RESEARCH NOTE

H3Africa: crucial importance of knowledge on human demographic history in strategies for data exploitation – an analysis of the *Luhya in Webuye, Kenya* population from the 1000 Genomes Project [version 1; referees: 2 approved with reservations]

Benard W. Kulohoma

Centre for Biotechnology and Bioinformatics, University of Nairobi, Nairobi, Kenya

v1 **First published:** 06 Jul 2018, **3**:82 (doi: 10.12688/wellcomeopenres.14692.1)
Latest published: 18 Sep 2018, **3**:82 (doi: 10.12688/wellcomeopenres.14692.2)

Abstract

Paucity of data from African populations has restricted understanding of the heritable human genome variation. Although under-represented in human genetic studies, Africa has sizeable genetic, cultural and linguistic diversity. The Human Heredity and Health in Africa (H3Africa) initiative is aimed at understanding health problems relevant to African populations, and titling the scales of data deficit and lacking expertise in health-related genomics among African scientists. We emphasise that careful consideration of the sampled populations in the H3Africa projects is required to maximise the prospects of identifying and fine-mapping novel risk variants in indigenous populations. H3Africa which considers national and within-continental cohorts must have well thought out documented protocols that carefully consider human demographic history.

Keywords

Africa, GWAS, Population substructure, H3Africa

Open Peer Review

Referee Status: ? ?

	Invited Referees 1	2
REVISED **version 2** published 18 Sep 2018		
version 1 published 06 Jul 2018	? report	? report

1 **Michèle Ramsay**, University of the Witwatersrand, South Africa

2 **Nicola J. Mulder**, University of Cape Town, South Africa

Discuss this article

Comments (0)

Corresponding author: Benard W. Kulohoma (bkulohoma@uonbi.ac.ke)

Author roles: Kulohoma BW: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The work was supported by the Wellcome Trust [087540] through a pump-priming grant to BWK from the Training Health Researchers into Vocational Excellence in East Africa (THRiVE) Initiative.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Kulohoma BW. **H3Africa: crucial importance of knowledge on human demographic history in strategies for data exploitation – an analysis of the *Luhya in Webuye, Kenya* population from the 1000 Genomes Project [version 1; referees: 2 approved with reservations]** Wellcome Open Research 2018, **3**:82 (doi: 10.12688/wellcomeopenres.14692.1)

First published: 06 Jul 2018, **3**:82 (doi: 10.12688/wellcomeopenres.14692.1)

Introduction

The 1000 Genomes Project (1000GP) is an invaluable resource that has improved understanding of global human genetic variation and its contribution to disease biology across multiple populations of distinct ethnicity[1]. This catalogue of over 88 million high-quality variants from 26 populations has enhanced power to screen for common and rare variants that depict geographic and demographic differentiation[2]. This represents 80% (80 million) of all variants contributed or validated in the public dbSNP catalogue, with recent major enhancements for genetic variation within several South Asian and African populations (24% and 28% of novel variants respectively)[2]. Most of the low-frequency (< 0.5%) variants likely to be of functional significance are disproportionately present in individuals with substantial African ancestry, indicating bottlenecks in non-African populations[2,3]. The *"Luhya in Webuye, Kenya"* (LWK) population has the most accentuated number of these rare variants.

Paucity of data from African populations has restricted understanding of the heritable human genome variation. Although under-represented in human genetic studies, Africa has sizeable genetic, cultural and linguistic diversity (> 2000 distinct ethno-linguistic groups)[4]. African populations are more genetically diverse, with considerable population substructure, and lower linkage disequilibrium (LD) compared to non-African populations[4,5]. Inclusion of more African populations will improve understanding of genetic variation attributed to complex population history, variations in climate, lifestyles, exposure to infectious diseases, and diets[4,6]. Diverse multi-ethnic imputation panels will undoubtedly improve fine-mapping of complex traits and provide detailed insights on disease susceptibility, drug responses, and improve therapeutic treatments. One such integrated panel, consisting of the phase 1 1000GP and African Genome Variation Project (AGVP) whole genome sequence panels, has shown marked improvement in detecting association signals in specific African populations poorly represented in the 1000GP[7]. AGVP also present a new genotype array design that captures genetic variation in African populations.

