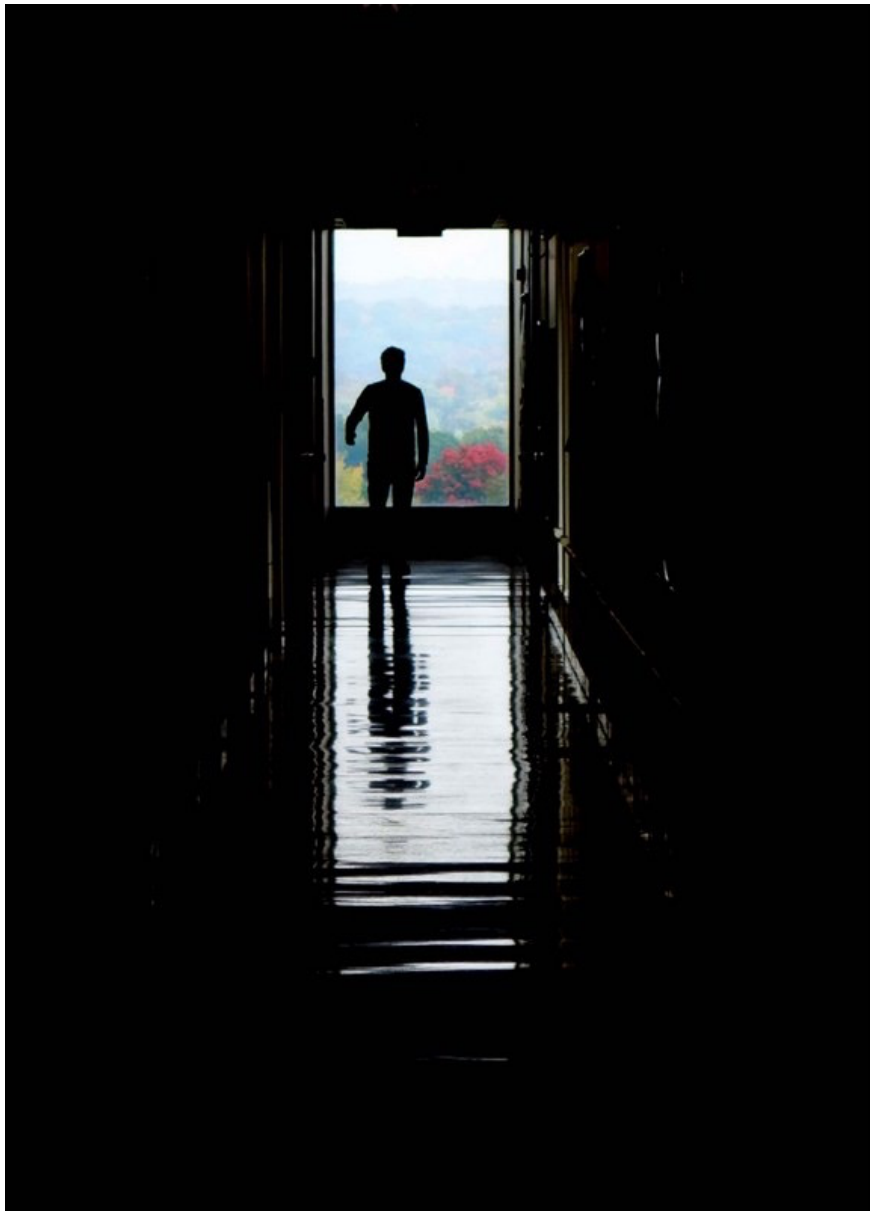


Human Genome Analysis -

SVs & Pseudogenes, Tricky but Crucial Genomic Features, Targeted by Long-read Sequencing:

Current Short-read Results & Future Prospects

Slides freely downloadable from
Lectures.GersteinLab.org
& “tweetable” (via @markgerstein).
See last slide for references & more info.

M Gerstein

Yale

Why we want long reads?

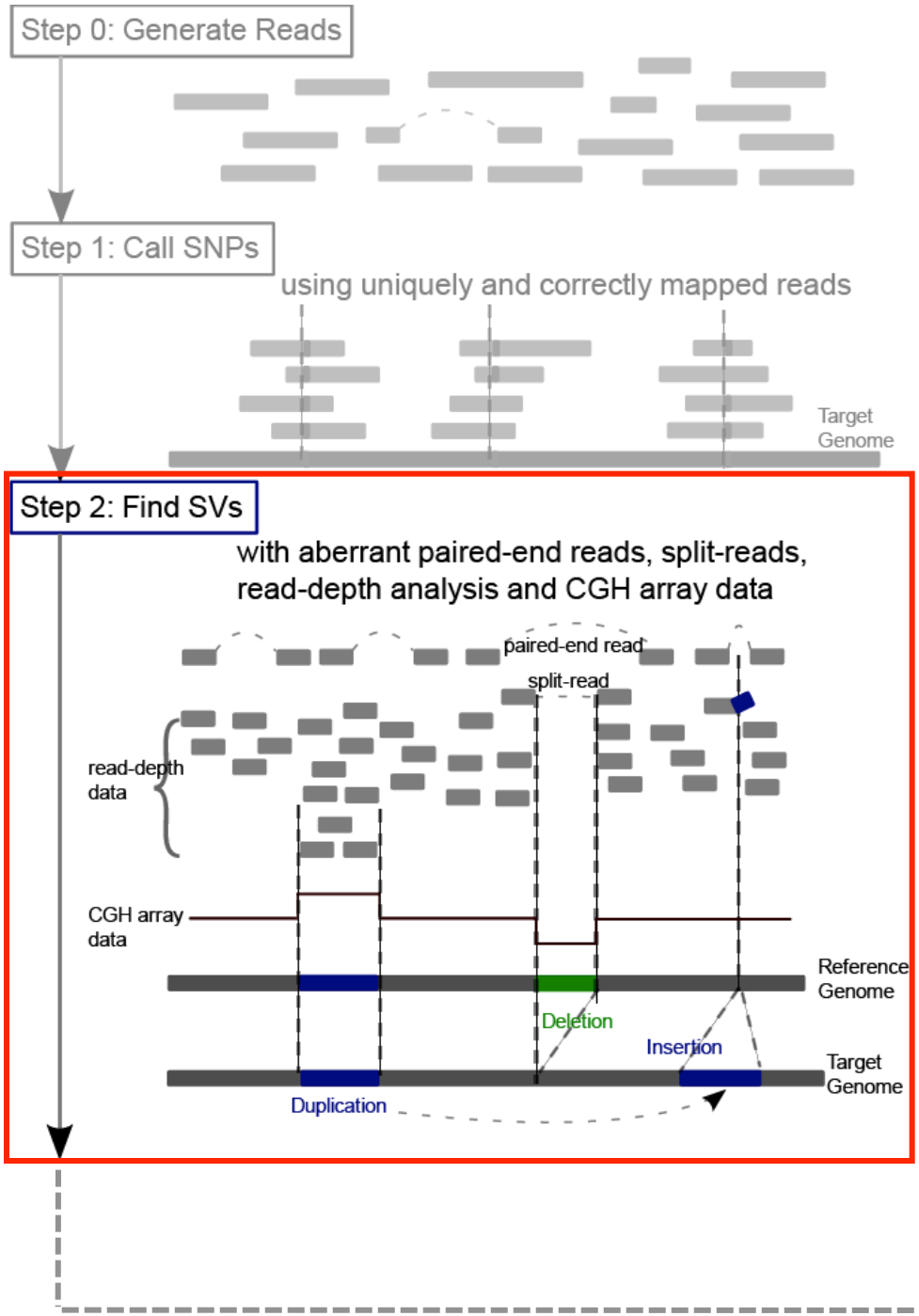
- Ability to better resolve SVs to a nucleotide resolution
  - Breakpoints
  - Complex events
- Ability to better study repetitive elements
  - Repeats not in reference
  - Activity (eg transcription) of repeats
  - Pseudogenes as a type of repeat
- Other stuff
  - Alt. splicing….

Human Genome Analysis – SVs & Pseudogenes, Tricky but Crucial Genomic Features, Targeted by Long-read Sequencing: Current Short-read Results & Future Prospects

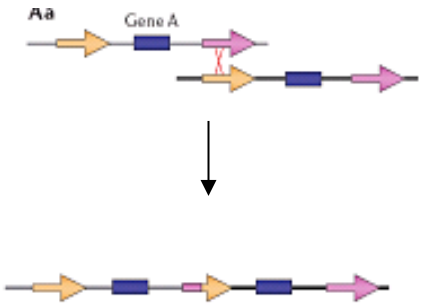
- **SV Breakpoints**
  - ~9K deletions with breakpoints & mechanism classification from 1000G
  - Small subset of tot. deletions, which could be greatly expanded by long reads
  - More nearby SNPs than genomic average.
  - From methylation, Hi-C , & his mods, NAHR breakpoints associated with open chromatin (perhaps occurring w/o replication & division)
  - NHR breakpoints associated w/ sequence microinsertions, templated from later replicating sites, spaced at 2 characteristic distances
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
- **Intersection of Pseudogenes & SVs**
  - Enrichment of SVs in pseudogenes v genes, particularly for NAHR
- **Novel Processed Pseudogenes as a Form of SV**
  - Not in reference but in human population – could be improved by long reads
  - Now found w/ splice junction mapping + clustering of unmapped PEs
  - ~8 per person, often pop. specific
  - Associated w/ G1/M expressed genes
- **Many Pseudogenes with Low Levels of Biochemical Activity**
  - Conservative assignment, mis-map issue, could be improved by long reads
  - ~15% transcribed & 80% w/ some activity

Main Steps in Genome Resequencing

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

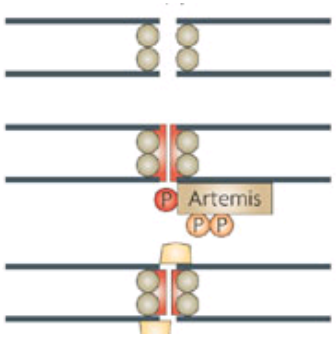


4 mechanisms for SV formation



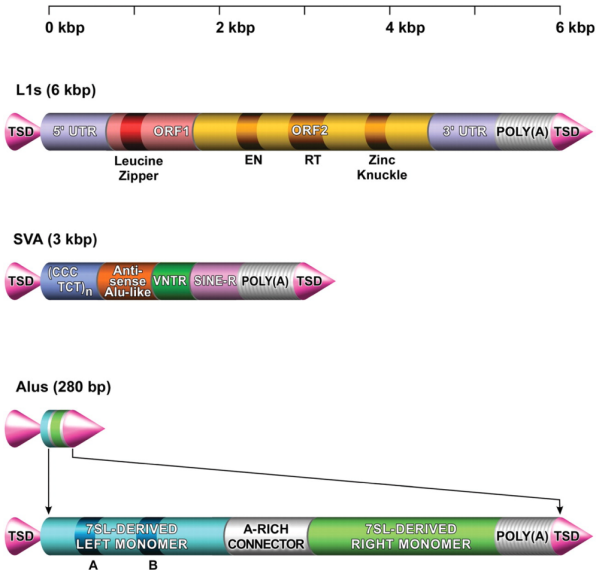
NAHR
(Non-allelic homologous recombination)

Flanking repeat
(e.g. Alu, LINE…)



NHEJ (NHR)
(Non-homologous-end-joining)

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

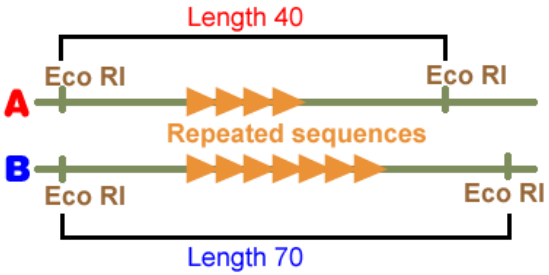


TEI
(Transposable element insertion)

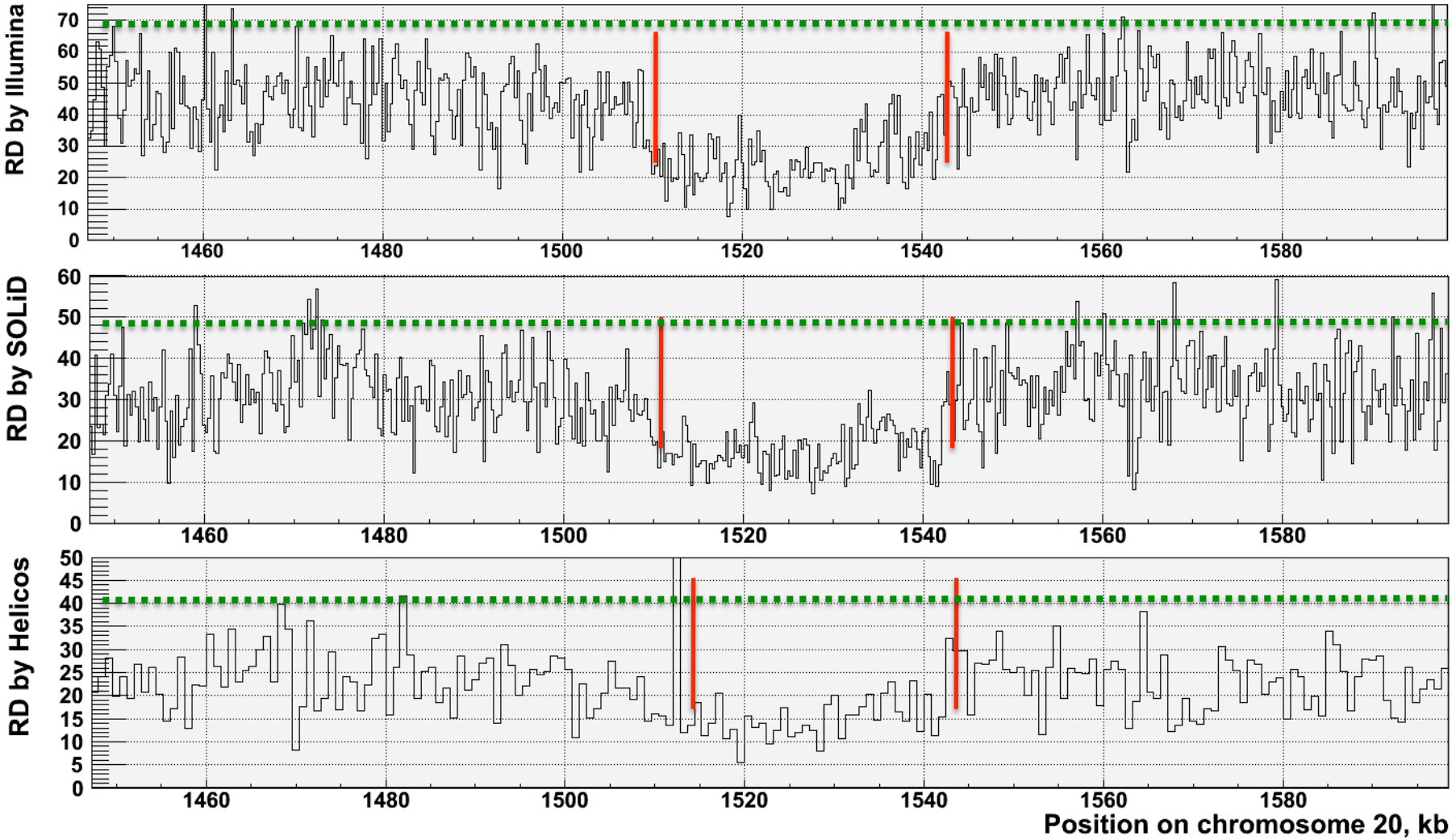
L1, SVA, Alus

VNTR
(Variable Number Tandem Repeats)

Number of repeats varies between different people



Read-depth works well on a variety of sequencing platforms but provides imprecise breakpoints



[Abyzov et al. Gen. Res. ('11)]

[NA18505]

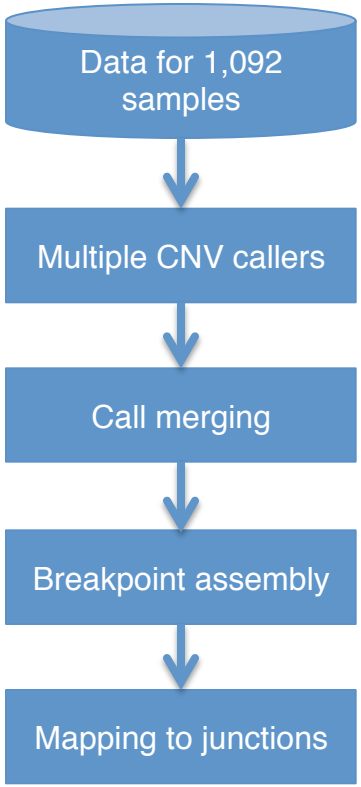
1000G SV (Pilot, **Phase I & III**)

- **Many** different callers compared & used
  - including SRiC & CNVnator but also VariationHunter, Cortex, NovelSeq, PEMer, BreakDancer, Mosaik, Pindel, GenomeSTRiP, mrFast….
- Merging
- Genotyping (GenomeSTRiP)
- Breakpoint assembly (AGE & Tigra_SV)
- Mechanism Classification

[1000 Genomes Consortium, Nature (2010, 2012); Mills et al., Nature (2011)]

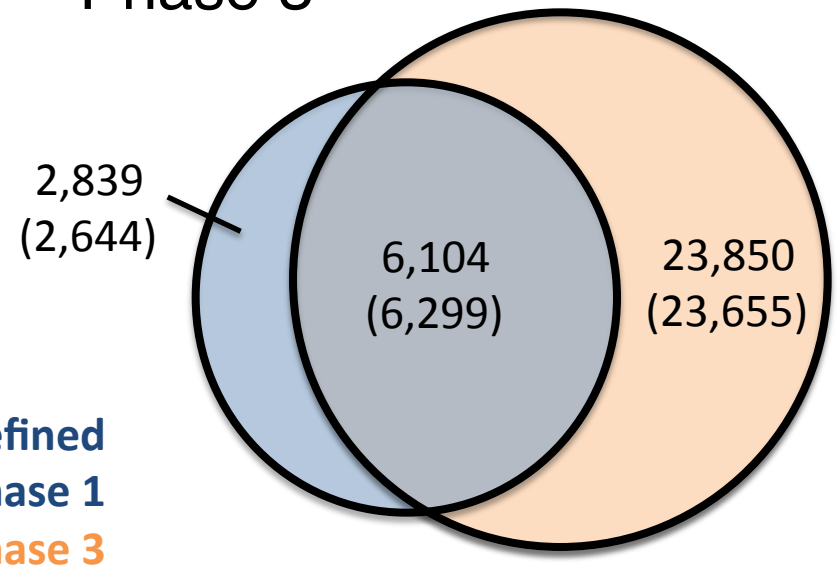
8,943 Deletion Breakpoints (Phase I Refined)

- 42K deletions in official Phase 1 release
  - ~20% w/ breakpt
- Breakpoint FDR from IRS, PCR, and high-coverage trios
  - ~7% for site existence
  - 13% for site existence + sequence precision



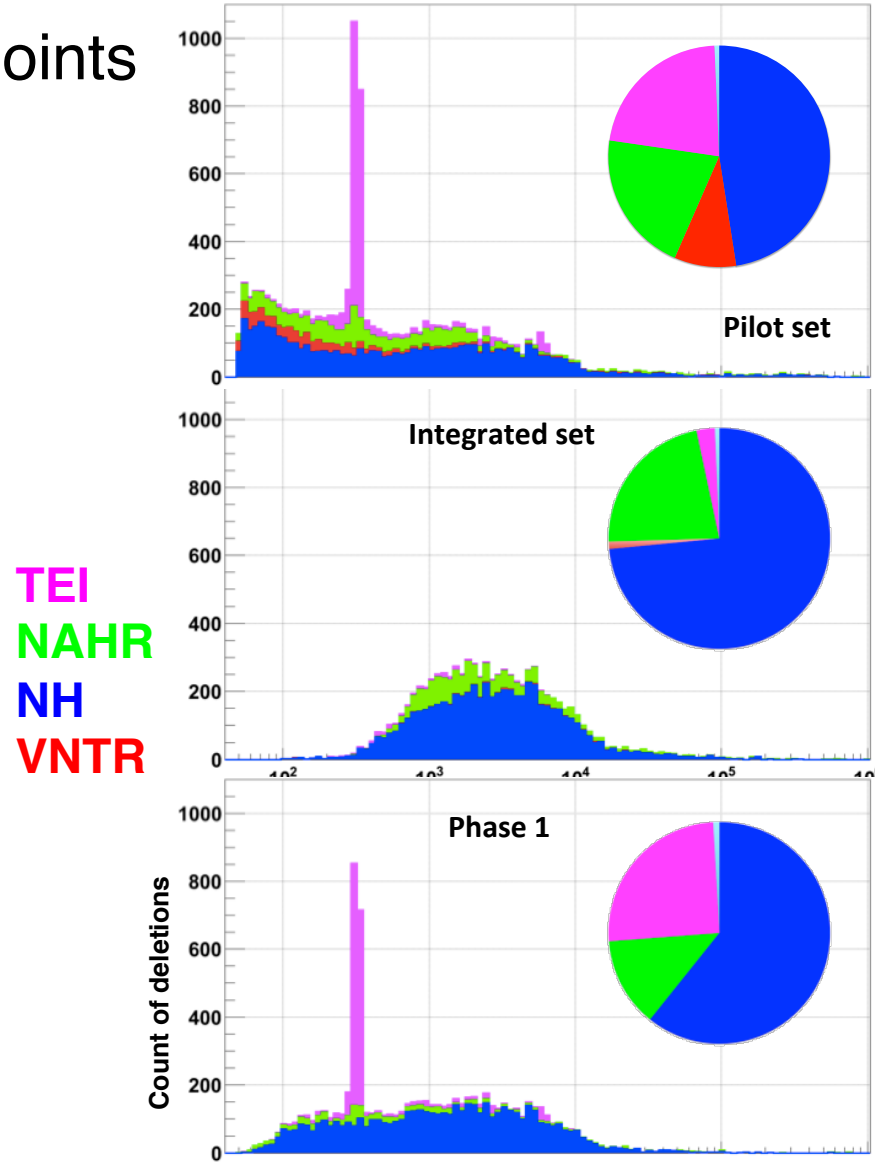
Breakpoint characterization in 1000G

- Breakseq #1 w/ ~2000 breakpoints [Lam et al. Nat. Biotech. (‘10)]
- Pilot
- Phase 1 “Integrated” & Phase 1 refined
- Phase 3



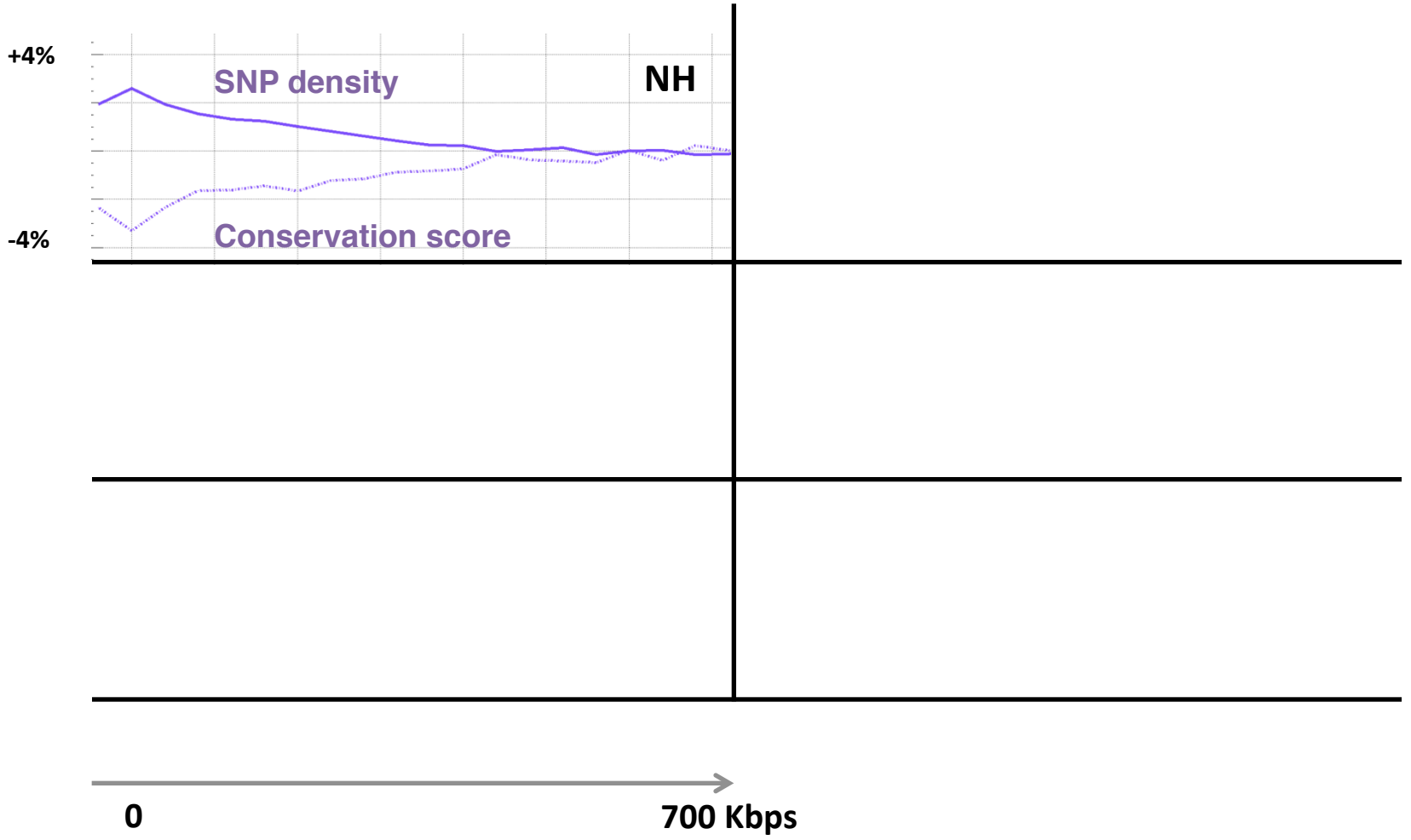
Refined
Phase 1
Phase 3

Exact match
Number in parentheses: >50% reciprocal match



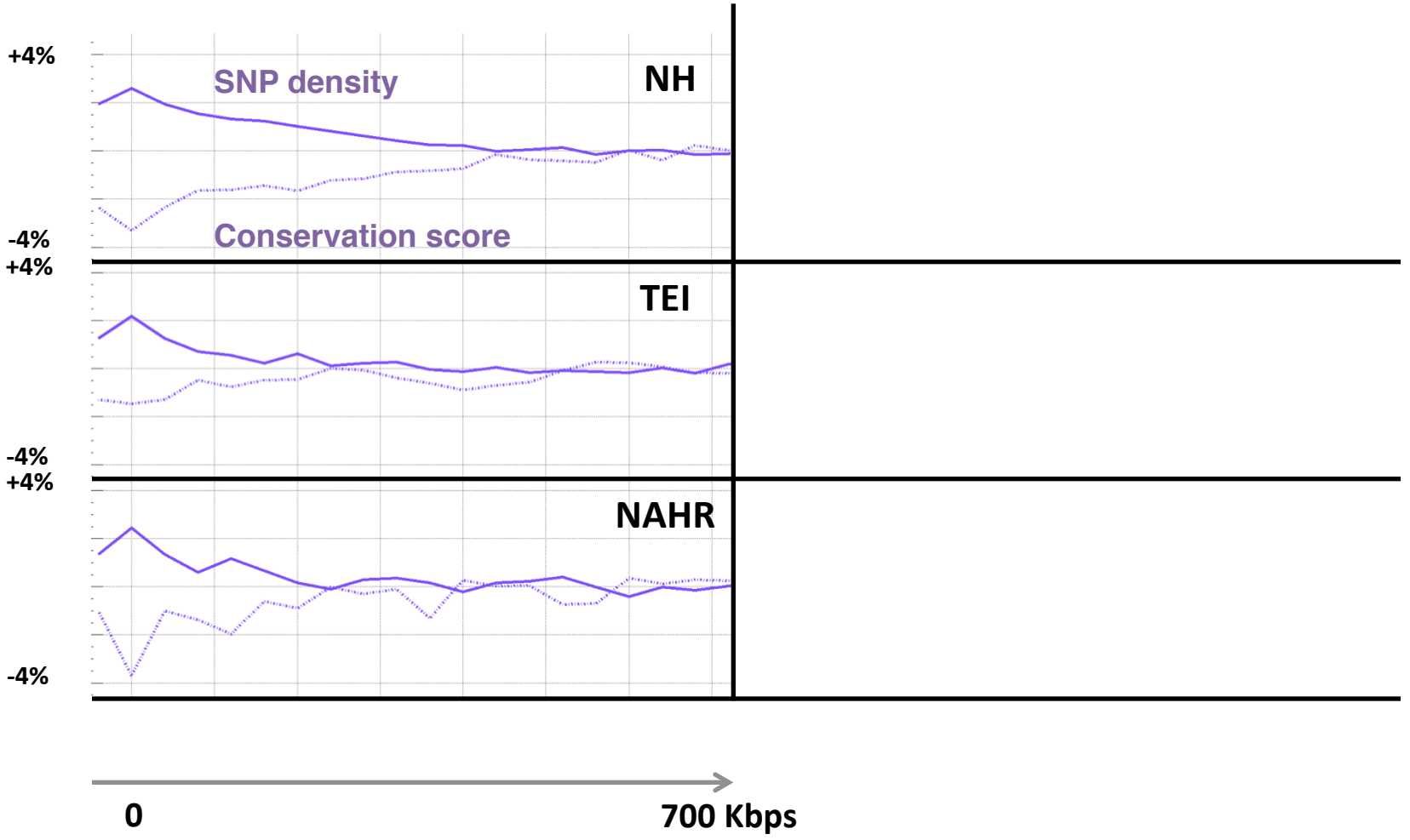
[Abyzov et al. (‘15) *Nature Comm.*]

Higher SNP Density and Relaxed Selection at NH Breakpoints



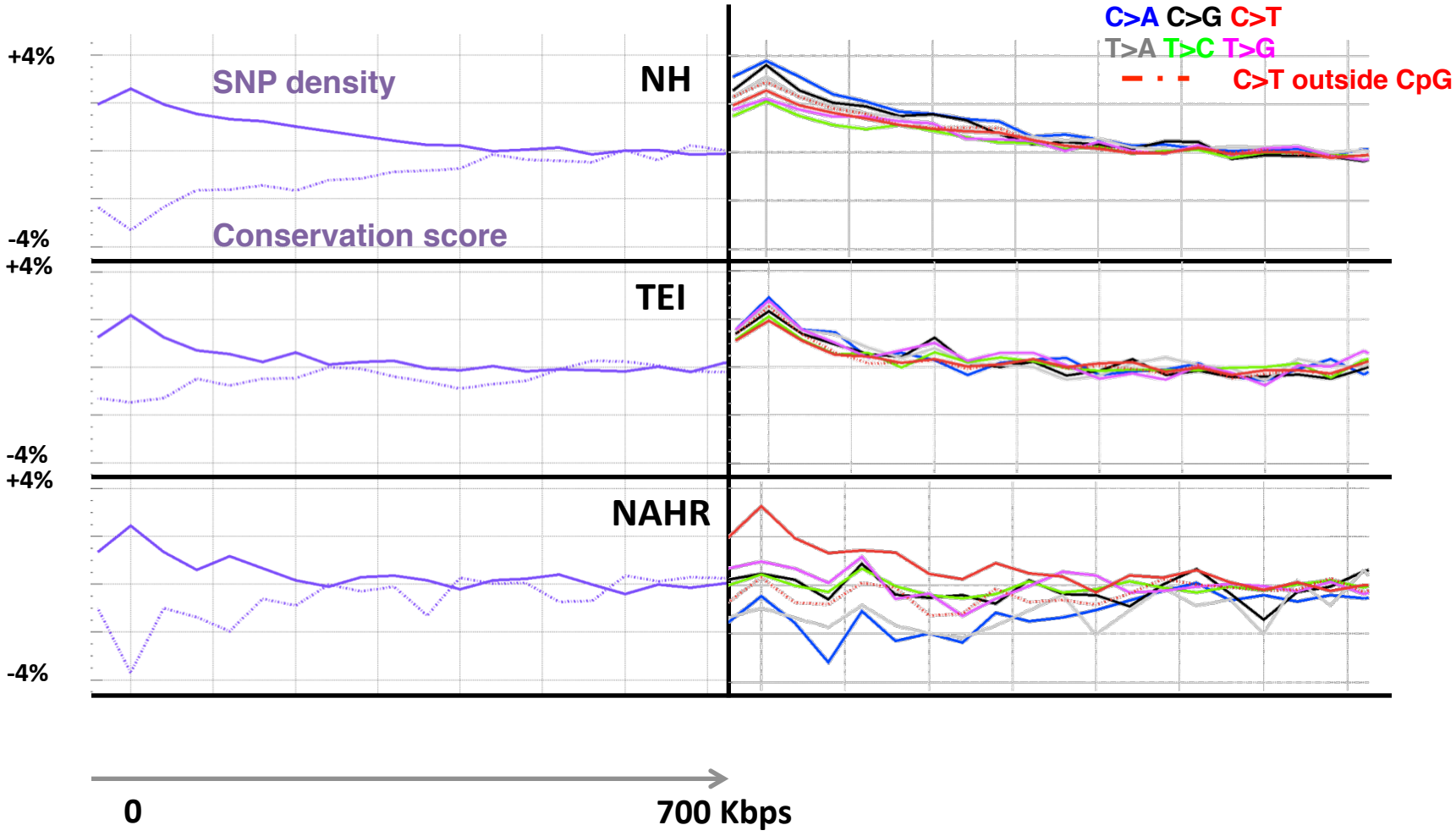
[Abyzov et al. (‘15) *Nature Comm.*]

Higher SNP Density and Relaxed Selection at all Breakpoints



[Abyzov et al. ('15) *Nature Comm.*]

SNP Density at NAHR is Driven by High C>T

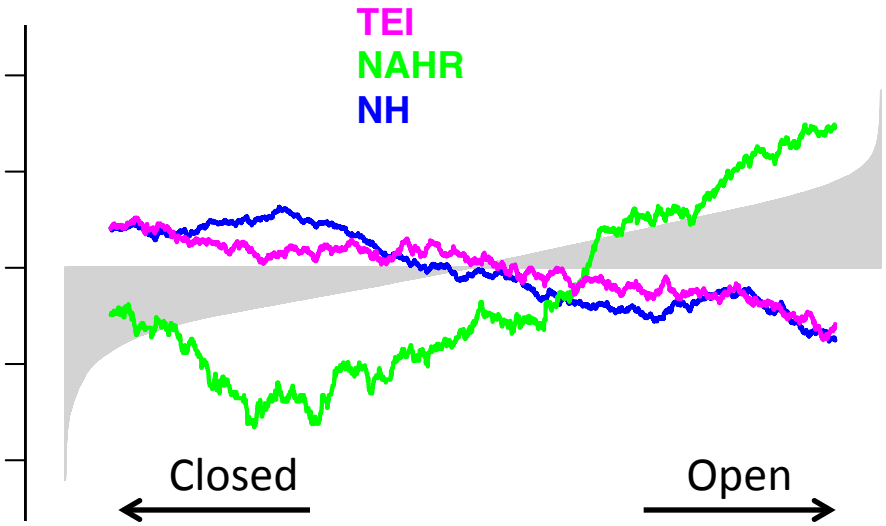


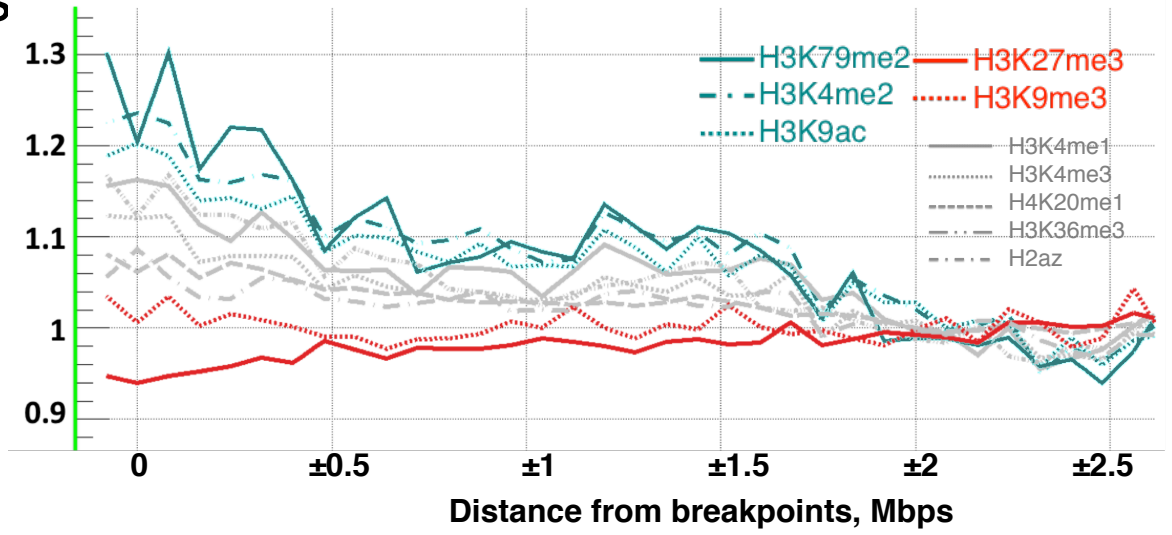
[Abyzov et al. ('15) *Nature Comm.*]

NAHR breakpoint are associated with open chromatin

- Supported by Hi-C and Histone modification
- Hypothesis: Some NAHR deletions occur w/o cell Replication

* H1 & GM12878 cells

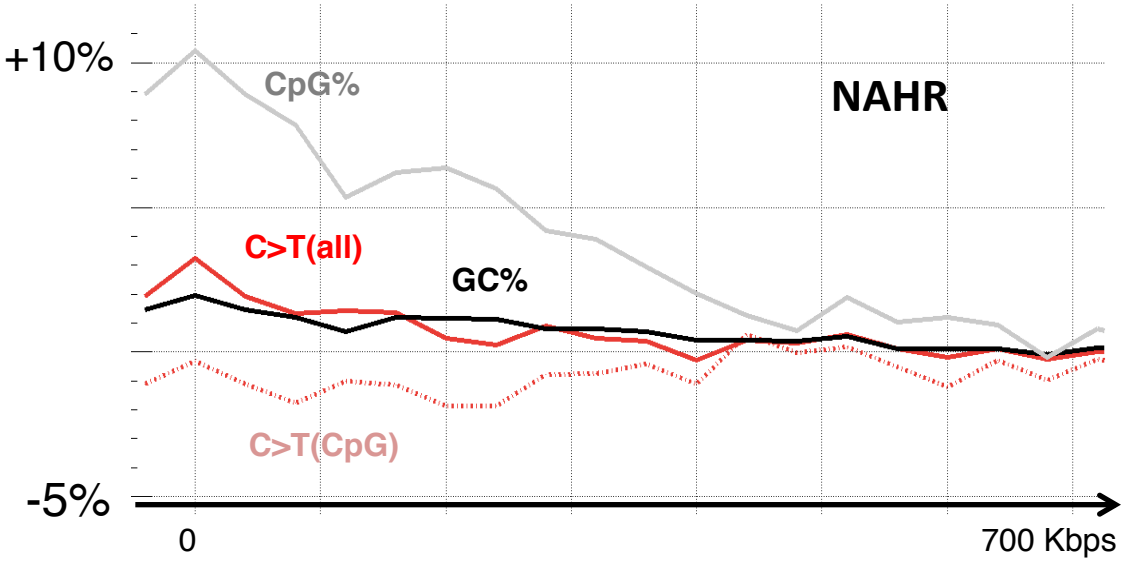


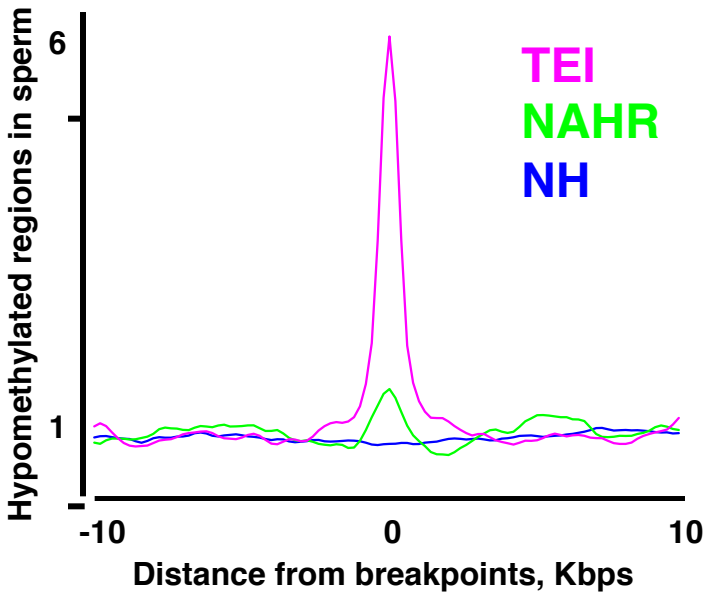


[Abyzov et al. ('15) *Nature Comm.*]

