Comparison of genomic biomarkers identified by the whole exome, RNASeq and whole genome sequencing pipelines developed for the PDMR

Li Chen1*, Rajesh Patidar1*, Biswajit Das1, Chris Karlovich1, Tomas Vilimas1, Corinne Camalier1, Vivekananda Dafta1, Shahanawaz Jiwani1, William Walsh1, Palmer Fliss1, Sean McDermott1, Justine N. Mc Culcheon1, Amanda Peach1, Michelle Ahalt-Gothholm1, Carrie Bonomi1, Kelly Dougherty1, John Carter1, Sergio Y. Alcoser2, Tiffanie Chase1, Raymond Dvelellas1, Marion Gibson1, Kelly Hedger1, Candace Mallow1, Chelsea McGlynn3, Maiorone Morris1, Marianne Radzyminski1, Howard Stotler1, Jesse Stöttlemeyer1, Debbie Trail1, Yvonne A. Evrard1, Melinda G. Hollingshead2, P. Mickey Williams1 and James H. Doroshow3


The National Cancer Institute (NCI) has developed a Patient-Derived Models Repository (PDMR; www.pdmr.cancer.gov) of patient-derived xenografts (PDXs) with clinical annotation and comprehensive genomic characterization using whole exome sequencing (WES) and RNASeq. An in-house data analysis pipeline has been developed and validated to call germline and somatic variants and to perform transcriptional profiling in these models. There is a need to incorporate additional biomarkers into a standard data analysis pipeline, including loss of heterozygosity (LOH), microsatellite instability (MSI), copy number variation (CNV) and structure variants (SVs)/fusions for identifying appropriate PDX models for preclinical drug studies.

The National Cancer Institute (NCI) has developed a Patient-Derived Models Repository (PDMR; www.pdmr.cancer.gov) of patient-derived xenografts (PDXs) with clinical annotation and comprehensive genomic characterization using whole exome sequencing (WES) and RNASeq. An in-house data analysis pipeline has been developed and validated to call germline and somatic variants and to perform transcriptional profiling in these models. There is a need to incorporate additional biomarkers into a standard data analysis pipeline, including loss of heterozygosity (LOH), microsatellite instability (MSI), copy number variation (CNV) and structure variants (SVs)/fusions for identifying appropriate PDX models for preclinical drug studies. Validation of the methods used for the assessment of these and other genomic biomarkers is a crucial aspect in the development of the PDMR data analysis pipeline.

**INTRODUCTION**

The National Cancer Institute (NCI) has developed a Patient-Derived Models Repository (PDMR; www.pdmr.cancer.gov) of patient-derived xenografts (PDXs) with clinical annotation and comprehensive genomic characterization using whole exome sequencing (WES) and RNASeq. An in-house data analysis pipeline has been developed and validated to call germline and somatic variants and to perform transcriptional profiling in these models. There is a need to incorporate additional biomarkers into a standard data analysis pipeline, including loss of heterozygosity (LOH), microsatellite instability (MSI), copy number variation (CNV) and structure variants (SVs)/fusions for identifying appropriate PDX models for preclinical drug studies. Validation of the methods used for the assessment of these and other genomic biomarkers is a crucial aspect in the development of the PDMR data analysis pipeline.

**METHODS**

**LOH calling from WGS:**
- Identify most common heterozygous SNPs (~795,000 SNPs) including:
  - 792,138 SNPs from a population level genomic database (gnomAD)
  - Present in gnomAD population (Frequency > 30%)
  - Number of Homozygotes in the population < 2500
  - ~3,000 SNPs from a Clovis clinically validated SNP array
- Use runs of homozygosity (BCFtools/RoH) to call LOH regions based on genotypes of ~795,000 SNPs
- Filter LOH regions to only include eligible LOH:
  - Copy number is not 0
  - Length of region > 1 million base pairs
  - Length of region < 90% of chromosome arm
- Calculate percent of genomic LOH (%LOH):
  - 100*(total length of eligible LOH)/(total length of genome - total length of excluded LOH)

**LOH calling from WES:**
- Call variants using germline pipeline from HaplotypeCaller and Platypus
- Filter homoyzgous variants annotated from the 1000 Genomes Project
- Use BCFtools/RoH to call LOH region on autosomes
- Filter LOH regions < 150,000 bases
- Calculate percent of genomic LOH

**MSI calling from WGS:**
- mSINGs was used to assign a microsatellite instability score based on the fraction of unstable microsatellite loci

**CNV calling from WGS:**
- CNVkit was used to call CNVs from WGS and WES

**SVs/Fusion calling from WGS/RNASeq:**
- Manta for WGS data
- Tophat-Fusion-catcher for RNASeq data

**GENE EXPRESSION**

**RESULTS**

**A. Percent of Genomic LOH is Highly Correlated Between WGS and WES**
- 52 PDX samples from 22 models were tested (including 3 paired germline specimens):
  - %LOH ranged from <1% to 50% – specimens within the same models have consistent %LOH data
  - %LOH is highly correlated between WGS and WES
  - Clinically relevant %LOH cut-offs are needed – highly dependent on assay platform and disease histology
  - Algorithm is under development to adjust for tumor purity in %LOH calls

**B. Strong Concordance of MSI Calling is Observed Between WGS and WES**
- Average Pearson correlation coefficient of CNV profiles between WGS and RNASeq among 55 samples is 0.94, with standard deviation of 0.05
- Interesting focal amplification/deletion events were consistently detected between the two assays, including MET and CDKN2A alterations

**C. CNV Assessment is Highly Correlated Between WGS and WES**
- Concordance of Structural Variants/Fusion Calls Between WGS and RNASeq

**REFERENCES**

1. NCI PDMR website: https://pdmr.cancer.gov
3. Genomic profiling data, SOPs, data analysis pipeline SOPs available at NCI PDMR website