Predicting which variants genotype well with array vs. sequence data

P.M. VanRaden[1], G.L. Spangler[1], C.P. Van Tassell[1], J. Jiang[2], L. Ma[2], S.K. DeNise[3] & J.R. O'Connell[4]

[1] USDA-ARS Animal Genomics and Improvement Laboratory, Building 5 BARC-West, Beltsville, Maryland 20705,USA
[2] University of Maryland, College Park, Maryland, USA
[3] Zoetis, Inc., Kalamazoo, Michigan, USA
[4] University of Maryland - Baltimore, 655 W Baltimore S, Baltimore, Maryland 21201, USA

paul.vanraden@ars.usda.gov *(Corresponding Author)*

Presented at the World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand, 2018.

Summary

Whole genome sequencing has identified millions of new variants, but many (about 35% in our experience) of the single nucleotide polymorphisms (SNPs) may not produce high quality genotypes from microarrays. Properties of SNPs can help predict which will pass or fail when designing arrays such as the customized version of Illumina's Bovine LD chip examined here. Genotypes for 26,970 reference bulls were imputed using 444 sequenced Holsteins from run 5 of the 1000 Bull Genomes Project, and 4,821 SNPs with largest effects for net merit were selected. When adding those to the Zoetis LD chip (version 5), the success rate was 96% for 3,220 SNPs from the Bovine HD chip, but only 64% for 1,601 new sequence SNPs not previously on any chip. To determine why SNPs failed, a pass/fail (1/0) indicator of sequence SNP conversion success was correlated with 1) Illumina design scores, 2) estimated heritabilities of the genotypes for 3,000 randomly selected bulls, and 3) the base distance that the SNP was inside a repetitive DNA segment as determined by RepeatMasker, using a minimum distance of 0 if outside a repeat and maximum of 50 bases if inside. The correlations were 0.51 for design scores, 0.14 for estimated heritabilities, and -0.15 for repeat distance. All three were highly significant ($P < 0.0001$), but repeat distance was less significant ($P = 0.04$) after fitting design score and heritability in multiple regression. Three other factors (minor allele frequency, SNP position with genes, and the reference/alternate allele combination pattern) were not associated with conversion success. In a reverse test, 56,815 SNPs from the Bovine 50K version 1 chip were matched with 38 million sequence SNPs. Previously 15,772 of the 50K SNPs had been declared not usable, and 11,969 (87%) of those were also either not identified or removed by sequence edits. However, 3,803 (9%) of the 43,053 currently used SNPs that produce high quality genotypes on the 50K chip were absent from the sequence data, and the absence was not associated with minor allele frequency or allele combination. If the goal is to select the best SNP subset for a chip, design scores could be pre-computed and examined before rather than after estimating SNP effects, allowing selection of other linked SNPs expected to perform better. Eventually targeted sequencing could provide genotypes for important SNPs that fail to convert, because many SNPs from sequence data are difficult to genotype using arrays.

Keywords: single nucleotide polymorphisms, genotyping arrays, design scores

Introduction

Genotyping with arrays vs. sequencing use different physical and chemical processes that can affect success rates for individual SNPs. Arrays (SNP chips) indicate the 3 genotypes by alleles attaching to beads, and each allele should have none, half, or all attached. Sequencing physically reads about 150 bases at both ends of a DNA segment that is about 1000 bases in length, and multiple reads are needed to detect both alleles. Some DNA patterns are easier to read than others because a single strand may loop on itself (such as a hairpin loop) or may attach to another DNA strand instead of attaching to the complementary bead. These DNA interactions can be predicted from the DNA base patterns (SantaLucia & Hicks, 2004).

Goals of this research were to understand why some SNPs can be accurately genotyped from sequence data but not on an array, or vice versa, and how this information can be used to design arrays and choose variant sets to use in selection. Early arrays for genotyping cattle usually included only high quality, polymorphic, equally spaced SNPs as markers. Newer arrays began adding preselected QTLs, not just markers, to better track biological effects. The same SNPs may not genotype as well on arrays as in sequence data because of the different chemistry used for genotyping.

