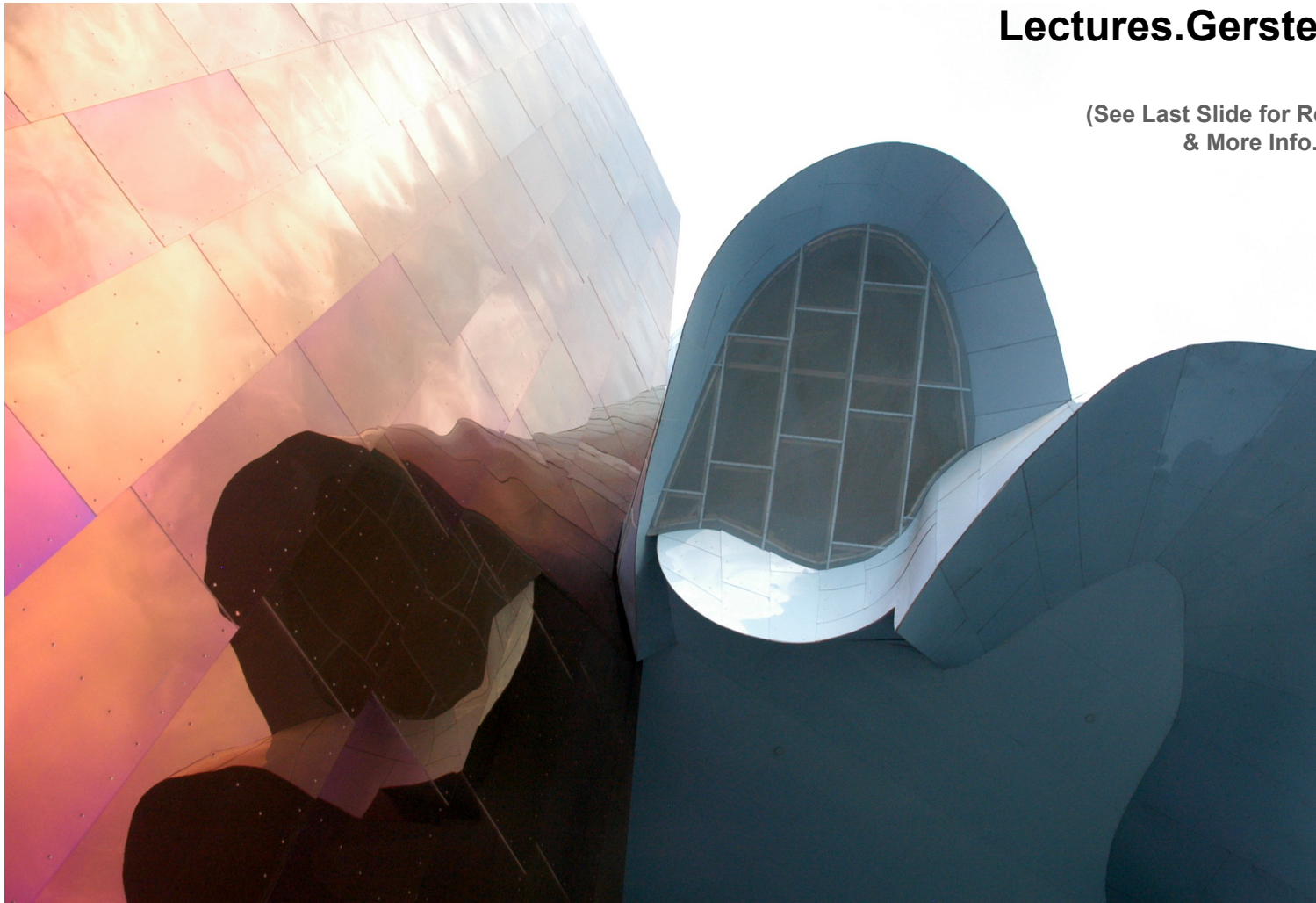


Human Genome Annotation

Mark B Gerstein
Yale

Slides at
Lectures.GersteinLab.org

(See Last Slide for References
& More Info.)

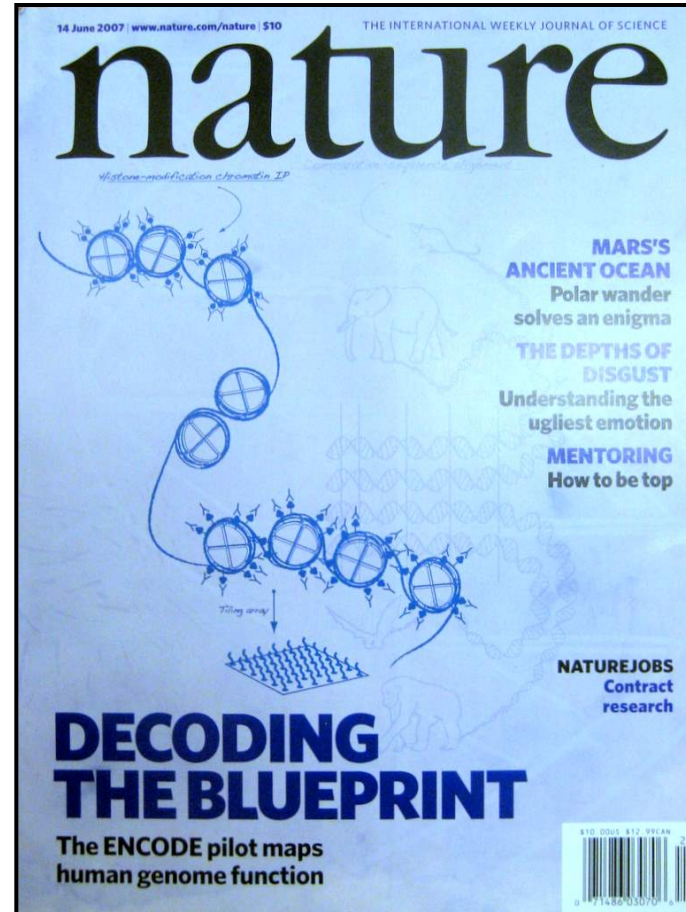




2001: Most of the genome is not coding (only ~1.2% exon). It consists of elements such as repeats, regulatory regions, non-coding RNAs, origins of replication, pseudogenes, segmental duplications....What do these elements do? How should they be annotated?

[IHGSC, *Nature* 409, 2001]

[Venter et al. *Science* 29, 2001]



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

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

[IHGSC, *Nature* 409, 2001]

[ENCODE Consortium, *Nature* 447, 2007]



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.

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,

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 of our work," Nozomu Yachie said in an e-mail.

"So might ET have inserted a message into the genome of the near teardrop-shaped, intelligent-looking alien life form that we call a cockroach?"

I stayed up all night with my friends playing the game of the near teardrop-shaped, intelligent-looking alien life form that we call a cockroach.

It is the relentless shifting and mutating, the probability of the near teardrop-shaped, intelligent-looking alien life form that we call a cockroach.

after all, that generates the raw material for evolution

sections of junk DNA seem to be markedly resistant to

Startibartfast.

Jimmy Turner

Using the same code that computer keyboards use, the Japanese group... wrote four copies of Albert Einstein's famous formula, $E=mc^2$... 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.

*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

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.

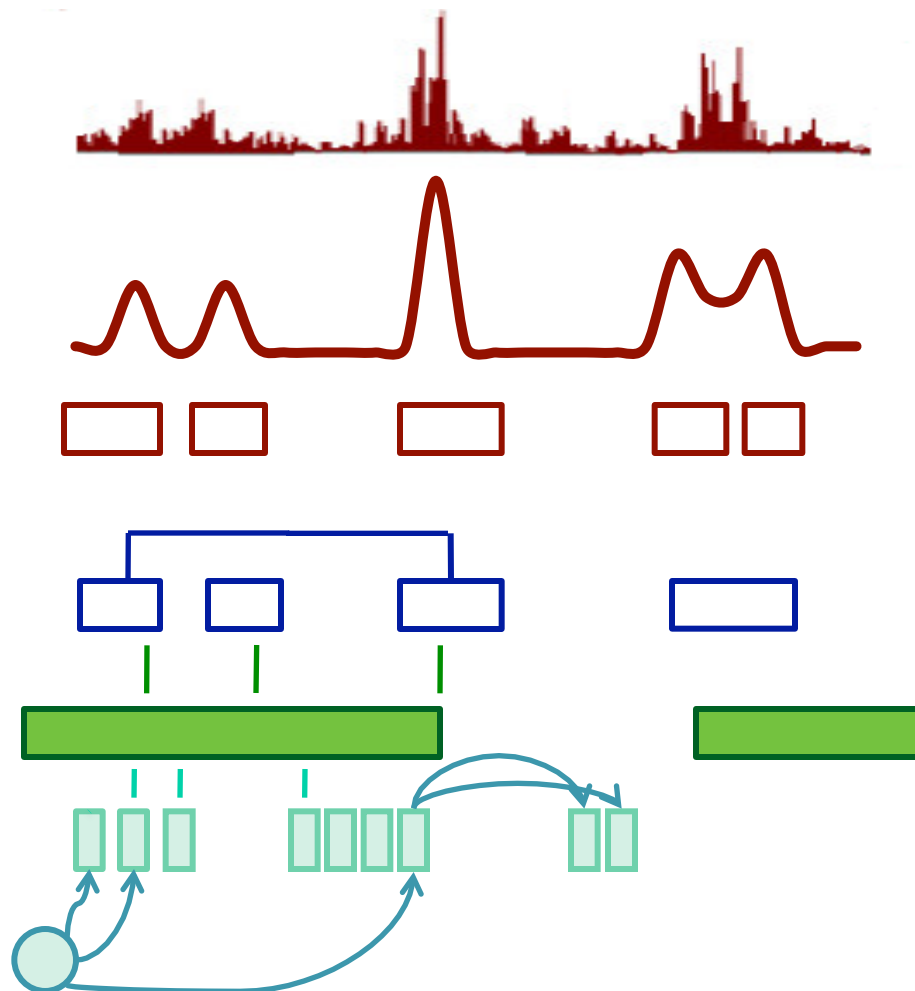
Determining experimental signals for activity (e.g. transcription) across each base of genome

- 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)

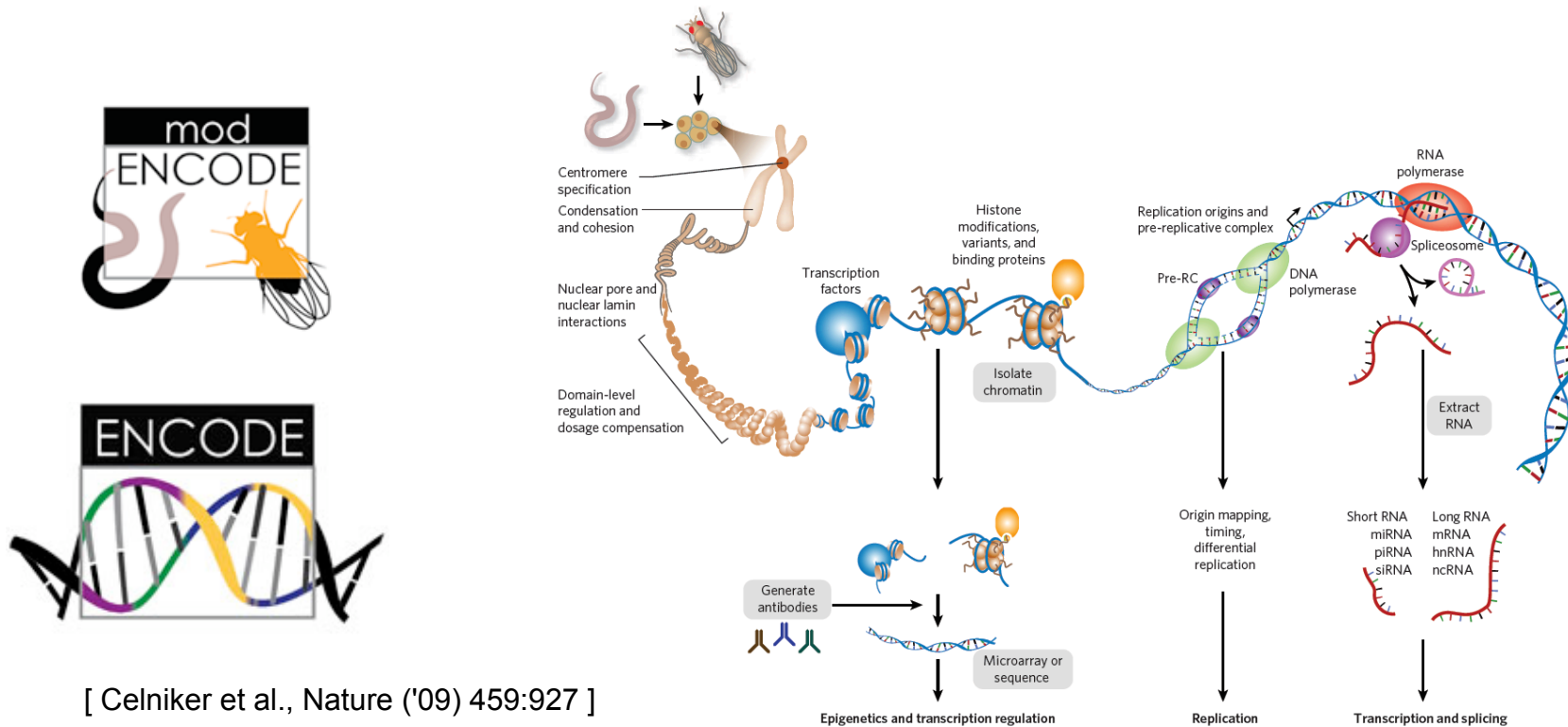
Overview of Functional Genomics Annotation Process



- **Determining experimental signals for biochemical activity (e.g. transcription) across each base of genome**

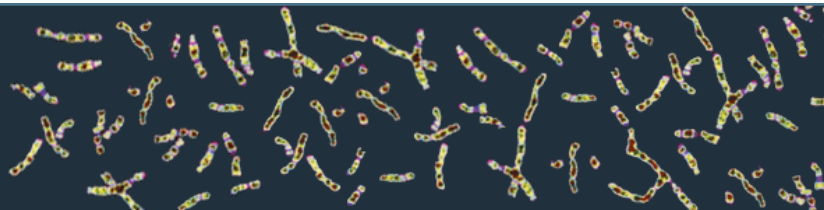
- **Development of Sequence (and Array) Technology**
 - Normalizing & Scoring Signal, Correcting Artifacts, Segmenting to create Small Annotation Blocks
- **Output of Production Pipelines and Surveying a Single Type of Annotation on a Large-scale**
 - Clustering Small Blocks into Larger Ones, Surveying
- **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



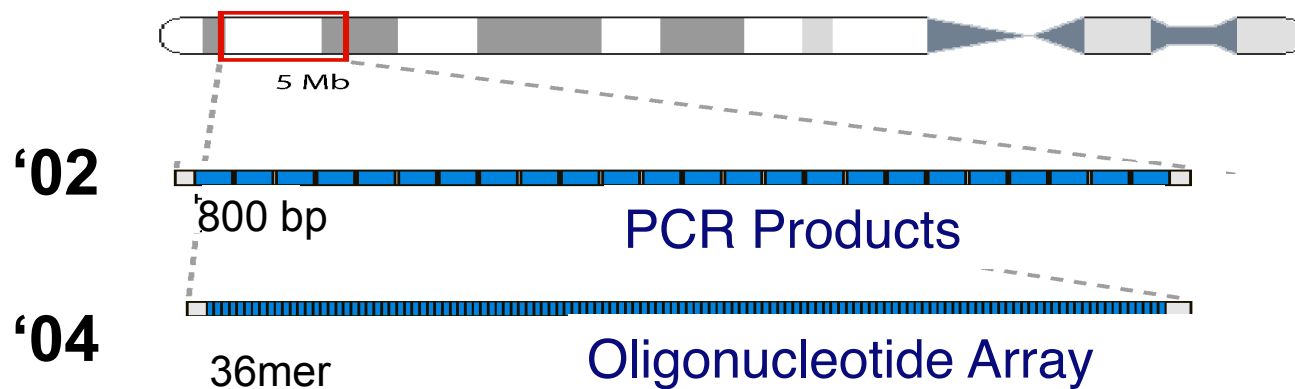
1000 Genomes

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



Application in a variety of contexts:

Transcription Mapping

DNA binding (inc. chromatin struc.)

Replication

Structural Variation

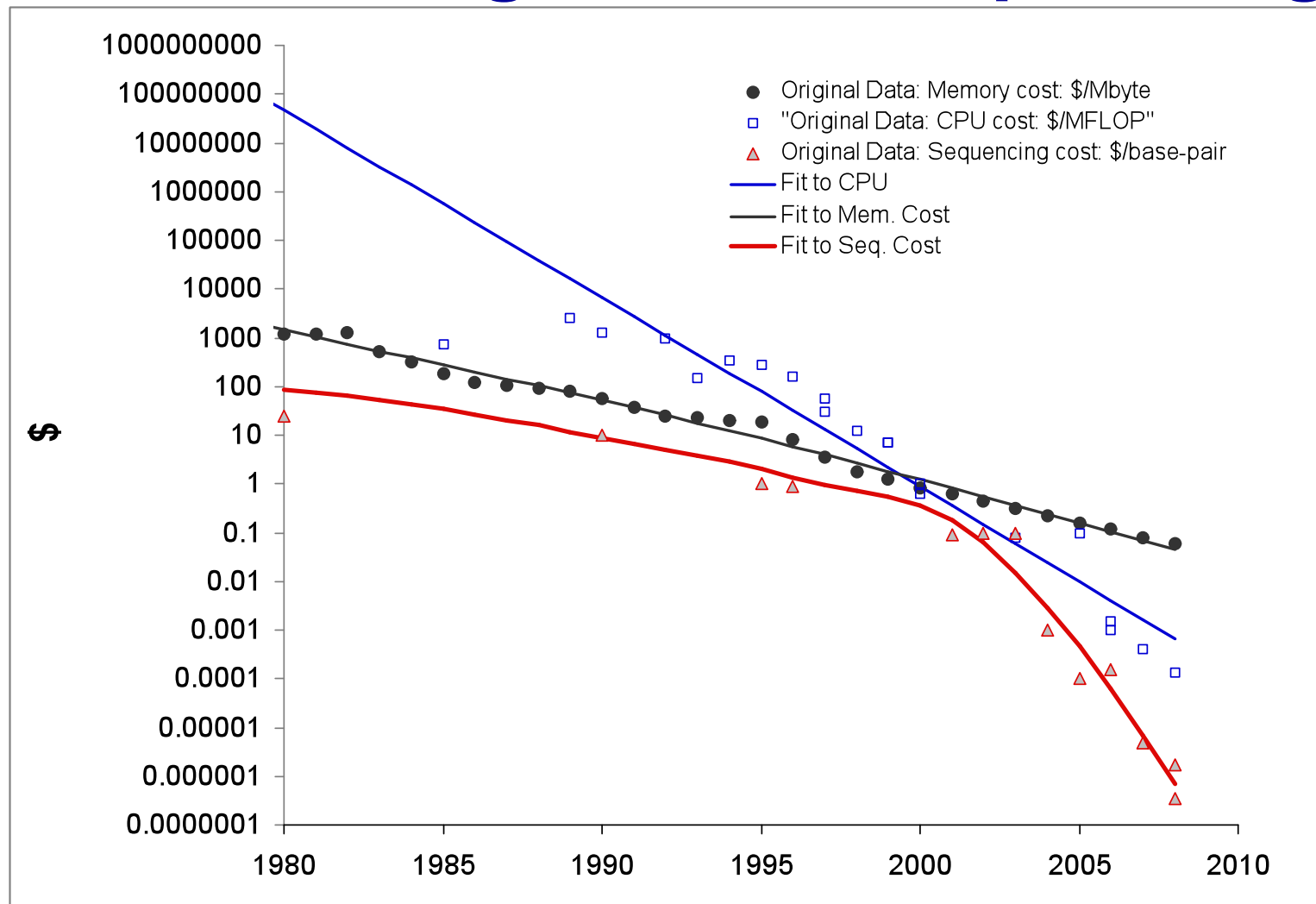
Massively Parallel Sequencing

'06+



AGTTCACCTAAGA...
CTTGAATGCCGAT...
GTCATTCCGCAAT...

Plummeting Cost of Sequencing



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

Outline

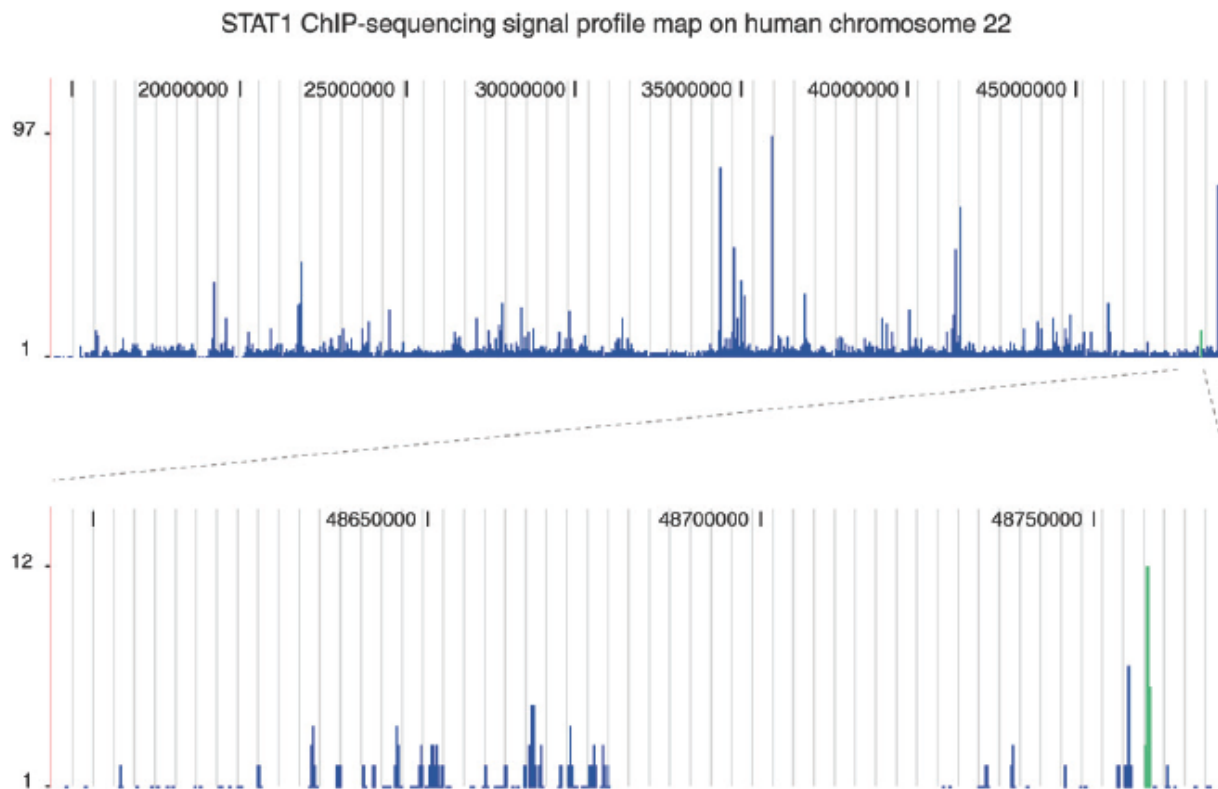


- Regulatory Sites
 - a. ChipSeq signal processing to call punctate "hits"
 - b. Clustering of hits into broader blocks and annotating them
- Variable Blocks in Genome (CNVs,SDs)
 - A/a. Calling them with various signal processing approaches
- Pseudogenes
 - A. Pattern-match tools for calling them
 -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

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

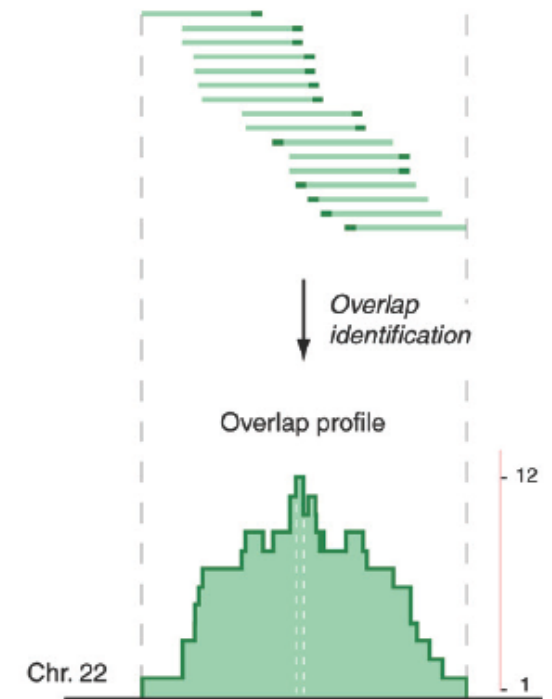


Representative Signal from Chip-Seq



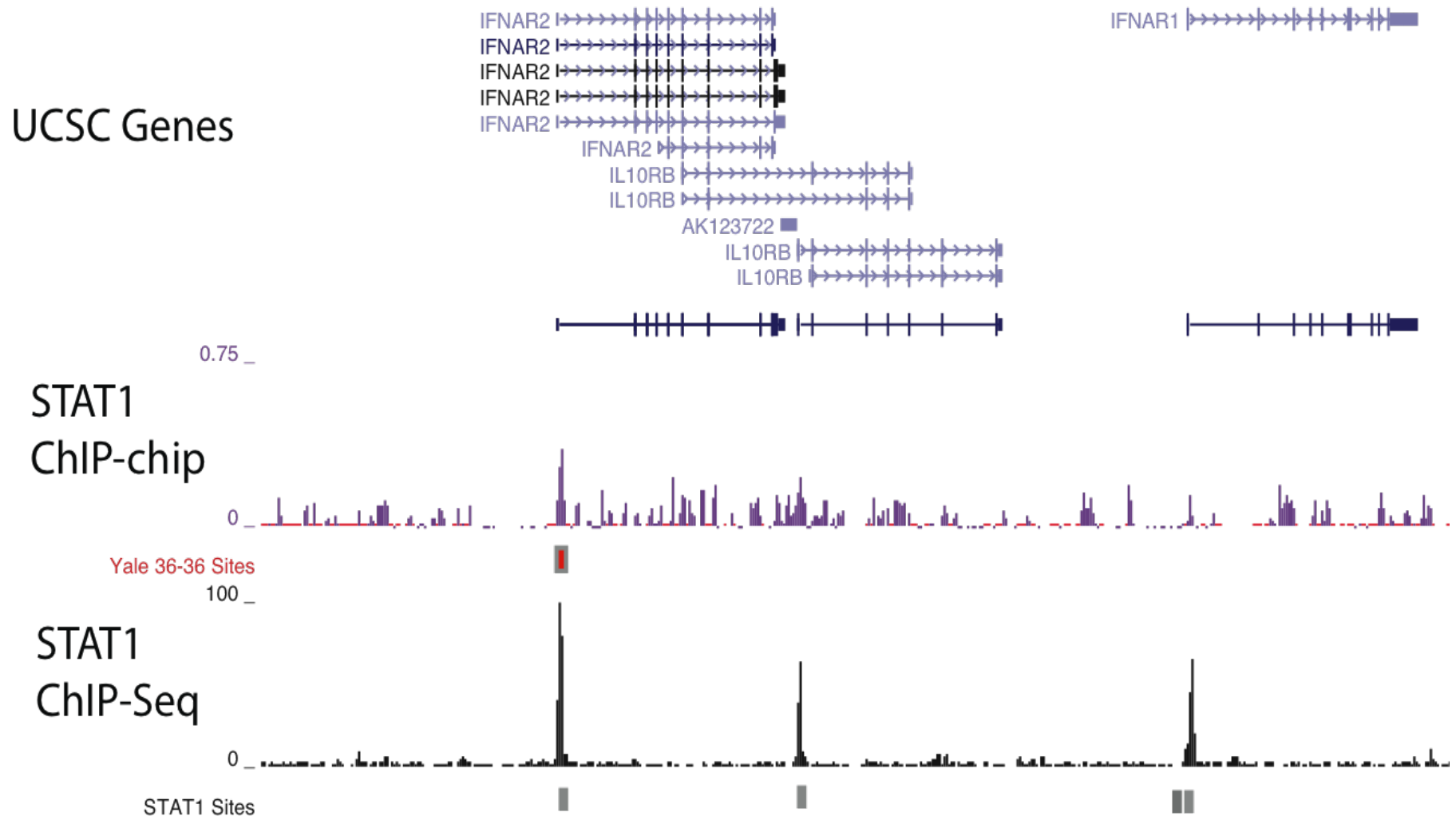
C

16 uniquely mapped sequence reads and their directional extension in a tag cluster



