Keep Garfagnina alive. An integrated study on patterns of homozygosity, genomic inbreeding, admixture and breed traceability of the Italian Garfagnina goat breed.

Christos Dadousis[1,*], Francesca Cecchi[2], Michela Ablondi[1], Maria Chiara Fabbri[3], Alessandra Stella[4,5], Riccardo Bozzi[3]

[1] Dipartimento di Scienze Medico-Veterinarie, Università di Parma, Parma, Italy.
[2] Dipartimento di Scienze Veterinarie, Università di Pisa, Italy.
[3] Dipartimento di Scienze e Tecnologie Agrarie, Alimentari, Ambientali e Forestali, Università di Firenze, Firenze, 50144, Italy.
[4] Istituto di Biologia e Biotecnologia Agraria, Consiglio Nazionale delle Ricerche, Milano, Italy.
[5] Fondazione Parco Tecnologico Padano, Lodi, Italy.

* Corresponding author:
Email: christos.dadousis@unipr.it (CD)
ORCID ID: https://orcid.org/0000-0002-7689-6443

Abstract

The objective of this study was to investigate the genomic background of the Garfagnina (GRF) goat breed that faces the risk of extinction. In total, 48 goats genotyped with the Illumina CaprineSNP50 BeadChip were analyzed together with 214 goats belonging to 9 Italian breeds (~25 goats/breed) from the AdaptMap project [Argentata (ARG), Bionda dell'Adamello (BIO), Ciociara Grigia (CCG), Di Teramo (DIT), Garganica (GAR), Girgentana (GGT), Orobica (ORO), Valdostana (VAL) and Valpassiria (VSS)]. We estimated i) runs of homozygosity (ROH), ii) admixture ancestries and iii) traceability success via discriminant analysis on principal components (DAPC) based on cross-validation. For GRF, an excess of frequent ROH (more than 45% in the GRF samples analyzed) was detected on CHR 12 at, roughly 50.25-50.94Mbp (ARS1 assembly), spanned between the CENPJ (centromere protein) and IL17D (interleukin 17D) genes. The same area was also present in DIT, while the broader region (~49.25-51.94Mbp) was shared among the ARG, CCG, and GGT. Admixture analysis depicted the uniqueness of the GRF breed, with a small part of common ancestry shared with BIO, VSS, ARG and CCG breeds. The DAPC model resulted in a 100% assignment success. We hope this work will contribute to the efforts of preventing the GRF from extinction and to add value to all the socio-agro-economic factors related with the farming of the GRF breed.

Introduction

Local breeds, that usually consist of a small number of animals, are increasingly recognized by E.U. action plans as a rule of rural land protection. There are several

reasons for this. To name some, local breeds are i) rustic and resistant to their local environment, ii) they represent a significant economic resource and have been used for the manufacture of niche products, especially in mountainous regions, iii) they represent an important and usually unique gene bank that could be essential to address the future climate changes, or potential disease outbreaks [1], and iv) they play an important role for the preservation of the human cultural inheritance. Especially for small ruminants, that can adapt in marginal and difficult areas, another reason of major importance is the provision of eco-system services. In mountainous regions of the Mediterranean basin, grazing can be used as a measure of protection against avalanches in winter and fire outbreaks during the summer period. Grazing, apart from cost-effective is also a nonpolluting, nontoxic, nearly carbon-neutral and an effective technique against fire propagation. In this context, goat grazing has been proposed as an alternative and eco-friendly solution [2]. Moreover, climate change has been identified as an additional pressure to the sustainability of livestock systems (e.g., health and productivity) and local breeds provide with an alternative through the adaptiveness in the regions they are reared. Despite this, usually the low productivity of local unimproved breeds, and thereby low farmer's income, endangers their existence.

Regarding goats (*Capra hircus*), their widespread presence and adaptation in a variety of agro-ecological conditions worldwide is well documented [3]. Goats, are closely related to the human kind history, since, together with sheep, cattle and pigs were from the earliest domesticated ungulates [3,4]. Nevertheless, based on the Domestic Animal Diversity Information System (DAD-IS) data, 21 goat breeds are extinct (18 from which were reared in the regions of Europe and Caucasus) and 41 are at critical situation (41 of

which from Europe and Caucasus) [6]. In Italy, 3 goat breeds are extinct and 12 are marked in a critical situation.

The Garfagnina breed (GRF) is one of those breeds that faces the risk of extinction. The latest recognition reported 1,468 animals spread in 29 different farms (ARAT, personal communication). The GRF is reared, mostly for dairy production, in central Italy, in the hills and mountains of the northwestern Tuscan Apennine area. The origins of this population are not clear. However, it is very likely that the breed was a result of crossings between native goats from Alps and from the Tuscan-Emilian Apennines. Moreover, the local breeders report that the population was reared for generations for its milk and meat production [4]. The breed is also closely linked to the production of typical products, such as the Controneria meat kid and the Caprino delle Apuane cheese. As it has been reported by Martini et al. [4], the milking of GRF goats is manual.

To support the management and conservation of the breed, and to provide support to the farmers and to the general region where the breed is reared (hills and mountains of the northwestern Tuscan Apennine area - central Italy), a few studies investigated various production characteristics [4,7], the adaptive profile (through physiological, haematological, biochemical and hormonal parameters) [8] and resistance to diseases [9,10] of GRF. Martini et al. [4] investigated various zootechnical characteristics of the GRF breed in comparison to other Italian and foreign goat breeds. Based on their results, the authors suggested the development of a breeding scheme based on pure bred animals. Nevertheless, no whole genome analysis has been conducted yet to investigate the genomic background of GRF and its ancestry. Genomic information, however, is essential for action measures to be taken for conservation purposes. The 50K SNP chip

(http://www.goatgenome.org; [11]) released in 2013 together with the recent results of the AdaptMap project [3] offered this opportunity.

Hence, the objective of the present study was to investigate the genomic background of the Garfagnina breed, relative to the native Italian goat breeds included in the AdaptMap dataset. A unified procedure on admixture, runs of homozygosity, and discriminant analysis was applied with results depicting the unique genetic structure of the breed.

Materials and methods

Ethics statement

Garfagnina goats belonged to commercial farms and blood sampling was conducted by veterinarians. No invasive procedures were applied. Thus, in accordance to the 2010/63/EU guide and the adoption of the Law D.L. 04/03/2014, n.26 by the Italian Government, an ethical approval is not required in our study.

Genomic data

Blood samples of forty eight female GRF goats were collected and animals were genotyped with the Illumina GoatSNP50 BeadChip (Illumina Inc., San Diego, CA) containing 53,347 Single Nucleotide Polymorpishms (SNPs) [12]. Genomic data of nine Italian autochthonous goat breeds, namely Argentata dell'Etna (ARG), Bionda dell'Adamello (BIO), Ciociara Grigia (CCG), Di Teramo (DIT), Garganica (GAR), Girgentana (GGT), Orobica (ORO), Valdostana (VAL) and Valpassiria (VSS) were downloaded from the online repository

(https://datadryad.org/stash/dataset/doi:10.5061/dryad.v8g21pt) of the ADAPTmap project [3,13]. The breeds were selected based on the breed abbreviation on the plink fam file downloaded from the repository and the breed description (code and country) reported in Table 1 of [13]. The two datasets were merged and quality control was conducted in PLINK v1.9 [3, 4] on the final dataset based on the following criteria: 1) only autosomes were kept, ii) call rate per SNP >95% and ii) missing values per sample <10%. After editing, 260 samples and 48,716 SNP were retained (Table 1). The distribution of the SNP per chromosome (CHR) is presented in S1 Fig.

Table 1. Name of breeds, breed code and number of animals analyzed before (pre-QC) and after (post-QC) quality control per breed.

Breed name	Breed code	No. pre-QC	No. post-QC
Argentata dell'Etna	ARG	25	24
Bionda dell'Adamello	BIO	24	24
Ciociara Grigia	CCG	19	19
Di Teramo	DIT	24	24
Garganica	GAR	20	20
Girgentana	GGT	30	30
Garfagnina	GRF	48	48
Orobica	ORO	24	23
Valdostana	VAL	24	24
Valpassiria	VSS	24	24

Runs of homozygosity

Analysis of runs of homozygosity (ROH) was conducted in the R (*v. 3.5.0*) package *detectRUNS v. 0.9.5* [16] using the consecutive method [17] that runs under the main function *consecutiveRUNS.run*. The required parameters were set to: i) minimum number of 15 SNPs/ROH, ii) 1 Mbp minimum length of ROH and iii) allow one heterozygous SNP

within an ROH (to account for genotyping errors). In addition, ROH lengths were split into five classes (0-2, 2-4, 4-8, 8-16 and >16 Mbp). For each of the class and breed, descriptive statistics of ROH per breed, per chromosome, per SNP and per length class were estimated. Principal component analysis (PCA) was used to identify (dis)similarities among breeds, relative to the average number of ROH identified per chromosome. In addition, genomic inbreeding (F_{ROH}) was calculated per breed. Regions with an excess of frequent ROH ($\geq 45\%$) were detected and surrounding genes (1 Mbp up/downstream) were identified using the *Capra hircus* ARS1 (http://www.ensembl.org/index.html) and the variant effect predictor (https://www.ensembl.org/Tools/VEP) Ensembl databases.

Population stratification and ancestry

PCA and admixture analysis were used to infer the presence of distinct populations based on the genomic data. The proportion of mixed ancestry in the breeds was assessed by the *ADMIXTURE 1.22* software [7, 8]. The number of ancestries (K) to be retained in the admixture analysis (K = 2 to 10) was evaluated via a 10-fold cross-validation (CV). The final selection on the number of ancestries was done by inspecting the CV error.

Discriminant analysis of principal components

Discriminant analysis was applied to assess the traceability of the GRF goats using genomic data. To achieve this, the methodology of discriminant analysis of principal components (DAPC) [20] implemented in the R package *adegenet* [5, 10, 11] was adopted. In brief, DAPC is a 2-step approach: firstly, a PCA on the matrix of the genotypes is conducted and then, a small number of selected PCs (instead of the original SNP

genotypes) is used as an input for the linear discriminant analysis (LDA). The selection on the optimal number of PCs to be further used in the LDA is done via cross-validation (CV) were the data is split in training and validation sets. For the selection of PCs the following criteria were implemented: i) 10-fold CV with 30 repetitions, ii) a maximum number of 300 PCs were tested, and iii) the number of PCs to be retained was based on number of PCs associated with the highest mean success. Three different scenarios of DAPC were applied as described below:

1. Scenario 1 (supervised learning). The full dataset was analysed simultaneously. In this scenario, all available data were used for model training and the discriminant functions were extracted based on all animals. This is not, however, a real case scenario, since the discriminant functions were developed utilizing the entire data set. The objective for a practical application is to identify an external individual membership to a group (i.e. external validation). Hence, two more scenarios were developed adopting a CV scheme also for the discriminant function.

2. Scenario 2 (semi-supervised learning). Assessment of correct assignment of GRF goats was done via a semi-supervised CV (CV_{SS}). Five GRF goats were sampled representing the testing set of the DAPC analysis. The reference population was constituted by the rest of 43 GRF goats plus all the goats from the other breeds. The five GRF samples were classified in one of the 10 breeds presented in the reference population via the function *predict.dapc*. The procedure was repeated 10 times and results were averaged over the 10 repetitions.