Data and methods

Genotypes were imputed for 26,970 reference bulls using 444 sequenced Holsteins from run 5 of the 1000 Bull Genomes Project in a previous study (VanRaden *et al.,* 2017). Allele effects for net merit were estimated for 481,904 sequence SNPs in or near genes plus 312,614 from the Illumina BovineHD chip (HD), following edits for low minor allele frequency (MAF) or high linkage disequilibrium. After hand edits to remove excess SNPs associated with the same QTL, a subset of 4,821 SNPs with large effects was selected. Those included 1,601 new sequence SNPs not previously on any chip plus 3,220 SNPs selected from the Bovine HD chip that were added to version 5 of the Zoetis LD chip (ZL5) and also provided to other companies for inclusion on custom chips. The SNPs with demonstrated large effects from sequence were selected with no pre-screening for SNP quality.

Many of the SNPs selected from sequence failed the quality control (QC) steps for use on arrays. To determine why SNPs failed, a pass/fail (1/0) indicator of sequence SNP conversion success onto the ZL5 was correlated with 1) Illumina design scores, 2) estimated heritabilities of the genotypes for 3,000 randomly selected bulls, and 3) the base distance that the SNP was inside a repetitive DNA segment as determined by RepeatMasker, using a minimum distance of 0 if outside a repeat and maximum of 50 bases if inside. Design scores were obtained online by uploading a file containing SNPs plus flanking sequence.

Other properties tested were MAF in Holsteins, location within genes (such as exonic vs. intergenic), and reference / alternate allele combination. Heritabilities were also estimated again for the array genotypes using 5,000 random animals genotyped with ZL5 for SNPs that passed QC. This test included mostly sibs because very few parents were genotyped for these new SNPs to detect parent-progeny conflicts.

In a reverse test to determine if the array SNPs were also present in the sequence data, 56,815 SNPs from the Illumina Bovine 50K version 1 chip (Matukumalli *et al.,* 2009) were matched with 38 million sequence SNPs. The 50K SNPs were edited for Mendelian conflicts, low call rate, etc. (Wiggans *et al.,* 2009), based on tens of thousands of genotyped animals.

Results

The conversion success rate was 96% for the 3,220 SNPs selected from HD, but only 64% for the 1,601 new sequence SNPs not previously on any chip. Success for these new SNPs was correlated by 0.51 with design scores, 0.14 with estimated heritabilities, and -0.15 with repeat distance. All three correlations were highly significant ($P < 0.0001$), but repeat distance was less significant ($P = 0.04$) after fitting design score and heritability in multiple regression. The F-test values from multiple regression were 358.4 for design score, 14.3 for heritability, and 4.2 for repeat distance. The multiple correlation was only 0.53, only slightly larger than 0.51 from design score alone. Figure 1 shows that most of the SNPs that passed QC had design scores above 0.90, whereas most that failed QC had design scores below 0.60.

Three other factors (MAF, SNP position with genes, and the reference/alternate allele combination pattern) were not associated with conversion success onto the ZL5. Heritabilities from the ZL5 genotypes averaged 95.1% for the new SNPs compared to 95.8% for the previous SNPs from HD. Those do not differ much, but are both lower than for sequence, perhaps because so few parent genotypes were available.

Array SNPs were not always in the sequence data in the reverse test. Previously 15,772 of the 50K SNPs had been declared not usable, and 11,969 (87%) of those were also either not identified or removed by sequence edits. However, 3,803 (9%) of the 43,053 currently used SNPs that produce high quality genotypes on the 50K chip were absent from the sequence data, and the absence was not associated with minor allele frequency or allele combination. Discovery of true QTLs from sequence does not guarantee quality genotypes from chips.

If two SNPs are highly correlated with similar effect sizes, choose the SNP with best design score (and heritability).

