Author’s response to reviews

Title: Validation of genotype imputation in Southeast Asian populations and the effect of single nucleotide polymorphism annotation on imputation outcome

Authors:

Worachart Lert-ithiporn (worachartlert@gmail.com)
Bhoom Suktitipat (bhoom.suk@mahidol.ac.th)
Harald Grove (harald.gro@mahidol.ac.th)
Anavaj Sakuntabhai (anavaj.sakuntabhai@pasteur.fr)
Prida Malasit (prida.mal@mahidol.ac.th)
Nattaya Tangthawornchaikul (nattaya90@hotmail.com)
Fumihiko Matsuda (fumi@genome.med.kyoto-u.ac.jp)
Prapat Suriyaphol (prapat.sur@mahidol.ac.th)

Version: 2 Date: 20 Oct 2017

Author’s response to reviews:

Dear Dr. Matteo Pasini,

We are pleased to resubmit for publication the revised version of manuscript number MGTC-D-16-00287 “Validation of genotype imputation in Southeast Asian populations and the effect of single nucleotide polymorphism annotation on imputation outcome”. We have revised the manuscript in keeping with all criticisms and comments raised by the reviewers. Responses point-by-point to reviewer’s comments are as follows.

We would like to thank the reviewers again. We greatly appreciate the reviewers for their complimentary comments and suggestions. We have thoughtful addressed the reviewer’s comments and feel the manuscript is substantially strengthened. We hope that you find our responses satisfactory and that the manuscript is now acceptable for publication. We are looking forward to a favorable response in this regard.

Sincerely Yours,
Author Responses to Reviewer Comments

Jodie Painter (Reviewer 1):

The authors have done a thorough job addressing all reviewer comments and I find this manuscript much improved. It is now very much clearer, and a nice concise study of issues with imputation in populations that are not well represented amongst reference panels. This will surely be a good reference for others seeking to impute SouthEast Asian datasets in the future (sorely needed given the Euro-centric nature of published GWAS so far).

I just have a few minor comments, mostly to do with language for clarity, although in my final point (with regards to manuscript lines 297-300). I outline why I think the authors are slightly off track with one of their reasons for wanting to impute data, and that this would be better if it was more directed to genome representation rather than statistical power or genotyping cost.

line 78 - 'variants' rather than 'variance'

Response: The sentence was changed as suggested.

line 117 - should this read 'whole genome genotype data'?

Response: The sentence was changed.

line 118 - could you define what you mean by 'gene'(and remove 'the' from in front of 'genes'. e.g. is there a cut-off in terms of base pairs as to how far beyond the exon/UTR boundaries you are extending this analysis?

Response: The description of the gene and their boundary was in “SNP selection and annotation” section. In brief, we defined the gene boundary based on Illumina sample sheets (Illumina Inc., San Diego, CA) and the NCBI database of genetic variation.
line 154 - remove the period after 5Mb

    Response: The period was removed as suggested.

line 213 - 'with ID slightly higher' rather than 'and that ID was slightly higher'

    Response: The sentence was changed as suggested.

lines 243 and 244: change to 'while complex regions, where SNPs have been associated with more than one gene'

    Response: The sentence was changed as suggested (line 240-241).

line 260 - 'complex regions', rather than 'the complex region'. Using 'the' in this context reads as though there was only 1. Do this throughout.

    Response: All relevant sentences were changed as suggested.

line 261 - 'UTR regions', rather than 'the UTR'. I would also write UTR out in full as this is the first time you've mentioned it. Do this throughout.

    Response: All relevant sentences were changed as suggested.

line 264 - replace 'got a lower MAF' with 'had even lower MAFs after imputation'? I'd also change the above line to read 'SNPs with low initial MAFs'

    Response: The sentence was changed as suggested (line 261).

line 267 - remove 'the' from in front of 'most imputed SNPs'

    Response: The sentence was changed as suggested (line 264).

line 271 onwards - 'Coding regions' instead of 'The coding region' - do this throughout
Response: All relevant sentences were changed as suggested.

line 290 - 'are' instead of 'is'

Response: The sentence was changed as suggested (line 286).

line 297 and 298 - it's the number of samples, not the number of SNPs, that influences statistical "power".

Response: The sentence was changed (line 293-294).

I still don't agree with the points you raise from here to line 300. I'd re-write the third point to indicate you want good representation across the genome. Also, There is a physical upper limit to the number of SNPs included on chips (especially now that imputation exists there is no incentive for companies to keep adding more and more SNPs, unless it's for custom designs). Depending on where you buy chip prices for denser arrays can be comparable to costs for less dense arrays. I still don't think this is an argument for imputation. You want imputation for full genome coverage, not to win out on chip costs - to get full coverage you'll need to impute regardless of which chip you use/how much you pay!

Response: The paragraph was revised as suggested (line 294-296).

Ali Torkamani (Reviewer 2):

The authors state in response to my concerns that different SNPs were selected per population for the imputation comparison. In my opinion, this makes any comparisons invalid. Each SNP has differing potential for imputation depending upon LD with other SNPs etc - comparison of imputation accuracy across differing sets of SNPs will likely lead to fluctuations in imputation accuracy that are related to the SNPs selected rather than the populations themselves. A similar criticism could be made regarding the 1000G vs HMII imputation results. These potential biases make it unclear whether the specifics of the presented results are valid.

Response: We thank Dr. Torkamani for his insight in using different SNP sets to evaluate the yield and accuracy across population. To address this problem, we have redone the imputation evaluation by selecting a new SNP set and used this set to evaluate the yield and accuracy across
the population and reference set. The results have been incorporated into the revised manuscript in the Figure 1 and 2. The method and result sections were rewritten and checked with the new result. The result is in consistent with previous result.

We have shown in this revised manuscript that choosing a fixed set of SNP or “randomly” sampling SNP sets yielded similar conclusion. Using PanSNPdb samples, the 1000G provided higher yield but lower accuracy. Thai genotypes showed the highest accuracy over other populations in both HMII and 1000G panels. The Philippine’s genotype also had lowest accuracy when compared to other population. These results could be validated in both randomly and fixed SNP.