Individual Ancestry Estimation from Whole Exome Sequencing Data in Patient-Derived Xenograft Samples

Li Chen1, Meredith Yeager1, Biswajit Das1, Chris Karlovich1, Diane Palmieri1, Eric Karlins1, Vivekananda Datta1, Corinne Camalier1, Yvonne Evrard1, Melinda Hollingshead2, P. Mickey Williams1, James Doroshow3

1Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, 2National Cancer Institute at Frederick, Developmental Therapeutics Program, Frederick, MD, 3National Cancer Institute, Division of Cancer Treatment and Diagnosis, Bethesda, MD

Abstract

The NCI Patient-Derived Models Repository (PDMR; pdmr.cancer.gov) provides the research community a useful resource of patient-derived models (PDMs) from primary and metastatic tumor tissues to facilitate pre-clinical drug studies. However, the repository has limited patient demographic data for race/ethnicity. An estimation of individual ancestry using whole exome sequencing (WES) data, which is available for early-passage PDMs and some originating patient specimens, will therefore be needed to stratify the patient population when investigating associations between genetic variants and phenotypes of interest. Here, we present a workflow to infer ancestry of patient-derived xenograft (PDX) samples from WES data using SNPweights, an algorithm to infer genetic ancestry from single nucleotide polymorphism (SNP) weights precomputed from large external reference panels and further estimates the fraction (%) ancestry. We have tested the workflow on PDX and patient samples from 134 PDMs. Overall, the workflow has been shown to be accurate for inferring ancestry of the PDX samples for the PDMR project.

NCI Patient-Derived Models Repository (PDMR)

Overview

- A national repository of Patient-Derived Models (PDMs) to serve as a resource for academic discovery efforts and public-private partnerships for drug discovery.
  - Clinically-annotated, Early-passage, Molecularly-characterized Patient-derived Xenografts (PDXs),
  - Patient-derived tumor cell cultures and cancer-associated fibroblast cultures (PDCs) developed from primary or metastatic tumors and/or PDXs,
  - Models have sequence data for a subset of PDXs including a targeted gene panel, whole exome, and RNASeq

PDMR samples and genomics

- 869 PDX samples (from 151 models) have whole exome sequencing
- 35 patients have both originating specimens and PDX samples
- 54 (36%) patients have self-reported race/ethnicity information (Figure 1)

Figure 1. Break-down of self-reported race. Initial data collection did not include race/ethnicity (Not Provided).

Results on PDMR Data

A. Comparison of the inferred genetic ancestry to self-reported race

Table 1. Inferred ancestry for self-reported and not provided/unknown patients

<table>
<thead>
<tr>
<th>Self-reported race (n)</th>
<th>Inferred ancestry (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black or African American: 1</td>
<td>European (CEU): 1</td>
</tr>
<tr>
<td>White: 53</td>
<td>European (CEU): 52 Mixed: 1</td>
</tr>
</tbody>
</table>

- For inferred ancestry, when a patient originating sample was available it was used exclusively for assessment; else the average of the ancestry assignment from all sequenced PDXs was used.
- Using 80% as a cutoff to assign ancestry, 52 out of 54 (96%) patients from the PDMR data have concordance between self-reported race and inferred ancestry by SNPweights.
- Of the remaining patients, one showed admixed ancestry (all <80%) and another patient showed a different ancestry from self-reported race/ethnicity (Table 2).

SNPweights (Chen CY et al., Bioinformatics, 2013)

- Infers ancestry using raw genotypes from the target samples and a set of genome-wide SNP weights pre-computed using external reference panels
- Extended to infer % ancestry using principal components

Reference panel

- HapMap3: West African (YRI), European (CEU), East Asian (EA) and Native American (NA)
- 364,458 SNPs

Conclusion

- With self-reported race as a reference, the SNPweights workflow has been shown to be accurate for inferring ancestry of samples from the PDMR.
- SNPweights will be an important tool for estimating ancestry in cases where race/ethnicity of the patient is unknown.

Table 2. Inferred genetic ancestry output examples

B. Comparison of the estimated ancestry from originating specimens and PDX samples

Figure 3. Scatter plot of inferred ancestry of YRI, CEU, EA and NA for PDMR models having data from both originating and PDX samples

C. Summary of inferred ancestry for all patients

From a total of 151 patients, 133 patients (88.08%) were inferred as European (CEU), 7 patients (4.63%) were inferred as West African (YRI), 2 patients (1.32%) were inferred as East Asian (EA) and no patient was inferred as Native American (NA). Of interest, there are 9 patients (5.96%) showing admixture of ancestry (Table 2).