ISSN (print): 2421-5783
ISSN (on line): 2421-5562

Consiglio Nazionale delle Ricerche

IRCrES

ISTITUTO DI RICERCA SULLA CRESCITA ECONOMICA SOSTENIBILE
RESEARCH INSTITUTE ON SUSTAINABLE ECONOMIC GROWTH

Rapporto Tecnico

Numero 1, Luglio 2015

GBrowse installation and customization to display the Gigaspora margarita BEG34 mitochondrial genome data

Stefano Ghignone, Francesco Venice,Giancarlo Birello, Paola Bonfante

Consiglio Nazionale delle Ricerche

IRCRES

ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
RESEARCH INSTITUTE on SUSTAINABLE ECONOMIC GROWTH

RAPPORTO TECNICO CNR-IRCRES
Anno 1, Numero 1, Luglio 2015

Direttore Responsabile
Secondo Rolfo

Direzione e Redazione
CNR-IRCRES
Istituto di Ricerca sulla crescita economica sostenibile
Via Real Collegio 30, 10024 Moncalieri (Torino), Italy
Tel. +39 011 6824.911
Fax +39 011 6824.966
segreteria@ircres.cnr.it
www.ircres.cnr.it

Sede di Roma
Via dei Taurini 19, 00185 Roma, Italy
Tel: 06 49937809
Fax: 06 49937808

Sede di Milano
Via Bassini 15, 20121 Milano, Italy
Tel: 02 23699501
Fax: 02 23699530

Sede di Genova
Università di Ge Via Balbi, 6 - 16126 Genova
Tel: 010-2465.459
Fax: 010-2099.826

Segreteria di redazione
Enrico Viarisio
enrico.viarisio@ircres.cnr.it

GBrowse installation and customization to display the *Gigaspora margarita* BEG34 mitochondrial genome data

Stefano Ghignone*
(IPSP-CNR)

Francesco Venice
(Università degli Studi di Torino)

Giancarlo Birello
(IRCrES-CNR)

Paola Bonfante
(Universita`degli Studi di Torino)

Summary

The Generic Genome Browser (GBrowse) is a simple but highly configurable web-based genome browser. GBrowse consists in a combination of database, interactive web pages and a rich set of utilities for manipulating and displaying annotations on genomes. The Gbrowse platform has been largely used as visualization tool for most of the model organisms. This tutorial shows how to configure a GBrowse genome browser installation to display the mitochondrial genome annotation of the Arbuscular Mycorrhizal Fungus *Gigaspora margarita* BEG34.

Key Words: open-source, NGS data, bioinformatics, GBrowse, Arbuscular Mycorrhizal Fungi

*Corresponding author: stefano.ghignone@ipsp.cnr.it

Contents

1 Introduction **6**

2 Base system **7**

3 Gbrowse installation **8**
3.1 Install and Configure a Basic Ubuntu Server 8
3.2 Installing BioPerl 9
3.2.1 Install as much as possible with apt-get 9
3.2.2 Install CPAN and Perl modules not available via apt-get 9
3.2.3 Install BioPerl 9
3.3 Bio::Graphics::Browser2 10
3.3.1 Prerequisites 10
3.3.2 Install GBrowse via the CPAN Shell 11
3.4 Samtools 13

4 Data processing **14**
4.1 Prerequisites 14
4.2 Run bowtie2 with reads 14
4.3 Run BWA with assembled transcripts 15
4.4 Convert GenBank file to Gff3 16

5 GBrowser configuration **17**
5.1 SQlite database creation 17
5.2 Configuration editing 17
5.2.1 Accessory files 17
5.2.2 Global configuration file 18
5.2.3 Data Source configuration file 19
5.3 Configuration test 21

6 Appendix **22**
6.1 *Gigaspora margarita* BEG34 transcriptome 22
6.2 Examples 23
6.2.1 Network interfaces configuration file 23

6.2.2 Batch bowtie2 *bash* script 24
6.2.3 GBrowse configuration file 26

Bibliography **32**

Chapter 1

Introduction

The Generic Genome Browser (**GBrowse**[1]) is a simple but highly configurable web-based genome browser. It is a component of the Generic Model Organism Systems Database project (**GMOD**[2]). GBrowse consists in a combination of database, interactive web pages and a rich set of utilities for manipulating and displaying annotations on genomes. The Gbrowse platform has been largely used as visualization tool for most of the model organisms, such as *Caenorhabditis elegans* (WormBase[3]), *Drosophyla melanogaster* (FlyBase[4]), Rice (Rice Genome Annotation Project[5]) and *Tuber melanosporum* (Tuber Genome Browser[6]).

In the context of Arbuscular Mycorrhizal Fungi (AMF) research, the only known application of such a tool was in BIOBITs Project[7], where this platform was used in data mining of the genome of the endobacterium *Candidatus* Glomeribacter Gigasporarum[8], a rod-shaped Gram negative beta-proteobacterium inhabiting isolates of the AMF *Gigaspora margarita* WN Becker & IR Hall, but it has never been used on fungal genome side. Within the framework of the Mycorrhizal Genomics Initiative, the JGI provides a Viewer for the analysis of the *Rhizophagus irregularis* DAOM 181602 genome[9], considered the model species for the Glomeromycotan lineage, but it's more likely based on **Vista** Tools[10] rather than on GBrowse.

