Human Genomics

PRIMARY RESEARCH Open Access

Architecture of polymorphisms in the human genome reveals functionally important and positively selected variants in immune response and drug transporter genes

Yu Jin[1,3], Jingbo Wang[2], Maulana Bachtiar[2,3], Samuel S. Chong[4] and Caroline G. L. Lee[1,2,3,5*]

Abstract

Background: Genetic polymorphisms can contribute to phenotypic differences amongst individuals, including disease risk and drug response. Characterization of genetic polymorphisms that modulate gene expression and/or protein function may facilitate the identification of the causal variants. Here, we present the architecture of genetic polymorphisms in the human genome focusing on those predicted to be potentially functional/under natural selection and the pathways that they reside.

Results: In the human genome, polymorphisms that directly affect protein sequences and potentially affect function are the most constrained variants with the lowest single-nucleotide variant (SNV) density, least population differentiation and most significant enrichment of rare alleles. SNVs which potentially alter various regulatory sites, e.g. splicing regulatory elements, are also generally under negative selection.

Interestingly, genes that regulate the expression of transcription/splicing factors and histones are conserved as a higher proportion of these genes is non-polymorphic, contain ultra-conserved elements (UCEs) and/or has no non-synonymous SNVs (nsSNVs)/coding INDELs. On the other hand, major histocompatibility complex (*MHC*) genes are the most polymorphic with SNVs potentially affecting the binding of transcription/splicing factors and microRNAs (miRNA) exhibiting recent positive selection (RPS). The drug transporter genes carry the most number of potentially deleterious nsSNVs and exhibit signatures of RPS and/or population differentiation. These observations suggest that genes that interact with the environment are highly polymorphic and targeted by RPS.

Conclusions: In conclusion, selective constraints are observed in coding regions, master regulator genes, and potentially functional SNVs. In contrast, genes that modulate response to the environment are highly polymorphic and under positive selection.

Keywords: Single-nucleotide variant, Natural selection, Potentially functional SNV, Immune response genes, Drug transporters

* Correspondence: bchleec@nus.edu.sg
[1]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 117456, Singapore
[2]Department of Biochemistry, National University of Singapore, Singapore 119077, Singapore
Full list of author information is available at the end of the article

Background

Genetic polymorphisms may contribute to the differences in disease risks and drug responses amongst different individuals. Different forms of genetic variants are found in the human genome. Single-nucleotide variants (SNVs) account for more than 90% of genomic variants and are the major form of genetic polymorphisms [1].

Some polymorphisms can affect phenotype. These polymorphisms are likely to alter gene expression or protein function leading to modulation of cellular function and influencing disease risk or drug response. However, to identify the single or a group of causal variants for a particular phenotype from a pool of more than 100 million polymorphisms is like 'finding a needle in a haystack' and remains a great challenge since not all genetic variants are functionally important.

While non-synonymous SNVs (nsSNVs) have been extensively investigated as they are the most likely to modulate phenotypes via changing the amino acid composition of proteins, synonymous SNVs (sSNVs) and non-coding variants can also account for phenotypic differences since these variants can affect mRNA stability and transcriptional or translational efficiency and have been associated with gene expression levels in various cell lines and tissues [2–10]. While it may not be feasible to experimentally test every single polymorphism for its function, a variety of bioinformatics tools is now available. These tools can reasonably predict the potential functions of genetic variants, including the likelihood of nsSNVs to disrupt protein structures and/or functions [11–19], SNVs that potentially modify splicing [20, 21] or transcription [22], and SNVs in 3′ untranslated regions (3′UTRs) with potential to alter miRNA target sites [23–25]. There are also comprehensive web tools for predicting various potential functions of both regulatory and coding SNVs, e.g. pfSNP [26] and PupaSNP finder [27]. They can facilitate our understanding of how polymorphisms can lead to phenotype change and help us prioritize the potentially functional SNVs (pfSNVs) for further investigation.

In addition to the above-mentioned predictive bioinformatics tools, signatures of natural selections can also facilitate the identification of causal variants since variants under natural selection are likely to be functionally significant. Patterns of population differentiation were employed to identify 174 candidate gene loci showing signatures of purifying or positive selection [28]. 'Long-range haplotype' methods have been employed to identify a list of targets under recent positive selection (RPS) [29]. Another study utilizing HapMap Phase II data found that negative selection preferentially targets non-synonymous sites, while both non-synonymous and 5′ untranslated regions (5′UTRs) show an excess of highly differentiated SNVs, suggesting the evidence of positive selection as well. The authors also reported that variants under selective pressures (either positive or negative) occur more frequently in disease-related genes and are more likely to contribute to disease phenotypes [30].

Although previous reports examined the association of SNVs in regulatory regions with natural selection, these studies were limited. They either merely focussed on only one class of regulatory SNVs (e.g. SNVs within miRBS) [31], on SNVs residing in non-coding regions [32] or within regulatory elements [33] without predicting whether these SNVs alter function (e.g. if a SNV will abolish or create a regulatory site).

In this study, we present the architecture of all genetic polymorphisms of the human genome, focusing on SNVs that are potentially functional and/or positively selected and the pathways that they reside.

Results

Polymorphisms are most constrained in coding regions

Of the > 14 million polymorphisms in the human genome validated in the dbSNV database (Build 131), 38% of the polymorphisms are within the protein-coding genes while 62% resides in the intergenic regions. More than 95% of the variants within human genes reside within introns (Fig. 1a). Coding polymorphisms constitute ~ 3% of the total polymorphisms within genes, of which 2.55% are SNVs while 0.35% are short insertion/deletions (INDELs) (Fig. 1a). Upon normalization against the length of each genic region, coding regions contain the lowest average densities of both SNVs and INDELs (Fig. 1b). Notably, frame-shift INDELs (i.e. length of INDELs is not in multiples of three) are significantly under-represented in the coding regions compared to non-coding regions in the human genes (p value < 0.001 by Fisher's exact test, Fig. 1c). These data suggest that both SNVs and INDELs are selectively constrained within coding sequences, especially the INDELs with potential to cause frame-shift.

To further investigate the regions within genes that may be most subjected to negative selection pressure, the derived allele frequencies (DAFs) of SNVs in different regions are further compared using allele frequency data of the International HapMap Project individuals. As evident in Fig. 1d, coding regions (red) contain a higher percentage of rare SNVs, defined as having DAF < 0.05 in all the three population groups, namely African, East Asian and European populations. nsSNVs (brown) within the coding region are also enriched with rare alleles compared to sSNVs (orange) (Fig. 1e). As negative selection increases the fraction of rare alleles [34], our results from the analysis of allele frequency data again suggest that coding SNVs, especially nsSNVs, tend to be targeted by negative selection.

