ADDRESSING ANCESTRY DISPARITIES IN GENOMIC MEDICINE: A GEOGRAPHIC-AWARE ALGORITHM

Anonymous authors
Paper under double-blind review

ABSTRACT

With declining sequencing costs a promising and affordable tool is emerging in cancer diagnostics: genomics [1]. By using association studies, genomic variants that predispose patients to specific cancers can be identified, while by using tumor genomics cancer types can be characterized for targeted treatment. However, a severe disparity is rapidly emerging in this new area of precision cancer diagnosis and treatment planning, one which separates a few genetically well-characterized populations (predominantly European) from all other global populations. Here we discuss the problem of population-specific genetic associations, which is driving this disparity, and present a novel solution–coordinate-based local ancestry–for helping to address it. We demonstrate our boosting-based method on whole genome data from divergent groups across Africa and in the process observe signals that may stem from the transcontinental Bantu-expansion.

1 INTRODUCTION

Cancer genomics depends upon the identification of variants that are associated with particular types of cancers. Because such variants are deleterious, they are not typically part of the ancient standing variation spread across all humans; instead they are more recent mutations specific to particular populations. Indeed, such variants are often present prominently only in particular ethnic groups due to genetic drift [2]. In addition, most associations are mapped not to causal variants, but to more common neighboring variants that are present on genotyping arrays. Since these neighboring variants are linked to the causal variant via correlation structures (linkage) that are specific to each population, the ancestry of the genomic segment in which the correlated variant is found becomes crucial. Indeed, as a result of linkage and epistatic effects, genomic variants that are associated with cancer in one ancestry maybe have no association [3], or may even have an opposite association [4], in another ancestry. This phenomenon persists even in admixed individuals possessing multiple ancestries, such as African Americans; in such individuals the ancestry (European or African) of the specific genomic fragment containing the associated variant has been found to reverse the association [5]. This phenomenon dubbed ”flip-flop,” is not an unusual case, rather ancestry-specific effects in genetic association studies are the rule. For this reason, polygenic-risk scores (PRS), increasingly important to genomic cancer prediction [6], have been found to be several times less accurate when used on populations of different ancestry from the one on which they were trained [7].

As a result of these ancestry specific effects, accurately identifying the ancestry of each segment of the genome is becoming increasingly crucial for genomic medicine. Such algorithms, known as local ancestry inference, have been developed both for historical population genetics [8–15] and for recreational consumer ancestry products [16], but none have been developed to date for the particular demands of clinical genomic medicine. Such an algorithm would need to provide ancestry not as a culturally defined label, but as continuous genetic coordinates that could be used as a covariate in predication and association algorithms. This method is also important for deconvolving ancestry effects in genetic association studies. To date, most genome-wide association studies (GWAS) are conducted in populations of single ancestry (typically European) to avoid confounding effects of ancestry on reversing associations. Researchers often avoid admixed populations, for instance African Americans or Hispanics, who encompass more than one ancestry, and avoid populations with too much genetic variation or too many diverse sub-populations, as is common within Africa. This has resulted in over 80% of the individuals in GWAS studies to date stemming from European ancestry (and only 2% from African ancestry) [17, 18]. A reliable coordinate-based local ancestry algorithm would allow such studies to embrace diversity, rather than intentionally eschewing it, by allowing an additional covariate along the genome to be used (ancestry) to remove the confounding effects of ancestry-dependent genomic associations. With such a tool, medical researchers would no longer need to avoid admixed and globally diverse genetic study cohorts.

2 Ancestry Inference

Here were present an accurate coordinate-based local ancestry inference algorithm, XGMix, that can be used for addressing ancestry-specific associations and predictions. XGMix uses modern single ancestry reference populations to accurately predict the latitude and longitude of the closest modern source population for each segment of an individual's genome. These coordinate annotations along the genome can then be used as covariates for genome-wide association studies (GWAS) and for polygenic risk score (PRS) predictions.

Estimation of an individual's ancestry, both globally and locally (i.e. assigning an ancestry estimate to each region of the chromosomal sequence), has been tackled with a wide range of methods and technologies [8–15]. Local ancestry inference has traditionally been framed as a classification problem using pre-defined ancestries. Classification approaches provide discrete ancestry labels but can be highly inaccurate for neighboring populations (or population gradients) and intractable for genetically diverse populations with multiple sources. Geographical regression along the genome, although a much more challenging problem, could provide a continuous representation of ancestry capable of capturing the complexities of worldwide populations.

