Annotating Non-coding Regions of the Human Genome

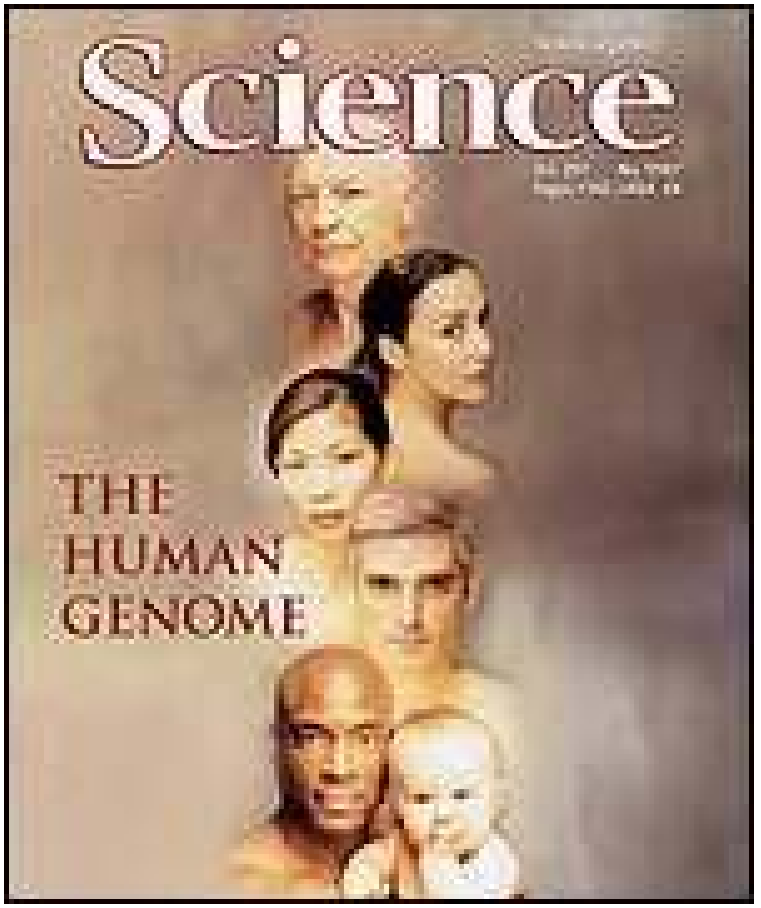
Robert Cedergren Bioinformatics Colloquium
Montreal, Canada
2008.11.03, 16:30-17:30

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

Slides from Lectures.GersteinLab.org

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

(Genome tech and Genome annotation talk, including:
mismatch, uniarray, DART, pgenes-general, pseudofam,
encode-pgenes, sdcnvcorr, gen-encode, sirna-pgene, what-is-gene
[I:CEDERGREN]. Fits into ~50' .)



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

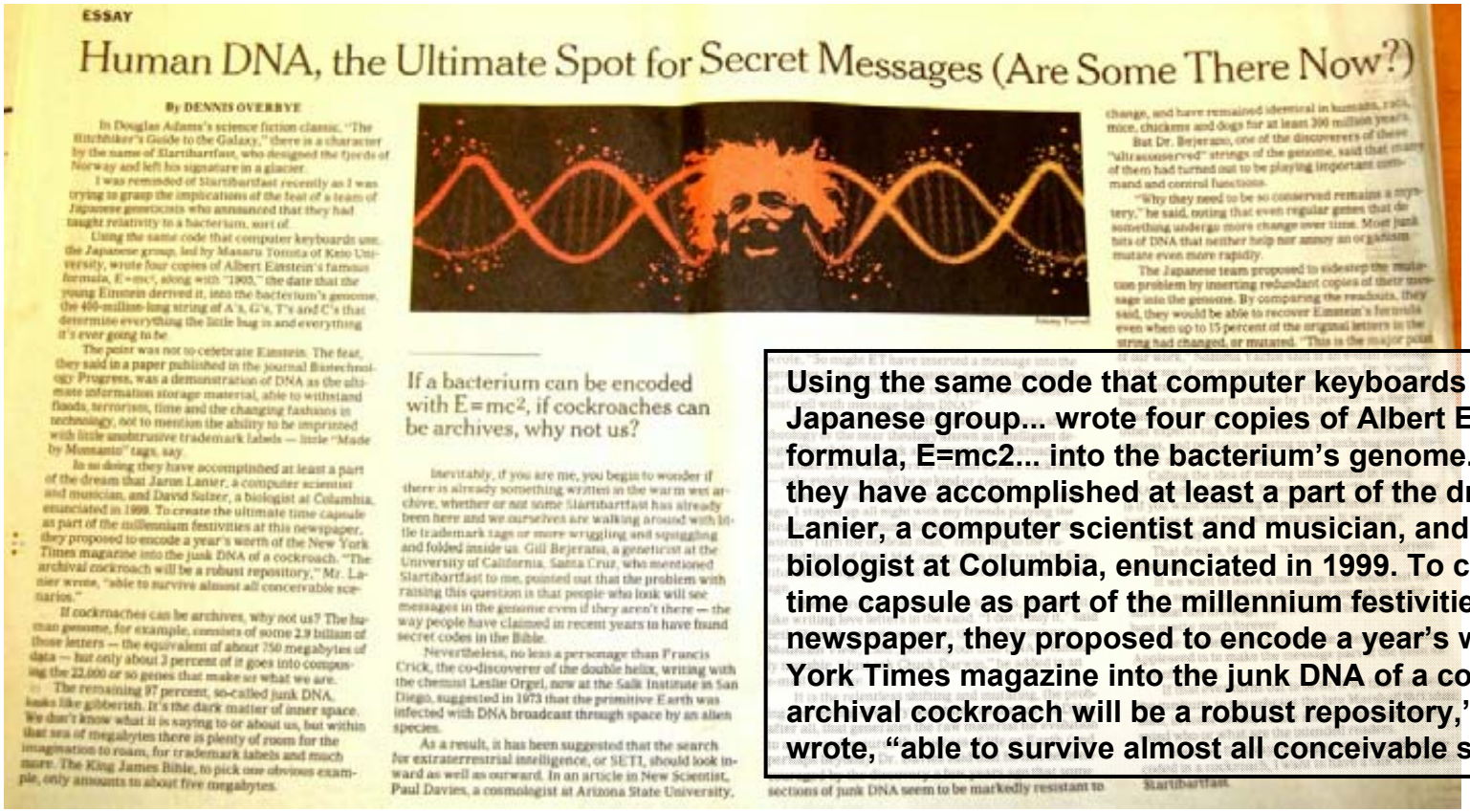
- 1% of Genome (30 Mb in 44 regions)
- Tiling Arrays to assay Transcription & Binding
- Multi-organism sequencing and alignment
- Careful Annotation
- Variation Data

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

- Binding Sites and Regulatory Nets
- Novel Transcribed Regions
- Pseudogenes
- SDs and SVs (Segmental Duplications & Structural Variations)

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 460-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, terrorists, 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.

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 struggling and squiggling and folded inside us. Gill 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."

How might we annotate a human text ?

Color is Function

Lines are Similarity

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

The Semicolon Wars

Brian Hayes

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

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

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

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

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

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

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

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

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

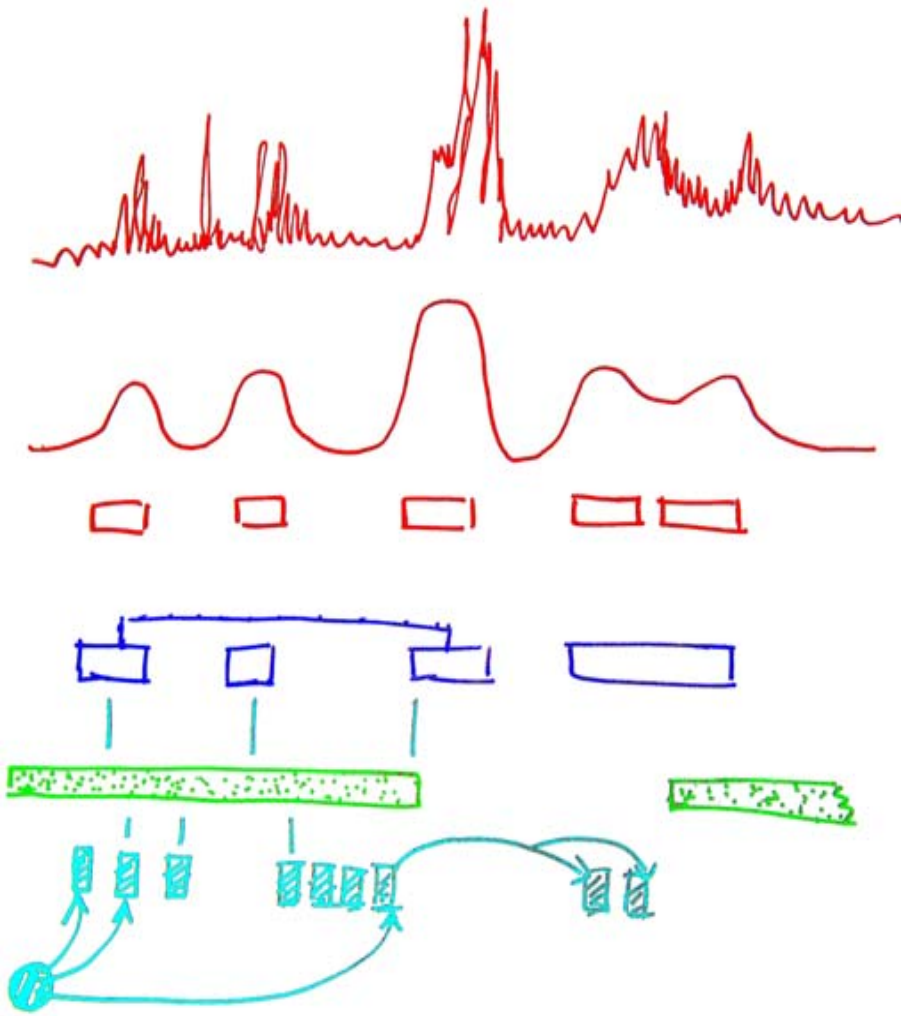
This modern echo of Swift's Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, "On holy wars and a plea for peace." The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

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

Overview of the Process of Annotation of non-coding Regions

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

Overview of Annotation Process



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

Outline

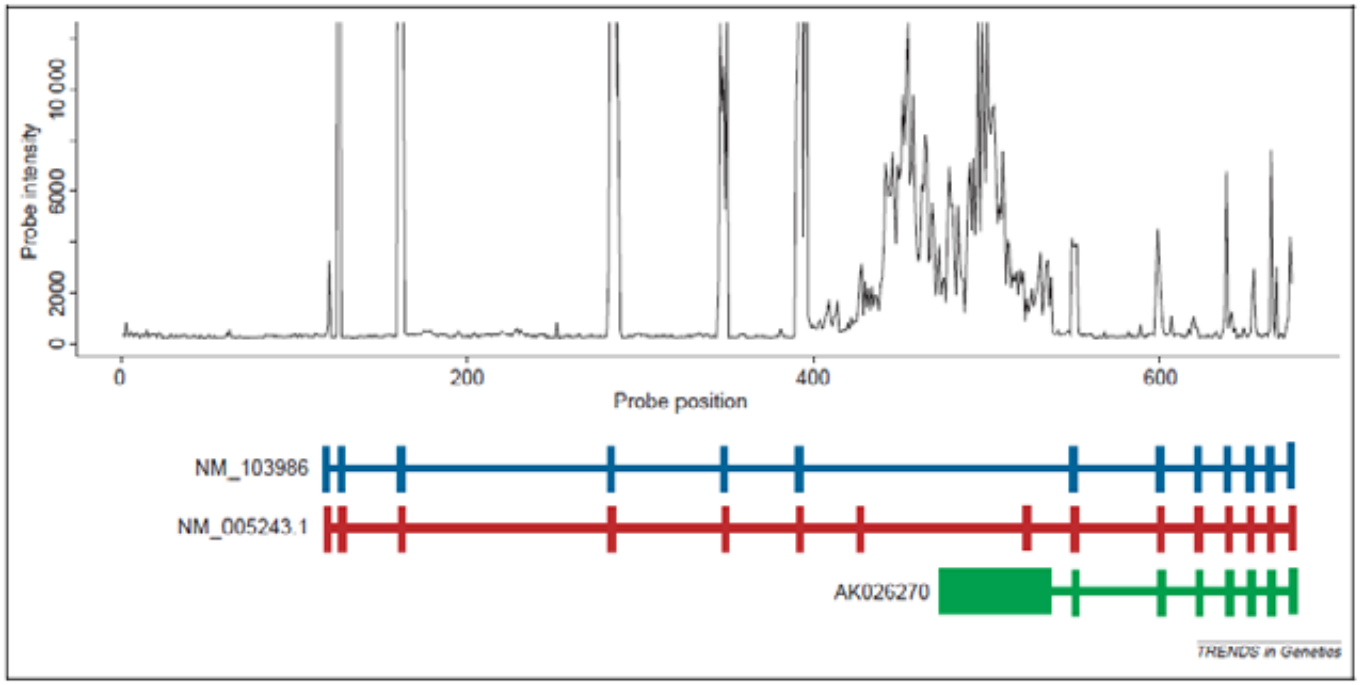
- Calling Blocks
  - ◊ Issues in Signal Processing Transcriptional Data from Tiling Arrays
- Clustering Blocks
  - ◊ Scoring transcribed regions and clustering these together into larger blocks
- Formal Annotation
  - ◊ Pseudogenes
- Integration with conservation
  - ◊ SDs and CNVs
  - ◊ pseudogenes and seq. constraint
- Integration with activity
  - ◊ pseudogenes and transcriptions
- Future of Gene Annotation
  - ◊ What is a gene post encode?

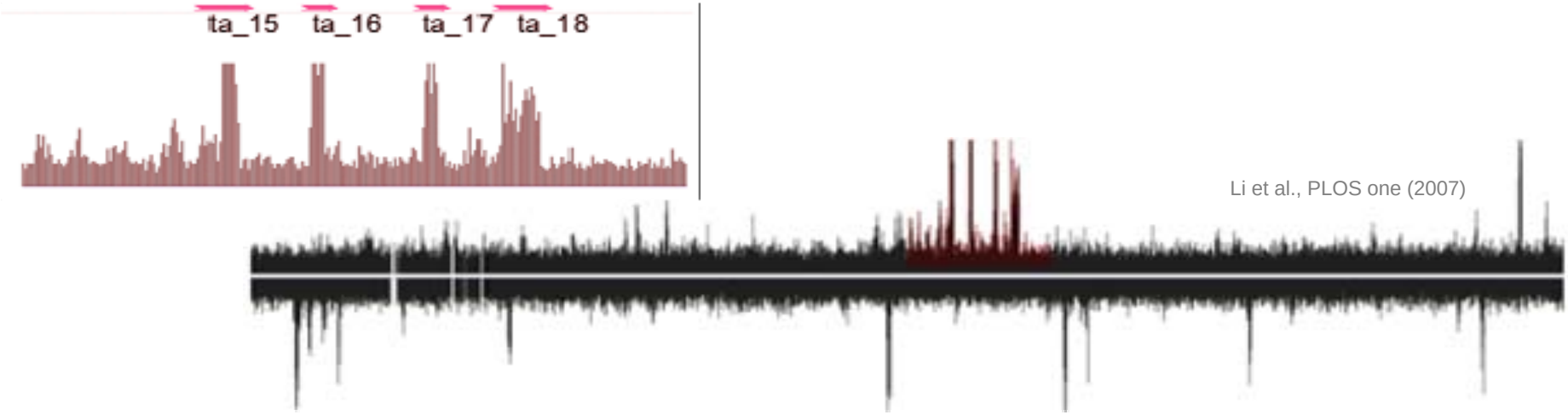
Signal Processing: Normalizing Signal and Finding Initial Annotation Blocks ("Hits")

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

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

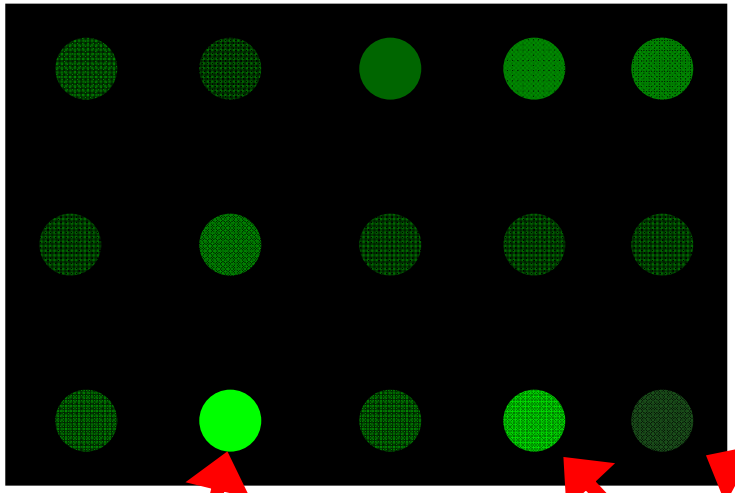
Johnson et al. (2005) TIG, 21, 93-102.





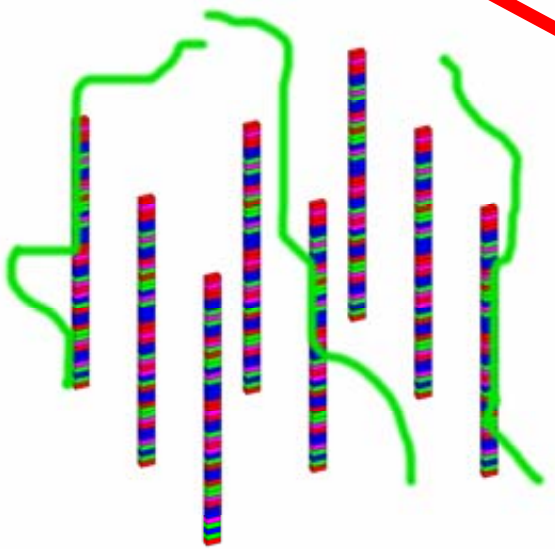
Li et al., PLOS one (2007)

Specific & Non-specific Cross-Hyb.

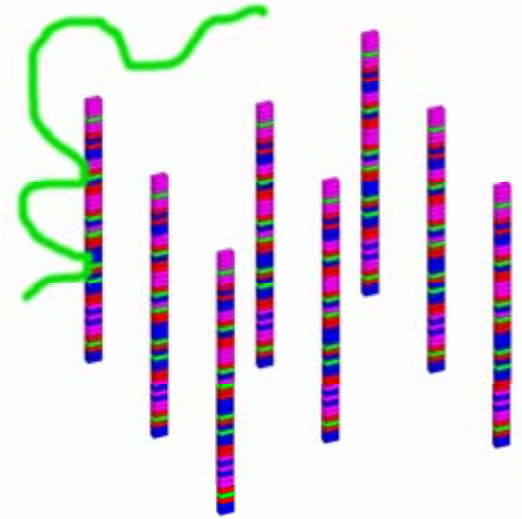


- Perfect match (PM): probe binding intended target
- Specific cross-hyb.: probes binding non-PM targets with a small number of mismatches
- Non-specific cross-hyb.: probes binding targets with many mismatches, due to general stickiness of oligos

Perfect Match



Specific Cross-hyb.

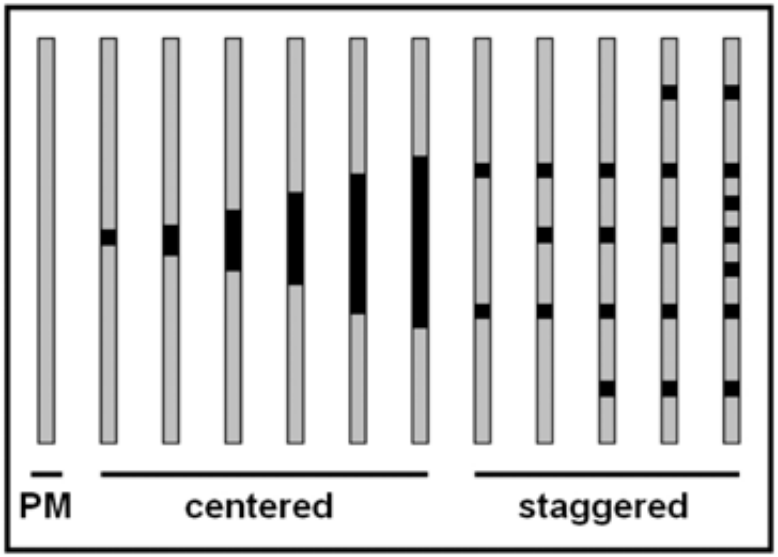


Non-specific Cross-hyb.

Non-Specific Cross Hyb. (Sequence Effects)

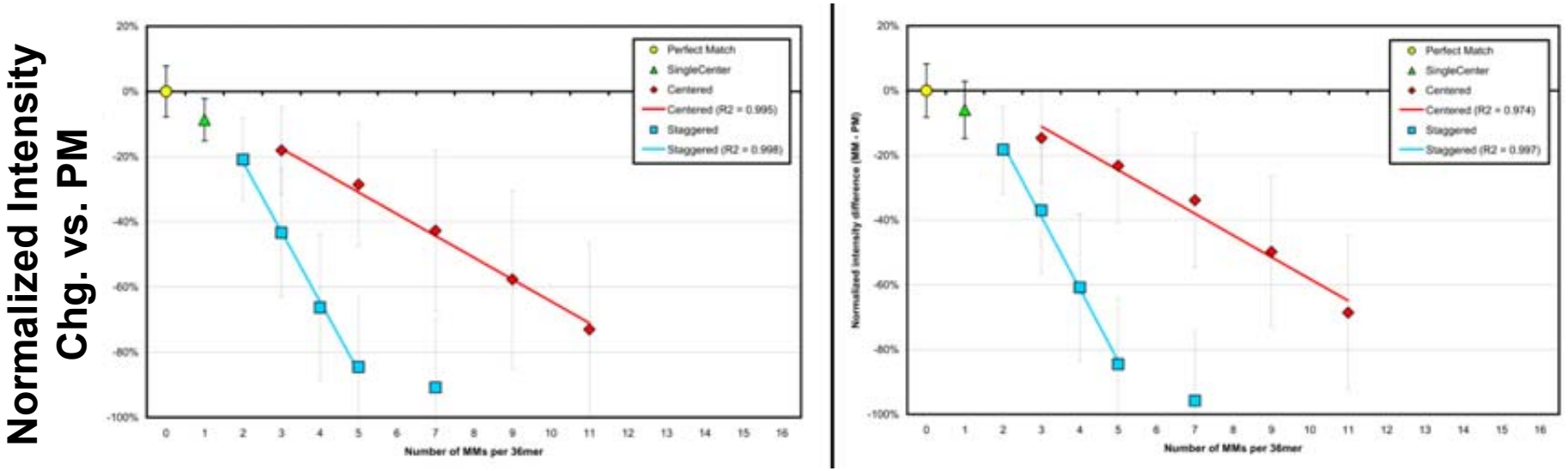
Creation of Standardized Datasets for Quantifying Effect of Mismatches

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



Yeast ACT1 Gene

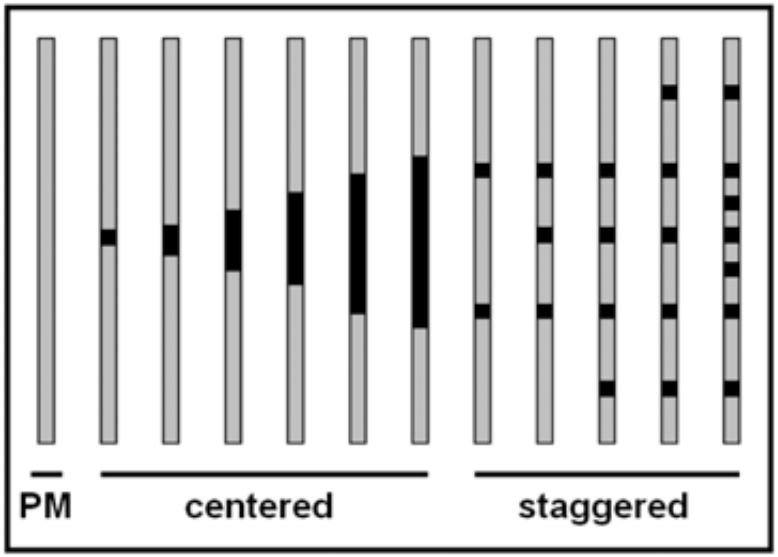
Human HBG2 Gene



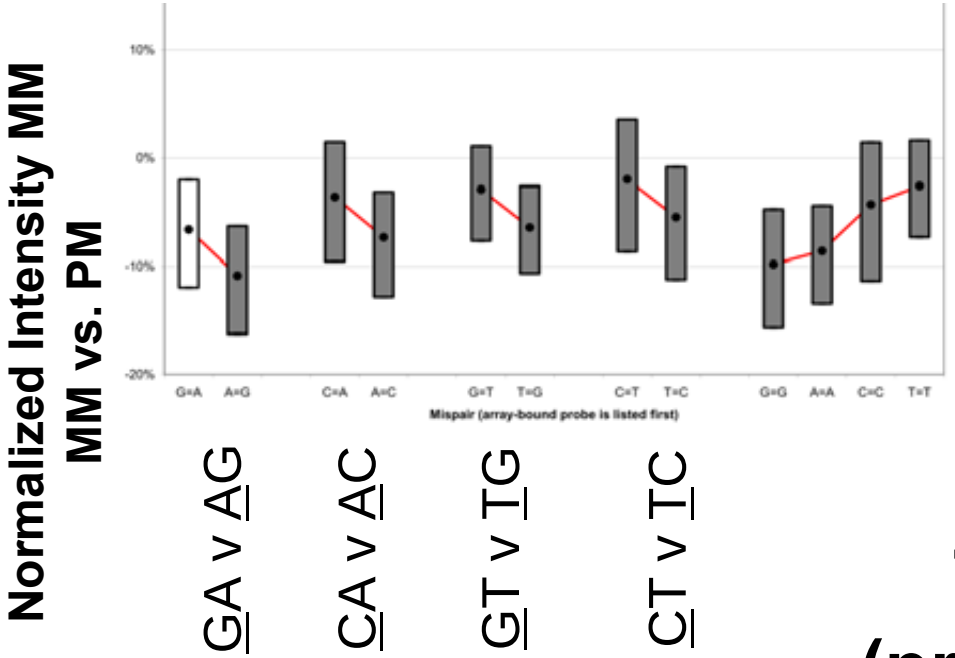
Number of Mismatches

Creation of Standardized Datasets for Quantifying Effect of Mismatches

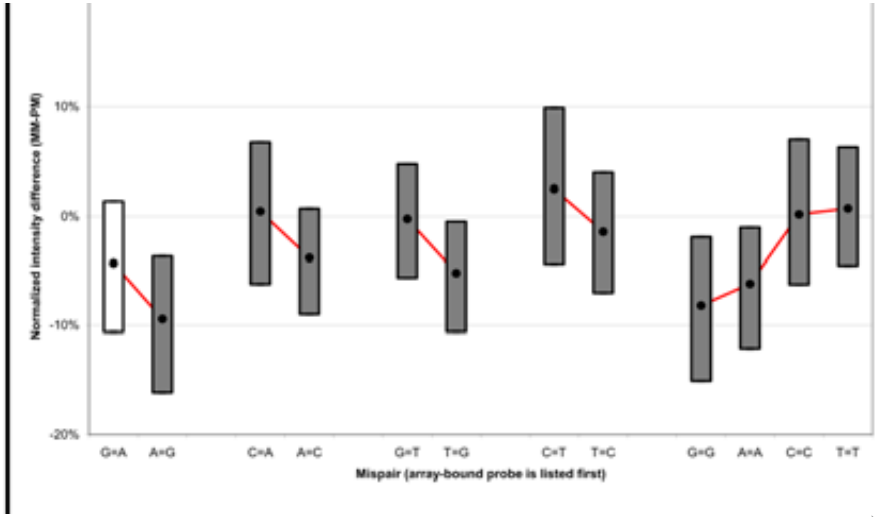
[Seringhaus et al., BMC Genomics (in press)]



Yeast

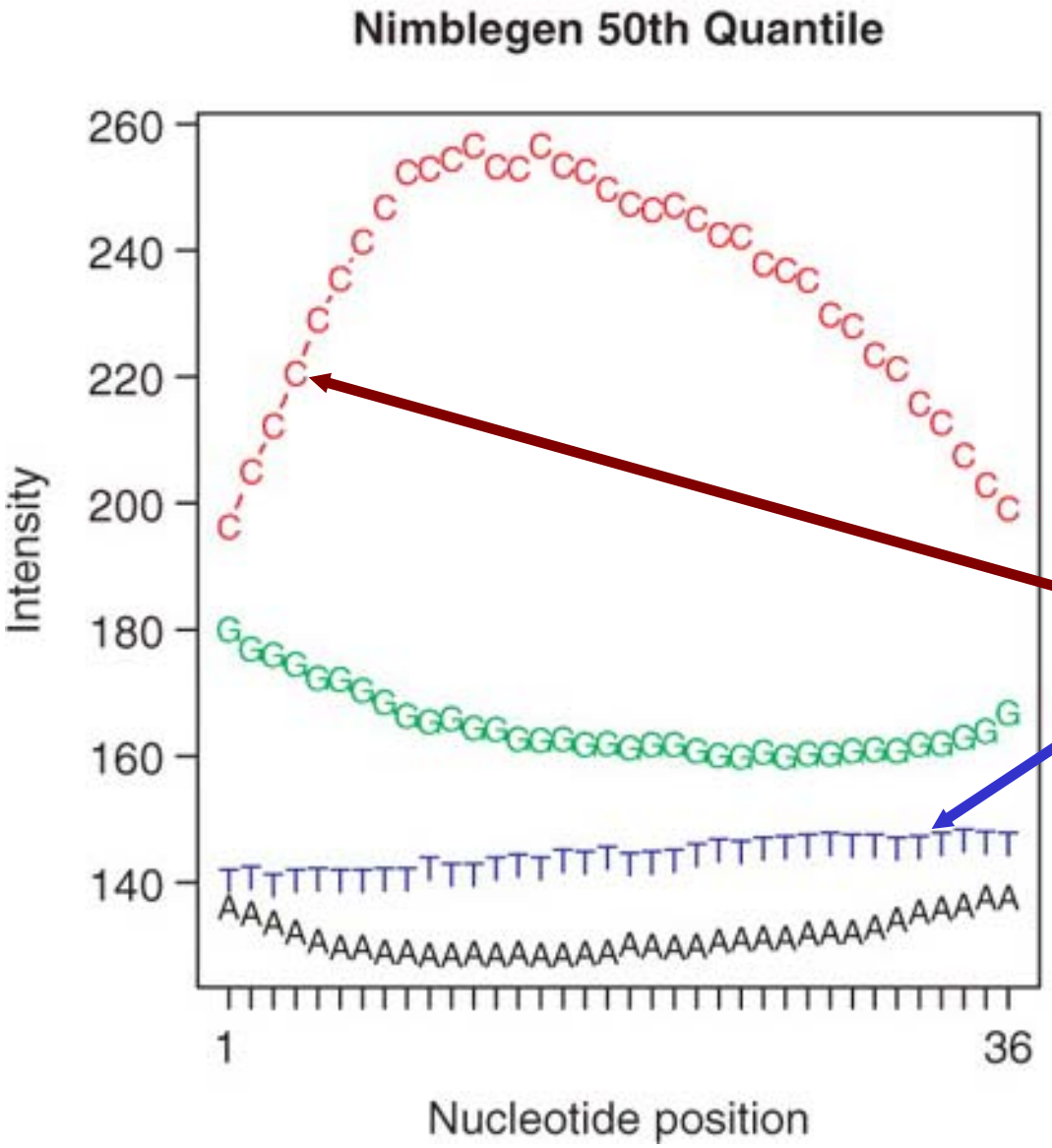


Human



Types of Mispairs (probe on array is first)

