Human Genome Annotation

Mark B Gerstein
Yale

Slides at **Lectures.GersteinLab.org**

(See Last Slide for References & More Info.)

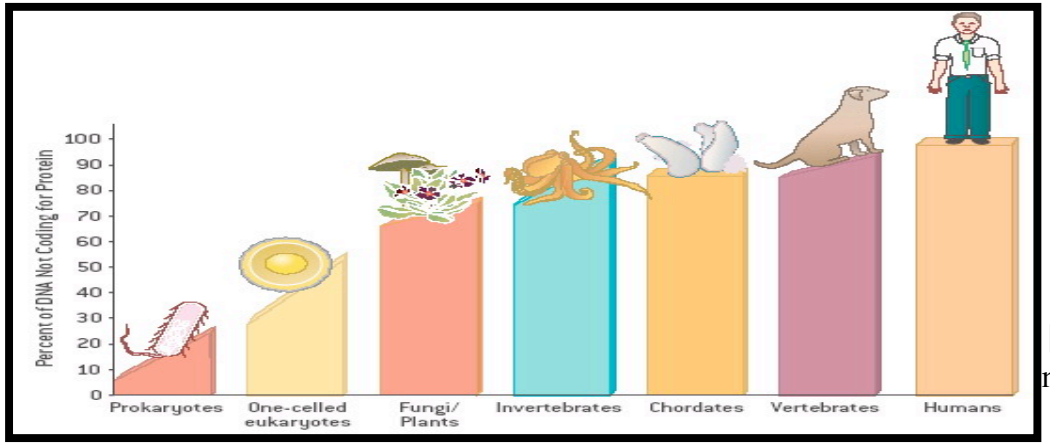
2001: Most of the genome is not coding (only ~1.2% exon).

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

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

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

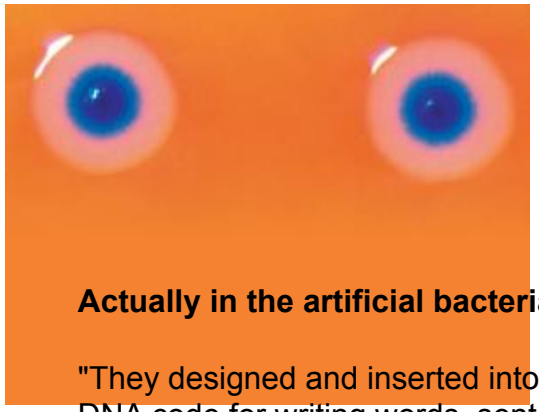
2001



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

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

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



Views on the Function of Junk DNA: Secret Messages

Actually in the artificial bacterial cell, Mycoplasma mycoides JCVI-syn1.0 [Gibson et al., '10]:

"They designed and inserted into the genome what they called watermarks.... Encoded in the watermarks is a new DNA code for writing words, sentences and numbers. In addition to the new code there is a web address... and three quotations: "TO LIVE, TO ERR, TO FALL, TO TRIUMPH, TO RECREATE LIFE OUT OF LIFE." - JAMES JOYCE; ... "

ESSAY

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

By DENNIS OVERBYE

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

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

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

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

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

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

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

Jimmy Turrell

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

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

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

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

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

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

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

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

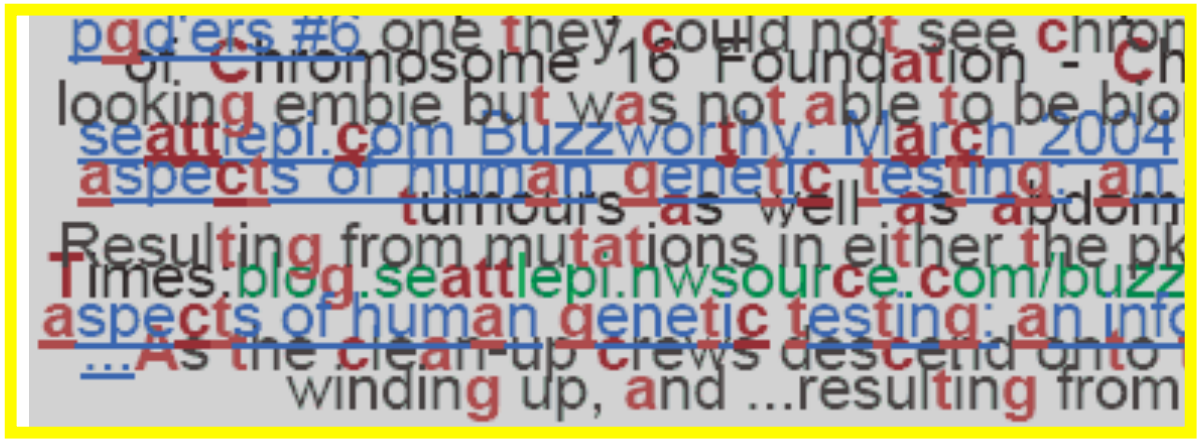
sections of junk DNA seem to be markedly resistant to

Startibartfast.

[NY Times, 26-Jun-07]

[M Gerstein ('10) Am. Sci.]

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

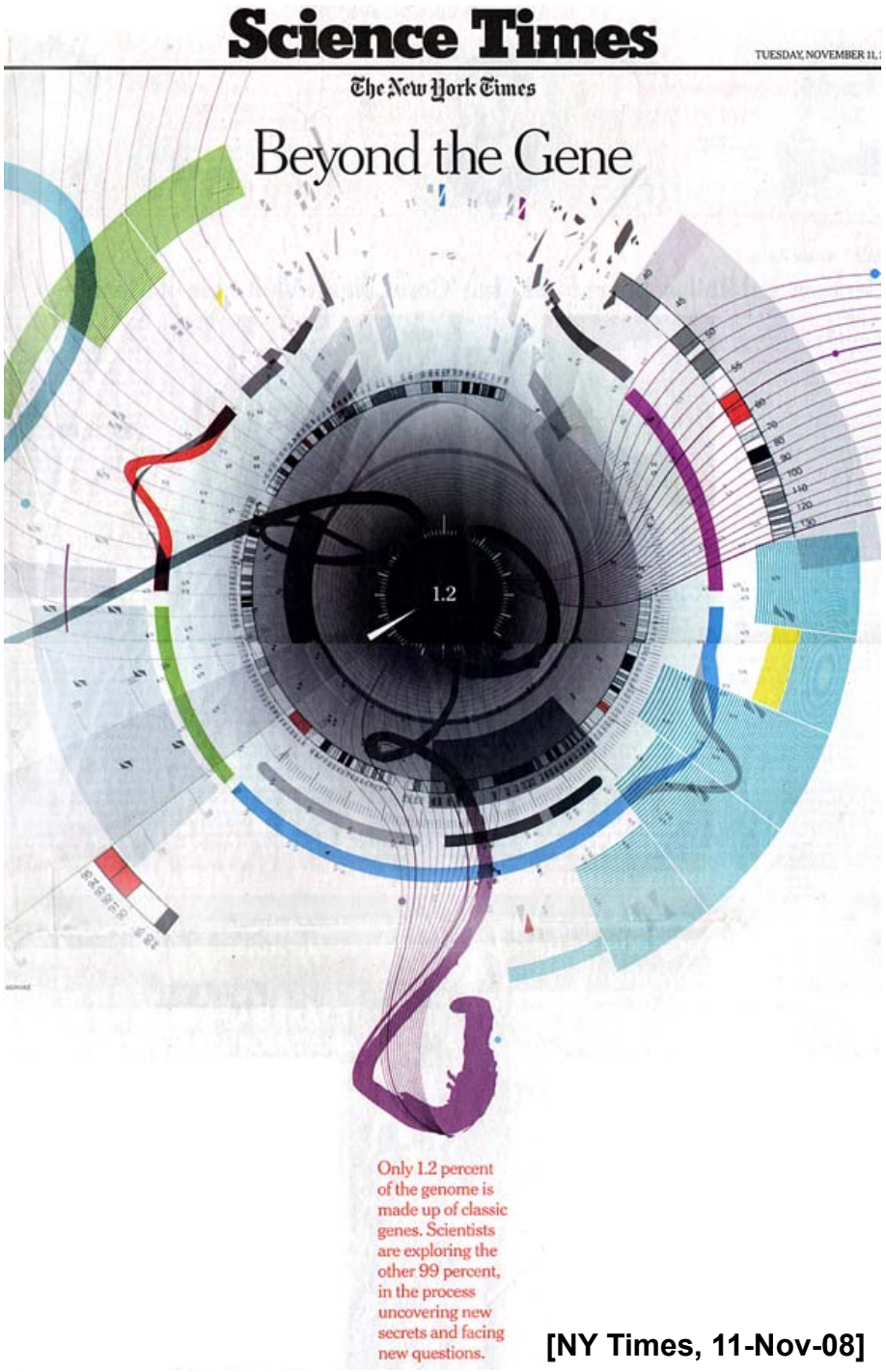


Junk DNA as Art

Significance of the "dark matter of the genome"

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

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



[NY Times, 11-Nov-08]

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

How might we annotate a human text?

Color is Function

Lines are Similarity

The Semicolon Wars

Brian Hayes

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

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

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

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

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity.

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

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

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

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

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

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

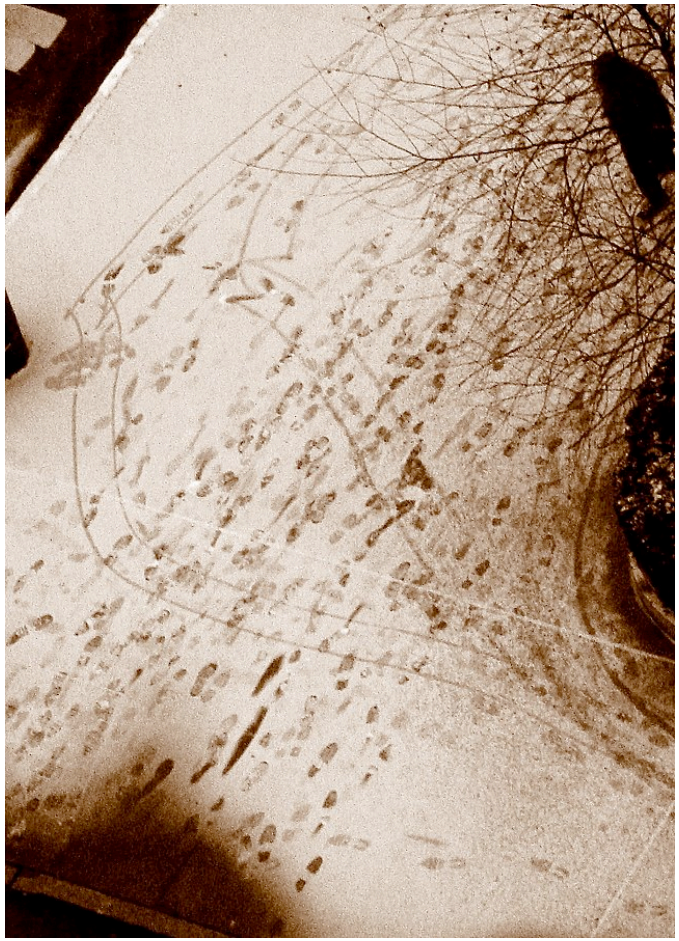
[B Hayes, Am. Sci. (Jul.- Aug. '06)]

Overview of the Process of Annotation of non-coding Regions

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

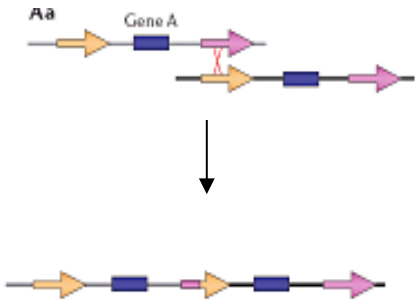
  Finding repeated or deleted blocks in the genome
  1. As a function of similarity (i.e. age, perhaps using explicit models)
  2. vs. other organisms, vs. human reference, or within the human population
     (synteny, SDs, and CNVs)
  3. Big and small blocks
     (duplicated regions and retrotransposed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

Outline



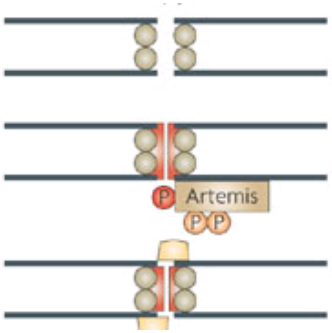
- Variable Blocks in the Genome (SVs,SDs)
  - Calling SVs with various approaches (MSB, PEMer, ReSeqSim, BreakSeq)
  - Analyzing mechanism of formation for precisely resolved breakpoints & on a large-scale over the genome
- Pseudogenes
  - Pattern-match assignment tools
  - Focus on different specific groups – glycolytic, unitary
  - Polymorphic Pseudogenes
  - Inter-relating Pseudogenes with SDs & SVs

SV Formation Mechanism



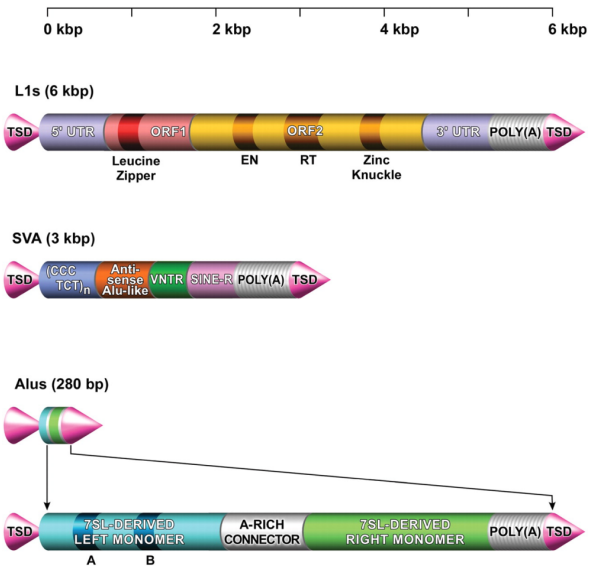
NAHR
(Non-allelic homologous recombination)

Flanking repeat
(e.g. Alu, LINE…)



NHEJ
(Non-homologous-end-joining)

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

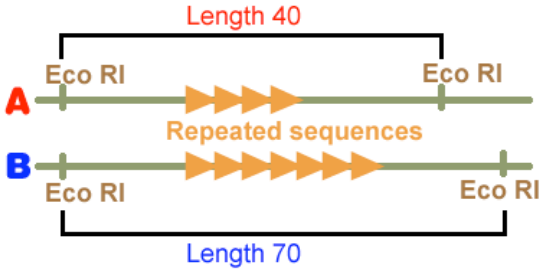


TEI
(Transposable element insertion)

L1, SVA, Alus

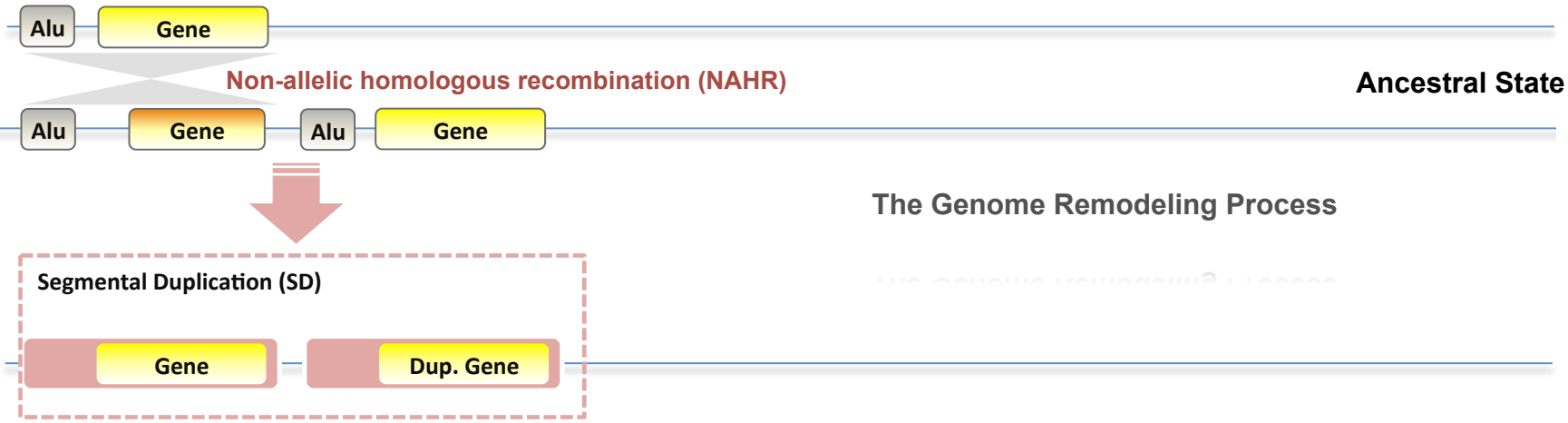
VNTR
(Variable Number Tandem Repeats)

