Advances in Farm Animal Genomic Resources (GENOMIC-RESOURCES)

Final Scientific Report (EX/3726)

Title: Identifying signatures of recent selection through the comparison of indicine and taurine cattle populations of beef and dairy types

Grantee: Yuri Tani Utsunomiya. Universidade Estadual Paulista - UNESP/Brazil

Host: Johann Sölkner. Universität für Bodenkultur Wien - BOKU/Austria

Exchange visit period: April - June 2012

Purpose of the visit

1) Perform genome-wide scans for footprints of recent positive selection in four cattle breeds (Angus, Brown Swiss, Gir and Nellore) using high-density SNP data, with special focus on breed-specific signatures.
2) Benefit the student's academic formation by exposing him to state of the art analyses and different working environment.
3) Contribute to the scientific community with more knowledge in the field of signatures of natural and artificial selection that may be related to milk and meat production or taurine and indicine fitness.

Description of the work carried out during the visit

Overview

Motivated to identify footprints of recent positive selection in cattle, I sought a strategy for integrating multiple tests in genome-wide SNP data. I applied a straightforward frequentist meta-analysis approach for combining P-values across tests, targeting common, moderate frequency variants. I covered two between and two within population tests for selection sweeps, divided into three different categories: extended haplotype homozygosity, change in the allele frequency spectrum and local heterozygosity depression. I also implemented strategies for assigning relevant SNPs to genes, allowing for exploration of the biological meaning of the findings and facilitating hypothesis generation. Additionally, I performed discovery of the ancestral *Bovinae* allele state of over 440,000 SNPs.

Samples description and quality control

Genotypes for Illumina® BovineHD Genotyping BeadChip assay of Angus (ANG), Brown Swiss (BSW), Gir (GIR) and Nellore (NEL) individuals were available for prospection of selection sweeps. Details on sample size and source for each breed can be found in **Table 1**. Only autosome markers were included into the analyses. SNPs were removed from the dataset if they did not exhibit: (1) minor allele frequency (MAF) greater than or equal to 0.03, (2) P-value for Hardy-Weinberg Equilibrium (HWE) greater than or equal to 1×10^{-6} or (3) Call rate (CR_{SNP}) greater than or equal to 90%. After the SNP QC, individuals exhibiting call rate (CR_{IND}) below

90% were also removed. This procedure was performed for each breed in parallel using PLINK (Purcell et al., 2007). In order to mitigate relatedness in the dataset, individuals were further investigated for proportions of alleles shared identically by descent using PLINK. I developed an algorithm for conservatively remove samples from potential parent-offspring, half-siblings and duplicate pairs. SNPs commonly passing QC in all four breeds were then overlapped. For each breed, a minor imputation procedure was performed with the fastPHASE software (Scheet & Stephens, 2006) to solve for remaining missing data.

Table 1. Description of samples available for analysis.

Breed	Code	Subspecies	Purpose	Source			Total
				HapMap[1]	BOKU[2]	ZGC[3]	
Angus	ANG	*Bos taurus*	Beef	27	0	0	27
Brown Swiss	BSW	*Bos taurus*	Dairy	24	48	0	72
Gir	GIR	*Bos indicus*	Dairy	30	0	0	30
Nellore	NEL	*Bos indicus*	Beef	35	0	691	726

[1] The Bovine HapMap Consortium, [2] Universität für Bodenkultur, [3] Zebu Genome Consortium

Ancestral allele discovery

For the ancestral allele discovery, there were available 2 Gaur (*Bos gaurus)*, 6 Water Buffalo (*Bubalus bubalis*) and 2 Yak (*Bos grunniens*) samples typed for the same assay. Genotypes for the three *Bovinae* species were pooled into a single dataset. I looked for markers with a CR_{SNP} of 100% and with a single allele present (100% AA or 100% BB). For each case, the allele was determined as ancestral. The final SNP set was defined as markers passing QC with ancestral allele information present.

