Proc. Assoc. Advmt. Anim. Breed. Genet. 20:129-129

COMBINING MULTIPLE TEST-STATISTICS INCREASES THE POWER OF SELECTIVE SWEEP ANALYSES IN CATTLE

I.A.S. Randhawa[1], P.C. Thomson[1], M.S. Khatkar[1] and H.W. Raadsma[1]

[1]ReproGen - Animal Bioscience Group, Faculty of Veterinary Science, University of Sydney, Camden, NSW 2570, Australia

SUMMARY

Technological advances in targeted DNA sequencing, SNP genotyping and biometrical tools, allow for accurate localization of selection signatures. We present a simple method of combining ranks (mean fractional ranks, MFR) of multiple test-statistics as evidence of selection from single (F_{ST}, ΔDAF) and multiple (XP-EHH) marker based tests. *P*-values and FDR (*q*-values) to assess significance of an association can be determined from MFR: this cannot be done for its constituent tests. MFR is validated in two datasets (grouped for the presence or absence of either polledness or double muscling) from 375 animals of 21 cattle breeds with genotypes on 38,610 SNP assays from an Illumina BovineSNP50 chip. Candidate regions under selection (CRS) on chromosomes 1 and 2 were localized to regions of 610 and 680 kb near the functional mutations causing polledness and double muscling in cattle, respectively. The existence of strong signals of low FDR (i.e., > 85% of SNPs in CRS have $q < 0.05$) close to the candidate genes confirms the robustness of MFR.

INTRODUCTION

Trait-specific signals of selection are very challenging to identify. Multiple methods have been developed for the detection of selection signatures from genome-wide single nucleotide polymorphism (SNP) data. These have been extensively implemented in population studies for many species. The specificity of each selection test is limited to certain aspects of selective forces operating under various models of selection. Hence, many tests being used to link genotypes with phenotypes often provide differing results for the same genomic data (Lin *et al.* 2010).

Non-neutral patterns of local genomic variation may reflect historical selective sweeps resulting in signatures of selection. A population undergoing positive selection for specific traits can exhibit signals of selection at the underlying genomic regions when measured by various selection tests of allele frequency spectrum and haplotype structures (Qanbari *et al.* 2011). Therefore, a combination of multiple strategies would appear to be a robust approach in localizing candidate regions under selection (CRS) and correlate them with phenotypic variation. Recently, several approaches have been developed (Grossman *et al.* 2010; Lin *et al.* 2010; Pavlidis *et al.* 2010) which combine multiple summary statistics in order to improve the power of detecting selection signatures. However, the complexity of methods, extensive range of computational resources and prior knowledge required to implement available combining approaches leaves researchers with limited resources at a disadvantage. To improve trait-specific genome-wide selection scans, we present a simple method of combining evidence from the ranks of several selection tests requiring no prior information and it is potentially ideal for outbred populations.

MATERIALS AND METHODS

This study was conducted on two well characterized traits under selection in cattle to validate the MFR method. We investigated multi-breed panels from 212 (dataset I: polled versus horned breeds) and 357 (dataset II: double muscle versus normal muscle breeds) cattle samples genotyped with the Illumina BovineSNP50 chip assays, available from Gautier *et al.* (2010). We used 38,600 SNPs that were mapped on the UMD3.1 bovine assembly. Imputation of missing genotypes and haplotyping were performed with BEAGLE 3.3 (Browning and Browning 2007). Ancestral and

derived allelic polarity was acquired from Decker *et al.* (2009) and Matukumalli *et al.* (2009).

