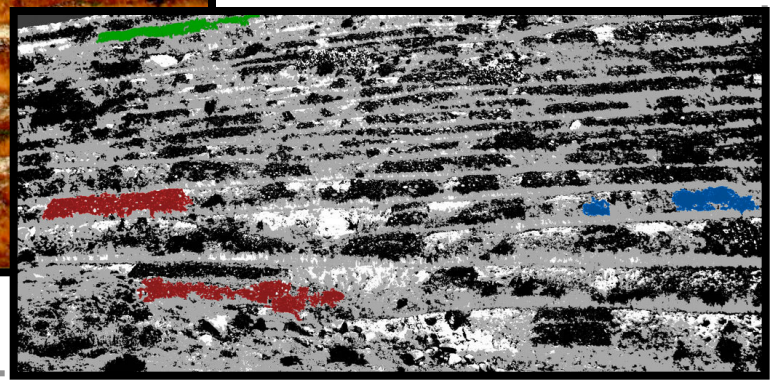
Human Genome Annotation

Mark B Gerstein
Yale

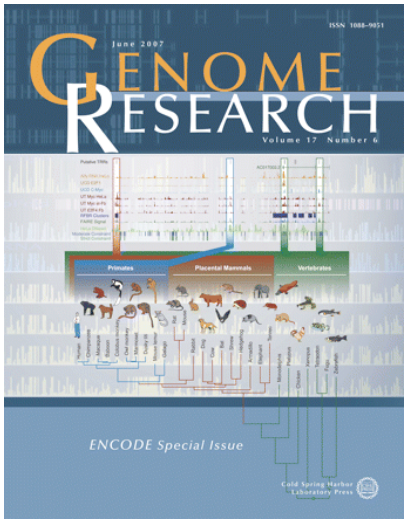
Slides at
Lectures.GersteinLab.org

(See Last Slide for References & More Info.)



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]

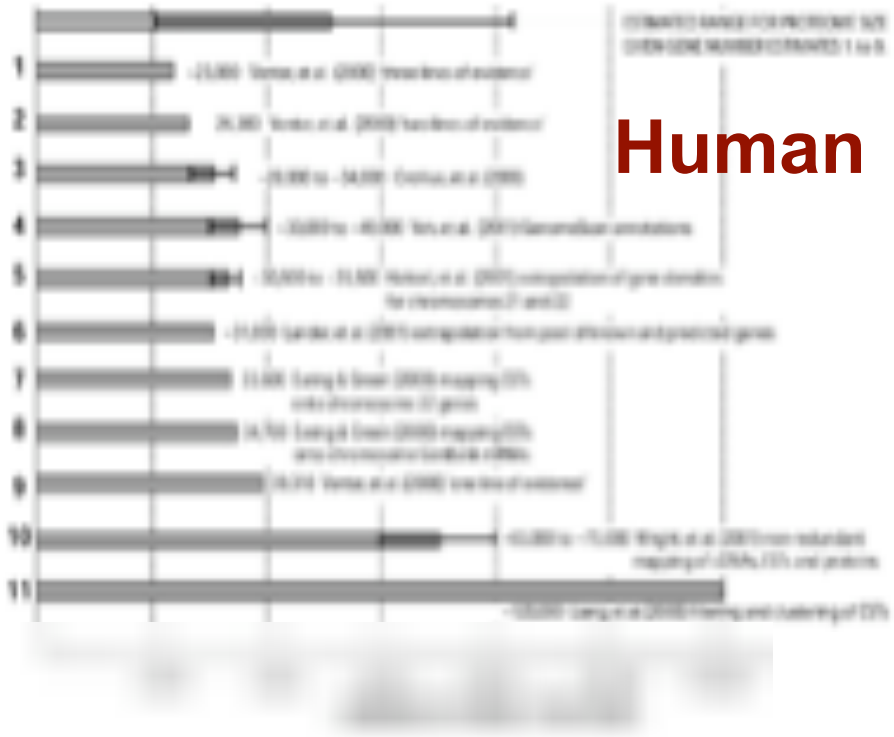


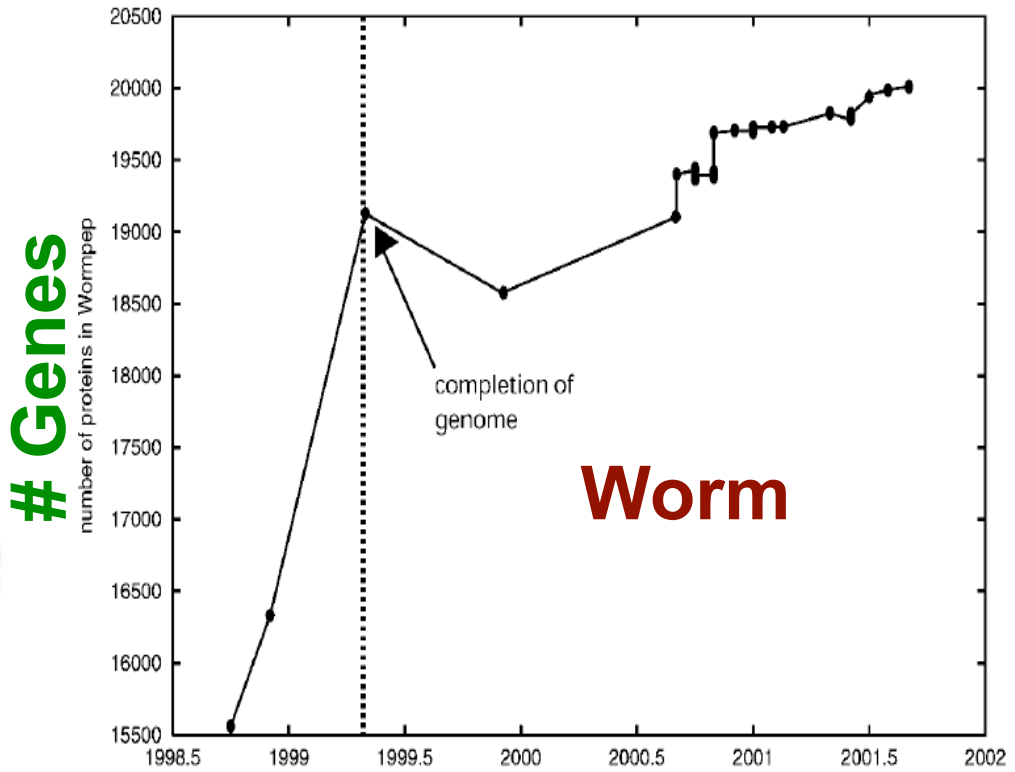
2007 : Pilot results from ENCODE Consortium on decoding what the bases do

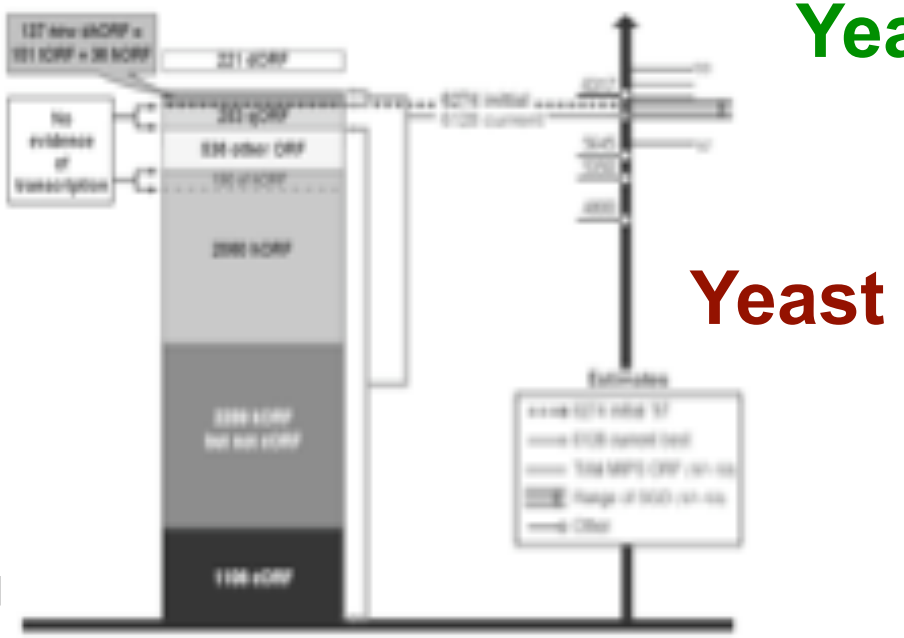
- 1% of Genome (30 Mb in 44 regions)
- Tiling Arrays to assay Transcription & Binding
- Multi-organism sequencing and alignment
- Careful Annotation
- Variation Data

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

Variation in Number of Genes Right Before or After Genome was Sequenced







[Harrison et al. ('02) NAR; Snyder & Gerstein ('03) Science]

So much “Junk” DNA

SO MUCH "JUNK" DNA IN OUR GENOME

Susumu Ohno
City of Hope National Medical Center

The mammalian genome (haploid chromosome complement) contains roughly 3.0×10^{-9} mg of DNA which represents about 3.0×10^{9} base pairs. This is at least 750 times the genome size of *E. coli*. If we take the simplistic assumption that the number of genes contained is proportional to the genome size, we would have to conclude that 3 million or so genes are contained in our genome. The falseness of such an assumption becomes clear when we realize that the genome of lowly lungfish and salamanders can be 36 times greater than our own (Ohno and Atkin, 1966).

In fact, there seems to be a strict upper limit for the number of gene loci which we can afford to keep in our genome. Consequently, only a fraction of our DNA appears to function as genes. The observations on a number of structural gene loci of man, mice and other organisms revealed that each locus has a 10^{-5} per generation probability of sustaining a deleterious mutation. It then follows that the moment we acquire 10^{5} gene loci, the overall deleterious mutation rate per generation becomes 1.0 which appears to represent an

unbearably heavy genetic load. Taking into consideration the fact that deleterious mutations can be dominant or recessive, the total number of gene loci of man has been estimated to be about 3×10^{4} (Muller, 1967; Crow and Kimura, 1970). Even if an allowance is made for the existence in multiplicates of certain genes, it is still concluded that, at the most, only 6% of our DNA base sequences is utilized as genes (Kimura and Ohta, 1971). Aside from conventional structural genes and regulatory genes, this 6% should include the *promotor* region and *operator* region which are situated adjacent to each structural gene, for these regions can certainly sustain deleterious mutations. More than 90% degeneracy contained within our genome should be kept in mind when we consider evolutional changes in genome sizes. What is the reason behind this degeneracy?

Certain untranscribable and/or untranslatable DNA base sequences appear to be useful in a negative way (the importance of doing nothing). If functional genes customarily occupied the region around the centromere, evolutional changes of chromosome complements would not have occurred as often as has been observed. Because the centromeric hetero-chromatin which represents a long tandem repeat of a short untranscribable sequence (Jones, 1970; Southern, 1970) can be lost or duplicated without deleterious consequences, speciation more often than not has been accompanied by chromosomal changes. The same can be said of those DNA base sequences which are used as partitions between the genes. It may be of selective advantage to space adjacent genes far enough apart by inserting a stretch of untranscribable and/or

Ohno S. So much "junk" DNA in our genome. Brookhaven Symp Biol. 1972;23:366-70.

Different Views of the Function of Junk DNA

[NY Times, 26-Jun-07]

ESSAY

Human DNA, the Ultimate Spot for Secret Messages (Are Some There Now?)

By DENNIS OVERBYE

In Douglas Adams's science fiction classic, "The Hitchhiker's Guide to the Galaxy," there is a character by the name of Slartibartfast, who designed the fjords of Norway and left his signature in a glacier.

I was reminded of Slartibartfast recently as I was trying to grasp the implications of the feat of a team of Japanese geneticists who announced that they had taught relativity to a bacterium, sort of.

Using the same code that computer keyboards use, the Japanese group, led by Masaru Tomita of Keio University, wrote four copies of Albert Einstein's famous formula, $E=mc^2$, along with "1905," the date that the young Einstein derived it, into the bacterium's genome, the 400-million-long string of A's, G's, T's and C's that determine everything the little bug is and everything it's ever going to be.

The point was not to celebrate Einstein. The feat, they said in a paper published in the journal Biotechnology Progress, was a demonstration of DNA as the ultimate information storage material, able to withstand floods, terrorism, time and the changing fashions in technology, not to mention the ability to be imprinted with little unobtrusive trademark labels — little "Made by Monsanto" tags, say.

In so doing they have accomplished at least a part of the dream that Jaron Lanier, a computer scientist and musician, and David Sulzer, a biologist at Columbia, enunciated in 1999. To create the ultimate time capsule as part of the millennium festivities at this newspaper, they proposed to encode a year's worth of the New York Times magazine into the junk DNA of a cockroach. "The archival cockroach will be a robust repository," Mr. Lanier wrote, "able to survive almost all conceivable scenarios."

If cockroaches can be archives, why not us? The human genome, for example, consists of some 2.9 billion of those letters — the equivalent of about 750 megabytes of data — but only about 3 percent of it goes into composing the 22,000 or so genes that make *us* what we are.

The remaining 97 percent, so-called junk DNA, looks like gibberish. It's the dark matter of inner space. We don't know what it is saying to or about us, but within that sea of megabytes there is plenty of room for the imagination to roam, for trademark labels and much more. The King James Bible, to pick one obvious example, only amounts to about five megabytes.

Jimmy Turrell

If a bacterium can be encoded with $E=mc^2$, if cockroaches can be archives, why not us?

Inevitably, if you are me, you begin to wonder if there is already something written in the warm wet archive, whether or not some Slartibartfast has already been here and we ourselves are walking around with little trademark tags or more wriggling and squiggling and folded inside us. Gil Bejerano, a geneticist at the University of California, Santa Cruz, who mentioned Slartibartfast to me, pointed out that the problem with raising this question is that people who look will see messages in the genome even if they aren't there — the way people have claimed in recent years to have found secret codes in the Bible.

Nevertheless, no less a personage than Francis Crick, the co-discoverer of the double helix, writing with the chemist Leslie Orgel, now at the Salk Institute in San Diego, suggested in 1973 that the primitive Earth was infected with DNA broadcast through space by an alien species.

As a result, it has been suggested that the search for extraterrestrial intelligence, or SETI, should look inward as well as outward. In an article in New Scientist, Paul Davies, a cosmologist at Arizona State University,

change, and have remained identical in humans, rats, mice, chickens and dogs for at least 300 million years.

But Dr. Bejerano, one of the discoverers of these "ultraconserved" strings of the genome, said that many of them had turned out to be playing important command and control functions.

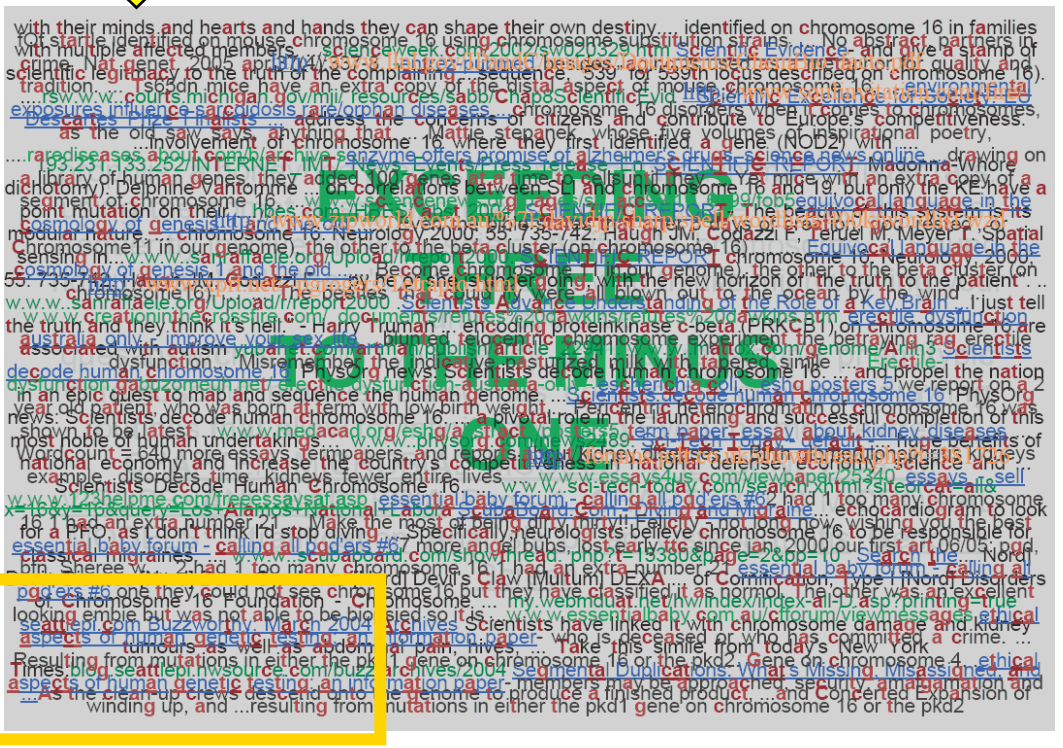
"Why they need to be so conserved remains a mystery," he said, noting that even regular genes that do something undergo more change over time. Most junk bits of DNA that neither help nor annoy an organism mutate even more rapidly.

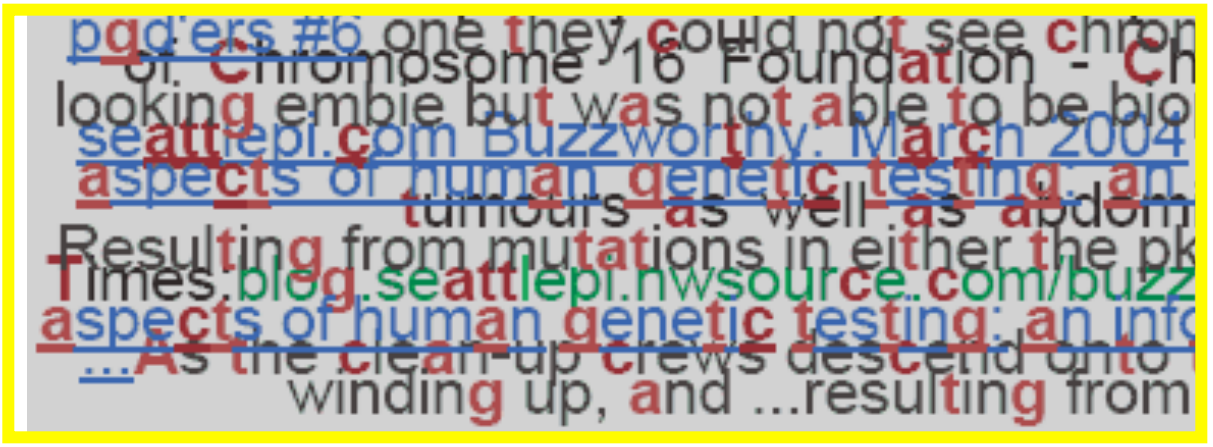
The Japanese team proposed to sidestep the mutation problem by inserting redundant copies of their message into the genome. By comparing the readouts, they said, they would be able to recover Einstein's formula even when up to 15 percent of the original letters in the string had changed, or mutated. "This is the major point

sections of junk DNA seem to be markedly resistant to

Slartibartfast.

Using the same code that computer keyboards use, the Japanese group... wrote four copies of Albert Einstein's famous formula, E=mc2... into the bacterium's genome... In so doing they have accomplished at least a part of the dream that Jaron Lanier, a computer scientist and musician, and David Sulzer, a biologist at Columbia, enunciated in 1999. To create the ultimate time capsule as part of the millennium festivities at this newspaper, they proposed to encode a year's worth of the New York Times magazine into the junk DNA of a cockroach. "The archival cockroach will be a robust repository," Mr. Lanier wrote, "able to survive almost all conceivable scenarios."





Junk DNA as Art

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.

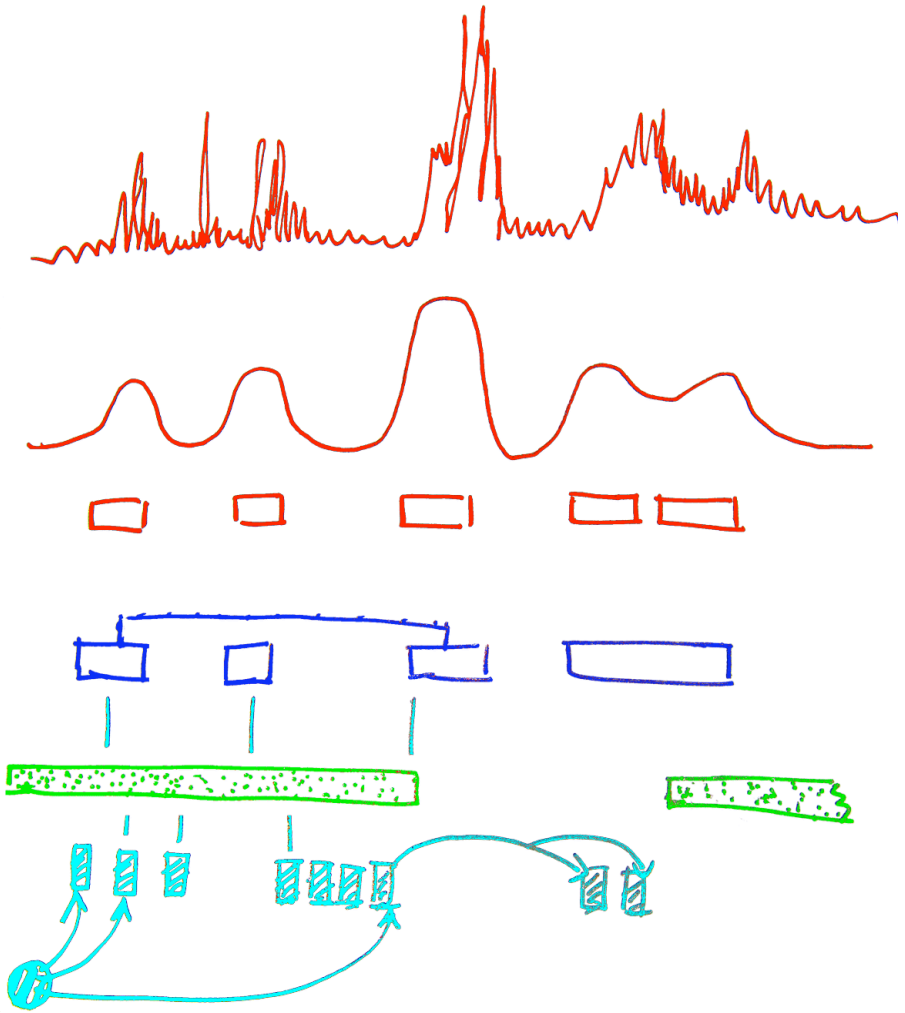
This modern echo of Swift's Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, "On holy wars and a plea for peace." The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

Another feud—largely forgotten, I 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

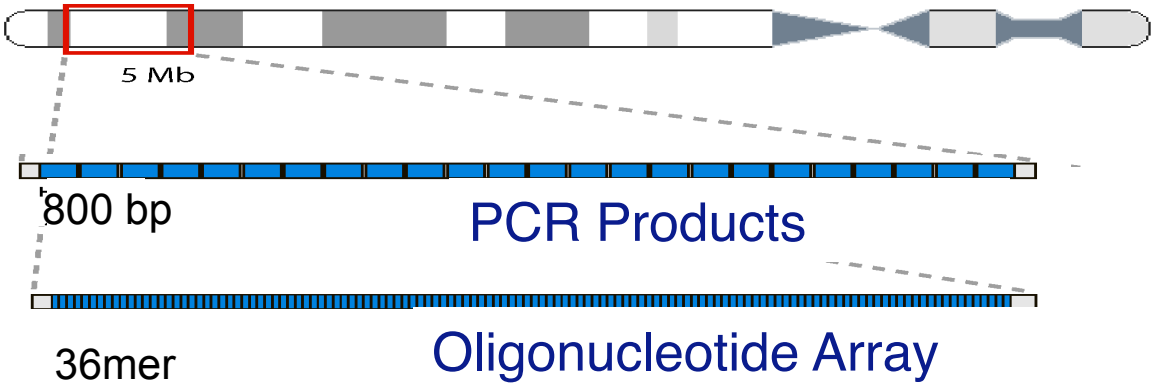


- Determining experimental signals for biochemical activity (e.g. transcription) across each base of genome
  - ◊ Development of Sequence (and Array) Technology
    - Normalizing & Scoring Signal, Correcting Artifacts, Segmenting to create Small Annotation Blocks
  - ◊ Output of Production Pipelines 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

Technologies used for Interrogating the Human Genome, over the past 6 years: Reading out "active" or "tagged" regions

Tiling Arrays

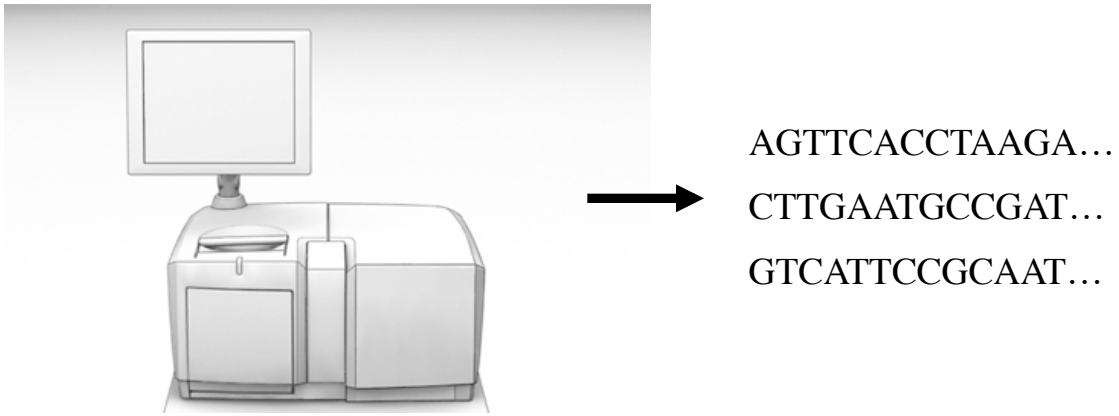
'02



'04

Massively Parallel Sequencing

'06+



Application in a variety of contexts:

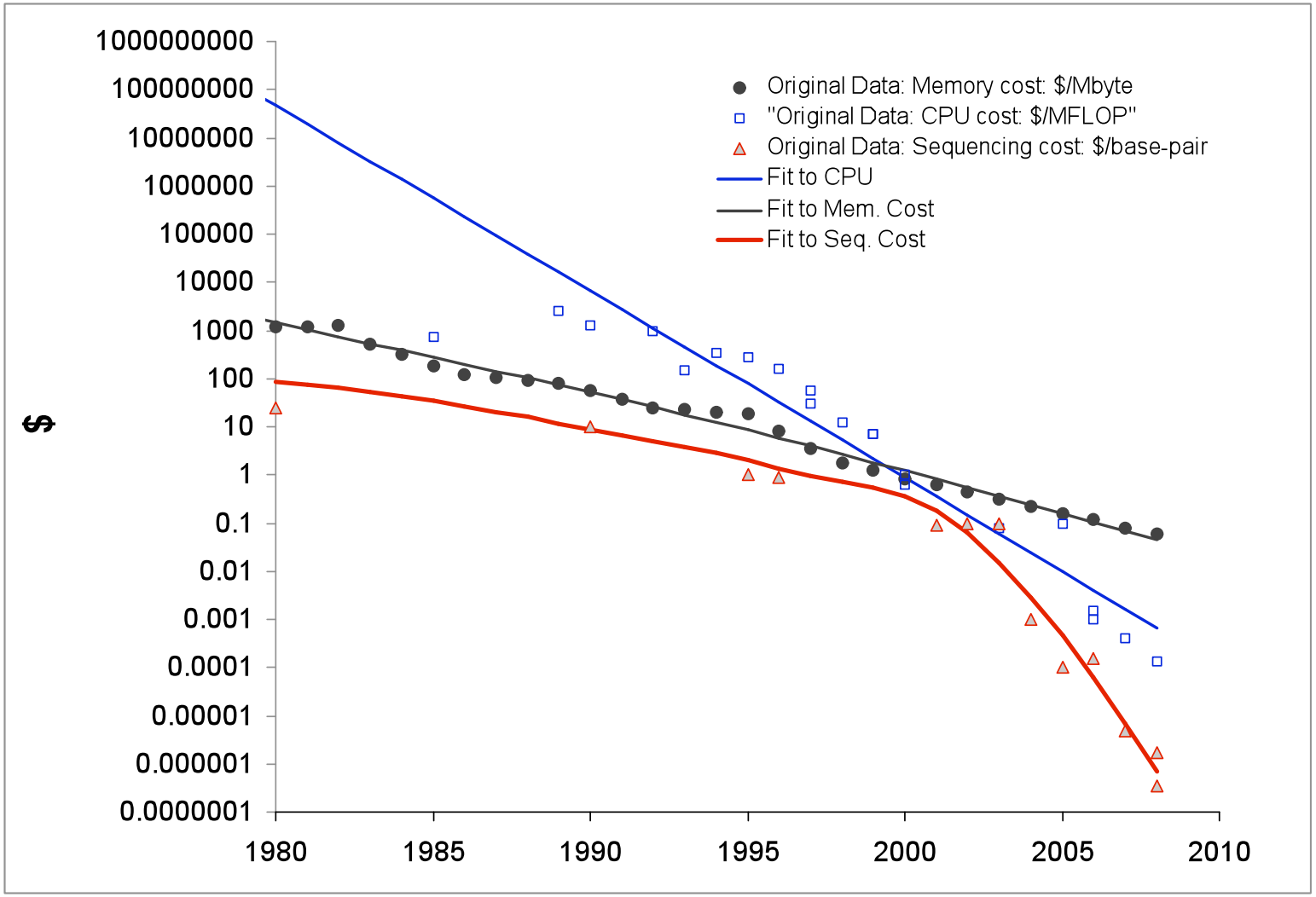
Transcription Mapping

DNA binding (inc. chromatin struc.)

Replication

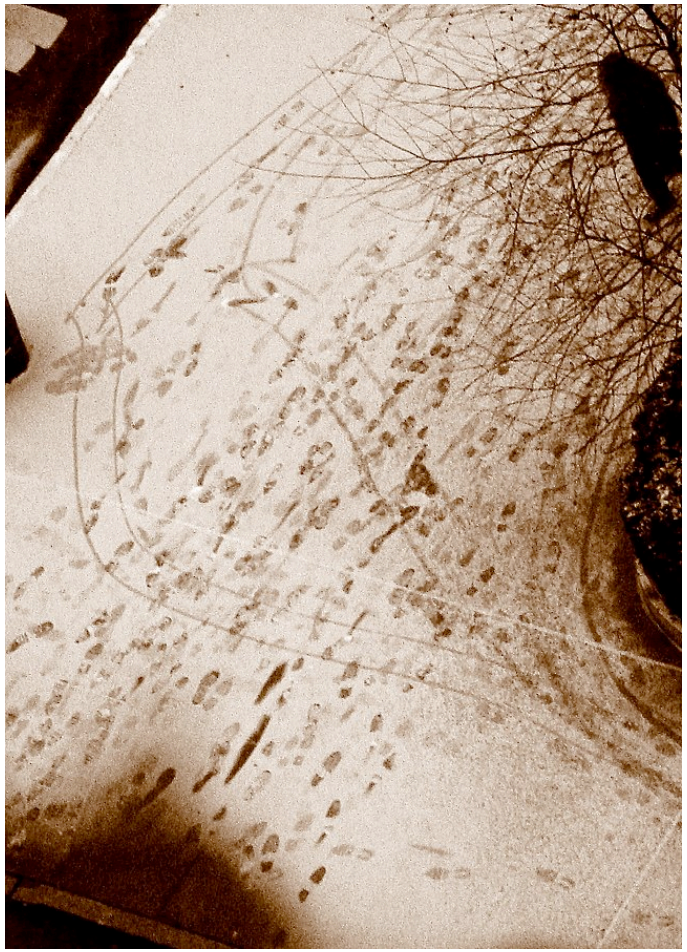
Structural Variation

Plummeting Cost of Sequencing



[Greenbaum et al., Am. J. Bioethics ('08)]

Outline

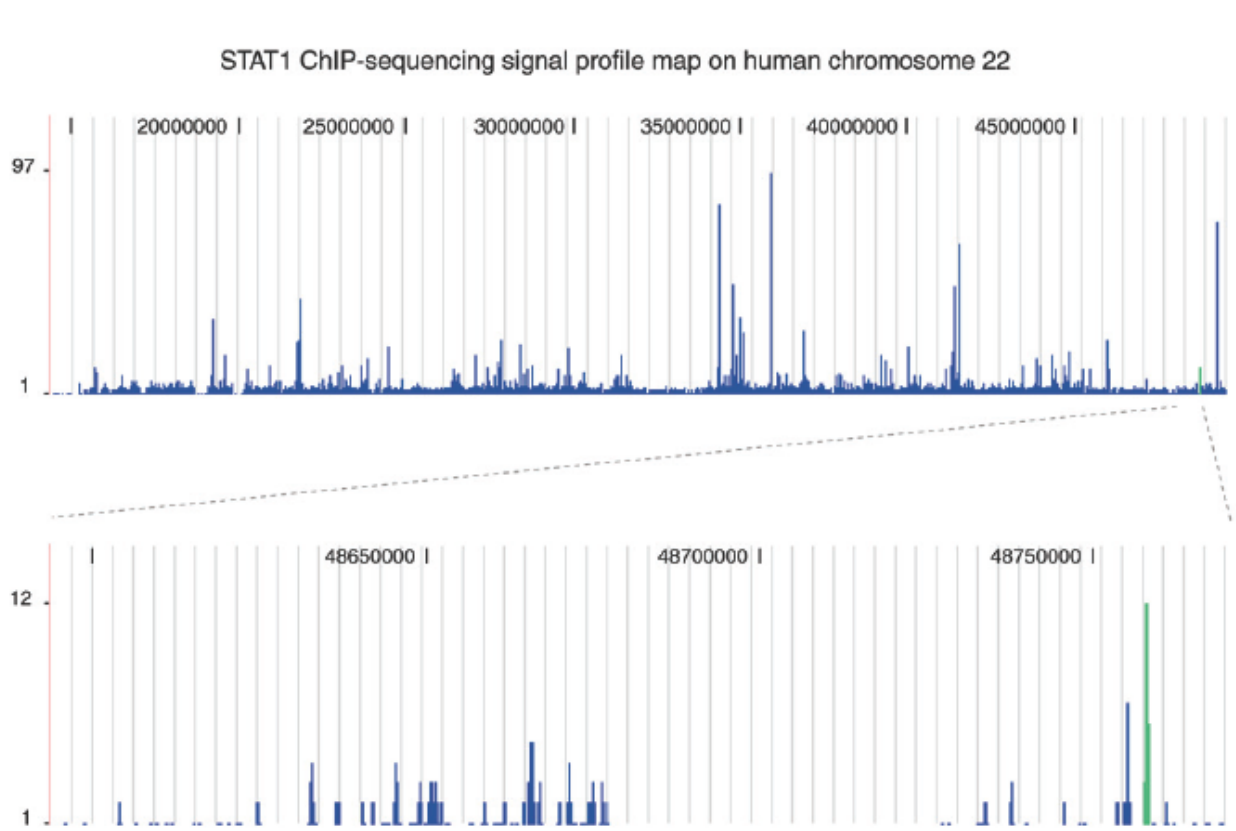


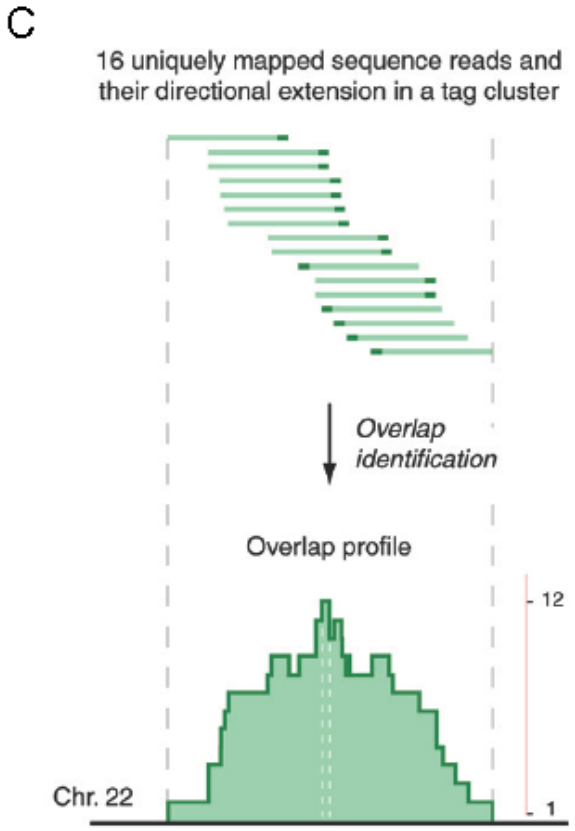
- Signal processing to call "Blocks"
  - ◊ Calling Punctate Blocks (ChipSeq)
  - ◊ Calling Broader Blocks (CNVs)
- Clustering "Blocks" into larger regions
  - ◊ Binding Sites
- Annotating Copied Regions in the Genome
  - ◊ SD and CNVs
  - ◊ Pseudogenes
- Integration of Pseudogenes with Other Annotations
- Future of Annotation
  - ◊ What is a "gene" post encode?

