SNPDoc User Guide

Richard T. Guy, David R. McWilliams, Wei Wang and Carl D. Langefeld

June 26, 2012

Contents

1 Summary

SNPDoc mines several on-line databases to record information about single nucleotide polymorphisms (SNPs) and genomic positions. There are three options for input: a list of SNPs in 'rs' format, a list of chromosomal regions, or a list of chromosomal positions. Output is to a tab-delimited file (for ease of spreadsheet import) or to an html file for browser viewing. Information captured by SNPDoc includes:

- Position

- Chromosome (and link to NCBI)
- Gene name, description, and aliases if the SNP is in a gene
- Nearest upstream and downstream genes if the SNP is not in a gene
- Risk score if the SNP is in a gene (see below for the risk algorithm)
- Position of the nearest CpG island
- Variation type if the SNP is in a variation region
- SNP merge information
- Links to the NCBI and UCSC information for the SNP

In all cases, user data in the file will be merged with SNPDoc output. This is of particular benefit analyzing and ranking the results of GWAS studies.

2 Installation

SNPDoc was constructed to be as self-contained as possible. The install script may complain of missing modules, however. Platform-specific instructions are given below for installing missing modules.

2.1 Linux/Unix/Mac

Download the archive file and unpack in a working directory. Change to the snpdoc directory and as root enter:

> perl installScript.pl

By default the software will be installed in *usr/local* with a link pointing to it in *usr/local/bin*. Type 'snpdoc.pl –help' to verify operation.

If you do not have root access, you can run snpdoc by giving the fully qualified path to the snpdoc.pl executable. For example:

> perl /home/username/path/to/snpdoc.pl <options>

If you receive errors describing missing modules, install them using CPAN. For example if the module 'DBI' is missing, enter as root:

> cpan install DBI

You may also be able to use your package manager to install packages.

2.2 Windows

If you do not have Perl installed on your machine obtain it from http://www.activestate.com/activeperl/. Once it installed, open a command window and change to the directory where you downloaded snpdoc and type:

> perl installScriptMSFT.pl

The script will create a folder called 'snpdoc' in your Program Files directory. Open the file 'test_results.txt' in the installation directory and check for any tests that failed. There will be instructions for installing perl modules that snpdoc needs.

You may have to add snpdoc to your command path. Do the following:

1. Click Start-> Control Panel
2. Click "Switch to Classic View"
3. Double click "System"
4. Click on the Advanced tab
5. Click Environment Variables
6. Double click on the variable Path. This will bring up a box with contents that look similar to:

   `C:\Perl\site\bin;C:\Perl\bin;\%SystemRoot\%\system32;\%SystemRoot\%;...`

7. Find the end of the text and append `';C:\Program Files\snpdoc\bin'` (include the semi-colon but not the quotation marks) and click OK.
8. Click OK several more times to close the "System" control panel

You should be able to run snpdoc by typing in a command window:
`C:\My_Project\snpdoc <options>`

3 Usage

3.1 General

Running snpdoc without command line options or with the –help flag prints the following usage information.

```
usage:
snpdoc [options] -infile FILE, where option is one or more of:

--help          print this help message

--infile        input file (required)

--search        search type; one of "snp", "reg", "pos" (default "snp")

--outfile       output file name; if not specified it will be created
                from the input file name.

--outformat     type of output; one of "text" or "html" (default "text")

--sep           field delimiter in the input file; currently tab and comma
                are recognized (supply with quotes as "\t" or ",");
                default comma

--db            use a database to save and retrieve results

--dbname        name of the database

--user          database username

--stamp         include a random number for use in temporary files

--verbose       print more information to the console as snpdoc runs

--ucsc\_version  set the UCSC database version; currently hg18 and hg19
                are recognized (default 19)

--restart       a snp designation; if given, processing will start at
                this snp in the file
```

Flags may also be given as '-h', '-o', etc., if the single letter uniquely specifies an option.

3.2 SNP Search

Running SNPDoc with the '–search snp' option (the default) will search a number of databases and aggregate this information with information supplied by the user (e.g. statistics from a GWAS study). The expected file format has a header line and data lines following, with the SNP in the first column. Only the 'rs#' format is currently recognized for a search. Empty results fields will be printed for non-standard names. Any further columns are retained and appended to the columns output by SNPDoc.

3.3 Positional or Regional Search

Running SNPDoc with the '–search reg' option performs a 'regional search'. The expected file format has a header line and the first column with chromosomal regions listed as 'chr2:2300-2500', for example. The region is searched and any SNPs found are output in a format suitable for the 'SNP' search described previously. If user data is supplied for the region, this data will be printed for each SNP found in the region. This may create a very large output file if your supplied region is large.

Running SNPDoc with the '–search pos' option performs a 'positional search'. The expected file format has a header line and the position description should be in the form 'chr2:1234', for example. If the position corresponds to a named snp, that name will be printed in the first column of output. If the snp is not named, risk and classification are not computed. User data is merged with results.

3.4 Summary of the Risk Score Algorithm

SNPDoc uses a modified version of the FASTSNP algorithm (Yuan et al., 2006) developed by Wei Wang. The algorithm proceeds as follows.

3.4.1 Collect the following information for each SNP

- SNP function annotations from UCSC
- Transcription factor binding site information from the TFSEARCH www.cbrc.jp/research/db/TFSEARCH.html database (Akiyama, 2011)
- Significant exonic splicing enhancer (ESE) motifs found by ESEfinder rulai.cshl.edu,(Cartegni, 2003)
- Significant exonic splicing enhancer motifs found by RESCUE-ESE genes.mit.edu/burge-lab/rescue-ess, (Fairbrother, 2002)

- Significant exonic splicing silencer motifs found by FAS-ESS genes.mit.edu/fas-ess, (Wang, 2004)

3.4.2 Examine the function for the SNP and score as follows

- If INTERGENIC then risk = 0
- If STOP_GAINED or STOP_LOST then risk = 5
- If INTRONIC
  - If the TFSEARCH results are equivalent for both alleles, risk = 0, classification = "Intronic with no known function"
  - If the TFSEARCH results are not equivalent, risk = 3, classification = "Intronic enhancer"
- If SPLICE_SITE then risk = 3 and classification = "Splice site"
- If 3PRIME_UTR then
  - If the TFSEARCH results are equivalent for both alleles, risk = 0, classification = "Downstream with no known function"
  - If the TFSEARCH results are not equivalent, risk = 3, classification = "Promoter/Regulatory region"
- If 5PRIME_UTR then proceed as for 3PRIME_UTR
- If UPSTREAM then
  - If the TFSEARCH results are equivalent for both alleles, risk = 0, classification = "Upstream with no known function"

- If the TFSEARCH results are not equivalent, risk = 3, classification = "Promoter/Regulatory region"

- If DOWNSTREAM then

  - If the TFSEARCH results are equivalent for both alleles, risk = 0, classification = "Downstream with no known function"

  - If the TFSEARCH results are not equivalent, risk = 3, classification = "Promoter/Regulatory region"

- If SYNONYMOUS_CODING then

  - If the ESE found by ESEfinder are equivalent for each allele, the ESE found by RESCUE-ESE are equivalent, and the splicing silencers found by FAS-ESE are equivalent then risk = 1 and classification = "Sense/Synonymous"

  - Otherwise risk = 3 and classification = "Sense/Synonymous; Splicing Region"

- If NON_SYNONYMOUS_CODING then

  - Get the number of SNP functions in Ensembl whose biotype is "protein coding."

  - If at least one function is of biotype protein coding then

    * If the ESE found by ESEfinder are equivalent for each allele, the ESE found by RESCUE-ESE are equivalent, and the splicing silencers found by FAS-ESE are equivalent then risk = 4 and classification = "Mis-Sense (Leading to Non-Conservative Change)."

* Otherwise risk = 4, classification = "Mis-Sense (Splicing Regulation, Protein Domain Abolished)

– If no function of biotype protein coding then

* If the ESE found by ESEfinder are equivalent for each allele, the ESE found by RESCUE-ESE are equivalent, and the splicing silencers found by FAS-ESE are equivalent then risk = 3, classification = "Mis-Sense (Leading to Conservative Change)"

* Otherwise risk = 3, classification = "Mis-Sense (Conservative); Splicing Regulation"
The final risk score is the maximum from the above heuristic and the classification is that associated with it

4 References

- Akiyama, Yutaka "TFSEARCH: Searching Transcription Factor Binding Sites", Computational Biology Research Center (CBRC), AIST , Japan. (Citation retrieved from the website in March 2011).
- Cartegni L., Wang J., Zhu Z., Zhang M. Q., Krainer A. R.; 2003. ESEfinder: a web resource to identify exonic splicing enhancers. Nucleic Acid Research, 2003, 31(13): 3568-3571.
- Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 2002 Aug 9;297(5583):1007-13.
- Guy, R.T., Wang, W., Marion, M.C, Ramos, P.S., Howard, T., and Langefeld, C.D., "SNPDoc: Integrating genomic data and statistical results." [Submitted]
- 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.

- Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M. and Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell 119, 831-845.