Across-kingdom optimization of Genotype by Sequence data analysis

Ruy Jauregui

AgResearch

eResearch NZ, Queenstown
Feb 2016

Motivation

While the underlying tenets of science are universal, specialization causes segregation of analysis methods due to the unique requirements of each research theme (i.e. human versus plant genetics).

Technology that spans broad research areas has great impact and facilitates interdisciplinary communication.

agresearch

GBS: Genotype By Sequence

- High-throughput screening method to identify genetic variations in plants and animals.
- Low cost per sample, fits hundreds of samples in a single run
- Lower bias than array based methods
- Allows the comparison of samples without a reference genome (higher accessibility)
- Facilitates genomic selection, kinship analysis, and determination of population structure
- The work shown here is based on data from two different models: an insect found as a pest in NZ, and a fungus of commercial interest
- The objective of this work is to develop methods that function in highly variable scopes of big data

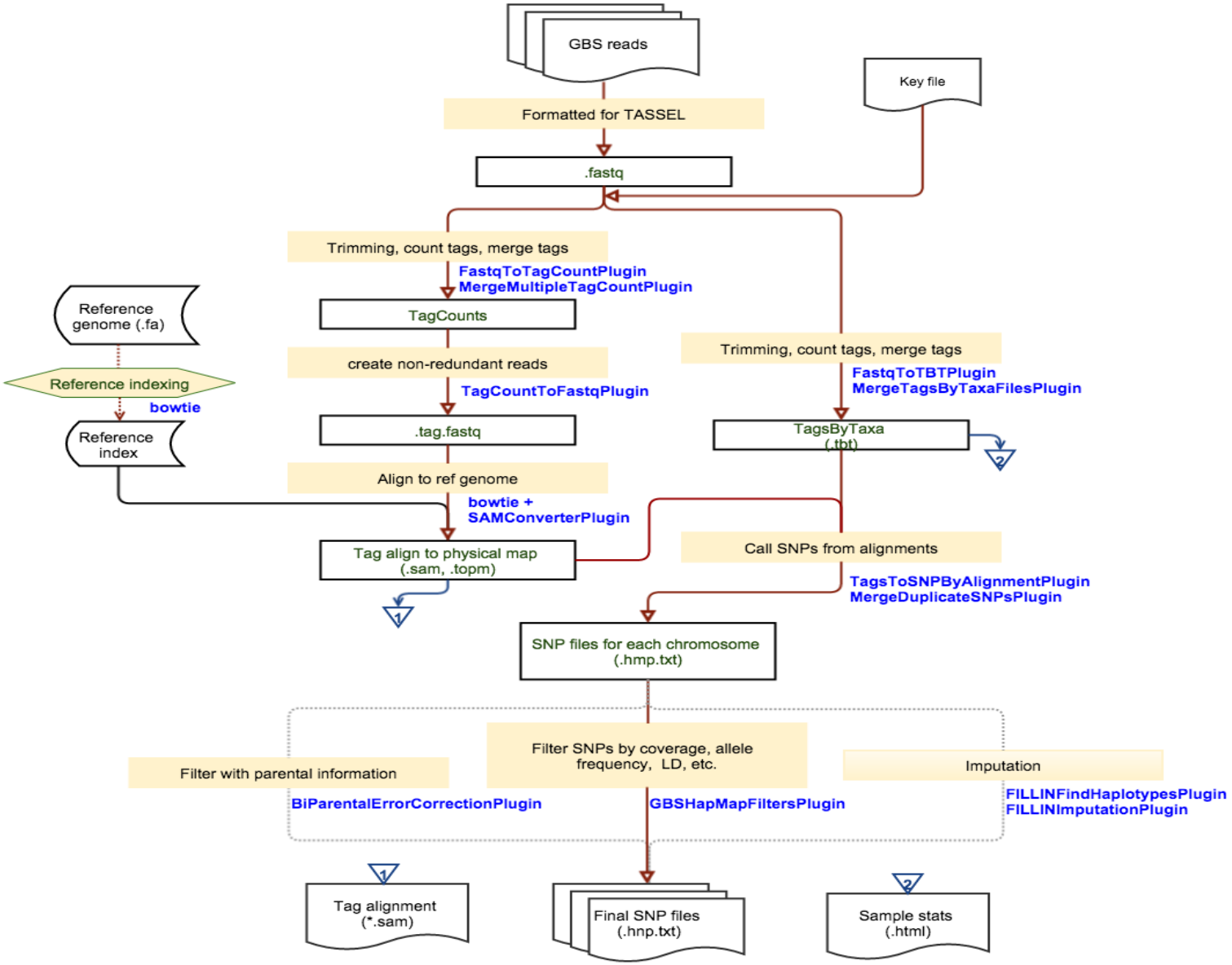
agresearch

GBS data

- A run with the Illumina HiSeq machine will produce about 300 million reads in a single flow cell run (about 50 Gb).
- The sequence output is highly redundant
- The reads have variable quality
- The path from raw data to understanding is arduous.

agresearch

Data analysis workflow



Read quality

- Reads are reported as lines of sequence/quality
- The quality line gives an error probability to each sequence letter
- Illumina tends to add more errors at the end of the read
- Reads can be trimmed according to the quality line
- There are public read trimmers already available

Aim

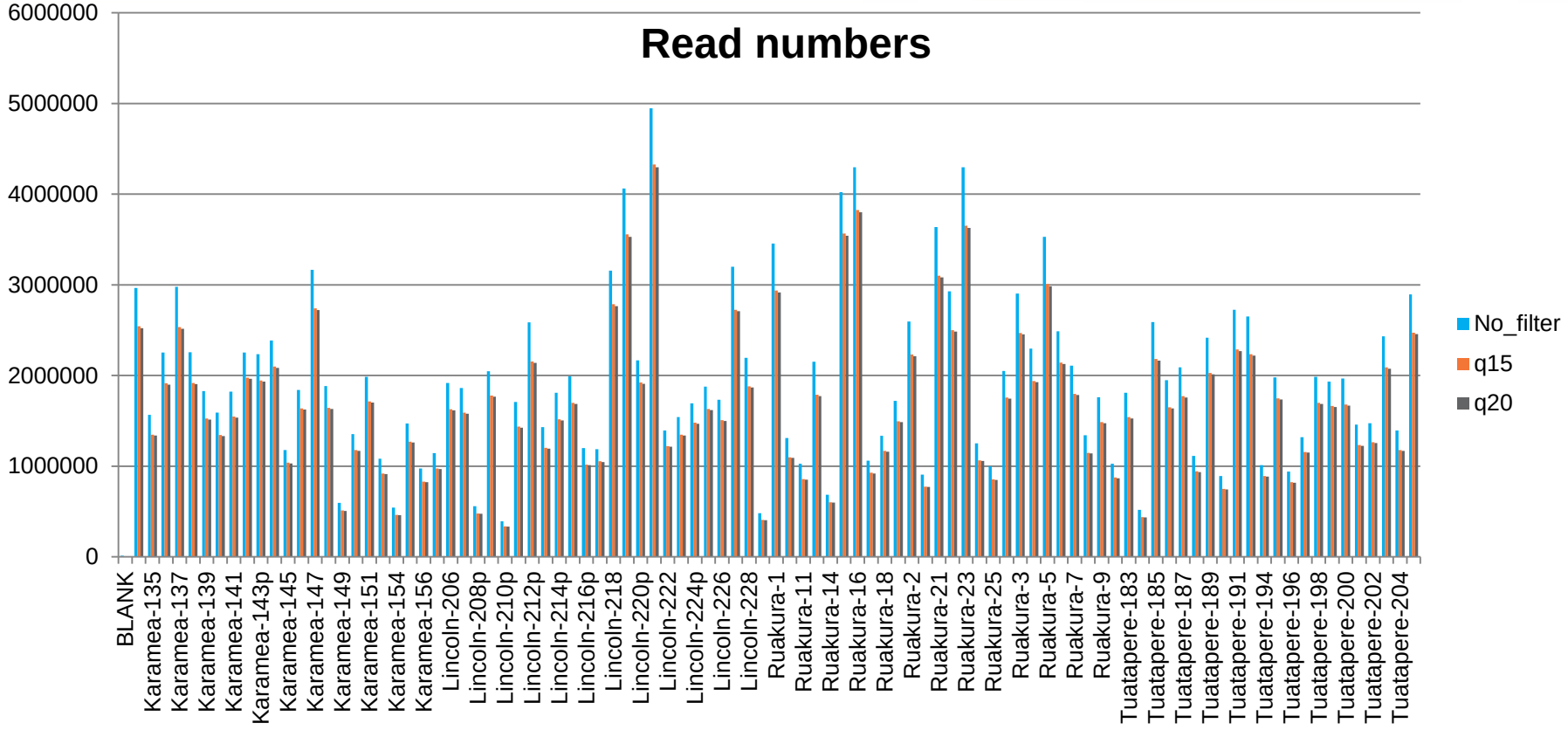
We aim to find early on action points to improve our interpretation of GBS data that are source-agnostic.