Long-range haplotype based methods

The set of methodologies cited here were calculated using the *rehh* package in R (Gautier & Vitalis, 2012), with minor modifications to the source code. As base for the actual tests, I calculated the integrated Extended Haplotype Homozygosity for the ancestral allele (iHH_A), derived allele (iHH_D) and SNP site (*iES*) for each marker. For the within breed test, I calculated the Integrated Haplotype Score (*iHS*) based on iHH_A and iHH_D, as described by Voight et al (2006). The scores were divided into 20 equally sized bins according to their derived allele frequencies, and then standardized to have mean 0 and variance 1. As both tails from the distribution are of interest, I derived two-sided P-values as $1-2|\phi(iHS)-0.5|$ from the Gaussian cumulative density function. For the between breeds tests, I computed the pairwise *Rsb* (Tang et al., 2007) from *iES*. Although the statistic is also standardized, the procedure recommended by the authors does not divide scores into bins and uses the median instead of the mean. As each comparison was performed twice, just shifting the direction of selection (Breed A x Breed B, with positive values representing selection towards Breed A; and Breed B x Breed A, with positive values expressing selection towards Breed B), I derived one-sided upper tail P-values from the normal cumulative density function.

Change in the allele frequency spectrum based method

Grossman et al (2010) describe a very simple method computed as the difference in the derived allele frequency between populations, called *ΔDAF*. Values range from -1 to 1 and

are also normally distributed. I standardized *ΔDAF* scores using the distribution mean and standard deviation, and retrieved one-sided upper tail P-values.

Local heterozygosity depression based method

Rubin et al (2010) defined and applied a Z-score test for local heterozygosity depression (*ZHp*) on whole genome sequence data of domestic chicken, which basically expresses how much the expected heterozygosity in chromosome windows deviate from the average genome heterozygosity. I adapted the approach to every SNP site and computed the observed instead of the expected heterozygosities. Values obtained resembled a normal distribution, and were standardized to produce mean 0 and variance 1. This time, negative values are of interest and the resulting site heterozygosity scores are multipled by -1 in order to switch their direction, yielding a new statistic called *SHp*. I derived one-sided upper tail P-values for each score obtained.

Combination of multiple tests

As all statistics approached have P-values retrieved from normal distributions with same parameters (mean 0 and variance 1), I adapted the weighted version of Stouffer method for meta-analysis of Z-transformed P-values (as reviewed by Whitlock, 2005). For each population, for each marker and each test *i*, the respective P-value is transformed into a Z-score by $Z_i = -\Phi^{-1}(1-p_i)$. Within population tests are performed only once per breed, hence their respective weight ω_i is set to 1. For each comparison of between population tests, the Z-score is weighted to $1/n$, where n is the number of comparisons. Then, the combined statistics of k tests, for each SNP in each breed, is defined as:

$$meta\text{-}SS = \frac{\sum_{i=1}^{k} \omega_i Z_i}{\sqrt[2]{\sum_{i=1}^{k} \omega_i{}^2}}$$

The *meta-SS* (stands for Meta-analysis of Selection Signals) scores are referred back to the standard normal distribution to obtain combined significance values, which are intended to address either the combination of information among different, independent tests can reject the shared null hypothesis (neutral marker) or not. Significance level for genome-wide *meta-SS* P-values was based on a Bonferroni threshold ($0.05/n_{SNP}$).

Functional annotation

For any peak crossing the significance line, I applied three different strategies for the annotation of functional features. The first approach consists on checking if any significant SNP is intragenic via mining the Ensembl Variation 67 database with the Ensembl Biomart tool (Kinsella et al., 2011). The second strategy comprehends isolating the most significant SNP from each visually identifiable peak and mapping the closest gene to it. For that matter, I downloaded the Bovine UMD3.1 gene set from Ensembl Genes 67 database via Biomart tool and used the ClosestBed algorithm from the BedTools software (Quinlan & Hall, 2010). The third approach is a LD-based window scheme, divided into three steps.

