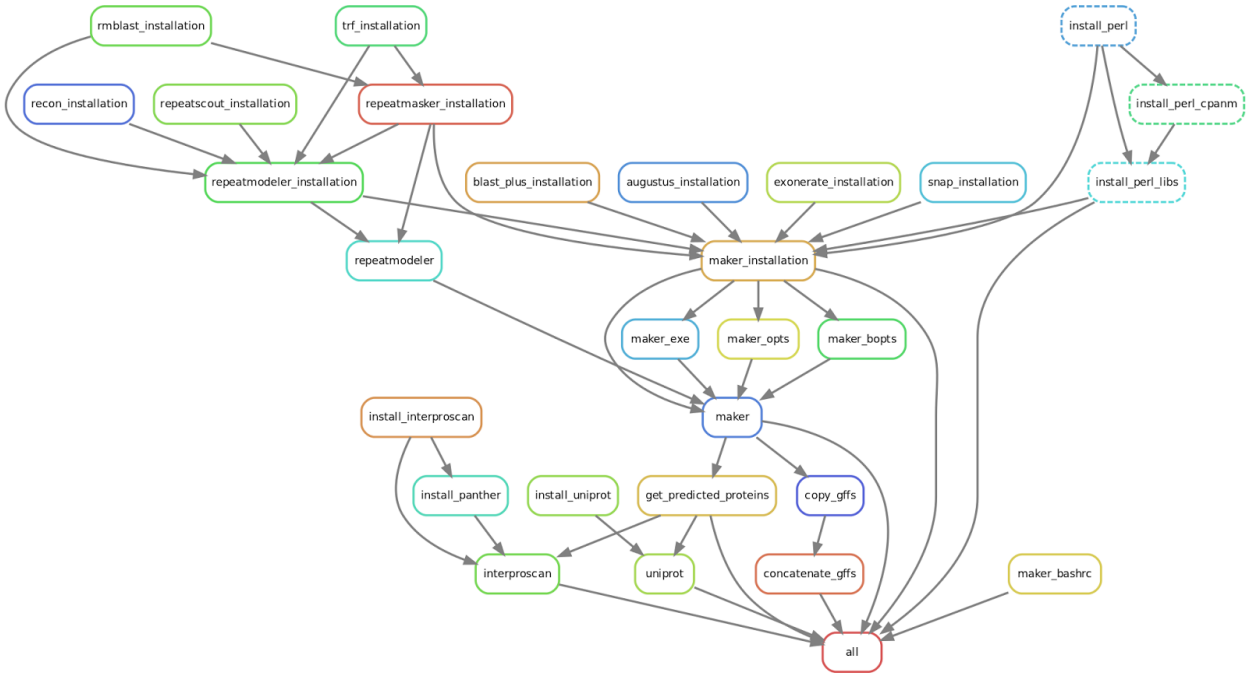
Introduction

An important step after the assembly of a genome is trying to identify structural elements and their functions within the genomic sequence. Some examples of these structural elements are genes, promoter regions and microRNAs. The roles they play are known as functional annotations.
For this project we focus on gene prediction, both structural and functional. The software package Maker is used for the prediction of genes in the genome assembly. Interproscan and Uniprot are then used to assign function to these predicted genes. The end result of the annotation process consists of a fasta file with the predicted proteins and their associated uniprot protein and GFF file with all genes identified in the genome assembly, including the protein domain prediction given by interproscan.

