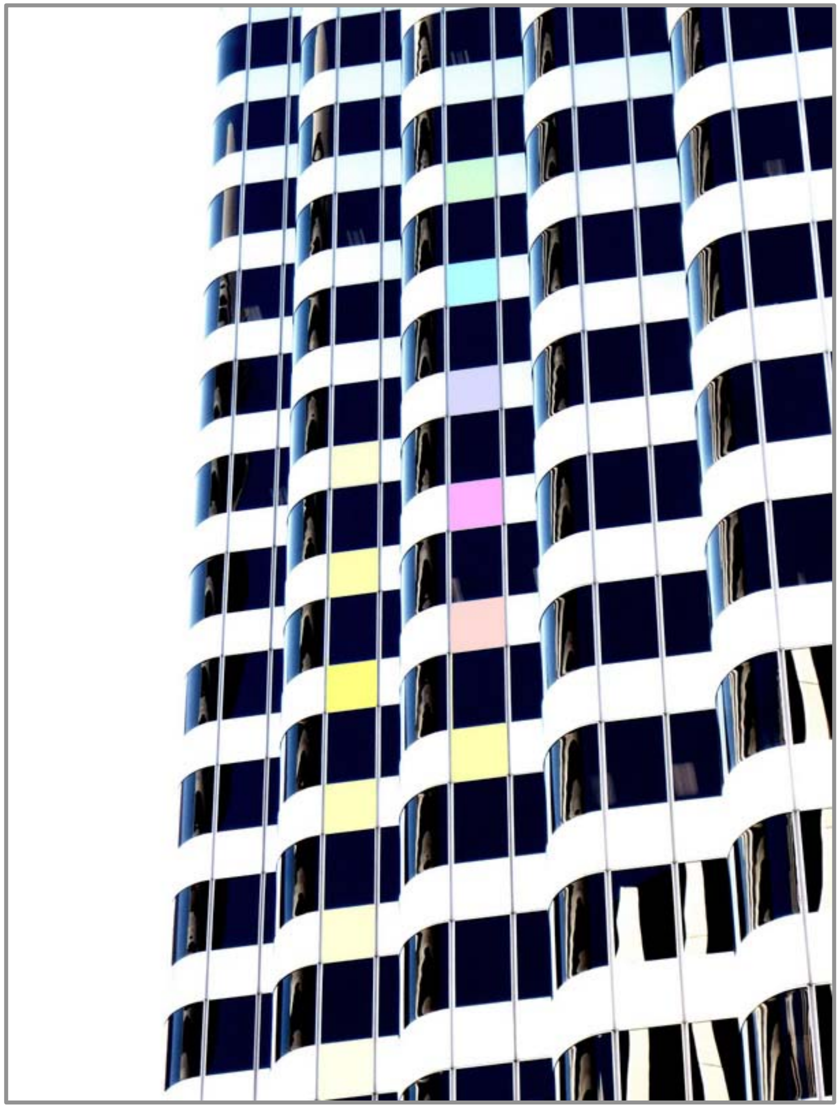
Human Genome Annotation

Mark B Gerstein
Yale

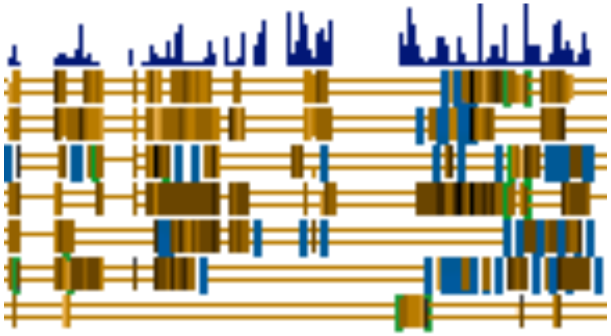
Slides at
Lectures.GersteinLab.org

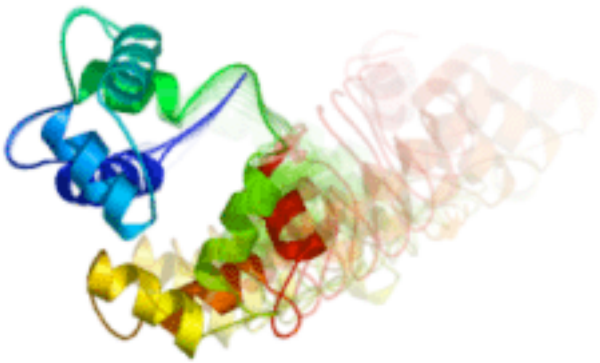
(See Last Slide for References & More Info.)



GersteinLab.org Research Overview: Bioinformatics

- **Genome Annotation**
  - ◊ Characterizing the function of non-coding regions of the genome, focusing on protein fossils and novel RNAs (Pseudogene.org + GenomeTech.GersteinLab.org)
- **Molecular Networks**
  - ◊ Using molecular networks to integrate & mine functional genomics information and describe genefunction on a large-scale (Networks.GersteinLab.org)
- **Macromolecular Motions**
  - ◊ Analyzing select populations of 3D-structures in detail, trying to understand their flexibility in terms of packing (MolMovDB.org)





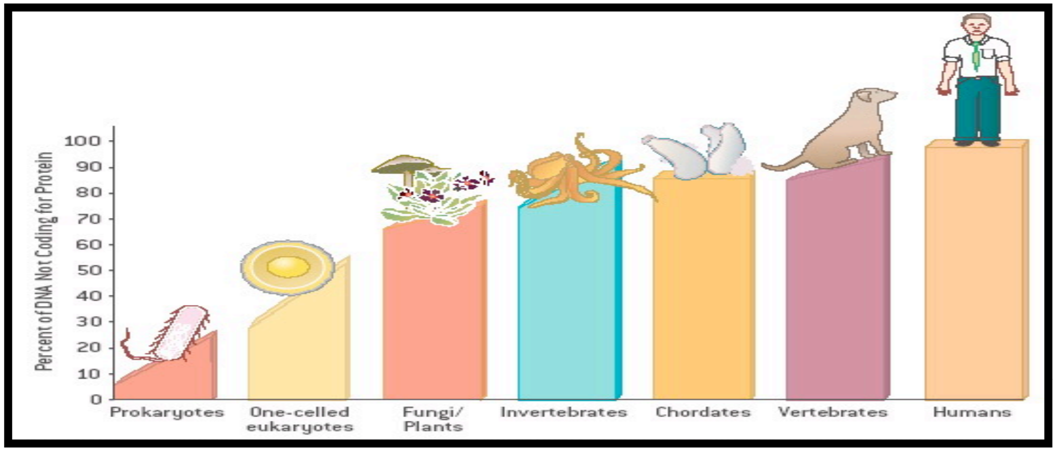
2001: Most of the genome is not coding (only ~1.2% exon).

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

iami.edu/~cmallery/150/gene/noncoding.genes.jpg

Humans have a comparatively large non-coding fraction of their genome

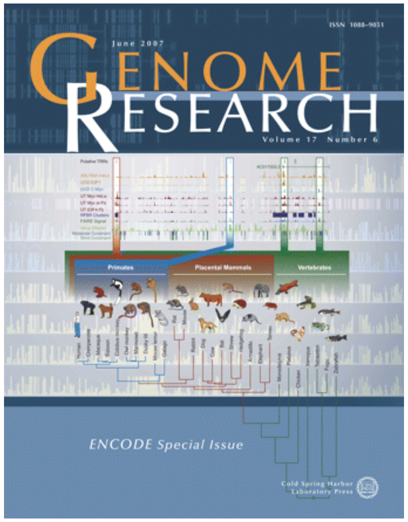
2001



[IHGSC, *Nature* 409, 2001]

nter et al. *Science* 29, 2001]

14 June 2007 | www.nature.com/nature | $10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
Histone-modification chromatin IP
Tiling array
MARS'S ANCIENT OCEAN
Polar wander solves an enigma
THE DEPTHS OF DISGUST
Understanding the ugliest emotion
MENTORING
How to be top
NATUREJOBS
Contract research
DECODING THE BLUEPRINT
The ENCODE pilot maps human genome function

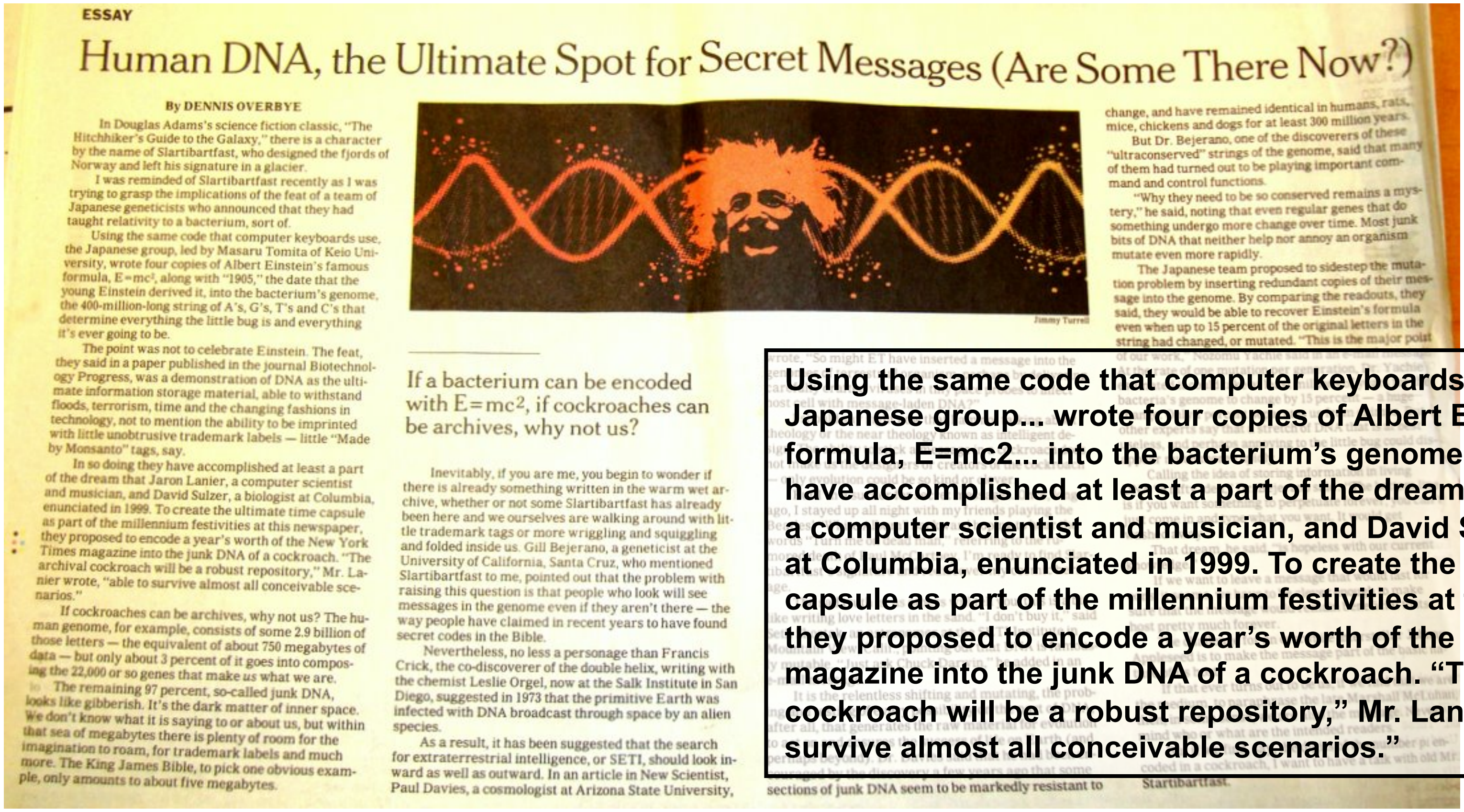


2007 : Pilot results from ENCODE Consortium on decoding what the bases do

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

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. 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,

sections of junk DNA seem to be markedly resistant to

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

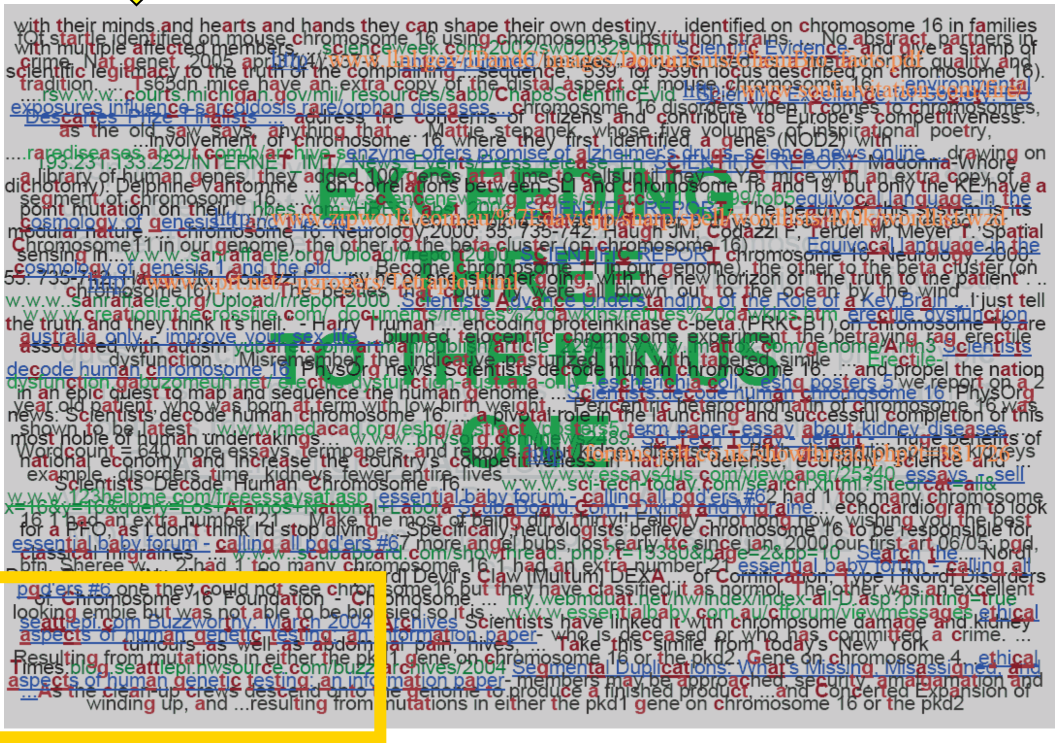
But Dr. Bejerano, one of the discoverers of these "ultraconserved" strings of the genome, said that many of them had turned out to be playing important command and control functions.

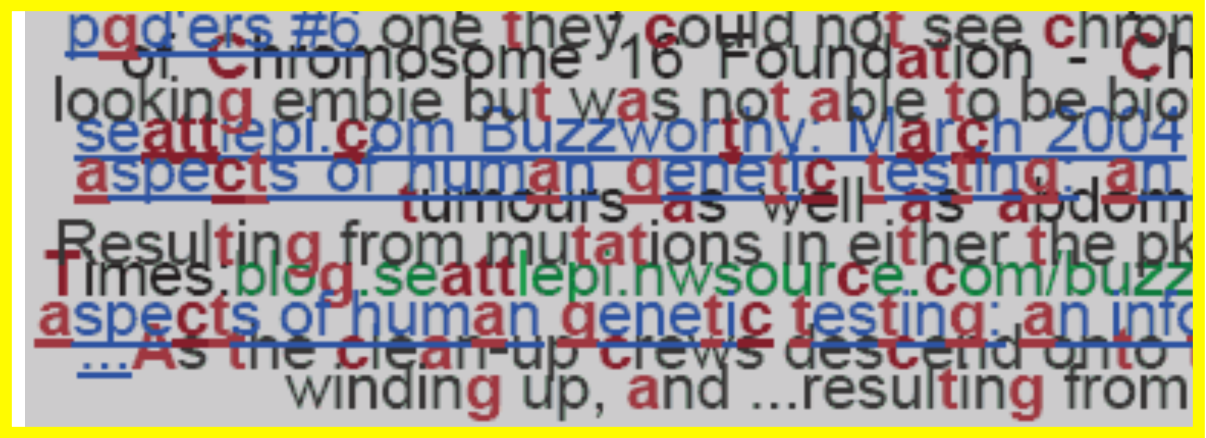
"Why they need to be so conserved remains a mystery," he said, noting that even regular genes that do something undergo more change over time. Most junk bits of DNA that neither help nor annoy an organism mutate even more rapidly.

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

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."




pgd'ers #6 one they could not see chro
of Chromosome 16 Foundation - Ch
looking embie but was not able to be bio
seattlepi.com Buzzworthy: March 2004
aspects of human genetic testing: an
tumours as well as abdom
Resulting from mutations in either the pk
Times.blog.seattlepi.nwsource.com/buzz
aspects of human genetic testing: an info
...As the clean-up crews descend onto
winding up, and ...resulting from

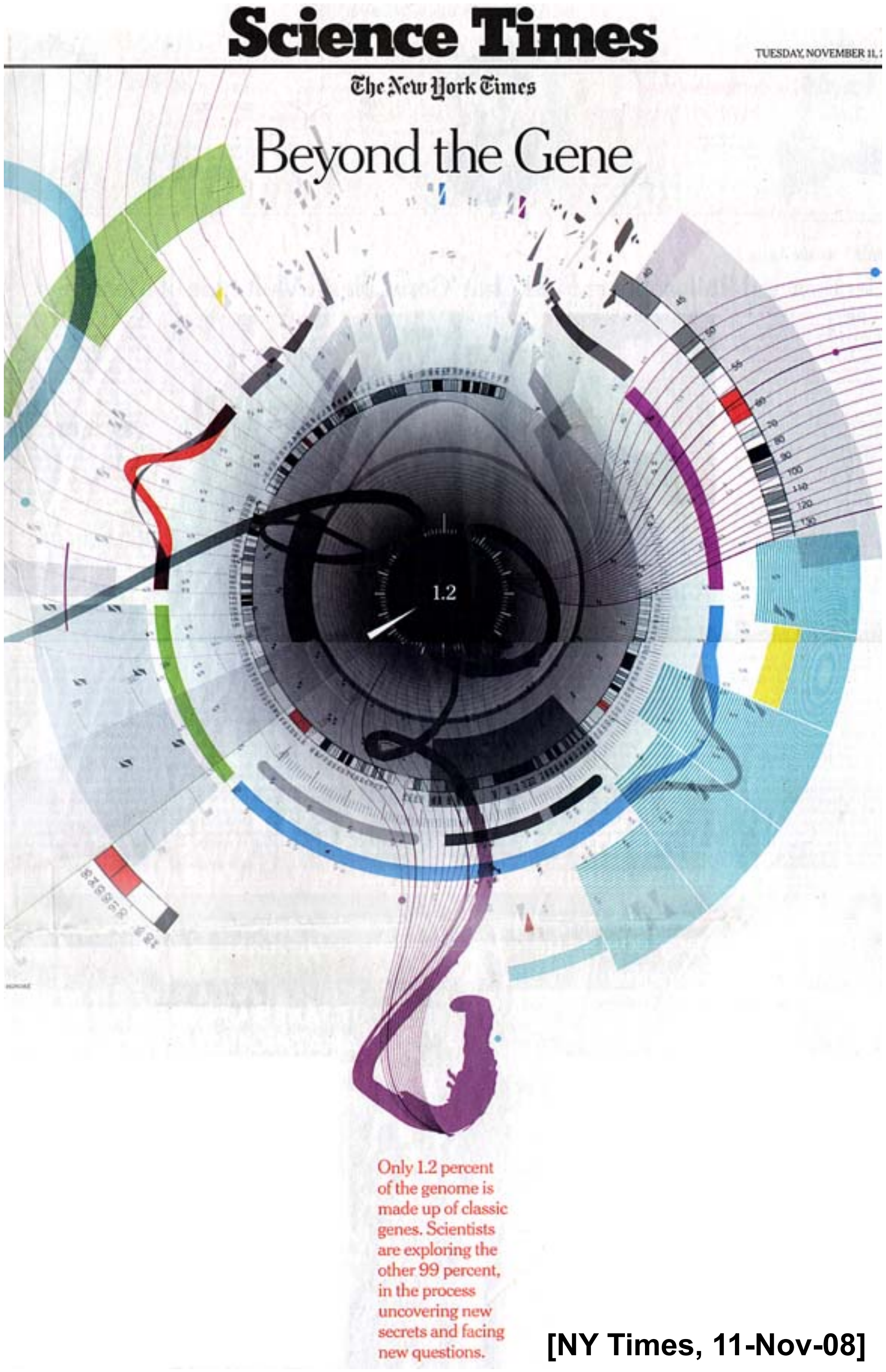
Junk DNA as Art

Significance of the “dark matter of the genome”

- Pervasive Activity
  - Encode pilot
- Association with Disease
  - Noncoding regions identified correlations with human diseases (GWAS)
- History
  - Historical record of genome, molecular clock

- **Personal Genomics**
  - Importance multipled by future need to interpret millions of personal genomes

References
http://www.nature.com/nature/journal/v461/n7261/full/nature08451.html
http://linkinghub.elsevier.com/retrieve/pii/S0002929707625403
http://www.springerlink.com/content/c3816334655h7844/
http://www.sciencemag.org/cgi/content/abstract/1138341v1
http://www.nature.com/nature/journal/v430/n7000/full/nature02697.html
http://www.ncbi.nlm.nih.gov/pubmed/7769622?dopt=Citation
http://www.springerlink.com/content/c8ptualwqby9pxr2/



Science Times

TUESDAY, NOVEMBER 11,

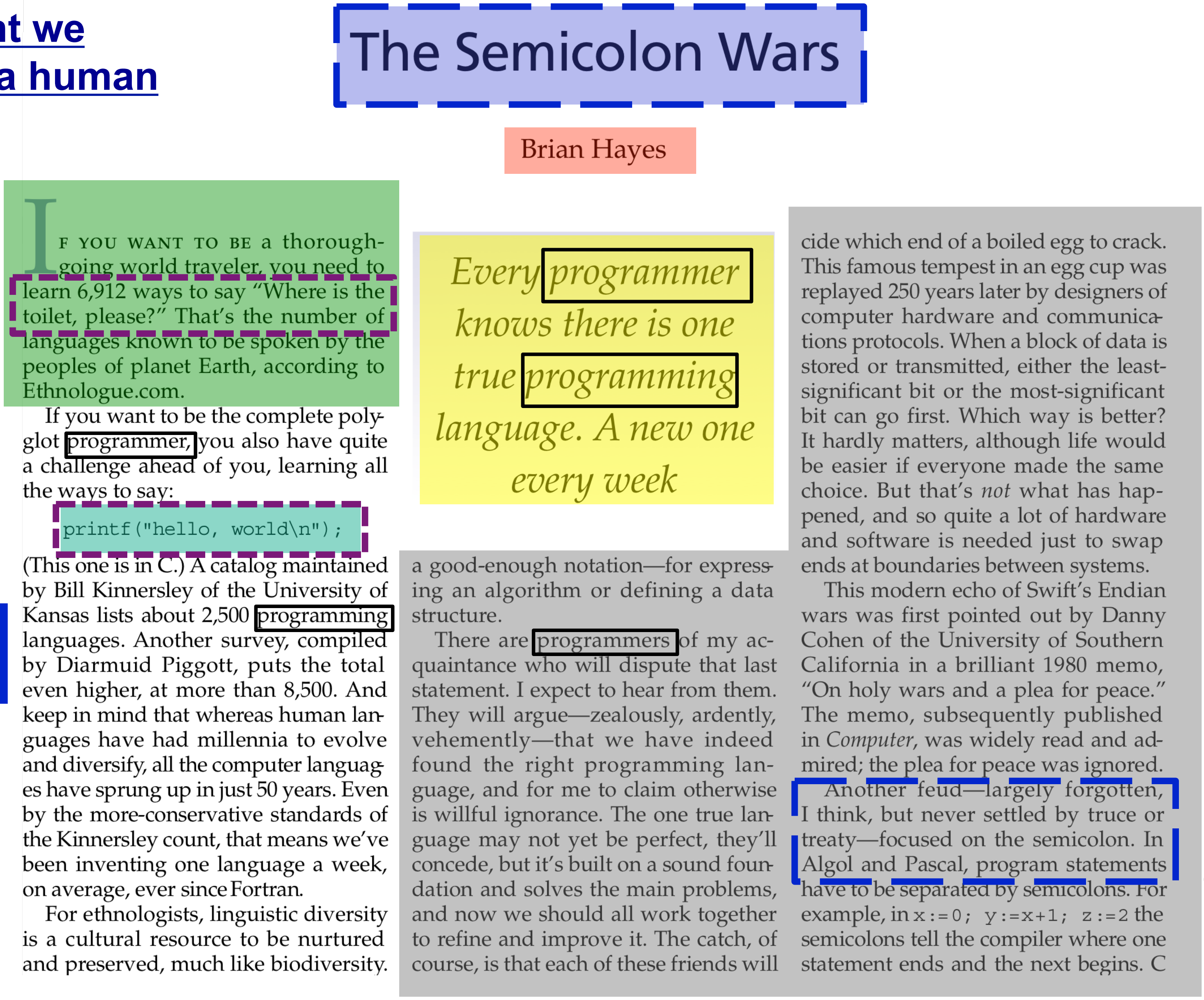
The New York Times

Beyond the Gene

Only 1.2 percent of the genome is made up of classic genes. Scientists are exploring the other 99 percent, in the process uncovering new secrets and facing new questions.

[NY Times, 11-Nov-08]

How might we annotate a human text?



The Semicolon Wars

Brian Hayes

If you want to be a thoroughgoing 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.

> *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 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

Color is Function

Lines are Similarity

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

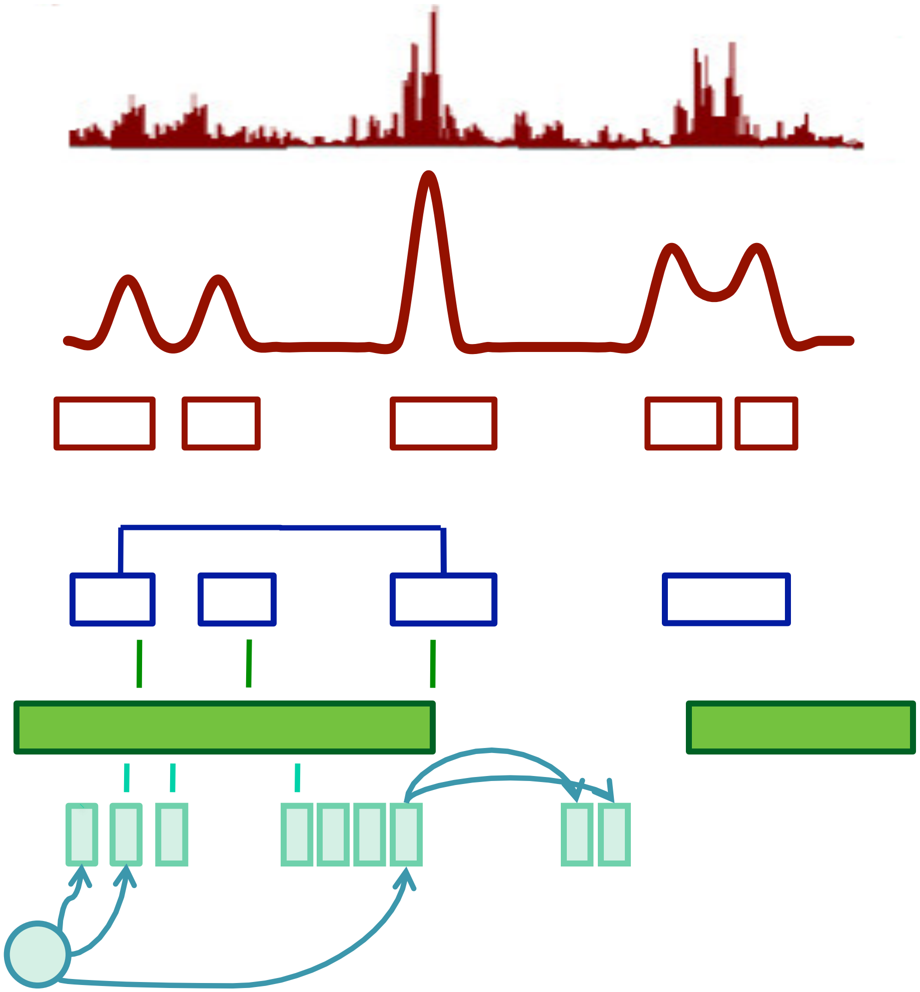
Overview of the Process of Annotation of non-coding Regions

- Basic Inputs
  1. Comparative Genomics.
     Doing large-scale similarity comparison, looking for repeated or deleted regions
  2. Functional Genomics.
     Determining experimental signals for activity (e.g. transcription) across each base of genome

- **Comparative Genomics**

  Finding repeated or deleted blocks in the genome
  1. As a function of similarity (i.e. age, perhaps using explicit models)
  2. vs. other organisms, vs. human reference, or within the human population (synteny, SDs, and CNVs)
  3. Big and small blocks
     (duplicated regions and retrotransposed repeats)
  4. Creation of formal annotations (e.g. genes and pseudogenes)

Overview of Functional Genomics Annotation Process

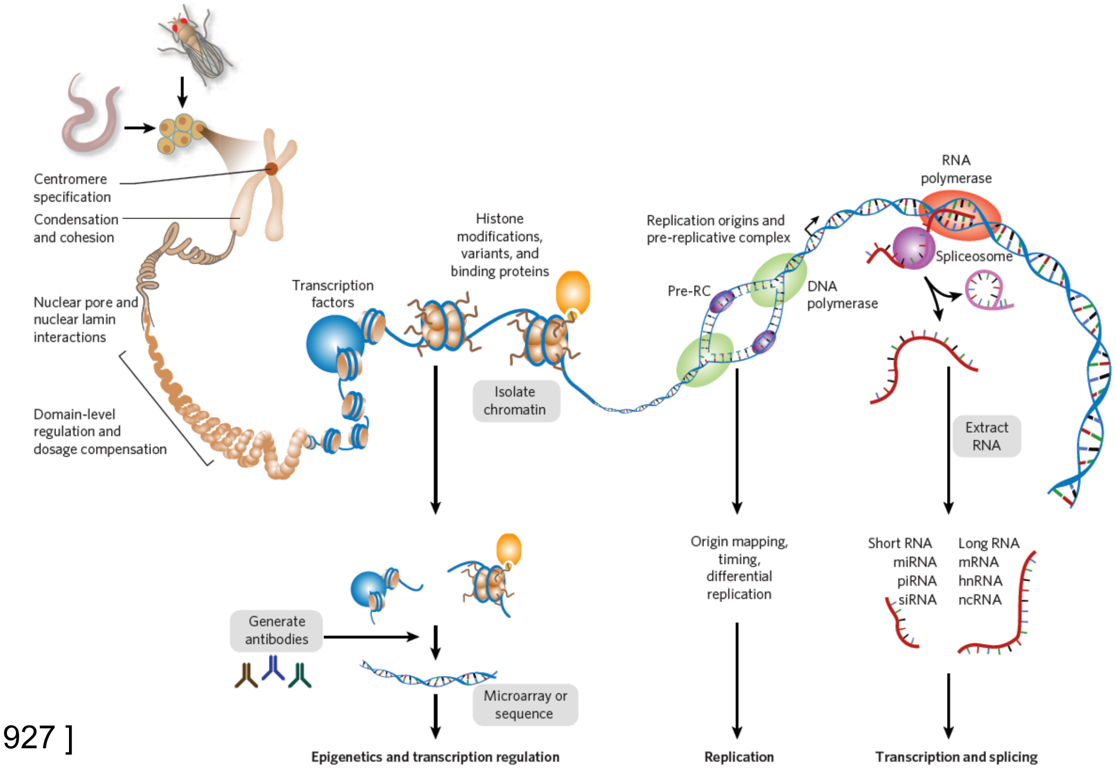


- **Determining experimental signals for biochemical activity (e.g. transcription) across each base of genome**
  - Development of Sequence (and Array) Technology
    - Normalizing & Scoring Signal, Correcting Artifacts, Segmenting to create Small Annotation Blocks
  - Output of Production Pipelines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

ENCODE + modENCODE Consortia for functional annotation & 1KG Consortium for variable blocks in human population

mod ENCODE

ENCODE


Centromere specification
Condensation and cohesion
Nuclear pore and nuclear lamin interactions
Domain-level regulation and dosage compensation
Transcription factors
Histone modifications, variants, and binding proteins
Isolate chromatin
Replication origins and pre-replicative complex
Pre-RC
DNA polymerase
RNA polymerase
Spliceosome
Extract RNA
Generate antibodies
Microarray or sequence
Origin mapping, timing, differential replication
Short RNA
miRNA
piRNA
siRNA
Long RNA
mRNA
hnRNA
ncRNA
Epigenetics and transcription regulation
Replication
Transcription and splicing

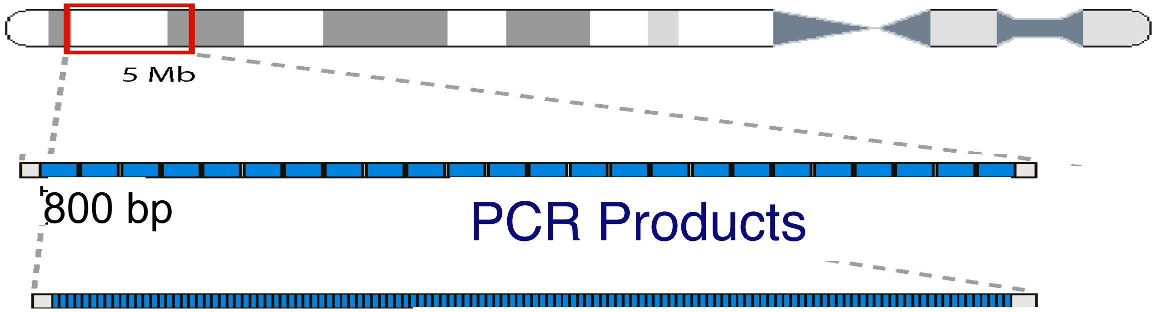
[Celniker et al., Nature ('09) 459:927]

1000 Genomes

A Deep Catalog of Human Genetic Variation

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

Tiling Arrays



‘02

800 bp

PCR Products

‘04

36mer

Oligonucleotide Array

Massively Parallel Sequencing

‘06+

→

AGTTCACCTAAGA…

CTTGAATGCCGAT…

GTCATTCCGCAAT…

Application in a variety of contexts:

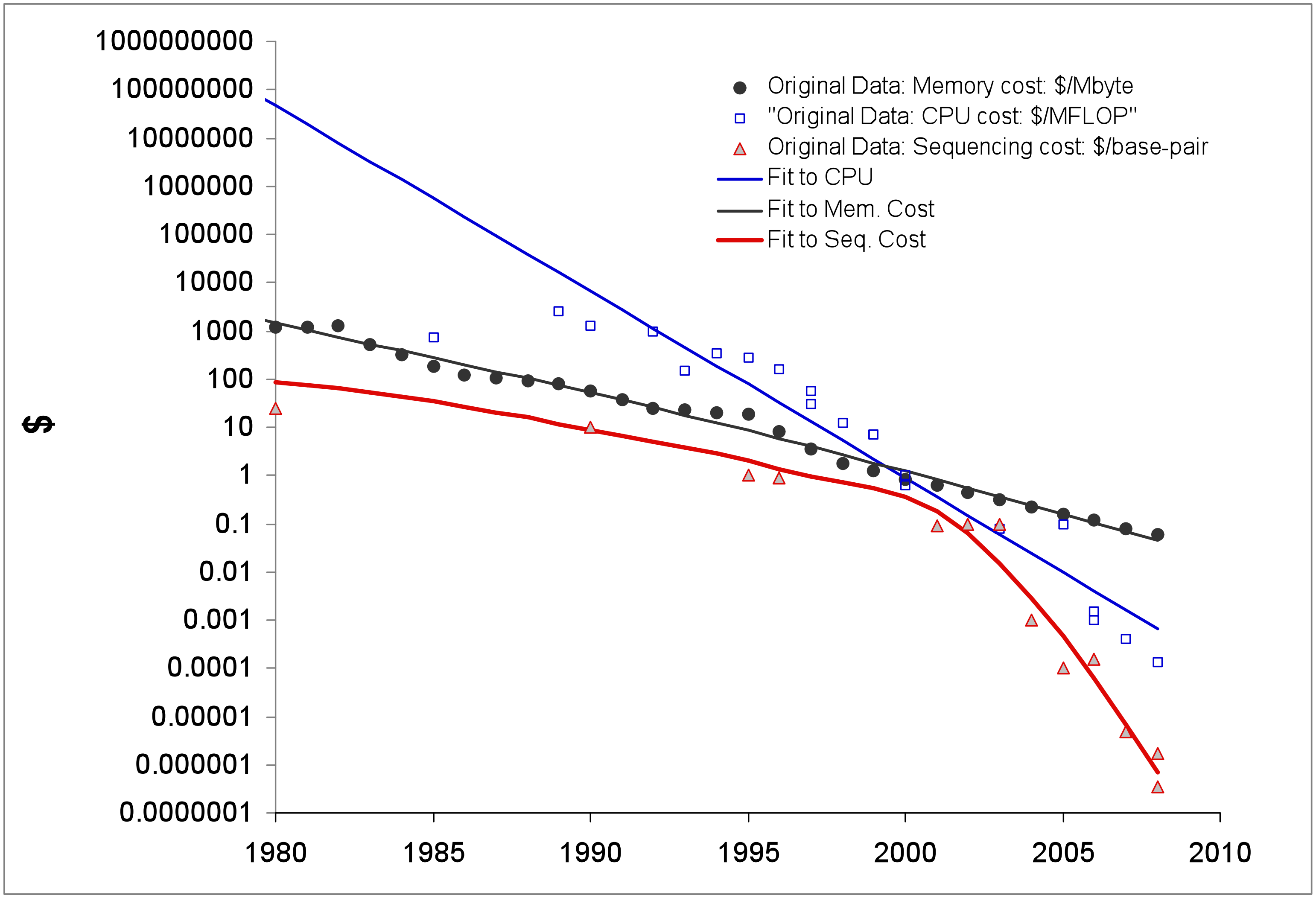
Transcription Mapping

DNA binding (inc. chromatin struc.)