First, every SNP crossing the significance line is defined as a 'core SNP'. Second, I walk down to proximal and distal chromosome positions calculating correlations between the core SNP and the neighbor markers, checking if they tag the core SNP or not based on r^2 values. The r^2 threshold adopted to declare that one marker tags the core SNP is set to 0.7. The positions of the last tag markers on both sides of the core SNP, i.e., positions from where r^2 decays below the defined threshold or the tag marker distance from the core position exceeds 1 Mb, are set as the boundaries of a window. This analysis was done in PLINK. Third, this window is interpreted as a single locus, and any gene overlapping it is considered to be in LD with the core SNP and is therefore annotated. For core SNPs which no window boundaries could be determined, I included the closest gene in the vicinity to the list. As some windows may also overlap, the gene list yielded is then parsed to exclude repeated gene names and then processed in DAVID (Huang et. al, 2009-A; Huang et. al, 2009-B) for annotation of functional terms. I used the default parameters for each breed gene list, pooling together all genes annotated across the genome to reveal over-represented functional terms. Finally, I used the Enrichment Map Cytoscape plugin (Merico et al., 2010) to build networks of inter-related enriched terms based on the number of overlapping genes.

Description of the main results obtained

Ancestral allele discovery

Assessing the outgroup species genotypes, I observed an average CR_{IND} of 83.79%, 96.93%, 94.87% and 88.63% for Water Buffalo, Yak, Gaur and pooled data, respectively. Considering only markers perfectly typed across the pooled outgroup samples (CR_{SNP} = 100%), I observed a total of 559,663 (71.94%) SNPs successfully genotyped, from which 111,376 were polymorphic (MAF > 0). Hence, a total of 448,287 SNPs (56.75%) had their ancestral allele determined.

Quality control

From the initial set of 742,910 autosome markers, numbers of SNPs passing QC were 579,470, 554,826, 485,655 and 461,702 for ANG, BSW, GIR and NEL, respectively. Overlapping of the four SNP lists retrieved a final set of 281,994 markers, from which 157,702 had ancestral allele information available. Even with the drastic drop in number of SNPs, the intermarker distance mean and median were 15.94 kb and 6.43 kb, respectively, superposing the median spacing of 37 kb declared for the BovineSNP5O assay (Matukumalli et al., 2009). The number of remaining samples for each breed after duplicates and first degree relationships removal were: 24 ANG, 44 BSW, 23 GIR and 581 NEL. As NEL exhibits a sample size much larger than the other breeds, 45 individuals were sampled from the total, in order to do fair comparisons.

Identification of selection signals and overview of functional annotation

All tests for footprints of selection performed resembled normal distributions and genome-wide Z-transformed P-values were weakly correlated, satisfying the independence condition for meta-analysis. After combining P-values, the number of SNPs crossing the genome-wide significance ($P < 3.17 \times 10^{-7}$) was: 153 for ANG, 212 for BSW, 3 for GIR and 13 for NEL. The most significant SNP was found in BSW ($P = 3.82 \times 10^{-12}$), and is an intronic variation

in Cornichon homolog 3 gene (CNIH3), located at BTA16:28478192. The genome-wide distribution of *meta-SS* P-values and the closest genes to the top of the peaks can be found in **Figure 1**.

In order to illustrate the potentiality of combining signals from different methodologies, P-values for each one of the individual tests for the CNIH3 region in BSW (candidate for being selected) and NEL (candidate for being neutral) are displayed in detail in **Figure 2**. EHH decay plots and a bifurcation diagram for the haplotypes containing the derived allele are also provided. It can be seen from BSW and NEL comparison that the composite test penalizes SNPs with little statistical support. The signal of the unusual derived allele long haplotype in BSW, revealed by the *meta-SS* statistic, is not detectable in NEL.

The number of genes directly harboring significant SNPs was ANG 20, BSW 27, GIR 1 and NEL 3. I found 2 synonymous exonic SNPs for ANG and BSW, 1 non-synonymous variation (Ala->Thr) for a gene of the olfactory receptor family (LOC524290 - OR2W3) in ANG ($P = 7.65 \times 10^{-9}$) and a 3'UTR variation for the KIF5C (kinesin family member 5C) gene in NEL ($P = 2.68 \times 10^{-7}$). All other variants within genes were located in introns. Total number of genes within LD-windows included in each breed specific list was: 309 ANG, 177 BSW, 4 GIR and 14 NEL.

