SUPPLEMENTARY RESULTS

Validation of Rare Variant Results in an Independent Dataset

The dataset used in this study involved 195 genes of pharmaceutical interest (Nelson et al. 2012), and therefore may not be representative of genome-wide patterns. To test this, we made use of publicly available data from the NHLBI Exome Sequencing Project (ESP). We applied logistic regression on 603,267 singletons in this dataset (DAF = 1.4×10^{-4}), limiting to sites with ≥ 10x depth of coverage. GC content and recombination rate were calculated as before in 1kb windows surrounding each site. The regression coefficients from the exome-wide rare variant data fell within the 99% confidence intervals of the coefficients estimated from the 195 gene data (Table 2), with the following exceptions. Recombination rate has a significantly larger effect on total variants in the ESP data (Table 2). Also, the proportion of CpG GC>AT transitions was positively influenced by recombination rate for ESP variants, but negatively for the previously described rare variants (although this negative influence was not statistically significant) (Table 2). Taken together, these results show that for most variant subtypes, there was no significant difference in the way that GC content, recombination rate or DTH influence variant patterns in the 195 gene dataset compared to a larger collection of genes. Therefore, we conclude that our analysis of rare variants in the 195 gene dataset is representative of a broader sampling of genes across the genome.

Robustness of the Logistic Regression

A central premise of this study is that natural selection has limited effects on rare variants. Our sequence data cover both targeted exons and 50 bp of flanking sequence, allowing us to compare between coding and intronic rare variants. While total and CpG GC>AT rare variants had a greater conditional variant proportion in coding compared to intronic regions, the proportion for all other variant subtypes was greater in intronic regions (Supplementary Table 4). While the differences in the conditional variant proportion between coding and noncoding sites were statistically significant for most subtypes, the magnitudes of the

differences were small (average across subtypes: 0.27%). Thus, while purifying selection may have slightly reduced the absolute number of rare variants in coding regions, the relative proportion of individual variant subtypes was not substantially affected. Importantly, with regard to the main conclusions of this study, there was no significant difference (based on 99% confidence intervals) in the coefficients for GC content, recombination rate, or DTH regressions performed on coding, intronic, or the total dataset (Supplementary Table 5).

The analysis presented above used GC content calculated in 1 kb windows. To test the dependence of our results on window size, we extended the analysis for windows ranging from 100 bp – 10 kb. With the exception of CpG sites, we observed no significant difference between regression coefficients for any other variant subtype across the range of window sizes tested (based on 95% confidence intervals) (Supplementary Figure 2).

The rare variants we analyzed were derived from exome sequencing and are distributed in tight clusters, corresponding to ~2,000 targeted exons in 195 autosomal genes. Genomic features of nearby sites are often not strictly independent. To evaluate the impact of spatial dependency on the regression results, we performed a subsampling analysis using 2,000 random sites (out of ~700K sites) in each run. All observed coefficients in the original analysis fell within the 25th-75th percentile range of the coefficients from 1,000 subsampling runs for GC content (Supplementary Figure 3A), recombination rate (3B), and DTH (3C). We also examined the potential impact of between-gene heterogeneity by performing a bootstrapping analysis involving variants in new sets of 195 genes, with each set obtained by random sampling of the original 195 genes with replacement. The distribution of the coefficients from 1,000 bootstrapping runs was symmetric around the original estimates for GC content (Supplementary Figure 4A), recombination rate (4B), and DTH (4C), confirming that there was no systematic bias due to outlier genes driving the results of the regressions. In addition, as the p-values in the logistic regression were model-based, we assessed potential bias of the reported p-values by running 1,000 rounds of permutations of the variant and invariant status across sites, and

found that the p-values calculated in the regressions were consistent for GC content, recombination rate, and DTH (Supplementary Table 6).

GC content and recombination rate are positively correlated (Kong et al. 2002). To determine the extent to which our results for recombination rate and distance to hotspot could be driven by GC content, and vice versa, we performed multivariate logistic regression with two models, one using GC content and recombination rate as covariates and another using GC content and DTH as covariates. We did not observe a significant difference between the regression coefficients (based on 99% confidence intervals) estimated from the univariate (presented above) and the multivariate models for GC content, recombination rate, or DTH in rare variants (Supplementary Table 7), common variants (Supplementary Table 8), or substitutions (Supplementary Table 9).

Because GC content influences read depth in high-throughput sequencing studies, especially following target capture (Albert et al. 2007; Porreca et al. 2007), we verified that the observed influence of GC content on rare variants was not an artifact of sequencing depth. In addition to the 10x coverage filter imposed on all sites in the rare variant analysis (see Methods), we first performed logistic regression using per-base coverage as the explanatory variable. Total, AT>GC, CpG GC>AT, and GC>AT variants were significantly affected by coverage (Supplementary Table 10). Including coverage as a covariate in the regression against GC content decreased the effect of GC content on CpG GC>AT transitions, but the coefficient was still negative (Supplementary Table 10). The estimated coefficients for other variant subtypes were not affected by including coverage in the model (based on 99% confidence intervals). We concluded that coverage was not driving the results regarding the influence of GC content on rare variants.