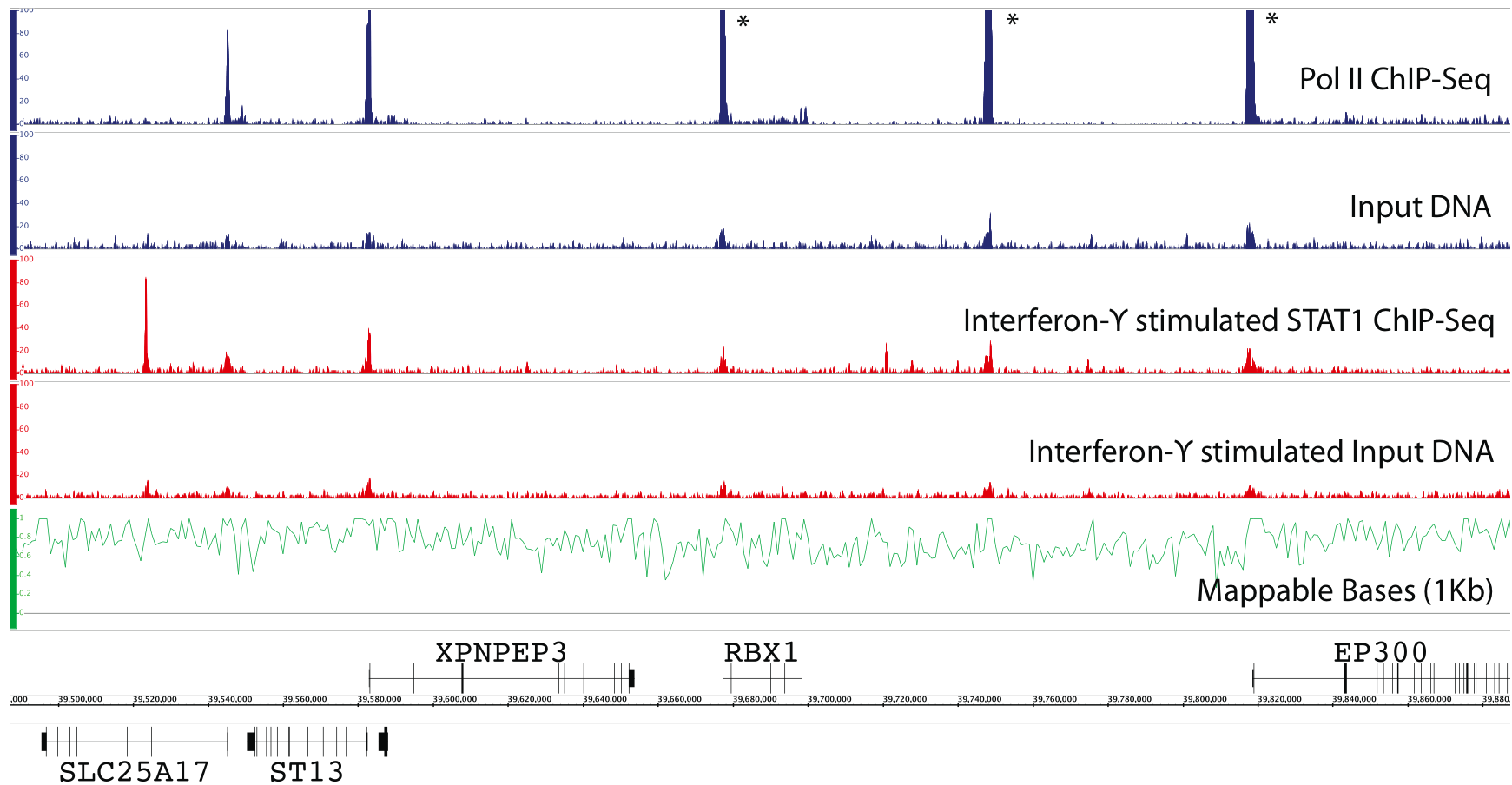
[Robertson et al., Nat. Meth. ('07); Zhang et al. PLOS Comp. Bio. (in revision, '08)]

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



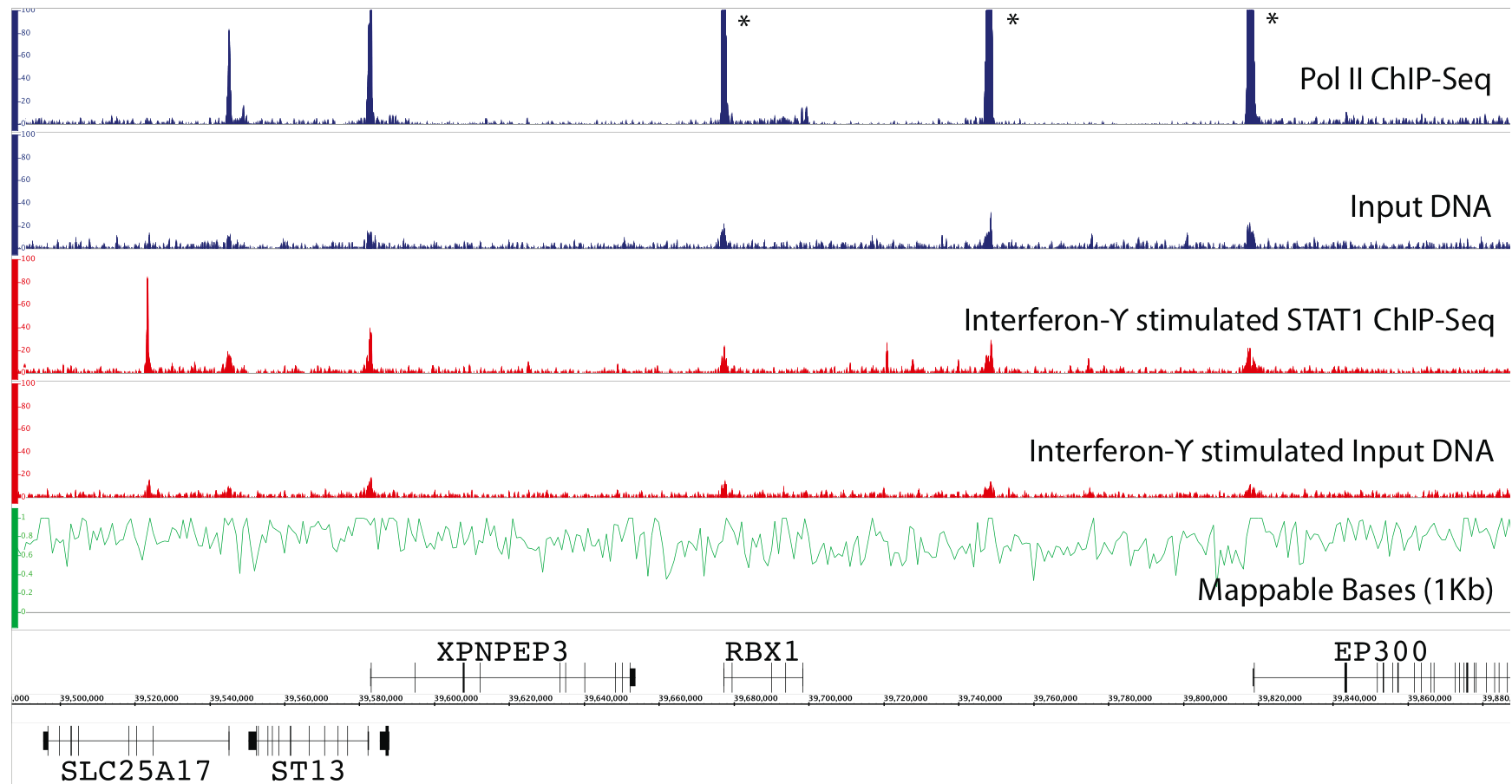
[Rozowsky et al. Nat. Biotech ('09)]

ChIP-Seq vs Input DNA Control



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

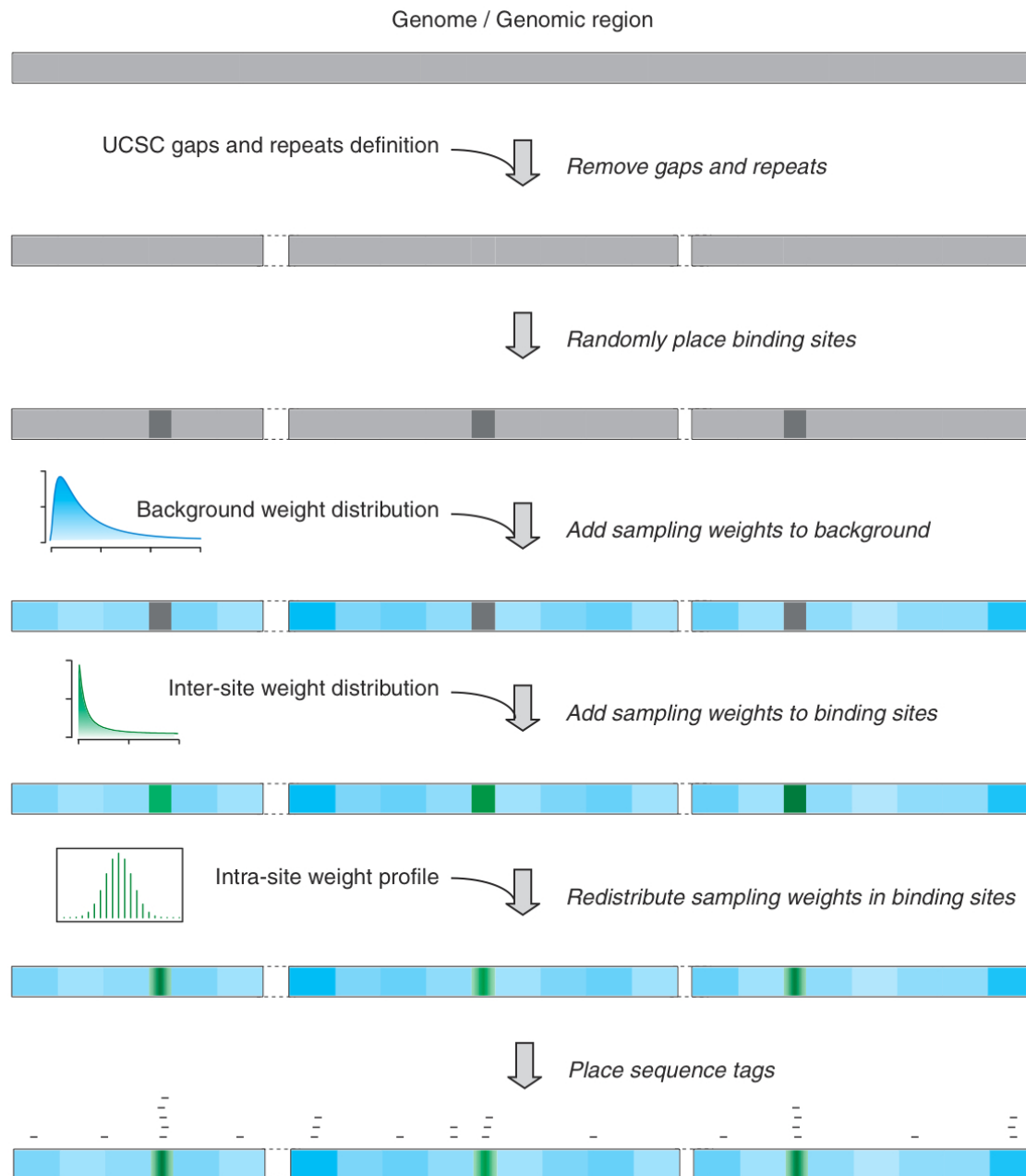
ChIP-Seq vs Input DNA Control



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

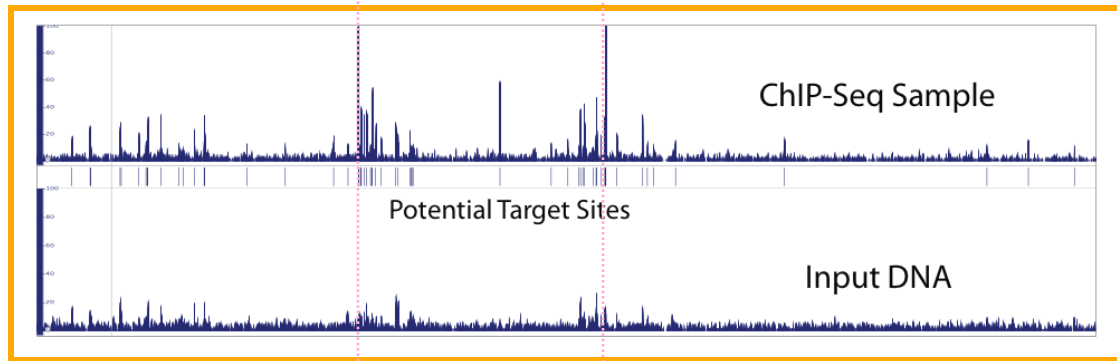
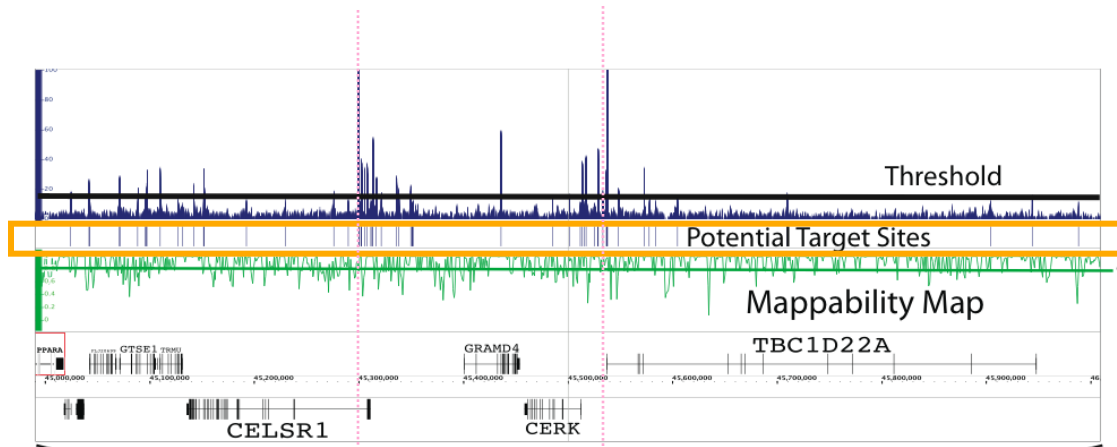
Correcting Chip-seq Signal by Simulating a Non- uniform Genomic Background

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



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

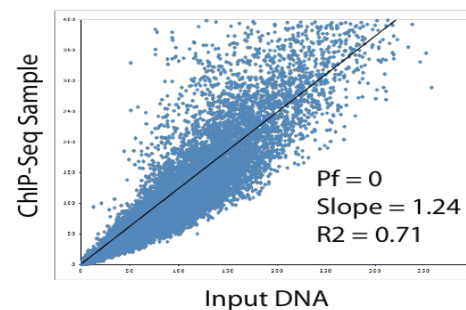
PeakSeq: Scoring Relative to Controls



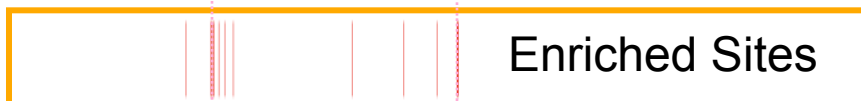
Filter for Potential
Targets based on
"Mappability"
Simulation

Scale Input
Relative to
ChIP

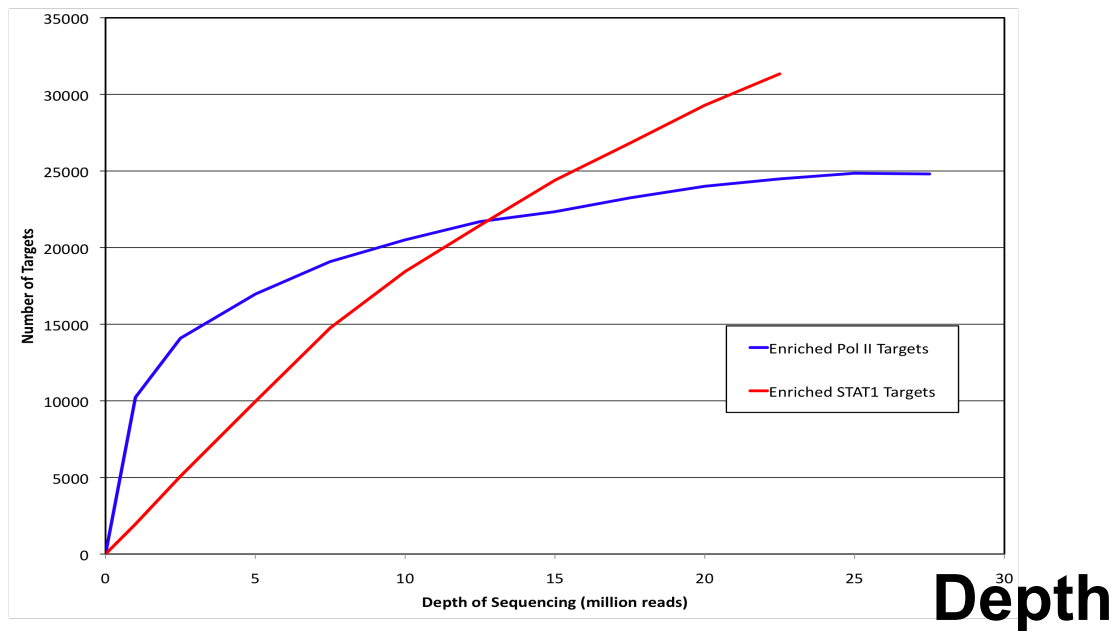
Score
Relative to
Bionomial
Expectation



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

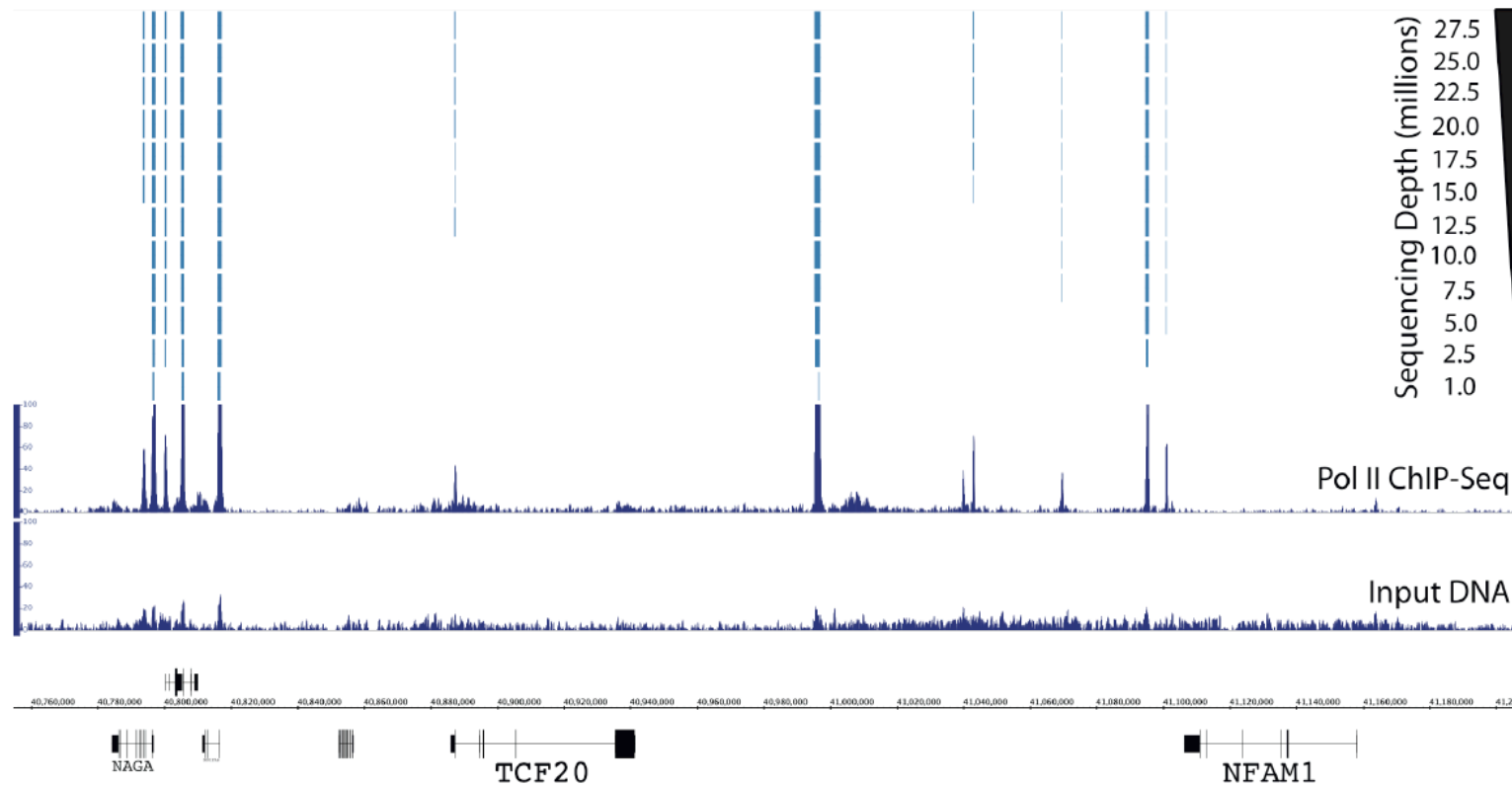


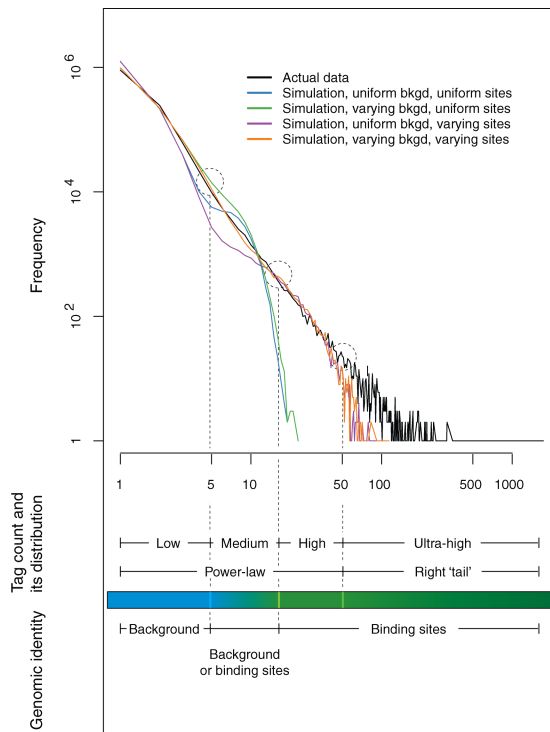
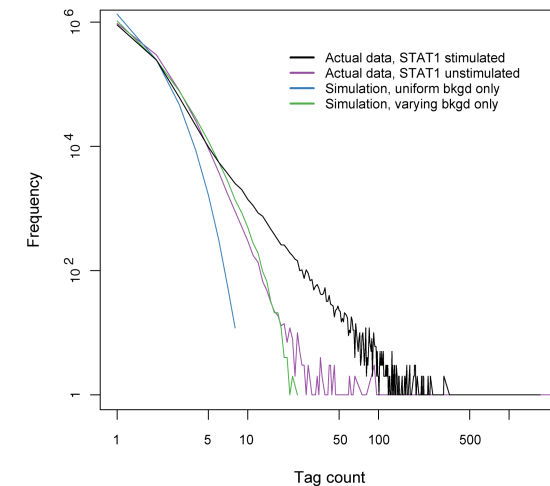
Binding Sites



Number of Reads for Saturation

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





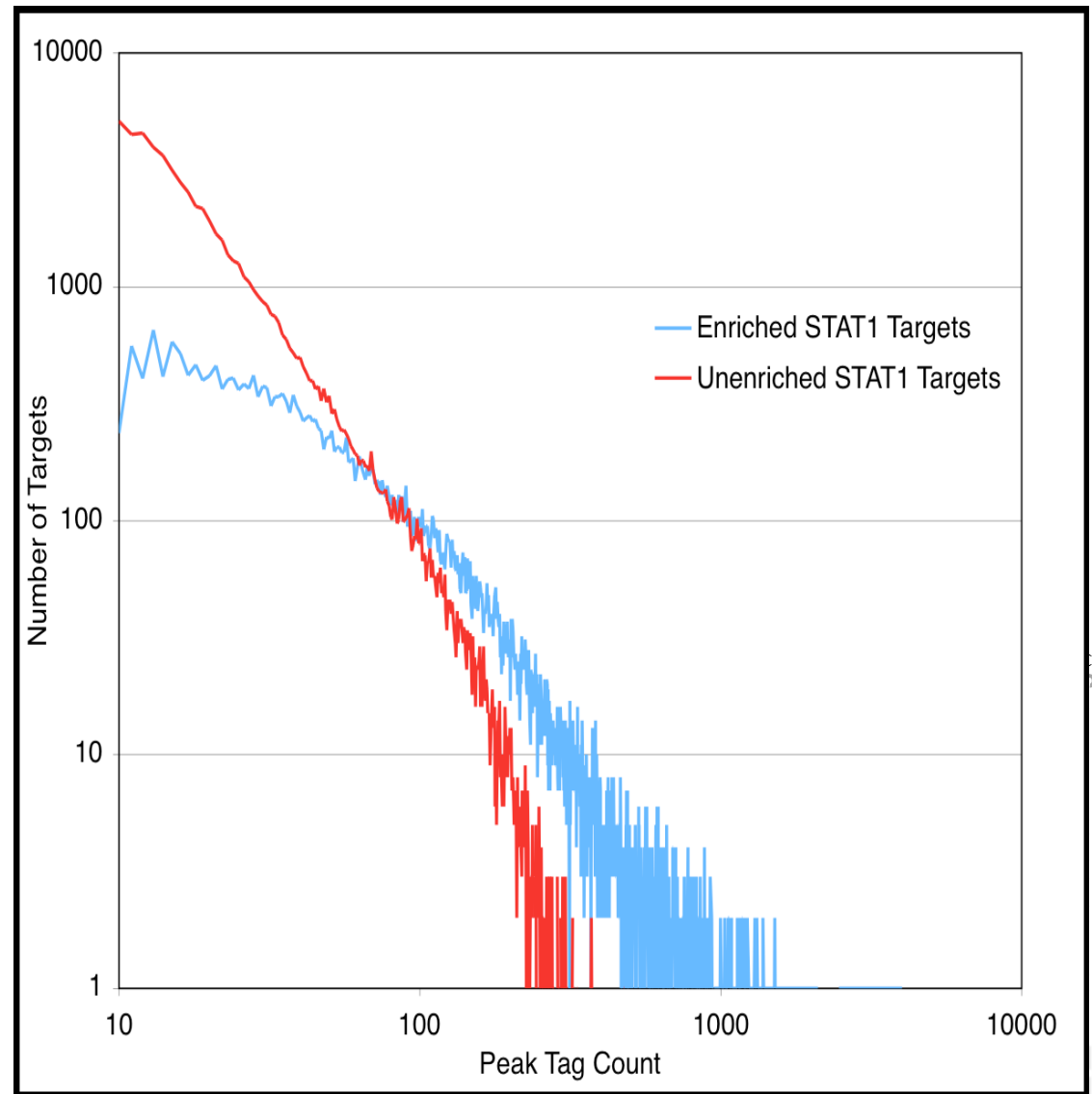
ChIP-sequencing simulation

- Contrary to the common belief, the background is mildly fluctuating and contains some 'hot' spots.
- Simple uniform background model does not count for all the variation in the background and thus leads to a serious underestimation of the background noise.
- Our study demonstrates that both the genomic background of ChIP and binding sites are not uniform.
- Simulated distributions segments the actual distribution into four sections.