The Human Heredity and Health in Africa (H3Africa) initiative is aimed at understanding health problems relevant to African populations, and titling the scales of data deficit and lacking expertise in health-related genomics among African scientists[8,9]. The H3Africa consortium consists of over 500 members, from more than 30 of the 55 African countries. H3Africa projects are focused on establishing genetic and environmental determinants associated with infectious (human African trypanosomiasis, tuberculosis, HIV, and other respiratory tract infections) and non-communicable diseases (kidney disease, diabetes, and cardiovascular diseases)[10]. H3Africa is driven by African investigators, and is anticipated to close the gaps of 'missing' heritability by increasing the number of causal variants identified within genes, from a dataset of over 70,000 individuals collected using standardized protocols[8,10]. This presents a unique opportunity for the investigators to not only develop and direct their independent research agendas, but also enrich the datasets using their extensive knowledge of the continent's history. However, careful consideration of the sampled populations in the H3Africa projects is required to maximise the prospects of identifying and fine-mapping novel risk variants in indigenous populations. In order to translate genomic research findings to useful resources for clinicians and drug development, substantial knowledge about reference populations that are relevant to the individuals being treated alongside the actionable variants is required[10]. This is in addition to harmonised and well curated phenotype data that will allow easy integration and direct comparison of data outputs across different cohorts and phenotypes. Attentiveness to the considerable genetic substructure in African population may reveal uncaptured variation and distinct ancestry[11]. This extensive genetic diversity would benefit from strategies that explore genomics datasets that put local populations in context to provide more detail from disease mapping efforts in Africa. An example is the LWK in the 1000GP who do not represent all the *"Luhya people"*, a Bantu-speaking Niger-Congo population with a complex population history composed of 17 tribes, each with a distinct dialect (Figure 1)[12]. We examined for possible substructure in LWK, from 1000GP, to establish its implication on association studies.

Methods

We used principal component analysis (PCA) to examine relationships within the LWK population (n=99) using 193,634 variants from the 1000GP phase 3[2]. The 1000GP call set was already filtered down using VCFtools and PLINK, and only contained biallelic, non-singleton SNV sites that are a minimum of 2KB apart from each other and a minor allele frequency > 0.05[2,13,14]. We considered just the first three principal components (PCs) computed to resolve the population substructure. We then used ADMIXTURE v1.3 to estimate ancestry for K values from 3 through 20[15]. Distruct plots of the output ancestry fractions were generated using STRUCTURE PLOT v2.0[16].

Results and discussion

Our PCA analyses reveal that all individuals cluster closely except five individuals along PC1 (n=2) and PC2 (n=3), possibly suggesting that the outliers are individuals from different Luhya tribes. We suggest that whereas the first two principal components, PC1 and PC2, distinguished individuals primarily on genetic ancestry, PC3 reflects the geographic distribution of the individuals (Figure 2). We propose that although a huge proportion of individuals sampled are actually from Webuye (Bukusu tribe), others hail from various settlements along major routes and smaller towns (Figure 1E and 1F). Unsupervised ADMIXTURE analysis suggests minimal substructure, and the cross-validation procedure identified K3 as the most plausible K (Figure 3).

GWAS studies largely rely on self-reported data on ethnic background. Genetic information is then used to confirm ancestral backgrounds and exclude outliers. Thus, in order to understand complex traits in say the entire "*Luhya people*", adequate sampling of underrepresented tribes would provide a high-resolution view of their ancestral history. Haphazard sampling would significantly reduce power to detect signal due to

