fGAP Documentation

Release 0.1

Byoungnam Min and In-Geol Choi

November 30, 2016

Contents

fGAP (fungal Genome Annotation Pipeline) performs gene prediction on given genome assembly and transcriptomic reads.

Source code is available via GitHub.

CHAPTER 1

License

fGAP is freely available only for non-commercial user. Please contact igchoi@korea.ac.kr for purchasing lisence for commercial usage.

CHAPTER 2

Contact

- Byoungnam Min: mbnmbn00@korea.ac.kr
- Prof. In-Goel Choi: igchoi@korea.ac.kr

2.1 Installation

2.1.1 0. Pre-requisites

fGAP requires several softwares installed before running the command. We provide simple installation guide for each software. The guide was intensely tested in Ubuntu 14.04 LTS (Unfortunately, we do not provide the supports for other platforms including Windows, OS X, or other Linux distributions in this version).

Download fGAP

Download fGAP using GitHub `clone`. Suppose we are installing fGAP in your `$HOME` directory, but you are free to change the location.

```
cd $HOME
git clone https://github.com/mbnmbn00/fGAP.git
```

BLAST+ installatioon

BLAST+ is used in Maker and BUSCO running.

```
sudo apt-get install ncbi-blast+
```

Trinity installation

Trinity performs efficient and robust *de novo* reconstruction of transcriptomes from RNA-seq data. https://github.com/trinityrnaseq/trinityrnaseq/wiki

Download and Install Trinity v2.2.0 using github.

```
cd $HOME/fGAP/external
git clone https://github.com/trinityrnaseq/trinityrnaseq.git
cd trinityrnaseq
make
```

Maker2 installation

Maker2 is an eay-to-use annotation pipeline designed for emerging model organism genomes. http://www.gmod.org/wiki/MAKER

Please note that you need a proper license to use Maker2.

```
# Move to install directory
cd $HOME/fGAP/tools/

# Download and unzip maker2 named maker-2.31.8.tgz
tar -zxvf maker-2.31.8.tgz

# Install Maker2 pre-requisites
cd maker/src
sudo apt-get install libpq-dev
sudo apt-get install exonerate # version 2.2.0
perl Build.PL
sudo ./Build installdeps
./Build installexes
./Build install

# Configure RepeatMasker
# First download repbase manually at http://www.girinst.org/server/RepBase/index.php
# Then move it to $HOME/fGAP/maker/exe/RepeatMasker/
tar -zxvf repeatmaskerlibraries-20150807.tar.gz # Or whatever you downloaded
cd $HOME/fGAP/tools/maker_2.31.8/exe/RepeatMasker/
./configure

# **TRF PROGRAM**
# This is the full path to the TRF program.
# This is now used by RepeatMasker to mask simple repeats.
# Enter path [  ]:
$HOME/fGAP/maker/exe/RepeatMasker/trf

# Add a Search Engine:
# 1. CrossMatch: [ Un-configured ]
# 2. RMBlast - NCBI Blast with RepeatMasker extensions: [ Un-configured ]
# 3. WUBlast/ABBlast (required by DupMasker): [ Un-configured ]
# 4. HMMER3.1 & DFAM: [ Un-configured ]

# 5. Done
```

```
# Enter Selection:
2

# **RMBlast (rmblastn) INSTALLATION PATH**
# This is the path to the location where
# the rmblastn and makeblastdb programs can be found.
# Enter path [  ]:
$HOME/fGAP/maker_2.31.8/exe/RepeatMasker/rmblast/bin
```

RepeatModeler installation

RepeatModeler is a de-novo repeat family identification and modeling package.