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

Scored results
consistent with
simulation

Actual peaks at tail of
power-law graph

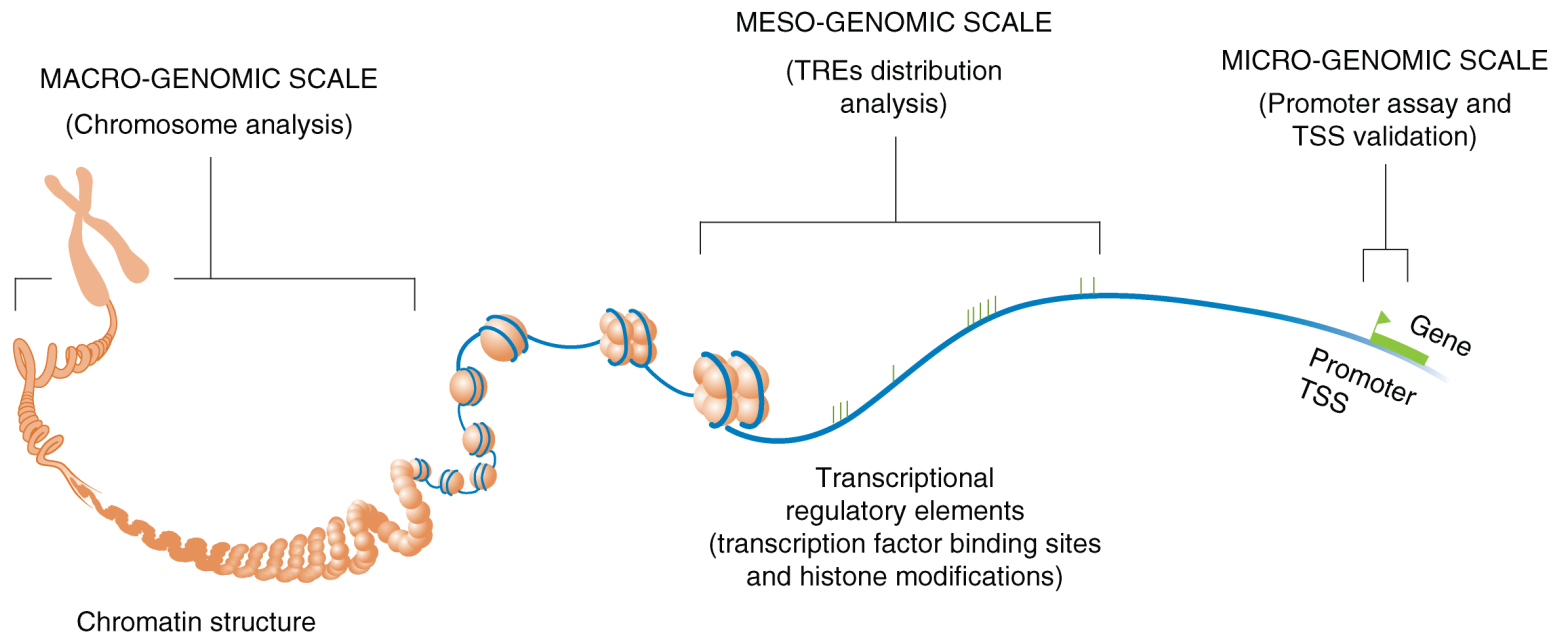


[Rozowsky et al. Nat. Biotech. ('09)]



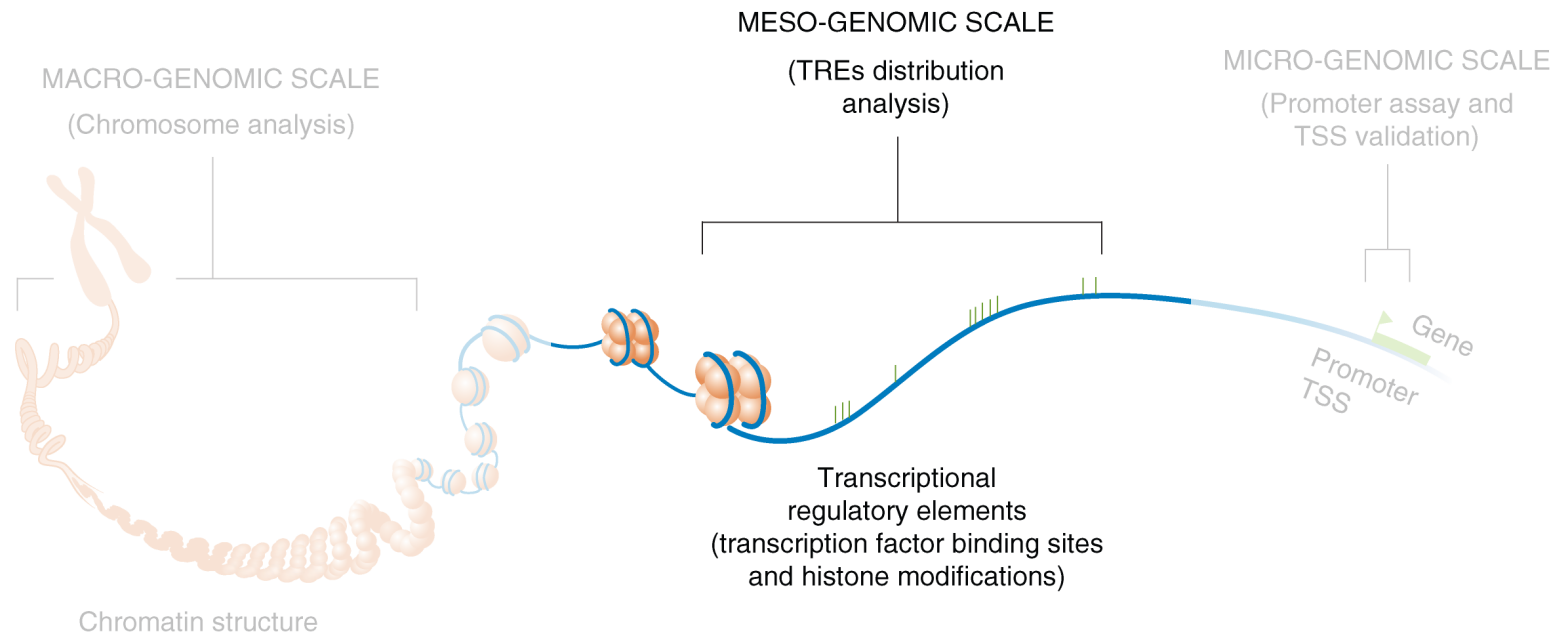
Annotating a single type of signal
on a large-scale:
Clustering and Characterizing
Binding Sites (TREs)

TRE analysis on the micro-genomic scale



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

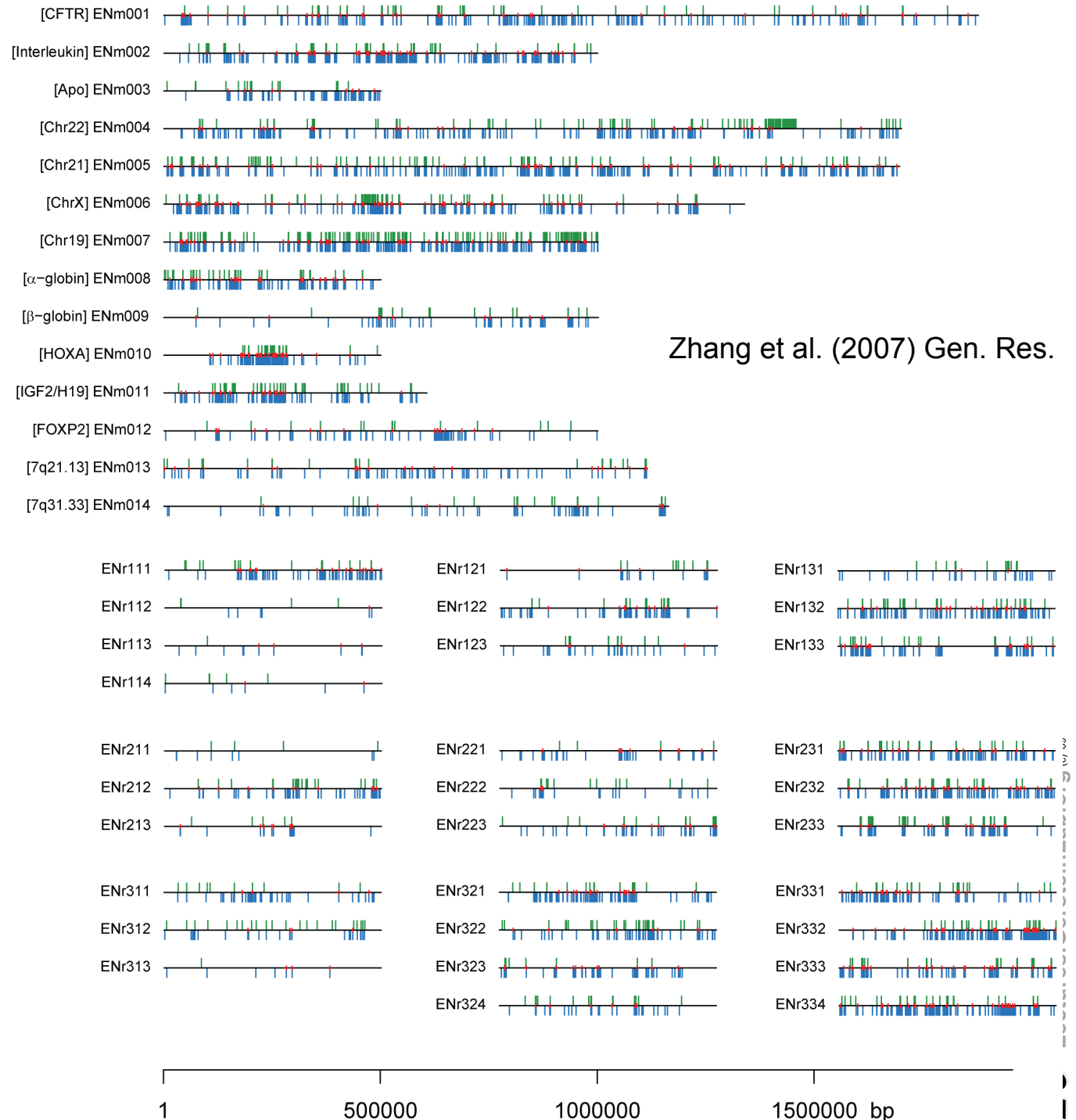
Clustering Binding Sites at ~50kb resolution



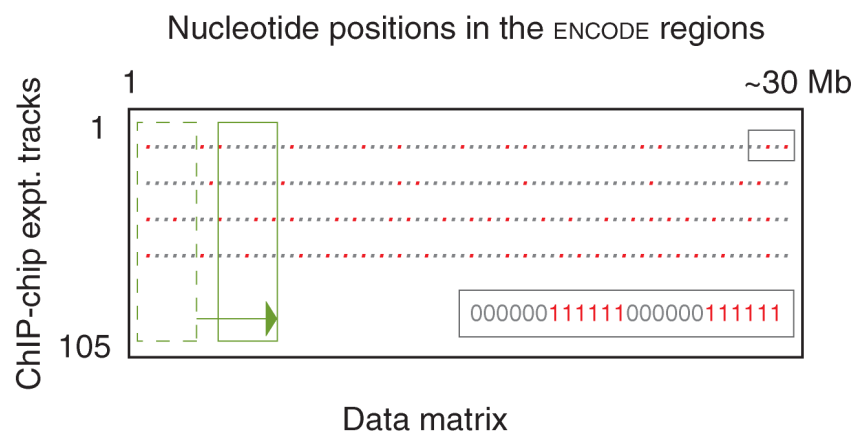
[Zhang et al. (2007) Gen. Res.]

Landscape of ENCODE Transcriptional Regulatory Elements

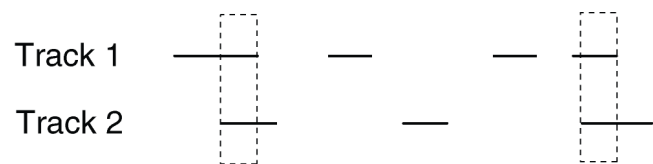
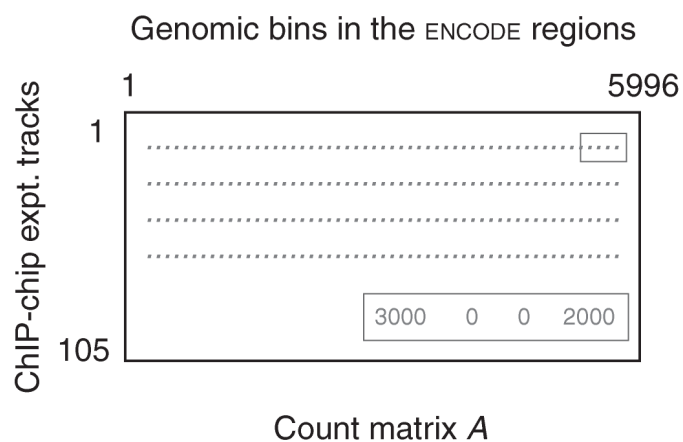
- Analyzed 105 lists of transcriptional regulatory elements in the encode regions
- 29 transcription factors, 9 cell lines, 2 time points
 - ◊ RNA Pol2
 - ◊ Histone modifications such as Ac & Me
 - ◊ Core promoters
 - ◊ Promoter proximal elements
 - ◊ Others such as enhancers, silencers, insulators,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



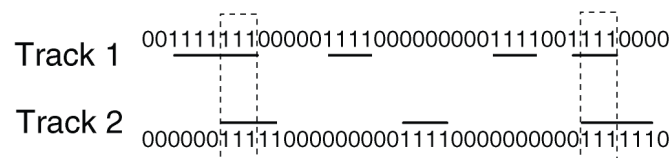
Collect Total Hits for Each Factor in ~6000 Bins of 10 to 100 kb and Compare to Random Control



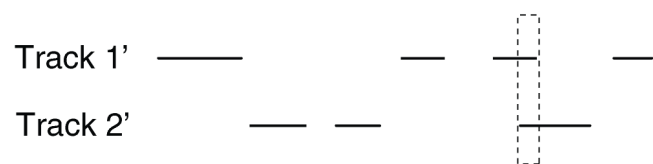
Sliding-window transformation



Binary coding



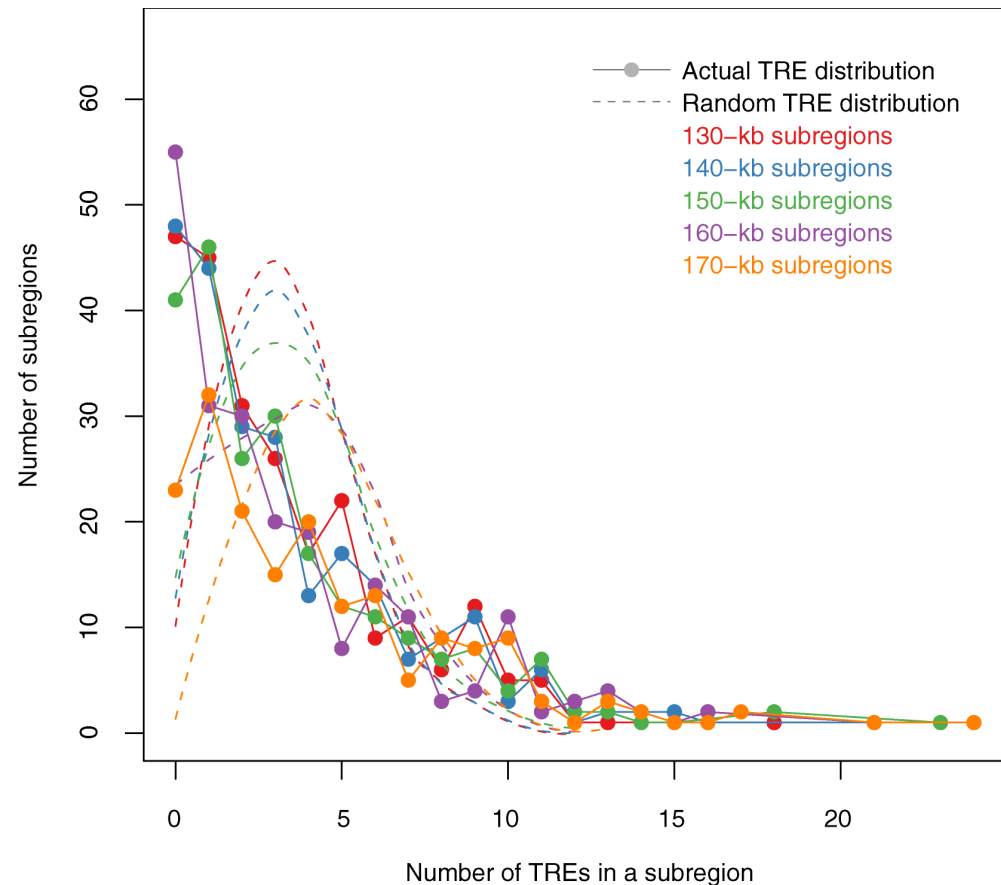
Randomization



Zhang et al. (2007) Gen. Res.

Non-random distribution of TREs

- TREs are not evenly distributed throughout the encode regions ($P < 2.2 \times 10^{-16}$).
- The actual TRE distribution is power-law.
- The null distribution is 'Poissonesque.'
- Many genomic subregions with extreme numbers of TREs.



Biplot to Show Overall Relationship of TFs and Genomic Bins

TFs: a, b, c...

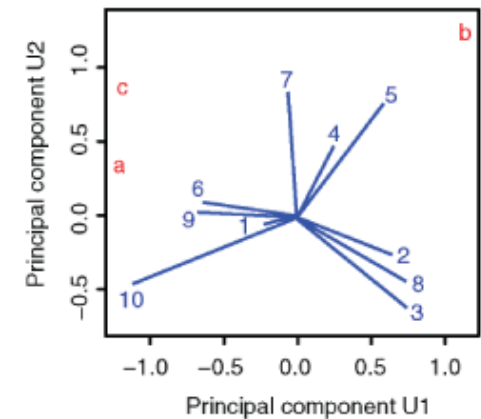
50kb Genomic Bins: 1,2,3...

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

$$A=USV^T$$

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

$$AA^T$$

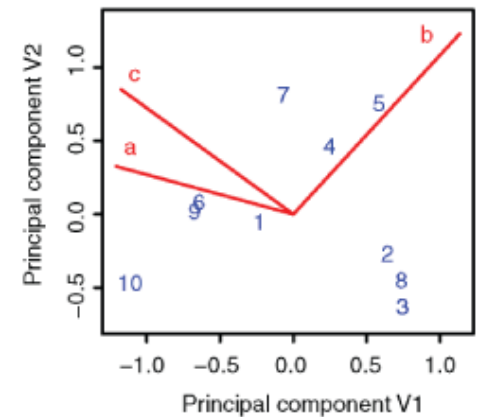


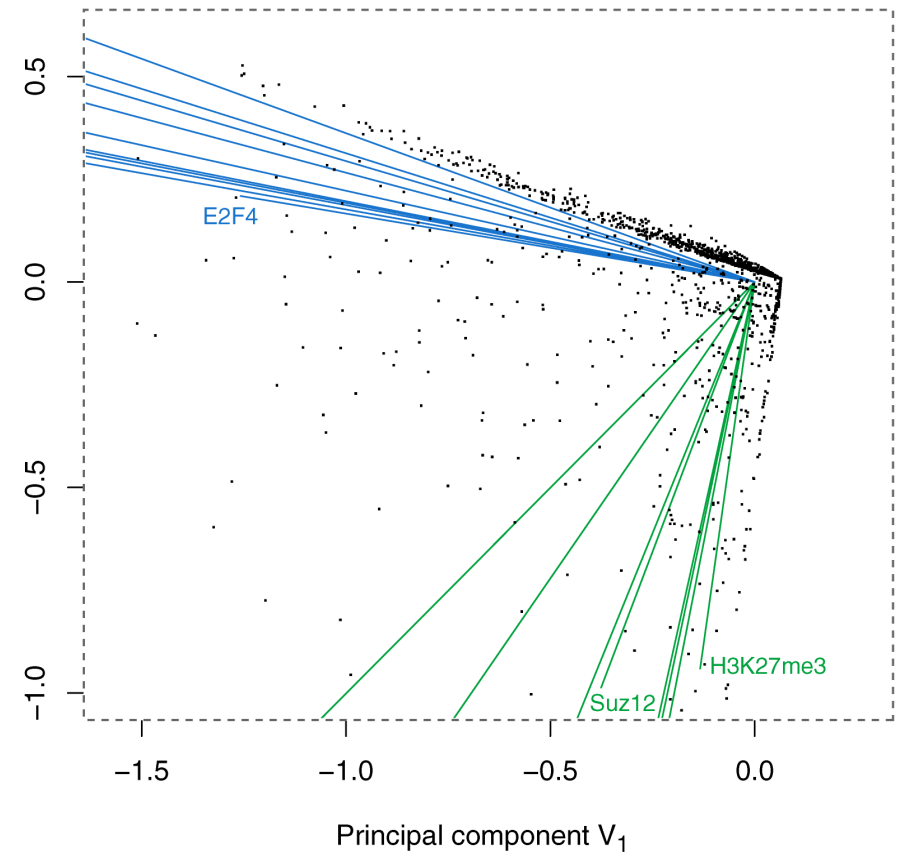
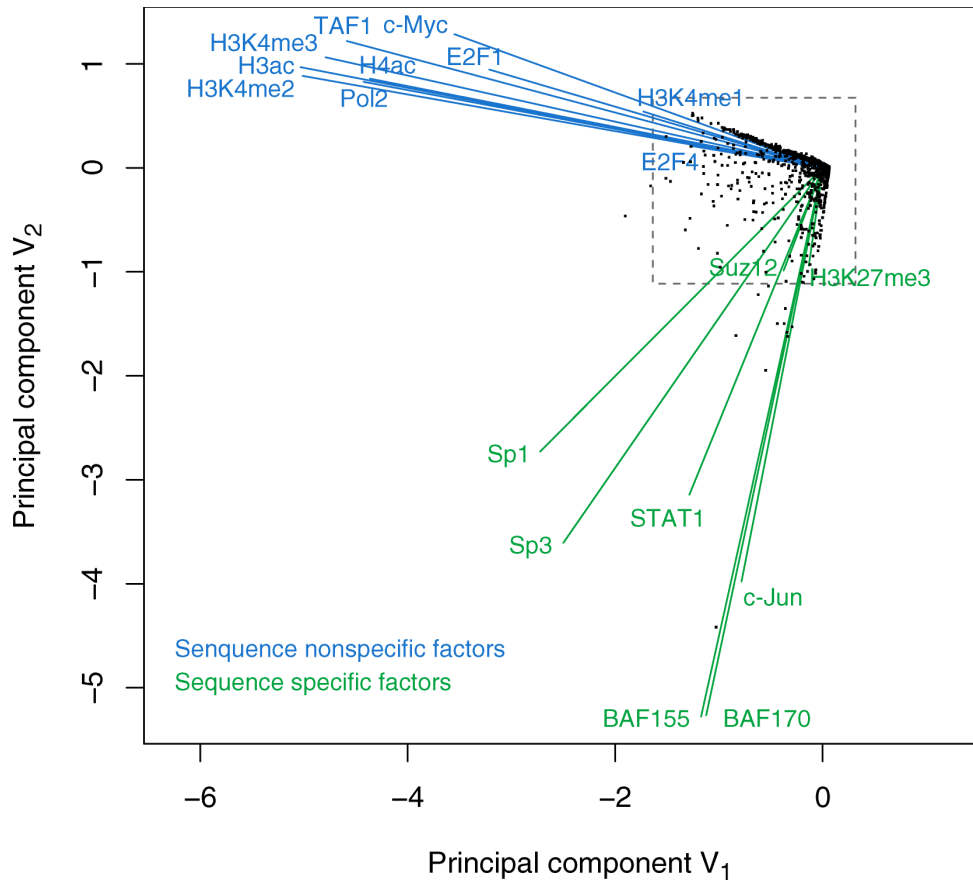
$$A^T$$

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

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

$$A^T A$$





Results of Biplot

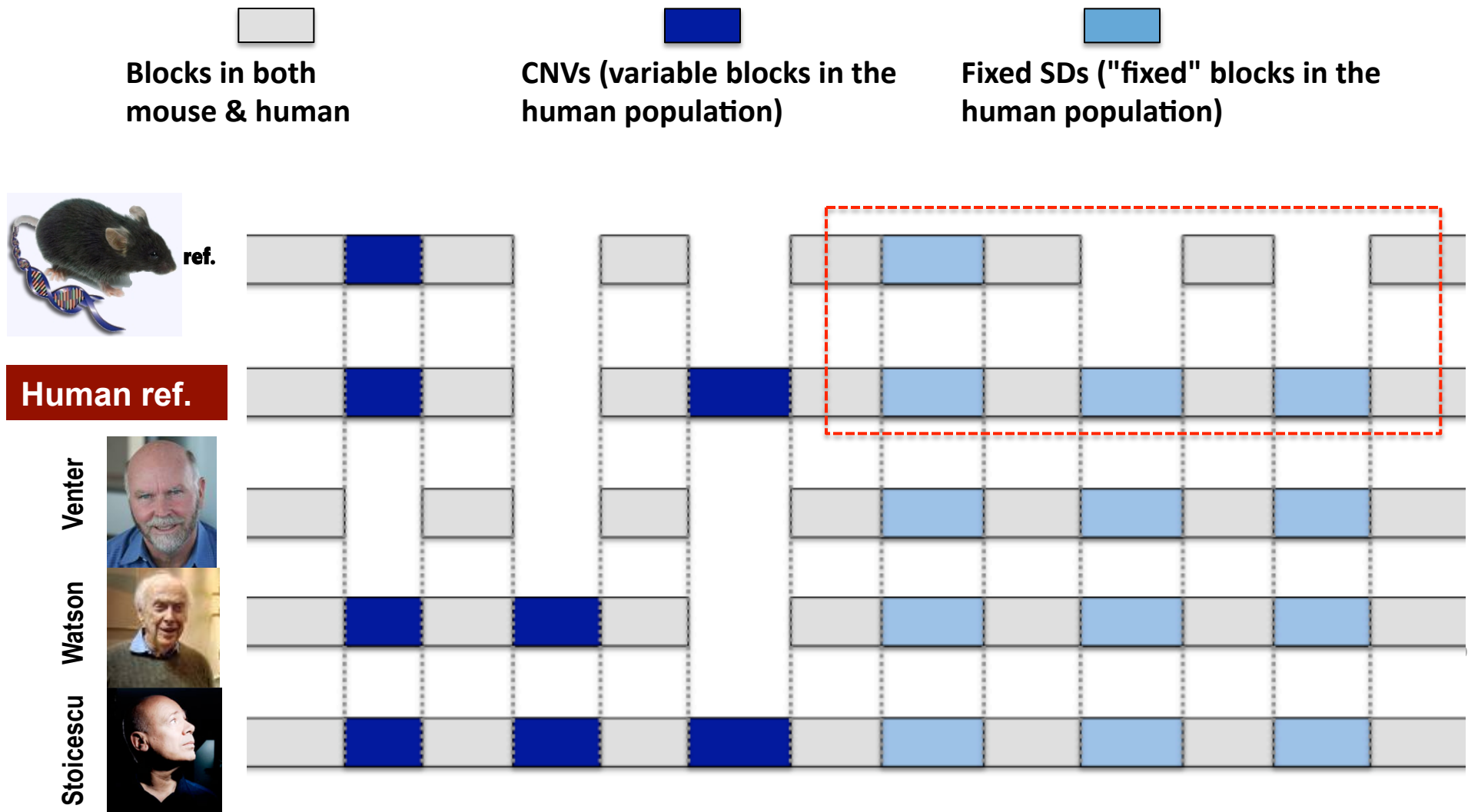
- Biplot groups TFs into sequence-specific and sequence-nonspecific clusters.
 - ◇ c-Myc may behave more like a sequence-nonspecific TF.
 - ◇ H3K27me3 functions in a transcriptional regulatory process in a rather sequence-specific manner.
- Genomic Bins are associated with different TFs and in this fashion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

Zhang et al. (2007)
Gen. Res.

