The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences

Philippe Cubry, Yves Vigouroux, Olivier François

frontiers in Genetics

METHODS
published: 29 September 2017
doi: 10.3389/fgene.2017.00139

The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences

Philippe Cubry[1], *Yves Vigouroux*[1] *and Olivier François*[2]*

[1] UMR DIADE, University of Montpellier, Montpellier, France, [2] TIMC-IMAG UMR 5525, Centre National de la Recherche Scientifique (CNRS), Université Grenoble-Alpes, Grenoble, France

Rare variants are important for drawing inference about past demographic events in a species history. A singleton is a rare variant for which genetic variation is carried by a unique chromosome in a sample. How singletons are distributed across geographic space provides a local measure of genetic diversity that can be measured at the individual level. Here, we define the empirical distribution of singletons in a sample of chromosomes as the proportion of the total number of singletons that each chromosome carries, and we present a theoretical background for studying this distribution. Next, we use computer simulations to evaluate the potential for the empirical distribution of singletons to provide a description of genetic diversity across geographic space. In a Bayesian framework, we show that the empirical distribution of singletons leads to accurate estimates of the geographic origin of range expansions. We apply the Bayesian approach to estimating the origin of the cultivated plant species *Pennisetum glaucum [L.] R. Br.* (pearl millet) in Africa, and find support for range expansion having started from Northern Mali. Overall, we report that the empirical distribution of singletons is a useful measure to analyze results of sequencing projects based on large scale sampling of individuals across geographic space.

Keywords: genetic diversity, singletons, geographic origin, range expansion, pearl millet

OPEN ACCESS

Edited by:
Samuel A. Cushman, United States Forest Service Rocky Mountain Research Station, United States

Reviewed by:
Pablo Orozco-terWengel, Cardiff University, United Kingdom
Ricardo T. Pereyra, University of Gothenburg, Sweden
Rita Rasteiro, University of Bristol, United Kingdom

***Correspondence:**
Olivier François
olivier.francois@univ-grenoble-alpes.fr

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *27 March 2017*
Accepted: *14 September 2017*
Published: *29 September 2017*

Citation:
Cubry P, Vigouroux Y and François O (2017) The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences. Front. Genet. 8:139. doi: 10.3389/fgene.2017.00139

1. INTRODUCTION

High-throughput sequencing technologies have enabled studies of genomic diversity in model and non-model species at a dramatically increasing rate. Conducted at population and at individual levels, those studies have provided comprehensive surveys of common and rare variation in model species genomes (Weigel and Mott, 2009; 1000 Genomes Project Consortium et al., 2010; International HapMap 3 Consortium, 2010; 1000 Genomes Project Consortium, 2015). For example, the 1000 Genomes Project Consortium (2015) reported that the majority of variants in human genomes are rare. During the last decade, the role that rare variants play in shaping complex traits has been hotly debated (Pritchard, 2001; Schork et al., 2009; Tennessen et al., 2012), and accurately determining their distribution has become important for medical applications and association studies (Lee et al., 2014; Auer and Lettre, 2015). Beyond humans, rare variation has attracted considerable interest from genome sequencing projects for model organisms, including plants (Zhu et al., 2011; Weigel, 2012; Memon et al., 2016).

Rare variants are also important for drawing inference about past demographic events in a species history (Schraiber and Akey, 2015). Studies of human populations have shown that our

species has experienced a complex demographic history, and that a recent period of explosive growth has resulted in an excess of those variants (Coventry et al., 2010; Keinan and Clark, 2012). The analysis of private and rare variation has been used to reveal signals of differential demographic history among populations, and to refine models of human evolution (Marth et al., 2004; Gravel et al., 2011; Mathieson and McVean, 2014). In addition, estimating rare allele frequencies has enabled estimates of gene flow between populations, and has facilitated inference of fine-scale population structure (Slatkin, 1985; Novembre and Slatkin, 2009; O'Connor et al., 2015).

In this study, we define the empirical distribution of singletons in a sample of chromosomes as the proportion of the total number of singletons that each chromosome carries, where a singleton is a uniquely represented allele in the sample (Fu and Li, 1993). We provide theoretical and empirical analyses of the distribution of singletons in a sample of chromosomes, and we evaluate the potential for this distribution to provide an accurate description of genetic diversity at the individual level. Using spatial data, we use the distribution of singletons as an individual-based estimate of genetic diversity in geographic space.

The theoretical background for the analysis of the empirical distribution of singletons rely on the distribution of external branch lengths for coalescent genealogies (Blum and François, 2005; Caliebe et al., 2007). First, we use coalescent and spatially explicit simulations to evaluate individual contributions to genetic diversity in the sample based on singletons. Then we evaluate the use of the distribution of singletons in an approximate Bayesian Computation (ABC) framework to estimate the geographic origin of range expansions (Beaumont, 2010; Csilléry et al., 2010). We eventually provide an illustration of our theory by applying the ABC approach to the plant species *Pennisetum glaucum [L.] R. Br.* (pearl millet). Pearl millet is a cereal cultivated in semi-arid regions of Africa and the Indian subcontinent, and it is known to originate in Africa (Clotault et al., 2012). We evaluate the geographic origin of its range expansion by using 146 inbred lines from the whole African range.

2. THEORY

We consider a sample of n chromosomes from a population of N haploid organisms. We assume that there are L polymorphic loci, and that for each locus, 0 represents the ancestral or reference allele and 1 is the derived allele. A singleton is defined as a derived allele carried by a single chromosome in the sample. The total number of singletons, ξ_1, is the number of uniquely represented derived alleles in the sample, and it corresponds to the first component of the site frequency spectrum. We assume that the singletons are distributed over the n chromosomes in the sample. More specifically, the number of singletons decomposes as follows

$$\xi_1 = \sum_{i=1}^{n} \xi_1^{(i)},$$

where $\xi_1^{(i)}$ is the number of singletons carried by chromosome i. For each i, we denote by p_i the conditional probability that a singleton is carried by i. The n values $p_1, \ldots, p_n$ sum up to one, and those values define the *empirical distribution of singletons* in the sample (see below).

