Washington University in St. Louis
School of Medicine

Washington University CPTAC3 Genomic Characterization Pipelines

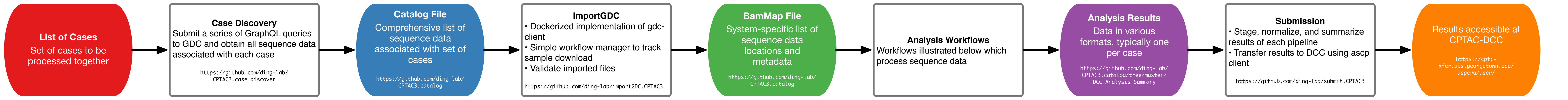
Matthew A. Wyczalkowski, Liang-Bo Wang, Daniel Cui Zhou, Lijun Yao, Ruiyang Liu, Yige Wu, Wen-Wei Liang, Song Cao, Houxiang Zhu, Li Ding

Department of Medicine and McDonnell Genome Institute, Washington University in St. Louis, MO 63108, USA.

DingLab
dinglab.wustl.edu

Data Acquisition and Processing

Each batch of cases is processed in a series of steps, beginning with the discovery of sample details at GDC, the import of data to local systems, running of the various workflows, and finally upload to DCC.



List of Cases
Set of cases to be processed together

Case Discovery
Submit a series of GraphQL queries to GDC and obtain all sequence data associated with each case
https://github.com/ding-lab/CPTAC3.case.discover

Catalog File
Comprehensive list of sequence data associated with set of cases
https://github.com/ding-lab/CPTAC3.catalog

ImportGDC
- Dockerized implementation of gdc-client
- Simple workflow manager to track sample download
- Validate imported files

https://github.com/ding-lab/importGDC.CPTAC3

BamMap File
System-specific list of sequence data locations and metadata
https://github.com/ding-lab/CPTAC3.catalog

Analysis Workflows
Workflows illustrated below which process sequence data

Analysis Results
Data in various formats, typically one per case
https://github.com/ding-lab/CPTAC3.catalog/tree/master/DCC_Analysis_Summary

Submission
- Stage, normalize, and summarize results of each pipeline
- Transfer results to DCC using ascp client

https://github.com/ding-lab/submit.CPTAC3

Results accessible at CPTAC-DCC
https://cptc-xfer.uis.georgetown.edu/aspera/user/

Processing Summary

Table below details batches which have been submitted to DCC to date.

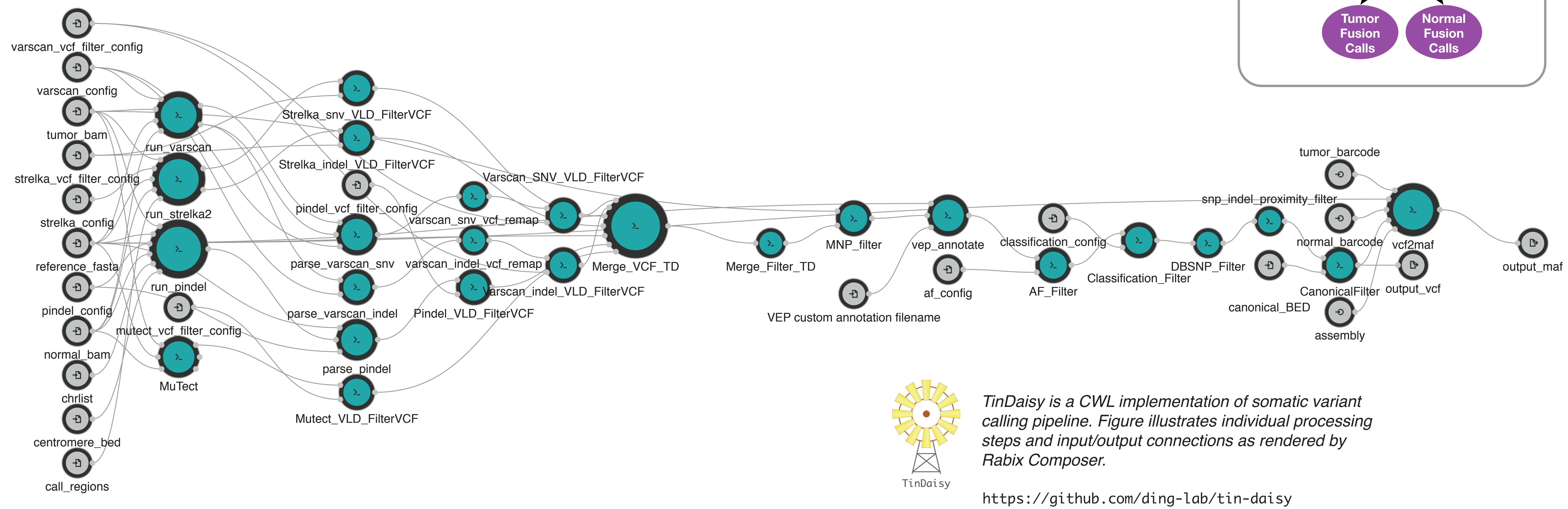
Pipeline	Data and Sample Type	CCRCC	GBM	HNSCC	LSCC	LUAD	PDA	UCEC	Total
Methylation Array	▦	222	116	111	113	229	164	246	1201
miRNA	▲	222	114	111	113	229	164	247	1200
Expression	◆ ◆	225	120	118	117	230	164	251	1225
Fusion	◆ ◆	222	119	111	113	164	164	222	1115
Transcript & Splicing	◆ ◆	222	119	111	113	53	164	222	1004
SV	● ●	222	59	109	113	111	166	217	997
Somatic CNV	● ●	222	59	109	113	121	166	39	829
MSI	○ ○	222	118	111	113	111	166	244	1085

Data Type ● WGS ○ WES ◆ RNA-Seq ▲ miRNA ▦ Methylation array

Sample Type 🟥 Tumor 🟦 Blood normal

Looking Ahead

- Import and storage of very large datasets is the principal bottleneck for high throughput genomic analysis.
- Containerization and workflow definition languages are technologies which enable cloud-based computing, bringing pipelines to the data rather than the other way around.
- We have implemented a number of pipelines in Docker / CWL. Somatic and germline variant calling is currently being processed using somatic calling pipeline on Cromwell workflow engine.