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

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

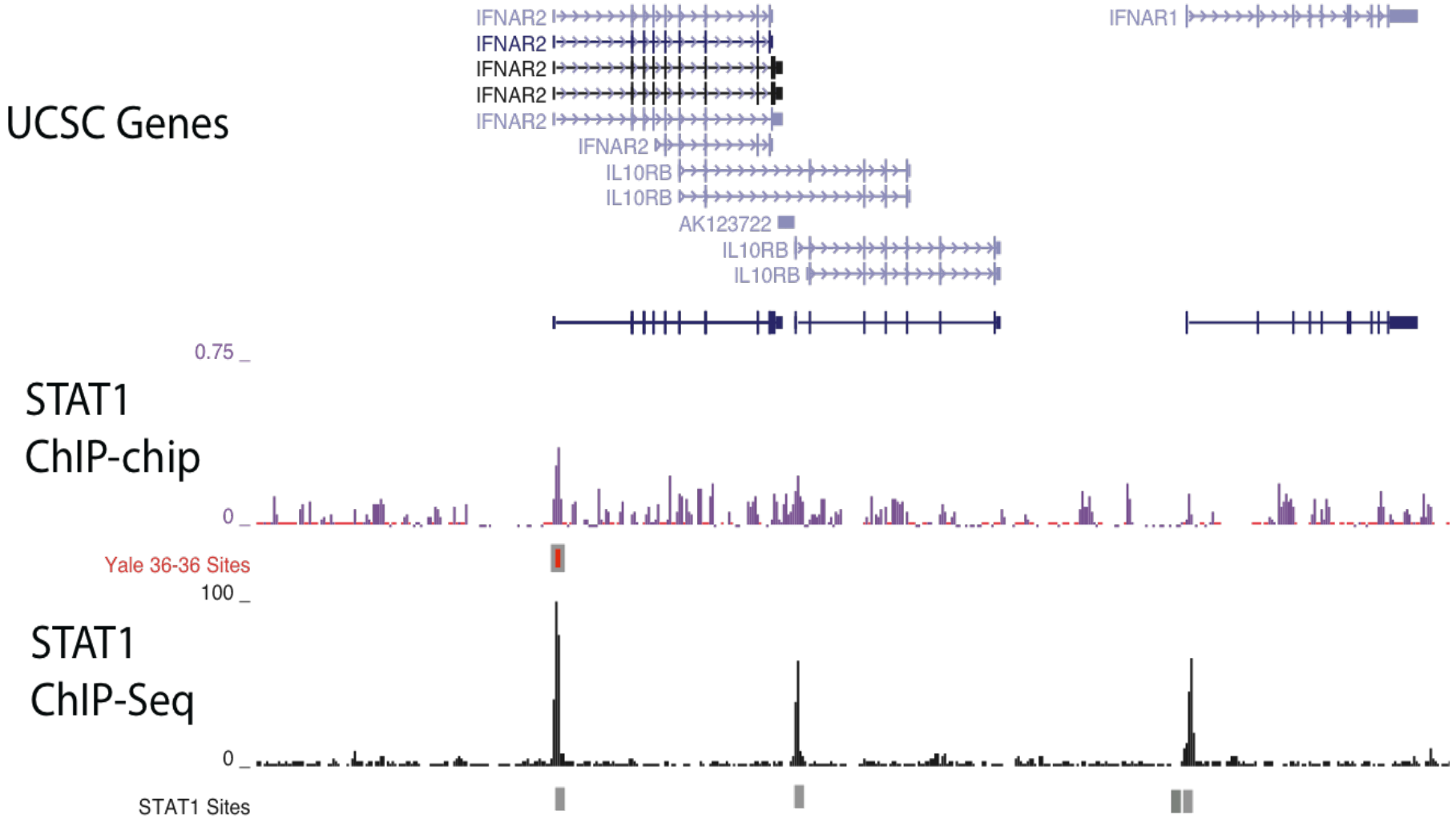
Representative Signal from Chip-Seq





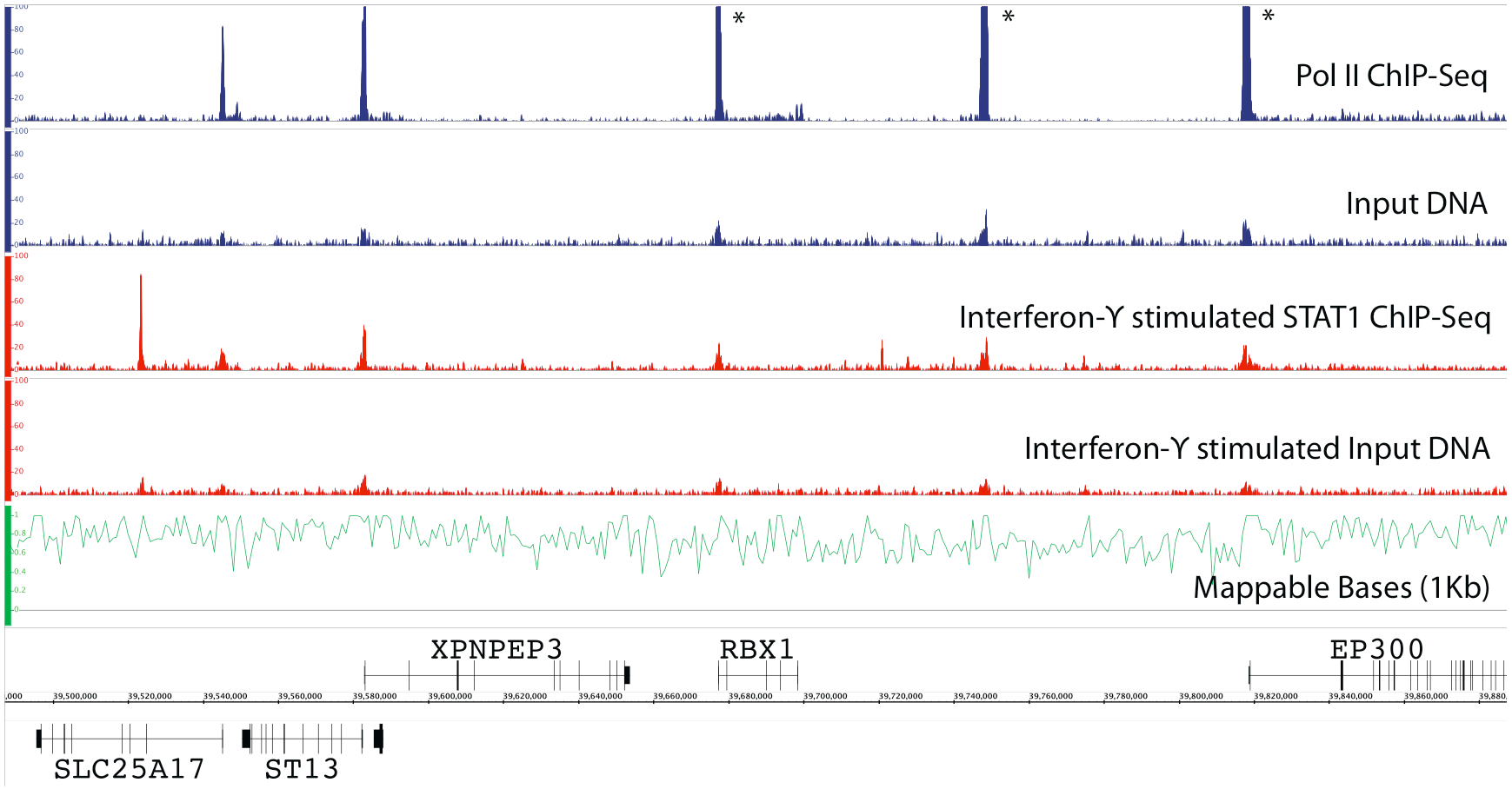
[Robertson et al., Nat. Meth. ('07); Zhang et al. PLOS Comp. Bio. (in revision, '08)]

ChIP-seq vs ChIP-chip: Much cleaner signal from sequencing than arrays



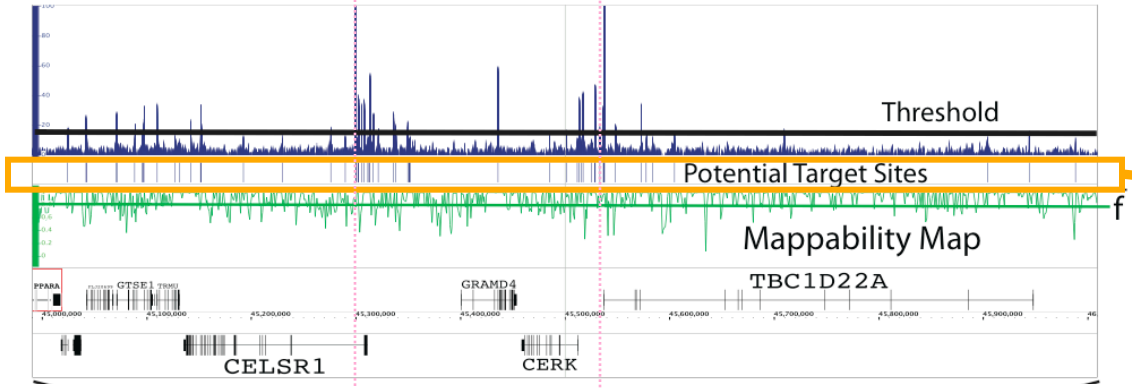
[Rozowsky et al. Nat. Biotech ('09)]

ChIP-Seq vs Input DNA Control



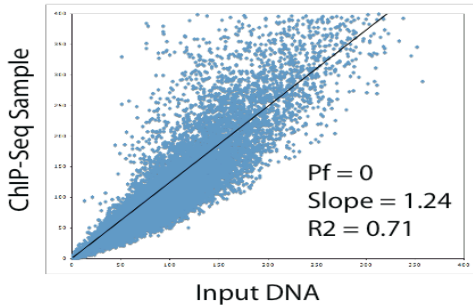
[Rozowsky et al. Nat. Biotech ('09)]

PeakSeq: Scoring Relative to Controls



Filter for Potential Targets based on "Mappability" Simulation

Scale Input Relative to ChIP

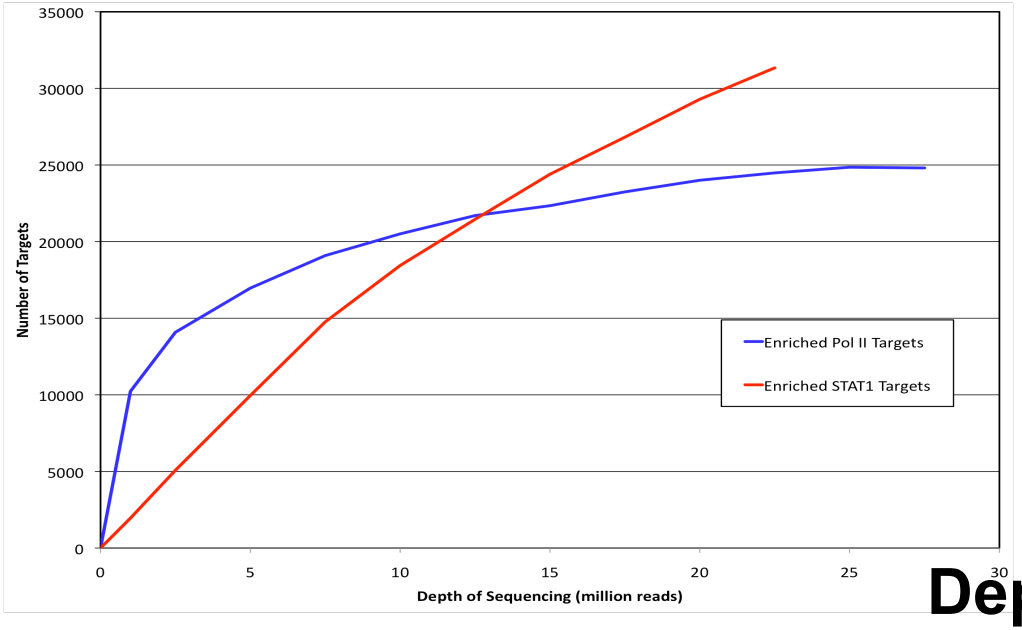


Score Relative to Bionomial Expectation

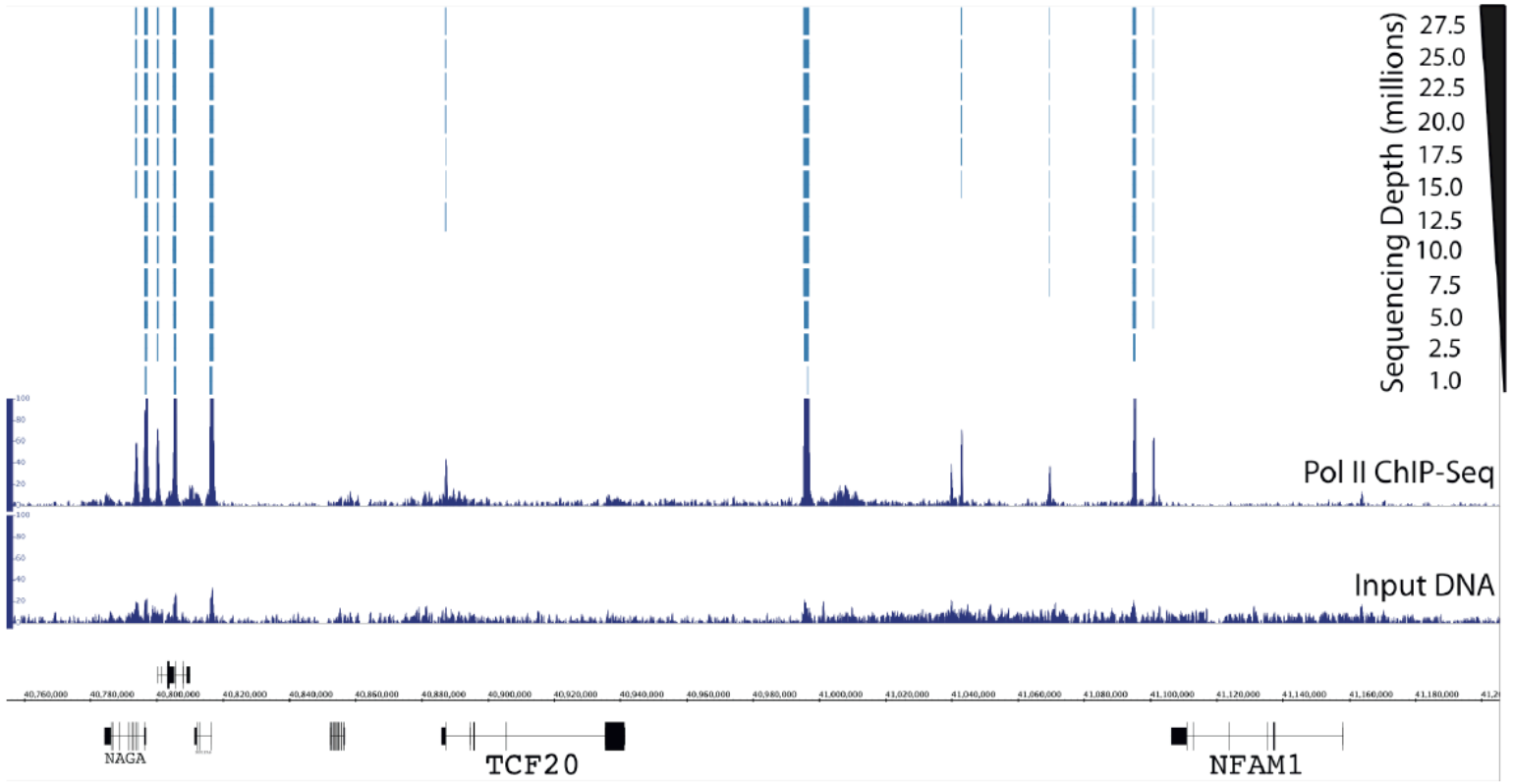
Enriched Sites

[Rozowsky et al. Nat. Biotech ('09)]

Number of Reads for Saturation

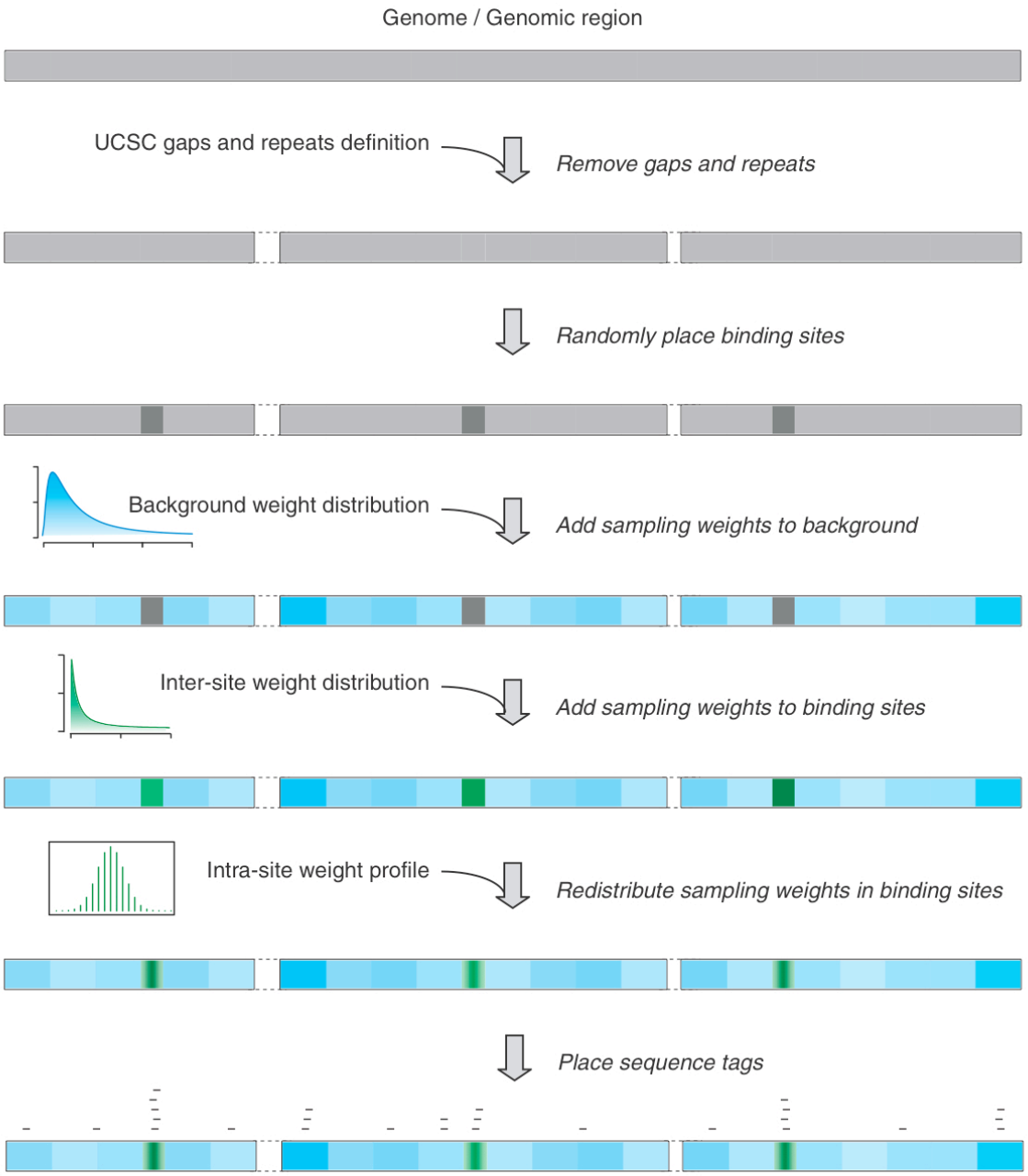


Depth



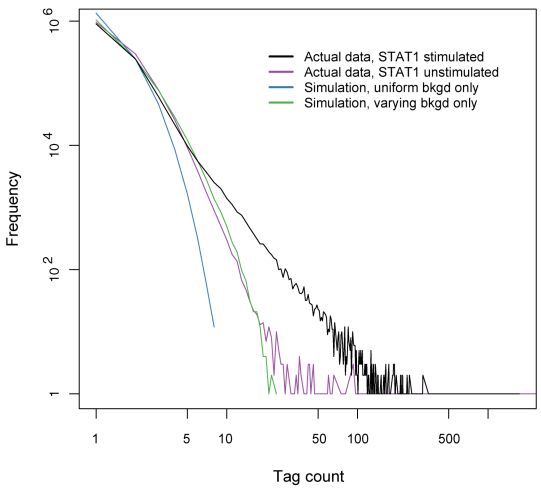
[Rozowsky et al. Nat. Biotech ('09)]

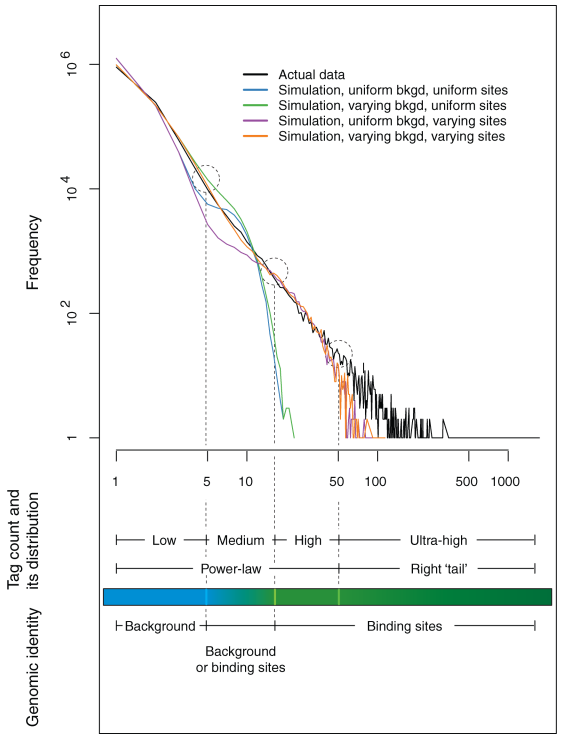
Correcting Chip-seq Signal by Simulating a Non-uniform Genomic Background



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

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





ChIP-sequencing simulation

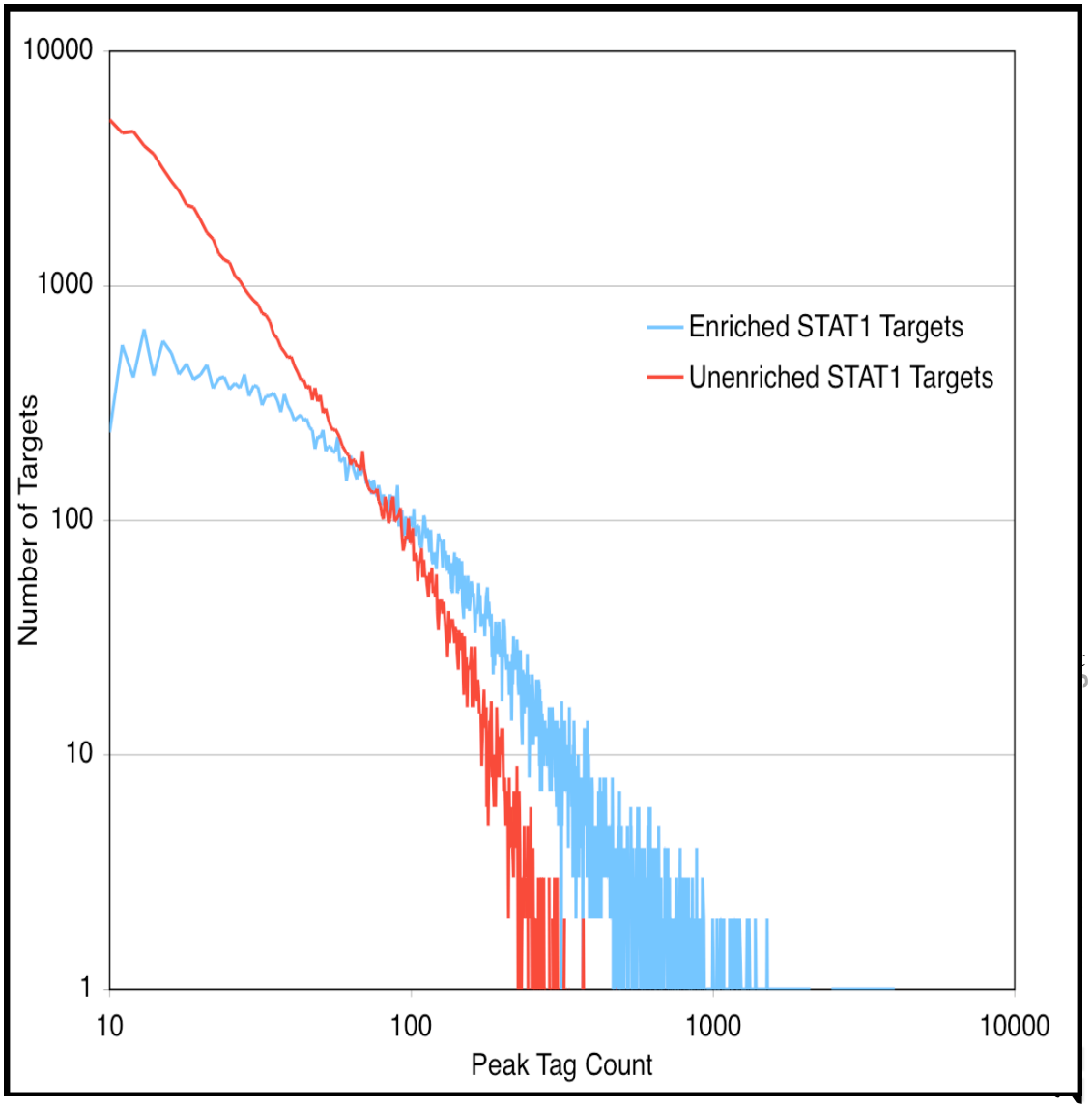
- Contrary to the common belief, the background is mildly fluctuating and contains some 'hot' spots.
- Simple uniform background model does not count for all the variation in the background and thus leads to a serious underestimation of the background noise.

- 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)]

Scored results consistent with simulation

Actual peaks at tail of power-law graph

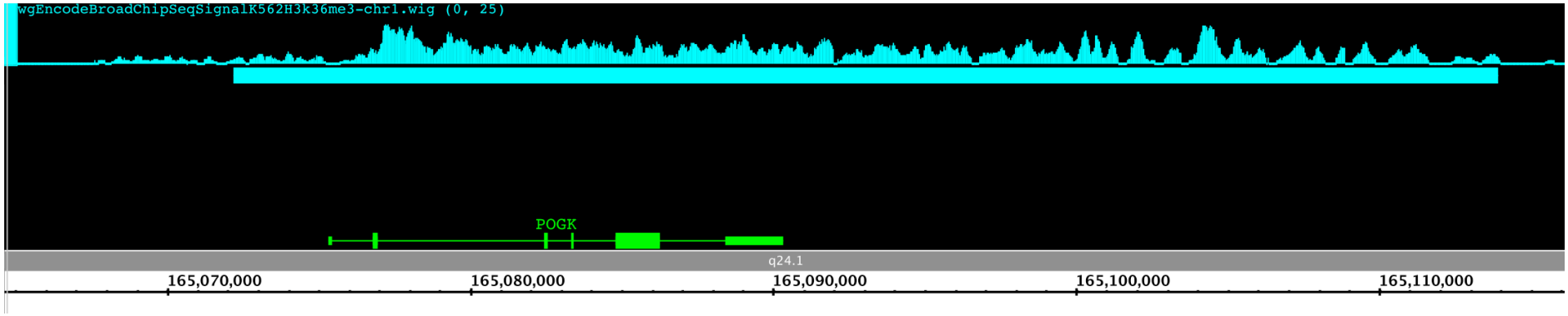


[Rozowsky et al. Nat. Biotech. ('09)]

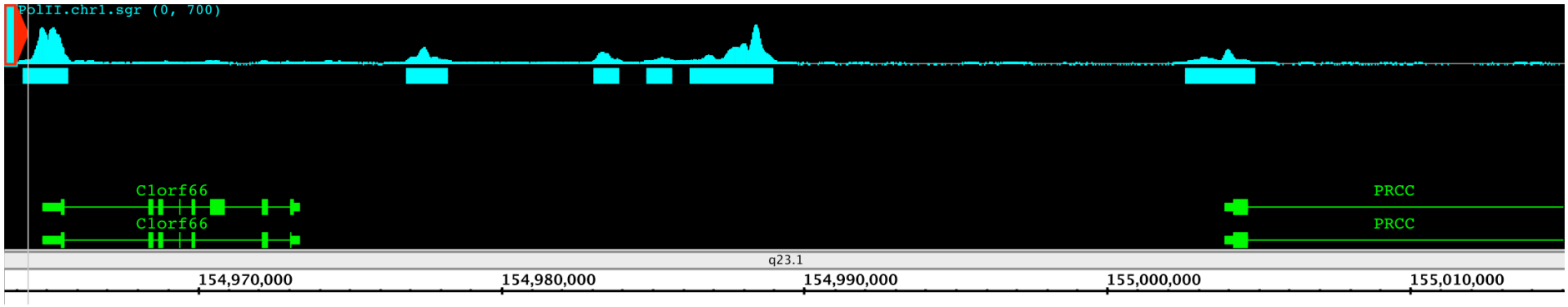
Punctate Blocks vs Broader Ones

- Peakseq - thresholding punctate chipseq peaks
- MSB - calling broader regions in CNV experiment (read depth)
- PEMer - using alternate paired-end strategy to find CNVs

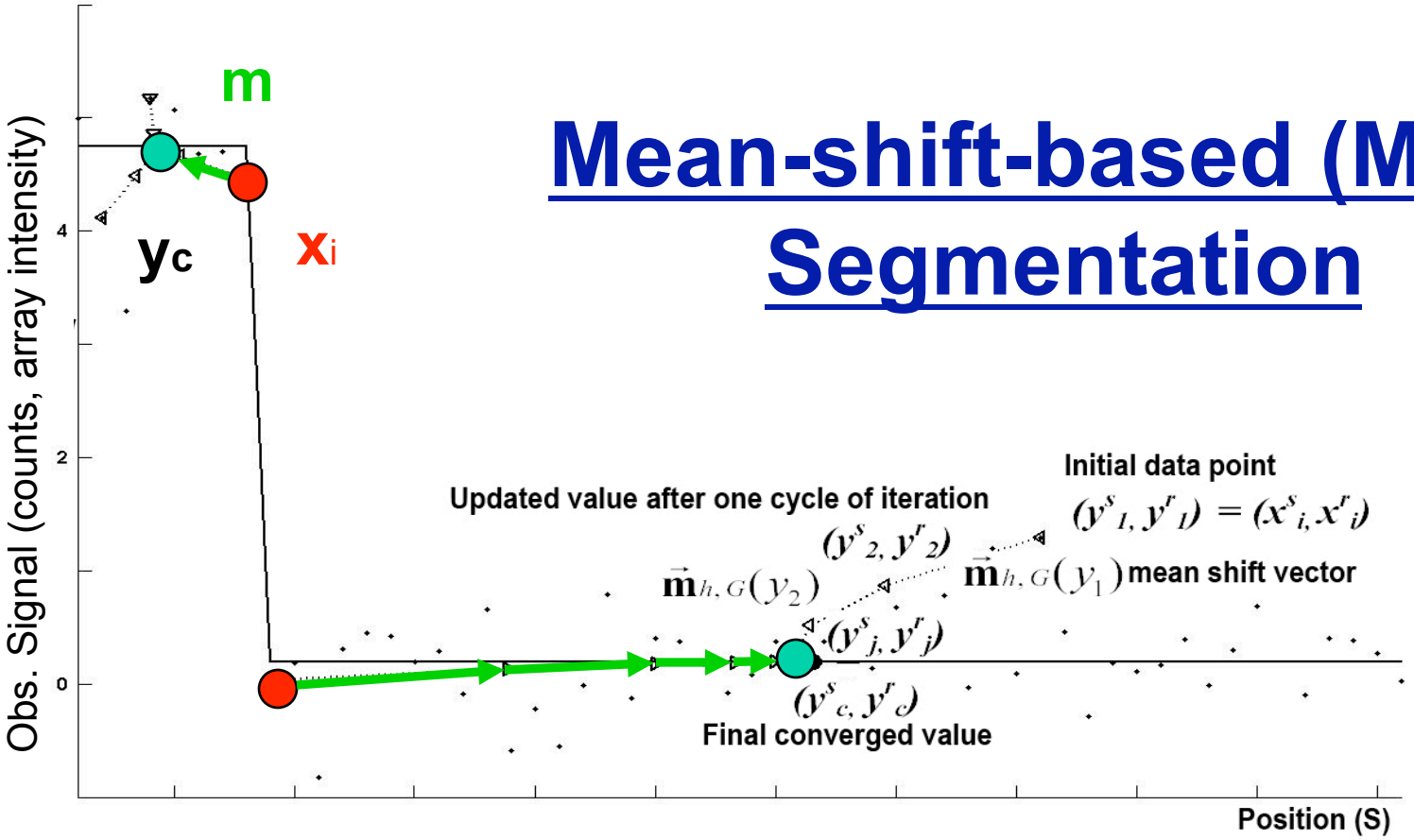
H3K36me3 (Broad Peak Example)



Pol II (Punctate Peak Example)