3. Scenario 3 (unsupervised learning). Assignment of GRF goats in a breed but without the presence of any GRF goats in the reference population and model training (unsupervised CV; CV_{US}). This scenario could also be viewed as a method to assess the genomic similarity of the GRF with the rest of the breeds (i.e., type of clustering). The approach was similar to Scenario 2 other than the testing population consisted of the entire GRF set and GRF samples had to be classified in one or more of the other 9 breeds. To increase the number of the tested samples in each round of the CV, 80% of the GRF breed was sampled. Moreover, to test for the effect of the size in training the model (TRN) in the assignment of the GRF, different proportions of the reference population were sampled (20, 30, …, 90%) 10 times each, and results were averaged over the 10 repetitions. In other words, the size of the reference population varied between 42 to 91 goats. All nine breeds were present in each scenario and all GRF goats were used in this scenario.

It should be noted that the terms (semi/un)-supervised should not be confused with the terminology used in machine learning. The introduction of these terms has been used in the manuscript to distinguish among the three approaches that have been used in the DAPC analysis, and, although they are, up to a point, analogous with the same terms used in machine learning they are not identical.

Results

Runs of Homozygosity

Summary results of the detected ROH regions as total counts or averaged based on the number of samples per breed are presented in Table 2 and Fig 1, respectively. A relative high number of ROH was detected for GRF (n=2,450), with the highest number being for GGT (n=2,762) followed by ORO (=2,693), while the smallest was found for ARG (n=465). For GRF, the number of ROH per *Capra hircus* chromosome (CHI) varied from 35 (CHI25) to 158 (CHI1). The maximum length of ROH per chromosome was found on CHI1 (568,887,711bp) and the minimum on CHI23 (113,323,345bp). In general, the total length of ROH per CHR followed the same pattern of the total ROH number per CHR (Fig 2).

For all breeds analyzed, except CCG and DIT, the number of ROH relative to the length on the genome was decreasing with an increased length (Fig 3α). In the DIT samples, the ROH were more frequent in length classes of 4-8 and 8-16 Mbp compared to the 2-4Mbp. The percentage of ROH with a length >16Mbp per breed varied between 0.89% to 13.53% for ORO and DIT, respectively. For small ROH length (<2Mbp) the proportion over the total number detected reached ~78% in ARG, while only 35% of ROH was observed for DIT. The pattern of ROH length class was similar among GRF, GGT and ORO with ~50% of the ROH having a length < 2Mbp, ~25% between 2-4Mbp, ~13% between 4-8Mbp, ~5% and ~2% between 8-16Mbp and > 16Mbp, respectively (S1 Table).

For GRF, an excess of frequent ROH (more than 45% in the GRF samples analyzed) was detected on CHI12, between ~34.6-35.3 Mbp (Fig 4, Table 3). In total, 14 SNP were contained in this region. The same excess area was also present in the DIT breed, while

197 the broader region (~33,9-36.5 Mbp) was shared among the ARG, CCG, and GGT
198 breeds. To identify similarities among breeds relative to the number of ROH per
199 chromosome, a PCA was conducted on the average number of ROH identified per
200 chromosome. In addition, a heatmap on the actual number of ROH per chromosome was
201 produced. Both approaches placed GRF closer to ORO and GGT in respect to the rest of
202 the breeds (Fig 5).

203
204 **Table 2. Total number of runs of homozygosity (ROH) detected per breed.**

Breed	No. ROH
ARG	465
BIO	814
CCG	496
DIT	813
GAR	730
GGT	2,762
GRF	2,450
ORO	2,693
VAL	1,613
VSS	768

205

206 **Fig 1**. **Average number of runs of homozygosity (ROH) detected per breed.**
207 ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di
208 Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL:
209 Valdostana and VSS: Valpassiria.

210

211 **Fig 2. Number and length of runs of homozygosity (ROH) per chromosome (CHR)**
212 **in Garfagnina breed.**

213

Fig 3. Frequency distribution of the number of runs of homozygosity (ROH) (a), in the breeds analyzed per length class, and (b) in different length classes per breed. ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

Fig 4. Number of times (%) each SNP was detected inside a run of homozygosity (ROH) in Garfagnina (GRF) goats.

Fig 5. a) Scatterplot of principal component analysis conducted on the average runs of homozygosity identified per chromosome and breed; b) heatmap on the number of runs of homozygosity identified per chromosome and breed. ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

Table 3. Most common (≥ 45% in each breed) runs of homozygosity (ROH) detected per breed on *Capra hircus* chromosome 12, with the start-end regions and number of SNP per ROH.

Breed	Start-SNP	End-SNP	No. SNP	Start-region (bp)	End-region (bp)
ARG	snp30406-scaffold335-807385/ rs268262652	snp30399-scaffold335-501928/ rs268262645	6	34,949,980	35,255,437
CCG	snp30421-scaffold335-1558038/ rs268262666	snp30391-scaffold335-171300/ rs268262637	27	34,199,327	35,586,065
DIT	snp30413-scaffold335-1113038/ rs268262659	snp30397-scaffold335-418126/ rs268262643	14	34,644,327	35,339,239
GGT	snp30428-scaffold335-1839052/	snp11142-scaffold140-760668/	53	33,918,313	36,518,132

	rs268262673	rs268243983			
GGT	snp17454-scaffold1805-39262/ rs268250096	snp55342-scaffold855-361968/ rs268286936	26	40,620,364	42,257,561
GRF	snp30413-scaffold335-1113038/ rs268262659	snp30397-scaffold335-418126/ rs268262643	14	34,644,327	35,339,239
VAL	snp50169-scaffold717-4207960/ rs268281880	snp3193-scaffold1095-4352995/ rs268236233	87	24,592,901	28,744,348

Genomic inbreeding coefficients (F_{ROH}) were found intermediate for GRF compared to the rest of the breeds analyzed, with a mean value of 0.069 (Fig 6a, S2 Table). The highest values were observed for GGT (0.143) and ORO (0.137). Moreover, the distribution of F_{ROH} calculated per CHR was similar, with some high values (>0.5) observed for CHI7, 9, 16, 22 and 25 (Fig 6b).

Fig 6 Summary of the genomic inbreeding coefficients (a) per breed and (b) of the Garfagnina breed per chromosome (Chr).

ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

Population Stratification and Ancestry

At a first step, a PCA was conducted on the complete data to visualize the general structure and relationships among breeds. The first axis distinguished GRF goats from ARG, CCG, DIT, GAR and GGT, while the second axis further separated GRF from the rest of the breeds (Fig 7a). An inspection of all the pairwise comparison between the first 10 axes (PCs) was carried out. By plotting the PC1 vs. PC6 (Fig 7b) four clusters were

252 observed, namely: i) DIT, GAR, and GGT, ii) ARG and CCG, iii) BIO and VSS while iv)
253 GRF was grouped together with ORO and VAL.
254
255 **Fig 7. Scatterplot of (a) the first two and (b) first and sixth principal components.**
256 ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di
257 Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL:
258 Valdostana and VSS: Valpassiria.
259
260 An admixture analysis was conducted to complement with the PCA results. A varying
261 number of group ancestries was investigated, from K=2 up to 10. The model with the
262 minimum CV error was the one with eight group ancestries (S2 Fig). In general, the
263 admixture results were in agreement with PCA, depicting the uniqueness of the GRF
264 genome. At K=4, the GRF was shown almost as a breed-specific ancestry, sharing a
265 small degree of ancestry primarily with ORO and VAL and further with GGT and DIT (Fig
266 8). At K=8, again, GRF had almost a breed-specific ancestry, with a small percentage of
267 the GRF goats sharing co ancestry with i) BIO and VSS and ii) ARG and CCG and to a
268 small extent with ORO, VAL, DIT and GGT. It should be noted that apart from GRF, group-
269 specific ancestries, at least to a great extent, existed almost for all breeds but ARG, CCG,
270 GAR and VSS.
271
272 **Fig 8**. **Admixture analysis with a) K=4 and b) K=8 coancestry groups.**

ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

Discriminant Analysis of Principal Components

In the first scenario of DAPC, all data were used. The first 40 PCs, explaining ~35.75% of the total variability in the SNP data (S3 Fig), were used in the final DAPC model, resulting in an assignment success rate of 100% for all the GRF goats to its breed of origin. The pattern of the genetic diversity based on the DAPC is presented in Fig 9, where a clear genetic distance of GRF from the rest of the breeds can be observed.

Fig 9 Scatterplot of the first two discriminant components of the DAPC. Breeds are presented by different colors and symbols.
ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

An external validation scenario, that better reflects a practical application of the discriminant model, was further assessed. In the first analysis (CV_{SS} scenario), the GRF breed had representative animals in the reference population where the model of DAPC was developed. Also, in this case, a 100% correct classification of the GRF goats was observed (S3 Table). Interestingly, the classification of the GRF was invariant to the number of PCs selected (ranged between 10 to 70) in DAPC (S4 Table). In the second

scenario (CV_{US}) there were no representative GRF samples in training the model of DAPC. In that case the majority of animals were classified as CCG while few were assigned to DIT in some of the CV replicates (S5 Table). Similar results were obtained with an increased size of the reference population. In the majority of the scenarios, the GRF goats were classified either as CCG or DIT, where there were few cases in which GRF goats were also assigned as VSS, GAR or BIO, but in none of the cases as ORO, GGT or VAL (Fig 10).

Fig 10. Percentage of assignment of the GRF goats in the CV_{US} scenario. Results were averaged over 10 replicates (CV) in each data subset (from 20 to 90%). CV_{US}: unsupervised CV, where GRF breed had no representative goats in model training, hence GRF goats had to be classified in one of the rest 8 breeds; ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

Discussion

The GRF breed is one of the Italian native goat breeds facing the risk of extinction, with a total number of registered animals lower than 1,500. Given the risk status of the breed, scientists have focused in a better description the GRF population. Characteristics on various zootechnical parameters of the breed, for instance, the milk total and fatty acid composition, milk coagulation properties and casein genotypes have been previously

318 investigated [4], with authors encouraging the development of a purebred breeding
319 scheme. Nevertheless, up to present, a whole-genome population analysis has not been
320 carried out to study the GRF genome in terms of genetic diversity.

321 Following from the above, the present study aimed at describing the genomic profile
322 of the GRF breed, relative to other Italian breeds for which genomic data were available.
323 To achieve this, a sample of 48 genotyped GRF goats was merged with 214 genotyped
324 goats coming from 9 Italian native breeds, 7 of them considered as dairy breeds and only
325 2 as meat breeds (ARG and VSS). These last have recently been analysed and presented
326 from the AdaptMap project [3], hence we focused in presenting the results for the GRF
327 breed. Our analysis was split in three parts, namely: i) runs of homozygosity, ii) principal
328 component and admixture ancestry and iii) discriminant analysis.
329

330 ## Runs of homozygosity

331 Runs of homozygosity can provide with a useful source of information on the historical
332 background and the breeding management of a population. For instance, large ROH
333 segments could be a result of recent intense selective breeding or of a potential bottleneck
334 effect. As it has been highlighted in the work of Bertolini et al. in goats [23], crossbred
335 populations tend to have smaller total ROH length and number compared to purebred
336 populations. Same pattern was observed comparing unselected vs. selected populations
337 undergoing breeding programs. Nevertheless, as it has been pointed out by [22,23] the
338 50k chip could not be considered adequate for an accurate detection of the small ROH,
339 resulting in underestimation of small ROH hits. Despite this, our analysis was based on
340 purebred goats, and thereby a smaller bias is expected. Results of ROH and F_{ROH} for the

9 Italian breeds of the AdaptMap project were in agreement with the estimates already reported [23]. Hence, our discussion on ROH is focused on highlighting the results on GRF in comparison to the rest of the 9 breeds analyzed.

