PopViz: a webserver for visualizing minor allele frequencies and damage prediction scores of human genetic variations

Supplementary Material

Peng Zhang, Benedetta Bigio, Franck Rapaport, Shen-Ying Zhang, Jean-Laurent Casanova, Laurent Abel, Bertrand Boisson, and Yuval Itan

Table of Contents

1. PopViz Data Collection, Processing, and Statistics

PopViz collected whole-exome sequencing (WES) data from 123,136 individuals in 7 populations (African/African American, Latino, Ashkenazi Jewish, East Asian, Finnish, Non-Finnish European, South Asian), from gnomAD r2.0.2 (Lek, et al., 2016), currently the most extensive publicly available population genetic database (compared to resources such as ExAC, BRAVO, 1,000 Genomes Project). The gnomAD was therefore selected as the reference database for PopViz webserver.

Supplementary Table S1: Databases of genetic variants

Database	Number of Individuals	Populations	Link
gnomAD	123,136	African/African American, Latino, Ashkenazi Jewish, East Asian, Finnish, Non-Finnish European, South Asian	http://gnomad.broadinstitute.org/
ExAC	60,706	African, East Asian, Finnish, Latino, Non-Finnish European, South Asian	http://exac.broadinstitute.org/
BRAVO hg38	62,784	N/A	https://bravo.sph.umich.edu/freeze5/hg38/
BRAVO hg19	14,559	N/A	https://bravo.sph.umich.edu/freeze3a/hg19/
1000 Genomes Project	2,504	African, American, East Asian, European, South Asian	http://www.internationalgenome.org/home

The variants downloaded from gnomAD were applied with the following inclusion/exclusion criteria to achieve a collection of 13,681,468 variants from 20,437 genes, to build up PopViz:

(1) the variant must have PASS in the VCF file's FILTER field;

(2) the variant must be in a canonical transcript;

(3) the variant must have a gene symbol annotated, and the gene is supported in Ensembl;

(4) the variant should be annotated as one of the 14 selected consequences;

(5) the indels that have more than 10 nucleotide changes were excluded.

Currently, the PopViz webserver is based on canonical transcripts, so the variants and their annotations on the alternative splicing transcripts are not yet supported.

PopViz provides five mutation deleteriousness prediction scores (**Supplementary Table S2**): CADD (Kircher, et al., 2014), EIGEN (Ionita-Laza, et al., 2016), LINSIGHT (Huang, et al., 2017), SIFT (Kumar, et al., 2009), and PolyPhen-2 (Adzhubei, et al., 2010). The pre-computed CADD, EIGEN, and LINSIGHT scores were obtained from the respective websites, and then matched with the variants stored in PopViz. These three sets of scores are available as plotting parameters. SIFT and PolyPhen-2 scores were extracted from the annotations from the gnomAD database, and these two scores are only displayed in the drop-down window for the details of the variants. Additional effective methods, such as DDIG for predicting the pathogenicity of frameshift, nonsense, and synonymous mutations (Folkman, et al., 2015), will be integrated into PopViz once the pre-computed scores are available.

Supplementary Table S2: Statistics of mutation damage scores supported in PopViz

Mutation Damage Scores	Number of Variants Matched with gnomAD	Number of Variants having Pre-Computed Scores in PopViz
CADD	13,681,468	509,303,619
EIGEN	5,641,271	81,366,024
LINSIGHT	4,708,171	403,864,908
SIFT	4,488,610	.
PolyPhen-2	4,875,989	.

The following 14 mutation consequences were extracted from the gnomAD and selected by PopViz: 3'UTR, 5'UTR, frameshift, inframe deletion, inframe insertion, intronic, missense, splice acceptor, splice donor, splice region, start lost, stop gained, stop lost, and synonymous. The following mutation consequences were excluded in PopViz: (1) upstream and downstream variants, as they have a lesser impact on gene/protein function, lesser association with pathogenicity, and would produce excessive visual noise in the plots of most genes; (2) the consequences with a very small number of cases (e.g. six variants annotated as 'transcript-ablation', and 52 variants annotated as

'non-coding-transcript'); and (3) consequences that were unclear or unexplained by the gnomAD database (e.g. 'protein-altering variant', 'stop-retrain variant', 'incomplete-terminal-codon variant'). See **Supplementary Table S3** for statistics.

Supplementary Table S3: Statistics of mutation consequences in gnomAD and PopViz

gnomAD Consequences	Counts in Raw Data	Counts in PopViz
3_prime_UTR_variant	412,522	388,811
5_prime_UTR_variant	249,068	231,281
coding_sequence_variant	1,208	.
downstream_gene_variant	2,765,994	.
frameshift_variant	232,926	204,977
incomplete_terminal_codon_variant	147	.
inframe_deletion	83,981	56,117
inframe_insertion	31,943	12,849
intron_variant	5,804,353	4,800,908
mature_miRNA_variant	3,794	.
missense_variant	5,005,108	4,884,607
non_coding_transcript_exon_variant	432,528	.
non_coding_transcript_variant	52	.
protein_altering_variant	1,382	.
splice_acceptor_variant	41,957	37,039
splice_donor_variant	52,101	46,541
splice_region_variant	566,884	526,300
start_lost	11,107	10,712
stop_gained	149,383	142,176
stop_lost	5,135	4,936
stop_retained_variant	2,617	.
synonymous_variant	2,389,968	2,334,214
transcript_ablation	6	.
upstream_gene_variant	2,466,042	.
TOTAL	20,710,206	13,681,468

