February 14, 2016

Re: Minor Revision for BIOINF-2016-1696.R1 "Identification of genetic outliers due to sub-structure and cryptic relationships"

Oliver Stegle
Associate Editor
Bioinformatics

Dear Dr. Stegle,

We are pleased to submit our manuscript with minor revisions for BIOINF-2016-1696 "Identification of genetic outliers due to sub-structure and cryptic relationships". Thank you and the three reviewers for the careful review. We appreciate the improvements that these comments have allowed us to make to our paper.

The typos and formatting error that were discovered by both reviews 1 and 3 have been fixed (along with an additional missing word on page 2). Additionally, we have expanded the our description in section 3.1 in response to the comment from reviewer 2, and added to the acknowledgements section.

Sincerely,
Daniel Schlauch
Heide Fier
Christoph Lange

Editor Comments

R2 has some remaining comments, which we would like to see addressed.

Reviewer Comments:

Reviewer: 1

Comments to the Author
The authors addressed satisfactorily all my comments. Despite the careful proofreading, I still was able to spot some typos (e.g. "it's" instead of "its", criterion "instead" of "criteria", reference style for the 1000GP papers), so I urge the authors to go through the text again if possible.

We thank Reviewer 1 for the encouraging review and valuable comments at the major revision stage. We have addressed these proofreading items and have further scoured the manuscript for any typos.

Reviewer: 2

Comments to the Author
Overall, the authors have responded to the comments OK. However, I believe they have misunderstood my second major comment (Line 50 of P4):

You compute your statistic using only rare variants, which you define as MAF <.01. My question asks how does the statistic vary if you varied the threshold, and instead, used only SNPs with, say, MAF < 0.001?
I believe you have partially answered this in one of your later responses, where you indicate the statistic would tend to increase. However, I was looking for you to demonstrate, that your statistic (or perhaps instead your conclusions about which individuals were related) did not (substantially) change if you varied the MAF threshold.

Signed Doug Speed

We thank reviewer 2 for valuable comments and apologize for our confusion regarding our initial response to the second major comment. As the reviewer points out, the expected value of our statistic is a function of both the kinship coefficient and the set of minor allele frequencies (Equation 6). It's therefore of interest to consider the sensitivity of our statistic to the MAF used. Informally, we attempted to find a balance between lower frequencies (more informative of recent ancestry) an quality control considerations. In our revision we added Supplemental Figure 5, which illustrates the concept that population separation improves with lower MAFs, except at the lowest MAF (<0.004). We believe this feature of the data attributable to technical artifacts which represent a greater proportion of the variants at these allele frequencies. While it is of interest to explore the sensitivity of the statistic in the 1000 Genomes Project data, we feel that any result we obtain will not be generalizable to other studies because of study-specific nature of the the technical artifacts.

To highlight the work we have done addressing this question, we have expanded paragraph 2 in section 3.1 and pointed the reader to supplemental materials which explain the reasons and methodology of our filtering decisions in greater detail.

Reviewer: 3

Comments to the Author
General comments:

The authors greatly cleaned up the manuscript, added appropriate derivations, examples, and intuition behind their statistic, and shifted the emphasis of the paper appropriately to whole genome sequencing studies. The manuscript is much more readable and understandable now.

Maybe give the manuscript one more pass to fix small stylistic issues like:

"to it’s preferential weighting" -> "to its preferential weighting"
spacing in reference: Nemesh, J. and McCarroll, S. (2012)
center justify in the title of fig 1
"quantitative utility for removal of outliers" -> "quantitative utility for the removal of outliers" or "quantitative utility for removing outliers"

We thank Reviewer 3 for their helpful comments at the major revision stage and agree that the manuscript is now more readable and understandable. The items listed above have been addressed in this revision.