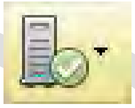
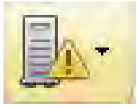
WILEY

Online Proofing System Instructions

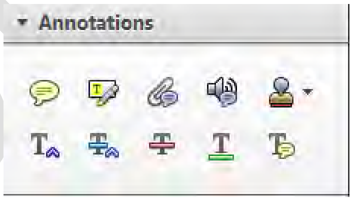
The Wiley Online Proofing System allows proof reviewers to review PDF proofs, mark corrections, respond to queries, upload replacement figures, and submit these changes directly from the locally saved PDF proof.

1. For the best experience reviewing your proof in the Wiley Online Proofing System ensure you are connected to the internet. This will allow the PDF proof to connect to the central Wiley Online Proofing System server. If you are connected to the Wiley Online Proofing System server you should see a green check mark icon above in the yellow banner.


Connected


Disconnected

2. Please review the article proof on the following pages and mark any corrections, changes, and query responses using the Annotation Tools outlined on the next 2 pages.



3. Save your proof corrections by clicking the "Publish Comments" button in the yellow banner above. Corrections don't have to be marked in one sitting. You can publish comments and log back in at a later time to add and publish more comments before you click the "Complete Proof Review" button below.

4. If you need to supply additional or replacement files bigger than 5 Megabytes (MB) do not attach them directly to the PDF Proof, please click the "Upload Files" button to upload files:

Upload files

Click Here

5. When your proof review is complete and all corrections have been published to the server by clicking the "Publish Comments" button, please click the "Complete Proof Review" button below:

IMPORTANT: Did you reply to all queries listed on the Author Query Form appearing before your proof?
IMPORTANT: Did you click the "Publish Comments" button to save all your corrections? Any unpublished comments will be lost.
IMPORTANT: Once you click "Complete Proof Review" you will not be able to add or publish additional corrections.

Complete Proof Review

Click Here

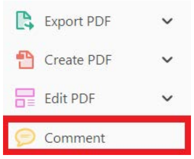
WILEY

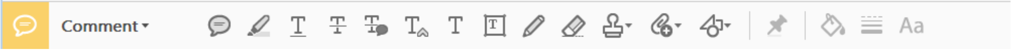
USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

Required software to e-Annotate PDFs: Adobe Acrobat Professional or Adobe Reader (version 11 or above). (Note that this document uses screenshots from Adobe Reader DC.)
The latest version of Acrobat Reader can be downloaded for free at: http://get.adobe.com/reader/

Once you have Acrobat Reader open on your computer, click on the Comment tab (right-hand panel or under the Tools menu).

This will open up a ribbon panel at the top of the document. Using a tool will place a comment in the right-hand panel. The tools you will use for annotating your proof are shown below:



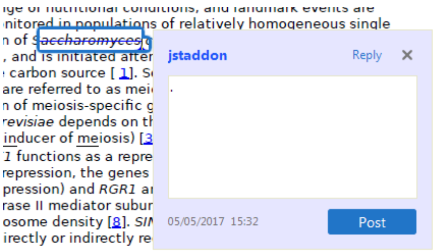


1. Replace (Ins) Tool – for replacing text.

Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it:

- Highlight a word or sentence.
- Click on .
- Type the replacement text into the blue box that appears.



nitored in populations of relatively homogeneous single
n of S~~accharomyces~~ c
, and is initiated after
carbon source [1]. S
are referred to as mei
n of meiosis-specific g
revisiae depends on th
inducer of meiosis) [3
1 functions as a repre
repression, the genes
pression) and *RGR1* ar
rase II mediator subur
osome density [8]. *SIN*
irectly or indirectly re

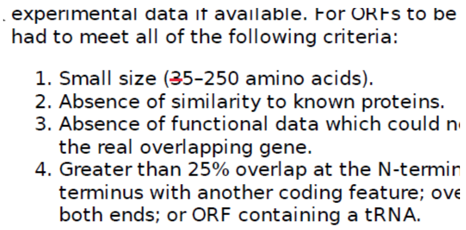
jstaddon
Reply ×
05/05/2017 15:32
Post

2. Strikethrough (Del) Tool – for deleting text.

Strikes a red line through text that is to be deleted.

How to use it:

- Highlight a word or sentence.
- Click on .
- The text will be struck out in red.



experimental data if available. For ORFs to be had to meet all of the following criteria:

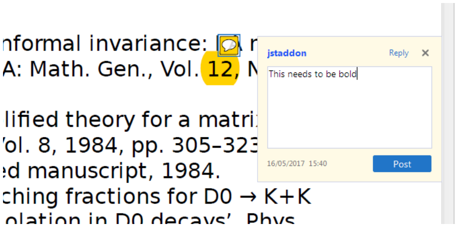
1. Small size (~~3~~5–250 amino acids).
2. Absence of similarity to known proteins.
3. Absence of functional data which could n the real overlapping gene.
4. Greater than 25% overlap at the N-termin terminus with another coding feature; ove both ends; or ORF containing a tRNA.

3. Commenting Tool – for highlighting a section to be changed to bold or italic or for general comments.

Use these 2 tools to highlight the text where a comment is then made.

How to use it:

- Click on .
- Click and drag over the text you need to highlight for the comment you will add.
- Click on .
- Click close to the text you just highlighted.
- Type any instructions regarding the text to be altered into the box that appears.



nformal invariance: r
A: Math. Gen., Vol. 12, N
lified theory for a matri
'ol. 8, 1984, pp. 305–323
ed manuscript, 1984.
ching fractions for D0 → K+K
olation in D0 decays', Phys

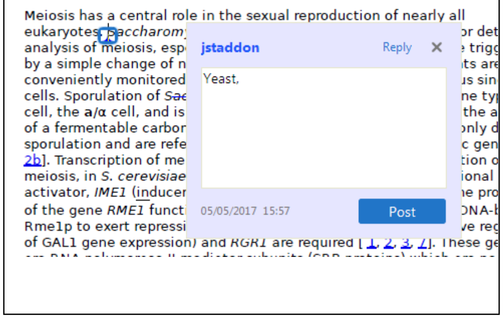
jstaddon
Reply ×
This needs to be bold
16/05/2017 15:40
Post

4. Insert Tool – for inserting missing text at specific points in the text.

Marks an insertion point in the text and opens up a text box where comments can be entered.

How to use it:

- Click on .
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the box that appears.



Meiosis has a central role in the sexual reproduction of nearly all
eukaryotes. *Saccharom* or det
analysis of meiosis, esp trigg
by a simple change of n ts are
conveniently monitored us sin
cells. Sporulation of *Sac* ne ty
cell, the **a**/α cell, and is the a
of a fermentable carbor only d
sporulation and are refe c gen
2b]. Transcription of me tion o
meiosis, in *S. cerevisiae* ional
activator, *IME1* (inducer he pro
of the gene *RME1* functi DNA-b
Rme1p to exert repressi ve reg
of GAL1 gene expression) and *RGR1* are required [1, 2, 3, 7]. These ge

jstaddon
Reply ×
Yeast,
05/05/2017 15:57
Post

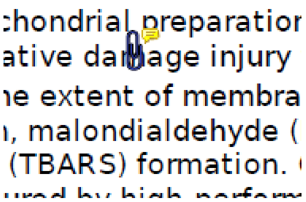
5. Attach File Tool – for inserting large amounts of text or replacement figures.

Inserts an icon linking to the attached file in the appropriate place in the text.

How to use it:

- Click on 📎.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.

  The attachment appears in the right-hand panel.



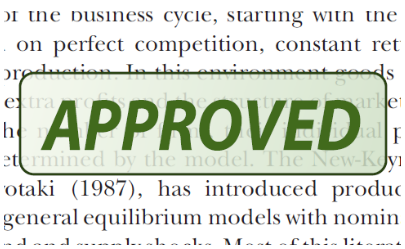
:hondrial preparatior
ative damage injury
ıe extent of membra
ı, malondialdehyde (
(TBARS) formation.

6. Add stamp Tool – for approving a proof if no corrections are required.

Inserts a selected stamp onto an appropriate place in the proof.

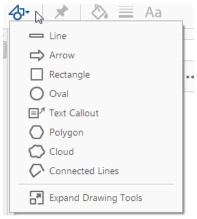
How to use it:

- Click on the stamp icon.
- Select the stamp you want to use. (The Approved stamp is usually available directly in the menu that appears. Others are shown under *Dynamic*, *Sign Here*, *Standard Business*).
- Fill in any details and then click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).



APPROVED

ɔr the business cycle, starting with the
. on perfect competition, constant ret
production. In this environment goods
extra profits and the structure of marke
he
etermined by the model. The New-Keyı
otaki (1987), has introduced produc
general equilibrium models with nomin



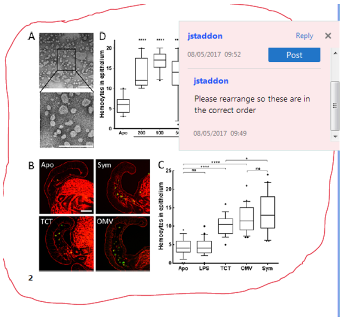
Drawing tools available on comment ribbon

7. Drawing Markups Tools – for drawing shapes, lines, and freeform annotations on proofs and commenting on these marks.

Allows shapes, lines, and freeform annotations to be drawn on proofs and for comments to be made on these marks.

How to use it:

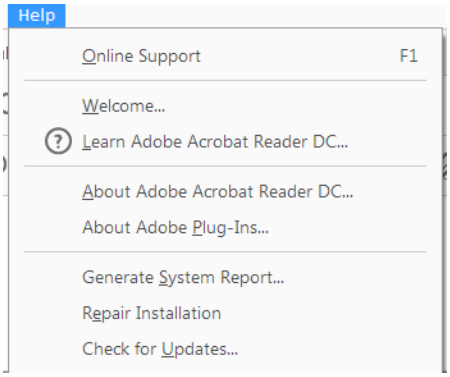
- Click on one of the shapes in the Drawing Markups section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, right-click on shape and select *Open Pop-up Note*.
- Type any text in the red box that appears.



jstaddon
08/05/2017 09:52

jstaddon
Please rearrange so these are in the correct order
08/05/2017 09:49

For further information on how to annotate proofs, click on the Help menu to reveal a list of further options:



Help

Online Support F1

Welcome...

