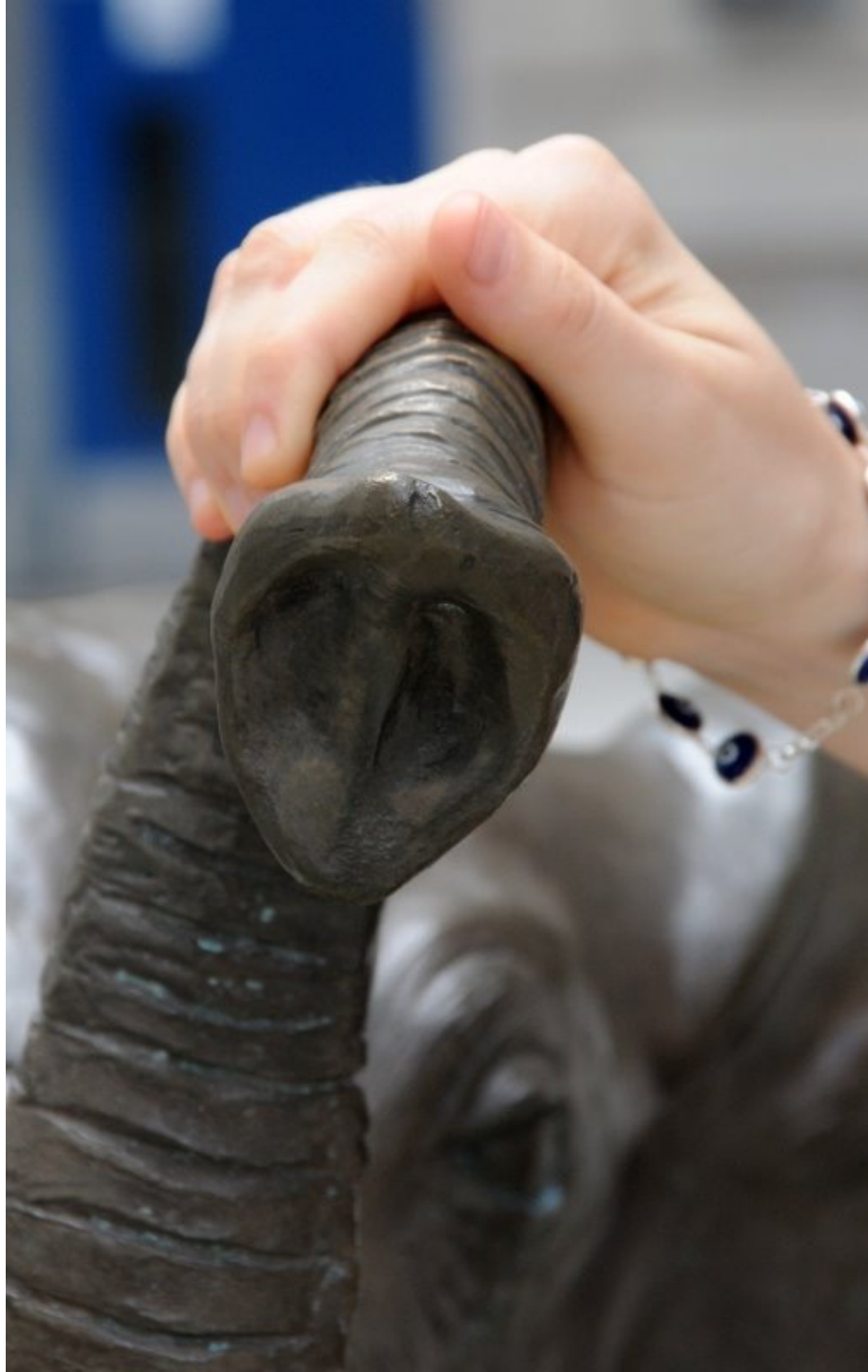
The Blind Men & the Elephant: Relating Pseudogenes, LOF, Retrodup Variation & Personalized Annotation

Mark Gerstein, Yale

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



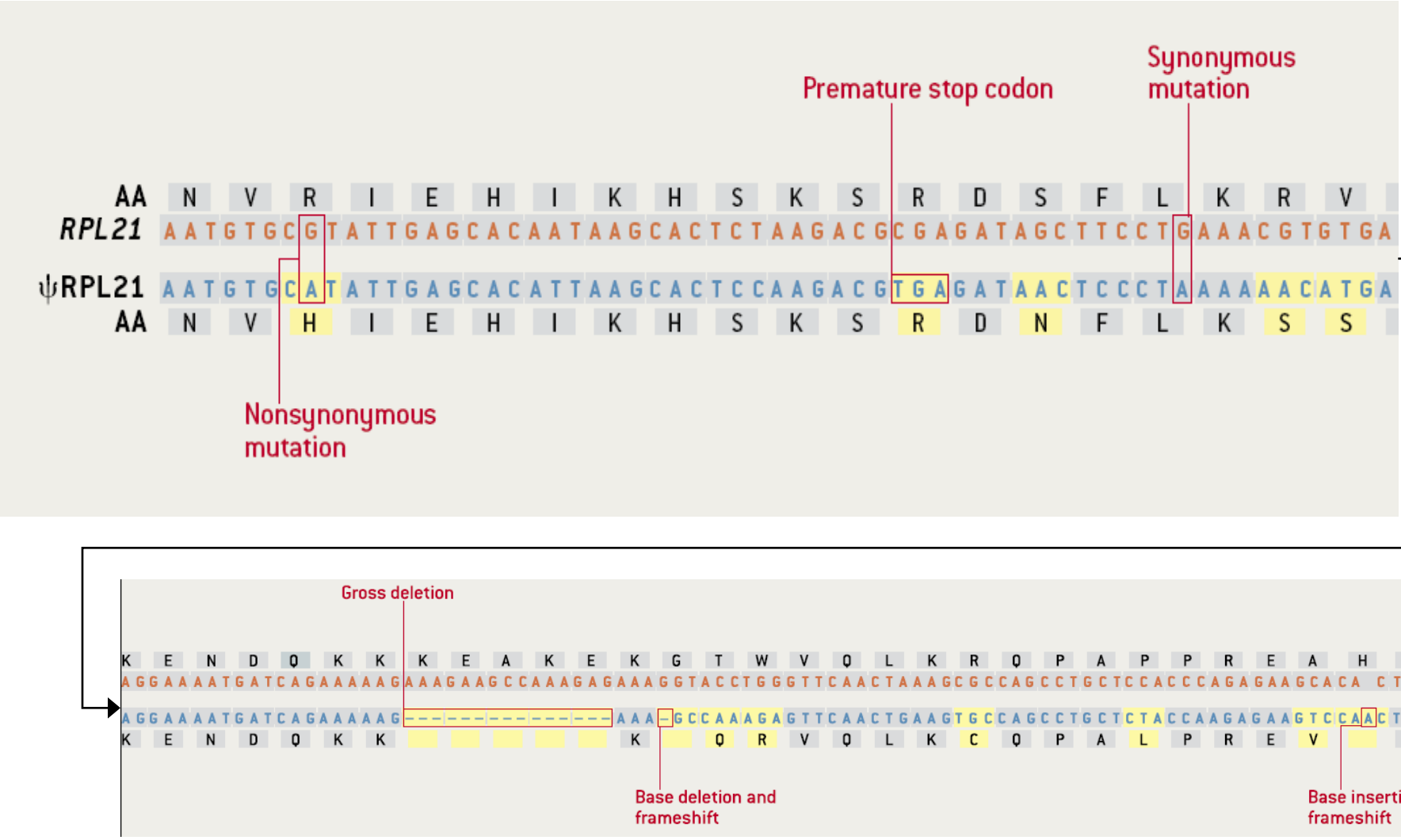
The Blind Men & the Elephant: Relating Pseudogenes, LOF, Retrodup Variation & Personalized Annotation

- LOFs
  - The spectrum:
    Genes=>LOFs=>
    Polymorphic Pseudogenes
    =>Fixed Pseudogenes
  - VAT & ALOFT
  - Large Diversity in 1000G-P3
- RDV
  - Further variation in polymorphic pseudogenes due to retroduplication
  - Absence of selection
- Personal Annotation
  - Personal Genomes
  - Best ref ?
  - No LOFs
- Mouse Strains
  - Putting it all together

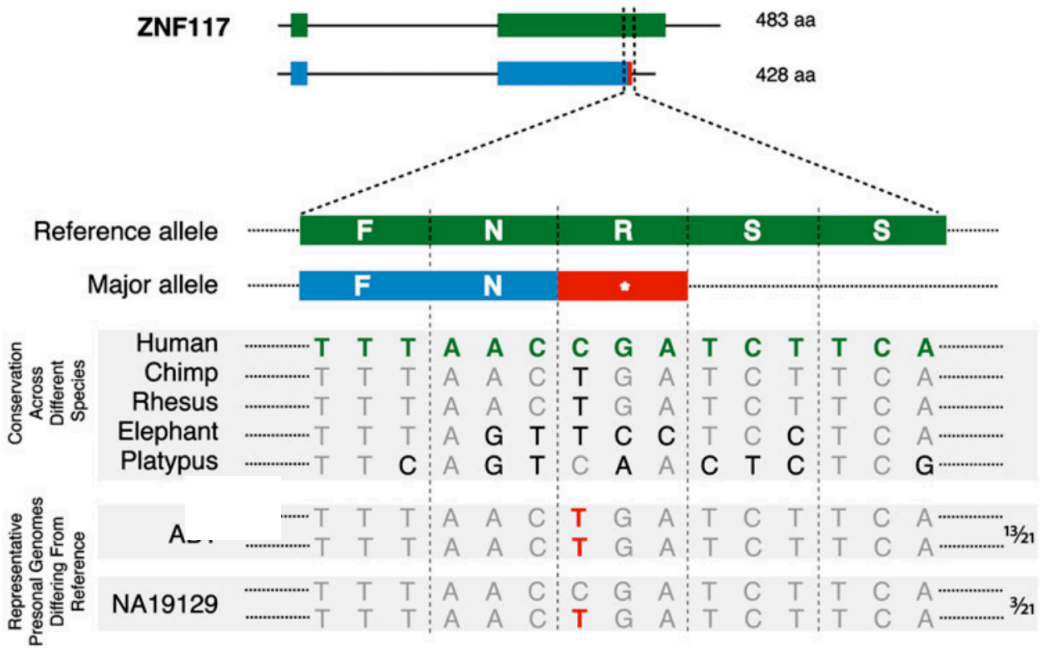
The Blind Men & the Elephant: Relating Pseudogenes, LOF, Retrodup Variation & Personalized Annotation

- LOFs
  - The spectrum: Genes=>LOFs=> Polymorphic Pseudogenes =>Fixed Pseudogenes
  - VAT & ALOFT
  - Large Diversity in 1000G-P3
- RDV
  - Further variation in polymorphic pseudogenes due to retroduplication
  - Absence of selection
- Personal Annotation
  - Personal Genomes
  - Best ref ?
  - No LOFs
- Mouse Strains
  - Putting it all together