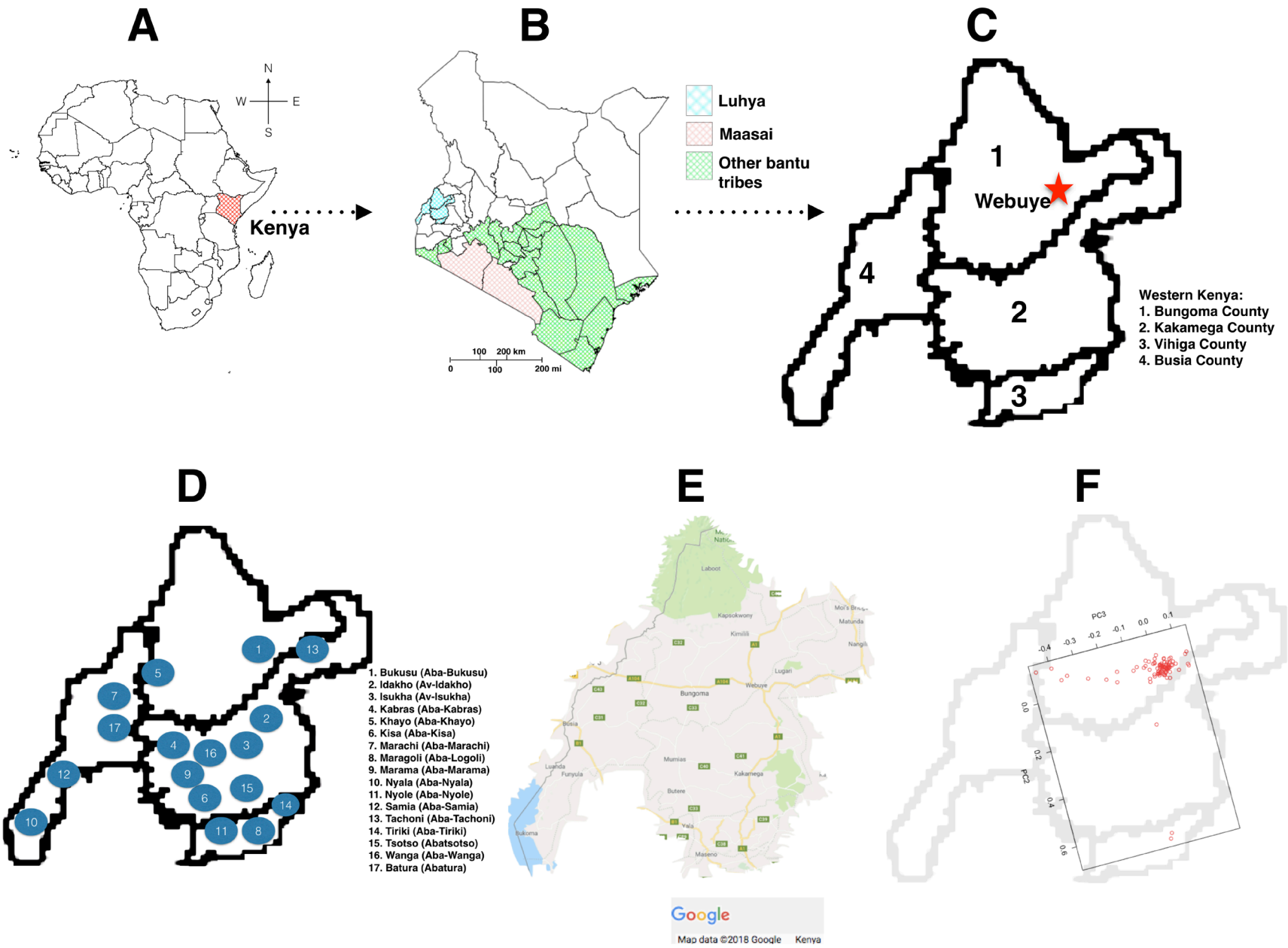

Figure 1. The geographic distribution of *Luhya in Webuye*, Kenya individuals. (**A**) Map of Africa showing the location of Kenya. (**B**) Map of Kenya showing the location of Western Kenya Counties. (**C**) Counties in Western Kenya inhabited by the "*Luhya people*". (**D**) The 17 tribes of the Luhya, and the locations they hail from in Western Kenya. (**E**) Map showing major routes in Western Kenya. Most settlements are along major routes, and in small towns at their intersections. (**F**) PC3 reflects the geographic distribution of the individuals, while the first two principal components, PC1 and PC2, distinguished individuals primarily on genetic ancestry.

