



LEIDEN UNIVERSITY MEDICAL CENTER

Phylogenetic reconstruction

Michiel van Galen

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics



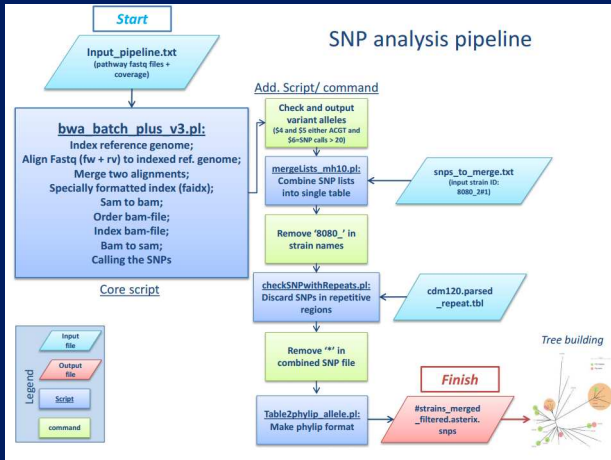
Input and goal

- Sequence data available for different strains of bacteria
- One FastQ file per strain

NGS throughput is much higher compared to conventional methods (Sanger sequencing). Increasing the chances on new insights.

However, there is little solutions available to accommodate the magnitude in the field of phylogenetic reconstruction.

Naive approach



Naive approach

Early workflow adapted from Sanger suffered from some limitations:

- Difficult to reproduce
- Poorly documented
- Using unconventional methods
- Not parallelized
- Susceptible to errors
- Customization or modification nearly impossible
- Stops at the tree construction

From bundle of scripts to pipeline

Re-factor the workflow into a complete pipeline

- Convert the workflow to an automated pipeline
- Replace custom scripts with maintained existing tools and methods
- Include cluster support
- Improve usability and customization

Breakdown of the pipeline

The workflow can be roughly broken down into two parts

- Per sample part - Analyze the samples separately
- Merged part - Combine output for each sample

Per sample part

These steps are for each sample the same and can be parallelized

- Add QC - Standard tools
- Alignment to canonical reference - BWA
- Variant calling and filtering - Samtools
- Mask variants in repeated regions - BEDtools

Merged part, combining the output

- Compare the variants between strains - Python
 - Merge the variant files into one matrix - VCFtools
- Use PHYLIP to infer a evolutionary tree
 - Create distance matrix (dnadist)
 - Create a phylogenetic tree

Implementation

The pipeline is designed to run on the LUMC Shark cluster

- All tools are available and maintained
- Pipeline is written in Make, compatible to run in parallel
- Reduced the number of custom scripts to just one
 - Not reinventing the wheel, outsource support for tools

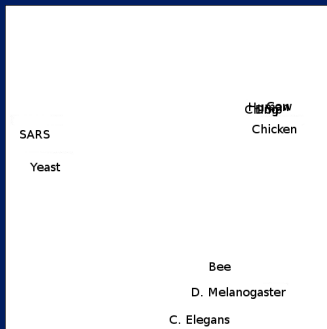
Possible expansions

- Improve usability even more
 - User friendly interface
 - More automation

- kMer analysis
 - Proven to work on meta-genomic datasets

kMer

- Calculate distance between samples based on occurrences of words of length k



Summarizing:

- Much room for pipeline development and automation
- Apply existing tools where possible reduce development time
- Data is relatively small compared to human data making our infrastructure well prepared

Acknowledgements:

Wilco Knetsch

Jeroen Laros

Martijn Vermaat

Jeroen Frank

LGTC