SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
DATA PROCESSING
RISK FACTORS

Received
25 March 2014

Accepted
7 July 2014

Published
23 July 2014

Correspondence and requests for materials should be addressed to Q.S. (qsong@msm.edu)

* These authors contributed equally to this work.

Accurate Inference of Local Phased Ancestry of Modern Admixed Populations

Yamin Ma[1]*, Jian Zhao[2]*, Jian-Syuan Wong[1], Li Ma[1], Wenzhi Li[1], Guoxing Fu[1], Wei Xu[1], Kui Zhang[3], Rick A. Kittles[4], Yun Li[5] & Qing Song[1]

[1]Cardiovascular Research Institute, Morehouse School of Medicine, Atlanta, Georgia, USA, [2]DNAncestree Inc, Atlanta, Georgia, USA, [3]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, USA, [4]Department of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA, [5]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

Population stratification is a growing concern in genetic-association studies. Averaged ancestry at the genome level (global ancestry) is insufficient for detecting the population substructures and correcting population stratifications in association studies. Local and phase stratification are needed for human genetic studies, but current technologies cannot be applied on the entire genome data due to various technical caveats. Here we developed a novel approach (aMAP, ancestry of Modern Admixed Populations) for inferring local phased ancestry. It took about 3 seconds on a desktop computer to finish a local ancestry analysis for each human genome with 1.4-million SNPs. This method also exhibits the scalability to larger datasets with respect to the number of SNPs, the number of samples, and the size of reference panels. It can detect the lack of the proxy of reference panels. The accuracy was 99.4%. The aMAP software has a capacity for analyzing 6-way admixed individuals. As the biomedical community continues to expand its efforts to increase the representation of diverse populations, and as the number of large whole-genome sequence datasets continues to grow rapidly, there is an increasing demand on rapid and accurate local ancestry analysis in genetics, pharmacogenomics, population genetics, and clinical diagnosis.

Population stratification is a growing concern in genetic-association studies[1–3]; it potentially leads to both spurious associations and reduced statistical power. Admixture has created mosaic chromosomes in human populations; even within the same individual, different segments and different homologous chromosomes may have different ancestral origins. Averaged ancestry at the genome level (global ancestry) is insufficient for detecting population substructures and correcting population stratifications in association studies. Although numerous approaches have been developed to infer local ancestry[4–18], several key issues remain unsolved. Here we report a new approach that exploits the sequence content information of each personal haplotype instead of using allele frequencies. This approach has been implemented into a software tool called aMAP (ancestry of Modern Admixed Populations).

Results

We first empirically evaluated the performance of aMAP on 6 HapMap populations, ASW (African-Americans), YRI (West Africans), CEU (Caucasians), CHB and CHD (Chinese), JPT (Japanese), and MEX (Mexican-Americans) (Supplementary Tables S1 & S2, Supplementary Fig. S1). The data not only showed the accuracy of aMAP at the global level, but also revealed substantial intra-individual variations, between different loci and between two homologous chromosomes (Supplementary Tables S3 & S4). About 21–31% of genomic regions are ancestrally heterozygous at the same locus between two homologous chromosomes. These observations suggest that local ancestry will not be sufficient for stratification; the ancestral information of each chromosome should also be considered.

To quantitatively measure the accuracy of aMAP, we simulated a 20-generation admixed population (50% YRI and 50% CHBCHD). On this dataset, the analysis with aMAP reached a mean haploid accuracy of 99.4% (Supplementary Tables S5 & S6), and the analysis with LAMP-HAP showed a mean haploid accuracy of 97.5% (Supplementary Table S5). We also analyzed the performance of aMAP when the ancestral blocks become smaller. Our results showed that aMAP maintained at a high level of accuracy over generations, but the accuracy

of LAMP-HAP decreased rapidly with the number of generations (Supplementary Table S5, Supplementary Fig. S2). There is no substantial difference on the aMAP performance between the SNPs in different MAF (minor allele frequency) ranges (Supplementary Fig. S3). To determine the resolution of aMAP, we calculated the haploid accuracies according to the sizes of ancestral segments. This showed that LAMP-HAP performed well on those large ancestral segments, but poorly on smaller segments; aMAP performed well on both large and small ancestral segments (Supplementary Fig. S4), indicating its potential application for older admixed populations.

To examine the performance of aMAP on closely related populations, we first analyzed the HapMap Chinese and Japanese populations. The results demonstrated the capacity of aMAP to stratify the Chinese and Japanese populations (Supplementary Fig. S5). To further quantitatively evaluate its performance on closely related populations, we simulated a CHBCHD-JPT admixed dataset (50% CHBCHD and 50% JPT). The mean haploid accuracy of the aMAP results was 98.5% (Supplementary Table S7), with sensitivity (98.62%), specificity (97.84%), PPV (97.86%) and NPV (98.61%) on those CHBCHD segments (Supplementary Table S8).

Existing approaches require *a priori*, a good proxy of reference panels to reveal the true ancestries of admixed individuals. In reality however, selecting good proxy reference panels for admixture deconvolution is highly challenging because the researchers and often the subjects themselves may not know their precise ancestral backgrounds, thus it is very likely that the reference panels fail to cover all of the ancestral origins of a subject. To examine the performance of aMAP under this scenario, we analyzed the ASW (African-Americans) individuals without the YRI (West Africans) reference (Fig. 1). The results showed that aMAP successfully detected the absence of a major population in the reference panel by showing a large portion of yellow segments ("others"). Furthermore, aMAP successfully labeled those African-originated loci as "others". We then created YRI-CHBCHD simulated dataset to quantitatively measure the performance of aMAP in this scenario (Supplementary Fig. S6). The results showed that aMAP could detect those YRI segments when the YRI reference was missing (reported as "others") with 99.5% PPV, 82.3% sensitivity, 91.7% NPV, and 99.6% specificity (Supplementary Table S9). Meanwhile, we evaluated the performance of LAMP-HAP with the data under this scenario. We found that when the YRI reference was missing, LAMP-HAP did not detect the absence of a major reference population, it assigned those YRI segments mainly into CEU (Supplementary Fig. S6). This advantage will enable aMAP to be applicable to more general or realistic scenarios in local ancestry analysis.

It is becoming increasingly clear that controlling for population stratification at the continental level is insufficient and that subcontinental ancestries must be considered in population studies[6]. The ideal reference panel should be composed of multiple continental, subcontinental, regional and ethnic populations; however, current approaches can only consider two or three ancestral populations at a time due to computational limitations. This caveat prevents the inclusion of more reference populations, and increases the chance of missing a major population reference in a real analysis. To examine if aMAP can overcome this multi-way ancestry inference challenge, we simulated a six-way admixed dataset with 6 HapMap populations (CEU 12.5%, GIH 12.5%, CHBCHD 25%, YRI 12.5%, MKK 25% and LWK 12.5%). It took aMAP 0.18 seconds to complete a six-way local ancestry inference for a single chromosome-1 haplotype (Supplementary Table S10). The mean haploid accuracy was 98.55% in this 6-way inference (Supplementary Table S11).

Despite the continuing contributions of organized efforts such as the International HapMap Project and the 1000 Genomes Project, and the growing availability of publically available population data, molecular haplotypes are still unavailable for many subcontinental and regional populations. This motivated us to explore whether aMAP could utilize unphased genotypes as references. We downloaded the genotypes and haplotypes from HapMap, and then inferred haplotypes statistically from the genotypes using the software Beagle. We observed highly accurate results when these statistically resolved haplotypes were used by aMAP as the reference to infer the ancestry, compared to the data using molecular haplotypes as references (Supplementary Table S12, Supplementary Fig. S7). We believe that it is because that these statistically inferred haplotypes are very accurate locally within the phasing range of tens to hundreds of kilobases, which is larger than the aMAP window size and enable

