RC25407 (WAT1308-063) August 28, 2013
Other

IBM Research Report

Effect of Sampling on the Extent and Accuracy of the Inferred Genetic History of Recombining Genome

Daniel E. Platt, Filippo Utro, Laxmi Parida
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 208
Yorktown Heights, NY 10598
USA

Effect of sampling on the extent and accuracy of the inferred genetic history of recombining genome

Daniel E. Platt[†] , Filippo Utro[†] , Laxmi Parida[*]

[1]Computational Biology Center, IBM T. J. Watson Research, Yorktown Heights, NY 10598, USA.
[†] These authors contributed equally to this work.

Email: watplatt@us.ibm.com; futro@us.ibm.com; *parida@us.ibm.com;

*Corresponding author

Abstract

Background: Accessible biotechnology is enabling the cataloging of genetic variants in individuals in populations at unprecedented scales. The use of phylogeny of the individuals within populations allows a model-based approach to studying these variations, which is important in understanding relationships between and across populations. For the somatic genome, however, the phylogeny must take recombinations (and other genetic mixing events) into account. Hence the resulting topology is more complex than a tree. Unlike a tree topology, it is not as apparent which events are visible from the extant samples. An earlier work presented a mathematical model (called the minimal descriptor) for teasing apart the inherent visible information from that which any specific algorithm might see. We use this framework to study the effect of sampling sizes on the overall inferred genetic history. In this paper, we seek to understand the extent, characteristics (in terms of recent versus ancient genetic events) and reliability of what was resolvable within field samples drawn from modern populations.

Results: We observed that most of the visible ancient events are recoverable from relatively small sample sizes. However, without identification of this relatively small minority of ancient genetic events, most of the signal will appear to reflect modern events and admixtures. We also found that the more ancient events are likely to be reproduced with higher fidelity between multiple samplings, and that the identified older events are less likely to yield false positive discrimination between populations.

Conclusions: We conclude that a recombinant phylogenetic reconstruction is necessary to identify which markers are most likely to discriminate ancient events, and to discriminate between populations with lower risk of false

positives. Secondly, on a broader note, this study also provides a general methodology for a critical assessment of the inferred common genetic history of populations (say, in plant cultivars or animal populations).

Background

The promise of higher resolution available from the vast number of mutations in the somatic genome to mark the impact on genetics of historic and prehistoric human migrations provides significant motivation to explore recombinant data. Unlike mtDNA or Y-chromosome non-recombinant signals, analysis of the somatic data has shown a persistent sensitivity to more recent migrations showing significant commonality even between otherwise apparently isolated populations [1–4]. Given the significant difference in the character of recombinant and non-recombinant signals, the question is whether the recombinant genome carries information of the deeper genetics reflecting more ancient population events.

We have therefore pursued the question of how much information is available in the recombinant genome to inform population events. Note that for a tree topology, all the events are considered visible from the extant samples. However, this does not hold for a general phylogeny, called ARG (ancestral recombination graph), that also models the mixing genetic events. We use the notion of minimal descriptor ARGs , which identifies the maximum historical information actually visible from the extant samples. We applied analytical techniques to the structure of ARGs, treating these as extended phylogenies that include recombination events, exploring how different lineages carried by segments are transmitted within populations, subject to mutation, recombination, and coalescence. This analysis was applied to data generated by COSI [5] a population genetic simulation application that produces populations bearing signatures of specific demographies, including Africans, Europeans, Asians, and African Americans. In this case, we generated multiple histories, each from the various demographic models, and computed the average number of reconstructable nodes both in total, and within specific time-sliced epochs. We have shown that as much as 30% of the information of significant genetic historical events is lost through coalescence [6]. In non-recombinant phylogenies, coalescence causes lineages marked by some mutations to be lost. As a result, the final record shows multiple SNPs accumulating along an edge marking a lineage between coalescence events keeping time ("equivalent SNPs"), but where the information of which SNP occurred in what time order is lost. In the case of recombinant phylogenies, not only does information of SNP order and whole lineages get lost, but the association between segments that evolved together within lineages of organisms may be completely lost through drift, with the only surviving record showing associations between segments which were only recently related to each other. Even given these losses, we were able to show significant alignment between numbers of nodes that were reconstructed, and the historical events (out-of-Africa migration, post last glacial period expansions, Neolithic agricultural revolution and expansion) that mark each of the demographies.

Since it appears ancient events are available within the somatic genome's ARGs, we sought to see how sampling impacts the availability of information of the more ancient events that may differentiate the expansions from different refugia that so boldly paint the non-recombinant genetics of human expansions. Given that information is reconstructable, we report on analysis of the impact of sampling in a population to see how much information is lost in the process of field sampling. In this set of experiments, we generated a single set of demographic populations using a simulator, and explored the reconstructable mdARGs available from different sizes of populations. We resampled multiple times to understand how consistently the various nodes in the mdARGs were identifiable solely from data carried within the samples, to understand whether more ancient information was lost in diffusion, or was reinforced as common information among samplings.

Methods

For the convenience of the reader here we briefly provide some general definition of the ARG. An ARG is a phylogenetic structure that encodes both duplication events, such as mutations, as well as genetic exchange events, such as recombinations: this captures the genetic dynamics of a population evolving over generations. From a topological point of view, an ARG is always a directed acyclic graph where the direction of the edges is toward the more recent generation, where the edge is annotated with genetic mutation events, such as single nucleotide polymorphisms. It is possible that there may be edges with empty labels. It is worth pointing out that the *length* of the edge, not to be confused with the edge label, represents the epoch defined by the age (or depth) of the two incident nodes. Finally, we recall that a chain node has a single incoming edge and a single outgoing edge.

In [7], a structure-preserving and samples-preserving core of an ARG G, called the minimal descriptor ARG (mdARG) of G, was identified. Its structure-preserving character ensures that the topology and the all the branch lengths of the marginal trees of the minimal descriptor ARG are identical to that of G, and the samples-preserving property asserts that the patterns of genetic variation in the samples of the minimal descriptor ARG are exactly the same as that of G. It was also shown that an unbounded G has a finite minimal descriptor, that continues to preserve critical graph-theoretic properties of G. Thus this lossless and bounded structure is well defined for all ARGs (including unbounded ARGs) and we use the same here. However, a minimal descriptor of an ARG may not be unique. An extensive study of the identification of the reconstructable fraction of the ARG is given in [6].

Simulating the populations

We used COSI [5], which is the only population simulator, to the best of our knowledge, that provides the ARG as well as produces populations that match the genetic landscape of the observed human populations, which we needed to explore whether such demographic events would be recoverable in the genetic signal. We used the *bestfit model* in COSI to simulate the samples with a calibrated human demography for different populations, proposed by Schaffner et al. [5]. This demography generates data matching three structured continental populations: Africans, Europeans and Asians. An admixed population, the Afro-Americans, can also be generated. This simulator has also been used in literature as a gold standard for generating the demographies [8–11]. We used the default parameters provided by the *bestfit model* for the mutation rate, sequence length and recombination rate, while we increase the sample size of 5000 to generate the the ARG used in this manuscript.

Sampling the populations and build the mdARG

Given an ARG G, we randomly select a fraction x of the leaves (representing a field sample) of the ARG, and remove the other leaves, for x representing 1.0%, 5.0%, 10%, and 30%. We then prune those internal nodes that do not transmit any genetic information to the selected leaves. This process leads to a new ARG G'. We compute the mdARG of G' removing the chain nodes and the non t-coalescent node (a coalescent node where the genetic material transmitted to his descendent has no intersection) as detailed in [6].

We replicated this field sampling simulation process 20 times for each x. Figure 1 presents an example to show, for a specific history represented as an ARG, how different samplings, modeling how a researcher in the field might happen to collect samples from an extant population, can lose information regarding historical genetic events (recombinations, mutations, or coalescences). Figure 1(a) shows a complete ARG showing all the historical events for a small population of 10 individuals. Figures 1(b) and 1(c) show the complete ARGs visible for each of two samplings of 3 individuals drawn from the population. Figure 1(d) shows the mdARG of all the reconstructable nodes given the information available in the modern population. Of these, the reconstructable nodes that are available to each of the two samples 1(b) and 1(c) are shown in Figures 1(e) and 1(f) respectively.

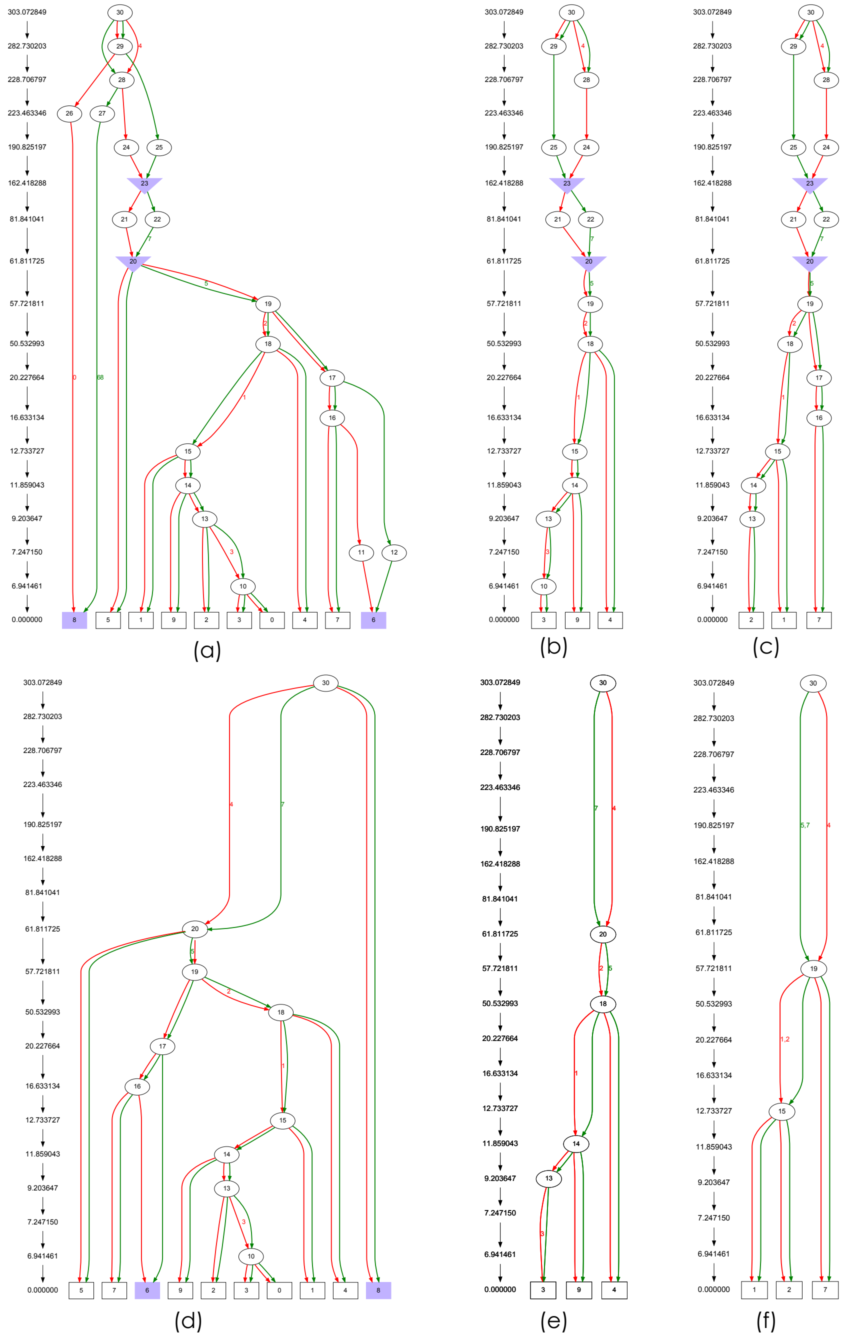

Figure 1: Example ARG (a) with 10 leaves, and the nodes that can be resolved from two distinct samplings (b) and (c) each with 3 samples drawn; also the mdARG (d) associated with the ARG is shown, together with the nodes resolvable from the two samples (e) and (f).

Results and Discussion

Results

We generated populations representing Europeans and Africans with populations of 5000 individuals. From these, we resampled 20 times with fractions of the European population shown in Figures 2(a)-2(d) ranging over (a) 1%, (b) 5%, (c) 10%, and (d) 30%, and the African populations shown in Figures 3(a) through 3(d) showing the same range of percentages sampled. The x-axis is measured in generations, while the y-axis counts the number of visible nodes reconstructable from each sampling. The vertical bars represent the modern increase in population due to medical and agricultural advances, the Last Glacial Period bottlenecks, and the out-of-Africa events. From each of those computations, we computed the number of nodes within each epoch slice of 200 generations thickness. We see in Figures 2(a) through 2(d) and Figures 3(a) through 3(d), that the various historical events marking each of the demographies were consistently visible over multiple histories in the reconstructable node counts, and that these signals were adequate to distinguish between European (Figures 2) and African (Figures 3) demographics (note there was no "Out-Of-Africa" founding effect for African populations). It is also important to note that most of the nodes are associated with sampled leaves. The ability to resolve the shape of the Kingman coalescent in this region depends on the breadth of sampling. Sampling at 1% and 5% does not provide sufficient detail to resolve the shape of the coalescent, while broader coverage at 10% and 30% does. By contrast, the ability to resolve older nodes is much more complete, and even amplified at lower sampling rates because the oldest nodes are the most likely to be supported by multiple samples.

