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

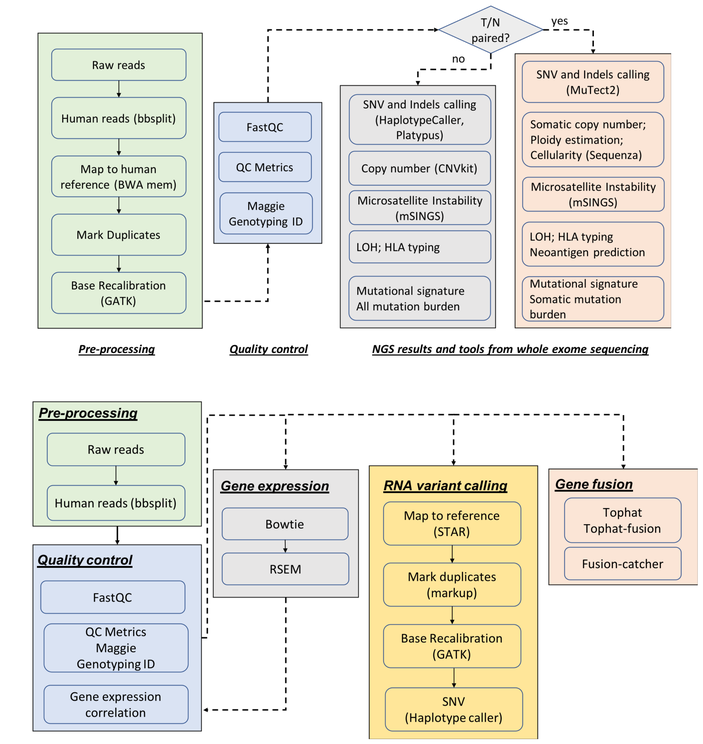
Li Chen[1*], Rajesh Patidar[1*], Biswajit Das[1], Chris Karlovich[1], Tomas Vilimas[1], Corinne Camalier[1], Vivekananda Datta[1], Shahanawaz Jiwani[1], William Walsh[1], Palmer Fliss[1], Sean McDermott[1], Justine N. McCutcheon[1], Amanda Peach[1], Michelle Ahalt-Gottholm[2], Carrie Bonomi[1], Kelly Dougherty[1], John Carter[1], Sergio Y. Alcoser[2], Tiffanie Chase[1], Raymond Divelbiss[1], Marion Gibson[1], Kelly Hedger[1], Candace Mallow[1], Chelsea McGlynn[1], Malorie Morris[1], Marianne Radzyminski[1], Howard Stotler[1], Jesse Stottlemyer[1], Debbie Trail[1], Yvonne A, Evrard[1], Melinda G. Hollingshead[2], P. Mickey Williams[1]and James H. Doroshow[3]

[1]Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, [2]National Cancer Institute at Frederick, Biological Testing Branch, Developmental Therapeutics Program, Frederick, MD and 3National Cancer Institute, Division of Cancer Treatment and Diagnosis, Bethesda, MD *: co-first author

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.

PDMR WES and RNASeq pipeline



OBJECTIVES

- Determine if whole exome sequencing (WES) or RNASeq can accurately call:
  - Loss of heterozygosity (LOH)
  - Microsatellite instability (MSI)
  - Copy number variation (CNVs)
  - Structural variants (SVs)/fusions
- Whole genome sequencing (WGS) is considered the gold standard
- WGS was performed on 58 PDX/Patient samples and assessment of LOH, MSI CNVs, and SVs was compared with matched WES (LOH, MSI, CNVs) and RNASeq (SVs) assessments

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[4]
- 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)[4]

LOH calling from WES:

- Call variants using germline pipeline from HaplotypeCaller and Platypus
- Filter homozygous variants annotated from the 1000 Genomes Project
- Use BCFtools/RoH to call LOH region on autosomes
- Filter LOH regions < 150000 bases
- Calculate percent of genomic LOH

MSI calling from WGS/WES:

- mSINGS was used to assign a microsatellite instability score based on the fraction of unstable microsatellite loci

CNV calling from WGS/WES:

- CNVkit was used to call CNVs from WGS and WES

SVs/Fusion calling from WGS/RNASeq:

- Manta for WGS data
- Tophat-fusion and Fusion-catcher for RNASeq data
- Clinically Relevant and Diagnostic Variants

CONCLUSION

We observed excellent consistency between WGS, WES and RNASeq data in the assessment of percent of genomic LOH, MSI score, CNVs, and SVs/fusions. Our data analysis pipeline can accurately call genomic biomarkers from WES and RNASeq data, which facilitates the molecular characterization and prioritization of PDMR models for preclinical drug treatment.