Replication

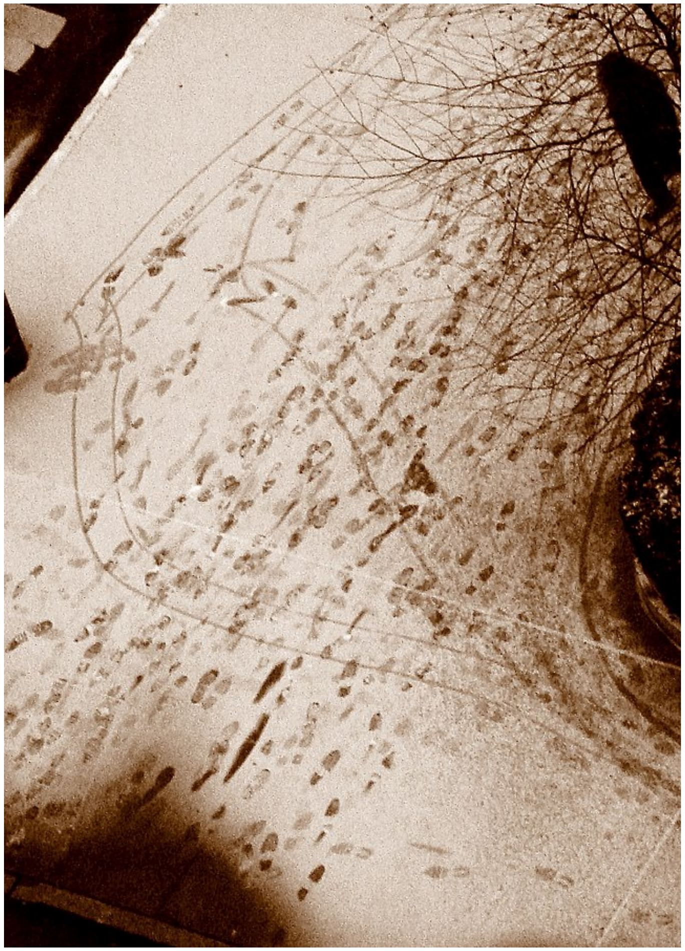
Structural Variation

Plummeting Cost of Sequencing



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

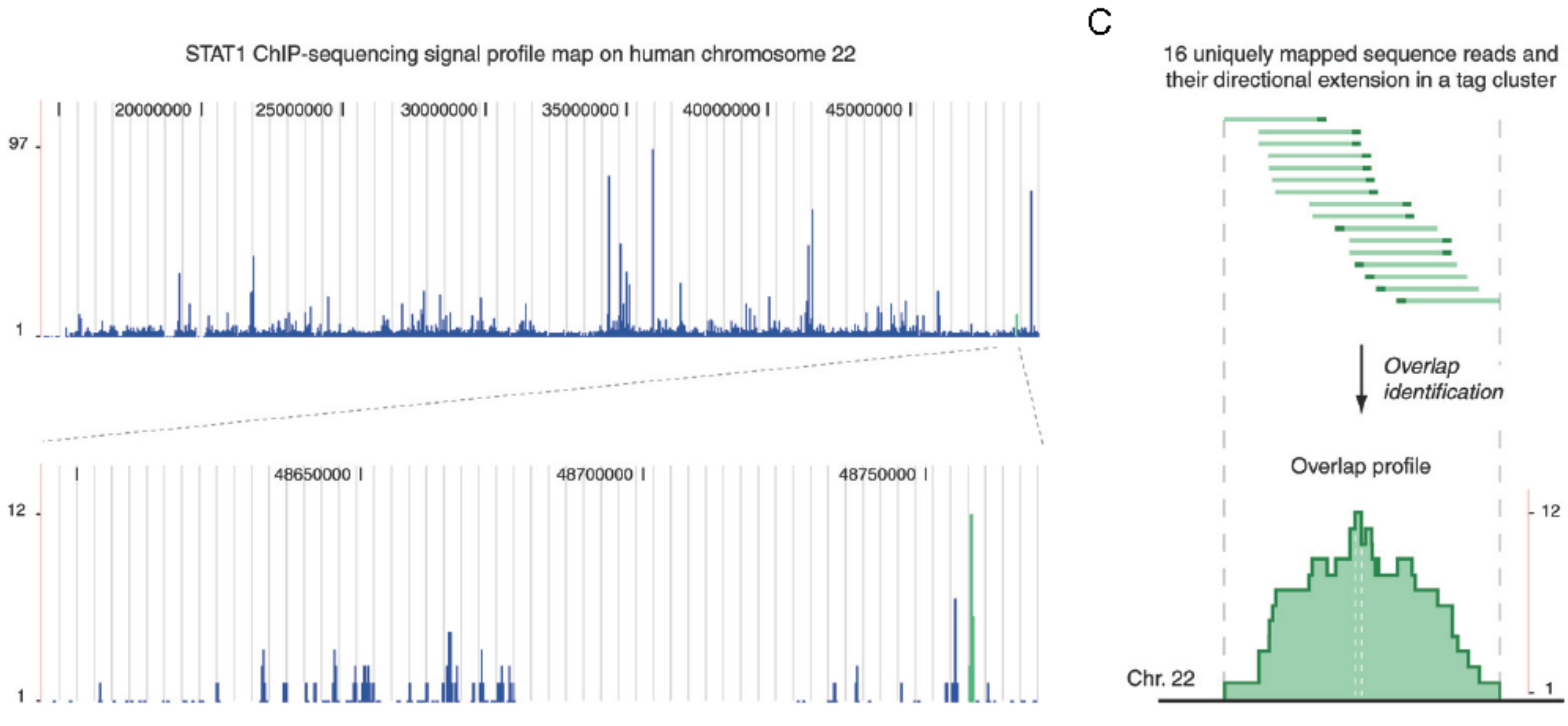
Outline



- Regulatory Sites
  - a. ChipSeq signal processing to call puncate "hits"
  - b. Clustering of hits into broader blocks and annotating them
- Variable Blocks in Genome (CNVs,SDs)
  - A/a. Calling them with various signal processing approaches (MSB, PEMer, ReSeqSim)
  - b. Grouping CNVs & SDs into larger features and inter-relating them
- Pseudogenes
  - A. Pattern-match tools for calling them
  - A. Focus on one group of pseudogenes
  - c. Integrating them with other annotations (transcription, regulation, CNVs, SDs)
- Future of Annotation
  - ◊ What is a "gene" post encode?

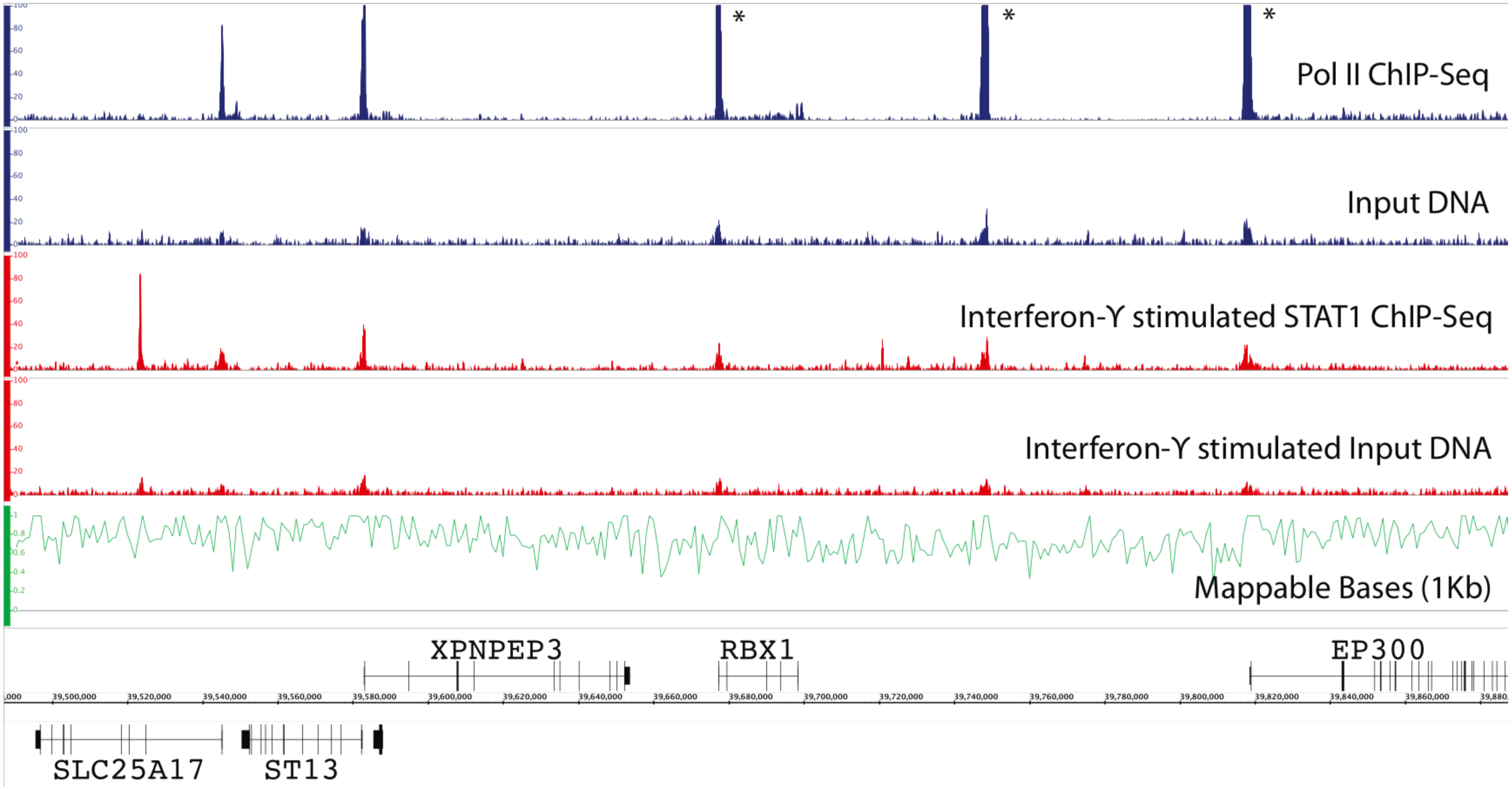
Signal Processing: Normalizing Signal and Finding Initial Annotation Blocks ("Hits")

Representative Signal from Chip-Seq

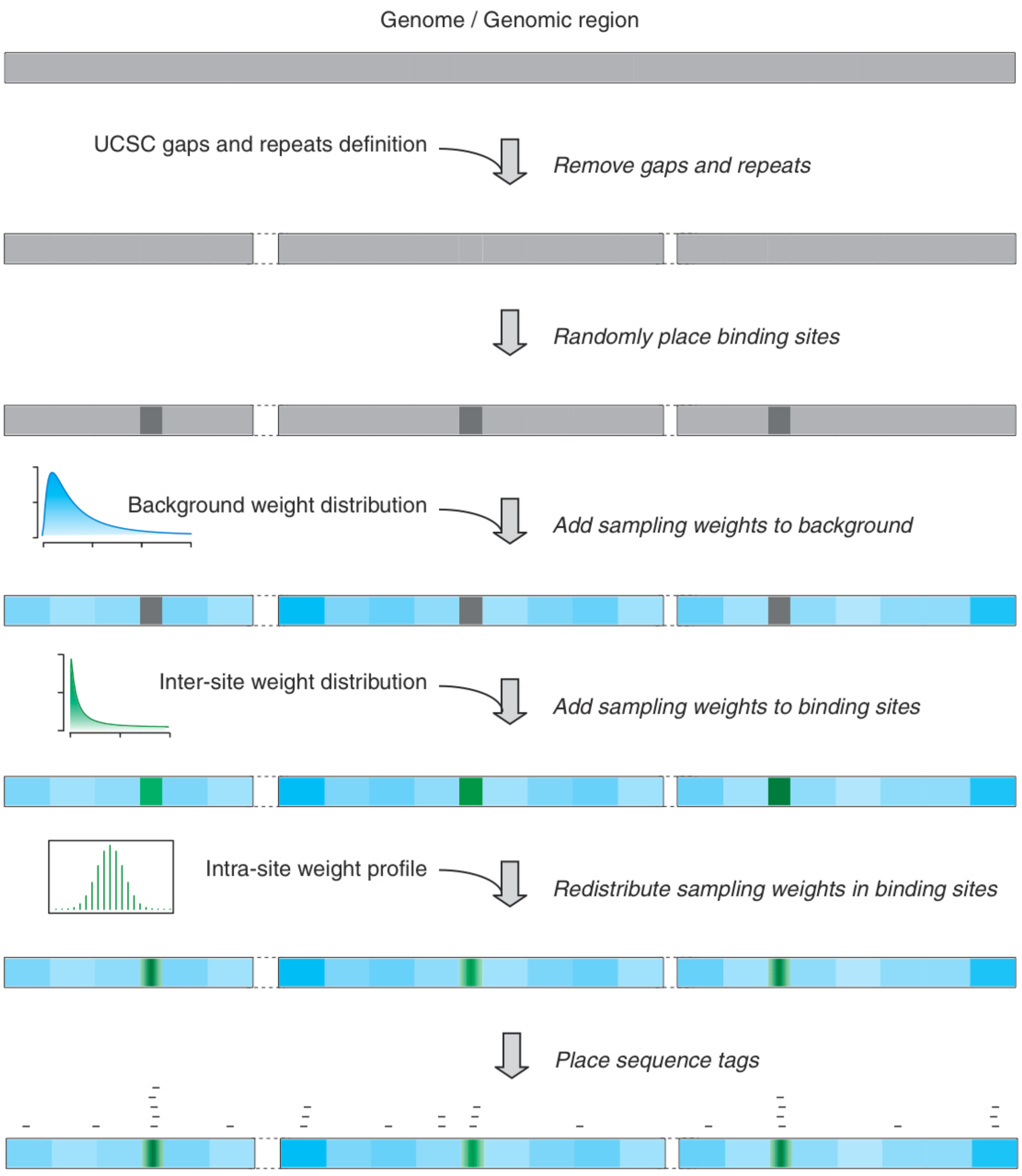


[Robertson et al., Nat. Meth. ('07); Zhang et al. PLOS Comp. Bio. (in revision, '08)]

ChIP-Seq vs Input DNA Control



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

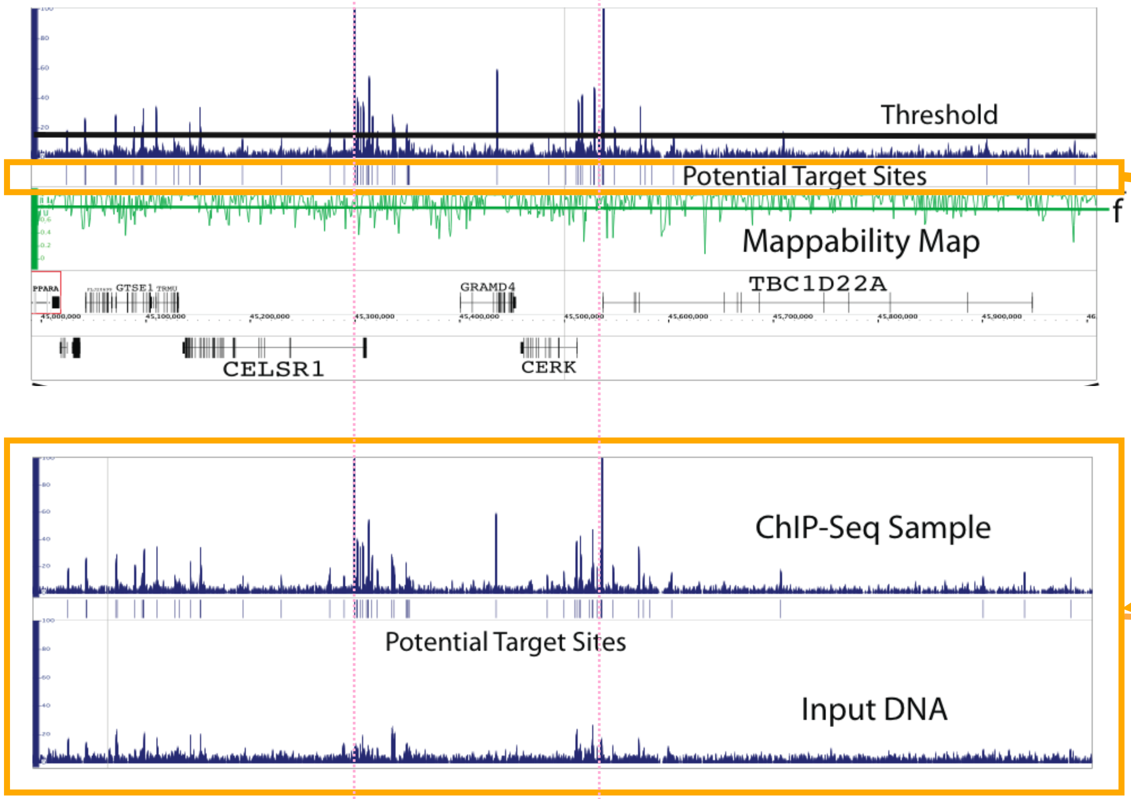

Genome / Genomic region
UCSC gaps and repeats definition
Remove gaps and repeats
Randomly place binding sites
Background weight distribution
Add sampling weights to background
Inter-site weight distribution
Add sampling weights to binding sites
Intra-site weight profile
Redistribute sampling weights in binding sites
Place sequence tags

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.

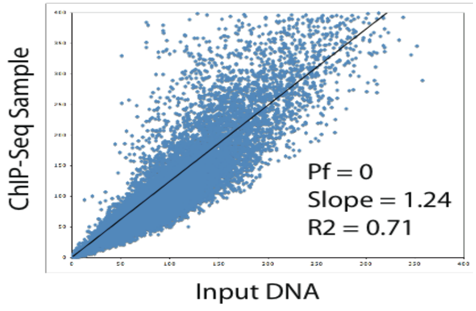
[Zhang et al. PLoS Comp Bio. ('08)]

PeakSeq: Scoring Relative to Controls

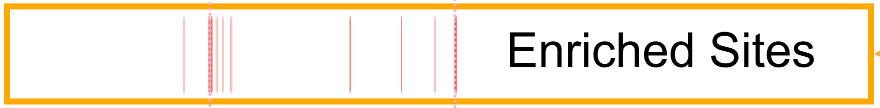


Filter for Potential Targets based on "Mappability" Simulation

Scale Input Relative to ChIP



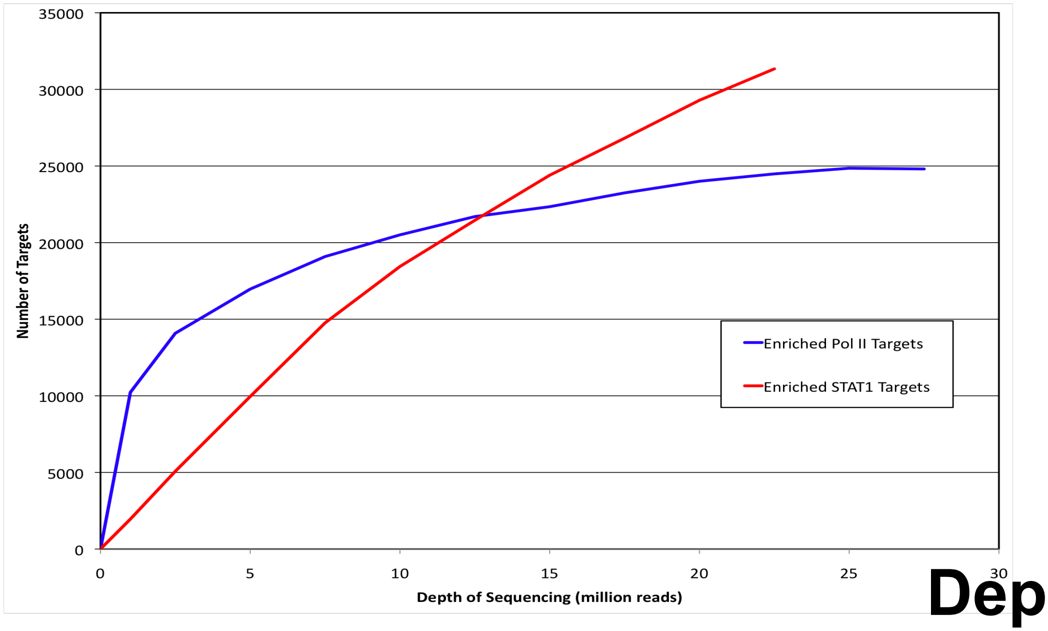
Score Relative to Bionomial Expectation



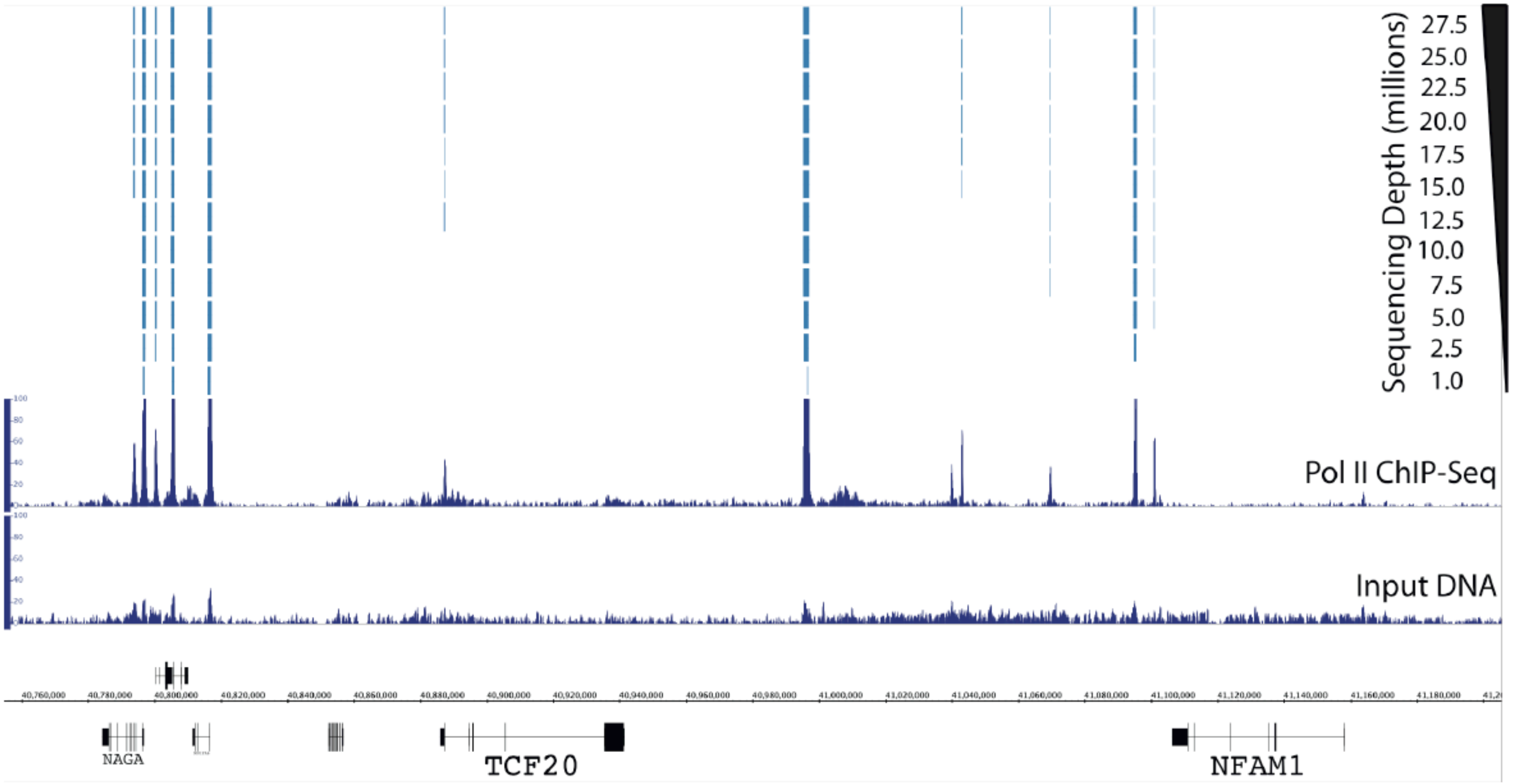
[Rozowsky et al. Nat. Biotech ('09)]

Number of Reads for Saturation

# Binding Sites



Depth

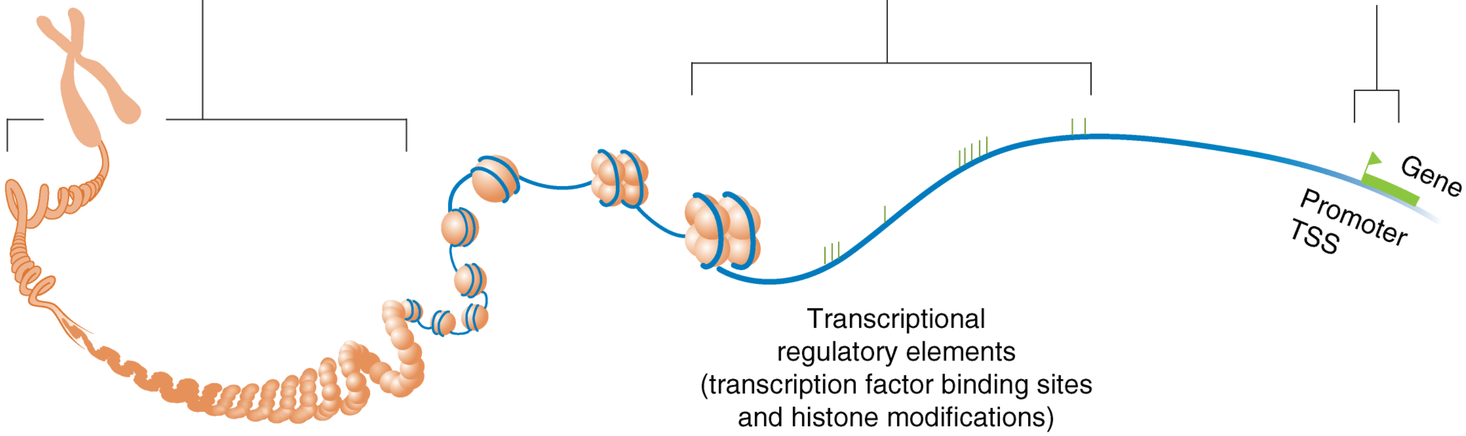


[Rozowsky et al. Nat. Biotech ('09)]

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



MACRO-GENOMIC SCALE

(Chromosome analysis)

MESO-GENOMIC SCALE

(TREs distribution analysis)

MICRO-GENOMIC SCALE

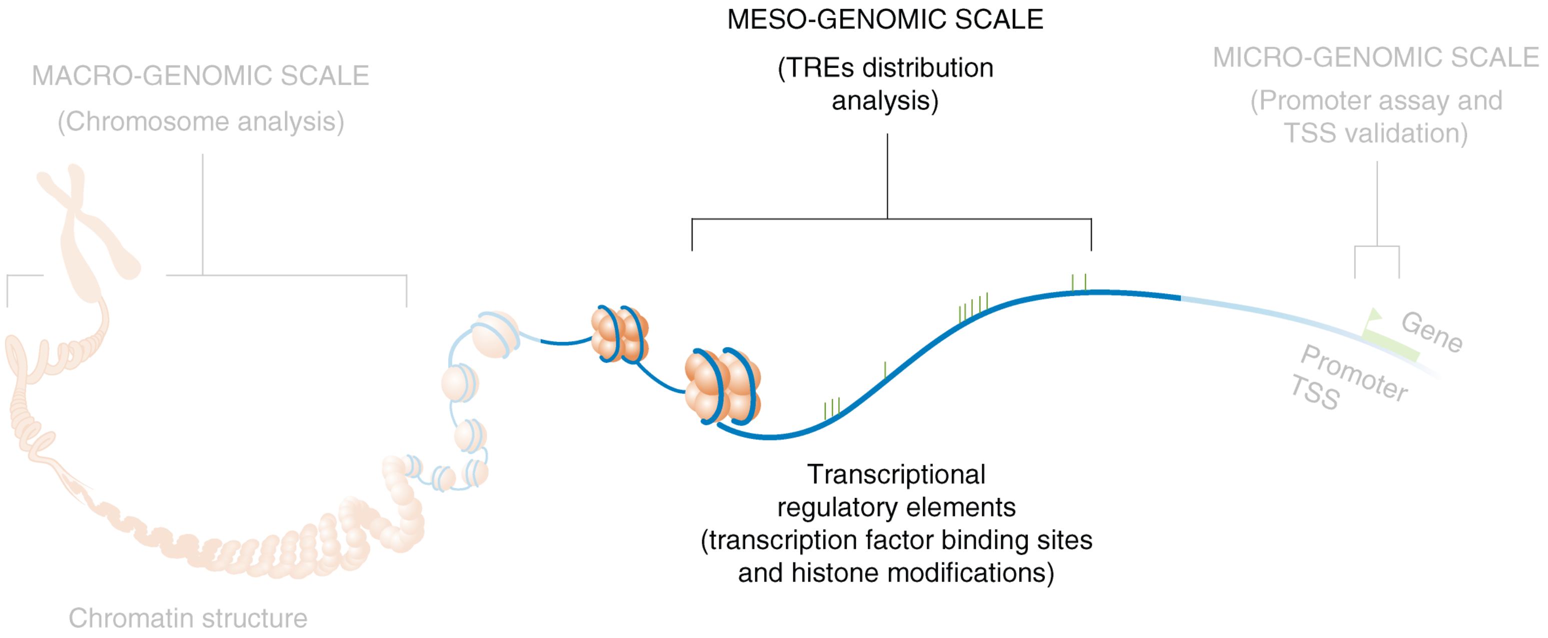
(Promoter assay and TSS validation)

Chromatin structure

Transcriptional regulatory elements (transcription factor binding sites and histone modifications)

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

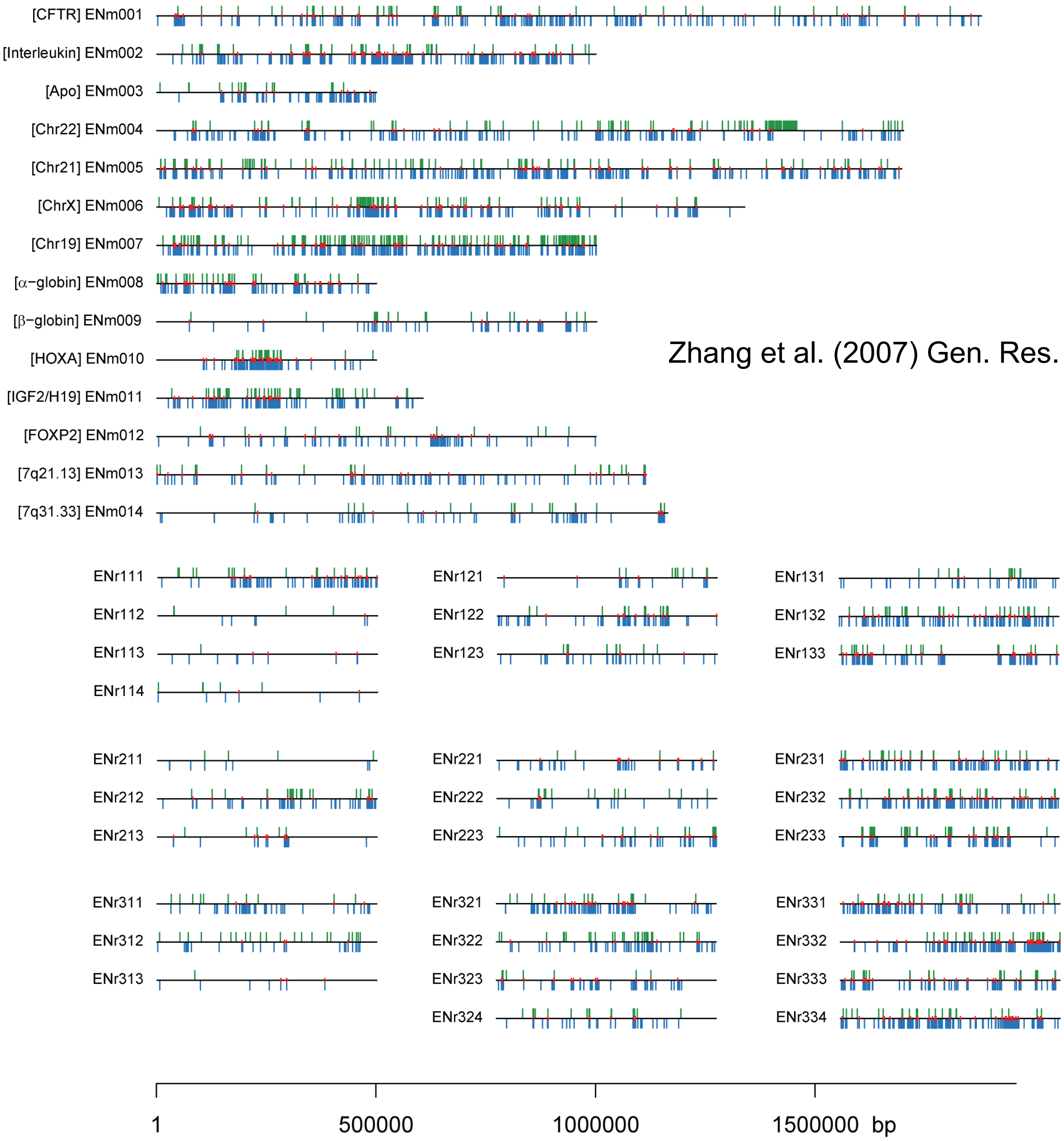
Clustering Binding Sites at ~50kb resolution


MESO-GENOMIC SCALE
(TREs distribution analysis)
MACRO-GENOMIC SCALE
(Chromosome analysis)
MICRO-GENOMIC SCALE
(Promoter assay and TSS validation)
Gene
Promoter
TSS
Transcriptional regulatory elements (transcription factor binding sites and histone modifications)
Chromatin structure

[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

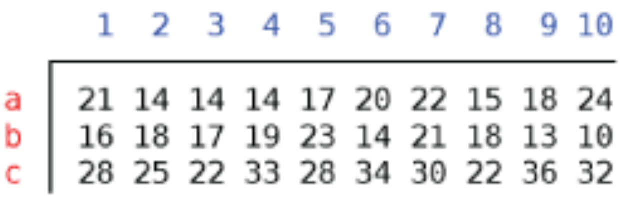


Zhang et al. (2007) Gen. Res.

Biplot to Show Overall Relationship of TFs and Genomic Bins

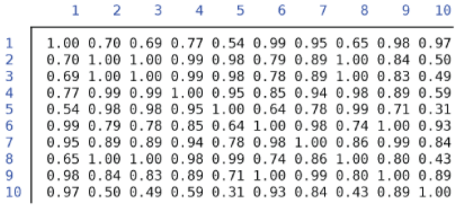
TFs: a, b, c...

