

### Human Genome Annotation

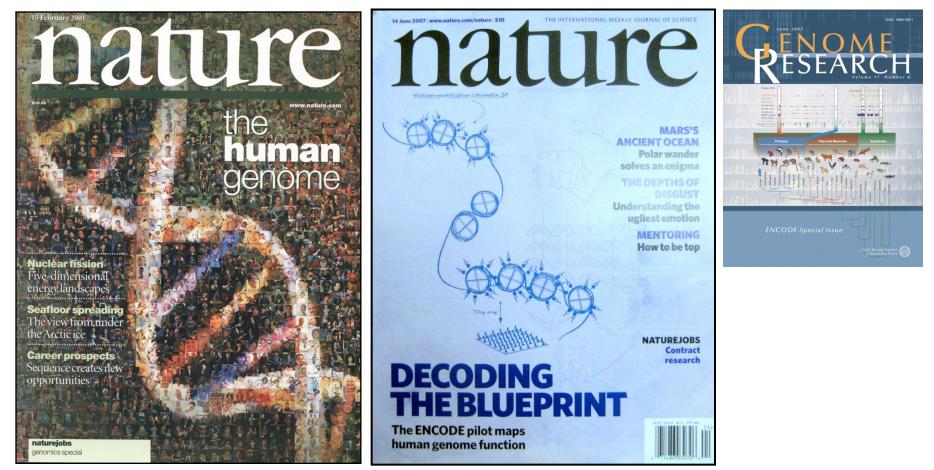
Mark B Gerstein Yale

#### 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 [IHGSC, *Nature* 409, 2001] they be annotated?



# 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]

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

#### **BV DENNIS OVERBYE**

ESSAY

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 aught 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=mc1, along with "1905," the date that the young Einstein derived it, into the bacterium's genome he 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 Biotechnol-ogy 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 sce-

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

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



If a bacterium can be encoded with E=mc<sup>2</sup>, 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,

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."

change, and have remained iden

mand and control functions.

mutate even more rapidly.

mice, chickens and dogs for at least 300 million years.

of them had turned out to be playing important com-

bits of DNA that neither help nor annoy an organism

But Dr. Bejerano, one of the discoverers of these

ultraconserved" strings of the genome, said that many

"Why they need to be so conserved remains a mystery," he said, noting that even regular genes that do

ething undergo more change over time. Most junk

The Japanese team proposed to sidestep the muta-

tion problem by inserting redundant copies of their mes-

sage into the genome. By comparing the readouts, they

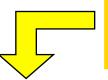
said, they would be able to recover Einstein's formula

even when up to 15 percent of the original letters in the

had changed or mutated "This is the

sections of junk DNA seem to be markedly resistant to





with their minds, and hearts and hands they can shape their own destiny. ... identified on chromosome 16 in families with the infinite of particles of particles and the soft and the soft

**Junk DNA** as Art



Ο

#### How might we annotate a human text ? F 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 **Color** is the ways to say: **Function** 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 Lines are languages. Another survey, compiled Similarity 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

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

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

been inventing one language a week,

Brian Hayes

The Semicolon Wars

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

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

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect, they'll concede, but it's built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will favor a different language. It's Lisp, cide which end of a boiled egg to crack. This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the leastsignificant 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 leud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal program statements have to be separated by semicolons. For example, in x := 0; y := x + 1; z := 2 the semicolons tell the compiler where one statement ends and the next begins. C programs are also peppered with semi-

# **Overview of the Process of**

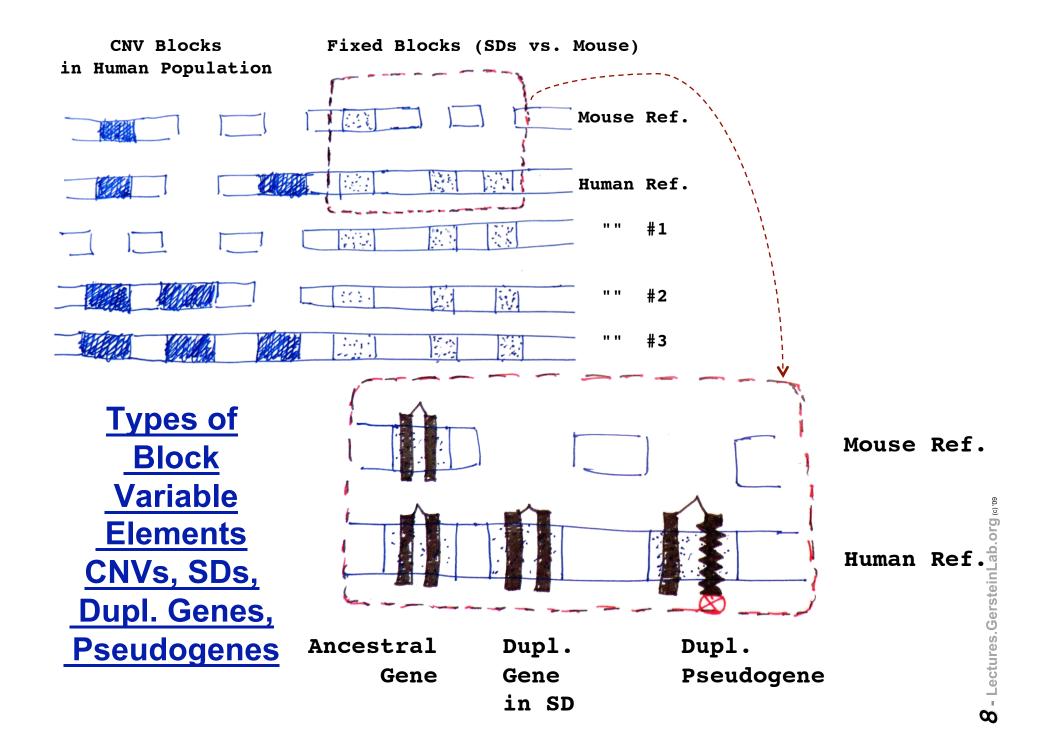
# **Annotation of non-coding Regions**

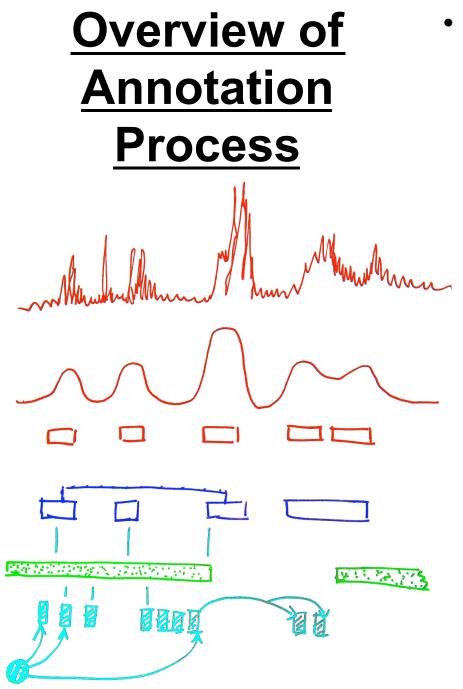
### Basic Inputs

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

# Results of Analyzing Similarity Comparison

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





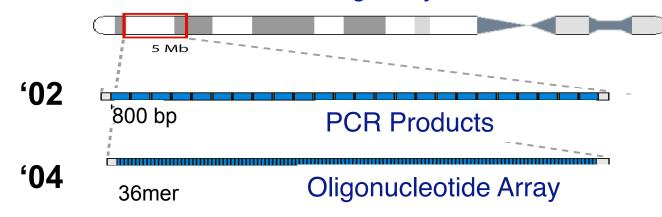
- Determining experimental signals for biochemical activity (e.g. transcription) across each base of genome
  - Oevelopment of Sequence (and Array) Technology
    - Normalizing & Scoring Signal, Correcting Artifacts, Segmenting to create Small Annotation Blocks
  - Output of Production Pipelines and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

#### Integrated Analysis Connecting Different Types of Annotation

Building networks and beyond

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

**Tiling Arrays** 

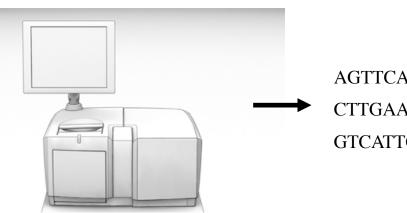


Application in a variety of contexts:

Transcription Mapping



**'06+** 

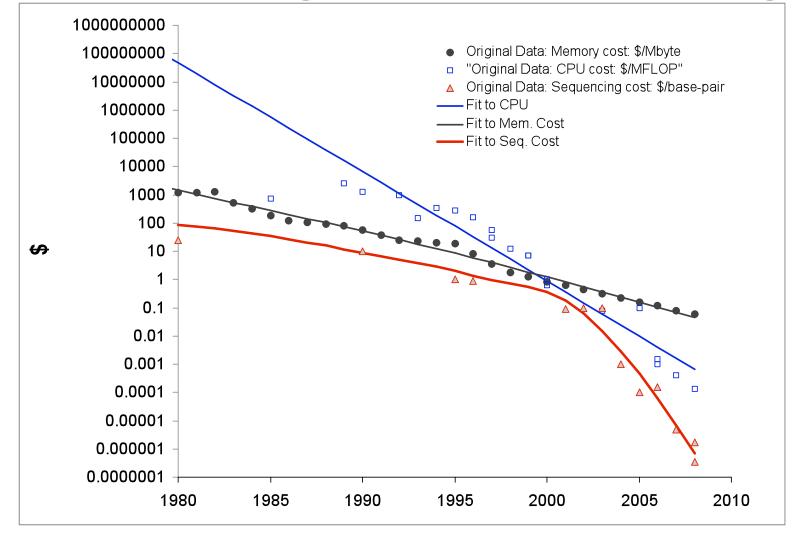


AGTTCACCTAAGA... CTTGAATGCCGAT... GTCATTCCGCAAT... DNA binding (inc. chromatin struc.)

Replication

Structural Variation

### **Plummeting Cost of Sequencing**



**11 -** Lectures.GersteinLab.org <sup>(c) 00</sup>

### **Outline**

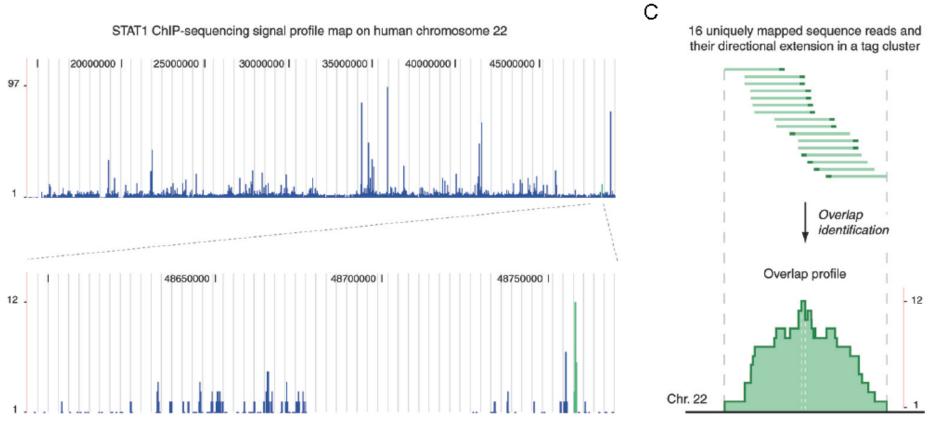


- Signal processing to call "Blocks"
  - ◊ Calling Punctate Blocks (ChipSeq)
  - ♦ Calling Broader Blocks (CNVs)
- Clustering "Blocks" into larger regions
   Ø Binding Sites
- Annotating Copied Regions in the Genome
  - $\Diamond$  SD and CNVs
  - $\Diamond$  Pseudogenes
- Integration of Pseudogenes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

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

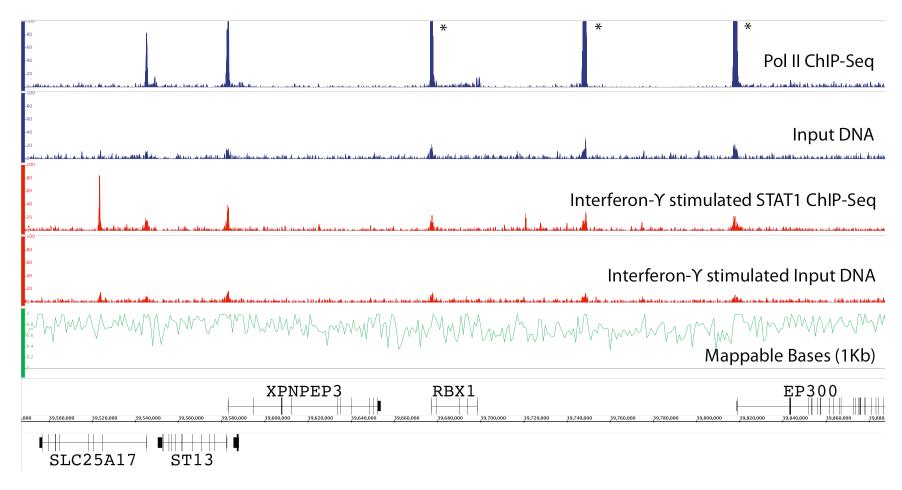


# Representative Signal from Chip-Seq

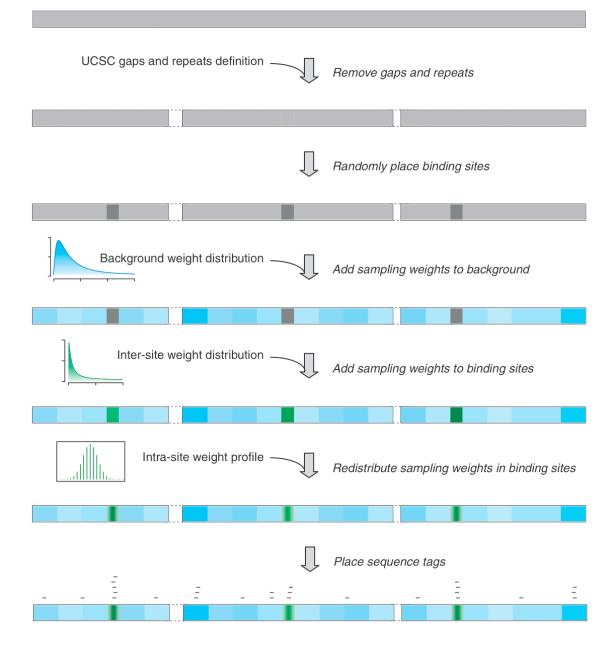


[Robertson et al., Nat. Meth. ('07); Zhang et al. PLOS Comp. Bio. ('08)]

