A systematic survey of loss-of-function variants in human protein-coding genes.

Mikk Eelmets

Journal Club

23.04.2012

loss of function (LoF) variants in protein-coding genes

- stop codon–introducing (nonsense),
- splice site–disrupting (SNVs),
- insertion/deletion (indel) variants predicted to disrupt a transcript's reading frame,
- larger deletions removing either the first exon or more than 50% of the proteincoding sequence of the affected transcript

loss of function (LoF) variants

- "less is less” - deleterious alleles (low freq)
- "less is nothing” - poorly evolutionarily conserved genes or belong to multigene families displaying high paralogous sequence identity (higher freq)
- "less is more” - positive selection regions

loss of function (LoF) variants

- How to distinguishing between:

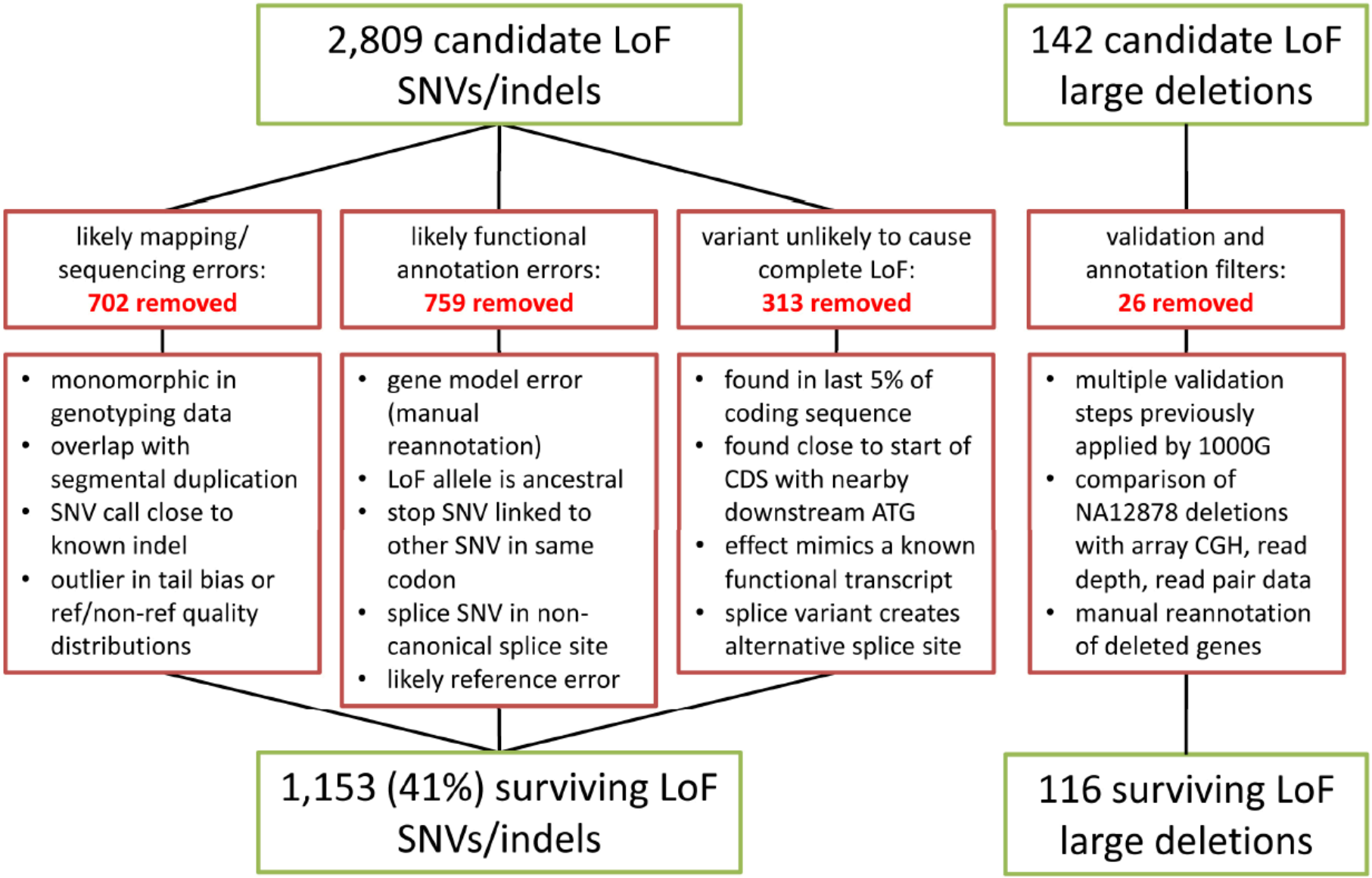
-severe recessive disease alleles in the heterozygous state;