XGMix consists of two layers of stacked gradient boosted trees (a genomic window-specific layer and a window aggregating smoother) and can infer local-ancestry with both classification probabilities and geographical coordinates along each phased chromosome. Here we demonstrate XGMix by training on whole genomes from real individuals from the five African populations included in the 1000 genomes project [19]. We simulate admixed individuals of various generations using Wright-Fisher simulation [13] to create ground truth labels of ancestry along the genome and split this data for training and testing. As these reference African populations lie close to a single arc along the globe we estimate along this arc, getting geographic assignments for each genomic segment.

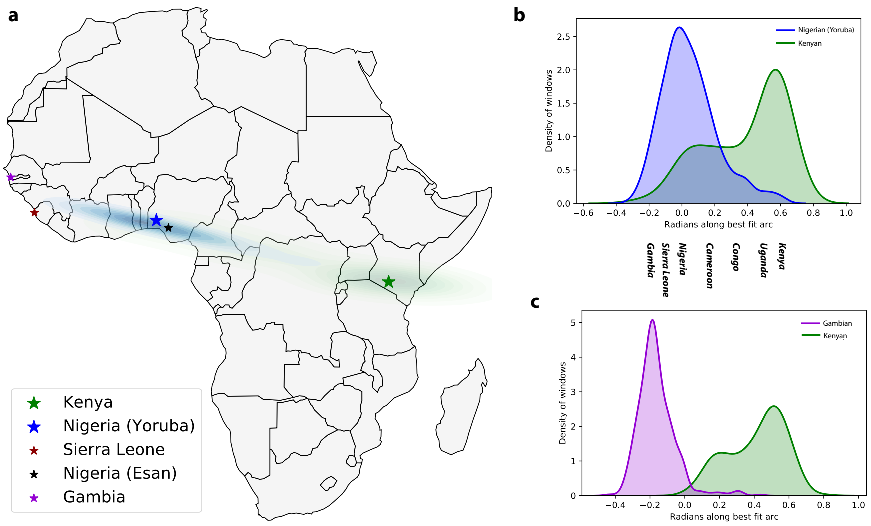

Figure 1: (a) The inferred coordinates for each genomic segment of an admixed Kenyan-Nigerian individual. The model was trained on all indicated African reference populations. (b-c) The inferred location of each genomic segment of a Kenyan-Nigerian (b) and Kenyan-Gambian (c) individual using the principal coordinate arc of the reference populations' locations. The bimodal distribution of Kenyan segments (green) may reflect the historical Bantu expansion from Cameroon into Kenya.

REFERENCES

[1] K. Schwarze, J. Buchanan, J. M. Fermont, H. Dreau, M. W. Tilley, J. M. Taylor, P. Antoniou, S. J. L. Knight, C. Camps, M. M. Pentony, E. M. Kvikstad, S. Harris, N. Popitsch, A. T. Pagnamenta, A. Schuh, J. C. Taylor, and S. Wordsworth, "The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom," *Genetics in Medicine*, vol. 22, no. 1, pp. 85–94, January 2020.

[2] W. D. Foulkes, I. Thiffault, S. B. Gruber, M. Horwitz, N. Hamel, C. Lee, J. Shia, A. Markowitz, A. Figer, E. Friedman, D. Farber, C. M. T. Greenwood, J. D. Bonner, K. Nafa, T. Walsh, V. Marcus, L. Tomsho, J. Gebert, F. A. Macrae, C. L. Gaff, B. B.-d. Paillerets, P. K. Gregersen, J. N. Weitzel, P. H. Gordon, E. MacNamara, M. C. King, H. Hampel, A. de la Chapelle, J. Boyd, K. Offit, G. Rennert, G. Chong, and N. A. Ellis, "The Founder Mutation MSH2*1906G–¿C Is an Important Cause of Hereditary Nonpolyposis Colorectal Cancer in the Ashkenazi Jewish Population," *The American Journal of Human Genetics*, vol. 71, no. 6, pp. 1395–1412, Dec. 2002.

[3] S. Wang, F. Qian, Y. Zheng, T. Ogundiran, O. Ojengbede, W. Zheng, W. Blot, K. L. Nathanson, A. Hennis, B. Nemesure, S. Ambs, O. I. Olopade, and D. Huo, "Genetic variants demonstrating flip-flop phenomenon and breast cancer risk prediction among women of African ancestry," *Breast Cancer Research and Treatment*, vol. 168, no. 3, pp. 703–712, Apr. 2018.