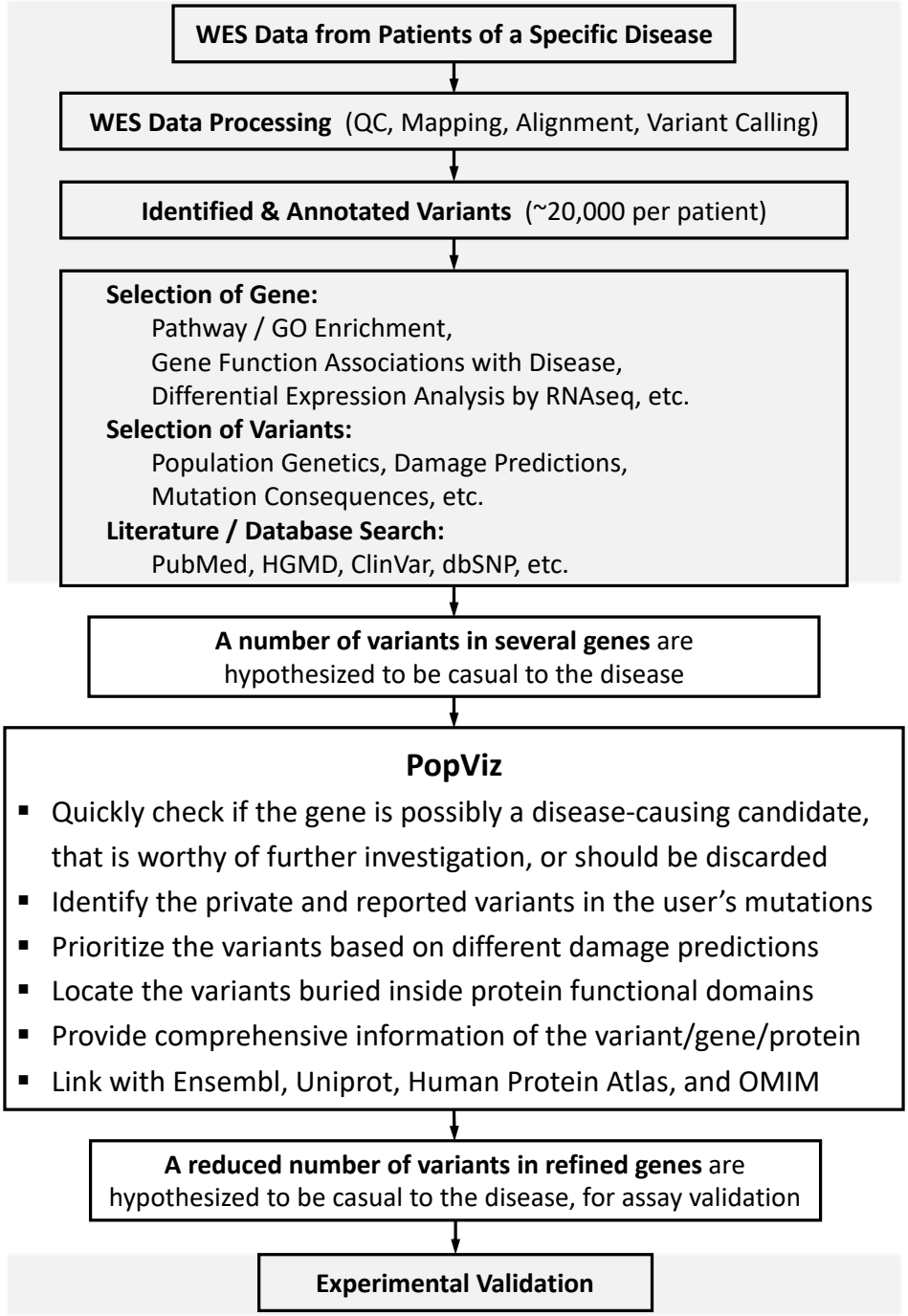
2. PopViz Application

2.1. Facilitate the Decision Making for Candidate Genes and Variants

PopViz, as a rapid approach to extract, integrate, and visualize population genetics, mutation damage prediction scores, and protein/gene information, would facilitate the selection and prioritization of disease-causing candidate genes and variants prior to experimental validation. PopViz is not designed as the first-step analysis to screen all the genes and variants from the entire WES / whole genome sequence (WGS) data of an individual or a cohort, but it will facilitate rapid check of whether the hypothesized disease-causing gene is a plausible candidate. A reinforcement or a quick rejection of the hypothesis would make the investigation more efficient. Once a gene is selected as a good candidate for further investigation, PopViz will help to guide the selection and prioritization of the candidate variants to be tested experimentally (**Supplementary Figure S1**). Without PopViz, investigators may lose time and effort, roving various databases, and tediously prioritizing candidate genes and variants manually.

To illustrate the value of PopViz in facilitating the decision making for candidate genes and variants, we are presenting one schematic example and three recent case studies in the following sections. Information about the definition and the calculation of MAF compatibility for variants in genetic diseases is provided in **Supplementary Box S1**, on basis of assuming the complete penetrance, and knowing the mode of inheritance and the prevalence of the disease.

Supplementary Figure S1: PopViz application in the discovery of disease-causing genes/variants



Supplementary Box S1: Computation of MAF compatibility for variants in genetic disorders

Consider a diploid autosomal locus with two alleles has '*AA*', '*Aa*', and '*aa*' phenotypes. Population genetic variation is usually described by allele frequency, by letting $p=f(A)$ and $q=f(a)$, where q is used for the common wild-type allele, and q (minor allele frequency, MAF) is used for the rare and deleterious allele.
The following computation of MAF compatibility is based on the hypothesis of complete penetrance.

Therefore, we have:

$\#AA + \#Aa + \#aa = N$ (population size)

$f(AA) = p^2$

$f(Aa) = 2pq$

$f(aa) = q^2$

$p + q = 1$

If a disease is autosomal dominant (AD), and its prevalence is $f(AD)$, then we have:

$f(AD) = (\#Aa + \#aa) / N$

$= f(Aa) + f(aa)$

$= 2pq + q^2$

$= 2(1-q)q + q^2$

$\rightarrow q^2 - 2q + f(AD) = 0$

$\rightarrow \boldsymbol{q = 1 - \sqrt{1 - f(AD)}}$, MAF cutoff for variants to be compatible with the AD disease prevalence

If a disease is autosomal recessive (AR), and its prevalence is $f(AR)$, then we have:

$f(AR) = \#aa / N$

$= f(aa)$

$= q^2$

$\rightarrow \boldsymbol{q = \sqrt{f(AR)}}$, MAF cutoff for variants to be compatible with the AR disease prevalence

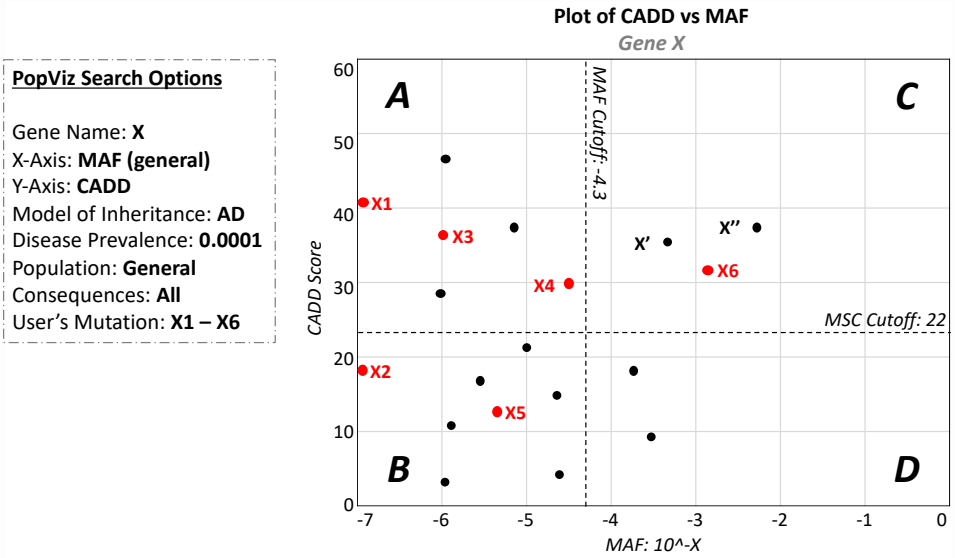
Since the PopViz webserver is based on WES data of 123,136 individuals from the gnomAD database, the lowest MAF that can be reached is: $1 / (2 \times 123136) = 4.06 \times 10^{-6}$.

2.2. A Schematic Example

Assume having a cohort with some patients with an autosomal dominant (AD) disease, and the disease prevalence is 1/10,000. The WES data processing, analysis, and annotation gives a number of variants in multiple genes as disease-causing candidates. As a schematic example, gene *X* and its six variants (X_1, X_2, X_3, X_4, X_5, X_6) are hypothesized to be deleterious. PopViz will first help to quickly check if gene *X* is a good candidate, and then help to prioritize its candidate variants for experimental test.

In the schematic plot of CADD vs MAF (**Supplementary Figure S2**), the black dots are variants from the gnomAD database, and the red dots are the user's variants, and some of the user's variants may overlap with gnomAD variants. In this case, X_1 and X_2 are private variants, so their MAFs are set to 10^{-7}. X_3 - X_6 have matches in PopViz, so their MAFs are based on gnomAD. This plot is split into quarters (A, B, C, D) by MAF cutoff ($10^{-4.3}$, according to **Supplementary Box S1**) and MSC cutoff (arbitrarily, 22). Here, the variants in Region A and B are compatible with the disease prevalence, and the variants in Region C and D are not.

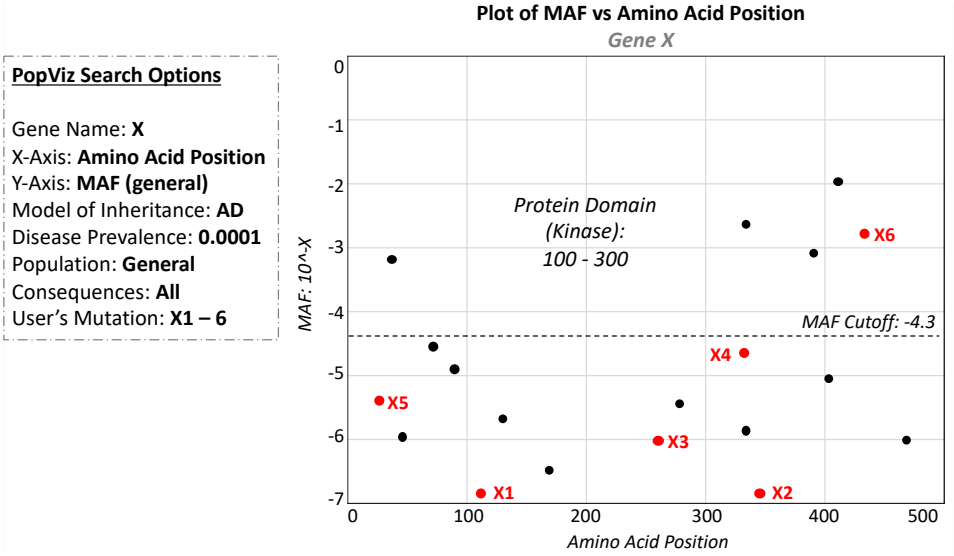
Supplementary Figure S2: Plot of CADD vs MAF in the schematic example



To study if gene *X* is a likely disease-causing candidate gene, we could examine if these three variants (X_6, *X'* and *X''*) are predicted as loss-of-function by multiple mutation damage scores, or quickly carry out assays to test these three variants. If yes, the hypothesis based on gene *X* could be rejected. This is because the deleterious mutation in gene *X* at this high frequency will prohibit this gene to be causal to this AD disease with a prevalence of 1/10,000. If not, the hypothesis based on gene *X* would be reinforced, and we could further investigate whether these candidate variants are disease-causing. Now, X_6 can be removed because the previous assay has proven that it is benign. Variants X_1, X_3, and X_4 are then prioritized higher than X_2 and X_5, due to their higher damage scores. Using the CADD score alone might be insufficient, so additional deleteriousness scores may be considered as well, such as EIGEN, LINSIGHT, SIFT, PolyPhen-2, DDIG, etc.