## **ChIP-Seq vs Input DNA Control**

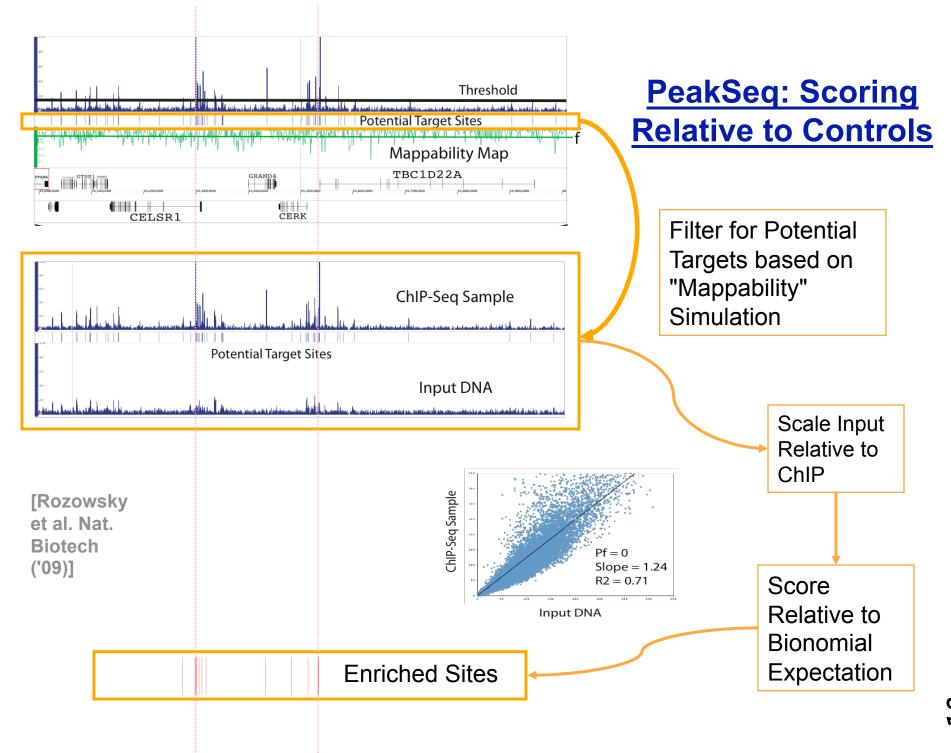


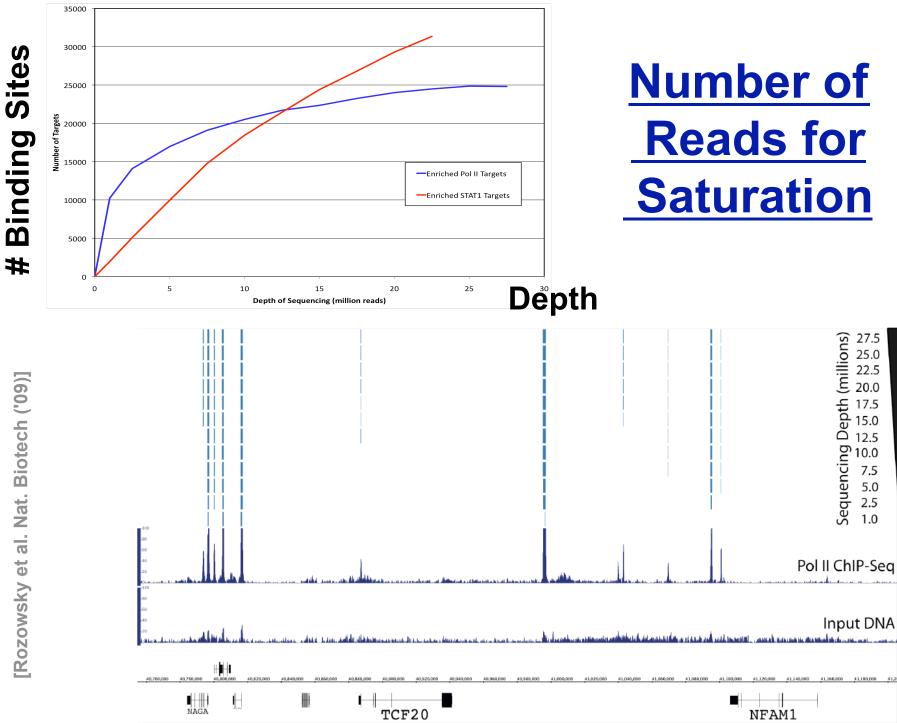
#### Genome / Genomic region



### <u>Correcting</u> <u>Chip-seq Signal by</u> <u>Simulating a Non-</u> <u>uniform Genomic</u> <u>Background</u>

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





# Signal Processing 2: Finding Variable Blocks in the Human Genome



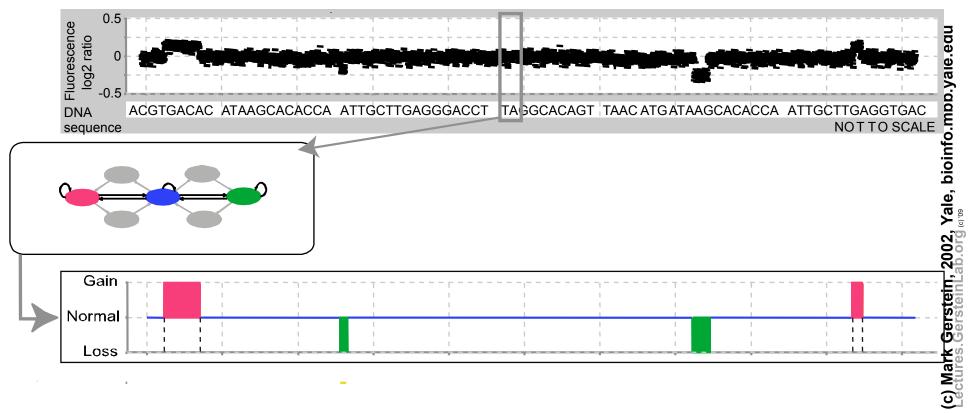
# Detection of Block Variation in Personal Genomics

- Main steps in Human Genome Resequencing
  - $\Diamond$  SNP detection
  - $\Diamond$  Haplotype phasing
  - Oeterming small indels
  - Reconstructing
     Large Structural
     Variants
     (most challenging)

- Different Techniques for SV Reconstruction
  - Segmenting Arrays and Sequencing Read-depth
  - Oiscordently placed paired-ends
  - $\Diamond$  Finding split reads
  - Obing small scale reassembly in presence of repeats

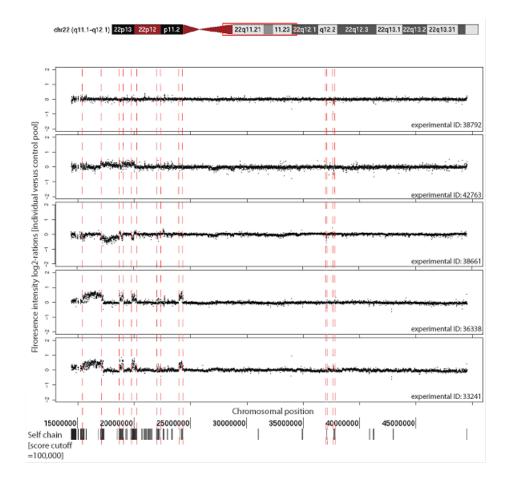
### **BreakPtr HMM**

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

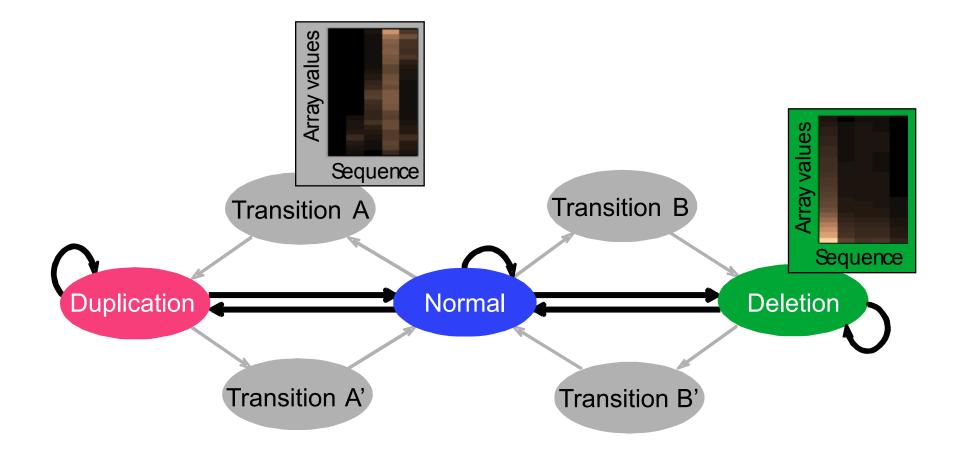


http://breakptr.gersteinlab.org

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

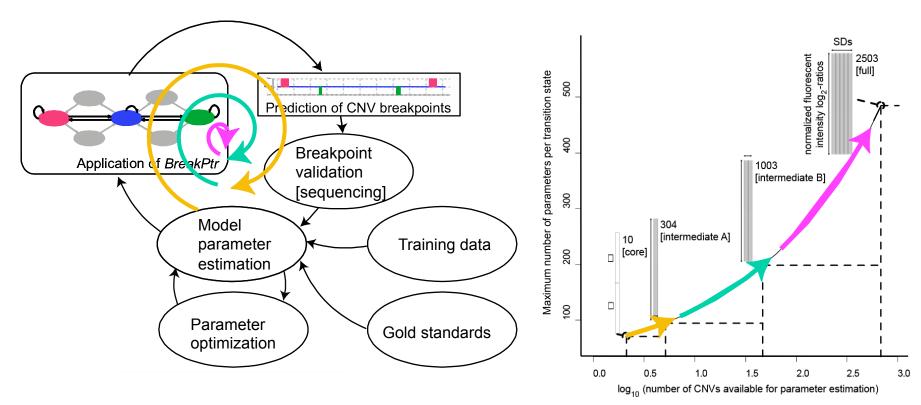


>4-fold enrichment of the breakpoints of copy number variants near segmental duplications (SDs) [e.g. Sharp *et al., Am. J. Hum. Genet.* 2005; 77:78-88]. *BreakPtr* statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)



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

### <u>'Active' approach for breakpoint identification: initial scoring</u> with preliminary model, targeted validation (with sequencing), retraining, and rescoring

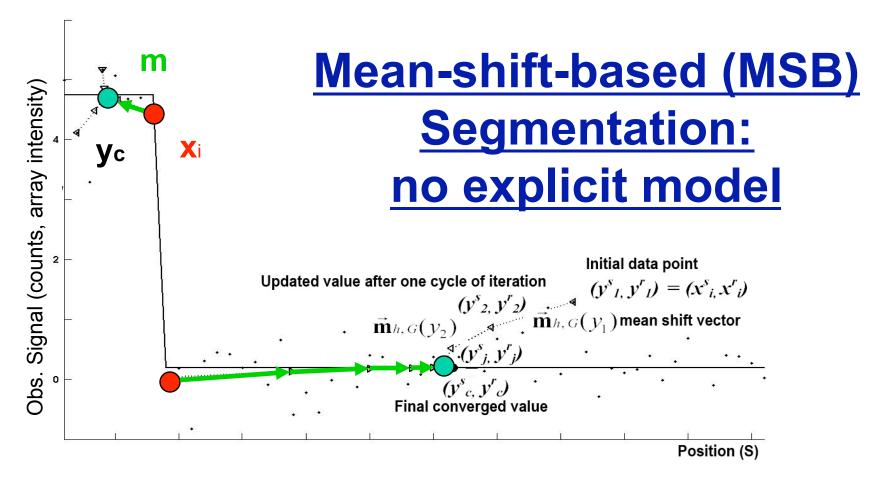


CNV breakpoints sequenced in ~10 cases following BreakPtr analysis;

#### Median resolution <300 bp

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

HMM optimized iteratively (using Expectation Maximization, EM) Korbel\*, Urban\* *et al.*, PNAS (2007)

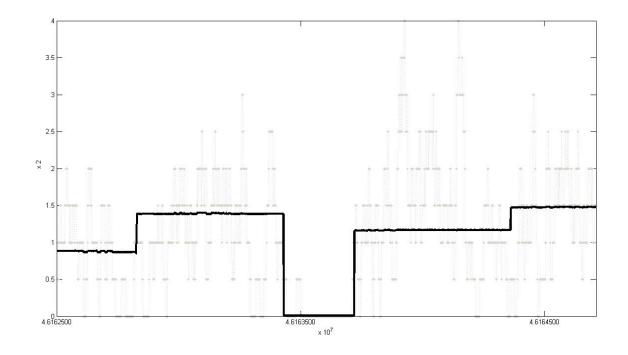


- (**x**<sub>i</sub>) Observed depth of coverage counts (or array signal) as samples from PDF
- (m) Kernel-based approach to estimate local gradient of PDF
- $(\mathbf{y}_{c})$  Iteratively follow grad to determine local modes

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

with global optimization, distr. assumption & parms. (e.g. num. of segments). Achieves discontinuity-preserving smoothing

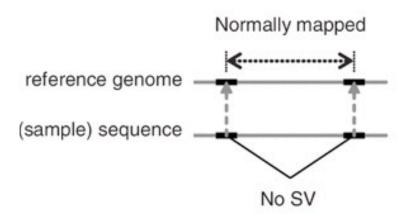
## Representative Result Showing Segmentation Based on Depth of Coverage



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)

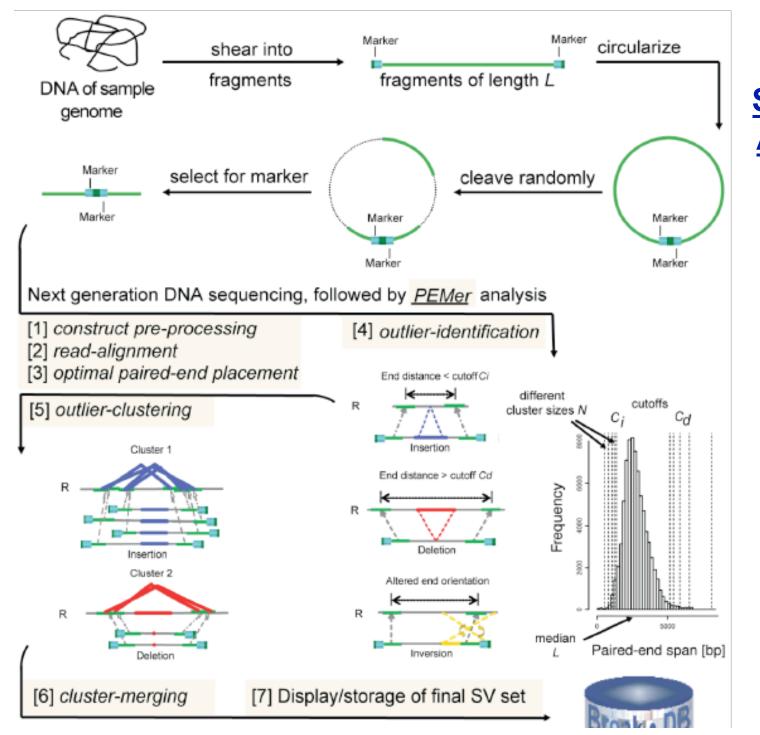
[Wang et al. Gen. Res (in press, '08)]