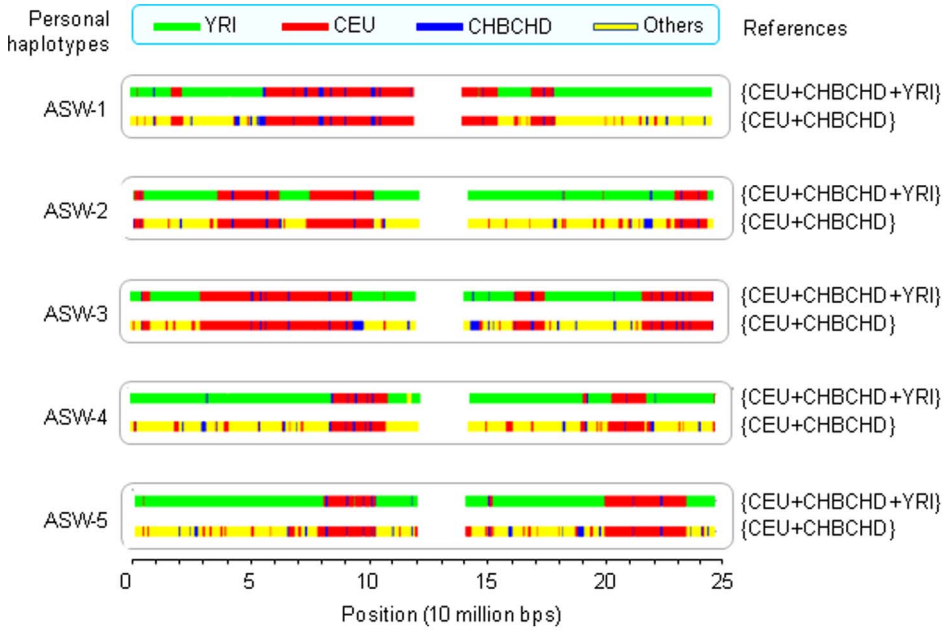

Figure 1 | The results of aMAP on ASW with proxy reference panels and imperfect reference panels. ASW individuals (chromosome-1) were analyzed by aMAP with and without the YRI reference panel. The results of 5 personal haplotypes are shown. When YRI was missing in the reference panel, those African-originated segments (green) could be detected and reported as "others" (yellow).

aMAP to tolerate the switching errors in statistically inferred haplotypes. This strongly suggests that aMAP can be applied on all populations that have genome-wide genotype data, such as GWAS data.

We examined the computing speed of aMAP on a regular desktop computer (an Intel Core, 3.40 GHz processor, 32 GB RAM). It took aMAP 55 seconds to finish an ancestry analysis on 20 ASW individual genomes (1.4 million SNPs) (Supplementary Table S13). The total runtime is linear to the number of SNPs (Fig. 2, Supplementary Table S13), the sample size (the number of sample haplotypes) (Supplementary Table S14, Supplementary Fig. S8), the reference size (the total number of reference haplotypes) (Supplementary Table S15, Supplementary Fig. S9), and the number of ancestral populations (Supplementary Fig. S10). Compared with LAMP-HAP, which is currently one of the fastest local ancestry inference software packages, aMAP was 923 times faster (Supplementary Table S13).

Discussion

In summary, we have presented a new method, aMAP, for locus-specific and haplotype-specific local ancestry inference. This new method is distinguished by high accuracy (99.4%), high-speed, high-resolution, and a scalability to larger datasets with respect to the number of SNPs, the number of samples, the total number of haplotypes in the reference panels, and the numbers of reference populations. Our method also exhibits a tolerance to missing ancestral reference panels, an applicability to genetically close populations, and a capacity for analyzing multi-way admixed individuals. As the biomedical community continues to expand its efforts to increase the representation of diverse populations, and as the number of large whole-genome sequence datasets continues to grow rapidly, there is an increasing demand on rapid and accurate local ancestry analysis in genetics, pharmacogenomics, population genetics, and clinical diagnosis.