Mean-shift-based (MSB) Segmentation

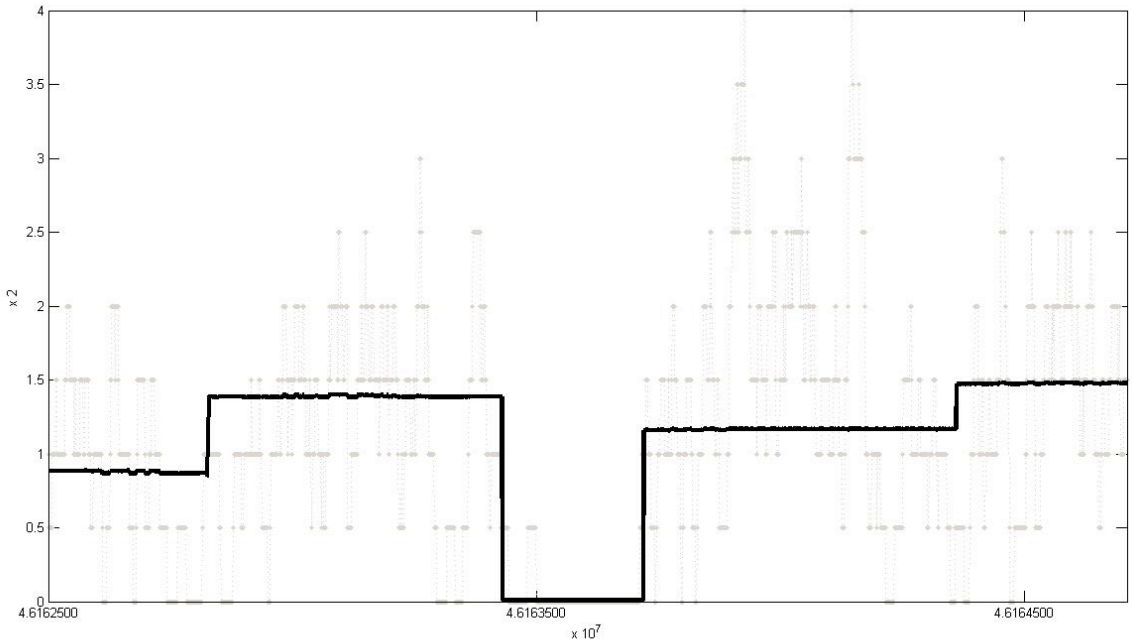


(x_i) Observed depth of coverage counts (or array signal) as samples from PDF
(m) Kernel-based approach to estimate local gradient of PDF
(y_c) Iteratively follow grad to determine local modes

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

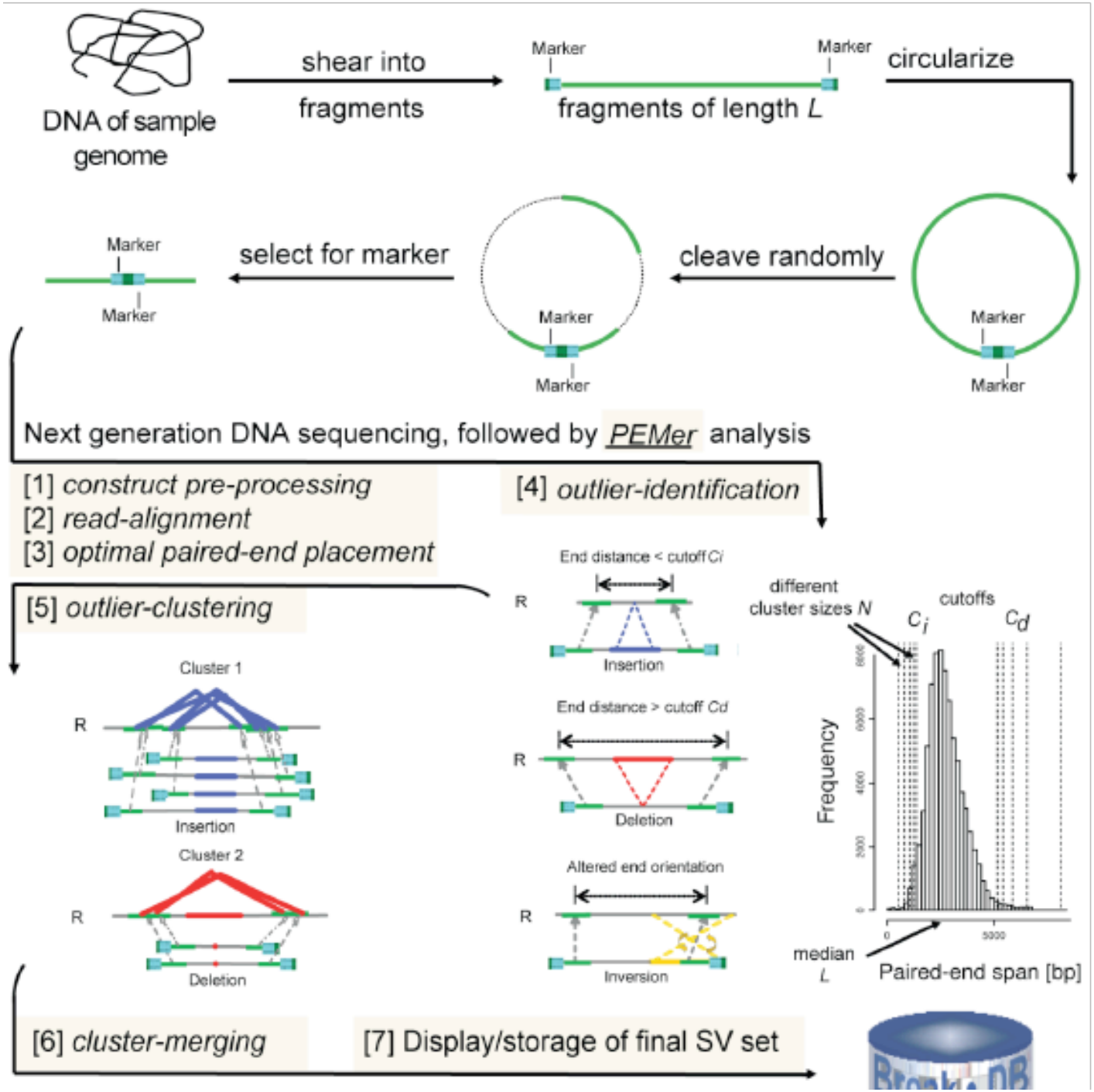


NA11995 (seq. by Sanger, MAQ mapping)
chr 21 (46162500 to 46164711)

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

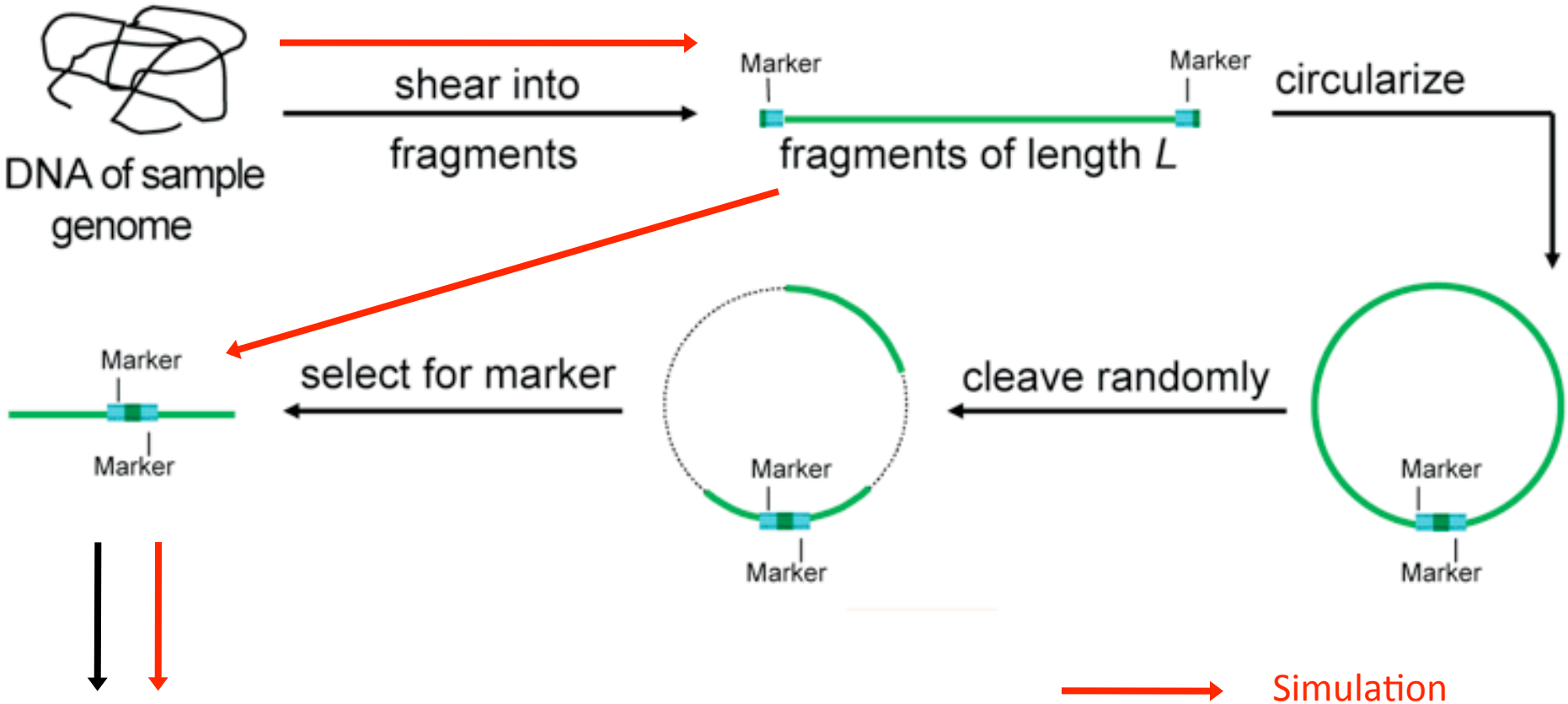
[Wang et al. Gen. Res (in press, '08)]

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



[Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]

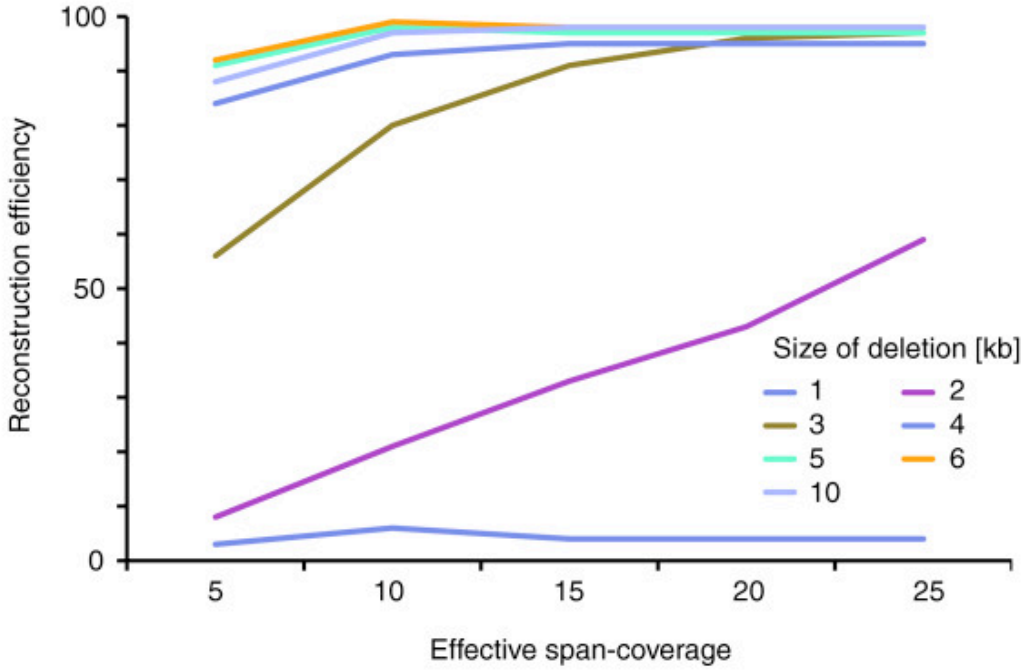
Simulation strategy



[Korbel et al., GenomeBiol. ('09)]

Reconstruction efficiency at different coverage

Deletion size	Reconstruction efficiency at 5x coverage by 2.5 kb inserts
1000	3
2000	11
3000	49
4000	80
5000	91
6000	92
10000	88
Total	414
False positives	5

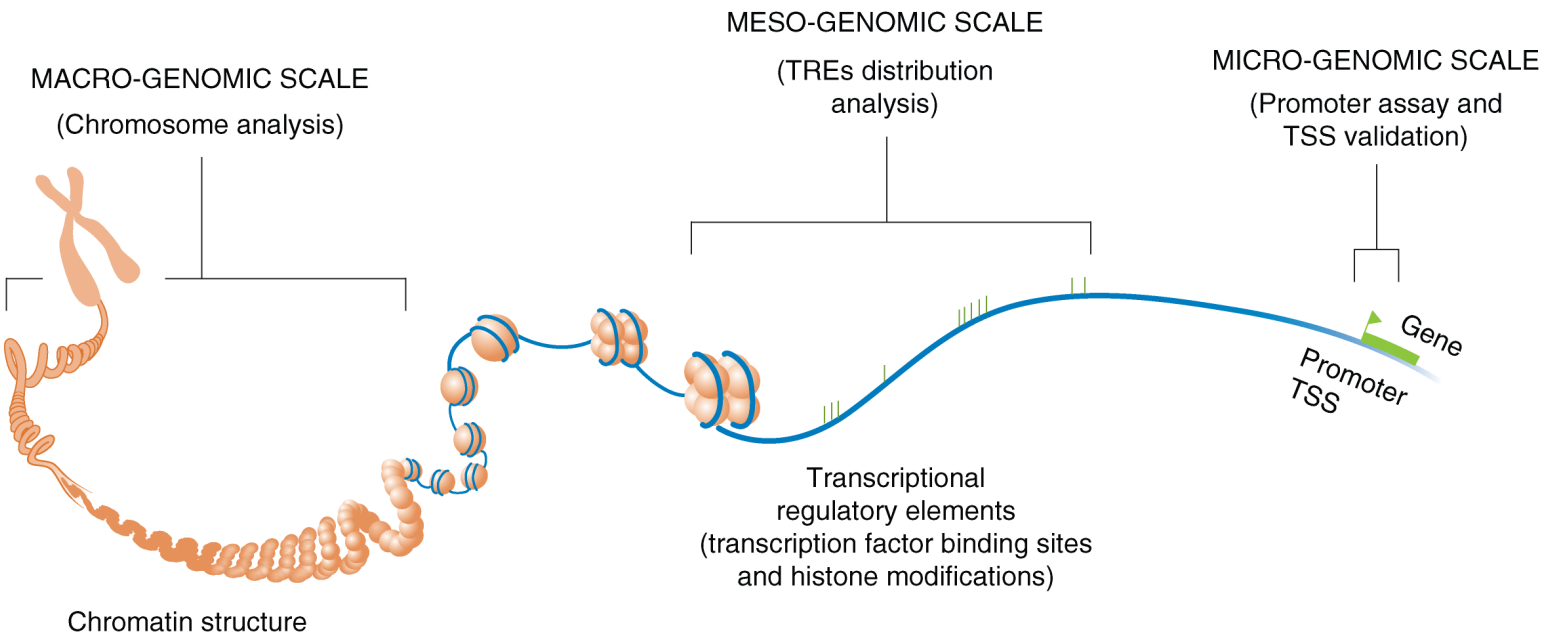


[Korbel et al., GenomeBiol. ('09)]

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

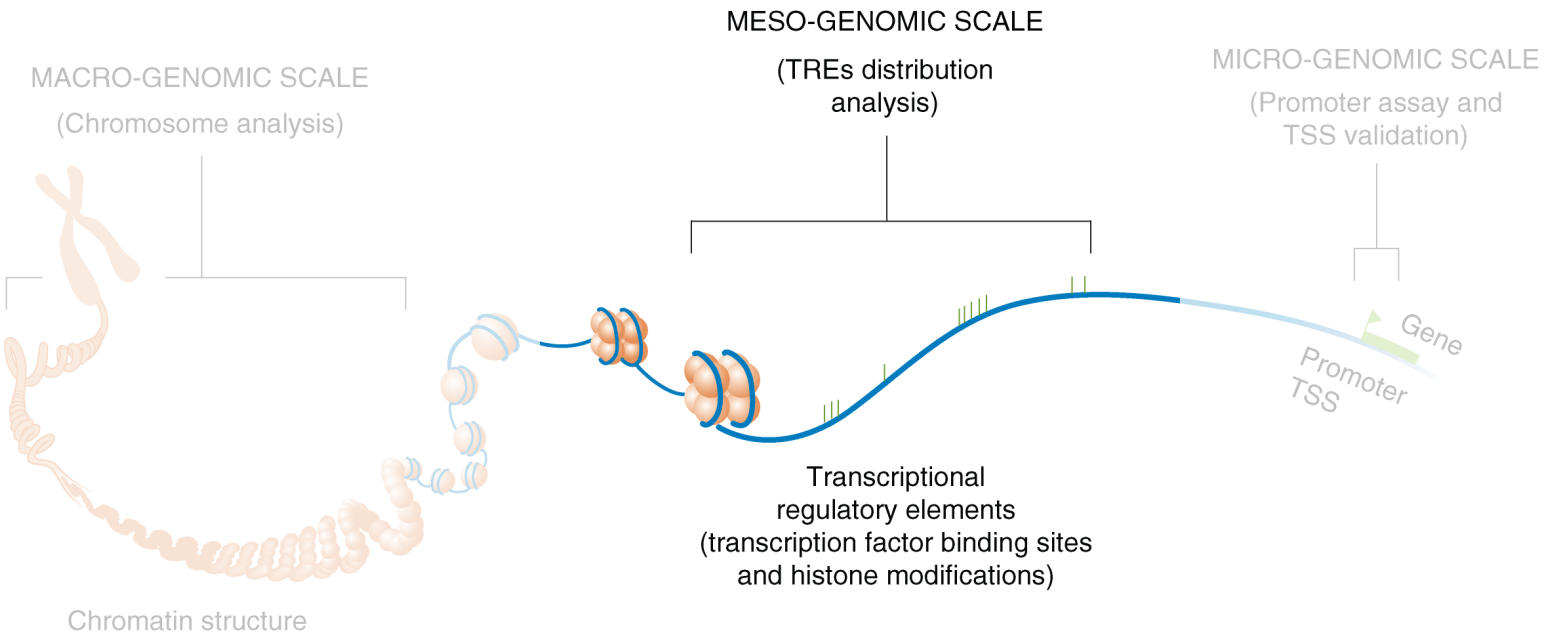
pers. **photo**, see streams.gerstein.info

TRE analysis on the micro-genomic scale



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

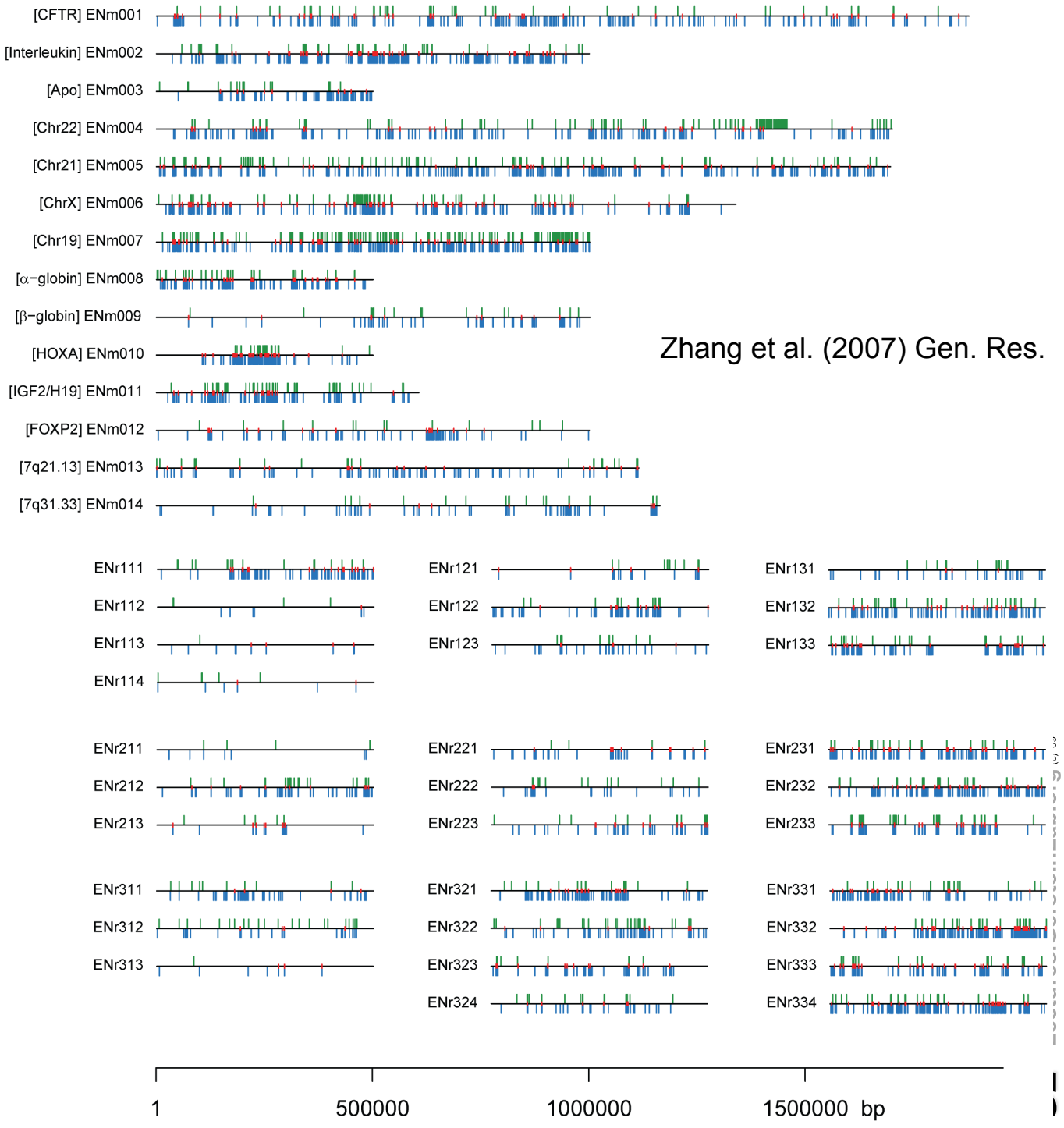
Clustering Binding Sites at ~50kb resolution



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

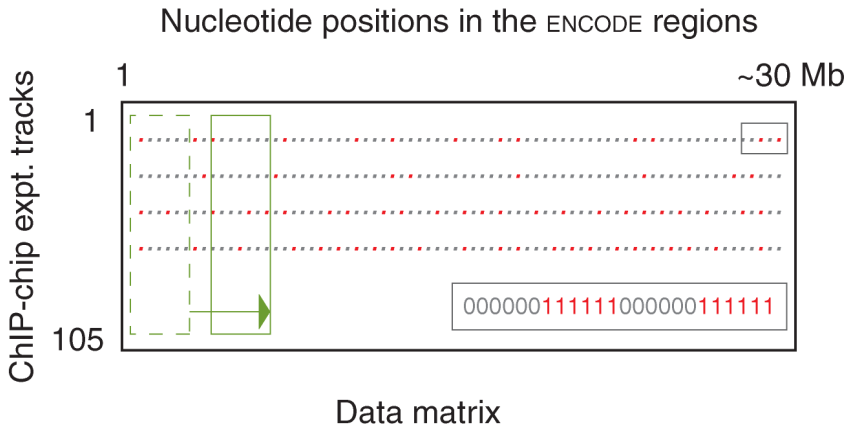
Landscape of ENCODE Transcriptional Regulatory Elements

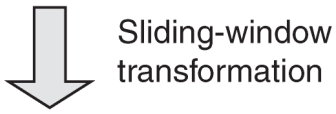
- Analyzed 105 lists of transcriptional regulatory elements in the encode regions
- 29 transcription factors, 9 cell lines, 2 time points
  - RNA Pol2
  - Histone modifications such as Ac & Me
  - Core promoters
  - Promoter proximal elements
  - Others such as enhancers, silencers, insulators, & response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

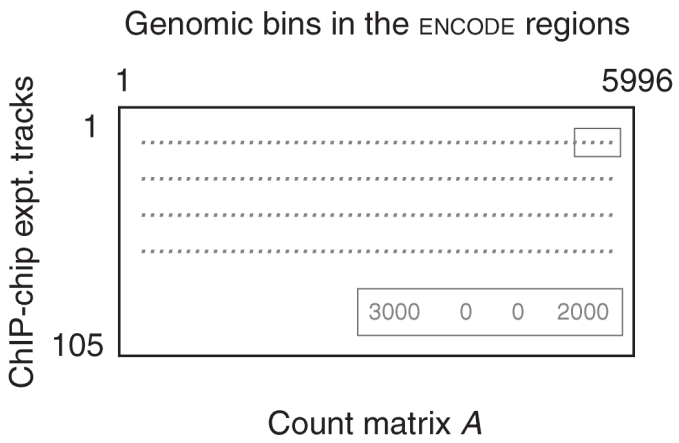


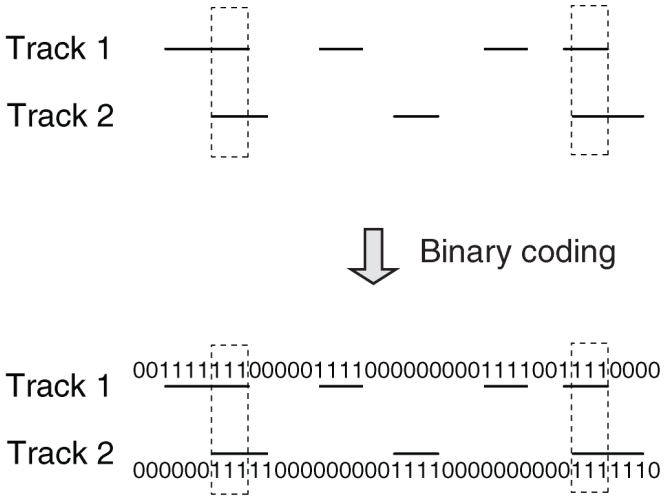
Zhang et al. (2007) Gen. Res.

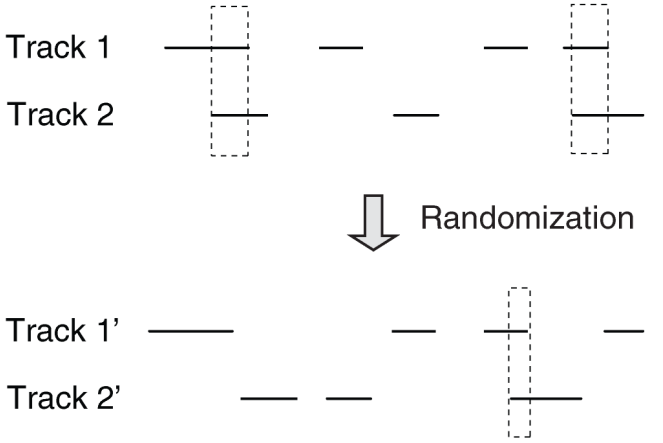
Collect Total Hits for Each Factor in ~6000 Bins of 10 to 100 kb and Compare to Random Control







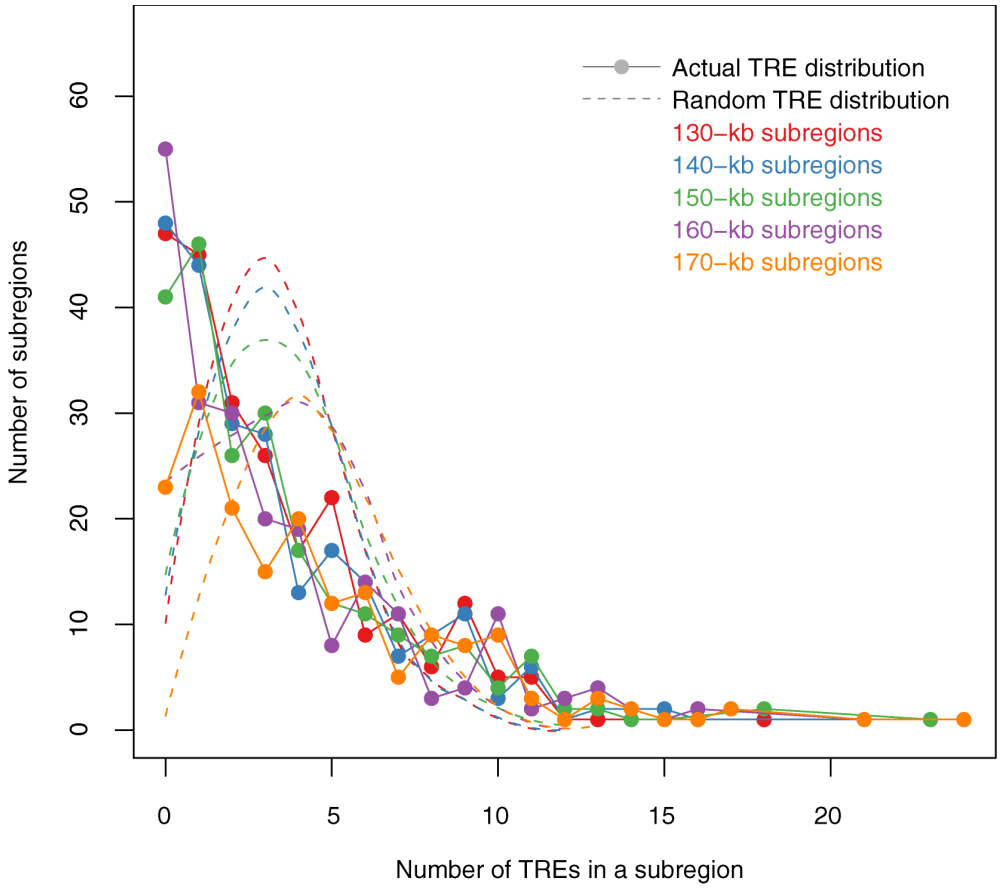




Zhang et al. (2007) Gen. Res.

Non-random distribution of TREs

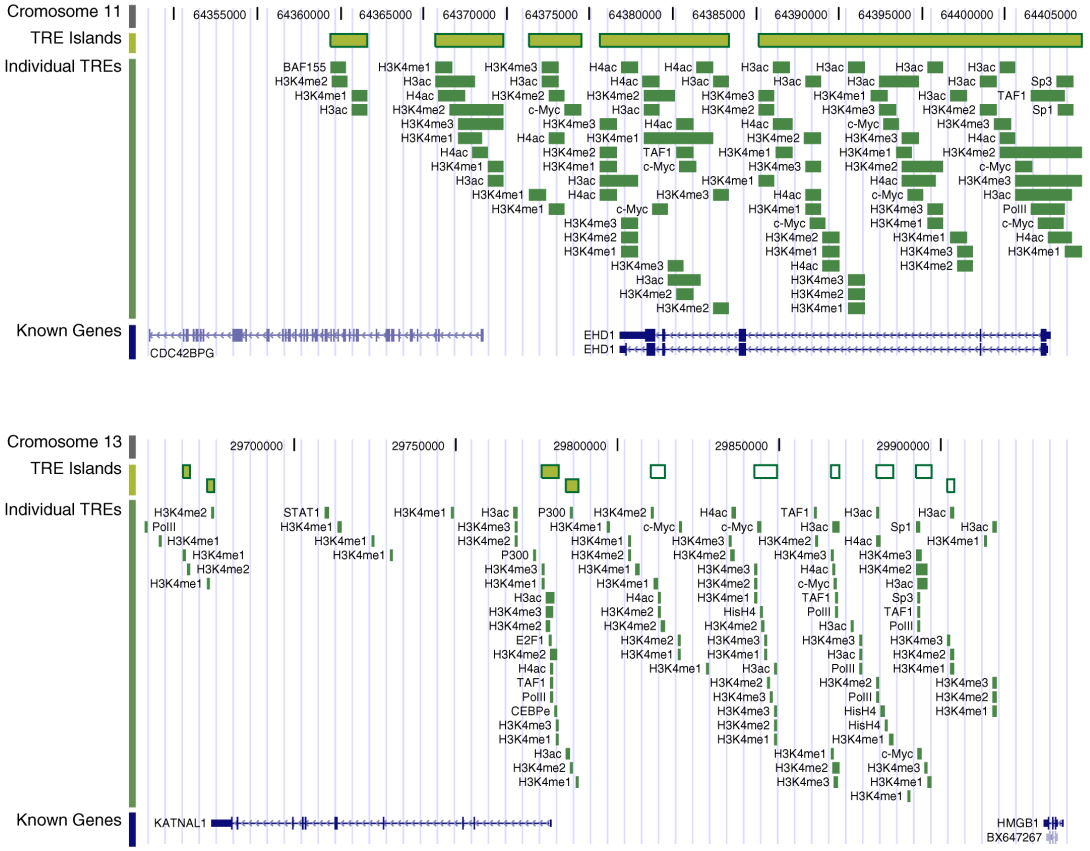
- TREs are not evenly distributed throughout the encode regions ($P < 2.2\times10^{-16}$).
- The actual TRE distribution is power-law.
- The null distribution is 'Poissonesque.'
- Many genomic subregions with extreme numbers of TREs.



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...

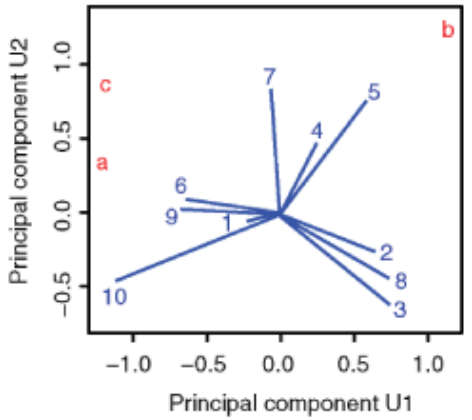
50kb Genomic Bins: 1,2,3...

