WILEY
Online Proofing System Instructions

The Wiley Online Proofing System allows authors and proof reviewers to review PDF proofs, mark corrections, respond to queries, upload replacement figures, and submit these changes directly from the PDF proof from the locally saved file or while viewing it in your web browser.

1. For the best experience reviewing your proof in the Wiley Online Proofing System please 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 the icon with a green check mark 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. To save your proof corrections, click the "Publish Comments" button appearing above in the yellow banner. Publishing your comments saves your corrections to the Wiley Online Proofing System server. Corrections don't have to be marked in one sitting, you can publish corrections and log back in at a later time to add more 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:

Click Here

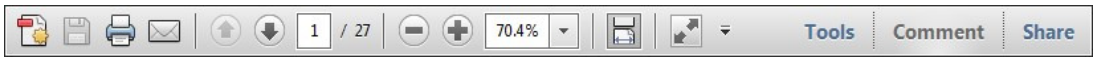
5. When your proof review is complete and you are ready to submit corrections to the publisher, please click the "Complete Proof Review" button below:

Click Here

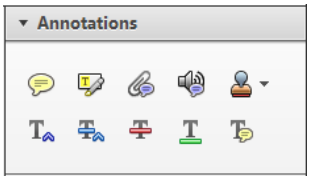
IMPORTANT: Do not click the "Complete Proof Review" button without replying to all author queries found on the last page of your proof. Incomplete proof reviews will cause a delay in publication.

IMPORTANT: Once you click "Complete Proof Review" you will not be able to publish further corrections.

Once you have Acrobat Reader open on your computer, click on the Comment tab at the right of the toolbar:



This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the Annotations section, pictured opposite. We've picked out some of these tools below:

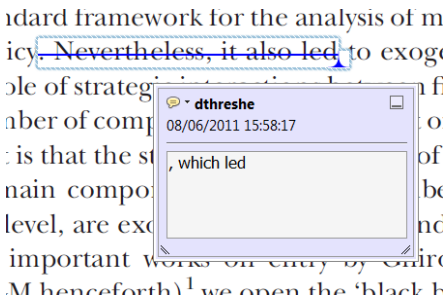


1. Replace (Ins) Tool – for replacing text.

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 the Replace (Ins) icon in the Annotations section.
- Type the replacement text into the blue box that appears.

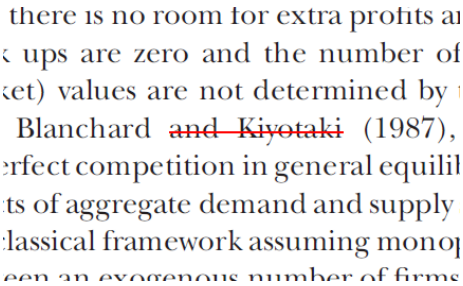


2. Strikethrough (Del) Tool – for deleting text.

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

How to use it

- Highlight a word or sentence.
- Click on the Strikethrough (Del) icon in the Annotations section.

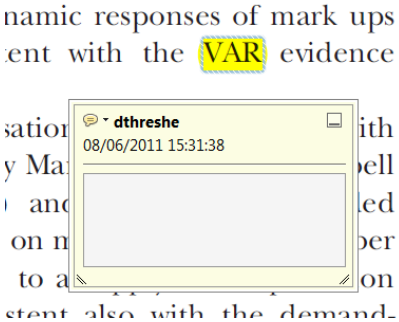


3. Add note to text Tool – for highlighting a section to be changed to bold or italic.

Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the Add note to text icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.

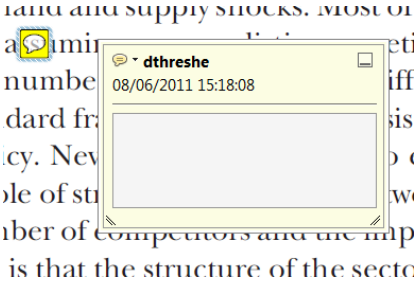


4. Add sticky note Tool – for making notes at specific points in the text.

Marks a point in the proof where a comment needs to be highlighted.

How to use it

- Click on the Add sticky note icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.

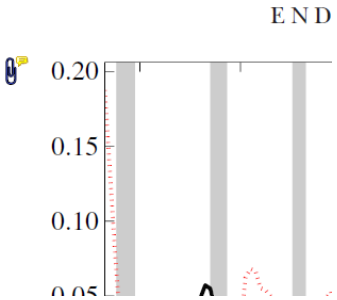


5. Attach File Tool – for inserting large amounts of text or replacement figures.

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

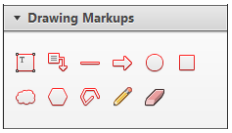
How to use it

- Click on the Attach File icon in the Annotations section.
- 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.



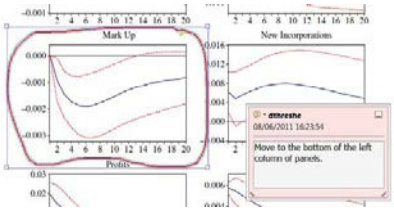
6. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

Allows shapes, lines and freeform annotations to be drawn on proofs and for comment 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, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.



Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12471

TESS3: fast inference of spatial population structure and genome scans for selection

KEVIN CAYE,* TIMO M. DEIST,* HELENA MARTINS,* OLIVIER MICHEL† and OLIVIER FRANÇOIS*
*Centre National de la Recherche Scientifique, Université Grenoble-Alpes, TIMC-IMAG UMR 5525, Grenoble 38042, France,
†Centre National de la Recherche Scientifique, Université Grenoble-Alpes, GIPSA-lab UMR 5216, Grenoble 38042, France

Abstract

Geography and landscape are important determinants of genetic variation in natural populations, and several ancestry estimation methods have been proposed to investigate population structure using genetic and geographic data simultaneously. Those approaches are often based on computer-intensive stochastic simulations and do not scale with the dimensions of the data sets generated by high-throughput sequencing technologies. There is a growing demand for faster algorithms able to analyse genomewide patterns of population genetic variation in their geographic context. In this study, we present TESS3, a major update of the spatial ancestry estimation program TESS. By combining matrix factorization and spatial statistical methods, TESS3 provides estimates of ancestry coefficients with accuracy comparable to TESS and with run-times much faster than the Bayesian version. In addition, the TESS3 program can be used to perform genome scans for selection, and separate adaptive from nonadaptive genetic variation using ancestral allele frequency differentiation tests. The main features of TESS3 are illustrated using simulated data and analysing genomic data from European lines of the plant species *Arabidopsis thaliana*.

Keywords: control of false discoveries, genome scans for selection, geographic variation, inference of population structure

Received 22 June 2015; revision received 4 September 2015; accepted 16 September 2015

Introduction

Since the early developments of population genetics, geography has been recognized as one of the major determinants of genetic variation in natural populations (Wright 1943; Malécot 1948; Kimura & Weiss 1964; Cavalli-Sforza *et al.* 1994; Epperson 2003). For these populations, spatial patterns of genetic variation can be influenced by landscape barriers, by geographical distances and by the processes of divergence and admixture resulting from the colonization of new areas. In addition, analysing spatial patterns of genetic variation has been a long-standing goal of evolutionary biogeography, molecular ecology, landscape genetics and conservation biology (Manel *et al.* 2010; Segelbacher *et al.* 2010).

Statistical approaches to analyse spatial patterns of genetic variation often rely on the inference of population genetic structure from multilocus genotype data, which is commonly performed using the Bayesian approach implemented in the computer program STRUCTURE (Pritchard *et al.* 2000). Assuming K unobserved ancestral gene pools, STRUCTURE computes allele frequencies in each pool and estimates individual *ancestry coefficients* representing the proportion of an individual genome that originates from each pool. Using STRUCTURE, ancestry coefficients are estimated without prior knowledge on geographic proximity among individuals.

The approach implemented in STRUCTURE has been substantially improved by a number of approaches that include spatial proximity information based on individual geographic coordinates [reviewed by François & Durand (2010)]. Among those spatially explicit approaches, the computer program TESS is one of the most frequently used algorithms (François *et al.* 2006; Chen *et al.* 2007). In the TESS model, ancestry proportions are continuously distributed over geographic space, and the parameters that specify the shape of the clines are estimated from the genetic and the geographic data. Using geographic information, TESS provides better estimates of ancestry coefficients than STRUCTURE when the levels of ancestral population divergence are low (Durand *et al.* 2009).

The Bayesian approaches implemented in STRUCTURE and TESS rely on Markov chain Monte Carlo algorithms. Monte Carlo algorithms are based on computer-intensive

Correspondence: Olivier François, Fax: +33-4-56-52-00-55;
E-mail: olivier.francois@imag.fr

MEN Journal Code | 12471 Manuscript No. | WILEY | Dispatch: 12.10.15 No. of pages: 9 | CE: Hemalatha PE: Iniya Selvi

stochastic simulations and have the advantage of sampling the posterior distribution of the model parameters. However, the application of stochastic algorithms can be difficult when the data include more than a few hundreds of individuals or a few thousands of allelic markers. With the availability of next-generation sequencing data, there is a need to analyse genotypic matrices that represent thousands of individuals and hundreds of thousands of markers. While fast versions of STRUCTURE have already been proposed (Alexander & Lange 2011; Frichot *et al.* 2014; Raj *et al.* 2014; Wollstein & Lao 2015), developing fast and accurate estimation algorithms for ancestry coefficients in a geographic framework remains an important computational challenge.

In this study, we present a spatially explicit algorithm that provides fast estimation of ancestry coefficients with accuracy comparable to TESS 2.3 (Durand *et al.* 2009). The new algorithms are based on least-squares optimization and on geographically constrained non-negative matrix factorization (Cai *et al.* 2011; Frichot *et al.* 2014). These improvements of TESS are implemented in the computer program TESS3. We show that TESS3 is substantially faster than TESS 2.3, with an increase in computational speed of one or two orders of magnitude. In addition, we show that ancestral allele frequencies are correctly estimated, and we illustrate the use of the TESS3 program to perform genome scans for selection based on ancestral allele frequency differentiation. To illustrate our approach, TESS3 was applied to genomic data from European lines of the model species *Arabidopsis thaliana* for which an individual-based sampling design was available (Atwell *et al.* 2010).

Materials and methods

The computer program TESS3 computes ancestry estimates for large genotypic matrices using the geographic coordinates of sampled individuals. The program also returns locus-specific estimates of ancestral genotypic frequencies and computes locus-specific estimates of a population-based differentiation statistic that can be used in genome scans for adaptive alleles. The TESS3 program is particularly suited to the analysis of large genomic data sets, for which the number of loci (L) ranges between thousands to hundreds of thousands genetic polymorphisms and the number of individuals (n) ranges between hundreds to thousands individuals.

Input data

TESS3 requires that the data consists of n multilocus genotypes and two geographic coordinates for each genotype. A genotypic matrix, X, records allelic data for each individual (i) and each locus (ℓ). With data representing single nucleotide polymorphisms (SNPs), the genotypic matrix records the number of derived or mutant alleles at each locus. Considering autosomes in a diploid organism, the genotype at locus ℓ corresponds to the number of derived alleles at this locus, which is encoded as an integer number 0, 1 or 2. For SNPs, the `geno` format is accepted by the program, which can also process other types of allelic data, such as short tandem repeats or amplified fragment length polymorphisms. Geographic coordinates can be expressed using several coordinate systems, for example longitude and latitude, and they are provided to the software in a separate input file.

Geographically constrained least-squares estimates of ancestry coefficients

Similar to TESS 2.3 or STRUCTURE, TESS3 supposes that the genetic data originate from the admixture of K ancestral populations, where K is unknown. TESS3 estimates a Q-matrix, $Q = (Q_{ik})$, which represents the individual ancestry coefficients ($n \times K$ dimensions), and a G-matrix, $G = (G_{k\ell}(j))$, which represents the ancestral genotypic frequencies. The dimension of G is equal to $K \times (p + 1)L$ where p is the ploidy of the studied organism genome. The ancestry coefficient Q_{ik} is the fraction of individual i's genome that originates from the ancestral population k, and the coefficient $G_{k\ell}(j)$ represents the frequency of genotype j at locus ℓ in population k.

The principle underlying the TESS3 algorithm differs from the likelihood methods implemented in STRUCTURE or in TESS 2.3, and it can be considered to be model-free. The main idea is that the probability that an individual i carries the genotype j at locus ℓ is determined by the *law of total probability*

$$P(X_{i\ell} = j) = \sum_{k=1}^{K} Q_{ik} G_{k\ell}(j).$$

The above formula establishes that each individual genotype is sampled from K pools of ancestral genotypes and that the sampling probabilities correspond to their admixture coefficients. The formula is equivalent to the factorization of the genotypic probability matrix, P, using the matrices Q and G as factors (Frichot *et al.* 2014). In the TESS3 algorithm, probabilities are replaced by zero/one values depending on the absence or the presence of each genotype at each locus, and the resulting matrix is denoted by $\tilde{X}$. Estimates of Q and G are obtained by factorizing $\tilde{X}$ as follows $\tilde{X} = \hat{Q}\hat{G}$. Matrix factorization is performed according to a least-squares minimization algorithm (see Appendix). During the minimization process, spatial constraints are introduced to ensure that individuals that are geographically close to each other are

more likely to share the same ancestral genotypes than individuals that are far apart. A regularization parameter, α, controls the regularity of ancestry estimates over the geographic space. Large values of α imply that ancestry coefficients have similar values for nearby individuals, whereas small values produce results close to STRUCTURE. The least-squares method leads to algorithms that are substantially faster than the Bayesian algorithms implemented in other programs. In addition, the approach makes no assumptions about linkage or Hardy–Weinberg equilibrium (HWE). The above framework is thus appropriate to deal with departures from HWE created by inbreeding or geographically restricted mating.

Number of populations

In TESS3, the number of ancestral populations, K, is chosen after the evaluation of a cross-entropy criterion for each K (Frichot *et al.* 2014). The choice of K is then based on a cross-validation method that partitions the genotypic matrix entries into a training set and a test set in which 5% of all entries are masked to the algorithm. The cross-entropy criterion compares the genotypic frequencies predicted from the training set to those computed from the test set at each locus. Smaller values of the criterion often indicate better estimates for TESS3. In practice, the best choice for K corresponds to a plateau in the cross-entropy plot (Frichot & François 2015).

Outlier locus tests

In addition to the inference of spatial population structure, TESS3 can perform genome scans for selection when the program is applied to large genomic data sets. More specifically, TESS3 uses the ancestral genotype frequency matrix, G, to derive the allele frequencies in the K ancestral populations. Then, the algorithm evaluates a locus-specific F_{ST}-statistic based on the estimated ancestral allele frequencies. Using standard population genetic theory, F_{ST}-statistics can be transformed into squared z-scores, and p-values can be computed using a chi-square distribution with $K-1$ degrees of freedom (Weir 1996). To correct for the test inflation statistic due to neutral population structure, the z-scores were recalibrated using estimates of the inflation factor. Here, inflation factors were determined using an 'empirical-null hypothesis' approach. The values of the inflation factor were determined graphically on the basis on quantile–quantile plots of P-values. This approach is less conservative than the method based on the median of the chi-square distribution with $K-1$ degrees of freedom (Devlin & Roeder 1999; Frichot & François 2015). Multiple testing issues were addressed by applying the Benjamini–Hochberg algorithm to the recalibrated p-values (Benjamini & Hochberg 1995).

Simulated data sets and program runs

We created simulated data sets containing 200 admixed genotypes with levels of ancestry that varied continuously across geographic space. To generate the data, we used the computer program MS to perform coalescent simulations of neutral and outlier SNPs under island models with two populations (Hudson 2002). One hundred genotypes were sampled from each source population, and admixed genotypes were created according to a longitudinal gradient of ancestry (Durand *et al.* 2009; François & Durand 2010). Individuals at each extreme of the longitudinal range were representative of ancestral populations, while individuals at the centre of the range shared intermediate levels of ancestry in the two source populations. The number of loci was varied in the range L = 1k-50k SNPs.

Our first series of simulations considered selectively neutral SNPs and used migration parameters, $M = 4mN_e$, between $M = 0.01$ and $M = 10$. The population differentiation statistic, F_{ST}, ranged from 0.007 to 0.42. Our second series of simulations included a proportion of outlier SNPs equal to 5%. Outlier loci were generated using two values of the effective migration rate $4m_sN = 0.1$ and $4m_sN = 1$. In simulations with outlier loci, the neutral migration rate was set to the value $4mN = 20$. The justification for using neutral migration–drift equilibrium models for simulating selection is that loci with selection have an effectively reduced migration rate, as compared to the neutral migration m in migration–selection–drift equilibrium models (Bazin *et al.* 2010).

