Annotation Non-coding Regions of the Human Genome

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

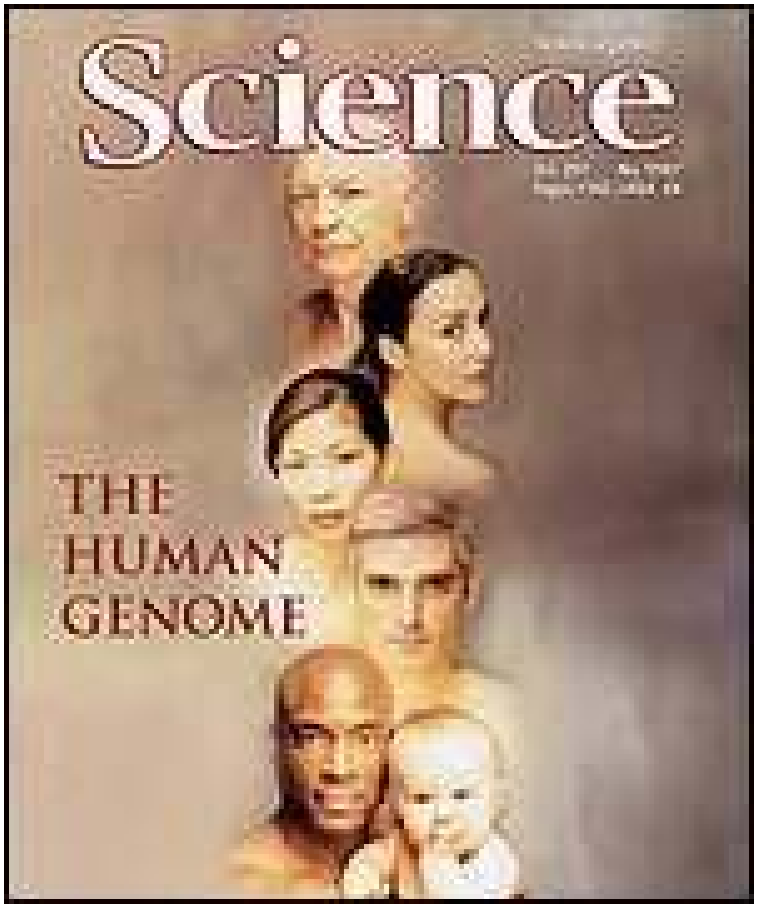
Bristol Myers

2008.08.15, 14:20-14:45

Slides downloadable from Lectures.GersteinLab.org.

(Please read permissions statement.)

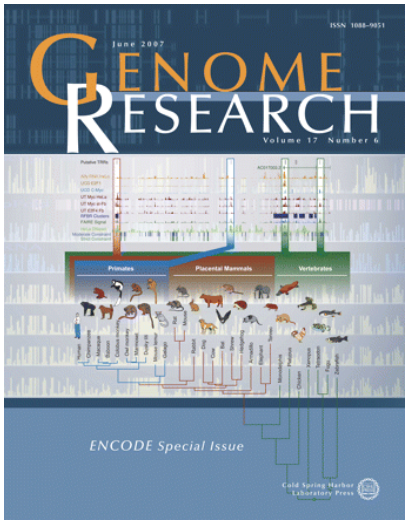
Paper references mostly from Papers.GersteinLab.org.



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]

3 Lectures.GersteinLab.org (c) 2007



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]

4 Lectures.GersteinLab.org (c) 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.

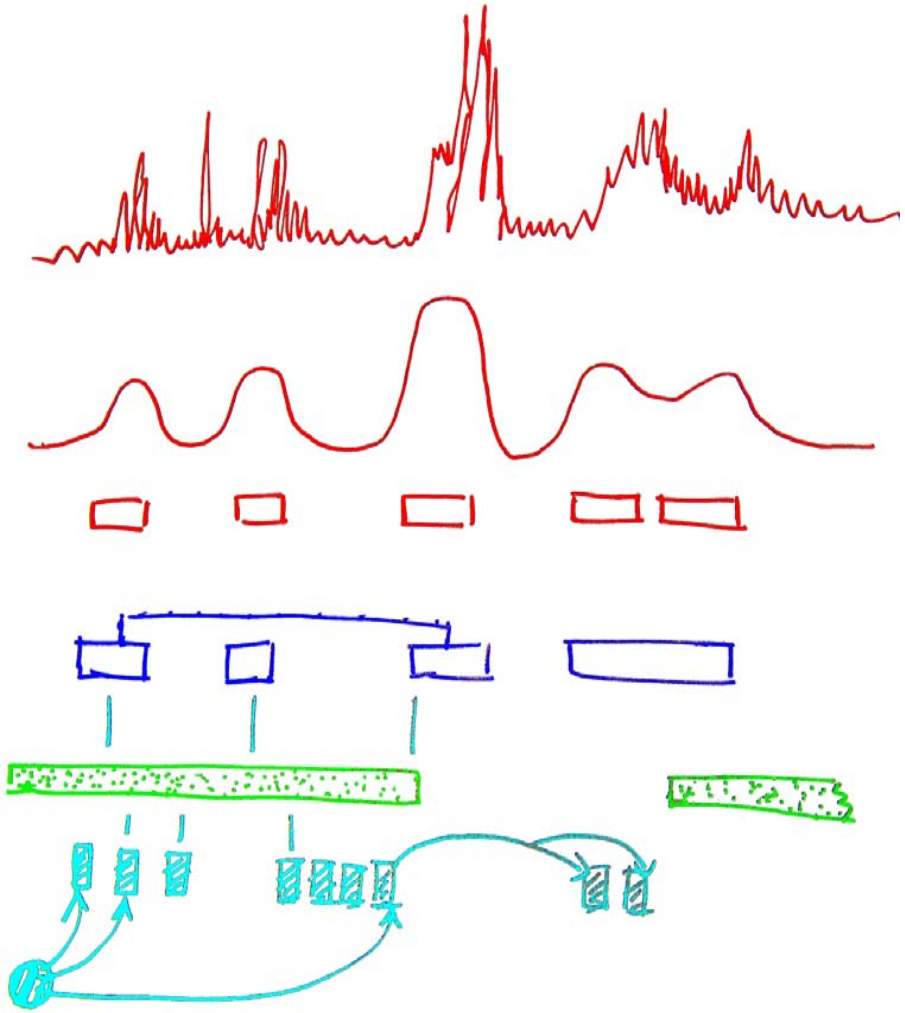
There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be 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-

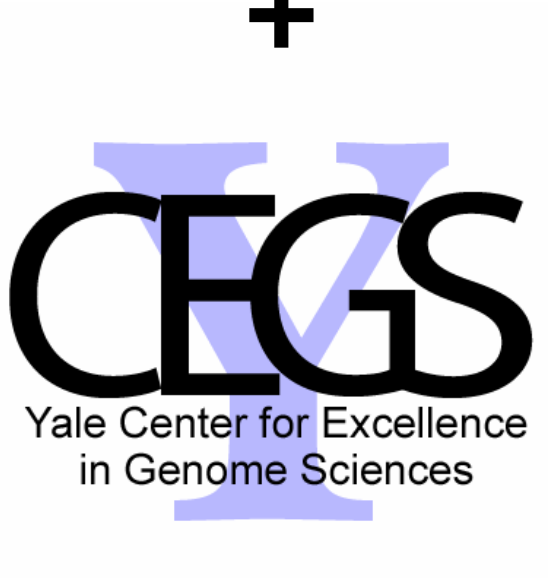
Overview of Annotation Process



- **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**

8 Lectures.GersteinLab.org (c) 2007

ENCODE & mod ENCODE @ Yale +



- Array and NextGen Seq. Experiments
  - Mike Snyder & Sherman Weissman
  - Interrogation of small fragments of chromosomes to determine their function
    - Large-scale hybridization to find transcribed regions in unbiased fashion
    - TF binding sites (via ChIP-chip)
    - CNVs and SDs (from hires-aCGH)
- 1st Pass Computational Annotation
  - Classification of Novel Transcribed Regions
  - Grouping and Classification of Binding Sites
  - Characterization of SDs and CNVs
- Integrative Annotation
  - Pseudogenes (Zheng et al., GR,GB)
    - Inter-relation with Transcription & CNVs

10 Lectures.GersteinLab.org (c)

A Starting Point: Raw Signal from Tiling Arrays and NextGen Sequencing Tags

11 Lectures.GersteinLab.org (c)

High-Resolution CGH with Oligonucleotide Tiling Microarrays

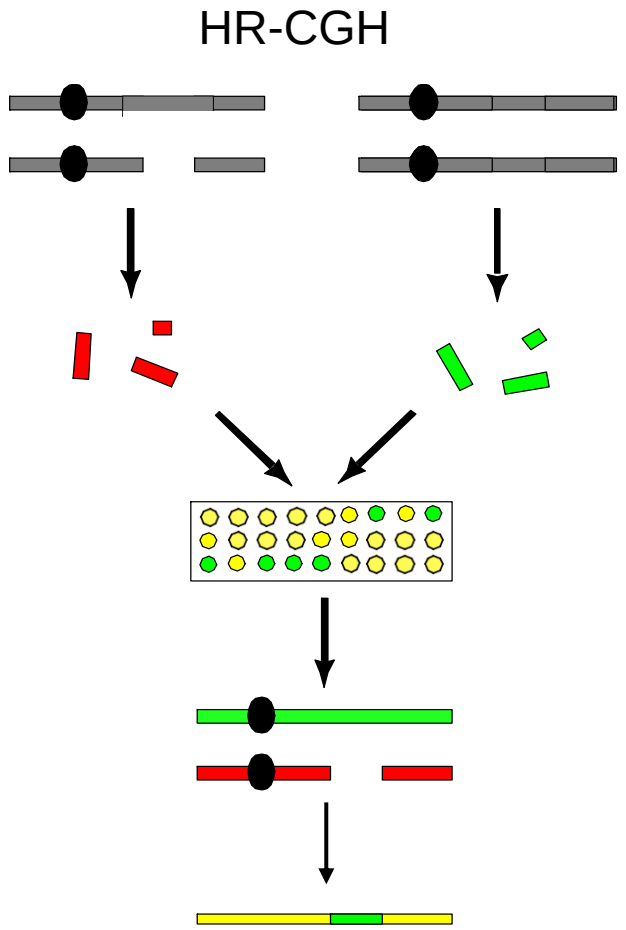
Maskless Array Synthesis

385,000 oligomers/chip

Isothermal oligomers, 45-85 bp

Tiling at ~1/100bp non-repetitive genomic sequence

Detects CNVs at 1 kb resolution

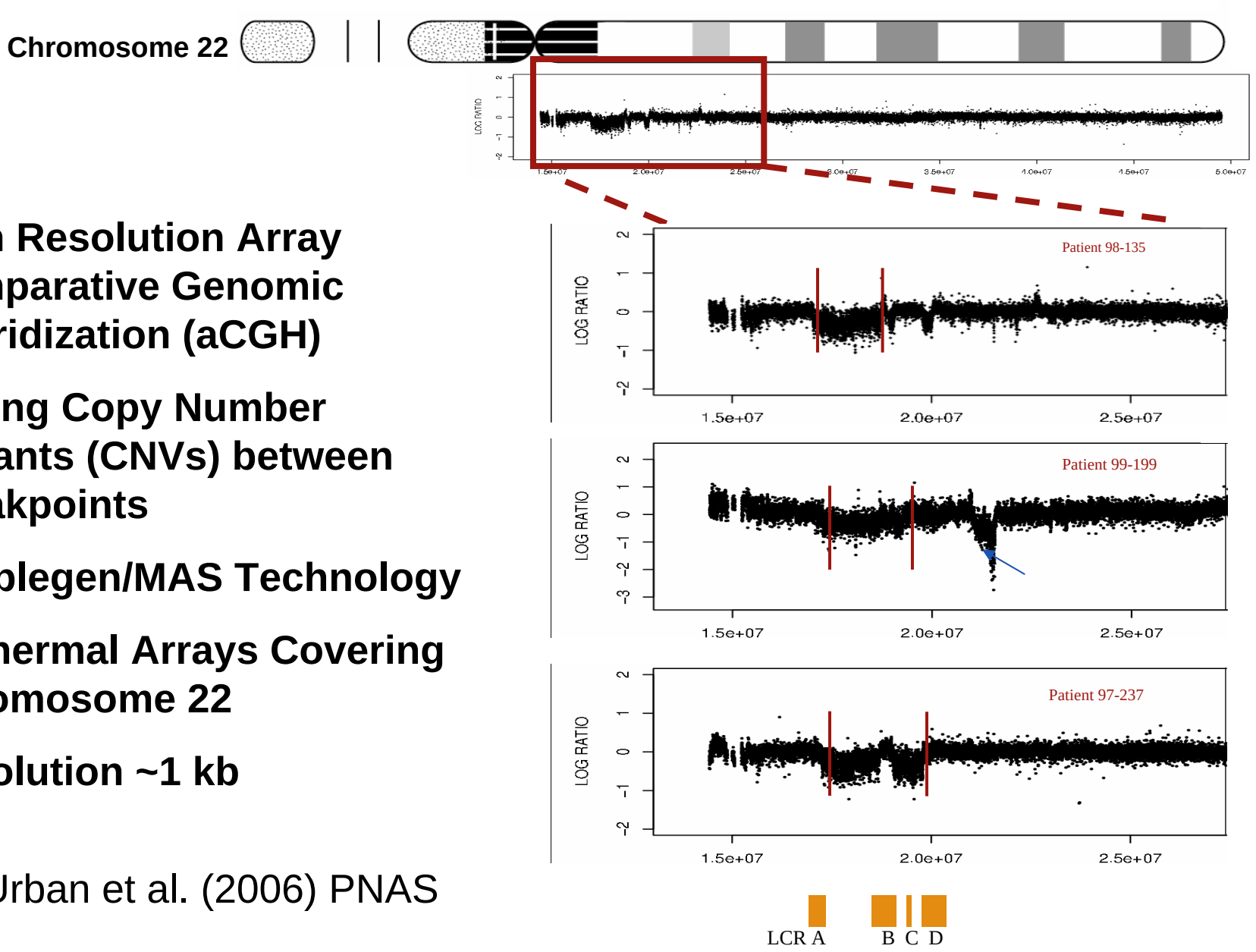


Urban et al., 2006

Nimblegen

12 Lectures.GersteinLab.org (c)

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

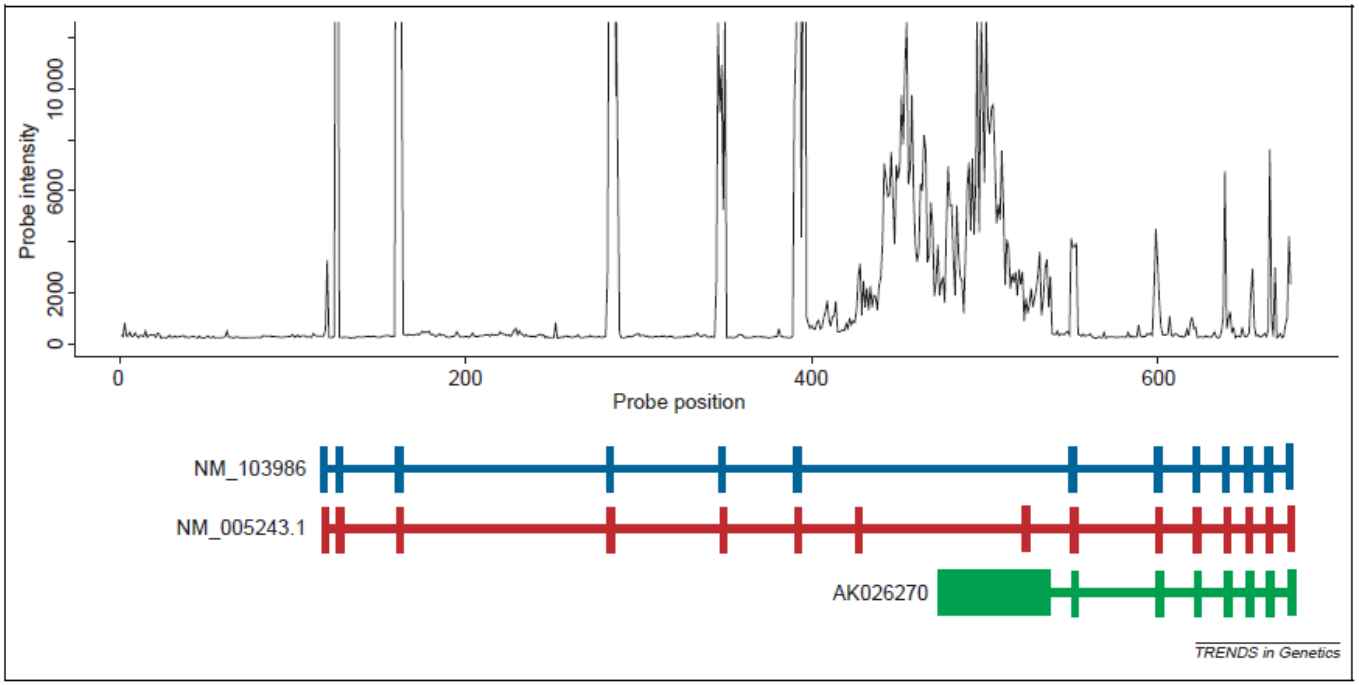
Urban et al. (2006) PNAS

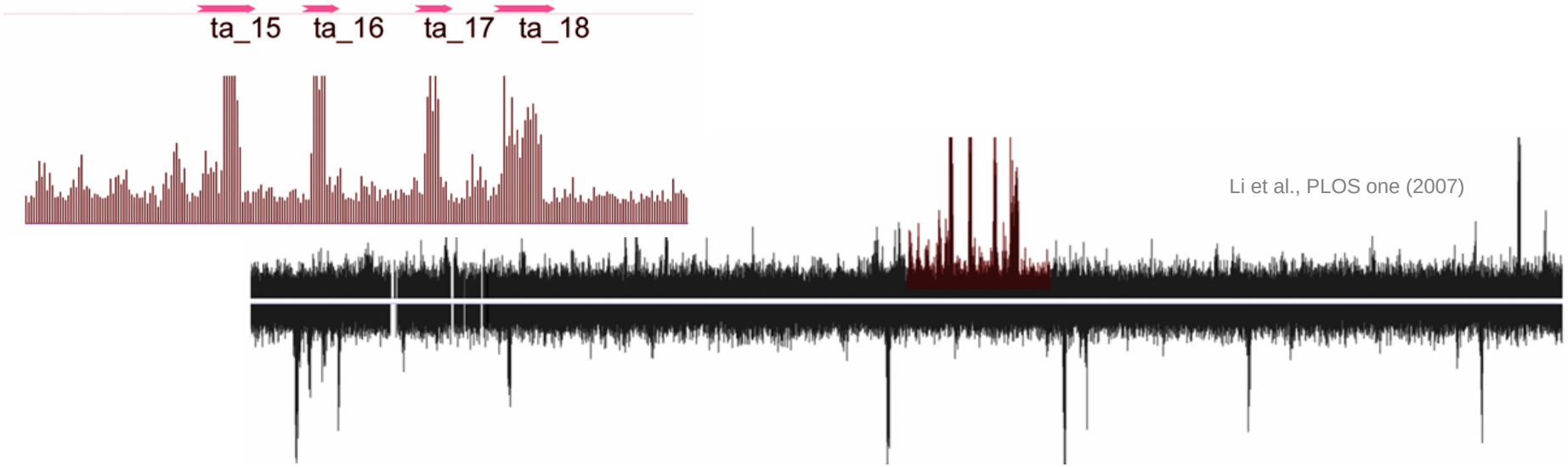
13 Lectures.GersteinLab.org (c)

Tech | Production

A Starting Point: Noisy Raw Signal from Tiling Arrays (Transcription)

Johnson et al. (2005) TIG, 21, 93-102.

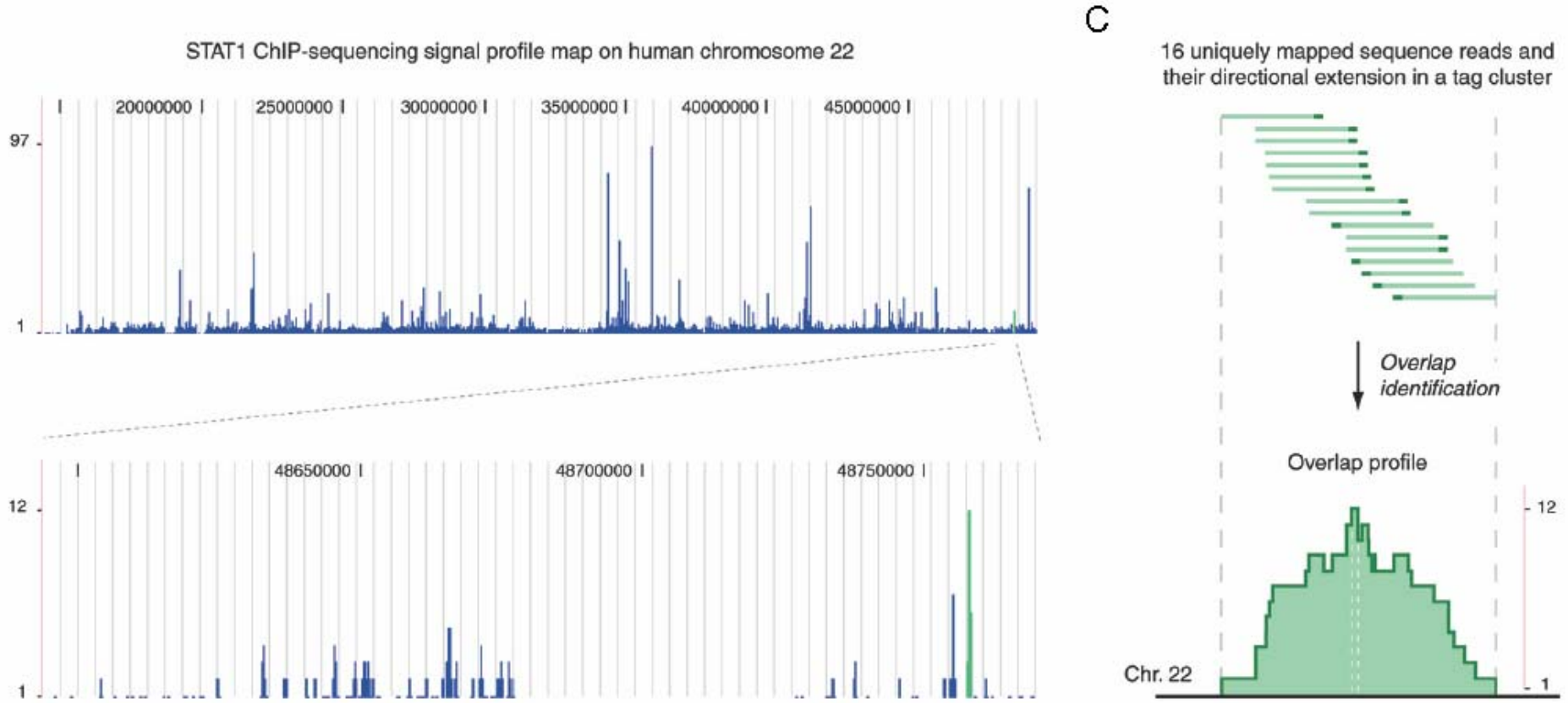




Li et al., PLOS one (2007)

14 Lectures.GersteinLab.org (c)

Representative Signal from Chip-Seq



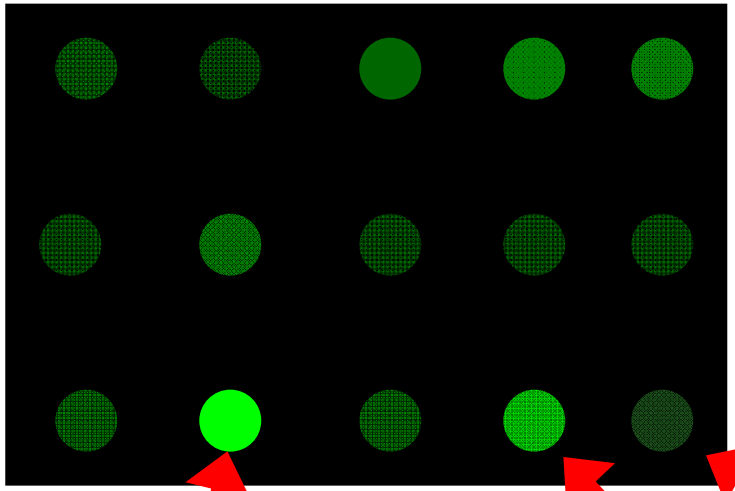
[Robertson et al., Nat. Meth. ('07); Zhang et al. PLOS Comp. Bio. (in revision, '08)]

15 Lectures.GersteinLab.org (c)

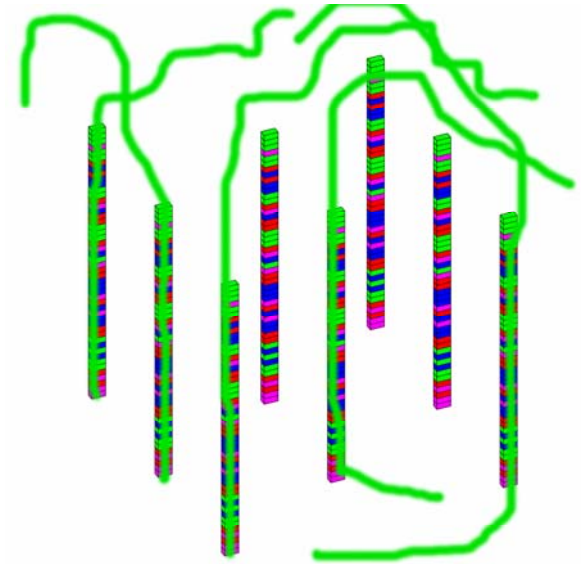
Signal Processing: Normalizing, Measuring & Correcting for Aspects of Hybridization)

36 Lectures.GersteinLab.org (c)

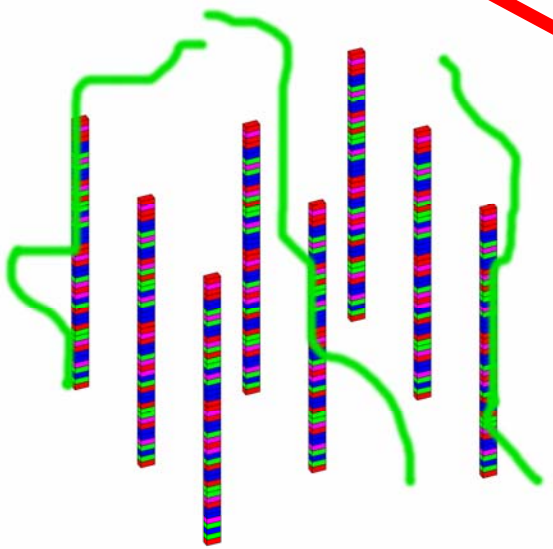
Specific & Non-specific Cross-Hyb.



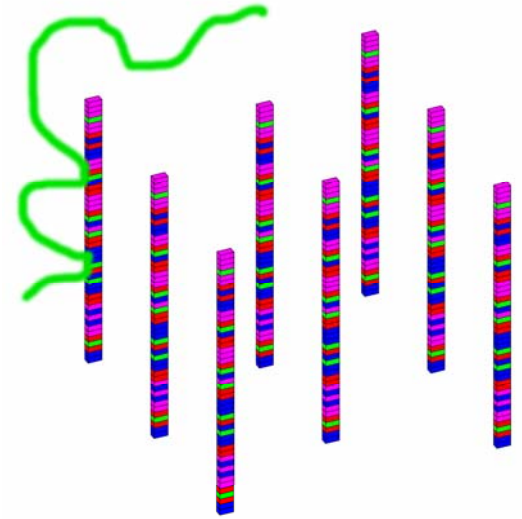
- Perfect match (PM): probe binding intended target
- Specific cross-hyb.: probes binding non-PM targets with a small number of mismatches
- Non-specific cross-hyb.: probes binding targets with many mismatches, due to general stickiness of oligos



Perfect Match



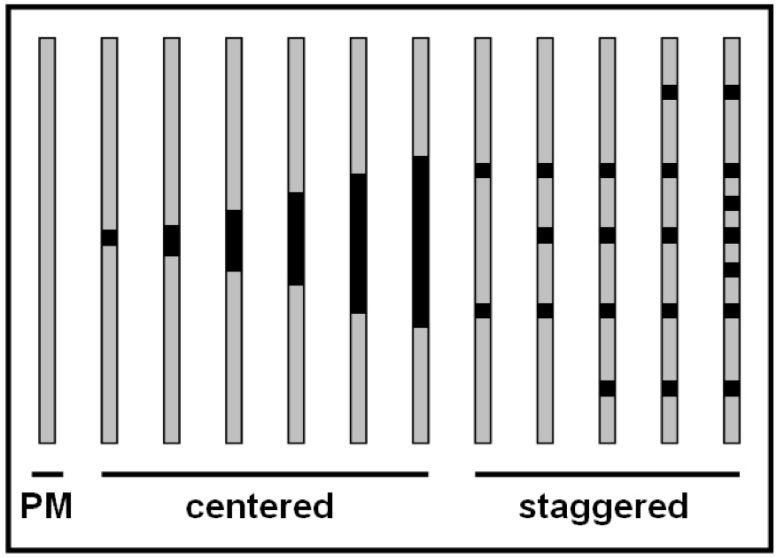
Specific Cross-hyb.



Non-specific Cross-hyb.

Creation of Standardized Datasets for Quantifying Effect of Mismatches

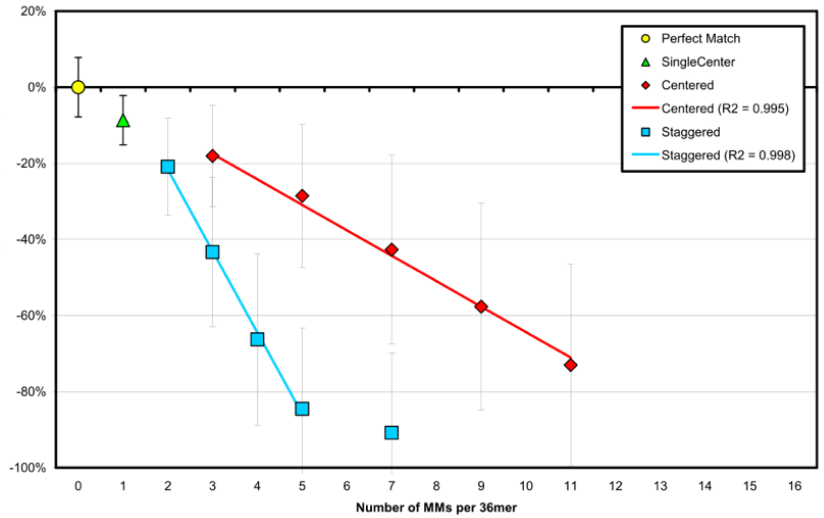
[Seringhaus et al., BMC Genomics (in press)]

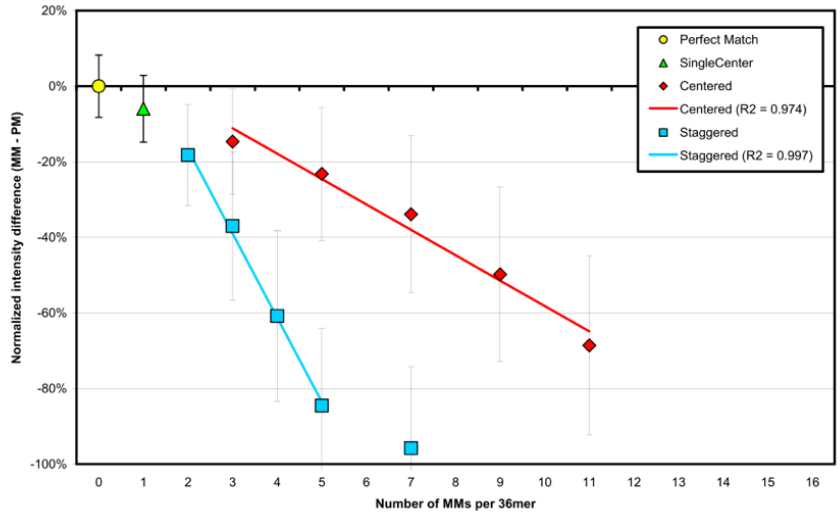


Yeast ACT1 Gene

Human HBG2 Gene

Normalized Intensity
Chg. vs. PM

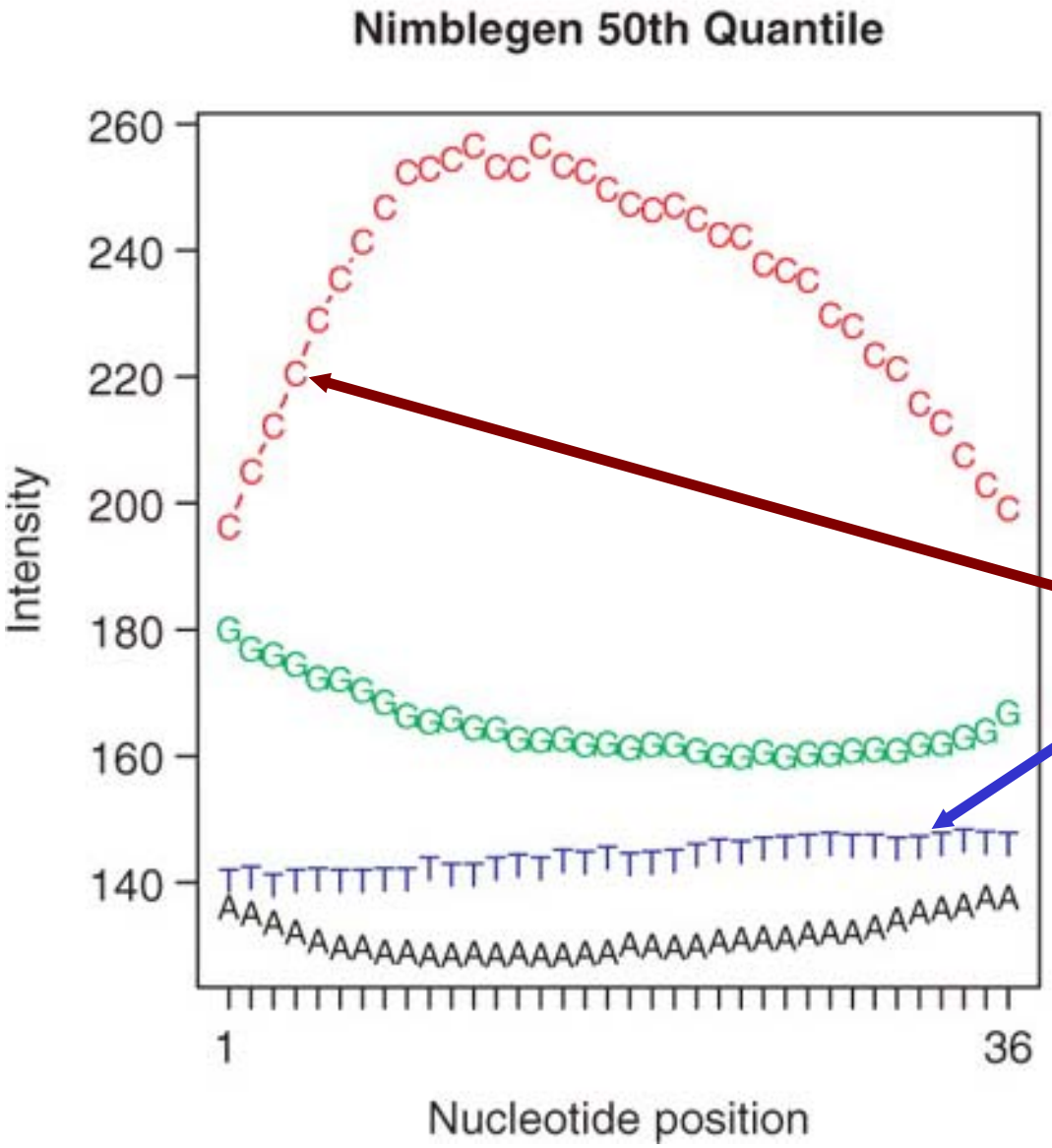




Number of Mismatches

38 Lectures.GersteinLab.org (c)

Observing Non-specific Cross-hyb. (Probe sequence effects)



Avg. intensity of all background probes with a C at position 4

Avg. intensity of all background probes with a T at position 33

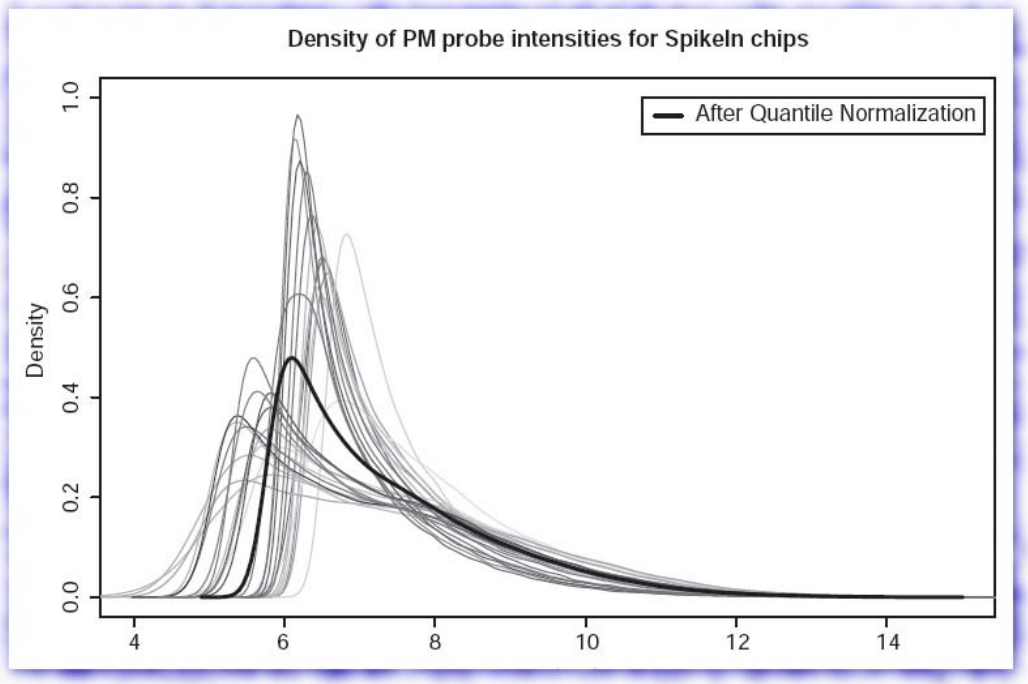
Source: Royce, T.E., et al (2007), *Bioinformatics*, **23**, 988-97

42 Lectures.GersteinLab.org (c)

Quantile Normalization

Gene expression quantile normalization

- Quantile normalization has proven to be the most effective way to normalize replicate gene expression arrays.



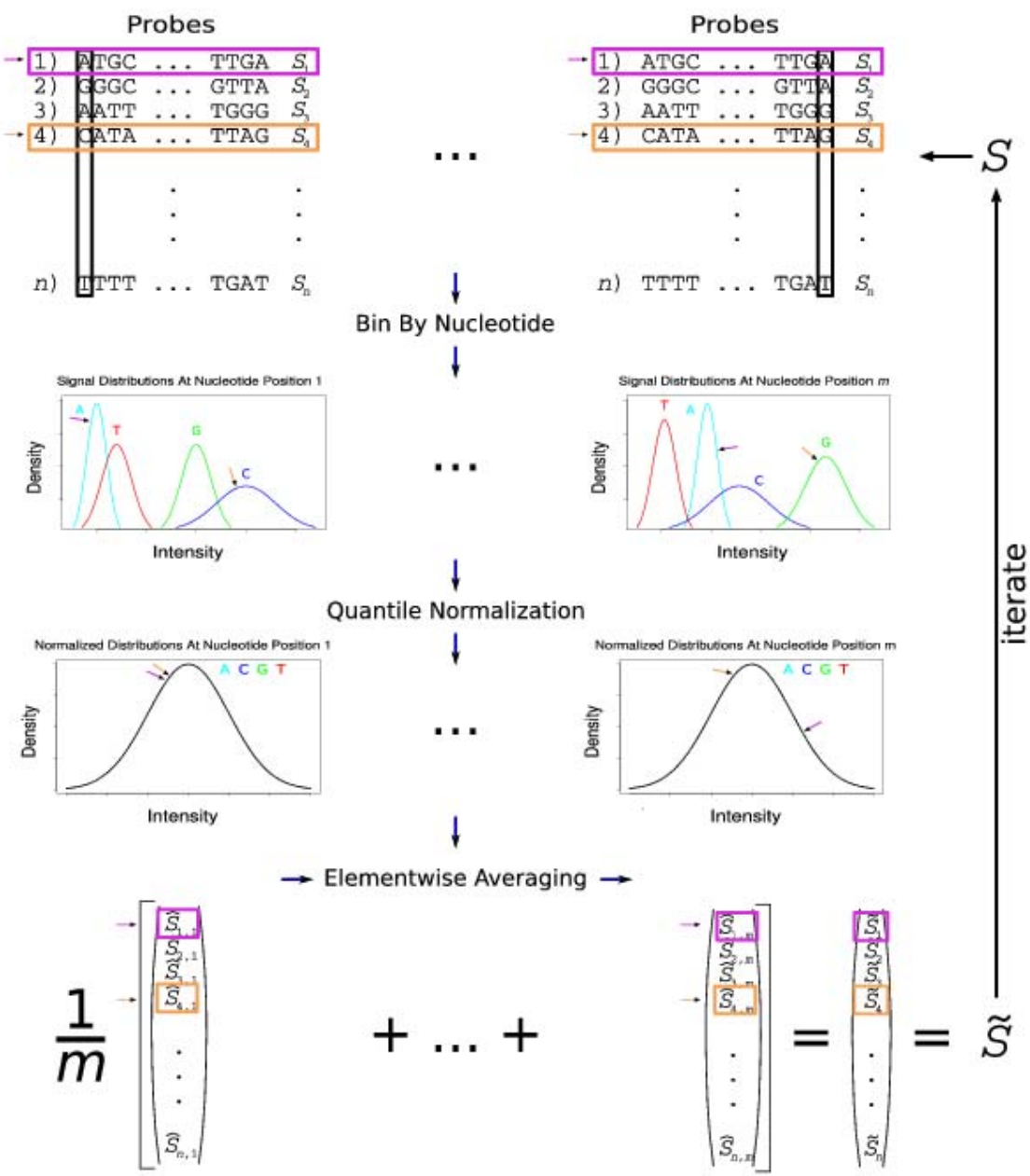
Source: Bolstad, B.M., et al (2003), *Bioinformatics*, **19**, 185-93.

Example

- Distributions that should be the same are forced to be the same (dark line).

43 Lectures.GersteinLab.org (c)

Iterated Quantile Normalization to Correct for Non-specific Cross-hyb.



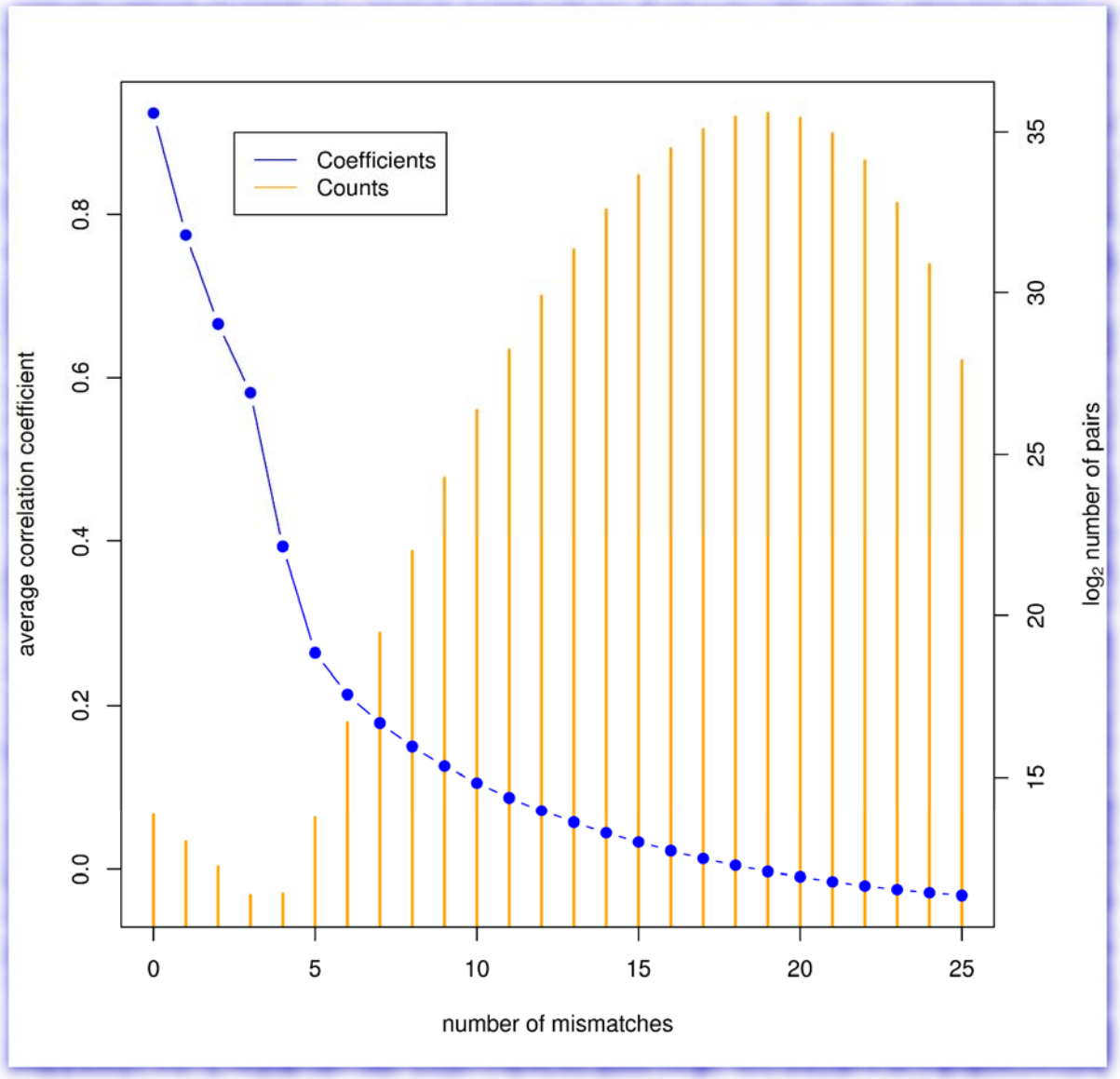
- Adapt Bolstad et al (2003) approach to tiling arrays
- Force distributions with a given nt at each position to be same
- Distributions at other positions now different so iterate
- Also, robust adaptation of Naef & Magnasco (2003)

T Royce et al (2007), *Bioinformatics*, **23**, 988-97

44 Lectures.GersteinLab.org (c)

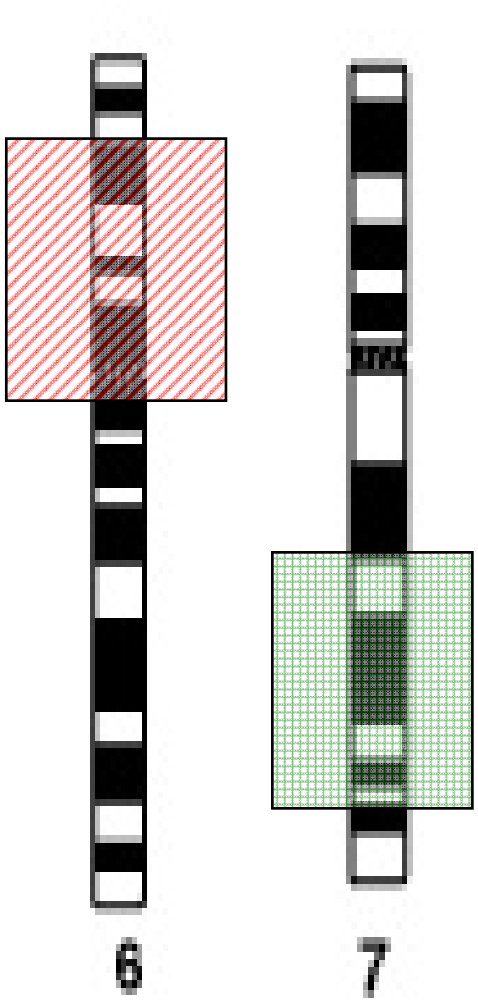
Measuring Specific Cross-Hyb

- Start with Cheng et al. (2005) tiling of human genome at 5 nt resolution giving expression profiles across various cell lines
- Correlation betw. probe pairs computed across cell lines' expression profiles and tabulated vs. number of mismatches
- The mean correlation coefficient was computed for each mismatch bin (blue series).
- The number of pairs is plotted as orange bars.



Source: Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 98-97

Proof of principle test to “exploit” this



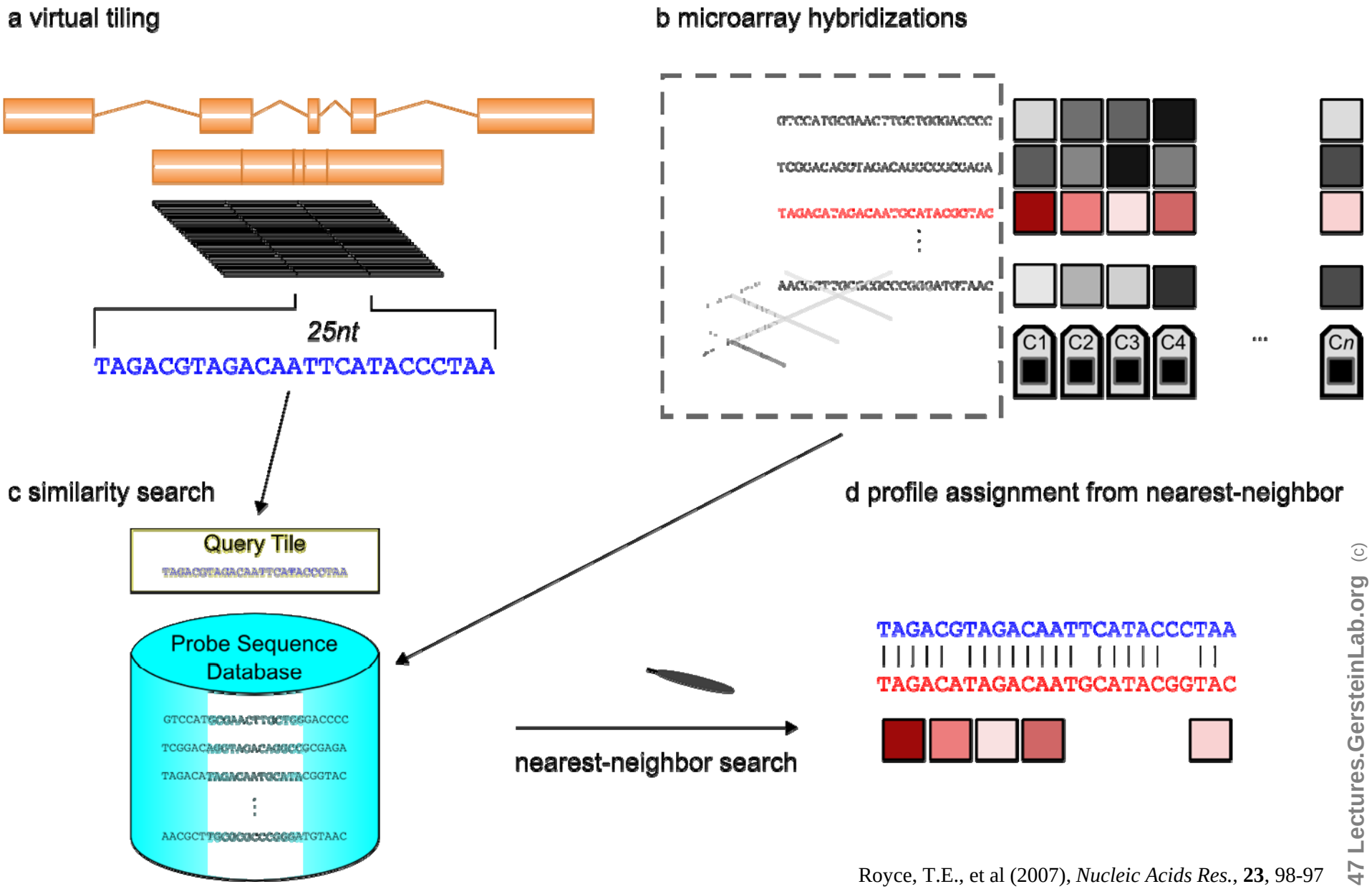
- Using Cheng et al. (2005), predict gene expression levels (and profiles across tissues) for genes on part of chr. #6
- ...Based on closest cross-hyb tiles on part of chr. #7
- Then compare to measured levels and profile on #6

Figure from http://www.members.cox.net/amgough/Fanconi-genetics-genetics-primer.htm

Source: Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 98-97

46 Lectures.GersteinLab.org (c)

Nearest Nbr Search on Virtual Tiling

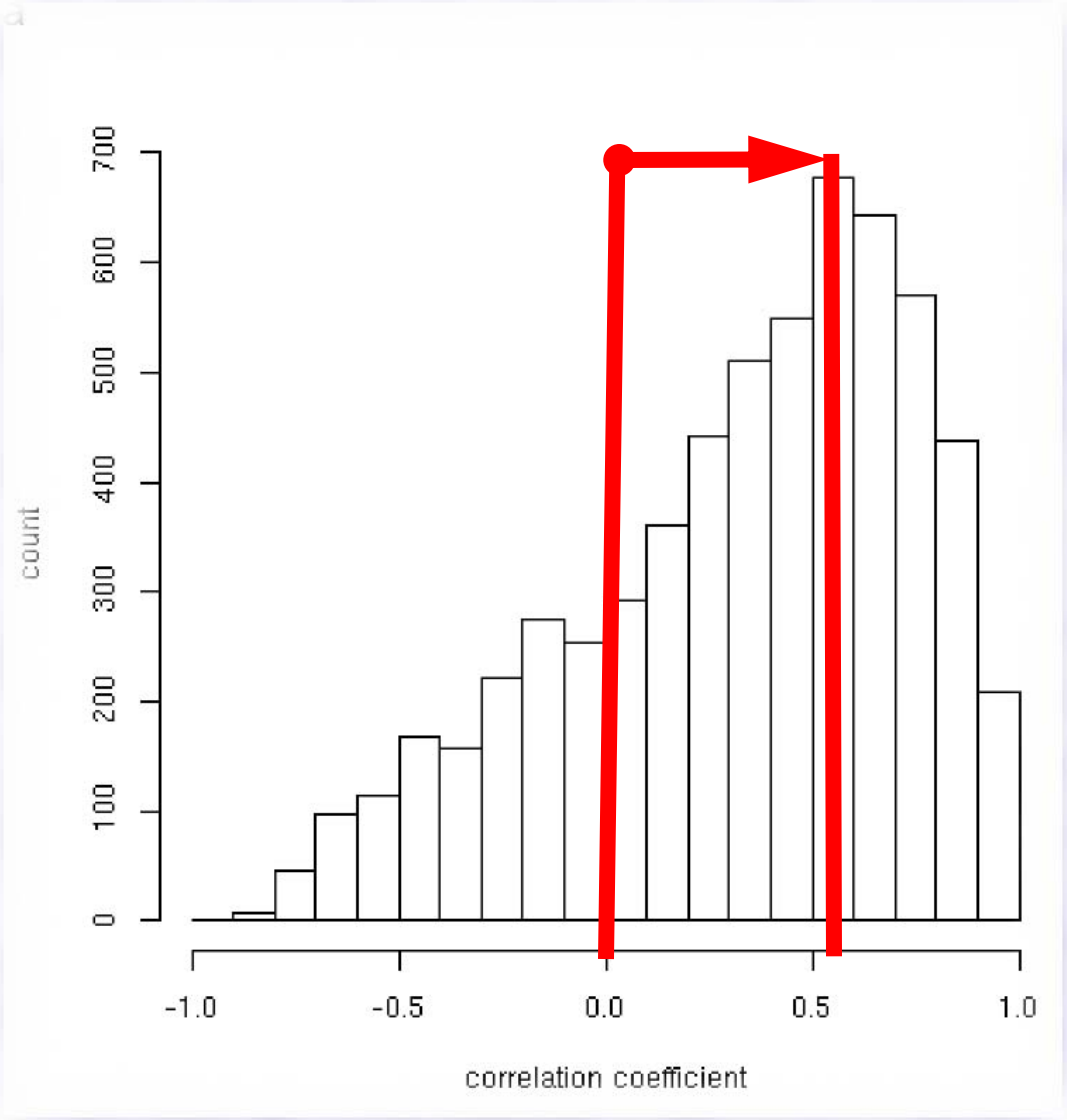


Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 98-97

47 Lectures.GersteinLab.org (c)

Agreement between predicted tile expression profile and actual one

- Correlated predicted profiles with the actual profiles of gene expression across cell lines
- Much more correlation than expected by chance (dist. centered on 0)



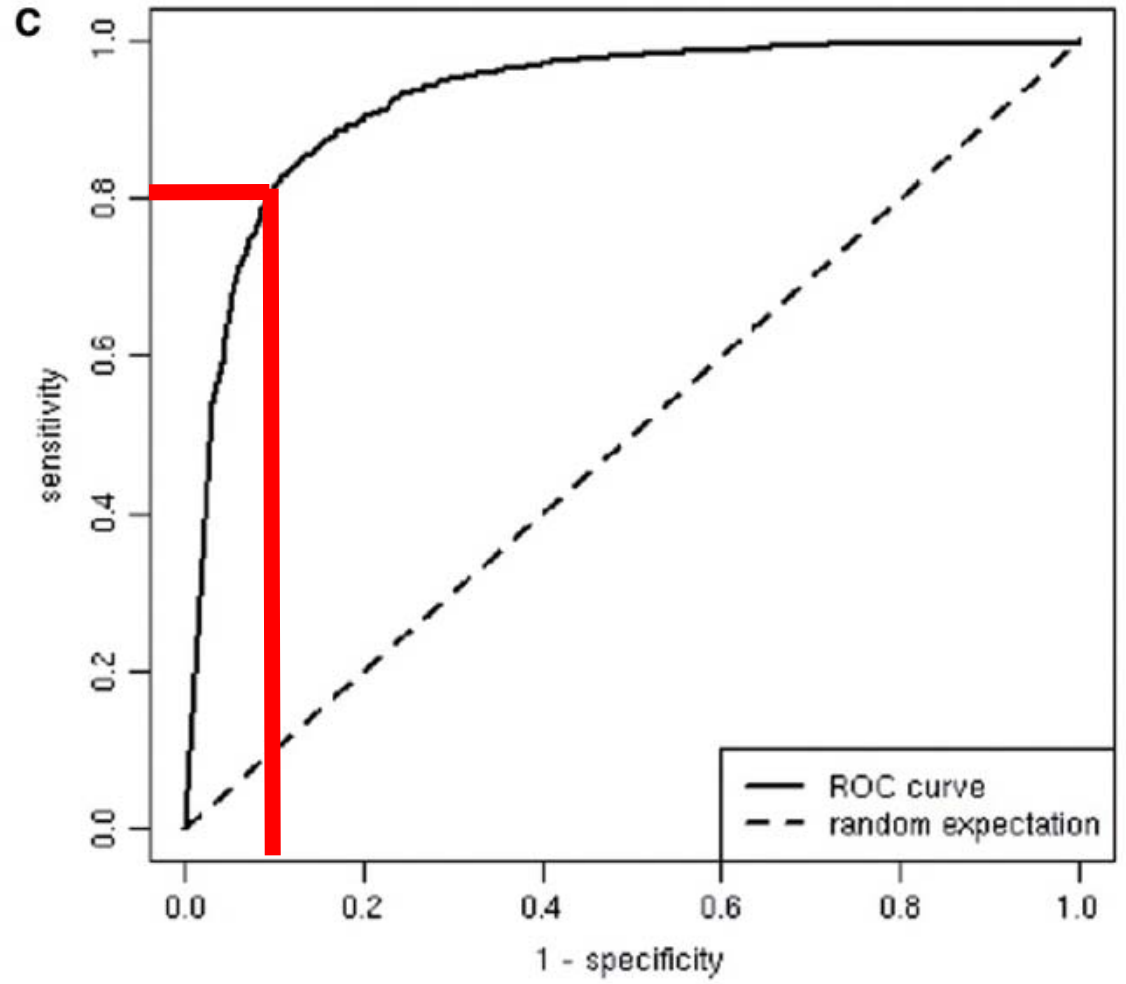
Source: Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 9

48 Lectures.GersteinLab.org (c)

Very Strong ROC Curve: Most genes are accurately detected using nearest-neighbor features' signals

- **Illustrates great magnitude of cross-hyb. on hi-density arrays**
- High feature density arrays inadvertently resurrecting generic n-mer concept (van Dam & Quake, 2003)
- Suggests that tiling arrays could be exploited to create **universal arrays**

- Gold std. set of known expressed genes. How well do we find.
- A set of known positives was defined as the Refseq genes with at least 75% transfrag coverage. A set of known negatives was constructed by permuting the sequences in the set of known positives. For various thresholds, sensitivity and specificity were computed and then plotted.



Royce, T. E. et al. Nucl. Acids Res. 2007 35:e99

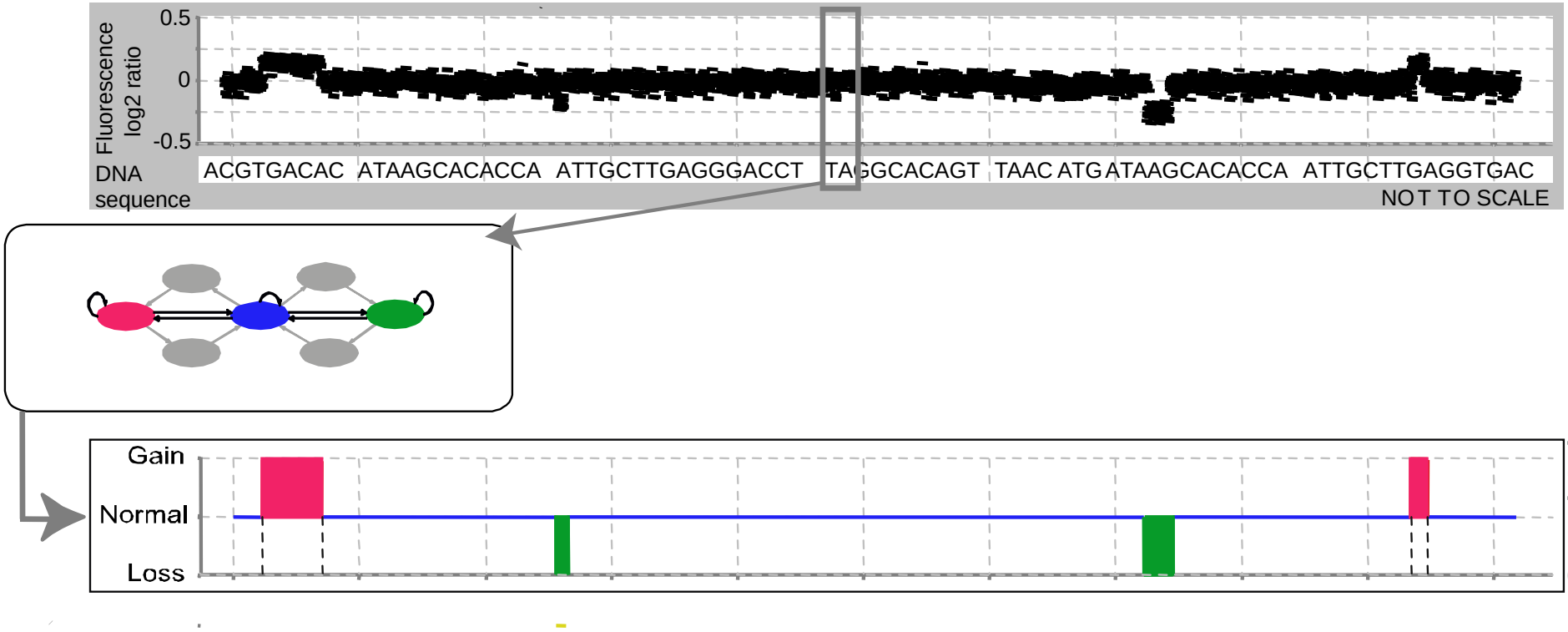
49 Lectures.GersteinLab.org (c)

HMM Segmenters, using "active learning" to find Annotation Blocks from Raw Signal

50 Lectures.GersteinLab.org (c)

BreakPtr HMM

- To get highest resolution on breakpoints need to smooth & segment the signal
- BreakPtr: prediction of breakpoints, dosage and cross-hybridization using a system based on Hidden Markov Models

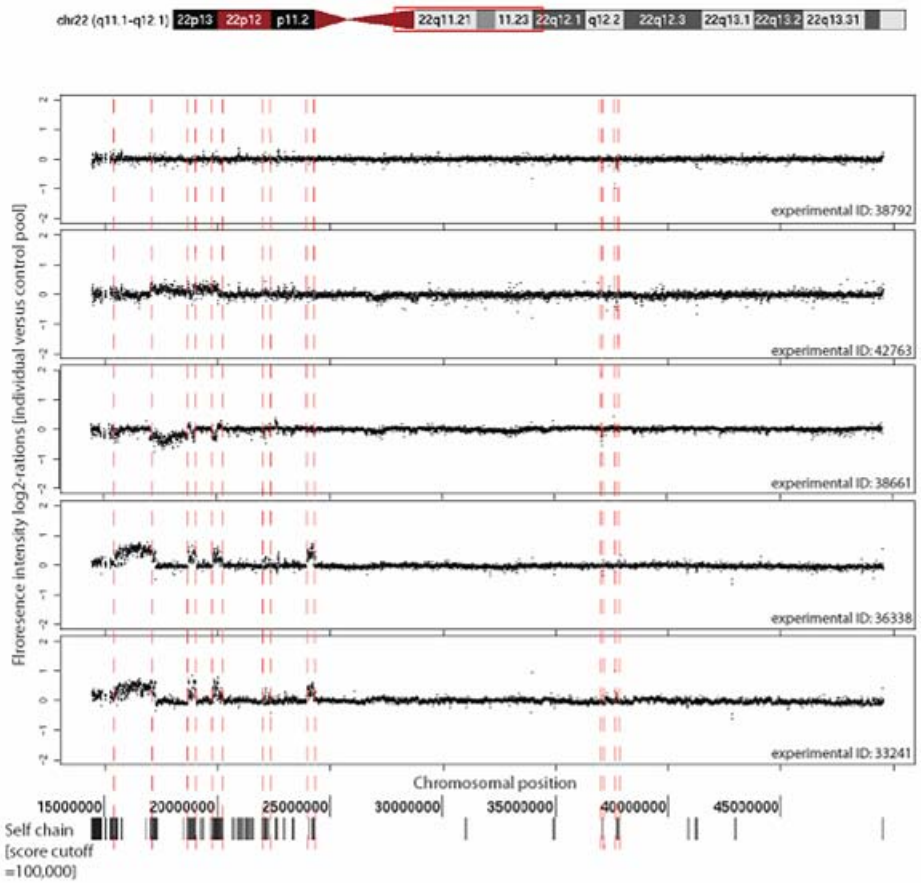


Korbel*, Urban* *et al.*, PNAS (2007)

http://breakptr.gersteinlab.org

53 Lectures.GersteinLab.org (c)

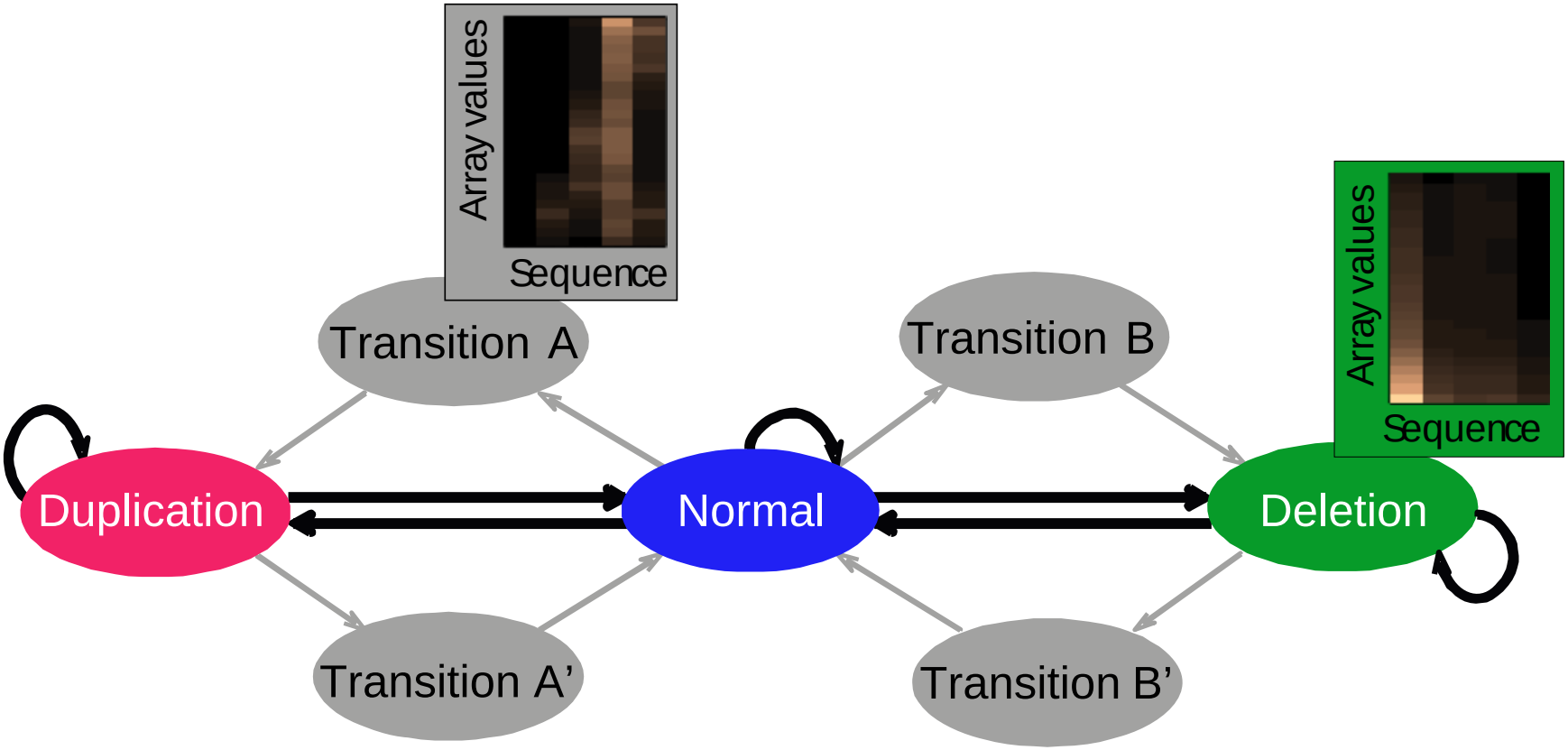
High resolution of tiling arrays allows statistical integration of nucleotide sequence patterns



>4-fold enrichment of the breakpoints of copy number variants near segmental duplications (SDs)
[e.g. Sharp *et al.*, *Am. J. Hum. Genet.* 2005; 77:78-88].

54 Lectures.GersteinLab.org (c)

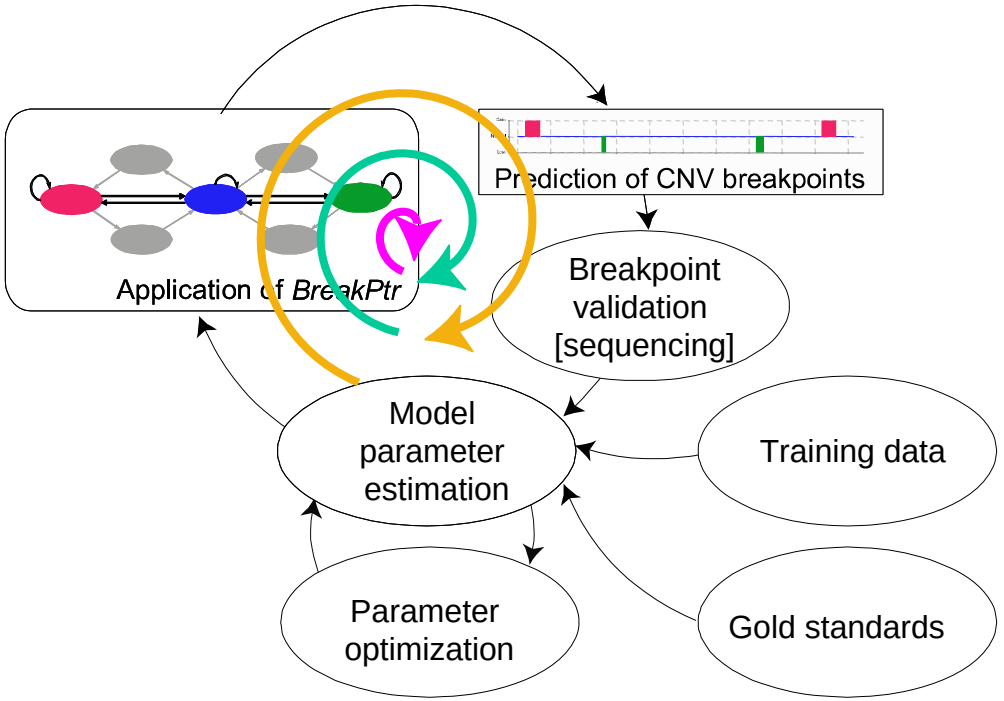
BreakPtr statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)

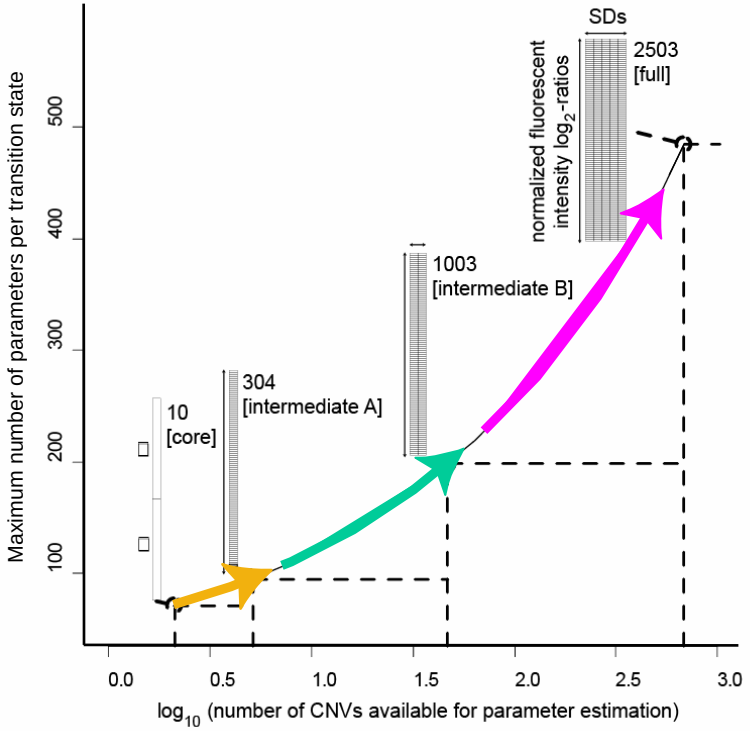


Korbel*, Urban* *et al.,* PNAS (2007)

55 Lectures.GersteinLab.org (c)

‘Active’ approach for breakpoint identification: initial scoring with preliminary model, targeted validation (with sequencing), retraining, and rescoring





CNV breakpoints sequenced in ~10 cases following BreakPtr analysis;

Median resolution <300 bp

No improvement in accuracy with higher resolution (9nt tiling)

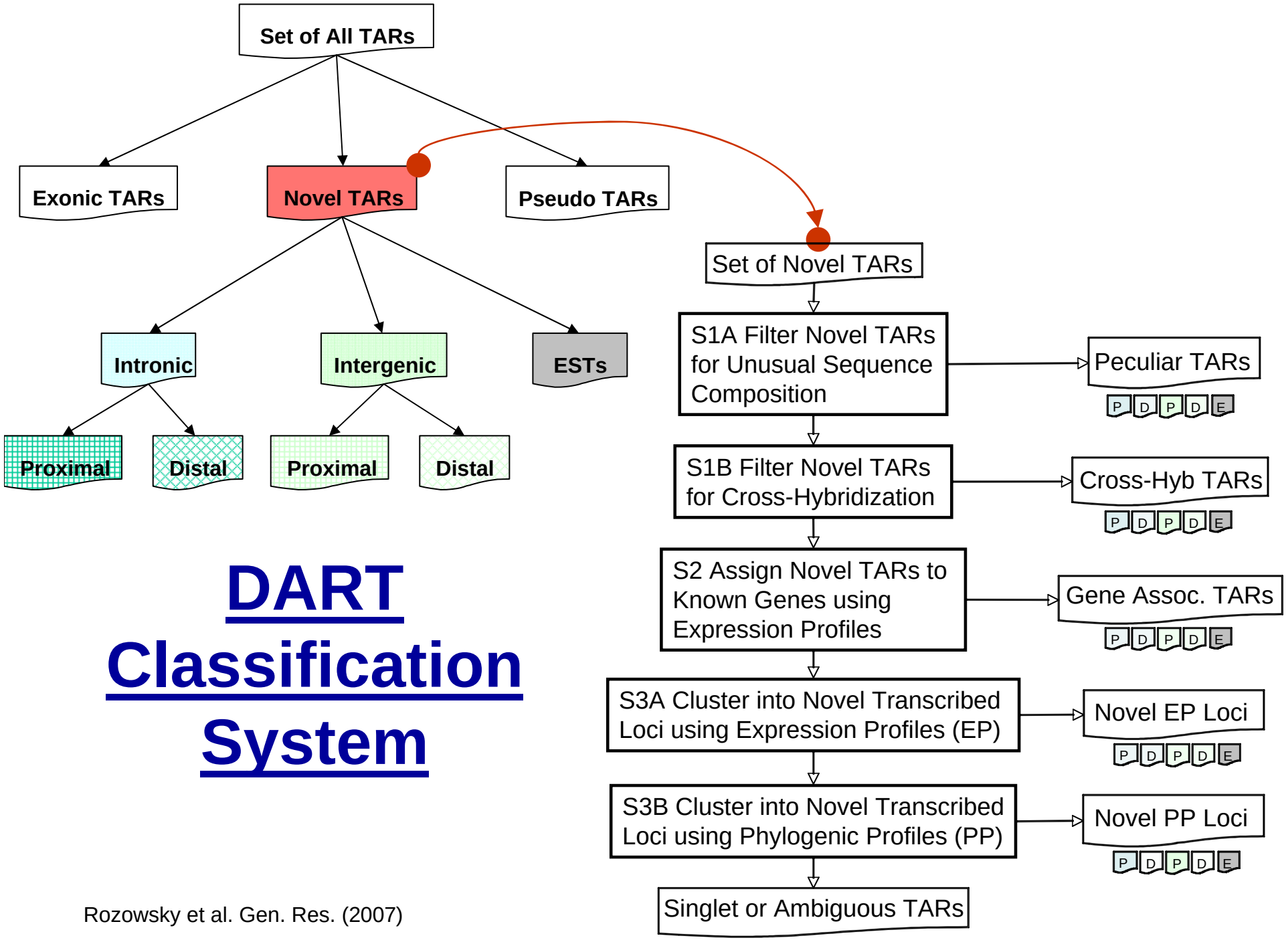
HMM optimized iteratively (using Expectation Maximization, EM)

Korbel*, Urban* *et al.,* PNAS (2007)

56 Lectures.GersteinLab.org (c)

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

57 Lectures.GersteinLab.org (c)

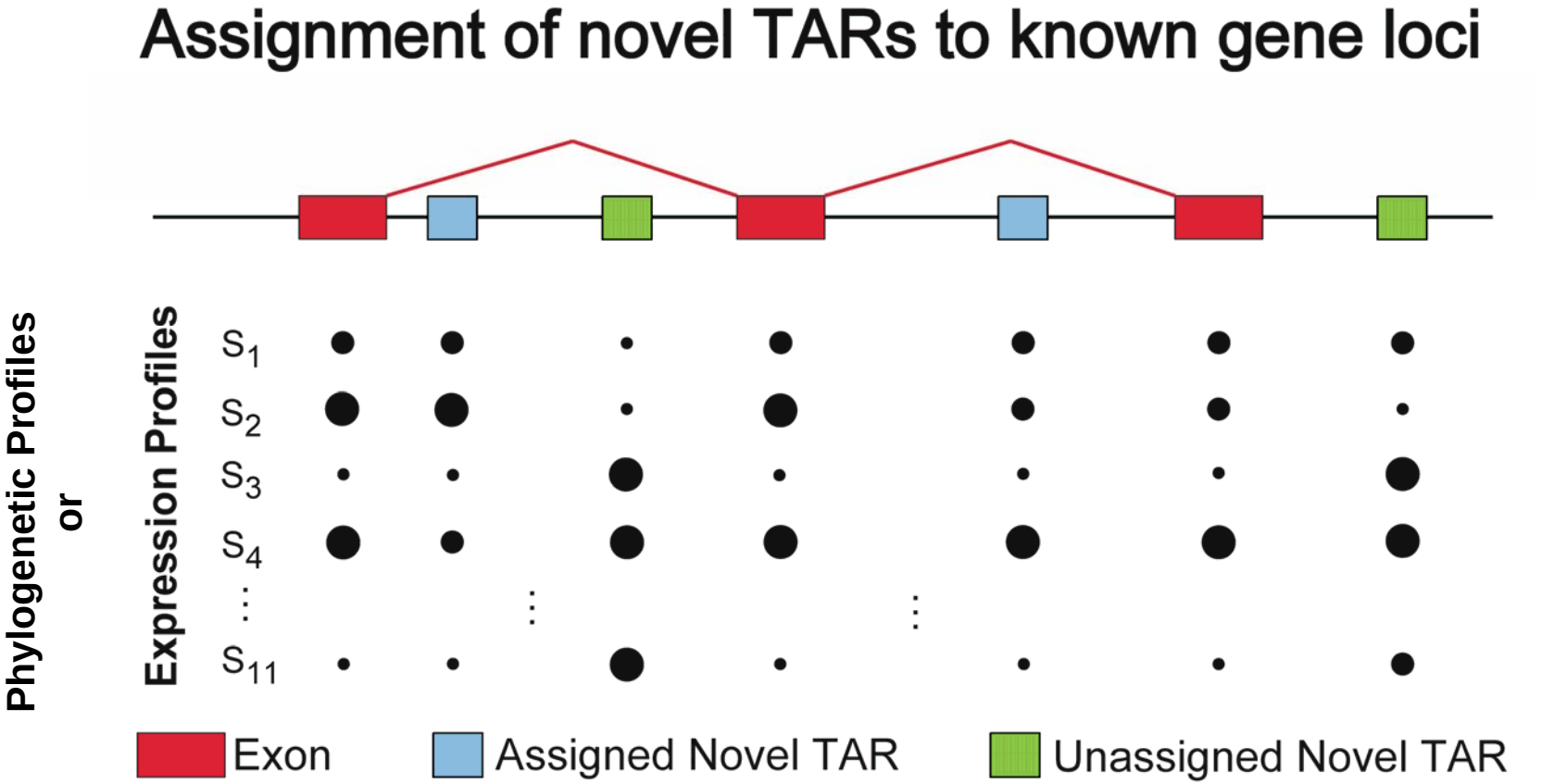


DART Classification System

Rozowsky et al. Gen. Res. (2007)

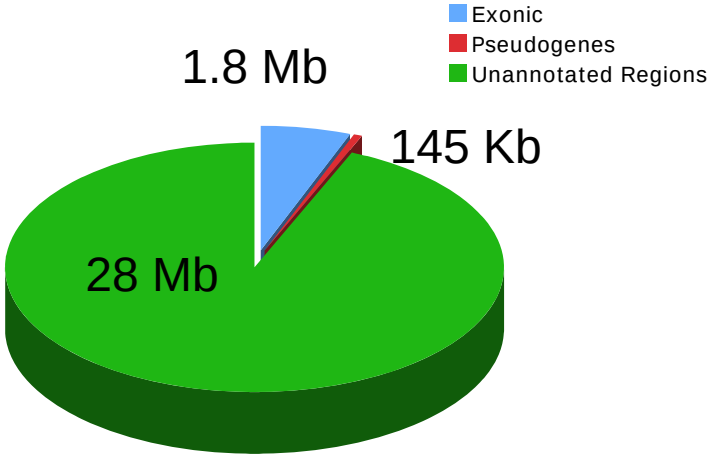
Production | Integrated

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

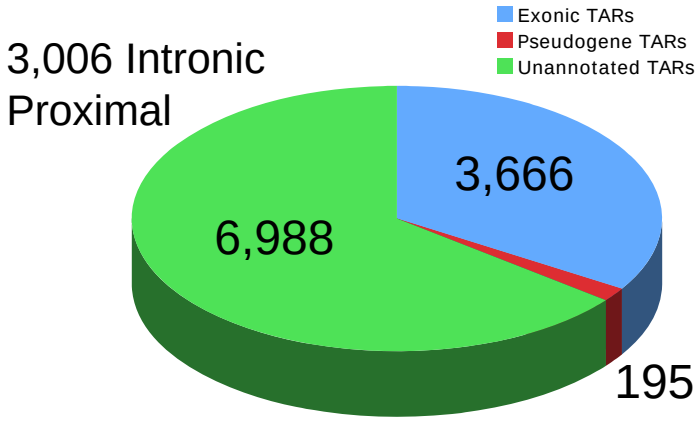


Rozowsky et al. Gen. Res. (2007)

60 Lectures.GersteinLab.org (c)



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

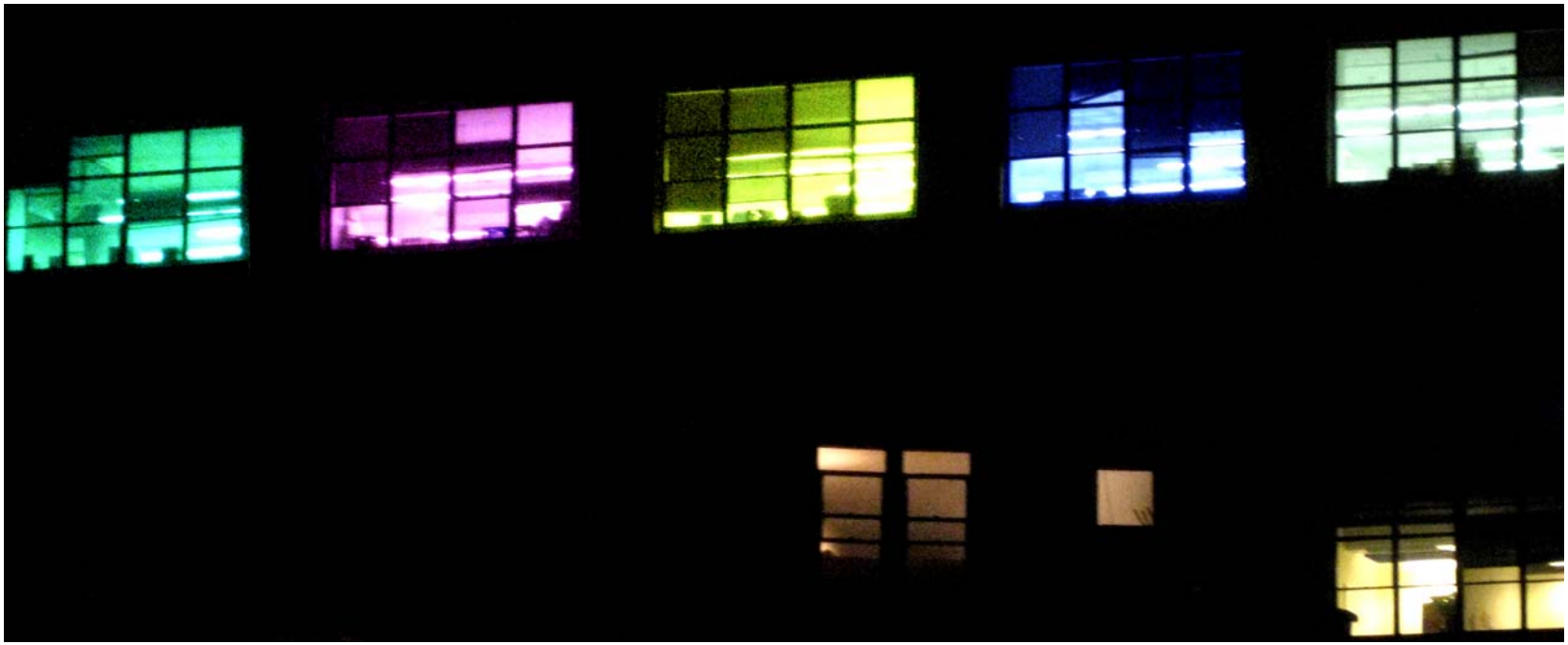
◊ 18/46 (39%) confirmed by RT-PCR

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

Rozowsky et al. Genome Research (2007)

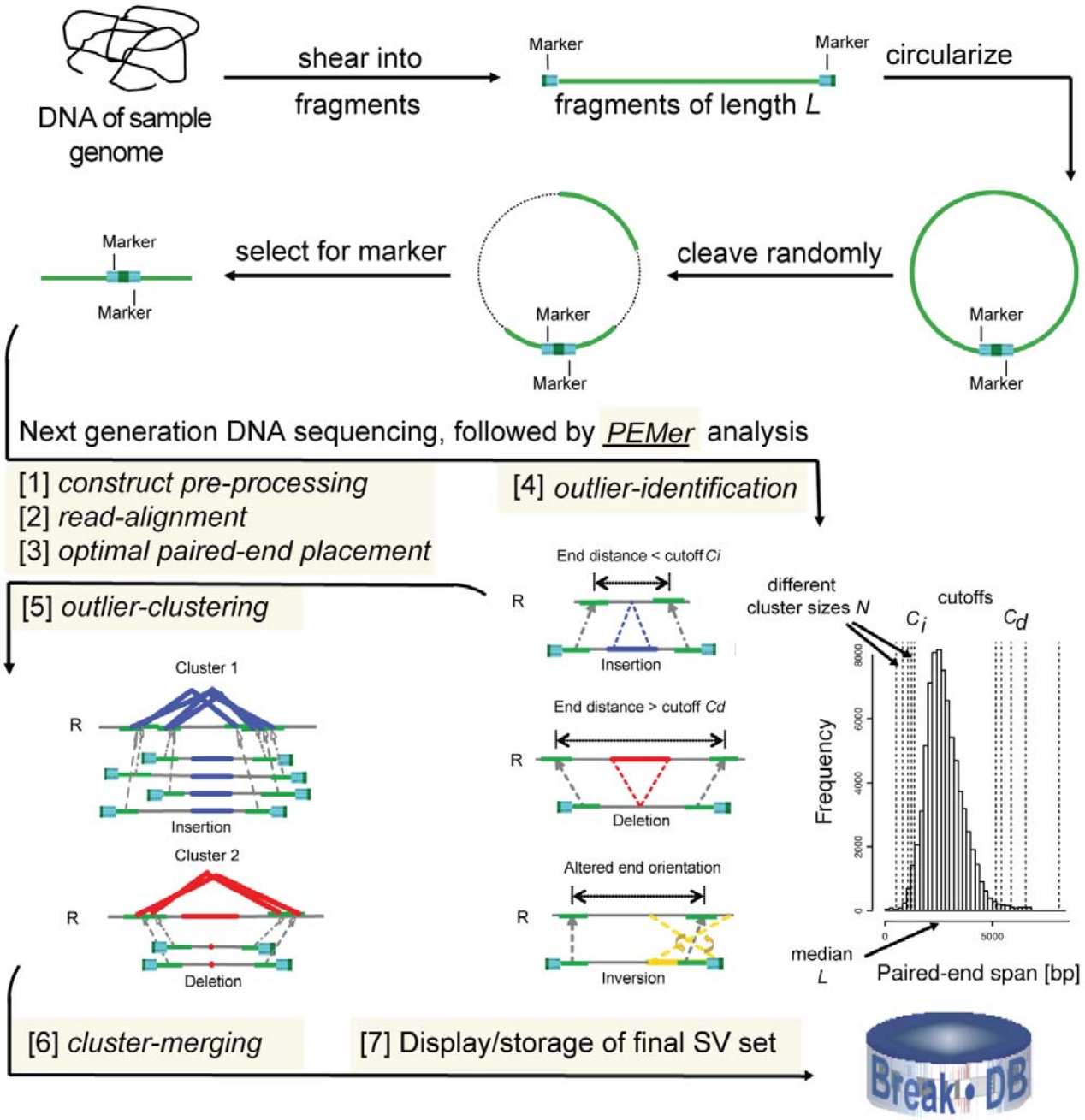
61 Lectures.GersteinLab.org (c)

Moving Beyond Arrays Next-Generation Sequencing strategy for characterizing genomes: Paired End Mapping to Find SVs



88 Lectures.GersteinLab.org (c)

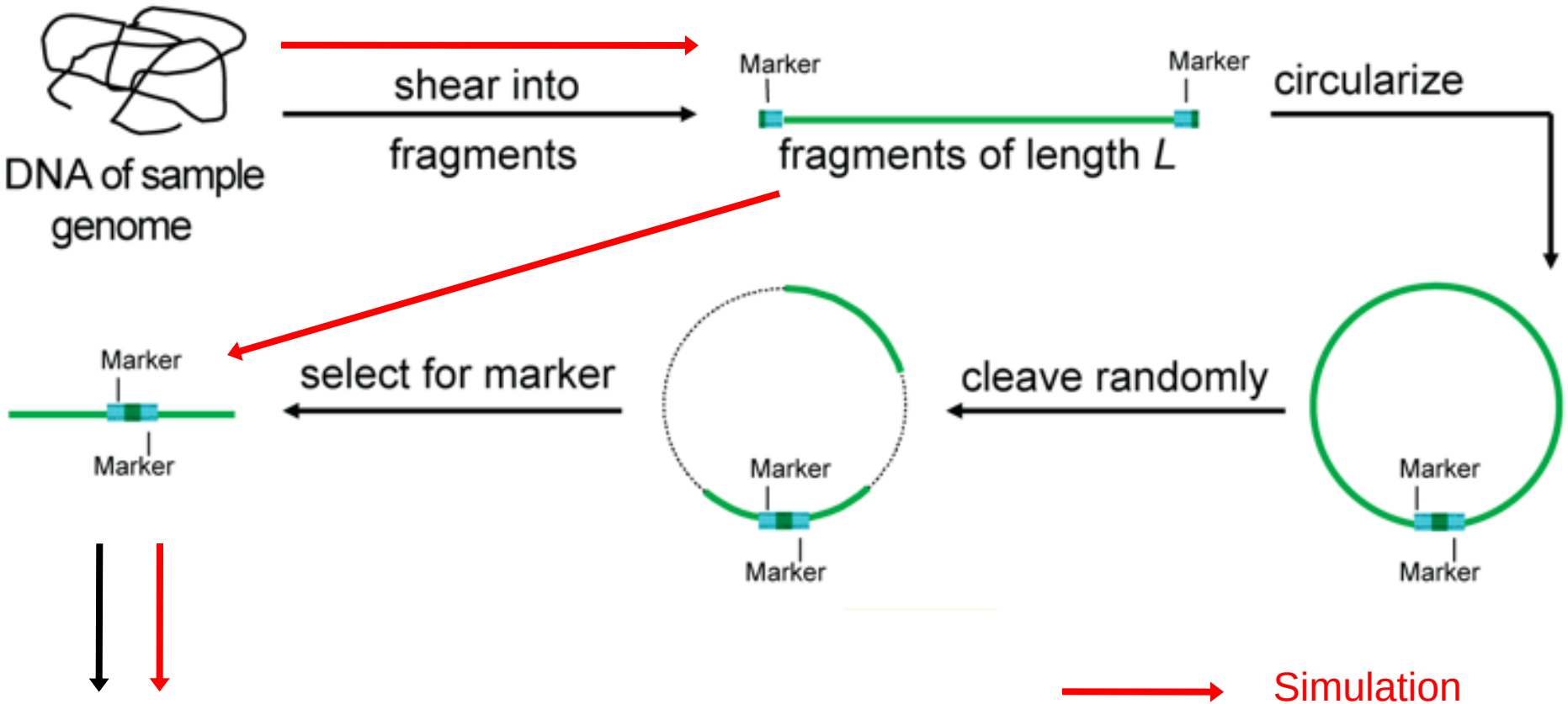
Overall Strategy for Analysis of NextGen Seq. Data to Detect Structural Variants



[Korbel et al., Science ('07); Korbel et al., GenomeBiol. (submitted)]

90 Lectures.GersteinLab.org (c)

Simulation strategy



[Korbel et al., GenomeBiol. (submitted)]

[Korbel et al., GenomeBiol. (submitted)]

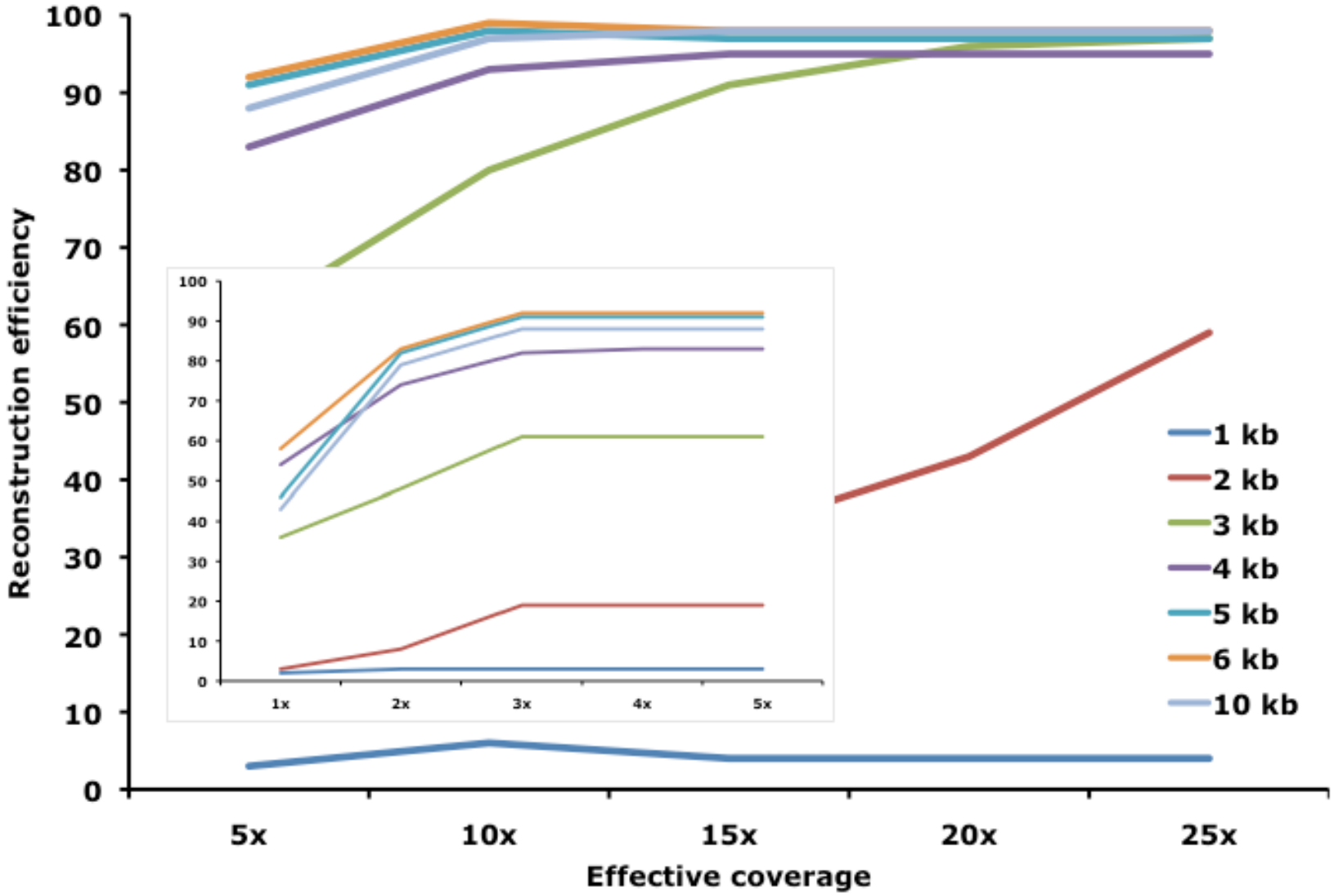
Reconstruction efficiency at 5x coverage

SV size	*Single cutoff*	*Multi-cutoff*	*Simplified multi-cutoff*	*Multi-cutoff(*)*	*Simplified multi-cutoff(*)*
1000	3(4)	3(4)	3(4)	3(4)	3(4)
2000	12(13)	23(26)	21(23)	11(13)	6(6)
3000	52(57)	61(68)	61(68)	49(52)	44(46)
4000	84(85)	85(86)	85(86)	80(82)	80(82)
5000	91(93)	91(93)	91(93)	91(93)	91(93)
6000	92(92)	92(92)	92(92)	92(92)	92(92)
10000	88(91)	88(91)	88(91)	88(91)	88(91)
Total	422(435)	443(460)	441(457)	414(427)	404(414)
False positives	31(31)	31(31)	26(31)	5(4)	2(1)

* optimal strategy; Multi-cutoff: overlaps 2-8; Simplified multi-cutoff: overlaps 2, 3, and 4

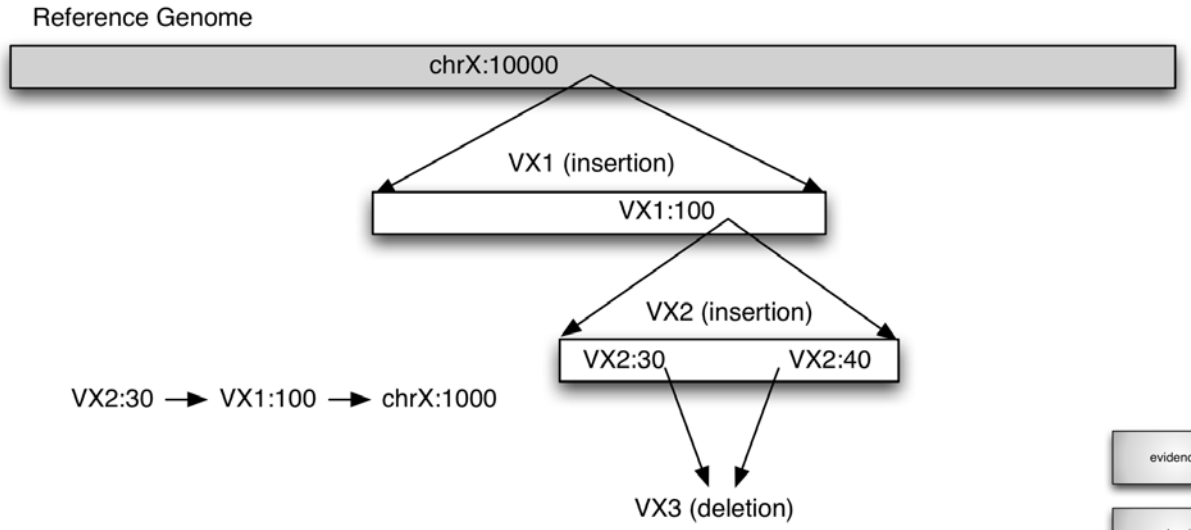
[Korbel et al., GenomeBiol. (submitted)]

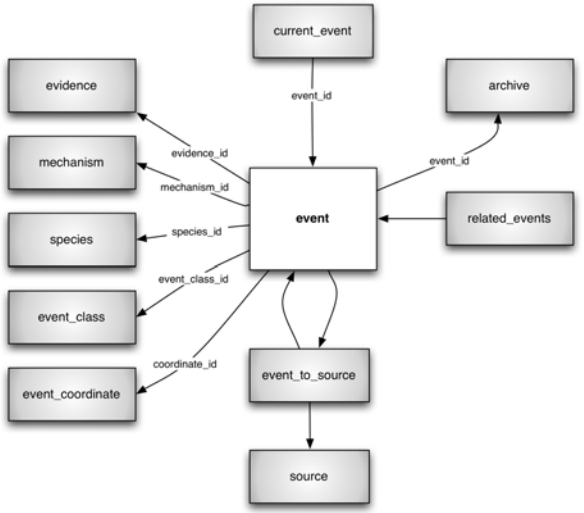
Reconstruction efficiency at different coverage



93 Lectures.GersteinLab.org (c)

Building a Database of Variants: Complexities





[Korbel et al., GenomeBiol. (submitted)]

94 Lectures.GersteinLab.org (c)

Summary of PEM Results

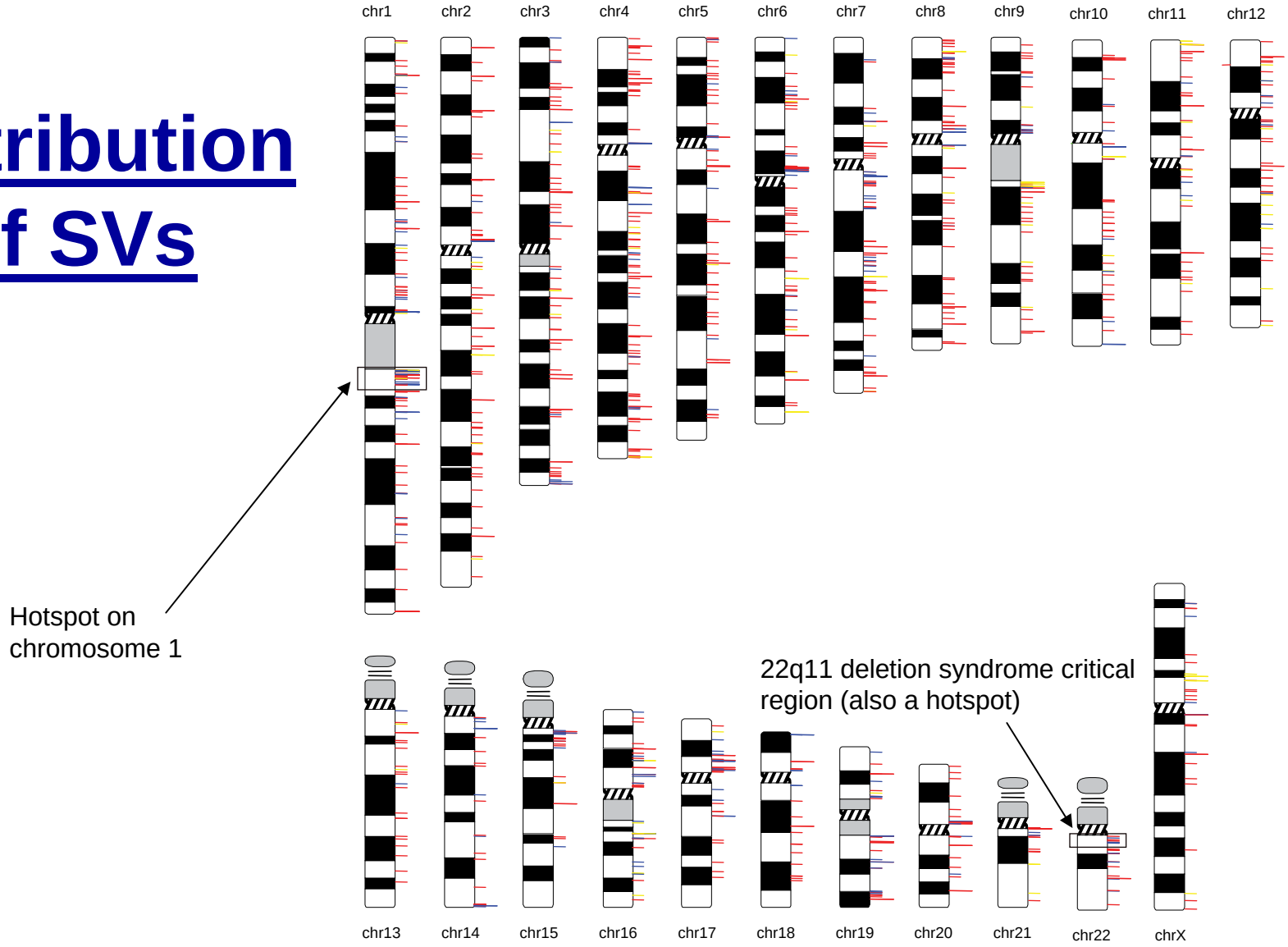
	NA15510 (Caucasian?, female)	**NA18505 (Yoruba, female)**
# of sequenced reads	> 10 M.	> 21 M.
Paired ends uniquely mapped	> 4.2 M.	> 8.6 M.
Fold coverage (on 6Gb)	~ 2.1x	~ 4.3x
Predicted Structural Variants*	**478**	**839**
Indels	*427*	*758*
Inversion breakpoints	*51*	*81*
Estimated total variants* with respect to ncbi36, genome-wide	**759**	**902**

*at this resolution

Korbel et al., 2007 *Science*

95 Lectures.GersteinLab.org (c)

Distribution of SVs

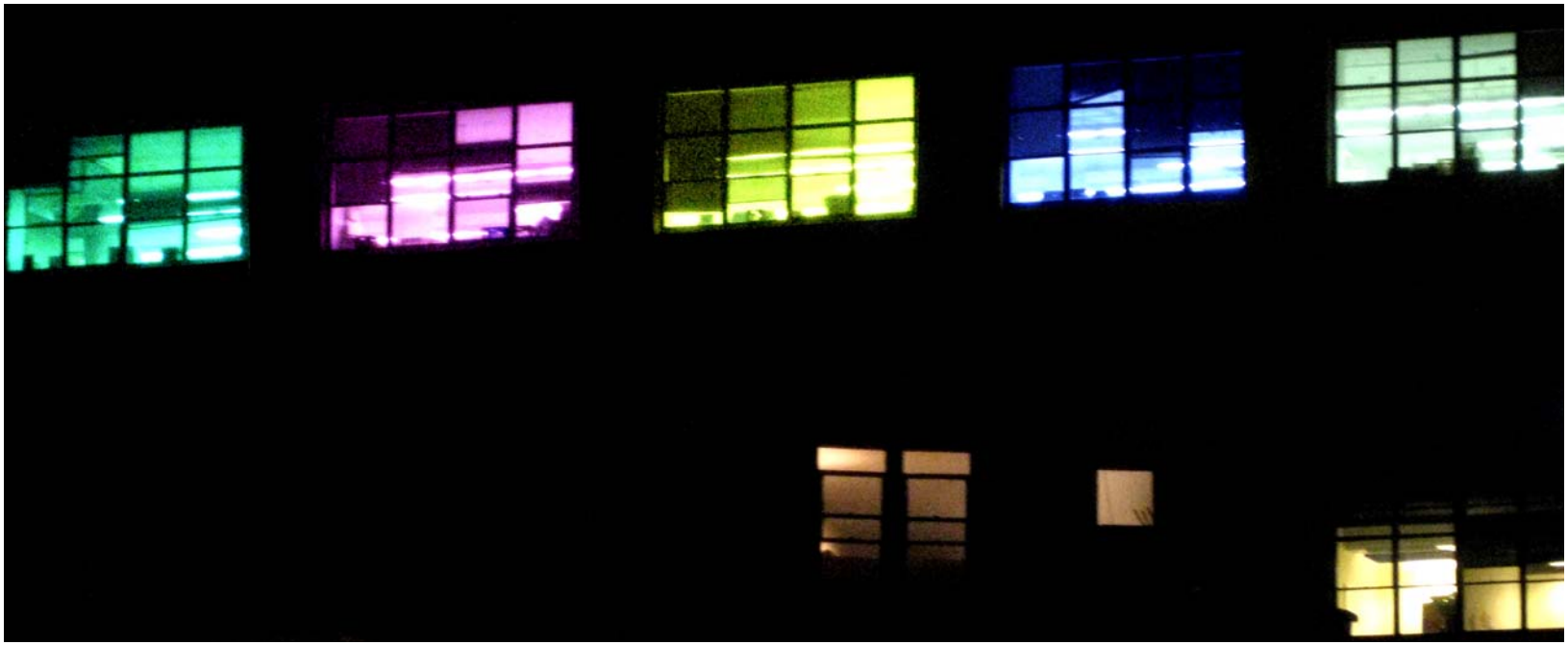


- red deletion; blue insertion; yellow inversion; double line length: same SV in both individuals.

Korbel et al., 2007 *Science*

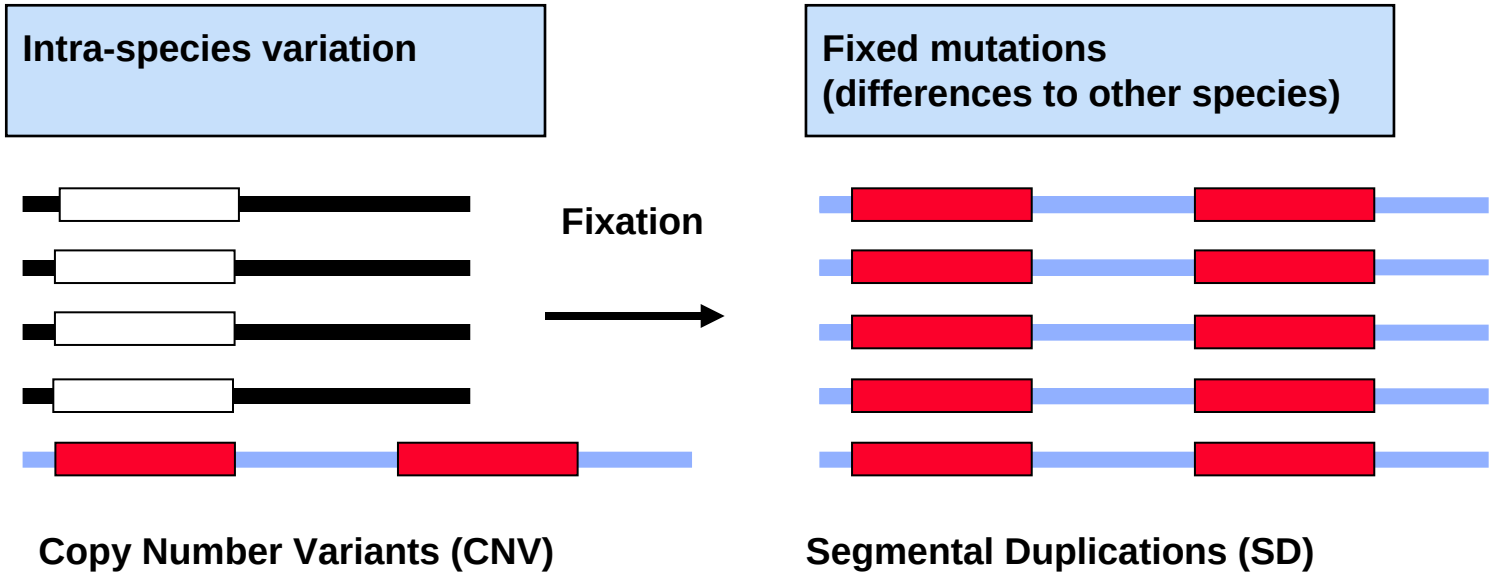
96 Lectures.GersteinLab.org (c)

Analyzing Duplications in the Genome (SDs & CNVs)



105 Lectures.GersteinLab.org (c)

SEGMENTAL DUPLCATIONS AND COPY NUMBER VARIANTS ARE RELATED PHENOMENA AND SHOULD HAVE BEEN CREATED BY SIMILAR MECHANISMS



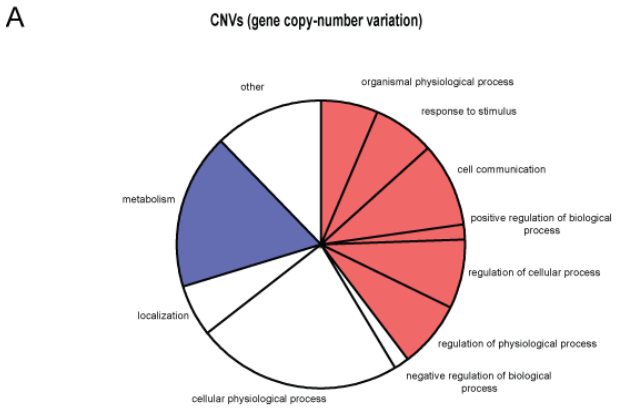
- SDs are the fixed forms of CNVs and arise when a CNV reaches fixation in the population.
- Hence, they should have been created by a similar mechanism

Association of SDs and CNVs with pseudogenes

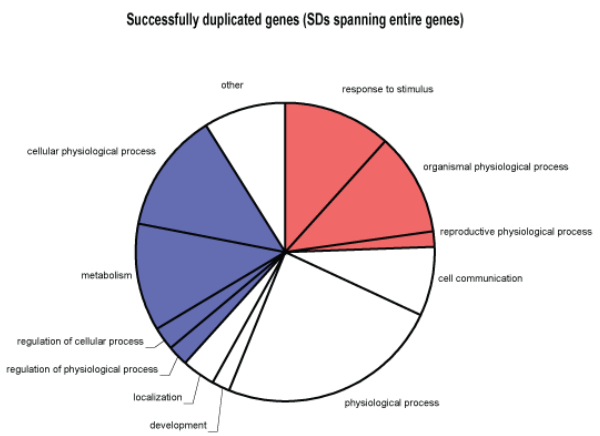
- CNVs are the raw form of variation producing duplicated elements
- Segmental Duplications (SDs) are fixed forms of CNVs/SVs. They give rise to duplicated genes and (eventually) protein protein families
- Thus, we expect, duplicated pseudogenes (failed duplications) to occur in SDs.
- CNVs and SDs tend to be enriched in environmental response genes, matching a patterns previously found for duplicated pseudogenes

[Korbel et al., COSB (in press, '08)]

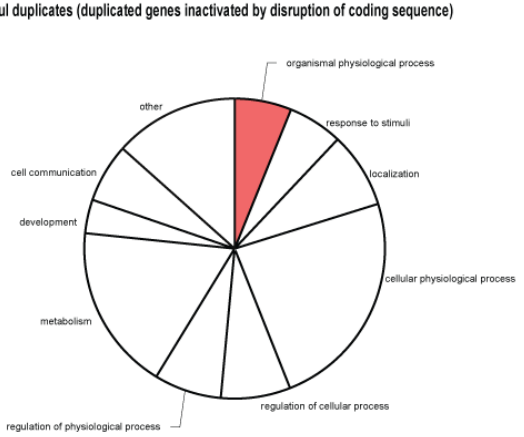
A



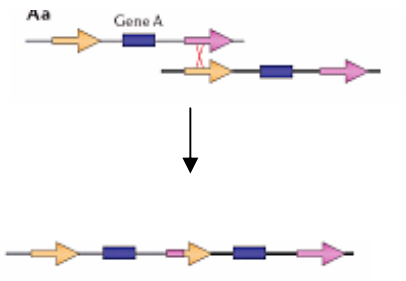
B



C

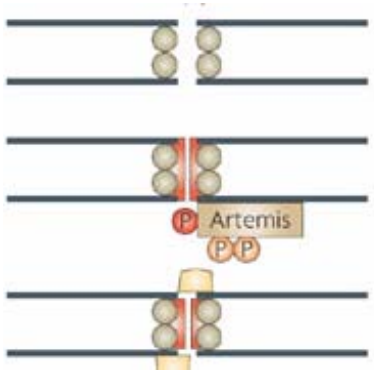


SEVERAL DIFFERENT MECHANISMS HAVE BEEN PROPOSED FOR THE GENESIS OF SDs AND CNVs



Non-allelic homologous recombination:

Flanking repeat signature should be at the breakpoint junctions



Non-homologous-end-joining

No repeats. In some cases <4bp microhomologies

- We can examine breakpoint sequences to determine the mechanism

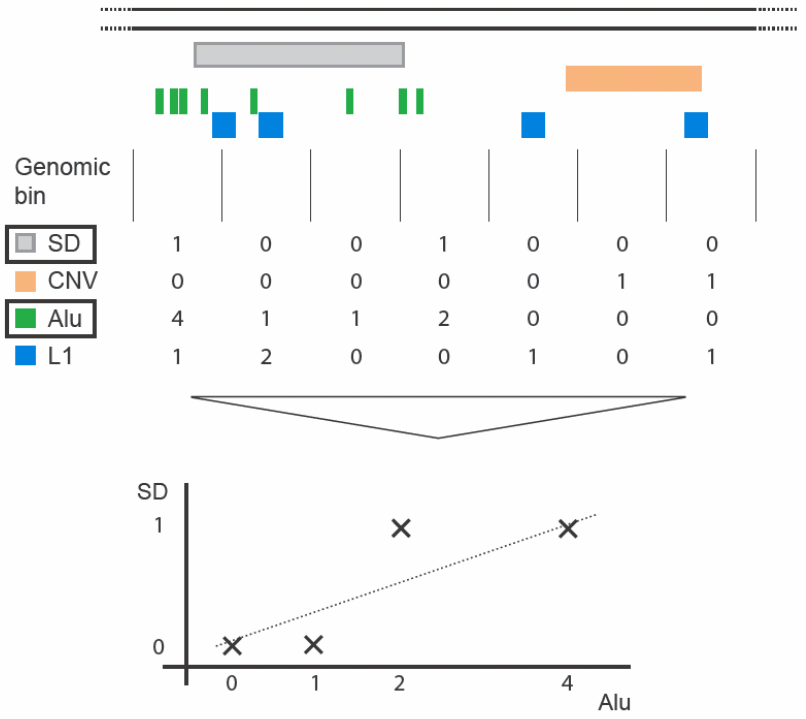
Problem:

- Most large-scale CNV data is of low resolution (50kb or worse)
- Cannot directly observe repeat signatures in the large-scale data!

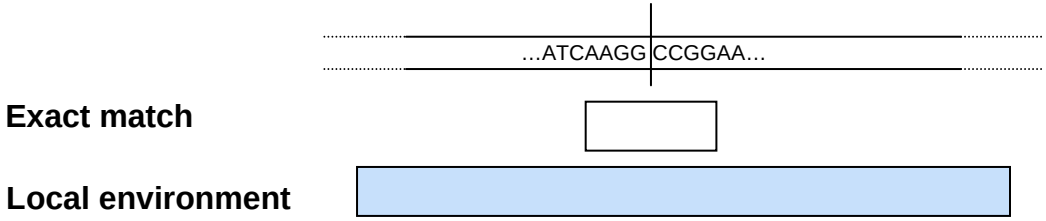
Source: Bailey and Eichler *Nat. Rev. Genetics* (2006), Korbel et al. *Science* (2007) and Kim et al. (submitted)

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

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

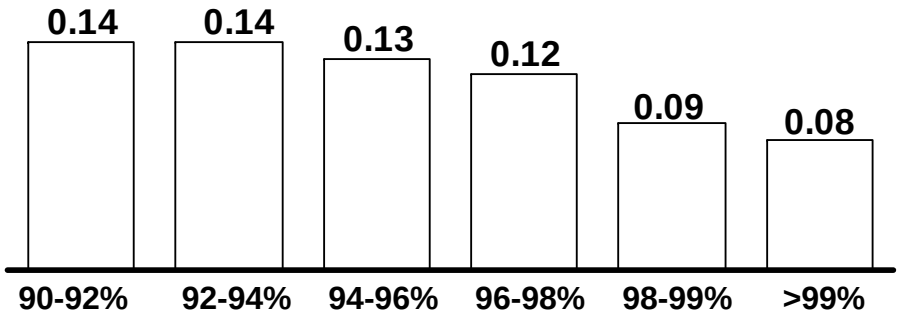


If exact CNV breakpoints are known, we can calculate the enrichment of repeat elements relative to the genome or relative to the local environment

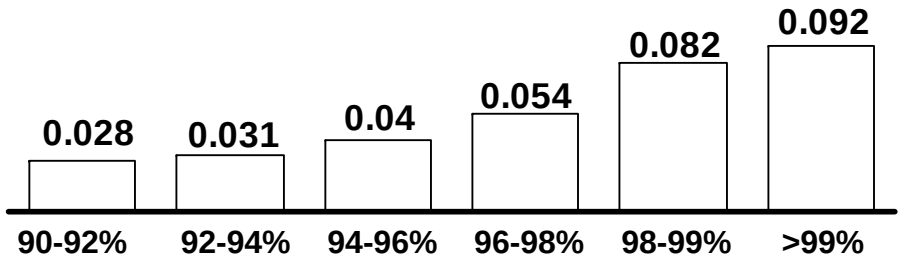


OLDER SDs ARE MUCH MORE LIKELY TO BE FORMED BY ALU ELEMENTS

Alu association with SDs by age



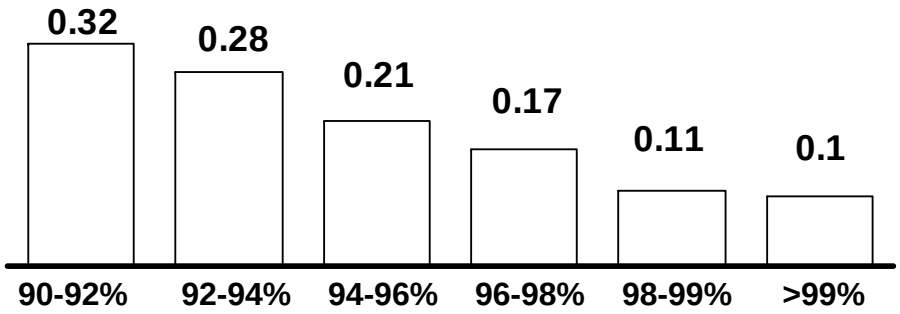
SD association with subtelomeres



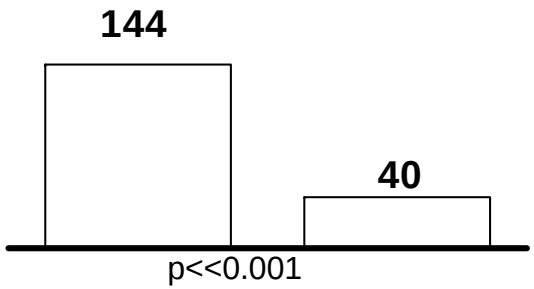
- The co-localization of Alu elements with SDs is highly significant.
- Older SDs have a much higher association with Alus than younger SDs.
- Hence it is likely, that Alu elements were more active in mediating NAHR in the past (consistent with the Alu burst)
- Younger SDs are more likely to be localized in subtelomeres (instable region susceptible to NHEJ)

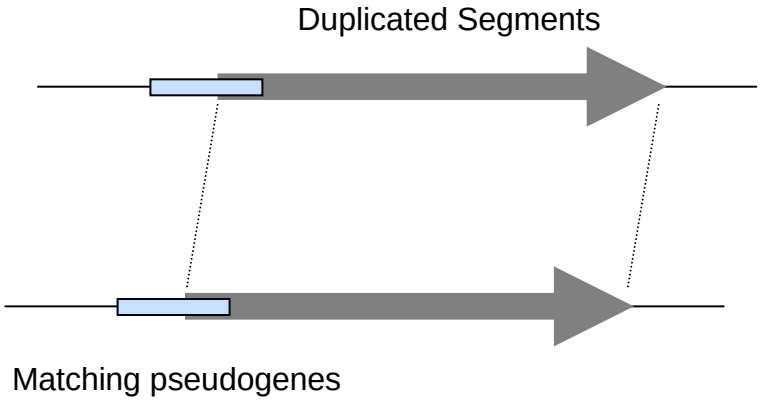
ANOTHER FUNCTION FOR PSEUDOGENES: SERVING AS REPEATS FOR MEDIATING NAHR

Processed pseudogene association with SDs by age



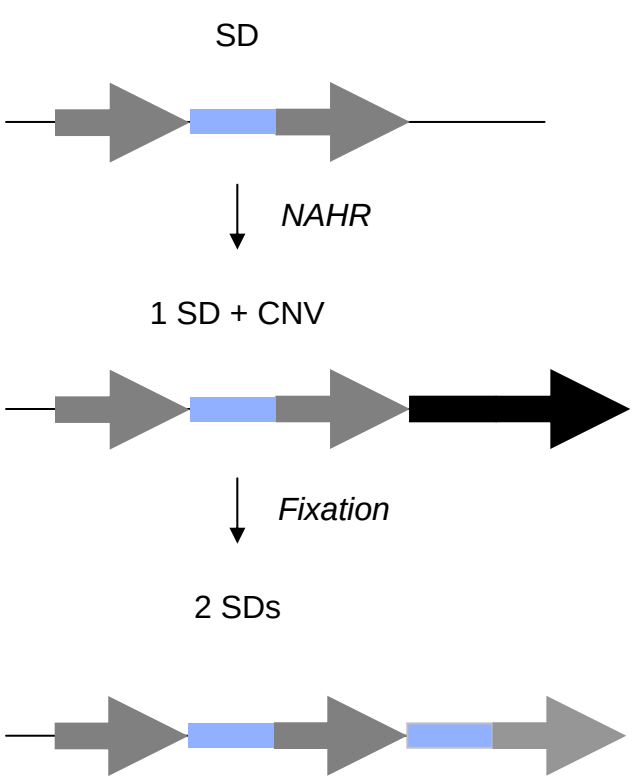
Processed pseudogenes at SD junctions





FOCUSSING ON SDS: SDS CAN PROPAGATE THEMSELVES, WHICH LEADS TO A POWER-LAW DISTRIBUTION

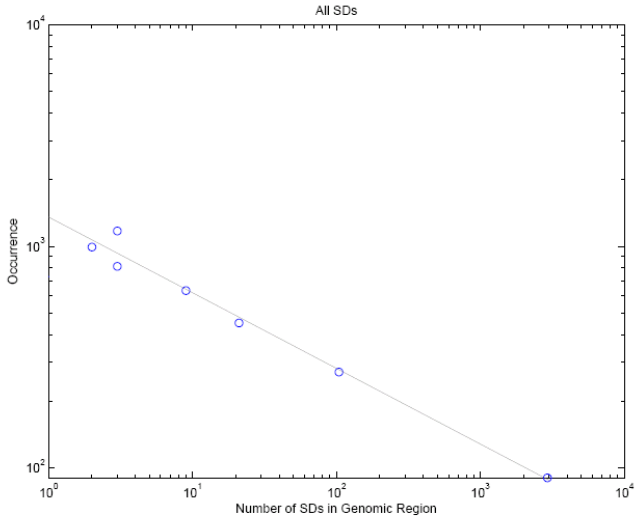
Hypothesis



"SD selfpropagation"

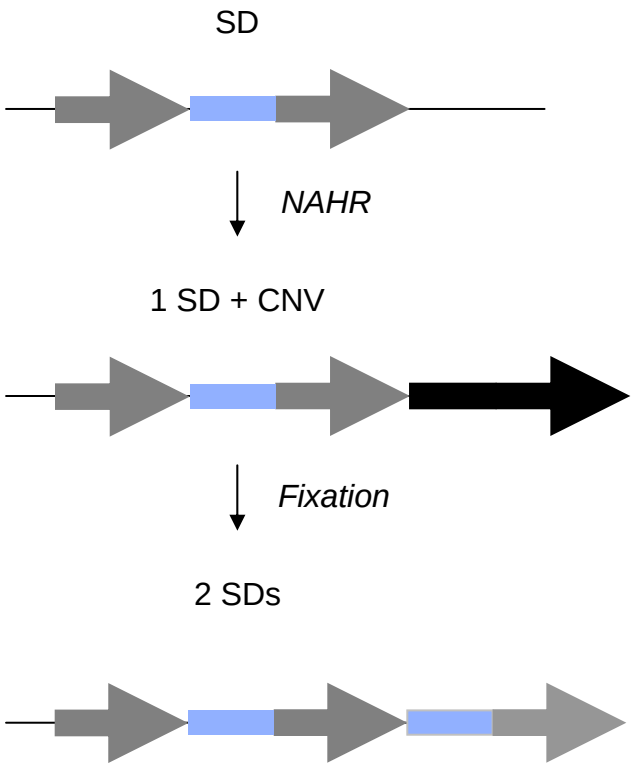
Corollary

- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- Such mechanisms ("preferential attachment") are well studied in physics and should leads a very skewed ("power-law") distribution of SDs.



FOCUSSING ON SDS: SDs COLOCALIZE WITH EACH OTHER

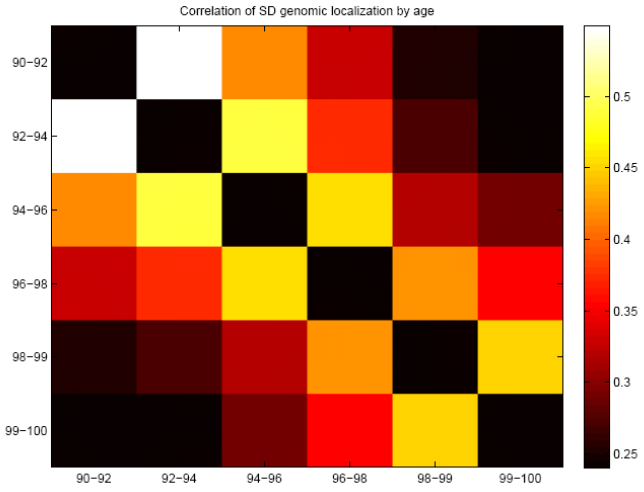
Hypothesis



"SD self-propagation"

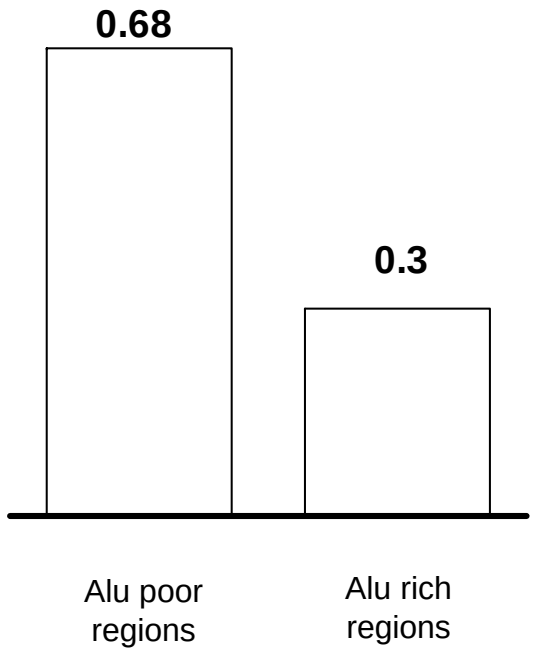
Corollary

- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- SDs of similar age should co-localize better with each other:



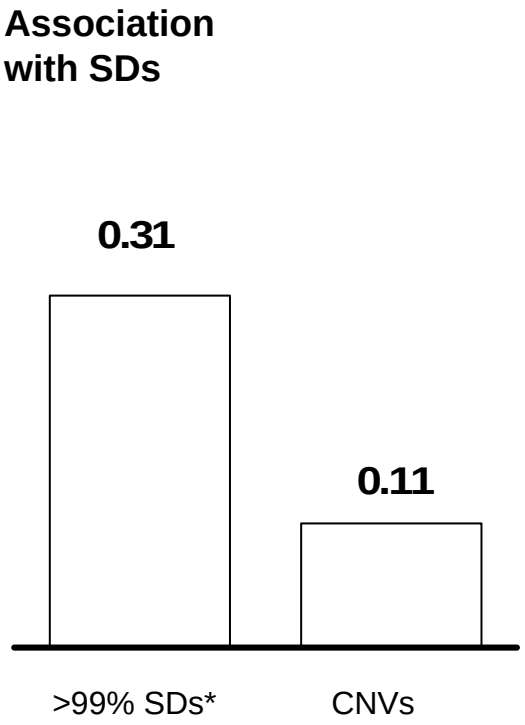
THE ASSOCIATION OF SDs WITH ALU ELEMENTS IS COMPLEMENTARY TO THE ONE WITH SDs

Correlation of young SDs (>99%) with older SDs



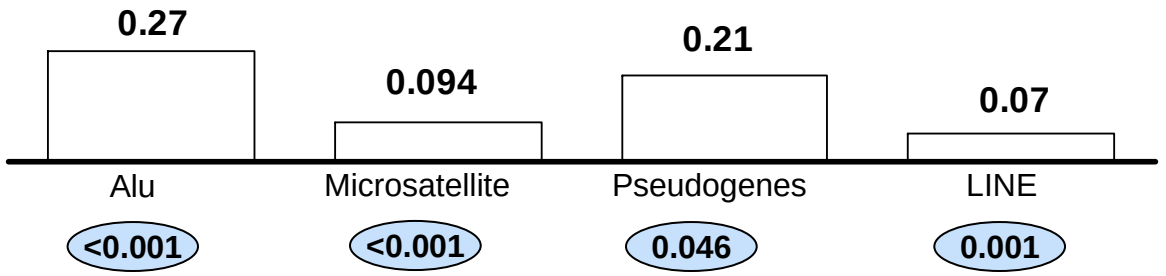
- SDs that lie in Alu rich regions are less likely to be associated with other SDs
- Hence, there is a certain complementarity to SD-mediated NAHR with Alu-metiated NAHR

CNVs ARE LESS ASSOCIATED WITH SDs THAN THE GENERAL SD TREND

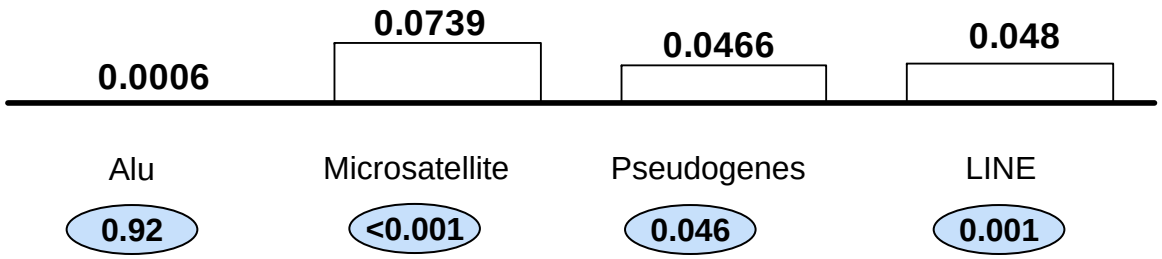


ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs

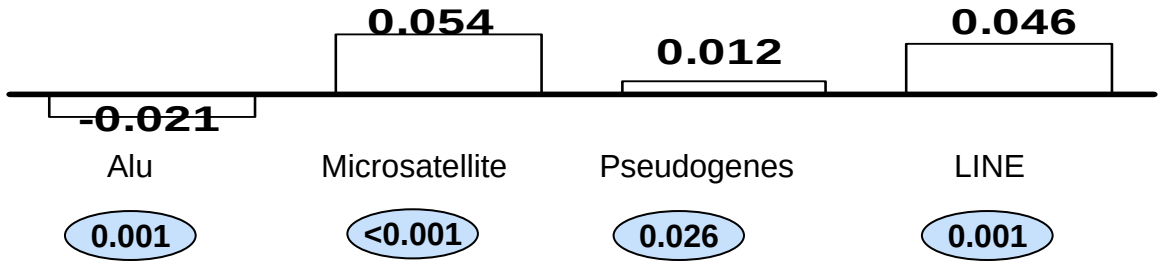
SD association with repeats



CNV association with repeats

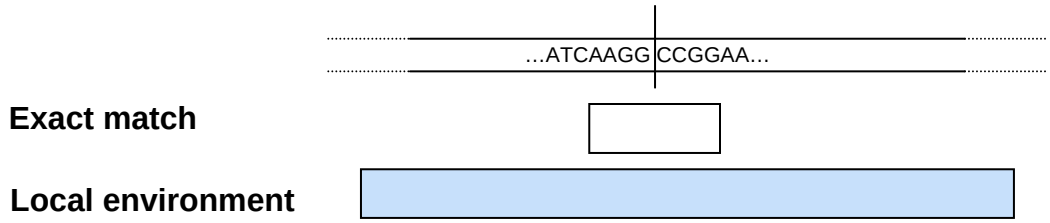


CNV association with repeats after correcting for SD content

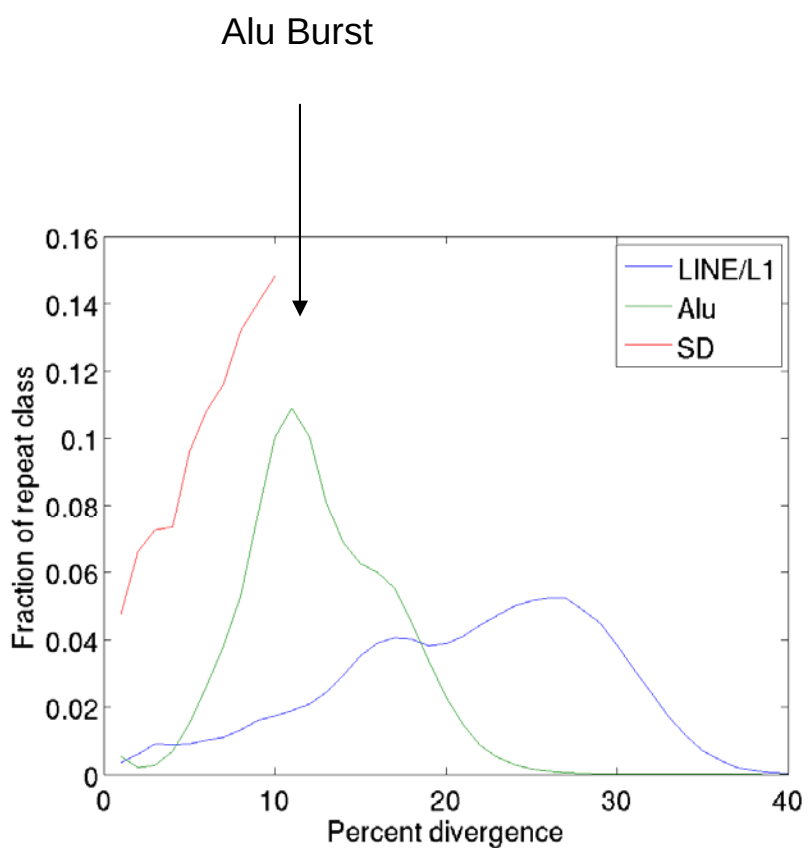


ANALYZING SEQUENCED BREAKPOINTS CONFIRMS THE RESULTS FROM THE COARSE GRAINED ANALYSIS

Repeat Type	Frequency	Global enrichment	p-value	Local enrichment	p-value
Alu	0.09	0.94	3.24E-01	1.13	1.74E-01
SD	0.41	2.57	2.14E-07	1.17	2.64E-01
L1	0.24	1.48	1.03E-07	1.12	7.16E-02
L2	0.01	0.47	1.72E-02	0.52	2.31E-02
Microsatellite	0.03	3.91	6.74E-11	3.11	2.99E-07
LTR	0.09	1.14	1.71E-01	0.89	1.97E-01
PPgene	0.01	2.08	9.55E-02	1.66	1.98E-01
GC	0.39	0.96	7.24E-03	0.97	3.00E-02

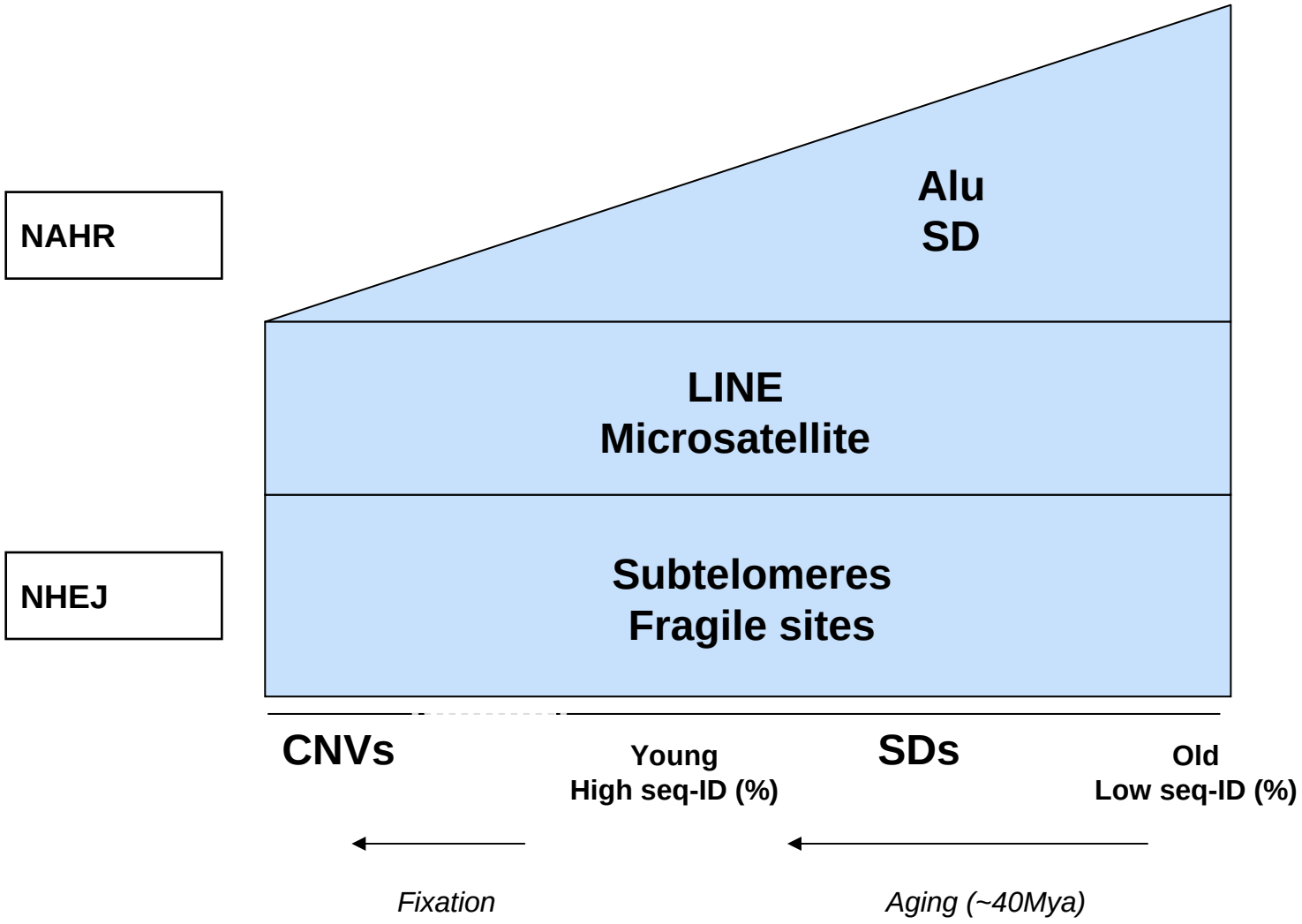


AFTER THE ALU BURST, THE IMPORTANCE OF ALU ELEMENTS FOR GENOME REARRANGEMENT DECLINED RAPIDLY



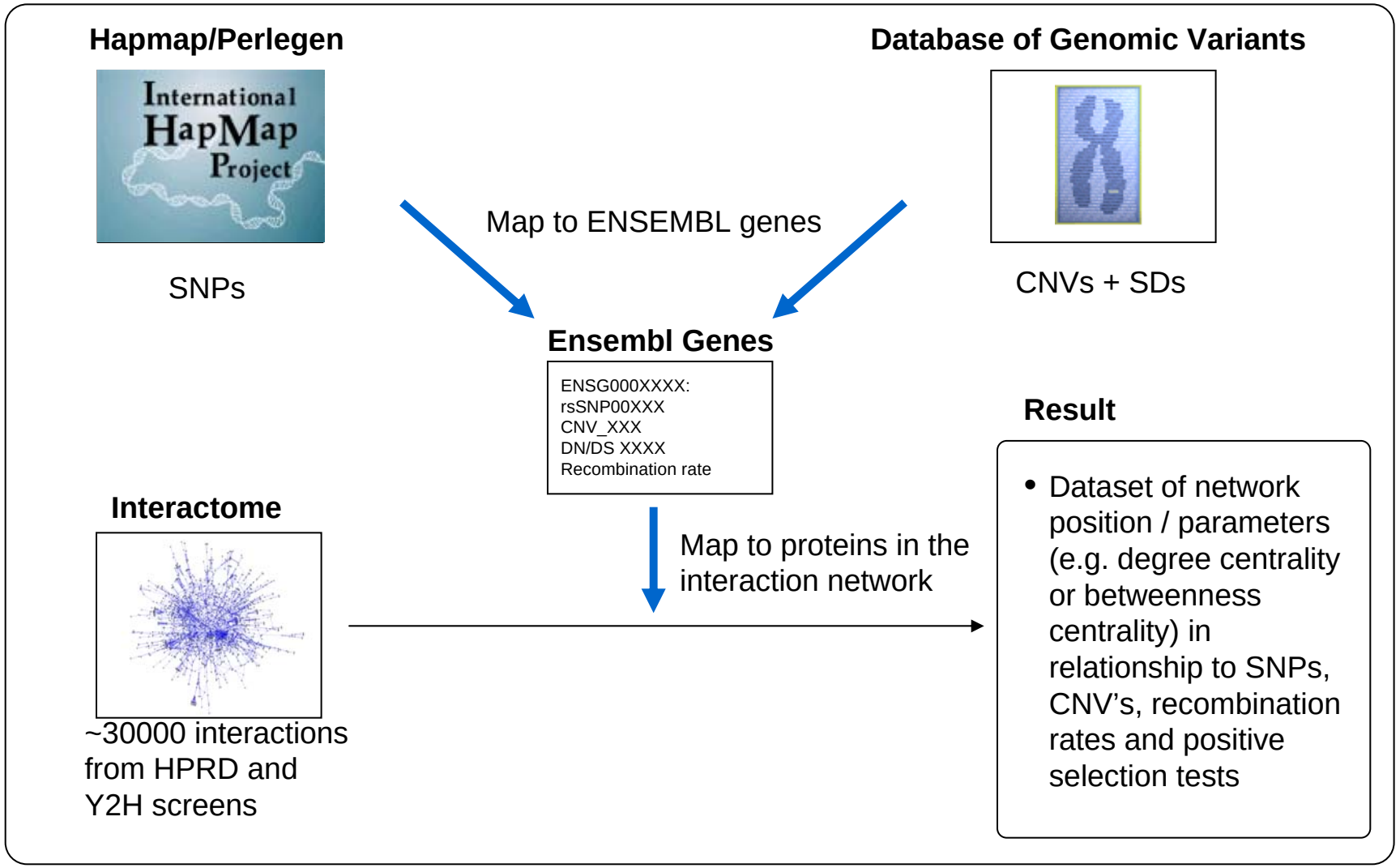
- About 40 million years ago there was a burst in retrotransposon activity
- The majority of Alu elements stem from that time
- This, in turn, led to rapid genome rearrangement via NAHR
- The resulting SDs, could create more SDs, but with Alu activity decaying, their creation slowed

THE MECHANISM DRIVING LARGE SCALE GENOME REARRANGEMENT UNDERWENT A MARKED SHIFT IN THE LAST 40 MYA



METHODOLOGY: MAP SNP AND CNV DATA ONTO ENSEMBL GENES, AND THEN MAP ENSEMBL GENES TO THE KNOWN INTERACTOME

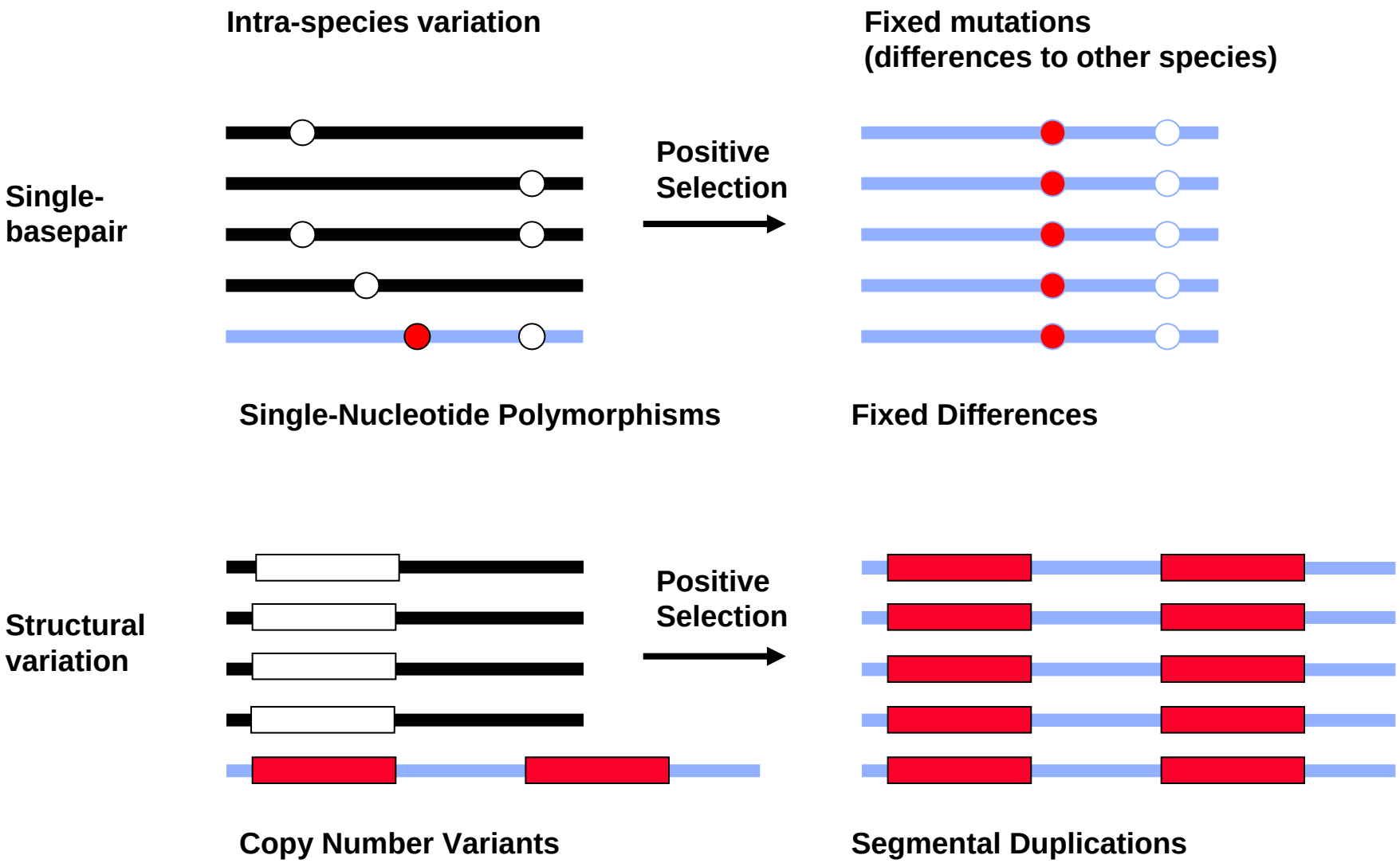
ILLUSTRATIVE



* From Nielsen et al. *PLoS Biol.* (2005) and Bustamante et al. *Nature* (2005)

Source: PMK

ADAPTIVE EVOLUTION CAN BE SEEN ON TWO DIFFERENT LEVELS

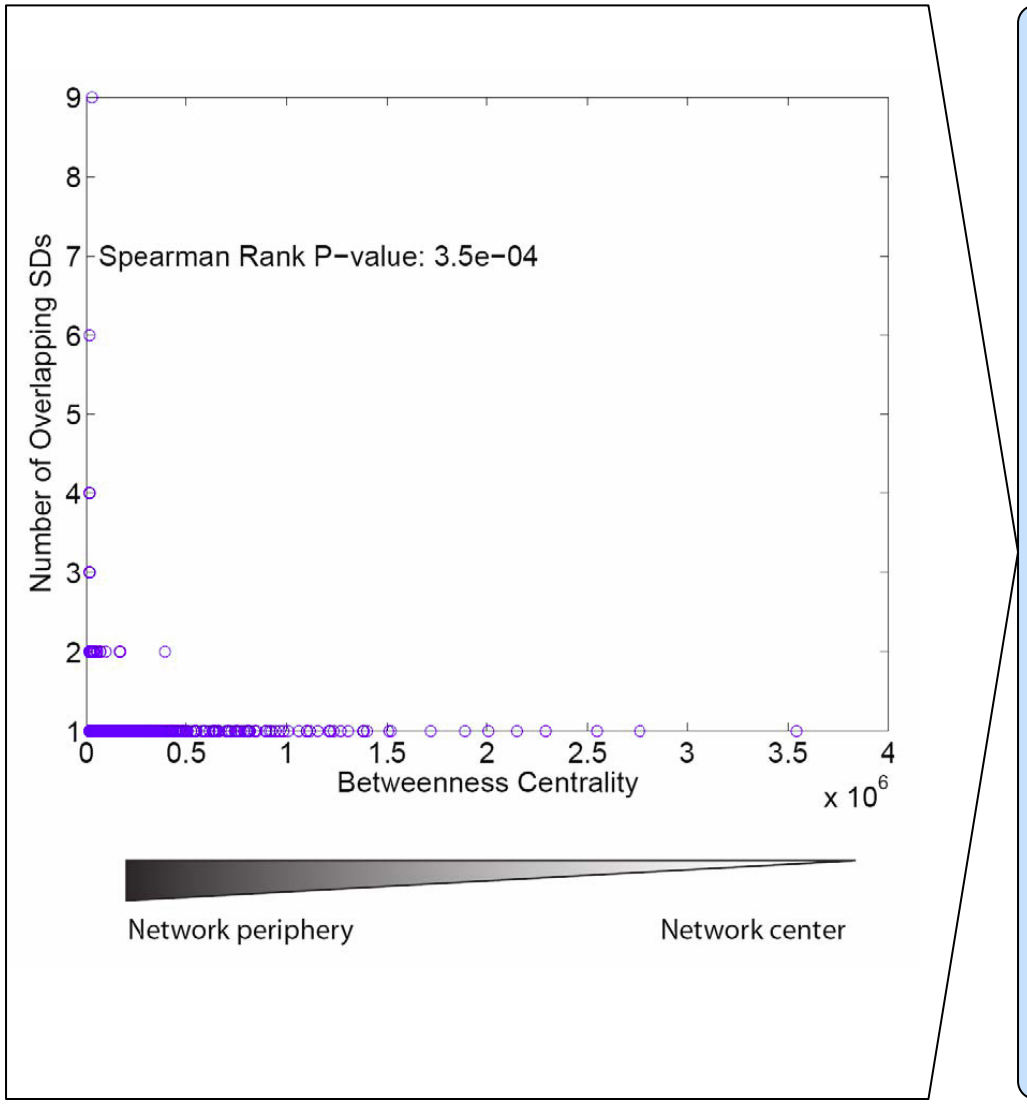


[Often but not always measured by dN/dS]

Source: PMK

CENTRAL NODES ARE LESS LIKELY TO LIE INSIDE OF SDs

Centrality vs. SD occurrence



Reasoning

- This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.
- Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.
- Note that despite the small size of our dataset for known SD's we get significant correlations. It is to be expected that the correlations will get clearer as more data emerges*

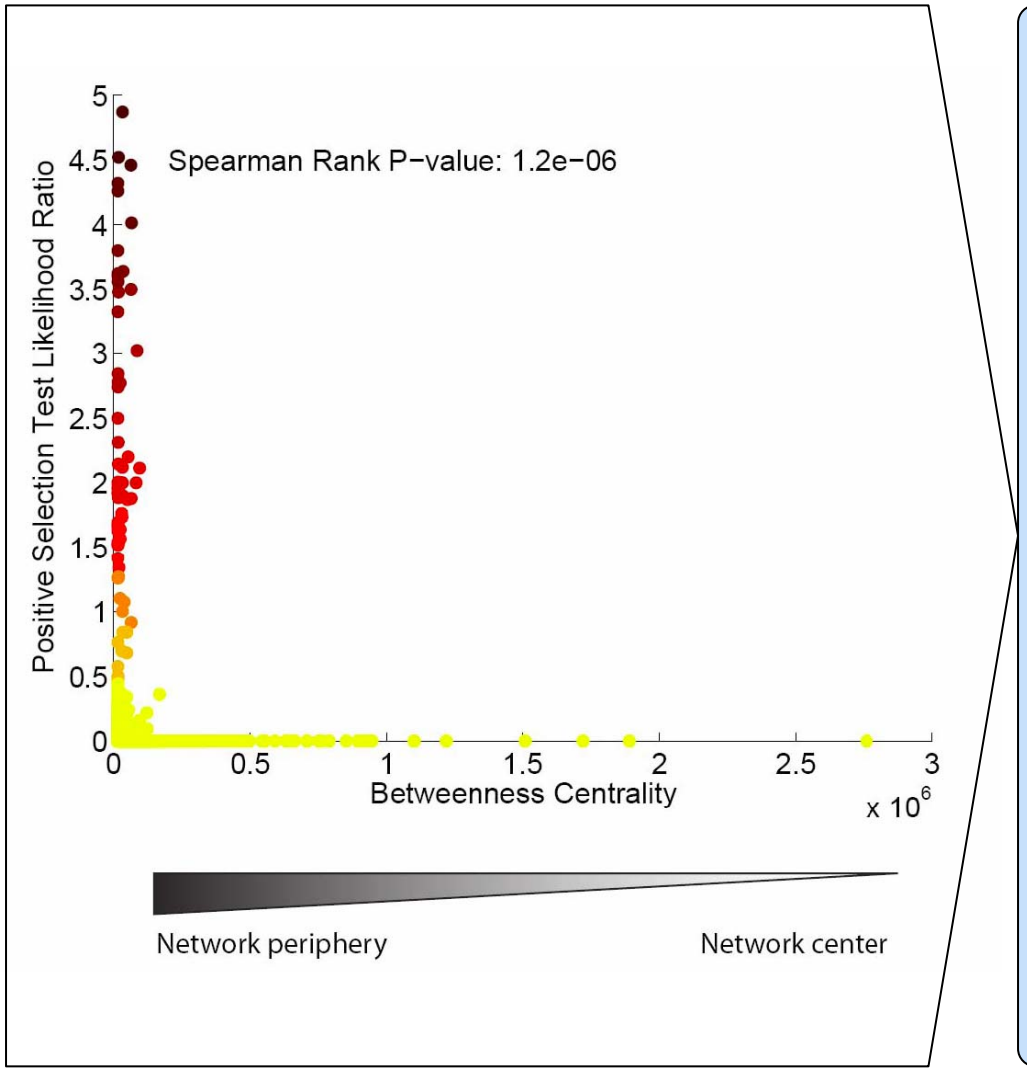
* Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome

Source: Database of genetic variation, HPRD, Rual et al. *Nature* (2005), and Kim et al. PNAS (2007)

CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

Hubs

Degree vs. Positive Selection



Reasoning

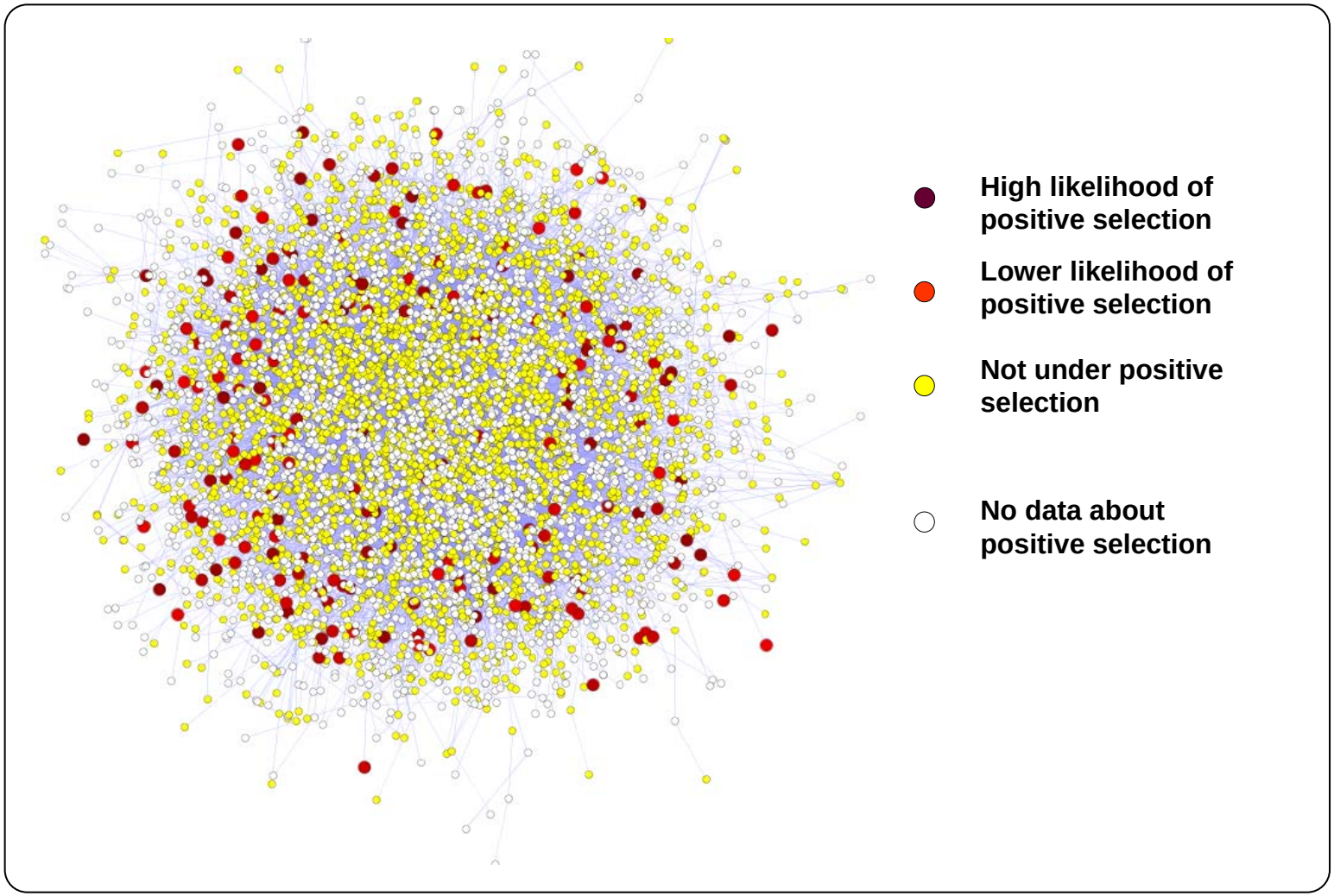
- Peripheral genes are likely to under positive selection, whereas hubs aren't
- This is likely due to the following reasons:
  - Hubs have stronger structural constraints, the network periphery doesn't
  - Most recently evolved functions (e.g. "environmental interaction genes" such as sensory perception genes etc.) would probably lie in the network periphery
- Effect is independent of any bias due to gene expression differences

* With a probability of over 80% to be positively selected as determined by Ka/Ks. Other tests of positive selection (McDonald Kreitmann and LDD) corroborate this result.

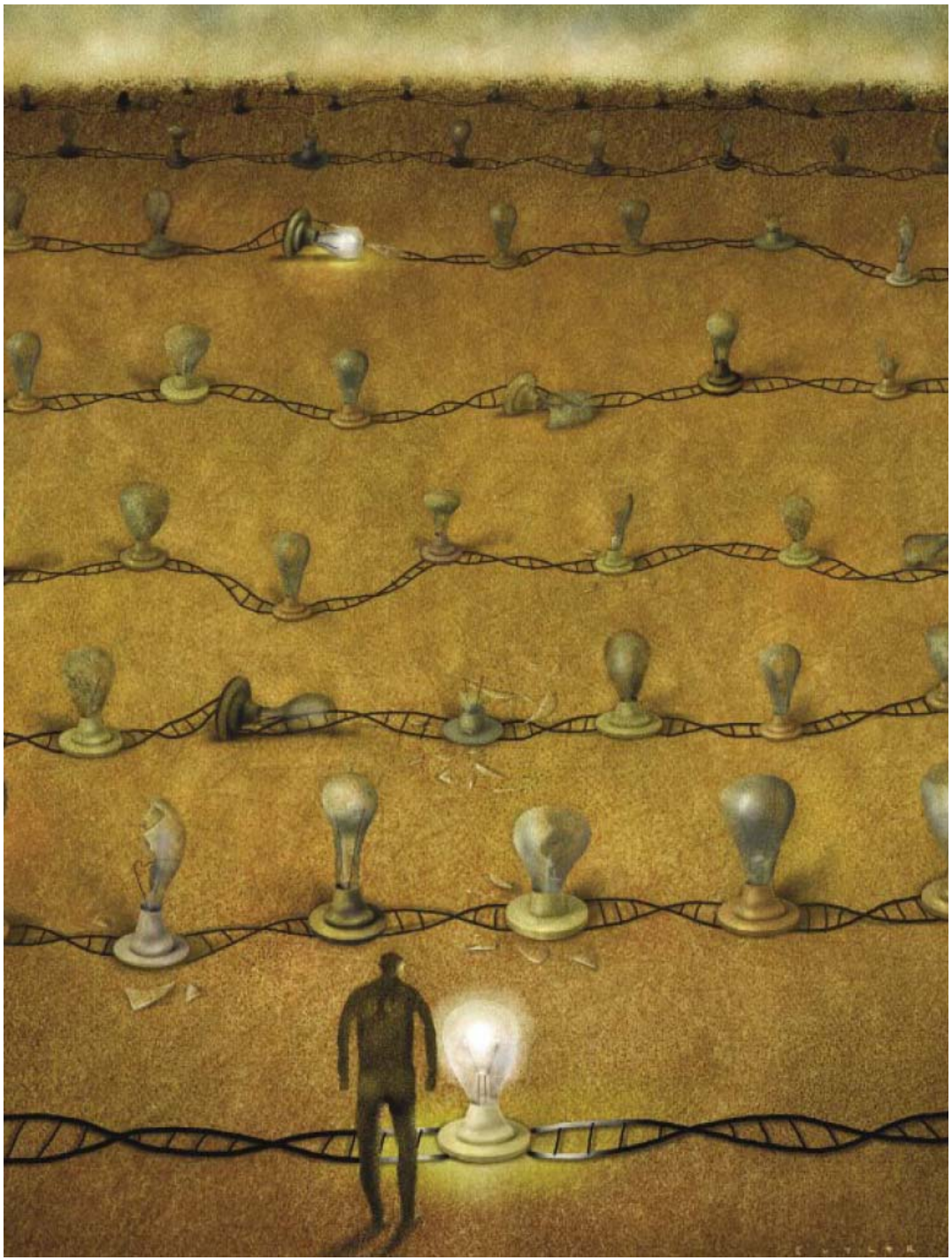
Source: Nielsen et al. *PLoS Biol.* (2005), Bustamante et al. *Nature* (2005), HPRD, Rual et al. *Nature* (2005), and Kim et al. PNAS (2007)

POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome



Source: Nielsen et al. *PLoS Biol.* (2005), HPRD, and Kim et al. PNAS (2007)



Integrative Analyses:

Annotating Pseudogenes and relating them to functional signal and measures of conservation

Illustration from Gerstein & Zheng (2006). Sci Am.