Errors in the definition of the ancestral allele could classify variants as incorrect variant subtypes. The ancestral definitions we use are the human ancestral sequences estimated by members of the 1000 Genomes Project, based on the 4-way alignment of the human,

chimpanzee, orangutan and rhesus macaque genomes. To estimate the potential effect of variant orientation errors on our regression results, we compared the 4-way ancestral definition with the naïve orientation method using only the chimpanzee reference allele. The orientation of 5.5%, 5.1%, 3.7%, and 5.4% of AT>GC, AT>CG, GC>AT, and GC>TA variants, respectively, are flipped between these two definitions, and we take these discordance rates as the range of worst possible errors. We ran 100 simulations on rare variants, common variants, and substitutions by randomly flipping a subset of variants based on those subtype-specific percentages, and found that the results for common variants and substitutions are consistent with the original results (Supplementary Figure 5 D-I). There were stronger deviations from the original analysis for rare variants (Supplementary Figure 5 A-C). However, this is not a major concern as rare variants are not as prone to erroneous orientation as common variants and substitutions. In all, these results show that even with gross errors in the orientation of variant subtypes, the results we report for common variants and substitutions are not affected.

SUPPLEMENTARY METHODS

Multinomial Logistic Regression Analysis

We used a multinomial logistic regression model to jointly analyze the probability of all possible variant subtypes for a given allele state. The model treats the alleles observed at a site with a given reference (or ancestral) allele state (AT, GC, or CpG) as a multinomial random variable with four potential outcomes. Variant sites are defined based on the ancestral and derived allele and are categorized according to the ancestral allele state. Invariant AT and GC sites are based on the human reference genome sequence (hg18) and CpG sites are based on the ancestral sequence. Sites with an AT allele state, for example, can have one of four possible derived states: AT reference (invariant), GC, CG, or TA. We ran separate multinomial regressions for each allele state and set the invariant allele as the baseline outcome. Running regressions for AT, GC, or CpG bases separately normalizes any discrepancy between the number of sites that could produce a given variant subtype, since all of the bases in the analysis

have the potential to give rise to any of the variant subtypes. From each of these regressions, we calculated unique slope and intercept parameters for each variant subtype. Let $X > Y_i$ denote a nucleotide site with ancestral allele X and derived allele Y_i. Then, the multinomial logistic regression for an ancestral allele state X to the derived state Y_i has the form,

$$\ln\left(\frac{\Pr(X > Y_i)}{\Pr(X > X)}\right) = \alpha_{X>Y_i} + \beta_{X>Y_i} z ,$$

where $\Pr(X > Y_i)$ is the probability that a site with ancestral allele X is variant with derived allele Y_i, $\Pr(X > X)$ is the probability that a site with ancestral allele X is invariant, z is the GC content, log(recombination rate), or log(DTH) at a given site, and $\sum_i \Pr(X > Y_i) = 1$ for each nucleotide site. We used a Wald test on the β slope parameter to assess significance. We fit separate multinomial logistic regression models for each allele state for rare variants, common variants, and substitutions in order to estimate the effect of genomic context on variant subtypes in these three distinct variant classes.

Analysis of Logistic Regression Robustness

We employed three strategies to assess the robustness of the logistic regression results on rare variants: two to assess the estimated coefficients (i.e., effect size) and another to analyze statistical significance. First, we used a subsampling strategy in which we randomly sampled 2,000 sites (out of ~700 kb) and ran total logistic regression on these 2,000 sites. There are 2,126 exons in our target regions; therefore sampling 2,000 sites will generate ~1 site per exon, on average. This analysis was repeated 1,000 times. We also performed this analysis using multinomial logistic regression separately on AT, GC, and CpG ancestral sites. To further analyze the potential impact of gene-gene heterogeneity on the logistic and multinomial logistic regressions, we performed a bootstrapping analysis. We randomly re-sampled 195 autosomal genes with replacement, repeatedly generating new gene sets with the same number of genes

(195) as the full analysis, but eliminating a random subset of genes in each run. For each of the 1,000 bootstrapping gene sets, we ran the logistic regression analysis on total rare variants and the multinomial logistic regression on AT, GC, and CpG sites. Finally, we used permutations to analyze the statistical significance reported by logistic regression. For the total logistic regression, we randomly shuffled the variant and invariant sites across the ~700kb target region and performed the regression 1,000 times. We performed this same analysis for the multinomial regression separately on AT, GC, and CpG ancestral sites.

SUPPLEMENTARY TABLES

Variant Subtypes	Rare Variants			Common Variants			Substitutions		
	β	SE	Sig	β	SE	Sig	β	SE	Sig
Total	0.10	0.028	**	0.24	0.024	***	0.072	0.025	*
AT>GC	-0.052	0.058		0.22	0.047	***	0.12	0.047	
AT>CG	-0.0010	0.11		0.29	0.092	*	-0.11	0.095	
AT>TA	0.090	0.12		0.21	0.10		0.055	0.11	
CpG GC>AT	-0.016	0.066		0.0042	0.065		-0.093	0.074	
GC>AT	0.097	0.052		0.10	0.048		-0.090	0.049	
GC>TA	0.039	0.096		0.067	0.086		-0.25	0.089	*
GC>CG	-0.13	0.095		0.032	0.086		0.11	0.083	

Supplementary Table 1: Logistic and Multinomial Regression Results from 2010 deCODE Recombination Rate

The observed sloped (β), standard error (SE), and statistical significance for regressions run using a high-resolution recombination map from deCODE (Kong et al. 2010) in rare variants, common variants and substitutions. ***p-value<0.0001, **p-value<0.001, *p-value<0.01

