

## **Supplementary Material**

Cui, P., *et al.* 2012. The transcript-centric mutations in human genomes. *Genomics Proteomics Bioinformatics* 10(1): 11-22.

DOI: 10.1016/S1672-0229(11)60029-6.

Tables S1-S5; Figures S1-S7



**Table S2 List of 9,732 ubiquitously-expressed genes and expression levels in 10 human tissues**

**Table S3 List of 646 EIGs**

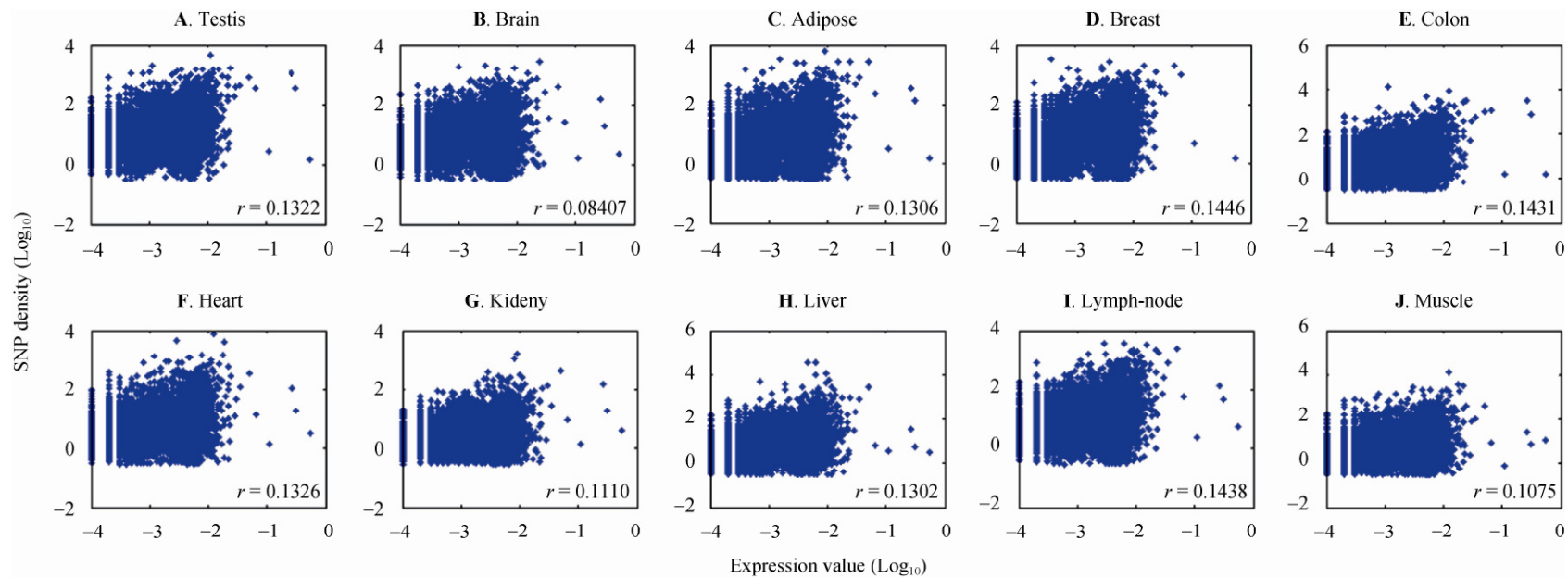
**Table S4 Gene ontology of 646 EIGs**

Category	GO	Term	<i>P</i>	Count	%
Molecular Function	GO:0003735	structural constituent of ribosome	2.93E-33	52	8.062016
	GO:0005515	protein binding	2.78E-18	393	60.93023
	GO:0008135	translation factor activity, nucleic acid binding	2.22E-04	13	2.015504
	GO:0016491	oxidoreductase activity	3.92E-04	44	6.821705
	GO:0004601	peroxidase activity	9.09E-04	7	1.085271
	GO:0003676	nucleic acid binding	0.002698	147	22.7907
	GO:0016787	hydrolase activity	0.002756	108	16.74419
	GO:0031202	RNA splicing factor activity, transesterification mechanism	0.010294	5	0.775194
Biology Process	GO:0044237	cellular metabolic process	2.68E-13	331	51.31783
	GO:0045184	establishment of protein localization	9.74E-12	70	10.85271
	GO:0033036	macromolecule localization	6.41E-11	85	13.17829
	GO:0022613	ribonucleoprotein complex biogenesis	1.52E-08	26	4.031008
	GO:0051641	cellular localization	1.50E-07	68	10.54264
	GO:0051649	establishment of localization in cell	1.56E-07	64	9.922481
	GO:0044238	primary metabolic process	2.95E-07	317	49.14729
	GO:0043170	macromolecule metabolic process	6.85E-06	264	40.93023
	GO:0051234	establishment of localization	1.71E-05	139	21.55039
	GO:0006810	transport	2.57E-05	137	21.24031
	GO:0016192	vesicle-mediated transport	2.92E-05	43	6.666667
	GO:0009056	catabolic process	5.32E-05	75	11.62791
	GO:0009058	biosynthetic process	1.23E-04	171	26.51163
	GO:0065009	regulation of molecular function	0.001331	56	8.682171
	GO:0034621	cellular macromolecular complex subunit organization	0.002077	26	4.031008
	GO:0008219	cell death	0.002915	43	6.666667
	GO:0055114	oxidation reduction	0.003355	39	6.046512
