Supplementary Information for

Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis

Patrick Maffucci[#], Benedetta Bigio[#], Franck Rapaport, Aurélie Cobat, Alessandro Borghesi, Marie Lopez, Etienne Patin, Alexandre Bolze, Lei Shang, Matthieu Bendavid, Eric M Scott, Peter D Stenson, Charlotte Cunningham-Rundles, David N Cooper, Joseph G Gleeson, Jacques Fellay, Lluis Quintana-Murci, Jean-Laurent Casanova, Laurent Abel, Bertrand Boisson[‡], Yuval Itan[‡]

[#,‡]: Equal contributions

Corresponding authors:
Jean-Laurent Casanova – casanova@rockefeller.edu,
Yuval Itan – yuval.itan@mssm.edu

This PDF file includes:
Figs. S1 to S19
Tables S1 to S10

www.pnas.org/cgi/doi/10.1073/pnas.1808403115

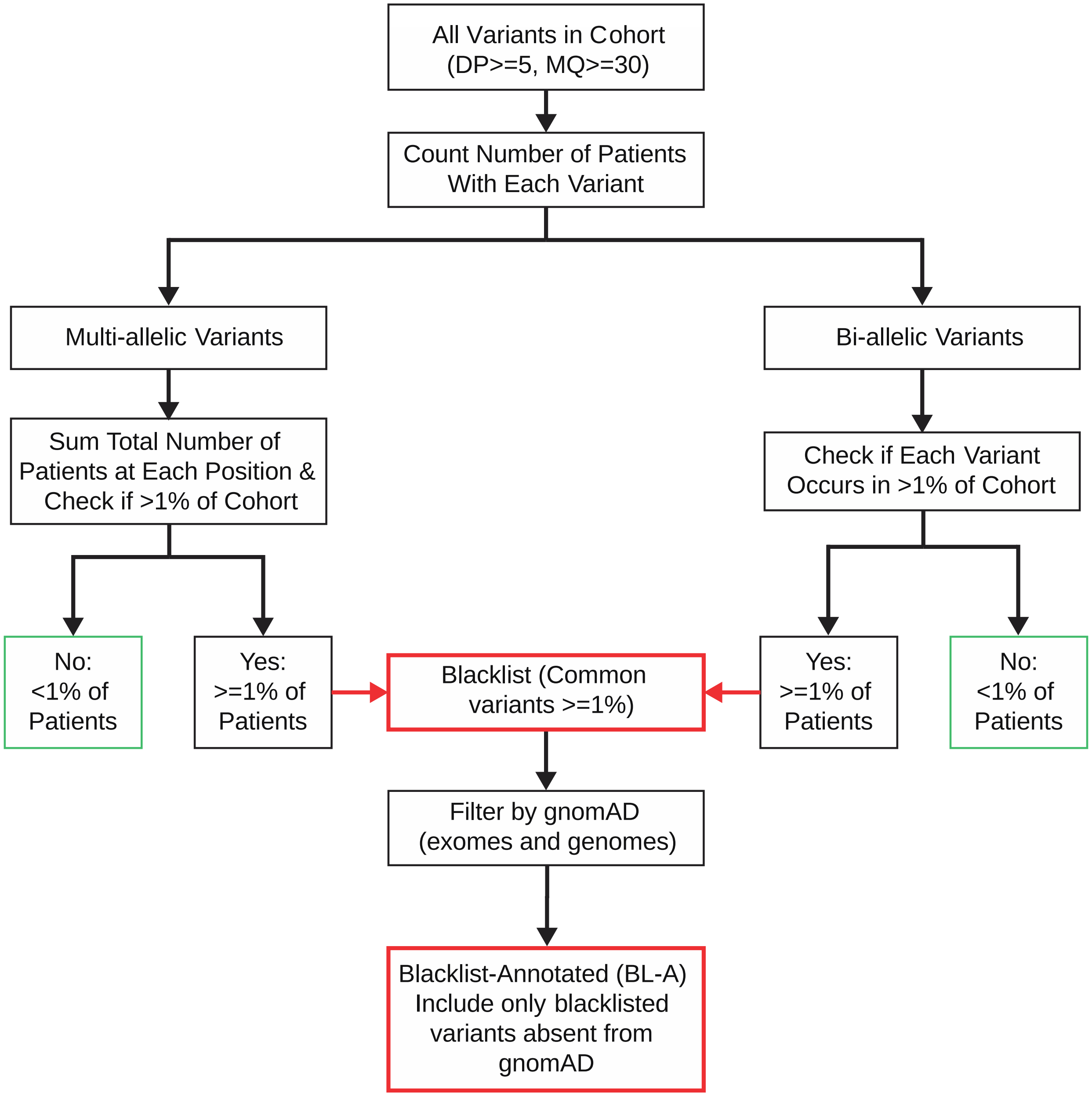

All Variants in Cohort (DP>=5, MQ>=30)
Count Number of Patients With Each Variant
Multi-allelic Variants
Bi-allelic Variants
Sum Total Number of Patients at Each Position & Check if >1% of Cohort
Check if Each Variant Occurs in >1% of Cohort
No: <1% of Patients
Yes: >=1% of Patients
Blacklist (Common variants >=1%)
Yes: >=1% of Patients
No: <1% of Patients
Filter by gnomAD (exomes and genomes)
Blacklist-Annotated (BL-A) Include only blacklisted variants absent from gnomAD

Figure S1. **Methodology for blacklist generation.** The blacklist was generated by first collecting unique high-quality variants (DP>=5, MQ>=30) from patient exomes and counting the occurrence of each variant. These variants were assembled into two classes: (1) biallelic, with a single alternative allele in our cohort; and (2) multiallelic, with two or more alternative alleles in the cohort, for which we collapsed all variants at a unique chromosomal position and summed the total number of patients containing these variants. We then collected the variants that had a frequency >=1% in the cohort (the Blacklist: "Common in-house variants"). Of these variants, 21.4% (167,144) were absent from gnomAD exome and genome databases. We considered these 167,144 variants to be "blacklist-annotated" (BL-A).

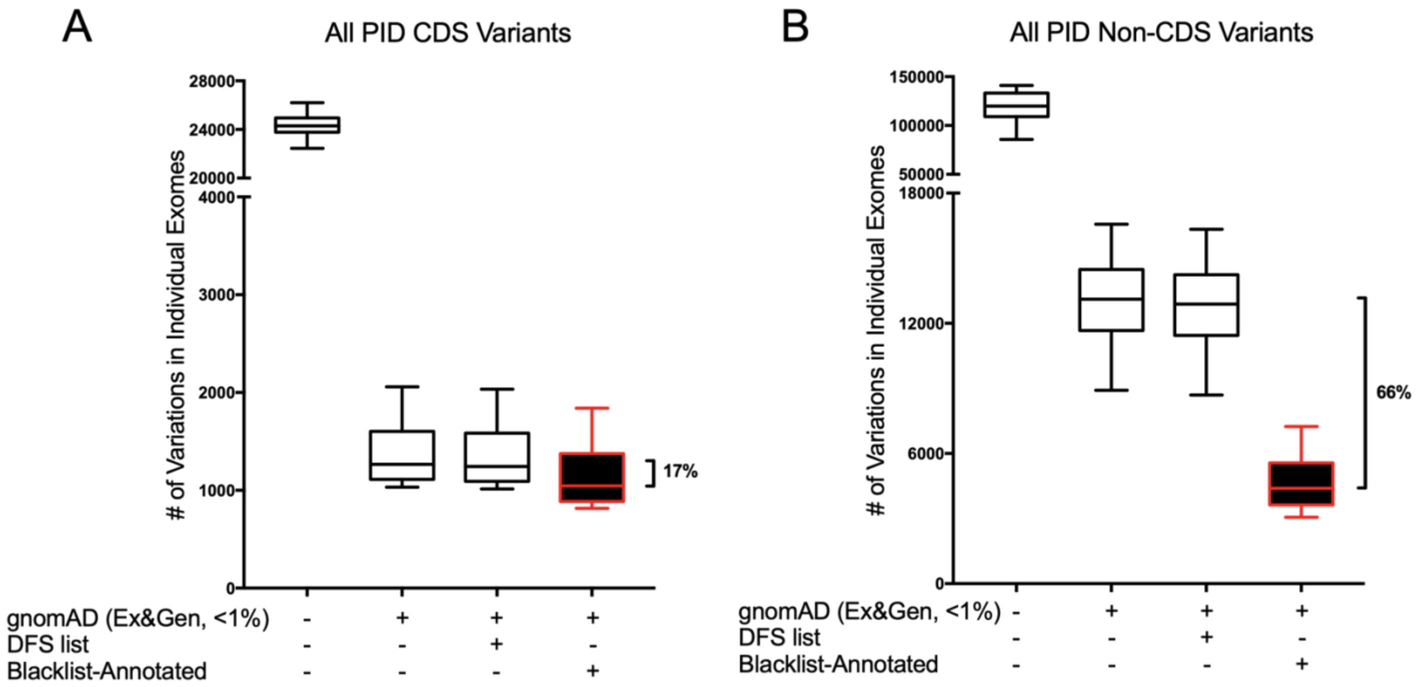


Figure S2. Filtering of coding sequence (CDS) or non-CDS variants in 3,104 PID exomes with the PID blacklist-annotated. Exomes were restricted to CDS (A) or non-CDS (B) variants and filtered by removing variants with a MAF greater than 0.01 in gnomAD. The remaining variants were filtered with the blacklist-annotated. Filtering with the DFS list is shown for comparison. Error bars represent the 10th-90th percentiles.