Wellcome Open Research 2018, 3:82 Last updated: 18 SEP 2018

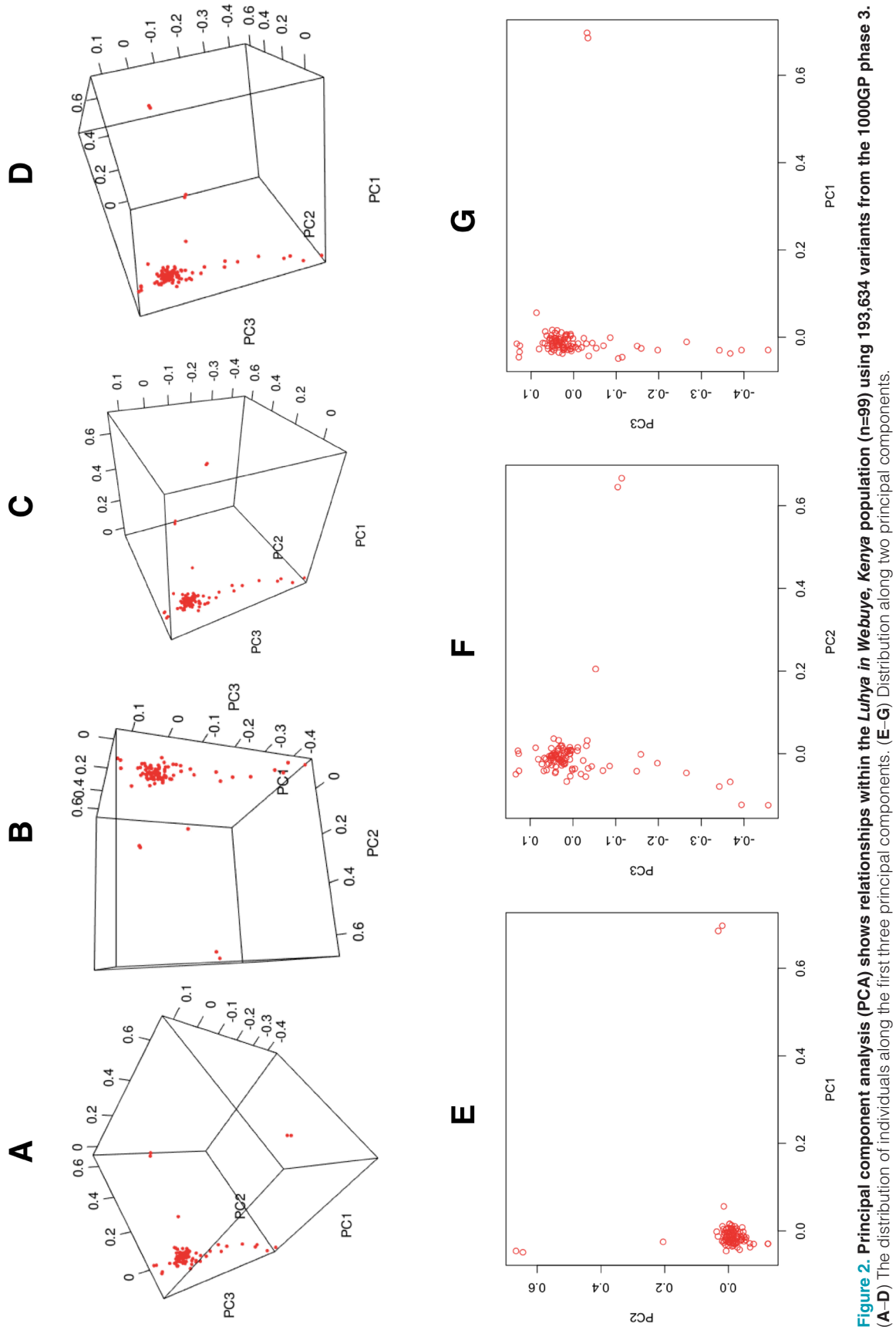

Figure 2. Principal component analysis (PCA) shows relationships within the *Luhya in Webuye, Kenya* population (n=99) using 193,634 variants from the 1000GP phase 3. (**A**–**D**) The distribution of individuals along the first three principal components. (**E**–**G**) Distribution along two principal components.

