

GDC WGS Variant Calling Workflow Updates

31 March 2025

Bill Wysocki, Ph.D – GDC Director of User Services
Zhenyu Zhang, Ph.D – GDC Director of Bioinformatics
Center for Translational Data Science
University of Chicago

Agenda

1. *Overview of Whole Genome Variants in the GDC*
2. *Available WGS Data*
3. *WGS Pipeline Overview*
4. *Downloading WGS Data*
5. *Documentation and Repositories*
6. *Questions*

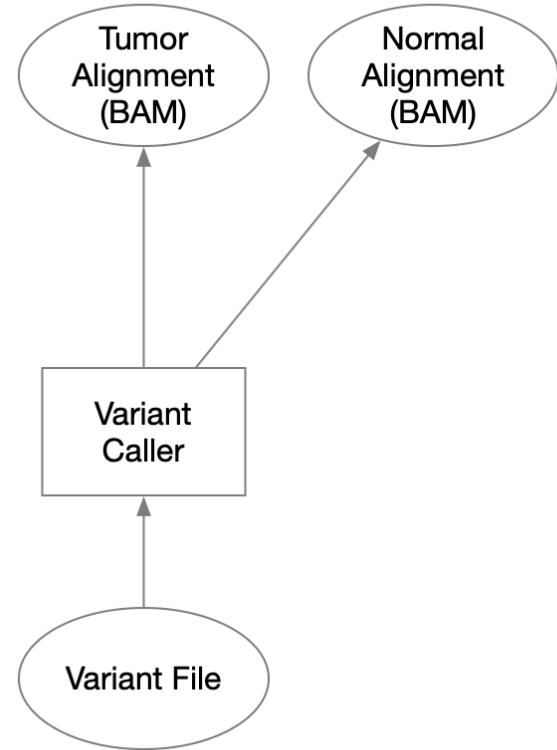


Overview of Whole Genome Variants in the GDC

Somatic Mutations

WGS Data in the GDC

- Tumor and normal samples are aligned
- The resulting alignment are processed with variant caller workflows of different types:
 - Point mutations/ Indels (SSM)
 - Copy number variants
 - Structural variants
- There are two WGS workflow sets -



Existing Sanger WGS Workflow Set

1. *Point Mutation: CaVEMan*
2. *Indel: Pindel*
3. *CNV: ascatNGS*
4. *SV: BRASS*

New GDC WGS Workflow Set

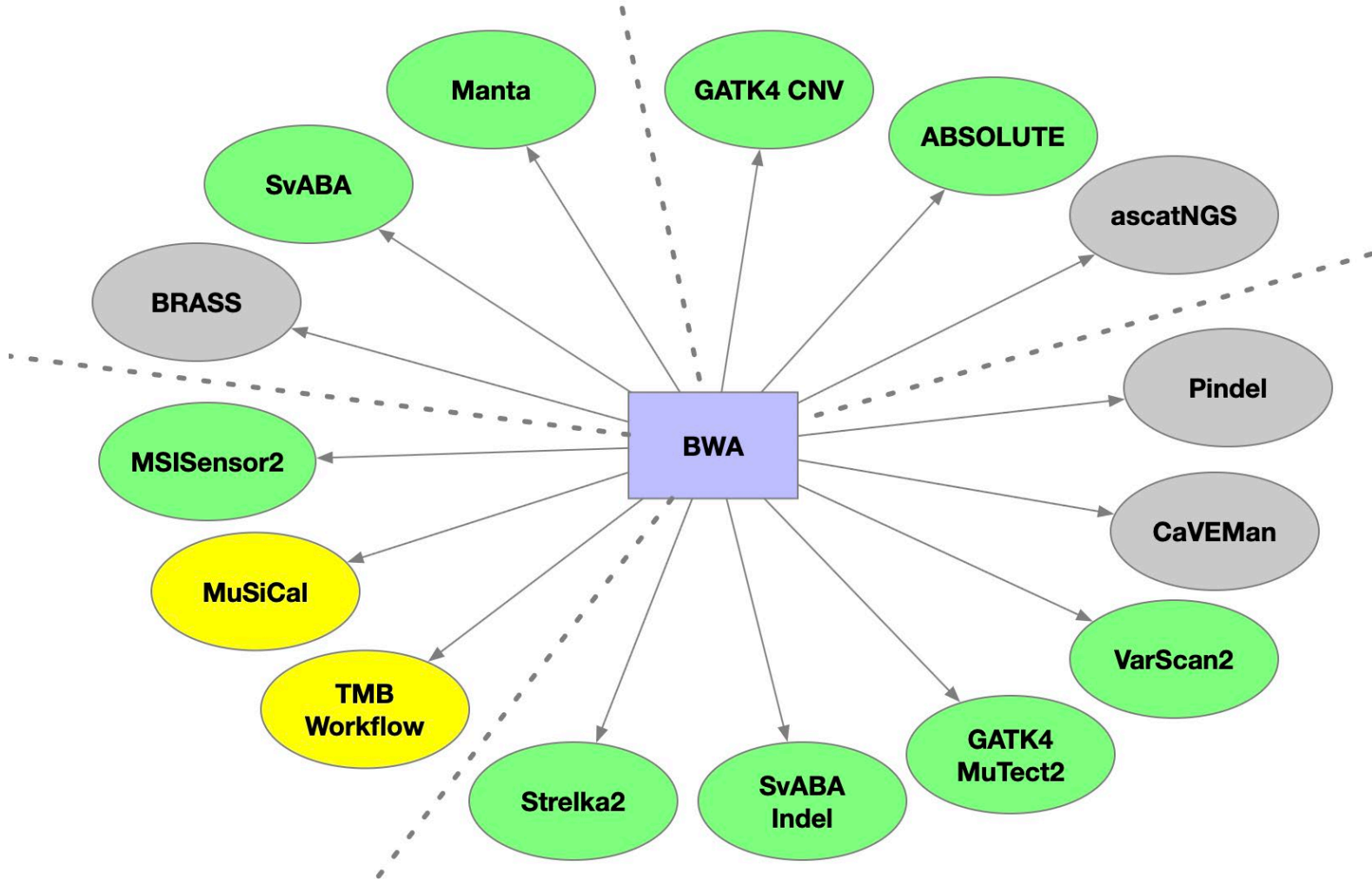
1. *Simple Somatic Mutation (SSM):*
 - *GATK4 MuTect2*
 - *VarScan2*
 - *SvABA Indel*
 - *Strelka2*
2. *Copy Number Variation (CNV):*
 - *GATK4 CNV + ABSOLUTE*
3. *Structural Variation (SV):*
 - *Manta*
 - *SvABA*

