15-Oct-2014

Dear Dr. Fedorov,

Your manuscript GBE-140909 entitled "Inference Of Distant Genetic Relations In Humans Using “1000 Genomes”" submitted to Genome Biology and Evolution, has now been reviewed. The comments of Prof. Partha Majumder, the Associate Editor who handled the peer review process, and the referees are included at the bottom of this letter.

While the overall evaluation of the MS is generally positive, some major revisions to your manuscript will be needed before it can be considered for publication. Therefore, I invite you to respond to the referees' comments and revise your manuscript, but please consider the comments carefully and take them into account during the revision process. Please ensure that you include both, a Word document of the manuscript text and tables, and a pdf file of the whole paper, including all tables and figures but without tracking or supplementary material, as well as the usual files.

To revise your manuscript, log into http://mc.manuscriptcentral.com/gbe and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.

You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript using a word processing program and save it on your computer. Please also highlight the changes to your manuscript within the document by using the track changes mode in MS Word or by using bold or coloured text.

Once the revised manuscript is prepared, you can upload it and submit it through your Author Centre.

When submitting your revised manuscript, please respond to the comments made by the referees in the space provided, documenting also changes you make to the original manuscript. The responses will be available to reviewers should there be a further review.

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.

Because we are trying to facilitate timely publication of work submitted to Genome Biology and Evolution, your revised manuscript should be uploaded as soon as possible. If it is not possible for you to submit your revision in a reasonable amount of time, we may have to consider your paper as a new submission.

Once again, thank you for submitting your manuscript to Genome Biology and Evolution and I look forward to receiving your revision.

Yours sincerely,

Bill Martin

Editor in Chief, Genome Biology and Evolution

w.martin@uni-duesseldorf.de

Associate Editor Recommendation:

Associate Editor: Majumder, Partha

Comments to the Author:

The manuscript needs to be revised thoroughly in the light of the comments made by the reviewers. The manuscript will be re-reviewed by the same reviewers. No guarantee can be made at this time regarding the acceptance of the manuscript for publication in GBE. GBE also does not encourage multiple rounds of revision of a manuscript. Therefore, careful attention to the reviewers' comments are crucial and the manuscript must be revised adequately for further consideration. It is essential to provide a point-by-point actions-taken report.

Reviewers' Comments:

Referee: 1

Comments to the Author

In this study, the authors demonstrate two different approaches to detect genetic/familial relationship between pairs of individuals. The first is based on counting the genomic differences that exist between the pair of individuals. And the second one uses the number of shared vrGVs (very rare genetic variants) to estimate the familial relationship between two individuals. The authors claim that the second approach is much more sensitive and it radically improves the detection of genetic relationship in humans.

I found their hypothesis interesting and promising. This appears to be a well-organized work and a new direction towards detection of human relationships on the basis of genomic data. In the context of the rapidly decreasing cost of genome sequencing, this looks like a timely piece of work with great potential.

However, I also felt that there are a few concerns that needs to be taken care of.

MAJOR POINTS:

While discussing the frequency and the distribution of rare SNPs the authors should cite a very important paper in this field by Keinan and Clark “Recent explosive human population growth has resulted in an excess of rare genetic variants” Science 336:740-3, 2012.

In the Discussion on pp19-20, the authors estimated the probabilities of shared vrGVs in non-related individuals. However, some of vrGVs may be in linkage disequilibrium with each other. This effect should be evaluated.

MINOR POINTS:

Firstly, some comments and/or claims appear difficult to understand mainly due to structural malformations of the sentence (typo?!). For example, it may be non-obvious to interpret the last sentence of the introduction (After accomplishment………. Genetic studies.). Technology race??

Secondly, there are some grammatical errors in the narrative. For example, usage of both past and present tense is noted in the paragraph under the heading ‘Statistics’ and also in the immediately following paragraph.

In the third last sentence on page 12, the number ‘2003’ appears inside parentheses with a lack of relevance. In line 7 on page 20, they refer to “one or two shared vrGVs” as negligibly small, but do not mention the frequency.

In addition to all these structural aberrations, I found the following discrepancy:

The Monte-Carlo simulation for calculating the maximum number of shared vrGVs in a virtual population with parameters similar to the LWK population gives 720 as the highest number of shared vrGVs. The authors consider this number as a “statistically validated threshold” and hypothesize that any African pair of individuals sharing more than 720 vrGVs should be considered distant relatives. When the same Monte-Carlo simulation gives a ‘threshold’ value of 128 for the GBR-like parameters, the authors claim that GBR pairs sharing more than 200 vrGVs should be distant relatives. They do not provide any explanation for choosing a value 1.56 times higher than the expected threshold obtained from the computer simulation.

Finally, on page 8 line 20th, the authors made a reference to a Wikipedia. This is very unusual for a respectful scientific journal.

But, overall, the text is organized and the figures are also explained elaborately.

Referee: 2

Comments to the Author

The paper by Al-Khudhair et al. entitled ‘Inference of distant genetic relations in humans using “1000 Genomes”’, examines relationships in 1,092 samples belonging to 14 populations that were whole genome sequenced at low (2-6X) coverage by the 1000 Genomes Project Consortium. They use several Perl scripts to interrogate the vcf files released by the project and count the number of genomic differences among 96,464 pairs of individuals to identify 271 distant pair-wise relations among these samples. Overall this is an interesting study with a few caveats that need to be addressed.

1. A cut-off of ≤0.2% is used to characterize very rare genomic variants. This could potentially include sequencing, mapping or calling errors, as the analysis is not restricted to the regions that were targeted by deep (50–100X) exome sequencing. Heterozygous calls would be enriched for errors in low coverage data and the authors do not describe what proportions of their rare variant calls are heterozygous. A figure showing the proportion of rare variants in high coverage exome versus low coverage sequenced regions and in genic versus intergenic regions should be included.

2. Data quality should be assessed by counting the number of high quality reads supporting the rare variant call, cross-validation with the SNP-chip genotypes and/or experimental validation of a proportion of the shared rare genomic variants.

3. The authors analyze SNVs, indels and 14,000 deletions. The heterozygous indels and large deletions would have an even higher error rate than SNVs and there is no data showing what proportion of their cryptic relatedness is based upon these rare indels and larger deletions.

4. The authors state that genomic differences were computationally assessed for 96,464 human pairs, but it’s not clear why did they not assess all possible pairs within and between populations.

5. Moore et al. (PLOS Genetics, 2013, vol 9:1003959) identified 75 cryptically

related individuals from the same data set using a minor allele frequency (MAF) cut-off of < 0.03 and the authors should compare whether all these samples were picked up in their analysis.

6 The authors mention haplotype sharing, but do not address what proportion of their rare variants are in linkage disequilibrium.

7. They discuss a model where 5,000 loci are in linkage equilibrium with each other and that corresponds with the distribution of the number of genetic variants in the GBR population, but they do not show whether this holds true for Africans where this number should be substantially reduced because of their population history and demography.

8. They give several examples of loci in strong linkage disequilibrium to support their model that suggests an average genomic size of 600 kb, but the region encompassing rs4988235 is a classic example of selective sweep in European populations. Such selective sweeps are known to reduce genomic diversity and this is not factored into their model.

Minor Points:

1. There are a number of grammatical and spelling errors in the manuscript for e.g.

a. “principle data” (Introduction, page 3, line 12).

b. “relay” instead of “rely” (Introduction, page 4, line 3).

2. Table 1 should be in supplementary data.

3. The three letter population codes are inconsistent in the main text and figure lends for e.g. Japanese are sometimes correctly referred to as “JPT” or often incorrectly as “JPN”. Similar errors are observed for TSI and CEU (See Discussion, Page 17, line 31).

4. Figure 3A is unclear and should be enlarged. For consistency all figures should show rare variants on left side (as in figures 1 and 2) and genetic relations marked by either arrows or stars.