RESEARCH ARTICLE

Replication of the principal component analyses of the human genome diversity panel [version 1; peer review: 1 approved, 1 approved with reservations]

Thomas Charlon ^{1,2}, Alessandro Di Cara¹, Sviatoslav Voloshynovskiy², Jérôme Wojcik¹

¹Quartz Bio, Geneva, 1202, Switzerland

²Stochastic Information Processing, University of Geneva, Geneva, 1227, Switzerland

V1 First published: 15 Mar 2017, 6:278
<https://doi.org/10.12688/f1000research.11055.1>

Latest published: 15 Mar 2017, 6:278
<https://doi.org/10.12688/f1000research.11055.1>

Abstract

Background. In 2008, several principal component analyses (PCAs) applied on 660,918 single-nucleotide polymorphisms (SNPs) from 938 individuals from 51 worldwide populations of the Human Genome Diversity Panel were published by Li *et al.* PCAs were applied on subsets of individuals sharing a common geographic origin and showed that in several geographic regions, genome-wide variations of SNPs grouped individuals by populations in the two first principal components. In this study, we replicated the PCAs applied on two geographic subsets, first on individuals from Europe and second on individuals from the Middle East & North Africa. **Methods.** Quality control, feature selection, and PCA were applied on each geographic subset. The results were displayed on the two first principal components and compared to the original figures. **Results.** The replicated figures were found to match closely to the original figures. **Conclusions.** Therefore, the main results were replicated and can be independently reproduced by using publicly available data, source code, and computing environment.

Keywords

Bioinformatics, Evolutionary/Comparative Genetics, Genomics

Open Peer Review

Approval Status

	1	2
version 1
15 Mar 2017	view	view

1. **Zoltán Kutalik**, University of Lausanne, Lausanne, Switzerland

2. **Michael G. B. Blum** , Grenoble Alpes University, Grenoble, France

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Preclinical Reproducibility and Robustness** gateway.

Corresponding author: Thomas Charlon (thomas.charlon38@gmail.com)

Competing interests: Thomas Charlon, Alessandro Di Cara, and Jérôme Wojcik are employees of Quartz Bio S.A., Switzerland. The authors declare no competing interests related to this commercial affiliation. This does not alter the authors' adherence to F1000Research policies on sharing data and materials.

Grant information: Quartz Bio S.A. provided support in the form of salaries for Thomas Charlon, Alessandro Di Cara, and Jérôme Wojcik, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work has received support from the EU/EFPIA/ Innovative Medicines Initiative Joint Undertaking PRECISEADS (grant no. 115565).

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

Copyright: © 2017 Charlon T *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Charlon T, Di Cara A, Voloshynovskiy S and Wojcik J. **Replication of the principal component analyses of the human genome diversity panel [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2017, 6:278 <https://doi.org/10.12688/f1000research.11055.1>

First published: 15 Mar 2017, 6:278 <https://doi.org/10.12688/f1000research.11055.1>

Introduction

Quartz Bio and the Stochastic Information Processing group are involved in the PRECISEADS project (<http://www.precisesads.eu/>), which aims at reclassifying Systemic Autoimmune Diseases (SADs), a group of chronic inflammatory conditions characterized by the presence of unspecific autoantibodies in the serum and resulting in serious clinical consequences, based on genetic and molecular biomarkers rather than clinical criteria.

In order to use genetic similarities to deliver personalized treatments to patients affected by SADs as well as other diseases, it is important to first understand the genetic structures in healthy populations.

In 2008, Li *et al.*¹ showed that although specific world regions have different genetic origins, all revealed population structures in principal component analyses (PCAs). Similar population structures were also observed in studies using other genome-wide variations datasets^{2,3}.

Li *et al.* applied PCAs on subsets of individuals from two geographic regions, Europe and the Middle East & North Africa, and displayed the results on the two first principal components in their article as Figures 2A and B, respectively, (with the latter labeled only Middle East).

In an attempt to replicate these two figures, we performed quality control, minor allele frequency filtering, tag SNP selection⁴, and PCAs on both regional subsets of the SNP microarray data. The PCAs were then displayed on the first two principal components.

The replicated figures were found to match closely to the original figures, and therefore confirmed a successful replication.

Methods

Genotype data

The dataset consisted of two files: a zip file including the genotype data of 660,918 SNPs from 1,043 individuals with the annotations of the SNPs, and a text file composed of the annotations of 953 individuals (see Data and software availability).

The annotations of individuals were used to create two subsets of the data. The first contained 157 individuals from Europe and the second contained 163 individuals from the Middle East & North Africa.