The general pattern of ROH (i.e. in terms of total – and by chromosome – number and length of ROH) for GRF was similar to GGT and ORO. Moreover, an excess of frequent ROH was found for GRF (more than 45% in the GRF samples analyzed) on CHR 12 at, roughly, between 34.6-35.3 Mbp (Table 3). The same region was also detected in the DIT breed, while the broader region (~33,9-36.5 Mbp) was shared among the ARG, CCG, and GGT. Further, a search of genes presented in the top ROH region identified for GRF (~50.25-50.94Mbp, updated on the ARS1 assembly) and 1Mbp up-downstream (~49.25-51.94Mbp, updated on the ARS1 assembly) was carried out. Interestingly, the region ~49-52Mbp has been previously reported in goats [22–26]. It is worth noting that within this region lay the genes of the general gap junction protein family *GJA3* (gap junction protein alpha 3 ; ~50.642-50.644Mbp), *GJB2* (gap junction protein beta 2; ~50.675-50.676Mbp), *GJB6* (gap junction protein beta 6; ~50.694-50.695Mbp). The *GJB2* and *GJB6* are associated with the nervous system, hearing functions and ectodermal [24,25]. Moreover, the *SAP18* (Sin3A associated protein 18; ~51.136-51.141Mbp) that is related to gonad development [28], is also mapped in this region.

The narrow region of the detected top ROH runs for GRF on CHI12 was spanned between the *CENPJ* (centromere protein J; 50.23-50.27Mb) and the *IL17D* (interleukin 17D; 50.91-50.93 Mb). More precisely, the snp30397-scaffold335-418126 was found to be an intron of the *CENPJ* gene, while the snp30413-scaffold335-1113038 was downstream the *IL17D*. There are a series of studies that have linked the *CENPJ* with

364 primary microcephaly in humans and in mice [28–31]. Moreover, *CENPJ* has been found
365 to regulate in mouse the neurogenesis and the cilia disassembly in the developing corex
366 [33]. Also in mouse, disruption of the *CENPJ* can cause the Seckel Syndrome [34].
367 Apart from the *CENPJ* and *IL17D* genes, within this genomic area, 9 more genes can
368 be found, namely *PARP4* (poly(ADP-ribose) polymerase family member 4; ~50.29-
369 50.35Mbp), *MPHOSPH8* (M-phase phosphoprotein 8; ~50.36-50.41Mbp), *PSPC1*
370 (paraspeckle component 1; ~50.44-50.48Mbp), *ZMYM2* (zinc finger MYM-type containing
371 2; ~50.56-50.63Mbp), as well as the *CRYL1* (crystallin lambda 1; ~50.77-50.83Mbp) and
372 the *IFT88* (intraflagellar transport 88; ~ 50.84-50.91Mbp).
373 Downstream this region, in 1Mbp expansion, the genes *ATP12A* (ATPase H+/K+
374 transporting non-gastric alpha2 subunit; ~50.08-50.11Mbp) and *RNF17* (ring finger
375 protein 17; ~50.11-50.23Mbp) are located. Moreover, upstream the region there are
376 mapped also the *EEF1AKMT1* (EEF1A lysine methyltransferase 1; ~50.94-50.95Mbp),
377 *LATS2* (large tumor suppressor kinase 2; ~51.07-51.09Mbp), *ZDHHC20* (zinc finger
378 DHHC-type containing 20; ~51.16-51.22Mbp), *MICU2* (mitochondrial calcium uptake 2;
379 ~51.23-51.28Mbp), and the *FGF9* (fibroblast growth factor 9; ~51.34-51.37Mbp).
380

381 ## Population Stratification and Ancestry

382 Two approaches, complementary to each other, have been used to infer the GRF
383 relationships with 9 native Italian breeds, namely principal component and admixture
384 analysis. PCA is widely used to identify structure in the data and to distinguish between
385 groups of the samples. In that sense, the objective of PCA is to summarize (dis)similarities
386 over the different groups in the data rather than the individual itself. On the other hand,

387 admixture is focusing on the individuals; it provides with probabilities for each individual
388 to be clustered in one of the pre-defined group ancestries. As such, the two approaches
389 could be viewed as complementary rather antagonistic to each other.

390 Both of the analysis confirmed the distinguished and unique genetic background of
391 the GRF breed and results were in general agreement. More precisely, the PCA
392 scatterplot of PC1 vs. PC2 (Fig 7a) placed the GRF closer to VSS, BIO and CCG. This
393 finding was similar to the admixture analysis with K=8 (Fig 8b), which was the final model
394 selected after CV comparison. At K=4 in admixture, GRF shared co-ancestry with the
395 ORO and VAL breeds (Fig 8a). In PCA this relationship has been visualized by plotting
396 PC1 vs. PC6 (Fig 7b).

397 As a further step, we investigated the potential of breed traceability based on genomic
398 data. To this purpose, an LDA model was used, where genotypes were firstly transformed
399 into PCs, and a small set of those (not more than 300) was fitted in LDA (DAPC analysis).
400 The DAPC model was able to classify with 100% success the GRF goats to its breed of
401 origin (S3 Table). Moreover, an unsupervised learning was applied, where the GRF had
402 no representative samples in the reference population. Results were consistent with the
403 PCA and admixture and assigned the majority of the GRF goats in the CCG breed with a
404 small number of goats (varied between 4 to 6) assigned as DIT (S5 Table).

405 As mentioned above and in M&M, the primary step of the DAPC analysis is to select
406 the number of PCs to be used in the discriminant model. Hence, a basic question was on
407 how robust the DAPC could be considered relative to the number of PCs used. Our
408 analysis showed that, although the assignment success was invariant to the number of
409 PCs in the semi-supervised DAPC analysis (number of PCs varied between 10-70), a

pattern was found in the case of the unsupervised model. More precisely, when the DAPC contained 40 PCs some of the goats were classified as DIT goats. In the rest of the cases, where 60 or 20 PCs were used, all of the GRF goats were assigned as CCG.

PCA analysis seems thus to individuate common ancestries between GRF goats and Alpine Arc goat breeds whereas the DAPC approach identifies similarities between GRF and the goat breeds from Central Italy. Both hypotheses are consistent with the history of the Tuscan goat populations that experienced migratory flows both from the North and from Central Italy. The genomic analysis confirms the hypothesis that the GRF breed is a result of crosses among goats from the Alpine Arc and Tuscan-Emilian Apennines regions. Nevertheless, at a great part, the GRF breed consists of a unique genetic pool, genetically distinguished from 9 other native Italian breeds, for which genomic information was available and analysed here. This, in turn, resulted in breed traceability with a 100% success rate after CV. To sum up, our analysis complements previous work on various zootechnical and adaptive characteristics [4,7,8] of the GRF population and provides with a more complete description of the breed.

Conclusions

Our genomic analysis suggests a unique genetic pool of the Garfagnina breed, with small parts of common ancestry shared with Bionda dell'Adamello, Valpassiria, Argentata dell'Etna and Ciociara Grigia. Moreover, GRF can be successfully discriminated by the rest of the breeds analysed using genomic information with a success rate of 100%. This could help in breed traceability and controlling the amount of crossbreeding in the future. A ROH on CHI12 associated with the *CENPJ* gene should be further investigated in the

433 population. We suggest conservation and breeding measures to be taken for the
434 Garfagnina goat. We hope our work will add value to the GRF farming and the local region
435 where the breed is reared.
436

437 References

438 1. Curone G, Filipe J, Cremonesi P, Trevisi E, Amadori M, Pollera C, et al. What we
439 have lost: Mastitis resistance in Holstein Friesians and in a local cattle breed. Res Vet
440 Sci. 2018;116:88–98.
441 2. Lovreglio R, Meddour-Sahar O, Leone V. Goat grazing as a wildfire prevention tool: a
442 basic review. IForest - Biogeosciences For. 2014;7:260.
443 3. Stella A, Nicolazzi EL, Van Tassell CP, Rothschild MF, Colli L, Rosen BD, et al.
444 AdaptMap: exploring goat diversity and adaptation. Genet Sel Evol. 2018;50:61.
445 4. Martini M, Salari F, Altomonte I, Rignanese D, Chessa S, Gigliotti C, et al. The
446 Garfagnina goat: A zootechnical overview of a local dairy population. J Dairy Sci. 93.
447 5. Zeder MA, Hesse B. The initial domestication of goats (Capra hircus) in the Zagros
448 mountains 10,000 years ago. Science. 2000;287:2254–7.
449 6. Domestic Animal Diversity Information System (DAD-IS) | Food and Agriculture
450 Organization of the United Nations [Internet]. 2019. Available from:
451 http://www.fao.org/dad-is/en/
452 7. Salari F, Altomonte I, Ribeiro NL, Ribeiro MN, Bozzi R, Martini M. Effects of season
453 on the quality of Garfagnina goat milk. Ital J Anim Sci. 2016;15:568–75.
454 8. Ribeiro NL, Costa RG, Pimenta Filho EC, Ribeiro MN, Crovetti A, Saraiva EP, et al.
455 Adaptive profile of Garfagnina goat breed assessed through physiological,

456 haematological, biochemical and hormonal parameters. Small Rumin Res.
457 2016;144:236–41.
458 9. Cecchi F, Russo C, Fratini F, Turchi B, Preziuso G, Cantile C. Preliminary association
459 analysis of microsatellites and Mycobacterium avium subspecies paratuberculosis
460 infection in the native Garfagnina goats. J Appl Anim Res. 2018;46:879–82.
461 10. Cecchi F, Dadousis C, Bozzi R, Fratini F, Russo C, Bandecchi P, et al. Genome
462 scan for the possibility of identifying candidate resistance genes for goat lentiviral
463 infections in the Italian Garfagnina goat breed. Trop Anim Health Prod. 2019;51:729–33.
464 11. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al.
465 Design and Characterization of a 52K SNP Chip for Goats. PLOS ONE. 2014;9:e86227.
466 12. Cecchi F, Russo C, Iamartino D, Galiero A, Turchi B, Fratini F, et al. Identification of
467 candidate genes for paratuberculosis resistance in the native Italian Garfagnina goat
468 breed. Trop Anim Health Prod. 2017;49:1135–42.
469 13. Bertolini F, Servin B, Talenti A, Rochat E, Kim ES, Oget C, et al. Signatures of
470 selection and environmental adaptation across the goat genome post-domestication.
471 Genet Sel Evol. 2018;50:57.
472 14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al.
473 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage
474 Analyses. Am J Hum Genet. 2007;81:559–75.
475 15. Purcell S. PLINK v1.9 [Internet]. Available from:
476 http://pngu.mgh.harvard.edu/purcell/plink/
477 16. Team RC. R: A language and environment for statistical computing. 2013.

478 17. Marras G, Gaspa G, Sorbolini S, Dimauro C, Ajmone-Marsan P, Valentini A, et al.
479 Analysis of runs of homozygosity and their relationship with inbreeding in five cattle
480 breeds farmed in Italy. Anim Genet. 2015;46:110–21.
481 18. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in
482 unrelated individuals. Genome Res. 2009;19:1655–64.
483 19. Alexander D, Shringarpure S, Novembre J, Lange K. ADMIXTURE Version 1.3.0
484 [Internet]. Available from: www.genetics.ucla.edu/software/admixture
485 20. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a
486 new method for the analysis of genetically structured populations. BMC Genet.
487 2010;11:94.
488 21. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers.
489 Bioinformatics. 2008;24:1403–5.
490 22. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide
491 SNP data. Bioinformatics. 2011;27:3070–1.
492 23. Bertolini F, Cardoso TF, Marras G, Nicolazzi EL, Rothschild MF, Amills M, et al.
493 Genome-wide patterns of homozygosity provide clues about the population history and
494 adaptation of goats. Genet Sel Evol. 2018;50:59.
495 24. Bertolini F, Gandolfi B, Kim ES, Haase B, Lyons LA, Rothschild MF. Evidence of
496 selection signatures that shape the Persian cat breed. Mamm Genome Off J Int Mamm
497 Genome Soc. 2016;27:144–55.
498 25. Kim E-S, Elbeltagy AR, Aboul-Naga AM, Rischkowsky B, Sayre B, Mwacharo JM, et
499 al. Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid
500 environment. Heredity. 2016;116:255–64.

