VitisOmics

Timothée Flutre

February 19, 2015

Contents

1 Overview

This document contains the documentation for the "VitisOmics" project. This project aims at handling "omics" data in the genus Vitis (e.g. grapevine) in an open and reproducible way.

Such data are available from various places, Genoscope, URGI, NCBI, EBI, etc, and several committes from the IGGP (International Grape Genome Program) strive at improving interoperability. But my attempt, via the usage of git and GitHub, could prove for the community to be a useful addition to these efforts.

The project directory is organized as advised by Noble (PLoS Computational Biology 2009):

On any Unix-like system, it is easily done with the following commands:

```
touch AUTHORS COPYING README; mkdir -p doc data src bin results
```

On any Unix-like system, it can also be easily compressed and transferred (ignoring large data files):

```
cd ..; tar -czvf VitisOmics.tar.gz \
--exclude=VitisOmics/data --exclude=VitisOmics/results \
--exclude="*~" --exclude=".*" VitisOmics
```

1.1 Contributors

As of today: Timothée Flutre, Charles Romieu, Gautier Sarah

2 Data

```
cd data/
```

TODO: retrieve genome data from other cultivars than PN40024, e.g. Sultanina and Tannat

2.1 URGI

- `https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences`

```
../../src/download_urgi.bash
```

When needed, the script decompresses zip files and compress them again but with gzip instead.

2.2 NCBI

- `http://www.ncbi.nlm.nih.gov/genome/401`
- `ftp://ftp.ncbi.nlm.nih.gov/genomes/Vitis_vinifera/`

```
../../src/download_ncbi.bash
```

Note the important file `scaffold_names` which provides the correspondence between original scaffold names (i.e. from the sequencing center) and various NCBI identifiers (RefSeq, GenBank, etc).

2.3 EBI

12X.0 as well as soft-masking by RepeatMasker

```
../../src/download_ebi.bash
```

3 Results

```
cd results/
```

On a computer cluster, indexed files could be copied for usage by everyone, e.g. in `/Genomics/Vitis` if on the CIRAD cluster of the SouthGreen platform.

TODO: use the 5-letter code VITVI to name files, as advised by Grimplet et al, 2014

TODO: keep the info about the source (URGI or NCBI) because differences in terms of N spacers and additional scaffolds (from Aegilops) at the NCBI

3.1 URGI

3.1.1 Reformat sequence headers of file `VV_chr12x.fsa.gz`

```
../../src/reformat_VV_chr12x.bash # take some time
zcat VV_chr12x.fsa.gz | wc -l # 8240706
zcat Vvin-PN40024-12x-chr.fa.gz | wc -l # 8240706
diff <(zcat Vvin-PN40024-12x-chr.fa.gz) <(zcat VV_chr12x.fsa.gz)
```

3.1.2 Reformat sequence headers of file `VV_12X_embl_102_Scaffolds.fsa.gz`

so that they comply with `transpose_annotation`: `https://github.com/SouthGreenPlatform/utils/tree/master/transpose_annotation/`

```
../../src/reformat_VV_12X_embl_102_Scaffolds.bash # take some time
zcat VV_12X_embl_102_Scaffolds.fsa.gz | wc -l # 8091565
zcat Vvin-PN40024-12x-scaff.fa.gz | wc -l # 8091565
diff <(zcat Vvin-PN40024-12x-scaff.fa.gz) <(zcat VV_12X_embl_102_Scaffolds.fsa.
    gz)
```

3.1.3 Format Vvin-PN40024-12x-chr.fa.gz for BLASTn

```
../../src/format_Vvin-PN40024-12x-chr_blastn.bash
```

3.1.4 Index Vvin-PN40024-12x-chr.fa.gz for BWA

```
../../src/index_Vvin-PN40024-12x-chr_bwa.bash
```