The simulated data were used to compare TESS3 estimates to those of TESS 2.3 (Durand *et al.* 2009). The number of ancestral populations ranged from $K = 1$ to $K = 6$. Each run was replicated five times for each computer program. The number of cycles in the Markov chain Monte Carlo algorithm of TESS 2.3 was set to 1000, and the optimal number of ancestral population was determined using the deviance information criterion. All other parameters were set to their default values. Statistical errors were measured as root-mean-squared errors (RMSE) between the estimated Q-matrix and the matrix of coefficients (Q^0) that were used to generate the data

$$\mathrm{RMSE} = \left(\frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} (Q_{ik} - Q_{ik}^0)^2 \right)^{1/2}.$$

A similar RMSE criterion was defined for comparing the estimates of G matrices obtained from TESS3 or TESS

2.3 to the estimates of the ancestral genotypic frequency matrix resulting from the coalescent simulations.

Arabidopsis thaliana data

We applied TESS3 to genomic data from 170 European lines of the model plant *Arabidopsis thaliana* genotyped for 216k SNPs (Atwell *et al.* 2010). For these data, we determined the number of ancestral populations using the cross-entropy criterion, and we computed ancestry estimates for the sample. The results were projected onto a map of the European continent using a raster file and R graphic functions (Jay *et al.* 2012). We also used TESS3 to perform a genome scan for selection on chromosome 5 using $K = 3$ ancestral populations (54k SNPs).

Results

Comparison of ancestry estimates

We used computer simulations of admixed populations to evaluate the ability of TESS3 to reproduce the ancestry estimates of TESS 2.3 using known individual ancestry proportions from two ancestral gene pools. Simulating 2k unlinked SNPs, we varied the level of ancestral population differentiation, measured by F_{ST}, to create difficult as well as easier data sets. For all data sets, the information criterion of each version of TESS led to $K = 2$ clusters. Statistical errors, measured by RMSEs for estimated Q and G matrices, ranged between 0.02 and 0.15 (Fig. 1). Statistical errors increased as the levels of differentiation between the two source populations decreased, but they remained in an acceptable range for values of $F_{ST} > 0.016$. Overall, the statistical performances were of the same order for both versions of TESS.

Run-time analysis

Next, we compared the run-times of TESS3 and TESS 2.3 for increasing values of the number of ancestral populations and increasing numbers of loci. For TESS 2.3, the total number of cycles in the MCMC algorithm was set to 1000, a value for which the Monte Carlo sampler reached its equilibrium state. Run-times were averaged over distinct random seed values for each K and number of loci. For both algorithms, the run-times increased with the number of loci and with the number of ancestral populations (Fig. 2). For $L = 10$k loci, TESS3 and TESS 2.3 runs took <6 min on an Intel Xeon 2.40 GHz CPU. With $L = 50$k loci and $K = 5$ ancestral populations, TESS 2.3 took on average 30 min to complete a single run, whereas the TESS3 average run-time was about 4 min.

Outlier locus tests

We evaluated the capacity of TESS3 to detect outlier loci on simulated data containing 5% of outlier loci. For each locus, we performed a population differentiation test

COLOR

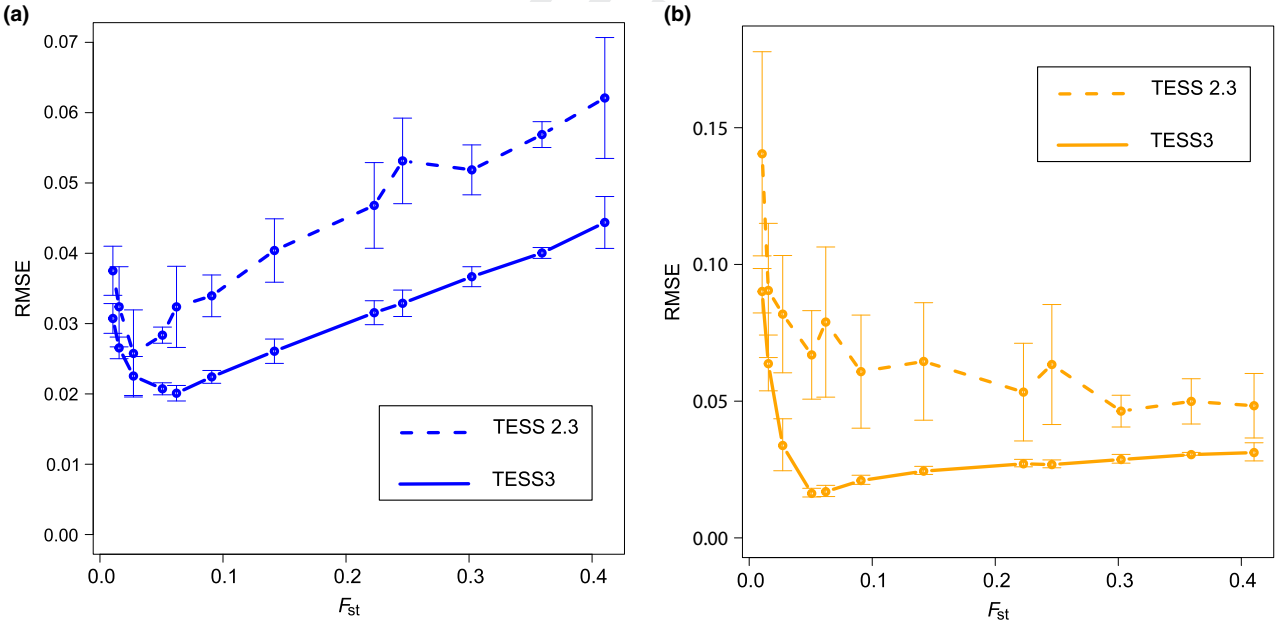


Fig. 1 Statistical errors of TESS3 and TESS 2.3 estimates. Computer simulations of admixed populations using known individual ancestry proportions from two ancestral gene pools. (a) RMSEs of G estimates as a function of the level of ancestral population differentiation (F_{ST}). (b) RMSEs of Q estimates as a function of the level of ancestral population differentiation (F_{ST}).