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

End distance < cutoff Ci

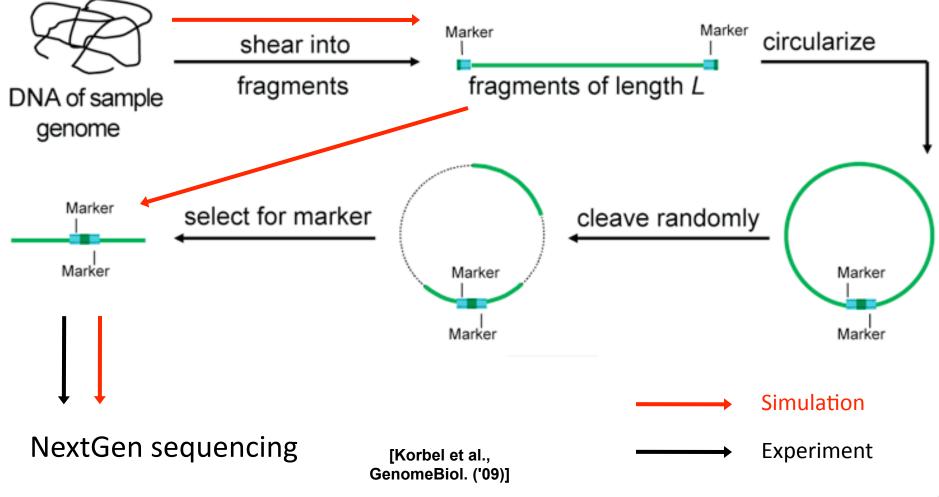
[Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]



<u>Overall</u> <u>Strategy for</u> <u>Analysis of</u> <u>NextGen</u> <u>Seq. Data</u> <u>to Detect</u> <u>Structural</u> Variants

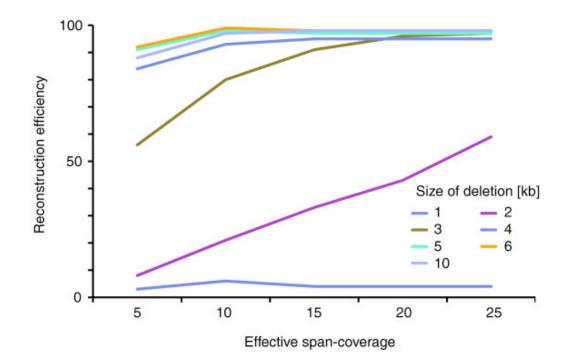
> [Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]

### **Simulation strategy**



### 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)] **36** - Lectures.GersteinLab.org

## 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

## Local Reassembly of Structural Variants

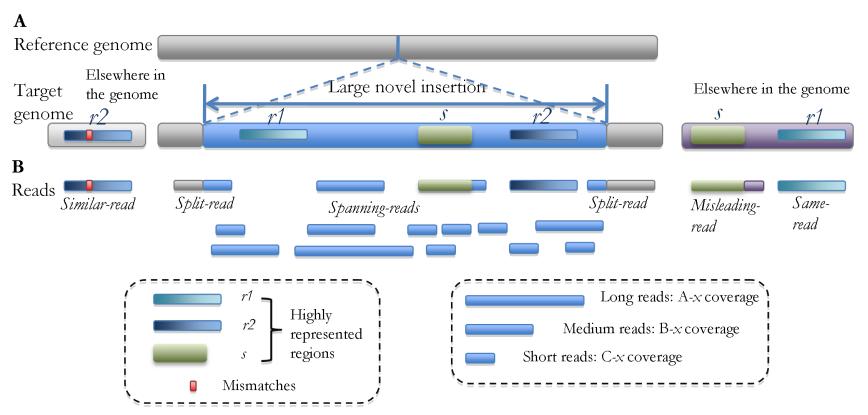
#### • Idea

- $\Diamond$  Works for long reads
- $\Diamond$  Not proven to be possible for short reads
- A Maybe small amount of long reads could be used to bootstrap re-assembly process
- Simulations
  - $\Diamond$  Expensive to test; very computational expensive to simulate
    - Each round of whole-genome assembly takes >100 CPU hrs; thus, simulation exploring 1K possibilities takes 100K CPU hr
  - Overlapped tool box for simulating reconstruction process with different combinations of long, medium & short reads (and arrays)
  - $\Diamond$  Find optimal low-cost combination with given parameters

### **Optimal integration of sequencing technologies:** *Reconstruction 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)?

#### Problem

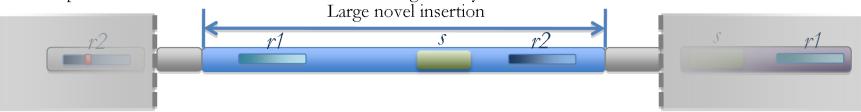


### Optimal integration of sequencing technologies: Efficient Simulation

Focus on the insertion region only

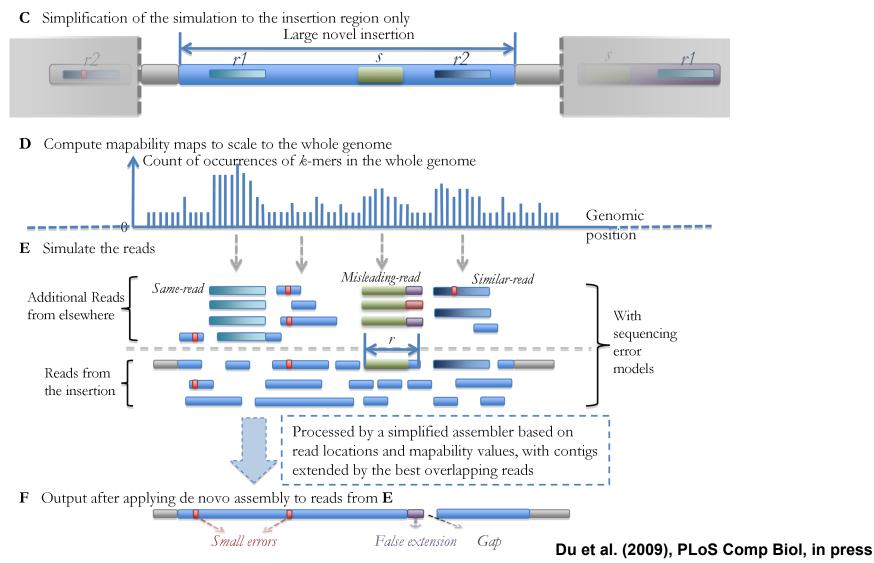
#### Simulation

**C** Simplification of the simulation to the insertion region only



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

#### Simulation



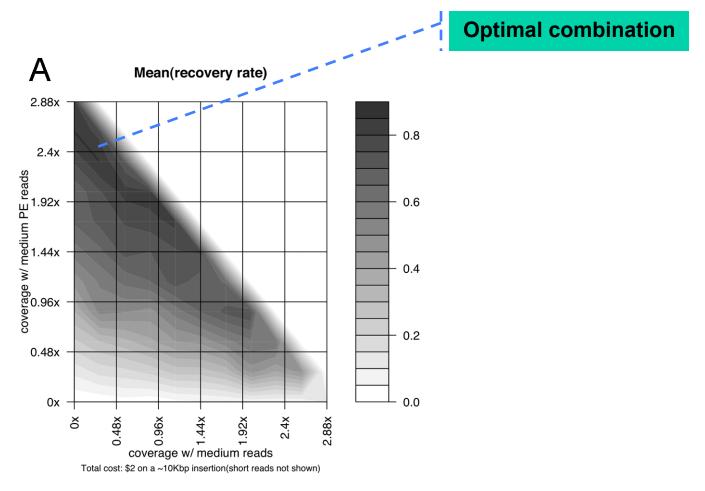
### **Tremendous Speedup**

G	Genome Size	3 Gb
С	Sequencing coverage	10 X
Ι	Novel Insert Size	10 Kb
R	(Avg) Read Length	50 bp
m	Avg. mapability of sub-seq.	3
	in insert	

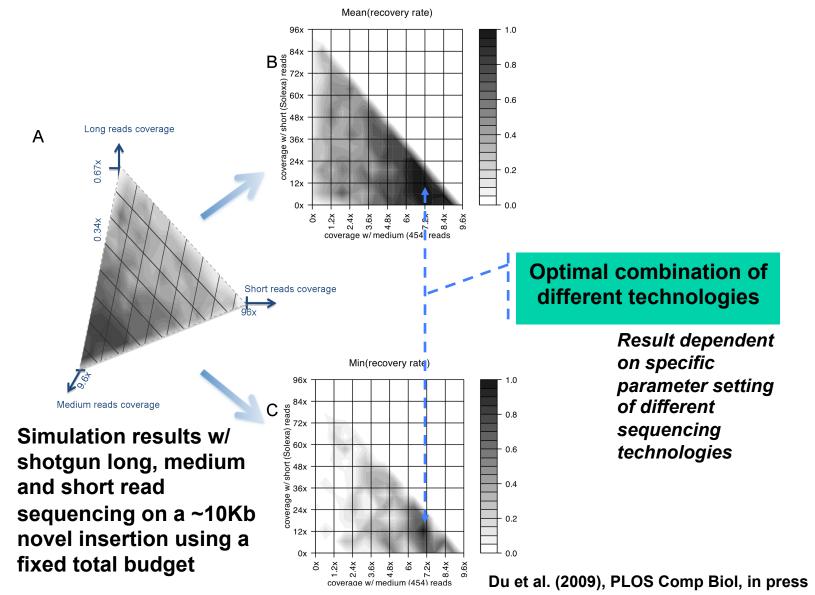
Number of reads generated for the reconstruction of a novel insertion		
Whole genome sequencing + hybrid assembly	O(Gc/R)	
Simulation w. mapability (just using insertion)	O(Icm/R)	
Reduction in complexity (fold)	~100K	

### Optimal integration of sequencing technologies: Simulation shows power of PEs

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



### <u>Optimal integration of sequencing technologies:</u> <u>Simulation shows combination better than single technology</u>



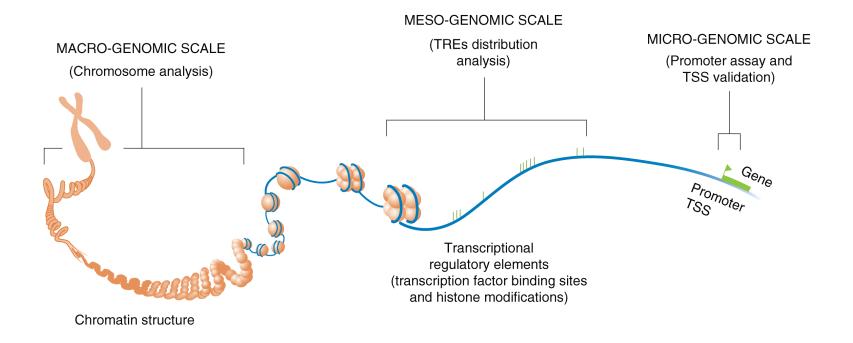
4



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

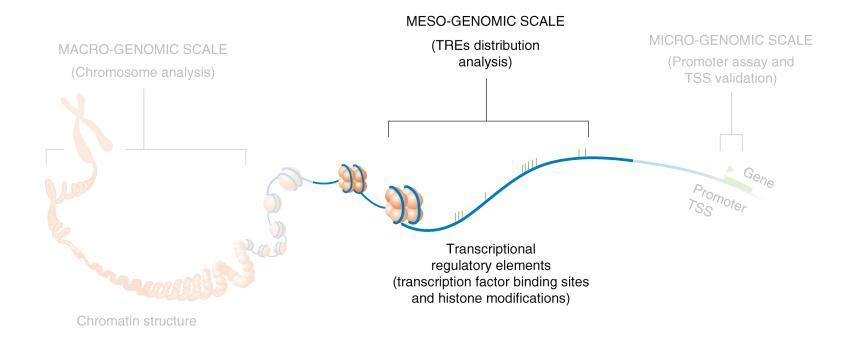
pers. **photo**, see streams.gerstein.info

## TRE analysis on the microgenomic scale



**4***B* - Lectures.GersteinLab.org<sub>@</sub>

## Clustering Binding Sites at ~50kb resolution



### Landscape of ENCODE Transcriptional Regulatory Elements

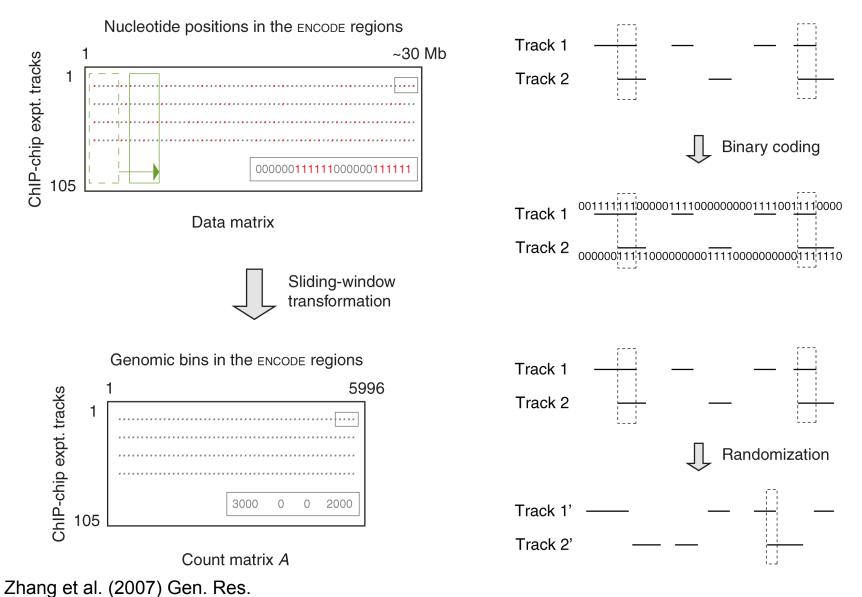
- Analyzed 105 lists of transcriptional regulatory elements in the encode regions
- 29 transcription factors, 9 cell lines, 2 time points

 $\Diamond \mathsf{RNA} \ \mathsf{Pol2}$ 

- Object to the second descent for the second descent descen
- $\langle \rangle \text{Core promoters}$
- Others such as enhancers, silencers, insulators, & response elements