Here we investigate the effect of read trimming on two different data sources: An insect and a fungus.

These organisms are non-model, and there is no genomic information about them.

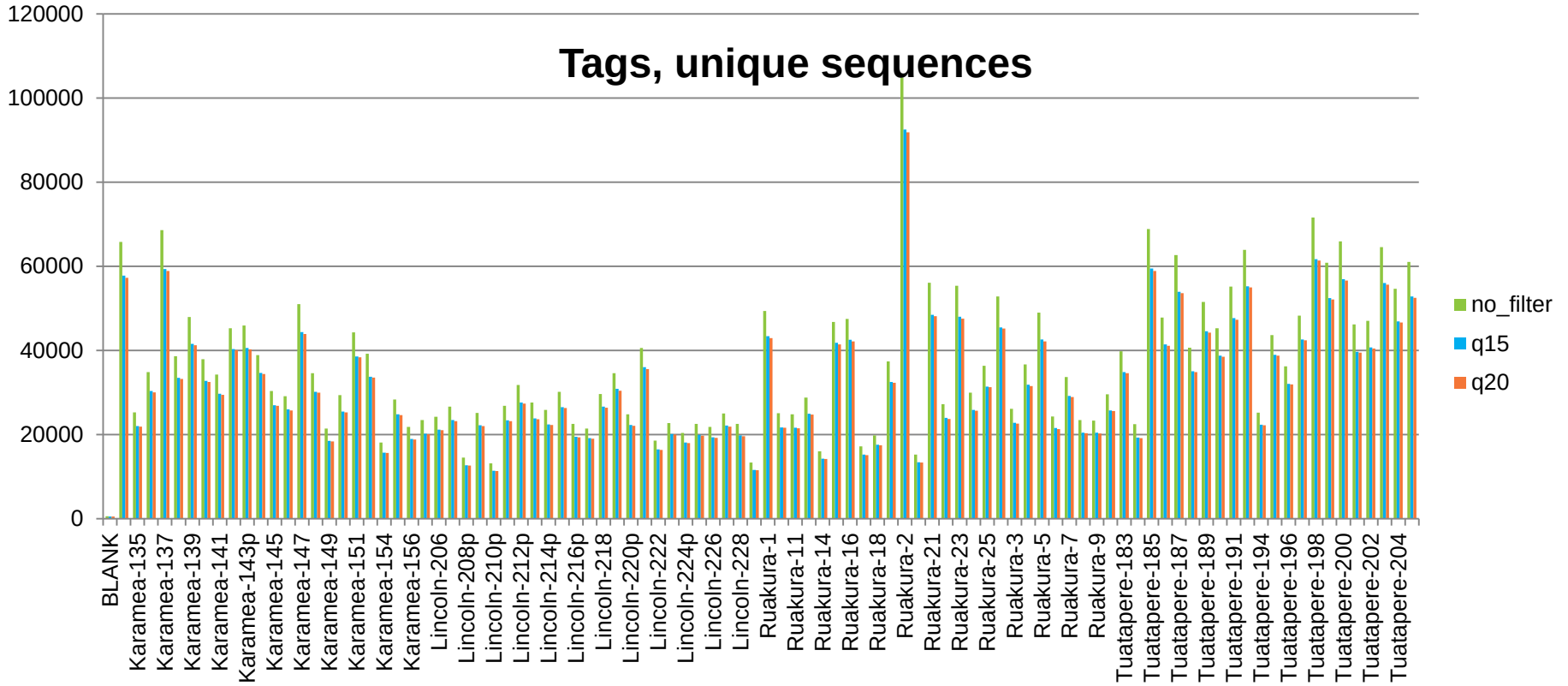
agresearch

Insect samples: trim effect on read number



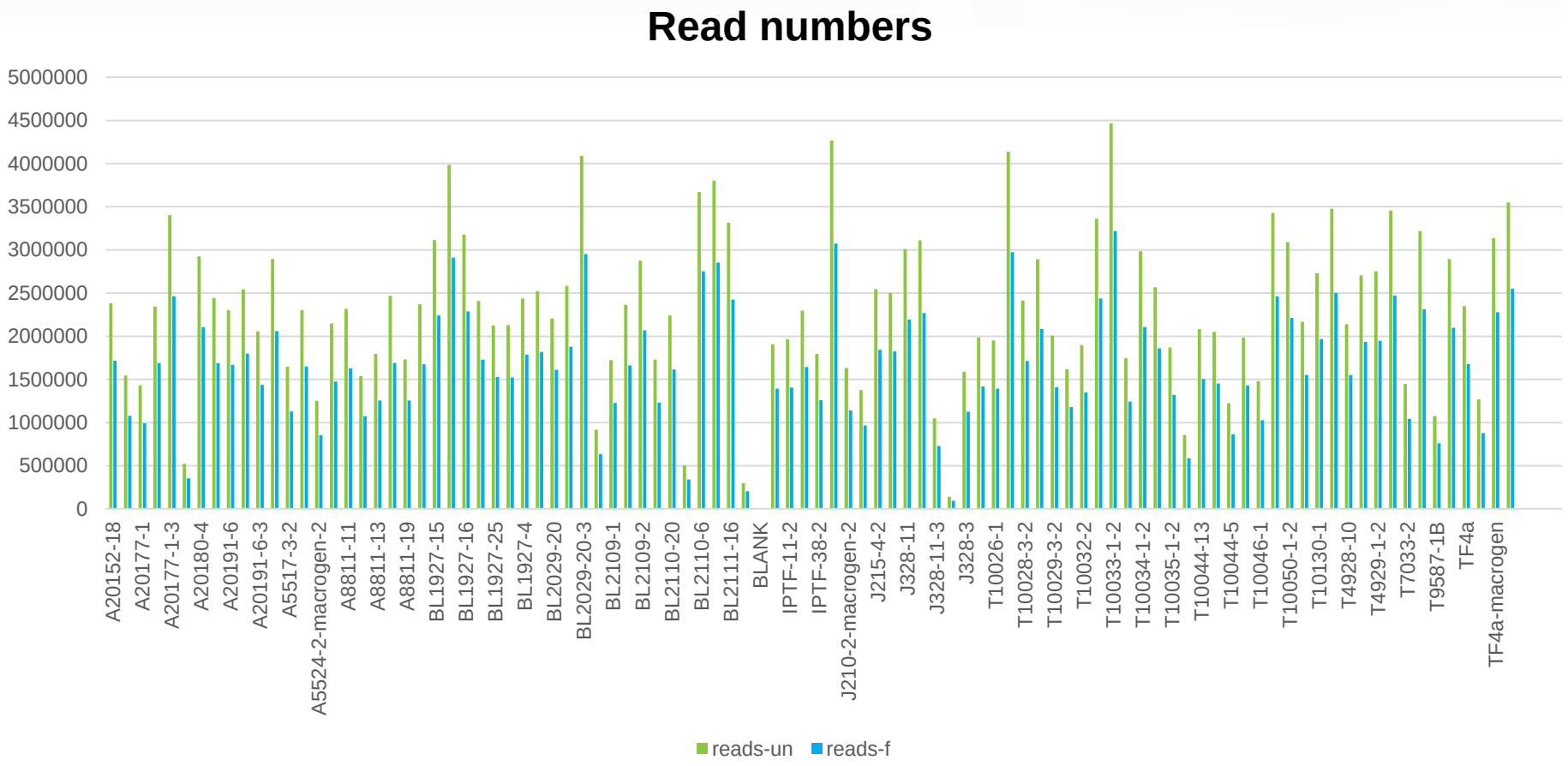
agresearch

Insect samples: trim effect on tag number