	GO:0051128	regulation of cellular component organization	0.004188	30	4.651163
	GO:0006413	translational initiation	0.006358	7	1.085271
	GO:0042221	response to chemical stimulus	0.007318	66	10.23256
	GO:0016044	membrane organization	0.009302	25	3.875969
	GO:0030029	actin filament-based process	0.009466	18	2.790698
	GO:0048523	negative regulation of cellular process	0.010434	81	12.55814
	GO:0043933	macromolecular complex subunit organization	0.011104	40	6.20155
	GO:0006996	organelle organization	0.011243	67	10.3876
	GO:0022607	cellular component assembly	0.015403	47	7.286822
	GO:0044087	regulation of cellular component biogenesis	0.017935	12	1.860465
	GO:0048522	positive regulation of cellular process	0.018202	87	13.48837
	GO:0048519	negative regulation of biological process	0.028858	84	13.02326
	GO:0051716	cellular response to stimulus	0.036536	42	6.511628
	GO:0044419	interspecies interaction between organisms	0.038341	18	2.790698
	GO:0009892	negative regulation of metabolic process	0.040596	40	6.20155
	GO:0070271	protein complex biogenesis	0.040773	28	4.341085
GO:0051656	establishment of organelle localization	0.044385	7	1.085271	
GO:0048518	positive regulation of biological process	0.049274	91	14.10853	
Cell Component	GO:0044424	intracellular part	1.27E-35	519	80.46512
	GO:0030529	ribonucleoprotein complex	1.08E-33	88	13.64341
	GO:0005622	intracellular	2.32E-33	525	81.39535
	GO:0044446	intracellular organelle part	9.42E-29	281	43.56589
	GO:0044422	organelle part	2.78E-28	281	43.56589
	GO:0043229	intracellular organelle	3.50E-27	455	70.54264
	GO:0043227	membrane-bounded organelle	1.45E-21	408	63.25581
	GO:0031967	organelle envelope	3.45E-14	66	10.23256
	GO:0031090	organelle membrane	6.78E-13	91	14.10853
	GO:0043233	organelle lumen	2.07E-11	124	19.22481
	GO:0043228	non-membrane-bounded organelle	1.43E-08	150	23.25581
	GO:0043234	protein complex	1.16E-07	146	22.63566
	GO:0031982	vesicle	1.01E-04	46	7.131783
	GO:0044464	cell part	0.001136	566	87.75194
	GO:0031252	cell leading edge	0.034063	11	1.705426
	GO:0012505	endomembrane system	0.040236	40	6.20155

