Published online 4 October 2021

Nucleic Acids Research, 2022, Vol. 50, No. 1 ***e4***
https://doi.org/10.1093/nar/gkab877

SBSA: an online service for somatic binding sequence annotation

Limin Jiang[1,2,3], Fei Guo[2], Jijun Tang[3], Hui Yu[4], Scott Ness[4], Mingrui Duan[4], Peng Mao[4], Ying-Yong Zhao[1,*] and Yan Guo[4,*]

[1]Faculty of Life Science & Medicine, Northwest University, No. 229 Taibai North Road, Xi'an 710069, China, [2]School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, [3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China and [4]Comprehensive cancer center, Department of Internal Medicine, University of New Mexico, Albuquerque, NM 87109, USA

Received April 20, 2021; Revised September 10, 2021; Editorial Decision September 16, 2021; Accepted September 17, 2021

ABSTRACT

Efficient annotation of alterations in binding sequences of molecular regulators can help identify novel candidates for mechanisms study and offer original therapeutic hypotheses. In this work, we developed Somatic Binding Sequence Annotator (SBSA) as a full-capacity online tool to annotate altered binding motifs/sequences, addressing diverse types of genomic variants and molecular regulators. The genomic variants can be somatic mutation, single nucleotide polymorphism, RNA editing, etc. The binding motifs/sequences involve transcription factors (TFs), RNA-binding proteins, miRNA seeds, miRNA-mRNA 3′-UTR binding target, or can be any custom motifs/sequences. Compared to similar tools, SBSA is the first to support miRNA seeds and miRNA-mRNA 3′-UTR binding target, and it unprecedentedly implements a personalized genome approach that accommodates joint adjacent variants. SBSA is empowered to support an indefinite species, including preloaded reference genomes for SARS-Cov-2 and 25 other common organisms. We demonstrated SBSA by annotating multi-omics data from over 30,890 human subjects. Of the millions of somatic binding sequences identified, many are with known severe biological repercussions, such as the somatic mutation in *TERT* promoter region which causes a gained binding sequence for E26 transformation-specific factor (ETS1). We further validated the function of this *TERT* mutation using experimental data in cancer cells. Availability: http://innovebioinfo.com/Annotation/SBSA/SBSA.php.

INTRODUCTION

Nucleic acid sequence altering mechanisms, such as somatic mutation, RNA editing, and single nucleotide polymorphism (SNP), can have devastating biological consequences, including tumorigenesis, if they alter pivotal binding sequences for transcription factors (TFs), RNA binding proteins (RBPs), microRNA (miRNA) seeds and miRNA-mRNA 3′-UTR binding targets. Somatic mutations are the acquired mutations and the well-known culprit in tumorigenesis. RNA editing refers to the enzymatic modification of RNA sequence after the genetic code has been transcribed by the RNA polymerase. Abnormal RNA editing activity, either increase (1,2) or decrease (3,4), have been identified in various tumors. SNPs are a representative type of germline variants that can regulate gene expression and thus affect disease risk. It has been found that certain SNPs, known as expression quantitative trait loci (eQTL), regulate gene expression by affecting regulatory binding sequences (5). In an extreme scenario, a single SNP can lead to a severe Mendelian disease.

By binding with certain motifs in their target sequences, TFs, RBPs and miRNAs work at distinct levels to coordinate a proper, functional cellular transcriptome. Genomic variants of various types can occur in the binding motifs for TFs, RBPs, miRNA seeds and miRNA–mRNA 3′-UTR binding targets. TF is a protein that controls the rate of transcription of genetic information from DNA to mRNA by binding to a specific motif in regulatory DNA. For example, the oncogenic E26 transformation-specific (ETS) factor may bind with a cryptic binding site triggered by a well-known somatic mutation in the *TERT* promoter region. This mutation creates a binding sequence, TTCCGG, for ETS proteins and thereby upregulates *TERT* expression, which leads to uncontrolled cell proliferation and eventually results in cancer (6). RBPs are proteins that bind to

*To whom correspondence should be addressed. Tel: +1 505 925 0099; Fax: +1 505 925 4459; Email: yanguo@gmail.com
Correspondence may also be addressed to Ying-Yong Zhao. Tel: +86 29 88305273; Fax: +86 29 88303572; Email: zyy@nwu.edu.cn

the double or single-stranded RNA sequences to regulate mRNA turnover or splicing. The impact of RBPs on cancer has been well studied (7). On average, RBPs have 3 mutations per Mb in cancers (8), and these immense mutations can cause somatic binding sequence changes and disrupt regulatory functions of RPBs. MiRNAs are a type of small non-coding RNAs that regulate the translation of mRNAs through binding between seed regions (of miRNAs) and target sites in mRNA 3′-UTRs. Somatic mutations in the seed regions or target sites of miRNAs can cause disturbance of the normal miRNA-mRNA binding relationship, which may lead to a disease. For example, a mutation in *hsa-miR-96* seed region is responsible for nonsyndromic progressive hearing loss (9). An SNP in the seed region of human *miR-184* causes the EDICT syndrome (10). Considerable efforts have been made to curate somatic mutations in miRNAs and the consequential impacts (11).

There has been a plethora of evidence of severe consequences resulting from mutated binding sequences (6,9–14). Previous studies mostly rely on individualized data mining techniques to identify candidate somatic binding sequences; however these fragmented or in-house bioinformatics solutions cannot be reutilized by other research groups. In this work, we developed Somatic Binding Sequence Annotator (SBSA) as a full-capacity online tool to annotate altered binding motifs/sequences, addressing diverse types of genomic variants and molecular regulators (Figure 1). SBSA annotates the precise gain, loss, or disruption of a binding sequence resulting from an arbitrary type of genomic variation. SBSA has been used to analyze multi-omic data from over 30 890 subjects to curate the results into the database SMDB (15). Here, a detailed protocol for SBSA is presented, including software implementation, input data requirements, and demonstrative analysis results.

MATERIALS AND METHODS

For a given reference genome, SBSA expects one dataset of genomic variants and another dataset of motifs, binding sequences, or target genomic regions (Figure 1). Depending on the manifestation format of the input files, SBSA performs a relevant annotation of the input variants in terms of binding disruption to three types of molecular regulators: TF, RBP and miRNA. SBSA offers a plethora of variant datasets for user's exploration, and it also curates a large collection of binding motifs and binding regions for TFs, RBPs and miRNAs. Other than these pre-loaded datasets and libraries, SBSA can also tackle variant files and motifs/binding sequences that are originally generated by users. In theory, SBSA can analyze binding sequence data generated in any organism, as long as the user supplies the reference genome sequence. With pre-loaded reference genomes, SBSA provides full support for a total of 26 species, covering common mammals (e.g. human and mouse), vertebrates (e.g. chicken and *fugu*), insects (e.g. fruit fly and bee), nematodes, plants, and viruses. In particular, the reference genome for SARS-CoV-2, with variant information from 19 307 variant strains, is also supported by SBSA.

Binding motifs and sequences

For TFs, we obtained 11 761 distinct binding motifs from a total of 11 databases (3D-footprint (16), CISBP (17), HT-SELEX2 (18), humanC2H2ZF-ChIP (19), HumanTF (20), HumanTF2 (21), JASPAR (22), SMILE-seq (23), UniPROBE (24), footprintDB (25) and HOCOMOCO (26)). For RBP, a total of 1,867 binding motifs were downloaded from four source databases (ATtRACT (27), ORNAment (28), RBPDB (29) and RBPmap (92) (30)). For miRNA seed information, we parsed human miRNA seed region files from miRBase v22.1 (31). For miRNA's 3′-UTR targets, we imported target sequences from starBase v2.0 (32).

Testing variant data

SBSA is designed to annotate genomic variants of an indefinite type. That is, SBSA can tackle both point variants and indels, and can handle all common types of genomic variants related to SNPs, somatic mutations, and RNA-editing. To demonstrate SBSA, we obtained the following five groups of testing variant datasets from diverse sources. The first and second groups originated from The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), respectively, both relating to somatic mutations. The TCGA dataset group covered 3,037,744 somatic mutations in total, which were identified from 10 182 subjects of 33 cancer types. The ICGC dataset group covered a total of 80 015 947 somatic mutations sequenced from 19,729 subjects of 57 cancer types compiled by 81 different projects. The third and fourth groups pertained to germline mutations, originating from the authoritative SNP source dbSNP (v152) and eQTL source GTEx (v8), respectively. The generic SNPs totaled 660 146 174 in the dbSNP dataset group, and the cis-eQTLs totaled 71 478 479 as combined from 49 tissue sites. Lastly, SBSA provides a dataset of 4.67 million A-to-I RNA editing events derived from REDIportal (33). The complete list of cancer names, abbreviations, tissue names and sample sizes for ICGC, TCGA and GTEx are available in Supplementary Tables S1, S2 and S3, respectively.

Identifying variant-affected binding sequences

As outlined in Figure 1, SBSA integrates two primary inputs with respect to a given reference genome, in order to identify variant-affected binding sequences for a certain type of molecular regulators. The GRCh38 human reference genome was utilized for all demonstrative analyses, and sequences in this manuscript or in general output are shown in the 5′ to 3′ orientation.

As one primary input to SBSA, a variant can be either a single nucleotide variant or an indel, and can fall into but is not limited to the following categories: somatic mutation, RNA editing event, and SNP. As the other primary input to SBSA, a target motif or binding sequence refers to consecutive nucleotides forming a short sequence (<25 nt) that can be potentially recognized and bound by a molecular regulator. Anchoring at the variant site, SBSA extracts a short sequence from the reference genome by extending symmetrically in both 5′- and 3′-directions. This extended sequence

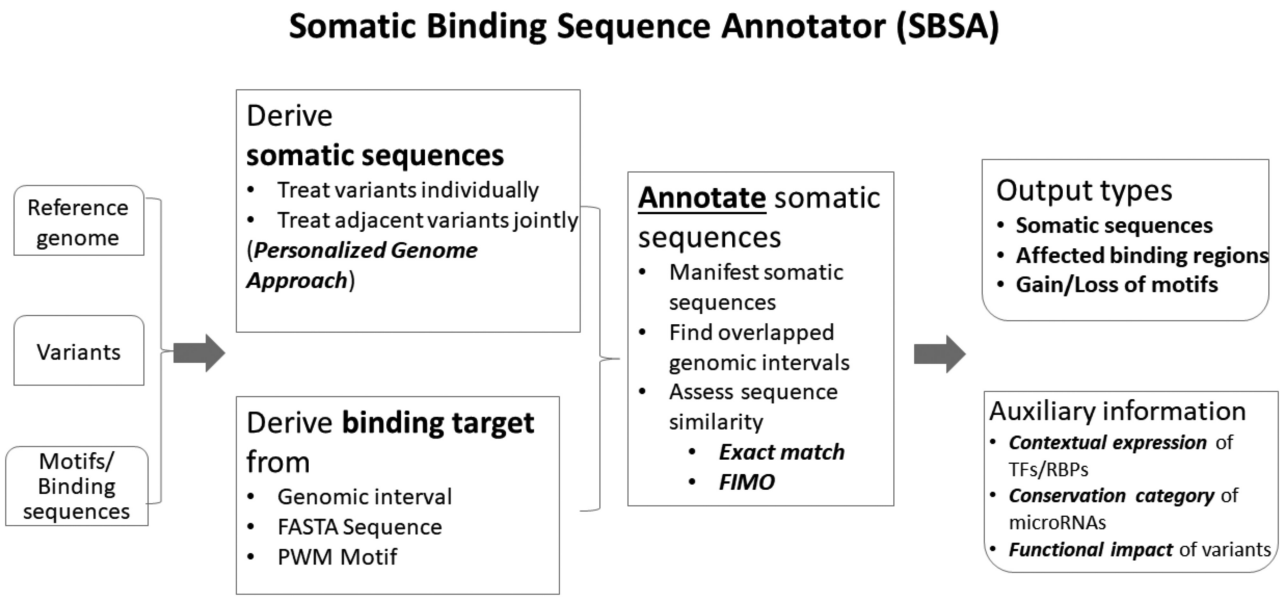

Figure 1. Somatic binding sequence annotator (SBSA) annotates genomic variants with respect to binding motifs of molecular regulators. Of the three inputs, variant specification file is mandatory, while the other two can be preloaded datasets. SBSA supports 26 reference genomes and harbors abundant motifs/binding sequences for three major types of transcription regulators: TFs, RBPs, and microRNAs. Depending on the type and specification of inputs, SBSA performs a relevant annotation of the input variants and generates a corresponding output. For TF/RBPs, gain or loss of binding motifs is affirmed by either FIMO or an intuitive Exact match method; for microRNA seeds or microRNA binding targets, the affected genomic regions are identified; in the simplest work mode, somatic sequences centered upon the input variants are generated. TF, transcription factor; RBP, RNA-binding protein; FIMO, the bioinformatics algorithm Find Individual Motif Occurrences; FASTA, the sequence file format for Fast Adaptive Shrinkage/Thresholding Algorithm; PWM, Position-Weight Matrix.

is of the same length as the target motif to be compared to. A 'reference sequence' is the target motif sequence without any sequence change, and a 'somatic sequence' refers to the mutated motif sequence including the concerned variant.

A binding motif/sequence is commonly represented in a Position-Weight Matrix (PWM). In such cases, the key is to comparing the derived somatic sequence against the PWM in terms of sequence similarity. Several methods, such as FIMO (34), motifbreakR (35) and RSAT (36), were previously developed for estimating the binding potential between a sequence and a motif. SBSA implemented FIMO for this PWM-sequence similarity assessment, and also developed an Exact match method as an alternative approach. With FIMO, if the binding potential *P*-value decreases from the reference sequence to the somatic sequence (binding propensity increases), we term that the variant causes a Gain of this binding motif; conversely, when the binding potential *P*-value increases from the reference sequence to the somatic sequence (binding propensity decreases), we term that the variant causes a Loss of this binding motif. The new Exact match method uses a simple yet intuitive strategy. Given the PWM of a binding motif, we approve all nucleotides at each individual position that exceed the minimum background probability threshold (default: 0.25), and generate all combinatorial binding sequences by stringing these position-wise nucleotides (Figure 2A). The pair of reference sequence and somatic sequence are checked against all motif-derived binding sequences. If exact match occurs between a binding sequence and the reference sequence, yet not between a binding sequence and the somatic sequence, a Loss of the binding motif is asserted; conversely, if exact match occurs between a binding sequence and the somatic sequence, yet not between a binding sequence and the reference sequence, a Gain of the binding motif is asserted.

Of note, SBSA applied different strandedness strategies with respect to different types of molecular regulators. For TF, sequence matching is sought from both the sense and antisense strands; for RBP, miRNA seeds and miRNA-mRNA 3′-UTR targets, only the sense strand is interrogated.

Personalized genome approach

SBSA allows an input file containing different types of variants,. The input file can include hundreds or thousands rows (truncated to the first 25 000 rows for certain intensive calculations). By default, these variants are treated mutually independently, leading to somatic sequences that each incorporates a single variant. Nevertheless, SBSA offers an optional Personalized Genome approach to analyze somatic sequences, where multiple variants in close vicinity are jointly accommodated in one somatic sequence (Figure 2B). Indels are considered as well as point variants. With the Personalized Genome approach, SBSA derives $2^k - 1$ somatic sequences for k adjacent variants, with each representing one combination of these adjacent mutations. After enumerating all possible somatic sequences, Gain or Loss of the target motif is inferred by comparing the reference sequence against the group of somatic sequences. As the term implies, Personalized Genome approach attempts to accommodate multiple adjacent variants manifested in an individualized genome, so it is only valid when the input variant file is summarized from a single subject (rather than from a cohort).

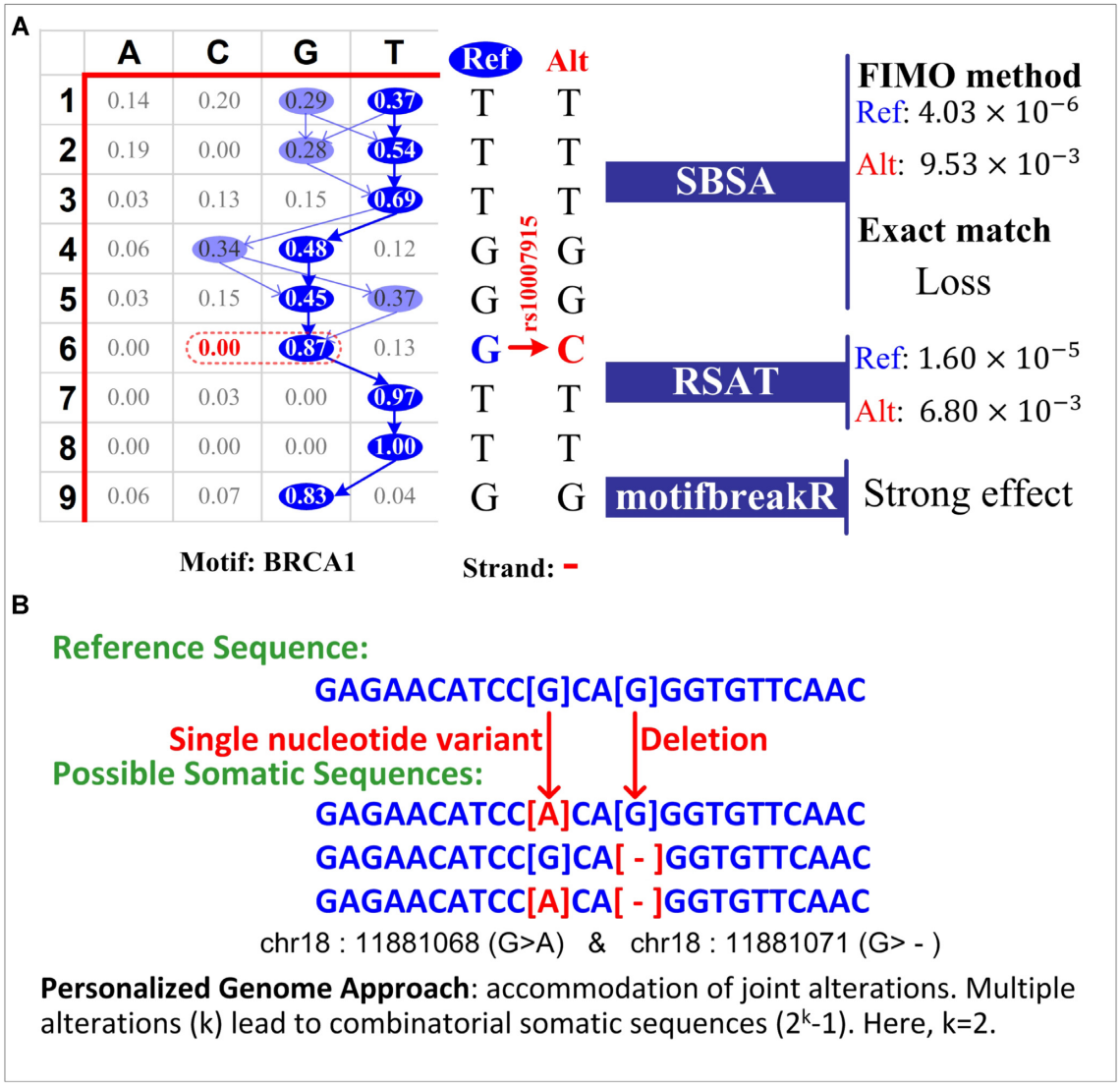

Figure 2. Exact match and personalized genome approaches of SBSA. (**A**) The exact match approach with which SBSA compares variant-derived sequences against a target motif (BRCA1′s motif is used as an example here). The focal variant is rs10007915, which leads to a pair of a reference sequence and a somatic sequence that differs precisely at chr4:105144151 (G > C variant). Given the Position-Weight Matrix (PWM) of BRCA1, traditional motif similarity algorithms, such as FIMO, RSAT, and motifbreakR, assesses the fitness of the reference/somatic sequence to the PWM, and all of them predict that a binding relationship retain for the somatic sequence, despite a diminishment of the statistical significance. With the Exact match approach, SBSA derives all likely target sequences (blue-arrow-connected paths) based on a position-wise probability threshold (default at 0.25), and seeks exact matching of the reference/somatic sequence with any derived target sequence. In this example, the reference sequence finds a hit within the group of target sequences whereas the somatic sequence does not, so the Exact match method concludes a Loss of the binding motif in the somatic sequence. (**B**) Illustration of the Personal Genome approach to generating combinatorial somatic sequences. A subject in TCGA (TCGA-BF-AAP1-01A) carries two adjacent variants on chromosome 18, one being a single nucleotide variant and the other a deletion. For this genomic segment, three ($3 = 2^2 - 1$) somatic sequences can be generated for analysis by Exact match or FIMO.

TERT mutation *in vivo* experiment

To validate the gain-of-function phenotype of the mutations in the *TERT* gene promoter, we used luciferase reporter plasmids (i.e. pGL4.0-TERT-wt and pGL4.0-TERT-mutant) in which the reporter gene is linked with either a wild-type (WT) or mutant *TERT* promoter fragment (G228A), as described in the previous study (14). The reporter plasmid (90 ng) was mixed with pRLCMV (renilla control) and transfected into A375 melanoma cells, using the XtremeGene-HP (Roche) transfection reagent. Transfected cells were incubated for 24 h and the luciferase activity (normalized by renilla) was measured using the Dual-Luciferase Reporter assay system (Promega). Experiments were repeated three times independently, with three technical repeats in each experiment.

RESULTS

Online application implementation

Research on somatic mutations in binding sequences has been accelerated since high-throughput sequencing technology became available. However, the research field still lacks

a tool that focuses on genome-wide variants in important gene regulatory sequences for TFs, RBPs and miRNAs. To promote the identification of functional variants in these binding sequences, we developed SBSA to enable fast and easy detection of somatic binding sequences at the genome-scale. The core program of SBSA is developed in Python, and the web interface is designed in PHP and Javascript. SBSA provides online service at http://innovebioinfo.com/Annotation/SBSA/SBSA.php. The overview of SBSA is displayed in Figure 1. SBSA annotates common genomic variants such as single nucleotide substitutions and indels in the binding motifs of important transcription regulators: TFs, RBPs, miRNA seeds, and miRNA-mRNA targets.

The primary input to SBSA is a file containing numerous genomic variants. As explained in the Materials and Methods above, SBSA treats variants that are not limited in form (either single nucleotide variants or indels) or type (SNPs, RNA editing events, or somatic mutations). This variant file can use either a standard Variant Call Format (v4.0) or a special Comma-Separated-Values format with specified chromosome positions and variant alleles. Along with the variant file, the user needs to inform SBSA on the species of the investigated biological sample. In the background, reference genomes for 26 species are pre-installed. Users must make sure the chromosome names in the variant file are of the same form as those of the background reference genome file (a list of standard chromosome names of all pre-installed genomes is provided as a reference). SBSA combines the variant specification and the reference genome to derive a pair of somatic sequence and reference sequence, which carries and lacks the specified variant, respectively.

The other input to SBSA defines the concerned binding target(s). This input can be provided by the user *ab initio*, or be chosen from built-in libraries. User-provided information can take the form of nucleotide sequences (in FASTA format), PWMs in format of Multiple Em for Motif Elicitation, or merely genomic intervals (start and end positions on chromosomes). By design, inputting a genomic interval invalidates a sequence similarity search, so SBSA does not invoke either FIMO or exact match in this scenario; instead, it seeks any overlapping between the variant-derived somatic sequences and the target genomic intervals, and annotates overlaid miRNA seeds if there are any. For user's convenience, SBSA has built in a diverse collection of motifs/binding sequences and accredited binding regions, which relate to TFs, RBPs, miRNA seeds, and miRNA-mRNA 3′-UTR targets. More details on these built-in motif/region datasets are provided in Materials and Methods.

While SBSA is capable of annotating thousands of input variants against thousands of regulators' motifs in a single session, we understand that users can be overwhelmed by an exceedingly large amount of results. To increase the manageability and validity of our annotation results, we implemented functional impact inference for variants, contextual expression quantification for TFs/RBPs, and conservation categorization for miRNAs. For input variants, we leveraged mature algorithms including SIFT (37), CADD (38) and MetaSVM (39) to assess the functional impact severity of the input variants. For TFs/RBPs, we obtained transcriptomes of GTEx tissues and sorted the expressed genes by their expression values, thereby returning the expression ranks of the associated molecular regulators. For miRNAs, we categorized them into four conservation categories (broadly conserved, conserved, poorly conserved, or 'other') according to the annotation done by TargetScan (40). In addition, from GTEx we downloaded fine-mapping results done by CAVIAR (41), CaVeMaN (42) and DAP-G (43), and annotated such posterior probability data for eQTL variants. Last but not the least, we leveraged the powerful tool ANNOVAR (44) to furnish the input variants with the most basic annotations. All such auxiliary annotation information can help guide a validity or creditability based prioritization of the immense annotation results.

The primary output from SBSA analysis takes the form of a comma-separated spreadsheet, where each row represents one somatic sequence of potential regulator-binding disruption. For each record, certain fields pertain to the somatic sequence only, such as variant location (chromosome, position), reference/somatic sequence, genomic features of the variant (host gene and the alteration effect to gene coding), and inferred functional impact significance. A particular set of fields are devoted to the coupling between the somatic sequence and the fetched molecular regulator, which includes the binary disruption effect (Gain/Loss), strandedness of the sequence matching, and sequence-motif similarity *P*-value (when FIMO is invoked). The last set of fields inform on identification of the plausible regulator (motif ID and gene symbol), contextual expression (for TFs/RBPs only), and conservation category (for miRNAs only).

In the simplest work mode, SBSA allows null input for motif/binding sequence specification, and consequentially outputs all somatic sequences derived from the input variants. For certain analysis scenarios, SBSA enables a secondary output file consisting of PWMs in format of Multiple Em for Motif Elicitation. To guard against a prolonged standby waiting time, we allow the user to leave an email address to receive a download link to the analysis result. Because SBSA performs diverse annotation modules for diverse types of input (Figure 1), which might entail distinct sets of parameters, the input files and advanced parameters are fed in a step-by-step gradient, and dynamic inactivation of irrelevant parameters is rendered based on the inputs at prior steps. To demonstrate major application contexts and output templates, we provide a few input examples with discrete sets of pre-populated inputs and parameters, so that demonstrative analyses can be readily exerted and representative results can be generated momentarily. Lastly, we rendered a comprehensive documentation that provides a detailed manual of all inputs and outputs.

Comparison with other tools

Many tools and studies have been dedicated to analyzing binding sequences of TFs. However, they are not purported for straightforward annotation. For example, FIMO (34) fits a PWM to a DNA/protein sequence through dynamic programming or alike techniques to obtain a *P*-value for assessing the TF's binding propensity to the concerned sequence. The input for FIMO comprises one part for DNA/Protein sequences and the other part for motifs. The output of FIMO is the location of likely binding sites

within the input sequences. Jayaram *et al.* (45) comprehensively evaluated ten prediction methods concerning transcription factor binding sites and four motif discovery tools. For motif discovery tools, the input comprises ChIP-seq data, and the output is a consensus PWM. For the transcription factor binding site prediction tools, the input is a ChIP-seq-derived PWM, and the output consists of the candidate transcription factor binding sites. Lee *et al.* (46) devised a computational model gkm-SVM to evaluate the impact of regulatory variants in DNA sequences. Similarly, another tool GERV (47) evaluates the effect of regulatory variants for transcription factor binding, tackling input of ChiP-seq data. In yet another work, Reshef *et al.* (48) proposed to investigate the signed effect of a SNP on transcription factor binding with the concept of polygenic disease risk.

Given the survey of related works above, we found three tools were most related to SBSA, namely SNP2TFBS (49), motifbreakR (35) and RSAT (36). SNP2TFBS is a database of SNPs that affect TF binding sequences. It does not enable novel discovery. motifbreakR can be used as a motif annotation tool, but the applicable variants are limited to only one type, namely SNP (identified with RefSNP ID). RSAT is a multi-function genome analysis suite; when its two specific functions are performed sequentially, the user can achieve annotation of variants with respect to motifs. The detailed comparison of the functionalities between SBSA and related tools is made in Table 1. For the purpose of annotating variants with respect to binding motifs/sequences incorporating variants, SBSA provides the most flexible and complete functionalities. With 37 source TF databases, RSAT supports the most abundant TF binding motifs; however, a large portion of the TF motifs are found exclusively in plant species. By contrast, SBSA supports 11 non-plant TF databases. The major advantage of SBSA lies in the support for miRNA seeds and miRNA-mRNA 3′-UTR binding targets, and the novel Personalized Genome approach (Figure 2B).

Runtime

The runtime of SBSA can range from seconds to minutes, and it scales with two parameters: number of variants and number of binding sequences. Processing 50 000 variants against the 2297 RBP binding sequences from ATtRACT took ∼20 minutes' runtime. Due to the potential long runtime, users can choose to be notified of the result download link via email. An overall evaluation of SBSA web server runtime can be seen in Supplementary Figure S1.

SBSA annotation demonstration

To demonstrate SBSA, we conducted analyses with the following five datasets (accessible on the SBSA website): (i) somatic mutation data from 10 182 subjects of 33 cancer types from TCGA, consisting of 33 variant files; (ii) somatic mutation data from 19 729 subjects of 57 cancer types from 81 projects within The ICGC, consisting of 81 variant files; (iii) 4.67 million A-to I RNA editing events in REDIportal, consisting of one variant file; (iv) 660.15 million SNPs in dbSNP 152 and (v) 71 478 479 *cis*-eQTL data from 980 subjects of 49 tissue sites in Genotype-Tissue Expression (GTEx), consisting of 49 variant files. Thorough analyses of these data using SBSA identified 1 255 503 863 consequential somatic binding sequences. All these identified mutations have been curated into our companion database SMDB (15) and can be queried and downloaded freely. Table 2 illustrates a few output examples of SBSA analysis using RNA editing from TCGA's BRCA cohort (Number of subjects: 942) against the four built-in RBP databases.

High-frequency annotation examples

Our annotation identified many known somatic binding sequences and many high-frequency novel results. The best proof-of-concept example is the well-established *TERT* promoter mutation that creates a new ETS1 protein binding motif. From the ICGC skin cancer (Australia) cohort, SBSA identified the somatic mutation C→T in *TERT* promoter at chromosome 5 position 1 295 135 with a frequency of 11.48%, which causes a gain of ETS1 binding sequence TTCCGG (Figure 3A). The function of this mutation in driving *TERT* expression has been studied in human cancer cell lines (6,12,13). We also conducted our own luciferase reporter experiment in melanoma cells to further validate the functional impact of this *TERT* mutation. Our data shows a roughly 7-fold increase in the promoter activity with the G228A mutation identified in *TERT* promoter relative to wild-type *TERT* (Figure 3B). This observation is consistent with previously published data (14) and suggests that the gain of an ETS binding site in the mutant promoter activates *TERT* gene expression.

Another similar yet novel example is the mutation C→T in *RPS20* promoter on chromosome 8 at position 56,074,582, with a frequency of 14.75% in the ICGC skin cancer (Australia) cohort. This mutation also potentially causes a gain of ETS1 binding sequence TTCCGG (Figure 3C). It is currently unknown whether this mutation has a similar functional effect on *RPS20* as the analogous mutation on *TERT*.

Regarding RBP regulation disruption, an excellent example entails the insertion of TT in *ANKRD33B* on chromosome 5 at position 10,634,463, with a frequency of 11.94% in the ICGC leiomyosarcoma (French) cohort. This mutation potentially causes a gain of PTBP1′s binding sequence (Figure 3D). PTBP1 is a well-known cancer-related RBP (50). For example, PTBP1 was observed to promote breast cancer cell growth by downregulating *PKM1*, a cancer suppressor (51). The SBSA-identified somatic sequence suggests potentially unexplored functional effects of *ANKRD33B*. As another example of RBP regulation disruption, the RNA editing event occurring in *POLR1E* on chromosome 9 at position 37 503 395 potentially causes a gain of ESRP2′s binding sequence (Figure 3E). This particular RNA editing event is ubiquitously observed in human cancers. For example, in TCGA's breast invasive carcinoma cohort, 941 of the 942 subjects tested have this RNA editing. ESRP2 is another cancer-related RBP (50) with known functionality such as sup-

Table 1. Functionality comparison between SBSA and similar tools

Reference	Work mode	Organism	TF database	RBP database	miRNA database	Custom motif sequence	Variant Type	Personalized Genome	Primary purpose
FIMO [1]	offline	any	NA	NA	NA	NA	NA	NA	Estimation of binding potential between a sequence and a motif
Jayaram et al [2]	offline	1	NA	NA	NA	NA	NA	NA	Detecting new TFs
deltaSVM [3]	offline	2	NA	NA	NA	NA	any	X	Estimating variant impact on binding sequence
SNP2TFBS [4]	online	1	1	NA	NA	X	SNP	X	Annotation of SNP affected motif
GERV [5]	offline	2	NA	NA	NA	NA	NA	NA	Estimating SNP impact on TF binding sequence
Reshef et al [6]	offline	1	NA	NA	NA	TF	NA	NA	Estimating SNP impact on TF binding sequence
motifbreakR [7]	offline	31	8	0	0	√	SNP	X	Annotation of SNP affected motif
RSAT [8]	online	72	37	2	0	√	any	X	Multiple functions. Can annotate any variants in any motifs
SBSA	online	any (26 pre loaded)	11	4	2	√	any	√	Annotation of any variants in any motifs/sequences

NA (not applicable) indicates the tool is not equipped with the referred functionality. For example, FIMO aims to compute the binding potential between a sequence and a motif, and it is not populated with a TF database. The √ sign denotes the support of the functionality. X sign denotes that the functionality is not supported.

Table 2. Example output from SBSA annotation results of altered RBP binding sequence resulting from RNA editing

Chr[1]	Location[2]	Gene[3]	RBP[4]	Edit[5]	Reference[6]	P ref[7]	Alternative[8]	P alt[9]	Effect[10]	Region[11]
chr13	52629800	HNRNPA1L2	ZFP36	A > G	aagaaagaAag	7.82×10^{-6}	aagaaagaGag	$>10^{-5}$	Loss	intronic
chr3	139355363	MRPS22	SNRPA	A > G	Aggaatgctg	$>10^{-5}$	Gggaatgctg	2.10×10^{-6}	Gain	intronic
chr6	43618165	POLH	ELAVL2	A > G	attAttttttttg	6.65×10^{-7}	attGttttttttg	$>10^{-5}$	Loss	UTR3
chr8	42287647	IKBKB	SNRPA	A > G	Aatacctgcta	$>10^{-5}$	Gatacctgcta	9.98×10^{-7}	Gain	intronic
chrX	48575833	RBM3	SFRS1	A > G	cAgacagagc	$>10^{-5}$	cGgacagagc	9.82×10^{-6}	Gain	intronic
chr17	43028783	RND2	HNRNPA1	A > G	tAgggcaggc	6.13×10^{-7}	tGgggcaggc	3.63×10^{-6}	Loss	UTR3
chr14	23325660	PABPN1	SFPQ	A > G	tggAaggac	3.31×10^{-6}	tggGaggac	6.03×10^{-6}	Loss	UTR3

The input for this analysis is the RNA editing information from TCGA's BRCA cohort ($N = 942$) against the four RBP databases. [1] Chromosome. [2] Chromosome coordinate position of the RNA editing event in GRCh38. [3] The host gene whose body encloses or approximates the RNA editing event. [4] The RBP whose binding sequence is disrupted by the RNA editing event. [5] Nucleotide alteration (A > G RNA editing). [6] Reference sequence surrounding the variant location reflecting the reference genome. [7] *P*-value for the approximity between reference sequence and the RBP's motif. [8] Altered sequence in parallel to Reference Sequence reflecting the variant in question (distinguished in uppercase). [9] *P*-value for the approximity between altered sequence and the RBP's motif. [10] Gain or loss of the plausible binding sequence resulting from the variant. [11] Classification of the genomic region enclosing the variant.

pressing cell motility in head and neck carcinoma cell lines (52) and driving alternative splicing patterns in prostate cancer (53).

Regarding repercussions in miRNA binding targets, a representative example entails the insertion of TC in the 3′-UTR of *SRSF7* on chromosome 2 at position 38 744 499, with a frequency of 26.87% in the ICGC leiomyosarcoma (French) cohort. This mutation causes an altered binding sequence for miR-409-3p (Figure 3F). Regarding consequential variants in miRNA seeds, a representative example entails the RNA editing in *miR-4477b*'s seed on chromosome 9 at position 63 819 626. This RNA-editing event leads to an altered miRNA seed (Figure 3G), and it occurs at a high population frequency—134 of the 154 (87.01%) subjects of Glioblastoma Multiforme cohort in TCGA were observed with this variant. Based on TargetScan (40) prediction, ∼66% mRNA targets of miR-4477b is susceptible to this altered seed. On average, the mRNA targets susceptible to variant-altered miRNA seeds account for 70% (ICGC) or 72% (TCGA) original mRNA targets. These results show that somatic mutations in miRNA seeds can lead to a substantial mRNA target shift. The biological effects of such mRNA target alterations have been demonstrated previously (9,10).

DISCUSSION

Identification of binding motifs for transcriptional regulators has been a fundamentally important topic in molecular biology, and for this sake numerous computational algorithms have been continually developed in the past several decades. The abundant algorithms were based on a variety of statistical frameworks, including dynamic programming (34), hidden Markov chain (54), deep learning (55), etc. Our web server SBSA does not aim to innovate in the methodological aspect of this important topic; rather, we developed SBSA as an easy-to-use online tool that streamlines large-scaled motif alteration annotations for genomic variants, leveraging a mature motif scanning approach FIMO or an intuitive exact sequence matching method. The identification of altered binding motifs resulting from major variant types such as somatic mutation and RNA editing has imminent scientific benefits. A myriad of studies have been conducted based on independent cases of such altered binding motifs/sequences (6,9–14). The cascading biological effects resulting from gain of an important binding sequence are relatively easier to observe than the effects of loss of a binding sequence. Because a binding sequence may have many targets, losing one may not cause a strong detrimental

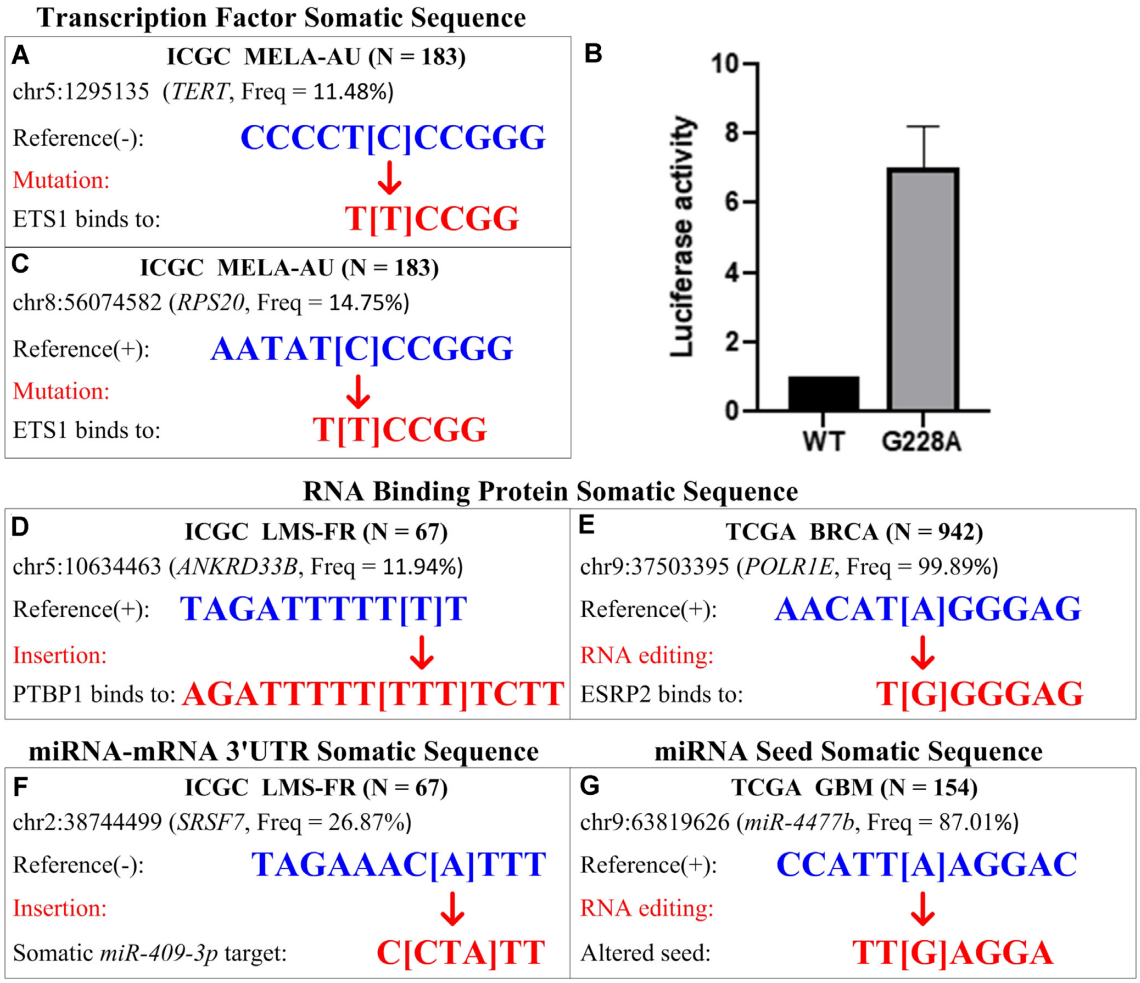

Figure 3. High-frequency examples from SBSA demonstrative analysis. Plus (+) and minus (–) signs denote the forward strand and the reverse strand of the GRCh38 reference genome, respectively. Blue and red nucleotide strings denote the non-altered reference sequences and the altered somatic sequences, respectively. (**A**) Example of TF ETS1 binding sequence formed by somatic mutation in *TERT* promoter region. (**B**) *In-vivo* experiments validated the expression activation function caused by the cryptic mutation G228A in *TERT* promoter. (**C**) Gain of ETS1 binding sequence due to somatic mutation in *RPS20* promoter region. (**D**) Gain of PTBP1 binding sequence due to insertion in *ANKRD33B*. (**E**) Gain of ESRP2 binding sequence due to RNA editing in *POLR1E*. (**F**) *hsa-miR-409–3p's* binding sequence in 3′-UTR of *SRSF7* mRNA is affected by an insertion. (**G**) Alteration of miRNA seed of *hsa-miRNA-4477b* due to RNA editing.

effect. However, some transcriptional effects can still be detected. For example, somatic mutations in the *SDHD* promotor region disrupted a ETS1 binding motif and significantly reduced *SDHD* gene expression (56). Using SBSA, we identified this motif loss in *SDHD,* with a frequency of 1.64% in the ICGC skin cancer (Australia) cohort. Annotation of variants using real somatic mutation, RNA editing, and SNP data from large consortiums revealed well-known somatic motifs as well as novel ones. Many of the novel altered motifs/sequences are of high frequencies, warranting follow-up studies to examine functional mechanisms in more detail. Furthermore, SBSA annotation of *cis*-eQTLs helps to explain the regulation mechanism of eQTLs. In our demonstrative analysis results, among the 1 875 338 *cis*-eQTLs, 1 354 071 (72.20%) caused at least one altered motif/sequence in miRNA seed regions and target sequences of TFs, RPBs, and miRNAs. These identified altered motifs/sequences might primarily or partially explain the eQTL regulation mechanisms.

SBSA has an enormous online calculation capability of analyzing genome-scaled input variants against vast motifs in a single session. SBSA also provides secondary functional annotations to enhance manageability and validity of the annotation results, which include functional impact inference for variants, contextual expression quantification for TFs/RBPs, fine-mapping information for eQTLs, and conservation categorization for miRNAs. Overall, we demonstrated the effectiveness of SBSA, a powerful tool that empowers the researchers to interrogate the functional effects of variants on binding motifs/ sequences in a wide range of species.

DATA AVAILABILITY

All omics data used in this study were from public repositories. The SBSA webserver can be accessed at http://innovebioinfo.com/Annotation/SBSA/SBSA.php.

The Python source code of SBSA is available at https://github.com/Limin-Jiang/SBSA.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Cancer Center Support Grant from National Cancer Institute [P30CA118100]; Bioinformatics, Biostatistics, and Analytical & Translational Genomics Shared Resources at the University of New Mexico Comprehensive Cancer Center; Y.G. was supported by National Cancer Institute Grant [R01ES030993-01A1]; L.J. and J.T. were supported by a grant from the National Natural Science Foundation of China (NSFC) [61772362, 61972280]; Shenzhen KQTD Project [KQTD20200820113106007]. Funding for open access charge: Institutional fund.
Conflict of interest statement. None declared.

REFERENCES

1. Chan,T.H., Lin,C.H., Qi,L., Fei,J., Li,Y., Yong,K.J., Liu,M., Song,Y., Chow,R.K., Ng,V.H. *et al.* (2014) A disrupted RNA editing balance mediated by ADARs (Adenosine DeAminases that act on RNA) in human hepatocellular carcinoma. *Gut*, **63**, 832–843.
2. Fu,L., Qin,Y.R., Ming,X.Y., Zuo,X.B., Diao,Y.W., Zhang,L.Y., Ai,J., Liu,B.L., Huang,T.X., Cao,T.T. *et al.* (2017) RNA editing of SLC22A3 drives early tumor invasion and metastasis in familial esophageal cancer. *PNAS*, **114**, E4631–E4640.
3. Maas,S., Patt,S., Schrey,M. and Rich,A. (2001) Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *PNAS*, **98**, 14687–14692.
4. Chen,Y.B., Liao,X.Y., Zhang,J.B., Wang,F., Qin,H.D., Zhang,L., Shugart,Y.Y., Zeng,Y.X. and Jia,W.H. (2017) ADAR2 functions as a tumor suppressor via editing IGFBP7 in esophageal squamous cell carcinoma. *Int. J. Oncol.*, **50**, 622–630.
5. Li,Q.Y., Seo,J.H., Stranger,B., McKenna,A., Pe'er,I., LaFramboise,T., Brown,M., Tyekucheva,S. and Freedman,M.L. (2013) Integrative eQTL-Based analyses reveal the biology of breast cancer risk loci. *Cell*, **152**, 633–641.
6. Chiba,K., Lorbeer,F.K., Shain,A.H., McSwiggen,D.T., Schruf,E., Oh,A., Ryu,J., Darzacq,X., Bastian,B.C. and Hockemeyer,D. (2017) Mutations in the promoter of the telomerase gene TERT contribute to tumorigenesis by a two-step mechanism. *Science*, **357**, 1416–1420.
7. Kim,M.Y., Hur,J. and Jeong,S. (2009) Emerging roles of RNA and RNA-binding protein network in cancer cells. *BMB Rep*, **42**, 125–130.
8. Neelamraju,Y., Gonzalez-Perez,A., Bhat-Nakshatri,P., Nakshatri,H. and Janga,S.C. (2018) Mutational landscape of RNA-binding proteins in human cancers. *RNA Biol*, **15**, 115–129.
9. Mencia,A., Modamio-Hoybjor,S., Redshaw,N., Morin,M., Mayo-Merino,F., Olavarrieta,L., Aguirre,L.A., del Castillo,I., Steel,K.P., Dalmay,T. *et al.* (2009) Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat. Genet.*, **41**, 609–613.
10. Iliff,B.W., Riazuddin,S.A. and Gottsch,J.D. (2012) A Single-Base substitution in the seed region of miR-184 causes EDICT syndrome. *Invest. Ophth. Vis. Sci.*, **53**, 348–353.
11. Bhattacharya,A. and Cui,Y. (2016) SomamiR 2.0: a database of cancer somatic mutations altering microRNA-ceRNA interactions. *Nucleic Acids Res.*, **44**, D1005–1010.
12. Li,X.J., Qian,X., Wang,B., Xia,Y., Zheng,Y.H., Du,L.Y., Xu,D.Q., Xing,D.M., DePinho,R.A. and Lu,Z.M. (2020) Programmable base editing of mutated TERT promoter inhibits brain tumour growth. *Nat. Cell Biol.*, **22**, 282–288.
13. 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.
14. Bell,R.J., Rube,H.T., Kreig,A., Mancini,A., Fouse,S.D., Nagarajan,R.P., Choi,S., Hong,C., He,D., Pekmezci,M. *et al.* (2015) Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science*, **348**, 1036–1039.
15. Jiang,L., Duan,M., Guo,F., Tang,J., Oybamiji,O., Yu,H., Ness,S., Zhao,Y.-Y., Mao,P. and Guo,Y. (2020) SMDB: pivotal somatic sequence alterations reprogramming regulatory cascades. *NAR Cancer*, **2**, zcaa030.
16. Contreras-Moreira,B. (2009) 3D-footprint: a database for the structural analysis of protein–DNA complexes. *Nucleic Acids Res.*, **38**, D91–D97.
17. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
18. Yin,Y., Morgunova,E., Jolma,A., Kaasinen,E., Sahu,B., Khund-Sayeed,S., Das,P.K., Kivioja,T., Dave,K., Zhong,F. *et al.* (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, **356**, eaaj2239.
19. Najafabadi,H.S., Mnaimneh,S., Schmitges,F.W., Garton,M., Lam,K.N., Yang,A., Albu,M., Weirauch,M.T., Radovani,E., Kim,P.M. *et al.* (2015) C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.*, **33**, 555–562.
20. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
21. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.
22. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
23. Isakova,A., Groux,R., Imbeault,M., Rainer,P., Alpern,D., Dainese,R., Ambrosini,G., Trono,D., Bucher,P. and Deplancke,B. (2017) SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods*, **14**, 316–322.
24. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
25. Contreras-Moreira,B. and Sebastian,A. (2016) FootprintDB: analysis of plant Cis-Regulatory elements, transcription factors, and binding interfaces. *Methods Mol. Biol.*, **1482**, 259–277.
26. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
27. Giudice,G., Sanchez-Cabo,F., Torroja,C. and Lara-Pezzi,E. (2016) ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)*, **2016**, baw035.
28. Benoit Bouvrette,L.P., Bovaird,S., Blanchette,M. and Lecuyer,E. (2020) oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic. Acids. Res.*, **48**, D166–D173.
29. Berglund,A.C., Sjolund,E., Ostlund,G. and Sonnhammer,E.L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.*, **36**, D263–266.
30. Paz,I., Kosti,I., Ares,M. Jr, Cline,M. and Mandel-Gutfreund,Y. (2014) RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **42**, W361–W367.
31. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
32. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**, D92–D97.

33. Lo Giudice,C., Tangaro,M.A., Pesole,G. and Picardi,E. (2020) Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. *Nat. Protoc.*, **15**, 1098–1131.
34. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
35. Coetzee,S.G., Coetzee,G.A. and Hazelett,D.J. (2015) motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics*, **31**, 3847–3849.
36. Santana-Garcia,W., Rocha-Acevedo,M., Ramirez-Navarro,L., Mbouamboua,Y., Thieffry,D., Thomas-Chollier,M., Contreras-Moreira,B., van Helden,J. and Medina-Rivera,A. (2019) RSAT variation-tools: an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. *Comput. Struct. Biotechnol. J.*, **17**, 1415–1428.
37. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
38. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
39. Dong,C.L., Wei,P., Jian,X.Q., Gibbs,R., Boerwinkle,E., Wang,K. and Liu,X.M. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137.
40. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
41. Hormozdiari,F., Kostem,E., Kang,E.Y., Pasaniuc,B. and Eskin,E. (2014) Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**, 497–508.
42. Brown,A.A., Vinuela,A., Delaneau,O., Spector,T.D., Small,K.S. and Dermitzakis,E.T. (2017) Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.*, **49**, 1747–1751.
43. Wen,X., Pique-Regi,R. and Luca,F. (2017) Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLos Genet.*, **13**, e1006646.
44. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
45. Jayaram,N., Usvyat,D. and Martin,A.C.R. (2016) Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, **17**, 547.
46. Lee,D., Gorkin,D.U., Baker,M., Strober,B.J., Asoni,A.L., McCallion,A.S. and Beer,M.A. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.
47. Zeng,H., Hashimoto,T., Kang,D.D. and Gifford,D.K. (2016) GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*, **32**, 490–496.
48. Reshef,Y.A., Finucane,H.K., Kelley,D.R., Gusev,A., Kotliar,D., Ulirsch,J.C., Hormozdiari,F., Nasser,J., O'Connor,L., van de Geijn,B. *et al.* (2018) Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.*, **50**, 1483–1493.
49. Kumar,S., Ambrosini,G. and Bucher,P. (2017) SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.*, **45**, D139–D144.
50. Wang,Z.L., Li,B., Luo,Y.X., Lin,Q., Liu,S.R., Zhang,X.Q., Zhou,H., Yang,J.H. and Qu,L.H. (2018) Comprehensive genomic characterization of RNA-Binding proteins across human cancers. *Cell Rep.*, **22**, 286–298.
51. He,X., Arslan,A.D., Ho,T.T., Yuan,C., Stampfer,M.R. and Beck,W.T. (2014) Involvement of polypyrimidine tract-binding protein (PTBP1) in maintaining breast cancer cell growth and malignant properties. *Oncogenesis*, **3**, e84.
52. Ishii,H., Saitoh,M., Sakamoto,K., Kondo,T., Katoh,R., Tanaka,S., Motizuki,M., Masuyama,K. and Miyazawa,K. (2014) Epithelial splicing regulatory proteins 1 (ESRP1) and 2 (ESRP2) suppress cancer cell motility via different mechanisms. *J. Biol. Chem.*, **289**, 27386–27399.
53. Munkley,J., Li,L., Krishnan,S.R.G., Hysenaj,G., Scott,E., Dalgliesh,C., Oo,H.Z., Maia,T.M., Cheung,K., Ehrmann,I. *et al.* (2019) Androgen-regulated transcription of ESRP2 drives alternative splicing patterns in prostate cancer. *Elife*, **8**, e47678.
54. Wheeler,T.J., Clements,J., Eddy,S.R., Hubley,R., Jones,T.A., Jurka,J., Smit,A.F. and Finn,R.D. (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, **41**, D70–D82.
55. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
56. Weinhold,N., Jacobsen,A., Schultz,N., Sander,C. and Lee,W. (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.*, **46**, 1160–1165.