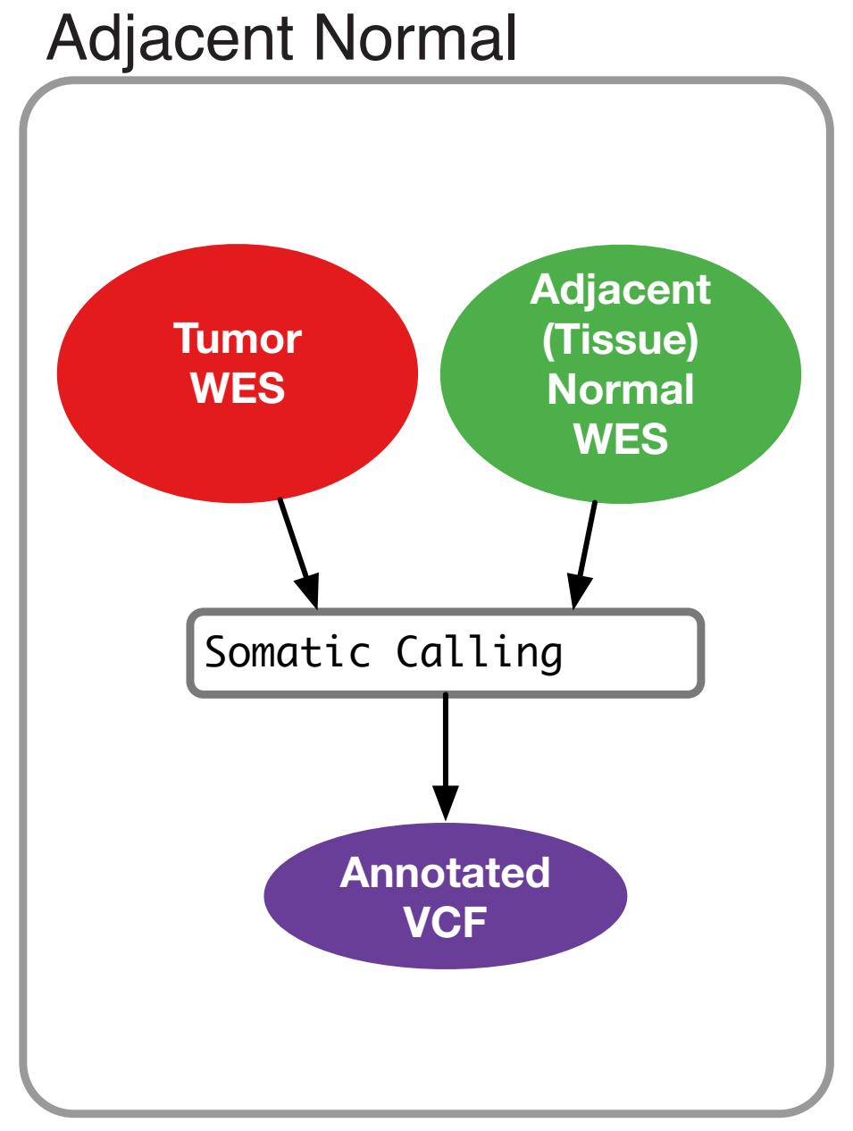


TinDaisy is a CWL implementation of somatic variant calling pipeline. Figure illustrates individual processing steps and input/output connections as rendered by Rabix Composer.