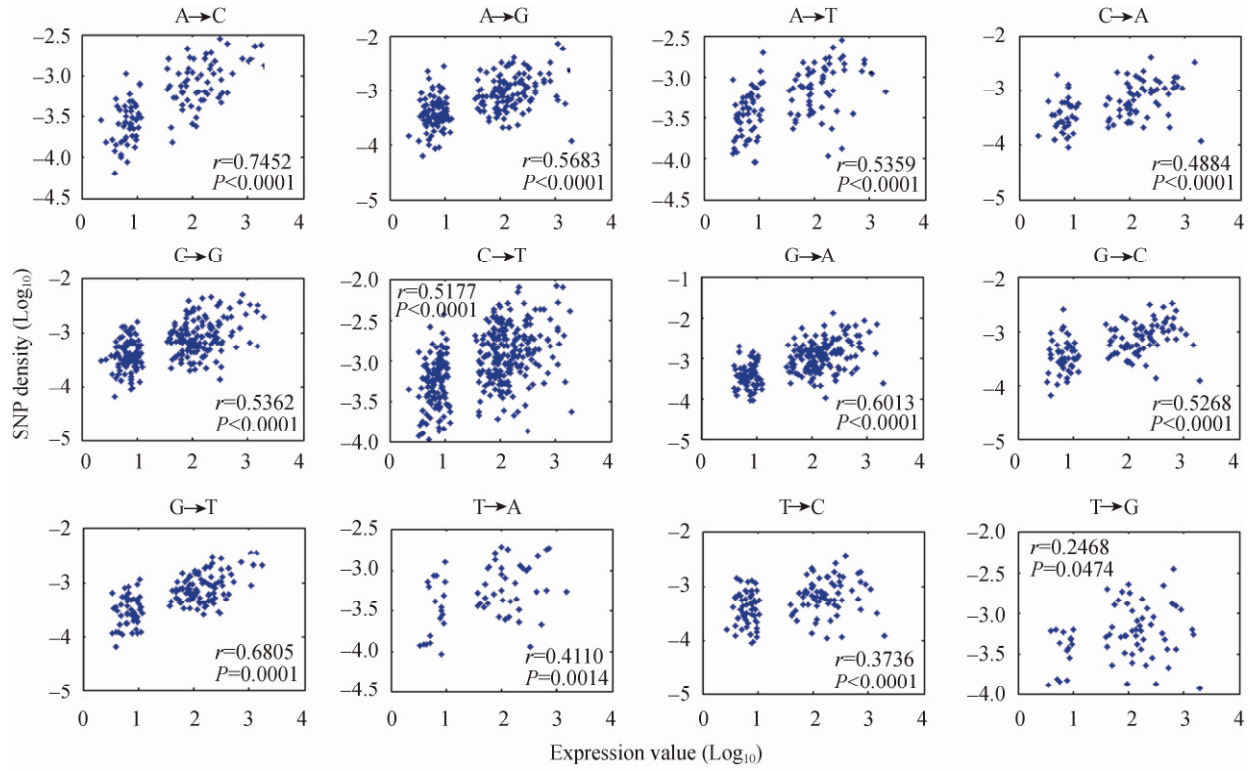
Note: list was sorted by *P*-value. Only the second classification out of the five GO classes was chosen for illustration.

