The American Journal of Human Genetics, Volume *89*
Supplemental Data

Epistatic Selection between Coding and Regulatory Variation in Human Evolution and Disease

Tuuli Lappalainen, Stephen B. Montgomery, Alexandra C. Nica, and Emmanouil T. Dermitzakis

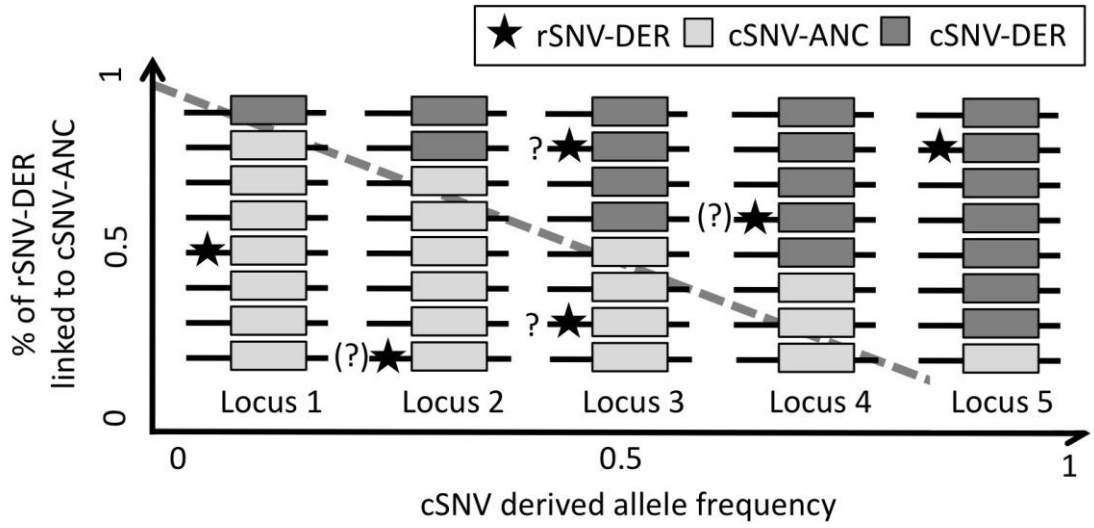

Figure S1. The approach for predicting the direction of expression change from allele-specific expression (ASE) data. In ASE data, only the coding variants (boxes) are observed. A new random regulatory mutation leading to a rare ASE event (stars) will hit the haplotype carrying the ancestral allele of a coding variant with the probability that equals the ancestral allele frequency. Thus, especially for rare ASE putatively caused by recent regulatory mutation, if the cSNVs DAF is low, the new derived rSNV allele is more likely to be linked to the common cSNV ancestral allele. In such a case, whether the cSNV ancestral allele is higher or lower expressed is informative of whether the new rSNV increases or decreases gene expression. Similarly, for cSNVs with high DAF, the common cSNV derived allele is likely to be linked to the new rSNV derived allele, and be informative of the direction of effect. This approach is expected to be informative of the direction of gene expression change induced by regulatory mutations especially when analyzing rare ASE events of neutral cSNVs – for common regulatory variants the linkage has sometimes been broken by recombination, and functional cSNVs may be affected by selection.

