Supplementary Material for

Prioritizing Genes for X-Linked Diseases Using Population Exome Data

Xiaoyan Ge,[1,2] Pui-Yan Kwok,[2,3,4] Joseph T.C. Shieh[1,2]*

[1]Division of Medical Genetics, Department of Pediatrics

[2]Institute for Human Genetics

[3]Department of Dermatology and

[4]Cardiovascular Research Institute, University of California San Francisco, San Francisco, California, 94143, USA

*Correspondence to:
Joseph Shieh, M.D., Ph.D.
Assistant Professor, Division of Medical Genetics, Department of Pediatrics
Institute for Human Genetics, University of California San Francisco
UCSF Benioff Children's Hospital
San Francisco, CA 94143-0793, USA
E-mail: shiehj2@humgen.ucsf.edu

Figure S1: Comparison of the non-synonymous ratio (dN) distribution and the synonymous ratio (dS) distribution.

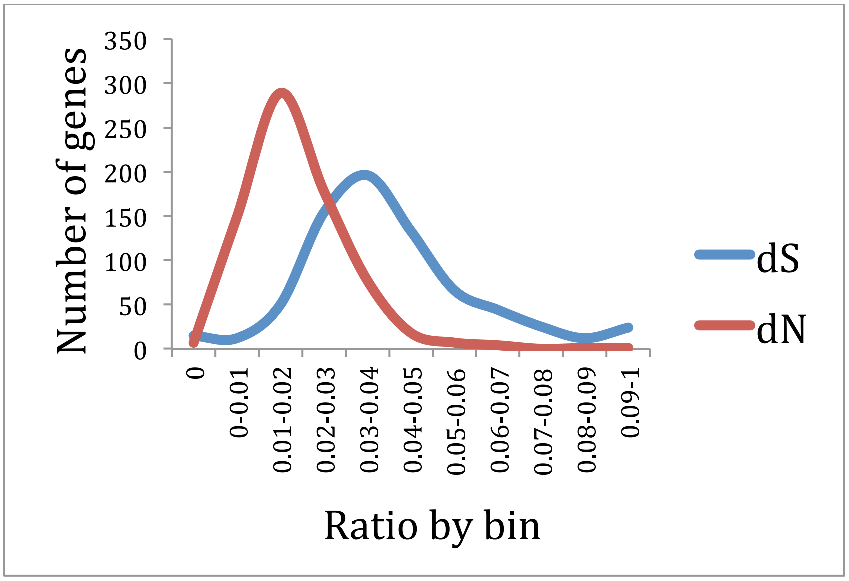

Figure S2: Genes with higher median CADD scores, higher median conservation scores (phyloP and phastCons) and higher network scores have lower dN/dS ratios. (A) dN/dS ratios for genes with high or low CADD scores. p-value=1.135e-15 (two-sample Wilcoxon test). (B) dN/dS ratios for genes with high or low combined network scores. p-value=1.188e-08 (two-sample Wilcoxon test). (C) dN/dS ratios for genes with high or low phyloP scores. p-value=7.041e-3 (two-sample Wilcoxon test). (D) dN/dS ratios for genes with high or low phastCons scores. p-value=7.131e-15 (two-sample Wilcoxon test).