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



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

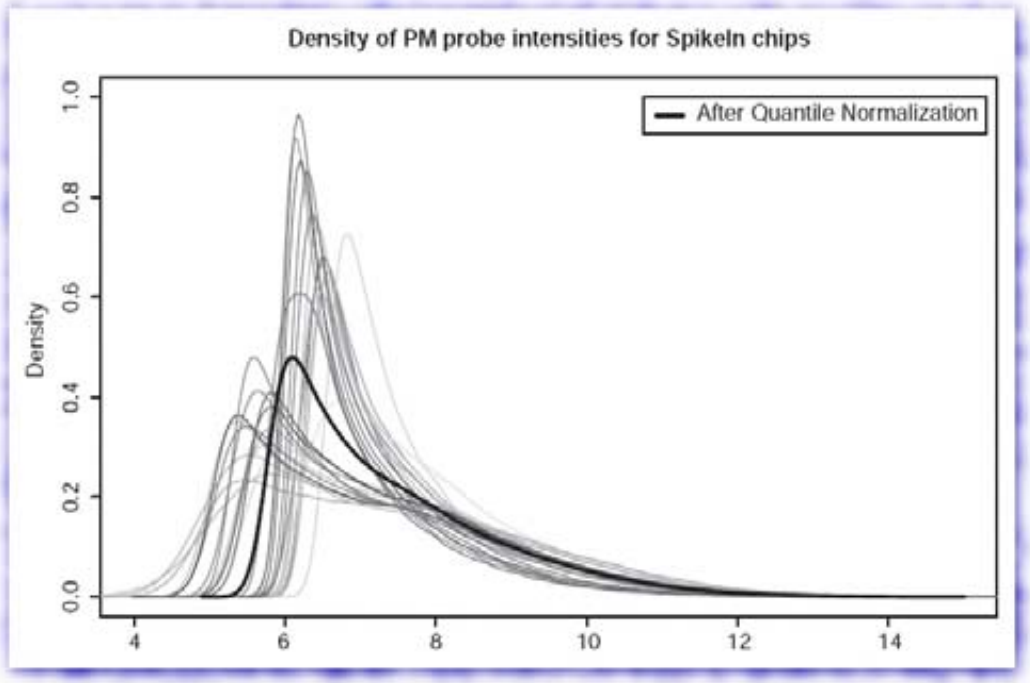
Avg. intensity of all background probes with a T at position 33

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

Quantile Normalization

Gene expression quantile normalization

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

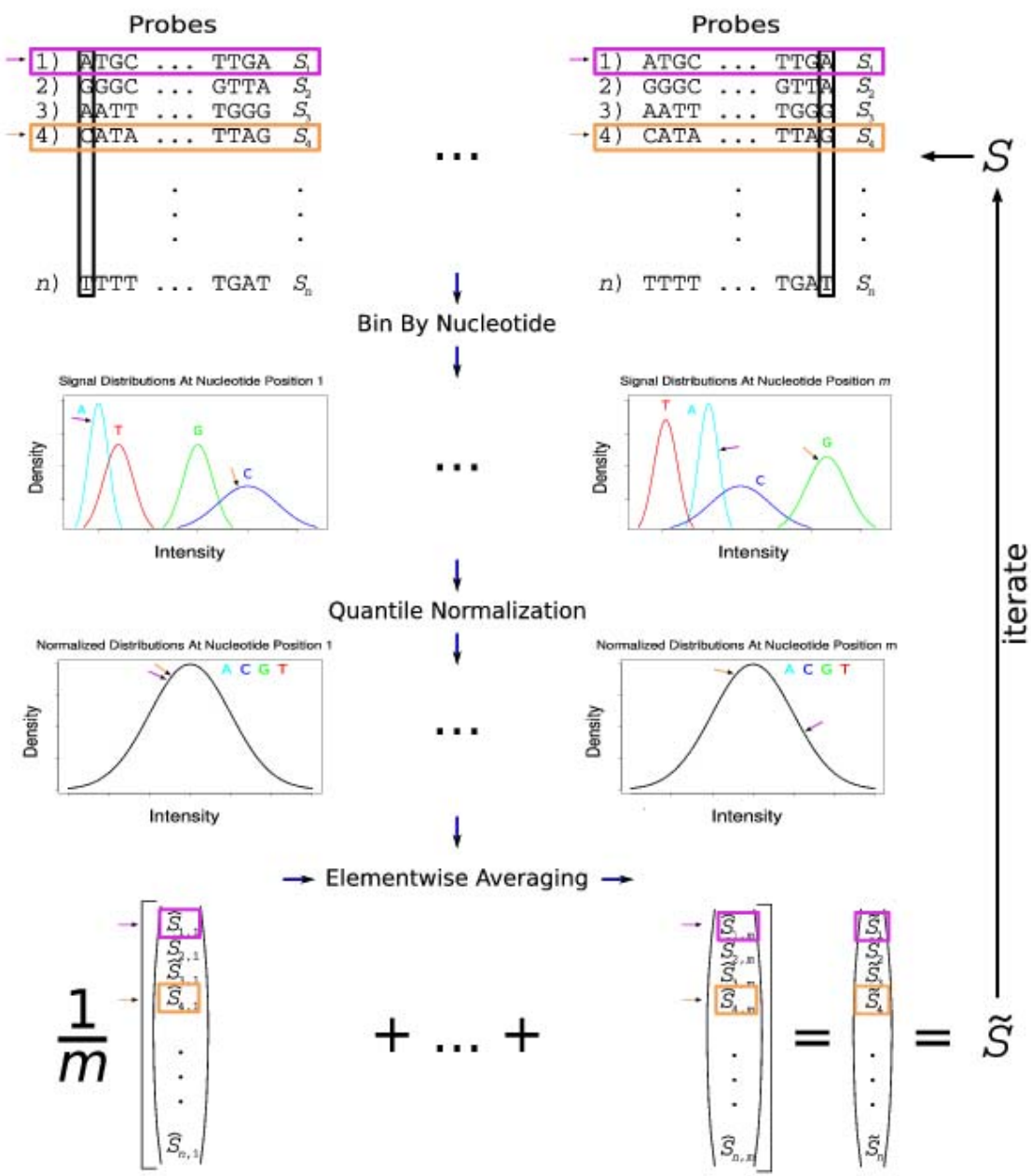


Source: Bolstad, B.M., et al (2003), *Bioinformatics*, **19**, 185-93.

Example

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

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)

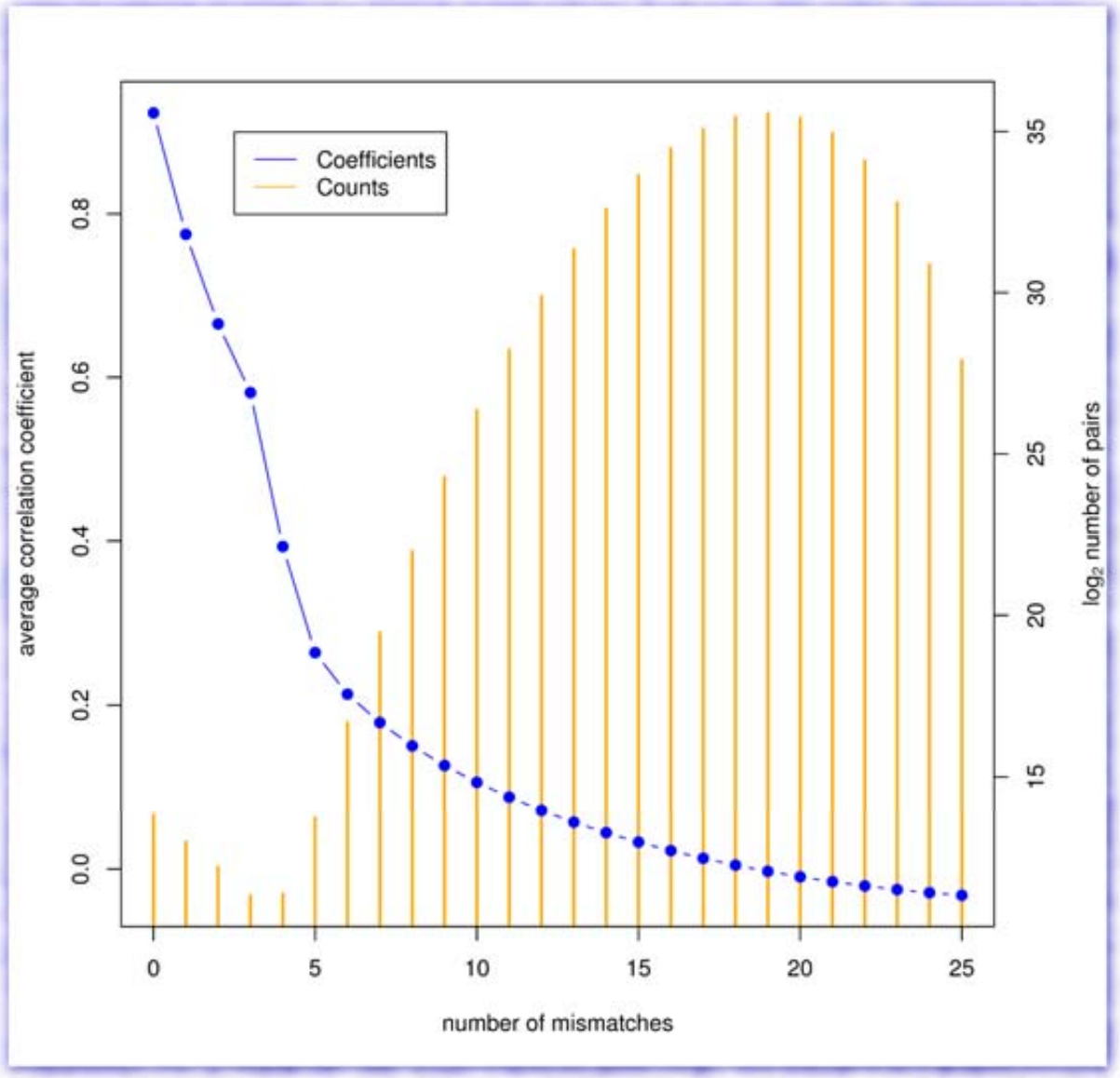
T Royce et al (2007), *Bioinformatics*, **23**, 988-97

Measuring Specific Cross-Hyb

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

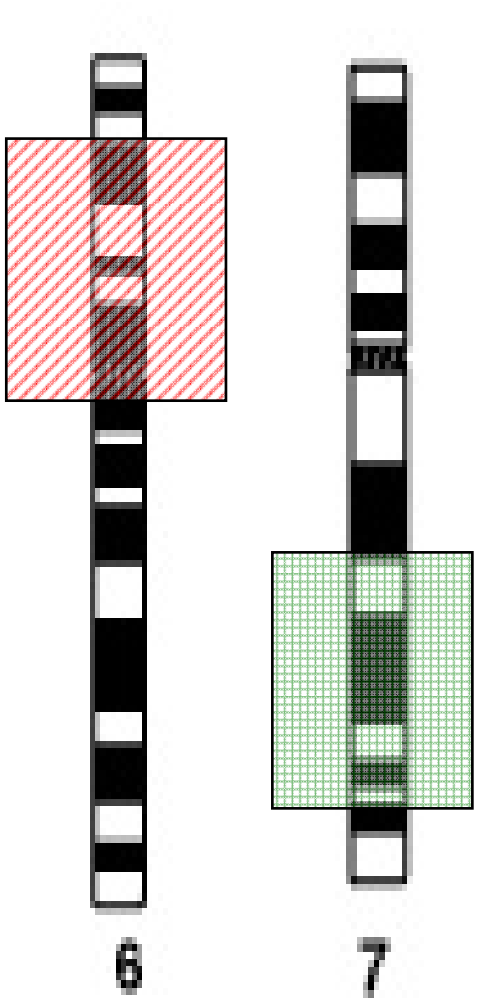
Measuring Specific Cross-Hyb

- 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

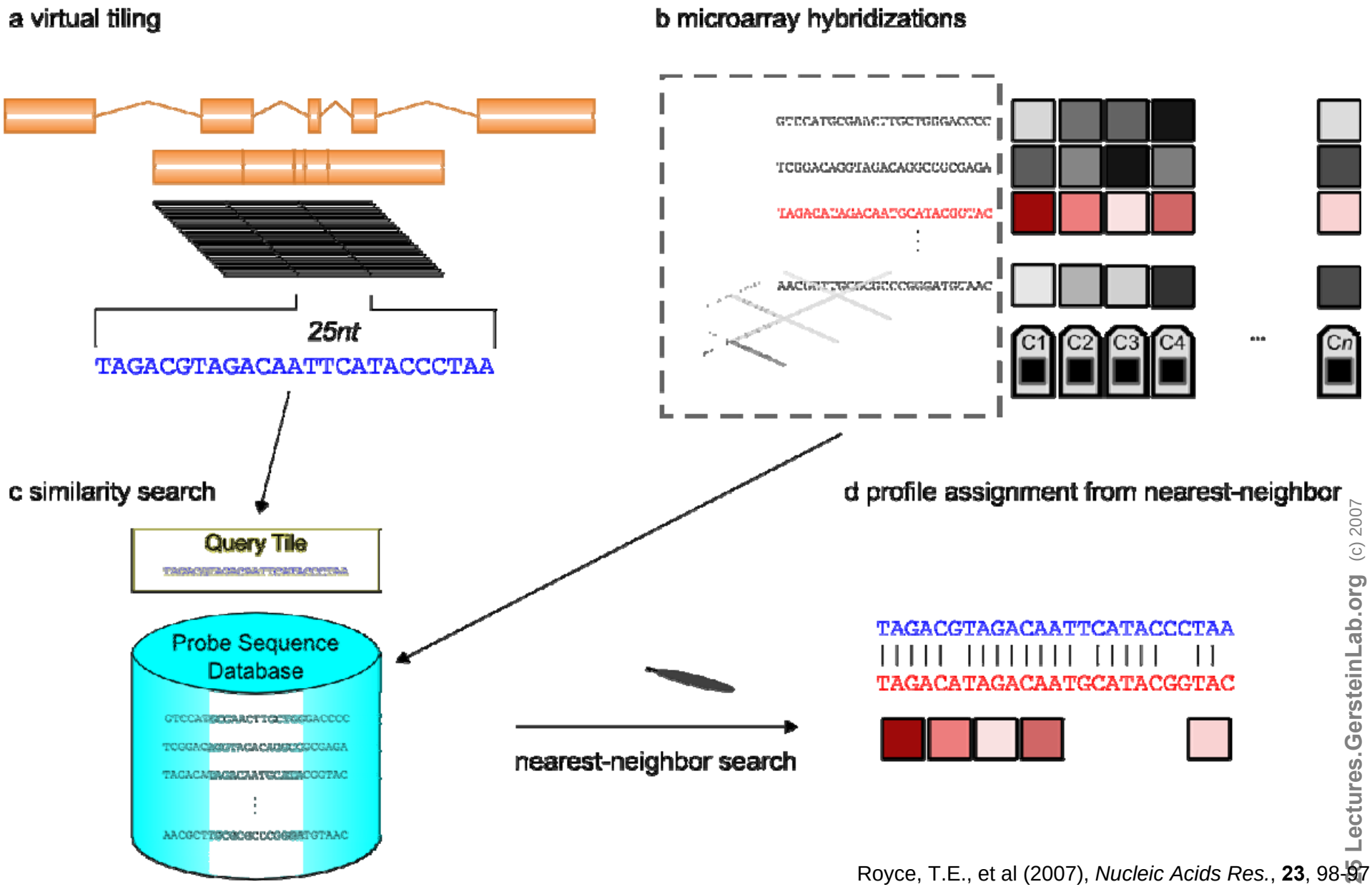


- Using Cheng et al. (2005), predict gene expression levels (and profiles across tissues) for genes on part of chr. #6
- ...Based on closest cross-hyb tiles on part of chr. #7
- Then compare to measured levels and profile on #6

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

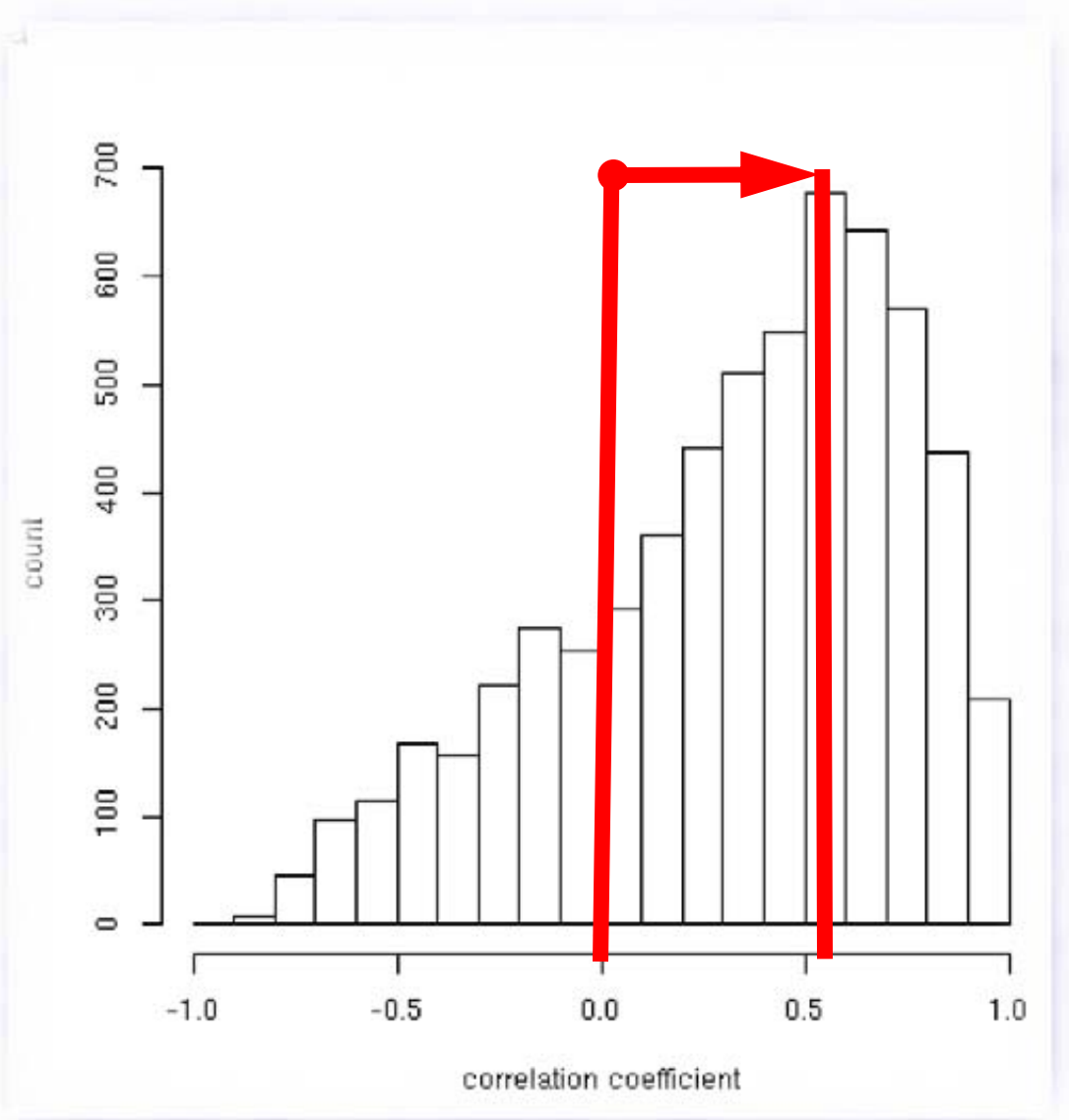
Nearest Nbr Search on Virtual Tiling



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

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)

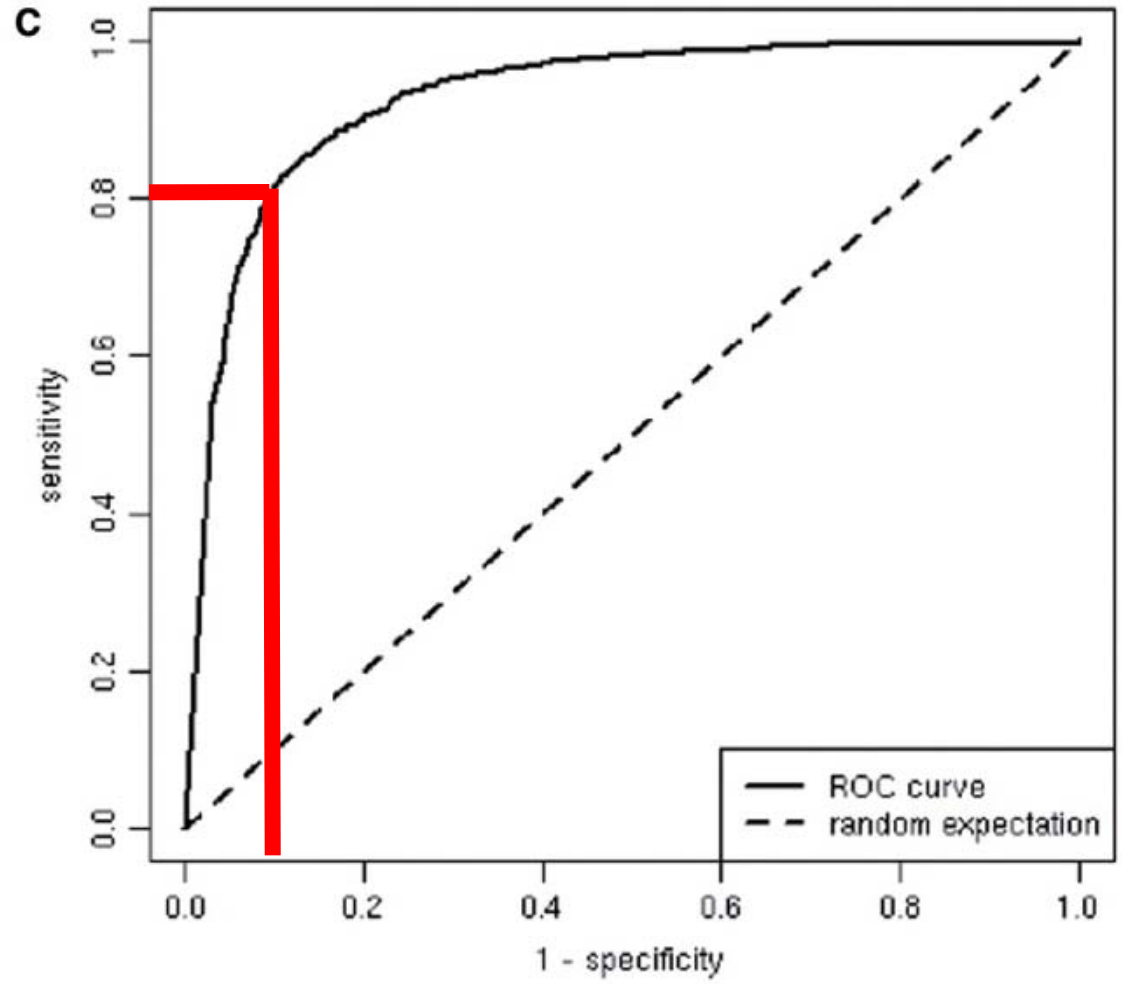


Source: Royce, T.E., et al (2007), *Nucleic Acids Res.*,

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**

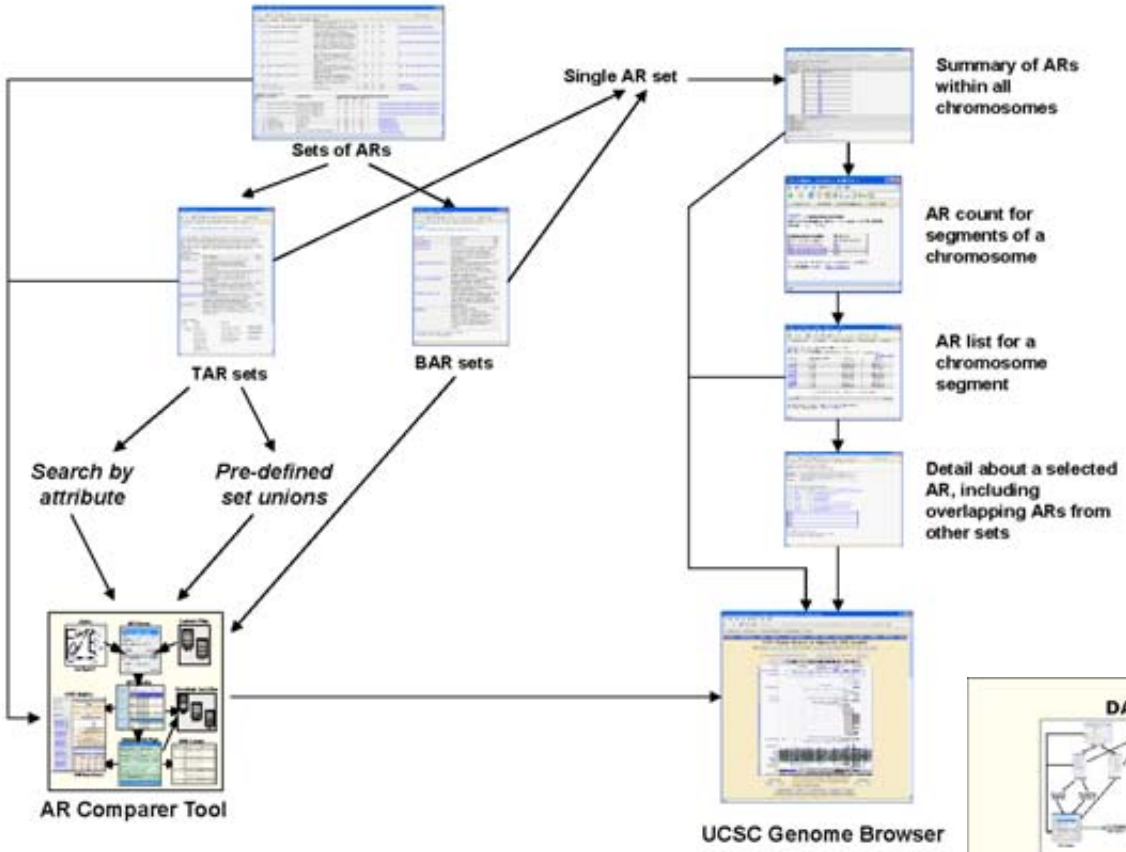
- Gold std. set of known expressed genes. How well do we find.
- A set of known positives was defined as the Refseq genes with at least 75% transfrag coverage. A set of known negatives was constructed by permuting the sequences in the set of known positives. For various thresholds, sensitivity and specificity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

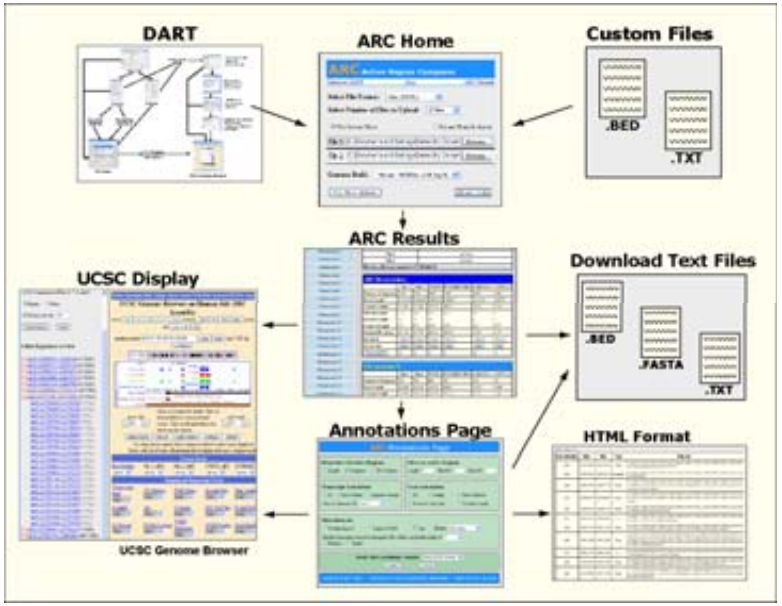


Royce, T. E. et al. Nucl. Acids Res. 2007 35:e99

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

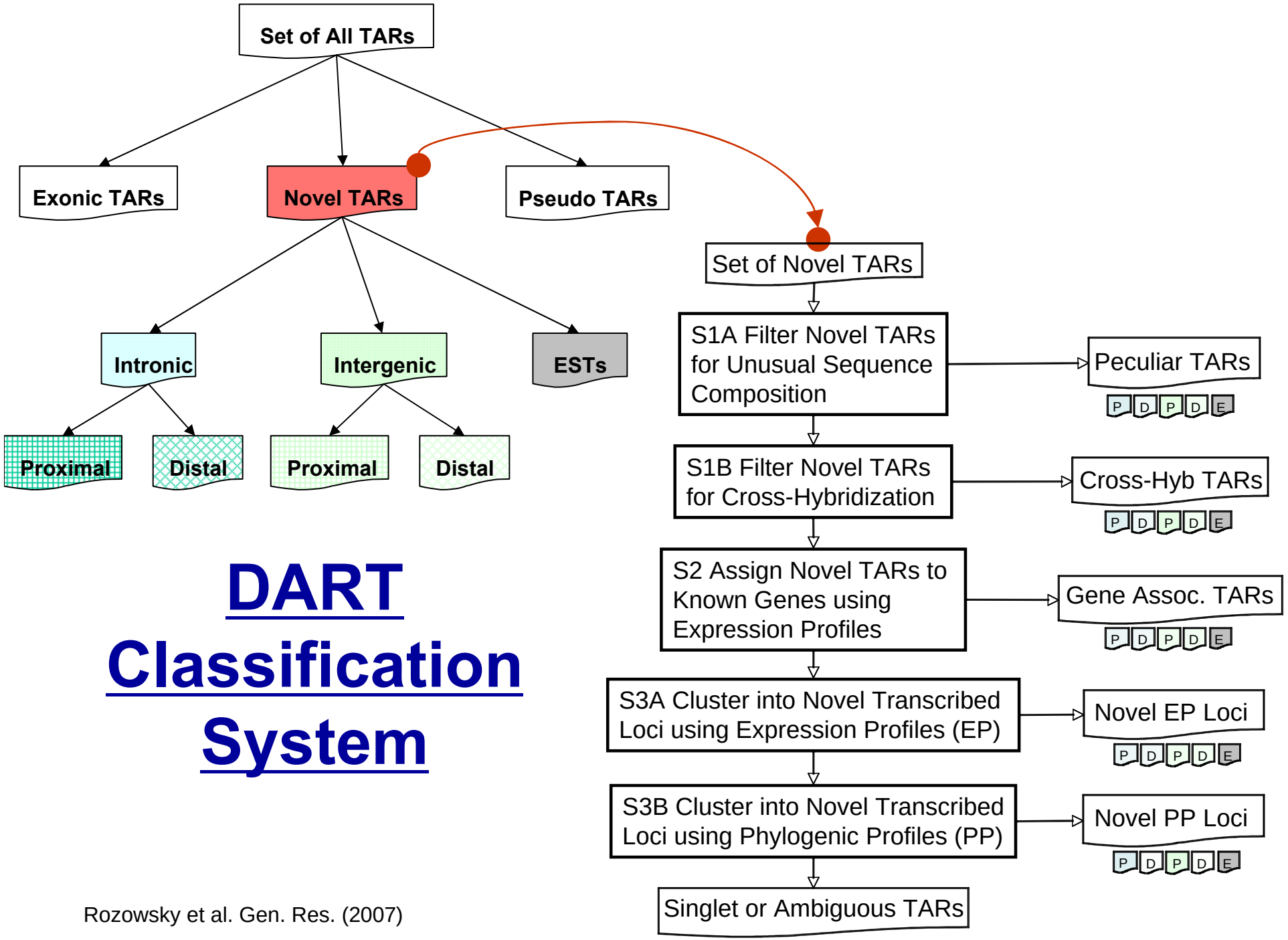
pers. **photo**, see streams.gerstein.info





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

Rozowsky et al. Genome Research (2007)

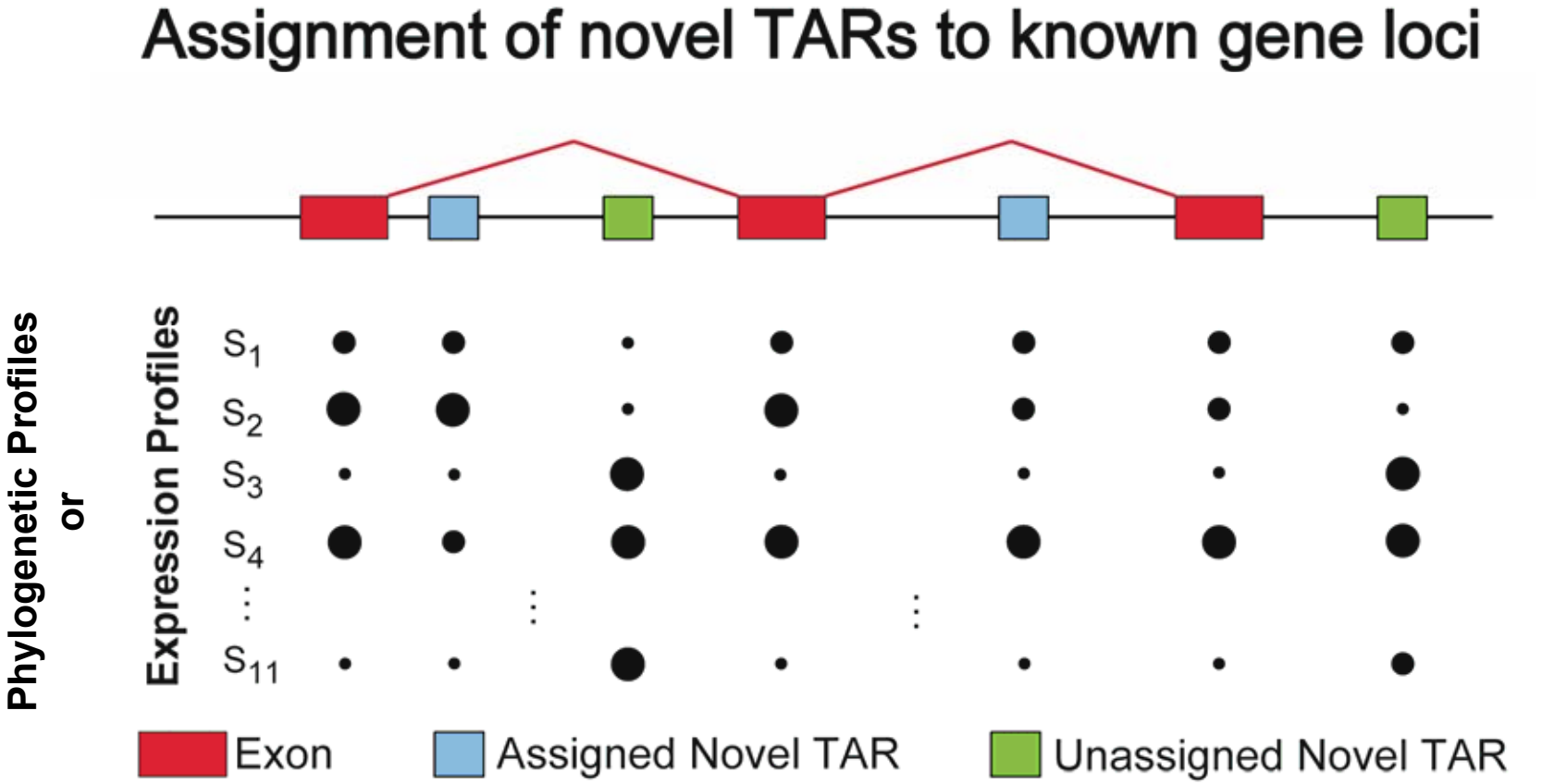


DART Classification System

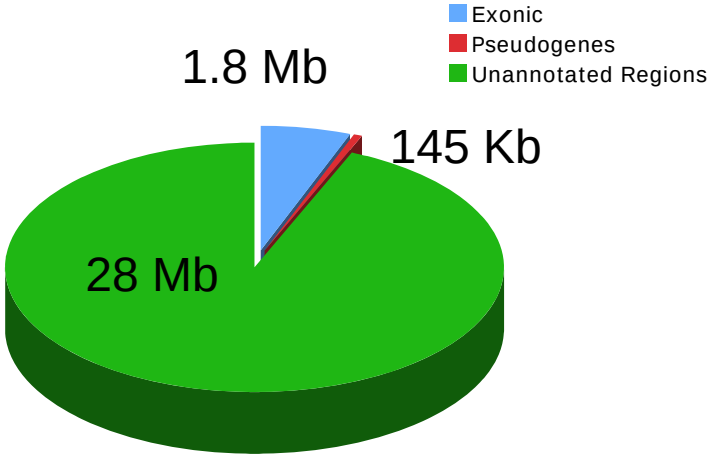
Rozowsky et al. Gen. Res. (2007)

Production | Integrated

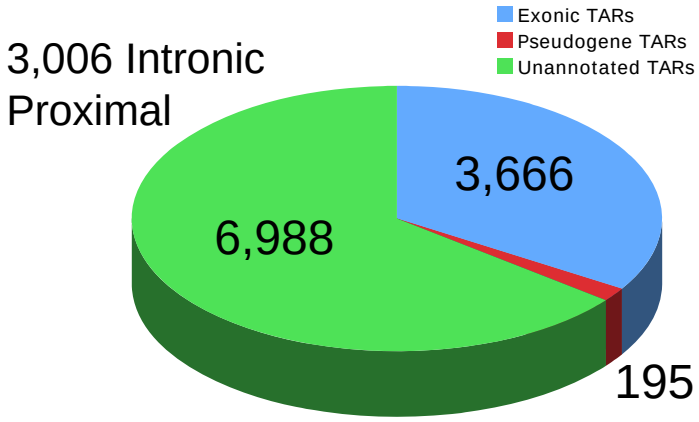
More Developed Annotation: Clustering and Classifying Blocks of Un-annotated Transcription into larger units



Rozowsky et al. Gen. Res. (2007)



ENCODE Regions (30 Mb)



Locations of TARs

Of the approx 7,000 Novel TARs

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

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

◊ RT-PCR & Sequencing

◊ 18/46 (39%) confirmed by RT-PCR

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

Rozowsky et al. Genome Research (2007)

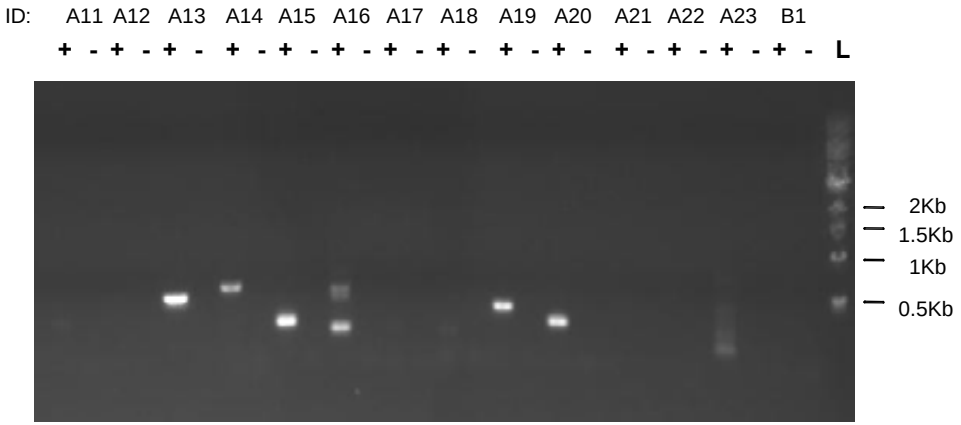
DART Classification has been experimentally validated with some small scale experiment

- RT-PCR & Sequencing

Results:

18/46 (39%) confirmed by RT-PCR

4/5 Sequenced Products Map uniquely to correct genomic region



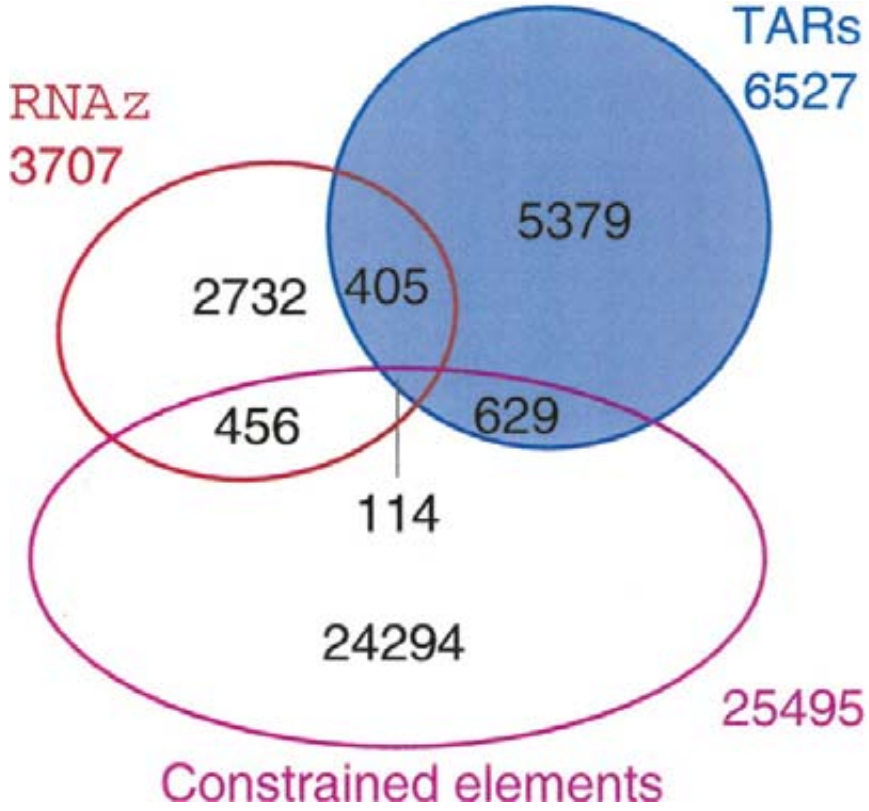

```
TAR 1
                                  Forward Primer
PCR Sequence     1 ttcttcggaaaagcacatgaactctttggagtctcctgttccacttggtaaatttcctat 60
                   ||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
Chr21   34,270,569 ttcttcggaaaagcacatgaactcttcggagtctcctgttccacttggtaaatttcctat 34,270,628

PCR Sequence    61 agctccgcactgaaagtccctgctgccctccttcctctgagcttgtggggcccacagatc 120
                   |||  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Chr21   34,270,629 agccacgcactgaaagtccctgctgccctcctcctctgagcttgtggggcccacagatc 34,270,688

PCR Sequence   121 ccctgctccacttcctgcttcatttcagctgat 153
                   |||||||||||||||||||||||||||||||
Chr21   34,270,689 ccctgctccacttcctgcttcatttcagctgat 34,270,721


TAR 2

PCR Sequence   154 ggatgacactccctcgttctaataccatctgaatgcctgagcaattacatcttacaacct 213
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Chr21   34,270,898 ggatgacactccctcgttctaataccatctgaatgcctgagcaattacatcttacaacct 34,270,957

PCR Sequence   214 catgaaaaacacagcagcttgtcacgatgaatg 246
                   ||||||||||||||||||||||||||||||
Chr21   34,270,958 catgaaaaacacagcagcttgtcacgatgaatg 34,270,990
                                  Reverse Primer
```

Rozowsky et al. Genome Research (2007)

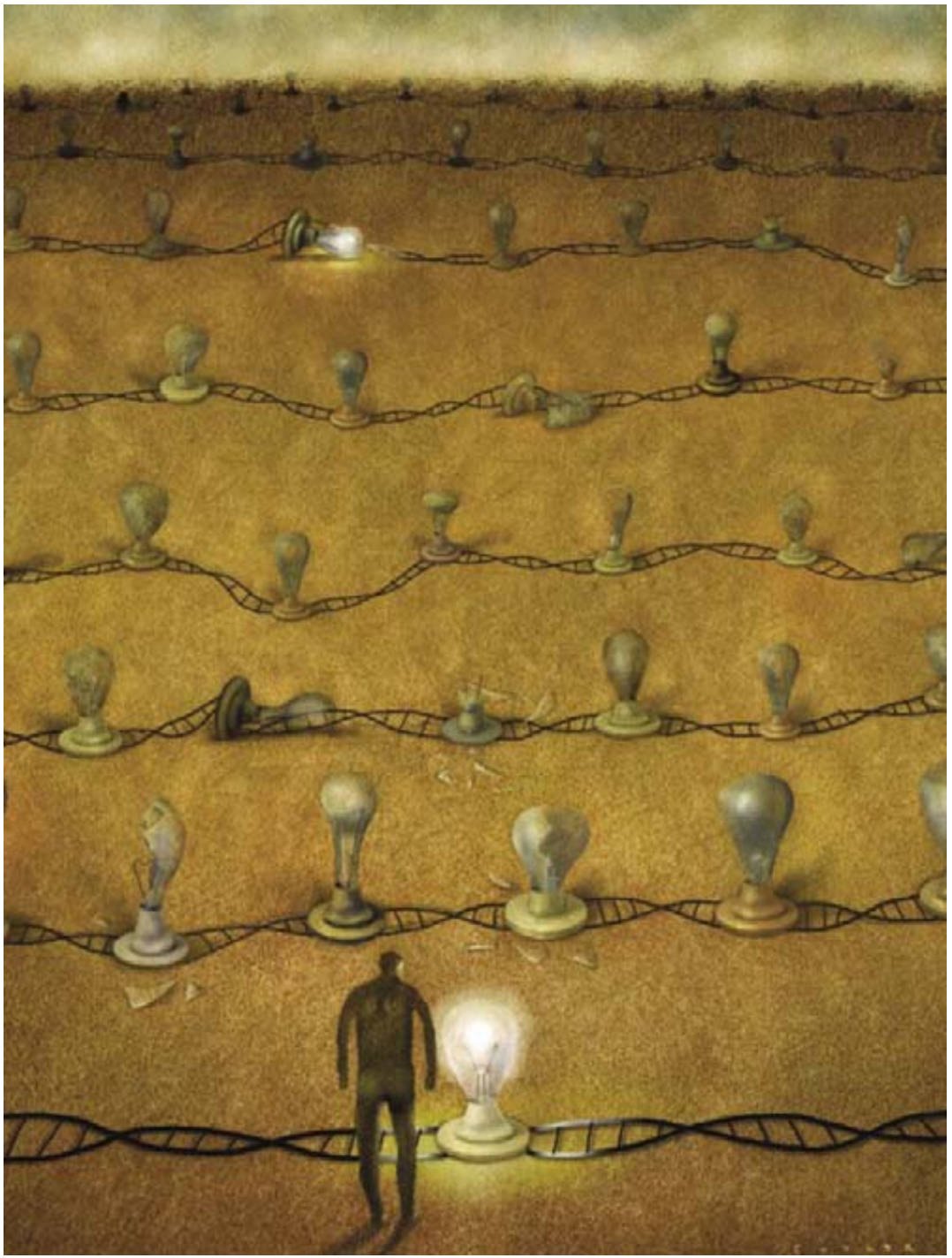
Example predicted structured RNAs (using RNAz)

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

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



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



Integrative Analyses:

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

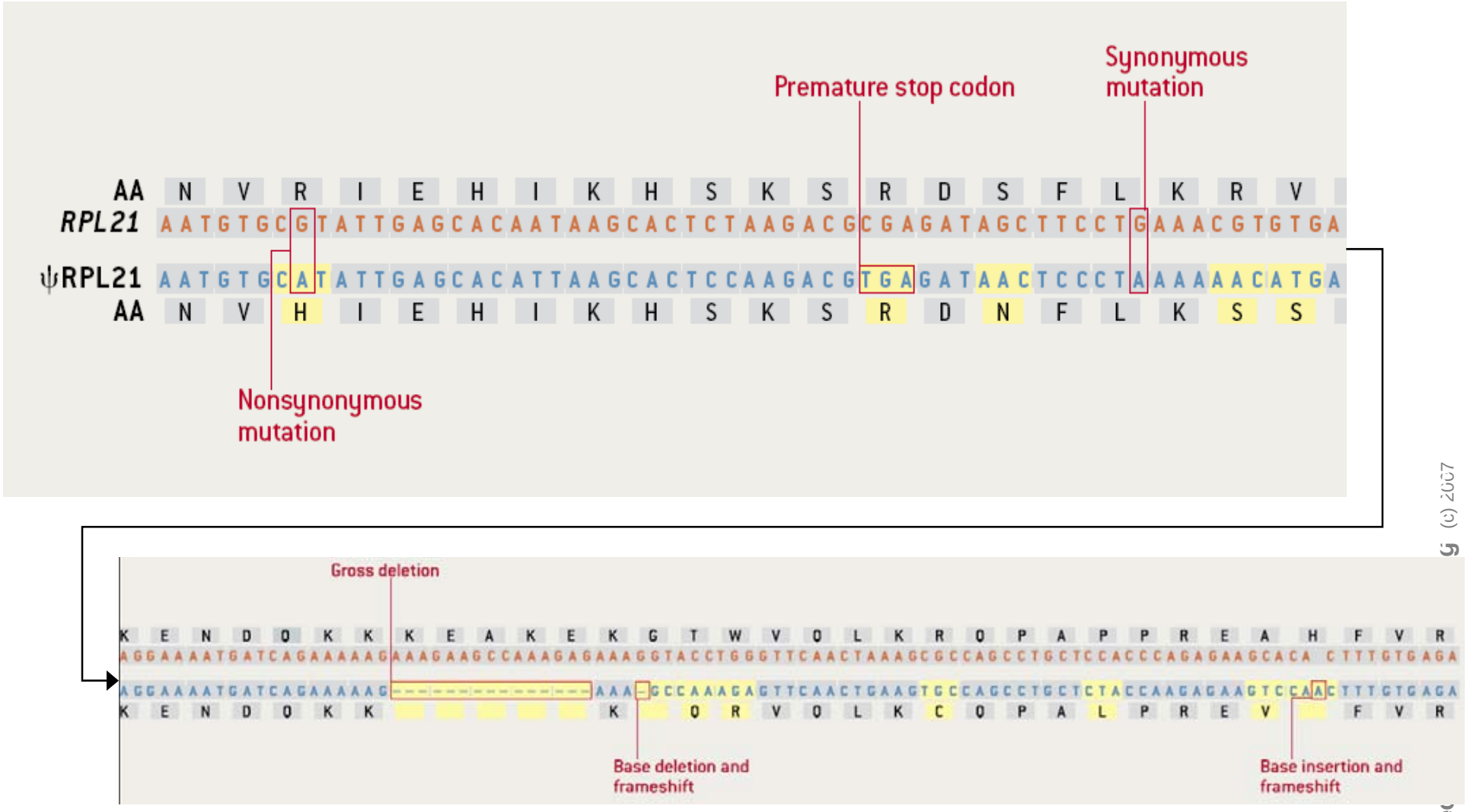
Illustration from Gerstein & Zheng (2006). Sci Am.

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?…

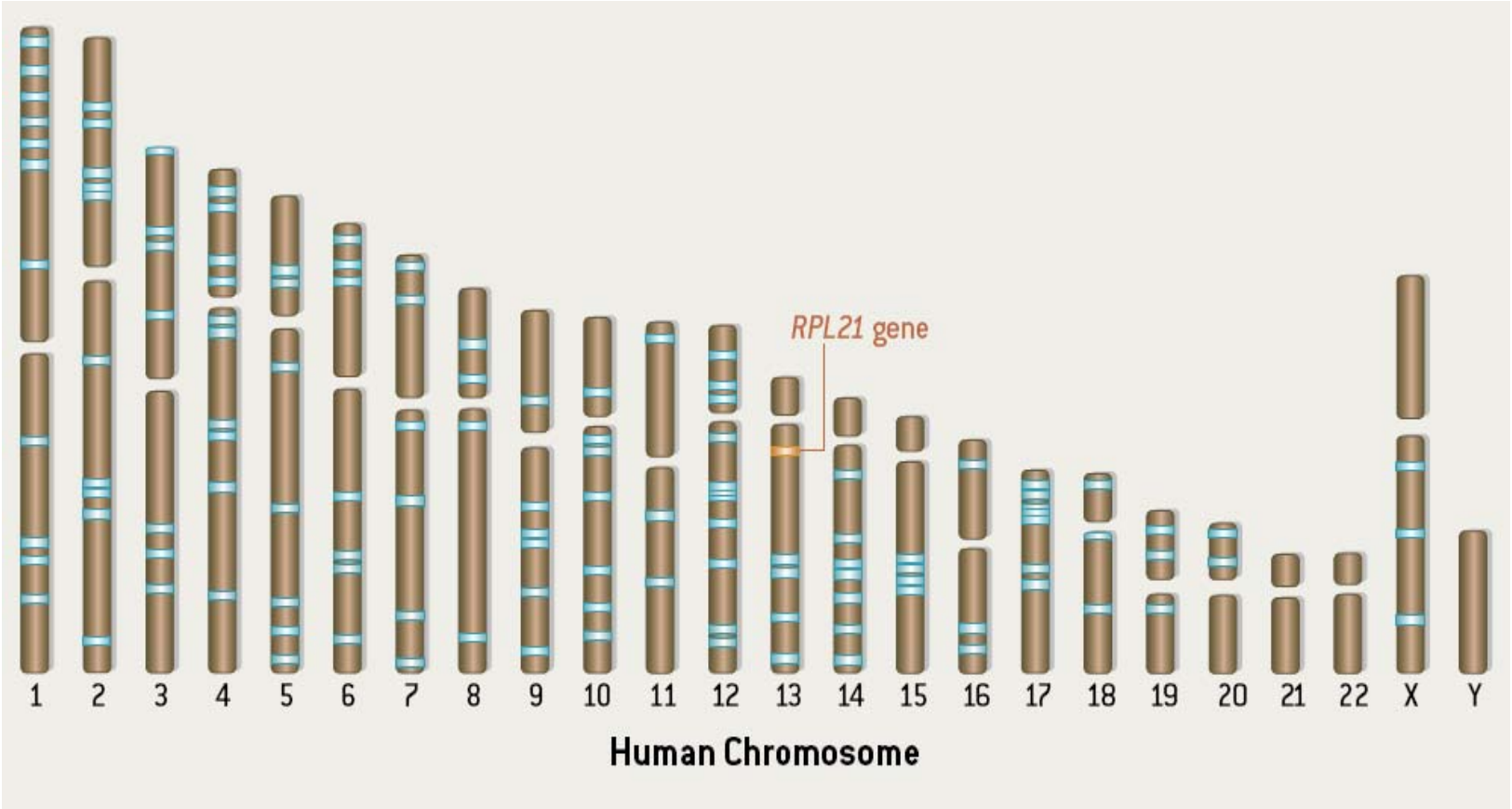
[Mighell et al. *FEBS Letts*, 2000]

Identifiable Features of a Pseudogene (ψRPL21)



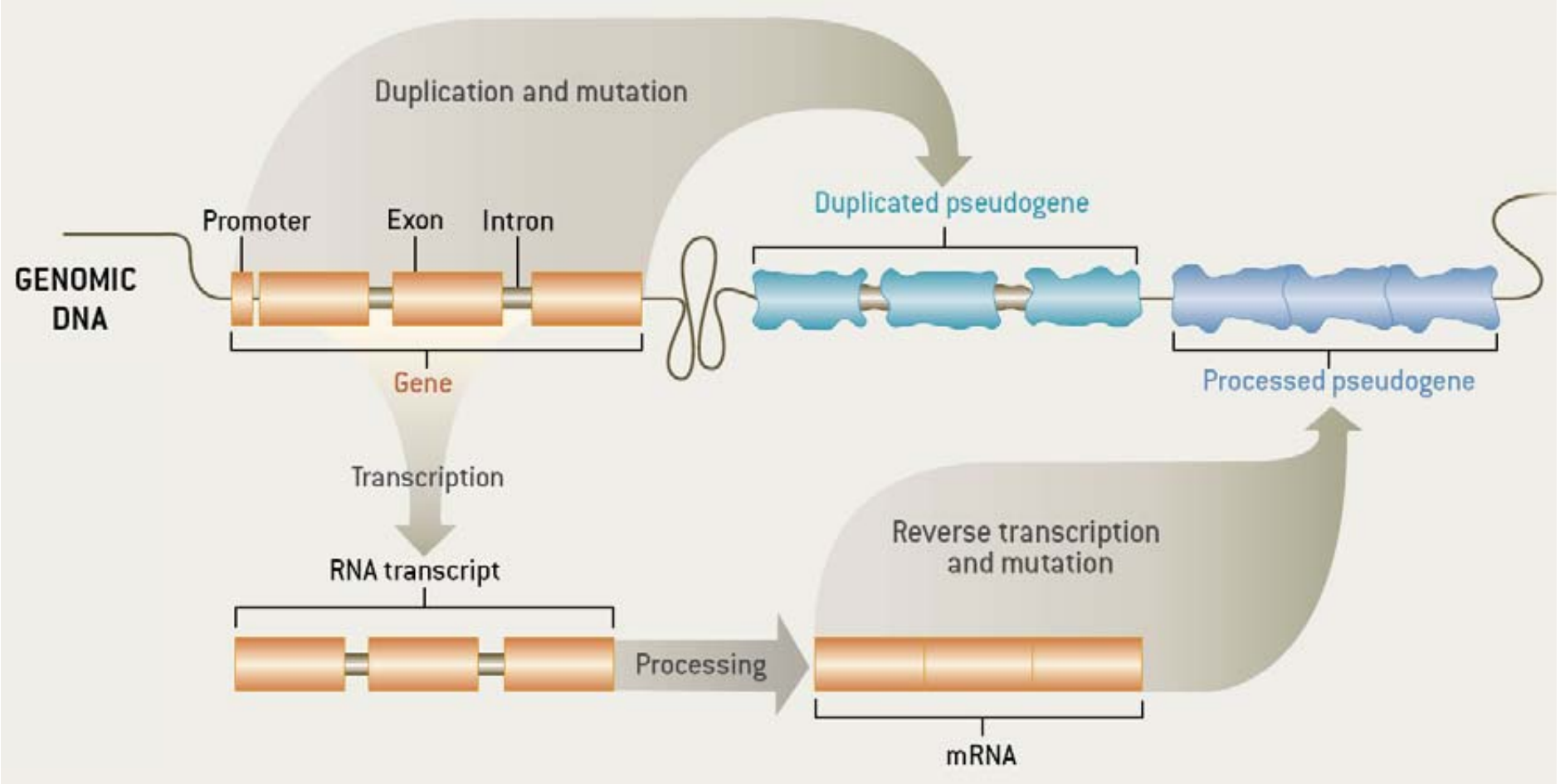
Gerstein & Zheng. Sci Am 295: 48 (2006).

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).