Variant Location	Measure	AT>GC	CpG GC>AT	GC>AT	AT>CG	GC>TA	CpG GC>TA	AT>TA	GC>CG	CpG GC>CG	Total
Total	**Number of Variants**	4,778	3,951	5,338	1,215	1,594	202	1,023	1,751	201	20,053
	Number of Sites	373,983	30,956	312,986	373,983	312,986	30,956	373,983	312,986	30,956	717,925
	Conditional Variant Proportion	1.28%	12.76%	1.71%	0.32%	0.51%	0.65%	0.27%	0.56%	0.65%	2.79%
In Hotspot	**Number of Variants**	361	335	500	103	118	23	73	111	12	1,636
	Number of Sites	28,719	2,705	25,935	28,719	25,935	2,705	28,719	25,935	2,705	57,359
	Conditional Variant Proportion	1.26%	12.38%	1.93%	0.36%	0.45%	0.85%	0.25%	0.43%	0.44%	2.85%
Outside Hotspots	**Number of Variants**	4,417	3,616	4,838	1,112	1,476	179	950	1,640	189	18,417
	Number of Sites	345,264	28,251	287,051	345,264	287,051	28,251	345,264	287,051	28,251	660,566
	Conditional Variant Proportion	1.28%	12.80%	1.69%	0.32%	0.51%	0.63%	0.28%	0.57%	0.67%	2.79%

Supplementary Table 2: Rare Variant Counts Inside versus Outside of Recombination Hotspots

The counts, number of available sites, and the conditional variant proportion of all rare variants, as well as those identified inside and outside of recombination hotspots. The number of sites indicates the number of nucleotides that could produce the given variant subtype.

Variant Subtypes	Rare Variants			Common Variants			Substitutions		
	β	SE	Sig	β	SE	Sig	β	SE	Sig
Total	0.023	0.026		0.11	0.022	***	0.071	0.022	*
AT>GC	-0.018	0.055		0.15	0.043	**	0.17	0.042	***
AT>CG	0.11	0.10		0.16	0.085		-0.040	0.088	
AT>TA	-0.079	0.12		0.028	0.10		-0.13	0.11	
CpG GC>AT	-0.038	0.061		0.019	0.051		0.018	0.059	
GC>AT	0.13	0.047	*	0.064	0.042		-0.042	0.044	
GC>TA	-0.12	0.096		-0.21	0.086		0.086	0.075	
GC>CG	-0.29	0.098	*	0.14	0.073		-0.033	0.076	

Supplementary Table 3: Logistic and Multinomial Regression Results from Inside vs. Outside Recombination Hotspots

The observed sloped (β), standard error (SE), and statistical significance for regressions run using a “inside” versus “outside” of a recombination hotspot as the explanatory variable in rare variants, common variants and substitutions. ***p-value<0.0001, **p-value<0.001, *p-value<.01

Variant Type	Coding		Intronic		p-value	
Total	8,738	(2.85%)	4,642	(2.70%)	0.0033	*
AT>GC	1,764	(1.19%)	1,245	(1.32%)	0.0028	*
AT>CG	398	(0.27%)	324	(0.34%)	0.00077	**
AT>TA	362	(0.24%)	249	(0.26%)	0.32	
CpG GC>AT	2,525	(13.28%)	551	(11.98%)	0.020	
GC>AT	2,147	(1.55%)	1,328	(1.81%)	0.0000037	***
GC>TA	746	(0.47%)	451	(0.58%)	0.00062	**
GC>CG	796	(0.50%)	494	(0.64%)	0.000056	***

Supplementary Table 4: Comparison of Rare Variant Counts in Coding and Intronic Sequences

Counts of variants identified in coding and flanking intronic regions. Numbers in parenthesis show the conditional variant proportion of each variant subtype, defined as the number of variants of the subtype divided by the number of total sites that could produce the given variant. The p-values from a two-proportion t-test performed in conditional variant proportion are also presented.

Variant Subtype	Model								
	All Sites			Coding Sites			Intronic Sites		
	GC Content								
Total	0.68	(0.069)	***	0.61	(0.10)	***	0.74	(0.15)	***
AT>GC	-1.048	(0.15)	***	-1.14	(0.24)	***	-1.14	(0.31)	**
AT>CG	-0.56	(0.29)		-1.41	(0.50)	*	-0.19	(0.58)	
AT>TA	-0.98	(0.32)	*	-0.68	(0.51)		-0.82	(0.67)	
CpG GC>AT	-2.64	(0.17)	***	-2.62	(0.20)	***	-1.91	(0.48)	***
GC>AT	0.024	(0.14)		0.40	(0.21)		-0.39	(0.28)	
GC>TA	-0.80	(0.25)	*	-0.77	(0.38)		-0.22	(0.49)	
GC>CG	-0.53	(0.24)		-1.10	(0.37)	*	-0.45	(0.47)	
	Recombination Rate								
Total	0.15	(0.043)	**	0.15	(0.063)		0.29	(0.087)	**
AT>GC	0.014	(0.089)		-0.091	(0.14)		0.27	(0.17)	
AT>CG	-0.014	(0.18)		-0.55	(0.30)		0.15	(0.33)	
AT>TA	-0.065	(0.19)		-0.46	(0.32)		0.32	(0.38)	
CpG GC>AT	-0.13	(0.10)		-0.073	(0.12)		-0.12	(0.25)	
GC>AT	0.19	(0.081)		0.33	(0.12)	*	0.18	(0.16)	
GC>TA	0.024	(0.15)		-0.14	(0.23)		0.15	(0.28)	
GC>CG	0.054	(0.14)		0.16	(0.22)		0.14	(0.27)	
	DTH								
Total	-0.042	(0.011)	**	-0.069	(0.017)	***	-0.027	(0.022)	
AT>GC	-0.025	(0.023)		-0.0086	(0.038)		0.014	(0.044)	
AT>CG	-0.060	(0.044)		0.0076	(0.079)		-0.043	(0.086)	
AT>TA	0.023	(0.049)		0.093	(0.084)		-0.015	(0.10)	
CpG GC>AT	-0.047	(0.025)		-0.11	(0.031)	**	0.067	(0.062)	
GC>AT	-0.089	(0.021)	***	-0.085	(0.034)		-0.096	(0.040)	
GC>TA	-0.054	(0.039)		-0.088	(0.061)		-0.035	(0.072)	
GC>CG	0.025	(0.037)		0.012	(0.060)		0.0042	(0.068)	