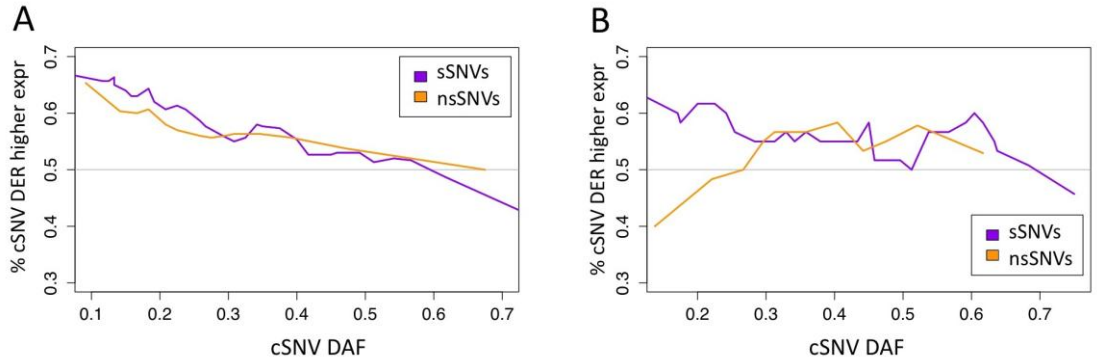

Figure S2. Expression of coding variants based on allele-specific expression data. The proportion of cSNVs with ASE where the derived allele is higher expressed for rare ASE (a; ≤ 2 heterozygotes with ASE of ≥ 6 heterozygote individuals), and for common ASE (b; ≥ 5 heterozygotes with ASE of ≥ 6 heterozygote individuals) – see Figure 3c for all cSNVs with ASE. The decreasing trend especially in (a) but also in the other cases is likely to result from the phenomenon illustrated in Supplementary Figure 3, indicating that the cSNV allele that is more likely to be linked to a rare, recent regulatory variant is usually lower expressed (in 78 ± 12% of the cases at cSNV DAF=0 according to a linear regression model of sSNVs in (a), $p = 2.1x10^{-10}$). This may be due to new regulatory mutation being more likely to lead to loss of gene expression, and/or due to a selective process leading to accumulation of rare loss-of-expression rSNVs. The overall difference between sSNVs and nsSNVs is not significant in (a) and (b), based on linear regression. However, since the trend in (b) does not appear linear for nsSNVs, we also tested sSNV-nsSNV difference with a Fisher test for cSNVs with DAF<0.15, yielding p-values of 0.92 for (a) and 0.015 for (b).

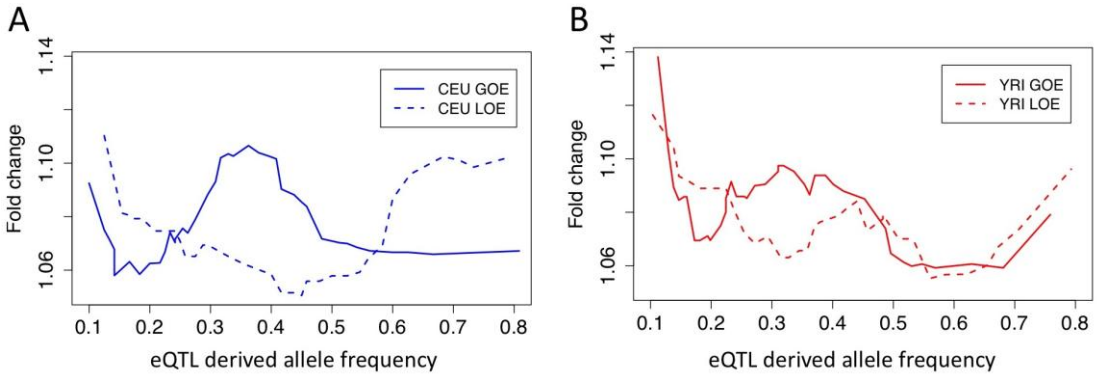

Figure S3. eQTL impact size. Fold change of gene expression for gain-of-expression (GOE) and loss-of-expression (LOE) eQTLs in CEU (a) and YRI (b) as a median in sliding windows of 50 SNPs with an overlap of 5 SNPs. The figure shows how eQTLs especially in CEU have very high fold changes in the parts of the frequency spectrum where they putatively lead to protective, beneficial epistatic effects. The enrichment of high fold changes in very low derived allele frequencies is probably due to cases where gene expression level change itself is deleterious, with selection preventing eQTLs causing major changes from reaching high frequencies. Of note, especially in CEU this signal of selective constraint on expression levels is not any stronger than the patterns putatively caused by epistatic selection, suggesting that epistasis has a profound effect on regulatory variation.