501 26. Pandya A, Arnos KS, Xia XJ, Welch KO, Blanton SH, Friedman TB, et al. Frequency
502 and distribution of GJB2 (connexin 26) and GJB6 (connexin 30) mutations in a large
503 North American repository of deaf probands. Genet Med. 2003;5:295–303.
504 27. Lamartine J, Essenfelder GM, Kibar Z, Lanneluc I, Callouet E, Laoudj D, et al.
505 Mutations in GJB6 cause hidrotic ectodermal dysplasia. Nat Genet. 2000;26:142–4.
506 28. Feng Y, Peng X, Li S, Gong Y. Isolation and characterization of sexual dimorphism
507 genes expressed in chicken embryonic gonads. Acta Biochim Biophys Sin.
508 2009;41:285–94.
509 29. Naveed M, Kazmi SK, Amin M, Asif Z, Islam U, Shahid K, et al. Comprehensive
510 review on the molecular genetics of autosomal recessive primary microcephaly
511 (MCPH). Genet Res [Internet]. 2018 [cited 2019 Sep 30];100. Available from:
512 https://www.cambridge.org/core/journals/genetics-research/article/comprehensive-
513 review-on-the-molecular-genetics-of-autosomal-recessive-primary-microcephaly-
514 mcph/4B562746F743B4976802467EA7498A20
515 30. Faheem M, Naseer MI, Rasool M, Chaudhary AG, Kumosani TA, Ilyas AM, et al.
516 Molecular genetics of human primary microcephaly: an overview. BMC Med Genomics.
517 2015;8 Suppl 1:S4.
518 31. Duerinckx S, Abramowicz M. The genetics of congenitally small brains. Semin Cell
519 Dev Biol. 2018;76:76–85.
520 32. Bond J, Roberts E, Springell K, Lizarraga S, Scott S, Higgins J, et al. A centrosomal
521 mechanism involving CDK5RAP2 and CENPJ controls brain size. Nat Genet.
522 2005;37:353–5.

523 33. Ding W, Wu Q, Sun L, Pan NC, Wang X. Cenpj Regulates Cilia Disassembly and
524 Neurogenesis in the Developing Mouse Cortex. J Neurosci. 2019;39:1994–2010.
525 34. McIntyre RE, Chavali PL, Ismail O, Carragher DM, Sanchez-Andrade G, Forment
526 JV, et al. Disruption of Mouse Cenpj, a Regulator of Centriole Biogenesis, Phenocopies
527 Seckel Syndrome. PLOS Genet. 2012;8:e1003022.

528

529 **Funding:** This work was supported by grants of the University of Pisa (PRA 2016) and
530 grants of the University of Firenze (RICCARDOBOZZIRICATEN19).

531

532 Supporting information

533 **S1 Fig. Number of SNPs per chromosome after quality control. The plot has been**
534 **produced with the *synbreed R* package [15].**
535 **S2 Fig. Cross-validation results for assessing the number of ancestry groups in the**
536 **admixture analysis.**
537 **S3 Fig. Cross-validation results for the selection of principal components to be**
538 **retained in the DAPC analysis.**

539

540 **S1 Table. Percentage of the number of runs of homozygosity per length class and**
541 **breed.**
542 ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di
543 Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL:
544 Valdostana and VSS: Valpassiria.

545

S2 Table. Descriptive statistics of the genomic inbreeding coefficients per breed.

ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

S3 Table. Assignment results of the Garfagnina goats in the semi-supervised cross-validation (CV_{SS}) scenario with 10 repetitions.

ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

S4 Table. Number of principal components selected via cross-validation (CV) in the 1st scenario of the DAPC analysis of the Garfganina breed (GRF).

CV_{SS}: semi-supervised CV, where some GRF goats are present in model training; CV_{US}: unsupervised CV, where GRF breed had no representative goats in model training.

S5 Table. Assignment results of the Garfagnina goats in the external cross-validation (CV_{US}) scenario with 10 repetitions.

ARG: Argentata dell'Etna; BIO: Bionda dell'Adamello; CCG: Ciociara Grigia; DIT: Di Teramo; GAR: Garganica; GGT: Girgentana; GRF: Garfagnina; ORO: Orobica; VAL: Valdostana and VSS: Valpassiria.

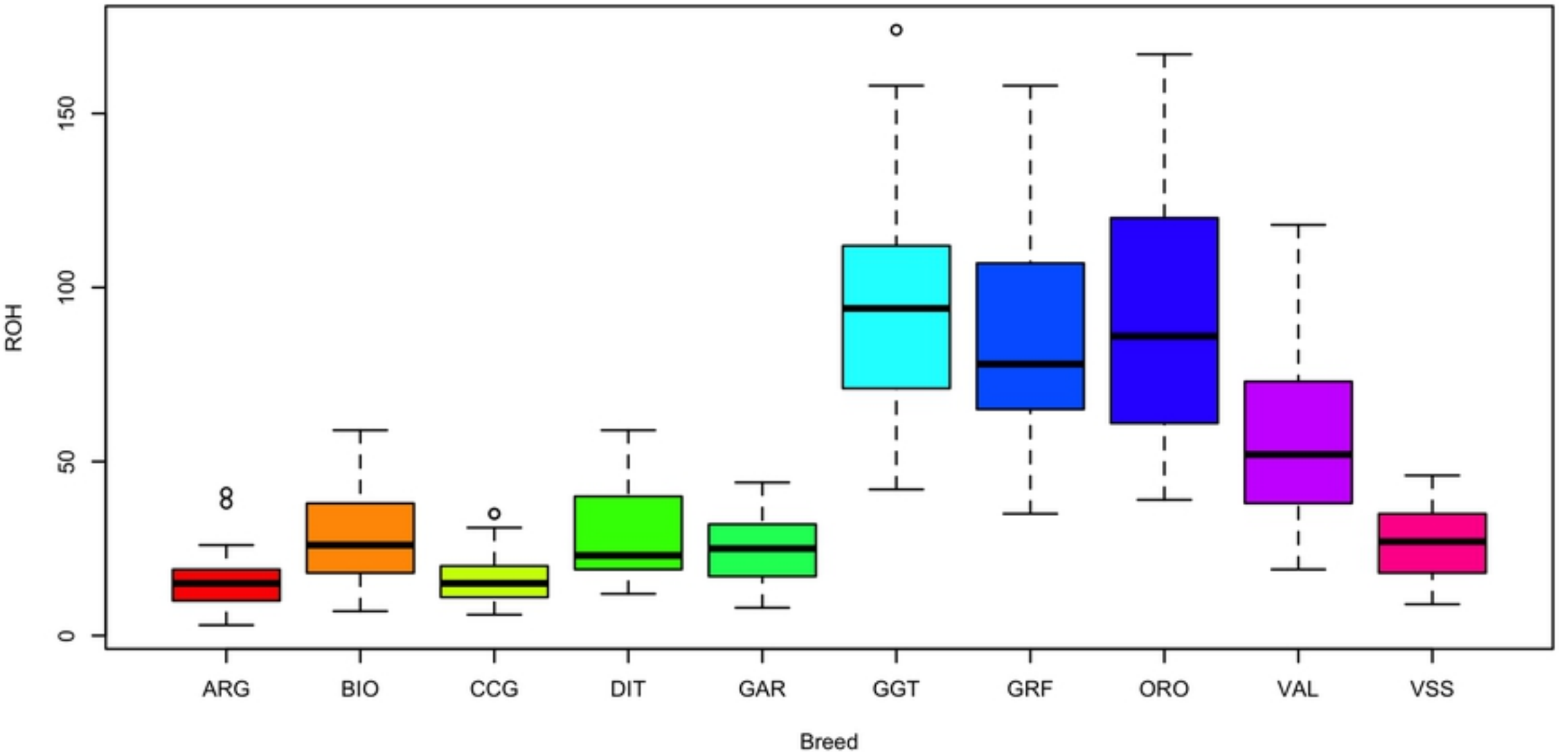

Fig 1

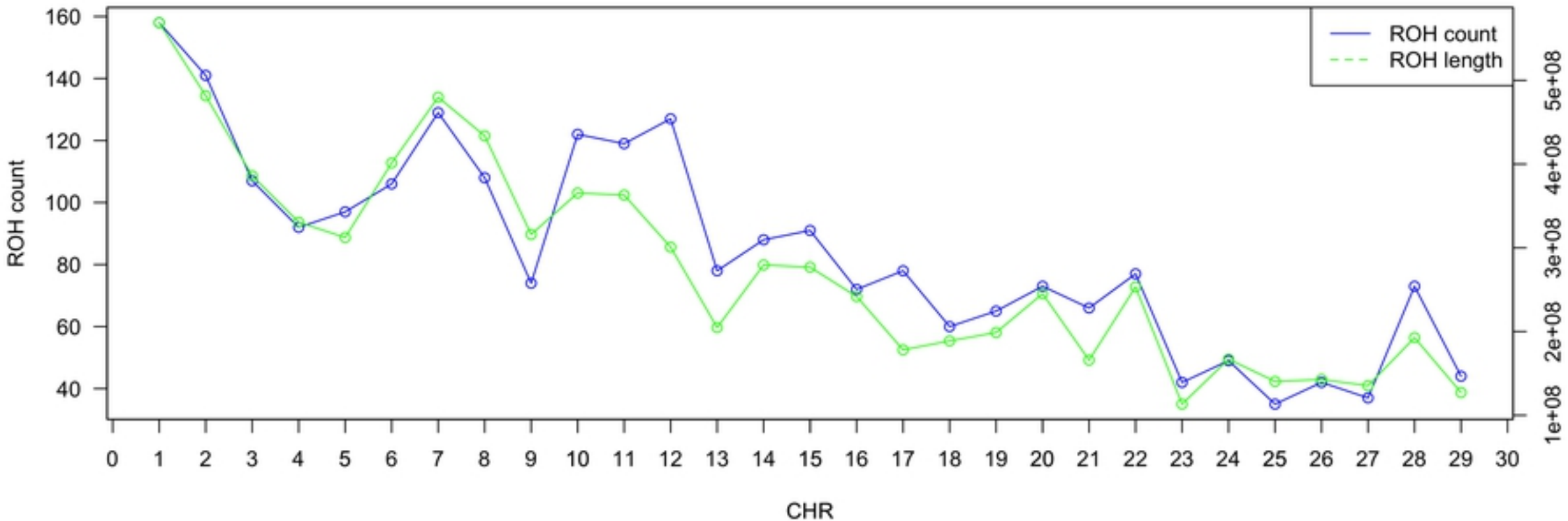
ROH count
ROH length
ROH count
CHR
160
140
120
100
80
60
40
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
1e+08
2e+08
3e+08
4e+08
5e+08