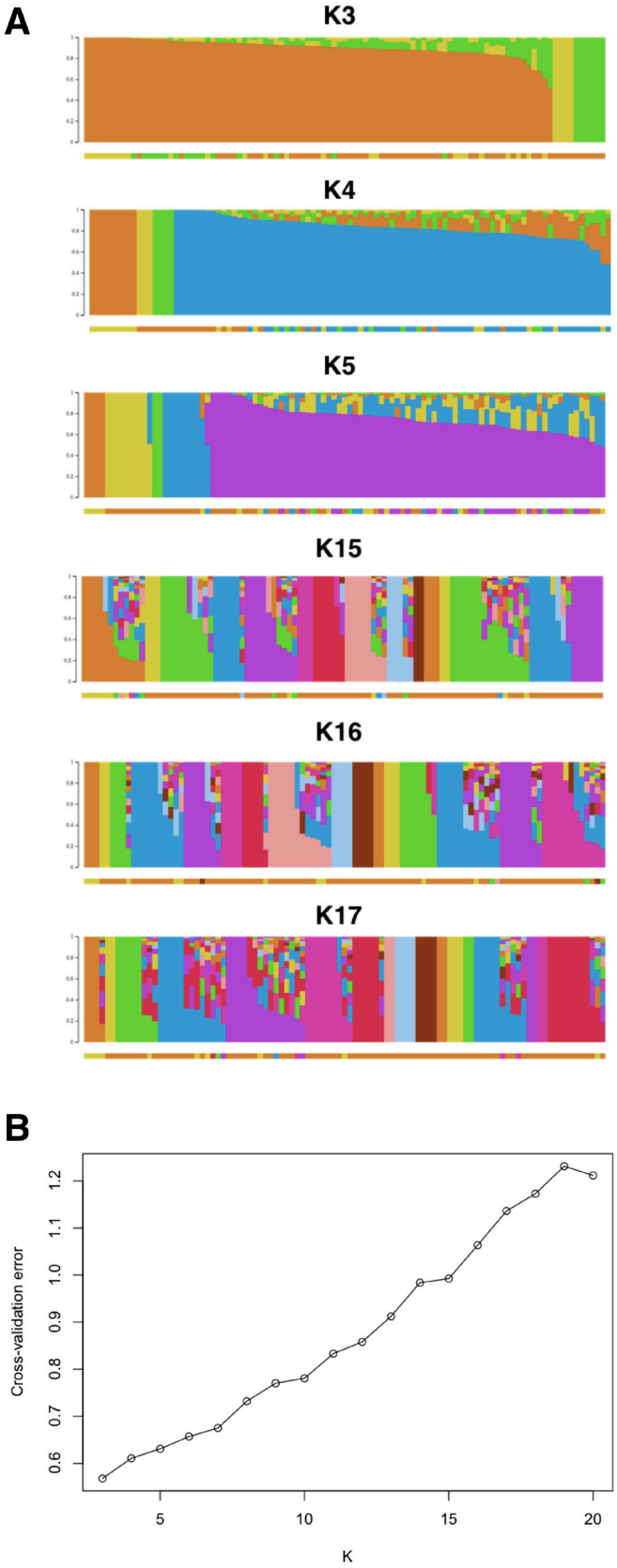

Figure 3. Population structure within the *Luhya in Webuye, Kenya*. (**A**). We show the first and last three K values for K=3 through K=20. (**B**). The cross-validation procedure identified K=3 as the most plausible K since it had the smallest error value.

population substructure, even within this single community. We speculate that this was largely circumvented at recruitment when sampling LWK by asking the participants whether all four of their grandparents were of the Bukusu tribe. Whereas projects covering relatively small geographical areas are able to overcome such challenges, national and within-continental cohorts in efforts like H3Africa must have well thought out documented protocols that carefully consider human demographic history.

Data availability

The LWK dataset was obtained from the European Bioinformatics Institute 1000 Genomes Project website http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/admixture_files/

Competing interests

No competing interests were disclosed.

Grant information

The work was supported by the Wellcome Trust [087540] through a pump-priming grant to BWK from the Training Health Researchers into Vocational Excellence in East Africa (THRiVE) Initiative.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgement

This work is published with permission from the University of Nairobi.