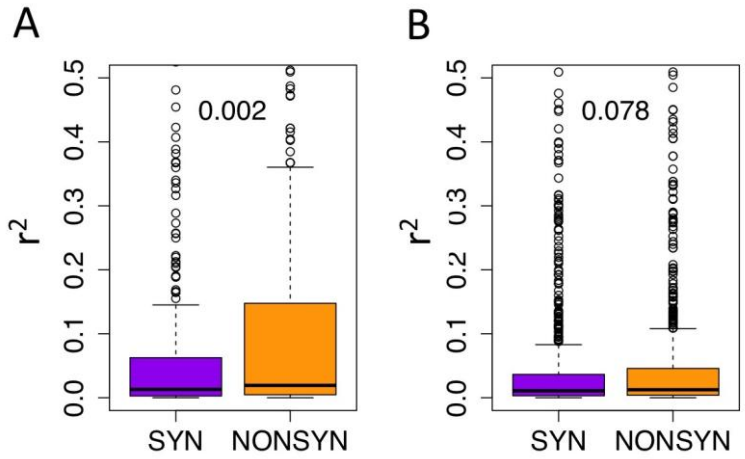

Figure S4. Linkage disequilibrium (r^2) between eQTLs and coding variants. LD between eQTLs and synonymous or nonsynonymous coding SNVs r^2 in CEU (a) and YRI (b). The synonymous SNVs were sampled to the derived allele frequency distribution of nonsynonymous SNVs, and the numbers denote the Mann-Whitney p-values for sSNV–nsSNV comparisons.

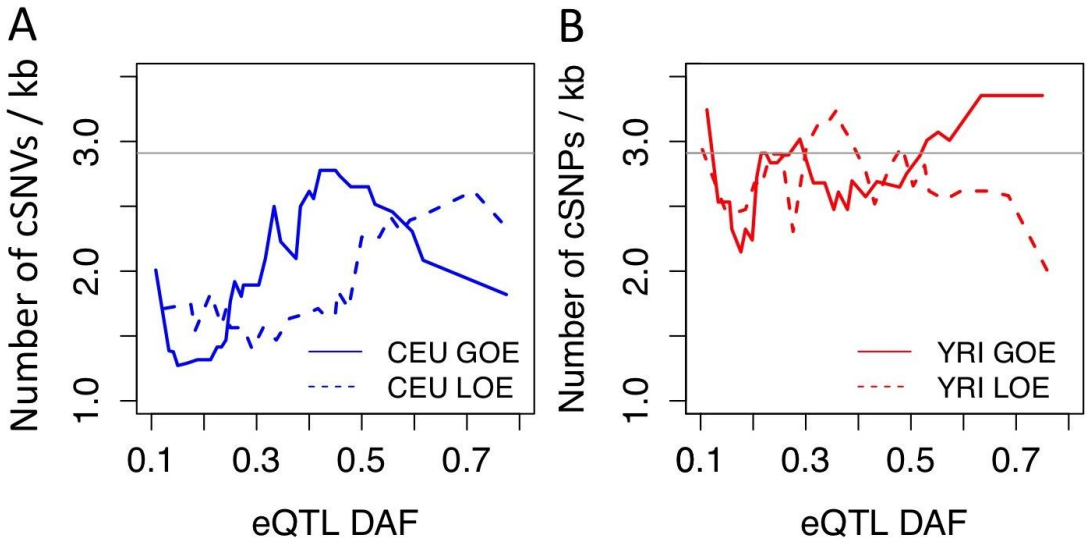

Figure S5. Deficiency of coding variation in eQTL genes. The number of segregating coding sites in CEU (a) and YRI (b) genes with gain-of-expression (GOE) and loss-of-expression (LOE) eQTLs, as a mean in sliding windows on 50 SNPs with an overlap of 5 SNPs. The vertical line shows the median cSNV density of genes without eQTLs in this dataset. The lower number of cSNVs in eQTL genes ($p < 2.2 \times 10^{-16}$ in CEU and $p = 6.4 \times 10^{-3}$ in YRI) suggests increased purifying selection against cSNVs in the presence of regulatory variation, a phenomenon that appears even more pronounced in genes with high frequency of the higher expressed eQTL haplotype, when many coding mutations hit the higher expressed haplotype. The pattern is similar for both nonsynonymous and synonymous variants (data not shown); even though epistatic selection is expected to act mainly on the former, it may increase background selection that affects neutral synonymous variants as well.