Mean Fractional Ranks (MFR). We combined three popular constituent tests to capture evidence for selection across multiple populations from genetic polymorphism data namely change in allele diversity by F_{ST} (Weir and Cockerham 1984), across population extended haplotype homozygosity (XP-EHH) test (Sabeti *et al.* 2007) and change in derived allele frequencies (ΔDAF) (Grossman *et al.* 2010). We derived composite test statistics (i.e., MFR) by combining 3 tests statistics at the same SNP, as well as determine *P*-values for these composite tests, to test the presence of a common signal as follows:

Let T_{ij} be the test statistic using method $i, (i = 1, \dots, m)$ calculated at SNP $j, (j = 1, \dots, n)$. Then for each test statistic type i obtain the rank of each observed test statistic across all n SNPs, say $R_{ij} = \text{rank}\,(T_{ij})$, which take values $1, \dots, n$ (using R program's `rank` function with default options so that it averages the sequential ranks for equal scores on multiple SNPs of a test). Next, these ranks are converted to fractional ranks by re-scaling them to lie between 0 and 1, i.e. $R'_{ij} = R_{ij}/(n+1)$, giving values from $1/(n+1)$ through $n/(n+1)$. Next, the MFR of the test statistics at each SNP is calculated, averaging over all the test statistic methods, $\bar{R}'_j, j = 1, \dots, n$. If there is a common signal across the multiple test statistics, this will show up as an excess in the $\bar{R}'_j$ value at that point, otherwise, $\bar{R}'_j$ may be dampened down, i.e. regressed to the average. Under the null hypothesis of no common signal, we can regard the values of R'_{ij} as m independent observations from a uniform $U(0,1)$ distribution, and using the results of Sadooghi-Alvandi *et al.* (2009) for the sum of m $U(0,1)$ random variables, we can derive the distribution of the mean $\bar{R}'_j$ as follows.

The probability density function (PDF) of $\bar{R}'_j$ is obtained as

$$f(r) = \frac{1}{(n-1)!}\sum_{k=0}^{n}(-1)^k \binom{n}{k}[(rn-k)_+]^{n-1}, 0 \le r \le 1$$

where $x_+ = x$ if $x > 0$, or 0 otherwise. By integration, the cumulative distribution function (CDF) is obtained as

$$F(r) = \frac{1}{n!}\sum_{k=0}^{n}(-1)^k \binom{n}{k}[(rn-k)_+]^{n}, 0 \le r \le 1$$

So for a mean scaled rank of $\bar{R}'_j$, the p-value for a test of no common signal would be calculated as $p = 1 - F(\bar{R}'_j)$.

The top 0.1% of $-\log_{10}$ of the empirical p-values were used to declare a SNP to be significant relative to the rest of the genome. The effectiveness of multiple tests was also compared gradually at various thresholds. Further, empirical *p*-values were calibrated using the ConReg-R method (Li *et al.* 2011) and the tail area based false discovery rate (FDR) i.e., q-values were estimated.

RESULTS AND DISCUSSION

Genome-wide distribution of empirical scores (non-smoothed) indicates that the highest $-\log_{10}(p)$ of MFR values above various thresholds were in the candidate regions in both datasets (Figure 1). The three component tests (FST, ΔDAF and XP-EHH) were found significant in the candidate gene regions but with fewer and lower ranked SNPs as compared to the MFR test (results not shown). To reduce spurious signals, the test statistics were smoothed by averaging statistics over SNPs within 1 Mb sliding windows centered at each SNP (Figure 2). Putative regions under selection (PRS) were defined from windows containing at least 3 significant SNPs and first to last SNP (top 0.1 %) positions as its boundaries. In total, 9 and 12 PRSs detected by at

least one of the constituent selection tests were substantially reduced at common signals to 3 and 4 PRSs by the MFR method in datasets I and II, respectively (Table 1). Genes located within the PRS ± 0.5 Mb positions were investigated for previously reported candidates of selection to localize CRS. MFR shows clusters of significant SNPs as peaks of selection signatures in CRSs on bovine autosome (BTA) 1 and 2 (Figure 2). The presence of non-candidate selection signals was much lower in MFR as compared to constituent tests (results not shown). The strategy of combining multiple test statistics has neutralized the unique patterns of each constituent selection test. In the empirical MFR distribution, the significant scores have an FDR < 0.0001, and after smoothing > 85% of SNPs in CRSs have $q < 0.05$. Additional peaks at PRSs by MFR also indicate the presence of genes under selection, for example; in the dataset I, a strong phenotypic diversity also exists for stature on BTA13 and 14, see Randhawa *et al.* (2013).