**Table S5 Parameters for nonlinear regression analysis**

mutation type	Parameters in Figure 5A				Parameters in Figure 5B (High exp.)				Parameters in Figure 5B (Low exp.)			
	C→T	A→G	C→G	G→T	C→T	A→G	C→G	G→T	C→T	A→G	C→G	G→T
e	10	5.5	6	5	1.2	1	0.8	0.8	1.2	0.5	0.4	0.4
h	200	175	20	170	0	160	230	10	30	10	250	160
f	176	176	233	176	176	176	233	176	176	176	233	176
B0	30.51	5.03	13.12	11.73	5.07	1.47	2.16	1.33	4.05	0.31	1.04	0.70
B1	-0.1	7.5E-4	-2.7E-3	-0.05042	-0.01135	-0.002601	-0.005453	-0.004128	-0.01564	0.0003108	-0.004825	-0.003314
B2	0.0001	-1.1E-05	1.2E-05	6.6E-05	9.4E-6	2.1E-06	5.2E-06	4.9E-06	1.9E-05	-1.3E-06	7.5E-06	4.2E-06
B3	-4E-08	6.7E-09	3.7E-09	-2.5E-08	-1.8E-09	-6.7E-10	-1.3E-09	-1.6E-09	-7.3E-09	7.9E-10	-3.3E-09	-1.5E-09
R <sup>2</sup>	0.22	-1.112	0.06983	0.2411	0.4329	-0.5063	0.2161	-0.06369	0.731	-1.038	0.2909	-0.1742