https://github.com/ding-lab/tin-daisy

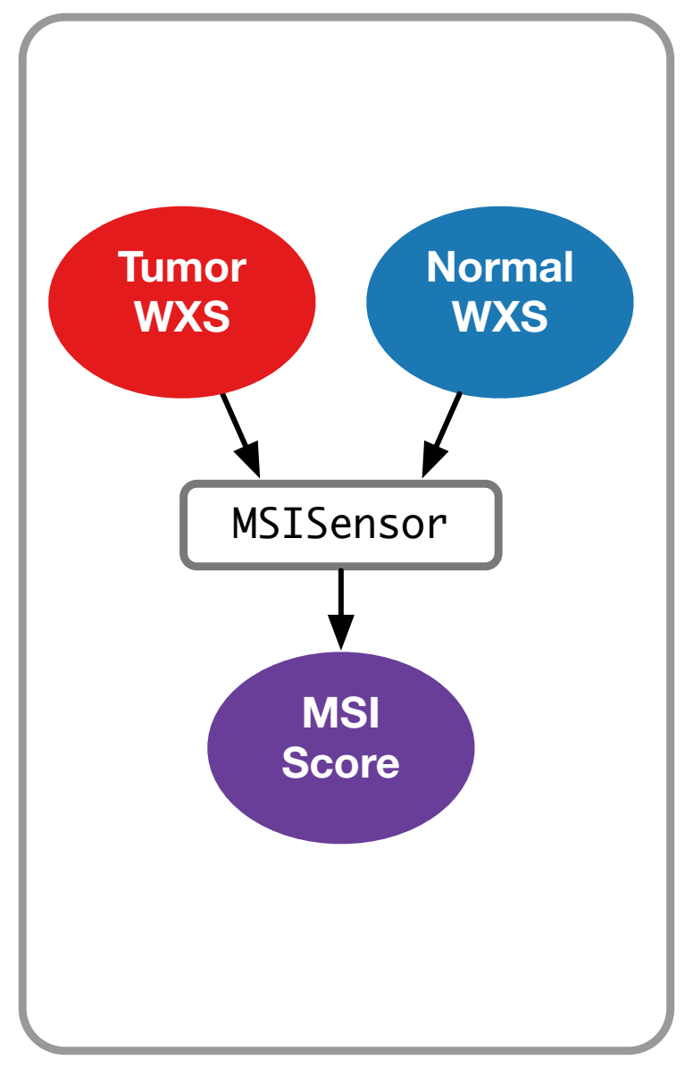
Pipeline Details

Adjacent Normal



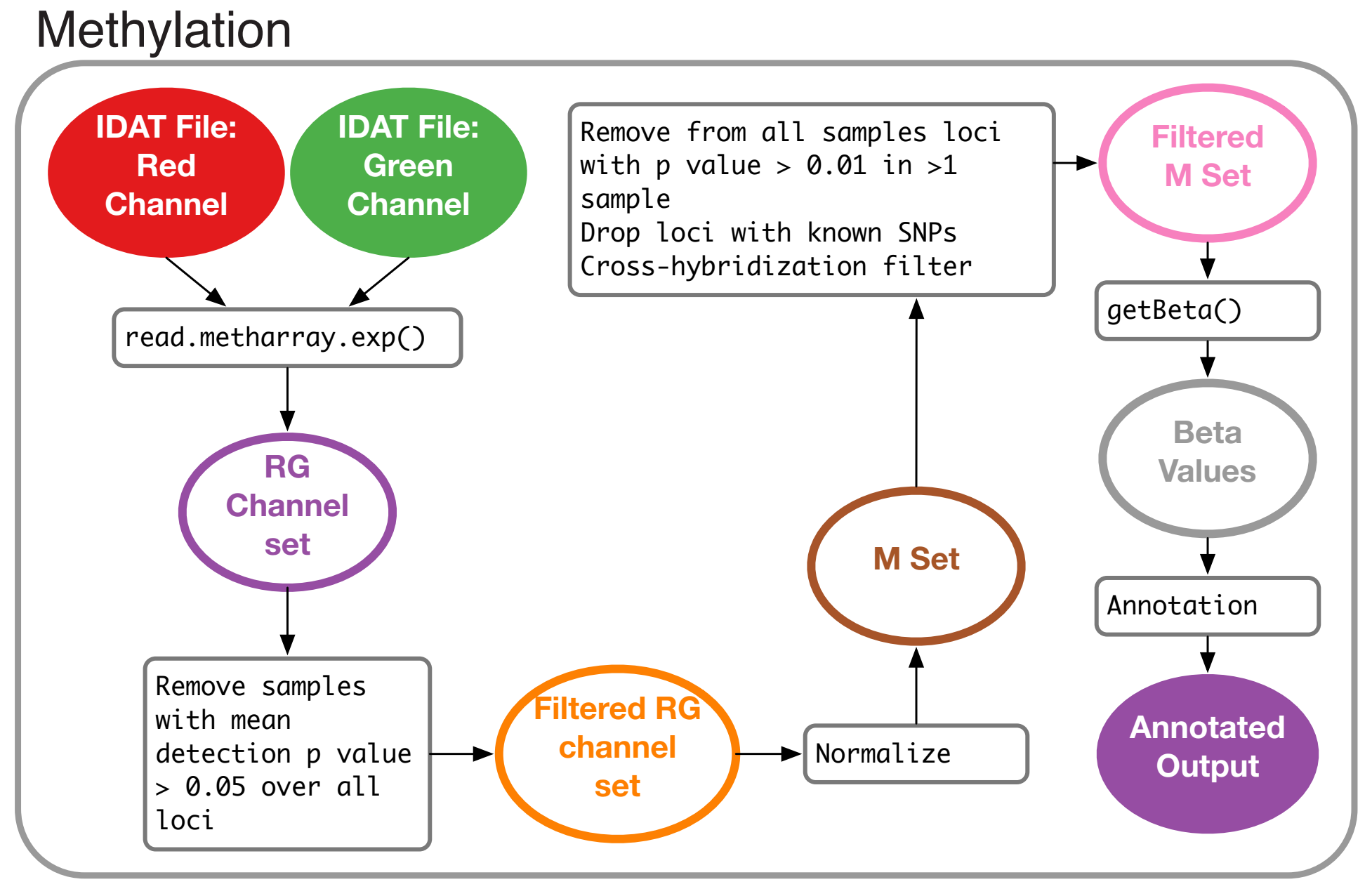
Tumor WES
Adjacent (Tissue) Normal WES
Somatic Calling
Annotated VCF

MSI



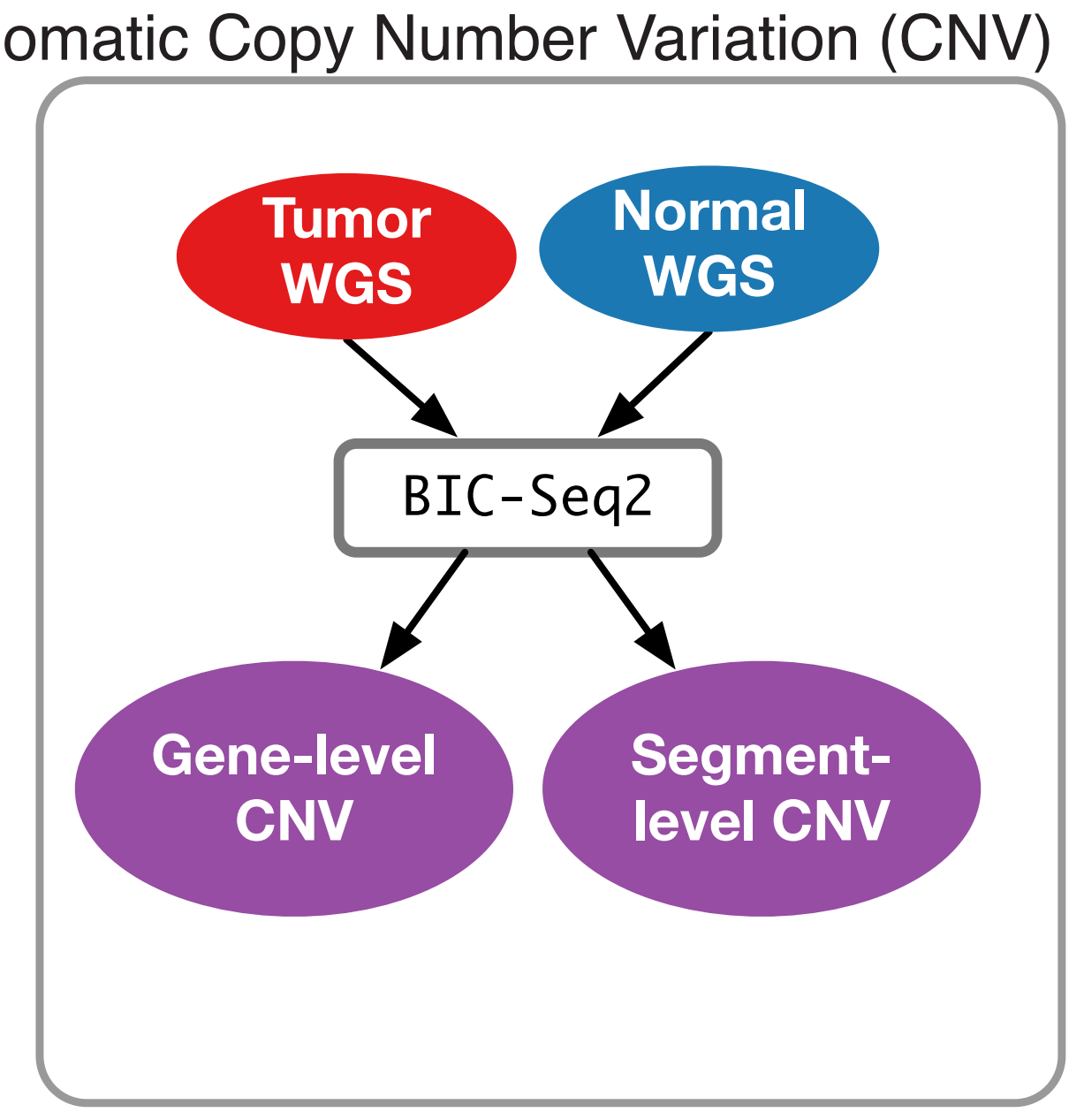
Tumor WXS
Normal WXS
MSISensor
MSI Score

Methylation



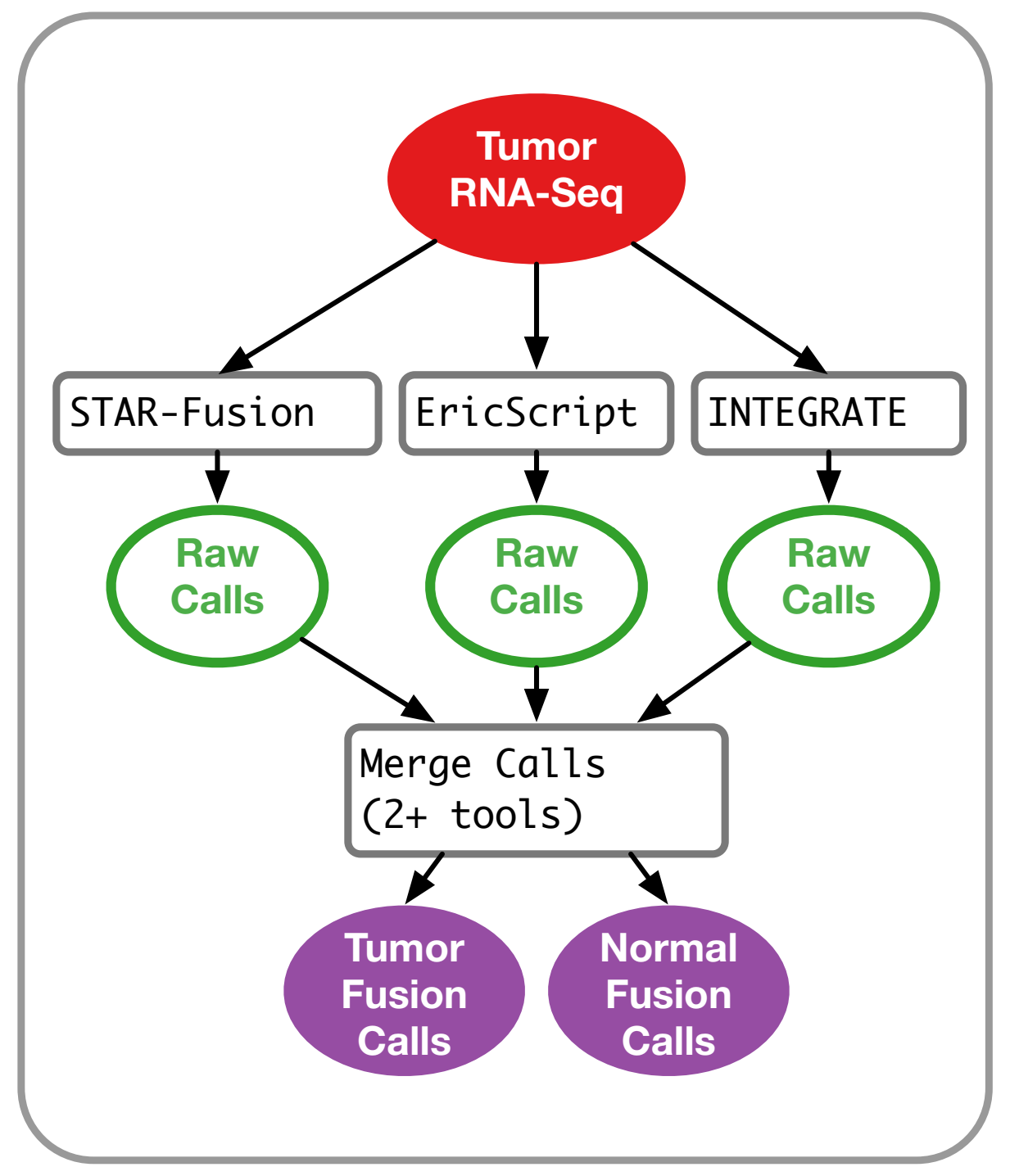
IDAT File: Red Channel
IDAT File: Green Channel
read.metharray.exp()
RG Channel set
Remove samples with mean detection p value > 0.05 over all loci
Filtered RG channel set
Normalize
M Set
Remove from all samples loci with p value > 0.01 in >1 sample
Drop loci with known SNPs
Cross-hybridization filter
Filtered M Set
getBeta()
Beta Values
Annotation
Annotated Output