Moreover, by plotting MAF vs Amino Acid Position (AA_Pos) (**Supplementary Figure S3**), we found that variants X_1 and X_3 are inside the protein functional domain (for instance, protein kinase domain from amino-acid position 100 to 300), therefore the variants X_1 and X_3 should be prioritized, followed by X_4, X_2 and X_5, for experimental validation. In the PopViz webserver, users can always customize their search by restricting the population, the consequence, the loss-of-function prediction, the homozygosity, and other available features, to facilitate their decision making in the investigation of the disease-causing candidate genes or variants.

Supplementary Figure S3: Plot of MAF vs Amino Acid Position in the schematic example



If the disease is autosomal recessive (AR) with a prevalence of 1/10,000, then the MAF cutoff will be 10^{-2} (**Supplementary Box S1**). The searching option '*Only Variants Having Homozygous Alleles'* is suggested to be checked, as it will only extract and visualize the variants that have been identified as homozygous in at least one individual in the gnomAD database.

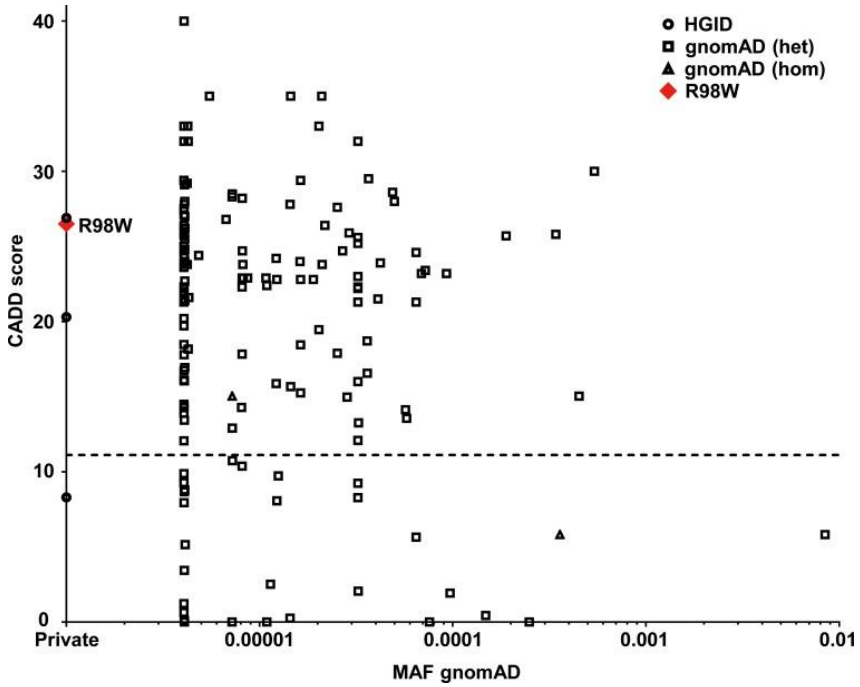
2.3. Case Studies

To further illustrate the value of PopViz webserver, we present three recent case studies that used the CADD vs MAF plot to help to: 1) remove the predicted benign variants in gene *IRF4* in Whipple's disease (Guerin, et al., 2018); 2) select the potentially disease-causing variants in gene *IKZF1* in common variable immunodeficiency (Kuehn, et al., 2016); and 3) select the possibly damaging variants in gene *DBR1* in viral encephalitis for experimental characterization (Zhang, et al., 2018).

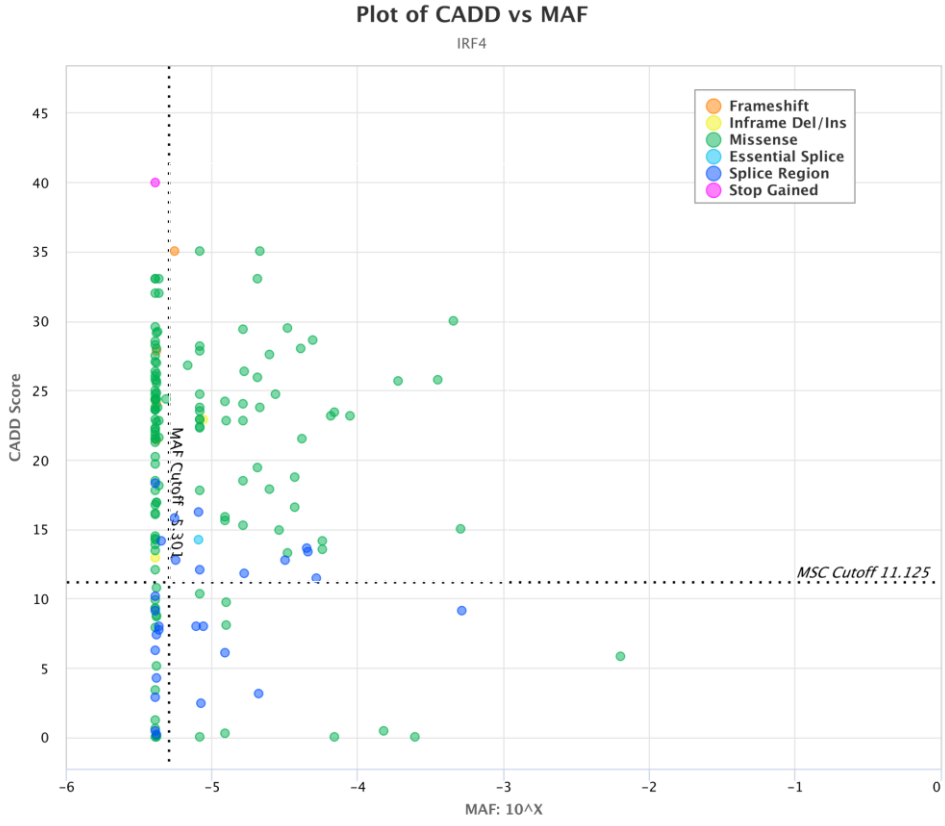
2.3.1. *IRF4* Mutations in Whipple's Disease

By combining linkage analysis with WES data in a large multiplex family, Guerin et al. identified heterozygous R98W *IRF4* mutation as the strongest candidate to underlie Whipple's disease (WD), an extremely rare disease, inherited in an AD manner, with prevalence $<1/10^5$ (Guerin, et al., 2018). They plotted CADD vs MAF with a MSC cutoff at 95% confidence interval, and identified 156 high-confidence heterozygous non-synonymous coding or splice variants of *IRF4* in the gnomAD database and their in-house database. The first interesting finding was that R98W had the second highest CADD score of the four variants with the lowest MAF$<4\times10^{-6}$ (**Supplementary Box S1**). In addition, their CADD vs MAF plot provided 156 variants that were tested experimentally for their function. Interestingly, only seven of these variants were found to be truly loss-of-function or hypomorphic, and the cumulative frequency was $<4\times10^{-5}$, fully consistent with the low prevalence of WD which occurs only in the individuals infected by the bacteria *Tropheryma whipplei*.