Methylation pattern associated with breakpoints mechanisms

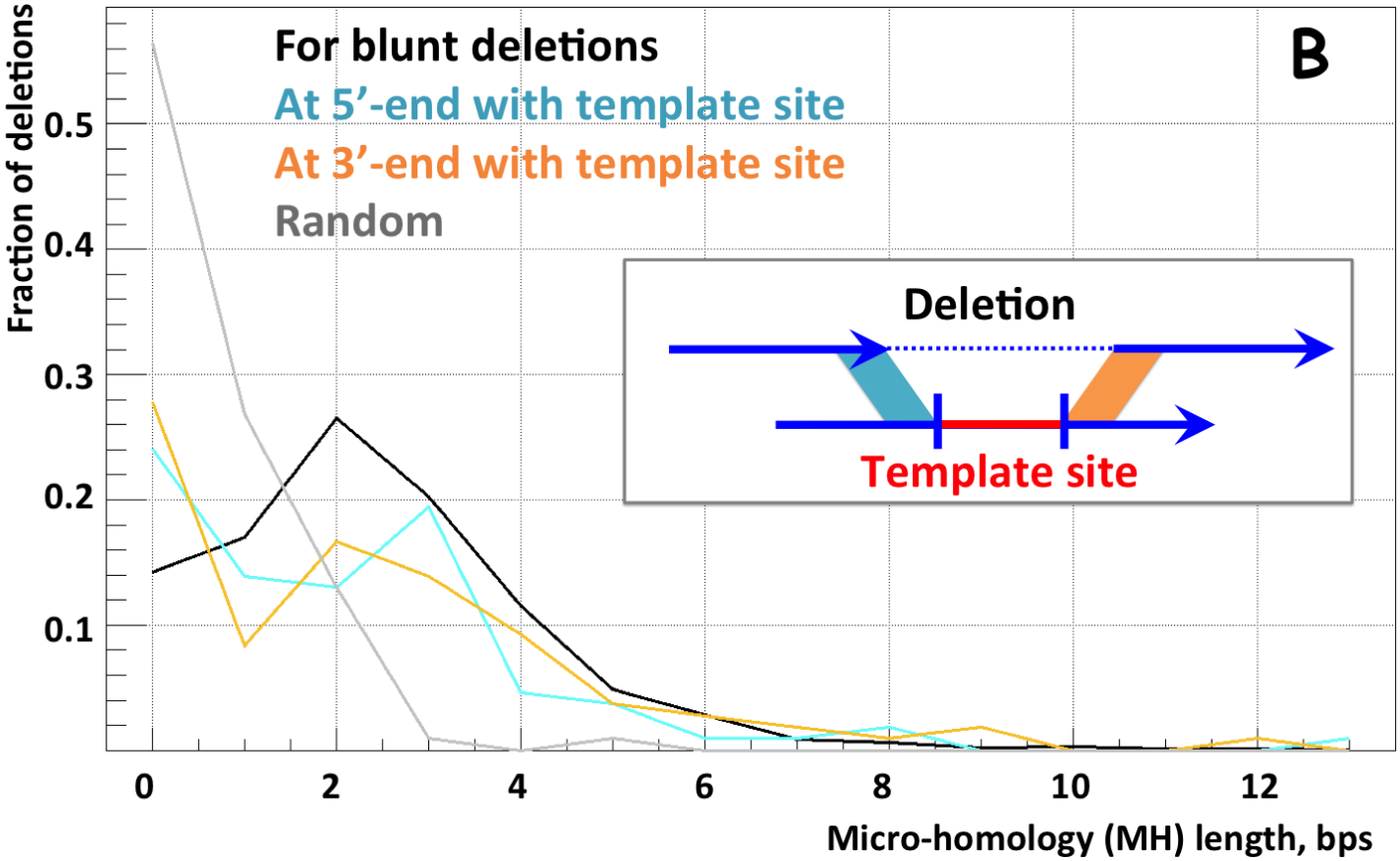
- Lower C>T in CpG around NAHR breakpoints
  - indicates lower methylation level in germline & embryonic cells
- Confirmed in male gamete





[Abyzov et al. ('15) *Nature Comm.*]

Micro-homologies Identified around NH Breakpoints

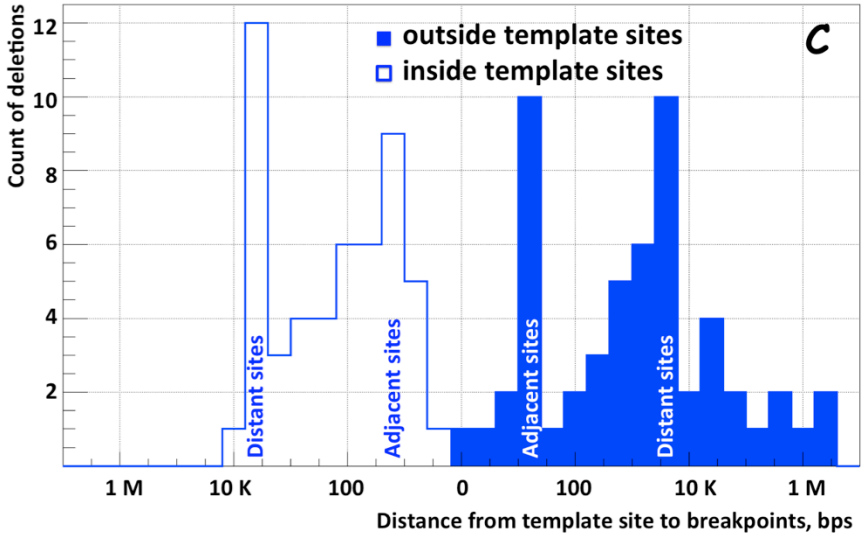


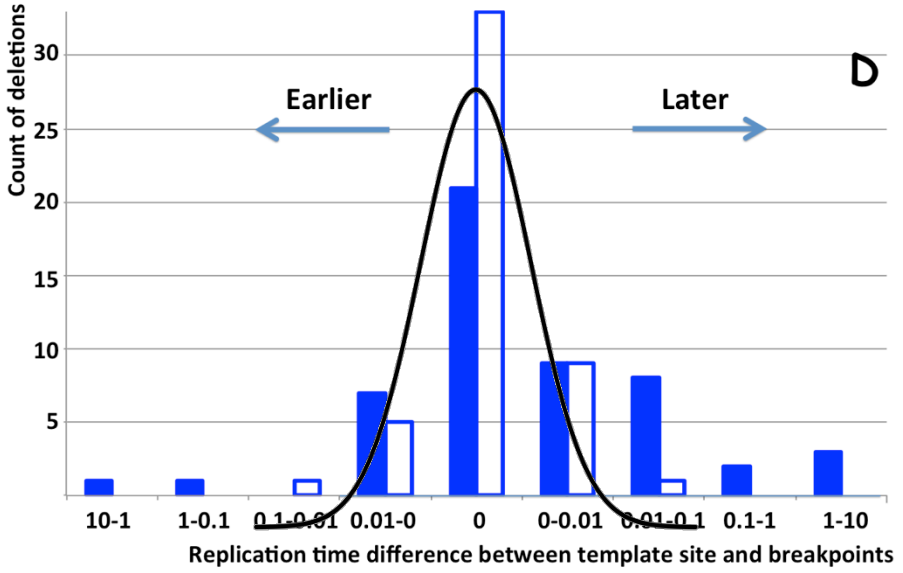
- Breakpoints have Micro-homologous sequences with the template sites.

[Abyzov et al. ('15) *Nature Comm.*]

NH deletions are often coupled with micro-insertions

- Templates located at 2 characteristic distances from breakpoints, which tend to replicate late
- Suggests spatial & temporal configuration of DNA during template switching





[Abyzov et al. ('15) *Nature Comm.*]

Human Genome Analysis – SVs & Pseudogenes, Tricky but Crucial Genomic Features, Targeted by Long-read Sequencing: Current Short-read Results & Future Prospects

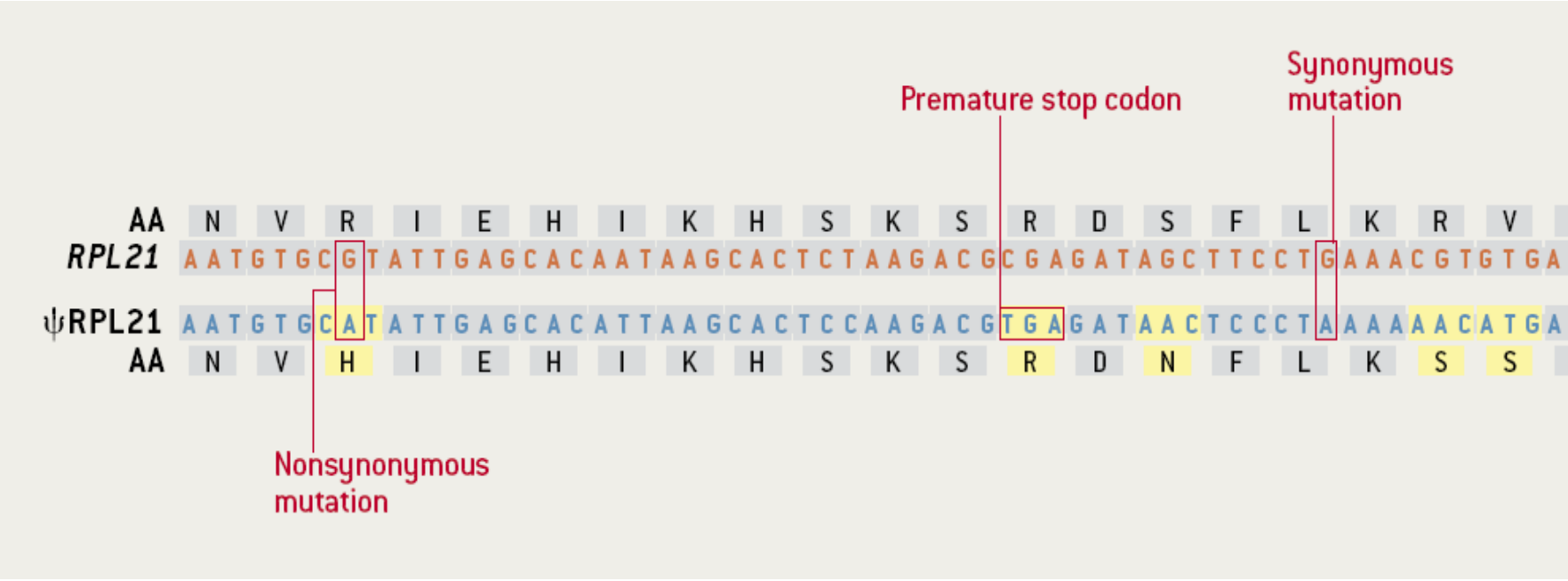
- **SV Breakpoints**
  - ~9K deletions with breakpoints & mechanism classification from 1000G
  - Small subset of tot. deletions, which could be greatly expanded by long reads
  - More nearby SNPs than genomic average.
  - From methylation, Hi-C , & his mods, NAHR breakpoints associated with open chromatin (perhaps occurring w/o replication & division)
  - NHR breakpoints associated w/ sequence microinsertions, templated from later replicating sites, spaced at 2 characteristic distances
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
- **Intersection of Pseudogenes & SVs**
  - Enrichment of SVs in pseudogenes v genes, particularly for NAHR
- **Novel Processed Pseudogenes as a Form of SV**
  - Not in reference but in human population – could be improved by long reads
  - Now found w/ splice junction mapping + clustering of unmapped PEs
  - ~8 per person, often pop. specific
  - Associated w/ G1/M expressed genes
- **Many Pseudogenes with Low Levels of Biochemical Activity**
  - Conservative assignment, mis-map issue, could be improved by long reads
  - ~15% transcribed & 80% w/ some activity

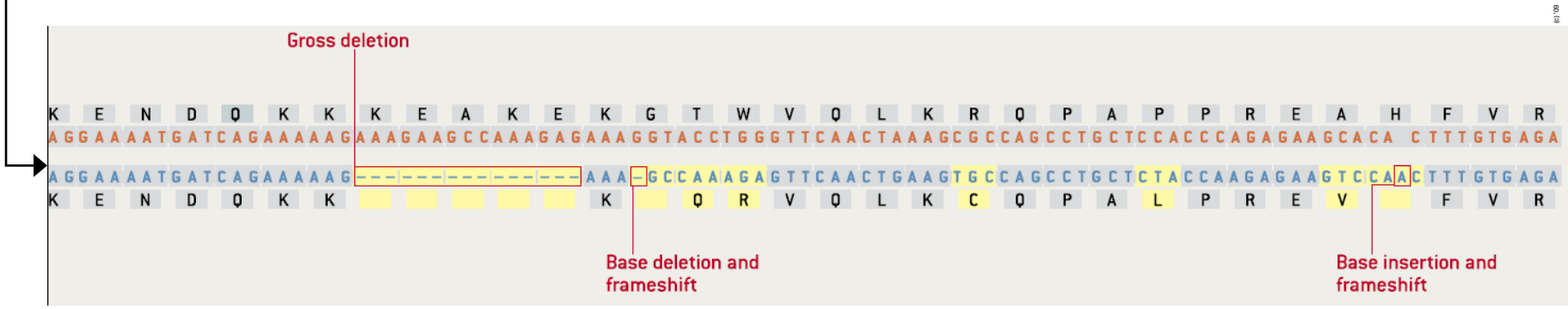
Pseudogenes are among the most interesting intergenic elements

- Formal Properties of Pseudogenes (ΨG)
  - Inheritable
  - Homologous to a functioning element – ergo a repeat!
  - 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]

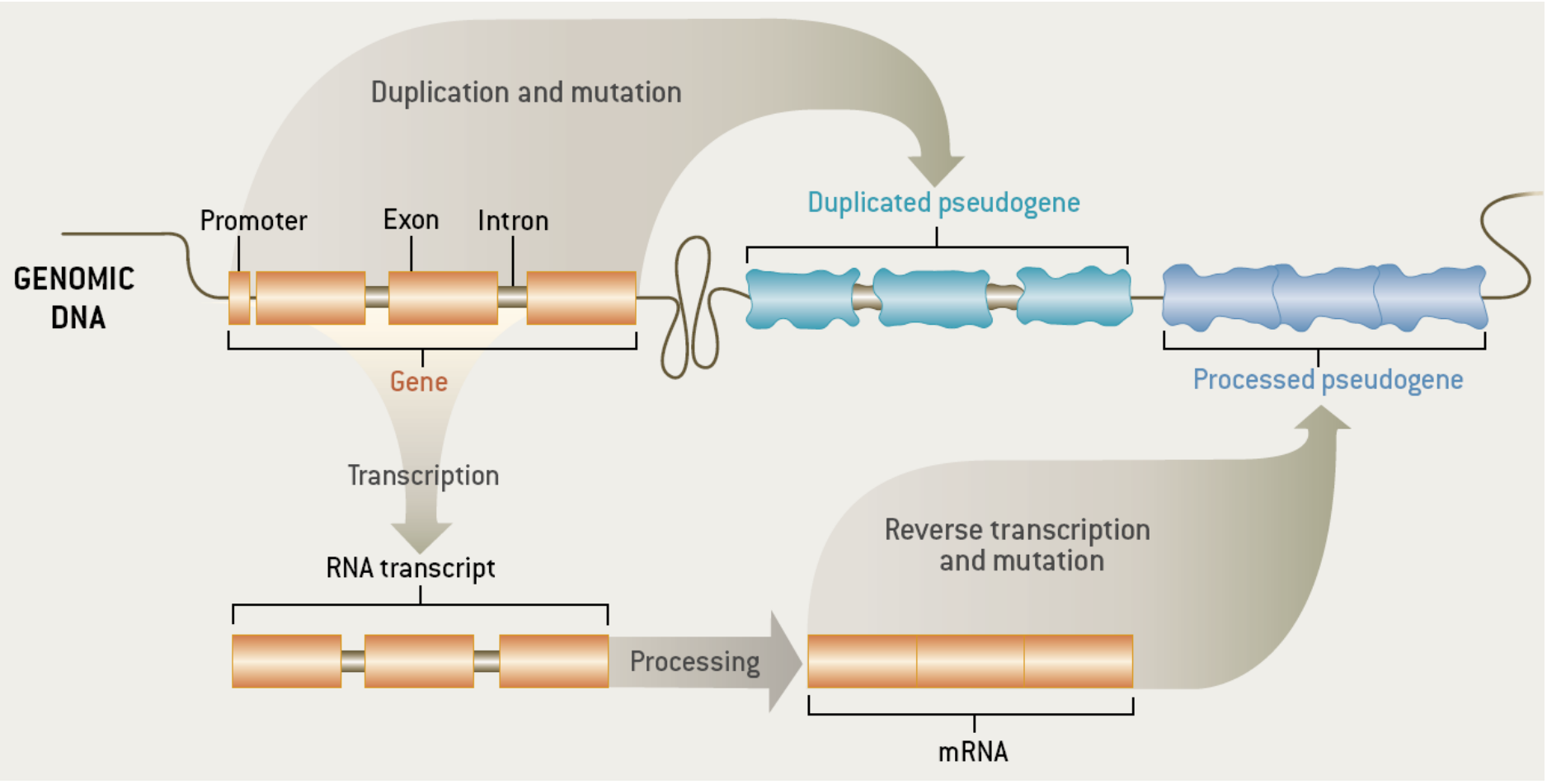
Identifiable Features of a Pseudogene (ψRPL21)





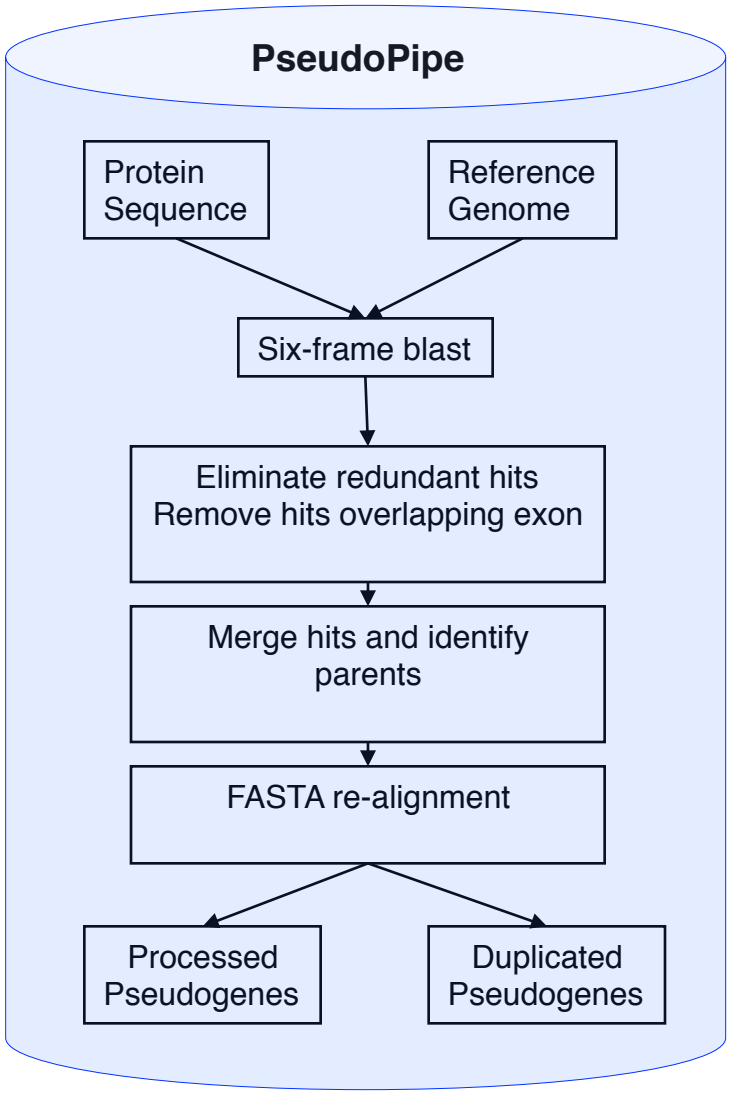
[Gerstein & Zheng. Sci Am 295: 48 (2006).]