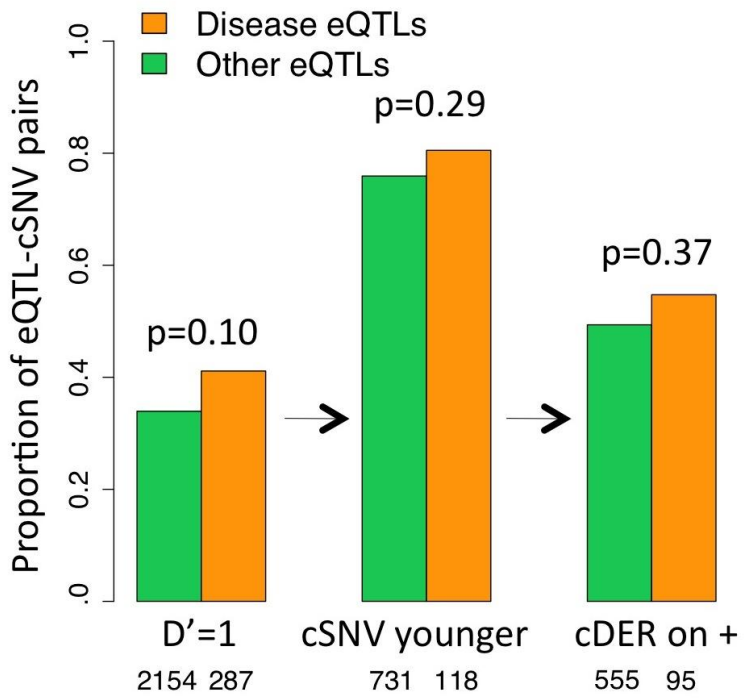

Figure S6. Coding variation in disease-associated eQTLs. Disease-associated eQTLs show a slight, nonsignificant trend of having more coding SNVs in full LD with the eQTL, having a bigger proportion of these cSNVs younger than the eQTL variant, and of these having the cSNV mutation more often occurring on the higher expressed haplotype. The numbers below the bars denote the total number of coding variants tested in each category (denominators of the proportions); the p-values are from Fisher tests.

Table S1. The epistasis model. The two-locus haplotypes have regulatory variation with gain-of-expression (G) and loss-of-expression (L) alleles with frequencies p_G and $p_L = 1-p_G$, and coding variation with wild type (w) and deleterious mutant (m) alleles with frequencies p_w and $p_m = 1-p_w$. Selection coefficient of the m allele is denoted by s, dominance by h, mutation rate w $\rightarrow$ m by μ, and the penetrance effect of the regulatory variant by i.

Genotype	**Fitness**	**Frequency**
$\mathbf{G_w / G_w}$	1	$[p_G(p_w - \mu p_G)]^2$
$\mathbf{L_w / L_w}$	1	$[p_L(p_w - \mu p_L)]^2$
$\mathbf{G_w / L_w}$	1	$2p_G(p_w - \mu p_G)\, p_L(p_w - \mu p_L)$
$\mathbf{G_w / G_m}$	$1 - hs$	$2p_G(p_w - \mu p_G)\, p_G(p_m + \mu p_G)$
$\mathbf{L_w / L_m}$	$1 - hs$	$2p_L(p_w - \mu p_L)\, p_L(p_m + \mu p_L)$
$\mathbf{G_m / L_w}$	$1 - [i + (1 - i)h]s$	$2p_G(p_m + \mu p_G)\, p_L(p_w - \mu p_L)$
$\mathbf{G_w / L_m}$	$1 - (1 - i)hs$	$2p_G(p_w - \mu p_G)\, p_L(p_m + \mu p_L)$
$\mathbf{G_m / L_m}$	$1 - s$	$2p_G(p_m + \mu p_G)\, p_L(p_m + \mu p_L)$
$\mathbf{L_m / L_m}$	$1 - s$	$[p_L(p_m + \mu p_L)]^2$
$\mathbf{G_m / G_m}$	$1 - s$	$[p_G(p_m + \mu p_G)]^2$

Table S2. Comparison of coding variation on the eQTL haplotypes. The table shows the distribution of coding variation for cSNVs that were in full LD (D' = 1) with the eQTL variant and likely more recent than that, assessed from the haplotype counts. The data suggests a stronger deficiency of derived alleles on the higher expressed (+) haplotype for nonsynonymous variants in CEU, compared to synonymous variants. However, also synonymous variants show underrepresentation of derived alleles on the higher expressed haplotype. This is likely due to both new nonsynonymous and synonymous mutations hitting more often the lower expressed eQTL haplotype due to its higher frequency (Figure 2), as well as increased background selection on the higher expressed haplotype. The p-values are from chi-squared tests for whether the observations are significantly different from even proportions of derived alleles on higher and lower expressed haplotypes, and the p-values for comparison between sSNVs and nsSNVs are based on Fisher's exact test.

