Reviewer’s report

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

Version: 1 Date: 14 Jul 2017

Reviewer: Jodie Painter

Reviewer's report:

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'

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

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?

line 154 - remove the period after 5Mb

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

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

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

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.

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

line 271 onwards - 'Coding regions' instead of 'The coding region' - do this throughout

line 290 - 'are' instead of 'is'

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

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!

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

Yes

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Yes

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

Yes

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

**Quality of written English**
Please indicate the quality of language in the manuscript:

Acceptable
**Declaration of competing interests**

Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license ([http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/)). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal