Selection signatures in worldwide Sheep populations

Maria-Ines Fariello[1], Bertrand Servin[1], Gwenola Tosser-Klopp[1], Rachel Rupp[2], Carole Moreno[2], Magali SanCristobal[1], Simon Boitard[3,4]

1 Laboratoire de Génétique Cellulaire, INRA, Castanet-Tolosan, France

2 Station d'Amélioration Génétique des Animaux, INRA, Castanet-Tolosan, France

3 Génétique Animale et Biologie Intégrative, INRA & AgroParisTech, Jouy-en-Josas, France

4 Origine, Structure et Evolution de la Biodiversité, Museum National d'Histoire Naturelle & EPHE & CNRS, Paris, France

Abstract

The diversity of populations in domestic species offer great opportunities to study genome response to selection. The recently published Sheep Hapmap dataset is a great example of characterization of the world wide genetic diversity in the Sheep. In this study, we re-analyzed the Sheep Hapmap dataset to identify selection signatures in worldwide Sheep populations. Compared to previous analyses, we make use of statistical methods that (i) take account of the hierarchical structure of sheep populations, (ii) make use of Linkage Disequilibrium information and (iii) focus specifically on either recent or older selection signatures. We show that this allows to pinpoint several new selection signatures in the sheep genome and to distinguish those related to modern breeding objectives and to earlier post-domestication constraints. The newly identified regions, together with the one previously identified, reveal the extensive genome response to selection on morphology, color and adaptation to new environments.

Introduction

Domestication of animals and plants played a major role in human history. With the advance of high-throughput genotyping and sequencing technologies, the analysis of large datasets in domesticated species offer great opportunities to study genome evolution in response to phenotypic selection [1]. Sheep was the first grazing animal to be domesticated [2] in part due to its manageable size and an ability to adapt to different climates and poor nutrition diets. A large variety of breeds with distinct morphology, coat color or specialized production (meat, milk or wool) were subsequently shaped by artificial selection. Since the realease of the 50K SNP array [3], it is now possible to scan the genetic diversity in Sheep in order to detect loci that have been involved in these various adaptative selection events. The Sheep HapMap dataset, which includes 50K genotypes for 3000 animals from 74 breeds with diverse world-wide origins, provides a considerable ressource for deciphering the genetic bases of phenotype diversification in Sheep. In the first analysis of this data set [4], the authors looked for selection by computing a global F_{ST} among the 74 breeds at all SNPs in the genome. They identified 31 genomic regions with extreme differentiation between breeds, which included candidate genes related to coat pigmentation, skeletal morphology, body size, growth, and reproduction. Further studies took advantage of the Sheep HapMap ressource to detect genetic variants associated with pigmentation [5], fat deposition [2], or microphtalmia disease [6]. An other study [7] performed a genome scan for selection focused on American synthetic breeds, using an

F_{ST} approach similar to that in [4].

The 74 breeds of the Sheep HapMap dataset have a strong hierarchical structure, with at least 3 distinct differentiation levels: an inter-continental level (e.g. European breeds vs Asian breeds), an intra-continental level (e.g. Texel vs Suffolk European breeds), and an intra-breed level (e.g. German Texel vs Scottish Texel flocks). Recent studies [8, 9] showed that, when applied to hierarchically structured data sets, F_{ST} based genome scans for selection may lead to a large proportion of false positives (neutral loci wrongly detected as under selection) and false negatives (undetected loci under selection). This statistical issue is also compounded by the heterogeneity of effective population size among breeds, implying that some breeds are more prone to contribute large locus-specific F_{ST} values than others [9]. Apart from these statistical considerations, merging populations with various degrees of shared ancestry can limit our understanding of the selective process at detected loci. Indeed, the regions pointed out in [4] can be related to either ancient selection, as the poll locus which has likely been selected for thousands of years, or fairly recent selection, as the myostatin locus which has been specifically selected in the Texel breed. But in most situations the time scale of adaptation can not be easily determined.

Another limit of genome scans for selection based on single SNP F_{ST} computations is that they do not sufficiently account for the very rich linkage disequilibrium information, even when the single SNP statistics are combined into windowed statistics. Recently, we proposed a new strategy to evaluate the haplotypic differentiation between populations [10]. We showed that using this approach greatly increases the detection power of selective sweeps from SNP chip data, and enables to detect also soft or incomplete sweeps. These latter selection scenarios are particularly relevant in the case of breeding populations, where selection objectives have likely varied along time and where the traits under selection are often polygenic.

In this study we provide a new genome scan for selection based on the Sheep HapMap data set, where we distinguish selective sweeps between and within 7 broad geographical groups. The within group analysis aims at detecting recent selection events related to the diversification of modern breeds. It is based on the single marker FLK test [9] and on its haplotypic extension [10], that both account for population size heterogeneity and for the hierarchical structure between populations. The between group analysis focuses on older selection events and is only based on FLK. Overall, we confirm 19 of the 31 sweeps discovered in [4], while providing more details about the past selection process at these locus. We also identify 68 new regions under selection, with candidate genes related to coloration, morphology or

production traits.

Results and discussion

We detected selection signatures using methods that aim at identifying regions of outstanding genetic differentiation between populations, based either on single SNP, FLK [9], or haplotype, hapFLK [10], information. These methods have optimal power when working on closely related populations so we analyzed separately seven groups of breeds, previously identified as sharing recent common ancestry [4] and corresponding to geographical origins of breeds. Before performing genome scans for selection signatures, we studied the population structure of each group to identify outlier animals as well as admixed and strongly bottelnecked populations, using both PCA and model-based approaches [11, 12]. hapFLK was found robust to bottlenecks or moderate levels of admixture, but these phenomena may affect the detection power so we preferred to minimize their influence by removing suspect animals or populations. Details of these corrections are provided in the methods section. The final composition of populations groups are given in table S1.

Overview of selected regions

An overview of selection signatures on the genome across the different groups is plotted on Figure 1 and Table 1 provides their detailed description. We found 40 selection signatures with hapFLK and 24 with FLK, although we allowed a slightly higher false discovery rate for FLK than hapFLK (10% vs 5%). This result is consistent with a higher power for hapFLK than FLK, as was shown before [10]. Four regions are found with both the single SNP and the haplotype test and harbor strong functional candidate genes: NPR2, KIT, RXFP2 and EDN3 (see below). The overlap is thus small, illustrating that the two tests tend to capture different signals. In particular, hapFLK will fail to detect ancient selective sweeps where the mutation-carrying haplotype is small, and not associated with many SNPs on the chip. On the other hand, single SNP tests will fail to capture selective sweeps when a single SNP is not in high LD with the causal mutation. Six regions were detected in more than one group of breeds. They all contain strong candidate genes. Three of these genes are related to coat color (KIT, KITLG and MC1R), and could correspond to independent selection events (see discussion below). One region harbors a gene (RXFP2) for which polymorphisms have been shown to affect horn size and polledness

in the Soay [13] and Autralian Merino [14]. The signatures of selection in this region exhibit different patterns among groups. The signal is very narrow in the SWE and SWA groups, and is in fact not detected by the hapFLK test, whereas it affects a large genomic region in the CEU group where it is detected by hapFLK. In the ITA group, the FLK statistics do not reach significance, and the hapFLK signal is not high (minimum qvalue of 0.04). Together, the selection signatures suggest that selection on RXFP2, most likely due to selection on horn phenotypes, was carried out worldwide at different times and intensities. The last two regions harbor the HMGA2 gene, involved in selection for stature in dogs [15], and ABCG2, a strong QTL for milk production in cattle [16]. Populations selected for ABCG2 variants belong to different European regions (SWE, ITA and CEU).