Supplementary Figure S4: CADD vs MAF plot of gene *IRF4* from Guerin's paper



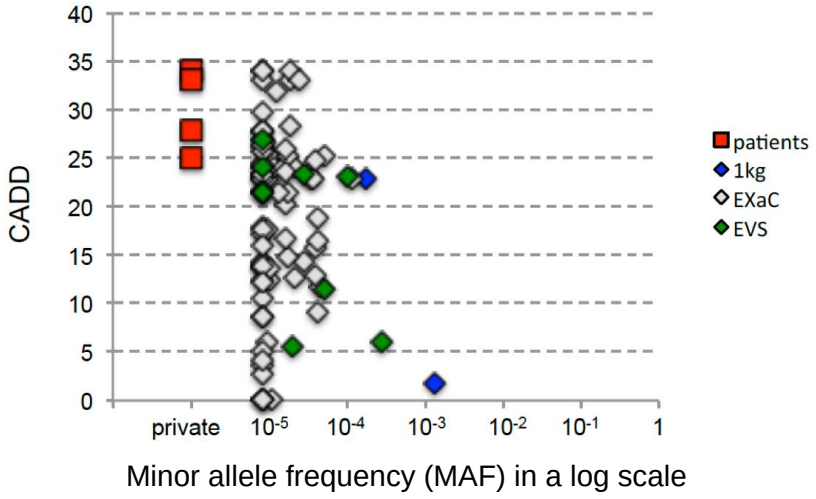
Supplementary Figure S5: CADD vs MAF plot of gene *IRF4* by PopViz, where the variants inside the yellow region were experimentally proven to be mostly benign, by Guerin et al.



2.3.2. *IKZF1* Mutations in Common Variable Immunodeficiency

Kuehn et al. investigated the loss of B cells in patients (29 individuals from 6 families) with common variable immunodeficiency (CVID), that have heterozygous mutations in gene *IKZF1*, which is a transcriptor regulator in the development of lymphocytes, B and T cells (Kuehn, et al., 2016). They used the CADD vs MAF plot together with protein domain information, to validate whether the four private mutations could be disease-causing by taking into account the CADD scores and the amino acid positions of these mutations compared to the reported variants in the worldwide population. Their assays examined the proteins bearing these missense mutations failed to bind target DNA sequences, as these damaging mutations resided inside the important zinc-finger regions of *IKZF1*. Hence, these *IKZF1* mutations were associated with CVID patients who showed a striking decrease in B cells.

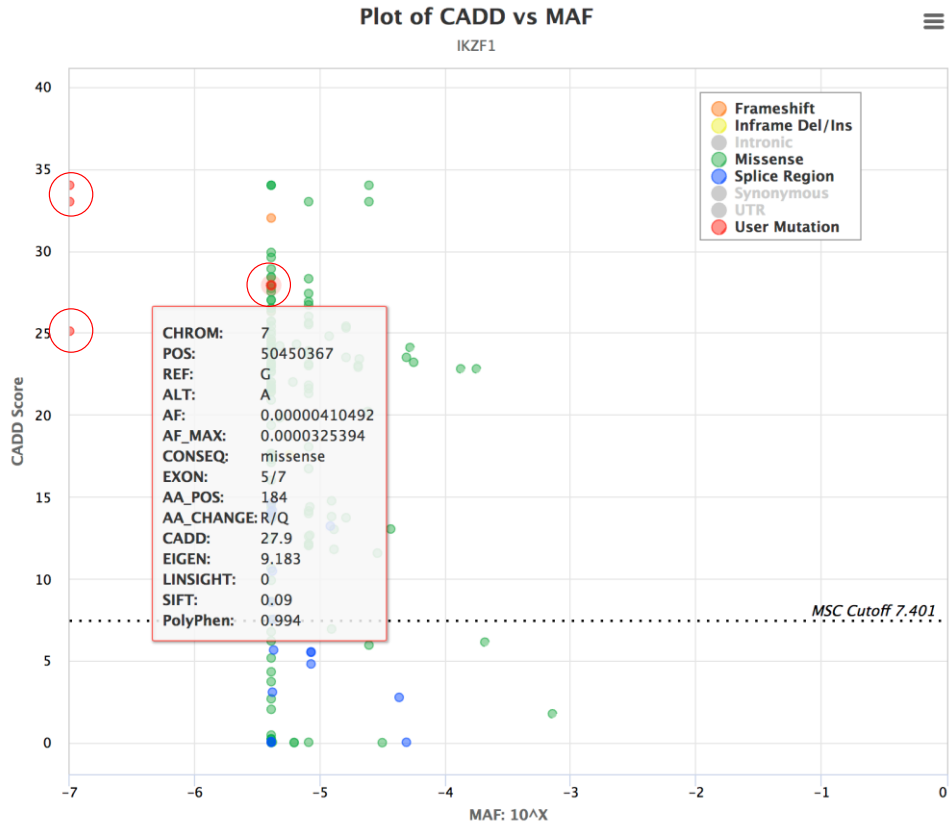
Supplementary Figure S6: CADD vs MAF plot of gene *IKZF1* from Kuehn's paper



Kuehn used the 1000 Genomes Project, ExaC and EVS databases as the population genetics reference, and all four mutations were found as private. PopViz is based on gnomAD (a population genetic database with many more individuals across more populations), and now one of the

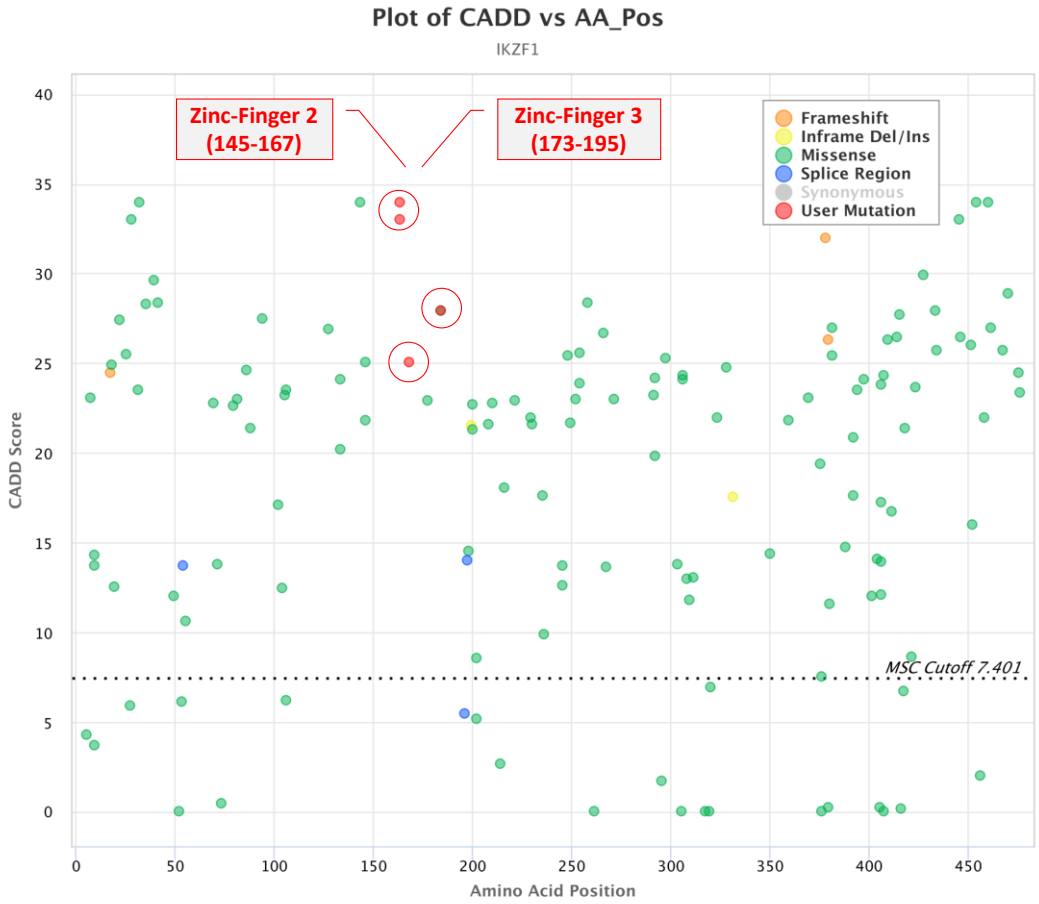
mutations is reported in one individual and still compatible with the reported incomplete penetrance.

Supplementary Figure S6: CADD vs MAF plot of gene *IKZF1* by PopViz,



Additionally, PopViz could help to identify the distribution of the variants of interest, based on the amino acid position. **Supplementary Figure S7** shows the plot of CADD vs amino acid positions for gene *IKZF1*, where the four mutations from the user are in red. It clearly highlights that these mutations are localized in the conserved domains, zinc-finger 2 (from position 145 to 167) and zinc-finger 3 (from position 173 to 195), which seems intolerant to missense variations. To have more information about the protein domain/region/family, users can easily find the cross-reference link to Uniprot database, in the table below the plot.

Supplementary Figure S7: CADD vs Amino Acid Position plot of gene *IKZF1* by PopViz,

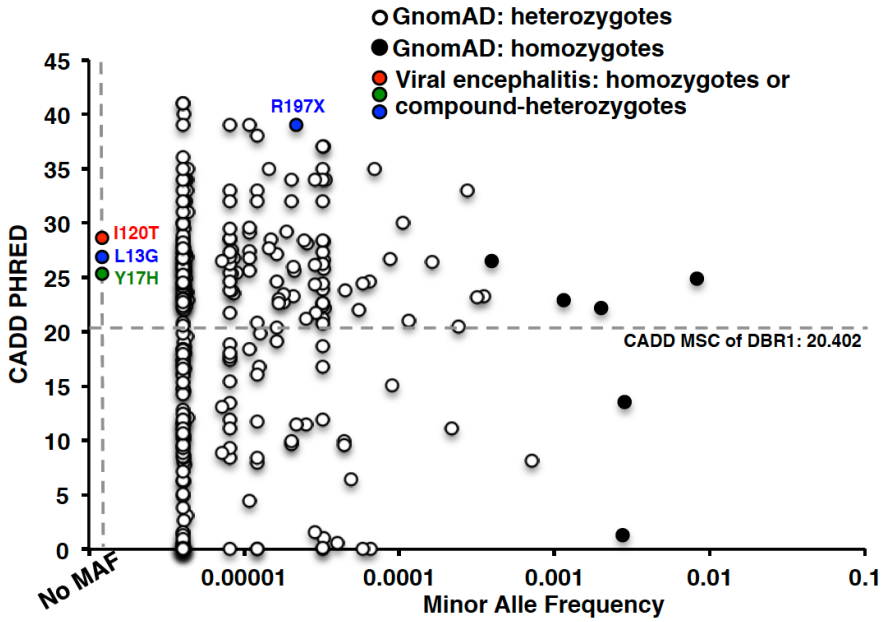


Gene Information	
Gene :	IKZF1
Gene Synonyms :	IKZF1, IK1, IKAROS, LYF1, ZNFN1A1
Gene Description :	IKAROS family zinc finger 1 (Ikaros)
UniProt :	Q13422
Ensembl (hg19) :	ENSG00000185811
Ensembl (hg38) :	ENSG00000185811
Protein Expression :	IKZF1 (Human Protein Atlas)
Protein Length :	519
GO Slim (Biological Process) :	[GO:0007049] cell cycle
GO Slim (Cellular Component) :	[GO:0005737] cytoplasm; [GO:0005654] nucleoplasm; [GO:0005634] nucleus; [GO:0043234] protein complex
GO Slim (Molecular Function) :	[GO:0003677] DNA binding; [GO:0003700] DNA binding transcription factor activity
OMIM :	603023
OMIM Phenotype :	Immunodeficiency, common variable, 13, 616873

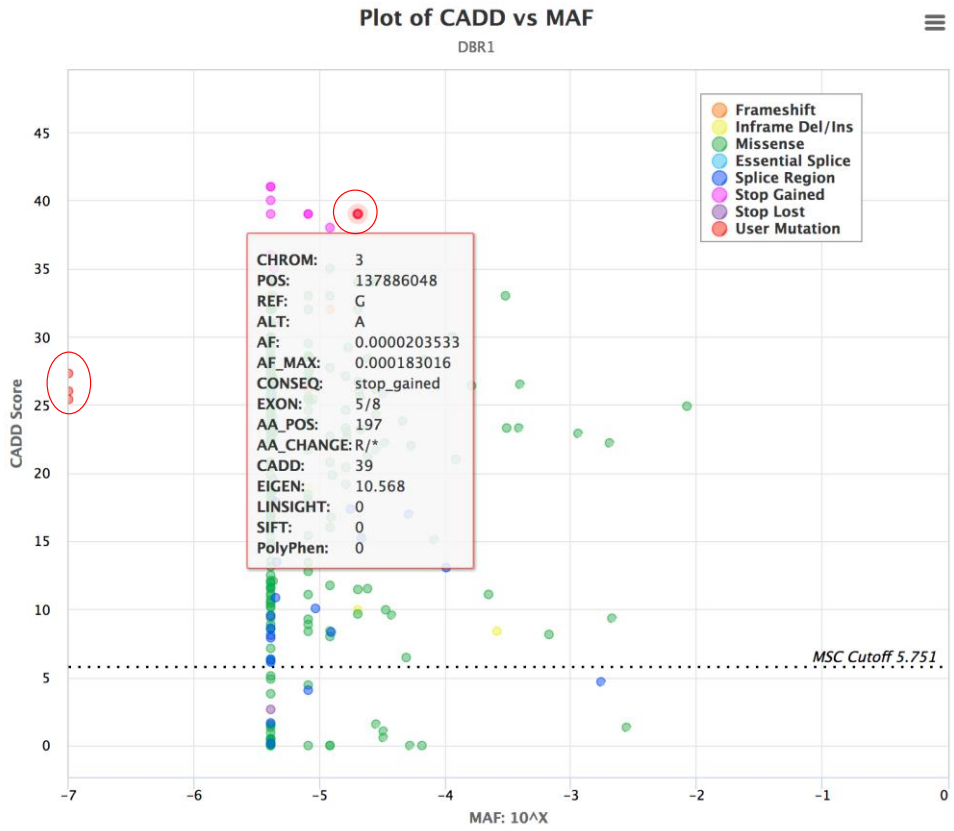
2.3.3. *DBR1* Mutations in Viral Encephalitis

Zhang et al. identified AR *DBR1* deficiency underlying the accumulation of RNA lariats, resulting in patient susceptibility to severe viral infections in the brainstem (Zhang, et al., 2018). In their study, they mapped all non-synonymous variants of *DBR1* from gnomAD, as well as the variants from their patients, onto the CADD vs MAF plot. The plot clearly shows that all the four patient-specific variations have very low MAF but high CADD scores beyond the gene-specific MSC of *DBR1*. Homozygosity or compound-heterozygosity of these variations would thus very likely lead to AR *DBR1* deficiency.

