SNPdat: rapid annotation of novel SNPs for model and non-model organisms

Anthony G Doran[1] and Christopher J Creevey[1]*

[1] Teagasc Animal and Bioscience Research Department, Animal & Grassland Research and Innovation Centre, Teagasc, Grange, Dunsany, Co. Meath, Ireland.Contact details and corresponding author

*corresponding author: chris.creevey@teagasc.ie and anthony.doran@teagasc.ie

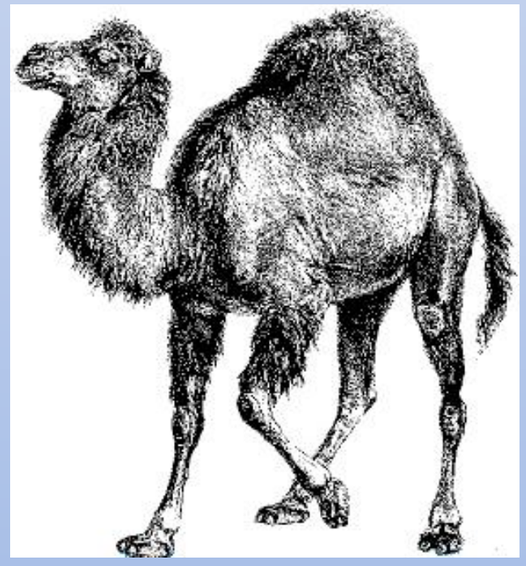
What is SNPdat?

SNPdat (**SNP D**ata **A**nalysis **T**ool) is a command line tool for rapid annotation of both novel and known single nucleotide polymorphisms (SNPs) for any organism with a reference sequence and annotation (even those in draft status)

Before running SNPdat?

SNPdat can be run on any machine that has a recent version of Perl installed.

Perl is a high-level, general-purpose, interpreted, dynamic programming language. Current versions of Perl are available from www.perl.org



Why use SNPdat?

SNPdat

•is an easy to use command-line tool for annotating the results of *de novo* SNP discovery projects.

•is specifically designed for use with organisms which are either not supported by other tools or have a small number of annotated SNPs available

•can also be used to analyse datasets from organisms which are densely sampled for SNPs.

•is a high throughput analysis tool that can provide a comprehensive annotation of both novel and known SNPs for any organism with a draft sequence and annotation.

SNPdat will be of most use to researchers involved in de novo SNP discovery studies for organisms with poor (or non at all) coverage of SNPs in databases such as dbSNP

Where to get SNPdat?

SNPdat is freely available from:

http://code.google.com/p/snpdat

You will also find sample files, further information on SNPdat and a manual here.

snpdat

SNPdat - A Simple High Throughput Analysis Tool for Annotating SNPs

Project Home | Downloads

Summary | People

Project Information

Recommend this on Google

Project feeds

Code license
GNU GPL v2

Labels
Perl, SingleNucleotidePolymorphism, Annotation, SNP, SNPanalysis, gtf, gff, fasta, sequenceanalysis

Members
d1an...@gmail.com, anthony....@teagasc.ie, chris.cr...@gmail.com

SNPdat (**SNP** **D**ata **A**nalysis **T**ool) is a high throughput analysis tool that can provide a comprehensive annotation of both novel and known single nucleotide polymorphisms (SNPs). It is specifically designed for use with organisms which are either not supported by other tools or have a small number of annotated SNPs available, however it can also be used to analyse datasets from organisms which are densely sampled for SNPs. It can be used for analysis of any organism with a draft sequence and annotation. SNPdat makes possible analyses involving non-model organisms that are not supported by the vast majority of SNP annotation tools currently available.

You can easily download the software, manual, sample dataset and short tutorial on SNPdat using the downloads tab on the left (http://code.google.com/p/snpdat/downloads/). Additional scripts and a sample input file for SNPdat are available on the downloads page.

Table of Contents

Input files

SNPdat requires three mandatory input files:

1. User input file
2. Gene annotation file (GTF)
3. FASTA sequence file

And one optional input file:

4. A processed FLAT file

User input file

SNPdat accepts two file formats as user input.
This file must be specified on the command line using the -i switch when using SNPdat. This will be explained in greater detail later.

- **The first user input file format** accepted by SNPdat is a simple tab-delimited text file.
  This file should not have any header information. Any line beginning with # will be skipped.
  This file should contain three columns:

```
chromosome_id    position        mutation
```

e.g. tab-delmited text file

```
25      1234568 A
X       1234568 T
19      1234568 G
1       1234568 C
```

The chromosome IDs used in the input file must be the same format in ALL input files

User input file

- **The second user input file format** accepted is a Variant Calling Format (VCF) file.
  This file should be tab-delimited. This file must have a header line that begins with **##fileformat=vcf**.
  Any line after the first that begins with # will be skipped. This file should have as its first five columns:

```
chromosome_id   position        snp_id  ref_base        mutation
```

Any column after the first five will be ignored by SNPdat

e.g. VCF file

```
##fileformat=vcfv4.0
##_This line will be ignored by SNPdat_
#_This line will also be ignored_
##_ This line will be ignored_
25      1234568 SNPid1  T       A
X       1234568 SNPid2  C       T
19      1234568 SNPid3  C       G
19      1234568 SNPid4  C       A
1       1234568 SNPid5  G       C
```

The chromosome IDs used in the input file must be the same format in ALL input files

FASTA format sequence File

The second file is a DNA FASTA format sequence file for the species of interest. FASTA format is a text based format that is standard for storing sequence information.

FASTA files can be downloaded from Ensmbl at:
http://www.ensembl.org/info/data/ftp/index.html

Or using GTF_FASTA_finder.pl

This file must be specified on the command line using the -f switch when using SNPdat. This will be explained in greater detail later.

The chromosome IDs used in the FASTA file must be the same format in ALL input files

```
>25
TGATCGATAAGATGTCCCAGCAGGGATTCTCTCTCTGGGCAGGTGAGGTCTTCTCTTAAA
GAAACCCCATATATAAGGGATTTTTAGGCCTTTTTGGATCAAAATTATCCCAGTTTTTCA
AGATATTATTCAAGGGGCGTGGTTCTGGACTGTTAGGCCCCATCTGTAAGAGAGAAAAAA
GCAGTGCCCAGTGCCATGGCTTTTTTTCTACCAGAGGTGTCTGCCCCTGCTCTAGACGGGG
TTGGAGACTGATTCTCCAACTAAGCTTCCTTCCCTGGCCAGAGCTAACCTGTGTCTTACA
GGTGAAGGTGTCCTACTCCCAACCCCCCTCGTTCTACCACCCAGGTATGGTGGAGATGCA
CCAGGGTTAGACCTGGTGGCATCCCAGATTGATGTCCAGGTCCTCACCATTAAAACCTTG
GTTTGATTTCCCTTTTTATTCCGGCACCACCTGGGGTGGGACAGAGTAGTTTTCCGGATGC
TTGCTGAGCCAAGACCACTAGGGGAAGCACCCAACTTCTGTGTACCTGTGCACATTCCTC
```

Gene Transfer File

The final mandatory input file is a gene annotation file. This file should be in gene transfer format (GTF). GTF files are a standard format for storing information on gene structure. Further information can be found here: http://genome.ucsc.edu/FAQ/FAQformat.html#format4.

GTF files can be downloaded from Ensembl at:
http://www.ensembl.org/info/data/ftp/index.html

Or using GTF_FASTA_finder.pl

This file must be specified on the command line using the -g switch when using SNPdat. This will be explained in greater detail later.

The chromosome IDs used in the GTF file must be the same format in ALL input files

```
25  protein_coding  exon  47304  47778  .  +  .  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "1"; gene_name "LOC789014"; transcript_name "LOC789014";
25  protein_coding  CDS   47304  47778  .  +  0  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "1"; gene_name "LOC789014"; transcript_name "LOC789014"; protein_id "ENSBTAP00000048803";
25  protein_coding  exon  47791  48026  .  +  .  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "2"; gene_name "LOC789014"; transcript_name "LOC789014";
25  protein_coding  CDS   47791  48026  .  +  2  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "2"; gene_name "LOC789014"; transcript_name "LOC789014"; protein_id "ENSBTAP00000048803";
25  protein_coding  exon  48031  48204  .  +  .  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "3"; gene_name "LOC789014"; transcript_name "LOC789014";
25  protein_coding  CDS   48031  48204  .  +  0  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "3"; gene_name "LOC789014"; transcript_name "LOC789014"; protein_id "ENSBTAP00000048803";
25  protein_coding  exon  48795  48818  .  +  .  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "4"; gene_name "LOC789014"; transcript_name "LOC789014";
25  protein_coding  CDS   48795  48818  .  +  0  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "4"; gene_name "LOC789014"; transcript_name "LOC789014"; protein_id "ENSBTAP00000048803";
25  protein_coding  exon  48900  48908  .  +  .  gene_id "ENSBTAG00000037750"; transcript_id "ENSBTAT00000054257"; exon_number "5"; gene_name "LOC789014"; transcript_name "LOC789014";
```

Ensembl Downloads Page

http://www.ensembl.org/info/data/ftp/index.html

FTP Download

API Code

If you do not have access to CVS, you can obtain our latest API code as a gzipped tarball:

Download complete API for this release

Note: the API version needs to be the same as the databases you are accessing, so please use CVS to obtain a previous version if querying older databases.

Database dumps

Entire databases can be downloaded from our FTP site in a variety of formats. Please be aware that some of these files can run to many gigabytes of data.

Looking for MySQL dumps to install databases locally? See our web installation instructions for full details.

Each directory on ftp.ensembl.org contains a README file, explaining the directory structure.

Custom data sets

If you want to filter or customise your download, please try Biomart, a web-based querying tool.

Multi-species data

Database						
Comparative genomics	MySQL	EMF	BED	XML		Ancestral Alleles
BioMart	MySQL	-	-	-		-

Single species data

Show All entries — Show/hide columns — Filter

Species	DNA (FASTA)	cDNA (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (EMF)	Variation (GVF)	Variation (VEP)	Regulation (GFF)	Data files	BAM
Ailuropoda melanoleuca (Panda)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-
Anolis carolinensis (Anole lizard)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-
Bos taurus (Cow)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	GVF	VEP	-	-	-
Caenorhabditis elegans (Caenorhabditis elegans)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-
Callithrix jacchus (Marmoset)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-
Canis familiaris (Dog)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	GVF	-	-	-	BAM
Cavia porcellus (Guinea Pig)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-
Choloepus hoffmanni (Sloth)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-
Ciona intestinalis (C.intestinalis)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-
Ciona savignyi (C.savignyi)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-
Danio rerio (Zebrafish)	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	GVF	VEP	-	-	-
Dasypus novemcinctus	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-

GTF_FASTA_finder.pl

- This is an interactive script designed to retrieve FASTA (dna) and GTF files from Ensembl. You will need to be connected to the internet to use this script. This is written in Perl but uses the system call cURL to retrieve the information from Ensembl. cURL is a part of most Linux distributions and Mac OS X and can also be provided for windows through cygwin, which is a collection of tools that provide a Linux-like environment for windows. To run this script simply type **'perl GTF_FASTA_finder.pl'** into your terminal and follow the prompts.

e.g.

```
user@server $HOME/directory/with_all/SNPdat/files/
$ perl GTF_FASTA_finder.pl
|36| /pub/release-56     |37| /pub/release-57     |38| /pub/release-58
|39| /pub/release-59     |40| /pub/release-60     |41| /pub/release-61
|42| /pub/release-62     |43| /pub/release-63     |44| /pub/release-64
|45| /pub/release-65     |46| /pub/release-66     |47| /pub/release-67

Please select a release to choose from by typing its number:
47
You chose release-67

|0| ailuropoda_melanoleuca
|1| ancestral_alleles
|2| anolis_carolinensis
|3| bos_taurus
.
.
.
|56| tursiops_truncatus
|57| vicugna_pacos
|58| xenopus_tropicalis

Please select an organism (enter the corresponding number) to retrieve the FASTA file for:
46

You chose organism Saccharomyces_cerevisiae
Now retrieving the relevant FASTA file for this organism
############################################################################## 100.0%

|0| ailuropoda_melanoleuca
|1| anolis_carolinensis
|2| bos_taurus
.
.
.
|55| tursiops_truncatus
|56| vicugna_pacos
|57| xenopus_tropicalis

Please select an organism (enter the corresponding number) to retrieve the GTF file for:
45

Now retrieving the selected GTF file
Retrieving GTF: Saccharomyces_cerevisiae.EF4.67.gtf.gz
############################################################################## 100.0%
FASTA information has been retrieved for saccharomyces_cerevisiae from release release-67 of ensembl
GTF information has been retrieved for saccharomyces_cerevisiae from release release-67 of ensembl

.gtf.gz and .fa.gz files need be unzipped using the command 'gzip -d filename'

If you have any queries regarding SNPdat or the additional scripts please consult the website:
http://code.google.com/p/snpdat/
```

Optional Input file

SNPdat has one optional input file. This file is a processed dbSNP FLAT file. This file can be used to identify SNPs that already have annotations in dbSNP. ASN_FLAT files can be downloaded from ftp://ftp.ncbi.nih.gov/snp/organisms/.

Or using dbSNP_finder.pl

This FLAT file must first be processed using the SNPdat_parse_dbsnp.pl script. This will create a file in the input format required for SNPdat. Scripts for processing FLAT files from other databases are available from the authors on request.

Once this file has been processed it can supplied as an input file for SNPdat using the –d switch. This will be explained in greater detail later.

```
10      86180529        rs17870235
10      86180332        rs17870236
10      37342771        rs17870237
10      83618673        rs17870252
10      83618951        rs17870253
10      83619862        rs17870254
10      83619857        rs17870255
4       71663812        rs17870258
4       71663942        rs17870259
4       71664184        rs17870260
25      21649618        rs42767860
25      21649613        rs42767861
25      14097644        rs43615187
25      21649549        rs43639289
```

dbSNP_finder.pl

Additional script for retrieving FLAT files from dbSNP

you can use '**dbSNP_finder.pl**' to get dbSNP flat files for organisms contained in ftp://ftp.ncbi.nih.gov/snp/organisms/

e.g.

```
user@server $HOME/directory/with_all/SNPdat/files/
$ perl dbSNP_finder.pl

|0| Alectoris_9077/
|1| Bos_29061/
|2| almond_3755/
.
.
.
|109| zebrafish_7955/
|110| zebu_9915/
|111| zostera_29655/

Please select and organism by typing its corresponding number
6
You have chosen: arabidopsis_3702/

Retrieving file: 'ds_flat_ch1.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_ch2.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_ch3.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_ch4.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_ch5.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_chMasked.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_chMulti.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_chNotOn.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_chPltd.flat.gz'
################################################################################ 100.0%
Retrieving file: 'ds_flat_chUn.flat.gz'
################################################################################ 100.0%


Now you will need to unzip these files.
This can be done using gzip -d
e.g.
gzip -d ds_*.gz

If you wish to use these (unzipped) files for SNPdat please run 'SNPdat_parse_dbSNP.pl'  to process them first
You may want to join these files before using SNPdat_parse_dbSNP.pl
To join these files you can use the 'join' command
e.g.
join ds*.flat > ds_arabidopsis_3702.all.flat
```

Output file

SNPdat creates a single output file with upto 25 columns of information about each queried SNP. This output file is tab-delimited and can easily be opened with any spreadsheet software such as Microsoft Excel and OpenOffice.

An output file can be specified using the -o switch. This will be explained in greater detail later. If no output file is specified SNPdat will create one with the same name as the input file but with the suffix .result

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Chromsor	SNP Posit	Within a	Region	Distance	Feature	Number c	Number c	Start of cu	End of cu	gene ID c	gene nan	transcript	transcript	Annotate	Strand se	Annotate	Estimate	Estimate	Codon	Amino Ac	synonym	Protein ID	RSidentif	Addition
2	25	117035	Y	Exonic	NA	exon	2	[1/1]	117033	117108	ENSBTAG	RHDF1_B	ENSBTATC	RHDF1_B	[15/18]	-	NA	*-2	0	[C/T]AG	[Q/-]	N	NA	NA	
3	25	117035	Y	Exonic	NA	CDS	2	[1/1]	117033	117108	ENSBTAG	RHDF1_B	ENSBTATC	RHDF1_B	[15/18]	-	-1	NA	0	[C/T]AG	[Q/-]	N	ENSBTAPC	NA	
4	25	639	N	Intergeni	216	exon	NA	NA	304	423	ENSBTAG	NA	ENSBTATC	NA	[1/1]	+	NA	NA	NA	NA	NA	NA	NA	NA	
5	25	646	N	Intergeni	223	exon	NA	NA	304	423	ENSBTAG	NA	ENSBTATC	NA	[1/1]	+	NA	NA	NA	NA	NA	NA	NA	rs132899415	
6	25	769	N	Intergeni	346	exon	NA	NA	304	423	ENSBTAG	NA	ENSBTATC	NA	[1/1]	+	NA	NA	NA	NA	NA	NA	NA	NA	
7	25	782	N	Intergeni	359	exon	NA	NA	304	423	ENSBTAG	NA	ENSBTATC	NA	[1/1]	+	NA	NA	NA	NA	NA	NA	NA	NA	
8	25	1211	N	Intergeni	788	exon	NA	NA	304	423	ENSBTAG	NA	ENSBTATC	NA	[1/1]	+	NA	NA	NA	NA	NA	NA	NA	NA	
9	25	1254	N	Intergeni	831	exon	NA	NA	304	423	ENSBTAG	NA	ENSBTATC	NA	[1/1]	+	NA	NA	NA	NA	NA	NA	NA	NA	
10	25	1288	N	Intergeni	865	exon	NA	NA	304	423	ENSBTAG	NA	ENSBTATC	NA	[1/1]	+	NA	NA	NA	NA	NA	NA	NA	NA	
11	25	2125	N	Intergeni	1702	exon	NA	NA	304	423	ENSBTAG	NA	ENSBTATC	NA	[1/1]	+	NA	NA	NA	NA	NA	NA	NA	rs136577265	

Column Number	Description
1	The queried SNPs chromosome ID
2	The queried SNPs genomic location
3	Whether or not the SNP was within a feature
4	Region containing the SNP; either exonic, intronic, or intergenic
5	Distance to nearest feature
6	Either the closest feature to the SNP or the feature containing the SNP
7	The number of different features that the SNP is annotated to
8	The number of annotations of the current feature
9	Start of feature (bp)
10	End of feature (bp)
11	The gene ID for the current feature
12	The gene name for the current feature
13	The transcript ID for the current feature
14	The transcript name for the current feature
15	The exon that contains the current feature and the total number of annotated exons for the gene containing the feature
16	The strand sense of the feature
17	The annotated reading frame (when contained in GTF)
18	The reading frame estimated by SNPdat
19	The estimated number of stop codons in the estimated reading frame
20	The codon containing the SNP, position in the codon and reference base and mutation
21	The amino acid for the reference codon and new amino acid with mutation in place
22	Whether or not the mutation is synonymous
23	The protein ID for the current feature
24	The RS identifier for queries that map to known SNPs
25	Error messages, warnings etc

An example

Sample input files can be downloaded from
http://www.code.google./p/snpdat/downloads

From here you can download SNPdat. You can also, download the SNPdat package, which includes all additional scripts, the manual, a tutorial and sample data.

snpdat

SNPdat - A Simple High Throughput Analysis Tool for Annotating SNPs

Project Home | Downloads | Administer

New download | Search | Current downloads | for | Search

	Filename ▼	Summary + Labels ▼
☆ ⬇	SNPdat.pl	The is the main SNPdat program Featured SNPdat SNP annotation
☆ ⬇	SNPdat_package.tar.gz	This contains the entire SNPdat package incl sample data, additional scripts and all documentation Featured Perl single nucleotide polymorphism SNP annotation SNPdat sample dataset
☆ ⬇	SNPdat_package.zip	This contains the entire SNPdat package incl sample data, additional scripts and all documentation Featured Perl Single nucleotide polymorphism SNP annotation SNPdat sample dataset

