Exploring GDC Copy Number Variation (CNV) Analysis Workflows and Tools

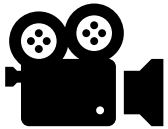
25 August 2025

- Bill Wysocki, Ph.D – Director of User Services
- Zhenyu Zhang, Ph.D – Director of Bioinformatics
  Center for Translational Data Science, University of Chicago
- Xin Zhou, Ph.D - Director of Data Visualization
  Computational Biology Department, St. Jude Children's Research Hospital

NIH NATIONAL CANCER INSTITUTE

Webinar Logistics

- *Webinar will be recorded*



- *Recording and slides will be made available soon*

- *Type any questions in the Q&A panel – they will be addressed at the end*

Agenda

1. *Overview of GDC copy number variation harmonization workflows*
2. *GDC CNV files overview*
3. *GDC Copy Number Segment tool*
4. *GDC CNV API*
5. *Questions*

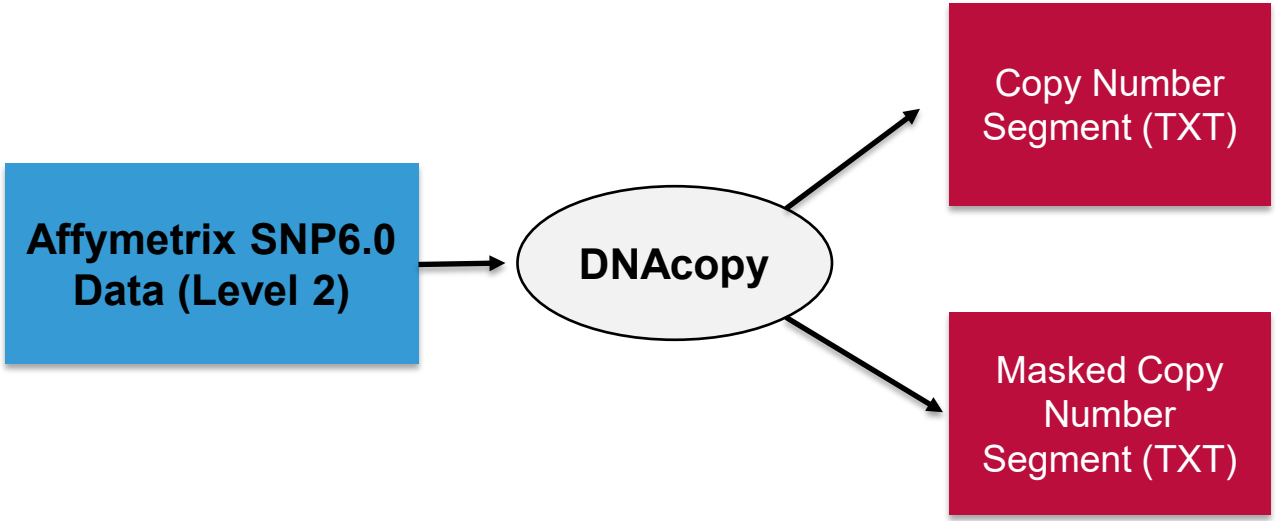
Overview of GDC Copy Number Variation Harmonization Workflows

Copy Number Variation Harmonization – Two Categories

- Array-based CNV harmonization
  - Uses Affymetrix SNP6.0 array
  - Currently only used for TCGA and TARGET programs

