Genotype call for chromosomal deletions using read-depth from whole genome sequence variants in cattle

M. Mesbah-Uddin¹,², B. Guldbrandtsen¹, M.S. Lund¹ & G. Sahana¹

¹ Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark,
² Animal Genetics and Integrative Biology, UMR 1313 GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
mdmesbah@gmail.com (Corresponding Author)

Summary

We presented a deletion genotyping (copy-number estimation) method that leverages population-scale whole genome sequence variants data from 1K bull genomes project (1KBGP) to build reference panel for imputation. To estimate deletion-genotype likelihood, we extracted read-depth (RD) data of all the bi-allelic variants within a given deletion locus, and fitted a Gaussian mixture model to the observed RD. We validated our method on brachyspina associated deletion of chromosome 21 (Chr21:21,184,869-21,188,202), which was segregating in our deletion-discovery population of Holstein cattle. We analysed the RD data of 55 progeny tested Holstein bulls with published recessive code for brachyspina (8 carriers and 47 non-carriers) along with 5 carriers from the discovery population (confirmed by assembling the breakpoint sequences). Using our approach we were able to genotype the carriers and non-carriers with 95% accuracy, and a false discovery rate of 18.8%.

Keywords: read-depth genotyping, Gaussian mixture model, deletion, copy number variation, dairy cattle

Introduction

The rate of genetic-gains in cattle increased substantially since the introduction of genomic selection in 2008. In the US dairy cattle industry, for example, the rate of yearly gains ranges from ~50-100% for production traits and 3 to 4 fold for fitness traits (Garcia-Ruiz et al., 2016). In such selection program, it is also very important to optimize the balance between the rate of genetic-gains and inbreeding. Copy-number variations (CNVs) are a class of DNA polymorphisms that changes gene-dosages and thus could affect a trait. The functional impact of CNVs in cattle populations could be understood analysing the relationship between CNV-genotype and phenotype, such as, using genome-wide association study (GWAS) approach. While recent studies reported the identification of CNVs in many cattle breeds (Shin et al., 2014; Boussaha et al., 2015; Chen et al., 2017), the accumulation of large reference population remains a challenge for including CNVs in GWAS or genomic-estimation of breeding values (GEBVs). In this study, we presented an approach to extend reference population for imputing deletions (CNV-loss) using Bos taurus animals from 1K bull Genomes Project (1KBGP).

Material and methods
Samples

We analysed a ~3.3Kb deletion on chromosome 21 (Chr2:21:184,869-21,188,202) segregating in our deletion-discovery population of 67 Holsteins with variant allele frequency of 5.2% (Mesbah-Uddin et al., 2017). From Run-6 of 1KBGP (Daetwyler et al., 2014), we extracted whole genome sequence variants within this deletion, and retrieved read-depth (RD) data from the VCF file (DP-tag).

Deletion genotyping from read-depth

To estimate genotype (copy-number) likelihood (GLs), we modelled observed RD ($x_i \sim N(\mu_j, \text{var}_j)$) at each variant locus within a deletion assuming a linear relationship between RD and copy-number status of that locus. We fitted a Gaussian mixture model (GMM) to the observed RD. Assuming a pure deletion locus and diploid genome, we constrained GMM to fit exactly three copy-number (CN) classes, such as CN0=homozygous for deletion, CN1=hemizygous, and CN2=homozygous for reference allele (Handsaker et al., 2011). GLs will provide the probabilities of observing the RDs given the underlying genotype, and can be expressed as $p(RD_i=\mu_j | CN_j)$, where $RD_i$ is the read-depth at SNP $i$ within the deletion locus, and $CN_j$ is the (unobserved) true deletion genotype (CN=0, 1 or 2). We implemented expectation-maximization (EM) algorithm for estimating GMM parameters and learning most likely combination of models to explain the data (Figure 1). The weights ($w_j$) and variances ($\text{var}_j$) for each CN class were updated iteratively until convergence, while keeping the mean RD fixed (such as, $\mu_{CN0}=0$, $\mu_{CN1}=0.5 \times$ average genome-wide RD, $\mu_{CN2}=$ average genome-wide RD).

Figure 1. Graphical illustration of read-depth genotyping at a deletion locus using Gaussian mixture model.

Figure 2. Estimated weights of copy-number 1 for carrier (13) and non-carrier (47) animals of the ~3.3Kb deletion on chromosome 21:21,184,869-21,188,202.
Results and discussion

We were interested to know whether it is possible to infer the copy-number status of a known chromosomal deletion from the auxiliary read-depth information provided with the variant-genotypes, such as in the VCF file of 1KBGP, where final variant-calls (not raw genome-sequences) are shared among the collaborators. Furthermore, this approach could also provide a formal way to interrogate putative CNV regions with fewer computational resources. Hence, we analysed a ~3.3Kb deletion on chromosome 21 (Chr21:21,184,869-21,188,202); this is a recessively inherited genetic-defect responsible for fetal death (brachyspina) in Holstein cattle (Charlier et al., 2012). From the 1KBGP samples, we found 55 progeny tested Holstein bulls with published recessive code for brachyspina (8 carriers and 47 non-carriers) in the US Council on Dairy Cattle Breeding database (https://www.usedcb.com/CF-queries/index.cfm, last accessed on 25 September, 2017). There were also five carrier animals (confirmed by targeted breakpoint-sequence assembly) in our discovery population (Mesbah-Uddin et al., 2017), totalling to 60 animals (13 carriers and 47 non-carriers) with confirmed genotype status. We fitted a constrained Gaussian mixture model to read-depth data from each of the 60 animals, and presented the estimated weights of copy-number 1, i.e. the probability of being a carrier of brachyspina, here in Figure 2. Using a naïve threshold of 0.5 we could correctly classify all the 13 carriers of the deletion and 44 of 47 non-carriers. Thus, our approach can classify the carriers and non-carriers with an accuracy of 95% and a false discovery rate of 18.8%.

For improving the model further, adjusting the expected read-depth for biases arise from sequence features, such as GC-content and mapability of the genomic region, could be taken into consideration.

Conclusions

We presented a deletion genotyping method that leverages population-scale variant-calls from 1KBGP to build reference panel for imputation, which could be extended to duplication or copy-number gains. This will facilitate inclusion of CNVs in GWAS and genomic prediction.

Acknowledgements

Md Mesbah-Uddin benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG". This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B).

List of References