[4] F. Rajabli, B. E. Feliciano, K. Celis, K. L. Hamilton-Nelson, P. L. Whitehead, L. D. Adams, P. L. Bussies, C. P. Manrique, A. Rodriguez, V. Rodriguez, T. Starks, G. E. Byfield, C. B. S. Lopez, J. L. McCauley, H. Acosta, A. Chinea, B. W. Kunkle, C. Reitz, L. A. Farrer, G. D. Schellenberg, B. N. Vardarajan, J. M. Vance, M. L. Cuccaro, E. R. Martin, J. L. Haines, G. S. Byrd, G. W. Beecham, and M. A. Pericak-Vance, "Ancestral origin of ApoE $\varepsilon4$ Alzheimer disease risk in Puerto Rican and African American populations," *PLoS Genetics*, vol. 14, no. 12, pp. e1007791, December 2018.

[5] F. Rajabli et al., "Ancestral origin of ApoE $\varepsilon4$ Alzheimer disease risk in Puerto Rican and African American populations," *PLoS Genetics*, vol. 14, no. 12, pp. e1007791, December 2018.

[6] N. Mavaddat, K. Michailidou, J. Dennis, M. Lush, L. Fachal, A. Lee, J. P. Tyrer, T.-H. Chen, Q. Wang, M. K. Bolla, X. Yang, M. A. Adank, T. Ahearn, K. Aittomäki, J. Allen, I. L. Andrulis, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, P. L. Auer, P. Auvinen, M. Barrdahl, L. E. Beane Freeman, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, L. Bernstein, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, A.-L. Børresen-Dale, H. Brauch, M. Bremer, H. Brenner, A. Brentnall, I. W. Brock, A. Brooks-Wilson, S. Y. Brucker, T. Brüning, B. Burwinkel, D. Campa, B. D. Carter, J. E. Castelao, S. J. Chanock, R. Chlebowski, H. Christiansen, C. L. Clarke, J. M. Collée, E. Cordina-Duverger, S. Cornelissen, F. J. Couch, A. Cox, S. S. Cross, K. Czene, M. B. Daly, P. Devilee, T. Dörk, I. dos Santos-Silva, M. Dumont, L. Durcan, M. Dwek, D. M. Eccles, A. B. Ekici, A. H. Eliassen, C. Ellberg, C. Engel, M. Eriksson, D. G. Evans, P. A. Fasching, J. Figueroa, O. Fletcher, H. Flyger, A. Försti, L. Fritschi, M. Gabrielson, M. Gago-Dominguez, S. M. Gapstur, J. A. García-Sáenz, M. M. Gaudet, V. Georgoulias, G. G. Giles, I. R. Gilyazova, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. González-Neira, G. I. Grenaker Alnæs, M. Grip, J. Gronwald, A. Grundy, P. Guénel, L. Haeberle, E. Hahnen, C. A. Haiman, N. Håkansson, U. Hamann, S. E. Hankinson, E. F. Harkness, S. N. Hart, W. He, A. Hein, J. Heyworth, P. Hillemanns, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, A. Howell, G. Huang, K. Humphreys, D. J. Hunter, M. Jakimovska, A. Jakubowska, W. Janni, E. M. John, N. Johnson, M. E. Jones, A. Jukkola-Vuorinen, A. Jung, R. Kaaks, K. Kaczmarek, V. Kataja, R. Keeman, M. J. Kerin, E. Khusnutdinova, J. I. Kiiski, J. A. Knight, Y.-D. Ko, V.-M. Kosma, S. Koutros, V. N. Kristensen, U. Krüger, T. Kühl, D. Lambrechts, L. Le Marchand, E. Lee, F. Lejbkowicz, J. Lilyquist, A. Lindblom, S. Lindström, J. Lissowska, W.-Y. Lo, S. Loibl, J. Long, J. Lubiński, M. P. Lux, R. J. MacInnis, T. Maishman, E. Makalic, I. Maleva Kostovska, A. Mannermaa, S. Manoukian, S. Margolin, J. W. M. Martens, M. E. Martinez, D. Mavroudis, C. McLean, A. Meindl, U. Menon, P. Middha, N. Miller, F. Moreno, A. M. Mulligan, C. Mulot, V. M. Muñoz-Garzon, S. L. Neuhausen, H. Nevanlinna, P. Neven, W. G. Newman, S. F. Nielsen, B. G. Nordestgaard, A. Norman, K. Offit, J. E. Olson,

