Applying Runs of Homozygosity to the Detection of Associations Between Genotype and Phenotype in Farm Animals

F. Biscarini,[1] S. Biffani,[2] E.L. Nicolazzi,[1] N. Morandi,[1] and A. Stella[2]
[1]CeRSA-Parco Tecnologico Padano (PTP), Lodi, Italy; [2]IBBA-CNR, Lodi, Italy

ABSTRACT: Runs of homozygosity (ROH) are contiguous stretches of homozygous genotypes which likely reflect transmission from common ancestors and can be used to track the inheritance of haplotypes of interest. In this paper, ROH were extracted from 50 K SNPs and used to detect genomic regions associated with susceptibility to diseases in 468 Holstein-Friesian cows. Diagnosed diseases were categorised as infectious, metabolic, reproductive, locomotive diseases and mastitis. ROH associated with infectious and locomotive diseases and mastitis were found on BTA 12. A region of homozygosity linked with reproductive diseases was detected on BTA 15. ROH associated with infectious, metabolic and reproductive diseases, and mastitis were observed on chromosomes 3, 5, 7, 13 and 18. Previous studies reported QTLs for milk production traits on these regions, thus substantiating the known negative relationship between selection for milk production and health in dairy cattle.

Key words: dairy cattle; ROH; association study.

INTRODUCTION

Runs of homozygosity (ROH) are contiguous stretches of homozygous genotypes which likely reflect transmission from a common ancestor, and can therefore be considered as being IBD (identical by descent). ROH have been used to estimate inbreeding both in humans (McQuillan et al. (2008)) and cattle (e.g. Purfield et al. (2012)). Hildebrandt et al. (2009) applied a similar concept to map recessive disease genes in human populations, and Biscarini et al. (2013) used ROH to identify causal mutations for arthrogryposis and macroglossia in Piedmontese cattle. ROH may be an alternative approach to genome-wide scans for signals of genotype-phenotype association.

Hypothesizing that genetic variants associated with increased risk of disease are more likely to be recessive than dominant, looking for associations between disease status (case/control) and homozygous segments of the genome appears to be a reasonable strategy. In this study, ROH were applied to the detection of genomic regions associated with susceptibility to 5 categories of diseases (infectious, metabolic and reproductive diseases, mastitis and locomotive disorders) in dairy cattle.

MATERIALS AND METHODS

Available data. A population of 468 Holstein-Friesian cows between the first and fifth lactation distributed over 4 herds from the Po Valley region in Northern Italy was analysed. Cows were farmed under intensive conditions in high-yielding dairy farms and were genotyped with the Illumina BovineSNP 50 beadchip version 2 (50k), mapped on the UMD 3.1 reference assembly of the *Bos taurus* genome.

Genotypic data were edited for individual and SNP call rate (> 90%). Unassigned SNPs were removed, while those on the sex chromosome were used. This left 458 cows and 53457 SNPs available for the analysis. Phenotype recording was carried out by veterinary practitioners within the framework of the regional project Prozoo (Biscarini et al. (2012)): diagnosis, onset and treatment for each disease were recorded. Given the limited number of cases for each specific disease, diseases were grouped together in five homogeneous categories in order to increase the statistical power of the analysis: infectious diseases, metabolic syndromes, reproductive diseases, mastitis and locomotive disorders. Categories were partially overlapping (for instance, infectious mastitis was classified both as mastitis and infectious disease). Table 1 reports the classification of diagnosed diseases into the above mentioned categories. There were 189, 152, 117, 163 and 74 cases respectively for infectious diseases, metabolic syndromes, mastitis, reproductive diseases and locomotive disorders. For each analysis, all animals not in the disease group were used as controls (e.g. for infectious diseases there were 189 cases and 458 - 189 = 269 controls).

Runs of homozygosity. Under the hypothesis that complex diseases have a genetic component made up of several recessive variants distributed throughout the genome, each with a small effect (McQuillan et al. (2008)), runs of homozygosity (ROH) were applied to detect genetic regions associated with susceptibility to diseases. Single-SNP GWAS, which compares allele frequency at each locus, would in fact detect also an excess of dominant alleles. ROH are defined in diploid organisms as contiguous stretches of homozygous genotypes, which reflect transmission of identical haplotypes from common ancestors. Instead of focusing on a single locus, ROH consider also the surrounding regions, thus accounting for the hitch-hiking effect. The observed homozygosity was therefore estimated at each SNP locus and averaged along a sliding window spanning 1000 kbps and progressing SNP by SNP. A maximum of 5 missing genotypes and 1 heterozygous genotypes were permitted for a contiguous stretch of DNA to be considered a ROH.

Software. The software PLINK v1.07 (Purcell et al. (2007)) was used for the analysis. No restriction on the minimum number of SNPs in a ROH was applied, and the default (1000 kbps) maximum gap between consecutive SNPs was used, in order to account for the lower SNP density in the 50k SNP chip compared to the HD (~777k) SNP chip. Data preparation, graphical plots and post-processing analyses were produced within the open source programming environment R (R Core Team, 2013).

