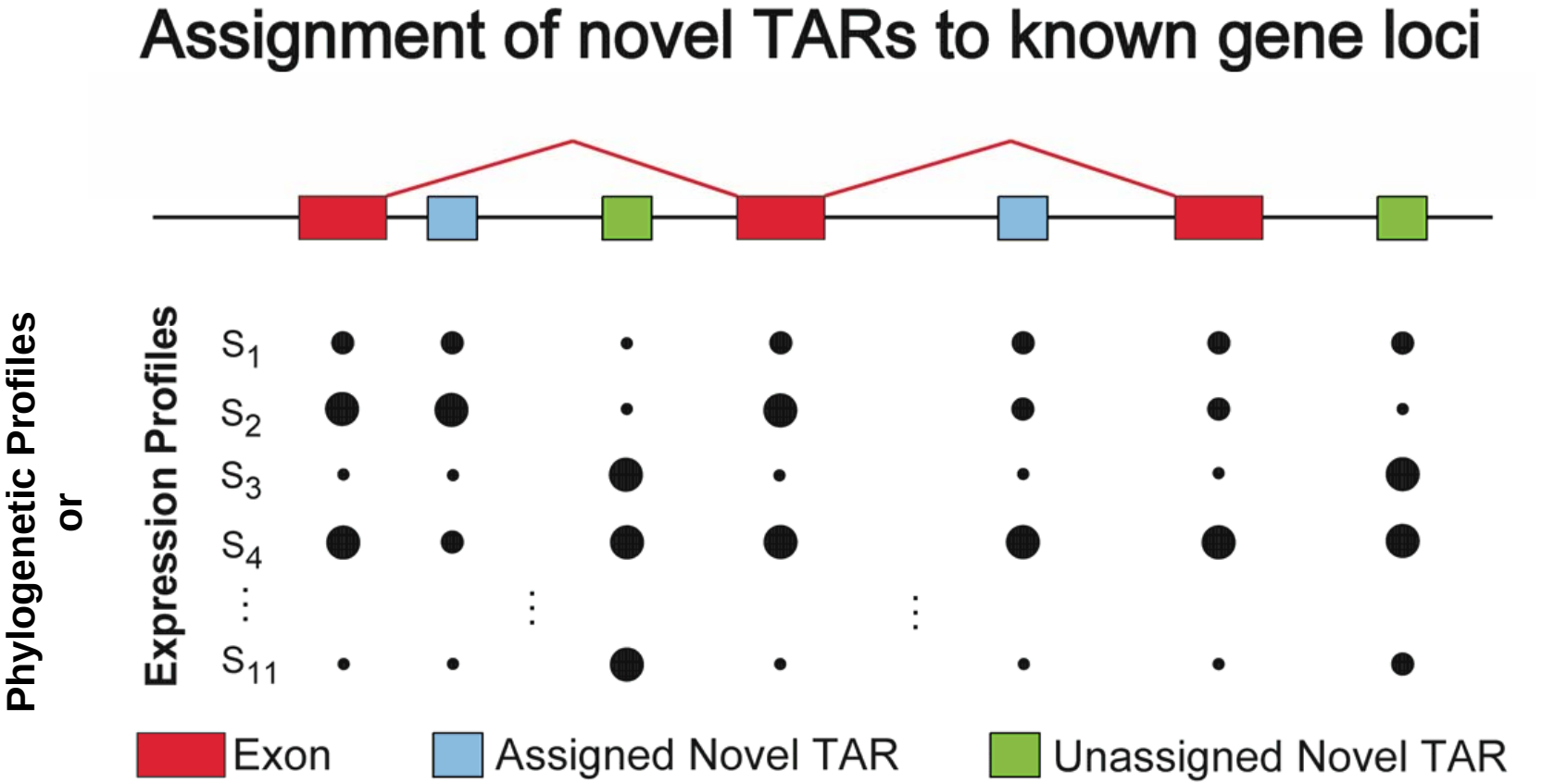


DART Classification System

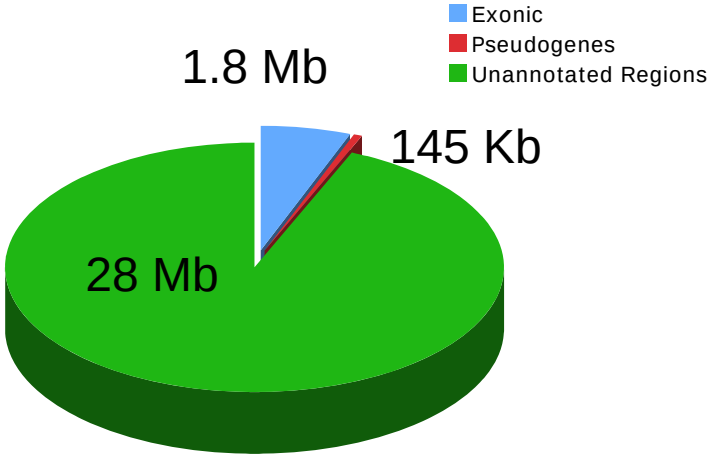
Rozowsky et al. Gen. Res. (2007)

Production | Integrated

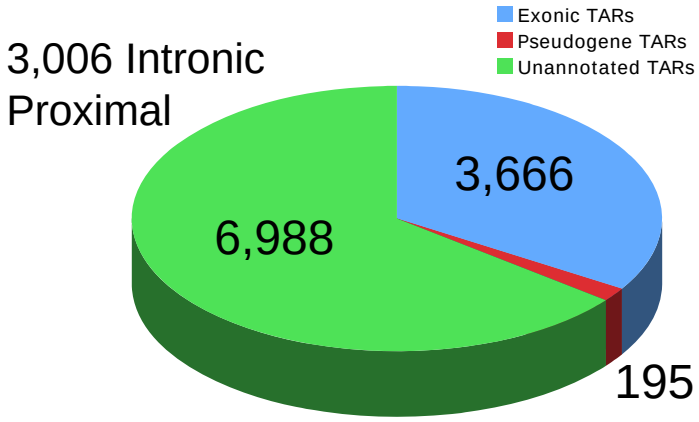
More Developed Annotation: Clustering and Classifying Blocks of Un-annotated Transcription into larger units