-alleles that are less deleterious but nonetheless have an impact on phenotype and disease risk;

-benign LoF variation in redundant genes;

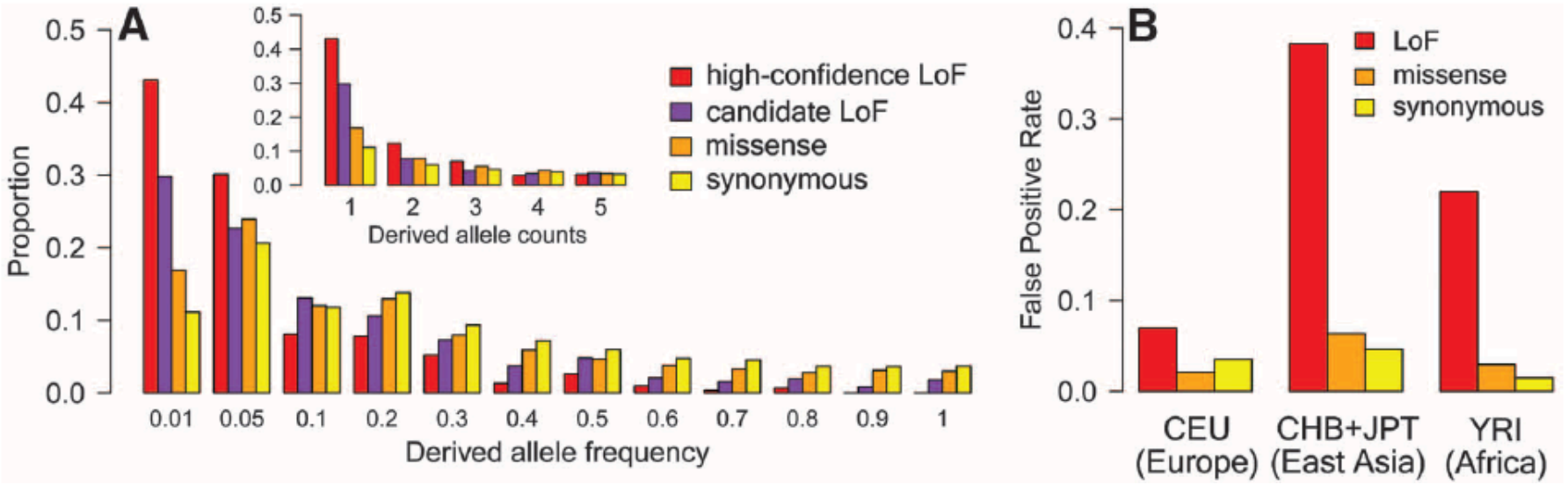
-genuine variants that do not seriously disrupt gene function;

-sequencing and annotation artifacts


2,809 candidate LoF SNVs/indels
142 candidate LoF large deletions
likely mapping/ sequencing errors: 702 removed
likely functional annotation errors: 759 removed
variant unlikely to cause complete LoF: 313 removed
validation and annotation filters: 26 removed
• monomorphic in genotyping data
• overlap with segmental duplication
• SNV call close to known indel
• outlier in tail bias or ref/non-ref quality distributions
• gene model error (manual reannotation)
• LoF allele is ancestral
• stop SNV linked to other SNV in same codon
• splice SNV in non-canonical splice site
• likely reference error
• found in last 5% of coding sequence
• found close to start of CDS with nearby downstream ATG
• effect mimics a known functional transcript
• splice variant creates alternative splice site
• multiple validation steps previously applied by 1000G
• comparison of NA12878 deletions with array CGH, read depth, read pair data
• manual reannotation of deleted genes
1,153 (41%) surviving LoF SNVs/indels
116 surviving LoF large deletions