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



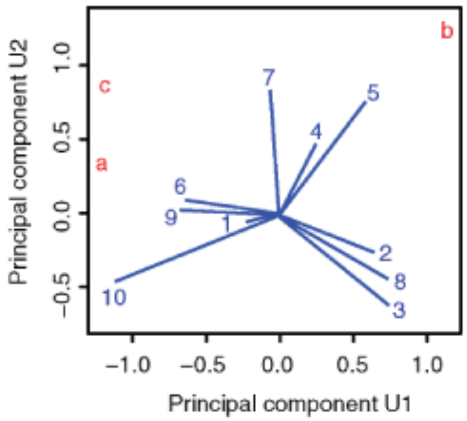
	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

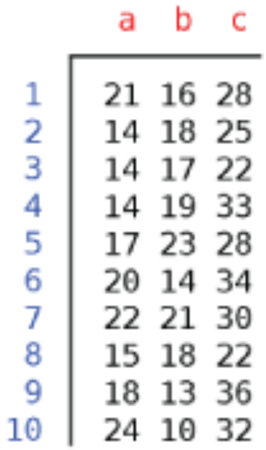
$A=USV^T$



	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

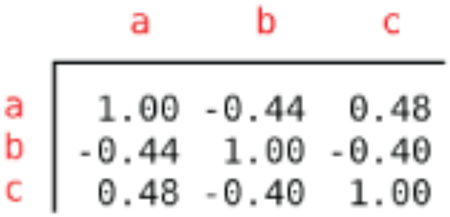
AA^T





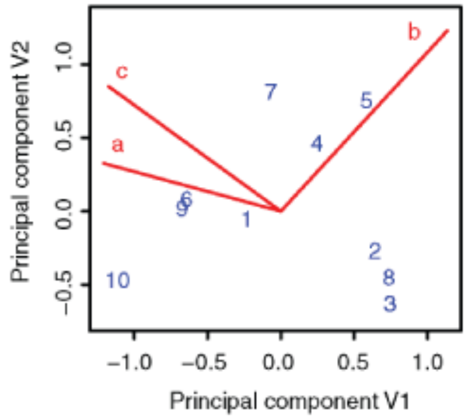
	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

A^T

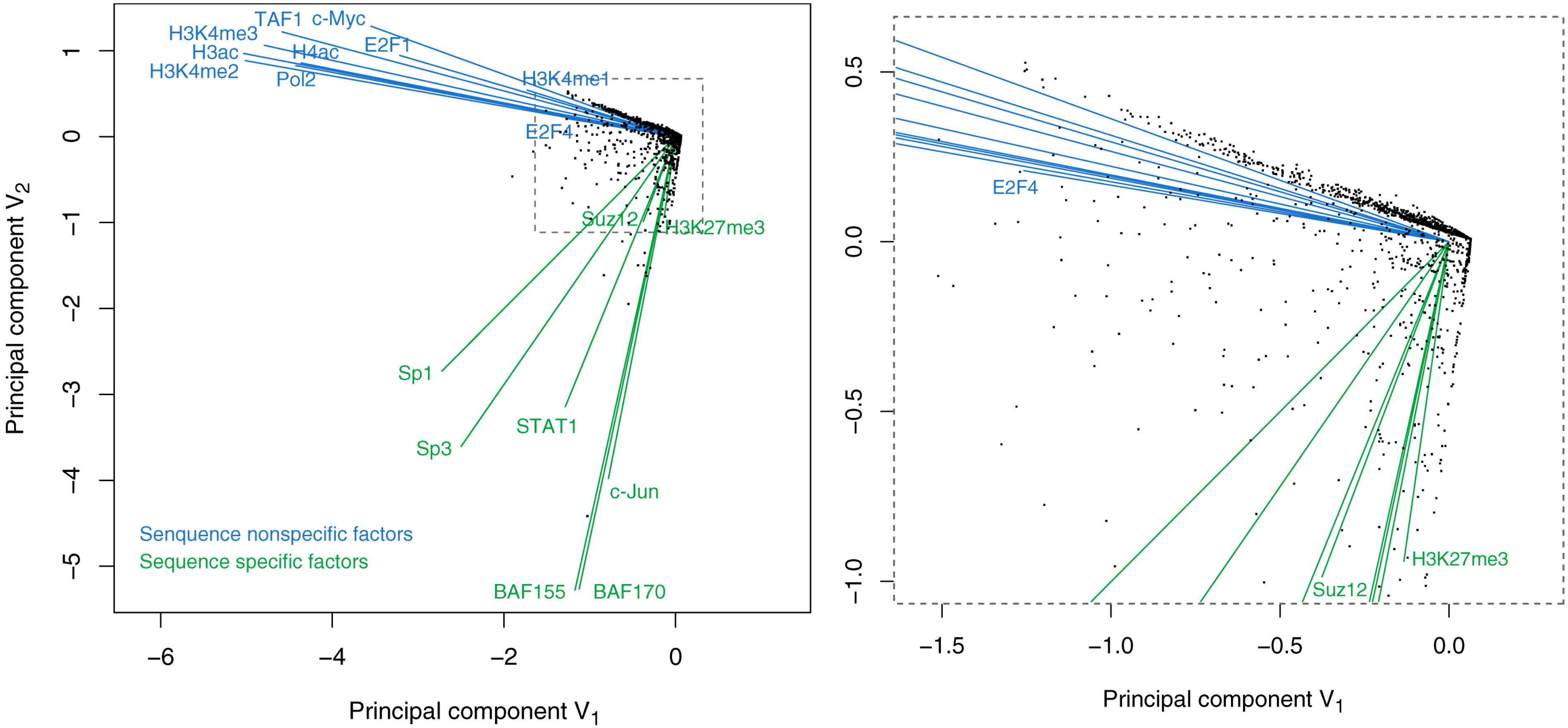


	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

A^TA



Lectures.GersteinLab.org (c) '09



Results of Biplot

Zhang et al. (2007)
Gen. Res.

- 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

Signal Processing 2: Finding Variable Blocks in the Human Genome

Terminology for Variable Elements in the Human Genome

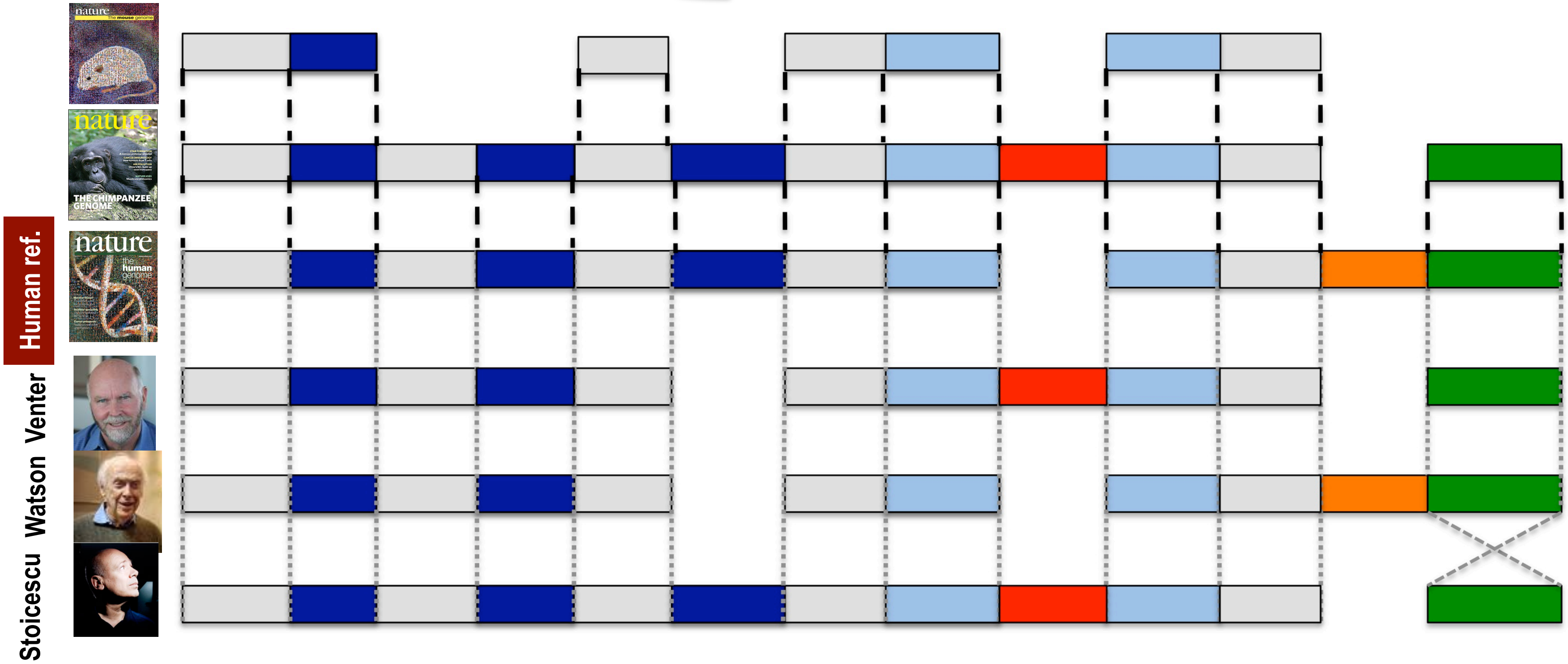
– – – – Syntenic regions

🟦 Old duplications

🟦 SDs: New duplications; during primate evolution

Structural variants (SVs)

- 🟩 Inversions
- Copy number variants (CNVs)
  - 🟦 Copy number polymorphic SDs
  - 🟥 Deletions in ref. w.r.t ancestral allele
  - 🟧 Insertions of novel seq. w.r.t ancestral allele



SDs ref : Bailey et al, Science, 2002

Terminology for Variable Elements in the Human Genome

- - - - Syntenic regions

🟦 Old duplications

🟦 SDs: New duplications; during primate evolution

SVs

Structural variants, potentialy the largest source of human variation

🟩 Inversions

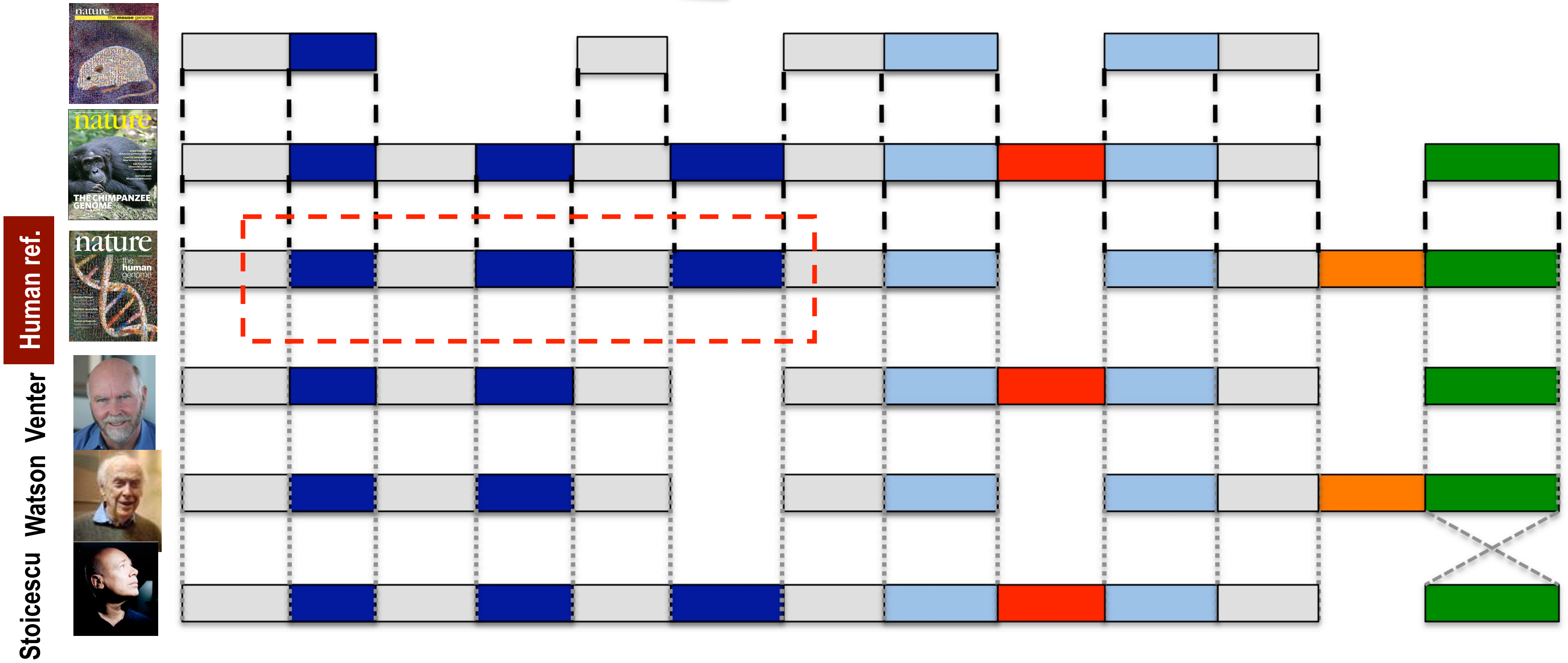
CNVs

Copy number variants

🟦 Copy number polymorphic SDs

🟥 Deletions in ref. w.r.t ancestral allele

🟧 Insertions of novel seq. w.r.t ancestral allele



SDs ref : Bailey et al, Science, 2002

Terminology for Variable Elements in the Human Genome

- - - - Syntenic regions

🟦 Old duplications

Structural variants

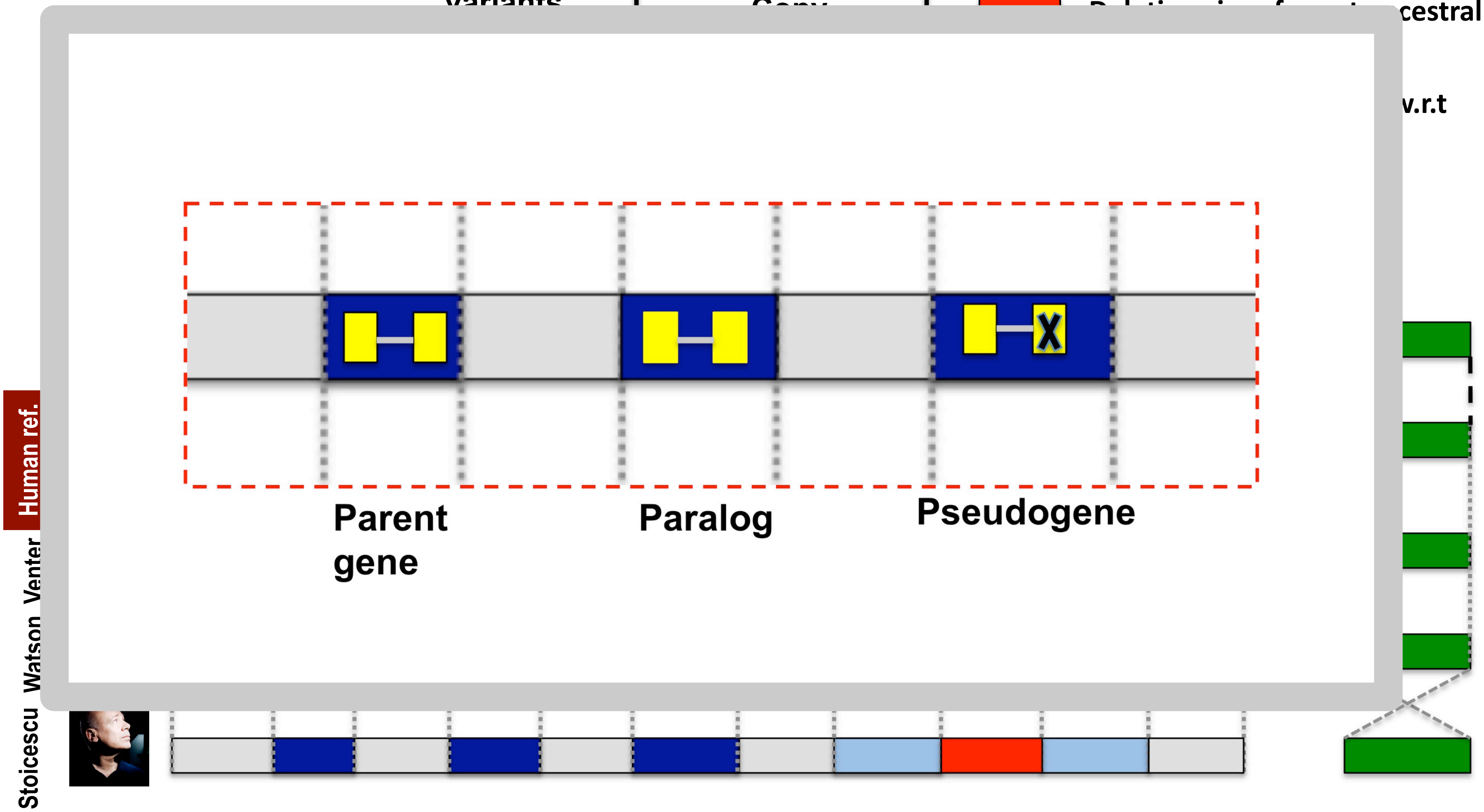
🟩 Inversions

Copy

🟦 Copy number polymorphic SDs

🟥 cestral

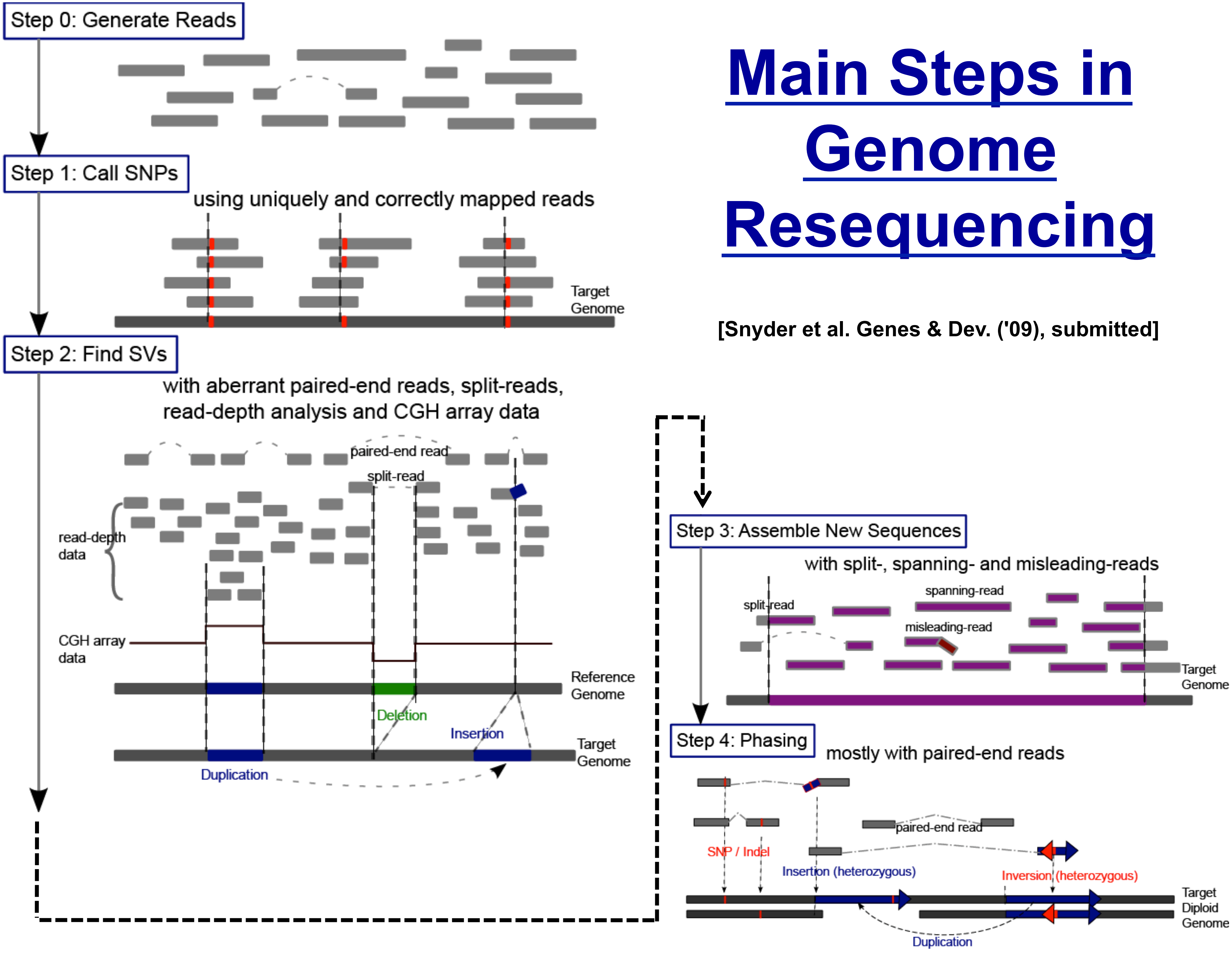
v.r.t



SDs ref : Bailey et al, Science, 2002

Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('09), submitted]



Step 0: Generate Reads

Step 1: Call SNPs

using uniquely and correctly mapped reads

Step 2: Find SVs

with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

Step 3: Assemble New Sequences

with split-, spanning- and misleading-reads

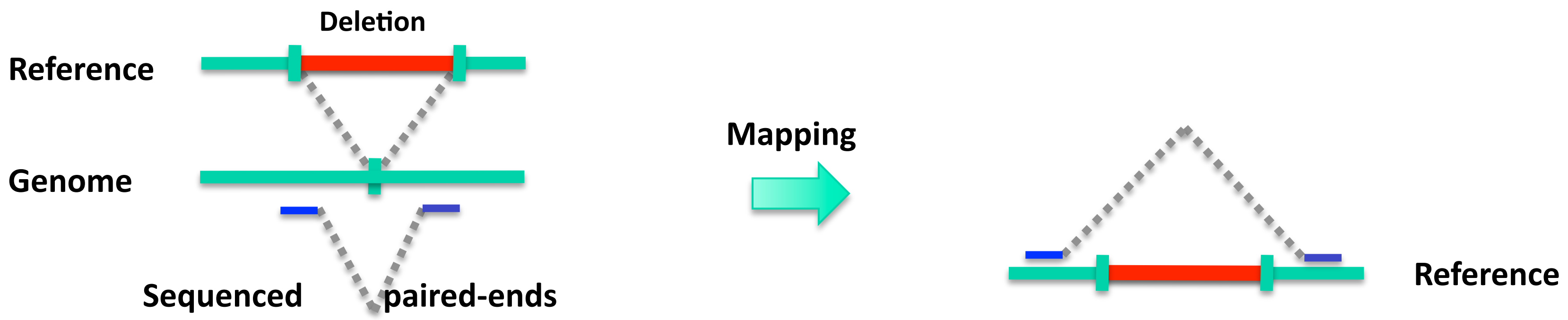
Step 4: Phasing

mostly with paired-end reads

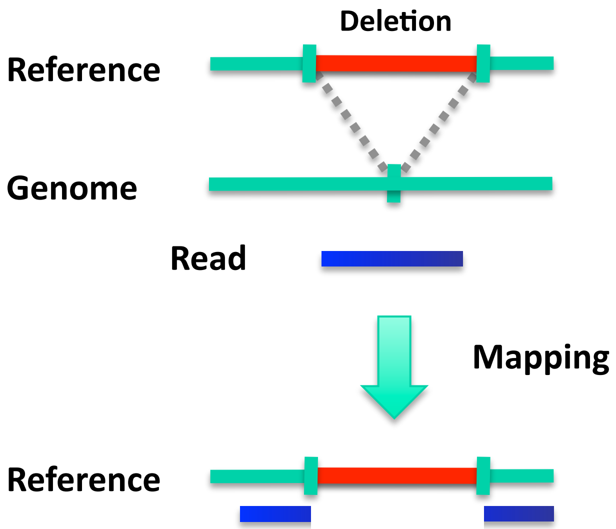
(c) '09

Methods to **Find SVs**

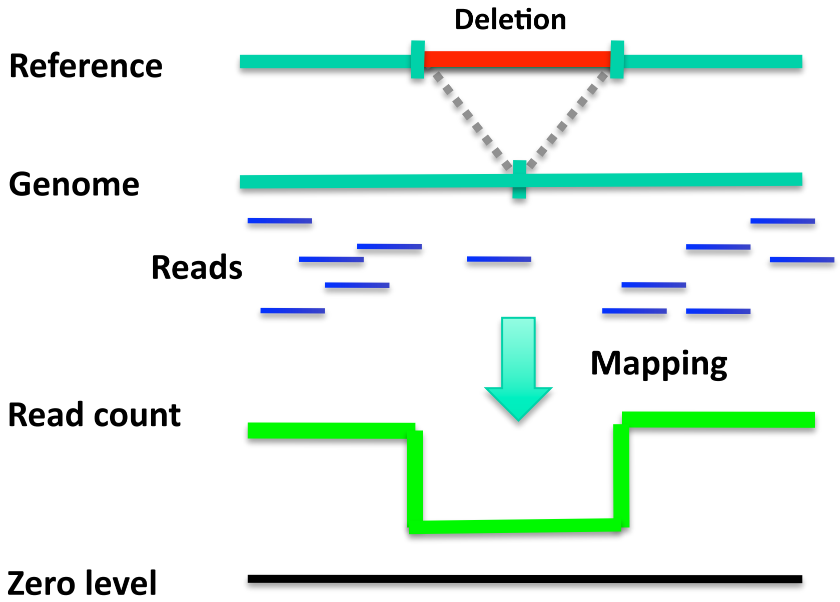
1. Paired ends



2. Split read



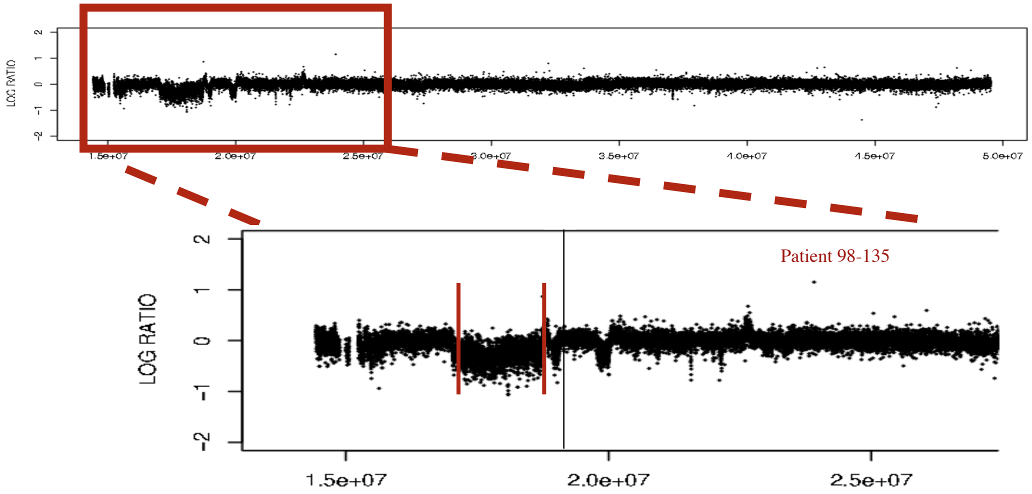
3. Read depth (or aCGH)



4. Local Reassembly

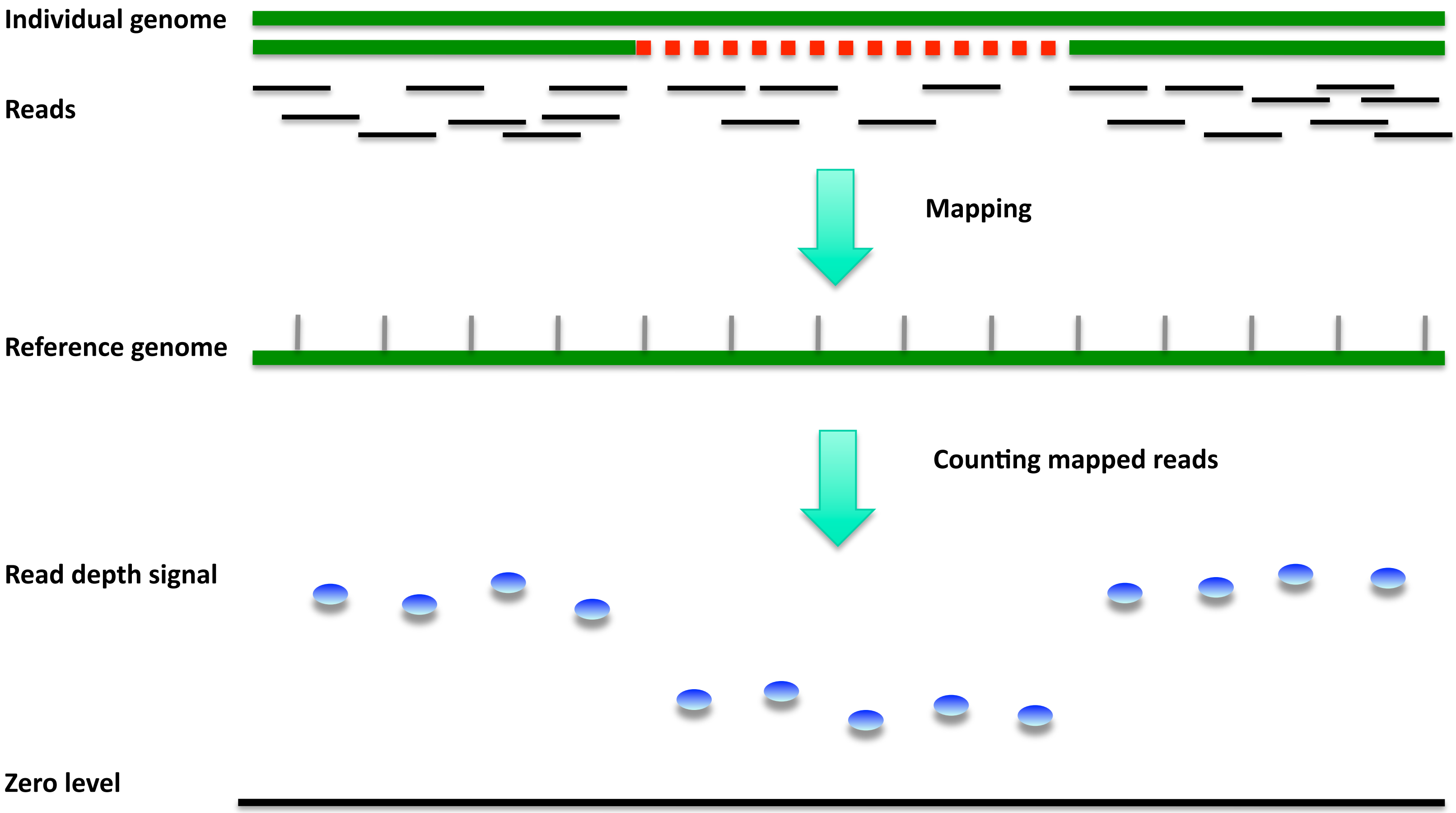
[Snyder et al. Genes & Dev. ('09), submitted]

Breakpointer: Segmentation of Array Signal as precursor to Read Depth



Array Signal

Read depth



Individual genome

Reads

Mapping

Reference genome

Counting mapped reads

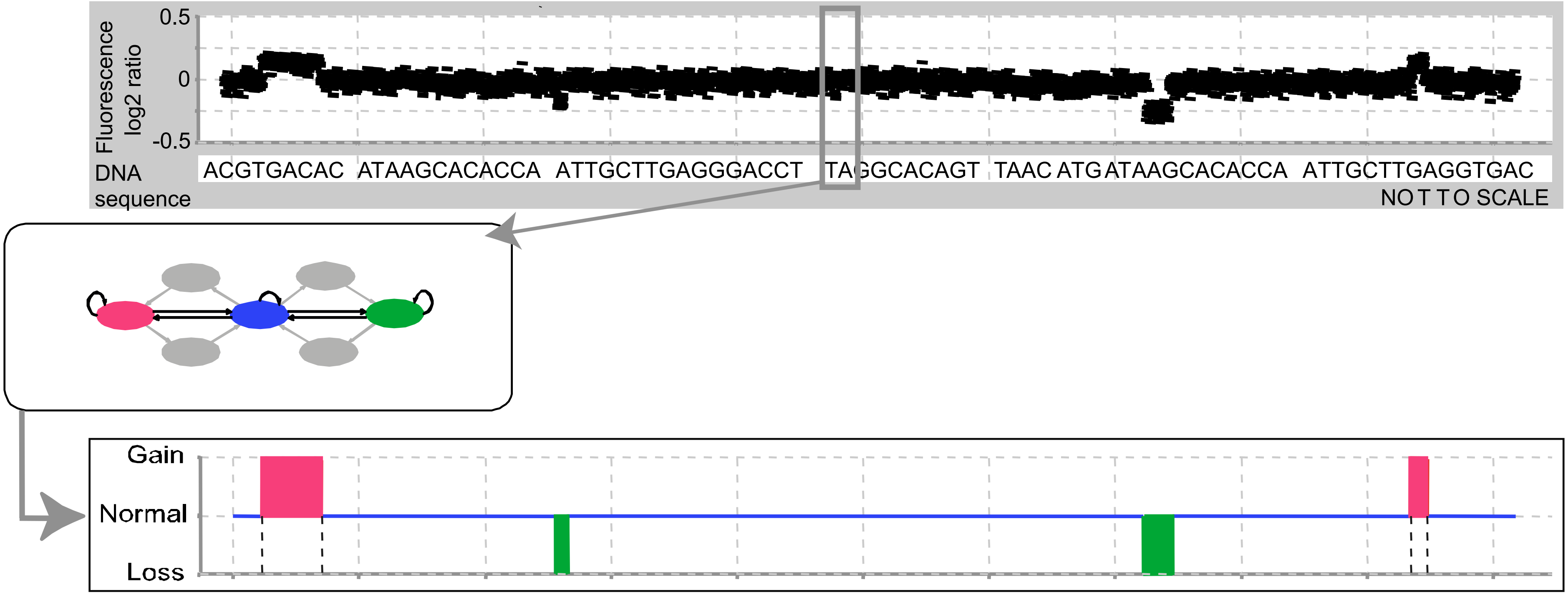
Read depth signal

Zero level

[Urban et al. ('06) PNAS; Wang et al. Gen. Res ('09)]

BreakPtr HMM

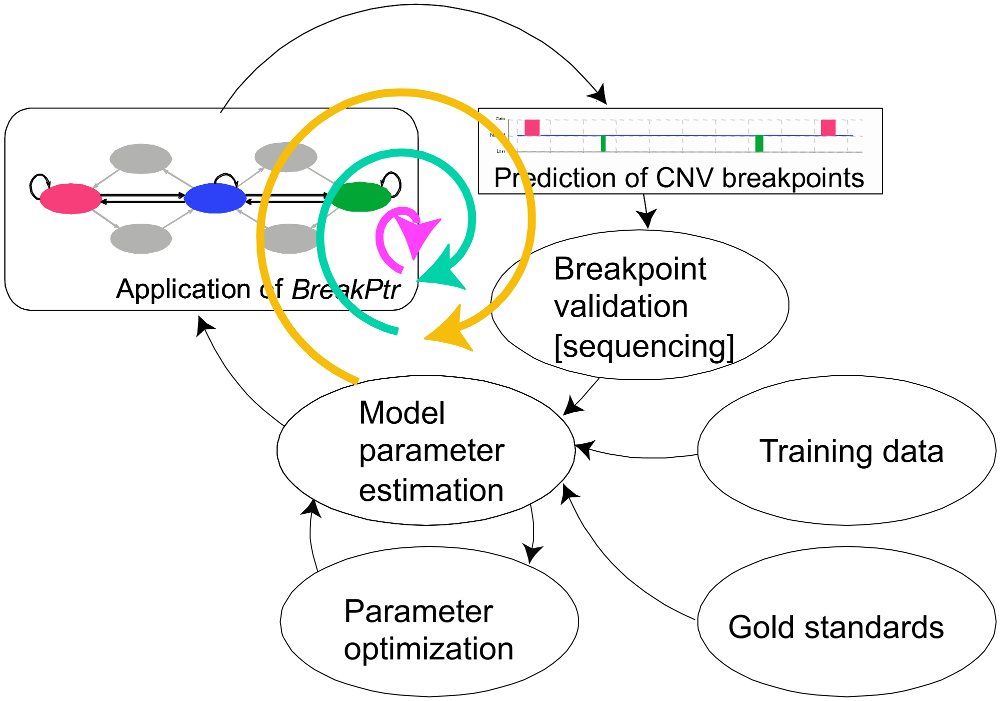
- To get highest resolution on breakpoints need to smooth & segment the signal
- BreakPtr: prediction of breakpoints, dosage and cross-hybridization using a system based on Hidden Markov Models