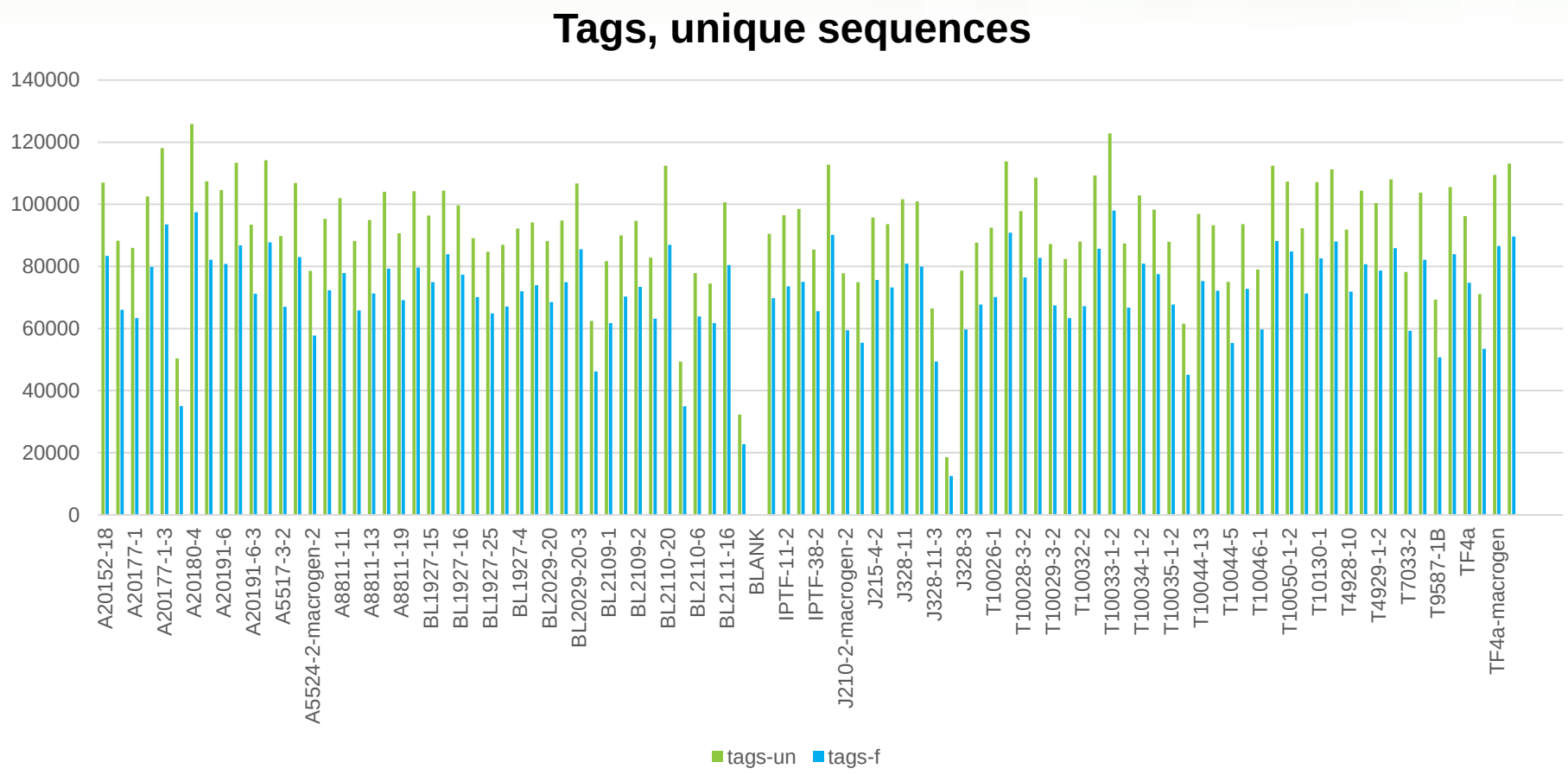
agresearch

Fungi samples: trim effect on read number



agresearch

Fungi samples: trim effect on tag number



agresearch

Effect of stringent filtering