The species *G. margarita* represents an additional model in AMF research, for which very little nuclear and organellar sequence information is currently available. To date, the mitochondrial genome is the larger nucleotide sequence available[11]. While waiting for other sequencing projects provide complete genome sequences of the fungus, we wanted to enhance the Gbrowse visualization of the annotation of the mitochondrial genome with data from the ongoing RNA-seq sequencing project.
This tutorial walks through how to configure a classical GBrowse genome browser installation to display Next Generation Sequencing (NGS) data using the SAMtools GBrowse adaptor, Bio::DB::Sam.

Chapter 2

Base system

The main server acting as hypervisor (*"Towanda"*) is equipped with 64 cores and 256 GB RAM due to application requirements. Virtualization framework is KVM (Kernel- based Virtual Machine) installed over an Ubuntu Linux operative system. Full hypervisor installation guide can be found in our technical report KVM: an open-source framework for virtualization, RT44 Ceris-CNR[12].

Virtual machine disks are LVM based for a flexible storage management (see details in RT 44). Towanda is equipped with 1 TB RAID-1 logical drive and 3.7 TB RAID-5 logical drive, both on local server storage. In addition a 1 TB iSCSI partition is available to virtual machines for backup and temporarily storage, the partition is located on two-nodes HA cluster (see technical reports RT37[13] and RT41 Ceris-CNR[14]).

Applications are hosted on a virtual machine (*"gbrowser"*) based on Ubuntu server 12.04 LTS. We made 2GB RAM, 4 cores and 8GB HDD space available to *gbrowser* virtual machine, while remaining resources are reserved for hypervisor base functions and other hosted virtual machines.

Operative system was installed from ISO image of standard distribution with default values.

Next paragraph starts from a fresh installation with all system packages updated to last version available.

Chapter 3

Gbrowse installation

3.1 Install and Configure a Basic Ubuntu Server

Do install, then configure.

```
$ sudo apt-get clean
$ sudo rm -rf /var/lib/apt/lists/*
$ sudo apt-get update
$ sudo apt-get upgrade
$ sudo apt-get install ssh screen htop apt-file
$ sudo apt-get clean
$ sudo reboot
```

Extend the Hardware Enablement Stack (HWE) support.

```
$ sudo apt-get install linux-generic-lts-trusty linux
   -image-generic-lts-trusty
```

Set up the Network interface, editing the *interfaces* file to obtain a Static IP Configuration.

```
$ sudo nano -w /etc/network/interfaces
```

A working configuration file, including IPv6 Static Addressing, is included in Appendix.

Optionally, provide the system with a minimal graphic interface. In this case, we use the Gnome Classic desktop environment.

```
$ sudo apt-get install lightdm gnome-terminal
   synaptic
$ sudo apt-get install gnome-core gnome-session-
   fallback
$ sudo reboot
```

3.2 Installing BioPerl

3.2.1 Install as much as possible with apt-get

Enable Universe and Multiverse in /etc/apt/sources.list

```
$ sudo apt-get install libexpat-dev libexpat1-dev
    zlibc zlib1g-dev libncurses5-dev lynx unzip zip
    ncftp gcc libc6-dev make build-essential mysql-
    server apache2 perl libgd-gd2-perl libcgi-session-
    perl libclass-base-perl sqlite gedit
```

3.2.2 Install CPAN and Perl modules not available via apt-get

```
$ sudo cpan -i 'Text::Shellwords'
```

Configure CPAN with defaults if first time.)

3.2.3 Install BioPerl

- Upgrade CPAN:

  ```
  $ sudo perl -MCPAN -e shell
  cpan>install Bundle::CPAN
  cpan>q
  ```

- Install/upgrade Module::Build, and make it your preferred installer:

  ```
  $ sudo perl -MCPAN -e shell
  cpan>install Module::Build
  cpan>o conf prefer_installer MB
  cpan>o conf commit
  cpan>q
  ```

  This will enable recording commands in cpan history

- Installing using CPAN
  Find the name of the most recent BioPerl version:

  ```
  $ sudo perl -MCPAN -e shell
  cpan>d /bioperl/
  ```

  As in December 2014, the most recent version is 1.6.924. Now install:

  ```
  cpan>install CJFIELDS/BioPerl-1.6.924.tar.gz
  ```

As there are over 800 modules in BioPerl and the test suite is running more than 12000 individual tests, a few failed tests may not affect your usage of BioPerl. Usually, failed tests are numerous and full BioPerl install is aborted. If you decide that the failed tests will not affect how you intend to use BioPerl and you'd like to install anyway do:

```
cpan>force install CJFIELDS/BioPerl-1.6.924.tar.
    gz
```

- Complete install with Bundle::BioPerl module:

```
cpan>install CJFIELDS/Bundle-BioPerl-2.1.9.tar.
    gz
```

3.3 Bio::Graphics::Browser2

3.3.1 Prerequisites

Detailed description of GBrowse prerequisites are available here: `http://gmod.org/wiki/GBrowse_2.0_Prerequisites`.