Supplementary Table 5: Comparison of Regression Results for Rare Variants In All Sites, Coding Sites, and Intronic Sites

β coefficients, standard error (in parenthesis), and significance from the regression on all sites, coding sites, and intronic sites. ***p-value<0.0001, **p-value<0.001, *p-value<.01

Variant Subtype	Model	
	Model-Based P-Value	Empirical (One-Sided) P-Value
	GC Content	
Total	$<2x10^{-16}$	$<1x10^{-3}$
AT>GC	$2.51x10^{-12}$	$<1x10^{-3}$
AT>CG	0.054	0.025
AT>TA	$2.28x10^{-3}$	0.001
CpG GC>AT	$<2x10^{-16}$	$<1x10^{-3}$
GC>AT	0.86	0.46
GC>TA	$1.15x10^{-3}$	0.001
GC>CG	0.024	0.009
	Recombination Rate	
Total	$3.58x10^{-4}$	0.001
AT>GC	0.87	0.47
AT>CG	0.94	0.49
AT>TA	0.74	0.38
CpG GC>AT	0.16	0.082
GC>AT	0.019	0.012
GC>TA	0.87	0.46
GC>CG	0.70	0.40
	DTH	
Total	$1.61x10^{-4}$	$<1x10^{-3}$
AT>GC	0.27	0.17
AT>CG	0.18	0.10
AT>TA	0.65	0.33
CpG GC>AT	0.059	0.028
GC>AT	$2.39x10^{-5}$	$<1x10^{-3}$
GC>TA	0.16	0.087
GC>CG	0.49	0.23

Supplementary Table 6: Comparison of Model-Based and Empirical P-values calculated from 1000 Permutations of Variant and Invariant Sites in Rare Variants

Variant Subtype	Model								
	Univariate			GC + Recombination			GC + DTH		
	GC Content								
Total	0.68	(0.069)	***	0.66	(0.070)	***	0.69	(0.069)	***
AT>GC	-1.048	(0.15)	***	-1.090	(0.15)	***	-1.05	(0.15)	***
AT>CG	-0.56	(0.29)		-0.58	(0.30)		-0.57	(0.29)	
AT>TA	-0.98	(0.32)	*	-0.99	(0.33)	*	-0.98	(0.32)	*
CpG GC>AT	-2.64	(0.17)	***	-2.64	(0.17)	***	-2.64	(0.17)	***
GC>AT	0.024	(0.14)		-0.014	(0.14)		0.027	(0.14)	
GC>TA	-0.80	(0.25)	*	-0.82	(0.25)	**	-0.80	(0.25)	*
GC>CG	-0.53	(0.24)		-0.55	(0.24)		-0.53	(0.23)	
	Recombination Rate								
Total	0.15	(0.043)	**	0.094	(0.043)		-	-	-
AT>GC	0.014	(0.089)		0.13	(0.092)		-	-	-
AT>CG	-0.014	(0.18)		0.048	(0.18)		-	-	-
AT>TA	-0.065	(0.19)		0.042	(0.20)		-	-	-
CpG GC>AT	-0.13	(0.010)		-0.044	(0.098)		-	-	-
GC>AT	0.19	(0.081)		0.19	(0.081)		-	-	-
GC>TA	0.024	(0.15)		0.086	(0.15)		-	-	-
GC>CG	0.054	(0.14)		0.095	(0.14)		-	-	-
	DTH								
Total	-0.042	(0.011)	**	-	-	-	-0.042	(0.011)	**
AT>GC	-0.025	(0.023)		-	-	-	-0.028	(0.023)	
AT>CG	-0.060	(0.044)		-	-	-	-0.061	(0.045)	
AT>TA	0.023	(0.049)		-	-	-	0.020	(0.049)	
CpG GC>AT	-0.047	(0.025)		-	-	-	-0.029	(0.026)	
GC>AT	-0.089	(0.021)	***	-	-	-	-0.089	(0.021)	***
GC>TA	-0.054	(0.039)		-	-	-	-0.054	(0.039)	
GC>CG	0.025	(0.037)		-	-	-	0.026	(0.037)	

Supplementary Table 7: Comparison of Logistic Regression Results for Rare Variants between Univariate and Multivariate Models

β coefficients, standard error (in parenthesis), and significance for GC content, recombination rate, and DTH. Results are shown for univariate and multivariate logistic regression models. ***p-value<0.0001, **p-value<0.001, *p-value<.01