Figures 4 show sample coverage by rows of the European population, with the first row showing 1%, second row 5%, third row 10%, and fourth row 30%. Each row shows three columns. In the first column, the number of times each node was visible is displayed. In this case, the x-axis marks each node id number. Since they are numbered sequentially as COSI generates them, they are roughly chronological, but the x-axis is radically expanded in some regions according to the number of nodes that are present in any given epoch. Specifically, the new nodes that are visible in the leaves are very large in number compared to more historic nodes, as marked by the vertical red line. It is clear that historic nodes find more and more consistent reconstructability from among various possible field sample sets. As sample sizes are increased, the number of leaf nodes identified increases. The second column shows the distribution of how many nodes are sampled by multiple re-samplings of the population within each 200 year epoch, with the x-axis showing time in generations. It is seen that almost all of the older nodes are completely supported by each resampling, while the new leaves are more weakly characterized. The third column expands the saturation seen at the far left of the second column figures, showing the distribution of node samplings per epoch as a whisker diagram.

Figures 5 show the total number of nodes available to the sampling within (a) Europeans, and (b) Africans. It is seen that, within the population, the recent nodes completely dominate the genetic history of the population. While the actual history of the recent nodes are most poorly defined in terms of resolving the Kingman coalescent, they still tend to dominate the genetic signal.

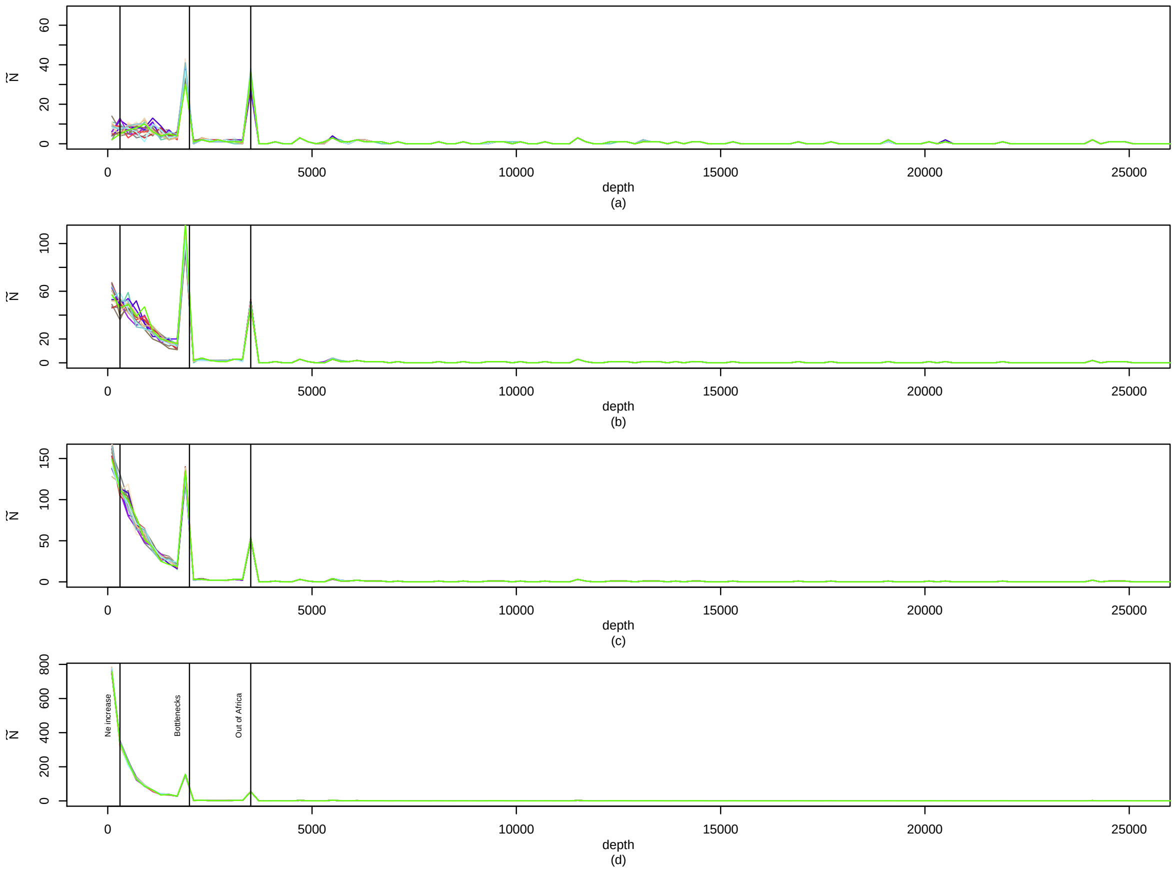

Figure 2: The number of resolvable nodes found within 200 generation epochs from (a) 1%, (b) 5%, (c) 10%, and (d) 30% of the European population sampled.