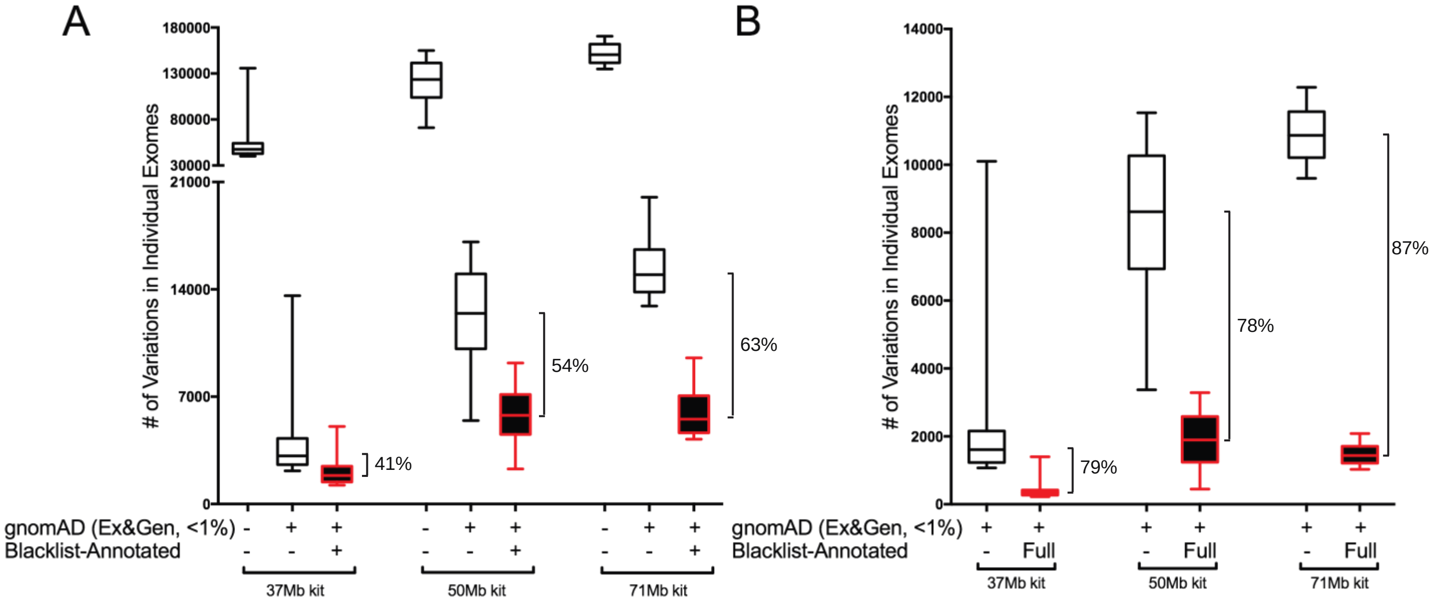


Figure S3. Filtering of 3,104 PID exomes broken down by the exome capture kit. PID exomes were captured with one of three SureSelect kits: 37 Mb (*n* = 96), 50 Mb (*n* = 727), or 71 Mb (*n* = 2,281). (A) Filtering of all variants in each exome, using gnomAD and the blacklist-annotated. gnomAD filtering performed by removing variants with a minor allele frequency greater than 0.01 in the databases. (B) Filtering of exomes restricted to cohort-specific variants with the blacklist-annotated. Error bars represent the 10th-90th percentiles.

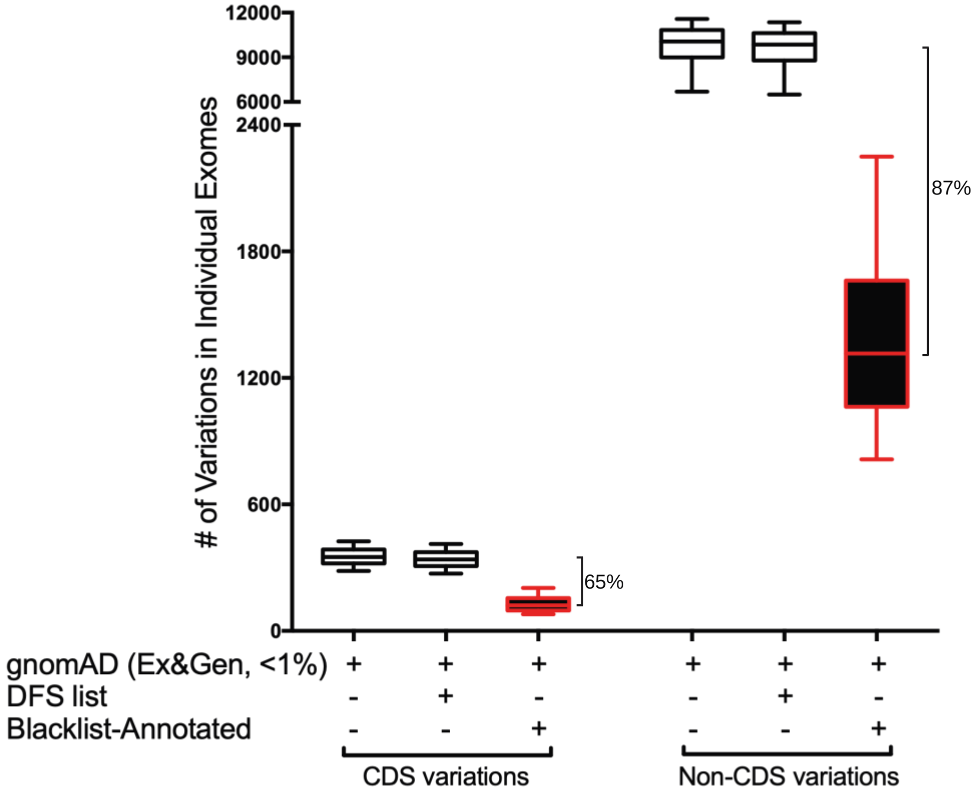


Figure S4. Filtering of coding sequence (CDS) and non-CDS variants in 3,104 PID exomes restricted to cohort-specific variations using the blacklist-annotated. DFS list shown for comparison. Error bars represent the 10th-90th percentiles.

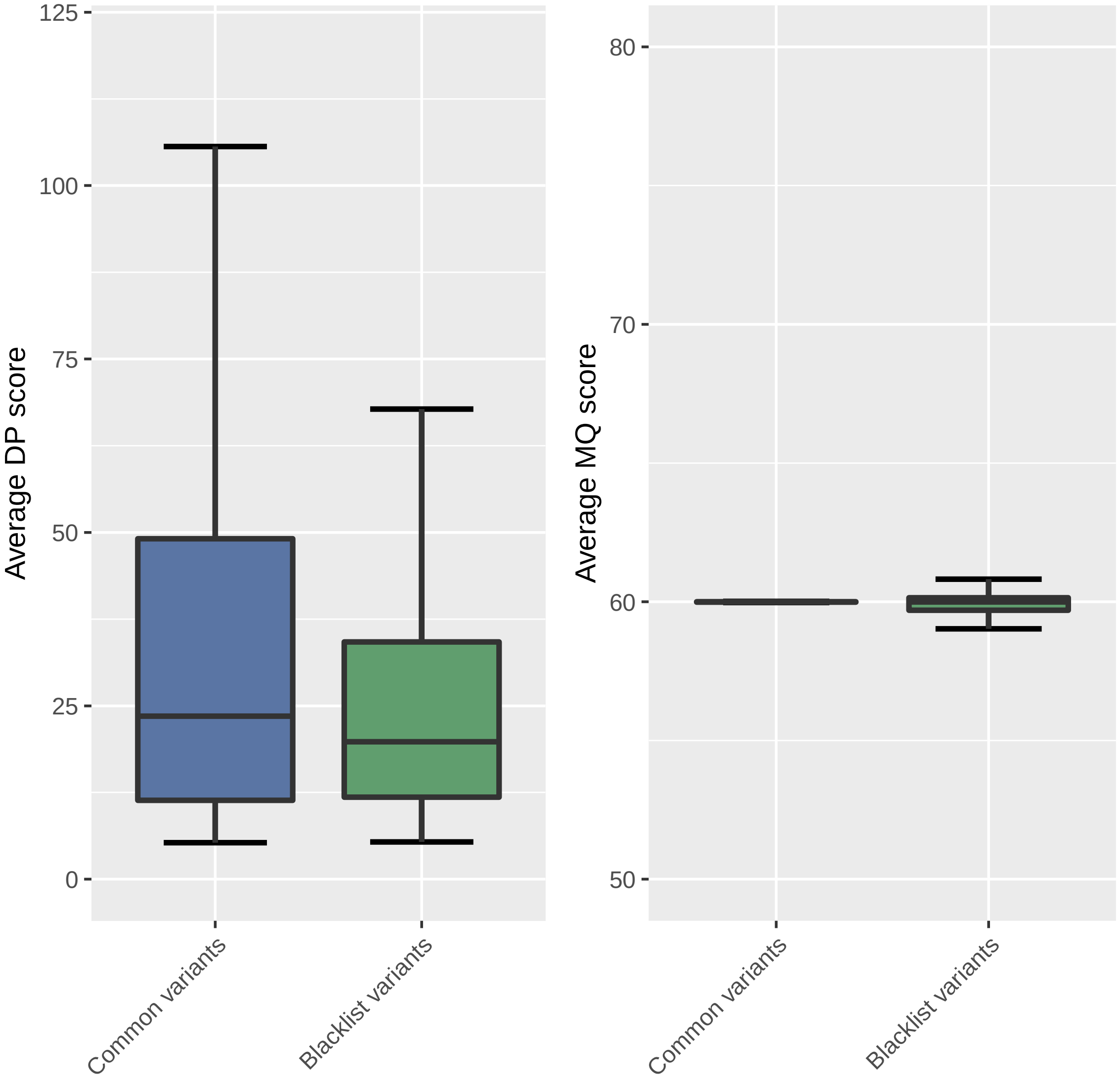


Figure S5. **Comparison of quality metrics for blacklisted and non-blacklisted variants.** Mean (A) read depth (DP) and (B) mapping quality (MQ) were calculated for common variants present in gnomAD with a MAF>1% (blue bar), and for blacklist-annotated variants (green bar). Error bars represent the upper and lower limits of 1.5 times the interquartile range.