Somatic Copy Number Variation (CNV)



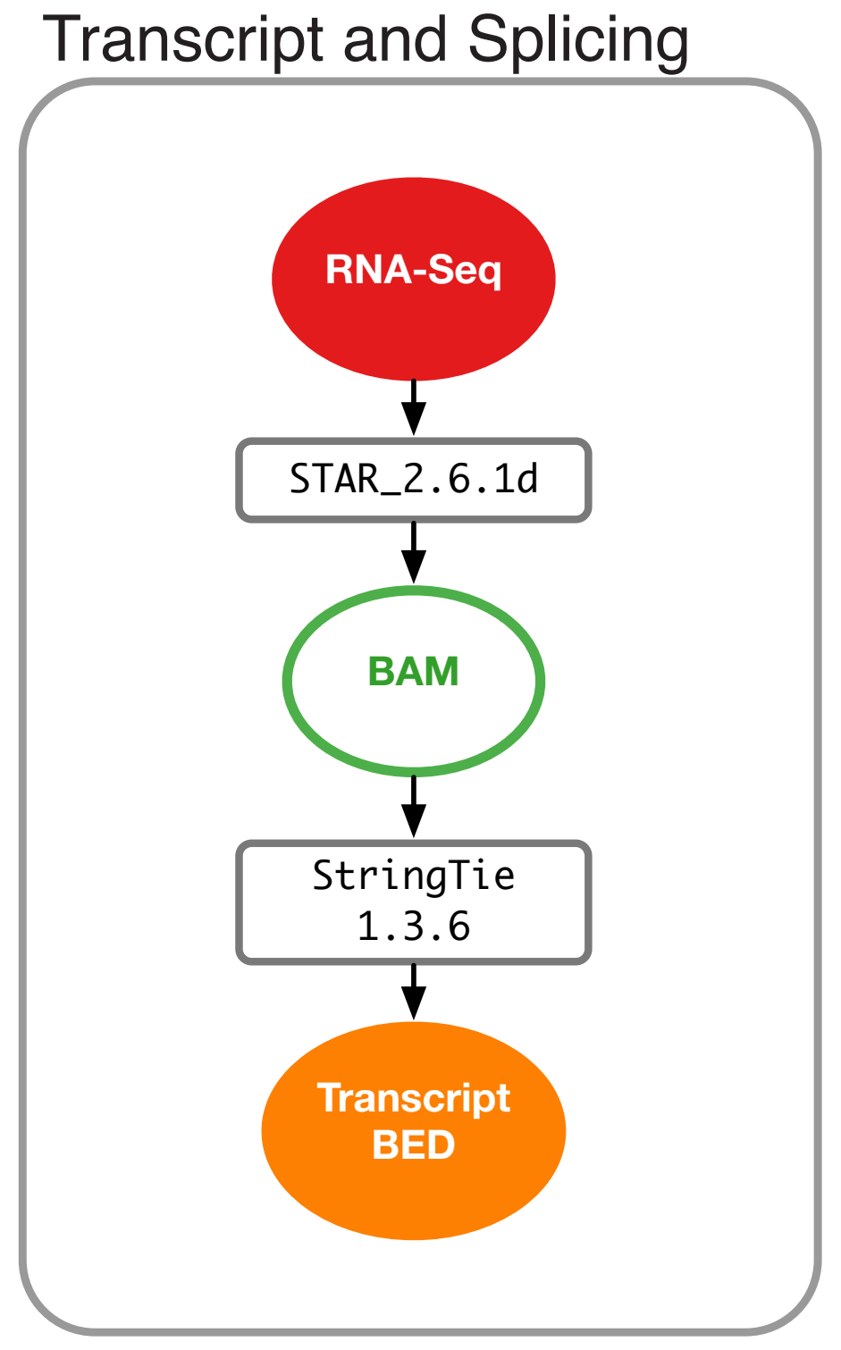
Tumor WGS
Normal WGS
BIC-Seq2
Gene-level CNV
Segment-level CNV

Fusion



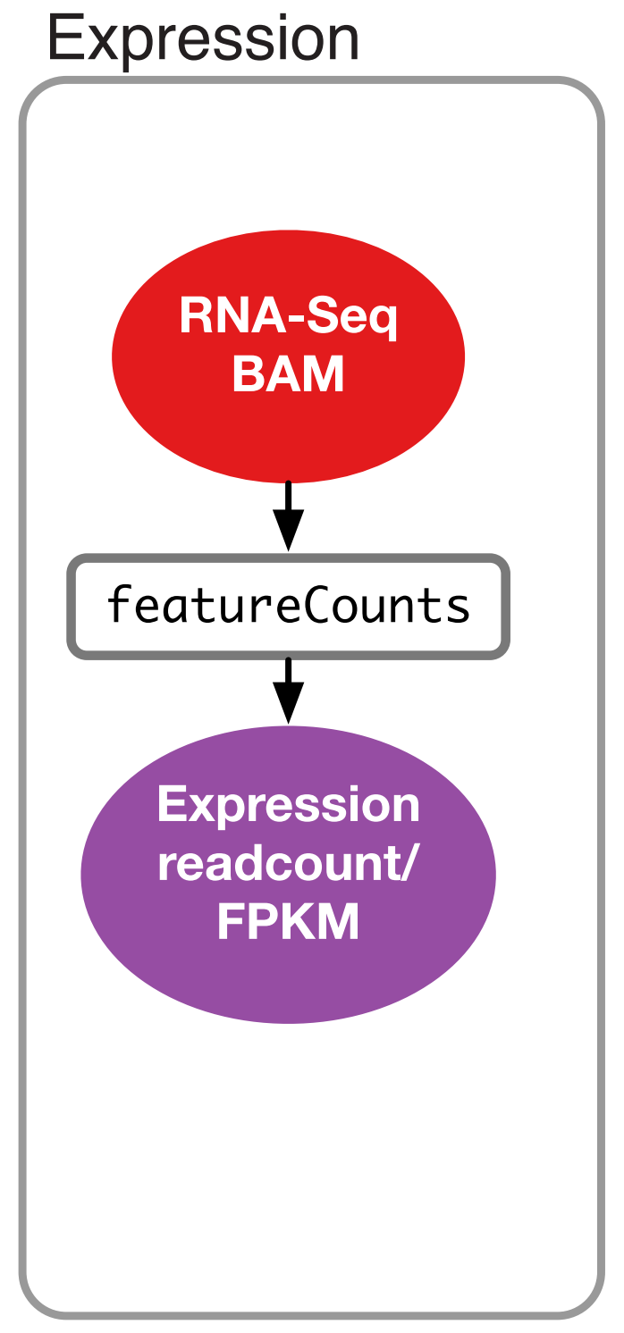
Tumor RNA-Seq
STAR-Fusion
EricScript
INTEGRATE
Raw Calls
Raw Calls
Raw Calls
Merge Calls (2+ tools)
Tumor Fusion Calls
Normal Fusion Calls

Transcript and Splicing



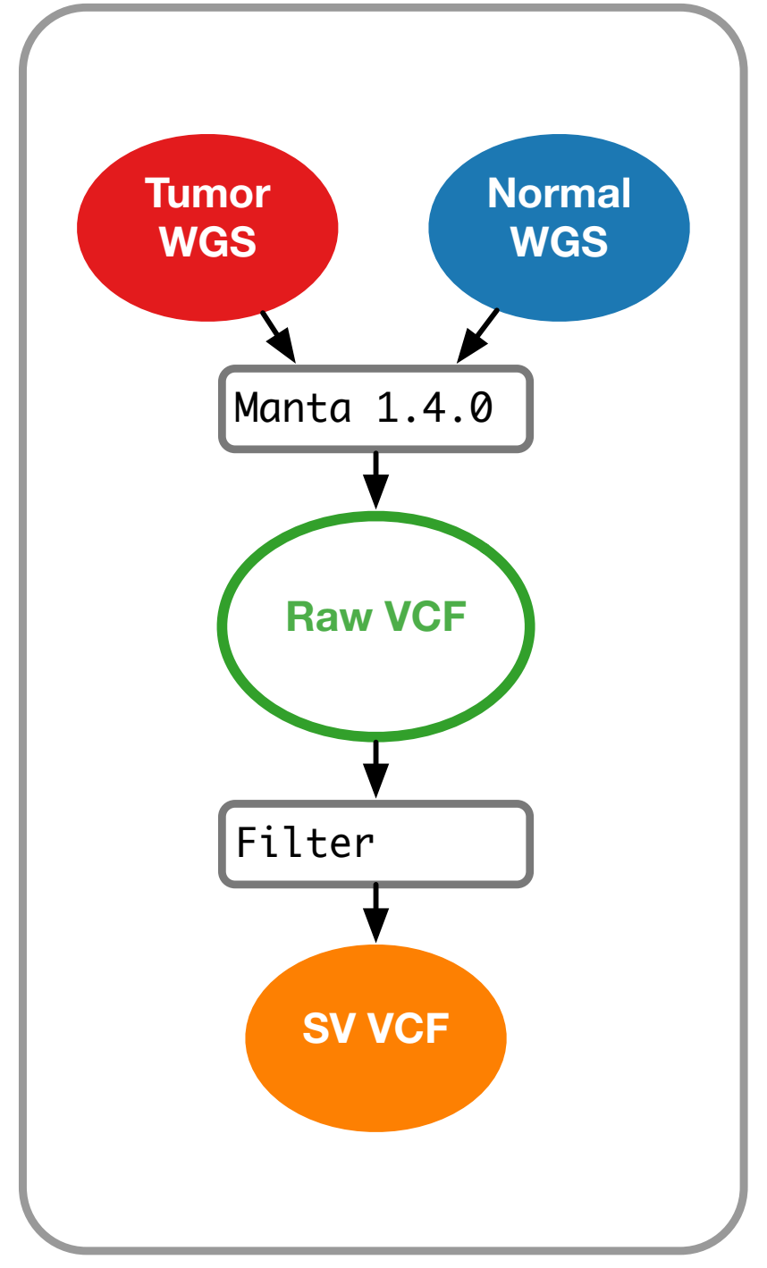
RNA-Seq
STAR_2.6.1d
BAM
StringTie 1.3.6
Transcript BED