Fig 2

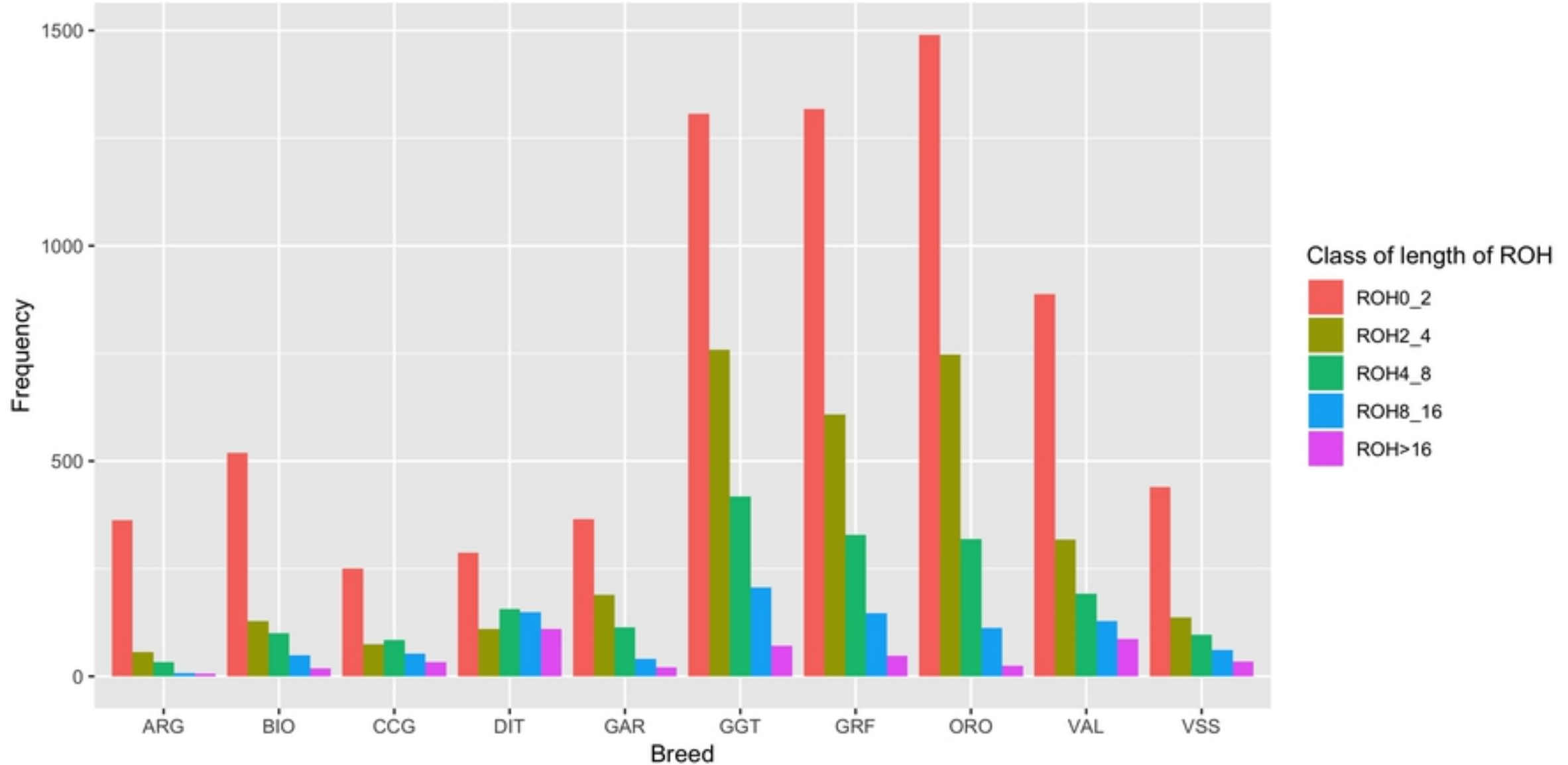
1500
1000
500
0
Frequency
ARG
BIO
CCG
DIT
GAR
GGT
GRF
ORO
VAL
VSS
Breed
Class of length of ROH
ROH0_2
ROH2_4
ROH4_8
ROH8_16
ROH>16

Fig 3a

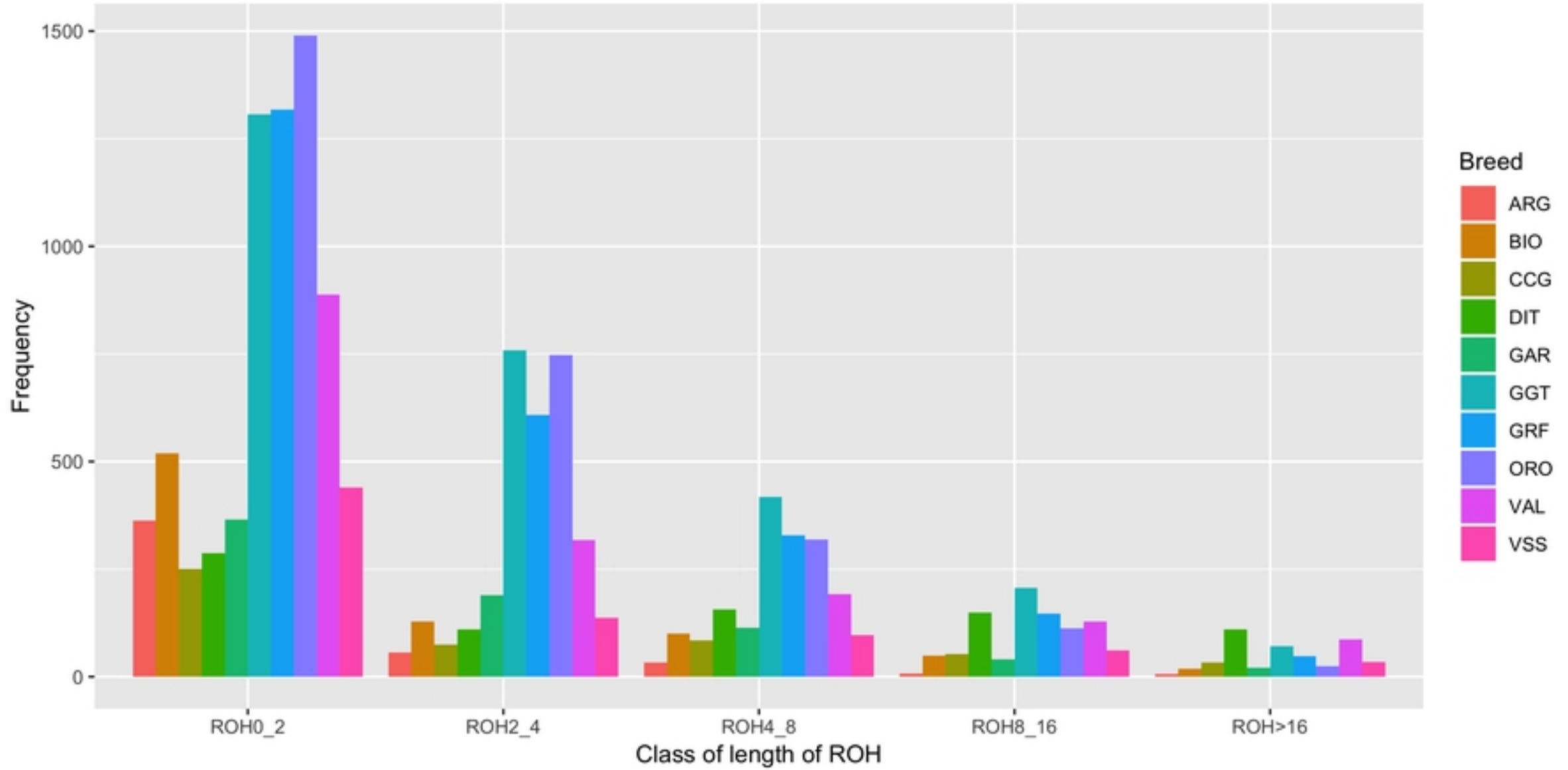
1500
1000
500
0
Frequency
ROH0_2
ROH2_4
ROH4_8
ROH8_16
ROH>16
Class of length of ROH
Breed
ARG
BIO
CCG
DIT
GAR
GGT
GRF
ORO
VAL
VSS

Fig 3b

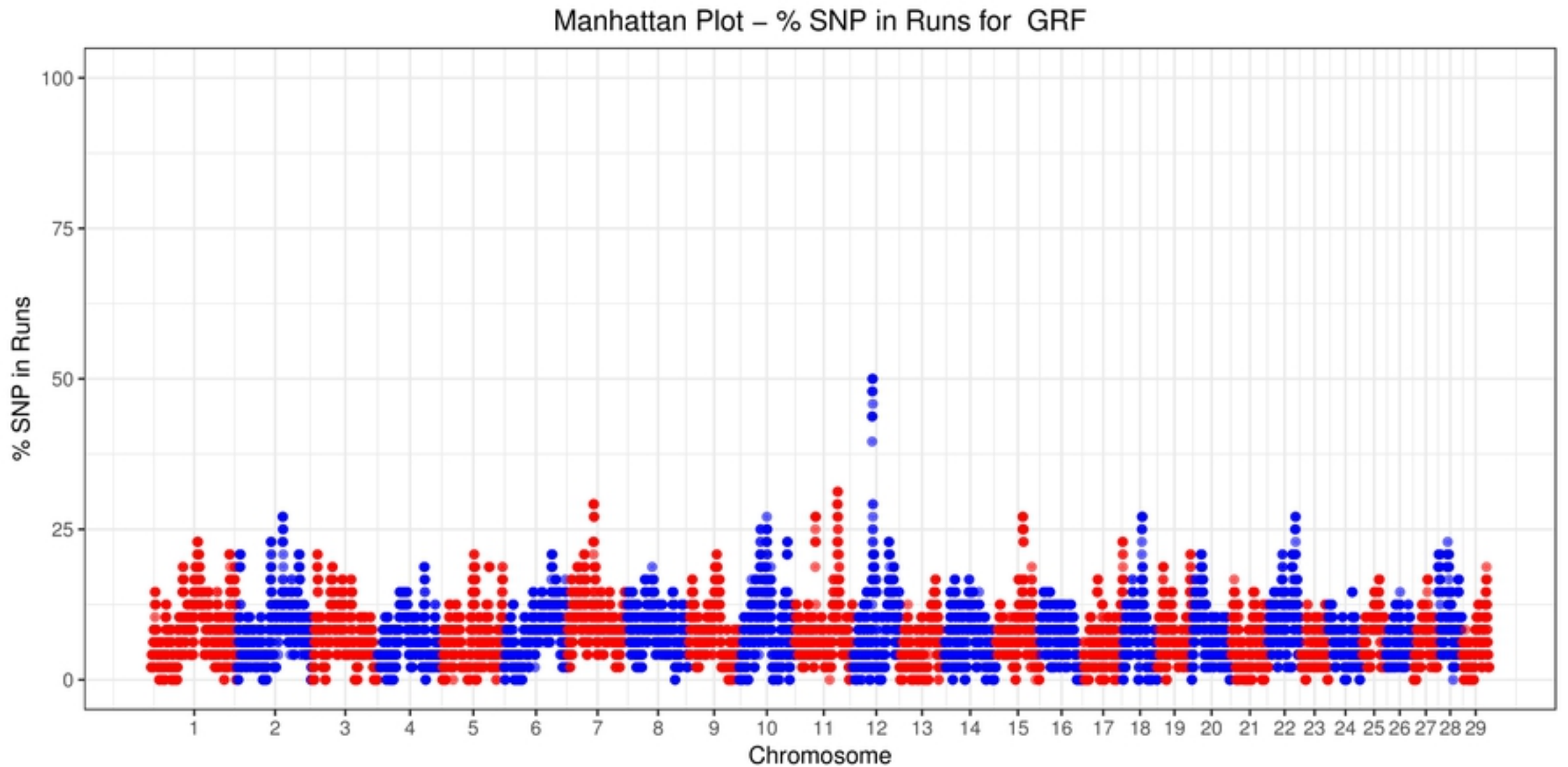
Manhattan Plot – % SNP in Runs for GRF
% SNP in Runs
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
Chromosome