Summary of Major Changes

1. *Increase SSM calling from 2 callers to 4, which enables proper variant ensemble.*
2. *Increase SV calling from 1 caller to 2 callers.*
3. *Add ABSOLUTE for better CNV inference.*
4. *Stop Sanger workflow production, and run new WGS workflows across the GDC.*

WGS Data Releases in the GDC

- Data Release 27 (2020):
 - WGS variants were first released from the Sanger workflow set
- Data Release 42 (2025):
 - WGS variants were first released from the new GDC workflow set
- **Transition:**
 - All data that was processed with the Sanger workflow set will be processed with the new GDC workflow set
 - WGS processing is currently in progress, any pipeline that is completed will be released in the subsequent data release
 - For SSMs (point + indels), this means that an annotated VCF is available



Available WGS Data

Currently Available in the GDC Data Portal

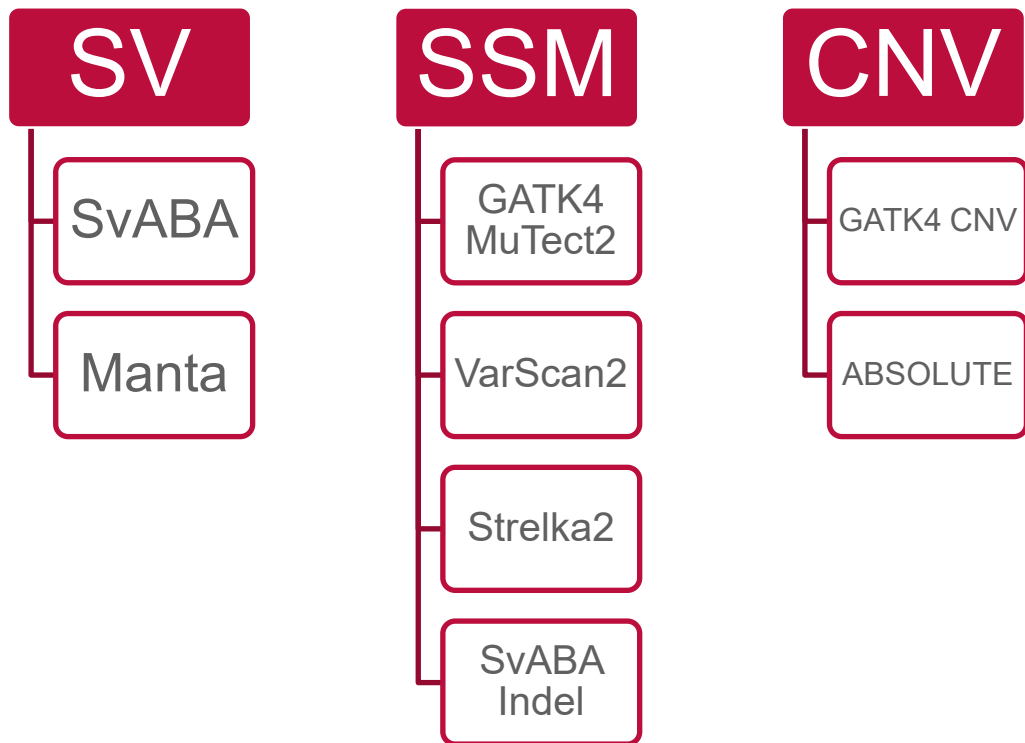
1. *More than 15,000 cases with aligned WGS BAMs*
2. *5,000+ cases with Sanger calls. We no longer use Sanger for production and will gradually release the rest of the Sanger calls that have already been completed.*
3. *Data from <14,000 cases of Manta, GATK4 CNV, GATK4 MuTect2, SvABA Indel callers have been released. We will release more in future releases and hopefully cover all WGS data in the GDC soon.*

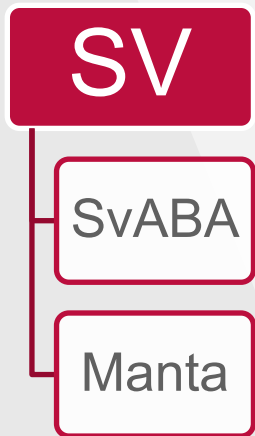
WGS Pipeline Overview

Why Switch Pipelines?