	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

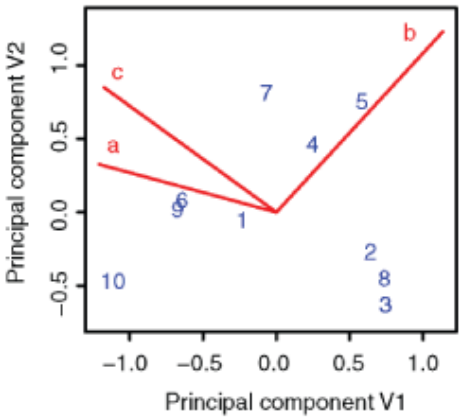


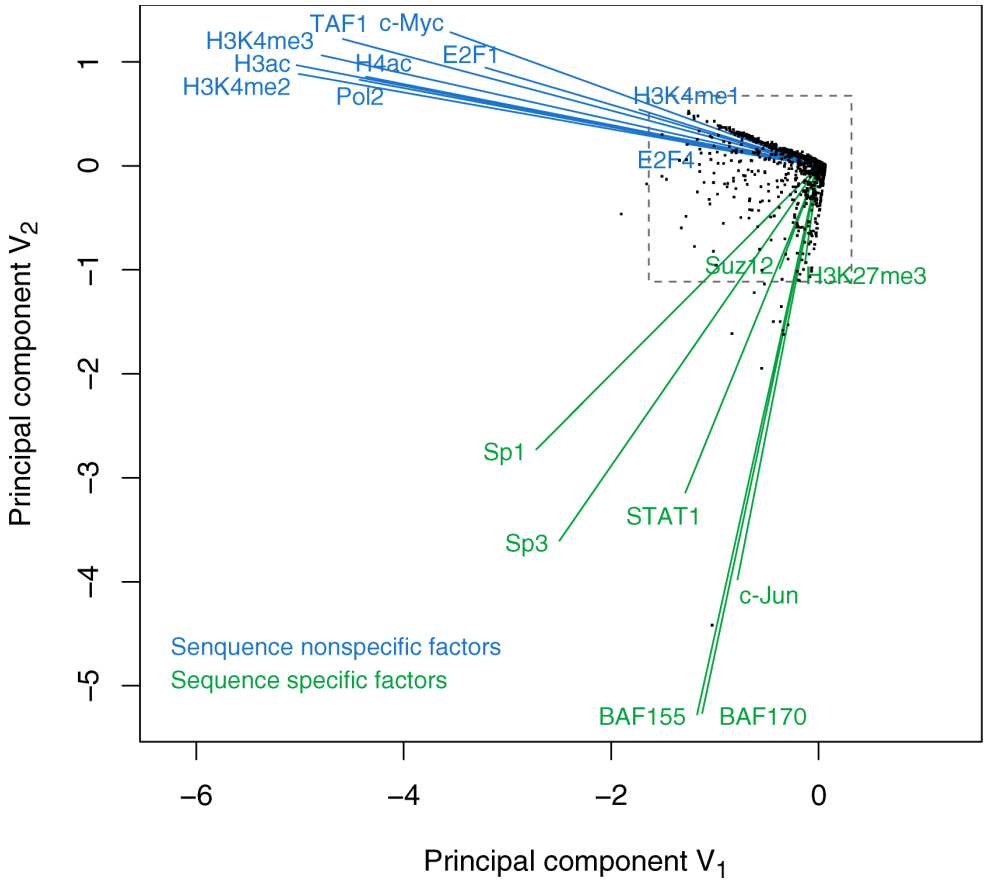
A^T

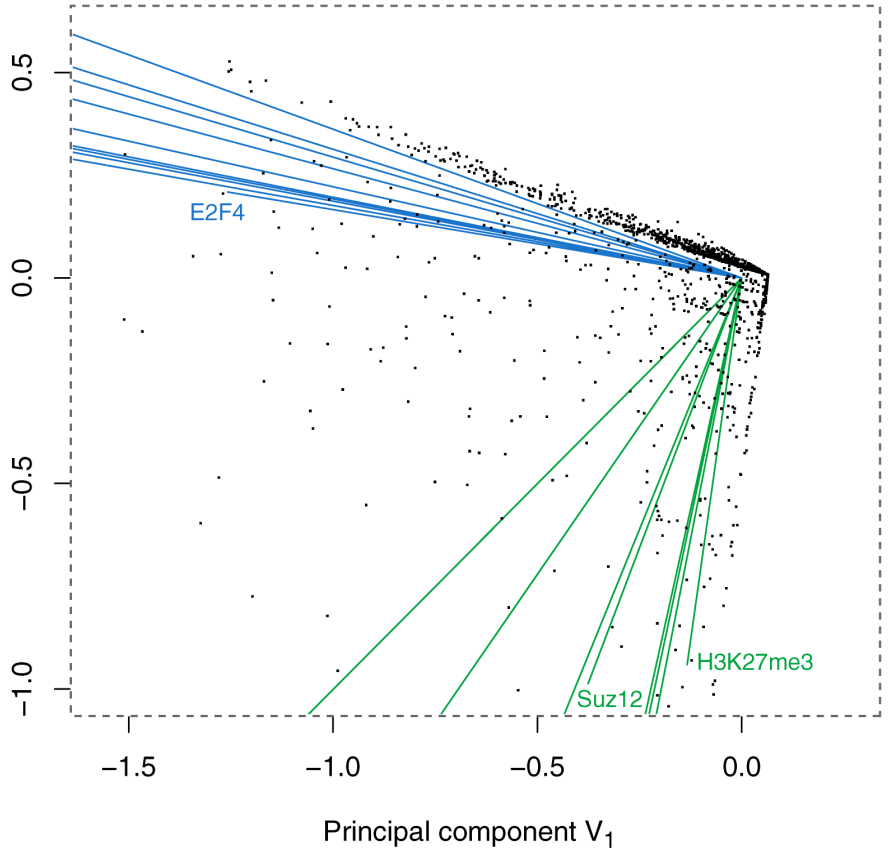
	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

	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







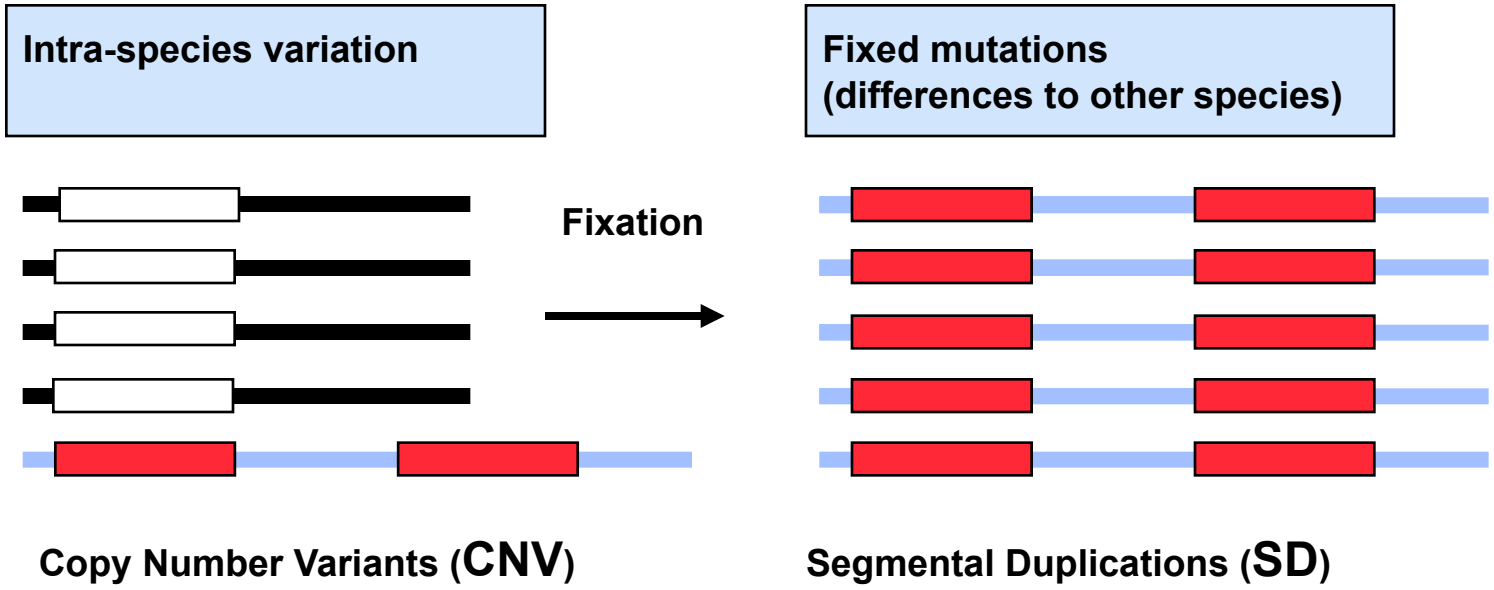
Results of Biplot

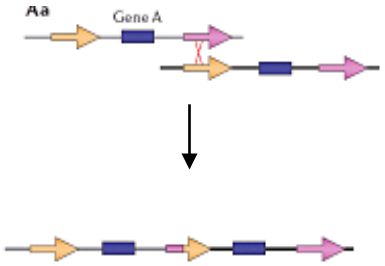
- Biplot groups TFs into sequence-specific and sequence-nonspecific clusters.
  - ◊ c-Myc may behave more like a sequence-nonspecific TF.
  - ◊ H3K27me3 functions in a transcriptional regulatory process in a rather sequence-specific manner.
- Genomic Bins are associated with different TFs and in this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

Zhang et al. (2007) Gen. Res.

Analyzing Repeated Blocks in the Genome (SDs & CNVs)

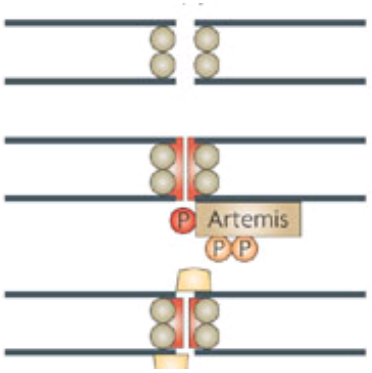
SEGMENTAL DUPLCATIONS AND COPY NUMBER VARIANTS ARE RELATED PHENOMENA AND HAVE BEEN CREATED BY SEVERAL DIFFERENT MECHANISMS





NAHR
(Non-allelic homologous recombination)

Flanking repeat
(e.g. Alu, LINE…)

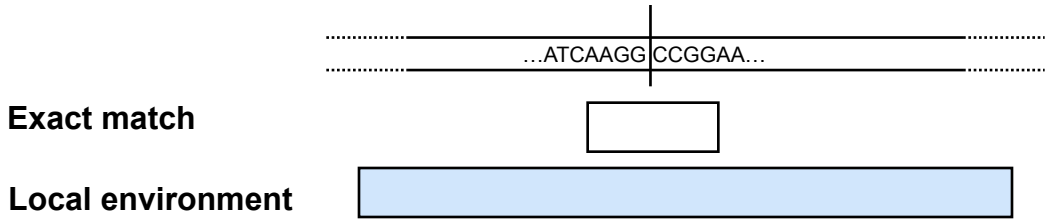


NHEJ
(Non-homologous-end-joining)

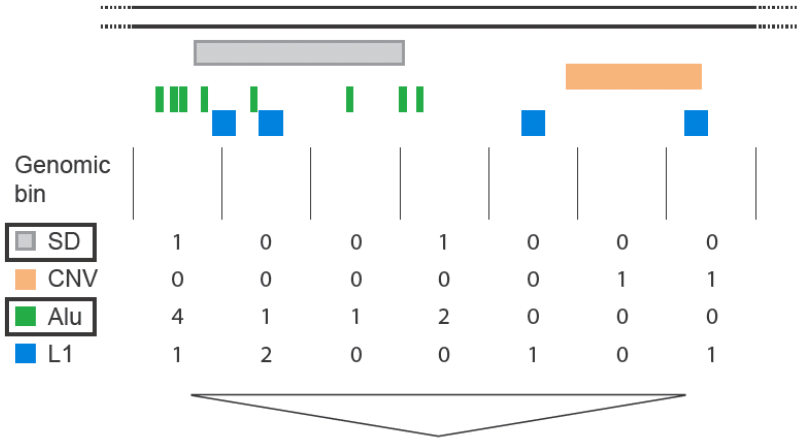
No (flanking) repeats.
In some cases <4bp microhomologies

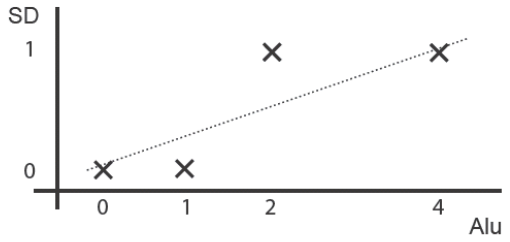
PERFORM LARGE SCALE CORRELATION ANALYSIS TO DETECT REPEAT SIGNATURES OF SDs AND CNVs

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



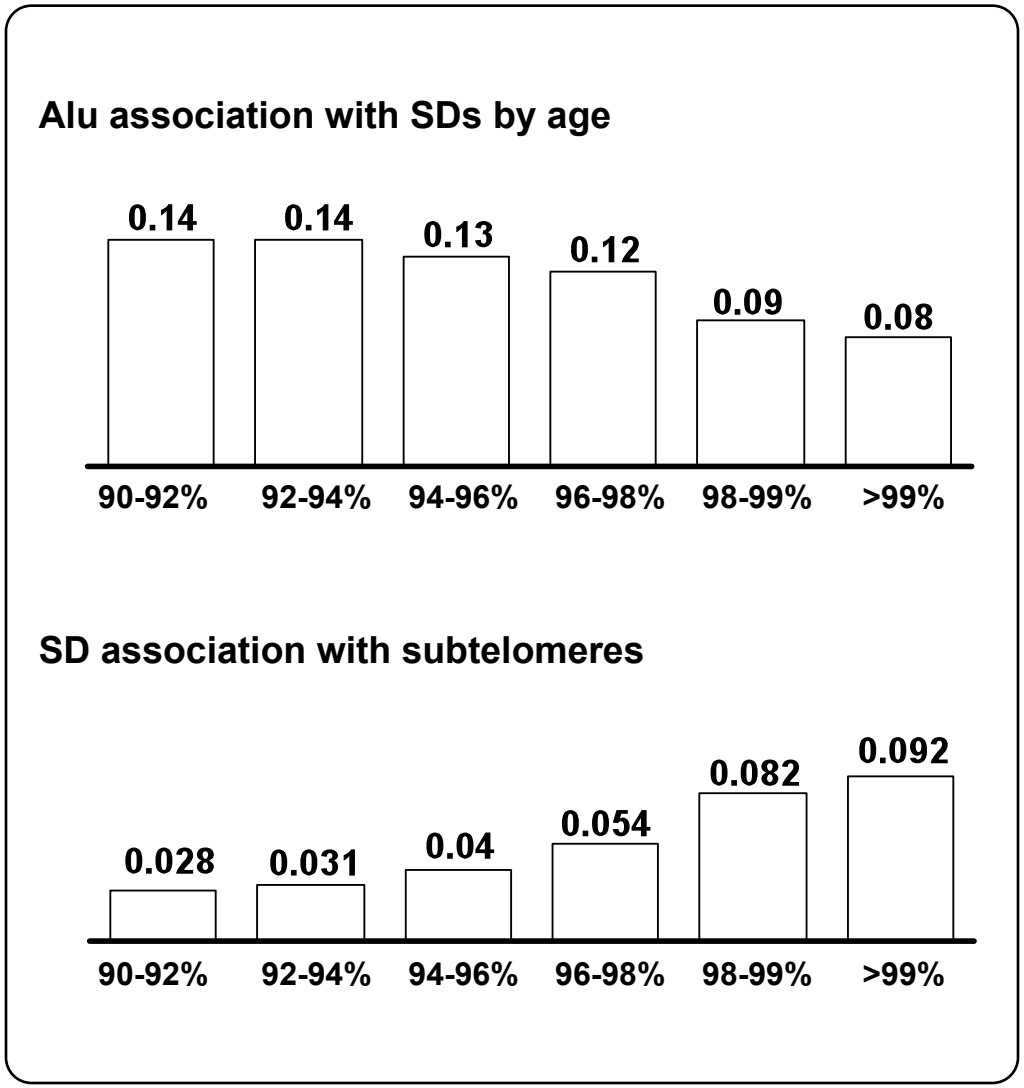
1. Survey a range of genomic features
2. Count the number of features in each genomic bin (100kb)
3. Calculate correlations / enrichments using robust stats





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

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

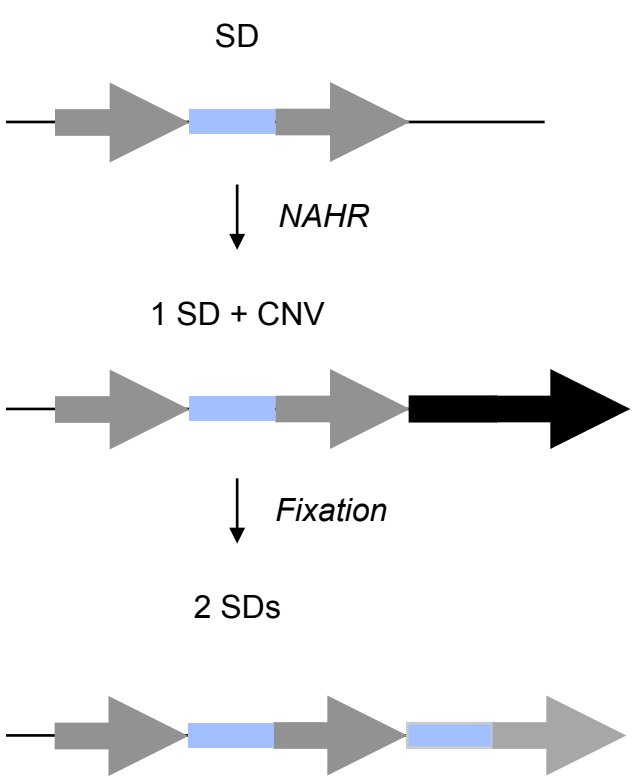


- The co-localization of Alu elements with SDs is highly significant.
- Older SDs have a much higher association with Alus than younger SDs.
- Hence it is likely, that Alu elements were more active in mediating NAHR in the past (consistent with the Alu burst)
- Younger SDs are more likely to be localized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

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

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

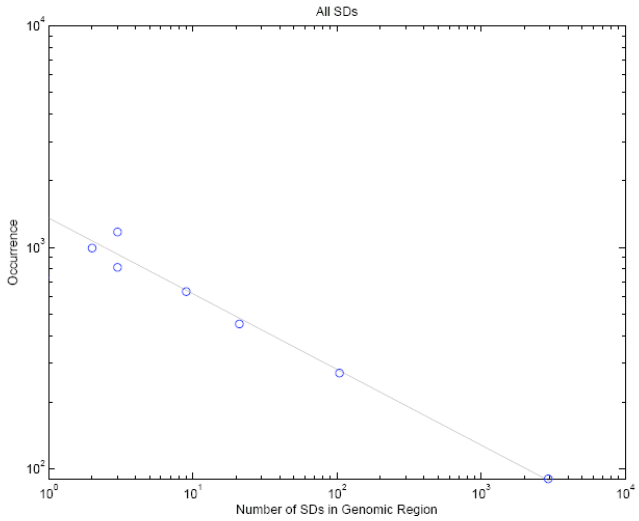
Hypothesis



"SD selfpropagation"

Corollary

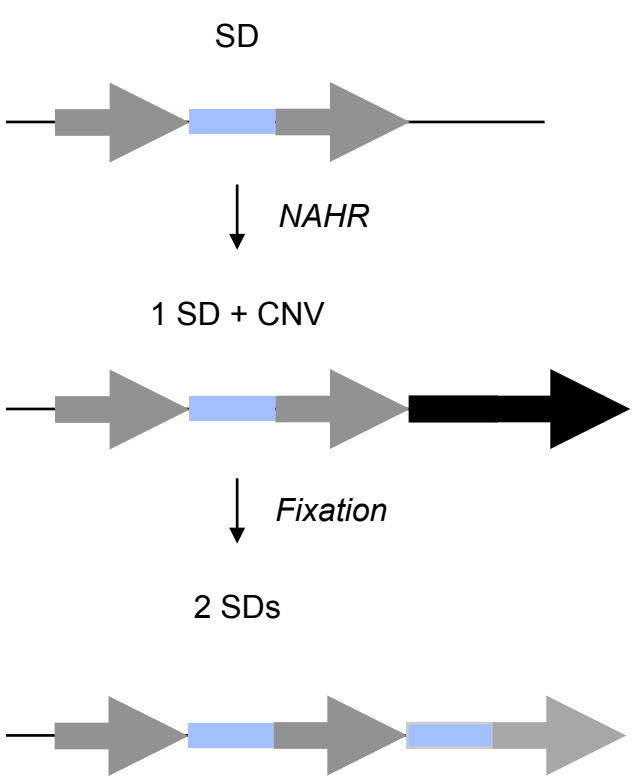
- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- Such mechanisms ("preferential attachment") are well studied in physics and should leads a very skewed ("power-law") distribution of SDs.



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

FOCUSSING ON SDS: SDs COLOCALIZE WITH EACH OTHER

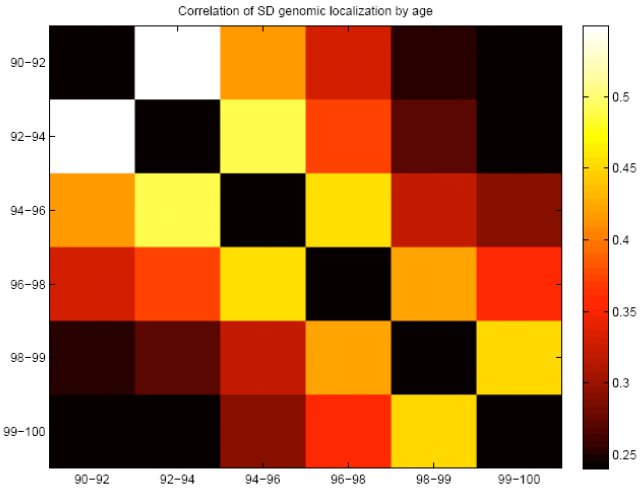
Hypothesis



"SD self-propagation"

Corollary

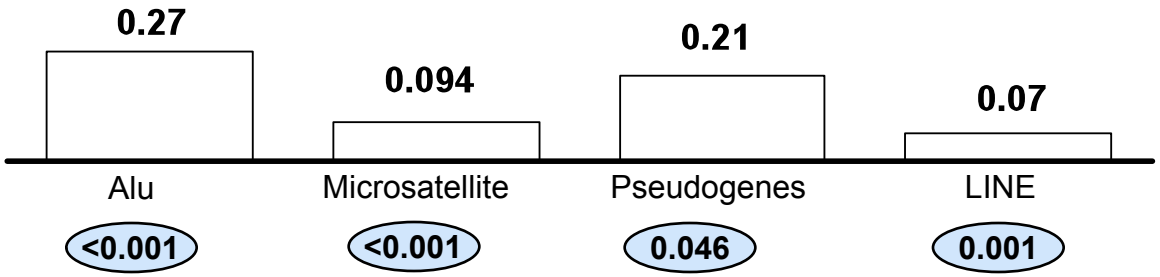
- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- SDs of similar age should co-localize better with each other:



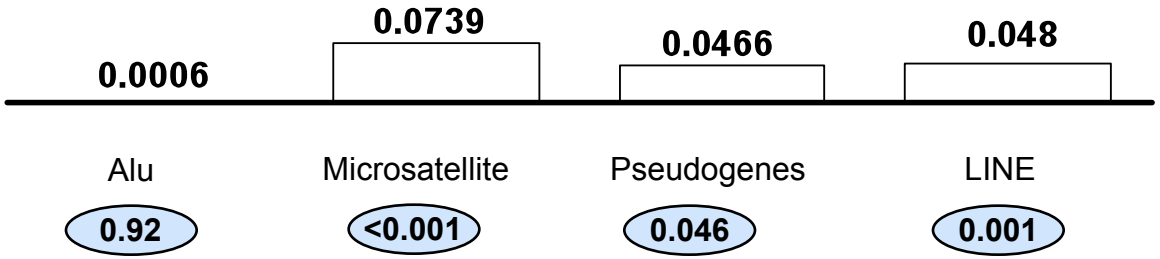
[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs

SD association with repeats

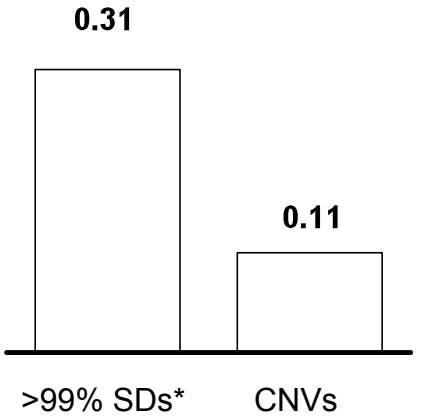


CNV association with repeats

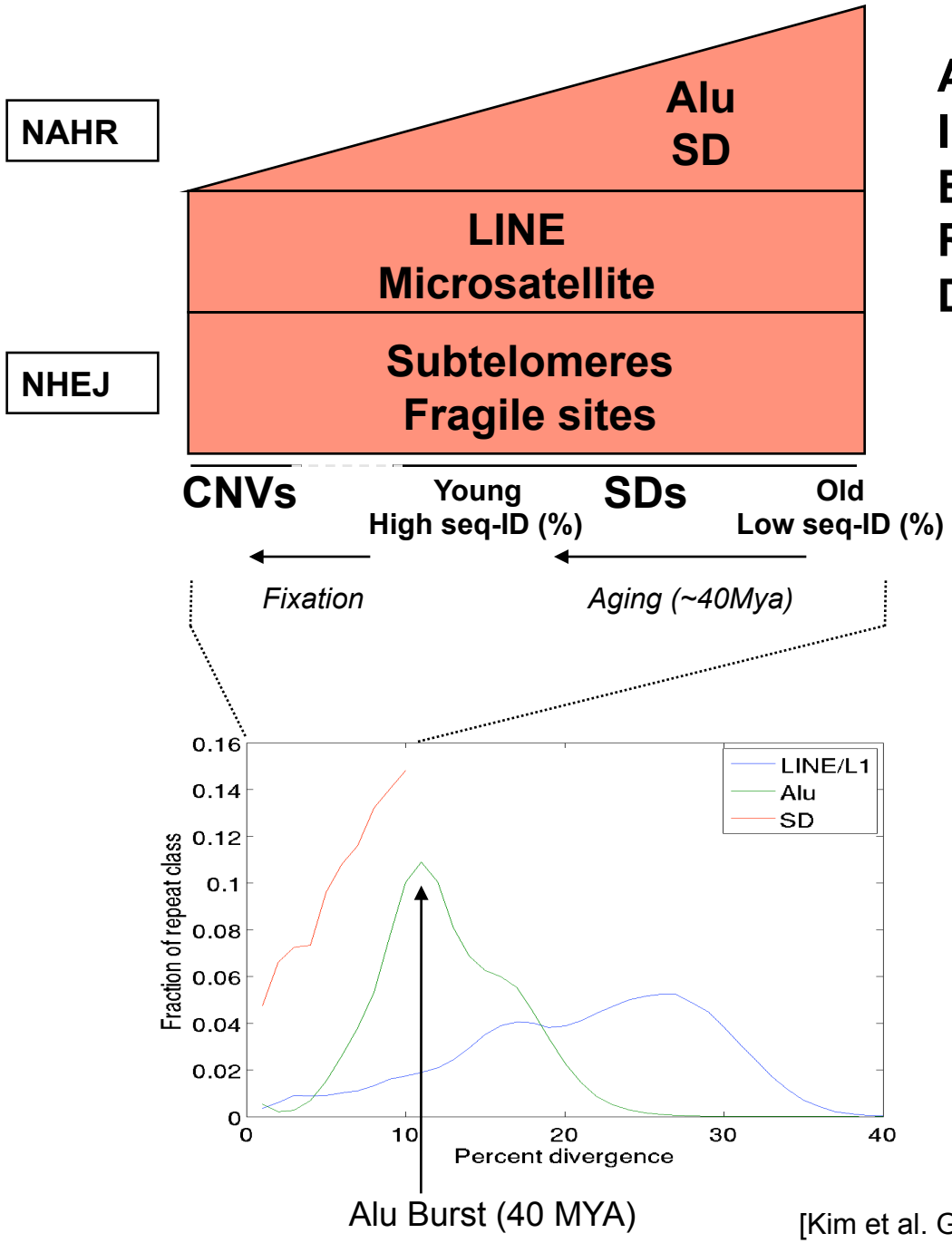


CNVs ARE LESS ASSOCIATED WITH SDs THAN THE GENERAL SD TREND

CNV Association with SDs



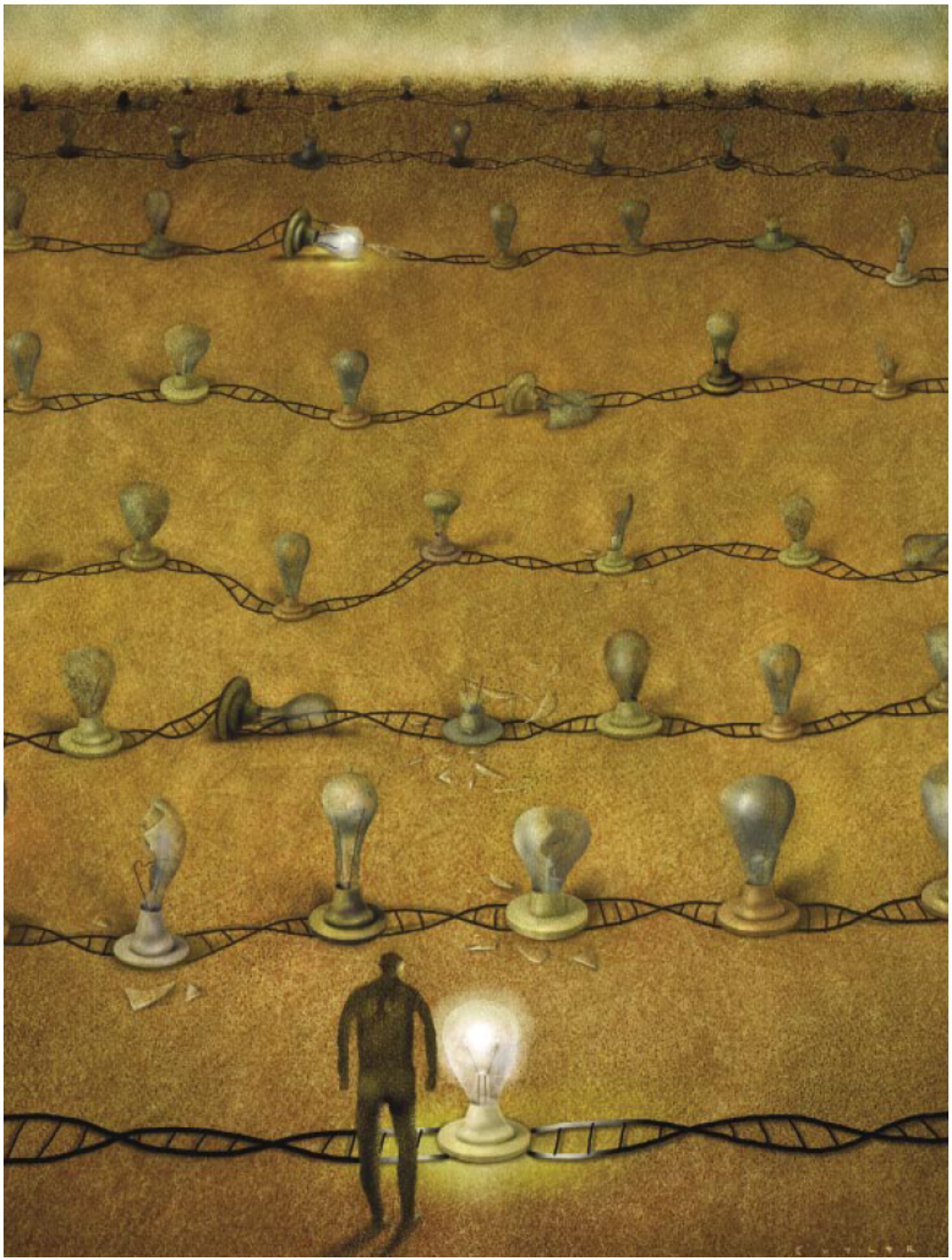
[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]



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. ('08), arxiv.org/abs/0709.4200v1]



Integrative Analyses:

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

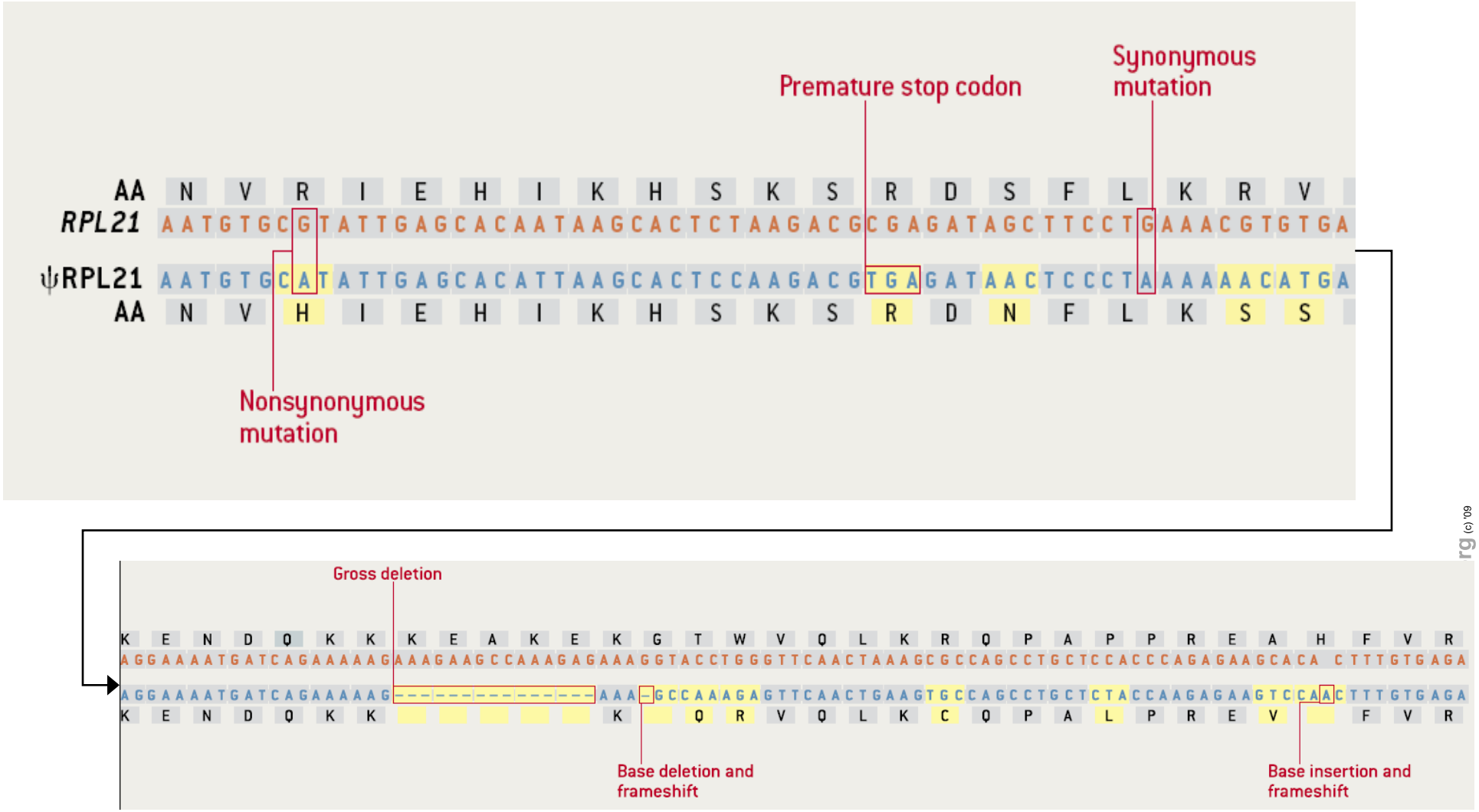
Illustration from Gerstein & Zheng (2006). Sci Am.