Signal Processing 2: Finding Variable Blocks in the Human Genome

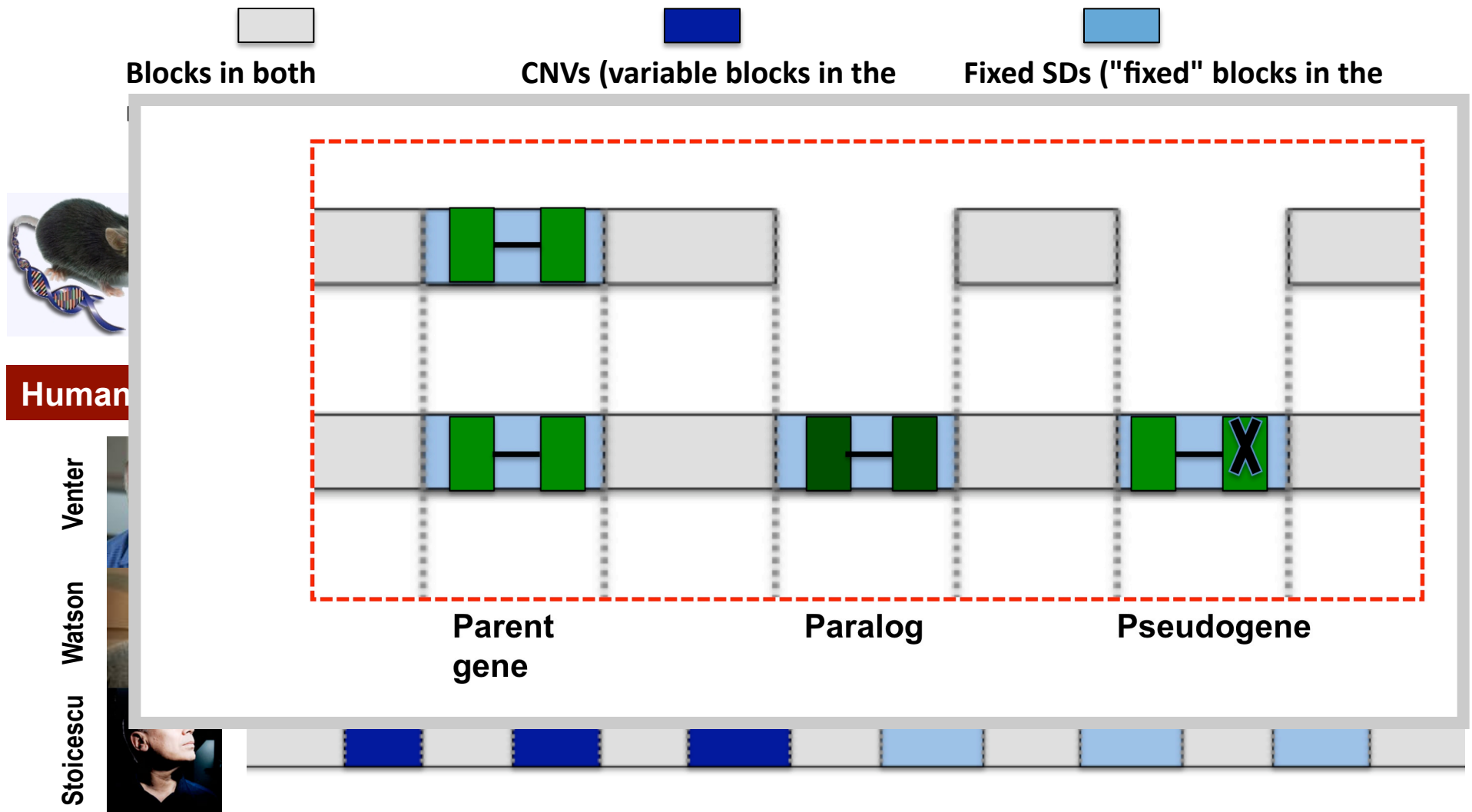


Terminology for Variable Duplicated Elements in the Human Genome



Segmental duplications (SDs) - Recent duplications (~40 million years and younger)

Terminology for Variable Duplicated Elements in the Human Genome



Segmental duplications (SDs) - Contain Duplicated Paralogs and Duplicated Pseudogenes

Detection of Block Variation in Personal Genomics

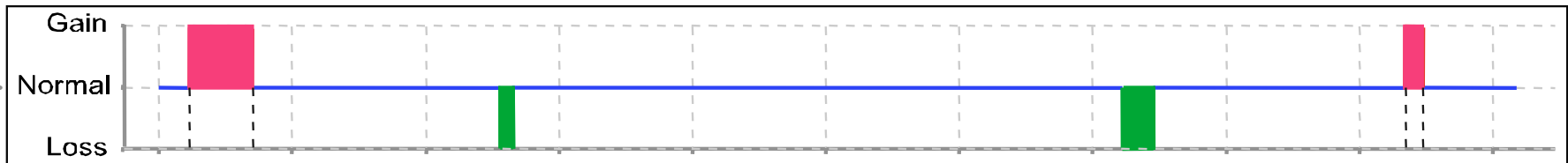
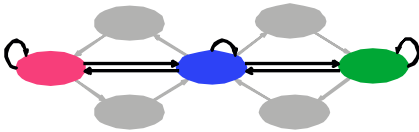
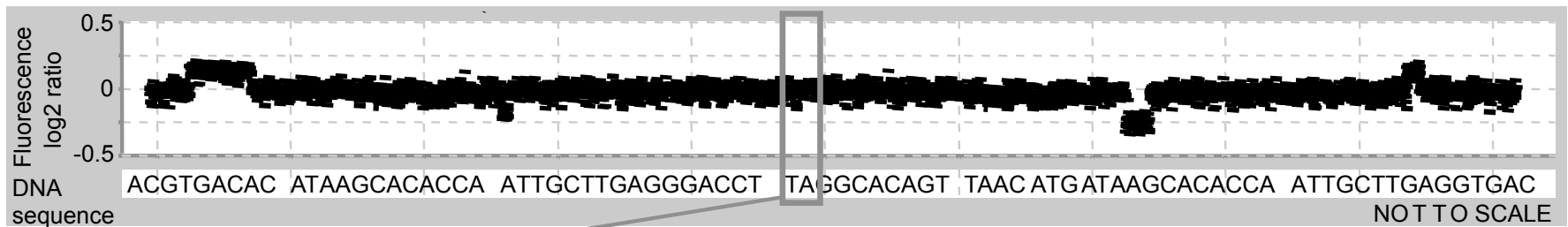
- Main steps in Human Genome Resequencing
 - ◇ SNP detection
 - ◇ Haplotype phasing
 - ◇ Determining small indels
 - ◇ **Reconstructing Large Structural Variants (most challenging)**
- Different Techniques for SV Reconstruction
 - ◇ Segmenting Arrays and Sequencing Read-depth
 - ◇ Discordantly placed paired-ends
 - ◇ Finding split reads
 - ◇ Doing small scale reassembly in presence of repeats

Segmentation of Read Depth or Array Signal

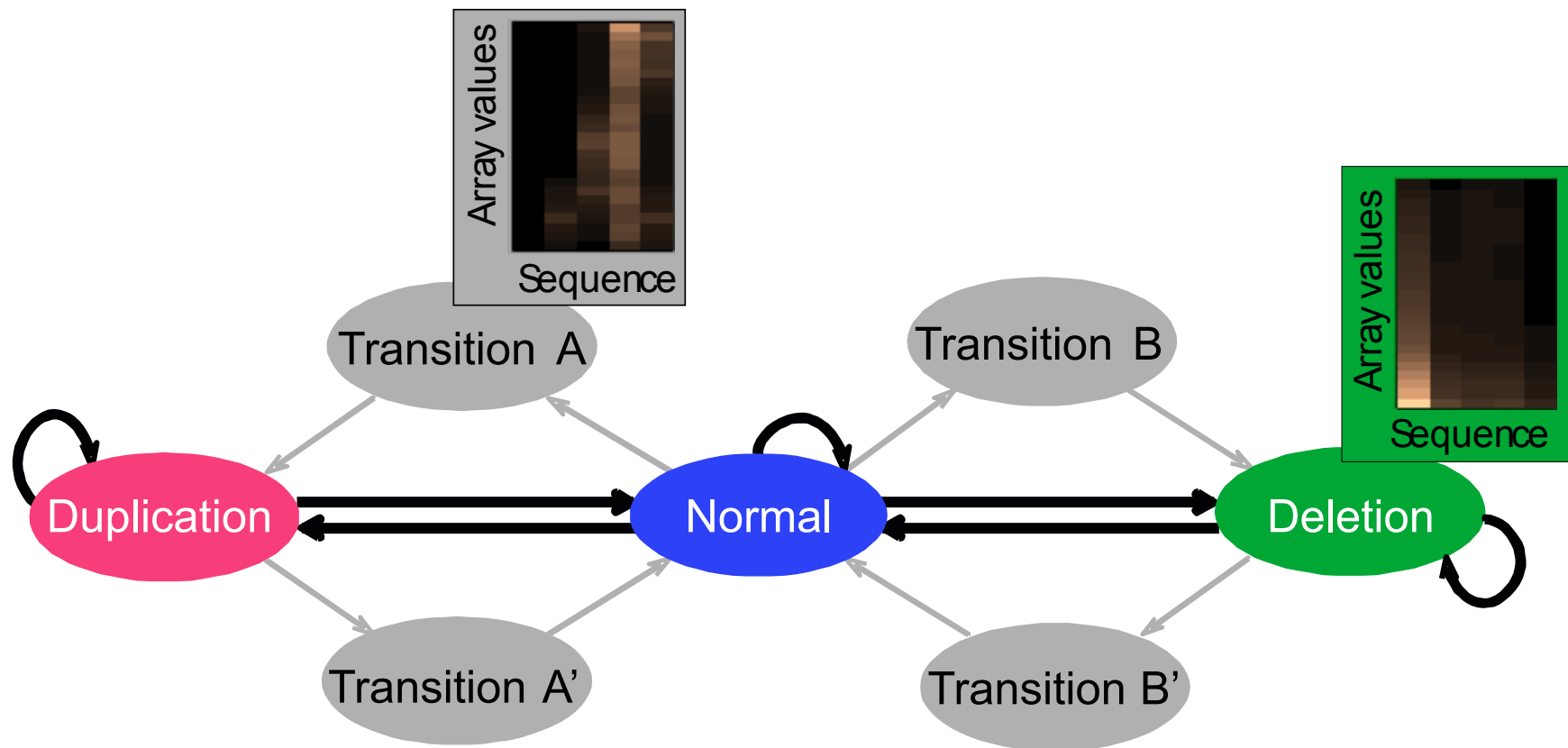


BreakPtr HMM

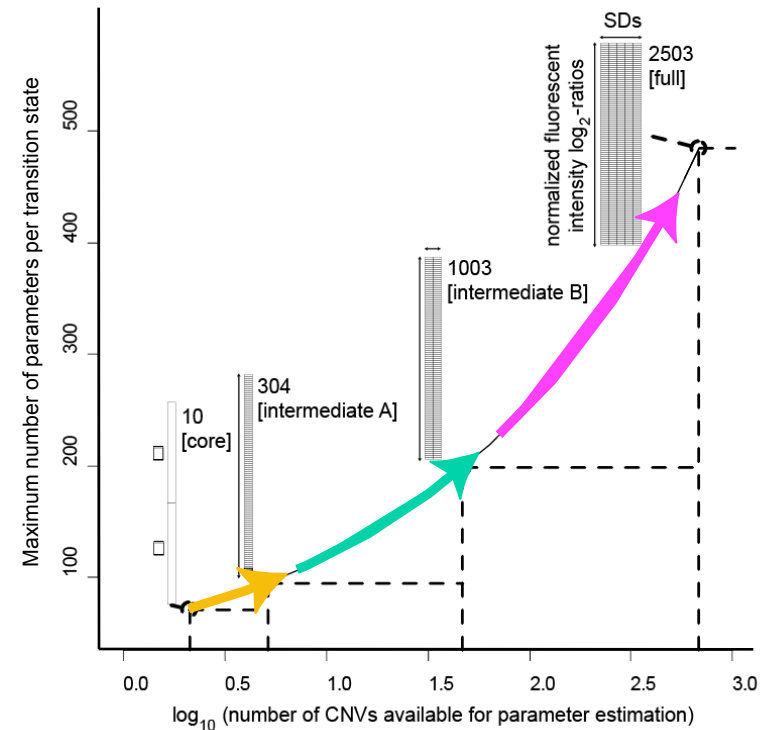
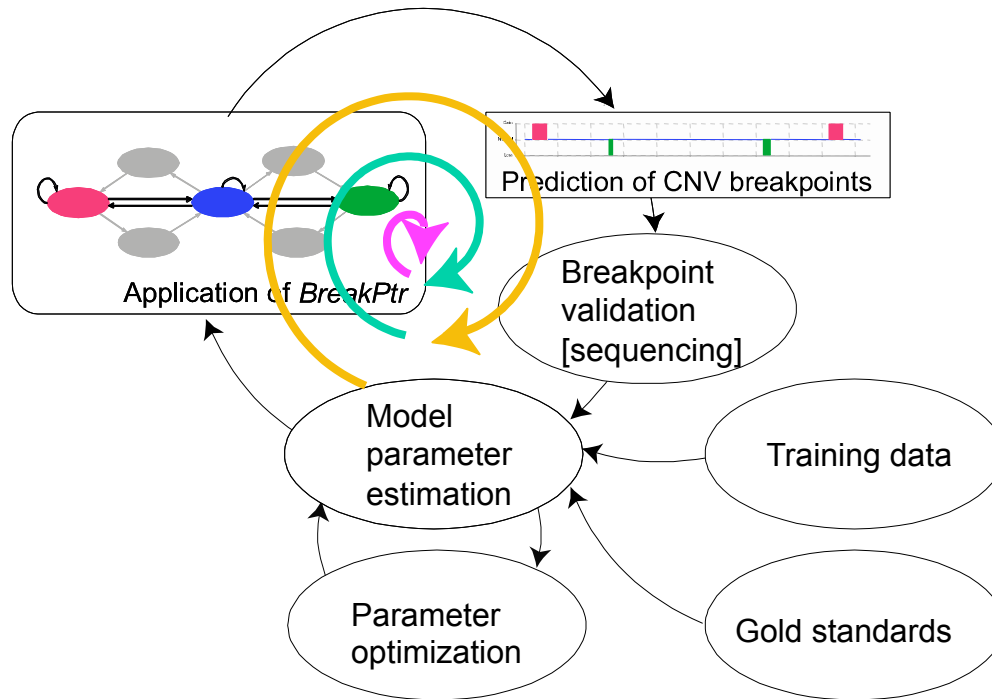
- To get highest resolution on breakpoints need to smooth & segment the signal
- BreakPtr: prediction of breakpoints, dosage and cross-hybridization using a system based on Hidden Markov Models



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



'Active' approach for breakpoint identification: initial scoring with preliminary model, targeted validation (with sequencing), retraining, and rescoreing



CNV breakpoints sequenced in ~10 cases following BreakPtr analysis;

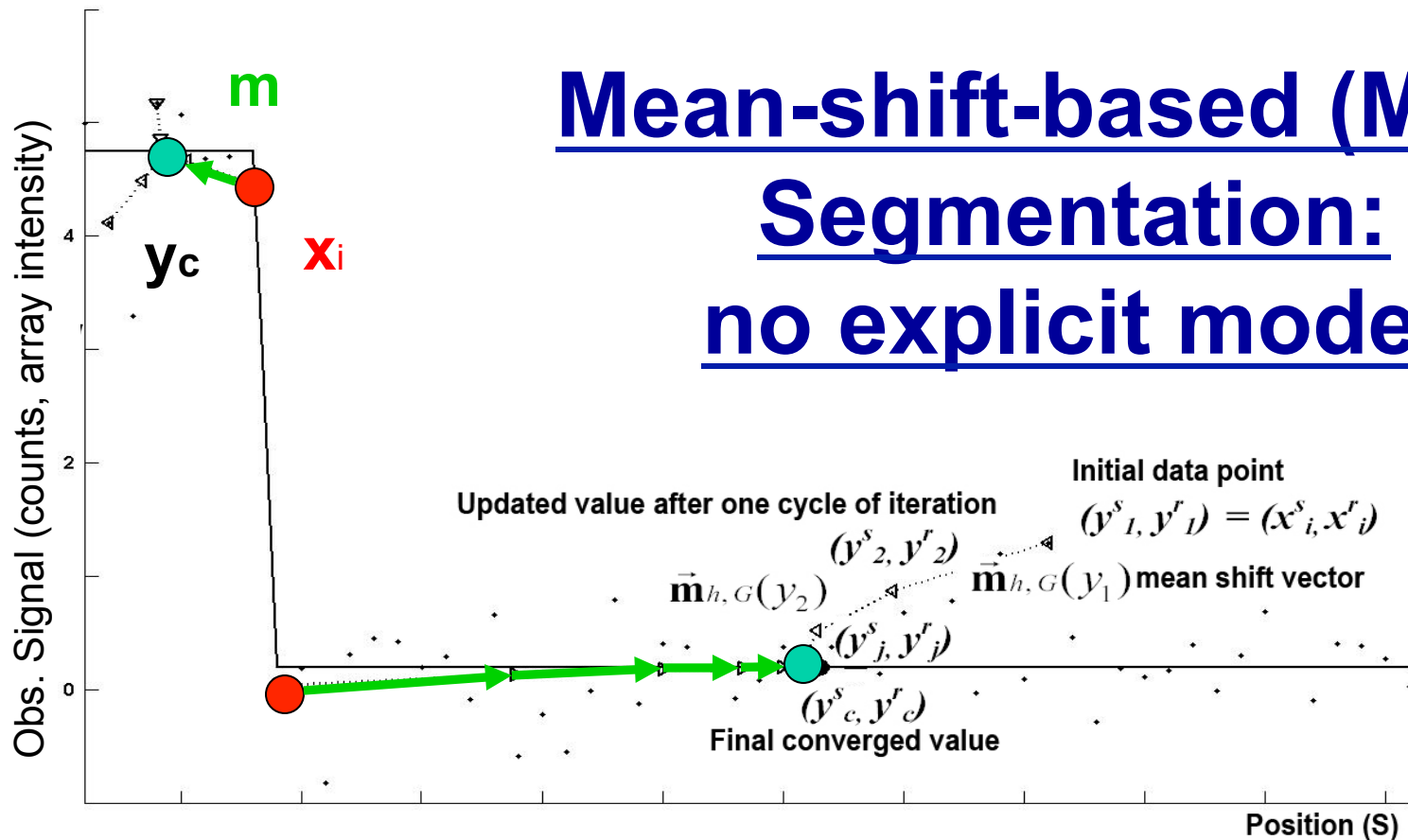
Median resolution <300 bp

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

HMM optimized iteratively
(using Expectation Maximization, EM)

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

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



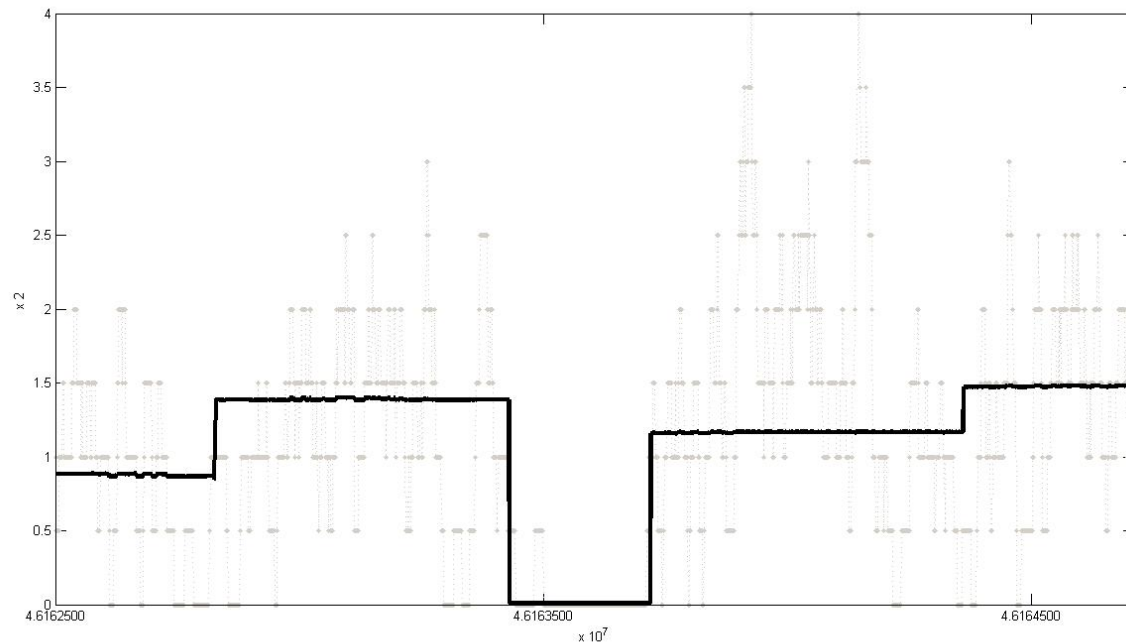
- (x_i) Observed depth of coverage counts (or array signal) as samples from PDF
- (m) Kernel-based approach to estimate local gradient of PDF
- (y_c) Iteratively follow grad to determine local modes

Not Model-based (e.g. like HMM)

with global optimization, distr. assumption & parms. (e.g. num. of segments).

Achieves discontinuity-preserving smoothing

Representative Result Showing Segmentation Based on Depth of Coverage



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

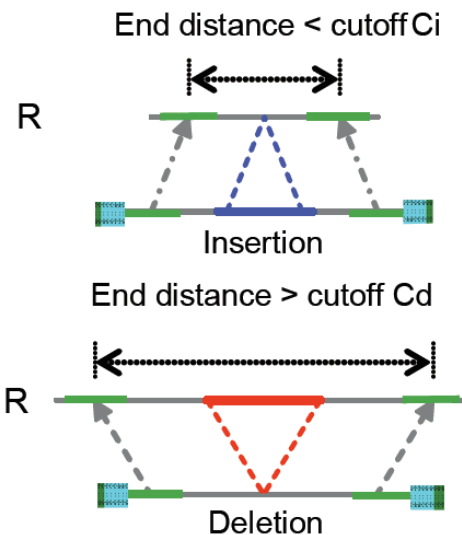
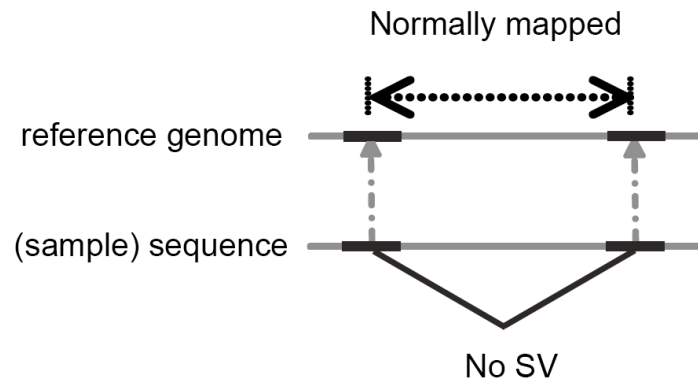
NA11995 (seq. by Sanger, MAQ mapping)
chr 21 (46162500 to 46164711)

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

Looking for Aberrantly Placed Paired Ends

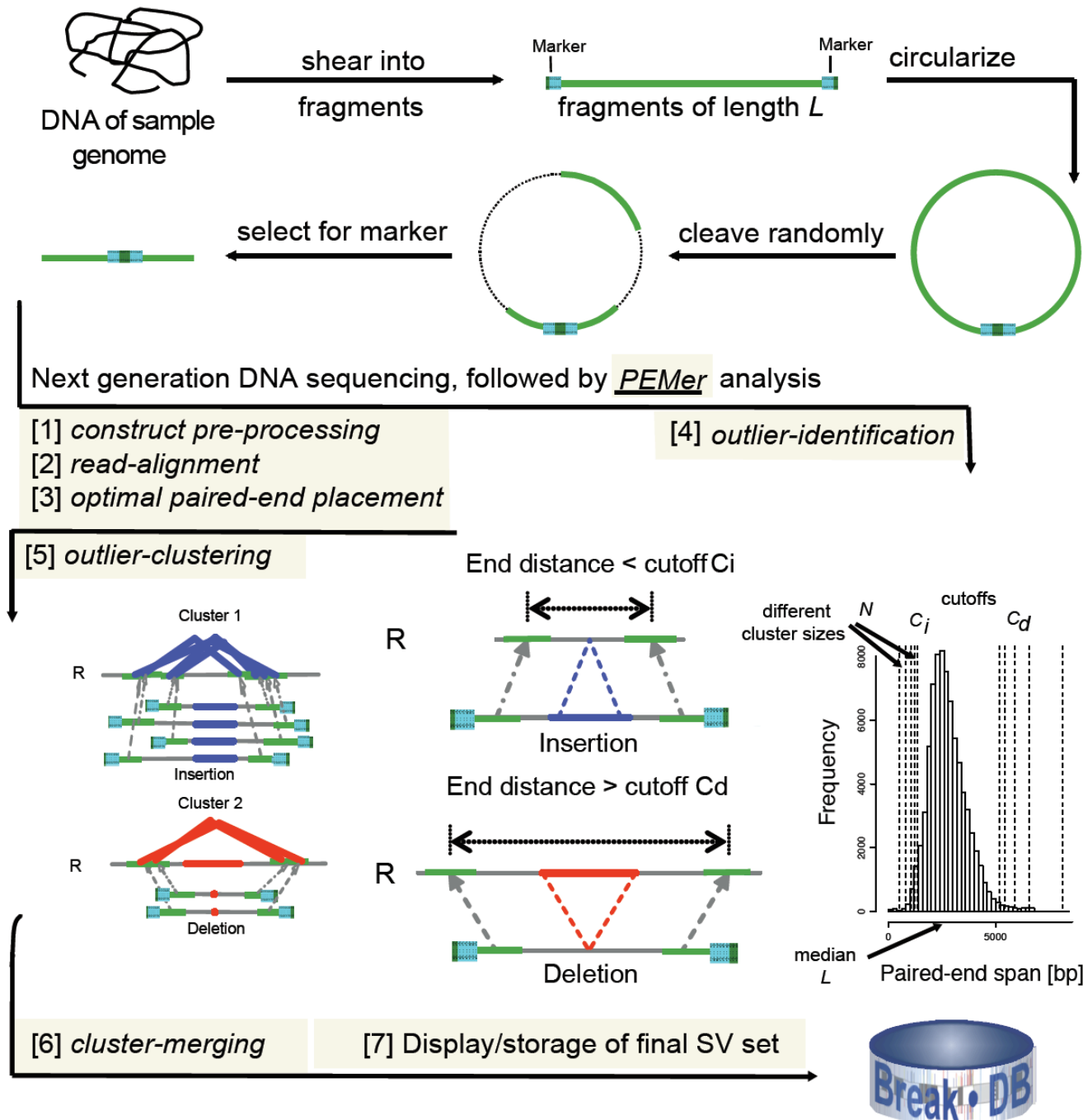


PEMer: Detecting Structural Variants from Discordant Paired Ends in Massive Sequencing