Analysis sets

For each geographic region subset of the data, we verified that no individuals had missing value rates above 3% and excluded SNPs with missing value rates above 1%. An additive genetic model was then used to encode each A/B SNP ($A/A = 0$, $A/B = 1$, $B/B = 2$), which converts categorical SNP values to numerics by assuming that the effect of the A/B heterozygote and B/B homozygote are proportional to the number of B alleles. SNPs with minor allele frequency below 5% were excluded to remove rare variants, which are more prone to genotyping errors. In addition, in order to decrease the required computation time and memory usage, redundant SNPs

were removed by applying TagSNP⁴ ($r^2 > 0.8$, window of 500,000 base pairs). The missing values were imputed by random sampling of each SNP. Then each SNP was centered and scaled to unit variance. All steps were performed using the SNPclust R package v1.0.0².

For the Europe subset, a total of 375,164 SNPs from 157 individuals were selected for analysis. This defines our Europe analysis set.

For the Middle East & North Africa subset, a total of 412,979 SNPs from 163 samples were selected for analysis. This defines our Middle East & North Africa analysis set.

For comparison, the supporting online material of Li *et al.* reported that individuals with missing value rates above 2.5% and SNPs with missing value rates above 5% were excluded. Table S1 of Li *et al.* reports that 156 individuals from Europe and 160 from the Middle East & North Africa were used and the supporting online material reports that 642,690 SNPs were used.

Principal component analyses

PCAs were applied on the two analysis sets and displayed using the SNPclust R package v1.0.0². Principal component analysis (PCA) is a dimensionality reduction method, which projects SNPs by linear combination to maximize the variance on successive axes, *i.e.* principal components, while constraining the axes to be orthogonal.

The supporting online material of Li *et al.* reports that they first computed the Identity-by-State (IBS) matrix among the 938 individuals by using PLINK (version not provided)⁵ and then performed PCAs on the IBS matrix for each region separately. In this study, PCAs were applied on the analysis sets and not on IBS matrices.

Results

PCA of the Europe analysis set

The PCA of the Europe analysis set was displayed on the two first principal components (Figure 1). Individuals were grouped by population and the replicated figure matched closely with Li *et al.*'s Figure 2A.

The explained variance was almost identical, as the replication stated 2.1% in PC1 and 1.6% in PC2, while Li *et al.*'s Figure 2A stated 2.4% and 1.6%, respectively.

PCA of the Middle East & North Africa analysis set

The PCA of the Middle East & North Africa analysis set was displayed on the two first principal components (Figure 2). Individuals were grouped by populations and the replicated figure matched closely with Li *et al.*'s Figure 2B.

Two differences from Li *et al.*'s analysis were noted, first the Bedouin and Druze populations exhibited a larger spread on PC1 in the original figure. Second, one Bedouin individual was located with Mozabite individuals, which did not appear in Li *et al.*'s Figure 2B.

Figure 1. Two first principal components of the Europe analysis set. Visualization of the principal component analysis on 375,164 SNPs from 157 individuals from Europe. Individuals from North and South were differentiated in the first principal component and located in the lower and upper sides, respectively. Individuals from East and West were differentiated in the second and located in the right and left sides, respectively.

Figure 2. Two first principal components of the Middle East & North Africa analysis set. Visualization of the principal component analysis on 412,979 SNPs from 163 individuals from the Middle East & North Africa. Individuals from East and West were differentiated in the first principal component and located in the right and left sides, respectively. Individuals from North and South were differentiated in the second and located in the lower and upper sides, respectively.

The explained variance was slightly smaller, as the replication stated 3.1% in PC1 and 2.2% in PC2, while Li *et al.*'s Figure 2B stated 5.0% and 2.6%, respectively.

Discussion

The replicated figures matched closely to the original figures, although two differences appeared when examining the Middle East & North Africa subset: the smaller spread of two populations and the presence of an outlier.

Therefore, the main results were replicated and can be independently reproduced by using publicly available data, source code, and computing environment.

We successfully confirmed that although the two geographic regions studied had different genetic origins, both exhibited population structures in PCAs.

Understanding the genetic structure of healthy populations will enable us to use genetic similarities to deliver personalized treatments to patients affected by SADs. Using this replication, the PRECISESADS project will be able to compare clusters of patients affected by SADs to clusters of healthy individuals, independently from their ancestry-driven genetic structure².

Data and software availability

As stated in Li *et al.*¹, the data sets are freely available online. Although the links that were provided are now outdated, the two data files are available from HGDP-CEPH: <http://www.hagsc.org/hgdp/files.html> (download link: <http://www.hagsc.org/hgdp/data/hgdp.zip> and http://www.cephb.fr/en/hgdp_panel.php#serie2; ftp link: ftp://ftp.cephb.fr/hgdp_v3/hgdp-ceph-unrelated.out).

The PCAs were computed and displayed using the previously published R package SNPclust v1.0.0².

Computing environment in a Docker container is available from: <https://hub.docker.com/r/thomaschln/reproducible-hgdp>

Source code required to generate this article and the definition of the corresponding computing environment, in which all required

software are installed: <https://github.com/ThomasChln/reproducible-hgdp>

Archived source code as at time of publication: doi, [10.5281/zenodo.3451376](https://doi.org/10.5281/zenodo.3451376)

License: GNU General Public License version 3.0

Ethical statement

The data were previously published¹ and approved by ethics committees. No samples were used and records were de-identified.

