December 31, 2016

Re: Major Revisions for BIOINF-2016-1696 "Identification of genetic outliers due to sub-structure and cryptic relationships"

Oliver Stegle
Associate Editor
Bioinformatics

Dear Dr. Stegle,

Thank you for the opportunity to revise our manuscript, BIOINF-2016-1696 "Identification of genetic outliers due to sub-structure and cryptic relationships". We appreciated the careful review and valuable comments of the Associate Editor and reviewers and have addressed each of the them below resulting in a improved manuscript.

In this letter, we have included the comments from the Associate Editor and the three reviewers, with our responses embedded in red.

Sincerely,
Daniel Schlauch
Heide Fier
Christoph Lange

Editor Comments

Your paper has now been seen by three referees. All reviewers highlight that the proposed method will in principle be of interest, however in particular reviewers 1 & 3 highlight major concerns that currently preclude publication. If you are preparing a revision, we would request compelling evidence that the method adds practical utility compared to the state of art. Both reviewers 1&3 request additional comparisons with existing strategies to detect outlying samples. Also, the referees highlight complementary applications to dissect cryptic structure, which may be worth considering.

We are grateful for the Editor's comments and helpful reviews. We have diligently considered each of the reviewers comments, and in particular those which were highlighted here. As we agree with all of them, we have revised our manuscript accordingly and have made the following major changes to address the points raised by the reviewers. We hope you will find this aides in the demonstration and readability of our manuscript and add clear evidence of utility over existing strategies.

For the application of our proposed methodology to genome-wide association studies (GWAS) in which common variants are analyzed, we agree with the reviewers that our approach adds relatively little in terms of practical utility, although it provides a more sounds theoretical framework in terms of the statistical analysis, e.g. its relationship to the kinship coefficient, testing for outliers.

However, for the application to whole-genome sequencing studies in which millions of rare variants are analyzed our approach has distinct theoretical, numerical and practical advantages over existing variance/covariance approaches that present substantial advances in terms of utility. Our method can take full advantage of the information contained in rare variant data, our approach is well suited to detect both subtle differences and similarities in population substructure, while variance/covariance approaches are not tailored to do this. Given the advantages, our approaches will provide much needed practical utility/advantages for the application to whole-genome sequencing studies and was designed with this purpose in mind. The application to GWAS data/common variant is really secondary and the main point that we aimed to make here was that our approach is very versatile and, when applied to GWAS data, does at least as well standard approach in terms of practical utility. We apologies for this misunderstanding and have revised the manuscript accordingly, stressing the targeted application area, whole-genome sequencing studies. Furthermore, we made the following key changes in the manuscript:

(1) A comparison of the performance of STEGO with the use of a variance-covariance approach (as in PCA) in groups of recently related pairs of populations of the 1000 Genomes Project.
(2) An toy example in the supplement which will aide in reader intuition and understanding.

(3) A demonstration, using 1000 Genomes data, of the motivating basis for our method, specifically that rare variants provide the ability to achieve higher resolution ancestry inference.
(4) A simulation, illustrating the improved performance of our method compared to PCA (EIGENSTRAT) in the context of detecting very subtle population stratification (Supplementary Materials 1.7).
(5) Expansion and edits of the discussion section, mathematical derivations, and methods section as suggested by the reviewers.
(6) Reduction of the length by moving section 3.1 of the Results section (1.8 "Simulations demonstrate power to detect heterogeneity") to Supplementary Materials 1.7.

A pdf containing the latexdiff comparison of the revisions versus the original submission is included in this submission.