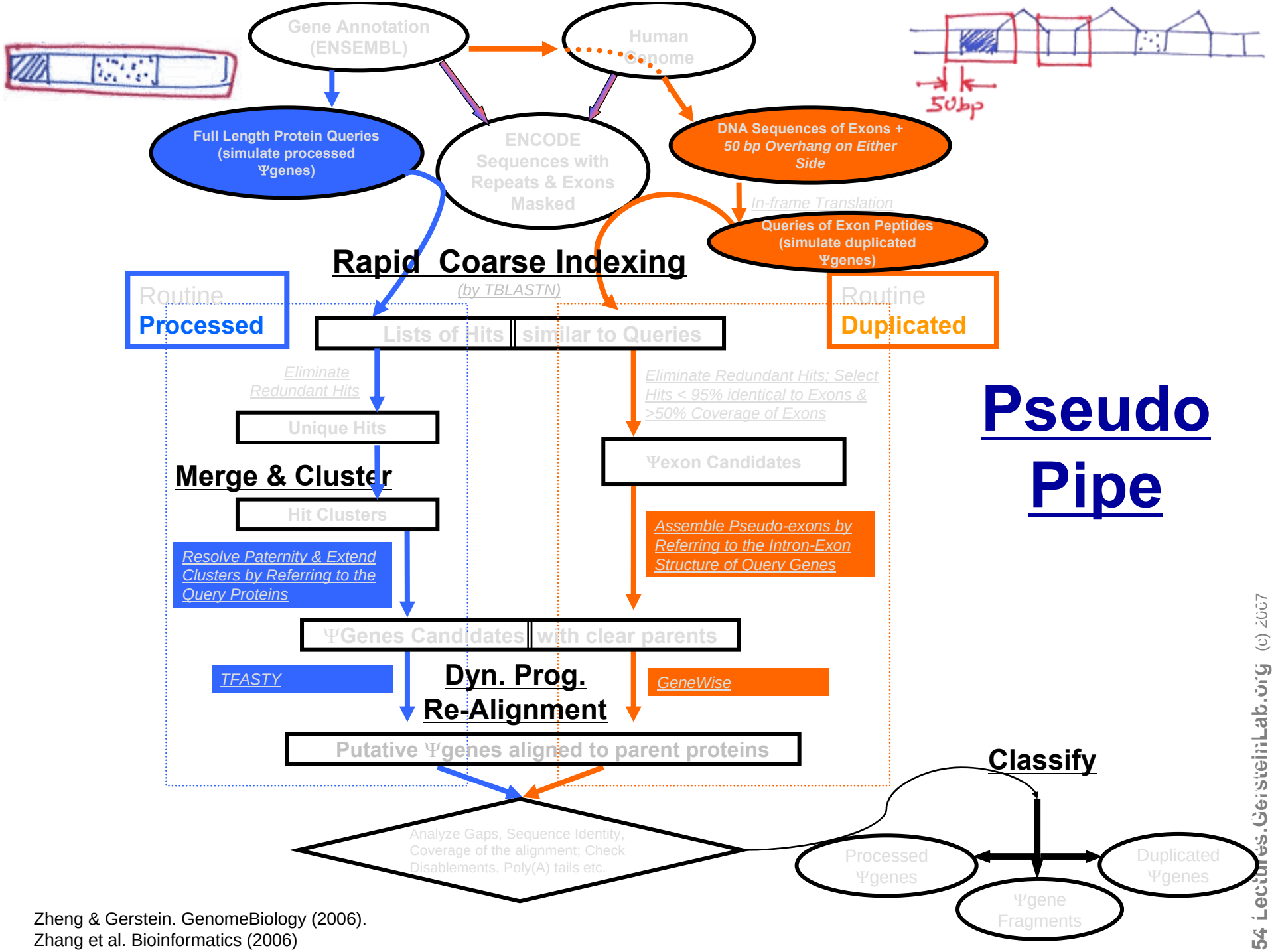
Why Study Pseudogenes?

- 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

Pseudogenes: Tools

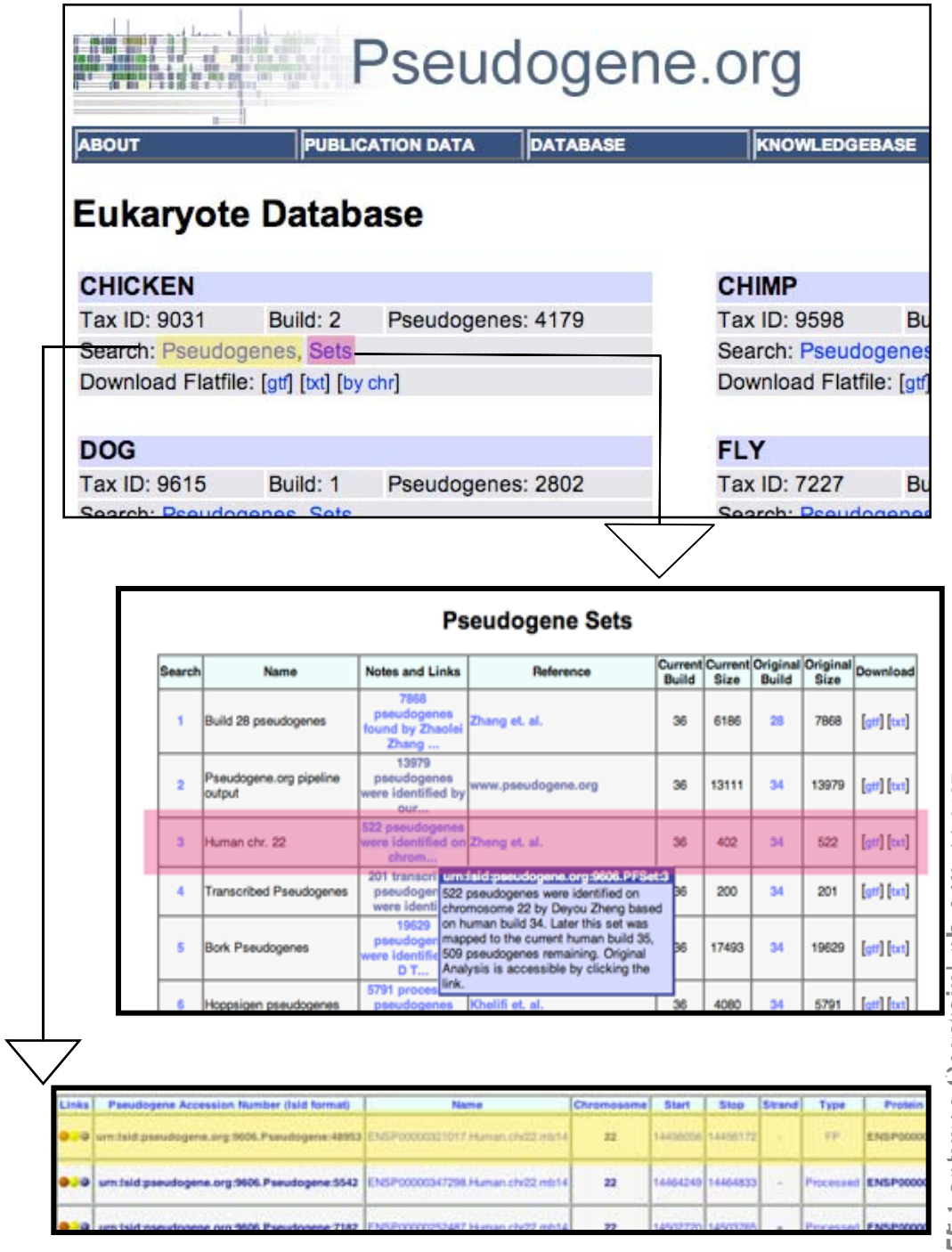


Zheng & Gerstein. GenomeBiology (2006).
Zhang et al. Bioinformatics (2006)

- 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

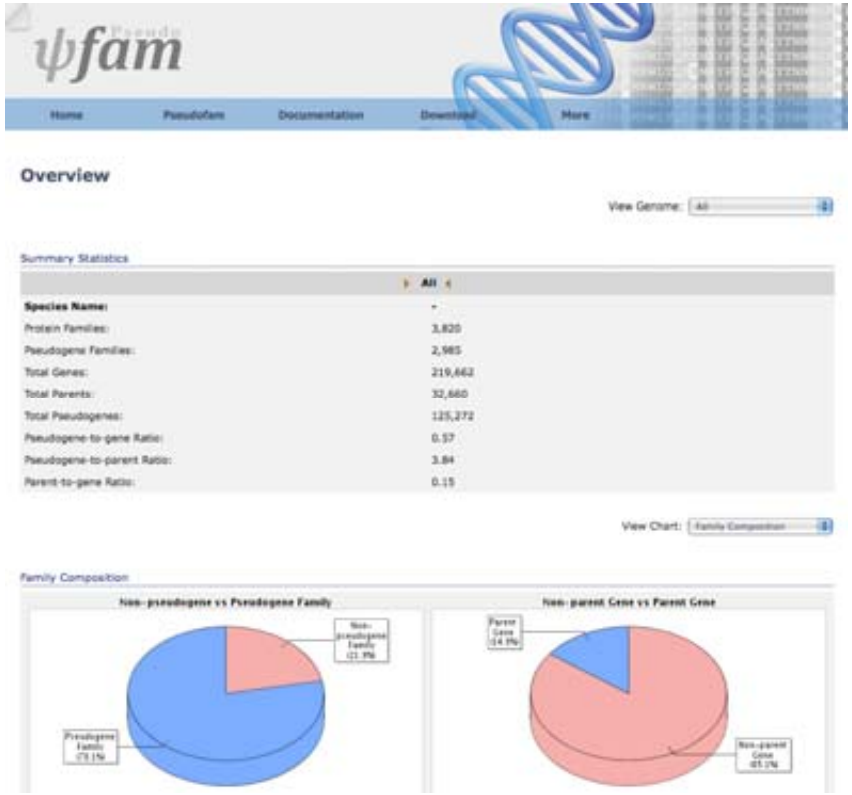
Karro et al., NAR (2007)

DB



Pseudofam

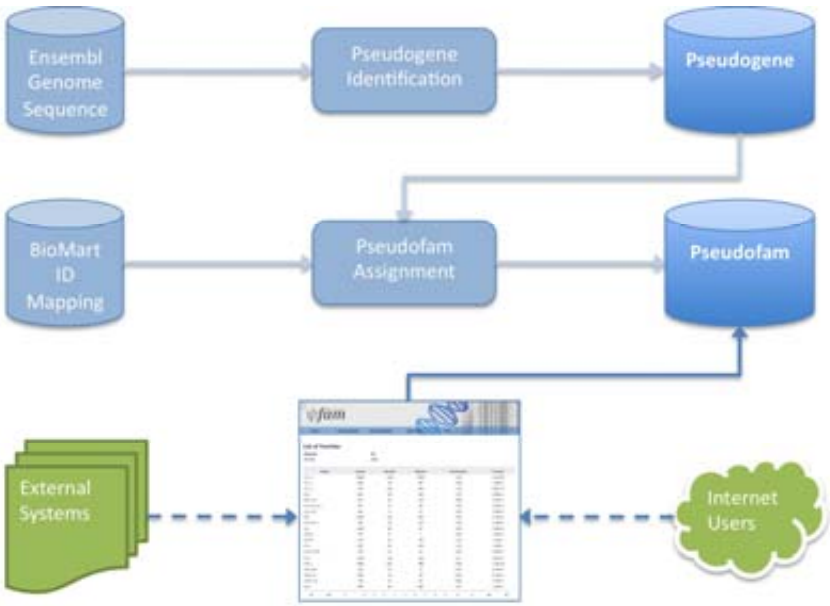
- A Pseudogene Families Database
- Highlighted Features:
  - Browse families
    - Statistics
      - Enrichment
      - P-value, etc
  - Search families
    - Pfam ID/Acc
    - Ensembl ID
    - Pgene ID
  - Correlate families
    - Genes
    - Pgenes
    - Parents



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

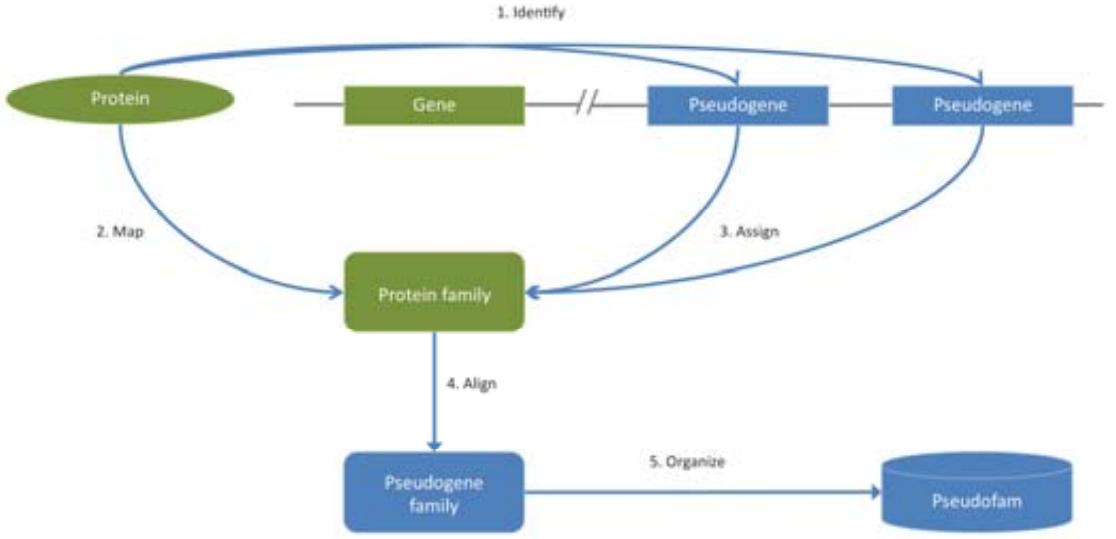
Pseudofam Data Sources

- Ensembl Database
  - Peptides
  - Exons
  - DNAs
- BioMart
  - ENSP-Pfam Mapping
- Pfam
  - Pfam Alignments
- PseudoPipe
  - Pseudogene Identification



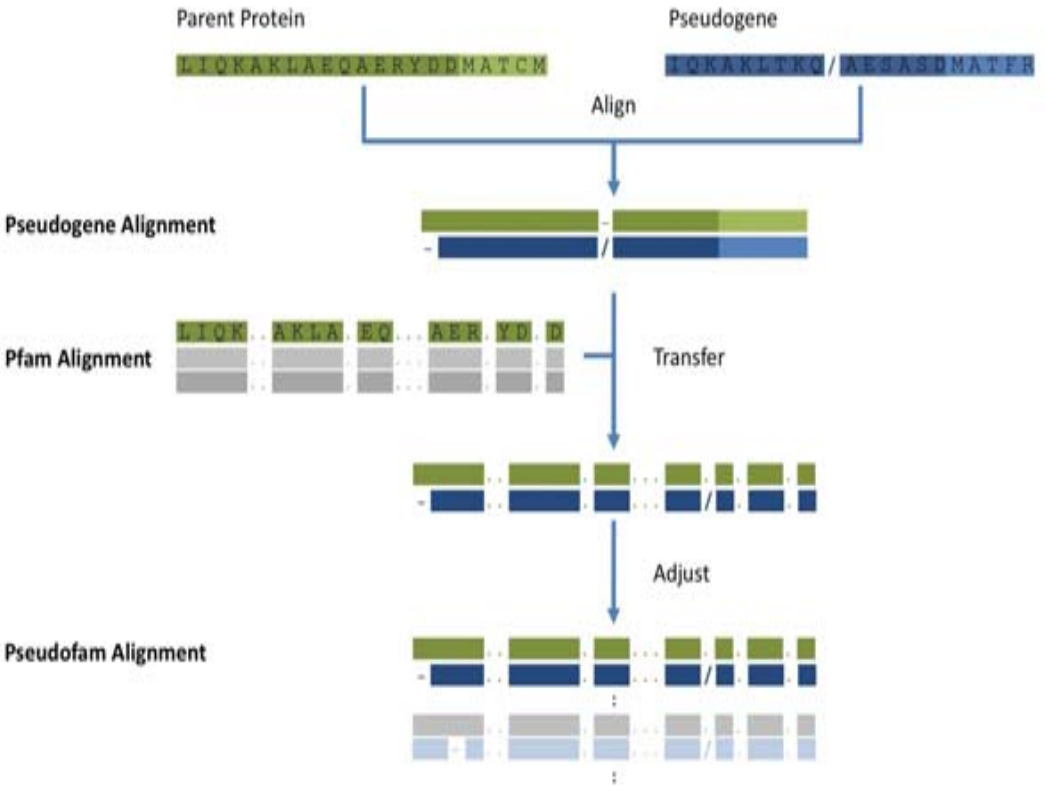
[Lam et al., NAR DB Issue (in press, '09)]

Pseudofam Data Generation & Alignment



- Identify pseudogenes by proteins
- Map parent proteins to protein families
- Assign pseudogenes to their parent families
- Align the pseudogenes in pseudogene families.
- Calculate the key statistics and organize the data into database.

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

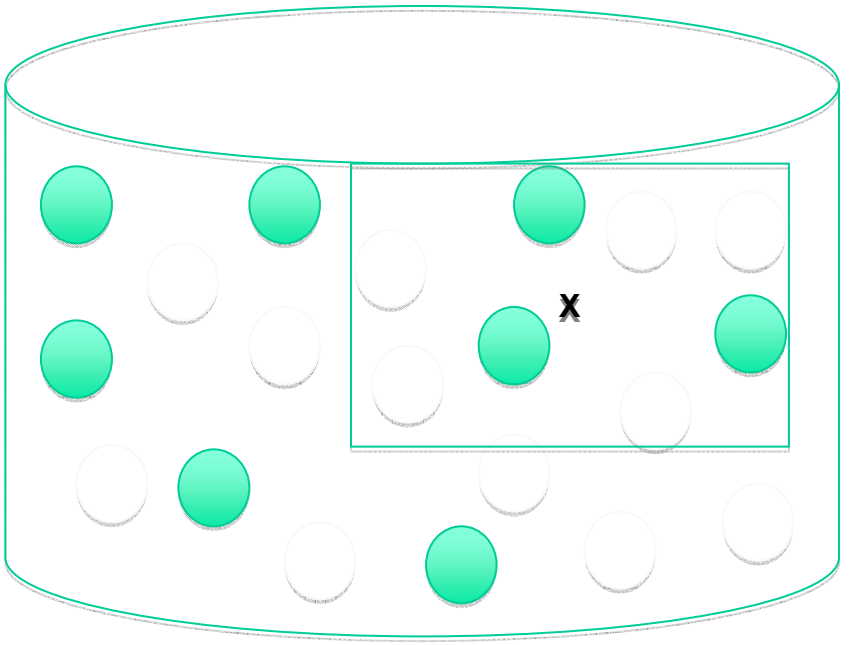


The Enrichment Statistics

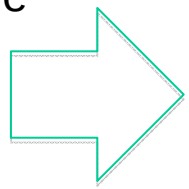
- Hypergeometric Dist.

$$f(k; N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

  - ◊ k: parent genes in a fam
  - ◊ n: genes in a fam
  - ◊ m: overall total parent genes
  - ◊ N: overall total genes



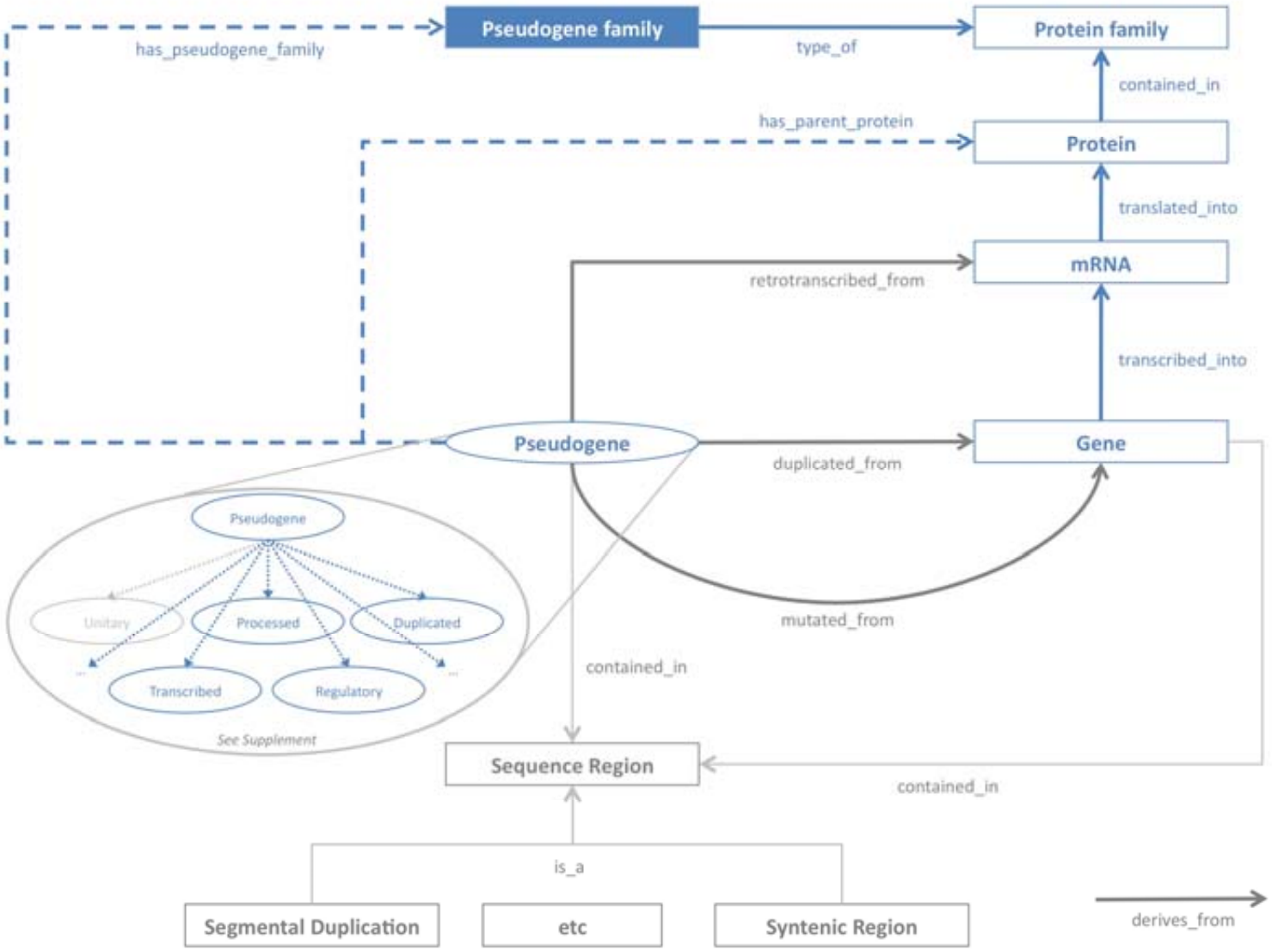
- For example, in fam x:
  - ◊ 3 parent genes in x
  - ◊ 5 genes in x
  - ◊ 8 total parent genes in genome
  - ◊ 20 genes in genome



  - ◊ Expected parents: 2
  - ◊ Enrichment: 3/2=1.5
  - ◊ P-value: Pr(k>=3)=0.3

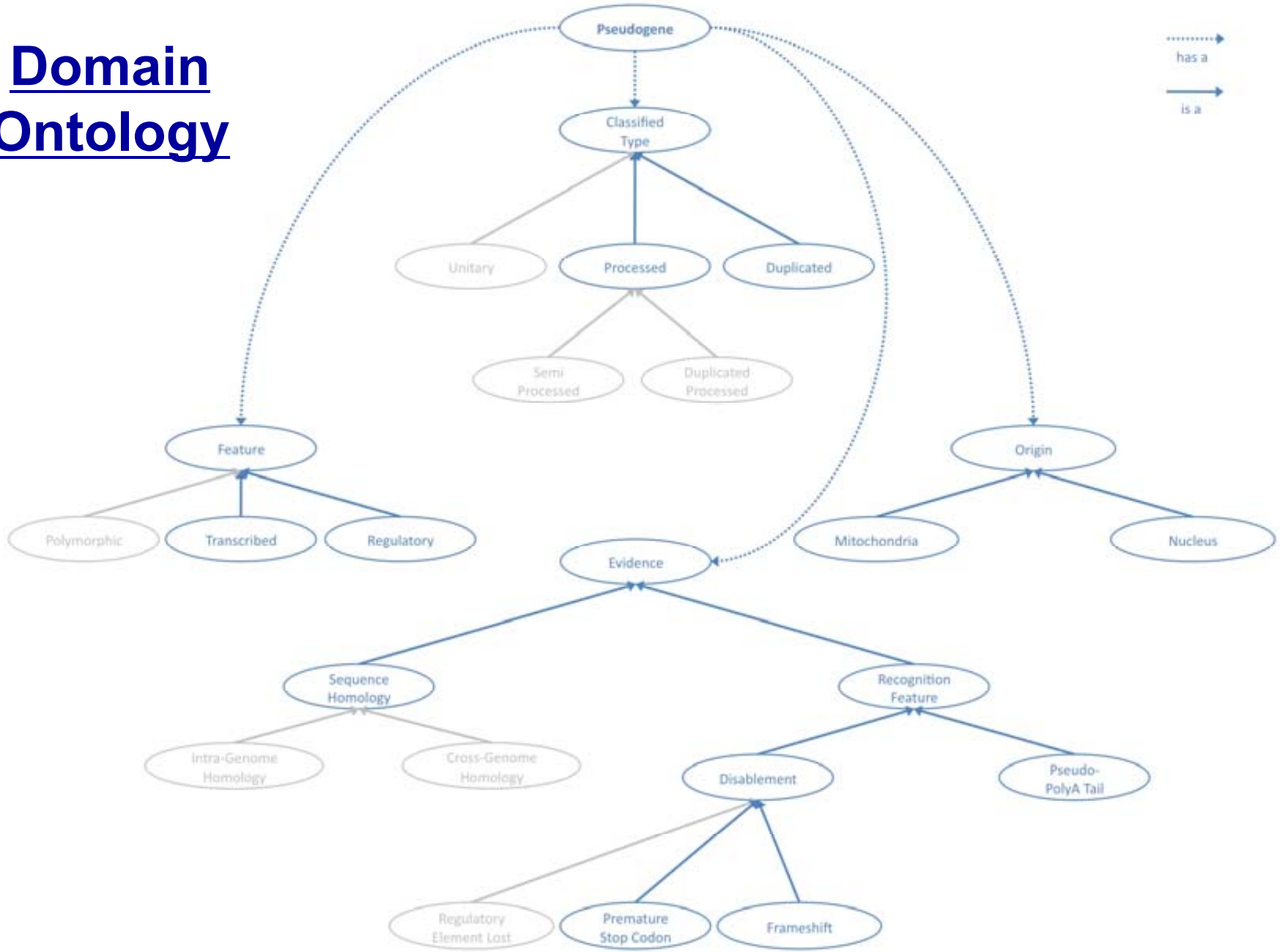
[Lam et al., NAR DB Issue (in press, '09)]

Upper Ontology



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

Domain Ontology



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

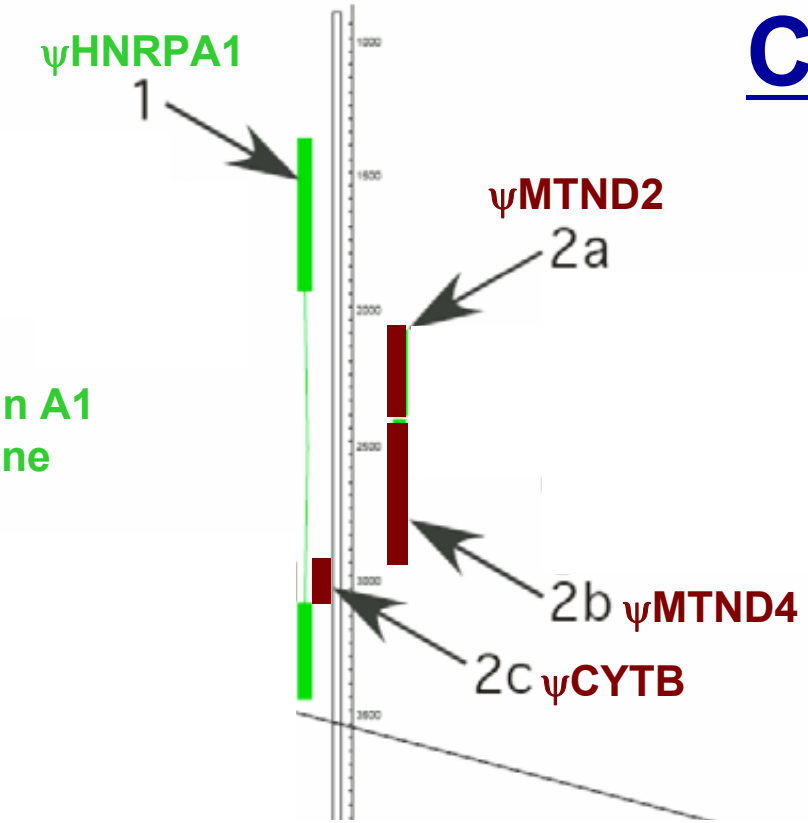
Pseudogene Assignments to the Human Genome

- Draft Assignment to Whole Genome
  - ◊ Automatic pipeline currently gives ~22644 [pseudogene.org]
- Eventual Annotation Process in framework of ENCODE project
  - ◊ Hybrid
  - ◊ Pipeline runs, manual curation flagging difficult cases, pipeline improvement, further curation...
- Pilot ENCODE Annotation
  - ◊ 4 automatic pipelines: retroFinder+pseudoFinder (UCSC), PseudoPipe (Yale), GIS
  - ◊ HAVANA manual
  - ◊ 201 pseudogenes vs ~400 genes

- Issues in Pilot
  - ◊ Comparing protein or transcript v genomic DNA, filtering, application of rules
  - ◊ What is a pseudogene?
  - ◊ Different criteria
  - ◊ Conservative approach here
  - ◊ Can't overlap gene annotation
  - ◊ Need to have a protein alignment

Zheng et al. (2007) Gen. Res.

Complexities in Pseudogene Annotation

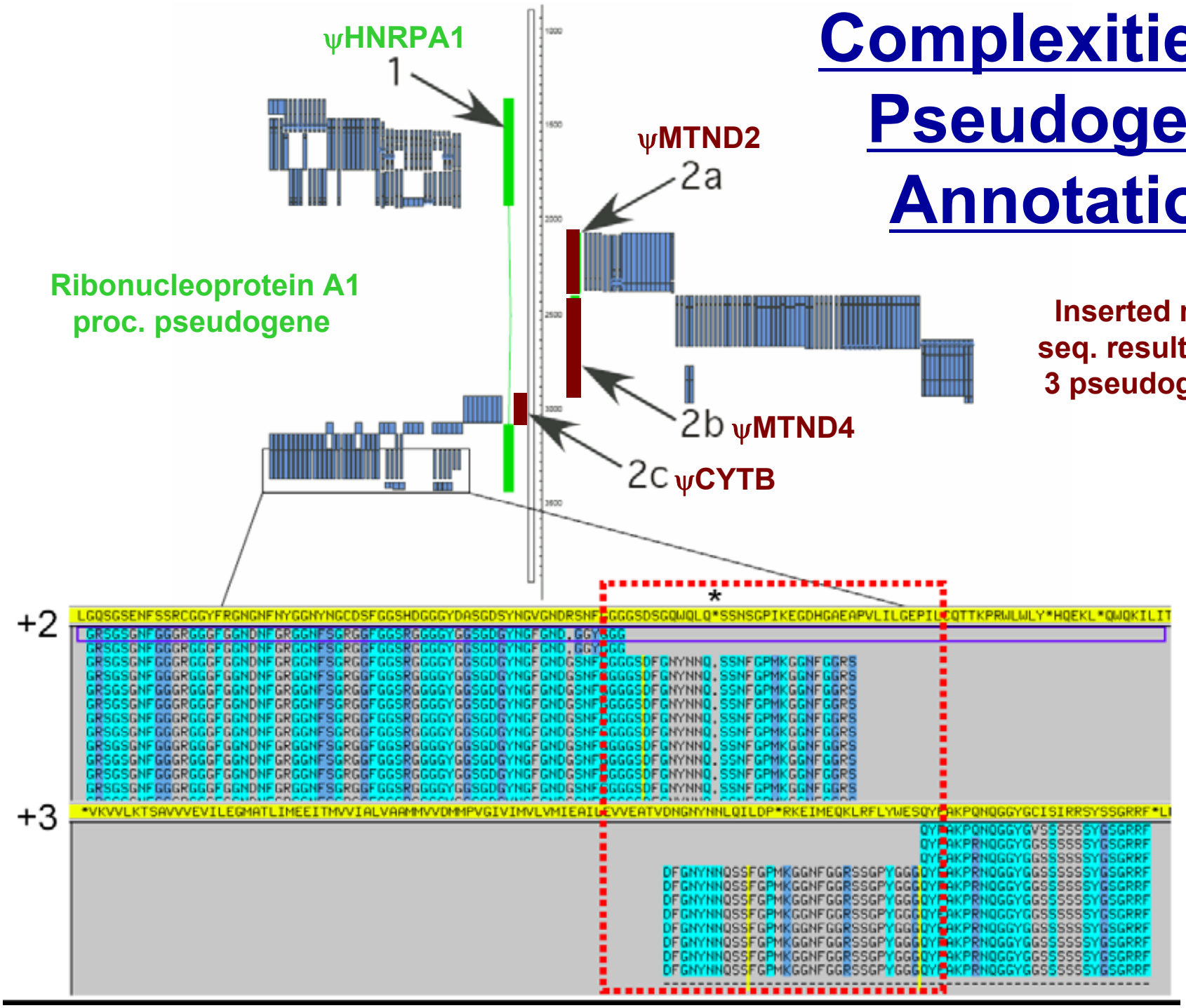


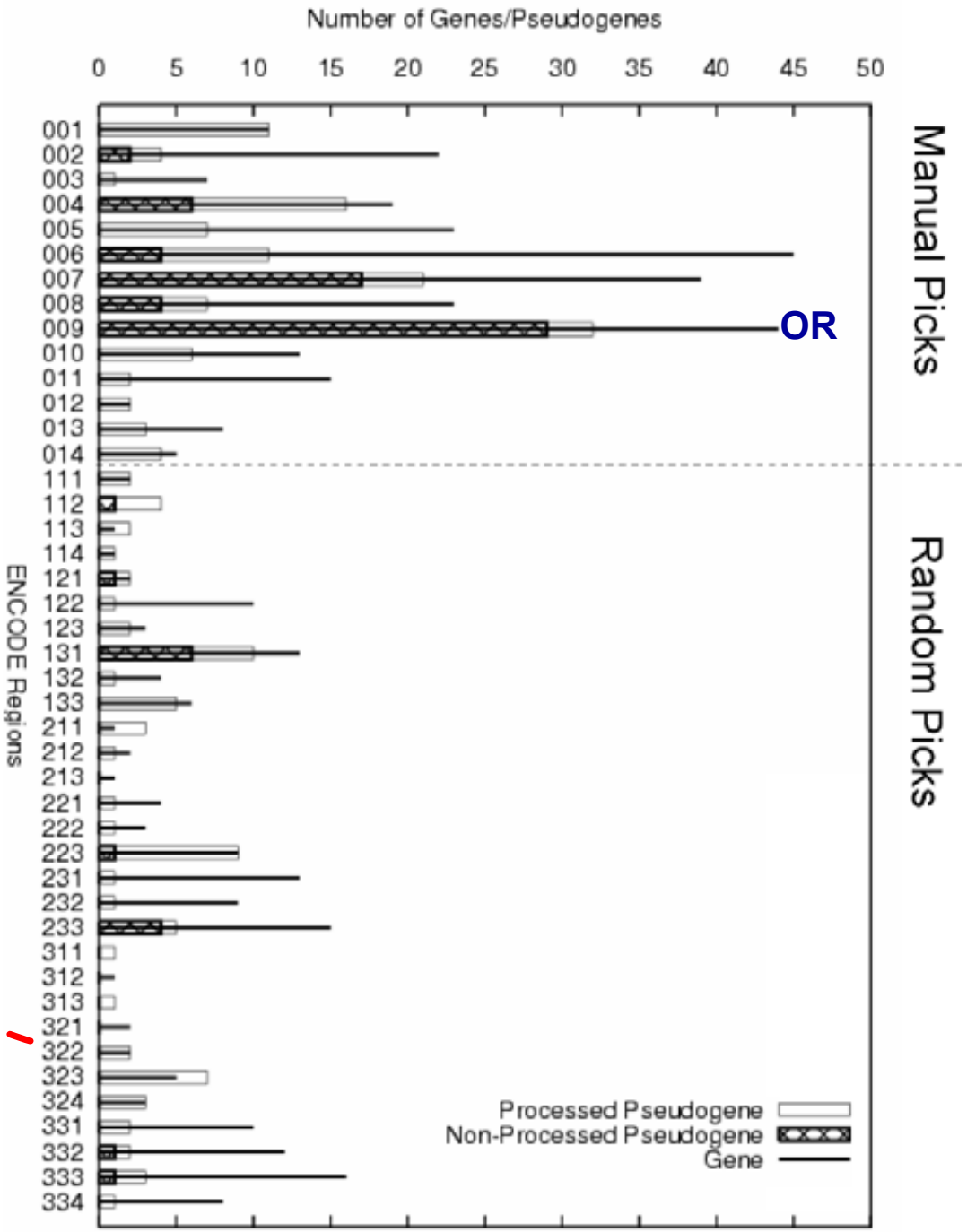
Ribonucleoprotein A1 proc. pseudogene

Inserted mito. seq. resulting in 3 pseudogenes

Zheng et al. (2007) Gen. Res.

Complexities in Pseudogene Annotation





Overall Results: Regional Distribution

201 pseudogenes
77 non-processed
124 processed

Zheng et al. (2007) Gen. Res.

browser
+
pseudogene.org/ENCODE

Integrated

Vast Amounts of Different Data Types to Integrate in pilot ENCODE

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

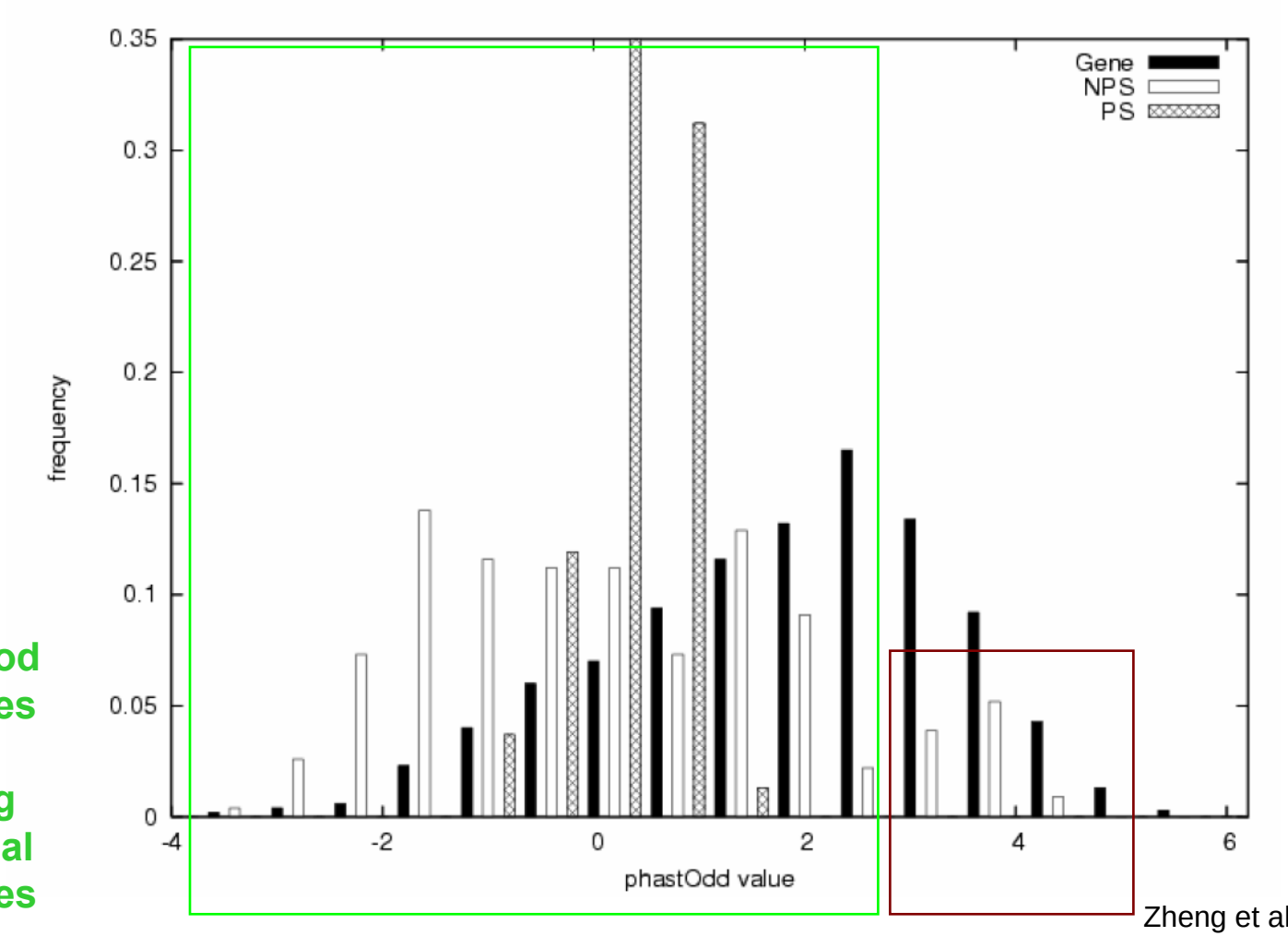
- Large-scale sequence comparison in relation to the human genome

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]

Integration of Different Types of Annotation: Pseudogenes with Sequence Constraint

Using phastOdd value to examine neutral evolution of pseudogenes

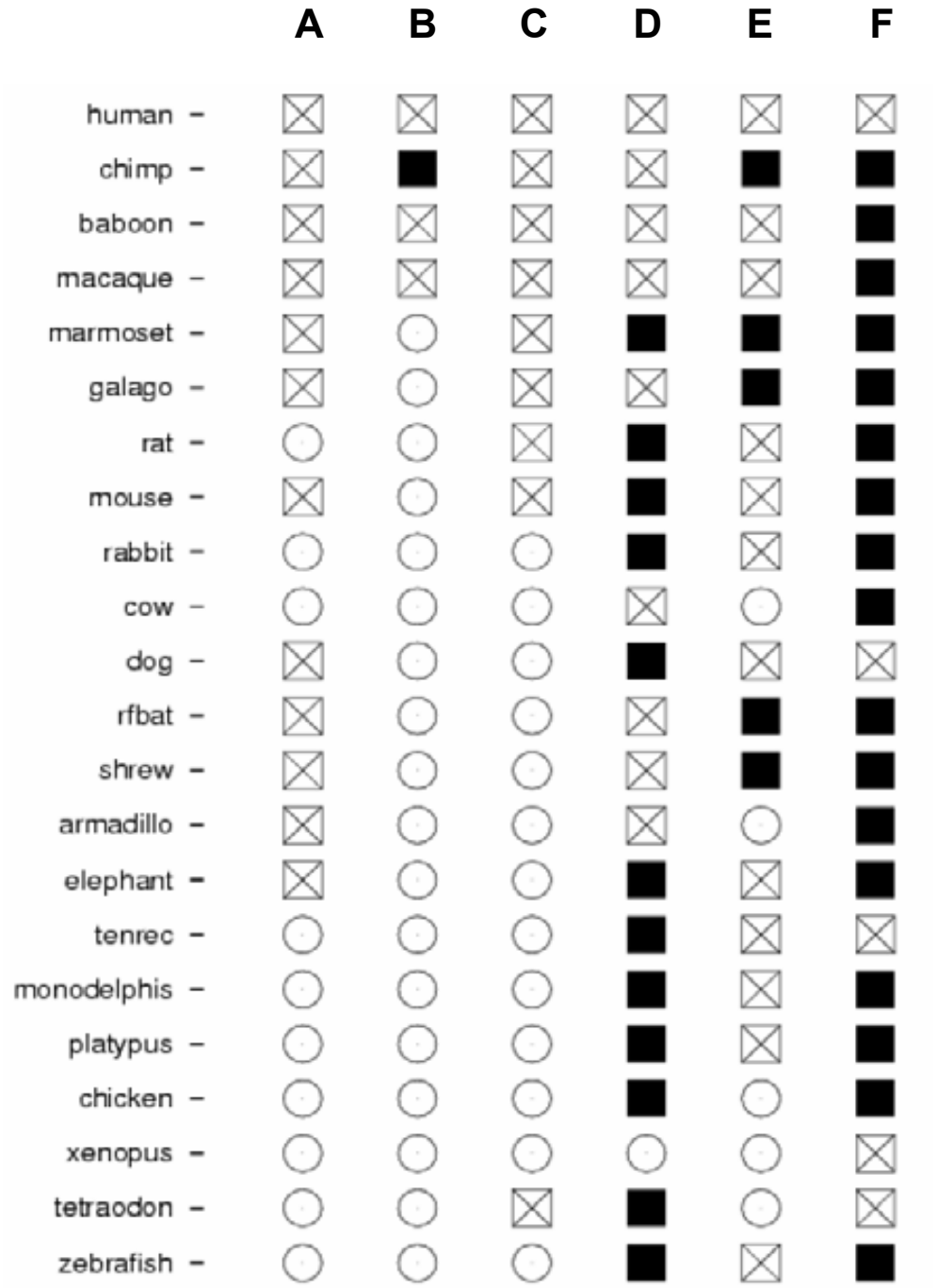


most good candidates for studying mutational processes

a few non-proc. ψG under constraint

Zheng et al. (2007) Gen. Res.

representative pseudogenes drawn from 201 total



History of Pseudogene Preservation

Based on alignment from ENCODE MSA group

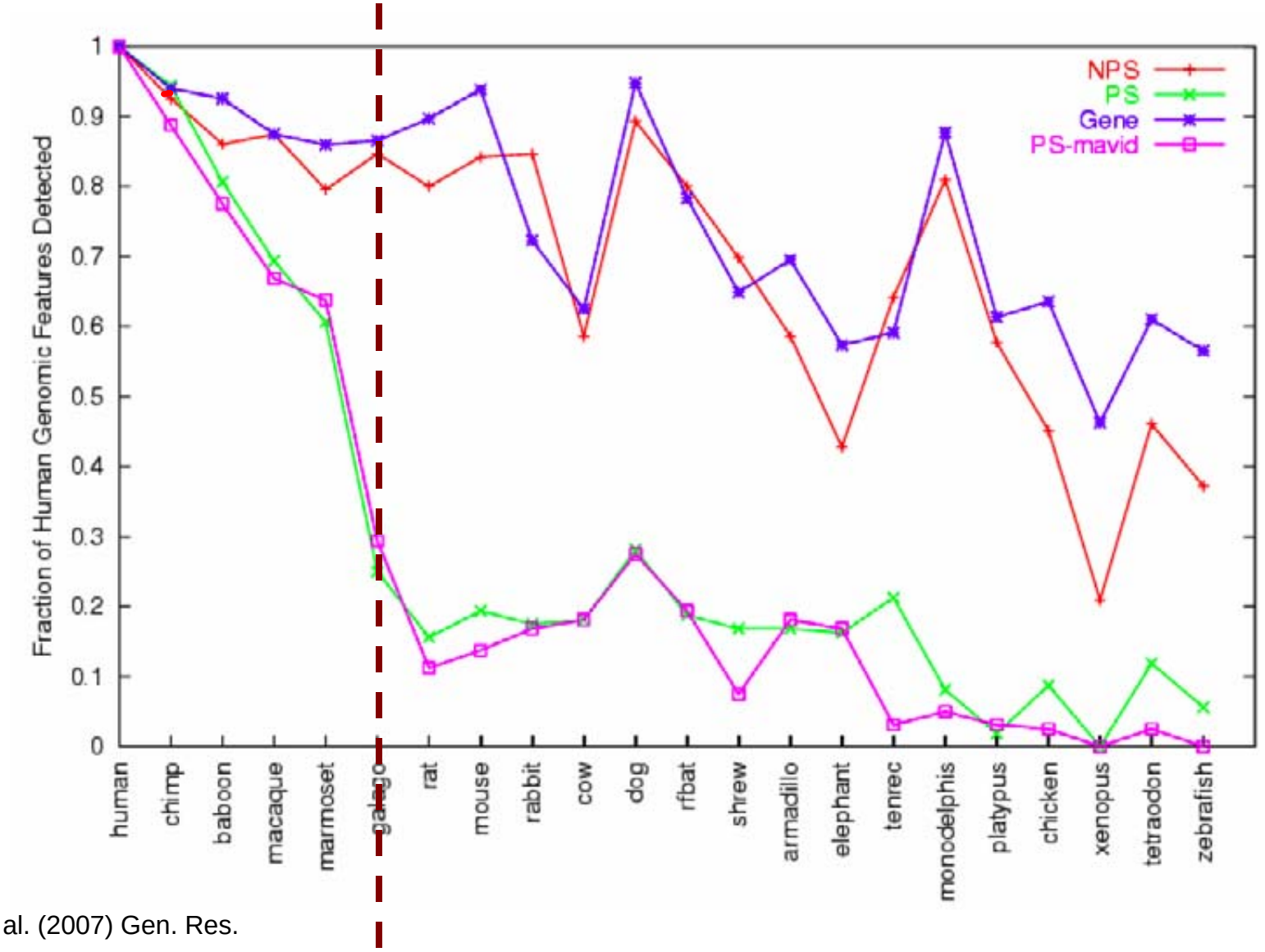
Zheng et al. (2007) Gen. Res.

Absent

Present with Disablement

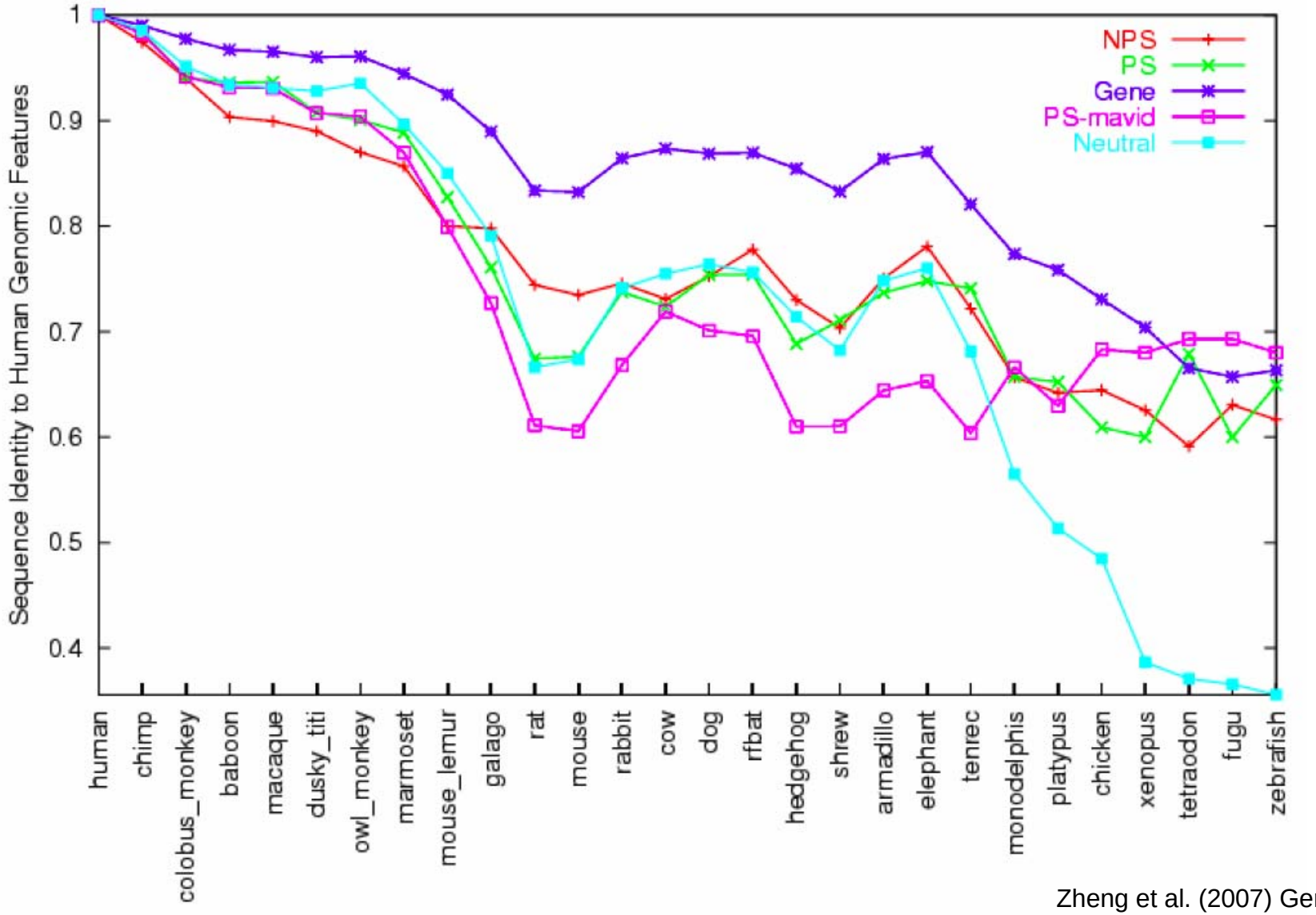
Present without Disablement

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



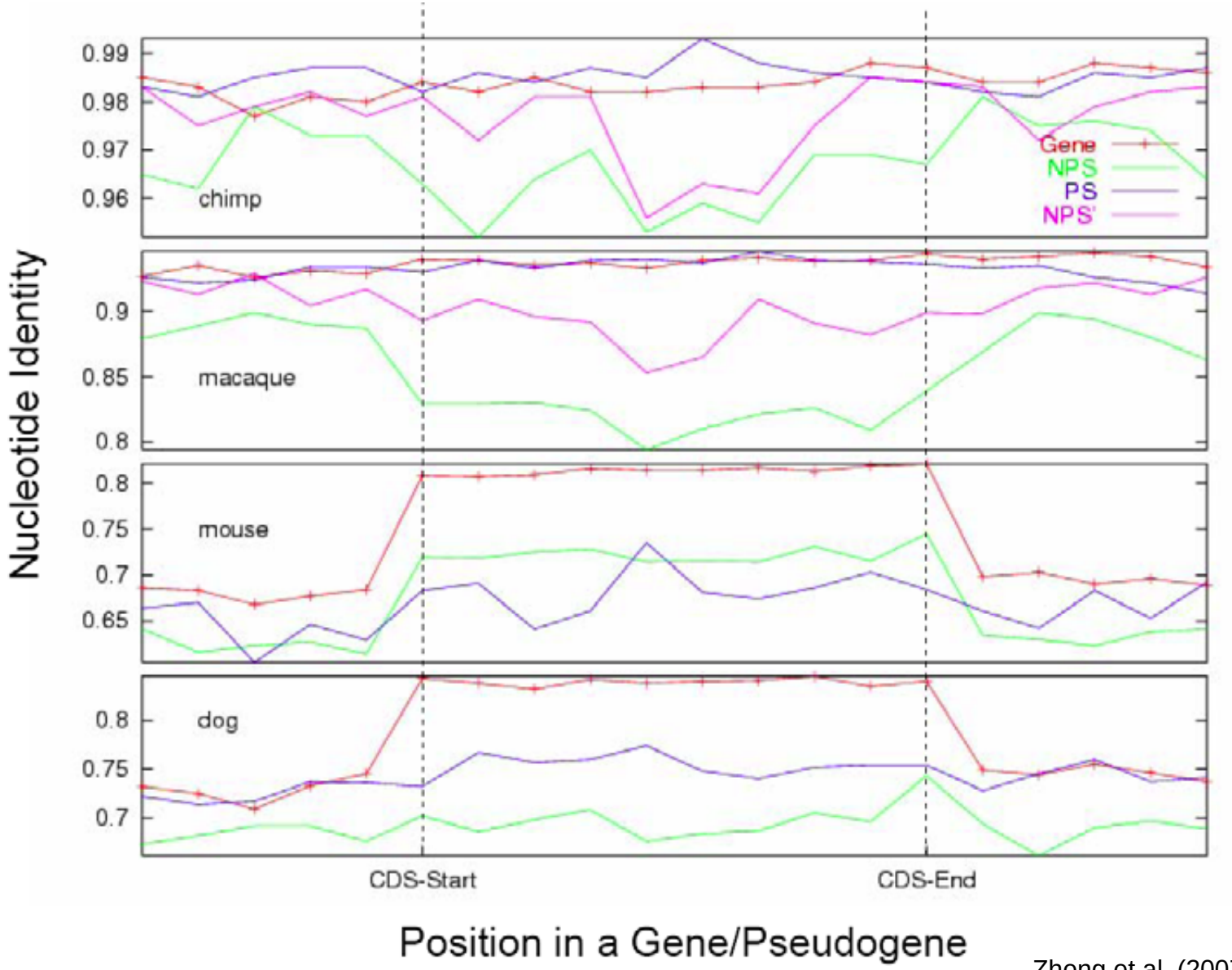
Zheng et al. (2007) Gen. Res.

Sequence Decay of Pseudogenes, Approximately Neutral



Zheng et al. (2007) Gen. Res.

Sequence Decay of Pseudogenes Relative to their Immediate Genomic Context

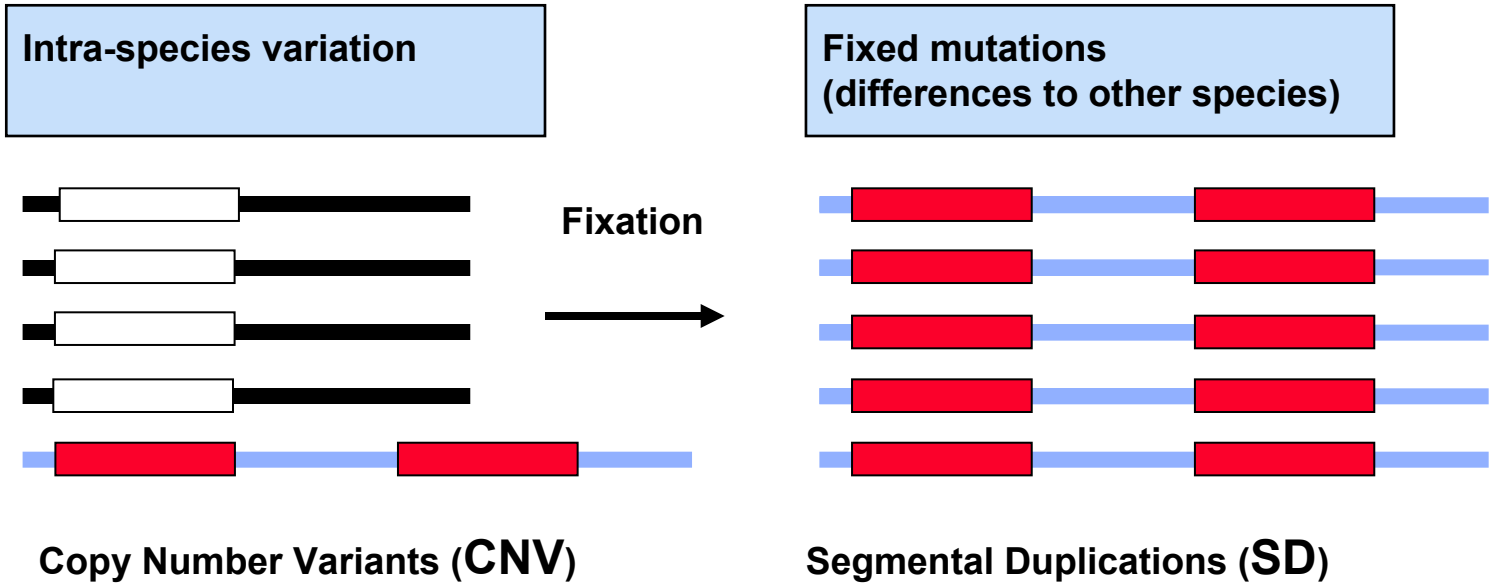


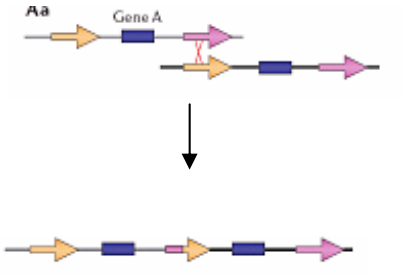
Zheng et al. (2007) Gen. Res.