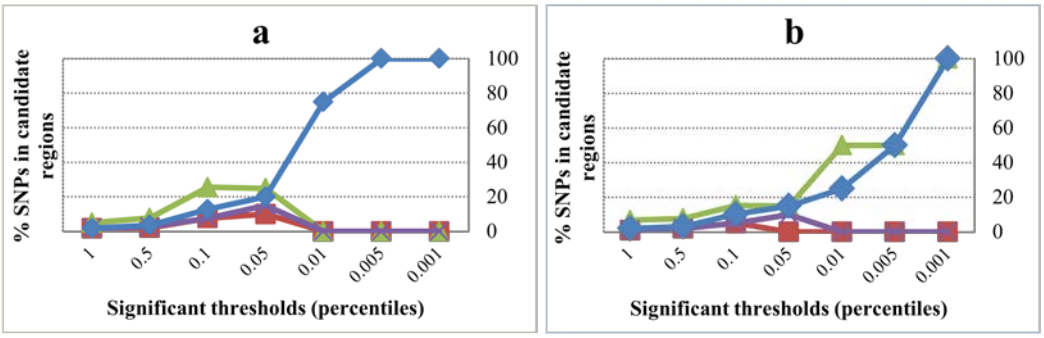

Figure 1: Percentage of significant SNPs present within the candidate gene regions (*y*-axis) identified by MFR (◆), XP-EHH (▲), F_{ST} (■) and ΔDAF (×) in a) polled and b) double muscle cattle at various thresholds (*x*-axis).

Polledness in cattle. In dataset I, out of 39 SNPs above the top 0.1% MFR scores, 10 SNPs within 610 kb span were found in the CRS harbouring *POLL* locus on BTA1 (Figure 2a). The *POLL* locus contains candidate mutations at the proximal end of BTA1 (1.65–2.05 Mb) where dominant alleles cause the polledness in cattle (Matukumalli *et al.* 2009; Allais-Bonnet *et al.* 2013).

Double muscling in cattle. In dataset II, among 39 significant MFR scores, a cluster of 10 SNPs was localized at CRS of 680 kb flanking myostatin (MSTN) gene at BTA2 (Figure 2b). Bovine MSTN gene (6.21–6.22 Mb) harbours various loss-of-function mutations or an 11 bp deletion in its third exon that underlie the muscular hypertrophy in some beef cattle (Piedmontese, Belgian Blue, South Devon and Asturiana de los Valles) breeds (Georges 2010).

Table 1: Regions under selection (putative = PRS, candidate = CRS) and significant SNPs in constituent and composite tests, and FDR of MFR in both datasets of cattle

Dataset	Total number of			Number of PRS and (SNPs* in CRS) in				% FDR¶ of MFR in Genome and (CRS)
	PRS	SNPs* in PRS	SNPs† in CRS	Constituent tests			Composite MFR	
				XPEHH	F_{ST}	ΔDAF		
I	9	105	14	3 (9)	5 (1)	5 (0)	3 (10)‡	9.8 (86.0)
II	12	129	10	5 (10)‡	4 (3)	5 (0)	4 (10)‡	6.2 (90.0)