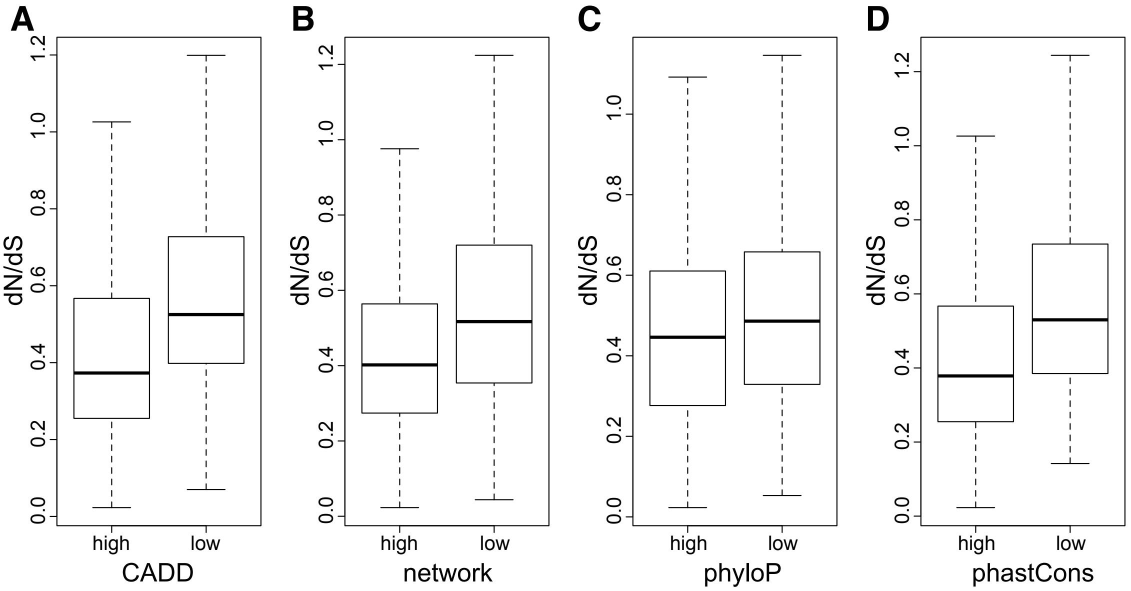

Figure S3: ROC curves of individual parameters used to predict OMIM disease genes. The dN/dS ratio and individual scores from network scoring are shown for X chromosome genes.

The performance of dN/dS ratio is compared to the single scores included in the combined network score. AUC for the corresponding ROC curves are shown. See methods for details.