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

http://breakptr.gersteinlab.org

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



Application of *BreakPtr*

Prediction of CNV breakpoints

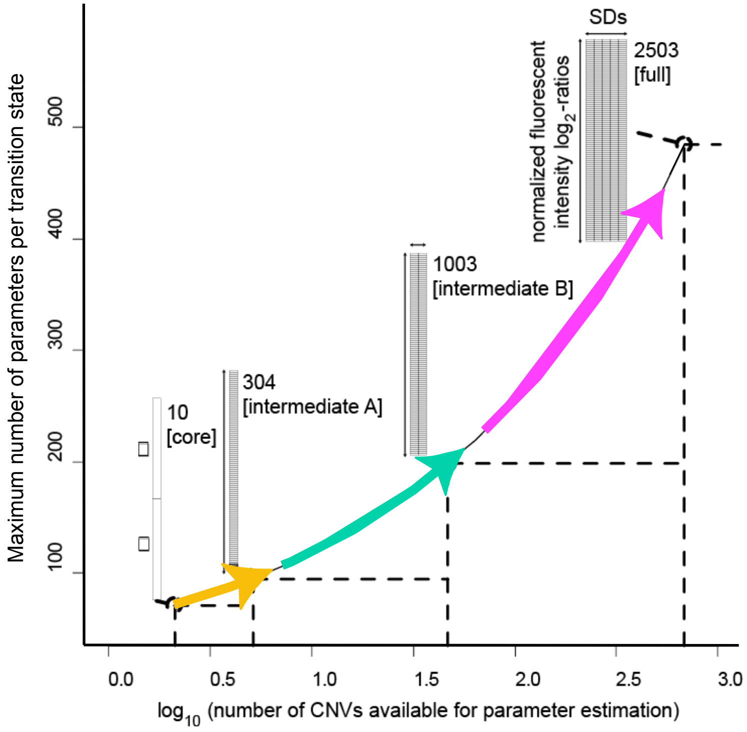
Breakpoint validation [sequencing]

Model parameter estimation

Training data

Parameter optimization

Gold standards



CNV breakpoints sequenced in ~10 cases following BreakPtr analysis;

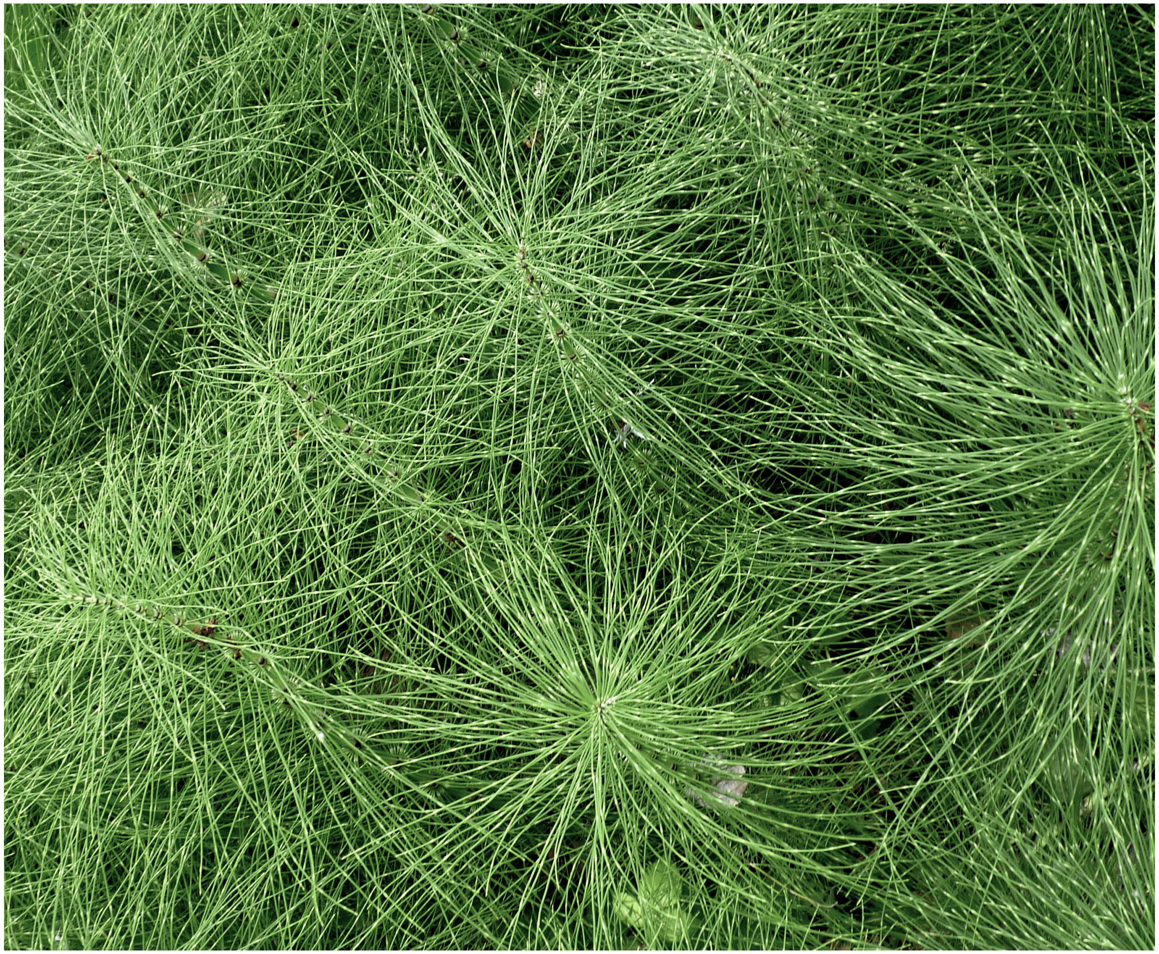
Median resolution <300 bp

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

HMM optimized iteratively
(using Expectation Maximization, EM)

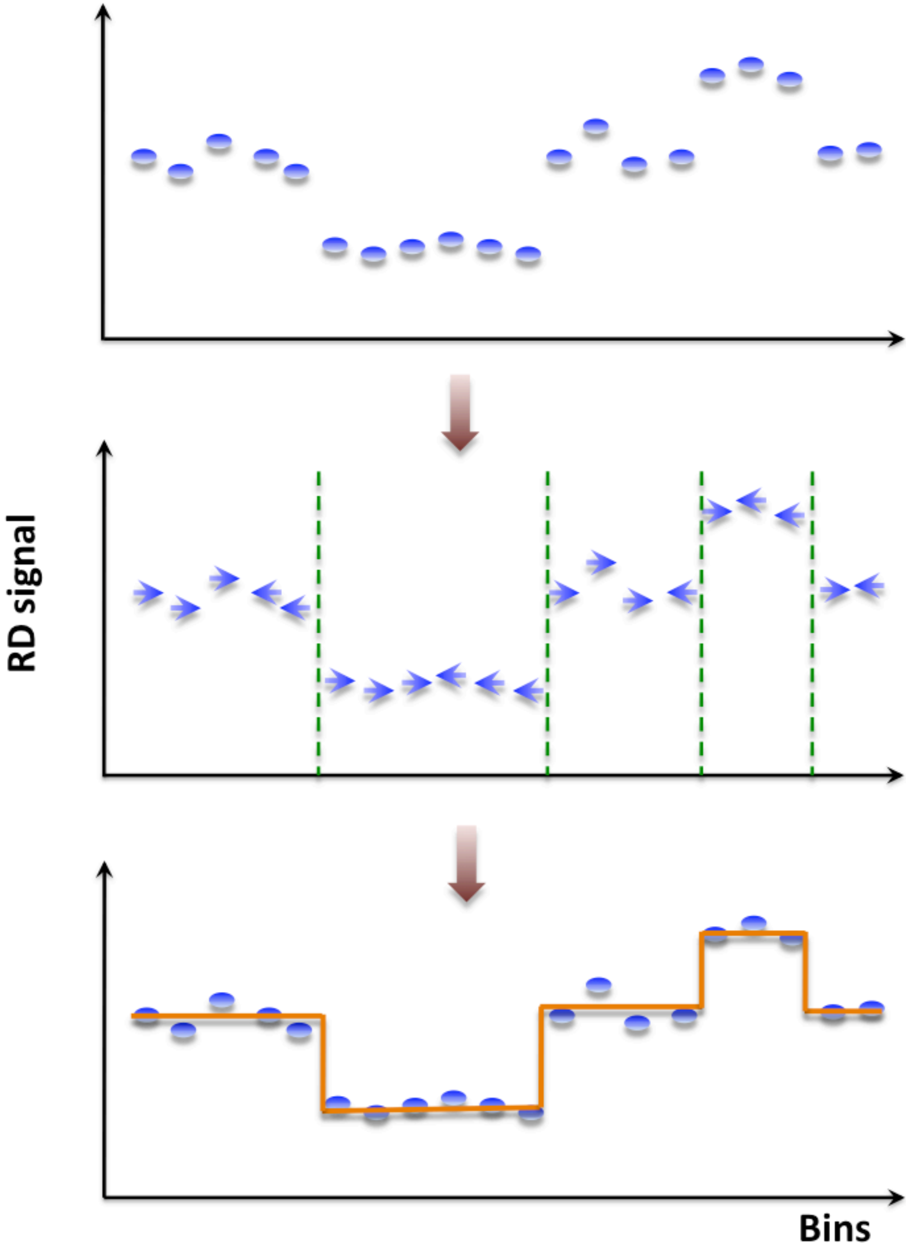
Korbel*, Urban* *et al.*, PNAS (2007)

MSB: Read-Depth Segmentation



Mean-shift-based (MSB) Segmentation: no explicit model

- For each bin attraction (mean-shift) vector points in the direction of bins with most similar RD signal
- No prior assumptions about number, sizes, haplotype, frequency and density of CNV regions
- Not Model-based (e.g. like HMM) with global optimization, distr. assumption & parms. (e.g. num. of segments).
- Achieves discontinuity-preserving smoothing
- Derived from image-processing applications



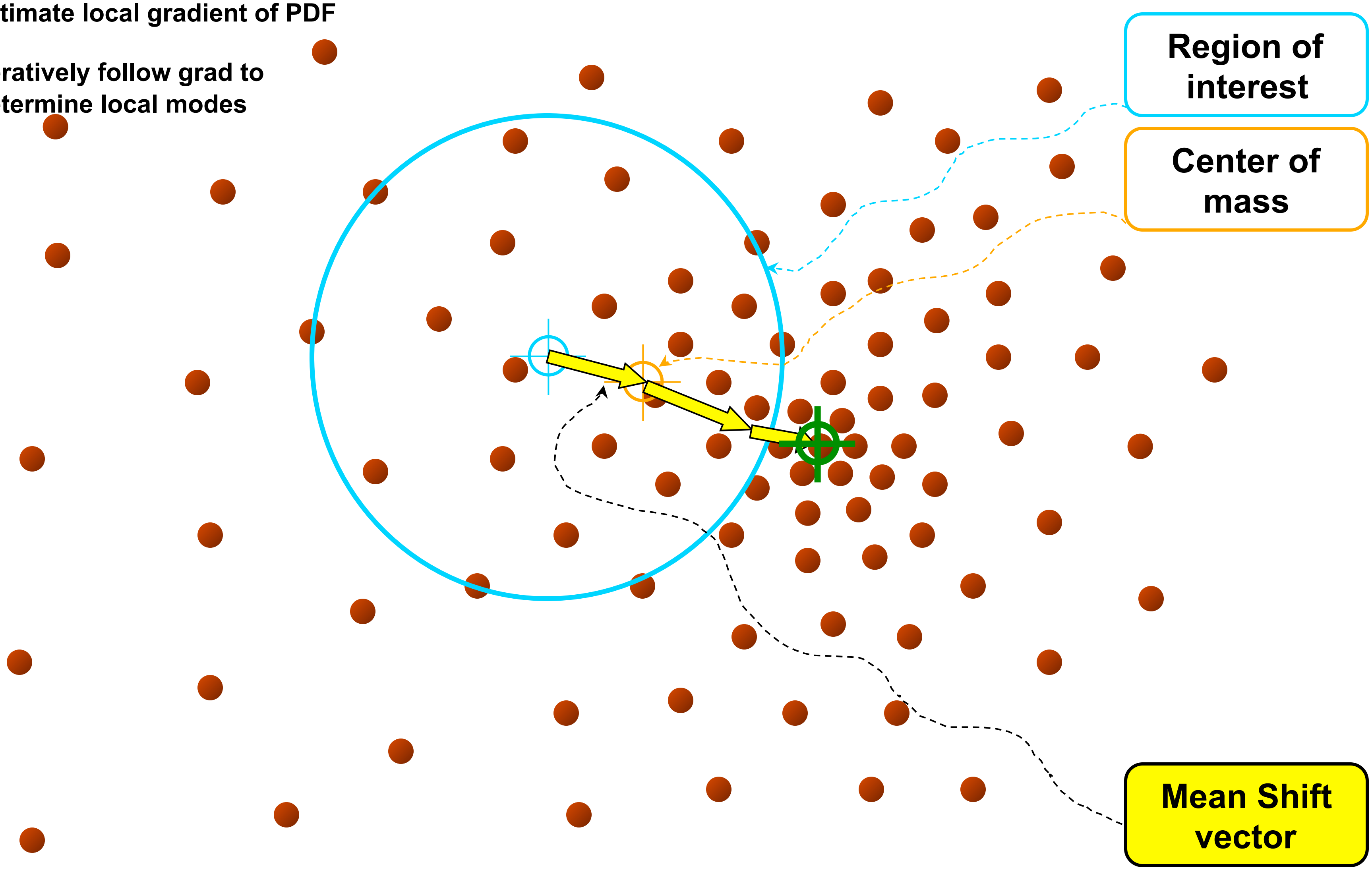
[Wang et al. Gen. Res ('09) 19:106]

Intuitive Description of MSB

● Observed depth of coverage counts as samples from PDF

➔ Kernel-based approach to estimate local gradient of PDF

⌖ Iteratively follow grad to determine local modes



Region of interest

Center of mass

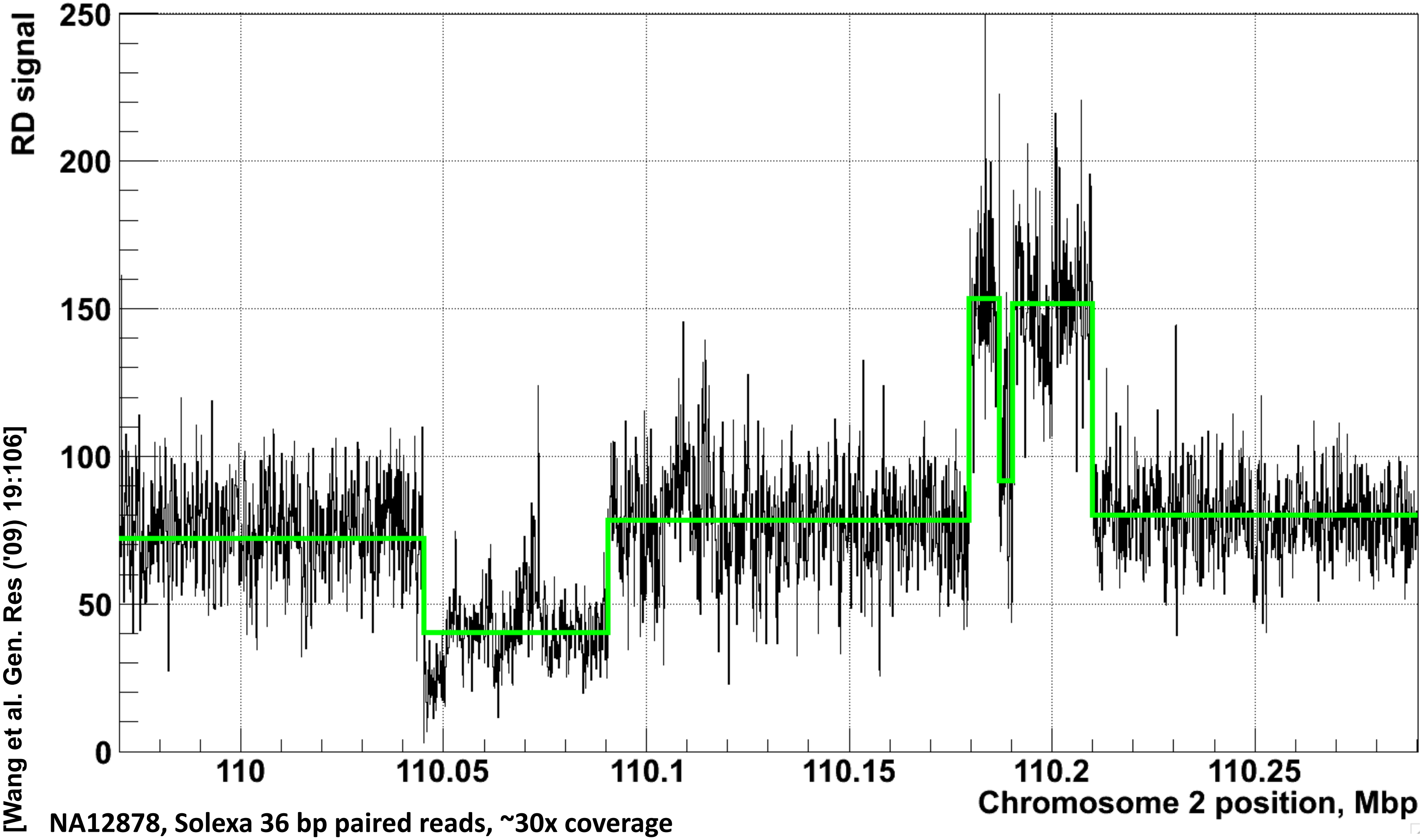
Mean Shift vector

Objective : Find the densest region

Distribution of identical billiard balls

[Adapted from S Ullman et al. "**Advanced Topics in Computer Vision,**" www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Example of Application of MSB to RD data

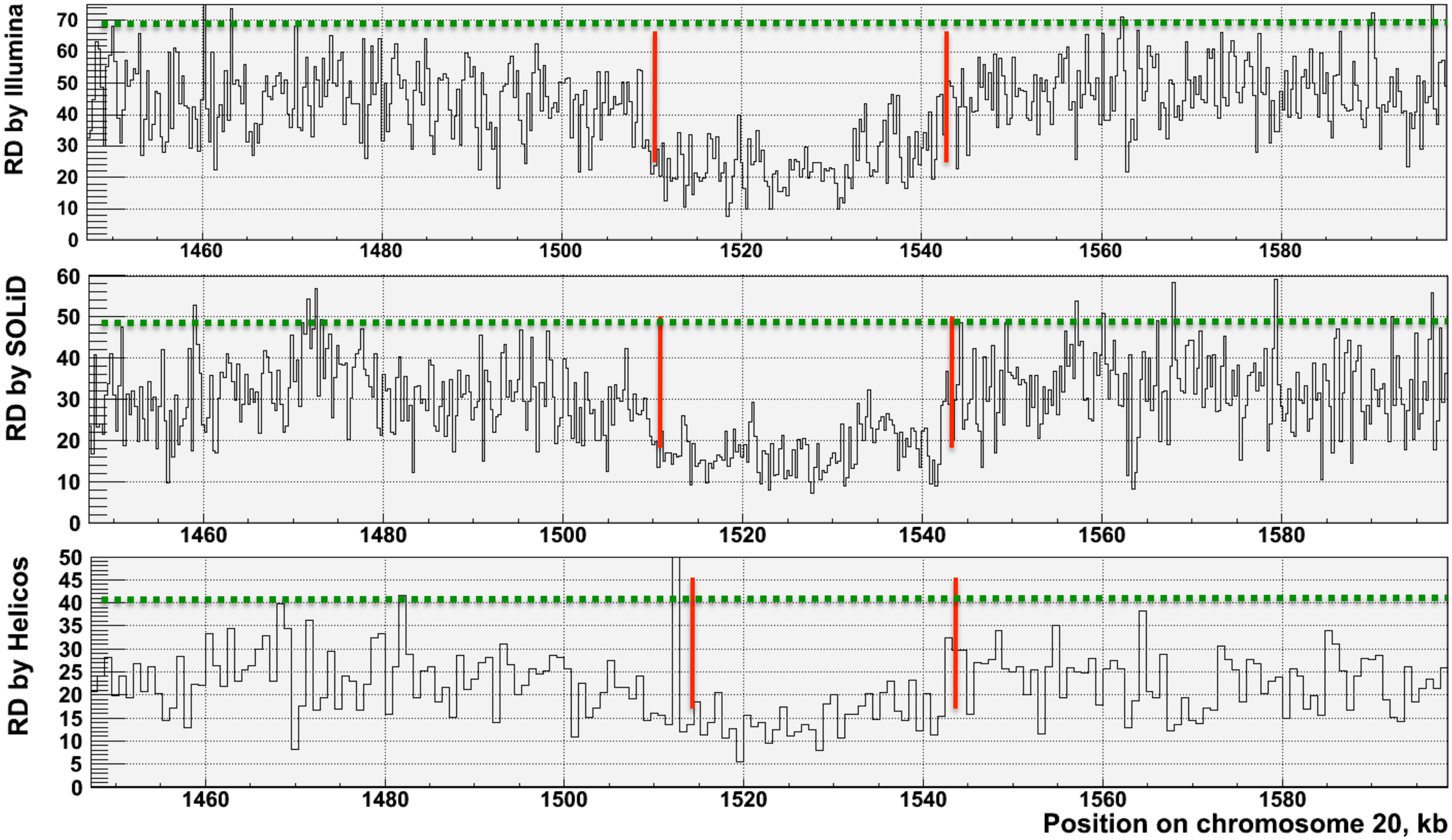


[Wang et al. Gen. Res ('09) 19:106]

NA12878, Solexa 36 bp paired reads, ~30x coverage

Lectures.GersteinLab.org (c) '09

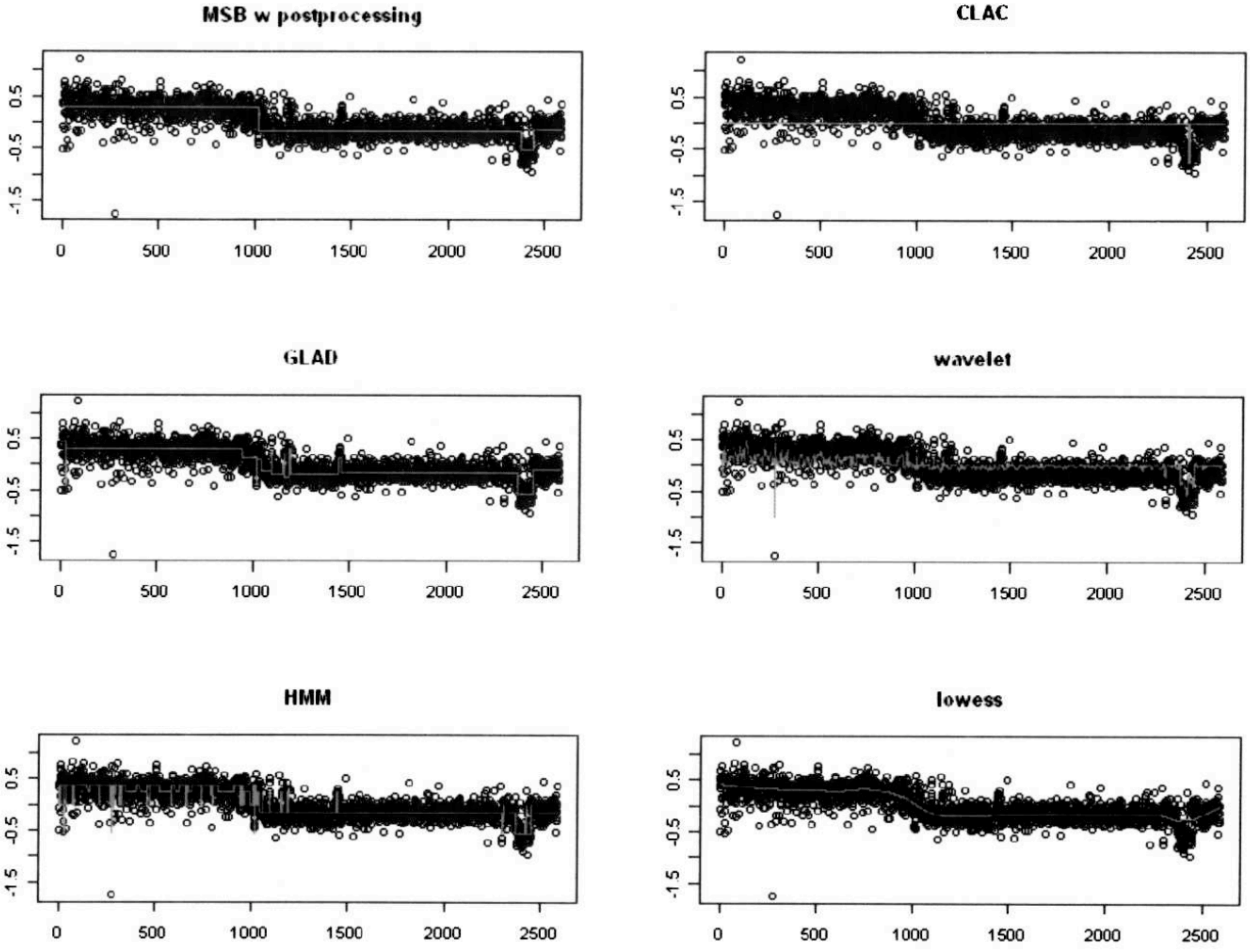
RD works well on a variety of sequencing platforms



[NA18505]

Lectures.GersteinLab.org (c) '09

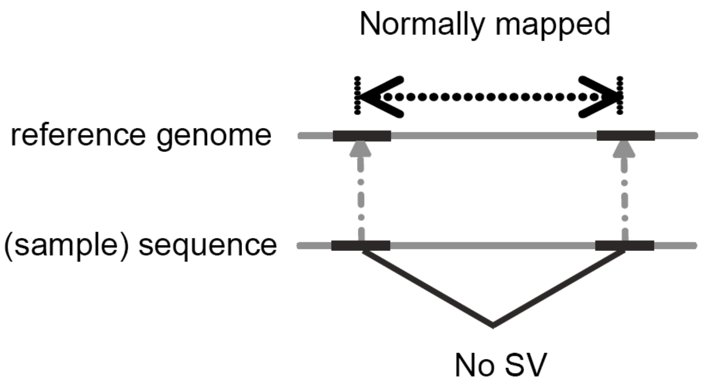
MSB works well on array data too

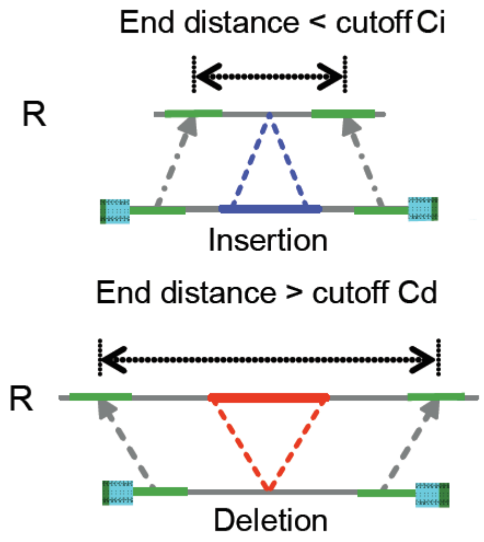


[Wang et al. Gen. Res ('09) 19:106]

Looking for Aberrantly Placed Paired Ends

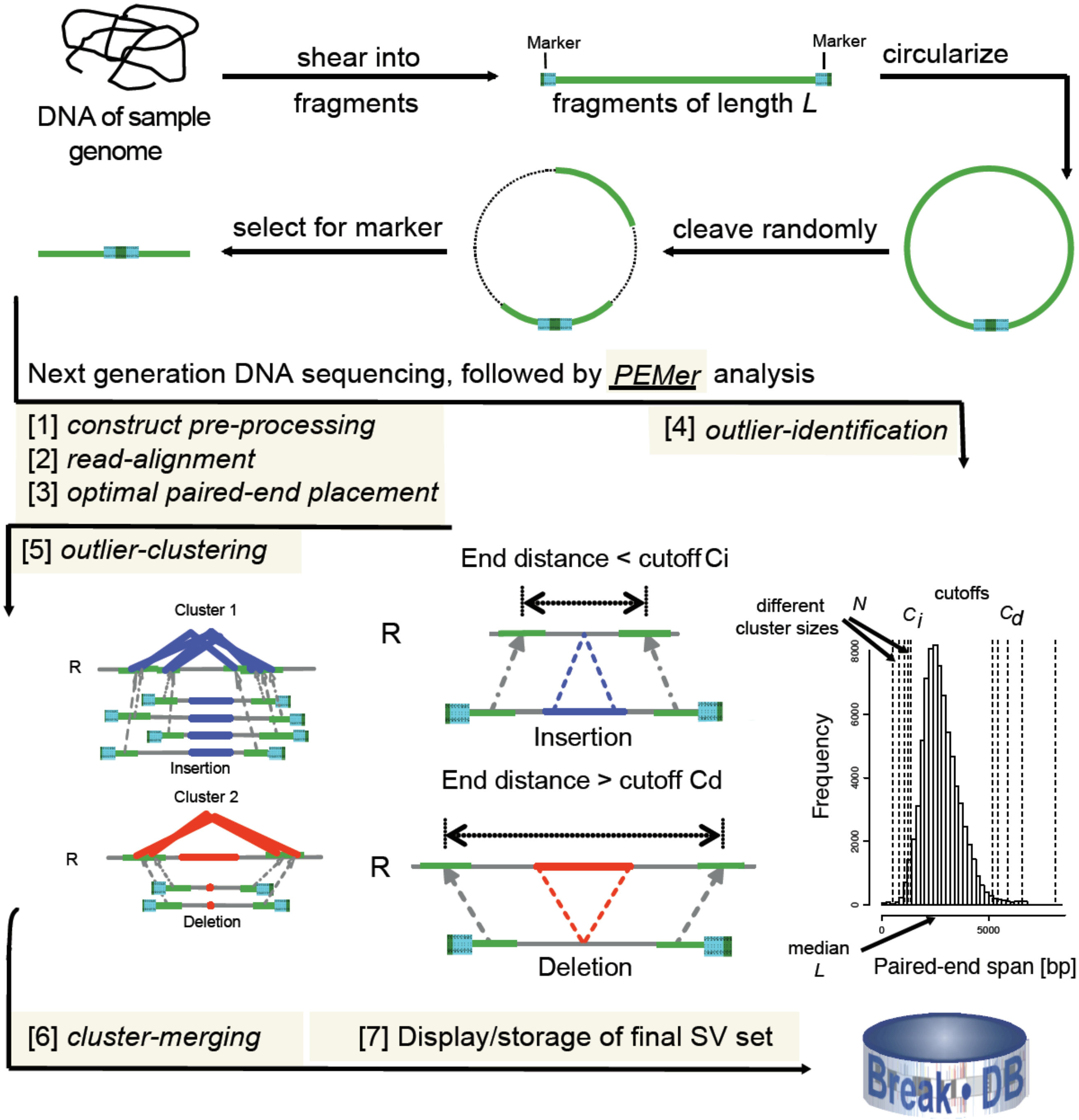
PEMer: Detecting Structural Variants from Discordant Paired Ends in Massive Sequencing





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

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



DNA of sample genome

shear into fragments

fragments of length L

circularize

cleave randomly

select for marker

Next generation DNA sequencing, followed by *PEMer* analysis

[1] *construct pre-processing*

[2] *read-alignment*

[3] *optimal paired-end placement*

[4] *outlier-identification*

[5] *outlier-clustering*

End distance < cutoff Ci

End distance > cutoff Cd

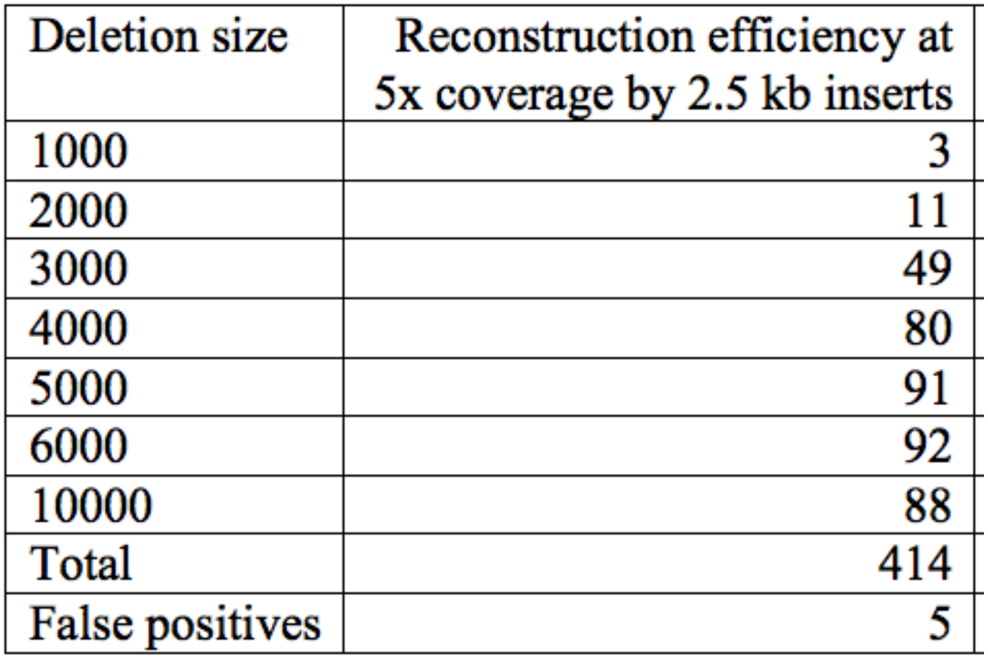
[6] *cluster-merging*

[7] Display/storage of final SV set

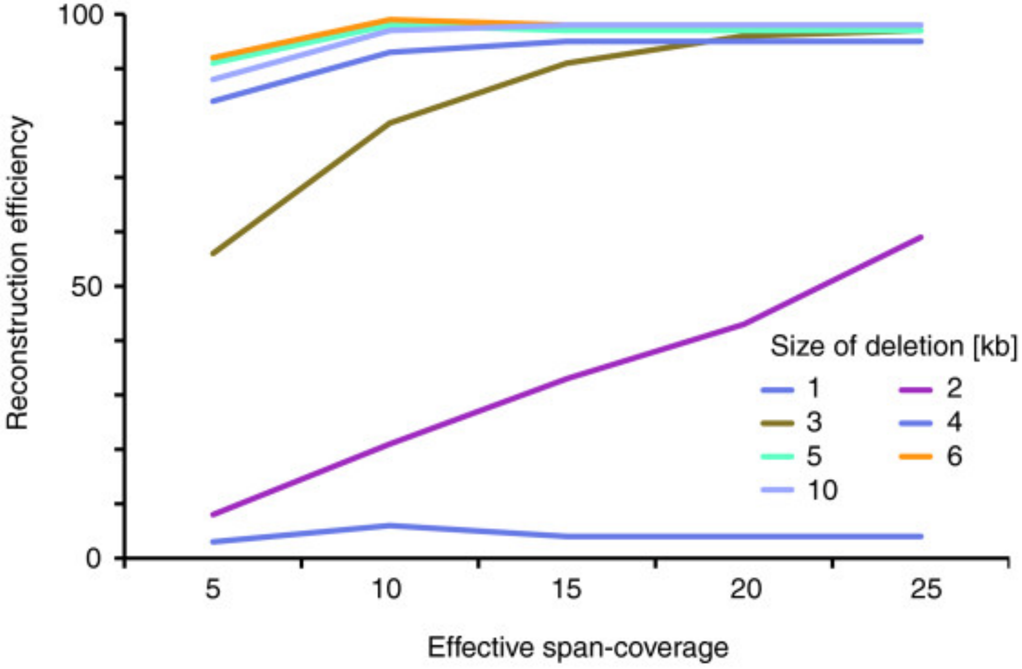
[Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]

