Characterisation and functional analysis of variations in the ADME genes in African populations

Overview or Project 2018

1. Introduction

The Human Health and Heredity in Africa Consortium (H3A) has sequenced approximately 350 full genome sequences of African individuals, which will be used for various projects within the consortium. Approximately 300 high-coverage genomes will be available for this joint project between H3A and GSK's Africa NCD Open Lab to mine the H3A data and other data that may become available for variation in the ADME genes in African populations, in order to understand possible functional effects of these variations.

It is GSK's Africa NCD Open Lab's intention that all the outputs resulting from this project should be put in the public domain for the benefit of the research community.

H3A comprises a number of autonomous projects, funded by the major funders (presently the NIH and Wellcome Trust). This project is proposed as a Consortium project, but will be administered and run by the AWI-Gen Project based at the University of the Witwatersrand, Johannesburg, as represented by the Wits Health Consortium (Pty) Ltd. The WHC will be the formal partner on this project, and responsible for financial and administrative management.

Project Principal Investigators (PIs)

Scott Hazelhurst (H3A AWI-Gen Collaborative Centre, University of the Witwatersrand)
Matt Hall (GSK)

This project will be done as a collaboration between GSK's Africa NCD Open Lab and H3A. GSK's Africa NCD Open Lab will support the project both financially and with scientific expertise.

2. Project Goals

The project has two complementary goals

(1) Building capacity for pharmacogenomics research in the H3A Consortium

(2) Mining the data generated by H3A Projects for pharmacogenomically relevant data, particularly in the ADME genes

We would like to extend the capacity of the H3A Consortium in translational science and a very important step toward this goal is the identification of pharmacogenomically relevant variants. The ADME genes are a class of genes that play an important role in the absorption, distribution, metabolism and excretion of drugs. Understanding the variation in African populations in these genes and identifying important candidate variations for future study will lead to scientifically interesting projects with significant translational impact. The importance of building pharmacogenomics capacity and resources for Africa has long been recognised

[Masimirembwa and Hasler 2013; Matimba et al 2008], and this project will accelerate work in the area through its research findings and capacity development.

Other projects within H3A that are exploring other genes of interest could collaborate with respect to computational pipelines and training opportunities.

The main questions are to

(1) Characterise variation in ADME genes in African populations, including understanding the relationship of African population sub-structure to variant frequency in ADME genes;
(2) Identify both common and rare ADME gene variants in African populations that may have a functional effect.
(3) Model a sub-set of potentially functionally relevant variants (e.g. protein modelling, docking studies)
(4) Develop capacity for pharmacogenomics in Africa

The data we intend to use includes (a) all full genome sequence data generated by H3A projects (as part of the chip design project); and Public data sets from African populations. Other H3A data sets may become available and we can explore the possibility of their use in the later stages of the project. For example, several H3A projects will be genotyping participants on the H3A SNP array as part of GWAS projects. The GWAS data from the AWI-Gen project will be available and other consortium partners may choose to contribute too.

The focus of the work will be the high coverage full genome sequence data that was generated for the H3A population study and was used in chip design. The quality and novelty of this data are outstanding and H3A needs to mine these data timeously, so there is some urgency and this collaboration with GSK will be invaluable in accelerating analysis. The use of additional data such as 1000 Genomes Project data and African Genome Variation Project (AGVP) data will also add value to understanding differences across African populations. We have had experience in genome data analysis from the work of the H3A population study and the chip design project and have developed strategies to merge data for studies with different levels of sequencing coverage.

We propose three components to the work plan: (1) Analysis of the data to answer the research questions; (2) The construction of portable and robust pipelines that can be applied to similar data sets by other groups; (3) Training.