The Canonical Pseudogene, with disablements fixed in the population (ψRPL21)

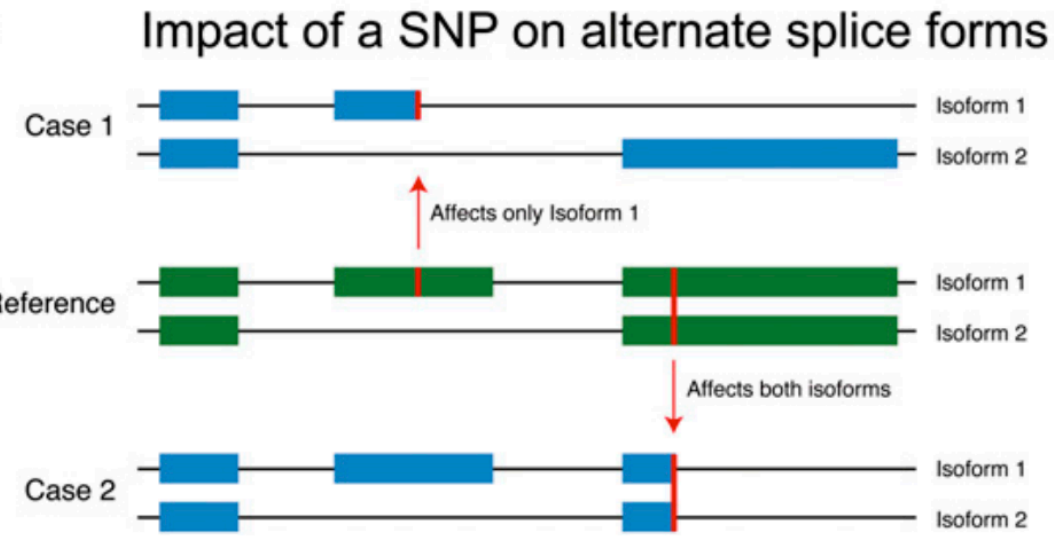


[Gerstein & Zheng. Sci Am 295: 48 (2006).]



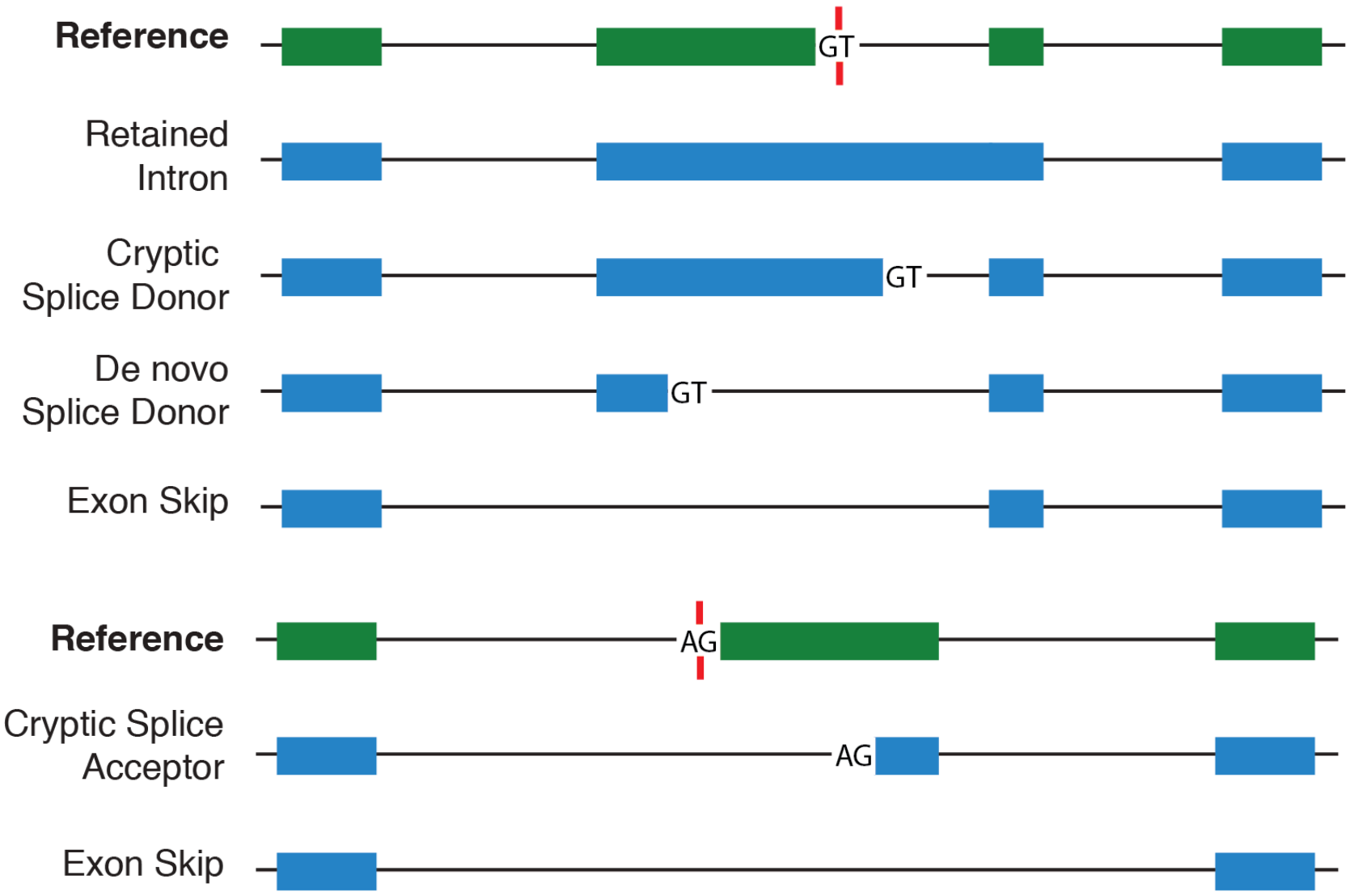
LOF variants

creation of "pseudogenes" from annotated genes



[Balasubramanian et al., *Genes Dev.*, '11]

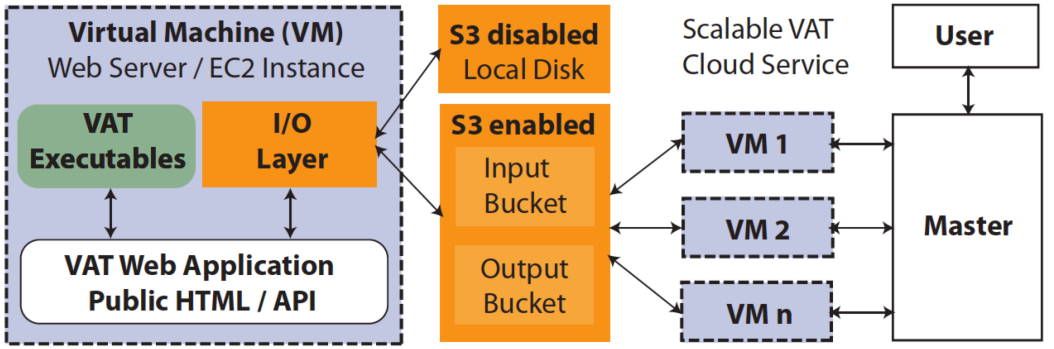
SNP IN SPLICE SITE



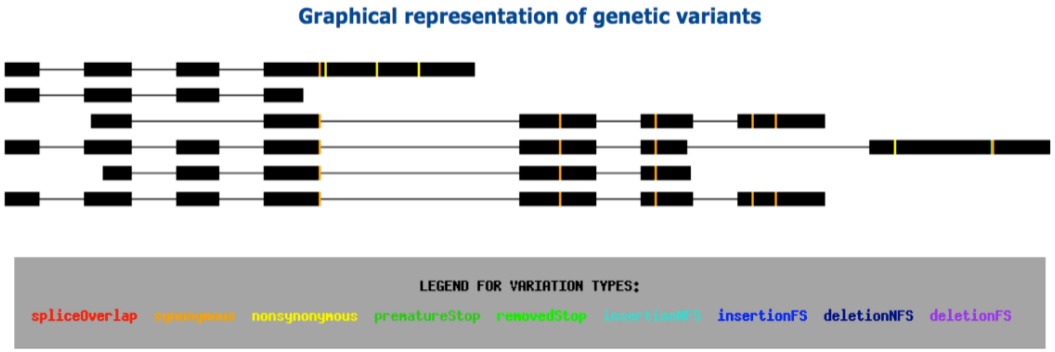
Balasubramanian S., et al. Genes Dev 25: 1-10, 2011

Variant Annotation Tool (VAT)

- Similar to VEP
- Input
  - Uses GENCODE (with option o CCDS & other annotations)
  - Overlaps with 1000G SNPs, MNPs, indels & SVs (other input VCFs possible)
- Output
  - Annotated VCFs
  - Graphical representations of functional impact on transcripts
- Access
  - Source freely available
  - Webserver
  - AWS cloud instance