Learn Adobe Acrobat Reader DC...

About Adobe Acrobat Reader DC...

About Adobe Plug-Ins...

Generate System Report...

Repair Installation

Check for Updates...

Author Query Form

Journal: MEN

Article: 12746

Dear Author,

During the copyediting of your manuscript the following queries arose.

Please refer to the query reference callout numbers in the page proofs and respond to each by marking the necessary comments using the PDF annotation tools.

Please remember illegible or unclear comments and corrections may delay publication.

Many thanks for your assistance.

AUTHOR: Please note that missing content in references have been updated where we have been able to match the missing elements without ambiguity against a standard citation database, to meet the reference style requirements of the journal. It is your responsibility to check and ensure that all listed references are complete and accurate.

Query reference	Query	Remarks
1	AUTHOR: Please verify that the linked ORCID identifier is correct for this author.	
2	AUTHOR: Please confirm that given names (red) and surnames/family names (green) have been identified correctly.	
3	AUTHOR: Please check whether the term “harens“is OK in the sentence “Paternal half-sibs. . .that of harens.”	
4	AUTHOR: Please provide the volume number for reference Peixoto et al. (2009).	
5	AUTHOR: Figure 3 has been saved at a low resolution of 224 dpi. Please resupply at 600 dpi. Check required artwork specifications at https://authorservices.wiley.com/asset/photos/electronic_artwork_guidelines.pdf	
6	AUTHOR: Figure 4 has been saved at a low resolution of 172 dpi. Please resupply at 600 dpi. Check required artwork specifications at https://authorservices.wiley.com/asset/photos/electronic_artwork_guidelines.pdf	
7	AUTHOR: Figure 5 has been saved at a low resolution of 256 dpi. Please resupply at 600 dpi. Check required artwork specifications at https://authorservices.wiley.com/asset/photos/electronic_artwork_guidelines.pdf	
8	AUTHOR: Figure 1 is of poor quality. Please check required artwork specifications at https://authorservices.wiley.com/asset/photos/electronic_artwork_guidelines.pdf	
9	AUTHOR: Figure 6 is of poor quality. Please check required artwork specifications at https://authorservices.wiley.com/asset/photos/electronic_artwork_guidelines.pdf	

10	**AUTHOR: If you would like the figures in your article to appear as colour in print, please promptly post or courier the completed hard copy of the Colour Work Agreement Form (including payment information) to this mailing address:** **Customer Services (OPI)** **John Wiley & Sons Ltd** **European Distribution Centre** **New Era Estate, Oldlands Way** **Bognor Regis** **West Sussex** **PO22 9NQ** **The form and charge information can be found online at: http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1755-0998/homepage/ForAuthors.html**	

Funding Info Query Form

Please confirm that the funding sponsor list below was correctly extracted from your article: that it includes all funders and that the text has been matched to the correct FundRef Registry organization names. If a name was not found in the FundRef registry, it may not be the canonical name form, it may be a program name rather than an organization name, or it may be an organization not yet included in FundRef Registry. If you know of another name form or a parent organization name for a "not found" item on this list below, please share that information.

FundRef name	**FundRef Organization Name**
Fundação de Amparo a Pesquisa de Minas Gerais	,

Received: 20 April 2017 | Revised: 4 December 2017 | Accepted: 14 December 2017

DOI: 10.1111/1755-0998.12746

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Reducing cryptic relatedness in genomic data sets via a central node exclusion algorithm

Pablo A. S. Fonseca[1] | Thiago P. Leal[1] | Fernanda C. Santos[1] | Mateus H. Gouveia[1] | Samir Id-Lahoucine[2] | Izinara C. Rosse[1] | Ricardo V. Ventura[2,3] | Frank A. T. Bruneli[4] | Marco A. Machado[4] | Maria Gabriela C. D. Peixoto[4] | Eduardo Tarazona-Santos[1] | Maria Raquel S. Carvalho[1]

[1]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[2]Center for Genetic Improvement of Livestock, University of Guelph, Guelph, ON, Canada

[3]Beef Improvement Opportunities, Guelph, ON, Canada

[4]Embrapa Dairy Cattle, Juiz de Fora, MG, Brazil

Correspondence
Maria Raquel S. Carvalho, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil.
Email: mraquel@icb.ufmg.br

Funding information
This study was supported by funding from Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Empresa Brasileira de Pesquisa Agropecuária (Embrapa). MG was supported by FAPEMIG-CVZ APQ 01353 and CVZ APQ 3182-5.04/07. MR has a fellowship from CNPq—312068/2015-8 and was supported by CNPq—505338/2008-A and 481018/2008-5 projects. MG, RV and MA have fellowships from FAPEMIG. PA, FC and IC have CAPES fellowships.

Abstract

Cryptic relatedness is a confounding factor in genetic diversity and genetic association studies. Development of strategies to reduce cryptic relatedness in a sample is a crucial step for downstream genetic analyses. This study uses a node selection algorithm, based on network degrees of centrality, to evaluate its applicability and impact on evaluation of genetic diversity and population stratification. 1,036 Guzerá (*Bos indicus*) females were genotyped using Illumina Bovine SNP50 v2 BeadChip. Four strategies were compared. The first and second strategies consist on a iterative exclusion of most related individuals based on PLINK kinship coefficient (*φij*) and VanRaden's *φij*, respectively. The third and fourth strategies were based on a node selection algorithm. The fourth strategy, *Network G matrix*, preserved the larger number of individuals with a better diversity and representation from the initial sample. Determining the most probable number of populations was directly affected by the kinship metric. Network *G matrix* was the better strategy for reducing relatedness due to producing a larger sample, with more distant individuals, a more similar distribution when compared with the full data set in the MDS plots and keeping a better representation of the population structure. Resampling strategies using VanRaden's *φij* as a relationship metric was better to infer the relationships among individuals. Moreover, the resampling strategies directly impact the genomic inflation values in genomewide association studies. The use of the node selection algorithm also implies better selection of the most central individuals to be removed, providing a more representative sample.

KEYWORDS
bovine, cryptic relatedness, genetic diversity, inbreeding, population genetic structure

1 | INTRODUCTION

Recently, the problems to obtain a truly random sample from a natural population and the consequences of this problem in the downstream genetic analyses have been highlighted (Peterman, Brocato, Semlitsch, & Eggert, 2016). Natural populations are composed of networks of individuals that are characterized by differences in gene flow. The presence of population stratification or cryptic relatedness in a sample used for genetic diversity estimates or genetic association studies can result in spurious results. Cryptic relatedness is an important confounding factor in genetic diversity studies, resulting in false bottleneck signals and erroneous estimates of the effective population size (Chikhi, Sousa, Luisi, Goossens, & Beaumont, 2010). In genetic association studies, cryptic relatedness is a problem for populations which have grown rapidly and recently from founder populations with small effective population sizes (Voight & Pritchard, 2005). For bovine populations, this is a common problem to be considered. Moreover, the presence of cryptic relatedness in a sample used for Genome-Wide Association Study (GWAS) violates the assumption of independence among the genetic variants observed in individuals that compose the sample. In recent years, some methodologies have been developed to correct the problem of cryptic relatedness in genetic association studies, mainly for GWAS (Astle & Balding, 2009; Hoffman, 2013; Kirkpatrick & Bouchard-Côté, 2016; Morrison, 2013; Price, Zaitlen, Reich, & Patterson, 2010; Tucker, Price, & Berger, 2014; Wang, Hu, & Peng, 2013). However, eliminating the effect of cryptic relatedness in a sample is not a simple process (Sillanpää, 2011). For example, the use of principal components in linear models, a very common strategy to correct the effect of population stratification, does not correct for the presence of cryptic relatedness (Price et al., 2006). In addition, most methodologies used to estimate genetic diversity in populations do not correct for cryptic relatedness.

Several studies have already described that SNPs used for genomic selection can, in addition to capturing the linkage disequilibrium (LD) between SNPs and quantitative trait loci (QTL), also capture family relationships among individuals (Clark, Hickey, Daetwyler, & van der Werf, 2012; Habier, Tetens, Seefried, Lichtner, & Thaller, 2010; Yee, Rogell, Lemos, & Dowling, 2015). It has also been demonstrated that the reliability of genomic predictions is subject more to effects of the level of family relationship in the sample than to LD (Wientjes, Veerkamp, & Calus, 2013). Therefore, developing strategies to reduce relatedness levels in samples, particularly when extracted from inbred populations, becomes important for reducing spurious results in the genomic selection. However, it is important to highlight that the level of relatedness of the individuals excluded is directly related to the genetic architecture of the trait and the population genetic structure.

Cattle offer an interesting model for evaluating methods for reducing relatedness in a sample. Bovine breeding programmes are based on the extensive use of specific animals. Frequently, sires in one generation descend from the most important sires in the previous generations. However, many bulls in one generation do not contribute to the next. Paternal half-sibs are common, and the population genetic structure resembles that of harens. Usually, cows have a much smaller number of progenies. Due to artificial selection, bovine pedigrees are usually highly complex and the impacts depend on the size of the breed and the selection intensity. In addition, reproductive life is long in both sexes and there is generation overlapping. Conservation efforts have been taken to preserve genetic diversity in commercial herds by the inclusion of less related bulls in the reproduction schemes. However, as breeding values evolve, it is increasingly difficult to insert animals that are not related to top ranked bulls, without losing breeding values.

For example, milk selection programmes are frequently based on the evaluation of the larger number of daughters or granddaughters of specific sires. In systems based on *multiple ovulation and embryo transfer* (MOET), an even smaller number of animals are selected to contribute to the next generation (Nicholas & Smith, 1983; Pedersen et al., 2012; Peixoto, Verneque, Teodoro, Penna, & Martinez, 2006). In 1994, a nation-wide breeding programme for the Guzerá (*Bos indicus*), based on progeny testing and a MOET selection nucleus scheme, was implemented in Brazil to improve milk production (Peixoto, Verneque, Pereira, Machado, & Carvalho, 2009; Somashekar, Selvaraju, Parthipan, & Ravindra, 2015; Speizer & Lance, 2015). The breed was subjected to an intense selection process that could potentially have resulted in inbreeding. Indeed, the breed had already been subjected to a series of bottlenecks, including its importation to Brazil in the 19th century, the extensive use of the breed to produce cross-breds in the 1930s and the closure of the registry books in the 1980s. Therefore, the Guzerá provides an interesting model for genetic diversity and population stratification studies due to their recent history of genetic diversity. In this context, obtaining an unrelated, or at least distantly related, sample is a hard task.

The selection of the individuals that will reproduce is a sampling process itself. In this context, methodologies such as *best linear unbiased predictor* (BLUP), which is based on the *best linear unbiased estimator* (BLUE), are used and may result in an increase of the inbreeding For example, it has been shown that using BLUP, without a correction for inbreeding levels, may increase the inbreeding in an intensity which is inversely proportional to the heritability of the trait (Khaw, Ponzoni, & Bijma, 2014). Alternative strategies for evaluating and reducing relatedness levels in the sample are needed.

In this study, we evaluate four strategies for selecting least related individuals in a sample. The final samples obtained using each strategy were compared to each other and to the initial sample, in order to evaluate the impact of these strategies on genetic diversity estimates. Moreover, the samples were also compared to each other to identify the strategy which best represents the genetic structure of the initial sample, however, with no significant relatedness among individuals. The heuristic strategy proposed by (Kehdy et al., 2015), based on the exclusion of the most central individuals present in a kinship coefficients network, provided the best resampling strategy. This strategy helps to identify the most endogamic individuals present in the sample and to select the individuals which retain the greatest part of the genetic variability. Furthermore, resampling allows the development of breeding strategies to reduce inbreeding and, consequently, decreases the effects of inbreeding depression observed in populations subjected to intensive artificial selection.

2 | MATERIAL AND METHODS

2.1 | Ethics statement

This study was performed following approval by the Embrapa Dairy Cattle Ethical Committee of Animal Use (CEUA-EGL), under Protocol

Number 09/2014. In addition, all experimental procedures were conducted in accordance with the recommendations of the Embrapa Dairy Cattle Ethical Committee of Animal Use.

2.2 | Sample and genotyping

One thousand and thirty-six (1,036) cows, the full data set, from the six main herds of the Guzerá Progeny Test and MOET Milk Selection Programs, were included in this sample. These animals are part of a selection scheme using the granddaughter design, in which a bull is mated to several cows. Therefore, the most frequents relationships are half-sisters, half-aunts, half-nieces, granddaughters and cousins. As some of the bulls descend from common ancestors, relatedness is even more complex. The animals were genotyped using the Illumina Bovine SNP50 v2 BeadChip (Illumina Inc., San Diego, CA). The bovine genome is distributed in 31 chromosomes (29 autosomes and the sexual pair). A detailed description on the structure of the bovine genome can be found in the NCBI genome ID:82 (https://www.ncbi.nlm.nih.gov/genome/?term=82).

2.3 | Identity by descent (IBD) estimates

To calculate the IBD estimates for the full data set, markers were excluded from the analyses when: the map position was unknown or nonautosomal, MAF < 0.01, Call Rate < 0.95 and they presented linkage disequilibrium (r^2) $> .2$ with any other marker from the whole data set. After this filtering for the 1,036 individuals, full data set sample, 11,264 markers were kept. This subset of markers was used in the IBD estimates, using the function in PLINK v1.07 (Purcell et al., 2007) and the methodology proposed by (VanRaden, 2008).

2.4 | Relatedness analyses

After the IBD was estimated, four different strategies were compared in the assessment of family structure in the sample. These strategies were chosen to reduce the level of family structure in the data and to eliminate the smallest possible number of individuals.

The first and second strategies were based on the pairwise kinship coefficients (φij) estimated using PLINK v1.07 (Purcell et al., 2007) and VanRaden's formula (VanRaden, 2008), respectively. For both strategies, a threshold of $\varphi ij \geq 0.1$ was assumed as a criterion for considering pairs of individuals to be closely related. This threshold allows identification, from the full data set, of pairs of first-, second- and third-degree relatives. Individuals were excluded in an iterative way, where individuals with higher numbers of $\varphi ij \geq 0.1$ values with other subjects in the sample were eliminated in each step (adapted from: Reed et al., 2015). The samples obtained using these strategies were named *Threshold IBD* and *Threshold G matrix*.

The third and fourth strategies for reducing family structure in the sample were based on a network approach shown by (Kehdy et al., 2015) and implemented in the NaTora software (unpublished). The approach used in the third and fourth strategies works in a multistep process. First, the relationship among the individuals of the sample is represented in a network, where each node is an individual and each edge is the relationship metric between two individuals. Second, the degree of centrality (a metric that represents the number of nodes connected to this node) of each node in the network is calculated and the node with the highest degree of centrality is excluded. At this point, we randomly select the node to be eliminated in those cases where the nodes have the same degree of centrality. Finally, when only pairs of nodes and disconnected nodes exist, the algorithm returns to the initial network and, for each pair of nodes, verifies which of the two nodes had more edges initially and eliminates it. In the end, only unrelated individuals remain in the sample. The third and fourth strategies are distinct; however, in that in the third strategy, the families within the sample were modelled like a network, where each node is an individual connected to the others by edges, representing PLINK $\varphi ij > 0.1$ (Figure 1a). In the fourth strategy, the edges between the individuals were based on the values obtained using VanRaden's φij (VanRaden, 2008). VanRaden's φij was divided by 2 to facilitate the comparison among the results obtained from the kinship coefficient estimates in PLINK (0-0.5). Consequently, similar to the third strategy, a threshold of $\varphi ij \geq 0.1$ was stipulated to connect individuals (Figure 1b).

Using the third and fourth strategies, we could eliminate all family clusters by successively eliminating higher central nodes. Thus, third and fourth new samples were generated and named as *Network IBD* and *Network G matrix*, respectively. In the resampling process, only the 11,264 markers that fulfilled the criteria adopted for IBD estimates in the *All Animals* sample were used. In the subsequent

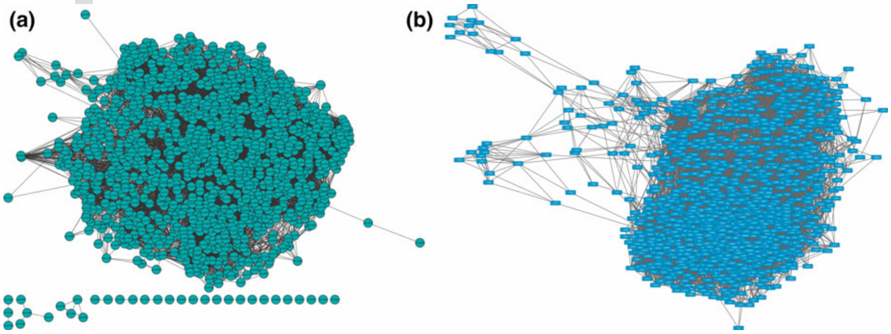


FIGURE 1 Network clustering all individuals in family groups. Nodes represent individuals and edges represent kinship coefficients higher than .1 for the IBD (a) and G matrix (b) approaches

analysis, all markers in the 50k panel were used and specific filtering was performed for each analysis.

2.5 | Intra- and intersample comparisons

2.5.1 | Similarity among individuals in the full data set vs. each sample

We calculated the degree of similarity among the individuals based on a multidimensional scaling (MDS) method (Kruskal, 1964). First, the number of opposite homozygotes was estimated for each pair of individuals in the sample matrix. Second, the position of each pair in this matrix was used to calculate the Euclidean distances between individuals. At the end of this analysis, the matrix with the Euclidean distance between each pair of individuals was used to plot the MDS distance among the individuals in the sample. For this analysis, only markers in autosomes, having MAF > 0.01 and Call Rate > 0.95, were used (at this moment, no LD pruning for the markers was performed).

The R packages pvclust, ape and the R base function hclust (Paradis, Claude, & Strimmer, 2004; Suzuki & Shimodaira, 2006) were used to represent the hierarchical clustering among the animals in the full data set and in each sample. The number of opposite homozygotes was also used to construct this clustering. The most probable number of clusters for each sample was defined as the step with the largest increase in height values.

2.5.2 | Linkage disequilibrium decay and effective population size (Ne) in the full data set vs. each sample

The r^2 fast algorithm in the GenABEL package (Aulchenko, Ripke, Isaacs, & Van Duijn, 2007) was used to estimate linkage disequilibrium (LD) by the r^2 statistic (Hill & Robertson, 1968). Moreover, the patterns of LD decay in the full data set and in each sample were calculated using the following distance intervals between markers (in Kb): 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-40, 40-50, 50-75, 75-100, >100. In these analyses, only syntenic markers were evaluated. Additionally, the portions of markers in strong LD ($r^2 > .3$) were measured in the full data set and in each sample for each distance interval between markers. The subset of markers used for this analysis was composed by markers with autosomal positions, MAF > 0.01 and Call Rate > 0.95 (*All Animals:* 32,194 markers; *Threshold IBD:* 32,802 markers; *Threshold G matrix:* 32,680 markers; *Network IBD:* 32,809 markers; *Network G matrix* 32,022 markers). Additionally, the effective population size (Ne) was estimated for each sample from five to 30 generations ago using the relationship between the distance c, r^2 and Ne, assuming absence of mutation (Sved, 1971).

2.5.3 | Detection of population structure and influence of relatedness level

The software Admixture 1.23 (Alexander, Novembre, & Lange, 2009) was used to evaluate the genetic structure of the samples, using fivefold cross-validation to identify the most likely number of components. The most probable number of populations (*K*) was defined by the smallest cross-validation error value. For each sample, the subset of markers used for this analysis was composed by markers with autosomal positions, MAF > 0.01, Call Rate > 0.95 and that presented linkage disequilibrium (r^2) < .2 with any other marker in the whole data set.