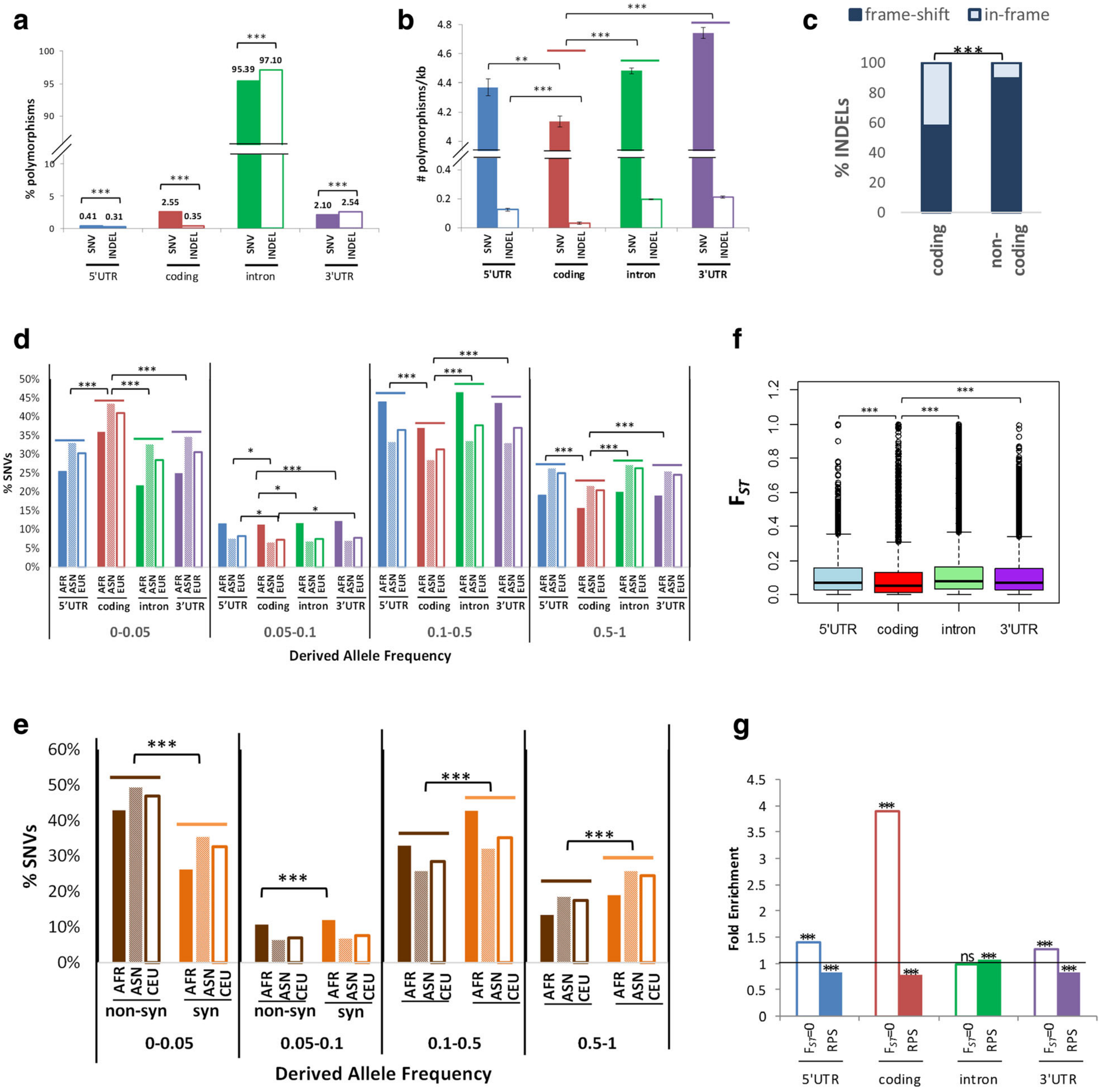


Fig. 1 Architecture of polymorphisms in the human genome. **a** Percentage of polymorphisms in different regions (5′UTRs, coding regions, introns, 3′UTRs) of the human genes. **b** Average SNV and INDEL densities (# polymorphisms/kb) in the different regions (5′UTR, coding region, intron and 3′UTR) of a gene in the human genome. Error bars represent the standard errors of the mean SNV and INDEL densities. **c** Percentage of frame-shift and in-frame INDELs in coding and non-coding regions in human genes. Frame-shift INDELS are defined as INDELs whose lengths are not in multiples of three while in-frame INDELs are those whose lengths are in multiples of three. **d** Percentage of SNVs with different DAFs in the four genic regions, as measured in HapMap individuals from African, Asian and European populations. **e** Percentage of synonymous and non-synonymous variants with different DAFs, as measured in HapMap individuals from African, Asian and European population groups. **f** Distribution of F_{ST} statistics in four genic regions. SNVs in coding regions show significantly lower median F_{ST} compared to the other non-coding regions. **g** Fold enrichment of SNVs showing signatures of negative selection ($F_{ST} = 0$) (open bar) or RPS (shaded bar) in the genic regions. Fold enrichment is determined by the percentage of SNVs with $F_{ST} = 0$ or under RPS in a specific region (e.g. coding region) divided by the percentage of all genotyped SNVs in that region. Coding, coding region; non-syn, non-synonymous; syn, synonymous. AFR, African; ASN, Asian; EUR, European. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$; ns, not significant

Signatures of natural selection are also examined through determining population differentiation using the F_{ST} statistics [28] across the different population groups (African, East Asian and European) since high F_{ST} is associated with a positive selection [34], while low F_{ST} is associated with a negative selection [30]. As shown in

Fig. 1f, coding SNVs have lower median F_{ST} than SNVs in other regions including 5′UTRs, 3′UTRs and introns (Bonferroni corrected p values < 0.001 by Mann-Whitney test). In fact, zero-F_{ST} SNVs are significantly over-represented in coding exons (Bonferroni corrected p value < 0.001 by Fisher's exact test) (Fig. 1g, non-shaded bars). Patterns of RPS are examined using linkage disequilibrium (LD) and haplotype-based methods. As shown in Fig. 1g (shaded bars), exonic regions, i.e. 5′ UTRs, coding regions and 3′UTRs, are significantly less enriched with RPS SNVs (Bonferroni corrected p values < 0.001 by Fisher's exact test), while introns are more enriched with RPS SNVs (Bonferroni corrected p value < 0.001 by Fisher's exact test).

Taken together, coding regions are generally under strong negative selection pressures as they show the lowest densities of SNVs and INDELs (especially frame-shift INDELs), the highest proportion of rare alleles with less enrichment of RPS SNVs. Notably, coding SNVs are also the least population differentiated.

Potentially functional SNVs are under natural selections

The putative functions of SNVs in the various genic and promoter regions are predicted using a variety of bioinformatics algorithms (see Additional file 1: Supplementary Methods). Approximately four hundred thousand (7%) pfSNVs in genic and promoter regions can potentially modulate gene expression and/or function. More than 93% of genes in the human genome contain at least one pfSNV (Fig. 2a). Each gene is predicted to contain an average of seven promoter SNVs capable of altering transcription factor binding sites (TFBS); eight intronic and two coding SNVs that may modulate splicing regulatory elements, i.e. intronic splicing regulatory element (ISRE) and exon splicing enhancers or silencers (ESE/ESS); and one coding SNV that is potentially deleterious to protein function and one SNV in 3′UTR that may alter miRNA binding site(s) (miRBS) (Fig. 2b).

To evaluate if pfSNVs are selectively constrained, the proportions of rare alleles (DAF < 0.05) of pfSNVs and non-functional SNVs (nfSNVs) in a specified region are compared. As evident in Fig. 2c, except for pfSNVs predicted to alter TFBS, most pfSNVs are enriched with rare alleles. Notably, pfSNVs predicted to be deleterious to protein function are more than 1.5-fold more enriched with rare alleles compared to nfSNVs in coding regions indicating that these pfSNVs may be under the strongest negative selection pressure.

Conversely, pfSNVs predicted to be deleterious to protein function are found to be the least significantly enriched with RPS SNVs (Fig. 2d, brown) (Bonferroni corrected p value < 0.001 by Fisher's exact test) consistent with the earlier observation indicating that these pfSNVs are under the strongest negative selection. The other pfSNVs are not under any significant RPS except for pfSNVs predicted to alter ISRE (Bonferroni corrected p value = 0.019 by Fisher's exact test) (Fig. 2d). In addition, more than half of the RPS pfSNVs are predicted to affect ISRE (~ 53%) followed by TFBS (~ 30%) while least RPS pfSNVs (3%) are predicted to be deleterious to protein function (Fig. 2e).

Highly polymorphic vs conserved genes in the human genome

Amongst > 20,000 genes in the human genome, beta haemoglobin (*HBB*) gene is the most polymorphic gene, containing approximately 176 SNVs per kilobase (kb) with the highest density of SNVs within its coding region (Fig. 3a, red) (570 SNVs/kb). Several other haemoglobin genes (in green boxes) are also amongst the most polymorphic genes in the human genome with the majority of their SNVs residing within coding exons (red). Other highly polymorphic genes include the *MHC* family of genes (blue box) with most of their SNVs residing within introns (Fig. 3a, green) as well as the olfactory receptor (*OR*) gene family (orange box) where all the SNVs are also found within the coding region (Fig 3a, red).

The density of SNVs within each gene, normalized against their length, is determined for all > 20,000 protein-coding genes in the human genome. Most genes have approximately four SNVs per kilobase. Although ~ 97% of genes carry at least one SNV, 149 genes do not contain any polymorphism as SNV or INDEL. More than half of these 149 non-polymorphic genes were yet to be annotated. Nonetheless, the annotated non-polymorphic genes are significantly over-represented in histone 2A and 2B families and involved in nucleosome assembly (Fig. 3b, grey shaded; Additional file 1: Table S1).

We then focus on polymorphisms within the coding region since this region encodes the functional protein. Nearly 20% (4389) of genes in the human genome are found to be functionally conserved with no nsSNVs nor coding INDELs. These genes are enriched in various categories including GTPases and translational elongation (Fig. 3b, unshaded; Additional file 1: Table S2). Seventy genes are found to carry ultra-conserved elements (UCEs) [35] in their coding regions; hence, these genes are evolutionarily conserved. These ultra-conserved genes include the homeobox proteins and are primarily involved in the transcription factor activity, RNA splicing and pattern specification (Fig. 3b, shaded black; Additional file 1: Table S3).

Taken together, genes involved in the basic fundamental biological process, for example, gene regulation, are highly conserved during evolution, being least polymorphic within the human species as well as between species.

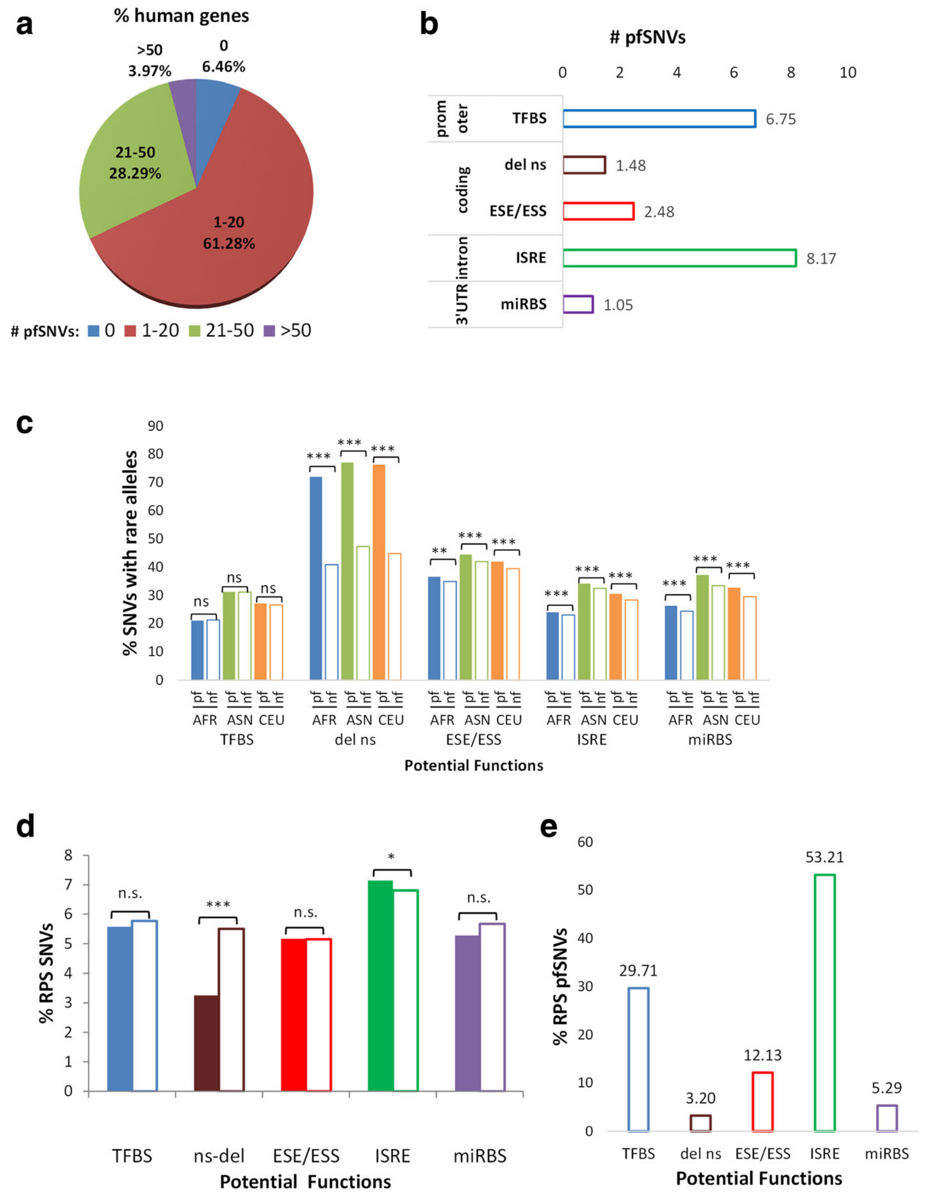


Fig. 2 Potentially functional SNVs. **a** Percentage of human genes containing a different number of pfSNVs. **b** Average numbers of pfSNVs with different potential functions in each transcript. **c** Percentage of SNVs with rare alleles (DAF < 0.05) amongst the pfSNVs and nfSNVs in the same genic regions, as determined in HapMap individuals from African (AFR), Asian (ASN) and European (EUR) population groups. **d** Percentage of pfSNVs or nfSNVs under RPS over all pf- or nfSNVs in that specific group. **e** Percentage of SNVs with different functions amongst all the RPS pfSNVs. pf, potentially functional; nf, non-functional; TFBS, SNVs that alter transcription factor binding sites; del ns, potentially deleterious nsSNV; ESE/ESS, SNVs that alter exon splice enhancers/silencers; ISRE, SNVs that alter intronic splicing regulatory elements; miRBS, SNVs that alter miRNA binding sites

Highly polymorphic genes are mainly involved in immune responses

A total of 512 highly polymorphic genes (see Additional file 1: Supplementary Methods) are identified. These genes are the most significantly over-represented in immune response pathways, as well as in the pathogenesis of a number of autoimmune diseases, including Graft-versus-host disease and type I diabetes mellitus (Fig. 3c, shaded grey; Additional file 1: Table S4). They are primarily in the *MHC* class I-related and class II-related protein families, involved in antigen presentation and processing. Majority of the *MHC* class I and class II genes are located on chromosome 6q21.3 which is the most polymorphic region in the human genome and facilitate the generation of diverse antigens to confer a selective advantage to fight infection [36]. The number of pfSNVs in the *MHC* genes (2234) is higher than the average number of pfSNVs in the