Next, we assume that the sample genealogies can be described by coalescent trees (Tavaré, 2004). For a particular locus, a tree is described by n tips and $n-1$ ancestral nodes. An external branch of the tree connects a tip to an ancestral node. For a given tree, we denote by $\tau^{(i)}$ the length of the external branch connecting chromosome i to its first ancestor node. The L coalescent trees exhibit complex patterns of statistical dependency along the chromosomes due to recombination among loci (Hudson, 1990). Measuring lengths in units of twice the total population size (N), and assuming a molecular clock model for mutations, the number of mutations falling on a particular branch of the tree has a Poisson distribution of rate $\theta/2$, where $\theta = 2\mu N$ and μ is the per generation mutation rate (Tavaré, 2004). Let ℓ be an arbitrary singleton locus. For all i, we write

$$\xi_1^{(i)} = \sum_{k=1}^{\xi_1} X_{i\ell},$$

where $X_{i\ell} = 1$ if singleton ℓ is carried by chromosome i, 0 otherwise. In the above formula, the summation runs over all singletons in the sample. Using mathematical properties of conditional distributions for the Poisson process, we have

$$p_i = \mathrm{P}(X_{i\ell} = 1) = \mathrm{E}\left[\frac{\tau_1^{(i)}}{\tau_1}\right],$$

where $\tau_1 = \sum_{i=1}^{n} \tau^{(i)}$. In this formula, the conditional probability that chromosome i carries a singleton at locus ℓ is given by the ratio of its external branch length to the total length of external branches in the sample genealogy at this locus. The distribution of singletons can be estimated by counting the number of singletons carried by each chromosome and normalizing as follows

$$\hat{p}_i = \xi_1^{(i)}/\xi_1, \quad i = 1, \ldots, n,$$

and the estimate is unbiased

$$\mathrm{E}\left[\hat{p}_i\right] = p_i\,.$$

In addition, the number of singletons carried by chromosome i, $\xi_1^{(i)}$, estimates the proportion of genetic diversity carried by chromosome i

$$\mathrm{E}[\xi_1^{(i)}] \approx \theta p_i, \quad i = 1, \ldots, n.$$

As a consequence of the theory presented in this section, the individual-based estimates of genetic diversity are unbiased

quantities regardless of demographic history, deviations from Hardy-Weinberg equilibrium and linkage disequilibrium. Limitations of the theory include the presence of closely related individuals, which should be removed from the sample prior to analysis. The approach is appropriate for modern sequencing data as soon as a few hundreds of DNA sequences are generated.

The rest of this study will evaluate the use of the empirical distribution of singletons in mapping genetic diversity in geographic space. To provide an elementary example, let us consider a sample of n chromosomes from a random mating population of size N. Using mathematical results for the neutral coalescent in a random mating population, the expected value of the number of singletons is an unbiased estimator of the genetic diversity in the sample (Fu and Li, 1993)

$$\mathrm{E}[\xi_1] = \theta.$$

For the lengths of external branch lengths, we have

$$\mathrm{E}[\tau^{(i)}] = 2/n, \quad i = 1, \ldots, n,$$

and $\mathrm{E}[\tau_1] = 2$ (Blum and François, 2005). Here, we expect that each chromosome contributes to genetic diversity equally. The above calculations show that, in a sample of size n from a random mating population, the distribution of singletons is uniform over the n chromosomes

$$p_i = 1/n, \quad i = 1, \ldots, n,$$

and we have

$$\mathrm{E}[\xi_1^{(i)}] = \mathrm{E}[\xi_1]\mathrm{P}(X_{ik} = 1) = \theta/n.$$

In other words, each individual contributes the same amount of genetic variation to the total sample diversity.

3. SIMULATION METHODS AND DATA SETS

3.1. Coalescent Simulations of Splitting Populations

We used the computer program *ms* to perform coalescent simulations for a two-population model (Hudson, 2002). In our simulations, we considered a population split model, in which two populations of sizes $N_1 = 50,000$ and $N_2 = sN_1$ ($s \in (0.01; 0.5)$, shrink rate) diverged t generations ago ($t \in (1,000; 10,000)$, split time). Population 1 expanded from an ancestral population of size $N_A = 5,000$, and the expansion started 10,000 generations ago. Samples of size $n = 100$ were considered and subdivided into subsamples of size 50 from each population. We simulated $L = 1,000$ unlinked haplotypes using the infinite-site model and an effective mutation rate $\theta \in (5; 10)$. The *ms* command line was written as follows: ./ms 100 1,000 -t theta -I 2 50 50 -g 1 46.05 -n 2 shrink.rate -eg 0.2 1 0.0 -ej split.time 2 1. The simulated data sets were processed by using the ".geno" format in the `R` package LEA (Frichot and François, 2015). We summarized the distribution of singletons by computing mean values and standard errors for each subsample. For all simulated samples, we used the `R` package *ape* to extract the coalescent trees generated by *ms*, and analyze the distribution of their external branch lengths (Paradis et al., 2004). We used the external branch length distribution to build a theoretical prediction for the distribution of singletons from each tree (see section 2), and summarized the theoretical distributions by computing mean values and standard errors for each subsample. The L coalescent simulations were replicated 200 times.

3.2. Range Expansions in Africa

Simulations of range expansions were performed by using the computer program SPLATCHE2 based on an array of 87 by 83 demes modeling the African continent (Currat et al., 2004). The demographic scenarios corresponded to range expansions from a single origin, simulated for a total duration of 1,600 generations. For each deme, the migration rate was equal to $m = 0.07$, and the growth rate was equal to $r = 0.1$. Additional parameters included an ancestral effective population size of 200 individuals, 200 generations before onset of expansion, and an effective mutation rate of 10^{-5} per base pair per generation.