Reviewer: 1

Comments to the Author
In this work, Schlauch et al. "propose a formal statistical test [STEGO] that assesses whether two study subjects come from the same population and whether they are unrelated". The proposed similarity index ingeniously gives more weight to rare variants (as they are more informative about recent population separations on the sub-continental level) and it can be calculated for both haploid and diploid genomes. The authors also propose a formal test for heterogeneity; quantify the relation between the new statistic s and kinship coefficient φ; and also provide a useful power calculation formula (which is validated by well-designed simulations). Finally, the authors apply their method to the 26 1000 Genomes Project populations and discover cases of substructure and cryptic relatedness beyond what has been previously reported using established methodologies. Finally, the genetic relationship matrix based on the new method seems to separate closely related populations with better resolution compared to the standard variance-covariance matrix.
The manuscript is well written, the argument easy to follow and the testing of the algorithm properly carried out. However, the impact of the method on the overall improvement of a hypothetical QC pipeline is somewhat undermined by the fact that, after cleaning the 1000GP dataset for related individuals with established protocols, the discovery of additional structure/relatedness by STEGO fell down significantly. This is not to discredit the overall superiority of STEGO compared to established methods, but it seems to provide little incremental improvement compared to standard QC analysis – at least for the 1000GP data.
Having said that, I believe that the most intriguing finding of this work is the one reported in section 3.3. Given that the decomposition of the STEGO-based GRM seems to separate more efficiently closely related populations, I think STEGO can find some really interesting applications to ancestry prediction and mixed models for gene mapping. The paper could benefit greatly by a slight shift of the focus on these two aspects.

We thank the reviewer for the encouraging and positive assessment of our work. We have incorporated all of the comments below, resulting in an improved manuscript. In particular, we agree with the potential of STEGO-based GRMs to more efficiently separate more closely related populations and have expanded on those portions of the manuscript.
As the reviewer points out, after using the methods of (Gazal et al 2015) to filter cryptically related individuals, only 4 of the 26 populations showed significant evidence of heterogeneity. We consider this to be strong evidence of population homogeneity among 22 of the populations, owing to the state-of-the-art collection and sequencing protocols in the 1000 genomes project. We contend that any identified structure within a 1000 genomes project population is noteworthy and demonstrates the sensitivity of STEGO.

Major comments:
1. As mentioned above, I believe that the true value of a better variance-covariance matrix by STEGO and subsequent PCA lies in that it could benefit ancestry studies whereby one tries to predict the ethnic background of an unknown sample based on a set of pre-calculated SNP weights (see Chen et al., bioinformatics 2013). In such case, a better PCA could increase the resolution of the prediction. It would be nice if the authors added this insight in their discussion.

We think the idea of inferring higher resolution ancestry of an unknown sample based on SNP weights has great promise by using STEGO. As suggested, we have added this idea to the discussion and referenced Chen et al.

"Moreover, our approach involves the estimation of a GRM which, due to it's preferential weighting towards rare variants, provides higher resolution for distinguishing populations which have recently diverged. As sequencing costs have plummeted and our ability to measure rare variants has increased, there will be increased demand for tools which make use of the differential informativeness of variants according to frequency. Recent work (Chen 2006) has already demonstrated the use of pre-calculated SNP weights to infer the ancestry of samples of unknown origin, and STEGO's GRM in

combination with large scale sequencing projects, such as the TGP, promises to further improve the resolution of this approach."

2. In the third paragraph of the Introduction, there is an inaccurate statement about the process of PCA-based outlier removal. Outliers can of course be removed after visual inspection, but EIGENSTRAT's SMARTPCA actually has a utility that removes outliers across a given number of PC's and for a given number of iterations (see their manual for details). Please acknowledge this in the text.

We agree with the reviewer that acknowledging SMARTPCA here is important. We have changed the wording in this paragraph from "The standard practice is..." to "A common practice is..." to reflect the fact that visual inspection is not the only approach taken here. Furthermore, we have referenced Patterson et al and acknowledged how their method uses principal components to identify and remove outliers.
"...Alternatively, a software tool SMARTPCA (\citealp{patterson2006population}), provides a more quantitative utility for removal of outliers by iteratively recomputing PCs in the genetic data. The method assumes a set of unrelated individuals and uses the covariance-based genetic relatedness matrix to identify these individuals."

3. In the last paragraph of Methods the authors state that [the variance] is independent of samples i and j. The way the statement is written is not clear whether it refers to si,j or var(si,j). I had to read the supplement in order to understand that this statement was about the variance. Please clarify.

We have clarified this statement with an equation reference to make clear to readers that we are referring to the variance (equation 3 in revision) .

