1000 Genomes Phase3 SV Analyses
(Group Meeting Report)

Yan Zhang

Gerstein Lab

The 1000 Genomes Project SV Group & Analysis Group

9/24/2015

Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Human Genetic Variations

- Single nucleotide variants (SNVs)
- Short insertions and deletions (Indels)
- Structural variations (SVs)
  - Sequence variations of at least 50bp in size

[1] Weischenfeldt J, et al. Nat Rev Genet, 2013.
[2] 1000GP Phase3 SV paper. Submitted to Nature, 2015.

A Typical Genome

- A typical genome differs from the reference genome at 4.09 – 5.02 million sites.
- The typical genome contains 2,100 – 2,500 SVs, covering ~20 million bases.
- A typical genome contains 149 – 182 sites with protein truncating variants, 10 – 12 thousand sites with peptide sequence altering variants, and 459 – 565 thousand variant sites overlapping regulatory regions.

[3] 1000GP Consortium. Submitted to Nature, 2015.

Structural Variations (SVs)

- SVs make up the majority of varying nucleotides among humans.

- More base pairs are altered as a result of SVs, than of single-nucleotide variations.
  - On the haploid reference assembly, a median of 8.9 Mbp are affected by SVs, while 3.6 Mbp affected by SNPs.

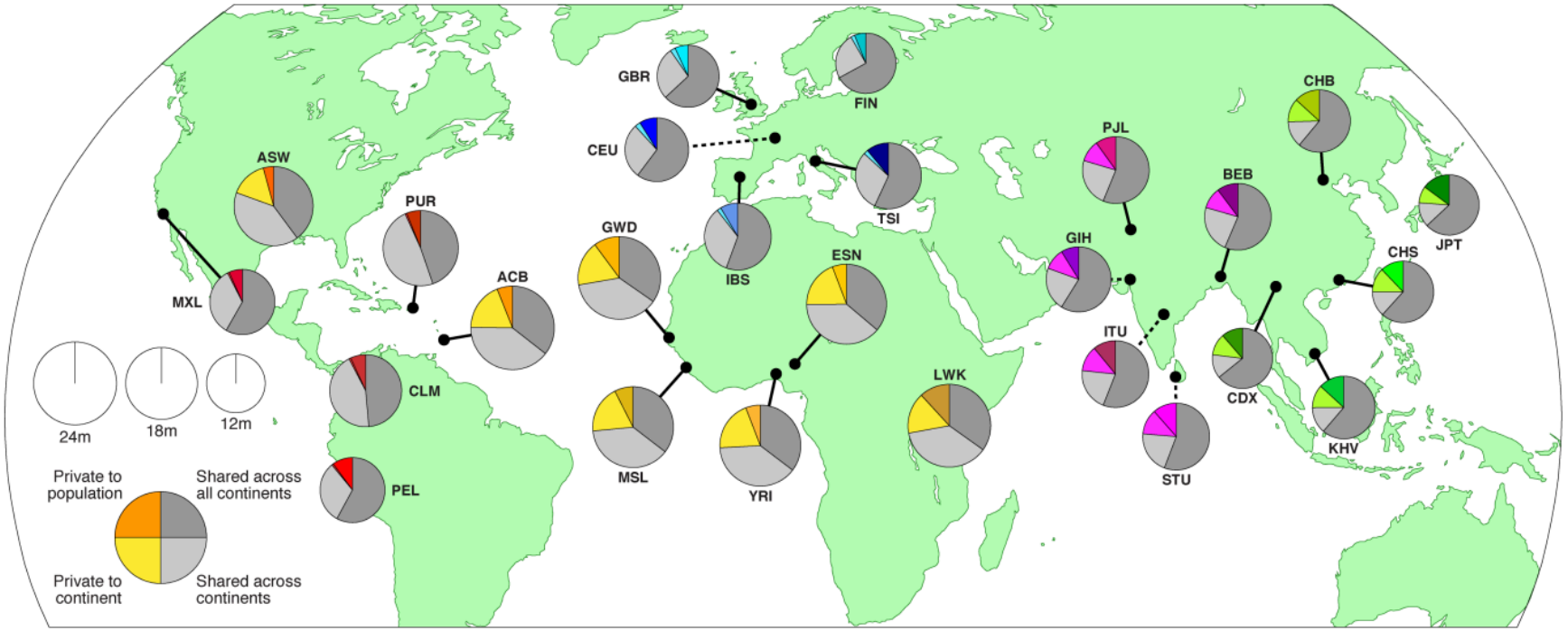
[1] Weischenfeldt J, et al. Nat Rev Genet, 2013.
[2] 1000GP Phase3 SV paper. Submitted to Nature, 2015.

Objective of 1000GP SV Analysis

- Discover and genotype major classes of SVs
- Enable integration of these SVs into phased reference panel for population and genetic studies

[2] 1000GP Phase3 SV paper. Submitted to Nature, 2015.

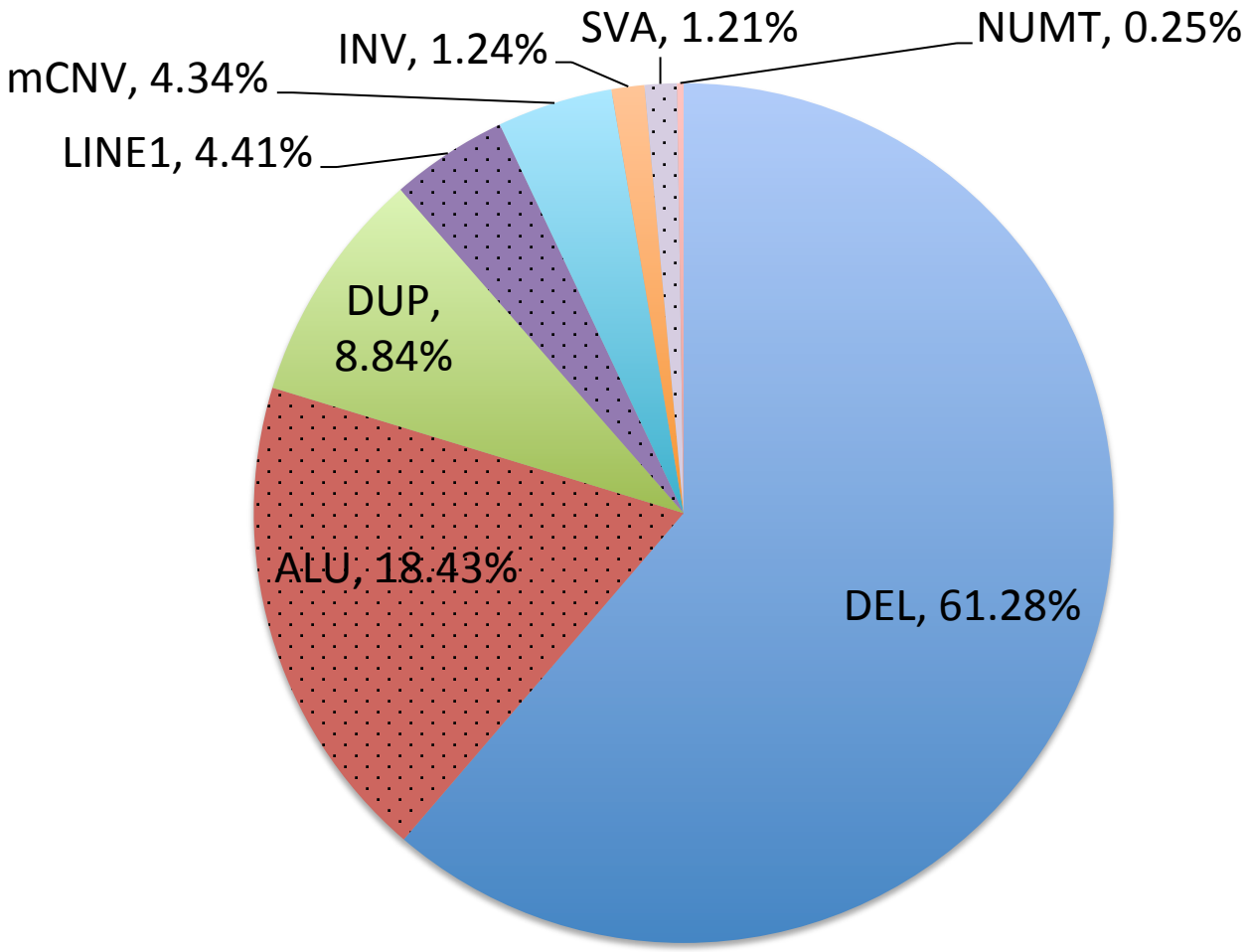
Summary Statistics of 1000GP SV Phase3



- 68,818 SVs
- 2,504 unrelated individuals
- 26 populations
- 37,250 SVs with resolved breakpoints

