Title: Functional characterization of genomic variants from 1,092 humans

Authors:
Ekta Khurana *[,1,2], Yao Fu *[,1], Vincenza Colonna *[,3,4], Xinmeng Jasmine Mu *[,1], Hyun Min Kang [5], Tuuli Lappalainen [6,7,8], Lucas Lochovsky [1], Jieming Chen [1,9], Alexej Abyzov [1,2], Suganthi Balasubramanian [1,2], Kathryn Beal [10], Daniel Challis [11], Yuan Chen [3], Declan Clarke [12], Laura Clarke [10], Fiona Cunningham [10], Jishnu Das [13,14], Emmanouil T. Dermitzakis [6,7,8], Uday Evani [11], Paul Flicek [10], Robert Fragoza [14, 15] , Erik Garrison [16], Richard Gibbs [11], Arif Harmanci [1,2], Javier Herrero [10], Naoki Kitabayashi [17], Yong Kong [2,18], Kasper Lage [19,20,21,22,23], Steven Lipkin [24], Daniel G. MacArthur [20,25], Gabor Marth [16], Donna Muzny [11], Tune H. Pers [22,26,27], Graham R. S. Ritchie [10], Jeffrey A. Rosenfeld [28,29], Mark A. Rubin [17], Andrea Sboner [17,30], Cristina Sisu [1,2], Xiaomu Wei [13,24], Michael Wilson [1,31], Yali Xue [3], Fuli Yu [11], Haiyuan Yu [13,14] 1000 Genomes Project Consortium, Chris Tyler-Smith[¶ ,3], Mark Gerstein[¶ ,1,2,32]

*These authors contributed equally to this work
[¶] Co-corresponding authors (Chris Tyler-Smith: cts@sanger.ac.uk and Mark Gerstein: pi@gersteinlab.org)

Affiliations:

1. Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA
2. Dept of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA
3. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK.
4. Institute of Genetics and Biophysics, National Research Council (CNR), 80131 Naples, Italy.
5. Center for Statistical Genetics, Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
6. Dept of Genetic Medicine and Development, University of Geneva Medical School, Geneva, 1211 Switzerland.
7. Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland.
8. Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland.
9. Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT
10. European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK.
11. Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas 77030, USA.
12. Dept of Chemistry, Yale University, New Haven, Connecticut 06520, USA.
13. Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA.
14. Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA.

15. Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA
16. Dept of Biology, Boston College, Chestnut Hill, Massachusetts 02467, USA.
17. Department of Pathology and Laboratory Medicine, Weill Cornell Medical College of Cornell University, New York, NY, USA.
18. Keck Biotechnology Resource Laboratory, Yale University, New Haven, CT, USA.
19. Pediatric Surgical Research Laboratories, MassGeneral Hospital for Children, Massachusetts General Hospital, Boston, MA, US
20. Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, US
21. Harvard Medical School, Boston, Massachusetts, USA.
22. Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark
23. Center for Protein Research, University of Copenhagen, Copenhagen, Denmark
24. Department of Medicine, Weill Cornell College of Medicine, New York, NY 10021, USA.
25. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
26. Division of Endocrinology and Center for Basic and Translational Obesity Research, Children's Hospital, Boston, USA
27. Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, USA.
28. Department of Medicine, New Jersey Medical School, Newark, NJ 07101
29. IST/High Performance and Research Computing, University of Medicine and Dentistry of New Jersey, Newark, NJ 07101
30. Institute for Computational Biomedicine, Weill Cornell Medical College of Cornell University, New York, NY, USA.
31. Child Study Center, Yale University, New Haven, Connecticut 06520, USA
32. Dept of Computer Science, Yale University, New Haven, Connecticut 06520, USA.

One sentence summary:
Prioritization of non-coding variants in disease studies using patterns of polymorphisms in functional elements.

In several places of the text & figures, values for log odds ratios are reported. In many studies, the log base 10 is used. However, here it appears as if you used log base e. Shouldn't the be stated explicitly somewhere in the text or supp?

Abstract

Thousands of personal genomes are now available; however, we do not understand the consequences of most of their variants, especially the non-coding ones. Here, we relate the full spectrum of 1000-Genomes-Phase-I variants to functional annotation, finding patterns aiding personal-genome interpretation. Specifically, we identify non-coding regions under strong, "coding-like" negative selection (“ultra-sensitive” elements). Next, we show variants breaking transcription-factor binding-motifs are selected against. We also examine the interplay between selection and connectivity in molecular networks. Indels and structural variants largely follow the same trend as SNPs, exhibiting depletion in functional elements. However, the large size of structural variants can lead to exceptions, such as enrichment of deletions formed by non-allelic homologous recombination in enhancers. In addition, we find positive selection is prevalent in many regulatory elements, particularly in promoters of hub genes. Finally, applying workflows based on these patterns of selection to cancer genomes allows identification of candidate drivers in non-coding regions.

Introduction

Whole-genome sequencing has revealed millions of variants per individual in personal genomes. However, the functional implications of the vast majority of these variants remain poorly understood (*1*). It is well established that variants in protein-coding genes play a crucial role in human disease. Although it is known that non-coding regions are under purifying selection and variants in them have been linked to disease, their role is generally less well understood (*2-9*). Sequence conservation has been used to identify ultra-conserved elements under very strong negative selection in mammals (*10*), but an analogous study to identify the specific elements under strong purifying selection amongst humans is missing. Signatures of purifying selection identified using population-scale variation data should provide better insights into the significance of a genomic region in humans than evolutionary conservation. This genome show human-specific purifying selection, whil mammals show lack of functional activity and selectio

Does the "They" at the start of this sent refer to both indels and SVs, or just SVs? This is ambiguous. The implication is that you mean both of these, but the ref (12) provided seems to discuss SVs only...

Besides SNPs, the human genome also contains other variants including small insertions and deletions (indels) and structural variants (SVs) (*11*). They actually account for more nucleotide differences amongst humans than SNPs, hence an understanding of their relationship with functional elements is crucial (*12*).

It is also well known that there is a close relationship between positive selection on sequence polymorphisms amongst modern-day humans and disease (*13*) (for example, positively selected variants in *Hemoglobin-B* and *DARC* genes are associated with malaria resistance (*14, 15*)). While previous studies have used extreme population differentiation of polymorphic sites from HapMap to analyze positive selection in coding regions, a similar genome-wide analysis of the relationship of positive selection and non-coding regulatory regions remains to be done (*16*).

One of the primary aims of sequencing healthy genomes is to facilitate interpretation of disease genomes. Though recent studies have demonstrated a link between common variants from genome-wide association studies (GWAS) and regulatory regions (*2*), the deleterious effects of rare inherited variants and somatic cancer mutations in non-coding regions have not been explored to the same extent. In particular, availability of many cancer genome sequences (*17-21*) and presence of many somatic mutations in non-coding regions call for an integrated framework to facilitate their functional interpretation.

Here we use the full range of sequence polymorphisms (ranging from SNPs to SVs) from 1,092 humans across fourteen populations to study patterns of selection (negative and positive) in various functional categories, especially non-coding regulatory regions (*22*). We identify specific genomic regions where variants are more likely to have strong phenotypic impact. The list of these regions includes a group of coding genes and specific sites within them, and, importantly, particular non-coding elements. By further comparing patterns of polymorphisms with disease mutations, we show how this list can aid interpretation of disease variants, leading to identification of cancer drivers. We use multiple experimental methods for validations, including yeast two-hybrid experiments and Sanger sequencing of independent cancer samples.

Results

Details of all results and methods are provided in the *Supplement*, which is organized in a parallel fashion to the main text for easy cross-referencing.

Genomic elements under strong purifying selection

Enrichment of rare variants can be used to estimate the strength of purifying selection in different functional categories (*22*). As expected, we find that sequence variants from 1,092 individuals allow us to detect specific functional categories under strong purifying selection with much greater power than ~100 samples (sample size used for most previous studies) (*2, 7, 9*). In particular, the increased number of samples provides a better estimate of allele frequencies of SNPs in different categories (Fig S4), allowing identification of differential negative selection constraints amongst specific categories (for example, motifs of transcription factor families HMG and MADs-box).

Purifying selection in coding genes
Estimates of purifying selection obtained using enrichment of rare non-synonymous SNPs (derived allele frequency or DAF<0.5%) show that different gene categories exhibit differential selection constraints consistent with their known phenotypic consequences. Genes tolerant to loss-of-function (LoF) mutations are under weakest selection while cancer causal genes exhibit the strongest constraints (Fig 1A and Table S1). GWAS genes associated with complex disorders lie in between these two extremes, consistent with the presence of common genetic variants in them. Thus, genes seen to be under strong selection here may be prioritized in future disease studies (Data file S1).

Purifying selection in non-coding regulatory elements
After observing variable selection in different gene groups, we analyze heterogeneous selection constraints in non-coding regions. With the power of 1000 Genomes Phase I data, we aim to find the regulatory elements under very strong selection.

First we estimate the strength of negative selection in broad categories (for example, transcription factor binding sites (TFBSs), DNaseI hypersensitive sites (DHSs), ncRNAs, and enhancers) (Fig 2A). We find that, as observed before, most of these categories show slight but statistically significant enrichment of rare SNPs compared to the genomic average, while pseudogenes demonstrate a depletion (Fig 2A and Data File S2) (*2*).

By the way, why no captions devoted to Supp figs?

Next, we divide the broad categories into 677 specific, high-resolution ones. These span various genomic features that are likely to influence the extent of selection acting on the element. For example, TFBSs of different TF families are further divided into proximal vs distal and cell-line–specific vs –non-specific (Fig S5). For specific categories, we find heterogeneous degrees of negative selection (Fig 2B and Data File S2). For example, core motifs in the binding sites of the TF families HMG and Forkhead are under particularly strong selection, whereas those of the CBF-NFY family do not exhibit selection constraints (relative to genomic average) (Fig 2B). Similarly, amongst all the pseudogenes, polymorphic ones have the highest fraction of rare alleles consistent with their functional coding roles in some individuals (*23*). Overall, we find that 102 of the 677 categories show statistically significant selection constraints (Data File S2). We also find

But notice how the TFs which are not enriched in eQTLs are generally those w/more conserved motifs (see Fig 2B).

that eQTLs are enriched in the binding sites of many TF families, many of which are under purifying selection based on our analysis (Fig 2B). The association of TF binding and gene expression at these loci provides a plausible explanation for their phenotypic effects.

Identification of "sensitive" and "ultra-sensitive" regulatory regions

Amongst the 102 categories under significantly strong selection we define the top ~0.02% and ~0.4% regions of the genome as "ultra-sensitive" and "sensitive" respectively (Data File S3). These regions possess

This should really be "ultrasensitive regions", not "sensitive" (according to Fig 2B).

comparable to that for coding sequences (Fig 2C). comparison with Complete Genomics data. The sensitive regions include binding sites of some chromatin and general TFs (e.g. *BRF1* and *FAM48A*) and core motifs of some important TF families (e.g. JUN, HMG, Forkhead and GATA). Interestingly, for some TFs there is a strong difference between their proximal and distal binding sites; for example, *ZNF274* proximal binding sites are under strong selection and belong to the ultra-sensitive category, while its distal sites are not under negative selection.

Why interestingly though? Isn't this to be expected?

Motif-breaking and allelic SNPs

We next examine selection constraints at nucleotide-level resolution in TFBSs. First, we analyze sites at which SNPs break or conserve the core binding motifs. As expected, we find that disruptive motif-breaking SNPs (those that decrease the motif-matching score to the position weight matrix) are significantly enriched for rare alleles compared to motif-conserving SNPs (those that decrease the motif-matching score) for all TF families (p value < 2.2e-16) (Fig 2D). Interestingly, the difference between selection constraints on motif-breaking vs. -conserving SNPs varies for different TF families, possibly reflecting differences in the topology of their DNA binding domains (Data File S4).

We also examine SNPs from a personal genome (NA12878) that show allele-specific TF binding in ChIP-Seq data or allele-specific expression in noncoding regions in RNA-seq data (with the allele-specific "activity" tagging a difference between the maternal and paternal chromosomes at the genomic region in question). We find that matched sites lacking allele-specific activity are enriched for rare variants (Fig 2E). This suggests that regions where allelic regulatory differences are not tolerated may be under stronger purifying selection (*24*).

Tissue-specificity vs purifying selection

We find that core motif regions bound by TFs only in a single cell-line show weaker negative selection than those bound in multiple cell lines (Data File S2). This is consistent with the greater functional importance of ubiquitously bound regions.

We also examine how purifying selection constraints vary amongst coding genes and DHSs with tissue-specific activity (Fig 1B). We find pronounced differences between tissues, with brain- (p value = 5.08e-10) and ovary- (p value = 6e-03) specific genes showing significantly higher selection constraints than other genes (Fig 1B and Table S2). Of tissue-specific DHSs, the strongest signal of purifying selection is in connective tissue, foreskin and spinal cord. Our results suggest that deleteriousness of both coding and regulatory variants depends not only on how they affect the function of the cell, but also on which tissues are affected.

Consider replacing w/ something else -- is "deleteriousness" a word?

Purifying selection in the human proteome and regulome

Genes do not act in isolation, but interact with each other to perform specific tasks. Here we investigate the interplay between gene interactions and selection.

Relationship with network connectivity
We find a significant negative correlation between the DAF of SNPs and the degree centrality of genes in physical protein-protein interaction (PPI) (rho = -0.05; p value < 2.2e-16) and regulatory networks (rho = -0.02; p value = 7.3e-10). Thus, consistent with previous studies, we find that in general hub genes tend to be under stronger negative selection constraints (*24, 25*). Indeed, we find that centralities of different gene categories in the PPI network follow the same trend as differential selection constraints for SNPs: causal cancer genes show the highest connectivity and LoF-tolerant genes show the least with GWAS genes in the middle (Figs 1A and 3A). These results show that the interactions of a gene likely influence the strength of purifying selection acting upon it.

Strong selection at interaction interfaces
Hub proteins tend to have more interaction interfaces in the PPI netwo
corollary of the above is that interaction interfaces are themselves unc
selection, in turn leading to stronger constraints on hub proteins. Indee
SNPs disrupting interaction interfaces are enriched for rare alleles con
missense SNPs (p value < 2.2e-16) (Fig 3B). To further substantiate this conclusion, we test a specific case of Wiskott-Aldrich syndrome protein (*WASP*) using yeast two-hybrid (Y2H) experiments (*27*). We find that all three single nucleotide variants (SNVs) at *WASP* interaction interfaces disrupt its interactions with other proteins (Fig 3C). This observation provides biochemical support for the observed statistical trend of strong selection at interaction interfaces.

Are these SNPs in WASP "rare" by your definitions though? I think it's implied, but is it stated explicitly that you classify all 3 SNPs as rare? If not, do you at least classify the most delterious (based on the Y2H expt) SNP as rare?

Relationship of functional elements with indels and larger SVs

Next we analyze the relationship of small indels (<50 bp) as well as SVs (large deletions) with functional annotations. Similar to the results for non-synonymous SNPs, we find that genes linked with diseases show stronger selection against indels while LoF-tolerant genes show weaker constraints (relative to all genes), with a consistent trend for indels overall and frameshift indels in particular (Fig 4A, Fig S10 and Table S1).

The wide range of SV sizes (from ~50 bp to ~1Mb) leads to their diverse modes of intersection with functional elements; for example, a single SV breakpoint can split an element, a smaller SV can cut out a portion of a single element, or a large SV can engulf an entire element. To analyze the diverse effects of SVs, we computed the number of SVs that overlap with each functional category relative to a randomized control set. As expected, we find that coding genes (both coding sequences and gene elements including UTRs and introns) are depleted for SVs, suggesting that SVs that affect gene function are in general deleterious in the genome (Fig 4B) (*11*). However, when we further break down the mode of SV intersection with genes into partial (where an SV breakpoint splits a gene) and whole (where SV engulfs the entire gene), we find that, surprisingly, SVs are enriched for whole but depleted for partial gene overlap. This

suggests that partial gene overlaps with SVs are under stronger selection than whole gene overlaps, possibly since whole gene deletions and duplications may facilitate gene shuffling in the genome. Furthermore, another category of gene-related elements, pseudogenes, is enriched for SV intersection, consistent with pseudogene formation mechanisms by either duplication or retrotransposition.

Consistent with our expectations from analysis of SNPs, we find that SVs tend to be depleted for non-coding regulatory elements such as binding site motifs and enhancers (Fig 4B). However, surprisingly, enhancer elements are enriched for SVs formed by non-allelic homologous recombination (NAHR). This observation is further supported by high aggregation signal of activating histone marks (which are associated with enhancers, e.g. H3K4me1) around NAHR breakpoints (Fig 4C and Fig S11). The association of enhancers and NAHR deletions may be explained as follows: the 3D chromosomal structure in the nucleus brings enhancer elements into close proximity with the transcription start site of a gene (involving DNA "looping") – if these two 'non-allelic' loci contain homologous sequences, is favorable for NAHR to occur.

Before the word "is", insert "it"

Functional implications of positive selection amongst human populations

Negative selection is widespread in the genome and few regions escape its influence; nevertheless, some positions within negatively selected regions also experience positive selection (*13*). We have previously identified and experimentally validated one category of variants that are strong candidates for positive selection: sites at which pairs of continental populations show extreme differences in the frequency of the derived allele (HighD sites) (*22*). Here we examine the functional signatures of positive selection in the same fashion as we have done for negative selection – in coding genes, non-coding regulatory elements and networks of gene interactions. We note that the functional analysis of positive selection using highly differentiated sites is limited to SNPs, due to the low numbers of such indels and SVs in functional categories.

Positive selection in coding genes
Among coding elements, we observe enrichment of HighD sites in missense SNPs and UTRs (Fig 5A). Next, we examine different gene categories and observe that some disease gene groups (those in the OMIM, HGMD and GWAS catalogs) are enriched for HighD SNPs (Fig S12). Mutations in disease genes are likely to have strong phenotypic impacts and it is possible that some of these mutations confer advantage for local adaptation. For example, while loss-of-function mutations in *ABCA12* lead to the severe skin disorder Harlequin Ichthyosis (*28*), we find that a SNP within the second intron of this gene is a HighD site (DAF >90% in Europe and East Asia; 13% in Africa), possibly reflecting adaptations of the skin to lower levels of sunlight outside of Africa.

Positive selection in non-coding regulatory regions
Similar to our analysis of negative selection, we analyzed the enrichment of HighD sites in broad as well as specific non-coding categories. We find that HighD sites are significantly enriched in many non-coding categories (Fig 5A). These include regions of open chromatin in multiple cell-lines (cell-line non-specific DHSs), distal DHSs and binding sites of sequence-specific TFs (specifically those in ZNF and NR families).

Why would this be expected? Maybe after reading further into the discussion or supp the rationale for this is fuller, but at face value, there's no clear biological rationale for why genes w/HighD SNPs have lower degree in any network.

Positive selection and gene interaction networks

We find that, as expected, coding genes with HighD SNPs tend to have lower degree centralities in both PPI and regulatory networks (although the small number of these cases does not give rise to statistical significance) (Fig 5B and Fig S13) (*25*). The availability of a comprehensive genome-wide catalog of HighD sites also allows us to relate positive selection in TFBSs to the topology of the human regulatory network. In an opposite trend to genes (for which positive selection occurs on the network periphery), we find that HighD sites in TFBSs tend to occur in hub promoters (p value = 0.02) (Fig 5B). While it has been proposed previously that mutations in cis-elements in regulatory networks might play a significant role in development, our study shows that indeed hub promoters have undergone adaptive evolution (*29, 30*).

Polymorphisms vs. disease variants

One of the major aims of understanding functional properties of sequence polymorphisms is to identify disease-causing variants in personal genomes. In this section we show how the patterns of selection in functional elements described above can practically help in disease genome interpretation. In particular, we summarize our results in the form of a workflow for identifying harmful variants in any genome.

Inherited disease mutations

First, we examine the presence of inherited regulatory disease-causing mutations from HGMD in regulatory regions classified as “sensitive” and “ultra-sensitive” (*31*). We find significant enrichment of disease-causing mutations in these regions (compared to the entire non-coding sequence, p value < 2.2e-16) (Fig 6A). Thus, these documented disease-causing variants provide strong independent validation for the high functional importance of sensitive regions. For example, causal mutations in congenital erythropoietic porphyria (a rare disorder) occurring upstream of Uroporphyrinogen III synthase function through disruption of *GATA1* binding motif, classified as sensitive here (*32*). Similarly, the well-known disease-causing ncRNA *RMRP* is in the binding site of *BRF2*, classified as ultra-sensitive here (*33*).

Before "classified", insert "is"?

Before "classified", insert "is"?

Somatic cancer variants

After validating the importance of sensitive regions using known inherited disease mutations, we examine the functional properties of somatic cancer variants. Since somatic variants from diverse cancer types might exhibit very different sets of properties, we analyzed variants from a wide range of cancer types: prostate, breast and medulloblastoma (*34-36*).

(a) Prevalence in non-coding regions

We find that an overwhelming number of somatic cancer variants occur in non-coding regulatory regions, including TFBSs, ncRNAs and pseudogenes (Fig S14). Indeed, we find that some non-coding elements from our functional categories show recurrent mutations (occurring in multiple samples), pointing to the possible role of these mutations as drivers (i.e. mutations providing a selective advantage to the tumor cells) (Fig S15). For example, the pseudogene *RP5-857K21.6* is mutated in three prostate cancer samples, and the promoter of *RP1* is mutated in two prostate cancer samples.

Are you implying that a variant in this pgene acts as a driver mutation? This seems very counter-intuitive though... use a different example?

(b) Enrichment of deleterious mutations
Analysis of healthy and cancer tissues from the same individuals shows that somatic variants tend to show an enrichment of missense, LoF, sensitive and ultra-sensitive variants (Fig 6B, Fig S16 and Table S3). Moreover, consistent with this trend, we find higher TF-motif-breaking/conserving ratios for somatic variants than germline variants across many different samples and cancer types (Fig 6C and Table S4). Thus, somatic cancer variants are generally enriched for functionally deleterious mutations.

Functional interpretation of disease genomes
This enrichment of functionally deleterious mutations amongst somatic variants is likely because they are not under organism-level natural selection (unlike inherited disease mutations, including GWAS variants). Indeed, amongst all somatic mutations, those most deviating from patterns of natural polymorphisms are most likely to be cancer drivers. Consistent with this reasoning, our analysis has shown that amongst all disease mutations, those causing cancer occur in genes under strongest negative selection (and with highest network connectivity) (Fig 1A and Fig 3A). Based on this, we argue that somatic variants in non-coding elements under strongest selection (and those associated with hubs) are also likely to be cancer drivers. Thus, cancer variants provide a particularly suitable case study to contrast with trends of selection in functional elements observed here and identify the most deleterious variants. Below we discuss filtering of thousands of somatic variants to pick candidate drivers for further experimental follow-up.

Our general workflow is as follows:

Step I. Screening somatic variants against 1000 Genomes variants since driver mutations are not likely to be present as polymorphisms.

Step II. Filtering out the variants that are not functionally annotated and dividing the remaining ones into coding and non-coding.

Step III. Picking those that occur in regions under strong negative selection and/or are particularly disruptive (for example, LoF and motif-disrupting SNVs).

Step IV. Retaining those in a network hub (for coding) or associated with a hub (for non-coding).

Step V. Further prioritizing recurrent mutations (where the same functional element is mutated in multiple cancer samples).

We demonstrate the application of this scheme to two sets of somatic variants from breast and prostate cancer (Fig 6D). In a breast cancer sample, this approach yields one non-coding SNV that is likely to have strong phenotypic consequences due to the following reasons: (1) It occurs in a region classified as ultra-sensitive (*BRF2* binding site). (2) It breaks a *PAX-5* TF binding motif. (3) It is associated with a network hub (Specifically, this locus has been annotated as a distal regulatory module with three potential targets genes, out of which one is a hub in the PPI network (*ERCC1*, degree centrality = 63)) (*37*). (4) It is recurrent – i.e. the regulatory module contains somatic mutations in multiple breast cancer samples.

A similar approach applied to a prostate cancer sample points to two non-coding SNVs predicted to have strong functional consequences (Fig 6D). One of these variants in the ultra-sensitive category (*FAM48A* binding site) lies in the promoter of the *WDR74* gene

(a hub in PPI network with degree centrality = 56). We further tested the presence of mutations in this binding site by PCR followed by Sanger sequencing in an independent cohort of 19 prostate cancer samples (*38*). Interestingly, we find that one sample in the cohort also harbors mutations in this region (Fig 6E and Fig S17). Due to the rarity of recurrent mutations in prostate cancer samples, this finding provides strong support for a likely functional role of the *FAM48A* binding site.

But notice that in both Fig 6 and its respective caption, part "E" is never identified or labelled explicitly. Is it necessary to include the "E" label in the Fig and the figure caption?

Discussion

In this study we present a comprehensive functional characterization of a wide spectrum of genomic sequence variants – ranging from SNPs to large SVs. By using polymorphisms from 1,092 individuals forming Phase 1 of the 1000 Genomes Project (*22*), we are able to discern patterns of natural selection in functional regions. These patterns can then be used to infer functional consequences of variants discovered in personal genomes. Our approach is especially useful for non-coding regions because of the vast landscape of regulatory variants and lack of standard ways to prioritize them in disease studies. Since somatic cancer variants are not under organism-level natural selection, they are particularly informative cases to demonstrate the utility of our approach.

Firstly, we identify the specific regulatory regions under very strong selection in humans: the “sensitive” and “ultra-sensitive” elements. These regions comprise ~0.4% and ~0.02% of the genome and show strong enrichment of known, inherited disease-causing mutations. Since they cover a small fraction of the entire genome, we propose that these regions can be easily probed alongside exome sequences in clinical studies. Secondly, we find that functionally disruptive mutations tend to be under strong selection: in an analogous manner to LoF variants in coding genes, variants which break motifs in TF binding sites are selected against. Thirdly, there is a close relationship between connectivity in biological networks and negative selection: higher connectivity is generally associated with stronger selection.

We find that overall, selection against indels and large SVs acts in similar ways as against SNPs, though the large size of SVs sometimes leads to a complex relationship with functional elements. For example, though overall, functional elements are depleted for SVs – whole gene deletions and certain enhancer deletions are enriched.

Finally, in addition to examining trends of negative selection in various functional elements we also analyze the occurrence of positive selection. We find that many functional categories, in particular certain regulatory regions, are enriched for potentially positively selected SNPs. Positive selection in regulatory regions is understandable: mutations in cis-regulatory regions are likely to alter the binding of few specific TFs -- often in a tissue-specific manner, thereby modulating gene expression -- as opposed to mutations in the coding sequence, which generally tend to have ubiquitous and disruptive consequences. In personal genome studies, variants in these positively selected regions may be probed further for their functional and possibly advantageous roles.

Based on the patterns of selection in functional elements, we develop a practical workflow for personal genome interpretation. Application of this workflow to breast

cancer and prostate cancer genome leads to identification of candidate driver mutations, particularly in non-coding regions. Sanger sequencing of an independent cohort of prostate cancer samples provides further support for a regulatory variant identified as a possible driver. Though we use cancer genomes as specific case studies, the workflow presented here can be broadly used to identify the most harmful non-coding variants in the multitude of personal genomes expected to be sequenced in the near future.

References:

1. B. Yngvadottir, D. G. Macarthur, H. Jin, C. Tyler-Smith, The promise and reality of personal genomics. *Genome Biol* **10**, 237 (2009).
2. I. Dunham *et al.*, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (Sep, 2012).
3. M. T. Maurano *et al.*, Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190 (Sep, 2012).
4. L. D. Ward, M. Kellis, Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095 (Nov, 2012).
5. A. Visel *et al.*, Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409 (Mar, 2010).
6. W. Lee, P. Yue, Z. Zhang, Analytical methods for inferring functional effects of single base pair substitutions in human cancers. *Hum Genet* **126**, 481 (Oct, 2009).
7. L. D. Ward, M. Kellis, Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**, 1675 (Sep, 2012).
8. 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).
9. B. Vernot *et al.*, Personal and population genomics of human regulatory variation. *Genome Res* **22**, 1689 (Sep, 2012).
10. G. Bejerano *et al.*, Ultraconserved elements in the human genome. *Science* **304**, 1321 (May, 2004).
11. R. E. Mills *et al.*, Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59 (Feb, 2011).
12. R. Redon *et al.*, Global variation in copy number in the human genome. *Nature* **444**, 444 (Nov, 2006).
13. P. C. Sabeti *et al.*, Positive natural selection in the human lineage. *Science* **312**, 1614 (Jun, 2006).
14. J. Ohashi *et al.*, Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *Am J Hum Genet* **74**, 1198 (Jun, 2004).
15. M. T. Hamblin, A. Di Rienzo, Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* **66**, 1669 (May, 2000).
16. L. B. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, Natural selection has driven population differentiation in modern humans. *Nat Genet* **40**, 340 (Mar, 2008).
17. C. G. A. R. Network, Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609 (Jun, 2011).

18. C. G. A. Network, Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330 (Jul, 2012).
19. P. S. Hammerman *et al.*, Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519 (Sep, 2012).
20. C. G. A. Network, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61 (Oct, 2012).
21. T. J. Hudson *et al.*, International network of cancer genome projects. *Nature* **464**, 993 (Apr, 2010).
22. G. R. Abecasis *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (Nov, 2012).
23. Z. D. Zhang, A. Frankish, T. Hunt, J. Harrow, M. Gerstein, Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol* **11**, R26 (2010).
24. M. B. Gerstein *et al.*, Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91 (Sep, 2012).
25. P. M. Kim, J. O. Korbel, M. B. Gerstein, Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* **104**, 20274 (Dec, 2007).
26. P. M. Kim, L. J. Lu, Y. Xia, M. B. Gerstein, Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938 (Dec, 2006).
27. X. Wang *et al.*, Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol* **30**, 159 (Feb, 2012).
28. M. Akiyama *et al.*, Mutations in lipid transporter ABCA12 in harlequin ichthyosis and functional recovery by corrective gene transfer. *J Clin Invest* **115**, 1777 (Jul, 2005).
29. R. Haygood, O. Fedrigo, B. Hanson, K. D. Yokoyama, G. A. Wray, Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet* **39**, 1140 (Sep, 2007).
30. S. B. Carroll, Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25 (Jul, 2008).
31. P. D. Stenson *et al.*, The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13 (2009).
32. C. Solis, G. I. Aizencang, K. H. Astrin, D. F. Bishop, R. J. Desnick, Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria. *J Clin Invest* **107**, 753 (Mar, 2001).
33. P. Hermanns *et al.*, Consequences of mutations in the non-coding RMRP RNA in cartilage-hair hypoplasia. *Hum Mol Genet* **14**, 3723 (Dec, 2005).
34. M. F. Berger *et al.*, The genomic complexity of primary human prostate cancer. *Nature* **470**, 214 (Feb, 2011).
35. S. Nik-Zainal *et al.*, Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979 (May, 2012).
36. T. Rausch *et al.*, Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59 (Jan, 2012).
37. 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).

38. C. E. Barbieri *et al.*, Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* **44**, 685 (Jun, 2012).

Acknowledgements
We thank Gunther Boysen and Catherine O'Reilly for help with experimental validation of SNVs in prostate cancer samples and Kevin Yip for identifying target genes of DRMs. T.H.P. is supported by The Danish Council for Independent Research Medical Sciences (FSS). Funding support at the EBI is provided by the Wellcome Trust (grants WT085532 and WT095908) and the European Molecular Biology Laboratory

Figure legends

Change "legends" to "captions"? (Isn't a "legend" a convention key embedded within a figure, whereas a "caption" is a worded description accompanying a figure?)

Figure 1. Fraction of rare (DAF<0.5%) SNPs in coding genes. **(A)** In various gene categories. **(B)** In DHSs and coding genes which show tissue-specific behaviour. Matching tissues for which both DHSs and gene expression data are available are shown in the same colors: with shades of green for endodermal, grey for mesodermal and blue for ectodermal origin of tissues. Error bars in both (A) and (B) denote 95% binomial confidence intervals.

Figure 2. Fraction of rare SNPs in non-coding categories. Red dotted line represents genomic average. Error bars denote 95% binomial confidence intervals. **(A)** Broad categories. “Ultra-sensitive” and “sensitive” regions are those under very strong negative selection. TFSS, Sequence-specific transcription factors. **(B)** Example of specific high-resolution categories: TFBS binding motifs separated into 15 families. “e” (superscripts in red) denote the enrichment of eQTLs in TFBSs of specific families. **(C)** Examples of TFBSs included in “ultra-sensitive” category. **(D)** SNPs which break TF motifs show an excess of rare alleles compared to those that conserve motifs. Examples of motifs for two families are also shown. **(E)** SNPs which do not show allele-specific behavior (-) show enrichment of rare alleles compared to SNPs which show allele-specific behavior (+)**.** Red dotted line represents genomic average of fraction of rare SNPs in NA12878 genome.

Figure 3. SNPs in protein-protein interaction (PPI) network. **(A)** Degree centrality of coding gene categories in PPI network. **(B)** Fraction of rare missense SNPs at protein interaction interfaces is higher than all rare missense SNPs (error bars show 95% binomial confidence intervals) **(C)** Effects of SNVs at interaction interfaces on interactions of *WASP* with other proteins tested by Y2H experiments. Wild-type (WT) *WASP* interacts with all proteins shown, while each SNV disrupts its interaction with at least one protein.

Figure 4. Functional annotations of indels and SVs. **(A)** Fraction of rare indels in coding gene categories. **(B)** Enrichment of the number of SVs intersecting each category of functional annotation is computed relative to a randomized background. Enrichments are shown in green and depletions in red. Asterisks indicate significant enrichment or depletion with p value < 0.05 after Bonferroni correction for multiple hypothesis testing. SVs intersecting various functional categories in different modes (e.g. whole/partial) are

shown in the schematics on the right. SVs are broken into four different formation mechanisms: NAHR (non-allelic homologous recombination), NH (non-homologous), TEI (transposable element insertion) and VNTR (variable number of tandem repeats) **(C)** Aggregation of histone signal around breakpoints of deletions formed by different mechanisms. Breakpoints are centered at zero. Aggregation for upstream/downstream regions corresponds to negative/positive distance. Signals for an activating histone mark (H3K4me1) and a repressive mark (H3K27me3) are shown.

Figure 5. Functional characterization of positive selection. **(A)** Left panel shows frequency of HighD sites vs. matched sites for various categories. Right panel shows the ratio for TFBSs of specific TF families. Asterisk denotes significant enrichment after Benjamini-Hochberg correction for multiple hypothesis testing in both panels. “e” (superscripts in red) denote the enrichment of eQTLs. **(B)** Top left panel shows that the in-degree of genes with HighD missense SNPs is lower than that of all genes. Bottom left panel shows that in-degree of genes with HighD SNPs in their promoters is higher than all genes. Right panel shows the human regulatory network with edges in grey. Edges of some TFs are colored in light yellow for visualization. Blue nodes represent genes with HighD SNPs in their promoters and red nodes represent genes with HighD missense SNPs. Size of nodes is scaled based on their degree centrality. Nodes with higher centrality are bigger and tend to be in the center while those with lower centrality are smaller and tend to be on the periphery.

Figure 6. Functional interpretation of disease variants. **(A)** Enrichment of HGMD regulatory disease-causing mutations in ultra-sensitive, sensitive and annotated regions compared to all non-coding regions. **(B)** Enrichment of functional mutations amongst somatic SNVs compared to germline SNVs. Mean values from seven prostate cancer samples are shown (variation shown in Fig S16). **(C)** Ratios of number of SNVs that conserve vs. number of SNVs that break TF-binding motifs are depicted for NA12878, average of 1000 Genomes Phase I samples and the average of somatic and germline samples from a few different cancers. Error bars represent one standard deviation. MB medulloblastoma. **(D)** Filtering of somatic variants from a breast cancer (left) and a prostate cancer (right) sample leading to identification of candidate drivers. A part of the *FAM48A* binding site sequenced by Sanger sequencing in an independent cohort of 19 prostate cancer samples is shown in green (with the coordinates of mutations observed in one sample).

Here, is it necessary to insert (E) ?

Supplementary Materials

Details of all the materials and methods can be found in the Supplementary material. The Supplementary file also includes Figures S1 to S17 and Tables S1 to S4. Data files S1 to S4 are provided separately.

Figure 1

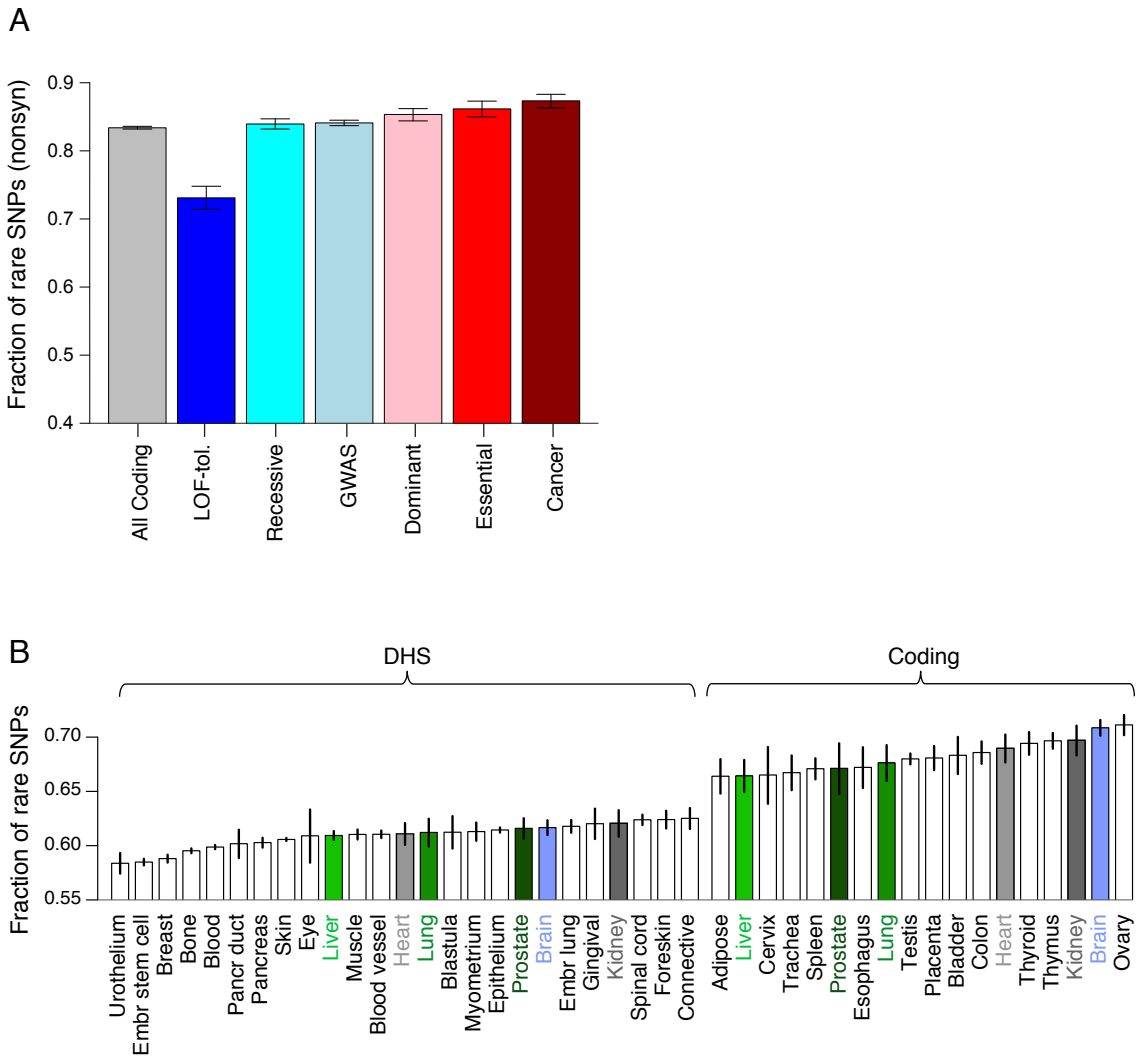
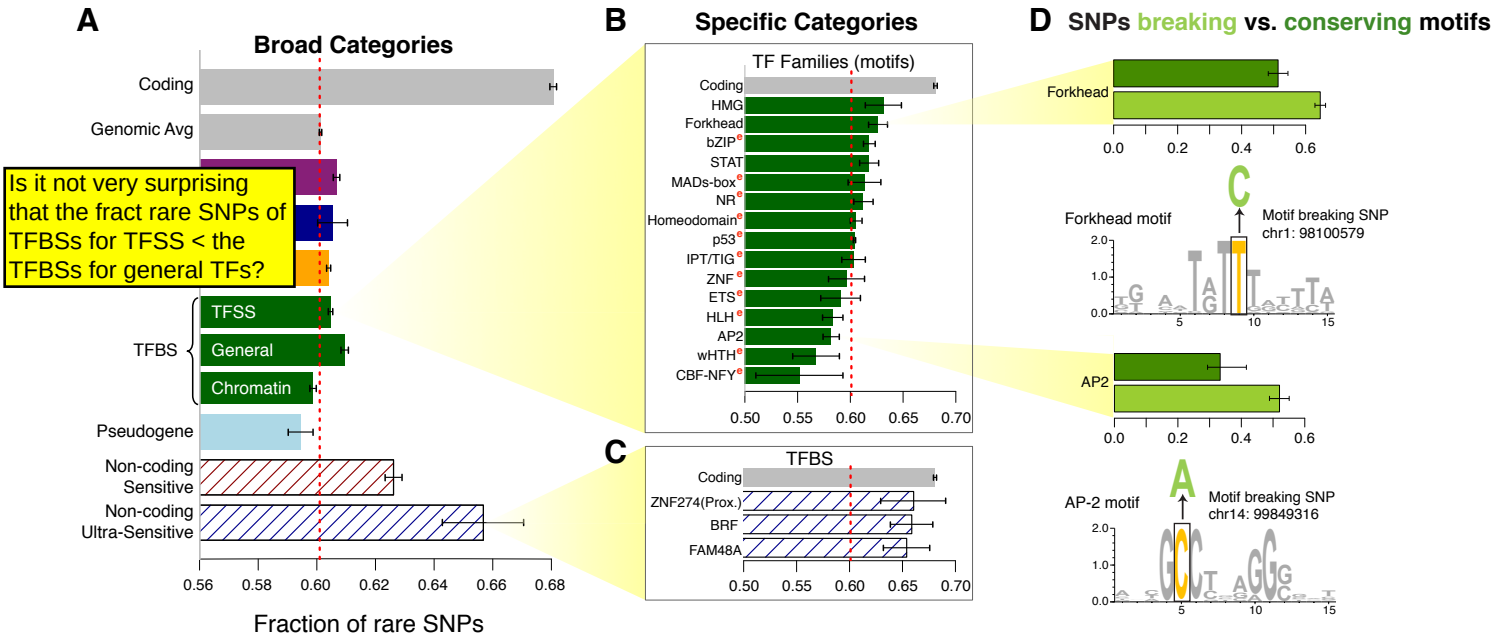

A
Fraction of rare SNPs (nonsyn)
0.9
0.8
0.7
0.6
0.5
0.4
All Coding
LOF-tol.
Recessive
GWAS
Dominant
Essential
Cancer
B
DHS
Coding
Fraction of rare SNPs
0.70
0.65
0.60
0.55
Urothelium
Embr stem cell
Breast
Bone
Blood
Pancr duct
Pancreas
Skin
Eye
Liver
Muscle
Blood vessel
Heart
Lung
Blastula
Myometrium
Epithelium
Prostate
Brain
Embr lung
Gingival
Kidney
Spinal cord
Foreskin
Connective
Adipose
Liver
Cervix
Trachea
Spleen
Prostate
Esophagus
Lung
Testis
Placenta
Bladder
Colon
Heart
Thyroid
Thymus
Kidney
Brain
Ovary

Figure 2


A
Broad Categories
Coding
Genomic Avg
TFBS
TFSS
General
Chromatin
Pseudogene
Non-coding Sensitive
Non-coding Ultra-Sensitive
0.56 0.58 0.60 0.62 0.64 0.66 0.68
Fraction of rare SNPs
Is it not very surprising that the fract rare SNPs of TFBSs for TFSS < the TFBSs for general TFs?
B
Specific Categories
TF Families (motifs)
Coding
HMG
Forkhead
bZIP*
STAT
MADs-box*
NR*
Homeodomain*
p53*
IPT/TIG*
ZNF*
ETS*
HLH*
AP2
wHTH*
CBF-NFY*
0.50 0.55 0.60 0.65 0.70
C
TFBS
Coding
ZNF274(Prox.)
BRF
FAM48A
0.50 0.55 0.60 0.65 0.70
D
SNPs breaking vs. conserving motifs
Forkhead
0.0 0.2 0.4 0.6
Forkhead motif
Motif breaking SNP
chr1: 98100579
AP2
0.0 0.2 0.4 0.6
AP-2 motif
Motif breaking SNP
chr14: 99849316

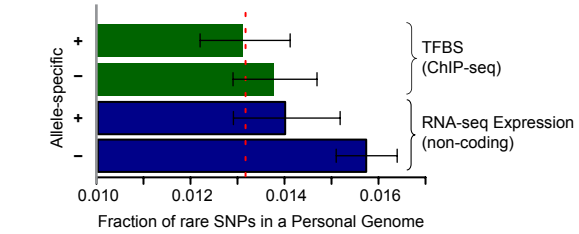

E
Allele-specific
+
−
+
−
TFBS
(ChIP-seq)
RNA-seq Expression
(non-coding)
0.010 0.012 0.014 0.016
Fraction of rare SNPs in a Personal Genome

Figure 3

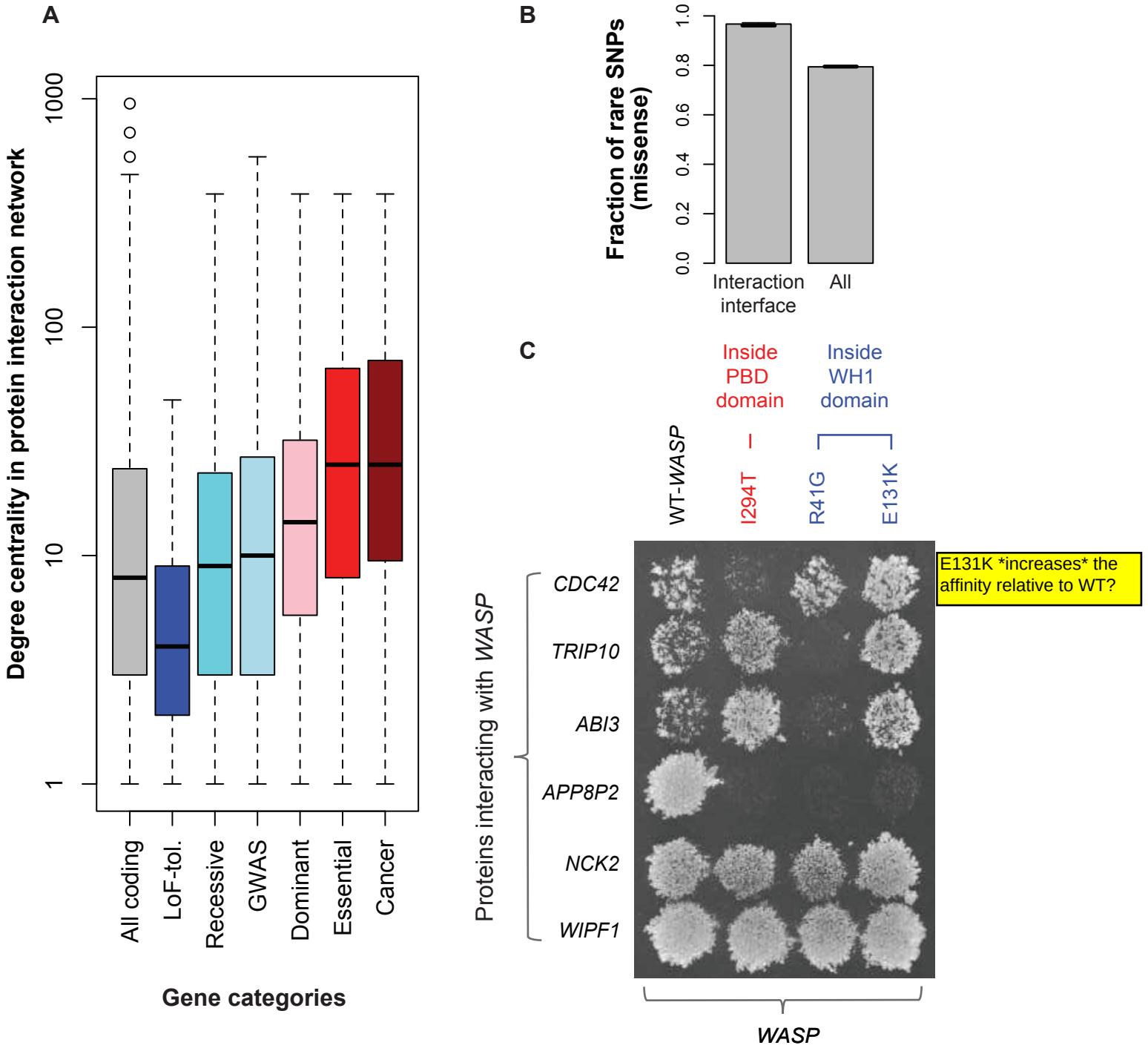

A
Degree centrality in protein interaction network
1000
100
10
1
All coding
LoF-tol.
Recessive
GWAS
Dominant
Essential
Cancer
Gene categories
B
Fraction of rare SNPs (missense)
0.0
0.2
0.4
0.6
0.8
1.0
Interaction interface
All
C
Inside PBD domain
Inside WH1 domain
WT-WASP
I294T
R41G
E131K
Proteins interacting with WASP
CDC42
TRIP10
ABI3
APP8P2
NCK2
WIPF1
WASP
E131K *increases* the affinity relative to WT?

Figure 4

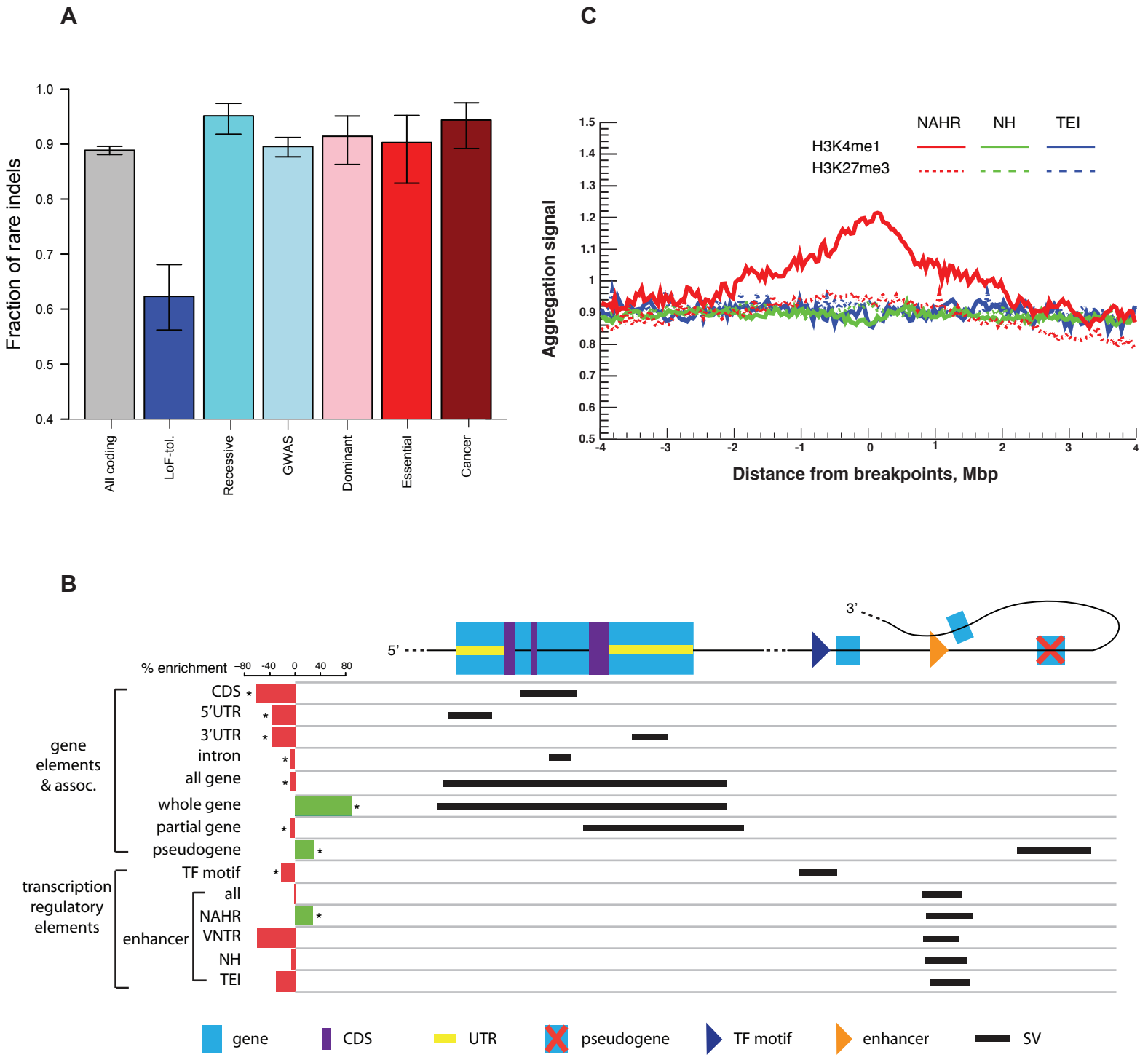
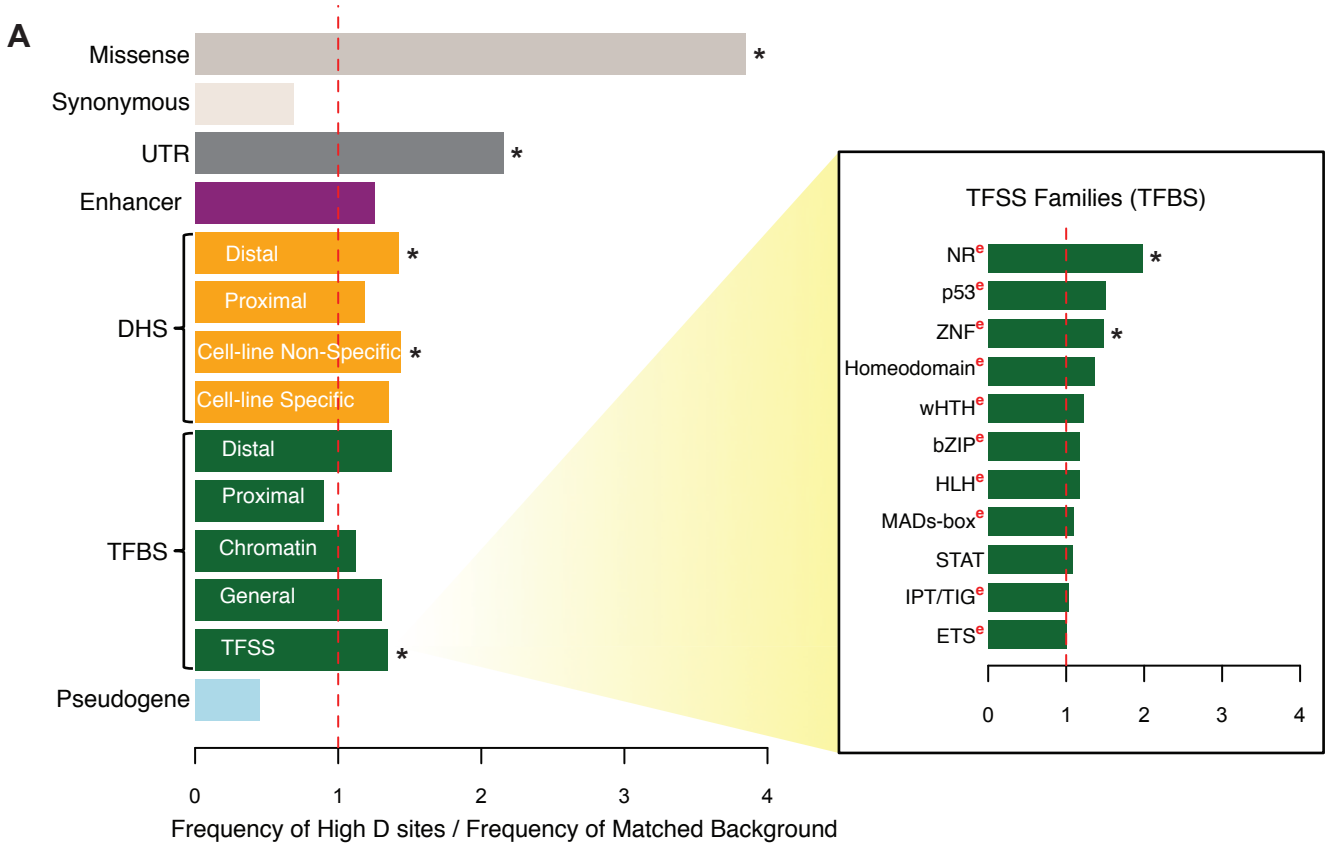

A
Fraction of rare indels
1.0
0.9
0.8
0.7
0.6
0.5
0.4
All coding
LoF-tol.
Recessive
GWAS
Dominant
Essential
Cancer
C
NAHR
NH
TEI
H3K4me1
H3K27me3
Aggregation signal
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
-4
-3
-2
-1
0
1
2
3
4
Distance from breakpoints, Mbp
B
5'
3'
% enrichment
−80
-40
0
40
80
CDS
5'UTR
3'UTR
intron
all gene
whole gene
partial gene
pseudogene
TF motif
all
NAHR
VNTR
NH
TEI
gene elements & assoc.
transcription regulatory elements
enhancer
gene
CDS
UTR
pseudogene
TF motif
enhancer
SV