http://www.repeatmasker.org/RepeatModeler.html

Install RepeatModeler and its dependencies.

```
# Check perl version (ensure version > 5.8.8)
perl -v

# Now install RepeatModeler
cd $HOME/fGAP/tools/
wget http://www.repeatmasker.org/RepeatModeler-open-1-0-8.tar.gz
tar -zxvf RepeatModeler-open-1-0-8.tar.gz
cd RepeatModeler/
perl ./configure

# **REPEATMASKER INSTALLATION PATH**
# This is the path to the location where
# the RepeatMasker program suite can be found.
# Enter path [  ]:
$HOME/fGAP/maker/exe/RepeatMasker/

# **RECON INSTALLATION PATH**
# This is the path to the location where
# the RECON program suite can be found.
# Enter path [  ]:
$HOME/fGAP/tools/RECON-1.08/bin

# **RepeatScout INSTALLATION PATH**
# This is the path to the location where
# the RepeatScout program suite can be found.
# Enter path [  ]:
$HOME/fGAP/tools/RepeatScout-1/

# **TRF INSTALLATION PATH**
# This is the path to the location where
# the TRF program can be found.
# Enter path [  ]:
$HOME/fGAP/maker/exe/RepeatMasker

# Add a Search Engine:
# 1. RMBlast - NCBI Blast with RepeatMasker extensions: [ Un-configured ]
# 2. WUBlast/ABBlast: [ Un-configured ]

# 3. Done
# Enter Selection:
1
```

```
# **RMBlast (rmblastn) INSTALLATION PATH**
# This is the path to the location where
# the rmblastn and makeblastdb programs can be found.
# Enter path [  ]:
$HOME/fGAP/maker/exe/RepeatMasker/rmblast/bin
```

Braker1 installation

Braker1 is an unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS.

http://exon.gatech.edu/genemark/braker1.html

Install Braker1 and its dependencies.

```
# Copy gm_key
cp $HOME/fGAP/tools/gm_et_linux_64/gm_key ~/.gm_key

# Perl modules
sudo cpan YAML
sudo cpan App::cpanminus
sudo cpanm File::Spec::Functions
sudo cpanm Hash::Merge
sudo cpanm List::Util
sudo cpanm Logger::Simple
sudo cpanm Module::Load::Conditional
sudo cpanm Parallel::ForkManager
sudo cpanm POSIX
sudo cpanm Scalar::Util::Numeric
sudo cpanm YAML

# For bamtools
sudo apt-get install zlib1g-dev
```

InterProScan installation

InterProScan scans a sequence for matches against the InterPro protein signature databases.

https://github.com/ebi-pf-team/interproscan/wiki

Install InterProScan.

```
cd $HOME/fGAP/tools/
wget ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/5.18-57.0/interproscan-5.18-57.0-64-bit.tar.gz
tar -zxvf interproscan-5.18-57.0-64-bit.tar.gz
```

Install Python modules

fGAP requires several python modules and they can be installed by *pip*.

```
# Install pip
sudo apt-get install python-pip

# Install needed modules
sudo pip install biopython
sudo pip install numpy
sudo pip install intervaltree
```

You can check if fGAP is correctly installed.

```
python $HOME/fGAP/check_dependencies.py -o tmp
```

2.2 Usage

2.2.1 1. Prepare protein database

fGAP requires protein database in FASTA file. We recommend 3 ~ 4 organisms' proteome to save running time. For this, we provide a script to build your database using NCBI API.

Usage:

```
# Assume you downloaded fGAP in $HOME/fGAP
python $HOME/fGAP/fgap/download_sister_orgs.py\
 --download_dir <download_directory>\
 --taxon <taxon>\
 --num_sisters <number_of_sisters>\
 --email_address <your_email_address>
```

E-mail address is needed for NCBI Entrez.

Example command:

```
python $HOME/fGAP/fgap/download_sister_orgs.py\
 --download_dir sister_orgs\
 --taxon "Schizophyllum"\
 --num_sisters 3\
 --email_address mbnmbn00@gmail.com
```

All taxon levels are allowed for *–taxon* argument, but genus level is appropriate. Now make protein database:

```
cd sister_orgs/
gunzip *.faa.gz
cat ./*faa > prot_db.faa
```

You can now input *prot_db.faa* to fGAP.

2.2.2 2. Run fGAP

To run fGAP, you need three main arguments

- Genome assembly (FASTA)
- Transcriptomic reads (FASTQ)
- Protein database (FASTA)

Currently, fgap gets only **Illumina paired-end** reads files. The file names of two paired-end reads should have suffix like 'XX_1.fastq' and 'XX_2.fastq'. The prefixes should be same without '_' character. For example, File names would be like *hyphae_1.fastq* and *hyphae_2.fastq*.

