1. Details on ‘Genomic elements under strong purifying selection’

Metrics used to quantify negative selection

Common metrics used to quantify negative selection, such as SNP density, heterozygosity and evolutionary conservation score (GERP score), are examined in this paper. One concern with SNP density is the potential GC bias in low-coverage sequencing data. Generally, GC enriched regions have lower sequencing coverage compared to GC depleted regions. Thus, the low SNP density in GC enriched categories is an artificial effect because of lacking power to detect variants rather than a sign of strong negative selection [Supp Figure S 7]. The same bias affects heterozygosity metrics. On the contrary, the GC bias has an opposite effect on fraction of rare variants. We tend to underestimate fraction of rare variants in GC enriched categories [Supp Figure S 8]. From this perspective, using fraction of rare variants is a better metrics to quantify recent negative selection in human populations, as we do not tend to have false positives. We also compared the fraction of rare variants with average GERP scores of underlying sequences and they demonstrate significantly positive correlation (r= 0.49, p=3e-4), suggesting the recent negative selection in human populations aligned well with evolutionary constraints [Supp Figure S 3].

Impact of sample size on identification of differential purifying selection in various functional categories

Individuals were randomly selected in sets of 100, 200, 300 ….1000 from the 1,092 samples. For each set the fraction of rare alleles in various functional categories was computed. The calculation was repeated 100 times for each set [Supp Figure S 1]. At small sample sizes, it is harder to distinguish between truly common and rare SNPs in the general population because the estimate of the prevalence of derived alleles is limited by the sample size. Thus, the occurrence of common polymorphisms in the population is underestimated in small samples due to incomplete sampling. By sub-sampling individuals from Phase I samples, we find that different functional categories are effected to varying degrees by sample size [Supp Figure S 1]. As sample size increases, the fraction of SNPs with low allele count in categories under weak purifying selection (for example, pseudogenes) decreases. This is because categories under weak selection harbor mostly common SNPs, whose derived allele count is severely underestimated at small sample size. In strong contrast, the fraction of rare SNPs in categories under strong selection (for example, SNPs introducing prematureStop codons) remains relatively constant with increasing sample size. This is because these categories harbor mostly rare SNPs whose derived allele count remains largely unaffected by greater sampling. As a result, with increasing sample size, the gap between the strength of negative selection (fraction of rare alleles) amongst different categories increases. The availability of 1,092 samples allows a clear separation of functional categories whose differences were either absent or only subtle at smaller sample sizes (for example, motifs of TF families HMG and MADs-box).

Sources of various gene categories

Most of the analyses are based on a consistent set of gene categories from a defined set of databases and/or references. The ‘LoF-tol’ gene category is a curated list of loss-of-function genes from phase 1 of the 1000 Genomes Project (*6*) that are non-pathogenic even in the homozygous state (hence **L**oss-**o**f-**F**unction **tol**erant). ‘Recessive’ and ‘Dominant’ genes are derived from Blekhman et. al. (*7*), which are curated from the Online Mendelian Inheritance in Man (OMIM) database. The ‘GWAS’ genes are extracted from the NHGRI catalog of published genome-wide association studies (*8*). The ‘essential’ genes are obtained from the Database of Essential Genes (DEG) version 5.0 (*9-11*). Genes in the ‘Cancer’ category are obtained from the Cancer Gene Census (*12*).

Identification of non-coding categories under purifying selection

Non-coding functional categories

Non-coding annotations used include ncRNAs, UTRs, transcription factor (TF) peaks, TF motifs, DNase I hypersensitivity sites (DHSs), enhancers and pseudogenes. ncRNAs are further divided into miRNA, snRNA, snoRNA, rRNA, lincRNA and miscellaneous RNA. ncRNAs, UTRs and pseudogenes are obtained from Gencodev7 (*1*). TF peaks, motifs, DHSs and enhancers are obtained from Encode Integrative paper release (*2*). In total, there are 88 sequence-specific TFs (TFSSs), 16 general TFs (like Pol2- and Pol3-associated factors), and 15 chromatin-associated factors. The classification of TFs into different families is as described by Vaquerizas et al and is based on the presence of DNA binding domains from the Interpro database (*3*). Details of the classification are also discussed in Gerstein et al (*4*). A conservative set of enhancer elements is used which consists of intersection of those obtained using combined ChromHMM/Segway segmentation (*2*) with distal regulatory modules obtained by discriminative training (*5*). A schema of the various sub-categories is presented in [Supp Figure S 2].

Quantify purifying selection of non-coding categories using fraction of rare variants.

Fraction of rare variants based on 1000 genomes phase 1 SNP data is used to estimate recent negative selection in human populations on non-coding categories. Higher fraction of rare variants suggests higher selection constraint. To reduce allele frequency bias due to sequencing coverage, we limited our analysis to 1000 genomes phase 1 low coverage SNPs found in ‘P’ sites of strict mask (*13*). Rare variants are defined as those with derive allele frequency less than 0.5% (DAF < 0.5%). Fraction of rare variants for each category is calculated as number of rare ones divided by total number of variants. Variants without ancestral state assignment are excluded from our analysis.

Ekta Khurana 12/12/12 5:16 PM
Deleted: or

Ekta Khurana 12/12/12 5:16 PM
Deleted: s

Significance estimation and Randomization

With the goal of identifying strongly negative-selected non-coding categories, we compared the fraction of rare variants with that of non-coding average. In addition to directly compare the relative values, the size effect is another important consideration. Categories with less SNPs will demonstrate high variation of fraction of rare variants compared to categories with more SNPs. Thus to quantify significance against non-coding average, a binomial test is used to capture of the effect of data scarcity, assuming the possibility of a SNP classified as rare or not following binomial distribution. Enrichment of rare variants of 677 categories against non-coding average were tested. To deal with multiple hypothesis correction, a randomization procedure was developed considering the specific dependency structure of different categories, for example, faction of binding peaks of BRF1 overlap with that of ZZZ3. Instead of randomly shuffling coordinates, all categories slide together along the genome to retain the relative positions [Supp Figure S 4]. For each sliding process, fraction of rare variants is recalculated based on the new coordinates for each category. This process is repeated 1,000 times resulting a distribution of fractions of rare variants for each category. Empirical P values were obtained comparing original fraction of rare variants with randomized distribution. We found that the randomized P values correlate well with the binomial P values [Supp Figure S 5], suggesting the binomial distribution assumption of rare variants is appropriate.

Ekta Khurana 12/12/12 5:18 PM
Deleted: [Supplement Figure 4]

Ekta Khurana 12/12/12 5:18 PM
Deleted: [Supplement Figure 5]

False discovery rate of the multiple hypothesis testing was calculated as (*14*):

$$FDR = \frac{E(R^0)}{R}$$

$$E(R^0) = \frac{1}{B}\sum_{b=1}^{B} R^0$$

with $R^{0b} = \#\{FP^b\}$, B is the number of randomization and R is the number of categories passing cut-offs. After setting FDR to 1.3%, 101 categories are found to be significantly enriched of rare variants.