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

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



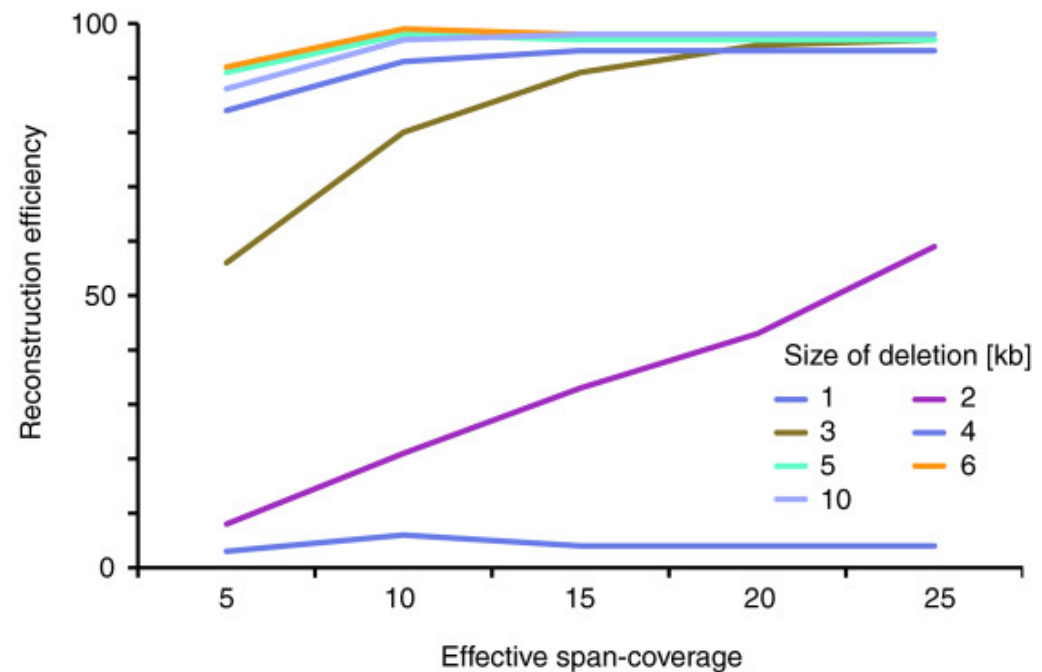
[Korbel et al.,
Science ('07);
Korbel et al.,
GenomeBiol. ('09)]

Parameterize Error Models through Simulation

Reconstruction efficiency at different coverage

[Korbel et al.,
GenomeBiol.
(‘09)]

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



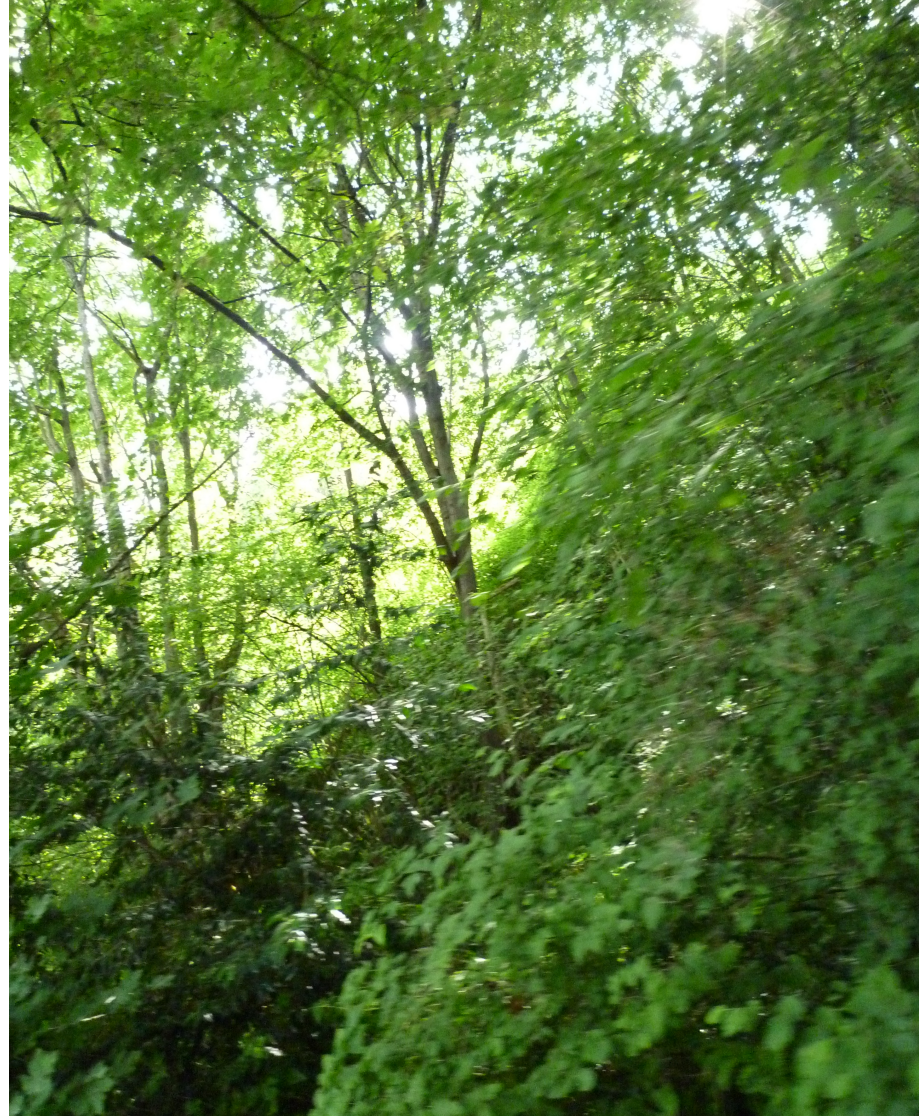
Reconstruction of heterozygous insertions

5x coverage by 2.5 kb inserts		5x coverage by 10 kb inserts	
Insertion size	Reconstruction efficiency	Insertion size	Reconstruction efficiency
250	0	1000	8
500	1	2000	42
750	2	3000	72
1000	1	4000	69
1250	8	5000	61
1500	3	6000	55
1750	3	7000	37
2000	1	8000	23
2250	1	9000	4
2500	0	10000	1
2750	0		
3000	0		
False positives	4		4

Better coverage and fewer reads allow to relax cutoff on outlier lengths and reconstruct more insertions

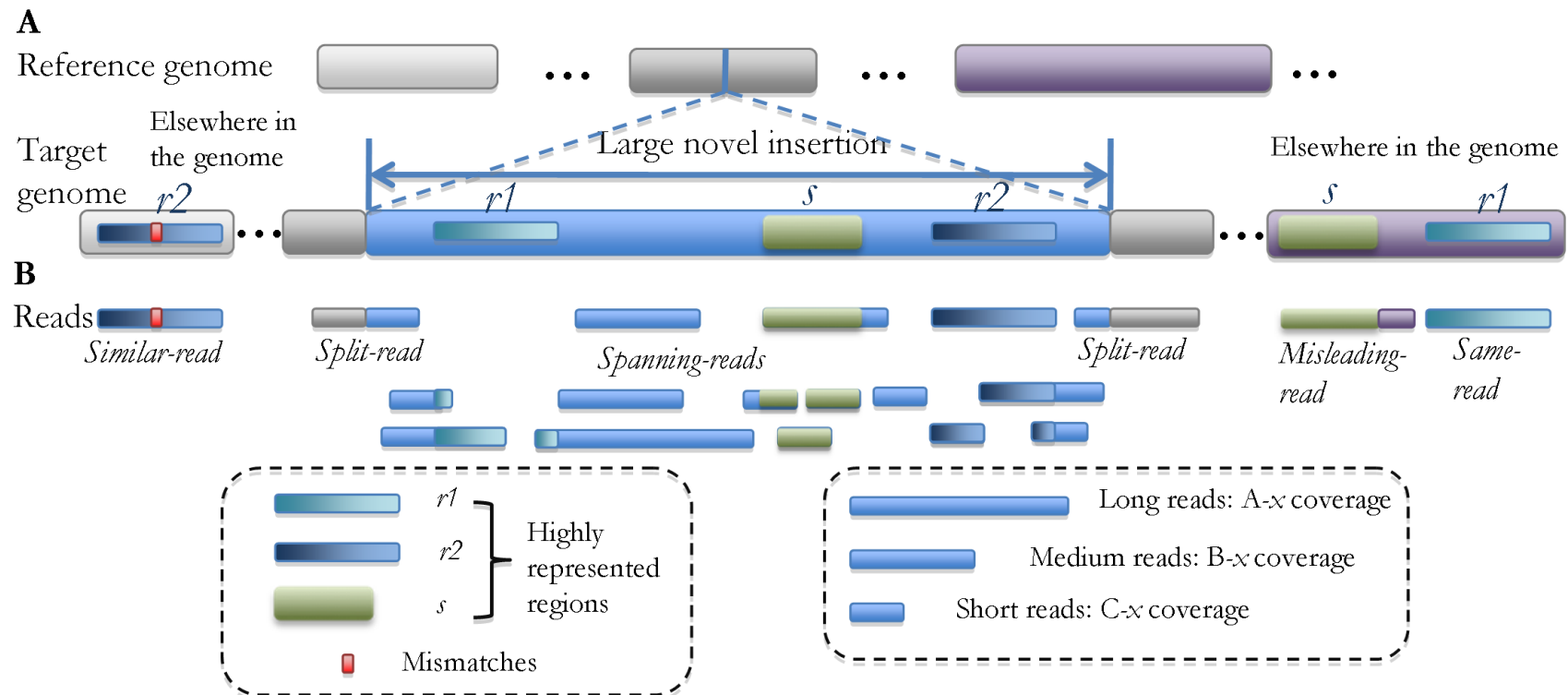
[Korbel et al., GenomeBiol. ('09)]

Local Reassembly



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.

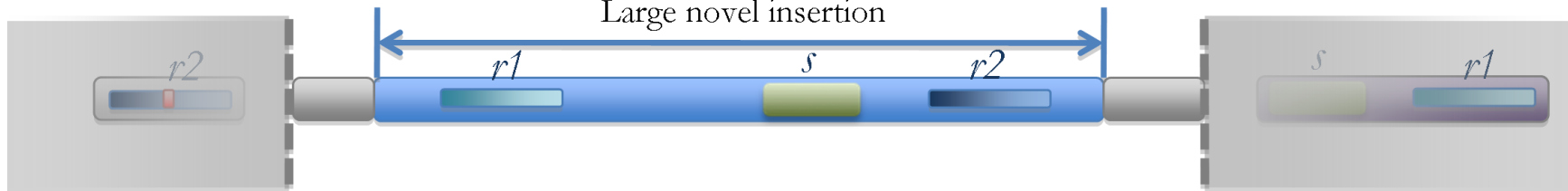


Optimal integration of sequencing technologies: *Need Efficient Simulation*

Different combinations of technologies (i.e. read lengths) very expensive to actually test.
Also computationally expensive to simulate.

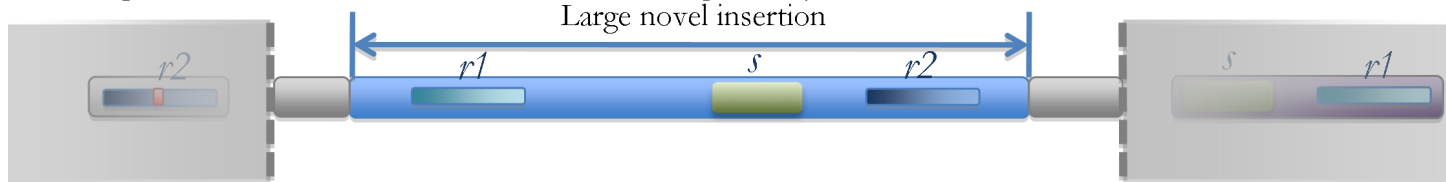
(Each round of whole-genome assembly takes >100 CPU hrs; thus, simulation exploring 1K possibilities takes 100K CPU hr)

C Simplification of the simulation to the insertion region only
Large novel insertion

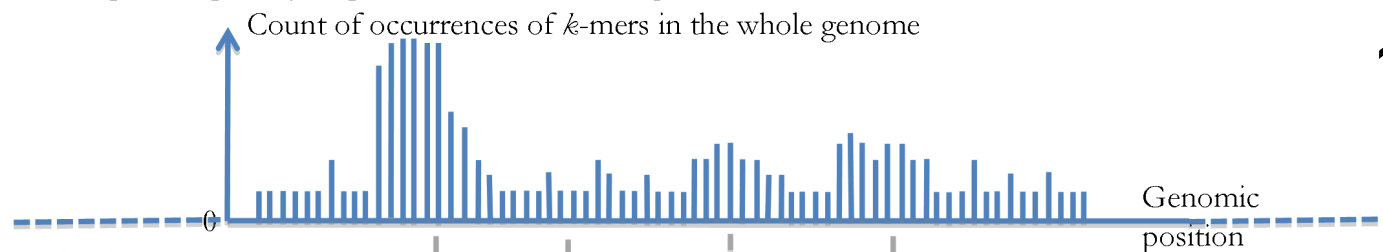


Optimal integration of sequencing technologies: *Efficient Simulation Toolbox using Mappability Maps*

C Simplification of the simulation to the insertion region only

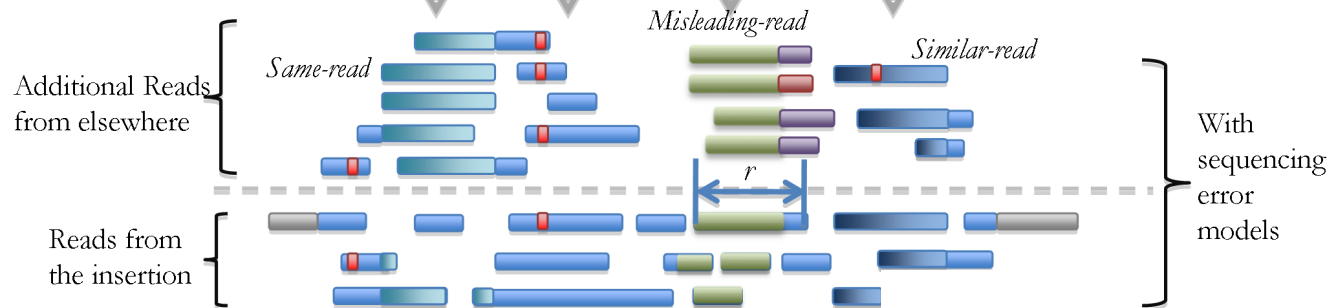


D Compute mappability maps to scale to the whole genome

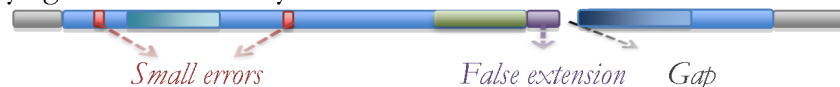


**~100,000X
speedup**

E Simulate the reads



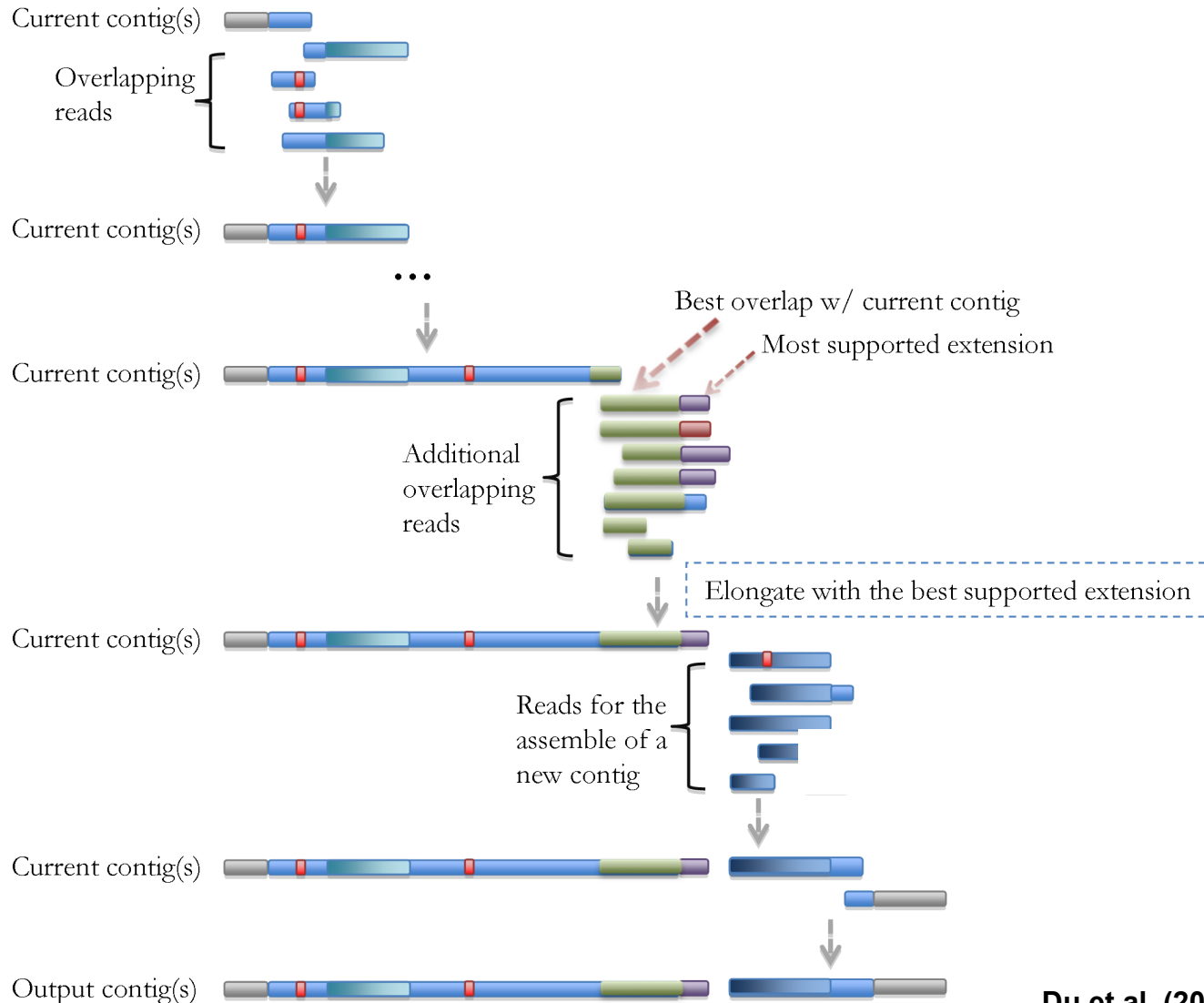
F Output after applying de novo assembly to reads from E



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

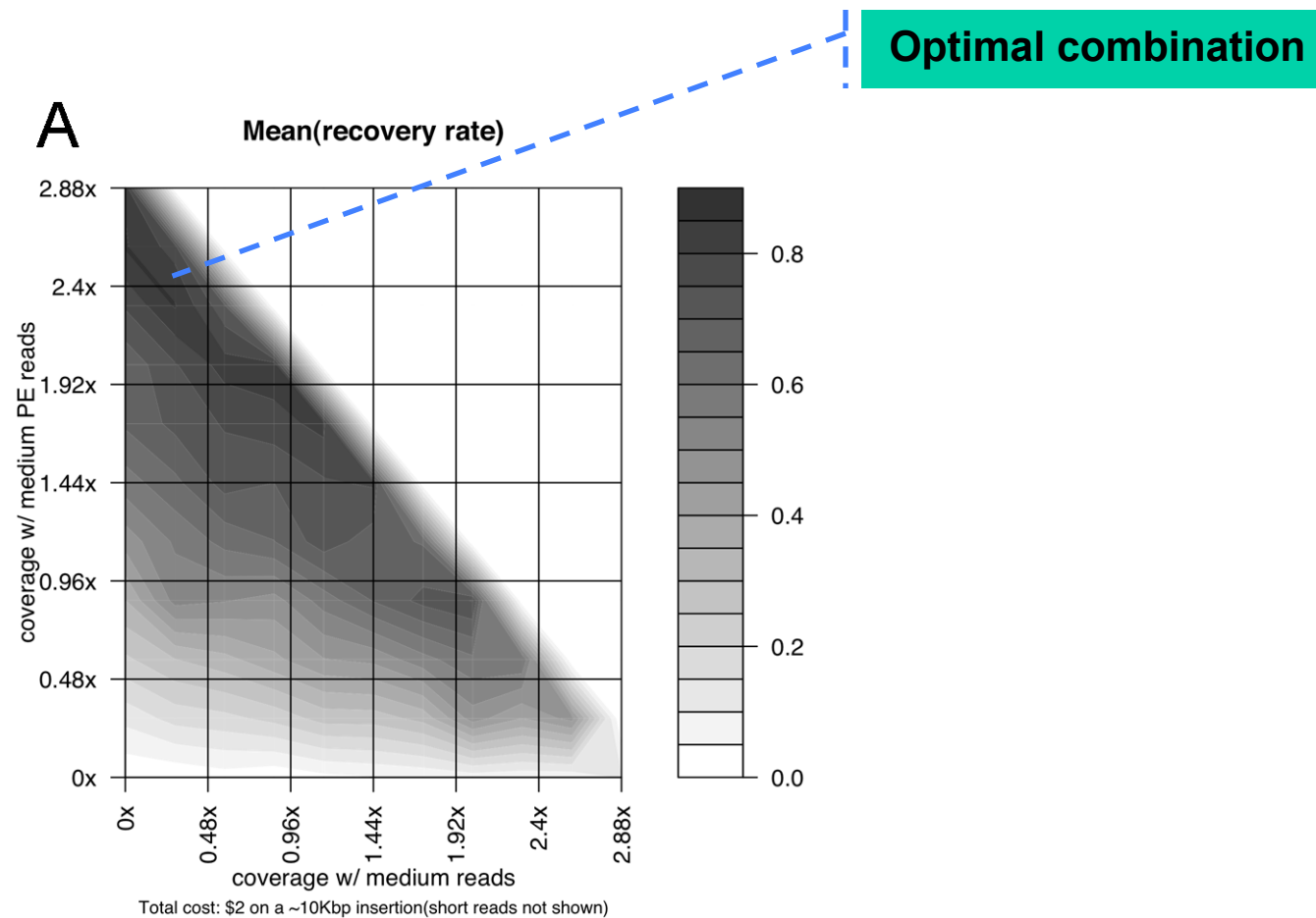
Optimal integration of sequencing technologies: *Efficient Simulation using A Simplified Assembler*

G Iterative contig elongation with the best supported extension



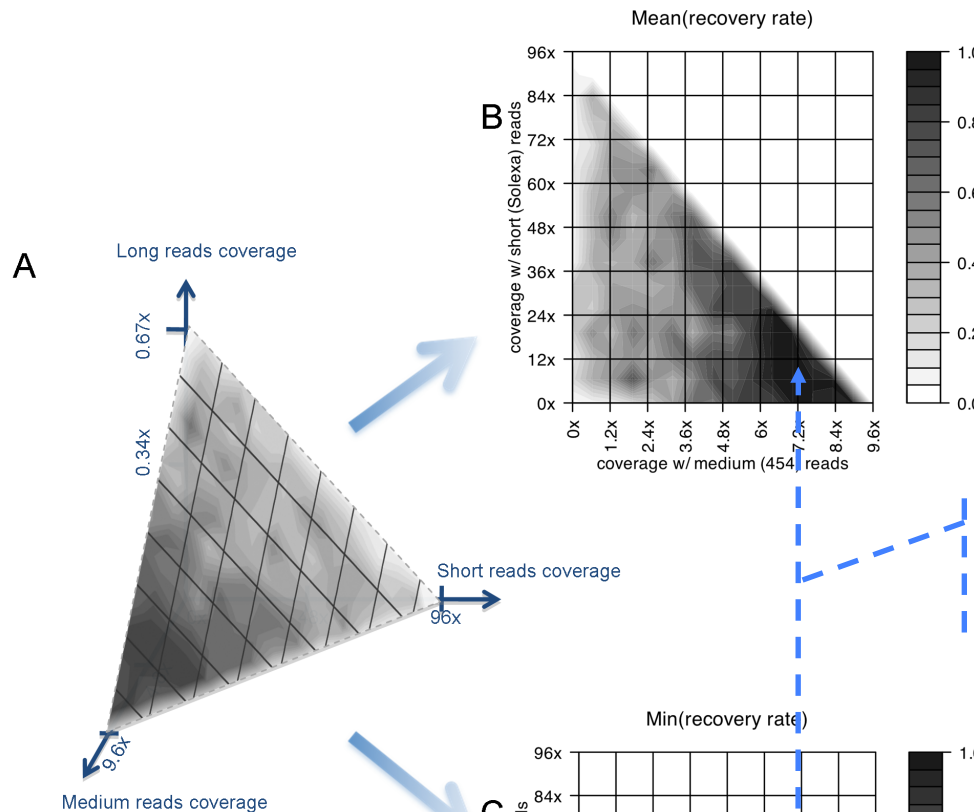
Optimal integration of sequencing technologies: Simulation shows power of PEs

Simulation results w/ shotgun & paired-end reads on the same ~10Kb insertion

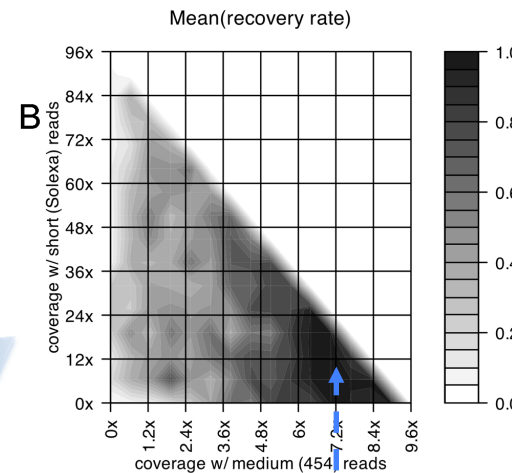


Source: Du et al. (2009), PLOS Comp Biol, in press

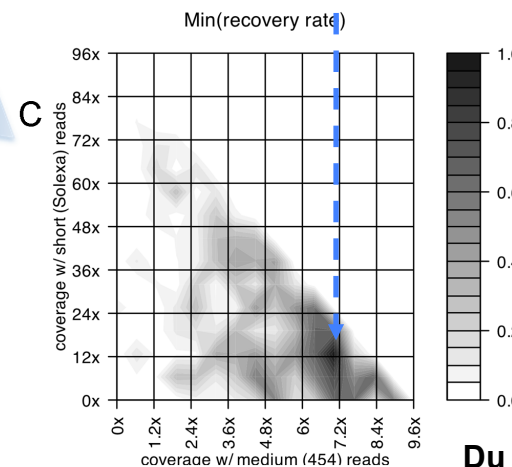
Optimal integration of sequencing technologies: Simulation shows combination better than single technology



