Cell Genomics Transparent Peer Review Record

Summary

Initial submission:	Received : 3/22/2024
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 2 Revision invited : 4/24/2024 Revision received : 8/14/2024
Second round of review:	Number of reviewers: 2 Accepted : 10/9/2024
Data freely available:	Yes
Code freely available:	Yes

Referees' reports, first round of review

Reviewer #1: Borda V. et. al. develop a comprehensive resource for the genomic analysis of Latin American populations. First of all, I want to congratulate the authors for this amazing effort and development of the resource. Neverthless, I have several comments about the paper.

Introduction:
Line 18: How do you defined a Latin American population, please expand.
Line 32: This resource aims to uncover fine-scale patterns of population structure across the Americas and boost statistical power for the discovery of genetic factors contributing to LAm health and disease. Neverthless, in the paper only 2 Native American populations were included, which clearly is not a fine-scale analysis, my suggestion is not to use the word fine-scale. Also, how the resource will improve power if you will use only a defined set of individuals, is not clear.
Line 51: Secondly, we address issues about data availability when performing large-scale analyses in LAm populations? How was this performed if all the data came from dbGAP or other repositories?
Line 53: Moreover, association studies in LAm populations face additional challenges, such as smaller sample sizes compared to Europeans and other populations. We propose a matching procedure and sharing summary statistics based on GLAD individuals to overcome these issues. How the sharing os allelic frequency from a fixed number of individuals will increase sample size?
Line 62: We demonstrated the effectiveness of GLAD-match compared to another matching algorithm, PCAmatchR36 63 , regarding genomic control variation. Is not clear.
Line 67: Finally, we demonstrate the potential of GLADdb as a critical resource for evaluating the performance of statistical genetic software in the presence of admixture. How was this adressed, there was not simulations of different levels of admixture and the application of the dataset.
Results:
Line 107: Regarding sample sizes, the best-represented regions in GLADdb included Brazil, Central America, Mexico, Peru, and the United States, Why was Central America considered as a group? This is inline with my previous comment about the definition of Latin America.

Methods:
For ancestry classification a clustering analysis, was performed, please describe in mroe detail the ancestry classification. Was this performed in only USA based individuals or it was also performed in the Latin America based cohorts (individuals recruited in Latin America). How this method converged with other methods of ancestry classification, for example the random forest approach used by GnomAD and PanUKB (https://pan-dev.ukbb.broadinstitute.org/docs/qc/index.html#ancestry-definitions). This is crucial, for the number of samples included in the database as Latin America.

Only Natives American from Peru and Guatemala, Juustify, why you did not performed the analysys using the merged called 1000G and HGDP, free resources could be availabel here (PMID: 36747613) and https://gnomad.broadinstitute.org/downloads.

For the PRS analysis expand using multiethnic summary statistics, example for T2D: https://www.nature.com/articles/s41588-022-01058-3; and evaluate the performance of the PRS in the cohorts.

Line 521: The TOPMed imputation panel contained over 90K individuals and was shown to accurately impute Latin Americans. The Brazilian population is not well represented in TopMed, please, evaluated the imputation accuracy across all the cohorts.

Line 554: We calculated the Fst among individuals sampled by different projects but of the same sample region. Please, justify the use of Fst value as metric for clustering.

Line 639: Please justify the usa of Plink for PCA and not other techniques that are more aware of admixture as PCAiR (https://rdrr.io/bioc/GENESIS/).}

Line 655: You included only Peruvian individuals as NAT, for Local ancestry, please run the analysis using the merged set of 1000G and HGDP and compared the calling of local ancestry. This could clearly impact downstream analysis.

Line 749: For the SNPs included in the down-sampled to Hapmap Phase, where all the same for all cohorts or there where differences, please add a Sup Table describing.

Line 758: Heritability estimation. How did you considered the potential environmental effects. How did you adjust for the optential bias in the GREML estimation

(https://pubmed.ncbi.nlm.nih.gov/37577588/, PMID: 25383972, PMID: 37131817)

In the matching for selection of controls for the target study, how did you control for potential gene-environment interactions that could deviate the summary statistics? (For example: PMID: 32888427)

To all Supplementary Tables add Notes describing the headers and appropiate data for interpretation.
Supplementary Table 1. Include the country of residence and recruitment of the individuals. Include information about the original genome reference and information about SNPs included in each level of analysis.
Supplementary Table 2. Add an analysis of unsepervised ADMIXTURE.