Methods

The aMAP algorithm. The procedure of the aMAP method is composed of five steps, 1) pre-analysis of reference haplotypes, 2) parallel window scans, 3) horizontal data integration, 4) vertical data integration, and 5) border refinement.

Pre-analysis and pre-treatment of reference haplotypes. Briefly, aMAP first scans the reference haplotypes using a set of parallel non-overlapping sliding windows and compares sequence contents of those reference haplotypes in each sliding window within each given reference population and between different reference populations.

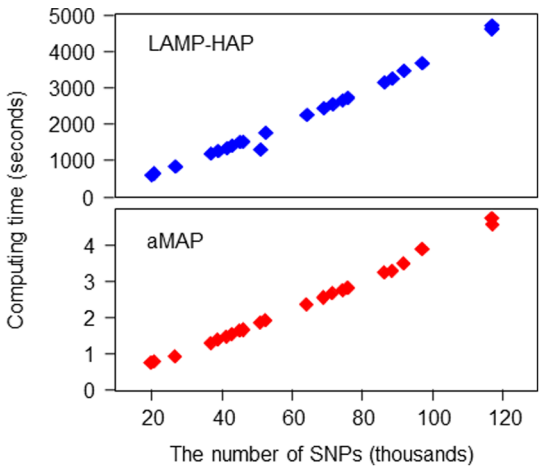

Figure 2 | The computing time of aMAP and LAMP. The whole-genome of 20 HapMap ASW individuals (African-Americans) were analyzed with three references (CEU, YRI, and CHBCHD). The computing speeds of aMAP and LAMP-HAP are compared, and both are linear to the total number of SNPs; the speed of aMAP is about 923 times faster than the speed of LAMP-HAP.

In this process, exact sequence matches are searched for among all haplotypes in the reference populations. Based on the results of these exact sequence comparisons, aMAP divides reference haplotypes of each window into several non-redundant groups, including one "population-unique" group for each of those populations given by users and a list of "common" (shared) groups in different combinations of those given reference populations (Supplementary Fig. S11). When a haplotype sequence in a window is observed in only one of the given reference populations, it will be placed into the population-specific group of the corresponding population; when a haplotype sequence in a window is observed in more than one given reference populations, it will be placed in a "common" group. To further illustrate how this process is done in practice, we provide a simplified example of this process in the Supplementary Table S16, and an actual case to show how this process is executed in the aMAP running on HapMap sample haplotypes in a window of the window-20 scan (Supplementary Table S17). When two reference populations are analyzed, in any window, all of those haplotypes will be divided into three groups, common group (shared by A and B), population-A group, and population-B group; when 4 populations are used as references (A, B, C, D), the reference haplotypes in each window will be divided into 15 groups, 4 population-unique groups (A, B, C, D), and 11 "common" groups (AB, AC, AD, BC, BD, CD, ABC, ABD, ACD, BCD, ABCD) (Supplementary Fig. S11). When two ancestral populations are genetically closer, they will share more haplotypes in their "common" pool and less haplotypes in the population-specific pools.

Parallel window scans. Then, aMAP scans each sample haplotype with a series of non-overlapping sliding windows simultaneously and seeks an exact sequence match to a reference haplotype in each window (Fig. 3, Supplementary Fig. S12). It records the scanning results by the group ID (ancestry calls) of each window, such as "A" (unique to population A), "B" (unique to population B), "AB" (shared between A and B), "ABCD" (shared among A, B, C, D). If no exact match is found, aMAP documents it as "other" (no match). The results will fall into three non-overlapping possibilities, 1) the sample haplotype matches to a reference haplotype in one of the given ancestral populations; 2) it matches to a reference haplotype in one of the "common" groups shared by any two or more populations; and 3) it does not match to any of the reference haplotypes. These parallel windows have different sizes and different number of windows; the default setting of the number of windows is 19, aMAP will scan each sample haplotype by 19 window tracks; the default window size of these 19 windows are, window-20 (the haplotypes of 20 consecutive SNPs are analyzed in each window), window-30 (30 consecutive SNPs), ……, and window-200 (200 consecutive SNPs).