There should be 4 folders and the SNPdat script within this download

- Tutorial – two files
  - SNPdat_tutorial.ppt
  - SNPdat_tutorial.pdf

- additional_scripts – three files
  - dbSNP_finder.pl
  - GTF_FASTA_finder.pl
  - SNPdat_parse_dbsnp.pl

- Manual – two files
  - SNPdat_manual.doc
  - SNPdat_manual.pdf

- Sample dataset – 6 files
  - SNPdat_sample_input.txt
  - SNPdat_sample_input.txt.output
  - SNPdat_sample_UMD3.1.gtf
  - SNPdat_sample_UMD3.1.fa
  - SNPdat_dbSNP_example.flat
  - dbSNP_processed.txt

SNPdat.pl

An example

Once these files have been downloaded, extract them to a location you can navigate to using the terminal. Ensure all files are in the same directory.

Once this has been done, the first thing you will want to do is process the dbSNP FLAT file using **SNPdat_parse_dbsnp.pl**
This is only a sample of the full FLAT file available from ftp://ftp.ncbi.nih.gov/snp/organisms/. dbSNP files may contain information on multiple genome builds for the same organism, so you will also need to know what build you are interested in.
In the below example we are interested in Bos_taurus_UMD_3.1
e.g.

```
adoran@server ~/directory/with_all/SNPdat_files/
$ perl SNPdat_parse_dbsnp.pl SNPdat_dbSNP_example.flat dbSNP_processed.txt

This is a perl script to process a dbSNP FLAT file of SNP data and produce
an output file that can be supplied as input for the software SNPdat

Please enter the assembly you want to map rs ids to
:Bos_taurus_UMD_3.1

The resulting output file is dbnsp_processed.txt
This file can be used as input for SNPdat using the -d switch
```

SNPdat_parse_dbsnp.pl

SNPdat_parse_dbsnp.pl is and additional script provided with the SNPdat package to parse dbSNP FLAT files and create a file in a format suitable for use with SNPdat

```
adoran@server ~/directory/with_all/SNPdat_files/
$ perl SNPdat_parse_dbsnp.pl SNPdat_dbSNP_example.flat dbSNP_processed.txt

This is a perl script to process a dbSNP FLAT file of SNP data and produce
an output file that can be supplied as input for the software SNPdat

Please enter the assembly you want to map rs ids to
:Bos_taurus_UMD_3.1

The resulting output file is dbnsp_processed.txt
This file can be used as input for SNPdat using the -d switch
```

The parsed output file (dbSNP_processed.txt in the above example), is a simple tab-delimited file with 3 columns
The first column is the chromosome name
The second column is the coordinate on the chromosome of the SNP
The third column is the mutation

Running SNPdat

Now you have created the optional input file and can run SNPdat.

When using SNPdat you must specify each file type on the command line.

NOTE – Chromosome Identifiers must be the same format across all input files

To specify file types for SNPdat a unique switch is used

-i input file
-g GTF file
-f FASTA file
-o output file
-d processed dbSNP file
-s summary report file

```
adoran@server ~/directory/with_all/SNPdat_files/
$ perl SNPdat.pl -i SNPdat_sample_input.txt -g SNPdat_sample_UMD3.1.gtf -f SNPdat_sample_UMD3.1.fa -d dbSNP_processed.txt -o SNPdat_sample_input.txt.output
```

