A flexible framework to annotate and prioritize noncoding 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

Name for the tool:

[CMAP] – cancer mutation annotation and prioritization
or
[SMAP] – somatic mutation annotation and prioritization

CATTG.

Abstract [100 words…]

We developed a flexible framework to annotate and prioritize cancer noncoding somatic alterations. We analyzed functional impact of variants, such as gain-of function, and systematically associated regulatory elements to target genes. We further investigated associated genes, incorporating network topologies, gene functions, differential gene expression and epigenetic profiles. Together with other features (e.g. conservations), we developed a weighted scoring scheme to predict 'high-impact' variants. Using COSMIC and 570 whole-genome sequenced samples, a recurrence database has been created. Our method performed well based on various criteria – somatic recurrent and germline pathogenic variants, as well as a case study on an individual genome.

Keywords

Annotation, Prioritization, Noncoding driver, Somatic alterations, Cancer

Background

Cancer genome sequencing generally identifies thousands of somatic alterations in individual tumor genomes. A few of them, called drivers, contribute to oncogenesis, whereas the majority are passenger mutations accumulated during cancer progression [1]. Systematic studies of human cancer genomes have discovered a wide range of driver genes [2-6]. However, comparatively less effort has been invested in the noncoding portions of the genome. Recent discovery of somatic mutations in telomerase reverse transcriptase (*TERT*) promoter shows that regulatory variants may constitute driver events [7-10]. With the decrease of sequencing cost, international cancer consortia, such as TCGA (The Cancer Genome Altas) and ICGC (The International Cancer Genome Consortium), plan to perform large-scale cancer whole-genome sequencing in the near future. Thus, there is a great demand for high-throughput computational methods analyzing those variants.

The important role of regulatory variants in various diseases has generated a great deal of interest in studying noncoding sequences [11-14]. In contrast to coding variants, the functional impact of noncoding variants is more difficult to evaluate. Projects aiming to uncover potential regulatory sequences, such as The Encyclopedia of DNA Elements (ENCODE) [15] and 29 Mammals Project [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 [19]. A number of tools have been developed using ENCODE data that suggest most likely causal variants in linkage disequilibrium with GWAS SNPs or annotate

YAO FU 4/5/14 3:27 PM
Formatted: Font:12 pt

YAO FU 4/5/14 3:27 PM
Formatted: Normal

YAO FU 4/4/14 2:49 PM
Deleted: Somatic alterations in regulatory regions can cause oncogenesis. Here we

YAO FU 4/4/14 2:49 PM
Deleted: report a

YAO FU 4/4/14 2:50 PM
Deleted: noncoding cancer drivers

YAO FU 4/5/14 3:02 PM
Deleted: Our method integrates large-scale genomics resources, such as conservation and functional studies, analyzes loss-of- and gain-of-function events, systematically associates regulatory elements with likely target genes, examines network topologies, creates a recurrence database from publicly available whole genome cancer sequences and uses a weighted scoring scheme to prioritize 'high-impact' variants. Cancer-specific knowledge, such as known cancer genes and sample-specific open-chromatin and transcriptome profiles are also incorporated. To evaluate performance, our method can predict deleterious variants from individual genomes. [[EK: Need to put a bit more about how good is performance and how evaluated.]]

YAO FU 4/5/14 3:20 PM
Deleted:

YAO FU 4/5/14 3:19 PM
Formatted: Font:(Default) MS Reference Sans Serif

YAO FU 4/1/14 1:53 PM
Deleted: cancer

YAO FU 4/5/14 4:02 PM
Deleted: cancer

YAO FU 4/3/14 2:31 PM
Deleted: recurrent

YAO FU 4/2/14 9:45 PM
Deleted: dramatic

YAO FU 4/5/14 5:18 PM
Formatted: Font:+Theme Body, 12 pt

YAO FU 4/3/14 10:12 AM
Deleted: sequence conservation studies [17, 18],

noncoding variants. These include Haploreg [20], RegulomeDB [21], ANNOVAR [22], GEMINI [23], FunciSNP [24] and VEP [25]. Recently, two computational approaches – GWAVA and CADD were published to predict the deleterious effect of variants genome-wide [26, 27]. These two methods utilized machine-learning models trained on potential pathogenic variants or nearly fixed/fixed human derived alleles to distinguish deleterious variants from neutral ones.

While much work has been done for germline variants, this is not the case for cancer somatic mutations. Through analyzing the variation patterns of inherited polymorphisms, we have published a prototype approach to identify potential noncoding drivers [18]. Here, we report a more comprehensive and flexible framework - XXX - to annotate and prioritize somatic alterations integrating various resources from genomic and cancer studies. We analyzed both evolutionary and human-specific conservations to identify noncoding regions that are less likely to tolerant mutations. Our method predicts nucleotide-level loss-of and gain-of function events. To explore the functional impact of noncoding variants, we systematically associated regulatory elements to target genes using data from Roadmap Epigenomics Projects. We further investigated the essentiality of associated genes, by incorporating network topology analysis, gene function prioritization, differential gene expression detection and sample-specific epigenetic or open chromatin profiles. To prioritize 'high-impact' variants, we developed a weighted scoring scheme that takes into account the relative importance of various features, based on the mutation patterns observed in natural polymorphisms. Using 570 whole-genome sequenced cancer samples together with COSMIC data, we have also created a recurrence database, to which user-input genome will be compared. ~~The usage of the database will further benefit the prioritization of potential noncoding drivers.~~ The framework consists of two modules – (1) optional rebuilding of the data context by processing large-scale data resources and (2) efficient and high-throughput variants prioritization pipeline. Users could either use the pre-existing or customized data context to prioritize variants.

Besides mutations in the *TERT* promoter, no other regulatory drivers are currently characterized. Thus, due to a lack of gold standard for regulatory cancer drivers, we used somatic recurrent mutations and known germline pathogenic variants to evaluate the performance of our method. Our method has good prediction power for both somatic recurrent and germline pathogenic regulatory variants, and more importantly it contains multiple cancer-specific features, such as differential gene expression detection between tumor and matched normal samples. As a test case, we also applied our method to an individual cancer genome with the known *TERT* promoter mutation. Our method is able to prioritize the variant and provides a hypothesis of its potential functional impact. This shows that our method can help researchers and clinicians to prioritize a few somatic regulatory mutations for further studies.

Results and discussion

High-throughput technologies have generated huge amount of genomics data in the past decades. How to mine and integrate these data to tackle particular

YAO FU 4/3/14 12:51 PM
Deleted: regulatory

YAO FU 4/4/14 11:06 AM
Deleted: cancer

YAO FU 4/1/14 1:54 PM
Deleted: ???

YAO FU 4/4/14 11:07 AM
Deleted: cancer

YAO FU 4/4/14 11:05 AM
Deleted: variants

YAO FU 4/5/14 5:32 PM
Deleted: [28]

YAO FU 4/5/14 4:08 PM
Formatted: Font:+Theme Body

YAO FU 4/5/14 4:08 PM
Formatted: Font:+Theme Body

YAO FU 4/5/14 4:08 PM
Formatted: Font:+Theme Body

YAO FU 4/3/14 2:46 PM
Deleted: [28]Our method goes beyond the prototype in the following aspects. It 1) analyzes patterns of inherited polymorphisms among humans and evolutionary conservation across species to identify regions that are less likely to tolerate mutations; 2) integrates functional annotations and systems-level information from various biological networks; 3) uses functional essentiality and prior knowledge of known cancer genes; 4) predicts loss-of- and gain-of- function mutations for transcription-factor (TF) binding; 5) identifies distal associations between regulatory sequences and target genes using various histone modifications; 6) estimates recurrence of somatic alterations in publicly available cancer whole-genome sequencing data; 7) uses a weighted scoring scheme based on natural polymorphisms to prioritize potential 'high-impact' variants and 8) has a flexible framework for users to incorporate case-specific data, such as open-chromatin and transcriptome profiles from corresponding samples. The framework consists of two modules – (1) a complex-to-regenerate data context created by processing large-scale data resources and (2) efficient and high-throughput variants prioritization pipeline. Users could ei ... [1]

YAO FU 4/2/14 9:49 PM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/2/14 9:49 PM
Deleted: in TF binding sites

YAO FU 4/2/14 9:49 PM
Deleted: known

YAO FU 4/2/14 12:01 PM
Deleted: functional

scientific question remains a challenge. In this study, we first build an organized data context by processing large-scale genomics and cancer resources into small-scale informative data and then use them to annotate and prioritize cancer somatic alterations. The workflow is depicted in Figure 1 and the detailed description is in Material and Methods.

Variants in conserved regions

Sequences that are under strong negative or purifying selection are thought to have important biological functions [17]. In previous studies, oncogenes or tumor suppressor genes are found to experience higher intensity of evolutionary selective pressure than other disease-related and non-disease genes [28]. We utilized conservation data from different resources – cross-species conservation from GERP scores [29] and ultra-conserved elements [17], as well as population-level conservation from 1000 Genomes [18, 30] to detect potential damaging variants in noncoding regions. Each variant will be annotated with the corresponding conservation information. Our framework also implements the method used in *Khurana et al.,* [18] for users to find novel population-level conserved regions using depletion of common variants with user input polymorphisms or annotation data (Additional file 1).

Variants in potential regulatory elements

Regulatory elements, especially promoters and enhancers, are capable of regulating the expression of specific genes. We collected functional annotations from ENCODE [15] (transcription factor binding sites and the high-resolution motifs within them, enhancers from genome segmentations, ncRNAs and DNase I hypersensitive sites) and regions that are highly occupied by transcription factors (HOT) from *Yip et al.,* [31] to annotate potential regulatory sequences. Functional impact of variants in regulatory regions is generally restricted to *cis*-acting effects that control the spatial and temporal patterns of gene expression [13]. Activation of regulatory elements is associated with the underlying epigenetic or open chromatin landscape, which is largely cell-line specific [32]. Therefore, we provide a module to incorporate sample-specific epigenetic or open chromatin profiles to denote corresponding activation or inactivation of regulatory sequences; for example, enrichment of H3K27ac may indicate active state of enhancers in the sample.