2.5.4 | Genomewide association study (GWAS) simulation

The impact of the relatedness level on the GWAS results was estimated using a simulation approach. Thirty QTLs were simulated across the 30 chromosomes in the bovine genome for one thousand replications using two heritability values, h^2 = 0.2 and h^2 = 0.5, separately. The simulation was performed twice. In the first approach, the simulated phenotypic values were obtained only for the *All Animals* sample (one thousand phenotypes with h^2 = 0.2 and one thousand phenotypes with h^2 = 0.5). After this step, the correspondent phenotypic value for each simulation was extracted for the animals present in each of the resampling samples (*Threshold IBD*, *Threshold G matrix*, *Network IBD* and *Network G matrix*). In the second approach, the simulation was performed independently for each of the samples to verify biases caused by the different sampling strategies tested. Furthermore, for each of the five samples, two thousand more groups of simulated phenotypes (one thousand phenotypes with h^2 = 0.2 and one thousand phenotypes with h^2 = 0.5) were also obtained. A schematic representation of the two simulation scenarios is shown in Figures S1 and S2, respectively. The additive allelic effect of each QTL was sampled from a standard Gaussian distribution, and the sum of all QTL effects was rescaled to generate an additive genetic variance, adjusted to each simulated h^2 (Casellas & Piedrafita, 2015). Phenotypic records were obtained by adding a residual from a normal distribution with mean of 0 and variance equal to the environmental variance (*Ve*) of the QTL effects.

The GWAS was performed for each replicate in each group using the –assoc function implemented in Plink v.1.07. At this moment, the markers present in the 30 simulated QTLs were removed from the GWAS to estimate the GWAS inflation value (lambda) created by secondary associated signals. The lambda is the ratio between the observed median of the GWAS *p*-values and the expected median of the gwas *p*-values. In addition, the descriptive statistic for the lambda values obtained in each simulated GWAS was calculated for the first simulation scenario.

3 | RESULTS

3.1 | Identity by descendent estimates

After IBD estimation of the *All Animals* sample, 536,130 pairwise combinations were obtained, of which 14,207 had a $\varphi ij \geq 0.1$. Using the *Threshold IBD* strategy, after eliminating individuals having $\varphi ij \geq 0.1$, only 203 of the 1,036 individuals in the *All Animals* sample were retained in the *Threshold IBD* sample. For the *Threshold G matrix* strategy, the final sample had 286 individuals.

Network centrality analysis, implemented in the NaTora software, resulted in the retention of 210 individuals in the *Network IBD* sample and 286 individuals in the *Network G matrix* sample. The VanRaden's φ*ij* obtained for the *All Animals* sample showed that, for all combinations among individuals, 9,743 had a φ*ij* ≥ 0.1.

The individuals remaining in each sample were compared to evaluate the final composition obtained using each approach. Regarding the individuals of the four new samples, results of this comparison indicated that only 22 cows were the same in the four samples (Figure 2). In addition, the higher number of shared individuals was observed between the samples obtained using the same kinship metric (*Threshold IBD Region* and *Network IBD*: 68 individuals; *Threshold G matrix* and *Network G matrix*: 121 individuals). Furthermore, in the *All Animals* sample, the mean of φ*ij* = 0.0187 ± 0.028. As expected, a decrease in the mean of φ*ij* was observed in the four filtered samples, as shown in Table 1.

3.2 | Intra- and intersample comparisons

The MDS plots show that the *Threshold G matrix* and *Network G matrix* individuals are more widely distributed. These results are shown in Figure 3, where the number of opposite homozygous genotypes is evaluated. Hierarchical cluster analysis shows that the basic structure of the dendrogram is retained independent of the resampling approach (Figure 4). The numbers of clusters present in each sample were as follows: *All Animals*, 8 (largest height increase = 12982.16); *Threshold IBD*, 47 (largest height increase = 3066.28); *Threshold G matrix*, 26 (largest height increase = 4987.06); *Network IBD*, 47 (largest height increase = 3156.46); and *Network G matrix*, 13 (largest height increase = 6275.18). In Figure 4, the red lines indicate the point, where the largest increase in the height of each dendrogram was observed, highlighting the point where the best separation of the groups was obtained. A similar number of clusters between the hierarchical cluster analysis and admixture analysis were observed only for the *Network G matrix* sample. In Figure 4, the groups identified by hierarchical cluster analysis in the *All Animals* sample were plotted in the MDS plot for each subsample. This correspondence analysis points that the *Network G matrix* sample was the only sample that retained individuals from all eight clusters. Moreover, the proportion of individuals in each cluster was similar to the proportion observed in the *All Animals* sample (Table 1).

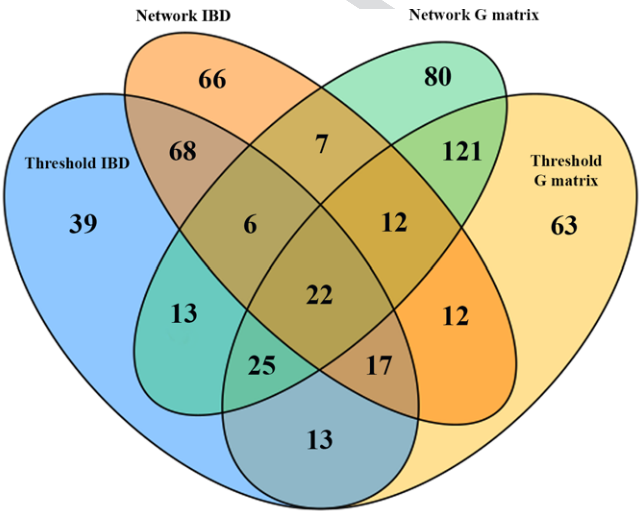


FIGURE 2 Venn diagram showing individuals shared among samples

When the linkage disequilibrium (LD) was evaluated, it was noted that the value of r^2 at distances between markers>100 Kb reached 0.1. It is important to observe that the *Threshold IBD* and *Network IBD* samples produce very similar effects on r^2 and, consequently, on LD decay (Figure 5). The *Threshold G matrix* and *Network G matrix* samples also produce very similar results. However, in general, the five samples produce very similar LD decay (Figure 5). For all five samples, a higher percentage of markers in strong LD (% $r^2 > .3$) was observed in the intervals 0-5, 10-15 and 15-20 Kb. Furthermore, Figure 5 shows that the % $r^2 > .3$ follows the LD decay pattern across the different distances between markers. Additionally, there was not observed substantial differences for the Ne across generations among all the samples.

The most probable number of populations, estimated by the ADMIXTURE 1.23 software, shows a strong impact of the cryptic relatedness in the detection of population stratification. For the *All Animals* sample, the most probable number of populations (smallest value of Cross-validation error) is *K* = 75. However, for the *Threshold IBD* and *Network IBD* samples, the most probable numbers of populations are *K* = 3 and *K* = 2, respectively. For the *Threshold G matrix* and *Network G matrix* samples, the smallest value of Cross-validation error is *K* = 14, but the difference observed among the values from the 11-16 populations was small. These results are shown in Figure 6. Therefore, any one of these populations in the *Threshold G matrix* and *Network G matrix* samples have virtually the same probability to be correct.

3.3 | Impact of the relatedness level on GWAS—Simulation analysis

The impact of the relatedness level on the GWAS results is shown in Figure 7. In the GWAS, the expected value of lambda is 1, in the absence of association. As associated markers were removed, only secondary effects, such as LD caused by relatedness between individuals in the sample, would increase lambda values. The highest lambda values were identified for the *All Animals* sample for both heritability values (lambda = 1.57 for h^2 = 0.2 and lambda = 2.371 for h^2 = 0.5). These results indicate strong inflation of the GWAS results. This further indicates that this inflation is related to the heritability values. When the heritability of the trait increases, the lambda also increases. However, for all other samples, there were no differences between the lambda values obtained for both simulations (Figure 7a), h^2 = 0.2 and h^2 = 0.5, in the first approach (Figure S1).

TABLE 1 Number of animals (and proportion) present in each group (1-8) identified by the hierarchical cluster analysis for the *All Animals* sample, in each one of the subsamples

	Groups							
	1	2	3	4	5	6	7	8
All Animals	419 (0.4)	293 (0.28)	48 (0.005)	116 (0.11)	39 (0.04)	110 (0.11)	6 (0.005)	5 (0.005)
Threshold IBD	57 (0.28)	77 (0.38)	14 (0.07)	6 (0.03)	21 (0.1)	24 (0.12)	4 (0.02)	0
Threshold G matrix	117 (0.41)	70 (0.24)	17 (0.06)	42 (0.15)	7 (0.02)	29 (0.10)	4 (0.02)	0
Network IBD	56 (0.27)	83 (0.4)	14 (0.07)	7 (0.03)	21 (0.1)	24 (0.11)	5 (0.2)	0
Network G matrix	108 (0.4)	89 (0.3)	13 (0.04)	30 (0.1)	9 (0.03)	33 (0.12)	3 (0.016)	1 (0.003)

Additionally, the lambda values for the four resampling samples were close to 1, as expected. The simulations performed in the second approach (Figure S2) retained the relationship between the lambda values and the heritability, in the *All Animals* sample and for the four resampling strategies (Figure 7b). In addition, the lambda values obtained in this scenario for all the resampling samples were higher than the values obtained in the first scenario. In both scenarios, the smallest lambda values were found for the resampling samples obtained using the IBD values calculated using PLINK, independently of the resampling strategy (Threshold IBD and Network G matrix).

4 | DISCUSSION

In cattle, artificial reproduction technologies allow the reduction in the number of animals needed to produce the next generation and the reduction in generation intervals. Consequently, the increase in the relatedness levels in the population is usually a result of the selection process (Macedo et al., 2014; Panetto, Gutiérrez, Ferraz, Cunha, & Golden, 2010). Samples may capture such phenomena and association studies, and may be affected by the population genetic structure. For this reason, it is necessary to evaluate and adjust the relatedness in samples generated by the intense selection process. To conduct this study, we selected a particularly complex sample, composed of large families with large numbers of cousins, half-nieces, half-sisters and granddaughters.