Rozowsky et al. Gen. Res. (2007)



ENCODE Regions (30 Mb)



Locations of TARs

Of the approx 7,000 Novel TARs

- 955 are assigned to known genes
- 1,463 are clustered into ~200 Novel Loci

•DART Classification has been experimentally validated with some small scale experiments

◊ RT-PCR & Sequencing

◊ 18/46 (39%) confirmed by RT-PCR

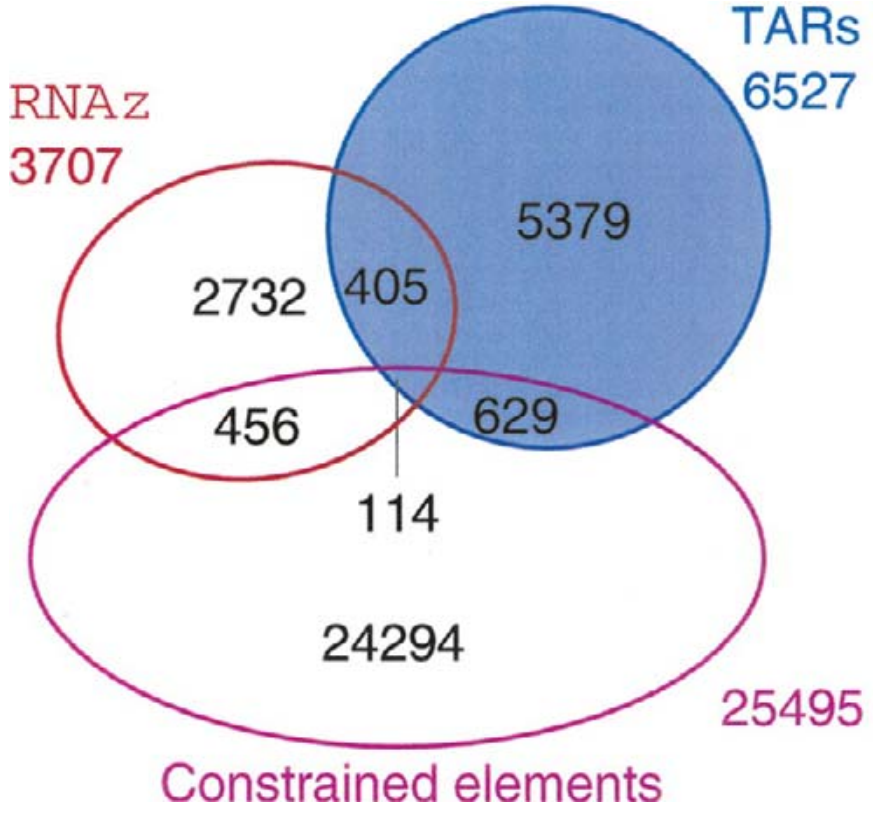
◊ 4/5 Sequenced Products Map uniquely to correct genomic region

Rozowsky et al. Genome Research (2007)

Example predicted structured RNAs (using RNAz)

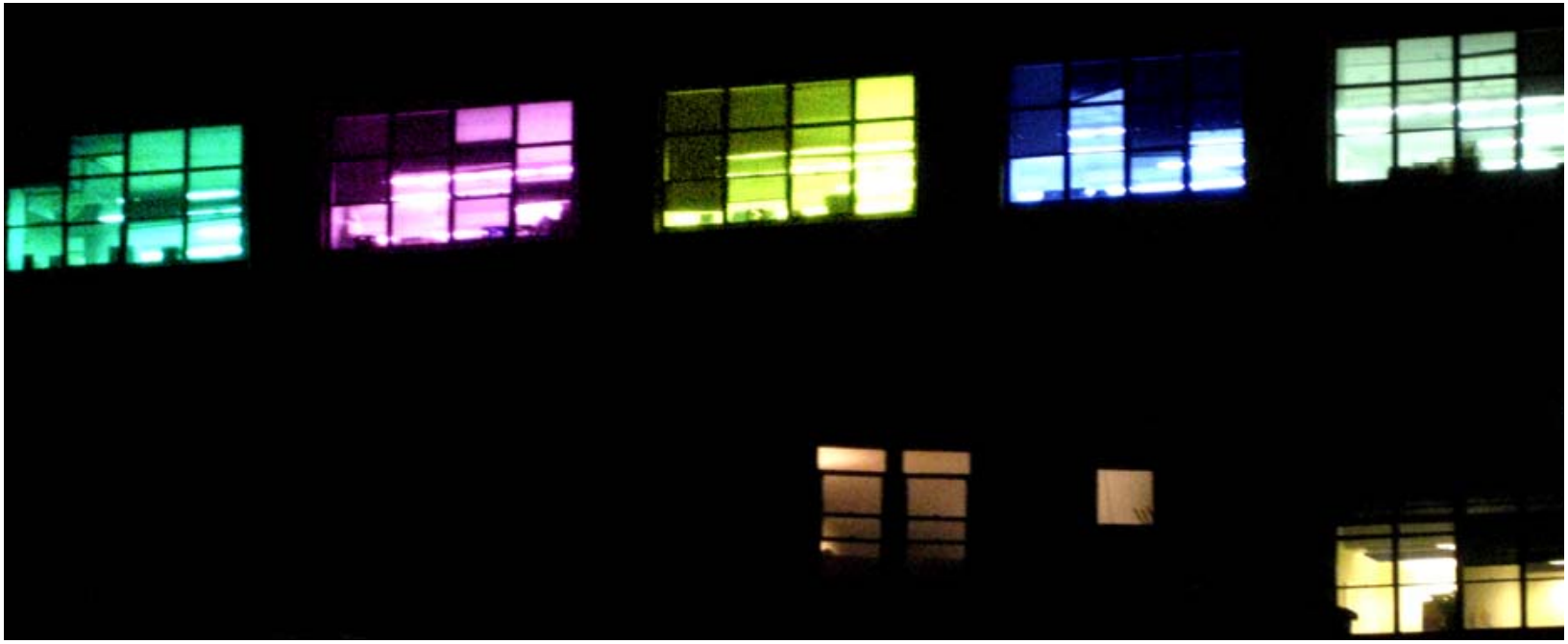
[>700 candidate structured RNAs predicted in 1% of the reference genome]