Analysis

- mining the H3A population data to characterise and annotate ADME gene variation in the H3A population data (see [Hovelson et al, 2016] for examples of the analyses that could be done);
- this mining will explore all types of variations, but one sub-project will aim to analyse structural variants and particularly copy number variation, which are recognised as being important for specific gene loci in terms of drug metabolism [He *et al*, 2011].
- *In silico methods for* functional interpretation of variants (through understanding of drug metabolism and pharmacokinetics) leading to identification of candidate variants which are promising for further study;
- modelling selected candidate proteins to explore the potential functional impact of

variations;
- small pilot laboratory validation of selected candidate variants focussing on copy-number variants;
- exploration of GWAS data (both raw and imputed) both for replication and greater understanding of differences in African populations. We expect ~40k participants genotyped on the H3A chip. This can be used in particular for understanding frequency differences of common ADME genes across different African populations. Although the participants are well phenotyped, it is not clear whether any of the current phenotypes will help in this study.

Laboratory validation. This project is primarily a bioinformatics project. However, some targeted wet-lab work for validation will increase the impact of the work. There are several validations that could take place, including limited target capture resequencing of candidate genes of interest emanating from analysis of the H3A data (these would be from individuals who do not have WGS data). However, at this point, the most useful and productive validation appears to be the validation of CNVs found in the bioinformatics analysis. If during the course of the project it appears that more productive validation can be done, the parties may agree to change the workplan and outcomes.

Training

We will use the expertise within the H3A network and from GSK's Africa NCD Open Lab for capacity development. The training activities will include

- Mentorship in data analysis: we will use the model we developed for the H3A Population Study as a mode of working, with a core analysis team and volunteers from the consortium, and there will be some capacity for mentorship.
- A short online course on CNV analysis from WGS short reads. In recent projects (e.g., H3A Population Study and SAHGP), the CNV analysis was relatively superficial, partly because of lack of capacity for this analysis, so this will be an important contribution to H3A capacity development.
- A monthly journal club across the working group members (open to all H3A members)
- Data analysis and ideas-generating Jamborees. We plan on running two jamborees during the two-year project. This will be an opportunity for analysis of interim data, as well as short tutorials on analysis techniques.
- An internship for two younger researchers at the SBIMB.
- An internship for two researchers at GSK's Africa NCD Open Lab.

Pipeline development

Pipelines will be developed to assist with the analysis and as an outcome for future projects. In conjunction with the H3A Bioinformatics Network we will develop portable and robust analysis pipelines.

Resource development

An outcome of the project will be resources such as databases of ADME variation in African populations. We will cooperate with the Informatics Network for the housing of these resources in such a way that they are accessible to the broadest number of researchers.

3. Summary — outcomes

In summary, the core outcomes of the project will be:

- An analysis of variation in core ADME genes in African populations, including the production of a database cataloguing variation (including annotation of possible functional effects of the variation), a paper and pipeline for the analysis;
- Protein modelling of 6 key variants and interpretation of the impact of variation;
- Validation of copy number variants found in the bioinformatics data;
- Capacity building: two Jamborees, student graduations and increased expertise in the H3A Consortium.

References

Y. He, J. M. Hoskins, and H. L. McLeod, “Copy number variants in pharmacogenetic genes,” *Trends Mol. Med.*, vol. 17, no. 5, pp. 244–251, 2011.

D. H. Hovelson, Z. Xue, M. Zawistowski, M. G. Ehm, E. C. Harris, S. L. Stocker, A. S. Gross, I.-J. Jang, I. Ieiri, J.-E. Lee, L. R. Cardon, S. L. Chissoe, G. Abecasis, and M. R. Nelson, “Characterization of ADME gene variation in 21 populations by exome sequencing,” *Pharmacogenet. Genomics*, p. 1, 2016.

C. Masimirembwa and J. A. Hasler, “Pharmacogenetics in Africa, an Opportunity for Appropriate Drug Dosage Regimens: on the Road to Personalized Healthcare,” *CPT Pharmacometrics Syst. Pharmacol.*, vol. 2, no. 5, p. e45, 2013.

A. Matimba, M. N. Oluka, B.U. Ebeshi, J. Sayi, O. O. Bolaji, A. N. Guantai, and C. M. Masimirembwa, “Establishment of a biobank and pharmacogenetics database of African populations.,” *Eur. J. Hum. Genet.*, vol. 16, no. 7, pp. 780–783, 2008.