In the present work, four methodologies were compared with the aim of reducing the relatedness level in a sample. The first and second methodologies eliminate, iteratively, the individuals with the largest number of relationships with a PLINK and VanRaden's φij greater than 0.1, respectively. The third and fourth methodologies use a more elaborate approach and perform a network relationship analysis that allows the elimination of the more central individuals of each network formed by a $\varphi ij \geq 0.1$ and VanRaden's $\varphi ij \geq 0.1$. This was carried out using a node selection algorithm based on a network's degree centrality statistic. The samples obtained with the centrality algorithm are expected to be more representative of the original genetic variability, as compared to the *Threshold IBD* and *Threshold G matrix* approaches. This happens because the more central animals, that is, those with more relatives in the sample, which would have been eliminated from the network, share a portion of the genome with the remaining animals. Thus, a portion of the genetic variability of the animals that have been eliminated will remain in the final sample. The results obtained in this study reinforce the necessity of adjusting the relatedness level in a sample. Moreover, it shows that a methodology already demonstrated to be efficient, for studies with human populations (Kehdy et al., 2015), works satisfactorily with a structured livestock sample.

The MDS analysis for the four samples shows that the resampling strategies retained individuals with fewer genetic similarities when compared with the *All Animals* sample. This result was expected because, in the four samples, only individuals with a PLINK φij or a VanRaden's $\varphi ij < 0.1$ were retained. The higher genetic similarity in the *All Animals* sample contributes to the higher mean φij and a higher r^2 average at all distances between markers in the *All Animals* sample. The MDS plots reflect the Euclidian distances among individuals calculated using the number of opposite homozygous markers (Figure 3). The relationship metric strongly affects the distance among individuals (Figure 3). The samples obtained using the PLINK φij show similar patterns; the same is observed among the samples obtained using VanRaden's φij, independent of the resampling approach. It is important to highlight that, although *Threshold G matrix* and *Network G matrix* samples retaining the same sample size, only *Network G matrix* retained individuals for all groups identified in the hierarchical clustering analysis performed on *All animals*. These results suggest that *Network G matrix* sample retains a more diverse and representative group of individuals.

The process used to determine the kinship coefficient among individuals in the *Threshold G matrix* and *Network G matrix* approaches might explain the results shown in Figure 3. Both approaches use the VanRaden's φij, which uses allelic frequency as a weight to estimate the relationship coefficient among individuals (VanRaden, 2008). Otherwise, the PLINK φij only takes into account the number of alleles shared among them (*Threshold IBD* and *Network IBD*). This way, two pairs of individuals that share the same number of alleles will have the same kinship coefficient. However, pairs sharing higher numbers of rare alleles will have higher relationship coefficients, obtained using the *Threshold G matrix* and *Network G matrix* approach, when compared with pairs sharing predominantly common alleles. These results suggest that the *Network G matrix* approach provides a better choice of the most central and most related individuals in the network.

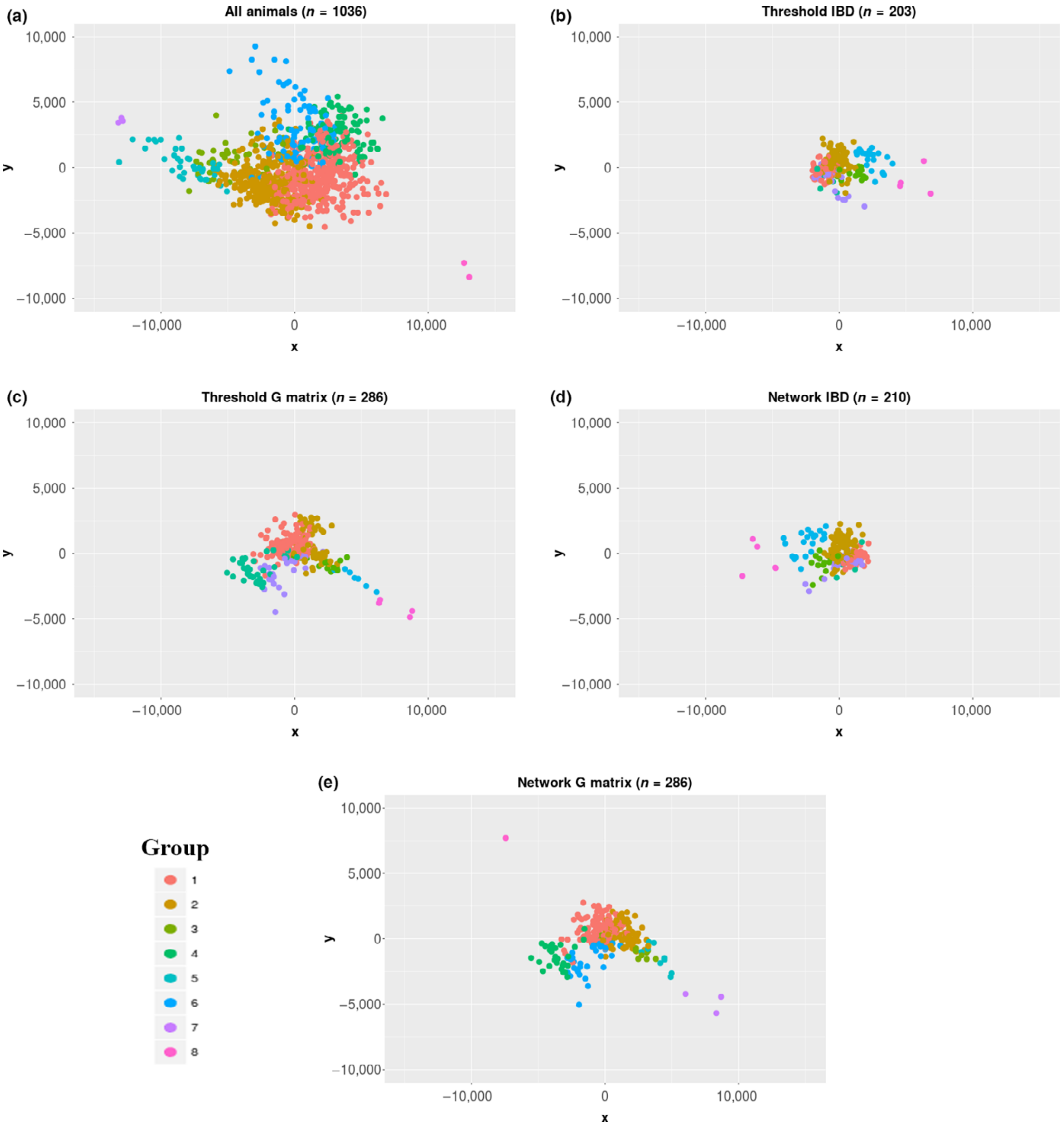


FIGURE 3 Multidimensional scaling (MDS) plots of individuals for each sample. *X* and *Y* coordinates are the output values of the MDS plots and were calculated using the Euclidian distances among the individuals, obtained through the number of opposite homozygous genotypes for each locus. (a) Coordinates *X* and *Y* for the individuals in the All Animals sample were obtained using 1,036 cows and 32,194 markers; (b) Coordinates *X* and *Y* for the individuals in the *Threshold IBD* sample were obtained using 203 cows and 32,802 markers; (c) Coordinates *X* and *Y* for the individuals in the *Threshold G matrix* sample were obtained using 286 cows and 32,680 markers; (d) Coordinates *X* and *Y* for the individuals in the *Network IBD* sample were obtained using 210 cows and 32,809 markers; (e) Coordinates *X* and *Y* for the individuals in the *Network G matrix* sample were obtained using 286 cows and 32,022 markers. The colours in the plot represent the eight groups identified in the hierarchical clustering analysis performed on the *All Animal* sample

The differences observed between the MDS plots of *Threshold G matrix* and *Network G matrix* samples, even with the same sample size and the same relationship metric (VanRaden's φ*ij*), might be explained by the resampling process. The difference is produced by the algorithm in the NaTora software. Initially, the NaTora software identifies the nets of related individuals and sequentially excludes