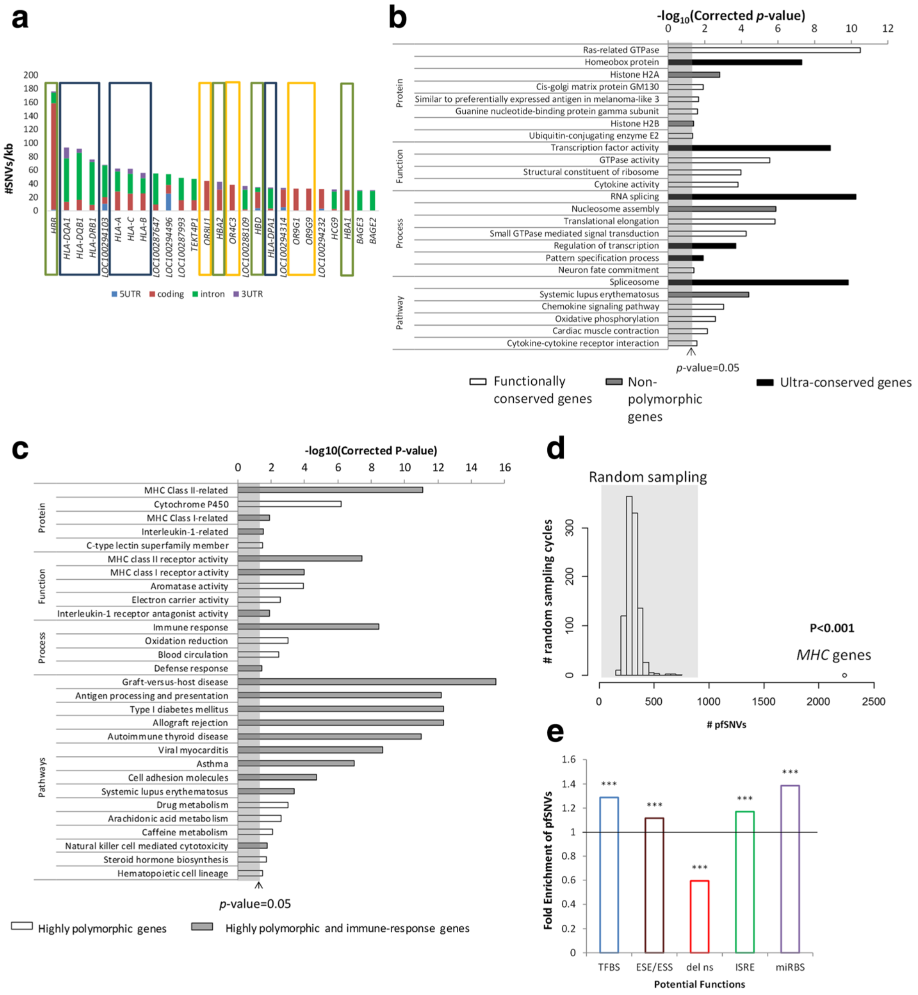


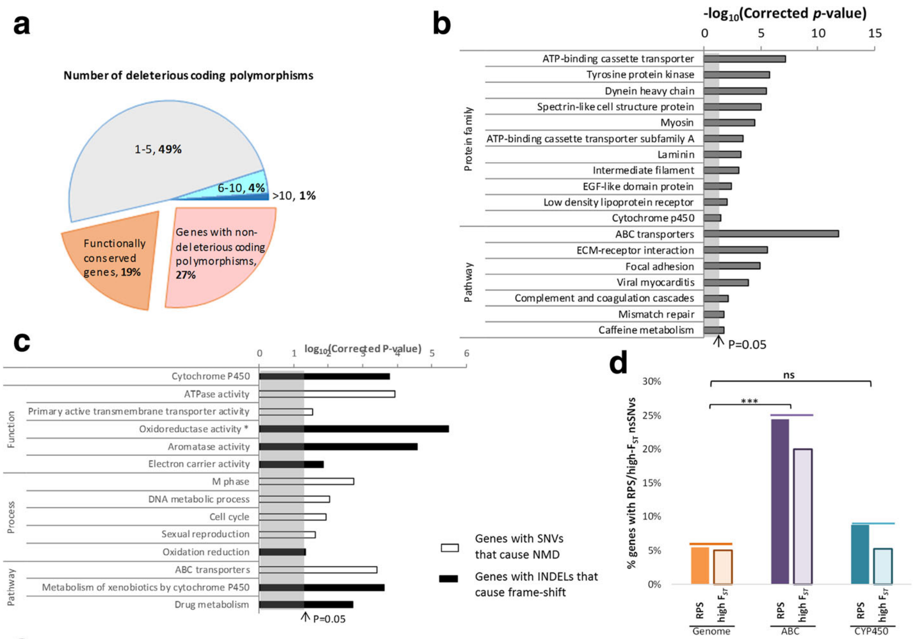
Fig. 3 Most polymorphic vs conserved genes in the human genome. **a** Genes having the highest number of SNVs normalized against gene length (> 30 SNVs/kb per gene). Green box: haemoglobin genes; blue box: *MHC* genes; orange box: *OR* genes. **b** Benjamini-corrected *p* values for the significantly enriched functional terms for non-polymorphic genes (grey bars), functionally conserved genes (white bars) which are genes without nsSNVs and coding INDELs as well as ultra-conserved genes (black bars) which are the genes with UCEs within their coding regions. **c** Benjamini-corrected *p* values for the functional terms that are significantly enriched by the highly polymorphic genes. Grey bars: the functional terms related to immune responses. **d** Empirical distribution of the numbers of pfSNVs obtained from 23 genes that are randomly sampled from all the human genes with lengths 3–15 kb for 1000 times. The number of pfSNVs in the 23 *MHC* genes is significantly higher than that in the randomly sampled gene sets (empirical *p* value < 0.001). **e** Enrichment of SNVs with different potential functions in *MHC* class I and class II genes. Fold enrichment is calculated as the percentage of pfSNVs in specific genic regions (e.g. coding SNVs that may alter ESE/ESS) in the *MHC* class I and class II genes against that for all the human genes. Deviation from one indicates that the pfSNVs are over- or under-represented in the *MHC* genes

other human genes (306). To determine if pfSNVs are significantly over-represented in this *MHC* gene family of 23 genes, sampling of 23 random human genes with similar gene length is performed 1000 times. The number of pfSNVs in the 23 random genes for each cycle is plotted to obtain an empirical distribution. The *MHC* family of genes, with 2234 pfSNVs, is found to carry significantly more pfSNVs than 1000 different sampling of 23 random human genes (empirical *p* value < 0.001 by random sampling test) (Fig. 3d). Interestingly, despite the enrichment of pfSNVs in the *MHC* gene family, there are significantly fewer nsSNVs predicted to be deleterious in the *MHC* family of genes, compared to other human genes (Bonferroni corrected *p* value < 0.001 by Fisher's exact test) (Fig. 3e). Hence, this family of proteins can nimbly respond to different infection, through a diversity of different regulatory mechanism, including differential transcription factor/miRNA binding/splicing.

Lastly, highly polymorphic genes are also significantly enriched in drug metabolism, cytochrome P450 (*CYP450*), arachidonic acid and caffeine metabolism pathways (Fig. 3c, Additional file 1: Table S4).

Drug response genes are most affected by potentially deleterious polymorphisms in coding regions

Although > 90% (20,890/22,333) of the genes in the human genome have pfSNVs, ~ 54% contain at least one potentially deleterious coding polymorphism (Fig. 4a, shaded blue, dark blue and grey) while ~ 19% are functionally conserved with no nsSNVs nor coding INDELS (Fig. 4a, shaded orange). Potentially deleterious coding polymorphisms are under the strongest negative selection as suggested earlier (Fig. 2c) since they can potentially have a drastic effect on protein function. Approximately 5% (1104/22,333) of all genes in the human



