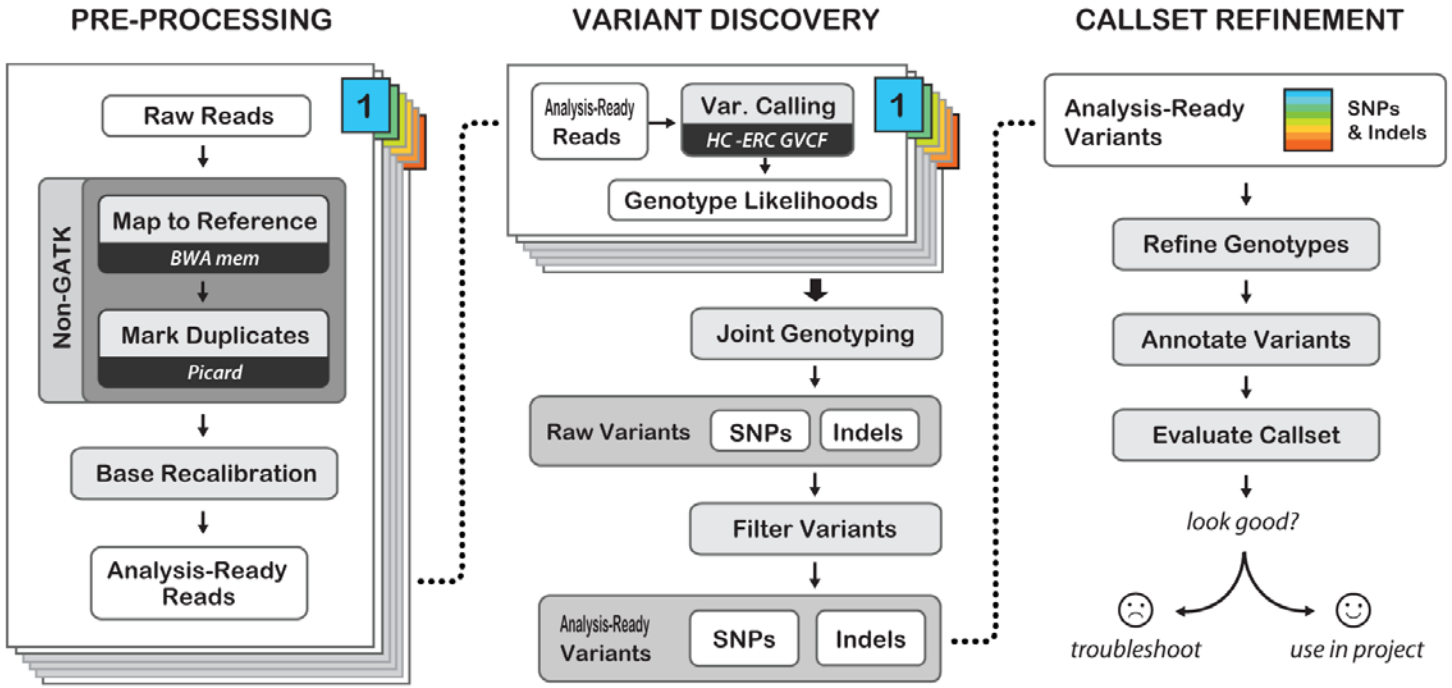
Allelic imbalance in sequence-based genotyping data

Torben Asp
Molecular Biology and Genetics
Aarhus University

Outline

- Introduction
- Allele balance analysis in mink genotyping-by-sequencing data
- Conclusions