Figure 5


A
Missense
Synonymous
UTR
Enhancer
DHS
Distal
Proximal
Cell-line Non-Specific
Cell-line Specific
TFBS
Distal
Proximal
Chromatin
General
TFSS
Pseudogene
0
1
2
3
4
Frequency of High D sites / Frequency of Matched Background
TFSS Families (TFBS)
NR^e
p53^e
ZNF^e
Homeodomain^e
wHTH^e
bZIP^e
HLH^e
MADs-box^e
STAT
IPT/TIG^e
ETS^e
0
1
2
3
4

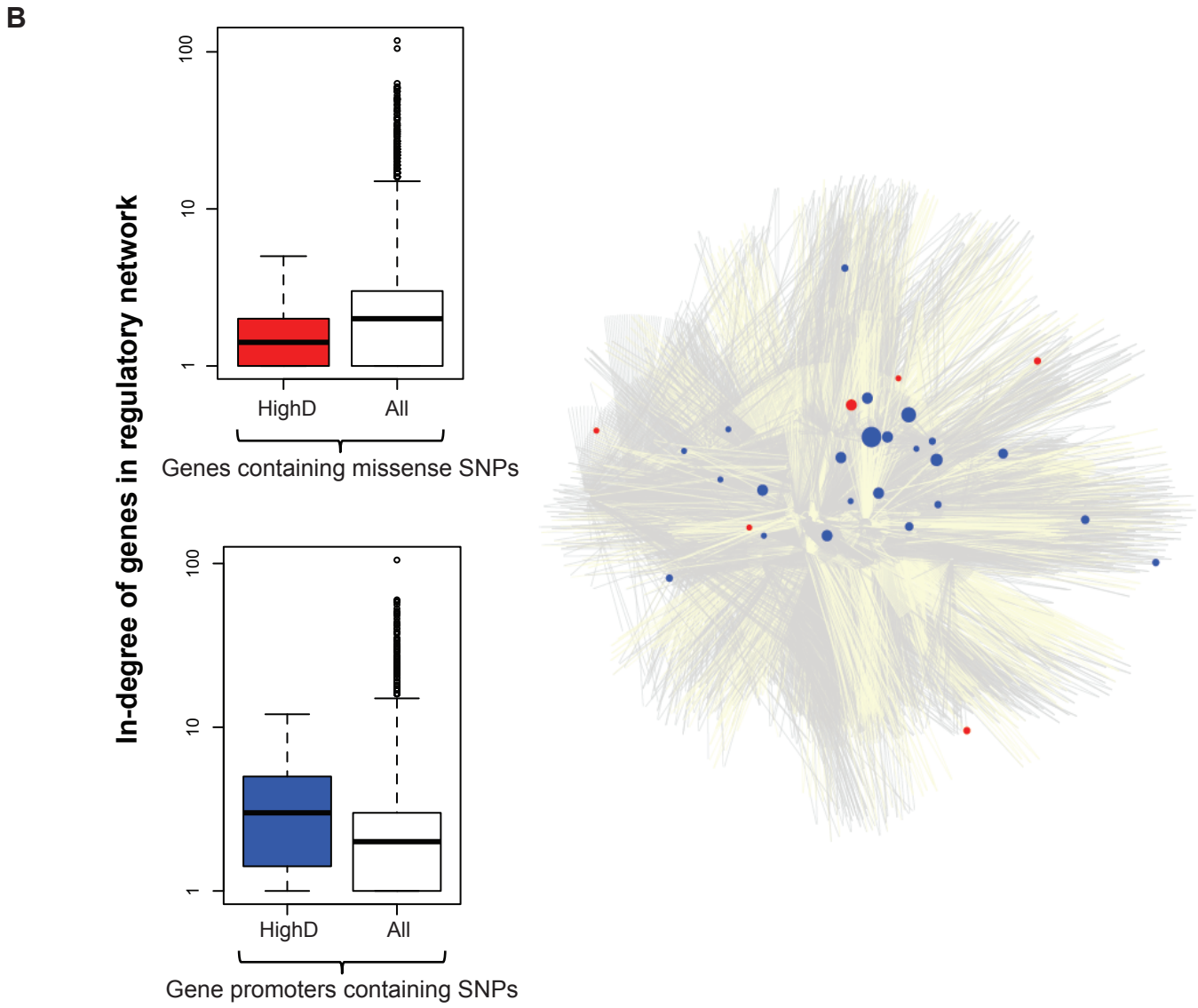
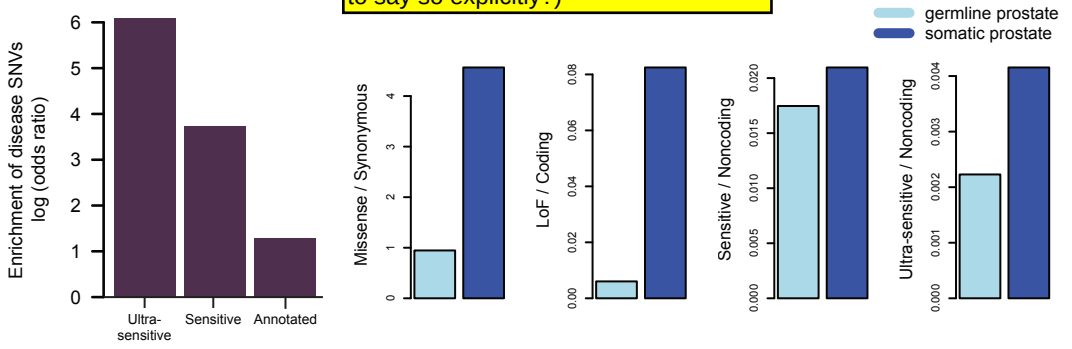
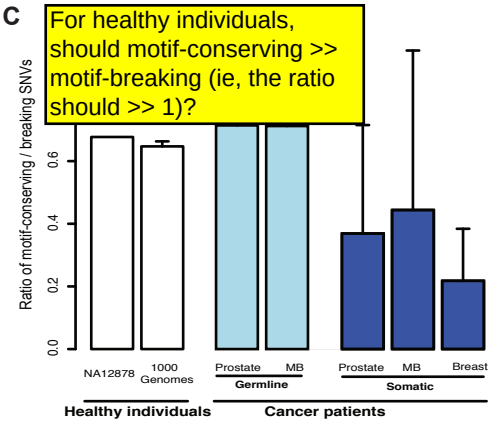

B
In-degree of genes in regulatory network
100
10
1
HighD
All
Genes containing missense SNPs
100
10
1
HighD
All
Gene promoters containing SNPs

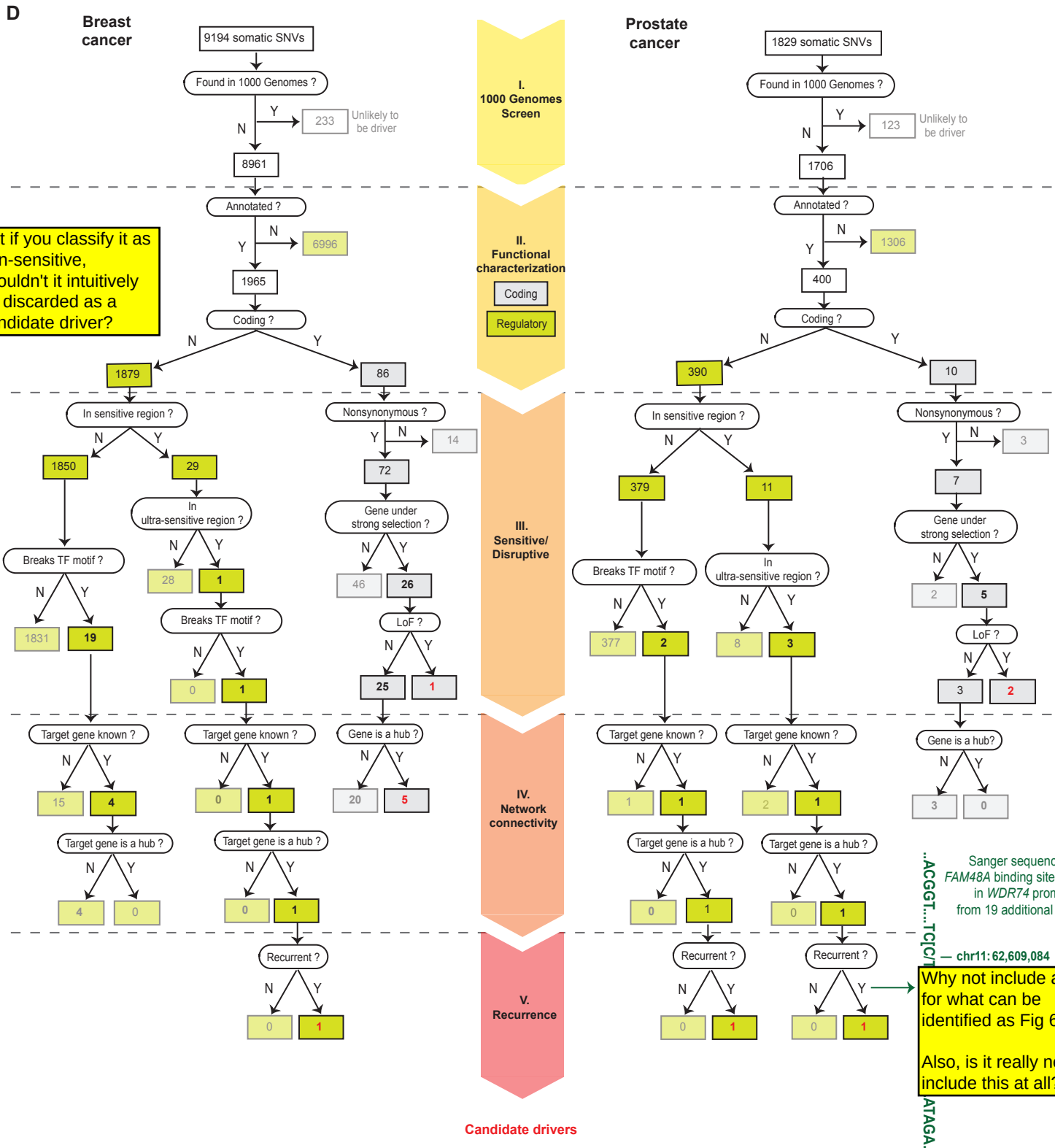
Figure 6

Why would misssence = synon in a healthy individual's germline? (or is this just a normalization? If so, necessarty to say so explicitly?)





For healthy individuals, should motif-conserving >> motif-breaking (ie, the ratio should >> 1)?



But if you classify it as non-sensitive, shouldn't it intuitively be discarded as a candidate driver?

Sanger sequencing of *FAM48A* binding site (~570 bp) in *WDR74* promoter from 19 additional samples

..ACGGT...TC[C/T]ATAGA..

— chr11: 62,609,084

Why not include a label for what can be identified as Fig 6E?

Also, is it really nec. to include this at all?

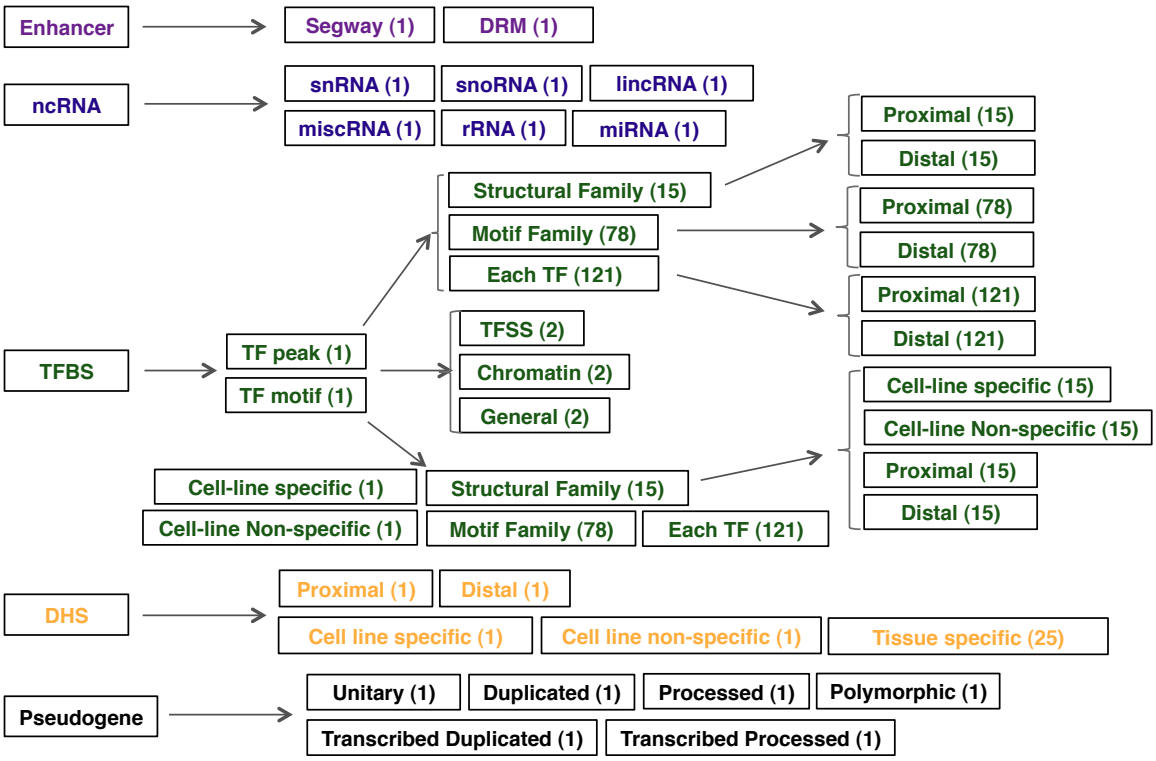


Figure S5 Broad and high-resolution categories. The numbers of sub-categories within each category are shown in brackets.

Nec. to change "brackets" to "parentheses"?

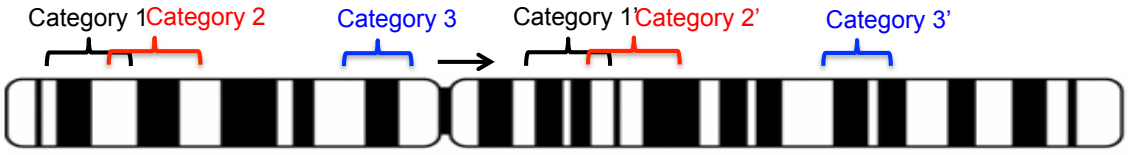
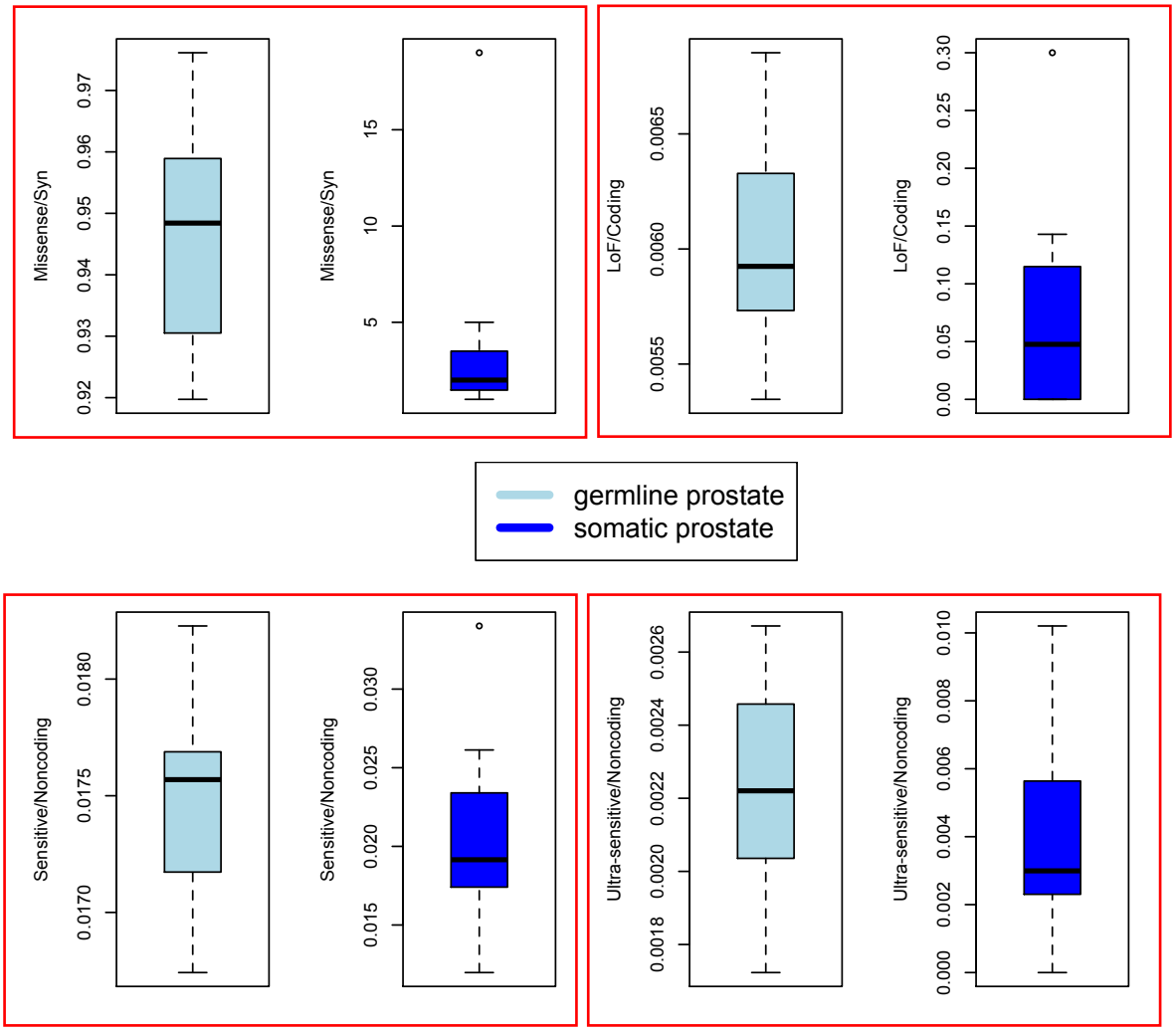


Figure S6 Schematic randomization procedures. Null distribution is obtained by sliding category coordinates along the genome for 1,000 times.

Would it be clearer to more clearly group or delineate each pair (since they share metrics)?



why diff scales? Better to share the same scale to accentuate the disparities?

Figure S16 Distributions of per sample ratios for somatic and matching germline SNVs in various functional categories across seven prostate cancer samples.

Supplementary Tables

Table S1 Binomial test p values for comparison of different gene categories with all genes. Significant p values are colored in grey.

These labels really need to be changed. It's not fract rare non-syn SNPs and fract rare indels that you're showing. It's the p-values for these values.

Gene category	Fraction of rare non-synonymous SNPs	Fraction of rare indels
LoF-tolerant	< 2.2e-16	< 2.2e-16
OMIM Recessive	1.566 e-01	5.878 e-04
GWAS	1.251 e-04	4.693e-01
OMIM Dominant	3.004e-05	3.359e-01
Essential	6.235e-06	7.55e-01
Cancer	1.17e-13	3.282e-02