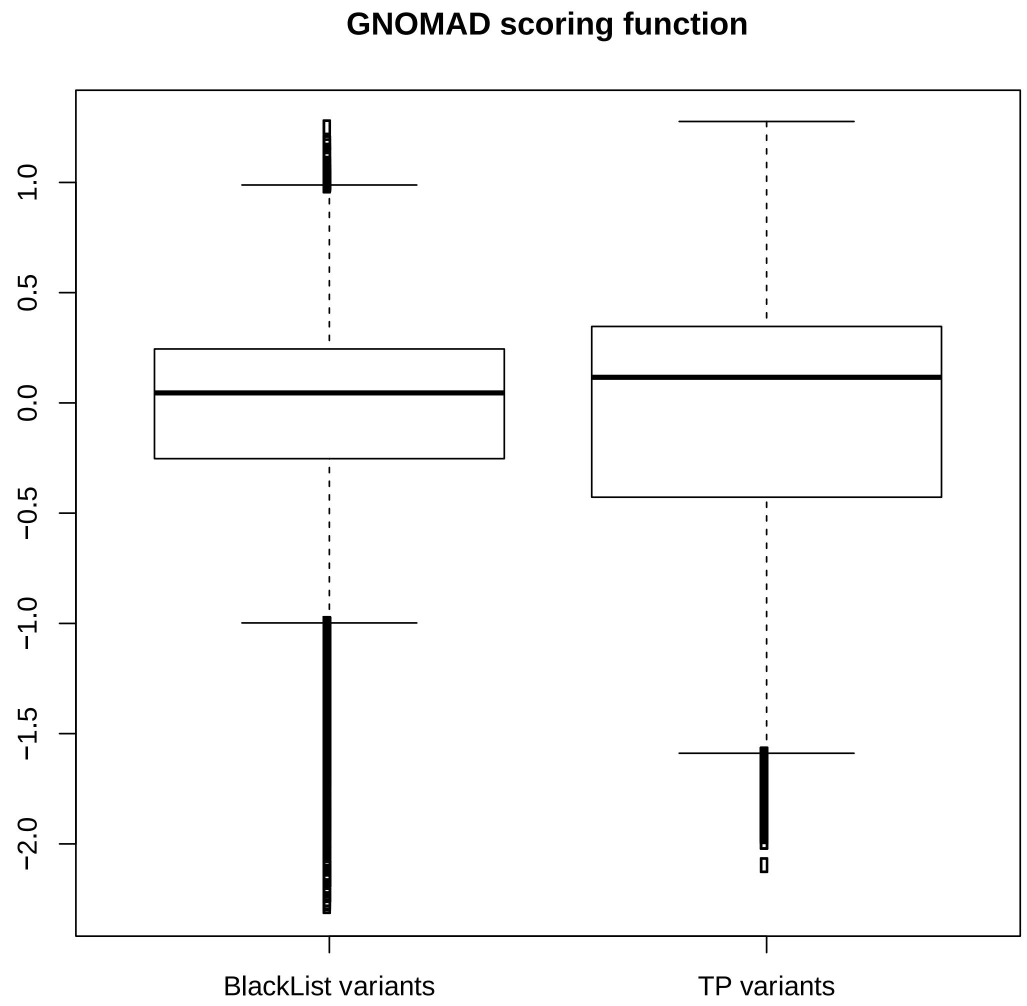


Figure S6. **Comparison with machine learning-based filtering methods.** We applied random forest scoring functions to blacklist-annotated variants and to a set of true-positive (TP) variants present in both the gnomAD dataset and our cohort with a MAF exceeding 1% in each dataset. The score distributions are almost identical, indicating that the blacklist-annotated variants are not distinguishable from TP variants according to this standard classification method.

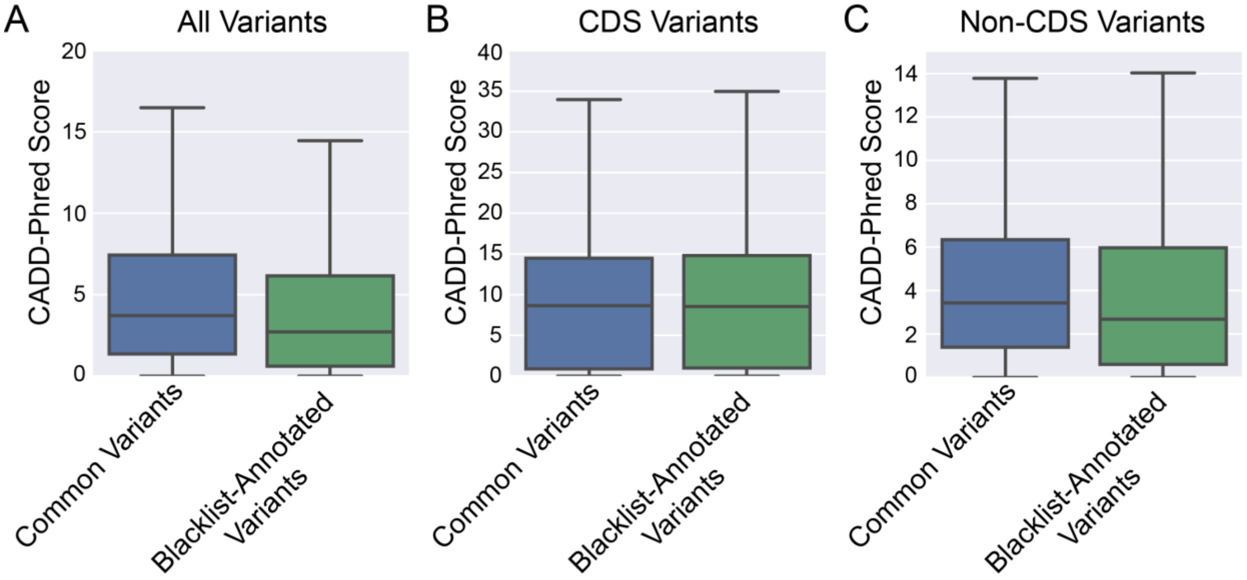


Figure S7. Comparison of CADD scores between blacklisted and non-blacklisted variants. Mean CADD scores were calculated for common variants present in gnomAD exome and genome databases with a MAF>1% (blue bar), or blacklist-annotated variants (green bar). Calculations were performed for all (A), CDS (B), and non-CDS (C) variants. Error bars represent the upper and lower limits of 1.5 times the interquartile range.

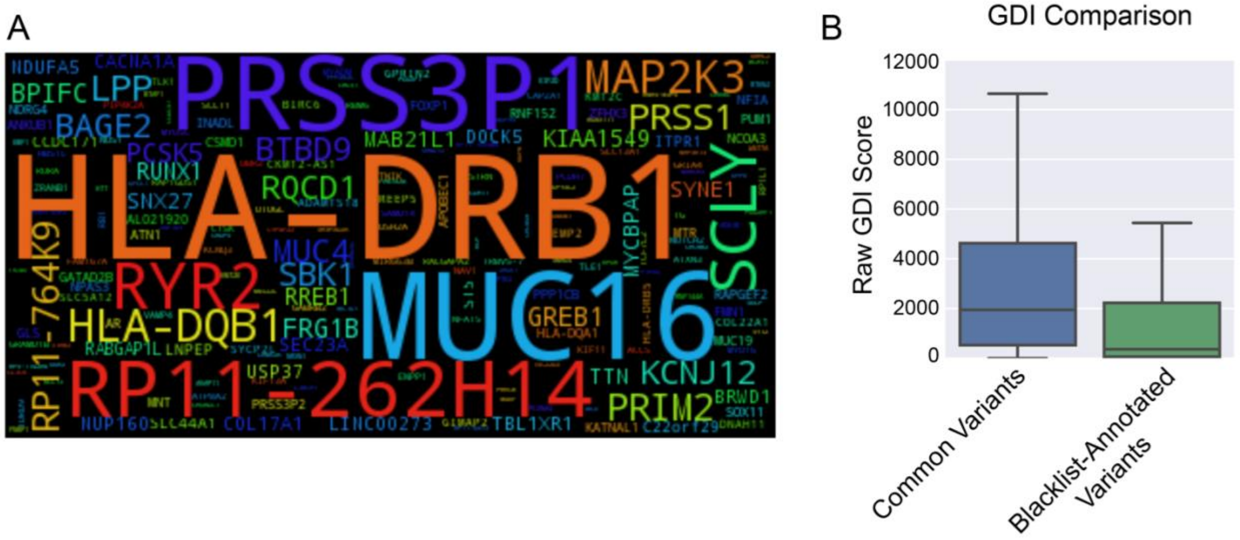


Figure S8. Characteristics of the most frequent genes in the blacklist-annotated. (A) Depiction of the top ranking genes in the blacklist-annotated according to the number of variants. The size of the text is proportional to the number of variants of the gene in the blacklist-annotated. (B) Comparison of GDI scores between the 1,000 most common genes in all the common in-house variants (gnomAD) and blacklist-annotated variants. Error bars represent the upper and lower limits of 1.5 times the interquartile range.