Parameterize Error Models through Simulation

Reconstruction efficiency at different coverage

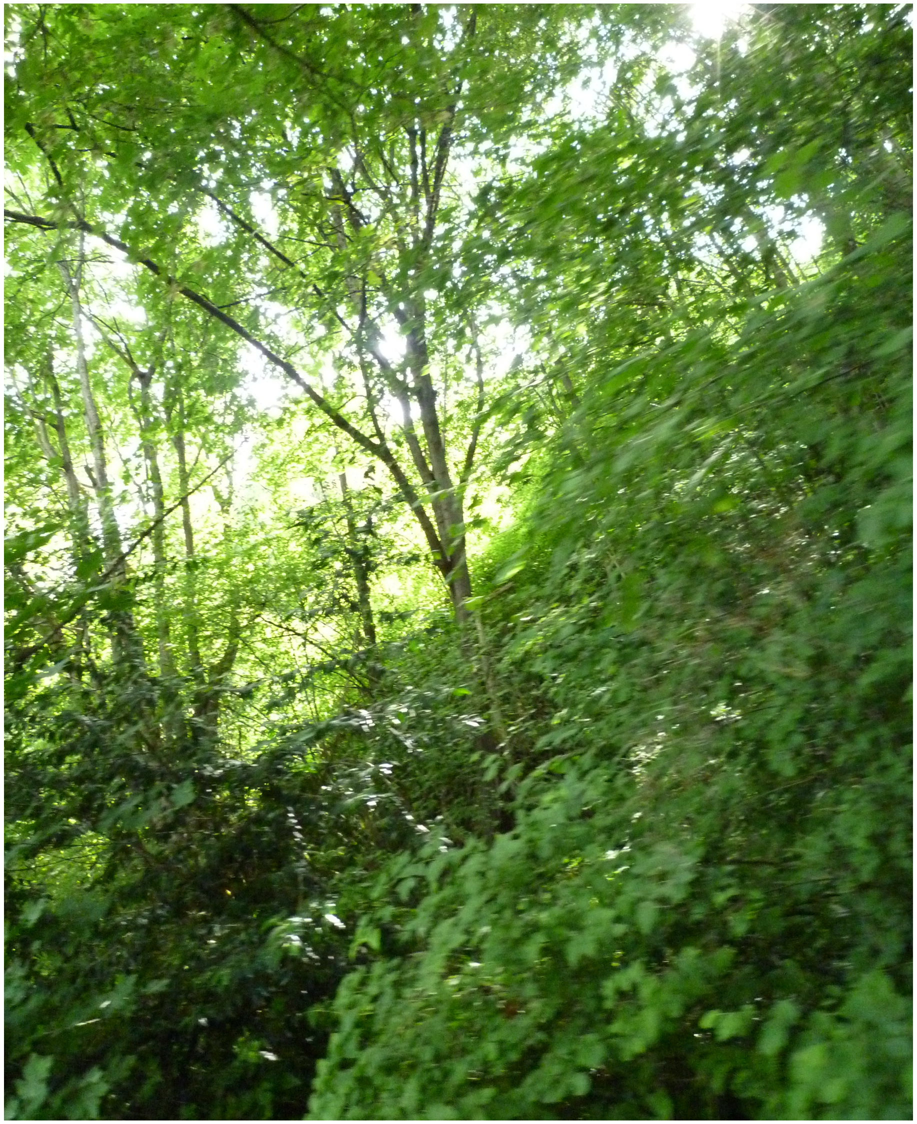


Deletion size	Reconstruction efficiency at 5x coverage by 2.5 kb inserts
1000	3
2000	11
3000	49
4000	80
5000	91
6000	92
10000	88
Total	414
False positives	5



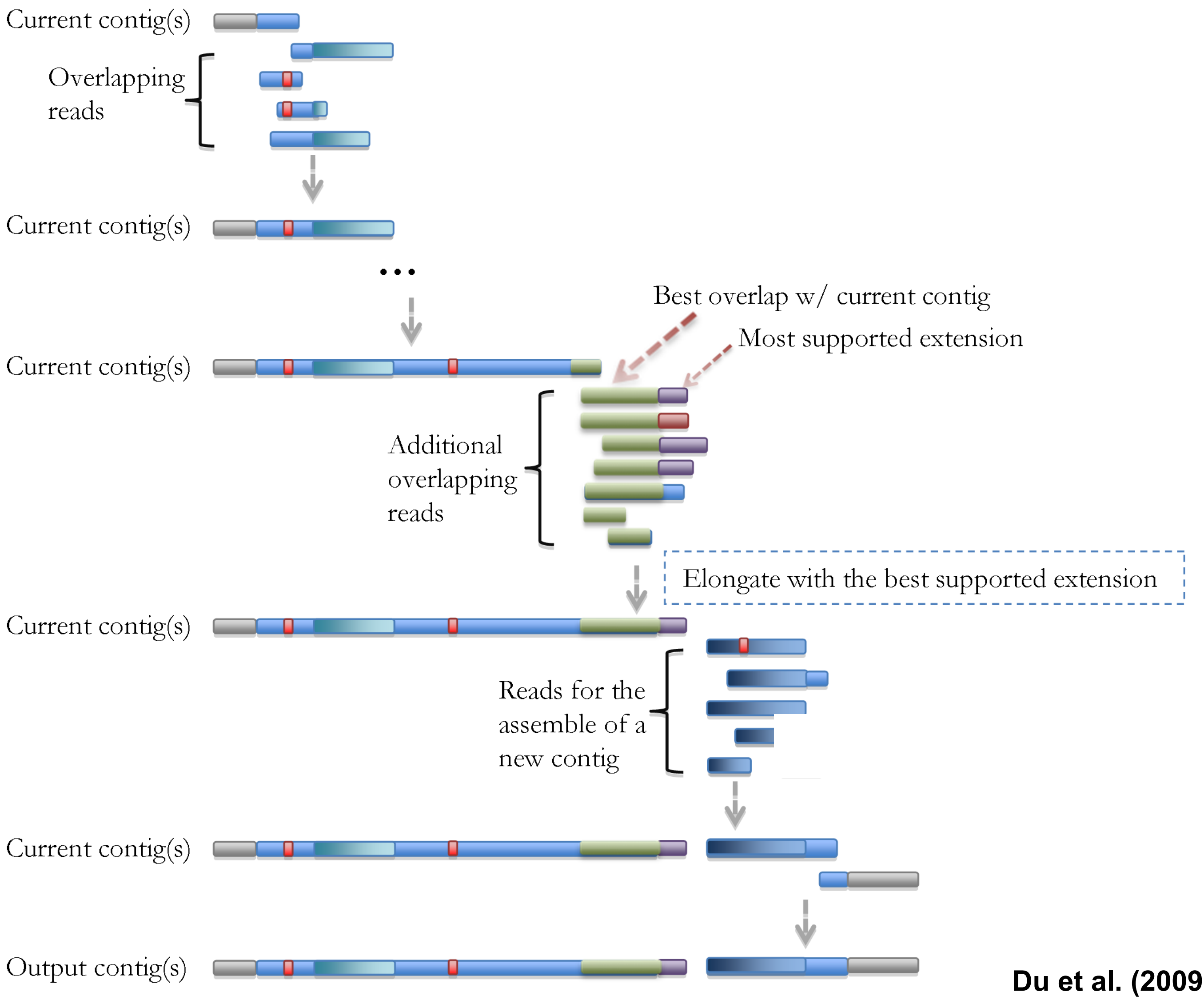
[Korbel et al., GenomeBiol. ('09)]

Local Reassembly



Simple Local Assembly: iterative contig extension

G Iterative contig elongation with the best supported extension -- a mostly greedy approach



Current contig(s)

Overlapping reads

Current contig(s)

…

Best overlap w/ current contig

Most supported extension

Current contig(s)

Additional overlapping reads

Elongate with the best supported extension

Current contig(s)

Reads for the assemble of a new contig

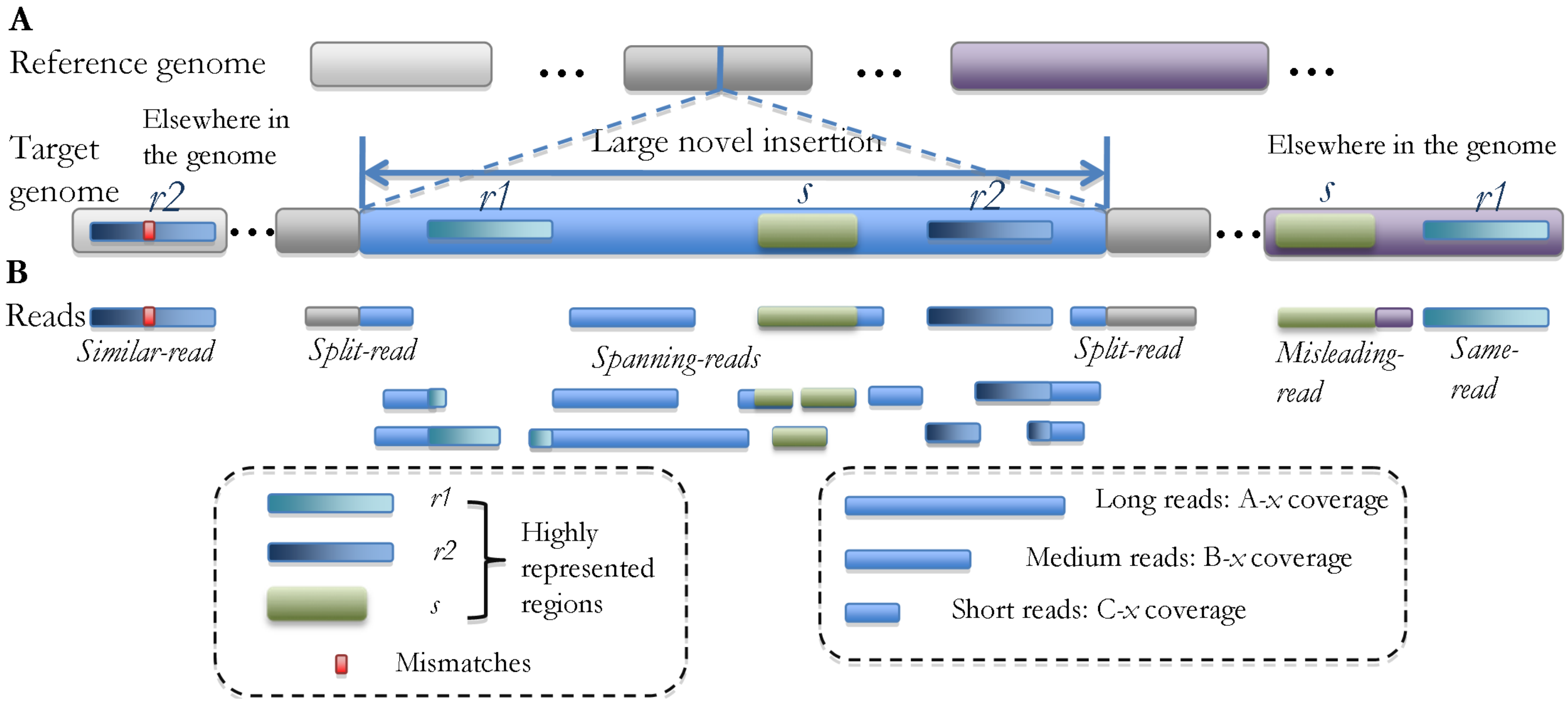
Current contig(s)

Output contig(s)

Du et al. (2009), PLoS Comp Biol.

Optimal integration of sequencing technologies:
Local Reassembly of large novel insertions

Given a fixed budget, what are the sequencing coverage A, B and C that can achieve the maximum reconstruction rate (on average/worst-case)? Maybe a few long reads can bootstrap reconstruction process.



Du et al. (2009), PLoS Comp Biol, in press

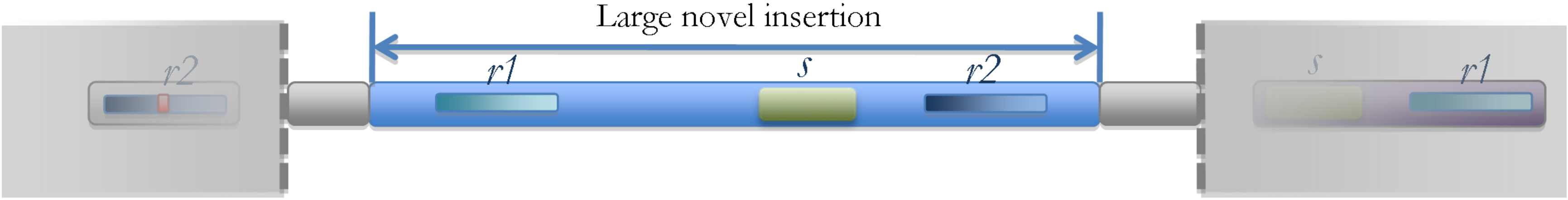
50 - Lectures.GersteinLab.org (c) '09

Optimal integration of sequencing technologies:

Need Efficient Simulation

Different combinations of technologies (i.e. read lenghs) very expensive to actually test.
Also computationally expensive to simulate.
(Each round of whole-genome assembly takes >100 CPU hrs; thus, simulation exploring 1K possibilities takes 100K CPU hr)

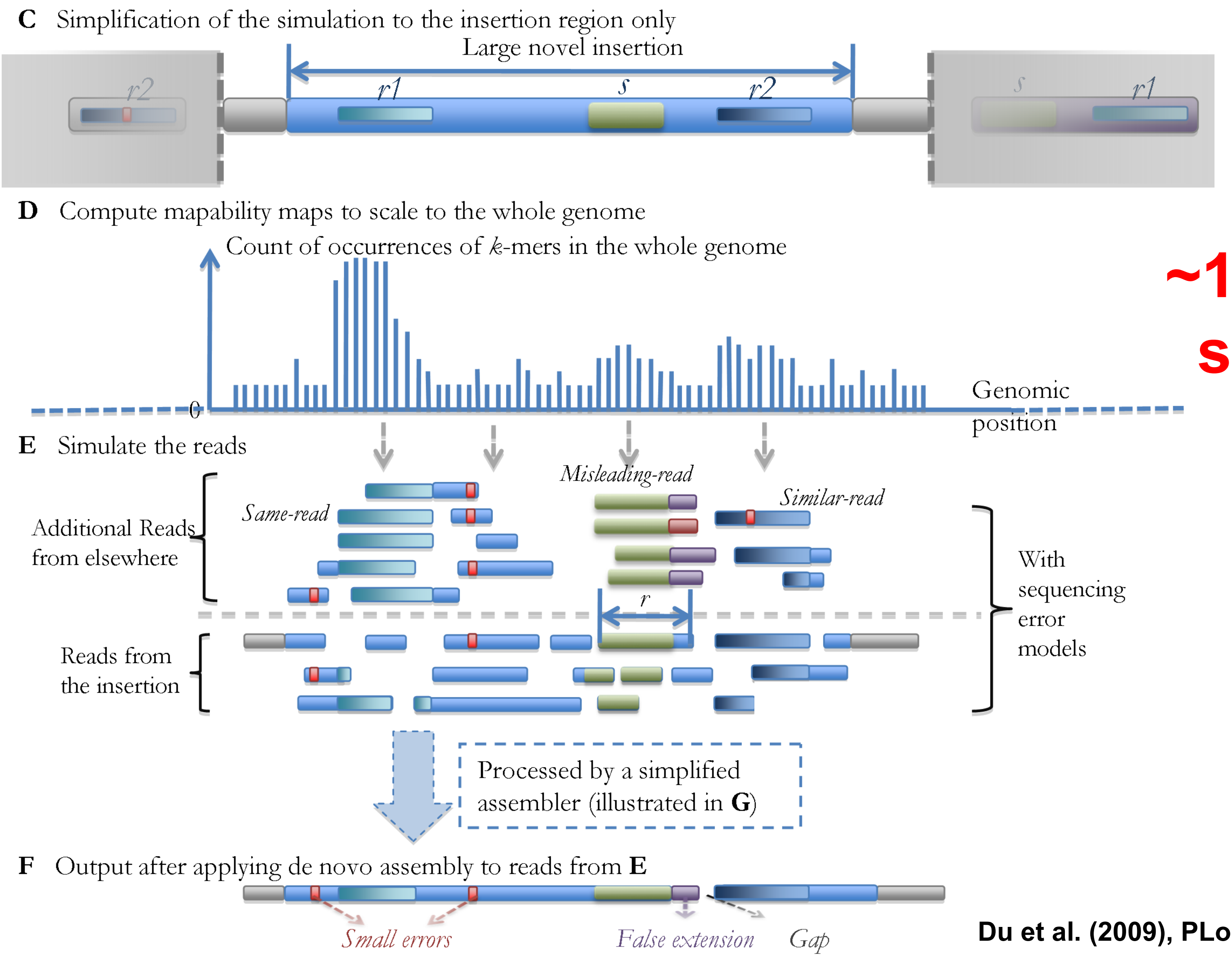
C Simplification of the simulation to the insertion region only



Du et al. (2009), PLoS Comp Biol, in press

Optimal integration of sequencing technologies:

Efficient Simulation Toolbox using Mappability Maps

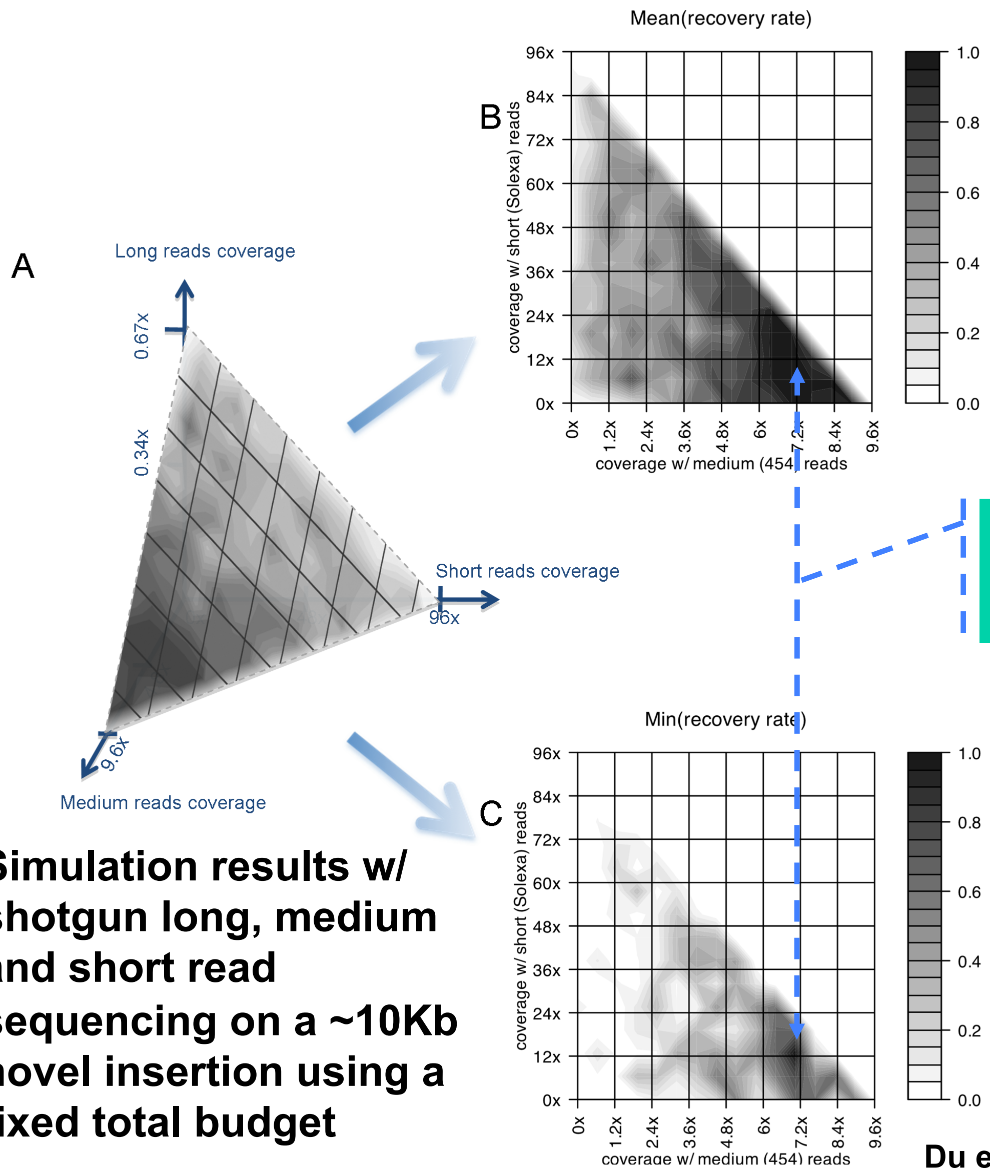

C Simplification of the simulation to the insertion region only
Large novel insertion
r2
r1
s
r2
s
r1
D Compute mapability maps to scale to the whole genome
Count of occurrences of k-mers in the whole genome
0
Genomic position
E Simulate the reads
Misleading-read
Same-read
Similar-read
Additional Reads from elsewhere
r
Reads from the insertion
With sequencing error models
Processed by a simplified assembler (illustrated in G)
F Output after applying de novo assembly to reads from E
Small errors
False extension
Gap

~100,000 X speedup

Du et al. (2009), PLoS Comp Biol, in press

Optimal integration of sequencing technologies:

Simulation shows combination better than single technology



Optimal combination of different technologies

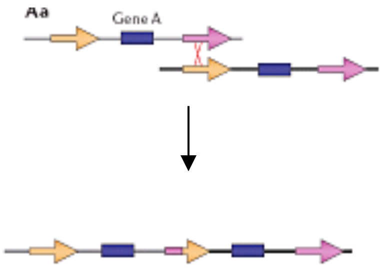
Result dependent on specific parameter setting of different sequencing technologies

Simulation results w/ shotgun long, medium and short read sequencing on a ~10Kb novel insertion using a fixed total budget

Du et al. (2009), PLOS Comp Biol, in press

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

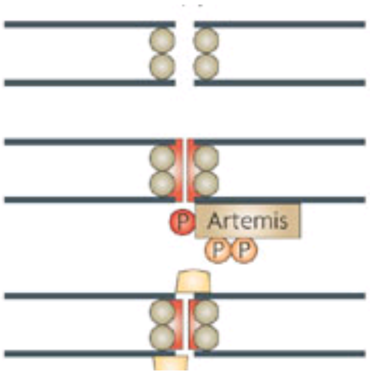


NAHR

(Non-allelic homologous recombination)

Flanking repeat

(e.g. Alu, LINE…)



NHEJ

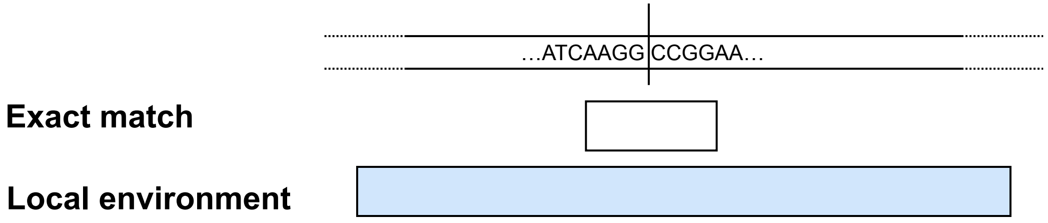
(Non-homologous-end-joining)

No (flanking) repeats.

In some cases <4bp microhomologies

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

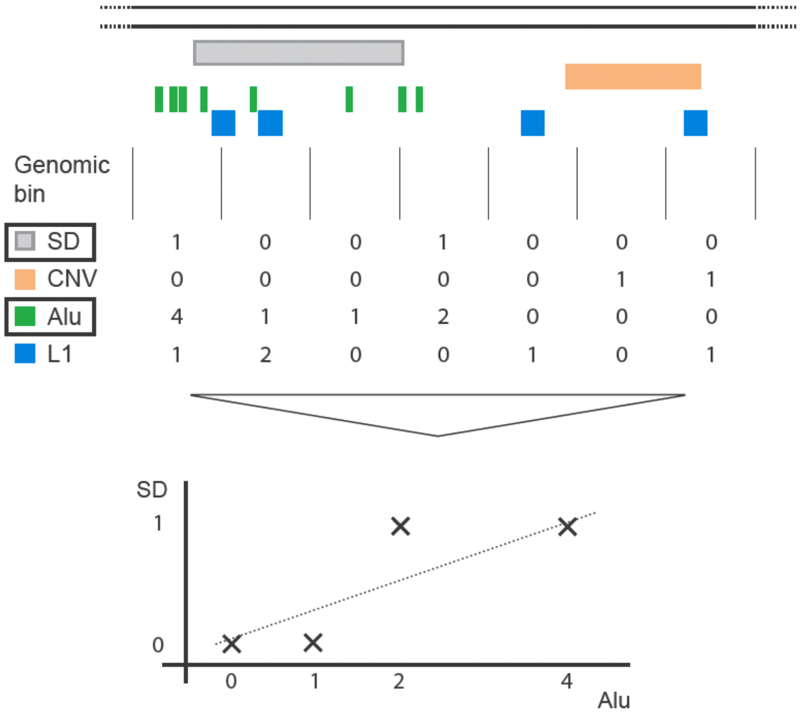
If exact CNV breakpoints are known, we can calculate the enrichment of repeat elements relative to the genome or relative to the local environment



1. Survey a range of genomic features

2. Count the number of features in each genomic bin (100kb)

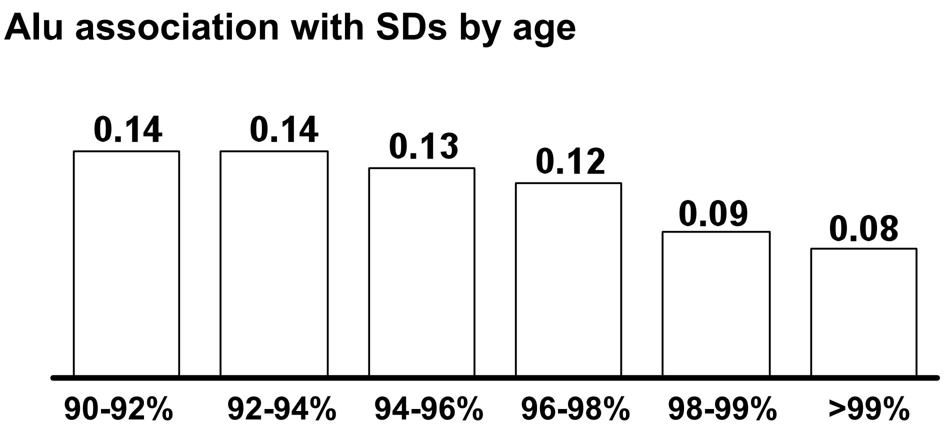
3. Calculate correlations / enrichments using robust stats



▨ SD	1	0	0	1	0	0	0
🟧 CNV	0	0	0	0	0	1	1
🟩 Alu	4	1	1	2	0	0	0
🟦 L1	1	2	0	0	1	0	1

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

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

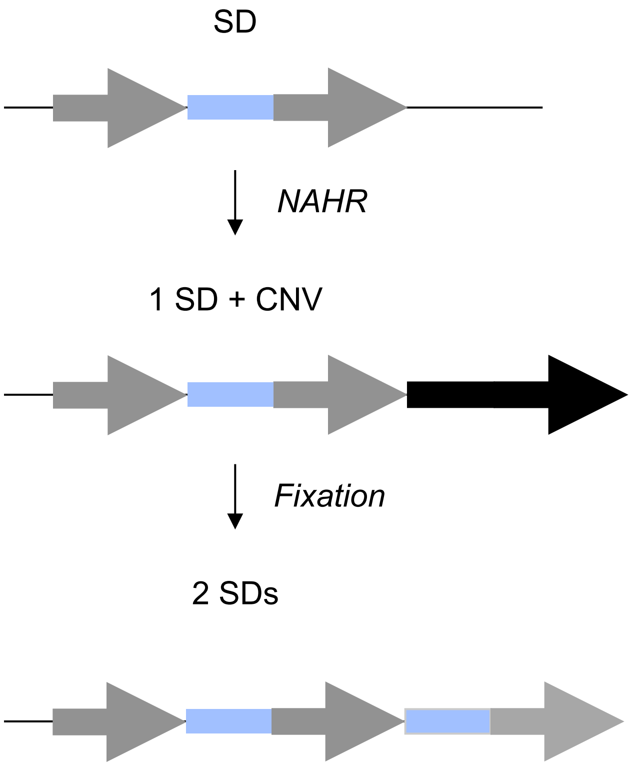


- The co-localization of Alu elements with SDs is highly significant.
- Older SDs have a much higher association with Alus than younger SDs.
- Hence it is likely, that Alu elements were more active in mediating NAHR in the past (consistent with the Alu burst)

[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

NAHR

1 SD + CNV

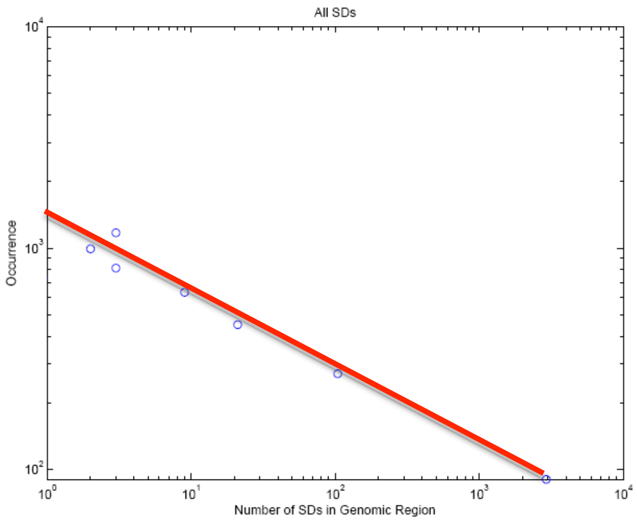
Fixation

2 SDs

“SD selfpropagation”

Corollary

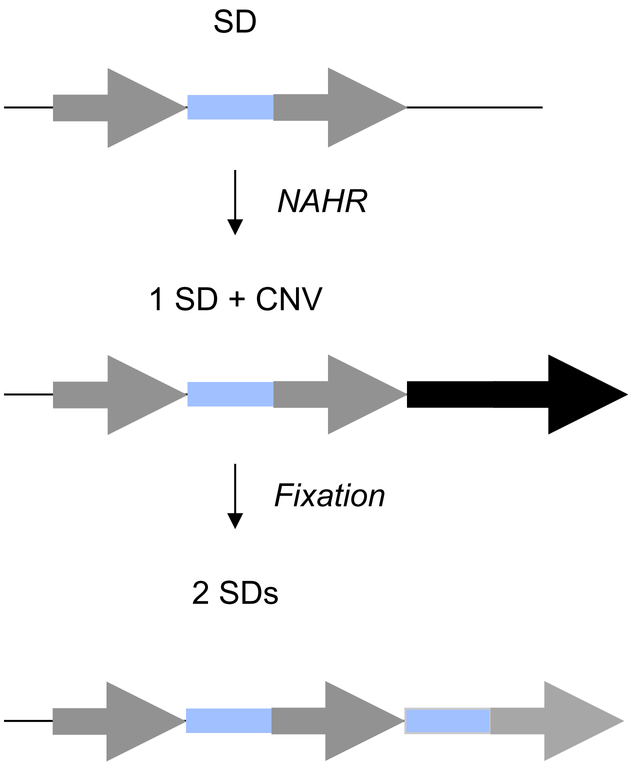
- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs
- Such mechanisms (“preferential attachment”) are well studied in physics and should leads a very skewed (“power-law”) distribution of SDs.



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

FOCUSSING ON SDS: SDs COLOCALIZE WITH EACH OTHER

Hypothesis



SD

NAHR

1 SD + CNV

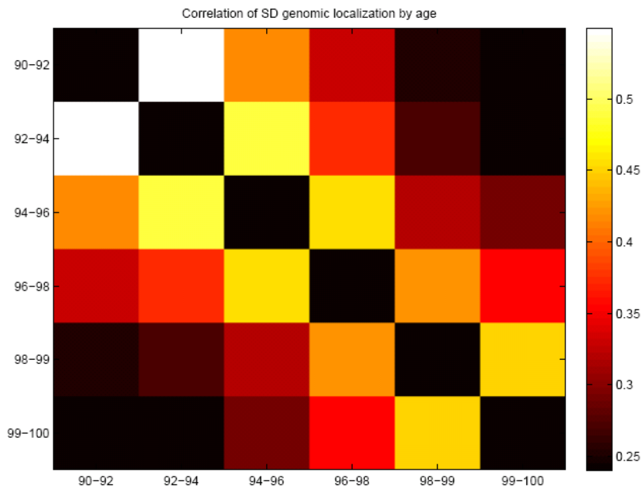
Fixation

2 SDs

“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:



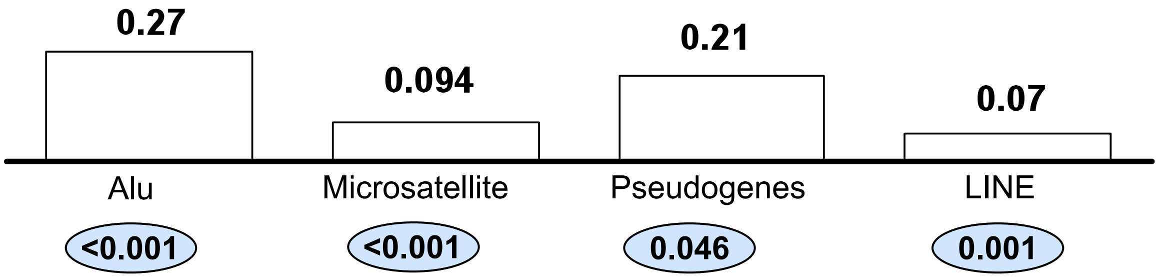
[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

080907_SD_CNV_Slides_MBG_CEGS_PMK

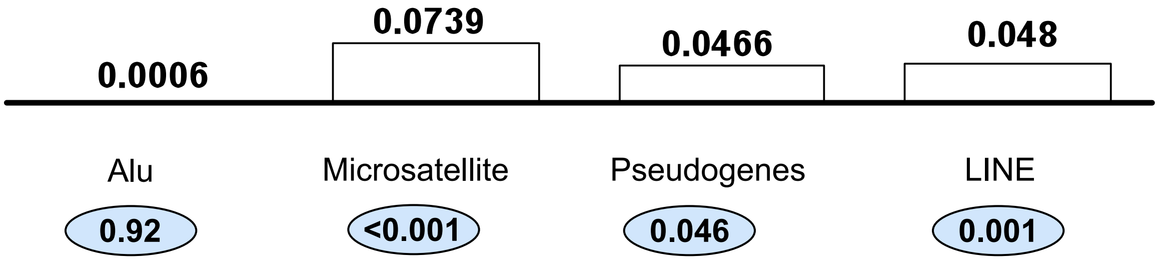
ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs

CNVs ARE LESS ASSOCIATED WITH SDs THAN THE GENERAL SD TREND

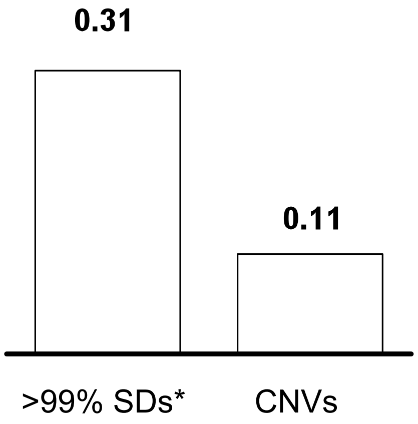
SD association with repeats



CNV association with repeats

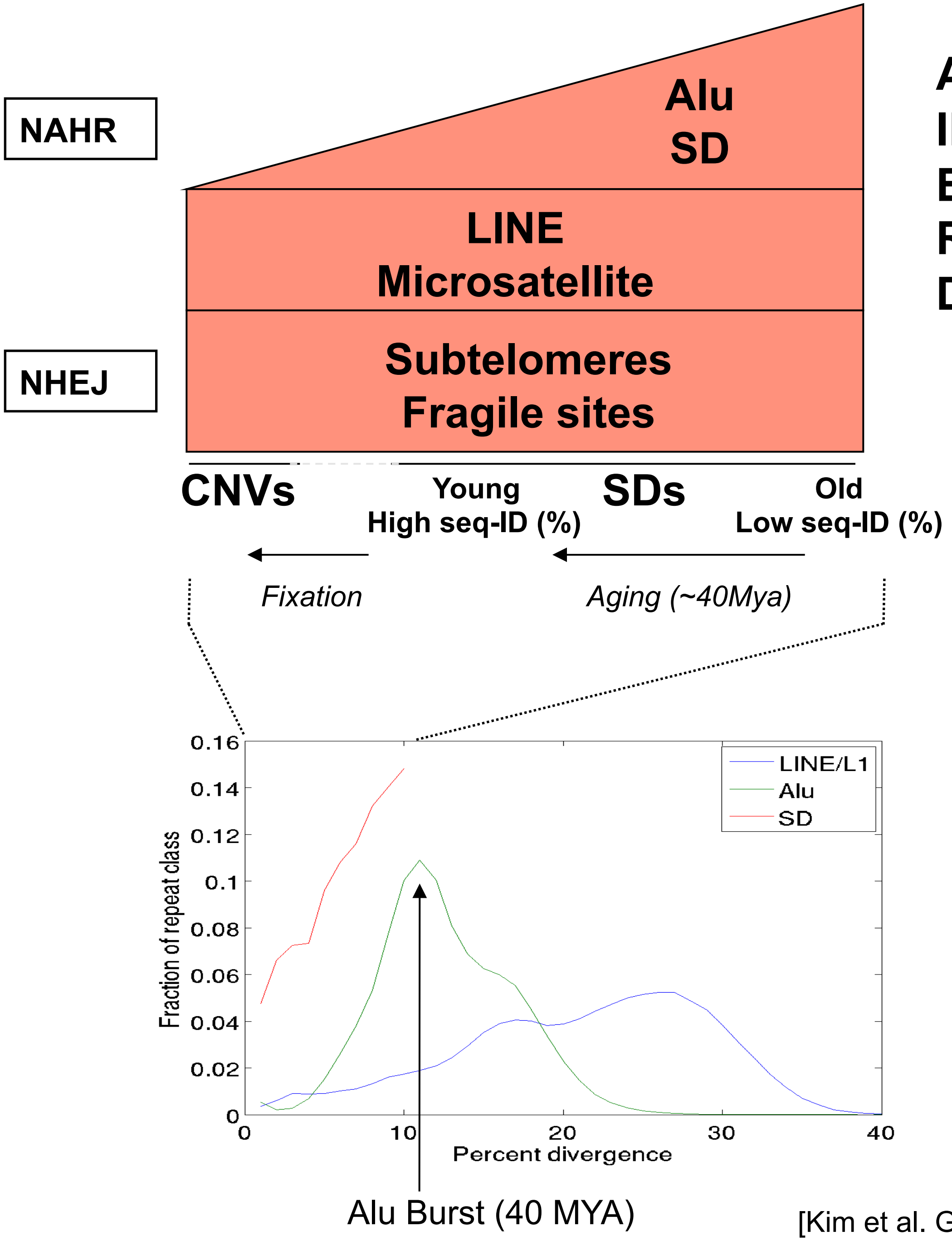


CNV Association with SDs



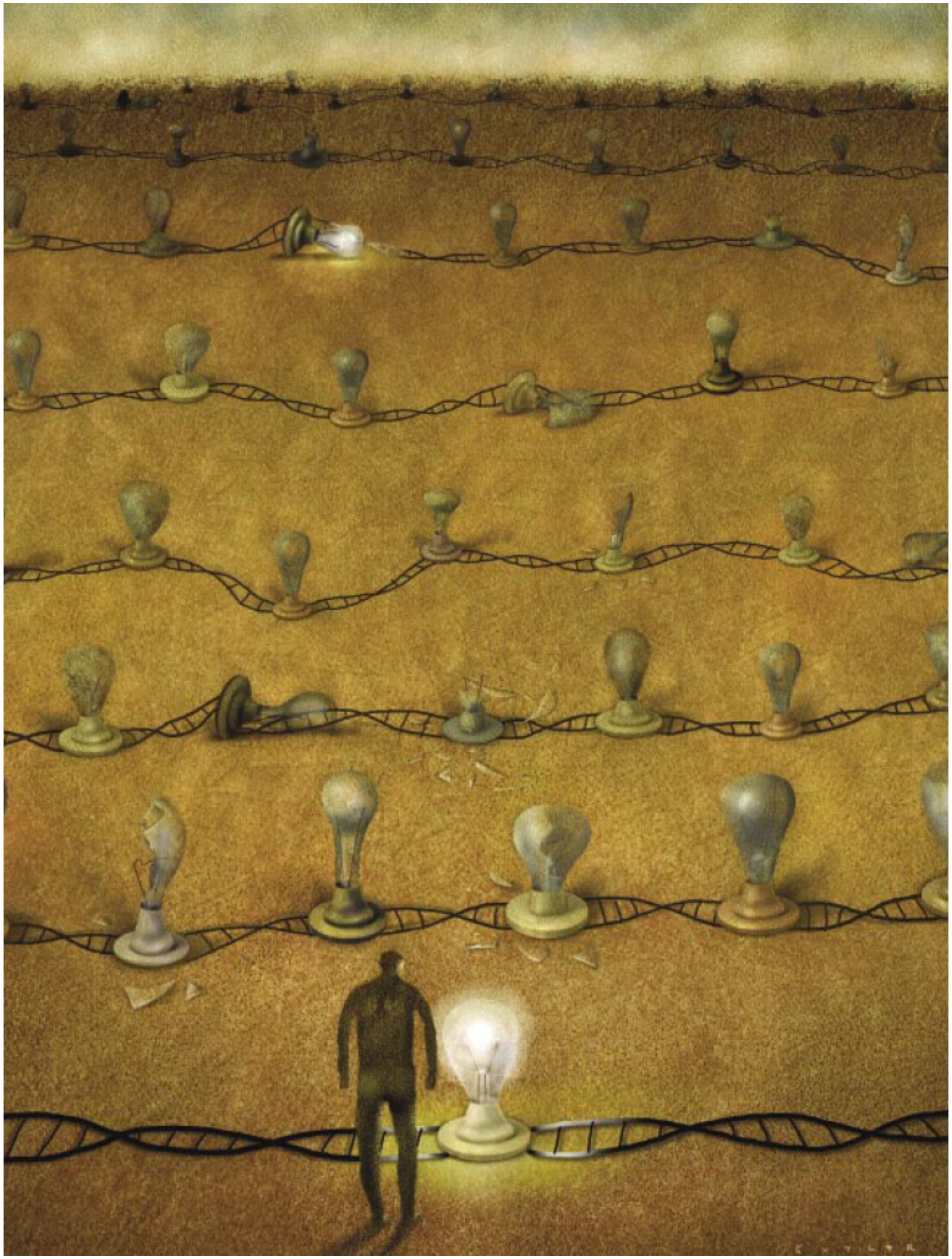
[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]

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]



Formal Annotation based on Comparative Genomics: Pseudogenes

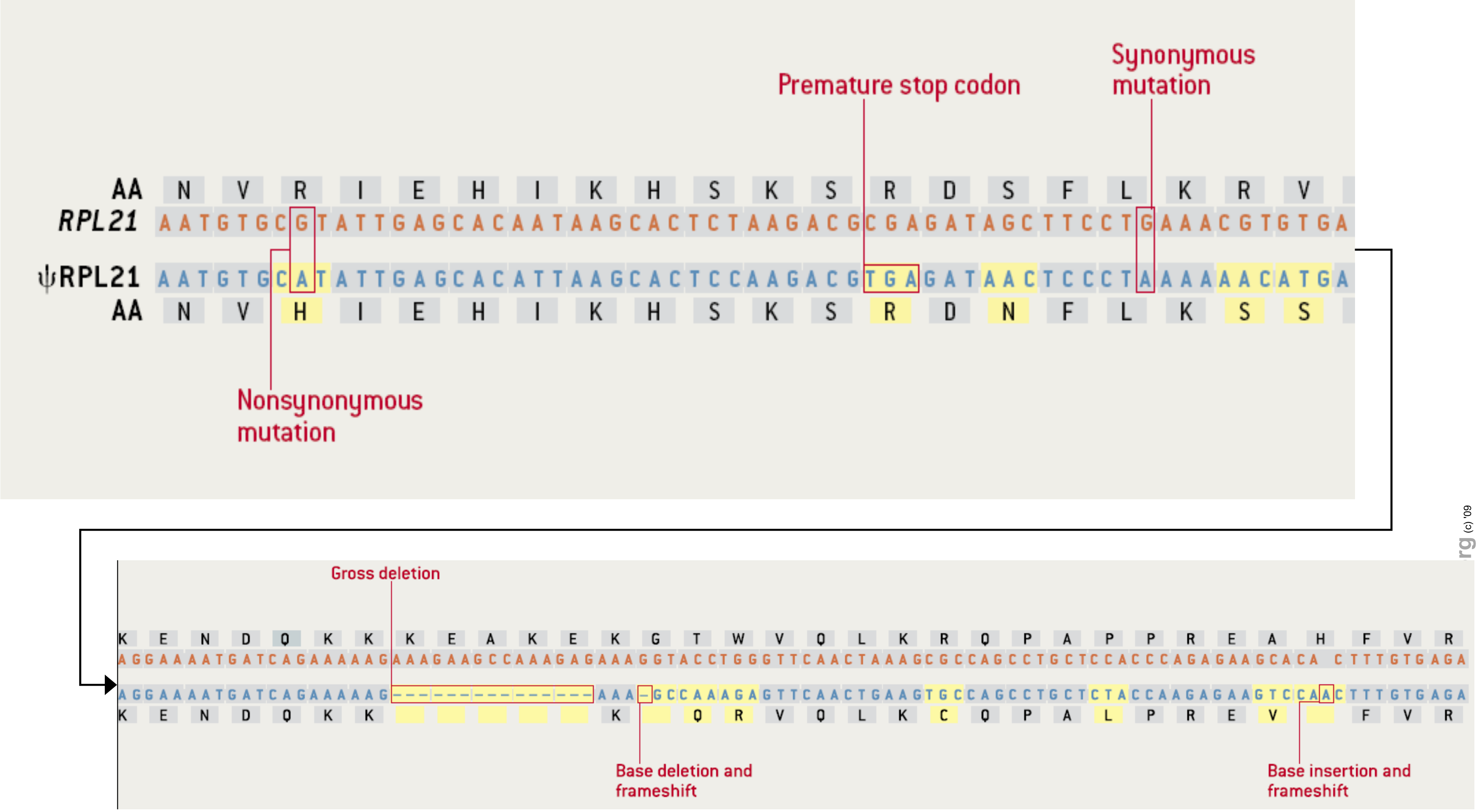
Illustration from Gerstein & Zheng (2006). Sci Am.

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]

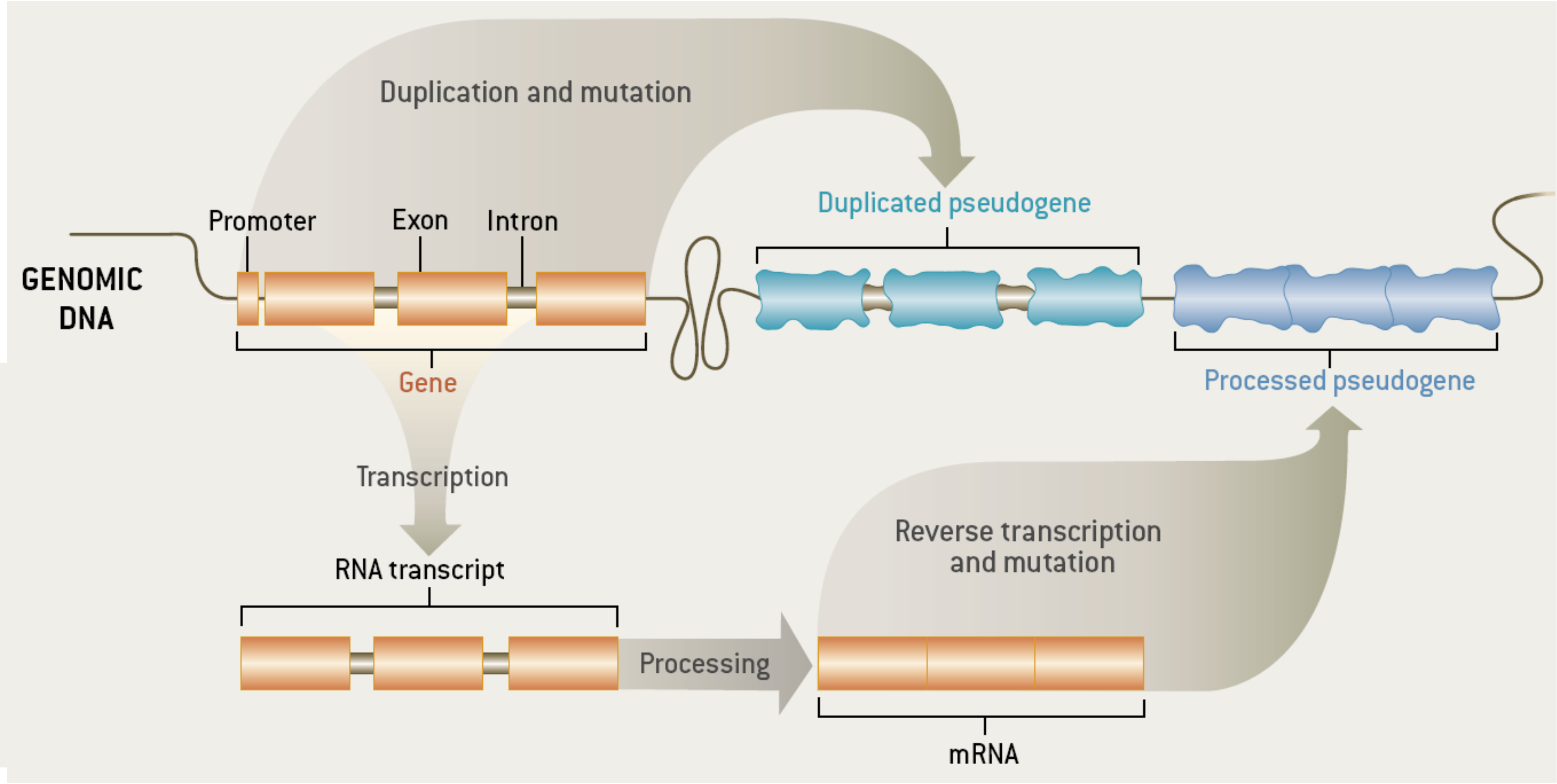
Identifiable Features of a Pseudogene (ψRPL21)


Premature stop codon
Synonymous mutation
AA N V R I E H I K H S K S R D S F L K R V
RPL21 AATGTGCGTATTGAGCACAATAAGCACTCTAAGACGCGAGATAGCTTCCTGAAACGTGTGA
ψRPL21 AATGTGCATATTGAGCACATTAAGCACTCCAAGACGTGAGATAACTCCCTAAAAAACATGA
AA N V H I E H I K H S K S R D N F L K S S
Nonsynonymous mutation
Gross deletion
K E N D Q K K K E A K E K G T W V Q L K R Q P A P P R E A H F V R
AGGAAAATGATCAGAAAAAGAAAGAAGCCAAAGAGAAAGGTACCTGGGTTCAACTAAAGCGCCAGCCTGCTCCACCCAGAGAAGCACA CTTTGTGAGA
AGGAAAATGATCAGAAAAAG---------------AAA-GCCAAAGAGTTCAACTGAAGTGCCAGCCTGCTCTACCAAGAGAAGTCCAACTTTGTGAGA
K E N D Q K K K Q R V Q L K C Q P A L P R E V F V R
Base deletion and frameshift
Base insertion and frameshift

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

rg (c) '09

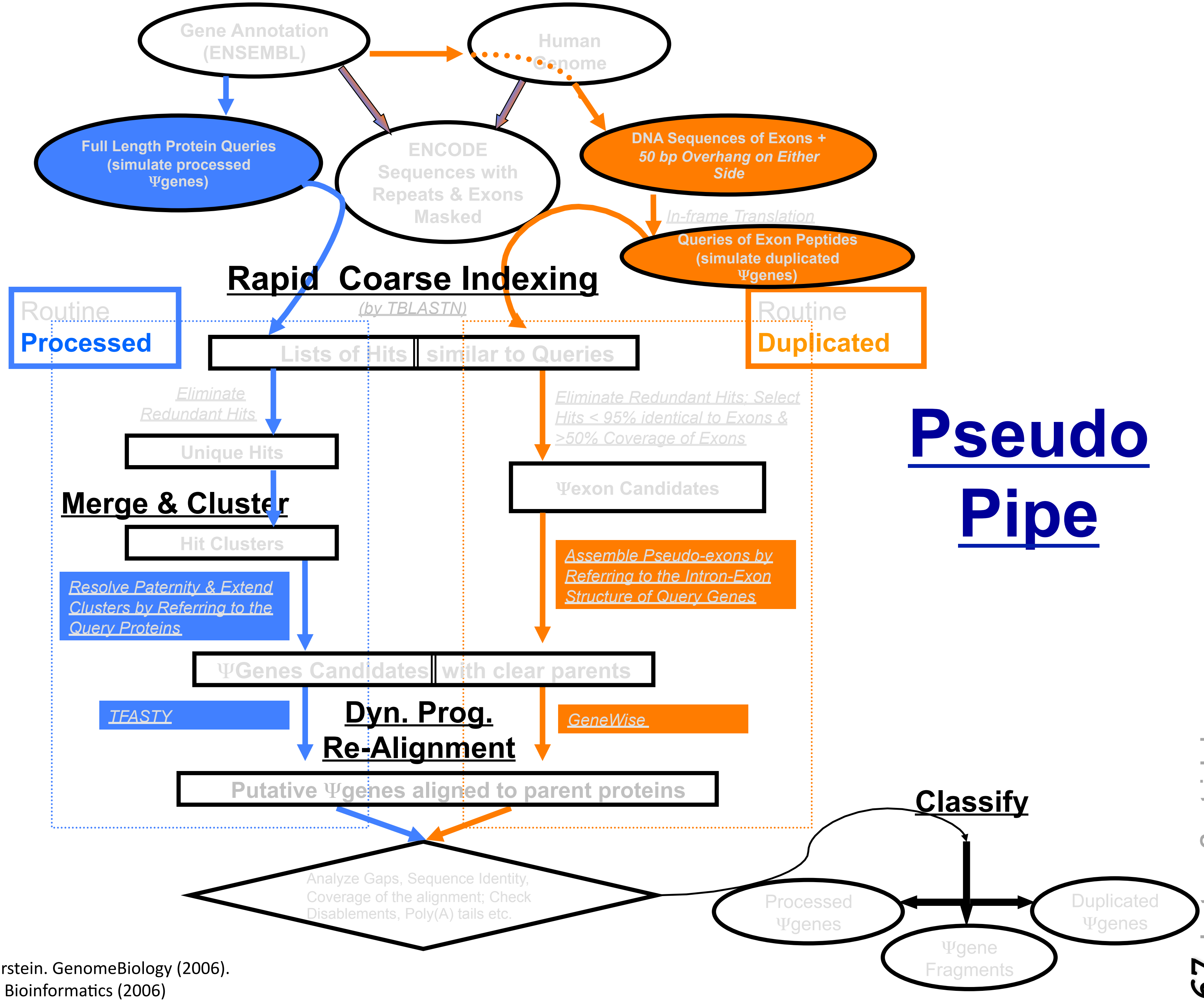
Two Major Genomic Remodeling Processes Give Rise to Distinct Types of Pseudogenes



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

Pseudogene Tools: Assignment Pipeline & DB

Pseudo Pipe



Gene Annotation (ENSEMBL)

Human Genome

Full Length Protein Queries (simulate processed Ψgenes)

ENCODE Sequences with Repeats & Exons Masked

DNA Sequences of Exons + *50 bp Overhang on Either Side*

In-frame Translation

Queries of Exon Peptides (simulate duplicated Ψgenes)

Rapid Coarse Indexing

(by TBLASTN)

Routine **Processed**

Routine **Duplicated**

Lists of Hits similar to Queries

Eliminate Redundant Hits

Unique Hits

Merge & Cluster

Hit Clusters

Resolve Paternity & Extend Clusters by Referring to the Query Proteins

Eliminate Redundant Hits; Select Hits < 95% identical to Exons & >50% Coverage of Exons

Ψexon Candidates

Assemble Pseudo-exons by Referring to the Intron-Exon Structure of Query Genes

ΨGenes Candidates with clear parents

TFASTY

Dyn. Prog. Re-Alignment

GeneWise

Putative Ψgenes aligned to parent proteins

Analyze Gaps, Sequence Identity, Coverage of the alignment; Check Disablements, Poly(A) tails etc.

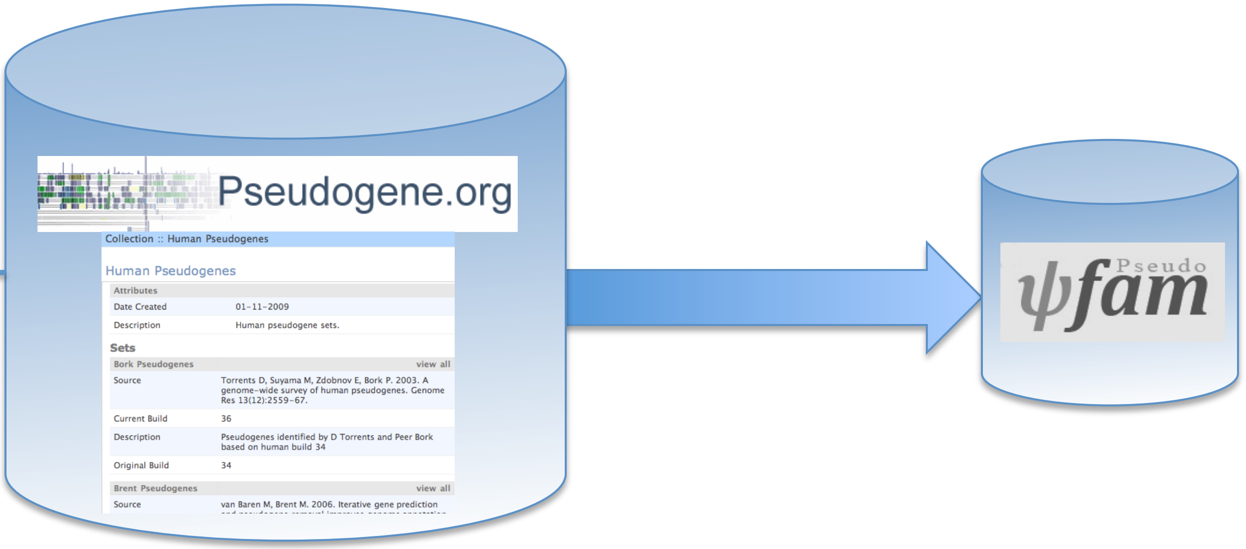
Classify

Processed Ψgenes

Duplicated Ψgenes

Ψgene Fragments

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



Collection :: Human Pseudogenes

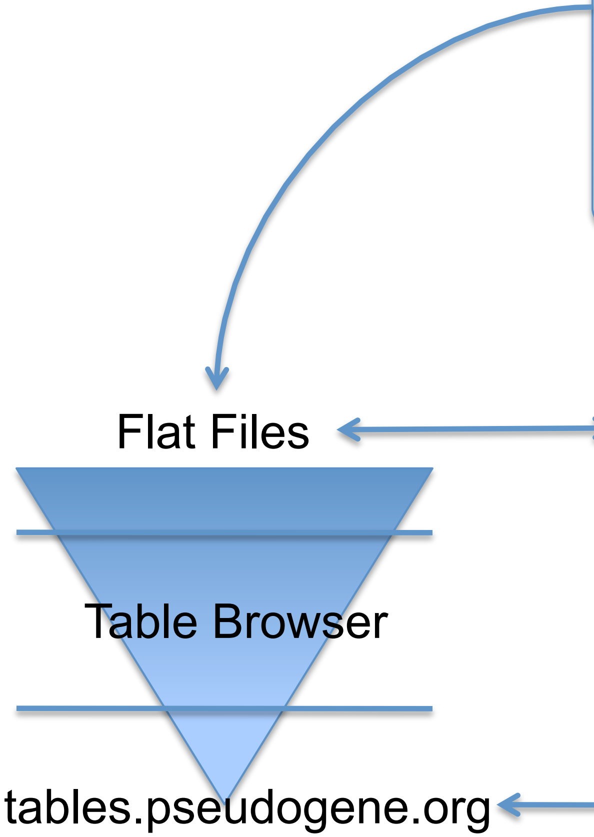
Human Pseudogenes

Attributes	
Date Created	01-11-2009
Description	Human pseudogene sets.

Sets

Bork Pseudogenes	view all
Source	Torrents D, Suyama M, Zdobnov E, Bork P. 2003. A genome-wide survey of human pseudogenes. Genome Res 13(12):2559-67.
Current Build	36
Description	Pseudogenes identified by D Torrents and Peer Bork based on human build 34
Original Build	34

Brent Pseudogenes	view all
Source	van Baren M, Brent M. 2006. Iterative gene prediction



Flat Files ↔ DAS

Table Browser

tables.pseudogene.org ↔ UCSC Genome Browser

- 12 eukaryotic species
  - Human, mouse, rat, chimp…
  - 100,052 pseudogenes

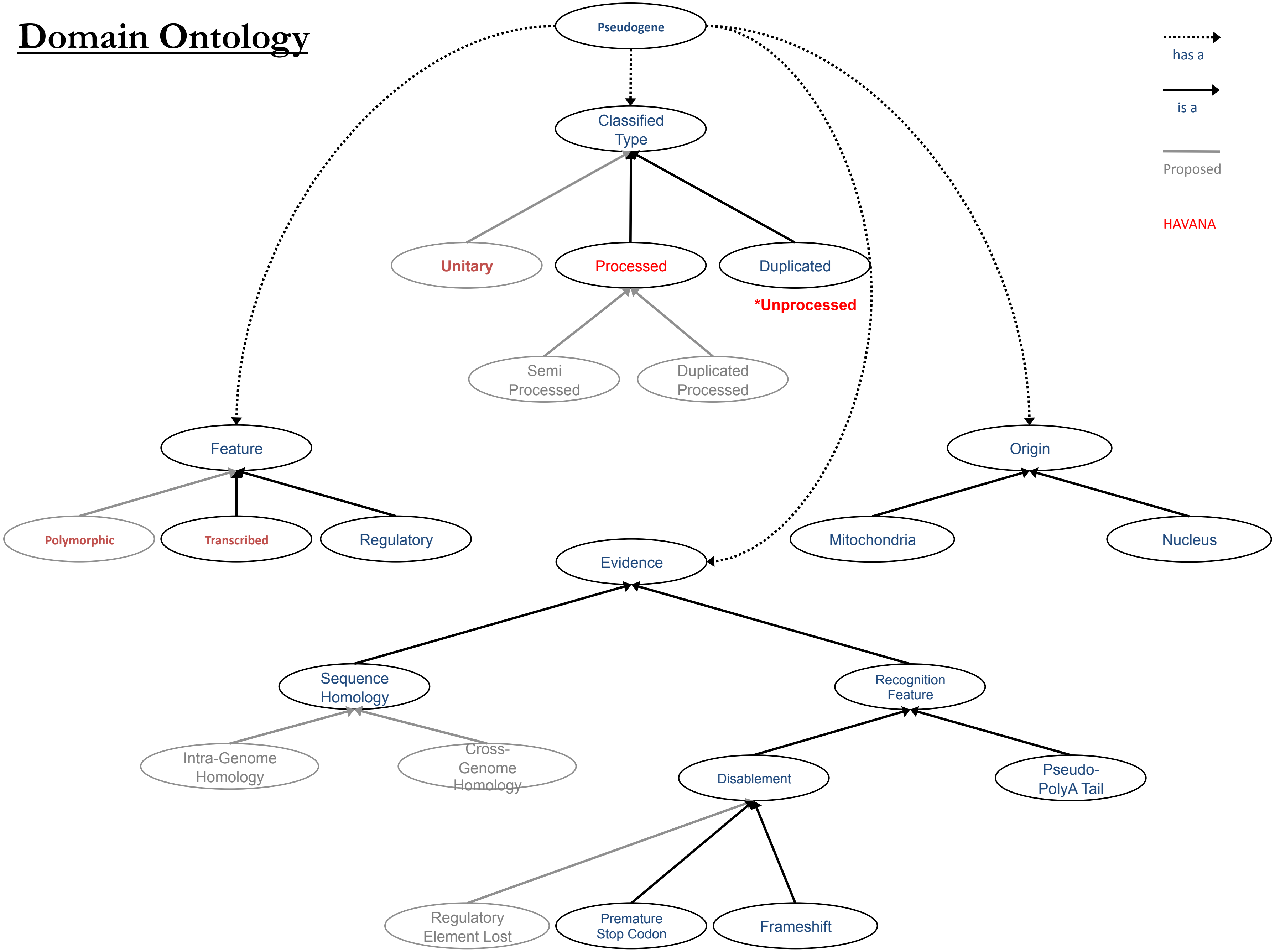
- 64 prokaryotic species
  - 6,412 pseudogenes

28,237 human pseudogenes total ~23K in recent pipeline run

- 13+ unique human sets

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

Domain Ontology



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

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

- Chronology of Sets
  1. Encode Pilot 1%
  2. Unitary pseudogenes (Hard)
  3. Ribosomal Protein pseudogenes
  4. **Glycolytic Pseudogenes**
  5. ....
- Totals (May '09)

◊ Automatic pipeline currently gives ~23K

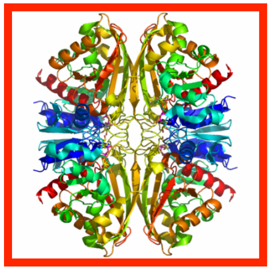
◊ Manually Annotated ~8K

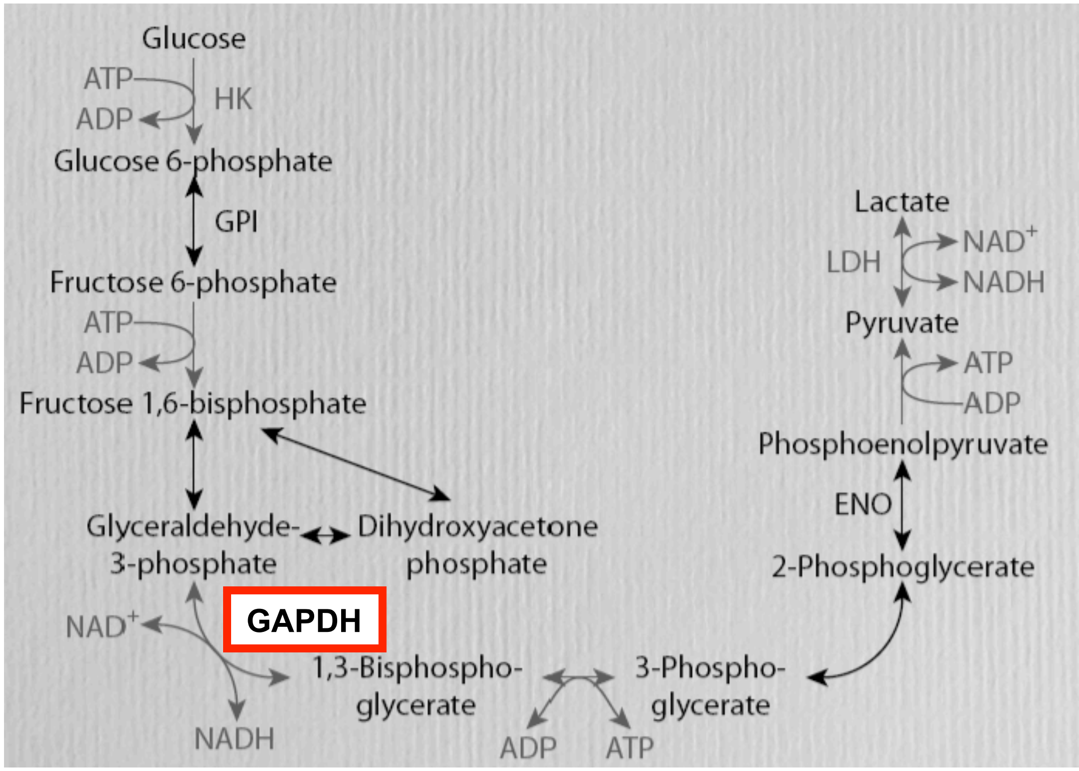
Specific Pseudogene Assignments: Glycolytic Pseudogenes

Number of pseudogenes for each glycolytic enzyme

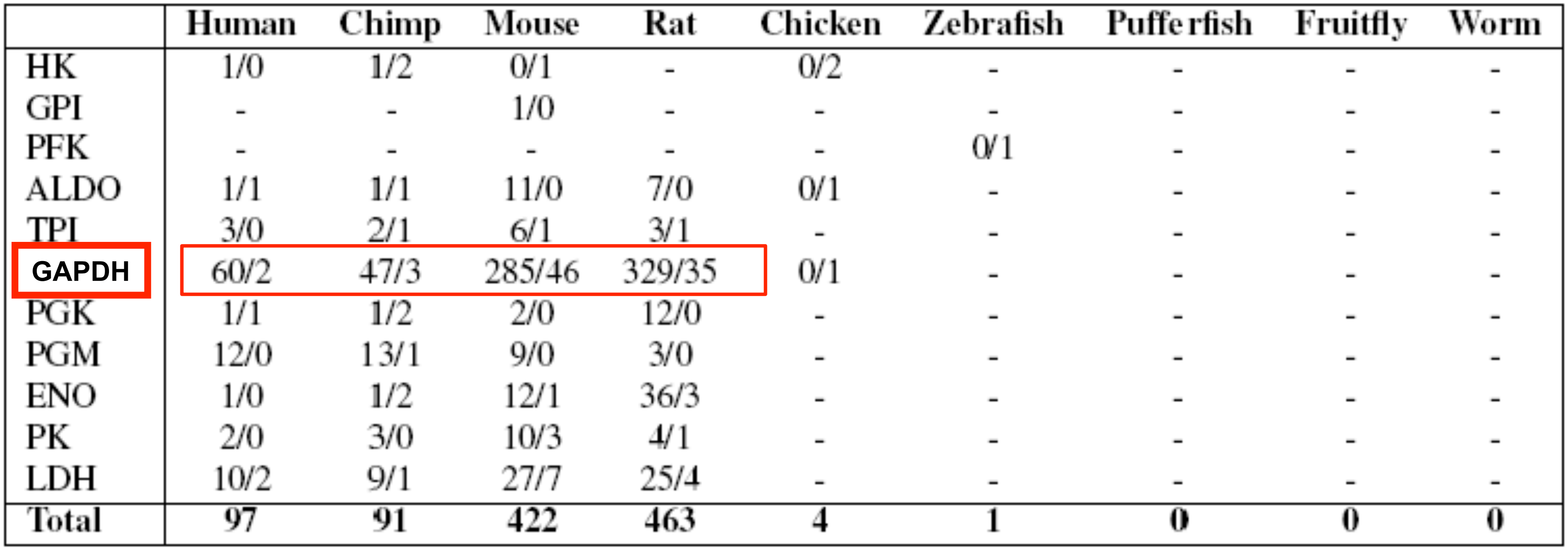
[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.