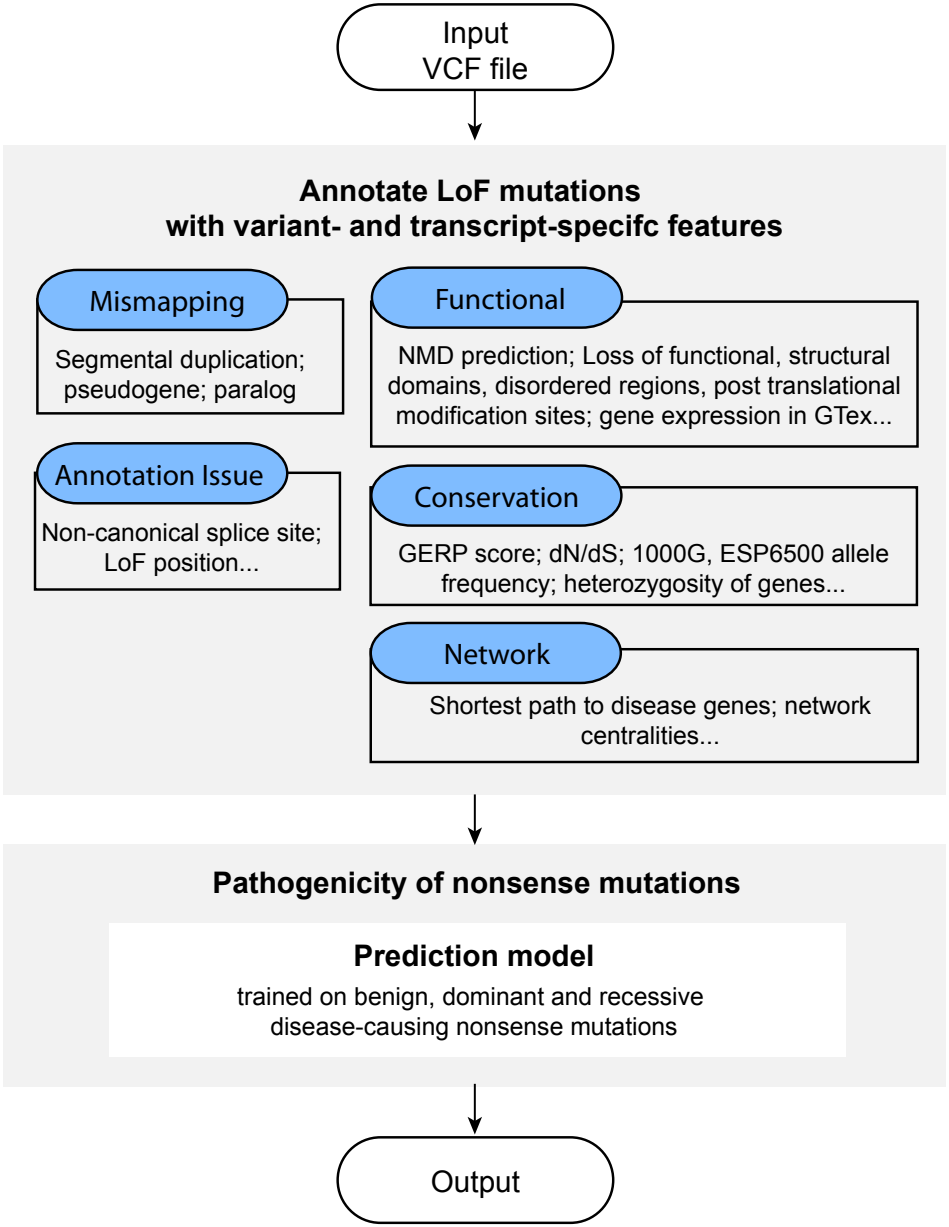
CLOUD APPLICATION



vat.gersteinlab.org

Habegger L., Balasubramanian S.*, et al. Bioinformatics, 2012*

ALoFT workflow



Annotate variant with Mismapping, Functional, Conservation and etc.

Predict disease-causing potentials

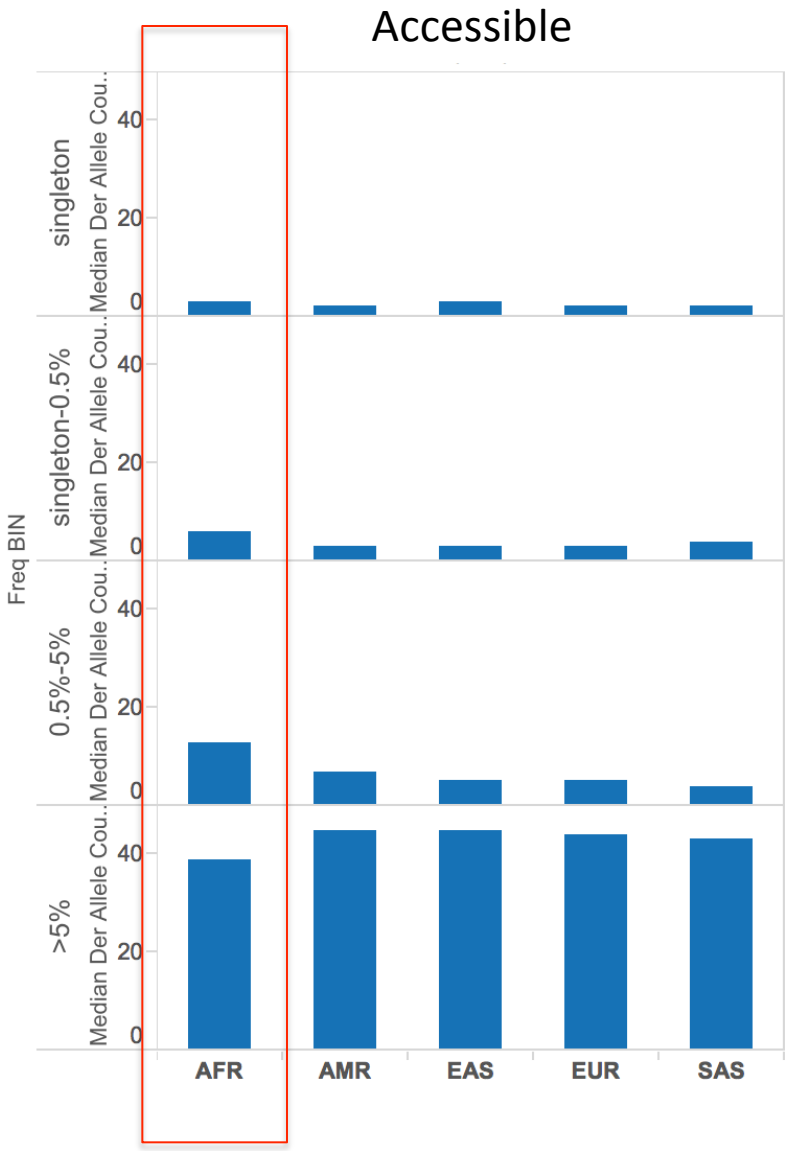
Yao Fu, Suganthi Balasubramanian

Median Autosomal Variant Sites Per Genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean Coverage	8.2		7.6		7.7		7.4		8.0	
	Var. Sites	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large Deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (LINE1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
NonSynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBS	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

[1000 Genomes Consortium (2015). *Nature*]

Stop-gain (median derived allele counts)



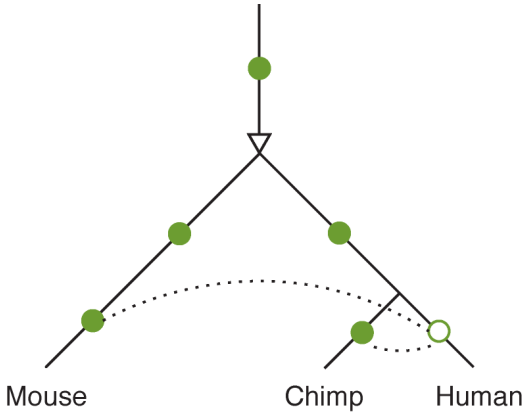
Aloft counts for 1000G Phase3

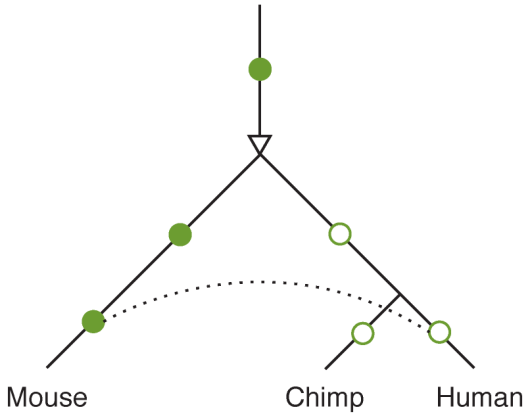
Considerably larger number than in the pilot

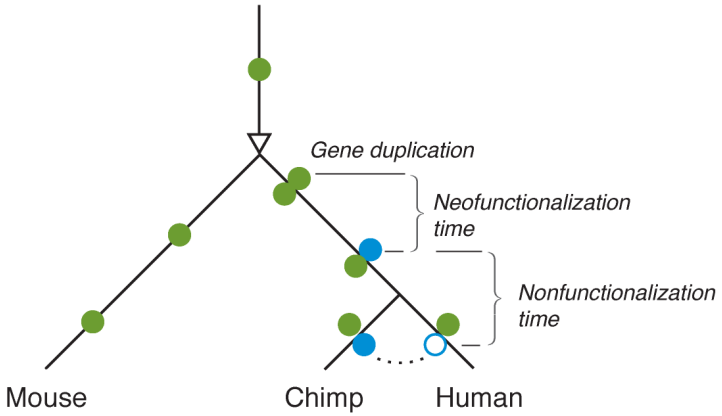
Largest amount for AFR, mostly rare

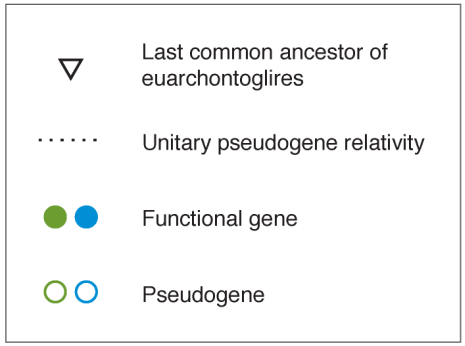
[1000 Genomes Consortium (2015). *Nature*]

Unitary Pseudogene Assignment Pipeline







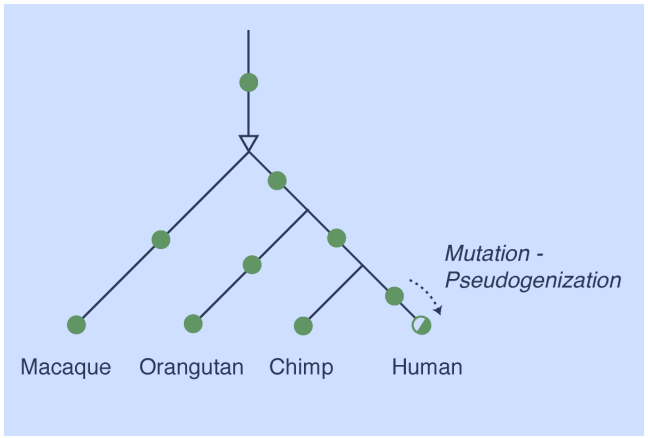


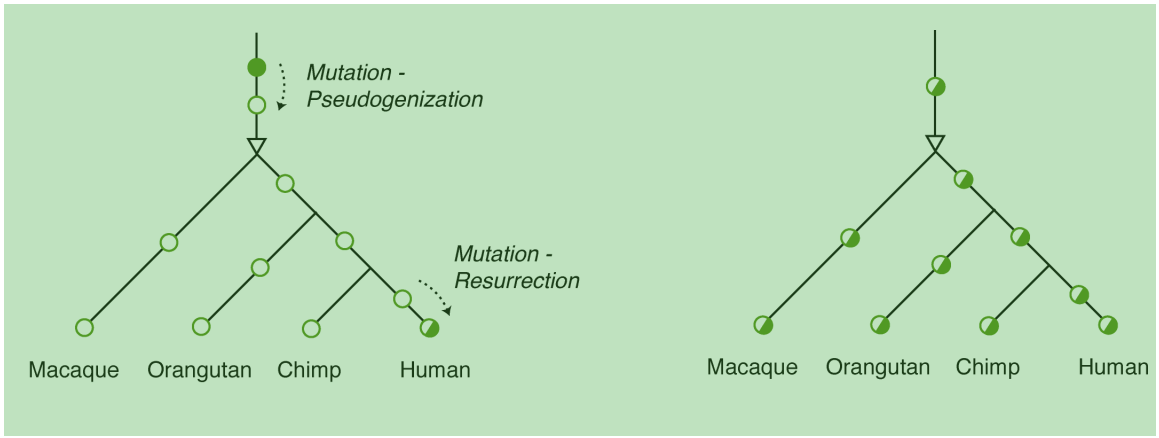
Gene **Polymorphic** **Pseudogene**

[Zhang et al. GenomeBiology ('10)]