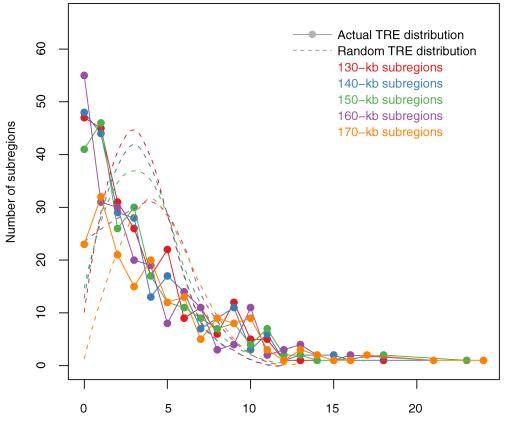
[CFTR] ENm001	╷╴╋┙╶┞╶┼╷┧╴╶╴┶┼╶┰┟┨╓┼╘╈┙╖┪╴┧┸╄┼╴╴╴╋╴╓╌╄╖╴╶┼┸╋╖╓╌┉╈┿╌┍╴╶┶┅╍┉╅╴╷┼╶╌┾┅╌┼╷╌╷╓╴╴╴┼╴┎╌┼╴┨╴╴╴╴┼╴┲╌╴					
[Interleukin] ENm002						
[Apo] ENm003						
[Chr22] ENm004	╾╌┰╶╫┙┍┿╓┙┍┱╶╖╬╋╪┲╴╴┶┫╴╴╾┍╴╴╴┍╴╓╢╅╌┉┽╌╗╋╌╢┥╌╢┿╌╢┿╌╢┿╌╗┽┥╋╌┥┥╌╎┙╢┥┥╋╝╋╋╸╴╴╴╴┼╖╸╴╴┧╢╢┙					
[Chr21] ENm005	<del>╽╡┫╹╝╕╗╷╵╶╎╢╴╝╎╡┛┍╢╎╢╎╎╎╎╢╹╎╎╖╝╎┧╹╍┍╵╷<mark>╢┛╘┉</mark>╧┍╽┤╶╎╧╘┉┉╸╘╸╴╞╸╞╴╘<mark>┝</mark>╌╖┍╷╴╷╎╴╘┙┙╍╶┍┼╧╍┺╶┲╸╼┙╴</del>					
[ChrX] ENm006	┴╷ <u>┪┝╉╝┥╋┎╷╓╴╴╴┤╴╶╵╷╎╶╎╴┥┠╫┺╝┨┧┾╴──┟┝┧╵┎╢╴┍╅╎╖╷╴─┤┝┼┧╵╽╴╴╴╴╴╴╴╴╴╴</u>					
[Chr19] ENm007	<u>╶┼╖╋╃╫╝╸╷┧╕┍╷╝┼╷╴╶╷┧┑╴┫╶╷╪╝╽╎╓╴╬╬╝┨╢╢╊╋┼╶┙╸┧┍┽╢╢╝╅┶┼┼╫┧╶┫╷┩╴╼╿╵╙╋╫╋</u> ╉╎┦					
[α−globin] ENm008						
[β−globin] ENm009		┞──┼╽┟╷┼╹┎┎╺╸╺┝┼┼╴				
[HOXA] ENm010		Zł	nang et al. (2007) Gen. Res.			
[IGF2/H19] ENm011						
[FOXP2] ENm012	· · · · · · · · · · · · · · · · · · ·					
[7q21.13] ENm013						
[7q31.33] ENm014		┼┼┼┼┶╴╫╷╴┶╫				
		ENr121 + + + + + + + + + + + + + + + + + +				
		ENr122				
		ENr123 ++				
ENr114	├ <u></u> └ <u></u> └ <u></u>					
ENr211		ENr221	ENr231			
ENr212		ENr222				
ENr213		ENr223	ENr233			
ENr311	<del></del>	ENr321	ENr331			
ENr312	<del>┍╶╶╽<mark>╓</mark>╶╴┍╶┙┇╹╷╿╴┍╗</del>	ENr322	ENr332			
ENr313		ENr323	— ENr333 - ENr333			
		ENr324	— ENr334 <b>411, 11, 11, 14, 14, 14, 14, 11, 14, 14,</b>			
			í			
	1 500000	1000000	1500000 bp			
		1000000	, 300000 bp			

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



## **Non-random distribution of TREs**

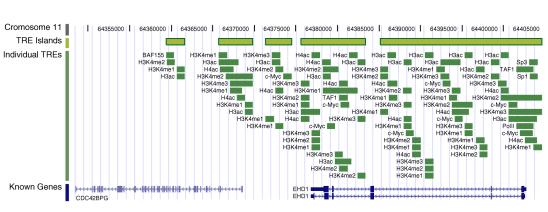
- TREs are not evenly distributed throughout the encode regions (*P* < 2.2×10<sup>-16</sup>).
- The actual TRE distribution is power-law.
- The null distribution is 'Poissonesque.'
- Many genomic subregions with extreme numbers of TREs.

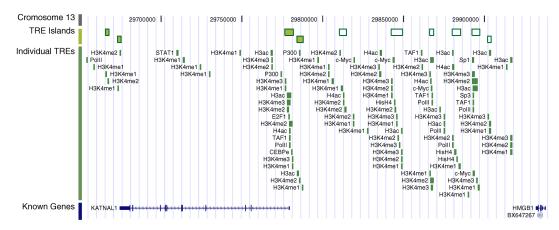


Number of TREs in a subregion

### Local TRE enrichment and depletion: Annotation of Desserts and Forests

- Hundreds of TRE 'forests' and 'deserts' are identified in ENCODE regions.
- The entirety of *ehd1* on chromosome 11 is covered by TRE islands.
- Some of islands are located in the intergenic regions in the genome.





dart.gersteinlab.org/encode/tr/

## Biplot to Show Overall Relationship of TFs and Genomic Bins

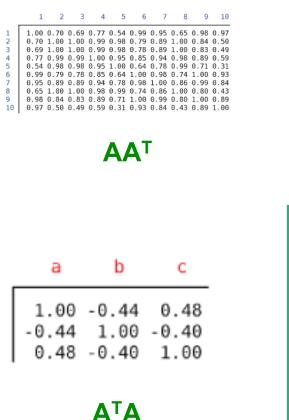
	「Fs: a, b, c										
50kb Genomic Bins: 1,2,3											
	1	2	3	4	5	6	7	8	9	10	
	16	18	17	14 19 33	23	14	21	18	13	10	
A=USV <sup>T</sup>											
					a	b	с				
	<b>Ą</b> ⊤		1 2 3 4 5 6 7 8 9 10	1 1 1 2 2 1 1	.4 1 .4 1 .7 2 .7 2 .7 1 .2 1 .5 1 .8 1	16 2 18 2 17 2 19 3 23 2 14 3 21 3 21 3 21 3 21 3 21 3 21 3 21 3 21	25 22 33 28 34 30 22 36				

5

а

b

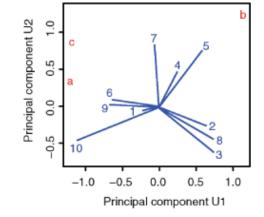
С

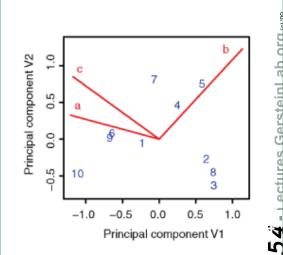


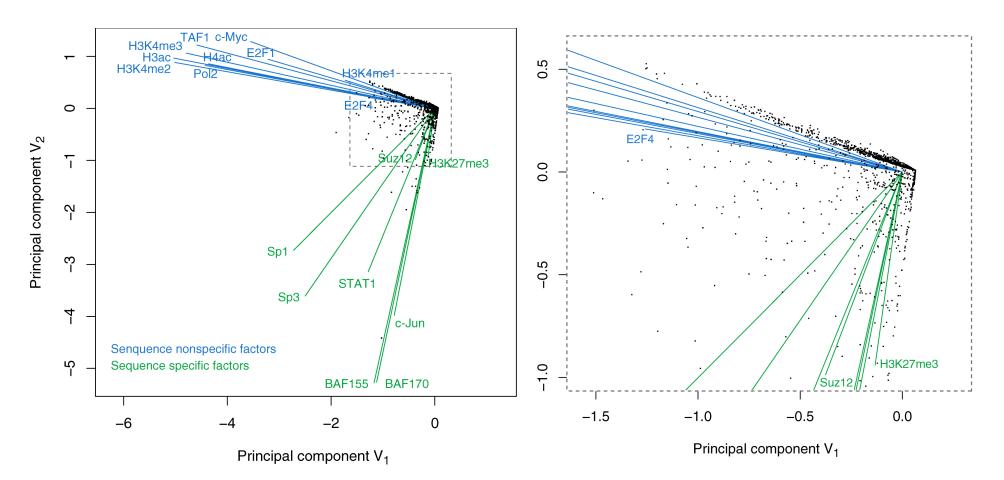
а

b

С









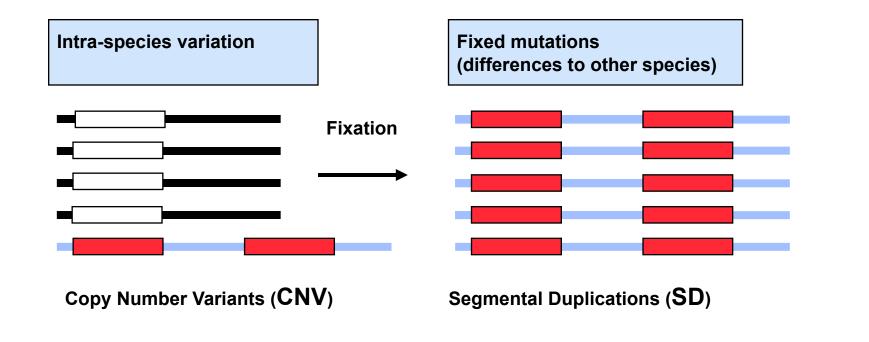
Zhang et al. (2007) Gen. Res.

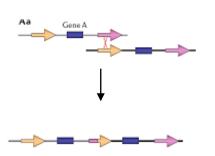
- Biplot groups TFs into sequence-specific and sequence-nonspecific clusters.
  - $\diamond$  c-Myc may behave more like a sequence-nonspecific TF.
  - A H3K27me3 functions in a transcriptional regulatory process in a rather sequence-specific manner.
- Genomic Bins are associated with different TFs and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

# Analyzing Repeated Blocks in the Genome (SDs & CNVs)



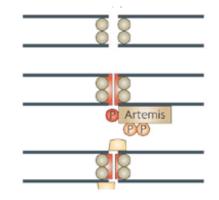
### SEGMENTAL DUPLCATIONS AND COPY NUMBER VARIANTS ARE RELATED PHENOMENA AND HAVE BEEN CREATED BY SEVERAL DIFFERENT MECHANISMS





### **NAHR** (Non-allelic homologous recombination)

Flanking repeat (e.g. Alu, LINE...)

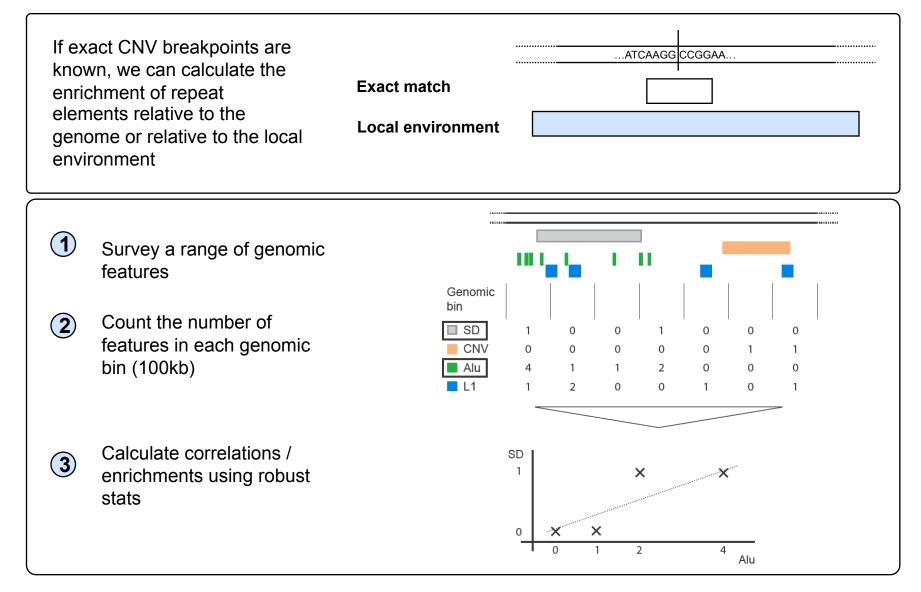


### NHEJ

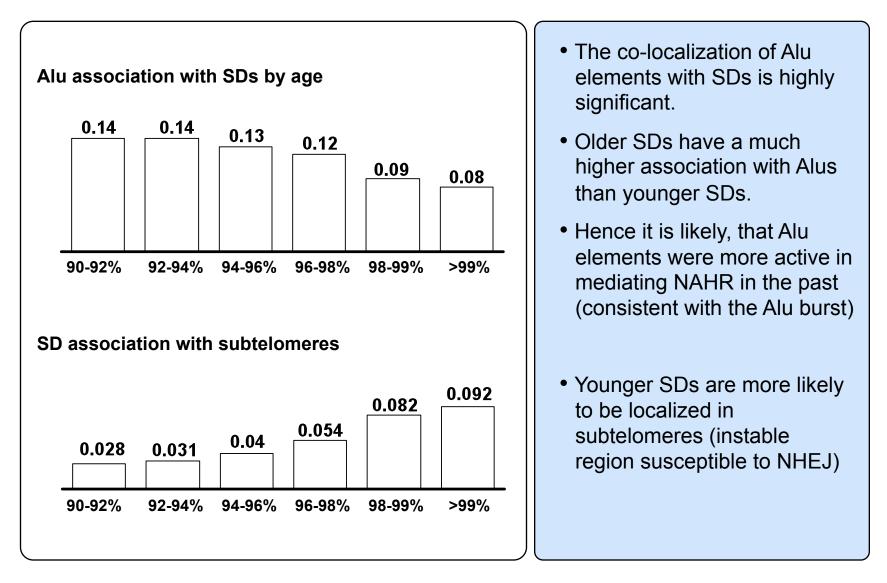
(Non-homologous-endjoining)

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

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

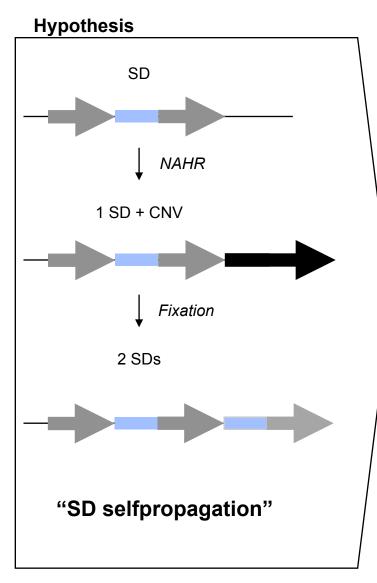


### OLDER SDs ARE MUCH MORE LIKELY TO BE FORMED BY ALU ELEMENTS



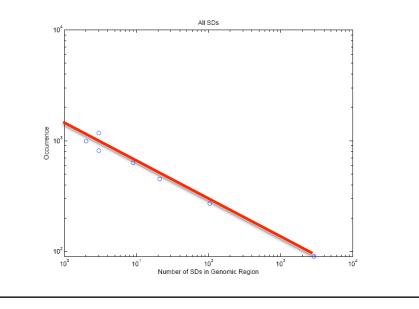
[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1 ]