		CEU				YRI			
		% on + haplo	Total count	p-value	p-value s/nsSNV	% on + haplo	Total count	p-value	p-value s/nsSNV
allele counts	All cSNP	45.0	38223	$1.3x10^{-85}$	NA	48.3	43665	$1.3x10^{-12}$	NA
	sSNP	45.9	20535	$2.0x10^{-32}$	$2.1x10^{-4}$	48.0	23706	$1.4x10^{-9}$	0.222
	nsSNP	44.0	17688	$6.1x10^{-58}$		48.6	19959	$1.0x10^{-4}$	
	sSNP DAF < 0.05	50.4	446	0.850	0.1277	46.1	848	0.023	0.423
	nsSNP DAF < 0.05	45.4	465	0.046		48.0	964	0.221	
SNP counts	All cSNP	48.6	327	0.619	NA	47.3	656	0.160	NA
	sSNP	51.9	158	0.633	0.2695	45.0	300	0.083	0.308
	nsSNP	45.6	169	0.249		49.2	356	0.751	
	sSNP DAF < 0.05	55.4	92	0.297	0.0590	43.5	177	0.084	0.309
	nsSNP DAF < 0.05	41.2	97	0.084		49.1	214	0.785	

Table S3. Candidate loci for epistatic disease associations. Disease-associated eQTLs with a high frequency of the higher expressed haplotype that may be good candidate loci for disease association arising from increased penetrance of deleterious coding variants. eQTL effects GOE and LOE denote gain and loss-of-expression, and the tissue abbreviations stand for fibroblasts, T-cells and LCLs.

eQTL SNP	eQTL gene	eQTL effect	+ allele freq	GWAS SNP	Phenotype	Tissue	RTC score
rs4759113	ENSG00000094914_AAAS	GOE	0.5282	rs10876432	Bone mineral density (spine)	FIBR	0.88
rs10774967	ENSG00000111445_RFC5	GOE	0.6294	rs10444502	Biochemical measures	FIBR	0.92
rs2073643	ENSG00000197375_SLC22A5	GOE	0.5647	rs2188962	Crohns disease	FIBR	0.84
rs12477063	ENSG00000138376_BARD1	GOE	0.5627	rs6435862	Neuroblastoma (high-risk)	TCELL	0.97
rs1321311	ENSG00000146192_FGD2	GOE	0.7606	rs1321311	Electrocardiographic traits	TCELL	1.00
rs7117022	ENSG00000185507_IRF7	GOE	0.7817	rs4963128	Systemic lupus erythematosus	TCELL	0.86
rs2524094	ENSG00000204371_EHMT2	GOE	0.6161	rs9264942	HIV-1 control	TCELL	1.00
rs12603332	ENSG00000073605_GSDML	GOE	0.548	rs2290400	Type 1 diabetes	BCELL	0.97
rs6060355	ENSG00000101019_UQCC	GOE	0.5857	rs6060369	Height	BCELL	1.00
rs3843935	ENSG00000164978_NUDT2	GOE	0.6743	rs216345	Bipolar disorder	BCELL	1.00
rs6918423	ENSG00000181315_ZNF322A	GOE	0.7163	rs13194053	Schizophrenia	BCELL	0.81
rs3812558	ENSG00000187796_CARD9	GOE	0.6426	rs10781500	Ulcerative colitis	BCELL	1.00
rs3130474	ENSG00000204498_NFKBIL1	GOE	0.6089	rs9264942	HIV-1 control	BCELL	0.95
rs270607	ENSG00000053108_FSTL4	LOE	0.7145	rs1016988	Fibrinogen	FIBR	0.80
rs744166	ENSG00000126561_STAT5A	LOE	0.5446	rs744166	Multiple sclerosis	FIBR	1.00
rs12150376	ENSG00000141200_KIF2B	LOE	0.787	rs8073783	Conduct disorder (interaction)	FIBR	0.98
rs1567438	ENSG00000146457_WTAP	LOE	0.6544	rs6919346	Plasma Lp (a) levels	FIBR	0.81
rs7644602	ENSG00000153551_CMTM7	LOE	0.6303	rs4380451	Bipolar disorder	FIBR	0.91
rs758642	ENSG00000172146_OR1A1	LOE	0.6813	rs758642	Smoking behavior	FIBR	1.00
rs214053	ENSG00000197061_HIST1H4C	LOE	0.5246	rs742132	Serum uric acid	FIBR	0.95
rs2270196	ENSG00000204920_ZNF155	LOE	0.6373	rs2191566	Acute lymphoblastic leukemia (childhood)	FIBR	0.91
rs10946208	ENSG00000071242_RPS6KA2	LOE	0.6962	rs2301436	Crohns disease	TCELL	0.80
rs2732588	ENSG00000073969_NSF	LOE	0.625	rs199533	Parkinsons disease	TCELL	0.91
rs11577406	ENSG00000133059_RIPK5	LOE	0.7668	rs1668873, rs1668873	Mean platelet volume	FIBR, TCELL	0.83, 0.85
rs11953346	ENSG00000133302_BRCTD1	LOE	0.7993	rs17418283	Bipolar disorder	TCELL	0.90
rs4310388	ENSG00000142606_MMEL1	LOE	0.5739	rs3890745	Rheumatoid arthritis	TCELL	0.98
rs3812584	ENSG00000148384_INPP5E	LOE	0.5972	rs10781500	Ulcerative colitis	TCELL	0.90
rs10035791	ENSG00000158985_CDC42SE2	LOE	0.8445	rs7731657	Fasting plasma glucose	TCELL	1.00
rs10851872	ENSG00000167173_C15orf39	LOE	0.8785	rs1378942	Diastolic blood pressure	TCELL	0.91
rs12534124	ENSG00000172209_GPR22	LOE	0.8612	rs3815148	Osteoarthritis	TCELL	0.98
rs10962263	ENSG00000173068_BNC2	LOE	0.8564	rs1927702	Body mass index	TCELL	0.93
rs803686	ENSG00000184779_RPS17	LOE	0.7465	rs783540	Chronic lymphocytic leukemia	TCELL	0.83
rs13219354	ENSG00000186470_BTN3A2	LOE	0.9208	rs13194053	Schizophrenia	TCELL	0.96