Variant type	Before filtering				
	Total	1000G low-coverage average per individual			NA12878
		CEU	CHB+JPT	YRI	
Stop	1111	85.7 (21.8)	113.4 (26.7)	109.1 (23.7)	115 (25)
Splice	658	80.5 (29.5)	98.1 (35.6)	89.0 (30.4)	95 (32)
Frameshift indel	1040	217.8 (112.1)	225.5 (121.7)	247.2 (118.7)	348 (159)
Large deletion	142	32.4 (12.2)	31.2 (11.8)	31.4 (9.7)	31 (5)
Total	2951	416.4 (175.6)	468.2 (195.8)	476.7 (316.0)	654 (286)

Variant type	After filtering				
	Total	1000G low-coverage average per individual			NA12878
		CEU	CHB+JPT	YRI	
Stop	565	26.2 (5.2)	27.4 (6.9)	37.2 (6.3)	23 (2)
Splice	267	11.2 (1.9)	13.2 (2.5)	13.7 (1.9)	12 (1)
Frameshift indel	337	38.2 (9.2)	36.2 (9.0)	44.0 (8.0)	38 (11)
Large deletion	116	28.3 (6.2)	26.7 (5.9)	26.6 (5.5)	24 (4)
Total	1285	103.9 (22.5)	103.5 (24.3)	121.5 (21.7)	97 (18)



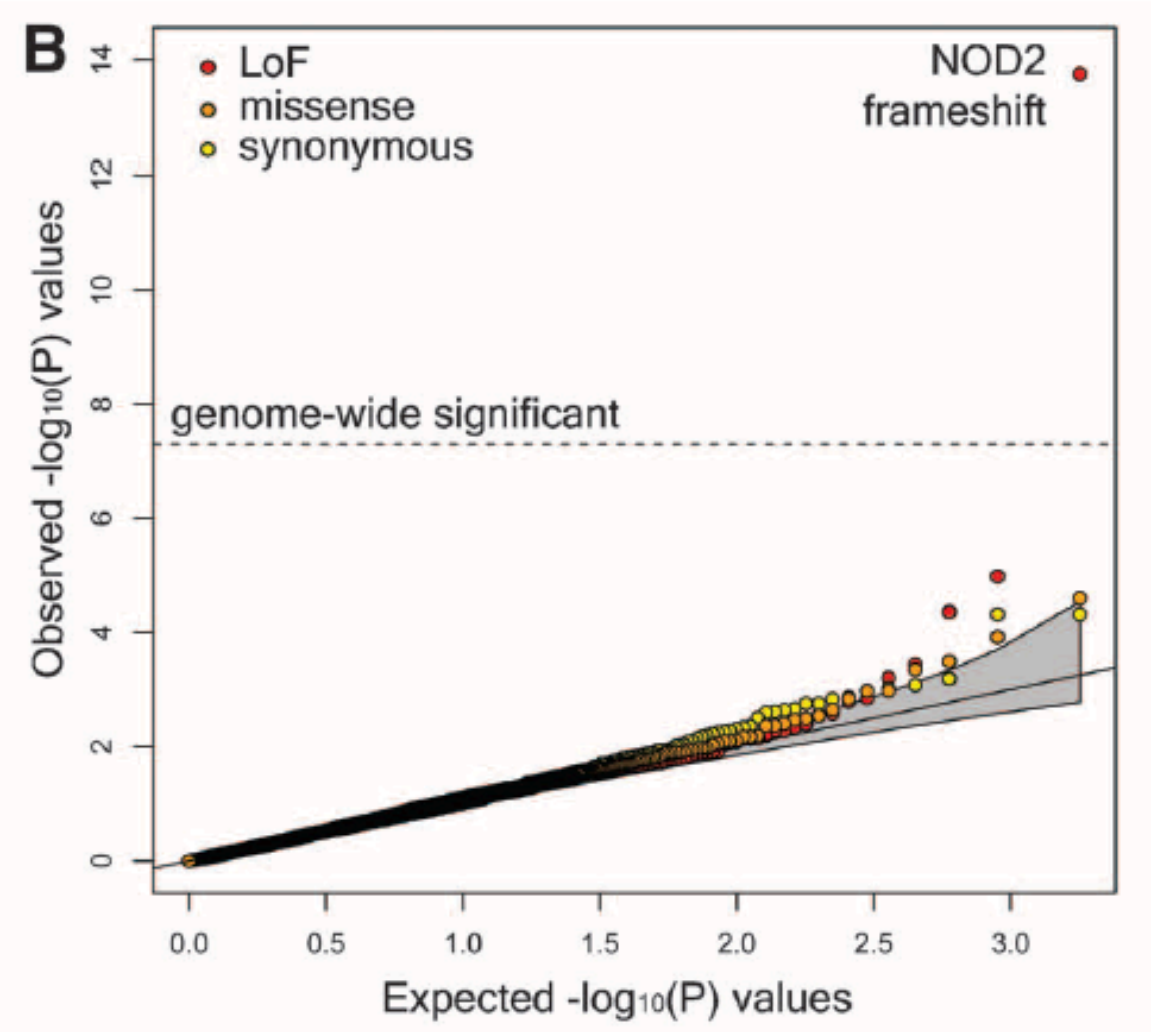
(A) Derived allele frequency distribution in the CEU population for raw and high-confidence LoF variants, compared to missense and synonymous coding variants. (Inset) Distribution of the proportion of SNVs in each class at low allele counts (1 to 5).