Variant Subtype	Model								
	Univariate			GC + Recombination			GC + DTH		
	GC Content								
Total	-0.18	0.059	*	-0.77	0.064	***	-0.27	0.060	***
AT>GC	-0.46	0.12	**	-1.053	0.13	***	-0.56	0.12	***
AT>CG	0.070	0.24		-0.51	0.26		-0.053	0.24	
AT>TA	-1.63	0.28	***	-2.086	0.30	***	-1.73	0.28	***
CpG GC>AT	-3.82	0.15	***	-4.32	0.16	***	-3.88	0.15	***
GC>AT	-1.65	0.12	***	-2.21	0.13	***	-1.73	0.12	***
GC>TA	-2.46	0.22	***	-2.94	0.23	***	-2.52	0.22	***
GC>CG	-1.48	0.21	***	-2.0082	0.22	***	-1.58	0.21	***
	Recombination Rate								
Total	0.95	0.039	***	1.12	0.042	***	-	-	-
AT>GC	0.78	0.076	***	1.021	0.082	***	-	-	-
AT>CG	0.90	0.15	***	1.017	0.16	***	-	-	-
AT>TA	0.28	0.17		0.76	0.18	***	-	-	-
CpG GC>AT	0.30	0.10	*	1.036	0.11	***	-	-	-
GC>AT	0.65	0.077	***	1.11	0.082	***	-	-	-
GC>TA	0.30	0.14		0.92	0.15	***	-	-	-
GC>CG	0.64	0.14	***	1.049	0.14	***	-	-	-
	DTH								
Total	-0.15	0.011	***	-	-	-	-0.16	0.011	***
AT>GC	-0.14	0.021	***	-	-	-	-0.15	0.021	***
AT>CG	-0.19	0.041	***	-	-	-	-0.19	0.041	***
AT>TA	-0.10	0.046		-	-	-	-0.14	0.046	*
CpG GC>AT	-0.10	0.028	**	-	-	-	-0.13	0.027	***
GC>AT	-0.14	0.021	***	-	-	-	-0.17	0.021	***
GC>TA	-0.075	0.039		-	-	-	-0.11	0.039	*
GC>CG	-0.18	0.037	***	-	-	-	-0.20	0.037	***

Supplementary Table 8: Comparison of Univariate and Multivariate Logistic Regression Results for Common Variants

β coefficients, standard error (in parenthesis), and significance GC content, recombination rate, and DTH. Results are shown for univariate and multivariate logistic regression models. ***p-value<0.0001, **p-value<0.001, *p-value<.01

Variant Subtype	Model								
	Univariate			GC + Recombination			GC + DTH		
	GC Content								
Total	0.056	(0.059)		-0.13	(0.064)		0.033	(0.060)	
AT>GC	0.35	(0.12)	*	0.038	(0.13)		0.31	(0.12)	*
AT>CG	3.14E-04	(0.23)		-0.14	(0.24)		0.026	(0.23)	
AT>TA	-1.27	(0.28)	***	-1.16	(0.30)	**	-1.28	(0.28)	***
CpG GC>AT	-3.74	(0.17)	***	-4.045	(0.18)	***	-3.77	(0.17)	***
GC>AT	-1.41	(0.12)	***	-1.42	(0.12)	***	-1.41	(0.12)	***
GC>TA	-1.97	(0.21)	***	-1.94	(0.22)	***	-1.97	(0.21)	***
GC>CG	-0.49	(0.20)		-0.66	(0.21)	*	-0.52	(0.20)	*
	Recombination Rate								
Total	0.34	(0.040)	***	0.37	(0.043)	***	-	-	-
AT>GC	0.57	(0.076)	***	0.56	(0.081)	***	-	-	-
AT>CG	0.21	(0.15)		0.24	(0.16)		-	-	-
AT>TA	-0.47	(0.18)	*	-0.19	(0.19)		-	-	-
CpG GC>AT	-0.061	(0.12)		0.63	(0.12)	***	-	-	-
GC>AT	-0.29	(0.078)	**	0.014	(0.083)		-	-	-
GC>TA	-0.49	(0.14)	**	-0.063	(0.15)		-	-	-
GC>CG	0.22	(0.13)		0.35	(0.14)		-	-	-
	DTH								
Total	-0.047	(0.011)	***	-	-	-	-0.046	(0.011)	***
AT>GC	-0.086	(0.021)	***	-	-	-	-0.080	(0.021)	**
AT>CG	0.042	(0.041)		-	-	-	0.042	(0.042)	
AT>TA	0.014	(0.049)		-	-	-	-0.013	(0.049)	
CpG GC>AT	-0.048	(0.032)		-	-	-	-0.080	(0.031)	
GC>AT	0.014	(0.022)		-	-	-	-0.0066	(0.022)	
GC>TA	0.031	(0.039)		-	-	-	0.0021	(0.039)	
GC>CG	-0.069	(0.037)		-	-	-	-0.076	(0.037)	

Supplementary Table 9: Comparison of Univariate and Multivariate Logistic Regression Results for Substitutions

β coefficients, standard error (in parenthesis), and significance for GC content, recombination rate, and DTH. Results are shown for univariate and multivariate logistic regression models. ***p-value<0.0001, **p-value<0.001, *p-value<.01

Variant Type	Model					
	Univariate Model			Multivariate Model		
	GC Content					
Total	0.68	(0.069)	***	0.86	(0.072)	***
AT>GC	-1.048	(0.15)	***	-1.011	(0.15)	***
AT>CG	-0.56	(0.29)		-0.57	(0.29)	
AT>TA	-0.98	(0.32)	*	-0.98	(0.32)	*
CpG GC>AT	-2.64	(0.17)	***	-1.42	(0.20)	***
GC>AT	0.024	(0.14)		0.22	(0.14)	
GC>TA	-0.80	(0.25)	*	-0.81	(0.26)	*
GC>CG	-0.53	(0.24)		-0.44	(0.25)	
	Coverage					
Total	6.39E-03	(8.32E-04)	***	8.72E-03	(8.51E-04)	***
AT>GC	9.020E-03	(1.74E-03)	***	8.22E-03	(1.75E-03)	***
AT>CG	-2.88E-05	(3.41E-03)		-6.18E-04	(3.43E-03)	
AT>TA	1.58E-03	(3.72E-03)		6.91E-04	(3.75E-03)	
CpG GC>AT	3.41E-02	(1.82E-03)	***	2.65E-02	(2.11E-03)	***
GC>AT	6.31E-03	(1.58E-03)	***	7.081E-03	(1.66E-03)	***
GC>TA	2.42E-03	(2.87E-03)		-3.88E-04	(3.015E-03)	
GC>CG	5.073E-03	(2.74E-03)		3.54E-03	(2.88E-03)	

Supplementary Table 10: Comparison of Rare Variant Regression Results for Univariate and Multivariate GC Content and Coverage Regressions
β coefficients, standard error (in parenthesis), and significance from the univariate regression models for GC content and coverage and multivariate model, using GC content and coverage as covariates in the regression model. ***p-value<0.0001, **p-value<0.001, *p-value<0.01

SUPPLEMENTARY REFERENCES

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 4(11): 903-905.