11 Polymorphic pseudogenes

CDS-disrupted gene	GPR33	SERPINB11	TAAR9
Disruptive mutation [1]	Cga (R) → Tga	Gaa (E) → Taa	Aaa (K) → Taa
Allele frequency [4]			
Test statistic for HWE in the meta-population [5]	0.285 ($P = 0.867$)	8.659 ($P = 0.013$)	0.071 ($P = 0.965$)



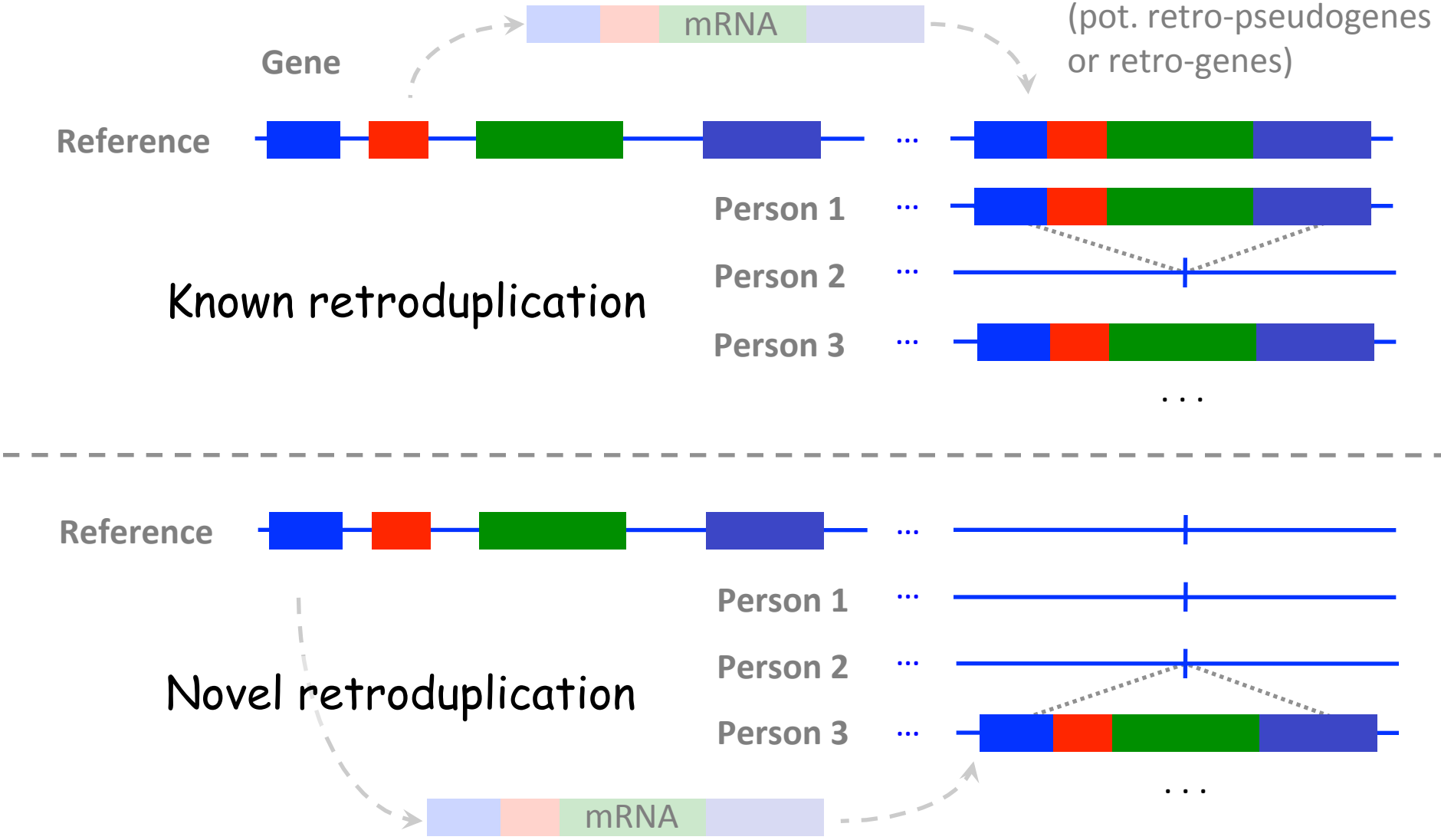


[Zhang et al. ('10) GenomeBiology]

The Blind Men & the Elephant: Relating Pseudogenes, LOF, Retrodup Variation & Personalized Annotation

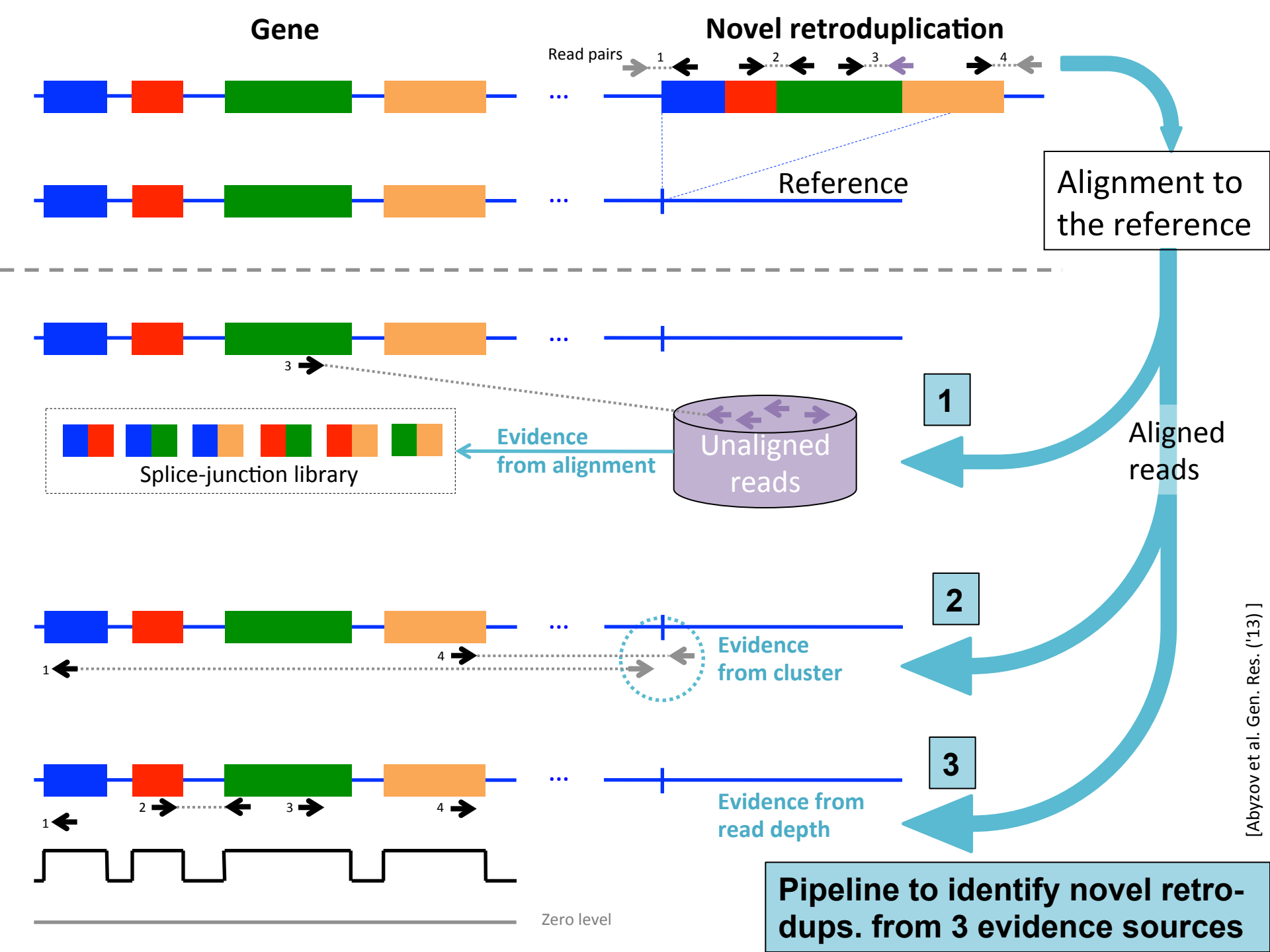
- LOFs
  - The spectrum: Genes=>LOFs=> Polymorphic Pseudogenes =>Fixed Pseudogenes
  - VAT & ALOFT
  - Large Diversity in 1000G-P3
- RDV
  - Further variation in polymorphic pseudogenes due to retroduplication
  - Absence of selection
- Personal Annotation
  - Personal Genomes
  - Best ref ?
  - No LOFs
- Mouse Strains
  - Putting it all together

Retroduplication variation (RDV)



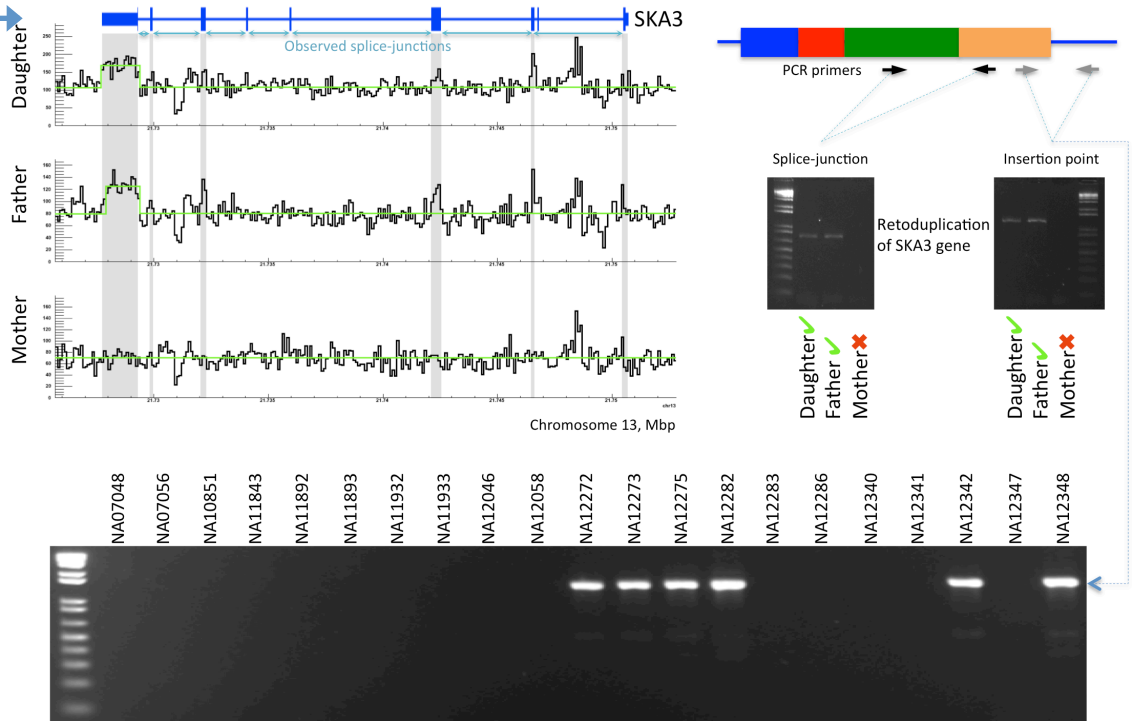
Schrider, D. R., **Navarro, F. C. P.**, Galante, P. A. F., Parmigiani, R. B., Camargo, A. A., Hahn, M. W., & de Souza, S. J. (2013). Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genetics,

Abyzov, A., Iskow, R., Gokcumen, O., Radke, D. W., Balasubramanian, S., Pei, B., et al. (2013). Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Research.


