Author’s response to reviews

Title: Clinical Utility of the Low-Density Infinium QC Genotyping Array in a Genomics-based Diagnostics Laboratory

Authors:

Petr Ponomarenko (ponomare@usc.edu)
Alex Ryutov (aryutov@chla.usc.edu)
Dennis Maglinte (dmaglinte@gmail.com)
Ancha Baranova (abaranov@gmu.edu)
Tatiana Tatarinova (tatiana.tatarinova@usc.edu)
Xiaowu Gai (xgai@chla.usc.edu)

Version: 1 Date: 19 Jul 2017

Author’s response to reviews:

Dear Editor

Please find enclosed our revised manuscript “Clinical Utility of the Low-Density Infinium QC Genotyping Array in a Genomics-based Diagnostics Laboratory”. We have prepared a revised manuscript and point-by-point response to reviewers’ comments.

Please address all correspondence concerning this manuscript to Tatiana.tatarinova@gmail.com

Thank you for your consideration of this manuscript.

Sincerely,

Tatiana Tatarinova, PhD
Reviewer reports:

(Reviewer 1):

Comment 1. The first major issue is that the manuscript was poorly written. In page 6, it says "we first extracted genotyping calls of 664 individuals from the Infinium QC, the OMMI..." In page 7, it says "we genotyped 48 DNA samples." In page 8, it says "we obtained from illumine the infinium qc data of 503 individuals previously also studied by the 1000 genomes project." How many individuals did the authors use for comparison?

Answer 1.

We used multiple publicly available and proprietary datasets in this study. For example, 48 clinical samples were genotyped by us with the Infinium QC array at the Center of Personalized Medicine, CHLA. A separate Infinium QC dataset of 664 individuals was obtained from the Illumina website (https://www.illumina.com/). These 664 samples are a part of a larger set studied by the 1000 Genomes Project. Also, different types of genotyping, array (e.g. Affymetrix) and Whole-genome sequencing (WGS), data can be obtained only for different number of 1,000 Genomes samples.

We thank the reviewer for identifying this a source of potential confusion. We have modified the manuscript to clarify this. We also presented a detailed description of datasets in the supplement.

Comment 2. The second major issue is the concordance analysis. How many markers are not shared among the platforms? When we use Infinium QC Array, we don't know which markers should be excluded or not. If there are huge difference in excluded markers between the platforms, even though there are highly consistent concordance for the shared markers, it is difficult to say the platforms have similar potential utility.

Answer 2. The number of markers shared between the platforms are in the Supp. Table 1. Thanks again the reviewer for this very important comment. We have included the list of excluded markers is in supplement, and state on the page 8 of the manuscript. Overall, only 319 markers were excluded., which represent a small fraction of total SNPs (2%) on the chip. They are listed in the Supplementary File 2.

Comment 3. The third major issue is prediction analysis about gender, ethnicity, gps, family relationship. The 1000 genome was used as gold-standard. When the training model was built on the Infinium QC data, were the testing set independent of the training set and blind? Please
provide information about data split for training set and testing set, and the prediction performance such as sensitivity and specificity for each group.

Answer 3. We did not build or train the model, so there was no training or testing set. We used the existing tools with default parameters.

Comment 4. The minor issue: 1000 genomes project is not a platform. Therefore, please provide data source when the authors compare Infinium QC Array-24 with 1000 genomes project.

Answer 4. We have description of 1KG data and links to download it in the Materials section on page 5. Specifically: «1000 Genomes Project (1KG) dataset was downloaded from EBI (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502) and for related individuals from (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/related_samples_vcf). It contains genotypes of 2,504 individuals merged from multiple sets of genotyping and NGS data experiments, and is considered a gold standard. The family information was extracted from the pedigree file available on the 1,000 Genomes website (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130606_sample_info/20130606_g1k.ped) [2]. »

Reviewer 2

Comment 1.

>> Below are some typos/errors I found that should be corrected before publication:

1. In page 6, line 27, "sample identify" should be "sample identity". fixed

2. In page 9, line 13, there is a Word Reference error: "Error! Reference source not found.". fixed

3. In page 9, line 13, "These types if sampling pairs" might be "These types of sampling pairs". fixed

4. In Figure 3, proper axis labels should be added. We replaced this figure with the text explaining the kinship calculation.

5. In page 10, line 56, "Suppl. Figures 5-10" is inconsistent with the labels of supplementary figures. fixed
Answer 1. Thank you very much for these suggestions – the typos are fixed.

Comment 2.

In page 8, line 27 to 34, the authors considered markers that are discordant between Infinium QC and 1000 Genomes as "under-performing markers" and recommended excluding them for further analysis. However, 1000 Genomes data is not an ideal gold standard due to several reasons such as mapping issue, nearby indels. The discordance between Infinium QC and 1000 Genomes does not necessarily mean the genotype calls of those markers are incorrect in Infinium QC. The authors should consider showing performance improvement before VS after such filtering.

Answer 2.

The reviewer is correct. The discordance between Infinium QC and 1000 Genomes Project data does not necessarily mean the genotype calls of those markers are incorrect in Infinium QC. We defined these underperforming markers as the ones that were consistently discordant between different platforms across at least 10% of samples. They are more likely therefore to provide erroneous genotyping calls. Most importantly, for our stated QC purposes (kinship determination, gender determination, etc.), they are safer to remove or not included in the analysis. Exclusion of 319 markers out of 15K represent a small perturbation to the dataset.

Comment 3.

In page 8, line 50, the authors should consider to report the p-value since they stated that they used Kolmogorov-Smirnov statistics to determine the significance in page 7, line 6.

Answer 3. Thank you very much for this suggestion. We agree, and we have added an additional supplemental file to show KS statistics and p-values.