# Vvin-PN40024

Timothée Flutre

February 9, 2015

## Contents



## 1 Overview

This document contains the documentation for the "Vvin-PN40024" project. This project aims at handling the reference genome sequences of cultivar Pinot Noir 40024 of *Vitis vinifera* (grapevine) in a reproducible way.

The original genome sequences, from Jaillon *et al.* were sequenced at 8x. The 12x data are not yet officially published, even though they are available at various places: Genoscope, URGI, NCBI, EBI, etc.

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 Vvin-PN40024.tar.gz \
--exclude=Vvin-PN40024/data --exclude=Vvin-PN40024/results \
--exclude="*~" --exclude=".*" Vvin-PN40024
```

# 2 Data

```
cd data/
```

## 2.1 URGI

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

```
wget https://urgi.versailles.inra.fr/download/vitis/VV_12X_embl_102_WGS_contigs
   .fsa.zip
wget https://urgi.versailles.inra.fr/download/vitis/VV_12X_embl_102_Scaffolds.
   fsa.zip
wget https://urgi.versailles.inra.fr/download/vitis/VV_chr12x.fsa.zip
wget -O 12x0_chr.agp https://urgi.versailles.inra.fr/content/download
   /1028/8244/file/chr.agp
wget -O 12x0_chrUn.agp https://urgi.versailles.inra.fr/content/download
   /1029/8248/file/chrUn.agp
wget -O 12x0_chr.agp.info https://urgi.versailles.inra.fr/content/download
   /2149/19329/file/chr.agp.info
wget -O 12x0_chr.lg https://urgi.versailles.inra.fr/content/download
   /2150/19333/file/chr.lg
wget -O 12x0_scaffolds.lg https://urgi.versailles.inra.fr/content/download
   /1093/8684/file/scaffolds.lg
wget https://urgi.versailles.inra.fr/download/vitis/12
   Xv2_grapevine_genome_assembly.fa.gz
wget https://urgi.versailles.inra.fr/content/download/3044/26115/file/
   golden_path_V2_111113_allChr.csv
wget https://urgi.versailles.inra.fr/content/download/3043/26111/file/
   chr_size_V2.txt
```

When needed, re-compress with gzip instead of zip.

Download annotation data for 12X.0 (Genoscope) and 12X.2 (CRIBI):

```
wget https://urgi.versailles.inra.fr/content/download/2160/19388/file/chrAll.
   jigsawgaze_NR.gff.gz
wget https://urgi.versailles.inra.fr/content/download/2157/19376/file/
   Vitis_vinifera_annotation.gff.gz
```

### 2.2 NCBI

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

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

### 2.3 EBI

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

```
wget ftp://ftp.ensemblgenomes.org/pub/plants/release-21/fasta/vitis_vinifera/
   dna/Vitis_vinifera.IGGP_12x.21.dna.genome.fa.gz
wget ftp://ftp.ensemblgenomes.org/pub/plants/release-21/fasta/vitis_vinifera/
   dna/Vitis_vinifera.IGGP_12x.21.dna_sm.genome.fa.gz
```

## 3 Results

```
cd results/
```

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

### 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
```