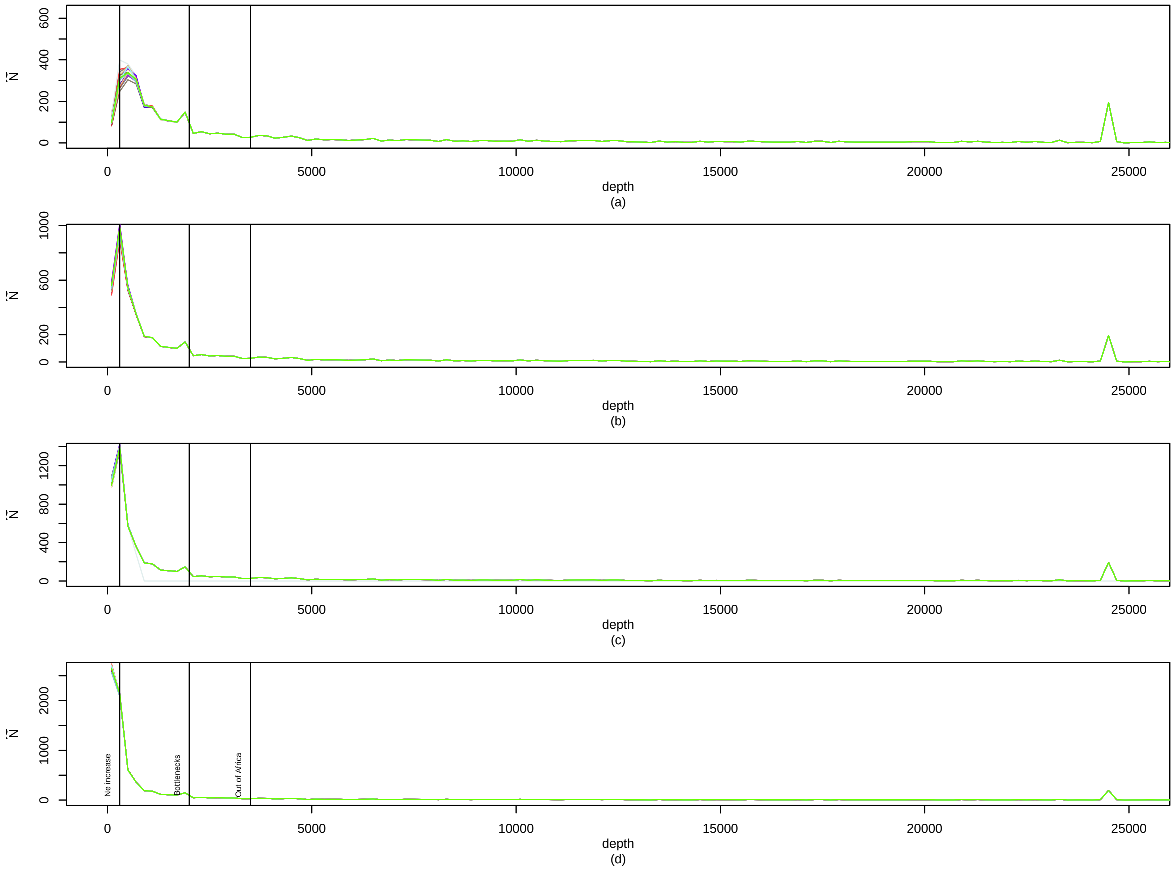

Figure 3: The number of resolvable nodes found within 200 generation epochs from (a) 1%, (b) 5%, (c) 10%, and (d) 30% of the African population sampled.

Discussion

Noting that a number of recent publications suggested that the somatic genome tends to inform relatively shallow population events [1–4], we set out to understand what kind of information was available in the somatic genome, how much fidelity would be available for resolving more ancient events compared to modern events, and, in this study, how much of that resolvable information was impacted by sample sizes. We therefore used a population simulation application, COSI [5], that simulated recombinant events, generated ARGs, and provided mechanisms for modeling specific human populations, such as Africans, Europeans, Asians, and African-Americans. In prior work [7], we explored how much information was inherently available after coalescence and drift removed information, which we call an mdARG. In this study, we extended the methodology to explore how much of that information is represented within multiple instances of "field sampling" studies, by exploring how often mdARG nodes were represented in networks that terminated in