Four types of demographic scenarios were considered. Two scenarios considered a "homogeneous" environment, for which the deme carrying capacities were set to a constant value $C = 100$ everywhere in Africa. Two other scenarios considered a heterogeneous environment linked to vegetation. In tropical semi-desert areas, the carrying capacities were set to $C = 60$, and in tropical extreme deserts and rain forests, the carrying capacities were set to $C = 30$. Demographic histories also differed by their geographic source of expansion. Range expansions were started either from an origin in West Africa (Mali, -4° E, 13° N) or from an origin in the Sahel area (Chad, 22° E, 20° N).

Ten haploid chromosomes were simulated for 30 population samples through the geographic range considered (300 chromosomes). Genetic variation was surveyed at 30,000 loci, and filtered out for monomorphic loci. From the resulting data sets, we computed the empirical distribution of singletons in each population sample, and compared this measure to expected heterozygosity for each population sample. Data files for running the SPLATCHE2 simulations are provided in Supplementary File 1. We reproduced the four scenarios by using individual sampling instead of population sampling. Here, individual genotypes were recorded at 300 distinct geographic sites, each obtained from a Gaussian perturbation of population centers with standard error of 2°. The Kriging method was used to interpolate the values of the expected heterozygosity and the empirical distribution of singletons on a geographic map of Africa (Cressie, 2015).

3.3. Pearl Millet Data

Whole genome sequencing data were obtained for 146 cultivated accessions of pearl millet (*Pennisetum glaucum [L.] R. Br.*) from the species range in Africa (International Pearl Millet Genome Sequencing Consortium, Varshney et al., 2017). A total of 169,095 SNPs were sampled after filtering out low quality

variants, and were used to estimate the distribution of singletons (Supplementary Material 1).

3.4. Approximate Bayesian Computation

We used Approximate Bayesian Computation (ABC) to evaluate the ability of the distribution of singletons to correctly estimate the onset of expansion in a range expanding species, and to estimate a posterior distribution for the location of this origin for cultivated pearl millet. We performed 20,000 range expansion simulations by considering a heterogeneous environment using the computer program SPLATCHE2. The deme carrying capacities were equal to $C = 100$ for tropical semi-desert areas, $C = 20$ for tropical extreme deserts and $C = 10$ for rain forests. Additional parameters included an ancestral effective population size of 200 individuals, 200 generations before onset of expansion, and an effective mutation rate of 10^{-5} per base pair per generation.

Prior distributions allowed the geographic coordinates of the origin of expansion to vary over the Sahel region. Longitude ranged between −16°E and 40°E, and latitude ranged between 5°N and 30°N. Lower prior probabilities were given to extreme latitudes and longitudes as a consequence of unsuitable habitats (water regions). Uninformative prior distributions were considered for the migration rate, the growth rate, the total duration of the demographic phase, the ancestral population size and the time before onset of expansion (Supplementary Table 1). In simulations, genetic variation was surveyed at 146 geographic sites corresponding to the exact sampling locations of pearl millet accessions. Ten thousands SNPs were simulated for each genotype. When evaluating summary statistics, a fraction of SNPs were removed from the simulated data in order to match with the amount of missing values observed in the original data set.

To define the summary statistics for ABC, we used a histogram for the distribution of singletons in the sample. The 146 accessions were grouped into spatial clusters according to a *k*-means algorithm and individual geographic information (Hartigan and Wong, 1979). The *k*-means algorithm resulted in 14 groups with more than 6 accessions in each group (**Figure 1**). To obtain a histogram, we computed the mean number of singletons in each group, and divided this value by the total number of singletons in the sample (Supplementary Table 2). Then ABC analysis was performed with the `R` package `abc` (Blum and François, 2010; Csilléry et al., 2012). Neural network models were used to estimate posterior distributions for the latitude and longitude of the geographic onset of expansion whereas the other parameters were considered as nuisance parameters without any interpretable unit. The tolerance rate was set to 0.05 and 250 neural networks were used in the `abc` function.

We first tested the accuracy of our estimates by using simulated data sets as inputs to the inference method. The sampling procedure and the ABC estimation were replicated 100 times, and we evaluated the correlation between coordinates of true origins and their estimated values. Then we considered the pearl millet data, and represented the prior and posterior densities of the geographic onset parameters by using

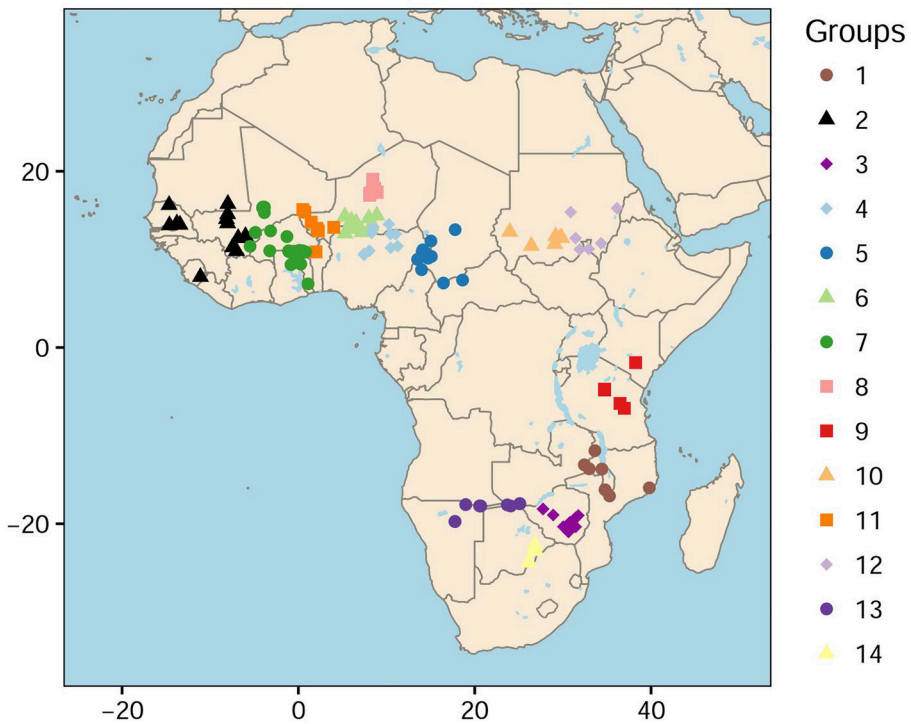

FIGURE 1 | Geographic distribution of 146 cultivated accessions of pearl millet. Fourteen geographic classes were defined as a result of a *k*-means procedure.

two-dimensional kernel density estimation with 100 grid points in each direction.