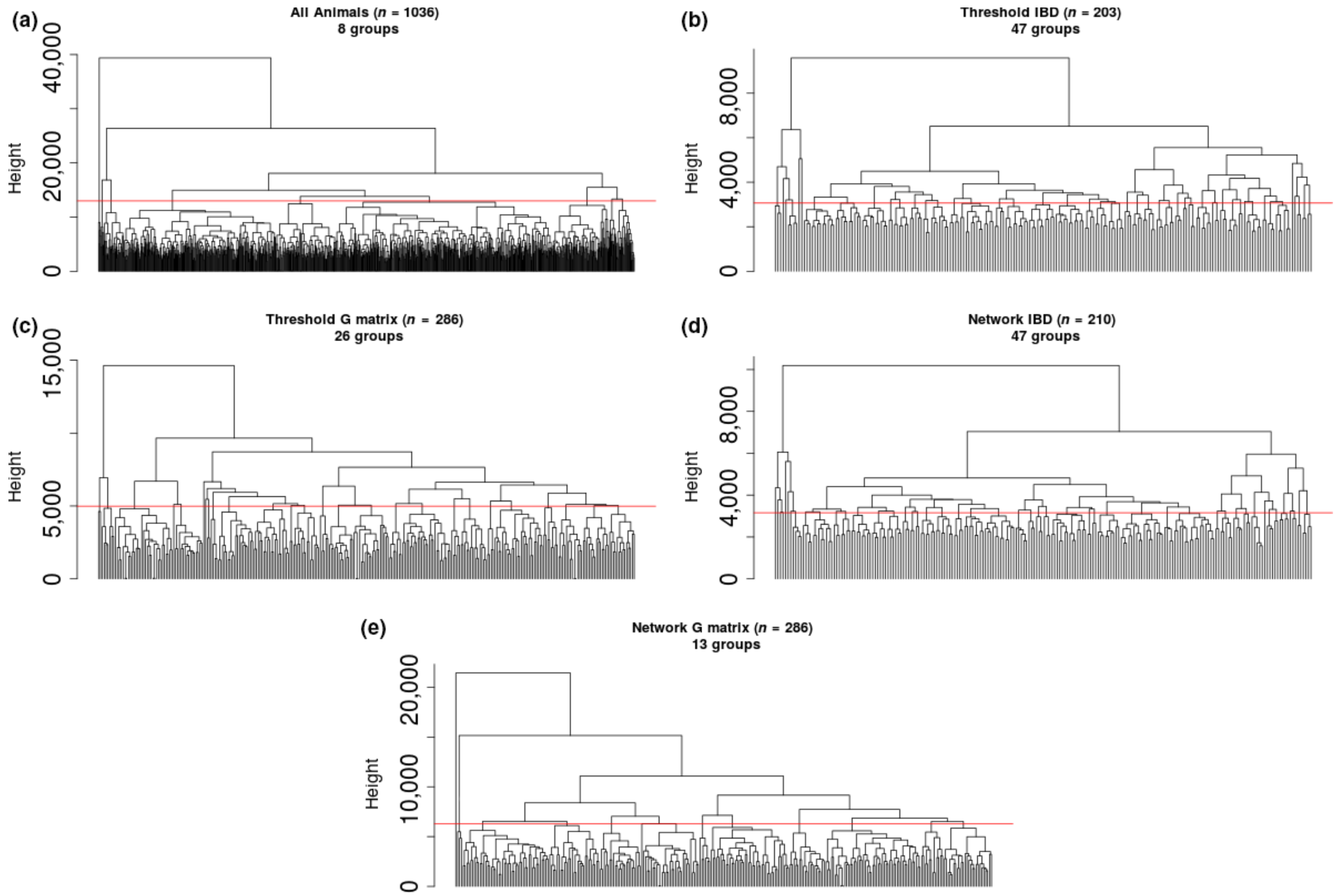


FIGURE 4 Dendrograms showing hierarchical clustering of individuals for (a) *All Animals*, (b) *Threshold IBD*, (c) *Threshold G matrix*, *(d) Network IBD* and (e) *Network G matrix*. The red lines indicate the point where the largest increase in the height of each dendrogram was observed, highlighting the point where the best separation of the groups was obtained. Dendrograms were generated using the number of opposite homozygous genotypes between individuals in each sample

the most central individuals. When the nets are deconstructed, and only pairs of related individuals are present in the sample, the algorithm returns to the original network to choose better individuals to be eliminated based on the initial degree of centrality of each individual. This characteristic of NaTora allows better representation of the initial sample. Differently, the iterative exclusion of individuals used in the *Threshold IBD* and *Threshold G matrix* samples only takes into account the degree of relatedness among individuals in the current step of exclusion. These methodological differences explain why both samples (*Threshold G matrix* and *Network G matrix*) have the same size, but are composed of different individuals.

LD decay analysis shows that the r^2 means and the percentages of markers in strong linkage disequilibrium were very similar for the *All Animals* and *Network G matrix* samples at almost all distances between pairs of markers. These results suggest a higher similarity between these two samples in relation to the other samples (*Threshold IBD* and *Network IBD*). Thus, they point to better representation of the original sample in the *Network G matrix* sample. In addition, it was possible to observe a similar pattern of LD decay among the samples obtained after the use of the same kinship metric (*Threshold IBD* and *Network IBD* for PLINK *φij;* and *Threshold G matrix* and *Network G matrix* for VanRaden's *φij*). Moreover, the Venn diagram in Figure 2 shows that samples obtained using the same kinship metric share more individuals. Consequently, these samples are more genetically similar. These results reinforce the impact of the different kinship metrics on the genetic diversity estimates.

A fivefold, cross-validation analysis was performed using the Admixture 1.23 software to identify the more likely numbers of populations (*K*) in each sample. The smallest cross-validation error value points to *K* = 75 for the *All Animals* sample, *K* = 3 for *Threshold IBD* sample, *K* = 2 for *Network IBD* sample, and *K* = 14 (*K* = 11-16, equally probable) for both *Threshold G matrix* and *Network G matrix* samples. We hypothesize that *K* = 75 reflects a macrofamilial structure because individuals coming from a MOET nucleus are included in the sample. Indeed, it has been shown that Admixture 1.23 detects familial structures in human populations (Kehdy et al., 2015). *K* = 3 and *K* = 2, observed for the *Threshold IBD* and *Network IBD* samples, may reflect selection purposes. Originally, Guzerá was selected only for meat production. In the last few decades, some of the herds have begun to be used for dual-purpose selection (milk and meat) and some lineages have started to be specialized for milk production (Peixoto et al., 2009). On the other hand, for the *Threshold G matrix*

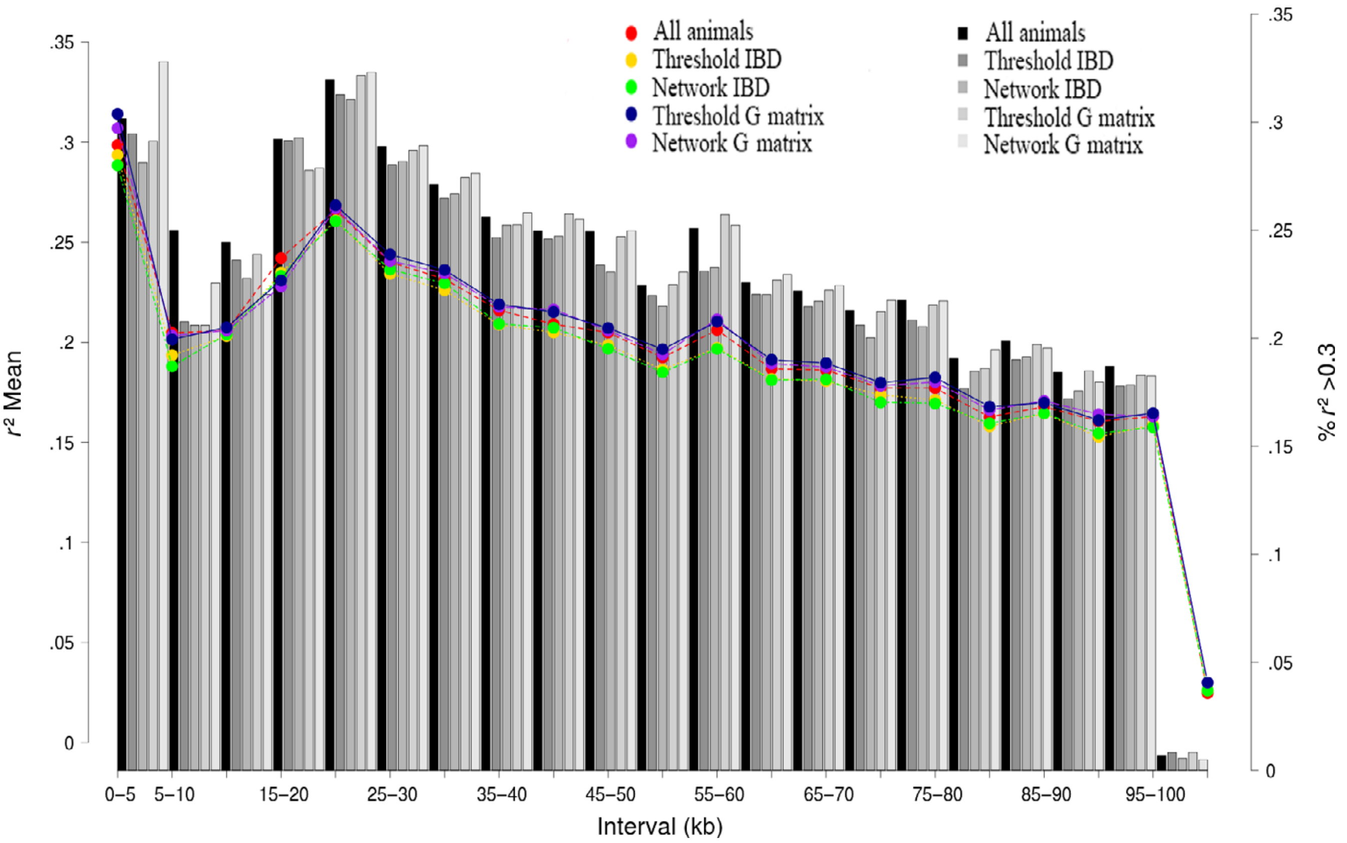


FIGURE 5 LD decay for Guzerá. The left *y*-axis shows the mean of r^2 at different distances between pairs of markers for *All Animals* (red), *Threshold IBD* (yellow), *Network IBD* (green), *Threshold IBD* (blue) and *Network G matrix* (purple). The right *y*-axis shows the percentage of markers in strong LD (r^2 > .3) for *All Animals* and for each sample (grey scale) at each distance between markers

and *Network G matrix* samples, *K* = 14 (*K* = 11-16 equally probable) was the number of populations with the smallest value of cross-validation error. These values, *K* = 11-16, are close to the number of clusters shown in Figure 6. Interestingly, we obtained similar results for the more likely number of Guzerá lineages using microsatellite markers in another sample (15 lineages) (results not shown). These results reinforce the impact of the kinship metric on the resampling processes and the representation of the initial population. *Threshold G matrix* and *Network G matrix*, obtained using two different approaches, resulted in a sample of the same size but with different individuals. However, the population structure detected was similar in the two samples (*K* = 11-16).