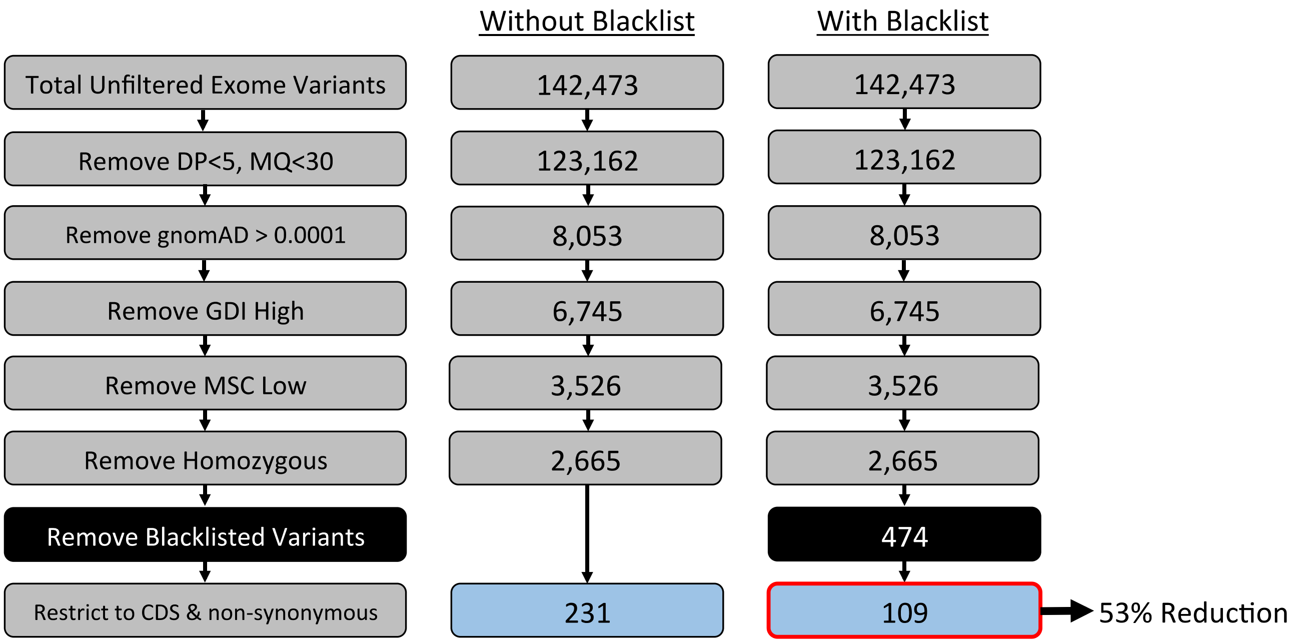


Figure S9. Practical analysis of a single patient exome by blacklisting. The practical utility of the blacklist approach was demonstrated with the exome of a patient with a published disease-causing mutation. The patient's exome was filtered with a standard pipeline with and without application of the blacklist-annotated. The numbers in each box represent the number of variants remaining in the exome after each filtering step. GDI: gene damage index; MSC: mutation significance cutoff.

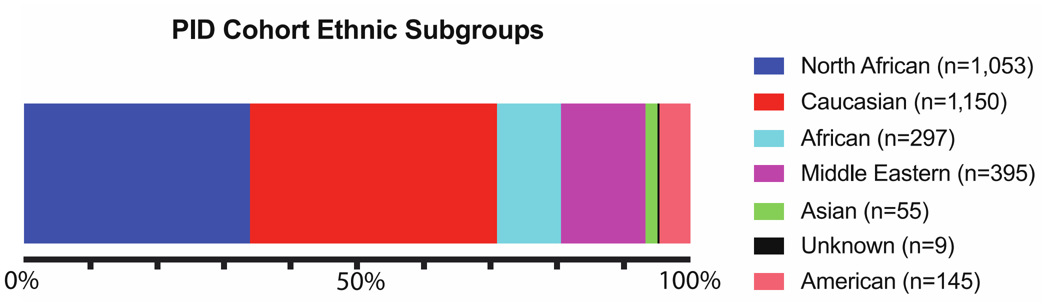


Figure S10. Representation of ethnic subgroups in 3,104 PID exomes. The distribution of the genetic ancestry groups in the PID cohort, as determined by PCA analysis.

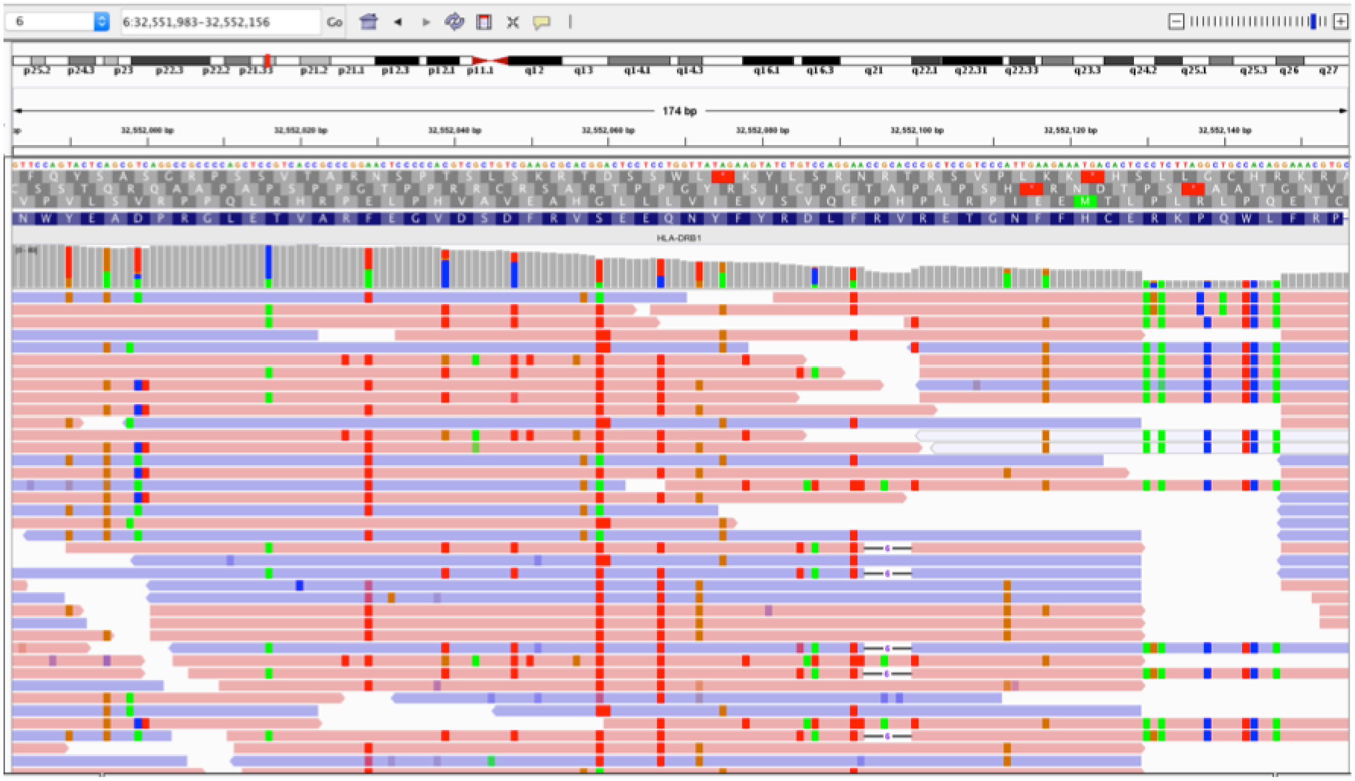
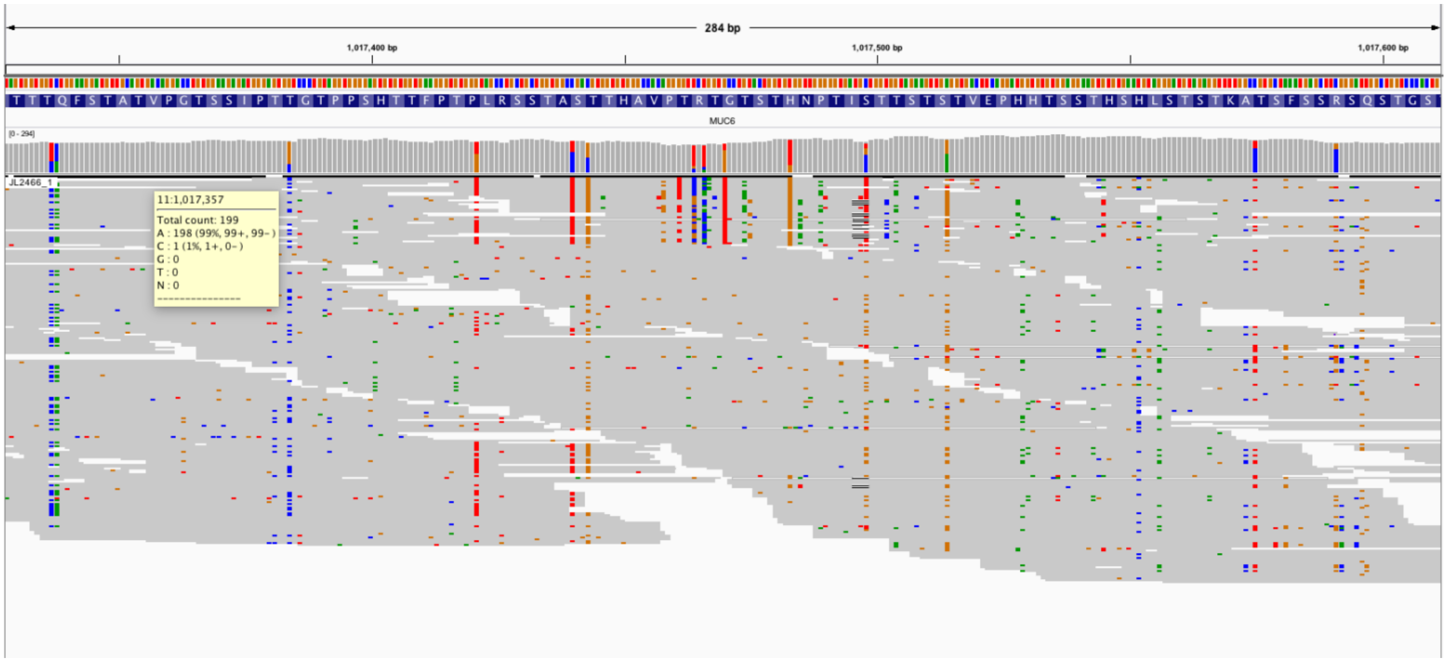


Figure S11. **Investigation of a biallelic *HLA-DRB1* variant: 6-32551960-T-TCC**
IGV screenshot of the WES alignment surrounding position 32,551,960 on chromosome 6.


284 bp
1,017,400 bp
1,017,500 bp
1,017,600 bp
T T T Q F S T A T V P G T S S I P T T G T P P S H T T F P T P L R S S T A S T T H A V P T R T G T S T H N P T I S T T S T S T V E P H H T S S T H S H L S T S T K A T S F S S R S Q S T G S
MUC6
[0 - 294]
JL2466_1

```
11:1,017,357
Total count: 199
A : 198 (99%, 99+, 99- )
C : 1 (1%, 1+, 0- )
G : 0
T : 0
N : 0
---------------
```

Figure S12. Investigation of a biallelic *MUC6* variant: 11-1017470-G-T
IGV screenshot of the WES alignment surrounding position 1,017,280 on chromosome 11.

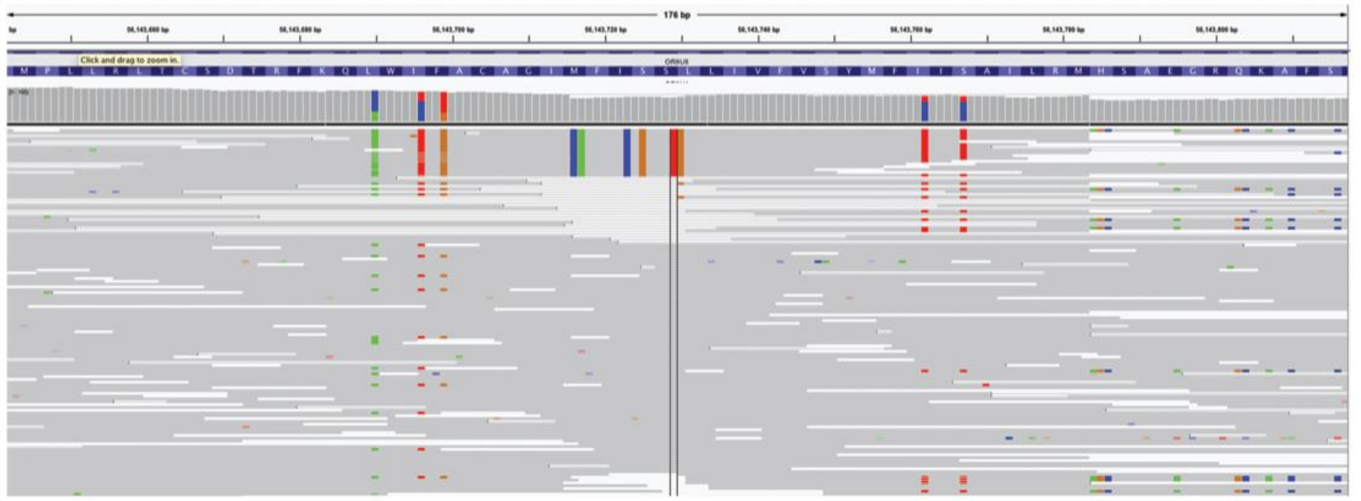

176 bp
56,143,660 bp
56,143,680 bp
56,143,700 bp
56,143,720 bp
56,143,740 bp
56,143,760 bp
56,143,780 bp
56,143,800 bp
Click and drag to zoom in.
OR8U8
M P L L R L T C S D T R F K Q L W I F A C A G I M F I S S L L I V F V S Y M F I I S A I L R M H S A E G R Q K A F S

Figure S13. Investigation of biallelic *OR8U1* variants: 11,56143784,C,T and 11,56143803,A,G
IGV screenshot of the WES alignment surrounding position 11,56143784 on chromosome 11.

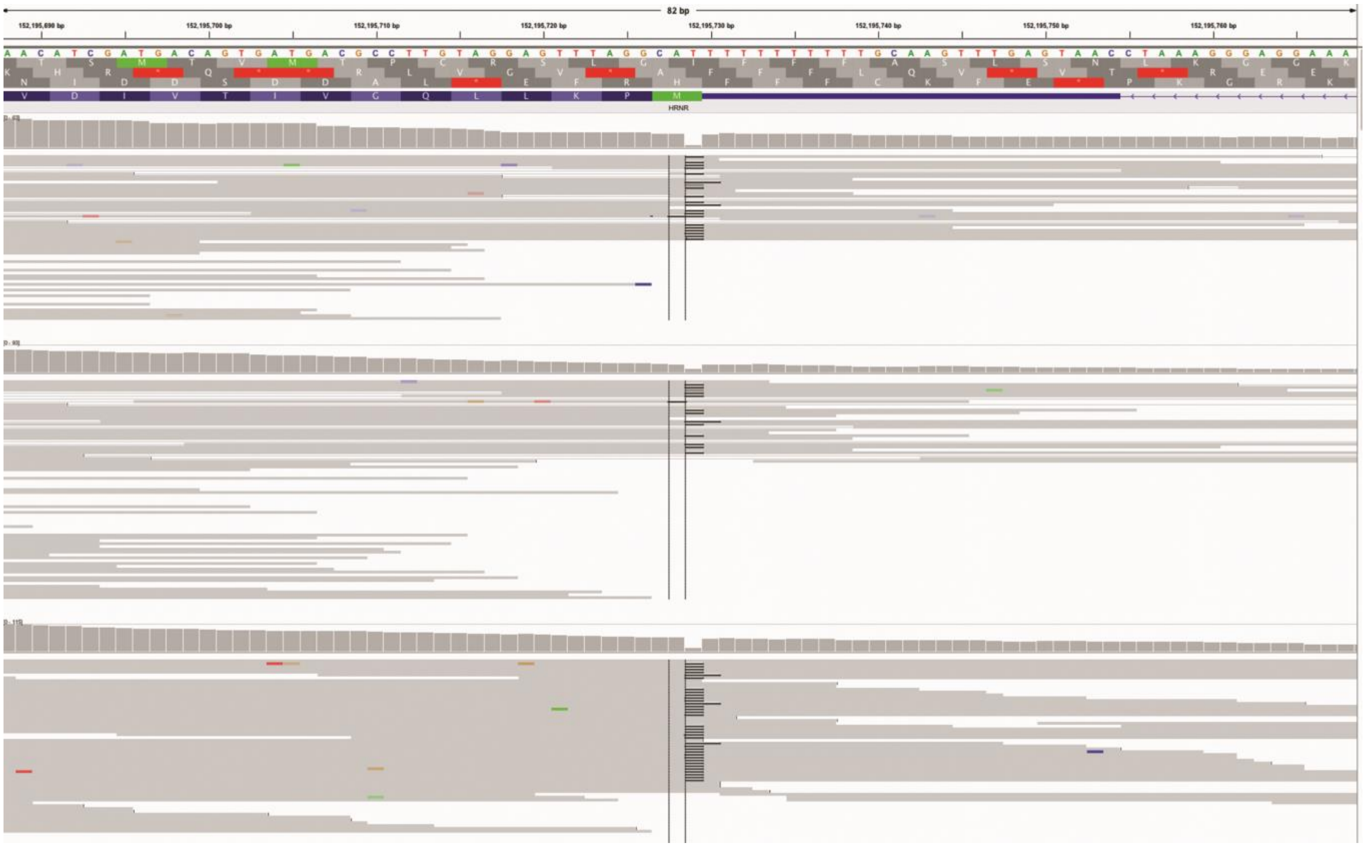

82 bp
152,195,690 bp
152,195,700 bp
152,195,710 bp
152,195,720 bp
152,195,730 bp
152,195,740 bp
152,195,750 bp
152,195,760 bp
A A C A T C G A T G A C A G T G A T G A C G C C T T G T A G G A G T T T A G G C A T T T T T T T T T T T G C A A G T T T G A G T A A C C T A A A G G G A G G A A A
V D I V T I V G Q L L K P M
HRNR

Figure S14. Investigation of a biallelic *HRNR* variant: 1-152195728-AT-A
IGV screenshot of the WES alignment at position 152195728 on chromosome 1.