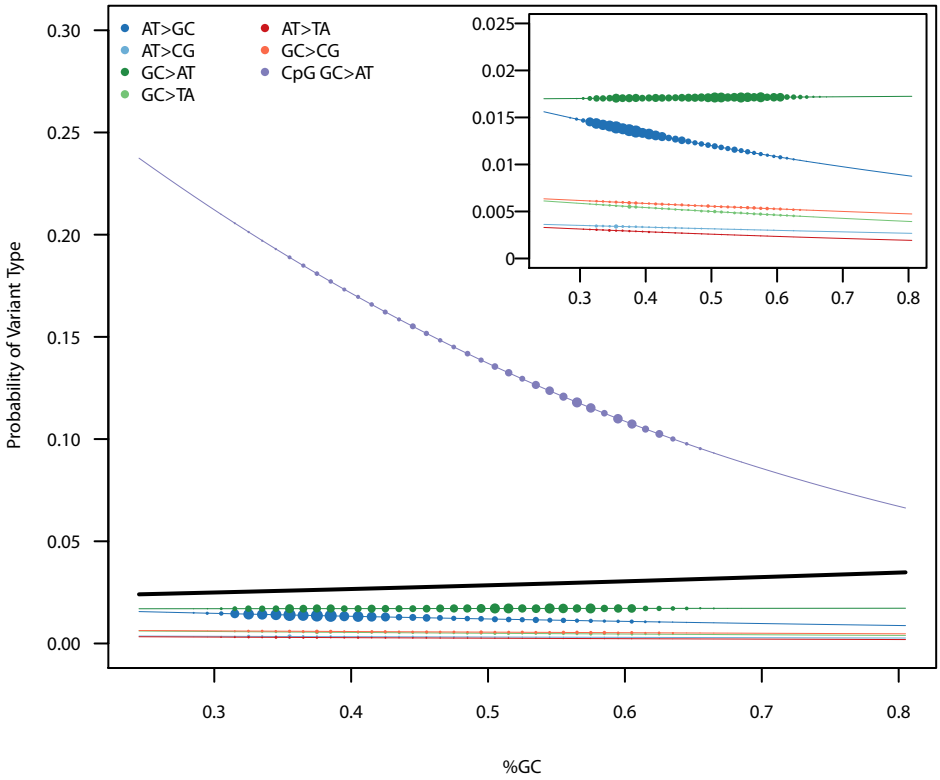
Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* 31(3): 241-247.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Gylfason A, Kristinsson KT, Gudjonsson SA et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319): 1099-1103.

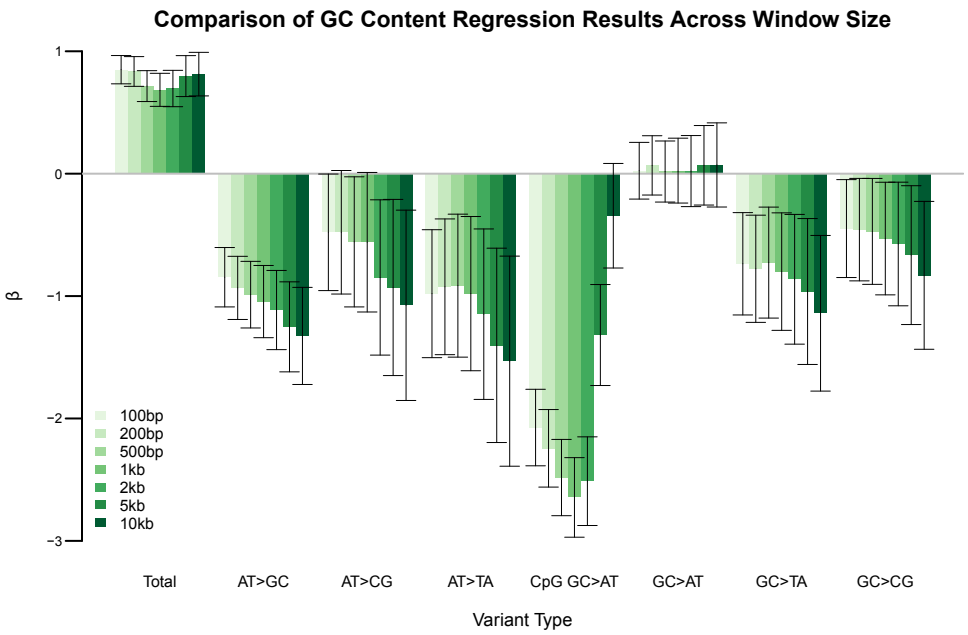
Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D et al. 2012. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* 337(6090): 100-104.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F et al. 2007. Multiplex amplification of large sets of human exons. *Nat Methods* 4(11): 931-936.

