

VitisOmics

Timothée Flutre

March 30, 2015

Contents

| | | |
|----------|--|----------|
| 1 | Overview | 1 |
| 1.1 | Contributors | 2 |
| 2 | Data | 2 |
| 2.1 | URGI | 2 |
| 2.2 | NCBI | 2 |
| 2.3 | EBI | 3 |
| 3 | Results | 3 |
| 3.1 | URGI | 3 |
| 3.1.1 | Reformat sequence headers for VITVI_PN40024_12x_v0_chroms_URGI . . | 3 |
| 3.1.2 | Reformat sequence headers for VITVI_PN40024_12x_v0_scaffolds_EMBL_r102 | 4 |
| 3.1.3 | Reformat sequence headers for VITVI_PN40024_12x_v0_contigs_EMBL_r102 | 4 |
| 3.1.4 | Reformat sequence headers for VITVI_PN40024_12x_v2_chroms_URGI . . | 4 |
| 3.1.5 | Format VITVI_PN40024_12x_v0_chroms_URGI for BLASTn | 5 |
| 3.1.6 | Index VITVI_PN40024_12x_v0_chroms_URGI for BWA | 5 |
| 3.1.7 | Index VITVI_PN40024_12x_v2_chroms_URGI for BWA | 5 |

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 interop-

erability. 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.

One needs to 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

```
cd urgi/
```

3.1.1 Reformat sequence headers for VITVI_PN40024_12x_v0_chroms_URGI

Launch script:

```
ln -s ../../data/urg/UV_chr12x.fsa.gz .
echo "../../src/reformat_VV_chr12x.bash" \
| qsub -cwd -j y -V -N reformat_VV_chr12x -q bioinfo.q
```

Check:

```
zcat VV_chr12x.fsa.gz | wc -l # 8240706
zcat VV_chr12x.fsa.gz | grep -c ">" # 33
zcat VITVI_PN40024_12x_v0_chroms_URGI.fa.gz | wc -l # 8240706
zcat VITVI_PN40024_12x_v0_chroms_URGI.fa.gz | grep -c ">" # 33
diff <(zcat VV_chr12x.fsa.gz) <(zcat VITVI_PN40024_12x_v0_chroms_URGI.fa.gz)
```

Only the headers differ, not the sequences, so everything is fine.

Basic stats:

```
zcat VITVI_PN40024_12x_v0_chroms_URGI.fa.gz | md5sum #  
    eff315994faf35333462b9595e10ce5
```

3.1.2 Reformat sequence headers for VITVI_PN40024_12x_v0_scaffolds_EMBL_r102

Launch script:

```
ln -s ../../data/urgi/VV_12X_embl_102_Scaffolds.fsa.gz .  
echo "../../src/reformat_VV_12X_embl_102_Scaffolds.bash" \  
| qsub -cwd -j y -V -N reformat_VV_12X_embl_102_Scaffolds -q bioinfo.q
```

Check:

```
zcat VV_12X_embl_102_Scaffolds.fsa.gz | wc -l # 8091565  
zcat VV_12X_embl_102_Scaffolds.fsa.gz | grep -c ">" # 2059  
zcat VITVI_PN40024_12x_v0_scaffolds_EMBL_r102.fa.gz | wc -l # 8091565  
zcat VITVI_PN40024_12x_v0_scaffolds_EMBL_r102.fa.gz | grep -c ">" # 2059  
diff <(zcat VV_12X_embl_102_Scaffolds.fsa.gz) <(zcat  
    VITVI_PN40024_12x_v0_scaffolds_EMBL_r102.fa.gz)
```

Only the headers differ, not the sequences, so everything is fine.

Basic stats:

```
zcat VITVI_PN40024_12x_v0_scaffolds_EMBL_r102.fa.gz | md5sum # 4  
    fa2432d7a66c019c7cb41ee4d0cb7bc
```

3.1.3 Reformat sequence headers for VITVI_PN40024_12x_v0_contigs_EMBL_r102

TODO

3.1.4 Reformat sequence headers for VITVI_PN40024_12x_v2_chroms_URGI

Launch script:

```
ln -s ../../data/urgi/12Xv2_grapevine_genome_assembly.fa.gz .  
echo "../../src/reformat_12Xv2_grapevine_genome_assembly.bash" \  
| qsub -cwd -j y -V -N reformat_12Xv2_grapevine_genome_assembly -q bioinfo.q
```

Check:

```
zcat 12Xv2_grapevine_genome_assembly.fa.gz | wc -l # 8103449
zcat 12Xv2_grapevine_genome_assembly.fa.gz | grep -c ">" # 20
zcat VITVI_PN40024_12x_v2_chroms_URGI.fa.gz | wc -l # 8103449
zcat VITVI_PN40024_12x_v2_chroms_URGI.fa.gz | grep -c ">" # 20
diff <(zcat 12Xv2_grapevine_genome_assembly.fa.gz) <(zcat
    VITVI_PN40024_12x_v2_chroms_URGI.fa.gz)
```

Only the headers differ, not the sequences, so everything is fine.

Basic stats:

```
zcat VITVI_PN40024_12x_v2_chroms_URGI.fa.gz | md5sum # 4
    e487c28eaf19ef59b0b6128b73935af
```

3.1.5 Format VITVI_PN40024_12x_v0_chroms_URGI for BLASTn

TODO: change Vvin to VITVI

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

3.1.6 Index VITVI_PN40024_12x_v0_chroms_URGI for BWA

Launch:

```
echo "../../src/bwa_index_VITVI_PN40024_12x_v0_chroms_URGI.bash" \
| qsub -cwd -j y -V -N bwa_index_VITVI_PN40024_12x_v0_chroms_URGI -q bioinfo.
q
```

3.1.7 Index VITVI_PN40024_12x_v2_chroms_URGI for BWA

Launch:

```
echo "../../src/bwa_index_VITVI_PN40024_12x_v2_chroms_URGI.bash" \
| qsub -cwd -j y -V -N bwa_index_VITVI_PN40024_12x_v2_chroms_URGI -q bioinfo.
q
```