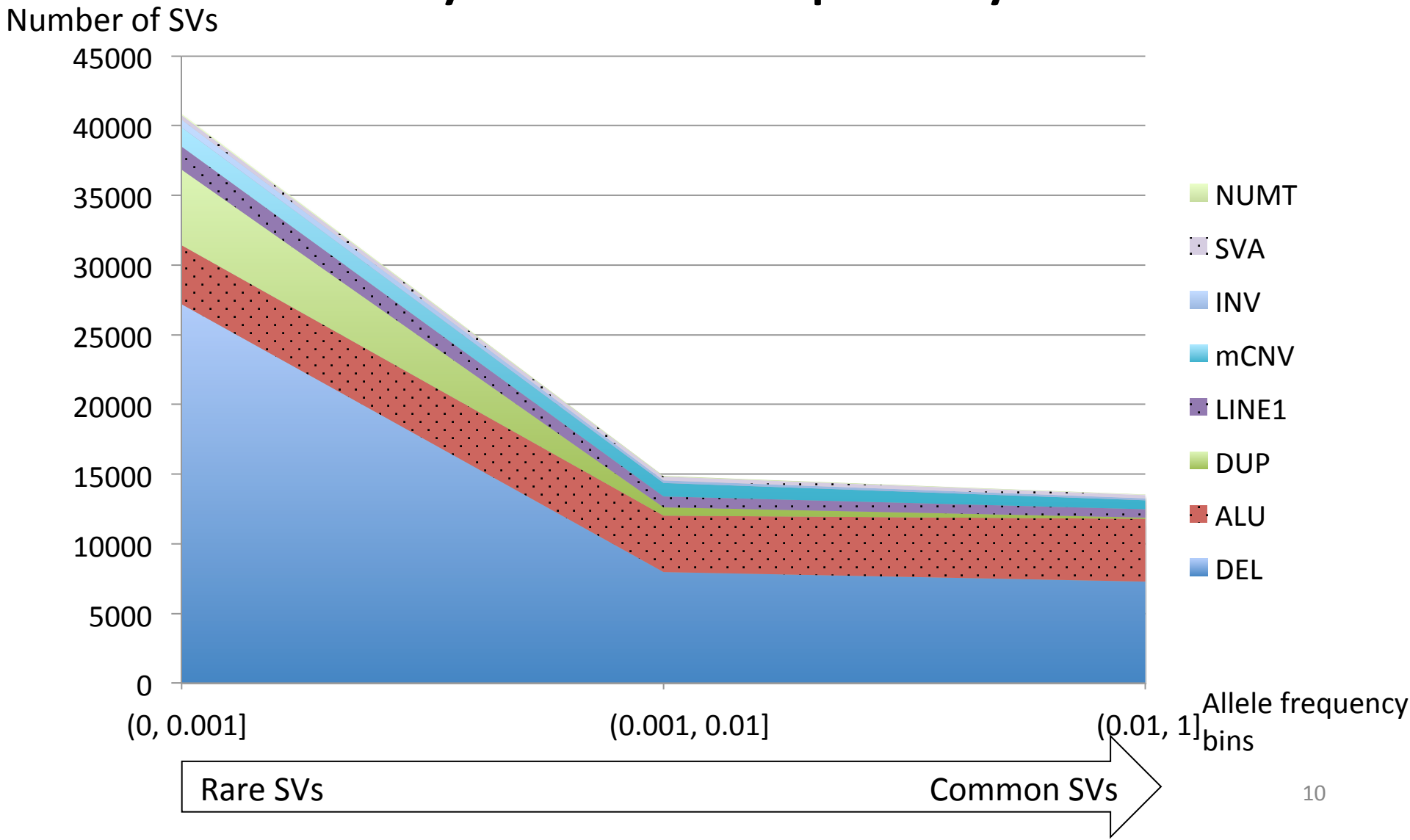
[2] 1000GP Phase3 SV paper. Submitted to Nature, 2015.
[3] 1000GP Consortium. Submitted to Nature, 2015.

Distribution of Different SVs in Normal Human Populations



Total ~70K SVs from over 2,500 normal individuals (the 1000 Genomes Project)

Distribution of Different SVs Stratified by Allele Frequency



Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

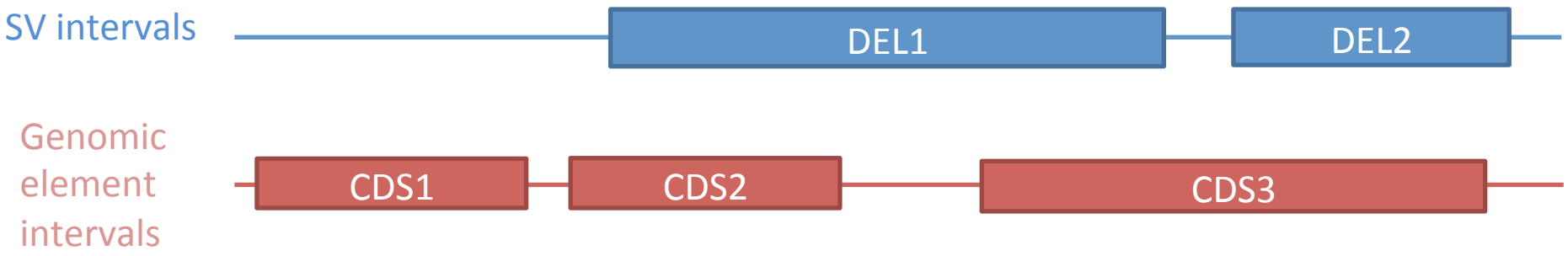
Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Enrichment Overlap Analysis

- Measure overlap between SVs and genomic elements
- Test statistical significance of the overlap

Measure of Overlap between SVs and Genomic Elements



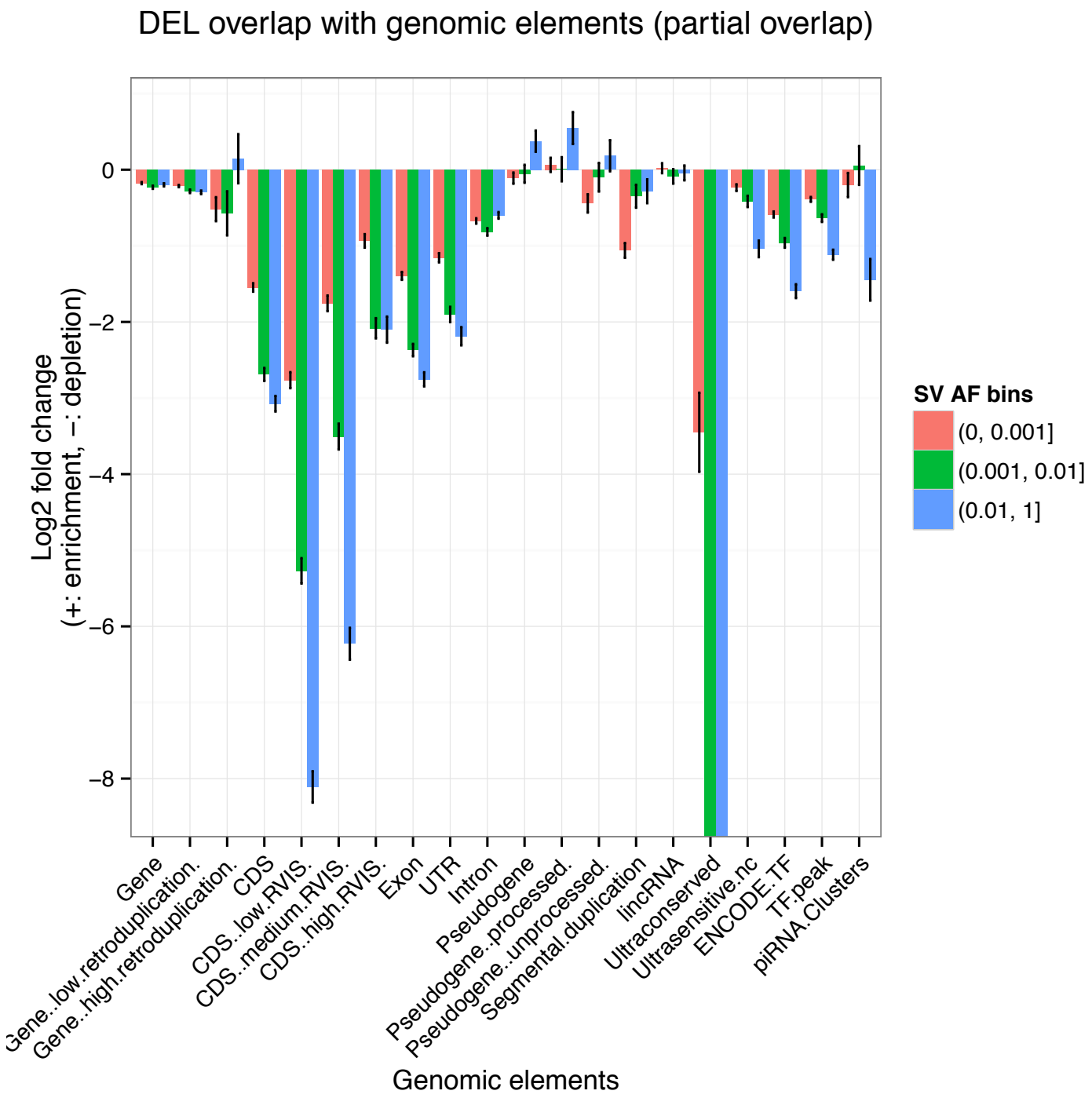
Partial overlap statistic:

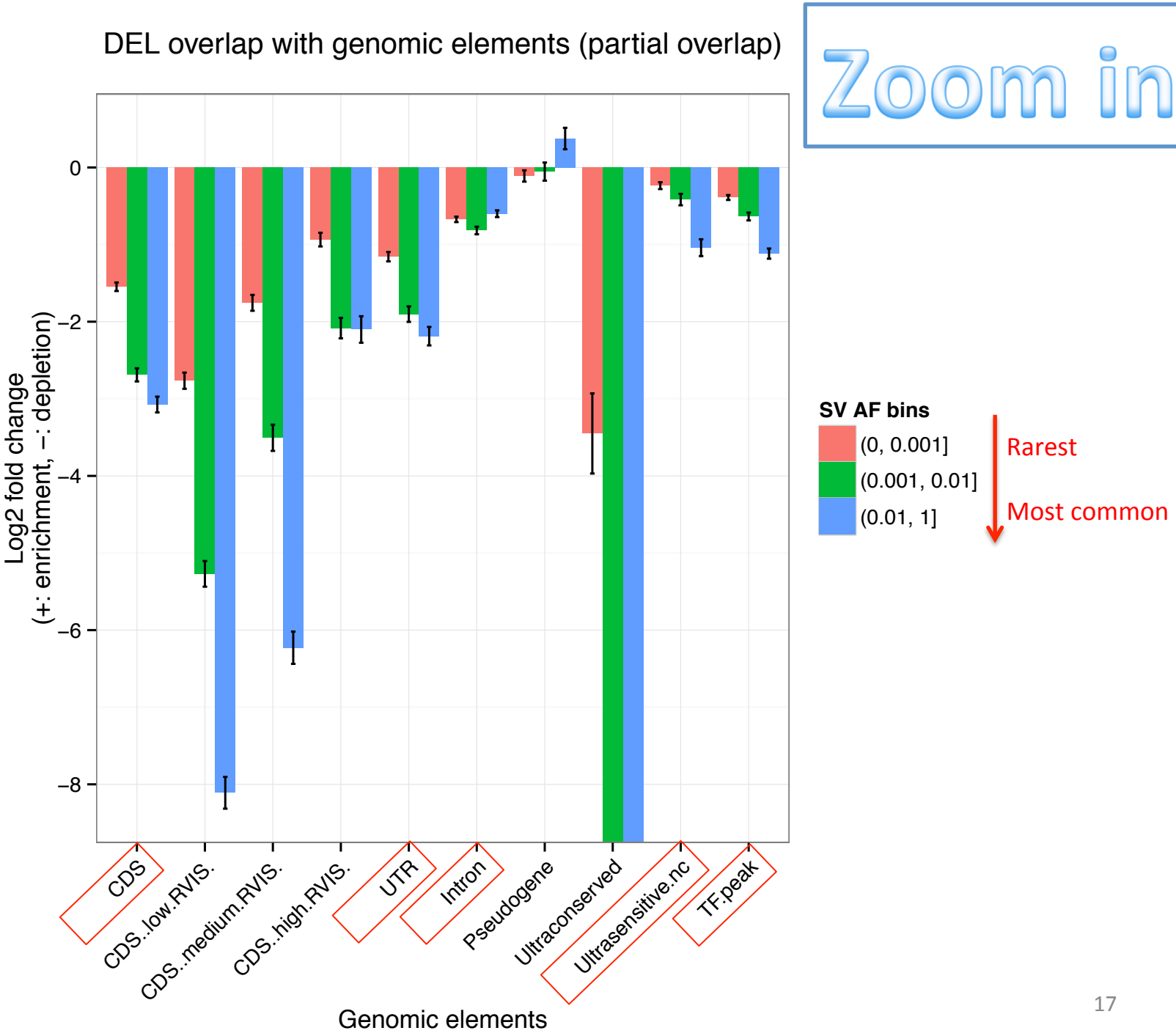
Count the number of genomic elements that have at least 1 bp overlap with SVs.

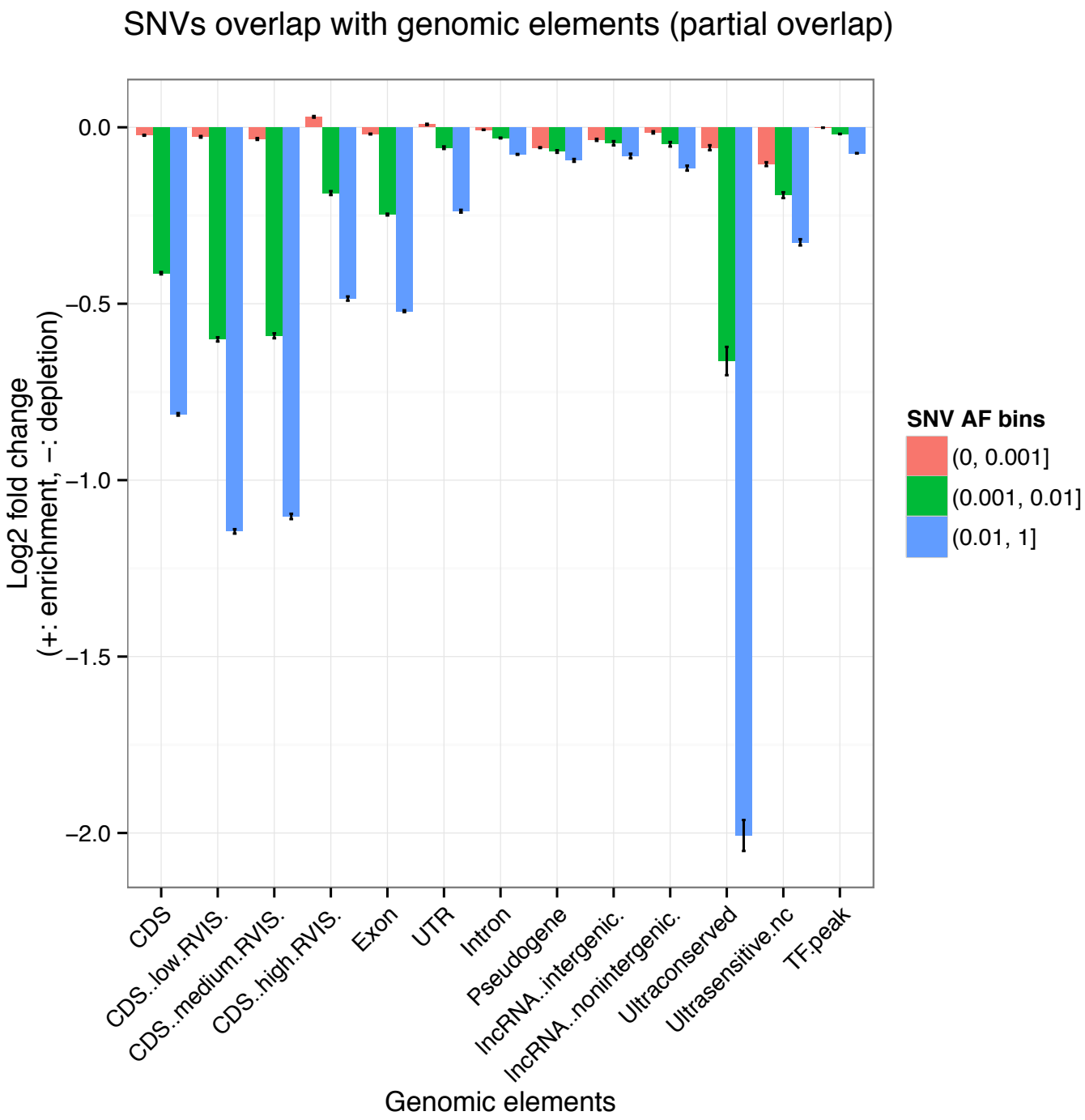
Permutation Tests

- Permutation scheme
  - Randomly shuffle SV locations while maintaining the local structure
    - Same number of SVs, same length distribution
    - Shuffled SVs still locate on the same chromosome
    - Hg19 gap removed
  - Log2 fold change and empirical p-values

