VitisOmics

(see contributors below)

October 13, 2015

Contents

1 Overview

This document describes the "VitisOmics" project. This project aims at handling "omics" data in the genus Vitis (e.g. grapevine) in an open and reproducible way. Nevertheless, it requires some basic knowledge and skills about bioinformatics on GNU/Linux computers.

A large amount of such "omics" data are already available, and several committes from the IGGP (International Grape Genome Program) strive at improving interoperability. However, several issues remain, among which:

- partially-overlapping data present at multiple locations (URGI, CRIBI, NCBI, EBI, etc);
- data downloadable as files in various formats not always easy to interchange (fasta, genbank, gff3, gtf, etc);
- data often available without meta-data inside the file;
- large files not always available in compressed form;
- main efforts dedicated to wet-labs grapevine biologists.

These have consequences in terms of ambiguity, inefficiency, potential mistakes, etc. Therefore, I hope that my attempt, via the usage of git and GitHub, could prove for the community to be a useful addition to the IGGP efforts.

The repository can be easily cloned from GitHub:

```
git clone https://github.com/timflutre/VitisOmics.git
```

The project directory is organized as advised by Noble (PLoS Computational Biology 2009). On any Unix-like system, it can be easily compressed and transferred (ignoring large data files):

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

In order to concretely promote collaborative editing in a distributed manner, the content of the "VitisOmics" repository should be based on plain text files. As a consequence, this document is written

in the org format, and can thus be automatically exported, best by emacs, but also by pandoc, into the pdf and html formats for easy reading. A choice is also made to use as much as possible softwares widely available on any GNU/Linux computer, such as bash, awk, etc, but of course it may sometimes be much easier to use R (with Bioconductor), Python (with Biopython), etc.

Last but not least, feel free to contribute, by reporting issues or forking the repository!

1.1 Contributors

The person roles comply with R's guidelines (The R Journal Vol. 4/1, June 2012).

- Timothée Flutre (cre,aut)
- Gautier Sarah (ctb)
- ...

1.2 References

As an introduction, the NCBI has open web pages about genome assembly:

- Assembly information;
- NCBI Genome Assembly Model;
- AGP specification.

Then, the original article for grapevine is a must read:

- Jaillon, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463-467 (2007).

Also of interest about genome annotation:

- Vitulo et al. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biology 14, 99+ (2014).

2 Data

```
mkdir -p data; cd data/
```

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

2.1 URGI

- `https://urgi.versailles.inra.fr/Species/Vitis`

```
mkdir -p urgi; cd urgi/
../../src/download_urgi.bash
```

Remarks:

- when needed, the script decompresses zip files and compress them again but with gzip instead;
- the AGP file for the PN40024 8x assembly is not available at URGI's web site.

2.2 NCBI

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

```
mkdir -p ncbi; cd ncbi/
../../src/download_ncbi.bash
```

Remarks:

- the important file `scaffold_names` provides the correspondence between original scaffold names (i.e. from the sequencing center) and various NCBI identifiers (RefSeq, GenBank, etc);
- in `ARCHIVE/`, `BUILD.1.1/` corresponds to the 8x genome sequences of PN40024.

2.3 EBI

```
mkdir -p ebi; cd ebi/
../../src/download_ebi.bash
```

Remarks:

- a genome soft-masked by RepeatMasker is available.

2.4 CRIBI

- http://genomes.cribi.unipd.it/grape/

```
mkdir -p cribi; cd cribi/
../../src/download_cribi.bash
```

3 Results

```
mkdir -p results; cd results/
```

TODO: compress fasta files with `bgzip` instead of `gzip`

3.1 Comparisons of original "assembly" files

Files from URGI:

```
cd urgi/
zcat VV_8X_embl_98_WGS_contigs.fsa.gz | grep -c ">" # 19577
zcat VV_8X_embl_98_Scaffolds.fsa.gz | grep -c ">" # 3514
zcat VV_chr8x.fsa.gz | grep -c ">" # 35
zcat VV_12X_embl_102_WGS_contigs.fsa.gz | grep -c ">" # 14642
zcat VV_12X_embl_102_Scaffolds.fsa.gz | grep -c ">" # 2059
zcat VV_chr12x.fsa.gz | grep -c ">" # 33
zcat 12Xv2_grapevine_genome_assembly.fa.gz | grep -c ">" # 20
```

