Methods 79–80 (2015) 32–40

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Current trend of annotating single nucleotide variation in humans – A case study on SNVrap

Mulin Jun Li, Junwen Wang *

Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China
Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China
Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, China

ARTICLE INFO

Article history:
Received 21 April 2014
Received in revised form 25 September 2014
Accepted 2 October 2014
Available online 13 October 2014

Keywords:
Single nucleotide variation
Next generation sequencing
Functional annotation
Functional prediction
Web server

ABSTRACT

As high throughput methods, such as whole genome genotyping arrays, whole exome sequencing (WES) and whole genome sequencing (WGS), have detected huge amounts of genetic variants associated with human diseases, function annotation of these variants is an indispensable step in understanding disease etiology. Large-scale functional genomics projects, such as The ENCODE Project and Roadmap Epigenomics Project, provide genome-wide profiling of functional elements across different human cell types and tissues. With the urgent demands for identification of disease-causal variants, comprehensive and easy-to-use annotation tool is highly in demand. Here we review and discuss current progress and trend of the variant annotation field. Furthermore, we introduce a comprehensive web portal for annotating human genetic variants. We use gene-based features and the latest functional genomics datasets to annotate single nucleotide variation (SNVs) in human, at whole genome scale. We further apply several function prediction algorithms to annotate SNVs that might affect different biological processes, including transcriptional gene regulation, alternative splicing, post-transcriptional regulation, translation and post-translational modifications. The SNVrap web portal is freely available at http://jjwanglab.org/snvrap.

1. Introduction

Recent advances in Next generation sequencing (NGS) have significantly facilitated the discoveries in various fields of biology especially in human genetics and genomics. The International HapMap Project and 1000 Genomes Project have revealed many common, rare and *de novo* human genetic variations (including single nucleotide polymorphisms (SNPs), insertions and deletions (Indels) and copy number variations (CNVs)) on over thousands individuals from different populations [1,2]. It is an urgent task to predict the functional impact and disease susceptibility of these variants by computational methods. To comprehensively annotate the variants in the genome-wide scale, we need sufficient knowledge in the entire biological process which spans from gene transcription to post translational modification. The newly emerging genomic, epigenomic and proteomic data, generated by the ENCODE Project, the Roadmap Epigenomics Project and The Human Proteome Project, provide an unprecedented opportunity for us to deeply annotate these genetic variants [3–5].

Function annotation of genetic variants is an old but important topic since the birth of molecular genetics. There are many outstanding bioinformatics tools and resources developed in the last decades. In the post-genomic era, ANNOVAR is the first batch of variant annotation tool for the large-scale NGS-based data. It incorporates many useful gene-based features and human genetic information [6]. Tools such as VEP [7], SeattleSeq [8], NGS–SNP [9], AnnTools [10], SVA [11], TREAT [12], SnpEff [13], VARIANT [14], GWASrap [15], VAT [16], GEMINI [17] and AVIA [18] were developed to annotate genetic variants in standalone applications or on web servers. These tools contain both general features and distinct functions in pre-processing, filtering and annotating variants [19]. However, the annotations collected by these tools largely focus on variant attributes (such as synonymous or non-synonymous) and genomic features surrounding each target variant, and usually offer limited information regarding the function consequence of the variants. The genetic variants have been shown to be involved in almost all aspects of gene regulation, from transcription to post-translation [20,21], but studies mostly concentrate on the functional effect of non-synonymous mutations that directly change protein function. However, other types of variants can also

* Corresponding author at: Room 1-05E, The Hong Kong Jockey Club Building for Interdisciplinary Research, 5 Sassoon Road, Pokfulam, Hong Kong Special Administrative Region, China. Fax: +852 28551254.
E-mail address: junwen@hku.hk (J. Wang).

http://dx.doi.org/10.1016/j.ymeth.2014.10.003
1046-2023/

be deleterious and pathogenic in disease etiology and evolution. Recently algorithms were developed to predict their functional effects for many other variant types. Therefore, incorporating those information into variant annotation can greatly facilitate the understanding of disease-causal variants in full spectrum. In addition, many previous tools require command line operation and relatively long execution time, and only provide plain text or simple interface to present the results. This may impede instant knowledge acquisition and the interpretation of variant function in an intuitive manner. Furthermore, there is no systematic survey for available tools in the functional prediction of genetic variants across multiple biological processes.

Here, we first review the current progress and trend of the genetic variants annotation field. We then introduce a versatile and interactive annotation web portal, SNVrap (http://jjwang-lab.org/snvrap) for human genetic variations, with emphasis on the function annotation of single nucleotide variation (SNV). We incorporated a large number of gene-based features and the latest genomic/epigenomic datasets to annotate all human SNVs. We also applied functional prediction algorithms to calculate the functional scores as well as related interpretations for the variants that affect different biological processes.

2. The classification of variants annotation and prediction

Many genetic, genomic and epigenomic information can be used for variants annotation. According to the features adopted in existed annotation tools, we classified the current methods into gene-based annotation, knowledge-based annotation and function prediction.

2.1. Gene-based annotation

The variant types and the surrounding genomic elements are the most direct and useful information for interpreting the underlying biological function of the investigated variants. Conventionally, researchers utilized RefGene (or KnownGene and Ensembl Gene) as the gold standard to locate the variant, indicating whether the investigated variant is resided in a gene body with the potential to disrupt the protein sequence and therefore affecting its function. Attentions are specially focused on the non-synonymous mutation altering the protein sequence and the splice mutation disrupting the transcript's splicing pattern. Recent RNA-Seq technology boosts the discovery of new splicing events and novel transcripts [22], and the GENCODE project annotates all evidence-based gene features on entire human genome at high accuracy [23]. This advance led to 18% increase of non-synonymous variants compared with RefGene annotation by the latest statistics [24]. In addition, annotations for many important non-coding RNAs (such as long non-coding RNA (lncRNA) and microRNA (miRNA)) are also new additions in gene-based annotation.

2.2. Knowledge-based annotation

Other than gene attributes, recent in-depth annotation mostly focuses on protein function and its metabolism. Researchers pay close attention to the genetic variants that disturb the protein function domain, protein–protein interaction and biological pathways. Universal Protein Resource (UniProt) is usually adopted to annotate protein sequence and functional information in non-synonymous SNV (nsSNV) locus [25]. However, the scope of functional elements in the human genome is continuously enlarged with the evolving understanding of human DNA. The non-coding regions in the human genome contain many important regulatory elements including promoter, enhancer and insulator [23]. The DNA mutations can also change the RNA sequence and then influence the RNA secondary structure [26,27], RNA-binding protein recognition [28], miRNA binding activity [29], etc.

Recently, the NGS coupled technologies greatly facilitate the discovery of those genome-wide novel functional elements. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) as an effective method for analyzing protein's interaction with DNA, has been frequently used to investigate the genome-wide binding pattern of a specific transcription factor (TF), histone mark and important regulator [30]. The Encyclopedia of DNA Elements (ENCODE) project has performed over 500 ChIP-Seq experiments of hundreds TFs across a number of human cell lines [3,31]. These data provide unprecedented resources to study TF's dynamic activities, particularly in considering the specificity of TF binding in different cell types as well as in specific condition. Also, chromatin marks (like H3K4me1, H3K27ac and H3k27me3) are frequently used to pinpoint the distinct functional elements (such as promoter, enhancer, silencer and other distal *cis*-acting regulatory elements) in different cell lines, which present active or repressive transcription activities in euchromatin [32,33]. Active chromatin captured by DNase I hypersensitive sites (DHSs) sequencing usually exposes the DNA and produces accessible chromatin zones that are functionally relate to transcriptional activity [34]. There are urgent need of an integrative resource to collect and organize aforementioned new knowledge, especially for the regions that contain both regulatory sites and functional signals.

Fortunately, several databases and web servers were recently developed to collect these evolving genomic data. HaploReg [35] explores annotations of variant in the non-coding human genome on haplotype blocks, such as candidate regulatory SNVs at disease-associated loci. It predicts chromatin state and variant effect on regulatory motifs by using ENCODE data. RegulomeDB [36] includes high-throughput, experimental datasets from ENCODE and other projects, and uses signal-combined scores to identify putatively regulatory potential of functional variants. GWAS3D [37] recruits cell-type specific genomic data, including histone modification, long range interaction and ChIP-Seq motif, to detect and prioritize regulatory SNVs especially these in the enhancer regions. rSNPBase [38] provides large-scale genomic mapping for human regulatory SNVs in a wide range of regulation types including proximal and distal transcriptional regulation and post-transcriptional regulation. These resources significantly promote the discovery of regulatory variants by information integration, and greatly enlarge the searching of functional variants outside the gene body.

2.3. Functional prediction

Aforementioned annotation methods mainly consider whether the variant loci are in the known functional units that harbor genomic or epigenomic signals. But simply linking variant function by a shared genomic location is not enough, researchers expect to know the exact function and the significance that variants exert. The functions of genetic variants are extensive in terms of the affected genomic region, and they can involve in almost all processes of gene regulation, from transcriptional to post-translational level [20,21] (Fig. 1). Here, we briefly introduce current studies on different function domains and focus on the bioinformatics tools for predicting the variant function in each domain.

2.3.1. Transcriptional gene regulation

Gene transcription is an elaborate process governed by many spatial and temporal factors in the nucleus such as global or local chromatin states, nucleosome positioning, TF binding, enhancer/promoter activities. Regulatory SNPs altering any one of these biological processes may change gene regulation and result in phenotypic

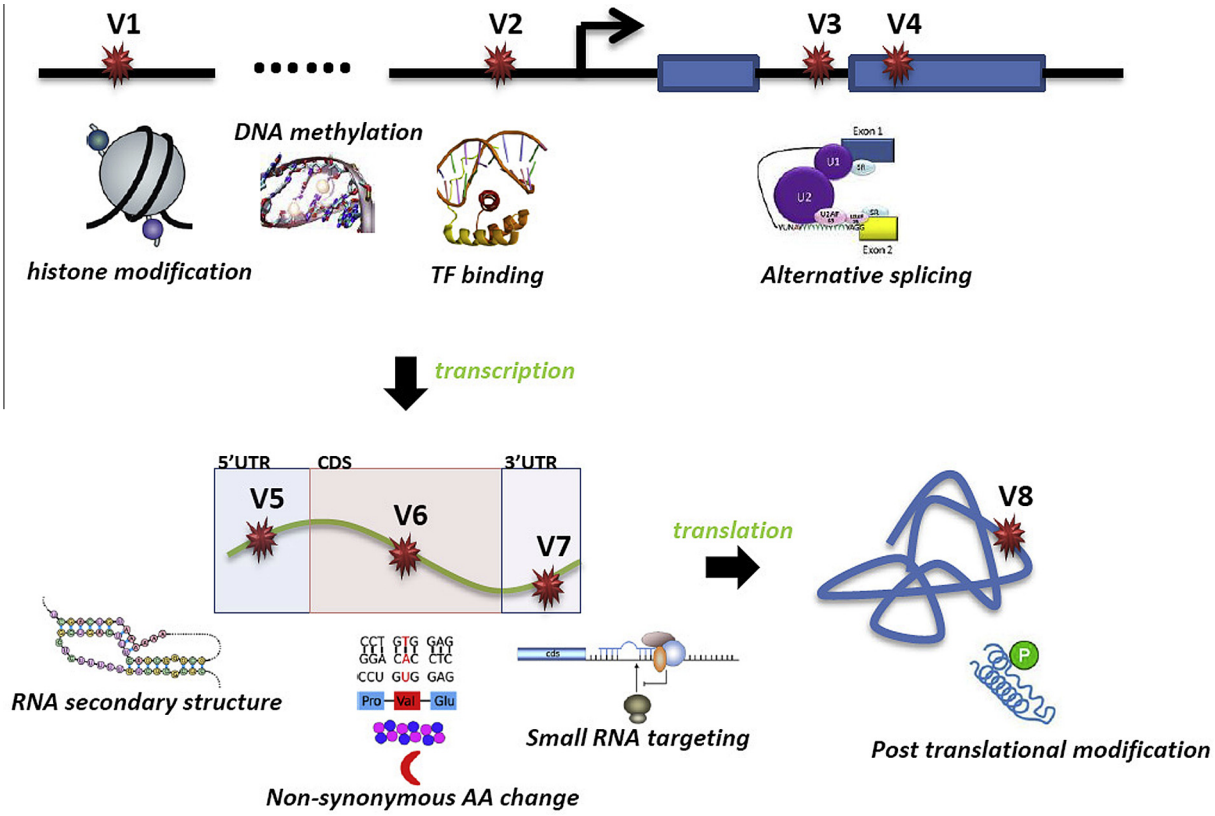

Fig. 1. The functional domains (from transcriptional gene regulation to post-translational modification) that can be disrupted by variants. (V1) the variant locates in the distal *cis*-acting regulatory elements such as enhancer and silencer, which may affect the activities of TF binding, nucleosome positioning, histone modification; (V2) the variant locates in the promoter region of gene, which may affect the transcription initiation; (V3) the variant locates in the intronic region of gene, which may affect the function of intronic splicing site; (V4) the variant locates in the exonic region of gene, which may affect the function of exonic splicing elements; (V5) the variant locates in the 5′UTR of mRNA, which may change the secondary structure of mRNA and affect translation initiation. (V6) the variant locates in the protein coding region of mRNA, which may change the codon and result in the substitution of amino acid or abnormal translation efficiency; (V7) the variant locates in the 3′UTR of mRNA, which may affect the interaction with miRNAs; (V8) the variant locates in the post-translational modification site of protein, which may influence the effect size of post-translational modification.

abnormality [39,40]. Variant can repress or promote transcription initiation and the assembly of transcriptional factory by directly affecting the binding of critical activators, repressors and other transcriptional units [41–44]. Genetic variants located in distal regulatory region can affect the binding motif of TFs, chromatin regulators and other distal transcriptional factors, which disturb the interaction between enhancer/silencer and its target gene [45–47].

Many diseases and traits can be attributed to the allele-specific TF binding and sequence variations in the DNA functional elements (include promoters, enhancers, silencers and insulators), which can either disrupt or create a TF-DNA binding event [48,49]. To quantitatively measure the variant's ability on affecting TF-DNA bindings (gain or loss) between different alleles, bioinformatics tools have been developed to calculate the difference of binding affinity given DNA sequences contain different alleles and to further measure the significance by log-odds of binding probabilities under different statistical models [50–52]. A recent review systematically summarized approaches for identifying human regulatory variants affecting gene expression [40]. Although the role of regulatory variants is far more complicated, current bioinformatics tools mainly focus on variant effect on TFBS. We listed some recent tools to predict the functional consequence of regulatory SNVs in Table 1.

2.3.2. Alternative splicing

Another type of variants that are intensively investigated locate near the splicing sites in either exon or intron. They may directly affect the splicing events and result in aberrant transcript isoform abundance by altering the landscape of *cis*-acting elements including exonic splicing enhancers (ESE), intronic splicing enhancers (ISE), exonic splicing silencers (ESS) and intronic splicing silencers (ISS) [53–55]. Different splicing events, such as exon skipping, 3′ or 5′ exonic splicing and intron retention, may produce abnormal transcripts and affect the original biological function. To predict the functional consequence of variants in the splice sites, some bioinformatics tools were developed to utilize the sequence information of splicing *cis*-acting elements under the motif scanning or build the statistics model using various genomic attributes (Table 1).

2.3.3. RNA processing and post-transcriptional regulation

Mutations in the untranslated region have been reported to affect many post-transcriptional regulations. Distinctive structural features are required for many RNA molecules and *cis*-acting regulatory elements to execute effective functions during gene regulation. SNVs can alter the secondary structure of RNA molecules and then disrupt the proper folding of RNAs, such as tRNA/mRNA/lncRNA folding [56–58] and miRNA binding recognition regions [59]. There are quite a lot prediction tools for estimating the effect of SNV in RNA molecules, and we summarized some of these tools in Table 1.

2.3.4. Translation and post-translational modifications

SNV can also affect the *cis*-acting regulatory elements in mRNA's 5′UTR to inhibit/promote the translation initiation [60]. Mutations that change synonymous codons may affect the translation

Table 1
The recent tools for variant functional prediction in different biological domains.

Function domain	Tools	Type	Prediction method and description
Transcriptional gene regulation	RAVEN [119]	Database	Directly scoring the difference of position weight matrix (PWM)
	is-rSNP [120]	Web server	The binding affinity scoring and Fourier transform for significance measurement
	TRAP [121]	Web server	Calculating the distribution PWM scores via direct computation of convolution
Alternative splicing	Human Splicing Finder [122]	Web server	Scoring the motifs of splicing site
	Skippy [96]	Software	Modeling the gain or loss effect of SNP on large numbers of ESE and ESS site with evolutionary constraint
	MutPred Splice [97]	Software	Trained model with large features to predict all exonic nucleotide substitutions that disrupt pre-mRNA splicing
	SilVA [101]	Software	Random forest model to predict the deleterious synonymous mutation, mostly for splicing effect
miRNA targeting	dbSMR [123]	Database	Measuring degree of change of SNP-caused intramolecular structure change at the particular target site
	miRNASNP [124]	Database	Counting the loss and gain for the change of predicted targets that affected by SNP
	Mirsnpscore [125]	Database	Predicting the miRNA–mRNA interaction score by two-step SVM classifier and measuring the differences of haplotypes effect
	PolymiRTS [100]	Database	Scoring the difference of context score by Targetscan [126]
RNA structure	Rchange [127]	Software	Computing the changes of the ensemble free energy, mean energy and the thermodynamic entropy of RNA secondary structure for mutation
	RNAsnp [114]	Web server	Calculating local regions of maximal structural change by a screening mode
Protein function (only list combined methods)	dbNSFP [24]	Database	Computing deleterious scores and annotations for all nsSNVs using seven prevalent prediction tools
	Condel [70]	Software	Weighted average of the normalized scores of five tools
	KGGSeq [71]	Software	Logit model to iteratively combine the best score using 13 tools
	eXtasy [72]	Software	Random forest classifier to predict deleteriousness the by combing multiple annotations
	SPRING [73]	Software	Fisher's combined probability test to integrate 6 tools and association scores derived from a variety of genomic data sources
Post-translational modification	PhosSNP [102]	Database	Scoring different SNP effect on kinase-specific phosphorylation site

efficiency due to codon usage biases [61]. The translation elongation can also be retarded by mutations along the ramp of ribosomal movement [62]. In the post-translational level, genetic variants can contribute to proteostasis [63] and amino acid modifications [64]. However, mechanisms of variant effect in this field are complicated and there are only a few tools available to predict variant's effect on translation related modifications (Table 1).

2.3.5. Protein function

The direct impacts of the non-synonymous variants on protein functions have been extensively investigated, and many algorithms have been developed to predict the deleteriousness and pathogenesis of nsSNVs [65]. Classical bioinformatics tools, such as SIFT [66], Polyphen [67] and MutationTaster [68], successfully predict the functional consequence of non-synonymous substitution. The latest version of dbNSFP contain the deleterious scores and annotations for all nsSNVs using seven prevalent prediction tools [24]. It is notable that prediction results of different algorithms are not always consistent, which complicates the usage of deleterious nsSNVs prediction for prioritization [69]. To overcome this defect, some up-to-date statistical methods combined the deleterious score from different tools to improve the sensitivity and specificity [70–73] (Table 1).

2.3.6. Evolutionary conservation and natural selection

Comparative genomics approaches were used to predict the function-relevant variants under the assumption that the functional genetic locus should be conserved across different species at an extensive phylogenetic distance. To this end, researchers have used evolutionarily conserved scores, like 64-way vertebrate alignment score or phastCon score, as the putative benchmark to identify genomic regions that may have biological importance, even if the functional annotation of these regions is unknown [74–76]. Furthermore, base-wise scores are adopted by GERP++ and PhyloP to detect conservation at a higher resolution measuring the exact evolutionary signature of SNV site [77,78]. On the other hand, some adaptive traits and the population differences are driven by positive selections of advantageous variants, and these genetic mutations are functionally relevant to population specific phenotypes. This make the positive selection signals, such as F_{ST} [79], Tajima's D [80] and iHS [81], useful to pinpoint the functional genetic variants that shape the population specific traits. In this level, dbPSHP [82] and 1000 Genomes Selection Browser [83] offer comprehensive selection scores for each known SNV.

Functional prediction of variants' effect in different biological processes is pivotal to pinpoint the molecular mechanism of diseases/traits and direct the experimental validation. To this end, we introduce a comprehensive web tool that provides fast and convenient annotation service for all of human SNVs, especially for their functions.

3. Materials and methods of the SNVrap web server

3.1. Data structure overview

We used the latest Ensembl Gene annotation (version 68) to catalog the variant attributes and classify variants to eleven different levels (including Intergenic, Downstream, Upstream, 3prime_utr, 5prime_utr, Intronic, Synonymous, Non_synonymous, Splice_site, Stop_gained and Stop_lost). Although the web portal focuses on SNV due to restriction of functional prediction of large size variants, we still can annotate Indels and CNVs with attribute (Frameshift and Inframe) as well as general genomic features for some known variants. We utilized dbSNP138 and the 1000 Genomes Project genotype and haplotype data for population specific annotation. For general gene-based genomic elements (such as

gene transcripts, non-coding RNA, validated enhancer/insulator) and disease-causal elements, we downloaded the data table from UCSC Genome Browser and other public resources (like OMIM [84], GAD [85], COSMIC [86], ClinVar [87], etc.). In addition, we processed the genomic and epigenomic data from ENCODE, ChromHMM [15] and seaway [88] to provide knowledge-based transcriptional annotation. Importantly, we pre-computed and incorporated the functional prediction scores and corresponding interpretation for SNVs that putatively alter transcriptional gene regulation, alternative splicing, post-transcriptional regulation, translational initiation and elongation, post-transcriptional modification by either internal or published tools (see Section 3.2 for data processing). We summarized all original or processed annotation data in our server. Users can download these data by fixed links in Table 2.

3.2. Web portal implementation and data processing

The SNVrap web server is constructed by Perl and Catalyst web framework. A backend MySQL database hosts most of gene-based annotation data. We also use plain text for some annotations, which are indexed and searchable by Tabix [89]. The overview of SNVrap architecture is shown in Fig. 2.

3.2.1. Genomic and epigenomic annotations

To efficiently incorporate the genomic and epigenomic data from ENCODE related projects, we collected ChIP-Seq peaks for transcription factor bind sites (TFBSs) and their related TF motifs to annotate non-coding regulatory SNVs that locate in important DNA functional elements. We also incorporated important enhancer marks (H3K4me1, H3K27ac, EP300), insulator marks (CTCF) and DHSs data for active chromatin into the annotation. Due to the data restriction for different cell type in ENCODE project, we only compile several well-annotated cell lines in current version, but other cell lines will be added when sufficient data are available. Second, to consider SNV function in long distance regulation, we collected DNA interaction data obtained from 5C, Hi-C and ChIA-PET experiments. We used very high resolution (10 Kb) to call genome-wide interaction data of Hi-C by ICE [90], which can improve the identification of variant that disrupts the chromatin structure.

Table 2
Sources of used dataset and tools in SNVrap annotation.

Type	Item	Description	Link
Variant annotation	HapMap I + II + III	HapMap phase 3 release version 2	ftp://ftp.ncbi.nlm.nih.gov/hapmap/phasing/2009-02_phaseIII/HapMap3_r2
	1000 Genomes	1000 Genomes Project phase 1 integrated release version 3	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521
	dbSNP 138	Genetic variant which collected in dbSNP 138 version	ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data
Gene-based annotation	Reference gene	Gene annotation from NCBI Refseq	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/refGene.txt.gz
	Ensemble gene	Gene annotation from ensemble	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/ensGene.txt.gz
	Known gene	Gene annotation from UCSC	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownGene.txt.gz
	GENCODE	Gene annotation from GENCODE V12	ftp://ftp.sanger.ac.uk/pub/gencode/release_12/gencode.v12.annotation.gtf.gz
	Small RNA	snoRNA and miRNA annotations from UCSC	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/wgRna.txt.gz
	MicroRNA Target	TargetScan generated miRNA target site predictions	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/targetScanS.txt.gz
Knowledge-based annotation	Transcriptional factor binding site	Transcription factor binding sites conserved in the human/mouse/rat alignment	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/tfbsConsSites.txt.gz
	Enhancer	Validated and predicted human enhancer	http://jjwanglab.org/jbrowse/data/human_eatdb/raw/enhancer_sorted.gff3
	Insulator	CTCF binding site database for characterization of human genomic insulators (CTCFBSDB)	http://insulatordb.uthsc.edu/download/CTCFBSDB_all_exp_sites_Sept12_2012.txt.gz
	HapMap eQTL	eQTLs of meta-analysis (consensus eQTLs) from HapMap human lymphoblastoid cell lines	http://www.bios.unc.edu/research/genomic_software/seeQTL/data/eQTL_Qvalue_cutoff_hapmap3_cis_hg19.txt http://www.bios.unc.edu/research/genomic_software/seeQTL/data/eQTL_Qvalue_cutoff_hapmap3_trans_hg19.txt
	GTEx eQTL	GTEx eQTL data for different human tissues	http://147.8.193.64/SNVrap/GTEx-eQTL.zip
	ENCODE factors	Epigenomic signals from ENCODE project	http://147.8.193.64/SNVrap/ENCODE.annotation.gz
	ENCODE TFBS	Transcription factors binding site from ENCODE ChIP-seq	http://147.8.193.64/SNVrap/ENCODE.tf.bound.union.bed.sort.gz
	Chromosome interaction	10 Kb resolution Hi-C, ChIA-PET and 5C data for ENCODE tier 1 cell lines	http://147.8.193.64/SNVrap/3D_ALL.gz
Functional prediction	Transcriptional factor binding site affinity	Prediction for the possible TFs binding affinity change caused by SNV	http://jjwanglab.org/gwas3d
	MicroRNA target site affinity	Prediction for the possible miRNA-Target binding affinity change caused by SNV (PolymiRTS)	http://147.8.193.64/SNVrap/target_miRSNP_human_PolymiRTS3.sort.txt.gz
	Splicing site affinity	Prediction for the possible splicing disruption caused by SNV (Skippy and MutPred Splice)	http://147.8.193.64/SNVrap/skippy.all.splice.sort.txt.gz http://147.8.193.64/SNVrap/MutPredSpliceBatchJob_Mulin.sort.csv.gz
	Non-synonymous SNV functional prediction	Prediction for the protein damaging caused by non-synonymous mutation (dbNSFP and KGGSeq)	http://dbnsfp.houstonbioinformatics.org/dbNSFPzip/dbNSFPv2.6.zip http://147.8.193.64/SNVrap/kggseq/kggseq.combined.score.kggseq.missense.txt.gz
	Synonymous SNV functional prediction	Prediction for the deleterious synonymous mutation (Silva)	http://147.8.193.64/SNVrap/silva-pathogenic.sort.txt.gz
	Phosphorylation functional prediction	Prediction for affection of kinase-specific phosphorylation site caused by non-synonymous mutation (PhosSNP)	http://phossnp.biocuckoo.org/

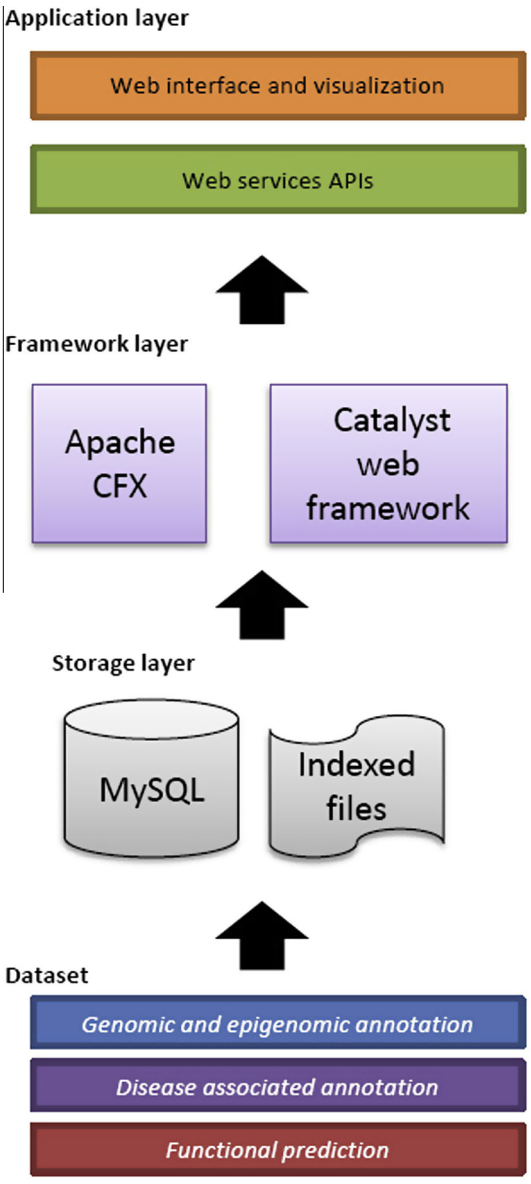

Fig. 2. The system architecture of SNVrap.

Last, SNVrap can also directly connect the target SNV to RegulomeDB [36], HaploReg [35], rSNPBase [38] and GWASdb [91] in internal page, which greatly facilitate the functional investigation of regulatory variants. Apart from collecting the genomic and epigenomic signals that underlie the putative DNA function elements, we also incorporated tissue specific expression quantitative traits locus (eQTLs) data from the Genotype-Tissue Expression (GTEx) project [92].

3.2.2. Disease associated annotation

Considering different disease classification (common, rare and cancer-related diseases) and associated/causal elements, we collected the disease-causal information from different resources. Disease causal genes for monogenetic disease were downloaded from OMIM [84] and NCBI Genetic Testing Registry (GTR) database [93]. Annotations for genetic association studies of complex diseases and disorders were retrieved from GAD [85], GWASdb [91] and NHGRI GWAS Catalog [94]. Data for relationships among human variations and observed health status were fetched from ClinVar [87]. For cancer related genes and somatic mutations, we used the latest COSMIC [86] data and a database for somatic mutations affecting miRNA function (SomamiR) [95]. We then cleaned up and synchronize these data in well-known standard formats with consistent information.

3.2.3. Functional prediction on multiple domains

We have reviewed many recent bioinformatics tools to predict the functional consequence of variants on different biological domains. However, it is difficult to pre-compute and integrate all of them in our system, either because some tools only provide web server or due to the heavy computing burden to run the algorithms for all known variants. Under this circumstance, we selectively run some workable and important tools or downloaded the compiled datasets from the developers' web site for each domain.

For variants affecting TF binding affinity, we adopted our previously described algorithm [37]. For alternative splicing, we utilized Skippy [96] and MutPred Splice [97] to get the prediction scores and annotations for all known SNV affecting splice site in dbSNP138. For SNVs influencing miRNA binding sites, we first used our own calculation that measures the hybridization energy change between miRNA and miRNA target (PITA [98] and miRanda [99]). We also downloaded and processed the predictions from PolymiRTS [100]. To measure the effect of synonymous mutation, we utilized the SilVA [101] to compute the deleterious scores. For protein coding domain, SNVrap also included dbNSFP [24] and computed the combined scores by KGGSeq [69]. Finally, we included PhosSNP [102] to estimate genetic variation that results in aberrant regulation of protein phosphorylation.

3.3. Web Services API

We provided programming interfaces for advanced user to retrieve the annotations in batch. We implemented a series of Web Service interfaces using Apache CXF. User can also simply query SNV through a specific web url, our system will immediately display the annotation page. If the user is only interested in a specific domain of queried SNV, he can utilize different Web Service functions to fetch the required information.

4. Results and examples

4.1. The usage of SNVrap

SNVrap presents two functions to annotate SNVs: a rapid mode and a batch mode. In the rapid mode, user can input a SNV locus and quickly receive comprehensive annotations in a well-designed web page. To graphically represent the annotation result, SNVrap provides three interactive panels including a dynamic Manhattan plot that displays the linkage disequilibrium (LD) proxy of targeted SNV using HapMap and 1000 Genomes Project data, a prioritization tree that describes functional hits according to our previous additive effect model [103], as well as annotation tabs that contain over 40 up-to-date annotations for the queried SNV. In the batched mode, user can input a list of SNVs (dbSNP ID or genomic coordinates), SNVrap will annotate these variants in a backend web server.

4.2. Case studies show the effectiveness of tool

We utilized several well-studied causal SNVs to verify the effectiveness of variant annotation in SNVrap. rs12740374 had been reported by many GWASs [104–109] and can alter the hepatic expression of the *SORT1* gene by creating a C/EBP (CCAAT/enhancer binding protein) transcription factor binding site [46]. SNVrap

annotates the function of this SNV in a very complete detail. In general information tab, SNVrap reports that rs12740374 can affect different genes. It locates in the DOWNSTREAM of ENSG00000134222 (*PSRC1*) and 3PRIME UTR of ENSG00000143126 (*CELSR2*). Knowledge-based annotation shows this mutation locus overlaps many regulatory regions of different function elements by ENCODE. It is also detected as a significant GTEx eQTL in muscle skeletal tissue. Importantly, SNVrap accurately annotates the chromosome interaction between rs12740374 locus and the promoter region of *SORT1* gene by Hi-C, which is consistent with functional validation. SNVrap also predicts several TFBS changes in the SNV locus, which include a significant C/EBP binding affinity alteration by different alleles. Interestingly, SNVrap predicts that rs12740374 can putatively affect the target (*CELSR2*) recognition of three miRNAs (hsa-miR-3663-3p, hsa-miR-6865-3p and hsa-miR-6871-3p), which indicates the SNV's potential role in post-transcriptional regulation. Furthermore, this SNV has been annotated in the conservative region, positively selected gene of some subpopulations and disease-related gene according to SNVrap. Therefore, the annotation for rs12740374 can be a representative case to show the multiple biological mechanisms that a unique SNV could affect. In this level, SNVrap can capture and explain maximal potential biological functions for query SNV by multiple mapping to different functional elements.

Another example is the annotations of rs16891982, which locates in human *SLC45A2* gene that affecting light skin pigmentation in Europeans [110]. SNVrap reports many related literatures of this mutation, in which the association of natural selection and phenotype in different human populations were studied. Also, functional annotation shows this non-synonymous SNV may disrupt regular protein structure or phosphorylation site, and therefore alters the protein function. Taken together, SNVrap can provide accurate and comprehensive annotations that are consistent with many well-studied cases.

4.3. Advantages of SNVrap in post-genomic variants functional annotation

SNVrap incorporates a large number of comprehensive and broad annotation features and has an easy-to-use interface, which greatly facilitate geneticists and clinicians to quickly retrieve related information in a compact web page. Comparing with existing tools, SNVrap focuses on complete annotation and has three distinct features. First, besides the common genomic features frequently adopted by many other tools, we put more emphasis on the functional prediction of SNVs that alter the biological functions spanning from transcription to post-translation. Previous annotation frequently utilized the functional prediction score on (nsSNVs) that directly disrupt protein sequence and function, but the functional predictions on synonymous and non-coding genetic variants were seldom evaluated. In SNVrap, we applied in-house and published software to calculate the functional scores in each domain and also reported sufficient annotations associated with these scores. Second, SNVrap is quick and intuitive. For a single query of one SNV, the system provides a well-structured visualization interface to display the annotation of target SNV. There are three interactive panels for representing the query result, which include detailed annotation tabs, LD proxy viewer and a variant prioritization viewer. Last, SNVrap also included many novel annotations referring to population genetics, high order epigenomics features and disease-related information, such as population haplotype structure, positive selection, long range chromosome interaction, eQTL. Therefore, SNVrap provides more extensive annotations than existing tools.

5. Trends in variants functional annotation and future directions

5.1. Condition specific annotations

Current human variant annotation tools tried to enumerate all functional relevant features across the whole human genome. However, there are in fact many dependent factors related to the exact biological function of a mutation, which hinders the rational usage of annotation. First, nature has shaped the population difference in which the allele frequency of functional genetic variants may be heterogeneous according to selection or genetic drift, which result in the population specific traits and promote human evolution [1,2]. Given a disease/trait associated variant in one population, the risk allele may not produce deleterious effect for another population due to haplotype structure, epistasis and other environment factors. To this end, the population-oriented information, such as disease/trait associated variant with subpopulation specificity and allele deviation from overall population, should be included in next phase annotation work. On the other hand, the temporal and spatial biological process will affect genetic variants to exert their function among different tissues/cell types [111]. The unbalanced transcriptional signals have been shown to express distinct pattern around the eQTL and GWAS loci in different tissues/cell types including gene expression, histone modification and DNA chromatin state [112,113]. Recent GTEx project has produced genome-wide transcript expression and individual genotype across as many as 26 human tissues from a large number of donors, which significantly facilitate the discovery of tissue specific eQTLs [114]. The Cancer Genome Atlas (TCGA) Pan-Cancer analysis project has uncovered many somatic drivers by aggregating events across tumor types [115]. Therefore, annotating tissue/cell type specific variants could be very useful work in the future.

5.2. Prioritization based on functional SNV annotations

Annotation is the starting point to decipher the functional role of a mutation. The following task will be focused on prioritizing the effect size of these variants according to their deleteriousness and pathogenesis in aberrant biological function and disease development. However, accurately estimating the function effect will be a tough work. Although most of functional prediction algorithms have provided scores as well as significance levels for each domain-qualified variant, the consistence of predictions by different algorithms could be poor. The reasons may be attributed to many factors like the model over-fitting, hidden covariants, small positive dataset, etc. To improve the performance of nsSNV functional prediction, recent algorithms have attained quite satisfactory result by combinatory strategy and information integration [70–73]. But in other functional domains, the accurate and effective evaluation methods are in urgent demand. Recently, as the constantly increasing of new genomic data, researchers have developed several systematic methods to score and prioritize non-coding genetic variants by data integration, which include GWAS3D [37], Funseq [116], CADD [117] and GWAVA [118]. Those significant works open a new door for understanding the full spectrum of variant function.

Furthermore, a critical question to be asked is whether one can prioritize the effect of functional variants across biological domains. For example, given an nsSNV and a non-coding variant both with high deleterious score, which is the most probable one that actually contribute the investigated disease? So far, none of the current methods can effectively distinguish and prioritize

them. Therefore, systematic solution and specific statistics model are needed to globally identify disease-causal variants.

6. Conclusion and future plan

We have presented a comprehensive web-based tool for annotating SNV in the human genome combined with a brief review of current trends in the field of variant annotation. To our knowledge, SNVrap is a unique tool that can provide functional prediction information in multiple domains. It also presented a one-stop web portal for quick search and interactive interface. For gene-based and knowledge-based annotations, SNVrap has many overlapping functions comparing with the prevalent software such as ANNOVAR, SeattleSeq and VEP. Although SNVrap try to incorporate recent functional genomics data, such as ENCODE and eQTL data, to annotate SNV, it could not attribute to a significant novelty. However, one of distinct merit of SNVrap is it equips with functional prediction across many biological processes other than only focusing on nsSNV functional prediction. Since SNVrap is a web-based tool, it also brings great convenience to many biologists who do not have enough computational skills.

There are many bioinformatics tools that can predict variant functional mechanism in different biological domains. SNVrap obviously did not include all of sophisticated prediction algorithms in each field instead of selectively picking up some representative ones. We have to face the challenges of inconsistent results between different tools in variant annotation and functional prediction. For example, even for well-studied nsSNVs, the pair-wise correlation of the five deleteriousness scores on both disease-causal and neutral rare variant sets are low (less than 0.5 Spearman's rank correlation coefficients for most pairs) [69]. To this end, researchers have adopted many integrated strategies to improve the performance. We strongly recommend user combining all-round evidences in determining the possible functional mechanism before performing experiment validation, especially when conflict predictions exist between different tools or weak score is reported. We will constantly keep updating information and adding new features into SNVrap. We have not incorporated much information about Indel and CNV in current version of SNVrap due to the complexity and the difficulty in functional prediction. This will be a critical work in next phase of SNVrap development.

Acknowledgements

This work was supported by the Research Grants Council (781511M, 17121414M) of Hong Kong and National Natural Science Foundation of China (91229105) of China.

References

[1] D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, P.E. Bonnen, P.I. de Bakker, P. Deloukas, S.B. Gabriel, et al., Nature 467 (7311) (2010) 52–58.
[2] G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, H.M. Kang, G.T. Marth, G.A. McVean, Nature 491 (7422) (2012) 56–65.
[3] B.E. Bernstein, E. Birney, I. Dunham, E.D. Green, C. Gunter, M. Snyder, Nature 489 (7414) (2012) 57–74.
[4] B.E. Bernstein, J.A. Stamatoyannopoulos, J.F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M.A. Marra, A.L. Beaudet, J.R. Ecker, et al., Nat. Biotechnol. 28 (10) (2010) 1045–1048.
[5] P. Legrain, R. Aebersold, A. Archakov, A. Bairoch, K. Bala, L. Beretta, J. Bergeron, C.H. Borchers, G.L. Corthals, C.E. Costello, et al., Mol. Cell Proteomics 10 (7) (2011) M111009993.
[6] K. Wang, M. Li, H. Hakonarson, Nucleic Acids Res. 38 (16) (2010) e164.
[7] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, F. Cunningham, Bioinformatics 26 (16) (2010) 2069–2070.
[8] S.B. Ng, E.H. Turner, P.D. Robertson, S.D. Flygare, A.W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E.E. Eichler, et al., Nature 461 (7261) (2009) 272–276.
[9] J.R. Grant, A.S. Arantes, X. Liao, P. Stothard, Bioinformatics 27 (16) (2011) 2300–2301.
[10] V. Makarov, T. O'Grady, G. Cai, J. Lihm, J.D. Buxbaum, S. Yoon, Bioinformatics 28 (5) (2012) 724–725.
[11] D. Ge, E.K. Ruzzo, K.V. Shianna, M. He, K. Pelak, E.L. Heinzen, A.C. Need, E.T. Cirulli, J.M. Maia, S.P. Dickson, et al., Bioinformatics 27 (14) (2011) 1998–2000.
[12] Y.W. Asmann, S. Middha, A. Hossain, S. Baheti, Y. Li, H.S. Chai, Z. Sun, P.H. Duffy, A.A. Hadad, A. Nair, et al., Bioinformatics 28 (2) (2012) 277–278.
[13] P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, Fly (Austin) 6 (2) (2012) 80–92.
[14] I. Medina, A. De Maria, M. Bleda, F. Salavert, R. Alonso, C.Y. Gonzalez, J. Dopazo, Nucleic Acids Res. 40 (2012) W54–W58.
[15] J. Ernst, M. Kellis, Nat. Methods 9 (3) (2012) 215–216.
[16] L. Habegger, S. Balasubramanian, D.Z. Chen, E. Khurana, A. Sboner, A. Harmanci, J. Rozowsky, D. Clarke, M. Snyder, M. Gerstein, Bioinformatics 28 (17) (2012) 2267–2269.
[17] U. Paila, B.A. Chapman, R. Kirchner, A.R. Quinlan, PLoS Comput. Biol. 9 (7) (2013) e1003153.
[18] H. Vuong, R.M. Stephens, N. Volfovsky, Bioinformatics 30 (7) (2013) 1013–1014.
[19] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M.R. Speicher, J. Zschocke, Z. Trajanoski, Brief. Bioinform. 15 (2) (2013) 256–278.
[20] Z.E. Sauna, C. Kimchi-Sarfaty, Nat. Rev. Genet. 12 (10) (2011) 683–691.
[21] P. Makrythanasis, S.E. Antonarakis, Clin. Genet. 84 (5) (2013) 422–428.
[22] Z. Wang, M. Gerstein, M. Snyder, Nat. Rev. Genet. 10 (1) (2009) 57–63.
[23] J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B.L. Aken, D. Barrell, A. Zadissa, S. Searle, et al., Genome Res. 22 (9) (2012) 1760–1774.
[24] X. Liu, X. Jian, E. Boerwinkle, Hum. Mutat. 34 (9) (2013) E2393–E2402.
[25] UniProt C, Nucleic Acids Res. 41 (2013) D43–D47 (Database issue).
[26] M. Halvorsen, J.S. Martin, S. Broadaway, A. Laederach, PLoS Genet. 6 (8) (2010) e1001074.
[27] J.S. Martin, M. Halvorsen, L. Davis-Neulander, J. Ritz, C. Gopinath, A. Beauregard, A. Laederach, RNA 18 (1) (2012) 77–87.
[28] Y. Wan, K. Qu, Q.C. Zhang, R.A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R.C. Spitale, M.P. Snyder, E. Segal, et al., Nature 505 (7485) (2014) 706–709.
[29] J. Carbonell, E. Alloza, P. Arce, S. Borrego, J. Santoyo, M. Ruiz-Ferrer, I. Medina, J. Jimenez-Almazan, C. Mendez-Vidal, M. Gonzalez-Del Pozo, Genome Med. 4 (8) (2012) 62.
[30] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Science 316 (5830) (2007) 1497–1502.
[31] J. Wang, J. Zhuang, S. Iyer, X.Y. Lin, M.C. Greven, B.H. Kim, J. Moore, B.G. Pierce, X. Dong, D. Virgil, et al., Nucleic Acids Res. 41 (2013) D171–D176 (Database issue).
[32] V.W. Zhou, A. Goren, B.E. Bernstein, Nat. Rev. Genet. 12 (1) (2011) 7–18.
[33] P.A. Jones, Nat. Rev. Genet. 13 (7) (2012) 484–492.
[34] R.E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M.T. Maurano, E. Haugen, N.C. Sheffield, A.B. Stergachis, H. Wang, B. Vernot, et al., Nature 489 (7414) (2012) 75–82.
[35] L.D. Ward, M. Kellis, Nucleic Acids Res. 40 (2012) D930–D934 (Database issue).
[36] A.P. Boyle, E.L. Hong, M. Hariharan, Y. Cheng, M.A. Schaub, M. Kasowski, K.J. Karczewski, J. Park, B.C. Hitz, S. Weng, et al., Genome Res. 22 (9) (2012) 1790–1797.
[37] M.J. Li, L.Y. Wang, Z. Xia, P.C. Sham, J. Wang, Nucleic Acids Res. 41 (2013) W150–W158 (Web server issue).
[38] L. Guo, Y. Du, S. Chang, K. Zhang, J. Wang, Nucleic Acids Res. 42 (1) (2014) D1033–D1039.
[39] T.J. Hudson, Nat. Genet. 33 (4) (2003) 439–440.
[40] M.J. Li, B. Yan, P.C. Sham, J. Wang, Brief. Bioinform. (2014).
[41] T. Niimi, M. Munakata, C.L. Keck-Waggoner, N.C. Popescu, R.C. Levitt, M. Hisada, S. Kimura, Am. J. Hum. Genet. 70 (3) (2002) 718–725.
[42] C. Phornphutkul, Y. Anikster, M. Huizing, P. Braun, C. Brodie, J.Y. Chou, W.A. Gahl, Am. J. Hum. Genet. 69 (4) (2001) 712–721.
[43] X.Z. Hu, R.H. Lipsky, G. Zhu, L.A. Akhtar, J. Taubman, B.D. Greenberg, K. Xu, P.D. Arnold, M.A. Richter, J.L. Kennedy, et al., Am. J. Hum. Genet. 78 (5) (2006) 815–826.
[44] J. Theuns, N. Brouwers, S. Engelborghs, K. Sleegers, V. Bogaerts, E. Corsmit, T. De Pooter, C.M. van Duijn, P.P. De Deyn, C. Van Broeckhoven, Am. J. Hum. Genet. 78 (6) (2006) 936–946.
[45] S. Tuupanen, M. Turunen, R. Lehtonen, O. Hallikas, S. Vanharanta, T. Kivioja, M. Bjorklund, G. Wei, J. Yan, I. Niittymaki, et al., Nat. Genet. 41 (8) (2009) 885–890.
[46] K. Musunuru, A. Strong, M. Frank-Kamenetsky, N.E. Lee, T. Ahfeldt, K.V. Sachs, X. Li, H. Li, N. Kuperwasser, V.M. Ruda, et al., Nature 466 (7307) (2010) 714–719.
[47] J.D. French, M. Ghoussaini, S.L. Edwards, K.B. Meyer, K. Michailidou, S. Ahmed, S. Khan, M.J. Maranian, M. O'Reilly, K.M. Hillman, et al., Am. J. Hum. Genet. 92 (4) (2013) 489–503.
[48] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S.M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A.E. Urban, et al., Science 328 (5975) (2010) 232–235.
[49] I. Williamson, R.E. Hill, W.A. Bickmore, Dev. Cell 21 (1) (2011) 17–19.
[50] J. Wang, Hum. Mutat. 33 (1) (2012) vii.

[51] G. Zhang, X. Chen, L. Chan, M. Zhang, B. Zhu, L. Wang, X. Zhu, J. Zhang, B. Zhou, J. Wang, Sci. Rep. 1 (2011) 20.
[52] M.J. Li, P.C. Sham, J. Wang, Bioinformatics 26 (22) (2010) 2897–2899.
[53] D. Benovoy, T. Kwan, J. Majewski, Nucleic Acids Res. 36 (13) (2008) 4417–4423.
[54] J.K. Pickrell, J.C. Marioni, A.A. Pai, J.F. Degner, B.E. Engelhardt, E. Nkadori, J.B. Veyrieras, M. Stephens, Y. Gilad, J.K. Pritchard, Nature 464 (7289) (2010) 768–772.
[55] K. Zhao, Z.X. Lu, J.W. Park, Q. Zhou, Y. Xing, Genome Biol. 14 (7) (2013) R74.
[56] J.W. Yarham, J.L. Elson, E.L. Blakely, R. McFarland, R.W. Taylor, Wiley Interdiscip. Rev. RNA 1 (2) (2010) 304–324.
[57] A.B. Glinskii, J. Ma, S. Ma, D. Grant, C.U. Lim, S. Sell, G.V. Glinsky, Cell Cycle 8 (23) (2009) 3925–3942.
[58] V. Kumar, H.J. Westra, J. Karjalainen, D.V. Zhernakova, T. Esko, B. Hrdlickova, R. Almeida, A. Zhernakova, E. Reinmaa, U. Vosa, et al., PLoS Genet. 9 (1) (2013) e1003201.
[59] L. Zhang, Y. Liu, F. Song, H. Zheng, L. Hu, H. Lu, P. Liu, X. Hao, W. Zhang, K. Chen, Proc. Natl. Acad. Sci. U.S.A. 108 (33) (2011) 13653–13658.
[60] W. Gu, T. Zhou, C.O. Wilke, PLoS Comput. Biol. 6 (2) (2010) e1000664.
[61] P.G. Higgs, W. Ran, Mol. Biol. Evol. 25 (11) (2008) 2279–2291.
[62] G. Cannarozzi, N.N. Schraudolph, M. Faty, P. von Rohr, M.T. Friberg, A.C. Roth, P. Gonnet, G. Gonnet, Y. Barral, Cell 141 (2) (2010) 355–367.
[63] D.A. Drummond, C.O. Wilke, Cell 134 (2) (2008) 341–352.
[64] L. Luna, V. Rolseth, G.A. Hildrestrand, M. Otterlei, F. Dantzer, M. Bjoras, E. Seeberg, Nucleic Acids Res. 33 (6) (2005) 1813–1824.
[65] J. Wu, R. Jiang, ScientificWorldJournal 2013 (2013) 675851.
[66] N.L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P.C. Ng, Nucleic Acids Res. 40 (2012) W452–W457 (Web server issue).
[67] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, Nat. Methods 7 (4) (2010) 248–249.
[68] J.M. Schwarz, C. Rodelsperger, M. Schuelke, D. Seelow, Nat. Methods 7 (8) (2010) 575–576.
[69] M.X. Li, H.S. Gui, J.S. Kwan, S.Y. Bao, P.C. Sham, Nucleic Acids Res. 40 (7) (2012) e53.
[70] A. Gonzalez-Perez, N. Lopez-Bigas, Am. J. Hum. Genet. 88 (4) (2011) 440–449.
[71] M.X. Li, J.S. Kwan, S.Y. Bao, W. Yang, S.L. Ho, Y.Q. Song, P.C. Sham, PLoS Genet. 9 (1) (2013) e1003143.
[72] A. Sifrim, D. Popovic, L.C. Tranchevent, A. Ardeshirdavani, R. Sakai, P. Konings, J.R. Vermeesch, J. Aerts, B. De Moor, Y. Moreau, Nat. Methods 10 (11) (2013) 1083–1084.
[73] J. Wu, Y. Li, R. Jiang, PLoS Genet. 10 (3) (2014) e1004237.
[74] E.S. Emison, A.S. McCallion, C.S. Kashuk, R.T. Bush, E. Grice, S. Lin, M.E. Portnoy, D.J. Cutler, E.D. Green, A. Chakravarti, Nature 434 (7035) (2005) 857–863.
[75] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, et al., Genome Res. 15 (8) (2005) 1034–1050.
[76] W. Miller, K. Rosenbloom, R.C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D.C. King, R. Baertsch, D. Blankenberg, et al., Genome Res. 17 (12) (2007) 1797–1808.
[77] E.V. Davydov, D.L. Goode, M. Sirota, G.M. Cooper, A. Sidow, S. Batzoglou, PLoS Comput. Biol. 6 (12) (2010) e1001025.
[78] K.S. Pollard, M.J. Hubisz, K.R. Rosenbloom, A. Siepel, Genome Res. 20 (1) (2010) 110–121.
[79] B.S. Weir, W.G. Hill, Annu. Rev. Genet. 36 (2002) 721–750.
[80] F. Tajima, Genetics 123 (3) (1989) 585–595.
[81] B.F. Voight, S. Kudaravalli, X. Wen, J.K. Pritchard, PLoS Biol. 4 (3) (2006) e72.
[82] M.J. Li, L.Y. Wang, Z. Xia, M.P. Wong, P.C. Sham, J. Wang, Nucleic Acids Res. 42 (2014) D910–D916 (Database issue).
[83] M. Pybus, G.M. Dall'Olio, P. Luisi, M. Uzkudun, A. Carreno-Torres, P. Pavlidis, H. Laayouni, J. Bertranpetit, J. Engelken, Nucleic Acids Res. 42 (2014) D903–D909 (Database issue).
[84] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Nucleic Acids Res. 33 (2005) D514–D517 (Database issue).
[85] K.G. Becker, K.C. Barnes, T.J. Bright, S.A. Wang, Nat. Genet. 36 (5) (2004) 431–432.
[86] S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, et al., Nucleic Acids Res. 39 (2011) D945–D950 (Database issue).
[87] M.J. Landrum, J.M. Lee, G.R. Riley, W. Jang, W.S. Rubinstein, D.M. Church, D.R. Maglott, Nucleic Acids Res. 42 (2014) D980–D985 (Database issue).
[88] M.M. Hoffman, O.J. Buske, J. Wang, Z. Weng, J.A. Bilmes, W.S. Noble, Nat. Methods 9 (5) (2012) 473–476.
[89] H. Li, Bioinformatics 27 (5) (2011) 718–719.
[90] M. Imakaev, G. Fudenberg, R.P. McCord, N. Naumova, A. Goloborodko, B.R. Lajoie, J. Dekker, L.A. Mirny, Nat. Methods 9 (10) (2012) 999–1003.
[91] M.J. Li, P. Wang, X. Liu, E.L. Lim, Z. Wang, M. Yeager, M.P. Wong, P.C. Sham, S.J. Chanock, J. Wang, Nucleic Acids Res. 40 (2012) D1047–D1054 (Database issue).
[92] Consortium GT, Nat. Genet. 45 (6) (2013) 580–585.
[93] W.S. Rubinstein, D.R. Maglott, J.M. Lee, B.L. Kattman, A.J. Malheiro, M. Ovetsky, V. Hem, V. Gorelenkov, G. Song, C. Wallin, et al., Nucleic Acids Res. 41 (2013) D925–D935 (Database issue).
[94] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al., Nucleic Acids Res. 42 (2014) D1001–D1006 (Database issue).
[95] A. Bhattacharya, J.D. Ziebarth, Y. Cui, Nucleic Acids Res. 41 (2013) D977–D982 (Database issue).
[96] A. Woolfe, J.C. Mullikin, L. Elnitski, Genome Biol. 11 (2) (2010) R20.
[97] M. Mort, T. Sterne-Weiler, B. Li, E.V. Ball, D.N. Cooper, P. Radivojac, J.R. Sanford, S.D. Mooney, Genome Biol. 15 (1) (2014) R19.
[98] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, E. Segal, Nat. Genet. 39 (10) (2007) 1278–1284.
[99] D. Betel, A. Koppal, P. Agius, C. Sander, C. Leslie, Genome Biol. 11 (8) (2010) R90.
[100] A. Bhattacharya, J.D. Ziebarth, Y. Cui, Nucleic Acids Res. 42 (2014) D86–D91 (Database issue).
[101] O.J. Buske, A. Manickaraj, S. Mital, P.N. Ray, M. Brudno, Bioinformatics 29 (15) (2013) 1843–1850.
[102] J. Ren, C. Jiang, X. Gao, Z. Liu, Z. Yuan, C. Jin, L. Wen, Z. Zhang, Y. Xue, X. Yao, Mol. Cell. Proteomics 9 (4) (2010) 623–634.
[103] M.J. Li, P.C. Sham, J. Wang, Cell Res. 22 (10) (2012) 1505–1508.
[104] S. Kathiresan, C.J. Willer, G.M. Peloso, S. Demissie, K. Musunuru, E.E. Schadt, L. Kaplan, D. Bennett, Y. Li, T. Tanaka, et al., Nat. Genet. 41 (1) (2009) 56–65.
[105] G. Lettre, C.D. Palmer, T. Young, K.G. Ejebe, H. Allayee, E.J. Benjamin, F. Bennett, D.W. Bowden, A. Chakravarti, A. Dreisbach, et al., PLoS Genet. 7 (2) (2011) e1001300.
[106] A.Y. Chu, F. Guilianini, H. Grallert, J. Dupuis, C.M. Ballantyne, B.J. Barratt, F. Nyberg, D.I. Chasman, P.M. Ridker, Circ. Cardiovasc. Genet. 5 (6) (2012) 676–685.
[107] H. Grallert, J. Dupuis, J.C. Bis, A. Dehghan, M. Barbalic, J. Baumert, C. Lu, N.L. Smith, A.G. Uitterlinden, R. Roberts, et al., Eur. Heart J. 33 (2) (2012) 238–251.
[108] M.S. Sandhu, D.M. Waterworth, S.L. Debenham, E. Wheeler, K. Papadakis, J.H. Zhao, K. Song, X. Yuan, T. Johnson, S. Ashford, et al., Lancet 371 (9611) (2008) 483–491.
[109] J. Kettunen, T. Tukiainen, A.P. Sarin, A. Ortega-Alonso, E. Tikkanen, L.P. Lyytikainen, A.J. Kangas, P. Soininen, P. Wurtz, K. Silander, et al., Nat. Genet. 44 (3) (2012) 269–276.
[110] I. Olalde, M.E. Allentoft, F. Sanchez-Quinto, G. Santpere, C.W. Chiang, M. DeGiorgio, J. Prado-Martinez, J.A. Rodriguez, S. Rasmussen, J. Quilez, et al., Nature 507 (7491) (2014) 225–228.
[111] A.S. Dimas, S. Deutsch, B.E. Stranger, S.B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, Science 325 (5945) (2009) 1246–1250.
[112] J. Fu, M.G. Wolfs, P. Deelen, H.J. Westra, R.S. Fehrmann, G.J. Te Meerman, W.A. Buurman, S.S. Rensen, H.J. Groen, R.K. Weersma, et al., PLoS Genet. 8 (1) (2012) e1002431.
[113] G. Trynka, C. Sandor, B. Han, H. Xu, B.E. Stranger, X.S. Liu, S. Raychaudhuri, Nat. Genet. 45 (2) (2013) 124–130.
[114] R. Sabarinathan, H. Tafer, S.E. Seemann, I.L. Hofacker, P.F. Stadler, J. Gorodkin, Hum. Mutat. 34 (4) (2013) 546–556.
[115] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, Nat. Genet. 45 (10) (2013) 1113–1120.
[116] E. Khurana, Y. Fu, V. Colonna, X.J. Mu, H.M. Kang, T. Lappalainen, A. Sboner, L. Lochovsky, J. Chen, A. Harmanci, et al., Science 342 (6154) (2013) 1235587.
[117] M. Kircher, D.M. Witten, P. Jain, B.J. O'Roak, G.M. Cooper, J. Shendure, Nat. Genet. 46 (3) (2014) 310–315.
[118] G.R. Ritchie, I. Dunham, E. Zeggini, P. Flicek, Nat. Methods 11 (3) (2014) 294–296.
[119] M.C. Andersen, P.G. Engstrom, S. Lithwick, D. Arenillas, P. Eriksson, B. Lenhard, W.W. Wasserman, J. Odeberg, PLoS Comput. Biol. 4 (1) (2008) e5.
[120] G. Macintyre, J. Bailey, I. Haviv, A. Kowalczyk, Bioinformatics 26 (18) (2010) i524–i530.
[121] M. Thomas-Chollier, A. Hufton, M. Heinig, S. O'Keeffe, N.E. Masri, H.G. Roider, T. Manke, M. Vingron, Nat. Protoc. 6 (12) (2011) 1860–1869.
[122] F.O. Desmet, D. Hamroun, M. Lalande, G. Collod-Beroud, M. Claustres, C. Beroud, Nucleic Acids Res. 37 (9) (2009) e67.
[123] M. Hariharan, V. Scaria, S.K. Brahmachari, BMC Bioinformatics 10 (2009) 108.
[124] J. Gong, Y. Tong, H.M. Zhang, K. Wang, T. Hu, G. Shan, J. Sun, A.Y. Guo, Hum. Mutat. 33 (1) (2012) 254–263.
[125] L.F. Thomas, T. Saito, P. Saetrom, Nucleic Acids Res. 39 (16) (2011) e109.
[126] B.P. Lewis, C.B. Burge, D.P. Bartel, Cell 120 (1) (2005) 15–20.
[127] H. Kiryu, K. Asai, Bioinformatics 28 (8) (2012) 1093–1101.