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

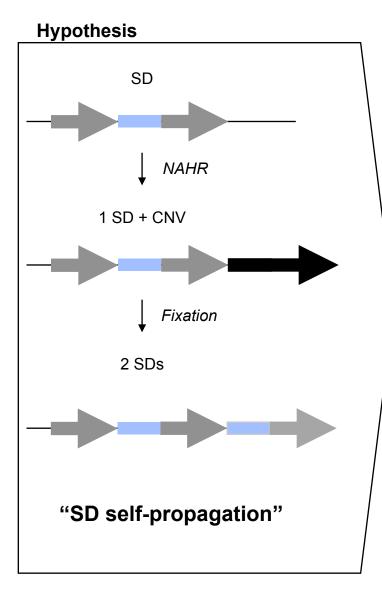


#### Corollary

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

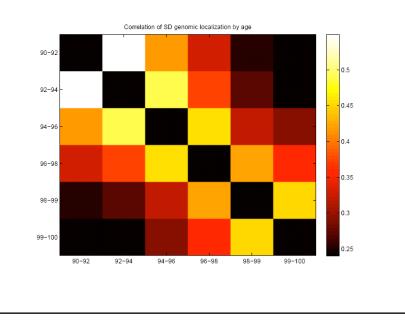


### FOCUSSING ON SDS: SDs COLOCALIZE WITH EACH OTHER



#### Corollary

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

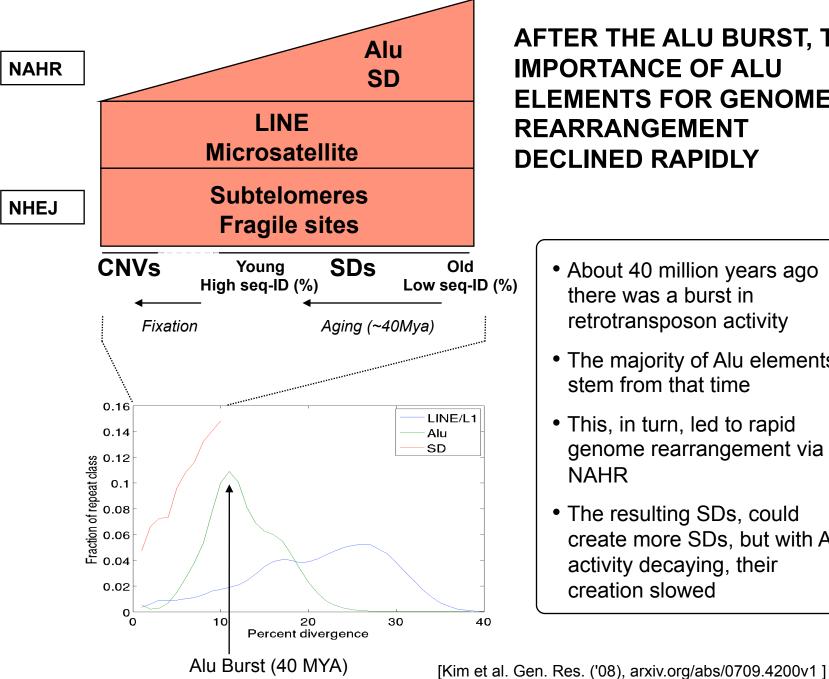


[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1]

### **ASSOCIATIONS ARE DIFFERENT FOR SDs AND CNVs**

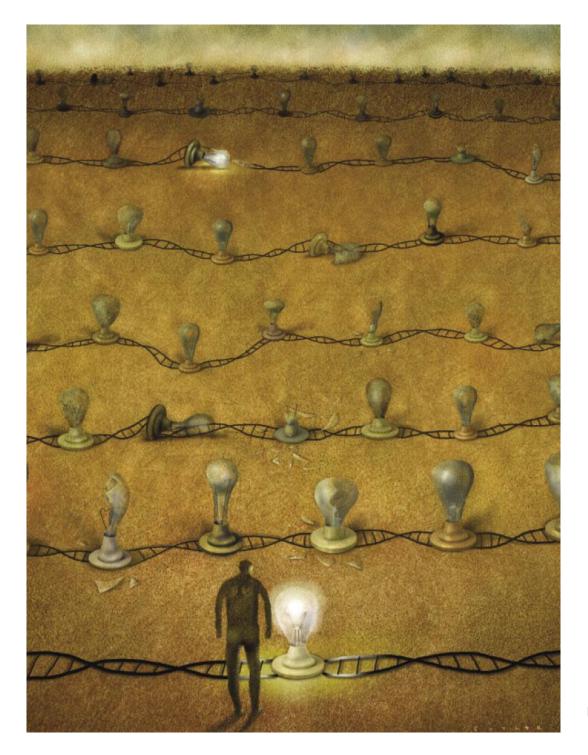
#### **CNVs ARE LESS ASSOCIATED WITH** SD association with repeats **SDs THAN THE GENERAL SD TREND** 0.27 CNV 0.21 0.094 Association 0.07 with SDs Alu Microsatellite Pseudogenes LINE 0.31 < 0.001 (<0.001) 0.001 0.046 0.11 **CNV** association with repeats 0.0739 0.048 0.0466 0.0006 >99% SDs\* CNVs Microsatellite Pseudogenes LINE Alu < 0.001 0.046 0.92 0.001

[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



### Integrative Analyses:

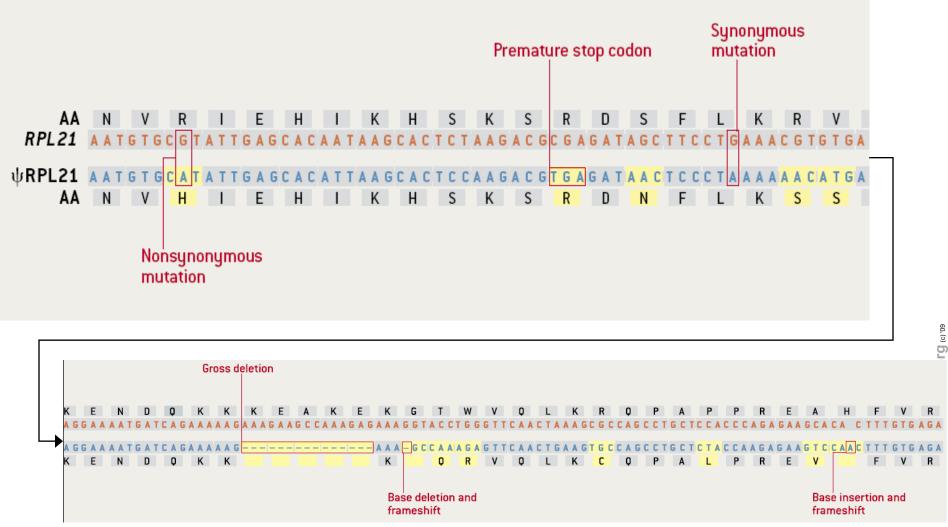
Annotating Pseudogenes and relating them to functional signals and measures of conservation

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

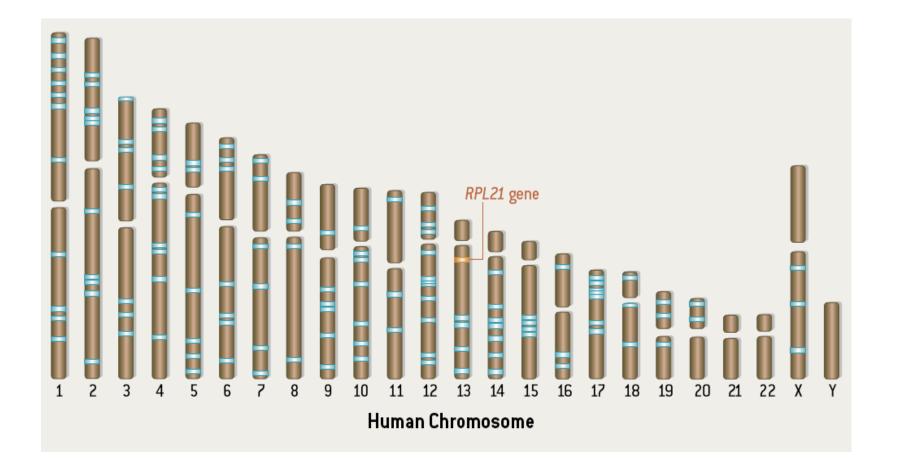
## <u>Pseudogenes are among the most</u> <u>interesting intergenic elements</u>

- Formal Properties of Pseudogenes ( $\Psi$ G)
  - $\Diamond$  Inheritable
  - $\Diamond\,$  Homologous to a functioning element
  - $\Diamond$  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?...

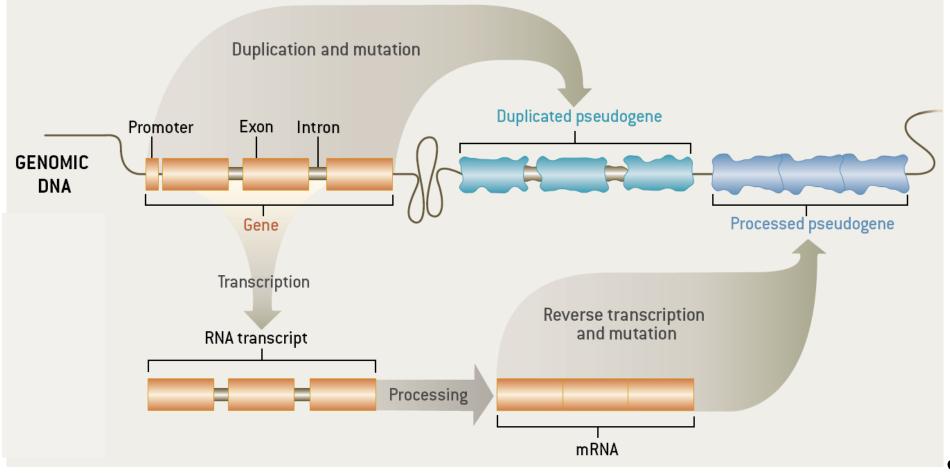
## Identifiable Features of a Pseudogene (ψRPL21)



Distribution of Human Pseudogenes (for RPL21) across the chromosomes



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



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

## **Overall Flow:**

## **<u>Pipeline Runs, Coherent Sets,</u> Annotation, Transfer to Sanger**

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

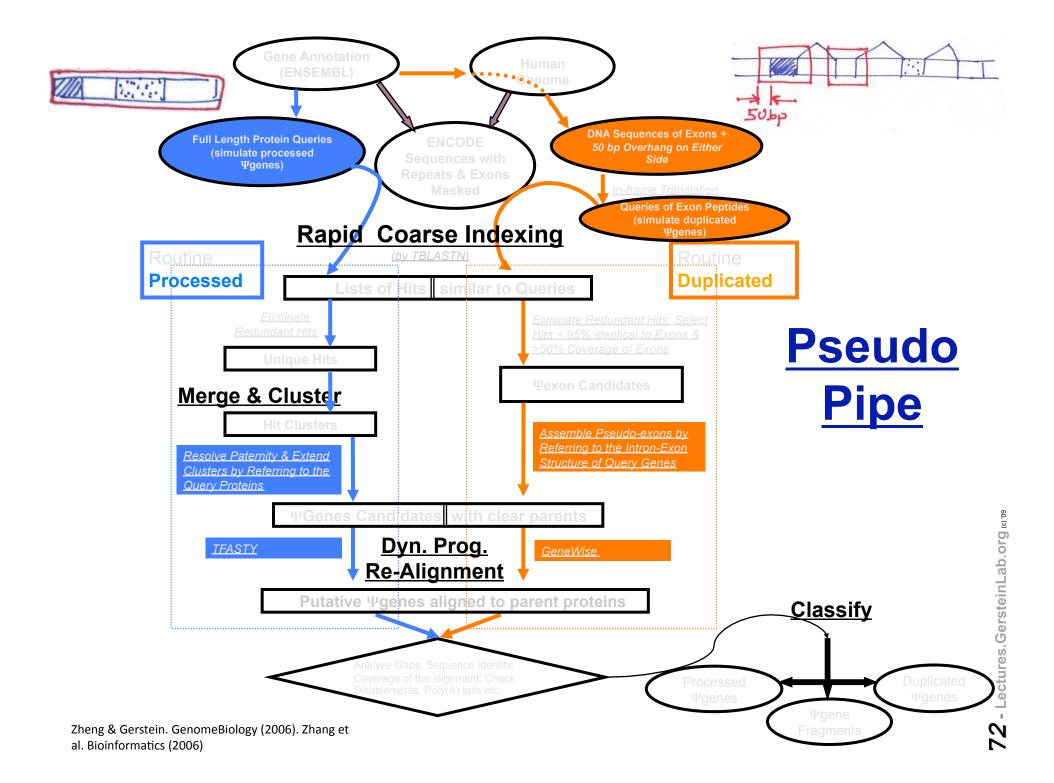
- Automatic pipeline currently gives ~23K [pseudogene.org]
- Chronology of Sets
  - 1. Encode Pilot 1%
  - 2. Unitary pseudogenes (Hard)

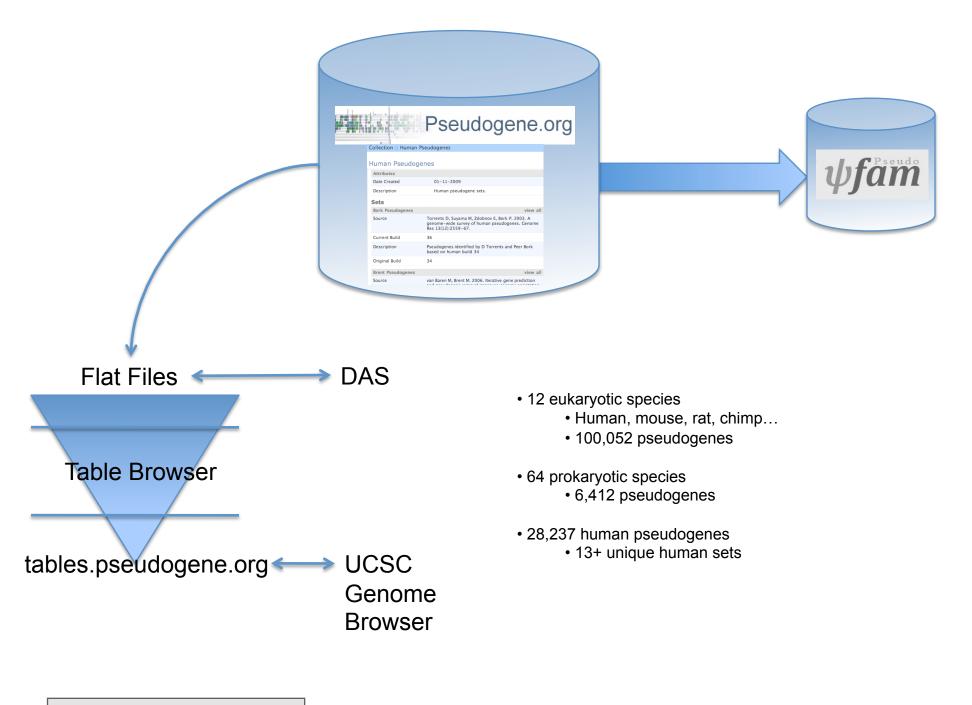
# 3. Ribosomal Protein pseudogenes

- 4. Glycolytic Pseudogenes
- 5. ....

