Using Alternate References for Transcriptome Analysis of an Indigenous Australian Study Cohort

Stevie Pederson[1,2], Yassine Souilmi[1,3] Hardip Patel[2], Alex Brown[1,2], Jimmy Breen[1,2]

[1]Black Ochre Data Labs, Telethon Kids Institute [2]John Curtin School of Medical Research, Australian National University
[3]School of Biological Sciences, University of Adelaide

Abstract

Alignment of transcriptomic data to the standard reference genome has recently been shown to be a problematic strategy.[1] This is particularly relevant for populations which are poorly represented within existing repositories that capture global diversity, and when these populations themselves are known to contain a poorly characterised amount of unique variation.

The use of an alternate reference genome which incorporates an appropriate set of variants has also been shown to improve alignments and quantifications at the gene level. We propose that we will incorporate variant sets from public repositories as well as a custom set of variants identified within our study cohort and assess the impact on gene quantification across each alternate reference genome. This is then extended into transcript-level quantification and the incorporation of user-defined sets of variants into an alternate reference transcriptome. In addition, a tool for creating custom reference transcriptomes has been developed.

Genomic Variability

The Indigenous Australian population contains a large amount of unique genetic diversity. Given that risk factors are increasingly being shown to be polygenic and dependent on the genetic background. ignoring diversity in a large cohort may limit our ability to address this contribution.

The historical relationship between researchers and the Aboriginal community has been problematic and as such, this population is poorly represented in databases such as the 1000 Genomes Project (1000GP), and the Human Genome Diversity Project (HGDP). Prepared in a highly consultative, and indigenous-led manner, the PROPHECY study present a unique opportunity to provide access to precision medicine for members of the Indigenous Australian community. The unique design of the study also presents a unique opportunity to include population-level diversity across all -omics layers in the study.

Variants from the 1000 Genomes Project

Taking the 1000GP variants as a proof of principle, an gene-level analysis (*STARconsensus*) was performed comparing the standard hg38 reference against one incorporating SNPs and InDels from the 1000GP.

The PROPHECY Study

The PROPHECY study (**P**reventing **R**enal, **OP**hthalmic and **H**eart **E**vents in **C**ommunit**Y**) consists of ∼1400 indigenous participants drawn from regional, remote and urban locations within South Australia. Amongst community, the study is colloquially known as the *Aboriginal Diabetes Study*.

The PROPHECY Study is a multi-omics study including genomic variants, DNA methylation, bulk RNA-Seq, proteomics, lipidomics, metabolomics and multiple other layers, all derived from blood samples taken from the same participants.

Haploid Reference Strategies

Stuff

Pan Genome Reference Graphs

Stuff

References

[1] Benjamin Kaminow et al. "Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses". en. In: *Genome Res.* 32.4 (Apr. 2022), pp. 738–749.