GBrowse depends on the following standard Perl libraries:

- Module::Build
- GD*
- Bio::Perl (version 1.6.0 or higher)
- Bio::Graphics
- JSON
- LWP
- Storable
- IO::String
- Capture::Tiny
- File::Temp
- Digest::MD5
- CGI::Session

- Statistics::Descriptive

* Bio::Graphics is strictly dependent on GD, a Perl module for generating bitmapped graphics. GD in turn is dependent on libgd, a C library. To use Bio::Graphics, both these software libraries must be installed. If you are on a Linux system, you might already have GD installed. To verify, run the following command:

```
$ sudo perl -MGD -e 'print $GD::VERSION, "\n"'
```

On Ubuntu Server 12.04 LTS, you must get '2.46'.
Most of the modules have been installed in previous steps. Use CPAN to install the missing ones:

```
$ sudo perl -MCPAN -e shell
>cpan install DBD::SQLite Bio::Graphics Storable IO
    ::String Digest::MD5 CGI::Session Statistics::
    Descriptive
```

3.3.2 Install GBrowse via the CPAN Shell

Install the latest released version of GBrowse by running the CPAN shell.

```
>cpan install Bio::Graphics::Browser2
```

The configuration process will ask you to confirm site-specific configuration options. Confirming the proposed options, the directories (and all their needed parents) will be created at install time.

The major configuration options are:

cgibin: The directory in which Apache's executable CGI scripts are located, for example /usr/lib/cgi-bin. This directory is set up for you when Apache is installed, and you must have the path correct in order for Build to install GBrowse's CGI scripts in the right place. GBrowse will be installed into the "gb2" subdirectory, so its path will be "*/usr/lib/cgi-bin/gb2/gbrowse*".

conf: The location of GBrowse's configuration files. The default is to place them in */etc/gbrowse2*. This is where you will go to customize GBrowse and add new data sources.

databases: The default location for GBrowse's in memory databases, and the place where the example databases will be stored (*/var/lib/gbrowse2/databases*).

htdocs: The directory in which to install GBrowse's Javascript libraries, static HTML pages and stylesheets. You can choose any location for this directory and it will be added to Apache's document tree. The default is to place the directory under the default document tree, such as */var/www/gbrowse2*.

tmp: The directory in which GBrowse will store its working data, including users' session information (such as preferred tracks), uploaded data files, and temporary image files. The default is to place these files into */var/tmp/gbrowse2*.

wwwuser: The account under which the system Apache runs, often "nobody", "www-data" (default) or "httpd."

The interactive configuration process should look like this.

```
**Beginning interactive configuration**
Directory for GBrowse's config and support files? [/
   etc/gbrowse2]
Directory for GBrowse's static images & HTML files?
   [/var/www/gbrowse2]
Directory for GBrowse's temporary data [/var/tmp/
   gbrowse2]
Directory for GBrowse's sessions, uploaded tracks and
    other persistent data [/var/lib/gbrowse2]
Directory for GBrowse's example databases [/var/lib/
   gbrowse2/databases]
Directory for GBrowse's CGI script executables? [/usr
   /lib/cgi-bin/gb2]
Internet port to run demo web site on (for demo)?
   [8000]
Apache loadable module directory (for demo)? [/usr/
   lib/apache2/modules]
User account under which Apache daemon runs? [www-
   data]
Automatically update Apache config files to run
   GBrowse? [y]
Automatically update system config files to run
   gbrowse-slave? [y]
```

3.4 Samtools

Samtools are needed in case you want enable GBrowser to deal with reads and mapped reads against reference sequences. The version of Samtools showing the best integration with GBrowser is the 0.1.19.

```
$ wget -c http://sourceforge.net/projects/samtools/
    files/samtools/0.1.19/samtools-0.1.19.tar.bz2
$ tar xvjf samtools-0.1.19.tar.bz2
$ sudo mv samtools-0.1.19 /opt/
$ sudo chmod -R 755 /opt/samtools-0.1.19
$ cd /opt/samtools-0.1.19/
```

Before compilation, the Makefile must be edited:

```
$ sudo nano -w Makefile
```

Add the flag -fPIC at line 4:

```
4 -* CFLAGS=  -g -Wall   -O2
4 +* CFLAGS=  -fPIC -g -Wall -O2
```

Compile with make:

```
$ sudo make
```

Compile also bcftools:

```
$ cd bcftools
$ sudo make
```

Export samtools executables to the PATH, by editing the .bashrc file

```
$ nano -w .bashrc
```

And adding following lines:

```
$ sudo export PATH=/opt/samtools-0.1.19:/opt/
    samtools-0.1.19/misc:/opt/samtools-0.1.19/
    bcftools:$PATH
```

Install Perl module Bio::Samtools via CPAN:

```
$ sudo perl -MCPAN -e shell
>cpan install LDS/Bio-SamTools-1.39.tar.gz
```

When asked about the location of bam.h and libbam.a files specify the install path of samtools (*/opt/samtools-0.1.19*).

Chapter 4

Data processing

4.1 Prerequisites

- Get the *Gigaspora margarita* mitochondrial complete genome sequence (acc. no. JQ041882)[11] as both fasta and GenBank (full) file from NCBI (`http://www.ncbi.nlm.nih.gov/nuccore/JQ041882`).
- **Install bowtie2**

  Refer to `http://devbioinfo.to.cnr.it/doku.php?id=rachaelx:bowtop2` for instructions to install bowtie2.
- **Install BWA**

  Refer to `http://devbioinfo.to.cnr.it/doku.php?id=rachaelx:bwa` for instructions to install BWA.

4.2 Run bowtie2 with reads

Merge reads files (replicates) belonging to the same experimental condition. File name prefixes used below are for example only, and do not represent actual file names.

```
$ cat cond1_rep1_R1.fastq cond1_rep2_R1.fastq >
    cond1_all_replicates_R1.fastq
```

In this case, in the input file name prefix, cond stands for '*Condition*', rep for '*replica*' and R1 identifies left (forward) reads group; in the output, the prefix is meant to identify all left reads from all replica from condition n.1. Create bowtie2 index

```
$ bowtie2-build JQ041882.fa JQ041882-index
```

Align reads belonging to the same condition to the reference sequence

```
$ bowtie2-align -p 50 JQ041882-index/JQ041882 -U
    cond1_all_replicates_R1.fastq -S
    cond1_all_replicates_R1.sam
```

Convert sam to bam

```
$ samtools view -F4 -bt JQ041882.fa.fai -o
    cond1_all_replicates_R1.bam
    cond1_all_replicates_R1.sam
```

Sort bam file and index it

```
$ samtools sort cond1_all_replicates_R1.bam
    cond1_all_replicates_R1.sorted
$ samtools index cond1_all_replicates_R1.sorted.bam
```

The procedure must be performed for each experimental conditions. It might be useful to write a *bash* script to run all these steps automatically. An example of such *bash* script is included in Appendix.

4.3 Run BWA with assembled transcripts

Gigaspora margarita BEG34 transcriptome (file Trinity.Cufffly.fasta) was obtained as described in Appendix.
Align contigs to reference sequence:

```
$ bwa index JQ041882.fa
$ bwa aln -f aligned_contigs JQ041882.fa Trinity.
    Cufffly.fasta
$ bwa samse -f aligned_contigs.sam JQ041882.fa
    aligned_contigs Trinity.Cufffly.fasta
```

Mapped contigs extraction with samtools:

```
$ samtools view -bS -F 4 aligned_contigs.sam -o
    aligned_contigs.bam
$ samtools sort aligned_contigs.bam  aligned_contigs
    .sort
$ samtools index aligned_contigs.sort.bam
```

4.4 Convert GenBank file to Gff3

GFF3 (`http://gmod.org/wiki/GFF3`) is a widely used standard format for genomic annotation.

```
$ bp_genbank2gff3.pl JQ041882.gb
```

Chapter 5

GBrowser configuration

5.1 SQlite database creation

Load reference sequence fasta and gff3 files into SQLite database

```
$ bp_seqfeature_load.pl -a DBI::SQLite -c -f -d ./
    Mito1.sqlite JQ041882.fa JQ041882.gb.gff
$ sudo ln -s Mito1.sqlite /var/lib/gbrowse2/
    databases/Mito1.sqlite
```

5.2 Configuration editing

5.2.1 Accessory files

Place a copy of sorted bowtie2 outputs and their indexes in GBrowser's database directory

```
$ sudo ln -s cond*_all_replicates_R1.sorted.bam* /
    var/lib/gbrowse2/databases/.
```

Place a copy of indexed bwa output and its indexed in GBrowser's database directory, and change files/directory permissions

```
$ sudo ln -s aligned_contigs.sort.bam* /var/lib/
    gbrowse2/databases/.
$ sudo chmod -R 755 /var/lib/gbrowse2/databases/
    gigaspora
```

```
$ sudo chown -R www-data:www-data /var/lib/gbrowse2/
    databases/gigaspora
$ sudo chmod -R 644 /var/lib/gbrowse2/databases/
    Mito1.sqlite
$ sudo chown -R www-data:www-data /var/lib/gbrowse2/
    databases/Mito1.sqlite
```

5.2.2 Global configuration file

Modify the file */etc/Gbrowse2/GBrowse.conf*, containing setting for all data sources, to tell GBrowse that the newly created database must be showed on startup in addition to the yeast default database. First add a data-source stanza to the "**DATASOURCE DEFINITIONS**" section of the file (bottom of the file):

```
[Mitochondrion]
description   = Gigaspora Mitochondrion
path          = Mito1.conf
```

The description will appear in the drop-down menu offered to the user in the navigation bar. The path is the name of the database .conf file we need to create (next section).

Once you added this stanza, make it the default one by searching the "**DEFAULT DATASOURCE**" section (right above the previous one) and replacing the existing with your stanza's name (in the example, "Mitochondrion").

In case User Account Registration is not wanted, edit the relative section, and turn off the option setting to 0 the values.

```
######## User Account Registration Database ######
# If no authentication plugin is defined, and
# "user_accounts" is true, then GBrowse
# will attempt to use its internal user accounts
   database
# to authenticate and/or register users.
user_accounts                = 0
user_accounts_registration   = 0
user_accounts_openid         = 0
```

5.2.3 Data Source configuration file

Create a data source configuration file in */etc/gbrowse2*:

```
$ sudo touch /etc/gbrowse2/Mito1.conf
$ sudo chmod 444 /etc/gbrowse2/Mito1.conf
$ sudo chown root:root /etc/gbrowse2/Mito1.conf
```