Reviewer #2: The authors present a broad study of genetic diversity across Latin America. This study is timely and showcases the vast diversity and history from understudied countries and understudied populations. The authors conducted multiple analyses that capture how human ancestry, migration, and importantly the GLAD platform can improve our understanding of both human history and health. Overall, I found the study to be well formulated and the analyses to be exciting and well executed. However, I only had two larger comments/questions about the chosen approaches, and I also include minor comments.

Major:

1) Perhaps I mis-understood this, but I was wondering why the authors chose to do the ancestry clines with UMAP? Why not do ancestry clines with PCA, where it is linear and we can see the mixtures of ancestry where we expect them?

We know that UMAP does strange things to the genotype data and distorts distances between individuals. The ancestry cline analysis highlighted this distortion for me.
The three major ancestry groups (IA, European, and African) look pretty distinct when we know that the LAm groups are a mixture of them, like the authors concluded (lines 99-108).
If the authors would like to keep the UMAP plots, it would make sense to remove the labels for tick marks on x and y since the distances are arbitrarily distorted and meaningless in UMAP space.

2) Figure 2 is pretty confusing and a bit messy. There are underscores in the country labels; no spaces for other labels (LatinoCountry; NativeAmerica, NotHispanic); and no labels for the axes in 2A. I think those are the pairs of individuals being compared, but I am not completely sure? Is each individual represented as a row in 2B?
I see what the authors are going for, perhaps the authors can distill the information down in a different way?
Could you maybe use Louvain clustering and build IBD networks? Then, you could label the individual with cluster, country, and ethnicity (2B). The network density would sort of recover the info for 2C.

3) The equations/summary statistics for the matching section did not render for any of the versions I downloaded. I also only see boxes for Results in text, e.g. Line 188
Can the authors double check those?

Minor:
The fonts do not match across the figures, which was distracting to me, but I will leave the choice to correct to authors.
1) Figure 3: remove underscores from country labels
2) Figure 4: The USA-NY label is blocking relevant parts of the plot can that be moved?
3) Figure 5: I might remove the plot titles and put them in the caption. Also, the font changes within the figure, which was rough, and it was difficult to read axes in B.

Authors' response to the first round of review

Dear Reviewers

We sincerely appreciate your dedication and meticulous review of our manuscript. In response to your valuable feedback, we have a thorough point-by-point response. The revisions made in this manuscript version effectively address the concerns you raised.

To enhance clarity, we have taken the liberty of highlighting reviewers' comments in bold, differentiating our responses with italics and indicating changes made based on the reviewers' suggestions in red font.

We are happy to provide an improved version of the manuscript.

Thank you

Response to Reviewers

Reviewers' Comments:

Reviewer #1: Borda V. et. al. develop a comprehensive resource for the genomic analysis of Latin American populations. First of all, I want to congratulate the authors for this amazing effort and development of the resource. Nevertheless, I have several comments about the paper.

Introduction:

R1-1. Line 18: How do you define a Latin American population, please expand.

Thank you for pointing this out. In the introduction, we stated, "Latin Americans, as an ethnic label, encompass diverse populations across the Americas with distinct ancestral composition resulting from admixture between various global populations." However, we recognize there are better definitions. Therefore, we include the following definition: a Latin American population is a group of people with heritage from Spanish-speaking or Portuguese-speaking countries in the Americas. We chose this definition because "heritage" encompasses various aspects, from culture and geography to genetics.

Now, we included this definition in the introduction to clarify the scope of this research. Considering the geographical aspect, we selected datasets from countries fitting this concept. However, US-based cohorts claiming to include US Hispanics/Latinos posed a challenge. For these cohorts, we consider the cultural or genetic aspect. Some cohorts included self-declared ethnicity (i.e., as Hispanic/Latino or not Hispanic/Latino), but others did not. For cohorts without self-declared ethnicity, we ran ADMIXTURE in a supervised mode with four parental ancestry groups: European, African, East Asian, and Indigenous Americans. We labeled an individual as Latin American if it shows at least 2% Indigenous American ancestry, indicating a signal of local admixture. This definition could be very restrictive, considering some Latin American groups might have no Native American ancestry (i.e., Brazilians or Argentinians). Still, it ensures the maximum number of individuals is collected without strong bias. In summary, Latin American individuals were selected based on two criteria: (i) self-definition (which means sampling country or self-declared ethnicity) or (ii) the presence of Indigenous American ancestry in the genome-wide composition of individuals when no ethnicity information is available for US cohorts.