- Datasets
  - 8 types of SVs from the 1000 Genomes Project
  - 20 types of genomic elements from GENCODE, ENCODE, and other literature


DEL overlap with genomic elements (partial overlap)
Log2 fold change
(+: enrichment, −: depletion)
0
−2
−4
−6
−8
Gene
Gene..low.retroduplication.
Gene..high.retroduplication.
CDS
CDS..low.RVIS.
CDS..medium.RVIS.
CDS..high.RVIS.
Exon
UTR
Intron
Pseudogene
Pseudogene..processed.
Pseudogene..unprocessed.
Segmental.duplication
lincRNA
Ultraconserved
Ultrasensitive.nc
ENCODE.TF
TF.peak
piRNA.Clusters
Genomic elements
SV AF bins
(0, 0.001]
(0.001, 0.01]
(0.01, 1]


DEL overlap with genomic elements (partial overlap)
Zoom in
Log2 fold change
(+: enrichment, −: depletion)
0
−2
−4
−6
−8
CDS
CDS..low.RVIS.
CDS..medium.RVIS.
CDS..high.RVIS.
UTR
Intron
Pseudogene
Ultraconserved
Ultrasensitive.nc
TF.peak
Genomic elements
SV AF bins
(0, 0.001]
(0.001, 0.01]
(0.01, 1]
Rarest
Most common


SNVs overlap with genomic elements (partial overlap)
Log2 fold change
(+: enrichment, −: depletion)
0.0
−0.5
−1.0
−1.5
−2.0
CDS
CDS..low.RVIS.
CDS..medium.RVIS.
CDS..high.RVIS.
Exon
UTR
Intron
Pseudogene
lncRNA..intergenic.
lncRNA..nonintergenic.
Ultraconserved
Ultrasensitive.nc
TF.peak
Genomic elements
SNV AF bins
(0, 0.001]
(0.001, 0.01]
(0.01, 1]

Conclusion

- Important biologically functional genomic elements are depleted with DELs.
- CDS regions under strong purifying selection are most depleted.
- This conclusion applies to other SV types; but less significant than DELs.
- We observed similar trend for SNVs binned by allele frequency.

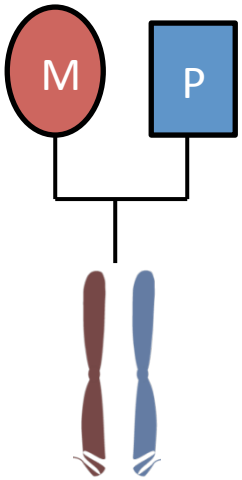
Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

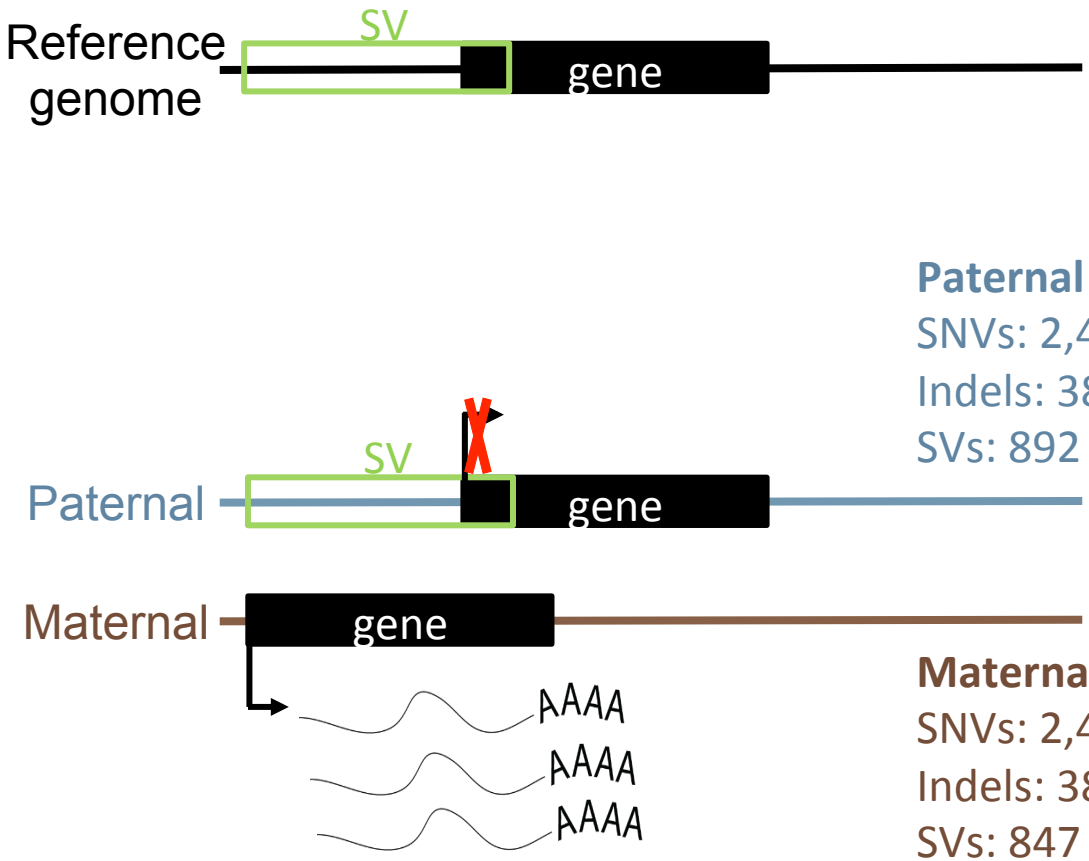
Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Personal Diploid Genome and Effects on SVs



NA12878



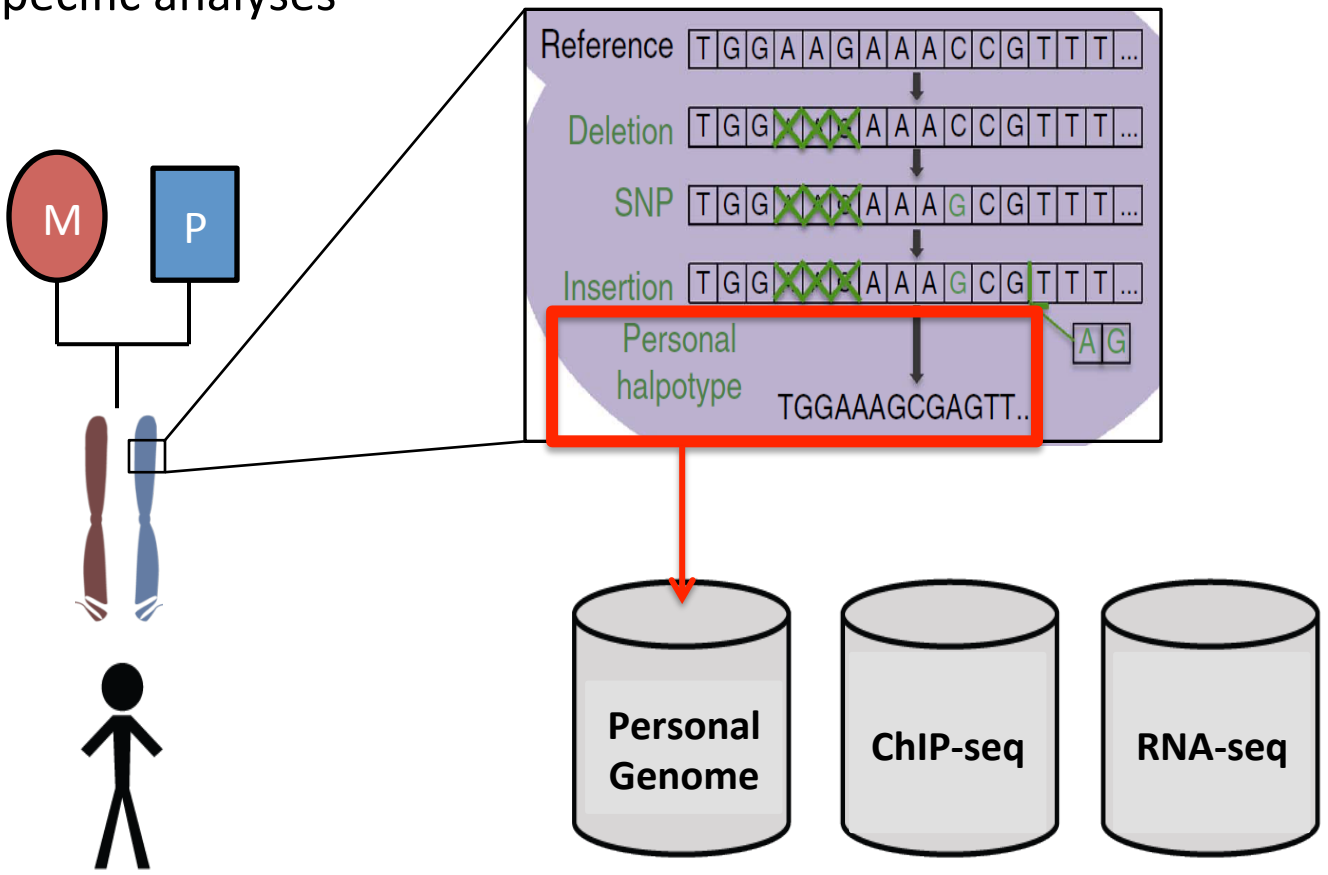
Paternal variation
SNVs: 2,480,064
Indels: 389,657
SVs: 892

Maternal variation
SNVs: 2,468,771
Indels: 388,114
SVs: 847

Jieming Chen, in collaboration with Oliver Stegle

Personal Genome Construction

- AlleleDB uses AlleleSeq pipeline that constructs a personal genome for allele-specific analyses



NA12878

[4] Rozowsky J, et al. Mol Syst Biol, 2011.
http://alleleseq.gersteinlab.org/

Outline

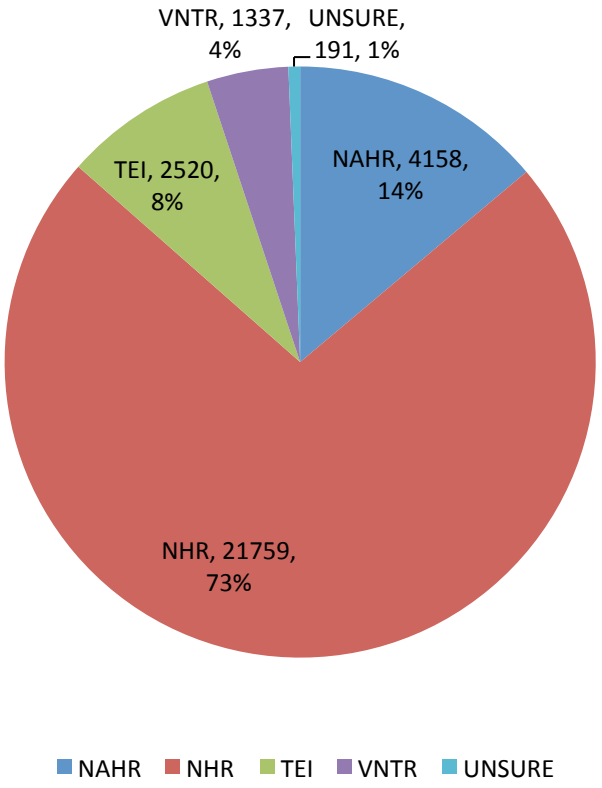
- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Outline

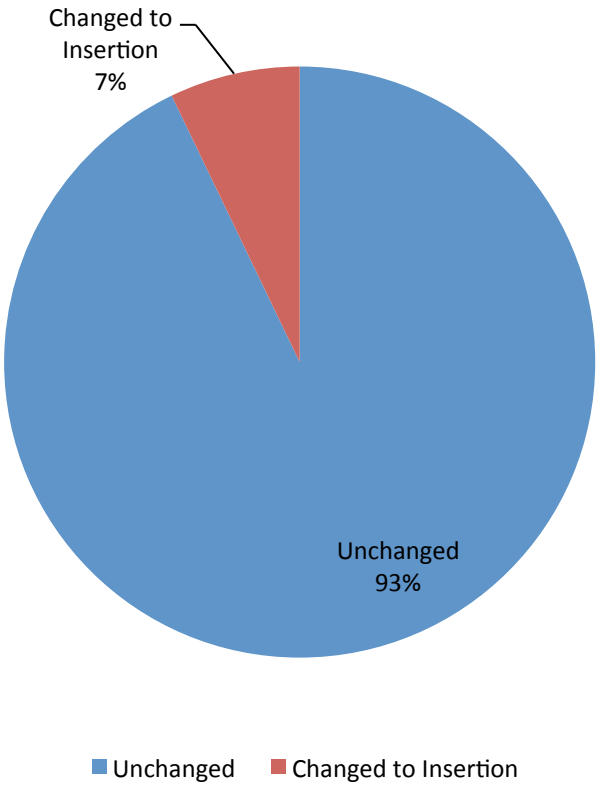
- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

BreakSeq Annotation

Formation Mechanisms



Ancestral States

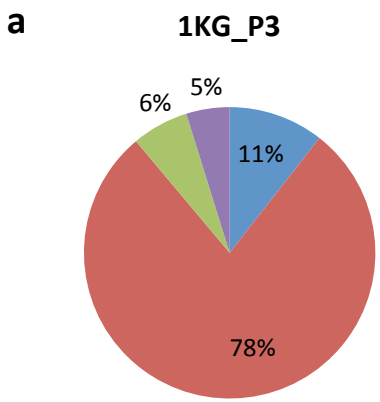


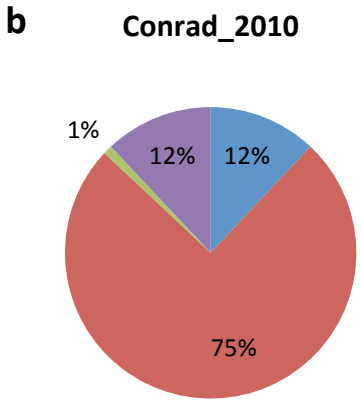
Remarks: There are 79 STEI_NAH events, i.e. 79 events were changed from NAHR to STEI based on our new criteria in the enhanced BreakSeq. Extended annotations from BreakSeq such as NAHR_EXT, STEI_NAH, etc are grouped into their corresponding mechanisms in the above.