Defining sensitive regions

Among the 101 significant categories, we defined categories constituting ~0.02% and ~0.4% of the genome with highest fraction of rare variants as “Ultra-Sensitive” and “Sensitive” regions (5 and 24 categories respectively) [Supp Figure S 6]. Mutations in these regions are more likely to be deleterious, as they are selected against variants.

Ekta Khurana 12/12/12 5:19 PM
Deleted: [Supplement Figure 6]

Ekta Khurana 12/12/12 5:19 PM
Deleted: We suspects

Ekta Khurana 12/12/12 5:19 PM
Deleted: m

Ekta Khurana 12/12/12 5:19 PM
Deleted: happened

Allelic SNPs and eQTLs

For the allele-specific analyses, we divided the SNPs found in the individual NA12878 into three categories: those that are allele-specific and found in ChIP-seq and/or RNA-seq peaks (AS), those non-allele-specific but found in peaks (non-AS) and those not found in peaks (non-peaks). The list of allele-specific SNPs found in ChIP-seq and RNA-seq peaks (specific to NA12878) are generated from the AlleleSeq pipeline by Rozowsky et al. (*15*).

Ekta Khurana 12/12/12 6:47 PM
Deleted: t.

The 14,812 eQTL SNPs are obtained from Montgomery et. al. (16). For comparison, the matched SNPs were selected to be located within 1Mb of a gene and matched for allele frequency and distance from transcription start site. For eQTL enrichment analysis, the eQTL SNPs were compared against the matched set of SNPs for all the functional categories; odds ratios and p-values are obtained from Fisher’s exact tests.

Ekta Khurana 12/12/12 6:47 PM
Deleted:

Genes showing tissue-specific expression

The tissue-specificity of protein-coding genes was analyzed by RNA-sequencing of 18 human tissue samples from the Ambion FirstChoice Human Total RNA Survey Panel. Library preparation and sequencing of 49bp paired-end read was done with Illumina HiSeq according to the manufacturer's instructions, the reads were mapped to hg19 with bwa, and gene expression levels were measured as RPKM from read counts in coding regions of Gencode v10 genes. Out of the total 19290 expressed genes in at least one tissue of this dataset and of these a total of 11719 (137-2626 per tissue) showed a tissue-specific pattern of expression defined as in (*17*).

Tissue-specific DNase hypersensitive sites

DNase hypersensitive sites of 125 cell types are obtained from Thurman et al. (*18*). After excluding cancer cell types, normal cell lines are grouped into 25 tissues according to Encode common cell types information. Tissue specific DHS sites for each tissue are defined as those that occur only in that tissue and are absent from other tissues.

Ekta Khurana 12/12/12 6:55 PM
Deleted: ,

Ekta Khurana 12/12/12 6:55 PM
Deleted: but

2. Details on 'Purifying selection in the human proteome and regulome'

Source of Interaction data

Binary protein-protein interactions were obtained from InWeb (*19*) and HINT (*20*). Regulatory interactions were obtained from Gerstein et al, 2012 (*4*).

Structural Interaction network (SIN) construction and analysis to find SNPs at interaction interfaces

For SIN construction, protein-protein interaction (PPI) network is curated and filtered from HPRD (Human Protein Reference Database) and MIPS database, containing 39,849 interactions between 7,432 proteins (*21*). For each protein, the domain information is obtained from Pfam. Pfam domain-domain interactions (DDI) and residue level interactions between protein domains in PDB are obtained from iPfam (release 20.0). Domain-domain interactions are mapped onto protein-protein interaction network through the protein-domain relationships. Interactions that are supported by both DDI and PPI are included in the SIN. Generally speaking, SIN has the interacting domain information in corresponding protein-protein interactions. SIN contains 11,433 domain interactions between 2,262 proteins. The presence of missense SNPs is then checked in the list of amino acid residues at interaction interfaces.

Atomic resolution structural interaction network

To construct an atomic-resolution human protein interactome network, we compiled all available high-quality co-crystal structures from Protein Data Bank (PDB) (*22*). Atomic-resolution interaction interfaces were identified using these co-crystal structures - we used a water molecule of diameter 1.4Å as the probe and calculated the relative solvent accessible surface areas of the interacting pair as well as the individual proteins involved in the interaction (*23*). Any residues whose relative accessibilities change by more than 1Å2 are considered as potential interface residues. Amino acids at the interface are on the surface of the corresponding proteins, but tend to be buried in the co-crystal structure where the two proteins are in a bound configuration. So, for all interface residues, there should be a significant change in accessible surface area when comparing the bound and unbound states (*24*). Furthermore, we required that interface residues be present at the surface of the corresponding proteins. We calculated the fraction of surface area for each residue in the corresponding proteins without their interaction partners accessible to the water molecule probe. If more than 15% of the total surface area is accessible to the water molecule for a particular residue, we define it to be at the surface, else it is considered to be buried (*24, 25*). Using these two criteria, we obtain a set of 89,075 residues that represent the interface for 2,069 interactions as determined by 5,549 atomic-resolution co-crystal structures (*26*). With our atomic-resolution interactome network, we calculated the enrichment of all phase I SNPs at the interaction interface, the remainder of the interacting domain, and the rest of the protein. We find that while rare variants are enriched significantly both at the interface and in the remainder of the interacting domain (odds ratio = 1.16, P = 0.001 for interface residues; odds ratio = 1.15, P < 10-5 for remainder of the interacting domain), common variants (DAF >= 0.5%) are enriched significantly outside interacting domains (odds ratio = 1.08, P = 0.0005 outside interacting domains). Our analysis provides a molecular mechanistic explanation for the differences in the way these two polymorphisms act – common variants are evolutionarily prone to remain away from the interaction interface as these occur frequently in the population and are unlikely to have any deleterious functional consequences.

Ekta Khurana 12/12/12 11:26 PM
Deleted: (AF < 0.05%)

Validation by yeast two-hybrid experiments

To test the functional consequences of rare variants, we cloned three different polymorphisms into Wiskott Aldrich Syndrome protein (*WAS*) – R41G, E131K and I294T. E131K is both a rare variant (DAF < 0.5%) and a known HGMD disease mutation, whereas I294T and R41G are known HGMD disease mutations but have not been detected as SNPs. E131K and R41G occur within the WH1 domain on *WAS* while I294T is in the PBD domain. We examined the effects of these polymorphisms on *WAS* interactions using a yeast two-hybrid (Y2H) system. We found that while wild-type *WAS* and *APPBP2* interact, all three polymorphisms disrupt the interaction. On the other hand, *WAS* uses the PBD domain for interacting with *CDC42* and regulating its auto-inhibition (*27*). Our Y2H results confirm that only the I294T mutation within the PBD domain disrupts the interaction, illustrating the specificity of our assay in detecting particular disruptions. Moreover, both wild-type *WAS* and all three variants interact with *WIPF1* and *NCK2*. Our results show that all the three mutations, including the rare SNP, have functional consequences as they disrupt specific interactions. To further explore such consequences, we examined the effect of these variants on the *WAS*-*TRIP10* and *WAS*-*ABI3* interactions. R41G disrupts both these interactions but I294T and E131K do not disrupt them. Our results suggest that the WH1 domain (and not the PBD domain) forms the interface for these two interactions, since previous analyses have shown that mutations at the interface are most likely to disrupt specific interactions (*26*). Moreover, while both R41G and E131K are within the WH1 domain, only the former disrupts the interaction. This can be explained by examining the severity of the functional consequences of these mutations. R41G has not been detected as a polymorphism in healthy individuals suggesting that it is a highly deleterious mutation occurring at extremely low allele frequencies within the population. On the other hand, E131K has been detected as a rare polymorphism in healthy individuals. This suggests that the functional consequences of this variant are less deleterious than that of R41G. Our analysis shows that rare variants have some functional consequences and disease mutations tend to be a more extreme case with more deleterious effects.