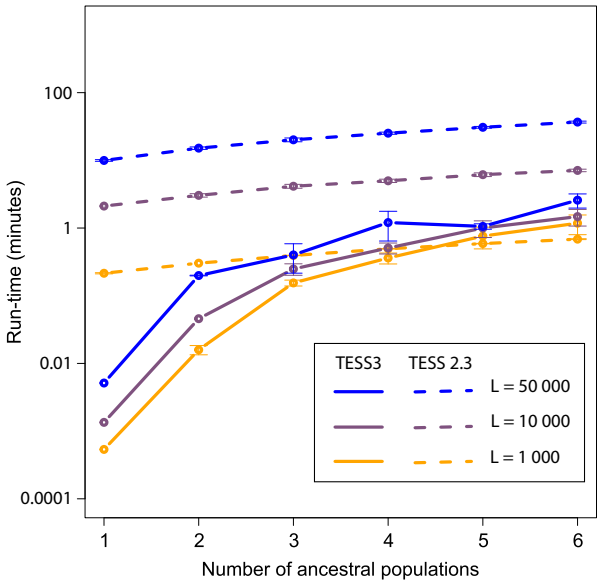


Fig. 2 Run-times for TESS3 and TESS 2.3. The number of ancestral population ranged between $K = 1$ and 6. Run-times were expressed in unit of minutes.

based on the estimated ancestral allele frequencies. Although the ratios m_s/m took large values, the probability distributions of F_{ST} statistics computed from neutral and selected ancestral allele frequencies overlapped substantially. Thus, the power of neutrality tests were expected to be low. For a data set with $m_s/m = 0.005$, the estimate of the genomic inflation factor was equal to $\lambda = 4.4$. For a data set with $m_s/m = 0.05$, this value was equal to $\lambda = 10.0$. After correction of the test statistic, the observed levels of the false discovery rate were close to their expected values. The power to reject the null hypothesis was lower when the intensity of selection was low (Table 1). For an expected FDR of $q = 0.1$, the power of the test was approximately equal to 60% for the higher selection rate and it was equal to 30% for the lower selection rate. The power values were close to those obtained when we applied outlier tests to the data before admixture. This experiment showed that the power to reject neutrality in continuous populations was similar to the power of traditional population differentiation tests applied to the discrete (ancestral) population data.

Biological data analysis

We applied TESS3 to a genomic data set of 170 European lines of *Arabidopsis thaliana* (216k SNPs). The cross-entropy curve exhibited a change in curvature for $K = 3$–4 clusters. For $K = 3$, the western cluster grouped all lines from the British Isles, France and Iberia. The eastern cluster grouped all lines from Central, Eastern Europe and

Table 1 Power to reject neutrality of TESS3 outlier tests for two simulated data sets

	Power			
FDR	After admixture	Before admixture	After admixture	Before admixture
0.05	**0.61**	0.63	**0.20**	0.26
0.10	**0.63**	0.66	**0.23**	0.29
0.15	**0.64**	0.67	**0.25**	0.32
0.20	**0.65**	0.69	**0.26**	0.33

Data set 1: $m_s/m = 0.005$. Data set 2: $m_s/m = 0.05$.

Southern Sweden. Fourteen northern Scandinavian accessions were grouped into a separate population (Fig. 3a). Those results were consistent with those obtained with TESS 2.3. The average run-time of TESS3 was about 5 min whereas each TESS 2.3 run took about 2 h. Then, we performed a genome scan for selection based on population differentiation in the three ancestral populations detected by TESS3. The genomic inflation factor was equal to $\lambda = 15.0$. The histogram of corrected *P*-values provided evidence that confounding errors were correctly removed (Fig. 4a). The Manhattan plot exhibited islands of strong differentiation around positions 8510, 6944, 6969 and 26 155 kb in the chromosome 5 (Fig. 3b). The top hits in the candidate list corresponded to genic SNPs. In particular, we discovered genes involved in defence response (*VSP1*), and in photoperiodism, flowering and root development (*WAV2*) (Mochizuki *et al.* 2005). The derived allele in the *VSP1* gene was present at high frequency in Eastern Europe and it was almost absent from Western Europe and Northern Scandinavia. The derived allele in the *WAV2* gene was present at high frequency in the Iberian peninsula and at low frequency in Eastern Europe and Northern Scandinavia (Fig. 4b).