Hugo Lam

Formation Mechanism Comparison

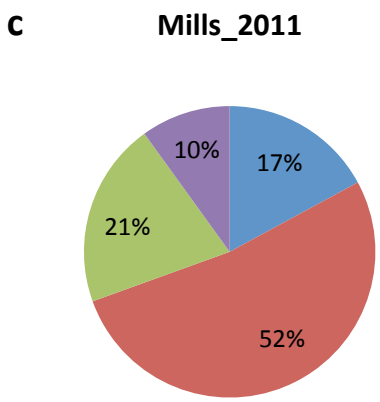
1KG Phase 3

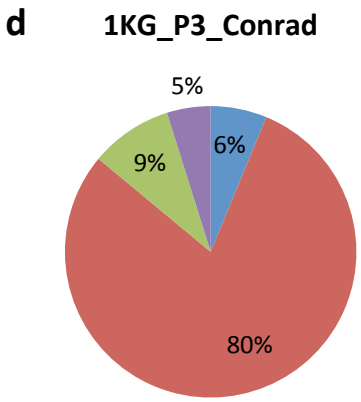




Conrad *et al.* 2010

Mills *et al.* 2011





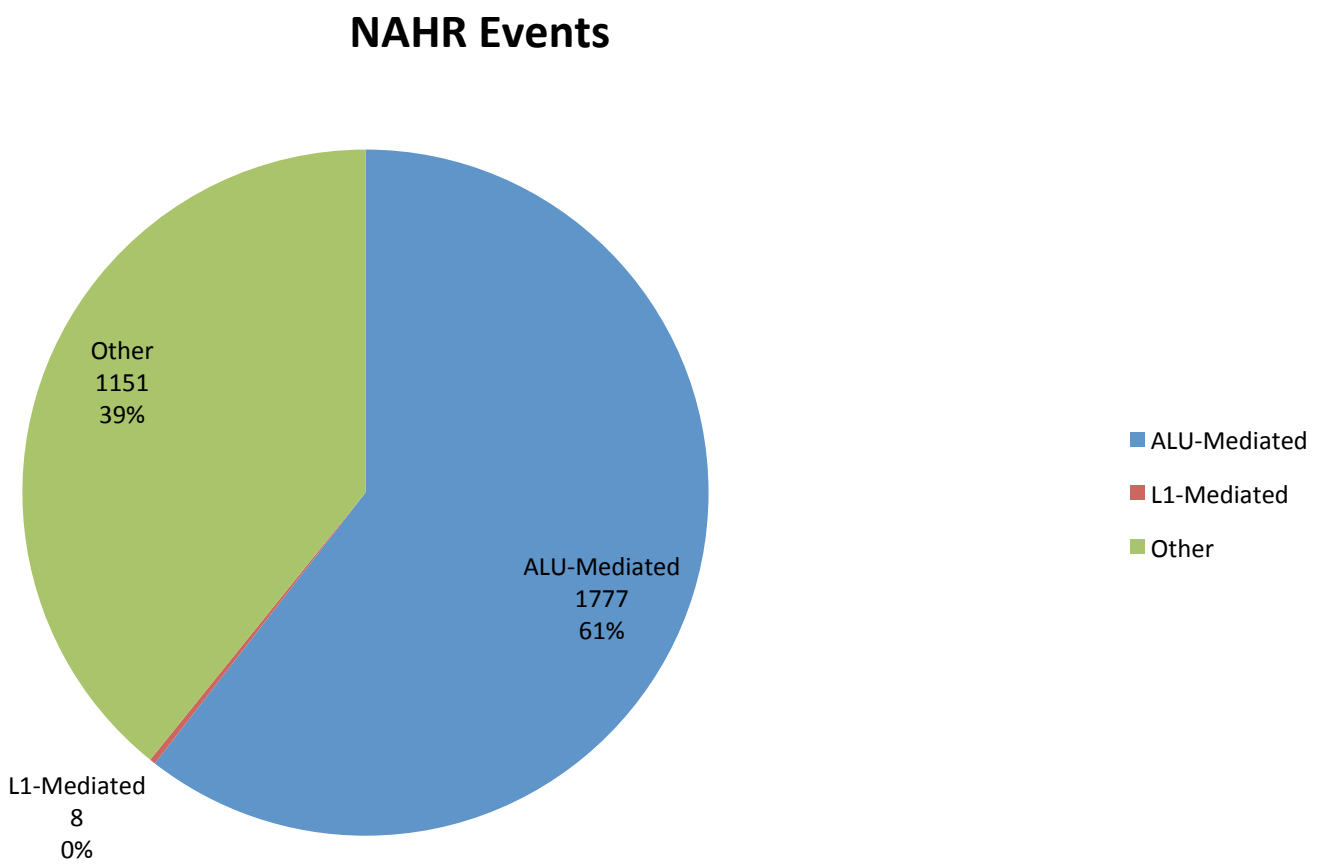
1KG Phase 3 using Conrad *et al.* 2010 criteria

NAHR NHR TEI VNTR

Remarks: For comparison purpose, extended annotations from BreakSeq such as NAHR_EXT, STEI_NAH, etc are not included in the above mechanisms.

Hugo Lam

Repeat-mediated NAHR Events



Hugo Lam

Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Outline

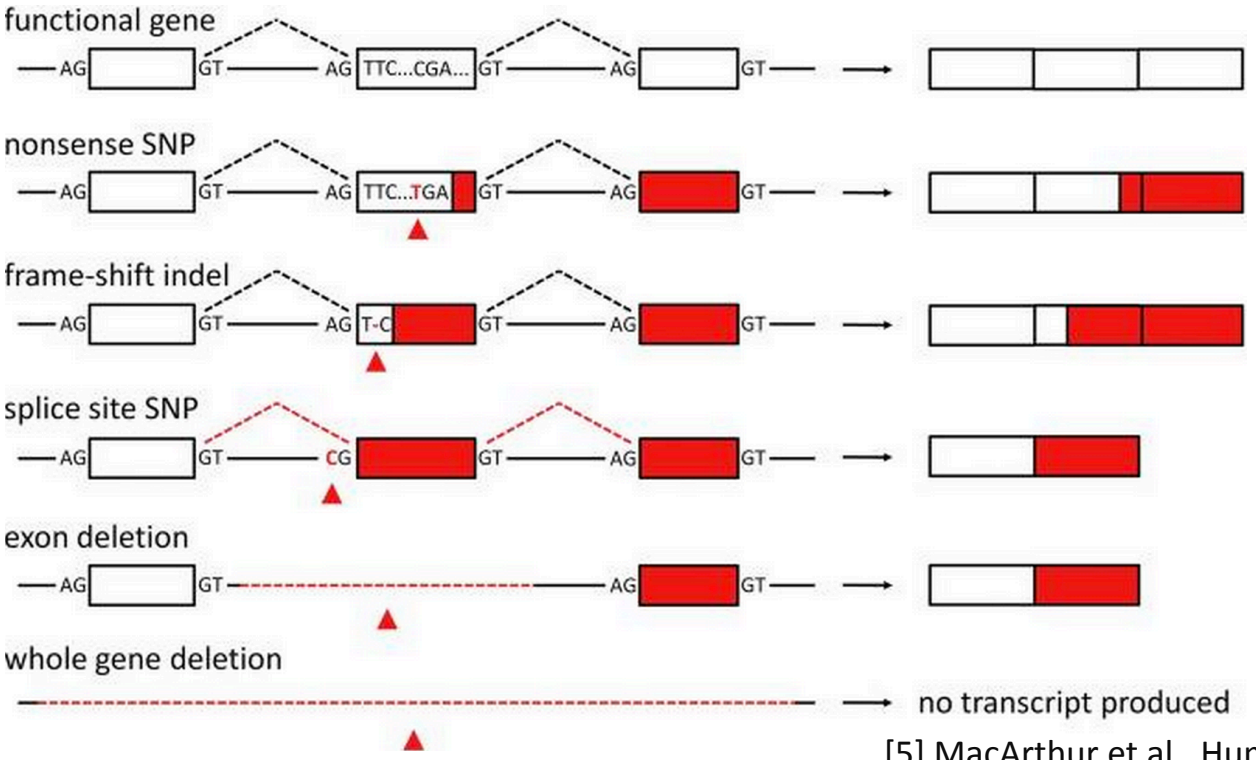
- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Impact of Genetic Variability: Loss-of-function

- Truncating nonsense SNPs
- Splice-disrupting SNPs
- Frameshift-causing indels
- Disrupting structural variants