e

SNV	Gene	Amino acid change	RPS	F_{ST}	deleterious	alter ESE/ESS	Phenotype Association	Pubmed ID
rs2230806	ABCA1	Arg219Lys	√	0.19		√	Late-onset Alzheimer's disease Plasma LDLc levels HDLc levels & coronary artery disease	16725228 20170916 21310416
rs2230808		Lys1587Arg		**0.35**	√	√	Alzheimer's disease Schizophrenia LDLc levels	17510949 21839797 22668585
rs2472458		Asp831Asn		**0.39**	√	√		
rs3112831	ABCA4	His423Arg	√	0.01		√	Age-related macular degeneration Congenital Toxoplasmosis	21106043 18523590
rs6666652		Ser2255Ile		**0.38**		√		
rs557491	ABCA5	Met960Val	√	0.05		√		
rs4968839	ABCA6	Val282Ile	√	0.08				
rs7212506		Met875Ile		**0.31**		√		
rs11890512	ABCA12	Asn237His		**0.35**		√		
rs16853238		Trp199Cys	√	0.24	√	√		
rs9282564	ABCB1	Asn21Asp	√	0.05		√	Tacrolimus or cyclosporine levels after heart transplantation Colorectal cancer	21839244 21317201
rs16945916	ABCC11	His1344Arg, His1306Arg	√	0.2		√		
rs16945988		Arg19His	√	0.14		√		
rs17822931		Gly180Arg	√	**0.83**	√		Earwax type Axillary osmidrosis Apocrine colostrum secretion	16444273 19650936 17394018
rs16945874	ABCC12	Ala102Glu	√	0.2		√	Body mass index	20017965
rs7193955		Arg1117Cys		**0.63**		√		
rs8187692	ABCC2	Arg1181Leu	√	0.04	√			
rs11708427	ABCC5	Gln285His		**0.37**	√	√		
rs6720173		Gln604Glu	√	0.11		√	Serum lipid levels Plasma HDLc levels	22655090 19005228
rs4148211	ABCG8	Tyr54Cys		**0.52**		√	Serum lipid levels Plasma total cholesterol and LDLc Plasma HDLc levels	22655090 20170916 19005228
rs213950	CFTR	Val470Met	√	**0.38**		√	Lung cancer Papillary thyroid cancer	20116881 22740931

Fig. 4 Distribution of potentially deleterious coding polymorphisms in human genes. **a** Percentage of genes with different numbers of potentially deleterious coding polymorphisms in their coding regions. Genes without any potentially deleterious coding polymorphisms are divided into two groups: (1) functionally conserved genes, i.e. genes with no nsSNV nor INDELs in coding regions; (2) genes carrying non-deleterious SNVs in their coding regions. **b** Benjamini-corrected *p* values for the functional terms that show enrichment of the genes with more than five potentially deleterious nsSNVs. **c** Benjamini-corrected *p* values for the functional terms that show enrichment of the genes with SNVs that cause NMD (non-shaded) and the genes with coding INDELs that cause frame-shift (black). **d** Percentage of genes with RPS nsSNVs and genes carrying nsSNVs with high F_{ST} (> 0.3) in the whole genome, *ABC* transporter and *CYP450* family. **e** Recently positively selected and/or population-differentiated nsSNVs in the *ABC* transporters. F_{ST} scores in bold indicate $F_{ST} > 0.3$. *Oxidoreductase activity: oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen

genome are highly enriched with more than five potentially deleterious nsSNVs in their coding regions (Fig. 4a, shaded blue and dark blue). These genes are significantly enriched in the ATP-binding cassette (*ABC*) transporter and the *CYP450* families, which play important roles in drug transport and metabolism (Fig. 4b, Additional file 1: Table S5). Notably, most of the common drug metabolizers including *CYP3A4* [37], *CYP1A1* [38] and *CYP2D6* [39] contain more than five potentially deleterious nsSNVs, while the *CYP* genes that metabolize endogenous substance, e.g. *CYP51A1* [40], are not affected by any potentially deleterious nsSNVs. Similarly, important xenobiotic transporters including *ABCB1*, *ABCC1* and *ABCG2* [41] have 8, 11 and 7 potentially deleterious nsSNVs, respectively, while genes in *ABCD* subfamily, which are peroxisomal transporters for very long chain fatty acids [42], contain fewer (1–3) potentially deleterious nsSNVs. In addition to drug metabolizer and transporter, other protein families enriched in genes with more than five predicted deleterious nsSNVs include tyrosine protein kinases, dynein heavy chains, spectrins and myosins.

Notably, not only are the *ABC* transporters significantly enriched in predicted deleterious coding SNVs (Fig. 4b), they are also enriched with nsSNVs that are predicted to cause nonsense-mediated decay (NMD) resulting in the degradation of the mRNA transcripts with premature stop codon (Fig. 4c, clear bars; Additional file 1: Table S6). Genes containing nsSNVs predicted to cause NMD are also significantly enriched in cell cycle processes including mitosis (Fig. 4c, clear bars; Additional file 1: Table S6).

While the *ABC* transporters are significantly enriched with nsSNVs predicted to cause NMD, the other family of genes involved in drug response, the *CYP450*, is significantly enriched with genes having another form of deleterious polymorphism, namely, INDELs that cause frame-shift, which have deleterious effect on protein function (Fig. 4c, black bars; Additional file 1: Table S7). Taken together, genes involved in the xenobiotic response, including drug transport and metabolism, are significantly enriched with potentially deleterious coding polymorphisms.

Signatures of natural selection on the nsSNVs in drug-response genes are investigated. Interestingly, unlike the *CYP450* family (5/57 genes), not only are the *ABC* transporters enriched with potentially deleterious coding polymorphisms, they are also significantly enriched (*p* value < 0.001 by Fisher's exact test) with genes carrying nsSNVs under RPS (11/45 genes) (*p* value = 0.24 by Fisher's exact test) compared to the other genes in the human genome (Fig. 4d). As genes under positive selection also show significant population differentiation [34], we evaluate if the drug response genes are also enriched with nsSNVs that show significant population differentiation ($F_{ST} > 0.3$). Similar to the above observations, Fisher's exact test revealed that the *ABC* transporters (9/45) (*p* value < 0.001 by Fisher's exact test) but not the *CYP450* genes (3/57) (*p* value = 0.76 by Fisher's exact test) are significantly enriched with nsSNVs that show significant population differentiation (Fig. 4d). Hence, the nsSNVs in the *ABC* transporter family are under strong positive selection pressure.

As evident from the table in Fig. 4e, all, except one (rs4968839), of the nsSNVs at the *ABC* transporter family, which showed evidence of RPS or significant population differentiation, are predicted to either have a potentially deleterious effect on protein function or alter ESE/ESS modulating the proportion of the different splice forms. Notably, > 40% of these nsSNVs have been reported to be significantly associated with various phenotypes including clinically relevant ones [43–59] (Fig. 4e), highlighting the functional importance of the nsSNVs under natural selection at the *ABC* transporter gene family.

Discussion

In this study, we comprehensively investigate the architecture of genetic polymorphisms in the human genome and demonstrate that polymorphisms in coding regions, especially those affecting protein sequences and/or functions, are the most constrained in the human genome, consistent with previous observations [30]. In particular, frame-shift INDELs in coding regions are under strong purifying selection, consistent with the previous observation of the strongest depletion of frame-shift INDELs in coding regions, which are enriched with gene expression association possibly contributed by NMD [60].

Through the interrogation of nine global populations, we demonstrate that the median F_{ST} of SNVs at the coding regions is lower than that of the other regions of the genome (Fig. 1f). Moreover, the observations that coding regions have the lowest SNV density (Fig. 1b), excess of rare alleles (Fig. 1d) and enrichment of SNVs with no population differentiation (Fig. 1g) all indicate that coding SNVs are constrained by purifying selections. This is further strengthened by the observation that potentially deleterious nsSNVs show enrichment of rare alleles, compared to non-deleterious nsSNVs (Fig. 2c). Furthermore, coding regions contain significantly fewer INDELs that cause frame-shift (Fig. 1c). Hence, polymorphisms predicted to be deleterious to protein functions are under the strongest purifying selection.

In addition to the potentially deleterious nsSNVs, the potential functions and signatures of natural selections in the other polymorphisms are also investigated. Through computational prediction of the potential functions of SNVs, we observe that significantly more SNVs are predicted to alter TFBS than to code for a potentially deleterious nsSNV (Fig. 2b). Additionally, except for SNVs affecting TFBS, the other pfSNVs show more significant enrichment of rare alleles than nfSNVs in the same regions (Fig. 2c). Hence, pfSNVs are more constrained than the other SNVs in the same region, perhaps because they affect the functionally important regulatory sites. This observation is congruent with previous studies that reported stronger negative selection on conserved miRBS than other conserved 3′UTR sequences [31], though different prediction algorithms and SNV data were used. Levenstien and Klein also reported similar observation that SNVs in a few functional classes, e.g. non-synonymous, methylation sites and miRBS, are under negative selection compared to genome, and suggested that they are promising candidates for functional characterization [61]. While these previous studies examined SNVs residing within regulatory consensus sites of the promoter, this study focuses on SNVs that are predicted to either disrupt or create regulatory sites. Hence, the negative selective pressure on several different classes of pfSNVs suggests that pfSNVs are likely to influence gene functions and contribute to phenotypic changes.

This study also highlights that ‘master regulators’ of gene expression tend to be functionally conserved and maintained during evolution, while regulation of specific target genes is less constrained and flexible. This is evident from the observation that ‘master regulator’ involved in general gene regulation including epigenetics (e.g. histones), transcription/translation and splicing is significantly enriched with non-polymorphic, functionally conserved and ultra-conserved genes (Fig. 3b). In contrast, an average gene contains more SNVs affecting its own regulation than altering its function as evident from the observed enrichment of SNVs predicted to alter TFBS at promoters, ESE/ESS in coding regions and ISRE within introns compared to SNVs predicted to result in deleterious non-synonymous amino acid changes (Fig. 2b). This is consistent with previous observation that genetic variants occur more frequently in the miRNA target regions, compared to the functional regions within miRNAs [62]. In addition, Hsiao et al. demonstrated that alternative splicing events regulated by intronic genetic variants tend to be under positive selection [63], which is consistent with our results that intronic SNVs that potentially affect splicing mechanisms show enrichment of RPS SNVs, compared to the other functional classes (Fig. 2d). On the other hand, splicing factors are more conserved during evolution [63], and our study demonstrates that UCEs were enriched in the genes involved in RNA splicing (Fig. 3b).

On the other hand, genes that modulate response to environmental changes are the most polymorphic. The immune response *MHC* class I and class II genes, implicated in the pathogenesis of several autoimmune diseases, reside in the most polymorphic region of the human genome [36, 64] and carry the highest density of SNVs (Fig. 3c). Notably, this family of genes is significantly enriched in SNVs predicted to alter various regulatory elements including TFBS, ESE/ESS, ISRE and miRBS rather than protein function (Fig. 3e). In fact, while none of the RPS pfSNVs in the *MHC* family is predicted to cause a deleterious effect on protein function, 87 pfSNVs are found to display signature of RPS (Additional file 1: Table S8). Hence, the regulatory regions of the *MHC* family of genes are likely to be under strong positive selection, as previously suggested [65], and are functionally significant, regulating gene expression to modulate phenotypes. For example, a very well-studied polymorphism, rs9378249 upstream of the *HLA-B* gene, has previously been associated with bipolar disorder [66, 67] and hypertension [67]. This polymorphism is predicted to alter TFBS and exhibits the signature of RPS; hence, it may be a causal variant for the various diseases although the underlying molecular mechanism requires further validation.

Another class of genes that modulates response to the environment is the drug/xenobiotic response families of genes including the *ABC* transporter and the *CYP450* metabolism families of genes. Unlike the *MHC* immune genes, which are significantly enriched in regulatory SNVs predicted to modulate gene expression, these drug response gene families are enriched in SNVs that affect the functions of the proteins, namely nsSNVs predicted to be deleterious (Fig. 4b). Previous reports also highlighted the high SNV density and excess of rare nsSNVs of the *CYP450* pathway [68, 69] with 90–95% individuals carrying at least one actionable variant in *CYP450* genes [70]. Another report predicted that ~ 32% (1949/6165) of SNVs at the *CYP450* loci are putatively functional with *CYP4F12* carrying amongst the most novel putatively functional variants [71] which is consistent with our observations that *CYP4F12* is enriched with the highest number of pfSNVs (Additional file 1: Table S9). In addition to the pfSNVs, RPS and highly population differentiated ($F_{ST} > 0.3$) SNVs are significantly represented in the *ABC* transporter genes but not in the *CYP450* genes (Fig. 4d) suggesting that the *ABC* transporter genes may be under stronger positive selection than the *CYP450* genes. For example, rs17822931, a coding variant at the *ABCC11* earwax determinant gene [59], is found to be highly differentiated amongst populations, and the *A* allele is positively associated with adaptation to cold climate [72]. Greater than 40% of these RPS and/or population differentiated SNVs in the *ABC* transporter genes have been associated with phenotype modulation and even diseases, e.g. Alzheimer’s and

Schizophrenia (table in Fig. 4e). Nearly all the coding SNVs at the *ABC* transporter gene family that display the signature of RPS or are significantly population differentiated are also predicted to alter ESE/ESS suggesting that differential splicing may also play an important role in the *ABC* genes to generate diverse splicing forms to respond to different environment.

Hence, adaptive genes that respond to environmental changes are likely to be highly polymorphic and subjected to strong positive selection pressures consistent with previous reports that variants associated with inflammatory diseases show evidence of RPS [73], and genes associated with pharmacogenomics show higher level of population differentiation, as a signature of positive selection [74]. Regulation of gene expression through variants that alter TFBS in the *MHC* gene family as well as modulation of protein function and/or splicing pattern in the *ABC* gene families highlight the different ways by different families of genes to adapt to the environment.

Conclusions

In conclusion, this study elucidates the overall architecture of the genetic polymorphisms, namely SNVs and INDELs, in the human genome. The coding region is found to be under strong negative selection, as being the least population differentiated, showing lowest densities of SNVs and INDELs (especially frame-shift INDELs), the highest proportion of rare alleles with less enrichment of RPS SNVs. SNVs predicted to be functional are found to be under negative selection with enrichment of rare alleles. Families of genes which are 'master regulators' of gene expression including those involved in epigenetics, transcription, translation or splicing are found to be least polymorphic, functionally conserved and/or enriched with ultra-conserved elements. Finally, genes that modulate response to the environment are the most polymorphic with the *MHC* gene family, which is involved in immune response, being the most polymorphic while genes involved in drug/xenobiotic response, including *ABC* transporter and *CYP450* genes, are the most enriched with functional nsSNVs.

Methods

Polymorphisms in the human genome

Polymorphisms from the dbSNP database (Build 131) were mapped to different genic regions (5′UTRs, coding regions, introns and 3′UTRs) of the human genes (NCBI Genome Build 37.1) with those residing outside genes classified as intergenic variants.

To minimize false-positive SNPs originating from highly paralogous sequences, which were estimated to be ~ 8% of biallelic coding SNVs in dbSNP129 [75], only polymorphisms, which mapped to a single location in the genome and have been validated using a non-computational method or have allele frequency information (e.g. from 1000 Genomes project), were included in this study. In the 1000 genomes project, the variant assignment was restricted to 'accessible genome', whereby ambiguously placed reads or unexpectedly high or low numbers of aligned reads were excluded (~ 15% genome) to minimize the detection of false-positive variants [76]. To evaluate if our data is valid, SNV density data of this study was compared and found to be comparable to the SNV density data calculated from whole-genome sequencing of 179 HapMap individuals [76] of 1000 genomes project. For example, similar to our observations using dbSNP data, the MHC gene loci from the 1000 genomes sequencing data were also found to be significantly more polymorphic than other human genes ($p < 0.001$ by Mann-Whitney test). Hence, results from sequencing data from the 1000 genomes project were consistent with the findings in this study using dbSNP data, suggesting that, in spite of the potential ascertainment biases and sequencing artefacts inherent in the dbSNP database, our findings about the enrichment of SNPs in MHC genes are valid.

Two major forms of genetic polymorphisms, SNVs and INDELs, were investigated. SNV/INDEL density within a particular genic region, e.g. 5′UTR, was calculated as the number of SNVs/INDELs divided by the length of that region. For genes with multiple transcripts, the mean densities were taken. Genes lacking polymorphism in all genic regions (promoter, 5′UTR, coding, intron, 3′UTR) were regarded as non-polymorphic genes. Highly polymorphic genes were identified based on a binomial model as described in Additional file 1.

Allele frequency of SNVs in the human genome was determined in the three population (East Asian, African and European) groups (HapMap release 28) as described in Additional file 1.

F_{ST}statistics [28] using the pooled allele frequencies in the three population groups was then calculated for each of the genotyped and polymorphic loci. Two groups of SNVs, namely, (1) zero-F_{ST} SNVs ($F_{ST} = 0$) and (2) high-F_{ST} SNVs ($F_{ST} > 0.3$), were further analysed. The fold enrichment of zero-F_{ST} or high-F_{ST} SNVs in a specific genic region (e.g. coding region) was determined by calculating the percentage of these SNVs in the coding region divided by the percentage of all the genotyped SNVs in the same region, and the significance of enrichment is determined using the Fisher's exact test. Fold enrichment, which significantly deviates from one, indicates that these SNVs are under- or over-represented in these regions.

Natural selections

Genic regions that display signatures of negative selection were previously reported to have excess rare derived alleles [31]. Hence, to identify the regions of genes subjected to negative selection, we determined if there is a statistical enrichment of rare SNVs (DAF < 0.05) in each genic region using the Fisher's exact test.

SNVs displaying signatures of RPS were identified using LD- or haplotype-based methods as described in [26, 77]. To identify the regions enriched in RPS SNVs, the percentage of RPS SNVs within the region was compared with the percentage of RPS SNVs in the whole genome, and significance of difference was determined using the Fisher's exact test.

UCEs are sequences within the genome that are 100% identical to the sequences with the mouse and the rat genomes [35], hence displaying evolutionary conservation and signatures of strong negative selection. A total of 481 UCEs have been identified [35], of which 70 are evolutionarily conserved coding sequences, overlapping with coding regions.

Potential functions of SNVs

The pfSNP database (http://pfs.nus.edu.sg/) [26], which integrates a variety of bioinformatics prediction algorithms, was used to evaluate potential functions of all the SNVs in the human genome that alter TFBS, protein functions, splicing events and miRBS. The prediction algorithms employed in this study are described in [26] and Additional file 1.

Functional annotation

The Database for Annotation, Visualization and Integrated Discovery [78, 79] was utilized for functional annotation of the genes of interest. The enrichment of the genes in PANTHER protein family, GO-molecular function, GO-biological process and KEGG pathway was investigated. Benjamini-Hochberg-corrected p value < 0.05 signifies statistical significance.

Additional file

Additional file 1: Supplementary Materials and Methods as well as **Tables S1-S9**. (PDF 830 kb)

Abbreviations

3′UTR: 3′ Untranslated region; 5′UTR: 5′ Untranslated region; *ABC*: ATP-binding cassette; *CYP450*: Cytochrome P450; DAF: Derived allele frequency; ESE/ESS: Exon splicing enhancer or silencer; *HBB*: Beta haemoglobin; INDEL: Insertion/deletion; ISRE: Intronic splicing regulatory element; *MHC*: Major histocompatibility complex; miRBS: MicroRNA binding site; miRNA: MicroRNA; nfSNV: Non-functional SNV; NMD: Nonsense-mediated decay; nsSNV: Non-synonymous SNV; *OR*: Olfactory receptor; pfSNV: Potentially functional SNV; RPS: Recent positive selection; SNV: Single-nucleotide variant; sSNV: Synonymous SNV; TFBS: Transcription factor binding site; UCE: Ultra-conserved element

Funding

This work was supported by a grant from the National Medical Research Council (NMRC) 1131/2007 and BioMedical Research Council - Science and Engineering Research Council (BMRC-SERC) 112 148 0008, as well as Block Funding from the National Cancer Centre Singapore and Duke-NUS Graduate Medical School Singapore to A/P GL Lee.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

CGLL and SSC conceived and designed the study. YJ conducted the statistical analyses. JW predicted the potential functions of the SNVs and identified the SNVs under RPS. MB analysed the population differentiation levels of the SNVs. YJ and CGLL wrote the manuscript, and SSC helped edit the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 117456, Singapore. [2]Department of Biochemistry, National University of Singapore, Singapore 119077, Singapore. [3]Division of Medical Sciences, National Cancer Centre, Singapore 169610, Singapore. [4]Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore. [5]Duke-NUS Graduate Medical School, Singapore 169547, Singapore.

Received: 20 May 2018 Accepted: 29 August 2018
Published online: 15 September 2018

References

1. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 1998;8:1229–31.
2. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009;325:1246–50.
3. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464:773–7.
4. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, et al. A survey of genetic human cortical gene expression. Nat Genet. 2007;39:1494–9.
5. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464:768–72.
6. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008;6:e107.
7. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet. 2005;1:e78.
8. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. Population genomics of human gene expression. Nat Genet. 2007;39:1217–24.
9. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008;4:e1000214.
10. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. PLoS One. 2010;5:e10693.

11. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
12. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. 2005;21:2814–20.
13. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006;7:166.
14. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002;12:436–46.
15. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25:2744–50.
16. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. Nucleic Acids Res. 2006;34:W645–50.
17. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22: 2729–34.
18. Masica DL, Karchin R. Towards increasing the clinical relevance of in silico methods to predict pathogenic missense variants. PLoS Comput Biol. 2016; 12:e1004725.
19. Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol. 2013;425:4047–63.
20. Faber K, Glatting KH, Mueller PJ, Risch A, Hotz-Wagenblatt A. Genome-wide prediction of splice-modifying SNPs in human genes using a new analysis pipeline called AASsites. BMC Bioinformatics. 2011;12(Suppl 4):S2.
21. Yang JO, Kim WY, Bhak J. ssSNPTarget: genome-wide splice-site single nucleotide polymorphism database. Hum Mutat. 2009;30:E1010–20.
22. Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J. SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. BMC Bioinformatics. 2008;9(Suppl 1):S2.
23. Bao L, Zhou M, Wu L, Lu L, Goldowitz D, Williams RW, Cui Y. PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits. Nucleic Acids Res. 2007;35:D51–4.
24. Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic Acids Res. 2010;38:D640–51.
25. Gong J, Tong Y, Zhang HM, Wang K, Hu T, Shan G, Sun J, Guo AY. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Hum Mutat. 2012;33:254–63.
26. Wang J, Ronaghi M, Chong SS, Lee CG. pfSNP: an integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. Hum Mutat. 2011;32:19–24.
27. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res. 2004;32:W242–8.
28. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 2002;12:1805–14.
29. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007; 449:913–8.
30. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet. 2008; 40:340–5.
31. Chen K, Rajewsky N. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet. 2006;38:1452–6.
32. Jha P, Lu D, Xu S. Natural selection and functional potentials of human noncoding elements revealed by analysis of next generation sequencing data. PLoS One. 2015;10:e0129023.
33. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. 2014;24:885–95.
34. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005; 39:197–218.
35. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. Ultraconserved elements in the human genome. Science. 2004;304:1321–5.
36. Vandiedonck C, Knight JC. The human major histocompatibility complex as a paradigm in genomics research. Brief Funct Genomic Proteomic. 2009;8: 379–94.
37. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther. 2013;138:103–41.
38. Beresford AP. CYP1A1: friend or foe? Drug Metab Rev. 1993;25:503–17.
39. Wang B, Yang LP, Zhang XZ, Huang SQ, Bartlam M, Zhou SF. New insights into the structural characteristics and functional relevance of the human cytochrome P450 2D6 enzyme. Drug Metab Rev. 2009;41:573–643.
40. Lepesheva GI, Waterman MR. Sterol 14alpha-demethylase cytochrome P450 (CYP51), a P450 in all biological kingdoms. Biochim Biophys Acta. 2007;1770: 467–77.
41. Dean M, Rzhetsky A, Allikmets R. The human ATP-binding cassette (ABC) transporter superfamily. Genome Res. 2001;11:1156–66.
42. Baker A, Carrier DJ, Schaedler T, Waterham HR, van Roermund CW, Theodoulou FL. Peroxisomal ABC transporters: functions and mechanism. Biochem Soc Trans. 2015;43:959–65.
43. Abellan R, Mansego ML, Martinez-Hervas S, Martin-Escudero JC, Carmena R, Real JT, Redon J, Castrodeza-Sanz JJ, Chaves FJ. Association of selected ABC gene family single nucleotide polymorphisms with postprandial lipoproteins: results from the population-based Hortega study. Atherosclerosis. 2010;211:203–9.
44. Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 2009;19:711–22.
45. Brion M, Sanchez-Salorio M, Corton M, de la Fuente M, Pazos B, Othman M, Swaroop A, Abecasis G, Sobrino B, Carracedo A, Spanish Multi-centre Group of AMD. Genetic association study of age-related macular degeneration in the Spanish population. Acta Ophthalmol. 2011;89:e12–22.
46. Chu LW, Li Y, Li Z, Tang AY, Cheung BM, Leung RY, Yik PY, Jin DY, Song YQ. A novel intronic polymorphism of ABCA1 gene reveals risk for sporadic Alzheimer's disease in Chinese. Am J Med Genet B Neuropsychiatr Genet. 2007;144B:1007–13.
47. Jamieson SE, de Roubaix LA, Cortina-Borja M, Tan HK, Mui EJ, Cordell HJ, Kirisits MJ, Miller EN, Peacock CS, Hargrave AC, et al. Genetic and epigenetic factors at COL2A1 and ABCA4 influence clinical outcome in congenital toxoplasmosis. PLoS One. 2008;3:e2285.
48. Jordan de Luna C, Herrero Cervera MJ, Sanchez Lazaro I, Almenar Bonet L, Poveda Andres JL, Alino Pellicer SF. Pharmacogenetic study of ABCB1 and CYP3A5 genes during the first year following heart transplantation regarding tacrolimus or cyclosporine levels. Transplant Proc. 2011;43:2241–3.
49. Junyent M, Tucker KL, Smith CE, Garcia-Rios A, Mattei J, Lai CQ, Parnell LD, Ordovas JM. The effects of ABCG5/G8 polymorphisms on plasma HDL cholesterol concentrations depend on smoking habit in the Boston Puerto Rican Health Study. J Lipid Res. 2009;50:565–73.
50. Kolovou V, Marvaki A, Karakosta A, Vasilopoulos G, Kalogiani A, Mavrogeni S, Degiannis D, Marvaki C, Kolovou G. Association of gender, ABCA1 gene polymorphisms and lipid profile in Greek young nurses. Lipids Health Dis. 2012;11:62.
51. Li Q, Yin RX, Wei XL, Yan TT, Aung LH, Wu DF, Wu JZ, Lin WX, Liu CW, Pan SL. ATP-binding cassette transporter G5 and G8 polymorphisms and several environmental factors with serum lipid levels. PLoS One. 2012;7:e37972.
52. Ma XY, Liu JP, Song ZY. Associations of the ATP-binding cassette transporter A1 R219K polymorphism with HDL-C level and coronary artery disease risk: a meta-analysis. Atherosclerosis. 2011;215:428–34.
53. Miura K, Yoshiura K, Miura S, Shimada T, Yamasaki K, Yoshida A, Nakayama D, Shibata Y, Niikawa N, Masuzaki H. A strong association between human earwax-type and apocrine colostrum secretion from the mammary gland. Hum Genet. 2007;121:631–3.
54. Nakano M, Miwa N, Hirano A, Yoshiura K, Niikawa N. A strong association of axillary osmidrosis with the wet earwax type determined by genotyping of the ABCC11 gene. BMC Genet. 2009;10:42.
55. Oh IH, Oh C, Yoon TY, Choi JM, Kim SK, Park HJ, Eun YG, Chung DH, Kwon KH, Choe BK. Association of CFTR gene polymorphisms with papillary thyroid cancer. Oncol Lett. 2012;3:455–61.
56. Ota M, Fujii T, Nemoto K, Tatsumi M, Moriguchi Y, Hashimoto R, Sato N, Iwata N, Kunugi H. A polymorphism of the ABCA1 gene confers susceptibility to schizophrenia and related brain changes. Prog Neuro-Psychopharmacol Biol Psychiatry. 2011;35:1877–83.
57. Sainz J, Rudolph A, Hein R, Hoffmeister M, Buch S, von Schonfels W, Hampe J, Schafmayer C, Volzke H, Frank B, et al. Association of genetic

polymorphisms in ESR2, HSD17B1, ABCB1, and SHBG genes with colorectal cancer risk. Endocr Relat Cancer. 2011;18:265–76.

58. Sundar PD, Feingold E, Minster RL, DeKosky ST, Kamboh MI. Gender-specific association of ATP-binding cassette transporter 1 (ABCA1) polymorphisms with the risk of late-onset Alzheimer's disease. Neurobiol Aging. 2007;28:856–62.
59. Yoshiura K, Kinoshita A, Ishida T, Ninokata A, Ishikawa T, Kaname T, Bannai M, Tokunaga K, Sonoda S, Komaki R, et al. A SNP in the ABCC11 gene is the determinant of human earwax type. Nat Genet. 2006;38:324–30.
60. Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. 2013;23:749–61.
61. Levenstien MA, Klein RJ. Predicting functionally important SNP classes based on negative selection. BMC Bioinformatics. 2011;12:26.
62. Saunders MA, Liang H, Li WH. Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci U S A. 2007;104:3300–5.
63. Hsiao YH, Bahn JH, Lin X, Chan TM, Wang R, Xiao X. Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. Genome Res. 2016;26:440–50.
64. Trowsdale J. The MHC, disease and selection. Immunol Lett. 2011;137:1–8.
65. Suo C, Xu H, Khor CC, Ong RT, Sim X, Chen J, Tay WT, Sim KS, Zeng YX, Zhang X, et al. Natural positive selection and north-south genetic diversity in East Asia. Eur J Hum Genet. 2012;20:102–10.
66. Jiang Y, Zhang H. Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. Genet Epidemiol. 2011;35:125–32.
67. Ross KA. Evidence for somatic gene conversion and deletion in bipolar disorder, Crohn's disease, coronary artery disease, hypertension, rheumatoid arthritis, type-1 diabetes, and type-2 diabetes. BMC Med. 2011;9:12.
68. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
69. Zhou Y, Ingelman-Sundberg M, Lauschke VM. Worldwide distribution of cytochrome P450 alleles: a meta-analysis of population-scale sequencing projects. Clin Pharmacol Ther. 2017;102:688–700.
70. Bank PCD, Swen JJ, Guchelaar HJ. Implementation of pharmacogenomics in everyday clinical settings. Adv Pharmacol. 2018;83:219–46.
71. Fujikura K, Ingelman-Sundberg M, Lauschke VM. Genetic variation in the human cytochrome P450 supergene family. Pharmacogenet Genomics. 2015;25:584–94.
72. Ohashi J, Naka I, Tsuchiya N. The impact of natural selection on an ABCC11 SNP determining earwax type. Mol Biol Evol. 2011;28:849–57.
73. Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. Common risk alleles for inflammatory diseases are targets of recent positive selection. Am J Hum Genet. 2013;92:517–29.
74. Amato R, Pinelli M, Monticelli A, Marino D, Miele G, Cocozza S. Genome-wide scan for signatures of human population differentiation and their relationship with natural selection, functional pathways and diseases. PLoS One. 2009;4:e7927.
75. Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Hum Mutat. 2010;31:67–73.
76. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
77. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:e72.
78. Huang d W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.
79. Huang d W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.