Files from NCBI (read first the README's):

```
cd ncbi/
ls ARCHIVE/BUILD.1.1/CHRS/vvi_ref_chr*.fa.gz | grep -v "Pltd" | while read f; do
    zcat $f; done | grep -c ">" # 3514
ls ARCHIVE/BUILD.1.1/Assembled_chromosomes/vvi_ref_chr*.fa.gz | while read f; do
    zcat $f; done | grep -c ">" # 19
zcat ARCHIVE/BUILD.1.1/allcontig.agp.gz | grep -v "#" | cut -f 5 | sort | uniq -c #
    F=1 N=16063 W=19577
cat data/ncbi/ARCHIVE/BUILD.1.1/scaffold_names | sed 1d | wc -l # 3514
```

See also the script `src/vitisomics.R` using R and Bioconductor.

For its build 1.1 (corresponding to the 8x sequences of the PN40024 variety), the NCBI has one file per assembled chromosome. However, all unlocalized and unplaced scaffolds are gathered in a single

file `chrUn`. This is not the case at URGI which has unlocalized scaffolds in files as `chr3_random` and a `chrUn_random` file with all unplaced scaffolds (and only them). Unfortunately, the NCBI has the annotation of the 8x (in the GenBank format), but the URGI hasn't.

3.2 Manipulations of files from URGI

```
mkdir -p urgi; cd urgi/
```

3.2.1 Reformat sequence headers for `VITVI_PN40024_8x_chroms_URGI`

Launch script:

```
ln -s ../../data/urgi/VV_chr8x.fsa.gz .
echo "../../src/reformat_VV_chr8x.bash" \
  | qsub -cwd -j y -V -N reformat_VV_chr8x -q normal.q
```

Check:

```
zcat VV_chr8x.fsa.gz | wc -l # 8291865
zcat VV_chr8x.fsa.gz | grep -c ">" # 35
zcat VITVI_PN40024_8x_chroms_URGI.fa.gz | wc -l # 8291865
zcat VITVI_PN40024_8x_chroms_URGI.fa.gz | grep -c ">" # 35
diff <(zcat VV_chr8x.fsa.gz) <(zcat VITVI_PN40024_8x_chroms_URGI.fa.gz)
```

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

Basic stats:

```
zcat VITVI_PN40024_8x_chroms_URGI.fa.gz | md5sum # 4b6ea1cb4ff189ac587fa269077885b5
```

Length of each sequence:

```
zcat VITVI_PN40024_8x_chroms_URGI.fa.gz \
  | awk 'BEGIN{RS=">"} {split($0,a,"\n");
if(length(a)==0)next;
sum=0; for(i=2;i<=length(a);++i){sum+=length(a[i])};
print a[1]": "sum; sumTot+=sum} END{print sumTot}'
```

header	length (bp)
chr1 CU462738|Vitis vinifera|PN40024|assembly8x|chromosome$_1$	15630816
chr10 CU462747|Vitis vinifera|PN40024|assembly8x|chromosome$_{10}$	9647040
chr10$_{\text{random}}$	2206354
chr11 CU462748|Vitis vinifera|PN40024|assembly8x|chromosome$_{11}$	13936303
chr11$_{\text{random}}$	1958407
chr12 CU462749|Vitis vinifera|PN40024|assembly8x|chromosome$_{12}$	18540817
chr12$_{\text{random}}$	2826407
chr13 CU462750|Vitis vinifera|PN40024|assembly8x|chromosome$_{13}$	15191948
chr13$_{\text{random}}$	1580403
chr14 CU462751|Vitis vinifera|PN40024|assembly8x|chromosome$_{14}$	19480434
chr14$_{\text{random}}$	5432426
chr15 CU462752|Vitis vinifera|PN40024|assembly8x|chromosome$_{15}$	7693613
chr15$_{\text{random}}$	4297576
chr16 CU462753|Vitis vinifera|PN40024|assembly8x|chromosome$_{16}$	8158851
chr16$_{\text{random}}$	4524411
chr17 CU462754|Vitis vinifera|PN40024|assembly8x|chromosome$_{17}$	13059092
chr17$_{\text{random}}$	1763011
chr18 CU462755|Vitis vinifera|PN40024|assembly8x|chromosome$_{18}$	19691255
chr18$_{\text{random}}$	5949186
chr19 CU462756|Vitis vinifera|PN40024|assembly8x|chromosome$_{19}$	14071813
chr19$_{\text{random}}$	1912523
chr1$_{\text{random}}$	5496190
chr2 CU462739|Vitis vinifera|PN40024|assembly8x|chromosome$_2$	17603400
chr2$_{\text{random}}$	60809
chr3 CU462740|Vitis vinifera|PN40024|assembly8x|chromosome$_3$	10186927
chr3$_{\text{random}}$	1343266
chr4 CU462741|Vitis vinifera|PN40024|assembly8x|chromosome$_4$	19293076
chr5 CU462742|Vitis vinifera|PN40024|assembly8x|chromosome$_5$	23428299
chr6 CU462743|Vitis vinifera|PN40024|assembly8x|chromosome$_6$	24148918
chr7 CU462744|Vitis vinifera|PN40024|assembly8x|chromosome$_7$	15233747
chr7$_{\text{random}}$	176143
chr8 CU462745|Vitis vinifera|PN40024|assembly8x|chromosome$_8$	21557227
chr8$_{\text{random}}$	12125
chr9 CU462746|Vitis vinifera|PN40024|assembly8x|chromosome$_9$	16532244
chrUn$_{\text{random}}$	154883714
total	497508771

3.2.2 Reformat sequence headers for `VITVI_PN40024_12x_v0_contigs_EMBL_r102`

TODO

3.2.3 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 normal.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
```

3.2.4 Reformat sequence headers for `VITVI_PN40024_12x_v0_chroms_URGI`

Launch script:

```
ln -s ../../data/urgi/VV_chr12x.fsa.gz .
echo "../../src/reformat_VV_chr12x.bash" \
  | qsub -cwd -j y -V -N reformat_VV_chr12x -q normal.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 #
   eff315994fafe35333462b9595e10ce5
```

3.2.5 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 normal.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
```

Length of each sequence:

```
zcat VITVI_PN40024_12x_v2_chroms_URGI.fa.gz \
  | awk 'BEGIN{RS=">"} {split($0,a,"\n");
if(length(a)==0)next;
sum=0; for(i=2;i<=length(a);++i){sum+=length(a[i])};
print a[1]": "sum; sumTot+=sum} END{print sumTot}'
```

header	length (bp)
chr1 Vitis vinifera|PN40024|assembly12x.2	24233538
chr2 Vitis vinifera|PN40024|assembly12x.2	18891843
chr3 Vitis vinifera|PN40024|assembly12x.2	20695524
chr4 Vitis vinifera|PN40024|assembly12x.2	24711646
chr5 Vitis vinifera|PN40024|assembly12x.2	25650743
chr6 Vitis vinifera|PN40024|assembly12x.2	22645733
chr7 Vitis vinifera|PN40024|assembly12x.2	27355740
chr8 Vitis vinifera|PN40024|assembly12x.2	22550362
chr9 Vitis vinifera|PN40024|assembly12x.2	23006712
chr10 Vitis vinifera|PN40024|assembly12x.2	23503040
chr11 Vitis vinifera|PN40024|assembly12x.2	20118820
chr12 Vitis vinifera|PN40024|assembly12x.2	24269032
chr13 Vitis vinifera|PN40024|assembly12x.2	29075116
chr14 Vitis vinifera|PN40024|assembly12x.2	30274277
chr15 Vitis vinifera|PN40024|assembly12x.2	20304914
chr16 Vitis vinifera|PN40024|assembly12x.2	23572818
chr17 Vitis vinifera|PN40024|assembly12x.2	18691847
chr18 Vitis vinifera|PN40024|assembly12x.2	34568450
chr19 Vitis vinifera|PN40024|assembly12x.2	24695667
chrUkn Vitis vinifera|PN40024|assembly12x.2	27389308
total	486205130

3.2.6 Format `VITVI_PN40024_12x_v0_chroms_URGI` for BLASTn

TODO: change Vvin to VITVI

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

3.2.7 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 normal.q
```

3.2.8 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 normal.q
```

3.2.9 Prepare VITVI_PN40024_12x_v2_chroms_URGI for SAMtools and Picard

Make an index as well as a SAM header.

Launch:

```
echo "../../src/samtools-picard_prep_VITVI_PN40024_12x_v2_chroms_URGI.bash" \
  | qsub -cwd -j y -V -N samtools-picard_prep_VITVI_PN40024_12x_v2_chroms_URGI -q
     normal.q
```

3.2.10 Index VITVI_PN40024_12x_v0_chroms_URGI for Bowtie2

Launch:

```
echo "../../src/bowtie2_index_VITVI_PN40024_12x_v0_chroms_URGI.bash" \
  | qsub -cwd -j y -V -N bowtie2_build_VITVI_PN40024_12x_v0_chroms_URGI -q normal.q
```

3.2.11 Index VITVI_PN40024_12x_v2_chroms_URGI for Bowtie2

Launch:

```
echo "../../src/bowtie2_index_VITVI_PN40024_12x_v2_chroms_URGI.bash" \
  | qsub -cwd -j y -V -N bowtie2_build_VITVI_PN40024_12x_v2_chroms_URGI -q normal.q
```

3.2.12 Index VITVI_PN40024_12x_v2_chroms_URGI for Bowtie2 compatible with Tassel

Tassel requires numbers as chromosome identifiers.

Launch:

```
echo "../../src/bowtie2_index_VITVI_PN40024_12x_v2_chroms_URGI_for_Tassel.bash" \
  | qsub -cwd -j y -V -N bowtie2_build_VITVI_PN40024_12x_v2_chroms_URGI_for_Tassel -
     q normal.q
```

3.2.13 Translate CRIBI annotations from 12X.0 to 12X.2

Requirement: use or write a script taking as input the 12X.0 GFF3 file as well as the 12.0-12.2 AGP file, and returns as output the 12X.2 GFF3 file

The URGI provides the following AGP file: `golden_path_V2_111113_allChr.csv`. Unfortunately, after looking at the official specification of the AGP format, the URGI file doesn't seem to be valid, neither for version 1.1, nor 2.2. After contacting URGI, they told me they were working on it (October 2015).

Another script was developped by G. Sarah, but it suffers from several issues.

TODO: test CrossMap

3.3 Manipulations of files from NCBI

```
mkdir -p ncbi; cd ncbi/
```

3.3.1 Reformat sequence headers for `VITVI_PN40024_8x_chroms_NCBI`

Launch script:

```
ln -s ../../data/ncbi/ARCHIVE/BUILD.1.1/Assembled_chromosomes/vvi_ref_chr*.fa.gz .
ln -s ../../data/ncbi/ARCHIVE/BUILD.1.1/CHRS/vvi_ref_chrUn.fa.gz
echo "../../src/reformat_chr_NCBI-build-1-1.bash ${i}" \
  | qsub -cwd -j y -V -N reformat_chroms_NCBI-build-1-1 -q normal.q
```