Furthermore, our definition does not include the Indigenous populations because this ethnicity label sometimes has a country-specific legal background. Moreover, genetic data for Indigenous American individuals is even more scarce than for Latin American individuals. However, one of our goals is to explore the migration dynamics between countries, so we are assuming that the contribution of self-declared Indigenous American individuals to this inter-regional gene flow is low. Finally, we included some self-declared Indigenous American groups from Peru, but we limited their use to a reference population for local ancestry analyses and used them to compare runs of homozygosity.

Actions:

We improve and clarify these concepts in the Introduction and Methods sections.

R1-2. Line 32: This resource aims to uncover fine-scale patterns of population structure across the Americas and boost statistical power for the discovery of genetic factors contributing to LAm health and disease. Nevertheless, in the paper only 2 Native American populations were included, which clearly is not a fine-scale analysis, my suggestion is not to use the word fine-scale. Also, how the resource will improve power if you will use only a defined set of individuals, is not clear.

Thank you for highlighting these two great points!

The first point, about 2 Native American populations, is related to our previous response in which we clarified some definitions, the scope, and our focus on non-Indigenous populations.

For the second question, how will the resource improve power if you use only a defined set of individuals? One of the resource's intended use cases is to, in conjunction with our matching algorithm and a set of query genomes composed of cases and/or controls for a study, provide a set of matched genomes from GLAD (a subset of GLAD) that share genetic similarities to be used as additional controls for that study. Our current means of preserving anonymity requires sharing the additional controls as summary statistics (allele and haplotype frequencies) to avoid sharing individual-level genomic data.

This sharing forces the end user to perform their study using a test that does not require access to covariates, such as a chi-squared test. This means we must account for some potential confounding effects, such as population structure and age, directly in the matching process. This is how we argue that this resource can boost the external study's power.

While these limitations narrow the applicability of the resource, smaller studies where no or very few controls are available may find the incurred limitations worth the additional statistical power gained through access to GLAD controls.

Actions:

• In the methods section, we added a paragraph explaining the criterion for Latin Americans and why we did not include Native Americans.

• We removed line 32 and added more details in the introduction and discussion sections about how GLADdb will support external studies.

R1-3. Line 51: Secondly, we address issues about data availability when performing large-scale analyses in LAm populations? How was this performed if all the data came from dbGAP or other repositories?

Thank you for this observation. For others to get access to this data is an arduous process, which we are mitigating through doing this for others. This took significant time even after pulling the data from publicly available sources, as each study had a different set of nongenetic data files, and some lacked information on self-described ethnicity or sampling location. This harmonization process was not straightforward and is leveraged in our matching algorithms and data analysis. Also, some measures of where people come from will be added in future implementations so that an external user can know the general source of the samples (e.g., 50% came from phsXXXX) so that if they find a specific cohort matches their own, they can target an application for that particular dataset knowing it will be a good fit to their samples. Finally, after matching, we provide the allele and ancestry haplotype counts of a subset of individuals so the external user can enhance their data without transferring individual data.

Action: We add more information to the paragraph corresponding to line 51 and the results section (subsection: Supporting external studies through the GLADdb matching algorithm and statistical genetic software benchmarking).

R1-4. Line 53: Moreover, association studies in LAm populations face additional challenges, such as smaller sample sizes compared to Europeans and other populations. We propose a matching procedure and sharing summary statistics based on GLAD individuals to overcome these issues. How the sharing of allelic frequency from a fixed number of individuals will increase sample size?

Thank you for highlighting this point. We considered the reply to comments R1-2 directly related to this question. Our goal is to provide allele and local ancestry counts to facilitate chi-square tests in a set of individuals in which population structure will not be a major confounding factor due to genetic proximity obtained with the matching. Additionally, even when GLAD has a fixed number of individuals, the number of matched individuals will depend on the external dataset and the parameters used by external users.

R1-5. Line 62: We demonstrated the effectiveness of GLAD-match compared to another matching algorithm, PCAmatchR, regarding genomic control variation. Is not clear.

Thank you for bringing this to our attention. The end of the sentence you highlighted was indeed unclear. We have rewritten it.

Before: "We demonstrated the effectiveness of GLAD-match compared to another matching algorithm, PCAmatchR, regarding genomic control variation"

Now: "We demonstrated the effectiveness of GLAD-match compared to another matching algorithm, PCAmatchR, using genomic control as a proxy for the quality of matched individuals."

The details for this comparison are fully described in the section titled "Supporting external studies through the GLADdb matching algorithm and statistical genetic software benchmarking." There, Table 1 contains comparisons of the genomic controls (for a dummy binary phenotype representing belonging to the query cohort) achieved by our matching algorithm compared to the one that PCAmatchR uses (which we refer to by its general name - bipartite matching).