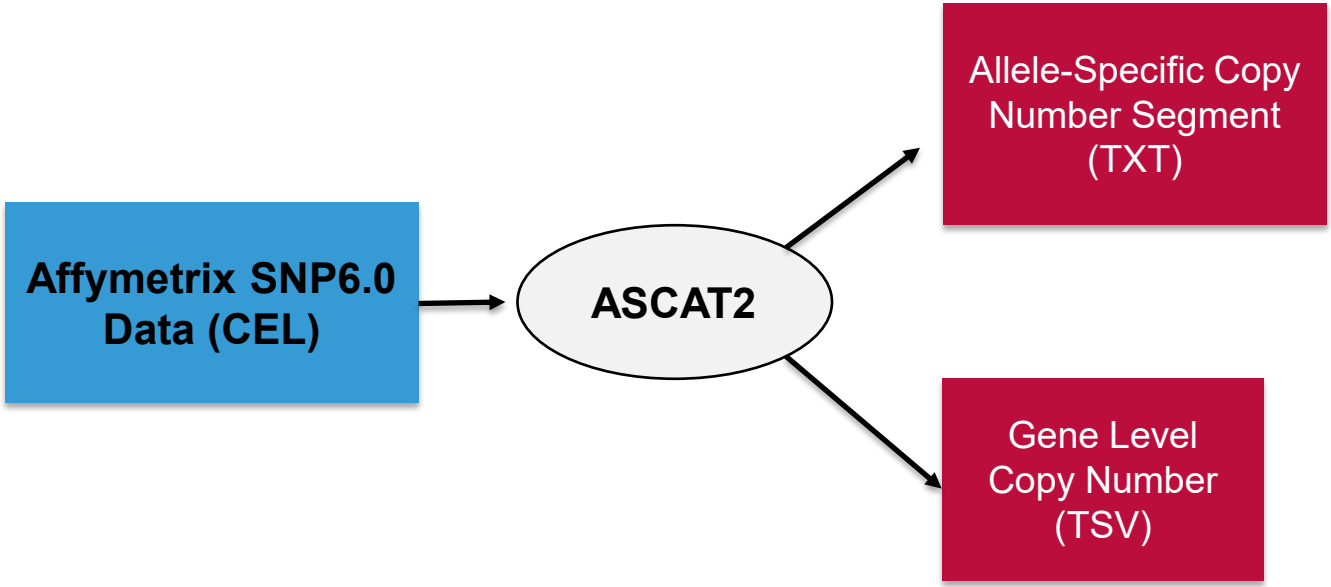
- WGS-based CNV harmonization
  - Uses aligned reads to infer copy number information
  - Ongoing project harmonization