(B) False-positive rates (based on independent array genotyping) for LoF variants filtered for annotation artifacts and frequency matched missense and synonymous SNVs.

Properties of LoF variants and affected genes

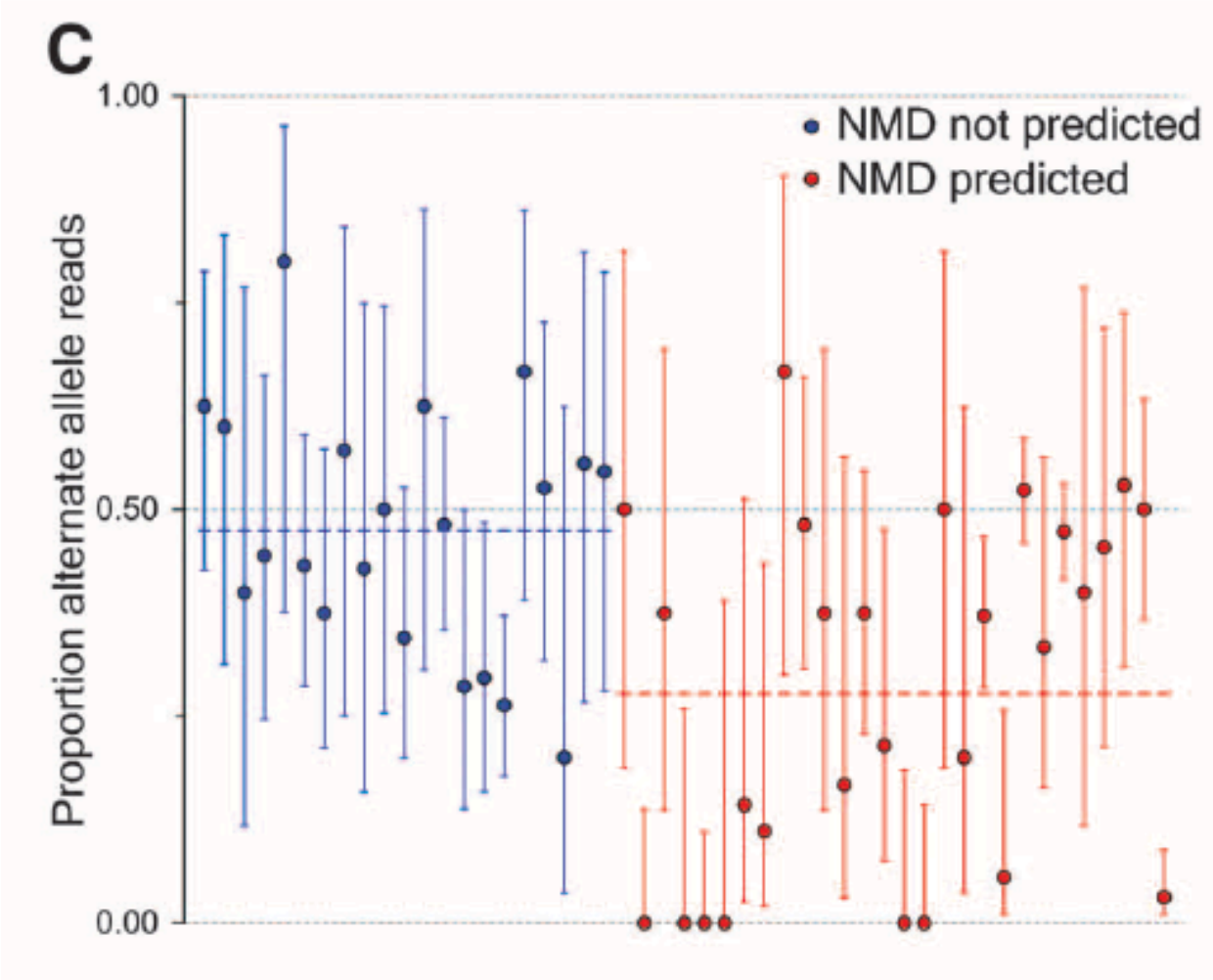
- Enriched for low-frequency alleles compared to synonymous and missense SNVs
- LoF variants per individual is 25% higher in the YRI
- Genes containing high-confidence LoF alleles are relatively less evolutionarily conserved and less evolutionary conservation in their promoter regions
- On average, they have more closely related gene family members than other genes and show greater sequence identity to paralogs
- They also have lower connectivity in both protein-protein interaction and gene interaction networks