Edit the configuration file, in order first to let GBrowse use SQLite databases, defining the suited adaptor in the "**GENERAL**" section, so to have:

```
[GENERAL]
description   = Output Database
db_adaptor    = Bio::DB::SeqFeature::Store
db_args       = -adaptor DBI::SQLite
                -dsn /var/lib/gbrowse2/databases/Mito1
                    .sqlite
```

In the same section, define the following arguments:

plugins, tells Gbrowse to activate pre-installed bioperl plugins;

default features, list of tracks to switch on by default;

and ***initial landmark***, name of the reference sequence (the name of the fasta file used with the *bp_seqfeature_load.pl* command) followed by the genomic range you wish to show on startup.

```
plugins  = BatchDumper
default features = JQ041882 Genes Clones DNA
   Translation EST 26_27 28_29_30 31_32 33_34 bwa
initial landmark = JQ041882:1..96,998
```

In this configuration, the BatchDumper plugin (to enable sequence download) is used, and the initial landmark value allows the visualization of the entire mitochondrial sequence.

Add a "**DATABASE**" section to the conf file, right below the "GENERAL" section, in which multiple database stanzas may be included (one for each aligned reads bam file to be shown):

```
[track_name:database]
db_adaptor      = Bio::DB::Sam
db_args         = -fasta /var/www/gbrowse2/databases/
   JQ041882.fa
```

```
                    --bam     /var/www/gbrowse2/databases/
                        track_name.bam
search options = default
```

In mitochondrial genome case, the track names are the outputs of bowtie2 (26_27, 28_29_30, 31_32, 33_34) and bwa.

Define the "**DEFAULT GLYPH SETTINGS**" section which describes the default graphic behaviour of feature tracks.

```
[TRACK DEFAULTS]
glyph          = generic
height         = 10
bgcolor        = lightgrey
fgcolor        = black
font2color     = blue
label density  = 25
bump density   = 100
link           = AUTO
```

More over, the default link option "AUTO" generates an automatic link to a helper script named "gbrowse_details", which lets the user visualize sequences and annotations loaded in the database for the selected feature.

Create the "**TRACK CONFIGURATION**" section, which defines the visualization options for generic genome features (e.g. Genes, DNA, Reading Frame) and for each database stanza in the "Database" section. For example, for the database entry 26_27:

```
[26_27]
feature        = coverage
glyph          = wiggle_xyplot
database       = mitochondrion
height         = 50
fgcolor        = black
bicolor_pivot  = 20
pos_color      = blue
neg_color      = red
key            = Coverage (xyplot)
category       = Reads
label          = 0
```

This stanza creates a track that shows coverage data. To show individual aligned contigs add the following track stanza:

```
[bwa]
```

```
feature          = match
glyph            = segments
draw_target      = 1
show_mismatch    = 1
mismatch_color   = red
database         = mitochondrion
bgcolor          = blue
fgcolor          = white
height           = 5
label density    = 50
bump             = fast
key              = Reads
category         = Reads
```

The working mitochondrion genome GBrowse configuration file is included in Appendix.

5.3 Configuration test

Point the preferred browser to the local path:

```
http://localhost/cgi-bin/gb2/gbrowse/
```

The application can also be browsed at the following URL:

```
http://gbrowse.to.cnr.it/cgi-bin/gb2/gbrowse
```

that is available via IPv4 and IPv6 protocols.

The *Gigaspora* annotated mitochondrial genome will be displayed with the tracks activated during configuration.

For advanced topics, such as configuring the user login and custom track upload system, and restricting access to certain databases and tracks via user authentication, see `http://gmod.org/wiki/GBrowse_2.0_Install_HOWTO/Advanced`

Chapter 6

Appendix

6.1 *Gigaspora margarita* BEG34 transcriptome

In the absence of a reference genome for *Gigasprora margarita* BEG34, a *de novo* assembly was generated using reads from 4 *in vitro* normalized paired end libraries (dataset 1) obtained from the wild type strain (B+ line) of *G. margarita* containing the endobacterium sampled in four moments of the fungal life cycle (quiescent spores, germinating spores, spores treated with strigolactone and extraradical mycelium), without replicates, and 14 single end libraries (dataset 2) obtained from both wild type strain and the cured line (B- line) sampled in three phases of the fungal life cycle (germinating spores, spores treated with strigolactone, and symbiotic mycelium thriving inside the roots). In total, five conditions were analyzed, leading to 18 libraries. Dataset 1 was composed by four *in vitro* normalised Paired-end libraries, obtained from the wild type strain (B+ line) of *G. margarita*, sampled in the following stages of the fungal life cycle: quiescent spores (GOU-13), germinating spores (GOU-14), spores treated with strigolactone (GOU-15) and extraradical mycelium (GOU-16). Dataset 2 was composed by 14 Single-end libraries, obtained from the wild type strain (B+ line) and the cured line (B- line) of *G.margarita*, sampled in the following stages of the fungal life cycle: germinating spores (B+: GDR-25/26/27; B-: GDR-28/29/30), spores treated with strigolactone (B+: GDR-31/32; B-: GDR-33/34) and mycorrhizal roots (B+: GDR-35/36; B-: GDR-37/38). The *de novo* assembly of dataset 1 and datase 2 libraries was performed on a 60 core and 256 GB RAM machine, running Ubuntu server 12.04 LTS, using Trinity v.Trinityrnaseq_r20131110[15]. First trials indicated that the available amount of memory was not sufficient to handle all the raw reads and, following the Trinity manual, we performed *in silico* reads normalization for