In the paper presenting the sheephapmap dataset [4], 31 selection signatures where found, corresponding to the 0.1% highest single SNP F_{ST}. Using FLK and hapFLK, we confirm signatures of selection for 11 of these regions. Considering the two analyses were performed on the same dataset, this overlap can be considered as rather small. Tow reasons can explain it.

First, the previous analysis was based on the F_{ST} statistic. Although this statistic is commonly used for selection scans, it is prone to produce false positives when the history of populations is characterized by population trees with unequal branch lengths (*i.e.* variation in the amount of drift experienced by different populations) [9]. In particular, strongly bottlenecked breeds will contribute high F_{ST} values preferentially, even under neutral evolution. With FLK and $hapFLK$, this varying amount of drift is accounted for, and populations with long branch lengths will not contribute to the signal more than others [10]. In fact they will tend to contribute less as it is harder to rule out the effect of drift alone in such populations.

Second, the previous analysis was performed using all breeds at the same time. It is therefore possible that some of these regions correspond to differentiation between groups of breeds rather than within groups. To investigate this question, we performed a genome scan for selection between the ancestors of the seven population groups using the FLK statistic computed on their estimated allele frequencies [9]. We did not include SNPs lying in regions detected within groups as selection biases their estimated ancestral allele frequencies. The population tree was reconstructed using SNPs for which we have unambiguous ancestral allele information (Figure 2). The tree is decomposed into two main lineages, one for European breeds and one for Asian and African breeds. The African group exhibits a slightly higher branch length. We note however that this could be due to ascertainment bias of SNPs on the SNP array.

This led to the identification of 23 new selection signatures (figure 3 and table 2), 9 of them being common to the previous analysis. Overall, we fail to replicate with this analysis 12 of the regions in [4].

Selection Signatures within population groups

Coloration Many selection signatures are located around genes that have been shown to be involved in hair, eye or skin color. In particular many genes underlying selection signatures are involved in the development and migration of melanocyte and in pigmentation: EDN3, KIT, KITLG, MC1R and MITF. We can add to this list SOX10 and ASIP that show some evidence of selection: in the ITA group, the q-value of hapFLK near SOX10 is 6.2%, while the closest SNP to ASIP (s66432 and s12884) present suggestive FLK p-values of respectively $7.5\,10^{-4}$ and $6.8\,10^{-5}$ in the ASI group, and is significantly differentiated between the ancestral groups. All these genes have previously been reported as being likely selection targets and/or associated to color patterns in different mammalian species. Finally we found a signal for selection around the BNC2 gene, that has recently been associated with skin pigmentation in Humans [17]. All population groups present at least a selection signature on one of these genes, reflecting the widespread importance of color patterns to define sheep breeds. Inferring a precise history of underlying causal mutations for color patterns in this dataset is hard for several reasons: the precise phenotypic characterizations of coat color patterns in the SheepHapMap breeds are not available; the 50K SNP array used does not offer sufficient density to associate a given selection signature to a specific set of polymorphisms; finally, from the litterature, it appears that coat color is a complex trait, with high genetic heterogeneity. In particular, mutations in different genes can give rise to the same phenotype (*e.g.* in Horse [18]). Also, within a gene different mutations can give rise to different phenotypes, *e.g* mutations in the MC1R gene (also named the extensions locus) have been associated to a large panel of skin or coat colors [19–21]. Studying more precisely selection signatures related to coat color and the underlying selected mutations will likely require further sequencing experiments targeted at these genes. This in turn will help to understand the evolutionary history of the breeds and the effect of selection [22]. To potentially help in this task, in table S2 we list, for each "color gene", the populations that have likely been selected for.

Morphology Another group of genes that are found in selection signatures have known effects on body morphology and development. NPR2, HMGA2 and BMP2 were identified previously [4], but we

also found selection signatures around IGF1, ALX4 or EXT2, WNT5A and two Hox gene clusters (HOXA and HOXC). IGF1 has been shown to be a major determinant of small body size in dogs [23]. WNT5A and ALX4 are two genes involved in the development of the limbs and skeleton. ALX4 loss of function mutations cause polydactily in the mouse, through disregulation of the sonic hedgehog (SHH) signaling factor [24, 25]. Moreover, the ALX4 protein has been shown to bind proteins from the HOXA (HOXA11 and HOXA3) and HOXC (HOXC4 and HOXC5) clusters [26], both of which are found under selection signatures (see below). Located just besides ALX4 and corresponding to the same selection signature EXT2 is responsible for the development of exostose in the mouse [27]. Mutations in WNT5A are causing the dominant Human Robinow syndrome, *characterized by short stature, limb shortening, genital hypoplasia and craniofacial abnormalities* [28]. An ancestral selection signature is found near the ACAN gene, which expression was shown to be upregulated by BMP2 [29], another candidate gene for selection. Mutations in the ACAN gene have been shown to induce osteochondrosis [30] and skeletal dysplasia [31]. The ACAN region has also been shown to be associated with Human adult height [32].

Two selection signatures are localized close to *Hox* genes clusters. *Hox* genes are responsible for antero-posterior development and skeletal morphology along the anterior-posterior axis in vertebrates. One is a recent selection signature in the SWA group near the HOXA gene cluster and the other is an ancestral signature near the HOXC gene cluster, with a high differentiation of the ASI ancestor compared to AFR and SWA at the most significant SNP (OAR3_141586525).

Traits of agronomical importance Sheeps have been raised for meat, milk and wool production. Under selection signatures, we found several genes associated with these production traits. Apart from the selection signature in Texels on the MSTN gene for increased muscularity [33], discussed in [10], selection on HDAC9 could also be linked to muscling. HDAC9 is a known transcriptional repressor of myogenesis. Its expression has been shown to be affected by the callypige mutation in the sheep at the DLK1-DIO3 locus [34]. The HDAC9 signal corresponds to a selection signature in the Garut breed from Indonesia, a breed used in ram fights. Two selection signatures contain genes shown to be underlying QTLs with large effects on milk production (yield and composition) in cattle: ABCG2 [16] and SREBP1 [35]. The SREBP1 gene is also found in a genome region associated with milk composition in the Lacaune breed (unpublished data). Also, one of the ancestral selection signatures lies close to the INSIG2 gene, in the SREBP1 signaling pathway and recently shown to be associated with milk fatty acid composition

in Holstein cattle [36]. Two selection signatures relate to wool characteristics, one in the CEU group near the FGF5 gene, partly responsible for hair type in the domestic dog [37], and an ancestral selection signature on chromosome 25 in a QTL region associated to wool quality traits in the sheep [38, 39].