* Significant SNPs † Total genomic SNPs ‡ Extreme scoring SNPs ¶ $q<0.05$

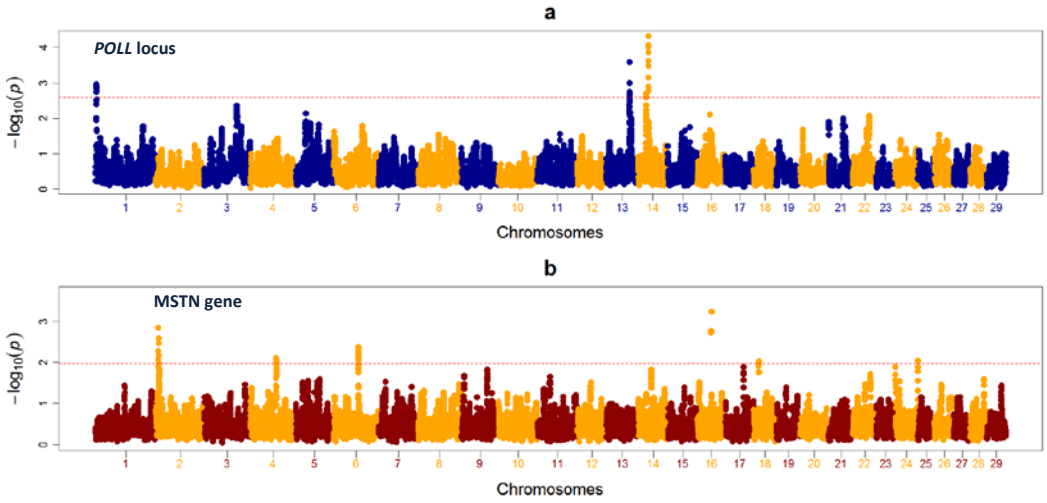

Figure 2: Manhattan plots of smoothed $-log_{10}(p)$ of MFR for a) polled and b) double muscle cattle. Dashed lines indicate genome-wide top 0.1% thresholds in both datasets.

Overall, MFR demonstrates its robustness even in the absence of any casual SNP in the genotype data. It provides an improvement for the predictions of positive selection as compared to its constituent tests of selection. MFR can be further improved by incorporating a strategy so that it can use the magnitudes of the actual test statistics. Moreover, MFR can easily accommodate additional selection tests given their sufficient power to distinguish selected and neutral loci in the genetic polymorphism data. This method can be used to identify the CRSs harbouring functional SNPs in genes for simple and potentially also for complex traits in domestic species.

REFERENCES

Allais-Bonnet A., Grohs C., Medugorac I., Krebs S., Djari A. *et al.* (2013) *PLoS ONE.* **8**:e63512.
Browning S.R. and Browning B.L. (2007) *Am. J. Hum. Genet.* **81**:1084.
Decker J.E., Pires J.C., Conant G.C., McKay S.D. *et al.* (2009) *Proc. Natl. Acad. Sci.* **106**:18644.
Gautier M., Laloë D. and Moazami-Goudarzi K. (2010) *PLoS ONE.* **5**:e13038.
Georges M. (2010) *Immun., Endoc. & Metab. Agents in Med. Chem.* **10**:240.
Grossman S.R., Shylakhter I., Karlsson E.K., Byrne E.H. *et al.* (2010) *Science.* **327**:883.
Li J., Paramita P., Choi K.P. and Karuturi R.K.M. (2011) *Biol. Direct.* **6**:27.
Lin K., Li H., Schlotterer C. and Futschik A. (2010) *Genetics.* **187**:229.
Matukumalli L.K., Lawley C.T., Schnabel R.D., Taylor J.F. *et al.* (2009) *PLoS ONE.* **4**:e5350.
Pavlidis P., Jensen J.D. and Stephan W. (2010) *Genetics.* **185**:907.
Qanbari S., Gianola D., Hayes B., Schenkel F., Miller S., Moore S., Thaller G. and Simianer H. (2011) *BMC Genomics.* **12**:318.
Randhawa I.A.S., Khatkar M.S., Thomson P.C. and Raadsma H.W. (2013) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **20**:(in press).
Sabeti P.C., Varilly P., Fry B., Lohmueller J., Hostetter E. *et al.* (2007) *Nature.* **449**:913.
Sadooghi-Alvandi S., Nematollahi A. and Habibi R. (2009) *Statistical Papers.* **50**:171.
Weir B.S. and Cockerham C.C. (1984) *Evolution.* **38**:1358.