The lambda values obtained in the simulation analyses performed in the present work indicate a strong influence of the relatedness over the GWAS inflation. For all GWAS simulations, the associated markers were removed. Therefore, in the absence of secondary signals, a lambda equals 1 was expected. Additionally, the lambda increase is stronger when the heritability of the trait is higher. The median lambda for all the resampling sample is close to the expected lambda (lambda = 1). This result indicates that, independently of the resampling strategy, the reduction in the relatedness level in the sample decreases the number of secondary signals obtained in the GWAS. The strong deviation from lambda = 1, observed for *All Animals*, could be explained by a strong LD present in this sample caused by the high relatedness level. However, the LD comparison performed in the present study demonstrated that there are no differences among the LD patterns among the samples. Additionally, there are no significant differences among the Ne across the generations, reinforcing the results obtained from the LD analysis (Table S1). The impact of LD and heritability over lambda values obtained in GWAS was already evaluated in the literature and follows the same pattern described here (Powell, Visscher, & Goddard, 2010; Speed & Balding, 2015). It is important to highlight that, in the second simulation scheme, where the simulations were performed independently for each sample, the lambda values were obtained using higher heritability (h^2 = 0.5). This suggests that the impact of relatedness reduction over the GWAS inflation is not by chance.

Although there was no significant difference among the lambda values observed in each resampling strategy, the results obtained in the present study reinforce the impact of relatedness level over the GWAS inflation. The *Network G matrix* sample has one of the largest sample sizes among the resampling samples and the more distant individuals, which may be a helpful characteristic for the GWAS analysis. The largest number of individuals and the genetic distances among them may influence the presence of less frequent alleles and increase the association power (Gibson, 2012). The two methodologies tested here, the iterative exclusion of most related individuals (*Threshold*) and the node selection algorithm based on degrees of

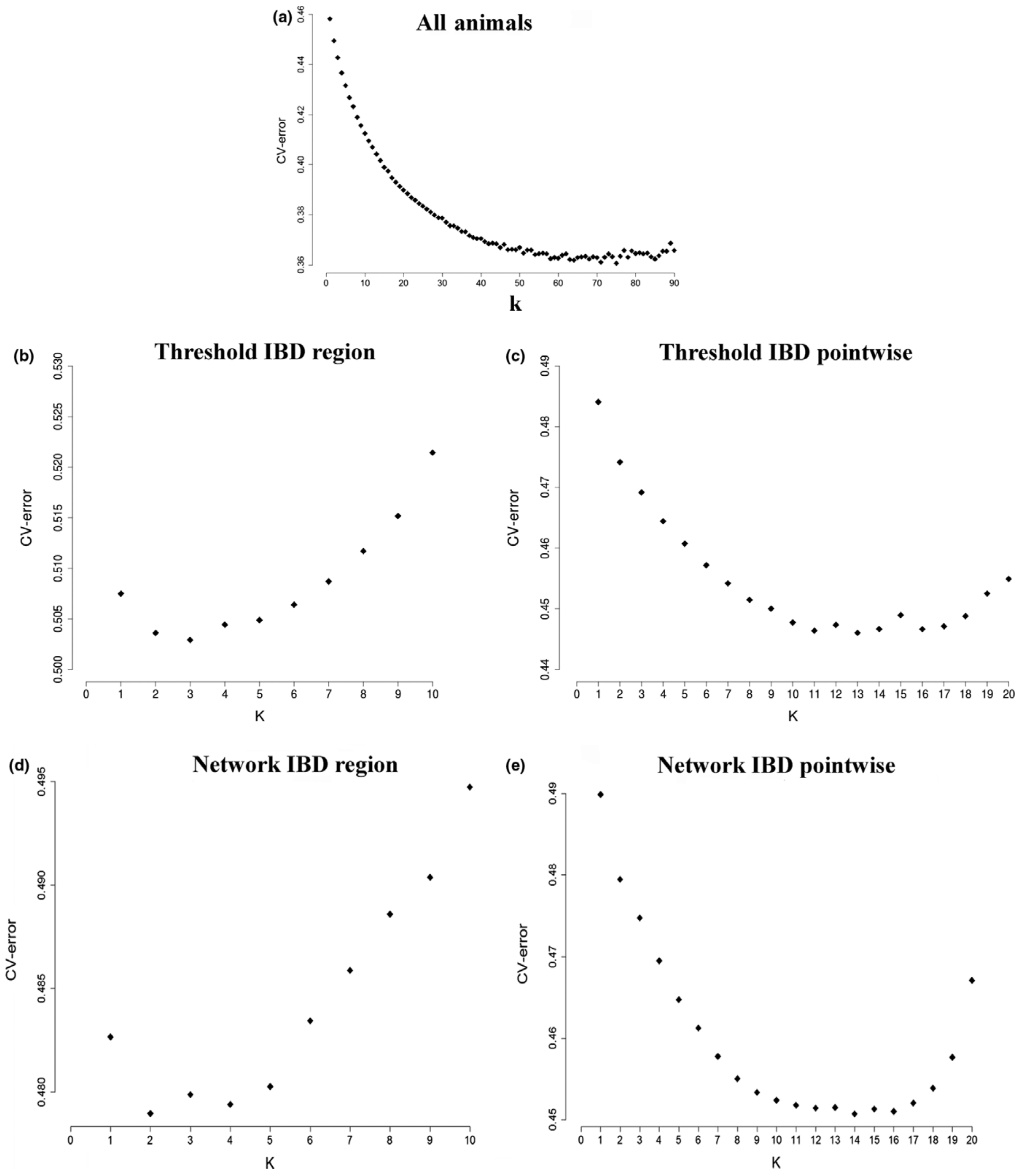


FIGURE 6 Cross-validation error values for each sample. For the *All Animals* sample, (a) the most probable number of populations was *K* = 75; for the *Network IBD Regions* (b) and *Threshold IBD Regions*, (c) the most probable number was *K* = 2; for the *Network IBD Pointwise*, the most probable number of *K* = 14

centrality (*Network*), in spite of performing very similar processes, are different. The threshold approaches needed more user time to complete the analysis. The *Threshold IBD* was obtained after approximately 391 min, and the *Threshold G matrix* was obtained after 302 min. Data S1 shows the R script used to perform the threshold approach. The NaTora approach was more computationally efficient for our data set. For the *Network IBD* sample, NaTora needed 17 s to finish the analysis. For the *Network G matrix* sample, it took less

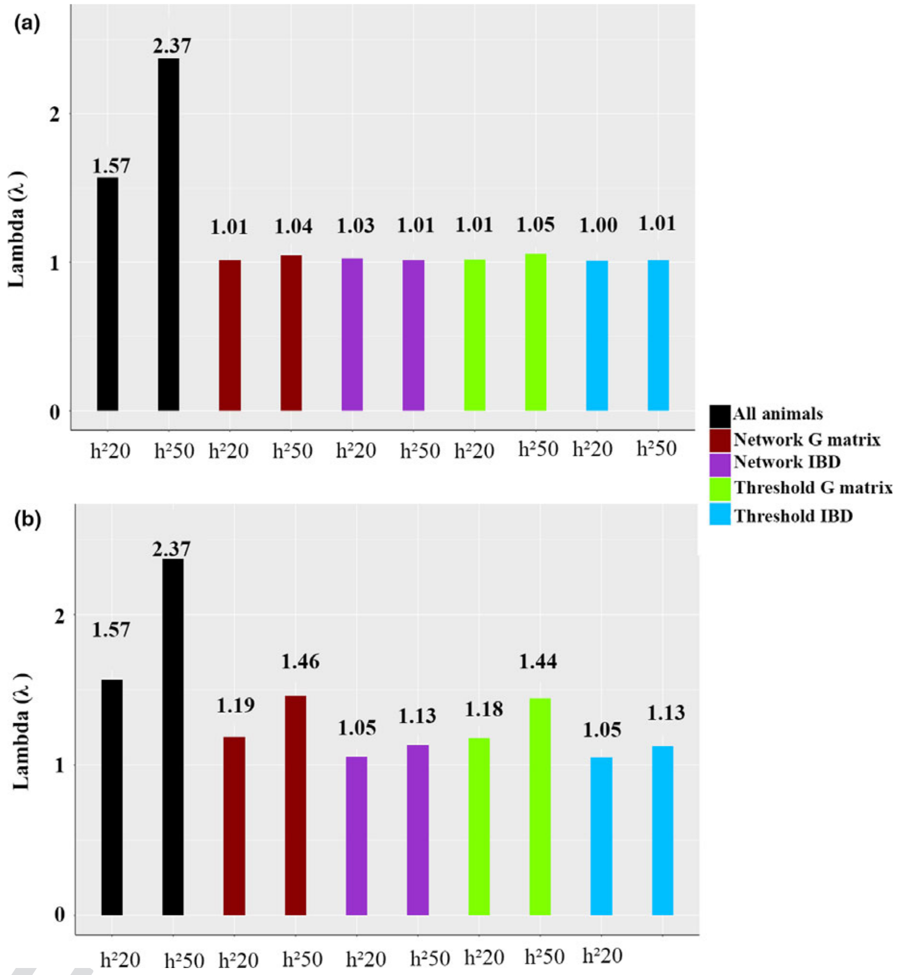


FIGURE 7 Results of genomic inflation (lambda) for each group of 1,000 replications performed by each sample. (a) Median lambda values for each sample using the phenotypic information simulated for the *All Animals* sample with two heritability values (h^2 = 0.2 and h^2 = 0.5). (b) Median lambda values for each sample obtained using the phenotypic information simulated independently for each sample with two heritability values (h^2 = 0.2 and h^2 = 0.5)

than 10 s to obtain the final sample. These results reinforce the higher computational and selection efficiency of the NATORA algorithm. Additionally, the same results also reinforce the impact of the relationship metric over the resampling processes. Using the G matrix coefficient (VanRaden, 208), it was possible to obtain a final sample with a size greater than or equal to all the other samples and taking less user time. All the analyses were performed using the same computer: Dell server with 4 Eight-Core Intel processors, 128 GB of RAM memory (16 × 8 GB) and 600 GB hard drive.