Network of enriched terms from ANG gene list revealed three groups: (1) immune response related genes, involved with chemokine and cytokine activity; (2) transcription activity, comprehending the biosynthesis of ribonucleoproteins, transcription activation and aminoacylation of tRNA with L-histidine residuals; and (3) glucolysis and gluconeogenesis pathways. For BSW, a cluster related to post-transcriptional modifications of RNAs (mostly methylation of adenosine residuals) and another involved with Calpain were observed. A significant intronic SNP (BTA16:27801014, $P = 2.61 \times 10^{-7}$) was detected in the Calpain 2 (m-Calpain) catalytic subunit, which may be capturing the signal of a causal untyped variant under selection. Due to a low number of genes mapped, the clustering of enriched terms for GIR and NEL retrieved no significant result. Across all lists, a total of 69 genes (13.69%) had no functional term associated to it, being either uncharacterized proteins or novel RNAs with no functional record available.

Future collaboration with host institution

Both host and guest institutions are already collaborating in the Zebu Genome Consortium, an international initiative to characterize indicine cattle genetic resources and develop adequate strategies for using genomic information in selection schemes, mainly in Nellore and Gir breeds. This exchange visit in particular allowed for stepping up current joint activities between my advisor Prof. José Fernando Garcia and the host researcher, Prof. Johann Sölkner. After this 3 months collaboration, Prof. Sölkner has kindly agreed to become my co-advisor in UNESP-Brazil.

Projected publications/articles resulting or to result from the grant

The results herein described have been intensively discussed with all parties, and a draft-paper was already composed. We plan to submit this work to a medium impact journal within the next months, under the provisory title of *Detecting loci under recent positive*

selection in dairy and beef cattle by integrating different genome-wide scan methods. The publication will acknowledge the European Science Foundation with the highest gratitude.

Other comments

This research was conceived by Prof. Johann Sölkner (BOKU-Austria) and Prof. José Fernando Garcia (UNESP-Araçatuba), who also provided genotypes of Brown Swiss and Nellore samples, respectively. PhD student Ana Maria Perez O'Brien (BOKU-Austria) has largely contributed to the study design, alongside with the grantee Yuri Tani Utsunomiya (UNESP-Brazil). The grantee developed the pipelines and performed all analyses, including the preparation of this report and the draft manuscript to be published. Research Geneticists Dr. Tad S. Sonstegard and Dr. Curtis P. Van Tassell (USDA-USA) provided Gaur, Yak, Water Buffalo, Gir, Angus and complementary Brown Swiss genotypes, and are currently evaluating the final manuscript. Thus, all people cited will be authors of the projected publication. Although not eligible as co-authors, many people have contributed to this research, either with direct fruitfully discussions on its content or with indirect insights. Among those, I would like to thank the Zebu Genome Consortium, the HapMap Consortium and personally Gabor Meszaros, Anamarija Frkonja, Ino Curik, Maja Ferencakovic and Patrik Waldmann. I would like to thank also all colleagues from BOKU for dedicating part of their time to make me feel home. Last but not least, my sincere appreciations to Prof. Sölkner, the European Science Foundation and the Advances in Farm Animals Genomic Resources Research Networking Programme for this great experience.

References

Gautier M, Vitalis R. rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics Advance Access. 2012.

Grossman SR et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science. 327, 883-886 (2010).

Huang DW, Sherman BT, Lempicki RA (A). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57.

Huang DW, Sherman BT, Lempicki RA (B). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). Vol. 2011 Published online Jul 23, 2011.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP,O'Connell J, Moore SS, Smith TP, Sonstegard TS, Van Tassell CP. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One.2009;4(4):e5350.

Merico D, Isserlin R, Stueker O, Emili A, Bader GD . Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation PLoS One. 2010 Nov 15;5(11):e13984.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

Rubin CJ et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 464, 587-593 (2010).

Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006; 78:629–644.

Tang K, Thornton KR, Stoneking M. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biology. 5, e171 (2007).

Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 4, e72 (2006).

Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J. EVO L. BI OL. 18 (2005) 1368–1373.

