Reviewer’s report

Title: Spectrum of mutations in monogenic diabetes genes identified from high-throughput DNA sequencing of 6,888 individuals

Version: 0 Date: 17 Jun 2017

Reviewer: Martine Vaxillaire

Reviewer’s report:

In their paper, Vikas Bansal and colleagues have screened a large cohort of individuals with a clinical diagnosis of type 2 diabetes (including 1346 individuals diagnosed before the age of 40 years) and controls (n=2872) for deleterious mutations in diabetes associated genes by performing high-throughput targeted sequencing from pooled DNA samples in three stages (in total 225 genes have been assessed in the study, and only 22 genes were assessed in this study). Although this sequencing strategy of DNA pooling is indeed highly cost-efficient to assess low-frequency genetic variants in a large cohort (~7,000 samples), there are some obvious limitations to use such a process as it becomes far less informative and even problematic for the detection of (very) rare or singleton variants which represent a large amount of the genomic variation that has to be assessed for disease association. This issue has to be balanced as regards the high quality and high efficiency of the sequencing technology used, as well as both overall and mean coverage rate of the targeted gene sequences.

From that point, another important issue of the paper is the relevance and accuracy of the global findings from this study, as it is reported in the conclusion. It is difficult to judge and comment the % of mutation carriers as given in the discussion/conclusion, which may be totally erroneous with regards to the poor sequencing data.

Moreover, these findings should be carefully discussed with regard to previous studies published in this field, in particular the recent reports from Flannick et al. (Nat Genet. 2013) and Najmi et al. (Diabetes 2017).

Comments on the paper:

- Introduction: the following sentence is quite misleading "Rare variants in monogenic diabetes genes such as HNF1A, HNF4A and WFS1 have also been associated with risk for type 2 diabetes (11; 12; 13) highlighting the genetic overlap between monogenic diabetes and T2D." Indeed, it is not the same variants, and so not the same genetic event, that lead to T2D and to monogenic diabetes. Actually, the missense variants in HNF1A, HNF4A and WFS1 found to be associated with T2D risk are low-frequent or frequent variants (whereas those involved in monogenic diabetes are very rare and highly penetrant mutations).

- Methods: the mean age of controls included in the first set is rather low (44.2 years of age) for a standard design of T2D case-control study. Otherwise, one can expect a high level of
heterogeneity in the clinical, metabolic features of the T2D cases (given the large number of diabetic individuals included in the study, and the wide range of age of onset for the diabetic patients).

- Methods Table S4: the clinical data shown in this table are quite poor. The authors should at least show the mean BMI and the mean fasting glucose levels (and/or glycated hemoglobin A1C) in cases and controls (as these parameters are key in the study of T2D).

- Results: importantly, the quality of the sequencing is very poor. Indeed, it is mentioned that about 79% (SD??) of the target was "well covered" according to the authors, which means that there was potentially more than 20% of the target sequences with false negative results. The authors should show more QC data as the present data are very worrying, knowing that the current targeted DNA-seq protocols (via NGS) [including whole-exome sequencing] enable an accurate sequencing of more than 99% of the target. From the data of Table S1, several target genes are very poorly covered in one or both sequencing stages (for ex. HNF1A, PDX1, INS, GATA6, NEUROG3).

- Results: this reviewer believes that the strategy of DNA pooling is not appropriate for the study of T2D as no adjustment is possible while sex, BMI, age and ethnicity have a huge effect on T2D association signals (even more for rare variants). This issue would not have been problematic if cases and controls were being matched for age, BMI, sex and ethnicity.

- Results: The authors should cautiously revise the numbering of the tables (for ex. for Table 1, 2 and 3, Table S5 and S6 ; Table 5 does not appear in the paper...). In table S5, there are 5 « recessive monogenic diabetes genes » (and not six genes as mentionned in the text on page 15).

- Discussion : the sentence « none of these participants fulfilled classical Tattersall criteria of monogenic diabetes mellitus » (on page 19) is somewhat fuzzy. The authors should clarify which are the clinical criteria defining these individuals with early onset diabetes (as they are mentionned as « non-MODY ») ; or the fulfilled clinical criteria could be included in one supplementary Table.

- In the tables, for each reported mutation, the gene transcript ID, the dbSNP ID, the MAF in ExAC and/or ESP should be systematically shown; as well as the in silico functional predictions for missense variants (following at least 3 different software programs - not only PolyPhen).

- All gene symbols have to be written in italics throughout the manuscript.

- The official nomenclature for mutation annotation should be used throughout the manuscript.

- The gene name KLF14 should be replaced by KLF11 (as a MODY related gene), or at least this should be clarified throughout the manuscript.

- References 52, 53, 54 mentionned on page 20 are not in accordance with the text.

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

No

**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.

No

**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 that 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/). 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