Construction of mutant clones

The wild-type *WAS* entry clone is obtained from the hORFeome 3.1 collection (*28*). Mutant clones were generated using PCR mutagenesis as previously described (*29, 30*). Briefly, wild-type genes in AD or DB vectors were used as templates in PCR reactions to generate N- and C-terminal fragments both containing the desired mutation in their overlapping regions. BP recombination reactions were done as per the manufacturer's manual (Gateway BP Clonase II enzyme mix) to clone mutant clones into the entry vector (pDONR223). Wild-type and mutant WAS clones were also PCR cloned into the mammalian expression vector pcDNA3 (Invitrogen Life Technologies) using XbaI and NotI restriction sites. A flag-tag was introduced into the C-terminal end of genes.
Primers used:
WAS_ cloning_F_XbaI
GCTGTCTAGAGCCACCATGAGTGGGGGCCCAATGGG
WAS_cloning_FLAG NotI_R
ATCAGCGGCCGCCTACTTATCGTCGTCATCCTTGTAATCGTCATCCCATTCATCATC
TTC

Yeast two-hybrid

Yeast two-hybrid (Y2H) was done as previously described (*25*). *CDC42*, *TRIP10*, *ABI3*, *APPBP2*, *WIPF1* and *NCK2* were transferred into AD vectors using Gateway LR reactions. Wild-type/mutant WAS was transferred into a DB vector. AD and DB constructs were transformed into Y2H strains MATa Y8800 and MATα Y8930, respectively. Transformed yeast was spotted onto YPD plates and incubated at 30 °C for ~20 h before replica plating onto SC-Leu-Trp plates. These plates were incubated at 30 °C for 24 h, then replica plated onto each of the four plates (SC-Leu-Trp-His, SC-Leu-His+CYH, SC-Leu-Trp-Ade, SC-Leu-Ade+CYH). 3 days later plates were scored for protein interactions.

3. Details on 'Relationship of functional elements with indels and larger SVs'

SVs of single nucleotide resolution are combined from the 1000 Genomes pilot data (*13*) and the phase I integrated call set. To removed redundancies, we take 50% reciprocal overlaps between SVs and preferentially keep the phase I SVs. This results in a dataset of 15,790 SVs of single nucleotide resolution. SV formation mechanisms are classified using the BreakSeq tool (*31*). The randomization test is performed as previously described (*32*). Gene and gene elements are taken from the longest transcript of protein-coding genes in Gencode v7 annotations. Whole gene intersection includes SV overlapping with one whole gene, as well as multiple whole genes. Partial gene intersection involves partial overlap of SVs with any gene.

Aggregation of histone marks around SV breakpoints
Yong Kong

4. Details on ‘Functional implications of positive selection amongst human populations’

Enrichment of sites with high population differentiation in categories of functional elements and genes

Sites presenting extreme population differences in the frequency of the derived allele were identified as described, and 604 were experimentally tested using Sequenom assays and concordance with Complete Genomics data revealing an average per-locus genotype concordance rate of 95% (*6 human genomes*)). We identified a control set of 1 million sites matched for allele frequency in the combined sample and calculated the ratio between the occurrence of functional annotations in the two data sets for the 27 functional categories where we would have a large enough sample size to detect enrichment (>100,000 SNPs). We performed a Fisher exact test for each category and applied two different corrections to take account of the 27 tests used: the Benjamini-Hochberg (BH) procedure and the Bonferroni correction. For the latter, we considered as significant p-values <0.0018 (0.05/27). This correction for multiple testing is conservative since the categories are not all independent. Since the definition of HighD sites requires setting a threshold for derived allele frequency difference (ΔDAF; threshold used = 0.7), we also investigated the consequences of varying this threshold between 0.5 and 0.8 and we found that the observed enrichments are stable across a broad range of ΔDAF thresholds [Supplementary figure XX].
In evaluating the prevalence of HighD sites among genes categories, we selected from Ensembl release 68 a set of control genes matched for number, GC content, gene length and recombination rate (from Hapmap Phase2).

We note that despite observing enrichment in some TF peaks, we do not observe enrichment in TF motifs, which might be expected to have stronger functional impact than TF peaks. This might be due to a limitation of the method used to identify HighD sites, since it only picks one SNP (with the highest ΔDAF value) in each cluster of highly differentiated SNPs. It is possible that the mostly highly differentiated SNP does not lie in a motif but it is in LD with another SNP which has a stronger impact on TF binding by its presence.

5. Details on ‘Natural germline vs disease variants’

Enrichment of known disease-causing mutations in sensitive regions

Disease-causing mutations in regulatory regions are obtained from HGMD database (*33*). After lifting over from hg18 to hg19, 566 mutations are found happened in non-coding regions using Gencode 7 annotations. Enrichments of these disease-causing mutations in “ultra-sensitive”, “sensitive”, “non-coding annotation” regions are compared to “non-coding” regions.

Source of cancer data sets

Cancer patient genome variant data was obtained from recently published whole genome sequence (WGS) cancer studies, including seven prostate cancer genomes (*34*), three medulloblastoma genomes (*35*), and 21 breast cancer genomes (*36*). Somatic variant calls for these cancers were obtained from the respective studies’ authors. Prostate germline variant calls were obtained using the Broad Institute’s Genome Analysis Toolkit (GATK) (*37*), and the medulloblastoma germline variant calls were obtained from the study authors. Germline variant calls were not available for the breast cancer data. For the purpose of uniform comparison and to remove artifacts arising from varied sequencing technologies and data processing pipelines, we have used germline variants from healthy tissues of matching tumour samples processed in a consistent manner whenever possible.