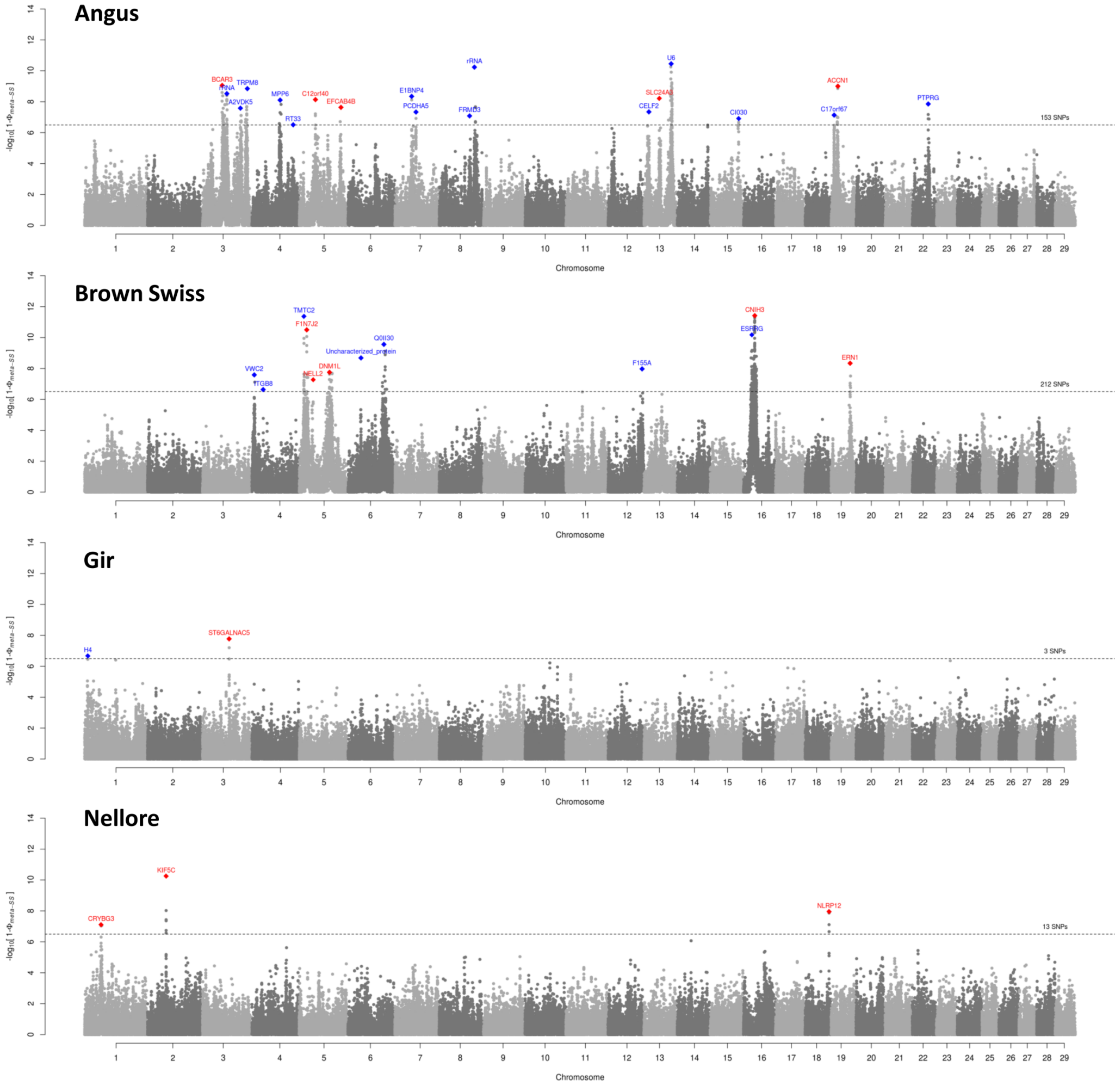

Figure 1. Manhattan plots for genome-wide *meta-SS* P-values. Diamonds represent top SNPs on peaks crossing the significance line ($P < 3.17 \times 10^{-7}$). Red and blue diamonds are intragenic and intergenic markers, respectively.