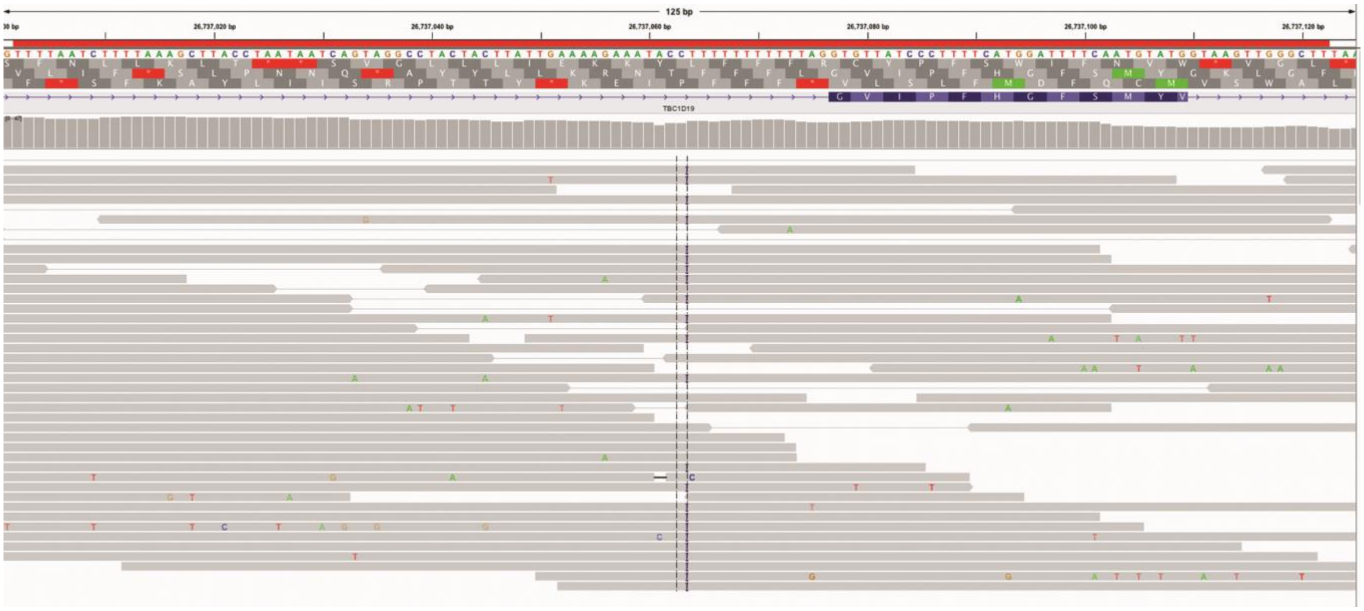


Figure S15. Investigation of a multiallelic *TBC1D19* variant: 4-26737063-C-CT
IGV screenshot of the WES alignment at position 26737063 on chromosome 4.

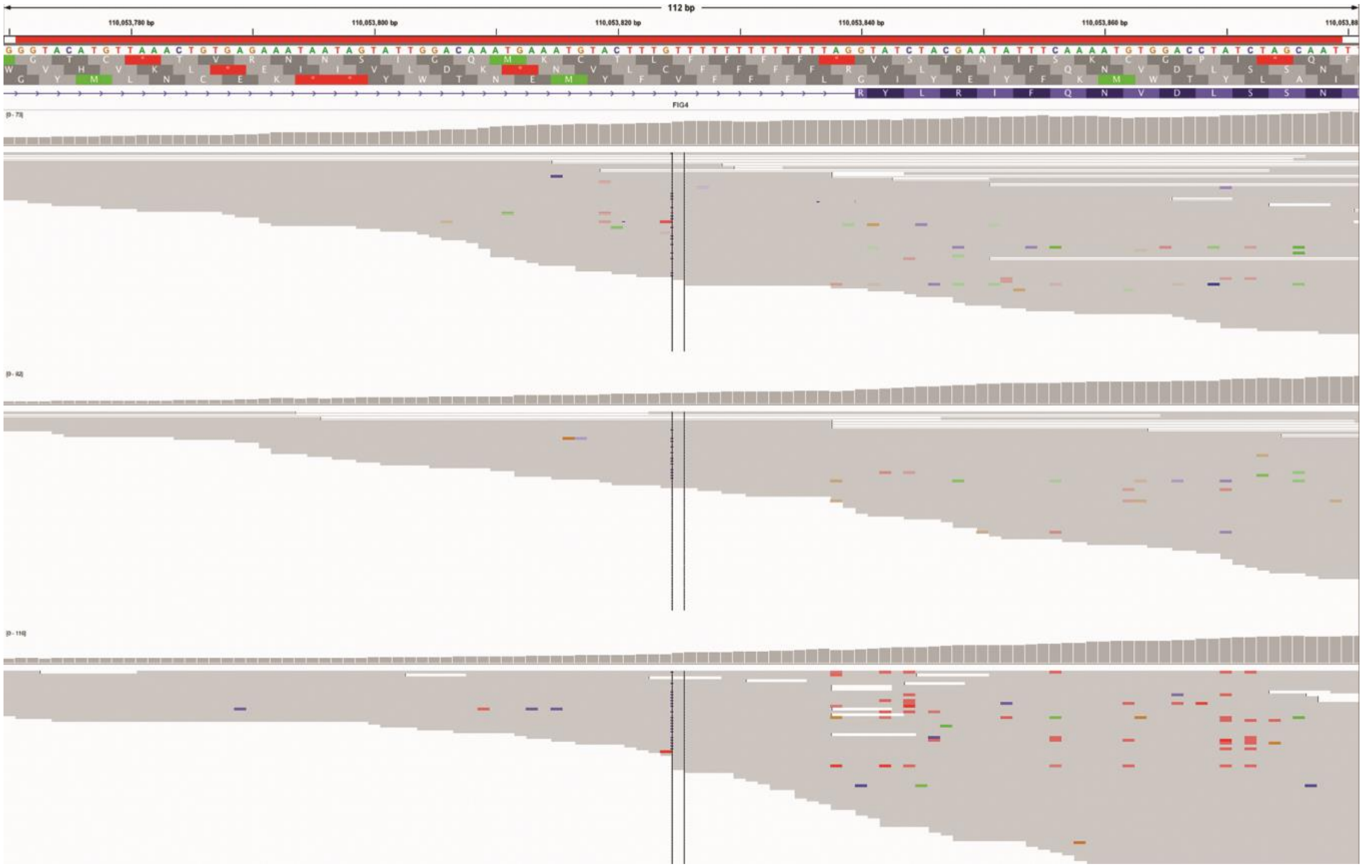


Figure S16. Investigation of a multiallelic *FIG4* variant: 6-110053824-G-GT
IGV screenshot of the WES alignment at position 110053824 on chromosome 6.

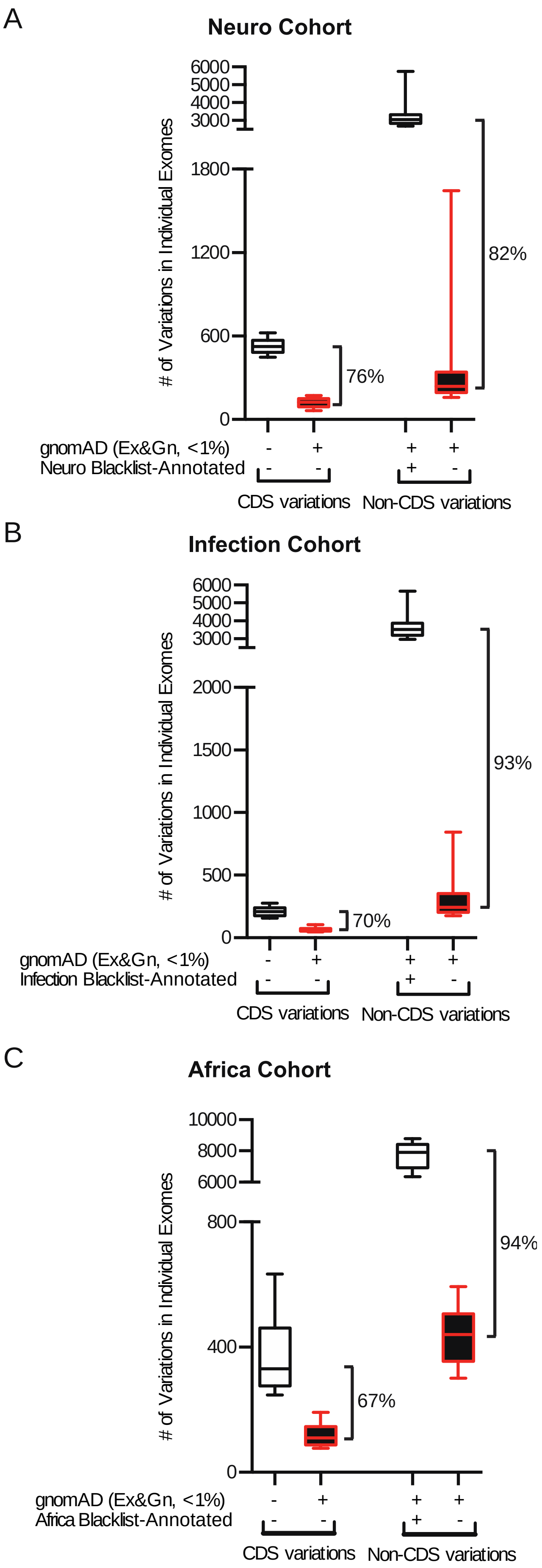


Figure S17. Filtering of coding and non-coding sequence variants in (A) 3,869 Neuro exomes restricted to cohort-specific variants with the Neuro blacklist-annotated, (B) 902 Infection exomes restricted to cohort-specific variants with the Infection blacklist-annotated, (C) 400 Africa exomes restricted to cohort-specific variants with the Africa blacklist-annotated. Error bars represent the 10th-90th percentiles.

