Referee reports

Catherine Hastings

4th September 2017

1 A probabilistic model to predict clinical phenotypic traits from genome sequencing (Chen et al., 2014)

Summary

Using the CAGI (Critical Assessment of Genome Interpretation) challenge as an incentive, the paper outlines a new Bayesian probabilistic model which aims to predict whether participants have a specific phenotype based on genotypic data. The Bayesian prior probabilities are calculated using a participant's age, gender and ancestry. The posterior distributions are calculated using known information of how single nucleotide variants (SNVs) relate to a given phenotype from multiple different sources. For the majority of phenotypes considered, the model here implemented does no better than random chance. However, predictions are much more accurate for phenotypes with a known genetic component. In addition, from the results of the CAGI challenge, it can be seen that this model is currently the most effective method of predicting phenotypes from genetic information.

Contribution

This paper is not the first to address the issue of how to predict phenotype from genetic information; multiple organisations have pipelines to allow the interpretation of genetic information. For example, the Personal Genome Project have developed the GET-evidence pipeline (Genome-Environment-Trait) (Ball et al., 2012), and companies like `understandyourgenome.com` and `23andme.com` trade on their ability to provide customers with an interpretation of the genome sequences. However, the method outlined here is novel due to the implementation of a Bayesian model. It also goes further than assessing simply the frequency and location of SNVs in affecting phenotype. Considered here are also SNVs with no current known effect, though located within a gene known to affect the phenotype. Additionally, the group's performance in the 2012 CAGI challenge justifies the publication of the methods they used.

Major comments

The model outlined is extremely complex, but every step and assumption is clearly outlined. Equally, the assumptions made seem reasonable. For example, supposing that data on SNVs taken from the Human Gene Mutation Database (HGMD) are generally more indicative of phenotype than SNVs located via genome-wide association (GWA) studies is logical due to the high level of curation in the (HGMD).

It is assumed that an area under ROC curve (AUC) greater than 0.7 shows that the phenotype is well predicted. There is no description given of why 0.7 was chosen for this threshold. Is it common for analyses of this type? The addition of a sentence justifying this decision would be helpful, perhaps citing other papers using similar methodology.

The data and how they are used are well-described. Equally, all data are available on request from the PGP. It is worth noting that there is likely to be an increase in the amount of data available for studies like this. I think this is likely to increase the utility of the methods outlined here.

p-values are calculated by performing permutation on the data, repeating the analysis and seeing how often random chance might account for the results achieved. In addition to this method providing p-values, it seems a rigorous way of checking that conclusions drawn are reflective of the data, and not introduced by this particular methodology.

Minor comments

Some of the terminology could be more clearly explained. For example, the term 'variant genotype' is often used synonymously with single nucleotide variant (SNV). This is not generally typical of other papers in this area. The term could still be used, but clarification should be supplied.

The text is well written and easily understood. The graphs are easy to understand with good captions, and add to the clarity of the arguments.

2 Improved fluorescence assays to measure the defects associated with F508del-CFTR allow identification of new active compounds (Langron et al., 2017)

Summary

The paper introduces two assays which aim independently to study the gating and the localization of F508del-CFTR. These assays are then used to identify compounds which overcome the defects caused by the F508del mutation. Both a corrector (to improve the localization of CFTR at the membrane) and a potentiator (increasing the open probability (P_o) of CFTR at the membrane) would be required to restore function.

Contribution

Previous studies have been performed introducing assays to quantify the localization of CFTR, reducing the impact of the CFTR-pHTomato assay described here. The results of the YFP assay are simple and easily understood, building upon the work of Galietta et al. (2001).

The identification of two new potentiators which increased the gating function of F508del is a significant finding. In particular, the potentiator which seem to improve the gating function even with prolonged exposure.

Major comments

In figure 6A, it appears that a third compound satisfies the requisite criteria of activity $> 50\%$ and SSMD > 2. Where any tests performed on this compound?

It is shown using single-channel patch clamping that the YFP tagged to the N-terminal of CFTR does reduce its gating function. Is it possible that the gating function of YFP-CFTR-WT was affected in a different way to that of YFP-CFTR-F508del? Could it be verified by performing patch clamping of CFTR-F508del with the addition of compounds that increasing its gating function? When shown in comparison with patch clamping experiments performed on CFTR-WT with potentiators, the effect of the YFP tag could perhaps be quantified.

The checks performed showing that the YFP tag has no effect on the localization of CFTR are convincing. It seems that no have been performed as to the effect of the introduction of the pHTomato gene, and the results of other studies have been relied upon. Some summary of what checks were performed in by these other studies might be appreciated.

The methods used to quantify the effects sought it good and provides the reader with a simple view results.

Minor comments

The introduction is well cited and gives a clear description of the aims of the paper. The methodology is described clearly and in great detail.

References

Madeleine P. Ball, Joseph V. Thakuria, Alexander Wait Zaranek, Tom Clegg, Abraham M. Rosenbaum, Xiaodi Wu, Misha Angrist, Jong Bhak, Jason Bobe, Matthew J. Callow, Carlos Cano, Michael F. Chou, Wendy K. Chung, Shawn M. Douglas, Preston W. Estep, Athurva Gore, Peter Hulick, Alberto Labarga, Je-Hyuk Lee, Jeantine E. Lunshof, Byung Chul Kim, Jong-Il Kim, Zhe Li, Michael F. Murray, Geoffrey B. Nilsen, Brock A. Peters, Anugraha M. Raman, Hugh Y. Rienhoff, Kimberly Robasky, Matthew T. Wheeler, Ward Vandewege, Daniel B. Vorhaus, Joyce L. Yang, Luhan Yang, John Aach, Euan A. Ashley, Radoje Drmanac, Seong-Jin Kim, Jin Billy Li, Leonid Peshkin, Christine E. Seidman, Jeong-Sun Seo, Kun Zhang, Heidi L. Rehm, and George M. Church. A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences*, 109(30):11920–11927, 2012.

Yun-Ching Chen, Christopher Douville, Cheng Wang, Noushin Niknafs, Grace Yeo, Violeta Beleva-Guthrie, Hannah Carter, Peter D Stenson, David N Cooper, Biao Li, Sean Mooney, and Rachel Karchin. A probabilistic model to predict clinical phenotypic traits from genome sequencing. *PLoS Computational Biology*, 10(9):e1003825, 2014.

Luis J.V Galietta, Peter M Haggie, and A.S Verkman. Green fluorescent protein-based halide indicators with improved chloride and iodide affinities. *FEBS Letters*, 499(3):220–224, 2001.

Emily Langron, Michela I Simone, Clémence Delalande, Jean-Louis Reymond, David L Selwood, and Paola Vergani. Improved fluorescence assays to measure the defects associated with F508del-CFTR allow identification of new active compounds. *British journal of pharmacology*, 174(7):525–539, 2017.