One of the strong outlying regions in the selection scan contains the PITX3 gene. Further analysis revealed that this signature was due to the German Texel population haplotype diversity differing from the other Texel samples (results not shown). It turns out that the German Texel sample consisted of a case/control study for microphtalmia [6], although the case/control status information in this sample is not given in the Sheep Hapmap dataset. The consequence of such a recruitment is to bias haplotype frequencies in the region associated with the disease, which provokes a very strong differentiation signal between the German Texel and the other Texel populations. This illustrates that our method for detecting selection has the potential to identify causal variants in case/control studies, while using haplotype information.

Ancestral signatures of selection

It is difficult to estimate how far back in time signatures of selection found in the ancestral tree appended. In particular, it would be interesting to place this population tree with respect to sheep domestication. Two genes lying close to ancestral selection signatures might indicate that the selection signatures captured could be rather old. First, we found selection near the TRPM8 gene, which has been shown to be a major determinant of cold perception in the mouse [40]. The pattern of allele frequency at the significant SNP (OAR1_6722309) is consistent with the climate in the geographical origins of the population groups. AFR, ASI and ITA, living in warm climates, have low frequency (0.04-0.16) of the A allele, while NEU and CEU, from colder regions, have higher frequencies (0.55-0.7), the SWE group having an intermediate frequency of 0.38. Overall, this selection signature might be due to an adaptation to cold climate through selection on a TRPM8 variant. Another selection signature lies close to a potential chicken domestication gene, TSHR [41], which signaling regulates photoperiodic control of reproduction [42]. This selection signature was identified before [4] and our analysis indicates that it happened in the ancestral population tree, consistent with an early selection event. Given its role, we can speculate that selection on TSHR gene is related to seasonality of reproduction. Under temperate climates, sheep experience a reproductive cycle under photoperiodic control. Furthermore, there is evidence that this control was altered during domestication [43] so our analysis suggests genetic mutations in TSHR may have contributed to this

alteration.

As discussed above, some of the genes found underlying ancestral selection signatures can be related to production or morphological traits (*e.g.* ASIP, INSIG2, ACAN, wool QTL), indicating that these traits have likely been important at the beginning of the sheep history. The other genes that we could identify as likely selection targets in the ancestral population tree relate to immune response (GATA3) and in particular to antirival response (TMEM154 [44], TRAF3 [45]). The most significant ancestral selection signature coincides with the NF1 gene, encoding neurofibromin. This gene is a negative regulator of the ras signal transduction pathway, therefore involved in cell proliferation and cancer, in particular neurofibromatosis. Due to this central role in intra-cellular signaling, mutations affecting this gene can have many phenotypic consequences so that its role in the adaptation of sheep breeds remains unclear.

Conclusions

We conducted a genome scan for selection in a large worldwide set of breeds from the Sheep Hapmap dataset. Using recently developed methods, we were able to detect a very large number of selection signatures in different geographical groups. We also found selection signatures that most likely predate the formation of contemporary breeds. This analysis reveals strong response of the genome diversity in sheep populations with respect to selection on morphology and color, and the influence of recent selection on production traits. We also pinpoint two strong candidate genes (TRPM8 and TSHR) most likely involved in selection response during the early history of domestic sheep.

Elucidating causal variation underlying these selection signatures will most likely require large scale sequencing projects, together with phenotypic characterization of individuals or populations. This study can help in targeting specific breeds and traits to be studied in priority in such projects.

Methods

Selecting populations and animals Seventy four breeds are represented in the Sheep HapMap data set, but we only used a subset of these breeds in our genome scan. We removed the breeds with small sample size (< 20 animals), for which haplotype diversity can not be determined with sufficient precision. Based on historical information, we also removed all breeds resulting from a recent admixture or having

experienced a severe recent bottleneck. Focusing on the remaining breeds, we then studied the genetic structure within each population group, in order to detect further admixture events. We performed a standardized PCA of individual based genotype data and applied the admixture software [12].

In two population groups (AFR and NEU) the different breeds were clearly separated into distinct clusters of the PCA and showed no evidence of recent admixture. These samples were left unchanged for the genome scan for selection. A similar pattern was observed in three other groups (ITA, SWA, ASI), except for a few outlier animals that had to be re-attributed to a different breed or simply removed (Figures S1, S2 and S3). In the two last groups (CEU and SWE), several admixed breeds were found and were consequently removed from the genome scan analysis (Figures S4 and S5).

We performed a genome scan within each group of populations listed in table S1, with a single SNP statistic FLK [9] and its haplotype version hapFLK [10].

Population trees Both statistics require estimating the population tree, with a procedure described in details in [9]. Briefly, we built a population tree for each group by first calculating Reynolds' distances between each population, and then applying the Neighbour Joining algorithm on the distance matrix. For each group, we rooted the tree using the Soay sheep as an outgroup. This breed has been isolated on an Island for many generations and exhibits a very strong differentiation with all the breeds of the Sheep hapmap dataset, making it well suited to be used as an outgroup.

FLK and hapFLK genome scans The FLK statistic was computed for each SNP within each group. The evolutionary model underlying the FLK statistic assumes that the mutation was present in the ancestral population. To consider only loci that most likely match this hypothesis, we restricted our analysis within each group to SNPs which estimated ancestral minor allele frequency p_0 was above 5%. Under neutrality, the FLK statistic should follow a χ^2 distribution with $n-1$ degrees of freedom (DF), where n is the number of populations in the group. Overall, the fit of the theoretical distribution to the observed distribution was very good (supporting information Text S1) with the mean of the observed distribution ($\overline{FLK}$) being very close to $n-1$ (table S4). Using $\overline{FLK}$ as DF for the χ^2 distribution provided a better fit to the observed data than the $n-1$ theoretical value. We thus computed FLK p-values using the $\chi^2(\overline{FLK})$ distribution. To compute the hapFLK statistic, we make use of the Scheet and Stephens LD model [46], a mixture model for haplotypes which requires specifying a number of haplotype clusters to be used. To choose this number, for each group, we used the fastPHASE cross-

validation based estimation of the optimal number of clusters. Results of this estimation are given in table S3. The LD model was estimated on unphased genotype data. The hapFLK statistic is computed as an average over 20 runs of the EM algorithm to fit the LD model. As in [10], we found that the hapFLK distribution could be modelled relatively well with a normal distribution (corresponding to non outlying regions) and a few outliers; we used robust estimation of the mean and standard deviation of the hapFLK statistic to eliminate the influence of outlying (*i.e.* potentially selected) regions. This procedure was done within each group, the resulting mean and standard deviation obtained are given in table S3. Finally, we computed at each SNP a p-value for the null hypothesis from the normal distribution.