Table 1. list of diagnosed diseases falling in each of five (partially overlapping) categories: infectious, metabolic reproductive, locomotive diseases and mastitis.

Disease category	*Included diseases*
Infectious diseases	mastitis, peritonitis, enteritis, traumatic reticuloperitonitis, digital dermatitis, interdigital dermatitis, foot rot, laminitis, pyometra, metritis, endometritis, clostridiosis
Metabolic syndromes	milk fever, ovarian cysts, persistent corpus luteum, ruminal atony, displaced abomasum, indigestion, mesenteric torsion, volvulus, ketosis, steatosis
Reproductive diseases	retained placenta, dystocia, ovarian cysts, hypoplastic ovaries, hypotrophic ovaries, persistent corpus luteum, abortion, embryo resorption, metritis, mummified fetus, stillbirth, endometritis, parauterine abscess, pyometra
Mastitis	mastitis, teat obstruction, teat lesions
Locomotive disorders	digital dermatitis, interdigital dermatitis, sole ulcer, foot rot, white line disease, laminitis, tyloma, carpal arthritis, femoral fracture

Table 2. subset of relevant ROH associated with disease traits (start and end positions in bps).

Disease group	BTA	start	end	#SNP
Infectious	7	21704630	22305419	13
Infectious	7	23756162	24014128	6
Infectious	7	24233634	24330857	3
Infectious	12	11414743	13005009	43
Infectious	12	25878820	30099199	84
Infectious	12	51834816	52573538	21
Metabolic	3	1737421	6813316	103
Mastitis	12	25878820	30099199	84
Mastitis	13	19816277	20687500	22
Reproduction	5	6656617	6976839	11
Reproduction	5	7832521	8063248	9
Reproduction	15	40889434	41169953	4
Reproduction	15	41780731	41971248	7
Reproduction	15	42715966	42758371	2
Reproduction	18	23224334	23253048	2
Locomotion	12	11414743	13005009	43

RESULTS AND DISCUSSION

Detected regions. A total of 1,273 distinct ROH were detected in the analysed population. The average ROH length was 285.7 kbps. In principle, stretches of homozygous DNA can appear in all animals, regardless of their health status; those associated with susceptibility to disease are however supposed to be more frequent in cases than in controls. Therefore, the longest ROH (less likely to be due to chance) that were more frequent in cases than controls were retained as results of interest. A subset of the results is reported in Table 2: their frequency in cases relative to controls ranged from 51.4% to 100%, and their average p-value and FDR were 0.398 and 0.612, respectively. A few results of interest are hereby described.

On BTA 12, three distinct regions were associated with infectious diseases. The first two of these regions were also found in cows with, respectively, locomotive disorders and mastitis. Both conditions often have a microbial etiology (e.g. infectious mastitis, foot rot). Interestingly, the ROH region at ~12 Mbps on BTA 12 contains the VWA8 gene (von Willebrand factor A domain), whose mutations might be implied in coagulation abnormalities (Sullivan et al. (1994)) and may be related with musculoskeletal disorders. Figure 1 shows the average observed heterozygosity in cases and controls for reproductive diseases along BTA 15: ROH are visible as regions of low heterozygosity (and conversely high homozygosity) in cases (black line) as compared to controls (grey line). In this study we identified genetic associations with infectious, metabolic and reproductive diseases, and mastitis on chromosomes 3, 5, 7, 13 and 18 where earlier works reported QTLs for milk production traits (Cole et al. (2011); Minozzi et al. (2013)): this confirms the known negative relationship between selection for milk production and health in dairy cattle.

Statistical power. An important issue in association studies is the expected statistical power of the analysis. This is known to depend on sample size, the heritability of the trait and linkage disequilibrium (LD) between markers and QTLs (Luo (1998)). Heritabilities for the diseases traits included in this study were estimated with a sire threshold model for binary traits fitting a random sire effect and different sets of systematic effects (herd, herd + calving_month, herd + calving_month + age). Heritability was not always estimable for all traits and models; when estimable, it ranged between 0.05 and 0.10. The average LD between adjacent markers in the available Holstein-Friesian population was estimated as r^2=0.23. With 459 individuals, the power to detect a QTL explaining 1% of the phenotypic variance would be therefore about 18%.