1. CNV Harmonization in the GDC – DNAcopy

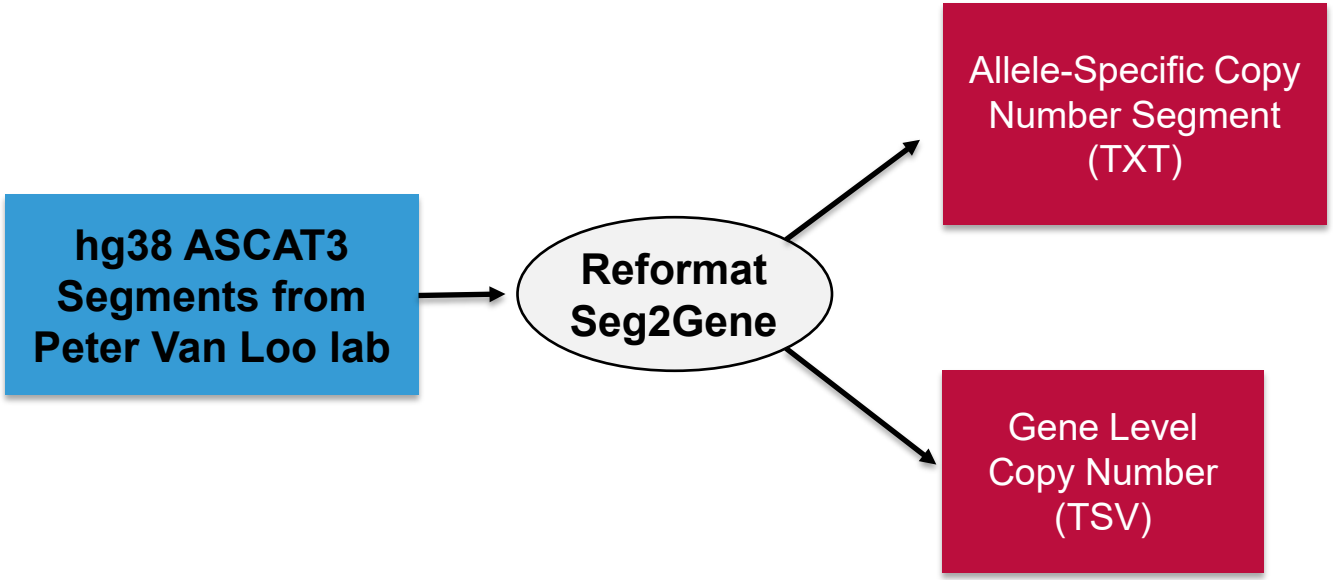


- SNP6.0 level 2 data, tangent copy number files, were generated by Birdsuite.
- "Masked" copy number segment has better quality compared to unmasked ones.

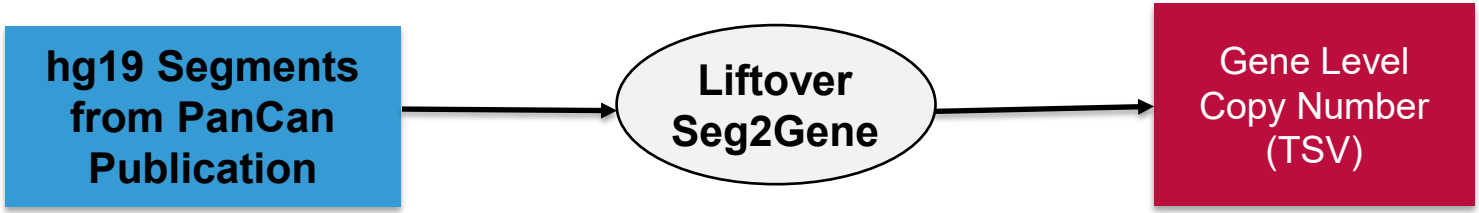
2. CNV Harmonization in the GDC – ASCAT2