4. RESULTS

4.1. Coalescent Simulations of Splitting Populations

To evaluate statistical bias in the estimation of the distribution of singletons, we performed coalescent simulations of samples from two populations with unequal genetic diversity. The two populations diverged from an ancestral population t generations ago (*split time*), and at split time, the size of population 2 shrinked to s times the size of population 1 (*shrink rate*).

For each simulation, the number of polymorphic loci ranged between 7,883 and 39,761 (average value: 25,265 loci). For a value of the shrink rate $s \approx 1/3$, the average proportion of singletons in population 1 was about $\pi_1 = 0.0122$, and the average proportion of singletons in population 2 was about $\pi_2 = 0.0078$ ($\pi_1 + \pi_2 = 2/n$). This result reflected that genetic diversity in population 1 was higher than in population 2. The ratio was about $\pi_1/\pi_2 = 1.55$ (**Figure 2A**). The individual proportions were concentrated around their mean values with relatively small standard deviations ($SD_1 = 0.0010$, $SD_2 = 0.0008$).

The results from 200 replicates provided clear evidence that the empirical distribution of singletons is an unbiased estimate of its theoretical distribution based on coalescent trees (**Figure 2B**). The split time parameter had a weak influence on the distribution of singletons (Pearson correlation test, $P = 0.64$). The ratio π_1/π_2 reached values between 10 and 40 when the shrink rate was below 10%, and this parameter had a strong influence on the empirical distribution of singletons (**Figure S1**).

4.2. Range Expansions in Africa

For data sets generated under range expansion scenarios, the number of polymorphic loci ranged between 25,453 and 29,321 loci. The number of singletons ranged between 8,835 and 12,653, and the site frequency spectrum showed an excess of rare alleles as expected under explosive population growth. When the onset of expansion was set in Western Africa (cross in **Figure 3**), the maps of the empirical distribution of singletons and expected heterozygosity exhibited similar large-scale geographic patterns (**Figure 3**, Pearson's correlation coefficient 0.78). Because the computation of expected heterozygosities

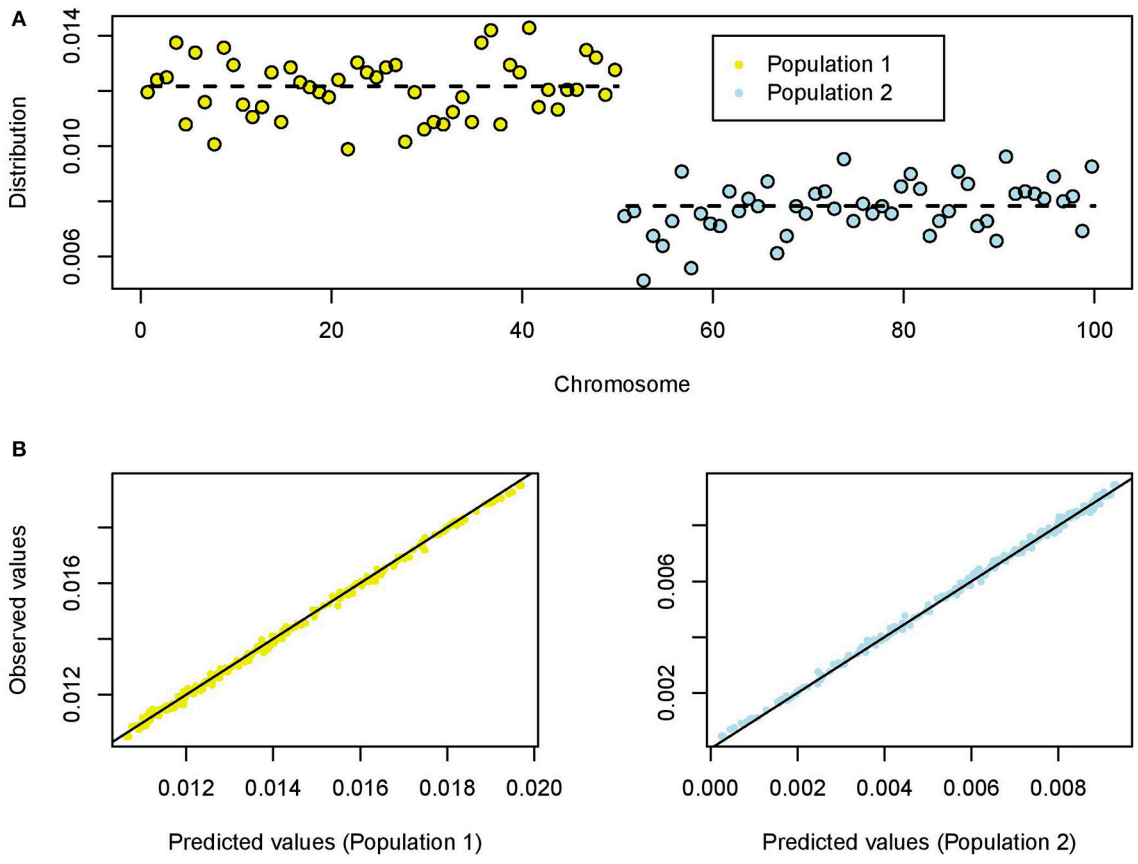

FIGURE 2 | Coalescent simulations of two splitting populations (100 chromosomes). **(A)** Empirical distribution of singletons for a value of the shrink rate $s = 0.33$. The dashed lines represent the averaged values for population 1 (expanding) and population 2 (skrinking). **(B)** Predicted and observed (empirical) values of the distribution of singletons for population 1 (left) and population 2 (right).

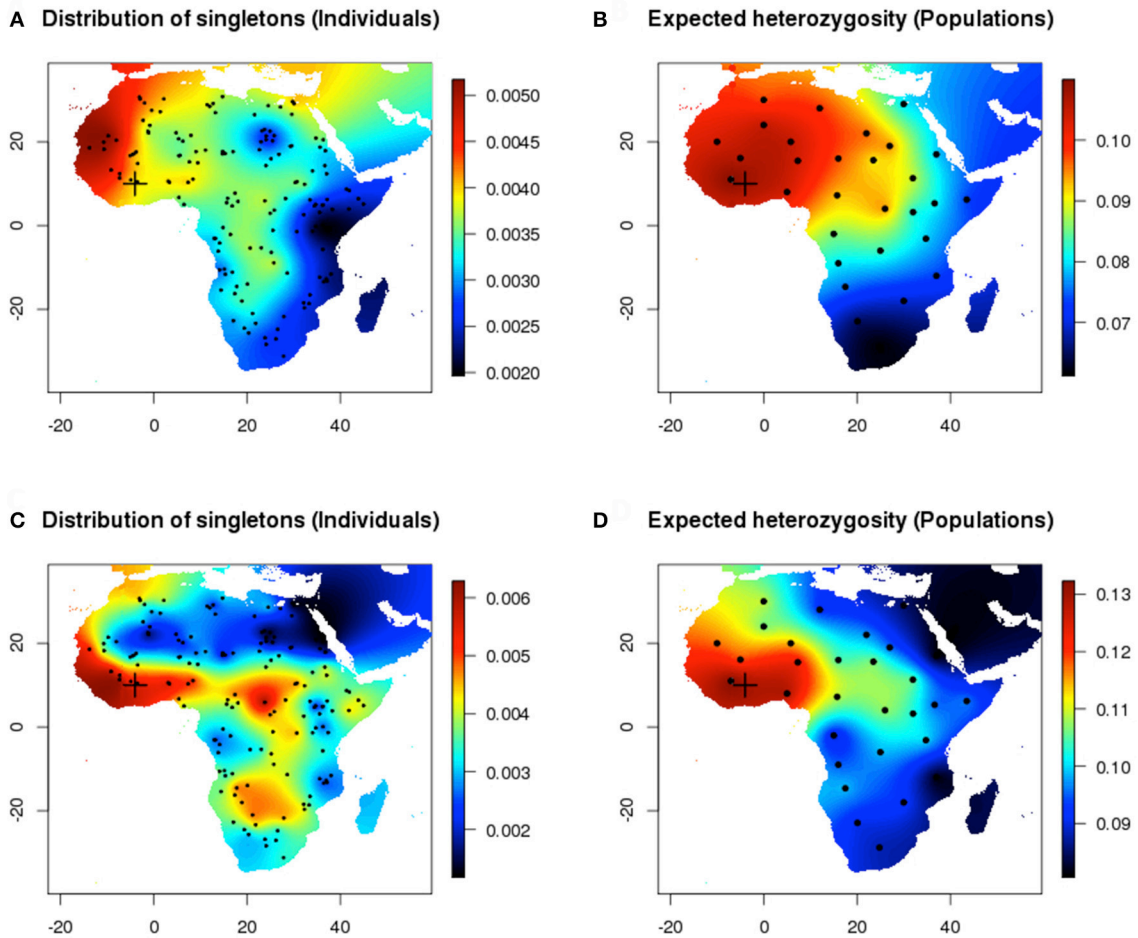

FIGURE 3 | Individual vs. population sampling after a range expansion simulation scenario (Western origin). **(A,B)** Homogeneous environment. Maps of the empirical distribution of singletons (individual sampling) and expected heterozygosity (true population sampling). **(C,D)** Inhomogeneous environment.

was based on a perfect assignment of samples to their true populations of origin, the interpolated maps corresponding to this measure (**Figures 3B,D**) contained less uncertainty than the maps of singletons (**Figures 3A,C**) that were based on random individual sampling. Considering environmental heterogeneity increased the variability of spatial estimates (**Figures 3C,D**).

Next, we compared estimates of heterozygosity for populations to the distribution of singletons in the same populations (**Figure 4**). Differences between maps produced with the empirical distribution of singletons and with expected heterozygosity decreased when the sampled chromosomes were perfectly assigned to their population of origin. The individual and population-based measures provided concordant estimates of genetic diversity in geographic space (Pearson's correlation coefficient 0.51). Similar results were observed when the onset of expansion was set in the Sahel area (20° E, 22° N) and were reported in **Figures S2**, **S3**.

4.3. Estimates of Expansion Onsets and Application to Pearl Millet

First, we used the distribution of singletons in ABC to infer origins of range expansion in 100 simulated data sets (**Figure 5**). The results provided evidence of the usefulness of the statistics to identify origins of range expansions. Estimated values for the longitude and latitude of the onset of expansion were highly correlated to the true values for these parameters. Pearson's squared correlation coefficients were equal to $R^2 = 0.950$ for the longitude and $R^2 = 0.948$ for the latitude (p-values < 0.01).