Association with risk of common, complex diseases



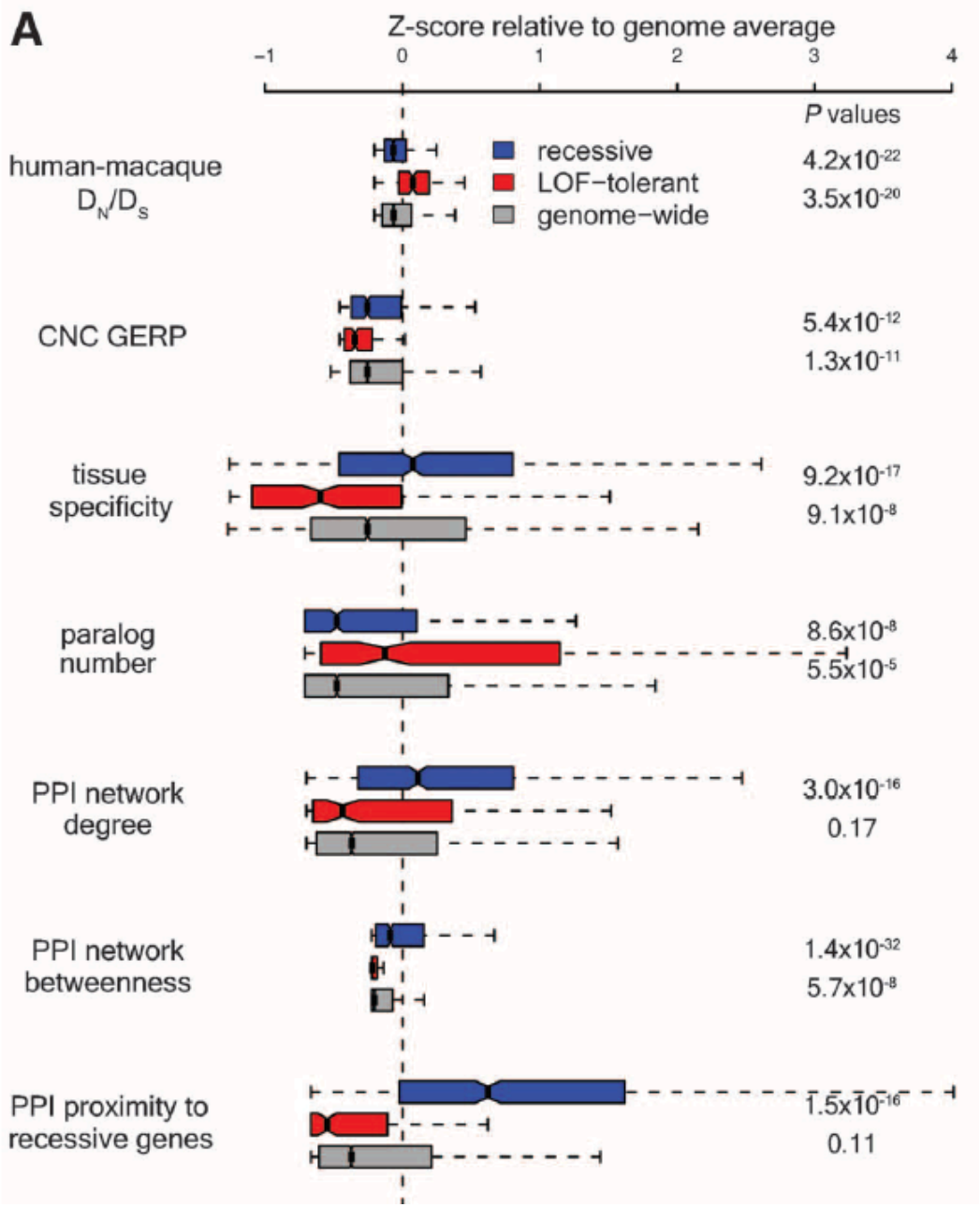
- Association of coding variants with complex disease risk. Observed –log10(P) values for disease association in 17,000 individuals from seven complex disease cohorts and a shared control group, following imputation of variants identified by the 1000 Genomes low-coverage pilot, are plotted against the expected null distribution for all LoF variants and frequency-matched missense and synonymous SNPs