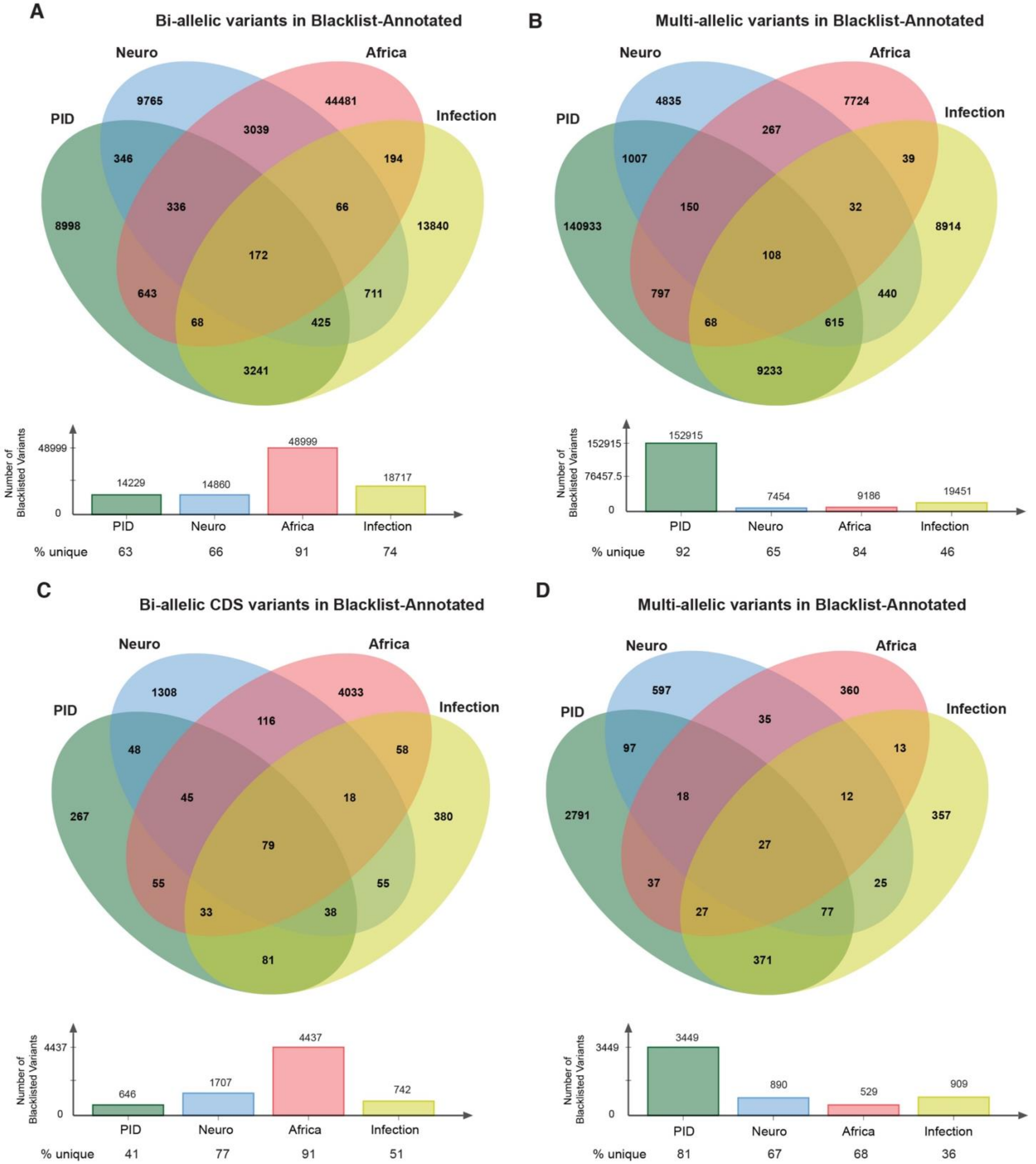


Figure S18. Relationship between the four blacklists. Common and unique biallelic (A), multiallelic (B), biallelic restricted to CDS (C), and multiallelic restricted to CDS (D) variants from the Blacklist-Annotated in the PID, Neuro, Africa and Infection cohorts.

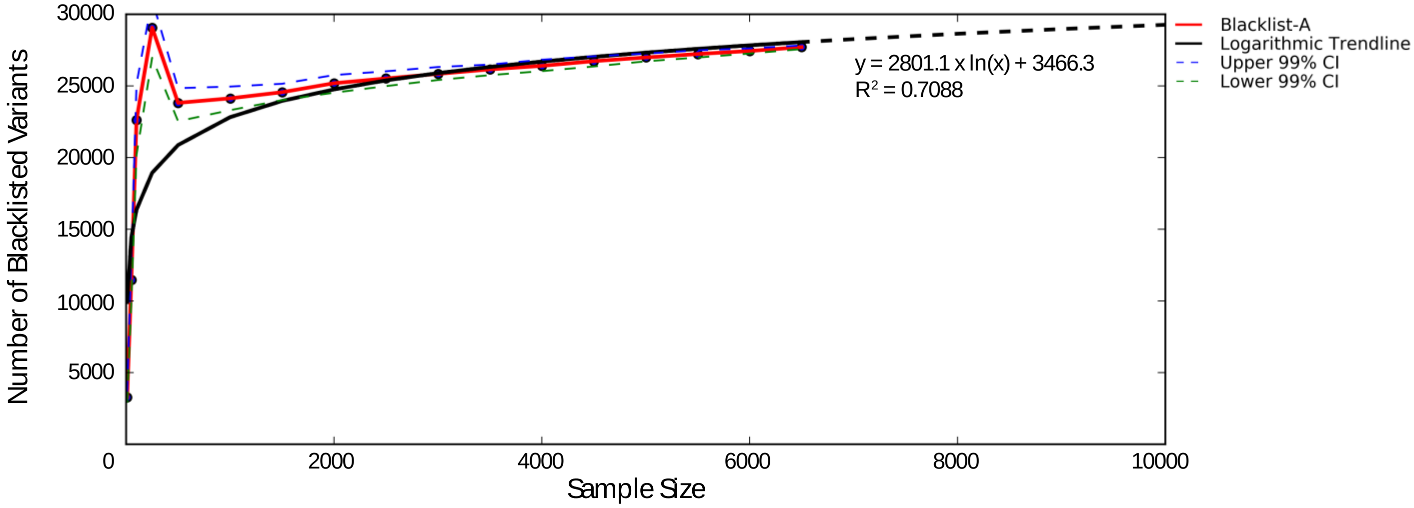


Figure S19. Relationship between sample size and number of blacklist variants. Estimation of the number of exomes required to create a saturated blacklist for CDS variants. Overlays in red, gray and green indicate that blacklist generation is unsafe, safe and optimal, respectively. The green vertical line indicates the suggested minimal sample size.

Table S1. VQSR status of blacklist-annotated (BL-A) variants

	# of VQSR PASS (%)	# of VQSR non-PASS (%)
Blacklist	125,614 (75.2%)	41,530 (24.8%)

Table S2. Blacklist-annotated variants in HGMD or ClinVar database

Chr	Position	Ref.	Alt.	HGMD	ClinVar	gnomAD	Gene	Disease	Status	Consequence	cDNA	Protein	rs ID	Publication (PMID)
4	88929173	C	CGAG	x		PASS	PKD2			inframe insertion	c.307_308insAGG	p.Glu102dup	rs547253972	
8	100844596	G	T	x	x	-	VPS13B	Cohen syndrome		splice acceptor variant	c.9406-1G>T		rs386834119	23188044, 16917849, 15154116
10	89720633	C	CT	x	x	PASS	PTEN	Hereditary cancer-predisposing syndrome		intron	c.802-18C>T		rs376702513	25394175, 18951446
12	102796022	A	T	x	x	PASS	IGF1	Insulin-like growth factor I deficiency	begign/likely benign	3' UTR variant	c.*297T>A		rs70961704	
13	20763685	A	AC	x	x	PASS	GJB2	Deafness, autosomal recessive 1	2 alleles one closed to 1%	frameshit	c.35dupG	p.Val13CysfsTer35	rs398123814	9482292, 24503448
21	47545369	A	AC	x		PASS	COL6A2			frameshit	c.1817-10_1817-9insC	p.Asp163ArgfsTer3	rs149954350	
X	66765161	A	T	x	x	PASS	AR	Infertility, male	Not tested.	Missense	c.173A>T	Gln58Leu	rs200185441	12801573, 24737579, 23637914
X	153006092	C	T	x		RF;AC0	ABCD1			stop gained	c.1699C>T	p.Gln567Ter	rs201114595	

Table S3. Biallelic and multi-allelic blacklist-annotated variants in the PID, Neuro, Infection and Africa cohorts

Blacklists						
	Biallelic		**Multiallelic**		**Total**	
	Count	**% of Total**	**Count**	**% of Total**	**Count**	**% of Total**
PID	14,229	8.5	152,915	91.5	167,144	100
Neuro	14,860	66.6	7,454	33.4	22,314	100
Infection	18,717	49.0	19,451	51.0	38,168	100
Africa	48,999	84.2	9,186	15.8	58,185	100

Table S4. Bi-allelic and multi-allelic blacklist-annotated variants by repetitive regions
(STR: short tandem repeats, Alu, GC-rich regions, other repetitive regions**)**