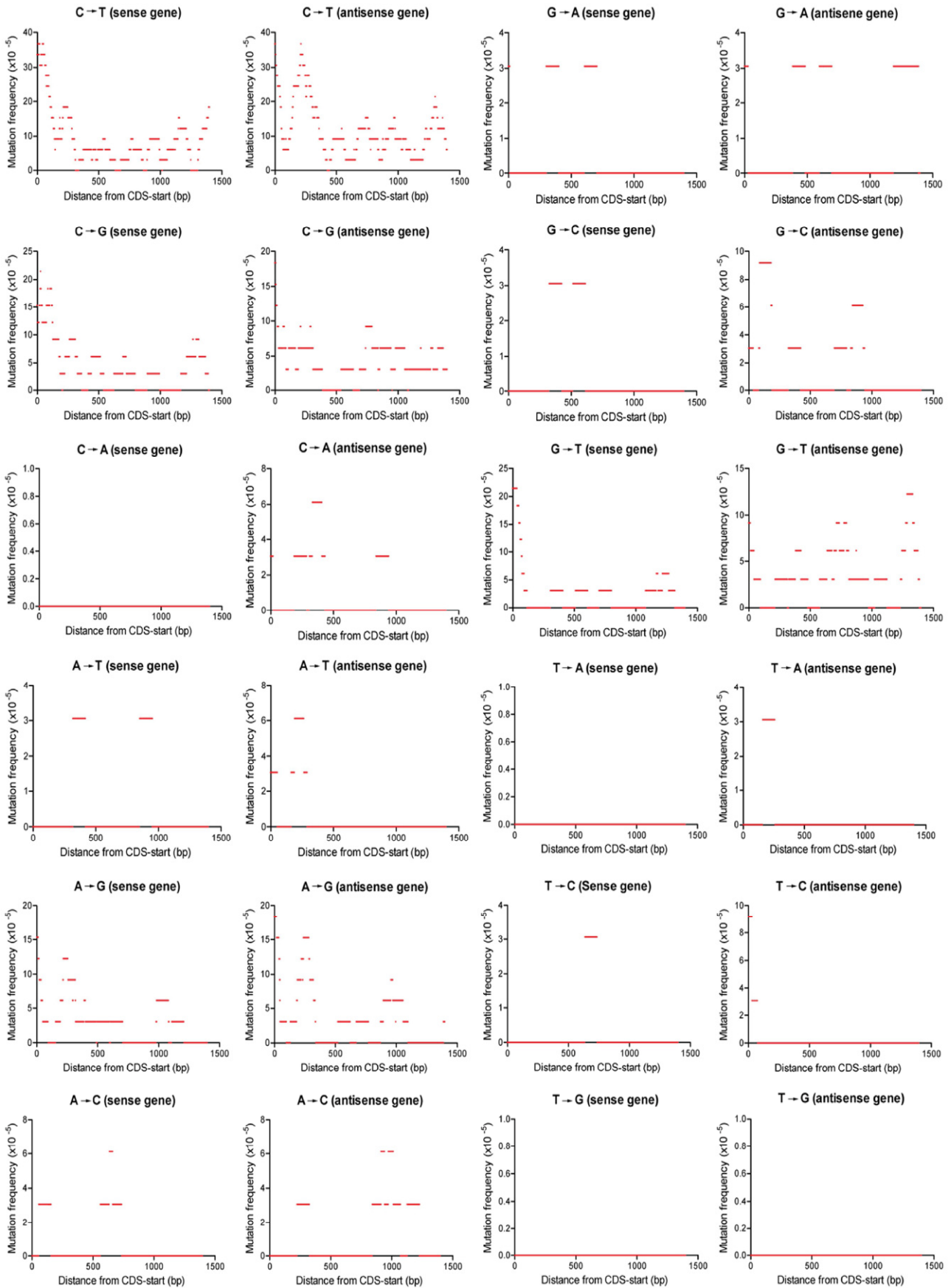


**Figure S1** Correlations between expression levels (Log<sub>10</sub>) and SNP densities of 9,732 ubiquitously-expressed genes in 10 human tissues examined. Panels A to J show data for different tissues including testis, brain, adipose, breast, colon, heart, kidney, liver, lymph-node and muscle. All  $P < 0.0001$ .

