Diagnostics for genotypes from low-depth sequencing

Ken Dodds[1] Timothy Bilton[1,2]

Alan McCulloch[1] Rudi Brauning[1] Matt Schofield[2]
Mik Black[2] Rachael Ashby[1] John McEwan[1]
Shannon Clarke[1] Jeanne Jacobs[1]

[1]AgResearch, New Zealand
[2]University of Otago, Dunedin, New Zealand

ken.dodds@agresearch.co.nz

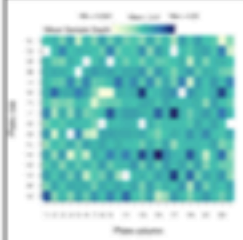
Introduction

- Sequencing is increasingly being used for genotyping. Costs can be reduced by sequencing a representative fraction of the genome at low depth.
- Low depth genotypes may be uncertain due to low read depth; statistical methods are being developed to account for this uncertainty.
- Diagnostics are important for checking the laboratory and computational workflows.
- We present some depth-aware diagnostics that should be used in conjunction with knowledge about the samples.
- These diagnostics are available in scripts in the AgResearch GitHub repository (https://github.com/AgResearch).

Aim

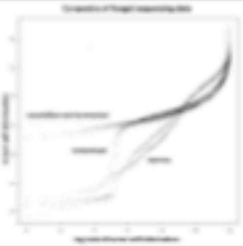
Demonstrate QC methods which account for read depth.

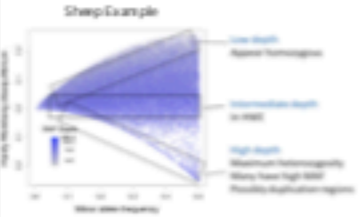
Diagnostic Tools



Graphical display of a measure of sequencing success against the physical arrangement of samples during lab processing. Here the sample depth (mean number of reads over detected SNPs) is displayed by 384-well plate layout and shows typical variation. Negative controls are marked with X. Any patterns in sequencing success may reflect sample mishandling.

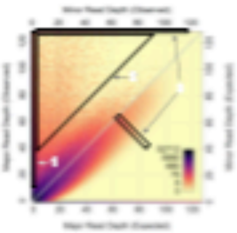
Genome sequences display characteristic signatures for species and contaminants. McCulloch *et al.* (2018) show how a kmer zipfian plot of self-information in k-mers against their rank can be used to diagnose contaminated and repetitive samples.

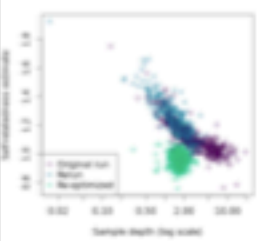




A *fin plot* (Dodds *et al.*, 2015) of Hardy-Weinberg disequilibrium (HWD) against minor allele frequency (MAF) for each SNP can help detect repeated regions erroneously assigned to the same SNP position—they have low HWD for their MAF and relatively high depth. The plot may also show that many SNPs with high apparent HWD have low depth so heterozygotes may appear homozygous.

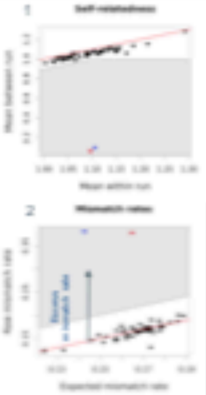
The *comet plot* (Bilton, 2020) shows the distribution of allele count pairs mirrored against that expected under binomial sampling. Example insights: 1) sequencing error rate (homozygotes), 2) extra-binomial sampling, 3) allele count limit in the software used.





Self-relatedness (1 + inbreeding) is normally unrelated to read depth. Higher estimated inbreeding at lower depth may indicate non-optimal laboratory processes. The plot shows this pattern being corrected after re-optimizing protocols such as enzyme concentrations.

Sometimes an individual is re-genotyped. Comparing the low-depth genotype calls can be problematic as one allele may not be observed. Two alternative approaches are: 1) compare relatedness estimates between the results to the mean self-relatedness (expected to be similar for the same individual) and 2) calculate the *excess mismatch rate*, the difference between the observed (raw) mismatch rate and the mismatch rate that is expected given the read depths. This is analogous to the parentage methods in Dodds *et al.* (2019). Both plots highlight the same two incorrectly labelled individuals (🔵,🔴).



Conclusion

Depth-aware diagnostics help QC-check genotypes from low depth sequencing.

References

Bilton (2020) PhD thesis, University of Otago
Dodds et al. (2015) doi: 10.1186/s12864-015-2252-3
Dodds et al. (2019) doi: 10.1534/g3.119.400501
McCulloch et al. (2018) doi: 10.1093/database/bay029

Acknowledgements

Funded by the Ministry of Business, Innovation and Employment (NZ), through the "Genomics for Production & Security in a Biological Economy" programme.