Author contributions

Conceptualization: JW SV; Formal analysis: TC; Funding acquisition: JW; Investigation: JW ADC; Methodology: TC JW; Project administration: JW; Software: TC; Supervision: JW SV; Validation: TC JW ADC; Visualization: TC; Writing - original draft: TC; Writing - review & editing: JW ADC SV.

Competing interests

Thomas Charlon, Alessandro Di Cara, and Jérôme Wojcik are employees of Quartz Bio S.A., Switzerland. The authors declare no competing interests related to this commercial affiliation. This does not alter the authors' adherence to F1000Research policies on sharing data and materials.

Grant information

Quartz Bio S.A. provided support in the form of salaries for Thomas Charlon, Alessandro Di Cara, and Jérôme Wojcik, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work has received support from the EU/EFPIA/ Innovative Medicines Initiative Joint Undertaking PRECISESADS (grant no. 115565).

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

Acknowledgments

We thank K. Forner for contributions on the software.

References

1. Li JZ, Absher DM, Tang H, *et al.*: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science*. 2008; 319(5866): 1100–1104.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Charlon T, Martínez-Bueno M, Bossini-Castillo L, *et al.*: **Single Nucleotide Polymorphism Clustering in Systemic Autoimmune Diseases.** *PLoS One*. 2016; 11(8): e0160270.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Novembre J, Johnson T, Bryc K, *et al.*: **Genes mirror geography within Europe.** *Nature*. 2008; 456(7218): 98–101.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Stram DO: **Tag SNP selection for association studies.** *Genet Epidemiol*. 2004; 27(4): 365–374.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Purcell S, Neale B, Todd-Brown K, *et al.*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet*. 2007; 81(3): 559–575.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. ThomasChln: **ThomasChln/reproducible-hgdp: Review release [Data set].** Zenodo. 2017.
[Data Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 18 April 2017

<https://doi.org/10.5256/f1000research.11923.r21151>

© 2017 Blum M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Michael G. B. Blum

TIMC-IMAG laboratory (Techniques for biomedical engineering and complexity management – informatics, mathematics and applications – Grenoble), Grenoble Alpes University, Grenoble, France

The authors replicate the ascertainment of worldwide population structure obtained by Li *et al.* (2008). They perform PCA to capture population structure. The PC axes closely match the ones obtained by Li *et al.*

However, the authors found that some Bedouin individuals don't belong to the population they should belong to. The authors should read and cite the 2 following papers that found related results

Jakobsson M, Scholz SW, Scheet P et al: Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 2008; 451: 998-1003.¹

Leutenegger, A.L., Sahbatou, M., Gazal, S., Cann, H. and Génin, E., 2011. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us?. *European Journal of Human Genetics*, 19(5), pp.583-587.²

Additionally, I run the provided docker command (docker pull thomaschln/reproducible-hgdp) to reproduce the analysis but I don't find the generated results. The webpage (<https://github.com/ThomasChln/reproducible-hgdp>) should be improved and should include a more detailed tutorial.

References

1. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, et al.: Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*. 2008; **451** (7181): 998-1003 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Leutenegger AL, Sahbatou M, Gazal S, Cann H, et al.: Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us?. *Eur J Hum Genet*. 2011; **19** (5): 583-7

[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?

Yes

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

Yes

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

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Population genetics, biostatistics, bioinformatics

I confirm that I have read this submission and 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.

Reviewer Report 28 March 2017

<https://doi.org/10.5256/f1000research.11923.r21333>

© 2017 Kutalik Z. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Zoltán Kutalik

Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

This manuscript reports on the re-running of two PCA analyses presented in an earlier publication Li *et al* 2008). The authors confirm the PCA results presented in the original paper and point out two minor differences.

The analysis looks solid and carefully executed. There a few aspects that could be improved:

1. What I missed a bit was the justification why only the middle Eastern and European subsets

were reanalysed. Also, the authors motivate their reanalysis so that they can use these individuals as controls for their PRECISEADS study. I was expecting the authors to go slightly further: do they have control samples? Where do they map on these PCA plots? If they match the location of those from the HGDP, I agree that it is an excellent indication to go further with their study cases. I think these points would further our understanding and go beyond the partial re-analysis of a published data and reporting identical findings.

2. Would be very helpful for the readers to see for every analysis step where did the authors use exactly the same tool as Li *et al* and where do they differ? If at some point different tools were used, were the parameters set to be identical? How close was the pruned subset of SNPs when analysed by them and by Li *et al*?

The title and abstract reflect well the study content. The methods and results are clearly explained, the data are available and the analysis is provided in full details in a Docker container. Study motivation could be better explained and the conclusions in terms of consequences for their future study could be more detailed.

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

Yes

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

Yes

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

Yes

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

Yes

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

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research