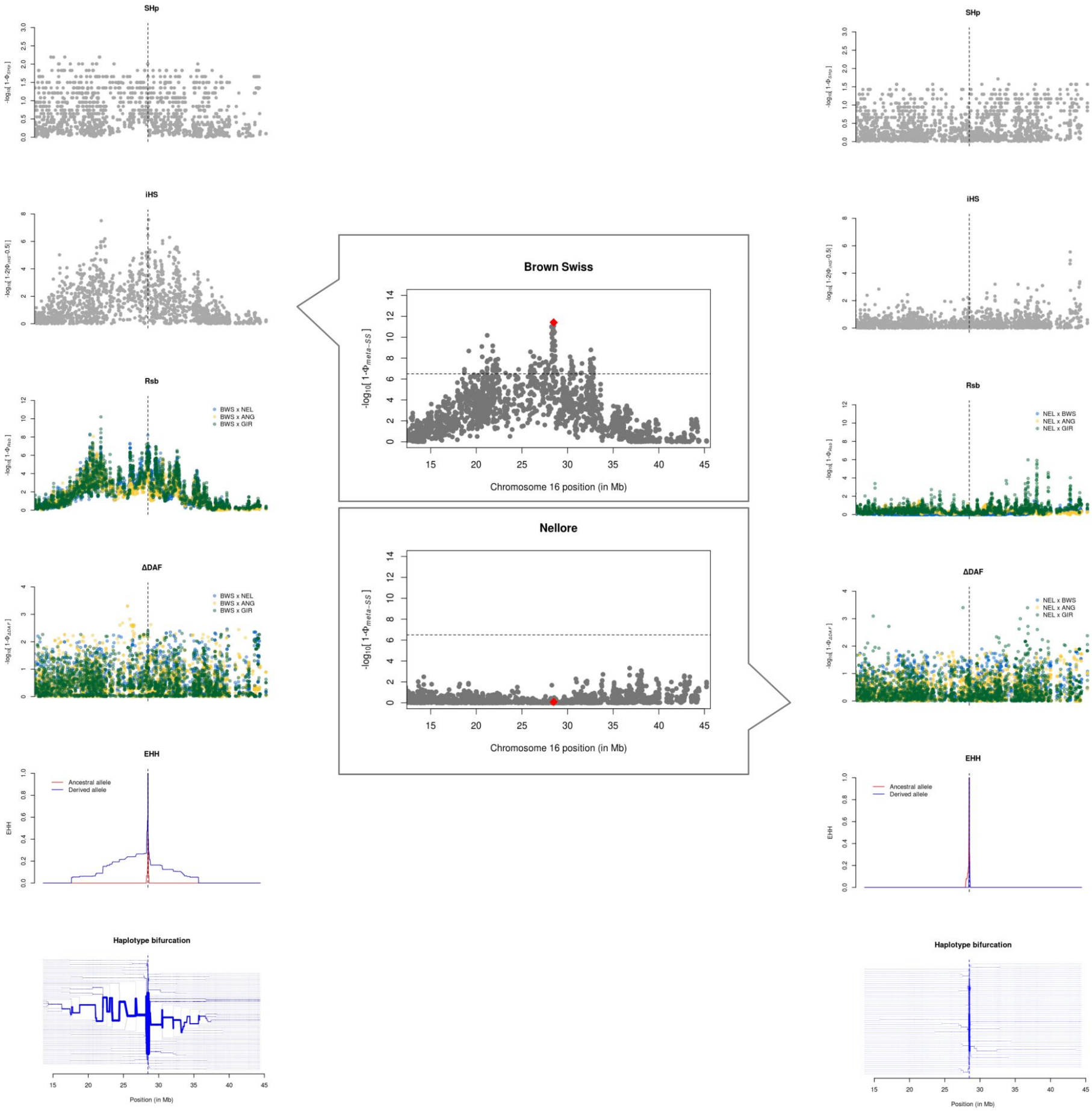

Figure 2. P-values for four individual tests (from top to bottom: *SHp*, *iHS*, *Rsb* and *ΔDAF*), EHH decay plot, derived allele bifurcation diagram and *meta-SS* P-values (center) for BSW and NEL for a region containing CNIH3. Vertical lines and red diamonds represent the position of the intronic SNP detected as highly significant in BSW ($P = 3.82 \times 10^{-12}$). Horizontal lines mark the Bonferroni significance threshold ($P < 3.17 \times 10^{-7}$).