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?…

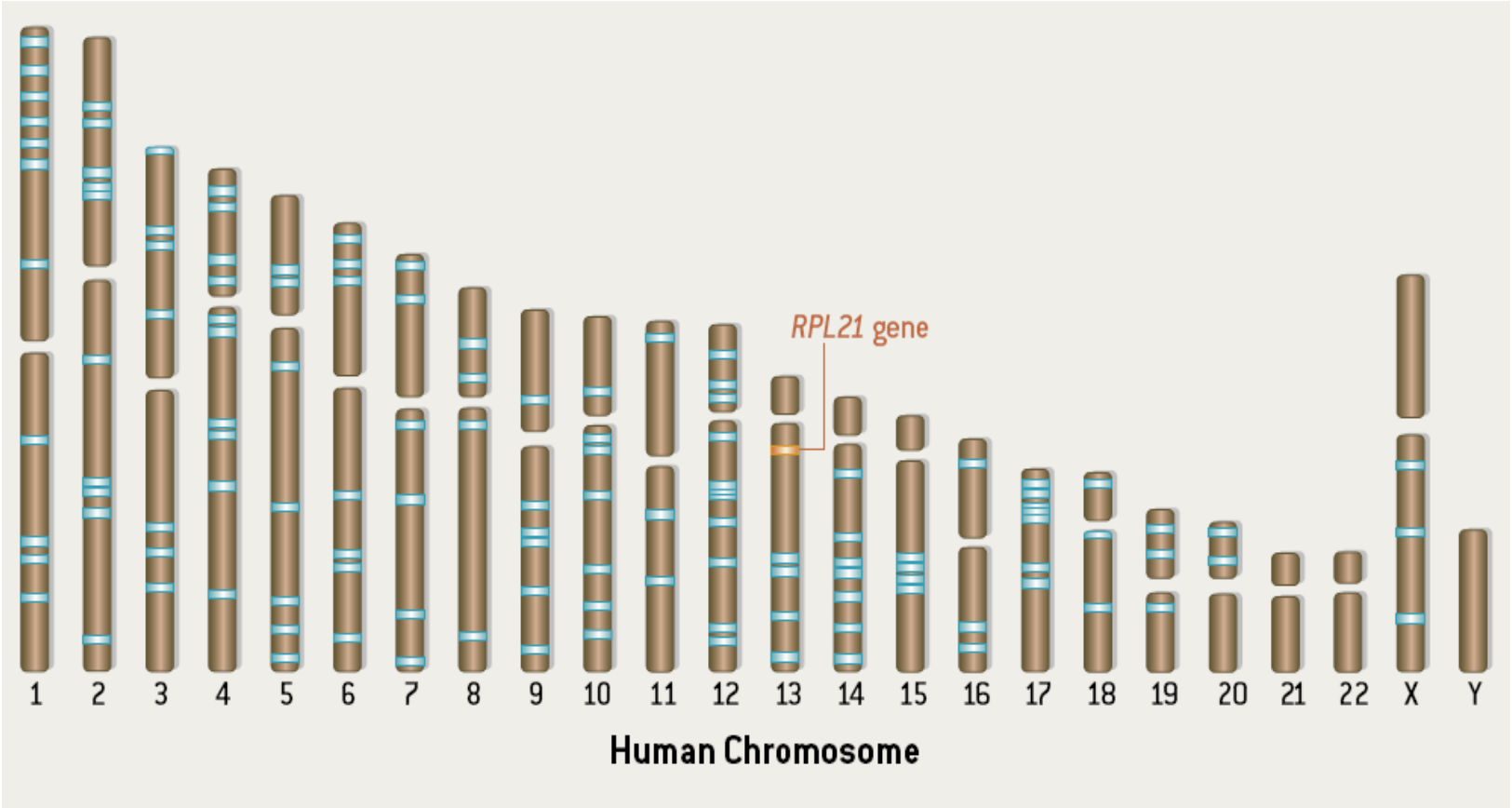
[Mighell et al. *FEBS Letts*, 2000]

Identifiable Features of a Pseudogene (ψRPL21)



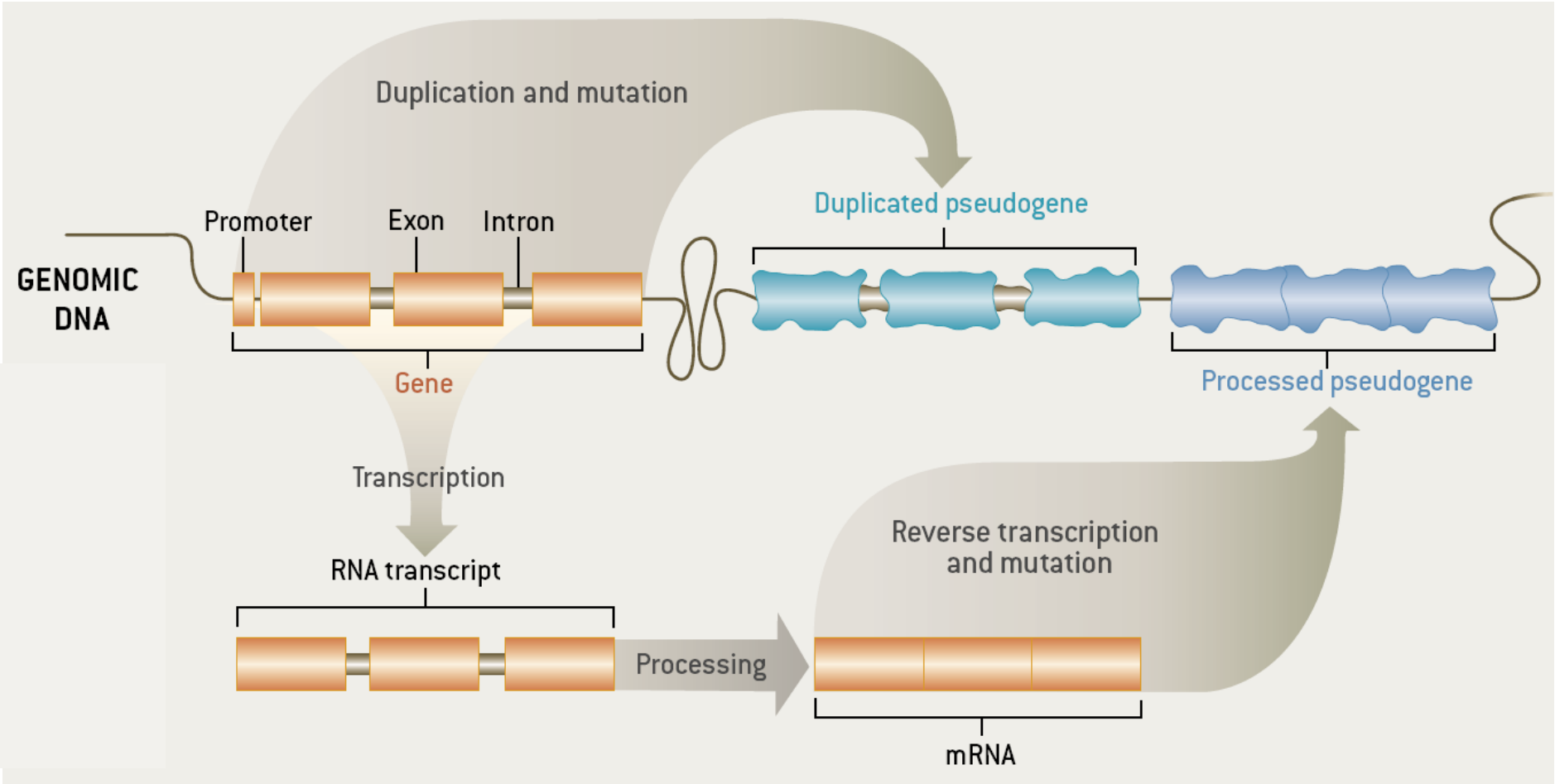
Gerstein & Zheng. Sci Am 295: 48 (2006).

Distribution of Human Pseudogenes (for RPL21) across the chromosomes



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

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



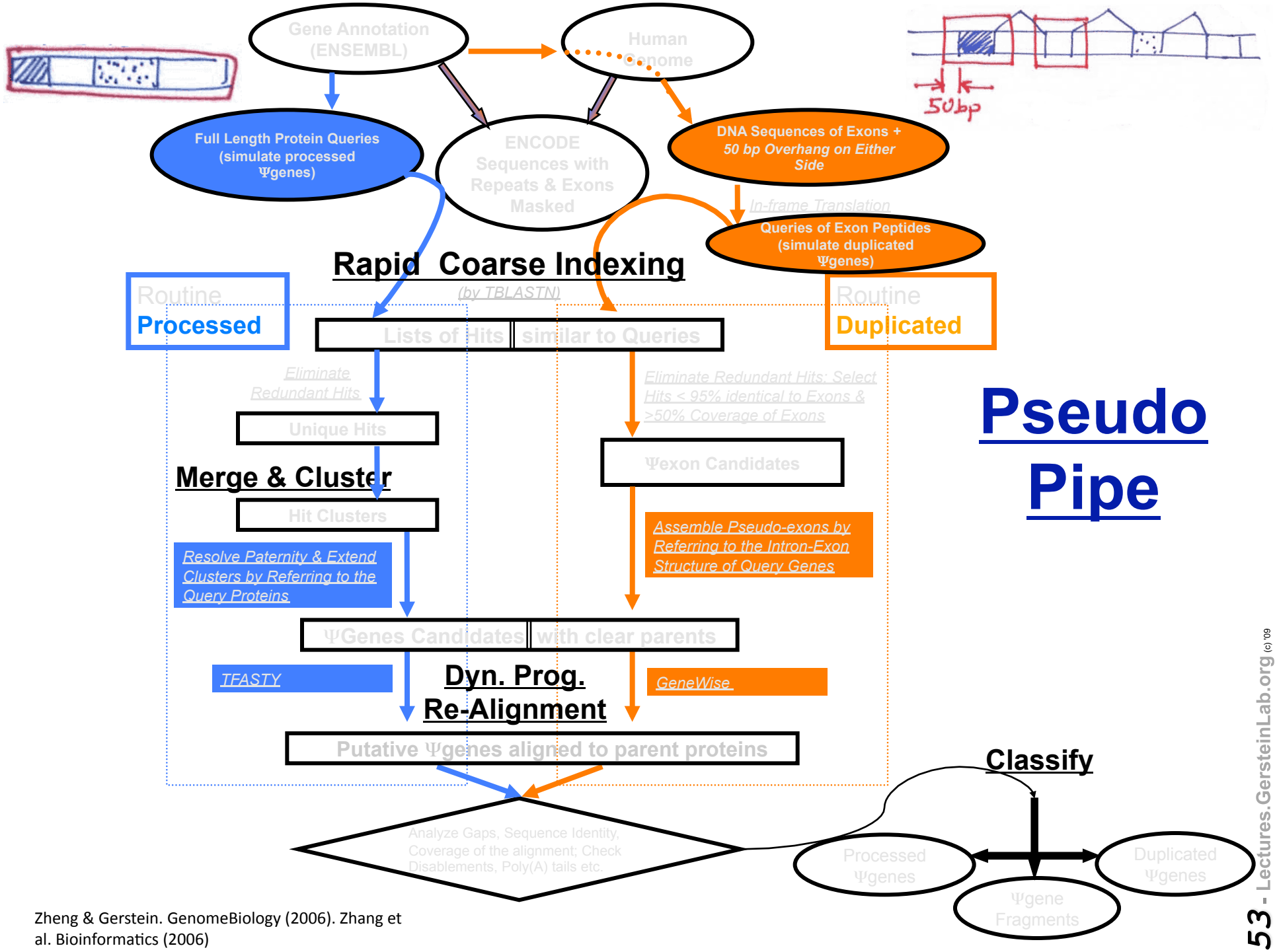
Gerstein & Zheng. Sci Am 295: 48 (2006).

Overall Flow: Pipeline Runs, Coherent Sets, Annotation, Transfer to Sanger

- Overall Approach
  1. Overall Pipeline runs at Yale and UCSC, yielding raw pseudogenes
  2. Extraction of coherent subsets for further analysis and annotation
  3. Passing to Sanger for detailed manual analysis and curation
  4. Incorporation into final GENCODE annotation
  5. Pipeline modification

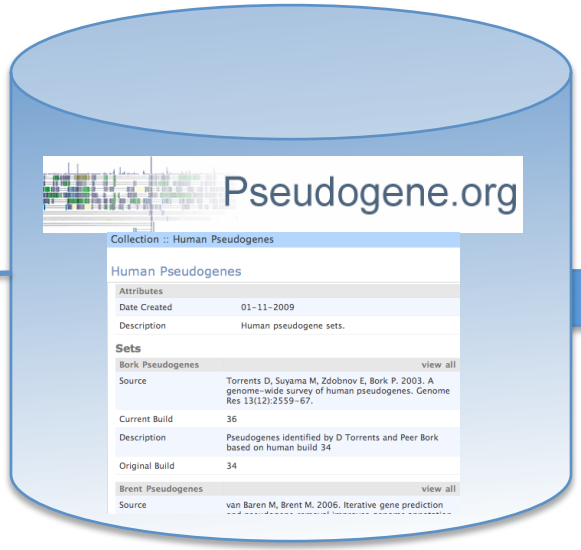
- Automatic pipeline currently gives ~23K [pseudogene.org]
- Chronology of Sets
  1. Encode Pilot 1%
  2. Unitary pseudogenes (Hard)
  3. **Ribosomal Protein pseudogenes (easy)**
  4. **Transcribed pseudogenes annotated earlier**
  5. **....**

Pseudogene Tools: Assignment Pipeline & DB



Pseudo Pipe

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



Flat Files ⟷ DAS

Table Browser

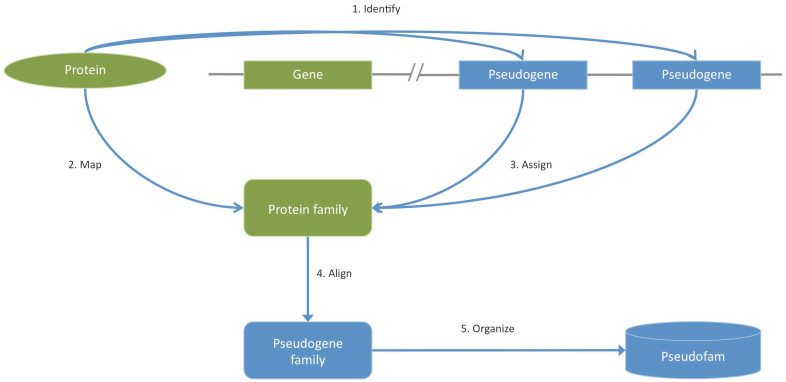
tables.pseudogene.org ⟷ UCSC Genome Browser

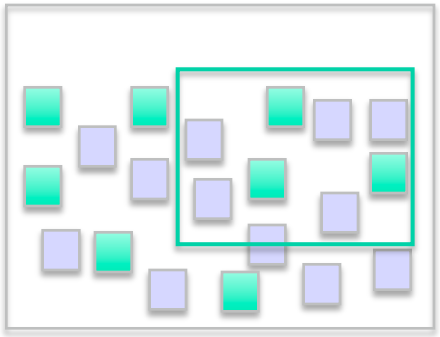
- 12 eukaryotic species
  - Human, mouse, rat, chimp…
  - 100,052 pseudogenes
- 64 prokaryotic species
  - 6,412 pseudogenes
- 28,237 human pseudogenes
  - 13+ unique human sets

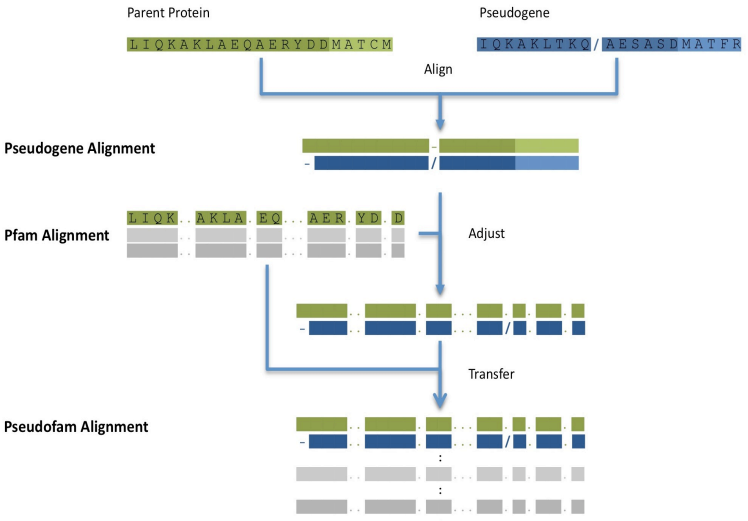
[Lam et al., NAR DB Issue ('09)]

Pseudofam Construction

- Data Generation
  - ◊ Identify pseudogenes by proteins and map parent proteins to protein families
- **Alignment**
  - ◊ Align pseudogene to parent
  - ◊ Transfer alignment from Pfam
  - ◊ Combine and adjust the alignments to build the pseudofam alignment
- **Statistics**
  - ◊ Enrichment





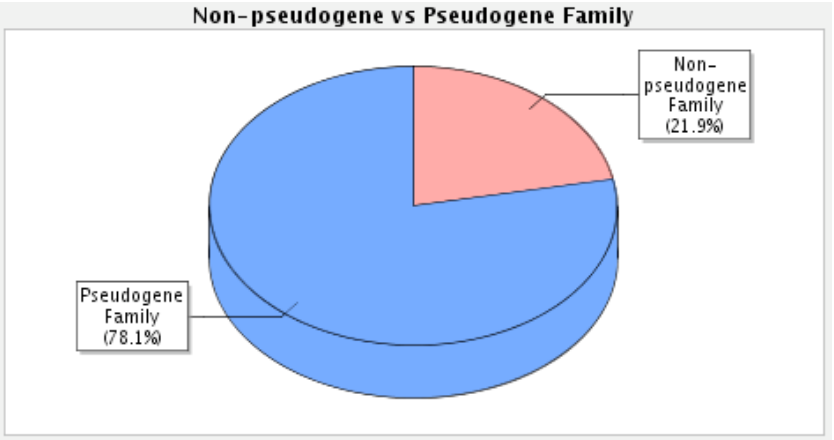


[Lam et al., NAR DB Issue (in press, '09)]

Pseudofam Statistics: Enrichment of pseudogenes within a family ("Living vs Dead")

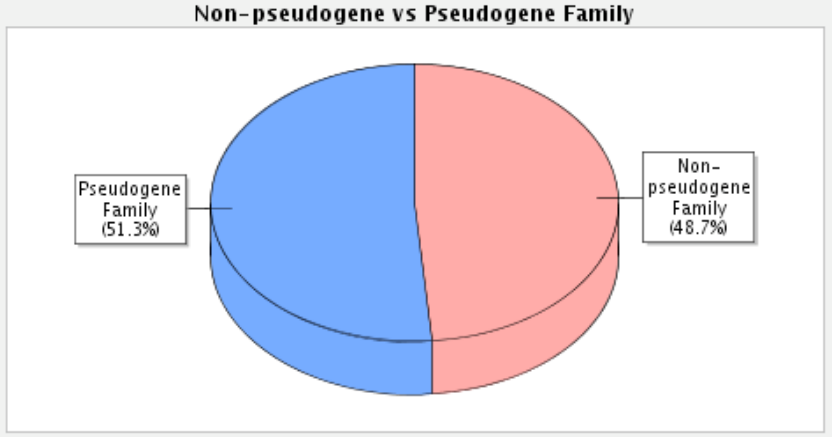
Total (10 Eukaryotes)

Protein Families:	3,820
Pseudogene Families:	2,985
Total Genes:	219,662
Total Parents:	26,679
Total Pseudogenes:	102,679
Pseudogene-to-gene Ratio:	0.47
Pseudogene-to-parent Ratio:	3.85
Parent-to-gene Ratio:	0.12

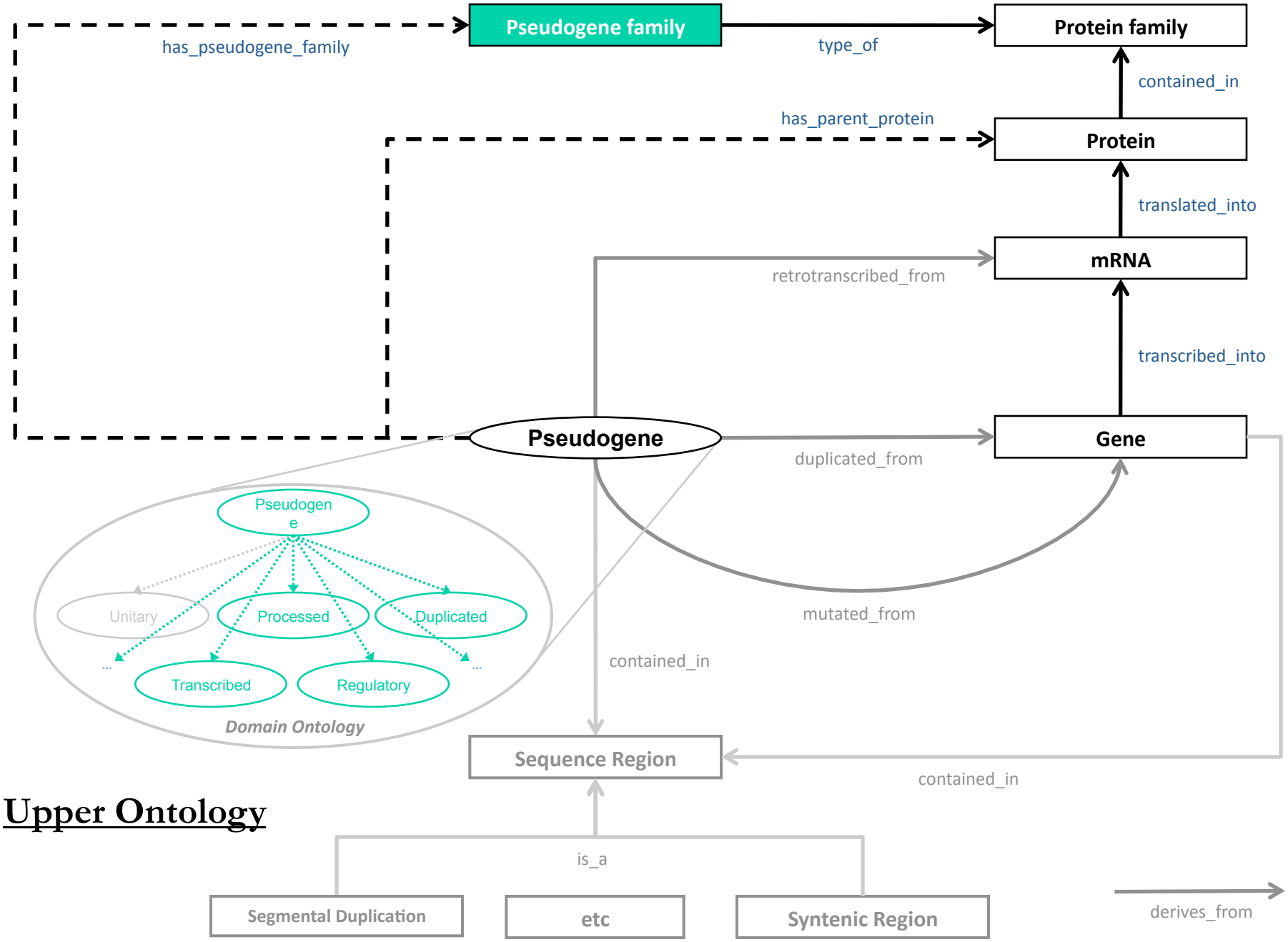


Human

Protein Families:	3,486
Pseudogene Families:	1,790
Total Genes:	34,686
Total Parents:	4,218
Total Pseudogenes:	12,534
Pseudogene-to-gene Ratio:	0.36
Pseudogene-to-parent Ratio:	2.97
Parent-to-gene Ratio:	0.12

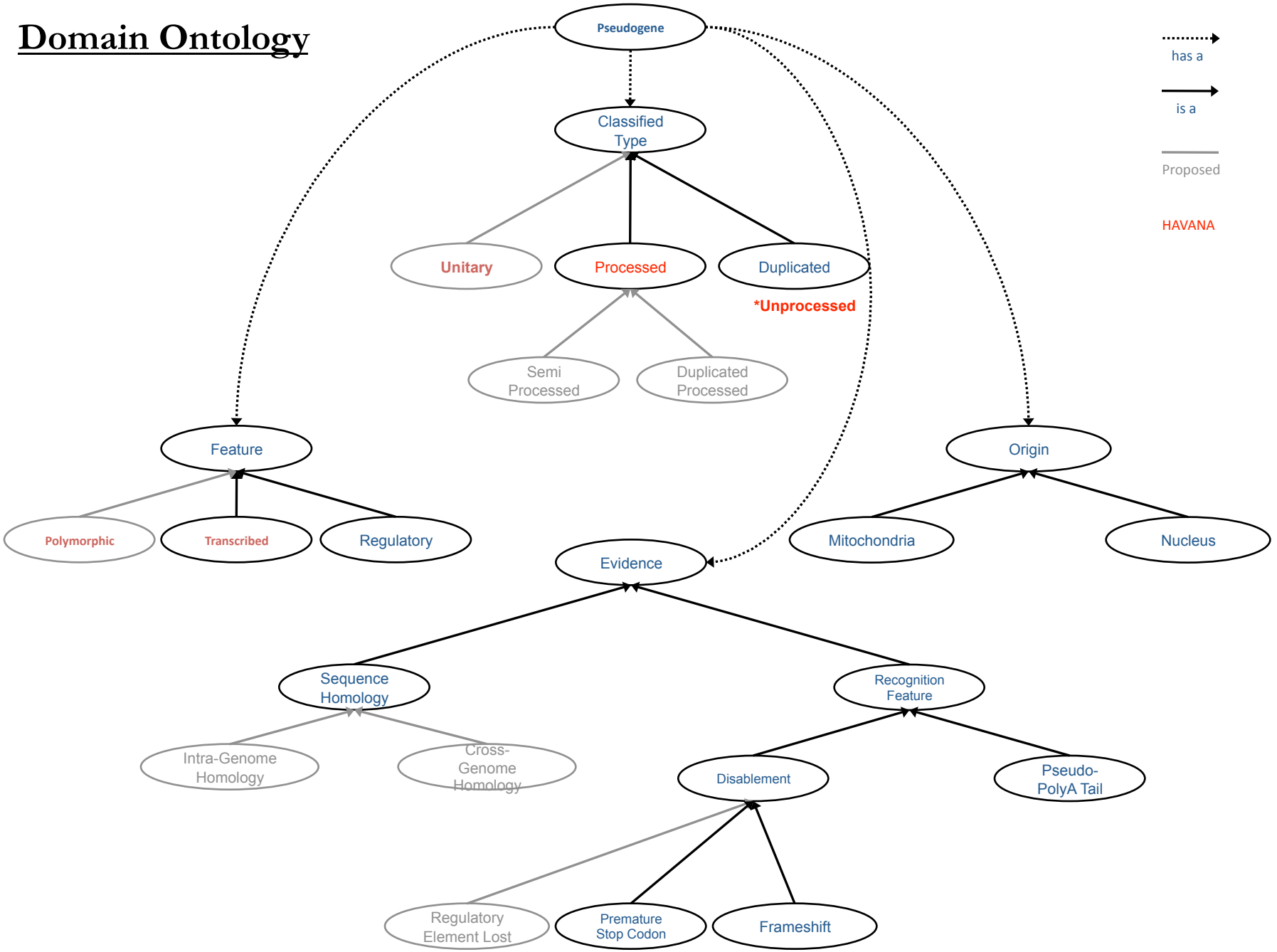


[Lam et al., NAR DB Issue (in press, '09)]



[Lam et al., NAR DB Issue (in press, '09)]

Domain Ontology



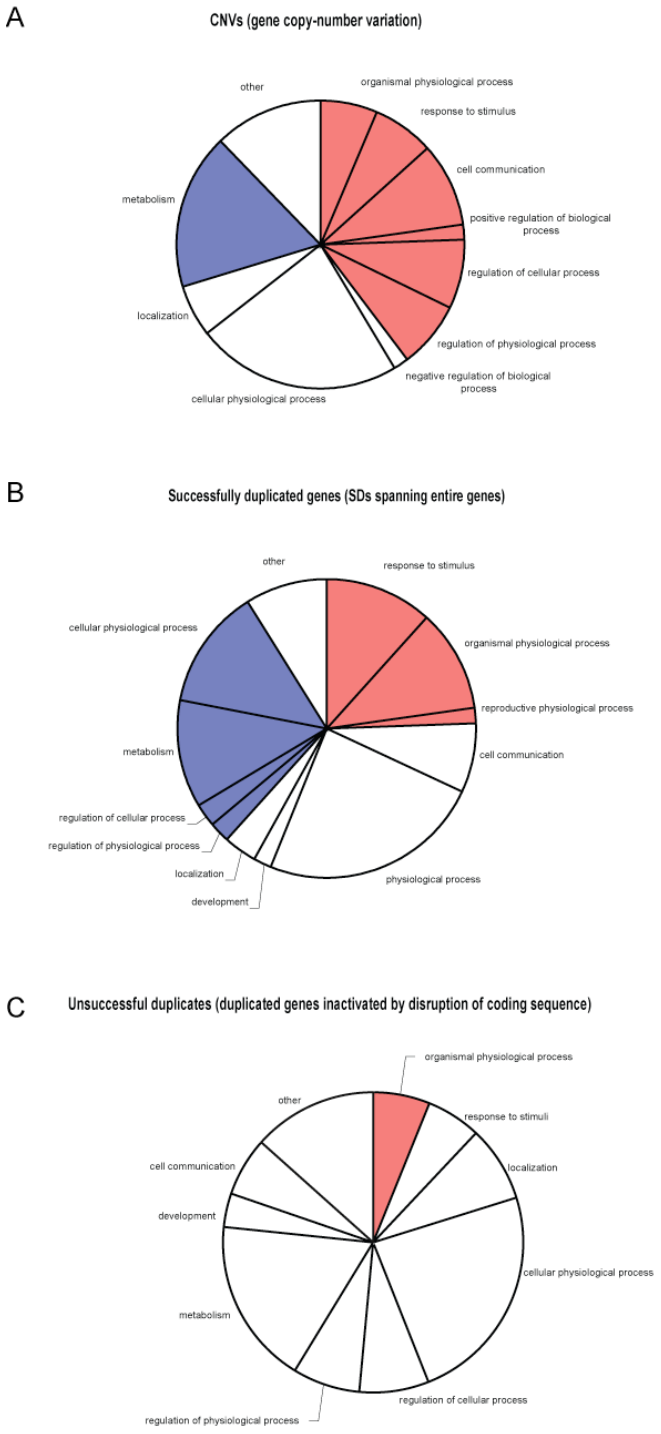
[Lam et al., NAR DB Issue (in press, '09)]

Relationship between pseudogenes and CNVs

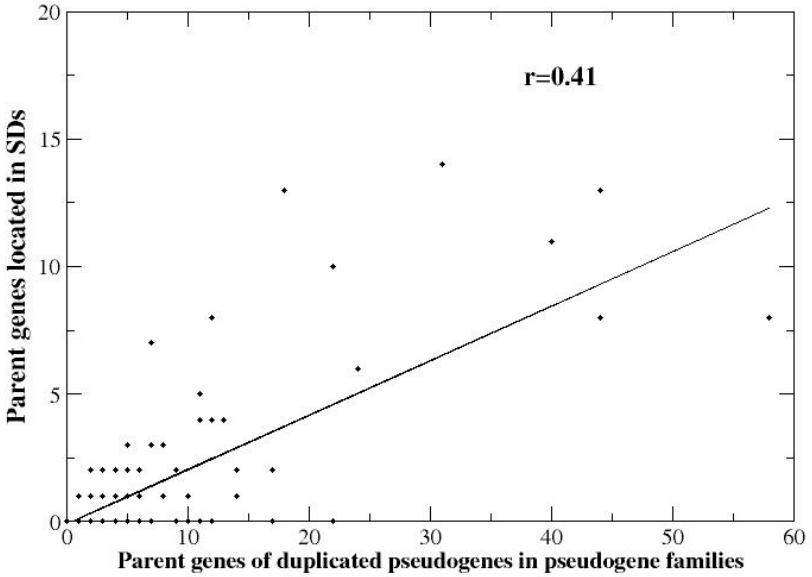
Association of SDs and CNVs with pseudogenes

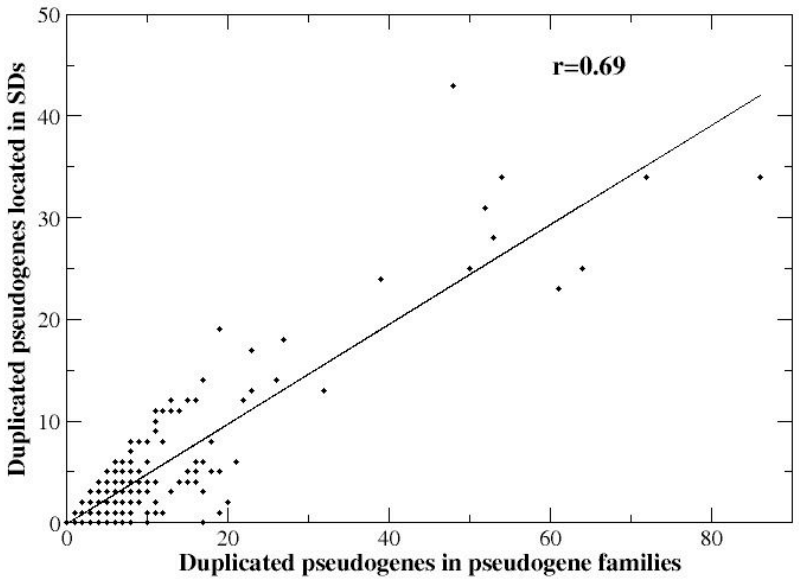
- CNVs are the raw form of variation producing duplicated elements
- Segmental Duplications (SDs) are fixed forms of CNVs/SVs. They give rise to duplicated genes and (eventually) protein protein families
- **Thus, we expect, duplicated pseudogenes (failed duplications) to occur in SDs.**
- CNVs and SDs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

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



Pseudogene families and Segmental Duplications (SDs)



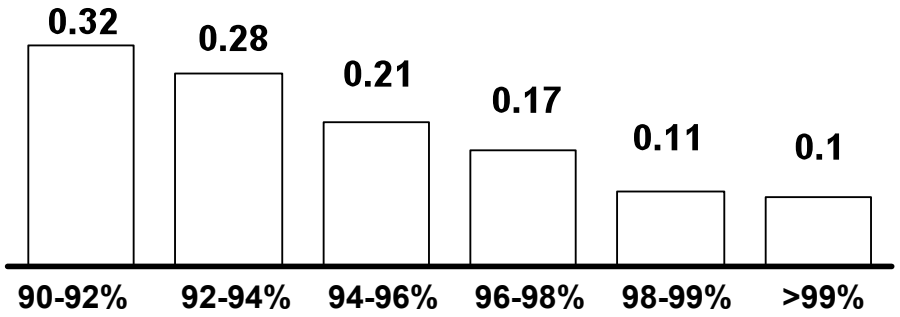


- **SDs comprise ~5% of the human genome but contain ~18% genes, 46% duplicated pgenes and 22% processed pgenes**
- Relative values of correlation coefficients in the plots above consistent with the observation that SDs contain more pgenes than parent genes

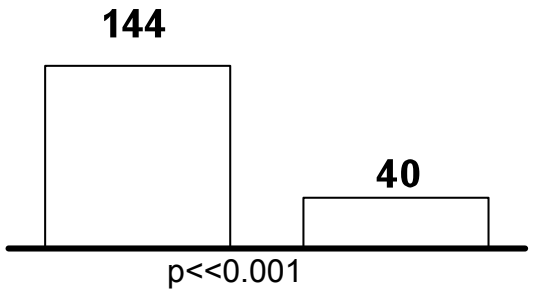
[Lam et al., NAR DB Issue (in press, '09)]