References

1. Sudmant PH, Rausch T, Gardner EJ, *et al.*: **An integrated map of structural variation in 2,504 human genomes.** *Nature.* 2015; **526**(7571): 75–81.
PubMed Abstract | Publisher Full Text | Free Full Text
2. 1000 Genomes Project Consortium; Auton A, Brooks LD, *et al.*: **A global reference for human genetic variation.** *Nature.* 2015; **526**(7571): 68–74.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Marth G, Schuler G, Yeh R, *et al.*: **Sequence variations in the public human genome data reflect a bottlenecked population history.** *Proc Natl Acad Sci U S A.* 2003; **100**(1): 376–381.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Campbell MC, Tishkoff SA: **African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping.** *Annu Rev Genomics Hum Genet.* 2008; **9**: 403–433.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Tishkoff SA, Dietzsch E, Speed W, *et al.*: **Global patterns of linkage disequilibrium at the CD4 locus and modern human origins.** *Science.* 1996; **271**(5254): 1380–1387.
PubMed Abstract | Publisher Full Text
6. Gomez F, Hirbo J, Tishkoff SA: **Genetic variation and adaptation in Africa: implications for human evolution and disease.** *Cold Spring Harb Perspect Biol.* 2014; **6**(7): a008524.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Gurdasani D, Carstensen T, Tekola-Ayele F, *et al.*: **The African Genome Variation Project shapes medical genetics in Africa.** *Nature.* 2015; **517**(7534): 327–332.
PubMed Abstract | Publisher Full Text | Free Full Text
8. H3Africa Consortium; Rotimi C, Abayomi A, *et al.*: **Research capacity. Enabling the genomic revolution in Africa.** *Science.* 2014; **344**(6190): 1346–1348.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Mulder NJ, Adebiyi E, Alami R, *et al.*: **H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa.** *Genome Res.* 2016; **26**(2): 271–277.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Mulder N, Abimiku A, Adebamowo SN, *et al.*: **H3Africa: current perspectives.** *Pharmgenomics Pers Med.* 2018; **11**: 59–66.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Retshabile G, Mlotshwa BC, Williams L, *et al.*: **Whole-Exome Sequencing Reveals Uncaptured Variation and Distinct Ancestry in the Southern African Population of Botswana.** *Am J Hum Genet.* 2018; **102**(5): 731–743.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Coriell Institute: **Luhya in Webuye, Kenya [LWK].** 2018; **2018**.
Reference Source
13. Danecek P, Auton A, Abecasis G, *et al.*: **The variant call format and VCFtools.** *Bioinformatics.* 2011; **27**(15): 2156–2158.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Chang CC, Chow CC, Tellier LC, *et al.*: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *GigaScience.* 2015; **4**: 7.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Alexander DH, Lange K: **Enhancements to the ADMIXTURE algorithm for individual ancestry estimation.** *BMC Bioinformatics.* 2011; **12**: 246.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Ramasamy RK, Ramasamy S, Bindroo BB, *et al.*: **STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface.** *Springerplus.* 2014; **3**: 431.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ? ?

Version 1

Referee Report 06 August 2018

doi:10.21956/wellcomeopenres.15998.r33497

? **Nicola J. Mulder**
Computational Biology Group, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

This paper uses the LWK population from the 1000 Genomes dataset to demonstrate the need for considering demographic data when sampling. The paper is very thin on results for a research paper as it only really includes a PCA on a small sample set. In order for it to be a reasonable research paper it would need to include other populations and data to verify the findings. It is also not clear to me whether there is support for the ethnic origin of the samples, is this information available?

Additional comments:

- The title refers to H3Africa but the paper is about the 1000 Genomes data so I recommend that the title changes, the results are relevant to all studies, not just H3Africa.
- In the abstract and later there is mention of "titling" was this supposed to "tilting"?
- There are too many figures to demonstrate a simple point.

Though the message of the paper is important, I think it needs more work. The PCAs and the paper in general need to include more populations to verify the point.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 10 Sep 2018

Benard Kulohoma,

We have made the following changes to this research note paper:

We have revised the analysis to include more populations in the analysis to verify our findings. We now also include anthropology and population genetics references on the history of the Bantu migration, and the Luhya population of Western Kenya. We have made revisions to the analysis, and now include other populations that share recent genetic ancestry from the 1000GP that belong to the Niger-Kordofanian Bantoid language group for comparison.

We have made revisions to the title to encompass all studies, not just H3Africa.

We have amended the word "titling" to "tilting". We have also revised the manuscript and reduced the number of figures.

We have made revisions to the analysis, and include other populations that share recent genetic ancestry from the 1000GP that belong to the Niger-Kordofanian language group for comparison.