4. In order to better understand the rejection criterion of the null, it would be helpful if the authors stated what the value range of their similarity index is; max(si,j) implies that only very similar individuals matter. What about the highly dissimilar individuals (possibly reflecting population sub-structure as nicely stated in the introduction)?
5. I am not sure I understand why the authors chose to use the Kolmogorov-Smirnoff test for normality when in section 2.3 they provide a proper test for heterogeneity? Also, is the admittedly conservative nature of the Kolmogorov-Smirnoff test a desirable "flaw" for this particular analysis? Please provide some comments on these issues.

To address comments 4 &5, we have expanded section 2.3 ("Tests of Heterogeneity").
The reviewer points out that there are many ways in which we can test for violations of homogeneity and that no way will be appropriate for all scenarios. In our manuscript, we use both the max(si,j) and the Kolmogorov-Smirnov statistic, but as the reviewer points out, readers would benefit from greater insight into each approach.
Essentially, we propose two methods for testing heterogeneity which address the complex nature in which it can arise. We recommend (1) the use of a simple Bonferroni approach for investigating cryptic relatedness, which will manifest itself as a small number of "right-tail" events, and (2) the use of the KS test for testing general population structure, which may appear in a wide range of forms and is likely to reveal itself as a large number of small deviations from the mean. We have also revised the corresponding sections in the text to clarify these issues.

Minor comments:
1. In the Abstract section, please remove the word "Abstract" after Motivation.

Thank you. We have fixed this issue.

2. The reference style does not seem to be appropriate, as most brackets seem to be consistently missing. Please amend.

Thank you. We have fixed this issue.

3. In Introduction, paragraph 4, there's an extra "for" in the first sentence. A bit further down, the word "recently" is also repeated twice in the same sentence.

Thank you. We have fixed this issue.

4. Please make some improvements in the mathematical notation. My understanding is that I[…] in expressions (1) and (2) is an indicator variable followed by its condition? There should be a comment explaining this notation somewhere in the

text. Also, please separate with commas or with ifs the values that wk takes from their corresponding conditions in expression (1).

As recommended by multiple reviewers, a note has been adding which defines the I[...] as the indicator function, evaluating to 1 if the condition is true and 0 if the condition is false. "Ifs" have been added to the function for w_k.

5. In section 3.2, second paragraph, please provide the missing year in the reference "(????)".

Thank you. We have fixed this issue.

6. As a general remark, I recommend a careful proofreading to spot/correct further typos.

We have carefully rechecked the manuscript both ourselves and with the help of external proofreaders.

Reviewer: 2

Comments to the Author
Schlauch propose a method to identify in gwas individuals who are related or population outliers. Specifically, they propose a measure of pairwise similarity based on identify by state between rare variants, then a statistical test for determining if this is significantly higher than expected. Overall, the paper is well structured and written. The method they propose seems sound. I have not tried their software package (STEGO) but as the method is straightforward, I would be surprised if it did not work, and their results indicate it is fast, especially relative to PCs. They have performed a simulation study and verified power of their test, as well as applied to real data from 1000G genomes. Moreover, they propose a way to distinguish between familial relatedness and pop structure - I'm not entirely sure how they separate these two phenomena (as in my mind they are indistinguishable), but this is a nice aspect if correct. My main (although not too severe) criticism is that it would be nice to get a better idea how necessary such a method is - my instinct is that in a gwas, one you remove obvious relatives or outliers, whether you filter out a couple of medium relatives or not is not too impactful on results.

Major comments

As I say, ideally for me (as I generally view relatedness as a nuisance variable which hinders gwas) it seems it would be easy to demonstrate the impact on association testing of using your method. I don't think required, but would be nice, say, to show the difference in p-values from your method compared to say just using 5 or 10 pcs or LMM - while the correlation is bound to be very high, does it have much impact, say, on which SNPs are declared associated?

We appreciate the interest in quantifying the specific effect on GWAS of improved genetic similarity matrices. As described in Mathieson 2012, the greatest impact of confounding due to fine scale population structure is in rare variants and believe that the greatest benefit of our method will be in that realm, as opposed to with SNPs. We feel that STEGO is important for detecting those subtle changes, but that the suggested analysis, while important, is outside the scope of this manuscript. However, we have added a simulation study in this revision which highlights the comparative abilities of STEGO vs PCA in separating subtle population stratification. Section 1.7 in the supplementary materials now illustrates the superior ability of STEGO in this context, which is important for properly carrying out gwas on these populations. Furthermore, we have mentioned this demonstration as future work in the discussion.

"Future work will include quantification of the specific gains achieved in controlling type I error and power in the context of rare variant association studies. Higher resolution population structure is always preferred, though the exact gains achieved in GWAS remain to be quantified."

As you propose using rare variants only, I think you should test sensitivity to MAF cutoff - ie does it matter too much if you include more common SNPs or reduce the threshold?

The reviewer suggests demonstrating that rare variants are more informative of fine scale ancestry than common SNPs. We agree that this is an important for the motivation of STEGO and greatly improves the basis for the method by demonstrating the increased information contained in rare variants.
To address this, we have added a new section to the supplementary materials, 1.6: Ancestry informativeness by allele frequency. To summarize: in this section we separate variants in the 1000 Genomes Project by MAF intervals and measure the jaccard similarity between individuals using only those variants. As expected, Supplementary Figure 5 clearly

demonstrates that the ratio of within group similarity to across group similarity increases as MAF decreases, with the notable exception of the lowest MAF interval, which is likely caused by QC limitations for such low frequencies.

Minor Comments

Page 2, Line 18 - Patterson, Price and Reich 2006 Plos Genetics proposed a test for significant eigenvalues (using the Tracy Widom distribution) which they suggested as a way to determine how many pcs to include in the regression. While this test never really caught on, as most prefer to "eyeball" pc plots, it has a lot in common with your method and I think deserves mention.

We have now acknowledged Patterson et al and the SMARTPCA software in the 3rd paragraph of the introduction. See response to Reviewer 1, Major Comment 2.

Page 2, Line 12 - would be nice to provide a reference

To support our claim: "...While the first approach is computationally fast and easy to implement, the direct modeling of the dependence structure between study subjects can be more efficient." We have cited Mathieson and McVean, 2012, who demonstrated that certain spatially structured phenotypes and genotypes lead to inflated type I errors when using standard PCA and LMM approaches.

The key equation of the paper is (1) - I found this a bit hard to define. To start with, it makes sense to define (say) T = sum_l=1^2n Gik rather than use this clunky expression repeatedly. Also, it would be nice to make clear here how rare variants are weighted more highly (because they have smaller denominator when computing wk).

As other reviewers commented, the key equations would benefit from greater explanation and insight into the effect of wk. We agree with these comments and have made enhancements, which include expanding methods section 2.1, defining our variables more explicitly and providing intuition for the use of the weight parameter (with more in the supplement). Additional improvements are described in more detail in our response to Major Comment 2 from Reviewer 3.

On this note, could you provide a plot (maybe supplement) showing how weight varies with MA count? Does it closely follow the standard 1/variance or some other obvious relationship?

We have created a plot demonstrating the relationship between w_k and minor allele count and added it to the supplement (Supplementary Figure 4).

Page 4 - "Given that STEGO weights more highly these rarer alleles, there is increased sensitivity to detection of relatedness" I don't see how this is necessarily true - yes, STEGO weights rarer alleles more highly, but I don't see how increased sensitivity necessarily follows

Please see our response to Major Comment 2 and the new supplementary section 1.6 which now addresses observed ancestry informativeness by allele frequency.

Page 3 - For example, in an otherwise homogeneous study group of unrelated individuals a pair of cousins ($\varphi = .0625$), with MAF ...
I think this is a good example, but related to my above point, how variable is 2.19 to allele frequencies of SNPs used in equation 1?

As noted by the reviewer, the value 2.19 cited in the text is a function of the allele frequency distribution. Equations 1,5 give the form of this calculation the new Figure 4 in Supplementary Materials now help establish to the reader that the greater prevalence of low frequency variants increases the expected value for s_ij in the presence of relatedness.

As you recommend using LE SNPs, do you have a recommended pruning threshold? As you are using rare SNPs, I imagine pruning has relatively little effect and only really removes duplicates? (Because correlation falls off fast for rarer variants)

As the reviewer notes, correlation falls off fast for rare variants. This mitigates much of the concern over LD relative to methods which use more common SNPS. With this in mind, we chose a sampling approach, described in more detail in

supplementary materials 1.4 which attempts to both preserve variant dependence and increase the relative value of each variant (by selecting those with low MAF).

Also Page 4, Line 62ish "promoted" seems a strange word to use

The wording has been changed to “increased”.

"Interestingly, not all related pairs belonged to the same population groups. We additionally discovered a pair of individuals, HG03998 from the STU"
Again, I think this is a good example.

My personal view is that the real data example is a bit too long - to me it serves only as a (nice) demonstration of the method, rather than providing particularly novel results. But then I'm a methods guy...

We thank the reviewer for the positive comment and add that we have reduced the Results section by moving section 3.1 to the Supplementary Materials.

Reviewer: 3

Comments to the Author
General comments
The statistic proposed in the manuscript is a clever way to incorporate allele frequency into a statistical test of relatedness. Parts of the paper would be improved by a careful re-write. (see minor comments below for some examples) It’s distracting to continually find syntax and grammar mistakes throughout a manuscript.

We thank the reviewer for the helpful and constructive comments. We have integrated each of the identified syntax and grammatical corrections as well as enlisted the help of external proofreaders to improve the quality of the manuscript.

Citations should probably be enclosed in brackets or parenthesis.

All citations have been formatted to be enclosed in parentheses.

Providing more details for the derivation of equations (perhaps in the supplement) would improve reader comprehension. This is only done for equation (2).

We agree that derivations help the reader understand our methods and have expanded the supplementary methods, which now include derivations of expectation of s_i_j and variance of s_i_j.

I would also appreciate more interpretation of the results instead of simply stating them (more on this below).

Please see responses to comments below.

Major comments

The method was motivated by using GWAS as an example but the method principally uses rare variants to identify recent shared ancestry. GWAS don’t tend to survey rare variants. And it’s not entirely clear what the end goal of this method is. It’s motivated with association studies but then doesn’t provide a meaningful way to incorporate the method into association studies (or compare with existing techniques of incorporating relatedness into association studies like LMMs). Is it advocating for removing individuals exhibiting cryptic relatedness? I would argue this is a much too conservative method for correcting for population structure.

We apologize for this misunderstanding, but our goal was to develop a method for whole-genome sequencing studies (WGS). In contrast to genome-wide association studies (GWAS) which focus on common variants, WGS data consist predominantly of rare variants. To avoid this misunderstanding in the text, we have revised the manuscript accordingly. Our example, the 1,000 genome dataset, is also whole-genome sequencing and not GWAS. Our approach can be used to detect

outliers in whole-genome sequencing studies as well as it can be used to adjust the association analysis for confounding due to population substructure (please see new simulation studies.). As current methodology for outlier detection/adjustment for population substructure focuses on common variants, we have outlined the reasons and motivation for the development of approaches that can handle/take full advantage of rare variant data/sequencing data.

The advantage with our method in regards to GWAS is in the superior identification of population structure and cryptic relatedness compared with existing methods. We agree with Reviewer 3 that most current GWAS don't tend to survey rare variants, but point to two critical reasons why our method is important in GWAS

1) As the availability of sequencing data grows with rapid decreases in cost, the focus on rare variants can be expected to rise.
2) Even in the context of common variant GWAS, it is beneficial to obtain the most precise quantification of population structure so as to limit the impact of confounding due to population structure.

For these reasons, we assert that the use of rare variants in our method will improve both common and rare variant GWAS.

As the reviewer points out, there are many approaches appropriate for handling shared ancestral history and cryptic relatedness once discovered, up to and including removal of samples. We do not advocate any particular method as the specific course of action will depend on the specifics of the study. But we wish to point out that the two most common approaches, PCA and LMM, each involve the estimation of a genetic relatedness matrix such as the one we have proposed here. Use of our relatedness matrix in these methods will serve to exploit the differential value obtained in less frequent variants that STEGO is designed to capture.

These are important considerations which require elaboration in our manuscript, and we have expanded our discussion section (particularly the addition of paragraph 2) to address them.

The readability of the manuscript could be greatly improved if you had an example to accompany the Methods section. Related to this, it would be helpful to explicitly define each variable you use in equations and explain the intuition behind the fomulas. E.g. "w_k is a weight which increases monotonically with the major allele count for site k". Is I in (1) the indicator function? It might be more clear to define the conditions of (1) in a discrete manner as you defined G as a binary matrix (unless, of course, G_{i,k} can actually take values in \mathbb{R}).

We agree with Reviewers 1 and 3, that readability is improved if we more explicitly define our variables.
1) Section 2.1 now includes sentences describing and interpreting w_k as well as an explanation of the I() function.
2) The supplement now includes an example distribution of the weight parameter.
3) As suggested, we have added an example of the calculation in (1) to the supplement which we believe to be very helpful to readers in understanding both the motivation and implementation of our method.

The assumptions required to obtain the functional form of $s{i,j}$ are not well supported in real data. And violations of this assumptions occur throughout the paper. E.g. "Intuitively, we can imagine two subjects which have a kinship coefficient, ϕ, indicating a probability of a randomly chosen allele in each person being identical by descent (IBD)." This seems true for a randomly selected allele, but alleles don't randomly sort by position. I understand that a method to reduce confounding introduced by LD is to subsample variants, but LD patterns change by population and no guidance is offered in the text on how, specifically, this was done.

The reviewer points out the independence of variants assumption in Section 2.1 with respect to LD.
LD patterns change by population and there is no "one-size fits all" method for LD pruning across multiple populations.
We have performed LD pruning based on subsampling informative (low frequency) variants across uniform blocks across the genome. We employed this approach in part because of the fact that due to their low frequency, the R^2 value between two nearby variants is substantially reduced.
To address this comment, we have expanded our section describing this procedure in Supplemental Materials 1.4.

It's impossible to get a sense of the functional form of the runtimes and how they scale with the number of sample or variants by looking at the table at the end of 3.1. I doubt the 30 seconds extra wait time going to prohibit someone from running PCA vs STEGO. Maybe increase the number of samples or variants and show this in a graph. No caption or label on table.

We have expanded this section in the supplement and run many more simulations to evaluate the runtimes. In addition to the cor() and princomp() functions, we have also added prcomp() for comparison, which uses an SVD instead of an eigendecomposition (as in princomp). We have replaced the table with a graph to more effectively convey the runtime of each method and how they scale. We agree with Reviewer 3 that this improves the readability of the comparison. (Please note that we have fixed a typo in the previous version that listed the number of simulated SNPs as 1,000,000 instead of the correct 100,000.)

How did you select the 100,000 variants in the 1KGP data? LD varies within and across populations and cannot be characterized by discrete intervals.

Please see response to comment 3 above.

How do I interpret Table 1? How does this compare to previous studies? Are these results expected? How do these results compare to methods with use the entire data instead of a subsample? What if you test the known relatives in 1KG data? In general the results lack appropriate comparisons that would inspire confidence in the method as compared to alternatives.

We agree with the reviewer that a comparison with other results and methods are difficult, as STEGO analysis is heavily based on rare variants and results/comparisons are not straight forward. We have therefore de-emphazied Table 1 and moved to the Appendix.

(In regards to Figure 6) It's not clear that PCA wouldn't perform just as well by restricting the variants to low minor allele frequency variants.

To clarify, we have emphasized that the plots generated in Figure 6 were each generated from the same set of variants. Due to the filtering step (described in supplementary materials 1.4), variants used in both methods were predominantly low frequency variants.

Minor comments
Page1, line 36: Remove "Abstract" after "Motivation:"
Line 44 "1000 genomes project" -> "1000 Genomes Project"
Line 46 "1000 Genomes Project, our" -> "1000 Genomes Project data, our". In fact, this entire final sentence of the Results part of the abstract could use a re-write.
Consider revision: "For candidate gene studies and later genome-wide association studies (GWAS), genomic control was developed."
Consider revision: "…account so directly for the dependence at the model-level".
???? on page 4
"However, both approaches benefit if, prior to the analysis, study subjects whose genetic profile is very different from the other study subjects, e.g. "genetic outliers", are removed from the data set." This statement must be supported.
"Many methods for exist for"
Fix Page 3, line 20
Equation 3 runs into the paragraph to the right.
Fig 3 comes before Fig 2 in the text.
Table 1 is referenced twice in the text in regards to two different preparations of the data. Which does it actually refer to?

We thank the reviewer for the list of minor comments and have integrated each of the above suggested changes into our manuscript, which has improved it's readability.