Effects of nonsense SNVs on gene expression



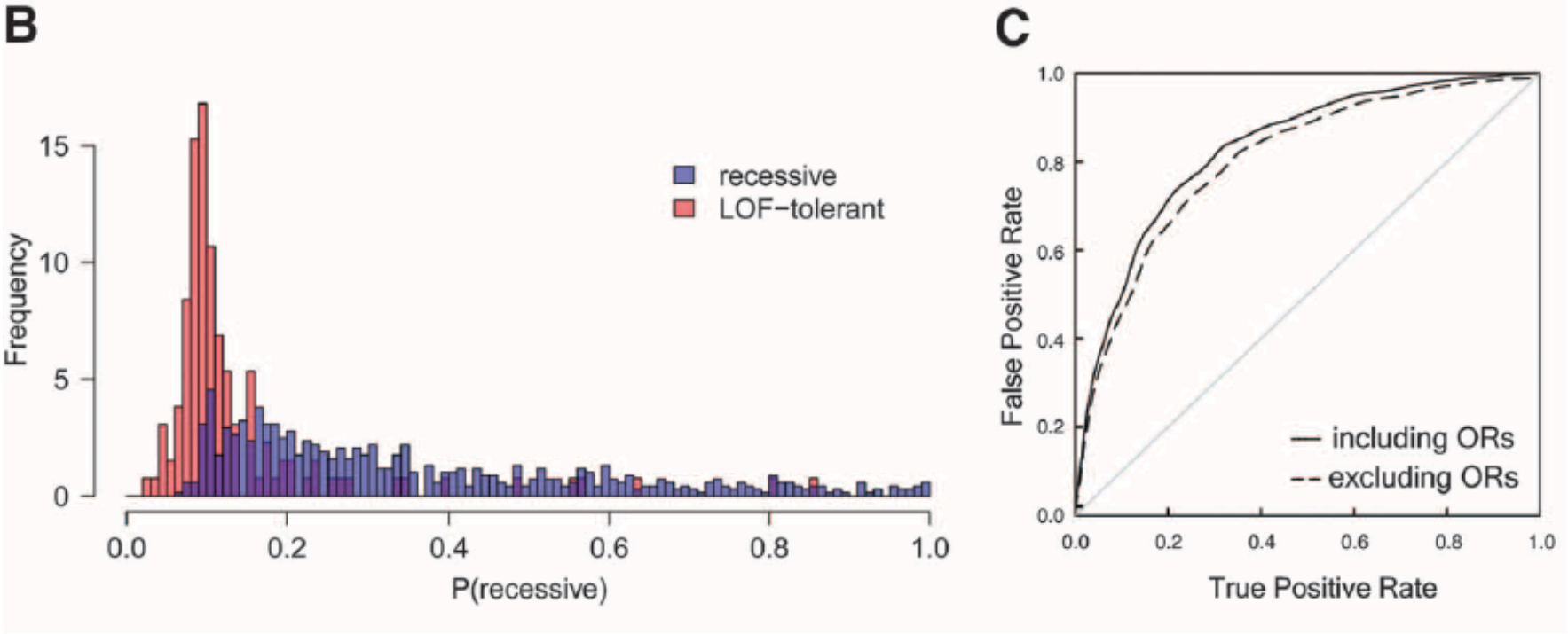
- Allele-specific expression analysis of nonsense variants, using RNA sequencing data from 119 lymphocyte cell lines. Circles show the proportion of LoF-carrying reads spanning each site across all heterozygous individuals. Variants predicted to cause nonsense-mediated decay (NMD, red) and those predicted to escape NMD (blue) are arbitrarily ordered by genome position within each class. Blue and red dashed horizontal lines indicate mean values in each class. Error bars, 95% confidence interval

LoF-tolerant genes VS recessive disease genes



- Distribution of selected evolutionary and functional parameters for recessive disease genes (blue) and LoF-tolerant genes (red) compared to all protein-coding genes (gray). Values are transformed to z scores to allow parameters to be plotted together. Boxes show interquartile range with medians indicated with a vertical black line, and whiskers terminate at themost extreme point less than 1.5 times the interquartile range from the box. For each pair of P values, the top value refers to the recessive versus LoF-tolerant comparison and the bottom value refers to the LoF-tolerant versus genome background comparison. Because many of the parameters are left-skewed, the medians typically fall below zero

Disease/LoF-tolerant genes classification



- P value distribution for linear discriminant model (LDM) trained using LoF-tolerant and recessive disease genes, based on human-macaque Dn/Ds ratio and PPI network proximity to known recessive disease genes.
- Receiver-operating characteristic (ROC) curve for LDM distinguishing between LoF-tolerant and recessive disease genes

Conclusion

- genomes typically contain ~100 genuine LoF variants with ~20 genes completely inactivated
- majority of LoF variants found in an individual genome are common variants in nonessential genes
- LoF-tolerant and recessive disease genes have differing functional and evolutionary properties, allowing us to develop a potential approach for prioritizing novel candidate recessive disease variants identified in patient samples

Reference

MacArthur DG, Balasubramanian S, Frankish A, et al.

" **A systematic survey of loss-of-function variants in human protein-coding genes.** “

Science 2012 Feb 17;335(6070):823-8

THANK YOU