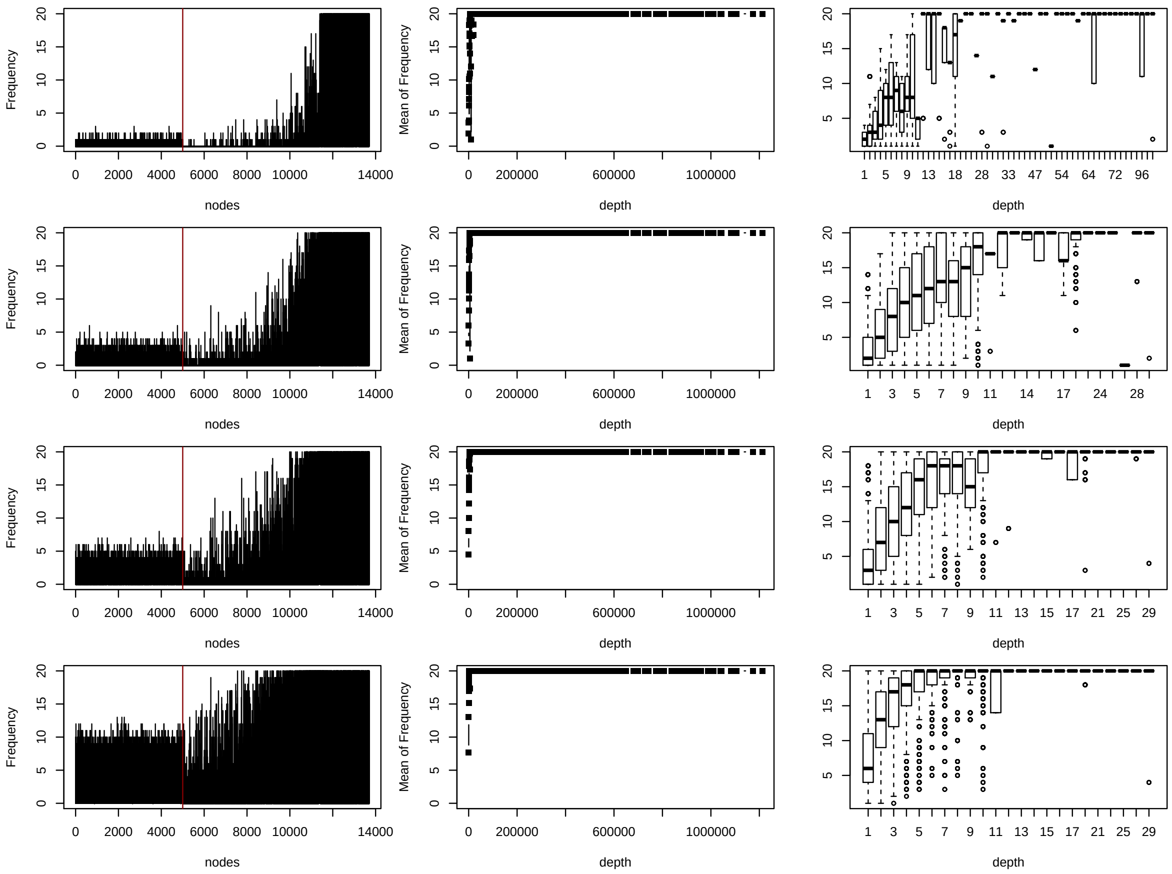

Figure 4: Multiple views of support for nodes by multiple resamplings of the European population. Rows show sampling fractions of 1%, 5%, 10%, and 30%. First column shows support by samplings of each individual node, x-axis as indexed by COSI, generally in chronological order. The leaf nodes are all to the left of the red vertical line, and all are referenced in a single time slice reflecting modern time. They represent more than 1/3 of the total number of nodes. Second column shows distribution of average frequencies the nodes are sampled by within the 200 generation epoch intervals. The third column expands the saturation region, showing the statistical characterization of the sampling frequencies within each 200 generation epoch interval by whisker diagram.

sampled leaves.

In Figures 4, it is possible to see the range of variation in how the more ancient information is more consistently constructable from relatively sparse field samples as representatives of their epochs. However, the identifiable nodes, as shown in Figures 2, 3, and 4, shows samples dominated by the most recent nodes. While Figures 4 show that these nodes are less likely to be consistent between resamplings, the SNPs associated with the recent nodes present a signal that apparently differentiates between populations and regions. Since the chances any given sampling of the population will clearly identify newer nodes is lower, the difficulty