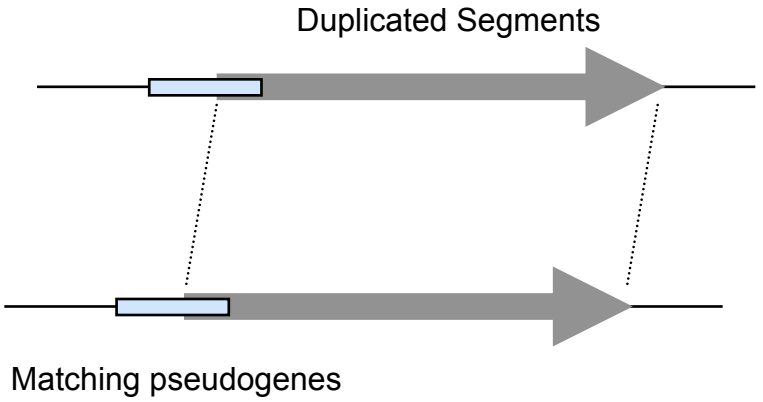
ANOTHER FUNCTION FOR PSEUDOGENES: SERVING AS REPEATS FOR MEDIATING NAHR

Processed pseudogene association with SDs by age



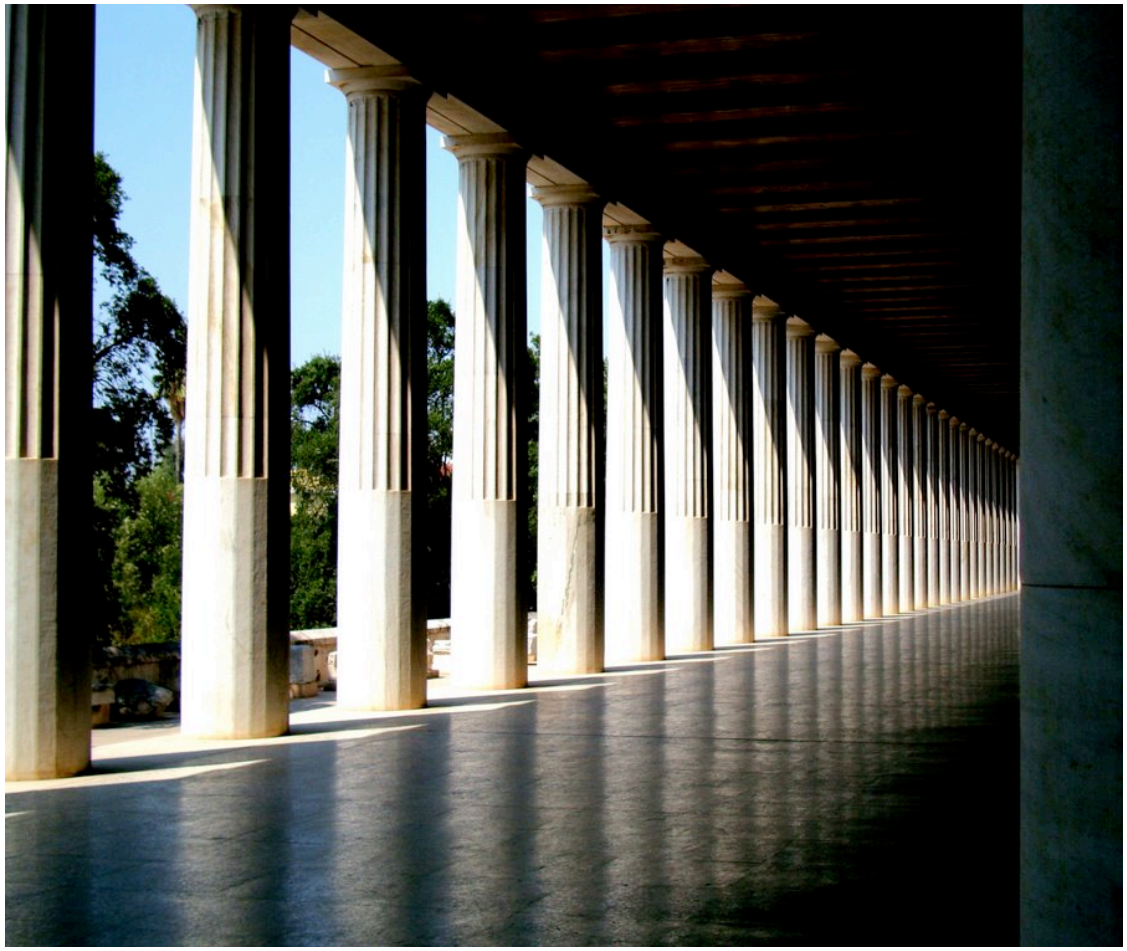
Processed pseudogenes at SD junctions



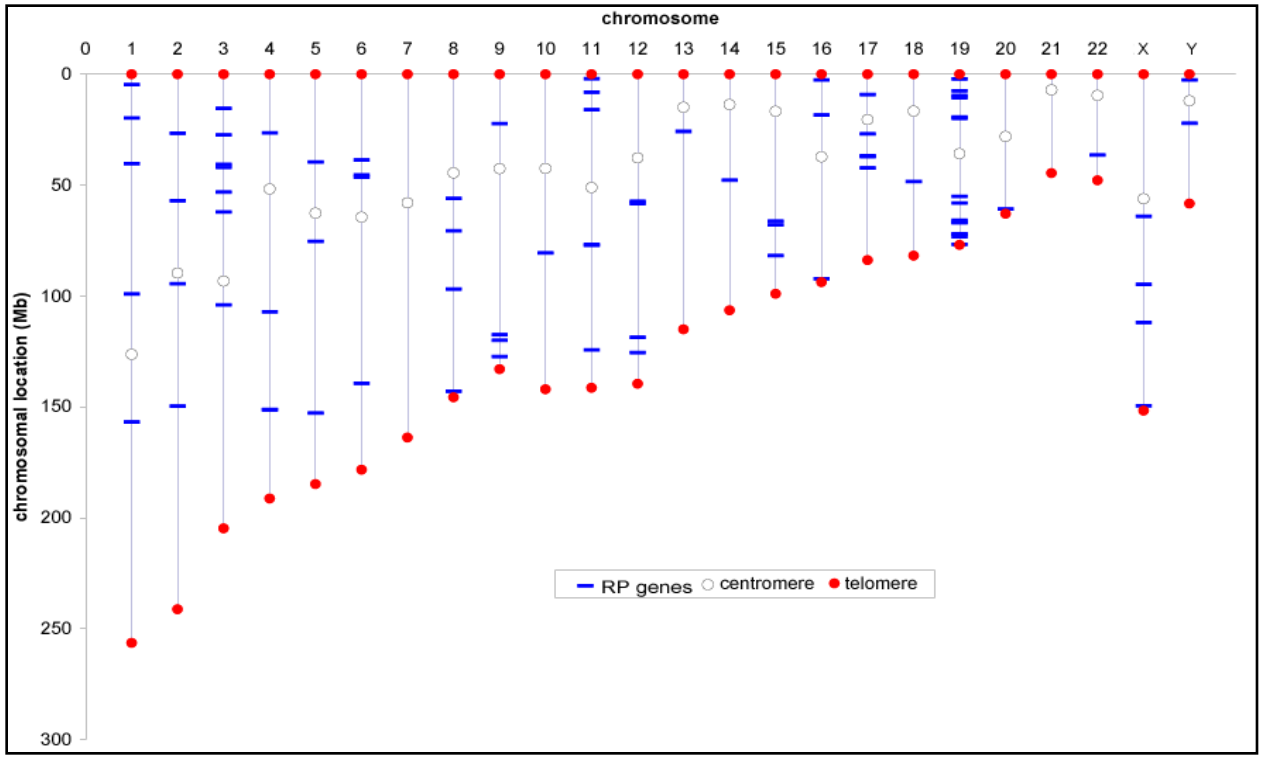


[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

Pseudogene Set #2: Ribosomal Protein Pseudogenes (Large Number)

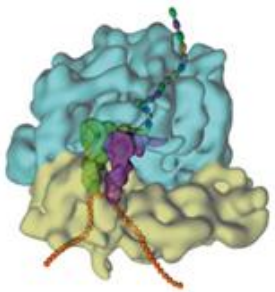


Human Ribosomal Proteins (RP)



79 Functional RP genes

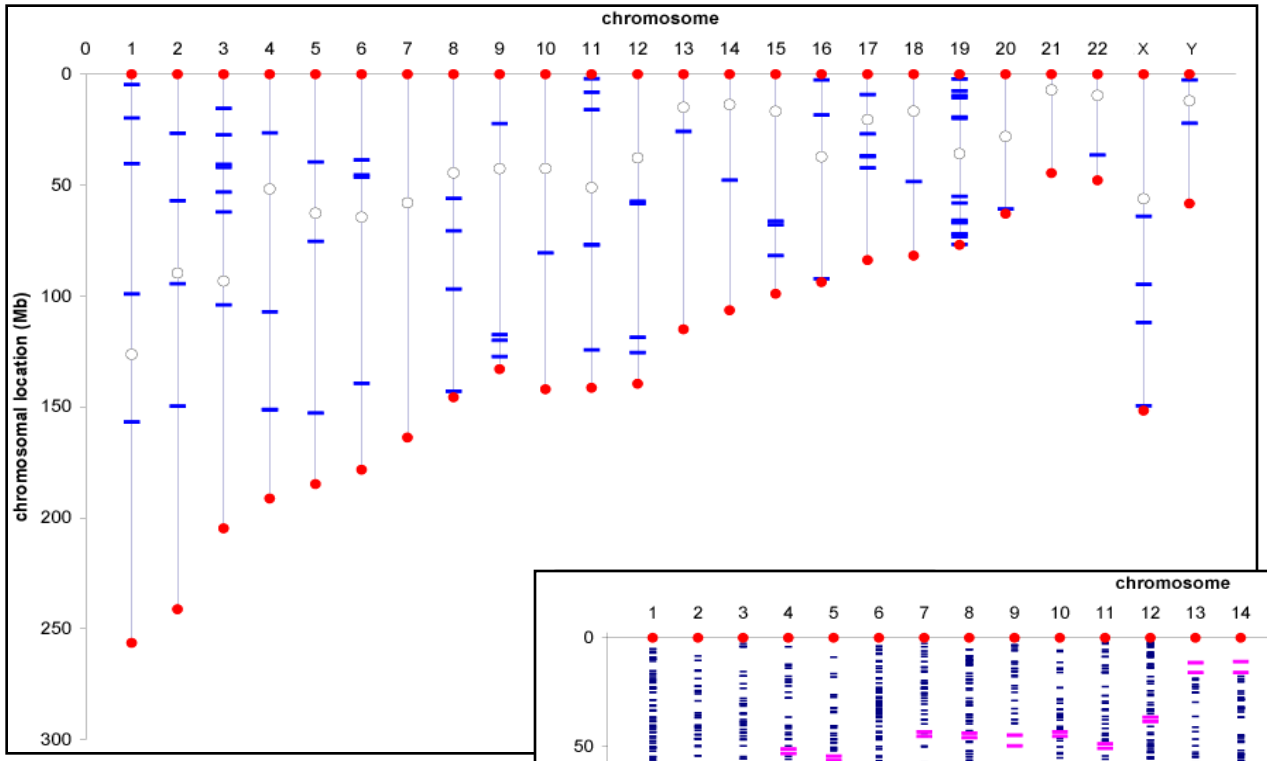
Nakao A, Yoshihama M, Kenmochi N: RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res 2004, 32:D168-170.*

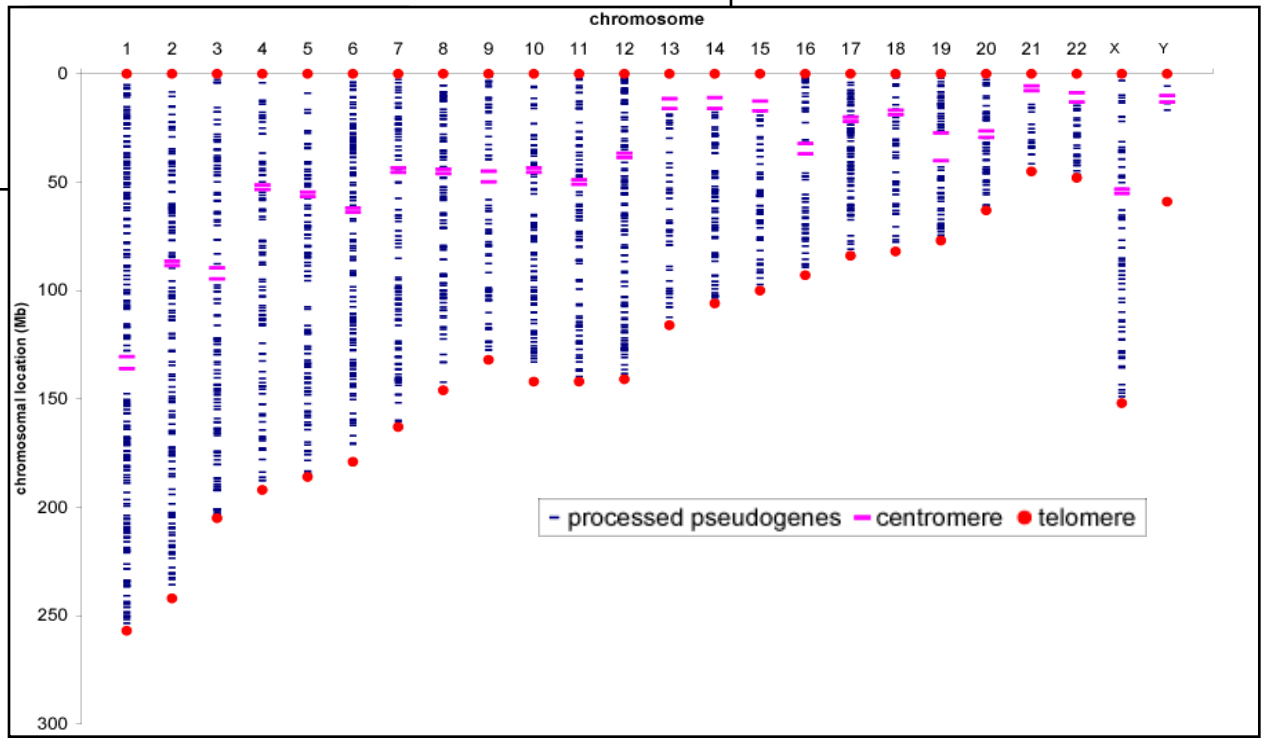


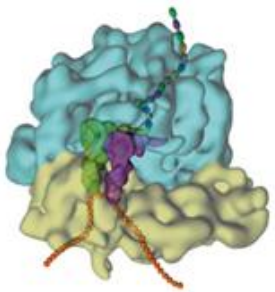
[Zhang *et al*. *Genome Research*, 2002]

Human Ribosomal Proteins (RP)

79 Functional RP genes







~ 2000 RP Ψgenes

[Zhang *et al*. *Genome Research*, 2002]

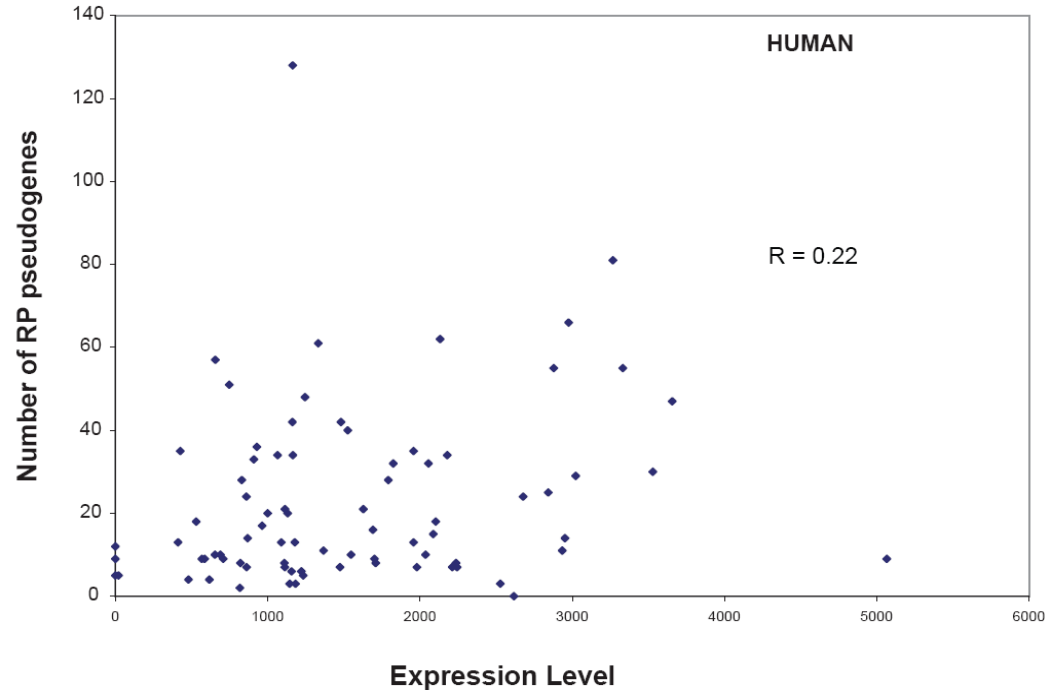
Number of RP pseudogenes

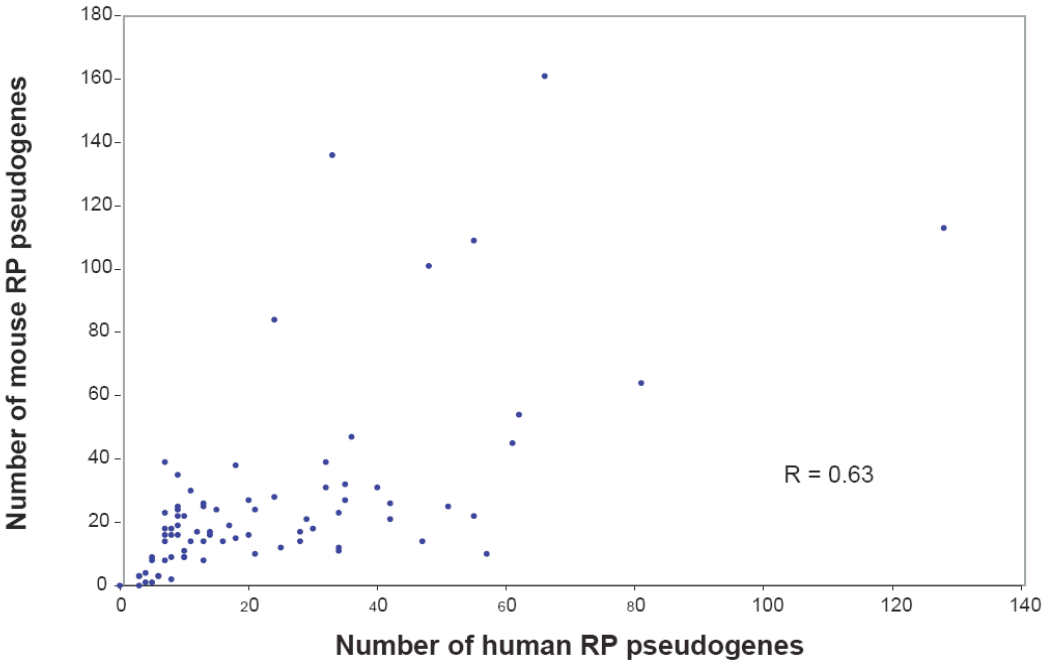
(identified by pipeline)

Organism	Processed	Fragments	Low confidence
Human	1822	218	212
Chimp	1462	219	160
Mouse	2092	326	413
Rat	2848	343	450

RP pseudogenes constitute the largest family of pseudogenes. Almost all are processed: There are ~90 clearly duplicated ones in the human genome

[Balasubramanian et al., *Genome Biol.* ('09)]

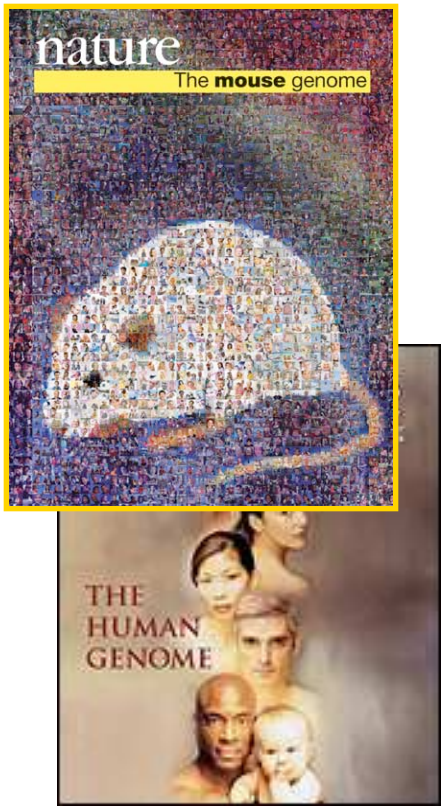


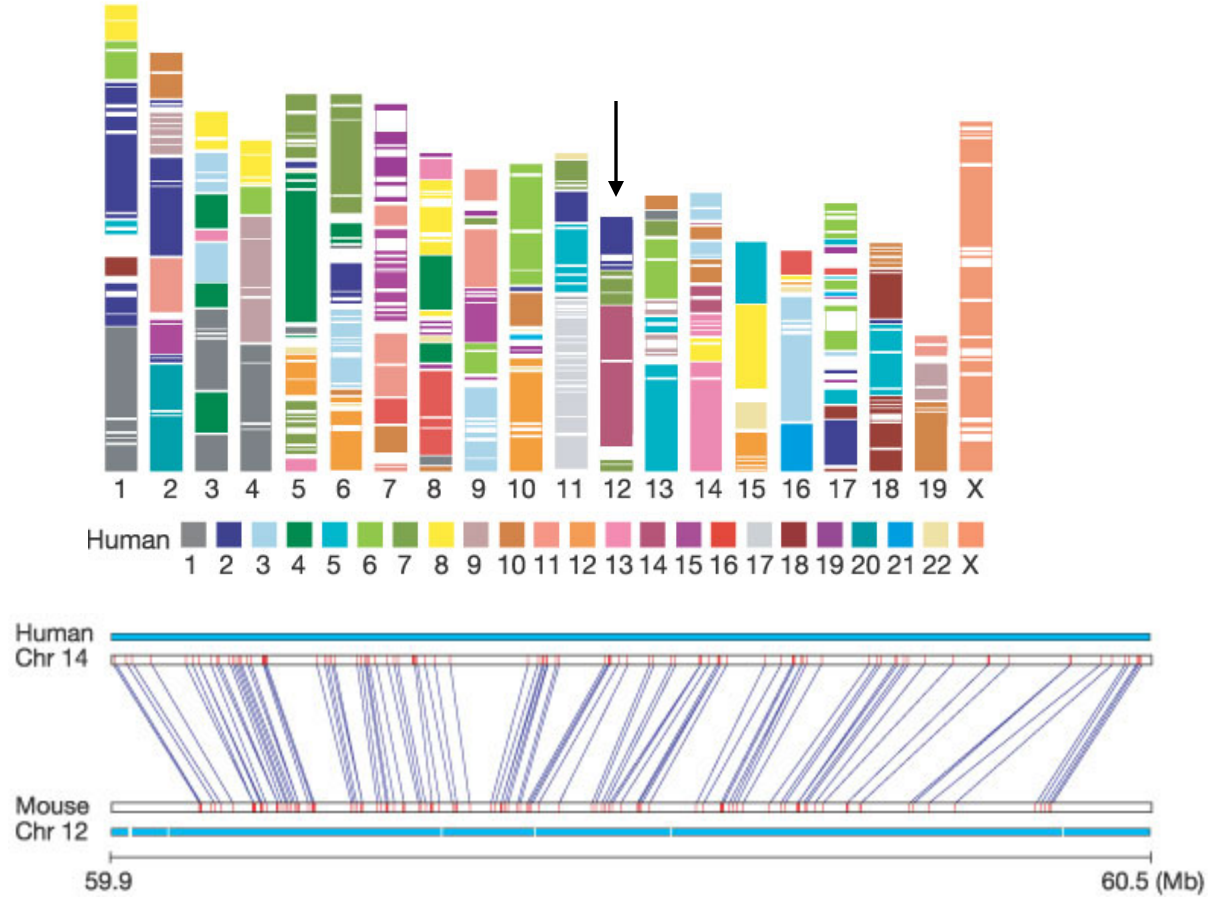


Number of each type of human ribosomal protein processed pseudogenes appears unrelated to expression level or to number in mouse

[Balasubramanian et al., *Genome Biol.* ('09)]

The Synteny Between Mouse and Man

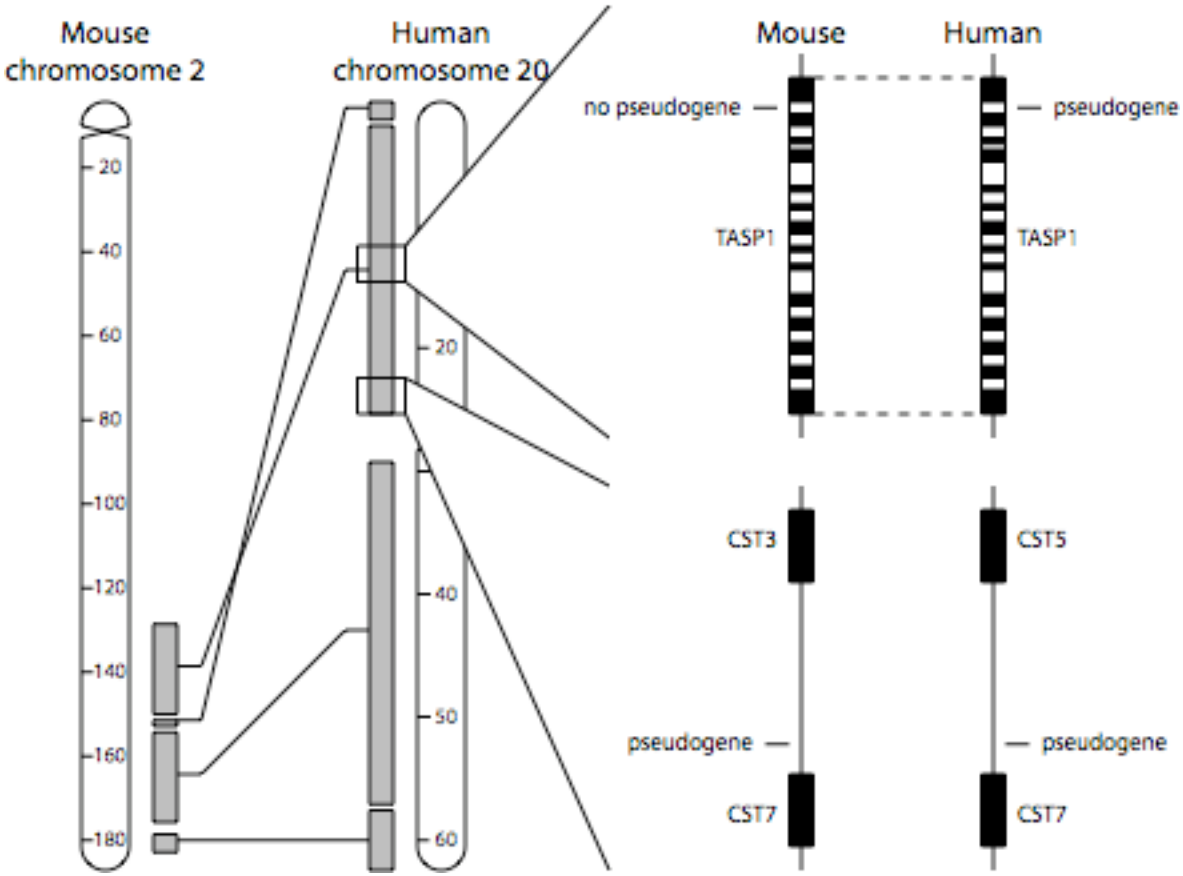




- About 90% of the mouse and human genomes are in syntenic blocks.

[Waterston *et al.* *Nature* 2002]

Using Synteny to Improve Annotation of RP pgenes



Synteny derived based on local gene orthology

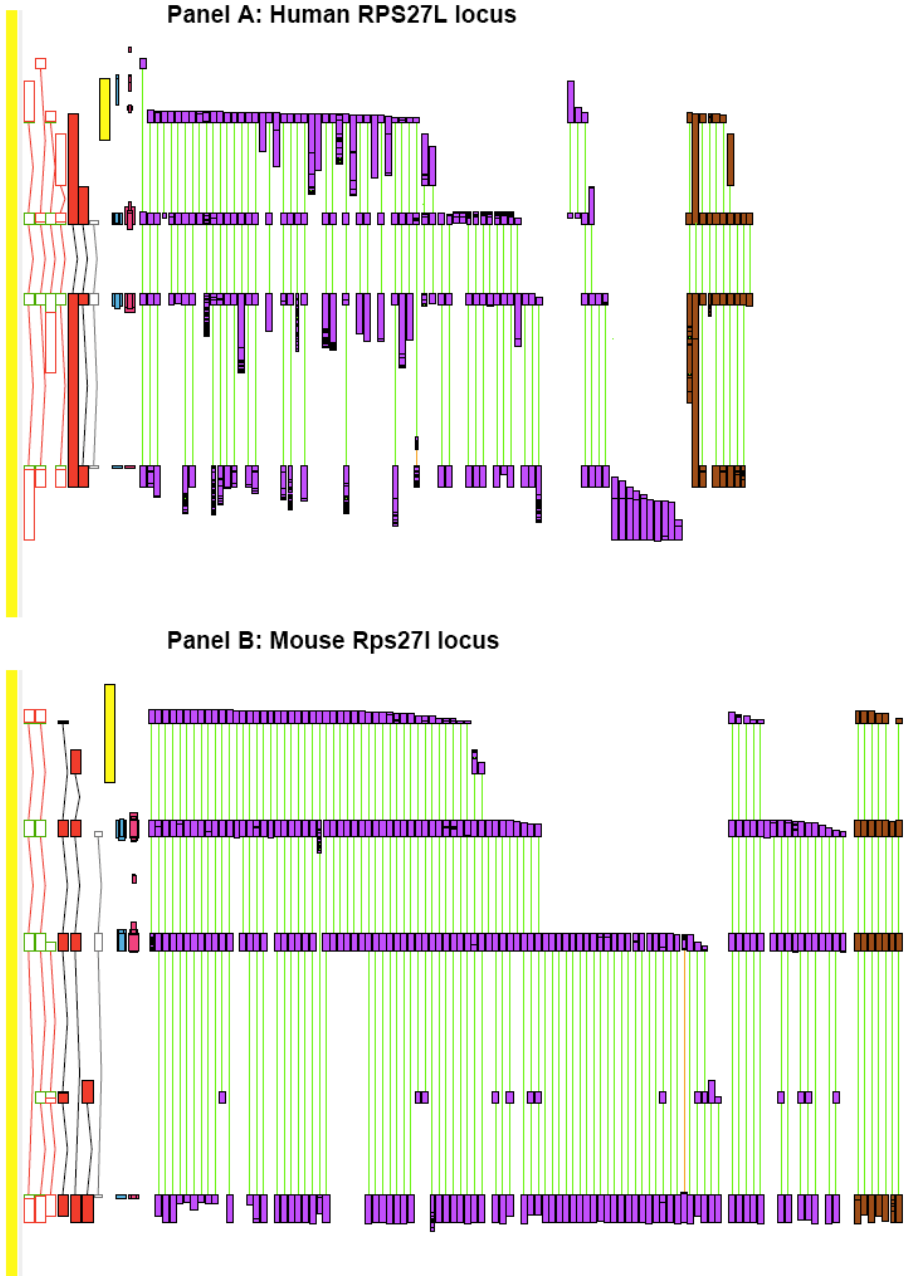
[Balasubramanian et al., *Genome Biol.* ('09)]

Syntenic proc pseudogenes

Species1- Species2	Number of syntenic pgenes
Human-chimp	1282
Human-mouse	6
Human-rat	11
Rat-mouse	394

[Balasubramanian et al., *Genome Biol.* ('09)]

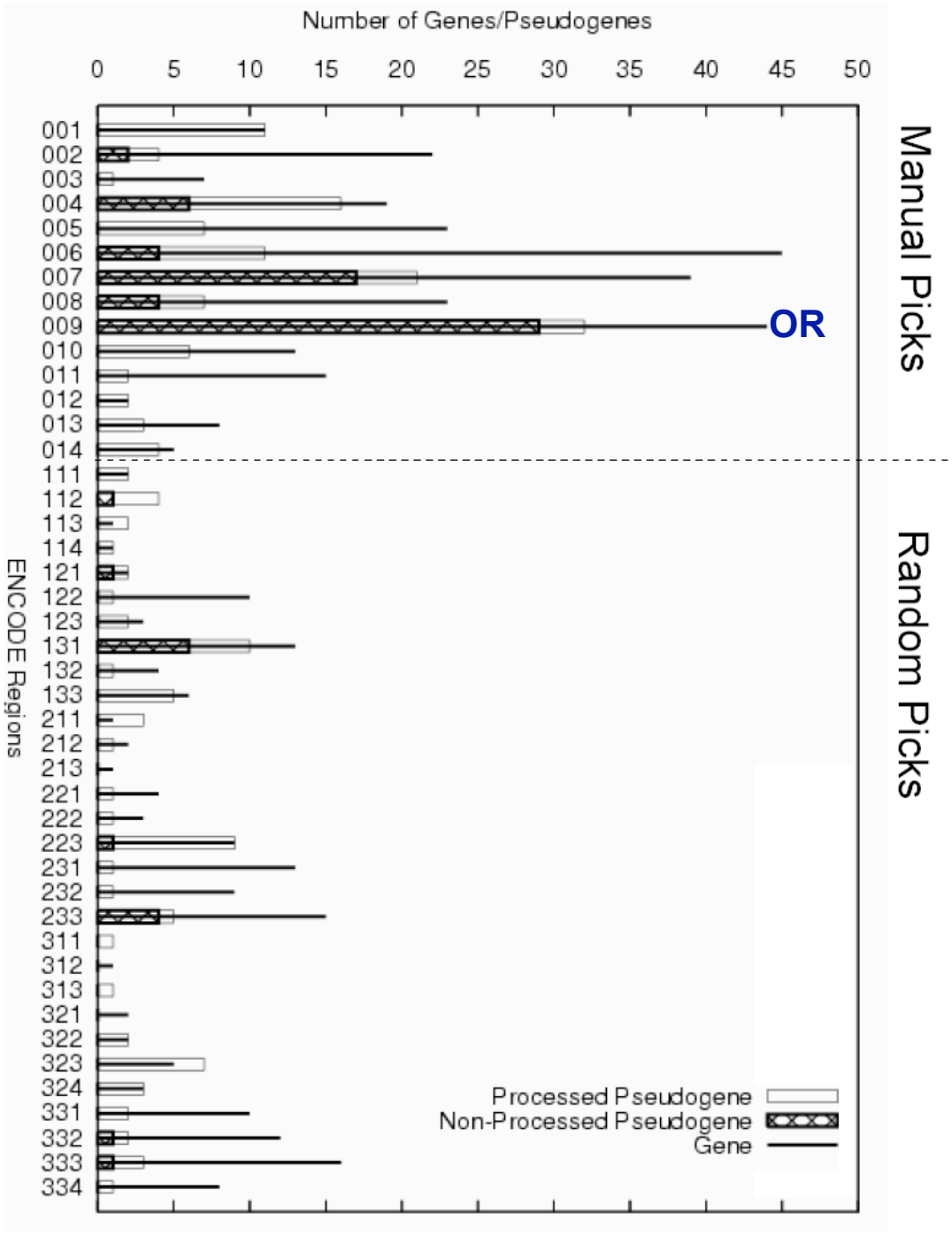
A human-mouse syntenic pgene, likely to be coding



[Balasubramanian et al., *Genome Biol.* ('09)]

Nakao A, Yoshihama M, Kenmochi N: RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res 2004, 32:D168-170.*

ENCODE Pilot Pseudogenes: Integration of Different Types of Annotation



Overall Results: Regional Distribution

201 pseudogenes
77 non-processed
124 processed

Zheng et al. (2007) Gen. Res.

browser
+
pseudogene.org/ENCODE

Vast Amounts of Different Data Types to Integrate in pilot ENCODE

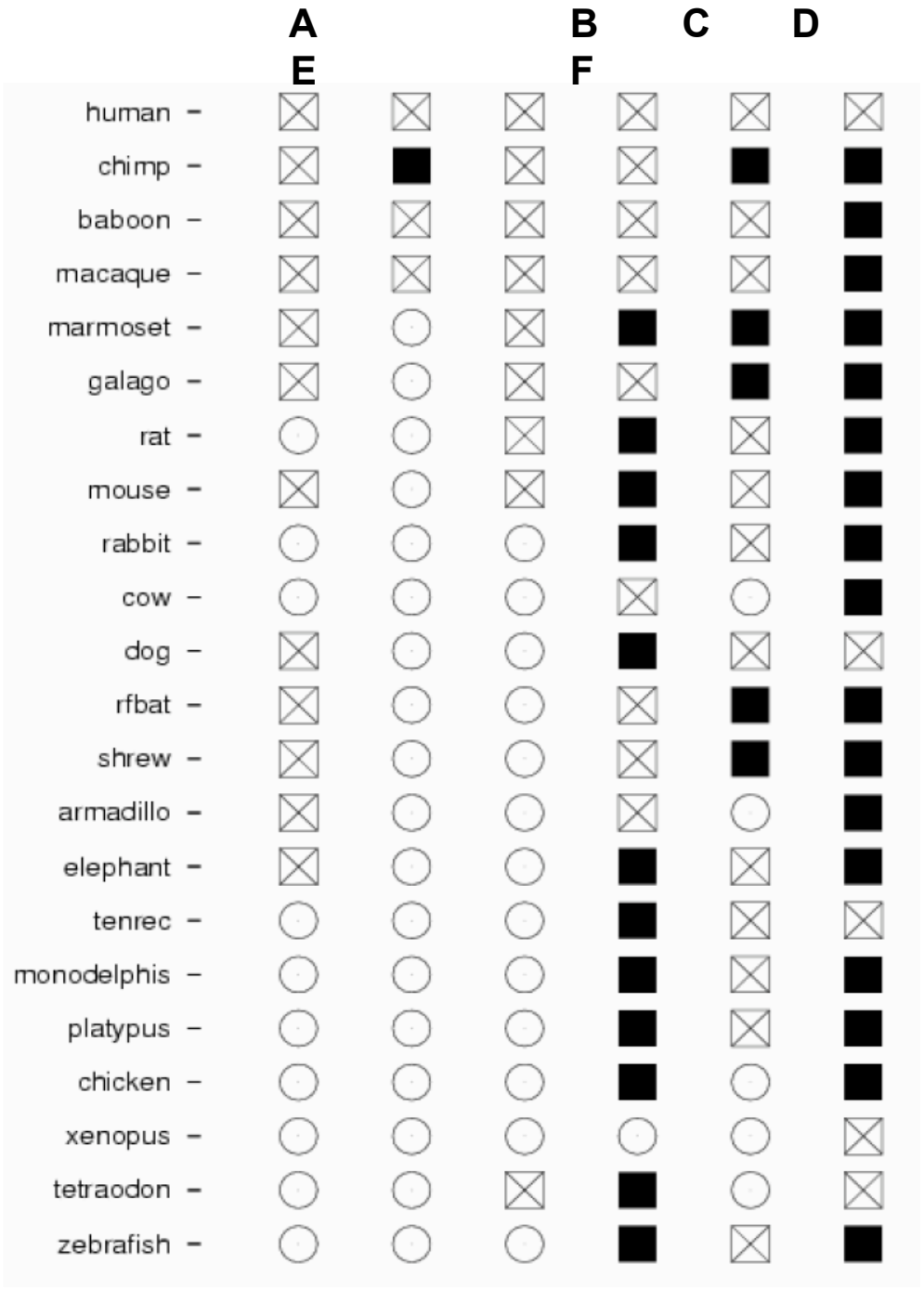
- Determining experimental signals for biochemical activity across each base of genome

- Large-scale sequence comparison in relation to the human genome

Feature Class	Expt. Tech.	Numb. Expt. Data Pts.
Transcription	Tiling array, Integrated annotation	**63,348,656**
5′ Ends of transcripts	Tag sequencing	**864,964**
Histone modifications	Tiling array	**4,401,291**
Chromatin structure	QT-PCR, Tiling array	**15,318,324**
Sequence-specific factors	Tiling array, tag sequencing, Promoter assays	**324,846,018**
Replication	Tiling array	**14,735,740**
Computational analysis	Computational methods	NA
Comparative sequence analysis	Genomic sequencing, multi- sequence alignments, computational analyses	NA
Polymorphisms	Resequencing, copy number variation	NA

[ENCODE Consortium, ***Nature*** 447, 2007]

representative pseudogenes drawn from 201 total



History of Pseudogene Preservation

Based on alignment from ENCODE MSA group

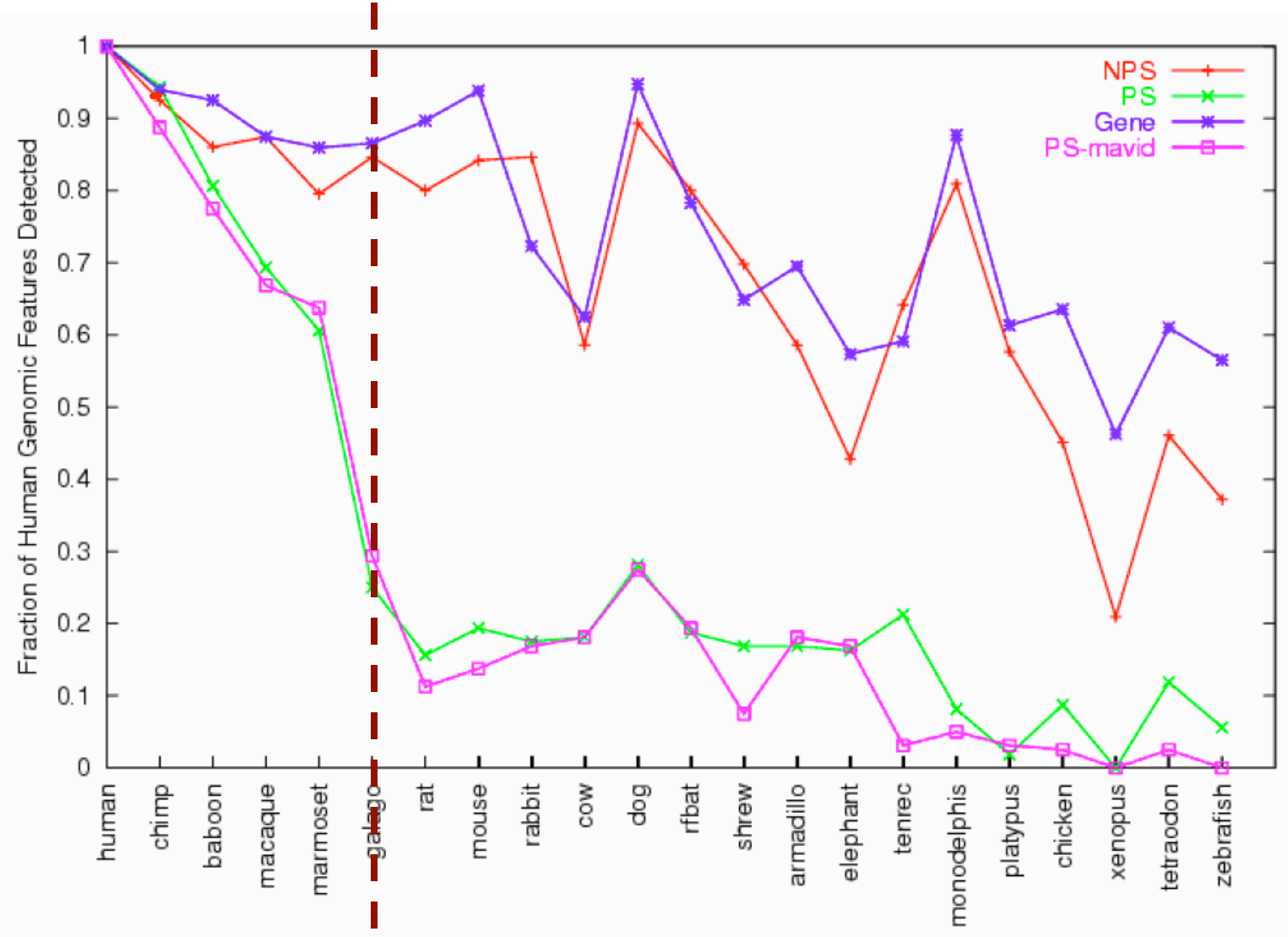
Zheng et al. (2007) Gen. Res.

Absent

Present with Disablement

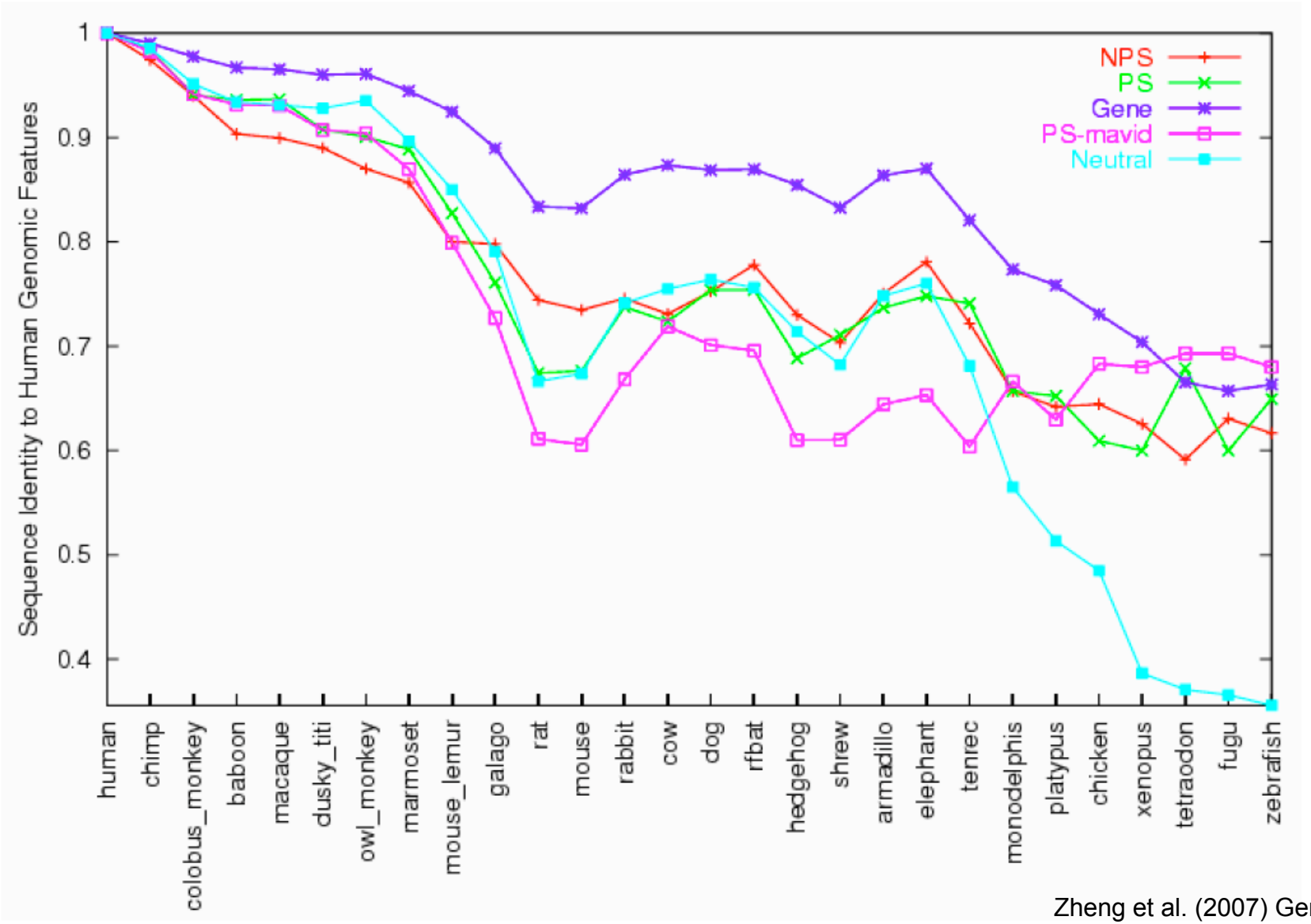
Present without Disablement

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



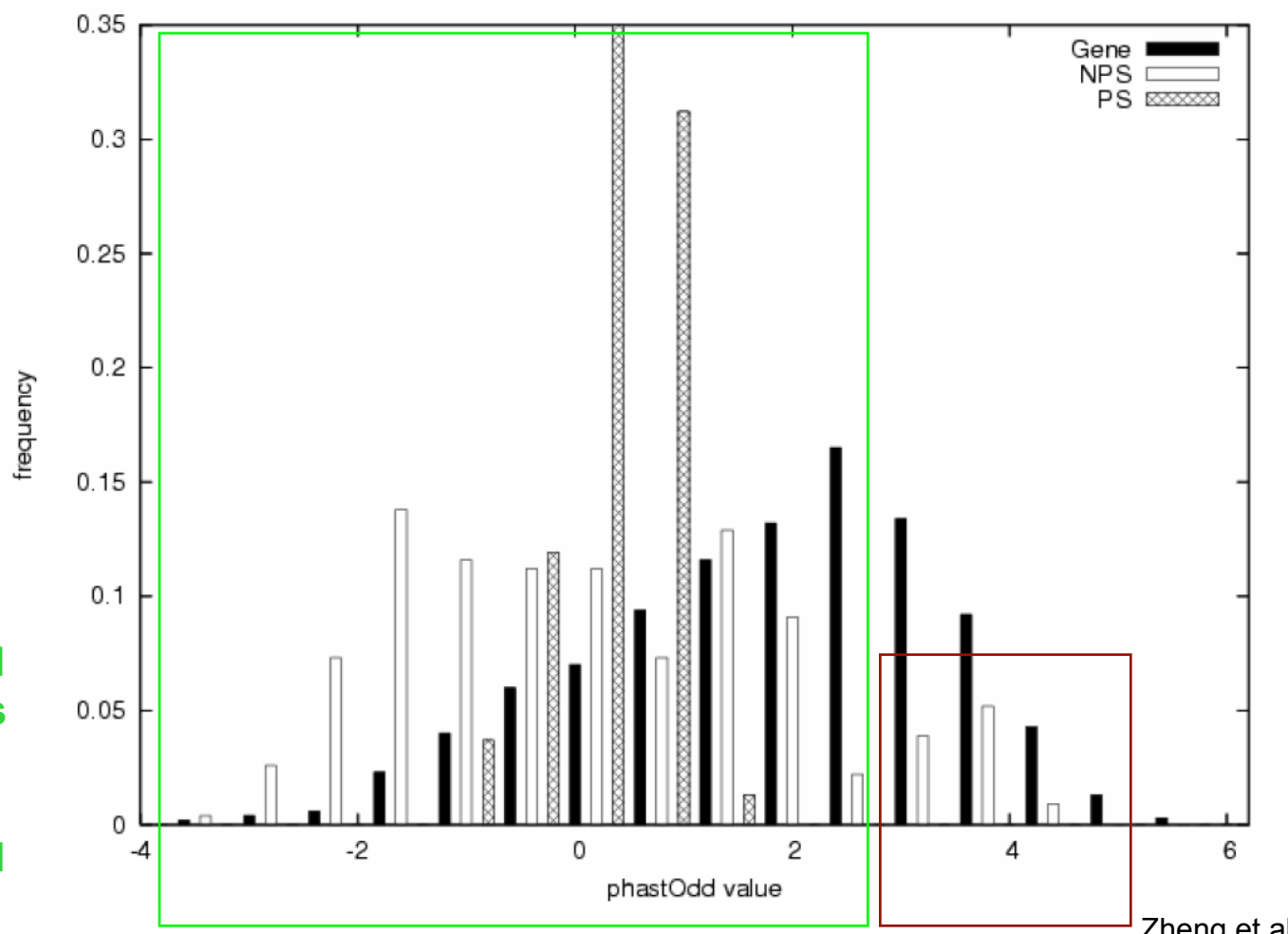
Zheng et al. (2007) Gen. Res.

Sequence Decay of Pseudogenes, Approximately Neutral



Zheng et al. (2007) Gen. Res.

Using phastOdd value to examine neutral evolution of pseudogenes

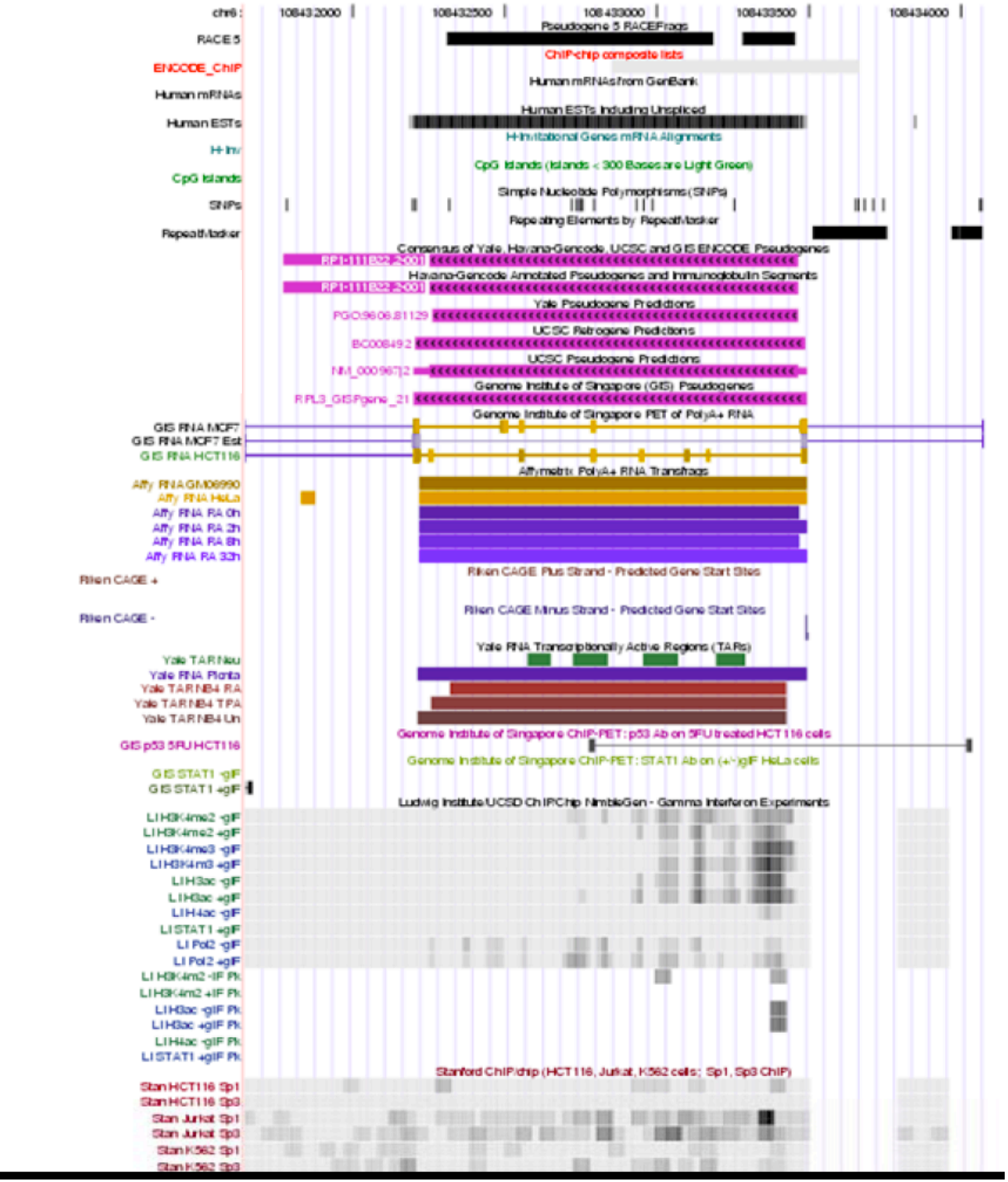


most good candidates for studying mutational processes

a few non-proc. ψG under constraint

Zheng et al. (2007) Gen. Res.

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.

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

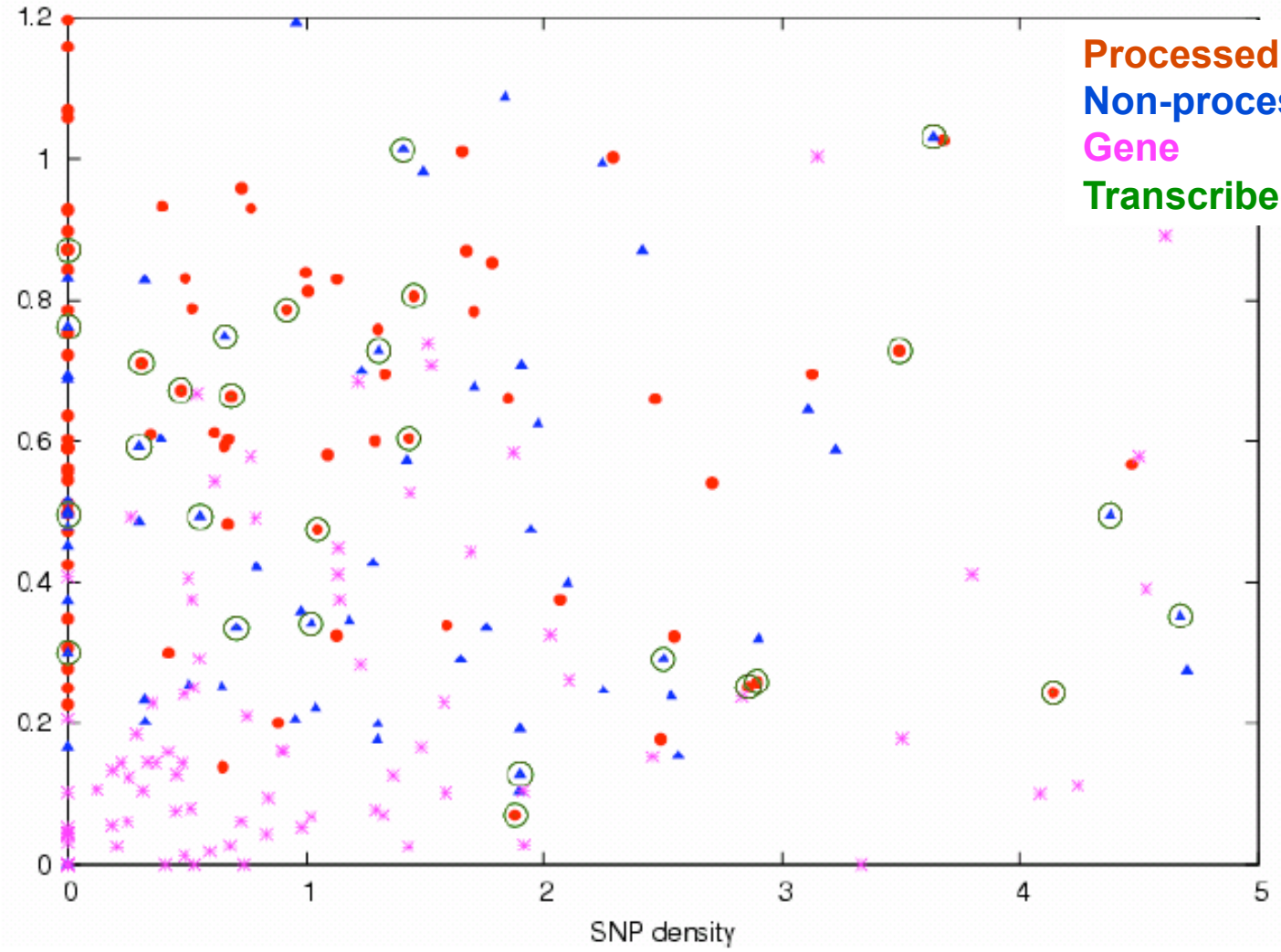
Excluding TARs (due to cross-hyb issues)

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.

Integrating Transcriptional Evidence with Gene Annotation and Sequence Constraints



Processed pseudogene
Non-processed pseudogene
Gene
Transcribed ○

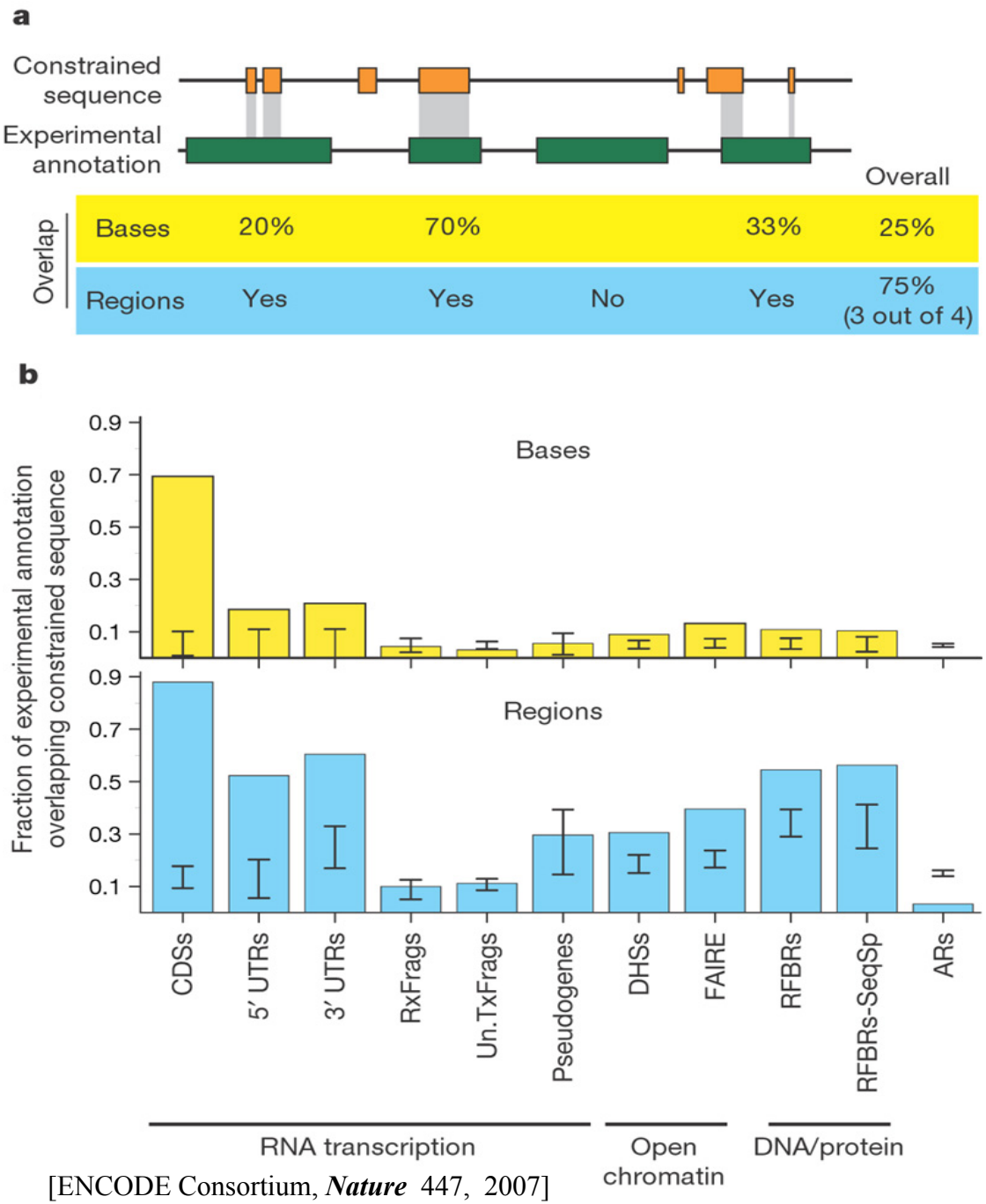
Avg. Integration over many instances

- 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.

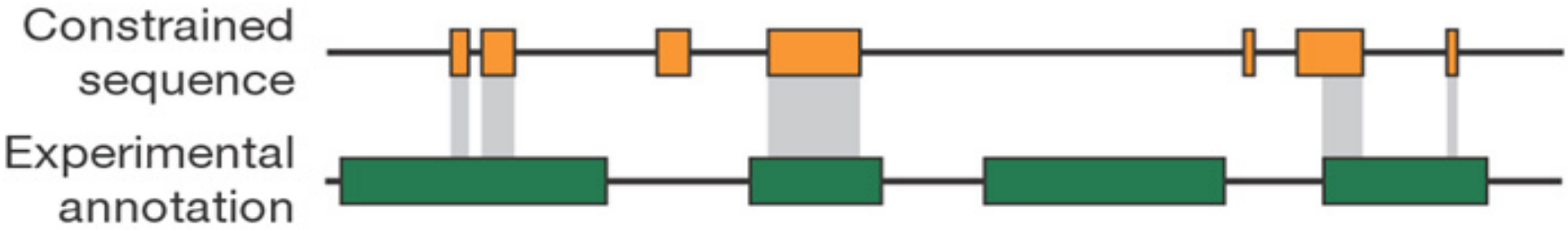
Biochemically Active Regions Don't all Appear to be Under Constraint



[ENCODE Consortium, ***Nature*** 447, 2007]

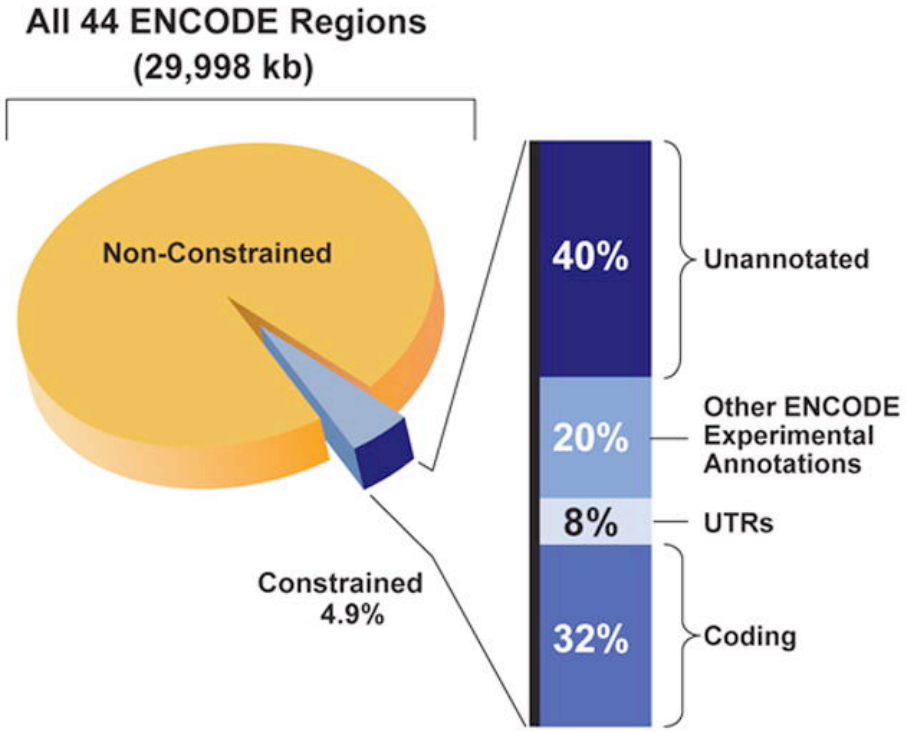
- Integrating & averaging results over larger and larger sets
- Comparison of integrated quantities

Grand Summary: Biochemical Activity vs. Sequence Constraints



- Not all constrained sequence annotated in some fashion
- Exactly how things are defined in terms of overlap?

- "At the outset of the ENCODE Project, many believed that the broad collection of experimental data would nicely dovetail with the detailed evolutionary information derived from comparing multiple mammalian sequences to provide a neat 'dictionary' of conserved genomic elements, each with a growing annotation about their biochemical function(s). In one sense, this was achieved; the majority of constrained bases in the ENCODE regions are now associated with at least some experimentally-derived information about function. However, we have also encountered a remarkable excess of unconstrained experimentally-identified functional elements, and these cannot be dismissed for technical reasons. This is perhaps the biggest surprise of the pilot phase of the ENCODE Project, and suggests that we take a more 'neutral' view of many of the functions conferred by the genome. "



[ENCODE Consortium, ***Nature*** 447, 2007]

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

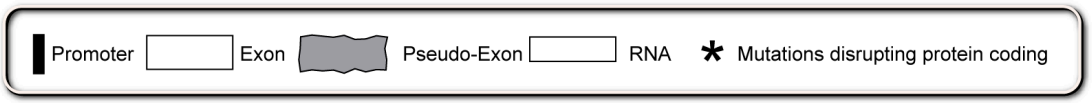
Genes & Pseudogenes

(a) Functional Gene

Ambiguous Cases

(b) Dead Pseudogene

Zheng & Gerstein, TIG (2007)

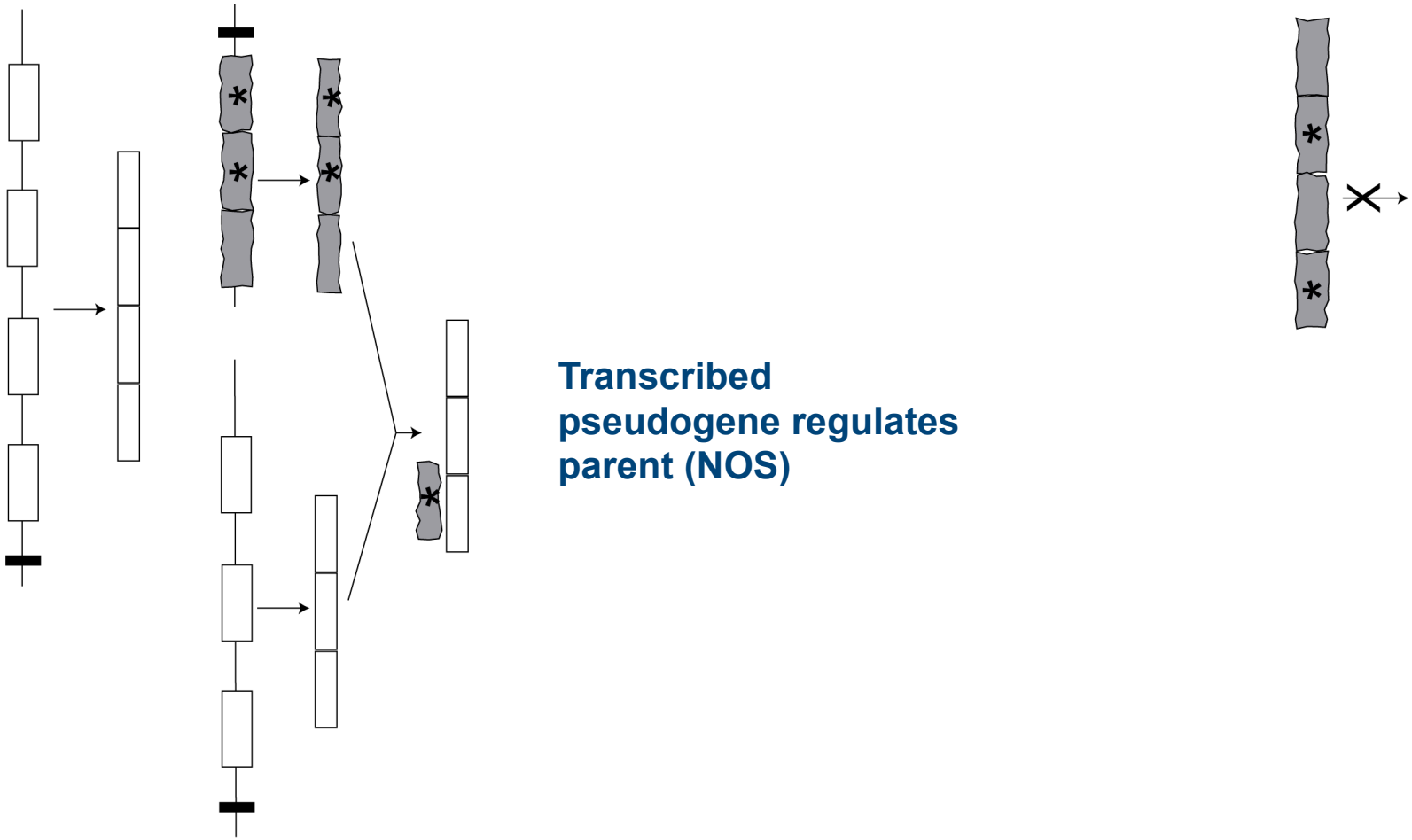


Genes or Pseudogenes?