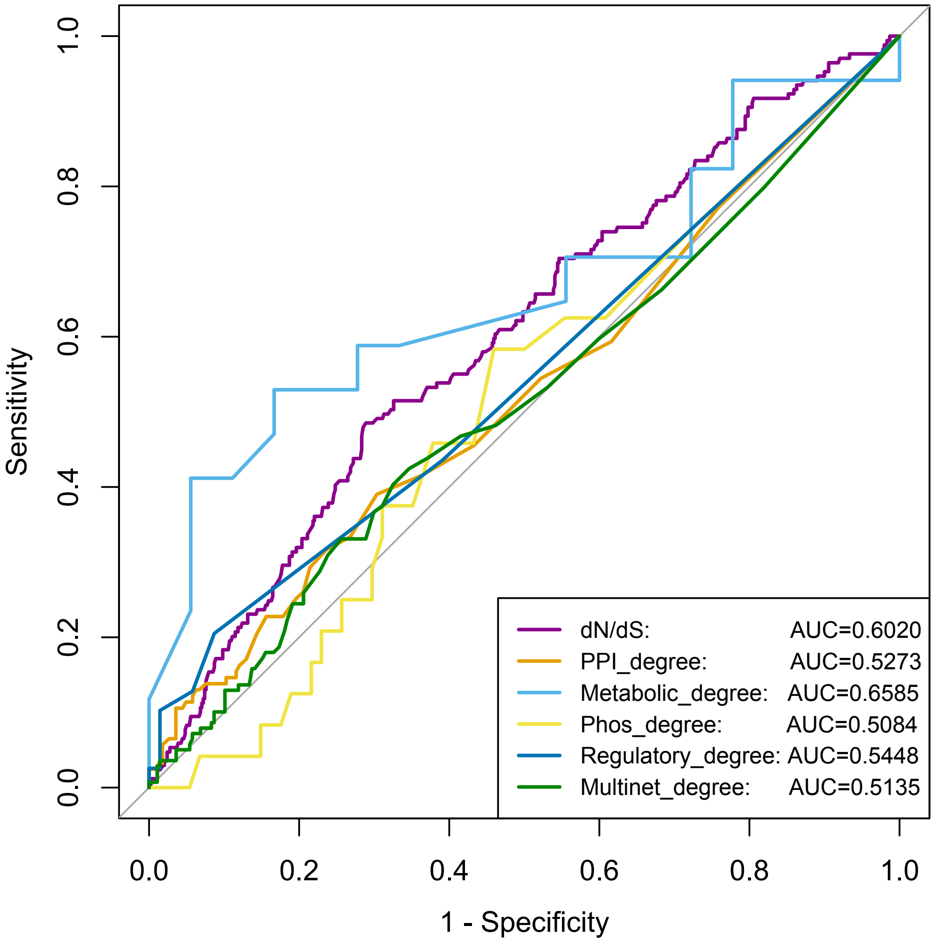

Figure S4: ROC curves of the dN/dS ratio, Median CADD scores, Median conservation scores, Network scores and the All combined scores, reflecting the capacity to predict the OMIM disease genes. Area under the curve (AUC) values are provided for each of the scores.

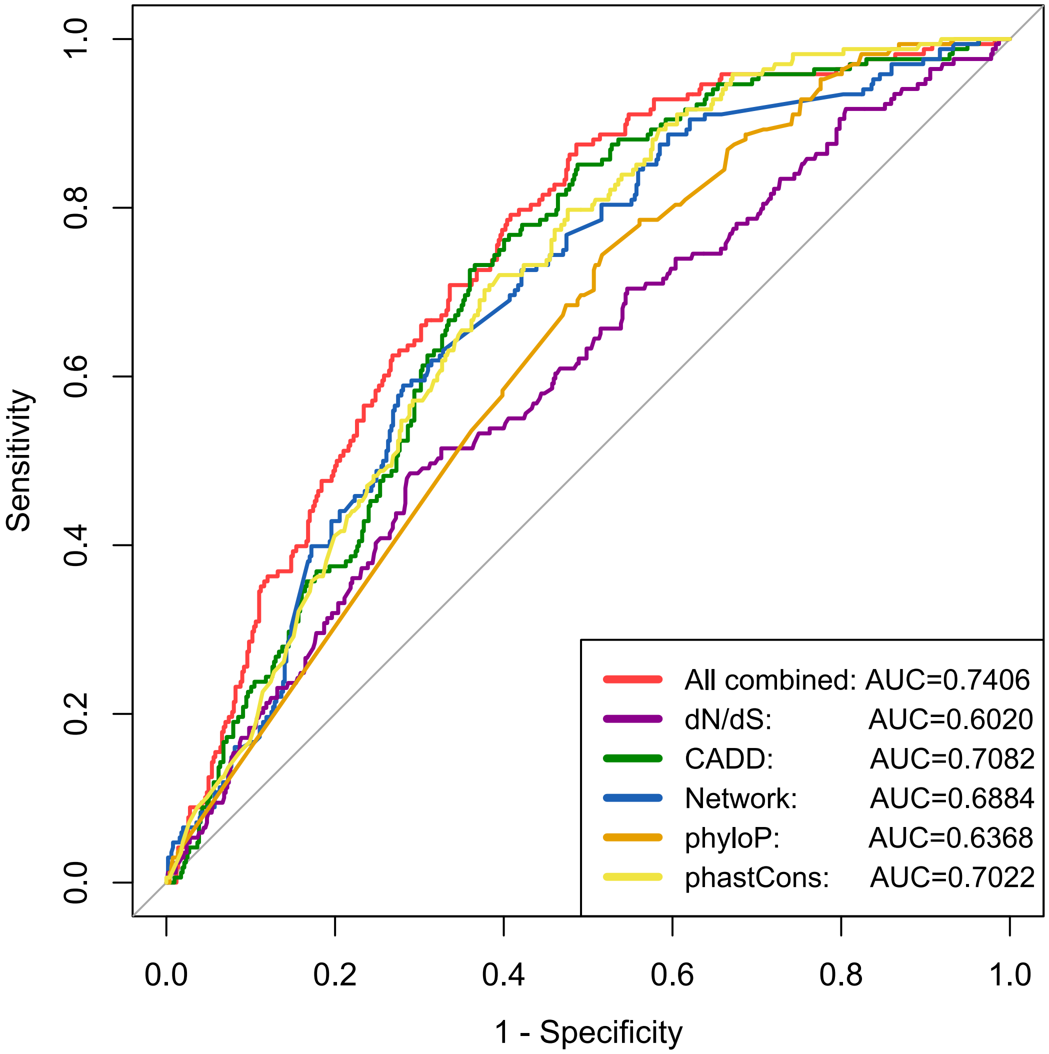

Figure S5: Comparison of All combined scores for OMIM disease genes (OMIM) or genes not yet annotated in OMIM disease database (non-OMIM). OMIM-disease genes have significantly higher All combined scores compared to non-OMIM genes. p-value<2.2e-16 (two-sample Wilcoxon test).