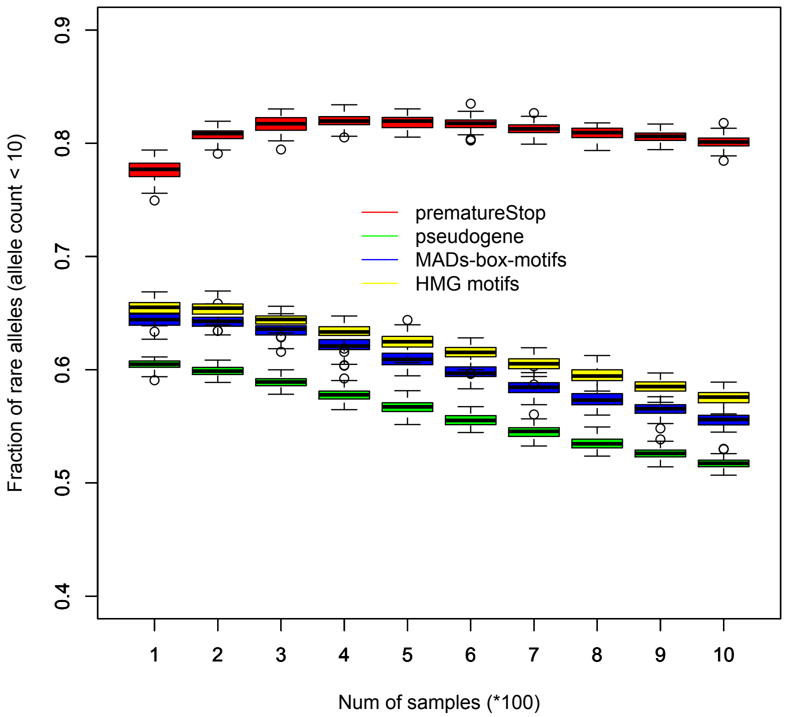
Enrichment/depletion of driver and passenger somatic SNVs with 1000 Genomes SNPs

Enrichment/depletion of cancer drivers and passengers was determined by comparing the observed intersection with an expected intersection computed by random simulation. These simulations involved randomizing the positions of the variants in each cancer dataset 10,000 times, producing 10,000 sets of random variants for each cancer. For each of these random datasets, an intersection analysis against 1000 Genomes variants was run. Hence, for each observed intersection, a distribution of intersections expected at random chance was created. These distributions were used to determine any significant enrichments or depletions of cancer variant intersections with the 1000 Genomes variants. We find that coding driver SNVs are significantly depleted for intersection with 1000 Genomes variants (p value= XX) while passenger SNVs are significantly enriched (p value= 1.19e-26).

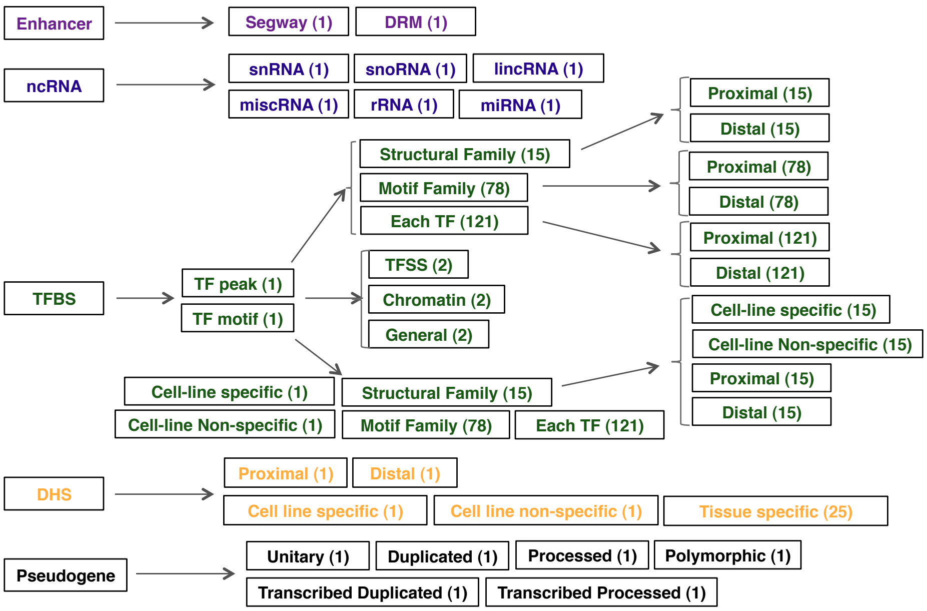
Annotation of SNV as breaking or conserving TF motif

A SNP that breaks a motif is defined as a mutation that decreases the motif-matching score of the TF-binding site to the PWM of the motif. Conversely, a SNP that conserves a motif is defined as a mutation that increases the motif-matching score of the TF-binding site to the PWM of the motif. Only polarized mutations, i.e. those whose ancestral states are known, are used for this analysis. For germline samples and deep-sequenced NA12878, only those variants whose ancestral states are determined in the 1000 Genomes Phase I data are used.

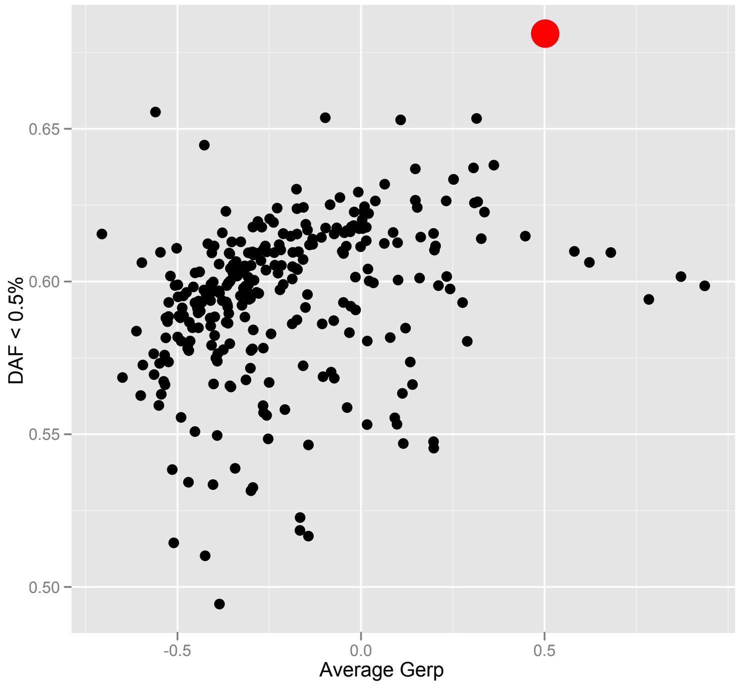
Supporting Figures



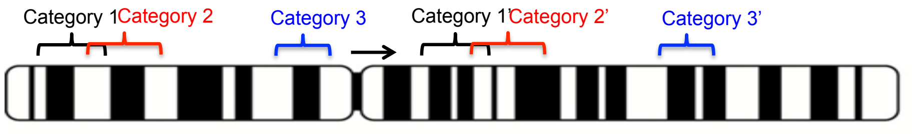
Supp Figure S 1 Impact of sample size on identification of differential purifying selection in various functional categories. As sample size increases, the fraction of SNPs with low allele count in categories under weak purifying selection (for example, pseudogenes) decreases. In strong contrast, the fraction of rare SNPs in categories under strong selection (for example, SNPs introducing prematureStop codons) remains relatively constant with increasing sample size.