Analyzing Repeated Blocks in the Genome (SDs & CNVs)

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

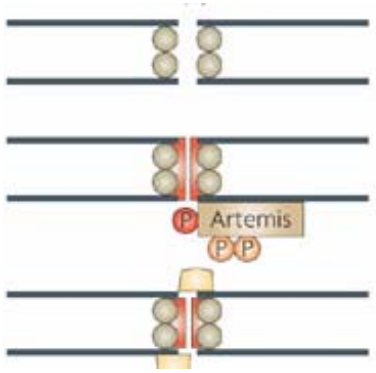
SEGMENTAL DUPLCATIONS AND COPY NUMBER VARIANTS ARE RELATED PHENOMENA AND HAVE BEEN CREATED BY SEVERAL DIFFERENT MECHANISMS





NAHR
(Non-allelic homologous recombination)

Flanking repeat
(e.g. Alu, LINE…)



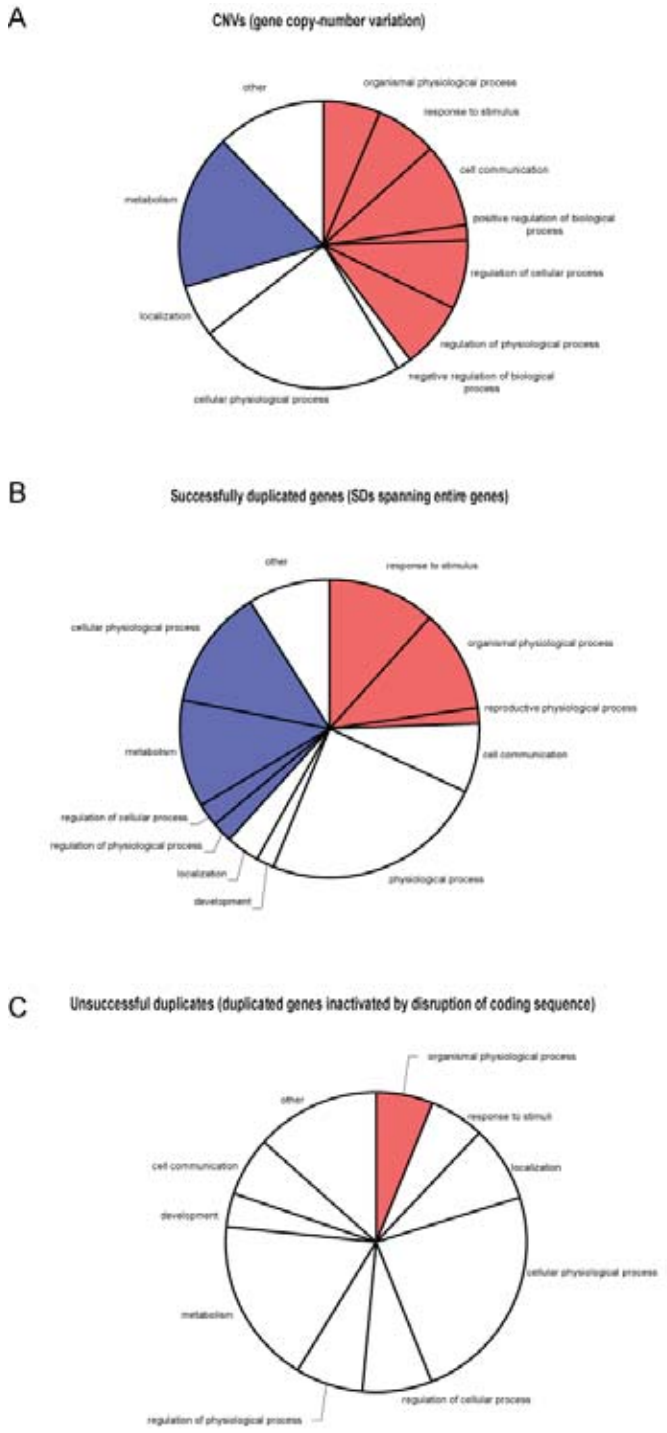
NHEJ
(Non-homologous-end-joining)

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

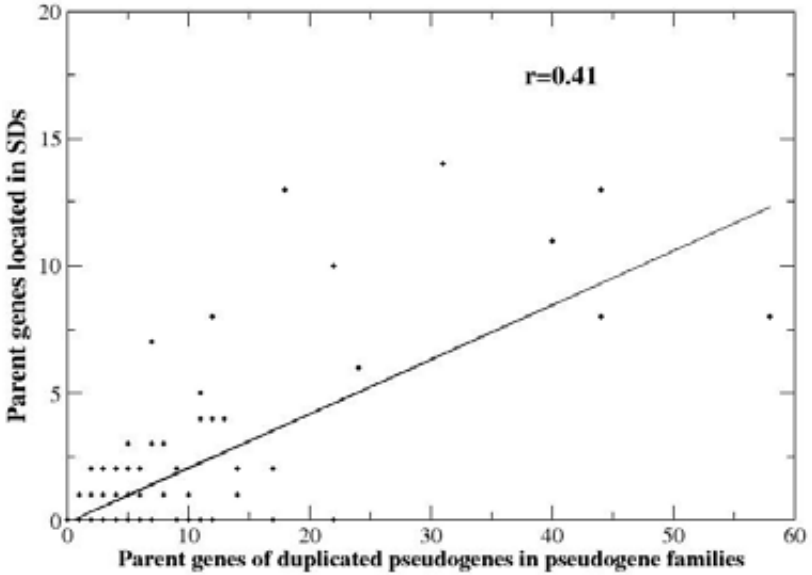
Association of SDs and CNVs with pseudogenes

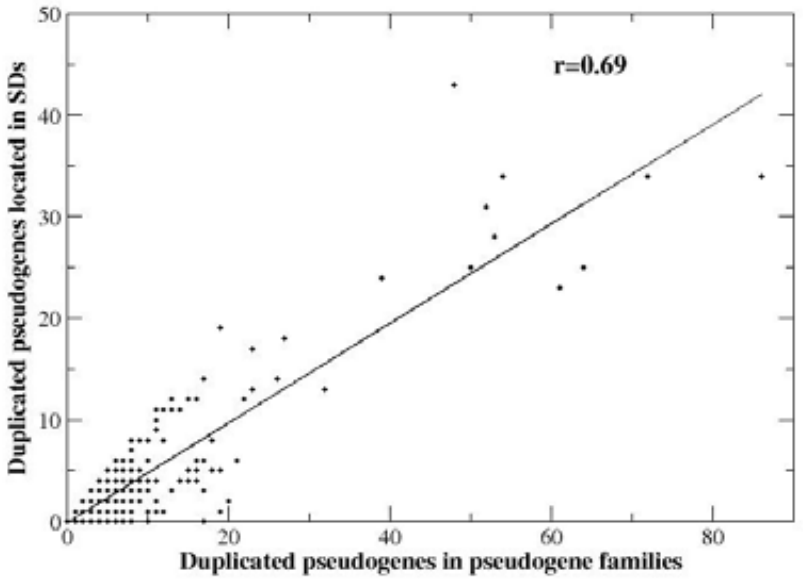
- CNVs are the raw form of variation producing duplicated elements
- Segmental Duplications (SDs) are fixed forms of CNVs/SVs. They give rise to duplicated genes and (eventually) protein protein families
- Thus, we expect, duplicated pseudogenes (failed duplications) to occur in SDs.
- CNVs and SDs tend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

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



Pseudogene families and Segmental Duplications (SDs)



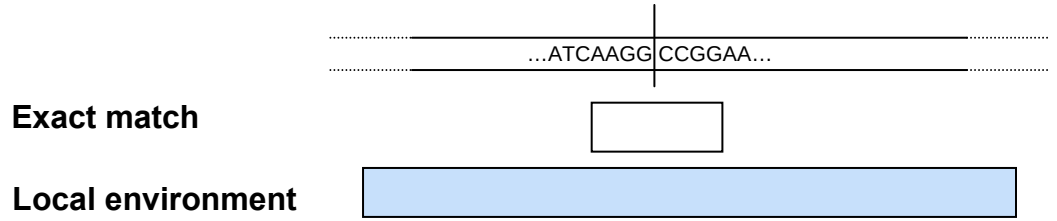


- SDs comprise ~5-6% of the human genome but contain ~17.8% genes, 45.7% duplicated pgenes and 21.6% processed pgenes
- Relative values of correlation coefficients in the plots above consistent with the observation that SDs contain more pgenes than parent genes

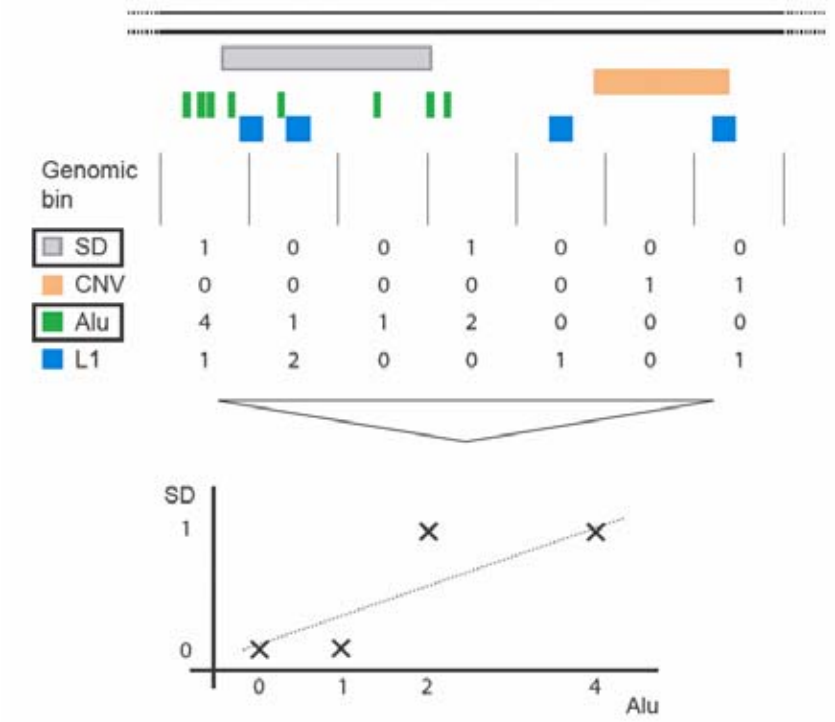
[Lam et al., NAR DB Issue (in press, '09)]

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

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

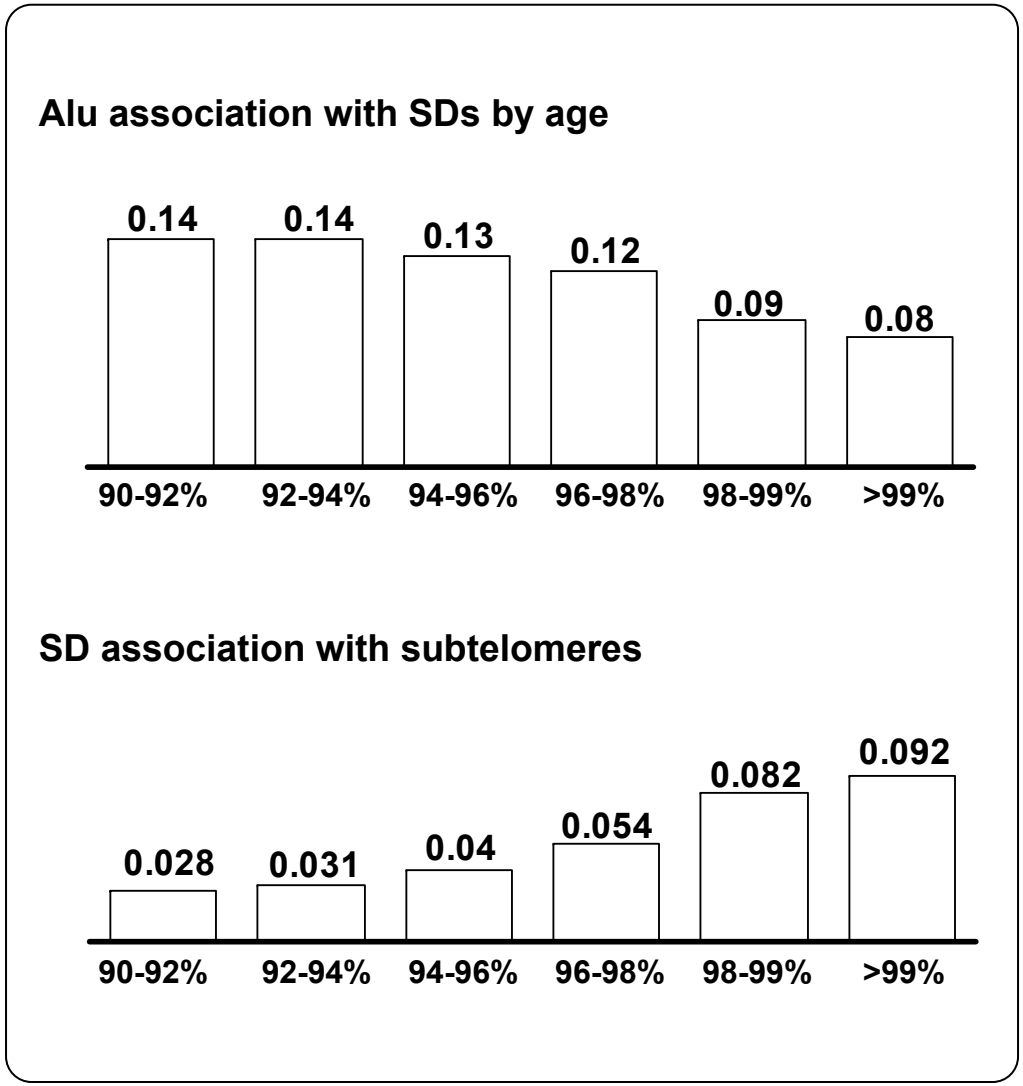


1. Survey a range of genomic features
2. Count the number of features in each genomic bin (100kb)
3. Calculate correlations / enrichments using robust stats



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

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

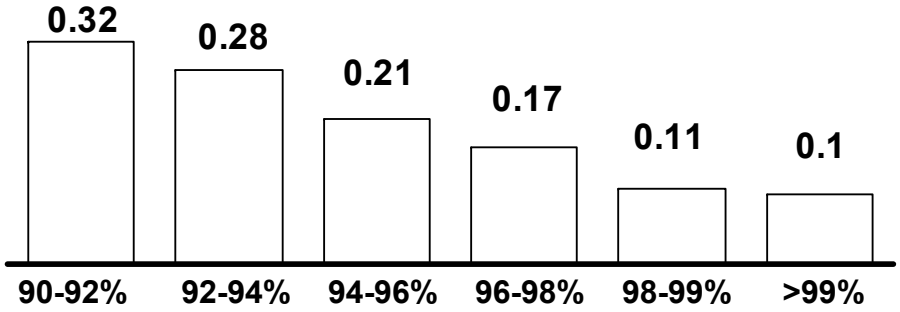


- The co-localization of Alu elements with SDs is highly significant.
- Older SDs have a much higher association with Alus than younger SDs.
- Hence it is likely, that Alu elements were more active in mediating NAHR in the past (consistent with the Alu burst)
- Younger SDs are more likely to be localized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

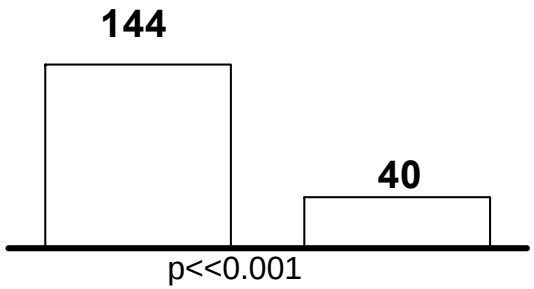
[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

ANOTHER FUNCTION FOR PSEUDOGENES: SERVING AS REPEATS FOR MEDIATING NAHR

Processed pseudogene association with SDs by age

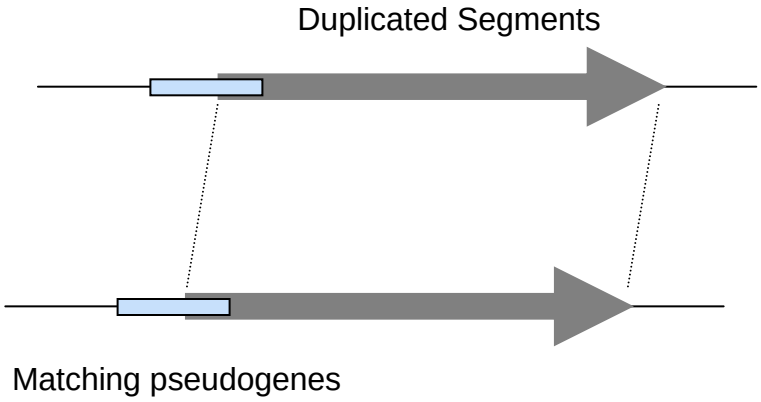


Processed pseudogenes at SD junctions



No. of SDs with matching pseudogenes at matching junctions

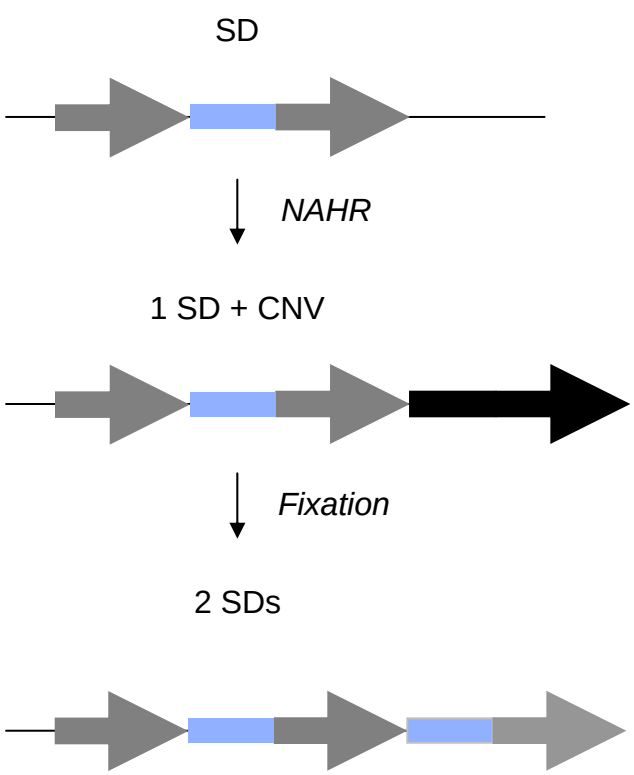
Number of matching pseudogenes expected at random



[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

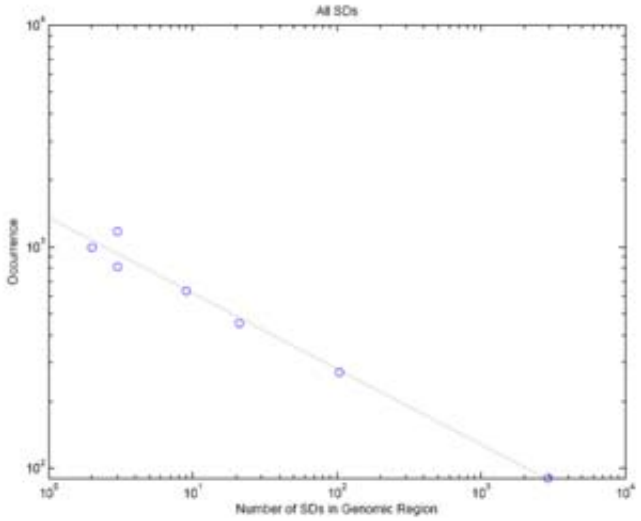
Hypothesis



"SD selfpropagation"

Corollary

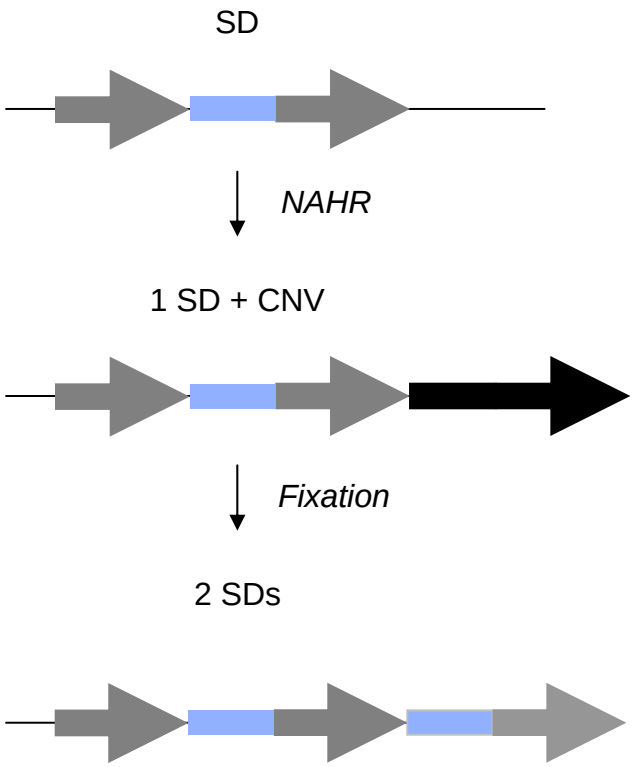
- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- Such mechanisms ("preferential attachment") are well studied in physics and should leads a very skewed ("power-law") distribution of SDs.



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

FOCUSSING ON SDS: SDs COLOCALIZE WITH EACH OTHER

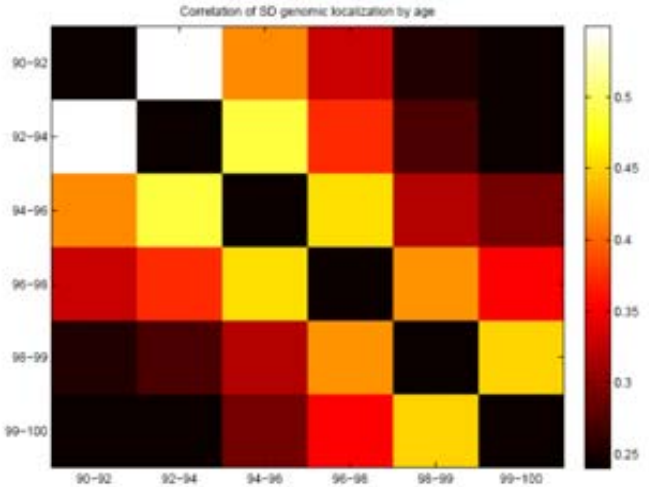
Hypothesis



"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:

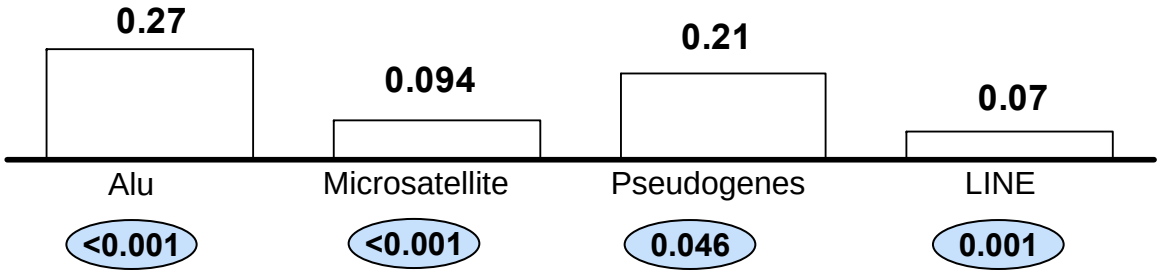


[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

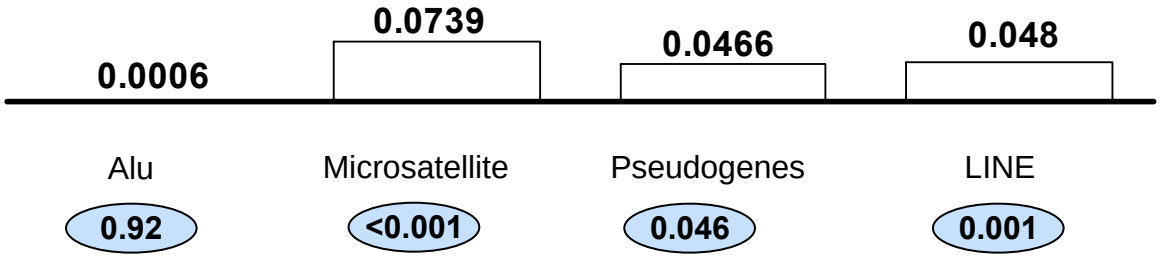
080907_SD_CNV_Slides_MBG_CEGS_PMK

ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs

SD association with repeats

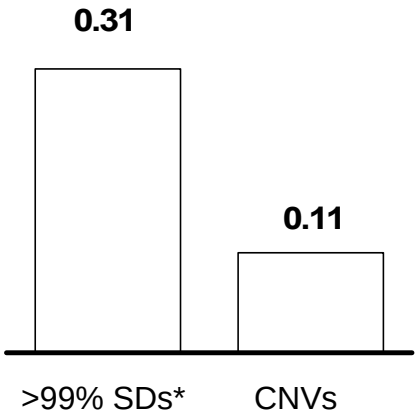


CNV association with repeats



CNVs ARE LESS ASSOCIATED WITH SDs THAN THE GENERAL SD TREND

CNV Association with SDs

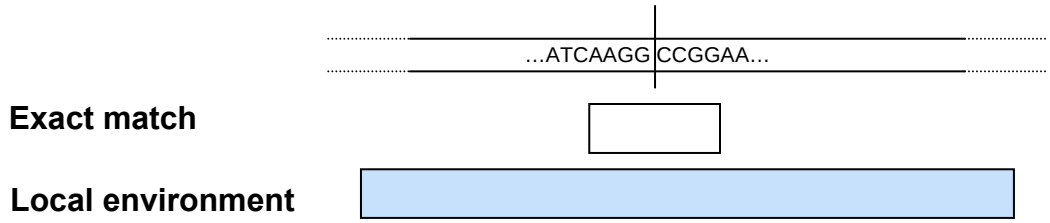


[Kim et al. Gen. Res. (in press, '08), arxiv.org/abs/0709.4200v1]

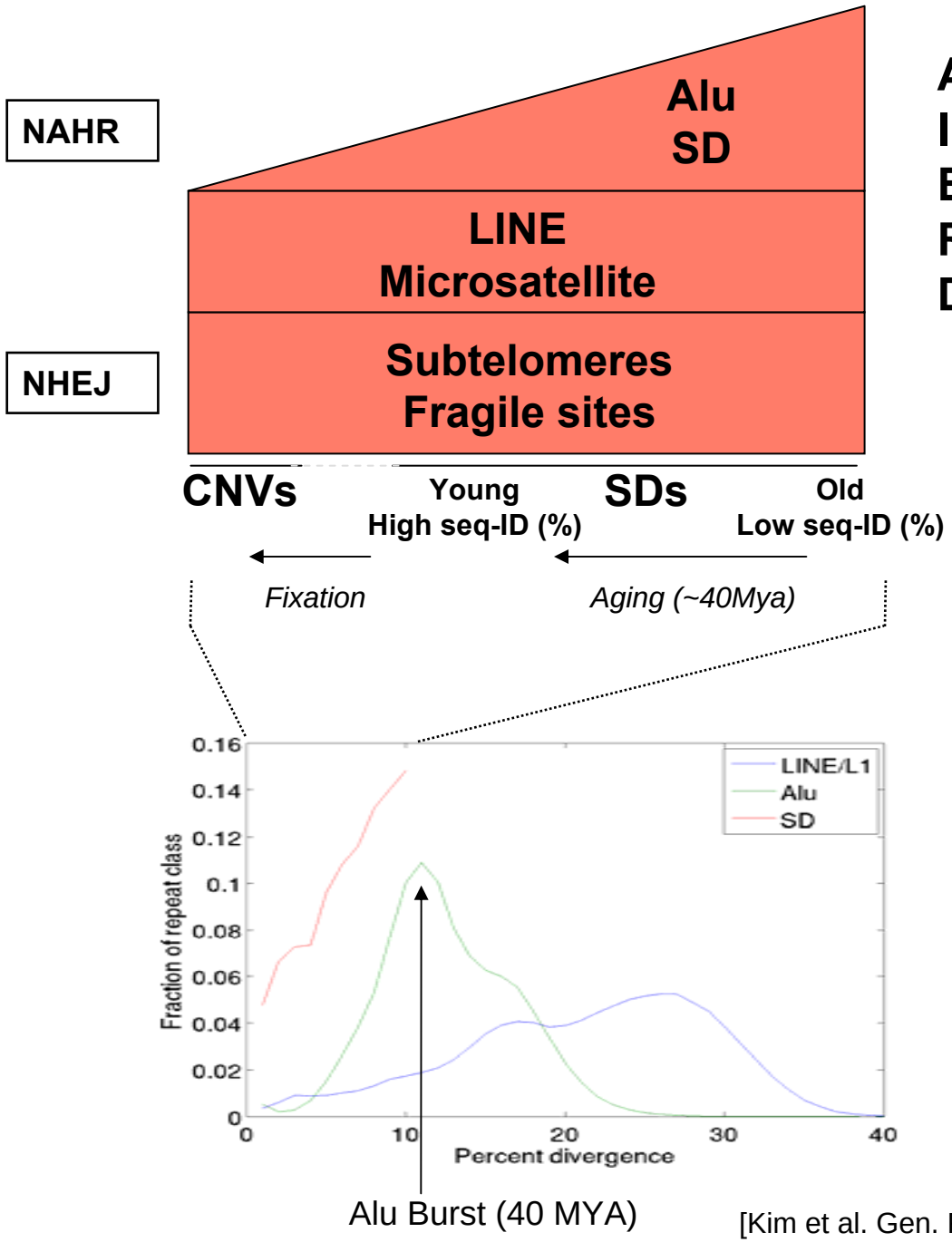
82

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



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

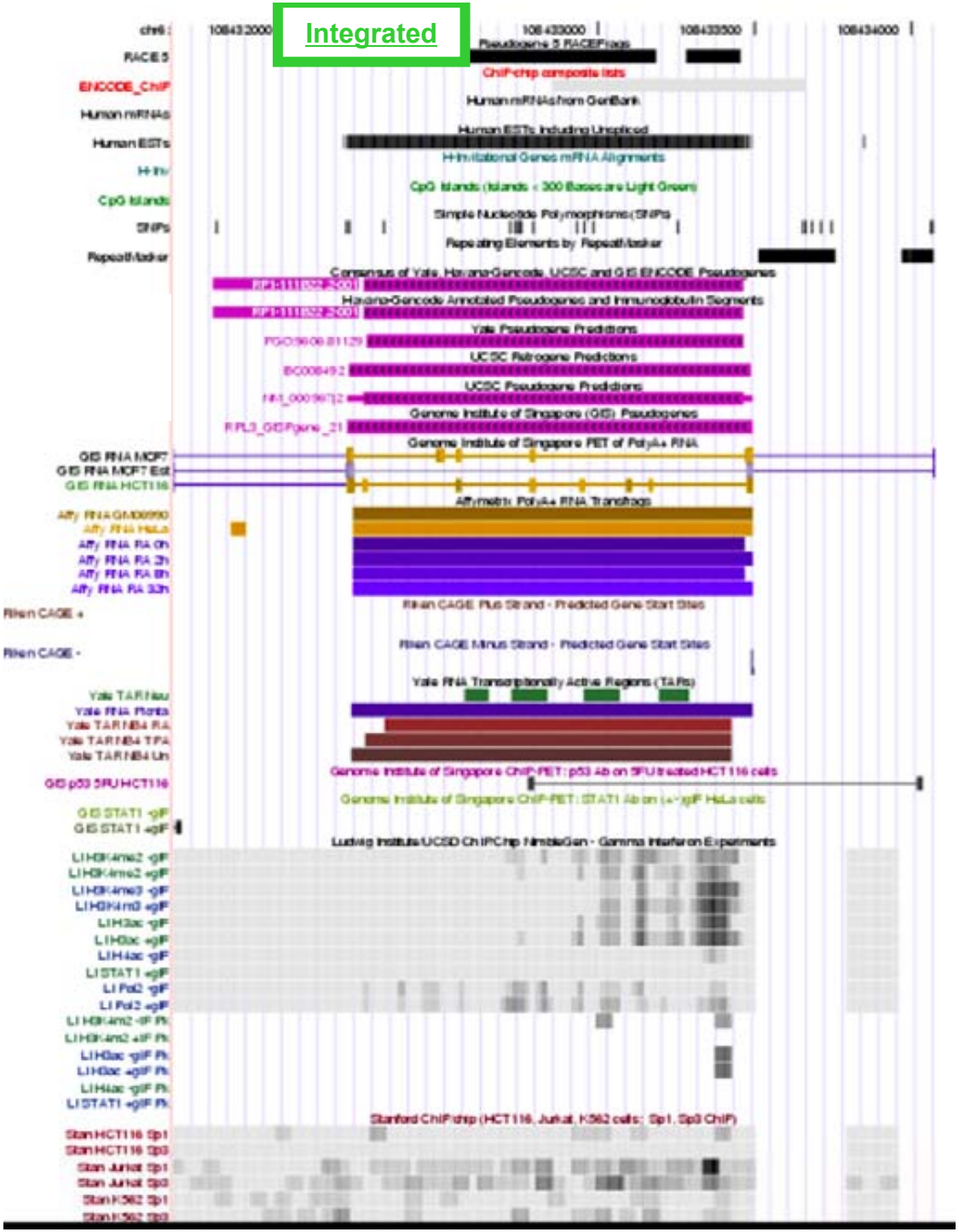


AFTER THE ALU BURST, THE IMPORTANCE OF ALU ELEMENTS FOR GENOME REARRANGEMENT DECLINED RAPIDLY

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

Integration of Different Types of Annotation: Pseudogenes with Intergenic Biochemical Activity



