Published in final edited form as:
Science. 2013 October 4; 342(6154): 1235587. doi:10.1126/science.1235587.

Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics

Ekta Khurana[#1,2], **Yao Fu**[#1], **Vincenza Colonna**[#3,4], **Xinmeng Jasmine Mu**[#1], **Hyun Min Kang**[5], **Tuuli Lappalainen**[6,7,8], **Andrea Sboner**[9,10], **Lucas Lochovsky**[1], **Jieming Chen**[1,11], **Arif Harmanci**[1,2], **Jishnu Das**[12,13], **Alexej Abyzov**[1,2], **Suganthi Balasubramanian**[1,2], **Kathryn Beal**[14], **Dimple Chakravarty**[9], **Daniel Challis**[15], **Yuan Chen**[3], **Declan Clarke**[16], **Laura Clarke**[14], **Fiona Cunningham**[14], **Uday S. Evani**[15], **Paul Flicek**[14], **Robert Fragoza**[13,17], **Erik Garrison**[18], **Richard Gibbs**[15], **Zeynep H. Gümüş**[10,19], **Javier Herrero**[14], **Naoki Kitabayashi**[9], **Yong Kong**[2,20], **Kasper Lage**[21,22,23,24,25], **Vaja Liluashvili**[10,19], **Steven M. Lipkin**[26], **Daniel G. MacArthur**[22,27], **Gabor Marth**[18], **Donna Muzny**[15], **Tune H. Pers**[24,28,29], **Graham R. S. Ritchie**[14], **Jeffrey A. Rosenfeld**[30,31,32], **Cristina Sisu**[1,2], **Xiaomu Wei**[13,26], **Michael Wilson**[1,33], **Yali Xue**[3], **Fuli Yu**[15], **1000 Genomes Project Consortium**[†], **Emmanouil T. Dermitzakis**[6,7,8], **Haiyuan Yu**[12,13], **Mark A. Rubin**[9], **Chris Tyler-Smith**[3,‡], and **Mark Gerstein**[1,2,34,‡]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 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, MI 48109, USA

[6]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, 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]Institute for Precision Medicine and the Department of Pathology and Laboratory Medicine, Weill Cornell Medical College and New York-Presbyterian Hospital, New York, NY 10065, USA

[10]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY 10021, USA

[11]Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT 06520, USA

‡Corresponding author. cts@sanger.ac.uk (C.T.-S.); mark.gerstein@yale.edu (M.G.).
†A full list of participants and institutions is available in the supplementary materials.

Supplementary Materials www.sciencemag.org/content/342/6154/1235587/suppl/DC1
Materials and Methods
Supplementary Text
Fig. S1 to S29
Tables S1 to S12
References (49-90)
Data S1 to S7

[12]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA

[13]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA

[14]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[15]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX 77030, USA

[16]Department of Chemistry, Yale University, New Haven, CT 06520, USA

[17]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

[18]Department of Biology, Boston College, Chestnut Hill, MA 02467, USA

[19]Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, 10065, USA

[20]Keck Biotechnology Resource Laboratory, Yale University, New Haven, CT 06511, USA

[21]Pediatric Surgical Research Laboratories, MassGeneral Hospital for Children, Massachusetts General Hospital, Boston, MA 02114, USA

[22]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

[23]Harvard Medical School, Boston, MA 02115, USA

[24]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark

[25]Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

[26]Department of Medicine, Weill Cornell Medical College, New York, NY 10065, USA

[27]Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, MA 02142, USA

[28]Division of Endocrinology and Center for Basic and Translational Obesity Research, Children's Hospital, Boston, MA 02115, USA

[29]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[30]Department of Medicine, Rutgers New Jersey Medical School, Newark, NJ 07101, USA

[31]IST/High Performance and Research Computing, Rutgers University Newark, NJ 07101, USA

[32]Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA

[33]Child Study Center, Yale University, New Haven, CT 06520, USA

[34]Department of Computer Science, Yale University, New Haven, CT 06520, USA

[#] These authors contributed equally to this work.

Abstract

Interpreting variants, especially noncoding ones, in the increasing number of personal genomes is challenging. We used patterns of polymorphisms in functionally annotated regions in 1092 humans to identify deleterious variants; then we experimentally validated candidates. We analyzed both coding and noncoding regions, with the former corroborating the latter. We found regions particularly sensitive to mutations ("ultrasensitive") and variants that are disruptive because of mechanistic effects on transcription-factor binding (that is, "motif-breakers"). We also found

variants in regions with higher network centrality tend to be deleterious. Insertions and deletions followed a similar pattern to single-nucleotide variants, with some notable exceptions (e.g., certain deletions and enhancers). On the basis of these patterns, we developed a computational tool (FunSeq), whose application to ~90 cancer genomes reveals nearly a hundred candidate noncoding drivers.

Whole-genome sequencing has revealed millions of variants per individual. 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 noncoding regions are under negative selection and that variants in them have been linked to disease, their role is generally less well understood (2-9).

In particular, whereas some studies have demonstrated a link between common variants from genome-wide association studies (GWASs) and regulatory regions (2, 3), the deleterious effects of rare inherited variants and somatic cancer mutations in noncoding regions have not been explored in a genome-wide fashion. Recently, three studies reported noncoding driver mutations in the *TERT* promoter in multiple tumor types, including melanomas and gliomas (10-12). In light of these studies and the growing availability of whole-genome cancer sequencing (13-20), an integrated framework facilitating functional interpretation of noncoding variants would be useful.

One may think to identify noncoding regions under strong selection purely through mammalian sequence conservation, and ultraconserved elements have been found in this fashion (21). However, signatures of purifying selection identified by using population-variation data could provide better insights into the importance of a genomic region in humans than evolutionary conservation. This is because many regions of the genome show human-specific purifying selection, whereas other regions conserved across mammals show a lack of functional activity and selection in humans (7). Thus, identifying the specific elements under particularly strong purifying selection among humans could provide novel insights.

Besides single-nucleotide polymorphisms (SNPs), the human genome also contains other variants, including small insertions and deletions (indels) and larger structural variants (SVs) (22). They account for more nucleotide differences among humans than SNPs; hence, an understanding of their relationship with functional elements is crucial (23).

We used the full range of sequence polymorphisms (ranging from SNPs to SVs) from 1092 humans to study patterns of selection in various functional categories, especially noncoding regulatory regions (24). We identified specific genomic regions where variants are more likely to have strong phenotypic impact. The list of these regions includes groups of coding genes and specific sites within them and, importantly, particular noncoding elements. By further comparing patterns of polymorphisms with somatic mutations, we show how this list can aid in the identification of cancer drivers. We used multiple experimental methods for validation, including yeast two-hybrid experiments, Sanger sequencing of independent cancer samples, and relevant gene-expression measurements. Furthermore, we provide a software tool that allows researchers to prioritize noncoding variants in disease studies.

Genomic Elements Under Strong Purifying Selection: Ultrasensitive Regions

Enrichment of rare variants can be used to estimate the strength of purifying selection in different functional categories (24). As expected, we found that having variants from 1092

individuals allowed us to detect specific functional categories under strong purifying selection with greater power than previously possible (2, 7, 9). In particular, the increased number of samples provided a better estimate of allele frequencies, making possible the measurement of differential selective constraints between specific categories [e.g., between motifs of transcription-factor (TF) families HMG and MADs box] (figs. S4 and S5).