The impact of cryptic relatedness in genetic diversity and genetic association studies has been previously described (Astle & Balding, 2009; Chikhi et al., 2010; Kehdy et al., 2015; Kirkpatrick & Bouchard-Côté, 2016; Tucker et al., 2014; Wang et al., 2013). Compared to the other strategies tested here, the *Network G matrix* is considered the best due to certain characteristics. First, one of the larger samples was obtained, composed of 286 animals (same number of individuals as *Threshold G matrix*). Second, it preserved the genetic diversity observed in the initial sample, indicating good representation of the full data set. Third, this strategy preserved the lineage connections among the individuals (number of populations identified *K* = 11-16) even after excluding closely related individuals. The sample used in the present study originates from a population which had been subjected to several recent bottlenecks (de Souza Fonseca et al., 2016). Guzerá breed was subjected to a strong founder effect during importation from India to Brazil. This is the main cause of these low Ne values in the oldest generations (Table S1). Additionally, an intensive trend to select a small group of sires in population was observed. This characteristic is observed in several bovine breeds. The strength of this trend is enhanced by both the intensive use of artificial insemination and the models applied in the genomic selection (e.g., BLUP). This is one of the main concerns regarding the development of new selection strategies to be applied in the genetic management of herds, or even breeds. These successive reductions in effective population size in the present study might explain the strong reduction in the sample size observed after each resampling strategy was performed. An additional, practical use of this strategy would be the selection of individuals for breeding programmes to preserve, as much as possible, the genetic diversity of the original population. This strategy may help to reduce the impact of inbreeding depression in herds in which genetic diversity levels are low.

Taken together, the results reported here suggest that the node selection algorithm, based on the degree of centrality of a network using VanRaden's φ*ij* as the connection among individuals, was the better strategy for reducing relatedness in a sample enriched by consanguineous individuals. The results obtained in the present study confirm the efficiency of the node selection algorithm in livestock populations and reinforce the impact of the level of relatedness in the sample on the evaluation of population structure and genetic association studies.

ACKNOWLEDGEMENTS

We thank the farmers, who allowed the development of this project on their farms. We thank Mr. Peter Laspina from ViaMundi Escola de Idiomas e Traduções for review language and valuable comments.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Embrapa Dairy Cattle Ethical Committee of Animal Use (CEUA-EGL) under Protocol Number 09/2014. In addition, all experimental procedures were conducted in accordance with the recommendations of the Embrapa Dairy Cattle Ethical Committee of Animal Use.

CONSENT FOR PUBLICATION

All authors have approved the manuscript and agree to its submission to Molecular Ecology Resources.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interests.

AUTHOR'S CONTRIBUTIONS

M.G., F,A., and M.A., were responsible for collecting and genotyping the biological materials. P.A., F.C., M.G., T.P., I.C., R.V., E.T., and M.R., developed and conducted the statistical and genetic diversity tests. S.I., and P.A., were responsible for the GWAS simulations. P.A., F.C., I.C., and M.R., were responsible for biological interpretation of the results and the literature review. P.A., F.C.S., M.G., T.P., and M.R., wrote the manuscript.

AVAILABILITY OF DATA AND MATERIAL

All relevant data are presented within the manuscript. The data sets used and/or analysed during the current study are available on Dryad (https://doi.org/10.5061/dryad.k8b8n).

ORCID

Maria Raquel S. Carvalho http://orcid.org/0000-0002-1744-448X

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109

Astle, W., & Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, *24*, 451–471. https://doi.org/10.1214/09-STS307

Aulchenko, Y. S., Ripke, S., Isaacs, A., & Van Duijn, C. M. (2007). GenABEL: An R library for genome-wide association analysis. *Bioinformatics*, *23*, 1294–1296. https://doi.org/10.1093/bioinformatics/btm108

Casellas, J., & Piedrafita, J. (2015). Accuracy and expected genetic gain under genetic or genomic evaluation in sheep flocks with different amounts of pedigree, genomic and phenotypic data. *Livestock Science*, *182*, 58–63. https://doi.org/10.1016/j.livsci.2015.10.014

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., & Beaumont, M. A. (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, *186*, 983–995. https://doi.org/10.1534/genetics.110.118661

Clark, S. A., Hickey, J. M., Daetwyler, H. D., & van der Werf, J. H. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution*, *44*, 1.

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*, *13*, 135–145. https://doi.org/10.1038/nrg3118

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., & Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution*, *42*, 5. https://doi.org/10.1186/1297-9686-42-5

Hill, W., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*, 226–231. https://doi.org/10.1007/BF01245622

Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PLoS ONE*, *8*, e75707. https://doi.org/10.1371/journal.pone.0075707

Kehdy, F. S., Gouveia, M. H., Machado, M., Magalhães, W. C., Horimoto, A. R., Horta, B. L. ... Rodrigues-Soares, F. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences*, *112*, 8696–8701. https://doi.org/10.1073/pnas.1504447112

Khaw, H. L., Ponzoni, R. W., & Bijma, P. (2014). Indirect genetic effects and inbreeding: Consequences of BLUP selection for socially affected traits on rate of inbreeding. *Genetics Selection Evolution*, *46*(1), 39. https://doi.org/10.1186/1297-9686-46-39

Kirkpatrick, B., & Bouchard-Côté, A. (2016). *Correcting for Cryptic Relatedness in Genome-Wide Association Studies*. arXiv preprint arXiv:1602.07956.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–27. https://doi.org/10.1007/BF02289565

Macedo, A. A., Bittar, J. F., Ronda, J. B., Bittar, E. R., Panetto, J. C., ... Martins-Filho, O. A. (2014). Influence of endogamy and mitochondrial DNA on immunological parameters in cattle. *BMC Veterinary Research*, *10*, 1.

Morrison, J. (2013). Characterization and correction of error in genome-wide IBD estimation for samples with population structure. *Genetic Epidemiology*, *37*, 635–641. https://doi.org/10.1002/gepi.21737

Nicholas, F., & Smith, C. (1983). Increased rates of genetic change in dairy cattle by embryo transfer and splitting. *Animal Science*, *36*, 341–353. https://doi.org/10.1017/S0003356100010382

Panetto, J., Gutiérrez, J., Ferraz, J., Cunha, D., & Golden, B. (2010). Assessment of inbreeding depression in a Guzerat dairy herd: Effects of individual increase in inbreeding coefficients on production and reproduction. *Journal of Dairy Science*, *93*, 4902–4912. https://doi.org/10.3168/jds.2010-3197

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290. https://doi.org/10.1093/bioinformatics/btg412

Pedersen, L. D., Kargo, M., Berg, P., Voergaard, J., Buch, L. H., & Sørensen, A. C. (2012). Genomic selection strategies in dairy cattle breeding programmes: Sexed semen cannot replace multiple ovulation and embryo transfer as superior reproductive technology. *Journal of Animal Breeding and Genetics*, *129*, 152–163. https://doi.org/10.1111/j.1439-0388.2011.00958.x

Peixoto, M. G., Verneque, R., Pereira, M., Machado, M., & Carvalho, M. R. (2009). Impact of milk production breeding program on the Guzerat (Bos indicus) population parameters in Brazil. *Interbull Bulletin*, [illegible]89.

Peixoto, [illegible]erneque, R., Teodoro, R., Penna, V., & Martinez, M. (2006). Genetic trend for milk yield in Guzerat herds participating in progeny testing and MOET nucleus schemes. *Genetics and Molecular Research*, *5*, 454–465.

Peterman, W., Brocato, E. R., Semlitsch, R. D., & Eggert, L. S. (2016). Reducing bias in population and landscape genetic inferences: The effects of sampling related individuals and multiple life stages. *PeerJ*, *4*, e1813. https://doi.org/10.7717/peerj.1813

Powell, J. E., Visscher, P. M., & Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, *11*, 800–805. https://doi.org/10.1038/nrg2865

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909. https://doi.org/10.1038/ng1847

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, *11*, 459–463. https://doi.org/10.1038/nrg2813

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559–575. https://doi.org/10.1086/519795

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P. & Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, *34*, 3769–3792. https://doi.org/10.1002/sim.6605

Sillanpää, M. (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity*, *106*, 511–519. https://doi.org/10.1038/hdy.2010.91

Somashekar, L., Selvaraju, S., Parthipan, S., & Ravindra, J. P. (2015). Profiling of sperm proteins and association of sperm PDC-109 with bull fertility. *Systems Biology in Reproductive Medicine*, *61*, 376–387. https://doi.org/10.3109/19396368.2015.1094837

de Souza Fonseca, P. A., dos Santos, F. C., Rosse, I. C., Ventura, R. V., Brunelli, F. Â. T., Penna, V. M., … Peixoto, M. G. C. D. (2016). Retelling the recent evolution of genetic diversity for Guzerá: Inferences from LD decay, runs of homozygosity and Ne over the generations. *Livestock Science*, *193*, 110–117. https://doi.org/10.1016/j.livsci.2016.10.006

Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nature Reviews Genetics*, *16*, 33–44.

Speizer, I. S., & Lance, P. (2015). Fertility desires, family planning use and pregnancy experience: Longitudinal examination of urban areas in three African countries. *BMC Pregnancy Childbirth*, *15*, 1.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, *22*, 1540–1542. https://doi.org/10.1093/bioinformatics/btl117

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, *2* (2), 125–141. https://doi.org/10.1016/0040-5809(71)90011-6

Tucker, G., Price, A. L., & Berger, B. (2014). Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics*, *197*, 1045–1049. https://doi.org/10.1534/genetics.114.164285

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Voight, B. F., & Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics*, *1*, e32. https://doi.org/10.1371/journal.pgen.0010032

Wang, K., Hu, X., & Peng, Y. (2013). An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. *Human Heredity*, *76*, 1–9. https://doi.org/10.1159/000353345

Wientjes, Y. C., Veerkamp, R. F., & Calus, M. P. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, *193*, 621–631. https://doi.org/10.1534/genetics.112.146290

Yee, W. K., Rogell, B., Lemos, B., & Dowling, D. K. (2015). Intergenomic interactions between mitochondrial and Y-linked genes shape male mating patterns and fertility in Drosophila melanogaster. *Evolution*, *69*, 2876–2890. https://doi.org/10.1111/evo.12788

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Fonseca PAS, Leal TP, Santos FC, et al. Reducing cryptic relatedness in genomic data sets via a central node exclusion algorithm. *Mol Ecol Resour*. 2017;00:1–13. https://doi.org/10.1111/1755-0998.12746