Process of annotation



The annotation pipeline first checks for installed software and databases. When software or a database is missing the pipeline will install it in the pre-defined location(s).
First step in annotating a genome is running Maker. This application is in itself a pipeline. It consists of repeat masking and gene prediction (augustus, exonerate, snap, blast). Maker can make use of different sources, but most importantly the selections of gene models and of a reference protein set are key for proper gene prediction. It is also highly recommended to use a RNAseq data set from the same organism: this will greatly improve the gene prediction.
When Maker has finished, the produced GFFs, predicted mRNA sequence and predicted protein sequences are collected.
The protein data set will then be search for known protein domains (interproscan) and known protein functions (BLAST analyses with the uniref50 database)

Required software

To run the annotation pipeline, you need:
wget: https://www.gnu.org/software/wget/
python 3: https://www.python.org/download/releases/3.0.1/
snakemake: https://bitbucket.org/johanneskoester/snakemake/wiki/Home
code from the VLPB git repository: https://github.com/vlpb3/NGS_snakemake_pipelines
On some Linux system, database development libraries are not available. These are required when installing BioPerl. This can be fixed by installing the libraries: *sudo apt-get install libdb6.0-dev libdb6.0* or ask your system administrator to do it for you.
The maker software (http://gmod.org/wiki/MAKER) can only be downloaded after requesting a license. You need to add the download URL to the JSON (see next section).

Configuring your pipeline

The configuration of the pipeline is located in two text files, stored in JSON. Any text editor can be used to edit these files. The first file is: *./src/workflows/annotation/maker/paths.json*. This file contains the paths on the file system for installation and data storage.

Key	**Value**	**Description**
base_dir	/tmp/	Base directory for installation, data storage, etc
executables	bin/maker_helpers/	Location to store the executables such as blast
databases	databases/	Location to store databases such as uniref
web_host	http://assembly.ab.wurnet.nl/~jvh/	Web host for some downloads
home_dir	/home/sven/	Home directory of the user. Required for perlbrew installation
download_dir	/tmp/	Location for storing downloads of software, database, etc. These downloads can be many GB and /tmp is then not always the best place to put in (file system may be filled up completely)

These paths are used in the general JSON config.json by substitution the name in {} with the value found in paths,json. In this file you can specify all options for Maker, perlbrew installation, databases, etc. The following table gives the most important options. Options not mentioned in this table should be left alone unless you are sure of what you are doing.

Key	Value	Description
Base		
working_dir	{base_dir}TEMP_MAKER	Location to store the Maker output
maker_opts		
genome	{base_dir}{executables}maker/data/dpp_contig.fasta	Location of genome assembly
organism_type	eukaryotic	Eukaryotic or prokaryotic
est	{base_dir}{executables}maker/data/dpp_est.fasta	Location of RNAseq / EST data set
protein	{base_dir}{executables}maker/data/dpp_protein.fasta	Location of set of protein sequence of closely related species
repeat_protein	"{base_dir}{executables}maker/data/te_proteins.fasta	Repeat protein data set
model_org	all	Gene models
snaphmm	{base_dir}{executables}snap/HMM/D.melanogaster.hmm	SNAP HMM gene models
augustus_species	fly	Augustus gene model
est2genome	1	Use est as gene evidence
protein2genome	1	Use proteins as gene evidence
executable_sources		
maker_URL	{web_host}maker-2.31.6.tgz	Download location for the maker executable. You will get this link after requesting a license.
functional_annotation		

uniprot_db	uniref50	Which uniprot database to use (uniref50, uniref90 or uniref100). Very large!

First run

For a first test run and to install all required software and databases, the default config.json will be sufficient: Maker supplies a test data test. The paths.json should modified to specify the correct locations.
Be aware: the software and databases required for annotation are very big. A basic installation requires 81 GB of storage!
To run the VLPB annotation pipeline, enter the directory *./src/workflows/annotation/maker/* and type *snakemake*.

Full run

After the first run has completed, which could take up to several hours, you can now configure the config.json file for your own genome assembly.
When something goes wrong, it is usually best to remove all temporary files. Type *snakemake clean_maker* to cleanup annotation files and maker configuration files. Note: when you have pointed 'working_dir' to an existing data folder, you will lose all your data!

Results

File	**Description**
concatenated.gff	Combined GFF. Maker produces a GFF per contig
predicted_proteins.faa	All protein sequences as predicted by Maker
protein_annotation.tsv	Interproscan annotation of predicted proteins (protein domains)
protein_uniprot.csv	Uniprot annotation of predicted proteins (based on protein BLAST)