rs363564	ENSG00000186924_KRTAP22-1	LOE	0.8246	rs363512	Hyperactive-impulsive symptoms	TCELL	0.80
rs3778709	ENSG00000187037_GPR141	LOE	0.6441	rs741301	Diabetic nephropathy	TCELL	0.98
rs2071286	ENSG00000204305_AGER	LOE	0.8163	rs2269426	Plasma eosinophil count	TCELL	0.84
rs4151672	ENSG00000204348_DOM3Z	LOE	0.9571	rs494620	Menopause (age at onset)	TCELL	0.82
rs9533101	ENSG00000023516_AKAP11	LOE	0.6316	rs9594759	Bone mineral density (spine)	BCELL	0.84
rs1946294	ENSG00000089225_TBX5	LOE	0.9435	rs3825214	Electrocardiographic traits	BCELL	0.81
rs5751901	ENSG00000100031_GGT4P	LOE	0.5985	rs5751901	Protein quantitative trait loci	BCELL	1.00
rs757110	ENSG00000110696_C11orf58	LOE	0.7165	rs5215	Type 2 diabetes	BCELL	0.96
rs10920046	ENSG00000118194_TNNT2	LOE	0.7376	rs7512898	Electrocardiographic conduction measures	BCELL	0.91
rs3789311	ENSG00000119396_RAB14	LOE	0.8292	rs3761847, rs3761847, rs3761847	Rheumatoid arthritis	FIBR, BCELL, TCELL	0.88, 0.89, 0.81
rs8018822	ENSG00000125375_ATP5S	LOE	0.7553	rs8020441	Cognitive performance	BCELL	1.00
rs535716	ENSG00000142102_ATHL1	LOE	0.7836	rs11602954	Mean platelet volume	BCELL	0.95
rs7127239	ENSG00000148943_LIN7C	LOE	0.7165	rs7481311	Weight	BCELL	1.00
rs770213	ENSG00000155158_C9orf52	LOE	0.6197	rs1927702	Body mass index	BCELL	0.92
rs9303277	ENSG00000172057_ORMDL3	LOE	0.5229	rs907092, rs7216389, rs2872507	Primary biliary cirrhosis,Asthma,Crohns disease	BCELL	0.89
rs9272346	ENSG00000179344_HLA-DQB1	LOE	0.5601	rs9272346, rs9268480, rs660895	Type 1 diabetes,Ulcerative colitis,Rheumatoid arthritis	BCELL	1.00