Processed/Duplicated

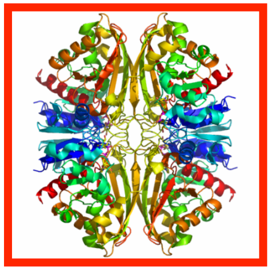


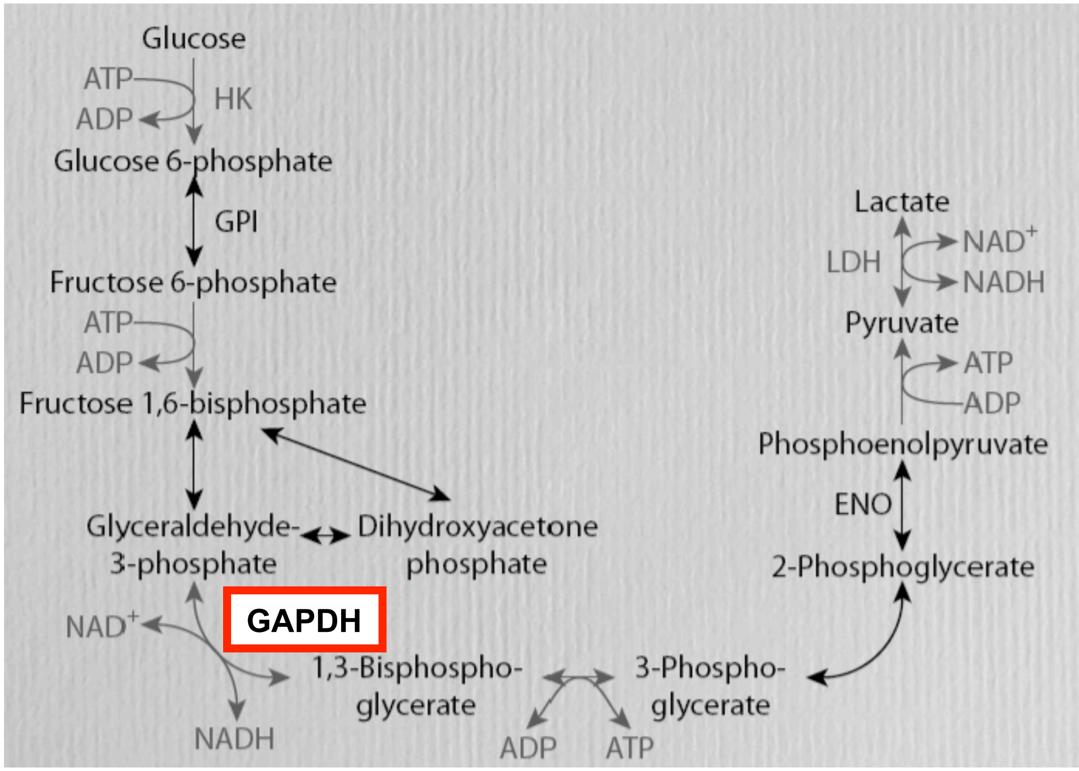
	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60/2	47/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0	-	-	-	-	-
ENO	1/0	1/2	12/1	36/3	-	-	-	-	-
PK	2/0	3/0	10/3	4/1	-	-	-	-	-
LDH	10/2	9/1	27/7	25/4	-	-	-	-	-
Total	**97**	**91**	**422**	**463**	**4**	**1**	**0**	**0**	**0**

Number of pseudogenes for each glycolytic enzyme

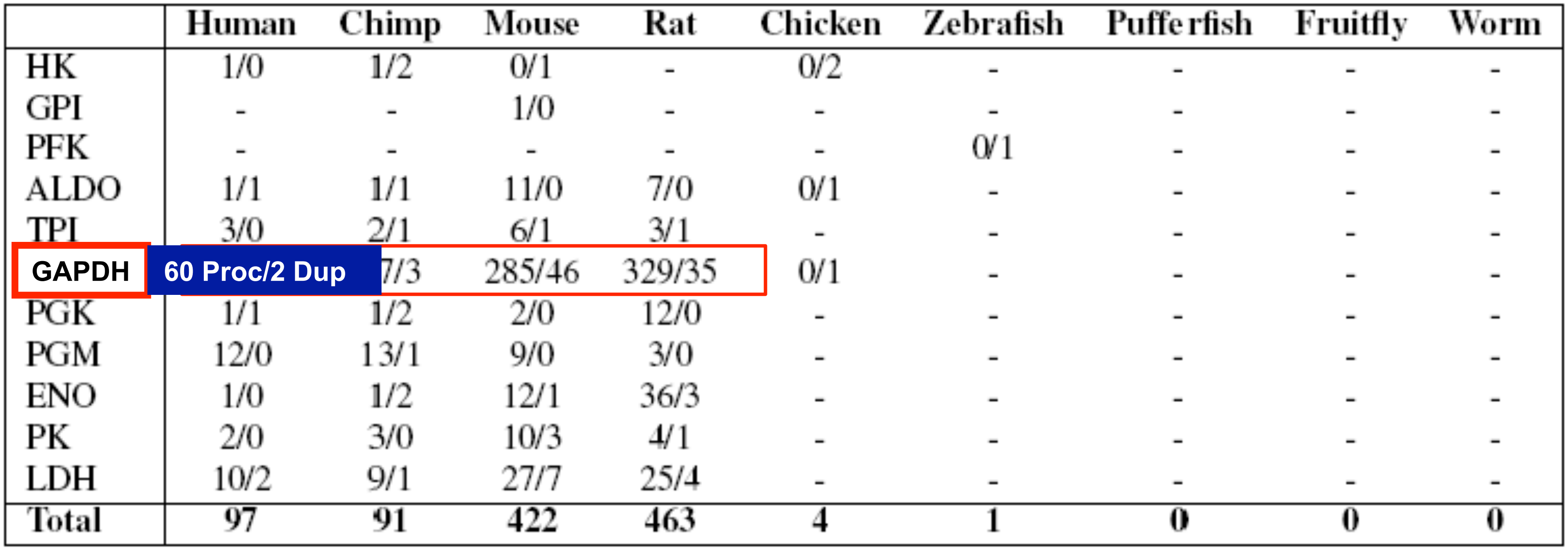
[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.





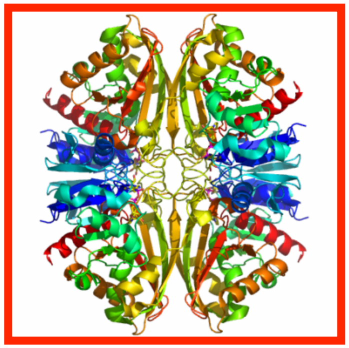
Processed/Duplicated



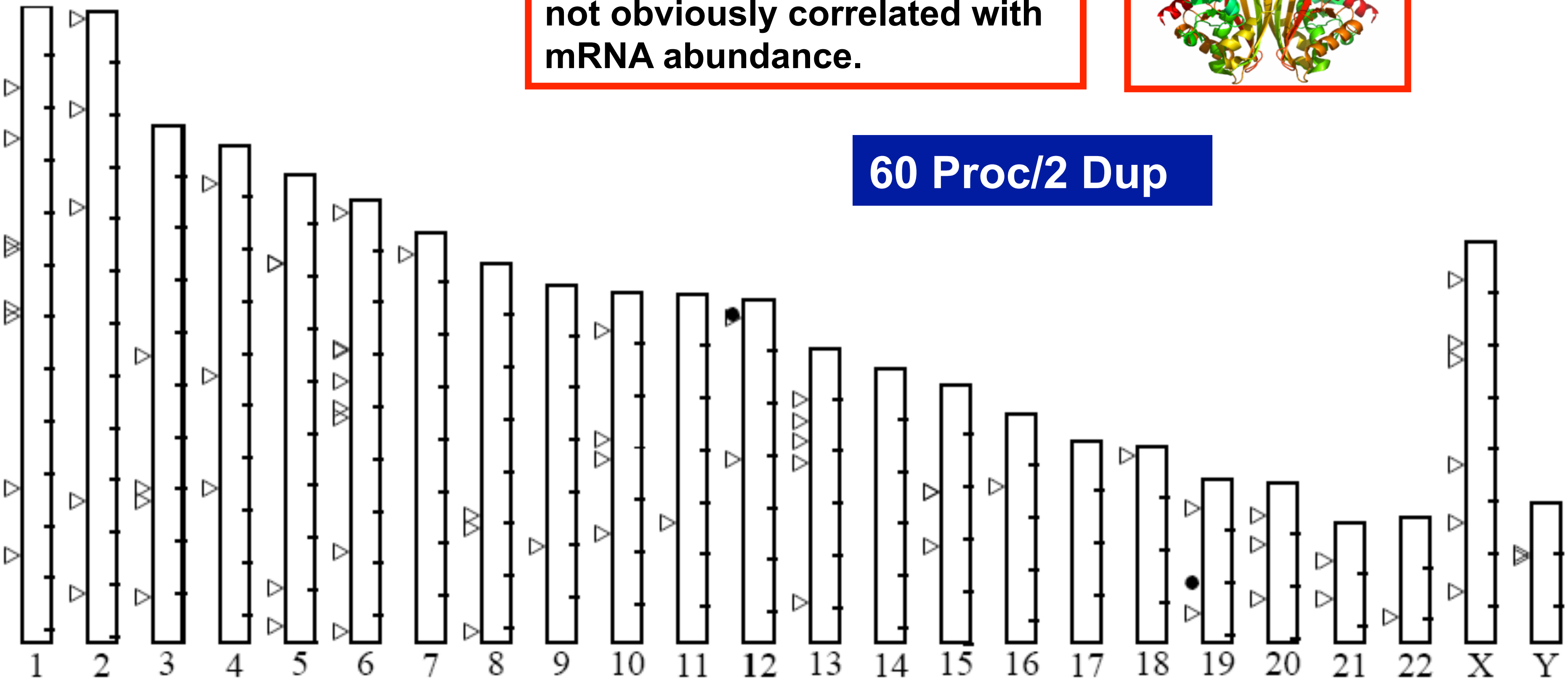
	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60 Proc/2 Dup	7/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0	-	-	-	-	-
ENO	1/0	1/2	12/1	36/3	-	-	-	-	-
PK	2/0	3/0	10/3	4/1	-	-	-	-	-
LDH	10/2	9/1	27/7	25/4	-	-	-	-	-
Total	**97**	**91**	**422**	**463**	**4**	**1**	**0**	**0**	**0**

Distribution of human GAPDH pseudogenes

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.

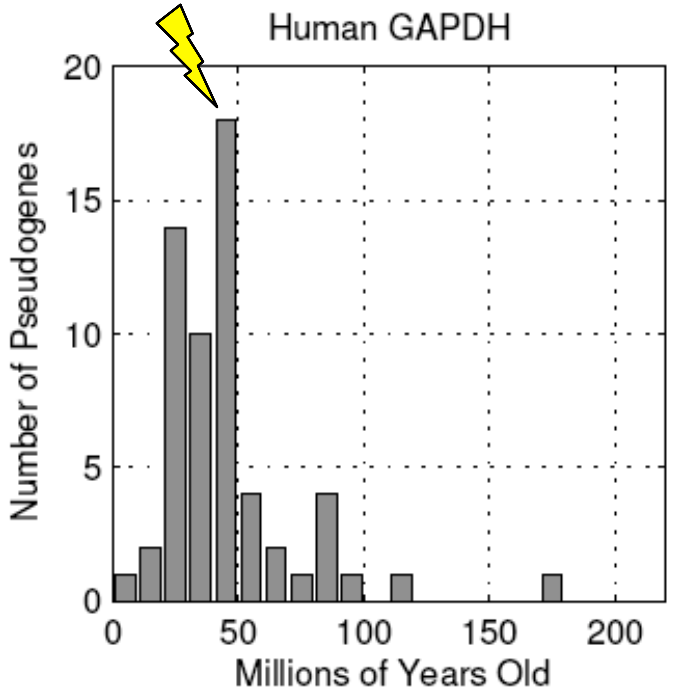


60 Proc/2 Dup



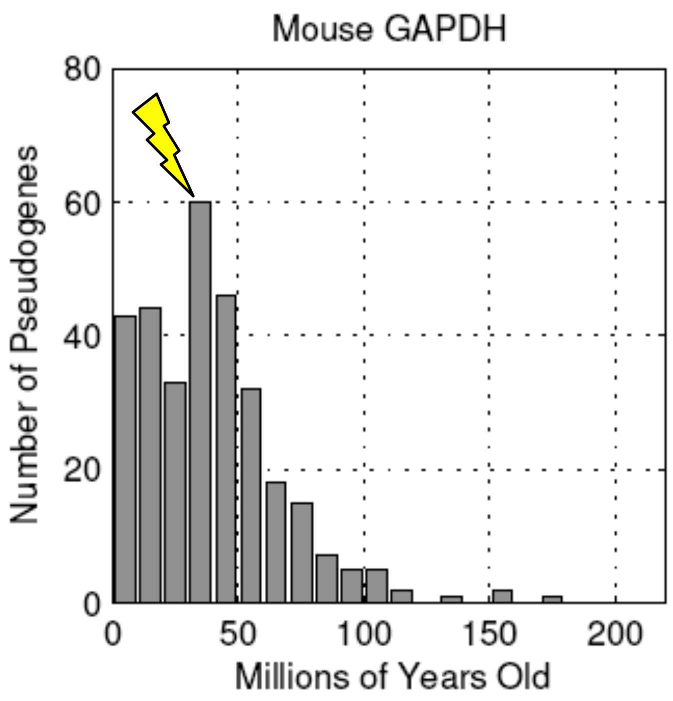
[Liu et al. BMC Genomics ('09, in press)]

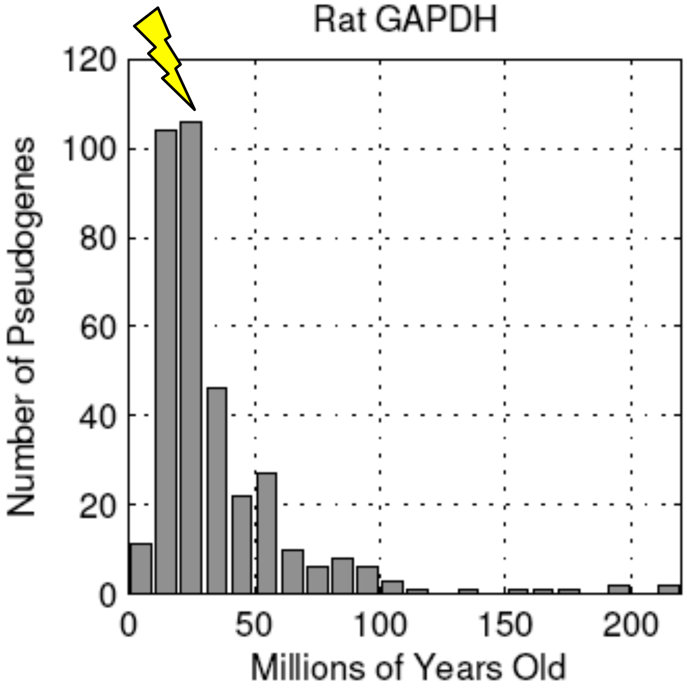
Aproximate Age of GAPDH pseudogenes



Burst of Retrotranspositional Activity

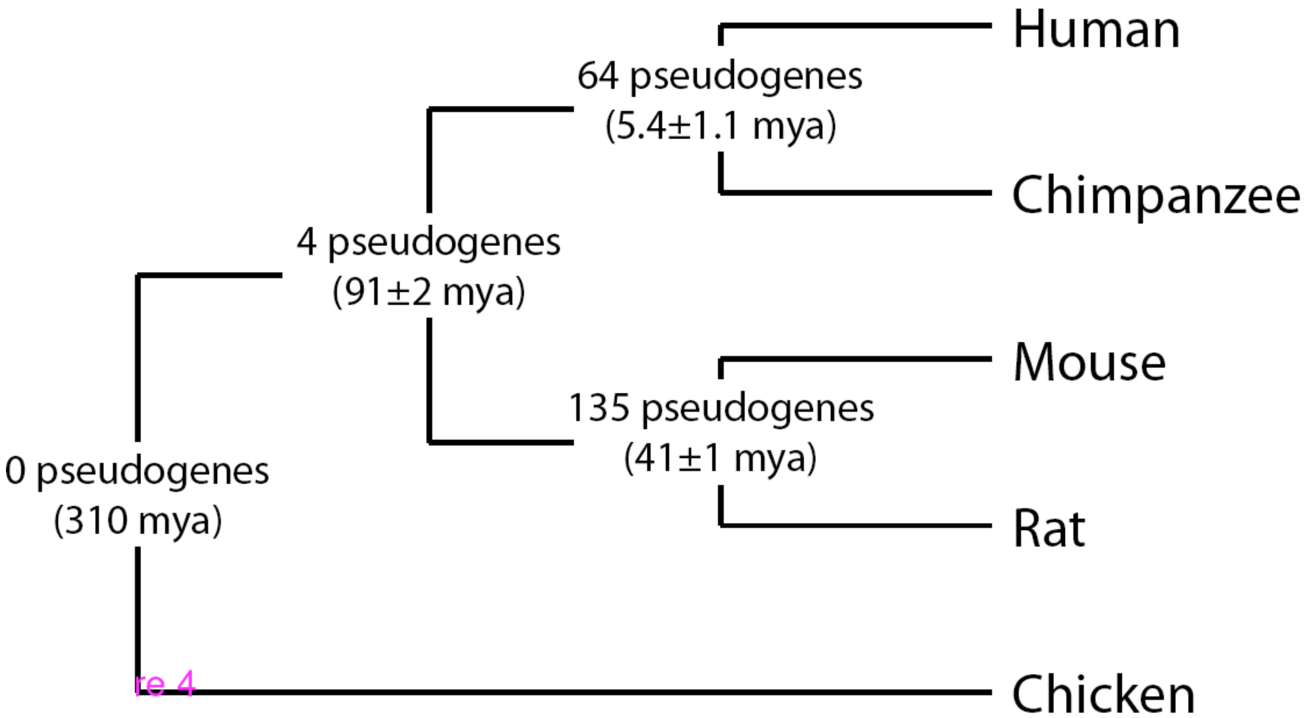
Age calculated based on Kimura-2 parameter model of nucleotide substitution

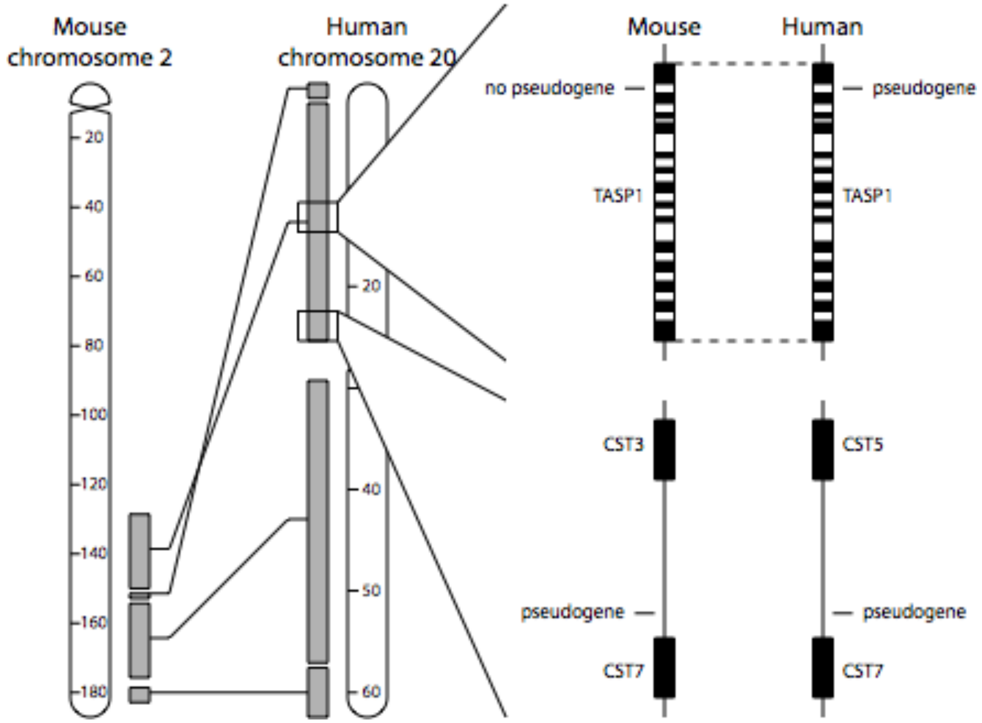




[Liu et al. BMC Genomics ('09)]

Synteny of GAPDH pseudogenes





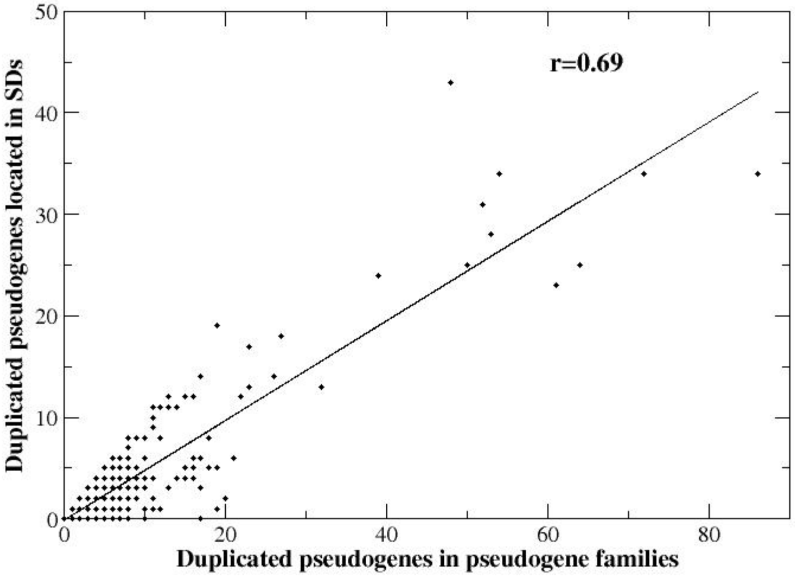
Synteny derived based on local gene orthology

[Liu et al. BMC Genomics ('09)]

Integration of Pseudogenes with Other Features (SDs & Measures of Biochemical Activity)

Pseudogene families and Segmental Duplications (SDs)

- CNVs are the raw form of variation producing duplicated elements
- Fixed CNVs/SVs create SDs, which in turn give rise to duplicated genes and (eventually) protein families
- Thus, we expect, duplicated pseudogenes (failed duplications) to occur in SDs

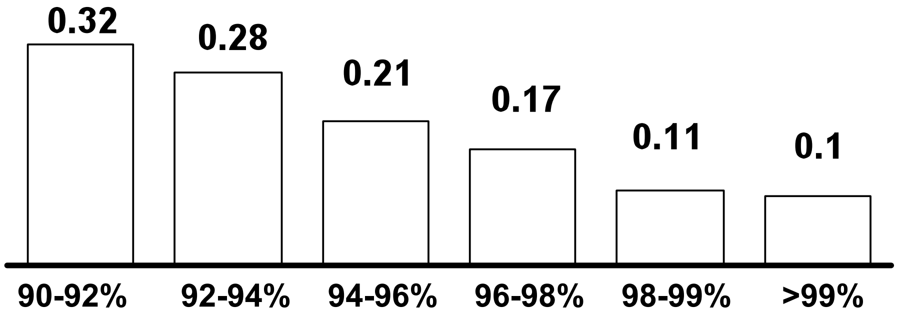


- **SDs comprise ~5% of the human genome but contain ~18% genes, 46% duplicated pgenes and 22% processed pgenes**
- Correlation above consistent with the observation that SDs contain more pgenes than parent genes

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

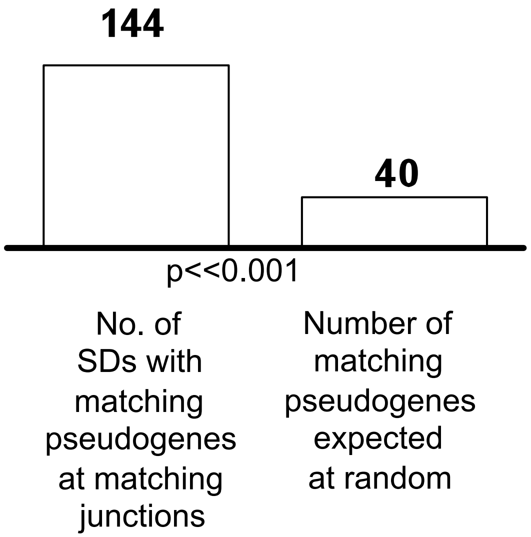
Pseudogenes & CNV/SDs (whole genome, not GAPDH)

Pseudogene association with SDs by age

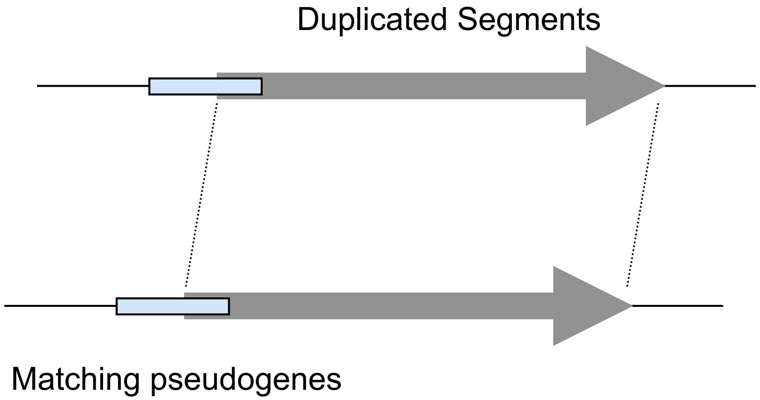


Duplicated pseudogenes associated with SDs, particularly older ones

Processed pseudogenes at SD junctions



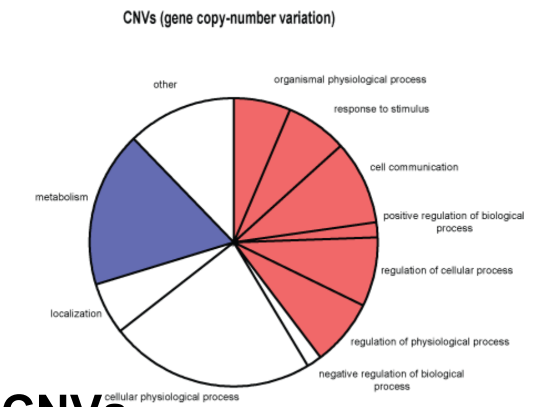
Processed Pseudogenes: serving as repeats for mediating NAHR



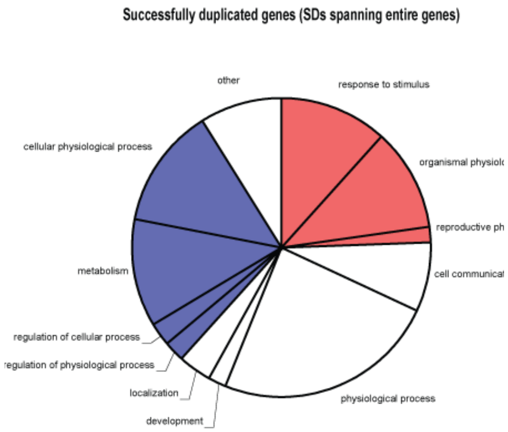
[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]

Association of SDs & CNVs with pseudogenes

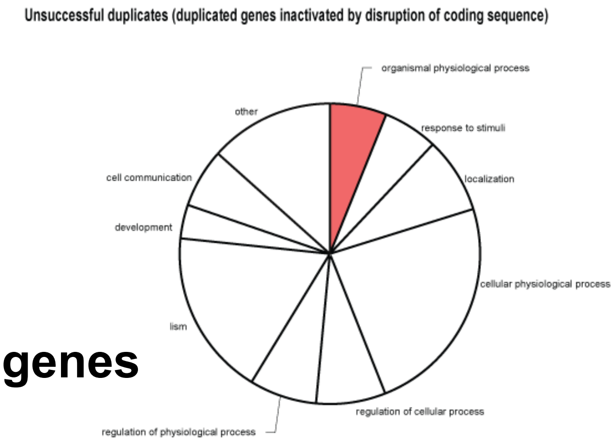
- CNVs & SDs tend to be enriched in environmental response genes, matching patterns found for duplicated pseudogenes



Genes in CNVs



Genes in SDs



Pseudogenes

GO Categories:
Environ-mental Response
Metabolism

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

Vast Amounts of Different Data Types to Integrate in pilot ENCODE

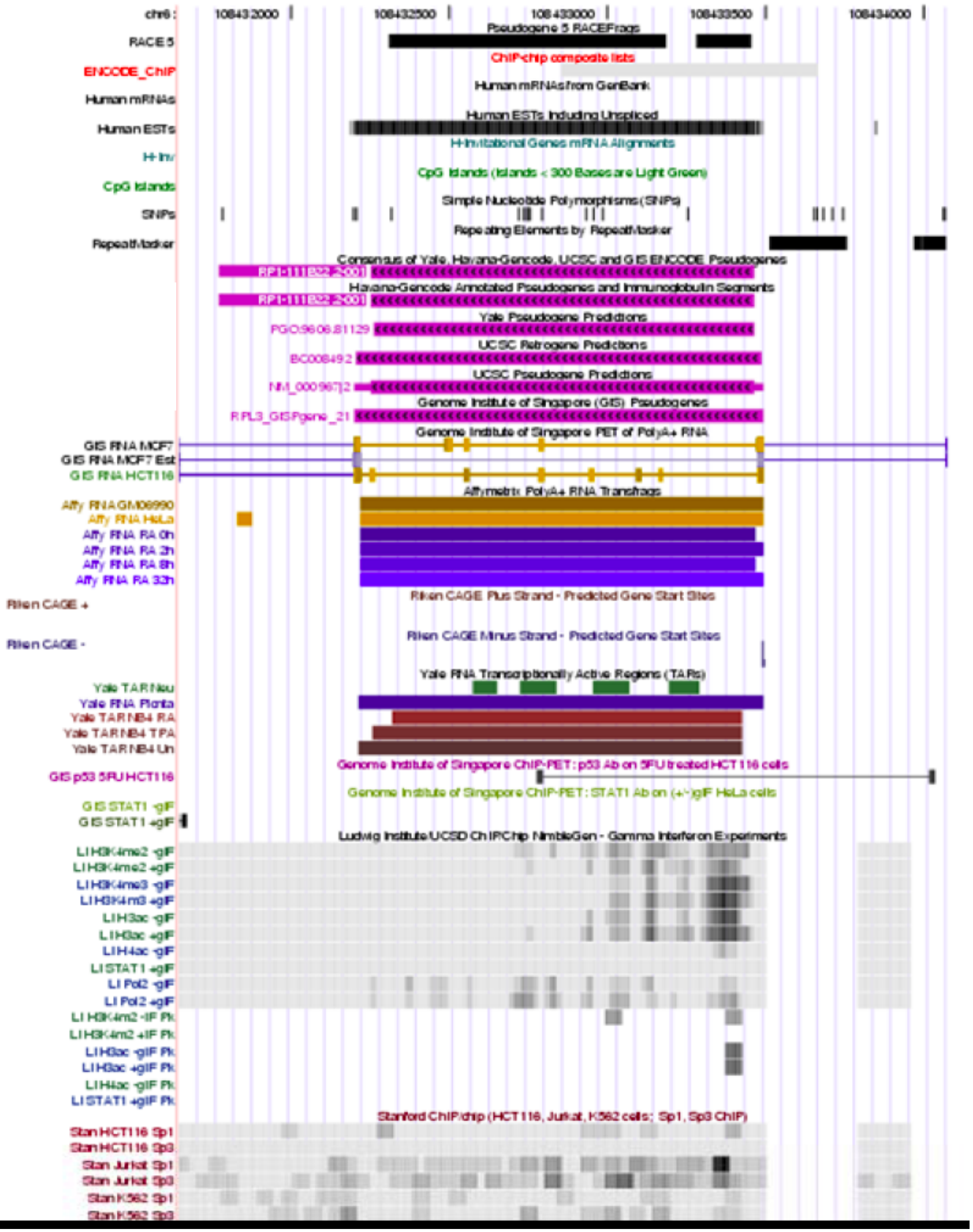
- Determining experimental signals for biochemical activity across each base of genome

- 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]

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


chr6:
108432000
108432500
108433000
108433500
108434000
RACE5
Pseudogene 5 RACEFrags
ENCODE_ChIP
ChIP-chip composite lists
Human mRNAs
Human mRNAs from GenBank
Human ESTs
Human ESTs Including Unspliced
H-Inv
H-Invitational Genes mRNA Alignments
CpG Islands
CpG Islands (Islands < 300 Bases are Light Green)
SNPs
Simple Nucleotide Polymorphisms (SNPs)
RepeatMasker
Repeating Elements by RepeatMasker
Consensus of Yale, Havana-Gencode, UCSC and GIS ENCODE Pseudogenes
RP1-111B22.2-001
Havana-Gencode Annotated Pseudogenes and Immunoglobulin Segments
RP1-111B22.2-001
Yale Pseudogene Predictions
PGOHS06.81129
UCSC Retrogene Predictions
BC008492
UCSC Pseudogene Predictions
NM_000987.2
Genome Institute of Singapore (GIS) Pseudogenes
RPL3_GISPgene_21
Genome Institute of Singapore PET of PolyA+ RNA
GIS RNA MCF7
GIS RNA MCF7 Est
GIS RNA HCT116
Affymetrix PolyA+ RNA Transfrags
Affy RNA GM06990
Affy RNA HeLa
Affy RNA RA 0h
Affy RNA RA 2h
Affy RNA RA 8h
Affy RNA RA 32h
Riken CAGE +
Riken CAGE Plus Strand - Predicted Gene Start Sites
Riken CAGE -
Riken CAGE Minus Strand - Predicted Gene Start Sites
Yale RNA Transcriptionally Active Regions (TARs)
Yale TAR Neu
Yale RNA Placta
Yale TAR NB4 RA
Yale TAR NB4 TPA
Yale TAR NB4 Un
Genome Institute of Singapore ChIP-PET: p53 Ab on 5FU treated HCT116 cells
GIS p53 5FU HCT116
Genome Institute of Singapore ChIP-PET: STAT1 Ab on (+/-)gIF HeLa cells
GIS STAT1 -gIF
GIS STAT1 +gIF
Ludwig Institute/UCSD ChIP/Chip NimbleGen - Gamma Interferon Experiments
LI H3K4me2 -gIF
LI H3K4me2 +gIF
LI H3K4me3 -gIF
LI H3K4m3 +gIF
LI H3ac -gIF
LI H3ac +gIF
LI H4ac -gIF
LI STAT1 +gIF
LI Pol2 -gIF
LI Pol2 +gIF
LI H3K4m2 -IF Pk
LI H3K4m2 +IF Pk
LI H3ac -gIF Pk
LI H3ac +gIF Pk
LI H4ac -gIF Pk
LI STAT1 +gIF Pk
Stanford ChIP/chip (HCT116, Jurkat, K562 cells; Sp1, Sp3 ChIP)
Stan HCT116 Sp1
Stan HCT116 Sp3
Stan Jurkat Sp1
Stan Jurkat Sp3
Stan K562 Sp1
Stan K562 Sp3

Composite ChIP hit

Special ψG tracks in browser

diTAG

CAGE

TARs

ChIP-chip

- 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 🟦

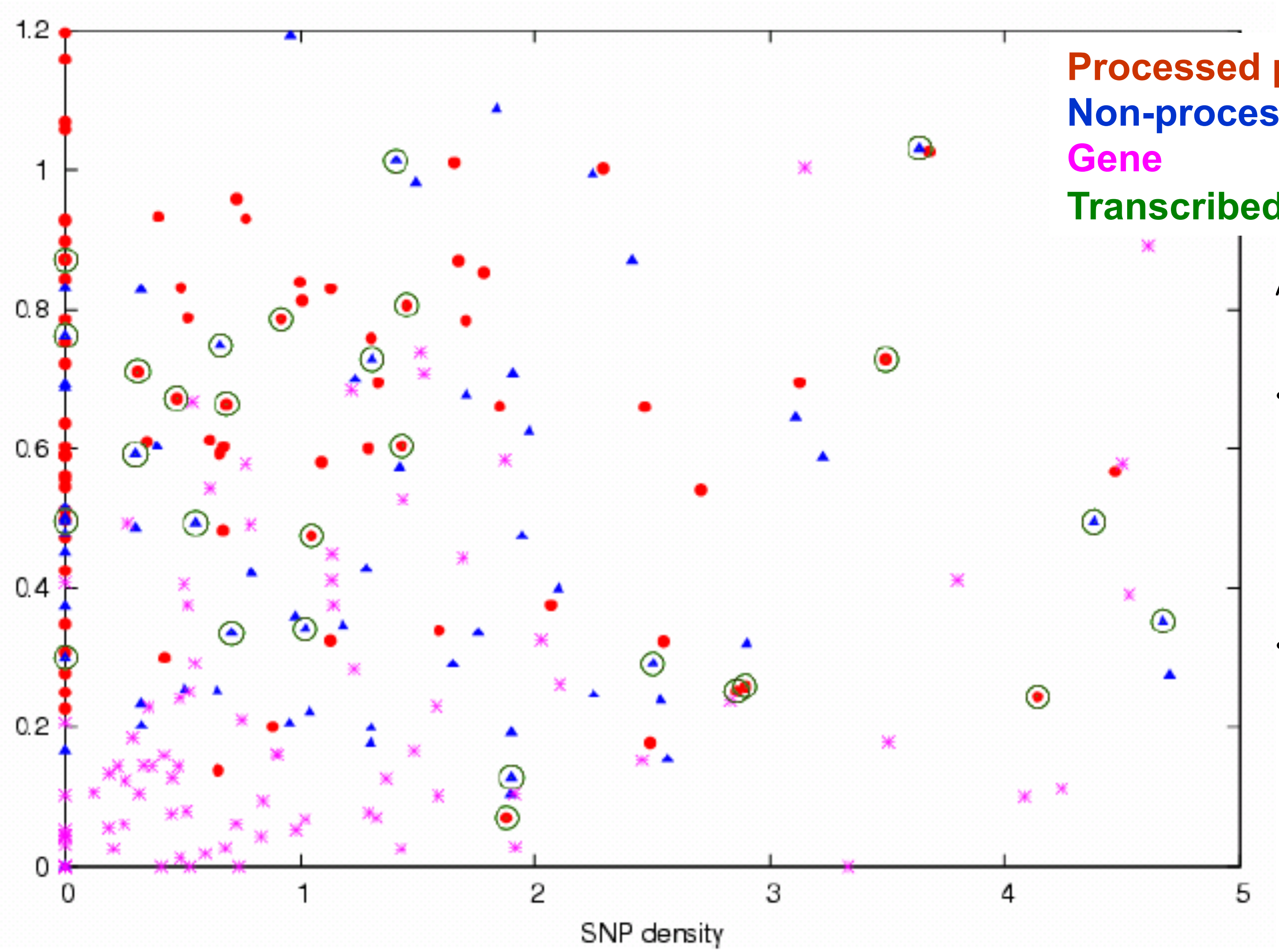
Excluding TARs (due to cross-hyb issues)

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 ○

Ka/Ks

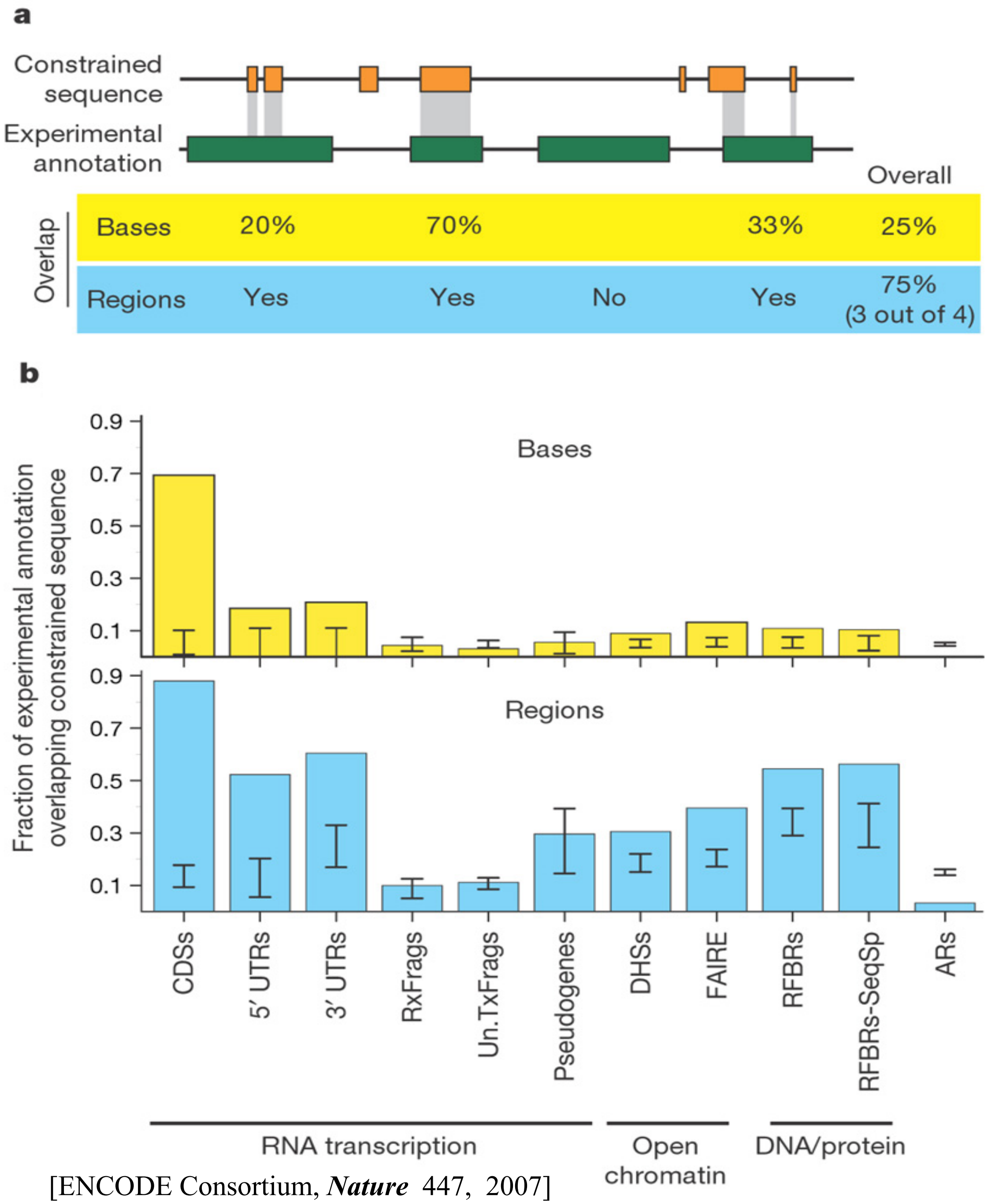
Measurement of Short-time variation (pN+pS)

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)

Zheng et al. (2007) Gen. Res.