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

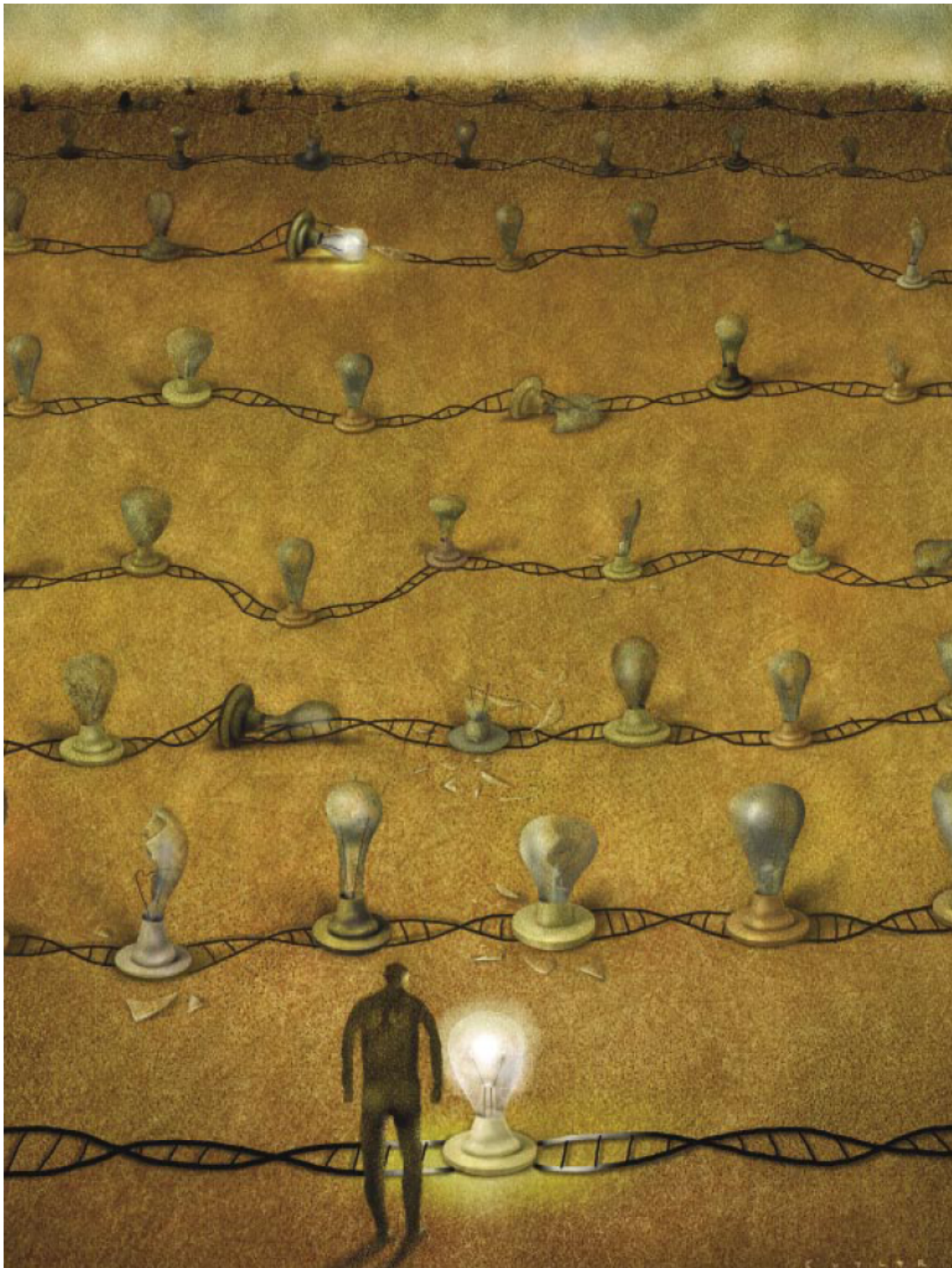


**Optimal combination of
different technologies**



**Result dependent
on specific
parameter setting
of different
sequencing
technologies**

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



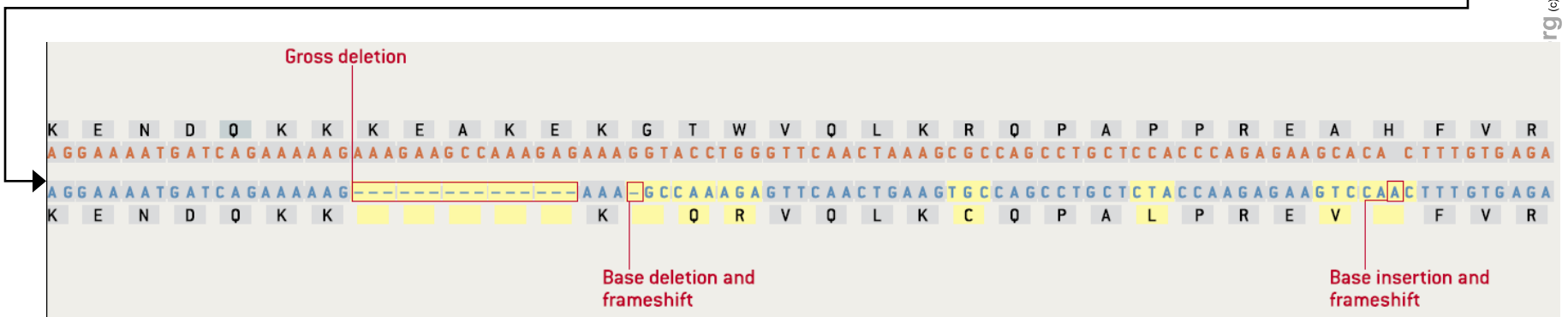
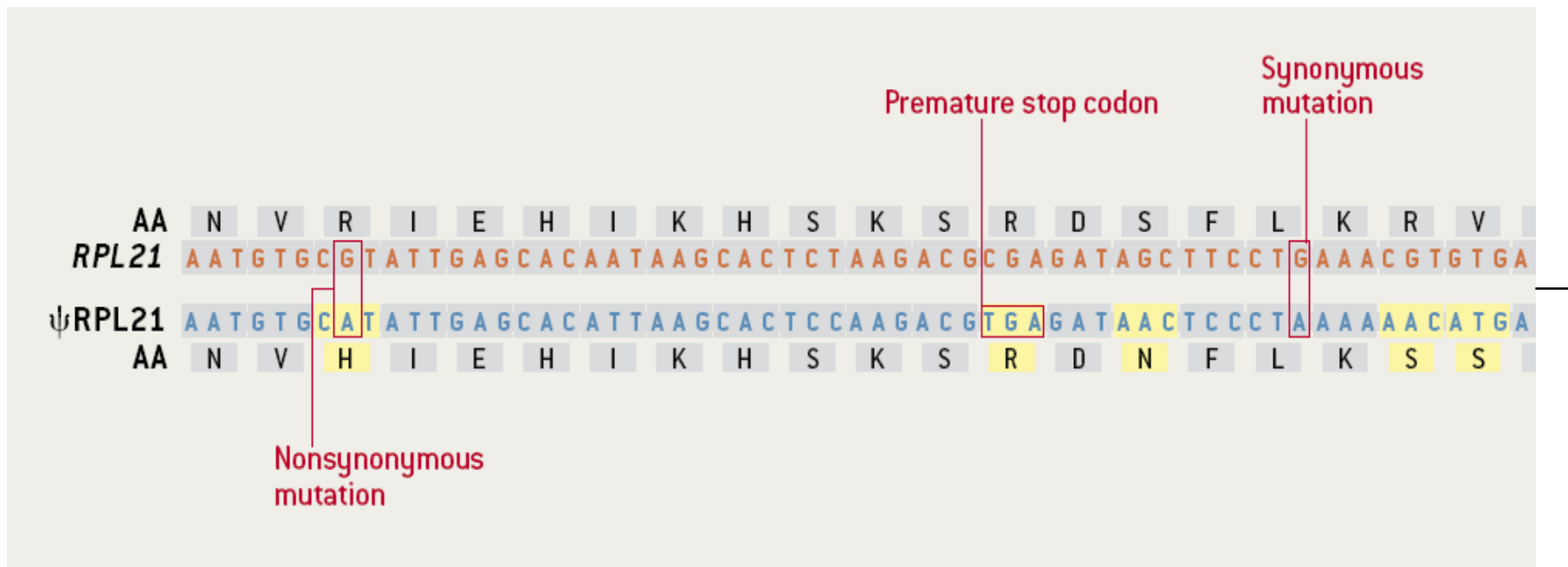
Formal Annotation based on Comparative Genomics: Pseudogenes

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

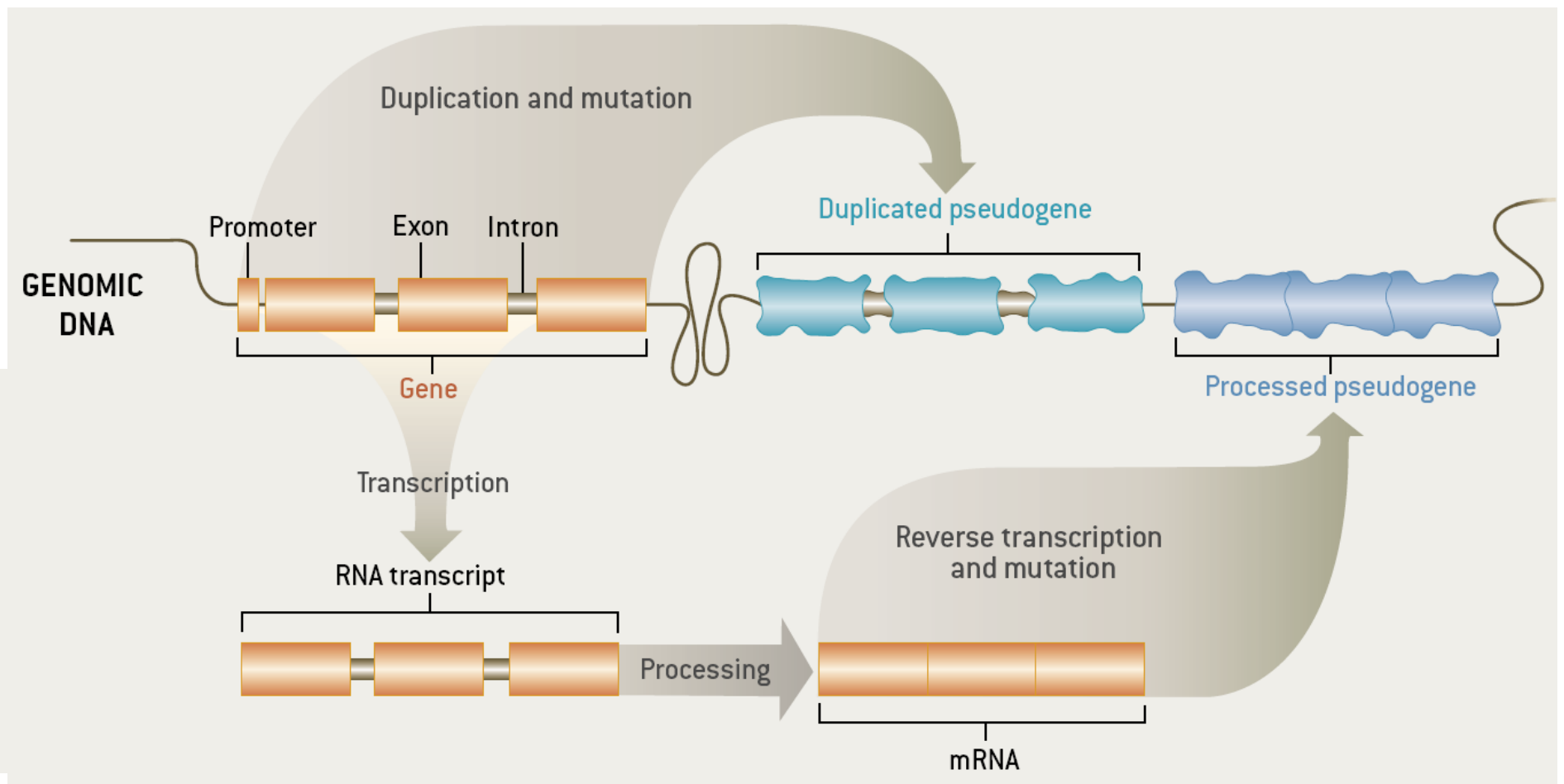
Pseudogenes are among the most interesting intergenic elements

- Formal Properties of Pseudogenes (Ψ G)
 - ◊ 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?...

Identifiable Features of a Pseudogene (ψ RPL21)

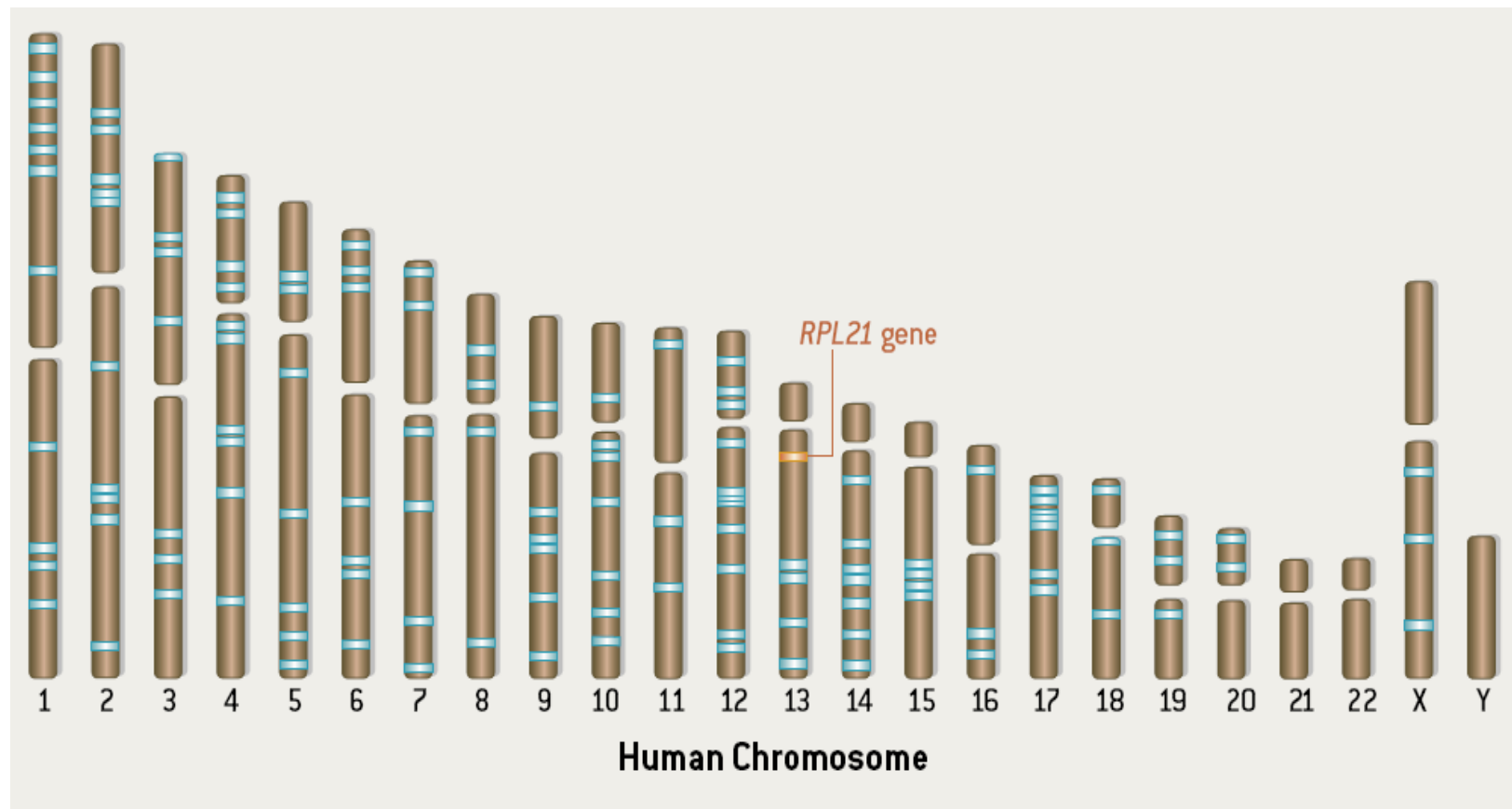


Two Major Genomic Remodeling Processes Give Rise to Distinct Types of Pseudogenes



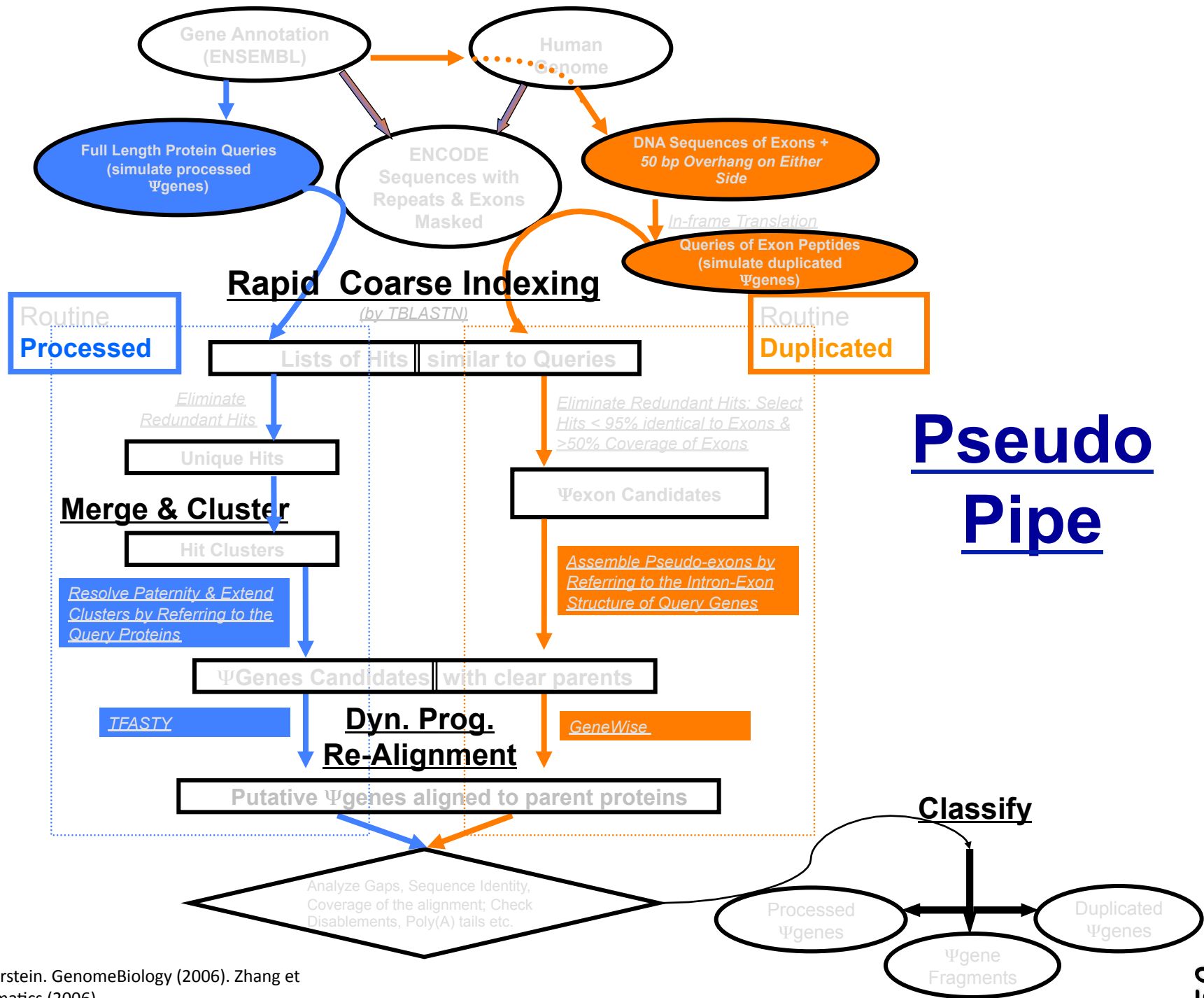
Gerstein & Zheng. Sci Am 295: 48 (2006).

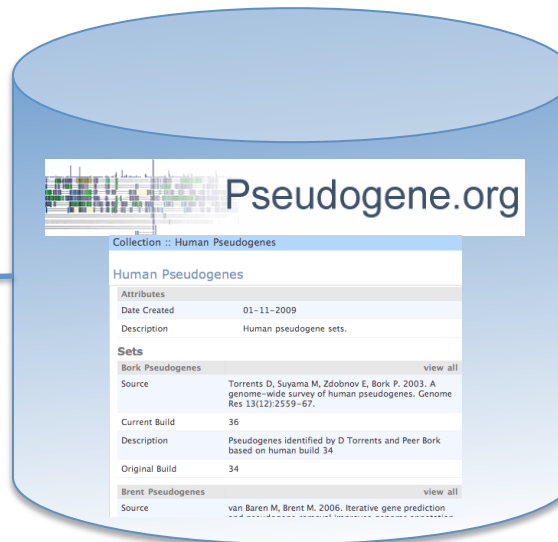
Distribution of Human Pseudogenes (for RPL21) across the chromosomes





Pseudogene Tools: Assignment Pipeline & DB





Flat Files

DAS

Table Browser

tables.pseudogene.org

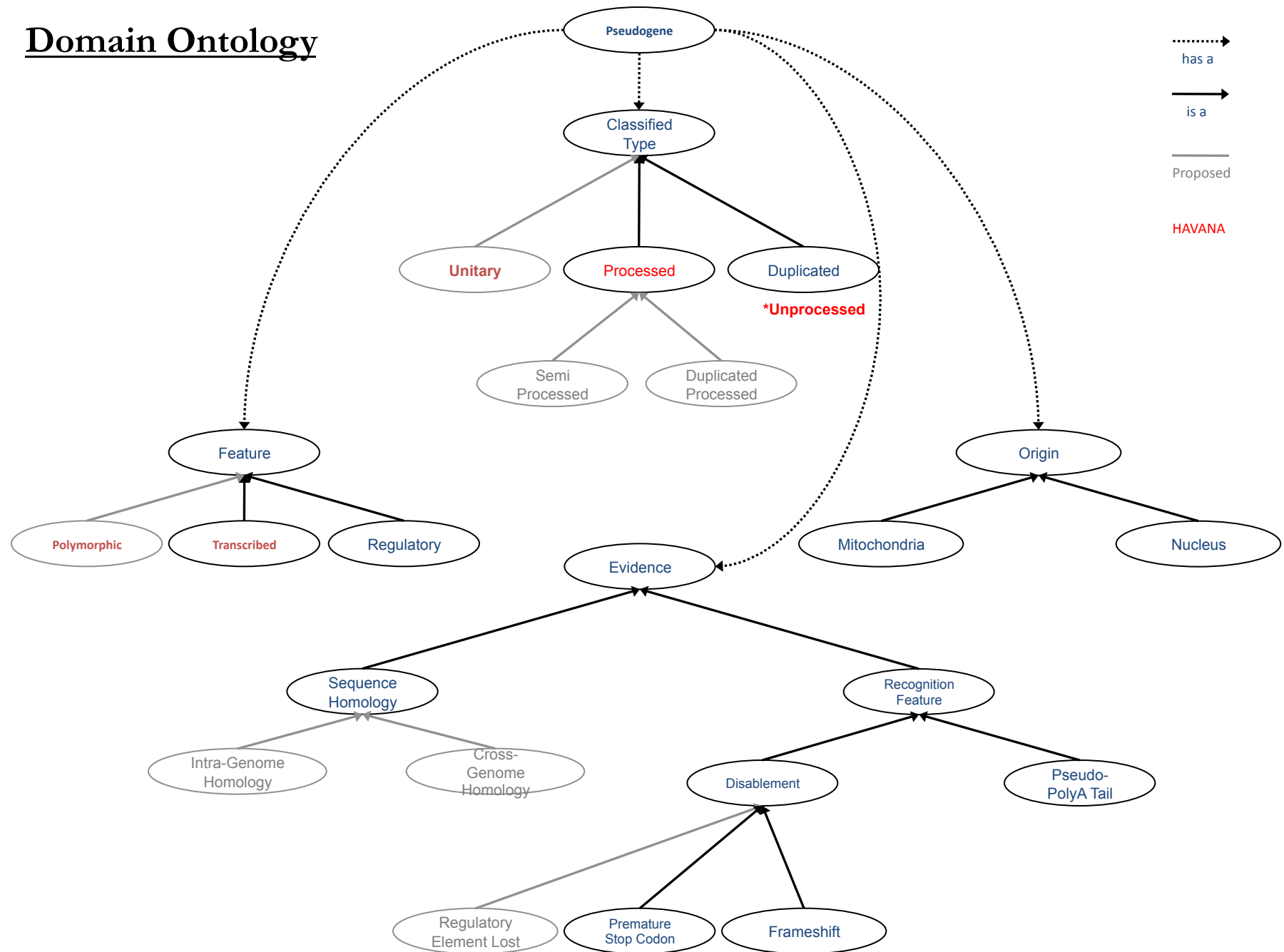
UCSC
Genome
Browser

- 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

Domain Ontology



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

Pseudofam Construction

- Data Generation

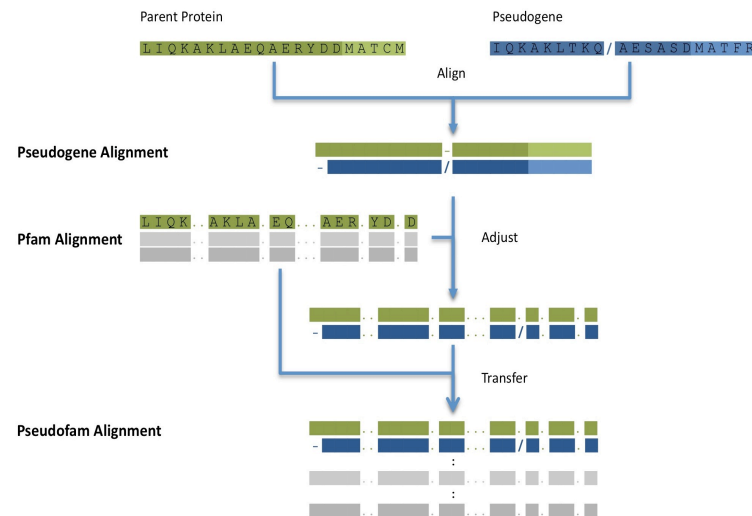
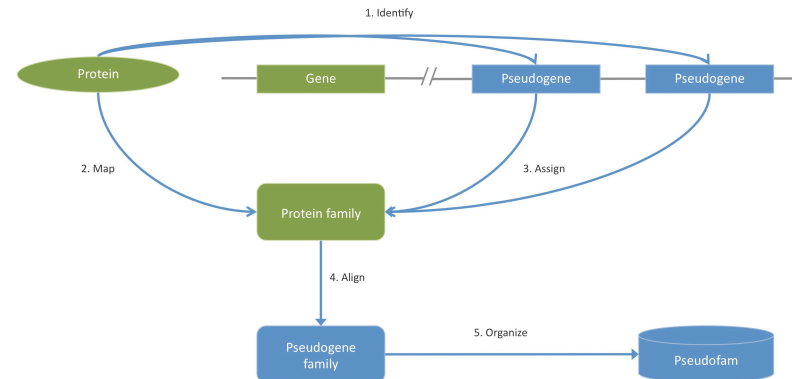
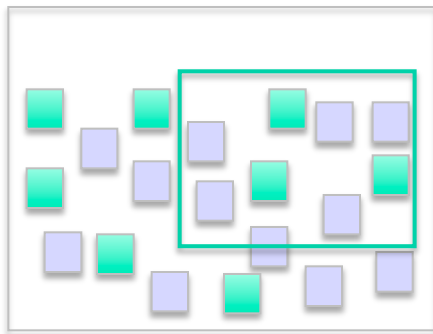
- ◇ Identify pseudogenes by proteins and map parent proteins to protein families

- Alignment

- ◇ Align pseudogene to parent
 - ◇ Transfer alignment from Pfam
 - ◇ Combine and adjust the alignments to build the pseudofam alignment

- Statistics

- ◇ Enrichment



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

Overall Flow:

Pipeline Runs, Coherent Sets,

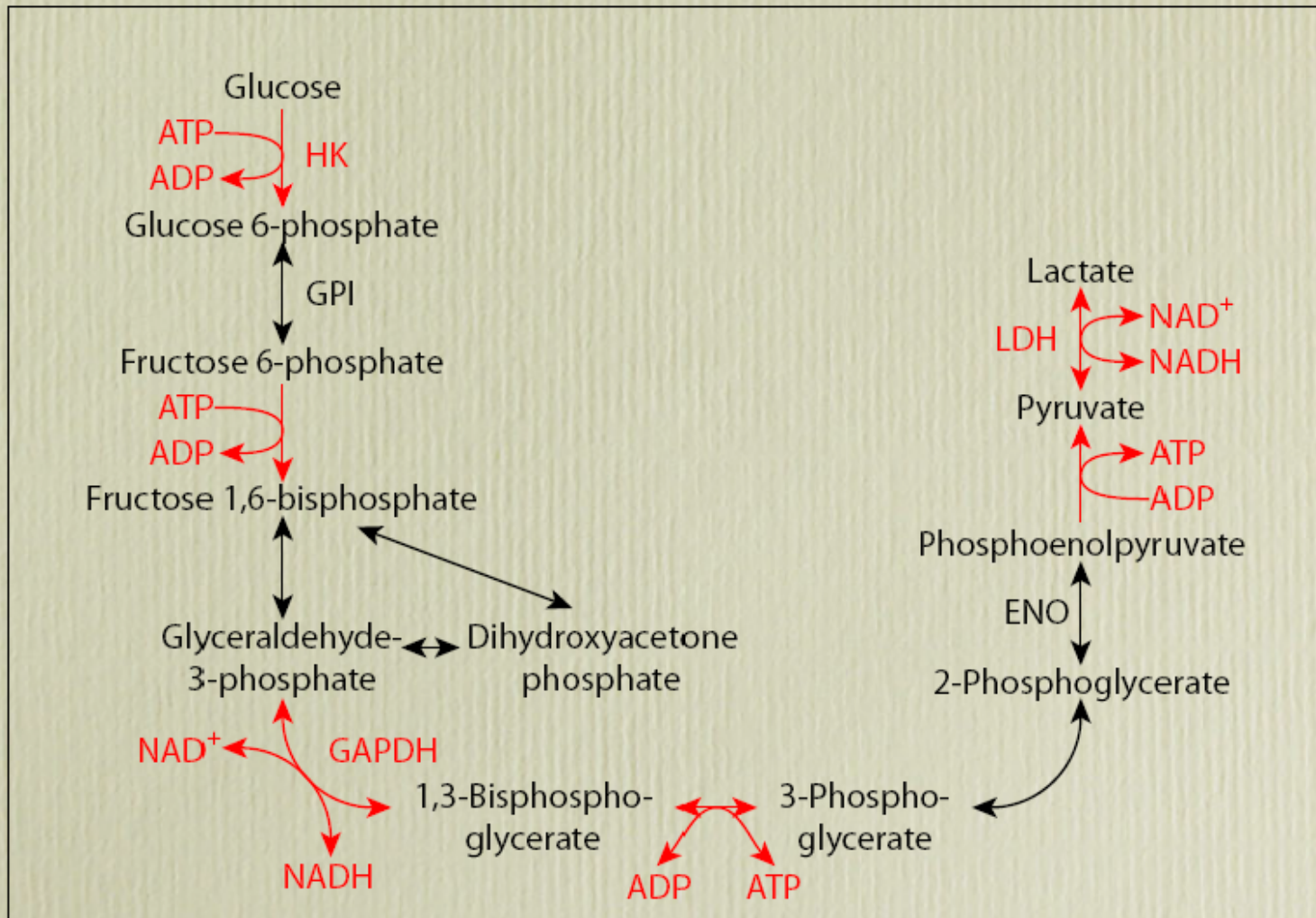
Annotation, Transfer to Sanger

- 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



Pseudogenes of glycolytic enzymes



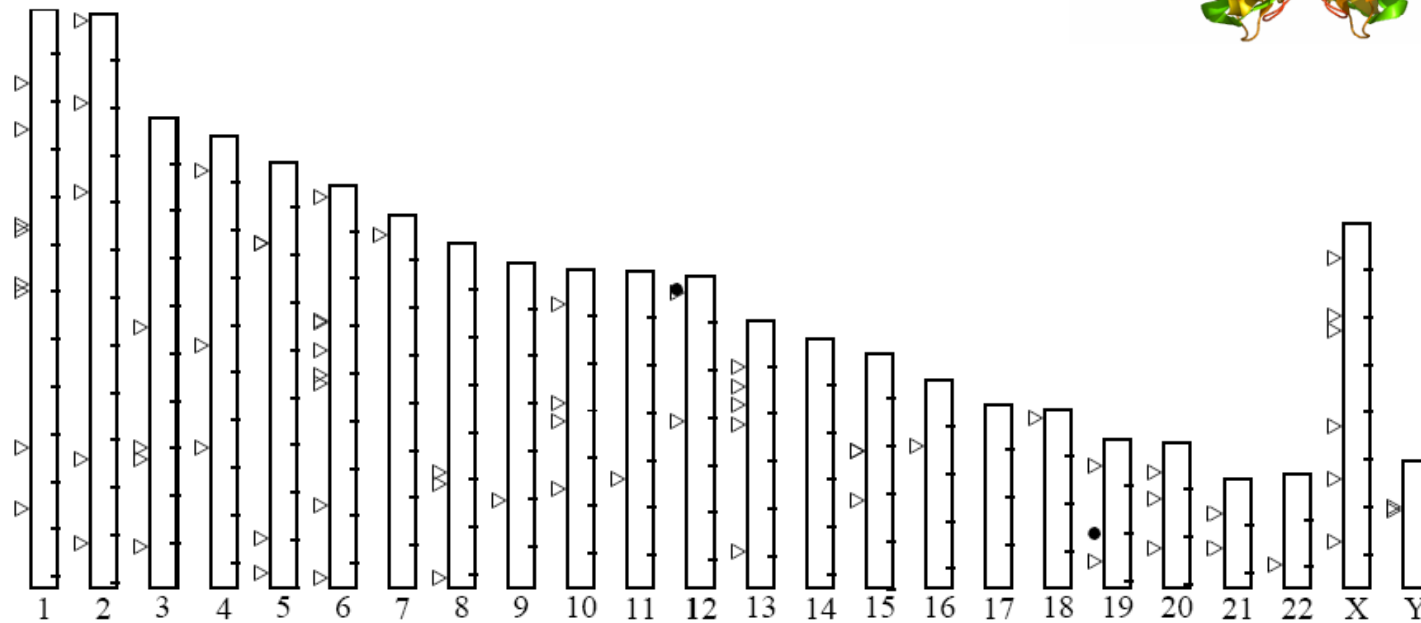
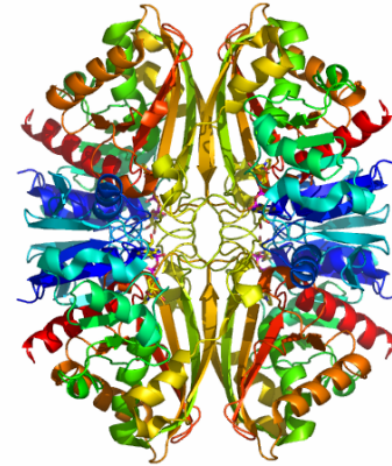
Number of pseudogenes for each glycolytic enzyme

	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

Processed/Duplicated

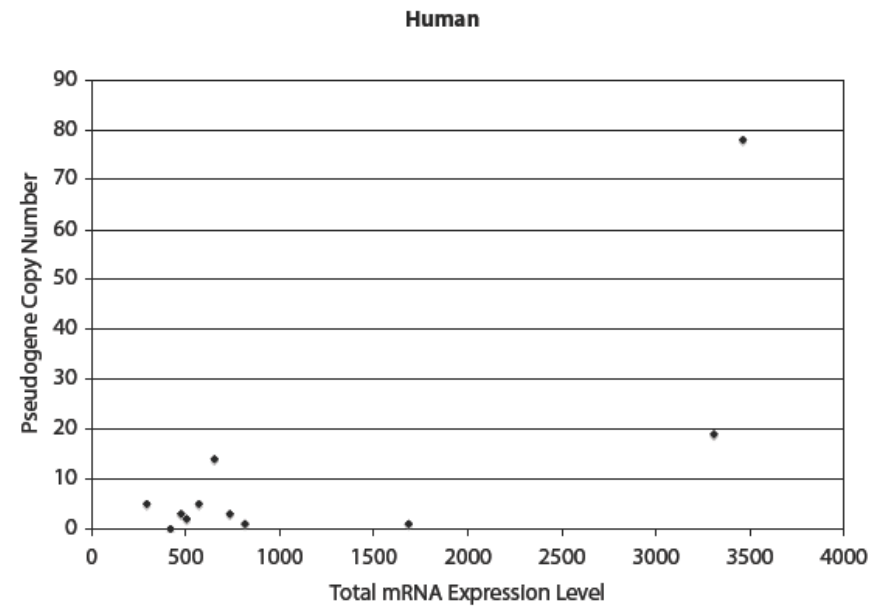
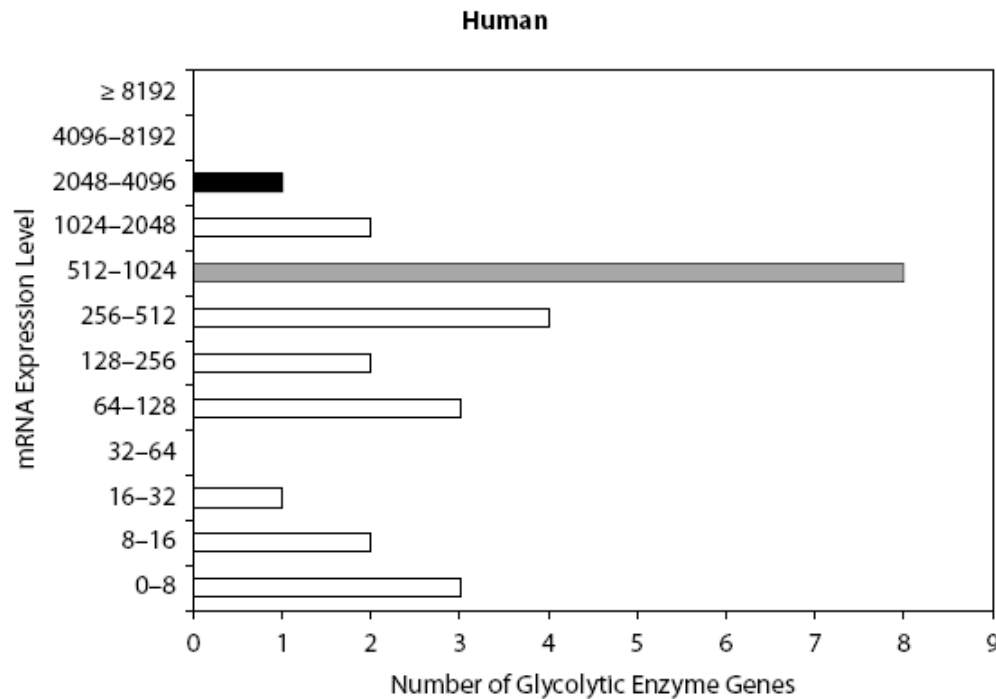
[Jong et al. BMC Genomics ('09, in press)]

Distribution of human GAPDH pseudogenes



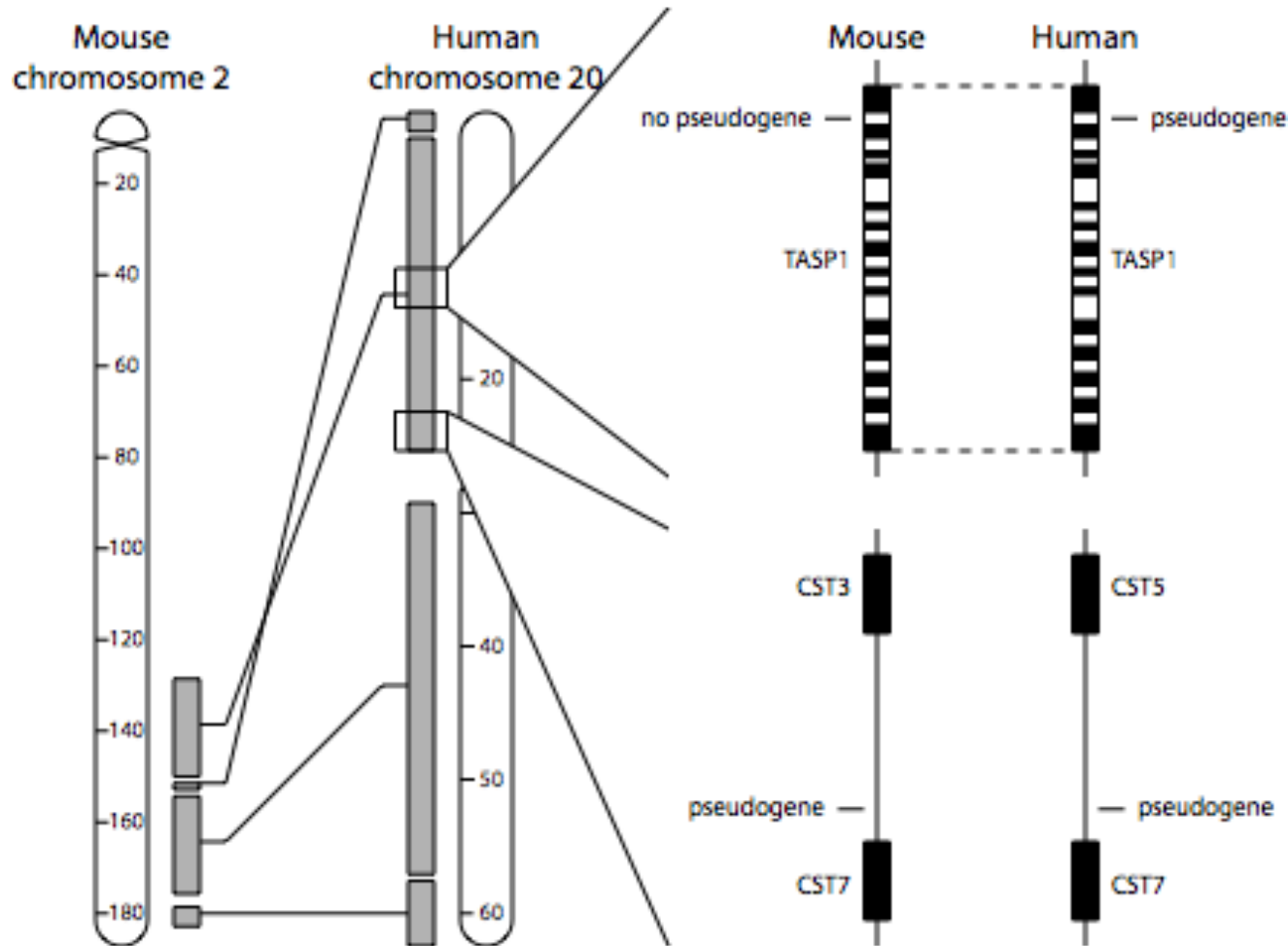
[Jong et al. BMC Genomics ('09, in press)]

Pseudogene abundance versus mRNA expression levels



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

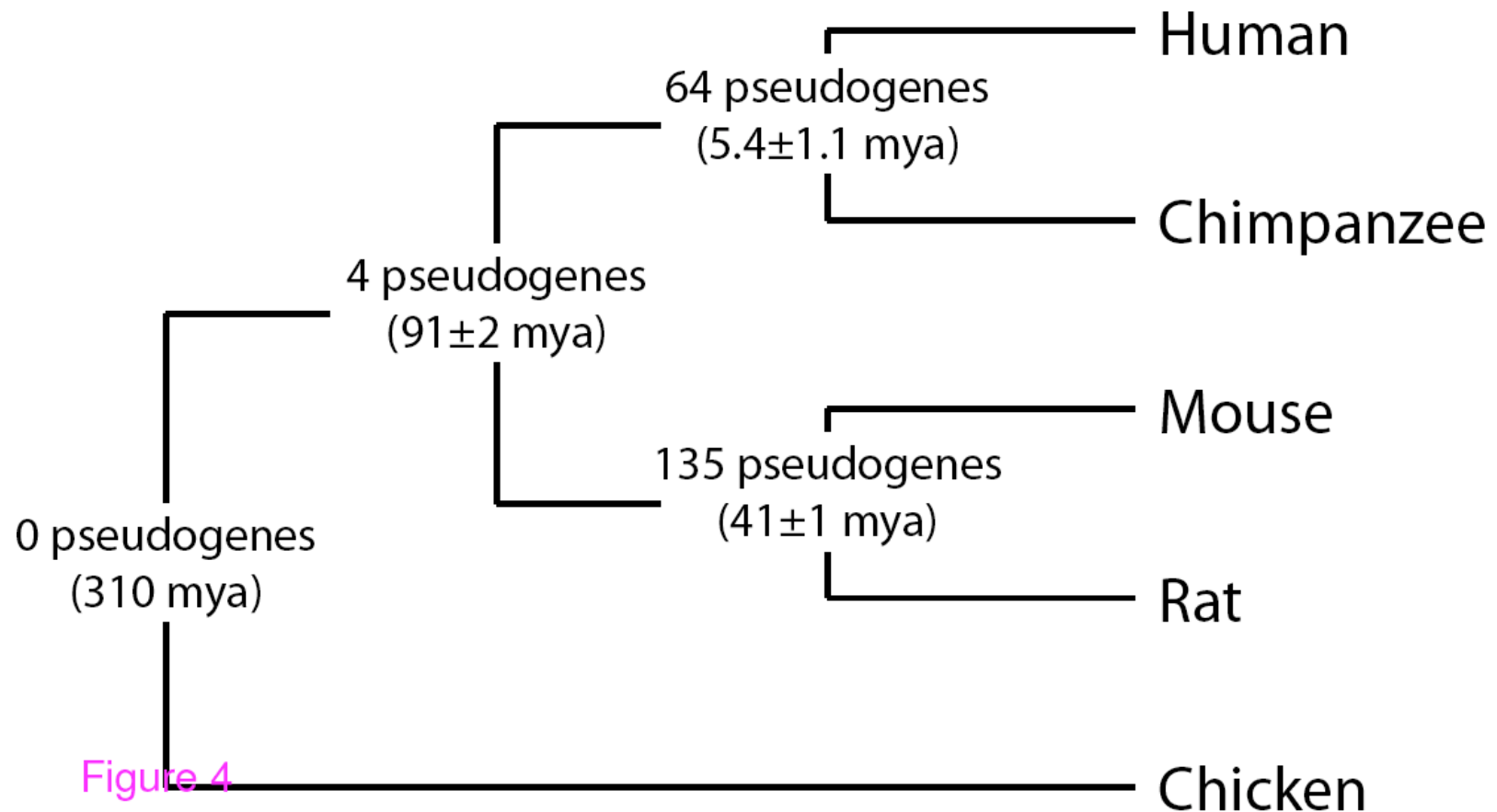
Using Synteny to Identify syntenic glycolytic pgenes

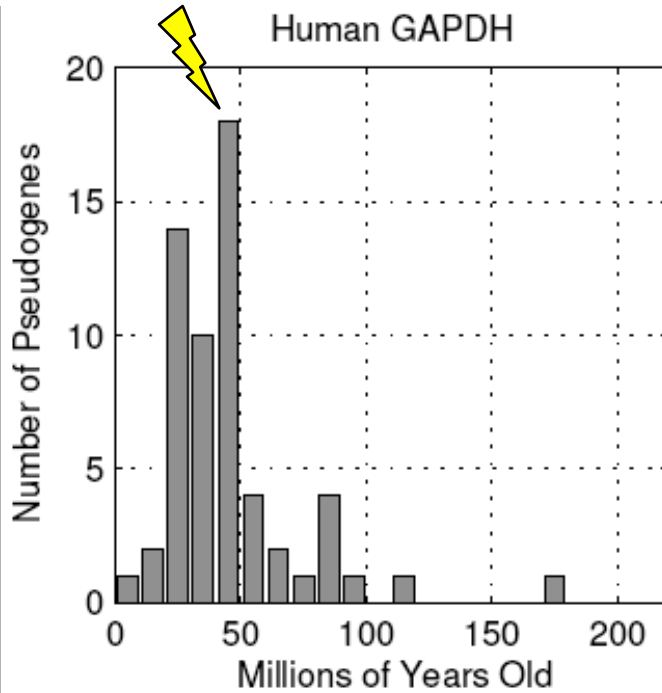


Synteny derived based on local gene orthology

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

Syntenic proc GAPDH pseudogenes

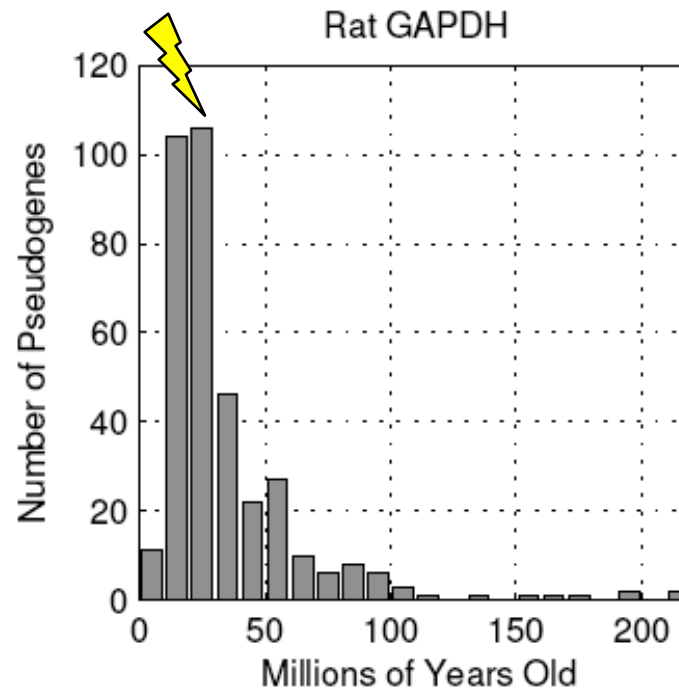
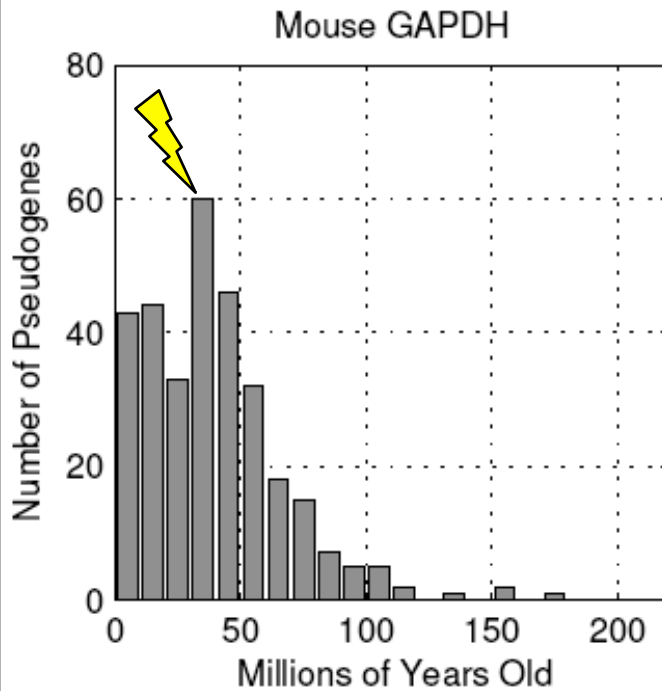




**Burst of
Retrotran-
spositional
Activity**

Age of GAPDH pseudogenes

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



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

ENCODE Pilot Pseudogenes: Integration with Measures of Biochemical Activity



Connecting TARs (TxFragments) 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.

Composite
ChIP
hit

Special
 ψ G
tracks in
browser

diTAG

CAGE






TARs

ChIP-
chip



Zheng et al. (2007) Gen. Res.