Selection in ancestral groups The within-group FLK analysis provides for each SNP an estimation of the allele frequency p_0 in the population ancestral to all populations of the group. We used this information to test SNP for selection using between groups differentiation, with some adjustments. First, the FLK model assumes tested polymorphisms are present in the ancestral population. SNPs for which the alternate allele has been seen in only one population group are likely to have appeared after divergence (within the ancestral tree) and were therefore removed of the analysis. Second, regions selected within groups affect allele frequency in some breeds and therefore bias our estimation of the ancestral allele frequency in this group. We therefore removed all SNPs that were included in within-group selection signatures. Finally, the FLK test requires a rooted population tree. For the within group analysis, we could use a very distant population to the current breeds (the Soay sheep). For the ancestral tree, we created an outgroup homozygous for ancestral alleles at all SNPs.

Identifying selected regions and candidate genes We defined significant regions for each statistic and within each group of populations. Using the neutral distribution (χ^2 for FLK and Normal for hapFLK), we computed the p-value of each statistic at each SNP. To identify selected regions, we estimated their q-value [47] to control the FDR. For FLK, we called significant SNPs with q-values less than 0.1 (therefore controlling the FDR at the 10% level). As the power of hapFLK is greater than that of FLK [10], we used an FDR threshold of 5%. For the FLK analysis in ancestral populations, we used an FDR threshold of 5%.

We then aimed at identifying genes that seem good candidates for explaining selection signatures. We proceeded differently for the single SNP FLK and hapFLK. For FLK, we considered that significant SNP less than 500Kb apart were capturing the same selection signal. Then, we considered as

potential candidate genes any gene that lie less than 500Kb of any significant SNP. For hapFLK, the genome signal is much more continuous than single SNP tests, because the statistic captures multi-point LD with the selected mutations. A consequence is that the significant regions can span large chromosome intervals. To restrict the list of potential candidate genes, and target only the ones closest to the most significant SNP, we restricted our search to the part of the signal where the difference in hapFLK value with the most significant SNP was less than 0.5σ. This allowed to take into consideration the profile of the hapFLK signal, *i.e.* if the profile ressembles a plateau, the candidate region will be rather broad while very sharp hapFLK peaks will provide a narrower candidate region. We listed all the genes present in the significant regions using the OAR3.1 genome browser at `http://www.livestockgenomics.csiro.au/cgi-bin/gbrowse/oarv3.1/`.

Some very likely candidate genes for selection were found in many of the significant regions. This is for example the case of the MSTN (GDF8) gene on chromosome 2 in the NEU group. In these cases, we did not list any other candidates in the region, *i.e.* we made a strong prior assumption of selection for these genes. Note however that we provide the position of the selected regions for the reader interested in knowing all the genes present in significant regions.

Acknowledgments

The ovine SNP50 HapMap data set used for the analyses described was provided by the International Sheep Genomics Consortium (ISGC) and obtained from http://www.sheephapmap.org in agreement with the ISGC Terms of Access. Data analyses were performed on the computer cluster of the bioinformatics platform Toulouse Midi-Pyrenees.

References

1. Andersson L (2012) How selective sweeps in domestic animals provide new insight into biological mechanisms. J Intern Med 271: 1-14.

2. Moradi MH, Nejati-Javaremi A, Moradi-Shahrbabak M, Dodds KG, McEwan JC (2012) Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. BMC Genet 13: 10.

322 3. Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, et al. (2009) A genome wide survey
323 of SNP variation reveals the genetic structure of sheep breeds. PLoS One 4: e4668.

324 4. Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, et al. (2012) Genome-wide analysis of
325 the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS
326 Biol 10: e1001258.

327 5. Garcia-Gámez E, Reverter A, Whan V, McWilliam SM, Arranz JJA, et al. (2011) Using regulatory
328 and epistatic networks to extend the findings of a genome scan: identifying the gene drivers of
329 pigmentation in merino sheep. PLoS ONE 6: e21158.

330 6. Becker D, Tetens J, Brunner A, Burstel D, Ganter M, et al. (2010) Microphthalmia in Texel sheep
331 is associated with a missense mutation in the paired-like homeodomain 3 (PITX3) gene. PLoS One
332 5: e8689.

333 7. Zhang L, Mousel MR, Wu X, Michal JJ, Zhou X, et al. (2013) Genome-Wide Genetic Diversity
334 and Differentially Selected Regions among Suffolk, Rambouillet, Columbia, Polypay, and Targhee
335 Sheep. PLoS One 8: e65942.

336 8. Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured
337 population. Heredity (Edinb) 103: 285-98.

338 9. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, et al. (2010) Detecting selection in
339 population trees: the Lewontin and Krakauer test extended. Genetics 186: 241-62.

340 10. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection
341 through haplotype differentiation among hierarchically structured populations. Genetics 193: 929-
342 41.

343 11. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus
344 genotype data. Genetics 155: 945-59.

345 12. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated
346 individuals. Genome Res 19: 1655-64.

13. Johnston SE, McEwan JC, Pickering NK, Kijas JW, Beraldi D, et al. (2011) Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. Mol Ecol 20: 2555-66.

14. Dominik S, Henshall JM, Hayes BJ (2012) A single nucleotide polymorphism on chromosome 10 is highly predictive for the polled phenotype in Australian Merino sheep. Anim Genet 43: 468-70.

15. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, et al. (2010) Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci U S A 107: 1160-5.

16. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, et al. (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res 15: 936-44.

17. Jacobs LC, Wollstein A, Lao O, Hofman A, Klaver CC, et al. (2013) Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans. Hum Genet 132: 147-58.

18. Hauswirth R, Haase B, Blatter M, Brooks Sa, Burger D, et al. (2012) Mutations in MITF and PAX3 cause "splashed white" and other white spotting phenotypes in horses. PLoS genetics 8: e1002653.

19. Lin JY, Fisher DE (2007) Melanocyte biology and skin pigmentation. Nature 445: 843–50.

20. Klungland H, Vå ge DI, Gomez-Raya L, Adalsteinsson S, Lien S (1995) The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. Mammalian genome 6: 636–9.

21. Joerg H, Fries HR, Meijerink E, Stranzinger GF (1996) Red coat color in Holstein cattle is associated with a deletion in the MSHR gene. Mammalian genome 7: 317–8.

22. Linnen CR, Poh YP, Peterson BK, Barrett RDH, Larson JG, et al. (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339: 1312–6.

23. Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, et al. (2007) A single IGF1 allele is a major determinant of small size in dogs. Science 316: 112-5.

24. Kuijper S, Feitsma H, Sheth R, Korving J, Reijnen M, et al. (2005) Function and regulation of Alx4 in limb development: complex genetic interactions with Gli3 and Shh. Developmental biology 285: 533–44.

25. Qu S, Tucker SC, Ehrlich JS, Levorse JM, Flaherty LA, et al. (1998) Mutations in mouse Aristaless-like4 cause Strong's luxoid polydactyly. Development 125: 2711-21.

26. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, et al. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140: 744 - 752.

27. Stickens D, Zak BM, Rougier N, Esko JD, Werb Z (2005) Mice deficient in Ext2 lack heparan sulfate and develop exostoses. Development 132: 5055–68.

28. Person AD, Beiraghi S, Sieben CM, Hermanson S, Neumann AN, et al. (2010) WNT5A mutations in patients with autosomal dominant Robinow syndrome. Developmental dynamics 239: 327–37.

29. Noguchi Ki, Watanabe Y, Fuse T, Takizawa M (2010) A new chondrogenic differentiation initiator with the ability to up-regulate sox trio expression. Journal of Pharmacological Sciences 112: 89-97.

30. Stattin EL, Wiklund F, Lindblom K, Onnerfjord P, Jonsson BA, et al. (2010) A missense mutation in the aggrecan C-type lectin domain disrupts extracellular matrix interactions and causes dominant familial osteochondritis dissecans. American journal of human genetics 86: 126–37.

31. Tompson SW, Merriman B, Funari Va, Fresquet M, Lachman RS, et al. (2009) A recessive skeletal dysplasia, SEMD aggrecan type, results from a missense mutation affecting the C-type lectin domain of aggrecan. American journal of human genetics 84: 72–9.

32. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nature genetics 40: 575–83.

33. Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, et al. (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet 38: 813-8.

34. Fleming-Waddell JN, Olbricht GR, Taxis TM, White JD, Vuocolo T, et al. (2009) Effect of DLK1 and RTL1 but not MEG3 or MEG8 on muscle gene expression in Callipyge lambs. PloS one 4: e7399.

35. Bouwman AC, Bovenhuis H, Visker MH, van Arendonk JA (2011) Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet 12: 43.

36. Rincon G, Islas-Trejo A, Castillo AR, Bauman DE, German BJ, et al. (2012) Polymorphisms in genes in the SREBP1 signalling pathway and SCD are associated with milk fatty acid composition in Holstein cattle. J Dairy Res 79: 66-75.

37. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, et al. (2009) Coat variation in the domestic dog is governed by variants in three genes. Science 326: 150-3.

38. Ponz R, Moreno C, Allain D, Elsen JM, Lantier F, et al. (2001) Assessment of genetic variation explained by markers for wool traits in sheep via a segment mapping approach. Mamm Genome 12: 569-72.

39. Bidinost F, Roldan D, Dodero A, Cano E, Taddeo H, et al. (2007) Wool quantitative trait loci in merino sheep. Small Ruminant Research 74: 113 - 118.

40. Bautista DM, Siemens J, Glazer JM, Tsuruda PR, Basbaum AI, et al. (2007) The menthol receptor TRPM8 is the principal detector of environmental cold. Nature 448: 204–8.

41. Rubin CJ, Zody MC, Eriksson J, Meadows JRS, Sherwood E, et al. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464: 587–91.

42. Nakao N, Ono H, Yamamura T, Anraku T, Takagi T, et al. (2008) Thyrotrophin in the pars tuberalis triggers photoperiodic response. Nature 452: 317-22.

43. Balasse M, Tresset A (2007) Environmental constraints on the reproductive activity of domestic sheep and cattle : what latitude for the herder ? Anthropozoologica 42: 71–88.

44. Heaton MP, Clawson ML, Chitko-Mckown CG, Leymaster Ka, Smith TPL, et al. (2012) Reduced lentivirus susceptibility in sheep with TMEM154 mutations. PLoS genetics 8: e1002467.

45. Oganesyan G, Saha SK, Guo B, He JQ, Shahangian A, et al. (2006) Critical role of TRAF3 in the Toll-like receptor-dependent and -independent antiviral response. Nature 439: 208–11.

46. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629-44.

427 47. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad
428 Sci U S A 100: 9440-5.

Figures Legends

Figure 1. Localisation of selection signatures identified in 7 groups of populations. Candidate genes are indicated above their genomic localisation. Only chromosomes harboring selection signatures are plotted.

Figure 2. Phylogenetic tree of the ancestral populations of geographical groups.

Figure 3. Genome scan for selection signature in ancestral populations of the geographical groups. Significant SNPs at the 5% FDR level are plotted in darker color.

Tables

Table 1. List of genome regions corresponding to selection signatures. Regions identified with the hapFLK and FLK test, with the corresponding population group and most differentiated populations (except for the AFR group). Overlapping regions in different groups or with different tests are grouped by background color. †: signatures of selection previously identified [4]. ‡: this outlying region is not due to evolutionary processes (see details in the main text). Full names of groups and populations are given in Table S1.

OAR	Begin (Mbp)	End (Mbp)	P-value	Q-value	Group	Test	Cand. gene	Diff. pop.
2	46.65	57.99	6.3e-10	7.1e-07	ITA	hapFLK	NPR2†	COM
2	51.41	53.44	4.1e-09	1.6e-04	ITA	FLK		COM
2	74	74.86	7.4e-04	3.7e-02	ITA	hapFLK		COM
2	81.27	87.32	4.1e-09	2.3e-06	ITA	hapFLK	BNC2	COM
2	110.08	112.08	1.5e-05	6.7e-02	ASI	FLK		SUM TIB GUR
2	113.36	122.24	7.0e-06	3.3e-03	NEU	hapFLK	MSTN†	GTX NTX STX
2	239.76	241.76	2.9e-05	9.3e-02	SWA	FLK	RH locus	AFS
3	84.4	86.4	2.5e-05	9.1e-02	ASI	FLK		–
3	120.91	125.49	5.3e-04	3.0e-02	ITA	hapFLK	KITLG	COM
3	122.07	130.85	6.8e-08	4.2e-04	AFR	hapFLK		
3	151.42	156.93	3.3e-16	3.1e-12	ITA	hapFLK	HMGA2†	COM SAB
3	154.79	154.93	5.9e-04	4.3e-02	AFR	hapFLK		
3	159.64	161.6	6.1e-04	3.3e-02	ITA	hapFLK		COM
3	167.85	171.67	1.5e-04	1.3e-02	ITA	hapFLK	IGF1	COM ALT SAB
4	4.61	6.61	5.3e-06	2.1e-02	SWA	FLK		MOG