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)

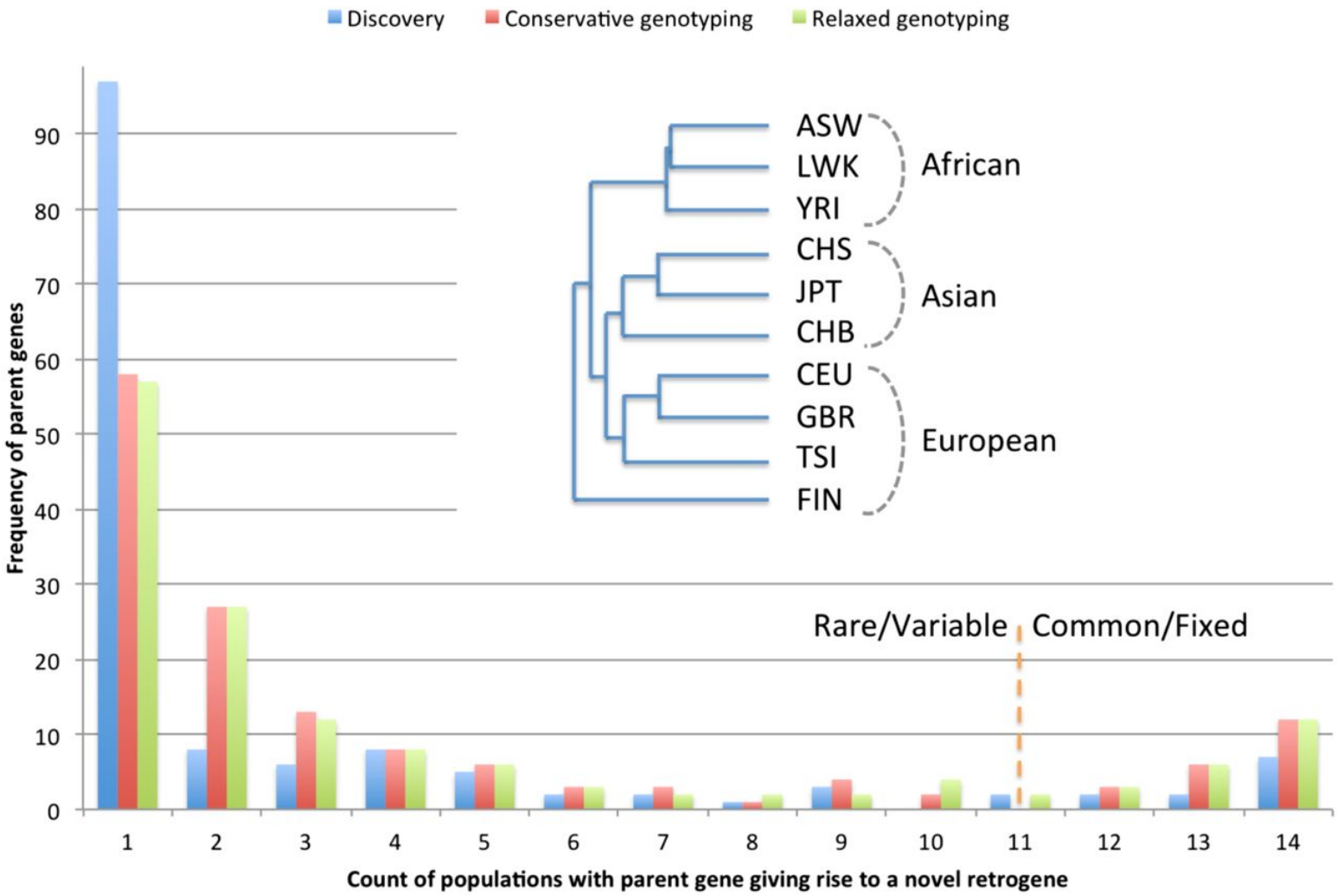
Parent gene with predicted novel retroduplication	Additional support			PCR validation
	Read depth support	Insertion point support	Found in Venter genome	
CDC27	Yes		Yes	UN
BCLAF1	Yes		Yes	UN
LAPTM4B	Yes	Yes		Yes
MTCH2				Yes
CBX3	Yes	Yes	Yes	Yes
TMEM66	Yes	Yes		Yes
TDG	Yes	Yes	Yes	Yes
BOD1				Yes
CACNA1B		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

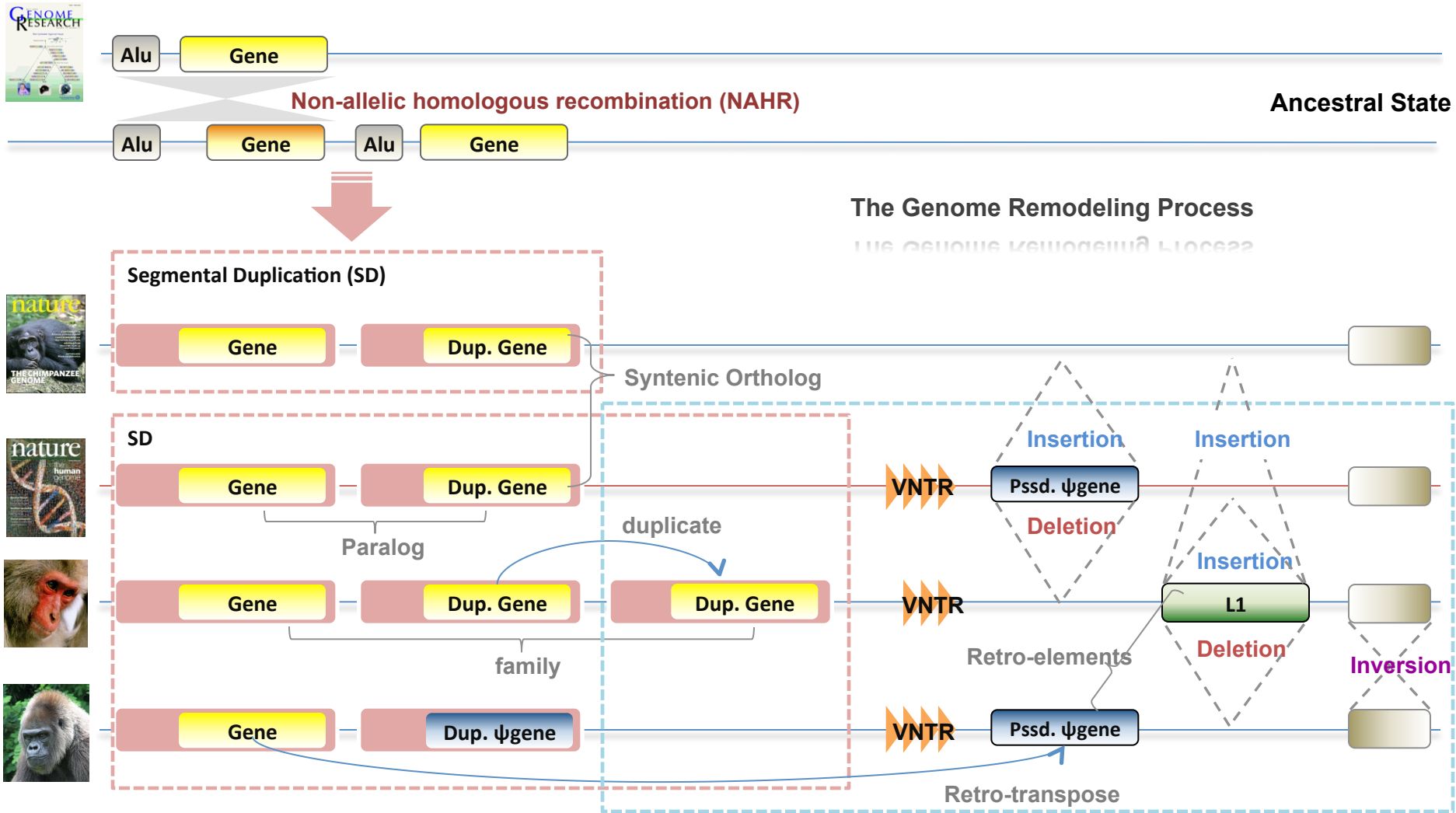
[Abyzov et al. Gen. Res. ('13)]

Frequency of novel retroduplications by populations.



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

Genomic Variation



Genomic Variation

"Polymorphic" Genes & Pseudogenes

The Blind Men & the Elephant: Relating Pseudogenes, LOF, Retrodup Variation & Personalized Annotation

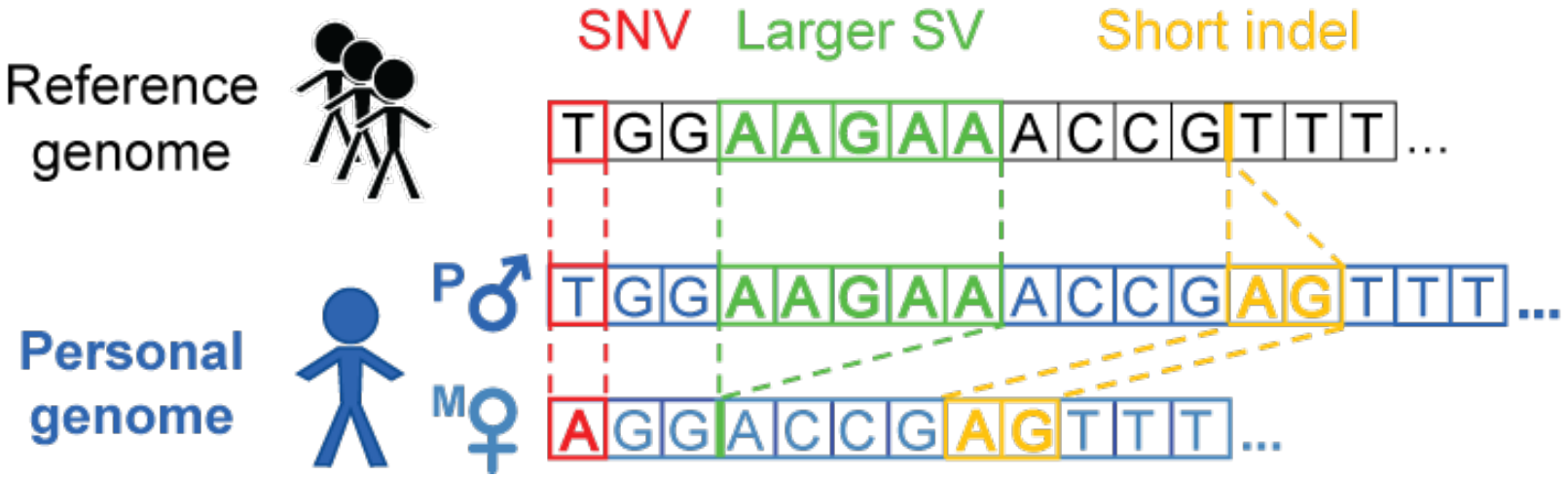
- LOFs
  - The spectrum: Genes=>LOFs=> Polymorphic Pseudogenes =>Fixed Pseudogenes
  - VAT & ALOFT
  - Large Diversity in 1000G-P3
- RDV
  - Further variation in polymorphic pseudogenes due to retroduplication
  - Absence of selection
- Personal Annotation
  - Personal Genomes
  - Best ref ?
  - No LOFs
- Mouse Strains
  - Putting it all together