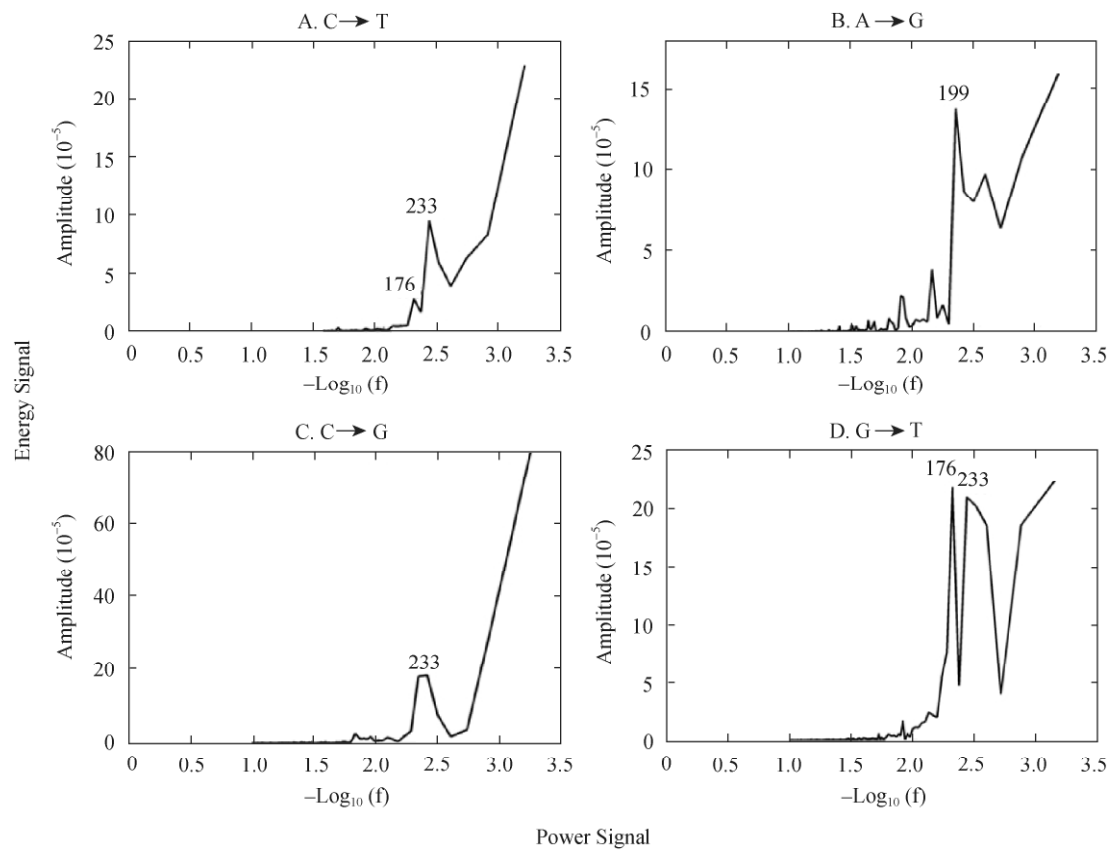


**Figure S2** Correlations between the expression intensity of EIGs and the density of each substitution type. The substitutions are categorized based on analyses of alleles in the human genome referenced to chimpanzee genome for ancestor alleles.