Expression



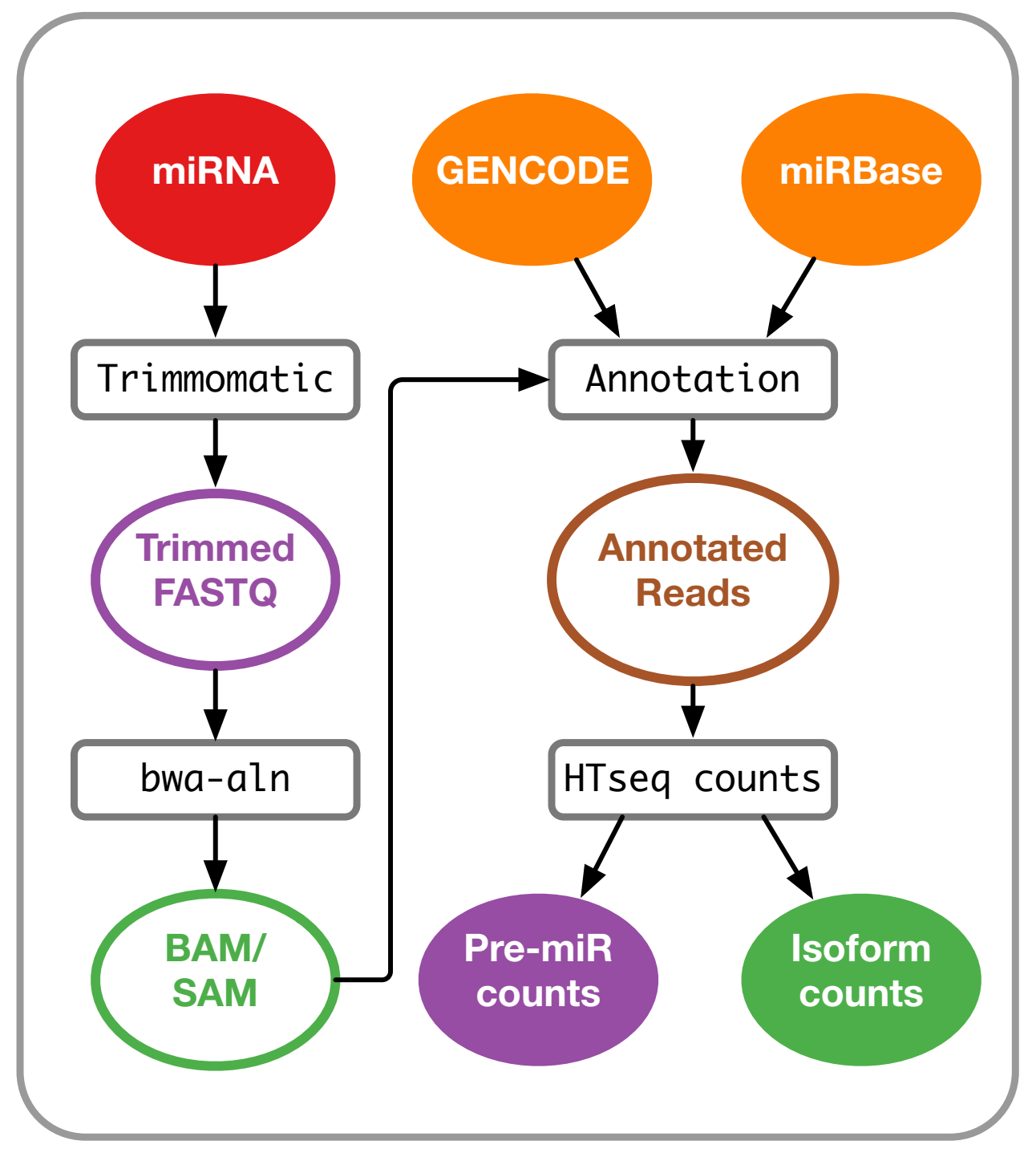
RNA-Seq BAM
featureCounts
Expression readcount/ FPKM

Structural Variation (SV)



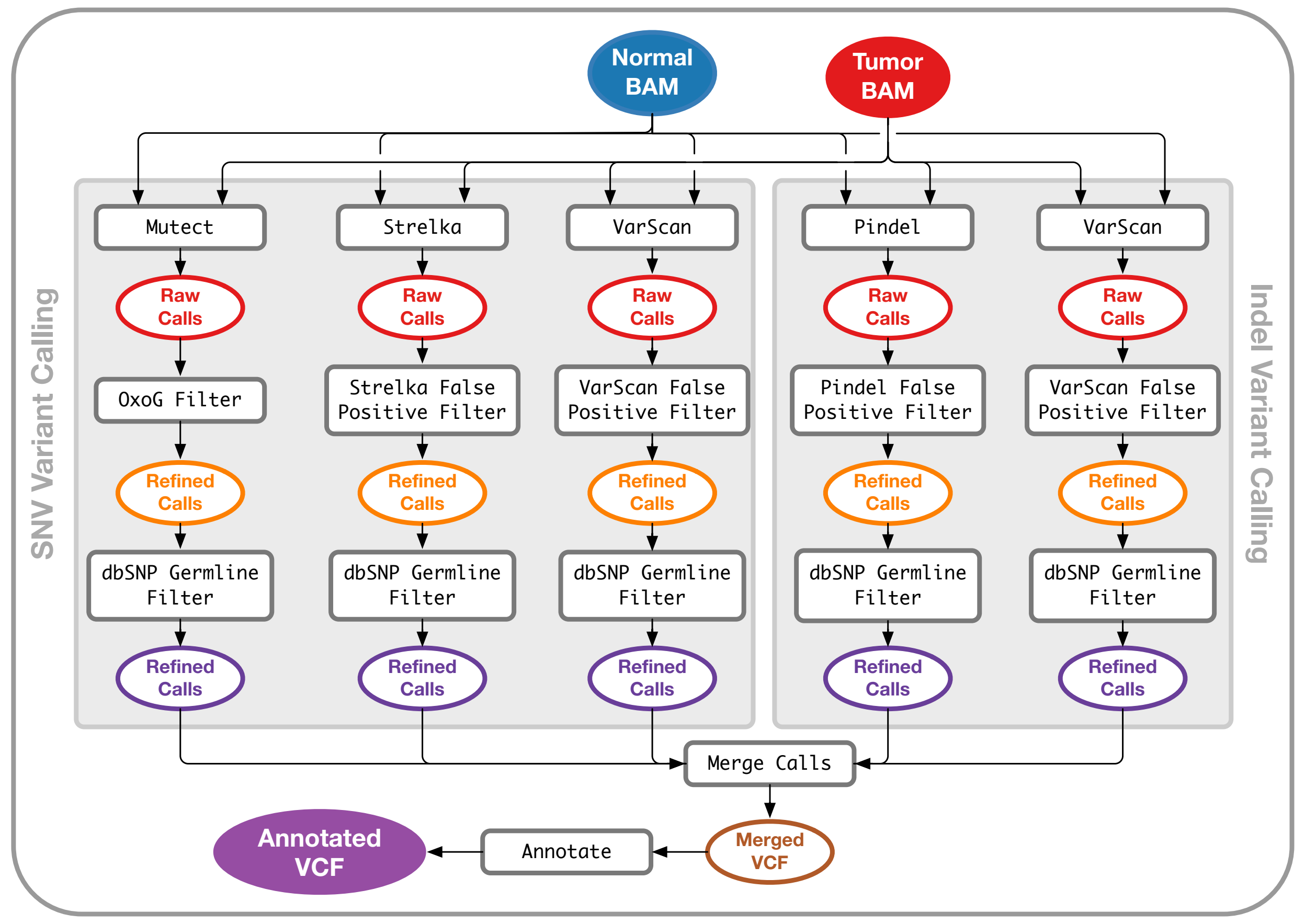
Tumor WGS
Normal WGS
Manta 1.4.0
Raw VCF
Filter
SV VCF