Intersection of Pseudogenes with Transcriptional Evidence

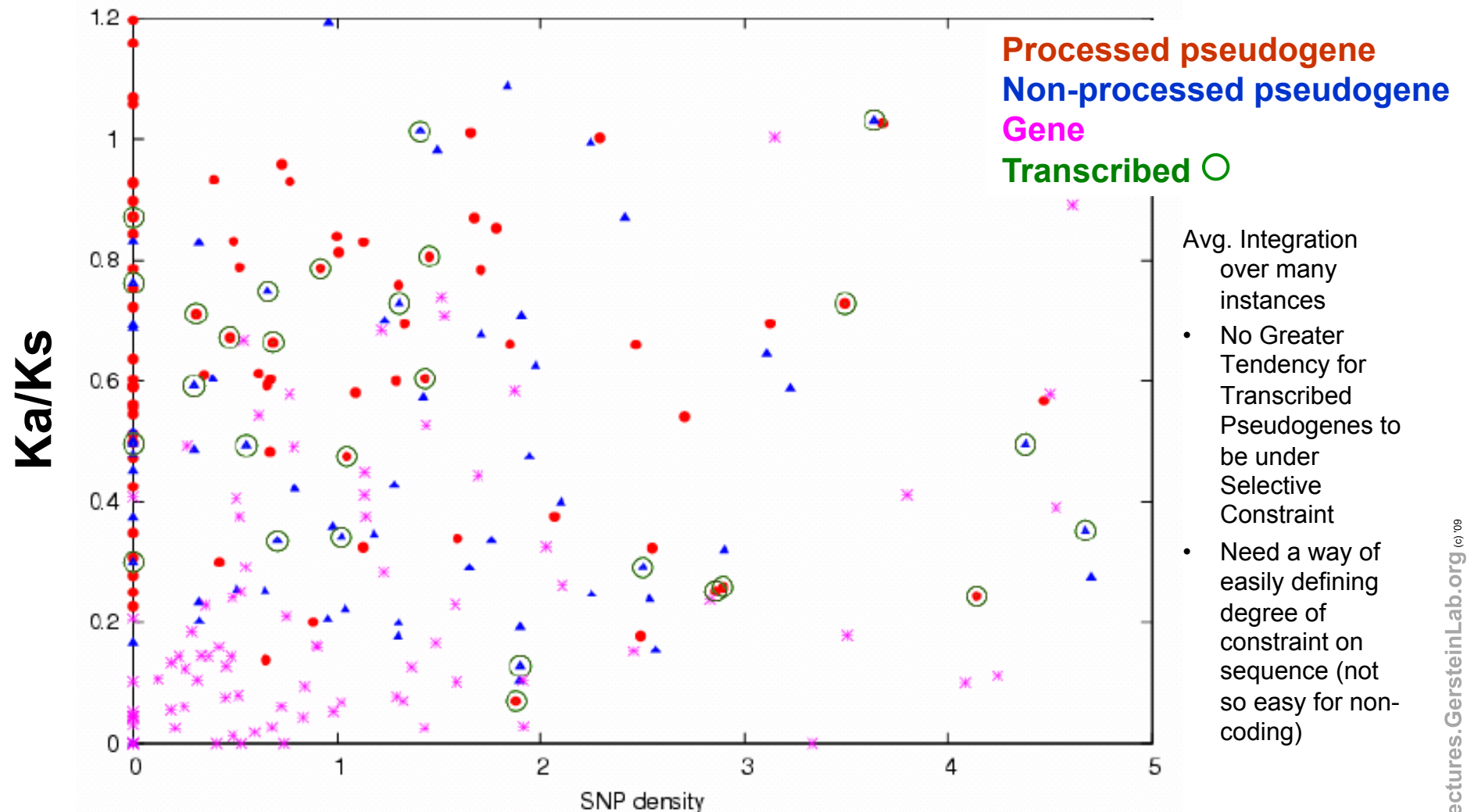
	TAR / transfrag	CAGE	DiTag	RACEfrag	EST / mRNA
TAR / transfrag	105 *	8	2	5	14
CAGE		8	1	0	1
DiTag			2	0	0
RACEfrag				<u>14</u>	5
EST / mRNA					21 

Excluding TARs (due to cross-hyb issues)

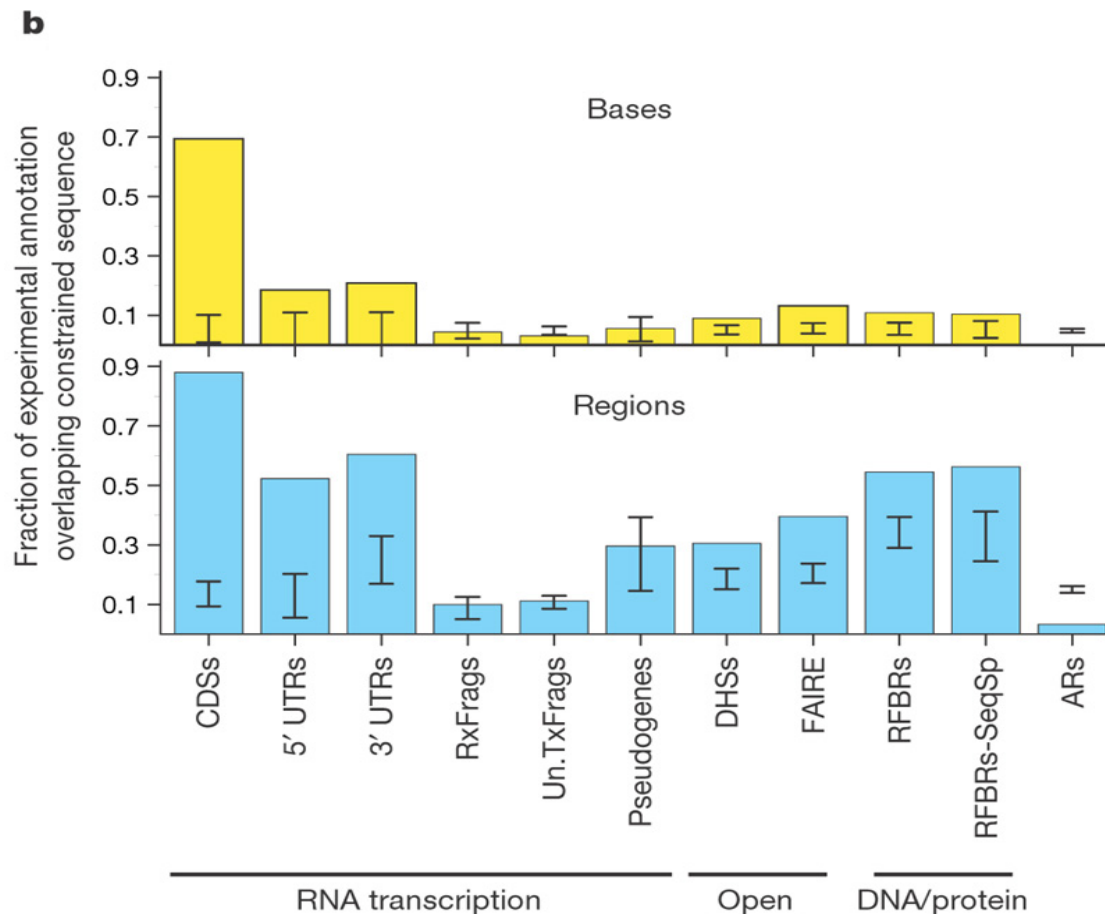
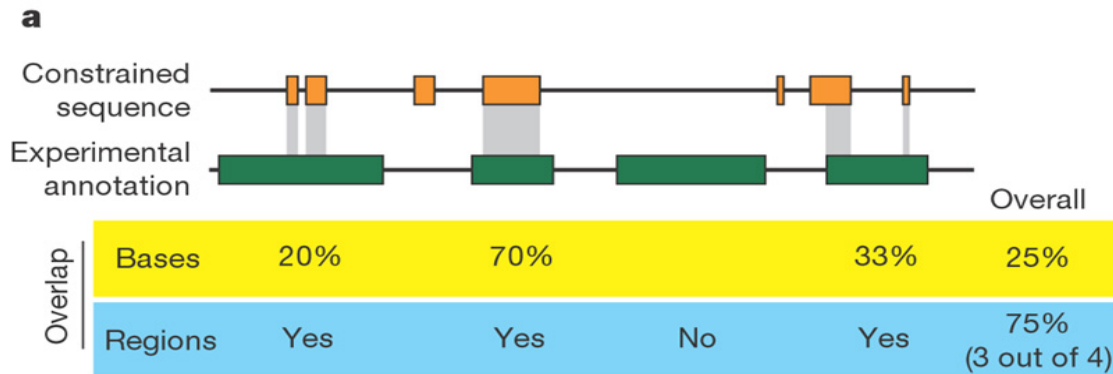
Targeted RACE expts to 160 pseudogenes, gives 14

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

Integrating Transcriptional Evidence with Gene Annotation and Sequence Constraints



Zheng et al. (2007) Gen. Res.

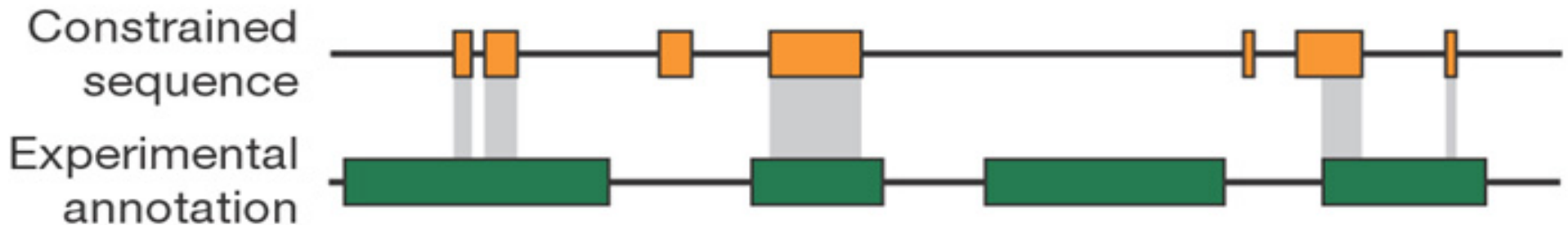


[ENCODE Consortium, *Nature* 447, 2007]

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

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

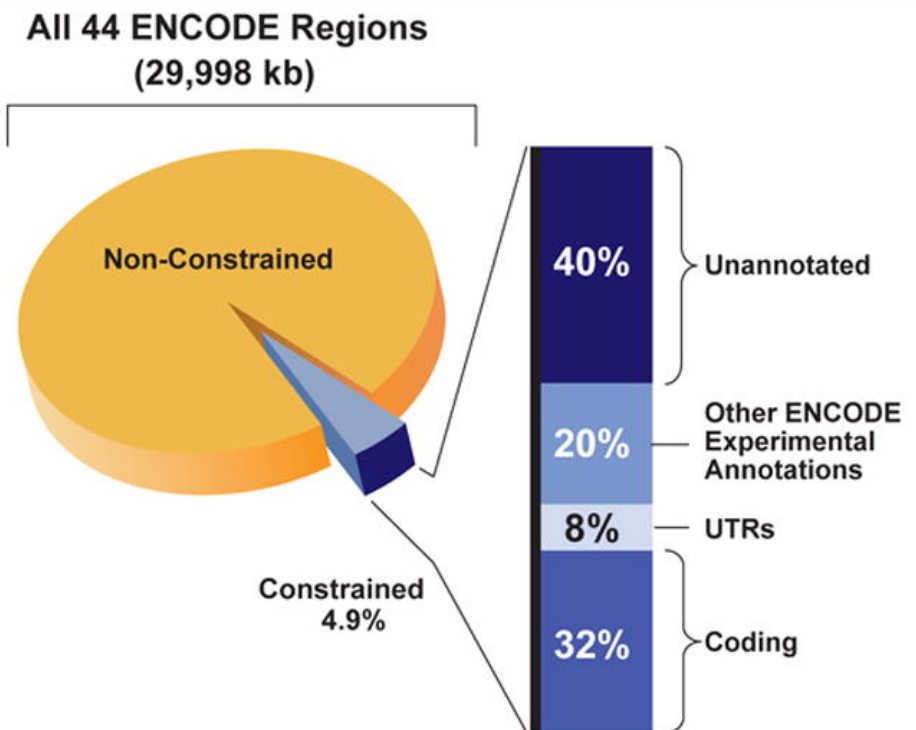
Grand Summary: Biochemical Activity vs. Sequence Constraints



- 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. "

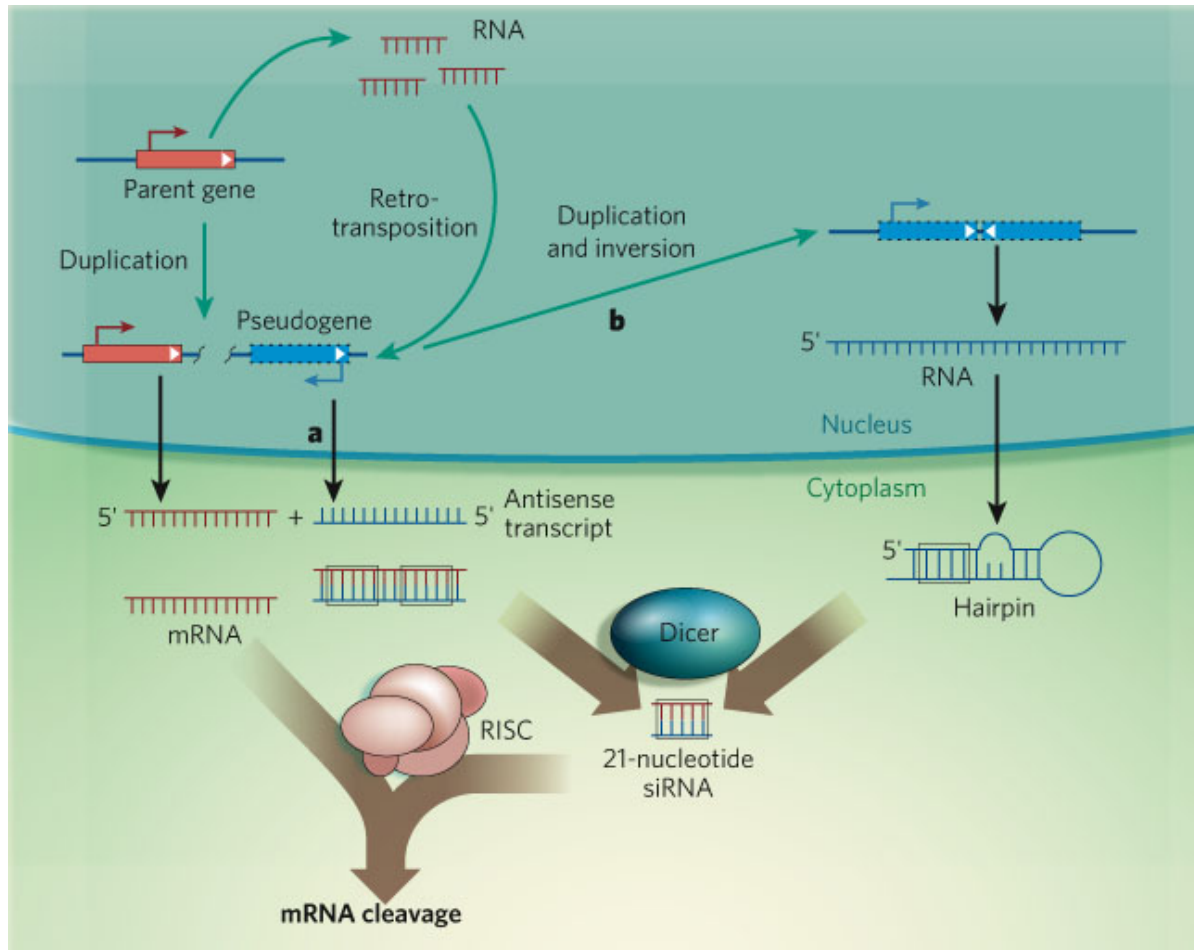


[ENCODE Consortium, *Nature* 447, 2007]

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).

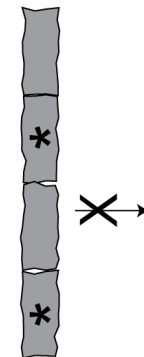
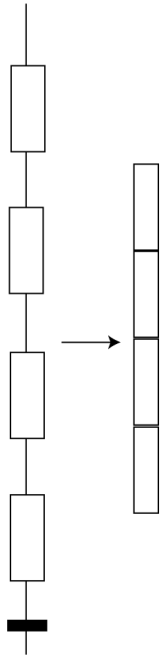
[Sasidharan & Gerstein, *Nature* ('08)]

Genes & Pseudogenes

(a) Functional Gene

Ambiguous Cases

(b) Dead Pseudogene

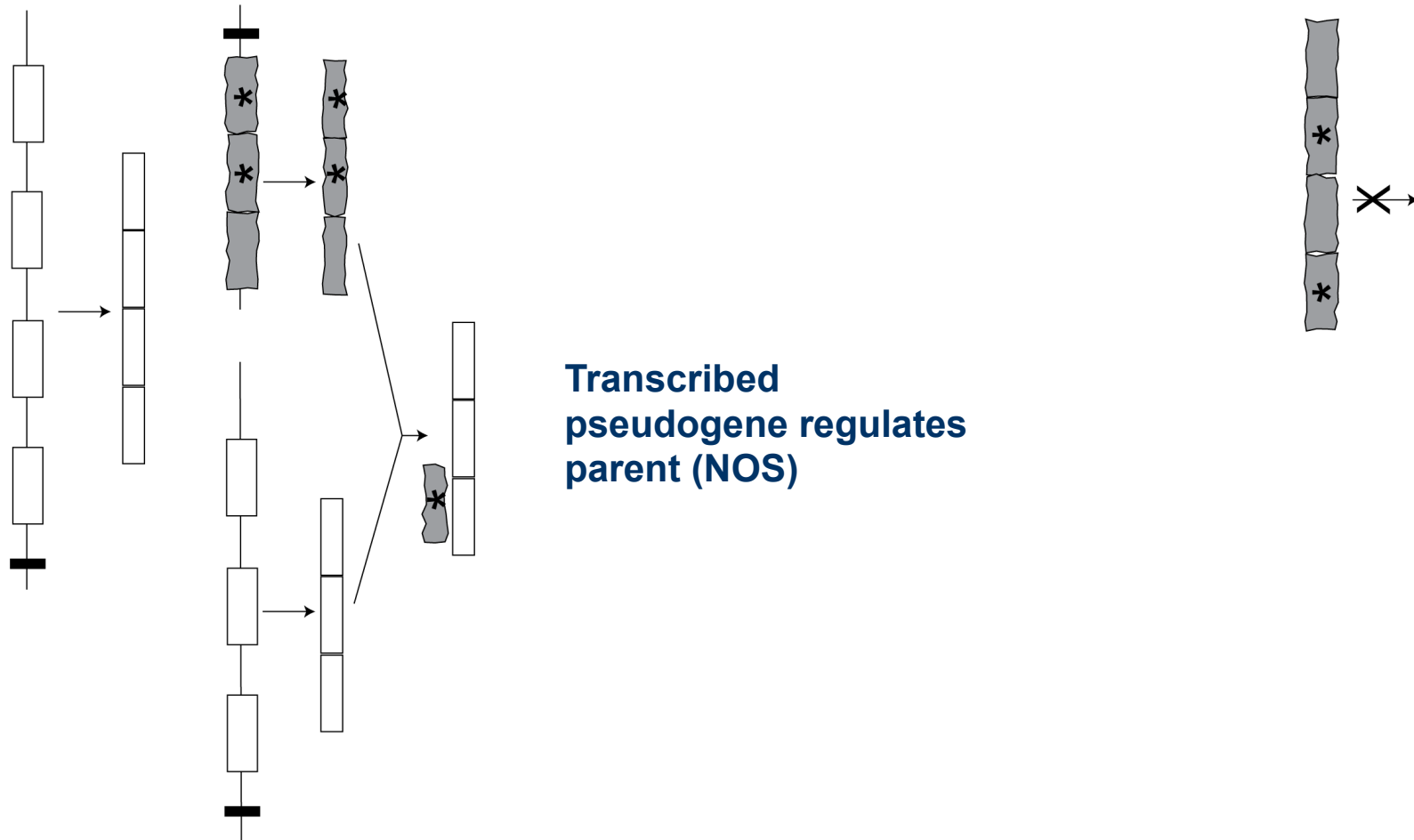


Genes or Pseudogenes?

(a) Functional Gene

Ambiguous Cases

(b) Dead Pseudogene

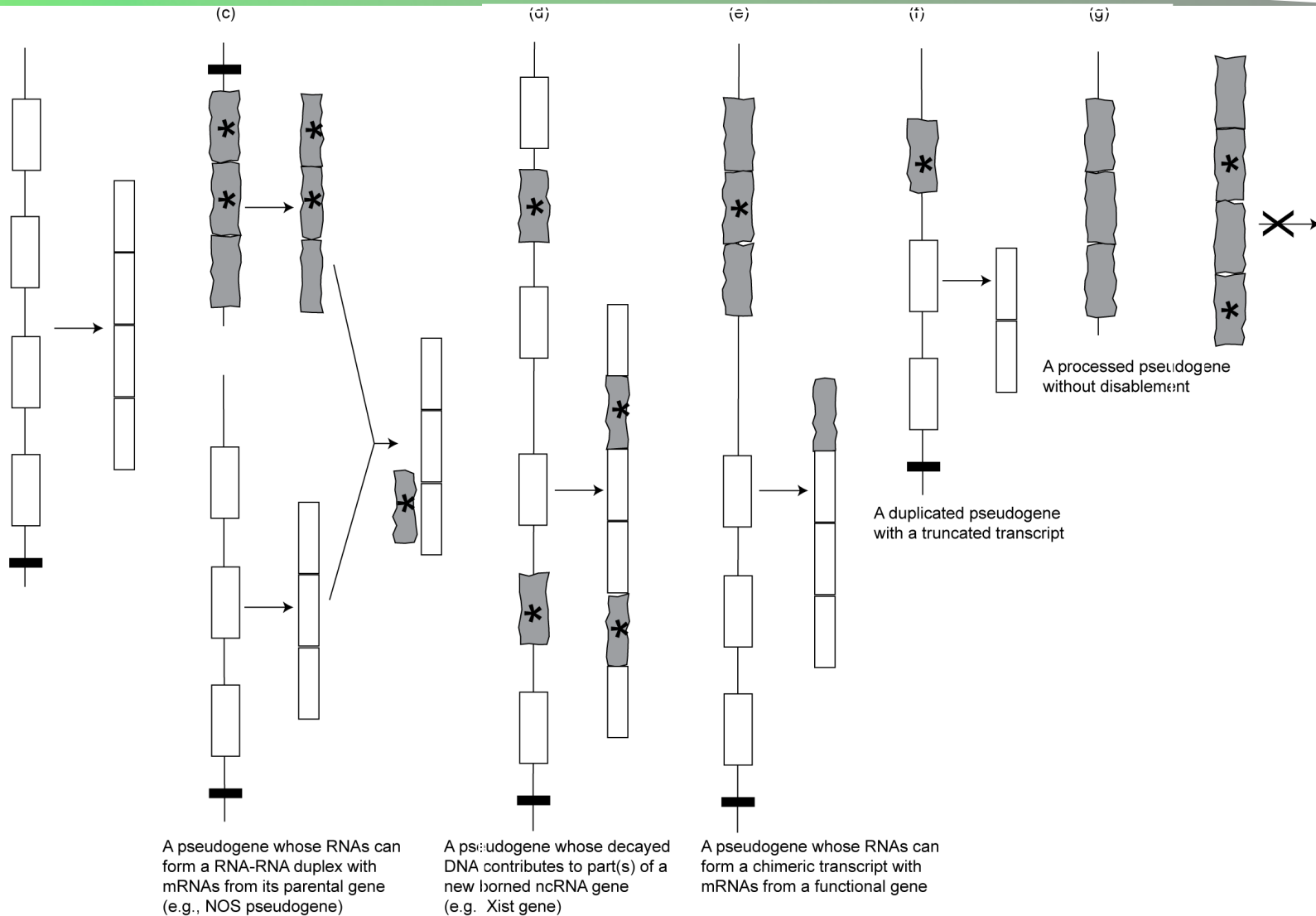


Genes or Pseudogenes?

(a) Functional Gene

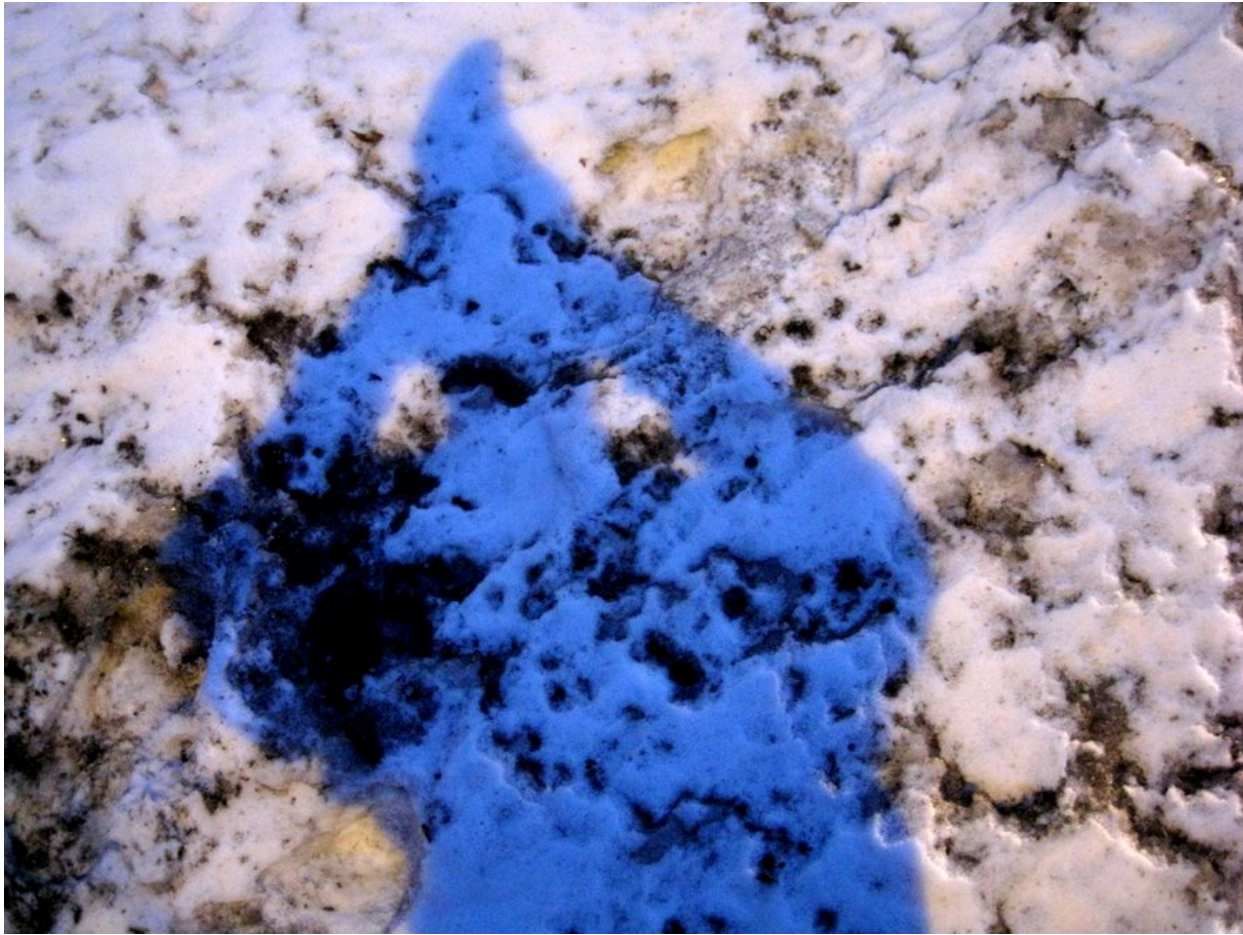
Ambiguous Cases

(b) Dead Pseudogene



Summary:

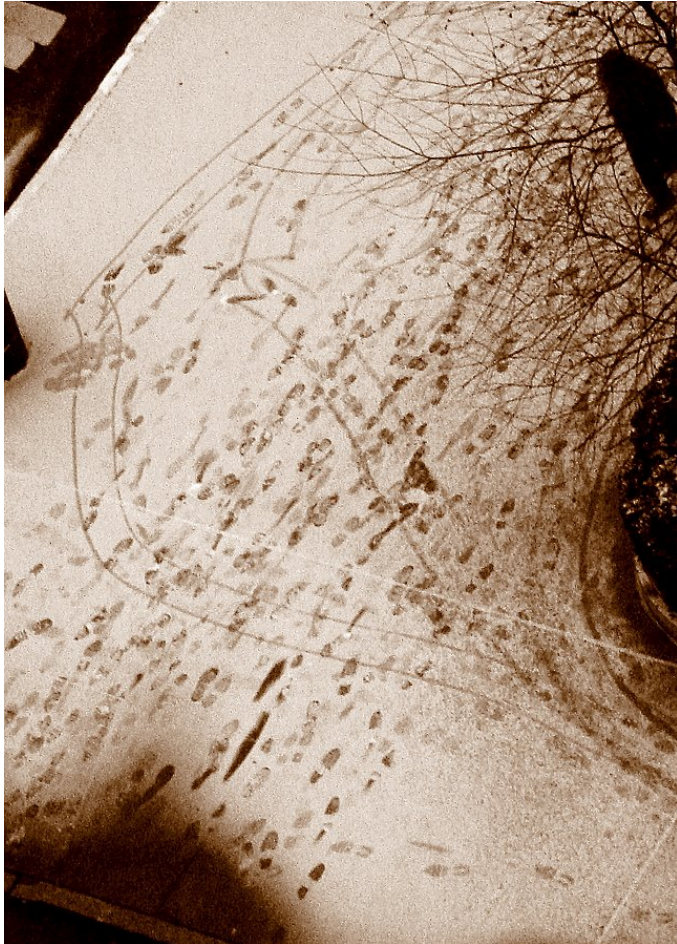
Looking Back Over the Talk



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 punctate "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
- Pseudogenes
 - A. Pattern-match tools for calling them
 - A. Focus on one group of pseudogenes
 - c. Integrating them with annotations of transcription and regulation
- Future of Annotation
 - ◇ What is a "gene" post encode?

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 Novel Transcribed Regions and Groups of Binding Sites

- Deserts and Forests of Binding Activity
 - ◇ on ~50kb scale
 - ◇ Biplot gives broad separation of seq. specific and non-specific factors and associated genomic bins



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 parameterize statistical model
- ReSeqSim
 - ◇ Efficiently simulating assembly of a representative variant
 - ◇ Shows that best reconstruction has a combination of long, med. and short reads

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, highlight outliers
- 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
- 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,
M Hurles, Tim Hubbard, Jennifer Harrow, J Affourtit, M Egholm

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

+

ENCODE, modENCODE, 1000 Genomes

D Zheng
Z_{hengdong} 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

Yale Acknowledgements



GenomeTECH.gersteinlab.org
Pseudogene.org

More Information on this Talk

TITLE: Human Genome Annotation

SUBJECT: GenomeTechAnnote

DESCRIPTION:

The ninth international conference for the Critical Assessment of Massive Data Analysis (CAMDA 2009), 2009.10.05, 10:00-10:50;
[I:**CAMDA**] (Long GenomeTechAnnote talk, building on [I:**UCSC**] with some subtractions and a first time addition **gapdh-pgenes*** .)

(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> .