A flexible framework to annotate and prioritize somatic variants from cancer whole-genome sequencing

Yao Fu[1], Zhu Liu[2], Shaoke Lou[3], Jason Bedford[1], Xinmeng Jasmine Mu[1,4], Kevin Y. Yip[3], Ekta Khurana[1,5, §], Mark Gerstein[1,5,6,§]

[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, 06520, United States of America
[2] School of Life Science, Fudan University, Shanghai, 200433, P.R. China
[3] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
[4] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[5] Molecular Biophysics and Biochemistry Department, Yale University, New Haven, Connecticut, 06520, United States of America
[6] Department of Computer Science, Yale University, New Haven, Connecticut, 06520, United States of America

[§]Corresponding author (Email: pi@gersteinlab.org or ekta.khurana@yale.edu)

Email addresses:

YF: yao.fu@yale.edu
ZL: 10300700070@fudan.edu.cn
SL: lousk@cuhk.edu.hk
JB: jason.bedford@yale.edu
XJM: xmu@broadinstitute.org
KY: kevinyip@cse.cuhk.edu.hk
EK: ekta.khurana@yale.edu
MG: pi@gersteinlab.org

[major changes are highlighted in RED]

Abstract [100 words…]

Somatic alterations in regulatory regions can cause tumorigenesis. We developed a flexible framework, integrating large-scale genomic and cancer resources, to annotate and prioritize potential regulatory cancer drivers. Our method has the ability to predict germline pathogenic and somatic deleterious variants. Applied to an individual cancer genome, our method could prioritize the *TERT* promoter mutation and provide the functional implications, such as gain-of a novel ETS motif. User-specific data, such as methylation and gene expression profiles, can be easily integrated. The framework would be useful for researchers to prioritize a few variants for further in-depth analysis to understand tumorigenic mechanism.

Keywords

Annotate, Prioritize, Noncoding driver, Somatic variants, Cancer

Background

Next generation sequencing usually identifies thousands of somatic alterations in individual cancer genomes. A few of them - called drivers, cause tumorigenesis, whereas the rest are passenger mutations accumulating during cancer progression. Systematic studies of human cancer genomes has discovered a wide range of cancer driver genes (1). However, mutations in noncoding genome are ignored in most cases. The important role of regulatory variants in various diseases has generated a great deal of interest in studying noncoding sequences (2-5). In particular, somatic mutations in telomerase reverse transcriptase (*TERT*) promoter have recently been implicated as cancer drivers (6-9). With the emerging number of cancer genomes being whole-genome sequenced, there is greater demand for high-throughput computation methods analyzing those variants. While several methods exist for identification of cancer driver genes (10-13), less effort has been invested in the investigation of noncoding drivers.

In contrast to coding variants, functional impact of noncoding variants is difficult to evaluate. Projects aiming to uncover potential regulatory sequences, such as The Encyclopedia of DNA Elements (ENCODE) (14) and sequence conservation studies (15,16), provide an unprecedented opportunity to interpret noncoding variants. Disease-associated single nucleotide polymorphisms (SNPs) identified by Genome-wide Association Studies (GWAS) are significantly enriched in ENCODE regions (17). A number of tools using ENCODE data to suggest most likely causal variants in linkage disequilibrium with associated SNPs or to annotate noncoding variants have been developed. These include Haploreg (18), RegulomeDB (19), ANNOVAR (20), GEMINI (21), FunciSNP (22) and VEP (23). Recently, two computational approaches – GWAVA and CADD were published to predict deleteriousness of variants genome-widely (24,25). The two methods utilize machine-learning models trained on potential pathogenic variants or nearly fixed/fixed human derived alleles to distinguish deleterious variants from neutral ones. Through analyzing variation patterns of inherited polymorphisms, we also reported a prototype approach to identify deleterious noncoding variants (16).

Here, we developed a flexible framework – FunXXX - to annotate and prioritize somatic cancer variants integrating various resources from genomic and cancer studies. It analyzes patterns of inherited polymorphisms among humans and evolutionary conservation across species to identify regions that are less likely to tolerate mutations; uses functional annotations from ENCODE and systems-level information from various biological networks; uses functional essentiality and prior knowledge of known cancer genes; predicts loss-of- and gain-of- function mutations for transcription-factor (TF) binding; associates distal regulatory regions with target genes using histone modifications; estimates recurrence of somatic alterations in publicly available cancer whole-genome sequencing data and developed a weighted scoring scheme based on natural polymorphisms to prioritize potential 'high-impact' variants. The framework consists of two modules – (1) a complex-to-regenerate data context generated by processing various data and (2) efficient and high-throughput variants prioritization run. Cancer genome sequencing usually couples with other experiments, such as RNA-Seq to quantify gene expression. We made our framework flexible for users to incorporate case-specific data, such as methylation and gene expression profiles from corresponding cancer samples.

We evaluated our framework using known germline pathogenic and somatic cancer variants. Our method has good prediction power for pathogenic regulatory variants, and more importantly it contains multiple modules that are specific for somatic variants prioritization. Applied to an individual cancer genome, our method is able to prioritize the *TERT* promoter mutation and provides a functional hypothesis of its potential impact. In general, our method can be directly used by researchers and clinicians to prioritize a few regulatory somatic variants for further studies.

Results and discussion

The framework first builds an organized data context processing various data resources and then annotates and prioritizes case-specific somatic variants, especially single nucleotide variations in noncoding regions. The workflow is depicted in Figure 1 and the detailed description is in Additional file 1.

Variants in ~~functional annotations~~ and conserved regions

We utilize functional annotations from ENCODE (transcription factor binding sites and the high-resolution motifs within them, enhancers, ncRNAs and DNase I hypersensitive sites) and conservation data from different resources – across-species conservation from GERP scores (26) and ultra-conserved elements (15) as well as population-level conservation from 1000 Genomes (16,27) to detect likely deleterious variants. Each variant will be annotated with corresponding functional annotations and conservation data. We also implemented the method used in *Khurana et al.,* (16) to a pipeline for users to find novel population-level conserved regions using depletion of common variants with user input polymorphism or annotation data (Additional file 1).

High-impact variants in motifs: Nucleotide resolution effect

Loss-of-function variants occurred in transcription factor binding motifs are more likely to cause deleterious impact (16,28,29). Variants decreasing the position weight matrix (PWM) scores could potentially alter the binding strength of transcription

factors, or even eliminate the binding. The framework consists of a module to detect motif-breaking events – defined as variants decreasing PWMs (Material and Methods). Meanwhile, gain of new binding sites caused by somatic mutations can constitute driver events (6-9). However, an automated tool to detect such events in whole tumor genomes is not available. We developed a gain-of-motif scheme to scan and statistically evaluate (30) all possible motifs created by variants compared to reference alleles. For each variant, we concatenate it with +/- 29bp reference sequences and calculate sequence score for each possible motif against the PWMs. Gain-of-motif events are identified when sequence score with mutated allele is significantly higher than the background ($p < 4e-8$), whereas that with reference allele is not. Our scheme is validated by the detection of ETS motifs created by the two cancer driver mutations in *TERT* promoter (Additional file 1: Table S2).

Correlating histone marks with gene-expression data to identify likely target genes of distal regulatory elements

To interpret likely functional consequences of noncoding variants, we comprehensively define associations between regulatory elements and genes through correlating various epigenetic modifications with expression levels of genes. We considered the enhancer marks H3K4me1 and H3K27ac as two types of activity signals, and DNA methylation as an inactivity signal. Using ChIP-Seq and RNA-Seq data from the Roadmap Epigenomics Mapping Consortium, for each regulatory element-candidate target gene pair, we computed the correlations of H3K4me1 and H3K27ac and anti-correlations of DNA methylation at the regulatory element with gene expression levels across ~20 tissue types (Material and Methods). In total, we identified ~769K distal regulatory elements significantly associated with ~17K genes. All noncoding variants in these regulatory elements could be associated with potential target genes. To incorporate the ever-increasing amounts of genomic data, FunSVPT offers a pipeline for users to extend the data context with their own data, for example, users can input annotation regions or chromatin marks to find novel associations between regulatory elements and coding genes (Additional file 1).

Network analysis of variants associated with genes

Disruption of highly connected genes or their regulatory elements is more likely to be deleterious (16,31). Cancer genes tend to have higher centralities than essential genes in biological networks (16). We use the regulatory element-target gene pairs to connect noncoding variants into a variety of networks. For each noncoding variant, we calculate scaled network centrality (the percentile after ordering centralities of all genes in a particular network) of the associated gene in various networks (Material and Methods). Amongst the different networks, we use the maximum as network disruptive measure of the variant. We make the scheme flexible so it can integrate user networks in addition to the pre-collected networks such as protein-protein interaction, regulatory and phosphorylation networks (16,32,33).

Gene prioritization: using expression and prior knowledge of target genes

Interpretation of the functional impact of variants can be greatly enhanced if the function of its target protein-coding genes is known. We incorporates prior knowledge of genes, such as known cancer-driver genes (1,34), genes involved in DNA repair (35) and actionable genes (‘druggable’ genes) (36) to annotate variants that are potentially involved in cancer or could be used as drug targets. In addition, user-specific gene lists can be easily inputted (Additional file 1). Variants in regulatory

sequences may disrupt the expression of coding genes. We provide a "differential gene expression analysis" module to detect differentially expressed genes in cancer samples (relative to matched normal) from RNA-Seq data. Differentially expressed gene list will be generated and used to annotate variants, as differential expression of target genes in cancer samples is an indication of potential effect of noncoding variants.

Recurrence database from whole-genome sequencing

One important approach to identify cancer driver genes is to examine their recurrence across multiple samples. We extended the concept to noncoding regulatory elements. Our method can detect recurrent same-site mutations, genes and regulatory elements from multiple cancer samples.

With the increasing number of cancer samples being whole-genome sequenced, we are able to study the recurrence pattern in regulatory sequences. Similar to the cancer recurrent gene database in cBio (37), we developed the recurrence database (coding genes, noncoding elements and same-site mutations) for whole-genome sequencing data. Currently, we have collected somatic mutations from 570 samples of 10 cancer types (38-40). For each cancer type, loci or sites with recurrent mutations in at least two samples are identified with our framework (Table 1). We also incorporated recurrent somatic noncoding mutations from COSMIC (41) into our database (Material and Methods). Variants in user-input tumor genome are compared to the recurrence database and the results in different cancer types are reported in the output. The database will be updated with newly available dataset.

Weighted scoring scheme to prioritize variants [details are moved to material and methods…]

All of the above features are used to annotate and score variants. In general, features can be classified into two classes - discrete and continuous (Figure 2). Discrete features are binary, such as in ultra-conserved elements or not. For continuous features, taking 'motif-breaking score' as an example, the values are the changes in PWMs.

We developed a weighted scoring scheme, based on the mutation patterns observed in the 1000 Genomes polymorphisms, to integrate features (Material and Methods, Additional file 1). Features that are frequently observed in polymorphisms are less likely to contribute to the deleteriousness of variants and thus are weighted less. We use information content to denote the relative importance of each feature. For each cancer variant, we score it by summing up the information contents of all its features (details in Material and methods). Variants ranked on top of the output are those with higher scores and are most likely to be deleterious.

Application to regulatory cancer somatic variants and germline pathogenic variants

We applied our method to predict functional impact of cancer somatic variants. Considering only two regulatory variants have been confirmed to be drivers, we use recurrence to proxy the deleteriousness of somatic variants. Recurrence is considered as one potential sign of positive selection amongst tumors and is more likely to constitute driver events. We examined recurrence from two different perspectives – recurrence at exact same-site and recurrence in same regulatory element. We obtained

regulatory somatic variants from COSMIC (41) and classified them as same-site recurrent or non-recurrent (Material and Methods) (25). Our method scored recurrent variants higher than non-recurrent ones (Wilcoxon rank-sum test: p-value < 2.2 e-16; Figure 3A). Variants occurring in more than 2 samples have higher scores than those in 2 samples. Data quality is one of our concerns with COSMIC data. As shown in Figure S2, percentage of variants in pseudogenes increases as the number of recurrent samples increases. We suspect that those variants should probably be mapped to parent genes of pseudogenes, instead of noncoding genome. Considering potential technical or mapping errors in these cancer studies, targeted sequencing is needed to confirm the existence of variants. Next we evaluated variants in recurrent regulatory elements. We ran our pipeline on 119 breast cancer samples (38), and classified variants as occurring in recurrent elements or not (Material and Methods). We found that variants in recurrent elements get significantly higher scores (Wilcoxon rank-sum test: p-value < 2.2e-16) (Figure 3B) than variants elsewhere. Similar patterns are observed with other cancer types (Additional file 1: Figure S3). In summary, our method could predict potential 'high-impact' somatic cancer variants. Moreover, it provides functional implications of corresponding variants. Considering the pervasive cancer molecular subtypes, our method has the ability to detect non-recurrent deleterious variant in each cancer sample.

Disease studies have discovered a number of regulatory pathogenic variants. We also evaluated the ability of our method to predict those germline deleterious variants. We obtained pathogenic regulatory variants from HGMD (42) and three sets of controls from *Ritchie et al* (25) – 'unmatched', 'matched TSS' and 'matched region' (Material and Methods). Our method scored HGMD variants higher than controls, with AUC scores - 0.62, 0.73, 0.86, respectively (Figure 3C and 3D). We compared our results with CADD using the same dataset (24) (Additional file 1: Figure S4). As negative sets are much larger than positive set, one concern with AUC scores is that the prediction power may come from the ability to predict negatives instead of positives. Thus we examined precision and recall to capture method ability to predict positives (Additional file 1: Figure S5). Generally speaking, our method has good prediction power for pathogenic regulatory variants. In addition, GWAS SNPs show higher scores than matched controls with our method (mean values: 0.41 vs. 0.34, p-value < 2.2e-16) (Material and Methods, Additional file 1: Figure S6).

Application to somatic variants from an individual cancer genome

High recurrence of somatic mutations in *TERT* promoter implicates their important roles in tumorigenesis. Among the 570 cancer samples we collected, 7 samples contain the *TERT* promoter mutation (chr5: 1295228). We use one Medulloblastoma sample as an example to prioritize regulatory variants from whole-genome sequencing. Amongst 2,183 somatic single nucleotide variants, our method ranked the *TERT* promoter mutation in top 0.64% (14^{th}). When taking into account recurrence across 100 Medulloblastoma samples, this mutation ranked the 2^{nd}. On the contrary, CADD ranked it as 224^{th} (10.3%) and GWAVA ranked it as 10^{th} (0.46%), 25^{th} (1.15%) and 129^{th} (5.92%) with 'unmatched', 'matched TSS' and 'matched region' models, respectively. Detailed analysis of GWAVA 'unmatched' model, we found that the high ranking of this mutation is not due to functional importance, but model bias caused by distance to TSS (Additional file 1: Figure S7). Both our method and CADD can predict deleterious variants distant to TSS.

Our method is the only one that could capture the potential functional impact of this variant. As shown in Table 2, this mutation occurs in ENCODE regulatory regions, creates a novel ETS binding motif and potentially affects a highly connected and known cancer gene –*TERT*. Our method also contains several cancer-specific features, such as filtering natural polymorphisms, detecting differentially expressed genes in cancer samples and recurrence database. Besides DNA sequences, epigenomics or open chromatin landscape could also be altered in cancer genomes. These data provide sample-specific activation or de-activation signatures of regulatory sequences; our framework is flexible in integrating those data into our annotation scheme (refer to Additional file 1 for details).

Output format and performance

FunSVPT is a Linux/Unix based tool with a web-server available at funseq2.gersteinlab.org. The code is also posted under GitHub - http://gersteinlab.github.io/FunSVPT/. It takes VCF or BED formatted cancer variants and generates results in either BED or VCF format (refer to Additional file 1 for examples). Users can retrieve or visualize results in concise tables through the web interface (Additional file 1: Figure S7, 8).

FunSVPT runs in a tiered fashion. Building data context from bulk of data resources is time-consuming. Currently FunSVPT takes about one week (on ~20 4-core 3.00-Ghz 16GB RAM PowerEdge 1955 nodes) to rebuild the data context based on pre-processed genomics data, such as ENCODE peak calls. The data context will be updated regularly to keep it up-to-date. Users can input additional data to customize the data context upon the existing one. Variant prioritization step is quite efficient. It takes ~2-3 mins to prioritize one genome with thousands of variants on a QEMU Virtual CPU version (cpu64-rhel6) @ 2.24-GHz 1 processor Linux PC with 20GB RAM, and a 500 GB local disk. Time comparison with CADD and GWAVA is in Additional file 1: Table S6 (FunSVPT is two times faster with equal number of variants). In addition, FunSVPT implements parallel-processing fork manager for efficient memory utilization to tackle multiple genomes in a single run. With a flexible and modularized structure, researchers can restructure the pipeline to incorporate more data and new features.

Conclusions

We have developed a method integrating various genomic and cancer resources to prioritize cancer somatic variants, especially noncoding ones. User data can be easily integrated into the framework. We believe that the software would be useful for researchers to identify a few somatic events among thousands for further in-depth analysis to understand the mechanisms underlying oncogenesis.

Materials and Methods

Data resources

We collects polymorphisms from 1000 Genomes Project Phase 1 (27), GERP scores and ultra-conserved elements from (15,26), sensitive/ultra-sensitive regions from (16), functional genomics data from ENCODE (14) and histone modifications and RNA-Seq data of 20 cell-lines from REMC (43). Cancer driver genes are the union of genes

from *Vogelstein et al*., cancer gene consensus and COSMIC (1,34). Actionable genes are from (36). Binary protein-protein interaction network is from InWeb (44) and HINT (45). Regulatory and phosphorylation networks are obtained from *Gerstein et al.,* (32) and *Lin et al.,* (33) respectively. Recurrent database uses somatic variants of 506 cancer genomes from *Alexandrov et al.,* (38) and 64 prostate cancer samples from (39,40).

High-impact variants in motifs: Nucleotide resolution effects

1. Motif breaking events

When variants hit transcription factor binding motifs under ENCODE Chip-Seq peaks, we examine their motif breaking or conserving effect using position weight matrixes (PWM). Motif-breaking events are defined as variants decreasing the PWM scores, whereas motif-conserving events are those that do not change or increase the PWM scores (29) (we calculate the difference between mutated and reference alleles in the PWMs). Variants causing motif-breaking events are reported in the output together with the corresponding PWM changes. Transcription factor PWMs are obtained from ENCODE project (14), including some of TRANSFAC, JASPAR motifs.

2. Motif gaining

We developed an automated module to detect gain-of-motif events. Whole genome motif scanning generally discovers millions of motifs, of which, a large fraction are false positives. To restrict motif scanning, we focused on variants occurred in promoters (defined as -2.5kb from transcription starting sites) or regulatory elements associated with genes. For each variant, +/- 29bp are concatenated from human reference genome (motif length is generally <30bp). For each PWM, we scan the 59bp sequence. For each candidate motif encompassing the variant, we evaluate the sequence score with mutated allele using TFM-Pvalue (30) (with respect to the PWM). Sequence score is computed by summing up the corresponding values at each position in the PWM. If the p-value with mutated allele <= 4e-8 and the p-value with reference allele > 4e-8, we define the variant creating a novel motif. The process is repeated for all PWMs and all variants.

Correlating histone marks with gene-expression data to identify likely target genes of distal regulatory elements

1. Definition of distal regulatory modules (DRMs)

We started with a list of regulatory regions from three different types, namely transcription factor binding peaks (TFP), DNase hypersensitive sites (DHS) and Segway/ChromHMM-predicted enhancers. All regulatory regions at least 1kb from the closest gene according to the Gencode v7 annotation (46) were defined as a distal regulatory module (DRM).

2. Identifying potential regulatory targets of each DRMs

We grouped different transcripts of a gene sharing the same transcription start site as a transcription start site expression unit (tssEU). For each DRM, we first considered all tssEUs within 1Mb from it as its candidate targets. We then correlated some activity/inactivity signals at a DRM and the expression of its candidate target tssEUs, and called the ones with significant correlation values as potential DRM-target pairs as follows.

At the DRMs, we considered the enhancer marks H3K4me1 and H3K27ac as two

types of activity signals, and DNA methylation as an inactivity signal. The activity level of each DRM was defined as the number of sequencing reads aligned to the DRM from the corresponding ChIP-seq experiments. The methylation level of a DRM was defined as follows. For each CpG site i within a DRM, we counted the number of reads that support the methylation of it (m_i), and the total number of reads covering it (n_i). The methylation level of the DRM was then defined as the ratio of their sums across all CpG sites in the DRM, $\frac{\sum_i m_i}{\sum_i n_i}$. For each tssEU, we defined its expression level as the number of RNA-seq reads aligned to the [TSS-50, TSS+50] window. Both the activity signal levels and gene expression levels were normalized by the total reads, then multiplied by one million to keep them within an easily readable range of values.

We collected all bisulfite sequencing, ChIP-Seq and RNA-Seq data from the Roadmap Epigenomics project website (43) (EDACC release 9[1]). We considered 19 tissue types with data for both the activity signals and gene expression, and 20 tissue types with data for both the inactivity signal and gene expression. For RNA-seq, we used the paired-end 100bp Poly-A enriched data sets. For experiments with replicates, we used the mean value across the replicates as the expression level of a gene.

For each DRM-candidate target pair, we computed the correlations of their activity/inactivity and expression levels across the different tissue types. We computed both value-based Pearson correlation and rank-based Spearman correlation. The statistical significance of each correlation value was evaluated by computing a p-value based on one-tailed tests using the built-in functions in R. Briefly, for Pearson correlation, the correlation values would follow a t distribution with $n-2$ degrees of freedom (where n is the number of tissue types) if the samples were drawn independently from normal distributions. The Fisher's Z transformation was used to compute the p-values. For Spearman correlation, the p-value was computed based on a procedure proposed by Hollander and Wolfe (47). For activity signals, we considered the right tail, which means we looked for correlations significantly more positive than would be expected by chance. For inactivity signals, we considered the left tail, which means we looked for correlations significantly more negative (i.e., strong anti-correlations) than would be expected by chance. All p-values were then adjusted for multiple hypotheses testing using the Bonferroni, Holm, Benjamini-Hochberg (BH) or Benjamini-Yekutieli (BY) methods.

Differential gene expression analysis

We incorporate a module to detect differentially expressed genes in cancer samples (relative to matched normal) from RNA-Seq data. When provided with gene expression files, our module calls NOISeq (48) when having RPKMs and DESeq (49) with raw read counts (from reads-mapping tools) to detect differentially expressed genes. Genes that are up- or down- regulated with FDR < 0.05 (with biological replicates) and FDR < 0.1 (without replicates) in cancer samples are identified and annotated in the output.

Network analysis of variants associated with genes

For each variant associated with genes, we examine their network properties in various networks. For each network, we calculate the centrality position (cumulative probability after ordering centralities of all genes increasingly) of the associated gene.

If one variant is associated with multiple genes or the associated gene participates in multiple networks, the maximum cumulative probability is used as the continuous value for centrality score.

Recurrence database from whole-genome sequencing

We used somatic variants from 570 samples of 10 cancer types to create the recurrence database. For each cancer type, recurrent genes, regulatory elements and mutations detected are stored as entries in the database. We also collected recurrent somatic regulatory variants from COSMIC (version 68). Recurrent variants are defined as identified in whole-genome sequencing data and observed in at least 2 samples.

Weighted scoring scheme

For more details, please refer to Additional file 1. For noncoding variants, we developed a weighted scoring scheme. We weight each feature based on the mutation patterns observed in the 1000 Genomes polymorphisms. We randomly selected 10% of the 1000 Genomes Phase 1 SNP (~3.7M) and run through our tool. For each discrete feature d, we calculate the probability p_d that it is observed in natural polymorphisms. Then we compute 1-Shannon entropy as its weighted value w_d *(1)*.

$$w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d) \quad (1)$$

$$p_d = \frac{number\ of\ polymorphisms\ with\ feature\ d}{total\ number\ of\ polymorphisms}$$

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in polymorphisms. Thus, one weight cannot suffice for varied feature values. For a continuous feature c, which is associated with a score v_c (e.g. motif-breaking score), we calculate feature weights for each v_c. In particular, we discretize at each v_c and compute 1-Shannon entropy using *(2)*. Then we fit a smooth curve for all v_c to obtain continuous $w_c^{v_c}$. Now, when we come to evaluate the continuous feature c for a particular variant, we calculate its weighted value (on the curve) using the actual v_c corresponding to the variant.

$$w_c^{v_c} = 1 + p_c^{\geq v_c} log_2 p_c^{\geq v_c} + \left(1 - p_c^{\geq v_c}\right) log_2 \left(1 - p_c^{\geq v_c}\right) \quad (2)$$

$$p_c^{\geq v_c} = \frac{number\ of\ polymorphisms\ with\ score \geq v_c\ for\ feature\ c}{total\ number\ of\ polymorphisms}$$

Taking 'motif-breaking score' as an example (Figure 2), for each score v, we calculated the probability of observing motif-breaking scores $\geq v$ in polymorphism data, then used *(2)* to fit the smooth function. We used 'nls' function in R to fit curves.

Finally, for each cancer variant, we score it by summing up the weighted values of all its features *(3)*.

$$score = \sum_d w_d + \sum_c w_c^{v_c} \quad (3)$$

In addition, we considered some of the feature dependencies when calculating the sum-up scores (described in details in Additional file 1).

Application to regulatory pathogenic and somatic cancer variants

1. HGMD and three sets of controls

Genome locations of pathogenic regulatory variants (HGMD) and negative controls are downloaded from GWAVA (25). The three control sets – 'unmatched', 'matched TSS' and 'matched region', contain regulatory polymorphisms from 1000 Genomes with minor allele frequency ≥ 1%. 'Unmatched' control has polymorphisms randomly selected from 1000 Genomes. 'Matched TSS' control only has polymorphisms within 2kb upstream of TSS. 'Matched region' control has polymorphisms within 1kb around HGMD regulatory variants. Allele information for these variants is obtained from ENSEMBL BioMart.

2. Noncoding somatic recurrent variants

We obtained noncoding somatic variants form COSMIC (version 68). Recurrent variants (10,041) are defined as identified in whole-genome sequencing data and observed in at least 2 samples. All other variants (1,311,389) are non-recurrent ones.

3. Noncoding somatic variants in recurrent regulatory elements

We first identified recurrent regulatory elements across multiple cancer samples. Then we classified variants either in recurrent regulatory elements or not. As recurrent regulatory elements are functional annotations, to be a fair comparison, we filtered variants in non-recurrent regulatory elements as those also with functional annotations. For example, from 119 breast cancer samples, there are 24,443 and 126,217 variants in recurrent and non-recurrent regulatory elements respectively. Recurrence feature is not added in calculating scores for recurrent variants.

4. Prediction power calculation

For each score threshold, we calculated TPR (true positive rate) = TP/(TP+FN); FPR (false positive rate) = FP/(FP+TN); Precision = TP/(TP+FP); Recall = TP/(TP+FN). TP: true positive; FP: false positive; TN: true negative; FN: false negative.

List of abbreviations

ENCODE: The Encyclopedia of DNA Elements; TF: transcription factor; PWM: position weight matrix; REMC: roadmap epigenomics mapping consortium; HGMD: the human gene mutation database; GWAS: genome-wide association studies; TSS: transcription starting site; TERT: Telomerase reverse transcriptase; SNP: single nucleotide polymorphisms; COSMIC: Catalogue of Somatic Mutations in Cancer.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YF, EK and MG designed the study and drafted the manuscript. ZL developed the web server. JB participated in differential gene expression analysis. SL and KY carried out studies associating regulatory elements with target genes. XJM participated in transcription factor binding motif analysis. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Institutes of Health, A L Williams Professorship and in part by the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center [Grant Number RR029676-01]. JB acknowledges support from the National Science Foundation - Graduate Research Fellowship Program [Grant Number 1346837]. Funding for open access charge: National Institutes of Health.

References

1. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nature reviews. Cancer*, **4**, 177-183.
2. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190-1195.
3. Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H. *et al.* (2013) Identifying recent adaptations in large-scale genomic data. *Cell*, **152**, 703-713.
4. Sakabe, N.J., Savic, D. and Nobrega, M.A. (2012) Transcriptional enhancers in development and disease. *Genome biology*, **13**, 238.
5. Ward, L.D. and Kellis, M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, **30**, 1095-1106.
6. Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L. and Garraway, L.A. (2013) Highly recurrent TERT promoter mutations in human melanoma. *Science*, **339**, 957-959.
7. Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K. *et al.* (2013) TERT promoter mutations in familial and sporadic melanoma. *Science*, **339**, 959-961.
8. Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L.A., Jr., Friedman, A.H., Friedman, H., Gallia, G.L., Giovanella, B.C. *et al.* (2013) TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 6021-6026.
9. Vinagre, J., Almeida, A., Populo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L. *et al.* (2013) Frequency of TERT promoter mutations in human cancers. *Nature communications*, **4**, 2185.
10. Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome research*, **22**, 1589-1598.

11. Reimand, J. and Bader, G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular systems biology*, **9**, 637.
12. Tamborero, D., Gonzalez-Perez, A. and Lopez-Bigas, N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238-2244.
13. Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M.S., Getz, G., Bader, G.D., Ding, L. *et al.* (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, **3**, 2650.
14. The Encode Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
15. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321-1325.
16. Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A. *et al.* (2013) Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science*, **342**, 1235587.
17. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. and Snyder, M. (2012) Linking disease associations with regulatory information in the human genome. *Genome research*, **22**, 1748-1759.
18. Ward, L.D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, **40**, D930-934.
19. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, **22**, 1790-1797.
20. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, **38**, e164.
21. Paila, U., Chapman, B.A., Kirchner, R. and Quinlan, A.R. (2013) GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS computational biology*, **9**, e1003153.
22. Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A. and Noushmehr, H. (2012) FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic acids research*, **40**, e139.
23. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. and Cunningham, F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069-2070.
24. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*.
25. Ritchie, G.R., Dunham, I., Zeggini, E. and Flicek, P. (2014) Functional annotation of noncoding sequence variants. *Nature methods*.
26. Cooper, G.M., Stone, E.A., Asimenos, G., Program, N.C.S., Green, E.D., Batzoglou, S. and Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, **15**, 901-913.

27. The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56-65.
28. Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S. and Kellis, M. (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome research*, **23**, 800-811.
29. Mu, X.J., Lu, Z.J., Kong, Y., Lam, H.Y. and Gerstein, M.B. (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic acids research*, **39**, 7058-7076.
30. Touzet, H. and Varre, J.S. (2007) Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms for molecular biology : AMB*, **2**, 15.
31. Khurana, E., Fu, Y., Chen, J. and Gerstein, M. (2013) Interpretation of genomic variants using a unified biological network approach. *PLoS computational biology*, **9**, e1002886.
32. Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91-100.
33. Lin, J., Xie, Z., Zhu, H. and Qian, J. (2010) Understanding protein phosphorylation on a systems level. *Briefings in functional genomics*, **9**, 32-42.
34. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr. and Kinzler, K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546-1558.
35. Ruark, E., Snape, K., Humburg, P., Loveday, C., Bajrami, I., Brough, R., Rodrigues, D.N., Renwick, A., Seal, S., Ramsay, E. *et al.* (2013) Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature*, **493**, 406-410.
36. Wagle, N., Berger, M.F., Davis, M.J., Blumenstiel, B., Defelice, M., Pochanard, P., Ducar, M., Van Hummelen, P., Macconaill, L.E., Hahn, W.C. *et al.* (2012) High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer discovery*, **2**, 82-93.
37. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, **2**, 401-404.
38. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415-421.
39. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M. *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666-677.
40. Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C. *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature*, **470**, 214-220.

41. Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research*, **39**, D945-950.
42. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M. and Cooper, D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Human mutation*, **21**, 577-581.
43. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**, 1045-1048.
44. Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tumer, Z., Pociot, F., Tommerup, N. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, **25**, 309-316.
45. Das, J. and Yu, H. (2012) HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, **6**, 92.
46. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, **22**, 1760-1774.
47. Wolfe, M.H.a.D.A. (1973) *John Wiley and Sons*, pages 185–194.
48. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome research*, **21**, 2213-2223.
49. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology*, **11**, R106.

Figures

Figure 1 - Schematic workflow of FunSVPT.

Figure 2 - Weighted scoring scheme.

A) Features used to score variants; B) Motif-breaking scores and corresponding weighted values.

Figure 3 - Application to pathogenic and cancer somatic noncoding variants.

A) Prediction scores of regulatory variants from HGMD and controls; B) ROC curves comparing HGMD with controls; C) Score distribution of variants based on their recurrence in COSMIC; D) Score distribution of variants based on recurrent regulatory elements in Breast cancer samples.

Tables

Table 1 - Summary of recurrence database.

Table 2 - Output for TERT promoter mutation in an Medulloblastoma sample.

Additional files

Additional file 1 – Supplementary information

This file contains detailed material and methods, supplementary figures, supplementary tables and software manual.

Table 1

Cancer Type	# Samples	# Somatic Mutations (SNV)	# Recurrent Genes/Elements/Mutations
AML	7	271~1068	1
Breast	119	1043~67347	69,140
CLL	28	522~3338	709
Liver	88	1348~25131	74,144
Lung Adeno	24	9284~297569	162,165
Lymphoma B cell	24	1502~37848	4,233
Medulloblastoma	100	44~47440	2,793
Pancreas	15	1096~14998	2,591
Pilocytic Astrocytoma	101	2~926	58
Prostate	64	1430~18225	36,327
COSMIC recurrent regulatory mutations	-	-	10,041

Table 2

Associated gene	Network	Recurrence in samples	Recurrence database	Score
TERT (promoter) [Cancer gene]	Protein-Protein Interaction Centrality: 0.798	2/100 Medulloblastoma samples	5/88 Liver samples; 54 COSMIC samples	2.6923

Variant	GERP	Functional annotations	Gain of motif
chr5: 1295228 G -> A	-1.46	DHS, Enhancer, TFP (E2F6, EGR1, ELF1, GABPA, HDAC2, MAX, MYC, SIN3A, TCF12, USF1, ZBTB7A, ZEB1)	Motif: Ets_known10 Position: 1295223 – 1295229 Strand: + Score: 1.893 -> 5.743

Figures (PDF)

Figure legends

short title of figure (maximum 15 words); detailed legend, up to 300 words.

Preparing tables

15 words. Detailed legends may then follow, but they should be concise. Tables should

Help and advice on scientific writing

The abstract is one of the most important parts of a manuscript. For guidance, please visit our page on Writing titles and abstracts for scientific articles.

Tim Albert has produced for BioMed Central a list of tips for writing a scientific manuscript. American Scientist also provides a list of resources for science writing. For more detailed guidance on preparing a manuscript and writing in English, please visit the BioMed Central author academy.

Typography

Please use double line spacing.

Type the text unjustified, without hyphenating words at line breaks.

Use hard returns only to end headings and paragraphs, not to rearrange lines.

Capitalize only the first word, and proper nouns, in the title.

All pages should be numbered.

Use the *Genome Biology* reference format.

Footnotes are not allowed, but endnotes are permitted.

Please do not format the text in multiple columns.

Greek and other special characters may be included. If you are unable to reproduce a particular special character, please type out the name of the symbol in full. **Please ensure that all special characters used are embedded in the text, otherwise they will be lost during conversion to PDF.**

Genes, mutations, genotypes, and alleles should be indicated in italics, and authors are required to use approved gene symbols, names, and formatting. Protein products should be in plain type.

Units

SI units should be used throughout (liter and molar are permitted, however).