Overlap of predicted structured RNAs with the union of TARs/Transfrags and the "moderate" set of sequence-constrained elements



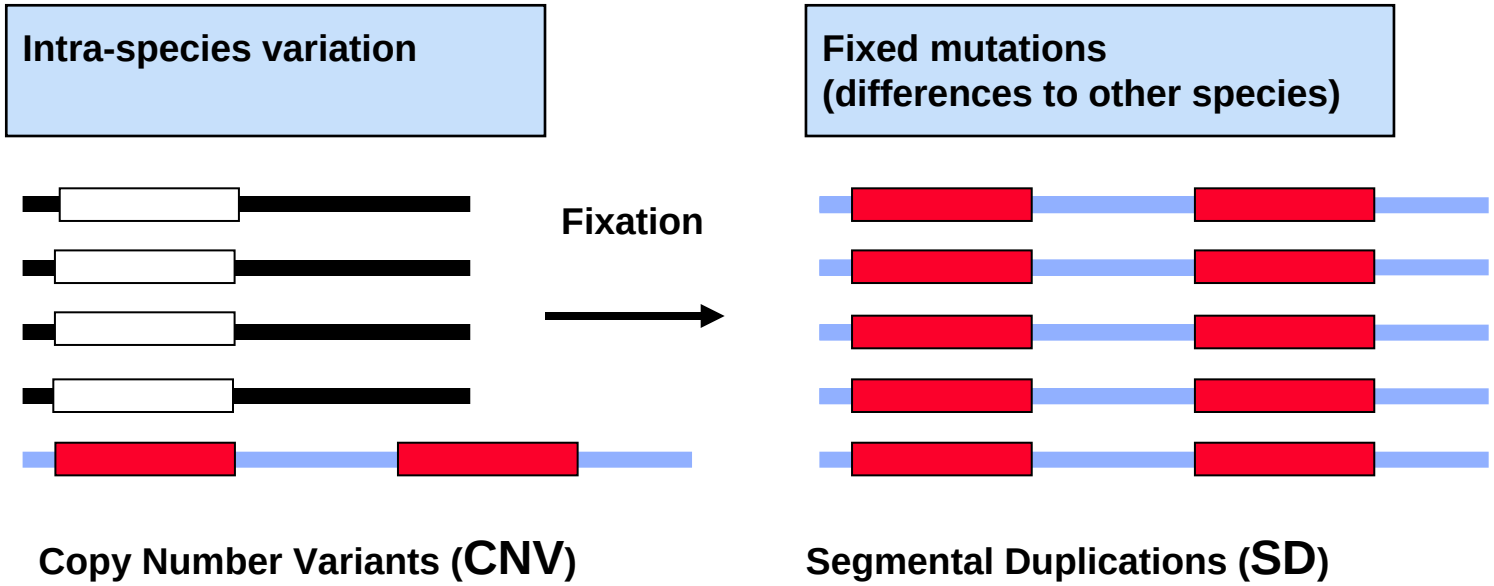
Stefan Washietl, Jakob Pedersen, Jan Korbel *et al.* (2007) *Genome Res* 17:852-864

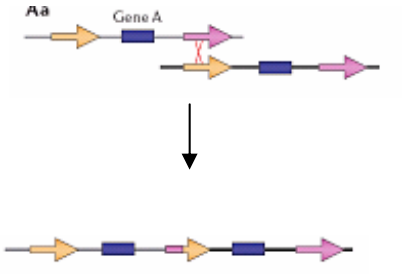
Analyzing Repeated Blocks in the Genome (SDs & CNVs)



pers. **photo**, see streams.gerstein.info

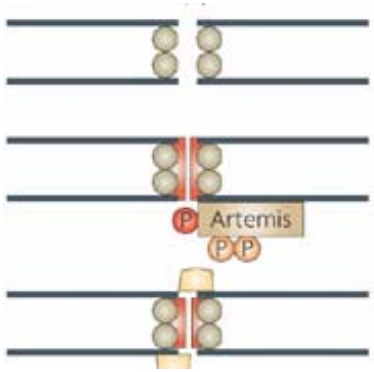
SEGMENTAL DUPLCATIONS AND COPY NUMBER VARIANTS ARE RELATED PHENOMENA AND HAVE BEEN CREATED BY SEVERAL DIFFERENT MECHANISMS





NAHR
(Non-allelic homologous recombination)

Flanking repeat
(e.g. Alu, LINE…)

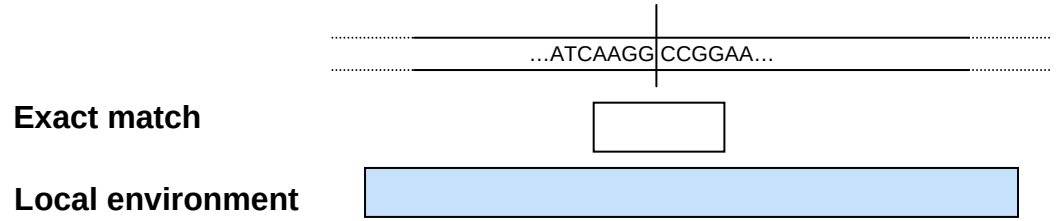


NHEJ
(Non-homologous-end-joining)

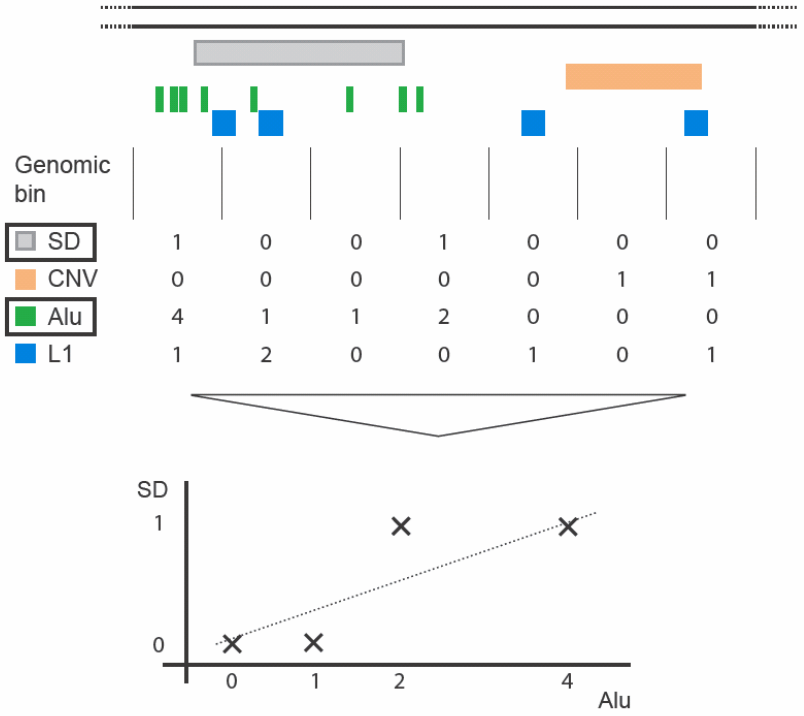
No (flanking) repeats.
In some cases <4bp microhomologies

PERFORM LARGE SCALE CORRELATION ANALYSIS TO DETECT REPEAT SIGNATURES OF SDs AND CNVs

If exact CNV breakpoints are known, we can calculate the enrichment of repeat elements relative to the genome or relative to the local environment

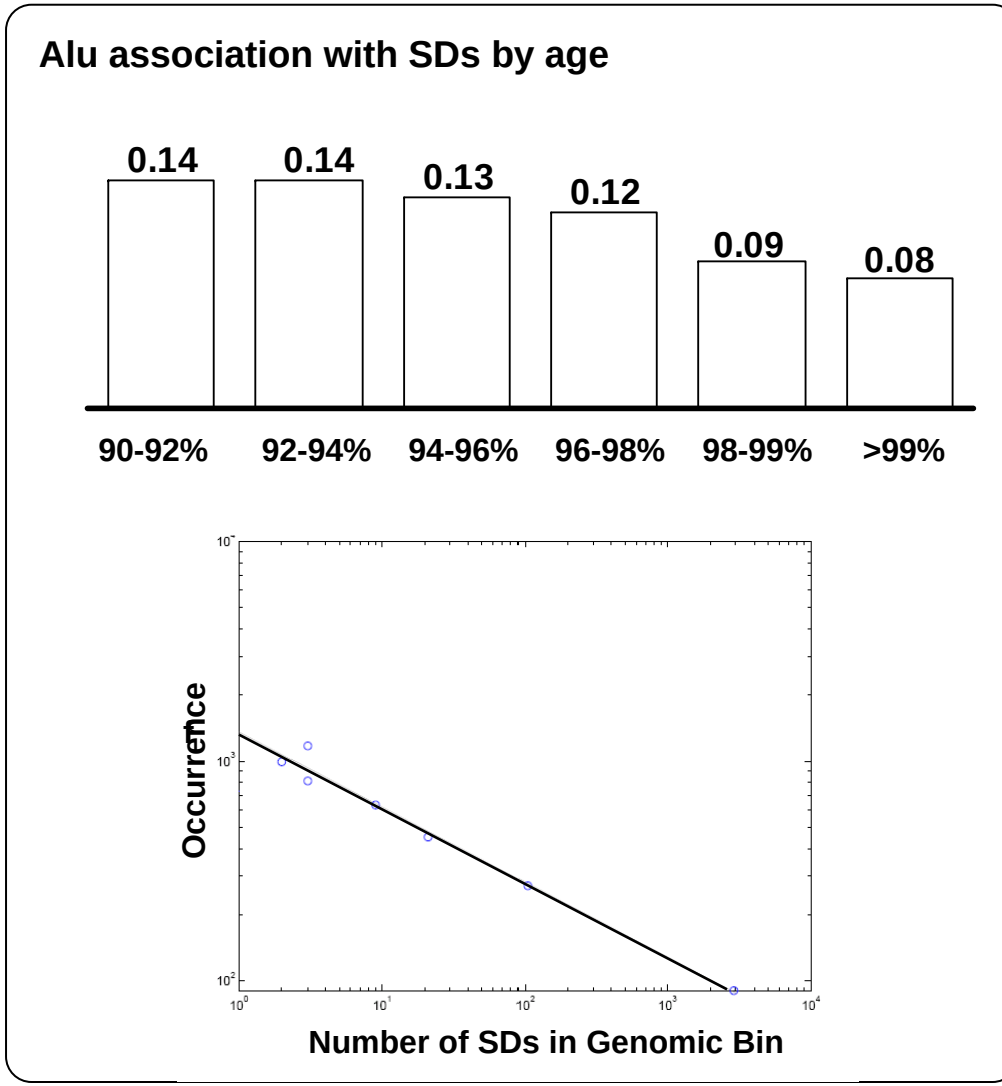


1. Survey a range of genomic features
2. Count the number of features in each genomic bin (100kb)
3. Calculate correlations / enrichments using robust stats



[Kim et al. Gen. Res. (in press, '08), arxiv.org/abs/0709.4200v1]

SDs ARE CORRELATED WITH ALUS AND OTHER SDs



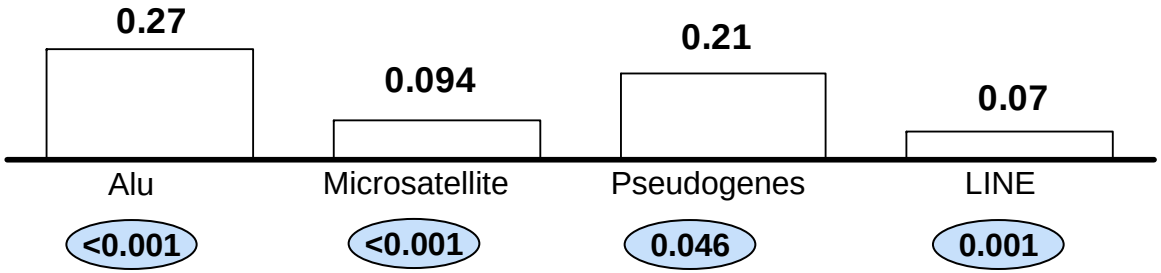
- The co-localization of Alu elements with SDs is highly significant.
- Older SDs have a much higher association with Alus than younger SDs.
- SDs can mediate NAHR and lead to the formation of CNVs
- Such mechanisms ("preferential attachment") are well studied in physics and should leads a very skewed ("power-law") distribution of SDs.
- Hotspots

