SNPDoc User Guide
Version 2.2.3

Richard T. Guy, Wei Weng, Dr. Carl D. Langefeld

Summary

SNPDoc mines several online databases to record information about SNPs and genomic positions. There are three modes of operation. In the most popular, a list of snps in rs# format are fed into the program and an html formatted output table is produced that can be viewed using any browser or pulled into Excel or OpenOffice for easy manipulation. Captured information includes:

- Position
- Chromosome (and URL at NCBI)
- Gene name, description, and aliases if SNP is in a gene
- Nearest up and downstream gene if SNP is not in a gene.
- Risk score if SNP is in a gene (see summary of risk algorithm below)
- Nearest CpG island
- Variation region type if SNP is in a variation region.
- SNPs that have been merged into the current SNP from previous builds as well as the current SNP ID if requested SNP has been merged into something else.
- UCSC and NCBI urls for the SNP.

For more information and an example use, see the Sample Use section of this guide.

Use Policy

SNPdoc is offered as open source under the GPL license. Any published use of this software should cite:

Richard T. Guy, Wei Wang, Miranda C. Marion, Paula S. Ramos, Timothy Howard, and Carl D. Langefeld, "SNPDoc: Integrating genomic data and statistical results." Submitted to Bioinformatics (OUP), 2010.

Installation Instructions

The installation package has been designed to be self contained with one exception. Most packages used are preinstalled on standard Perl loads, but you may need to install a few database packages from CPAN. Updating to at least Perl version 5 is recommended.

Linux/Unix/Mac

On Linux/Unix/Mac, open a command window, su to root, go to the same directory as this manual and type

>> perl installScript.pl

To run SNPdoc, you call
>> snpdoc <options>

By default, the installation will go to

/usr/local/

and a link will exist in /usr/local/bin/

If you do not have root access, you can use snpdoc by typing

>> perl /path/to/this/folder/bin/snpdoc.pl <options>

If you receive an error saying you need database drivers, the following commands will install what you need:

>> cpan install DBI
>> cpan install DBD::mysql

You will need to answer yes to a lot of options using CPAN, but defaults should suffice for almost anyone.

To remove SNPdoc, erase the files
/usr/local/bin/snpdoc
/usr/local/snpdoc
You will need root permission to access those files.

Windows

On Windows, you must have ActivePerl, available at
http://www.activestate.com/activeperl/

Once ActivePerl is installed, simply type

> perl installScriptMSFT.pl

The script will create a folder called snpdoc in your Program Files. If you add that folder to your path variable, you should be able to access the program by simply typing

> snpdoc <options>

at the command line.

To add the path to snpdoc to your path variable, following the following steps:

1) Click Start->Control Panel
2) Click "Switch to Classic View" available in the upper right hand corner.
3) Double click on "System"
4) Click on the Advanced tab
5) Click Environment Variables
6) Double click on variable Path under system variables. This will bring up a box with the variable value and will look something like:

C:\Perl\site\bin;C:\Perl\bin;%SystemRoot%\system32;%SystemRoot%;%SystemRoot%\System32\Wbem;C:\Program Files\ATI Technologies\ATI Control Panel;C:\Program Files\QuickTime\QTSystem\;C:\Program Files\Windows Imaging\;c:\Program Files\Microsoft SQL Server\100\Tools\Binn\;c:\Program Files\Microsoft SQL Server\100\DTS\Binn\

7) Append ;C:\Program Files\snpdoc\bin to that path and click Ok.
8) Click OK several more times to close the "System" control panel.
9) Open a DOS shell and type snpdoc. You should get an error because you didn't pass any commands to the program, but it should find the program.

Before using SNPdoc, open the file test_results.txt in the installation folder. Follow the instructions for any tests that came back no. This will guide you through the installation of several packages that snpdoc uses.

Mysql database access does NOT come standard with ActivePerl (at least according to the experience of the developers) and must be installed using the following command line option:

> cpan DBI DBD::mysql

This will commence several processes and take at least 5 minutes.

To remove SNPdoc from a Windows machine, simply erase the folder
C:\Program Files\snpdoc
and reverse the steps above to remove the path from your path variable.

Sample Use

SNPdoc works from a command line, so you will need either a shell open. On Windows, click Start->run then type cmd.

There are three types of input to snpdoc:
1) SNP file that looks like:

header, optional extra header, optional other header,…
rs123, optional column, optional column, …
rs245, something, something, …
…

Only the first column is necessary, but you must use rs# format. SNPdoc does not accept other formats at this time.

To run with SNP input, type

snpdoc –snp <snp file> [other parameters]

2) Regional format file that looks like:

name of temp file, optional header1, optional header2, …
chr2:2300-2500, optional col1, optional col2, …
chrX:1-10000, optional col1, optional col2, …
…

With this input, an auxiliary file will be created with the name given that will contain all SNPs in the region(s) specified. Extra columns will be appended to each extra SNP. SNPdoc will run on the SNPs identified as described for optional (1) above.

To run SNPdoc with this option type

snpdoc –reg <regional file> [options]

3) Pos format file that looks like:

name of temp file, optional header1, optional header2, …
chr2:2300, optional col1, optional col2, …
chrX:123, optional col1, optional col2, …
…

With this input, SNPdoc will return position specific information only. Risk and classification are not computed for this option.

To run SNPdoc with this option type

snpdoc –pos <position file> [options]

Note: You must use one of the three options described as the first two inputs to snpdoc. Other options are as follows:

-out Always use this option. This will cause files to be output.
-stamp Optionally include a random number for use in temporary files.
-v Include a lot more output about each SNP as the program runs.
-LD Compute 5 SNPs with highest r^2 with each SNP. This is slow.
-restart=[rs#] If the snp is found in the file, start with the next SNP in the file. Useful if a list stops running midway.

Summary of Risk Score Algorithm

SNPdoc uses a modified version of the FASTSNP algorithm (Yuan et. al. 2006) developed by Wei Weng. The algorithm proceeds as follows:

Risk Calculation

1. Capture all SNP functions from Ensembl through Perl API
2. Capture the following information
    i. TFSEARCH from www.cbrc.jp (Japanese language site). Get highly correlated sequence fragments, report number of such fragments.
    ii. ESEFINDER from rulai.cshl.edu. Get count of motifs that appear with highest threshold (2.676?).
    iii. RESCUEESUS from genes.mit.edu. Contains list of oligonucleotide sequences that enhance premRNA splicing when in exons. Check our sequence against the list.
    iv. FAS-ESE from genes.mit.edu. Get list of ESEs in given sequence. Return count.
3. Look at each function listed for given gene/snp.
    i. If INTERGENIC then risk = 0.
    ii. If STOP_GAINED or STOP_LOST then risk = SG or SL (not in FastSNP)
    iii. If INTRONIC then
        1. If up and down stream TFSEARCH hit counts are equal, risk = 0, classification = "Intronic with no known function."
        2. If up and downstream TFSEARCH hit counts are unequal, risk = 3, classification = "Intronic enhancer."
    iv. If SPLICE_SITE then risk = 3 and classification = "Splice Site".
    v. If 3PRIME_UTR then
        1. If up and down stream TFSEARCH hit counts equal, risk = 0 and classification = "Downstream with no known function."
        2. Otherwise risk = 3 and classification = "Promoter/Regulatory Region."
    vi. If 5PRIME_UTR then do same as 3-prime.

vii. If UPSTREAM then
    1. If up and down stream TFSEARCH hit counts equal then risk = 0, classification = "Upstream with no known function"
    2. Otherwise risk = 3, classification = "Promoter/Regulatory Region."

viii. If DOWNSTREAM then
    1. If up and down stream TFSEARCH hit counts equal then risk = 0, classification = "Upstream with no known function"
    2. Otherwise risk = 3, classification = "Promoter/Regulatory Region."

ix. If SYNONYMOUS_CODING then
    1. If up and down stream ESE counts equal, up and down stream RESCUEESE counts equal, and up and down stream RAS-ESE equal then risk = 1, classification = "Sense/Synonymous."
    2. Otherwise risk = 3, classification = "Sense/Synonymous; Splicing Region."

x. If NON_SYNONYMOUS_CODING
    1. Get the number of SNP functions in Ensembl whos biotype is "protein coding."
    2. If at least one function is of biotype protein coding then
        1. If up and down stream ESE counts equal, up and down stream RESCUEESE counts equal, and up and down stream RAS-ESE equal then risk = 4, classification = "Mis-Sense (Leading to Non-Conservative Change)."
        2. Otherwise risk = 4, classification = "Mis-Sense (Splicing Regulation, Protein Domain Abolished)."
    3. If no function of biotype protein coding then
        1. If up and down stream ESE counts equal, up and down stream RESCUEESE counts equal, and up and down stream RAS-ESE equal then risk = 3, classification = "Mis-Sense (Leading to Conservative Change)
        2. Otherwise risk = 3, classification = "Mis-Sense (Conservative); Splicing Regulation

4. The final risk score and classification is the maximum risk (and its classification) from the above list.

FASTSNP citation:

Hsiang-Yu Yuan, Jen-Jie Chiou, Wen-Hsien Tseng, Chia-Hung Liu, Chuan-Kun Liu, Yi-Jung Lin, Hui-Hung Wang, Adam Yao, Yuan-Tsong Chen, and Chun-Nan Hsu.

FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.*, 1 July 2006; 34: W635 - W641.