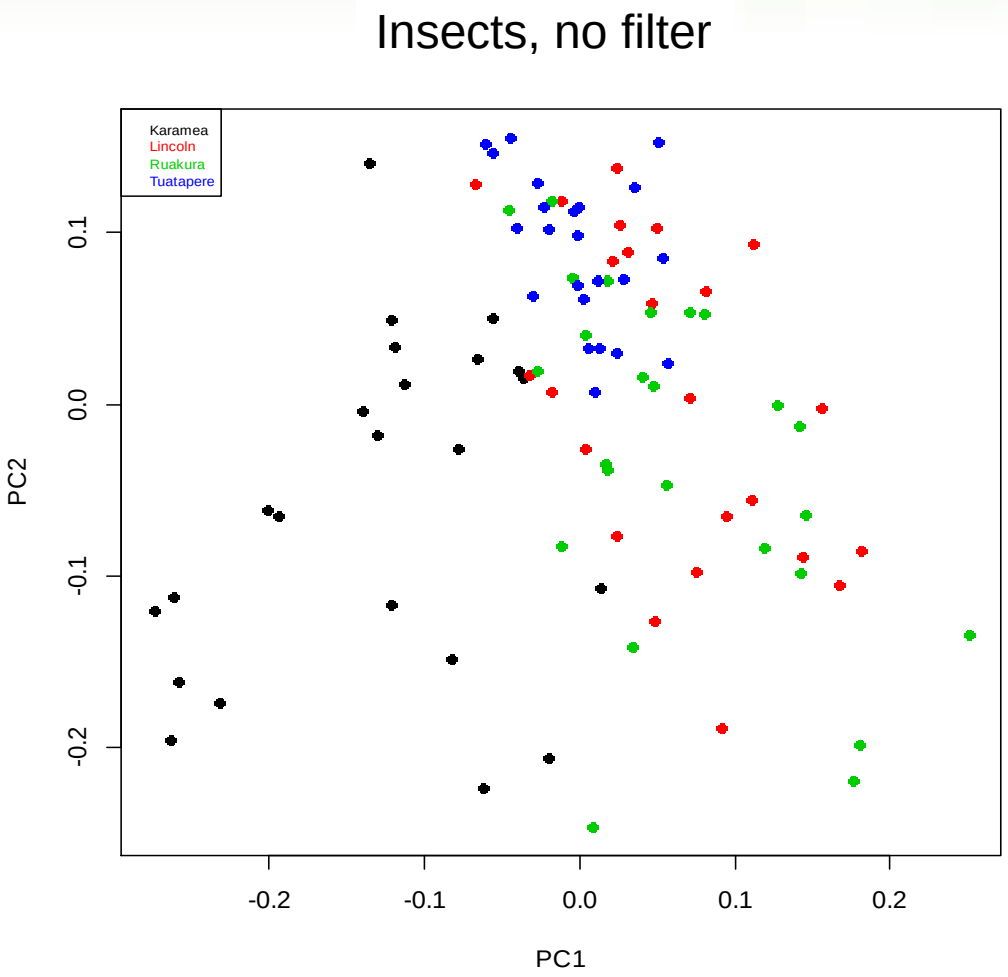
Tassel filter: tag count

	Insect	Fungi
Unfiltered	58712	48671
Filtered	653	9556
Trimmed - unfiltered	51537	44202
Trimmed - filtered	447	5254