Table 1 – continued from previous page

4	8.5	19.66	4.2e-06	1.1e-03	CEU	hapFLK		VBS VRS
4	15.11	17.11	8.4e-07	1.5e-02	CEU	FLK		VBS
4	26.46	28.46	2.4e-05	9.1e-02	ASI	FLK	HDAC9	GUR IDC SUM
4	44.49	45.76	2.7e-04	3.4e-02	NEU	hapFLK		NZR
4	45.57	47.57	1.8e-06	2.4e-02	ASI	FLK		SUM
4	67.75	69.8	3.5e-07	2.3e-03	SWA	FLK	HOXA	MOG
5	29.4	31.4	1.1e-05	6.7e-02	ASI	FLK		GAR
5	47.35	49.35	1.4e-05	6.7e-02	ASI	FLK		BGA
5	78.16	78.76	4.2e-04	4.2e-02	NEU	hapFLK		NZT
6	5.62	7.62	3.1e-06	6.0e-02	ITA	FLK		SAB
6	33.22	41.02	3.4e-08	8.0e-05	SWE	hapFLK	ABCG2†	LAC LAM
6	34.71	39.12	1.6e-07	4.1e-05	ITA	hapFLK		COM
6	35.94	38.31	2.1e-04	1.9e-02	CEU	hapFLK		VRS VBS
6	67.98	70.36	4.3e-06	1.1e-03	CEU	hapFLK	KIT†	VBS
6	68.9	70.95	9.6e-07	5.3e-03	SWA	FLK		
6	93.3	94.39	3.8e-04	2.7e-02	CEU	hapFLK	FGF5†	(VRS&VBS) or (ERS&BOS)
7	49.15	51.15	1.1e-05	9.7e-02	CEU	FLK		VRS
7	78.31	80.31	8.1e-07	1.5e-02	CEU	FLK		VRS ERS
8	23.97	25.97	2.9e-05	9.6e-02	ASI	FLK		TIB
9	29.46	31.55	3.7e-04	3.4e-02	SWE	hapFLK		CHU MER
9	37.79	46.03	1.9e-05	6.2e-03	NEU	hapFLK		NZT ISF