H. Olsson, N. Orr, V. S. Pankratz, T.-W. Park-Simon, J. I. A. Perez, C. Pérez-Barrios, P. Peterlongo, J. Peto, M. Pinchev, D. Plaseska-Karanfilska, E. C. Polley, R. Prentice, N. Presneau, D. Prokofyeva, K. Purrington, K. Pylkäs, B. Rack, P. Radice, R. Rau-Murthy, G. Rennert, H. S. Rennert, V. Rhenius, M. Robson, A. Romero, K. J. Ruddy, M. Ruebner, E. Saloustros, D. P. Sandler, E. J. Sawyer, D. F. Schmidt, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, F. Schumacher, P. Schürmann, L. Schwentner, C. Scott, R. J. Scott, C. Seynaeve, M. Shah, M. E. Sherman, M. J. Shrubsole, X.-O. Shu, S. Slager, A. Smeets, C. Sohn, P. Soucy, M. C. Southey, J. J. Spinelli, C. Stegmaier, J. Stone, A. J. Swerdlow, R. M. Tamimi, W. J. Tapper, J. A. Taylor, M. B. Terry, K. Thöne, R. A. E. M. Tollenaar, I. Tomlinson, T. Truong, M. Tzardi, H.-U. Ulmer, M. Untch, C. M. Vachon, E. M. van Veen, J. Vijai, C. R. Weinberg, C. Wendt, A. S. Whittemore, H. Wildiers, W. Willett, R. Winqvist, A. Wolk, X. R. Yang, D. Yannoukakos, Y. Zhang, W. Zheng, A. Ziogas, A. M. Dunning, D. J. Thompson, G. Chenevix-Trench, J. Chang-Claude, M. K. Schmidt, P. Hall, R. L. Milne, P. D. P. Pharoah, A. C. Antoniou, N. Chatterjee, P. Kraft, M. García-Closas, J. Simard, and D. F. Easton, "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes," *American journal of human genetics*, vol. 104, no. 1, pp. 21–34, Jan. 2019.

[7] A. R. Martin et al., "Clinical use of current polygenic risk scores may exacerbate health disparities," *Nature Genetics*, vol. 51, no. 4, pp. 584–591, April 2019.

[8] H. Tang, M. Coram, P. Wang, X. Zhu, , and N. Risch, "Reconstructing genetic ancestry blocks in admixed individuals," *The American Journal of Human Genetics*, vol. 79, pp. 1–12, May 2006.

[9] A. Sundquist, E. Fratkin, C. B. Do, and S. Batzoglou, "Effect of genetic divergence in identifying ancestral origin using HAPAA," *Genome research*, vol. 18, pp. 676–682, April 2008.

[10] A. L. Price et al., "Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations," *PLoS Genetics*, vol. 5, no. 6, pp. 1–18, June 2009.

[11] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin, "Estimating local ancestry in admixed populations," *The American Journal of Human Genetics*, vol. 82, no. 2, pp. 290–303, February 2008.

[12] E. Y. Durand, C. B. Do, J. L. Mountain, and J. M. Macpherson, "Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution," *bioRxiv*, October 2014.

[13] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante, "RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference," *The American Journal of Human Genetics*, vol. 93, no. 2, pp. 278–288, August 2013.

[14] D. Mas Montserrat, C. Bustamante, and A. Ioannidis, "Class-Conditional VAE-GAN for Local-Ancestry Simulation," *Machine Learning in Computational Biology*, December 2019, Vancouver, Canada.

[15] D. Mas Montserrat, C. Bustamante, and A. Ioannidis, "LAI-Net: Local-Ancestry Inference With Neural Networks," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2020, Barcelona, Spain.

[16] K. Bryc, E. Y. Durand, J. M. Macpherson, D. Reich, and J. L. Mountain, "The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States," *American journal of human genetics*, vol. 96, no. 1, pp. 37–53, January 2015.

[17] "The Missing Diversity in Human Genetic Studies," *Cell*, vol. 177, no. 1, pp. 26–31, Mar. 2019.

[18] A. B. Popejoy and S. M. Fullerton, "Genomics is failing on diversity," *Nature News*, vol. 538, no. 7624, pp. 161–164, October 2016.

[19] 1000 Genomes Project Consortium and others, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68, 2015.