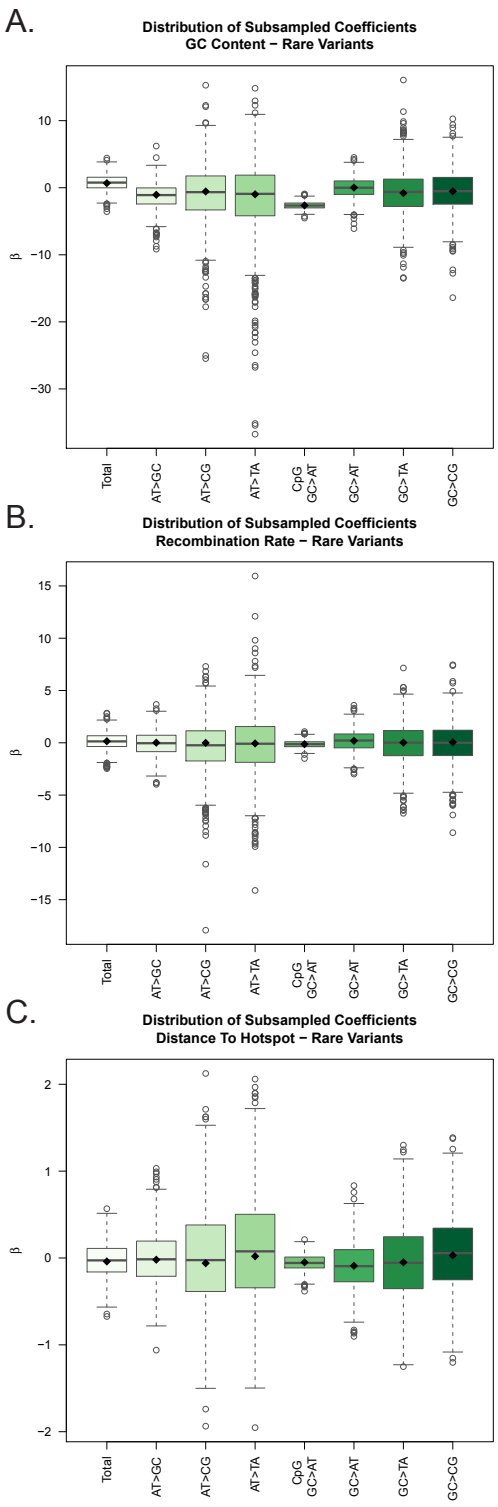
SUPPLEMENTARY FIGURES



Supplementary Figure 1: Difference in effect of GC content on rare variants between total variants and individual variant subtypes.
This plot shows the fitted logistic regression curves for a given variant subtype across observed GC content. The probability for total variants is shown in black. Point size corresponds to the proportion of the given variant subtype in each GC content bin. While most of the variant subtypes have a negative relationship between probability of occurrence and GC content, the trend between the overall probability of observing a rare variant and GC content is positive. This is driven by the increased mutation rate of CpG dinucleotides and the uneven distribution of CpG GC>AT and AT>GC variants across GC content. The inset shows the portion of the plot with variant probability ≤ 0.025 for all GC content bins to provide a better view of the probability across GC content for non-CpG-induced variants.

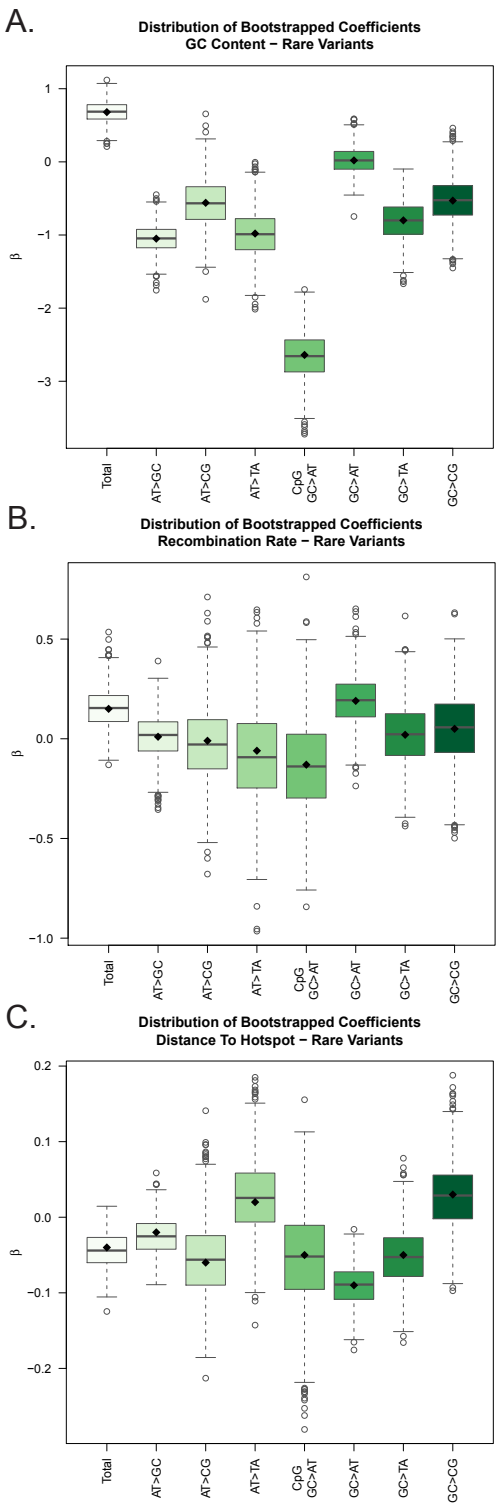


Supplementary Figure 2: Sensitivity analysis for rare variants with varying GC content and recombination rate window sizes.
We compared regression analysis for GC content using window sizes of 100 bp, 200 bp, 500 bp, 2 kb, 5 kb, and 10 kb to the original 1 kb analysis. The barplots show the estimated regression coefficients for each of the window sizes including the 1 kb described in the results. Error bars represent 95% confidence intervals for each regression coefficient.