Fig 4

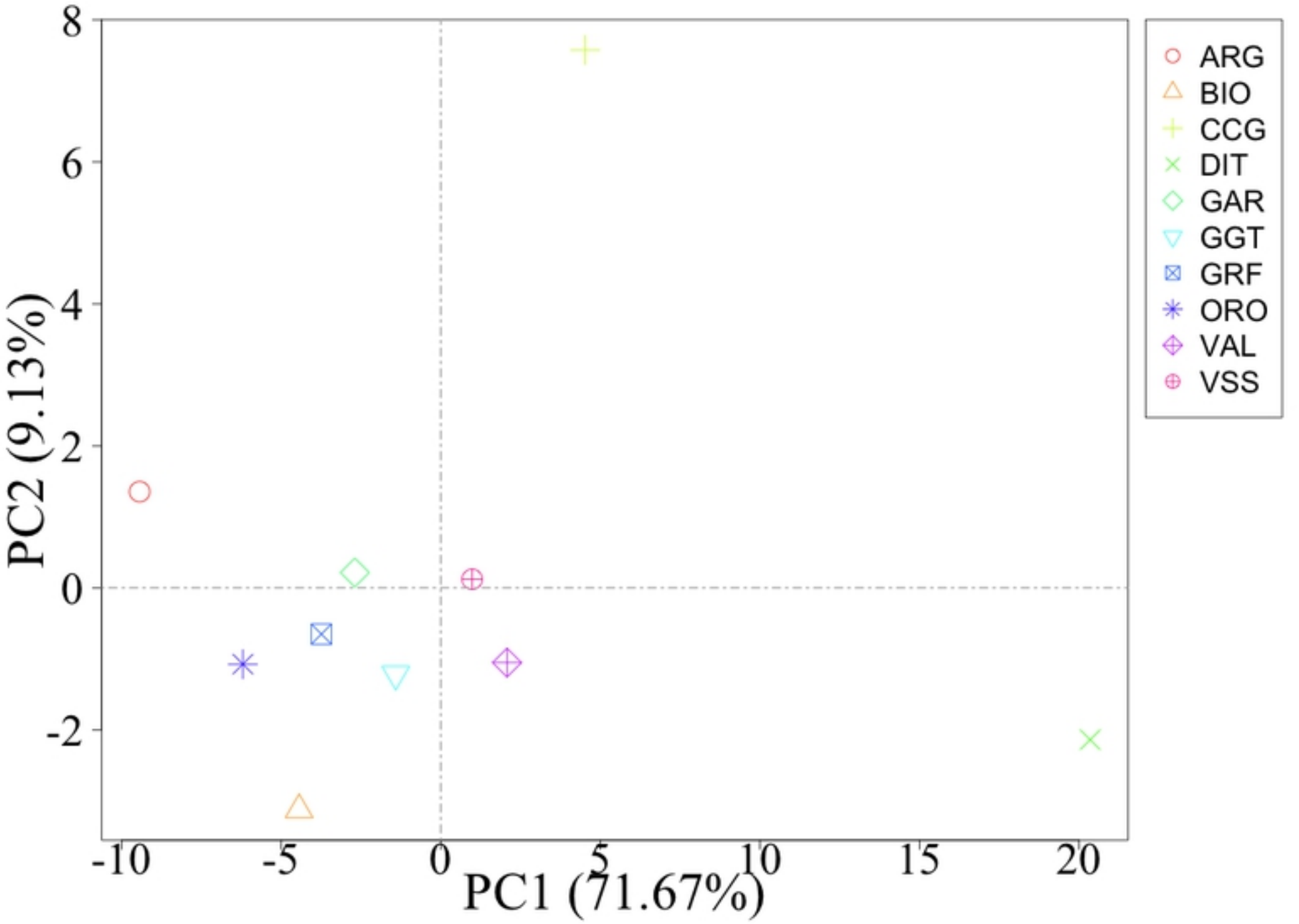
8
6
4
2
0
-2
PC2 (9.13%)
-10
-5
0
5
10
15
20
PC1 (71.67%)
ARG
BIO
CCG
DIT
GAR
GGT
GRF
ORO
VAL
VSS

Fig 5a

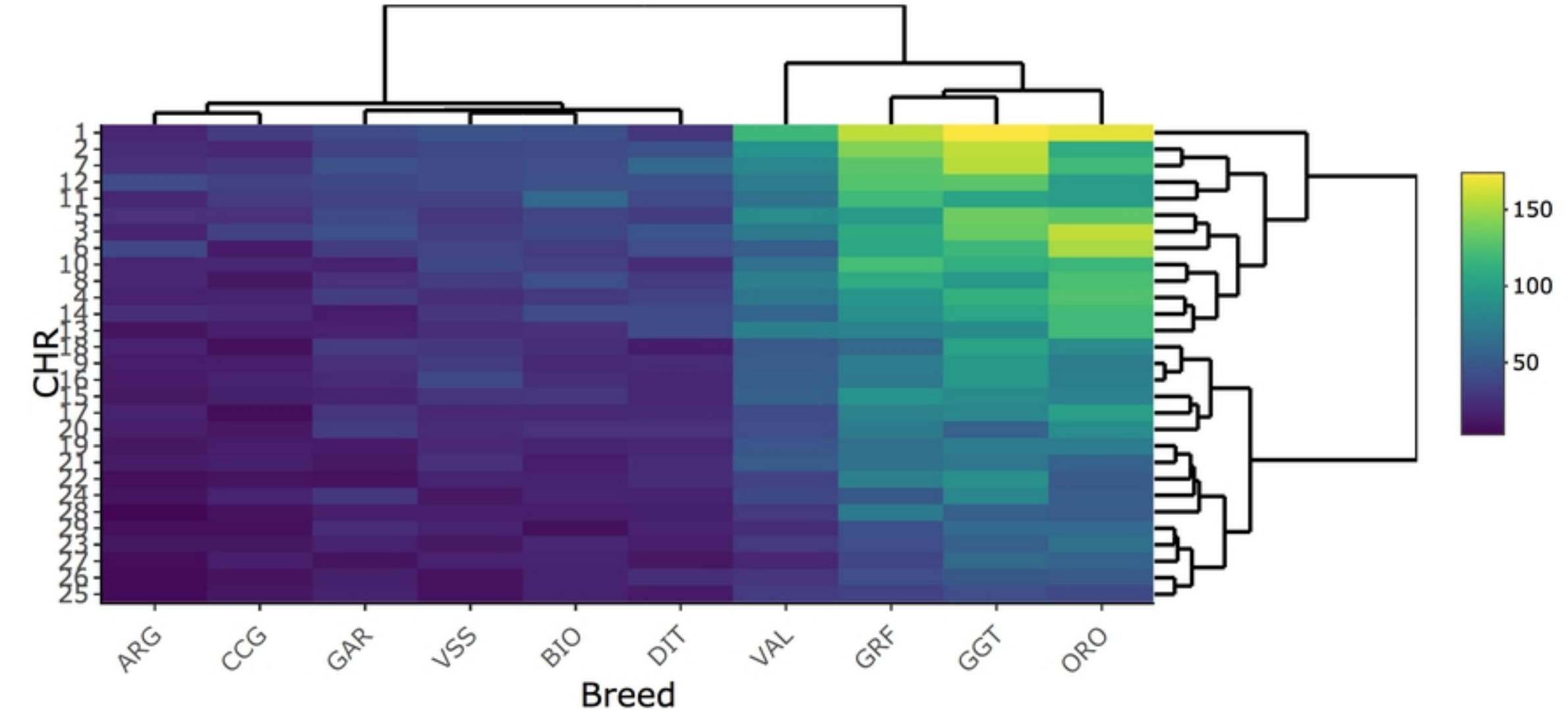

Fig 5b

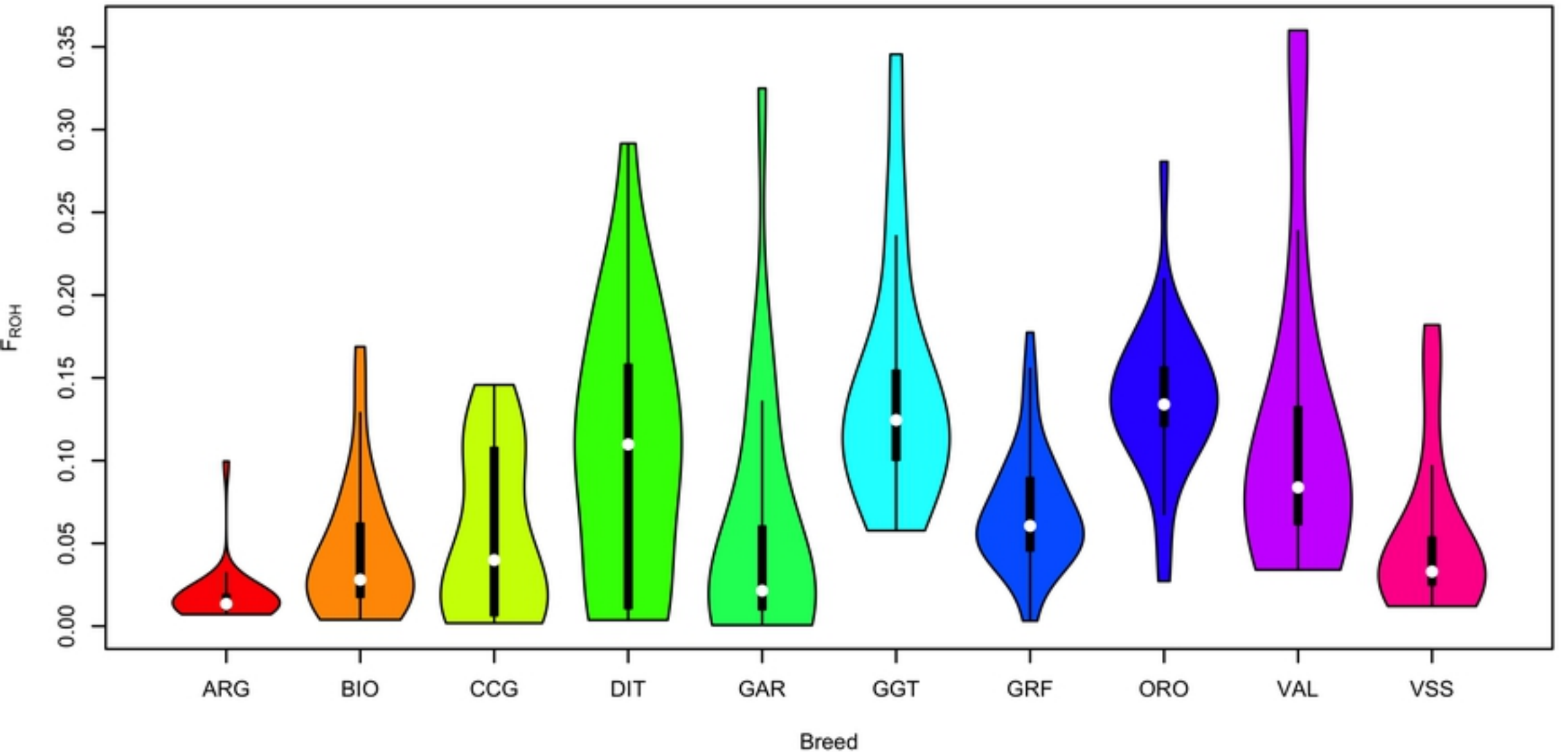
F_ROH
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
ARG
BIO
CCG
DIT
GAR
GGT
GRF
ORO
VAL
VSS
Breed