Supplementary Figure S8: CADD vs MAF plot of gene *DBR1* from Zhang's paper

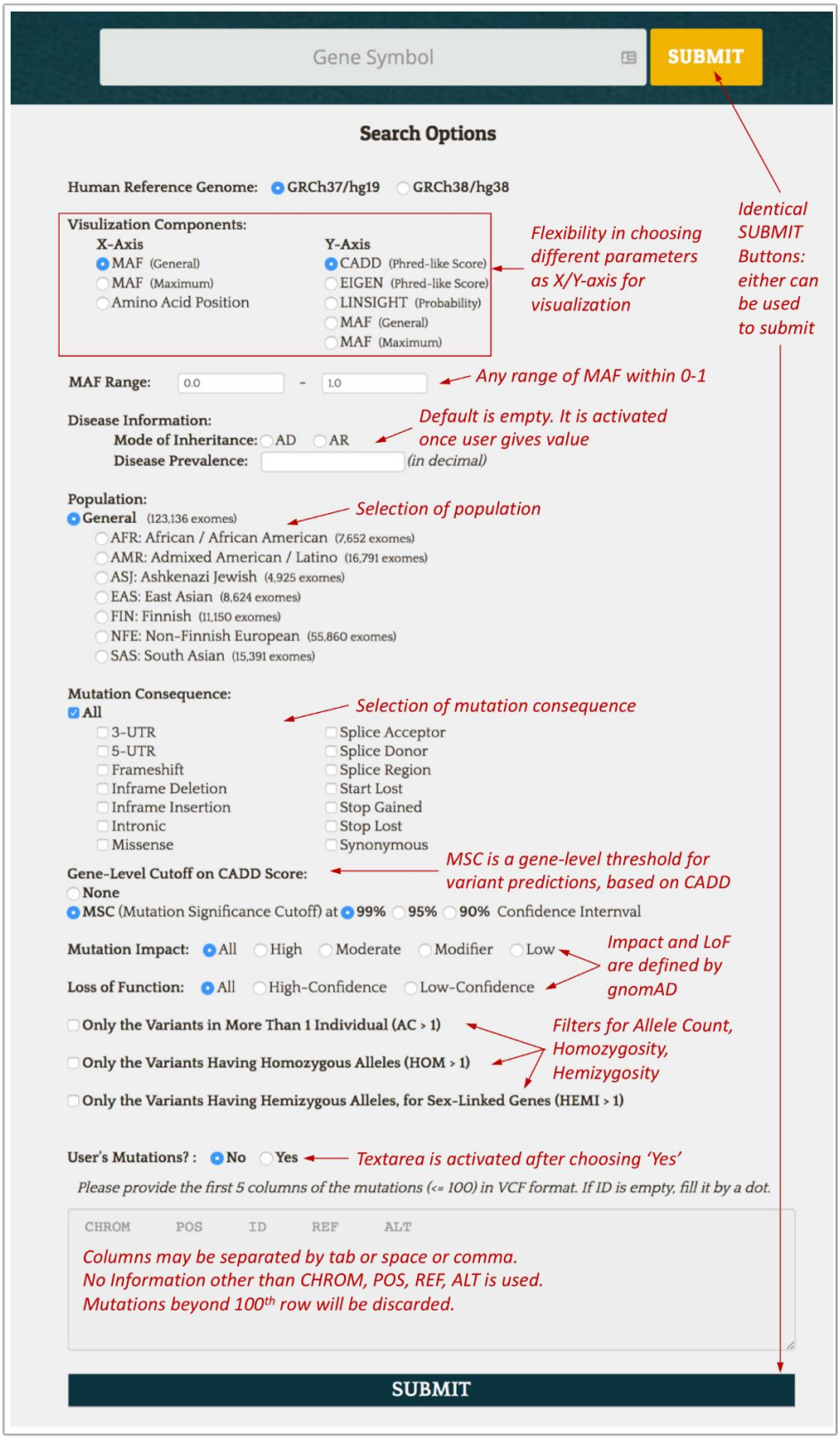


Supplementary Figure S9: CADD vs MAF plot of gene *DBR1* by PopViz, where the four mutations of the user are highlighted in red (one is matched in gnomAD, and three are private mutations)

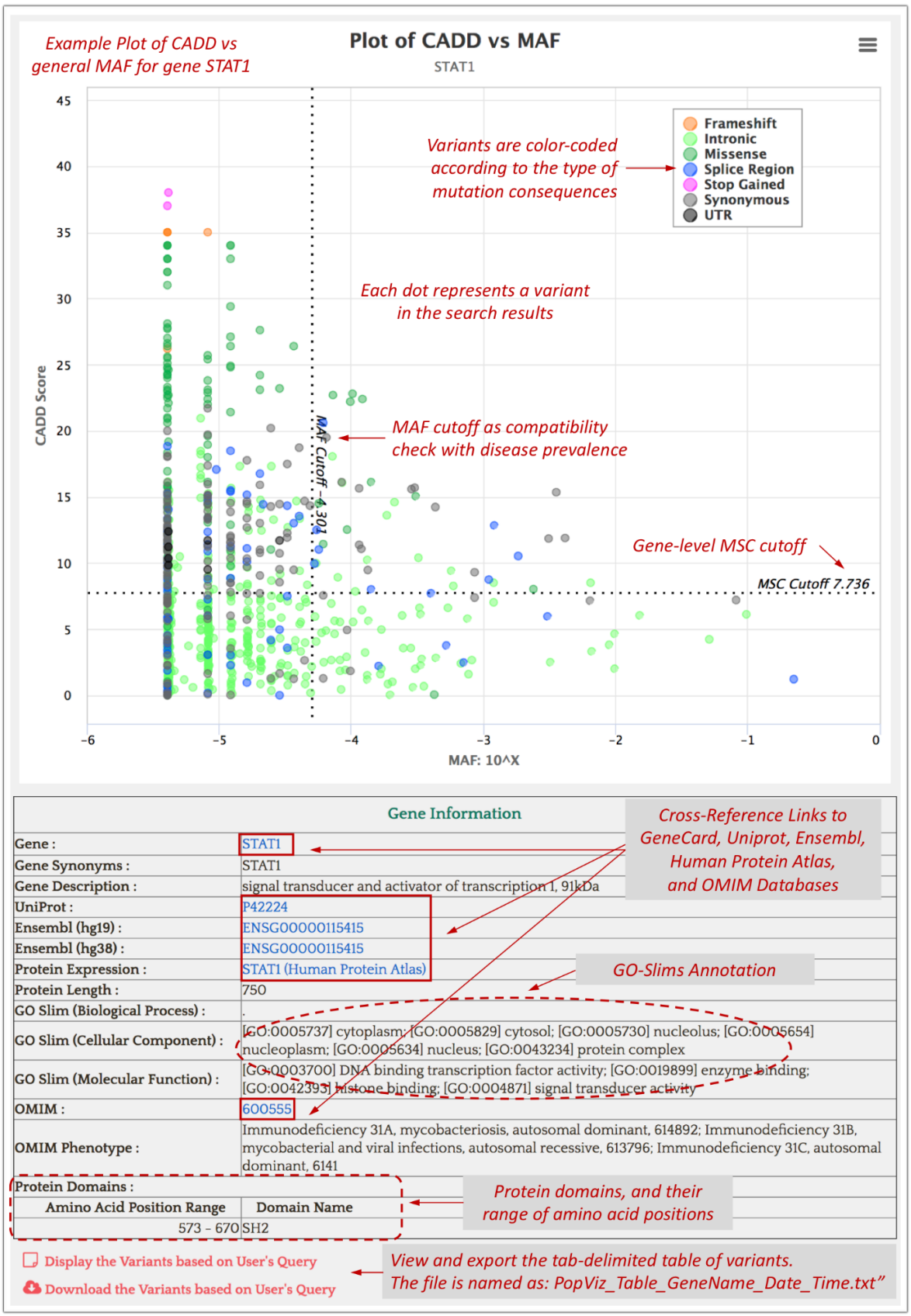


3. PopViz User Manual

Supplementary Figure S10: Search options in PopViz



Supplementary Figure S11: Example CADD vs MAF plot for gene *STAT1*



Gene Information	
Gene :	STAT1
Gene Synonyms :	STAT1
Gene Description :	signal transducer and activator of transcription 1, 91kDa
UniProt :	P42224
Ensembl (hg19) :	ENSG00000115415
Ensembl (hg38) :	ENSG00000115415
Protein Expression :	STAT1 (Human Protein Atlas)
Protein Length :	750
GO Slim (Biological Process) :	.
GO Slim (Cellular Component) :	[GO:0005737] cytoplasm; [GO:0005829] cytosol; [GO:0005730] nucleolus; [GO:0005654] nucleoplasm; [GO:0005634] nucleus; [GO:0043234] protein complex
GO Slim (Molecular Function) :	[GO:0003700] DNA binding transcription factor activity; [GO:0019899] enzyme binding; [GO:0042393] histone binding; [GO:0004871] signal transducer activity
OMIM :	600555
OMIM Phenotype :	Immunodeficiency 31A, mycobacteriosis, autosomal dominant, 614892; Immunodeficiency 31B, mycobacterial and viral infections, autosomal recessive, 613796; Immunodeficiency 31C, autosomal dominant, 6141
Protein Domains :	

Amino Acid Position Range	Domain Name
573 - 670	SH2

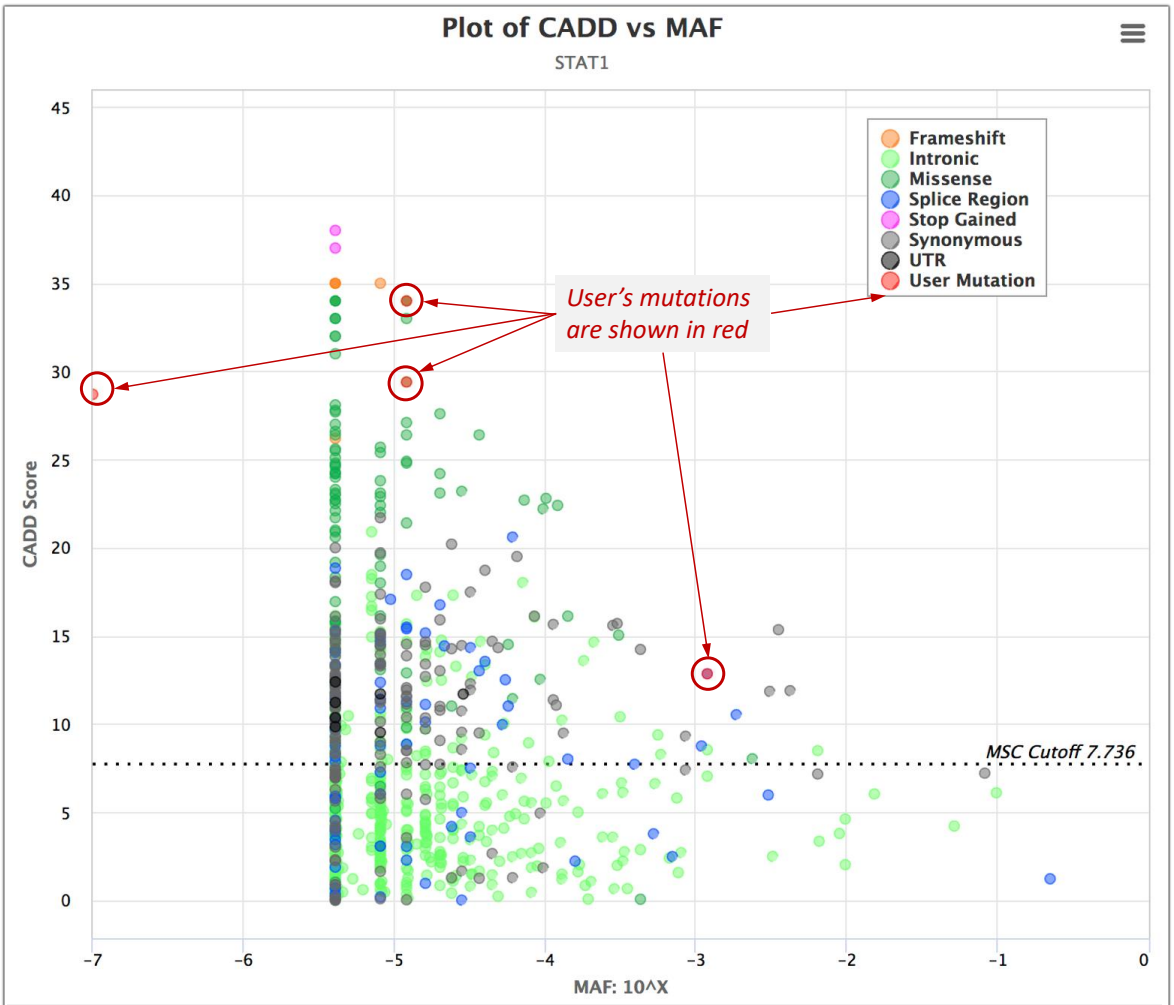
Supplementary Figure S12: Example CADD vs MAF plot for gene *STAT1*, with user's mutations

```
2       191859820       rs76397617      C       T
2       191843583       rs2230101       G       A
2       191862624       rs77937135      A       T

3       191859820       .       C       T
2       19184358300     .       G       C
2       191862624       .       A       C
```

3 true variants

3 mock variants

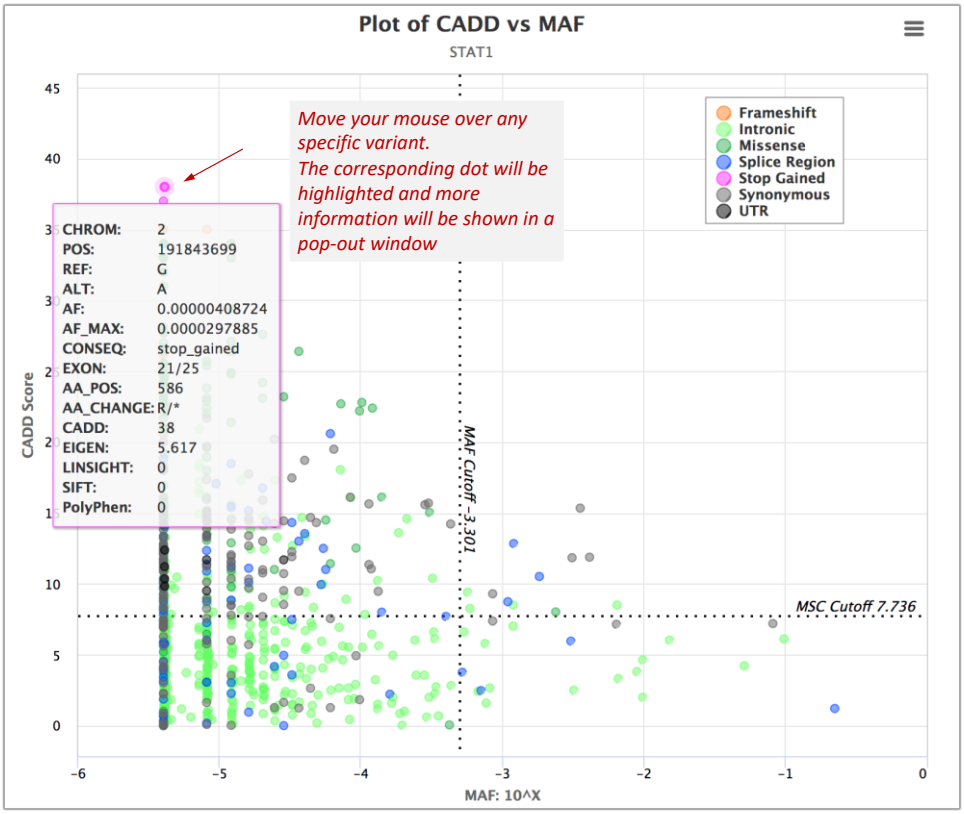


User's Mutations Information

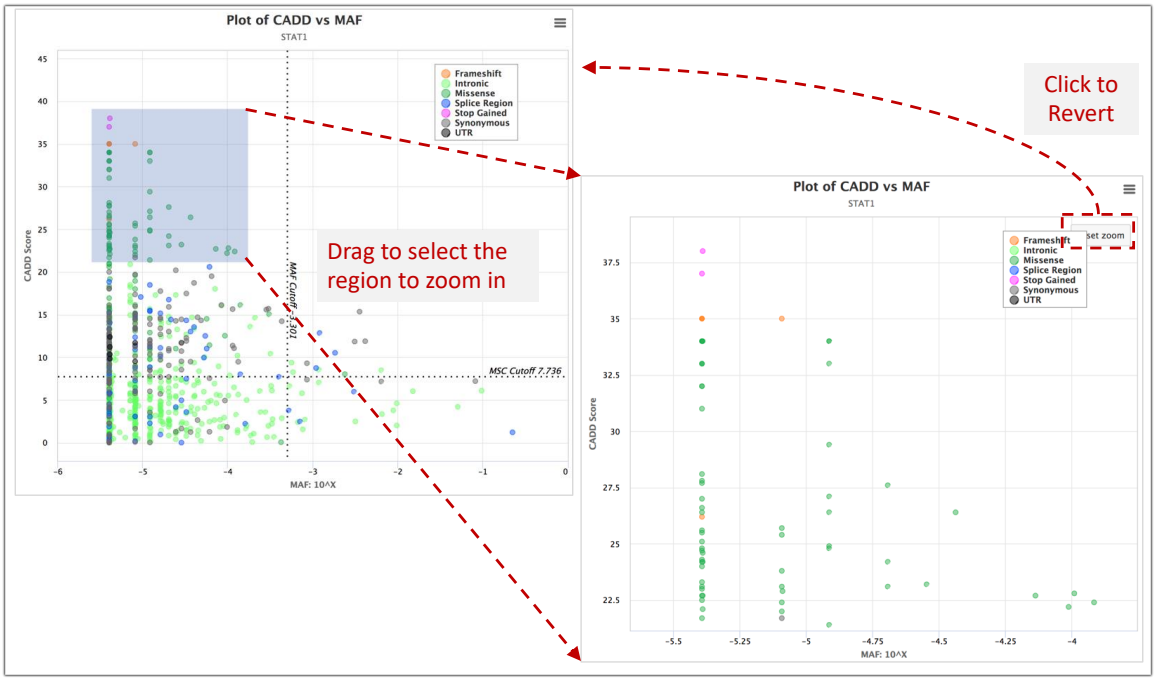
CHROM	POS	REF	ALT	MAF	CADD	Comment
2	191859820	C	T	-4.914	34	Exist in PopViz
2	191843583	G	A	-2.921	12.85	Exist in PopViz
2	191862624	A	T	-4.914	29.4	Exist in PopViz
3	191859820	C	T	.	.	Incorrect CHROM
2	19184358300	G	C	.	.	Incorrect POS
2	191862624	A	C	-7	28.7	Exist in PopViz