[Kim et al. Gen. Res. (in press, '08), arxiv.org/abs/0709.4200v1]

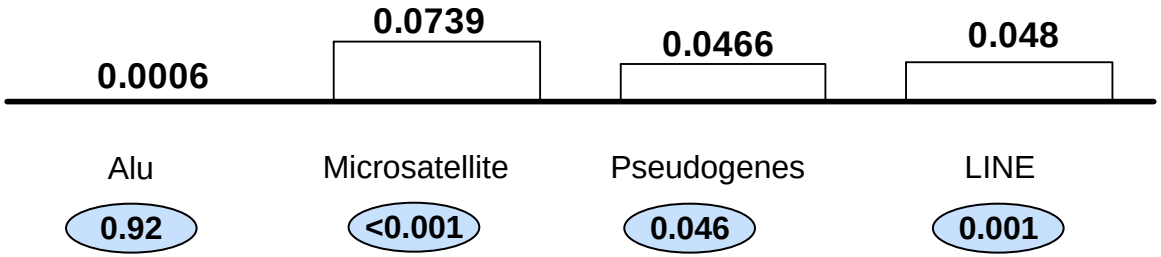
080907_SD_CNV_Slides_MBG_CEGS_PMK

ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs

SD association with repeats

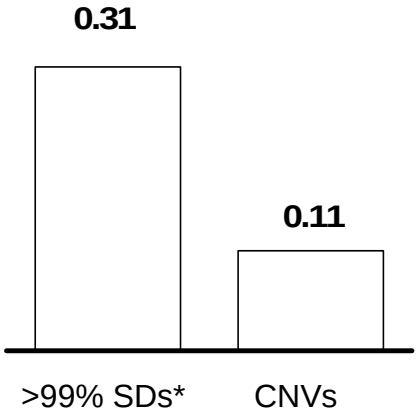


CNV association with repeats



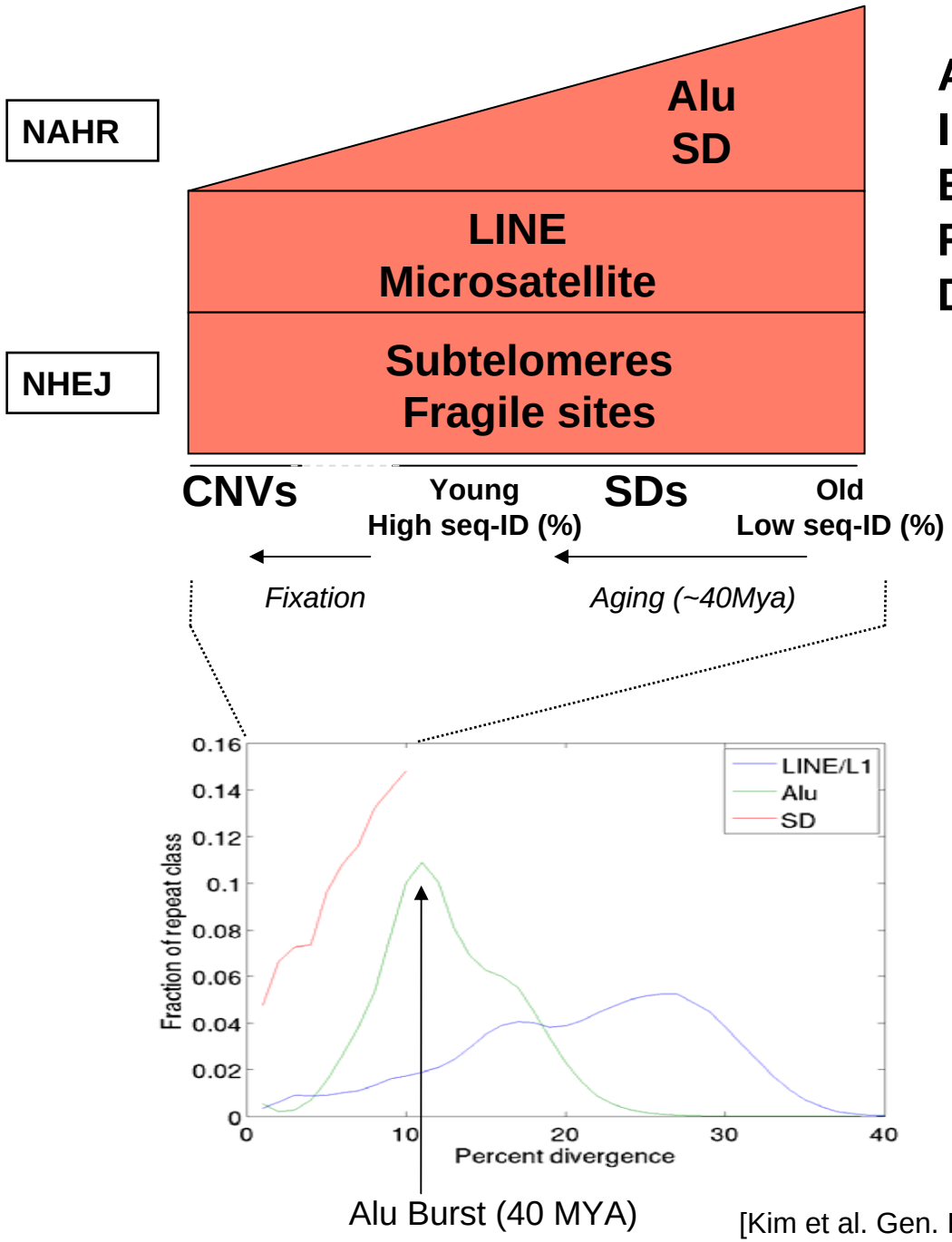
CNVs ARE LESS ASSOCIATED WITH SDs THAN THE GENERAL SD TREND

CNV Association with SDs



[Kim et al. Gen. Res. (in press, '08), arxiv.org/abs/0709.4200v1]

59



AFTER THE ALU BURST, THE IMPORTANCE OF ALU ELEMENTS FOR GENOME REARRANGEMENT DECLINED RAPIDLY

- About 40 million years ago there was a burst in retrotransposon activity
- The majority of Alu elements stem from that time
- This, in turn, led to rapid genome rearrangement via NAHR
- The resulting SDs, could create more SDs, but with Alu activity decaying, their creation slowed