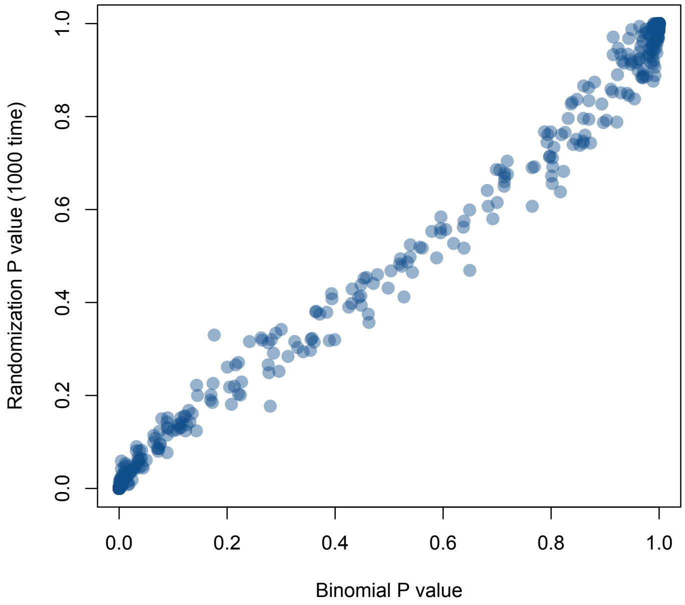
Supp Figure S 2 Broad and high-resolution categories. The numbers of sub-categories within each category are shown in brackets.



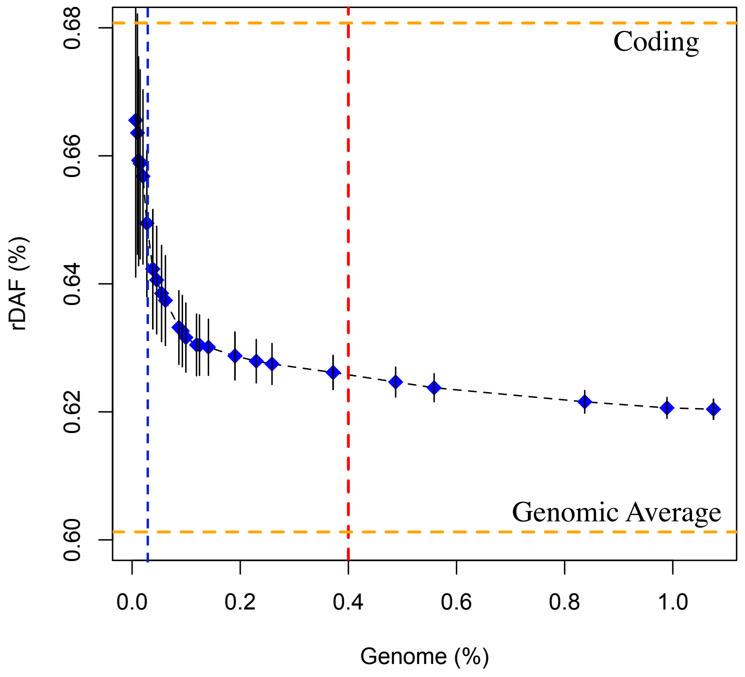
Supp Figure S 3 Relationship between fractions of rare SNPs and average Gerp scores of non-coding categories (rho= 0.49, p=3e-4; spearman rank test). Average Gerp score is calculated as the mean Gerp score of all bases underlying each category. Red dot is the coding region.



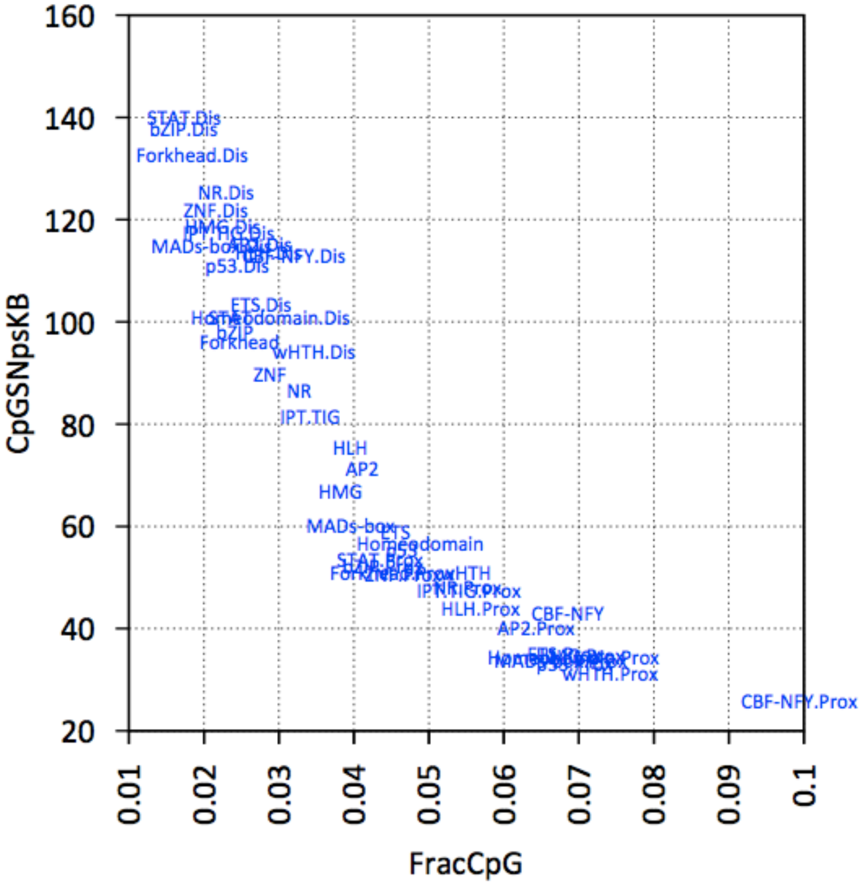
Supp Figure S 4 Schematic randomization procedures. Null distribution is obtained by sliding category coordinates along the genome for 1,000 times.



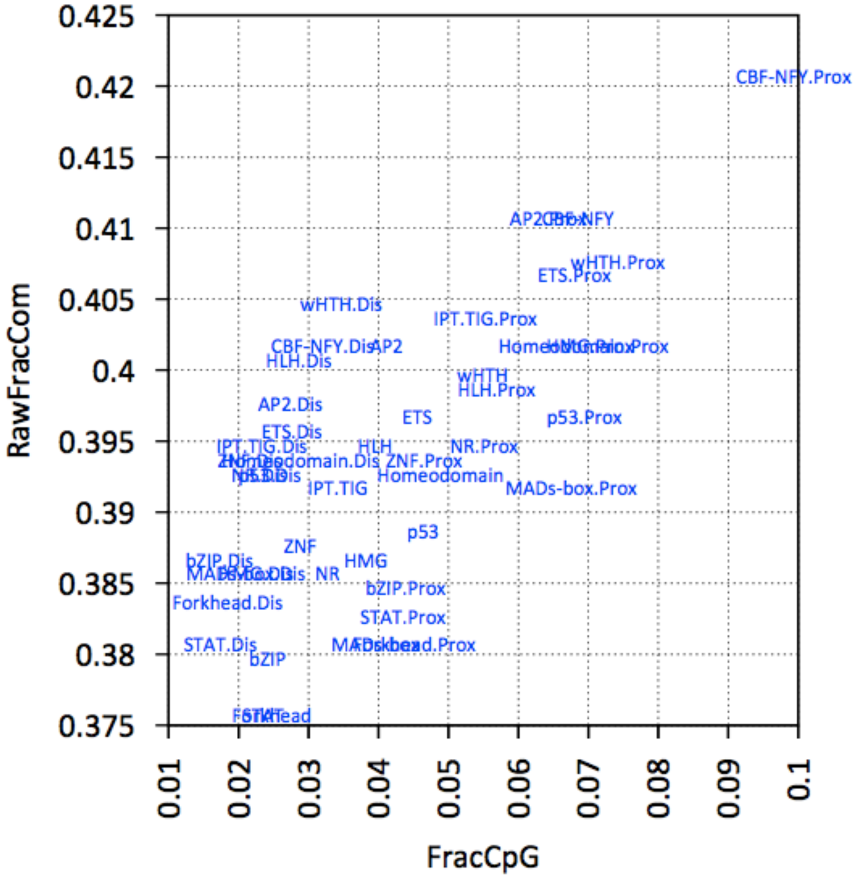
Supp Figure S 5 Comparison of randomized P values with binomial P values of fraction of rare SNPs for 677 categories. Binomial P value is calculated using fraction of rare variants of non-coding average as background.