Conclusions

Whole genome sequencing has identified millions of new variants, but many (about 35% in our experience) of the single nucleotide polymorphisms (SNPs) may not produce high quality genotypes from microarrays, and about 9% of usable chip SNPs were not in sequence data. Design scores were very helpful to predict which SNPs will pass or fail when designing arrays, whereas heritability & repetitive DNA location somewhat helpful. Gene location, allele pattern, and MAF were not helpful. If the goal is to select the best SNP subset for a chip, design scores could be pre-computed and examined before rather than after estimating SNP effects, allowing selection of other linked SNPs expected to perform better. Arrays are excellent for tracking marker SNPs, but some true QTLs may require targeted sequencing to provide genotypes for important SNPs that fail to convert, because many SNPs from sequence data are difficult to genotype using arrays.

Acknowledgments

The authors thank the Council on Dairy Cattle Breeding for genotypic, phenotypic, and pedigree data, the 1000 Bull Genotypes Project for sequence genotypes, and USDA-ARS project 1265-31000-101-00, “Improving Genetic Predictions in Dairy Animals Using Phenotypic and Genomic Information” for funding support.

percentage

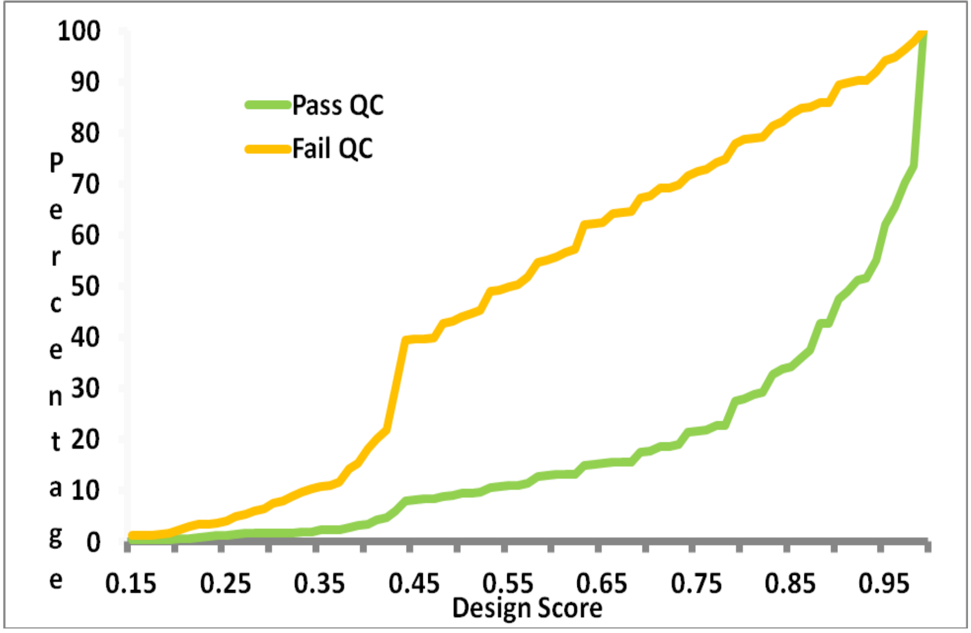

Figure 1. Cumulative frequency of design scores by pass/fail status.

List of References

Matukumalli, L.K., C.T. Lawley, R.D. Schnabel, J.F. Taylor, M.F. Allan, M.P. Heaton, J. O'Connell, S.S. Moore, T.P.L. Smith, T.S. Sonstegard & C.P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4:e5350.

SantaLucia, J., and D. Hicks. 2004. The thermodynamics of DNA structural motifs. Ann. Rev. Biophys. Biomol. Struct. 33:415–40.

VanRaden, P.M., M.E. Tooker, J.R. O'Connell, J.B. Cole, & D.M. Bickhart, 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet. Sel. Evol. 49:32.

Wiggans, G.R., T.S. Sonstegard, P.M. VanRaden, L.K. Matukumalli, R.D. Schnabel, J.F. Taylor, F.S. Schenkel, & C.P. Van Tassell, 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92:3431–3436.