Usage:

```
# Assume you downloaded fGAP in $HOME/fGAP
python $HOME/fGAP/fgap/fgap.py\
 --output_dir <output_directory>\
 --trans_read_files <transcriptome_reads_fastqs>\
```

```
--project_name <project_name_without_space>\
--genome_assembly <genome_assembly_fasta>\
--augustus_species <augustus_species>\
--org_id <organism_id>\
--sister_proteome <sister_proteome>\
--num_cores <number_of_cpus_to_be_used>\
```

- Augustus species: you should provide one augustus_species used in Augustus. This is the list what Augustus provides.

Phylum	Class	Species	augustus_species
Ascomycota	Eurotiomycetes	Aspergillus fumigatus	aspergillus_fumigatus
Ascomycota	Eurotiomycetes	Aspergillus nidulans	aspergillus_nidulans
Ascomycota	Eurotiomycetes	Aspergillus oryzae	aspergillus_oryzae
Ascomycota	Eurotiomycetes	Aspergillus terreus	aspergillus_terreus
Ascomycota	Leotiomycetes	Botrytis cinerea	botrytis_cinerea
Ascomycota	Saccharomycetes	Candida albicans	candida_albicans
Ascomycota	Saccharomycetes	Candida guilliermondii	candida_guilliermondii
Ascomycota	Saccharomycetes	Candida tropicalis	candida_tropicalis
Ascomycota	Sordariomycetes	Chaetomium globosum	chaetomium_globosum
Ascomycota	Eurotiomycetes	Coccidioides immitis	coccidioides_immitis
Basidiomycota	Agaricomycetes	Coprinus cinereus	coprinus
Basidiomycota	Agaricomycetes	Coprinus cinereus	coprinus_cinereus
Basidiomycota	Agaricomycetes	Cryptococcus neoformans gattii	cryptococcus_neoformans_gattii
Basidiomycota	Agaricomycetes	Cryptococcus neoformans gattii	cryptococcus_neoformans_neoformans_B
Basidiomycota	Agaricomycetes	Cryptococcus neoformans gattii	cryptococcus_neoformans_neoformans_JEC21
Ascomycota	Saccharomycetes	Debaryomyces hansenii	debaryomyces_hansenii
Microsporidia		Encephalitozoon cuniculi	encephalitozoon_cuniculi_GB
Ascomycota	Saccharomycetes	Eremothecium gossypii	eremothecium_gossypii
Ascomycota	Sordariomycetes	Fusarium graminearum	fusarium_graminearum
Ascomycota	Eurotiomycetes	Histoplasma capsulatum	histoplasma_capsulatum
Ascomycota	Saccharomycetes	Kluyveromyces lactis	kluyveromyces_lactis
Basidiomycota	Agaricomycetes	Laccaria bicolor	laccaria_bicolor
Ascomycota	Saccharomycetes	Lodderomyces elongisporus	lodderomyces_elongisporus
Ascomycota	Sordariomycetes	Magnaporthe grisea	magnaporthe_grisea
Ascomycota	Sordariomycetes	Neurospora crassa	neurospora_crassa
Basidiomycota	Agaricomycetes	Phanerochaete chrysosporium	phanerochaete_chrysosporium
Ascomycota	Saccharomycetes	Pichia stipitis	pichia_stipitis
Mucoromycotina	Mucorales	Rhizopus oryzae	rhizopus_oryzae
Ascomycota	Saccharomycetes	Saccharomyces cerevisiae	saccharomyces_cerevisiae_S288C
Ascomycota	Saccharomycetes	Saccharomyces cerevisiae	saccharomyces_cerevisiae_rm11-1a_1
Ascomycota	Schizosaccharomycetes	Schizosaccharomyces pombe	schizosaccharomyces_pombe
Basidiomycota	Ustilaginomycetes	Ustilago maydis	ustilago_maydis
Ascomycota	Saccharomycetes	Yarrowia lipolytica	yarrowia_lipolytica

- Organism ID will be used in naming gene ID

2.2.3 3. Output

Final output will be located in output directory you gave in the arguments

- fgap_output_prot.faa

- fgap_output.gff3
- fgap_output_stats.html

2.2.4 4. Trouble-shootings

This is very beta version of the software, so please don't hesistate reporting any bug or error you have encountered at mbnmbn00@korea.ac.kr or mbnmbn00@gmail.com.

2.3 Working principle