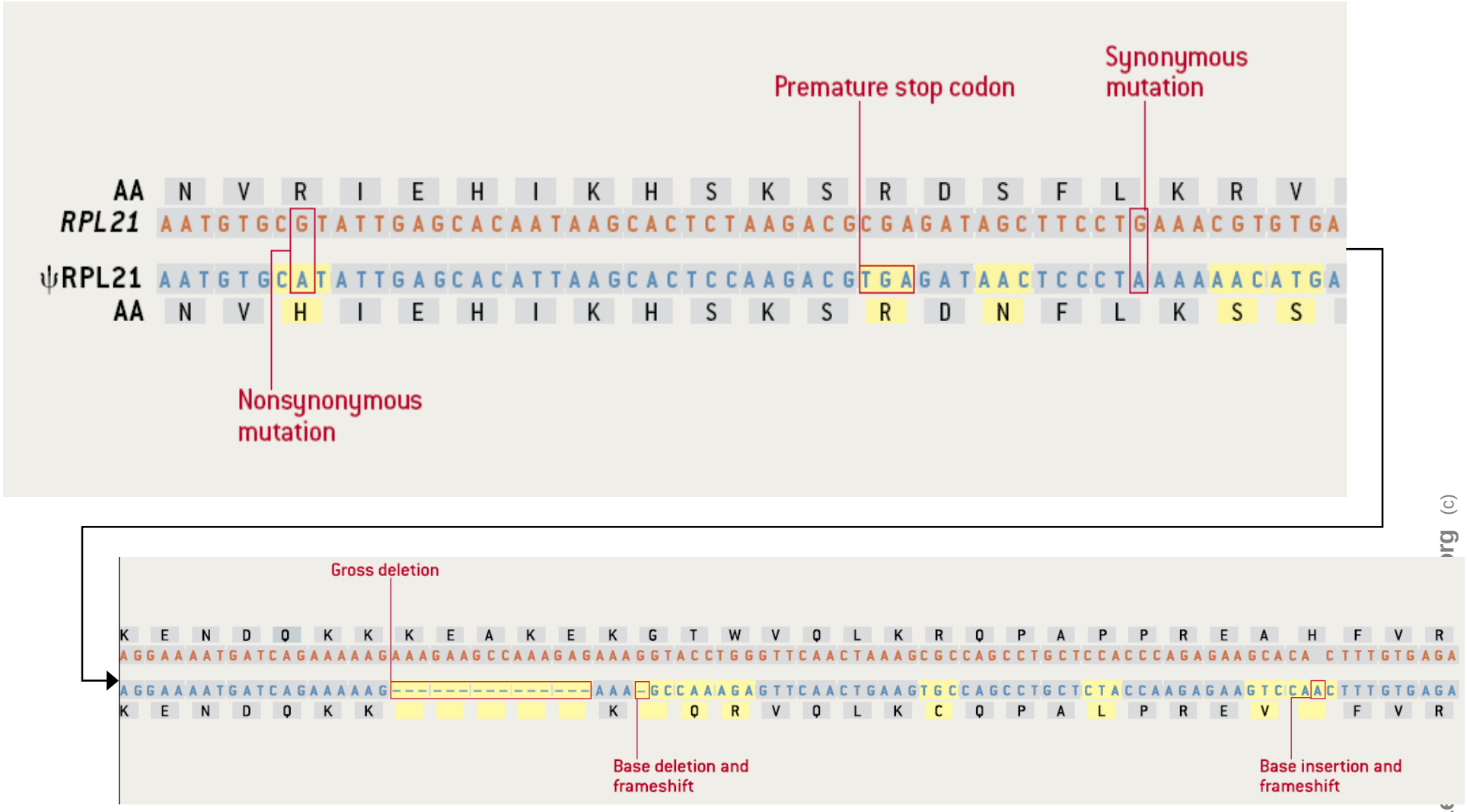
127 Lectures.GersteinLab.org (c)

Pseudogenes are among the most interesting intergenic elements

- Formal Properties of Pseudogenes (ΨG)
  - ◊ Inheritable
  - ◊ Homologous to a functioning element
  - ◊ Non-functional*
    - No selection pressure so free to accumulate mutations
      - – Frameshifts & stops
      - – Small Indels
      - – Inserted repeats (LINE/Alu)
    - **What does this mean?** no transcription, no translation?…

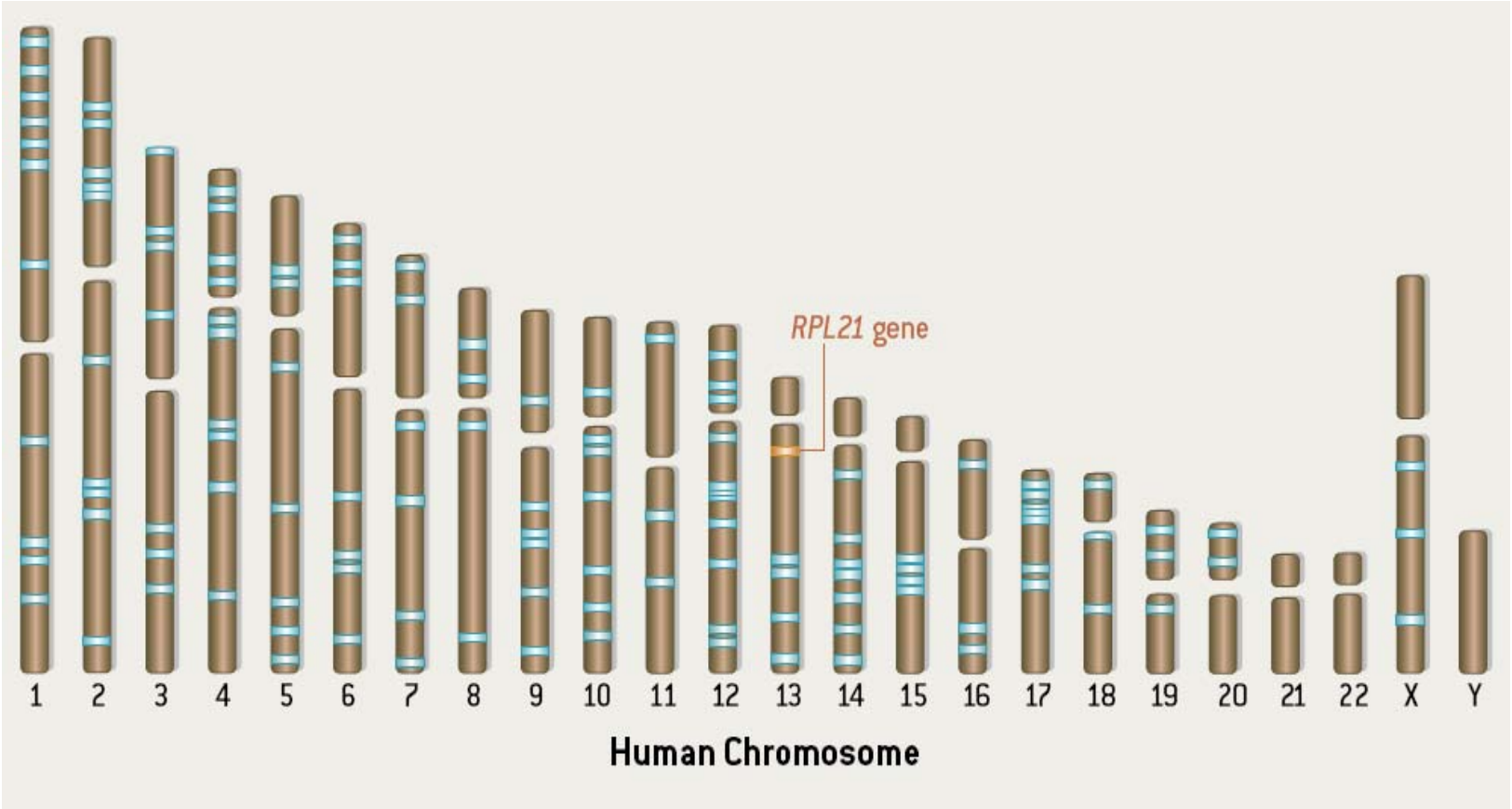
[Mighell et al. *FEBS Letts*, 2000]

Identifiable Features of a Pseudogene (ψRPL21)



Gerstein & Zheng. Sci Am 295: 48 (2006).

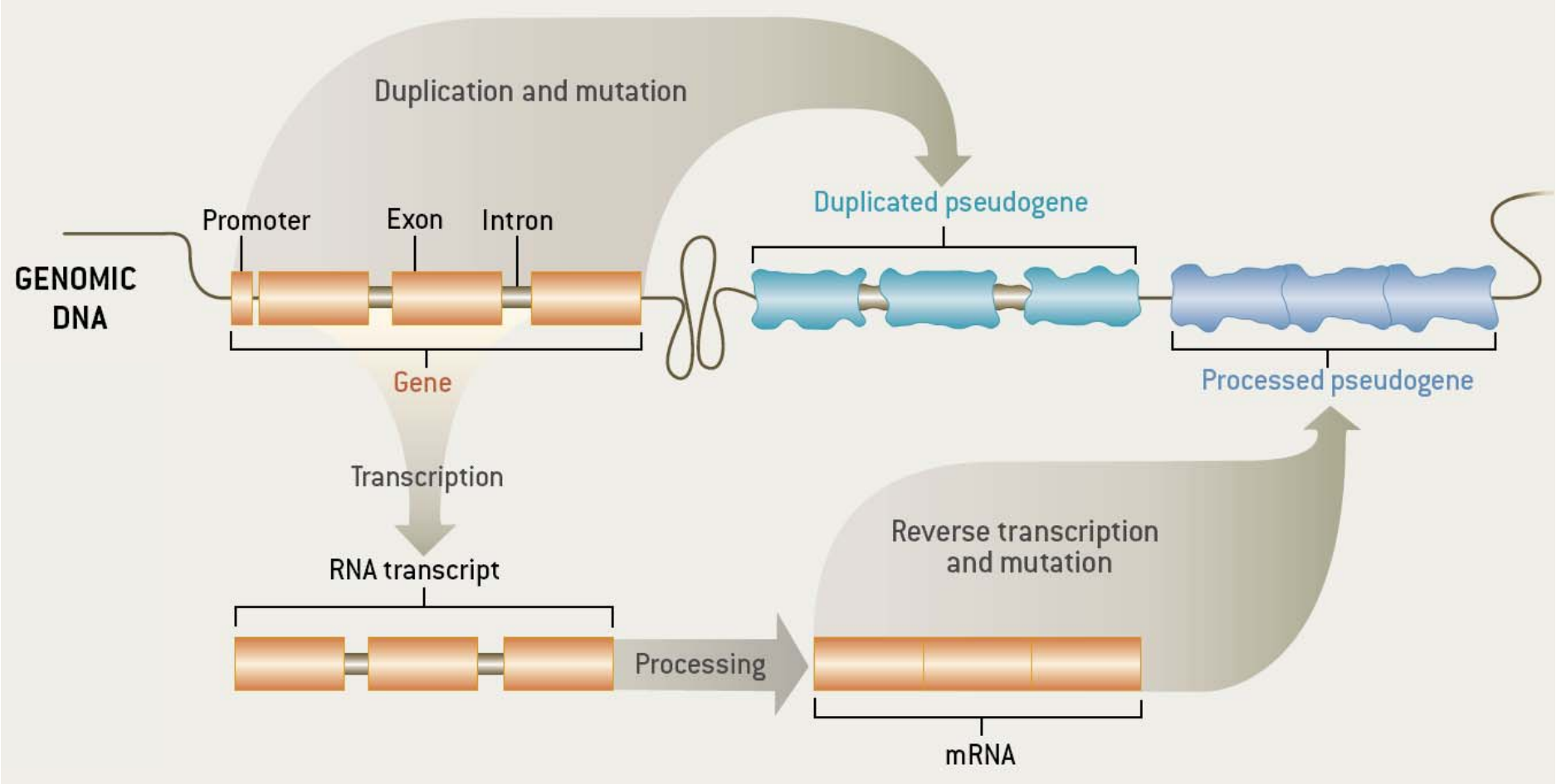
Distribution of Human Pseudogenes (for RPL21) across the chromosomes



Gerstein & Zheng. Sci Am 295: 48 (2006).

130 Lectures.GersteinLab.org (c)

Two Major Genomic Remodeling Processes Give Rise to Distinct Types of Pseudogenes



Gerstein & Zheng. Sci Am 295: 48 (2006).

131 Lectures.GersteinLab.org (c)

Why Study Pseudogenes?

- Important for Doing Accurate Gene Annotation
  - Abundant: > 8000 retropseudogenes in human
  - High sequence similarity with genes
  - 25% in *C. elegans* ? [Mounsay, *Genome Research*, 2002]

- Interfere with study on functional genes
  - Cross-hybridation in micro-array and RT-PCR. [Ruud, *Int. J. Cancer* 1999]
  - Some pseudogenes have regulatory roles

- ΨG are "genomic fossils"
  - Study the evolution of genes and genomes
  - Measure mutation/insertion rates

132 Lectures.GersteinLab.org (c)

Why Study Pseudogenes?

- Cause errors in sequence databases
  - > 8000 retropseudogenes in human
  - Contamination in Ensembl
  - 25% in *C. elegans* ? [Mounsay, *Genome Research*, 2002]

- "Interfere" with functional genes
  - Cross-hybridation in microarray and PCR (Cytokeratin 19, *Int. J. Cancer* 1999)
  - Very rarely this gives some pseudogenes regulatory roles

- ΨG are “genomic fossils”
  - Study the evolution of genes and genomes
  - Measure mutation/insertion rates

In mouse, a pseudogene up-regulates gene expression of *Makorin1* by binding to a transcriptional repressor or an RNA-digesting enzyme [Hirotsune *et al*. *Nature* **423** 2003]

133 Lectures.GersteinLab.org (c)

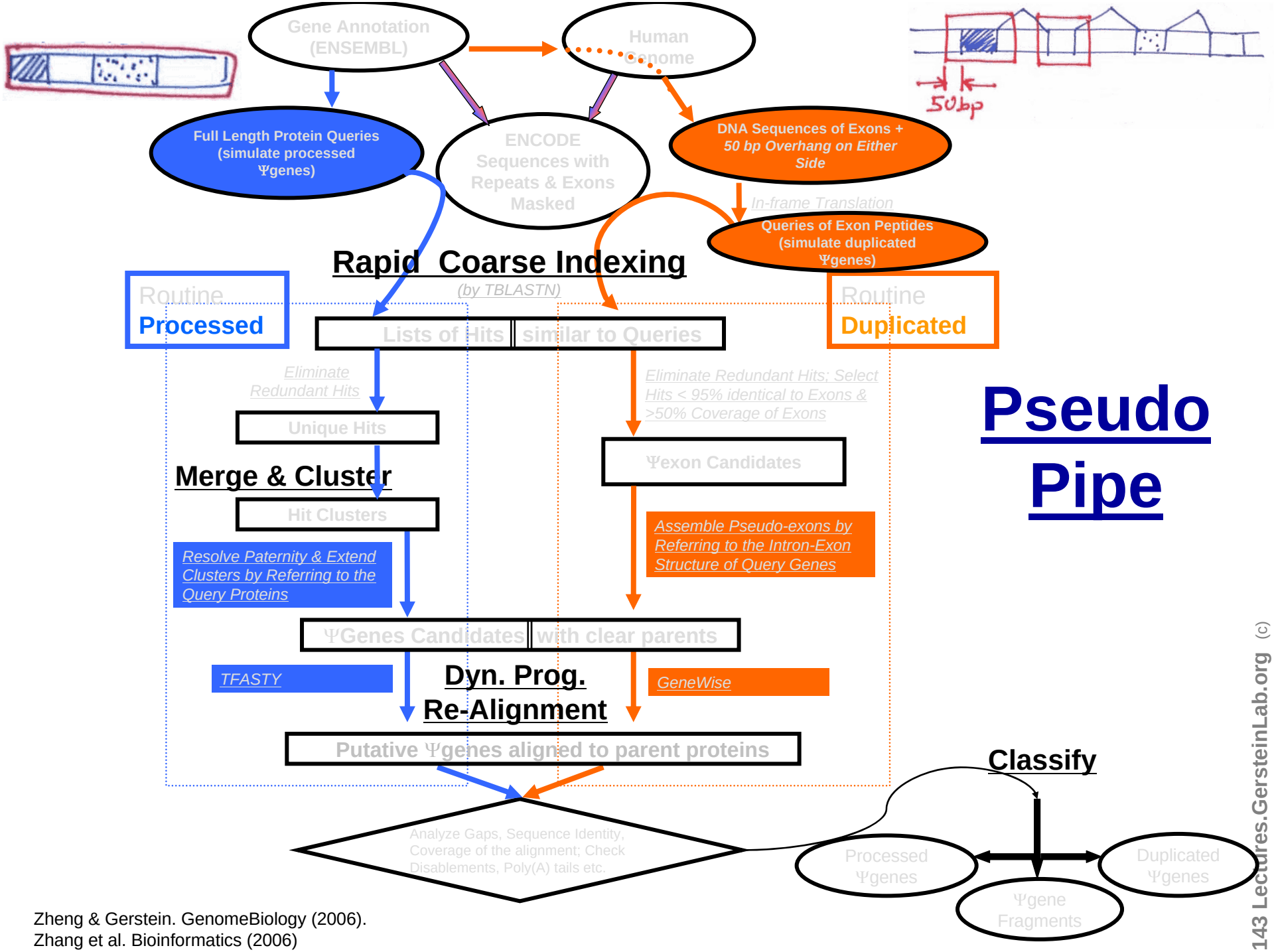
Why Study Pseudogenes?

- Cause errors in sequence databases
  - > 8000 retropseudogenes in human
  - Contamination in Ensembl
  - 25% in *C. elegans* ? [Mounsay, *Genome Research*, 2002]

- Interfere with study on functional genes
  - Cross-hybridation in micro-array and RT-PCR. [Ruud, *Int. J. Cancer* 1999]
  - Some pseudogenes have regulatory roles

- ΨG are "genomic fossils"
  - Study the evolution of genes and genomes
  - Measure mutation/insertion rates

134 Lectures.GersteinLab.org (c)



Pseudo Pipe

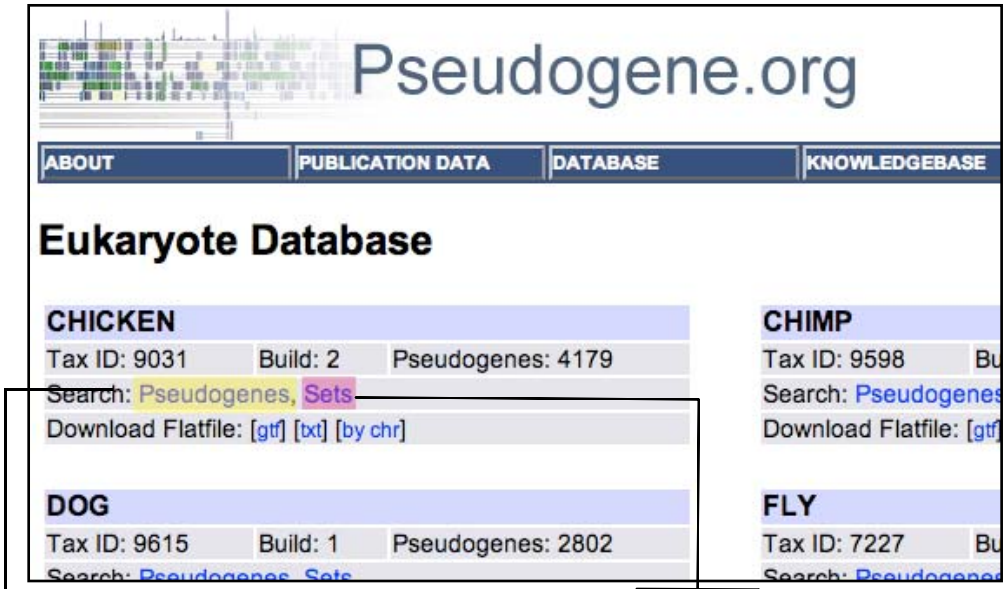
Zheng & Gerstein. GenomeBiology (2006).
Zhang et al. Bioinformatics (2006)

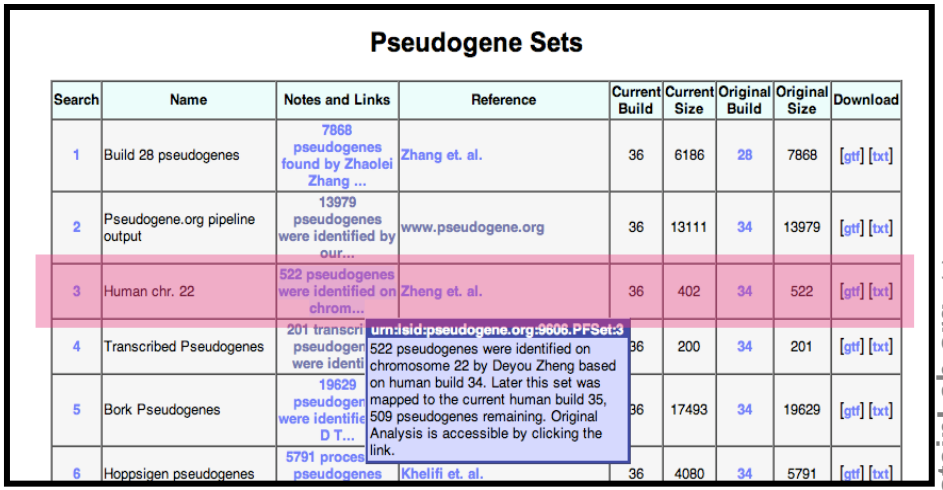
143 Lectures.GersteinLab.org (c)

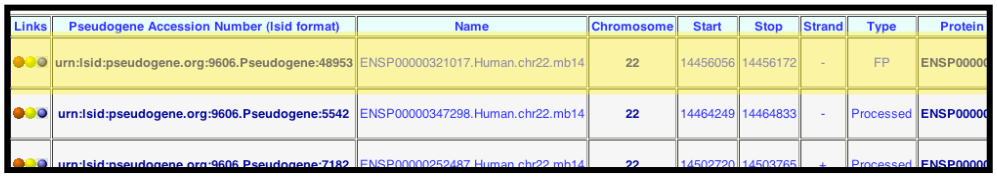
- Integrating heterogeneous, **Dynamically Changing Annotation**
  - ◊ Changing sequences, gene predictions, repeats
- Track (slightly) changing objects across genome builds
  - ◊ Versioning and exact temporal reconstructability
- Fixed **Sets** of Pseudogenes
  - ◊ Corresponding to particular types of analyses or papers
- Generalizable **Class** Structure
  - ◊ fragments, alignments, collections, pseudogenes
- **EAV**
  - ◊ Flexible Annotation for extended characteristics
- Interface with Uniprot & UCSC

Karro et al., NAR (2007)

DB







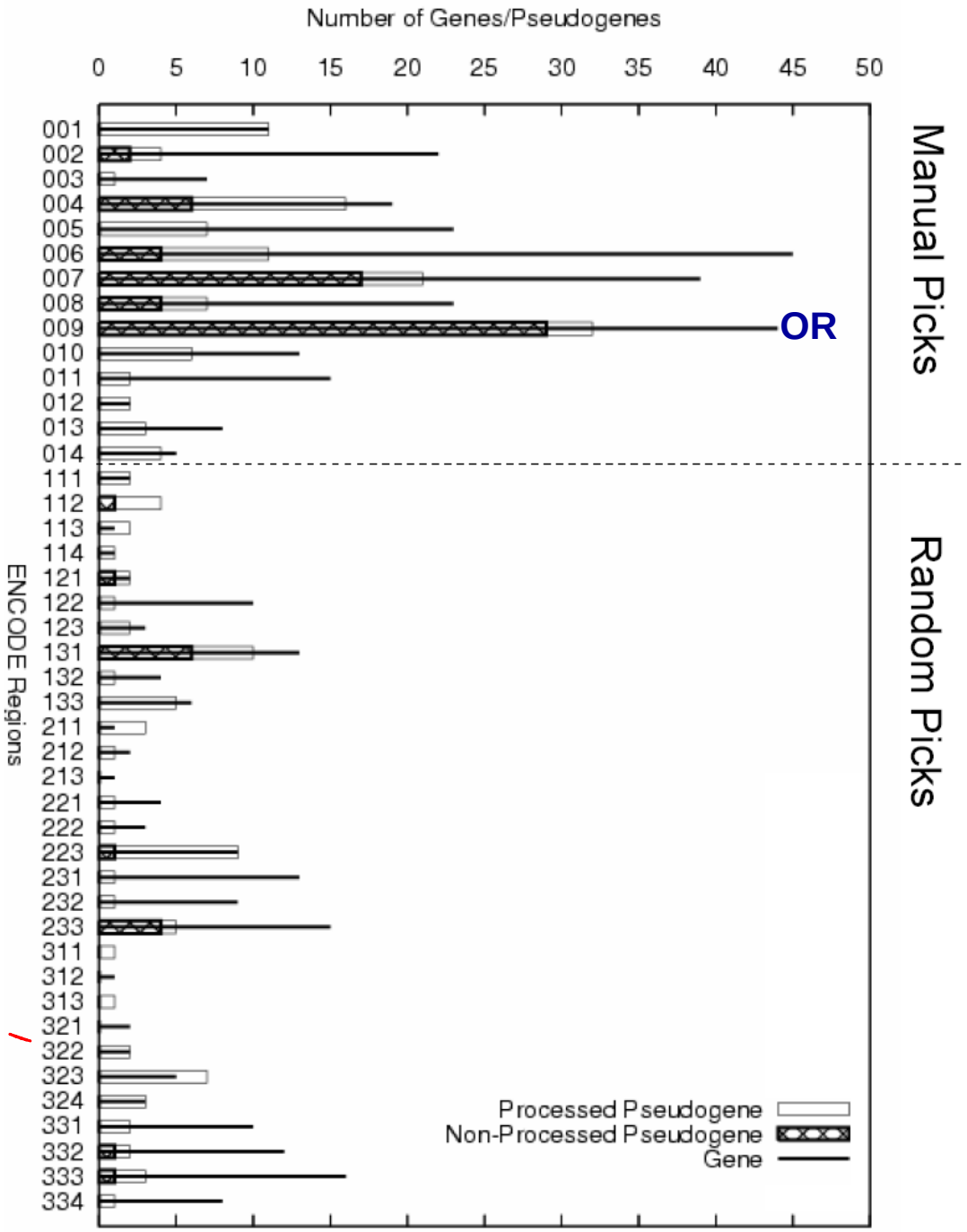
144 Lectures.GersteinLab.org (c)

5 Methods of Assignment

- **4 automatic pipelines**
  - ◊ **retroFinder+pseudoFinder (UCSC), PseudoPipe (Yale), GIS**
  - ◊ Comparing protein or transcript v genomic DNA, filtering, application of rules
- **HAVANA manual**
- What is a pseudogene?
  - ◊ Different criteria
- Conservative approach here
  - ◊ Can't overlap gene annotation
  - ◊ Need to have a protein alignment
  - ◊ 201 pseudogenes vs ~400 genes

Zheng et al. (2007) Gen. Res.

148 Lectures.GersteinLab.org (c)



Overall Results: Regional Distribution

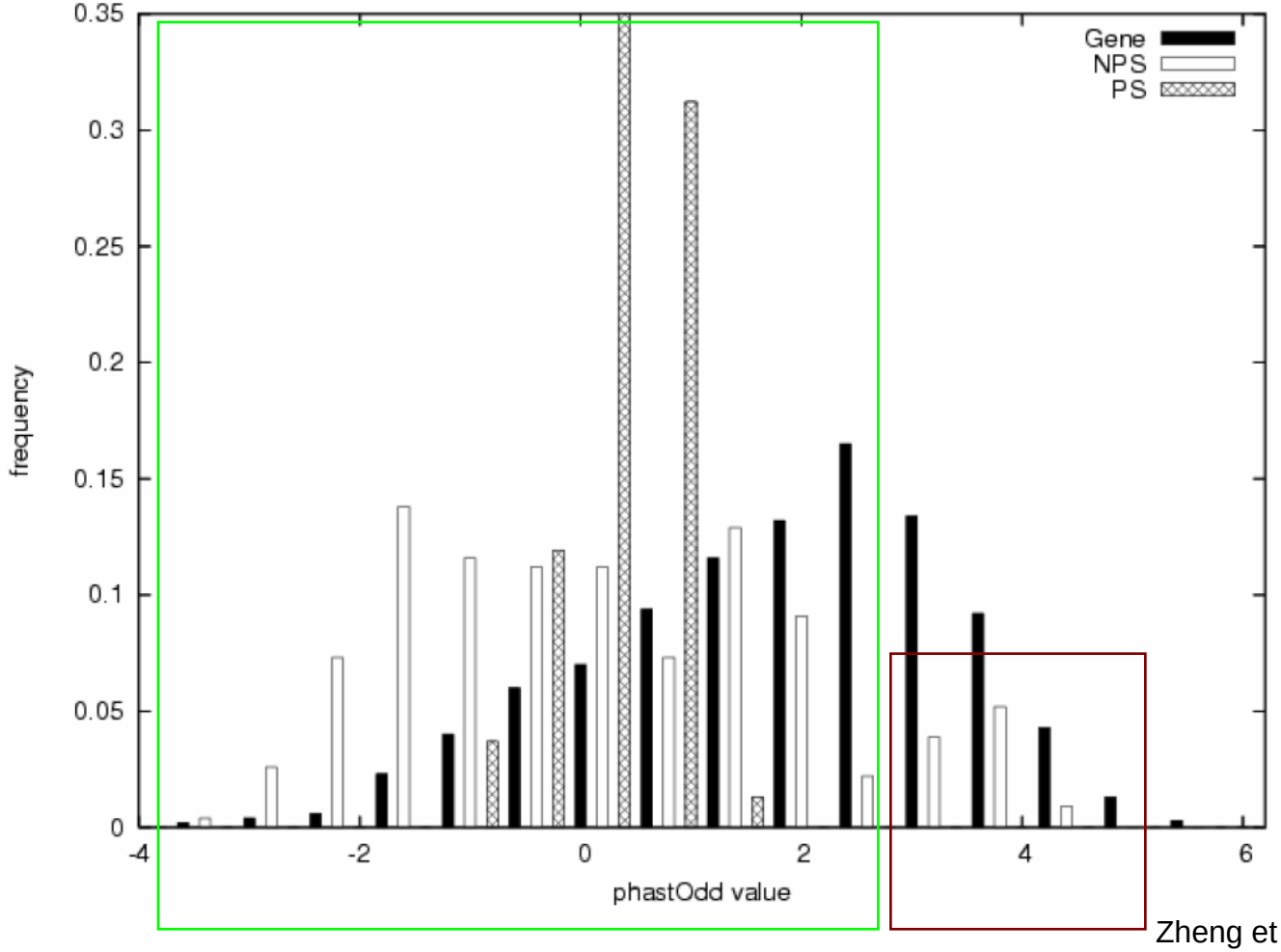
201 pseudogenes
77 non-processed
124 processed

Zheng et al. (2007) Gen. Res.

browser
+
pseudogene.org/ENCODE

151 Lectures.GersteinLab.org (c)

Using phastOdd value to examine neutral evolution of pseudogenes



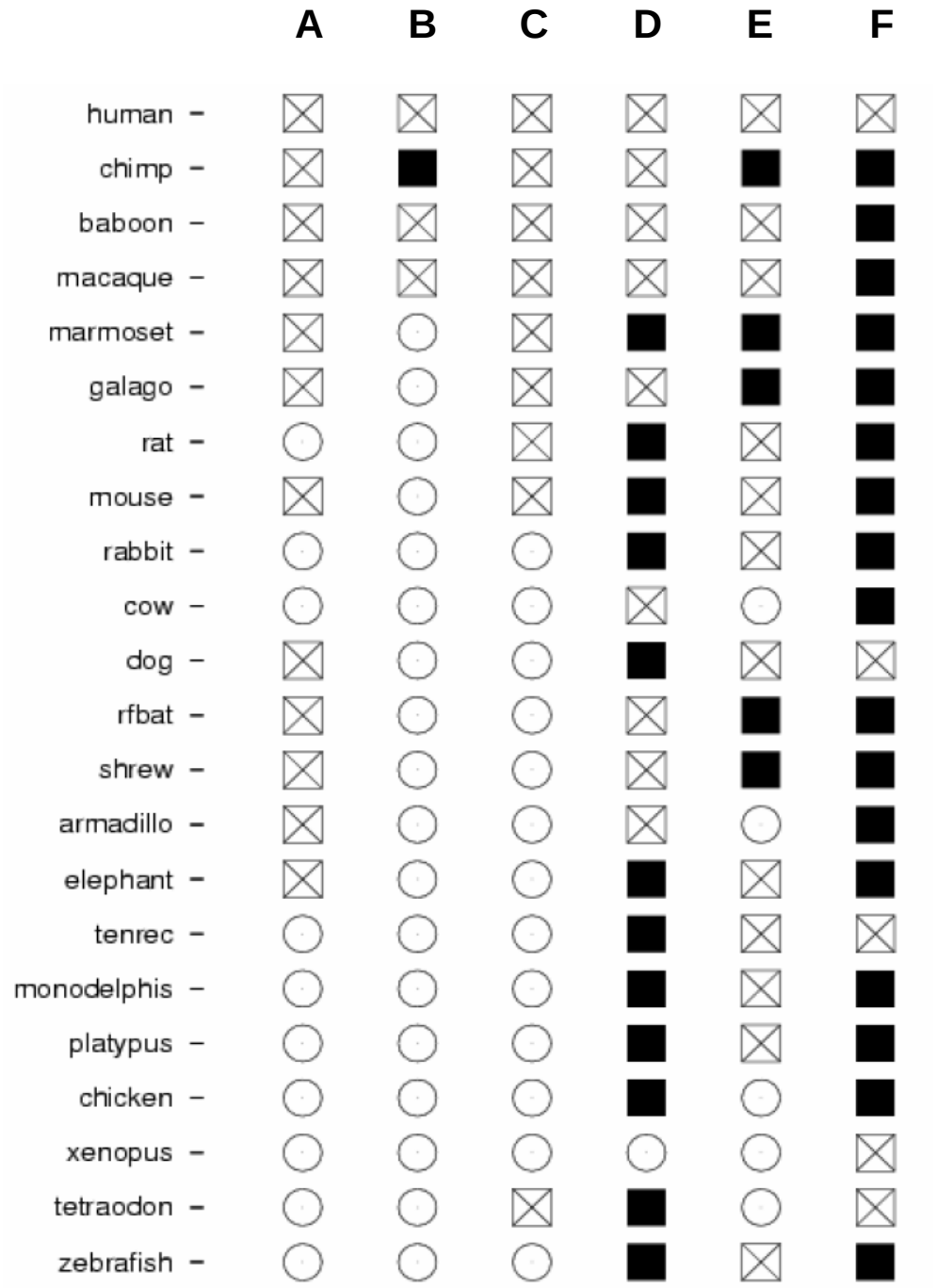
most good candidates for studying mutational processes

a few non-proc. ψG under constraint

Zheng et al. (2007) Gen. Res.

154 Lectures.GersteinLab.org (c)

representative pseudogenes drawn from 201 total



History of Pseudogene Preservation

Based on alignment from ENCODE MSA group

Zheng et al. (2007) Gen. Res.

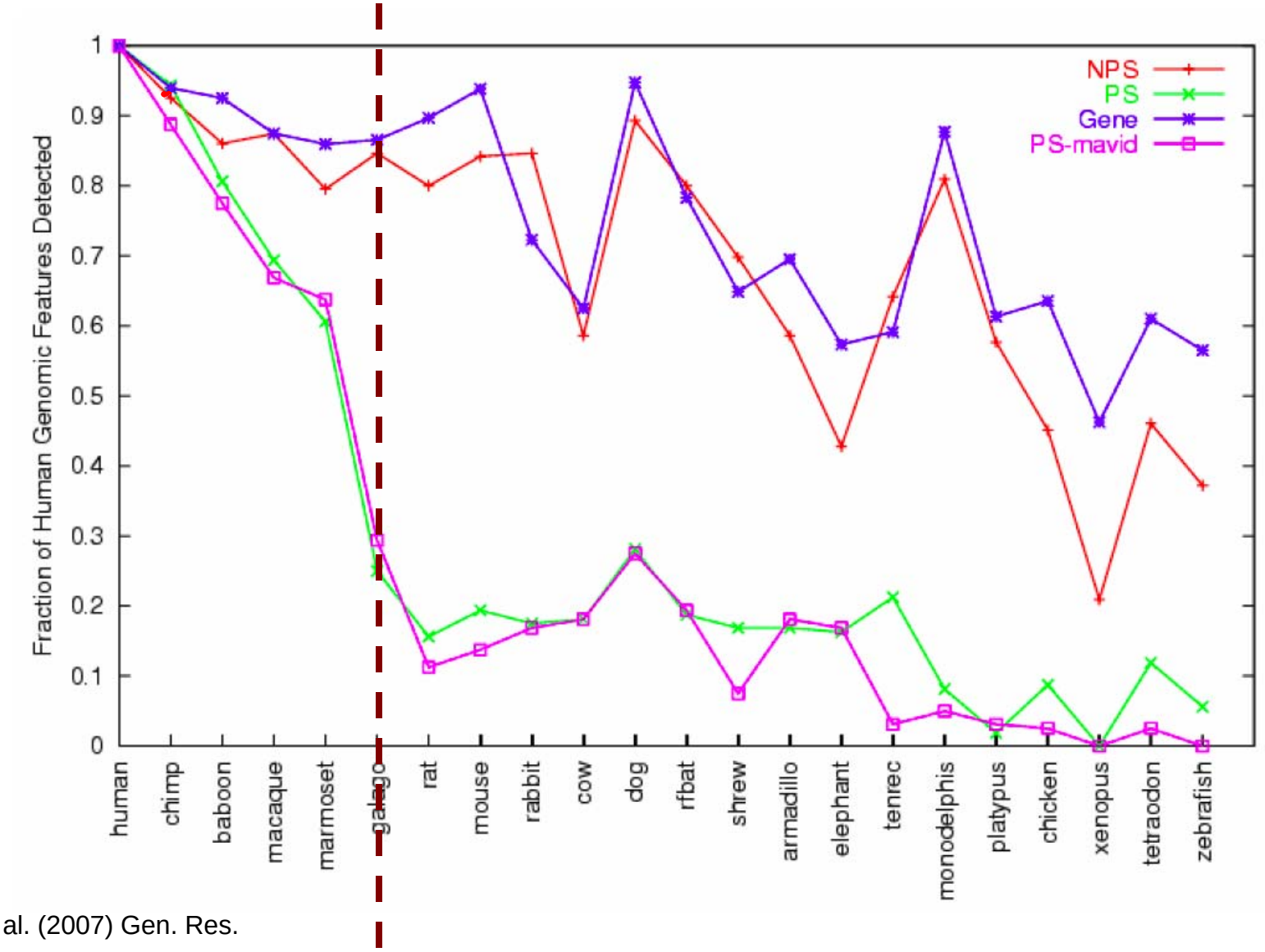
Absent

Present with Disablement

Present without Disablement

155 Lectures.GersteinLab.org (c)

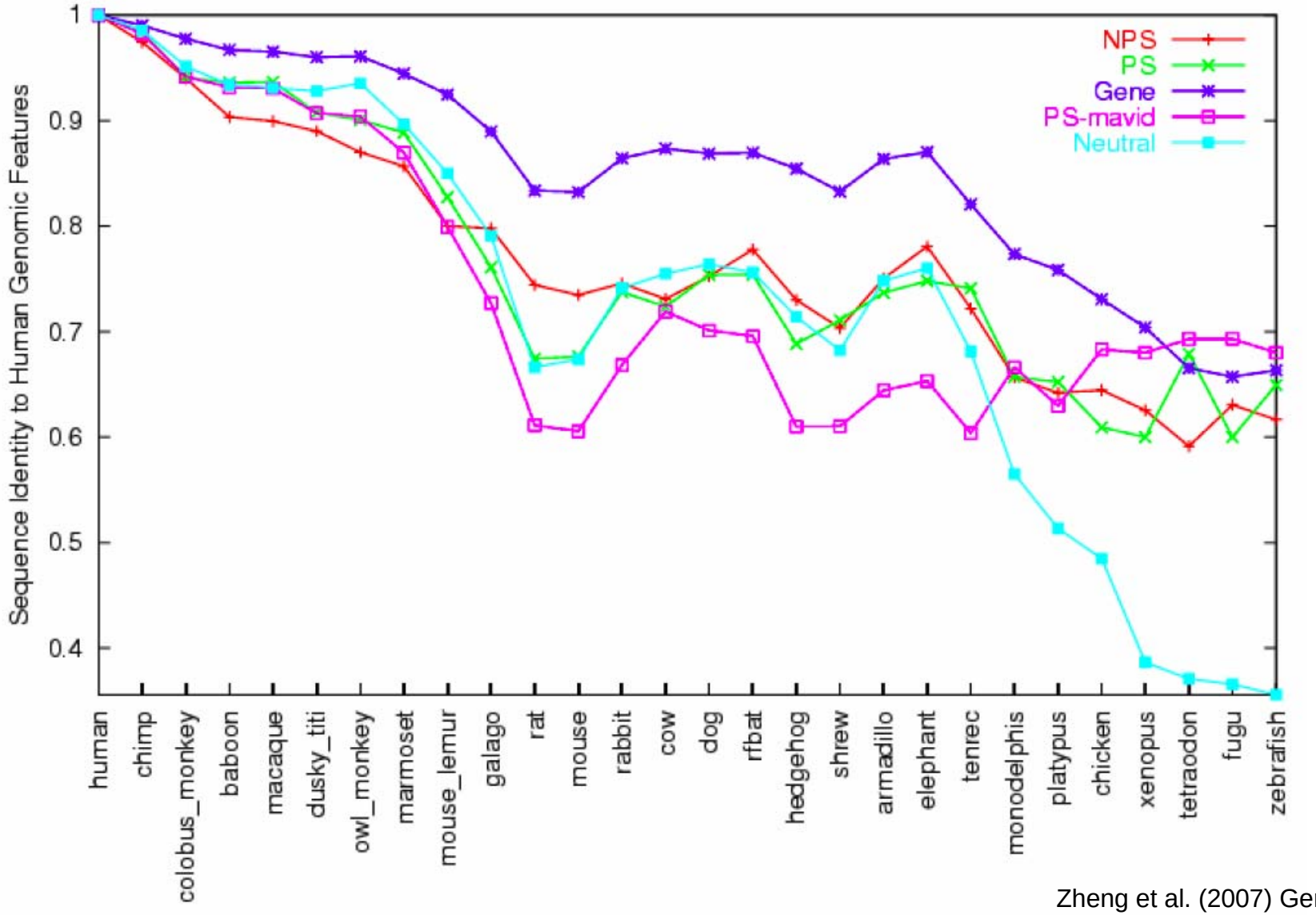
Most Processed Pseudogenes are Primate Specific Created by Recent (<45 MYA) Retrotranspositional Activity



Zheng et al. (2007) Gen. Res.