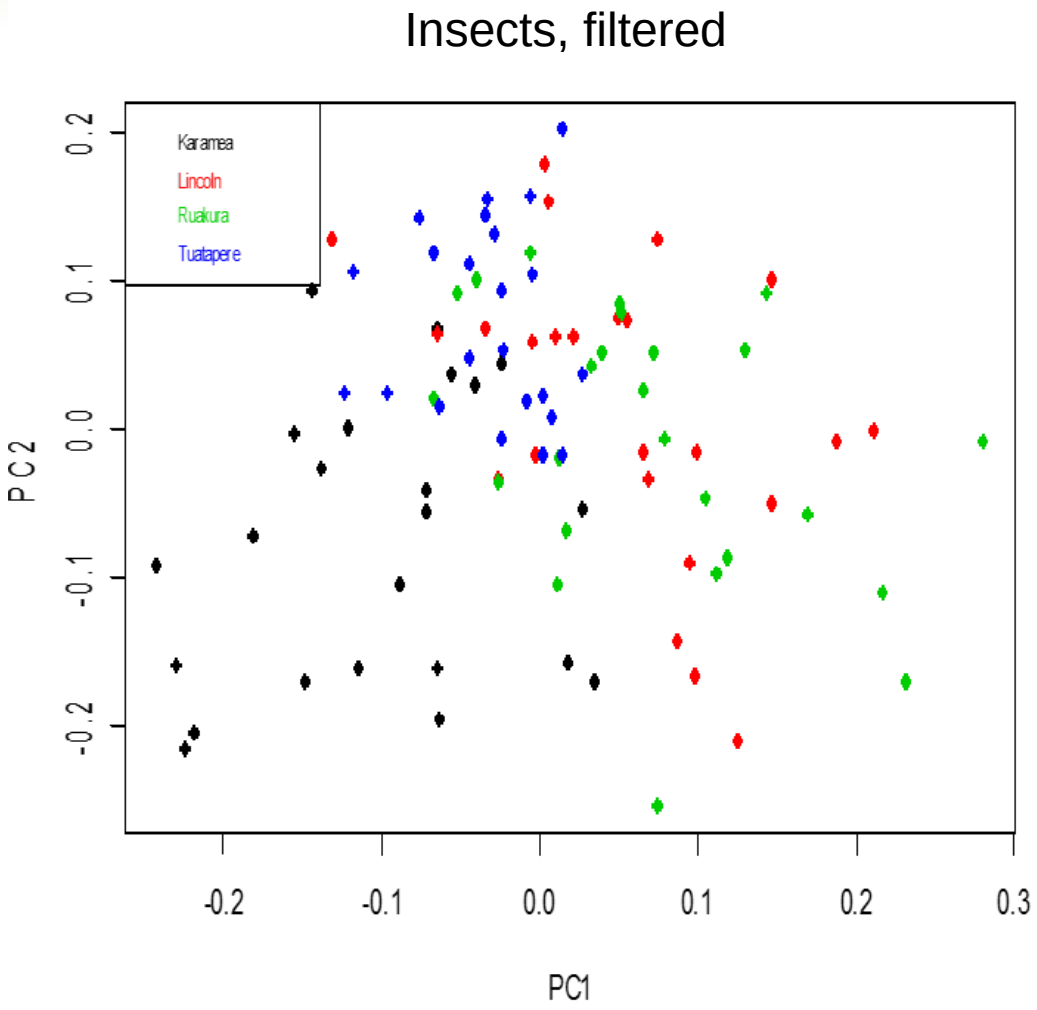
Filter: Present in 80% of the samples, Min. freq 0.1

agresearch

Sample structure via PCO



Sample structure via PCO



agresearch

Conclusions

- Data quality filter impacts end results
- This process, at the beginning of the pipeline, is source-agnostic

agresearch

Considerations

- Evaluate the effect of data quality trimming for tags mapped to a reference genome

- Establish the relationship between tag number and sequence depth (especially relevant for heterozygous organisms)

agresearch

Acknowledgements

Stephen Goldson

Jeanne Jacobs

Christine Voisey

agresearch