# MCCRD-SOP0058: Individual Ancestry Estimation from Whole Exome Sequencing Data

Effective Date: 7/3/2020

**Please check for revision status of the SOP at**

https://pdmr.cancer.gov/sops/

**PDMR** **NCI Patient-Derived Models Repository**
An NCI Precision Oncology Initiative$^{SM}$ Resource

**TABLE OF CONTENTS**

## VERSION INFORMATION

1. Change History

| Revision | Description |
|---|---|
| | Internal SOP used by MOCHA Laboratory |
| 7/3/2020 | Standardize SOP for posting to PDMR internal site for use by designated NCI intramural laboratories |

2. Related SOPs

| |
|---|
| MCCRD_SOP0011: Whole Exome Sequencing Data Analysis Pipeline and Specifications |

3. References

[1] Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. Bioinformatics. 2013;29(11):1399-1406

[2] Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. Bioinformatics. 2013;29(11):1399-1406

[3] https://github.com/mathii/gdc/blob/master/vcf2eigenstrat.py

## 1.0 PURPOSE/SCOPE

This Standing Operating Procedure (SOP) describes procedures for estimating individual ancestry using whole exome sequencing (WES) data for reporting in the NCI Patient-Derived Models database as performed by the Molecular Characterization Laboratory (MoCha) at the Frederick National Laboratory for Cancer Research. **This SOP is for research-use purposes only; do not use for clinical sample analysis.**

## 2.0 DESCRIPTION OF INDIVIDUAL ANCESTRY ESTIMATION

**2.1** The processed bam files are generated using whole exome sequence (WES) data following the WES data analysis pipeline in the SOP MCCRD_SOP0011.

**2.2** VCF files are generated using samtools mpileup on 364,458 SNPs using the SNPWeights algorithm[1].

**2.3** Ancestry information is estimated using SNPWeights for each PDX sample which outputs the fraction ancestry of four populations: West African (YRI), European (CEU), East Asian (EA), and Native American (NA)[2].

**2.4** Patient-level ancestry is determined based on the priority of available source material:

**2.4.1** If germline WES is available, it is used exclusively for assessment;

**2.4.2** Else, if WES from the originating patient sample is available, it is used exclusively for assessment;

**2.4.3** Else, the average of the ancestry assignment from all sequenced PDXs are used.

**2.5** One of the four populations are assigned as the inferred ancestry if cutoff > 80%, otherwise “Mixed (All < 80%)” is assigned.

## 3.0 CODE DESCRIPTION

**3.1** VCF file is generated from the bam file to call genotype information on pre-defined SNPs on snpwt.bed.gz[1].

```
o samtools mpileup -q 30 -Q 20 -v -f  genome.fa -l snpwt.bed ${file}.bam |
o bcftools call -c -Ov | bcftools filter -e 'ALT=\".\"' |
o bcftools annotate -c CHROM,FROM,TO,ID -a snpwt.bed.gz > ${file}_annot.vcf
```

**3.2** Convert a VCF file to eigenstrat format[3].

```
o python vcf2eigenstrat.py -v ${file}_annot.vcf -o ${file}
```

**3.3** Infer ancestry information using SNPweights[1].

```
o python SNPweights2.1/inferancestry.py --par ${file}.par
```

**3.4** Patient level ancestry information is obtained based on all files from samples from the patient using custom perl script (available upon request)