each of the libraries from dataset 1 and 2, to a max coverage of 30. Libraries from mycorrhizal roots (GDR-35 to GDR-38) were not subjected to normalization and were not used for the *de novo* assembly, since only a fraction of reads were ascribable to the fungal transcriptome (13.4%, 17.7%, 16.6%, 7.6%), whereas the larger part of the reads were from the plant host (*Lotus*). All the normalised single-end dataset 2 libraries were merged together with the paired-end dataset 1 left ends. Trinity was run with the following characterizing options, suited to assemble a gene-dense compact genome, such as a fungal genomes, and to minimize the number of isoforms per transcript:

```
$ Trinity.pl --seqType fq --CPU 30 --JM 150G --
    min_contig_length 350 --jaccard_clip --
    min_kmer_cov 2 --CuffFly --group_pairs_distance
    300 --extended_lock
```

6.2 Examples

6.2.1 Network interfaces configuration file

```
# This file describes the network interfaces available
    on your system
# and how to activate them. For more information, see
    interfaces(5).

# The loopback network interface
auto lo
iface lo inet loopback

# The primary network interface
auto eth0
iface eth0 inet static
        address 150.145.48.169
        netmask 255.255.255.0
        network 150.145.48.0
        broadcast 150.145.48.255
        gateway 150.145.48.1
        # dns-* options are implemented by the
            resolvconf package, if installed
        dns-nameservers 150.145.48.8 150.145.48.9
        dns-search to.cnr.it

iface eth0 inet6 static
```

```
        address 2a00:1620::169
        netmask 64
        gateway 2a00:1620::1
# Disable autoconf
    post-up echo 0 > /proc/sys/net/ipv6/conf/default/
       accept_ra
    post-up echo 0 > /proc/sys/net/ipv6/conf/all/
       accept_ra
    post-up echo 0 > /proc/sys/net/ipv6/conf/$IFACE/
       accept_ra
    post-up echo 0 > /proc/sys/net/ipv6/conf/default/
       autoconf
    post-up echo 0 > /proc/sys/net/ipv6/conf/all/
       autoconf
    post-up echo 0 > /proc/sys/net/ipv6/conf/$IFACE/
       autoconf
```

6.2.2 Batch bowtie2 *bash* script

```
#!/bin/bash
#bowtie2-build JQ041882.fa JQ041882-index

echo
echo "`date` :: Merging files ..."
cat ../RUN_2/130322_SN365_B_L004_GDR-26_R1.fastq ../
   RUN_2/130322_SN365_B_L004_GDR-27_R1.fastq > GDR
   -26-27_R1.fastq
echo "`date` :: done"
echo "`date` :: Running bowtie2: aligning"
bowtie2-align -p 50JQ041882-index/JQ041882 -U GDR
   -26-27_R1.fastq -S GDR-26-27_R1.sam
echo "`date` :: done"
echo "`date` :: Running bowtie2: converting files sam
   to bam"
samtools view -F4 -bt JQ041882.fa.fai -o GDR-26-27_R1.
   bam GDR-26-27_R1.sam
echo "`date` :: Running bowtie2: sorting bam file"
samtools sort GDR-26-27_R1.bam GDR-26-27_R1.sorted
echo "`date` :: Running bowtie2: indexing bam file"
samtools index GDR-26-27_R1.sorted.bam
echo "`date` :: done"
rm GDR-26-27_R1.fastq

echo
```

```
echo "`date` :: Merging files ..."
cat ../RUN_2/130322_SN365_B_L004_GDR-28_R1.fastq ../
    RUN_2/130322_SN365_B_L004_GDR-29_R1.fastq ../RUN_2
    /130322_SN365_B_L004_GDR-30_R1.fastq > GDR-28-29-30
    _R1.fastq
echo "`date` :: done"
echo "`date` :: Running bowtie2: aligning"
bowtie2-align -p 50JQ041882-index/JQ041882 -U GDR
    -28-29-30_R1.fastq -S GDR-28-29-30_R1.sam
echo "`date` :: done"
echo "`date` :: Running bowtie2: converting files sam
    to bam"
samtools view -F4 -bt JQ041882.fa.fai -o GDR-28-29-30
    _R1.bam GDR-28-29-30_R1.sam
echo "`date` :: Running bowtie2: sorting bam file"
samtools sort GDR-28-29-30_R1.bam GDR-28-29-30_R1.
    sorted
echo "`date` :: Running bowtie2: indexing bam file"
samtools index GDR-28-29-30_R1.sorted.bam
echo "`date` :: done"
rm GDR-28-29-30_R1.fastq

echo
echo "`date` :: Merging files ..."
cat ../RUN_2/130322_SN365_B_L004_GDR-31_R1.fastq ../
    RUN_2/130322_SN365_B_L005_GDR-32_R1.fastq > GDR
    -31-32_R1.fastq
echo "`date` :: done"
echo "`date` :: Running bowtie2: aligning"
bowtie2-align -p 50JQ041882-index/JQ041882 -U GDR
    -31-32_R1.fastq -S GDR-31-32_R1.sam
echo "`date` :: Running bowtie2: converting files sam
    to bam"
samtools view -F4 -bt JQ041882.fa.fai -o GDR-31-32_R1.
    bam GDR-31-32_R1.sam
echo "`date` :: Running bowtie2: sorting bam file"
samtools sort GDR-31-32_R1.bam GDR-31-32_R1.sorted
echo "`date` :: Running bowtie2: indexing bam file"
samtools index GDR-31-32_R1.sorted.bam
echo "`date` :: done"
rm GDR-31-32_R1.fastq

echo
echo "`date` :: Merging files ..."
```

```
cat ../RUN_2/130322_SN365_B_L005_GDR-33_R1.fastq ../
    RUN_2/130322_SN365_B_L005_GDR-34_R1.fastq > GDR
    -33-34_R1.fastq
echo "`date` :: done"
echo "`date` :: Running bowtie2: aligning"
bowtie2-align -p 50JQ041882-index/JQ041882 -U GDR
    -33-34_R1.fastq -S GDR-33-34_R1.sam
echo "`date` :: Running bowtie2: converting files sam
    to bam"
samtools view -F4 -bt JQ041882.fa.fai  -o GDR-33-34_R1
    .bam GDR-33-34_R1.sam
echo "`date` :: Running bowtie2: sorting bam file"
samtools sort GDR-33-34_R1.bam GDR-33-34_R1.sorted
echo "`date` :: Running bowtie2: indexing bam file"
samtools index GDR-33-34_R1.sorted.bam
echo "`date` :: done"
rm GDR-33-34_R1.fastq
```