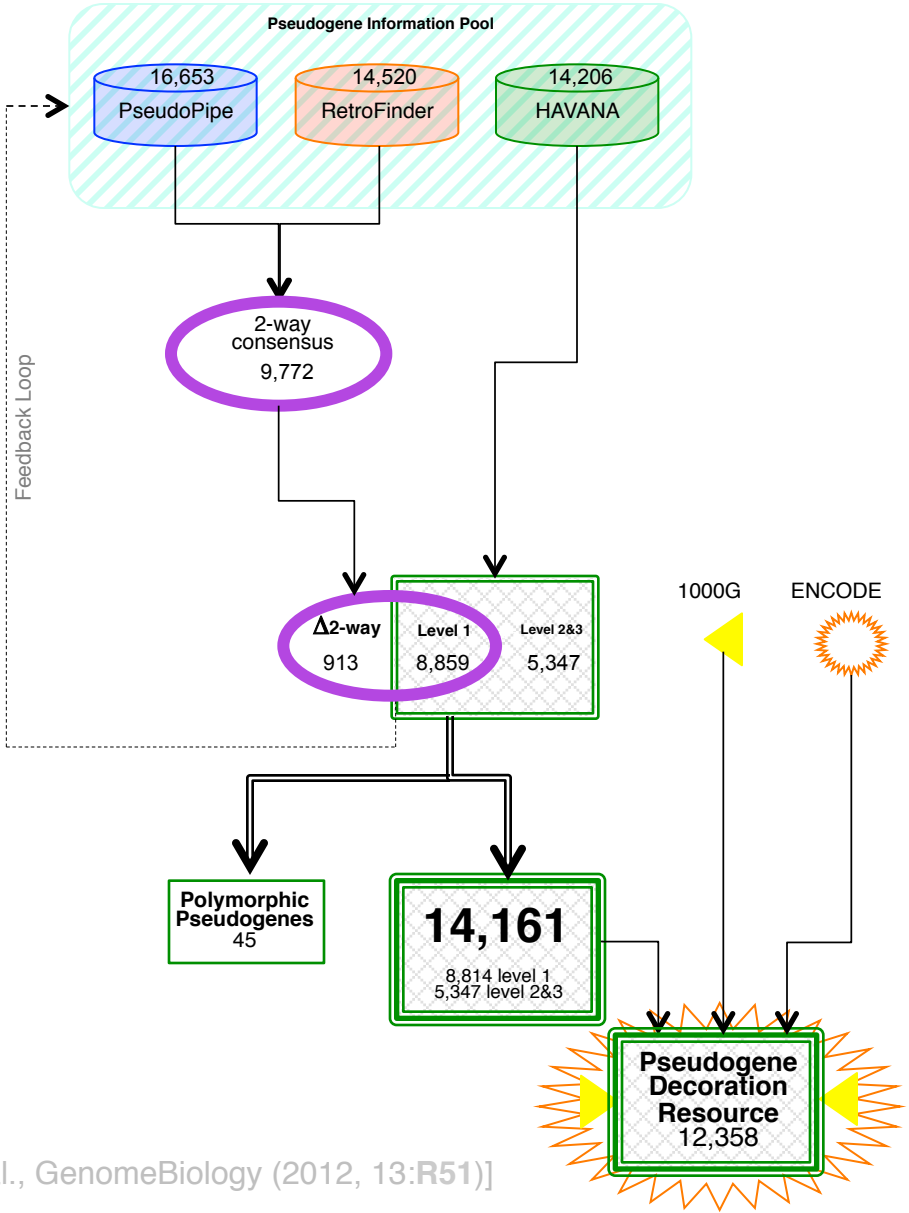
Two Major Genomic Remodeling Processes Give Rise to Distinct Types of Pseudogenes



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

Genome-wide Annotation of Pseudogenes



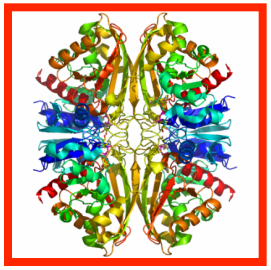


[Pei et al., GenomeBiology (2012, 13:**R51**)]

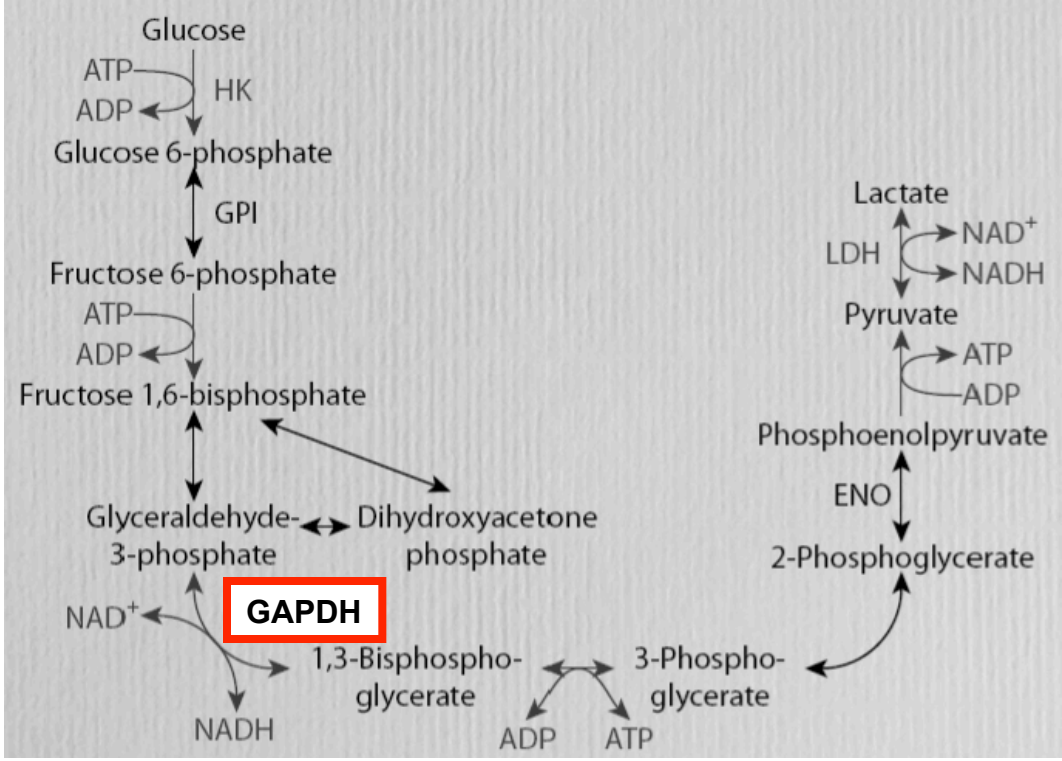
EX: Number of pseudogenes for each glycolytic enzyme

[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

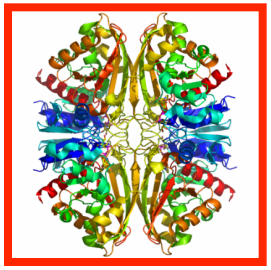


	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60/2	47/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0	-	-	-	-	-
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**

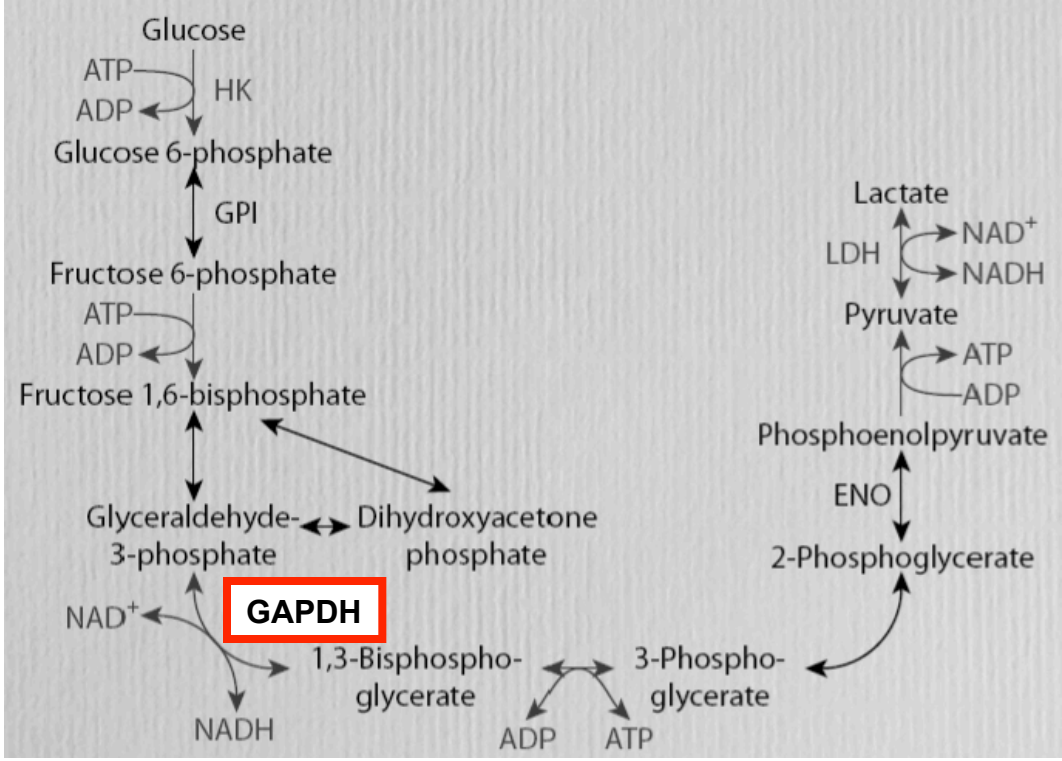
EX: Number of pseudogenes for each glycolytic enzyme

[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.



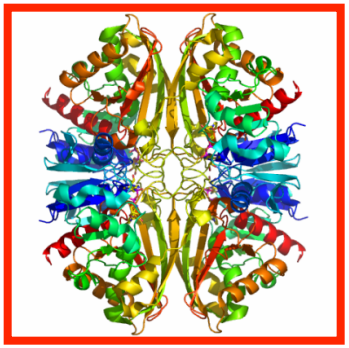
Processed/Duplicated



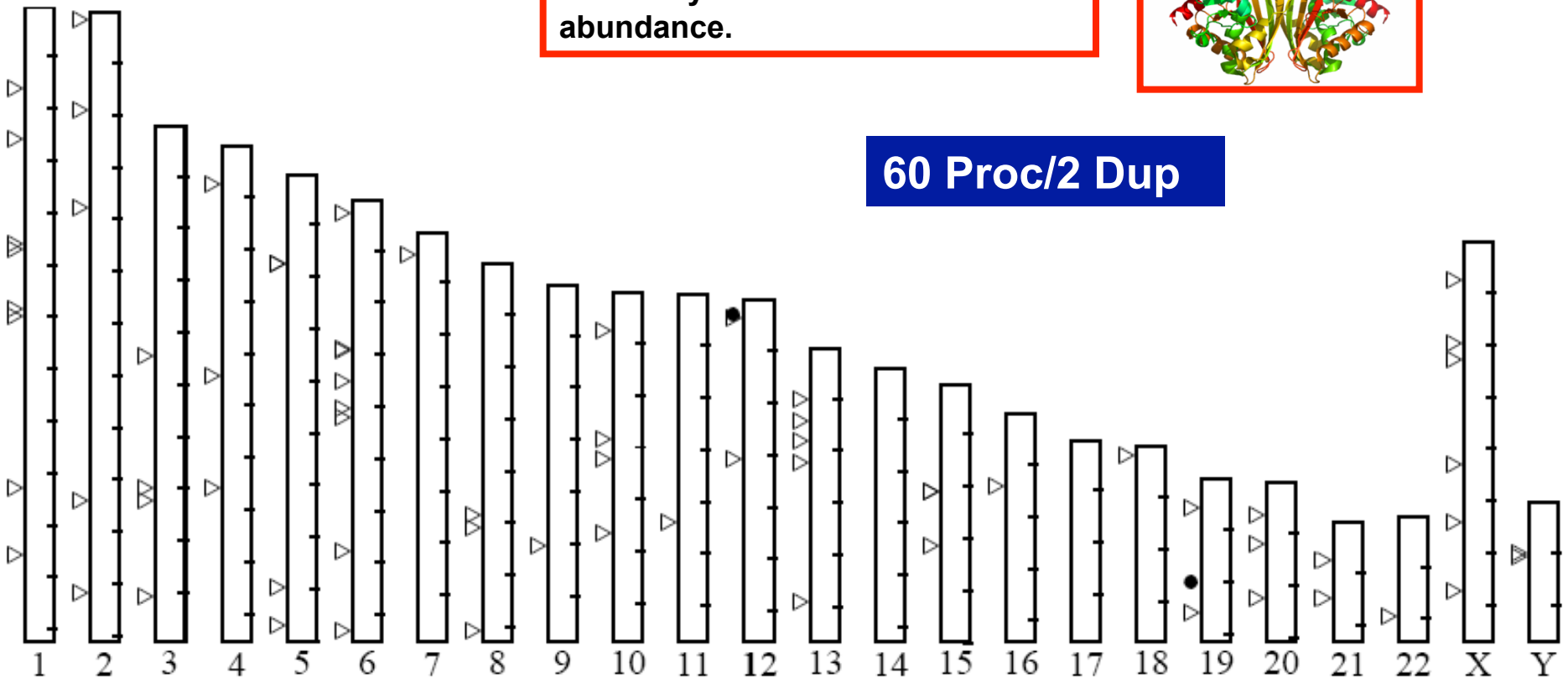
	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60 Proc/2 Dup	7/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	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



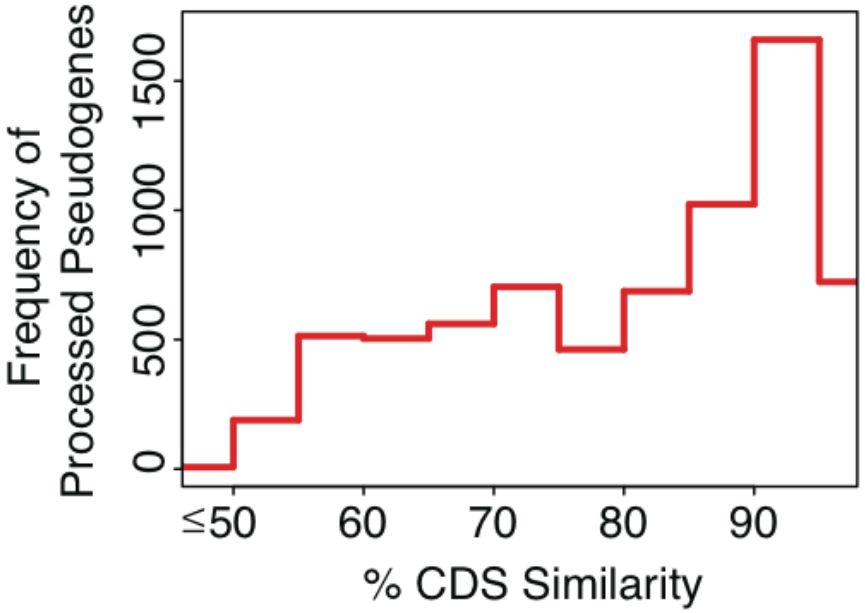
[Liu et al. BMC Genomics ('09, in press)]

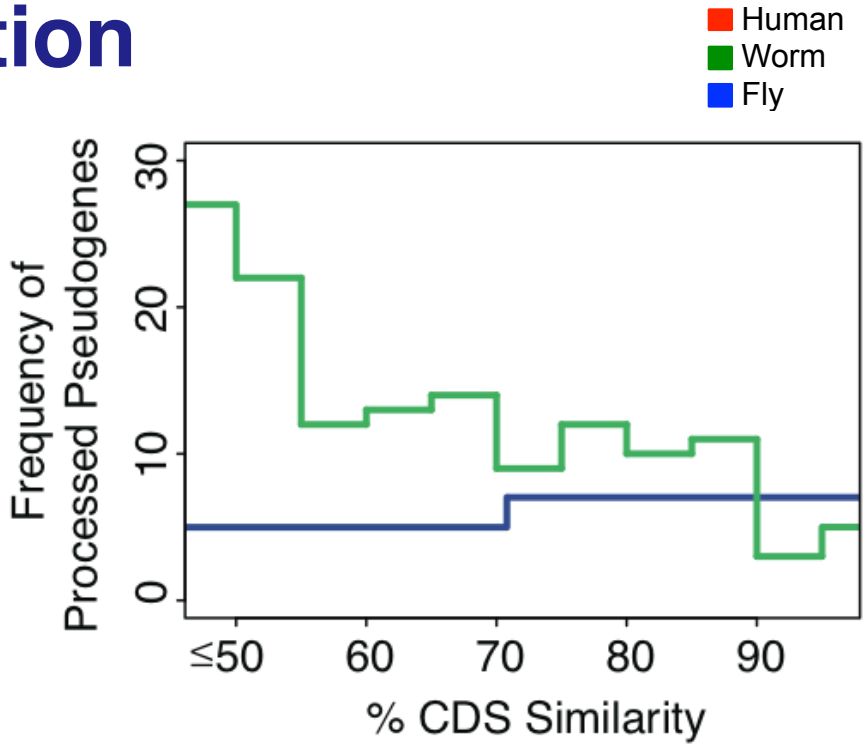
Annotation of Human Pseudogenes in Comparison to those in other Model Organisms

Organism	Total Pseudogenes	Biotype Distrbution (Processed / Duplicated)	ENCODE Functional Genomics Data	Completed Manual Annotation
Human	12,358	8908 / 2266	✔	✔
Worm	911	159 / 566	✔	✔
Fly	145	16 / 109	✔	✔
Zebrafish	229	21 / 177	✔	✔
Macaque	11,136	6570 / 1725	✗	✗
Mouse	13,169	7811 / 1827	✔	✗

[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Evolution

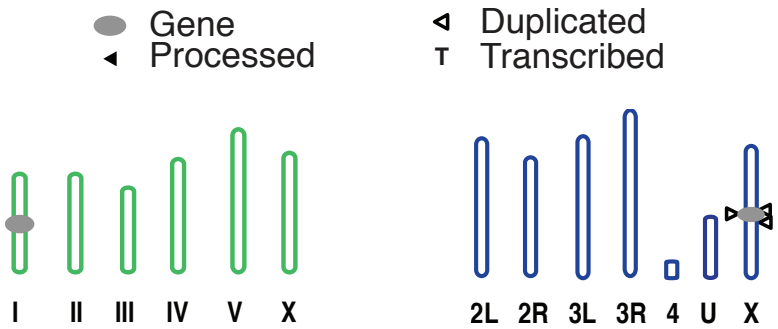


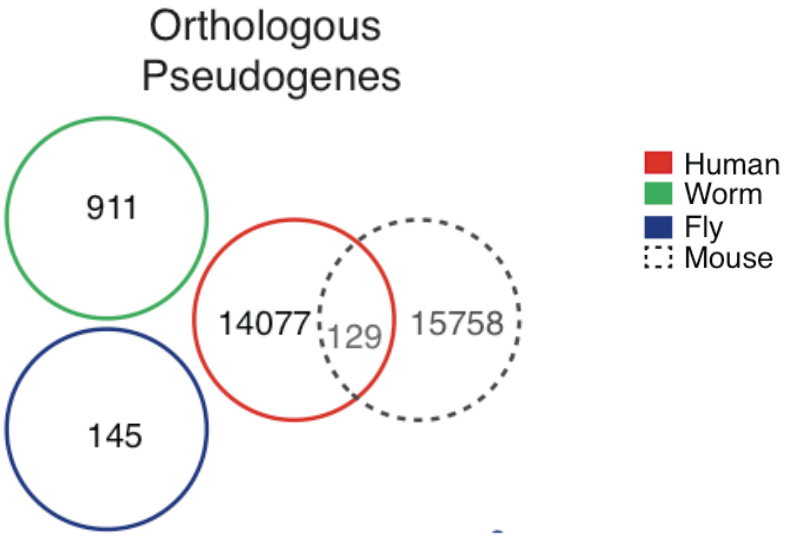


Organism	Defect / Pseudogene x MB		
	Insertion	Deletion	Stop
Human	4.4	4.9	2.4
Worm	**25.8**	7.45	2.5
Fly	7.9	**12.7**	1.1

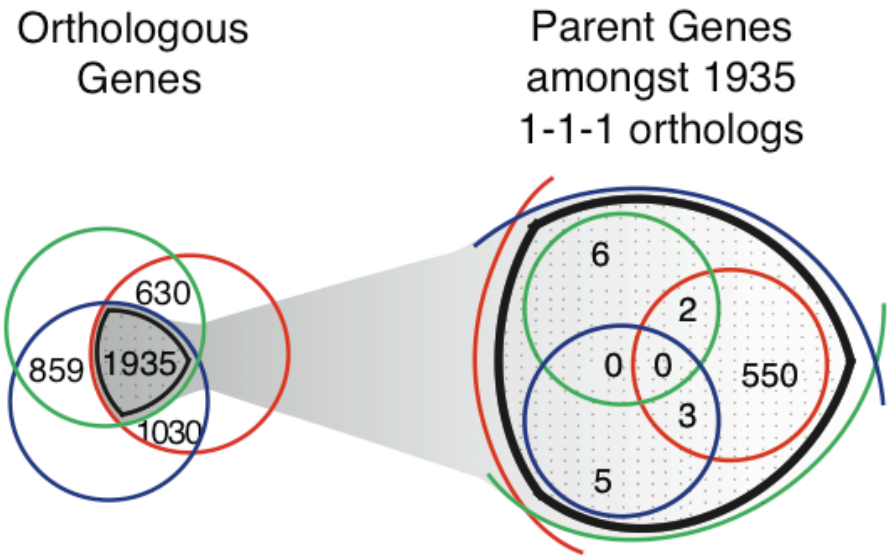
[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Case Study: Ribosomal Protein RpS6



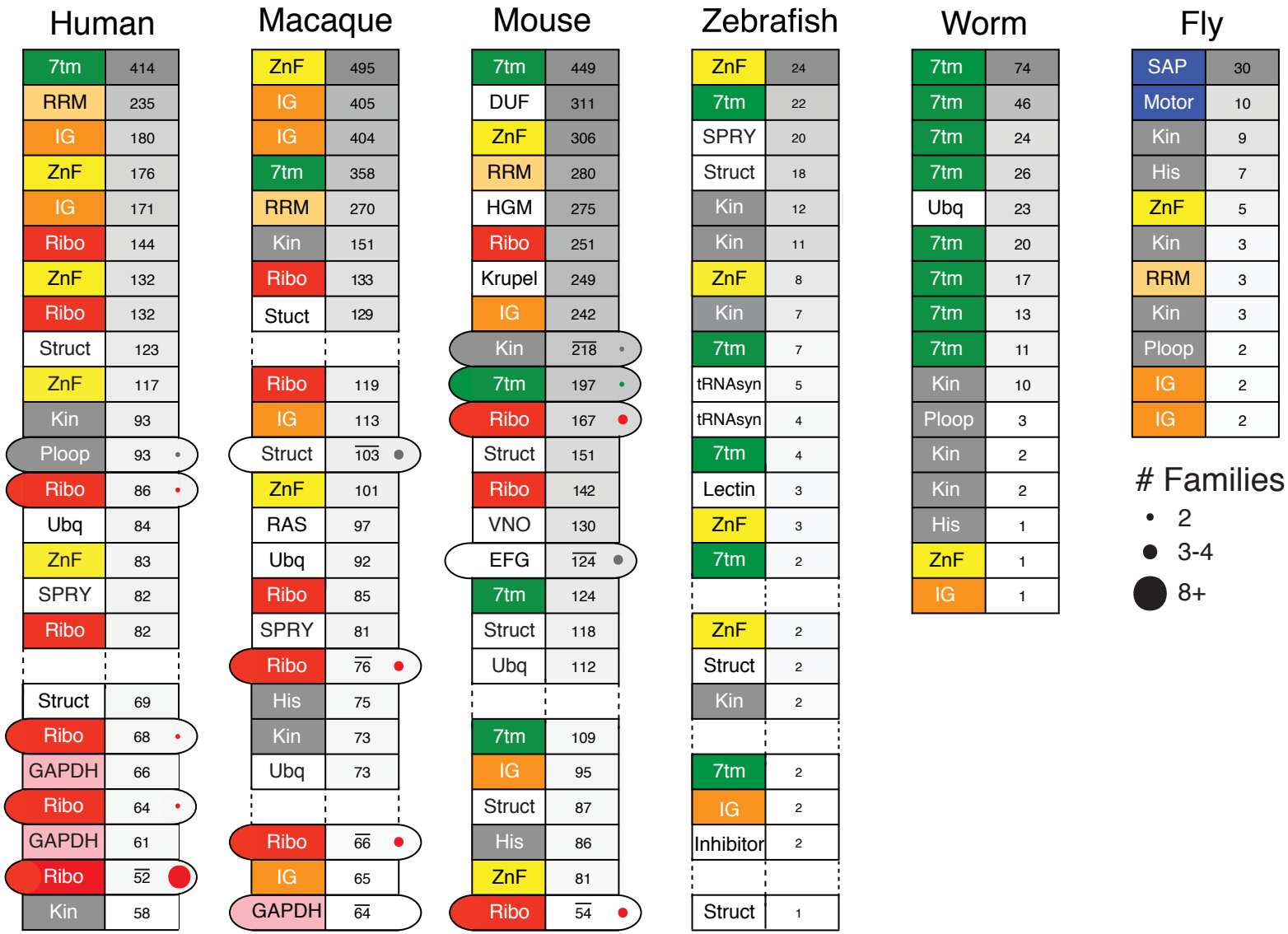


Great divergence in pseudogenes in terms of Orthologs & Paralogs



[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Divergence but
More interpretable Patterns in terms of Families

