Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Genetic and population analysis

SiNoPsis: Single Nucleotide Polymorphisms selection and promoter profiling

Daniel Boloc [5], Natalia Rodríguez [1], Patricia Gassó [1,6], Josep F Abril [3,4], Miquel Bernardo [2,5,6,7], Amalia Lafuente [1,6,7], Sergi Mas [1,6,7]*

[1]Dep. *Fonaments Clínics, Unitat de Farmacologia*, University of Barcelona, Barcelona, Catalonia, Spain;
[2]Unitat Esquizofrènia, Hospital Clínic de Barcelona, Barcelona, Catalonia, Spain;
[3]Dep. de *Genètica, Microbiologia i Estadística*, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain;
[4]Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Catalonia, Spain;
[5]Dep. de *Medicina*, University of Barcelona, Barcelona, Catalonia, Spain;
[6]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Catalonia, Spain;
[7]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed.

Abstract

Motivation: The selection of a Single Nucleotide Polymorphism (SNP) using bibliographic methods can be a very time-consuming task. Moreover, a SNP selected in this way may not be easily visualized in its genomic context by a standard user hoping to correlate it with other valuable information. Here we propose a web form built on top of `Circos` that can assist SNP-centred screening, based on their location in the genome and the regulatory modules they can disrupt. Its use may allow researchers to prioritize SNPs in genotyping and disease studies.

Summary: `SiNoPsis` is bundled as a web portal. It focuses on the different structures involved in the genomic expression of a gene, especially those found in the core promoter upstream region. These structures include transcription factor binding sites (for promoter and enhancer signals), histones, and promoter flanking regions. Additionally, the tool provides eQTL and linkage disequilibrium (LD) properties for a given SNP query, yielding further clues about other indirectly associated SNPs. Possible disruptions of the aforementioned structures affecting gene transcription are reported using multiple resource databases. `SiNoPsis` has a simple user-friendly interface, which allows single queries by gene symbol, genomic coordinates, Ensembl gene identifiers or RefSeq transcript identifiers. It is the only portal providing useful SNP selection based on regulatory modules and LD with functional variants in both textual and graphic modes (by properly defining the arguments and parameters needed to run `Circos`).

Contact: danielboloc@gmail.com

Supplementary information:

`SiNoPsis` is freely available at `https://compgen.bio.ub.edu/SiNoPsis/`

1 Introduction

Single nucleotide polymorphisms (SNPs) represent a difference in a DNA sequence, which define alternate nucleotide alleles at a given genomic position. They usually occur every 300 nucleotides, meaning that there are about 10 million SNPs in the human genome (Feuk *et al.*, 2006). They can act as biological markers, thus being associated with certain diseases or pharmacogenetic processes either directly (it is causal), or indirectly via linkage disequilibrium (LD) (associated with the causative one). We can find two types of SNPs in a genic coding region: synonymous (with no changes in amino acid) and non-synonymous (with changes in amino acid). However, nucleotide variants can also be found in non-coding regions, in 5'– and 3'–UTR of the transcript structure, as well as in promoters, enhancers, insulators, and post-translational histone modifications, and so on. Therefore, they can disrupt regulatory elements

and alter the expression of a gene by a variety of means (see Bonev *et al*., 2016).

To obtain an idea of the SNPs effects or their positions in the different regulatory elements, researchers have to search different databases manually, which is a time-consuming and difficult process. This difficulty and the lack of user-friendly tools for selecting SNPs based on their position in the different regulatory structures and the LD with a functional variant that contributes to the phenotype, have led us to create this online tool.

SiNoPsis is a simple web application that allows the characterization of SNPs in the context of known annotated human gene or a region defined by genomic coordinates (GRCh37/hg19). Its main goal is to present candidate SNPs related to the regulatory elements for genic/genomic loci that can be pinpointed for further functional assays. The tool includes multiple options in the input form to allow users to choose among a number of genomic features that may be relevant for the subsequent analysis, and generates different outputs: a summary web table, a PDF report and a Circos-based representation of all the selected features in different formats (HTML, PNG and SVG). The PDF report contains the users' input, the resulting image and its legend, tracks description and a table with all the SNPs classified in four categories (see Suppl. Material Table 2), based on the genomic structures that the corresponding SNPs disrupt (ecreSNP, eSNP, creSNP and normSNP), also mapped to the regulatory region and the number of structures they interfere with, ~~the cell lines that have that region,~~ their LD, and, finally, their eQTL properties. The generated image contains the selected gene/genomic locus (with additional genes identified in that region, highlighting exons), its upstream/downstream regions, the relevant regulatory features, as well as the SNPs mapping to that region organized into multiple tracks (see Suppl. Material Figure 1). The results page allows users to download the Circos (Krzywinski *et al*., 2009) configuration files generated along with the data selected, enabling them to customize the final figure layout manually.