Estimates of purifying selection obtained by using enrichment of rare nonsynonymous SNPs (derived allele frequency or DAF < 0.5%) showed that different gene categories exhibit differential selection consistent with their known phenotypic consequences (data S1). Genes tolerant of loss-of-function (LoF) mutations are under the weakest selection, whereas cancer-causal genes are under the strongest (Fig. 1A and table S1). GWAS genes associated with complex disorders lie in between these extremes, consistent with the presence of common genetic variants in them.

We then analyzed selective constraints in noncoding regions, trying to find elements under very strong selection (i.e., with a fraction of rare variants similar to that of coding genes, ~67%). We first estimated the strength of negative selection in broad categories [e.g., in all TF binding sites (TFBSs), deoxyribonuclease I (DNaseI)-hypersensitive sites (DHSs), noncoding RNAs (ncRNAs), and enhancers] (Fig. 2A). As observed previously, most of these categories show slight but statistically significant enrichment of rare SNPs compared with the genomic average; in contrast, pseudogenes demonstrate a depletion (Fig. 2A and data S2) (2).

We further divided the broad categories into 677 high-resolution ones. These span various genomic features likely to influence the extent of selection acting on the element. For example, TFBSs of different TF families are divided into proximal versus distal and cell-line–specific versus–nonspecific (fig. S7). We find heterogeneous degrees of negative selection for specific categories (Fig. 2B and data S2). For instance, core motifs in the binding sites of TF families HMG and Forkhead are under particularly strong selection, whereas those in the CBF-NFY family do not exhibit selective constraints (relative to the genomic average) (Fig. 2B). Among all the pseudogenes, polymorphic ones have the highest fraction of rare alleles, consistent with their functional coding roles in some individuals (25). Overall, we found that 102 of the 677 categories show statistically significant selective constraints (data S2) (figs. S8 to S10).

Among these 102 categories, we defined the top ones covering ~0.02% and ~0.4% of the genome as ultrasensitive and sensitive, respectively (fig. S11) (data S3). Thus, these regions were defined such that they possess a high fraction of rare variants comparable to that for coding sequences (67.2% for coding and 65.7% for ultrasensitive) (Fig. 2C). We validated the rare variants in them by comparison with Complete Genomics data. 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). For some TFs, there is a strong difference between proximal and distal binding sites—for example, for *ZNF274*, proximal binding sites are under strong selection and belong to the ultrasensitive category, whereas distal sites are not under negative selection.

In order to validate the functional importance of sensitive and ultrasensitive regions, we examined the presence of inherited disease-causing mutations from HGMD (Human Gene Mutation Database) in them (26). We found ~40- and ~400-fold enrichment of disease-causing mutations in sensitive and ultrasensitive regions, respectively (compared with the entire noncoding sequence, $P < 2.2 \times 10^{-16}$) (Fig. 2E). Thus, these documented disease-causing variants provide independent validation for the functional importance of sensitive regions. As a specific example, the disease congenital erythropoietic porphyria is caused by

disruption of a binding site classified as sensitive (the *GATA1* motif upstream of uroporphyrinogen-III synthase) (27). Similarly, the well-known disease-causing ncRNA *RMRP* is in the binding site of *BRF2*, classified as ultrasensitive (28).

Purifying Selection and Other Aspects of Regulatory Regions

We analyzed sites at which SNPs break or conserve core-binding motifs. As expected, we found that disruptive motif-breaking SNPs are significantly enriched for rare alleles compared with motif-conserving ones ($P < 2.2 \times 10^{-16}$; Fig. 2D; a motif-breaking SNP is defined as a change that decreases the matching score in the motif position weight matrix). This result is over all TF families; moreover, we find the difference between constraints on motif-breaking versus -conserving SNPs varies considerably for different TF families, possibly reflecting differences in the topology of their DNA binding domains (data S4).

We also found that expression quantitative trait loci (eQTLs) are enriched in the binding sites of many TF families (Fig. 2B); the association of TF binding and gene expression at these loci provides a plausible explanation for their phenotypic effects.

An analysis of SNPs from a personal genome (NA12878) exhibiting allele-specific TF binding in chromatin immunoprecipitation sequencing (ChIP-Seq) data or allele-specific expression in RNA-seq data (with the allele-specific "activity" tagging a difference between maternal and paternal chromosomes at the genomic region in question) showed that these sites are depleted for rare variants (relative to a matched control) (Fig. 2F). This suggests that regions where differential allelic activity is not observed may be under stronger purifying selection (29).

In a similar fashion, we found that core-motif regions bound in a "ubiquitous manner" (i.e., where differential cell-type-specific binding is not observed) are under stronger selection than those bound by TFs in a single cell line (data S2), consistent with the greater functional importance of ubiquitously bound regions. In relation to this, we further examined how selective constraints vary among coding genes and DHSs with tissue-specific activity (Fig. 1B). We found there are pronounced differences between tissues: For example, genes with ovary- and brain-specific expression are under significantly stronger selection than the average across all tissues (Fig. 1B and table S4). Similarly, some DHSs are under significantly stronger selection, whereas others are under relaxed constraints relative to the average (brain- and kidney-specific versus urothelium- and breast-specific, respectively; Fig. 1B and table S4). Last, matched expression and DHS data for six tissues indicate that purifying selection in tissue-specific genes and their corresponding regulatory regions is likely correlated (fig S15). Thus, our results suggest that the deleteriousness of both coding and regulatory variants depends on the tissues they affect.

Purifying Selection in the Interactome and Regulome

We found a significant positive correlation between the fraction of rare SNPs and the degree centrality of genes in networks: physical protein-protein interaction (PPI) (rho = 0.15; $P < 2.2 \times 10^{-16}$) and regulatory (rho = 0.07; $P = 6.8 \times 10^{-08}$). Thus, consistent with previous studies, we found that hub genes tend to be under stronger negative selection (29-31). Indeed, centralities of different gene categories in the PPI network follow the same trend as differential selective constraints on them: Cancer-causal genes show the highest connectivity, and LoF-tolerant genes, the least, with GWAS genes in the middle (Figs. 1A and 3A). These results indicate that the interactions of a gene likely influence the selection acting on it.

Hub proteins tend to have more interaction interfaces in the PPI network (31). A corollary of this is that interaction interfaces are themselves under strong selection, in turn leading to stronger constraints on hub proteins. Indeed, we found that SNPs disrupting interaction interfaces are enriched for rare alleles ($P < 2.2 \times 10^{-16}$) (Fig. 3B). To further corroborate this, we tested a specific case, the Wiskott-Aldrich syndrome protein (WASP), using yeast two-hybrid (Y2H) experiments (32). All of the three tested single-nucleotide variants (SNVs) at *WASP* interaction interfaces disrupted its interactions with other proteins (Fig. 3C). We observed similar behavior for two other proteins: Mutations at their interfaces disrupted specific protein interactions (fig. S16).

Relationship of Functional Elements with Indels and Larger SVs

We analyzed the association of functional annotations with small indels [<50 base pairs (bp)] and large SVs (deletions). Similar to the results for nonsynonymous SNPs, we found that genes linked with diseases show stronger selection against indels whereas LoF-tolerant genes show weaker constraints (relative to all genes), with a consistent trend for indels overall and frame-shift indels, in particular (Fig. 4A, fig. S17, and table S1).

The wide range of SV sizes (~50 bp to ~1 Mb) 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, and a large SV can engulf an entire element. To analyze the diverse effects of SVs, we computed the enrichment or depletion of SVs overlapping each functional category relative to a randomized control. As expected, we found that genic regions [coding sequences, untranslated regions (UTRs), and introns] are depleted for SVs, suggesting SVs affecting gene function are deleterious (Fig. 4B) (22). However, when we broke down the mode of SV intersection with genes into partial versus whole (an SV breakpoint splitting a gene versus an SV engulfing a whole gene), we unexpectedly found that SVs are enriched for whole-but depleted for partial-gene overlap. This suggests that partial-gene overlap is under stronger selection than whole-gene overlap, possibly because whole-gene deletions may be compensated by duplications. Furthermore, another category of gene-related elements, pseudogenes, are enriched for SVs, consistent with their formation mechanism involving either duplication or retrotransposition.

In relation to nongenic elements, we found that SVs tend to be depleted in regulatory elements such as binding-site motifs and enhancers (Fig. 4B), consistent with our expectations from SNPs. However, enhancer elements are enriched for SVs formed by nonallelic homologous recombination (NAHR). This observation is further supported by the high signal of activating histone marks associated with enhancers (e.g., H3K4me1) around NAHR breakpoints (Fig. 4C and fig. S18). The association of enhancers and NAHR deletions may be explained by the three-dimensional structure of chromatin bringing enhancer elements into close proximity with the gene transcription start site (via DNA "looping"). If these two "nonallelic" loci contain homologous sequences, it would be favorable for NAHR to occur.

Functional Implications of Positive Selection Among Human Populations

Negative selection is widespread in the genome; nevertheless, some positions within negatively selected regions also experience positive selection (33-36). We have previously identified and validated one category of variants that are strong candidates for positive selection: sites where continental populations show extreme differences in DAF (HighD sites) (24). By analyzing these HighD sites, we are focusing on positive selection under the classic selective-sweep model (37). Positive selection via other modes (such as selection on standing variation) likely also played a major role in recent human evolution (38).

Nonetheless, functional annotation of HighD sites can provide important insights about recent adaptations (39).

We examined positive selection in the same fashion as we have done for negative selection: in coding genes, noncoding regulatory elements, and networks of gene interactions. The functional analysis of positive selection using highly differentiated sites is limited to SNPs, because of the low numbers of such indels and SVs in functional elements.

We observed enrichment of HighD sites in UTRs and missense SNPs in coding regions (Fig. 5A). Next, we observed that some disease gene groups (Online Mendelian Inheritance in Man, HGMD, and GWAS) are enriched for HighD SNPs (fig. S20). Mutations in disease genes are likely to have strong phenotypic impact; thus, it is possible that some of these mutations confer advantage for local adaptation. For example, whereas LoF mutations in *ABCA12* lead to the severe skin disorder harlequin ichthyosis (40), we found 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 levels of sunlight outside of Africa.

Similar to our analysis of negative selection, we analyzed the enrichment of HighD sites in broad and specific noncoding categories, finding significant enrichment in many noncoding categories (Fig. 5A). These enriched categories include DHSs (particularly distal ones) and binding sites of sequence-specific TFs (specifically those in ZNF and NR families). Out of the seven enriched categories, five are also under significant negative selection (Figs. 2A and 5A and data S2). Thus, even though an entire category might be under negative selection, some particular sites within it can be targets of positive selection. In this respect, our results are consistent with previous studies for missense SNPs: Overall they are under strong negative selection, but a small group of them have been targets of positive selection (36).

We found that, as expected, coding genes with HighD SNPs tend to have lower degree centrality in both PPI and regulatory networks (although the small number of these cases does not produce statistical significance) (Fig. 5B and fig. S21) (41). In an opposite trend to genes (where positive selection occurs on the network periphery), HighD sites in TFBSs tend to occur in hub promoters ($P = 0.02$ with 23 promoters and $P = 3.2 \times 10^{-03}$ with37 proximal TFBSs) (Fig. 5B). It was previously proposed that mutations in cis elements in regulatory networks may play an important role in development (42, 43); our study supports this by suggesting that some hub promoters may have undergone recent adaptive evolution.

Contrasting Patterns of Somatic Mutations with Inherited Variants

After analyzing inherited polymorphisms in functional elements, we examined somatic variants. Because somatic variants from diverse tumors exhibit different sets of properties, we analyzed variants from a wide range of cancer types: prostate, breast, and medulloblastoma (17, 19, 20). We found that ~99% of somatic SNVs occur in noncoding regions, including TFBSs, ncRNAs, and pseudogenes (fig. S22).

Analysis of matched tumor and normal tissues from the same individuals showed that somatic variants tend to be enriched for missense (~5×), LoF (~14×), sensitive (~1.2×), and ultrasensitive (~2×) variants (Fig. 6A, fig. S24, and table S6). Consistent with this trend, we found higher TF-motif-breaking/conserving ratios for somatic variants compared with germline ones across many different samples and cancer types (~3 for somatic versus ~1.4 for germline) (Fig. 6B and table S7). Thus, somatic-cancer variants are generally enriched for functionally deleterious mutations.

This enrichment of functionally deleterious mutations among somatic variants is understandable because they are not under organism-level natural selection (unlike inherited-disease mutations, including GWAS variants). Indeed, among all somatic mutations, those most deviating from patterns of natural polymorphisms are the most likely to be cancer drivers. Consistent with this, our analysis has shown that, among all disease mutations, those causing cancer occur in genes under strongest negative selection (and with highest network connectivity) (Figs. 1A and 3A). Thus, we argue that somatic variants in the noncoding elements under strongest selection are the most likely to be cancer drivers.

Another feature of somatic mutations associated with their potential role as drivers is their recurrence in the same genomic element across multiple cancer samples. We found that some noncoding elements from our functional categories show recurrent mutations (fig. S23). For example, the pseudogene *RP5-857K21.6* is mutated in three out of seven prostate cancer samples, and the promoter of *RP1* is mutated in two (17).

FunSeq: Tool for Identification of Candidate Drivers in Tumor Genomes

On the basis of the integrative analysis above, we developed a tool to filter somatic variants from tumor genomes and obtain a short list of candidate driver mutations (funseq.gersteinlab.org). FunSeq first filters mutations overlapping 1000 Genomes variants and then prioritizes those in regions under strong selection (sensitive and ultrasensitive), breaking TF motifs, and those associated with hubs. It can score the deleterious potential of variants in single or multiple genomes and output the results in easy-to-use formats (i.e., “decorated” variant call format files, fig. S29 and data S6). The scores for each noncoding variant vary from 0 to 6, with 6 corresponding to maximum deleterious effect. When multiple tumor genomes are given as input, FunSeq also identifies recurrent mutations in the same element. Although our emphasis is on noncoding variants, it also outputs scores for coding variants.

We demonstrate the application of FunSeq as a workflow on representative breast and prostate cancer genomes (Fig. 6C). In the breast cancer sample, the workflow yielded one noncoding SNV likely to have strong phenotypic consequences: This SNV (i) occurs in an ultrasensitive region (*BRF2* binding site); (ii) breaks a *PAX-5* TF binding motif; (iii) is associated with a network hub (44); and (iv) is recurrent—that is, the regulatory module contains somatic mutations in multiple breast-cancer samples. In a similar fashion, the prostate-cancer sample revealed two noncoding SNVs predicted to have strong functional consequences (Fig. 6C). One of these is in an ultrasensitive region (*FAM48A* binding site) and lies in the promoter of *WDR74* gene (a hub in the PPI network with degree centrality = 56). We further tested the presence of mutations in this binding site by polymerase chain reaction followed by Sanger sequencing in an independent cohort of 19 prostate-cancer samples (45). We found that one sample in the cohort also harbors mutations in this region (Fig. 6D and fig. S25). Furthermore, we also observed increased expression of *WDR74* in the tumor relative to benign samples (fig. S26). These experimental results provide support for a likely functional role of this candidate driver.

A large-scale application of our tool to three medulloblastoma, 21 breast, and 64 prostate cancer genomes provided a total of 98 noncoding candidate drivers (table S8 and data S6) (17-20). Among these candidates, 68 occur in sensitive regions, 55 break TF motifs, and 90 target network hubs.

Generalized Identification of Deleterious Variants in Personal Genomes

Although we envision the most effective use of our tool for tumor genomes, it can also be applied to germline sequences to identify potentially deleterious variants. We applied it to

four personal genomes: Snyder, Venter, NA12878, and NA19240 (46-48). Out of ~3 million SNVs, we were able to identify ~15 (range from 6 to 26) noncoding SNVs per individual with high scores from FunSeq (>4), indicating their potential deleterious effects (Fig. 6E, tables S9 and S10, and data S6 and S7). Thus, our approach can be used to prioritize noncoding variants in personal genomes as well.

Discussion

We identified the sensitive and ultrasensitive noncoding elements, which exhibit depletion of common polymorphisms and strong enrichment of known, inherited disease-causing mutations. Because they cover a small fraction of the entire genome (comparable to the exome), these regions can be probed alongside exome sequences in clinical studies. We found that functionally disruptive noncoding mutations tend to be under strong selection: In an analogous manner to LoF variants in coding genes, variants that break motifs in TF binding sites are selected against. There is a close relation between connectivity in biological networks and selective constraints: Higher connectivity is generally associated with higher constraint. Furthermore, selection against indels and large SVs acts in a similar fashion as against SNPs overall; however, the large size of SVs sometimes leads to a complex relation with functional elements. On the basis of these patterns of negative selection in functional elements, we developed a workflow and a corresponding software tool to prioritize noncoding variants in disease studies.

The prioritization scheme presented in our paper can be readily extended by incorporation of genomic polymorphisms from larger populations and higher-resolution functional annotations. Moreover, with the availability of RNA-seq data from large cohorts, additional genomic features such as eQTLs can be folded in. Our approach can be immediately applied in precision medicine studies to prioritize noncoding variants for follow-up characterization, particularly candidate driver mutations in cancer, and it can be further extended in the future.

Materials and Methods

Details of all data sets and methods are provided in the supplementary materials. A brief summary of major data sets and methods is provided here. SNPs, indels, and SVs from 1000 Genomes Phase I release were used to investigate patterns of selection in DNA elements (24). Noncoding annotations were obtained from ENCODE Integrative paper release (2). Although we did analyze broad functional annotations, such as all TFBSs, we focused on highly specific categories such as distal binding sites of factor *ZNF274*. A randomization procedure, similar to the Genome Structure Correction (2), was developed by considering the dependency structure of different categories to deal with multiple hypothesis correction while identifying the categories under significantly strong selection. Patterns of somatic mutations were obtained from seven prostate cancer (17), three medulloblastoma (20), and 21 breast cancer genomes (19), whereas driver mutations were also identified in additional 57 prostate cancer genomes (18).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank G. Boysen and C. O'Reilly for help with SNV experimental validation, K. Yip for target-gene identification, and Z. Liu for Web site design. T.H.P. is supported by the Danish Council for Independent Research Medical Sciences (FSS). Funding at the European Bioinformatics Institute is provided by European Molecular

Biology Laboratory and the Wellcome Trust (WT085532 and WT095908). C.T.-S. acknowledges grant 098051 from the Wellcome-Trust Sanger Institute. Funding for the Institute for Precision Medicine (Weill Cornell Medical College/New York Presbyterian) is provided by National Cancer Institute (NCI) grant R01CA152057 (A.S., M.G., and M.A.R.) and Early Detection Research Network NCI U01 CA111275 (M.A.R.). M.A.R. also thanks the Prostate Cancer Foundation. M.G. also acknowledges grants HG005718 and HG007000. G.M. acknowledges National Human Genome Research Institute grants R01HG4719 and U01HG6513. H.Y. and S.M.L. are supported by NCI grant CA167824, National Institute of General Medical Sciences grant GM104424, and a Clinical and Translational Science Center Pilot Award and Cornell Seed Grant for intercampus collaborations.

References and Notes

1. Yngvadottir B, Macarthur DG, Jin H, Tyler-Smith C. The promise and reality of personal genomics. Genome Biol. 2009; 10:237. doi: 10.1186/gb-2009-10-9-237. pmid: 19723346. [PubMed: 19723346]
2. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. doi: 10.1038/nature11247; pmid: 22955616. [PubMed: 22955616]
3. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–1195. 10.1126/science.1222794. doi: 10.1126/science.1222794; pmid: 22955828. [PubMed: 22955828]
4. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. 2012; 30:1095–1106. doi: 10.1038/nbt.2422; pmid: 23138309. [PubMed: 23138309]
5. Visel A, et al. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature. 2010; 464:409–412. doi: 10.1038/nature08801; pmid: 20173736. [PubMed: 20173736]
6. Lee W, Yue P, Zhang Z. Analytical methods for inferring functional effects of single base pair substitutions in human cancers. Hum. Genet. 2009; 126:481–498. doi: 10.1007/s00439-009-0677-y; pmid: 19434427. [PubMed: 19434427]
7. Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. 2012; 337:1675–1678. 10.1126/science.1225057. doi: 10.1126/science.1225057; pmid: 22956687. [PubMed: 22956687]
8. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res. 2011; 39:7058–7076. doi: 10.1093/nar/gkr342; pmid: 21596777. [PubMed: 21596777]
9. Vernot B, et al. Personal and population genomics of human regulatory variation. Genome Res. 2012; 22:1689–1697. doi: 10.1101/gr.134890.111; pmid: 22955981. [PubMed: 22955981]
10. Horn S, et al. *TERT* promoter mutations in familial and sporadic melanoma. Science. 2013; 339:959–961. 10.1126/science.1230062. doi: 10.1126/science.1230062; pmid: 23348503. [PubMed: 23348503]
11. Huang FW, et al. Highly recurrent *TERT* promoter mutations in human melanoma. Science. 2013; 339:957–959. 10.1126/science.1229259. doi: 10.1126/science.1229259; pmid: 23348506. [PubMed: 23348506]
12. Killela PJ, et al. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc. Natl. Acad. Sci. U.S.A. 2013; 110:6021–6026. doi: 10.1073/pnas.1303607110; pmid: 23530248. [PubMed: 23530248]
13. Bell D, et al. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615. doi: 10.1038/nature10166; pmid: 21720365. [PubMed: 21720365]
14. Muzny DM, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337. doi: 10.1038/nature11252; pmid: 22810696. [PubMed: 22810696]
15. Hammerman PS, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–525. doi: 10.1038/nature11404; pmid: 22960745. [PubMed: 22960745]
16. Hudson TJ, et al. International network of cancer genome projects. Nature. 2010; 464:993–998. doi: 10.1038/nature08987; pmid: 20393554. [PubMed: 20393554]
17. Berger MF, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470:214–220. doi: 10.1038/nature09744; pmid: 21307934. [PubMed: 21307934]

18. Baca SC, et al. Punctuated evolution of prostate cancer genomes. Cell. 2013; 153:666–677. doi: 10.1016/j.cell.2013.03.021; pmid: 23622249. [PubMed: 23622249]
19. Nik-Zainal S, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012; 149:979–993. doi: 10.1016/j.cell.2012.04.024; pmid: 22608084. [PubMed: 22608084]
20. Rausch T, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell. 2012; 148:59–71. doi: 10.1016/j.cell.2011.12.013; pmid: 22265402. [PubMed: 22265402]
21. Bejerano G, et al. Ultraconserved elements in the human genome. Science. 2004; 304:1321–1325. 10.1126/science.1098119. doi: 10.1126/science.1098119; pmid: 15131266. [PubMed: 15131266]
22. Mills RE, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470:59–65. doi: 10.1038/nature09708; pmid: 21293372. [PubMed: 21293372]
23. Redon R, et al. Global variation in copy number in the human genome. Nature. 2006; 444:444–454. doi: 10.1038/nature05329; pmid: 17122850. [PubMed: 17122850]
24. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. doi: 10.1038/nature11632; pmid: 23128226. [PubMed: 23128226]
25. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. Identification and analysis of unitary pseudogenes: Historic and contemporary gene losses in humans and other primates. Genome Biol. 2010; 11:R26. doi: 10.1186/gb-2010-11-3-r26; pmid: 20210993. [PubMed: 20210993]
26. Stenson PD, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009; 1:13. doi: 10.1186/gm13; pmid: 19348700. [PubMed: 19348700]
27. Solis C, Aizencang GI, Astrin KH, Bishop DF, Desnick RJ. Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria. J. Clin. Invest. 2001; 107:753–762. doi: 10.1172/JCI10642; pmid: 11254675. [PubMed: 11254675]
28. Hermanns P, et al. Consequences of mutations in the non-coding RMRP RNA in cartilage-hair hypoplasia. Hum. Mol. Genet. 2005; 14:3723–3740. doi: 10.1093/hmg/ddi403; pmid: 16254002. [PubMed: 16254002]
29. Gerstein MB, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012; 489:91–100. doi: 10.1038/nature11245; pmid: 22955619. [PubMed: 22955619]
30. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science. 2002; 296:750–752. doi: 10.1126/science.1068696; pmid: 11976460. [PubMed: 11976460]
31. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLOS Comput. Biol. 2013; 9:e1002886. doi: 10.1371/journal.pcbi.1002886; pmid: 23505346. [PubMed: 23505346]
32. Wang X, et al. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat. Biotechnol. 2012; 30:159–164. doi: 10.1038/nbt.2106; pmid: 22252508. [PubMed: 22252508]
33. Sabeti PC, et al. Positive natural selection in the human lineage. Science. 2006; 312:1614–1620. doi: 10.1126/science.1124309; pmid: 16778047. [PubMed: 16778047]
34. Ohashi J, et al. Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. Am. J. Hum. Genet. 2004; 74:1198–1208. doi: 10.1086/421330; pmid: 15114532. [PubMed: 15114532]
35. Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. Am. J. Hum. Genet. 2000; 66:1669–1679. doi: 10.1086/302879; pmid: 10762551. [PubMed: 10762551]
36. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat. Genet. 2008; 40:340–345. doi: 10.1038/ng.78; pmid: 18246066. [PubMed: 18246066]
37. Xue Y, et al. Population differentiation as an indicator of recent positive selection in humans: An empirical evaluation. Genetics. 2009; 183:1065–1077. doi: 10.1534/genetics.109.107722; pmid: 19737746. [PubMed: 19737746]
38. Hernandez RD, et al. Classic selective sweeps were rare in recent human evolution. Science. 2011; 331:920–924. doi: 10.1126/science.1198878; pmid: 21330547. [PubMed: 21330547]

39. Grossman SR, et al. Identifying recent adaptations in large-scale genomic data. Cell. 2013; 152:703–713. doi: 10.1016/j.cell.2013.01.035; pmid: 23415221. [PubMed: 23415221]
40. Akiyama M, et al. Mutations in lipid transporter ABCA12 in harlequin ichthyosis and functional recovery by corrective gene transfer. J. Clin. Invest. 2005; 115:1777–1784. doi: 10.1172/JCI24834; pmid: 16007253. [PubMed: 16007253]
41. Kim PM, Korbel JO, Gerstein MB. Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. Proc. Natl. Acad. Sci. U.S.A. 2007; 104:20274–20279. doi: 10.1073/pnas.0710183104; pmid: 18077332. [PubMed: 18077332]
42. Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat. Genet. 2007; 39:1140–1144. doi: 10.1038/ng2104; pmid: 17694055. [PubMed: 17694055]
43. Carroll SB. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. Cell. 2008; 134:25–36. doi: 10.1016/j.cell.2008.06.030; pmid: 18614008. [PubMed: 18614008]
44. Yip KY, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. 2012; 13:R48. doi: 10.1186/gb-2012-13-9-r48; pmid: 22950945. [PubMed: 22950945]
45. Barbieri CE, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. 2012; 44:685–689. doi: 10.1038/ng.2279; pmid: 22610119. [PubMed: 22610119]
46. Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. doi: 10.1038/nature09534; pmid: 20981092. [PubMed: 20981092]
47. Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012; 148:1293–1307. doi: 10.1016/j.cell.2012.02.009; pmid: 22424236. [PubMed: 22424236]
48. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5:e254. doi: 10.1371/journal.pbio.0050254; pmid: 17803354. [PubMed: 17803354]

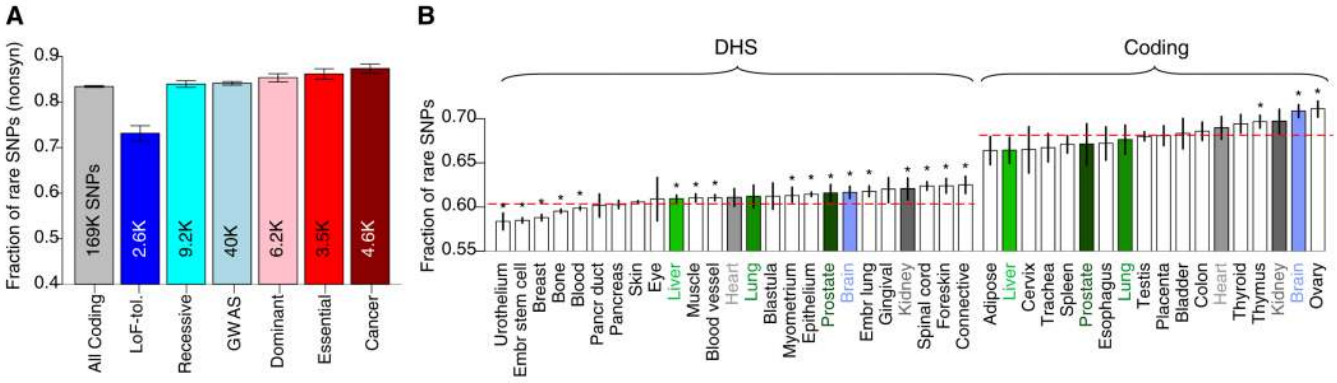

Fig. 1. Fraction of rare (DAF < 0.5%) SNPs

(**A**) In various gene categories. Total number of SNPs in each category shown. (**B**) In noncoding DHSs and coding genes, which show tissue-specific behavior. Matching tissues for which both DHS and gene expression data are available shown in same colors: shades of green for endodermal, gray for mesodermal, and blue for ectodermal origin of tissues. Red dotted lines show the total fraction for all DHSs and coding genes. Asterisks show significant depletion or enrichment after multiple-hypothesis correction. Error bars in both (A) and (B) denote 95% binomial confidence intervals.

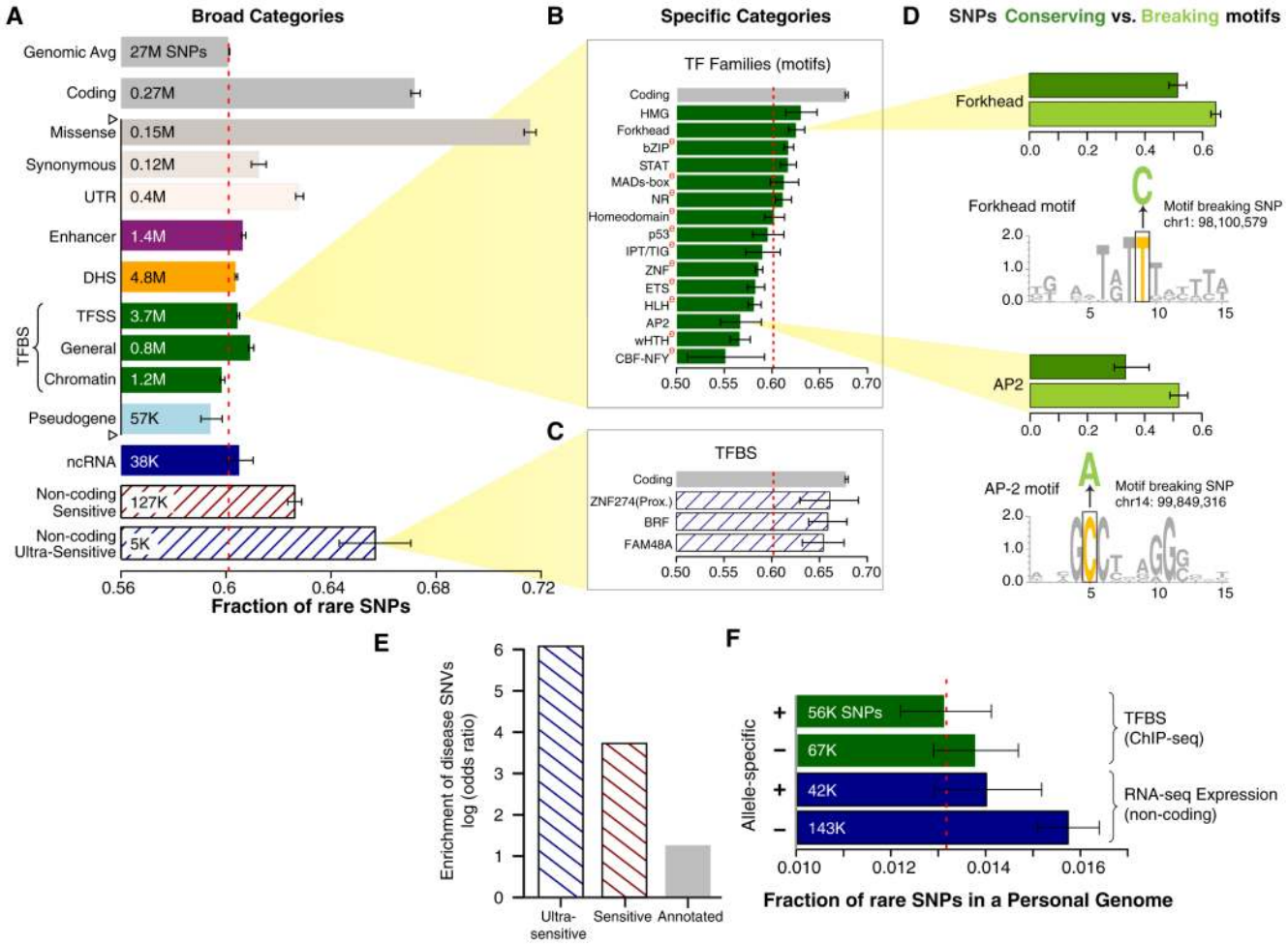

Fig. 2. Fraction of rare SNPs in noncoding categories

Red dotted lines represent genomic average. Error bars denote 95% binomial confidence intervals. Total numbers of SNPs in each category shown. (**A**) Broad categories. Ultrasensitive and sensitive regions are those under very strong negative selection. TFSS, sequence-specific TFs. Categories tested for enrichment of HighD sites (Fig. 5A) marked by using hollow triangles on the left. (**B**) Example of high-resolution categories: TFBS motifs separated into 15 families. e superscripts in red denote enrichment of eQTLs in TFBSs of specific families. (**C**) Examples of TFBSs included in ultrasensitive category. (**D**) SNPs breaking TF motifs show an excess of rare alleles compared with those conserving them. Representative motifs for two families are shown. (**E**) Enrichment of HGMD regulatory disease-causing mutations in ultrasensitive, sensitive, and annotated regions compared with all noncoding regions. (**F**) SNPs not exhibiting allele-specific behavior (−) are enriched in rare alleles compared with SNPs exhibiting allele-specific behavior (+).

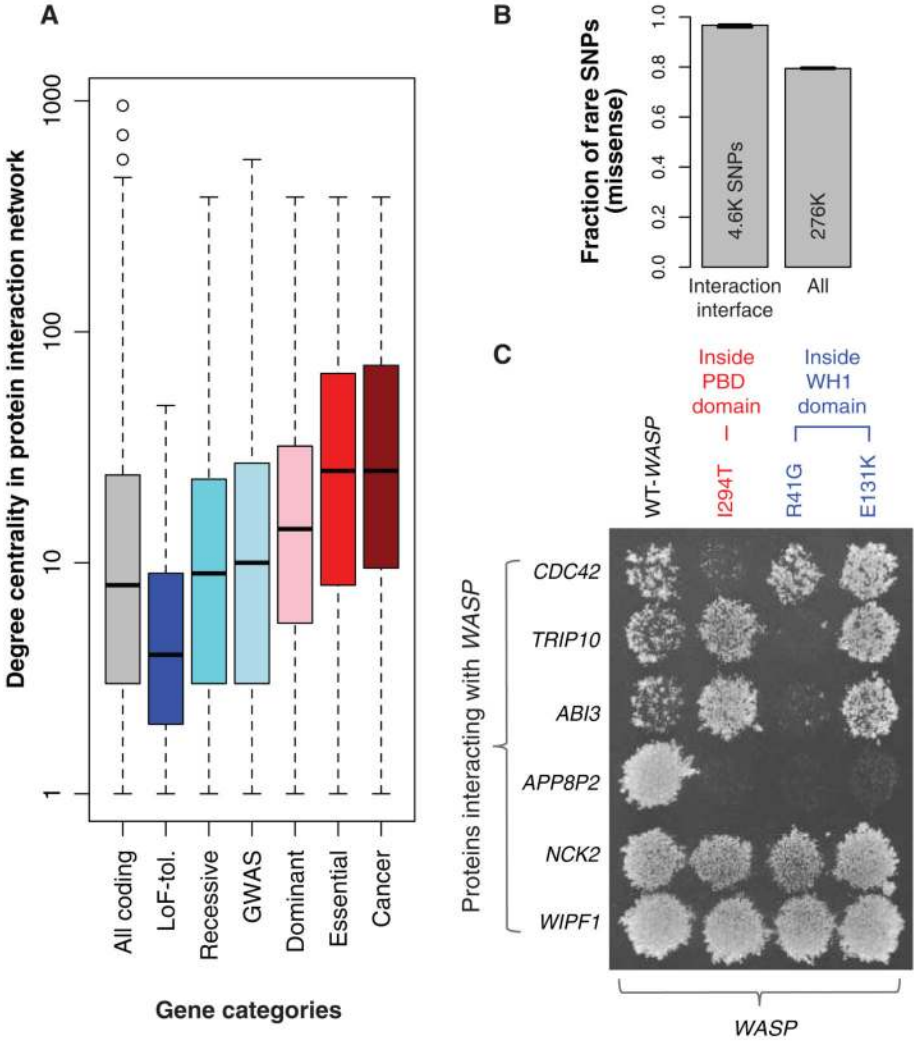

Fig. 3. SNPs in protein-protein interaction (PPI) network

(**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; total number of SNPs also shown). (**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, whereas each missense SNV disrupts its interaction with at least one protein.

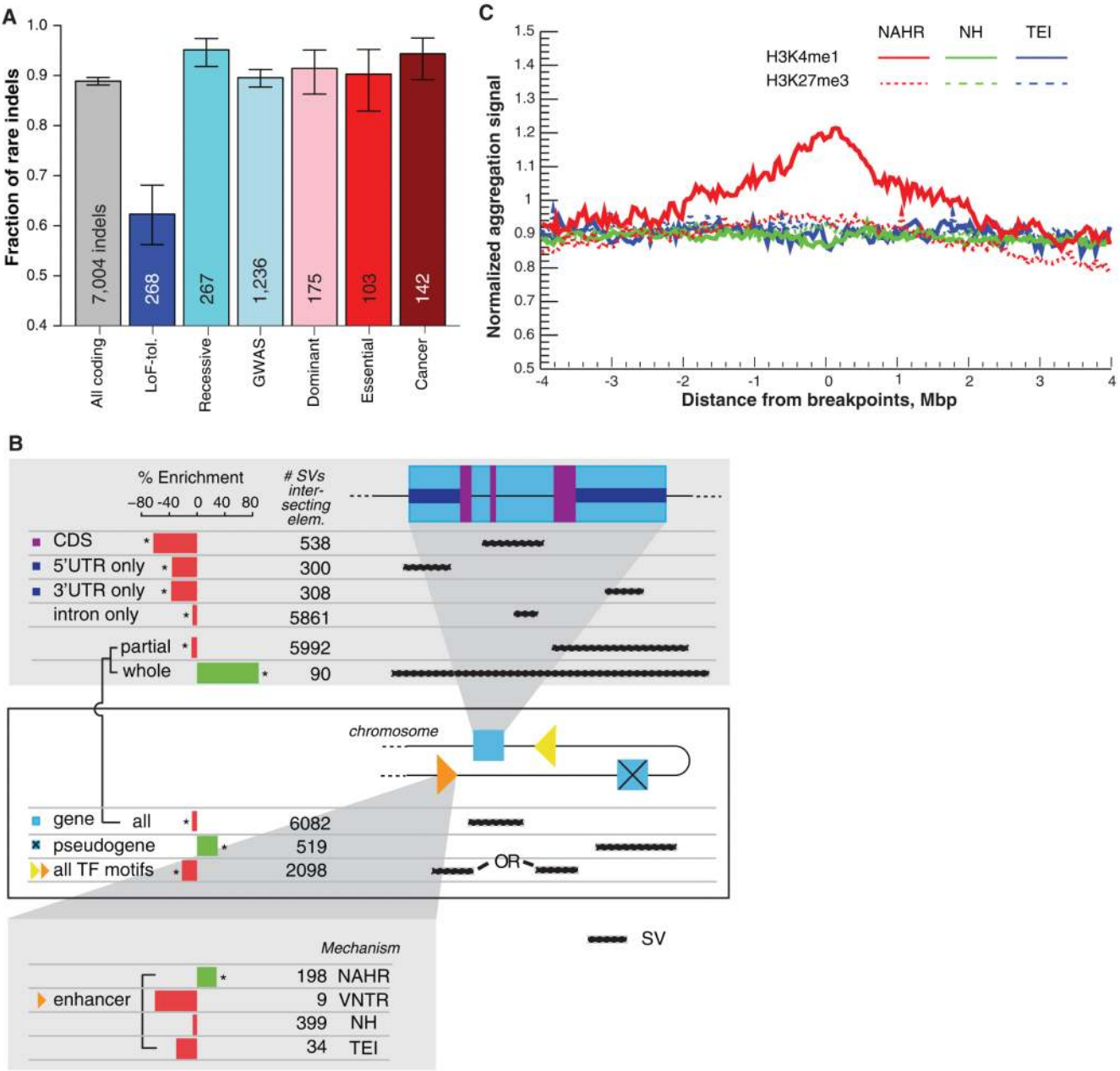

Fig. 4. Functional annotations of indels and SVs

(**A**) Fraction of rare indels in coding-gene categories. Total number of indels shown. (**B**) Enrichment of SVs affecting functional annotations. Middle box shows genes, pseudogenes, and TF motifs; upper blow-out shows gene parts in different modes, and bottom blow-out shows enhancers with different formation mechanisms, i.e., NAHR, NH (nonhomologous), TEI (transposable element insertion), and VNTR (variable number of tandem repeats). Asterisks indicate significant enrichment (green) or depletion (red) after multiple hypothesis correction. SVs intersecting various functional categories in different modes (e.g., whole/partial) are shown in the right-hand schematics. (**C**) Aggregation of histone signal around breakpoints of deletions formed by different mechanisms. Breakpoints centered at zero. Aggregation for upstream and downstream regions corresponds to negative and positive

distance, respectively. Signals for an activating histone mark (H3K4me1) and a repressive mark (H3K27me3) are shown.

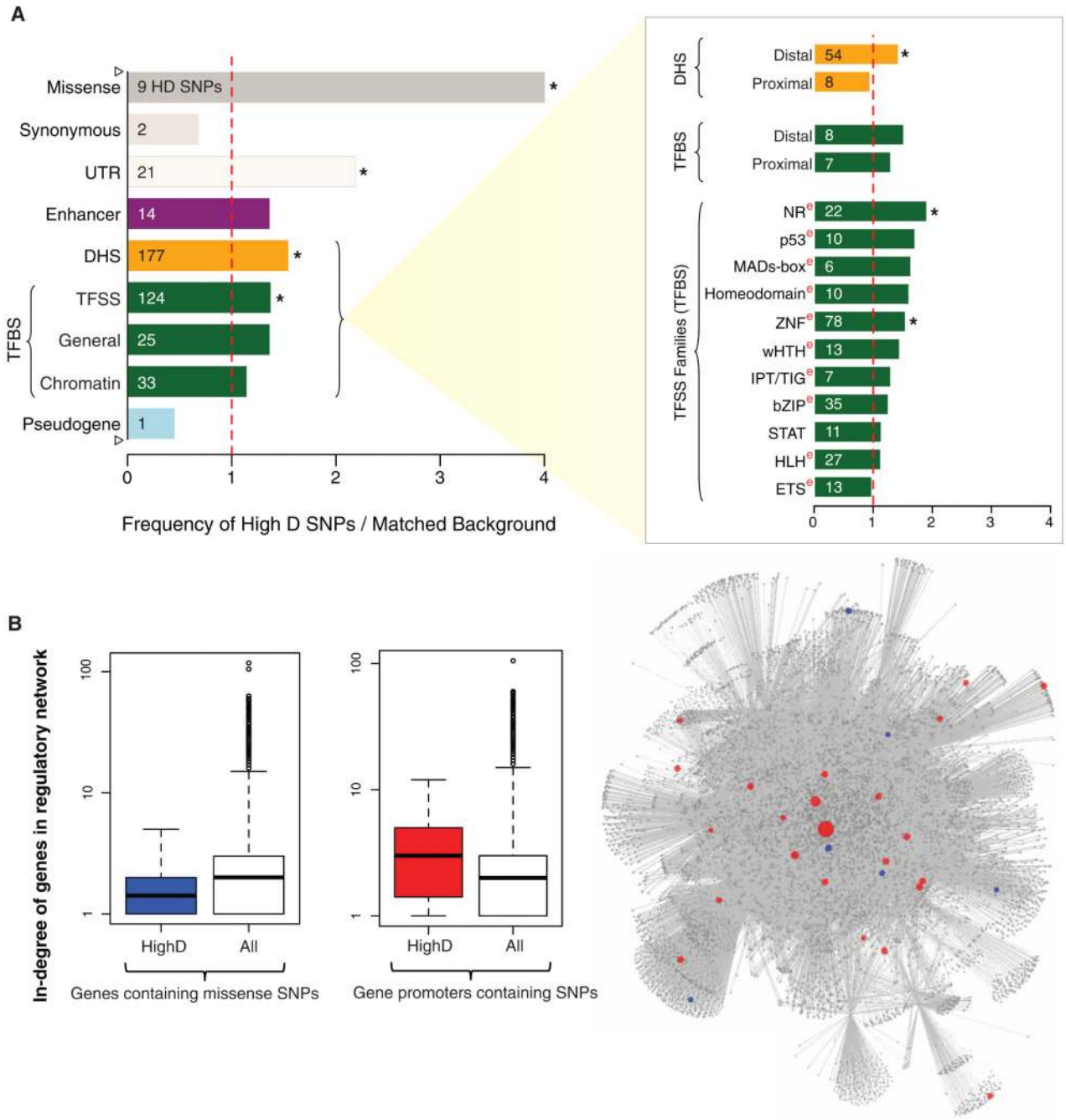

Fig. 5. Functional implications of positive selection

(**A**) (Left) Frequency of HighD SNPs versus matched sites for broad categories (marked by hollow triangles in Fig. 2A). (Right) Specific categories, e.g., specific TF families. Asterisk denotes significant enrichment after multiple-hypothesis correction. e superscripts in red denote the enrichment of eQTLs. (**B**) (Left) The in-degree of genes with HighD missense SNPs is lower than that of all genes. (Center) The in-degree of genes with HighD SNPs in their promoters is higher than all genes. (Right) The human regulatory network with edges in gray. Red nodes represent genes with HighD SNPs in their promoters, and blue nodes represent genes with HighD missense SNPs. Size of nodes scaled based on their degree

centrality. Nodes with higher centrality are bigger and tend to be in the center, whereas those with lower centrality are smaller and tend to be on the periphery.

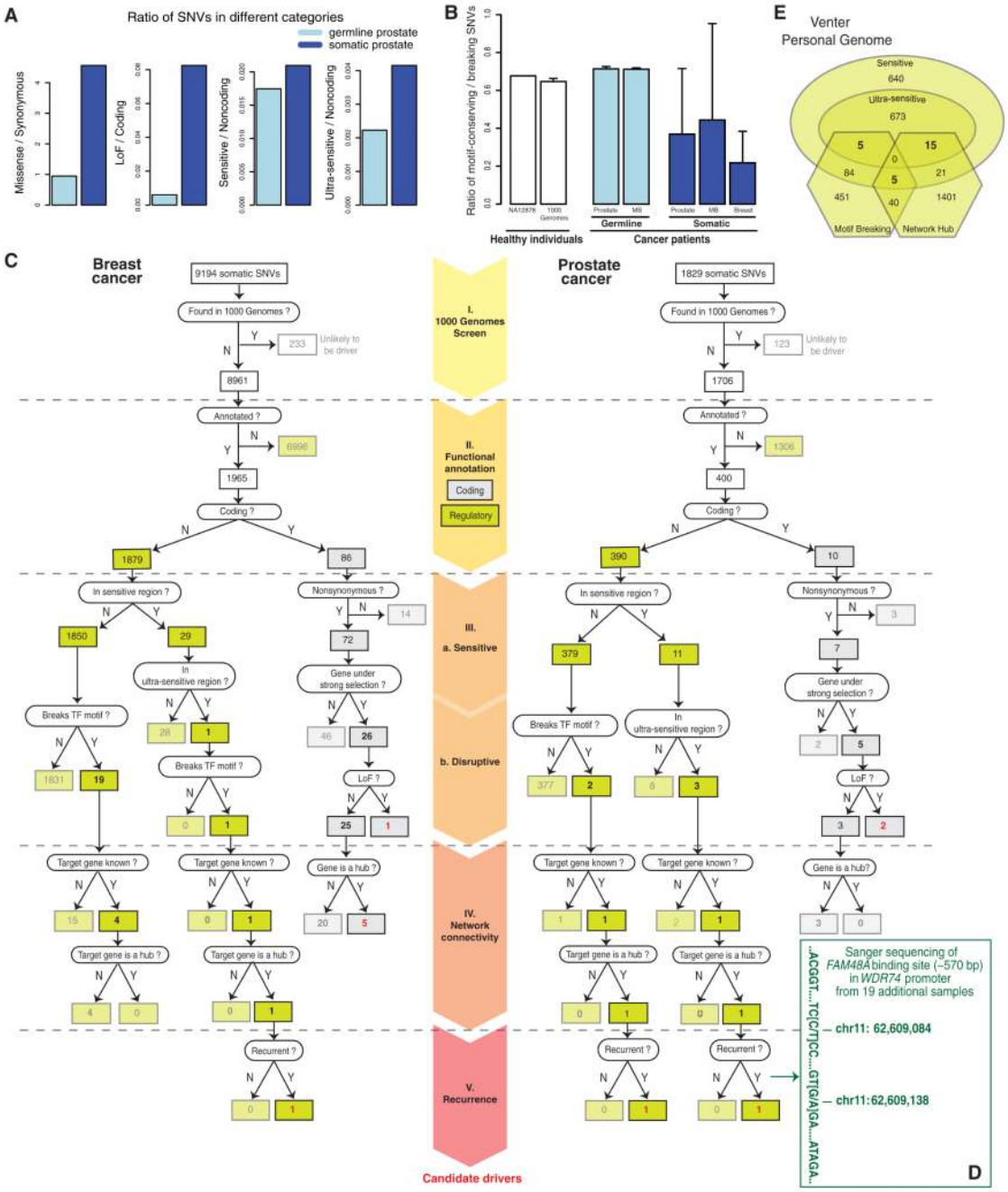

Fig. 6. Functional interpretation of disease variants

(**A**) Enrichment of functionally deleterious mutations among somatic compared with germline SNVs. Mean values from seven prostate cancer samples shown (variation shown in fig. S16). (**B**) Ratios for the number of SNVs that conserve versus break TF-binding motifs depicted for NA12878, the average of 1000 Genomes Phase I samples,and the average of somatic and germline samples from different cancers. Error bars represent 1 SD. MB, medulloblastoma. (**C**) Filtering of somatic variants from a breast (PD4006, left) and a prostate (PR-2832, right) cancer sample leading to identification of candidate drivers. (**D**) A part of the FAM48A binding site sequenced by Sanger sequencing in an independent cohort of 19 prostate cancer samples shown in green (with the coordinates of mutations observed in

one sample). (**E**) Application of variants filtering scheme to Venter personal genome. Number of SNVs in various categories shown.