Testing for significance. ROH are basically a model-free statistical technique. Unlike classic GWAS, there is no direct modeling of the phenotype of interest nor a straightforward way to test for the strength of detected associations. However, approaches can be conceived to assess the significance of the detected signals. In this experimental application of ROH to association studies, the significance of phenotype/genotype associations was tested by looking at the difference in homozygosity between cases and controls. The average homozygosity at each SNP locus within the ROH was computed for cases and controls separately, and the significance of the difference tested with a one-tailed t-test ($H_0: \mu_{cases} = \mu_{controls}$; $H_1: \mu_{cases} >$

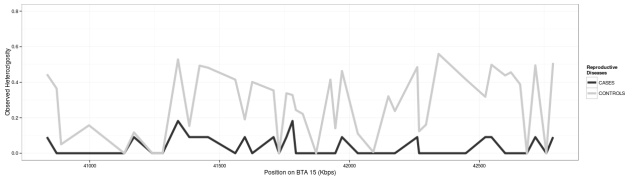

Figure 1. Observed heterozygosity in cases (black line) and controls (grey line) for reproductive diseases along BTA 15.

$\mu_{controls}$). Alternatively, the different frequency of ROH in cases and controls could be tested instead of the average homozygosity. With many loci, the issue of multiple testing may arise. This could be handled by permuting cases and controls several times in order to create a basis distribution of ROH under the null hypothesis of no association between genotype and phenotype against which the original ROH distribution can be benchmarked.

Non-inferiority. Drawing inspiration from the clinical concept of non-inferiority trials (D'Agostino et al. (2002)), the performance of ROH relative to standard GWAS in association studies can be evaluated. The false discovery rate (FDR) was computed for both ROH associations and single-SNP GWAS. The null hypothesis was that the FDR was actually larger for ROH than GWAS (H_0: $FDR_{ROH} - FDR_{GWAS} > M$, M: tolerance margin; ROH inferior to GWAS), and the alternative hypothesis that the FDR was equal for both methods (H_0: $FDR_{ROH} - FDR_{GWAS} < M$; ROH non-inferior to GWAS). Non-inferiority was tested for the 5 traits analysed (infectious, metabolic, reproductive, locomotion diseases and mastitis) with a M=0.01. In all cases H_0 was rejected, indicating that the ROH approach is not inferior to standard GWAS (average p-value: $2.27 \cdot 10^{-6}$).

Systematic effects. Population structure (e.g. pedigree) and systematic effects (e.g. herd, calving date) are known to affect the results of association studies and may lead to the detection of false positive signals. Systematic effects are traditionally accounted for by including them in the statistical model of analysis. In ROH analysis there is no explicit statistical model and accounting for systematic effects is not as straightforward. However, some approaches could be conceived. Following Pollott (2012) and Biscarini et al. (2013), the individual per-chromosome homozygosity could be modeled as the sum of chromosome, status (case/control) and interaction between chromosome and status. A significant interaction effect suggests the presence of mutations associated with the phenotype on that chromosome. Any number of systematic effects could be potentially added to such a model, and ROH analysis would then be restricted to those chromosomes still showing a significant chromosome x status interaction. Alternatively, residuals instead of the original observations could be used: phenotypes corrected for systematic effects would then be used in the ROH analysis. Another possibility would be to perform the ROH analysis within class of systematic effects and retain as valid results only ROH consistent across classes. The latter approach could be used also to deal with selection bias due to culling of animals for health or productive reasons (e.g. ROH analysis within parity/age class).

CONCLUSION

The analysis of ROH appear to be a valid alternative or complement to standard GWAS association studies. Here we presented an application to a case/control observational study. In principle, ROH could be used also to detect associations for continuous traits. In this case, either the extreme tails of the distribution would be used as contrasting groups, or the homozygosity of detected regions relative to that of the entire genome would be analysed. However, several aspects of the use of ROH in association studies still need to be further investigated (e.g. testing for significance, accounting for population structure and systematic effects). Hopefully, this communication might serve of inspiration and foster a fruitful line of research.

LITERATURE CITED

Biscarini, F, Nicolazzi, E.L., Stella, A. et al. (2012). 63rd EAAP Book of Abstracts, pag. 95

Biscarini, F., Del Corvo, M., Stella, A. et al. (2013). Actas XV Jor. Prod. Anim. AIDA, volume 2:538-540

Cole, J., Wiggans, G., Ma, L. et al. (2011). BMC Genom. 12:408

D'Agostino, R.B., Massaro, J.M. and Sullivan, L.M. (2002). Statis. Med. 22:169-186

Hildebrandt, F., Heeringa, S.F., Rüschendorf, F. et al. (2009). PLoS genet. 5(1),e10000353

Luo, Z.W. 1998. Heredity 80:198-208

McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R. et al. (2008). Am. J. Hum. Genet. 83:359-372

Minozzi, G., Nicolazzi, E.L., Stella, A. et al. (2013). PLos ONE 8(11), e80219

Pollott, G.E. (2012). EAAP Book of Abstracts, pag. 231

Purcell, S., Neal, B., Todd-Brown, L. et al. (2007). Am. J. Hum. Genet. 81:559-575

Purfield, D.C., Berry, D.P., McParland, S. et al. (2012). BMC Genet. 13:70

Sullivan, P.S., Grubbs, S.T., Olchowy, T.W. et al. (1994). J. Am. Vet. Med. Assoc. 205:1763-1766