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)

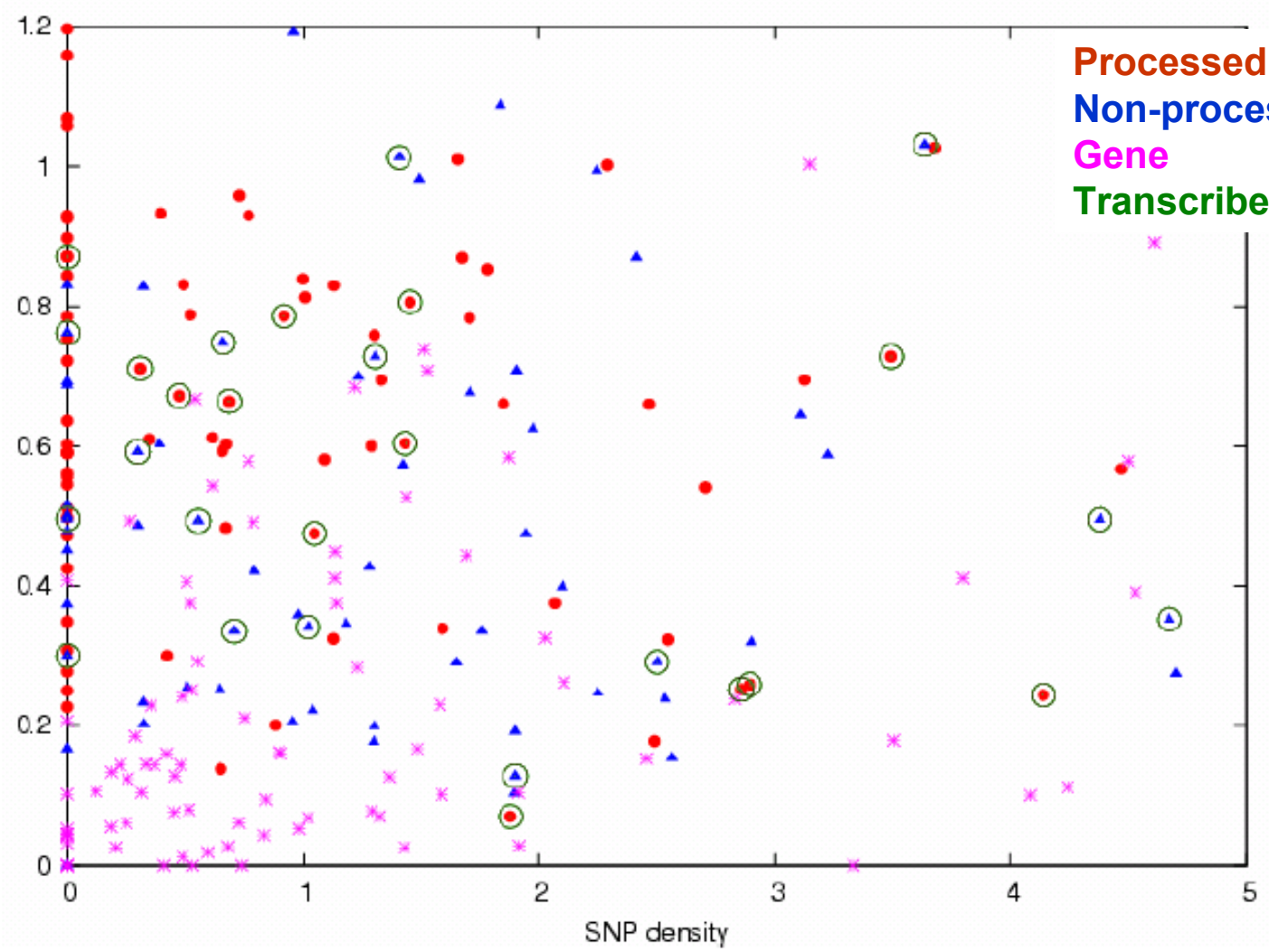
Targeted RACE expts to 160 pseudogenes, gives 14

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

Zheng et al. (2007) Gen. Res.

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)

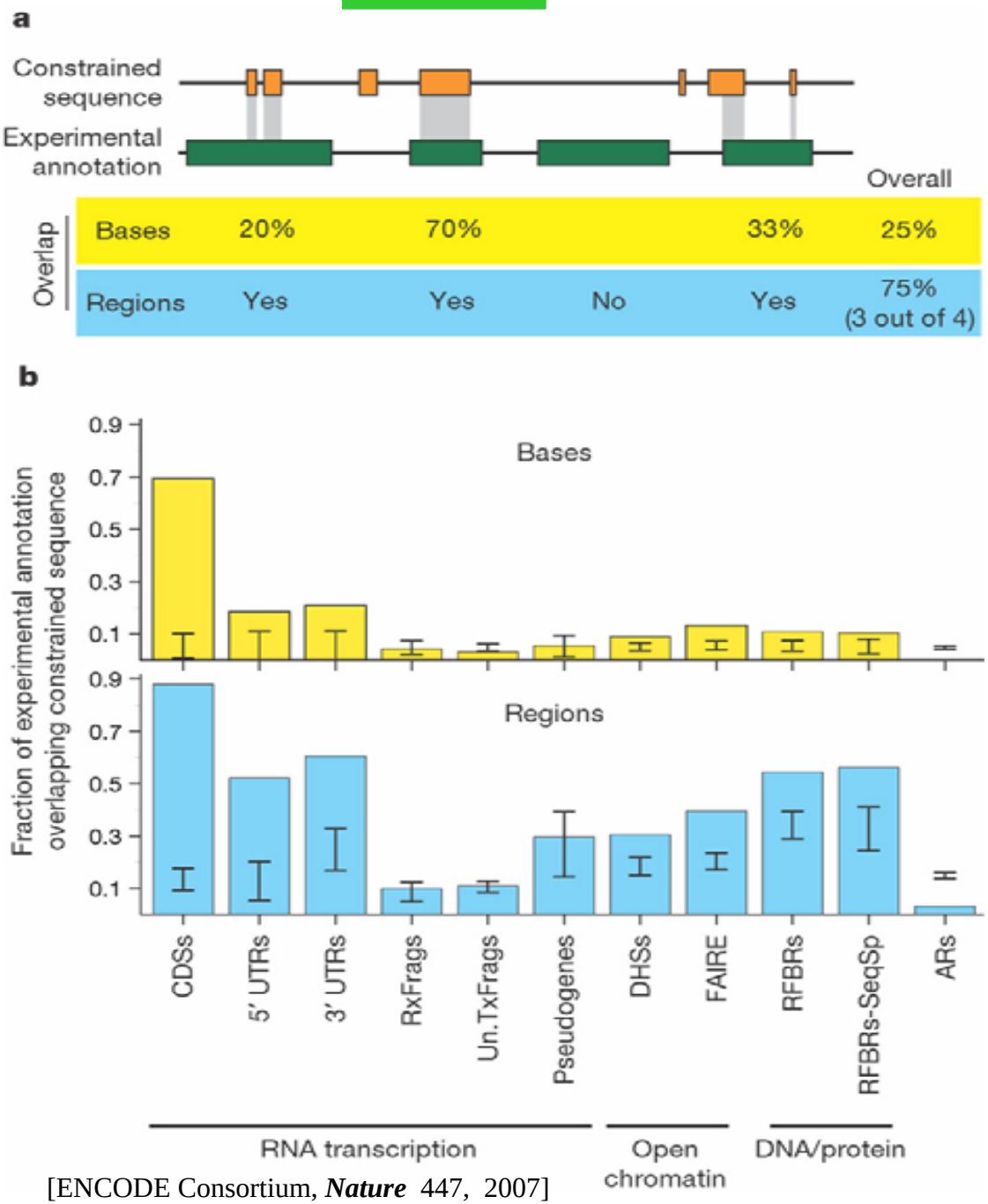
Zheng et al. (2007) Gen. Res.

Extension to Whole Genome

- **233** Transcribed from ~8000 Processed Pseudogenes
- Evidence for Transcription
  - ◊ 8% Refseq mRNAs
  - ◊ 32% Unigene consensus sequences
  - ◊ 72% dbEST expressed sequence tags
  - ◊ 32% Oligonucleotide microarray data (extra support)
- Highly decayed
  - ◊ Fraction with Ka/Ks ≥ 0.5 is 54%

Harrison et al. (2005) NAR

Integrated



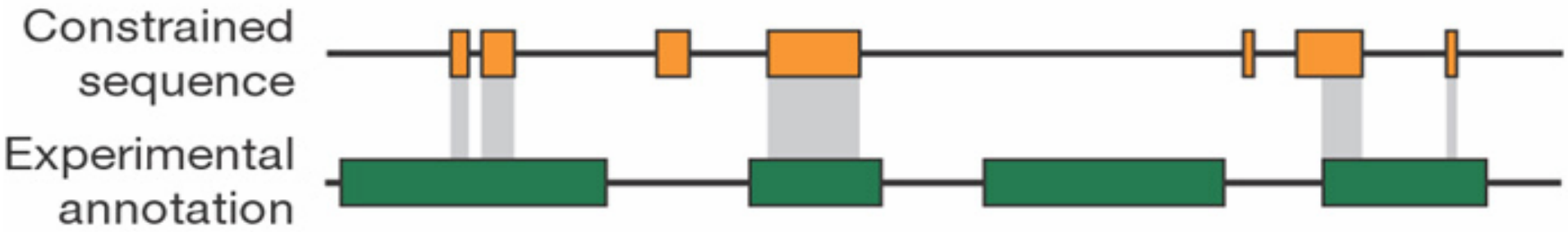
[ENCODE Consortium, ***Nature*** 447, 2007]

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

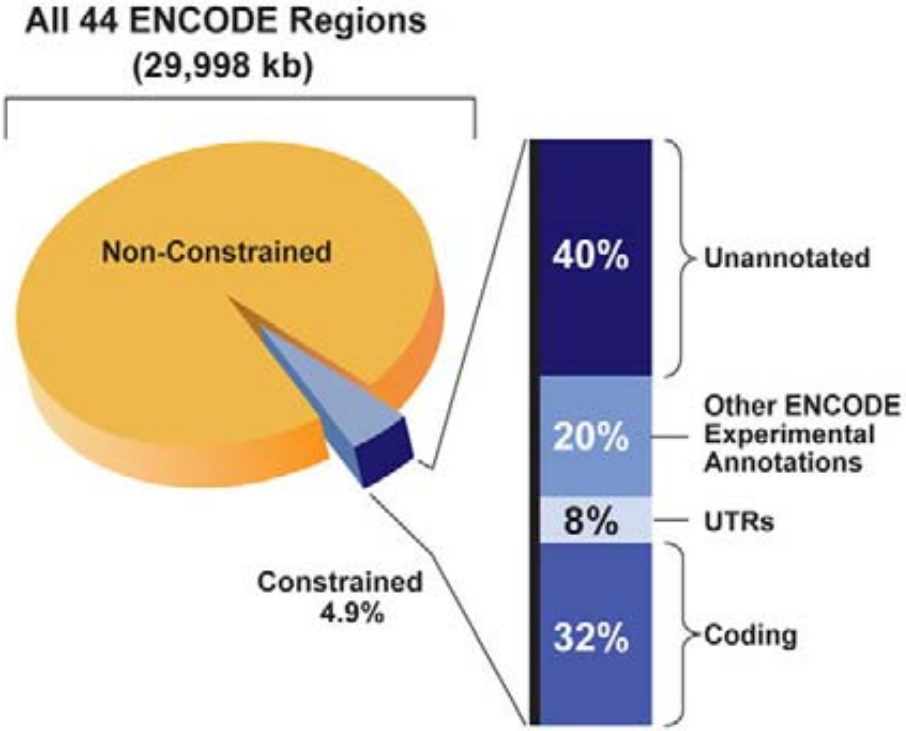
- Integrating & averaging results over larger and larger sets
- Comparison of integrated quantities

Integrated

Grand Summary: Biochemical Activity vs. Sequence Constraints

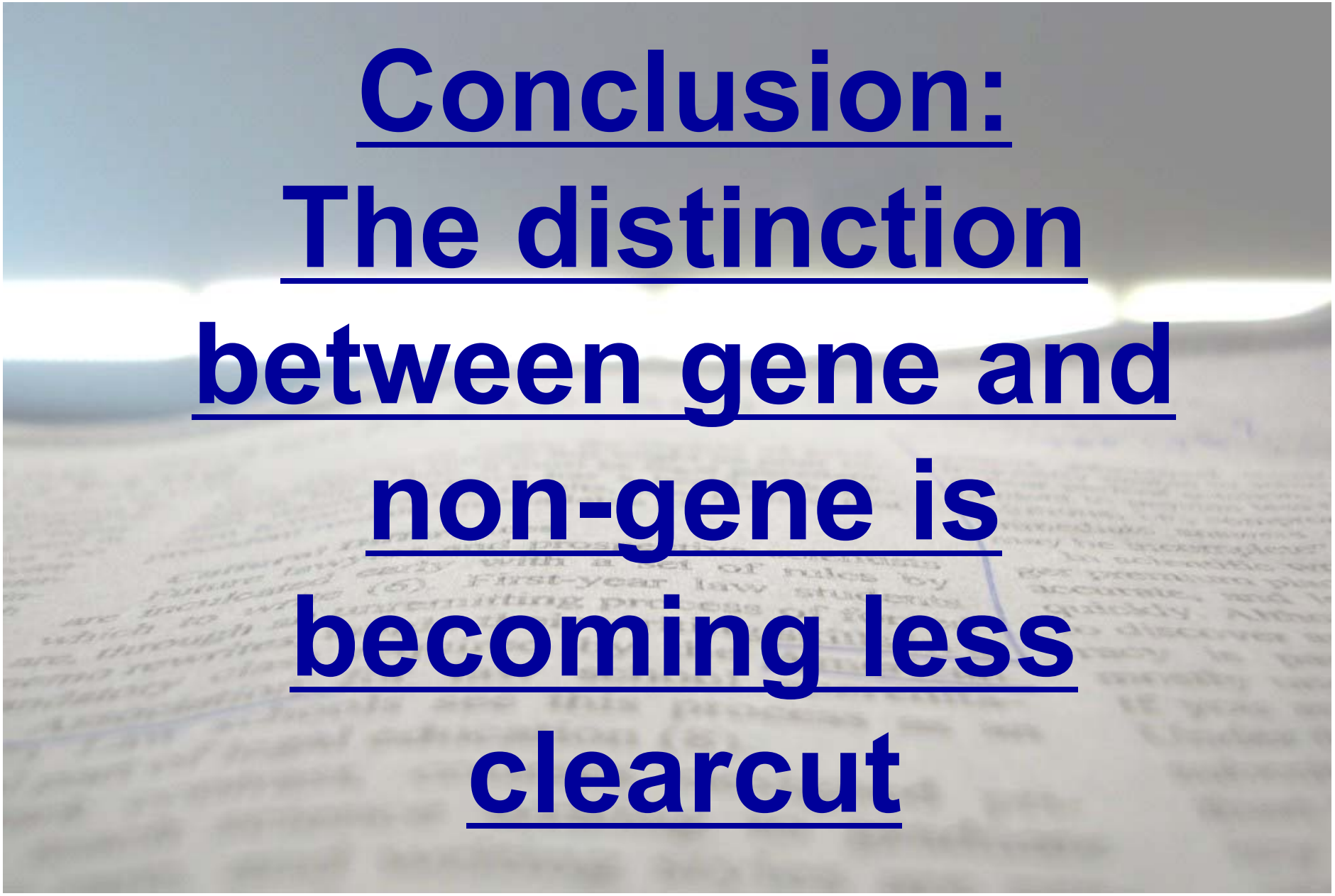


- Not all constrained sequence annotated in some fashion
- Exactly how things are defined in terms of overlap?

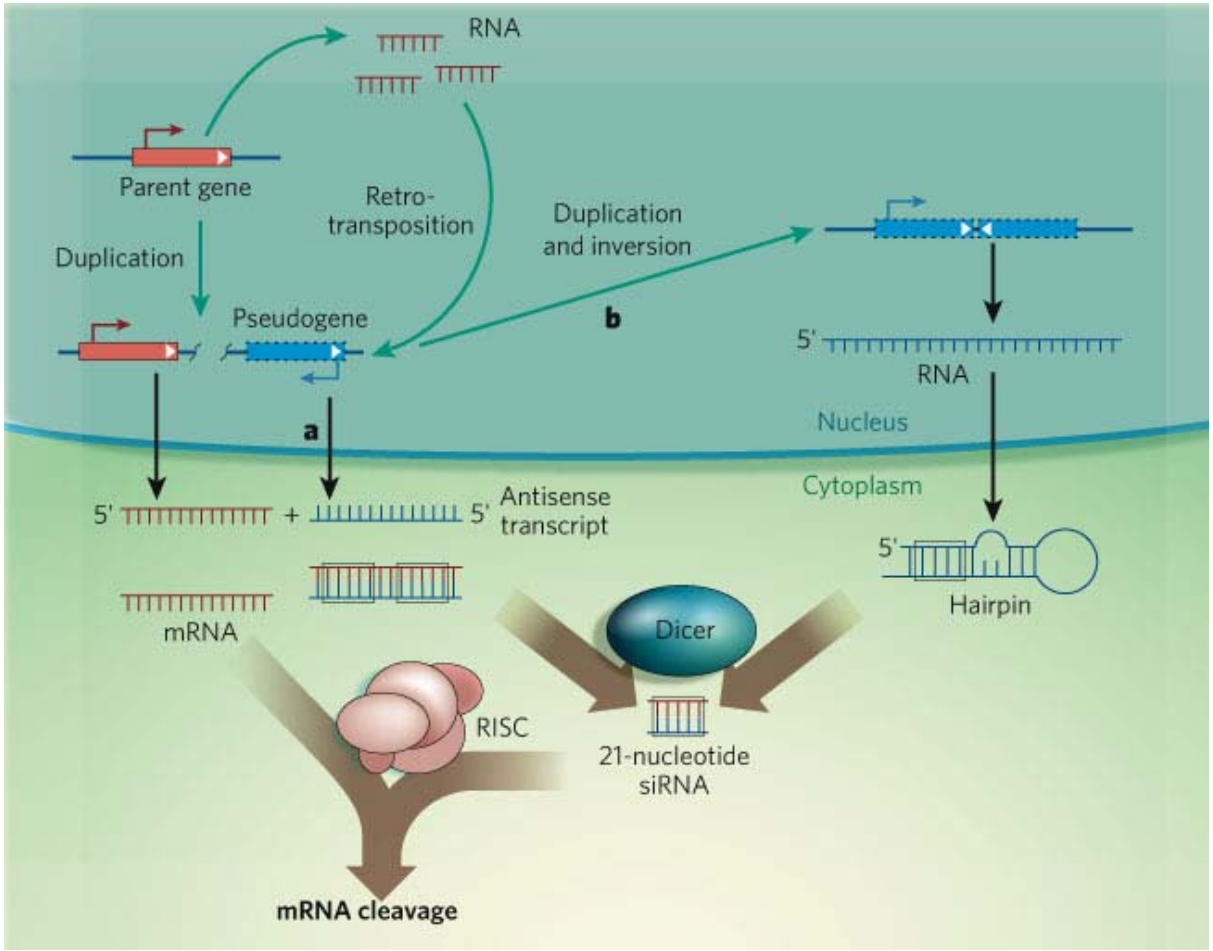


- "At the outset of the ENCODE Project, many believed that the broad collection of experimental data would nicely dovetail with the detailed evolutionary information derived from comparing multiple mammalian sequences to provide a neat 'dictionary' of conserved genomic elements, each with a growing annotation 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



What are Active Pseudogenes Doing?

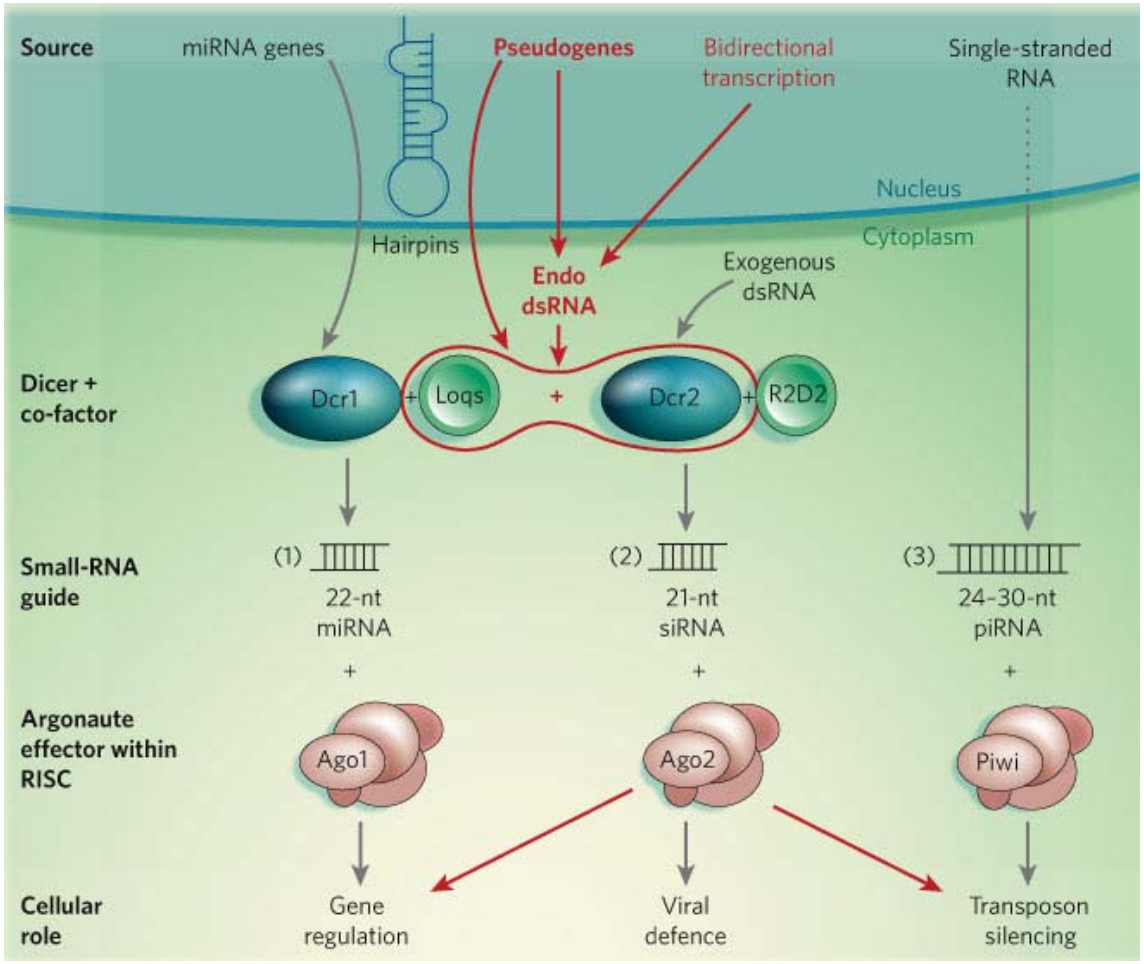
Potential for Gene Regulation via endo-siRNA

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

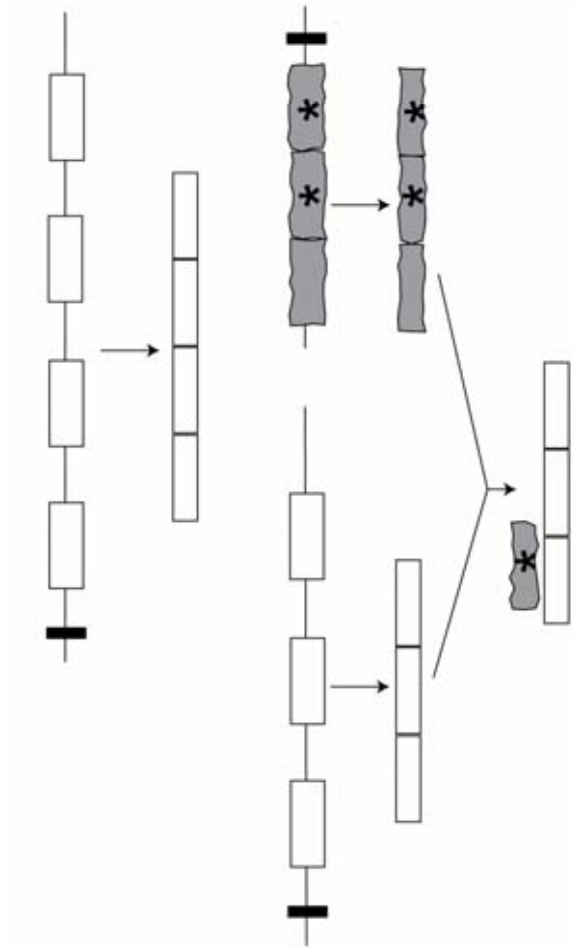


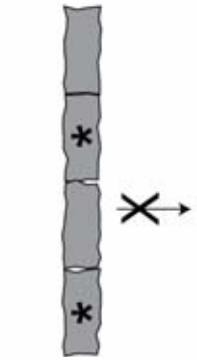
[Sasidharan & Gerstein, Nature ('08)]

Genes & Pseudogenes

Zheng & Gerstein, TIG (2007)

Genes or Pseudogenes?