(a) Functional Gene

Ambiguous Cases

(b) Dead Pseudogene

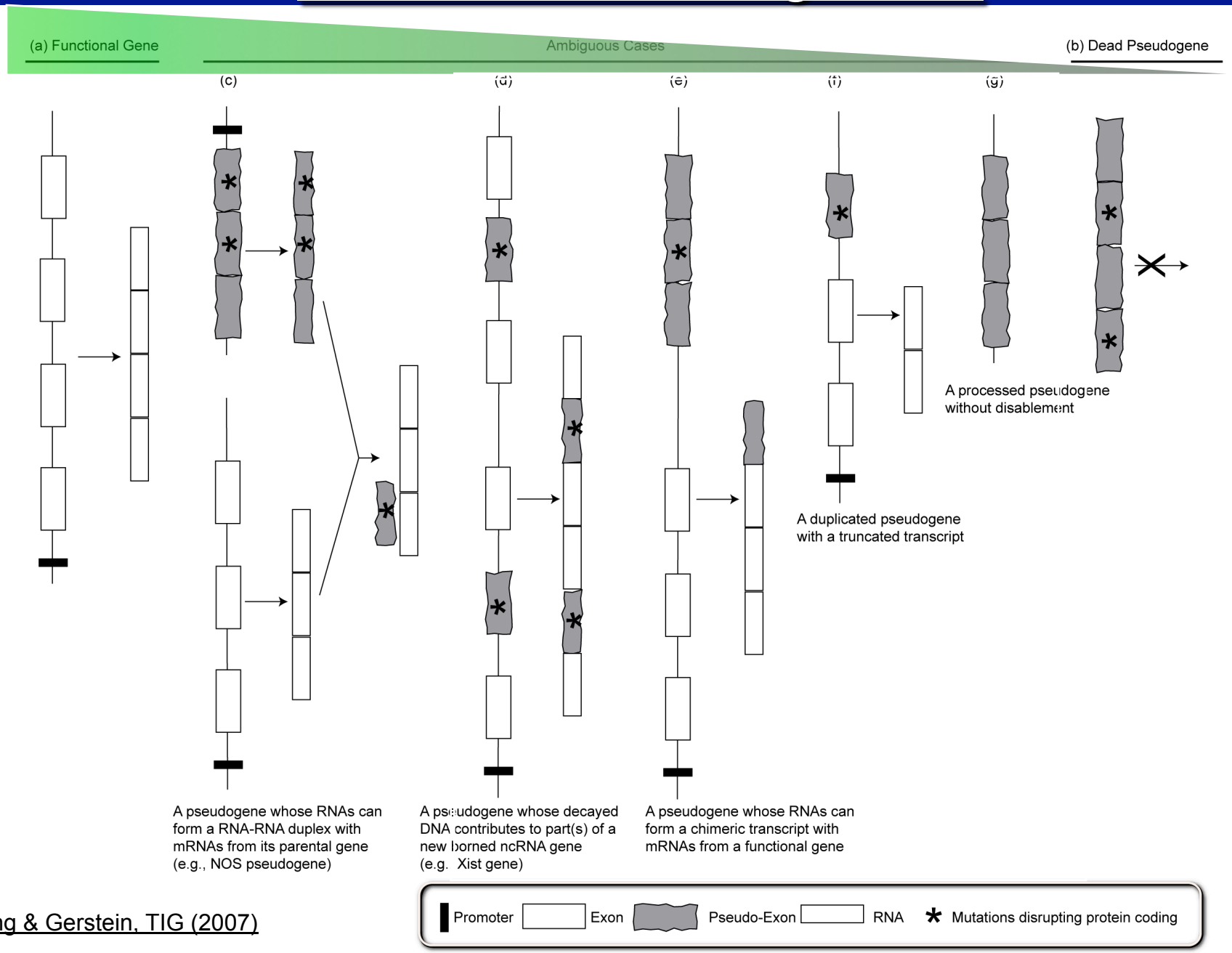


Transcribed pseudogene regulates parent (NOS)

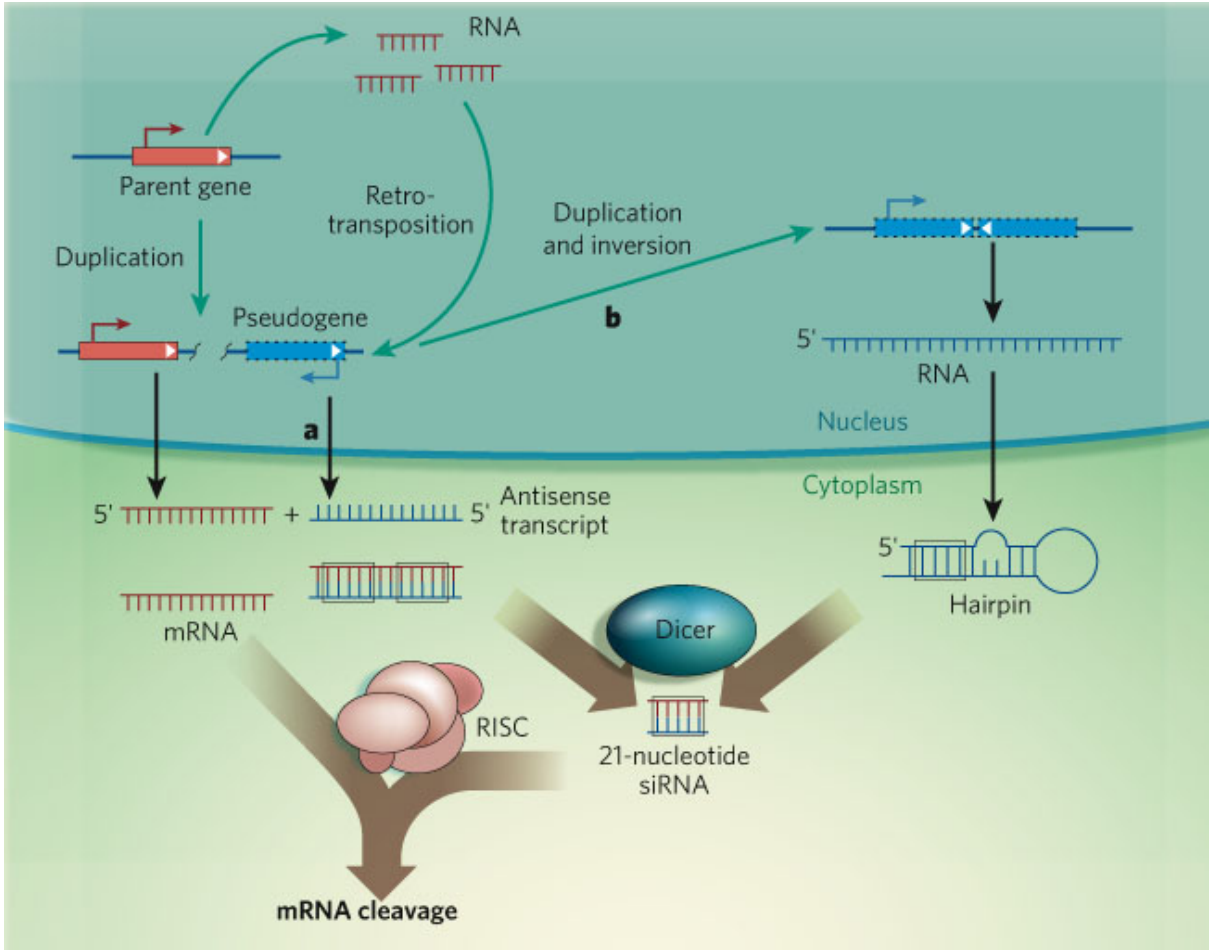
Zheng & Gerstein, TIG (2007)

Promoter | Exon | Pseudo-Exon | RNA | * Mutations disrupting protein coding

Genes or Pseudogenes?



Zheng & Gerstein, TIG (2007)



What are Active Pseudogenes Doing?

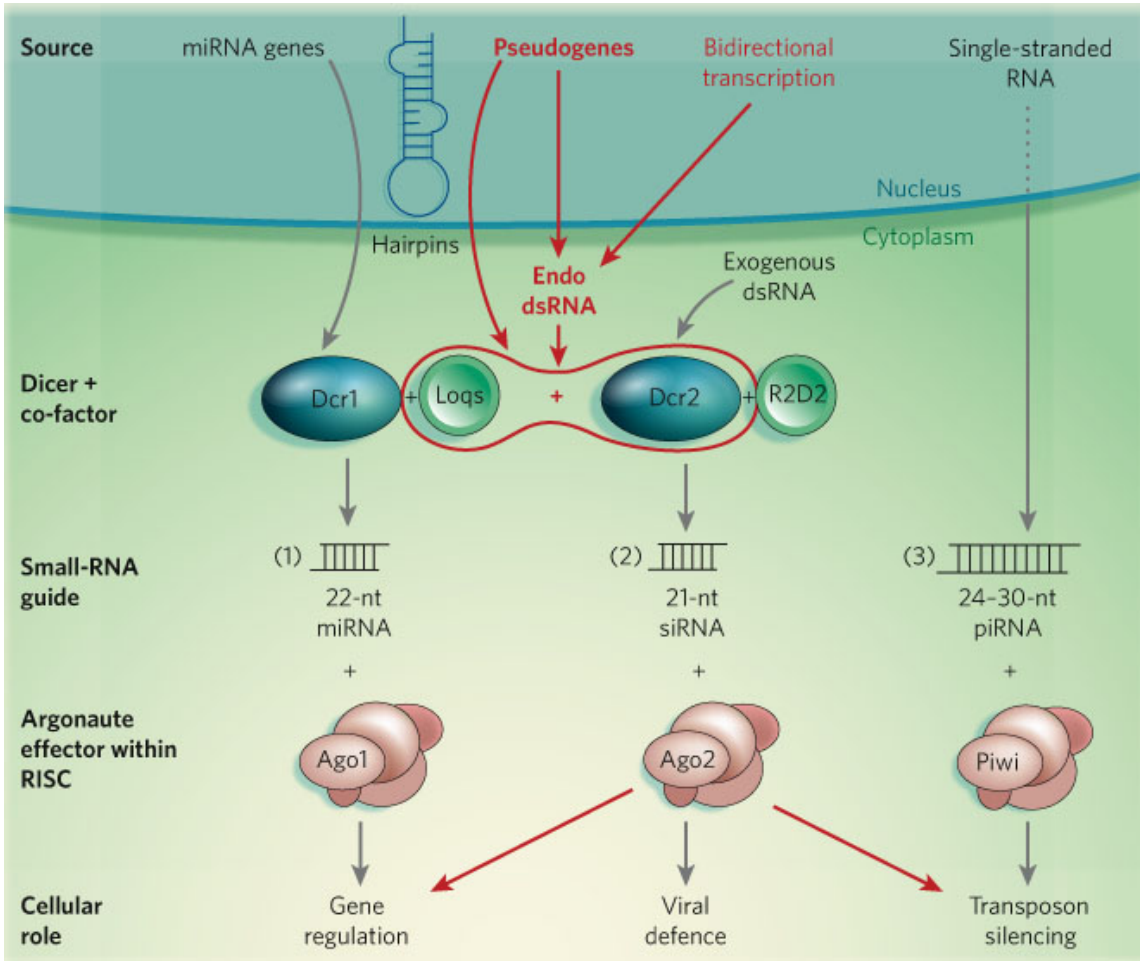
Potential for Gene Regulation via endo-siRNA

Recent Discoveries in Mouse & 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).

[Sasidharan & Gerstein, Nature ('08)]

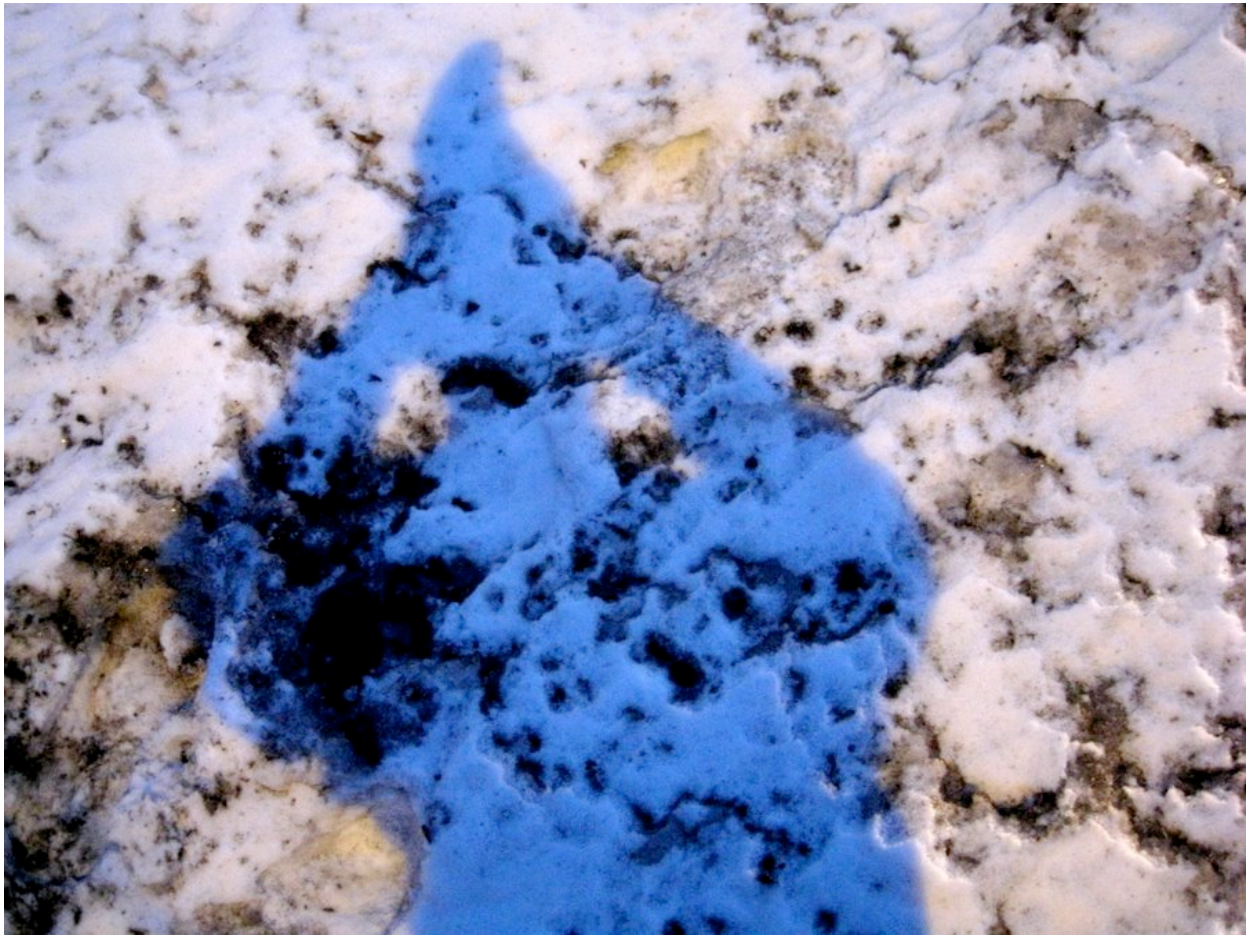
Very Speculatively, Papers Blur Boundaries betw. siRNAs and miRNAs



[Sasidharan & Gerstein, Nature ('08)]

Summary:

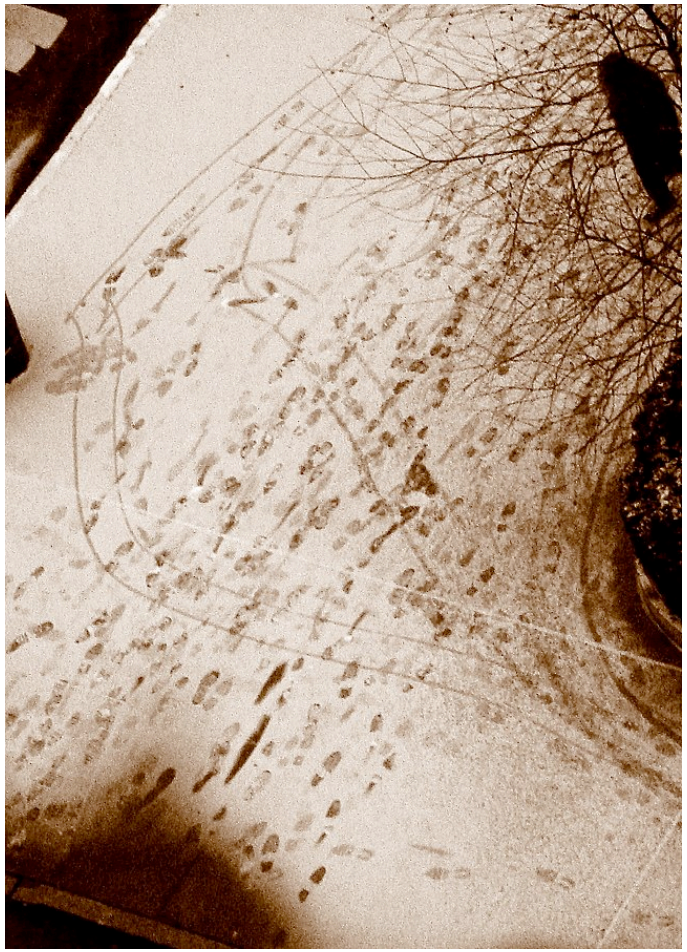
Looking Back Over the Talk



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....

Outline



- Signal processing to call "Blocks"
  - ◊ Calling Punctate Blocks (ChipSeq)
  - ◊ Calling Broader Blocks (CNVs)
- Clustering "Blocks" into larger regions
  - ◊ Binding Sites
- Annotating Copied Regions in the Genome
  - ◊ SD and CNVs
  - ◊ Pseudogenes
- Integration of Pseudogenes with Other Annotations
- Future of Annotation
  - ◊ What is a "gene" post encode?

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

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

Annotating the Human Genome: Integrative Annotation of Pseudogenes in Relation to Conservation, Transcription, and Duplication

- Pseudogene Assignment Technology
  - ◊ Pipeline + DB
  - ◊ Pseudofam analysis of Pseudogene Families, highlight outliers
  - ◊ Ontology
- Annotation of Human Genome
  - ◊ Pipeline draft (20K) + Hybrid Approach
  - ◊ Pilot Phase: Consensus annotation from automatic pipelines & manual curation gives 201
- Integration with Conservation and Seq. Constraint
  - ◊ ~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

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

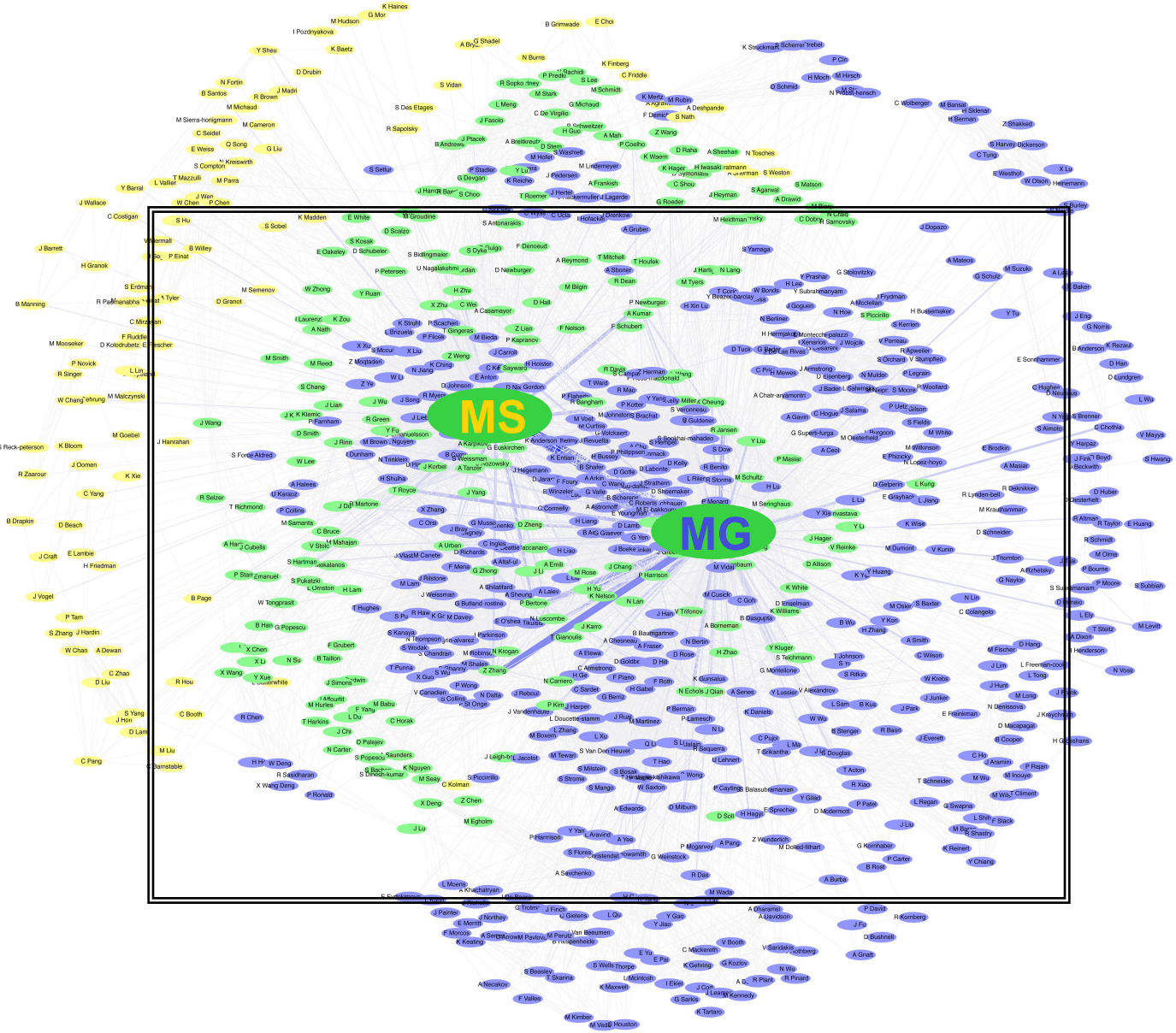
Sanger, UCSC, GIS, AFFX, Geneva, IMIM, BU + SU

+

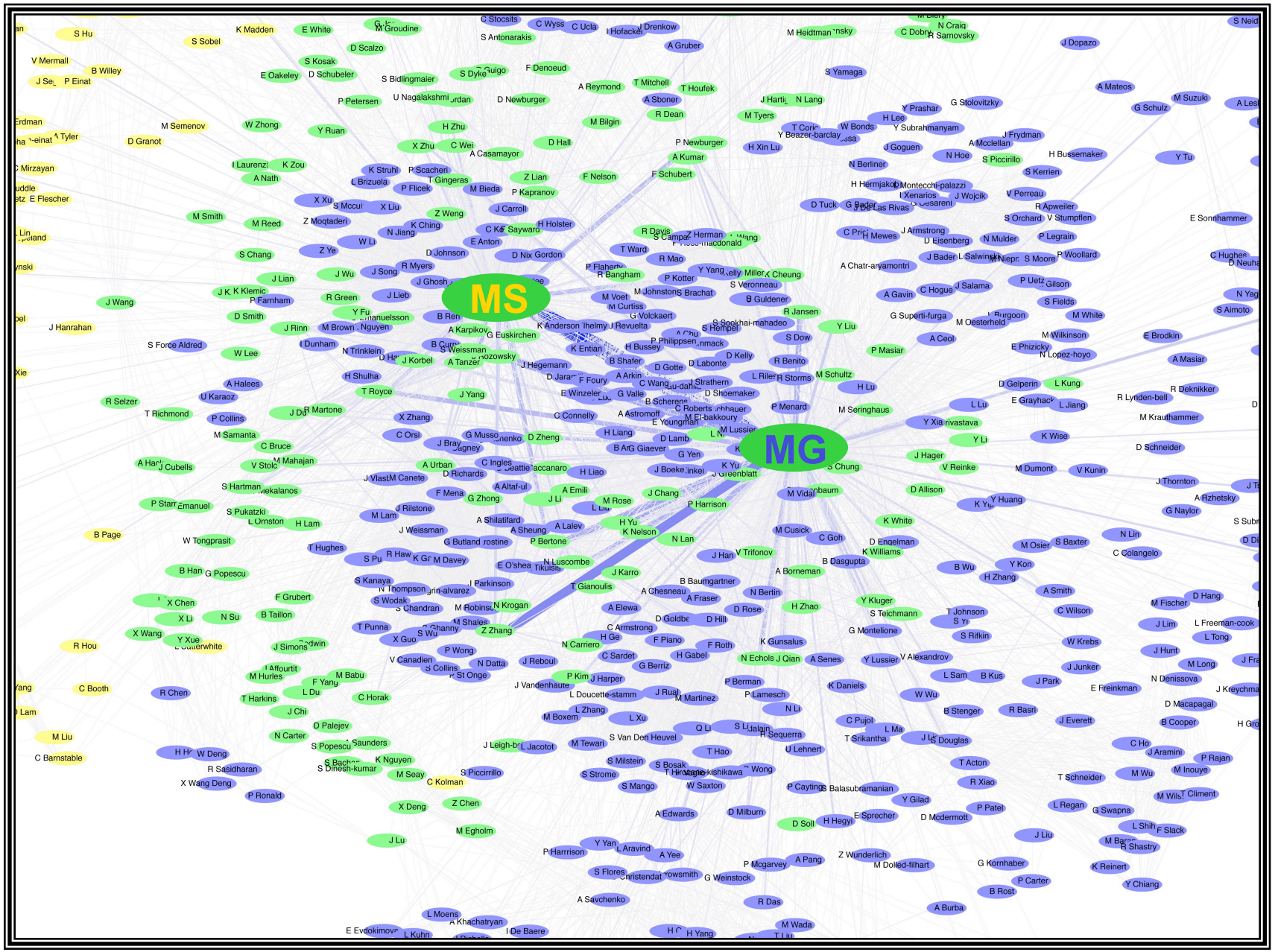
various consortia

(ENCODE, modENCODE, 1000 Genomes)

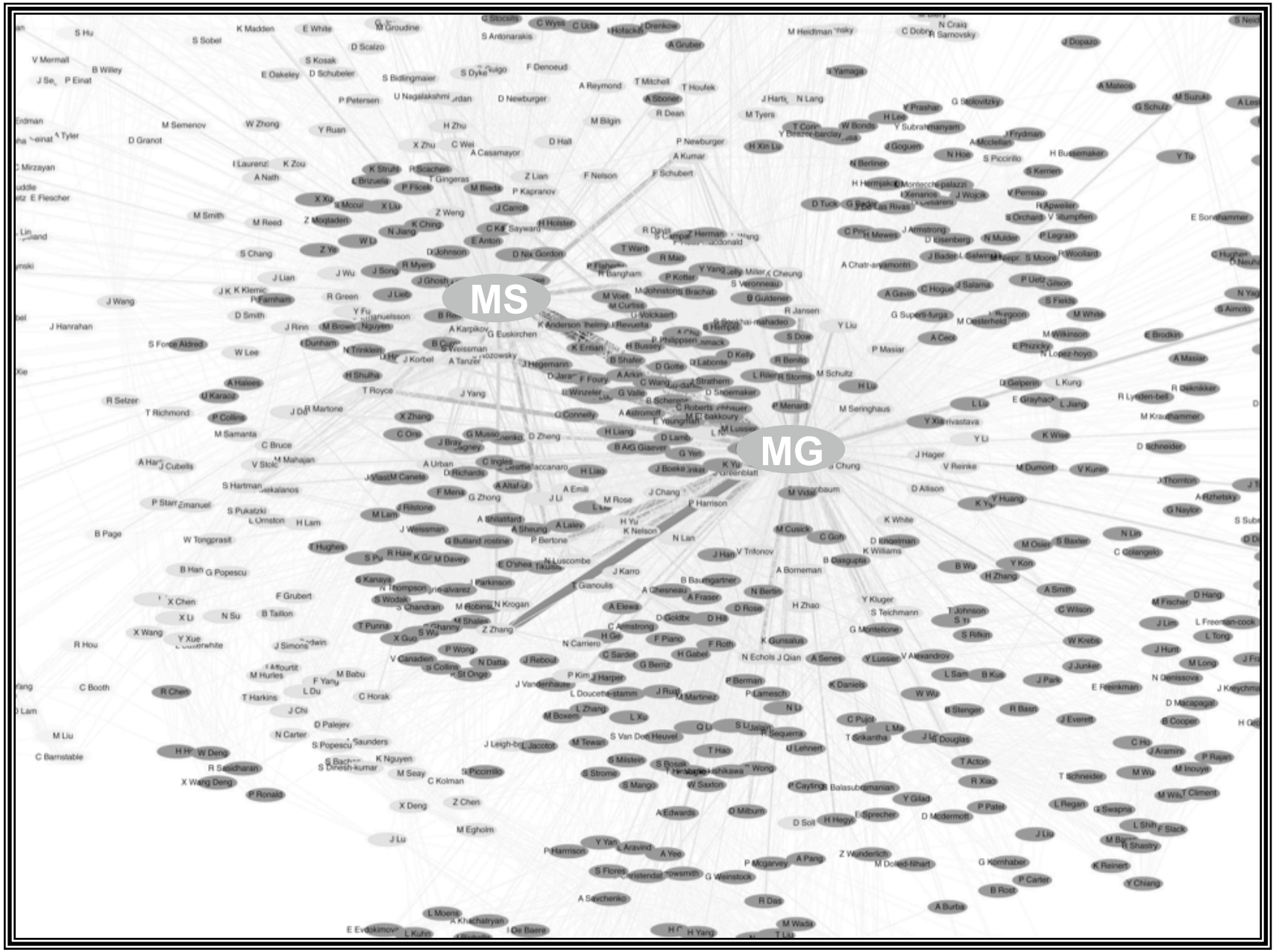
Acknowledgements



Acknowledgements



Acknowledgements



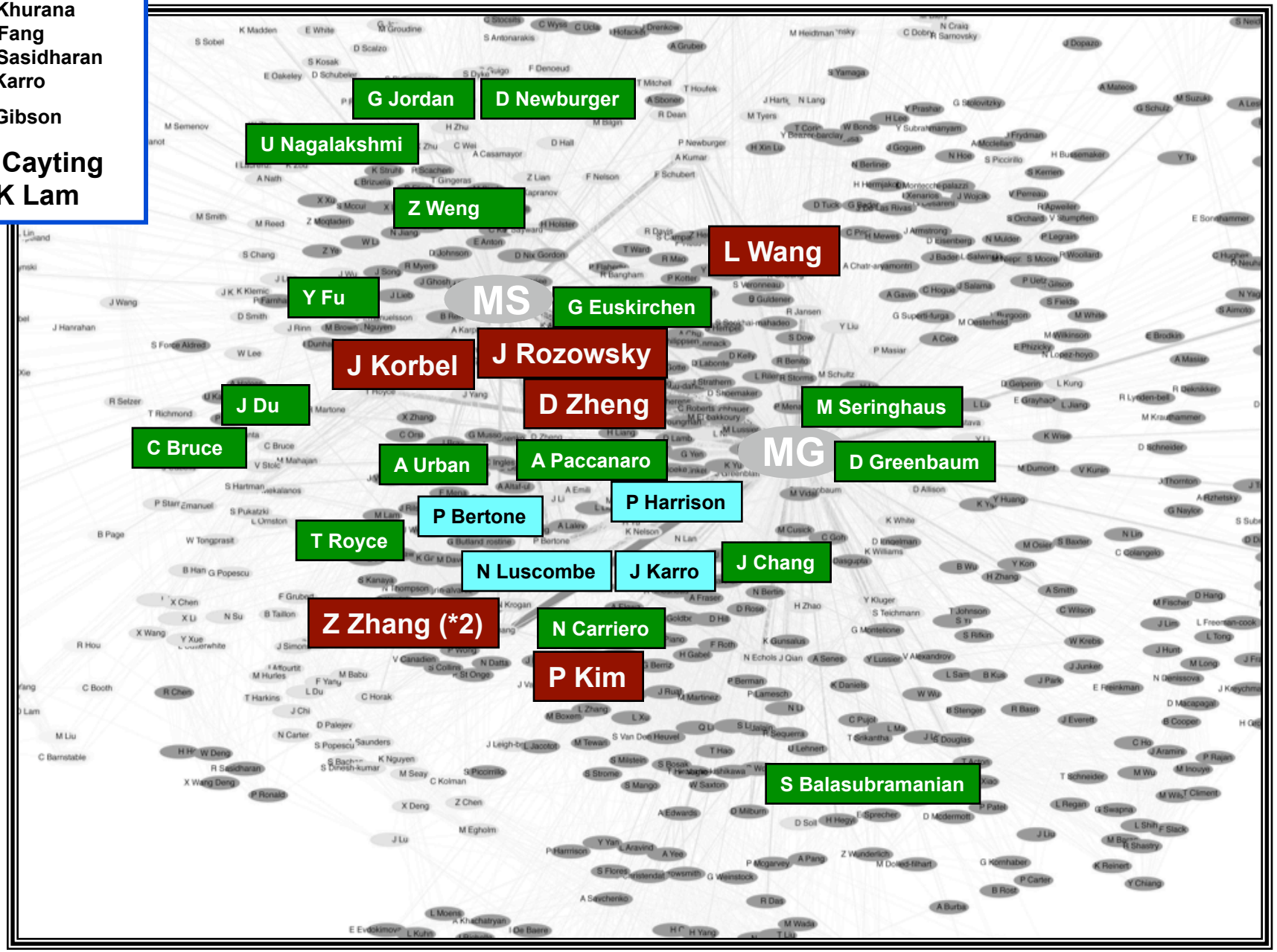
Acknowledgements

pseudogene.org + papers.gersteinlab.org

X Mu
A Abyzov
J Wu
E Khurana
G Fang
R Sasidharan
J Karro

T Gibson

P Cayting
YK Lam



More Information on this Talk

TITLE: Human Genome Annotation

SUBJECT: GenomeTechAnnote, Genomics, Intergenic, Pseudogenes

DESCRIPTION:
Sarkar Lecture at Sick Kids Hosp., U Toronto , 2009.03.16, 13:15-14:15; [I:**SARKAR**] (Long GenomeTechAnnote talk, incl. the following topics:
"junk DNA", **anonymity***, **pgenes-rev***, **whatisgene***, **chip-seq-simu***, **peakseq***, **msb***, **pemer***, **tredist***, **sdcnvcorr*** , **cosbcnv***, **pseudofam***, **pseudopipe***, **encodepgenes***, **rp-pgenes***, **sirnapgene***, **encode-pilot***, **pgene-classify***, & **pubnet*** . Too long, fits into 60' w. skipping most of pgene sect. PPT works on mac & PC and has many photos w. EXIF tags .)

(Paper references in the talk were mostly from Papers.GersteinLab.org. The above topic list can be easily cross-referenced against this website. Each topic abbrev. which is starred is actually a papers "ID" on the site. For instance,
the topic **pubnet*** can be looked up at
http://papers.gersteinlab.org/papers/pubnet)

.

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as **kwpotppt** , that can be easily queried from flickr, viz: **http://www.flickr.com/photos/mbgmbg/tags/kwpotppt** .