Regulatory mutations can cause transcriptional alterations by either loss-of or gain-of- function effects. Loss-of-function variants in transcription factor binding motifs are likely to cause deleterious impact [18, 33, 34]. Variants decreasing the position weight matrix (PWM) scores could potentially alter the binding strength of transcription factors, or even eliminate the binding. Our 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 [7-10]. To the best of our knowledge, there is no automated tool to detect such events in whole tumor genomes. We incorporated a gain-of-motif scheme to scan and statistically evaluate [35] all possible motifs created by mutated alleles in promoter or enhancer regions. For each variant (SNV or indel), we concatenate it with +/- 29bp nucleotides around it and calculate sequence score for each possible motif against the PWMs. Gain-

YAO FU 4/3/14 3:44 PM
Deleted: Our framework...first build ... [2]

YAO FU 4/1/14 11:30 AM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/3/14 9:58 PM
Deleted: Sequences that are under strong negative or purifying selection are thought to have important biological functions [17]. W... utilized conservation data from ... [3]

YAO FU 4/3/14 1:54 PM
Formatted: Font:(Default) +Theme Body, 14 pt

YAO FU 4/4/14 11:27 AM
Formatted: Tabs:Not at 2.13"

YAO FU 4/1/14 11:07 PM
Deleted: elements

YAO FU 4/1/14 11:02 PM
Deleted:

YAO FU 4/1/14 12:03 PM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/2/14 1:06 PM
Formatted: Font:+Theme Body, Italic

YAO FU 4/1/14 4:47 PM
Deleted: Different from variants in protein coding sequences, variants in regulatory regions exert their function in a spatiotemporal manner [13] [[EK: this could also be true for coding variants if gene shows spatiotemporal expression]]. Activation of regulatory elements is associated with the underlying epigenomic or open chromatin landscape, which is largely cell-line specific [32]. ...herefo ... [4]

YAO FU 4/4/14 11:27 AM
Formatted: Font:(Default) +Theme Body, Font color: Auto

YAO FU 4/2/14 2:48 PM
Deleted: developed

YAO FU 4/1/14 2:02 PM
Formatted ... [5]

YAO FU 4/1/14 2:02 PM
Deleted: [[EK: is this correct]].

YAO FU 4/1/14 11:30 AM
Formatted: Font:+Theme Body, Font color: Red

of-motif events are identified when sequence score with mutated allele is significantly higher than the background (p < 4e-8), whereas that with germline allele is not. Our scheme is validated by the detection of ETS motifs created by the two cancer drivers in *TERT* promoter (Additional file 1: Table S1).

Associating distal regulatory elements with likely target genes

Positioned distant to their target genes, distal regulatory elements regulate gene expression through long-range interactions [36]. The linkages between regulatory elements and target genes remain elusive. To explore likely functional consequences of noncoding variants, we comprehensively define regulatory element-target gene pairs through correlating various epigenetic modifications with gene expression levels. 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 (REMC), for each regulatory element-candidate target gene pair, we computed the correlations of H3K4me1 and H3K27ac and the 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 (with various association confidence). To incorporate the ever-increasing amounts of genomic data, we include 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).

YAO FU 4/3/14 12:15 PM
Deleted: Correlating histone modifications with gene-expression

YAO FU 4/1/14 2:03 PM
Deleted:

YAO FU 4/3/14 12:15 PM
Deleted: data to identify

YAO FU 4/3/14 12:15 PM
Deleted: of distal regulatory elements

YAO FU 4/5/14 9:20 PM
Deleted: interpret

YAO FU 4/4/14 2:43 PM
Deleted: associations between

YAO FU 4/4/14 2:43 PM
Deleted: s and

Network analysis of variants associated with genes

Unlike germline mutations, somatic alterations are not expected to be under organism-level evolutionary selection pressure and thus are more likely to affect functional centers in gene interaction networks [37]. Network studies have found that cancer genes possess high topological centralities, even higher than essential genes [18, 37]. We use the regulatory element-target gene pairs to connect noncoding variants into a variety of networks: protein-protein interaction, regulatory and phosphorylation networks [18, 36, 38]. For each noncoding variant, we calculate the scaled network centrality (the percentile after ordering centralities of all genes in a particular network) of the associated gene in each network (Material and Methods). Amongst the different network centralities, we use the maximum centrality as the network disruptive measure of the variant. The higher this value, the more likely the variant would be deleterious. We make the scheme flexible so it can integrate user networks in addition to the pre-collected networks.

Gene prioritization: using expression and prior knowledge of target genes

Interpretation of the functional impact of noncoding variant can be greatly enhanced if the function of its target protein-coding gene is known. Many cancer genes are known to play a crucial role in cell proliferation and DNA repair. We incorporate prior knowledge of genes, such as known cancer-driver genes [2, 39], genes involved in DNA repair [40] and actionable genes ('druggable' genes) [41] to annotate noncoding variants that are more likely to be involved in cancer development and growth or their associated genes could be used as drug targets. In addition, user-specific gene lists can be easily input (Additional file 1).

Variants in regulatory elements may disrupt the expression of coding genes. Differential expression of target genes in cancer samples indicates the potential effect of noncoding variants. We provide a “differential gene expression analysis” module to detect differentially expressed genes in cancer samples (relative to matched normal) from RNA-Seq data. Lists of differentially expressed genes can be generated and used to annotate variants.

Weighted scoring scheme to prioritize variants

All of the above features are used to annotate and score variants. To integrate the various features to predict ‘high-impact’ somatic alterations, we developed a weighted scoring scheme, taking into account the relative importance of each feature. 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 would be the changes in PWMs. We weighted each feature based on the mutation patterns observed in natural polymorphisms (Material and Methods). Constrained by selection pressure, natural variations tend to arise in functionally unimportant regions. Thus, features that are frequently observed are less likely to contribute to the deleteriousness of variants and are weighted less. We calculated the information content to denote the importance of each feature. For each cancer mutation, we scored 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.

YAO FU 4/3/14 10:35 PM
Deleted: F

YAO FU 4/5/14 4:27 PM
Deleted: thus

HEADING

Performance on regulatory cancer somatic variants and germline pathogenic variants

Currently only two known regulatory variants are thought to act as cancer drivers. Hence, to evaluate the performance of our scoring scheme, we used recurrence to approximate the deleteriousness of somatic variants. Recurrence is considered as one potential sign of positive selection amongst tumors and is more likely to be associated with driver events [3]. We examined recurrence from two different perspectives – recurrence at the exact same-site and recurrence in the same regulatory element. First we obtained regulatory somatic variants from COSMIC [42] and classified them as same-site recurrent or non-recurrent (Material and Methods) [27]. Our method scores recurrent variants higher than non-recurrent ones (Wilcoxon rank-sum test: p-value < 2.2 e-16; Figure 3A). Variants that occur in more than 2 samples have higher scores than those that are in 2 samples. As we know, COSMIC collects somatic alterations from diverse papers and studies. We noticed a potential artifact related to pseudogenes (Additional file 1: Figure S2). After removing variants in pseudogenes, the trend of prediction scores persists (Additional file 1: Figure S3). Next we evaluated variants in recurrent regulatory elements using a separate dataset. We ran our pipeline on 119 breast cancer samples [43], and classified variants as occurring in recurrent elements or not (Material and Methods). We found that variants in recurrent elements get significantly higher

YAO FU 4/2/14 1:28 PM
Deleted: in TF binding sites

?

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 S4).

We note that cancer is a very heterogeneous disease and distinct molecular subtypes may involve unique oncogenesis mechanisms. Thus, tumor samples from different patients may involve different driver events. These unique drivers would not show recurrence across samples. Furthermore, in the absence of large sample sizes, it might be impossible to detect recurrence of mutations. Our method would be especially useful in such scenarios, since it has the ability to prioritize deleterious variants in each cancer genome. In relation to this, we have also created a recurrence database (including regulatory elements, coding genes and the same-site mutations) with publicly available cancer whole-genome sequencing data. Currently, we have identified recurrent loci or sites from 570 samples of 10 tumor types [43-45] and from COSMIC [42] (Table 1; Material and Methods). The use of the database along with our framework would provide higher confidence in prioritization of noncoding drivers. Moreover, the functional relevance and hence the biological mechanism by which drivers act is largely unknown. Our method provides an in-depth annotation of such variants, including the relative contribution of each feature to its deleteriousness. This knowledge would greatly help understand the potential oncogenic mechanisms.

Do 'x' EARLIER

Though designed primarily for somatic mutations, our framework contains several features that are applicable to germline variants. Thus, we also tested the ability of our method to distinguish germline pathogenic variants from neutral ones. We obtained pathogenic regulatory variants from HGMD [46] and three sets of controls from *Ritchie et al* [27] – 'unmatched', 'matched TSS' and 'matched region' (Material and Methods). 'Unmatched' control consists of likely neutral polymorphisms randomly selected from 1000 Genomes Project. Restrictions of '2Kb around TSS' and '1Kb around HGMD variants' are applied to 'matched TSS' and 'matched region' controls, respectively. Our method scored HGMD variants higher than all controls, with AUC scores of 0.86 (for 'unmatched'), 0.73 (for 'matched TSS') and 0.62 (for 'matched region') (Figure 3C and 3D). Results from CADD [26] using the same dataset are shown in Additional file 1: Figure S5 (AUC scores: 0.75 ('unmatched'), 0.68 ('matched TSS') and 0.61 ('matched region')). 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 the ability of our method to predict positives (Additional file 1: Figure S6). Generally speaking, our method has good prediction power for pathogenic regulatory variants. In addition, GWAS SNPs show higher scores than matched common polymorphisms (mean values: 0.41 vs. 0.34, p-value < 2.2e-16; Material and Methods; Additional file 1: Figure S7).

A case study: somatic variants from an individual cancer genome

High recurrence of the *TERT* promoter mutations implicates their important roles in oncogenesis [7]. Among the 570 cancer samples we collected, 7 samples contain the *TERT* promoter mutation (chr5: 1295228). We used one Medulloblastoma sample as an example to prioritize regulatory variants from

YAO FU 4/4/14 11:59 AM
Deleted: In summary, our method can predict potential 'high-impact' somatic variants.

YAO FU 4/2/14 10:02 PM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/1/14 2:07 PM
Deleted:

YAO FU 4/4/14 12:04 PM
Deleted: contain

YAO FU 4/2/14 3:24 PM
Deleted: across multiple samples.

YAO FU 4/2/14 3:24 PM
Deleted: non-recurrent

YAO FU 4/2/14 3:22 PM
Deleted: sample

YAO FU 4/2/14 1:32 PM
Deleted: even when a noncoding driver event could be detected due to recurrence,

YAO FU 4/1/14 12:11 PM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/2/14 1:32 PM
Deleted: its

YAO FU 4/2/14 1:33 PM
Deleted: it

YAO FU 4/4/14 12:05 PM
Deleted: s

YAO FU 4/4/14 12:06 PM
Deleted: by which the variant acts as a driver

YAO FU 4/1/14 2:06 PM
Deleted: [[EK: This paragraph should be shortened to two sentences and added at the end of first para in this section. Details can go in supplement. The last sentence is not very clear to me.]] Lack of consistent variants calling procedure, quality of COSMIC somatic mutations varies. This is a potential caveat when considering recurrence with COSMIC data. As shown in Figure S2, percentage of variants in pseudogenes increases as the number of recurrent samples increases. We suspect that reads containing these variants should probably be mapped to parent genes of pseudogenes, instead of the noncoding genome. After excluding variants in ... [6]

YAO FU 4/1/14 11:31 AM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/2/14 3:28 PM
Deleted: potential

YAO FU 4/1/14 2:10 PM
Deleted: **[[EK: This section below needs to be connected with the first para** ... [7]

whole-genome sequencing. Amongst the 2,183 somatic single nucleotide variants, the *TERT* promoter mutation ranks the 2nd (0.09%). Our method further suggests 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*. It is also found in another 5 liver samples and 54 COSMIC samples in our recurrence database. Besides DNA sequences, epigenomics or transcriptome could also be altered in cancer genomes. These data provide sample-specific activation or inactivation signatures of regulatory sequences. If provided, our framework is flexible in integrating those data into our annotation scheme (refer to Additional file 1 for details).

We also applied CADD and GWAVA on the Medulloblastoma sample. CADD ranked the *TERT* promoter mutation as 224th (10.3%) and GWAVA ranked it as 25th (1.15%) with matched region model (Additional file 1: Figure S8). However, only our method shows that the mutation corresponds to gain-of an ETS binding motif in the promoter of a cancer driver gene.

Output format and performance

XXX is a Linux/Unix based tool with a web-server available at xxxx.gersteinlab.org. The code is also posted under GitHub - http://gersteinlab.github.io/XXX/. 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 S9 and S10).

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 S3 (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

YAO FU 4/1/14 2:07 PM
Deleted: among the top 0.64% (14th)

YAO FU 4/1/14 2:07 PM
Deleted:

YAO FU 4/1/14 2:07 PM
Deleted: When analyzing all 100 Medulloblastoma samples with our method, this mutation is found as a recurrent one. In our scoring scheme considering recurrence across samples, the mutation ranked the 2nd.

YAO FU 4/1/14 12:16 PM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/1/14 12:19 PM
Deleted: 10th (0.46%), 25th (1.15%) and

YAO FU 4/2/14 10:52 AM
Deleted: 129

YAO FU 4/2/14 10:52 AM
Deleted: 5

YAO FU 4/2/14 10:53 AM
Deleted: .

YAO FU 4/2/14 10:52 AM
Deleted: 92

YAO FU 4/1/14 12:20 PM
Deleted: with

YAO FU 4/1/14 12:19 PM
Deleted: 'unmatched', 'matched TSS' and '

YAO FU 4/1/14 12:19 PM
Deleted: '

YAO FU 4/1/14 12:19 PM
Deleted: s, respectively. Upon detailed analysis of GWAVA 'unmatched' model, we found that the high ranking of this mutation is not due to the functional importance, but model bias influenced by the feature - distance to TSS

YAO FU 4/1/14 12:17 PM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/1/14 12:17 PM
Formatted: Font:+Theme Body, Font color: Red

YAO FU 4/1/14 12:18 PM
Deleted: Compared to GWAVA, both our method and CADD can predict deleter ... [8]

YAO FU 4/1/14 12:17 PM
Formatted ... [9]

YAO FU 4/5/14 5:15 PM
Deleted: FunSVPT

YAO FU 4/5/14 5:15 PM
Deleted: funseq2

YAO FU 4/5/14 5:15 PM
Deleted: FunSVPT

YAO FU 4/2/14 11:07 AM
Deleted: 4

useful for researchers to identify a few somatic events among thousands for further in-depth analysis to understand the mechanisms underlying oncogenesis.

Material and Methods

Data resources

We collected polymorphisms from 1000 Genomes Project Phase 1 [30], GERP scores and ultra-conserved elements from [17, 29], sensitive/ultra-sensitive regions from [18], functional genomics data from ENCODE [15], highly occupied regions (HOT) from [31] and histone modifications ChIP-Seq and gene expression RNA-Seq data from REMC [47]. Cancer driver genes are the union of genes from *Vogelstein et al*., cancer gene consensus and COSMIC [2, 39]. DNA repair and actionable genes are from [40, 41]. Binary protein-protein interaction network is from InWeb [48] and HINT [49]. Regulatory and phosphorylation networks are obtained from *Gerstein et al.,* [36] and *Lin et al.,* [38] respectively. Whole-genome somatic alterations contain 506 cancer genomes from *Alexandrov et al.,* [43] and 64 prostate cancer samples from [44, 45].

High-impact variants in motifs: Nucleotide resolution effect

User-input variants are first filtered against natural polymorphisms based on user-defined minor allele frequency (MAF) threshold to get rid of unlikely somatic variants (hg19). Currently, SNVs and small indels (<=20bp) will be analyzed.

1. Motif breaking

When variants hit transcription factor binding motifs under ENCODE Chip-Seq peaks, we examined 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 [34] (we calculated the difference between mutated and germline 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 [15], including TRANSFAC, JASPAR motifs.

2. Motif gaining

Whole-genome motif scanning generally discovers millions of motifs, of which, a large fraction are false positives. We focused on variants occurred in promoters (defined as -2.5kb from transcription starting sites) or regulatory elements significantly associated with genes. For each variant, +/- 29bp are concatenated from human reference genome (motif length is generally <30bp). For each PWM, we scanned the 59bp sequence. For each candidate motif encompassing the variant, we evaluated the sequence scores using TFM-Pvalue [35] (with respect to the PWM). Given a particular PWM (frequencies are transformed to log likelihoods), sequence score is computed by summing up the relevant values at each position in the PWM. If the p-value with mutated allele <= 4e-8 and the p-value with germline allele > 4e-8, we define the variant creating a novel motif. The process is repeated for all PWMs and all variants. The sequence score changes are reported in the output.

Associating regulatory elements to likely target genes

We define both proximal and distal associations. For proximal associations, we assign variants in gene promoters, introns or UTRs to their nearby genes. For distal associations, in addition to those identified in [31], we further expanded the method to all ENCODE non-coding regulatory elements and identified ~769K regulatory elements significantly associated with ~17K genes (see below). The association confidence is reported in the output for each regulatory element - target gene pair.

Correlating histone modifications 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 [50] 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 [47] (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 [51]. 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 incorporated 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 [52] when having RPKM values and DESeq [53] 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 examined their network properties in various networks. For each network, we calculated 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. Scripts are provided to calculate network centralities (Additional file 1). User can easily incorporate other networks in this analysis.

Recurrence database from whole-genome sequencing

One criterion 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 patterns in regulatory sequences. We analyzed somatic alterations from 570 samples of 10 cancer types to create the recurrence database [43-45], similar to the cancer recurrent gene database in cBio [54]. For each cancer type, recurrent regulatory elements, coding genes and the same-site mutations are stored as entries in the database. We also incorporated recurrent somatic regulatory variants from COSMIC (version 68) into our database. Recurrent variants are defined as identified in whole-genome sequencing and observed in at least 2 samples. 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

1. Coding scoring scheme

Variants occurred in coding regions (GENCODE 16 for the current version; users can replace this with other GENECODE versions) are analyzed with VAT (variant annotation tool) [55]. Variants are ranked based on the following scheme (each criterion gets score 1): 1) non-synonymous; 2) premature stop; 3) is the gene under strong selection; 4) is the gene a network hub; 5) recurrent; 6) GERP score>2.

2. Noncoding scoring scheme (weighted scoring scheme)

Features used to score noncoding variants are shown in Additional file 1: Table S2. In general, features can be classified into two classes - discrete and continuous. Discrete features are binary, such as in ultra-conserved elements or not. Continuous features: 1. GERP score, 2. Motif-breaking score is the difference between germline and mutated alleles in PWMs, 3. Motif-gaining score is the sequence score difference between mutated and germline alleles, 4. Network centrality score (the cumulative probability, see ‘Network analysis of variants associated with genes’). If one variant has multiple values of a particular feature (e.g. breaks multiple motifs), the largest value is used.

We weighted each feature based on the mutation patterns observed in the 1000 Genomes polymorphisms. We randomly selected 10% of the 1000 Genomes Phase 1 SNPs (~3.7M) and run through our pipeline. For each discrete feature d, we calculated the probability p_d that overlaps a natural polymorphism. Then we computed 1-Shannon entropy *(1)* as its weighted value w_d. The value ranges from 0 to 1 and is monotonically decreasing when the probability is between 0 and 0.5 (in our study, the probability of observing each feature is below 0.5).

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

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

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in natural 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 calculated feature weights for each v_c. In particular, we discretized at each v_c and computed 1-Shannon entropy using *(2)*. Then we fitted 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.

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

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

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. 'nls' function in R is used to fit curves.

The criterion of 'GERP >2' has been commonly used to define conserved bases [15]. For GERP score, we chose to use sigmoid transformation to transform scores to the range 0~1. The parameters we chose make the sigmoid curve sharp at 'GERP = 2' (Figure S1). The sigmoid transformation preserves the 'GERP > 2' cut-off and makes the score continuous at the same time. We calculated *(1)* treating 'GERP > 2' as a discrete feature. Then we used $w_d * sigmoid\ transformed\ value$ to assign weighted value for each continuous GERP score.

Finally, for each cancer variant, we scored it by summing up the weighted values of all its features *(3)*. Considering the situation that some features are subsets of other features, to avoid overweighting similar features, we take into account feature dependencies when calculating the sum-up scores. As shown in Table S2, when having leaf features, the weighted values of root features are ignored. For example, when a variant occurs in sensitive regions, the score of 'in functional annotations' is not used in the sum-up. Leaf features are assumed independent. Variants ranked on top of the output are those with higher scores and are most likely to be deleterious.

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

Application to regulatory pathogenic and somatic cancer variants

1. 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 and observed in at least 2 samples. All other variants (1,311,389) are non-recurrent ones (with GENCODE 16). Because the same sample from different papers may have multiple ids, we also defined recurrence based on number of studies. Study-based recurrent variants also have higher scores than non-recurrent ones (Wilcoxon test: p-value = 2.16 e-07). We also calculated the percentage of variants in pseudogenes. As shown in Figure S2, the percentage increases with the number of recurrent samples or studies.

2. 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 in functional annotations. From 119 breast cancer samples, there are 24,443 and 126,217 variants in recurrent and non-recurrent regulatory elements respectively. The feature of recurrence is not considered in the weighted scoring scheme for variants in sections 1 and 2.

3. Germline pathogenic variants and matched controls
Genome locations of pathogenic regulatory variants (from HGMD [46] - 1,614) and three sets of negative controls were downloaded from GWAVA [27]. The control sets – 'unmatched', 'matched TSS' and 'matched region', contain regulatory polymorphisms from 1000 Genomes with minor allele frequency ≥ 1%. 'Unmatched' control has 161,400 randomly selected variants. 'Matched TSS' control includes 16,140 variants matched for distance to the nearest TSS. 'Matched region' control has variants within 1kb around HGMD regulatory variants (5,027). Allele information for HGMD variants was obtained from HGMD database. For controls, the alleles were from ENSEMBL BioMart, using reference SNP ids.

We downloaded pre-calculated CADD scores for 1000 Genomes variants and extracted corresponding scores for control sets. For HGMD variants, we used the online CADD server to obtain the scores.

We compared the prediction scores of HGMD variants with three sets of controls using various measures – TPR (true positive rate), FPR (false positive rate), precision and recall. We treated HGMD as positive set and controls as negative sets. For each possible score, we draw the cut-off to make predictions and calculated - TPR = TP/(TP+FN); FPR = FP/(FP+TN); Precision = TP/(TP+FP); Recall = TP/(TP+FN). TP: true positive; FP: false positive; TN: true negative; FN: false negative. AUC score is the cumulative area under the curve of TPR and FPR.

We also tested our method with GWAS SNPs (6,604) and matched controls (66,040) from [27]. Allele information was obtained from ENSEMBL BioMart.

Framework flexibility
User data can be easily incorporated into our framework. Cancer sample-specific studies, such as histone modifications and gene expression, are especially important to evaluate variants impact. Please refer to 'Additional file 1' for usage.

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 the 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. We also thank Zhixiang Lin for useful discussion and proofreading the manuscript.

References

1. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446:**153-158.
2. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4:**177-183.
3. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, et al: **MuSiC: identifying mutational significance in cancer genomes.** *Genome Res* 2012, **22:**1589-1598.
4. Reimand J, Bader GD: **Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.** *Mol Syst Biol* 2013, **9:**637.
5. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N: **OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.** *Bioinformatics* 2013, **29:**2238-2244.
6. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N: **Comprehensive identification of mutational cancer driver genes across 12 tumor types.** *Sci Rep* 2013, **3:**2650.
7. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA: **Highly recurrent TERT promoter mutations in human melanoma.** *Science* 2013, **339:**957-959.
8. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, Kadel S, Moll I, Nagore E, Hemminki K, et al: **TERT promoter mutations in familial and sporadic melanoma.** *Science* 2013, **339:**959-961.

9. Killela PJ, Reitman ZJ, Jiao Y, Bettegowda C, Agrawal N, Diaz LA, Jr., Friedman AH, Friedman H, Gallia GL, Giovanella BC, 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.

10. Vinagre J, Almeida A, Populo H, Batista R, Lyra J, Pinto V, Coelho R, Celestino R, Prazeres H, Lima L, et al: **Frequency of TERT promoter mutations in human cancers.** *Nat Commun* 2013, **4:**2185.

11. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al: **Systematic localization of common disease-associated variation in regulatory DNA.** *Science* 2012, **337:**1190-1195.

12. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al: **Identifying recent adaptations in large-scale genomic data.** *Cell* 2013, **152:**703-713.

13. Sakabe NJ, Savic D, Nobrega MA: **Transcriptional enhancers in development and disease.** *Genome Biol* 2012, **13:**238.

14. Ward LD, Kellis M: **Interpreting noncoding genetic variation in complex traits and human disease.** *Nat Biotechnol* 2012, **30:**1095-1106.

15. The Encode Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.

16. Lowe CB, Haussler D: **29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome.** *PLoS One* 2012, **7:**e43128.

17. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304:**1321-1325.

18. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, et al: **Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics.** *Science* 2013, **342:**1235587.

19. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M: **Linking disease associations with regulatory information in the human genome.** *Genome Res* 2012, **22:**1748-1759.

20. Ward LD, Kellis M: **HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants.** *Nucleic Acids Res* 2012, **40:**D930-934.

21. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al: **Annotation of functional variation in personal genomes using RegulomeDB.** *Genome Res* 2012, **22:**1790-1797.

22. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38:**e164.

23. Paila U, Chapman BA, Kirchner R, Quinlan AR: **GEMINI: integrative exploration of genetic variation and genome annotations.** *PLoS Comput Biol* 2013, **9:**e1003153.

24. Coetzee SG, Rhie SK, Berman BP, Coetzee GA, Noushmehr H: **FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with**

genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res* 2012, **40:**e139.

25. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *Bioinformatics* 2010, **26:**2069-2070.
26. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants.** *Nat Genet* 2014.
27. Ritchie GR, Dunham I, Zeggini E, Flicek P: **Functional annotation of noncoding sequence variants.** *Nat Methods* 2014.
28. Thomas MA, Weston B, Joseph M, Wu W, Nekrutenko A, Tonellato PJ: **Evolutionary dynamics of oncogenes and tumor suppressor genes: higher intensities of purifying selection than other genes.** *Mol Biol Evol* 2003, **20:**964-968.
29. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A: **Distribution and intensity of constraint in mammalian genomic sequence.** *Genome Res* 2005, **15:**901-913.
30. The 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491:**56-65.
31. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M: **Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.** *Genome Biol* 2012, **13:**R48.
32. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al: **Histone modifications at human enhancers reflect global cell-type-specific gene expression.** *Nature* 2009, **459:**108-112.
33. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M: **Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.** *Genome Res* 2013, **23:**800-811.
34. 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.
35. Touzet H, Varre JS: **Efficient and accurate P-value computation for Position Weight Matrices.** *Algorithms Mol Biol* 2007, **2:**15.
36. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al: **Architecture of the human regulatory network derived from ENCODE data.** *Nature* 2012, **489:**91-100.
37. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network.** *Proc Natl Acad Sci U S A* 2007, **104:**8685-8690.
38. Lin J, Xie Z, Zhu H, Qian J: **Understanding protein phosphorylation on a systems level.** *Brief Funct Genomics* 2010, **9:**32-42.
39. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW: **Cancer genome landscapes.** *Science* 2013, **339:**1546-1558.
40. Ruark E, Snape K, Humburg P, Loveday C, Bajrami I, Brough R, Rodrigues DN, Renwick A, Seal S, Ramsay E, et al: **Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer.** *Nature* 2013, **493:**406-410.

41. Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, Ducar M, Van Hummelen P, Macconaill LE, Hahn WC, et al: **High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing.** *Cancer Discov* 2012, **2:**82-93.
42. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, et al: **COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.** *Nucleic Acids Res* 2011, **39:**D945-950.
43. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al: **Signatures of mutational processes in human cancer.** *Nature* 2013, **500:**415-421.
44. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al: **Punctuated evolution of prostate cancer genomes.** *Cell* 2013, **153:**666-677.
45. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, et al: **The genomic complexity of primary human prostate cancer.** *Nature* 2011, **470:**214-220.
46. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN: **Human Gene Mutation Database (HGMD): 2003 update.** *Hum Mutat* 2003, **21:**577-581.
47. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al: **The NIH Roadmap Epigenomics Mapping Consortium.** *Nat Biotechnol* 2010, **28:**1045-1048.
48. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, et al: **A human phenome-interactome network of protein complexes implicated in genetic disorders.** *Nat Biotechnol* 2007, **25:**309-316.
49. Das J, Yu H: **HINT: High-quality protein interactomes and their applications in understanding human disease.** *BMC Syst Biol* 2012, **6:**92.
50. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22:**1760-1774.
51. Wolfe MHaDA: *John Wiley and Sons* 1973**:**pages 185–194.
52. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A: **Differential expression in RNA-seq: a matter of depth.** *Genome Res* 2011, **21:**2213-2223.
53. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11:**R106.
54. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer Discov* 2012, **2:**401-404.
55. Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A, Rozowsky J, Clarke D, Snyder M, Gerstein M: **VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment.** *Bioinformatics* 2012, **28:**2267-2269.

Figures

Figure 1 - Schematic workflow.

Figure 2 - Weighted scoring scheme.

A) Features used in weighted scoring scheme; B) Motif-breaking scores and corresponding weighted values.

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

A) Score distribution of variants based on their recurrence in COSMIC; B) Score distribution of variants based on recurrent regulatory elements in 119 breast cancer samples; C) Prediction scores of regulatory variants from HGMD and controls; D) ROC curves comparing HGMD with controls;

Tables

Table 1 - Summary of recurrence database.

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

Additional files

Additional file 1 – Supplementary information

This file contains supplementary figures, supplementary tables and software manual.

Table 1

Cancer Type	# Samples	# Somatic Mutations (SNVs)	# Recurrent Elements | Genes | 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

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

YAO FU 4/4/14 2:48 PM
Formatted: Font:Italic

Variant
chr5: 1295228 *G -> A*

YAO FU 4/4/14 2:48 PM
Formatted: Font:Italic

YAO FU 4/4/14 2:48 PM
Formatted: Font:Italic

Typography

Please use double line spacing.

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.