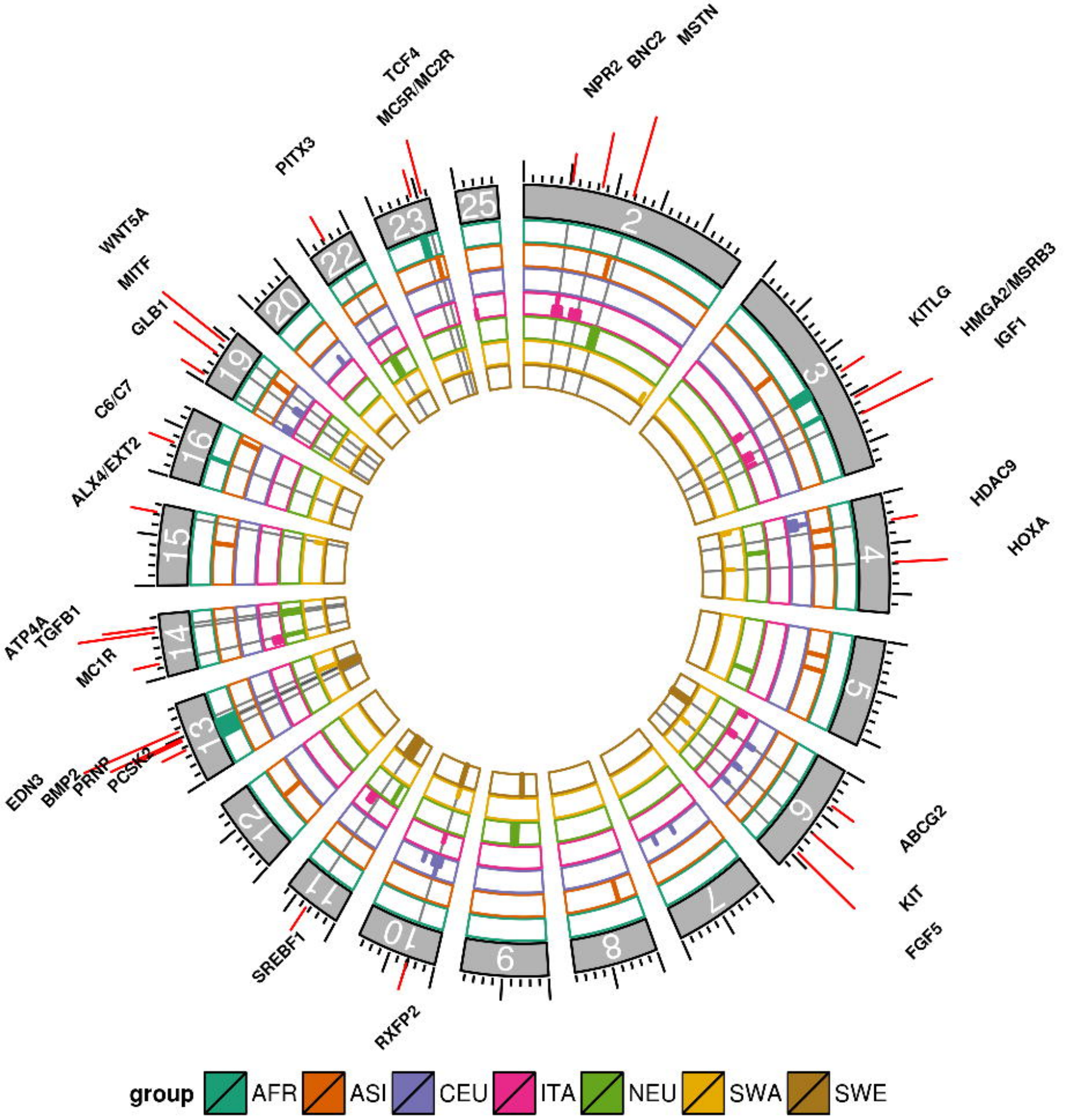
Table 1 – continued from previous page

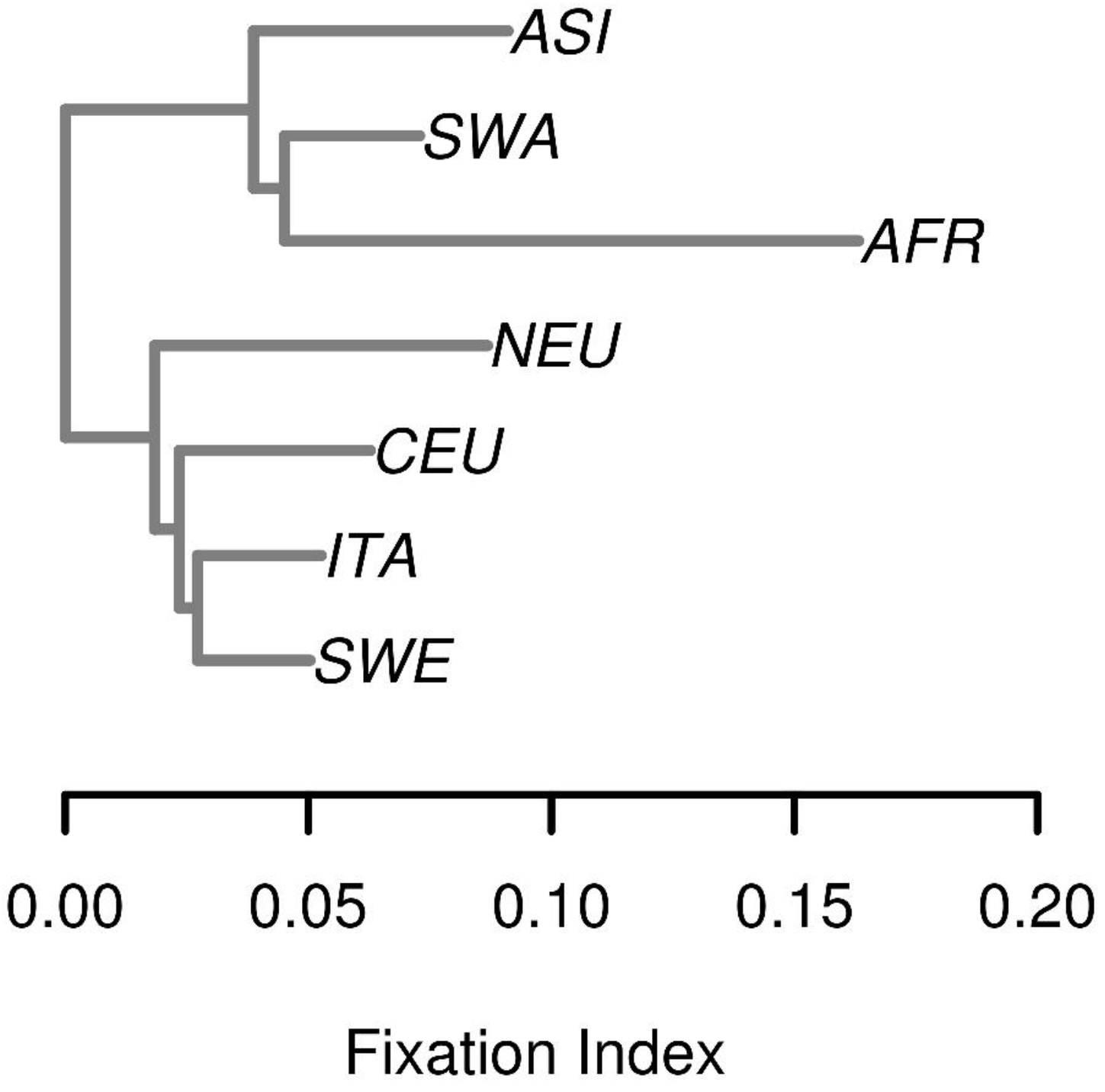
10	24.02	34.91	1.4e-14	1.1e-10	CEU	hapFLK	RXFP2†	BOS ERS VRS
10	29.42	29.71	9.6e-04	4.4e-02	ITA	hapFLK		COM ALT
10	28.5	30.5	6.3e-06	7.5e-02	CEU	FLK		BOS ERS
10	28.5	30.5	3.2e-05	9.7e-02	SWA	FLK		NDZ
10	28.5	30.5	1.3e-06	5.4e-02	SWE	FLK		MER
10	48.9	49.59	5.2e-04	3.1e-02	CEU	hapFLK		–
11	12.55	14.12	1.4e-04	2.2e-02	NEU	hapFLK		
11	24.18	38.74	9.8e-09	8.0e-05	SWE	hapFLK	SREBP1	LAC MER
11	40.31	46.7	3.3e-06	5.5e-04	ITA	hapFLK		SAB
12	42.66	44.66	3.4e-07	7.6e-03	ASI	FLK		SUM
13	33.1	40.02	5.7e-06	1.8e-03	AFR	hapFLK	PCSK2	
13	40.6	50.3	4.9e-07	4.9e-04	AFR	hapFLK	BMP2†	
13	43.34	51.28	2.7e-07	1.7e-04	SWE	hapFLK	PRNP	LAC LAM
13	56.11	57.17	2.5e-08	4.8e-04	SWA	hapFLK	EDN3	MOG
13	55.33	57.43	8.4e-11	1.1e-06	SWA	FLK		MOG
14	6.37	13.6	1.6e-04	1.4e-02	ITA	hapFLK		SAB
14	13.64	13.7	5.3e-04	4.9e-02	NEU	hapFLK	MC1R	ISF
14	13.7	16.46	1.2e-04	1.1e-02	ITA	hapFLK		SAB
14	45.49	50.09	1.6e-04	2.5e-02	NEU	hapFLK	TGFB1	NTX NZR
15	48.87	50.87	1.5e-05	6.7e-02	ASI	FLK		GAR IDC
15	71.71	73.71	3.8e-06	1.6e-02	SWA	FLK	ALX4 EXT2	MOG