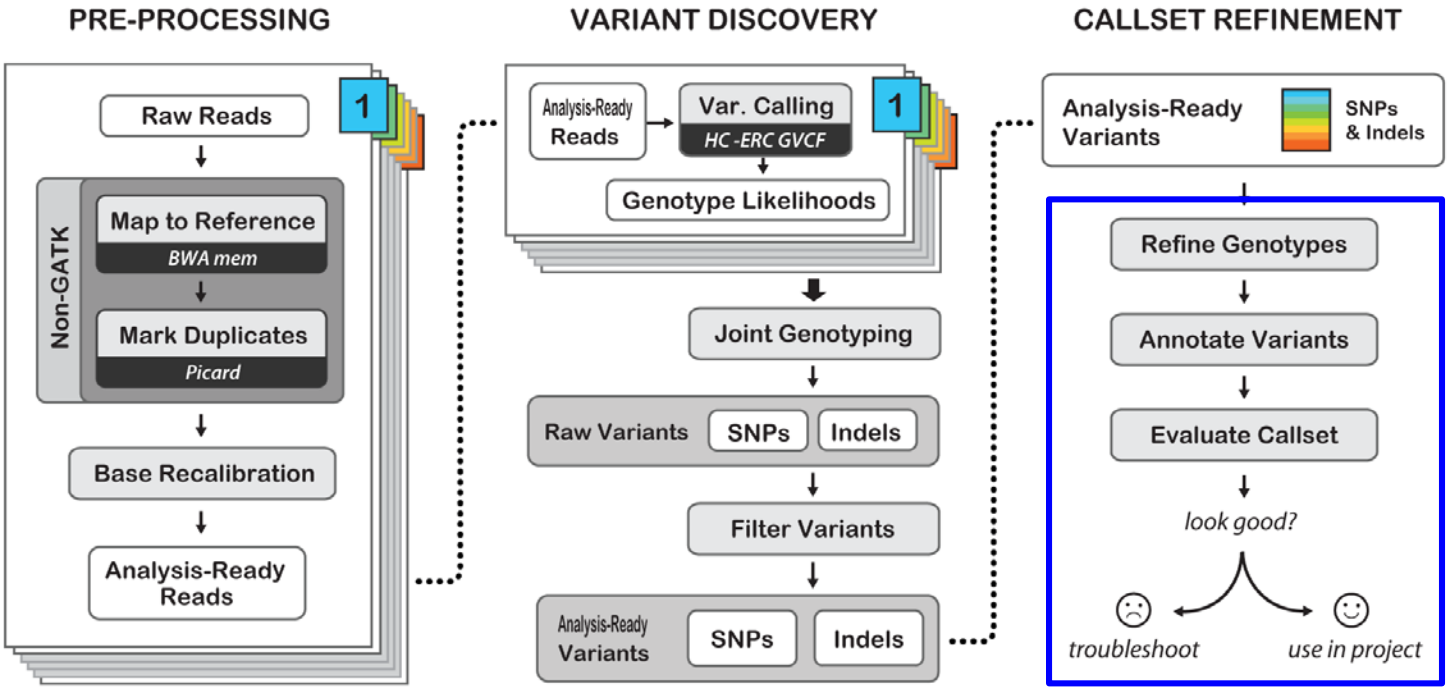
GATK variant calling workflow



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

https://software.broadinstitute.org/gatk/

GATK variant calling workflow



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

https://software.broadinstitute.org/gatk/

SNP filtering, why

- The quality control (QC) filtering of SNPs is an important step to minimize genotyping errors
- SNP QC commonly uses filters, e.g. Hardy–Weinberg equilibrium, missing proportion and minor allele frequency to remove SNPs with insufficient genotyping quality
- Implementation requires arbitrary thresholds and does not jointly consider all QC features