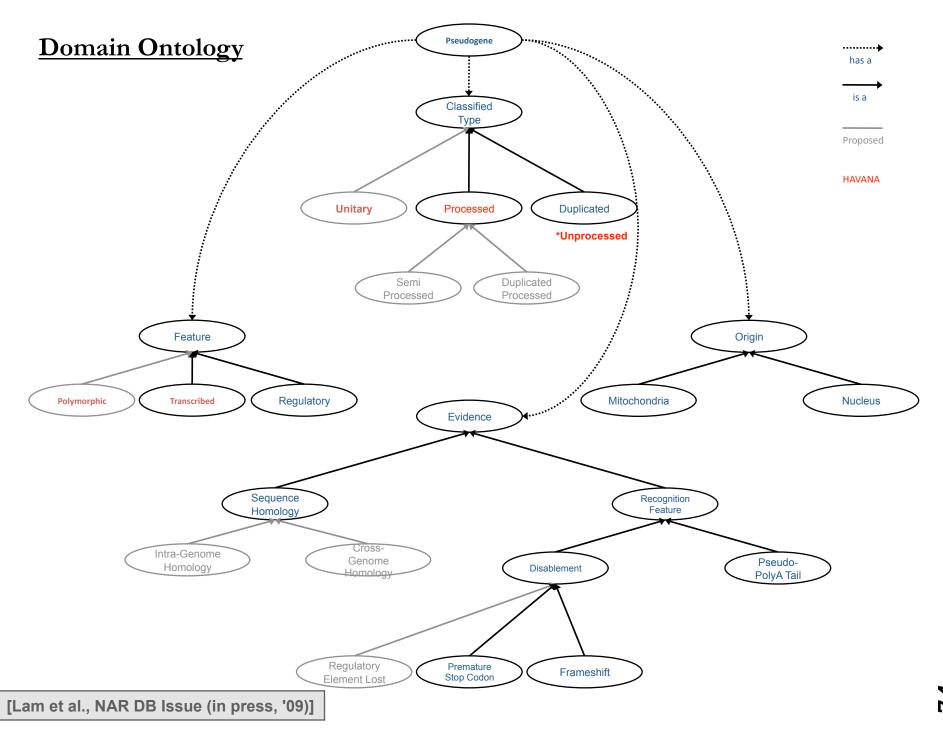


# Pseudogene Tools: Assignment Pipeline & DB





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



## **Pseudofam Construction**

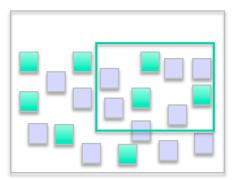
- Data Generation
  - Identify pseudogenes by proteins and map parent proteins to protein families

### • <u>Alignment</u>

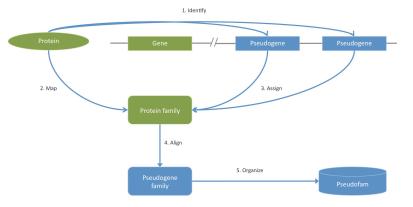
- Align pseudogene to parent
- ♦ Transfer alignment from Pfam
- Combine and adjust the alignments to build the pseudofam alignment

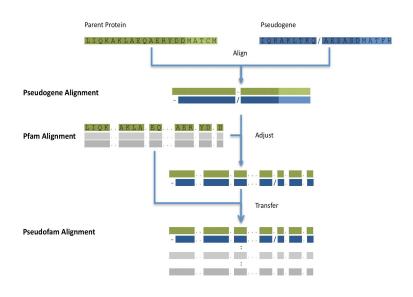
### • <u>Statistics</u>

♦ Enrichment



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





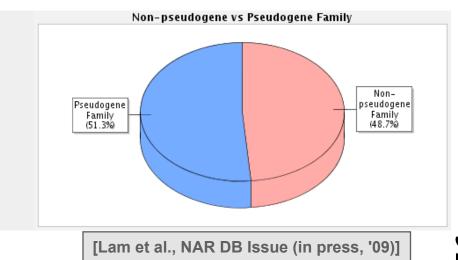
### **Pseudofam Statistics: Enrichment of** pseudogenes within a family ("Living vs Dead")

### Total (10 Eukaryotes)

Protein Families:	3,820	Non-pseudogene vs Pseudogene Family
Froten rannes.	5,620	Non-
Pseudogene Families:	2,985	pseudogene Family
Total Genes:	219,662	(21.9%)
Total Parents:	26,679	
Total Pseudogenes:	102,679	
Pseudogene-to-gene Ratio:	0.47	Pseudogene Family
Pseudogene-to-parent Ratio:	3.85	(78.1%)
Parent-to-gene Ratio:	0.12	

#### Human

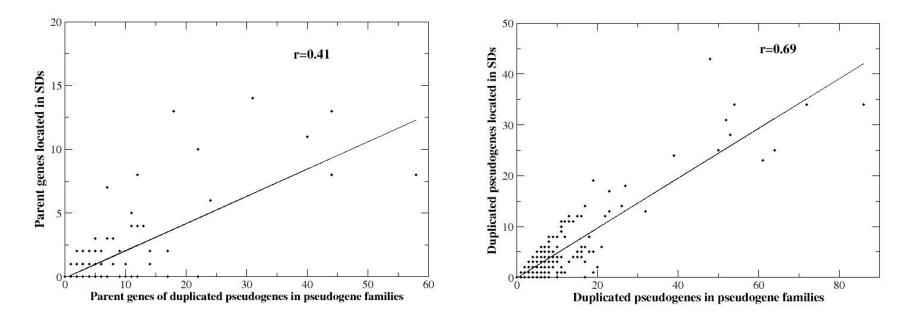
Protein Families:	3,486
Pseudogene Families:	1,790
Total Genes:	34,686
Total Parents:	4,218
Total Pseudogenes:	12,534
Pseudogene-to-gene Ratio:	0.36
Pseudogene-to-parent Ratio:	2.97
Parent-to-gene Ratio:	0.12





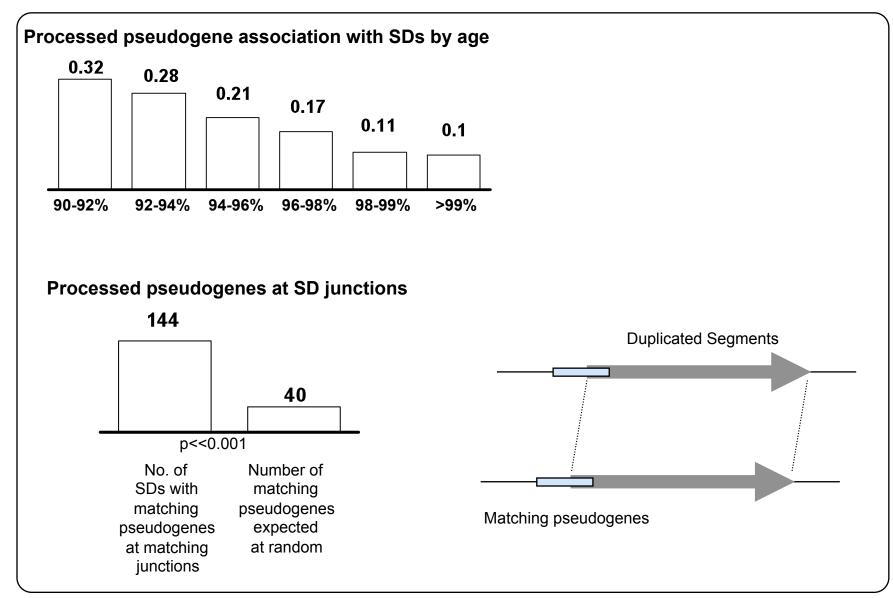
Relationship between pseudogenes and CNVs

### **Pseudogene families and Segmental Duplications (SDs)**



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

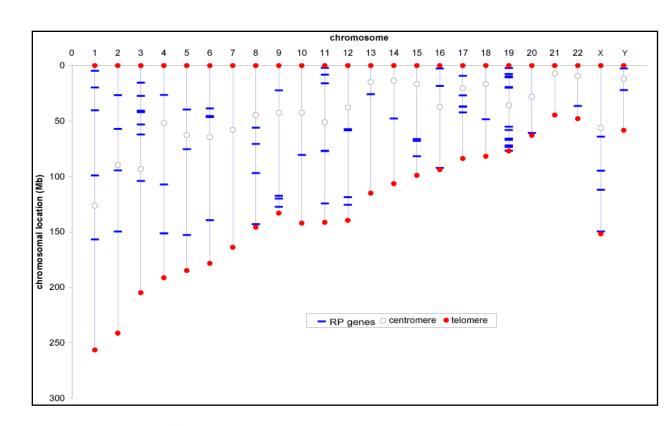
# ANOTHER FUNCTION FOR PSEUDOGENES: SERVING AS REPEATS FOR MEDIATING NAHR

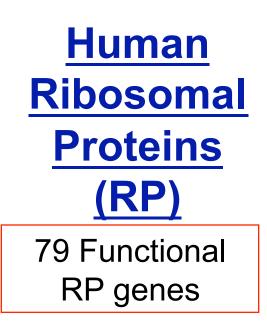


[Kim et al. Gen. Res. (submitted, '08), arxiv.org/abs/0709.4200v1 ]

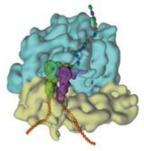
## Pseudogene Set #2: Ribosomal Protein Pseudogenes (Large Number)



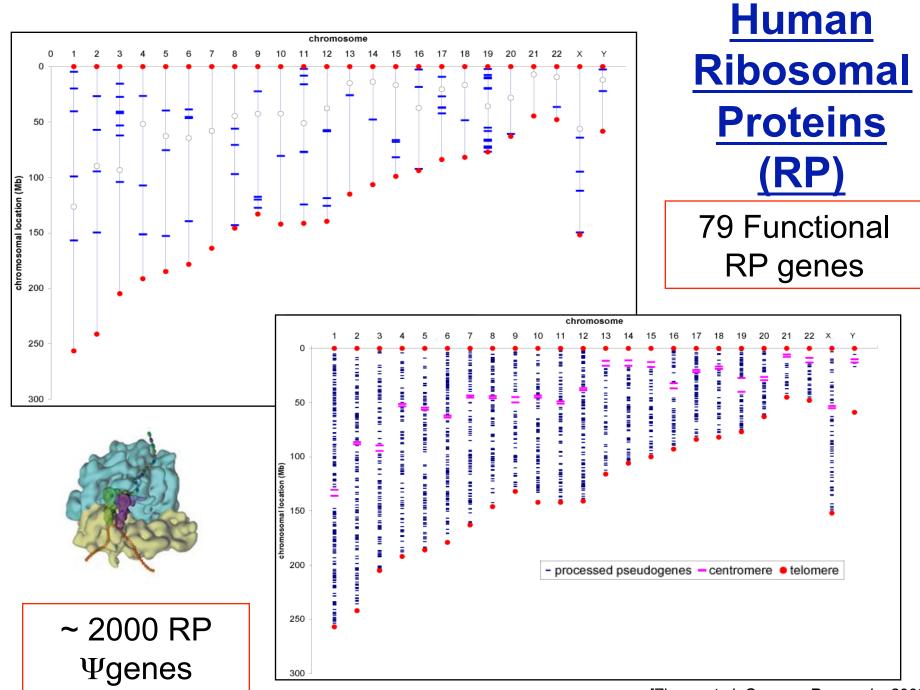




Nakao A, Yoshihama M, Kenmochi N: RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res 2004, 32:D168-170.* 



82



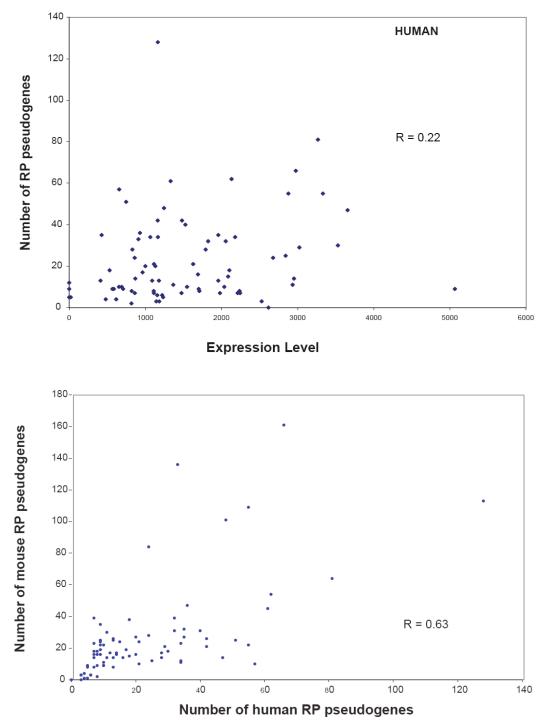
[Zhang et al. Genome Research, 2002]

# Number of RP pseudogenes

(identified by pipeline)

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

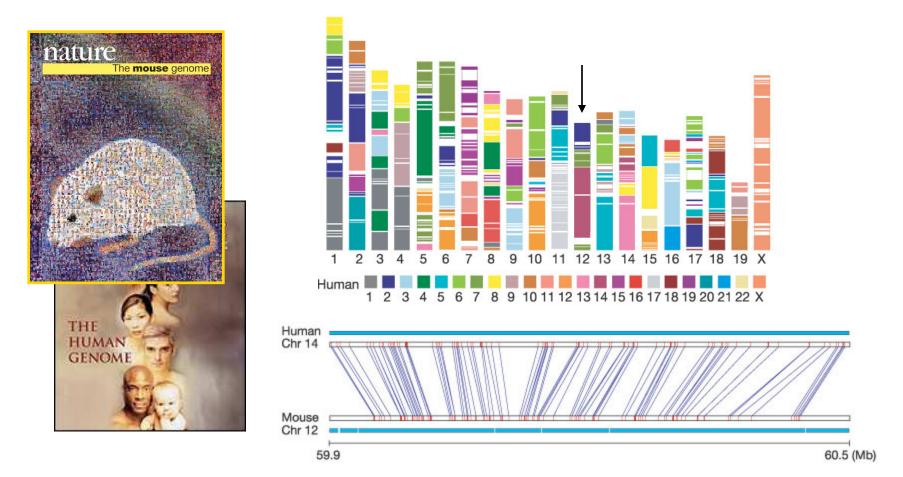
RP pseudogenes constitute the largest family of pseudogenes. Almost all are processed: There are ~90 clearly duplicated ones in the human genome



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

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

### **The Synteny Between Mouse and Man**



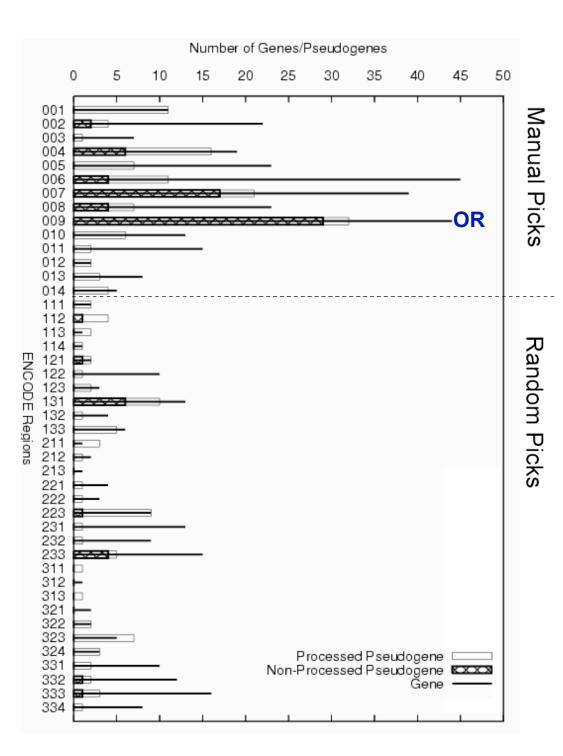
- About 90% of the mouse and human genomes are in syntenic blocks.