Number of repeats varies between different people

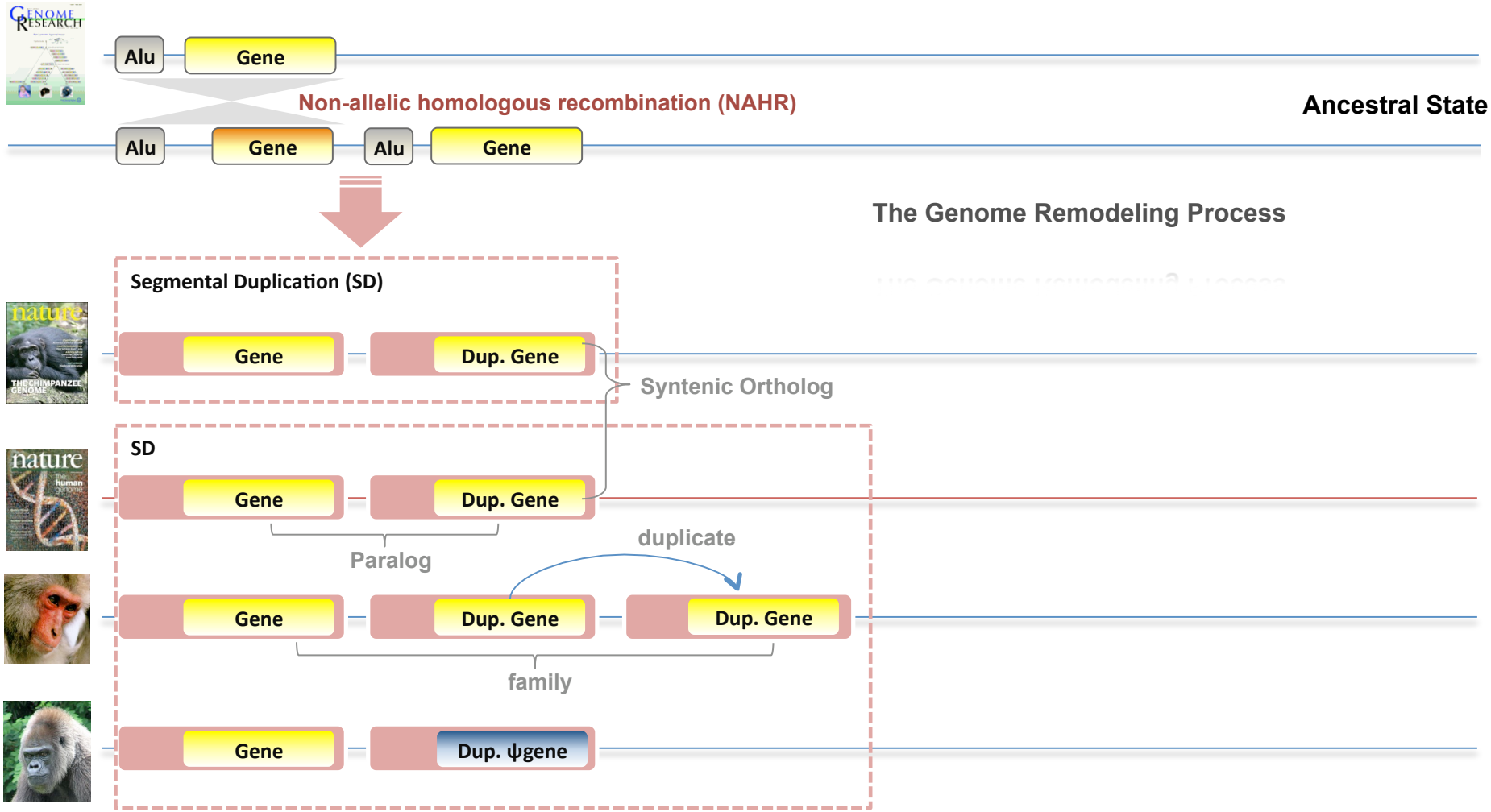


Genomic Variation

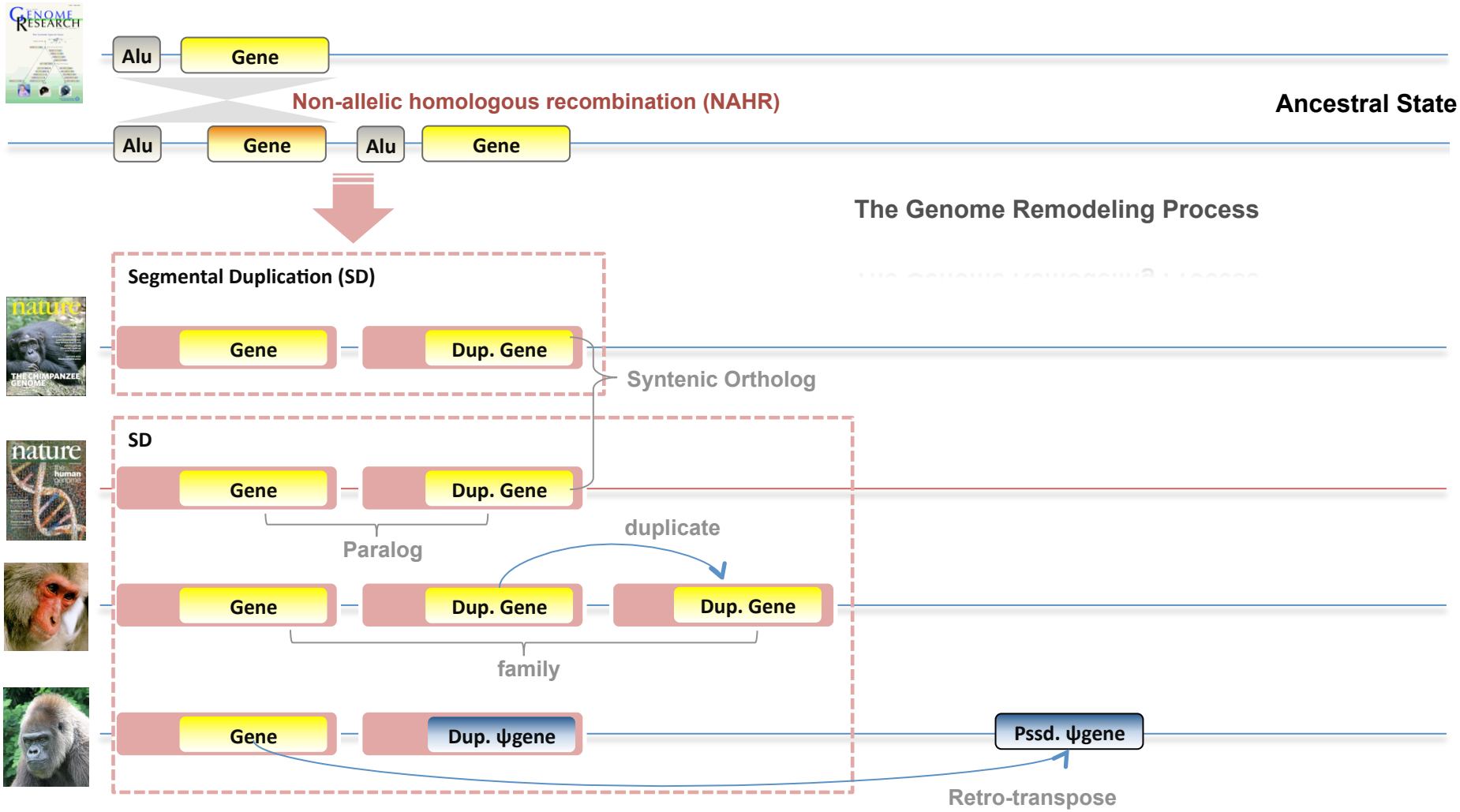
Genomic Variation



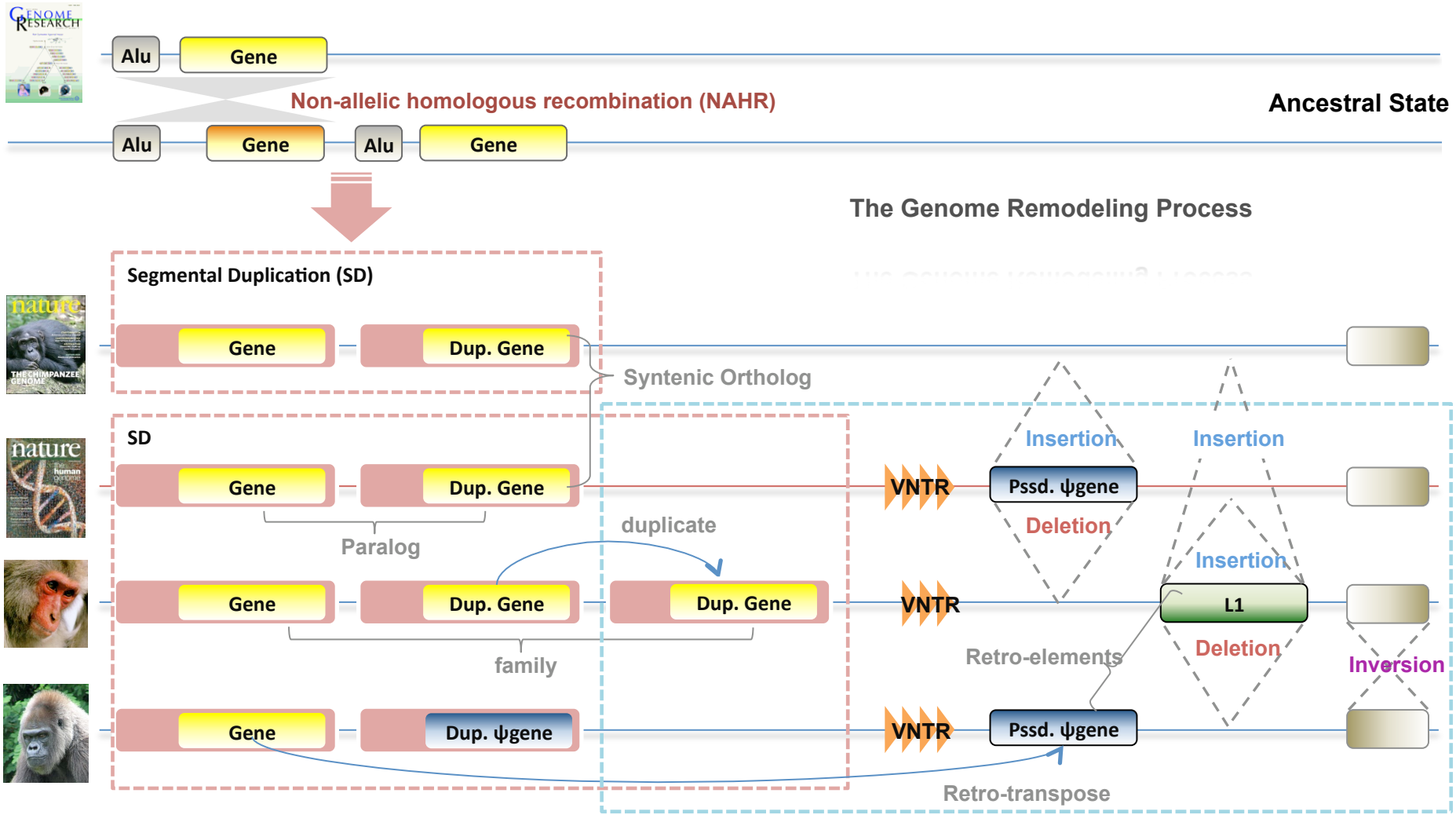
Genomic Variation



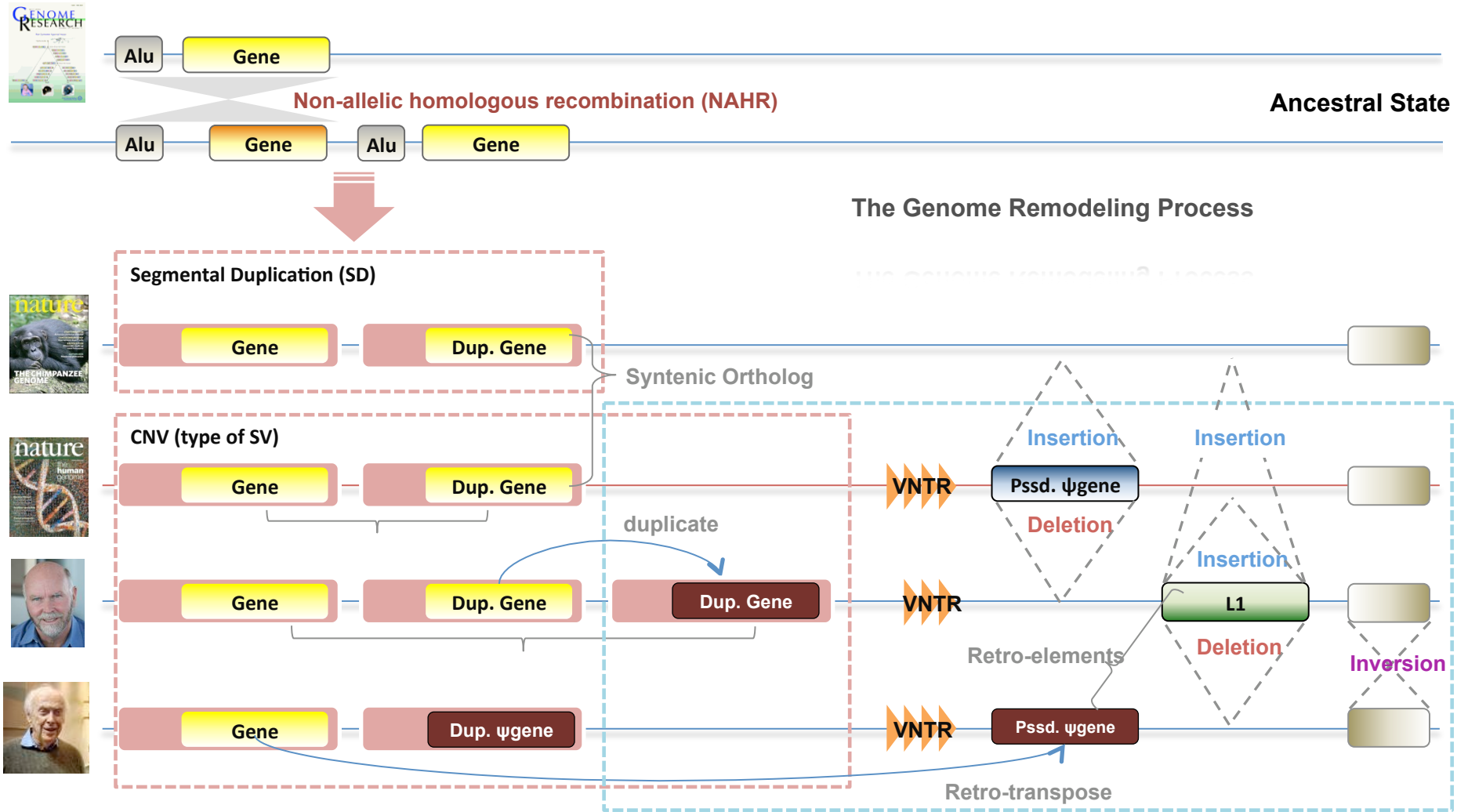
Genomic Variation



Genomic Variation



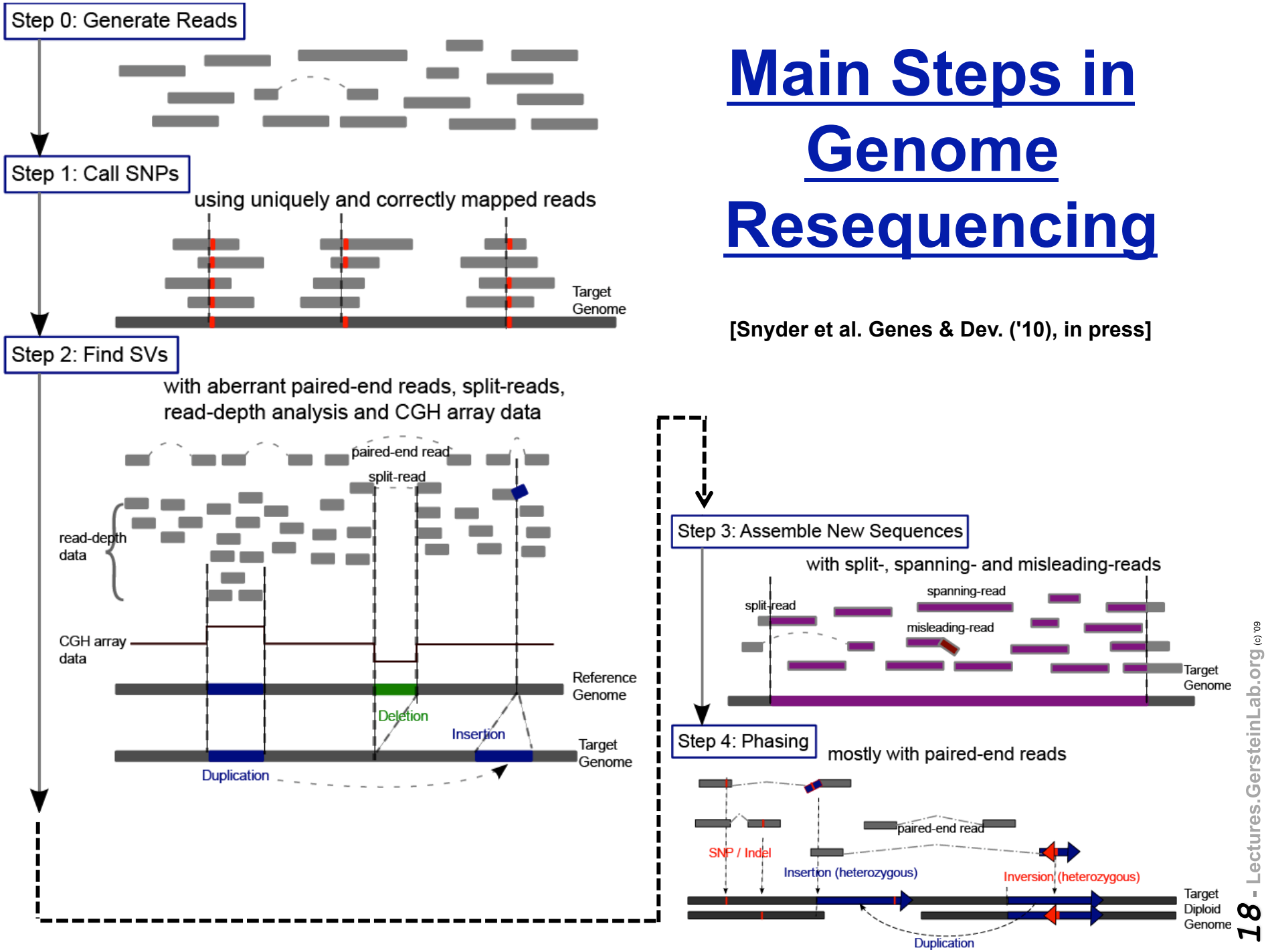
Genomic Variation



"Polymorphic" Genes & Pseudogenes

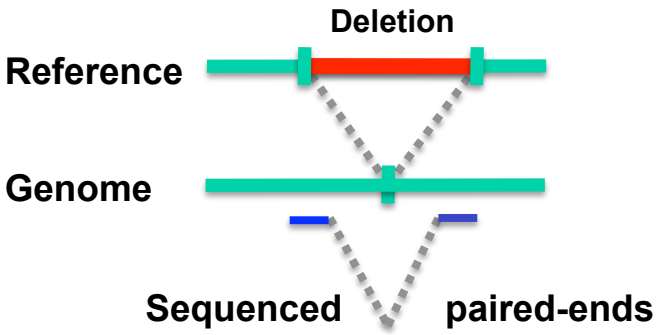
Main Steps in Genome Resequencing

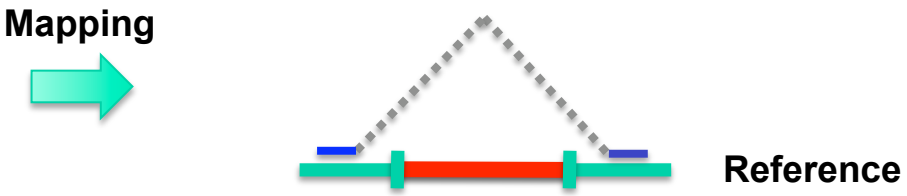
[Snyder et al. Genes & Dev. ('10), in press]



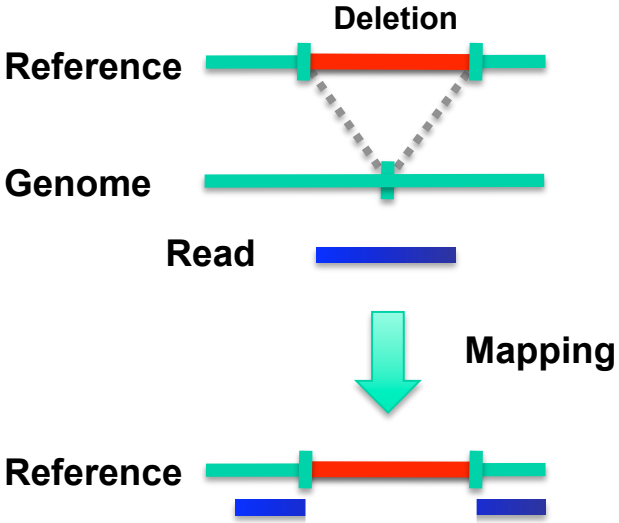
Methods to **Find SVs**

1. Paired ends

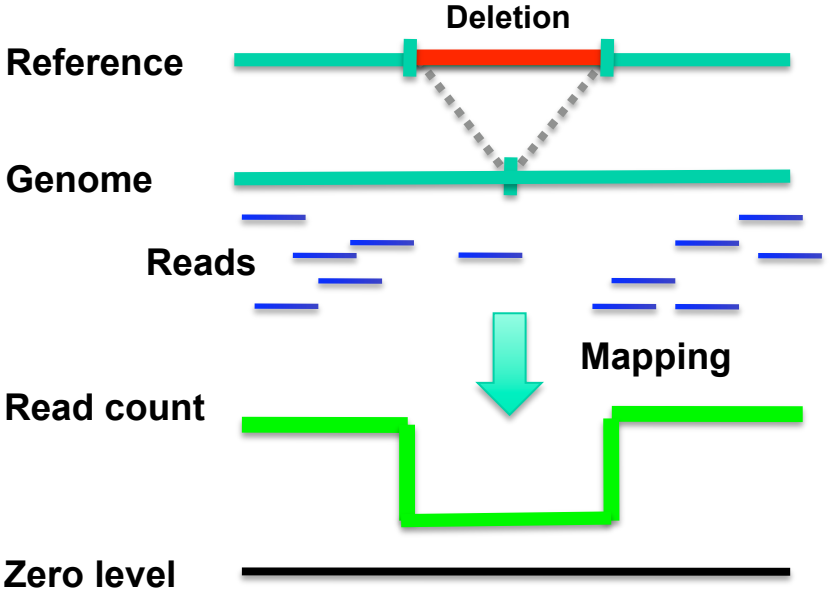




2. Split read



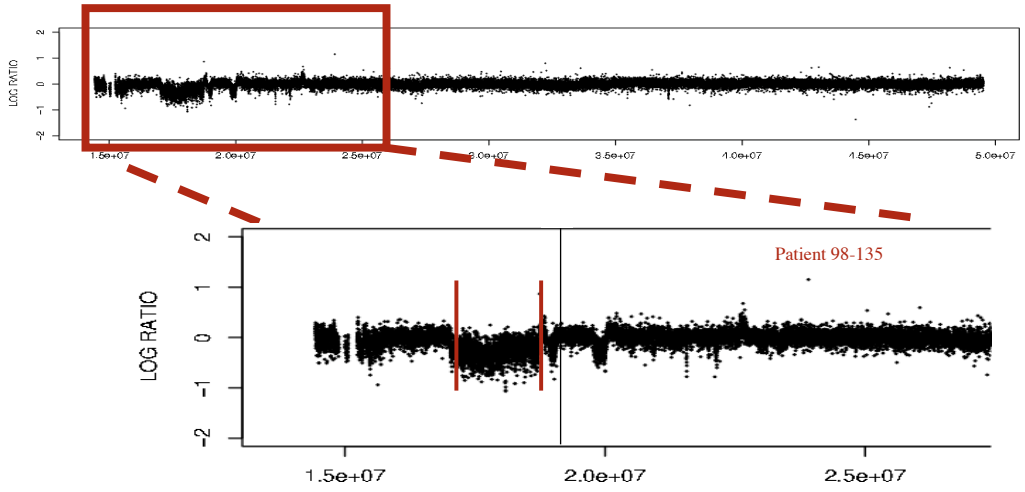
3. Read depth (or aCGH)



4. Local Reassembly

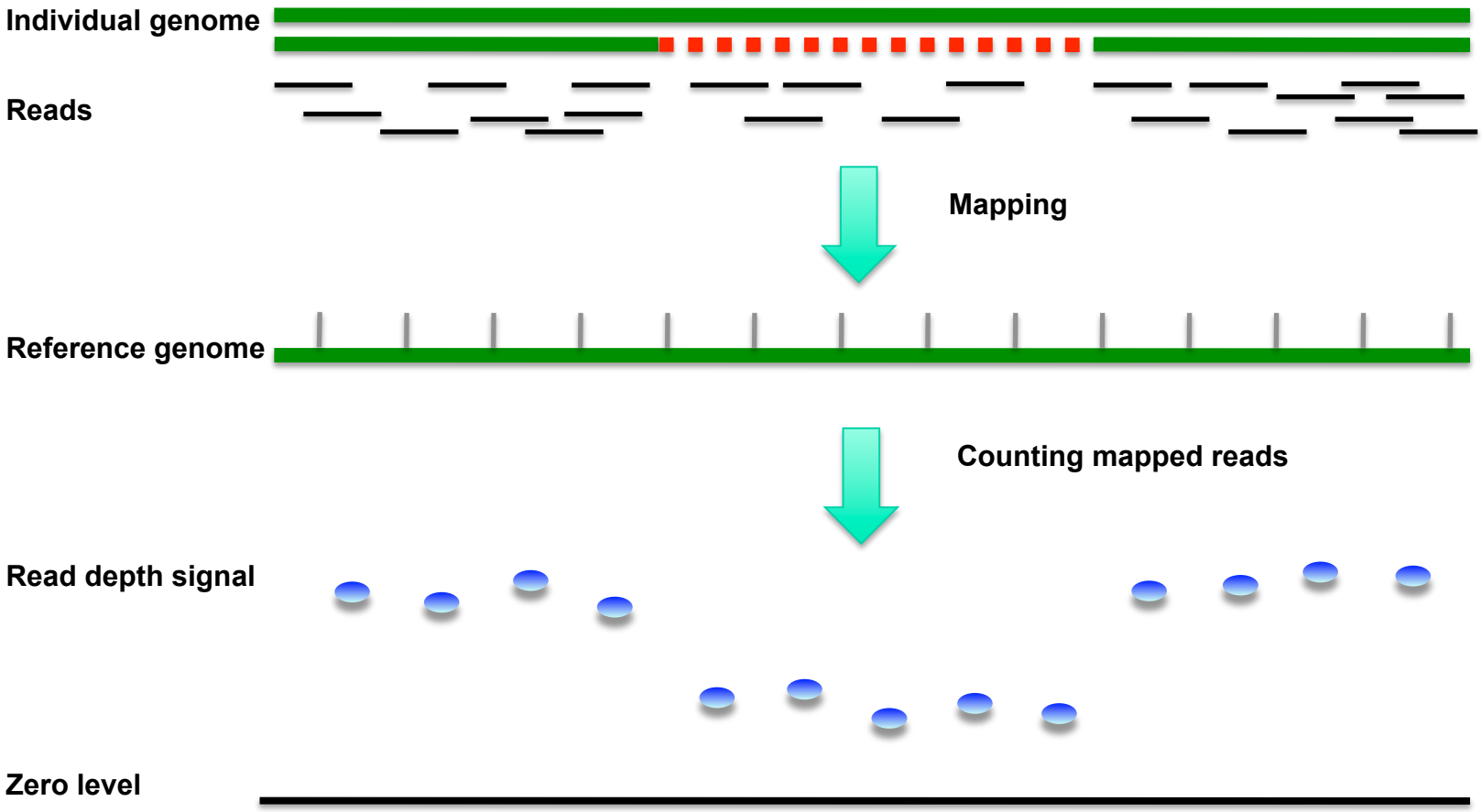
[Snyder et al. Genes & Dev. ('10), in press]

MSB: Read-Depth Segmentation



Array Signal

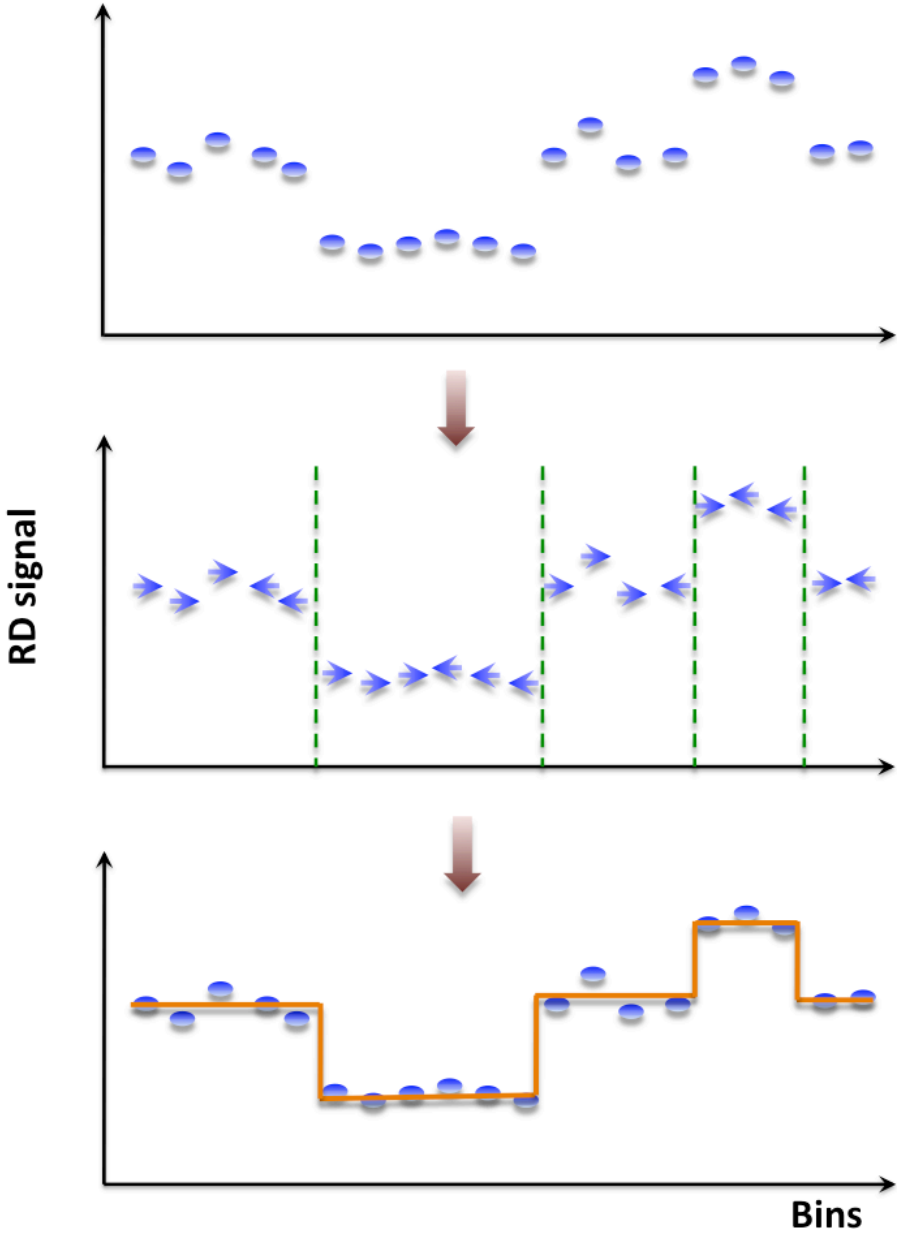
Read depth



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

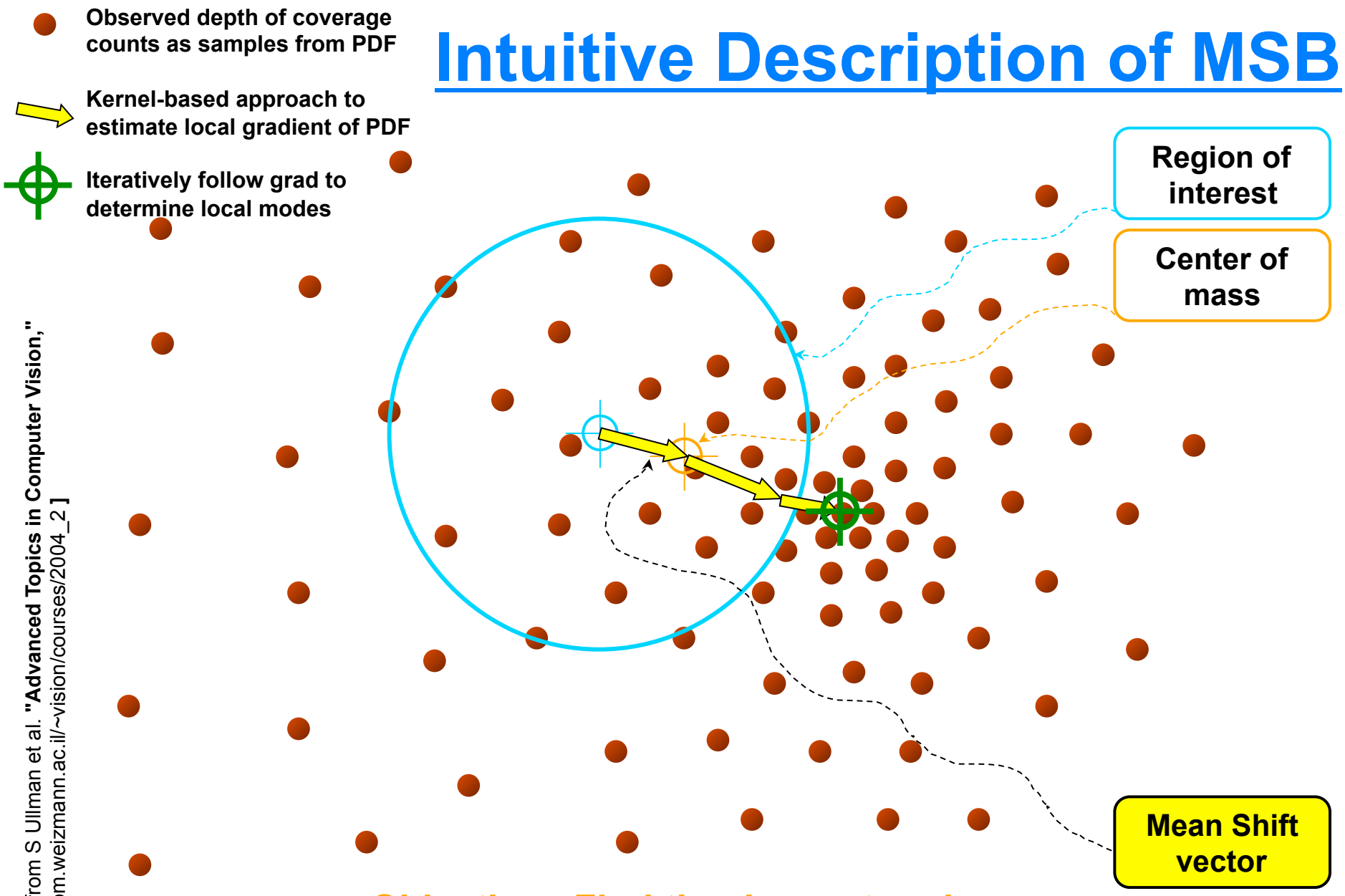
Mean-shift-based (MSB) Segmentation: no explicit model

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



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

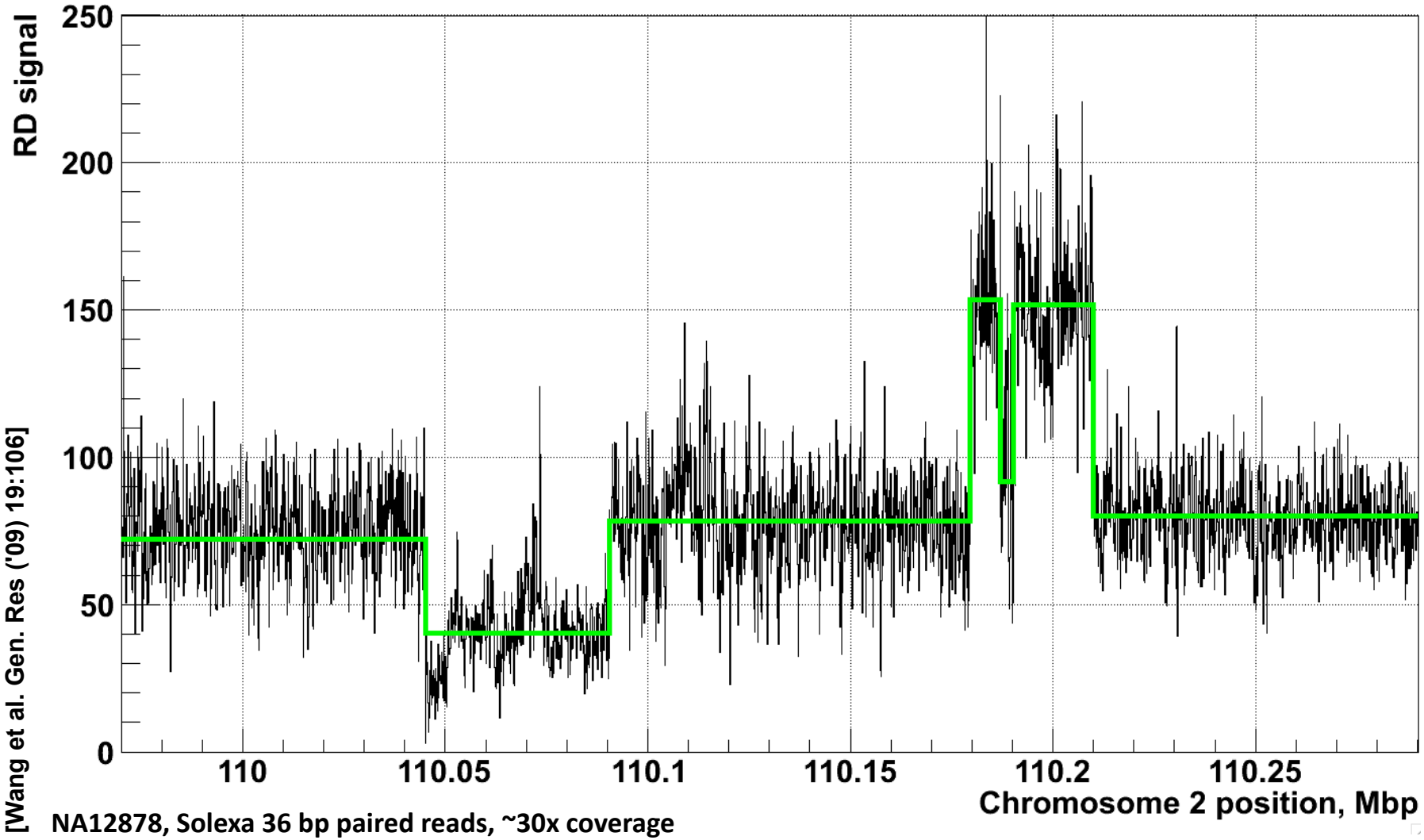
Intuitive Description of MSB



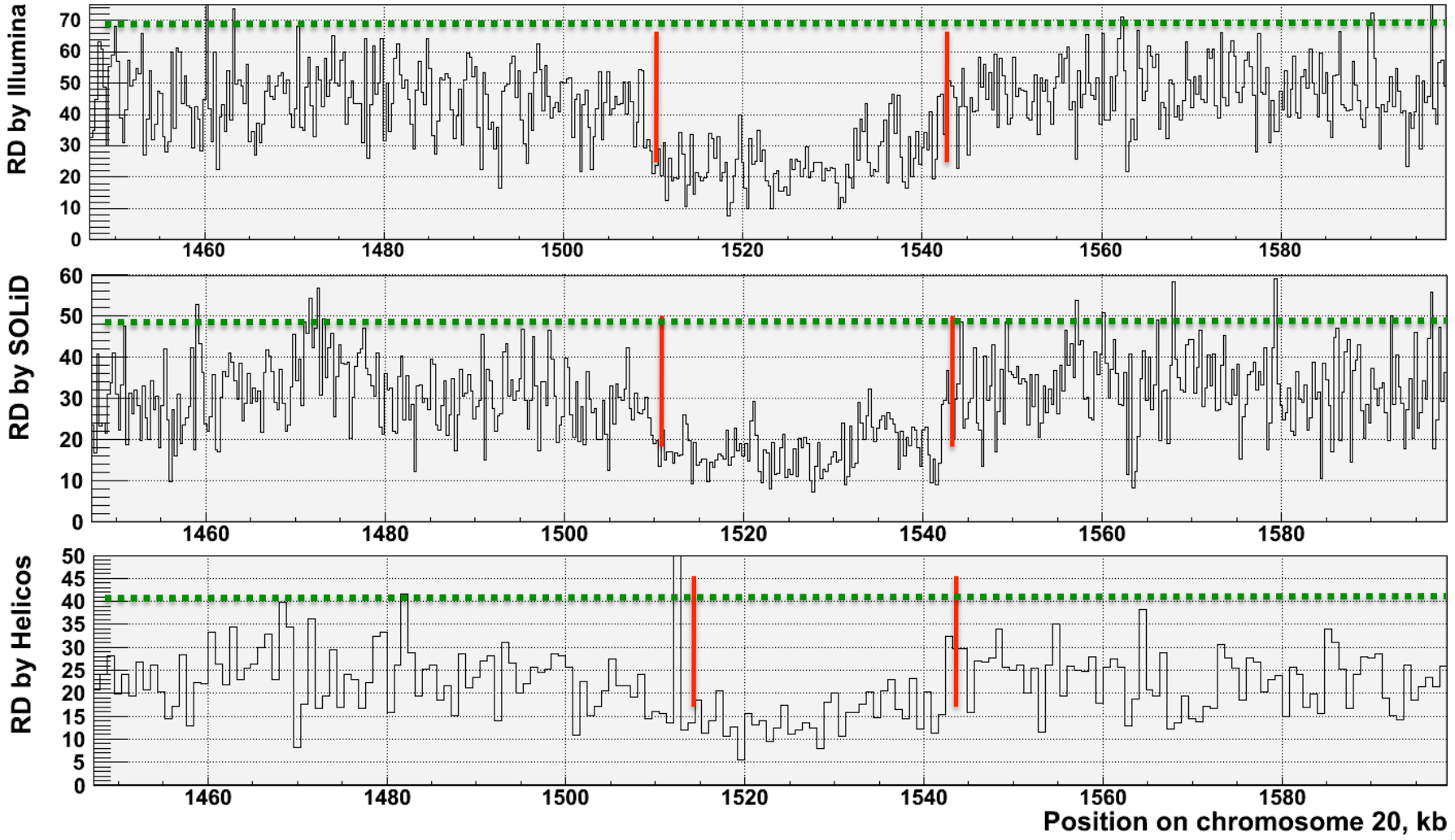
[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Objective : Find the densest region
Distribution of identical billiard balls

Example of Application of MSB to RD data



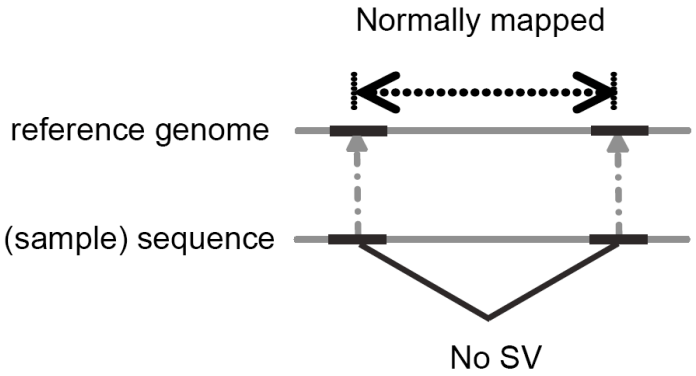
RD works well on a variety of sequencing platforms

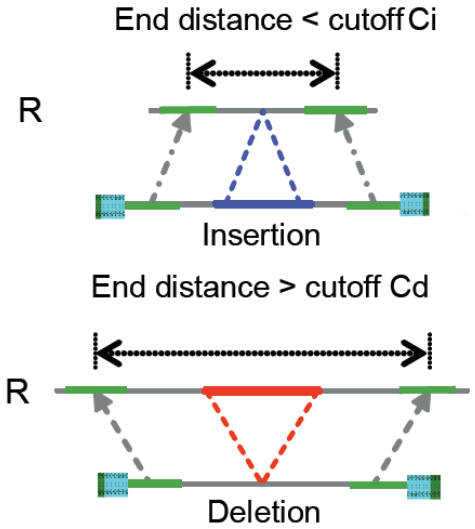


[NA18505]

Looking for Aberrantly Placed Paired Ends

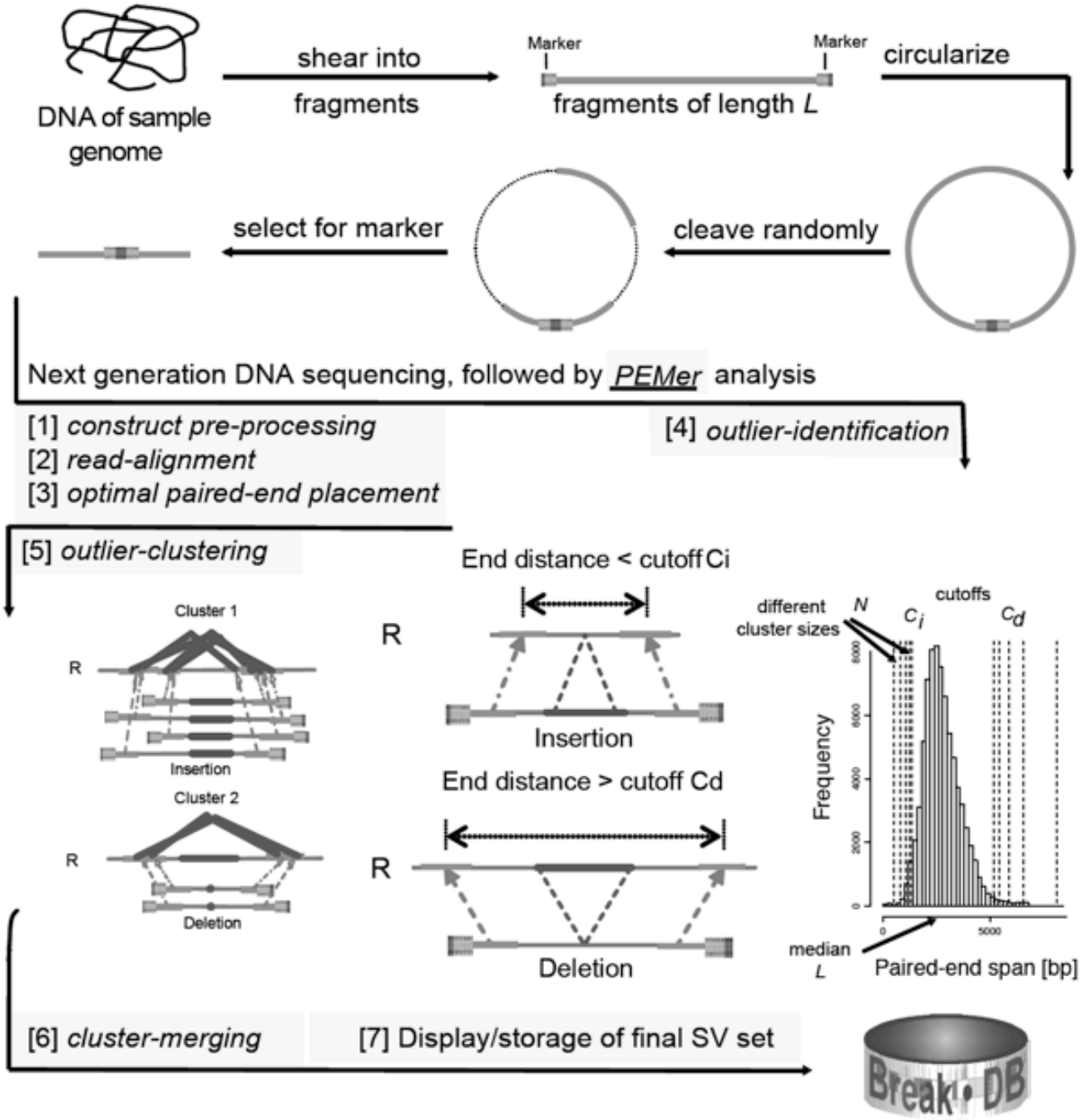
PEMer: Detecting Structural Variants from Discordant Paired Ends in NextGen Seq. Data





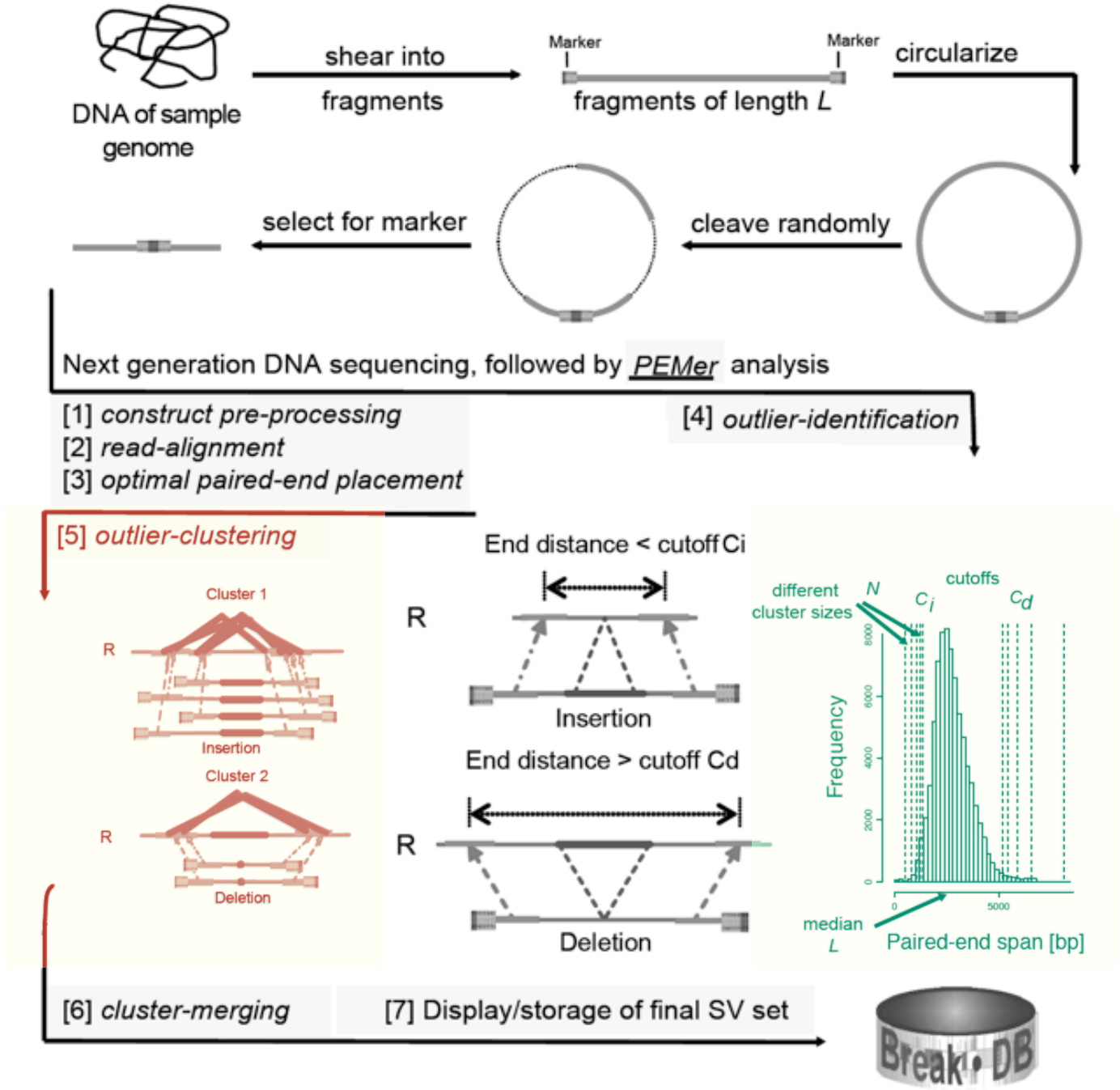
[Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]

PEMer: Detecting Structural Variants from Discordant Paired Ends in NextGen Seq. Data



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

PEMer: Detecting Structural Variants from Discordant Paired Ends in NextGen Seq. Data

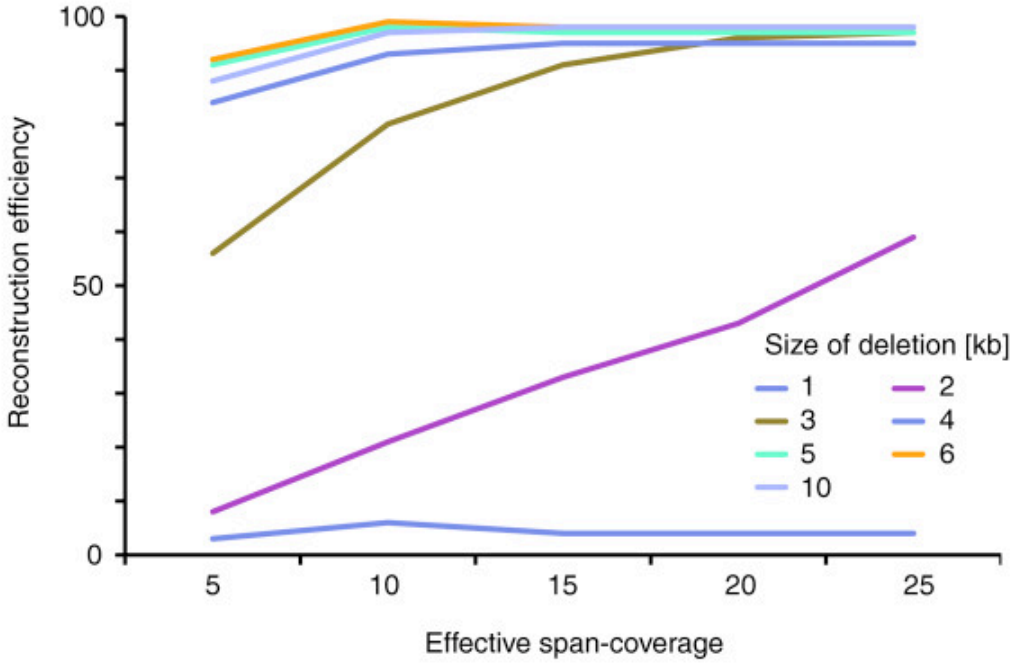


[Korbel et al., Science ('07); Korbel et al., GenomeBiol. ('09)]

Parameterize Error Models through Simulation

Reconstruction efficiency at different coverage

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

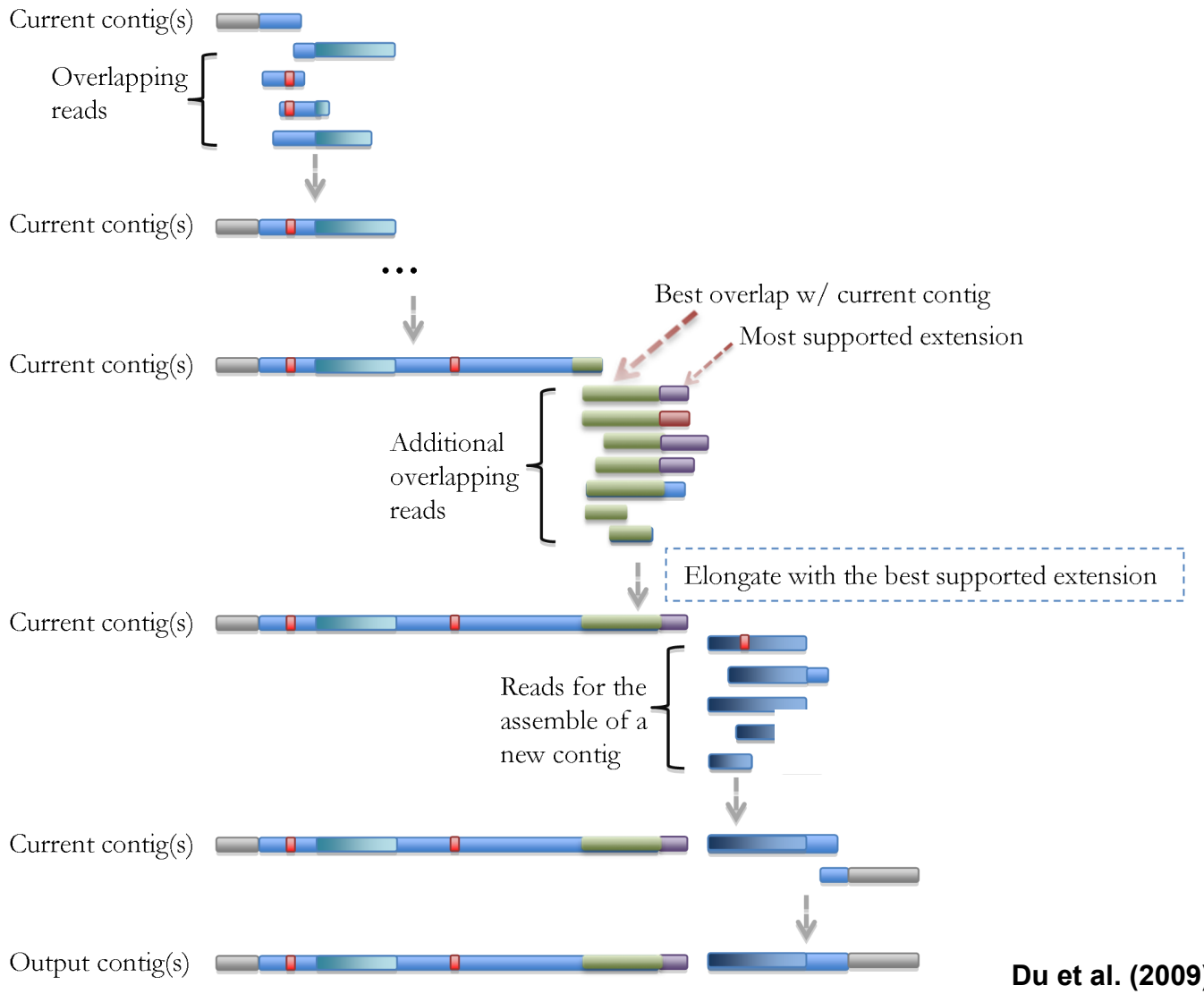


[Korbel et al., GenomeBiol. ('09)]

Local Reassembly

Simple Local Assembly: iterative contig extension

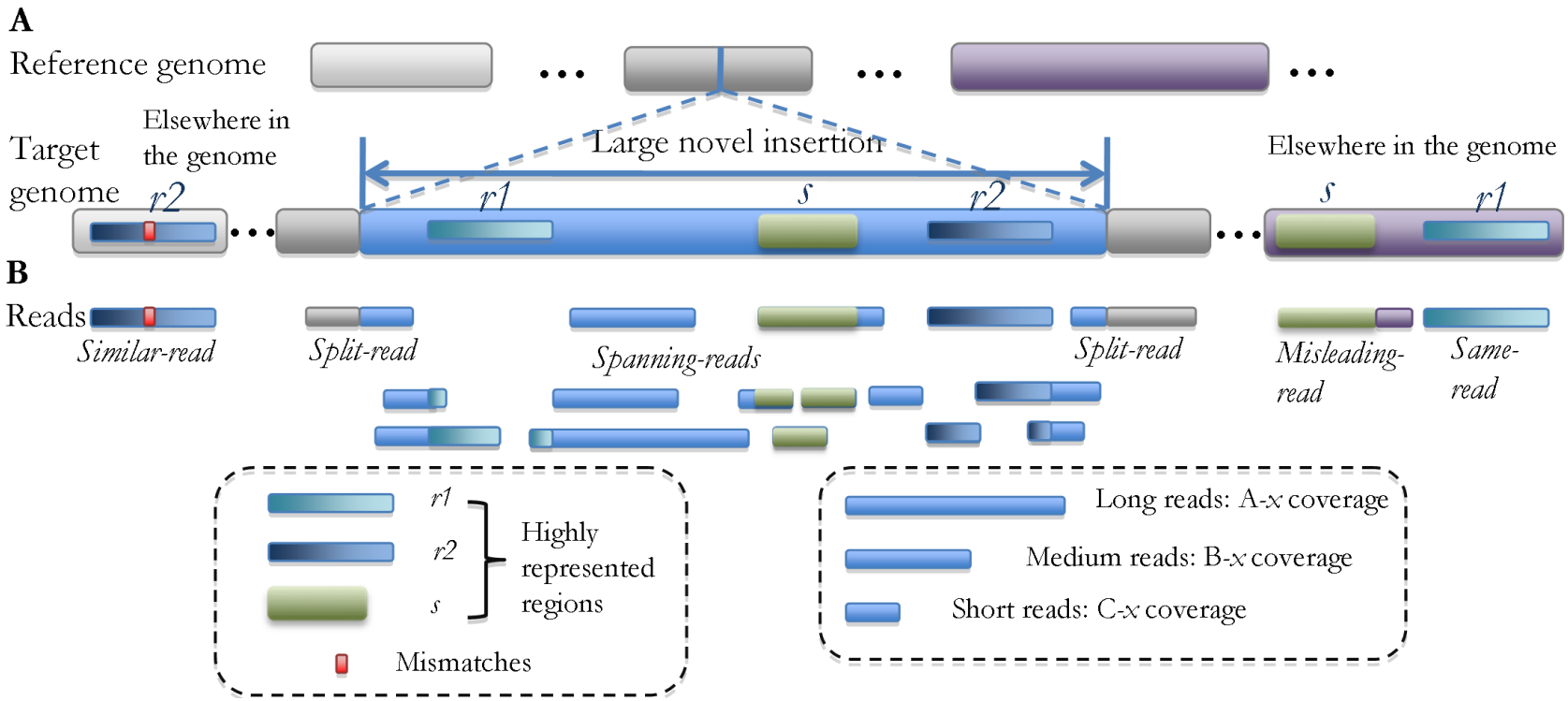
G Iterative contig elongation with the best supported extension -- a mostly greedy approach



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

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

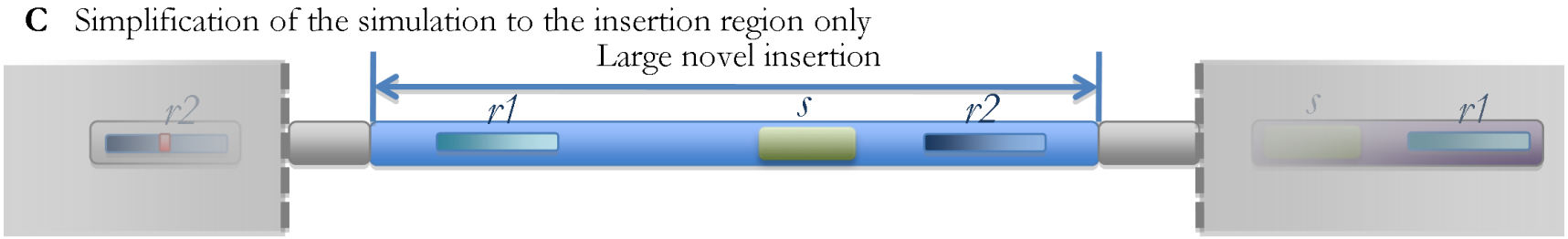
Given a fixed budget, what are the sequencing coverage A, B and C that can achieve the maximum reconstruction rate (on average/worst-case)? Maybe a few long reads can bootstrap reconstruction process.



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

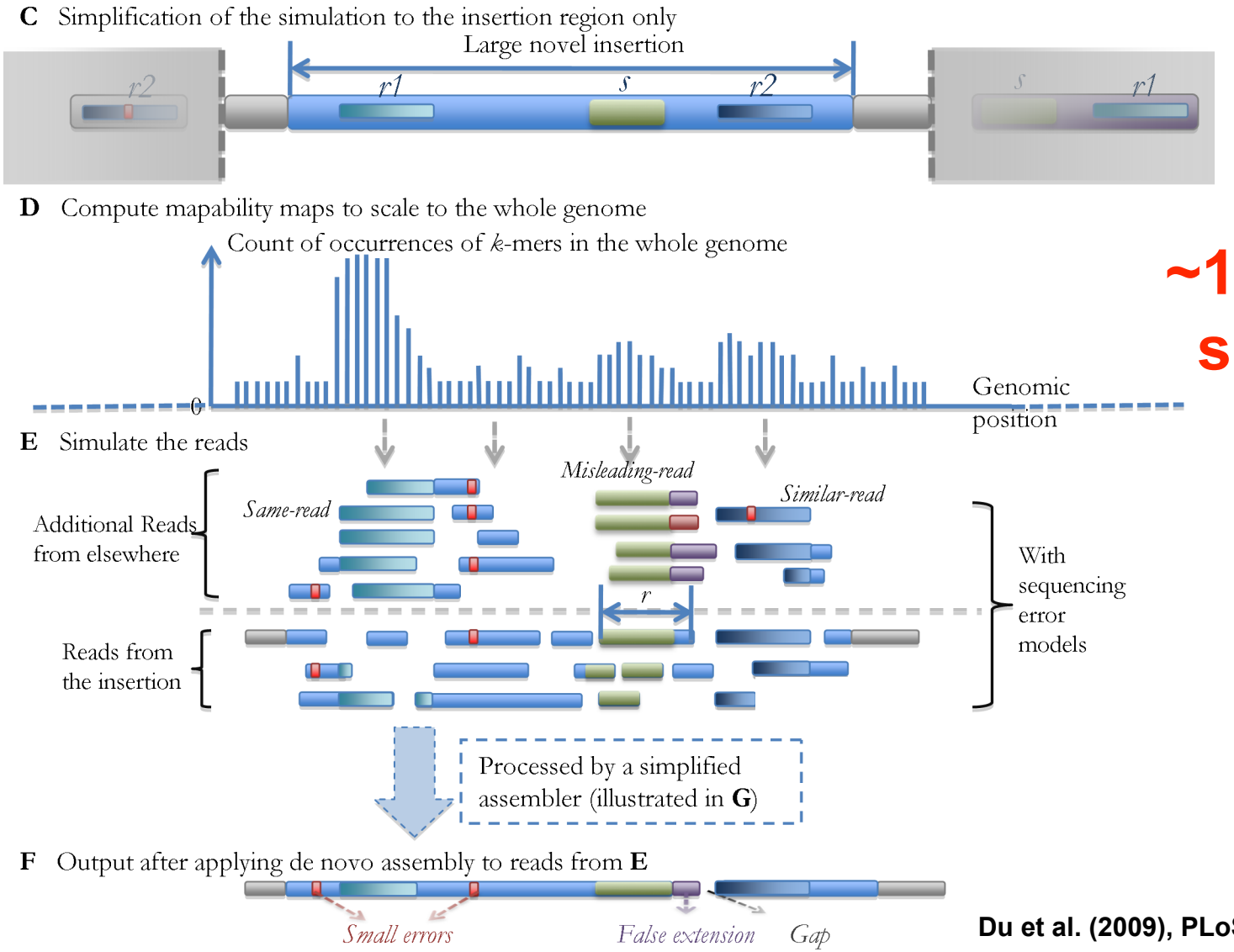
Optimal integration of sequencing technologies:
Need Efficient Simulation

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



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

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

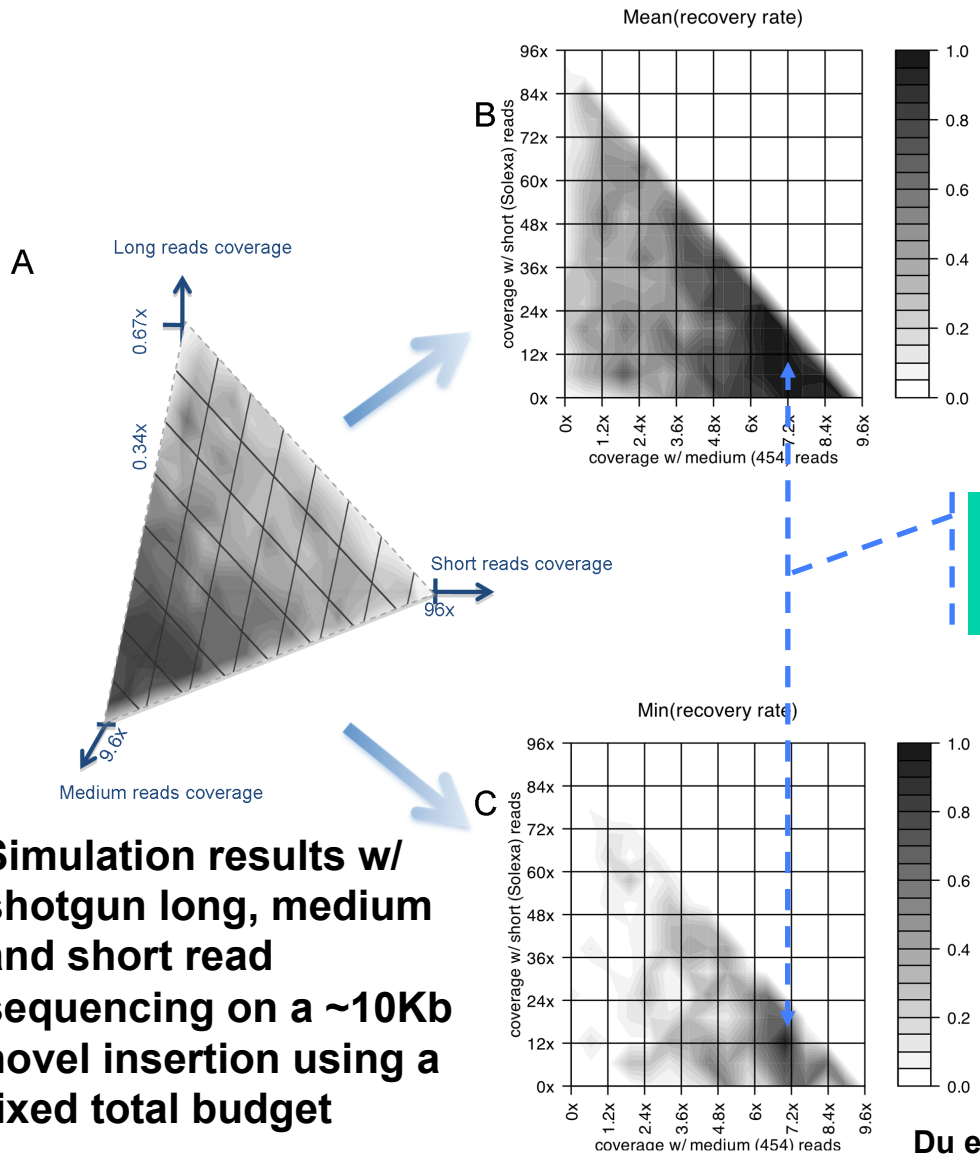


~100,000 X speedup

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

Optimal integration of sequencing technologies:

Simulation shows combination better than single technology



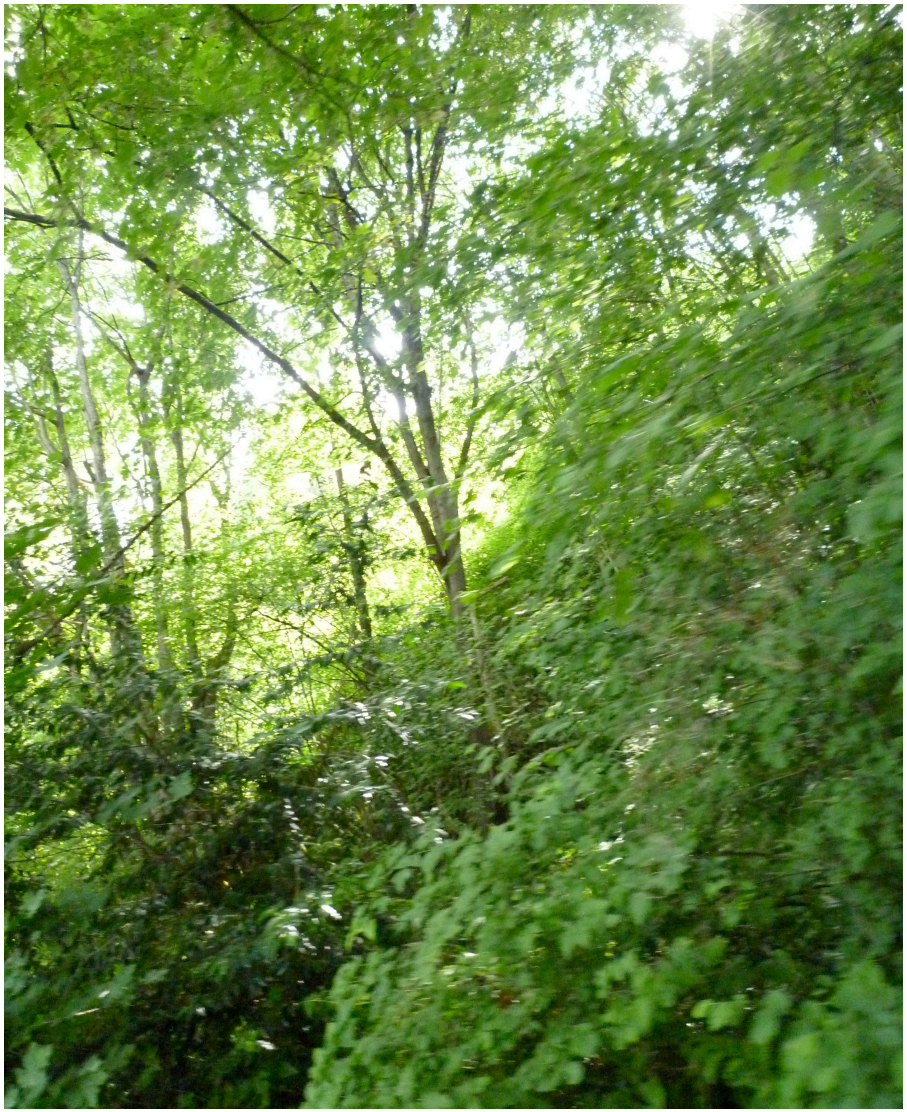
Optimal combination of different technologies

Result dependent on specific parameter setting of different sequencing technologies

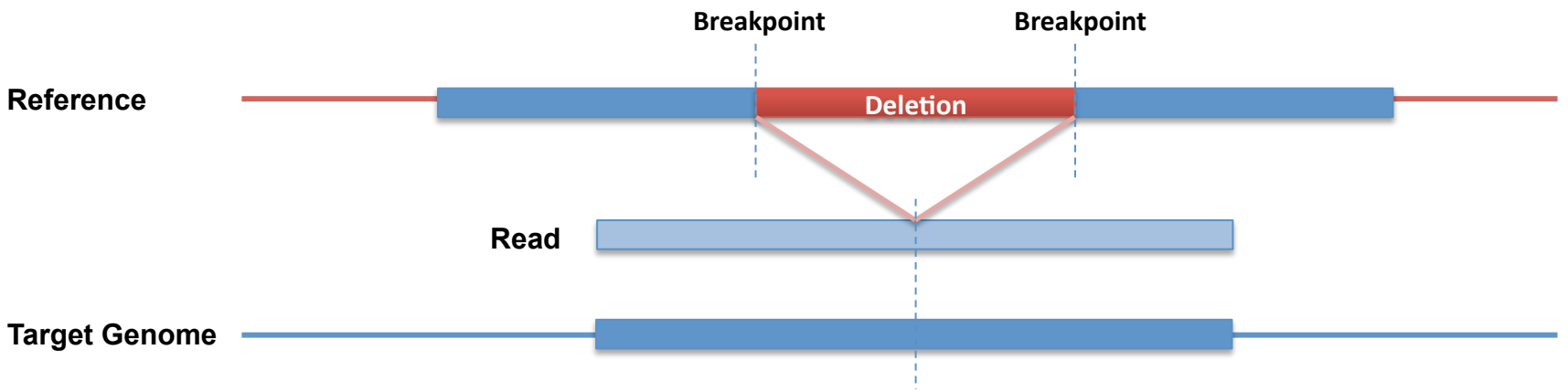
Simulation results w/ shotgun long, medium and short read sequencing on a ~10Kb novel insertion using a fixed total budget

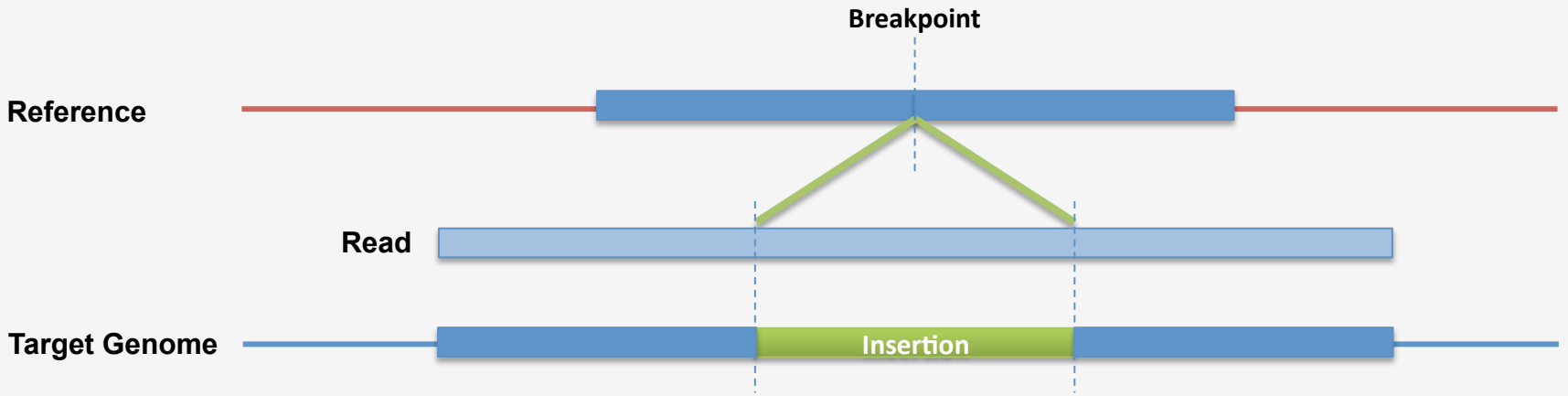
Du et al. (2009), PLOS Comp Biol, in press

Split Read



Split-read Analysis



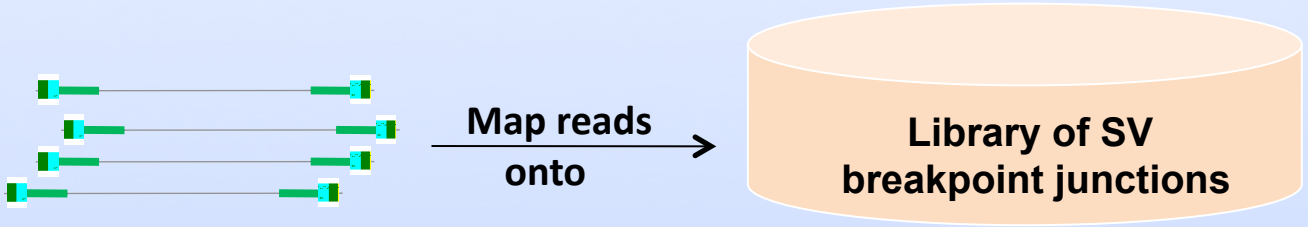


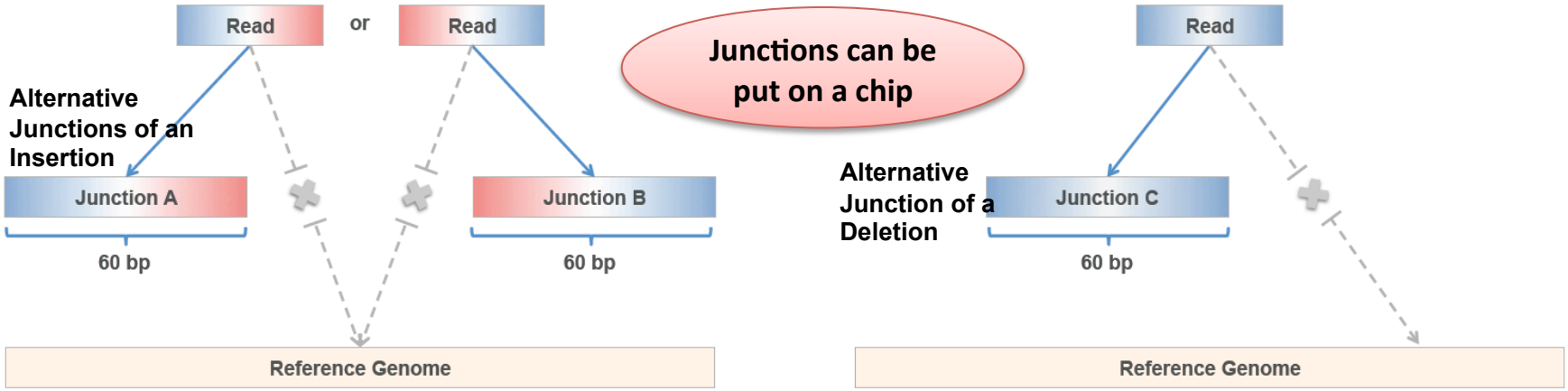
Zhang et al. Submitted

More: Breakpoint AssemblyAlt: BreakSeq

SV Detection and Genotyping

"BreakSeq" leverages the junction library to detect previously known SVs at nucleotide-level from short-read sequenced genome, which can hardly be achieved by methods such as split-read

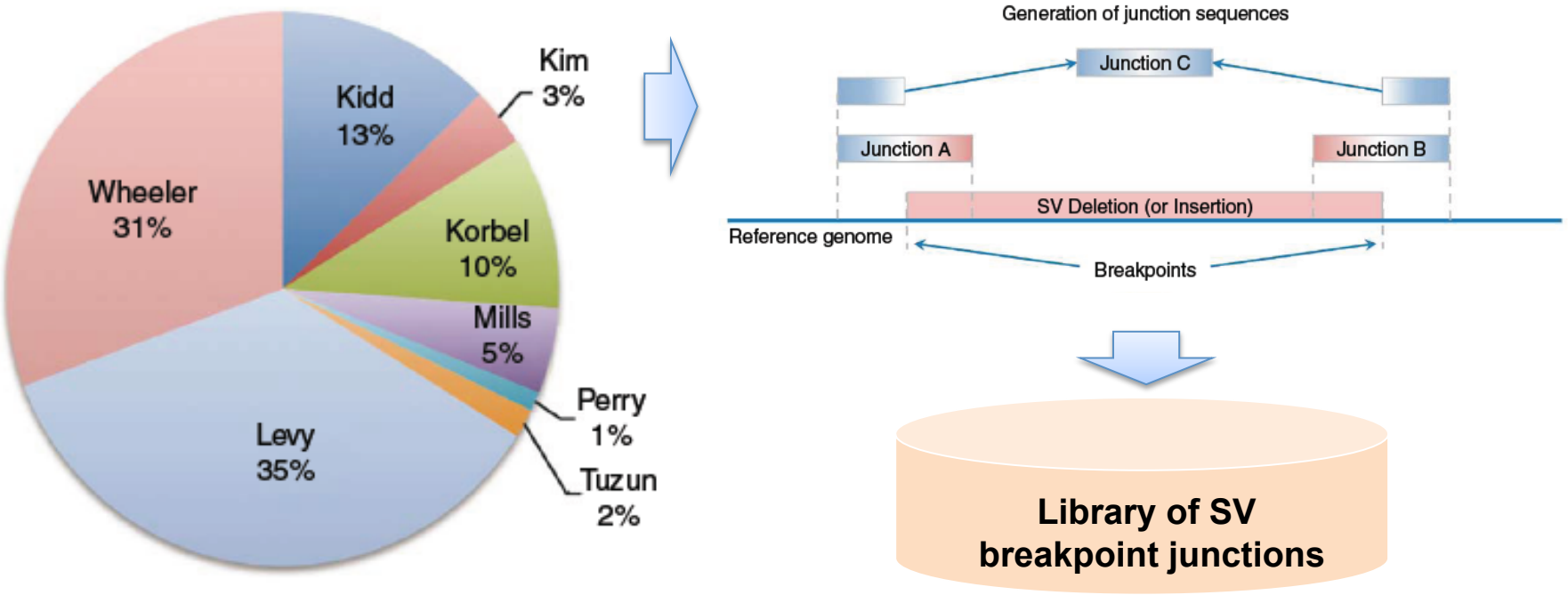




* Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

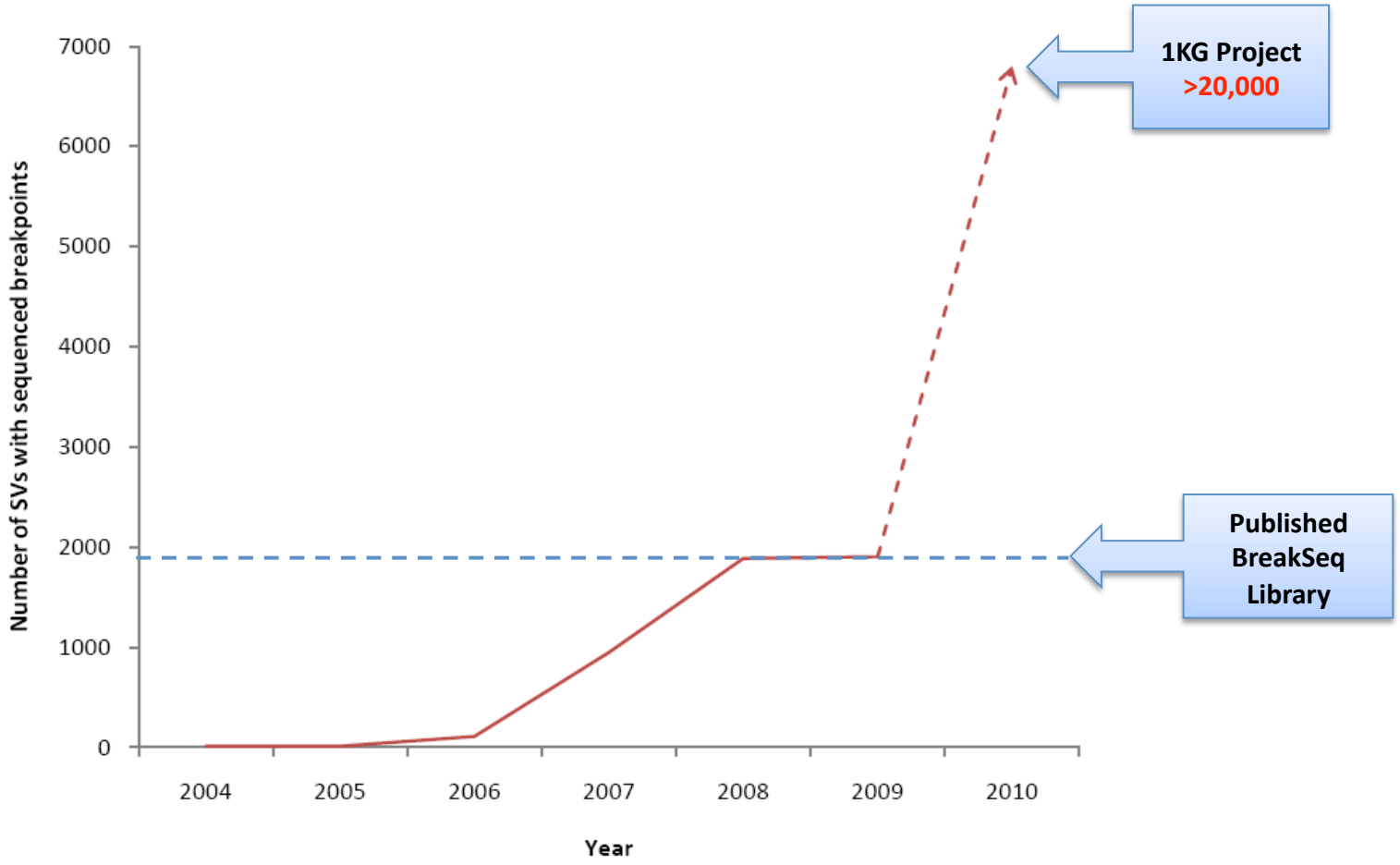
[Lam et al., ('10) *Nat. Biotech.*]

SV Breakpoint Library



[Lam et al., ('10) *Nat. Biotech.*]

SVs with sequenced breakpoints

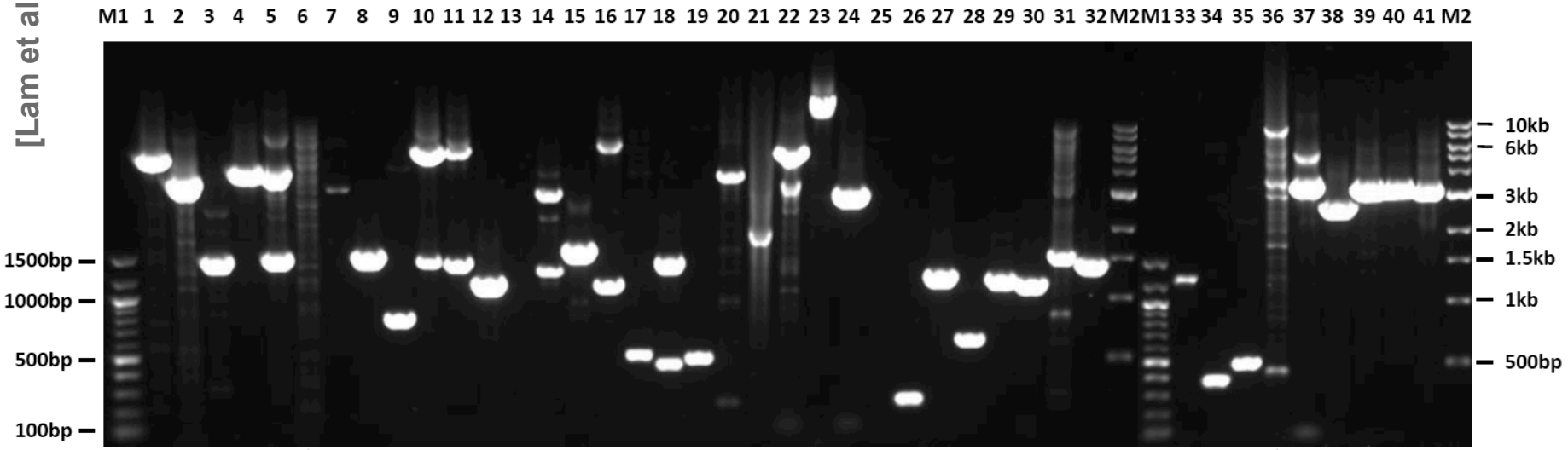


[Lam et al., ('10) *Nat. Biotech.*]

Validation for Identified SVs

[Lam et al., ('10) *Nat. Biotech.*]

Personal genome (ID)	Ancestry	High support hits (>4 supporting hits)	Total hits (incl. low support)
NA18507*	Yoruba	105	179
YH*	East Asian	81	158
NA12891 [1000 Genomes Project, CEU trio]	European	113	219

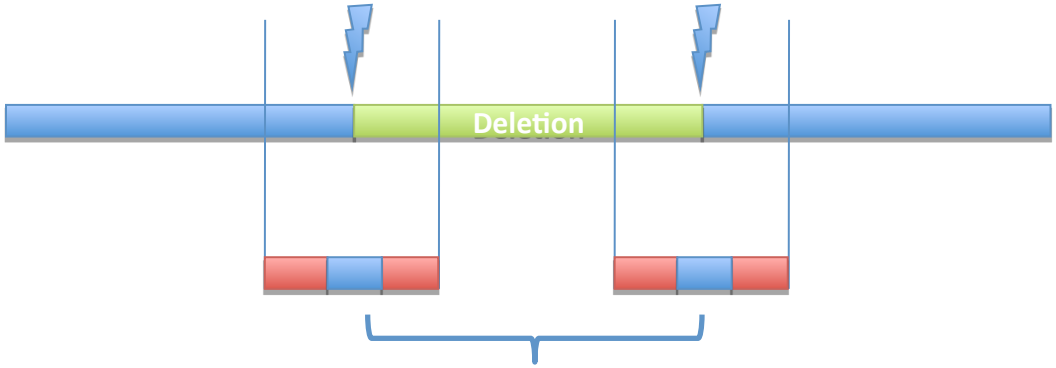


48 positive outcomes out of 49 PCRs that were scored in NA12891:
98% PCR validation rate (for low and high-support events)
12 amplicons sequenced in NA12891: all breakpoints confirmed

Mechanism Assignment Pipeline

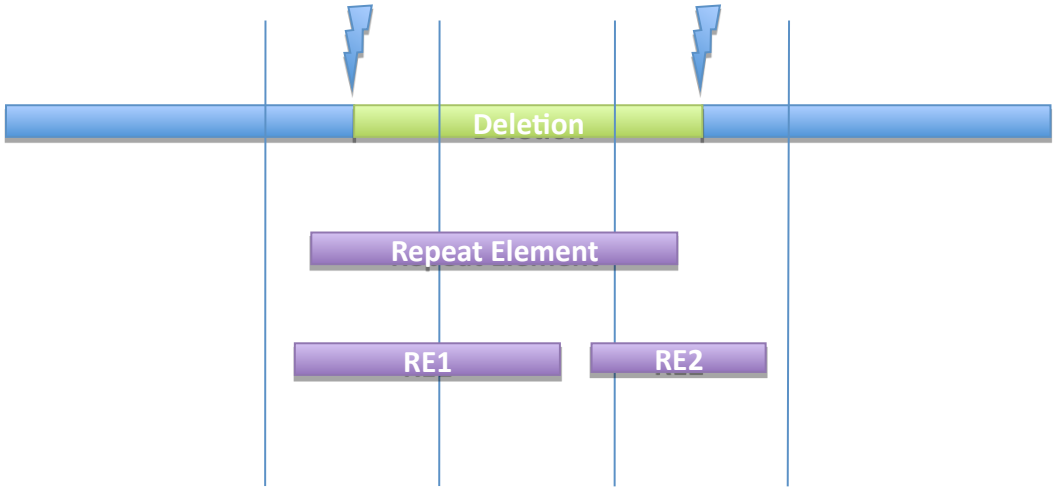
SV Mechanism Classification

NAHR



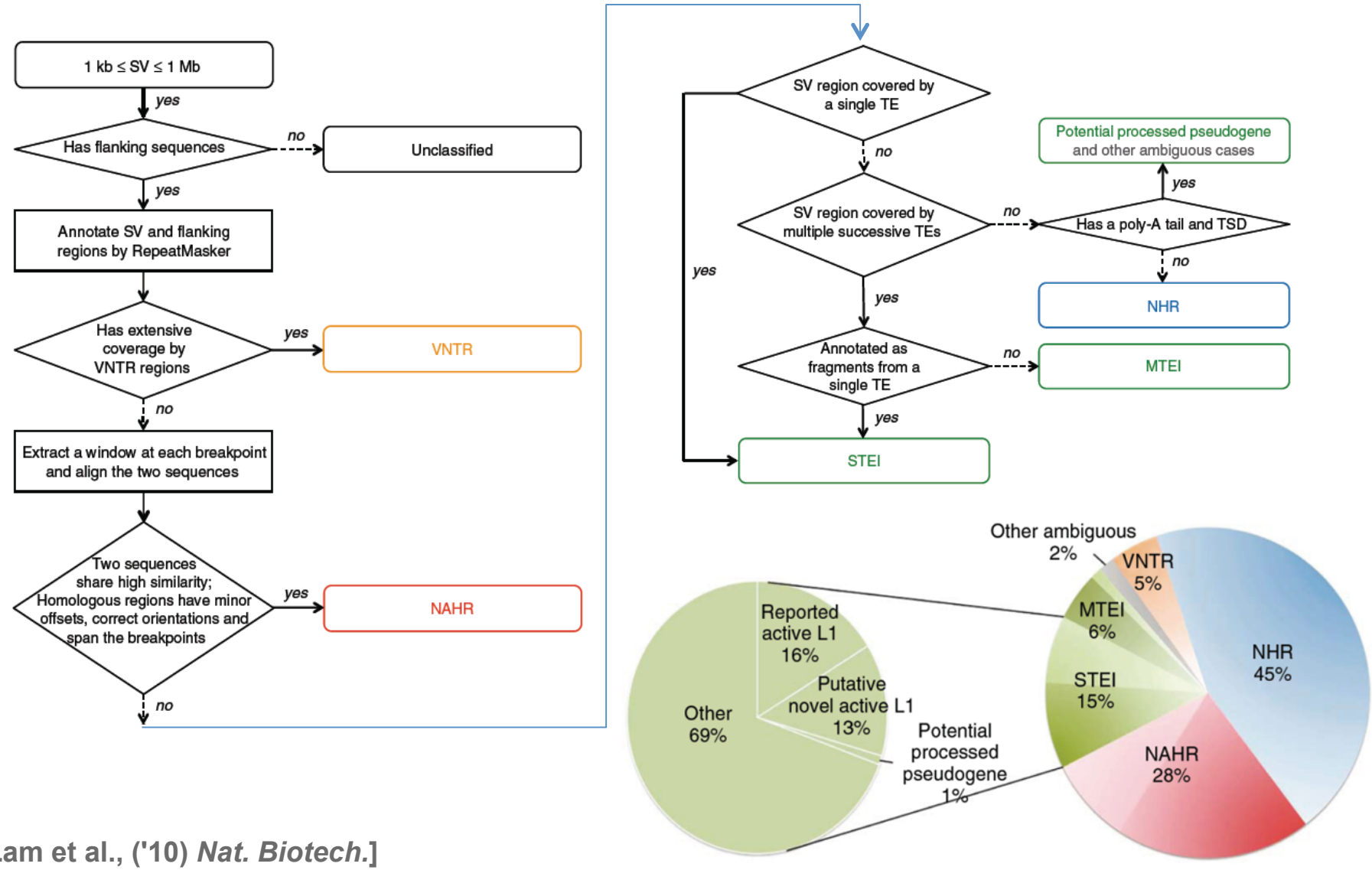
Single RETRO

Multiple RETRO



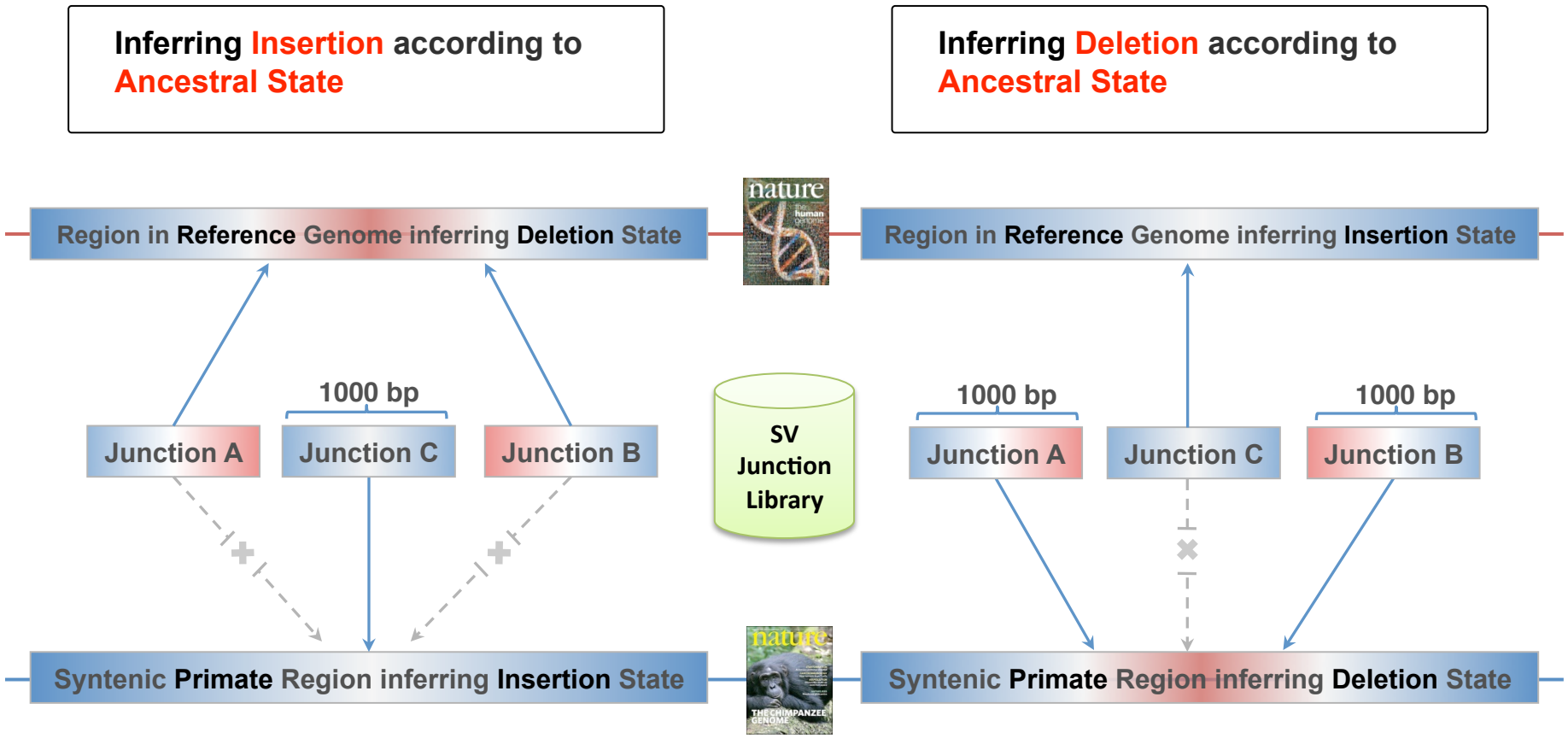
[Lam et al., ('10) *Nat. Biotech.*]

SV Mechanism Classification



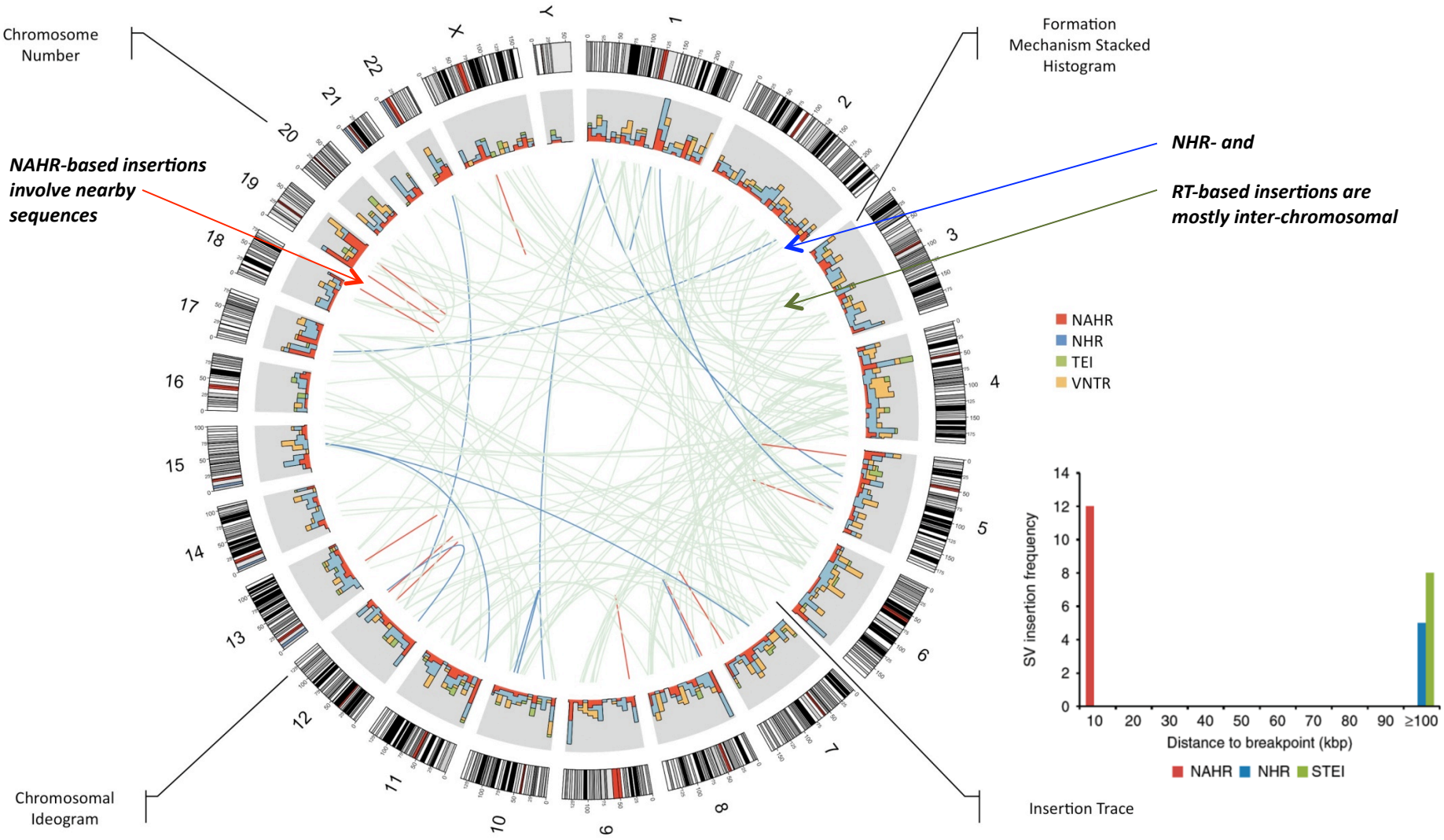
[Lam et al., ('10) *Nat. Biotech.*]

SV Ancestral State Analysis



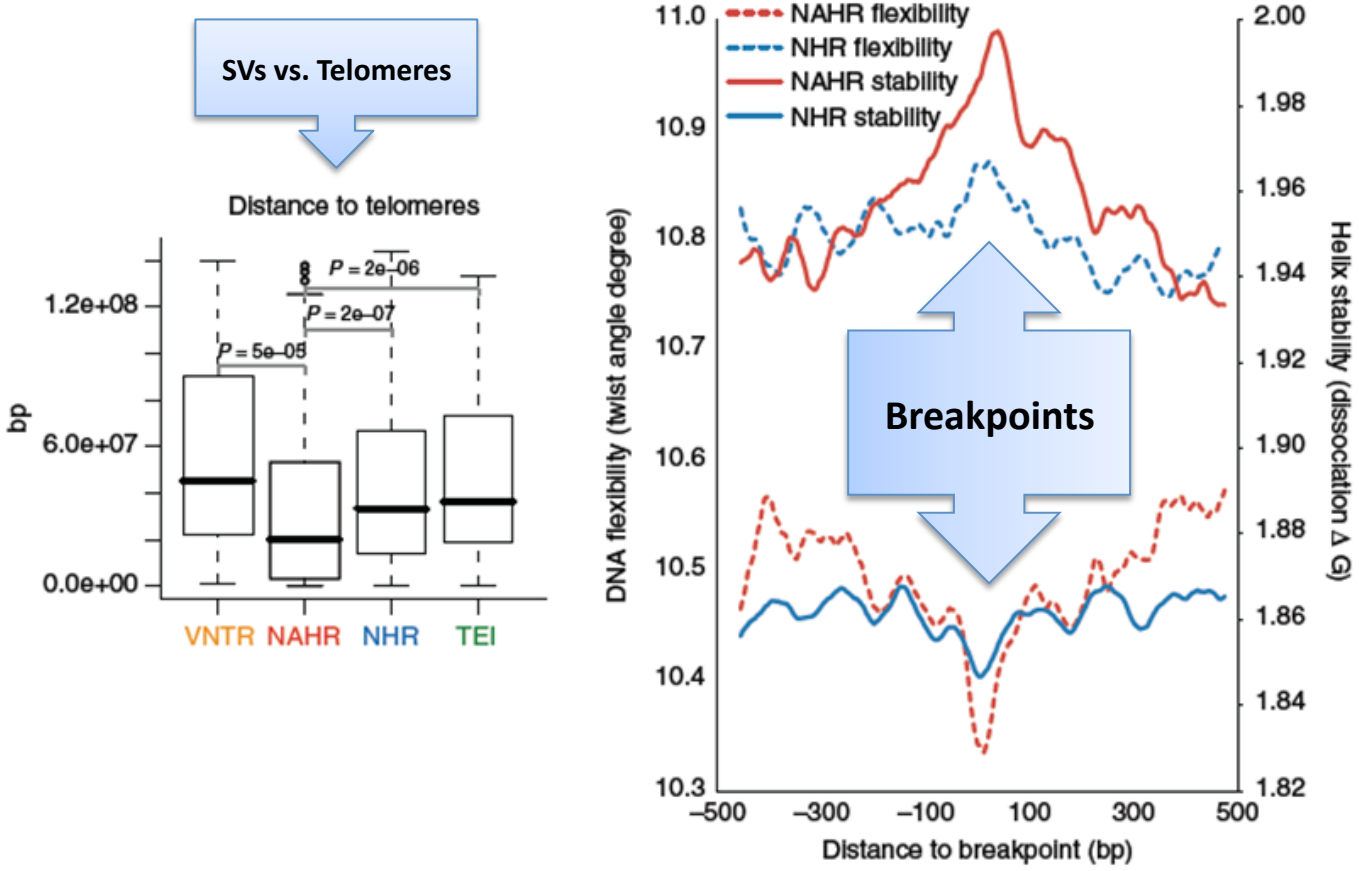
[Lam et al., ('10) *Nat. Biotech.*]

SV Insertion Traces



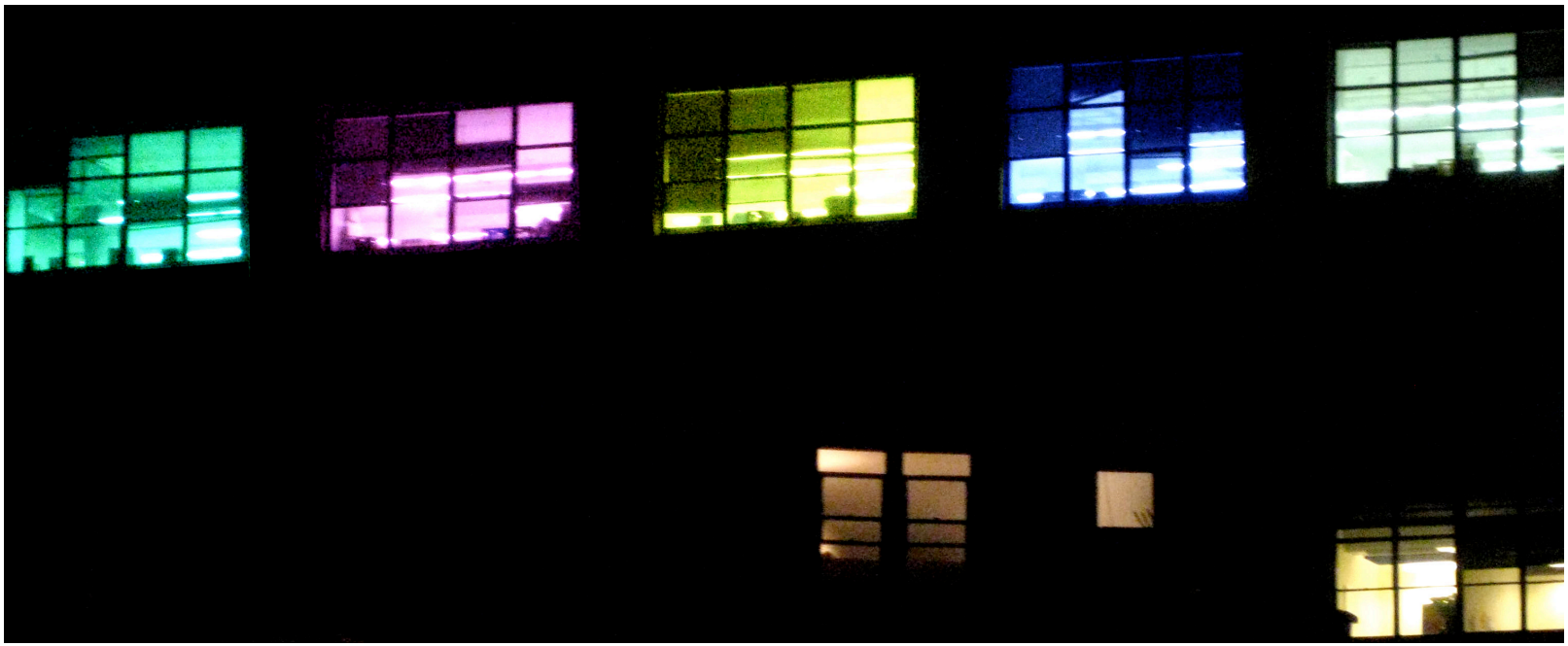
[Lam et al., ('10) *Nat. Biotech.*]

Breakpoint Features Analysis



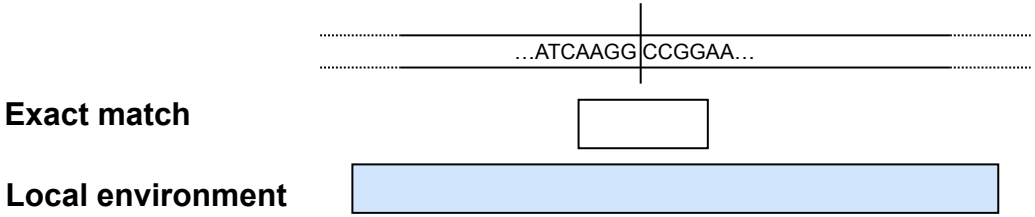
[Lam et al., ('10) *Nat. Biotech.*]

Large-scale Analysis of Repeated Blocks in the Genome (SDs & CNVs)

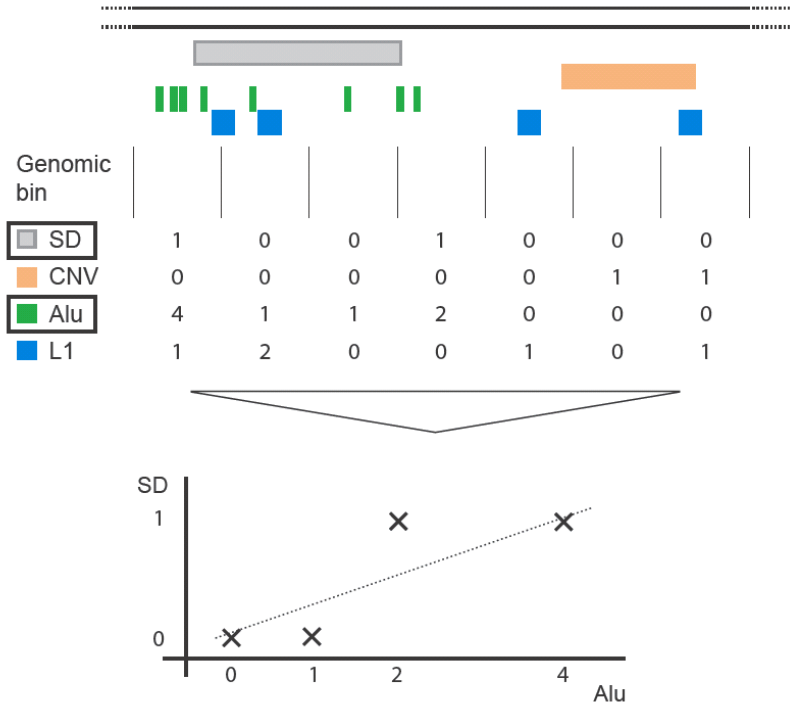


PERFORM LARGE SCALE CORRELATION ANALYSIS TO DETECT REPEAT SIGNATURES OF SDs AND CNVs

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

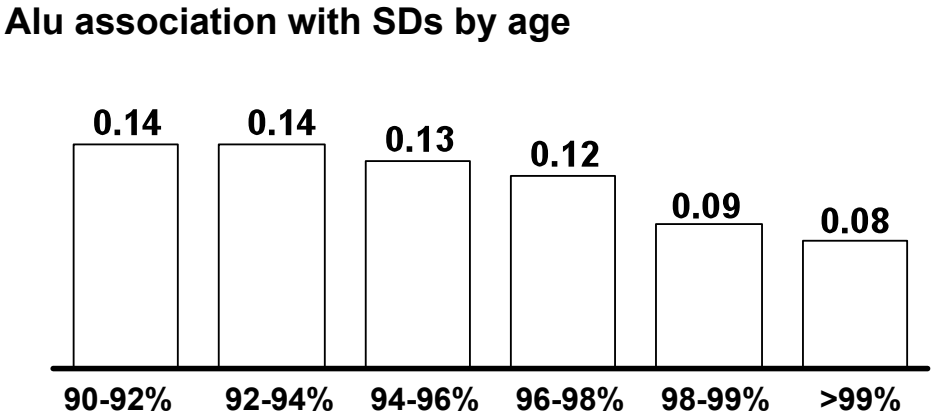


1. Survey a range of genomic features
2. Count the number of features in each genomic bin (100kb)
3. Calculate correlations / enrichments using robust stats



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

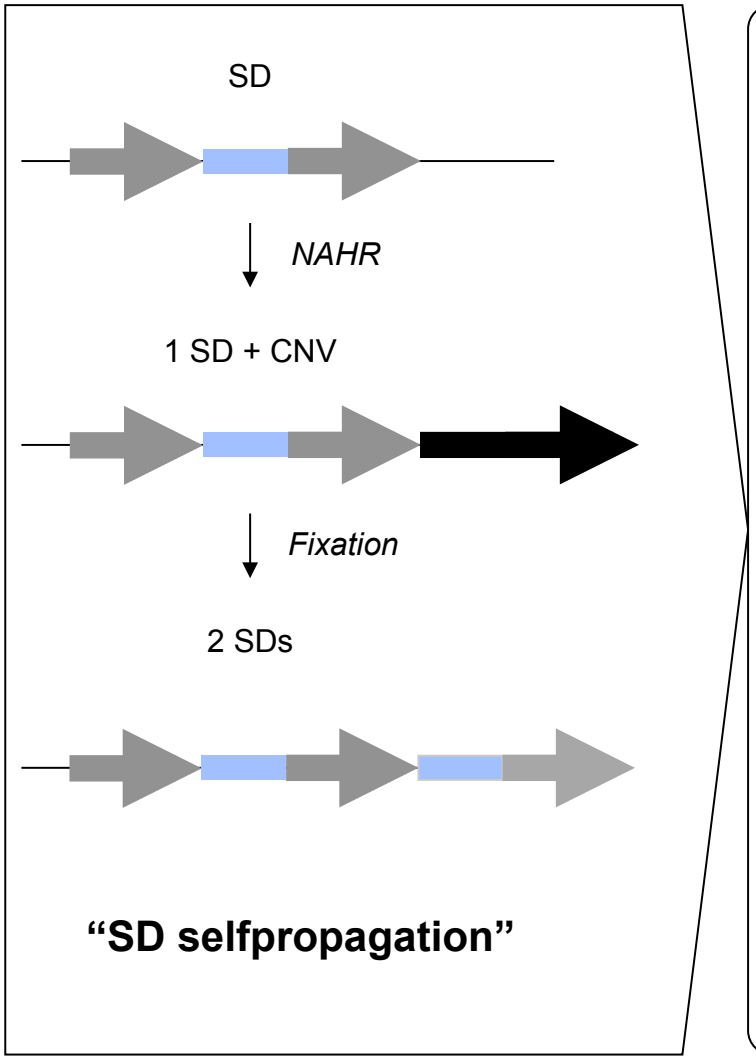
OLDER SDs ARE MUCH MORE LIKELY TO BE FORMED BY ALU ELEMENTS



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

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

SDs COLOCALIZE WITH EACH OTHER, PARTICULARY THOSE OF THE SAME AGE



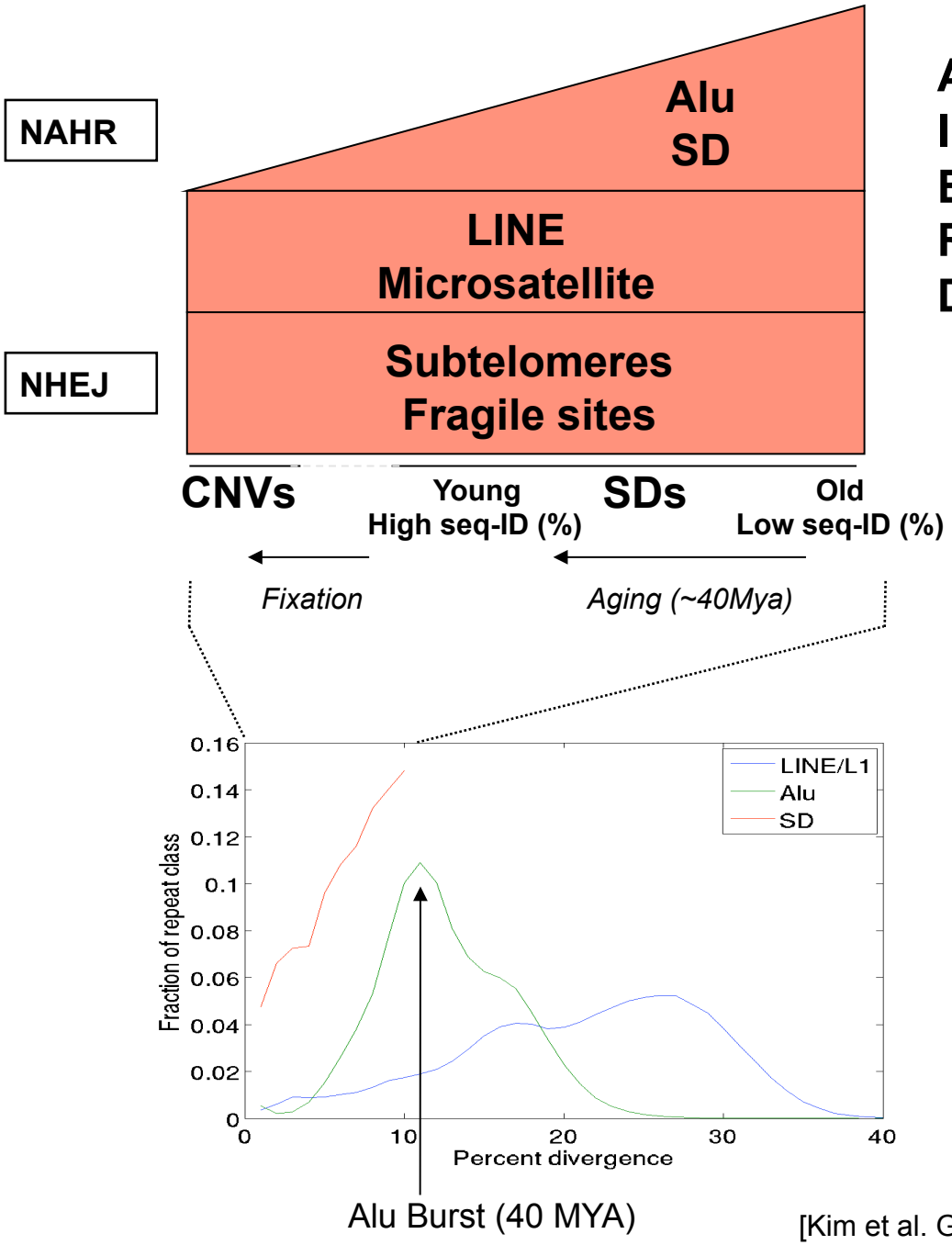
Corollary

- SDs can mediate NAHR and lead to the formation of CNVs
- CNVs can become fixed and then be SDs

- We find (not shown) that SD location tends to be correlated with other SDs
- Furthermore, SDs co-localize most with SDs of a similar age.

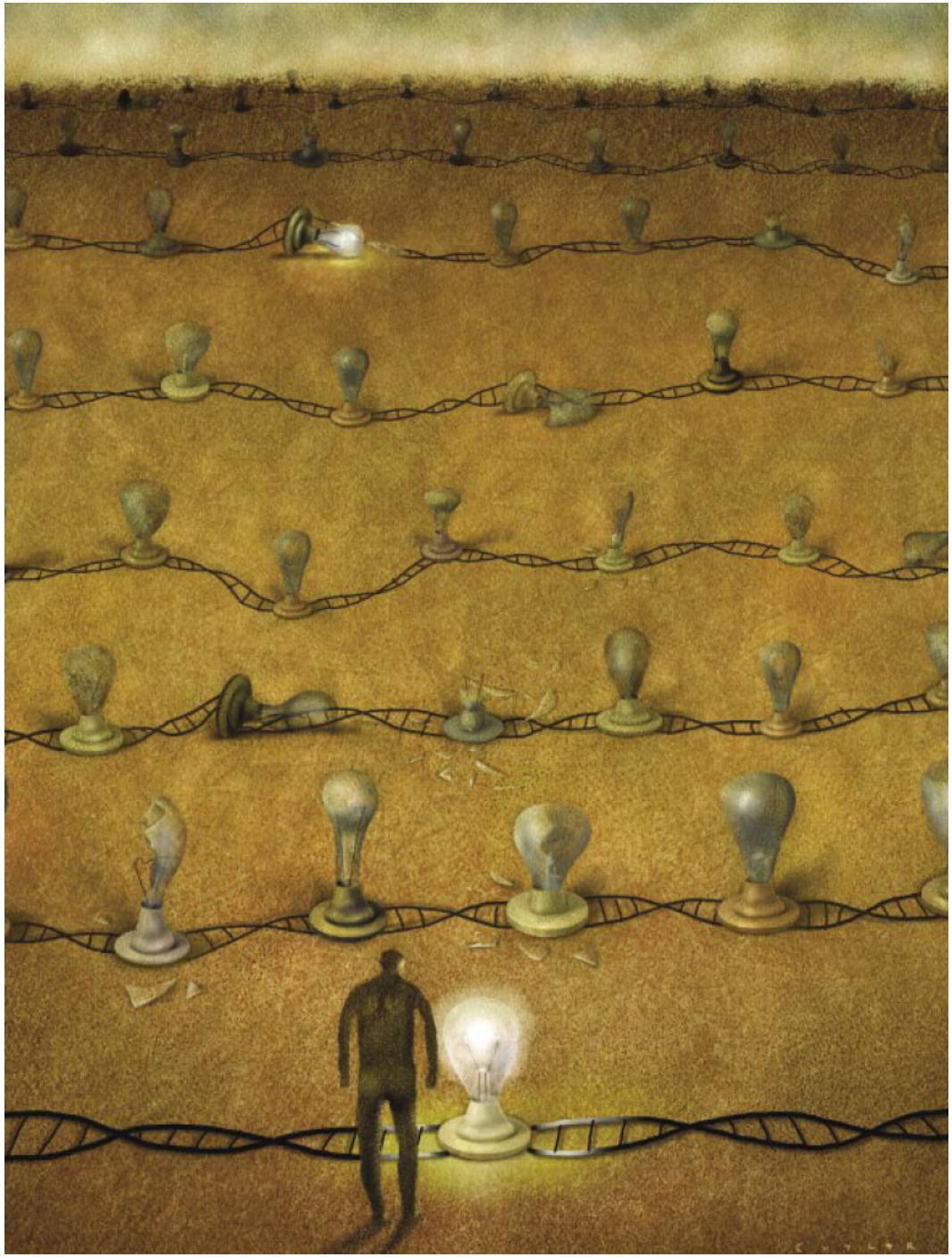
[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]

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



- About 40 million years ago there was a burst in retrotransposon activity
- The majority of Alu elements stem from that time
- This, in turn, led to rapid genome rearrangement via NAHR
- The resulting SDs, could create more SDs, but with Alu activity decaying, their creation slowed

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



Formal Annotation based on Comparative Genomics: Pseudogenes

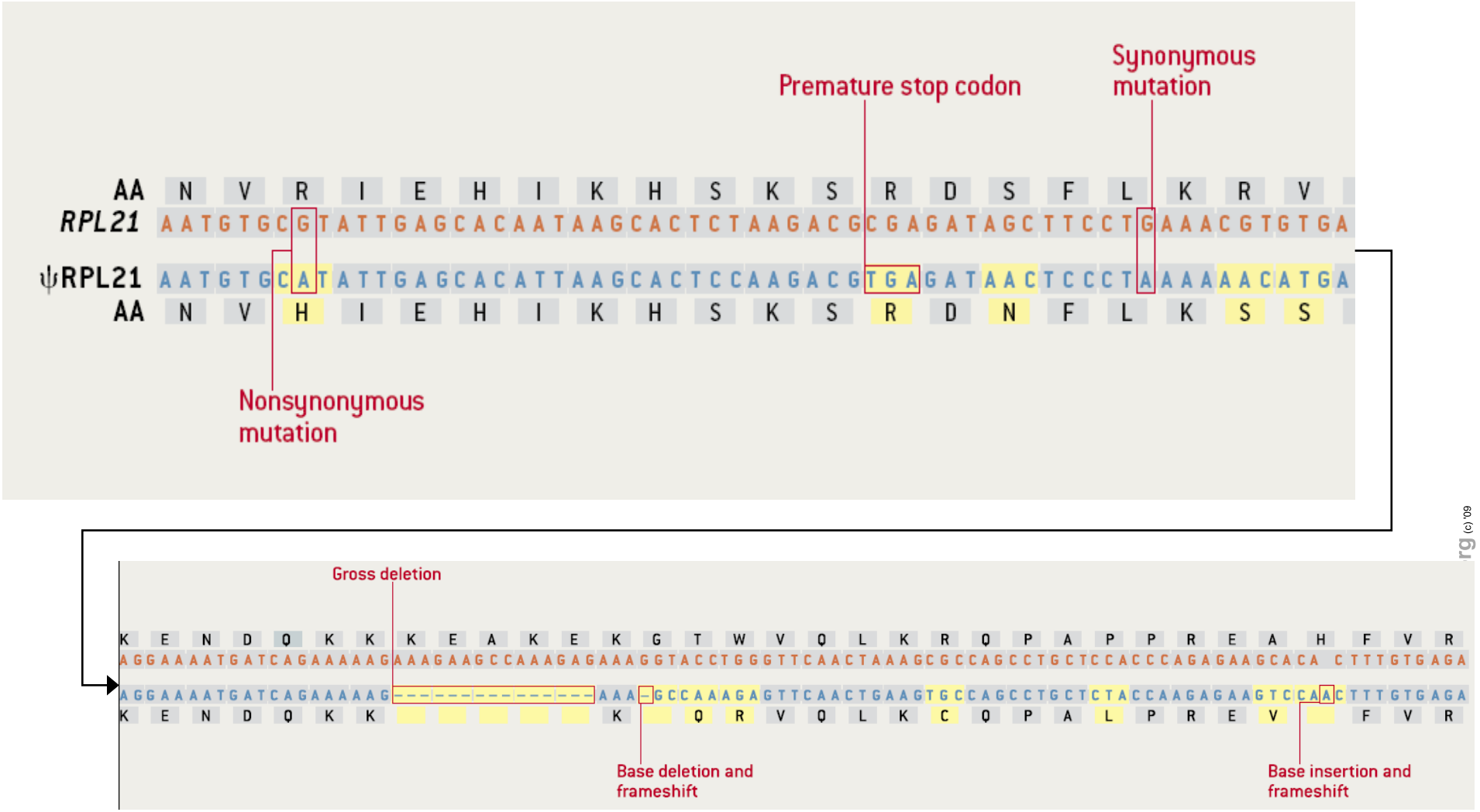
Illustration from Gerstein & Zheng (2006). Sci Am.

Pseudogenes are among the most interesting intergenic elements

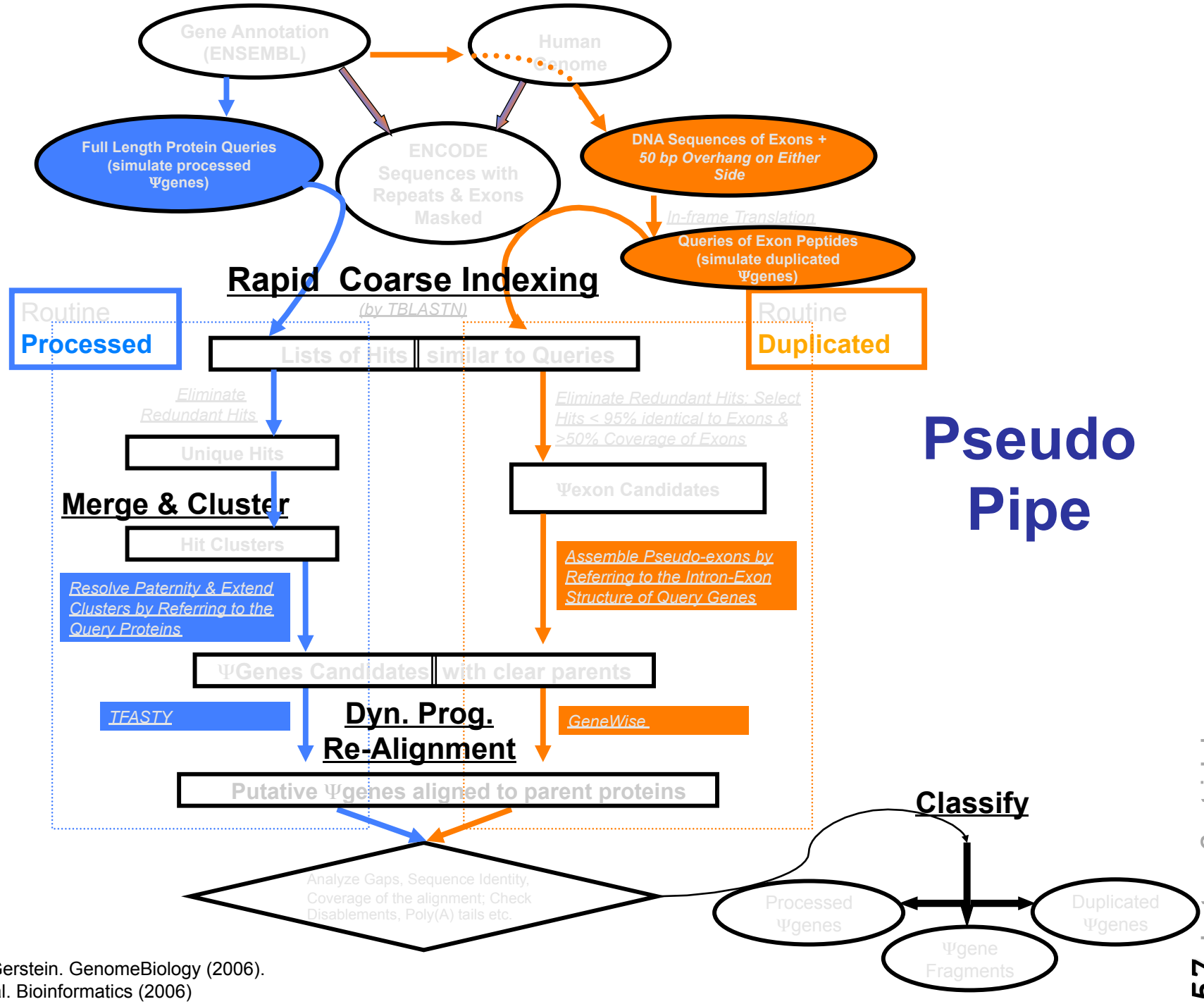
- Formal Properties of Pseudogenes (ΨG)
  - Inheritable
  - Homologous to a functioning element
  - Non-functional*
    - No selection pressure so free to accumulate mutations
      - Frameshifts & stops
      - Small Indels
      - Inserted repeats (LINE/Alu)
    - **What does this mean?** no transcription, no translation?…

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

Identifiable Features of a Pseudogene (ψRPL21)

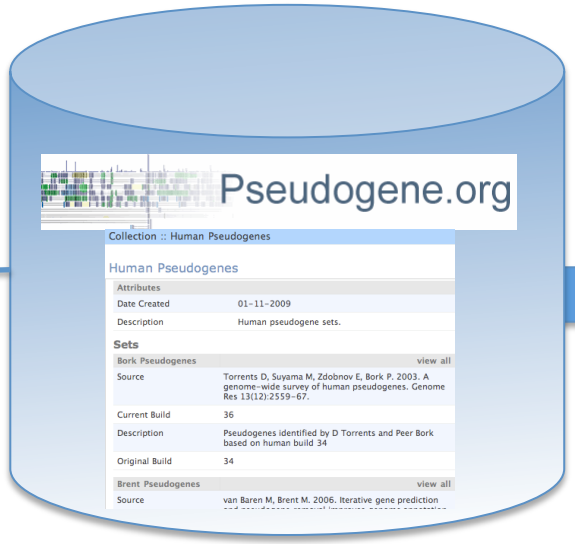


Gerstein & Zheng. Sci Am 295: 48 (2006).