Action: The paragraph corresponding to line 62 was edited, and more information related was added to the Methods section (subsection: Matching). We also updated the Table 1 caption.

R1-6. Line 67: Finally, we demonstrate the potential of GLADdb as a critical resource for evaluating the performance of statistical genetic software in the presence of admixture. How was this adressed, there was not simulations of different levels of admixture and the application of the dataset.

Thank you for your observation. Our goal was to use the GLADdb cohorts to observe the behavior of the PRS algorithm in highly admixed populations rather than engage in a thorough discussion about PRS. Interestingly, GLADdb includes populations with different levels of African, European, and Indigenous American ancestries that can be useful in identifying the limitations of PRS or other algorithms in genetic epidemiology.

Additionally, to evaluate the performance of GLAD-match, we performed an empirical test using several cohorts with a different ancestral background (Table 1). For this, we created a dummy binary phenotype that defined query individuals as cases and matched individuals as controls. Matched controls were identified using gladmatch and PCAmatch (referred to in the paper as bipartite matching). Then, a pseudo-GWAS was performed with the dummy phenotype, and finally, we estimated the genomic inflation parameter and compared which method provided the lowest.

Action: We edited line 67 and improved the method section (Matching subsection)

Results:

R1-7. Line 107: Regarding sample sizes, the best-represented regions in GLADdb included Brazil,

Central America, Mexico, Peru, and the United States, Why was Central America considered as a group? This is inline with my previous comment about the definition of Latin America.

Thank you for pointing this out. We agreed that our definition of Latin America is unclear. After considering this observation in the first comment and improving the text, we decided to remove broad terms when possible.

Action: We replaced Central America, and we specified country-level information.

Methods:

R1-8. For ancestry classification a clustering analysis, was performed, please describe in more detail the ancestry classification. Was this performed in only USA based individuals or it was also performed in the Latin America based cohorts (individuals recruited in Latin America). How this method converged with other methods of ancestry classification, for example the random forest approach used by GnomAD and PanUKB (https://pan-dev.ukbb.broadinstitute.org/docs/qc/index.html#ancestrydefinitions). This is crucial, for the number of samples included in the database as Latin America.

The clustering analysis to identify possible Latin American individuals was performed only in US cohorts with no ethnic information. This analysis was not performed to identify individuals related to a particular ethnicity but to infer ancestry proportions. Since some Latino individuals have Indigenous American ancestry, we set 2% as the threshold to identify as Latin American. We recognize this is not a perfect strategy, considering that some Latin American individuals can have only European, African, or both ancestries with no traces of Native American ancestry (i.e., Argentinian, Brazilian, etc). Interestingly, our PCA visualization showed no major bias when individuals were clustered. That is, clusters include self-described and ADMIXTURE-defined individuals (See Figure S7).

Action: We updated the data description in the Results section. Now, we include details about the criterion for identifying Latin American individuals in US cohorts without demographic information.

R1-9. Only Natives American from Peru and Guatemala, Juustify, why you did not performed the analysys using the merged called 1000G and HGDP, free resources could be availabel here (PMID: 36747613) and https://gnomad.broadinstitute.org/downloads.

The merged call, including 1000 Genomes Project (1KGP) high coverage and HGDP generated as part of GNOMAD resources, is an important reference panel for human genetic diversity (Koenig et al. 2024). It includes Latin American (MXL, PEL, CLM, and PUR from 1KGP) and Indigenous American populations (Maya, Pima, Surui, Karitiana, and Colombians from HGDP) that can be used as a reference for Indigenous American ancestries. However, we have two reasons for not including them in our reference panel.

First, we selected individuals with at least 99% Indigenous American ancestry for Indigenous American reference in our Local ancestry analyses. To determine the Indigenous American ancestry proportions, we ran ADMIXTURE in unsupervised mode on a subset of 1KGP + HGDP that included All European, African, East Asian, and Latin American populations from 1KGP and all Indigenous Americans from HGDP. Results from ADMIXTURE K=4 showed only 21 individuals met this criterion. Second, the overlapping between GLAD and 1KGP + HGDP reduced our SNP density (final merged=80%).

Even with these two issues, we merged them and compared the results of the local ancestry. For this analysis, we used rfmix version 2, which provides a confusion matrix that can be used to determine the accuracy of a reference panel. The confusion matrix resulting from our reference panel without the 21 individuals accumulated fewer errors in prediction compared to the reference panel with the 21 individuals (Figure below). For example, for Indigenous American ancestry in chromosome 20, our GLADdb showed 97.4% accuracy compared to 95.7% of the GLADdb.