SNP filtering considerations

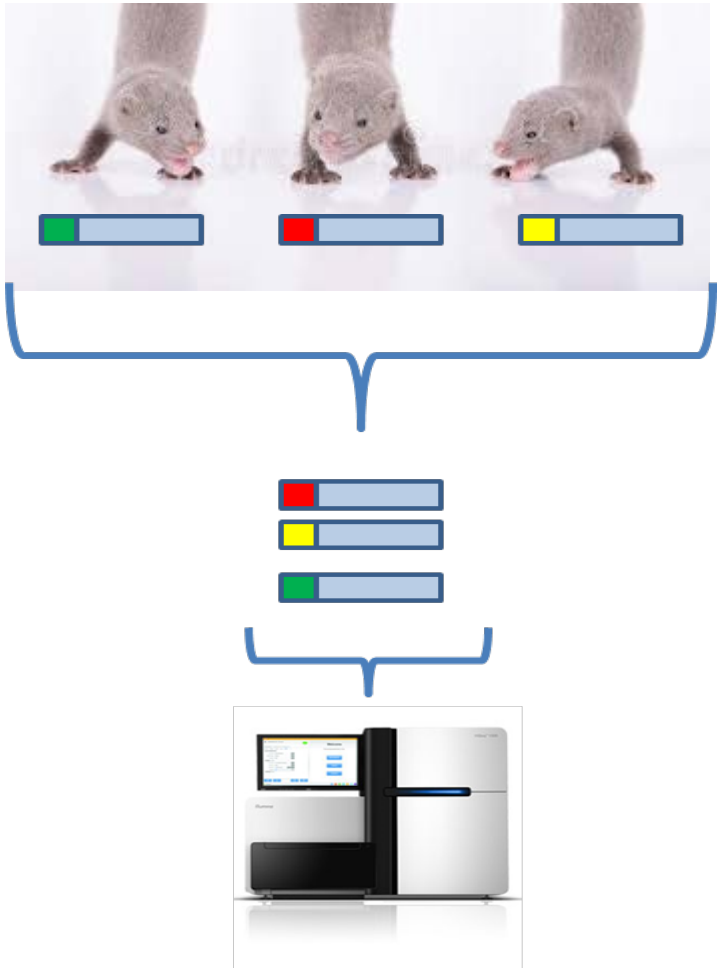
- Variant caller: methods, available info, VCF specific tags
- Sequencing technology
- Data type: DNA-Seq, Exome-Seq, RNA-Seq, GBS
- Reference genome: reliability of the reference sequence
- Genome features (Transposable Elements, Tandem Repeats)
- Available resources: reference variant sets

Sequence data and variant calling bias

- Systematic sequencing errors
  - Strand bias
  - Base Quality Rank Sum Test
- Local alignment problems
  - Distance from the end of read
  - Read Position Rank Sum
  - HaplotypeScore
- Mapping problems
  - Mapping Quality
  - CNV
- Abnormal allele balance or Quality/Depth