Fig 6a

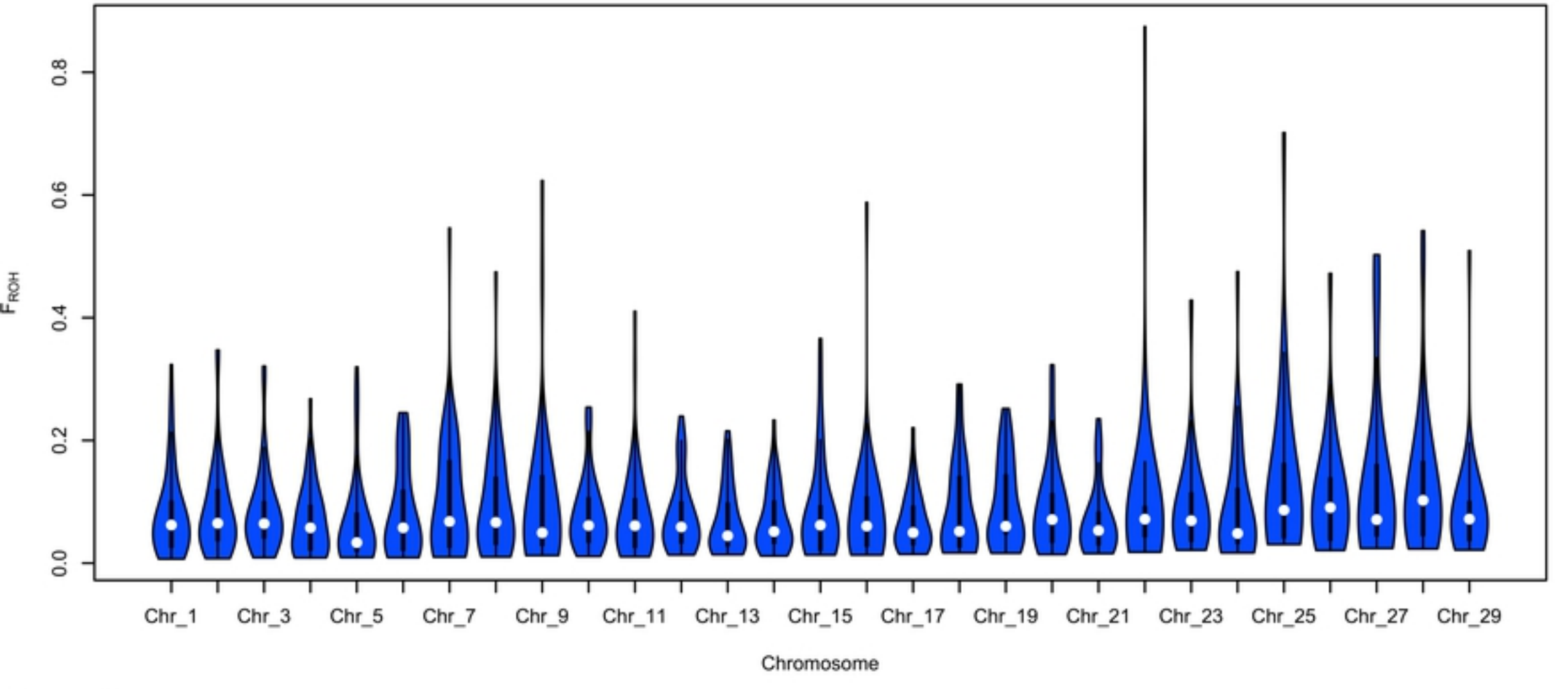
0.0
0.2
0.4
0.6
0.8
F_{ROH}
Chr_1
Chr_3
Chr_5
Chr_7
Chr_9
Chr_11
Chr_13
Chr_15
Chr_17
Chr_19
Chr_21
Chr_23
Chr_25
Chr_27
Chr_29
Chromosome

Fig 6b

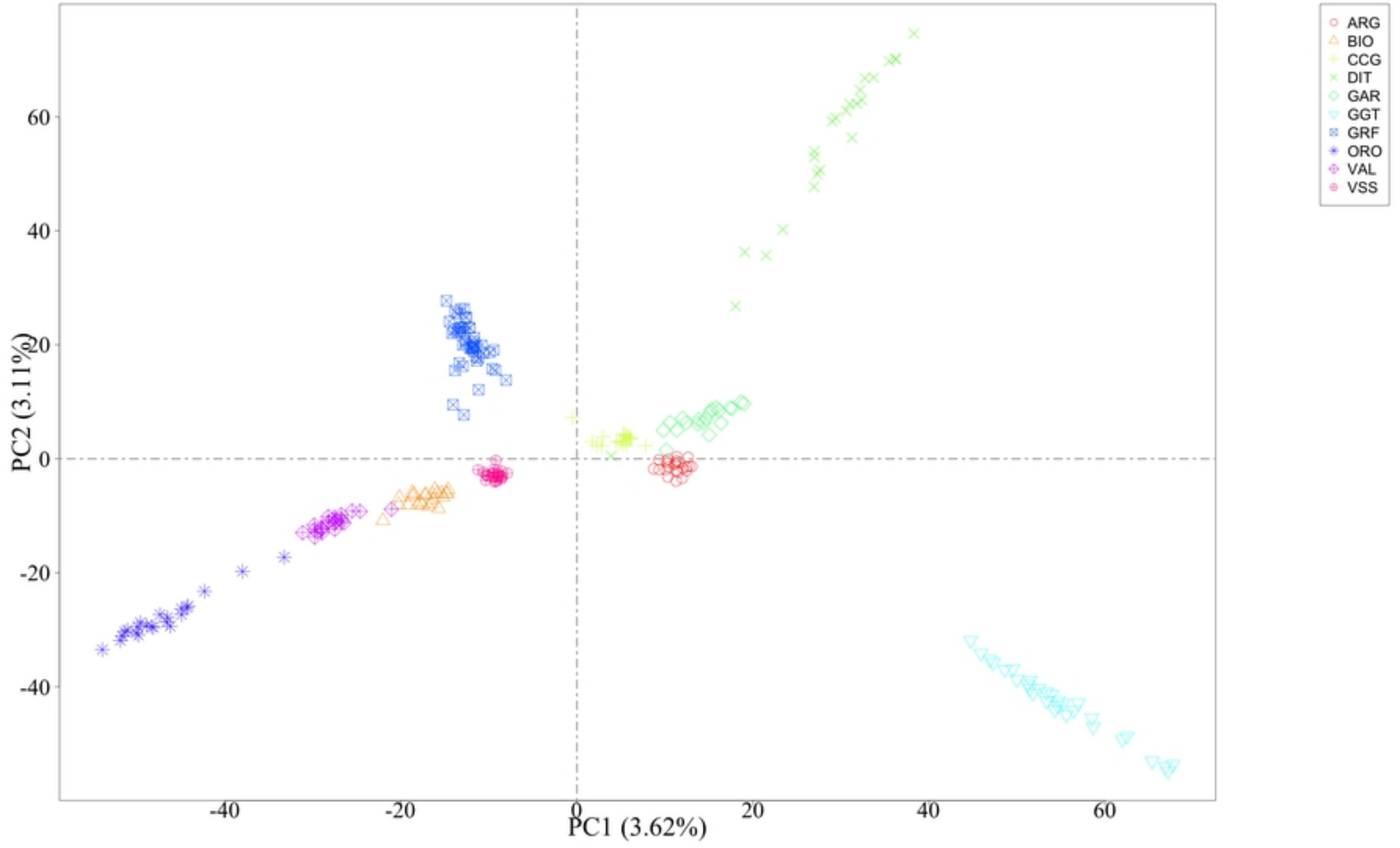
ARG
BIO
CCG
DIT
GAR
GGT
GRF
ORO
VAL
VSS
60
40
20
0
-20
-40
PC2 (3.11%)
-40
-20
0
20
40
60
PC1 (3.62%)

Fig 7a

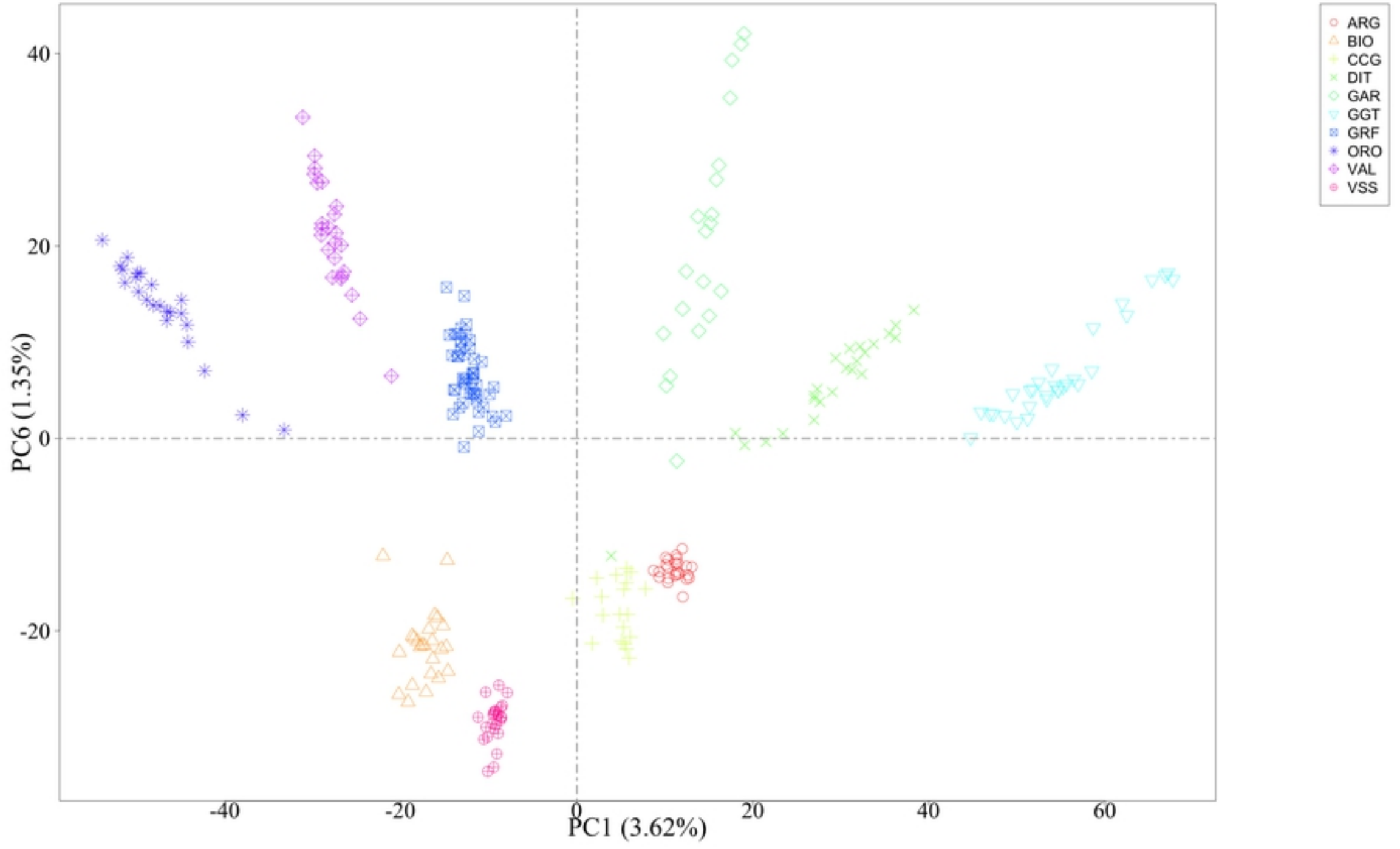
ARG
BIO
CCG
DIT
GAR
GGT
GRF
ORO
VAL
VSS
40
20
0
-20
PC6 (1.35%)
-40
-20
0
20
40
60
PC1 (3.62%)