Occurrence of blacklisted variants in complex regions						
	Multi-allelic		**Bi-allelic**		**Total**	
	Count	**% of Total**	**Count**	**% of Total**	**Count**	**% of Total**
In complex regions	118,154	77.3	6,711	47.2	124,865	74.7
Not in complex regions	34,761	22.7	7,518	52.8	42,279	25.3
Breakdown by complex regions						
	Multi-allelic		**Bi-allelic**		**Total**	
	Count	**% of Total**	**Count**	**% of Total**	**Count**	**% of Total**
STR	65,646	55.6	2,457	36.6	68,103	53.5
Alu elements	44,866	38.0	1,713	25.5	46,579	36.7
GC-rich regions	4,314	3.7	1,742	26.0	6,056	6.2
Other repeat regions	3,328	2.8	799	11.9	4,127	3.6

Table S5. Hardy-Weinberg of bi-allelic CDS blacklist-annotated (BL-A) variants in Caucasian individuals

CDS bi-allelic variants in Caucasian Individuals (n = 1150)			
Total	**$<10^{-8}$**	**$>=10^{-8}$**	**% Disequilibrium**
622	74	548	12
CDS bi-allelic variants in disequilibrium by excess genotype			
	excess het	**excess hom alt**	**excess hom WT**
Counts	35	28	11
%	47.3	37.8	14.9
DP	163.0	20.5	15.6

Table S6. Ethnicity distribution of bi-allelic CDS blacklist-annotated (BL-A) variants in Hardy-Weinberg equilibrium

Ethnicity Distribution of CDS bi-allelic variants in HW equilibrium

	Total	**$<10^{-8}$**	**$>10^{-8}$**	**Ethnical Disequilibrium (%)**
Counts	548	200	348	36.5

Causal Ethnicity for Disequilibrium

	Middle Eastern	**African**	**Caucasian**
Counts	20	20	6
%	43.5	43.5	13.0

Table S7: Biallelic blacklist annotated CDS variants in Hardy-Weinberg disequilibrium

Var	Gene	Unique	Exome_gnomAD	Genome_gnomAD	Obs het	Obs hom	Obs wt	HW_Disequilibrium	DP Avg	Figure
4,88536886,CAGTGACAGCAGCAACAGCAGTGACAGCAGCGAT,C	DSPP	unique	PASS	PASS	352	75	150	7.01E-09	50	
6,136599910,T,TGTATCGCTTCTTTCTAGAATGAGATCTTGATCTTGATCA	BCLAF1	unique	PASS	AC0;RF	348	0	797	1.33E-09	210	
6,31324025,G,GT	HLA-B	unique	PASS	PASS	689	43	401	3.1E-32	23	
6,31324603,C,T	HLA-B	unique	PASS	PASS	717	253	173	1.18E-18	61	
6,32489852,A,ACGG	HLA-DRB1	unique	PASS	RF	608	102	357	9.33E-12	49	
6,32551960,T,TCC	HLA-DRB1	multi-01	PASS	PASS	631	113	394	1.03E-09	90	Sup. Figure 11
6,32552056,A,G	HLA-DRB1	multi-01	RF	InbreedingCoeff;RF	720	0	425	2.62E-54	152	Sup. Figure 11
6,32552085,G,GC	HLA-DRB1	multi-01	PASS	InbreedingCoeff	950	47	148	1.35E-114	124	Sup. Figure 11
6,32552093,A,T	HLA-DRB1	multi-01	RF	RF	528	0	610	2.2E-24	109	Sup. Figure 11
6,32552140,T,A	HLA-DRB1	multi-01	PASS	PASS	846	16	253	3.37E-86	64	Sup. Figure 11
6,32552144,A,C	HLA-DRB1	multi-01	PASS	PASS	953	28	119	1.13E-134	58	Sup. Figure 11
6,32557610,T,C	HLA-DRB1	multi-01	.	.	451	0	693	1.01E-16	55	Sup. Figure 11
7,100550245,G,T	MUC3A	unique	InbreedingCoeff	InbreedingCoeff	533	0	192	3.3E-55	547	
7,100551331,G,T	MUC3A	unique	PASS	PASS	850	0	177	2.49E-113	508	
7,142470773,A,G	PRSS3P1	unique	.	.	992	0	153	1.85E-147	213	
7,142231826,T,C	TRBV10-1	unique	PASS	PASS	1046	0	99	4.59E-178	236	
10,94018,T,G	TUBB8	unique	RF;AC0	AC0;InbreedingCoeff;RF	404	0	729	2.81E-13	51	
11,1093430,C,CCACCACGGTGACCCCAACCCCAACACCCACCGGCACACAGACCCCAACAACGACACCCATCAGCACCAA	MUC2	unique	PASS	PASS	740	0	404	8.39E-59	171	
11,1016961,G,T	MUC6	multi-02	RF;AC0	AC0;RF	444	0	700	3.83E-16	306	Sup. Figure 12
11,1016972,G,A	MUC6	multi-02	.	InbreedingCoeff;RF	733	0	411	3.16E-57	280	Sup. Figure 12
11,1017040,G,GA	MUC6	multi-02	RF;InbreedingCoeff	InbreedingCoeff;RF	863	0	281	3.01E-93	237	Sup. Figure 12
11,1017458,A,G	MUC6	multi-02	RF;AC0	InbreedingCoeff;RF	1055	0	89	3.72E-184	231	Sup. Figure 12
11,1017470,G,T	MUC6	multi-02	.	.	908	0	70	1.12E-161	253	Sup. Figure 12
11,1018483,C,G	MUC6	multi-02	InbreedingCoeff	InbreedingCoeff	1015	0	129	3.5E-160	110	Sup. Figure 12
11,48387118,G,A	OR4C5	unique	InbreedingCoeff	InbreedingCoeff	1144	0	0	9.03E-251	125	
11,56143784,C,T	OR8U1	multi-03	InbreedingCoeff	InbreedingCoeff	1102	0	42	8.52E-217	122	Sup. Figure 13
11,56143803,A,G	OR8U1	multi-03	InbreedingCoeff	InbreedingCoeff	1071	0	73	1.1E-194	117	Sup. Figure 13
12,11244067,A,ATT	TAS2R43	multi-04	PASS	AC0;RF	660	203	266	6.36E-09	60	
12,11244070,T,C	TAS2R43	multi-04	PASS	PASS	665	210	251	8.68E-10	60	
15,23685604,TC,T	GOLGA6L2	unique	InbreedingCoeff	InbreedingCoeff	970	1	159	1.23E-140	308	
15,23686113,C,CTGCTCTTACATCTTCTCG	GOLGA6L2	unique	PASS	RF	342	0	785	1.92E-09	401	
15,90294306,C,A	MESP1	unique	PASS	PASS	649	172	270	4.5E-11	23	
19,8999561,G,C	MUC16	unique	RF	InbreedingCoeff;RF	619	0	525	4.27E-36	69	
19,4511350,T,A	PLIN4	unique	.	InbreedingCoeff	713	417	6	1.05E-50	140	
19,50463670,T,G	SIGLEC11	unique	PASS	PASS	404	9	730	3.97E-09	42	

Table S8. Sanger sequencing of 3 variants from blacklist annotated in patient exomes.

Variant					Characterization						Databases			Quality			WES Total			WES Genotype of 10 individuals			Sanger sequence of 10 individuals			Variant Status		
Gene	Chr	Pos	Ref	Alt	BL category	Diseq.	HW Eq. p-value[a]	Repeat region	CCDS	Ethnic Heterogenity	% of cohort with variant	ExAC 0.3.1	GnomAD r2.0.2	Mean DP	Mean MQ	Mean QD	WT[c]	Het	Hom	WT[c]	Het	Hom	WT	Het	Hom	Variant	Call problem	Suspected reason
HRNR	1	152,195,728	AT	A	Multi allelic	nd	nd	No	No	nd	98.3	-	Yes	42.3	60.2	18.3	44	170	2890	0	0	10	nc	nc	nc	nc	Yes	Short stretch of T
TBC1D19	4	26,737,063	C	CT	Multi allelic	nd	nd	No	No	nd	91.8	-	Yes	24.1	60.2	15.1	210	877	2017	0	5	5	nc	nc	nc	nc	Yes	Short stretch of T
FIG4	6	110,053,824	G	GT	Multi allelic	nd	nd	No	No	nd	88.6	-	Yes	28.4	60.0	13.9	349	1231	1524	0	6	4	nc	nc	nc	nc	Yes	Short stretch of T

nc : Not confirmed by Sanger sequencing due to poor quality.

Table S9. Summary of the technology employed for each cohort

Cohort							
	Size	**Kit**	**Sequencer**	**Aligner**	**Reference Genome**	**Caller**	**Annotator**
PID	3,104	Agilent 37, 50, 71 Mb	Hiseq 2000, 2500	bwa(v0.7.12)	hg19	GATK (v3.4-46)	snpEff
Neuro	3,869	Agilent 50 Mb	Hiseq 2000	bwa (v0.7.5)	GRCh37	GATK (v.3.1-1)	snpEff
Africa	400	Nextera Rapid Capture Expanded Exome 61 Mb	Hiseq 2500	bwa (v0.7.7)	GRCh37	GATK (v.3.5)	snpEff
Infection	902	Agilent 50 Mb, Illumina 65Mb	Hiseq 2000, 2500	bwa (v0.7.10)	hg19 decoy	GATK (v3.8)	snpEff

Table S10. Primers for PCR and sanger sequencing

Gene	Forward primer (5’ → 3’)	Reverse primer (5’ → 3’)
FIG4	CTGTCTTGCCCAAAGTCTGC	TTCTCATTCTGCTTTTACCCGC
HRNR	GGCGTGGAGTTCTTACCTTC	CACTCTCTTGCTACATGGCTTG
TBC1D19	CTTTCTGACATTTATGAACAGAG	GTGATTAGAAATAAAGTGGTG