Yao Fu, Suganthi Balasubramanian

Prevalence of Loss-of-function Variants in Healthy Individuals



[5] MacArthur et al., Hum Mol Genet, 2010

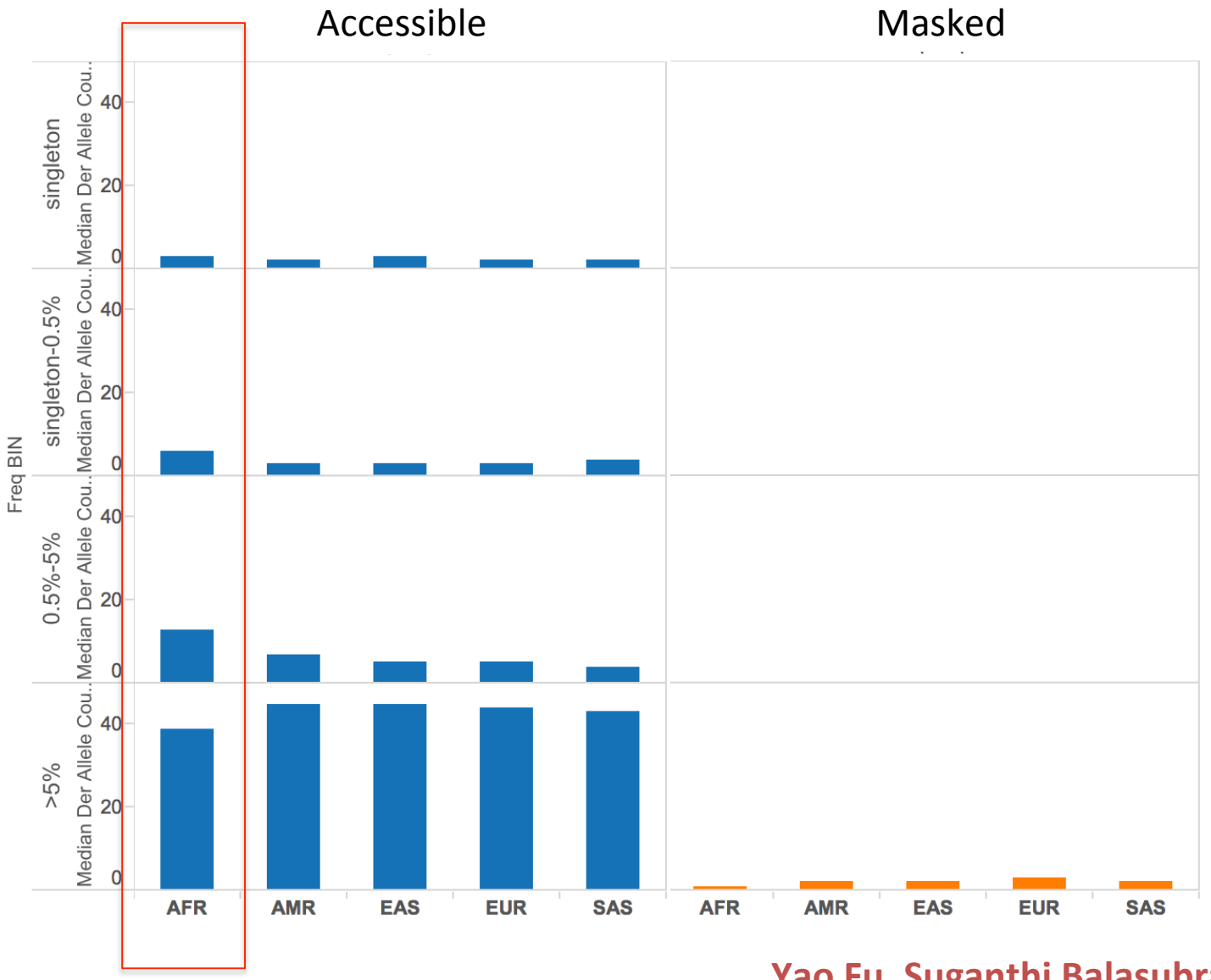
- Previous LoFs are considered as having high probability of being deleterious
- Surprisingly, ~ 100 LoF variants per genome, 20 genes are completely inactivated

Medium 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

[3] 1000GP Consortium. Submitted to Nature, 2015.

Stop-gain (median derived allele counts)



Yao Fu, Suganthi Balasubramanian

Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Outline

- Background
- Functional Impact of Various SVs
- Personal Diploid Genome and Effects on SVs
- SV Formation Mechanism Annotation
- Loss-of-function Annotation
- SVs and Disease Associated lncRNAs

Motivation

- It has been reported that deletion/CNV of lncRNA can be associated with a lethal lung development diseases.

Genome Res. 2013 Jan;23(1):23-33. doi: 10.1101/gr.141887.112. Epub 2012 Oct 3.

Small noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder.

Szafranski P[1], Dharmadhikari AV, Brosens E, Gurha P, Kolodziejska KE, Zhishuo O, Dittwald P, Majewski T, Mohan KN, Chen B, Person RE, Tibboel D, de Klein A, Pinner J, Chopra M, Malcolm G, Peters G, Arbuckle S, Guiang SF 3rd, Hustead VA, Jessurun J, Hirsch R, Witte DP, Maystadt I, Sebire N, Fisher R, Langston C, Sen P, Stankiewicz P.

- We look at functional impact of SVs (including CNVs) on known disease associated lncRNAs.

Three datasets

- **SVs**: 1000G phase 3 SV set.

- **"Conserved" lncRNAs**: A high-quality strict set of human lncRNAs (5413 transcripts) from Nitsche et al. 2015.

RNA. 2015 May;21(5):801-12. doi: 10.1261/rna.046342.114. Epub 2015 Mar 23.

Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved.

Nitsche A[1], Rose D[2], Fasold M[3], Reiche K[4], Stadler PF[5].

A little bit of detail:

GENCODE v14 (GRCh37) + a series of filters

- Remove transcripts that overlap with protein-coding sequences or pseudogenes in sense or antisense by at least one of GENCODE, ENSEMBL, UCSC, or RefSeq.
- Remove transcripts with putative coding regions.
- Remove unspliced entries
- Other cutoffs of PhyloCSF, possible ORF length, etc.

Three datasets

- **Disease associated lncRNAs**: The latest experimentally supported lncRNA-disease association data from LncRNADisease database (as of 4/27/2015).

Nucleic Acids Res. 2013 Jan;41(Database issue):D983-6. doi: 10.1093/nar/gks1099. Epub 2012 Nov 21.

LncRNADisease: a database for long-non-coding RNA-associated diseases.

Chen G[1], Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q.

Database summary from Chen et al. 2013:

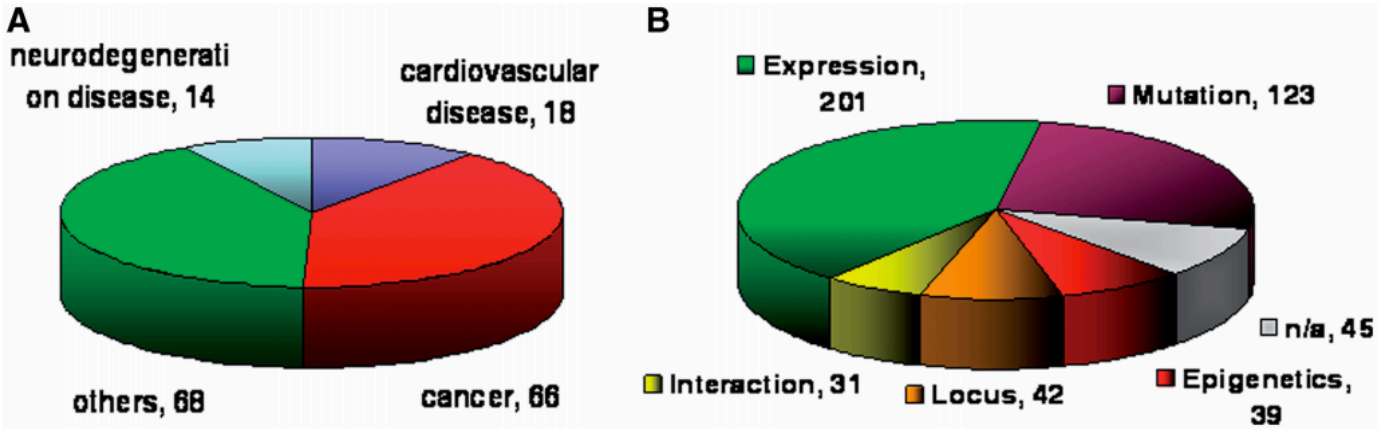
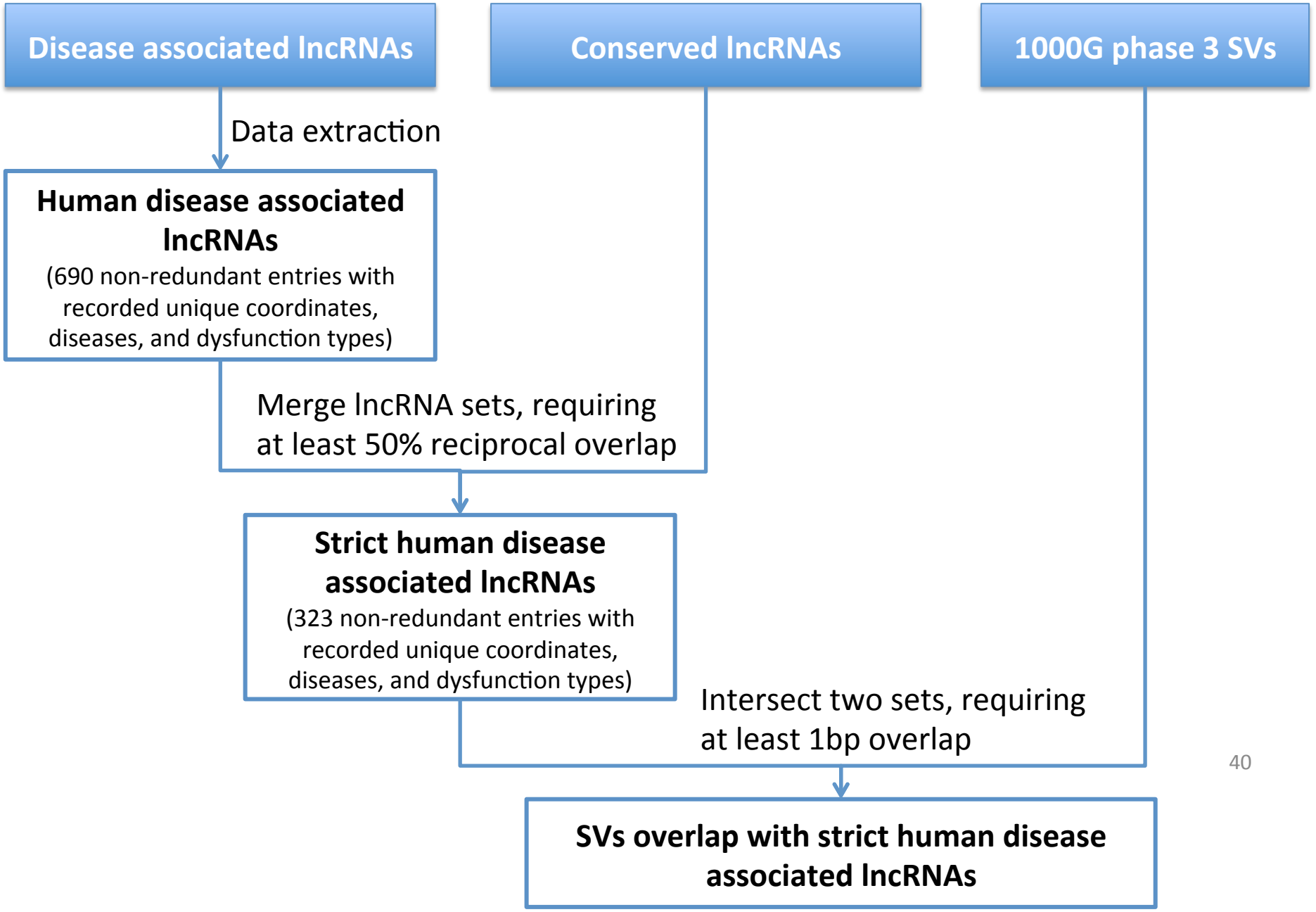