6.2.3 GBrowse configuration file

```
[GENERAL]
description   = Gigaspora mitocondrio
db_adaptor    = Bio::DB::SeqFeature::Store
db_args       = -adaptor DBI::SQLite
          -dsn     /var/lib/gbrowse2/databases/Mito1.
             sqlite


# just the basic track dumper plugin
plugins      = Blat BatchDumper

# list of tracks to turn on by default
default features = JQ041882 Genes Clones DNA
    Translation 26_27 28_29_30 31_32 33_34 bwa

# size of the region
region segment         = 10000

# feature to show on startup
#initial landmark = JQ041882:1..96,998
initial landmark = JQ041882:1..50000


##########################
```

```
# Database sections
##########################

[bwa:database]
db_adaptor      = Bio::DB::Sam
db_args         = -fasta /var/lib/gbrowse2/databases/
   gigaspora/sequence.fasta
                  -bam   /var/lib/gbrowse2/databases/
                     gigaspora/bwa.sort.bam
search options = none

[26_27:database]
db_adaptor      = Bio::DB::Sam
db_args         = -fasta /var/lib/gbrowse2/databases/
   gigaspora/sequence.fasta
                  -bam   /var/lib/gbrowse2/databases/
                     gigaspora/GDR-26-27_R1_corretto.
                     sorted.bam
search options = none

[28_29_30:database]
db_adaptor      = Bio::DB::Sam
db_args         = -fasta /var/lib/gbrowse2/databases/
   gigaspora/sequence.fasta
                  -bam   /var/lib/gbrowse2/databases/
                     gigaspora/GDR-28-29-30_R1_corretto
                     .sort.bam
search options = none

[31_32:database]
db_adaptor      = Bio::DB::Sam
db_args         = -fasta /var/lib/gbrowse2/databases/
   gigaspora/sequence.fasta
                  -bam   /var/lib/gbrowse2/databases/
                     gigaspora/GDR-31-32_R1.sorted.bam
search options = none

[33_34:database]
db_adaptor      = Bio::DB::Sam
db_args         = -fasta /var/lib/gbrowse2/databases/
   gigaspora/sequence.fasta
                  -bam   /var/lib/gbrowse2/databases/
                     gigaspora/GDR-33-34_R1.sorted.bam
search options = none
```

```
############################
# Default glyph settings
############################

[TRACK DEFAULTS]
glyph          = generic
height         = 10
bgcolor        = lightgrey
fgcolor        = black
font2color     = blue
label density  = 25
bump density   = 100
# where to link to when user clicks in detailed view
link           = AUTO

############################
# TRACK CONFIGURATION
# the remainder of the sections configure individual
    tracks
############################

[Genes]
feature          = gene
glyph              = gene
bgcolor            = peachpuff
label_transcripts  = 1
draw_translation   = 1
category           = Genes
key                = Protein-coding genes

[ReadingFrame]
feature            = mRNA
glyph              = cds
ignore_empty_phase = 1
category           = Genes
key                = Frame usage

[DNA]
glyph          = dna
global feature = 1
height         = 40
do_gc          = 1
gc_window      = auto
```

```
fgcolor          = red
axis_color       = blue
strand           = both
key              = DNA/GC Content

[Translation]
glyph            = translation
global feature   = 1
height           = 40
fgcolor          = purple
start_codons     = 0
stop_codons      = 1
translation      = 6frame
key              = 6-frame translation


[26_27]
feature          = coverage
glyph            = wiggle_xyplot
database         = 26_27
height           = 50
fgcolor          = black
bicolor_pivot    = 20
pos_color        = blue
neg_color        = red
key              = GDR-26-27_R1
category         = Reads
label            = 0 # Labels on wiggle tracks are
   redundant.

[28_29_30]
feature          = coverage
glyph            = wiggle_xyplot
database         = 28_29_30
height           = 50
fgcolor          = black
bicolor_pivot    = 20
pos_color        = blue
neg_color        = red
key              = GDR-28-29-30_R1
category         = Reads
label            = 0 # Labels on wiggle tracks are
   redundant.
```

```
[31_32]
feature         = coverage
glyph           = wiggle_xyplot
database        = 31_32
height          = 50
fgcolor         = black
bicolor_pivot   = 20
pos_color       = blue
neg_color       = red
key             = GDR-31_32_R1
category        = Reads
label           = 0 # Labels on wiggle tracks are
   redundant.

[33_34]
feature         = coverage
glyph           = wiggle_xyplot
database        = 33_34
height          = 50
fgcolor         = black
bicolor_pivot   = 20
pos_color       = blue
neg_color       = red
key             = GDR-33_34_R1
category        = Reads
label           = 0 # Labels on wiggle tracks are
   redundant.

[bwa]
feature         = match
glyph           = segments
draw_target     = 1
show_mismatch   = 1
mismatch_color  = red
database        = bwa
bgcolor         = blue
fgcolor         = white
height          = 6
label density   = 50
bump            = fast
key             = Trinity transcripts (Bwa)
category        = Reads
```

```
##########################
# Plugin configuration
##########################

[Aligner:plugin]
alignable_tracks    = EST
upcase_tracks       = CDS Motifs
upcase_default      = CDS

[Motifs:overview]
feature         = polypeptide_domain
glyph           = span
height          = 5
description     = 1
label           = 1
key             = Motifs
```