2 Methods

Multiple resources from several databases were used to produce the core dataset for the characterization of genes: gene names and Ensembl IDs from HGNC (Gray *et al*., 2015); epigenetics data from the ENCODE project (Rosenbloom *et al*., 2013); enhancers, open chromatin, promoter, promoter flanking regions were retrieved from Ensembl (Bronwen *et al*., 2016); transcription factor binding sites, exons and RefSeq transcripts IDs from UCSC' table browser utility (Karolchik *et al*., 2004); transcription start sites from FANTOM5 (Lizio *et al*., 2014); eQTL information from GTEX project (Lonsdale *et al*., 2013); SNPs from the 1000Genomes database (Sudmant *et al*., 2015); and linkage disequilibrium information from Ensembl. All those datasets were downloaded from the respective repositories and parsed by custom *Perl* scripts to populate a *MySQL*-based functional database. The application forms were built upon PHP scripts that communicate with the database and extract the information requested by the user. All the users' query information is mapped onto the specified genomic region using an upstream/downstream default flanking region of 5 kbp (users may customize this further). In this way, the promoter region together with any cis-regulatory region (i.e., enhancers) lying in the flanks will be included.

SiNoPsis procedure is set into three main steps:

1. SiNoPsis receives the official symbol of a gene, Ensembl gene ID, RefSeq transcript ID or chromosome coordinates and the different options that constitute the query (upstream/downstream flanking distance, as well as cell filters for the chromatin state, regulation and histone marks tracks and SNP population filters). Here, if required, a master SNP can also be provided. This SNP is used to calculate the LD between it and each of the mapped SNPs. It is usually a functional or statistically significant SNP. The form also accepts searching solely by SNP, leaving the gene and genomic loci form fields empty.
2. SiNoPsis uses the captured information to start *SQL* queries and retrieve information from the pre-computed functional database. *PHP* scripts are in charge of distributing the data into temporary files on the server, thus creating files to be used later on by Circos (tracks and configuration files), combining all the different data sources in between the coordinates of interest, and then color-coding each structure (according to Suppl. Material Table 1 color key).
3. SiNoPsis takes the files generated in step 2 and formats them to suit the Circos input requirements in order to produce the drawings and to provide data type-specific customizations. It produces the results page with links to the image in different formats, the web table, and to the PDF report; it also packs all the produced files into a zip file, so that the users can download it and manually adjusts the Circos customization files to improve the resulting maps.

3 Results and Discussion

We illustrate the functionality of this web tool by using the genes ***AKT1***, ***DDIT4***, and ***FCHSD1***, using both, gene name and coordinates as input. The partial results for the ***AKT1*** gene are shown in Suppl. Material Figure 1; the full results for the two other genes are available from Suppl. Material Figures 2 and 3. 94 SNPs were mapped onto the ***AKT1*** gene locus, of which 24 were classified as eSNP (25.5%), 7 as ecreSNP (7.4%), 43 as normSNP (45.7%), and 20 as creSNP (21.3%). Out of all those variants projected, four SNPs were selected for a posterior functional validation assay.

The crucial transfer of mere nucleotide changes to the context of laboratory experiments, treatments and predictors of disease requires the initial non-blind selection of SNPs. SiNoPsis integrates information from multiple biological resources, characterizing all the input and providing tables and visual images as a result. This will yield biological clues with regard to the importance of different SNPs, but leaves it up to the researchers to choose the ones to be analyzed in further experiments, such as genotyping or functional validation. Among the advantages of this web application are its multiple options for different outputs, its acceptance of official gene symbols (also Ensembl and RefSeq identifiers), its ability to allow any genomic region as input, and its generation of informative tables and high-quality images. This enables researchers to prioritize experiment design and analysis rather than time-consuming database searches.

SiNoPsis data should be considered in the SNPs selection and prioritizing for experimental validation and in the analysis of medium-scale sequencing and genotyping projects. SiNoPsis is freely available and will be updated regularly.

Funding

DB is supported by a FPU fellowship (FPU14/06834) from the "Ministerio de Educación, Cultura y Deporte". NR is supported by the University of Barcelona (APIF2015_24782). This study was supported by the Spanish Ministry of Health, Instituto de Salud Carlos III (FIS, Fondo de Investigación Sanitária PI10/02430), and the Catalan Innovation, Universities and Enterprise Authority (Grants DURSI 2014SGR436 and 2014SGR441).

Conflict of Interest: none declared.

References

Bonev, B. *et al.*, (2016) Organization and function of the 3D genome. *Nat Rev Genet.*, **17**, 661–678.

Bronwen, A. *et al.*, (2016) The EnsEmbl gene annotation system. *Database*, **baw093**.

Feuk, L. *et al.* (2006) Structural variation in the human genom. *Nat Rev Genet*, **7**, 85–97.

Gray, KA. *et al.* (2015) genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, **43(Database issue)**, D1079–85.

Karolchik, D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, **32(Database issue)**, D493–6.

Krzywinski, M. *et al.* (2009) `Circos`: an information aesthetic for comparative genomics. *Genome Res*, **19**, 1639–1645.

Lizio, M. *et al.* (2014) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology*, **16**, 22.

Lizio, M. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, **45**, 580–585.

Rosenbloom, K. *et al.* (2013) ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res*, **41(Database issue)**, D56–63.

Sudmant, P. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.