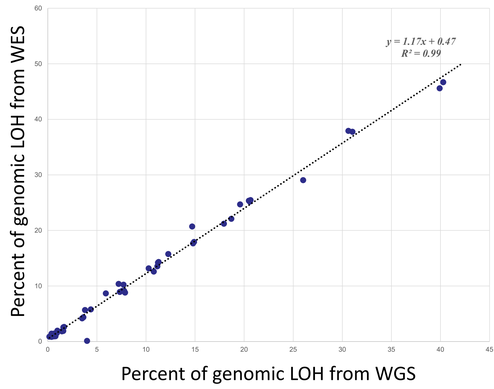
REFERENCES

1. NCI PDMR website: https://pdmr.cancer.gov
2. MoCha NGS pipeline: https://github.com/FNL-MoCha/nextgenseq_pipeline
3. Genomic profiling data, SOPs, data analysis pipeline SOPs available at NCI PDMR website
4. Swisher EM, Lin KK, Oza AM, et al., Lancet Oncol 2016

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

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

WGS \ WES	MSI-H	MSI-S	Total
MSI-H	22	3	25
MSI-S	0	24	24
Total	22	27	49

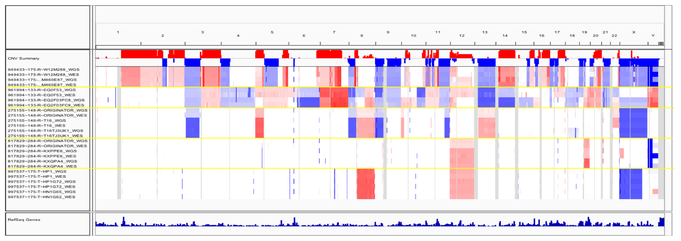
- 49 samples (22 models) were tested for both WGS and WES
- Concordance rate = 46/49 = 91%
- Same MSI score cut-off (0.2) derived from WES data was used to call MSI status, which may not be applicable to WGS

Discordance (3 samples):

- 2 samples had WES MSI score < 0.2; other samples from the same model are MSI-H for both WGS and WES -> likely false negative calls for WES
- 1 sample had a WGS MSI score of 0.23; other samples from the same model are MSI-S for both WGS and WES -> likely false positive call for WGS

C. CNV Assessment Is Highly Correlated Between WGS and WES

- Average Pearson correlation coefficient of CNV profiles between WGS and WES among 55 samples is 0.94, with standard derivation of 0.05
- Interesting focal amplification/deletion events were consistently detected between the two assays, including MET and CDKN2A alterations



D. Concordance of Structural Variants/Fusion Calls Between WGS and RNASeq

Model	Fusion	# samples compared	# samples detected in WGS	# samples detected in RNASeq
114868-125-R	FGFR3-TACC3	3	3	3
237351-077-R	EWSR1-FLI1	2	2	2
287954-098-R	EWSR1-FLI1	2	2	2
117519-064-T	TMPRSS2-ERG	2	0	0

PDX model 117519-064-T (prostate cancer, 62y male):

TMPRSS2-ERG fusion found in some samples from RNASeq or other fusion panels

Sample Name	# Reads(WGS)	# Reads (RNASEQ)	# Reads (Oncomine)	# Reads (Illumina TruSight Fusion Panel)
117519~064-T~ED0D46J42	Not Sequenced	1	56667	1793
117519~064-T~ED0D48	Not Sequenced	2	30329	986
117519~064-T~ED1D49	Not Detected	Not Detected	41723	5
117519~064-T~ED1D50YD8	Not Detected	Not Detected	54031	387
117519~064-T~ED1D51WF6	Not Sequenced	1	48777	25
117519~064-T~ED1D53YE1	Not Sequenced	Not Detected	43577	530

NIH NATIONAL CANCER INSTITUTE

Frederick National Laboratory for Cancer Research

sponsored by the National Cancer Institute