SOP Outlines **Version: 2.0.0**

1.0 PURPOSE/SCOPE

This Standing Operating Procedure (SOP) describes procedures for generating consensus/intersect variants calls 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 CAVEATS

2.1 Data should be considered representative of the major clone from the patient-derived models provided by the NCI Patient-Derived Models Repository and should not be considered to represent the entire model since intra-model heterogeneity in early-passage patient-derived models is expected and as this is intersection data, if a minor clone is not taken by a mouse it would not be represented here.

2.2 Common variants present in the population (e.g., germline) have not been removed from the reported non-synonymous variants. Population frequency databases (e.g., ExAC, 1000 genomes, NHLBI Exome Sequencing Project) should be used to remove variants present at high frequencies in the population.

3.0 DESCRIPTION OF CONSENSUS/INTERSECT VCF FILES

3.1 The consensus/intersect variant calls data are generated using whole exome sequence (WES) *.VCF files generated following the WES data analysis pipeline, version 2.0 (MCCRD_SOP0011).

3.2 VCF file contains all the variants present in all the PDX samples sequenced from a model.

3.3 MAF file contains only the variants which can alter protein sequence. This file is also filtered for Population frequency.

4.0 PROTOCOL

4.1 VCF files generated from the WES data analysis pipeline are converted to gz format.

```
bgzip -c ${file} >${file}.gz
tabix -p vcf ${file}.gz
```

4.2 All the files from a Model are merged using bcftools merge.

```
bcftools merge  *.vcf.gz -O z -o merge.vcf.gz
tabix -p vcf merge.vcf.gz
```

4.3 vcf-isec is used to filter out the variants which are not present in all the PDXs. A simple perl script is used to change the header of the 10th column (representative for VAF for Model).

```
vcf-isec -f -n +X merge.vcf.gz ${file}.gz |fixconsensesVCF.pl - Model |cut -f 1-10 >
{output}.vcf
```

4.4 vcf2maf is used to convert the VCF to MAF format.

```
module VEP/92 load vcf2maf/1.6.16
```

```
vcf2maf.pl --input-vcf {output}.vcf --output-maf ${output}.maf --tumor-id {modelID} --ref-fasta {input.ref} --filter-vcf ExAC.r0.3.1.sites.vep.vcf.gz --vep-path $VEP_HOME/ --vep-forks 2 --vep-data $VEP_CACHEDIR --custom-enst {isoforms}
```

isoforms file is downloaded from MSK GitHub account.

4.5 MAF file is filtered to carry nonsynonymous rare variants only using custom perl script (available upon request)

4.5.1 Common Variants (if allele count across at least one ExAC subpolutaiton is >10) are removed.