Horizontal data integration. It is well-known that haplotype diversity varies dramatically across the genome in different populations. The optimal window size may vary with the genomic positions, the ethnohistory of each population, and the genetic distance between two reference populations. Therefore, there is no universally optimal window size for different regions across the genome and for different combinations of reference populations. As a solution, the aMAP algorithm scans each sample haplotype simultaneously with a series of windows with different sizes (Supplementary Fig. S12). It is obvious that at small windows it will tend to yield "common" calls; to a certain value, when the windows become larger, it will yield "population-unique" calls; as the window size continues to increase, it will eventually yield "other" calls (Fig. 3, Supplementary Fig. S13). At any given SNP position, when a larger window contains completely a smaller window, a unidirectional transition from "common" to "population-unique" to "others" can be observed when the ancestry is called sequentially from the smaller windows gradually to the larger windows. From this unidirectional transition, aMAP takes those "population-unique" messages before the appearance of the "others" calls in the larger windows as the ancestral calls of each locus (horizontal integration); when it is transformed directly from "common" to "other", aMAP records it as "common"; when the "other" calls are received from all parallel windows horizontally, aMAP records it as "other" (Supplementary Table S18).

None of those single window scans could retrieve the ancestral background sufficiently and accurately (Supplementary Table S18); thus, the parallel scan with a series of windows at different sizes and the horizontal integration is necessary to report the ancestry at different locus. Although we selected the window series of 20-snp, 30-snp, ……, 190-snp, and 200-snp) as the default setting of the window size in the current version of aMAP, it allows the users to choose their own window size.

Vertical data integration. Next aMAP vertically integrates ancestral calls from adjacent windows along the chromosomal haplotypes and absorbs those "common" calls if possible (vertical integration) (Supplementary Fig. S14). For example, when an Common_AC segment is recorded between two population_A_unique segments along a sample haplotype, aMAP revises the ancestry call of this AC_shared segment to population_A because this segment is shared between the population A and the population C and may be inherited from the common ancestors of populations A and C.

Border refinement. At last, aMAP zooms into these common blocks between two ancestral segments and finalizes the borders (Supplementary Fig. S15). When a "common" segment is called between two different population calls vertically, aMAP does a border refinement by zooming into the common segment using a series of overlapping windows with one SNP as a moving step. The scans start from the two margins between the common segment and the population-ancestral segments.