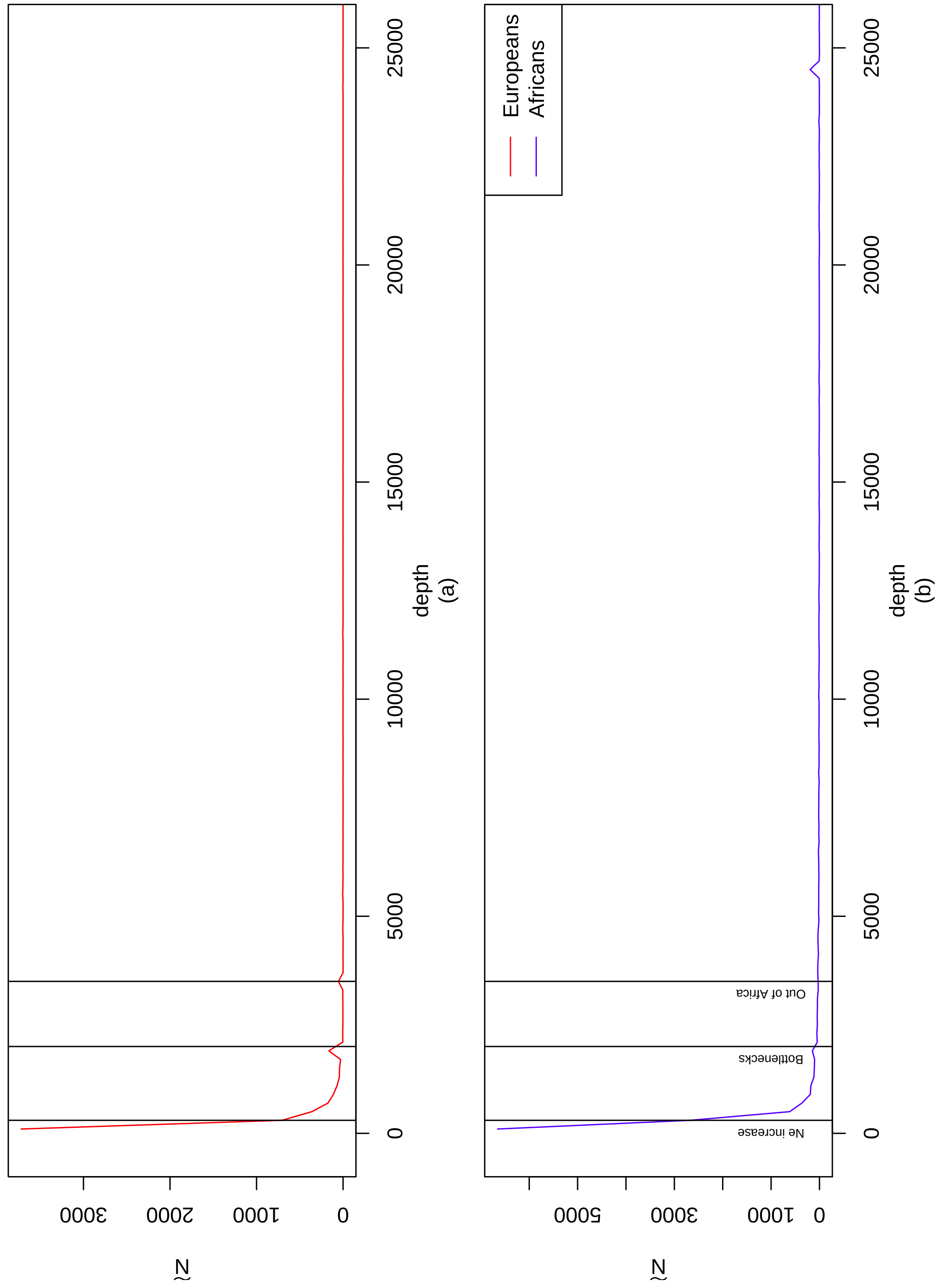

Figure 5: This shows the complete number of nodes represented in the mdARGs for (a) European and (b) African demographic simulations.

of recognizing a true differentiation between populations using those newer nodes due to chances of a false positive is relatively high. Also, their apparent time signature will tend to be significantly more recent.

Figures 2 show that the surviving ancient nodes mark demographic events with a sufficient fidelity capable of clearly distinguishing the various demographies' historic events. Specifically, there is a founder event signature for Europeans marking the Out-Of-Africa event that is not present among the African demographics generated by COSI. Further, these older nodes tend to have the most resolved signal compared to recent nodes, even though their numbers tend to be much lower in comparison to the more recent leaf nodes.

This may also shed light on why Principal Component Analysis (PCA) has been so successful in resolving geographic regions [12]. While many of the SNPs that contribute to distance are uncorrelated, edges between older nodes are marked by multiple "equivalent" SNPs. PCA is very sensitive to correlated variations, since these components tend to contribute disproportionately to distance. That implies that leading principal components are more likely to identify older genetic variations than the relatively new SNPs in leaf nodes.