Pseudo Pipe

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



Flat Files ⟷ DAS

Table Browser

tables.pseudogene.org ⟷ UCSC Genome Browser

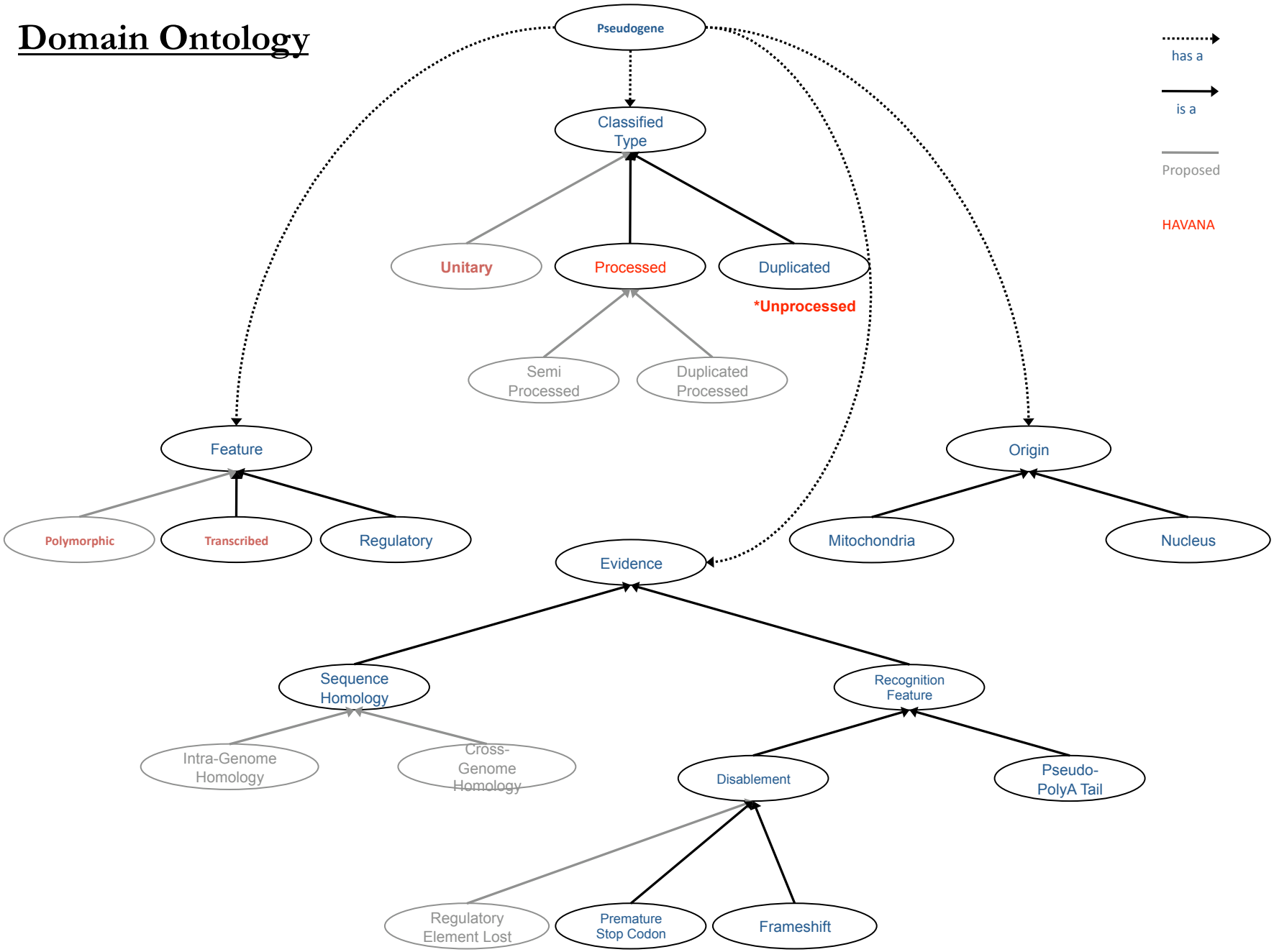
- 12 eukaryotic species
  - Human, mouse, rat, chimp…
  - 100,052 pseudogenes
- 64 prokaryotic species
  - 6,412 pseudogenes

~23K in recent pipeline run

- 13+ unique human sets

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

Domain Ontology



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

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

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

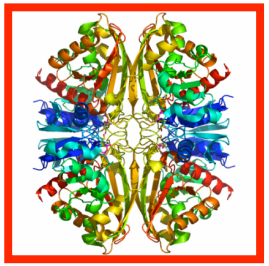
- Chronology of Sets
  1. Encode Pilot 1%
  2. Ribosomal Protein pseudogenes
  3. **Glycolytic Pseudogenes**
  4. **Unitary pseudogenes**
  5. Polymophic pseudogenes
- Totals (May '09)
  - Automatic pipeline currently gives ~23K
  - Manually Annotated ~8K

Specific Pseudogene Assignments: Glycolytic Pseudogenes

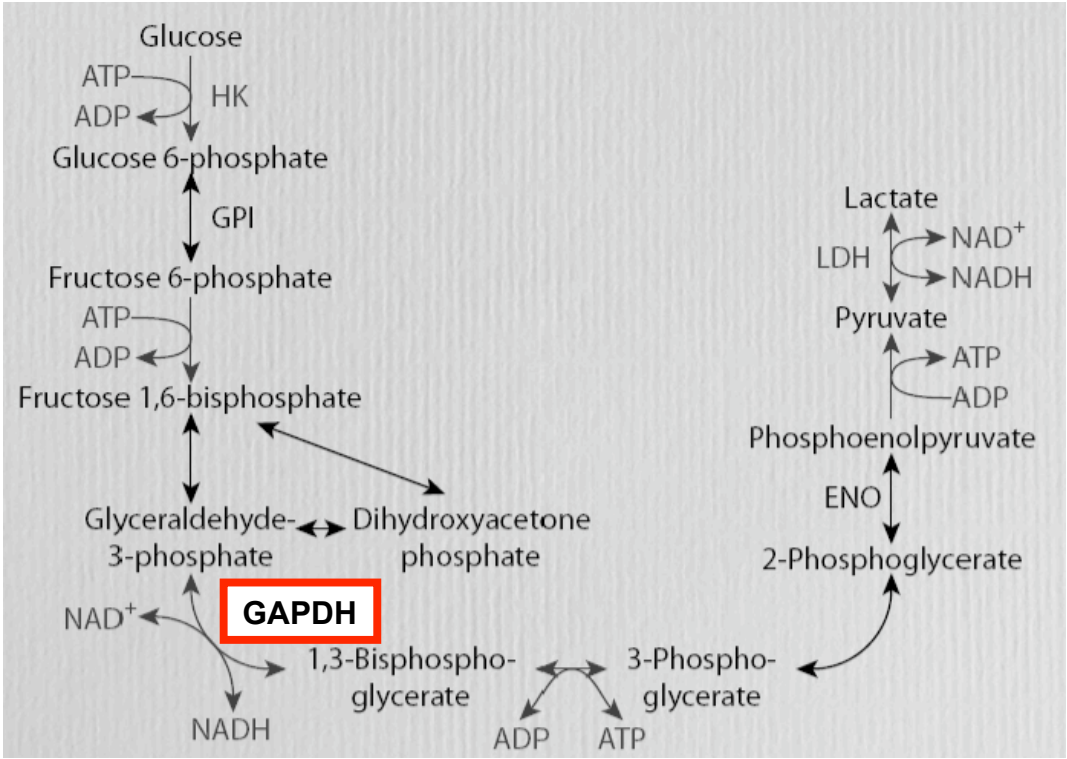
Number of pseudogenes for each glycolytic enzyme

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

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



Processed/Duplicated

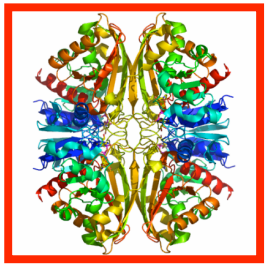


	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60/2	47/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0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

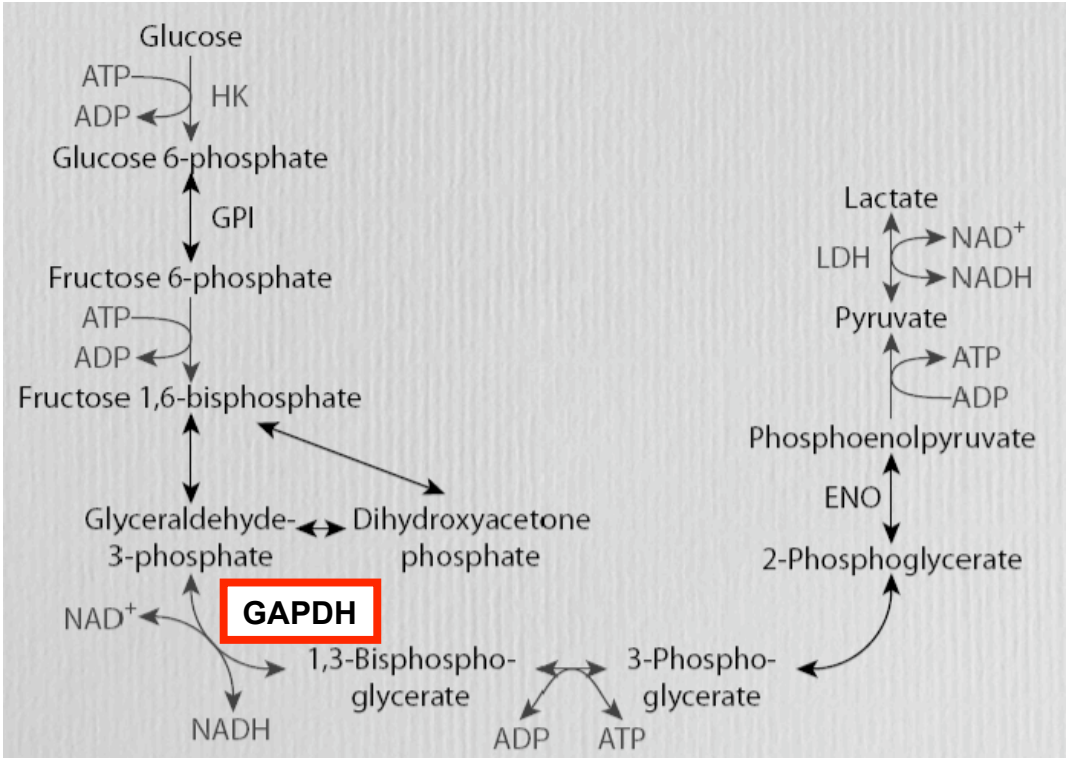
Number of pseudogenes for each glycolytic enzyme

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

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



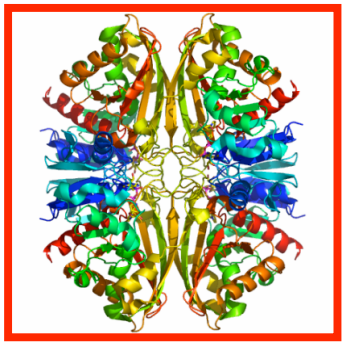
Processed/Duplicated



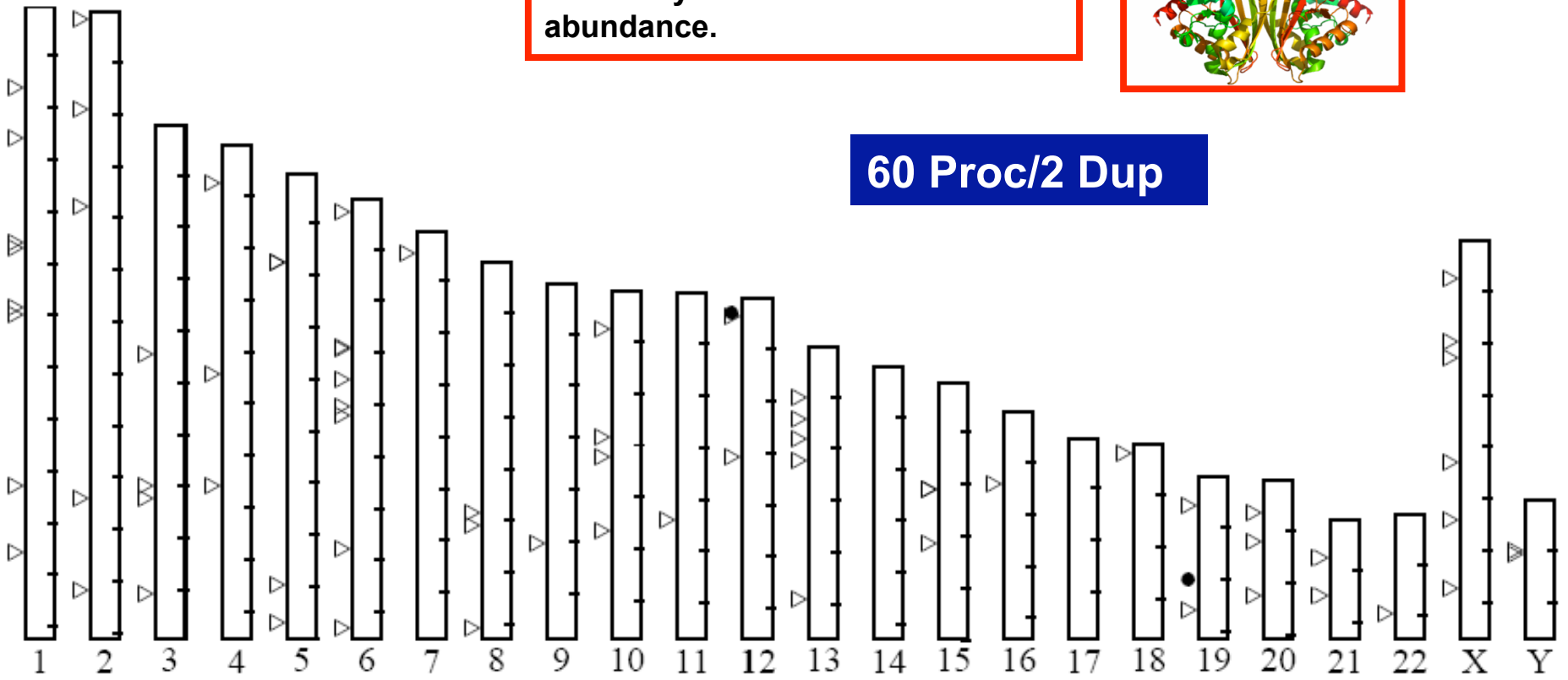
	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60 Proc/2 Dup	7/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1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

Distribution of human GAPDH pseudogenes

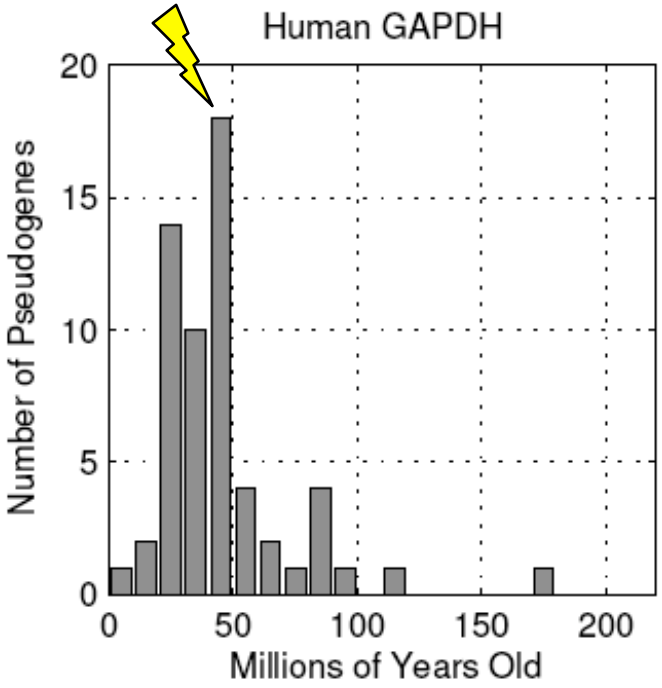
Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.



60 Proc/2 Dup



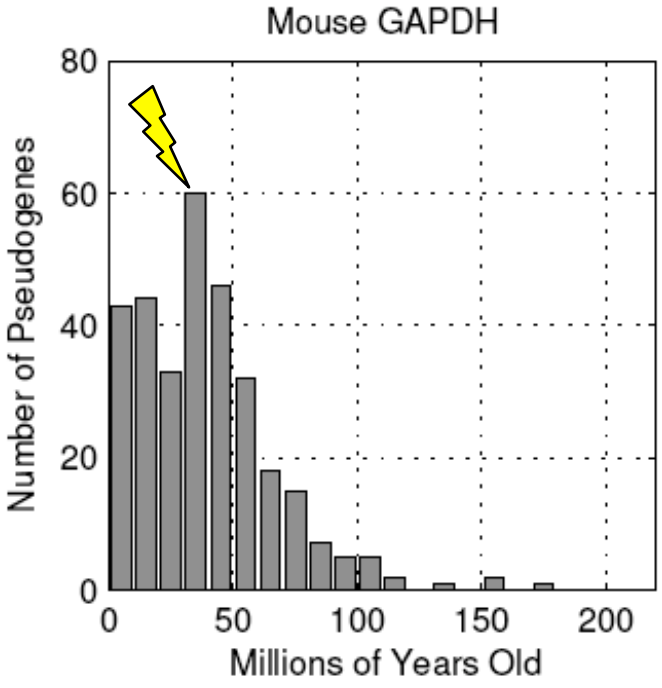
[Liu et al. BMC Genomics ('09, in press)]

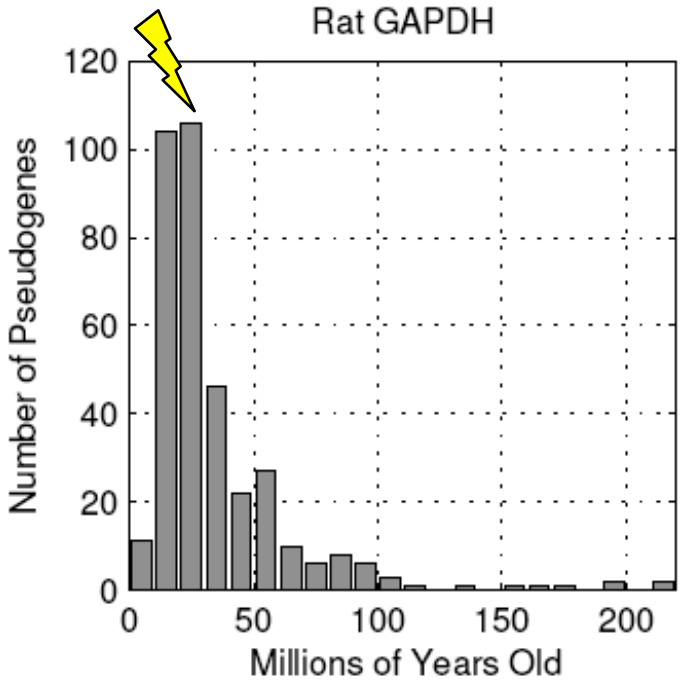


Burst of Retrotran-spositional Activity

Aproximate Age of GAPDH pseudogenes

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

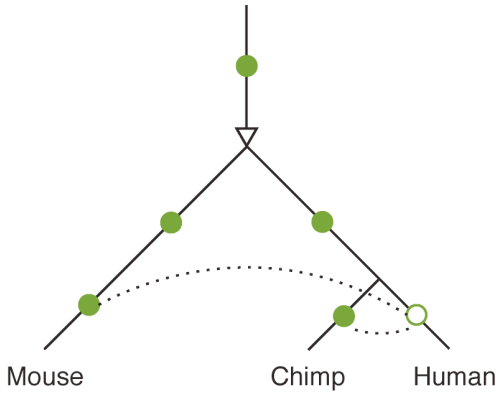


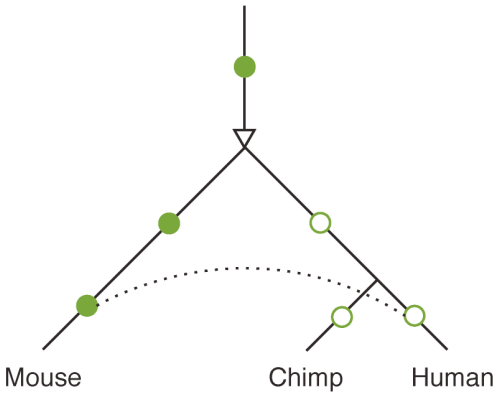


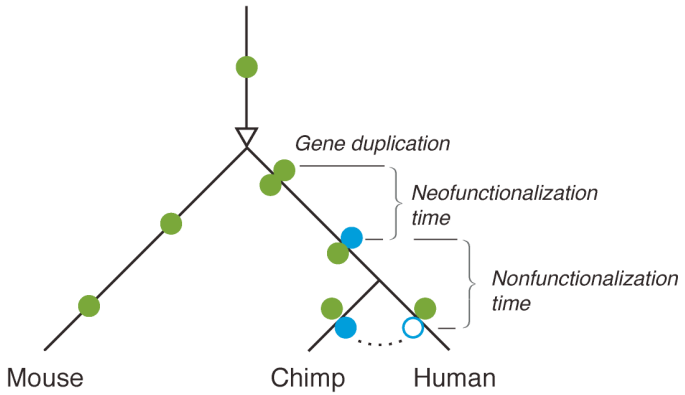
[Liu et al. BMC Genomics ('09)]

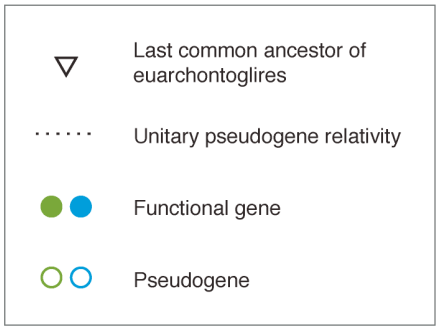
Specific Pseudogene Assignments: Unitary Pseudogenes

Unitary pseudogenes





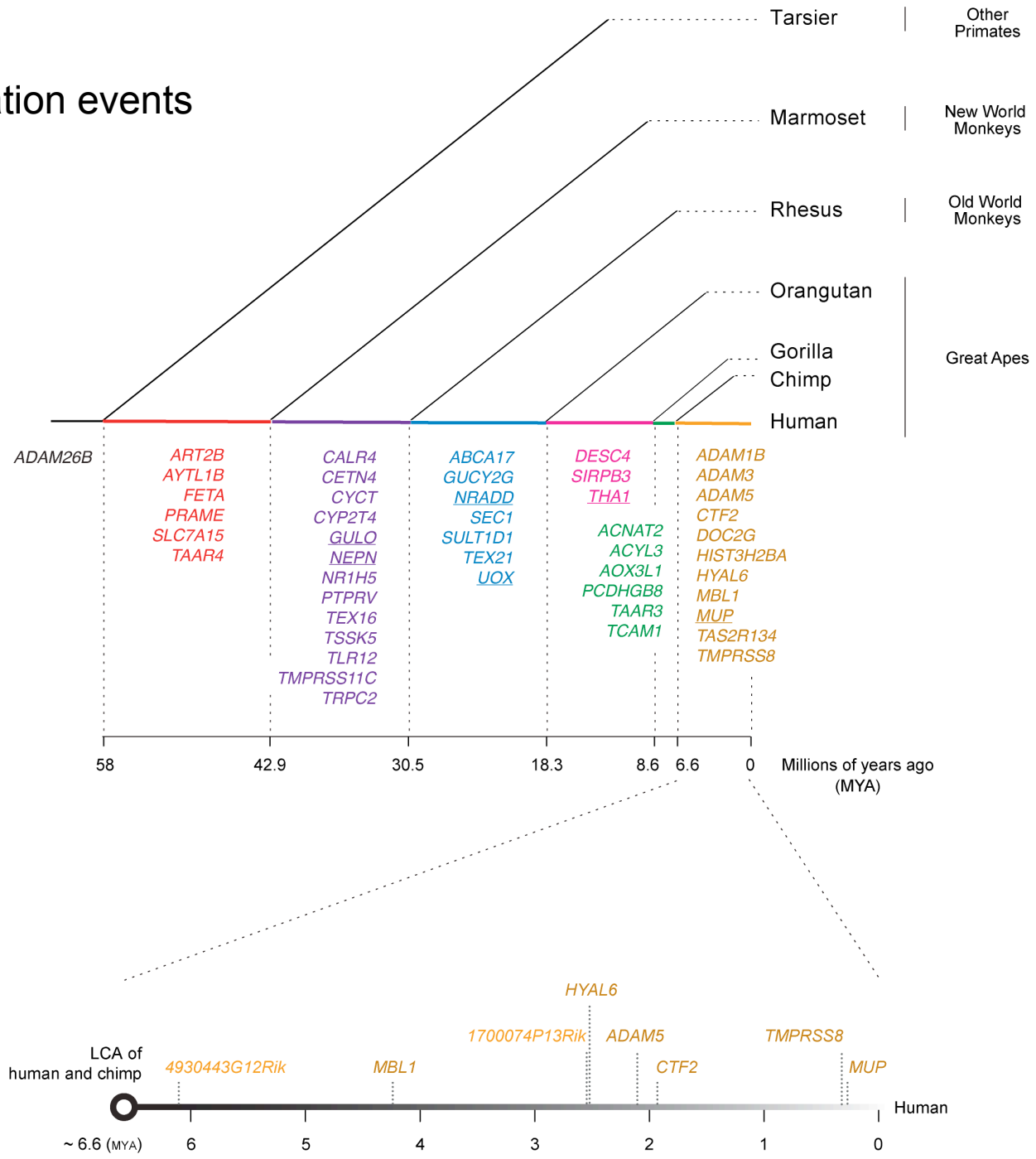




- Unprocessed pseudogenes with no functional counterparts in the same genome
- Assignment is "relative"
- 76 in the human genome relative to the mouse