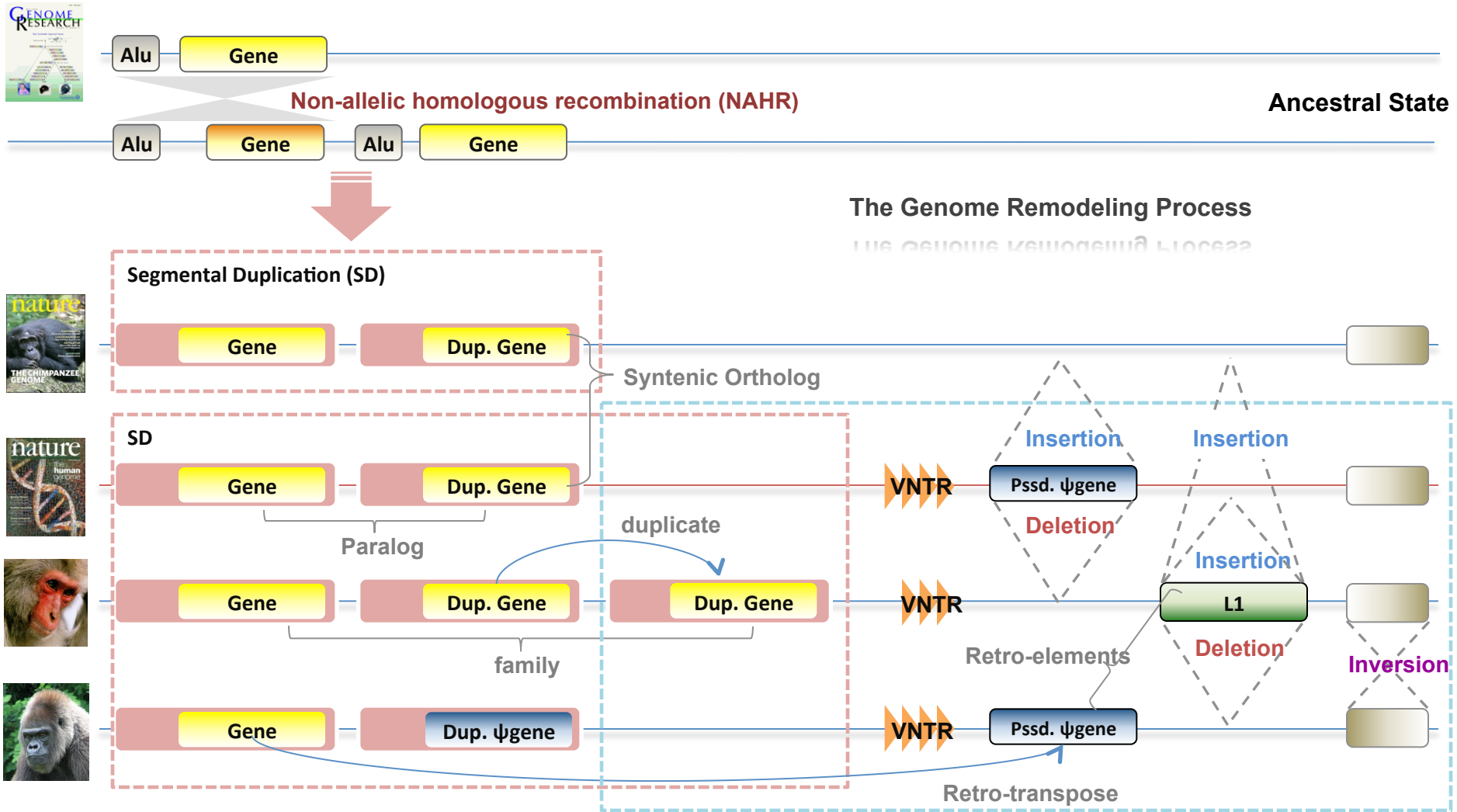


[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Human Genome Analysis – SVs & Pseudogenes, Tricky but Crucial Genomic Features, Targeted by Long-read Sequencing: Current Short-read Results & Future Prospects

- **SV Breakpoints**
  - ~9K deletions with breakpoints & mechanism classification from 1000G
  - Small subset of tot. deletions, which could be greatly expanded by long reads
  - More nearby SNPs than genomic average.
  - From methylation, Hi-C , & his mods, NAHR breakpoints associated with open chromatin (perhaps occurring w/o replication & division)
  - NHR breakpoints associated w/ sequence microinsertions, templated from later replicating sites, spaced at 2 characteristic distances
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
- **Intersection of Pseudogenes & SVs**
  - Enrichment of SVs in pseudogenes v genes, particularly for NAHR
- **Novel Processed Pseudogenes as a Form of SV**
  - Not in reference but in human population – could be improved by long reads
  - Now found w/ splice junction mapping + clustering of unmapped PEs
  - ~8 per person, often pop. specific
  - Associated w/ G1/M expressed genes
- **Many Pseudogenes with Low Levels of Biochemical Activity**
  - Conservative assignment, mis-map issue, could be improved by long reads
  - ~15% transcribed & 80% w/ some activity

Fixed Genomic Variation



Polymorphic Genomic Variation

Alu Gene

Non-allelic homologous recombination (NAHR)

Ancestral State

Alu Gene Alu Gene

The Genome Remodeling Process

Segmental Duplication (SD)

Gene Dup. Gene

Syntenic Ortholog

CNV (type of SV)

Gene Dup. Gene

duplicate

Gene Dup. Gene Dup. Gene

Gene Dup. ψgene

VNTR Pssd. ψgene

Insertion

Deletion

Insertion

VNTR L1

Insertion

Deletion

Retro-elements

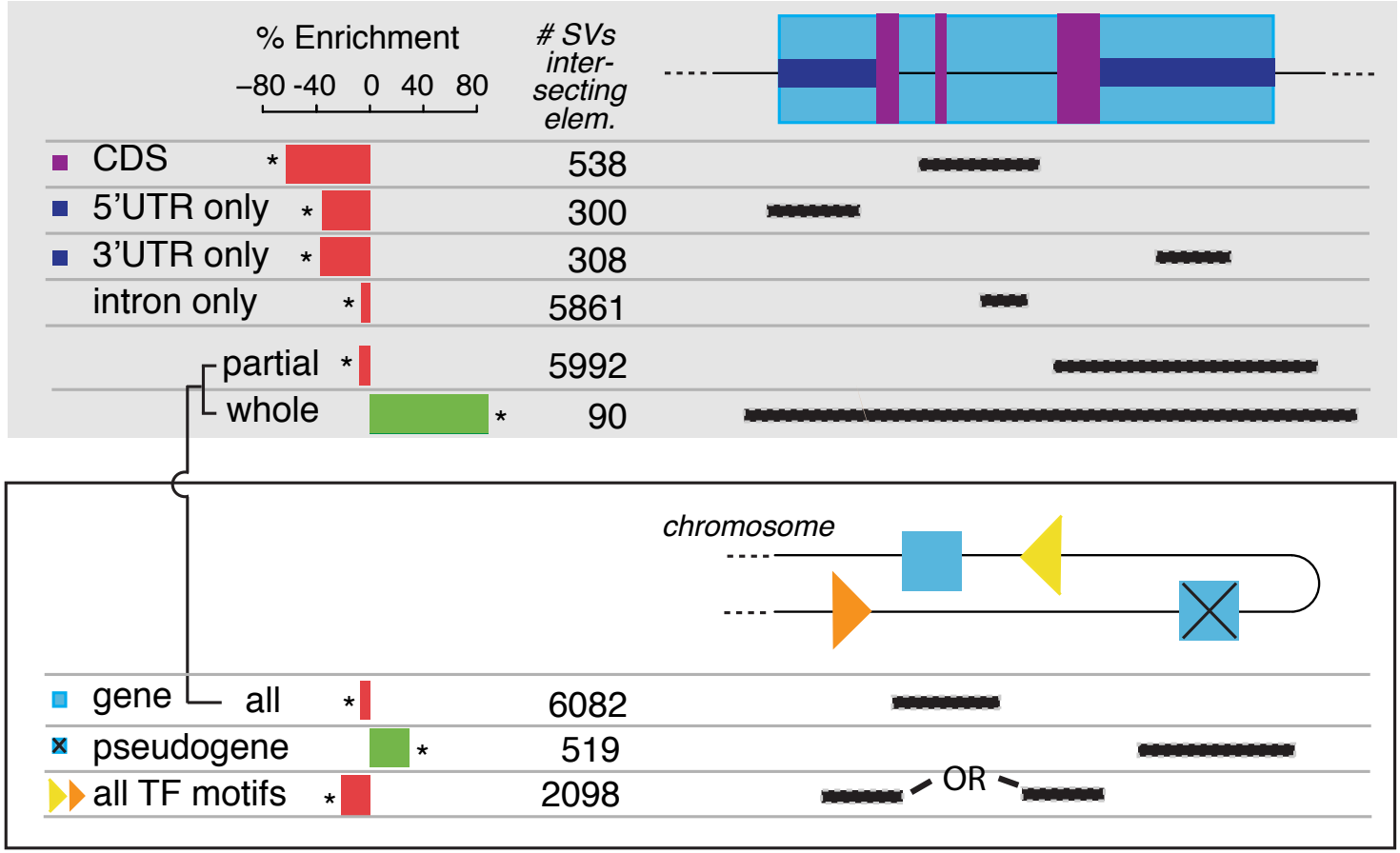
Inversion

VNTR Pssd. ψgene

Retro-transpose

"Polymorphic" Genes & Pseudogenes

In comparison to other genomic elements, pseudogenes tend to overlap SVs



[Khurana et al., *Science* ('13)]

More detail on pseudogene overlap with SVs (enrichment wrt randomized control)

Element	All SVs		NAHR		VNTR		NHR		TEI	
	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value	Enrichment	*P*-value
Gene	**0.90**	**8.68E-20**	**1.13**	**4.98E-08**	**0.84**	**6.50E-06**	**0.83**	**8.28E-27**	**0.87**	**6.96E-09**
CDS	**0.37**	**8.72E-85**	**0.68**	**1.94E-06**	**0.07**	**3.40E-11**	**0.37**	**5.82E-53**	**0.04**	**3.47E-24**
Pseudogene	**1.24**	**1.11E-05**	**1.56**	**3.37E-07**	**1.54**	**1.73E-02**	**1.24**	**6.94E-04**	**0.50**	**3.58E-03**
Whole Pseudogene	**1.51**	**1.15E-12**	**1.95**	**3.98E-13**	**2.50**	**1.22E-04**	**1.33**	**1.44E-04**	0.51	1.63E-01
Partial Pseudogene	0.93	2.39E-01	0.97	4.40E-01	1.05	4.37E-01	1.10	2.16E-01	**0.50**	**6.26E-03**
Duplicated Pseudogene	1.14	9.94E-02								
Processed Pseudogene	1.46	1.14E-08								

- SVs are shuffled in the whole genome.
- Significant *P*-values (<0.05) in black and bold
- Significant enrichments in green
- Significant depletions in red

[Mu et al., NAR (2011)]

Pseudogenes & CNVs

- CNVs are the raw form of variation producing duplicated elements (SDs)
  - SDs give rise to duplicated genes ➡ protein families
  - **SDs comprise ~5% of the human genome but contain ~18% genes, 46% duplicated and 22% processed pseudogenes**

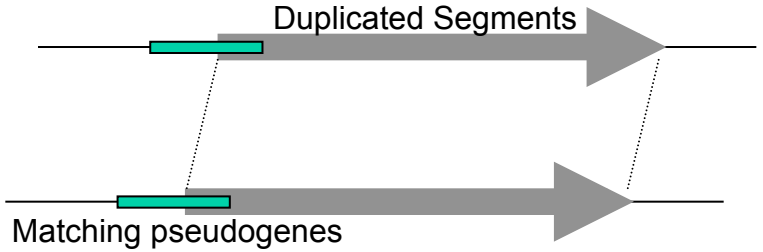
[Lam et al., NAR DB Issue ('09)]

Duplicated pseudogenes

- CNVs & SDs tend to be **enriched in environmental response genes**, matching patterns found for duplicated pseudogenes [Korbel et al., COSB ('08)]
- **Duplicated** pseudogenes are associated in general with **older SDs** [Kim et al. Gen. Res. ('08)]

Processed pseudogenes

- Matching **processed** pseudogenes (sharing the parent gene) are **enriched** at **SD junctions**
- Processed pseudogenes can serve as **repeats for mediating NAHR**

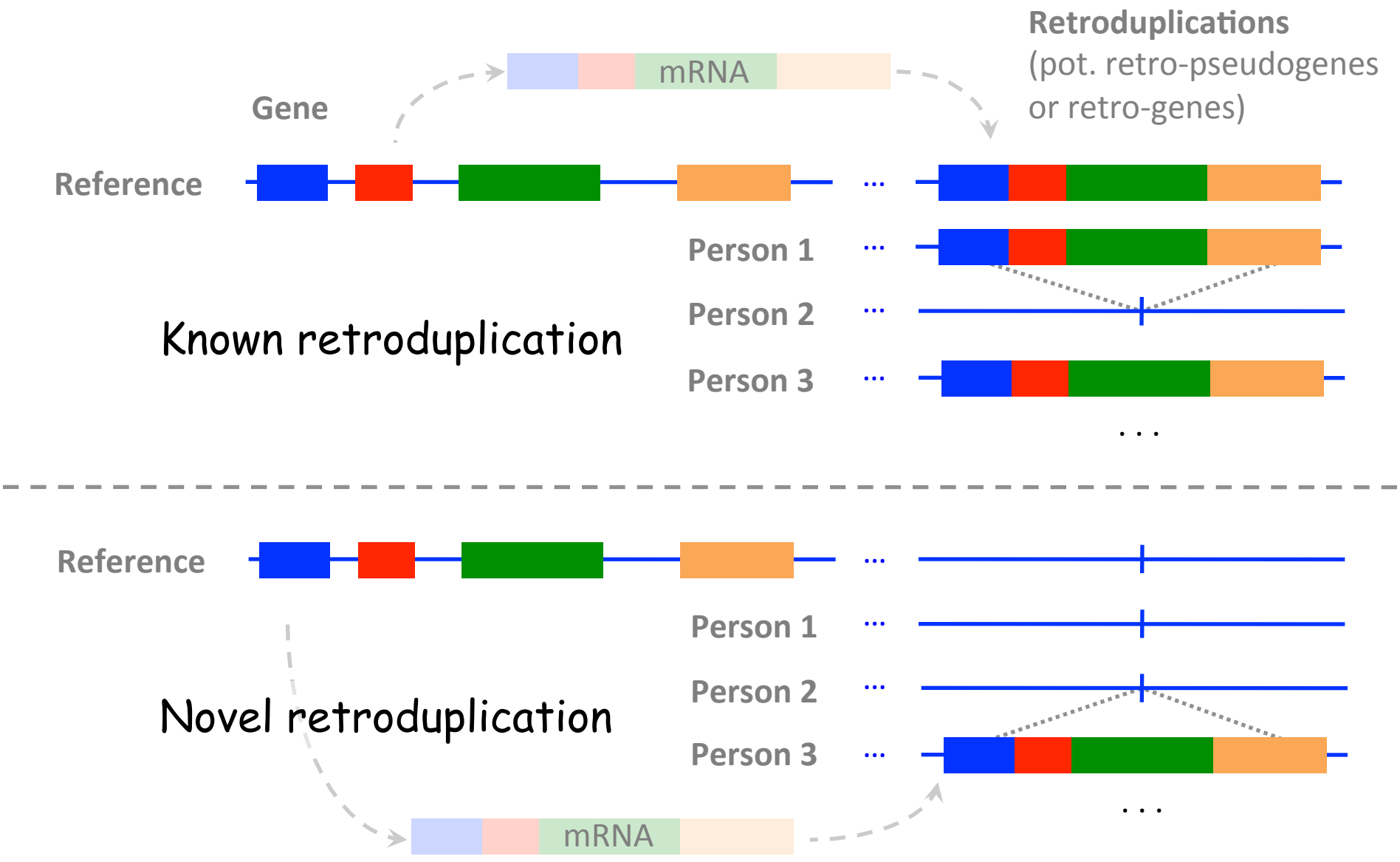


[Kim et al. Gen. Res. ('08)]

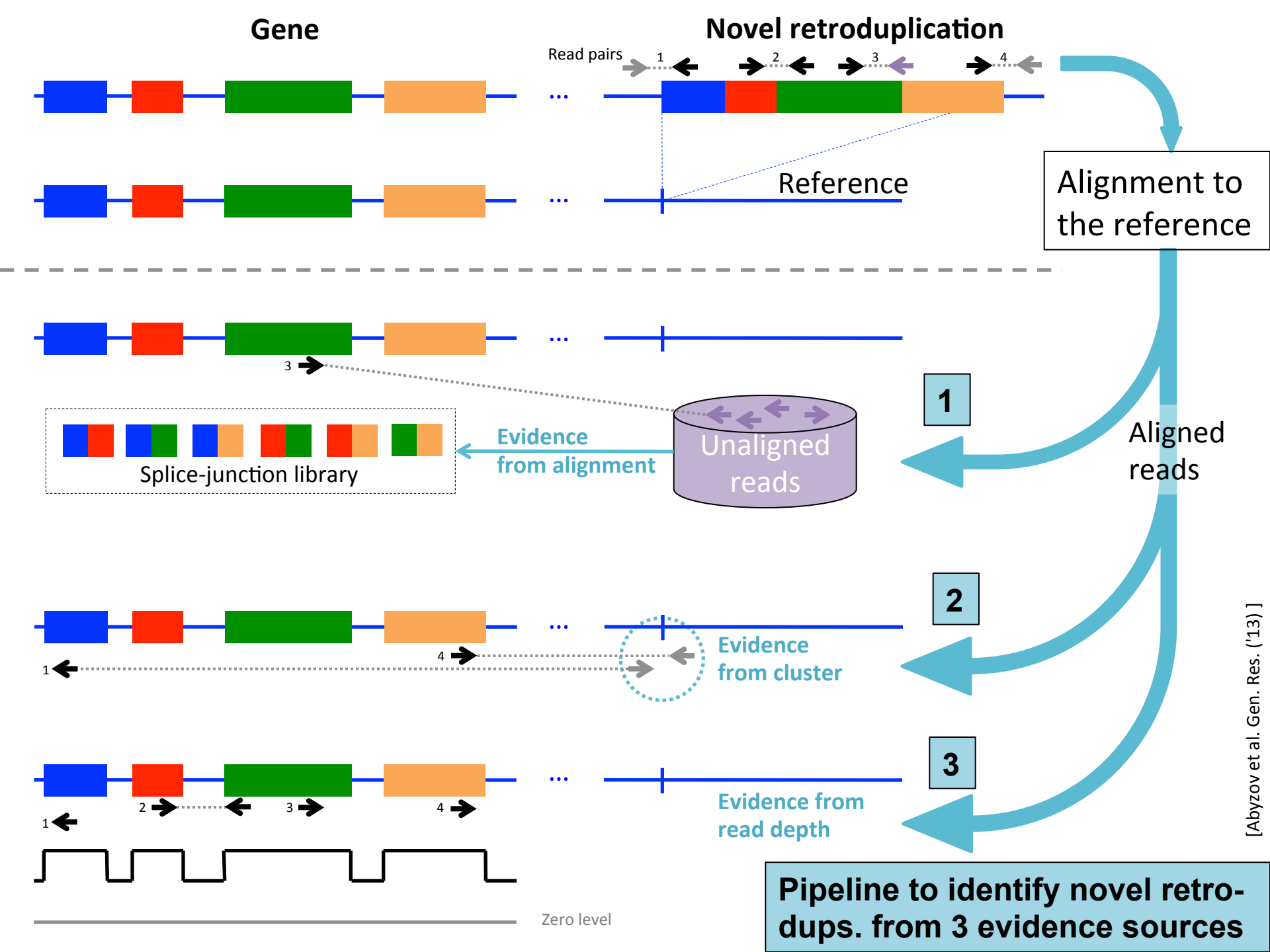
Human Genome Analysis – SVs & Pseudogenes, Tricky but Crucial Genomic Features, Targeted by Long-read Sequencing: Current Short-read Results & Future Prospects