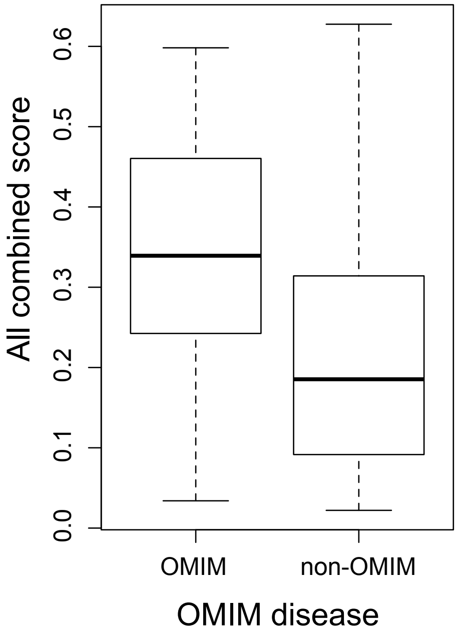

Figure S6: Comparison of dN/dS ratio for African-American and European-American populations in OMIM disease genes (OMIM) or genes not yet annotated in OMIM disease database (non-OMIM). (A) dN/dS ratio for European-American (dN/dS_EA) is lower for OMIM-disease genes compared to non-OMIM genes. p-value=3.12e-02 (two-sample Wilcoxon test). (B) dN/dS ratio for African_American (dN/dS_AA) is lower for OMIM-disease genes compared to non-OMIM genes. p-value=4.65e-05 (two-sample Wilcoxon test).

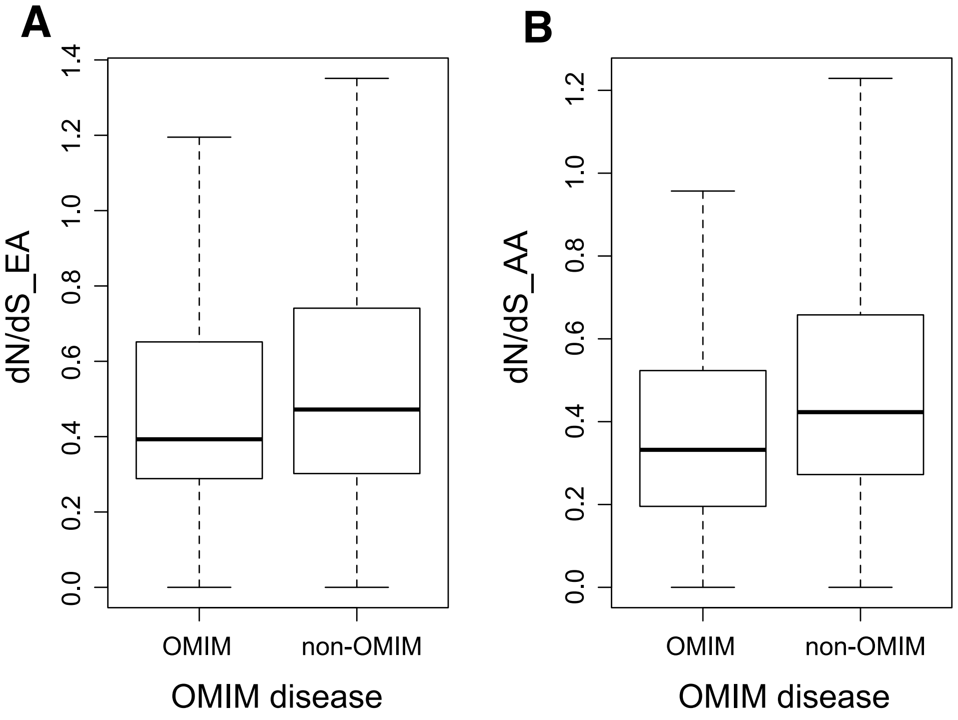

Figure S7: dN/dS ratio for genes with different coding sequence length by scatterplot. Inset: Boxplot of dN/dS ratio for different coding sequence lengths. X-axis is the coding sequence length. Y-axis is the dN/dS ratio. Median dN/dS ratio does not change significantly with coding sequence length, while more variation is observed in smaller genes (e.g. genes with coding sequence length less than 500bp).