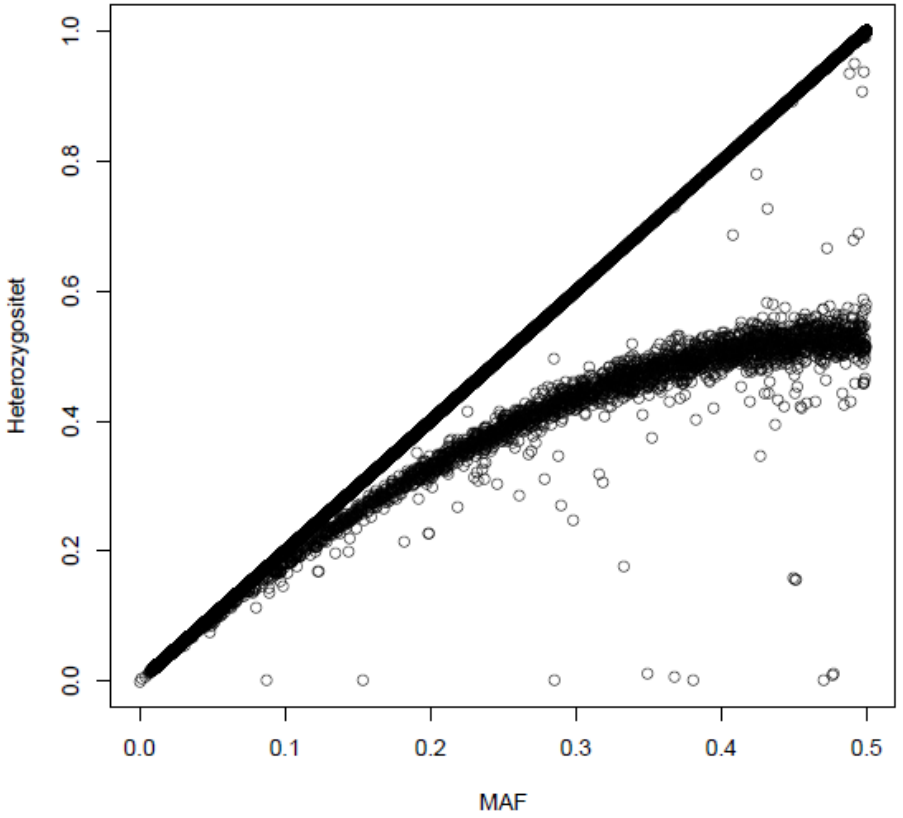
Mink genotyping-by-sequencing

- 2451 mink individuals
- Genome complexity reduction using PstI/MspI
- Alignment to the mink draft genome
- Variant calling
- SNP filtering



Mink GBS analysis – the problem

- Clearly there are two classes: the first class behaves much as expected, so He is approximately 2 pq
- The second class of points form a straight line
- The straight line reflects points where only heterozygotes and one homozygote are observed; the other homozygote is absent
- Then the allele frequency estimate is p=½He, or He=2p, where p is the allele frequency estimate



The problem…

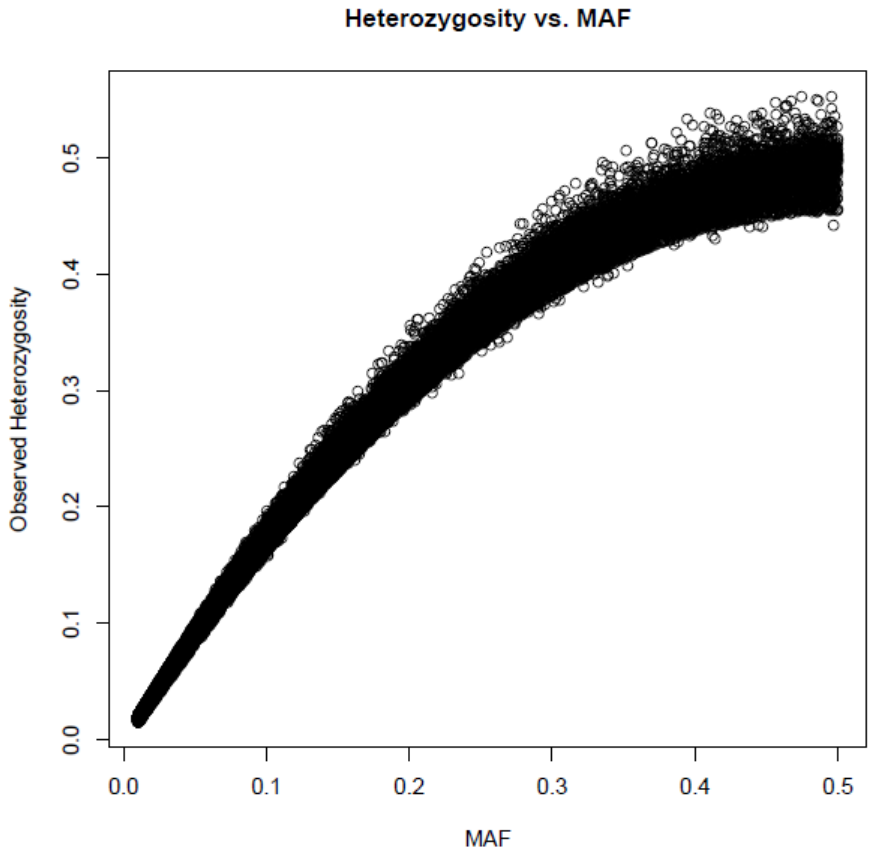
- Erroneous realignment in low-complexity regions
- CNVs
- Incomplete reference genome

Leading to allele imbalance genotyping errors (excess heterozygosity)