# **Syntenic proc pseudogenes**

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

# ENCODE Pilot Pseudogenes: Integration of Different Types of Annotation





Overall Results: Regional Distribution

### 201 pseudogenes 77 non-processed 124 processed

Zheng et al. (2007) Gen. Res.

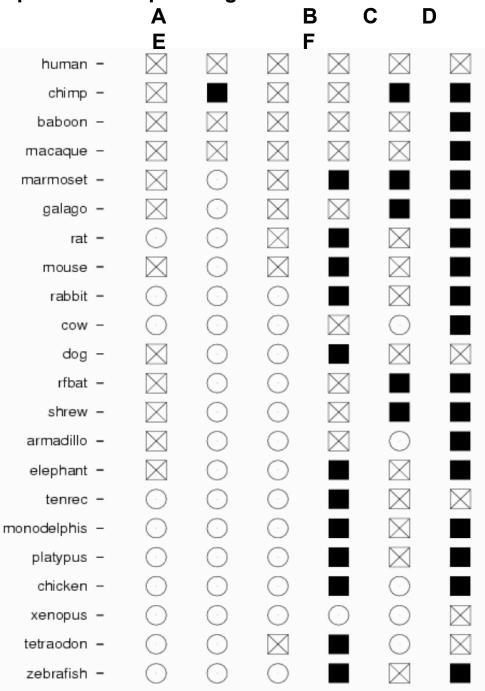
browser + pseudogene.org/ENCODE

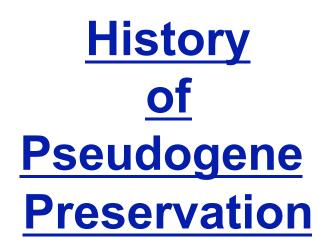
### Vast Amounts of Different Data Types to Integrate in pilot ENCODE

- Determining experimental signals for biochemical activity across each base of genome
- Large-scale sequence comparison in relation to the human genome

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

#### representative pseudogenes drawn from 201 total





#### Based on alignment from ENCODE MSA group

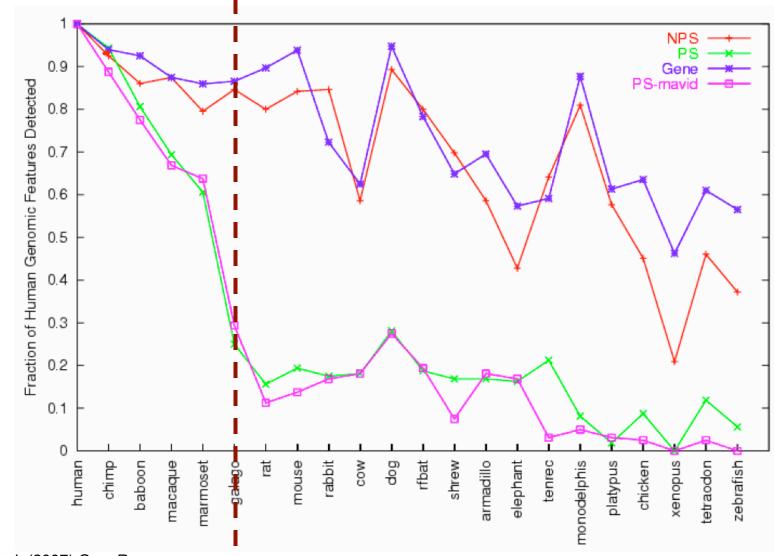
Zheng et al. (2007) Gen. Res.

Absent

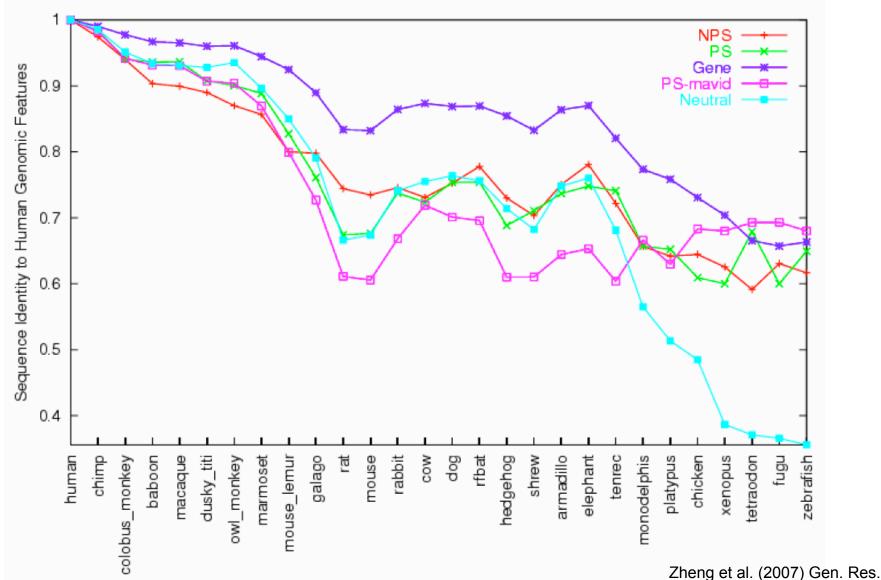
**Present with Disablement** 

**Present without Disablement** 

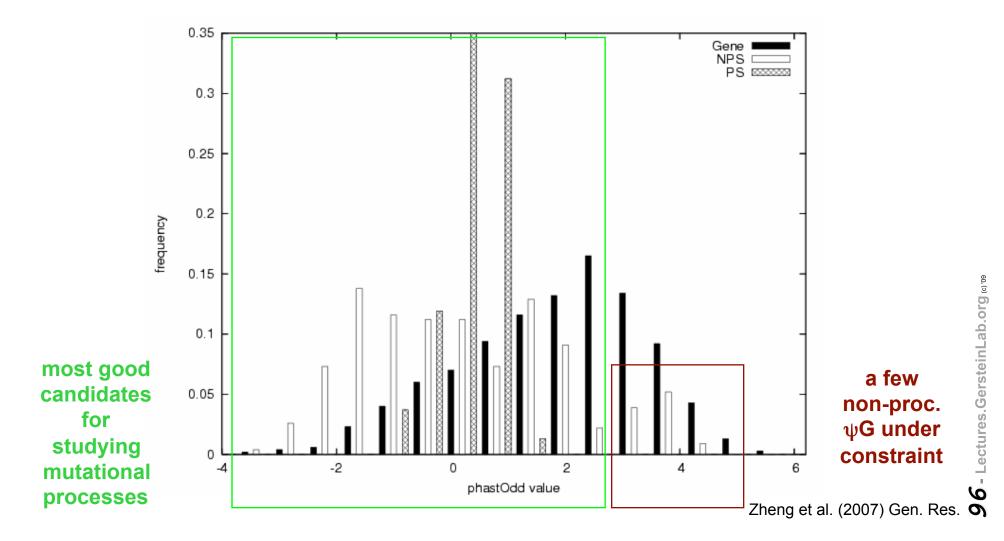
### <u>Most Processed Pseudogenes are Primate</u> <u>Specific Created by Recent (<45 MYA)</u> <u>Retrotranspositional Activity</u>

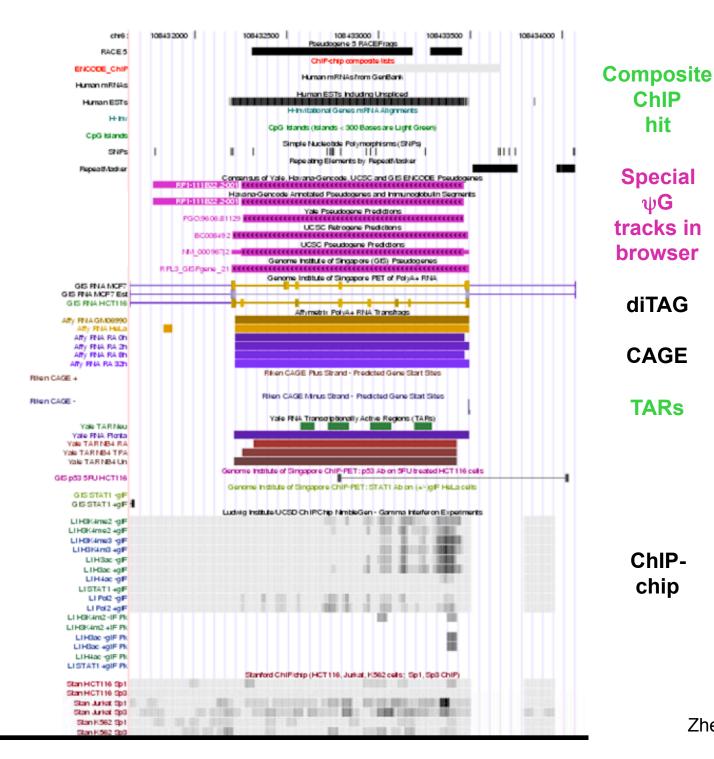


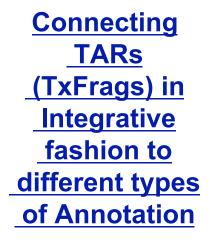
# <u>Sequence Decay of Pseudogenes,</u> <u>Approximately Neutral</u>



# Using phastOdd value to examine neutral evolution of pseudogenes







- 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

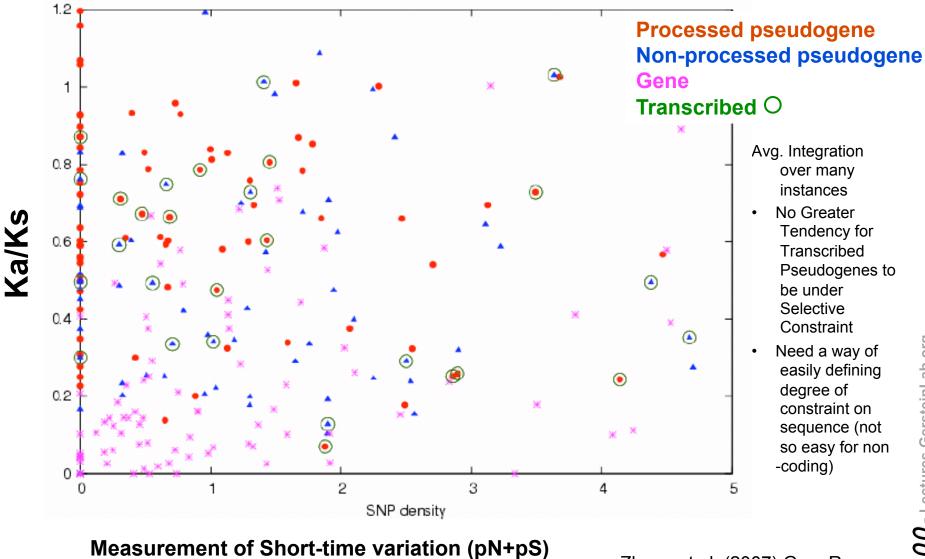
	TAR / transfrag	CAGE	DiTag	RACEfrag	EST / mRNA
TAR / transfrag	105 *	8	2	5	14
CAGE		8	1	0	1
DiTag			2	0	0
RACEfrag				<u>14</u>	5
EST / mRNA					21

**Excluding TARs (due to cross-hyb issues)** 

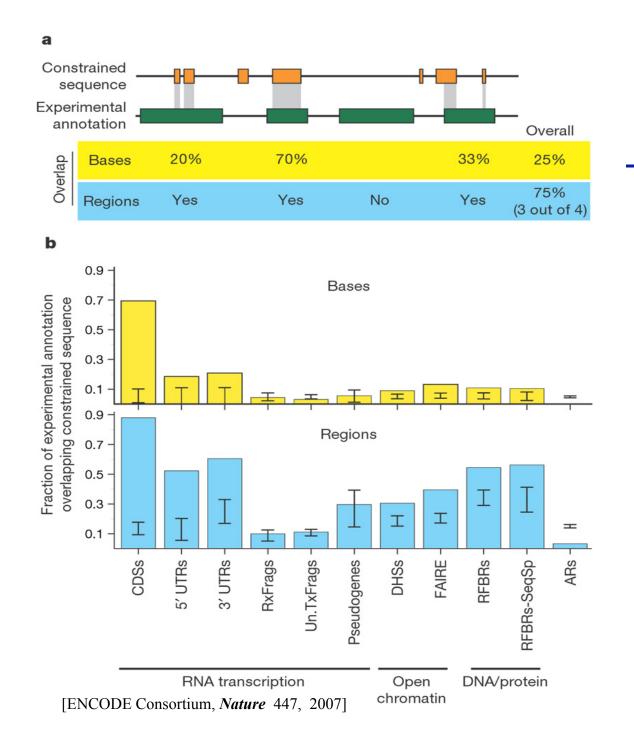
Targeted RACE expts to 160 pseudogenes, gives <u>14</u>

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

### **Integrating Transcriptional Evidence with Gene Annotation and Sequence Constraints**



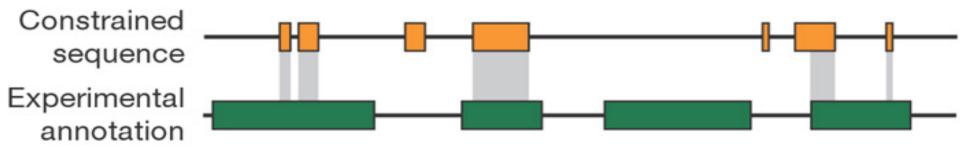
Zheng et al. (2007) Gen. Res.