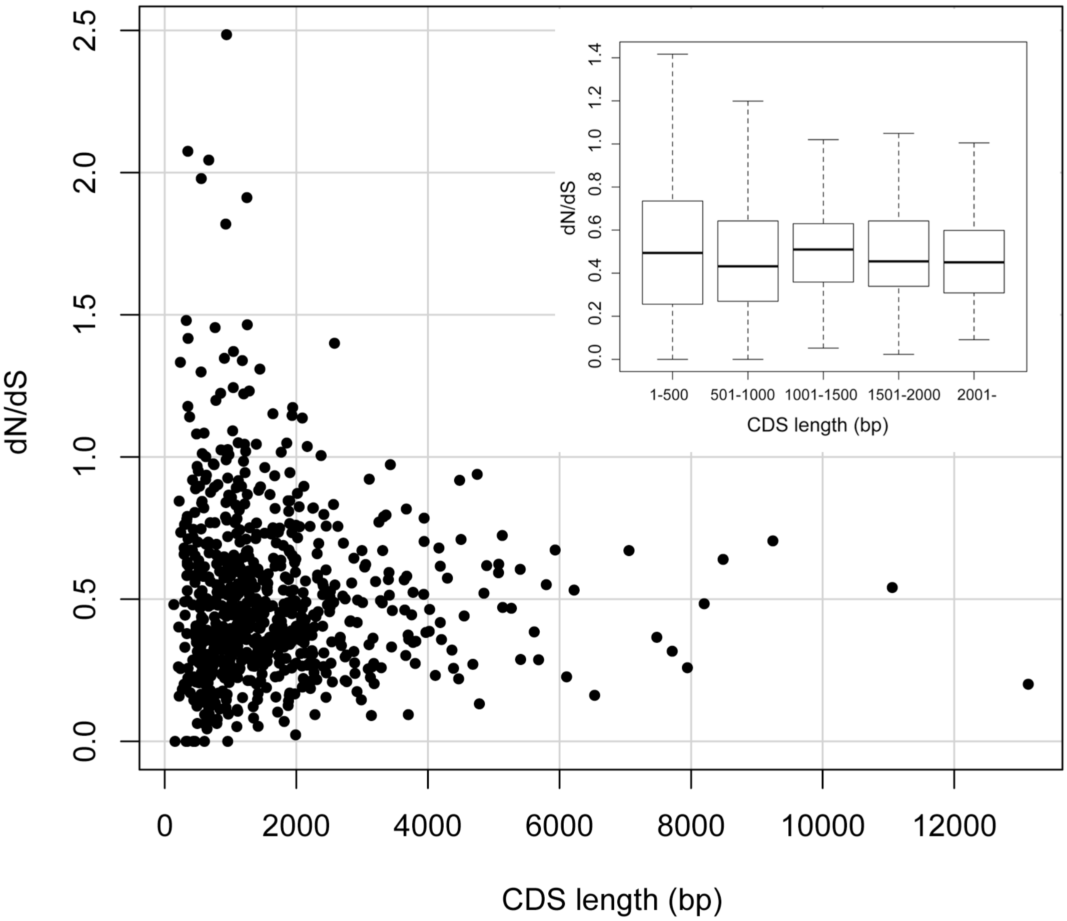

Figure S8: Variant distribution patterns in genes without annotated pathogenic variants. From top to bottom: Top: Variant site density plots for synonymous and nonsynonymous variants along the coding sequence demonstrate six genes with localized missense depletion or occurrence. X-axis shows the relative position of the synonymous and nonsynonymous variants in the coding sequence; the y-axis reflects the number of variants. Green: synonymous. Red: nonsynonymous. Bottom: Domain structures: dark grey rectangles show the protein domains and light grey bars show the whole protein. *HDAC6*: NM_006044; *RPGR*: NM_001034853; *TRPC5*: NM_012471; *UPF3B*: NM_080632; *USP51*: NM_201286; *VAMP7*: NM_001185183;