Check:

```
\ls vvi_ref_chr* | while read f; do zcat $f; done | wc -l # 6499942
\ls vvi_ref_chr* | while read f; do zcat $f; done | grep -c ">" # 3343
zcat VITVI_PN40024_8x_chroms_NCBI.fa.gz | wc -l # 6499942
zcat VITVI_PN40024_8x_chroms_NCBI.fa.gz | grep -c ">" # 3343
diff <(\ls -v vvi_ref_chr* | while read f; do zcat $f; done) <(zcat
   VITVI_PN40024_8x_chroms_NCBI.fa.gz)
```

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

Basic stats:

```
zcat VITVI_PN40024_8x_chroms_NCBI.fa.gz | md5sum # d8eeff80c824f1b5bd91b4274fddb696
```

3.4 Creation of R/Bioconductor packages

- http://www.bioconductor.org/
- Huber, W. et al. Orchestrating high-throughput genomic analysis with bioconductor. Nature Methods 12, 115-121 (2015). URL http://dx.doi.org/10.1038/nmeth.3252.

TODO: see AnnotationHub

3.4.1 BSgenome IGGP12Xv2 package

http://bioconductor.org/packages/release/bioc/html/BSgenome.html

Retrieve the sequence data from URGI:

```
cd results/
mkdir -p make_BSgenome_IGGP12Xv2
cd make_BSgenome_IGGP12Xv2/
ln -s ../../data/urgi/12Xv2_grapevine_genome_assembly.fa.gz .
```

Split into one chromosome per file (in the headers, discard everything after the first space):

```
zcat 12Xv2_grapevine_genome_assembly.fa.gz | awk 'BEGIN{RS=">"} {if(NF==0)next;
    split($0,a,"\n"); split(a[1],b," "); print b[1]; print ">"b[1] > b[1]".fa"; for(
    i=2;i<length(a);++i){print a[i] >> b[1]".fa"}}'
gzip chr*.fa
```

Prepare the seed file (IGGP12Xv2_seed.txt) by hand as indicated in the vignette as well as in the official R manual "Writing R extensions". Following this article, I chose the CC0 license (present in the R list of licenses in share/licenses/license.db).

Forge the target package from the seed file:

```
echo "date; echo \"library(BSgenome); forgeBSgenomeDataPkg(\\\"IGGP12Xv2_seed.txt\\\
    ")\" | R --vanilla; date" | qsub -cwd -j y -V -N forge_BSgenome -q normal.q
```

Build the package and check it:

```
echo "date; R CMD build BSgenome.Vvinifera.URGI.IGGP12Xv2; date" | qsub -cwd -j y -V
     -N build_BSgenome -q normal.q
echo "date; R CMD check BSgenome.Vvinifera.URGI.IGGP12Xv2_0.1.tar.gz; date" | qsub -
    cwd -j y -V -N check_BSgenome -q normal.q
```

The target package is now ready to be installed:

```
R CMD INSTALL BSgenome.Vvinifera.URGI.IGGP12Xv2_0.1.tar.gz
```

A.-F. Adam-Blondon (part of IGGP) and other colleagues from INRA gave positive feedback. I hence sent the package to the Bioconductor team. I hope it will soon be available for download via `biocLite()`.

3.4.2 BSgenome IGGP12Xv0 package

`http://bioconductor.org/packages/release/bioc/html/BSgenome.html`

Retrieve the sequence data from URGI:

```
cd results/
mkdir -p make_BSgenome_IGGP12Xv0
cd make_BSgenome_IGGP12Xv0/
ln -s ../../data/urgi/VV_chr12x.fsa.gz .
```

TODO: Prepare the seed file (`IGGP12Xv0_seed.txt`) by editing the one used for `IGGP12Xv2`.

3.4.3 TxDb IGGP12Xv0 package from NCBI annotations

`http://www.bioconductor.org/packages/release/bioc/html/GenomicFeatures.html`

TODO: use `makeTxDbFromGFF()`