Conclusions

In this study, we sought to understand how much was resolvable of the historical genetic information in extant population samples, and what is likely to be resolvable within field samples drawn from modern populations. We have seen from sampling mdARGs that most of the ancient nodes will be identified from relatively small samples drawn by field researchers from any given populations. However, we also note that, without identification of which of the relatively small minority of nodes marking ancient genetic events, most of the signal will appear to reflect modern events and admixtures. This argues strongly for a need for identification of which SNPs actually associate with ancient *vs.* recent genetic events, and which may mark the more ancient population expansion events. We also found that the more ancient nodes are likely to be produced with higher fidelity between multiple samplings, and that those older identified nodes are less likely to yield false positive discriminations between populations.

We therefore conclude that a recombinant phylogenetic reconstruction – as perhaps represented in ARGs – is necessary to identify which nodes and markers are perhaps most likely to discriminate ancient events, and to discriminate between populations.

Competing Interests

The authors declare they have no competing interests.

Author's contributions

LP designed the study; DP designed and FU carried out the experiments. All authors wrote the paper.

References

1. Botigué L, Henn B, Gravel S, Maples B, Gignoux C, Corona E, Atzmon G, Burns E, Ostrer H, Flores C, Bertranpetit J, Comas D, Bustamante C: **Gene flow from North Africa contributes to differential human genetic diversity in southern Europe**. *Proceedings of the National Academy of Sciences* 2013, **110**:11791–11796.

2. Moorjani P, Patterson N, Hirschhorn J, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price A, Reich D: **The History of African Gene Flow into Southern Europeans, Levantines, and Jews**. *PLoS Genet* 2011, **7**:e1001373+.

3. Ralph P, Coop G: **The Geography of Recent Genetic Ancestry across Europe**. *PLoS Biol* 2013, **11**:e1001555+.

4. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra M, Rojas W, Duque C, Mesa N, Garcia L, Triana O, Blair S, Maestre A, Dib J, Bravi C, Bailliet G, Corach D, Hunemeier T, Bortolini M, Salzano F, Petzl-Erler M, Acuna-Alonzo V, Aguilar-Salinas C, Canizales-Quinteros S, Tusie-Luna T, Riba L, Rodriguez-Cruz M, Lopez-Alarcon M, Coral-Vazquez R, Canto-Cetina T, Silva-Zolezzi I, Fernandez-Lopez J, Contreras A, Jimenez-Sanchez G, Gomez-Vazquez M, Molina J, Carracedo A, Salas A, Gallo C, Poletti G, Witonsky D, Alkorta-Aranburu G, Sukernik R, Osipova L, Fedorova S, Vasquez R, Villena M, Moreau C, Barrantes R, Pauls D, Excoffier L, Bedoya G, Rothhammer F, Dugoujon JM, Larrouy G, Klitz W, Labuda D, Kidd J, Kidd K, Di Rienzo A, Freimer N, Price A, Ruiz-Linares A: **Reconstructing Native American population history**. *Nature* 2012, **488**:370–374.

5. Schaffner S, Foo C, Gabriel S, Reich D, Daly M, Altshuler D: **Calibrating a coalescent simulation of human genome sequence variation**. *Gen. Res* 2005, **15**:1576–1583.

6. Utro F, Pybus M, Parida L: **Sum of parts is greater than the whole: inference of common genetic history of populations**. *BMC Genomics* 2013, **14**:S10.

7. Parida L, Palamara P, Javed A: **A minimal descriptor of an ancestral recombinations graph**. *BMC Bioinformatics* 2011, **12**:S6.

8. Javed A, Pybus M, Melè M, Utro F, Bertranpetit J, Calafell F, Parida L: **IRiS: Construction of ARG network at genomic scales**. *Bioinformatics* 2011, **27**:2448–2450.

9. Li H, Durbin R: **Inference of human population history from individual whole-genome sequences**. *Nature* 2011, **475**:493–496.

10. Pickrell J, Coop G, Novembre J, Kudaravalli S, Li J, Absher D, Srinivasan B, Barsh G, Myers R, Feldman M, Pritchard J: **Signals of recent positive selection in a worldwide sample of human populations**. *Genome Research* 2009, **19**:826–837.

11. Sabeti P, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne E, Mccarroll S, Gaudet R, Schaffner S, Lander E, Consortium TIH: **Genome-wide detection and characterization of positive selection in human populations**. *Nature* 2007, **449**:913–918.

12. Javed A, Melé M, Pybus M, Zalloua P, Haber M, Comas D, Netea M, Balanovsky O, Balanovska E, Jin L, Yang Y, ArunKumar G, Pitchappan R, Bertranpetit J, Calafell F, Parida L: **Recombination networks as genetic markers in a human variation study of the Old World**. *Human Genetics* 2012, **131**:601–613.