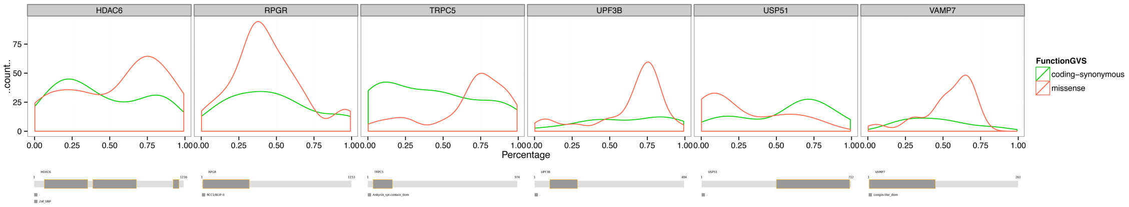

Figure S9: Relative position of stop-gain variants in X-chromosome OMIM gene coding sequence. X axis represents the relative position along coding sequence (e.g. 0.1 is 10% of the coding sequence). Y axis represents the number of genes.

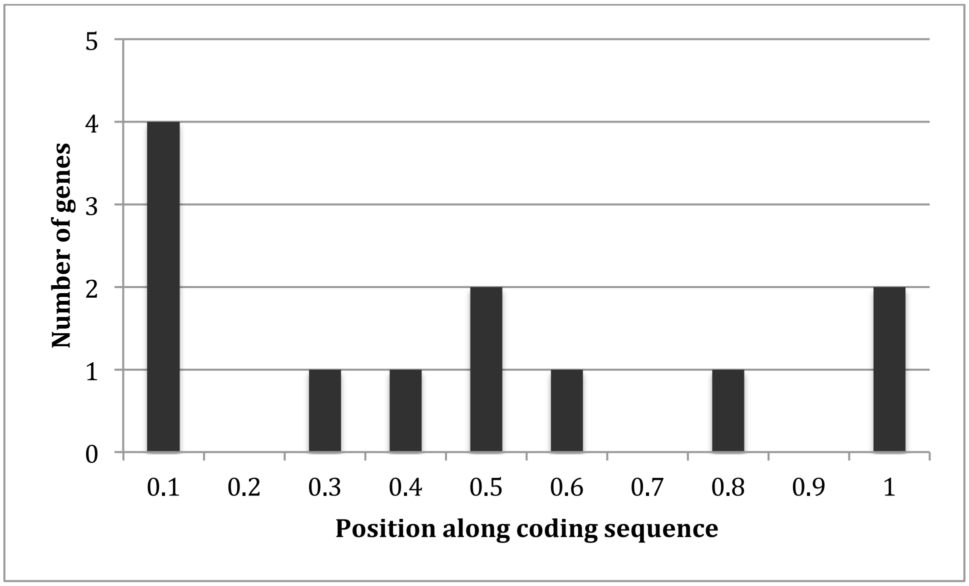

Table S1: Variant diversity for each chromosome.