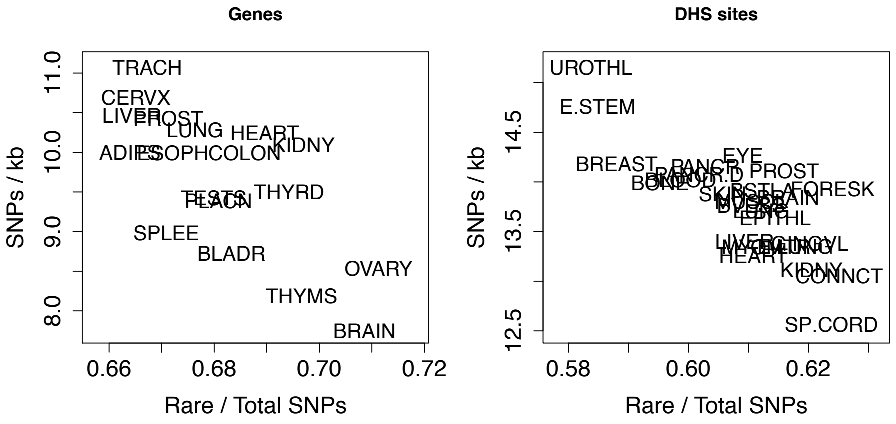
Supp Figure S 6 Changes of "Genomic percentage" and "Fraction of rare SNPs (rDAF)" by sequentially adding significant categories into examined regions. Significant categories are decreasingly ordered according to fraction of rare SNPs. Black lines are 95% confidence intervals. Orange lines denote rDAFs of "Coding" and "Genomic average". Blue dotted line is the ultra-sensitive cut-off; red dotted line is the sensitive cut-off.



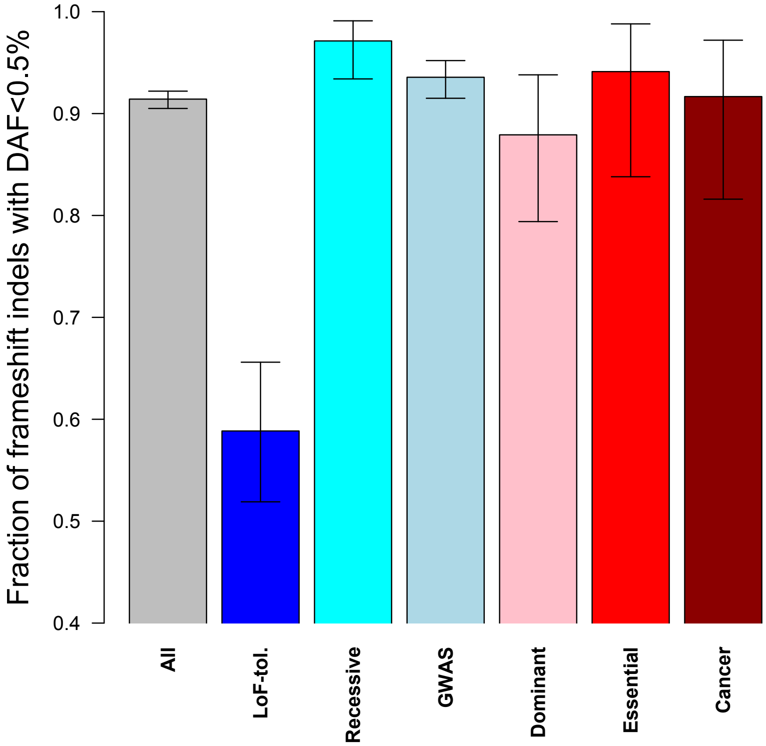
Supp Figure S 7 Relationship of ‘SNP density’ and ‘CpG’ content for various functional categories



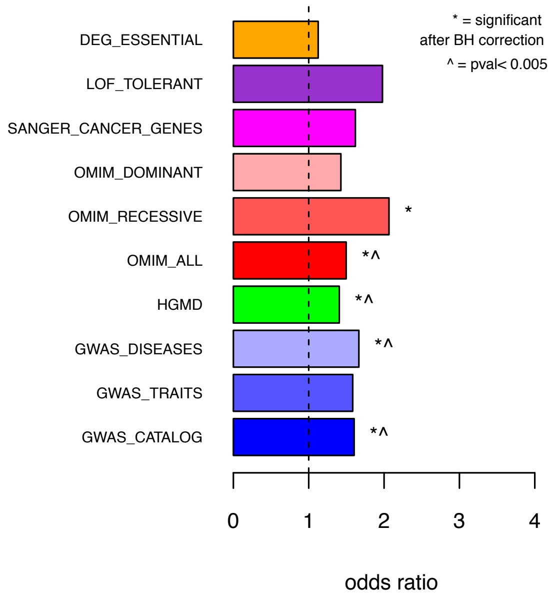
Supp Figure S 8 Relationship of ‘Fraction of common SNPs’ and ‘CpG content’ for various functional categories



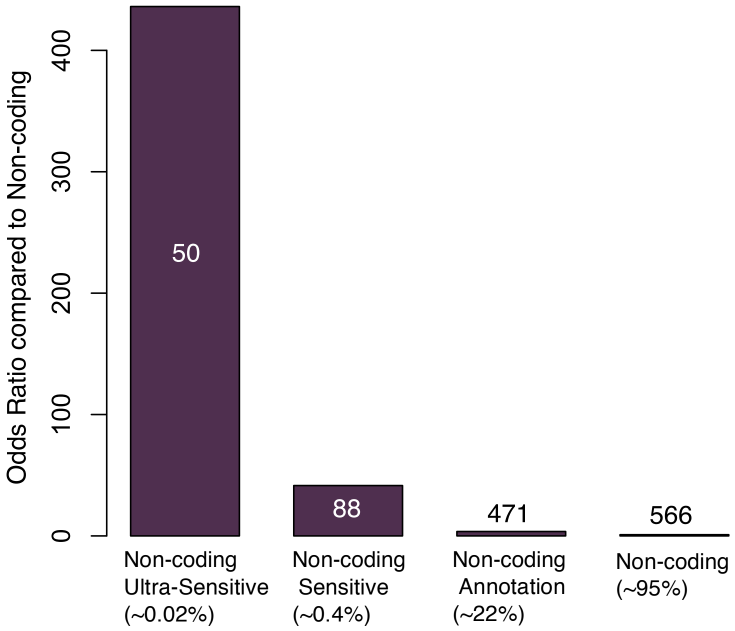
Supp Figure S 9 Left panel shows the relationship between 'SNP density' and 'Fraction of rare SNPs' for coding genes (rho=-0.6346749 , p value=0.005684). Right panel shows the relationship between SNP density and 'Fraction of rare SNPs' for DHSs (rho=-0.7361538, p value=4.383e-05).