Discussion

A fundamental objective of evolutionary biology is the evaluation of the distribution of genetic variation among populations in geographic space. During the last few years, high-throughput sequencing technologies have allowed population geneticists to make fast progress in this direction. The access to extensive data have opened the door to a deeper understanding of the spatial distribution of adaptive and nonadaptive genetic variation in model and nonmodel organisms (Manel *et al.* 2010). This transition from population genetics to population and ecological genomics is accompanied by a revolution of the principles and methods used to analyse the influence of landscape features on genetic variation. This revolution is made possible thanks to the availability of fast

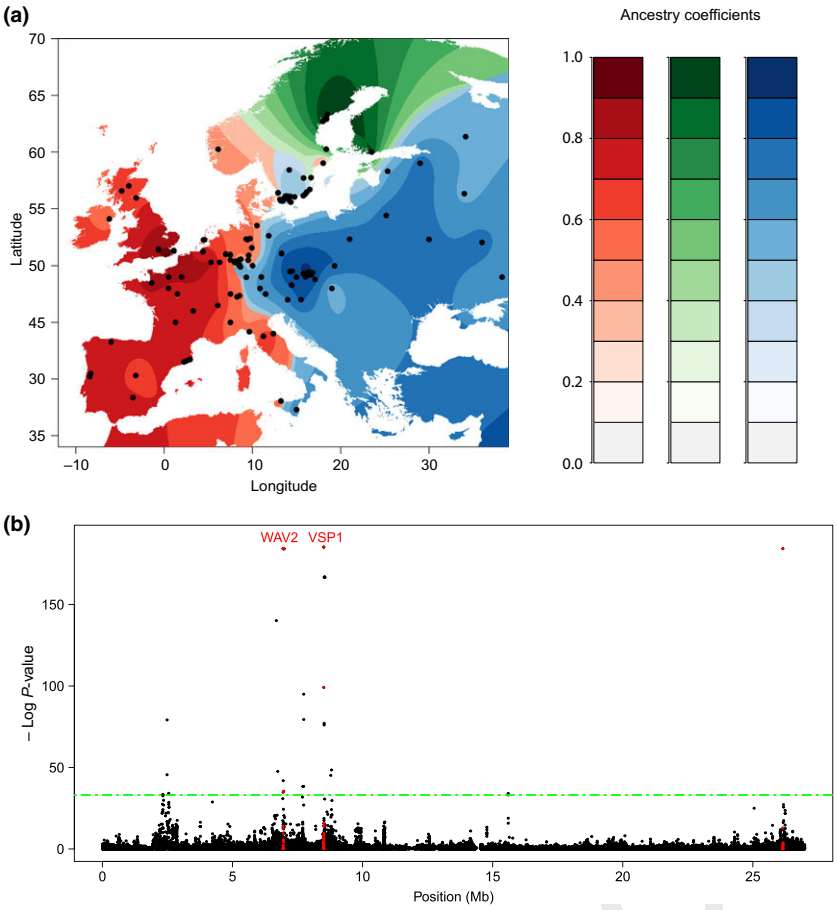


Fig. 3 Results of the *Arabidopsis thaliana* data analysis with TESS3. (a) Geographic maps of ancestry coefficients using $K = 3$ ancestral populations. (b) Manhattan plot of $\log_{10}(P\text{-values})$ for the plant chromosome 5. The horizontal line corresponds to an expected FDR value of $q = 10^{-30}$.

COLOR

computing programs than can deal with high dimension and heterogeneity in the data.

By combining matrix factorization and spatial statistical methods, the computer program TESS3 enabled fast analysis of geographic and genomewide patterns of genetic variation from large genomic data sets. In coalescent simulations of individuals with known ancestry, TESS3 produced accurate estimates of ancestry coefficients and ancestral allele frequencies. TESS3 results were statistically similar to those obtained with the Bayesian clustering program TESS 2.3, but TESS3 was about 30 times faster than TESS 2.3 when used with $K = 5$ ancestral populations and 50k binary loci. Although Bayesian approaches might be preferable for genotypic matrices of moderate dimensions, TESS3 generally outperformed TESS 2.3 when more than a few thousands of markers were used.

A novelty of TESS3 is the identification of outlier loci from the genotypic matrix. An important property of TESS3 outlier tests is that they do not require predefined populations, and that can be applied to individual sampling designs. Based on the estimations of the ancestral allele frequency matrix, the TESS3 algorithm computes a population differentiation statistic estimating a fixation index for each locus. If local adaptation favours a particular allele in some ancestral populations, the population differentiation statistic at that locus will be larger than at loci that are selectively neutral. Outliers in the distribution of the population differentiation statistic are usually considered as loci potentially targeted by local selection (Holsinger & Weir 2009). In addition, the program output allows population geneticists to determine candidate loci based on classical FDR control algorithms.

The study of European lines of *A. thaliana* illustrated the main steps of analysis using TESS3. These steps can be summarized as follows: (i) identifying the number of clusters using the cross-validation criterion, by launching multiple runs of the program for each value of *K*, (ii) displaying maps of ancestry coefficients using R scripts provided with the program and (iii) performing a genome scan for selection based on ancestral allele frequency differentiation statistics. Results for *A. thaliana* suggested that clinal variation occurs along an East–West gradient separating two ancestral populations in Central Europe. Those results were in very good agreement with previ-

