Experience from large scale use of the EuroGenomics custom SNP chip in cattle

Didier Boichard, Mekki Boussaha, Aurelien Capitan, Dominique Rocha, Chris Hoze, Marie-Pierre Sanchez, Thierry Tribout, Rabia Letaief, Pascal Croiseau, Cécile Grohs, et al.

Experience from large scale use of the EuroGenomics custom SNP chip in cattle

D. Boichard[1], M. Boussaha[1], A. Capitan[1,2], D. Rocha[1], C. Hozé [1,2], M.P. Sanchez[1], T. Tribout[1], R. Letaief[1], P. Croiseau[1], Grohs C. [1], W. Li[3], C. Harland[3], C. Charlier[3], M.S. Lund[4], G. Sahana[4], M. Georges[3], S. Barbier[5], W. Coppieters[3], S. Fritz[1,2], & B. Guldbrandtsen[4]

[1] GABI, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy-en-Josas, France
[2] Allice, 75012 Paris, France
[3] GIGA-R & Faculty of Veterinary Medicine, University of Liège, 4000 Liège, Belgium
[4]Aarhus University, Center for Quantitative Genetics and Genomics, 8830 Tjele, Denmark
[5] Valogene, 75012 Paris, France
didier.boichard@inra.fr *(Corresponding Author)*

Summary

This article presents the strategy to evaluate candidate mutations underlying QTL or responsible for genetic defects, based upon the design and large-scale use of the Eurogenomics custom SNP chip set up for bovine genomic selection. Some variants under study originated from mapping genetic defects, embryonic lethals, or QTL by GWAS. Other variants originated from a reverse genetics approach and were selected according to their annotation. Because of the limitation in chip size, these variants were severely selected. For instance, structural variants were required to affect a gene. Loss-of-function variants were preferred to deleterious non synonymous variants. An incremental process was used, with one or two chip versions each year, the less informative variants being replaced by new candidates. Many examples are presented. Expected results are: confirmation of effects in large independent populations; estimation of allelic frequencies in different breeds; accumulation of individual genotypes. When confirmed, a variant can be used in a straightforward way by the industry in breeding programs by switching its status on the chip.

Keywords: bovine, custom chip, causal variant, genetic defect, GWAS analysis, confirmation study

Introduction

Eurogenomics is a consortium created in 2009 and gathering initially the cattle breeding industry from France, the Netherlands, Germany, Denmark, Sweden, Finland, Poland and Spain. It was also closely associated with academic teams from Liege, Aarhus and Wroclav Universities, INRA, VIT, and INIA. The initial goals of Eurogenomics were data exchange to enlarge reference populations in Holstein (Lund et al, 2011), common design and purchase of large quantities of SNP chips for genomic selection (GS), and different R&D projects. This paper aims at presenting 5 years of experience in using the so-called EuroG10k custom chip.

Structure of the EuroG10k chip

The EuroG10k chip was based on the low density BovineLD BeadChip of Illumina (Boichard et al, 2012) and includes an add-on, the content of which evolves over time. The chip was composed of a common part and several private parts. In addition to the basic LD content of

6,909 SNPs, the common part included variants from the 50k or HD Illumina chip to improve imputation accuracy close to the end of the chromosomes, to fill some remaining gaps, and to impute microsatellites for parentage testing. It also included highly predictive variants in genomic evaluation. The functional variants of the common part were defined from the literature. They initially included 125 tests corresponding to genetic defects or genes of interest. To avoid any variant loss or genotyping errors, important commercial tests were duplicated.

The chip is physically the same for all partners and all their genotyping laboratories, but the Illumina manifest file used to decode fluorescence signals differs between laboratories, allowing customized content of the chip according to the partner. The name EuroG10k relates to the size of the common part, however the whole chip is larger.

Origin of the research variants

The variants were chosen on the basis of different arguments: (a) candidate variants for genetic defects or embryonic lethals; (b) variants within peaks of GWAS analyses; (c) variants with a strongly deleterious annotation, as predicted by VEP (McLaren et al, 2016); (d) variants believed to have a regulatory function; (e) boundaries of structural variants affecting some genes (designed as shown in Figure 1). Because the number of slots was limited, only a small proportion of the candidate variants were included. This resulted in an increasingly severe selection of the most appealing variants over time. The use of successive versions imposed a tight workplan for the research teams (Figure 2). An annual update of the chip was preferable to allow for a proper analysis of the results. In 2017, the teams used its 7th update.

Uses of the EuroG10k chip in R&D projects

Confirmation of genetic defects and estimation of allelic frequencies

As a genetic defect is identified from a limited number of cases, it is important to evaluate the effect of the candidate variant on a larger set of animals. For instance, for a recessive defect, homozygous unaffected individuals lead to exclusion of the candidate variant.