156 Lectures.GersteinLab.org (c)

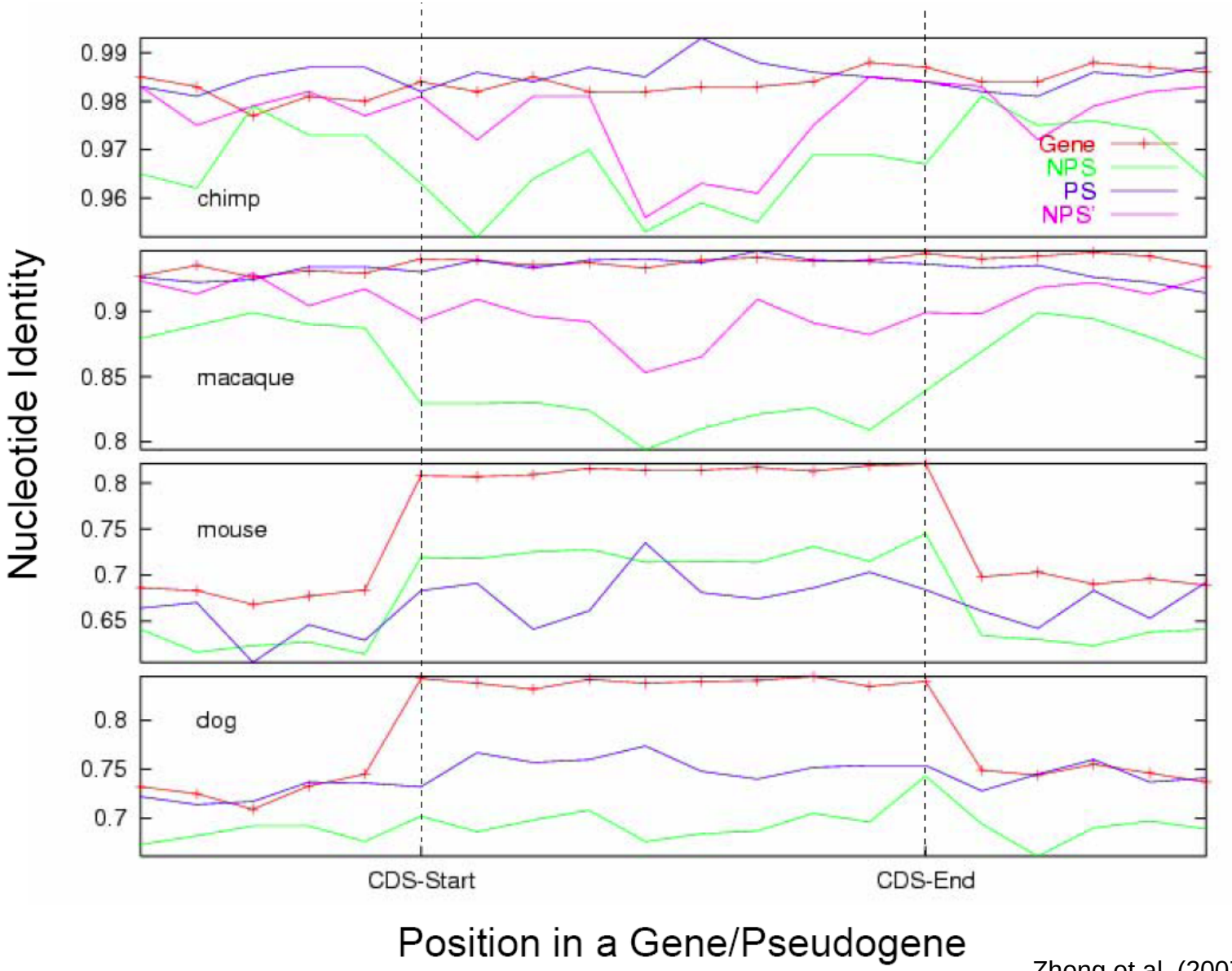
Sequence Decay of Pseudogenes, Approximately Neutral



Zheng et al. (2007) Gen. Res.

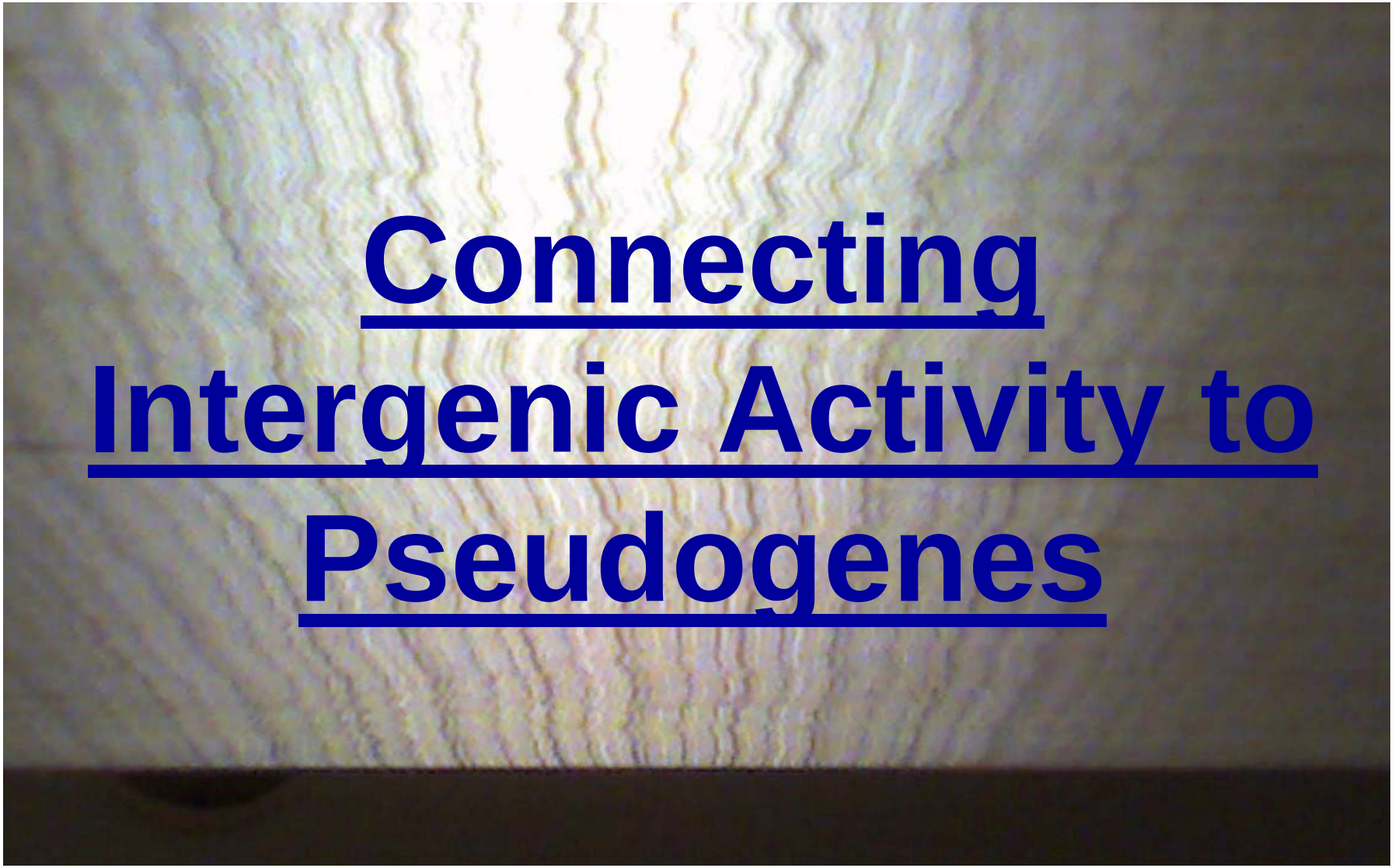
157 Lectures.GersteinLab.org (c)

Sequence Decay of Pseudogenes Relative to their Immediate Genomic Context

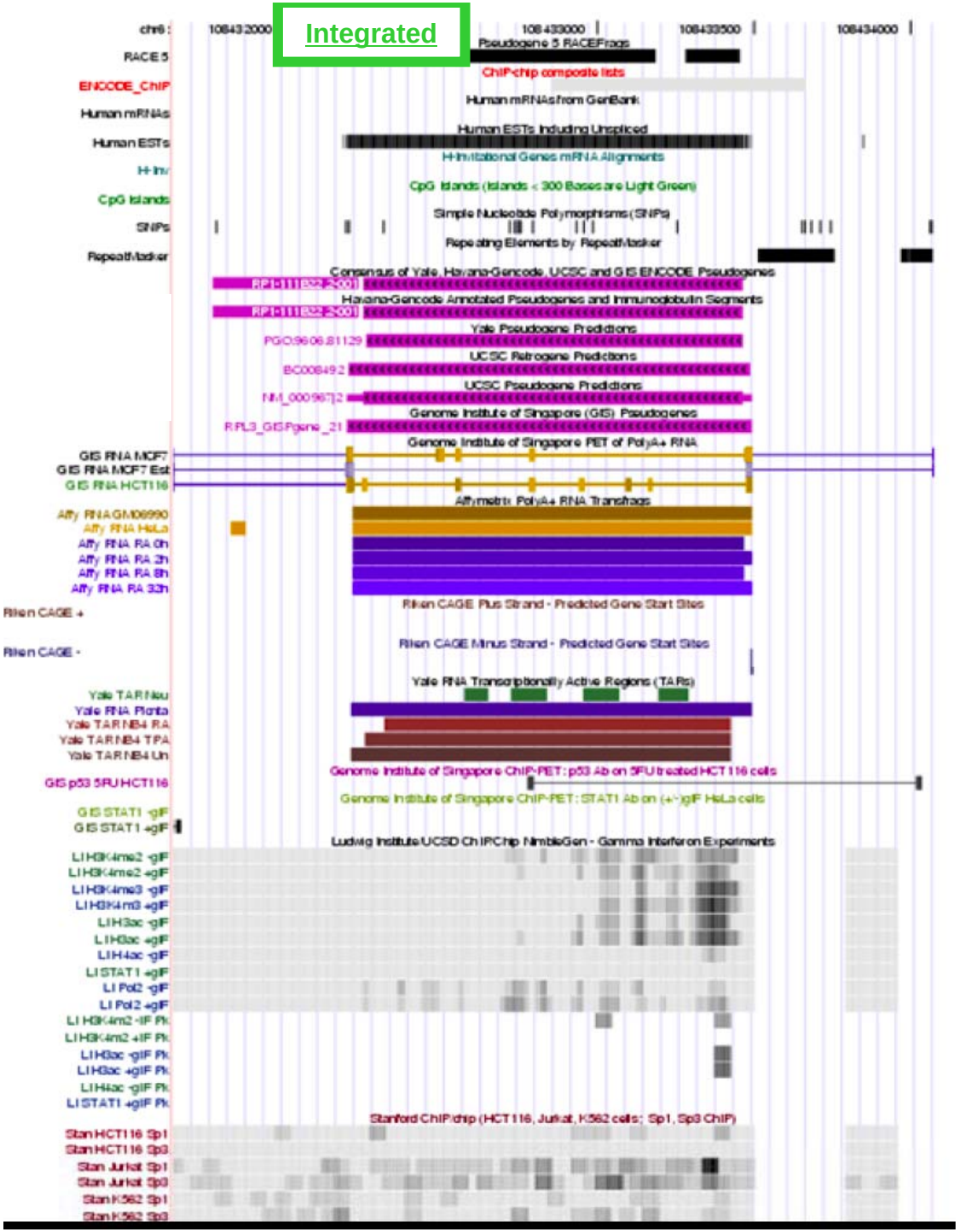


Zheng et al. (2007) Gen. Res.

158 Lectures.GersteinLab.org (c)



159 Lectures.GersteinLab.org (c)



Composite ChIP hit

Special ψG tracks in browser

diTAG

CAGE

TARs

ChIP-chip

Connecting TARs (TxFrags) in Integrative fashion to different types of Annotation

- Single Ex. of Pseudogene Intersecting with Transcriptional and Regulatory Evidence
- Are integrated experiments comparable -- i.e. done on consistent cell lines, on same coordinate sys., &c.

Zheng et al. (2007) Gen. Res.

160 Lectures.GersteinLab.org (c)

Intersection of Pseudogenes with Transcriptional Evidence

	TAR / transfrag	CAGE	DiTag	RACEfrag	EST / mRNA
TAR / transfrag	105 *	8	2	5	14
CAGE		8	1	0	1
DiTag			2	0	0
RACEfrag				14	5
EST / mRNA					21

Excluding TARs (due to cross-hyb issues)

Targeted RACE expts to 160 pseudogenes, gives 14

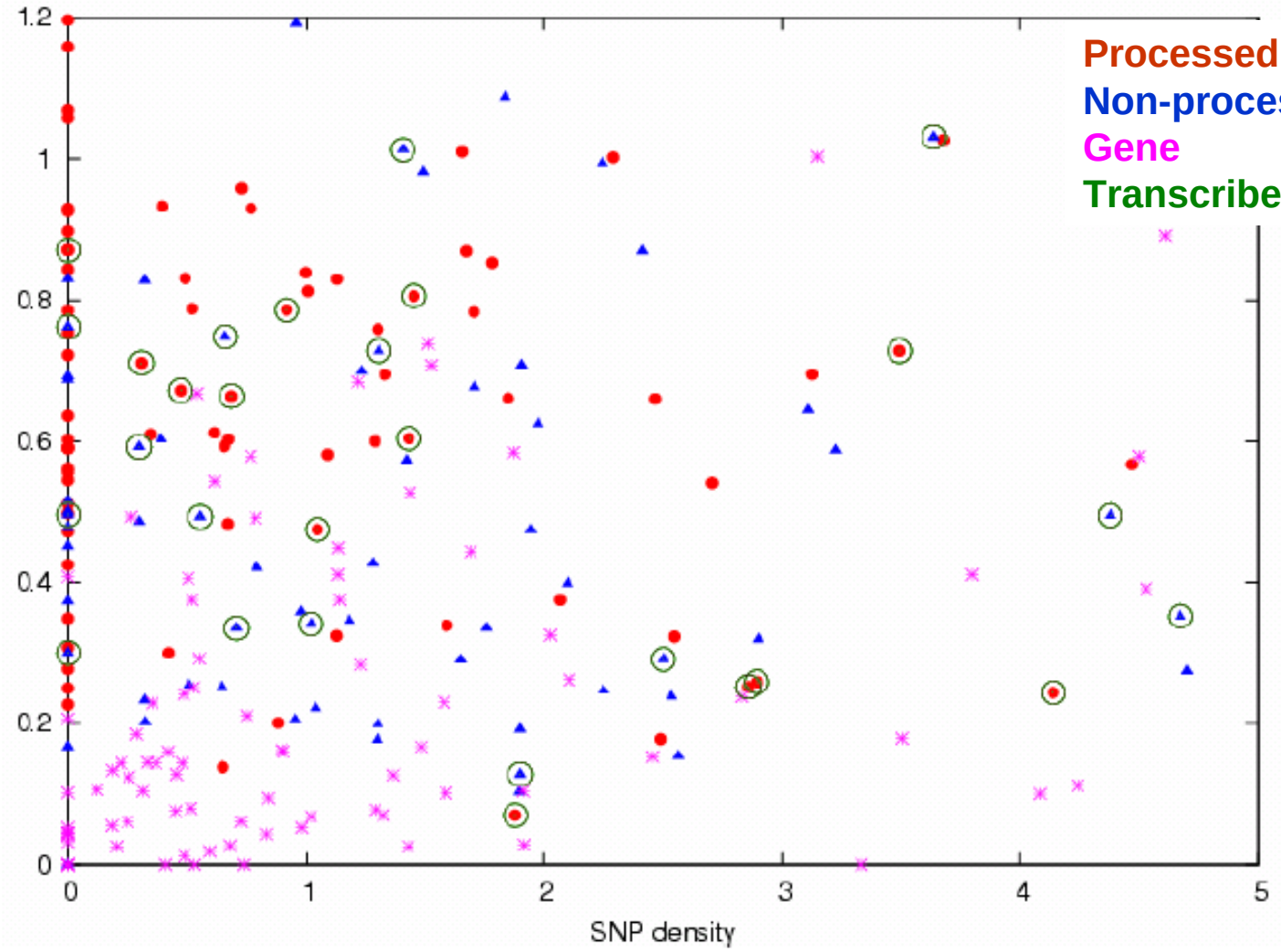
Total Evidence from Sequencing is 38 of 201 (with 5 having cryptic promotors)

Zheng et al. (2007) Gen. Res.

161 Lectures.GersteinLab.org (c)

Integrated

Integrating Transcriptional Evidence with Gene Annotation and Sequence Constraints



Processed pseudogene
Non-processed pseudogene
Gene
Transcribed ○

Avg. Integration over many instances

- No Greater Tendency for Transcribed Pseudogenes to be under Selective Constraint
- Need a way of easily defining degree of constraint on sequence (not so easy for non-coding)

Measurement of Short-time variation (pN+pS)

Zheng et al. (2007) Gen. Res.

162 Lectures.GersteinLab.org (c)

Conclusion: The distinction between gene and non-gene is becoming less clearcut

181 Lectures.GersteinLab.org (c)

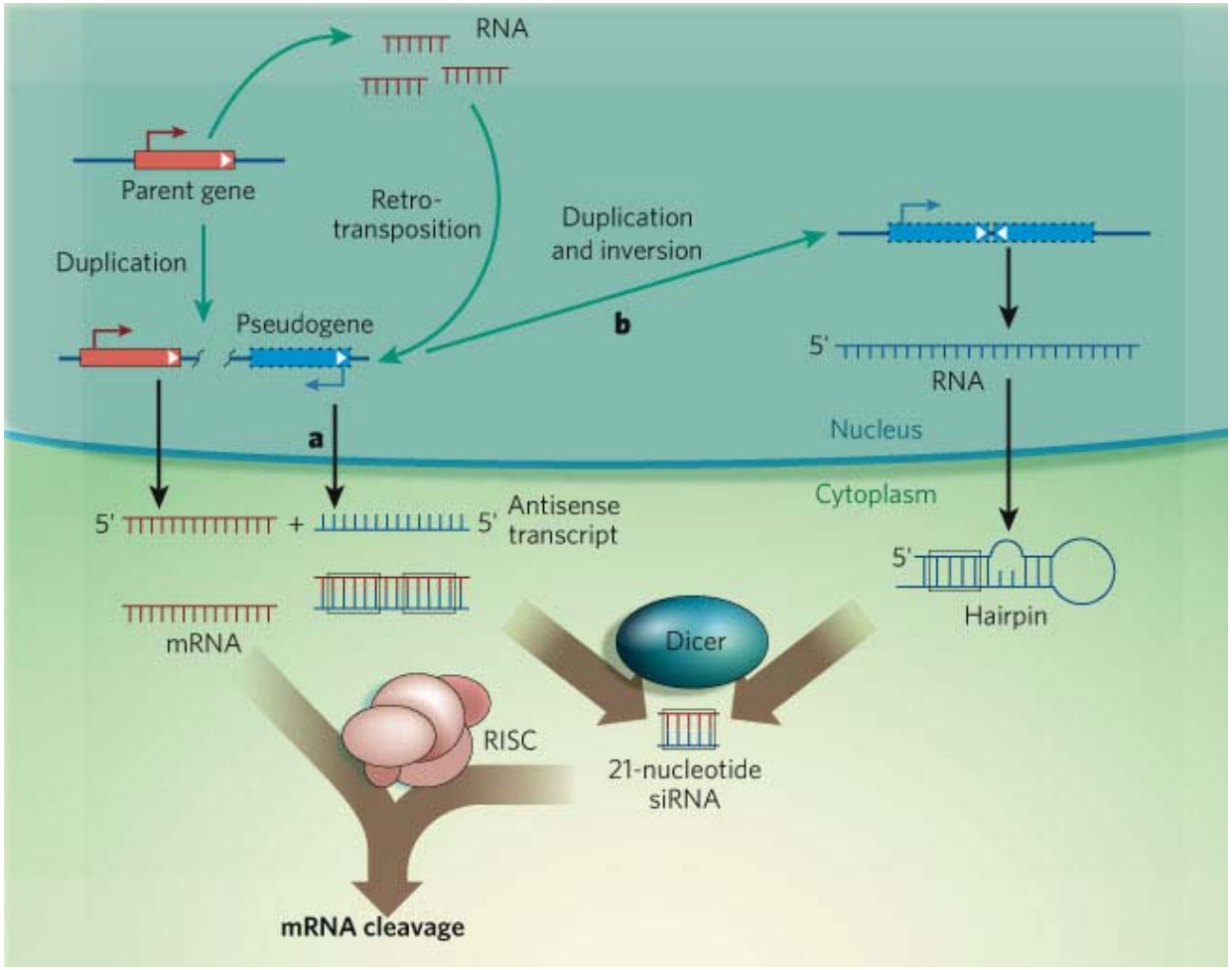
What are Active Pseudogenes Doing? Potential for Gene Regulation via endo-siRNA

- Recent Discovery in Mouse and Fly

- Czech, B. *et al. Nature* 453, 798–802 (2008).
- Ghildiyal, M. *et al. Science* 320, 1077–1081 (2008).
- Kawamura, Y. *et al. Nature* 453, 793–797 (2008).
- Okamura, K. *et al. Nature* 453, 803–806 (2008).
- Tam, O. H. *et al. Nature* 453, 534–538 (2008).
- Watanabe, T. *et al. Nature* 453, 539–543 (2008).

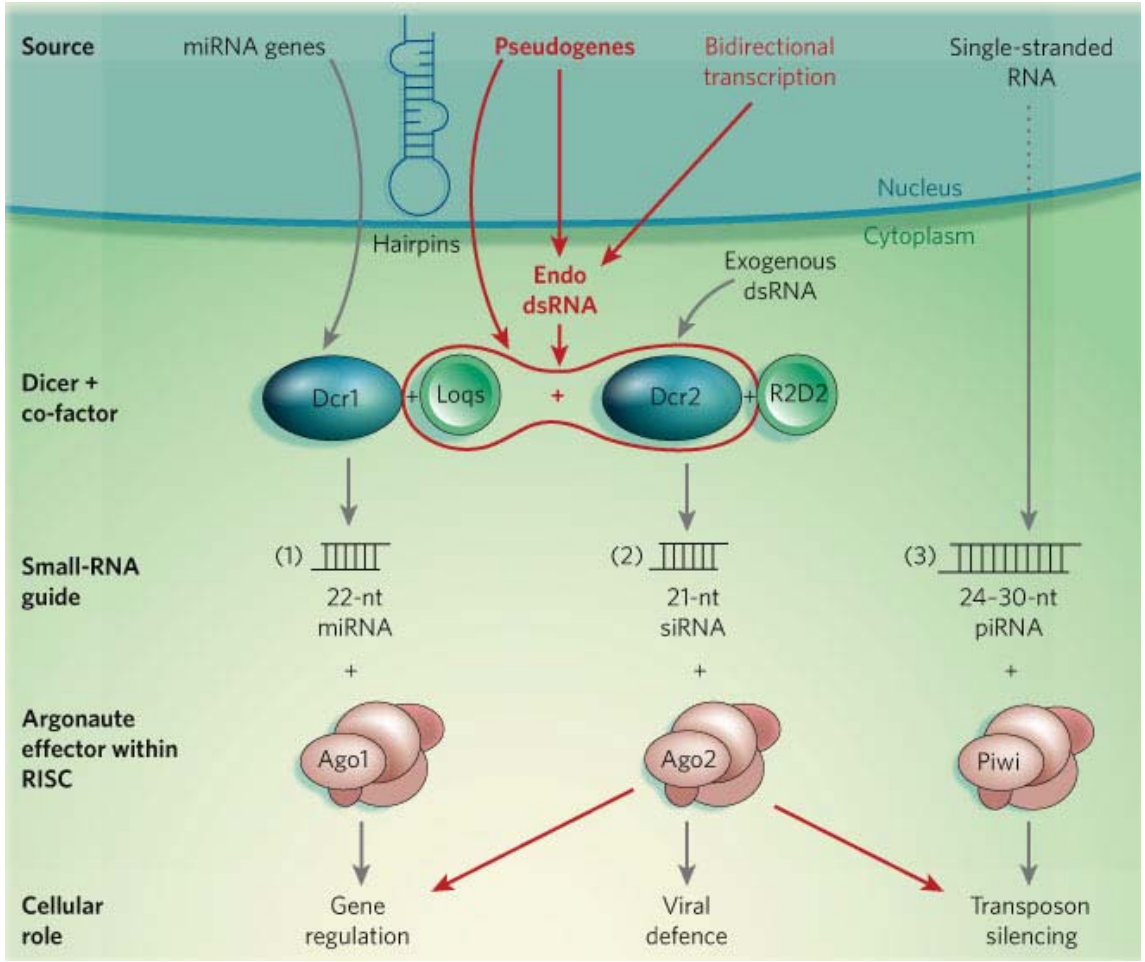
182 Lectures.GersteinLab.org (c)

How could a pseudogene be involved in RNAi?



183 Lectures.GersteinLab.org (c)

Very Speculatively, Papers Blur Boundaries betw. siRNAs and miRNAs

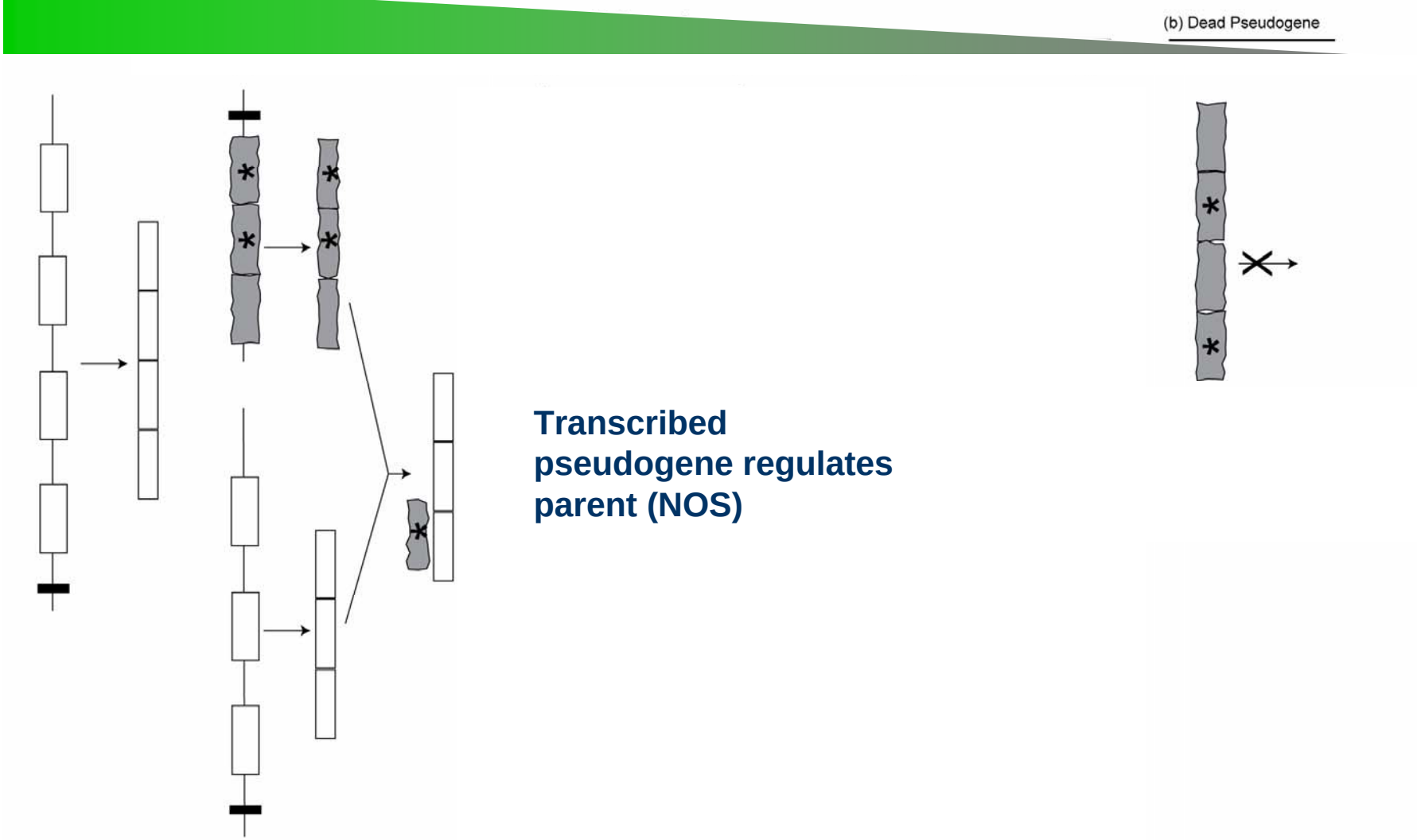


184 Lectures.GersteinLab.org (c)

Genes & Pseudogenes

Zheng & Gerstein, TIG (2007)

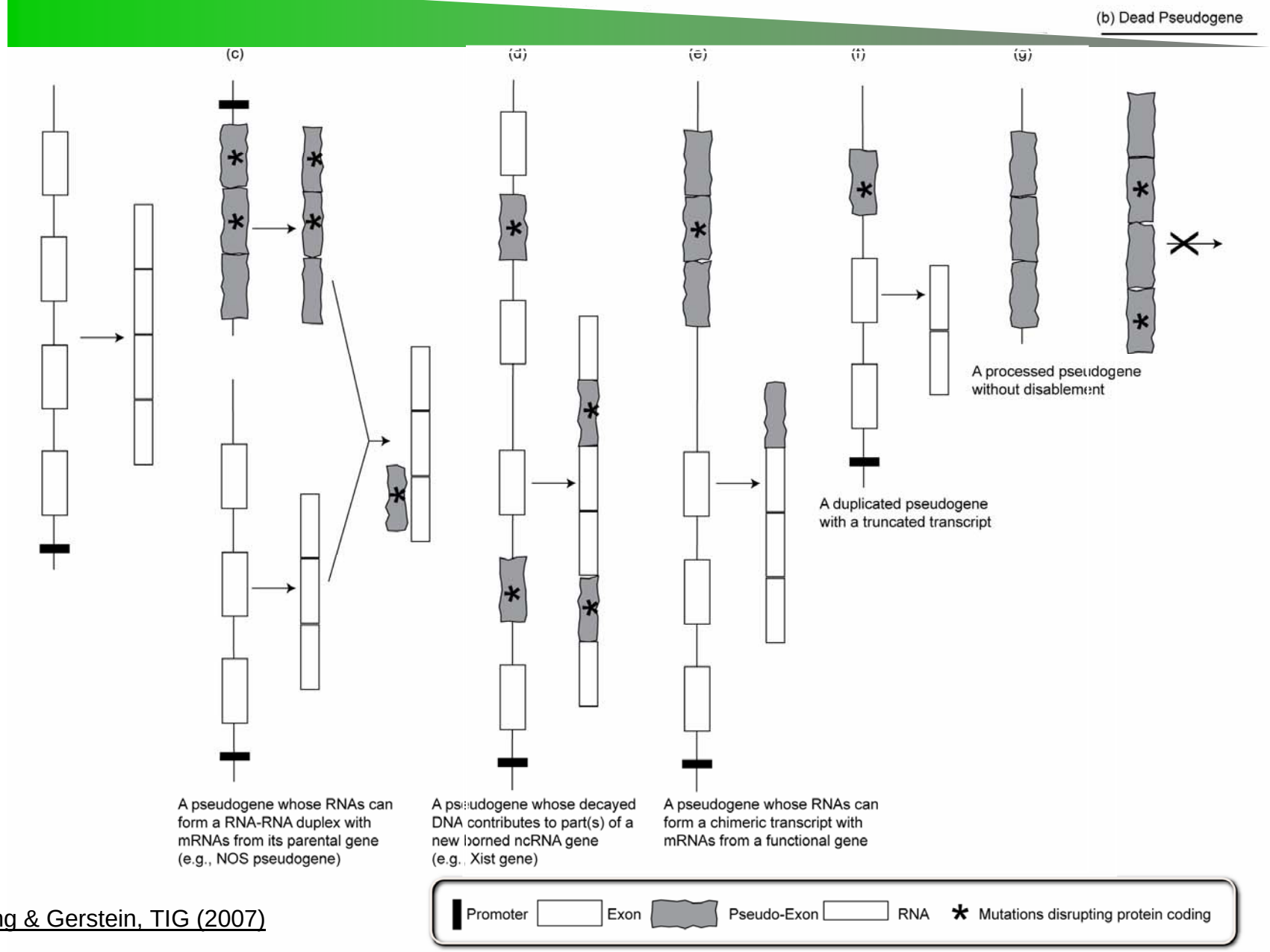
Genes or Pseudogenes?