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

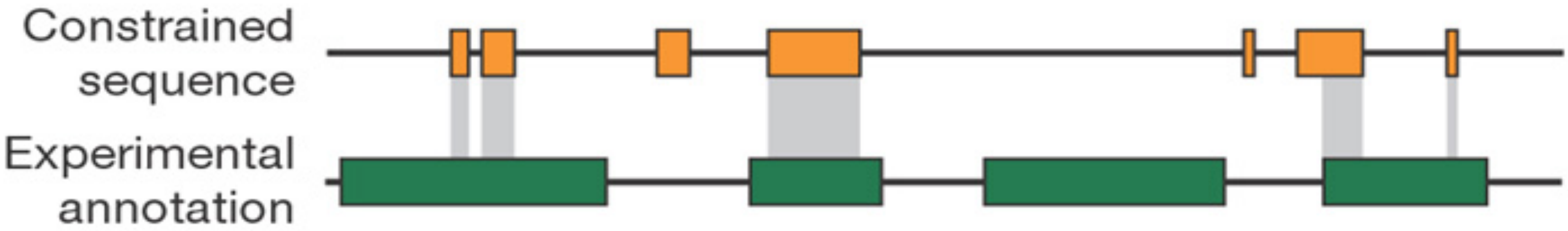


Overlap					Overall
Bases	20%	70%		33%	25%
Regions	Yes	Yes	No	Yes	75% (3 out of 4)

[ENCODE Consortium, *Nature* 447, 2007]

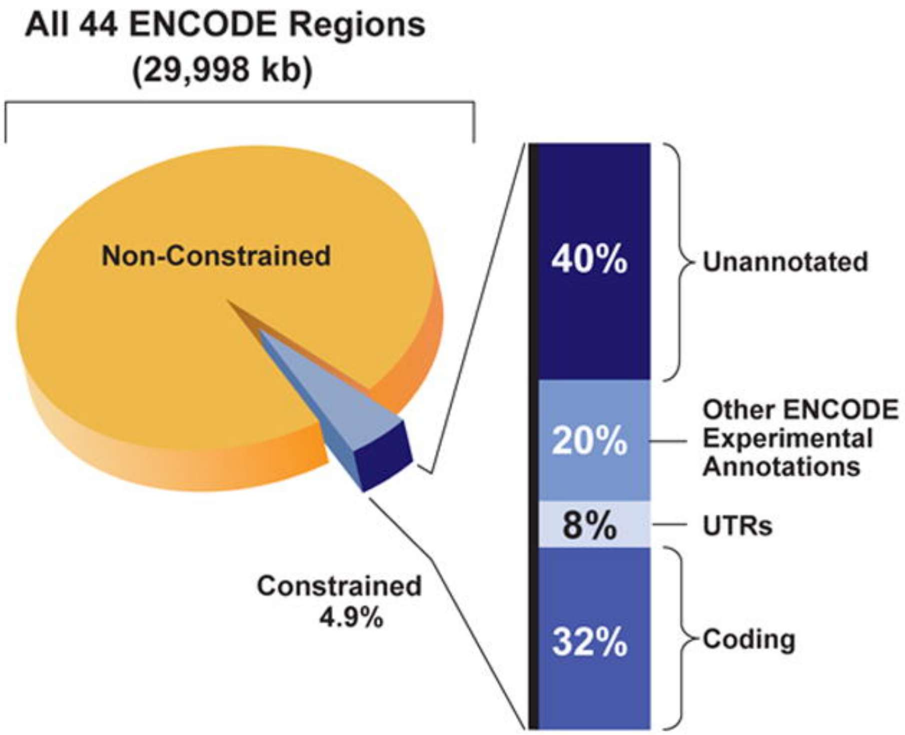
- Integrating & averaging results over larger and larger sets
- Comparison of integrated quantities

Grand Summary: Biochemical Activity vs. Sequence Constraints

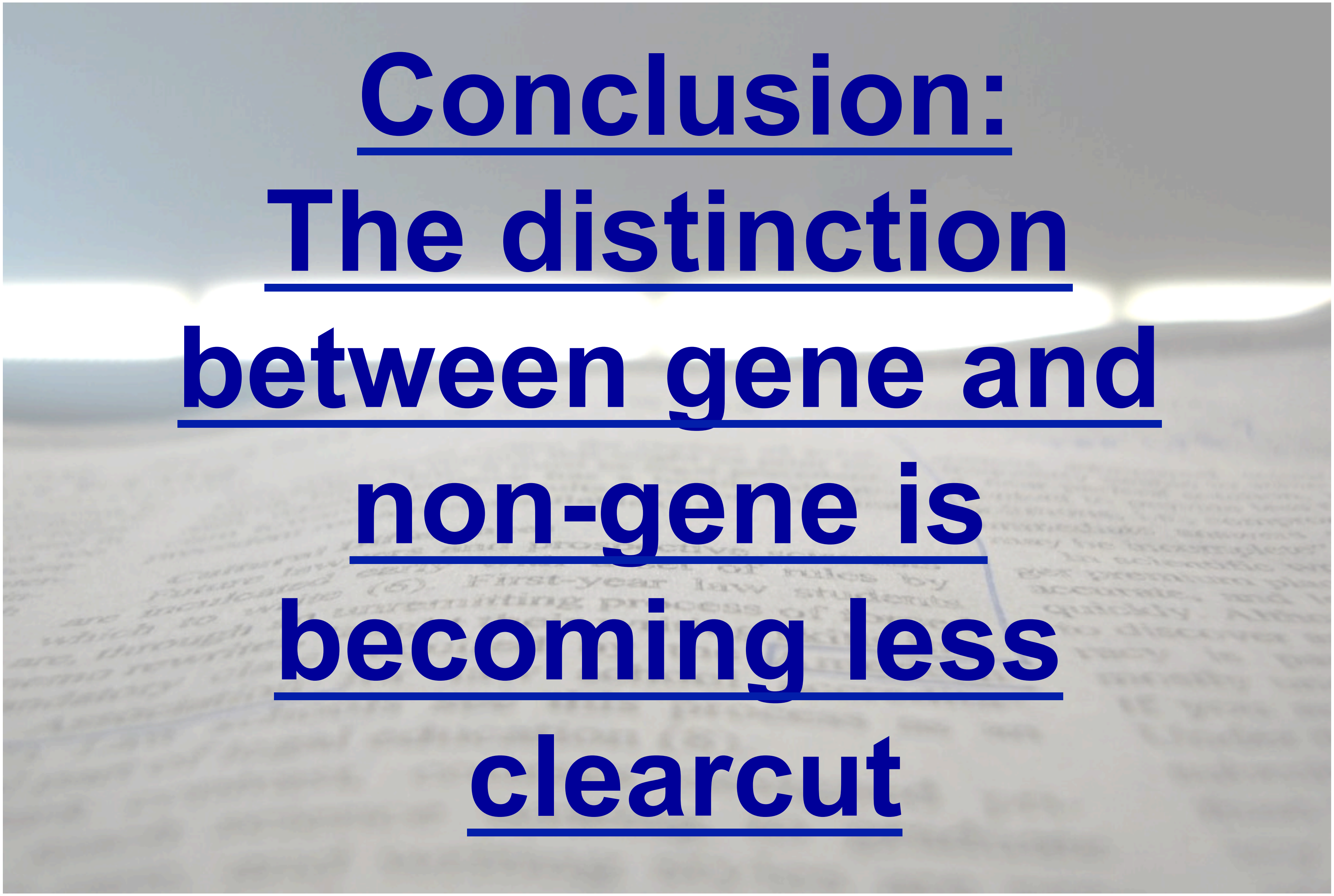


- Not all constrained sequence annotated in some fashion
- Exactly how things are defined in terms of overlap?

- "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. "

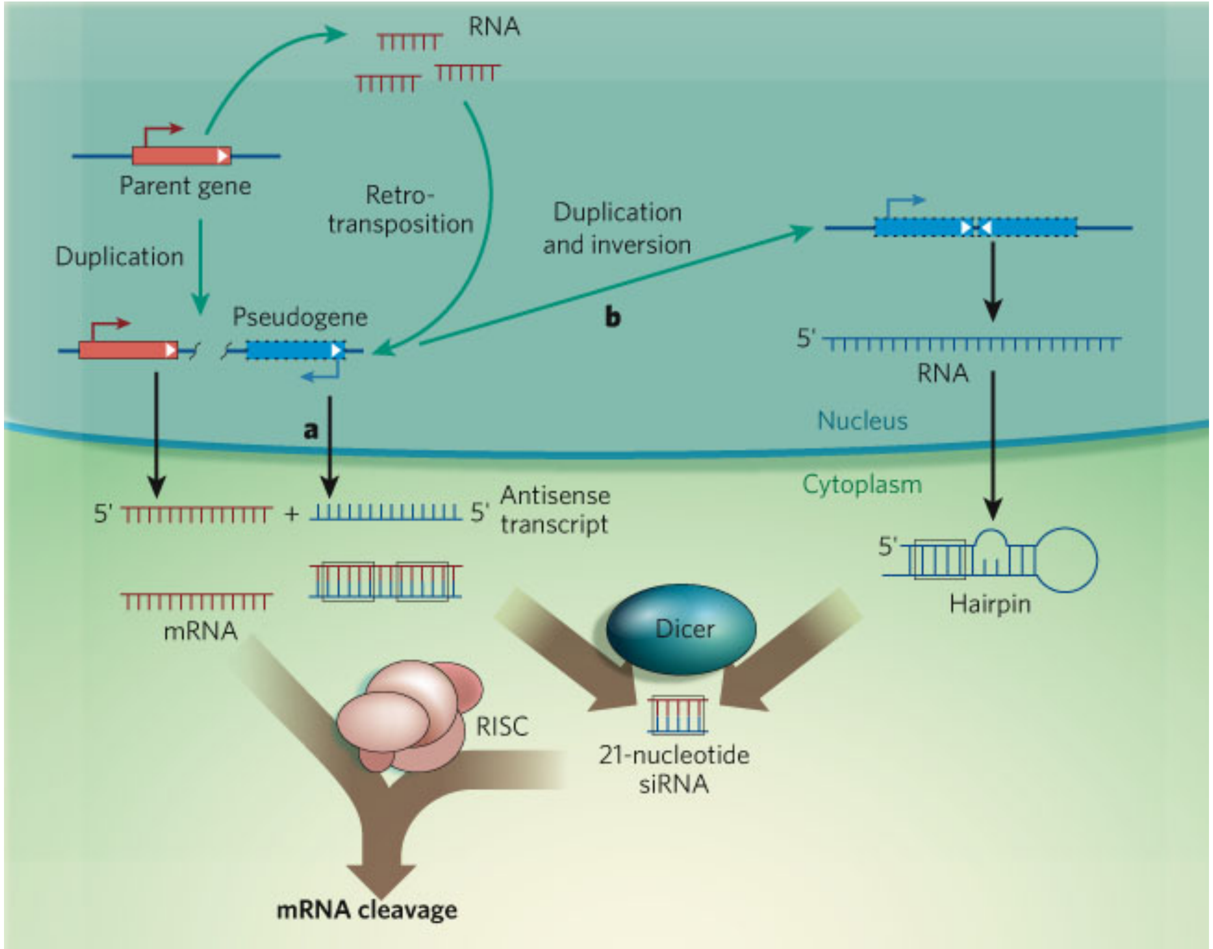


[ENCODE Consortium, *Nature* 447, 2007]



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

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



What are Active Pseudogenes Doing?

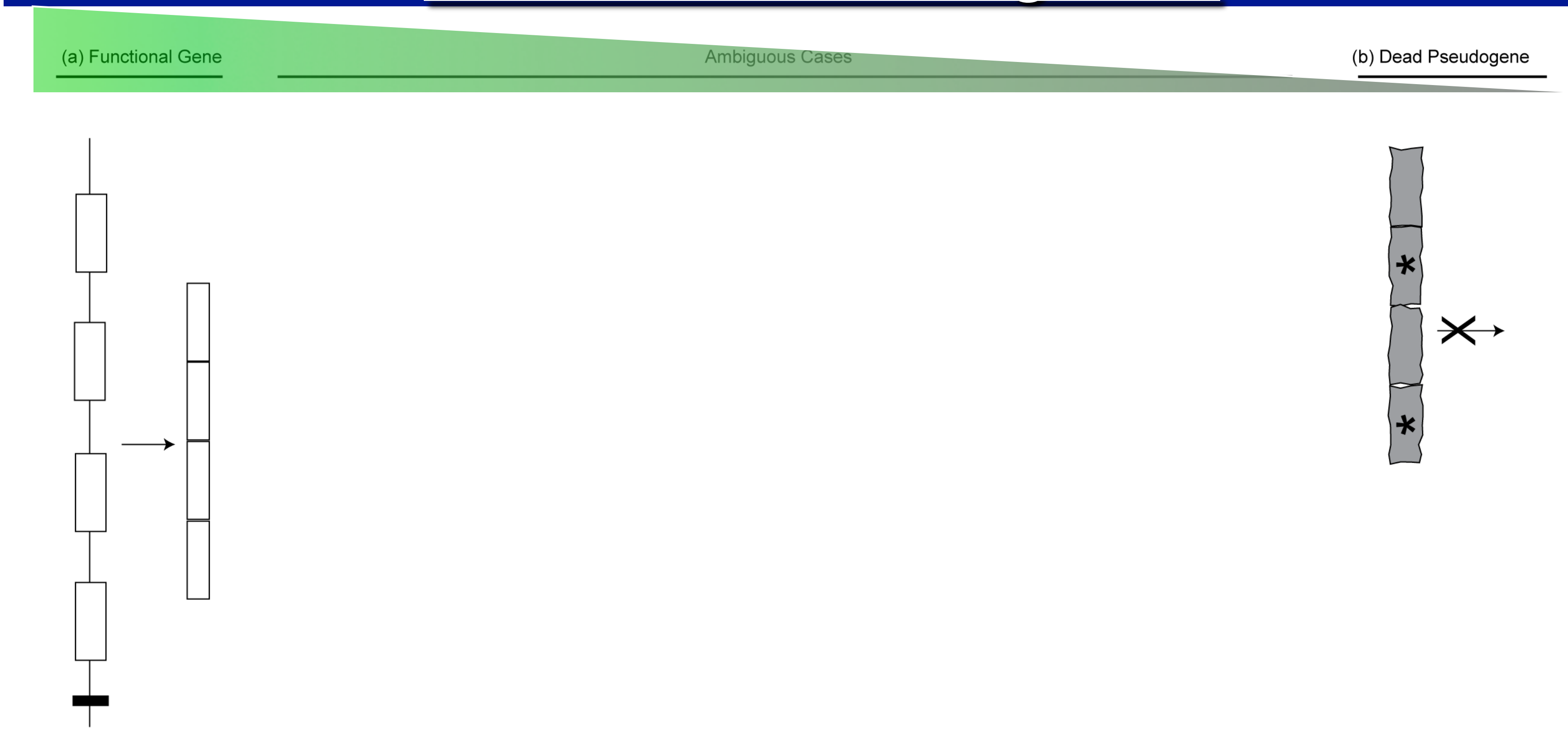
Potential for Gene Regulation via endo-siRNA

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)]

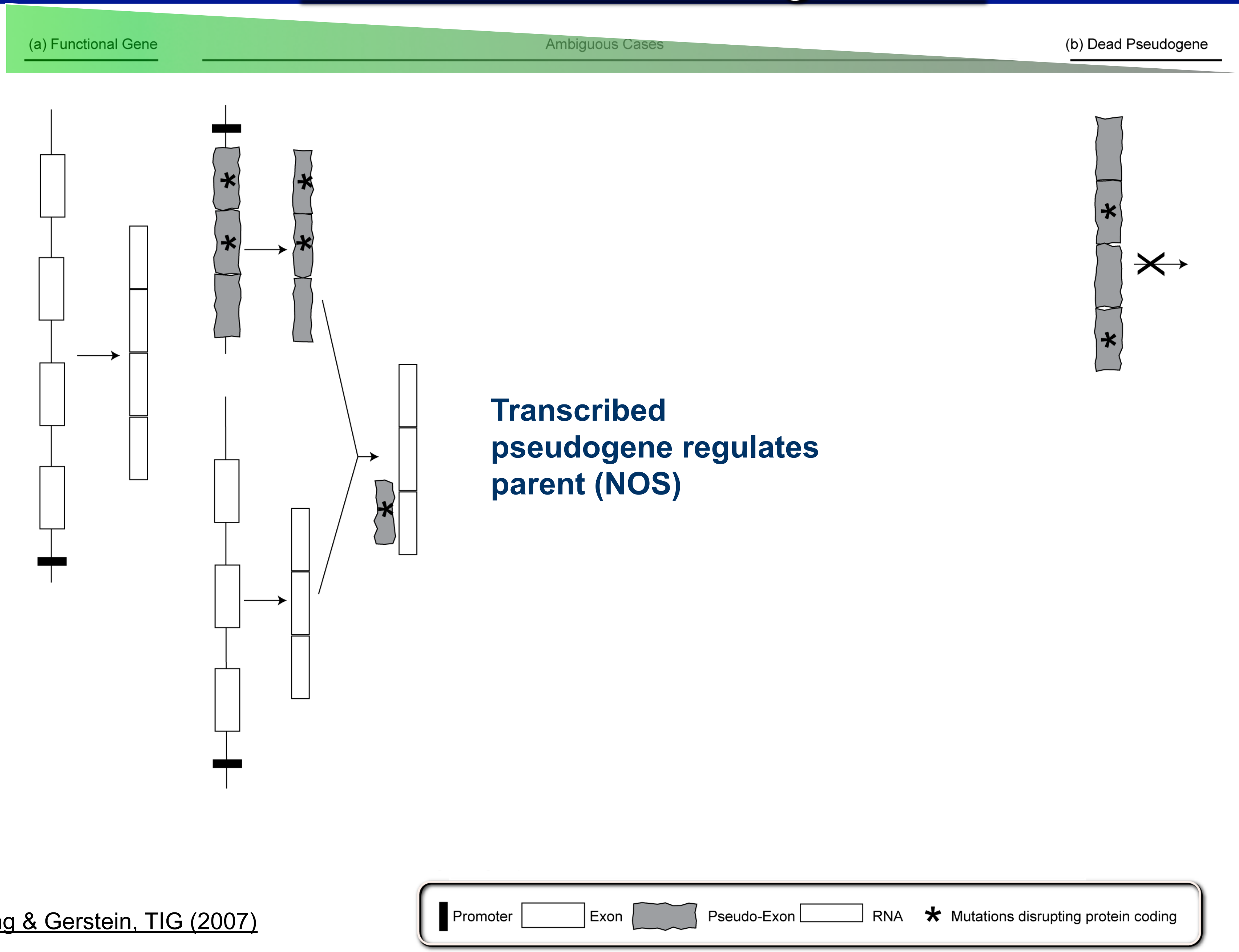
Genes & Pseudogenes



▌ Promoter □ Exon ▨ Pseudo-Exon □ RNA * Mutations disrupting protein coding

Zheng & Gerstein, TIG (2007)

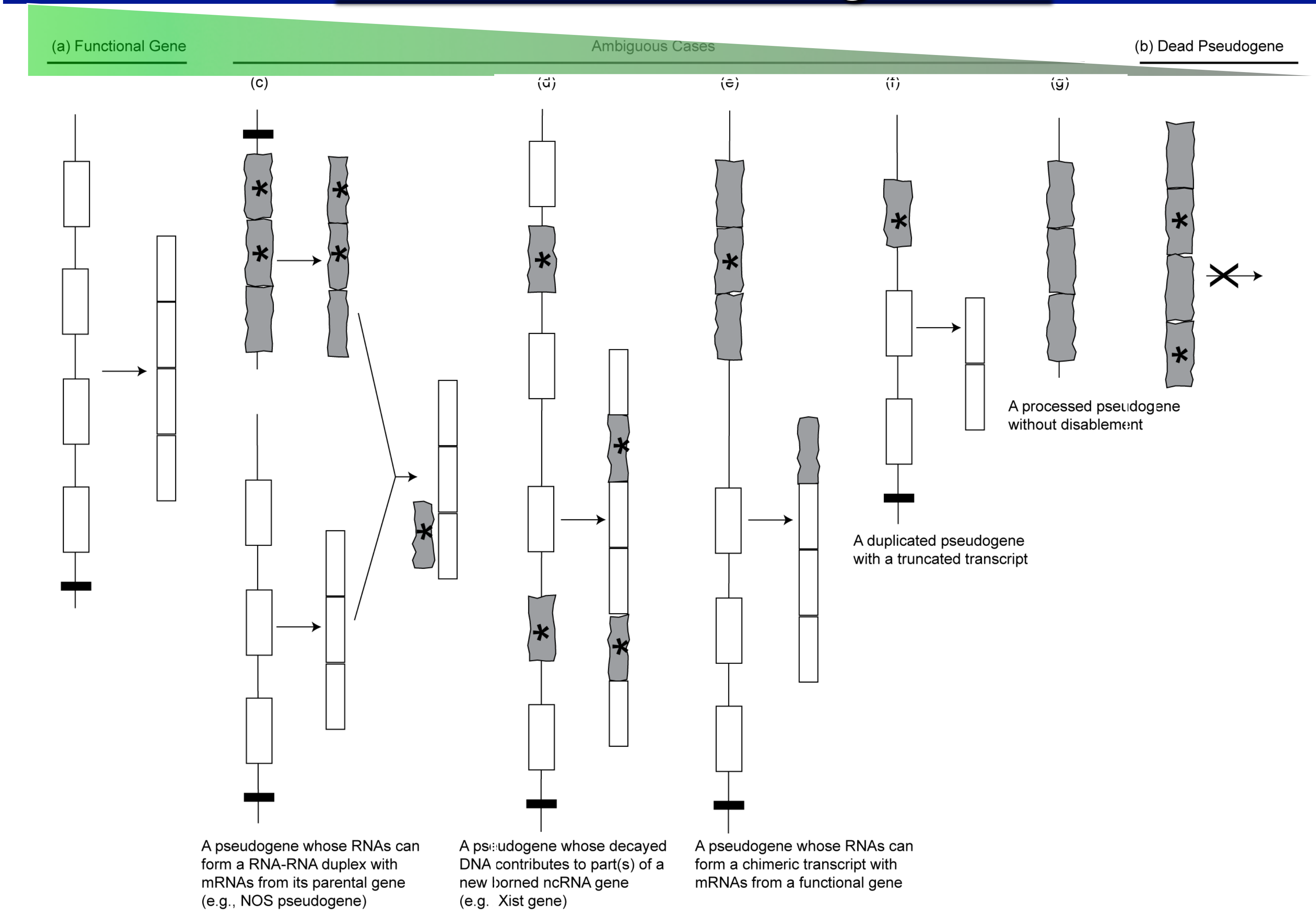
Genes or Pseudogenes?



Transcribed pseudogene regulates parent (NOS)

Zheng & Gerstein, TIG (2007)

Genes or Pseudogenes?



A pseudogene whose RNAs can form a RNA-RNA duplex with mRNAs from its parental gene (e.g., NOS pseudogene)

A pseudogene whose decayed DNA contributes to part(s) of a new borned ncRNA gene (e.g. Xist gene)

A pseudogene whose RNAs can form a chimeric transcript with mRNAs from a functional gene

A duplicated pseudogene with a truncated transcript

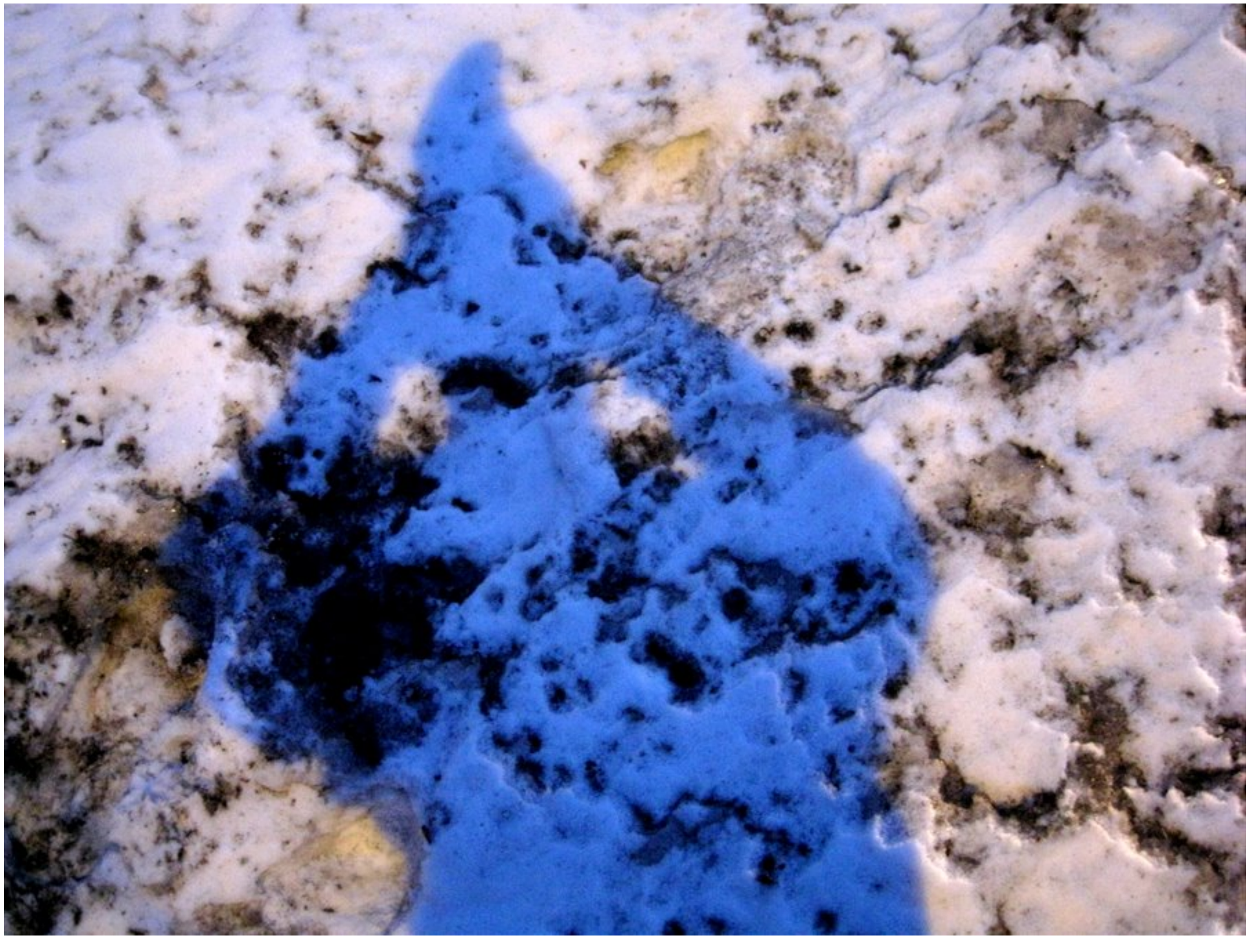
A processed pseudogene without disablement

▌ Promoter □ Exon ▨ Pseudo-Exon □ RNA * Mutations disrupting protein coding

Zheng & Gerstein, TIG (2007)

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 activity (e.g. transcription) across each base of genome

- **Results of Analyzing Similarity Comparison**

  A. Finding repeated or deleted blocks
     1. As a function of similarity (age)
     2. vs. other organisms or vs. human reference
     3. Big and small blocks (duplicated regions and retrotransposed repeats)

- Results of Processing Raw Expt. Signals

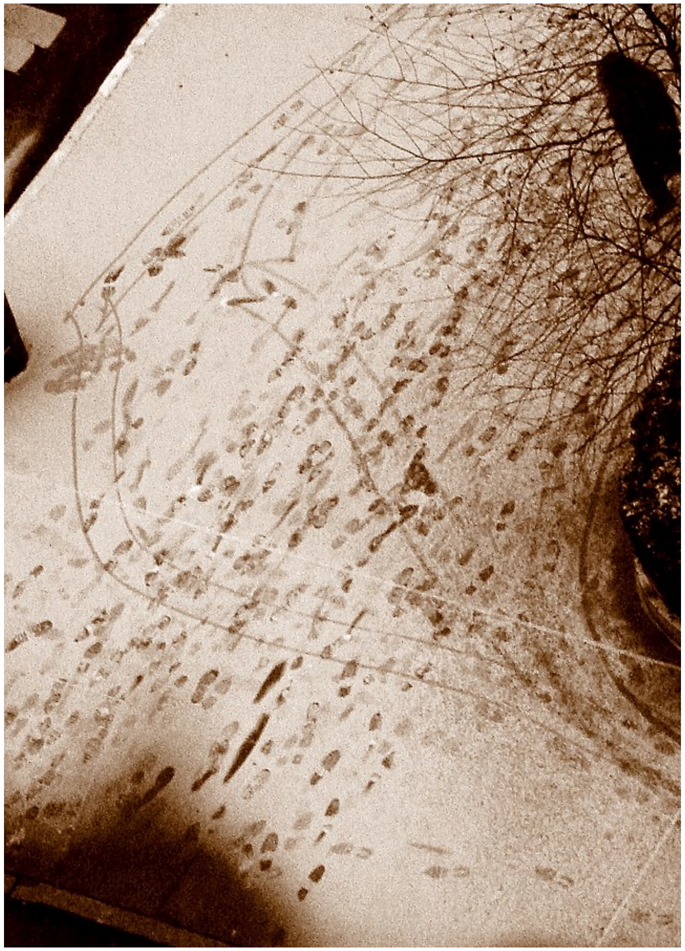
  a. Signal Processing: removing artifacts, normalizing, window averaging

  a. Segmenting signal into larger "hits"

  b. Clustering together active regions into even larger features at different length scales and classifying them

  c. Integrating Annotations, Building networks and beyond....

Outline



- Regulatory Sites
  - a. ChipSeq signal processing to call puncate "hits"
  - b. Clustering of hits into broader blocks and annotating them
- Variable Blocks in Genome (CNVs,SDs)
  - A/a. Calling them with various signal processing approaches (MSB, PEMer, ReSeqSim)
  - b. Grouping CNVs & SDs into larger features and inter-relating them
- Pseudogenes
  - A. Pattern-match tools for calling them
  - A. Focus on one group of pseudogenes
  - c. Integrating them with other annotations (transcription, regulation, CNVs, SDs)
- Future of Annotation
  - ◊ What is a "gene" post encode?

PeakSeq + Biplots

- Segmenting the Raw "Signal" from Next-generation Sequencing into Usable Annotation Blocks (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 Groups of Binding Sites (Biplots)
  - ◊ on ~50kb scale
  - ◊ Gives broad separation of seq. specific and non-specific factors and associated genomic bins

Signal Processing #2: Identifying Structural Variants in Human Population

- BreakPtr
  - ◊ Model-based segmentation using bivariate HMM
- MSB
  - ◊ Mean-shift segmentation approach following grad. of PDF
  - ◊ Equally applied to aCGH and depth of coverage of short reads
- PEMer
  - ◊ Detecting Variants from discordantly placed paired-ends
  - ◊ Simulation to paramaterize statistical model
- ReSeqSim
  - ◊ Efficiently simulating assembly of a representative variant
  - ◊ Shows that best reconstruction has a combination of long, med. and short reads

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
  - ◊ Ontology
  - ◊ Pseudofam analysis of Pseudogene Families
- Annotation of Human Genome
  - ◊ Pipeline draft (20K) + Hybrid Approach
- Glycolytic pseudogenes
  - ◊ Great variation in number, with GAPDH the largest
  - ◊ Synteny & dating shows most GAPDH ones are recent, resulting from retrotranspositional bursts
- Association with SDs
  - ◊ As expected, duplicated pseudogenes associated with SDs and processed pseudogenes like Alus are near SD junctions
- Pseudogene Activity
  - ◊ >20% appear to be transcribed (38/201)
  - ◊ No obvious selection on transcribed ones

Consortia Acknowledgements

Adam Frankish, Robert Baertsch, M Diekhans,R Harte,
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, J Affourtit, M Egholm

Sanger, UCSC, GIS, AFFX, 454, Geneva, IMIM, BU + SU

+

ENCODE, modENCODE, 1000 Genomes

Yale Acknowledgements

D Zheng
Zhengdong Zhang
Y J Liu
YK Lam
J Du
J Rozowsky
J Korbel
L Wang
M Snyder
S Weissman

P Kim
S Balasubramanian

E Khurana
G Fang
R Sasidharan
J Karro
G Euskirchen
J Chang
R Bjornson
N Carriero
X Mu
T Gibson
R Robilotto
Y Liu
D Greenbaum
A Urban
T Royce
P Cayting
R Auerbach
E Khurana
A Abyzov
J Wu
Zhaolei Zhang

GenomeTECH.gersteinlab.org
Pseudogene.org

More Information on this Talk

SUBJECT: GenomeTechAnnote

DESCRIPTION:
Structural Studies, LMB, Cambridge, UK, 2009.12.01, 10:15-11:15;
[I:**LMB**] (Long GenomeTechAnnote talk, building on [I:**UCSC**] .)

(Works equally well on mac or PC. 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**)

PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2008. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

.

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** .