A direct test can also be very useful to implement an experimental study for an accurate characterization of the different genotypes for a genetic defect: detailed phenotyping can be planned on targeted animals of the different genotypes. The whole range of phenotypes can be observed and not only the most obvious or extreme. Bourneuf et al. (2017) presented the effect of a dominant mutation in *CHD7* gene responsible for the CHARGE syndrome and showed from a large progeny group that the severity of the phenotype was influenced by modifier genes.

Large-scale genotyping also provides allelic frequency estimates in the breed of discovery but also in a broad range of breeds. As an example, Michot et al (2016) showed that the *RP1* mutation (responsible for progressive retinal degeneration) was very frequent in the Normande breed (27%), but also present at lower frequency in other European breeds.

It also provides an official status for the most important bulls, to inform the breeders and orientate matings. When the information is delivered to the industry, many animals have already been genotyped and the test is readily available. Genotyping can be used to develop haplotype tests or imputation, to predict the status of animals genotyped with older chip versions.

Confirmation of embryonic lethals

The situation is similar for embryonic lethals regarding the confirmation studies, the estimation of allelic frequencies, the experiments for detailed phenotyping, and the assessment of the status of the animals as well as their management in the population. Very large numbers of genotypes are required for embryonic lethals when their allelic frequency is low and therefore the expected number of homozygotes is very low. As the confirmation study relies on the absence of homozygotes, presence of normal homozygotes invalidates the candidate variant. This situation occurred for MH1 embryonic lethal where live homozygous animals for the *SHBG* variant were observed after a recombination occurred in the carrier haplotype. A follow-up analysis led to validate a variant in the *PFAS* gene (Michot et al, 2017).

Confirmation of GWAS peaks

A popular approach consists in performing GWAS analyses on large populations with genotypes imputed up to the sequence level (*e.g.,* Kadri et al, 2016; Sanchez et al, 2017). Combined with functional annotation, this approach produce a limited number of candidate variants per QTL. Therefore, in a short time, it can produce many candidate variants which need to be validated. After these candidate variants have been put on the chip, many genotypes are obtained in a few months. However, most animals genotyped for GS purpose are young and cannot be used immediately for confirmation studies. But these animals can be used to impute older animals with phenotypes. In France, around 200,000 dairy genotyped cows have phenotypic information. Backward imputation is not usual but practical results showed a very high accuracy. The best accuracy was obtained ignoring the pedigree (Sanchez et al, 2018). An example is described in Sanchez et al. (2018). We can also assess the stability of variant effects across populations, considering possible epistatic interactions with the genetic background.

A reverse genetics approach

The previous strategy is limited by the number of candidate variants that can be proposed, especially for genetic defects and embryonic lethals. A higher-throughput approach is to select variants only on the basis of their annotation. A difficulty is the large number of variants with an appealing annotation, much higher than the chip capacity. A combination of criteria were used to increase the chance of a phenotypic effect. Structural variants were required to affect an important gene. Loss of function variants were preferred to deleterious variants. Genes reported in OMIM or MGI were prioritized. Charlier et al. (2016) estimated that 15% of the loss-of-function and 6% of the deleterious missense variants would be responsible for embryonic loss. Michot et al. (2016) proposed over 1,000 such variants. A few of them have been studied, such as in *RP1* (vision deficiency) or *CAD* genes (embryonic mortality), but the majority of them remain to be investigated. Finally, these variants can also be studied by GWAS of common traits such as production, reproduction, disease resistance or conformation.

In reverse genetics, many variants are detected from sequencing projects. Because many interesting variants are rare and often breed-specific, they are detected on a limited number of sequenced individuals. The first result from large-scale genotyping is the confirmation of their existence and the estimation of the allelic frequency in different breeds. This is especially critical for structural variants which are detected with a rather high false-positive rate (Boussaha et al., 2015).

Difficulties encountered with this strategy

This strategy, although very efficient on average, presents some difficulties. Once a candidate variant is identified, it takes 8 months on average to obtain a large number of genotypes. If results are needed more quickly, another approach must be chosen. For breeds with limited genotyping work, the approach is not as efficient. 10-20% of the variants are lost during chip manufacture or due to low call rates. Therefore the strategy, efficient on average, may fail for individual important variants. Many variants have low MAF. The standard clustering procedure may be inappropriate, especially when one genotype is rare or absent, and may require a manual clustering. Last, the strategy with private research parts of the chip preserves the confidentiality of the research work but does not prevent redundancy causing some loss of efficiency. This problem is overcome if partners accept to share information.