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

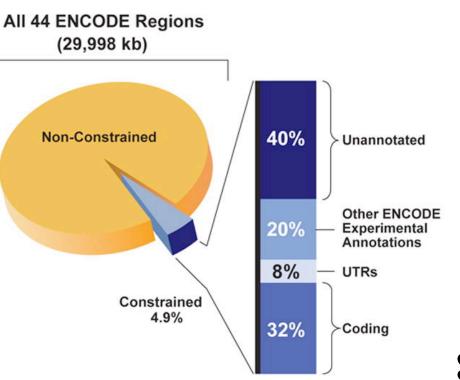
- Integrating & averaging results over larger and larger sets
- Comparison of integrated quantities

# Grand Summary: Biochemical Activity vs. Sequence Constraints

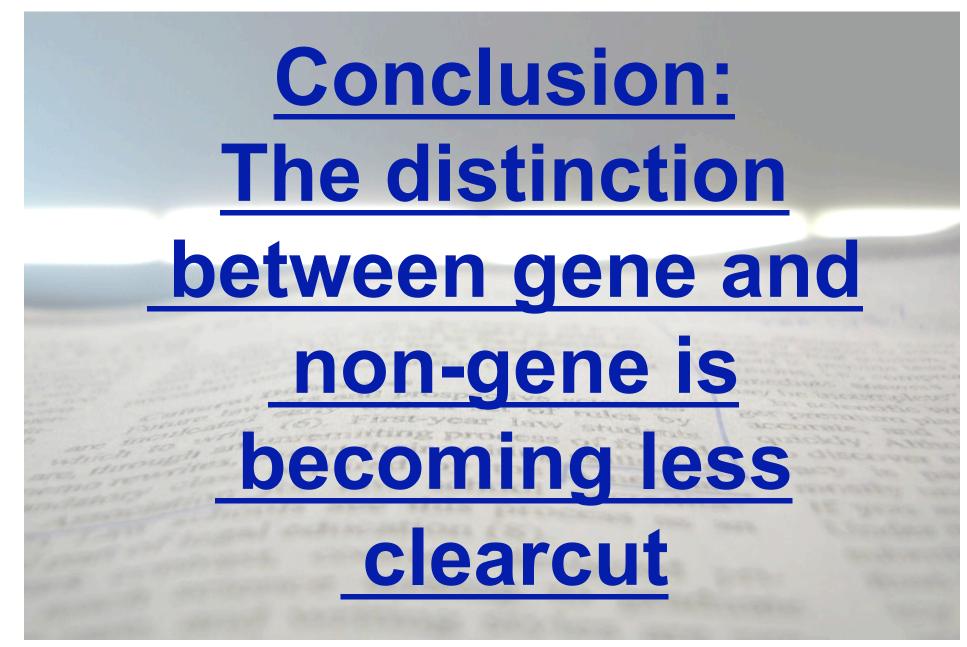


- Not all constrained sequence annotated in some fashion
- Exactly how things are defined in terms of overlap?

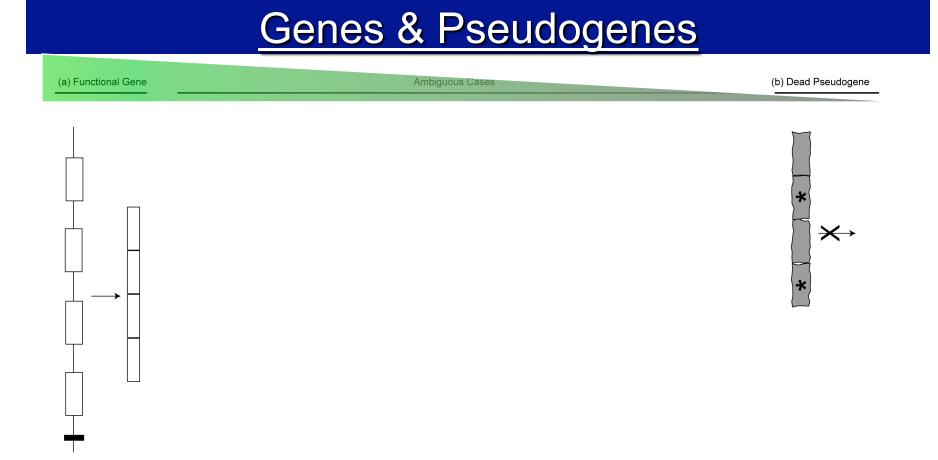
 "At the outset of the ENCODE Project, many believed that the broad collection of experimental data would nicely dovetail with the detailed evolutionary information derived from comparing multiple mammalian sequences to provide a neat 'dictionary' of conserved genomic elements, each with a growing annotation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



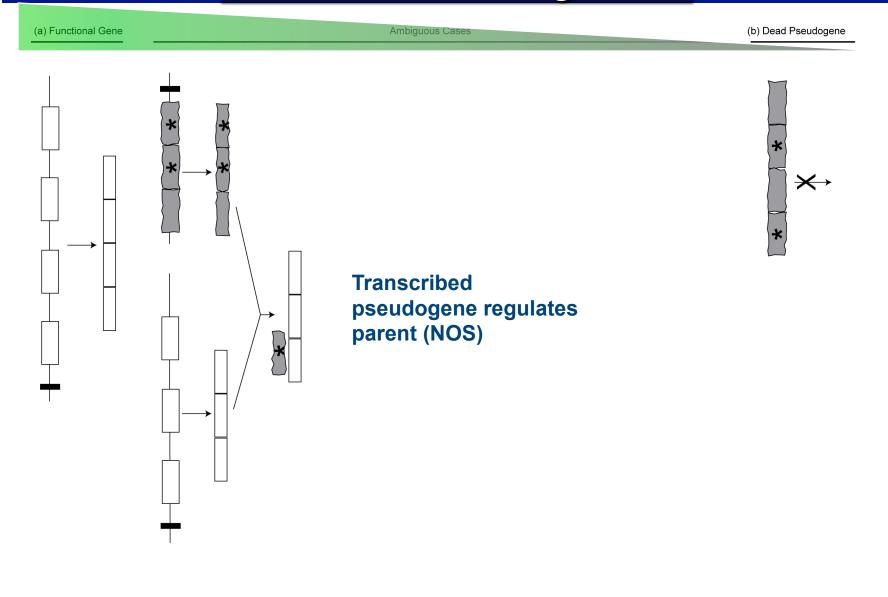
pers. photo, see streams.gerstein.info



Zheng & Gerstein, TIG (2007)

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

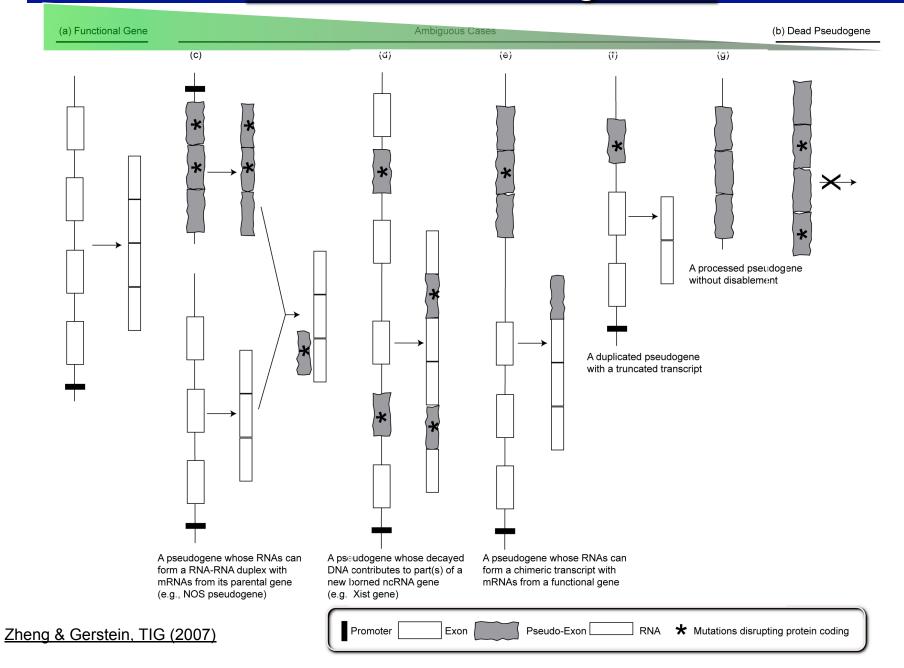
### **Genes or Pseudogenes?**

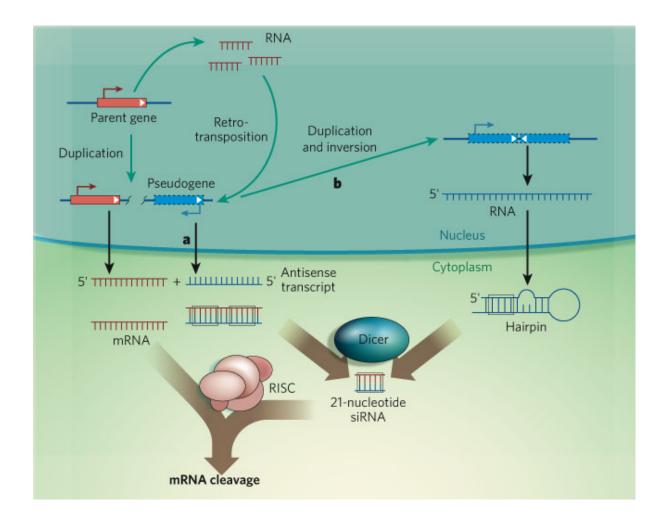


Zheng & Gerstein, TIG (2007)

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

### **Genes or Pseudogenes?**





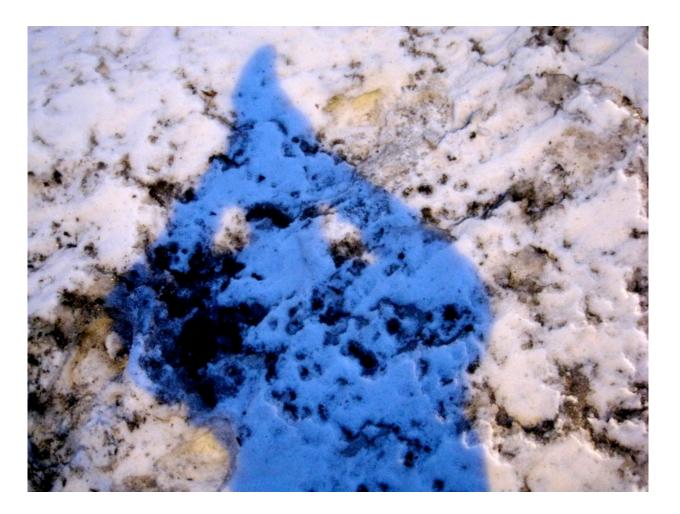
What are Active Pseudogenes Doing?

Potential for <u>Gene</u> 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).





# Overview of the Process of Intergenic Annotation

#### Basic Inputs

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

#### Results of Analyzing Similarity Comparison

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

- Results of Processing Raw Expt. Signals
  - 1. Signal Processing: removing artifacts, normalizing, window averaging
  - Segmenting signal into larger "hits" ("Active Regions" or ARs)
  - 3. Clustering together active regions into even larger features at different length scales and classifying them
  - 4. Building networks and beyond....

# **Outline**



- Signal processing to call "Blocks"
  - ◊ Calling Punctate Blocks (ChipSeq)
  - ♦ Calling Broader Blocks (CNVs)
- Clustering "Blocks" into larger regions
   Ø Binding Sites
- Annotating Copied Regions in the Genome
  - $\Diamond$  SD and CNVs
  - $\Diamond$  Pseudogenes
- Integration of Pseudogenes with Other Annotations
- Future of Annotation
  - $\Diamond$  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



## Signal Processing #2: Identifying Structural Variants in Human Population

- BreakPtr
  - Ø 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

### First-Pass Annotation Clustering and Characterizing Novel Transcribed Regions and Groups of Binding Sites

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



### Analysis of Duplication in the Genome: SVs and SDs

- Large-scale analysis of existing CNVs & SDs in human genome
- SDs assoc. with Alu, pseudogenes and older SDs
- CNVs assoc. other repeats (microsat.) and not as much with SDs
- Suggestion: Alu burst 40 MYA triggered much NAHR rearrangement, then dupl. feed on itself in hotspots but now dying down and NAHR assoc. with other repeats and CNVs also from NHEJ

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

- Pseudogene Assignment Technology
  - $\Diamond$  Pipeline + DB
  - $\Diamond$  Ontology
  - Pseudofam analysis of
     Pseudogene Families, highlight
     outliers
- Annotation of Human Genome
  - Original Operation of the Approach
    Output: Description of the Approach
  - Ribosomal Pseudogenes
  - Pilot Phase: Consensus annotation from automatic pipelines & manual curation gives 201

- Integration with Conservation and Seq. Constraint
  - ◊ ~2/3 processed are primate specific
  - Evidence for selection operating on a few but most neutral
- Pseudogene Activity
  - >20% appear to be transcribed (38/201)
  - No obvious selection on transcribed ones

# **Consortia Acknowledgements**

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

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

╋

### **ENCODE**, modENCODE, 1000 Genomes)

**D** Zheng Zhengdong Zhang **J Rozowsky J Korbel** L Wang YK Lam **P** Kim **J** Du

E Khurana G Fang R Sasidharan J Karro G Euskirchen J Chang R Bjornson N Carriero M Snyder S Weissman X Mu T Gibson

# S Balasubramanian

R Robilotto Y Liu D Greenbaum A Urban T Royce

P Cayting R Auerbach E Khurana A Abyzov J Wu Zhaolei Zhang

Tiling.gersteinlab.org Pseudogene.org SV.gersteinlab.org

# <u>Yale</u> Acknowledgements



# **More Information on this Talk**

TITLE: Human Genome Annotation

#### **SUBJECT:** GenomeTechAnnote

#### DESCRIPTION:

```
CSHL, 2009.04.29, 12:00-13:00; [I:CSHL] (Long GenomeTechAnnote talk,
incl. the following topics:
"junk DNA", anonymity*, chip-seq-simu*, peakseq*, breakptr*, msb*,
pemer*, reseqsim*,tredist*, sdcnvcorr*, cosbcnv*, pseudofam*,
pseudopipe*, encodepgenes*, rp-pgenes*, sirnapgene*, encode-pilot*,
pgene-classify*. Too long, fits into 70'. PPT works on mac & PC and
has many photos w. EXIF tags .)
```

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

```
the topic pubnet* can be looked up at
http://papers.gersteinlab.org/papers/pubnet )
```

**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).

<u>PHOTOS & IMAGES</u>. 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 .