Supplementary Figure S13: Functions embedded in PopViz plot: (A) view more details for each variant, (B) zoom in and out, (C) show/hide specific consequences, and (D) export plot as an image

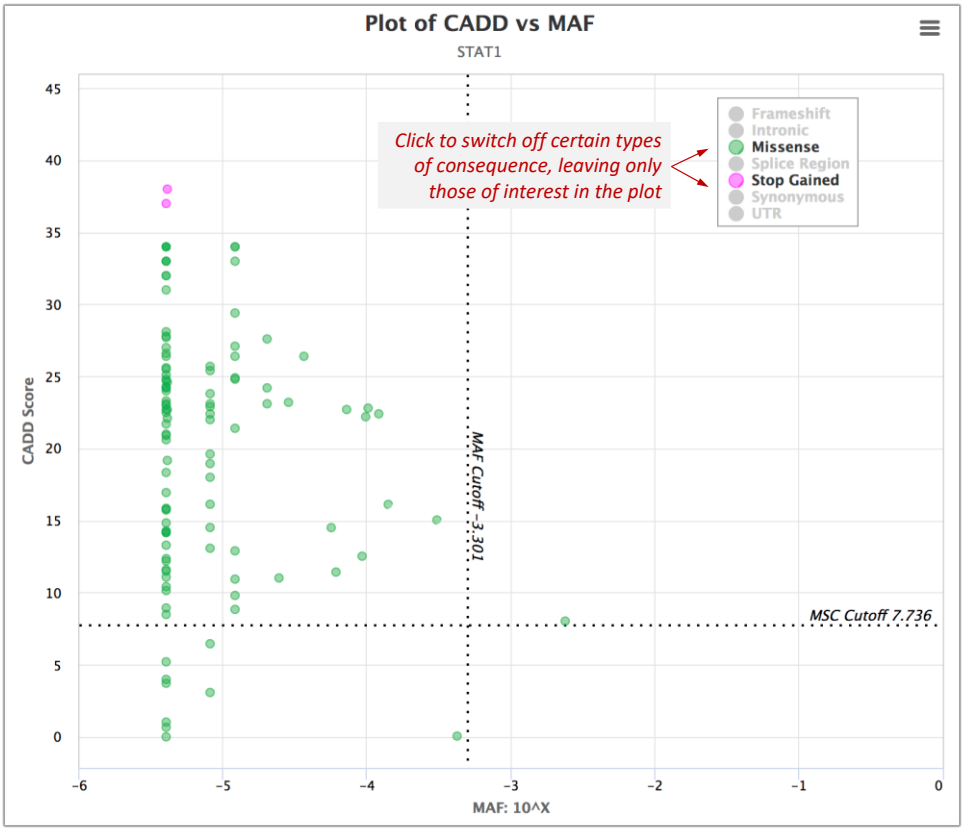
(A)



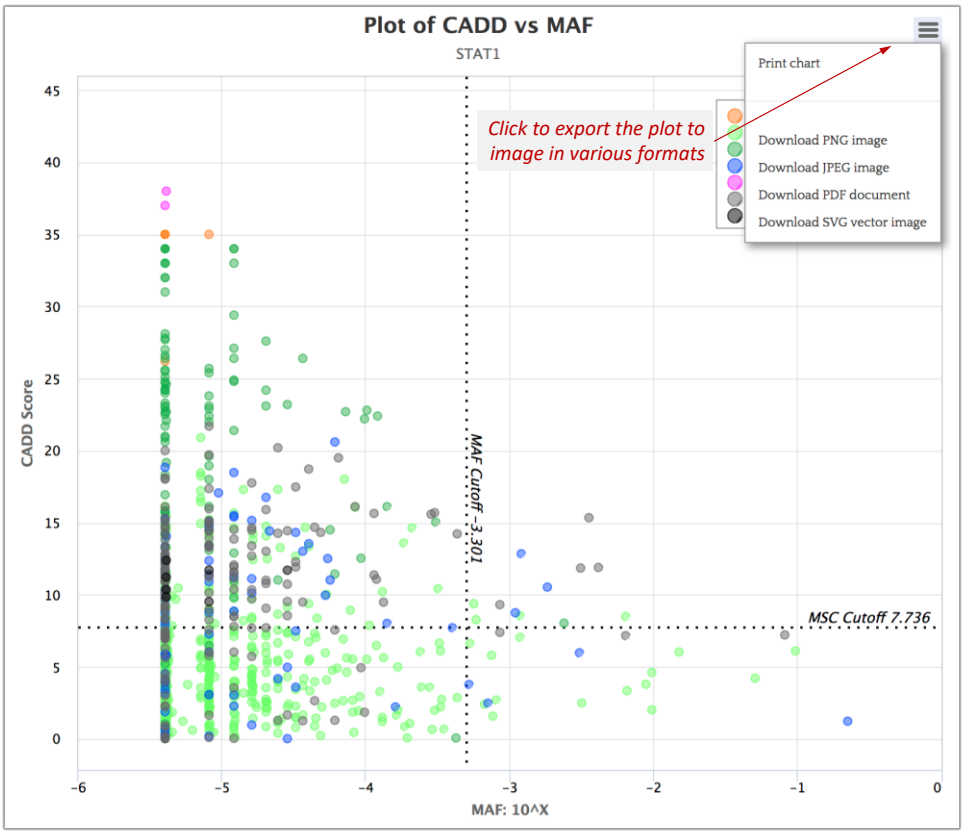
(B)



(C)


Plot of CADD vs MAF
STAT1
Click to switch off certain types of consequence, leaving only those of interest in the plot
Frameshift
Intronic
Missense
Splice Region
Stop Gained
Synonymous
UTR
MAF Cutoff -3.301
MSC Cutoff 7.736
CADD Score
MAF: 10^X

(D)


Plot of CADD vs MAF
STAT1
Click to export the plot to image in various formats
Print chart
Download PNG image
Download JPEG image
Download PDF document
Download SVG vector image
MAF Cutoff -3.301
MSC Cutoff 7.736
CADD Score
MAF: 10^X

Supplementary Figure S14: Color scheme for mutation consequences

Lastly, the PopViz webserver is also mobile-friendly, making it even more accessible and useful. PopViz enables the users to check any gene of interest during meetings and conferences quickly, visually, and conveniently. It can be accessed by scanning the QR code as below.

4. Reference

Adzhubei, I.A., *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7(4):248-249.

Folkman, L., *et al.* DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. *Bioinformatics* 2015;31(10):1599-1606.

Guerin, A., *et al.* IRF4 haploinsufficiency in a family with Whipple's disease. *Elife* 2018;7.

Huang, Y.F., Gulko, B. and Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* 2017;49(4):618-624.

Ionita-Laza, I., *et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;48(2):214-220.

Kircher, M., *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46(3):310-315.

Kuehn, H.S., *et al.* Loss of B Cells in Patients with Heterozygous Mutations in IKAROS. *N Engl J Med* 2016;374(11):1032-1043.

Kumar, P., Henikoff, S. and Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4(7):1073-1081.

Lek, M., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536(7616):285-291.

Zhang, S.Y., *et al.* Inborn Errors of RNA Lariat Metabolism in Humans with Brainstem Viral Infection. *Cell* 2018;172(5):952-965 e918.