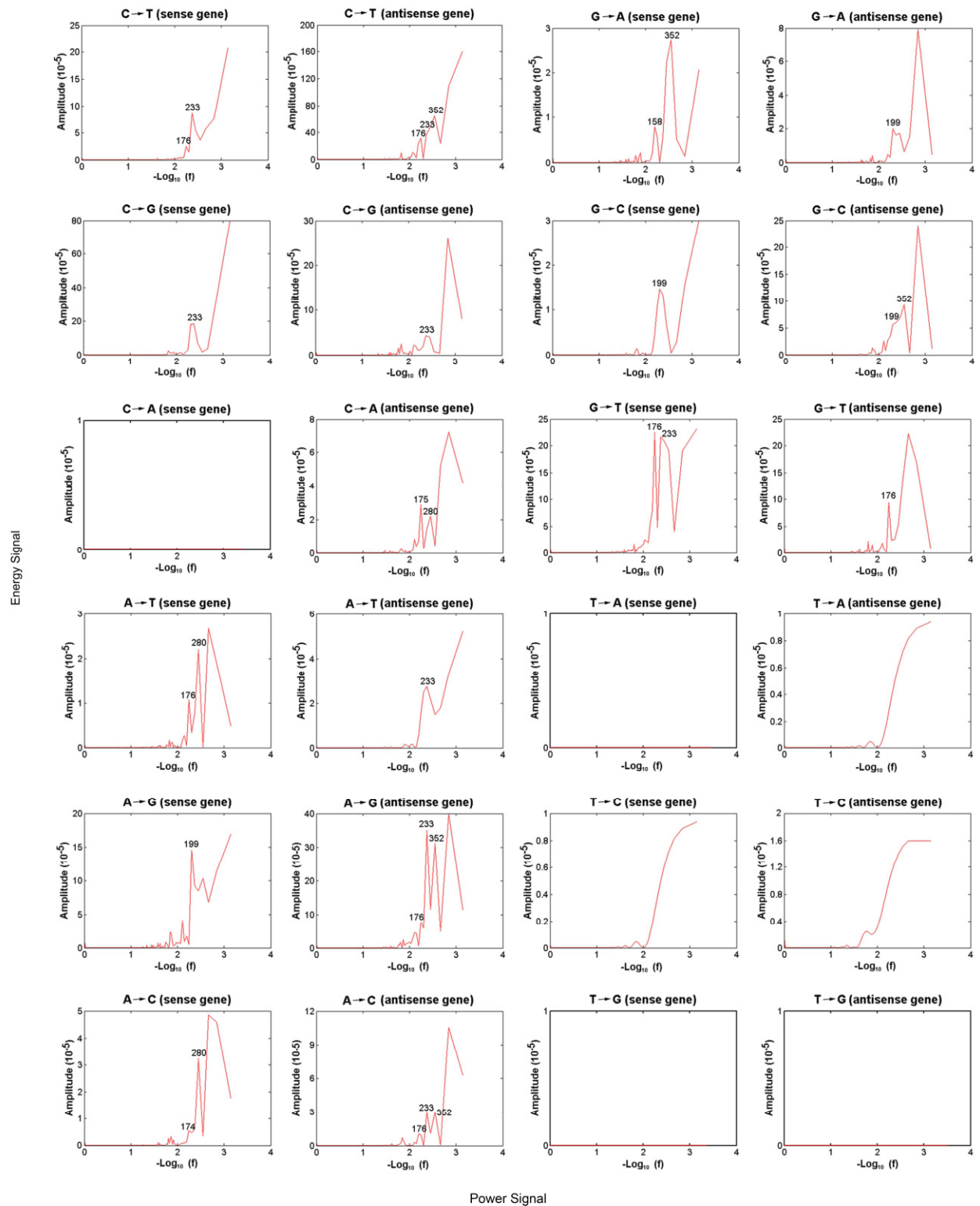




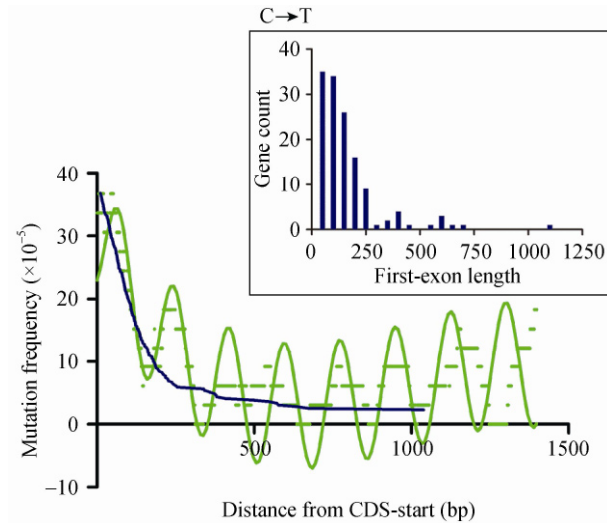
**Figure S3** Average mutation frequencies plotted across EIGs. The mutation types C→T, A→G, C→G and G→T show periodicities in these TUs.



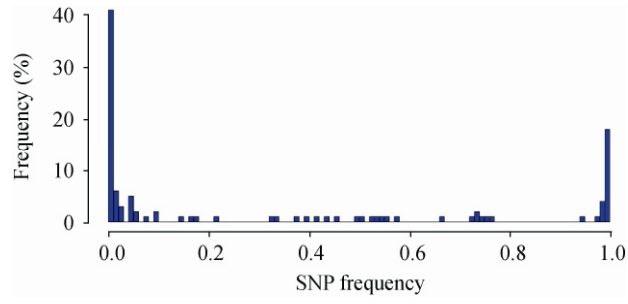
**Figure S4** Power spectrum analyses of mutation frequencies. Periodicities correspond to peaks above the baseline in the plots. The major peaks for C→T (**A**), A→G (**B**), C→G (**C**) and G→T (**D**) were found at a length of either 176 nt or 233 nt.



**Figure S5** Power spectrum analyses of mutation frequencies among EIGs. Periodicities are shown as peaks above the baseline and appear fluctuating when different gene sets (sense-strands and antisense-strands genes) are used for the analyses, probably caused by statistical errors. The orientation of the genes (Sense, 5' to 3' in a DNA strand; antisense, on the opposite strand) was determined based on the annotated chromosome strands.



**Figure S6** The mutation gradient and periodicity extending from CDS-start to intronic regions. The blue curve indicates the boundary of the first exons and introns, which suggest that the mutation gradient and periodicity are also present in intronic regions of TUs. The insets show the length distribution of the first exons.



**Figure S7** The distribution of allele frequencies in sampled populations. Most mutations are minor alleles in the populations.