Impact of Polymorphic Pseudogenes & LoF Events on Gene Sets

- What should be the reference gene set (& pseudogene set)?
  - Single individual
  - Ancestral individual
  - Current reference
  - Union of genes found in any individual
  - Intersection of genes found in everyone

[Balasubramanian et al., *Genes Dev.*, '11]

How we build a personal genome



reference, .fasta

+

personal variants, .vcf

→

personal diploid genome, .fasta

+

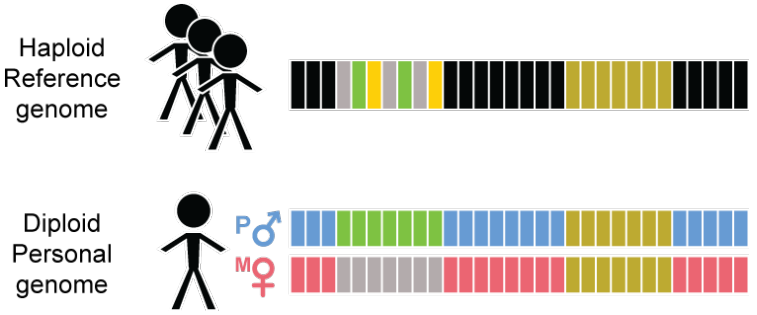
coordinate offset files, .chain

alleleseq.gersteinlab.org

Rozowsky *et al. Mol Syst Biol* (2011)

Why the personal genome (PG) should be a platform for functional genomics

1. **Diploid**
   --Ability to incorporate diverse variants of any size
   --exhibit phase information
2. **Scales easily with more samples & improve with development of sequencing technologies: longer reads and more accurate phase information**
3. **Demonstrably useful in functional genomic assay analyses**
   a) read alignment
   b) RNA-seq quantification
   c) allele-specific analyses



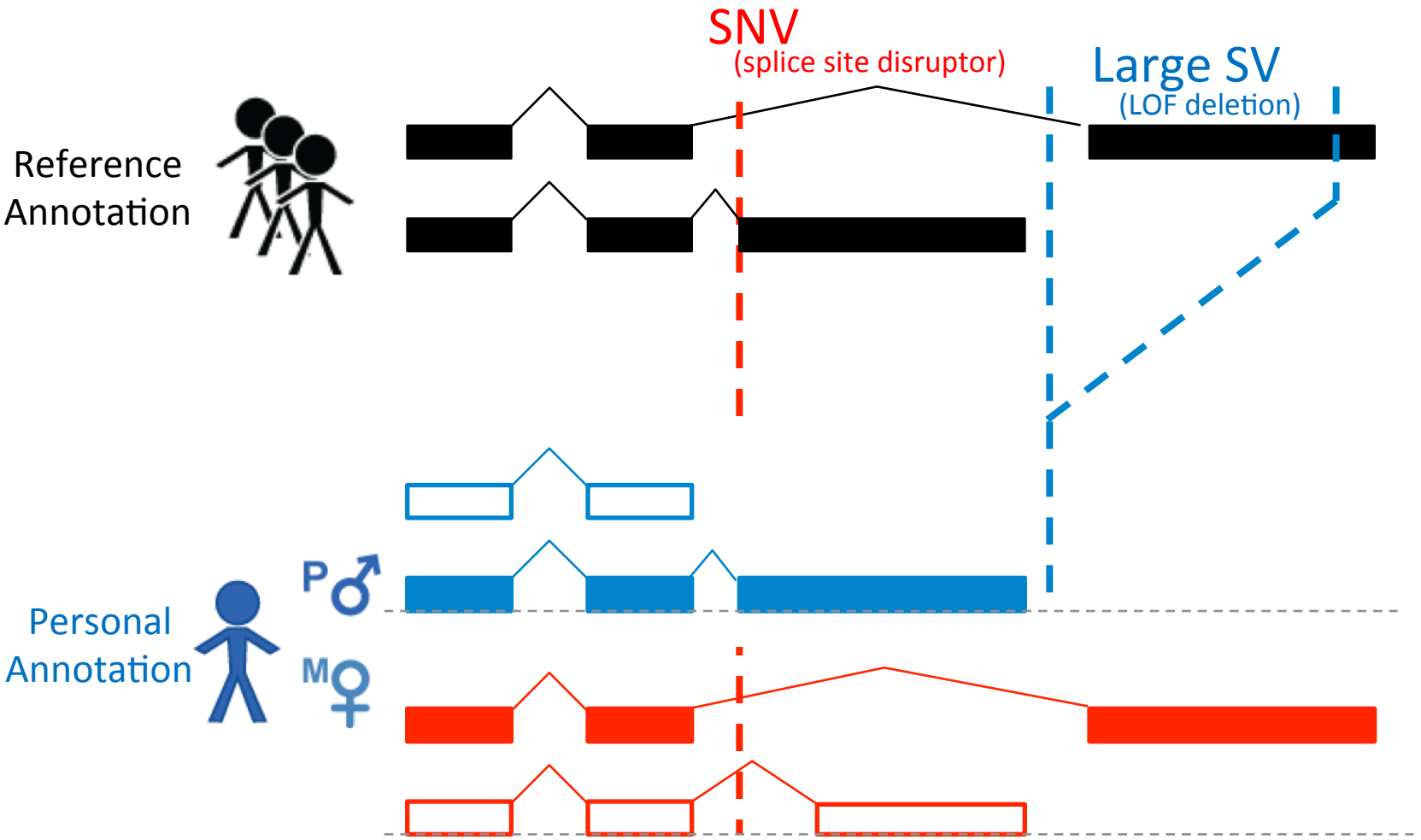
Evolution of NA12878 family of Personal Genomes

	Source	RefGen	Depth	Variants
1	1000 Genomes Project (1000GP) pilot (used for Rozowsky et al., ('11), alleleseq.gersteinlab.org)	hg18	60x	SNVs, indels, deletions (including 33 from fosmid sequencing)
2	GATK Best Practices v3 (UnifiedGenotype)	hg19	64x	SNVs, indels
3	GATK Best Practices v4 (HaplotypeCaller, PCR-free)	hg19	64x	SNVs, indels
4	1000GP Phase 3 SNVs-only	hg19	7.4x	SNVs
5	1000GP Phase 3 SNVs-indels	hg19	7.4x	SNVs, indels
6	1000GP Phase 3 SNVs-indels-SVs	hg19	7.4x	SNVs, indels, SVs
7	**1000GP Phase 3 SNVs-indels-SVs**	hg19	7.4x	SNVs, indels, SVs
8	**GIAB NA12878 pilot genome**	hg19	12x-190x	SNVs, indels, SVs

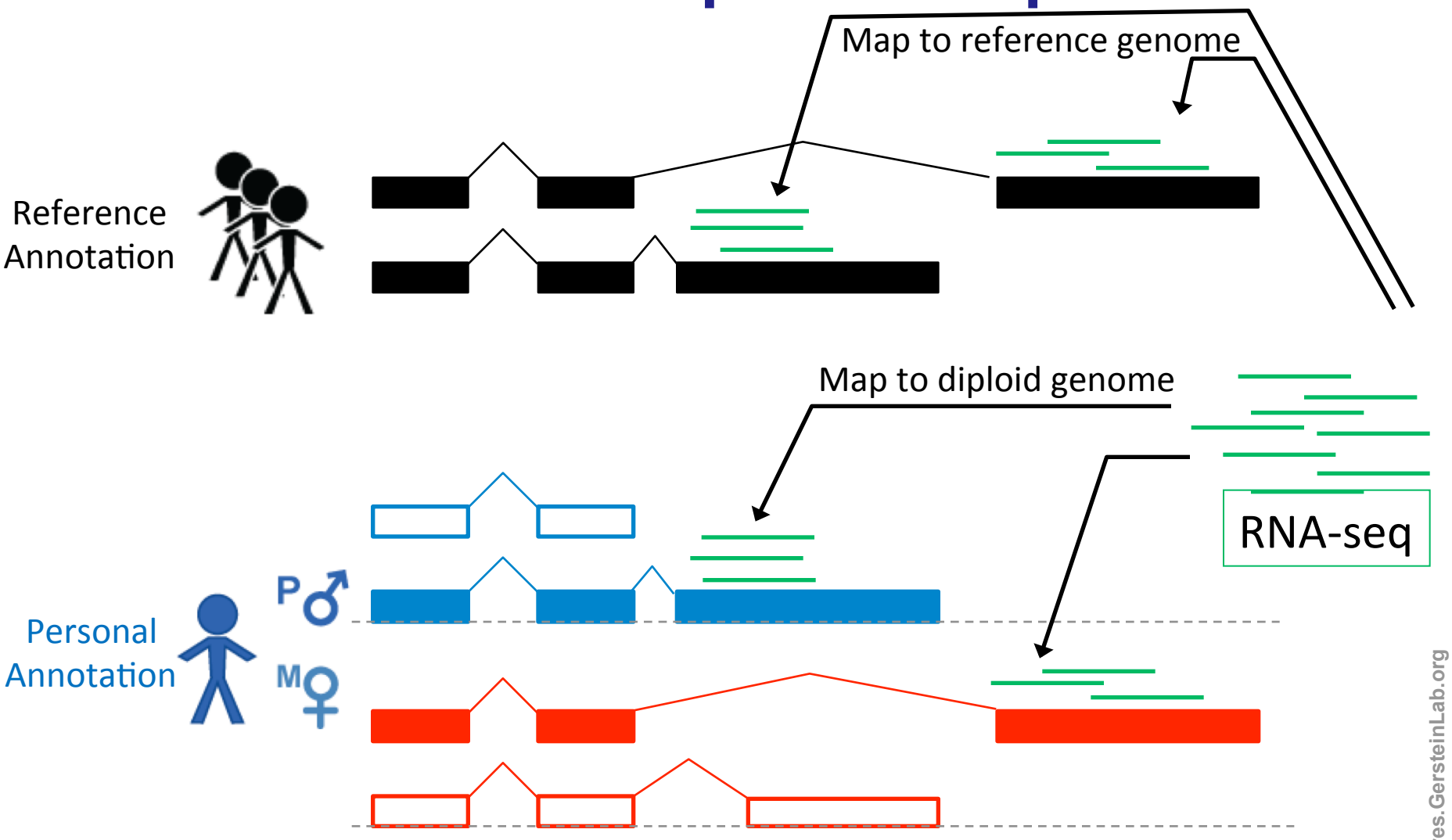
[7] Updated version of PG used in Sudmant et al, (Nature'15) [#6],
now with added complex SVs Pindel calls