- **SV Breakpoints**
  - ~9K deletions with breakpoints & mechanism classification from 1000G
  - Small subset of tot. deletions, which could be greatly expanded by long reads
  - More nearby SNPs than genomic average.
  - From methylation, Hi-C , & his mods, NAHR breakpoints associated with open chromatin (perhaps occurring w/o replication & division)
  - NHR breakpoints associated w/ sequence microinsertions, templated from later replicating sites, spaced at 2 characteristic distances
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
- **Intersection of Pseudogenes & SVs**
  - Enrichment of SVs in pseudogenes v genes, particularly for NAHR
- **Novel Processed Pseudogenes as a Form of SV**
  - Not in reference but in human population – could be improved by long reads
  - Now found w/ splice junction mapping + clustering of unmapped PEs
  - ~8 per person, often pop. specific
  - Associated w/ G1/M expressed genes
- **Many Pseudogenes with Low Levels of Biochemical Activity**
  - Conservative assignment, mis-map issue, could be improved by long reads
  - ~15% transcribed & 80% w/ some activity

Retroduplication variation (RDV)

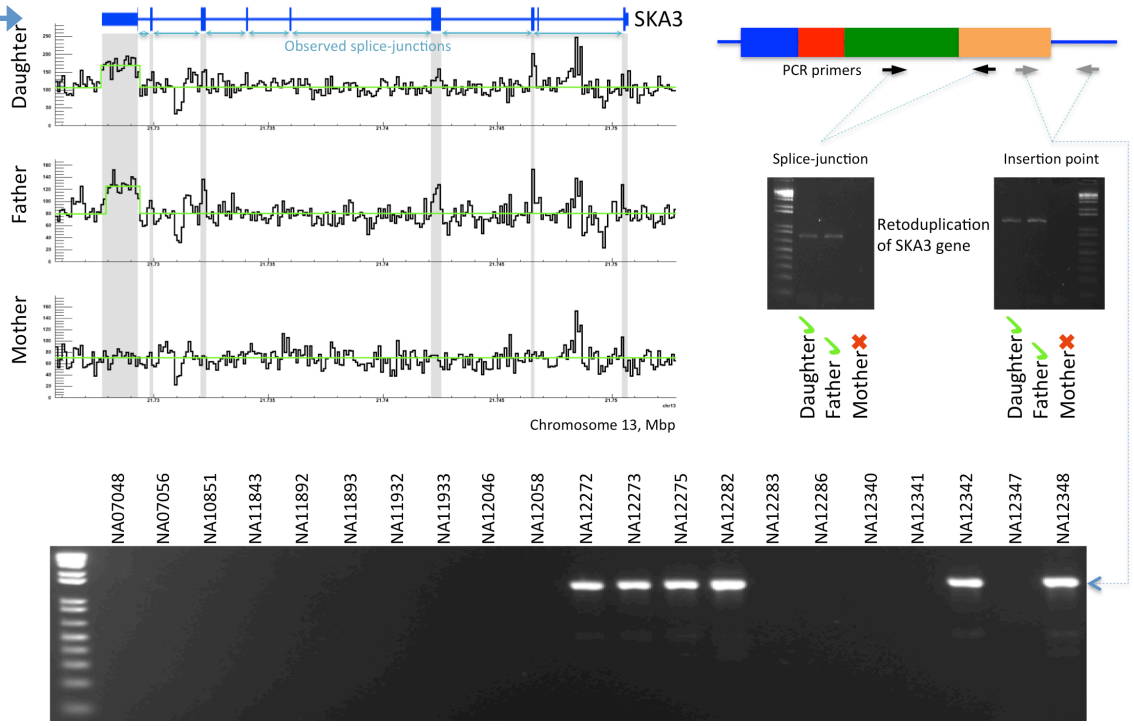


[Abyzov et al. Gen. Res. ('13)]


Gene
Novel retroduplication
Read pairs
1
2
3
4
Reference
Alignment to the reference
Aligned reads
3
Splice-junction library
Evidence from alignment
Unaligned reads
1
2
Evidence from cluster
3
Evidence from read depth
Zero level
[Abyzov et al. Gen. Res. ('13)]
Pipeline to identify novel retro-dups. from 3 evidence sources

A typical individual (NA12878) with 10 validated retrodups (by RD & PCR)

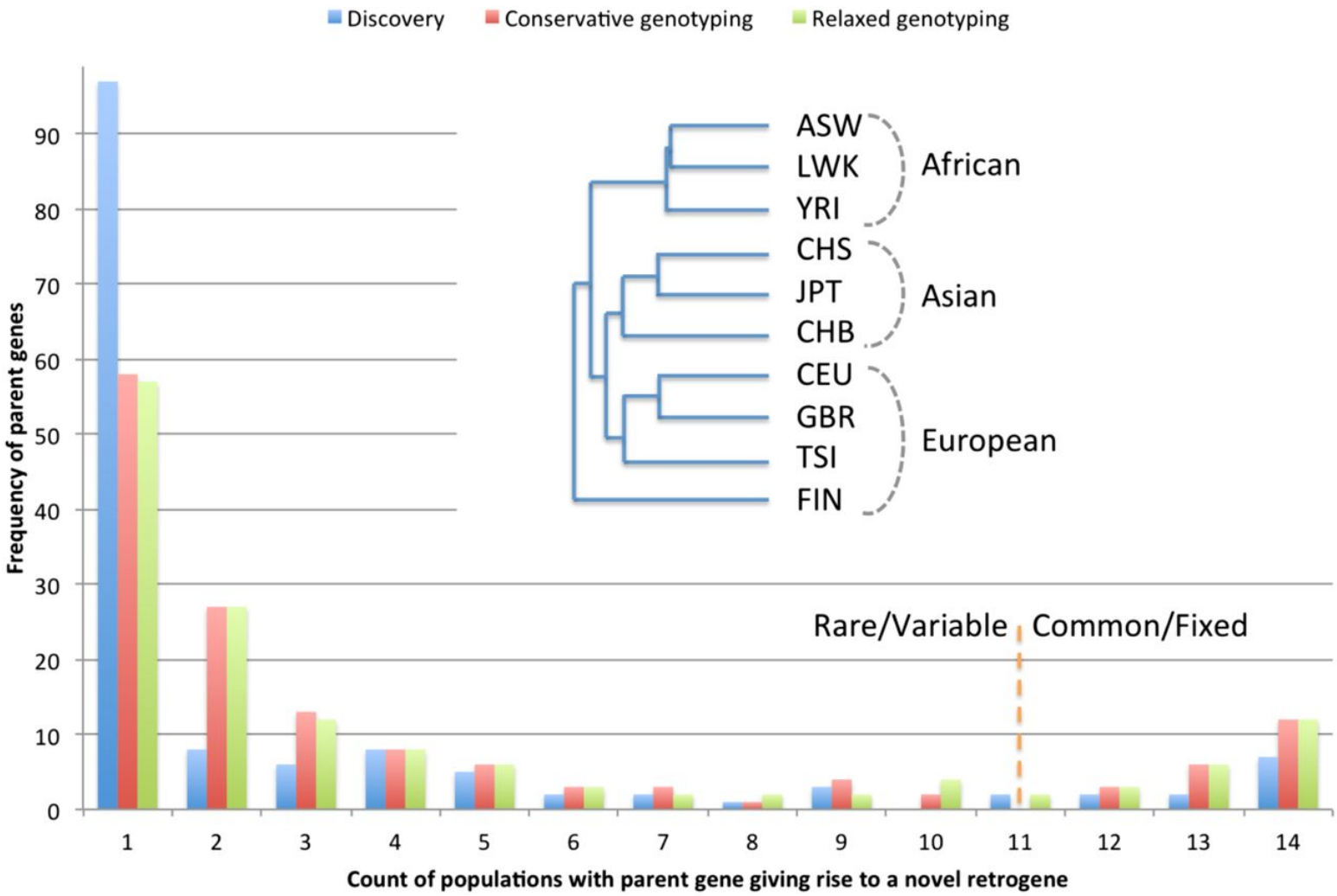
Parent gene with predicted novel retroduplication	Additional support			PCR validation
	Read depth support	Insertion point support	Found in Venter genome	
CDC27	Yes		Yes	UN
BCLAF1	Yes		Yes	UN
LAPTM4B	Yes	Yes		Yes
MTCH2				Yes
CBX3	Yes	Yes	Yes	Yes
TMEM66	Yes	Yes		Yes
TDG	Yes	Yes	Yes	Yes
BOD1				Yes
CACNA1B		Yes		Yes
SKA3	Yes	Yes		Yes
AP3S1	Yes		Yes	Yes
AC131157				N/A
AL590623	Centromere			



On avg. 6-10 novel Retrodups per person in 1000G dataset. Also, 147 total genes with retrodups

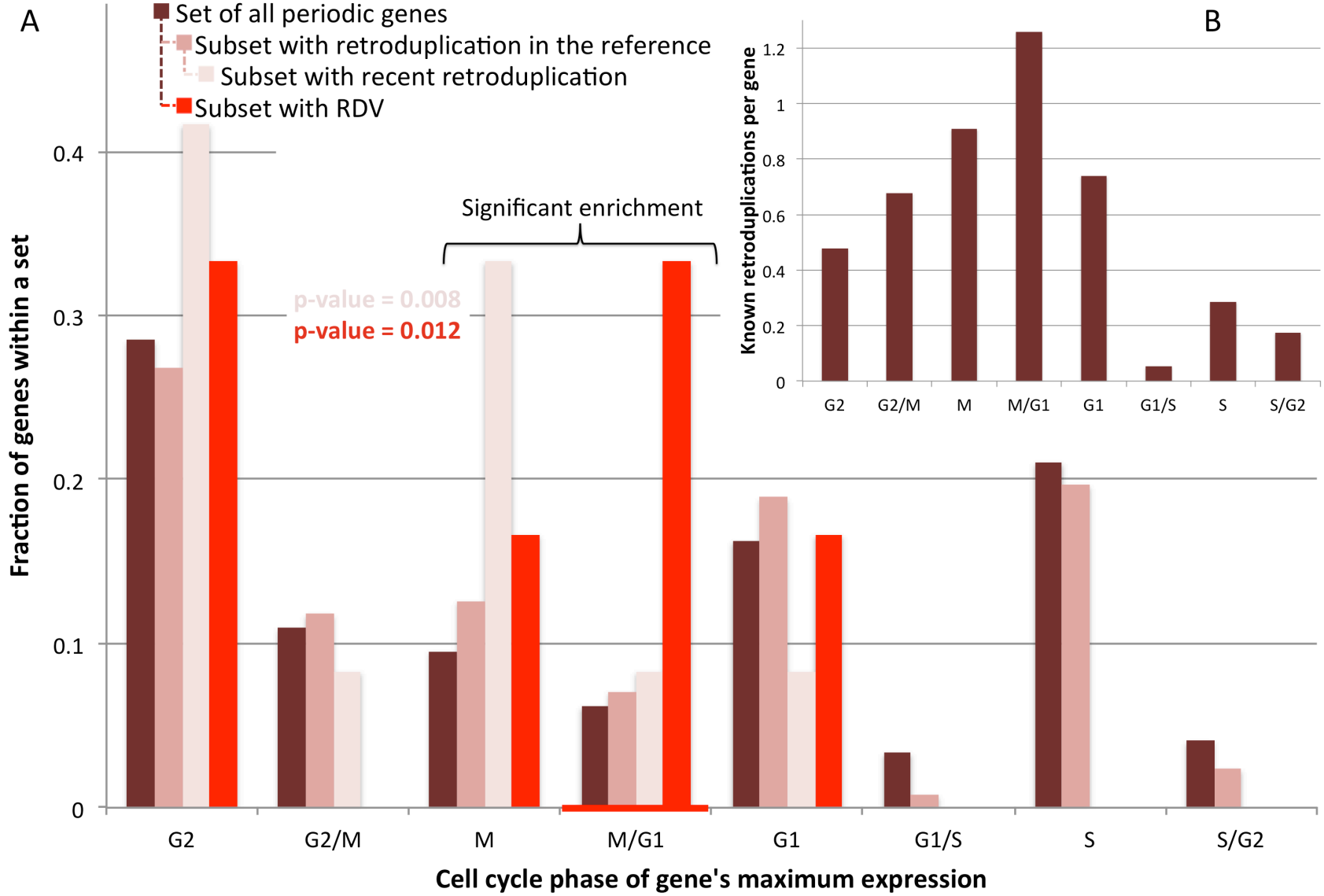
[Abyzov et al. Gen. Res. ('13)]

Frequency of novel retroduplications by populations.



Abyzov A et al. Genome Res. 2013;23:2042-2052

Hypothesis: **retrotransposition is coupled to cell division (in germline)**



[Abyzov et al. Gen. Res. ('13)]

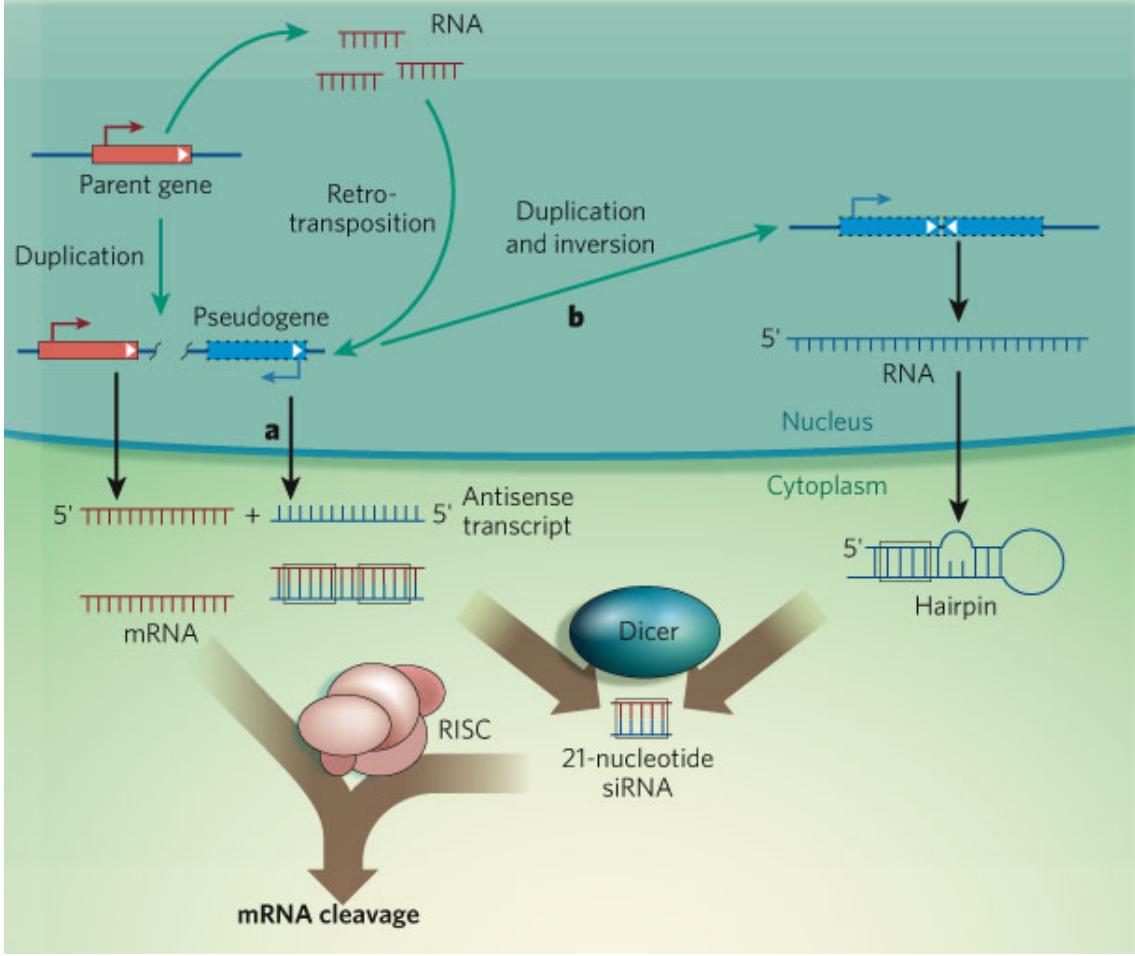
Human Genome Analysis – SVs & Pseudogenes, Tricky but Crucial Genomic Features, Targeted by Long-read Sequencing: Current Short-read Results & Future Prospects

- **SV Breakpoints**
  - ~9K deletions with breakpoints & mechanism classification from 1000G
  - Small subset of tot. deletions, which could be greatly expanded by long reads
  - More nearby SNPs than genomic average.
  - From methylation, Hi-C , & his mods, NAHR breakpoints associated with open chromatin (perhaps occurring w/o replication & division)
  - NHR breakpoints associated w/ sequence microinsertions, templated from later replicating sites, spaced at 2 characteristic distances
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
- **Intersection of Pseudogenes & SVs**
  - Enrichment of SVs in pseudogenes v genes, particularly for NAHR
- **Novel Processed Pseudogenes as a Form of SV**
  - Not in reference but in human population – could be improved by long reads
  - Now found w/ splice junction mapping + clustering of unmapped PEs
  - ~8 per person, often pop. specific
  - Associated w/ G1/M expressed genes
- **Many Pseudogenes with Low Levels of Biochemical Activity**
  - Conservative assignment, mis-map issue, could be improved by long reads
  - ~15% transcribed & 80% w/ some activity

Examples & speculation on the function of pseudogene ncRNAs:

Regulating their parents

- via acting as **endo-siRNAs** [ex. in fly & mouse, '08 refs.]
- via acting as **miRNA decoys** [PTEN]
- via **inhibiting degradation** of parent's mRNA [makorin]



[Sasidharan & Gerstein, Nature ('08)]

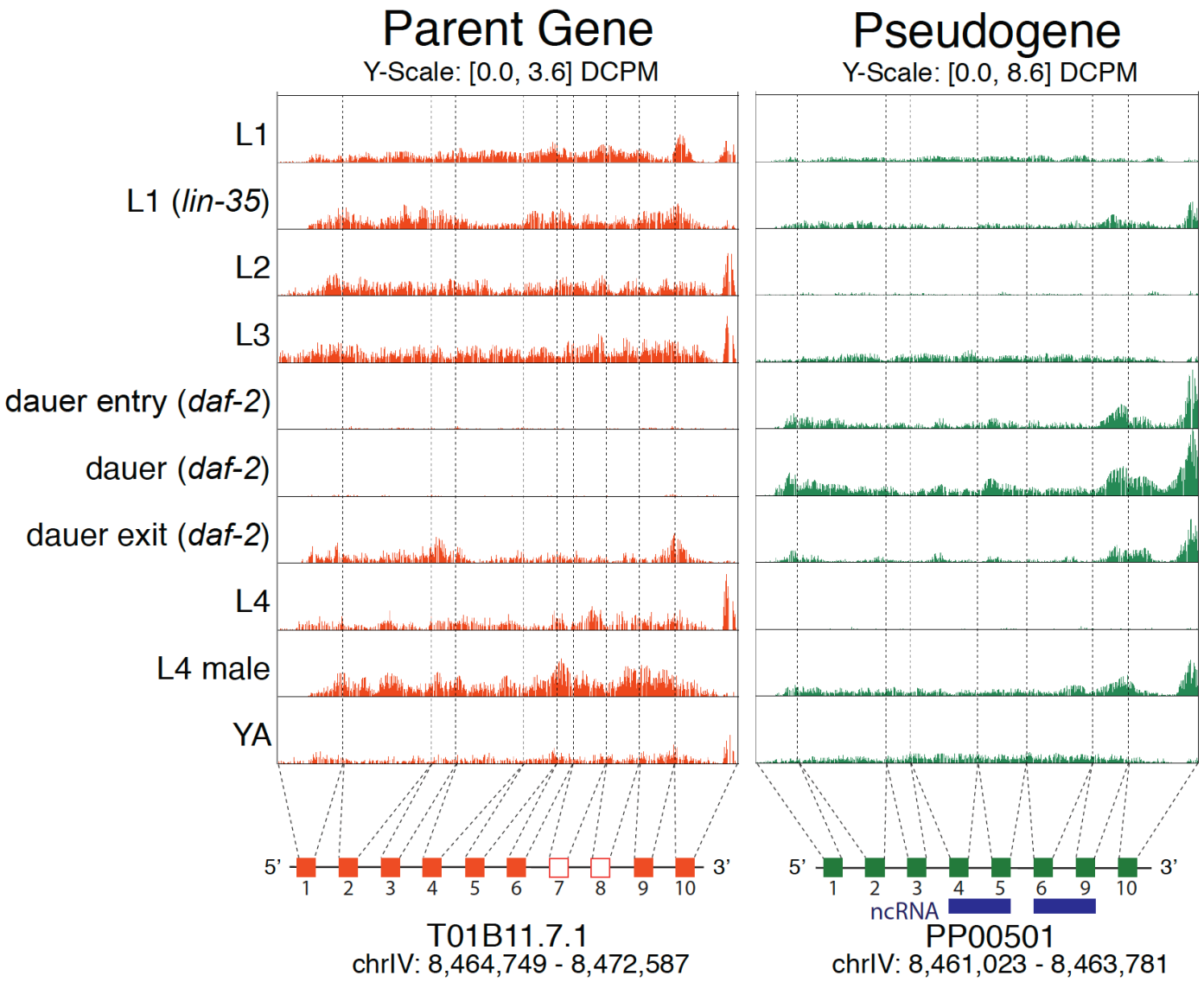
- Functional candidates:
  - SLIT-ROBO Rho GTPase activating protein 2B pseudogene
  - PRKY-004, Y-linked protein kinase pseudogene
  - Fer-1-like 4 (C. elegans), pseudogene

Alternatively, just last gasps of a dying gene

Czech *et al. Nature* 453: 798 ('08).
Ghildiyal *et al. Science* 320: 1077 ('08).
Kawamur *et al. Nature* 453: 793 ('08).
Okamura *et al. Nature* 453: 803 ('08).
Tam *et al. Nature* 453: 534 ('08).
Watanabe *et al. Nature* 453: 539 ('08).