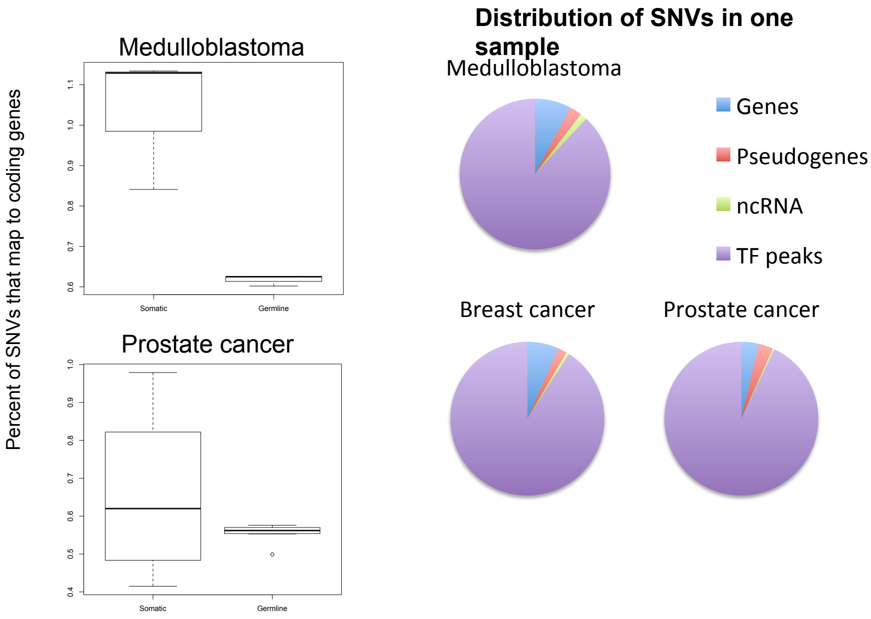
Supp Figure S 10 Fraction of rare frameshift indels for different gene categories



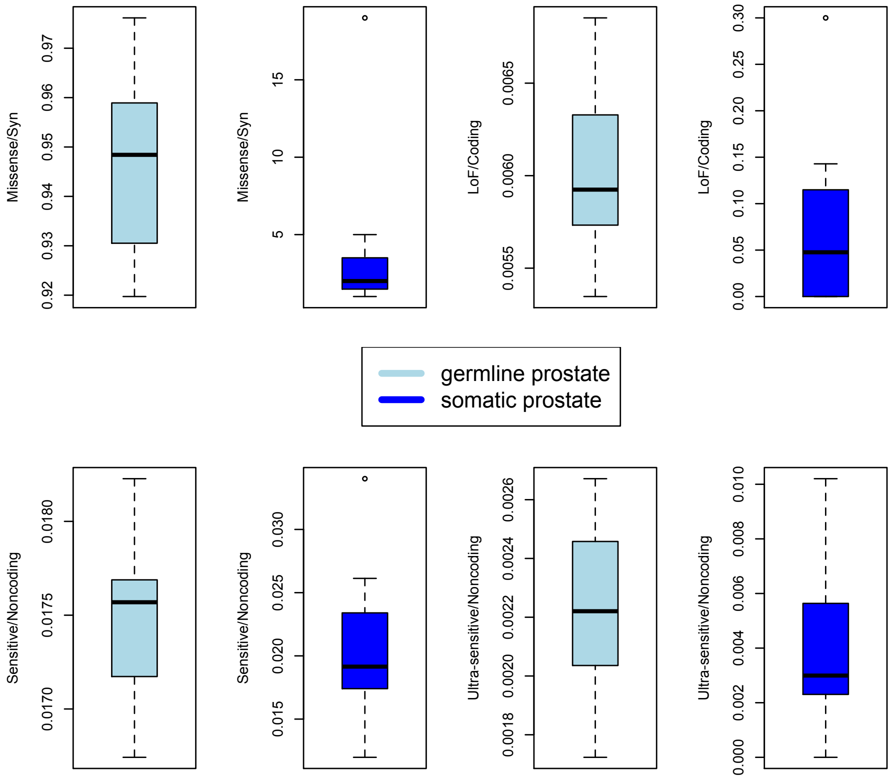
Supp Figure S 11



Supp Figure S 12 (HMGD) Enrichment of HGMD regulatory disease-causing mutations in non-coding ultra-sensitive, sensitive and annotated regions compared to whole non-coding regions.



Supp Figure S 13 Left panel shows that a higher percentage of somatic SNVs map to coding genes than germline SNVs for both medulloblastoma and prostate cancer. Right panel shows the distribution of somatic SNVs in different functional categories for one sample from different cancers.



Supp Figure S 14 Distributions of per sample ratios for somatic and matching germline SNVs in various functional categories across seven prostate cancer samples.

Supplementary Tables

Tabel S1. Tissue-specific expressed genes: Fraction of rare SNPs, Mann-Whitney p values for comparison with all expressed genes and direction of change of their median expression from all genes.

Tissue	Median rare/total SNPs	P-value	Median vs all
Adipose	0.639	0.167	lower
Bladder	0.667	0.267	lower
Brain	**0.668**	**5.08E-06**	**higher**
Cervix	0.625	0.442	lower
Colon	0.667	0.740	lower
Esophagus	0.646	0.739	lower
Heart	0.658	0.287	lower
Kidney	0.678	0.556	higher
Liver	0.644	0.058	lower
Lung	**0.643**	**0.031**	**lower**
Ovary	**0.680**	**0.006**	**higher**
Placenta	0.654	0.475	lower
Prostate	0.604	0.824	higher
Spleen	**0.642**	**0.002**	**lower**
Testes	**0.658**	**0.014**	**lower**
Thymus	0.658	0.121	higher
Thyroid	0.667	0.086	higher
Trachea	0.648	0.054	lower

Tabel S2. Fisher's exact test odds ratio (somatic:germline) and p values for enrichment of functionally deleterious mutations amongst somatic variants. Significant p values (<0.05) are colored in grey.

