NCBI - dbSNP

Nociones Básicas de Bioinformática y Genómica

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2015-05-04

dbSNP

- single nucleotide polymorphisms (SNPs)
  -> single nucleotide variants (SNV)
- small-scale variations
- insertions / deletions
- microsatellites
- ...

Classic site:
http://www.ncbi.nlm.nih.gov/SNP/

New Entrez:
http://www.ncbi.nlm.nih.gov/snp/

SNP Attributes

- chromosomal location
- strand (orientation): not to relevant but very confusing
- Allele Origin:
  - germline: population
  - somatic: individual mutation
  - unknown
- Clinical significance: OMIM
- Global minor allele frequency (**MAF**): estimated in the 1000 Genomes population (estimated over chromosomes, not individuals)

SNP Attributes

See details on attributes:
http://www.ncbi.nlm.nih.gov/projects/SNP/docs/rs_attributes.html

See types of variants:
http://www.ncbi.nlm.nih.gov/projects/SNP/snp_legend.cgi?legend=snpClass

Nomenclature

ss#

- Submitted SNP (ss)
- user submitted variants
- generally the *Flanking Sequence* is submitted
- may be reported in any strand

rs#

- RefSNP: reference SNPs
- curated by alignment to the genome, consensus ss#
- clustered form several ss#
- representative sequence: that of the longest ss#
- strand: that of the longest ss#

Nomenclature

IDs are stable but representatives may change or be “merged”

Example
http://www.ncbi.nlm.nih.gov/snp/?term=rs12345

See SNP Validation
http://www.ncbi.nlm.nih.gov/projects/SNP/snp_legend.cgi?legend=validation

SNP Viewer

Forward strand (5’ to 3’): the one on the top

Examples

- Cancer related SNP
- Non disease related snp

Nucleic acid notation and *ambiguity* codes

A C G T are generally used to represent nucleic acids ...

but further "ambiguity" codes are used to represent *uncertain* DNA bases.

See IUPAC notation for nucleic acids at:
http://droog.gs.washington.edu/parc/images/iupac.html
http://www.bioinformatics.org/sms/iupac.html
http://en.wikipedia.org/wiki/Nucleic_acid_notation

Notice that there is some differences in how to interpret "-" or "." symbols.
In dbSNP "-" represents a *gap* ... see dbSNP see Ensembl

FTP Downloads

General:
ftp://ftp.ncbi.nih.gov/snp

Useful:
ftp://ftp.ncbi.nih.gov/snp/organisms

Example: Drosophila Melanogaster
ftp://ftp.ncbi.nih.gov/snp/organisms/fruitfly_7227/

```
wget ftp://ftp.ncbi.nih.gov/snp/organisms/fruitfly_7227/ASN1_flat/ds_flat_ch2R.
wget ftp://ftp.ncbi.nih.gov/snp/organisms/fruitfly_7227/chr_rpts/chr_2R.txt.gz
```

FTP Files

- rs_fasta: reference SNP (rs) sequence data. Usually for BLASTing
- ss_fasta: all available submitted SNP (ss) sequences
- ASN1_flat: human readable text file. Not good for parsing
- **chr_rpts**: tabular file with most relevant information.
- database: SQL dump
- BED files

Explore Files

Explore the tabular files:

```
wc -l chr_2R.txt
```

```
head -n 1000 chr_2R.txt > mytabhead
```

Explore the text files:

```
head ds_flat_ch2R.flat
```

```
head ds_flat_ch2R.flat > myflathead
```