[8] Incl. PacBio-based SV call set from GIAB
SNVs and Indels: High-confidence call set based on 11 WGS & 3 ES datasets (Zook et al, Nat Biotech '14);
SVs: Preliminary PacBio-based call set from
ftp://ftp- trace.ncbi.nih.gov/giab/ftp/data/NA12878/analysis/BCM_PacBio_PBHoney_15.8.24_09012015/

How we build a personal annotation



How to map RNA-seq



The Blind Men & the Elephant: Relating Pseudogenes, LOF, Retrodup Variation & Personalized Annotation

- LOFs
  - The spectrum: Genes=>LOFs=> Polymorphic Pseudogenes =>Fixed Pseudogenes
  - VAT & ALOFT
  - Large Diversity in 1000G-P3
- RDV
  - Further variation in polymorphic pseudogenes due to retroduplication
  - Absence of selection
- Personal Annotation
  - Personal Genomes
  - Best ref ?
  - No LOFs
- Mouse Strains
  - Putting it all together

Mus castaneus

Mus spretus

129S1/SvImJ
129S5/SvEv
129P2/OlaHsd
LP/J
DBA/2J
C3H/HeJ
CBA/J
A/J
BALB/cJ
AKR/J
NZO/lLtJ
C57BL/6NJ
NOD/ShiLtJ
WSB/Eij
CAST/Eij
PWK/PhJ
SPRET/EiJ
CAROLI/Eij
PAHARI/EiJ
Rat

0.009MYA
0.5MYA
1.5MYA
3MYA
6MYA
12MYA

Classical laboratory strains

Wild strains

Outgroup *Mus* species

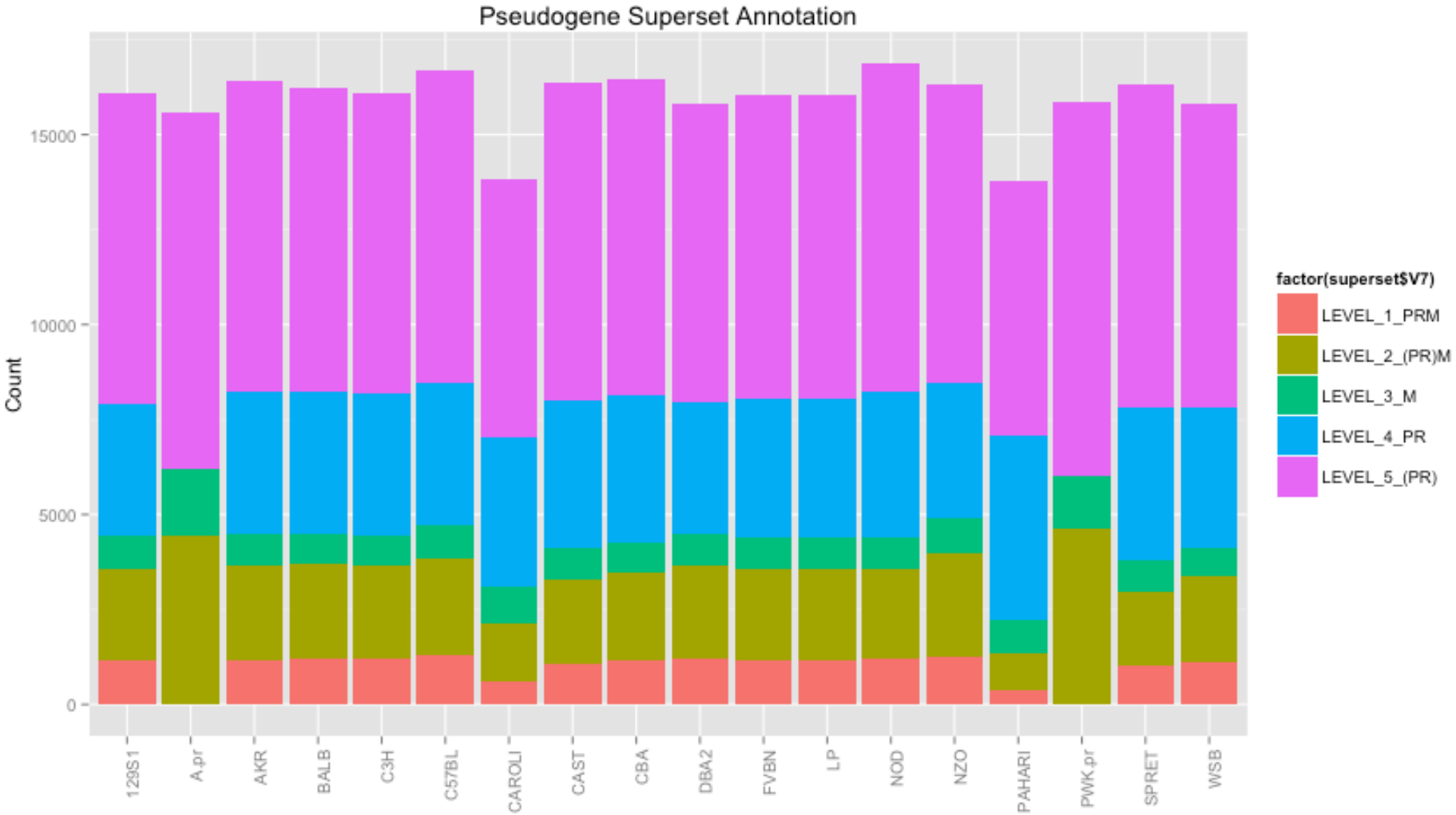
Rat

Mus pahari

Adapted from Ian Fiddes

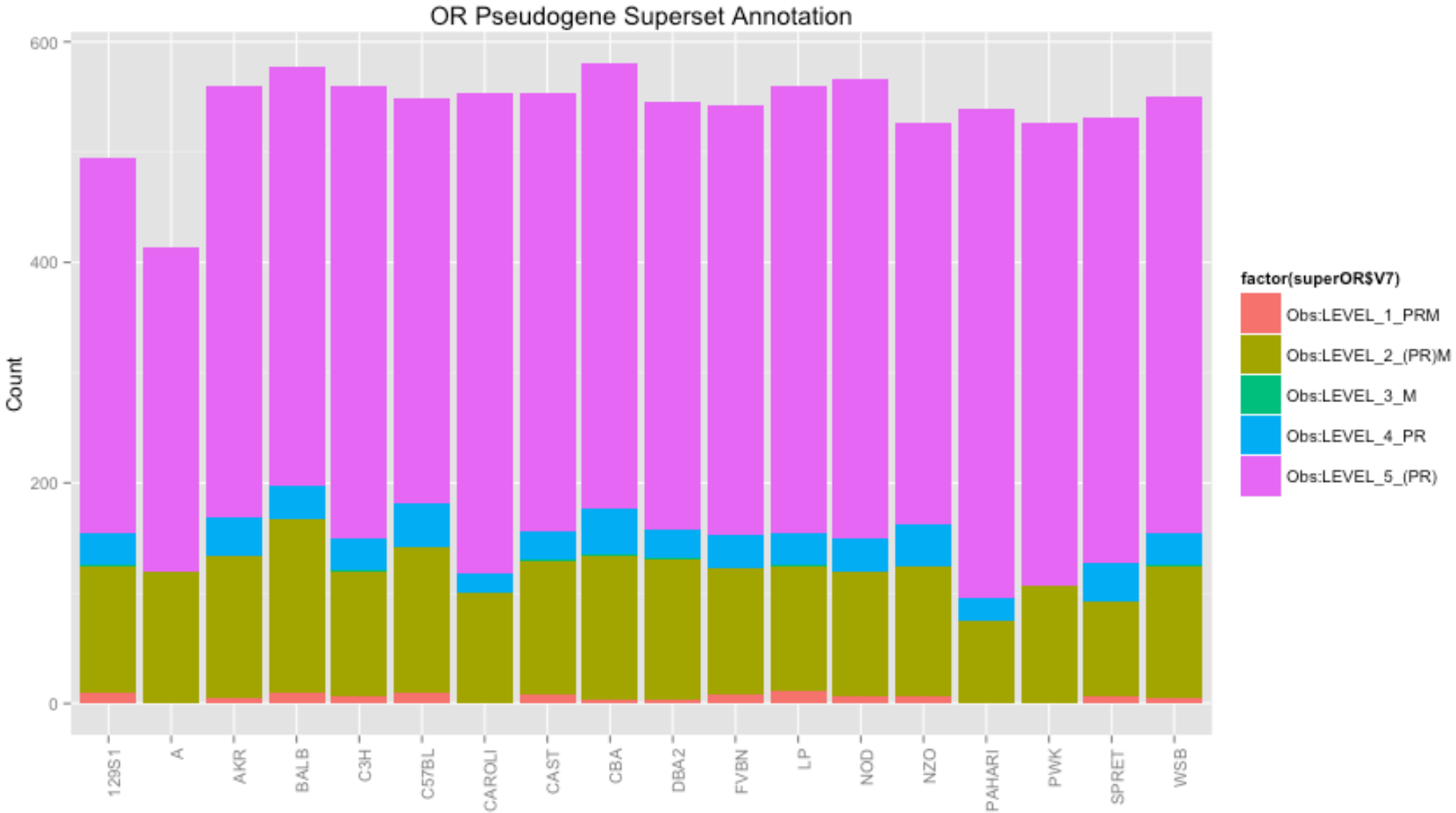
Pseudogene Annotation

- Using the union of RCPedia & Pseudopipe as well as overlapping with the liftover from manually annotated Mouse reference pseudogene sets.
  - LEVEL_1_RPM - pseudogenes identified by both pipelines and by manual annotation
  - LEVEL_2_(PR)M - pseudogenes identified by only 1 pipeline and by manual annotation
  - LEVEL_3_M- pseudogenes annotated only using the liftover of manually curated reference genome
  - LEVEL_4_PR - pseudogenes identified by both automatic pipelines
  - LEVEL_5_(PR) - pseudogenes identified by only an automatic annotation pipeline


Pseudogene Superset Annotation
Count
15000
10000
5000
0
129S1
A.pr
AKR
BALB
C3H
C57BL
CAROLI
CAST
CBA
DBA2
FVBN
LP
NOD
NZO
PAHARI
PWK.pr
SPRET
WSB
factor(superset$V7)
LEVEL_1_PRM
LEVEL_2_(PR)M
LEVEL_3_M
LEVEL_4_PR
LEVEL_5_(PR)