Next, we used the ABC approach to provide insights on the origin of range expansion of cultivated pearl millet in Africa. A total number of 41,032 singletons were found for 146 individuals, representing 24.27% of all variants. The posterior density for the longitude exhibited a mode around −7.52°E (CI:-11.26°E, 0.84°E) (**Figure 6**). For the latitude of origin, the posterior density exhibited a mode around 24.2°N and a large credible interval (CI: 11.03°N , 29.06°N) (**Figure 6**). The most probable

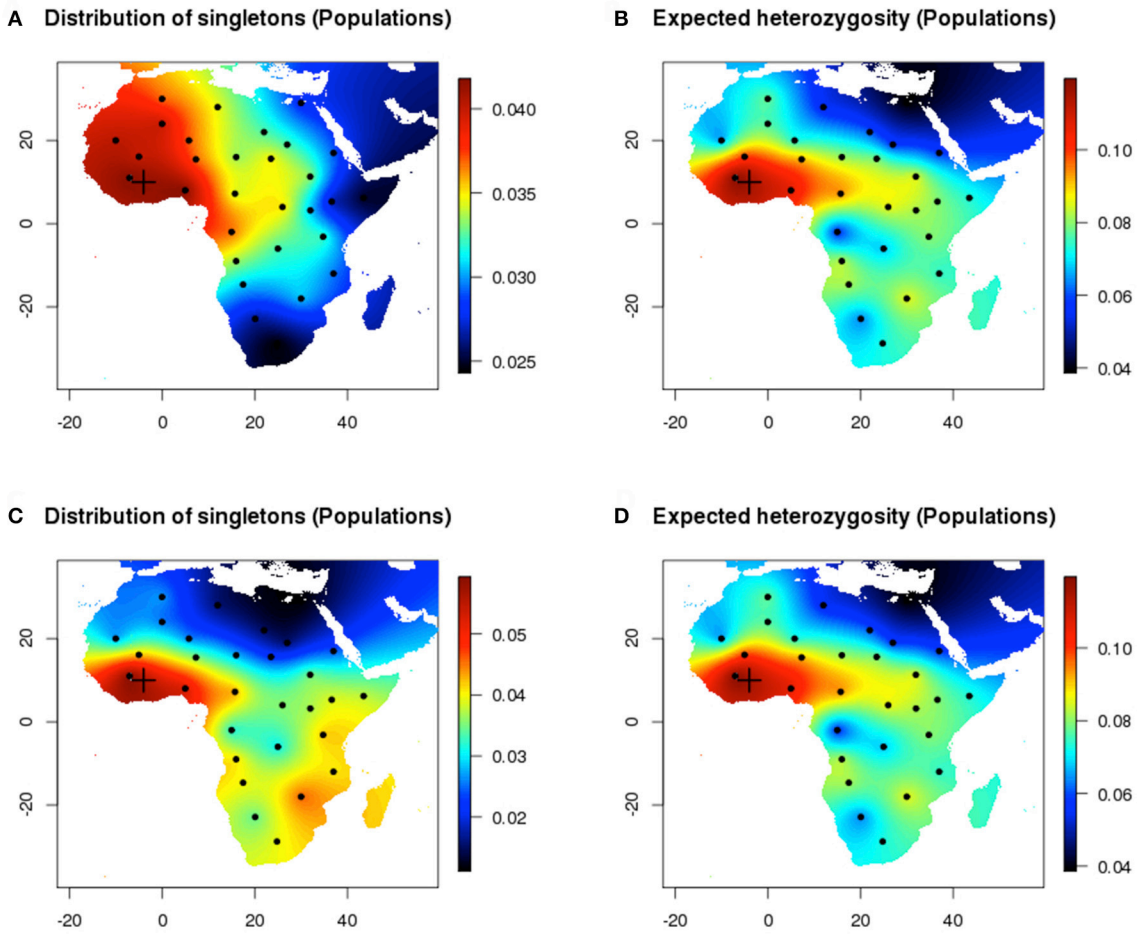

FIGURE 4 | Population sampling after a range expansion simulation scenario (Western origin). **(A,B)** Homogeneous environment. Maps of the empirical distribution of singletons (true population sampling) and expected heterozygosity (true population sampling). **(C,D)** Inhomogeneous environment.

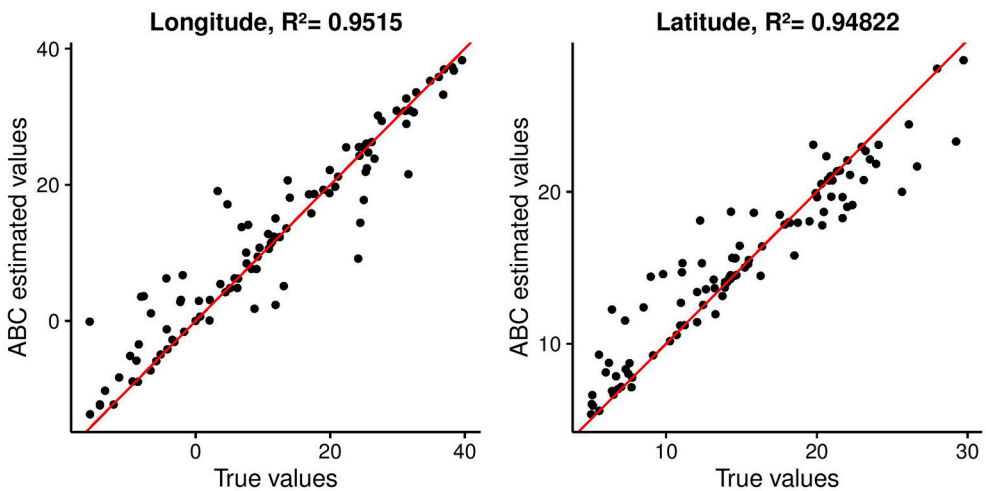

FIGURE 5 | Estimated coordinates of origin against their true values for 100 simulated data sets used as targets for ABC analysis. Pearson's correlation coefficients are reported.

location for the origin of expansion of pearl millet in Africa was found near the Mali-Mauritania border (**Figure 7**).

5. DISCUSSION

How singletons are distributed across geographic space provides a local measure of genetic diversity that can be measured at the individual level. In this study, we developed a theoretical background for the empirical distribution of singletons in a sample of chromosomes. We used simulations to provide evidence that the empirical distribution of singletons measures individual contributions to genetic diversity in the sample. The main advantage of this approach is to provide individual-based (local) estimates of genetic diversity that do not require the definition of populations.

Incorporated in an ABC framework, the empirical distribution of singletons led to accurate estimates of the geographic origin of range expansions in simulations. In ABC, the distribution of singletons was estimated by histograms obtained from clustering algorithms, and the histograms were used as summary statistics for Bayesian inference. Those statistics are appropriate to analyze the results of sequencing projects based on large scale sampling of individuals across geographic space. The method can be viewed as an interesting alternative to phylogenetic approaches when genomic sequences are used.