[Kim et al. Gen. Res. (in press, '08), arxiv.org/abs/0709.4200v1]

Summary

pers. **photo**, see
streams.gerstein.info

Overview of the Process of Intergenic Annotation

- Basic Inputs
  1. Doing large-scale similarity comparison, looking for repeated or deleted regions
  2. Determining experimental signals for biochemical activity (e.g. transcription) across each base of genome
- Results of Analyzing Similarity Comparison
  1. Finding large repeated or deleted blocks (e.g. CNVs) as a function of degree of similarity
     1. within reference human genome
     2. within human population
     3. between related organisms (e.g. mouse)
  2. Finding smaller "exon-level" similarities (e.g. pseudogenes)
- Results of Processing Raw Expt. Signals
  1. Signal Processing: removing artifacts, normalizing, window averaging
  2. Segmenting signal into larger "hits" ("Active Regions" or ARs)
  3. Clustering together active regions into even larger features at different length scales and classifying them
  4. Building networks and beyond....

Segmenting the Raw "Signal" from Next-generation Sequencing into Usable Annotation Blocks

- MSB
  - ◊ Mean-shift segmentation approach following grad. of PDF
  - ◊ Equally applied to aCGH and depth of coverage of short reads
- PeakSeq
  - ◊ Scoring chip-seq expt relative to input control
  - ◊ Simulating chip-seq expt anticipates & allows correction for non-uniformity

First-Pass Annotation Clustering and Characterizing Novel Transcribed Regions and Groups of Binding Sites

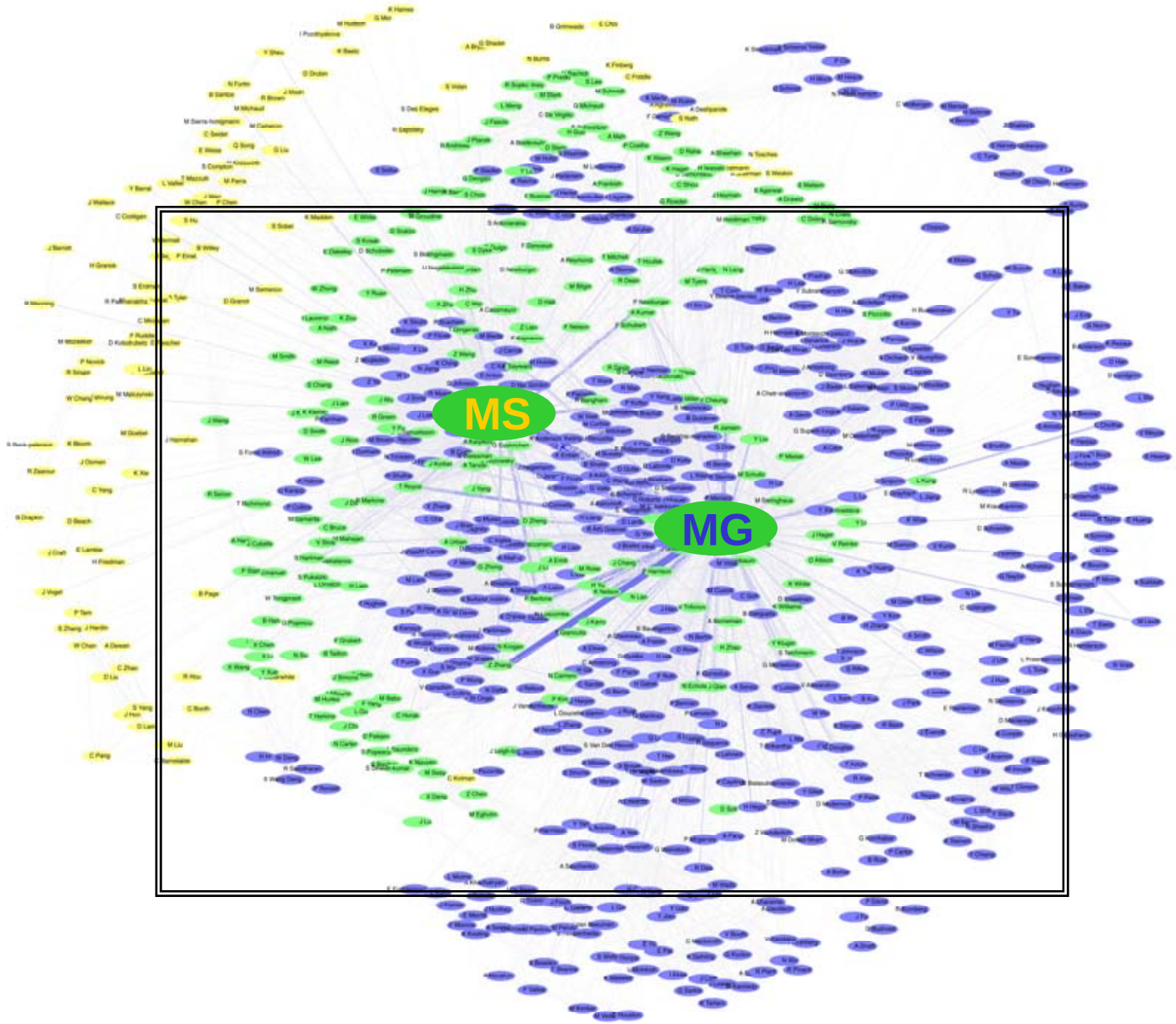
- Deserts and Forests of Binding Activity
  - ◊ on ~50kb scale
  - ◊ Biplot gives broad separation of seq. specific and non-specific factors and associated genomic bins
- Analyzing Promotors
  - ◊ BoCaTFBS: Refining binding site motifs based on the results of chIP-chip experiments
- DART classification of TARs
  - ◊ 1300 TARs in ~200 novel pilot ENCODE loci
    - based on expression and phylogenetic clustering

Analysis of Duplication in the Genome: SVs and SDs

- Large-scale analysis of existing CNVs & SDs in human genome
- SDs assoc. with Alu, pseudogenes and older SDs
- CNVs assoc. other repeats (microsat.) and not as much with SDs
- Suggestion: Alu burst 40 MYA triggered much NAHR rearrangement, then dupl. feed on itself in hotspots but now dying down and NAHR assoc. with other repeats and CNVs also from NHEJ