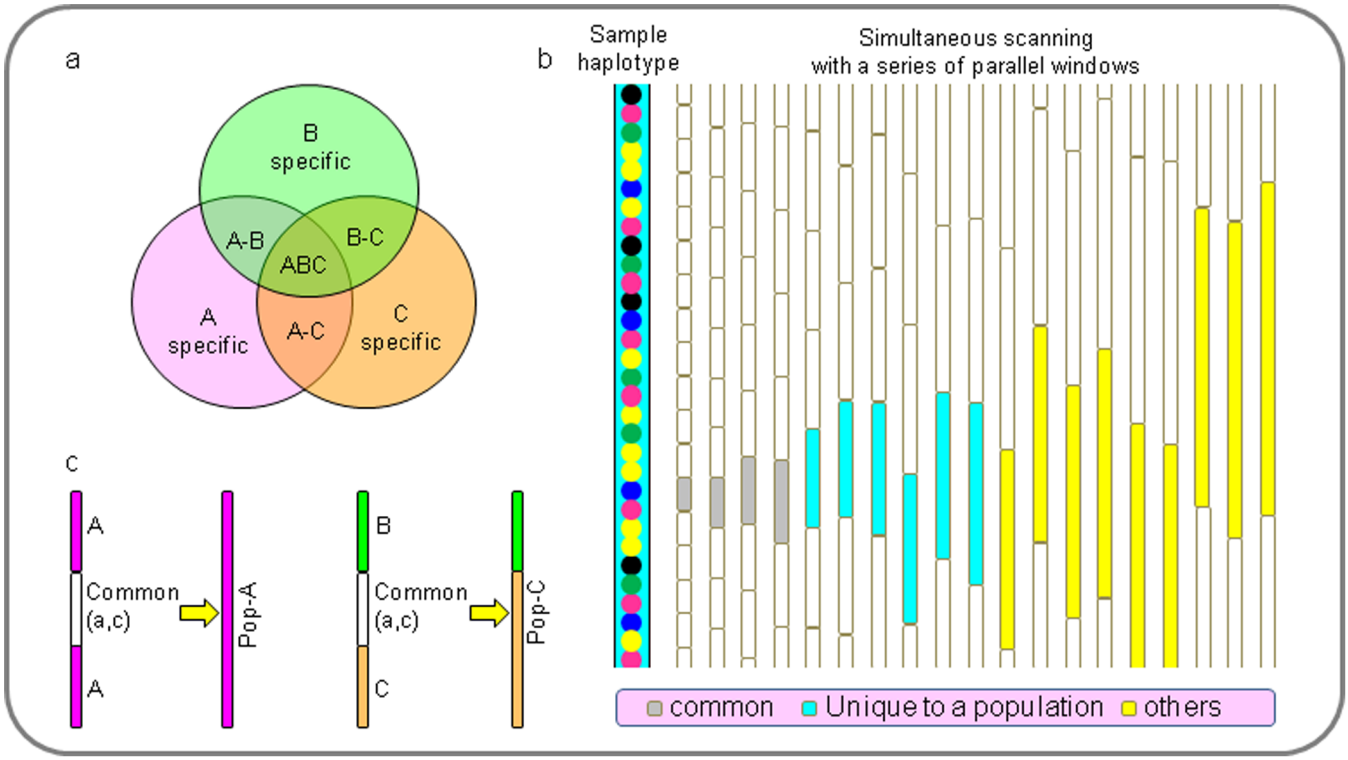

Figure 3 | The aMAP algorithm. (a) Reference analysis and pretreatment prior to use. (b) Parallel window scan using a set of parallel sliding windows and horizontal data integration. (c) Vertical data integration that integrates local ancestral calls from adjacent windows. Common calls are indicated by brackets.

The input and output of the aMAP software. The aMAP requires only two documents as the input; one is the sample haplotype file, the other one is the file with the sequences of the reference haplotypes. The output is a digital file that documents the ancestry of each segment along chromosomes.

Data downloaded from HapMap. All haplotype data were downloaded from HapMap, CEU (CEPH, U.S. Utah residents with ancestry from northern and western Europe), YRI (Yoruba in Ibadan, Nigeria), CHB(Han Chinese in Beijing, China), CHD (Chinese in Metropolitan Denver, Colorado), JPT (Japanese in Tokyo, Japan), MEX (Mexican ancestry in Los Angeles), TSI (Toscans in Italy), LWK (Luhya in Webuye, Kenya), MKK (Maasai in Kinyawa, Kenya), and GIH (Gujarati Indians in Houston) (Supplementary Table S19). The entire genome contains 1,437,974 SNPs on 23 chromosomes. CHB and CHD were combined throughout this study. These haplotypes were used to assess the performance of aMAP as sample haplotypes and as reference haplotypes, as well as original haplotypes to create simulated datasets. When a haplotype is used as a sample haplotype or to create a simulated dataset, this haplotype was always temporarily removed from the corresponding reference panel during the analysis.