A final consideration is that we ran local ancestry using rfmix with two expectationmaximization (EM) steps. These EM steps allow rfmix to include ancestry segments inferred from the query data to enhance the original reference panel, thereby improving the inference of ancestry segments in subsequent runs for other individuals (Maples et al., 2013). Moreover, in the discussion, we recognize this as one limitation of our ancestry analyses.

References

Koenig Z, Yohannes MT, Nkambule LL, Zhao X, Goodrich JK, Kim HA, Wilson MW, Tiao G, Hao SP, Sahakian N, Chao KR, Walker MA, Lyu Y; gnomAD Project Consortium; Rehm HL, Neale BM, Talkowski ME, Daly MJ, Brand H, Karczewski KJ, Atkinson EG, Martin AR. A harmonized public resource of deeply sequenced diverse human genomes. Genome Res. 2024 Jun 25;34(5):796-809. doi: 10.1101/gr.278378.123. PMID: 38749656; PMCID:

PMC11216312.

Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet. 93, 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020.

Referees' reports, second round of review

Reviewer #1: Comments enter in this field will be shared with the author; your identity will remain anonymous.

Reviewer #2: I felt the authors for addressed my comments thoughtfully, and appreciated their willingness to do the additional analyses to correct for structure as well with PC-AiR.
This resource will be helpful for future researchers, and my remaining comments are quite minor.

Minor Comments:
Link is broken https://github.com/umb-oconnorgroup/gladmatch
Maybe consider rewording to Line 2-3: Latin American populations are underrepresented in genetic studies, which increases disparities in personalized genomic medicine.
Line 41: Figure1A etc. not bold
Line 95: Typo Figures S2 and not bold
Line 100: Figure S3 and S4 not bold
Line 107: Local is capitalized
Line 109: 7 is not bold
Line 112: Figure S8B not bold
Line 129 & 164 Afro-Peruvians
Line 144: citation/stable doi for infomap?
Line 317: Individuals is capitalized

In the discussion, if space allows, the authors should consider citing/relating IBD network analyses to results from Baharian et al. 2016 since the approach and results are quite similar.
https://doi.org/10.1371/journal.pgen.1006059

Figures:
Can authors double check figure captions (Figure 2 & Figure S10) to make sure captions say self-described ethnicity when applicable?
Figure 1 USA is being cut-off by the dot and 1KG population abbreviations are not in in caption
Figure 6 remove titles in C & D

Authors' response to the second round of review

Reviewer #1: No comments.

Thank you so much for your effort and time during the whole reviewing process.

Reviewer #2: I felt the authors for addressed my comments thoughtfully, and appreciated their willingness to do the additional analyses to correct for structure as well with PC-AiR.
This resource will be helpful for future researchers, and my remaining comments are quite minor.
Thank you so much for all your suggestions during the whole reviewing process.
Minor Comments:
Link is broken https://github.com/umb-oconnorgroup/gladmatch
We apologize for this inconvenience. We have updated the permissions, and the link is now publicly available.
Maybe consider rewording to Line 2-3: Latin American populations are underrepresented in genetic studies, which increases disparities in personalized genomic medicine.
Thank you for this suggestion; we have updated these lines.
Line 41: Figure1A etc. not bold
Line 95: Typo Figures S2 and not bold
Line 100: Figure S3 and S4 not bold
Line 107: Local is capitalized
Line 109: 7 is not bold
Line 112: Figure S8B not bold
Line 129 & 164 Afro-Peruvians
Line 144: citation/stable doi for infomap?
Line 317: Individuals is capitalized
Thank you for your valuable suggestions. We have revised the text accordingly, addressing all these observations.

In the discussion, if space allows, the authors should consider citing/relating IBD network analyses to results from Baharian et al. 2016 since the approach and results are quite similar.

Response to Reviewers

https://doi.org/10.1371/journal.pgen.1006059

Thank you for this important detail. We added this reference in the discussion section in the context of ancestry-biased migrations.

Figures:

Can authors double-check figure captions (Figure 2 & Figure S10) to make sure captions say selfdescribed ethnicity when applicable?

Figure 1 USA is being cut-off by the dot and 1KG population abbreviations are not in in caption

Figure 6 remove titles in C & D

Thank you. We have improved the captions for Figures 1, 2, and S10. In the specific case of Figure 6, we removed the title just from plot D, as we believe the title in plot C is necessary to describe the phenotype. Thank you so much to both reviewers