miRNA



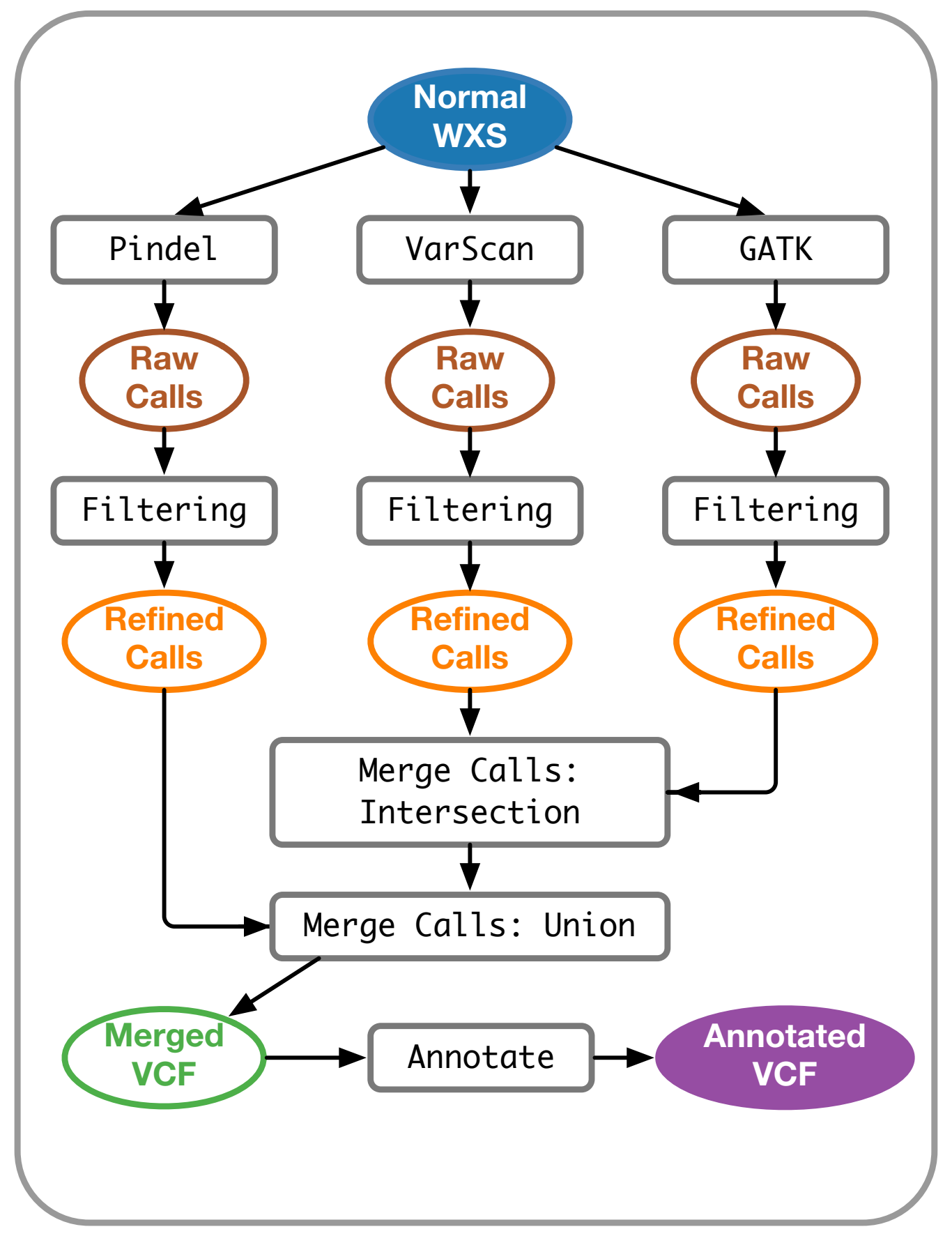
miRNA
GENCODE
miRBase
Trimmomatic
Trimmed FASTQ
bwa-aln
BAM/ SAM
Annotation
Annotated Reads
HTseq counts
Pre-miR counts
Isoform counts

Somatic Variant



Normal BAM
Tumor BAM
SNV Variant Calling
Mutect
Raw Calls
OxoG Filter
Refined Calls
dbSNP Germline Filter
Refined Calls
Strelka
Raw Calls
Strelka False Positive Filter
Refined Calls
dbSNP Germline Filter
Refined Calls
VarScan
Raw Calls
VarScan False Positive Filter
Refined Calls
dbSNP Germline Filter
Refined Calls
Indel Variant Calling
Pindel
Raw Calls
Pindel False Positive Filter
Refined Calls
dbSNP Germline Filter
Refined Calls
VarScan
Raw Calls
VarScan False Positive Filter
Refined Calls
dbSNP Germline Filter
Refined Calls
Merge Calls
Merged VCF
Annotate
Annotated VCF

Germline Variant



Normal WXS
Pindel
VarScan
GATK
Raw Calls
Raw Calls
Raw Calls
Filtering
Filtering
Filtering
Refined Calls
Refined Calls
Refined Calls
Merge Calls: Intersection
Merge Calls: Union
Merged VCF
Annotate
Annotated VCF

Acknowledgments

We acknowledge National Cancer Institute grants U24CA210972 and U24CA211006. We are grateful for contributions from the following individuals: NCI CPTAC program office: Mehdi Mesri, Henry Rodriguez, Christopher Kinsinger, Emily Boja, Tara Hiltke; Leidos: Mathangi Thiagarajan, Bill Bocik; NYU-WU-PNNL PGDAC: Fenyö lab, Ding lab, Payne lab, Kelly Ruggles; DCC: Karen Ketchum; Wash U: Sherri Davies.