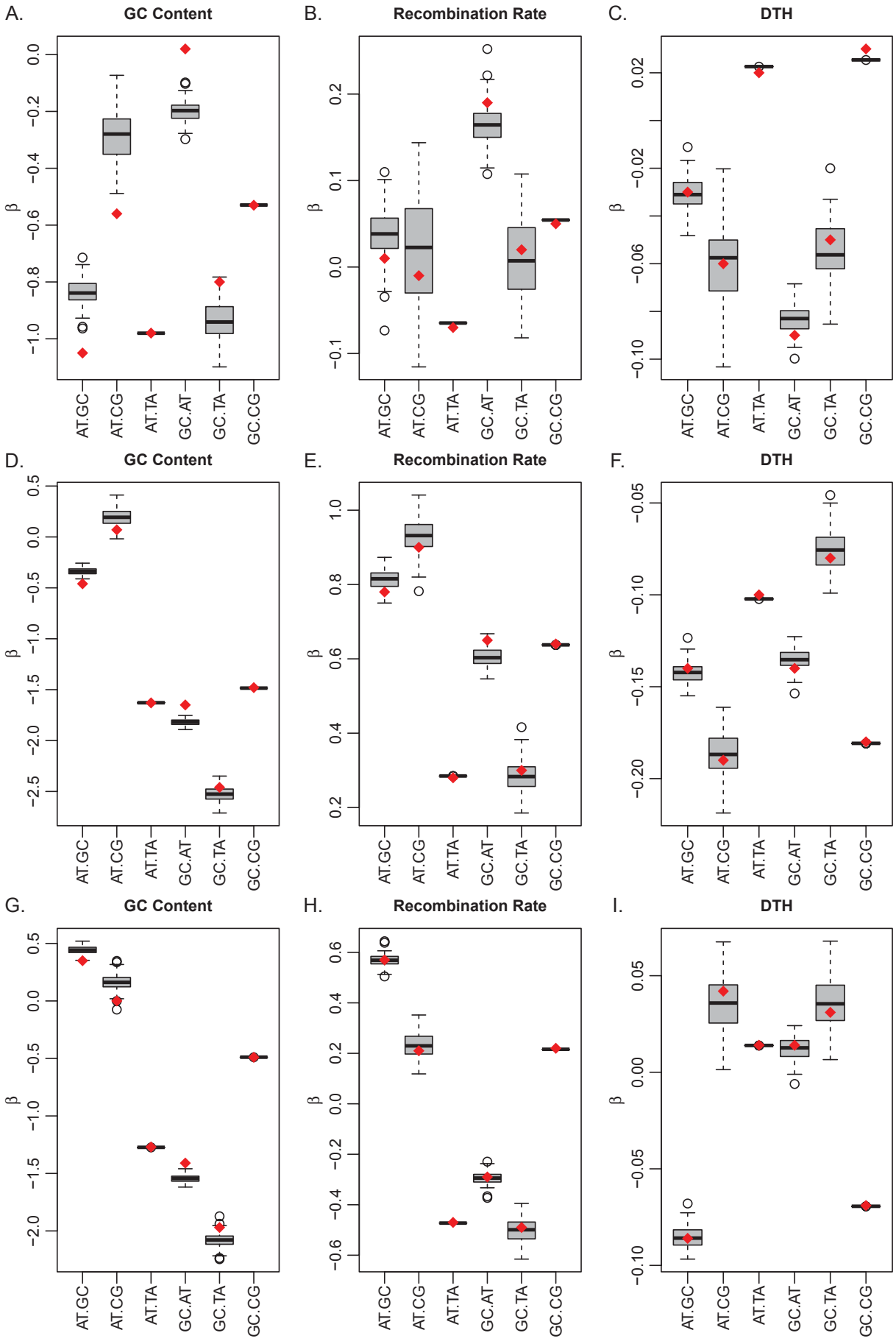


Supplementary Figure 3: Distribution of estimated regression coefficients from subsampling analysis.

This plot shows the distribution of estimated regression coefficients from the 1,000 subsampling analyses for (A) GC content, (B) recombination rate, and (C) DTH for rare variants. Black diamonds indicate the coefficients obtained in the original analysis.



Supplementary Figure 4: Distribution of estimated regression coefficients from bootstrapping analysis.
This plot shows the distribution of estimated regression coefficients over the 1,000 bootstrapping analyses for (A) GC content, (B) recombination rate, and (C) DTH for rare variants. Black diamonds show the coefficients obtained in the original analysis.


A.
GC Content
B.
Recombination Rate
C.
DTH
D.
GC Content
E.
Recombination Rate
F.
DTH
G.
GC Content
H.
Recombination Rate
I.
DTH
β
AT.GC
AT.CG
AT.TA
GC.AT
GC.TA
GC.CG

Supplementary Figure 5: Distribution of β Coefficients in Datasets Simulating Error in Variant Orientation

Results from the analysis simulating error in the orientation of the AT>GC, AT>CG, GC>AT, and GC>TA variants based on the chimpanzee allele. Barplots showing the distribution of the coefficients from the error-simulated regressions are shown for rare variants (A-C), common variants (D-F) and substitutions (G-I). The red diamonds show the coefficients estimated from the original analysis.