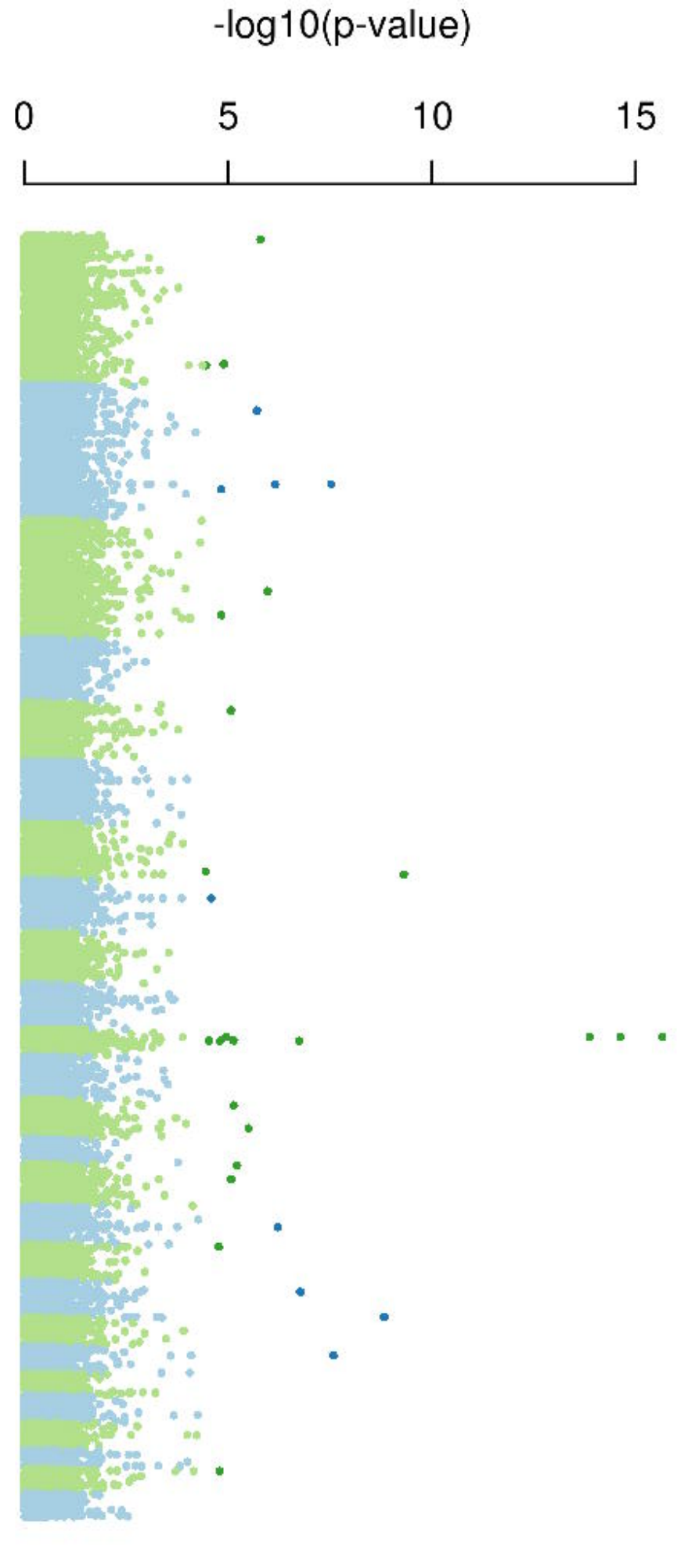
Table 1 – continued from previous page

16	33.2	35.1	1.8e-04	1.8e-02	AFR	hapFLK	C6/C7	
16	63.97	65.97	1.1e-05	6.7e-02	ASI	FLK		GAR
								IDC
19	4.42	7.43	2.2e-04	1.9e-02	CEU	hapFLK	GLB1†	VRS
								BOS
19	30.42	35.09	3.2e-05	4.2e-03	CEU	hapFLK	MITF†	VBS
								BOS
								ERS
19	44.6	46.6	3.9e-06	3.9e-02	ASI	FLK	WNT5A	GAR
								BGA
20	36.74	38.52	2.8e-04	2.3e-02	CEU	hapFLK		VRS
22	18.9	24.36	1.5e-11	7.4e-08	NEU	hapFLK	PITX3‡	GTX
23	42.5	46.96	2.2e-05	5.4e-03	AFR	hapFLK	MC5R	
							MC2R	
23	54.14	56.14	3.8e-07	7.6e-03	ASI	FLK		GAR
25	0.08	3.08	3.7e-04	2.4e-02	ITA	hapFLK		SAB

		Estimated ancestral allele frequencies									
chr	pos	AFR	ASI	SWA	NEU	CEU	ITA	SWE	P-value	Q-value	candidate gene
1	7192190	0.15	0.08	0.16	0.55	0.69	0.04	0.38	1.7e-06	5.3e-03	TRPM8
1	237070498	0.87	0.95	0.91	0.48	0.24	0.77	0.35	1.4e-05	2.5e-02	GYG1
1	239424807	0.46	0.68	0.06	0.21	0.15	0.11	0.17	3.4e-05	4.8e-02	
1	239491620	0.53	0.41	0.94	0.86	0.93	0.93	0.88	4.3e-05	5.6e-02	
2	45500785	0.43	0.91	0.23	0.76	0.87	0.87	0.93	2.2e-06	6.4e-03	LPL
2	182607165	0.99	0.97	0.18	0.64	0.73	0.83	0.64	3.4e-08	1.8e-04	INSIG2
2	182672296	0.99	0.94	0.32	0.9	0.86	0.89	0.81	7.7e-07	2.8e-03	
2	192231314	0.59	0.93	0.36	0.96	0.89	0.81	0.95	1.6e-05	2.8e-02	
3	132478420	0.24	0.89	0.18	0.93	0.81	0.84	0.82	1.2e-06	3.9e-03	HOXC †
3	180860403	0.71	0.53	0.28	0.82	0.31	0.12	0.13	1.7e-05	2.8e-02	
5	15522700	0.68	0.63	0.92	0.27	0.76	0.99	0.78	9.8e-06	2.0e-02	
7	89519883	0.63	0.61	0.19	0.89	0.18	0.6	0.95	6.1e-10	5.2e-06	TSHR †
8	31748642	0.84	0.93	0.94	0.16	0.63	0.47	0.19	2.8e-05	4.1e-02	PREP/BVES †
11	18248852	0.35	0.32	0.82	0.64	0.94	0.96	0.92	1.3e-05	2.5e-02	NF1 †
11	18325488	0.87	0.93	0	0.35	0.04	0.03	0.04	3.3e-16	7.2e-12	
11	18335747	0.87	0.93	0	0.35	0.04	0.03	0.04	3.3e-16	7.2e-12	
11	18433474	0.87	0.93	0.02	0.35	0.07	0.02	0.05	3.8e-15	5.4e-11	
11	18440783	0.78	0.93	0.02	0.34	0.07	0.02	0.05	2.0e-14	2.2e-10	
11	25704651	0.97	0.96	0.97	0.42	0.94	0.94	0.96	8.5e-06	1.9e-02	
11	26284826	0.99	0.97	0.94	0.38	0.93	0.95	0.79	3.2e-05	4.6e-02	
11	26571629	0.92	0.94	0.98	0.29	0.89	0.88	0.86	1.8e-05	2.8e-02	
11	26872280	0.78	0.71	0.93	0.15	0.89	0.9	0.9	2.2e-07	9.5e-04	
13	12120674	0.29	0.84	0.97	0.91	0.97	0.92	0.84	7.7e-06	1.8e-02	GATA3
13	62857560	0.52	0.62	0.65	0.98	0.67	0.92	0.36	3.6e-06	9.7e-03	ASIP †
15	3706790	0.71	0.22	0.96	0.28	0.27	0.34	0.21	6.8e-06	1.7e-02	
15	29856310	0.98	0.99	0.99	0.47	0.92	0.95	0.96	9.8e-06	2.0e-02	
16	38696505	0.95	0.98	0.95	0.99	0.68	0.31	0.3	6.8e-07	2.7e-03	PRLR †
17	4867509	0.91	0.95	0.85	0.54	0.18	0.58	0.17	1.8e-05	2.8e-02	TMEM154
18	19342316	0.9	0.79	0.67	0.35	0.75	0.1	0.09	1.9e-07	9.3e-04	ACAN †
18	66470371	0.99	0.97	0.9	0.9	0.18	0.04	0.08	1.9e-09	1.3e-05	TRAF3
20	17381047	0.24	0.61	0.97	0.98	0.93	0.99	0.91	3.1e-08	1.8e-04	VEGFA †
25	7517270	0.95	0.94	0.93	0.14	0.27	0.57	0.19	1.8e-05	2.8e-02	wool QTL †


TCF4
MC5R/MC2R
NPR2
BNC2
MSTN
PITX3
WNT5A
MITF
GLB1
KITLG
HMGA2/MSRB3
IGF1
C6/C7
ALX4/EXT2
HDAC9
HOXA
ATP4A
TGFB1
MC1R
EDN3
BMP2
PRNP
PCSK2
ABCG2
KIT
FGF5
SREBF1
RXFP2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
19
20
22
23
25
group
AFR
ASI
CEU
ITA
NEU
SWA
SWE


ASI
SWA
AFR
NEU
CEU
ITA
SWE
0.00
0.05
0.10
0.15
0.20
Fixation Index


-log10(p-value)
0
5
10
15