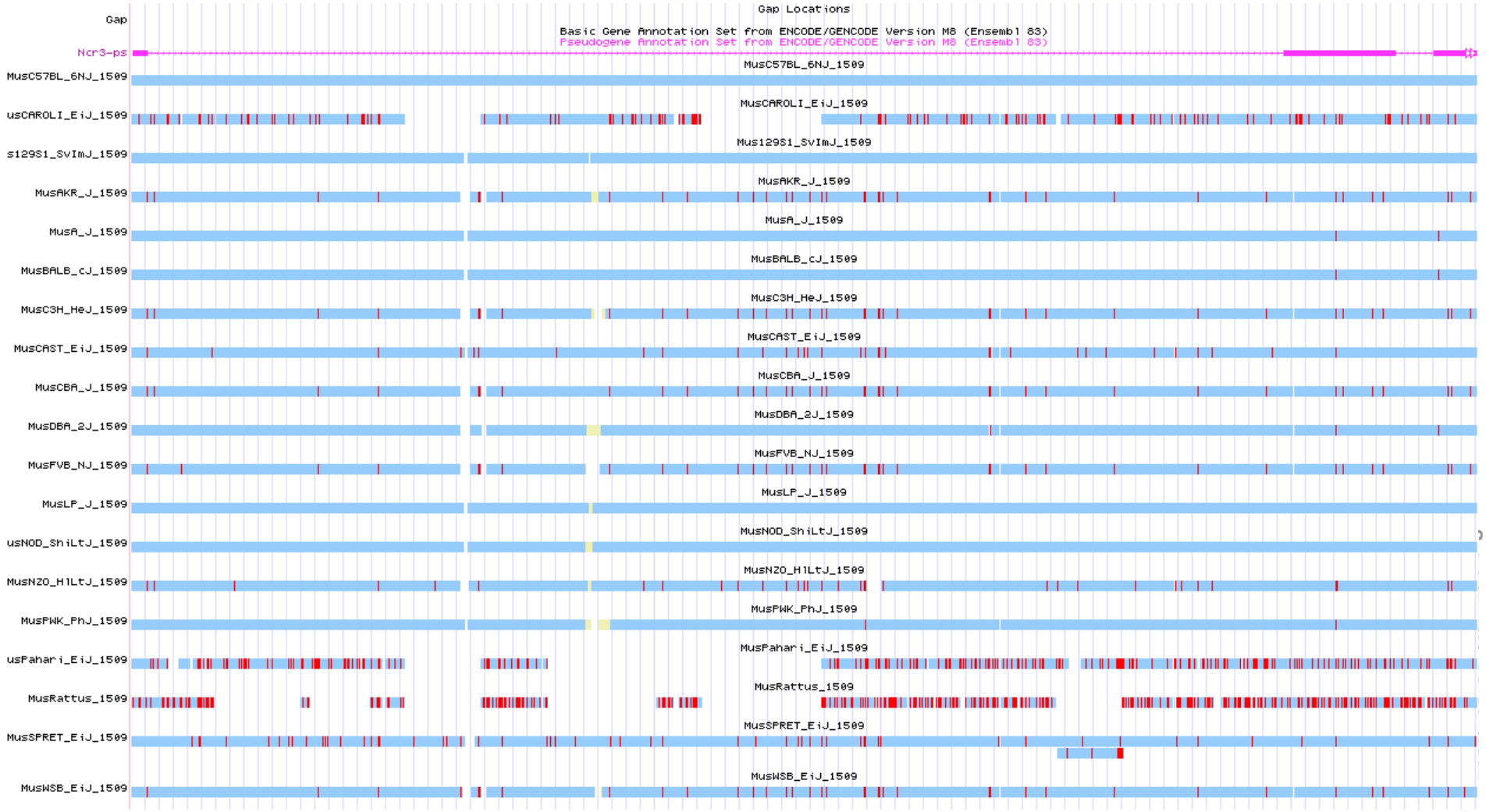
Case study: Olfactory Receptors

- 1294 transcripts & 1106 OR genes in mouse reference genome



ENSMUST00000134687 - NCR3 pseudogene

- Pseudogene in reference genome and laboratory strains
- Reversed stop codon mutations in CAROLI



```
Caroli  1     CATGGCCAAGGTGCTTCTGGTCATCTTCATCATGGTTTATCCAGG  45
              ||||||||||||||| ||||||||||||||||||||||||||||
Refer   1     CATGGCCAAGGTGCTCCTGGTCATCTTCATCATGGTTTATCCAGG  45

Caroli  3751        ATCCTGTGCTCTCTGGGTGTCTCAGCCCCCTGAGACTCCTGTGCAGAAGGACACT  3805
                    ||||||||||||||| |||||||||||||||||| || ||||||| |||  |||
Refer   46          ATCCTGTGCTCTCTGAGTGTCTCAGCCCCCTGAGATTCGTGTGCAGGAGGGGACT  100
Caroli  3806  ACTGCCTCCCTGCCCTGCCCCTTCAATGCCATCCGAGGAAAACCGGCCACTGGCTCTGTC  3865
              ||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Refer   101   ACTGCCTCCCTGCCCTGCTCCTTCAATGCCATCCGAGGAAAACCGGCCACTGGCTCTGTC  160
Caroli  3866  ACATGGTACCAAGACAAAGTGACCCTGGGGATGGAGTTGAGCAACGTGACTCCGGGGTTC  3925
              |||||  |||||||||||||||||||||||||||||||||||| ||||||||  ||||||
Refer   161   ACATGATACCAAGACAAAGTGACCCTGGGGATGGAGTTGAGCAAAGTGACTCCAGGGTTC  220
Caroli  3926  GGAGGCCGCTTGGTCTCCTTTTCTGTTTCTCAGTTCATTAGGGACCACAAGGCAGGGCTG  3985
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Refer   221   AGAGGCCGCTTGGTCTCCTTTTCTGTTTCTCAGTTCATTAGGGACCACAAGGCAGGGCTG  280
Caroli  3986  CTCATACAGGACACCCAAAGCTATGATGCTGGAATCTATGTGTGCAGGGTGGAGGTGCTA  4045
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Refer   281   CTCATACAGGACACCCAAAGCTATGATGCTGGAATCTATGTGTGCAGGGTGGAGGTGCTA  340
Caroli  4046  GGCCTGGGCGTCAGAACGGGAGATGGGACTTGGCTGCTGGTGGAGA  4091
              ||||||||||||||| ||||||  ||||| |||||||||||||||
Refer   341   GGCCTGGGCGTCAGAATGGGAGACAGGACTCGGCTGCTGGTGGAGA  386

Caroli  4209              CCTCCTCAGCAAGCCTCCAACACAGAGCAAGAGGCTCACACAAGTAT  4255
                          |||||||||||||||||||||||||||||||||||||||||||| |
Refer   387               CCTCCTCAGCAAGCCTCCAACACAGAGCAAGAGGCTCACACAAGTCT  433
Caroli  4256  CCTCCTCCGGGCTGGATTCTATGCCCTCAGCTTTCTCTCTGTGGCCACGGGCAGTACCAT  4315
              |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||
Refer   434   CCTCCTCCGGGCTGGATTCTACGCCCTCAGCTTTCTCTCTGTGGCCACGGGCAGTACCAT  493
Caroli  4316  CTATTACCAGGGCAAATGTGAGTTGCTGAGCA  4347
              |||||||||||||||||||||||||||||||
Refer   494   CTATTACCAGGGCAAATGTGAGTTGCTGAGCA  525
```

Reference

MAKVLLVIFIMVYPGSCAL-VSQPPEIRVQEGTTASLPCSFNAIRGKPATGSVT-YQDKV
TLGMELSKVTPGFRGRLVSFSVSQFIRDHKAGLLIQDTQSYDAGIYVCRVEVLGLGVRMG
DRTRLLVETSSASLQHRARGSHKSPPPGWILRPQLSLCGHGQYHLLPGQM-VAE

Caroli

MAKVLLVIFIMVYPGQDPVLSGCLSPLRLLCRRTLLPPCPAPSMPSEENRPLALSHGTKT
K-PWGWS-AT-LRGSEAAWSPFLFLSSLGTTRQGCSYRTPKAMMLESMCAGWRC-AWASE
REMGLGCWWRLL-GLLSKPPTQSKRLTQVSSSGLDSMPSAFSLWPRAVPSITRANVSC-A

The Blind Men & the Elephant: Relating Pseudogenes, LOF, Retrodup Variation & Personalized Annotation

- LOFs
  - The spectrum: Genes=>LOFs=> Polymorphic Pseudogenes =>Fixed Pseudogenes
  - VAT & ALOFT
  - Large Diversity in 1000G-P3
- RDV
  - Further variation in polymorphic pseudogenes due to retroduplication
  - Absence of selection
- Personal Annotation
  - Personal Genomes
  - Best ref ?
  - No LOFs
- Mouse Strains
  - Putting it all together

The Blind Men & the Elephant: Relating Pseudogenes, LOF, Retrodup Variation & Personalized Annotation

- LOFs
  - The spectrum:
    Genes=>LOFs=>
    Polymorphic Pseudogenes
    =>Fixed Pseudogenes
  - VAT & ALOFT
  - Large Diversity in 1000G-P3
- RDV
  - Further variation in polymorphic pseudogenes due to retroduplication
  - Absence of selection
- Personal Annotation
  - Personal Genomes
  - Best ref ?
  - No LOFs
- Mouse Strains
  - Putting it all together

Acknowledgments

Personal Genomes - J Chen, **J Rozowsky, TR Galeev**, A Harmanci, R Kitchen, J Bedford, A Abyzov, Y Kong, L Regan

1000G Phase 3 LOFs - Yuan Chen, **Suganthi Balasubramanian**, Yao Fu, Donghoon Kim, Vincenza Colonna, Heiko Horn, Jakob Berg Jespersen, Kasper Lage, Xiangqun Zheng-Bradley, **Fiona Cunningham**, Ian Dunham, Paul Flicek, Ekta Khurana, Daniel Zerbino, Laura Clarke, **Chris Tyler-Smith, Yali Xue**

Gencode Pgenes - **C Sisu, B Pei,** J Leng, **A Frankish**, Y Zhang, S Balasubramanian, R Harte, D Wang, M Rutenberg-Schoenberg, W Clark, **M Diekhans**, J Rozowsky, **T Hubbard, J Harrow**

RDVs - **A Abyzov,** R Iskow, O Gokcumen, DW Radke, S Balasubramanian, B Pei, L Habegger, The 1000 Genomes Project Consortium, C Lee

VAT+ALOFT - **L Habegger, S Balasubramanian**, DZ Chen, E Khurana, A Sboner, A Harmanci, J Rozowsky, D Clarke, M Snyder + **Y Fu** & D MacArthur

Ref. Gene Set - S Balasubramanian, L Habegger, A Frankish, DG MacArthur, R Harte, C Tyler-Smith, J Harrow

Unitary Pgenes - **ZD Zhang**, A Frankish, T Hunt, J Harrow

Current Team – **C Sisu, F Navarro**

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2016.
  - 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