SnpEff walk-through

Pablo Cingolani

1 Analysing fly mutations

1.1 Introduction

A program called SnpEff was created in order to analyze the effect of genetic variations (SNPs, insertions, deletions and MNPs). The program calculates which genomic regions each variant hits, such as: genes, transcripts, exons, UTRs, introns, etc. It also provides detailed information on the changes induced by each variant. For details, see `snpeff.sourceforge.net`.

SnpEff is also being used by the following institutions (this list is incomplete since is only based on personal emails):

- Princeton University,
- Massachusetts General Hospital,
- UC Berkeley,
- Kyoto University (Japan),
- Max Planck (Germany).
- McGill University (Canada),
- University of Cambridge,
- Genome Quebec (Canada),
- University of Wisconsin-Madison,
- UT Southwestern Medical Center,
- University of Southern California,
- Bologna University (Italy),
- Universität Wien (Vienna, Austria),
- Ottawa University (Canada),
- Peter MacCallum Cancer Centre (Australia),

2 Analysis example

In this walk-through we show how to analyze data form fly samples.

2.1 Genome data and annotations

The reference sequence and the annotations used were from FlyBase (dm5.30 was the latest version at the moment). First we download the genome and annotations from FlyBase and move the information to snpEff:

```
wget ftp://ftp.flybase.net/releases/FB2010_07/dmel_r5.30/gff/dmel-all-r5.30.gff.gz

mkdir snpEff/data/dm5.30
mv dmel-all-r5.30.gff.gz snpEff/data/dm5.30/genes.gff.gz
```

We need to update the main configuration file (`snpEff.config`). By inserting `",dm5.30"` (without the quotes) in the `genomes` section of the file we indicate we are adding a new genome.

A new configuration file needs to be created for each genome. In this case, we have to create a file `snpEff/data/dm5.30/snpEff.config`, that only has three lines:

```
# Drosophila melanogaster genes (dm5.30 FlyBase)
dm5.30.genome : Drosophila_melanogaster
    dm5.30.chromosomes : 2L , 2LHet , 2R , 2RHet , 3L , 3LHet , 3R , 3RHet \
                       , 4 , dmel_mitochondrion_genome , Uextra , U , X \
                       , XHet , YHet
```

Now we are ready to create the database by issuing the following command:

```
java -Xmx2G -jar snpEff.jar -b -g dm5.30
```

This command will create a database (`-b` option) from a GFF3 file (`-g` optionss). Running the command will probably show a few warnings about some inconsistent annotations (that is normal for most of the genomes). This command has to be run only once per genome to build the database that snpEff uses to analyze variations.

2.2 Analyzing the data

The sequenced data was mapped to the genome using BWA, then processed using SamTools and BcfTools, thus obtaining a 'vcf' file. A `vcf` file contains variation calls (e.g. SNPs calls) that be processed using snpEff. In this case, each lane was processed independently, so we issued the following command for each lane:

```
java -Xmx2G -jar snpEff.jar -vcf variants_lane1.vcf dm5.30 > snpEff_lane1.txt
```

This command analyses a vcf file (`-vcf` option) for variants from the dm5.30 reference genome.

2.3 Comparing samples

In this example each lane contained a different sample. Since we were interested in finding mutations common to two samples (named X1 and X2), we created a custom perl script (now is part of the snpEff's main distribution). The script joins different files from snpEff analysis and adds a column indicating which samples have each variation:

```
cd snpEff
./scripts/joinSnpEff.pl X1 snpEff_lane1.txt X2 snpEff_lane2.txt > snpEff_join.txt
```

Now we can easily filter the interesting variants using the last column of the joined file.

3 Variants from 1000 Genomes project

One of the goals of snpEff is the ability to analyze large data sets fast. As a test, we analyzed all the variants generated by the **1000 Genomes project**. The whole analysis was done in less than 25 minutes using a generic server by using the following commands:

```
cd snpEff
# Download data
wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20100804/ALL.2of4intersection.20100804.genotypes.vcf.gz

# Analyze variants
java -Xmx8G -jar snpEff.jar -vcf ALL.2of4intersection.20100804.genotypes.vcf.gz hg37.60 > snpEff_lane1.txt
```