Competing Interests: None

Referee Report 03 August 2018

doi:10.21956/wellcomeopenres.15998.r33495

Michèle Ramsay
Sydney Brenner Institute for Molecular Bioscience, Division of Human Genetics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

This study uses genotype data from the 1000 Genomes Project to examine population structure among the 99 individuals of the Luhya people sampled from Webuye in Kenya. It emphasizes the need for careful participant selection in disease-related genome-wide studies in African populations. Principal component analysis (PCA) and ADMIXTURE are used. I fully agree that we need to consider demographic histories when analyzing genomic data, but found the title misleading in two ways: Why focus on H3Africa as the target audience and not more generally research among Africans (and other populations) and, secondly, how does selection of this group of Luhya people demonstrate that the sampling was less than optimal. The only suggestion in the paper, related to the title, is to request data on the ethnolinguistic origins of the grandparents during the participant selection phase – I was expecting more insight into how to include knowledge of demographic history into a sampling framework (e.g. collaborate or consult with anthropologists, linguists and historians; select participants from rural areas that have less ethnolinguistic diversity and admixture; request ethnolinguistic identification of grandparent; etc.). Sampling frameworks

will of course be influenced by the aims and objectives of the research.

The study focused on a single population and no additional or neighboring populations were included in the PCA and ADMIXTURE analyses. It is therefore unclear what the three ancestral components (K=3) are likely to be and what the origins of the 5 outliers may be. The sampling appears to have been good in identifying a closely related group of individuals, which is not unexpected as they were from a single ethnolinguistic origin and geographic location. It would be very interesting to sample all 17 Luhya tribes to determine whether they are genetically distinguishable on PCA, but I appreciate that these data are not available. An understanding of their demographic history could provide some clues as to what one may expect. Perhaps refer to the recent paper on South African genomes by Choudhury et al.[1].

Currently there are good analysis programs that either rely on adjusting for ancestral component diversity or using meta-analysis when participants are from significantly different populations (e.g. GEMMA, METAL, BOLT-LMM and others). Since there is a vast amount of data on every participant in GWAS studies, PCA can effectively be used to identify outliers and to exclude them, if appropriate, or adjust for population sub-structure.

The paper would be strengthened by providing other examples and illustrating how some studies/populations have been successful at using demographic history in their sampling strategy whereas others fall short, as they have ignored this. The figures are numerous and do not all add additional value, no need to show structure plots at high K values. The structure plots would be enhanced by using a tool such as Genesis (www.bioinf.wits.ac.za/software/genesis) to ensure that the individuals are shown in the same order and that the same color is used throughout to represent the same ancestral component for the different K values.

There are several sentences that could be rewritten to improve meaning and direct repetition of whole sentences should be avoided between the abstract and paper content. Presumably the author meant “tilting the scales” and not “titling the scales”?

References

1. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, Chimusa ER, Christoffels A, Gamieldien J, Sefid-Dashti MJ, Joubert F, Meintjes A, Mulder N, Ramesar R, Rees J, Scholtz K, Sengupta D, Soodyall H, Venter P, Warnich L, Pepper MS: Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans.*Nat Commun*. 2017; **8** (1): 2062 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Genetics, genomics and population genetics.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 10 Sep 2018

Benard Kulohoma,

We have revised the title to make it more general to encompass all studies, not just H3Africa. We have made revisions to illustrate that these considerations should be made to all multi-ethnic GWAS studies.

We have amended the manuscript and discuss the importance of multi-disciplinary approaches that enlist the knowledge of anthropologists, linguists, geneticists, and historians to improve understanding on human history and migration of populations, genetics of complex traits and adaptive variations to modern environments, and language and cultural changes.

We have made revisions to the analysis, and include other populations that share recent genetic ancestry from the 1000GP that belong to the Niger-Kordofanian language group for comparison.

We agree that current analysis program that implement standard linear regression can use principal components from PCA analysis to avoid errors due to population stratification, as well as speed-up computation. Essentially, our manuscript is aimed at highlighting population stratification; and the importance of multi-disciplinary approaches when sampling underrepresented populations of interest to put local populations in the right context and provide more detailed and accurate information for disease mapping efforts in Africa. This will provide a more granular understanding on the genetic traits associated with these populations. Although GWAS studies largely rely on self-reported data on ethnic background, which is the verified using genetic information to confirm ancestral backgrounds and exclude outliers. The exclusion of individuals may lead to insufficient representation of some populations, and disease-association studies of low prevalence or late onset conditions, such as Alzheimer's disease, would be underpowered. Taking population history into account during study design my help to alleviate these challenges.

We have revised the manuscript to provide examples of how multi-disciplinary approaches could improve research findings.

We have revised the manuscript to scale down the number of figures. We have also redrawn the structure plot.

We have made revisions to the manuscript to remove repetitions and improve meaning. We have amended the word "titling" to "tilting".

Competing Interests: None