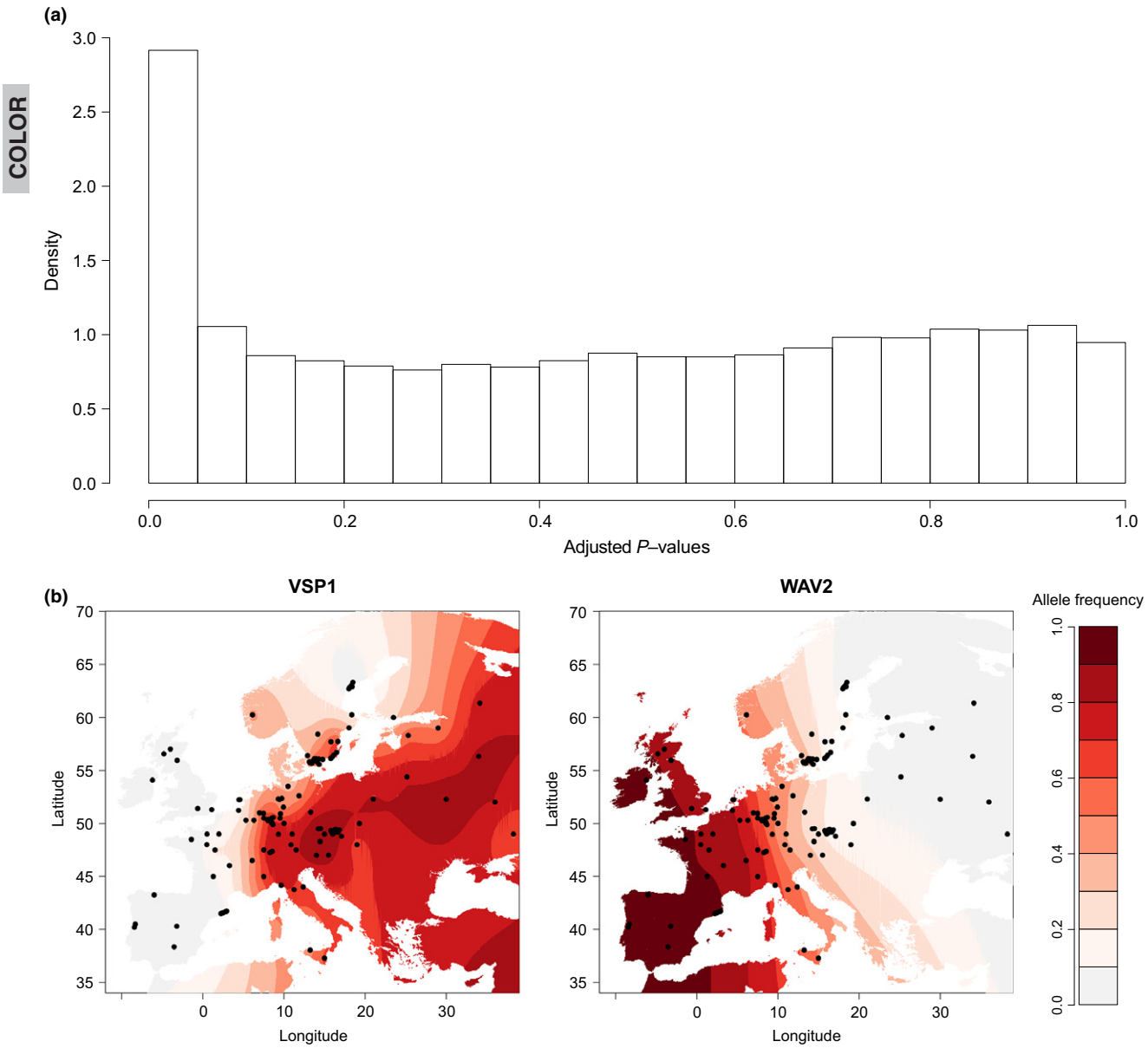


Fig. 4 Candidate SNPs from a genome scan of *A. thaliana* chromosome 5. (a) Histogram of adjusted *P*-values. (b) Spatial distribution of allele frequency for two top-hit SNPs located in the *VSP1* and *WAV2* genes.

ous findings using TESS 2.3, although these findings were obtained with a different set of markers (François *et al.* 2008). A genome scan for selection revealed contrasted patterns among European lines of *A. thaliana* and provided evidence of a substantial role for natural selection in shaping the genomewide variation of the plant species in Europe.

To conclude, the computer program TESS3 provides a major update of the TESS program enabling rapid ancestry coefficient estimation and genome scans for adaptive alleles. While preserving the accuracy of TESS 2.3, the least-squares algorithms of TESS3 ran vsubstantially faster than the Bayesian algorithms of TESS when analysing large population genomic data sets.

Acknowledgements

This work was supported by a grant from the Laboratoire d'Excellence Labex Persyval-lab to Kevin Caye. H. Martins acknowledges support from the 'Ciências sem Fronteiras' scholarship program from the Brazilian government. Olivier François acknowledges support from Grenoble INP and from the

'Agence Nationale de la Recherche' (project AFRICROP ANR-13-BSV7-0017).

References

Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, **12**, 246.

Atwell S, Huang YS, Vilhjálmsson BJ, *et al*. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Bazin E, Dawson KJ, Beaumont MA (2010) Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, **185**, 587–602.

Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, **15**, 1373–1396.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**, 289–300.

Cai D, He X, Han J, Huang TS (2011) Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 1548–1560.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) *The History and Geography of Human Genes*. Princeton University Press, ????, USA.

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.

Chung FR (1997) *Spectral Graph Theory, Vol. 92 of Regional Conference Series in Mathematics*. American Mathematical Society, ?????, USA.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Durand E, Jay F, Gaggiotti OE, François O (2009) Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, **26**, 1963–1973.

Epperson BK (2003) *Geographical Genetics (MPB-38)*. Princeton University Press, ?????, USA.

François O, Durand E (2010) Spatially explicit Bayesian clustering models in population genetics. *Molecular Ecology Resources*, **10**, 773–784.

François O, Ancelet S, Guillot G (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, **174**, 805–816.

François O, Blum MG, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genetics*, **4**, e1000075.

Frichot E, François O (2015) LEA: an R package for Landscape and Ecological Association studies. *Methods in Ecology and Evolution*, **6**, 925–929.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics*, **196**, 973–983.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, **10**, 639–650.

Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Jay F, Manel S, Alvarez N, *et al*. (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, **21**, 2354–2368.

Kim J, Park H (2011) Fast nonnegative matrix factorization: an active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, **33**, 3261–3281.

Kimura M, Weiss GH (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561.

Malécot G (1948) *Les Mathématiques de l'Hérédité*. Masson, Paris.

Manel S, Joost S, Epperson BK, *et al*. (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Mochizuki S, Harada A, Inada S, *et al*. (2005) The *Arabidopsis* WAVY GROWTH 2 protein modulates root bending in response to environmental stimuli. *The Plant Cell*, **17**, 537–547.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.

Segelbacher G, Cushman SA, Epperson BK, *et al*. (2010) Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation Genetics*, **11**, 375–385.

Weir (1996) *Genetic Data Analysis II*. Sinauer Associates Inc., Sunderland, Massachusetts.

Wollstein A, Lao O (2015) Detecting individual ancestry in the human genome. *Investigative Genetics*, **6**, 1–12.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114.

xxxxxxxxxxxxxx.

Data accessibility

Installing `TESS3`**.** Source codes, installation files and program documentation are available from Github (https://github.com/cayek/TESS3).

The Atwell *et al*. (2010) data used in this study are publicly available from the following link: https://github.com/Gregor-Mendel-Institute/atpolydb

Appendix

This section provides a detailed description of the `TESS3` algorithm. The first step of the algorithm builds a nearest-neighbour graph based on the geographic coordinates of the sampling sites. The number of neighbours in the graph was set to represent 5% of total connections. Then, the program runs a least-squares minimization algorithm. In this approach, the estimates of Q and G are obtained after solving the following constrained least-squares problem (Cai *et al*. 2011)

$$(\hat{Q}, \hat{G}) = \arg\min \mathrm{LS}(Q, G),$$

where

$$\mathrm{LS}(Q, G) = \|\tilde{X} - QG\|_{\mathrm{F}}^2 + \alpha \sum_{s_i \sim s_j} w_{ij} \|Q_{i\cdot} - Q_{j\cdot}\|^2, \quad \text{(eqn 1)}$$

and Q and G are non-negative matrices such that, for all i and ℓ, we have

$$\sum_{k=1}^{K} Q_{ik} = 1 \qquad \sum_{j=0}^{p} G_{i\ell}(j) = 1.$$

In this equation, $\|M\|_F$ denotes the Frobenius norm of a matrix M, $\| V \|$ is the Euclidean norm of a vector V, α is a non-negative *regularization parameter*. The summation on the right-hand side of the second term runs over all pairs of sites, $s_i \sim s_j$, sharing an edge in the nearest-neighbour graph. The quantity w_{ij} is a weight that decreases with geographic distance between sampling sites as follows

$$w_{ij} = \exp(-d(s_i, s_j)^2/\bar{d}^2), \qquad \text{(eqn 2)}$$

where d is the Euclidean distance, and $\bar{d}$ is the average distance computed over the neighbouring sites in the sample. More specifically, the weight of an edge in the nearest-neighbour graph is related to the Laplace–Beltrami operator on a manifold (Belkin & Niyogi 2003). In the algorithm, the regularization parameter α is equal to $c \times nL(p+1)/\sum w_{ij}$. The default value of c is 0.1%.

Least-squares minimization is performed using the alternating non-negativity-constrained least-squares (ANLS) algorithm with the active set (AS) method following the approach used in the computer program `sNMF` (Kim & Park 2011; Frichot *et al.* 2014). The ANLS-AS algorithm starts with the initialization of the Q matrix and then computes a non-negative matrix G that minimizes the quantity

$$\mathrm{LS}_1(G) = \|X - QG\|_F^2.$$

The obtained solution is normalized so that its entries satisfy the probabilistic constraints for genotypic frequencies. Given G, the Q-matrix is computed after minimizing the following quantity

$$\mathrm{LS}_2(Q) = \left\| \begin{pmatrix} \mathrm{Vec}(\tilde{X}^T) \\ 0 \end{pmatrix} - \begin{pmatrix} \mathrm{Id} \otimes G^T \\ \sqrt{\alpha}\,\Gamma \otimes \mathrm{Id} \end{pmatrix} \mathrm{Vec}(Q^T) \right\|_F^2,$$

where $\mathrm{Vec}(\tilde{X})$ denotes the vectorization of the matrix $\tilde{X}$ formed by stacking the columns of $\tilde{X}$ into a single column vector, Γ is the Cholesky decomposition of the graph Laplacian associated with the weights of the graph (Chung 1997), Id is the identity matrix, and $\otimes$ is a symbol for the Kronecker product. Iterations are stopped when the relative difference between two successive values of LS(Q,G) is lower than a tolerance threshold of ε. The default value for ε equals 10^{-7}.

Author Query Form

Journal: MEN

Article: 12471

Dear Author,
During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers on the query sheet if there is insufficient space on the page proofs. Please write clearly and follow the conventions shown on the attached corrections sheet. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.

Many thanks for your assistance.

Query reference	Query	Remarks
1	**AUTHOR: Please confirm that given names (red) and surnames/family names (green) have been identified correctly.**	
2	**AUTHOR: Belkin, Niyogi (2003) not cited. Please cite reference in text or delete from the list.**	
3	**AUTHOR: Please provide Publisher Location for reference Cavalli-Sforza et al. (1994).**	
4	**AUTHOR: Please provide Publisher Location for reference Chung (1997).**	
5	**AUTHOR: Please provide Publisher Location for reference Epperson (2003).**	
6	**AUTHOR: Kim, Park (2011) not cited. Please cite reference in text or delete from the list.**	
7	**AUTHOR: Please supply a short paragraph describing the specific contributions of the individual authors to the published work.**	
8	**AUTHOR: Please provide an appropriate table footnote to explain the bold values in Table 1.**	
9	**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://onlinelibrar-y.wiley.com/journal/10.1111/(ISSN)1755-0998/homepage/ForAuthors.html**	