Fig 7b

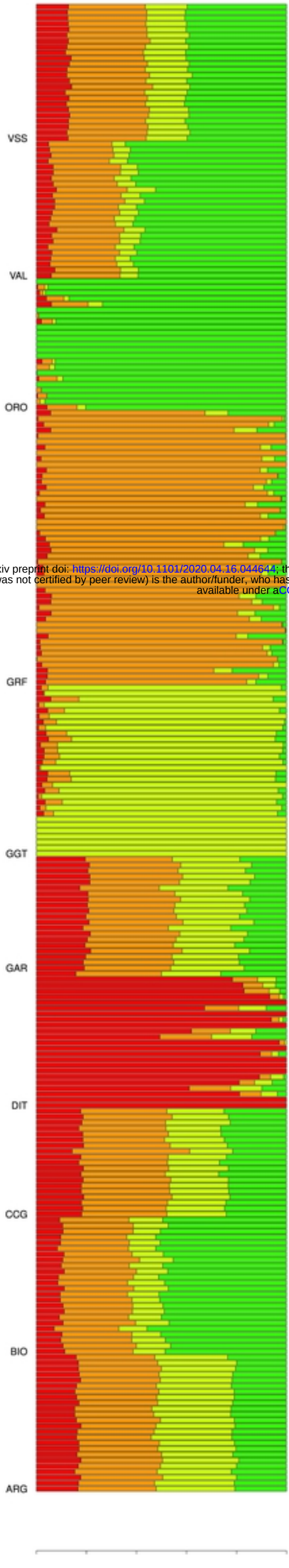

Fig 8a

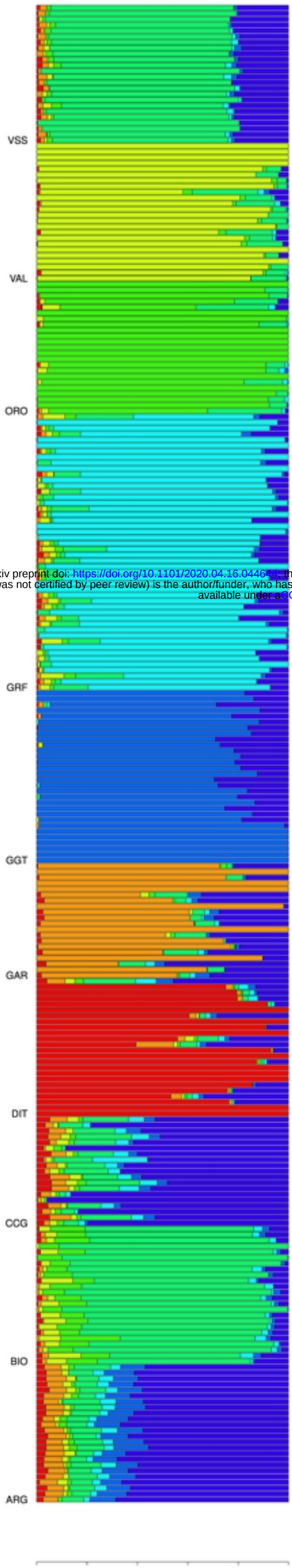

Fig 8b

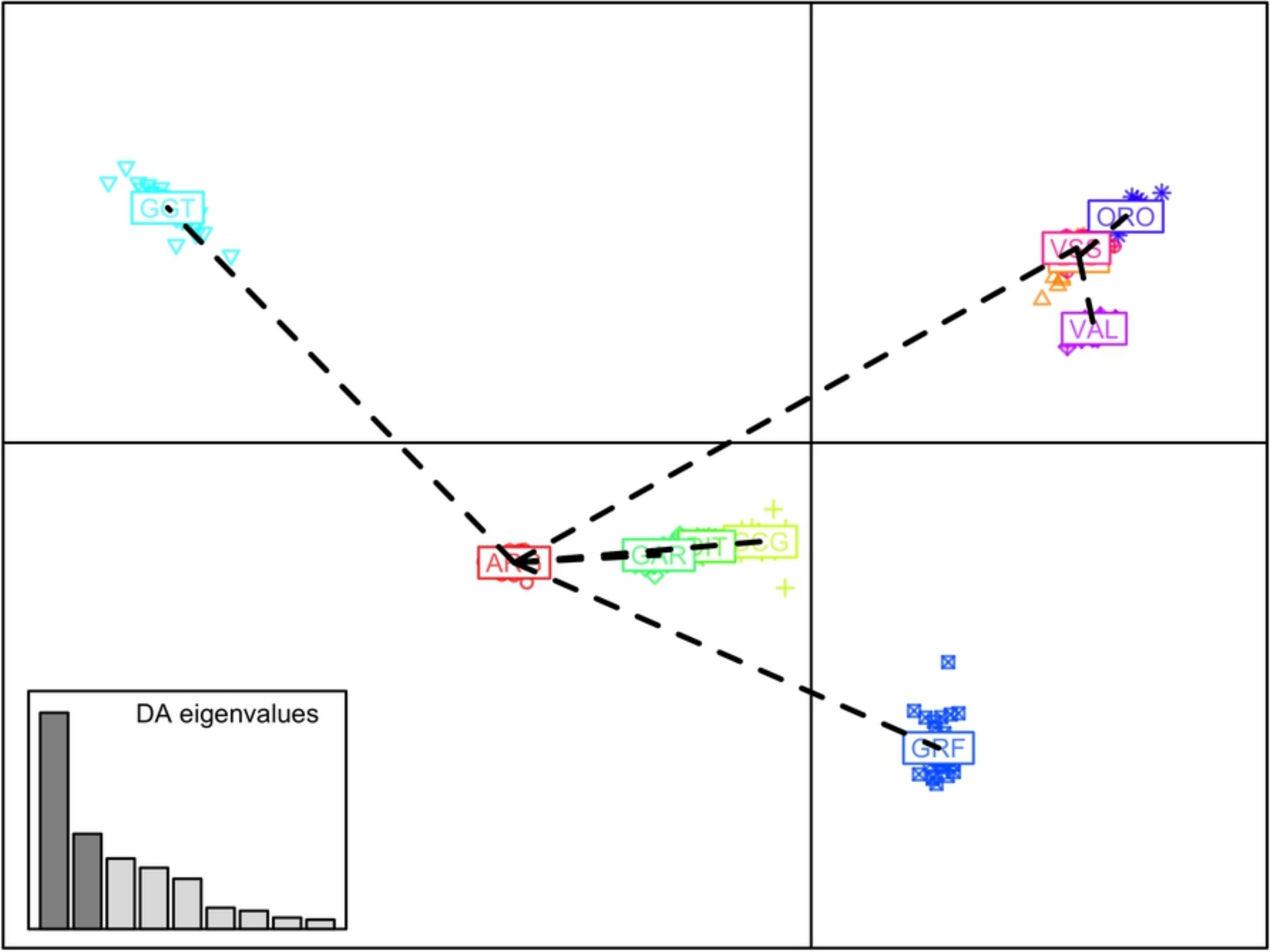

Fig 9

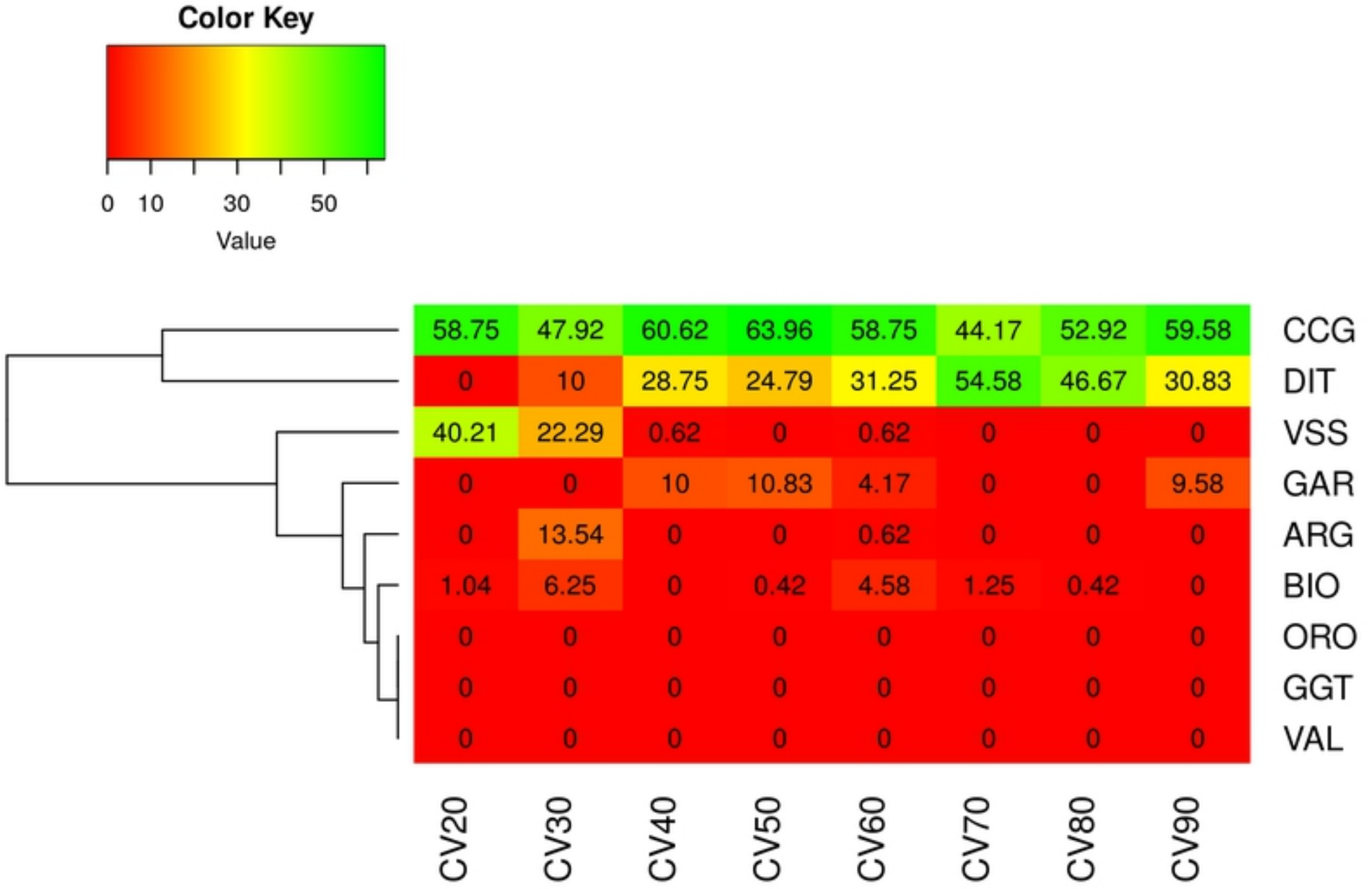
Color Key
0 10 30 50
Value
58.75 47.92 60.62 63.96 58.75 44.17 52.92 59.58 CCG
0 10 28.75 24.79 31.25 54.58 46.67 30.83 DIT
40.21 22.29 0.62 0 0.62 0 0 0 VSS
0 0 10 10.83 4.17 0 0 9.58 GAR
0 13.54 0 0 0.62 0 0 0 ARG
1.04 6.25 0 0.42 4.58 1.25 0.42 0 BIO
0 0 0 0 0 0 0 0 ORO
0 0 0 0 0 0 0 0 GGT
0 0 0 0 0 0 0 0 VAL
CV20 CV30 CV40 CV50 CV60 CV70 CV80 CV90

Fig 10