Figure 1. Statistics and distributions of diseases (**A**) and dysfunction types (**B**) of lncRNAs in the LncRNADisease database.

Analysis


Disease associated lncRNAs
Conserved lncRNAs
1000G phase 3 SVs
Data extraction
Human disease associated lncRNAs
(690 non-redundant entries with recorded unique coordinates, diseases, and dysfunction types)
Merge lncRNA sets, requiring at least 50% reciprocal overlap
Strict human disease associated lncRNAs
(323 non-redundant entries with recorded unique coordinates, diseases, and dysfunction types)
Intersect two sets, requiring at least 1bp overlap
SVs overlap with strict human disease associated lncRNAs

Result summary

- 44 unique SVs overlap with strict human disease associated lncRNAs.

DEL	DUP	mCNV	ALU	LINE1
30	4	1	7	2

- Example 1: The SV with the most (7) lncRNA entries

SV Information						lncRNA information						
Chr	Start (0-based)	End	Type	Frequency	ID	Chr	Start (0-based)	End	Strand	Symbol	Associated disease	Dysfunction type
chr11	65182225	65192548	DEL	0.0002	UW_VH_9761	chr11	65190268	65192232	+	NEAT-1	AIDS	expression
chr11	65182225	65192548	DEL	0.0002	UW_VH_9761	chr11	65190268	65192232	+	NEAT-1	amyotrophic lateral sclerosis	regulation
chr11	65182225	65192548	DEL	0.0002	UW_VH_9761	chr11	65190268	65192232	+	NEAT-1	frontotemporal lobar degeneration	Interaction
chr11	65182225	65192548	DEL	0.0002	UW_VH_9761	chr11	65190268	65192232	+	NEAT-1	Huntington's disease	expression
chr11	65182225	65192548	DEL	0.0002	UW_VH_9761	chr11	65190268	65192232	+	NEAT-1	Intrauterine Growth Restriction	expression
chr11	65182225	65192548	DEL	0.0002	UW_VH_9761	chr11	65190268	65192232	+	NEAT-1	TDP-43-associated pathological state	expression
chr11	65182225	65192548	DEL	0.0002	UW_VH_9761	chr11	65190268	65192232	+	NEAT-1	oral squamous cell carcinoma	expression

Result summary

- 135 unique disease associated lncRNA entries overlap with SVs.
- Example 2: The lncRNA overlap with the most SVs

SV Information						lncRNA information						
Chr	Start (0-based)	End	Type	Frequency	ID	Chr	Start (0-based)	End	Strand	Symbol	Associated disease	Dysfunction type
chr2	8170890	8182766	DEL	0.000599	BI_GS_CNV_2_8170891_8182766	chr2	8147900	8418214	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8210077	8210517	DEL	0.0002	DEL_pindel_2551	chr2	8147900	8418214	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8265735	8267776	DEL	0.000399	BI_GS_DEL1_B3_P0259_12	chr2	8147900	8418214	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8359006	8360475	DEL	0.0002	UW_VH_14482	chr2	8147900	8418214	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8383265	8383514	ALU	0.0002	ALU_umary_ALU_988	chr2	8147900	8418214	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8391683	8393675	DEL	0.000599	BI_GS_DEL1_B5_P0259_533	chr2	8147900	8418214	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8170890	8182766	DEL	0.000599	BI_GS_CNV_2_8170891_8182766	chr2	8147900	8464760	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8210077	8210517	DEL	0.0002	DEL_pindel_2551	chr2	8147900	8464760	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8265735	8267776	DEL	0.000399	BI_GS_DEL1_B3_P0259_12	chr2	8147900	8464760	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8359006	8360475	DEL	0.0002	UW_VH_14482	chr2	8147900	8464760	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8383265	8383514	ALU	0.0002	ALU_umary_ALU_988	chr2	8147900	8464760	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8391683	8393675	DEL	0.000599	BI_GS_DEL1_B5_P0259_533	chr2	8147900	8464760	-	LINC00299	Intellectual and developmental disability	mutation
chr2	8426073	8426304	ALU	0.000399	ALU_umary_ALU_989	chr2	8147900	8464760	-	LINC00299	Intellectual and developmental disability	mutation

Acknowledgements

Gerstein Lab @ Yale

Mark Gerstein

Alexej Abyzov (Mayo Clinic)

Jieming Chen

Hugo Lam (Bina)

Yao Fu

Jing Zhang

Xinmeng J. Mu

Suganthi Balasubramanian

Baikang Pei

And other members

P2-VAR

The 1000 Genomes Project

1000GP SV Group

P.H. Sudmant*, T. Rausch*, E.J. Gardner*, R.E. Handsaker*, **A. Abyzov***, J. Huddleston*, **Y. Zhang***, K. Ye*, G. Jun, M.H. Fritz, M.K. Konkel, A. Malhotra, A.M. Stütz, X. Shi, F.P. Casale, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M.J.P. Chaisson, K.Walter, S. Meiers, S. Kashin, E. Garrison, C. Alkan, D. Antaki, T. Bae, P. Chines, **J. Chen**, Z. Chong, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J.M. Kidd, **H.Y.K. Lam**, S. McCarthy, R.A. Gibbs, G. Marth, A. Menelaou, **X.J. Mu**, D.M. Muzny, B. Nelson, A. Noor, N.F. Parrish, A. Quitadamo, B. Raeder, E. Schadt, A. Schlattl, A. Shabalin, A. Untergasser, E. Lameijer, J.A.Walker, M.Wang, F. Yu, C. Zhang, **J. Zhang**, W. Zhou, T. Zichner, J. Sebat, M.A. Batzer, S.A. McCarroll, The 1000 Genomes Project Consortium, R.E. Mills, **M.B. Gerstein**, A. Bashir, O. Stegle, S.E. Devine, C. Lee, E.E. Eichler, J.O. Korbel.

1000GP Functional Interpretation Group