Conclusion

Widespread use of a customized SNP array for GS has simultaneously provided researchers with the opportunity to examine the effect of candidate variants across a range of cattle populations in a cost-effective fashion within a moderate time span. It allows validation of candidates and population-level application of reverse-genetics and it provides reliable estimates allele frequencies. Once the variants have been confirmed, genotypes for current breeding animals are immediately available for breed management.

Acknowledgments
We thank André Eggen (Illumina) for his help for the custom part of the EuroG10k chip.

List of References

Boichard D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, et al, 2012. Design of a Bovine Low-Density SNP Array Optimized for Imputation. Plos One 7: e34130.

Bourneuf E., P. Otz, H. Pausch, V. Jagannathan, P. Michot, C. Grohs, et al. 2016. Rapid discovery of de novo deleterious mutations in cattle using genome sequence data: Enhancing the value of farm animals as model species. Sci. Rep. 7: 11466.

Boussaha M., D. Esquerré, J. Barbieri, A. Djari, A. Pinton, R. Letaief et al. 2015. Genome-wide study of structural variants in bovine Holstein, Montbeliarde and Normande dairy breeds. Plos One 10: e0135931.

Charlier C., W.Li, C. Harland, M. Littlejohn, W. Coppieters, F. Creagh, et al. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Res. 26:1333-1341.

Kadri N.K., C. Harland, P. Faux, N. Cambisano, L. Karim, W. Coppieters, et al. 2016. Sequence-based association analysis identifies coding and non coding variants in *HFM1, MLH3, MSH4, MSH5, RNF212* and *RNF212B* with large effects on male and female recombination rate in cattle. Genome Res. 26: 1323-1332.

Lund M.S., A.P.W. de Roos, A.G. de Vries, T. Druet, V. Ducrocq, S. Fritz, et al. 2011. Common reference of four European Holstein populations increase reliability of genomic predictions. Genet. Sel. Evol. 43: 43.

McLaren W., L. Gil, S.E. Hunt, H.S. Riat, G.R. Ritchie, A. Thormann. 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17:122.

Michot P., S. Chahory, A. Marete, C. Grohs, D. Dagios, E. Donzel, et al. 2016. A reverse genetic approach identifies an ancient frameshift mutation in *RP1* causing recessive progressive retinal degeneration in European cattle breeds. Genet. Sel. Evol. 48: 56.

Michot P., S. Fritz, A. Barbat, M. Boussaha, M.C. Deloche, C. Grohs, et al. A missense mutation in *PFAS* is likely causal for embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle. J. Dairy Sci. 100: 8176–8187.

Sanchez M.P., A. Govignon-Gion, P. Croiseau, S. Fritz, C. Hoze, et al. 2017. Within-breed and multi-breed GWAS on imputed whole genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet. Sel. Evol. 49: 68.

Sanchez M.P., V. Wolf, M. El Jabri, E. Beuvier, Y. Gauzère, S. Minery, et al. 2018. Validation of candidate causative variants on fine milk composition and cheese-making properties in Montbéliarde cows. This meeting.

Figure 1. Design of the molecular test of a structural variant (example of an insertion)

Using the black arrow as a primer, the G allele reveals the insertion and the A allele the absence of insertion. A confirmation can be obtained with a second test on the other side of the insertion (red arrows as a primers). Allele C reveals the insertion and T the absence of insertion.

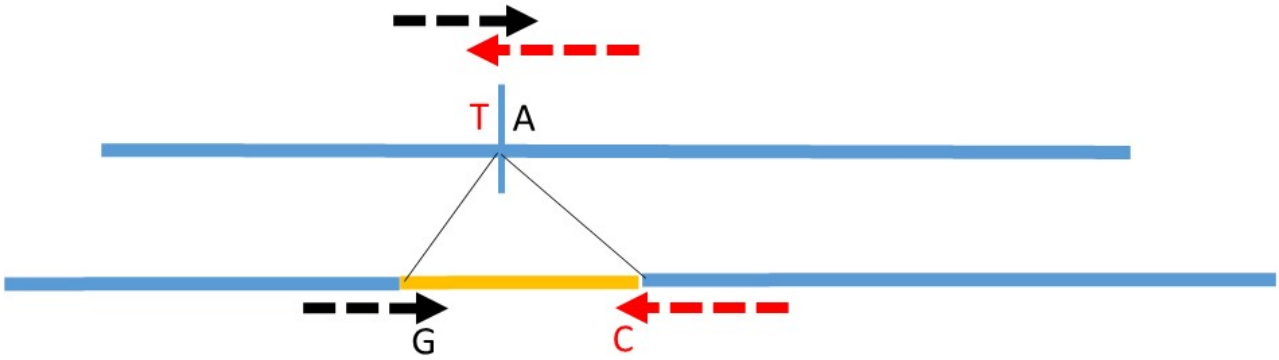

Figure 2. Workplan for chip design and use