SNPdat output

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Chromsom	SNP Positi	Within a Fe	Region	Distance to	Feature	Number of	Number of	Start of cur	End of curr	gene ID co	gene name	transcript II
2	CHR25	1818721	N	Intronic	470	exon	NA	NA	1818078	1818251	ENSBTAG	E1BCR1_E	ENSBTAT
3	CHR25	1818721	N	Intronic	470	CDS	NA	NA	1818078	1818251	ENSBTAG	E1BCR1_E	ENSBTAT
4	CHR25	1818938	N	Intronic	687	exon	NA	NA	1818078	1818251	ENSBTAG	E1BCR1_E	ENSBTAT
5	CHR25	1818938	N	Intronic	687	CDS	NA	NA	1818078	1818251	ENSBTAG	E1BCR1_E	ENSBTAT
6	CHR25	1819363	N	Intronic	547	exon	NA	NA	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
7	CHR25	1819363	N	Intronic	547	CDS	NA	NA	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
8	CHR25	1819436	N	Intronic	474	exon	NA	NA	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
9	CHR25	1819436	N	Intronic	474	CDS	NA	NA	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
10	CHR25	1819646	N	Intronic	264	exon	NA	NA	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
11	CHR25	1819646	N	Intronic	264	CDS	NA	NA	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
12	CHR25	1819682	N	Intronic	228	exon	NA	NA	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
13	CHR25	1819682	N	Intronic	228	CDS	NA	NA	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
14	CHR25	1819940	Y	Exonic	NA	exon	2	[1/1]	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT
15	CHR25	1819940	Y	Exonic	NA	CDS	2	[1/1]	1819910	1820030	ENSBTAG	E1BCR1_E	ENSBTAT

SNPdat output

	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	transcript I(	transcript r	Annotated I	Strand sen	Annotated (	Estimated	Estimated	Codon	Amino Acic	synonymou	Protein ID (	RSidentifie	Additional Not
2	ENSBTAT(	E1BCR1_E	[8/30]	-	NA	NA	NA	NA	NA	NA	NA	NA	
3	ENSBTAT(	E1BCR1_E	[8/30]	-	-2	NA	NA	NA	NA	NA	ENSBTAP(	NA	
4	ENSBTAT(	E1BCR1_E	[8/30]	-	NA	NA	NA	NA	NA	NA	NA	NA	
5	ENSBTAT(	E1BCR1_E	[8/30]	-	-2	NA	NA	NA	NA	NA	ENSBTAP(	NA	
6	ENSBTAT(	E1BCR1_E	[7/30]	-	NA	NA	NA	NA	NA	NA	NA	NA	
7	ENSBTAT(	E1BCR1_E	[7/30]	-	0	NA	NA	NA	NA	NA	ENSBTAP(	NA	
8	ENSBTAT(	E1BCR1_E	[7/30]	-	NA	NA	NA	NA	NA	NA	NA	NA	
9	ENSBTAT(	E1BCR1_E	[7/30]	-	0	NA	NA	NA	NA	NA	ENSBTAP(	NA	
10	ENSBTAT(	E1BCR1_E	[7/30]	-	NA	NA	NA	NA	NA	NA	NA	NA	
11	ENSBTAT(	E1BCR1_E	[7/30]	-	0	NA	NA	NA	NA	NA	ENSBTAP(	NA	
12	ENSBTAT(	E1BCR1_E	[7/30]	-	NA	NA	NA	NA	NA	NA	NA	NA	
13	ENSBTAT(	E1BCR1_E	[7/30]	-	0	NA	NA	NA	NA	NA	ENSBTAP(	NA	
14	ENSBTAT(	E1BCR1_E	[7/30]	-	NA	*-1	0	C[A/T]G	[Q/L]	N	NA	NA	
15	ENSBTAT(	E1BCR1_E	[7/30]	-	0	NA	0	C[A/T]G	[Q/L]	N	ENSBTAP(	NA	

Licence

SNPdat is freely available under a GNU Public License (Version 2 http://www.gnu.org/licenses/gpl-2.0.html) at: http://code.google.com/p/snpdat

This tutorial is only a brief summary and introduction to SNPdat and how to use SNPdat. For more information on using SNPdat or any of the available options (not discussed here) please read the manual

The manual is available from the website (http://code.google.com/p/snpdat/).

The published manuscript for SNPdat is available from BMC Bioinformatics (http://www.biomedcentral.com/1471-2105/14/45#).