The solution…

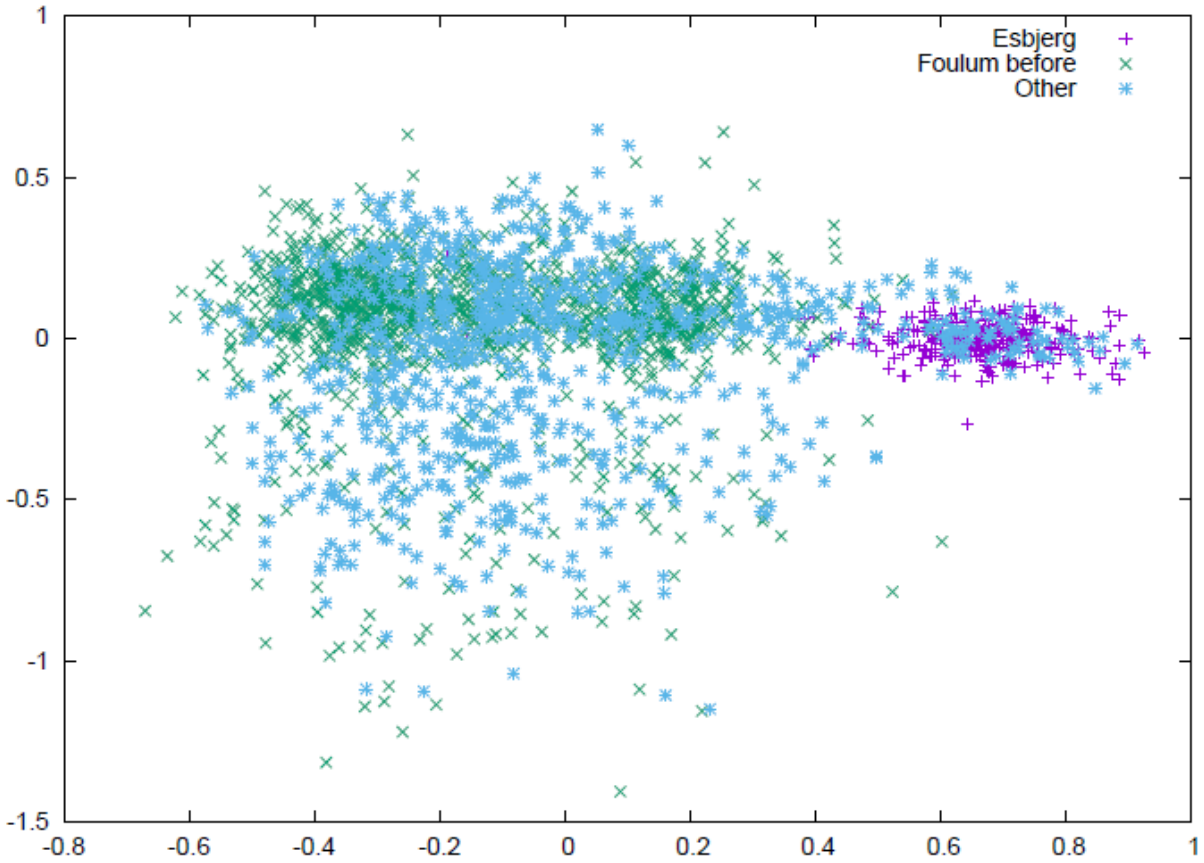
- The solution was a new SNP filtering strategy, of which allele balance (AB) filtering was the most important
- AB annotation attempts to estimate whether the data supporting a variant call fits allelic ratio expectations, or whether there might be some bias in the data

Het. vs MAF after AB filtering



Approximately 11% of all SNPs were removed

PCA plot


Esbjerg
Foulum before
Other
1
0.5
0
-0.5
-1
-1.5
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Conclusion

- We identified the erroneous realignment in low-complexity regions, uneven sequencing depth, and the incomplete reference genome with respect to the samples as the three major sources of errors
- Developed and new SNP filtering strategy – filtering for allele balance was the most important step
- All the markers with large excesses of proportions of heterozygotes have been removed, e.g. 11% (GBS) and 6% in grasses (resequencing; data not shown)
- Recommend to filter for allele balance

Acknowledgements

- Mogens Sandø Lund
- Guosheng Su
- Trine Villumsen
- Bernt Guldbrandsen