Chromsome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
stop_gained	1771	1081	970	667	716	823	700	564	605	598	1042	857	246	452	519	683	965	241	1306	358	160	313	233	4
stop_lost	57	27	29	27	24	32	23	23	19	25	41	26	4	10	19	19	28	3	43	15	7	11	10	0
splice_3	355	268	224	158	135	205	154	140	142	144	243	213	72	107	106	169	233	58	217	109	35	59	56	0
splice_5	490	299	283	187	160	233	217	185	161	180	270	285	76	124	147	197	271	54	279	116	40	103	62	2
frameshift	2337	1459	1292	889	961	1113	1074	732	925	777	1424	1143	420	677	708	887	1372	289	1934	500	220	471	380	2
missense	70833	48760	39807	27393	31893	34704	30989	24674	29156	26536	44310	34637	11628	21982	23636	31788	40124	10452	49045	17859	7918	15421	16624	147
coding_synonymous	41568	28736	23432	15555	19056	20173	19258	14938	18363	15935	25834	21228	7161	13150	13951	20722	25600	6528	31951	11690	5190	10427	11186	71
coding_other	1184	820	656	540	533	668	1254	401	471	437	699	660	235	377	395	649	761	170	1020	322	156	290	510	2
utr_3	4717	2653	2302	1532	2269	2218	2117	1370	1792	1521	2430	2356	689	1305	1273	2459	3106	619	3398	1279	608	1325	1113	5
utr_5	3003	1832	1597	1057	1204	1711	1286	1020	1061	1046	1839	1501	538	894	779	1174	1693	355	2164	743	277	630	821	5
intron	71115	54500	42209	26428	64873	33604	34609	25470	31718	29348	39536	38413	12697	21165	24162	33864	44520	10613	46689	19661	9876	17214	17683	139
total_protein_coding	1E+05	81450	66693	45416	53478	57951	53669	41657	49842	44632	73863	59049	19842	36879	39481	55114	69354	17795	85795	30969	13726	27095	29061	228
LOF	5010	3134	2798	1928	1996	2406	2168	1644	1852	1724	3020	2524	818	1370	1499	1955	2869	645	3779	1098	462	957	741	8
bps	3E+06	2E+06	2E+06	1E+06	2E+06	2E+06	2E+06	1E+06	1E+06	1E+06	2E+06	2E+06	6E+05	1E+06	1E+06	1E+06	2E+06	5E+05	2E+06	8E+05	3E+05	7E+05	1E+06	36858
No.genes	1925	1177	1028	707	825	988	848	625	738	698	1220	983	298	569	538	767	1095	256	1321	516	211	415	737	18
mis_by_syn	1.704	1.697	1.699	1.761	1.674	1.72	1.609	1.652	1.588	1.665	1.715	1.632	1.624	1.672	1.694	1.534	1.567	1.601	1.535	1.528	1.526	1.479	1.486	2.07
LOF_by_syn	0.121	0.109	0.119	0.124	0.105	0.119	0.113	0.11	0.101	0.108	0.117	0.119	0.114	0.104	0.107	0.094	0.112	0.099	0.118	0.094	0.089	0.092	0.066	0.113

Table S2: dN/dS ratios for X-chromosome genes, calculated using non-synonymous (N) and synonymous (S) variants in ESP.

See attached file Supplementary Table 2.xlsx

Table S3: High-confidence loss of function variants in males, females and by racial/ethnic group (European American and African American).

	LOF Total	frameshift	splice-3	splice-5	stop-gain	stop-lost
Female specific	36	6	8	9	12	1
Male specific	15	6	3	3	3	0
EA specific	36	13	6	7	10	0
AA specific	18	2	5	5	5	1

Table S4: OMIM disease genes on X chromosome that are deleted in at least one structural variant in healthy control samples.

	Total Gene Number	Number of genes with complete DGV deletion	Percentage
Total OMIM genes	174	71	40.8%
OMIM genes with LoF variants	57	27	47.4%
OMIM genes without LoF variants	117	44	37.6%

Table S5: Enriched pathways for X-chromosome OMIM disease genes compared to all X chromosome genes

KEGG Pathway	Number of Genes	Gene List	Adjusted P-value*
Metabolic pathways	20	*PDHA1, GK, OTC, NSDHL, SAT1, HSD17B10, C1GALT1C1, ALAS2, EBP, IDS, HPRT1, PGK1, ALG13, OCRL, NDUFA1, MAOA, ACSL4, PIGA, PRPS1, SMS*	9.45E-05
Small cell lung cancer	4	*COL4A5, IKBKG, COL4A6, XIAP*	0.0325
Ubiquitin mediated proteolysis	6	*UBA1, HUWE1, MID1, CUL4B, UBE2A, XIAP*	0.0325
Primary immunodeficiency	4	*IKBKG, CD40LG, IL2RG, BTK*	0.0325

*Benjamini-Hochberg