Bibliography

[1] GBrowse - GMOD; http://gmod.org/wiki/GBrowse. Visited April 2015.

[2] GMOD; http://gmod.org. Visited April 2015.

[3] WormBase; http://www.wormbase.org/tools/genome/gbrowse/c_elegans/. Visited April 2015.

[4] FlyBase; http://flybase.org/cgi-bin/gbrowse/dmel/. Visited April 2015.

[5] Rice Genome Annotation Project; http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/. Visited April 2015.

[6] Tuber Genome Browser; http://www.genoscope.cns.fr/externe/GenomeBrowser/Tuber. Visited April 2015.

[7] Francesca Cordero, Stefano Ghignone, Luisa Lanfranco, Giorgio Leonardi, Rosa Meo, Stefania Montani, and Luca Roversi. BIOBITS - A study on Candidatus Glomeribacter gigasporarum with a data warehouse. In C. Plant and C. Böhm, editors, *Database Technology for Life Sciences and Medicine*, pages 193–155. World Scientific Publishing, Singapore, 2010.

[8] Stefano Ghignone, Alessandra Salvioli, Iulia Anca, Erica Lumini, Giuseppe Ortu, Luca Petiti, Stéphane Cruveiller, Valeria Bianciotto, Pietro Piffanelli, Luisa Lanfranco, and Paola Bonfante. The genome of the obligate endobacterium of an AM fungus reveals an interphylum network of nutritional interactions. *The ISME journal*, 6:136–145, August 2012.

[9] Emilie Tisserant, Mathilde Malbreil, Alan Kuo, Annegret Kohler, Aikaterini Symeonidi, Raffaella Balestrini, Philippe Charron, Nina Duensing, Nicolas Frei dit Frey, Vivienne Gianinazzi-Pearson, Luz B

Gilbert, Yoshihiro Handa, Joshua R Herr, Mohamed Hijri, Raman Koul, Masayoshi Kawaguchi, Franziska Krajinski, Peter J Lammers, Frederic G Masclaux, Claude Murat, Emmanuelle Morin, Steve Ndikumana, Marco Pagni, Denis Petitpierre, Natalia Requena, Pawel Rosikiewicz, Rohan Riley, Katsuharu Saito, Hélène San Clemente, Harris Shapiro, Diederik van Tuinen, Guillaume Bécard, Paola Bonfante, Uta Paszkowski, Yair Y Shachar-Hill, Gerald A Tuskan, J Peter W Young, Peter W Young, Ian R Sanders, Bernard Henrissat, Stefan A Rensing, Igor V Grigoriev, Nicolas Corradi, Christophe Roux, and Francis Martin. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, 110(50):20117–22, December 2013.

[10] VISTA tools; http://genome.lbl.gov/vista/index.shtml. Visited April 2015.

[11] Adrian Pelin, Jean-François Pombert, Alessandra Salvioli, Linda Bonen, Paola Bonfante, and Nicolas Corradi. The mitochondrial genome of the arbuscular mycorrhizal fungus Gigaspora margarita reveals two unsuspected trans-splicing events of group I introns. *The New phytologist*, 194(3):836–45, May 2012.

[12] Giancarlo Birello, Ivano Fucile, Valter Giovanetti, and Anna Perin. Rapporto tecnico Ceris-CNR n.44 KVM: an open-source framework for virtualization. Technical report, CNR, Torino, 2013.

[13] Giancarlo Birello, Ivano Fucile, Valter Giovanetti, and Anna Perin. Rapporto tecnico Ceris-CNR n.37 - Storage in HA: cluster attivo/passivo open source, 2011.

[14] Giancarlo Birello, Ivano Fucile, Valter Giovanetti, and Anna Perin. Rapporto Tecnico Ceris-CNR n.41 Storage in HA: Manutenzione ordinaria e straordinaria, 2012.

[15] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–52, July 2011.