Potential factors that could bias our estimates of local genetic diversity includes missing data, genotyping errors, related individuals, and the use of a folded site frequency spectrum. Missing values or genotyping errors impacts individual data regardless of geography. By sharing genomic variation locally, related individuals reduce the number of unique variants drastically, and generate bias in global estimates of genetic diversity. Though those errors increase uncertainty in estimates, the biases on geographic estimates remain at small levels. Our ABC analysis took the potential biases into account by simulating the missing data, genotyping errors and the other issues. Alternative methods that could remove the biases would be based on genotype imputation and on the availability of genomic data from a closely related species.

We provided an illustration of the potential of singletons to inform demographic history by studying range expansion of pearl millet in Africa. Pearl millet is a widely grown staple

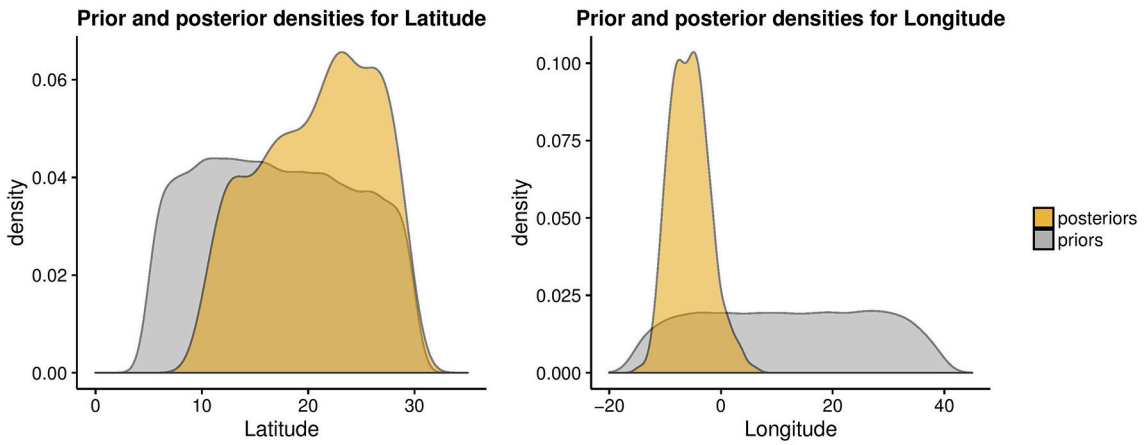

FIGURE 6 | Prior and posterior density estimates for the longitude and latitude of the expansion onset for cultivated pearl millet in Africa.

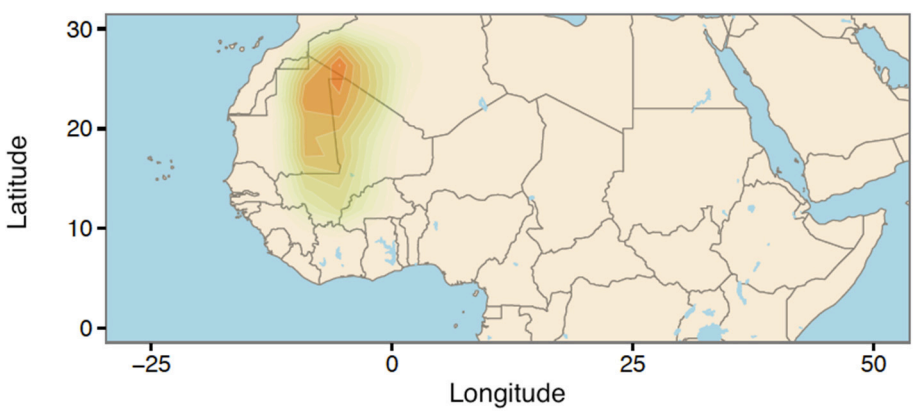

FIGURE 7 | Geographic origin of cultivated pearl millet expansion using kernel density estimation.

crop in Africa and India, but its precise origin is currently unknown (Tostain, 1992; Oumar et al., 2008; Clotault et al., 2012). When we applied an ABC approach to cultivated pearl millet genomes, we obtained a result supporting the Northern Mali region as the most probable geographic origin of expansion. Although the accuracy of the ABC approach was validated with extensive computer simulations of range expansion, the empirical results pointed out some limitations of our model for the data. The uncertainty around 18° reported for the latitude of origin was high, and improving our estimate would require supplementary information on past environmental conditions, carrying capacities and gene flow between pearl millet and related species. Interestingly, our results rejected an eastern origin for the expansion of the domesticated cereal. This result is consistent with recent archeological studies using both wild and cultivated samples, that pinpointed the Mali-Niger region as the most likely origin of domestication of pearl millet (Manning et al., 2011; Ozainne et al., 2014).

To conclude, singletons are a major component of the site frequency spectrum for many model and non-model species. The density of singletons in genomes has recently proven useful to detect selection in human genomes (Field et al., 2016). Here we showed that the density of singletons in geographic space is useful for providing local estimates of genetic diversity and key insights on the demographic history of a species.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This work has been partially supported by the Agence Nationale de la Recherche, project AFRICROP, ANR-13-BSV7-0017, and by the LabEx PERSYVAL Lab, ANR-11-LABX-0025-01, funded by the French program Investissement d'Avenir.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00139/full#supplementary-material

Figure S1 | Averaged proportion of singletons in population 1, and standard deviations in populations 1 and 2, as functions of the shrink rate.

Figure S2 | Individual vs. population sampling after a range expansion simulation scenario (Sahel origin). **(A,B)** Homogeneous environment. Maps of the empirical distribution of singletons (individual sampling) and expected heterozygosity (true population sampling). **(C,D)** Inhomogeneous environment.