Transcribed pseudogene regulates parent (NOS)

Zheng & Gerstein, TIG (2007)

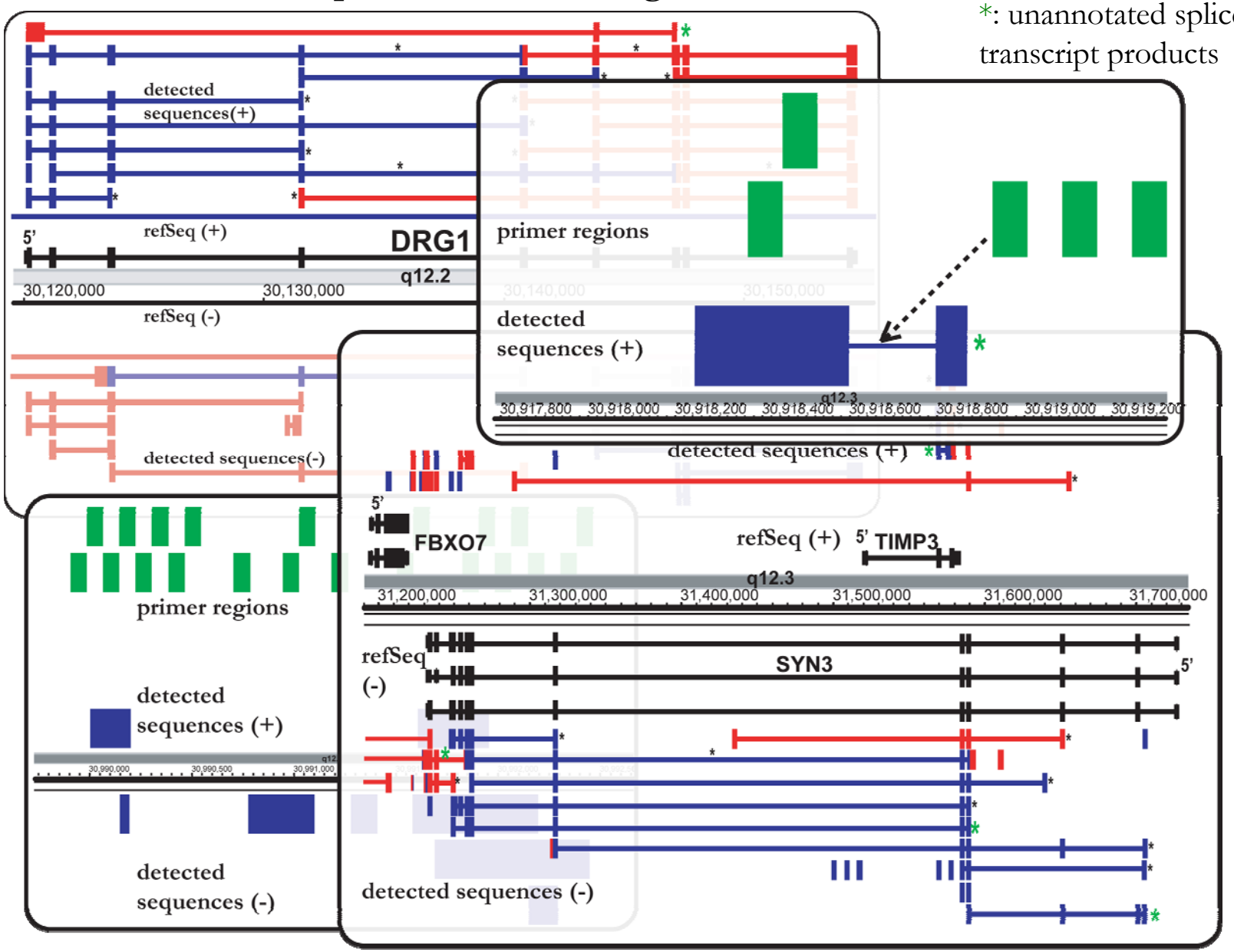
Genes or Pseudogenes?



Zheng & Gerstein, TIG (2007)

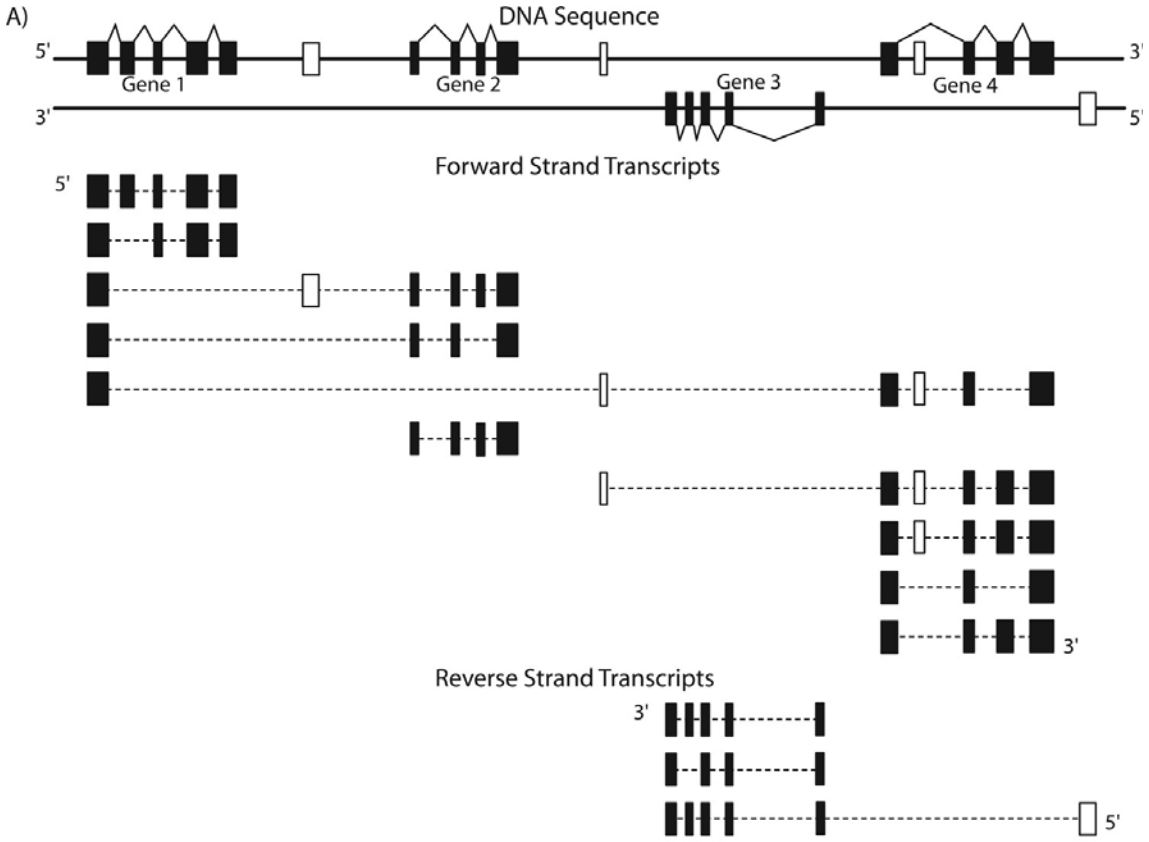
Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome

*: unannotated spliced transcript products

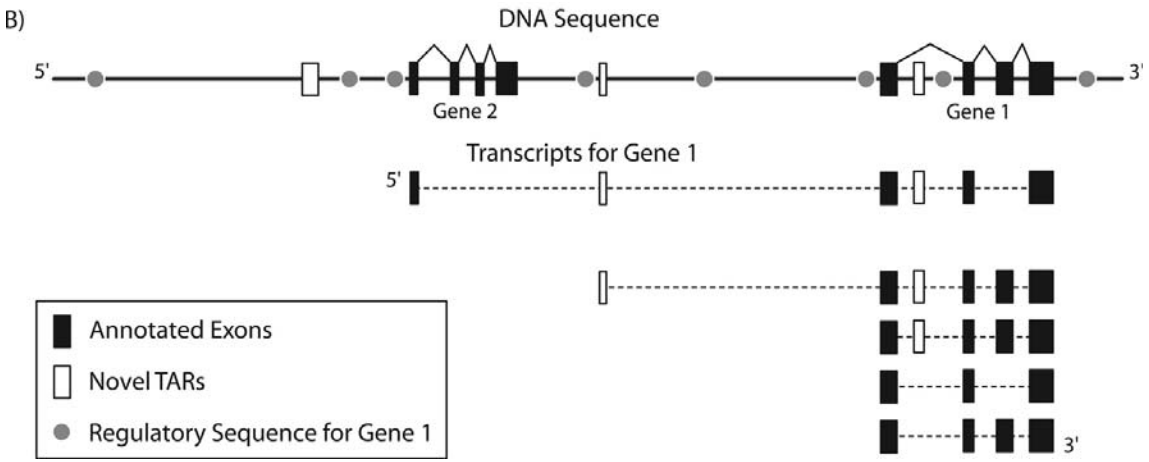


Source: Wu , Du, et al. (2007) Genome Biology

188 Lectures.GersteinLab.org (c)



Biological complexity revealed by ENCODE: Long Interleaved Transcripts and Distributed Regulation

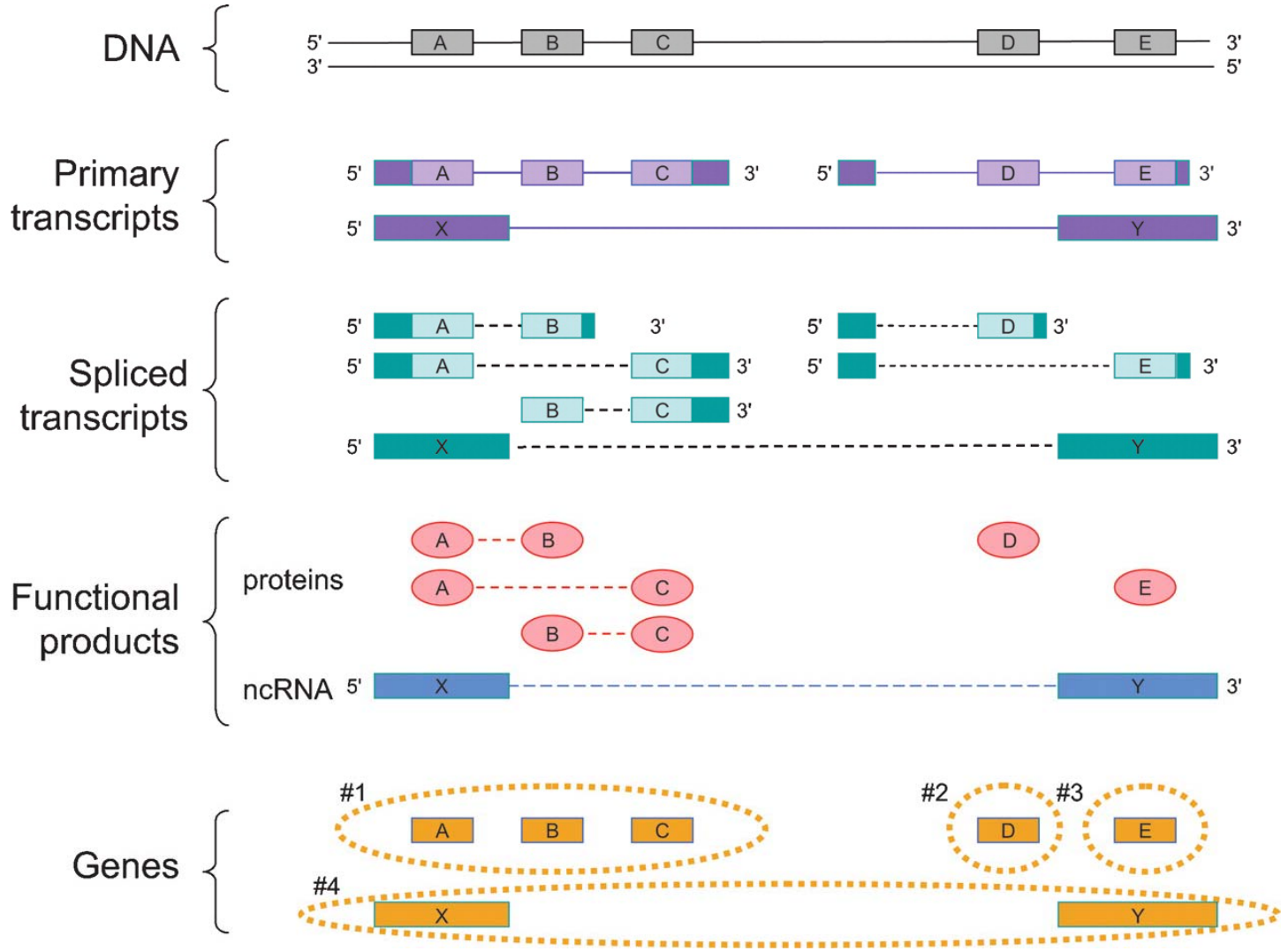


What is a Gene? and What is not a Gene?

[Gerstein et al. Genome Res. 2007; 17: 669-681]

189 Lectures.GersteinLab.org (c)

Proposed Re-definition of a Gene: "Gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products."

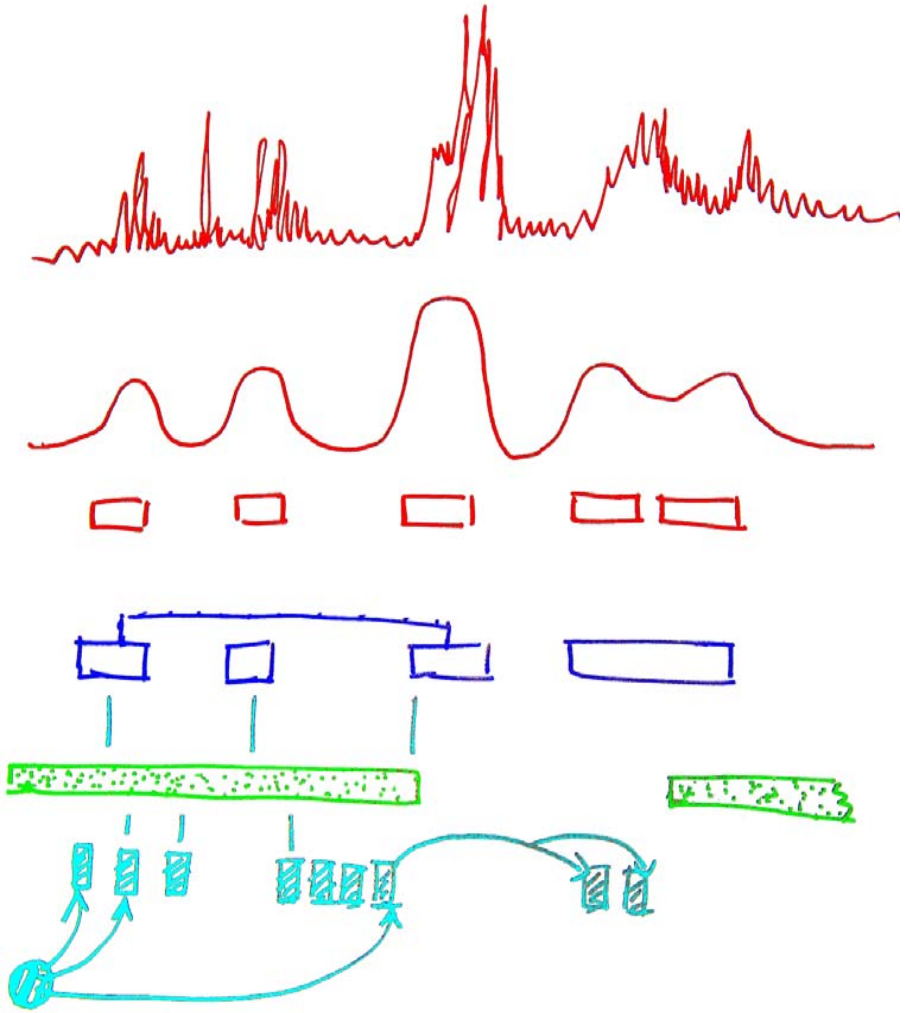


Gerstein et al. Genome Res. 2007; 17: 669-681

190 Lectures.GersteinLab.org (c)


Summary

Overview of Annotation Process



- Doing large-scale similarity comparison, looking for repeated or deleted regions
- **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**

193 Lectures.GersteinLab.org (c)

Processing the Raw Experimental Signal, developing scoring technology

- Simulating to correct for non-uniform coverage of the genome in Chip-seq experiments and using this to better score the experiments

195 Lectures.GersteinLab.org (c)

Large-scale analysis of a single type of "signal" : First-Pass Annotation Clustering and Characterizing Novel Transcribed Regions and Groups of Binding Sites

- DART classification of TARs
  - ◊ 1300 TARs in ~200 novel ENCODE loci
    - based on expression and phylogenetic clustering
- Deserts and Forests of Binding Activity
  - ◊ on ~50kb scale
  - ◊ Biplot gives broad separation of seq. specific and non-specific factors and associated genomic bins

196 Lectures.GersteinLab.org (c)

Integrative Annotation: Relating Pseudogenes to Conservation & Transcription

- Annotation: Pseudogene Assignment
  - ◊ Consensus annotation from automatic pipelines & manual curation gives 201 in ENCODE
    - ~2/3 processed are primate specific
  - ◊ Evidence for selection operating on a few but most neutral

- Pseudogene Activity
  - ◊ >20% appear to be transcribed (38/201)
  - ◊ No obvious selection on transcribed ones

197 Lectures.GersteinLab.org (c)

ENCODE Acknowledgements

Adam Frankish, Robert Baertsch,
Philipp Kapranov, Alexandre Reymond,
Siew Woh Choo, Y Fu, Yontao Lu, France Denoeud,
Stylianos Antonarakis, Yijun Ruan, Chia-Lin Wei, Z Weng, Thomas Gingeras, Roderic Guigo, Tim Hubbard, Jennifer Harrow

Sanger, UCSC, GIS, AFFX, Geneva, IMIM, BU + SU

+

N Trinklein, U Karaöz, A Halees, SF Aldred, PJ Collins, RM Myers

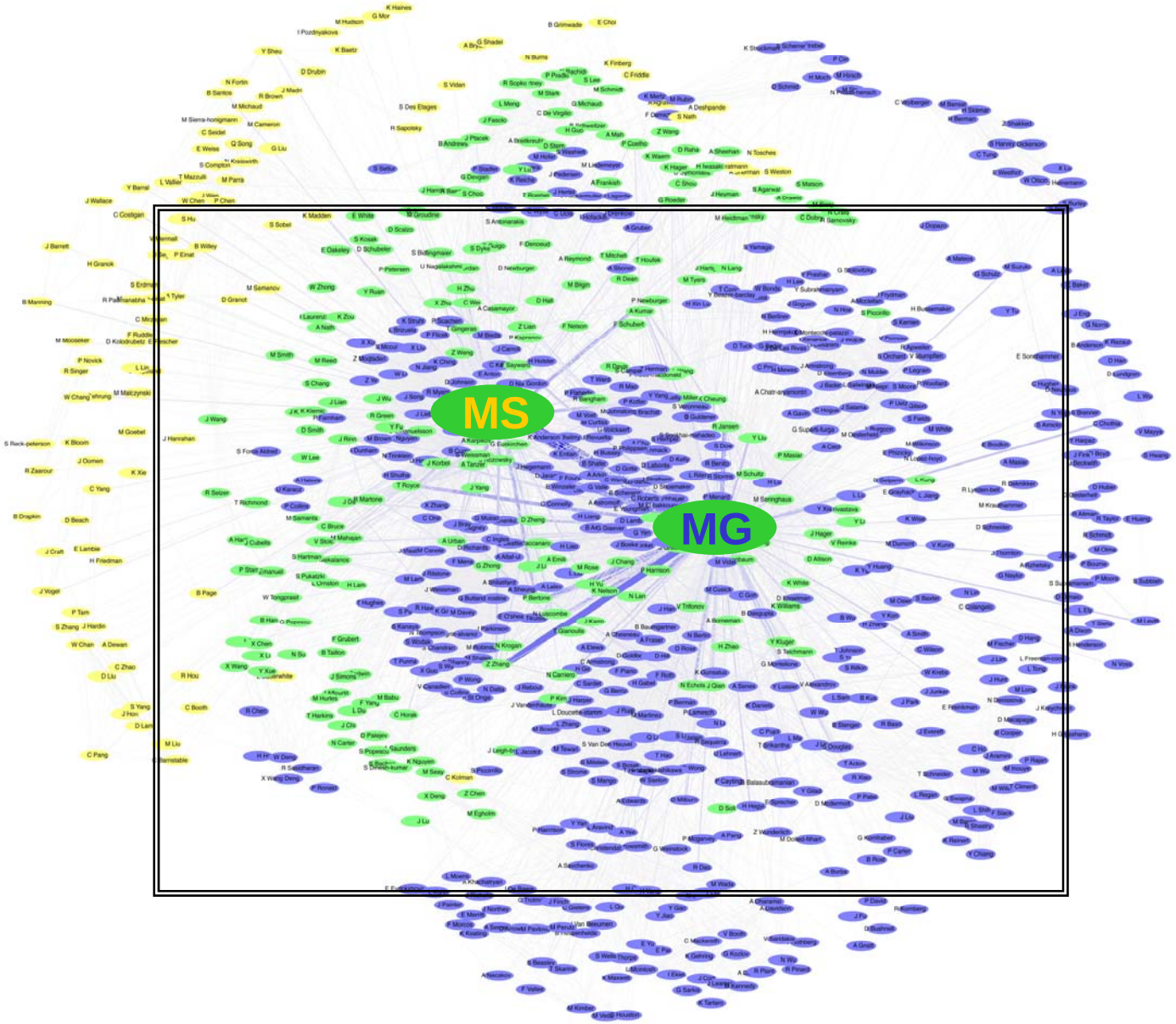
+

Consortium

198 Lectures.GersteinLab.org (c)

Acknowledgements

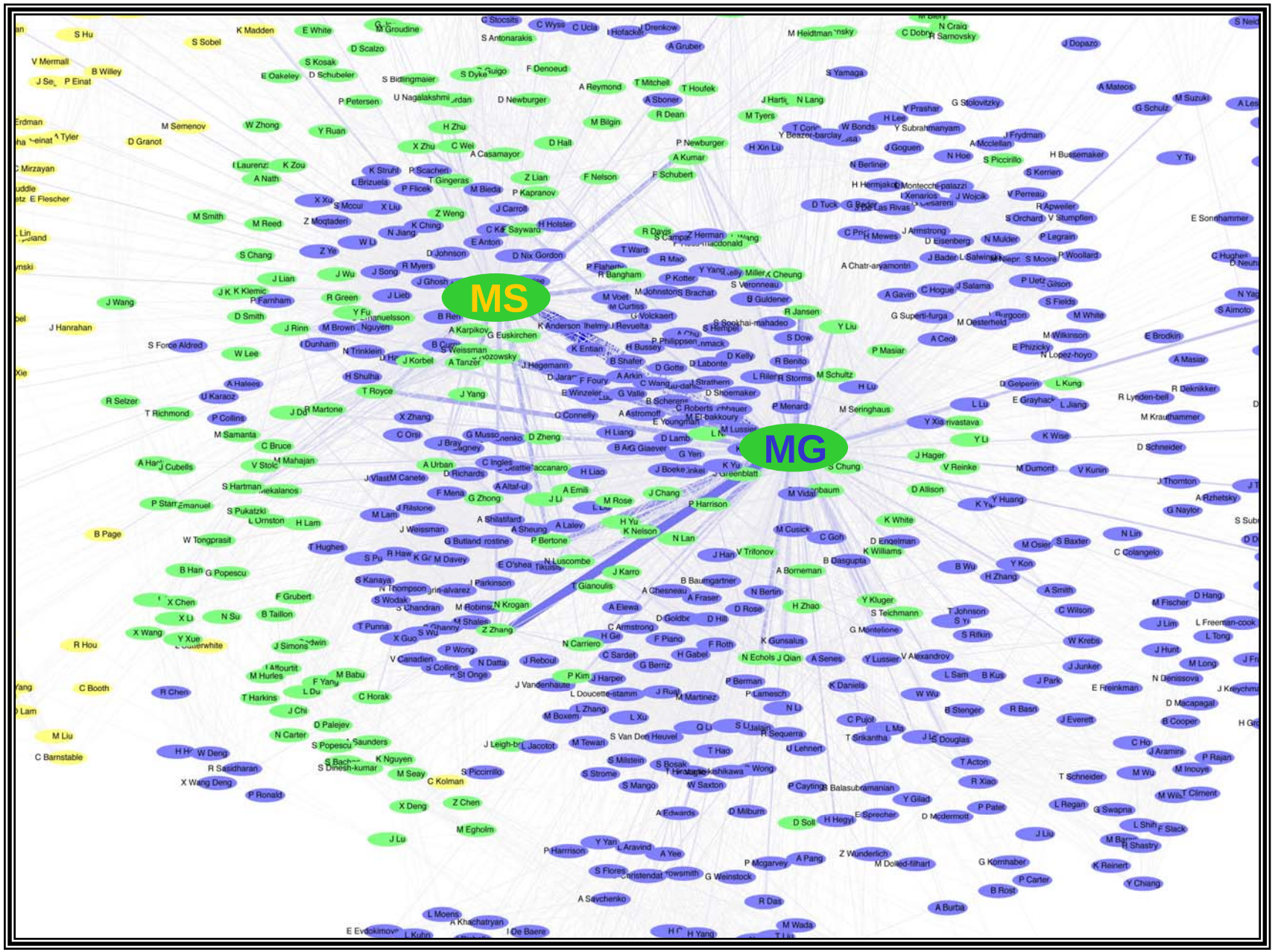
pseudogene.org, tiling.gersteinlab.org, sv.gersteinlab.org



199 Lectures.GersteinLab.org (c)

Acknowledgements

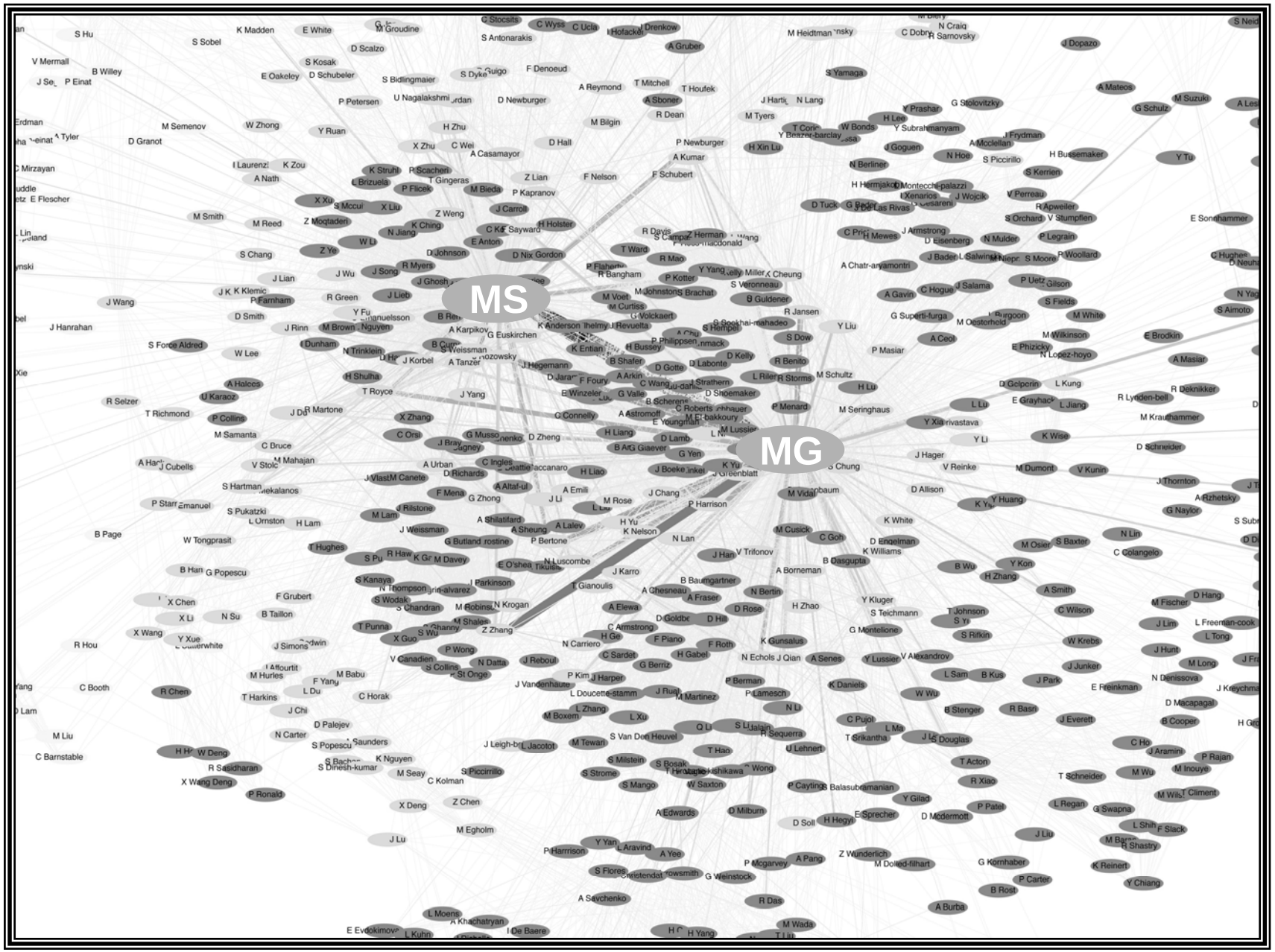
pseudogene.org, tiling.gersteinlab.org, sv.gersteinlab.org



2011 Lectures Gersteinlab.org (c)

Acknowledgements

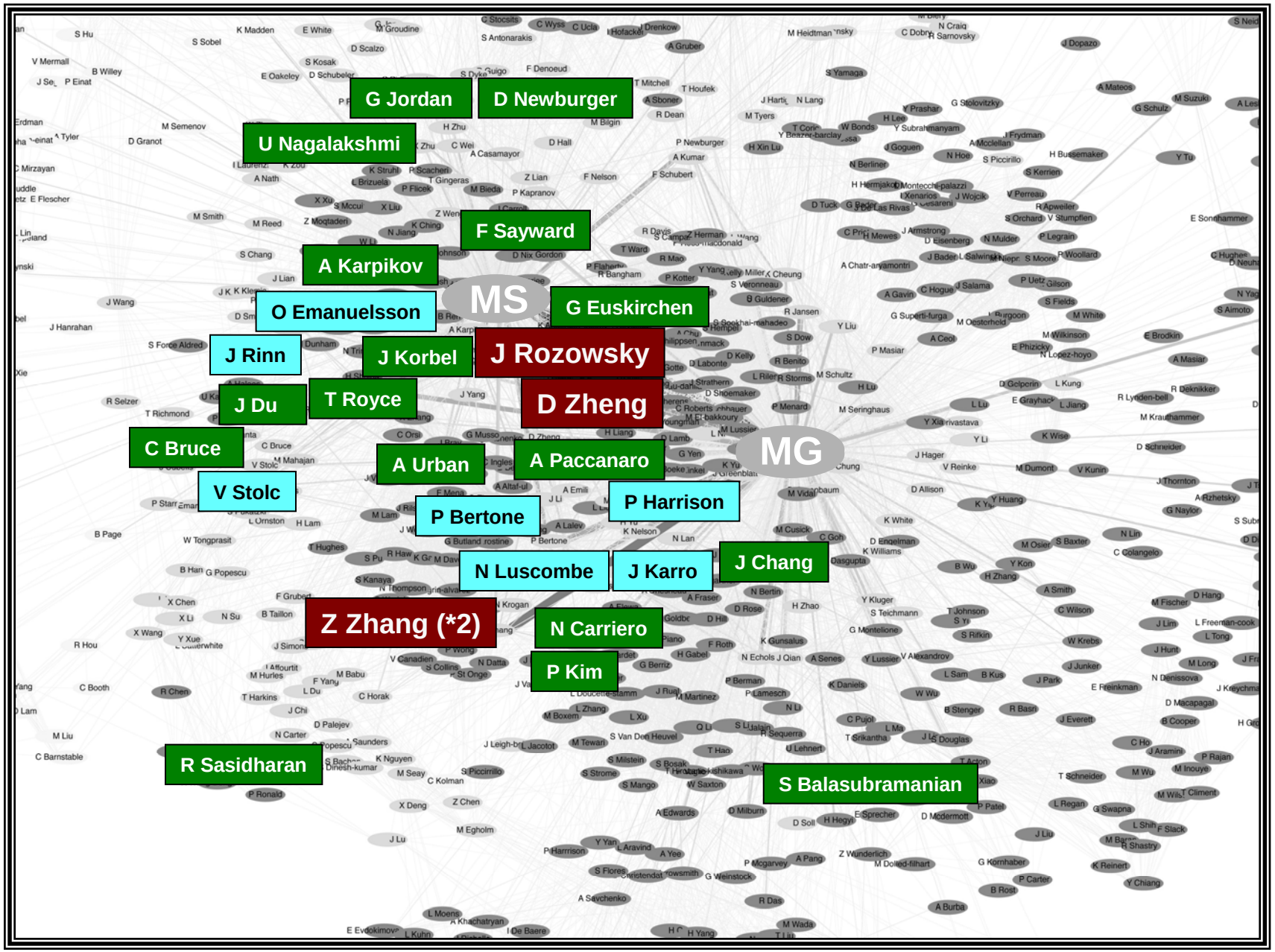
pseudogene.org, tiling.gersteinlab.org, sv.gersteinlab.org



261 Lectures.GersteinLab.org (c)

Acknowledgements

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



2012 Lectures.GersteinLab.org (c)

Permissions Statement

This Presentation is copyright
Mark Gerstein, Yale University, 2007.

Feel free to use images in it with
PROPER acknowledgement

(via citation to relevant papers or link to gersteinlab.org).