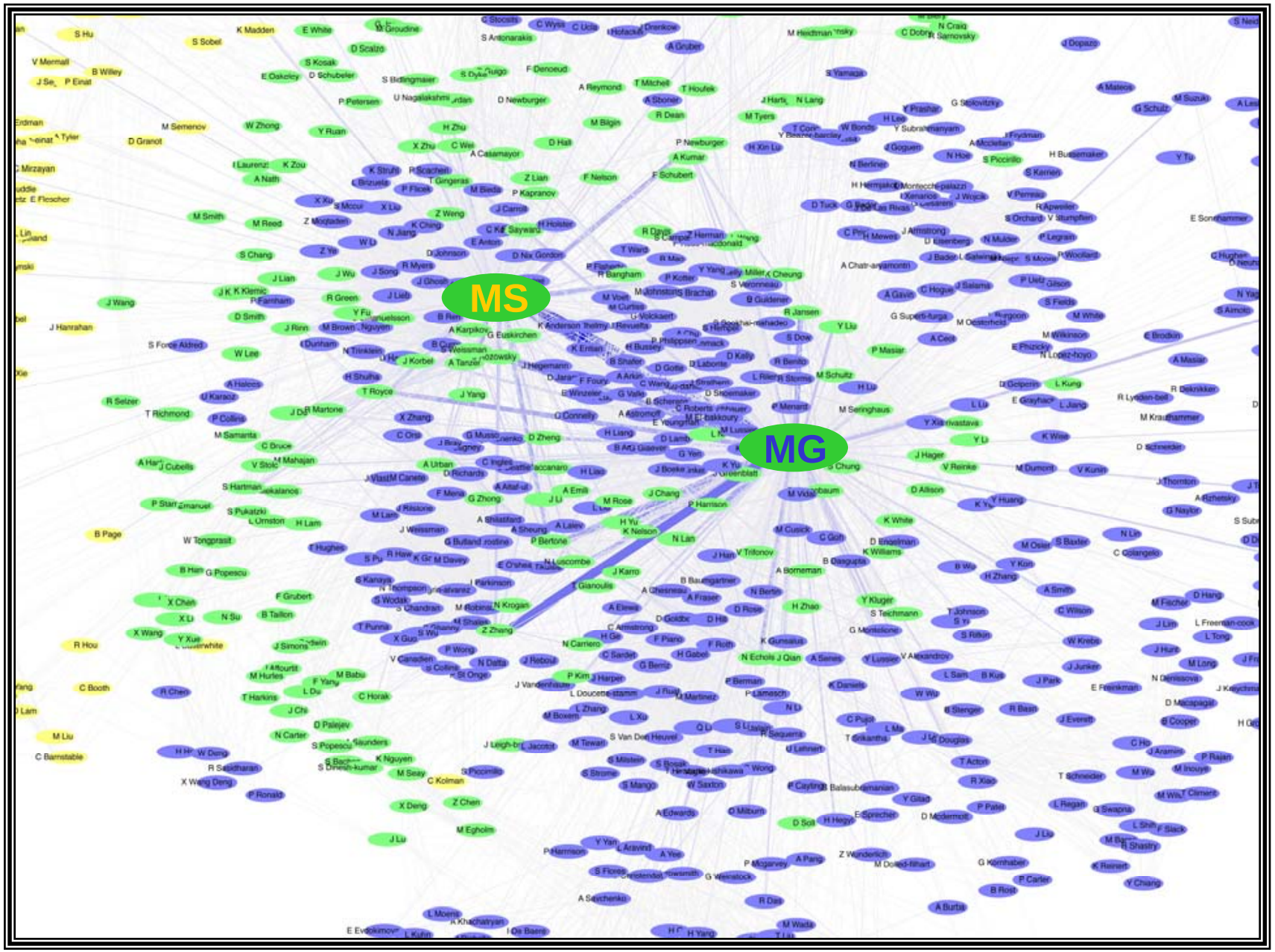
Acknowledgements

pseudogene.org, tiling.gersteinlab.org, sv.gersteinlab.org



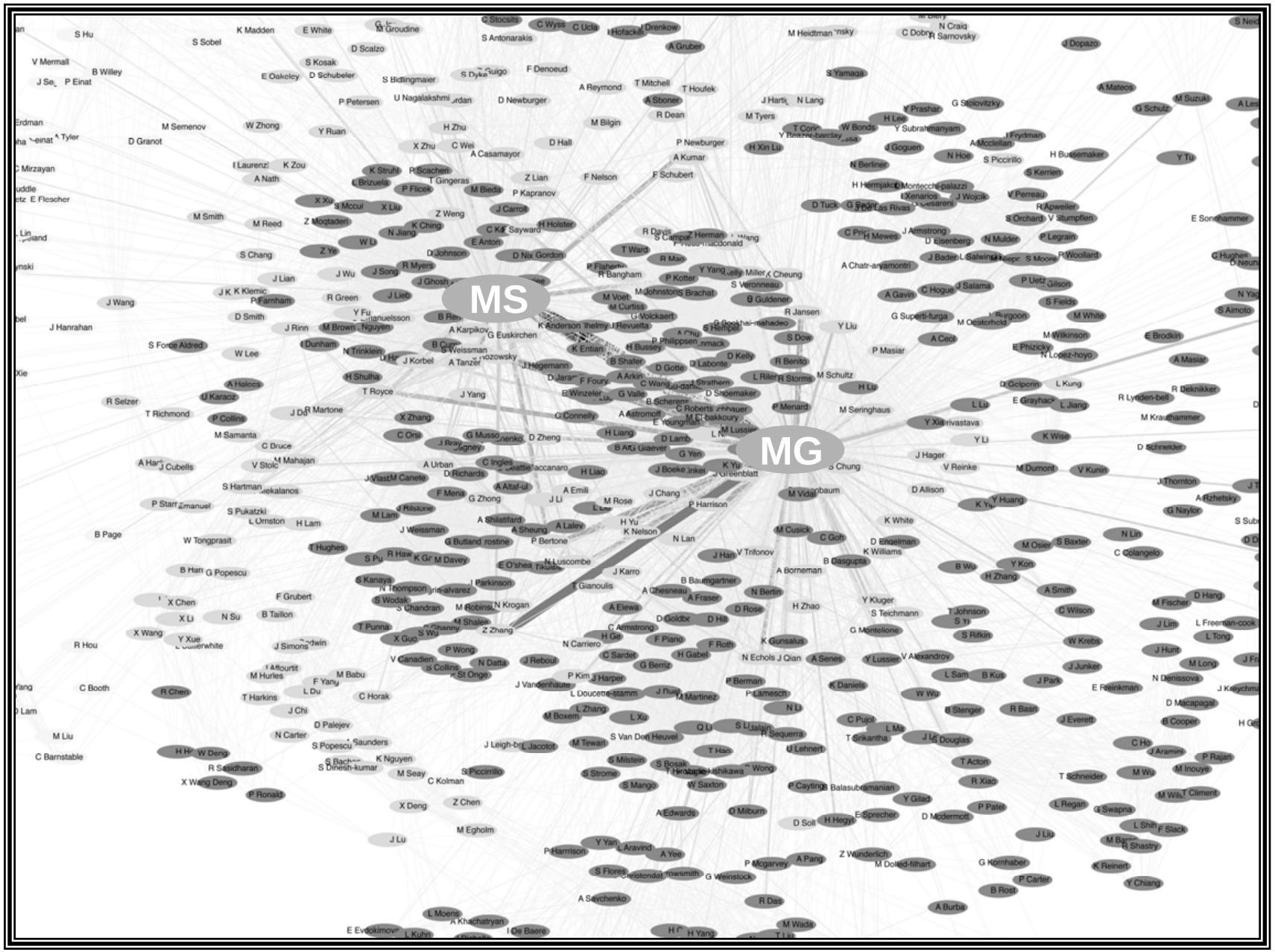
Acknowledgements

pseudogene.org, tiling.gersteinlab.org, sv.gersteinlab.org



Acknowledgements