Figure S3 | Population sampling after a range expansion simulation scenario (Sahel origin). **(A,B)** Homogeneous environment. Maps of the empirical distribution of singletons (true population sampling) and expected heterozygosity (true population sampling). **(C,D)** Inhomogeneous environment.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

Auer, P. L., and Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* 7:16. doi: 10.1186/s13073-015-0138-2

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Ann. Rev. Ecol. Evol. Syst.* 41, 379–406. doi: 10.1146/annurev-ecolsys-102209-144621

Blum, M. G. B., and François, O. (2005). Minimal clade size and external branch length under the neutral coalescent. *Adv. Appl. Probabil.* 37, 647–662. doi: 10.1017/S0001867800000409

Blum, M. G. B., and François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Stat. Comput.* 20, 63–73. doi: 10.1007/s11222-009-9116-0

Caliebe, A., Neininger, R., Krawczak, M., and Rösler, U. (2007). On the length distribution of external branches in coalescence trees: genetic diversity within species. *Theor. Popul. Biol.* 72, 245–252. doi: 10.1016/j.tpb.2007.05.003

Clotault, J., Thuillet, A. C., Buiron, M., De Mita, S., Couderc, M., Haussmann, B. I., et al. (2012). Evolutionary history of pearl millet (*Pennisetum glaucum* [L.] R. Br.) and selection on flowering genes since its domestication. *Mol. Biol. Evol.* 29, 1199–1212. doi: 10.1093/molbev/msr287

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* 1:131. doi: 10.1038/ncomms1130

Cressie, N. (2015). *Statistics for Spatial Data*. New-York, NY: John Wiley and Sons.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* 25, 410–418. doi: 10.1016/j.tree.2010.04.001

Csilléry, K., François, O., and Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* 3, 475–479. doi: 10.1111/j.2041-210X.2011.00179.x

Currat, M., Ray, N., and Excoffier, L. (2004). SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol. Ecol. Notes* 4, 139–142. doi: 10.1046/j.1471-8286.2003.00582.x

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., et al. (2016). Detection of human adaptation during the past 2000 years. *Science* 354, 760–764. doi: 10.1126/science.aag0776

Frichot, E., and François, O. (2015). LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol.* 6, 925–929. doi: 10.1111/2041-210X.12382

Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133, 693–709.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., et al. (2011). Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11983–11988. doi: 10.1073/pnas.1019276108

Hartigan, J. A., and Wong, M. A. (1979). A *K*-means clustering algorithm. *Appl. Stat.* 28, 100–108. doi: 10.2307/2346830

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7:44.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338. doi: 10.1093/bioinformatics/18.2.337

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi: 10.1038/nature09298

Keinan, A., and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740–743. doi: 10.1126/science.1217283

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95, 5–23. doi: 10.1016/j.ajhg.2014.06.009

Manning, K., Pelling, R., Higham, T., Schwenniger, J. L., and Fuller, D. Q. (2011). 4500-year old domesticated pearl millet (*Pennisetum glaucum*) from the Tilemsi Valley, Mali: new insights into an alternative cereal domestication pathway. *J. Archaeol. Sci.* 38, 312–322. doi: 10.1016/j.jas.2010.09.007

Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166, 351–372. doi: 10.1534/genetics.166.1.351

Mathieson, I., and McVean, G. (2014). Demography and the age of rare variants. *PLoS Genet.* 10:e1004528. doi: 10.1371/journal.pgen.1004528

Memon, S., Jia, X., Gu, L., and Zhang, X. (2016). Genomic variations and distinct evolutionary rate of rare alleles in *Arabidopsis thaliana*. *BMC Evol. Biol.* 16:25. doi: 10.1186/s12862-016-0590-7

Novembre, J., and Slatkin, M. (2009). Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. *Evolution* 63:2914. doi: 10.1111/j.1558-5646.2009.00775.x

O'Connor, T. D., Fu, W., Turner, E., Mychaleckyj, J. C., Logsdon, B., Auer, P., et al. (2015). Rare variation facilitates inferences of fine-scale population structure in humans. *Mol. Biol. Evol.* 32, 653–660. doi: 10.1093/molbev/msu326

Ozainne, S., Lespez, L., Garnier, A., Ballouche, A., Neumann, K., Pays, O., et al. (2014). A question of timing: spatio-temporal structure and mechanisms of early agriculture expansion in West Africa. *J. Archaeol. Sci.* 50, 359–368. doi: 10.1016/j.jas.2014.07.025

Oumar, I., Mariac, C., Pham, J.-L., and Vigouroux, Y. (2008). Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. *Theor. Appl. Genet.* 117, 489–497. doi: 10.1007/s00122-008-0793-4

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

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69, 124–137. doi: 10.1086/321272

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* 19, 212–219. doi: 10.1016/j.gde.2009.04.010

Schraiber, J. G., and Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* 16, 727–740. doi: 10.1038/nrg4005

Slatkin, M. (1985). Rare alleles as indicators of gene flow. *Evolution* 39, 53–65. doi: 10.1111/j.1558-5646.1985.tb04079.x

Tavaré, S. (2004). "Ancestral inference in population genetics," in *Lectures on Probability Theory and Statistics*, Lecture Notes Math. 1837, ed J. Picard (Berlin: Springer), 1–188.

Tennessen, J., Bigham, A., O'Connor, T., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69. doi: 10.1126/science.1219240

Tostain, S. (1992). Enzyme diversity in pearl millet (*Pennisetum glaucum* L.). *Theor. Appl. Genet.* 83, 733–742. doi: 10.1007/BF00226692

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., et al. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat. Biotechnol.* doi: 10.1038/nbt.3943. [Epub ahead of print]. Available online at: http://ceg.icrisat.org/ipmgsc/

Weigel, D. (2012). Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol.* 158, 2–22. doi: 10.1104/pp.111.189845

Weigel, D., and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.* 10:107. doi: 10.1186/gb-2009-10-5-107

Zhu, C., Li, X., and Yu, J. (2011). Integrating rare-variant testing, function prediction, and gene network in composite resequencing-based genome-wide association studies (CR-GWAS). *Genes Genomes Genet.* 1, 233–243. doi: 10.1534/g3.111.000364

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.