Poliseno et al. Nature 465:1033 ('10).

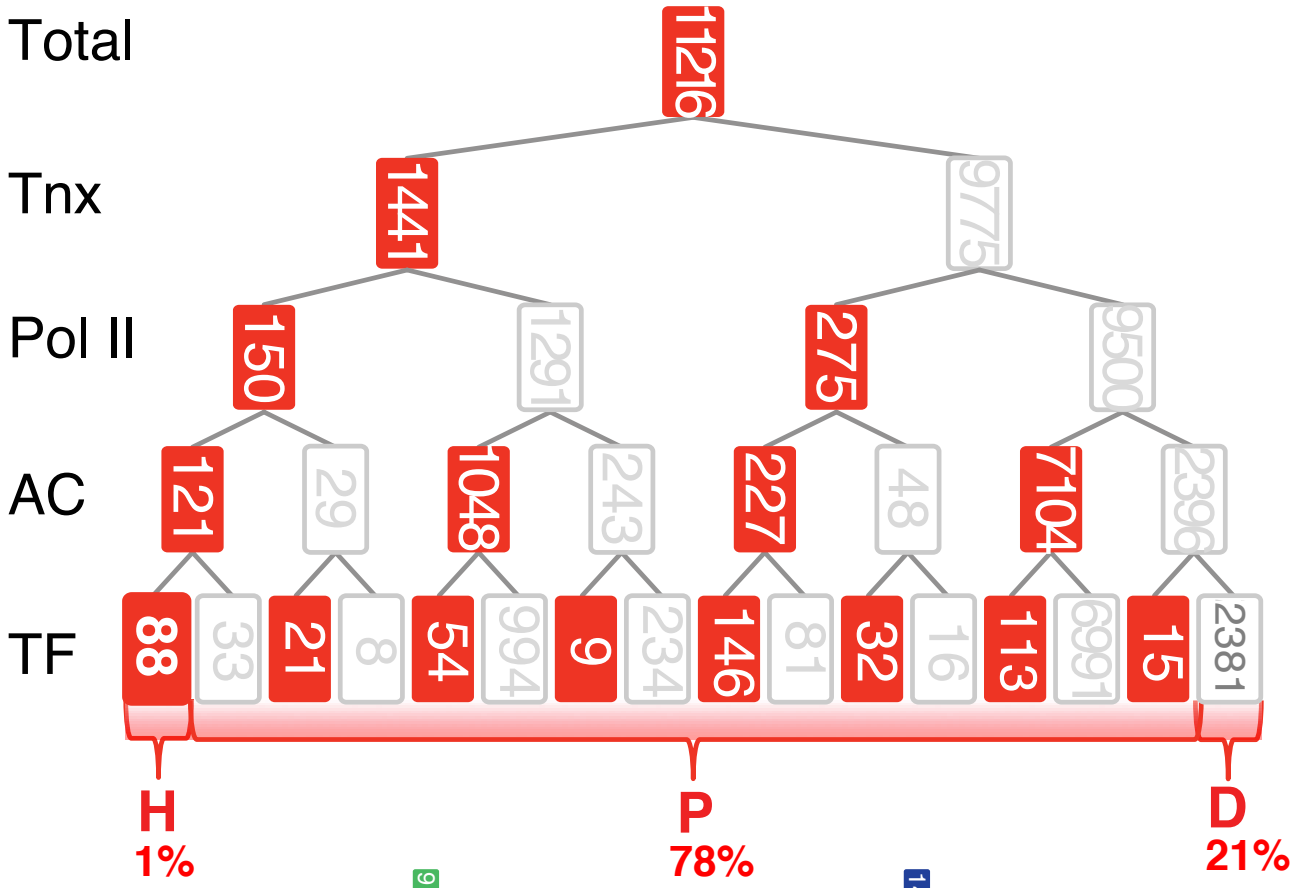
Pseudogene Transcription: interesting but tricky to ascertain

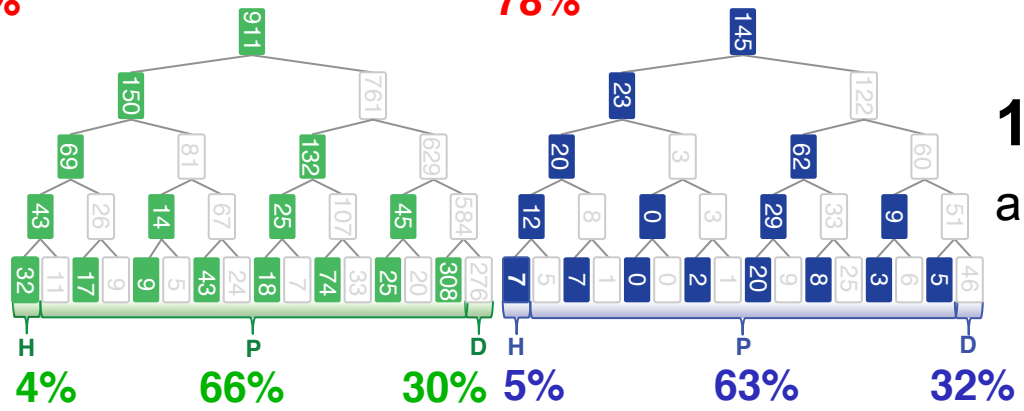


- Difficulty in ascertainment because of mis-mapping v parent
- One approach to this confound is look across mult. samples

[*Science* 330:6012]

Pseudogene Activity





H Highly-Active
P Partially-Active
D Dead
Yes
No
Human
Worm
Fly

15% of pseudogenes are **transcribed** in each organism

[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Human Genome Analysis – SVs & Pseudogenes, Tricky but Crucial Genomic Features, Targeted by Long-read Sequencing: Current Short-read Results & Future Prospects

- **SV Breakpoints**
  - ~9K deletions with breakpoints & mechanism classification from 1000G
  - Small subset of tot. deletions, which could be greatly expanded by long reads
  - More nearby SNPs than genomic average.
  - From methylation, Hi-C , & his mods, NAHR breakpoints associated with open chromatin (perhaps occurring w/o replication & division)
  - NHR breakpoints associated w/ sequence microinsertions, templated from later replicating sites, spaced at 2 characteristic distances
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
- **Intersection of Pseudogenes & SVs**
  - Enrichment of SVs in pseudogenes v genes, particularly for NAHR
- **Novel Processed Pseudogenes as a Form of SV**
  - Not in reference but in human population – could be improved by long reads
  - Now found w/ splice junction mapping + clustering of unmapped PEs
  - ~8 per person, often pop. specific
  - Associated w/ G1/M expressed genes
- **Many Pseudogenes with Low Levels of Biochemical Activity**
  - Conservative assignment, mis-map issue, could be improved by long reads
  - ~15% transcribed & 80% w/ some activity

Human Genome Analysis – SVs & Pseudogenes, Tricky but Crucial Genomic Features, Targeted by Long-read Sequencing: Current Short-read Results & Future Prospects

- **SV Breakpoints**
  - ~9K deletions with breakpoints & mechanism classification from 1000G
  - Small subset of tot. deletions, which could be greatly expanded by long reads
  - More nearby SNPs than genomic average.
  - From methylation, Hi-C , & his mods, NAHR breakpoints associated with open chromatin (perhaps occurring w/o replication & division)
  - NHR breakpoints associated w/ sequence microinsertions, templated from later replicating sites, spaced at 2 characteristic distances
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
- **Intersection of Pseudogenes & SVs**
  - Enrichment of SVs in pseudogenes v genes, particularly for NAHR
- **Novel Processed Pseudogenes as a Form of SV**
  - Not in reference but in human population – could be improved by long reads
  - Now found w/ splice junction mapping + clustering of unmapped PEs
  - ~8 per person, often pop. specific
  - Associated w/ G1/M expressed genes
- **Many Pseudogenes with Low Levels of Biochemical Activity**
  - Conservative assignment, mis-map issue, could be improved by long reads
  - ~15% transcribed & 80% w/ some activity

Consortium Acknowledgements

WashU - **Ken Chen**, Asif Chinwalla, Donald Conrad, Li Ding, Mike McLellan, John Wallis

WT Sanger Inst. – Ben Blackburne, Richard Durbin, Matt Hurles, Heng Li, Zemin Ning, Alywyn Scally, Klaudia Walter, Manuela Zanda, Yujun Zhang

Yale –Alexej Abyzov, Jieming Chen, Declan Clarke, Mark Gerstein, Rajini Haraksingh, Ekta Khurana, Joe Lee, Jing Leng, Cristina Sisu, Daifeng Wang

Stanford – Fabian Grubert Mark Kaganovich Phil Lacroute Hugo Lam Michael Snyder Alexander Urban

EMBL – Tobias Rausch, Andreas Schlattl, Adrian Stütz

Univ. of Washington – Tonia Brown, Arthur Ko, Peter Sudmant

EBI – Ewan Birney Laura Clarke Paul Flicek Matthias Haimel Paul Kersey Ilkka Lappalainen Lisa Skipper Richard Smith Daniel Zerbino Xiangqun Zheng-Bradley

Oxford – Zamin Iqbal, Gerton Lunter, Gil McVean

LSU – Mark Batzer, Miriam Konkel, Jerilyn Walker

Simon Fraser – Iman Hajirasouliha, Fereydoun Hormozdiari

Bilkent University – Can Alkan

Brigham – Xinghua Shi, Chengsheng Zhang

Cornell – Jeremiah Degenhardt

Harvard – Marcin Von Grotthuss

Rutgers – Jinchuan Xing

TGen – David Craig

AECOM – Kenny Ye

UCSD – Vineet Bafna, Jacob Michaelson, Jonathan Sebat

UCLA – Stan Nelson

Illumina – Bret Barnes, David Bentley, Michael Eberle, R. Keira Cheetham, Sean Humphray, Scott Kahn, Lisa Murray, Richard Shaw, Michael Stromberg

Life Tech. – Yutao Fu, Fiona Hyland, Heather Peckham, Yongming Sun, Daryl Thomas, Sowmi Utiramerur

BC – Erik Garrison Deniz Kural Wan-Ping Lee Gabor Marth Chip Stewart Alistair Ward Jiantao Wu

Broad Inst. – Guillermo del Angel, David Altshuler, Eric Banks, Mark DePristo, Menachem Fromer, **Robert Handsaker,** Chris Hartl, Steve McCarroll, James Nemesh, Khalid Shakir

Univ. of Michigan – Gonçalo Abecasis, Tom Blackwell, Jeffrey Kidd, **Ryan Mills,** Matthew Snyder

BGI – Yingrui Li, Srinka Ghosh, Aaron Halpern, Jason Laramie, Steve Lincoln

Leiden Univ. – Kai Ye

MS School of Medicine – Jayon Lihm, Vladimir Makarov, Elena Parkhomenko, Seungtai Yoon

Baylor – James Lu, Jeff Reid, Fuli Yu

Univ. of Maryland – Scott Devine

Univ. of Uath – David Witherspoon

Univ. of Virginia – Aaron Quinlan

NIH – Chunlin Xiao

Co-chairs – **Charles Lee, Jan Korbel, Evan Eichler**

GENCODE

Tim Hubbard, WT Sanger Inst.

Jennifer Harrow (lead PI), WT Sanger Inst.

Steve Searle, WT Sanger Inst.

Alexandre Reymond (PI), Univ. of Lausanne

Roderic Guigo (PI), CRG

David Haussler (PI), UCSC

Rachel Harte (Co-PI), UCSC

Manolis Kellis (PI), MIT

Mark Gerstein (PI), Yale

Alfonso Valencia (PI), CNIO

Michael Tress, CNIO

Acknowledgements

Breakpoints: **A Abyzov,** S Li, DR Kim, M Mohiyuddin, AM Stütz, NF Parrish, XJ Mu, W Clark, K Chen, M Hurles, JO Korbel, HY Lam, C Lee

GAPDH Pseudogene: Yuen-Jong Liu, Deyou Zheng, Suganthi Balasubramanian, Nicholas Carriero, Ekta Khurana, Rebecca Robilotto

Non-coding Variation: **X Mu,** Zhi J. Lu, Yong Kong, Hugo Y.K. Lam, Y Fu, E Khurana

Retroduplication Variation: **A Abyzov,** Yan Zhang, Shantao Li, Rebecca Iskow, Omer Gokcumen, David W. Radke, Suganthi Balasubramanian, Baikang Pei, Lukas Habegger, The 1000 Genomes Project Consortium, **C Lee**

Comparative Pseudogene: **C Sisu, B Pei,** Jing Leng, **A Frankish,** Yan Zhang, Suganthi Balasubramanian, Rachel Harte, Daifeng Wang, Michael Rutenberg-Schoenberg, Wyatt Clark, Mark Diekhans, Joel Rozowsky, Tim Hubbard, **J Harrow**

SV.gersteinlab.org + Pseudogene.org

Hiring Postdocs. See gersteinlab.org/jobs

Extra Slides

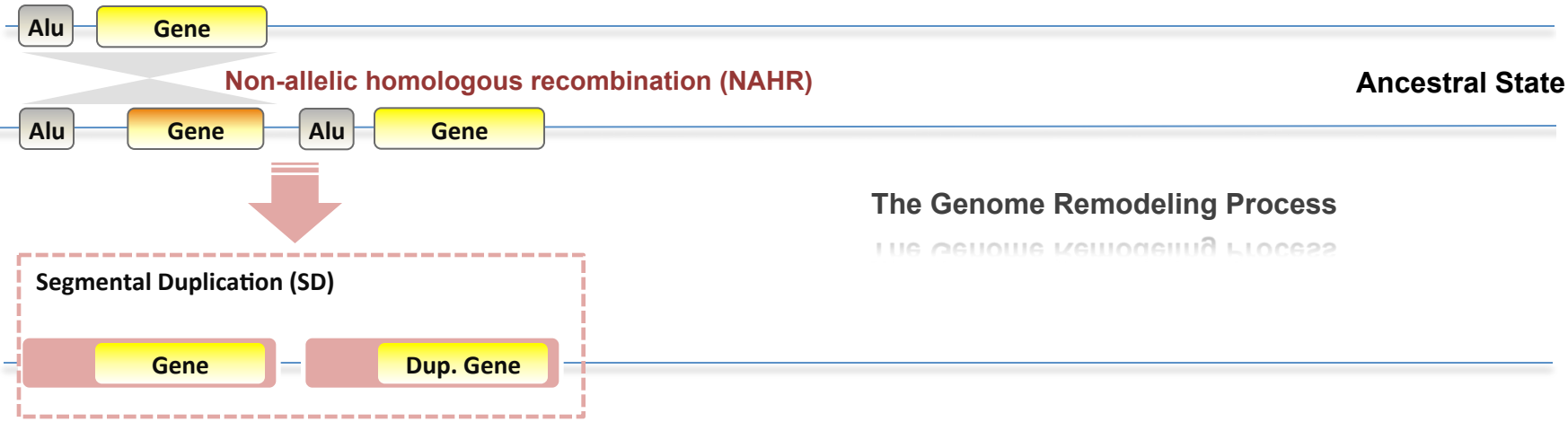
Lessons learned

from comparing the human, worm, and fly pseudogenes

- Mammalian pseudogenes are defined **independently** by a large event: the **retrotransposition burst 40 MYA**
- **Human** pseudogenes are defined by:
  - a majority of **young processed** pseudogenes
    - highly transcribed
    - located in regions of low recombination & near the centromeres
  - **old duplicated** pseudogenes
    - hints to the shared ancestry with worm & fly
- **Worm & Fly** pseudogenes are defined by:
  - selective sweeps
  - large population size
  - tandem duplications
- There are **NO pseudogenes orthologs** between distant species, however there are human-mouse orthologous pseudogenes
  - pseudogene **families are lineage specific**
  - few **universal** families across distant species
- **Activity** levels are **conserved** across all organisms
  - **15%** of pseudogenes **are transcribed** in all organisms

Genomic Variation

Genomic Variation



Genomic Variation

Alu Gene

Non-allelic homologous recombination (NAHR)

Ancestral State

Alu Gene Alu Gene

The Genome Remodeling Process

Segmental Duplication (SD)

Gene Dup. Gene

Syntenic Ortholog

SD

Gene Dup. Gene

Paralog

duplicate

Gene Dup. Gene Dup. Gene

family

Gene Dup. ψgene

Genomic Variation

Alu Gene

Non-allelic homologous recombination (NAHR)

Ancestral State

Alu Gene Alu Gene

The Genome Remodeling Process

Segmental Duplication (SD)

Gene Dup. Gene

Syntenic Ortholog

SD

Gene Dup. Gene

Paralog

duplicate

Gene Dup. Gene Dup. Gene

family

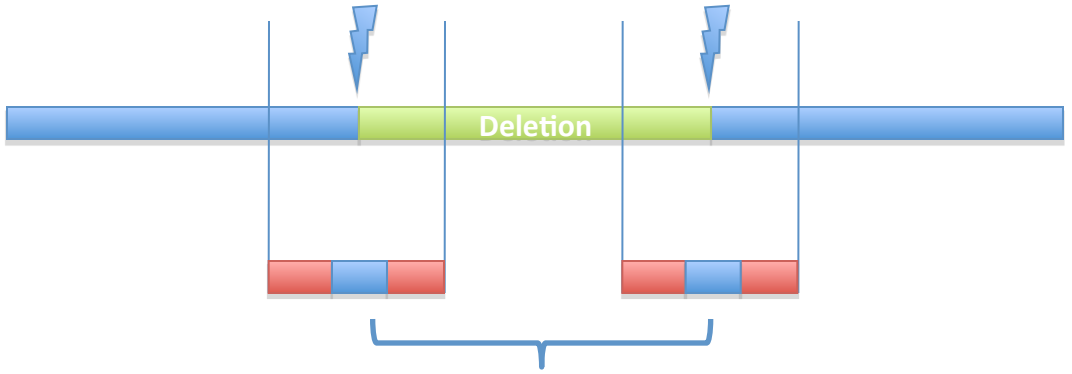
Gene Dup. ψgene

Pssd. ψgene

Retro-transpose

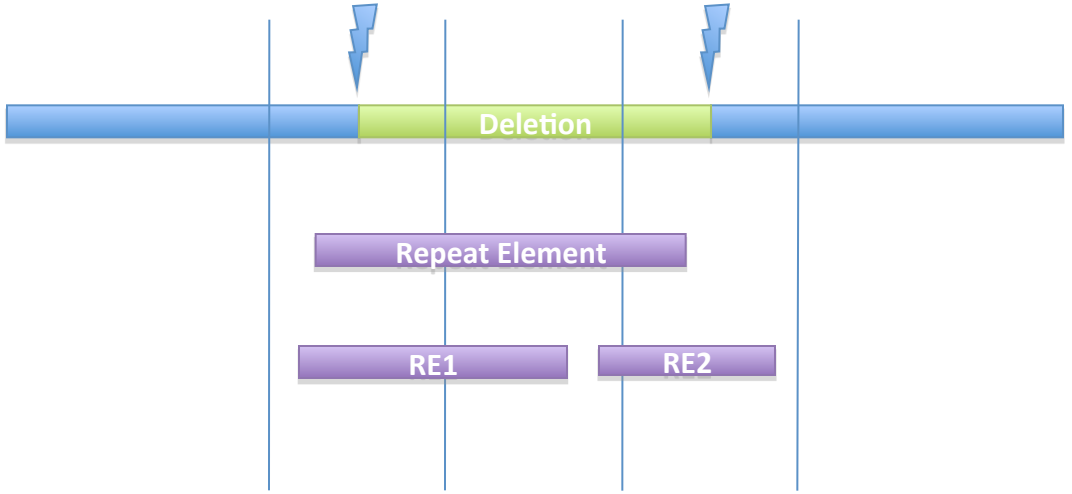
SV Mechanism Classification

NAHR



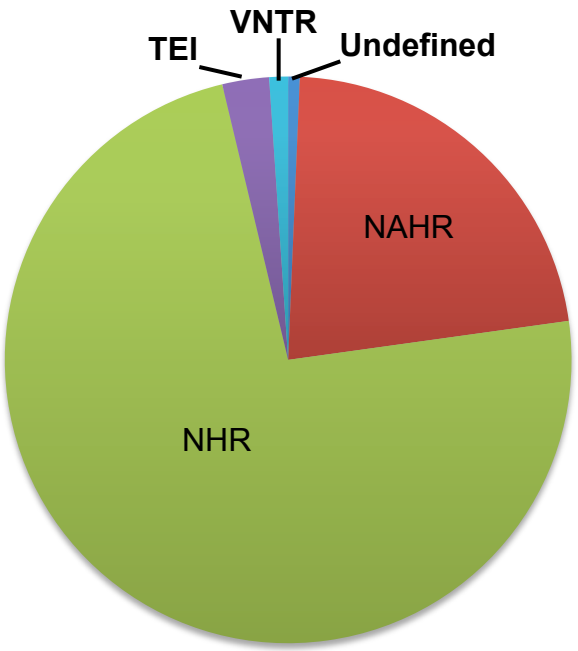
Single RETRO

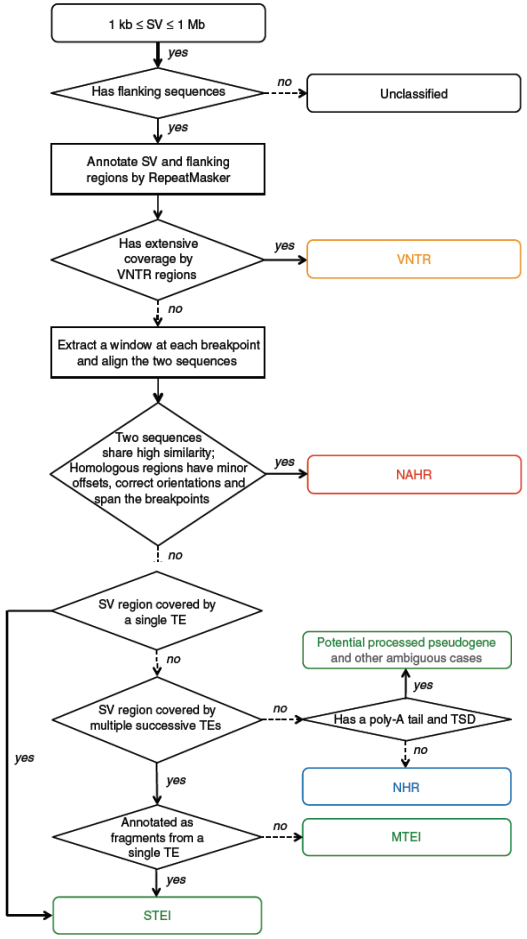
Multiple RETRO



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

Summary of Mechanism Classification of ~8900 Deletion Breakpoints in 1000G Phase I

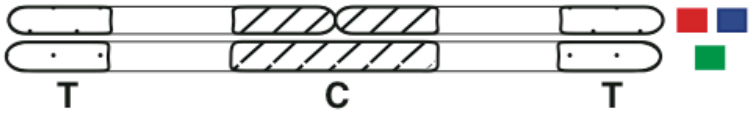




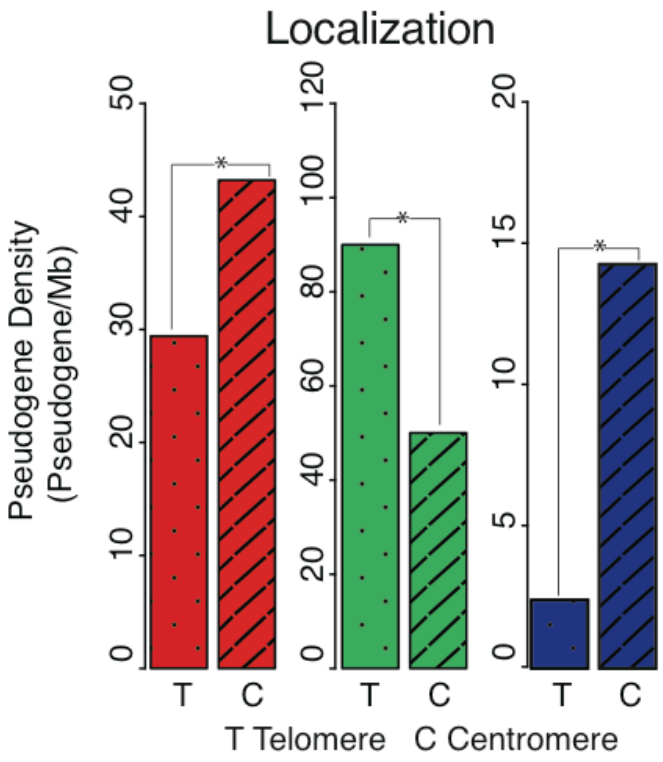
[1000 Genomes Consortium, Nature (2012)]
[Lam et al., ('10) *Nat. Biotech.*]

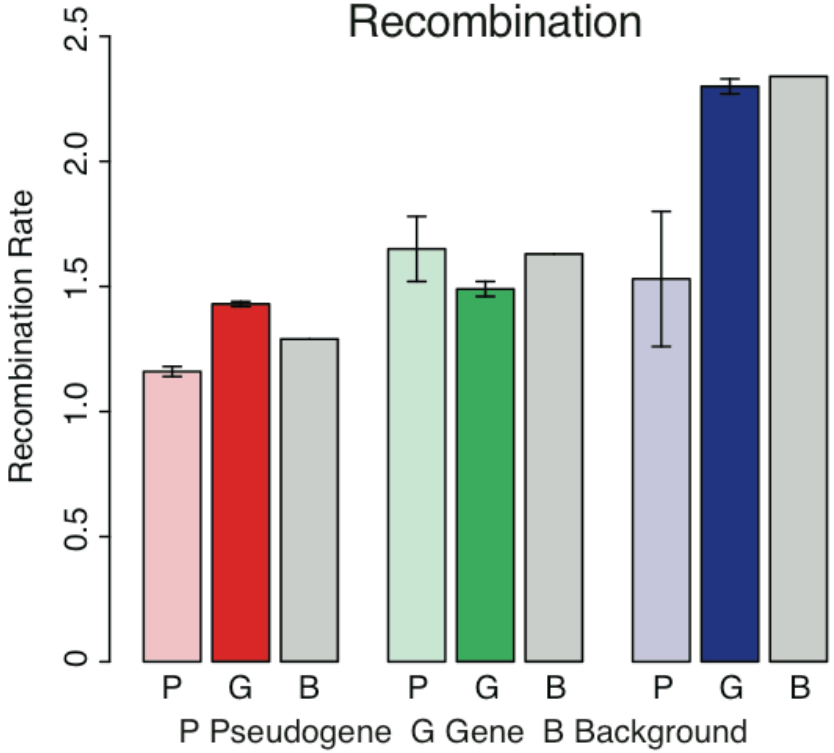
Mechanism	<500 bps	500-1000 bps	1-10 kbps	>10 kbps
NAHR	9 (2.6%)	294 (23.3%)	1420 (22.6%)	255 (24.7%)
NHR	284 (82.8%)	889 (70.4%)	4642 (73.7%)	748 (72.4%)
MEI	47 (13.7%)	67 (5.3%)	124 (2.0%)	0 (0%)
VNTR	2 (0.6%)	7 (0.6%)	64 (1.0%)	23 (2.2%)
Undefined	1 (0.3%)	6 (0.5%)	45 (0.7%)	7 (0.7%)
Total	343 (100%)	1263 (100%)	6295 (100%)	1033 (100%)

Localization



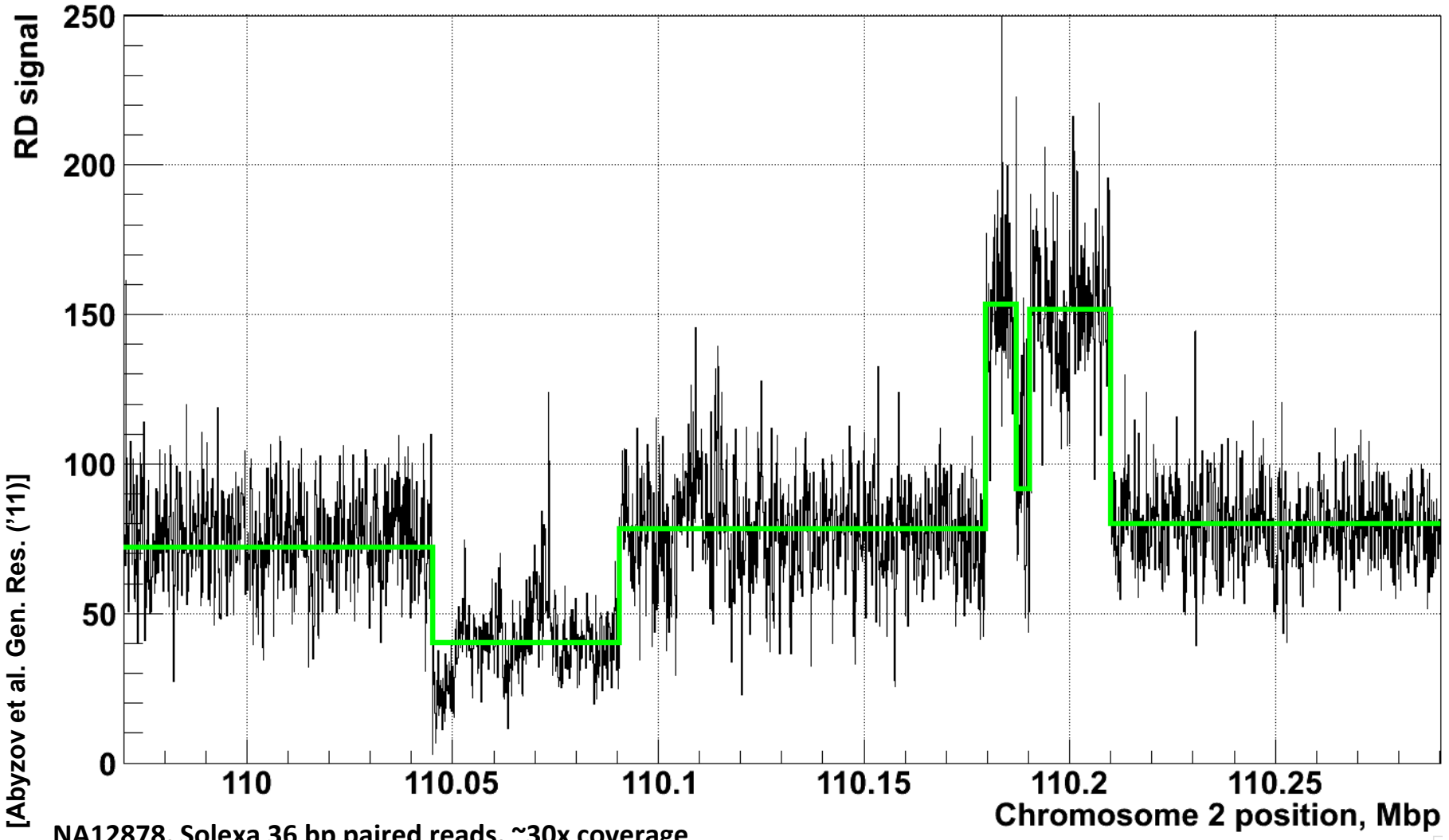
Human
Worm
Fly





[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

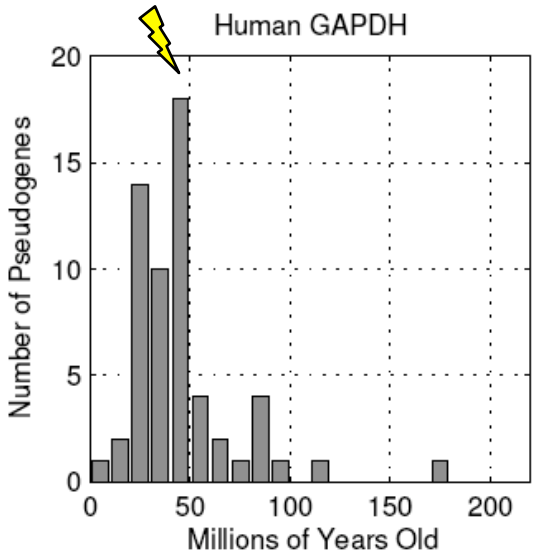
Example of Application of CNVnator to RD data

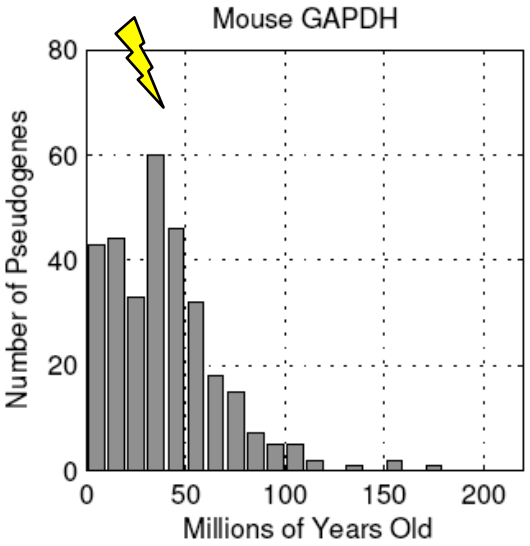


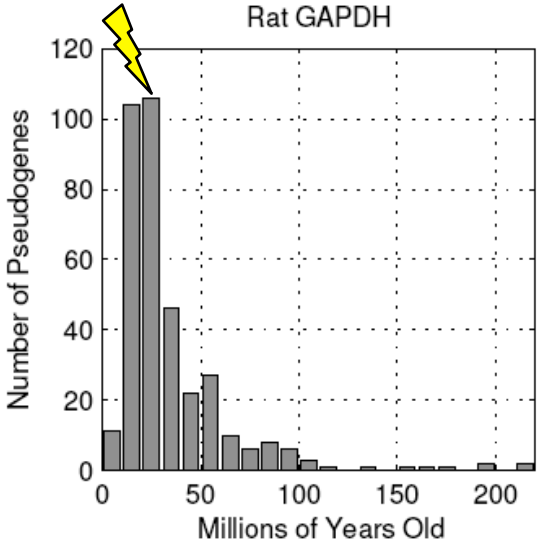
Mammalian GAPDH

Burst of Retrotranspositional Activity

40 MYA







GAPDH synteny

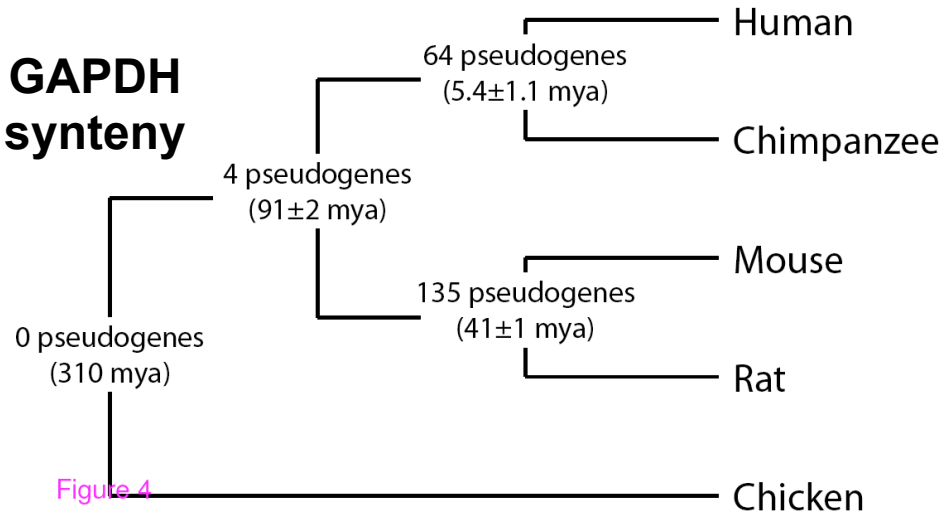


Figure 4

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

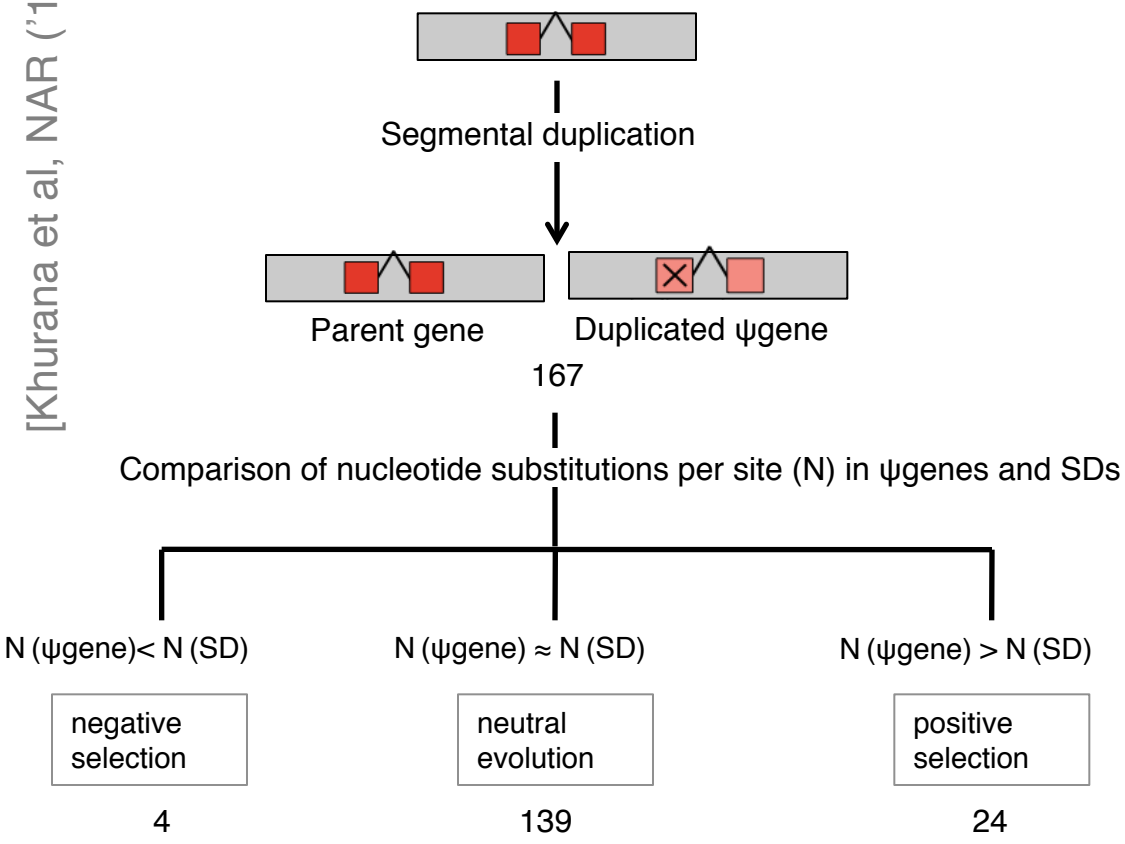
Similar patterns of pseudogene appearence across mammals.

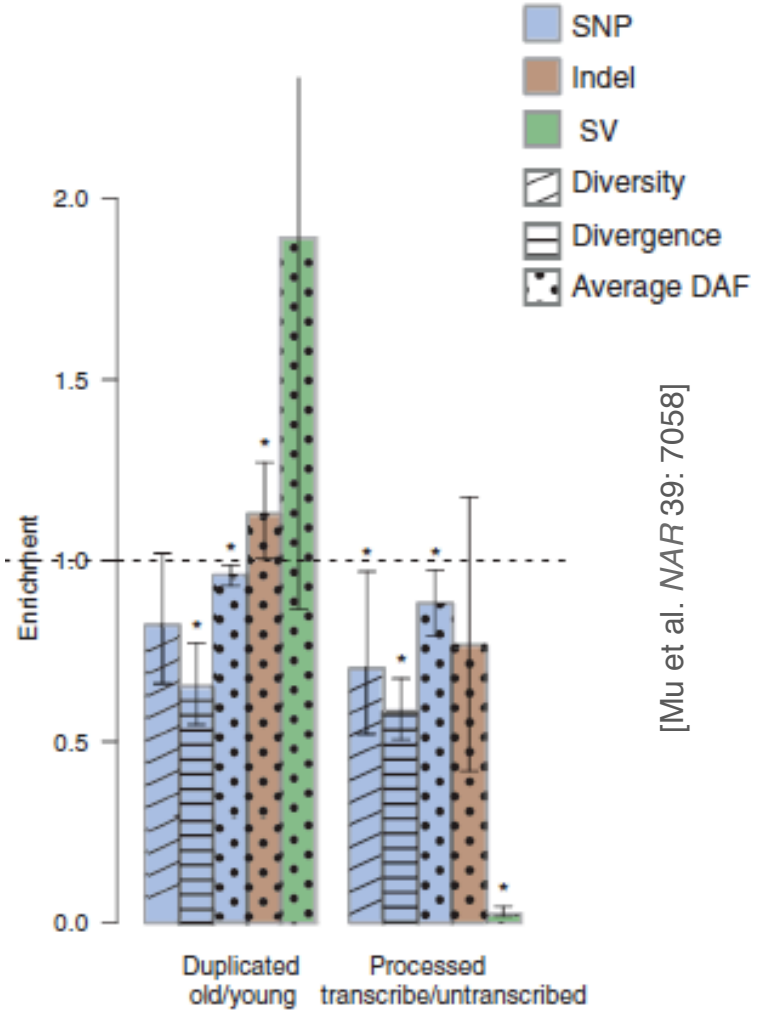
However retro events are **independent** for mammals.

Hints of Selection on Some Pseudogenes

- **Ka/Ks** conventional measure of selection for genes, shows no signal for pgenes
- Signature for selection on some SD pgenes (16%), derived from intersecting with UW SD DB & looking for differential conservation of neighborhood vs. center of pgene
- Weak signature for greater selection on transcribed pgenes using 1000G polymorphisms

[Khurana et al, NAR ('10)]





[Mu et al. *NAR* 39: 7058]

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2015.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt