Original Paper

Public Health Genomics

Public Health Genomics
DOI: 10.1159/000477561

Received: May 15, 2017
Accepted: May 17, 2017
Published online: July 13, 2017

DoriTool: A Bioinformatics Integrative Tool for Post-Association Functional Annotation

Isabel Martín-Antoniano[a, c] Lola Alonso[a, b] Miguel Madrid[d]
Evangelina López de Maturana[a, b] Núria Malats[a, b]

[a]Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), [b]Centro de Investigación Biomédica en red Cáncer (CIBERONC), [c]Instituto de Medicina Molecular Aplicada (IMMA), Facultad de Medicina, Universidad San Pablo CEU, and [d]Structural Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Keywords
Association · Cancer · Functional annotation · Genome-wide association study · Genomics · Next-generation sequencing

Abstract
The emergence of high-throughput data in biology has increased the need for functional in silico analysis and prompted the development of integrative bioinformatics tools to facilitate the obtainment of biologically meaningful data. In this paper, we present DoriTool, a comprehensive, easy, and friendly pipeline integrating biological data from different functional tools. The tool was designed with the aim to maximize reproducibility and reduce the working time of the researchers, especially of those with limited bioinformatics skills, and to help them with the interpretation of the results. DoriTool is based upon an integrative strategy implemented following a modular design pattern. Using scripts written in Bash, Perl and R, it performs a functional in silico analysis annotation at mutation/variant level, gene level, pathway level and network level by combining up-to-date functional and genomic data and integrating also third-party bioinformatics tools in a pipeline. DoriTool uses GRCh37 human assembly and online mode. DoriTool provides nice visual reports including variant annotation, linkage disequilibrium proxies, gene annotation, gene ontology analysis, expression quantitative trait loci results from Genotype-Tissue Expression (GTEx) and coloured pathways. Here, we also show DoriTool functionalities by applying it to a dataset of 13 variants associated with prostate cancer. Project development, released code libraries, GitHub repository (https://github.com/doritool) and documentation are hosted at https://doritool.github.io/. DoriTool is, to our knowledge, the most complete bioinformatics tool offering functional in silico annotation of variants previously associated with a trait of interest, shedding light on the underlying biology and helping the researchers in the interpretation and discussion of the results.

Introduction

In the last decade, high-throughput genomics technologies, including genotyping arrays and next-generation sequencing, have revolutionized the characterization of

Isabel Martín-Antoniano and Lola Alonso contributed equally to this work.

Isabel Martín-Antoniano and Lola Alonso
Genetic and Molecular Epidemiology Group
Spanish National Cancer Research Center (CNIO)
C/Melchor Fernandez Almagro, 3, ES–28029 Madrid (Spain)
E-Mail iamartin@ceu.es and lola.alonso@cnio.es

disease at molecular resolution. Large-scale sequencing projects, such as the 1000 Genomes Project (1000GP) [1], UK10K [2] and NHLBI GO Exome Sequencing Project (ESP) [3], are being followed by even larger projects, such as the 100,000 Genomes Project [4]. Although those datasets are of great interest to both researchers and clinicians, their ultimate value will depend not on the number of variants identified, but rather on their functional interpretation [5]. In fact, the interpretation of results from association analyses (genome-wide association studies [GWAS] and, more recently, next-generation sequencing [NGS]) remains challenging [6]. Too frequently, a comprehensive interpretation of association results based on functional in silico analysis characterizing the mechanism behind the association or identifying the real causal variant/gene [7] is missing. Therefore, once a set of variants has been found to be associated with a particular phenotype of interest, a post-association functional annotation should be a common strategy of every analysis in order to interpret mutations/variants lists. However, this task has several difficulties.

The first drawback concerns the selection of the bioinformatics tools to be used, which will largely depend on the analysis options we intend to cover. In the last decade, many alternatives have been developed to interpret the mutations/variants lists providing functional information. However, many popular pipelines have not been updated in years. Furthermore, while most tools focus on the convenient mapping of diverse gene identifiers, few provide functional enrichment analysis as part of a large platform, the majority being web services (i.e., Babelomics, g:Profiler, and DAVID) [8]. Functional enrichment analysis can also be performed using R packages (i.e., FGNet and KEGGREST) [9], Perl scripts and Java applications [10], but a more accessible and comprehensive tool would be desirable.

Another critical point is the selection of the databases that need to be consulted to find the information required to annotate the functional effect of synonymous or non-synonymous single-nucleotide polymorphisms (SNPs) on the genes, or to obtain their predicted functional effect. In addition, it is important to know whether those databases are curated and updated, or which kind of predictor (e.g., SIFT, PolyPhen-2 or Condel) is the most suitable one [11–14].

The last but not least difficulty is related to the number of mutations/variants to be explored. When the association analysis results in a short list of statistically significant variants, functional annotation can be performed and curated manually (extracting information from genome browsers, such as Ensembl or UCSC). However, this is a tedious work that becomes unfeasible with large numbers of mutations/variants. The research community increasingly requires automatic and programmatic access to bioinformatics tools, since few available modern web-based genomic resources offer sophisticated tools for further analysis of these data.

At present, there is a lack of a bioinformatics tool that provides complete functional information of the GWAS hits at a glance. Even Variant Effect Predictor (VEP), which is a software suite that performs annotation and analysis of most types of genomic variation in coding and non-coding regions of the genome, does not provide a complete functional annotation [15].

Here, we present DoriTool, a pipeline built to combine different bioinformatics algorithms and public databases into one comprehensive, easy, fast and friendly tool, integrating different functional instruments in order to perform a complete functional in silico assessment, taking as its starting point a list of GWAS- or NGS-derived variants. DoriTool helps maximizing the reproducibility and research timelines, reducing the working time of the researchers, especially of those with limited bioinformatics skills, and helping them in the interpretation of their results.

Materials and Methods

DoriTool is based upon an integrative strategy implemented following a modular design pattern (Fig. 1). It allows performing a functional in silico analysis at (1) mutation/variant level, performing annotation of a set of mutations/variants, reporting expression quantitative trait loci (eQTLs) results from Genotype-Tissue Expression (GTEx) and providing their linkage disequilibrium (LD) proxies; (2) gene level, performing annotation of the genes tagged by the set of input variants and reporting also their gene ontology (GO); (3) pathway level; and (4) network level.

Our tool requires as input data a mutation/variant call format file (VCF) or an rs identifier SNP list (Table 1), which does not need to meet any requirement regarding size. DoriTool uses GRCh37 human assembly and online mode. A more detailed description of DoriTool pipeline as well as the existing bioinformatics tools used follow below (Table 2).

Mutation/Variant Level

Variant Annotation. Input variants are functionally annotated using VEP [16], which is an open-source, free-to-use, actively maintained and developed toolset (https://github.com/Ensembl/ensembl-vep) for the analysis, annotation and prioritization of genomic variants in both coding and non-coding regions. VEP allows determining the effect of the variants (SNPs) on genes, transcripts and proteins, as well as regulatory regions, non-genic variants, and including transcription factor binding sites [17], using regularly updated data files that are distributed by Ensembl, and

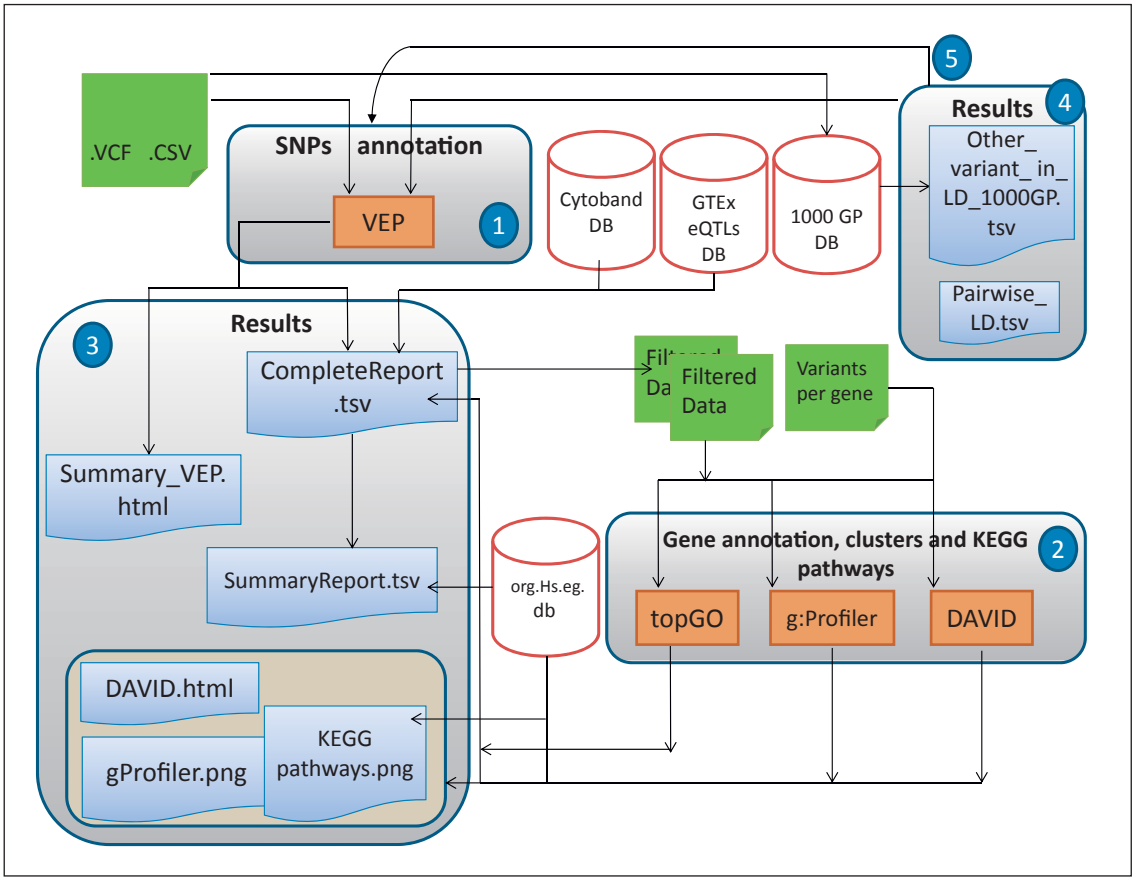


Fig. 1. Schematic diagram of the DoriTool architecture. The workflow starts with an input file (VCF format, list of chromosome positions of the variants and their alleles or list of variant rs identifiers). (1) Variant Effect Predictor (VEP) annotates the input variants with the tagged gene, the impact and some annotations described in the main text, querying the Ensembl database. The next step involves locating the variants in their corresponding cytobands, and the user has the option to query the Genotype-Tissue Expression (GTEx) database locally by providing a tissue-specific file with the expression quantitative trait loci (eQTLs). (2) DoriTool uses topGO (FGNet), g:Profiler and DAVID (FGNet) R packages to produce gene ontology terms, functional networks, enriched pathways and other annotations. (3) The main output files are the CompleteReport that contains all the annotations at transcript level provided by VEP and the SummaryReport that includes only selected information from the former, collapsing the results by gene and effect in the transcript. (4) As an option, the user may ask for Linked variants in a window size of 500 kb, with a linkage disequilibrium (LD) cut-off specified in the main call. (5) In case the user has asked for the previous task, a new CompleteReport profiling the linked variants will be produced. SNP, single-nucleotide polymorphism; GP, Genome Project; DB, database.

its output follows a standard form (VCF). Table 3 contains the plugins and default parameters of VEP in DoriTool. DoriTool also locates each variant in its corresponding cytoband, retrieving the information from the database downloaded from UCSC hg19 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/cytoBand.txt.gz).

Expression Quantitative Trait Loci. The user has the option to provide a database including tissue-specific significant SNP-gene pairs (with the same format as the ones GTEx provides) to obtain the effect size of the eQTLs as well as the eGene [18].

LD Proxies. DoriTool also explores proxy and putatively functional SNPs for a query SNP in a selected 1000GP population, using the Ensembl REST API in an integrated Perl script [19]. It computes and returns pairwise LD values (1) among the input variants

Table 1. An example of the VCF file that can be used as input of DoriTool

Chrom No.	Pos	ID	Ref	Alt	Qual	Filter	Info	Format
15	67468285	rs17294280	A	G	–	–	–	–
15	67450305	rs17228058	A	G	–	–	–	–

VCF, variant call format; Chrom, chromosome; Pos, position; Ref, reference; Alt, alternative; Qual, quality; Info, information.

Table 2. Annotation levels and the bioinformatics tools used in DoriTool

Annotation level	Tool	Input	Use	Script language
Variant annotation	VEP [17]	SNPs (VCF, CSV)	Variant effect prediction tool	Perl
	GTEx [18]	SNPs (LD proxies)	eQTLs	Bash
	LD proxy [19]	SNPs rs ID	Variants in LD with any overlapping existing variants from the Ensembl variation databases	Perl
Gene annotation	g:Profiler [8]	Symbol genes	Functional analysis	R
	org.Hs.eg.db [21]	ENS genes	Annotation	R
	TopGO (FGNet) [22]	ENS genes	Functional annotation	R
Pathway level	org.Hs.eg.db [21]	ENS genes	Annotation	R
	DAVID [26] (FGNet) [24]	Symbol genes	Cluster Colour pathways	R
	KEGGprofile [23]	KEGG ID	Colour pathways	R
Network level	DAVID (FGNet) [26]	Symbol genes	Cluster Colour pathways	R

VEP, Variant Effect Predictor; GTEx, Genotype-Tissue Expression; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism; VCF, variant call format; CSV, comma-separated values; ID, identifier; ENS, Ensembl Gene Stable IDs; eQTLs, expression quantitative trait loci.

Table 3. Default parameters of Variant Effect Predictor considered in DoriTool

Plugin options	Features	Description
Symbol		Adds the gene symbol to the output
Plugin UpDownDistance	10000, 5000	Changes the distance to call upstream and downstream consequence types
Plugin NearestGene	Limit = 1 Max_range = 1000000	Finds the nearest gene to an intergenic variant
Plugin Condel		Single-nucleotide polymorphism effect prediction

and (2) between each input variant and all other variants in a surrounding window. The default parameters in DoriTool for window size and strength of LD (r^2) are 500 kb and 0.90.

Gene Level

Gene Annotation. DoriTool uses 3 different tools to perform the gene annotation, therefore allowing interpreting and identifying the biological processes for the gene list tagged by the input list of mutations/variants [20]: (1) org.Hs.eg.db (Genome wide annotation for Human, R package version 3.4.1) [21]; (2) FGNet, which uses TopGO R package [22]; and (3) g:Profiler R package [8]. The Entrez Gene identifiers mapped to an Ensembl gene are obtained using the genome-wide annotation for human org.Hs.eg.db [21]. It is an R package maintained by Bioconductor at bioconductor.org (http://bioconductor.org/packages/org.Hs.eg.db/). DoriTool retrieves the following gene information to annotate: symbol, gene name, and Online Mendelian Inheritance in Man (OMIM) identifier, and it focuses primarily on inherited or heritable genetic diseases, GO identifiers, protein families (PFAM) and pathway.

topGO. topGO functions are included in FGNet R package (https://bioconductor.org/packages/release/bioc/html/topGO.html) to facilitate semi-automated enrichment analysis for GO terms [22], mapping the genes tagged by the input variants with the associated biological annotation terms (e.g., GO terms), and then statistically examine the enrichment of gene members for each of the annotation terms on the basis of gene counts. The default parameters in DoriTool are nodeSize = 1 (no prune) and pValThr = 0.01.

g:Profiler. g:Profiler is an open-source, free-to-use R package actively maintained on CRAN (https://CRAN.R-project.org/package=gProfileR) that performs functional enrichment analysis, including transcription factor binding site predictions, Mendelian disease annotations, information about protein expression and complexes, statistically significant GO terms, pathways and other gene function-related terms [8]. Multiple testing correction is performed by selecting the Benjamini-Hochberg false discovery rate. DoriTool manages the output given by g:Profiler as a visual report that complements and helps to interpret the results given by the other tools.

Table 4. Colour legend for the description of variants affecting genes in networks and pathways

Variants, *n*	Colour code FGNet	KEGG colour	DAVID colour
1	na	blue	green
2	−1	green	white
3	0	yellow	red
>3	1	red	red

na, not applicable.

Pathway Level

Pathway Annotation. It is performed using 3 R packages: FGNet, which uses DAVID, org.Hs.eg.db, which maps Entrez Gene identifiers to KEGG pathways (https://www.bioconductor.org/packages/release/data/annotation/manuals/org.Hs.eg.db/man/org.Hs.eg.db.pdf). Mappings were based on data provided by KEGG GENOME (http://www.genome.jp/kegg/genome.html). For the DoriTool pipeline, we included a code to obtain coloured KEGG pathways (Table 4) considering the number of variants per gene by using KEGGprofile [23]. Moreover, the offline mode of DAVID is considered, since it does not need specific considerations for the installation and guarantees always obtaining results even if the server is down. However, the offline mode does not allow to modify the default settings (default cut-off linkage: 0.5 and overlap = 4, initialseed = 4, finalseed = 4), resulting in a lower number of KEGG pathways and GO terms, which is compensated with the org.Hs.eg.db annotation.

Network Level

DoriTool also provides coloured functional networks of the list of genes tagged by the input variant list (Table 4). These functional connections between the different genes were based on annotations (GO) [24] and given by DAVID functions included in FGNet [25]. Building functional networks provides an overview of the biological functions of the genes/terms and permits links between genes, overlapping between clusters.

Installation

DoriTool modules are written in Perl, Bash, and R scripts and run on any UNIX-like operating system. To install DoriTool, simply download and run the installer script, which automatically downloads the necessary libraries, packages and annotation files. Alternatively, the user can install its dependencies manually or using the Docker container provided. The full source code and the container are freely available on the GitHub repository (https://github.com/doritool).

The DoriTool website includes general information about the purpose of the tool, instances and explanations about its uses, as well as the link to connect to the GitHub repository in order to download the tool directly (https://doritool.github.io/). The DoriTool is available under a GNU AGPLv3 license (https://choosealicense.com/licenses/agpl-3.0/#).

Results and Discussion

DoriTool is a pipeline designed to perform functional annotation using a range of reliable proven tools and databases. Therefore, it was designed for a post-association analysis (e.g., GWAS and NGS). It combines up-to-date functional and genomic data and serves the community through a pipeline with nice visual reports.

Here, we demonstrate the abilities of DoriTool by applying it to a dataset of 13 SNPs previously reported as associated with prostate cancer, which were selected ad hoc to show the scope and the full potential of DoriTool (see online suppl. File S1; for all online suppl. material, see www.karger.com/doi/10.1159/000477561).

Several output files were obtained: 4 text files (SummaryReport.tsv; CompleteReport.tsv; Pairwise_LD.tsv; and other_variants_in_LD_1000GP_EU.tsv), an image (gprofilerResults.png), 2 HTML files (DAVID.html and Summary_VEP.html) and 4 folders (DAVID, KEGG, topGO and VEP), containing the results obtained in each module of DoriTool. Online supplementary Files S2–S9 are the most important outputs obtained after running DoriTool with the example dataset.

The output file SummaryReport.tsv is a tab-delimited text file containing in each row, apart from the input information related to the mutation/variant, its consequence defined by the sequence ontology (http://www.sequenceontology.org/), its impact (high, moderate, low and modifier), transcript quality flags (cds_start_NF: CDS 5′ incomplete, cds_end_NF: CDS 3′ incomplete), the symbol of the gene tagged by that variant (considering a region of 10 kb upstream of the transcription start site and 5 kb downstream of the gene end), information regarding the transcription factor binding site related to the variant, the nearest gene in case of intergenic variants, the Condel score (i.e., a consensus score considering SIFT and PolyPhen-2, which ranges from 0 to 1, 0 being neutral and 1 deleterious), the cytoband, the KEGG pathway in which the tagged gene was annotated, the description of the gene, OMIM, PFAM, pathway, variants ID, Entrez Gene, the GO term, the eGene and the eQTLs effect size. Table 5 contains the first rows of our example, and online supplementary File S3 shows the complete file.

The output file CompleteReport.tsv is also a tab-delimited text file containing in each row, apart from the information reported in the SummaryReport.tsv file, additional information, such as feature identifier and feature type (e.g., transcript and regulatory feature), position of the input variant in cDNA, coding sequence and protein position of the input variant, changed amino acid,

Table 5. SummaryReport.tsv file which shows the first rows of our example

Uploaded variants	Consequence	Impact	Symbol	Condel	Cytoband	KEGG	Gene description	GO
rs17632542	3_prime_UTR_vriant, NMD_ transcipt_variant	Modifier	KLK3	–	q13.33	hsa5200	Kallikrein related peptidase 3	Antimicrobial peptide production
rs17632542	downstream_gene_ variant	Modifier	KLK3	–	q13.33	hsa5200	Kallikrein related peptidase 3	Antimicrobial peptide production
rs17632542	missense_variant	Moderate	KLK3	deleterious (0.487)	q13.33	hsa5200	Kallikrein related peptidase 3	Antimicrobial peptide production
rs17632542	non_coding_transcript_ exon_vriant,non_coding_ transcript_variant	Modifier	KLK3	–	q13.33	hsa5200	Kallikrein related peptidase 3	Antimicrobial peptide production
rs17632542	upstream_gene_variant	Modifier	KLK2	–	q13.33		Kallikrein related peptidase 2	Cellular component disassembly
rs115160117	upstream_gene_variant	Modifier	PDK1	–	q31.1	hsa4660	Pyruvate dehydrogenase kinase 1	Mitophagy in response to mitochondrial depolarization
rs115160117	intron_variant,non_ coding–transcript_variant	Modifier	AC093818.1	–	q31.1		na	
rs78647569	upstream_gene_variant	Modifier	PDK1	–	q31.1	hsa4660	Pyruvate dehydrogenase kinase 1	Mitophagy in response to mitochondrial depolarization

na, not applicable; GO, gene ontology.

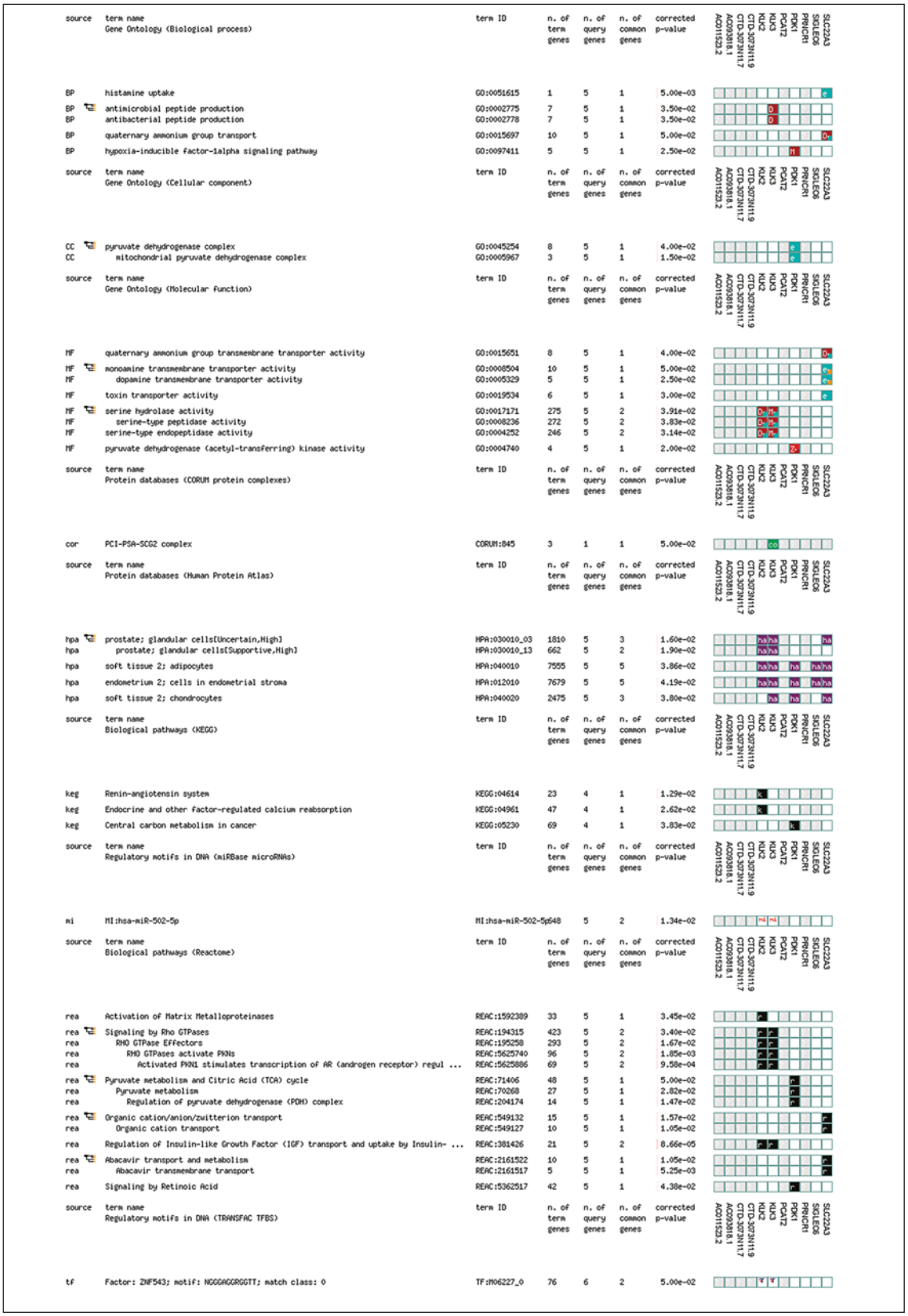
codon change (the alternative codons with the variant base in upper case), existing variation (dbSNP or COSMIC variations), distance to the transcription start site, strand of the feature, source of the symbol (e.g., VEGA, Ensembl or HGNC) and HGNC ID.

The LD proxies of the input variants (i.e., 1000GP variants in high LD considering the individuals of European descent) are provided in the pairwise_LD.tsv and other_variants_in_LD_1000GP_EU.tsv files. In addition to the IDs of the proxies, the *D'* (difference between the observed and the expected frequency of a given haplotype) and r^2 (correlation) for each pair are reported, as well as the functional annotation of the proxies (SummaryReport_LDproxies.tsv), in order to facilitate more information regarding the putative functional variant (see online suppl. File S6).

The image gprofilerResults.png, resulting from the implementation of g:Profiler package, is also reported as an output of DoriTool. This png file is a static image that does not provide the possibility to interact, which can affect the interpretability of the results when many genes are tagged by the input SNPs (Fig. 2).

In g:Profiler, evidence codes distinguish different types of associations that can occur between a gene and a property, for example GO term (stronger evidence is presented in red and weaker evidence in blue), KEGG pathway (strong evidence is presented in black) or TRANSFAC motif. DAVID.html file was generated by the DAVID version in FGNet (see online suppl. File S8). It contains a comprehensive report including different views of the functional network where the genes are interconnected, the cluster/meta-group legend and some further statistics derived from the genes tagged by the variants from the input list.

The functional network is based on 2 networks/incidence matrices: common clusters and common gene-term sets. These are their transitivity values, that is, the probability that adjacent vertices of a vertex are connect-



source	term name Gene Ontology (Biological process)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
BP	histamine uptake	GO:0051615	1	5	1	5.00e-03
BP	antimicrobial peptide production	GO:0002775	7	5	1	3.50e-02
BP	antibacterial peptide production	GO:0002778	7	5	1	3.50e-02
BP	quaternary ammonium group transport	GO:0015697	10	5	1	5.00e-02
BP	hypoxia-inducible factor-1alpha signaling pathway	GO:0097411	5	5	1	2.50e-02
source	term name Gene Ontology (Cellular component)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
CC	pyruvate dehydrogenase complex	GO:0045254	8	5	1	4.00e-02
CC	mitochondrial pyruvate dehydrogenase complex	GO:0005967	3	5	1	1.50e-02
source	term name Gene Ontology (Molecular Function)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
MF	quaternary ammonium group transmembrane transporter activity	GO:0015651	8	5	1	4.00e-02
MF	monoamine transmembrane transporter activity	GO:0008504	10	5	1	5.00e-02
MF	dopamine transmembrane transporter activity	GO:0005329	5	5	1	2.50e-02
MF	toxin transporter activity	GO:0019534	6	5	1	3.00e-02
MF	serine hydrolase activity	GO:0017171	275	5	2	3.91e-02
MF	serine-type peptidase activity	GO:0008236	272	5	2	3.83e-02
MF	serine-type endopeptidase activity	GO:0004252	246	5	2	3.14e-02
MF	pyruvate dehydrogenase (acetyl-transferring) kinase activity	GO:0004740	4	5	1	2.00e-02
source	term name Protein databases (CORUM protein complexes)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
cor	PCI-PSA-SCG2 complex	CORUM:845	3	1	1	5.00e-02
source	term name Protein databases (Human Protein Atlas)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
hpa	prostate; glandular cells[Uncertain,High]	HPA:030010_03	1810	5	3	1.60e-02
hpa	prostate; glandular cells[Supportive,High]	HPA:030010_13	662	5	2	1.90e-02
hpa	soft tissue 2; adipocytes	HPA:040010	7555	5	5	3.86e-02
hpa	endometrium 2; cells in endometrial stroma	HPA:012010	7679	5	5	4.19e-02
hpa	soft tissue 2; chondrocytes	HPA:040020	2475	5	3	3.80e-02
source	term name Biological pathways (KEGG)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
keg	Renin-angiotensin system	KEGG:04614	23	4	1	1.29e-02
keg	Endocrine and other factor-regulated calcium reabsorption	KEGG:04961	47	4	1	2.62e-02
keg	Central carbon metabolism in cancer	KEGG:05230	69	4	1	3.83e-02
source	term name Regulatory motifs in DNA (miRBase microRNAs)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
mi	MI:hsa-miR-502-5p	MI:hsa-miR-502-5p548	5	2	1.34e-02
source	term name Biological pathways (Reactome)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
rea	Activation of Matrix Metalloproteinases	REAC:1592389	33	5	1	3.45e-02
rea	Signaling by Rho GTPases	REAC:194315	423	5	2	3.40e-02
rea	RHO GTPase Effectors	REAC:195258	293	5	2	1.67e-02
rea	RHO GTPases activate PKNs	REAC:5625740	96	5	2	1.85e-03
rea	Activated PKN1 stimulates transcription of AR (androgen receptor) regul ...	REAC:5625886	69	5	2	9.58e-04
rea	Pyruvate metabolism and Citric Acid (TCA) cycle	REAC:71406	48	5	1	5.00e-02
rea	Pyruvate metabolism	REAC:70268	27	5	1	2.82e-02
rea	Regulation of pyruvate dehydrogenase (PDH) complex	REAC:204174	14	5	1	1.47e-02
rea	Organic cation/anion/zwitterion transport	REAC:549132	15	5	1	1.57e-02
rea	Organic cation transport	REAC:549127	10	5	1	1.05e-02
rea	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin- ...	REAC:381426	21	5	2	8.66e-05
rea	Abacavir transport and metabolism	REAC:2161522	10	5	1	1.05e-02
rea	Abacavir transmembrane transport	REAC:2161517	5	5	1	5.25e-03
rea	Signaling by Retinoic Acid	REAC:5362517	42	5	1	4.38e-02
source	term name Regulatory motifs in DNA (TRANSFAC TFBS)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
tf	Factor: ZNF543; motif: NGGGAGGRGGTT; match class: 0	TF:M06227_0	76	6	2	5.00e-02

Fig. 2. g:Profiler output.

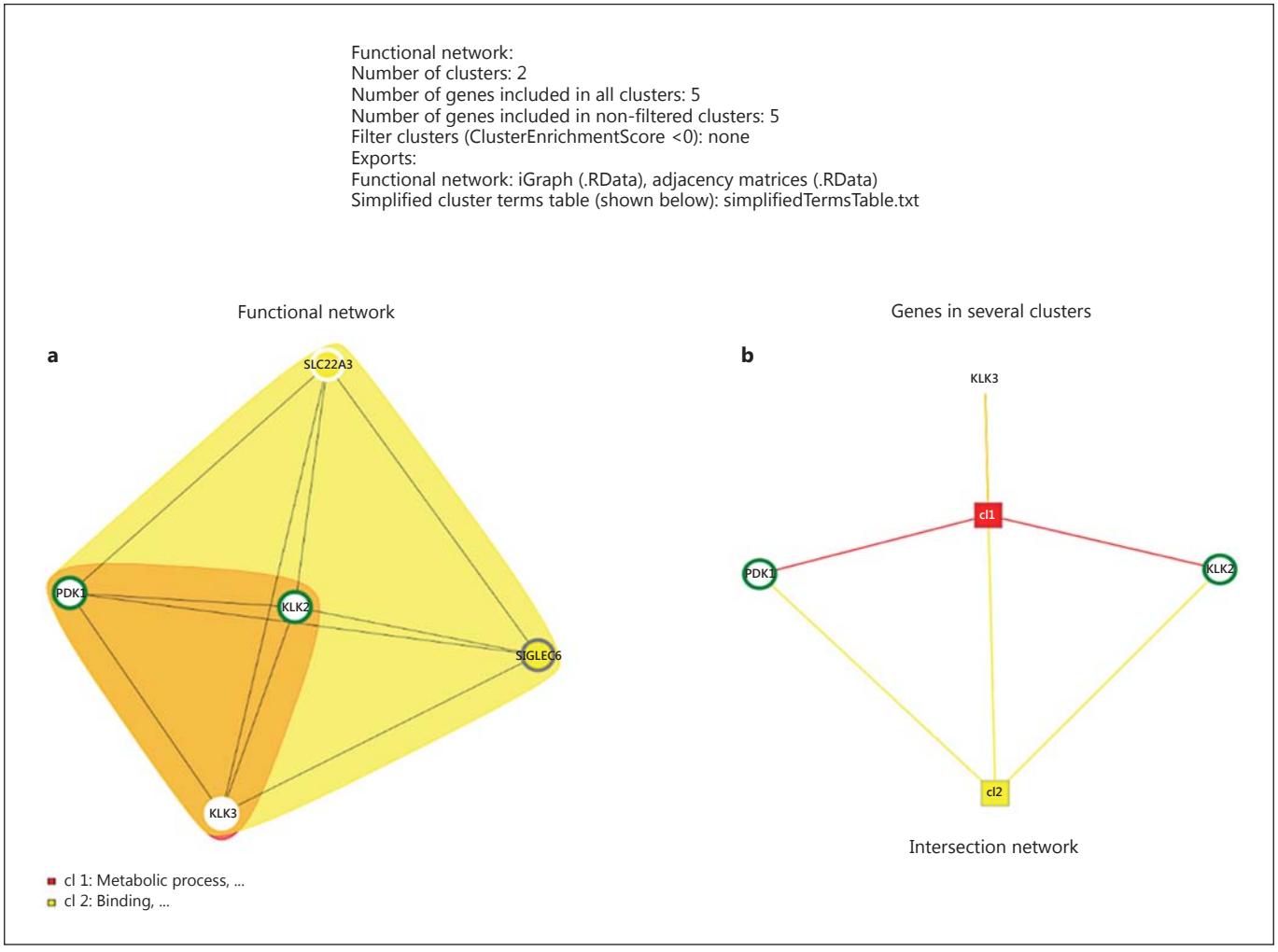


Fig. 3. DAVID results from our 13-variants set. Green circles represent genes affected by 1 variant, white circles are genes with 2 variants and red circles (not shown in the figure) would be genes affected by more than 2 variants (see Table 4). **a** Functional network: representation of the results as gene-term groups. Genes with a coloured background have a unique cluster (and their colour corresponds to the one shown in the legend), while genes with a white background are shared genes between clusters. **b** Intersection network: simplified functional network where all the genes that belong to only one meta-group are clustered into a single node. Clusters are represented by square boxes and genes by circles. Here, we show one of the clusters obtained.

ed. Figure 3 shows the cluster results from our 13-variants set. Figure 3a shows 2 clusters into gene-term groups in green (only 1 variant) and white circles (2 variants). Genes with a coloured background have a unique cluster (and their colour corresponds to the one shown in the legend), while genes with a white background are shared genes between clusters. Figure 3b shows the intersection network, which is a simplified functional network where all the genes that belong to only 1 meta-group are clustered into a single node. Clusters are represented by square boxes and genes by circles. The outputs reported by DAVID, KEGG, topGO and VEP are generated in 4 folders (Table 6).

Among them, the most relevant ones are Summary_VEP.html (VEP folder) and the png files starting with the hsa prefix, which will be described in the following. The output file Summary_VEP.html in the VEP folder shows general statistics (variants processed, novel/existing variants, genes, transcripts and regulatory features overlapped) on the top. The variant classes and consequences (most severe) are shown using pie graphs in a user-friendly form (see online suppl. File S9).

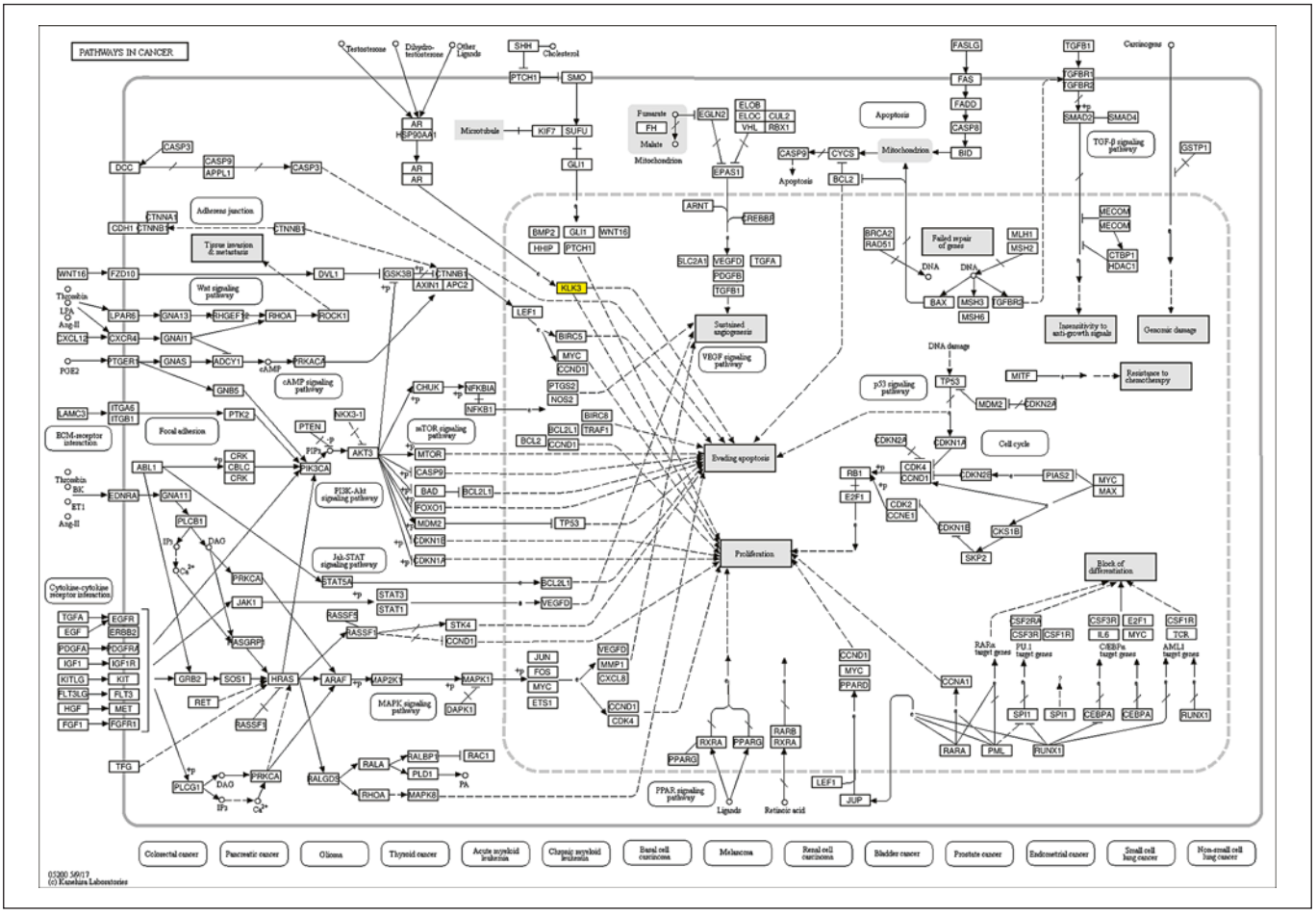


Fig. 4. Coloured pathways in cancer.

Table 6. Folders and the most important files generated by DoriTool

Folder	Files
DAVID	DAVID.html DAVID_formatted.txt
KEGG	hsa.png
VEP	Summary_VEP.html
TopGO	TopGO.txt

DoriTool provides the nearest gene for the intergenic variants as well as the consequence type. Most of the SNPs from our example were intron variants (38.5%) and upstream gene variants (38.5%) and were distributed differently by chromosome. Chromosome 6 and 19 harboured the largest number of variants for this instance (see online suppl. File S9). Figure 4 shows the online supplementary File S10, which represents the pathway in cancer, one of the KEGG pathways obtained after running DoriTool in the example dataset. Note that DoriTool reports the genes coloured according to the number of input variants tagging them.

Running DoriTool is not computationally demanding. Less than 5 min were needed to obtain the results for the in silico functional analyses of the 13 input variants on a 1,600-MHz laptop (running Linux).

Conclusions

DoriTool is a new and automated integrative bioinformatics pipeline that performs a functional in silico analysis of variants previously associated with a trait of interest through a comprehensive annotation. At present, it is the

most complete bioinformatics tool that allows obtaining, processing and interpreting the results from genetic association analyses in a timely manner. DoriTool is mainly a gene-centric tool to find ontologies and pathways, shedding light on the underlying biology and helping the researchers in the interpretation and discussion of the results in a comprehensive and friendly manner.

Acknowledgment

This work has been supported by Fondo de Investigaciones Sanitarias (FIS), Instituto de Salud Carlos III, Spain (Grant #PI1501573).

Disclosure Statement

The authors have no disclosures to declare.

References

1 McVean GA, et al: An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
2 UK10K Consortium: The UK10K project identifies rare variants in health and disease. Nature 2015;526:82–90.
3 NHLBI GO Exome Sequencing Project (ESP): Exome Variant Server. 2014. http://evs.gs.washington.edu/EVS/.
4 Genomics England: The 100,000 Genomes Project. 2014. http://www.genomicsengland.co.uk/the-100000-genomes-project/.
5 Harrow J, et al: GENCODE: producing a reference annotation for ENCODE. Genome Biol 2006;7(suppl 1):S4.
6 Huang Q: Genetic study of complex diseases in the post-GWAS era. J Genet Genomics 2015;42:87–98.
7 Freedman ML, et al: Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 2011;43:513–518.
8 Reimand J, et al: g:Profiler – a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res 2016;44:W83–W89.
9 Falcon S, Gentleman R: Using GOstats to test gene lists for GO term association. Bioinformatics 2007;23:257–258.
10 Bauer S, Grossmann S, Vingron M, Robinson PN: Ontologizer 2.0 – a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 2008;24:1650–1651.
11 González-Pérez A, López-Bigas N: Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 2011;88:440–449.
12 Kumar P, Henikoff S, Ng PC: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–1081.
13 Adzhubei IA, et al: A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–249.
14 Flicek P, et al: Ensembl 2012. Nucleic Acids Res 2012;40:D1.
15 McLaren W, et al: The Ensembl variant effect predictor. Genome Biol 2016;17:122.
16 Yates A, et al: Ensembl 2016. Nucleic Acids Res 2016;44:D710–D716.
17 McLaren W, et al: The Ensembl variant effect predictor. bioRxiv 2016;42374.
18 GTEx Consortium: The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–585.
19 Yates A, et al: The Ensembl REST API: Ensembl data for any language. Bioinformatics 2015;31:143–145.
20 Huang DW, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009;37:1–13.
21 Carlson M: org.Hs.eg.db: Genome Wide Annotation for Human. R Package 2015, version 3.1.2.
22 Alexa A, Rahnenfuhrer J: topGO: Enrichment Analysis for Gene Ontology. October, 2010.
23 Zhao S, Guo Y, Shyr Y: KEGGprofile: An Annotation and Visualization Package for Multi-Types and Multi-Groups Expression Data in KEGG Pathway. R Package, 2015, version 1.18.0.
24 Aibar S, Fontanillo C, Droste C, De Las Rivas J: Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. Bioinformatics 2015;31:1686–1688.
25 Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
26 Fresno C, Fernández EA: RDAVIDWebService: a versatile R interface to DAVID. Bioinformatics 2013;29:2810–2811.