[Zhang et al. GenomeBiology (in press, '10)]

Dating the pseudogenization events



[Zhang et al. ('10) GenomeBiology]

11 Polymorphic Pseudogenes

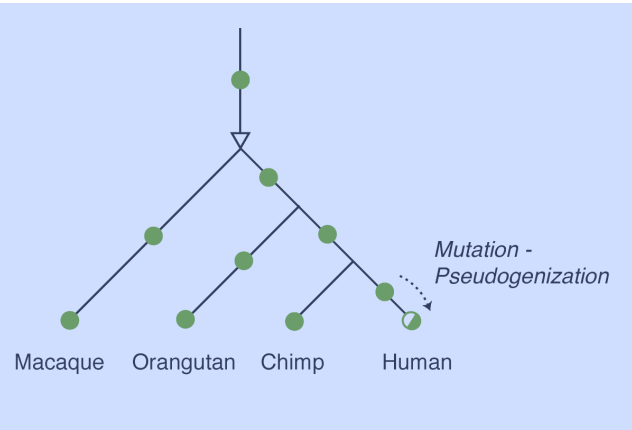
Table 2. Human polymorphic pseudogenes

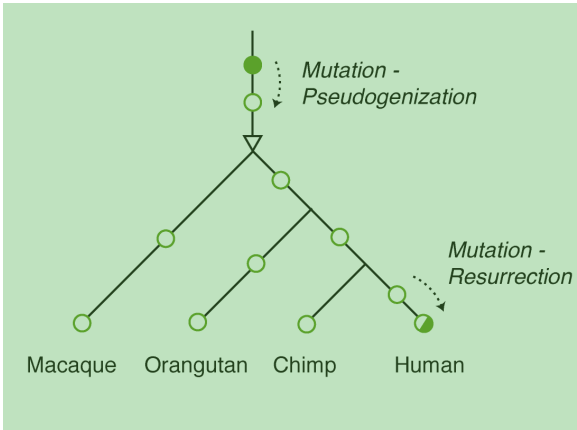
Gene	CDS-disruptive mutation		dbSNP ID [3]	HapMap SNP ID
	Change [1]	Location [2]		
Nonsense mutation				
FBXL21	taT (Y) → taA	chr5+:135,300,350	rs17169429 (+27)	rs17169429 (+27)
FCGR2C	Cag (Q) → Tag	chr1+:159,826,011	rs3933769 (−60)	rs3933769 (−60)
GPR33	Cga (R) → Tga	chr14−:31,022,505	rs17097921	rs17097921
SEC22B	Caa (Q) → Taa	chr1+:143,815,304	rs2794062	rs16826061 (+95)
SERPINB11	Gaa (E) → Taa	chr18+:59,530,818	rs4940595	rs4940595
TAAR9	Aaa (K) → Taa	chr6+:132,901,302	rs2842899	rs2842899
Frame-shift mutation				
CASP12	ΔCA	chr11−:104,268,394-5	rs497116 (−67)	rs497116 (−67)
KRTAP7-1	ΔT	chr21−:31123841	rs35359062	rs9982775 (−20)
PSAPL1	∇A	chr4−:7,487,457	rs58463471	rs4484302 (+441)
TMEM158	∇A	chr3−:45,242,396	rs11402022	rs33751 (+725)
TPSB2	ΔC	chr16−:1,219,240	rs2234647	rs2745145 (−1771)

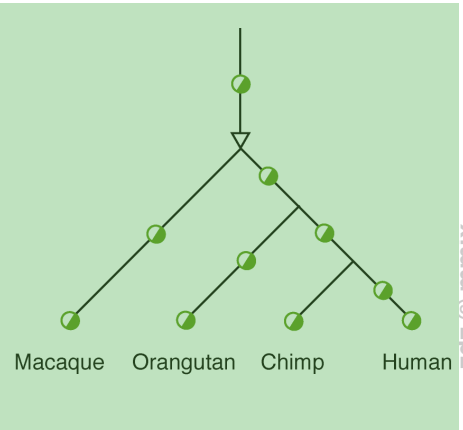
[Zhang et al. ('10) GenomeBiology]

Polymorphic pseudogenes (3 with allele frequency data)

CDS-disrupted gene	GPR33	SERPINB11	TAAR9
Disruptive mutation [1]	Cga (R) → Tga	Gaa (E) → Taa	Aaa (K) → Taa
dbSNP ID	rs17097921	rs4940595	rs2842899
Genomic location	chr14−:31,022,505	chr18+:59,530,818	chr6+:132,901,302
Disrupted codon position [2]	140 (332)	89 (388)	61 (344)
Reference allele in human	T	T	T
Reference allele in other primates [3]	C	T	T
Allele frequency [4]			
Test statistic for HWE in the meta-population [5]	0.285 ($P = 0.867$)	8.659 ($P = 0.013$)	0.071 ($P = 0.965$)



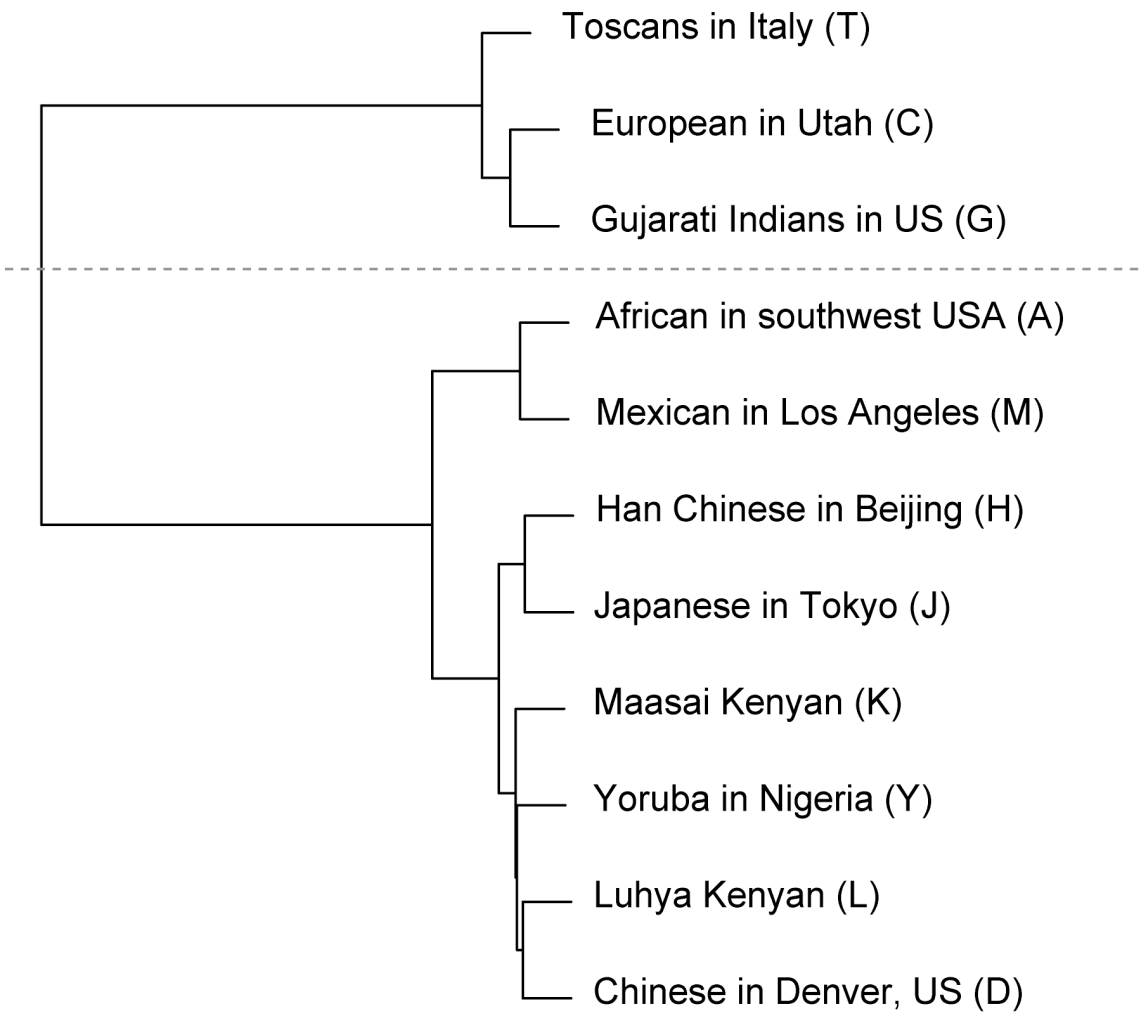




3 SNPs not found to be under recent positive selection....

[Zhang et al. ('10) GenomeBiology]

F_{st} hierarchical clustering for rs4940595 in SERPINB11

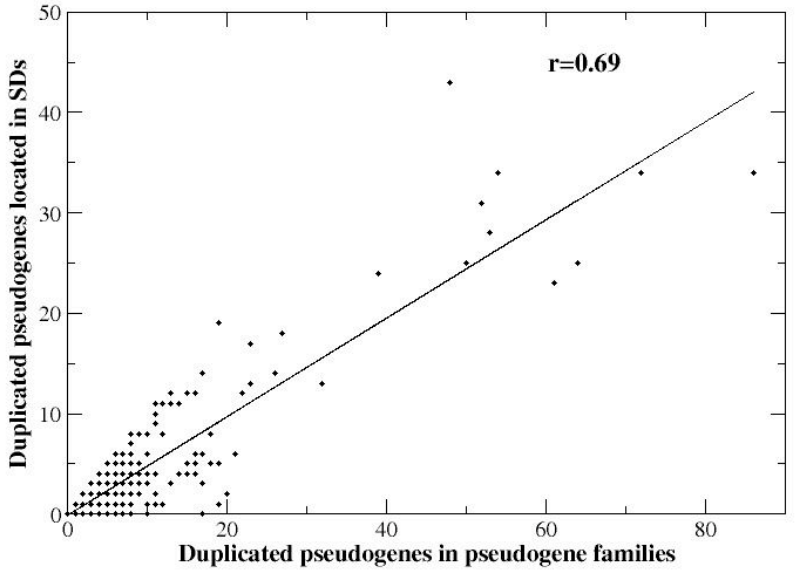


▪▪▪▪but population structure at rs4940595—the difference in the allelic frequencies in different populations—could be result of different selective regimes that the same allele at rs4940595 is subjected to in different population subdivisions.

Integration of Pseudogenes with Other Features

Pseudogene families and Segmental Duplications (SDs)

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

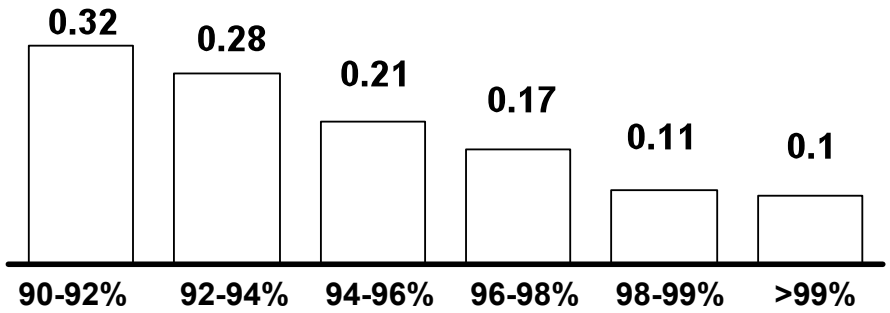


- **SDs comprise ~5% of the human genome but contain ~18% genes, 46% duplicated pgenes and 22% processed pgenes**
- Correlation above consistent with the observation that SDs contain more pgenes than parent genes
- Also, 431 fully rectifiable breakpoints overlapped with 8 pseudogenes identified by PseudoPipe

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

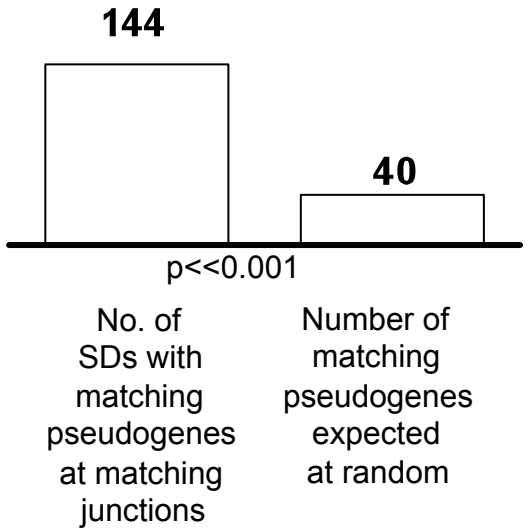
Pseudogenes & CNV/SDs (whole genome, not GAPDH)

Pseudogene association with SDs by age

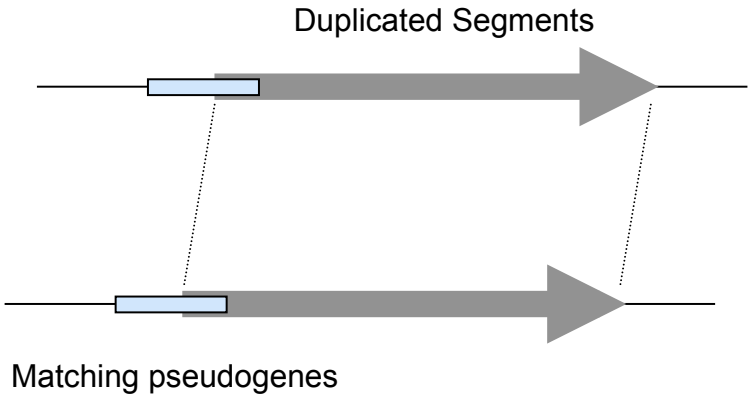


Duplicated pseudogenes associated with SDs, particularly older ones

Processed pseudogenes at SD junctions



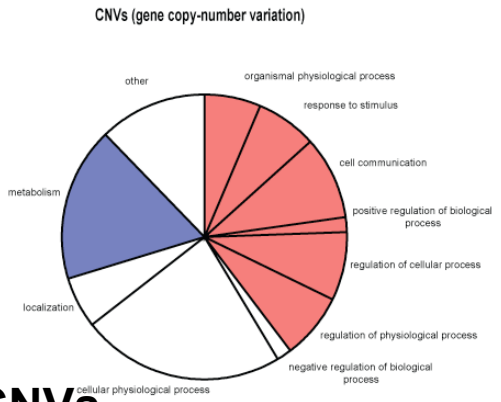
Processed Pseudogenes: serving as repeats for mediating NAHR



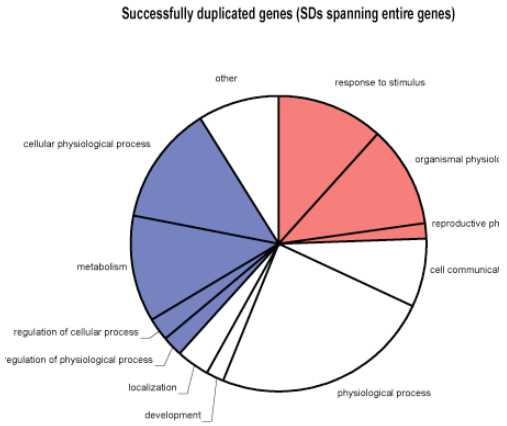
[Kim et al. Gen. Res. ('08), arxiv.org/abs/0709.4200v1]

Association of SDs & CNVs with pseudogenes

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

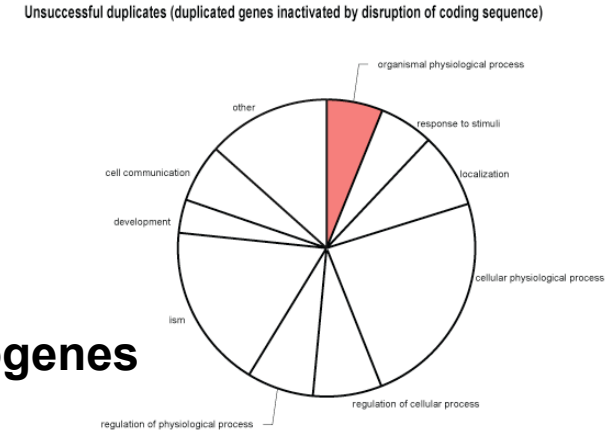


Genes in CNVs



Genes in SDs

GO Categories:
Environ-mental Response
Metabolism

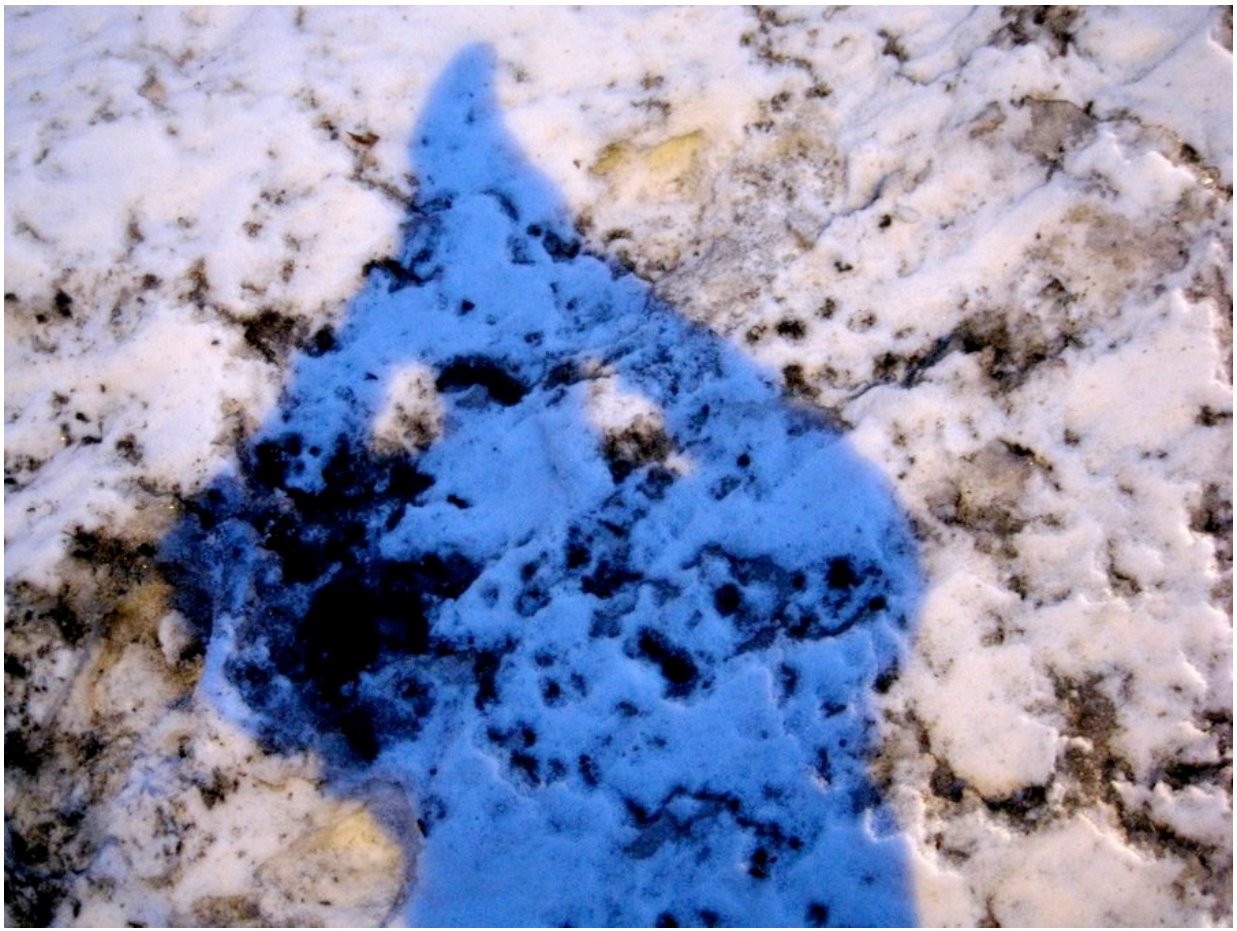


Pseudogenes

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

Summary:

Looking Back Over the Talk



Overview of the Process of Intergenic Annotation

- Basic Inputs
  1. Doing large-scale similarity comparison, looking for repeated or deleted regions
  2. Determining experimental signals for 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

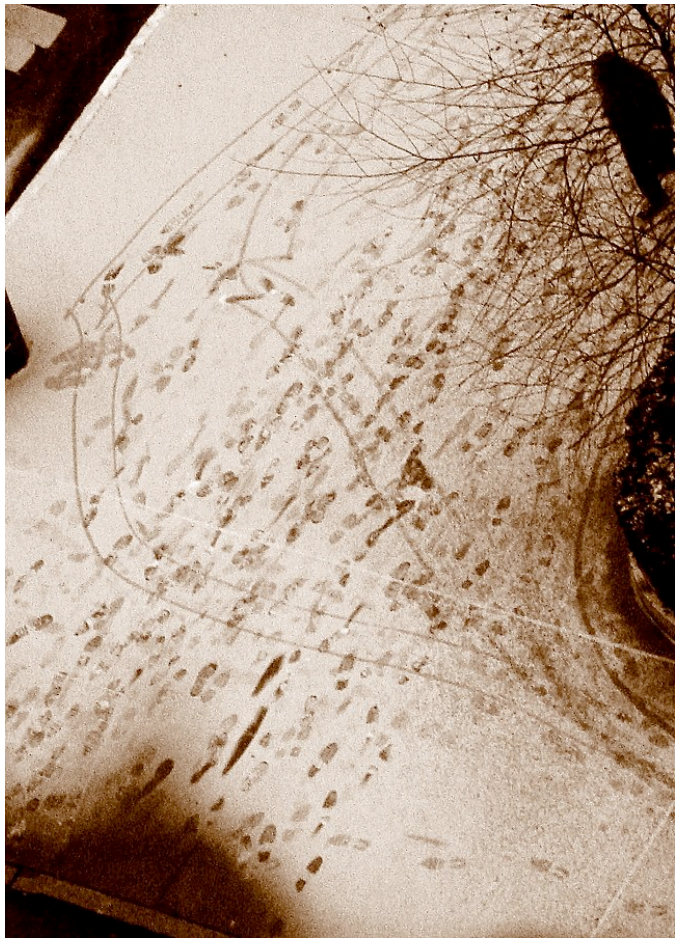
  a. Signal Processing: removing artifacts, normalizing, window averaging

  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



- Variable Blocks in the Genome (SVs,SDs)
  - Calling SVs with various approaches (MSB, PEMer, ReSeqSim, BreakSeq)
  - Analyzing mechanism of formation for precisely resolved breakpoints & on a large-scale over the genome
- Pseudogenes
  - Pattern-match assignment tools
  - Focus on different specific groups – glycolytic, unitary
  - Polymorphic Pseudogenes
  - Inter-relating Pseudogenes with SDs & SVs

Identifying Structural Variants in the Human Population

- MSB
  - Mean-shift segmentation approach following grad. of PDF
  - Equally applied to aCGH and depth of coverage of short reads
- ReSeqSim
  - Efficiently simulating assembly of a representative variant
  - Shows that best reconstruction has a combination of long, med. and short reads
- PEMer
  - Detecting Variants from discordantly placed paired-ends
  - Simulation to paramaterize statistical model
- BreakSeq
  - Building a breakpoint library
  - Running against reads in newly seq. genome to genotype new SVs
  - Building a pipeline for characterizing breakpoints according to SV mechanisms

Analysis of Duplication in the Genome: SVs and SDs

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

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

- Pseudogene Assignment Technology
  - ◊ Pipeline + DB
  - ◊ Ontology
  - ◊ Pseudofam analysis of Pseudogene Families
- Annotation of Human Genome
  - ◊ Pipeline draft (20K) + Hybrid Approach
- Glycolytic pseudogenes
  - ◊ Great variation in number, with GAPDH the largest
  - ◊ Synteny & dating shows most GAPDH ones are recent, resulting from retrotranspositional bursts
- Unitary pseudogenes
  - – Continuous disablement
  - – A few polymorphic in human population
- Association with SDs & SVs
  - – As expected, duplicated pseudogenes associated with SDs and processed pseudogenes like Alus are near SD junctions

Acknowledgements

Zhengdong Zhang
Y J Liu
YK Lam
J Du
X Mu
A Abyzov
J Korbel
L Wang

M Snyder
S Weissman
P Kim
S Balasubramanian
D Zheng

J Karro
R Bjornson
N Carriero
D Greenbaum
A Urban
P Cayting
E Khurana
Zhaolei Zhang
A Stuetz
A Tanzer
A Frankish
J Harrow
T Hubbard
R Harte

+ ENCODE, modENCODE, 1K Genomes

GenomeTECH.gersteinlab.org
Pseudogene.org

More Information on this Talk

SUBJECT: GenomeTechAnnote

DESCRIPTION:
6th Intl. Symp. on Bioinformatics Research & Applications (ISBRA), U Conn, Storrs, CT, 2010.05.24, 9:00-10:00; [**iOISBRA**] (Long GenomeTechAnnote talk, building on [I:**IBM**] . Should take 60' with questions.)

MORE DESCRIPTION:
Talk 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)

.

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** .