3. CNV Harmonization in the GDC – ASCAT3

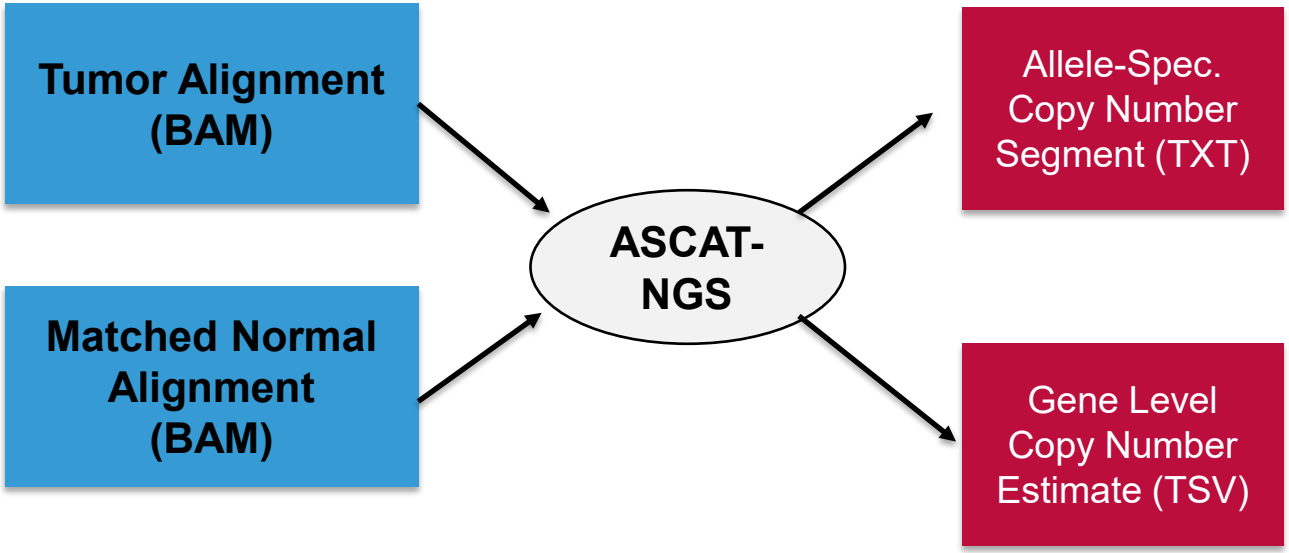


4. CNV Harmonization in the GDC – ABSOLUTE Liftover



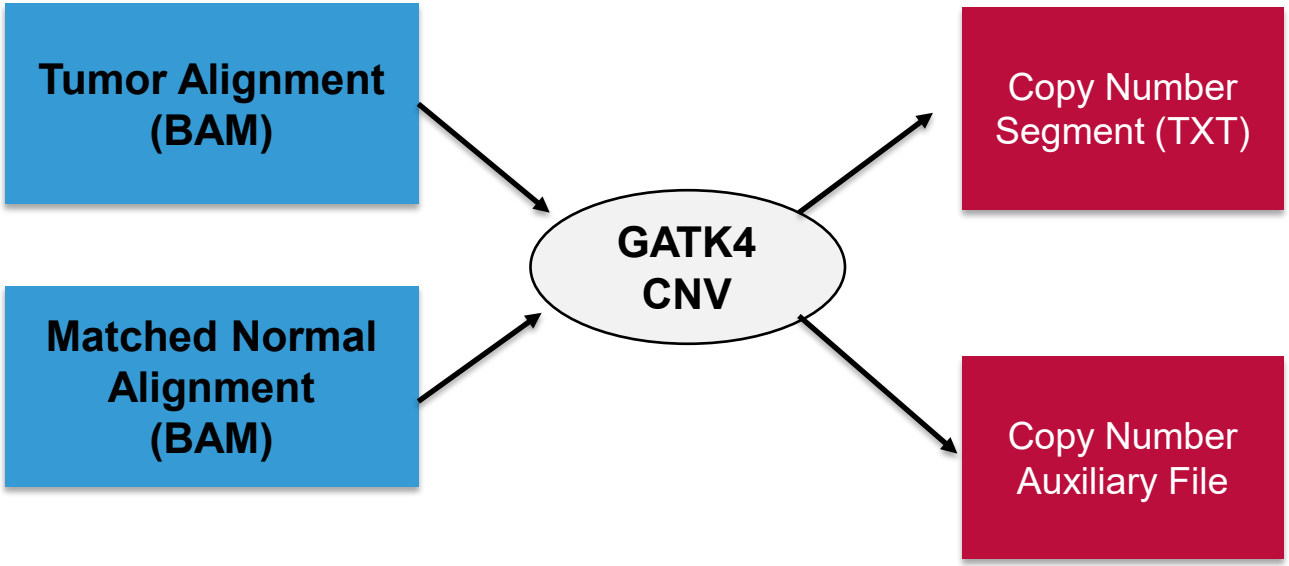
- The hg19 segments are from the TCGA PanCanAtlas publications.

5. CNV Harmonization in the GDC – Sanger Pipeline



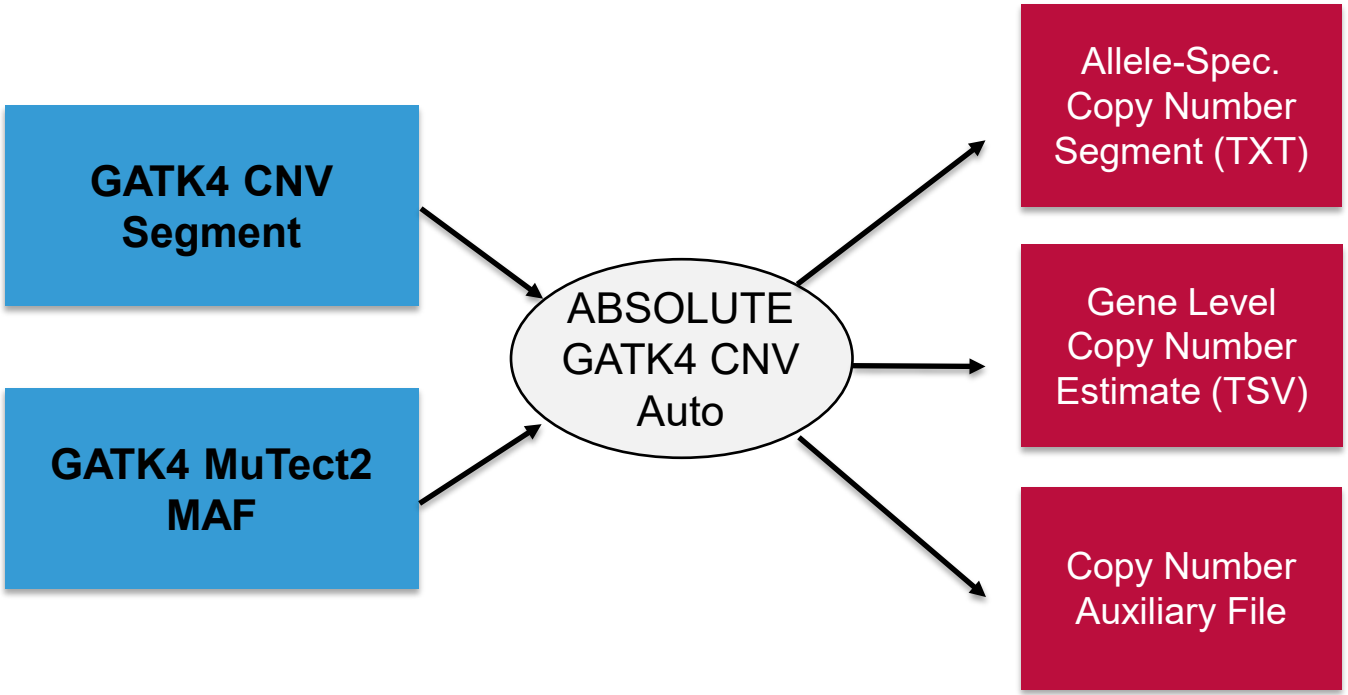
- This pipeline is deprecated - files will remain available but new files will not be generated

6. CNV Harmonization in the GDC – GATK4 CNV Pipeline



The auxiliary file contains intermediate calling products, including germline allele frequencies, to be consumed by experts for manual curation purpose.

7. CNV Harmonization in the GDC – ABSOLUTE Pipeline



The auxiliary file is a PDF that contains all potential CNV models (purity + ploidy combinations)

Summary of GDC CNV Workflows

Workflow	Strategy	Segment-Mean Segment	Integer (absolute) Copy Number Segment	Integer (absolute) Gene-Level Copy Number	Purity/ Ploidy Measurement
DNACopy	SNP6	✓			
ASCAT2	SNP6		✓	✓	
ASCAT3	SNP6		✓	✓	
ABSOLUTE LiftOver	SNP6			✓	
GATK4 CNV	WGS	✓			
ascatNGS	WGS		✓	✓	✓
ABSOLUTE	WGS		✓	✓	✓

- GATK4 CNV and ABSOLUTE are the only workflows currently in active production for new data.