Simulated datasets. *YRI-CHBCHD simulated dataset.* We first ran aMAP on the chromosome-1 YRI and CHBCHD using YRI and CHBCHD as the reference panels. Then we selected top 6 relatively pure haplotypes from YRI and 6 relatively pure haplotypes from CHBCHD, and simulated YRI-CHBCHD admixed individuals (50%–50%) by random mating between these original haplotypes. Six recombinations were introduced in each mating. We totally simulated 19 generations of offspring.

JPT-CHBCHD simulated dataset. We used the same approach to simulate CHBCHD-JPT admixed individuals (50%–50%) from 6 CHBCHD haplotypes and 6 JPT haplotypes.

6-way admixed dataset (CEU-YRI-CHBCHD-GIH-LWK-MKK). We used the same approach to simulate 6-way admixed individuals, including 12.5% CEU, 12.5% GIH, 25% CHBCHD, 12.5% YRI, 25% MKK and 12.5% LWK.

Measurement of accuracy.

$$Accuracy = \frac{\text{The number of SNPs with correct ancestry calls}}{\text{The number of all SNPs}}$$

1. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459–463 (2010).
2. Seldin, M. F., Pasaniuc, B. & Price, A. L. New approaches to disease mapping in admixed populations. *Nat Rev Genet* **12**, 523–528 (2011).
3. Kidd, J. M. *et al.* Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *Am J Hum Genet* **91**, 660–671 (2012).
4. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
5. Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**, 1359–1367 (2012).
6. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet* **93**, 278–288 (2013).
7. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**, e1000519 (2009).
8. Sundquist, A., Fratkin, E., Do, C. B. & Batzoglou, S. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res* **18**, 676–682 (2008).
9. Tang, H., Coram, M., Wang, P., Zhu, X. & Risch, N. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* **79**, 1–12 (2006).
10. Yang, J. J., Li, J., Buu, A. & Keoki Williams, L. Efficient inference of local ancestry. *Bioinformatics* **29**, 2750–2756 (2013).
11. Pasaniuc, B., Sankararaman, S., Kimmel, G. & Halperin, E. Inference of locus-specific ancestry in closely related populations. *Bioinformatics* **25**, i213–221 (2009).
12. Brisbin, A. *et al.* PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol* **84**, 343–364 (2012).
13. Fu, G., Sabnis, A. & Harrison, R. W. A deterministic-stochastic crossover algorithm for simulation of complex biochemical systems. *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference* 1–7 (2013).
14. Brown, R. & Pasaniuc, B. Enhanced methods for local ancestry assignment in sequenced admixed individuals. *PLoS Comput Biol* **10**, e1003555 (2014).
15. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & Francois, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973–983 (2014).
16. Lao, O., Liu, F., Wollstein, A. & Kayser, M. GAGA: a new algorithm for genomic inference of geographic ancestry reveals fine level population substructure in Europeans. *PLoS Comput Biol* **10**, e1003480 (2014).
17. Hu, Y., Willer, C., Zhan, X., Kang, H. M. & Abecasis, G. R. Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. *Am J Hum Genet* **93**, 891–899 (2013).

18. Elhaik, E. *et al*. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun* **5**, 3513 (2014).

Acknowledgments

This work was supported by NIH grants (HG006173, MD005964, HL095098, RR003034), American Heart Association grant (09GRNT2300003), RCMI Infrastructure for Clinical and Translational Research (U54MD07588), Baltimore Research Enhancement Award Program in Stroke, and the Baltimore Geriatrics Research, Education, and Clinical Center of the Department of Veterans Affairs.

Author contributions

Y.M., J.Z. and Q.S. developed the algorithm. J.Z. coded the program. Y.M., J.Z., J.W., L.M., W.L., G.F., W.X., K.Z., R.A.K., Y.L., Q.S. performed the data analysis. Y.M. and Q.S. wrote the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Ma, Y.M. *et al*. Accurate Inference of Local Phased Ancestry of Modern Admixed Populations. *Sci. Rep.* **4**, 5800; DOI:10.1038/srep05800 (2014).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/