pseudogene.org, tiling.gersteinlab.org, sv.gersteinlab.org



Acknowledgements

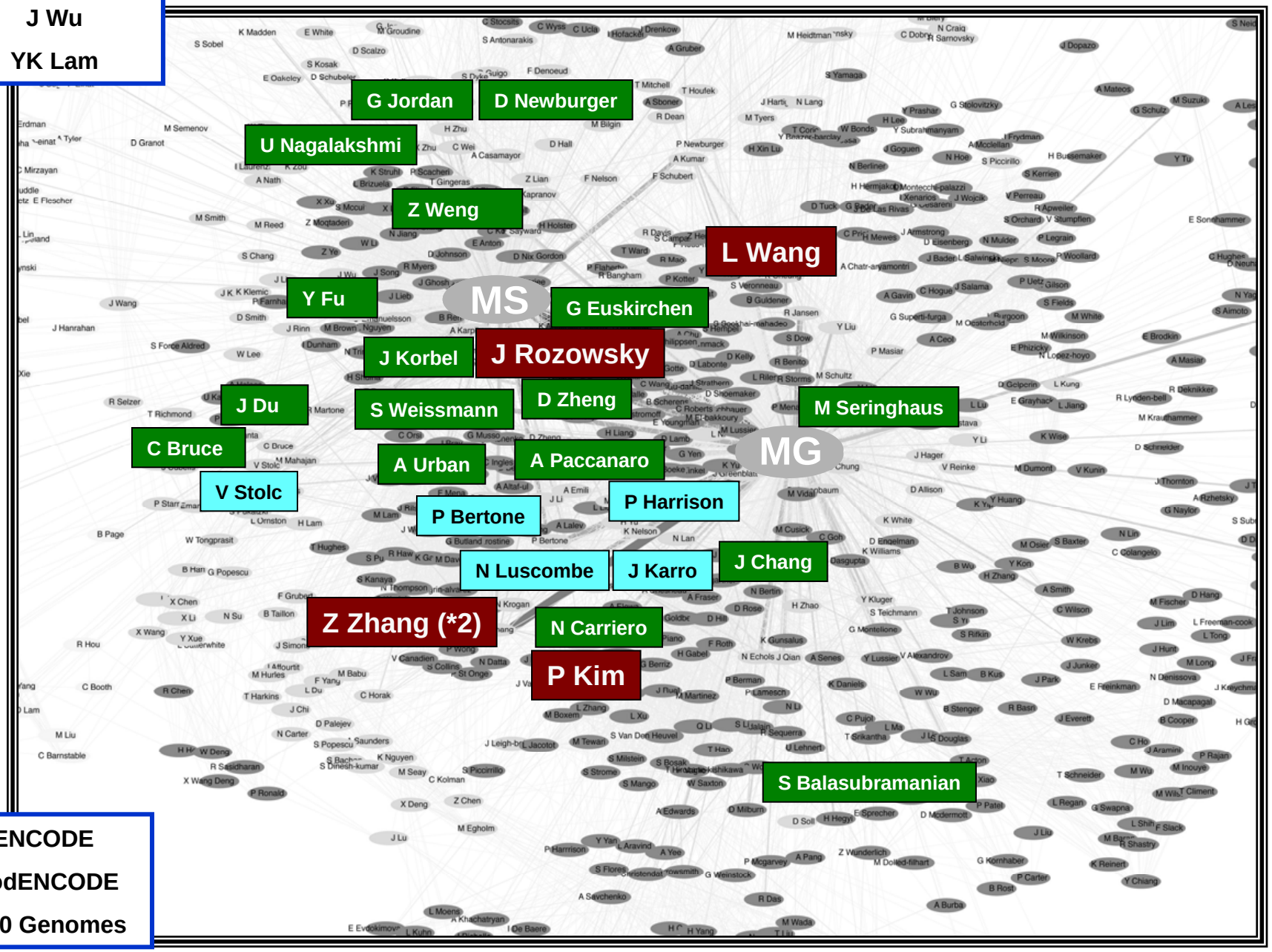
pseudogene.org, tiling.gersteinlab.org, sv.gersteinlab.org

X Mu

A Abyzov

J Wu

YK Lam



ENCODE

modENCODE

1000 Genomes

Permissions Statement