CNV File Overview

ASCAT2/ ASCAT3/ ascatNGS/ ABSOLUTE Allele-Specific Copy Number Segment – File format

- GDC_Aliquot – Aliquot UUID in the GDC
- Chromosome – Chromosome number
- Start/End – Segment coordinates
- Copy_Number – the sum of major and minor copy number
- Major_Copy_Number – the larger strand copy number
- Minor_Copy_Number – the smaller strand copy number

DNAcopy and GATK4 CNV Segment (Segment Mean) – File format

- GDC_Aliquot – Aliquot UUID for DNAcopy, submitter_id for GATK4
- Chromosome – Chromosome number
- Start/End – Segment coordinates
- Num_Probes – Number of probes that support the segment
- Segment_Mean – log2(copy_number/2)
  - Diploid regions will have a segment mean of zero
  - Amplified regions will have a positive value
  - Regions with copy number losses will be negative

Copy Number Estimate (Gene Level Copy Number) – File format

Applies to: ASCAT2/ ASCAT3/ ascatNGS/ ABSOLUTE

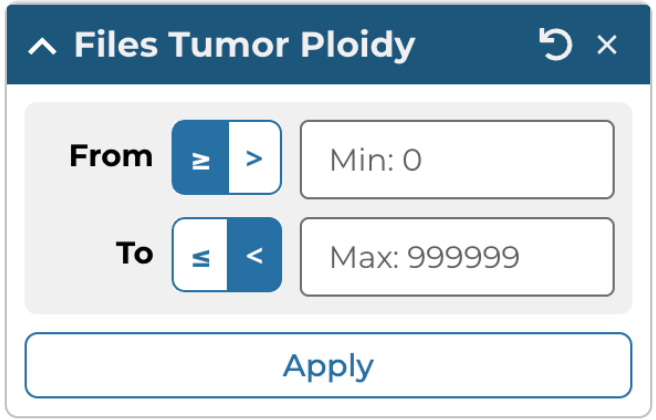
- gene_id – the Ensembl ID for the gene
- gene_name – the gene name in HUGO format
- Chromosome – chromosome number
- start/end – gene coordinates
- copy_number – weighted median of copy number values from overlapped regions
- min_copy_number - minimum value of overlapped segments
- max_copy_number - maximum value of overlapped segments

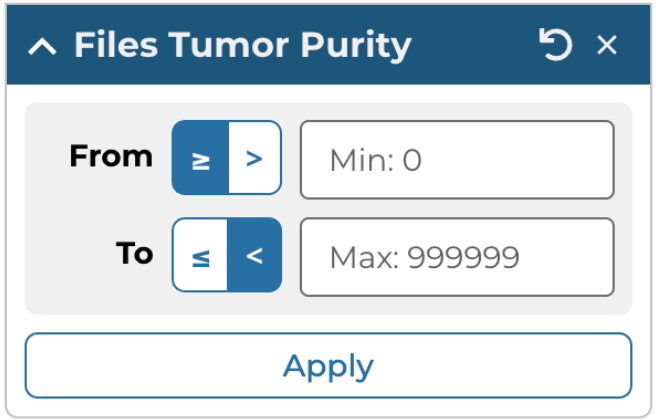
Copy Number Auxiliary File

- Produced from GATK4 CNV pipeline
  - A data bundle contains multiple intermediate files, including germline allele frequencies, to be consumed by experts for manual curation purpose only. We don't expect regular users to consume this file.
  - Controlled access
- Produced from ABSOLUTE GATK4 CNV Auto
  - A PDF that contains all potential CNV models (purity + ploidy combinations)

Other Copy Number Derived Data

- Property values on Copy Number Segment
  - Tumor purity – ratio of tumor cells to total cells.
  - Tumor ploidy – number of chromosome sets in a tumor sample





Tool Demo

CNV API Information and Demo

GDC Copy Number Variation API

The GDC API has two sets of copy number variation endpoints

- *cnv* and *cnv_occurrences:* provide information about specific gene-level CNV mutations
- *segment_cnvs* and *segment_cnv_occurrences:* provide information about specific CNV segments

CNVs in the GDC API are classified as:

- *cnv_change*: Gain, (Neutral), Loss
- *cnv_change_5_category*: Amplification, Gain, (Neutral), Heterozygous Deletion, Homozygous Deletion

Questions?

U.S. Department of Health & Human Services

National Institutes of Health | National Cancer Institute

https://www.cancer.gov/

1-800-4-CANCER

Produced August 2025