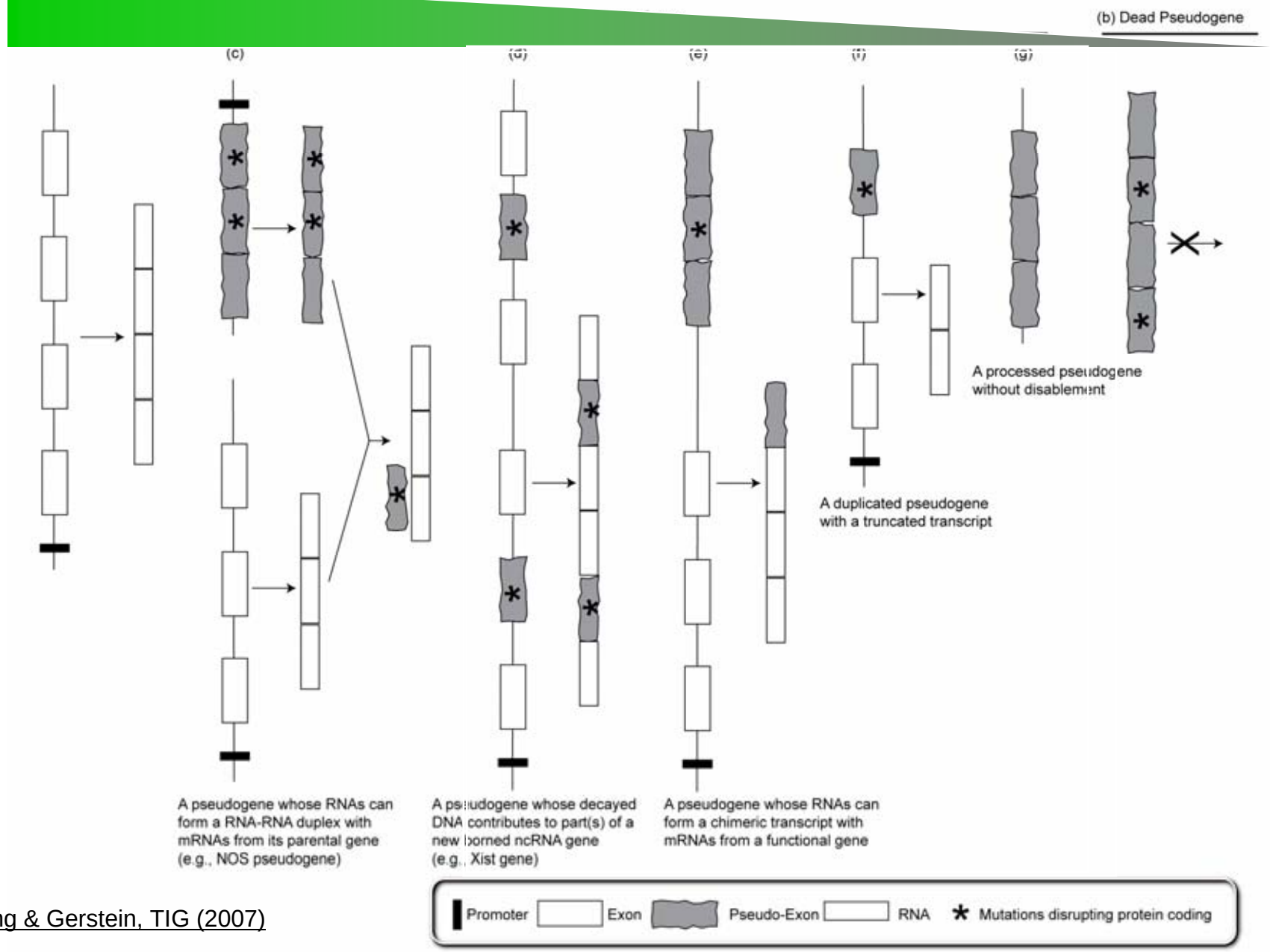




Transcribed pseudogene regulates parent (NOS)

Zheng & Gerstein, TIG (2007)

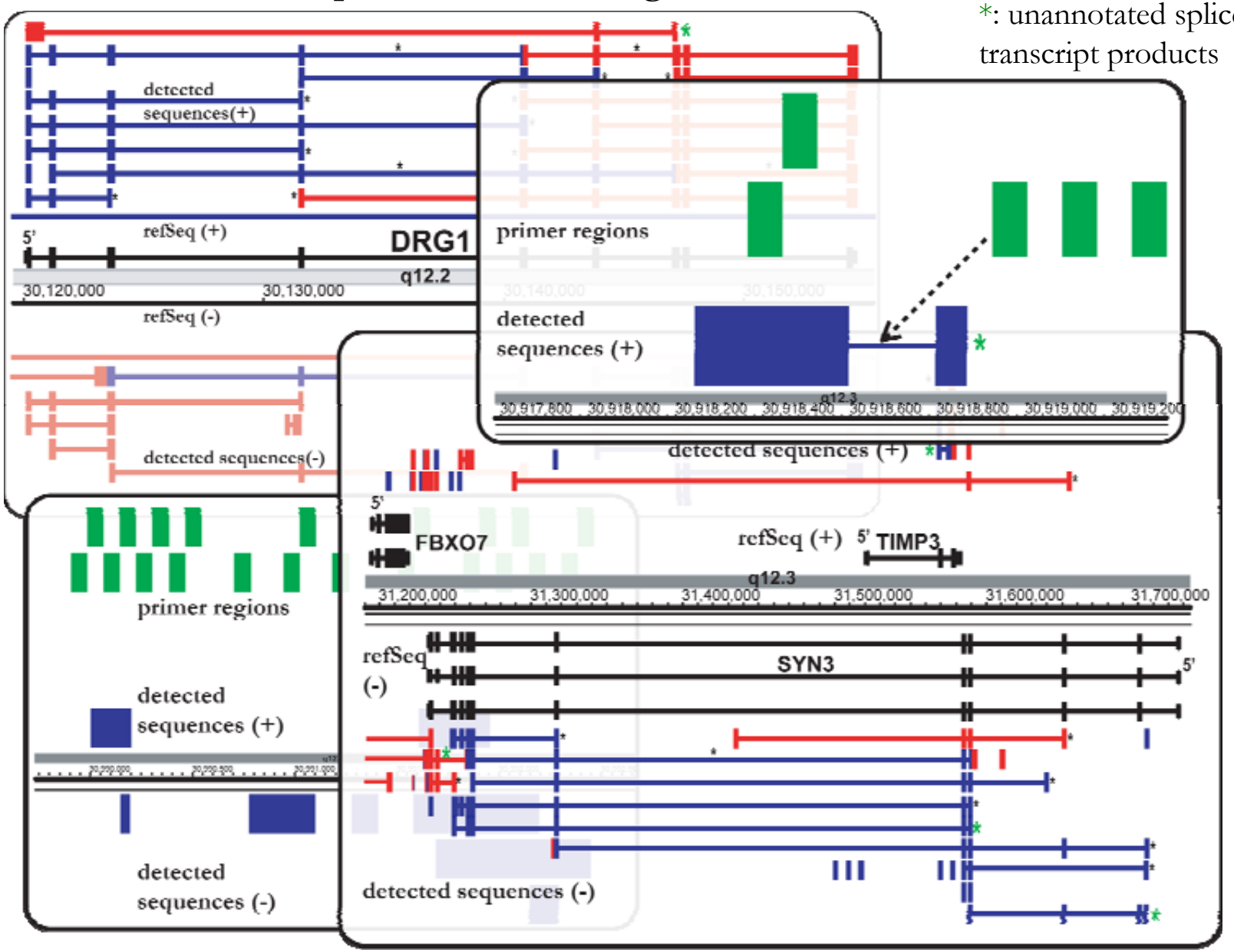
Genes or Pseudogenes?



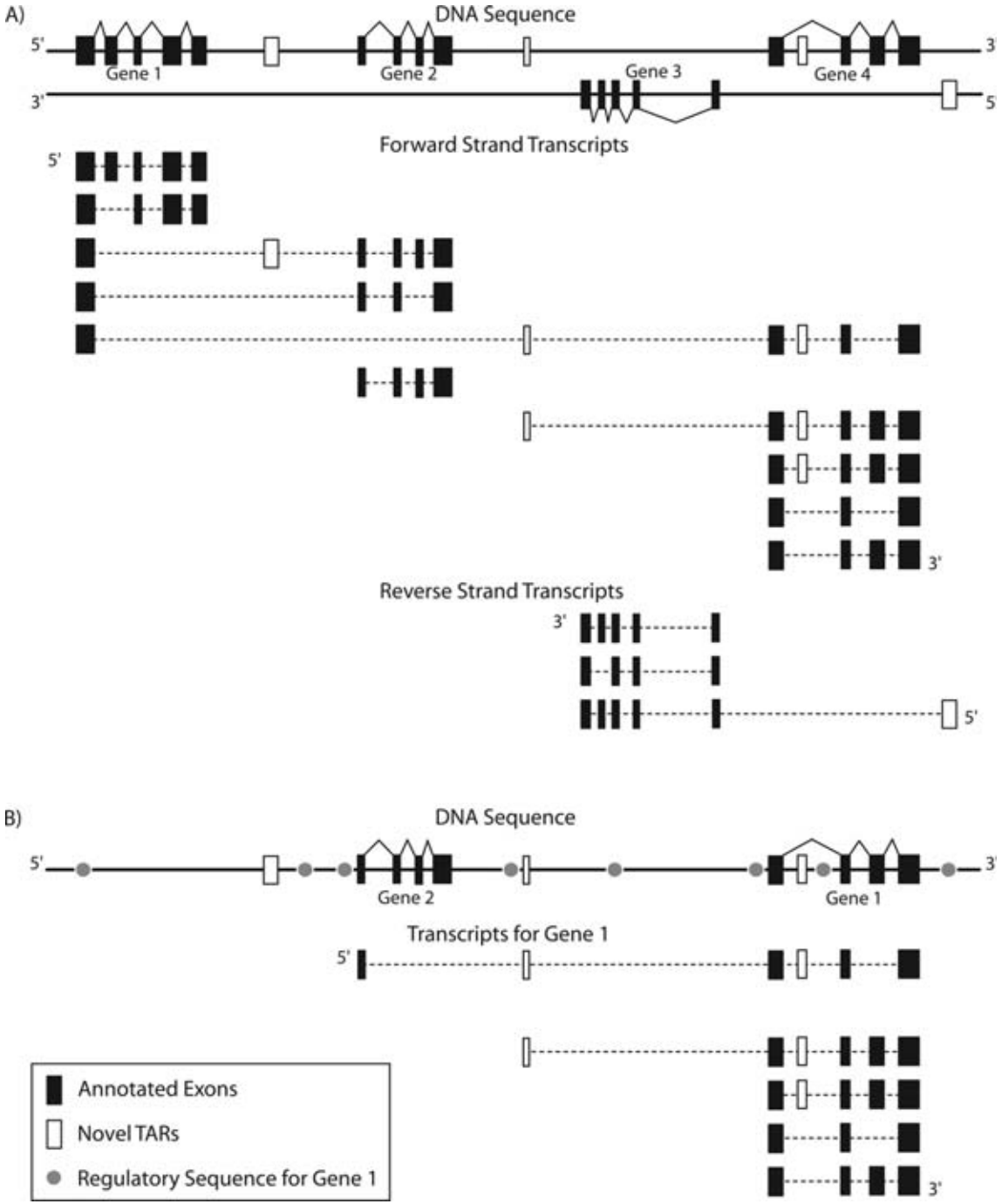
Zheng & Gerstein, TIG (2007)

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

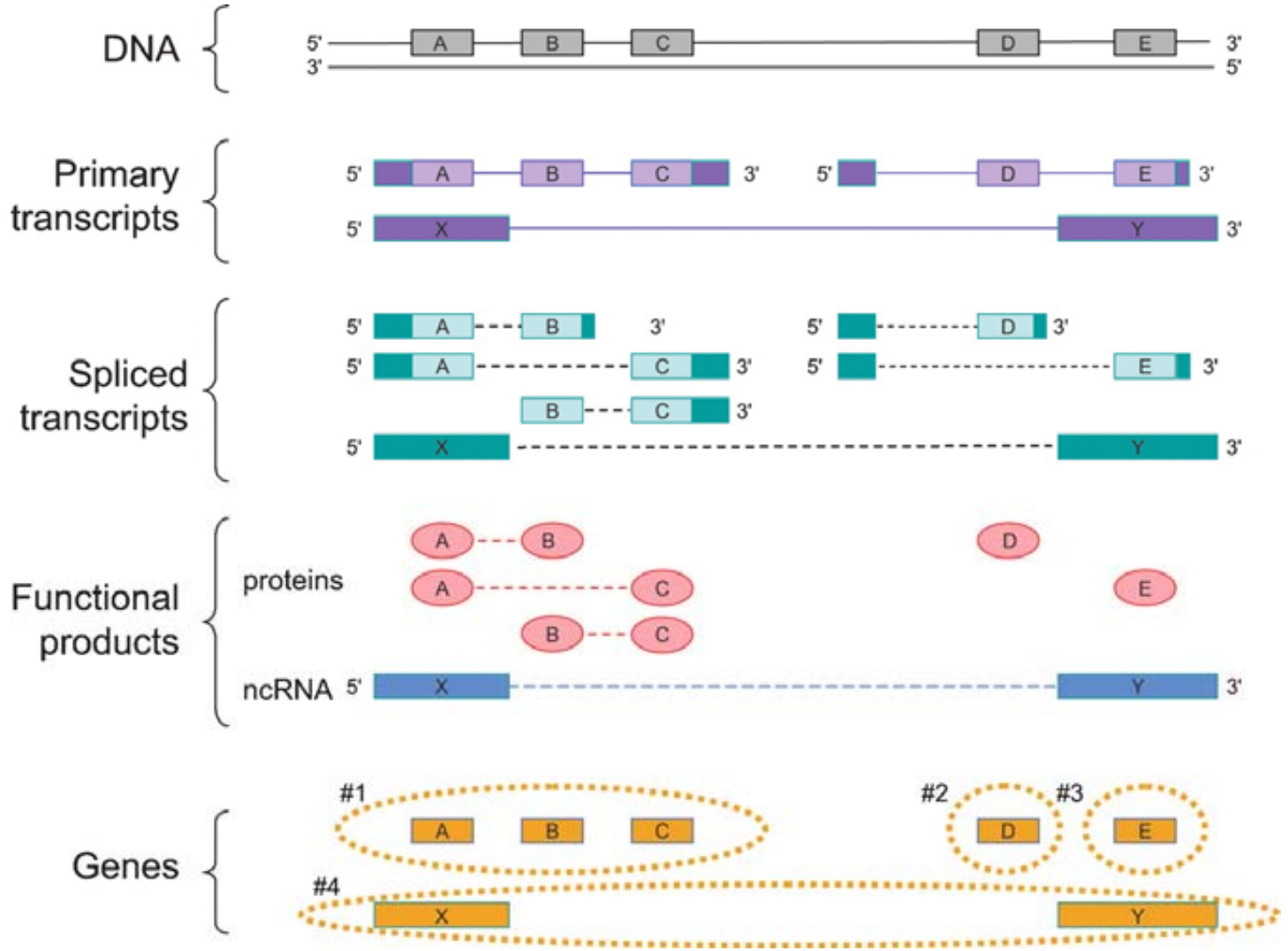


Biological complexity revealed by ENCODE: Long Interleaved Transcripts and Distributed Regulation

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

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

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

Summary

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

Overview of the Process of Intergenic Annotation

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

Outline

- Calling Blocks
  - ◊ Issues in Signal Processing Transcriptional Data from Tiling Arrays
- Clustering Blocks
  - ◊ Scoring transcribed regions and clustering these together into larger blocks
- Formal Annotation
  - ◊ Pseudogenes
- Integration with conservation
  - ◊ SDs and CNVs
  - ◊ pseudogenes and seq. constraint
- Integration with activity
  - ◊ pseudogenes and transcriptions
- Future of Gene Annotation
  - ◊ What is a gene post encode?

Processing the Raw Experimental Signal

- Normalizing and Correcting Artifacts
  - ◊ Characterizing, correcting and exploiting specific and non-specific cross-hybridization
    - Measuring the effect of types of mismatches
    - Iterated quantile norm. to correct non-specific cross
    - Towards a universal array based on specific cross-hyb.

Annotating the Human Genome: 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

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

Analysis of Duplication in the Genome: SVs and SDs

- Large-scale analysis of existing CNVs & SDs in human genome
  - ◊ Process giving rise to pseudogenes
- 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
  - ◊ Role for proc. pseudogenes in giving rise to duplicated 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

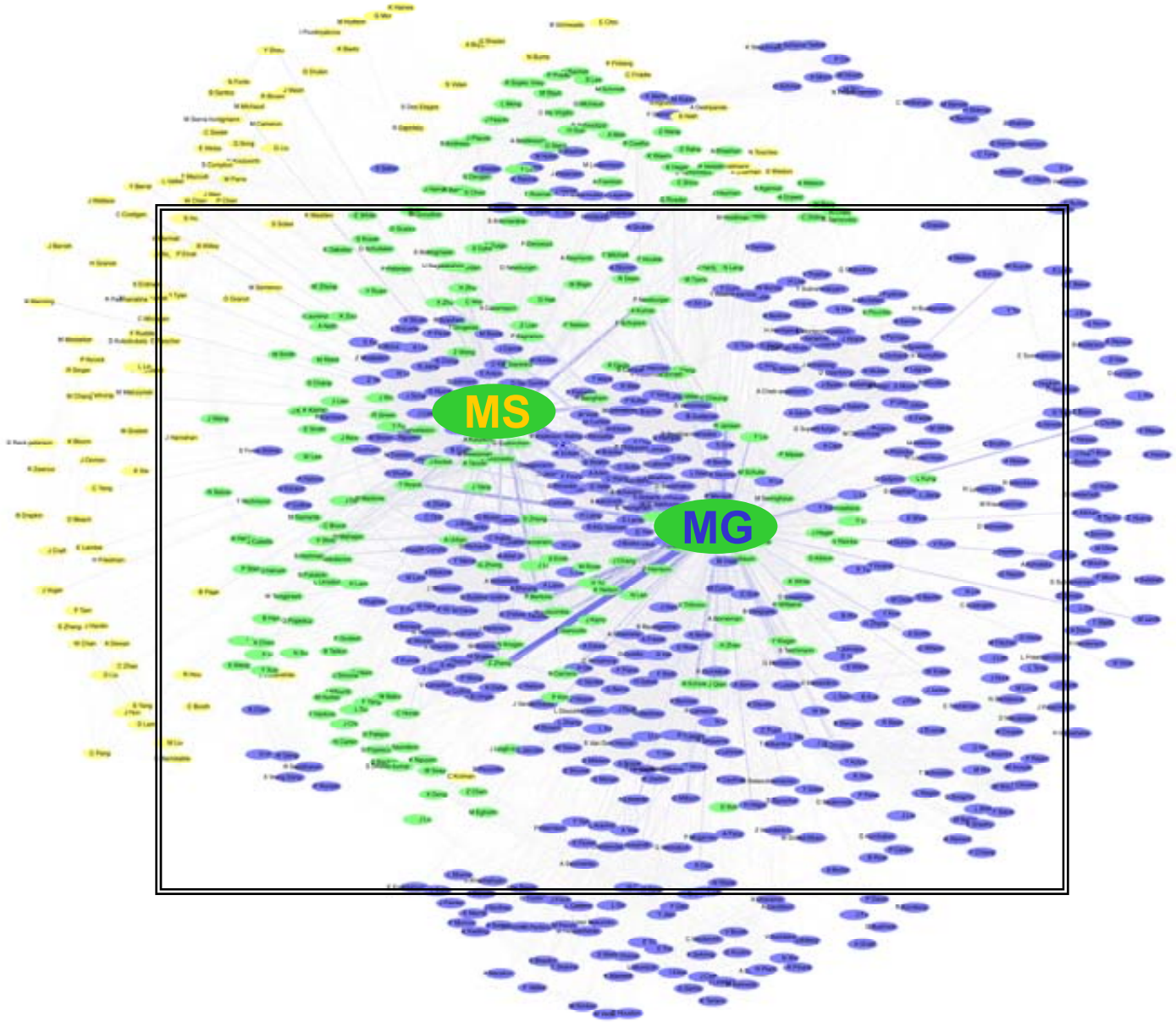
+

various consortia

(ENCODE, modENCODE, 1000 Genomes)

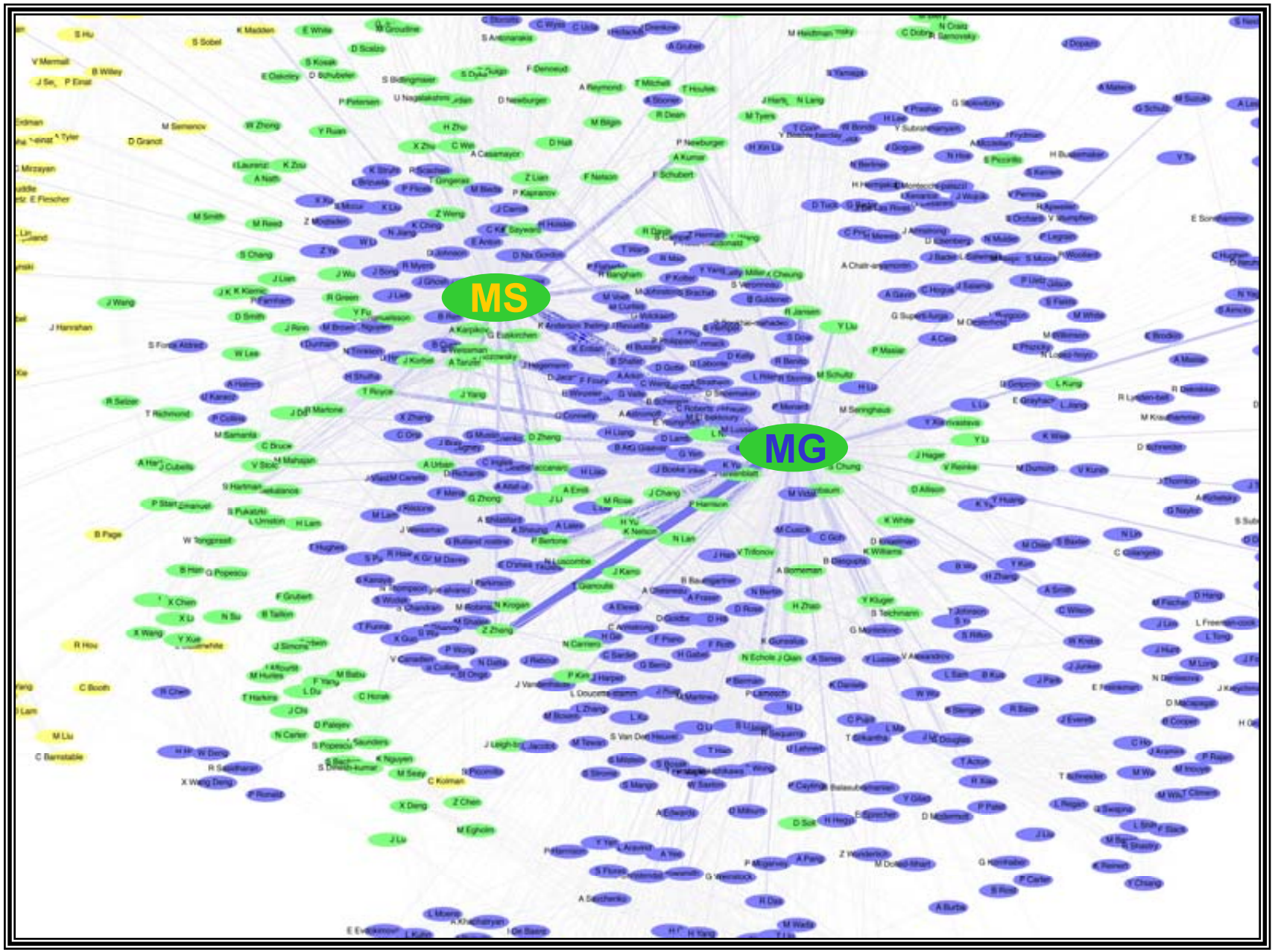
Acknowledgements

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



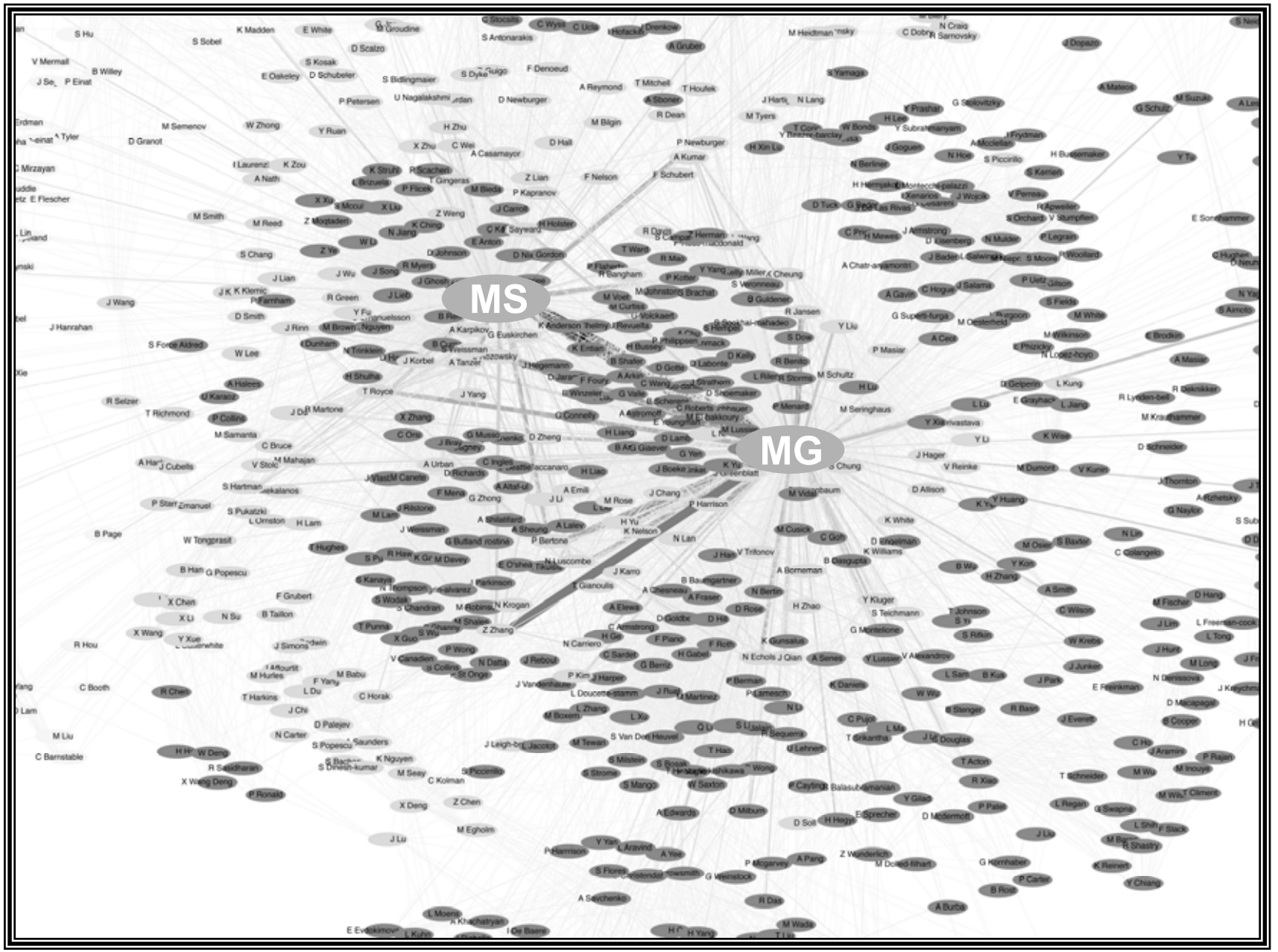
Acknowledgements

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



Acknowledgements

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

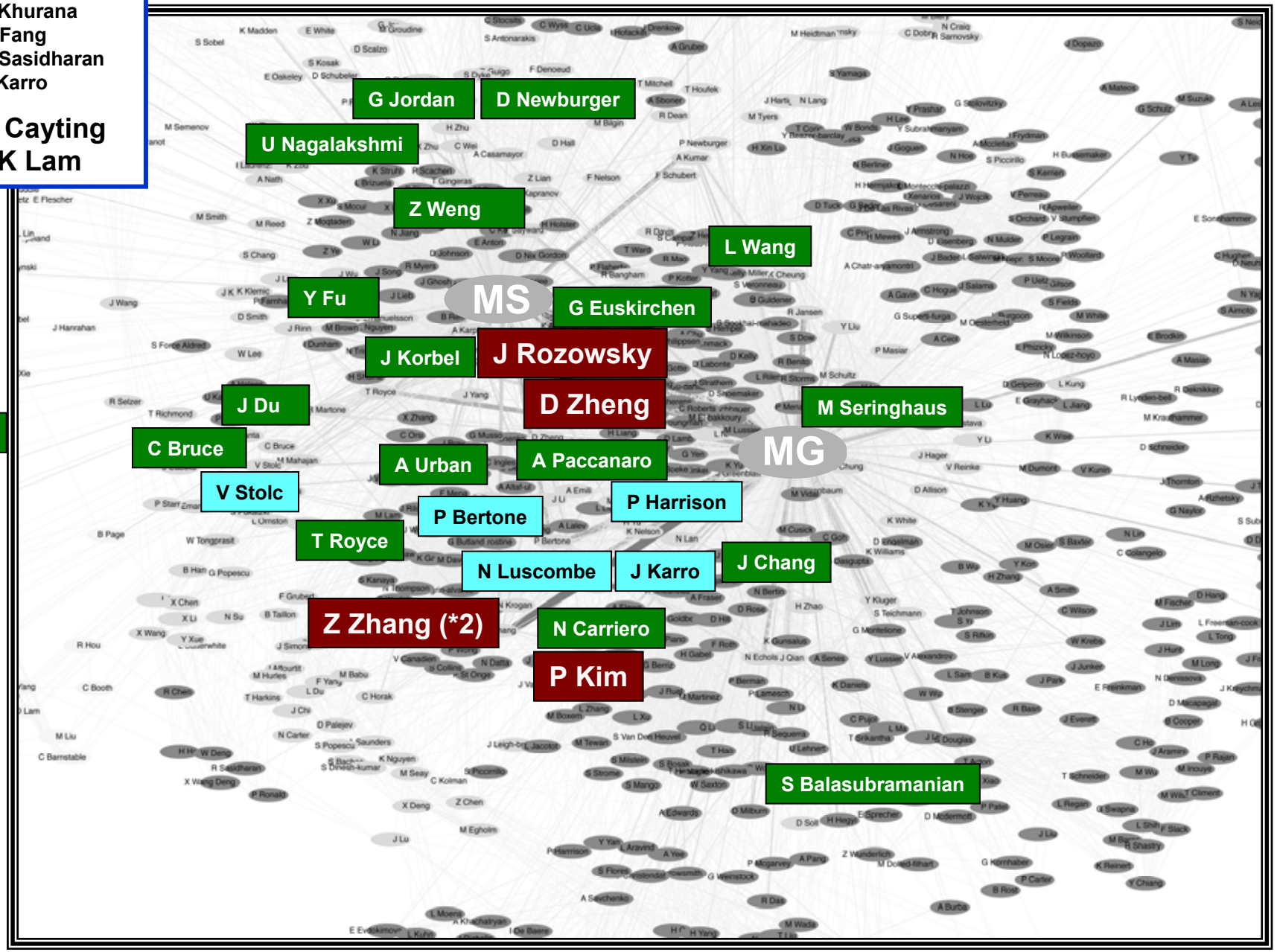


Acknowledgements

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

P Cayting
YK Lam

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



Permissions Statement