Sample	Missense/Syn	LoF/Coding	Sensitive/NonCoding	Ultrasensitive/Noncoding
PR-0508	2.11 (4.415e-1)	24.85 (4.521e-2)	1.94 (4.541e-2)	4.26 (3.521e-2)
PR-0581	1.05 (1)	0 (1)	0.93 (1)	1.65 (4.563e-1)
PR-1701	2.16 (2.523e-1)	0 (1)	1.18 (5.618e-1)	1.16 (5.772e-1)
PR-1783	20.66 (1.143e-5)	8.32 (1.189e-1)	1.07 (7.285e-1)	1.08 (6.053e-1)
PR-2832	1.43 (7.182e-1)	48.81 (6.522e-5)	1.56 (1.314e-1)	4.14 (3.779e-2)
PR-3027	1.68 (2.797e-1)	14.67 (9.7755e-3)	1.10 (7.221e-1)	0 (6.318e-1)
PR-3043	5.12 (2.102e-2)	0 (1)	0.66 (5.376e-1)	1.12 (5.91e-1)

Tabel S3. Number of SNVs conserving vs. breaking TF-binding motifs

Sample	# SNPs	# SNPs breaking motifs	# SNPs conserving motifs	Ratio of conserving: breaking
Deeply-sequenced NA12878 (Polarized SNPs in 1000 Genomes Phase I)	2562766	4410	2985	0.677
Average 1000 Genomes Phase I (Polarized SNPs)	3607334	6244	4030	0.647

Prostate cancer sample		# SNPs	# SNPs breaking motifs	# SNPs conserving motifs	Ratio of conserving: breaking
Somatic	PR-0508	1446	4	1	0.250
	PR-0581	1430	1	0	0.000
	PR-1701	1936	1	1	1.000
	PR-1783	2226	2	1	0.500
	PR-2832	1829	3	1	0.333
	PR-3027	2452	4	2	0.500
	PR-3043	1713	4	0	0.000
	Average	1862	3	1	0.333
Germline (Polarized SNPs in 1000 Genomes Phase I)	PR-0508	2697721	4004	2799	0.699
	PR-0581	2699101	3948	2860	0.724
	PR-1701	2759025	4041	2830	0.700
	PR-1783	2758799	3893	2784	0.715
	PR-2832	2757884	4024	2851	0.708
	PR-3027	2748298	3774	2706	0.717
	PR-3043	2756406	3465	2536	0.732
	Average	2739605	3878	2767	0.714

Medulloblastoma sample		# SNPs	# SNPs breaking motifs	# SNPs conserving motifs	Ratio of conserving: breaking
Somatic	MB1	2250	3	3	1.000
	MB2	1786	6	2	0.333
	MB4	1607	5	0	0.000
	Average	1881	5	2	0.444
Germline (Polarized SNPs in 1000 Genomes Phase I)	MB1	2753485	3893	2799	0.719
	MB2	2731595	4188	2,55	0.706
	MB4	2758545	4037	2871	0.711
	Average	2747875	4039	2875	0.712

Breast cancer somatic sample	# SNPs	# SNPs breaking motifs	# SNPs conserving motifs	Ratio of conserving: breaking
PD3851a	1782	5	1	0.200
PD3890a	6124	11	5	0.455
PD3904a	5608	7	2	0.286
PD3905a	4587	9	1	0.111
PD3945a	10308	21	6	0.286
PD4005a	6104	18	3	0.167
PD4006a	9194	25	2	0.080
PD4085a	2673	2	0	0.000
PD4086a	2199	5	0	0.000
PD4088a	1705	7	0	0.000
PD4103a	5360	6	3	0.500
PD4107a	10291	20	4	0.200
PD4109a	9888	23	3	0.130
PD4115a	9954	23	9	0.391
PD4116a	8026	12	6	0.500
PD4120a	70690	275	60	0.218
PD4192a	3919	7	3	0.429
PD4194a	1484	3	1	0.333
PD4198a	4552	7	1	0.143
PD4199a	6932	20	3	0.150
PD4248a	2536	5	0	0.000
Average	8758	24	5	0.222

Tabel S4. Ultra-sensitive and sensitive coordinates and functional annotations.

References:

1. J. Harrow *et al.*, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760 (Sep, 2012).
2. I. Dunham *et al.*, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (Sep, 2012).
3. J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, N. M. Luscombe, A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252 (Apr, 2009).
4. M. B. Gerstein *et al.*, Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91 (Sep, 2012).
5. K. Y. Yip *et al.*, Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48 (2012).
6. G. R. Abecasis *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (Nov, 2012).
7. R. Blekhman *et al.*, Natural selection on genes that underlie human disease susceptibility. *Curr Biol* **18**, 883 (Jun, 2008).
8. L. A. Hindorff *et al.*, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362 (Jun, 2009).
9. R. Zhang, H. Y. Ou, C. T. Zhang, DEG: a database of essential genes. *Nucleic Acids Res* **32**, D271 (Jan, 2004).
10. R. Zhang, Y. Lin, DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res* **37**, D455 (Jan, 2009).
11. B. Y. Liao, J. Zhang, Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**, 6987 (May, 2008).
12. P. A. Futreal *et al.*, A census of human cancer genes. *Nat Rev Cancer* **4**, 177 (Mar, 2004).
13. G. P. Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (Oct, 2010).
14. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440 (Aug, 2003).
15. J. Rozowsky *et al.*, AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).
16. S. B. Montgomery, T. Lappalainen, M. Gutierrez-Arcelus, E. T. Dermitzakis, Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* **7**, e1002144 (Jul, 2011).
17. M. B. Gerstein *et al.*, Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775 (Dec, 2010).

18. R. E. Thurman *et al.*, The accessible chromatin landscape of the human genome. *Nature* **489**, 75 (Sep, 2012).
19. K. Lage *et al.*, A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25**, 309 (Mar, 2007).
20. J. Das, H. Yu, HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* **6**, 92 (2012).
21. P. Pagel *et al.*, The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832 (Mar, 2005).
22. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic Acids Res* **28**, 235 (Jan, 2000).
23. S. J. Hubbard, J. M. Thornton. (1993).
24. E. A. Franzosa, Y. Xia, Structural principles within the human-virus protein-protein interaction network. *Proc Natl Acad Sci U S A* **108**, 10538 (Jun, 2011).
25. H. Yu *et al.*, High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104 (Oct, 2008).
26. X. Wang *et al.*, Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol* **30**, 159 (Feb, 2012).
27. A. S. Kim, L. T. Kakalis, N. Abdul-Manan, G. A. Liu, M. K. Rosen, Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein. *Nature* **404**, 151 (Mar, 2000).
28. P. Lamesch *et al.*, hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**, 307 (Mar, 2007).
29. Q. Zhong *et al.*, Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* **5**, 321 (2009).
30. Y. Suzuki *et al.*, A novel high-throughput (HTP) cloning strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. *Nucleic Acids Res* **33**, e109 (2005).
31. H. Y. Lam *et al.*, Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28**, 47 (Jan, 2010).
32. X. J. Mu, Z. J. Lu, Y. Kong, H. Y. Lam, M. B. Gerstein, Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* **39**, 7058 (Sep, 2011).
33. P. D. Stenson *et al.*, The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13 (2009).
34. M. F. Berger *et al.*, The genomic complexity of primary human prostate cancer. *Nature* **470**, 214 (Feb, 2011).
35. T. Rausch *et al.*, Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59 (Jan, 2012).
36. S. Nik-Zainal *et al.*, Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979 (May, 2012).
37. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491 (May, 2011).