PRECISION HEALTH
UNIVERSITY OF MICHIGAN

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN
DATA OFFICE FOR CLINICAL & TRANSLATIONAL RESEARCH

Whole Exome Sequencing in the Michigan Genomics Initiative Cohort

About this guide

This guide provides information about the whole exome sequencing in the Michigan Genomics Initiative (MGI) cohort.

High Level Summary

DNA from consented Michigan Genomics Initiative participants underwent whole exome sequencing covering ~300K variants (single nucleotide polymorphisms and insertion/deletions).

Example Use Case

Researchers could examine genetic variation in protein coding regions of the genome.

<table>
<tr>
<td>What methods were used to sequence the specimen?</td>
<td>DNA samples from whole blood were prepared for whole exome sequencing as outlined by the Northwest Genomics Center (NWGC, University of Washington).

Sequence Batch 1
DNA libraries underwent exome capture using Roche/Nimblegen S eqCap EZ v2.0 (~36.5 MB target). NWGC's sequencing pipeline is a combined suite of Illumina software and other industry standard software packages (i.e., Genome Analysis ToolKit [GATK], Picard, BWA-MEM, SAMTools, and in-house custom scripts) and consisted of base calling, alignment, local realignment, duplicate removal, quality recalibration, data merging, variant detection, genotyping, and annotation. Variant detection and genotyping were performed using the HaplotypeCaller tool from GATK4 and hard filtering was performed (GATK v3.4). Exome completion was defined as having > 90% of the exome target at > 8X coverage and >80% of the exome target at > 20X coverage. Exome completion and several metrics including capture efficiency, raw error rates, and sample contamination validation were used for standard quality control assessment. A total of 323,867 variants (single nucleotide polymorphisms and insertion/deletions) passed standard quality control and were released to researchers.

<table>
<tr><th colspan="2">Sequence Batch 2</th></tr>
<tr><td>Sequencing Provider</td><td>Macrogen</td></tr>
<tr><td>Type of Read</td><td>Paired-end</td></tr>
<tr><td>Read length</td><td>151</td></tr>
<tr><td>Library Kit</td><td>Sureselect V5-post</td></tr>
<tr><td>Library protocol</td><td>SureSelectXT Library Prep Kit/SureSelectXT Target Enrichment System for Illumina Version B.2, April 2015</td></tr>
<tr><td>Sequencer</td><td>Illumina</td></tr>
<tr><td></td><td>Base calls are converted into FASTQ using Illumina bcl2fastq.</td></tr>
<tr><td>Alignment</td><td>Gotcloud</td></tr>
<tr><td>Variant calling</td><td>Gotcloud</td></tr>
<tr><td>Filtering</td><td>Variants were filtered using vcftools using an iterative procedure to remove poorly sequenced variants and individuals. We restricted analysis to variants with a minimum depth of 10, removed individuals with a missingness rate >0.4, filtered to variants with a maximum missing rate of 0.2, and finally to individuals with a missing rate of 0.1. The final dataset included 615 individuals.</td></tr>
</table>
</td>
</tr>
</table>

In what format is the data available?	The data are available in BAM files (*.bam). These files are compressed binary versions of a tab-delimited text file that contains sequence alignment data.
How do I access the data?	Once the cohort is createdthrough DataDirect, please put in a request for genetic data through the Precision Health Research Scientific Facilitators.
How do I cite the data?	Clinical Implications of Identifying Pathogenic Variants in Individuals with Thoracic Aortic Dissection (https://pubmed.ncbi.nlm.nih.gov/31211624/)
Where can I get help with genetic data?	Contact the Precision Health Research Scientific Facilitators at PHDataHelp@umich.edu