1. *SSM calling: increased from 2 callers to 4, that enables multi-caller variant ensemble.*
2. *SV calling: increased from 1 caller to 2, that enables users to derive all confidence SV set.*
3. *CNV calling: added ABSOLUTE for better CNV inference.*
4. *Efficiency: the new workflows, particularly modern tools like Strelka2 and Manta, require significantly fewer compute resources than the existing Sanger workflows.*

WGS Workflows





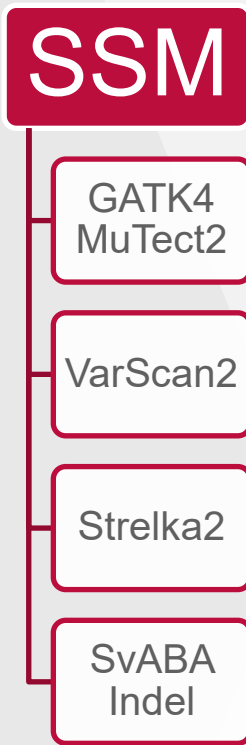
Output: VCF and BedPE

1. SvABA SV Caller

- *Generate both VCF and BedPE.*
- *Also generate an Indel VCF for SSM ensemble.*

2. Manta SV Caller

- *Generate both VCF and BedPE.*
- *Also generate a candidate SSM VCF. This VCF is then fed into Strelka2 calling to enhance the quality and coverage of Strelka2.*



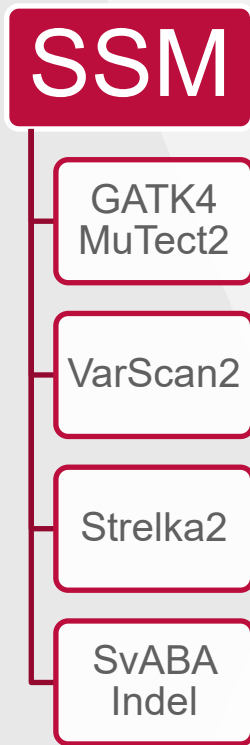
Output: raw VCF, annotated VCF, aliquot-level MAF, 4-caller ensemble MAF

1. GATK4 MuTect2

- *Different from the MuTect2 used in WXS calling (GATK3).*
- *Known issue: Missing variants on chr10 and chr20 in currently released VCFs. We expect to fix in later data releases.*

2. VarScan2

- *The same caller used in WXS calling.*
- *Not released yet.*



Output: raw VCF, annotated VCF, aliquot-level MAF, 4-caller ensemble MAF

3. *Strelka2*

- *Implemented as Strelka2 – Manta joint calling. Manta candidate VCFs is further filtered by Strelka2 calling to enhance the quality and coverage of Strelka2.*

4. *SvABA Indel*

- *Indel output from the SvABA SV calling.*



Output: Segmentation, gene-level CNV, purity/ploidy estimation

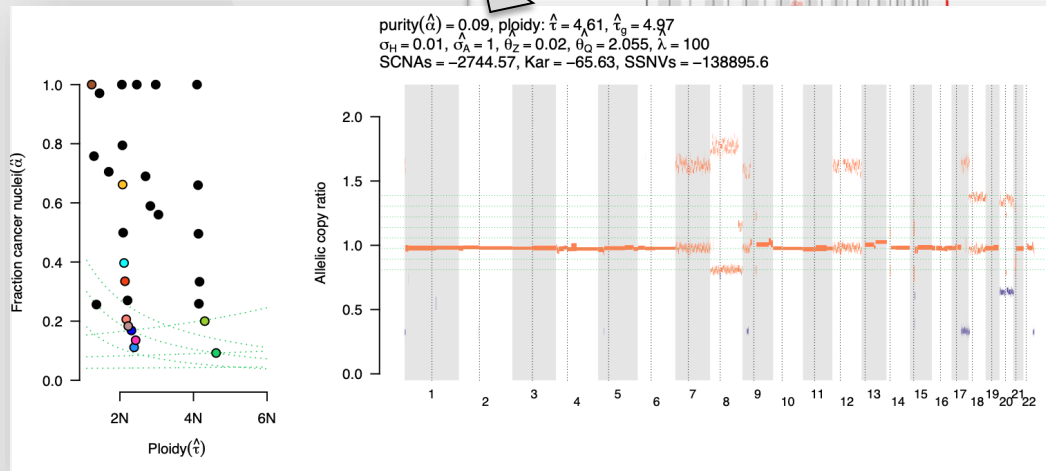
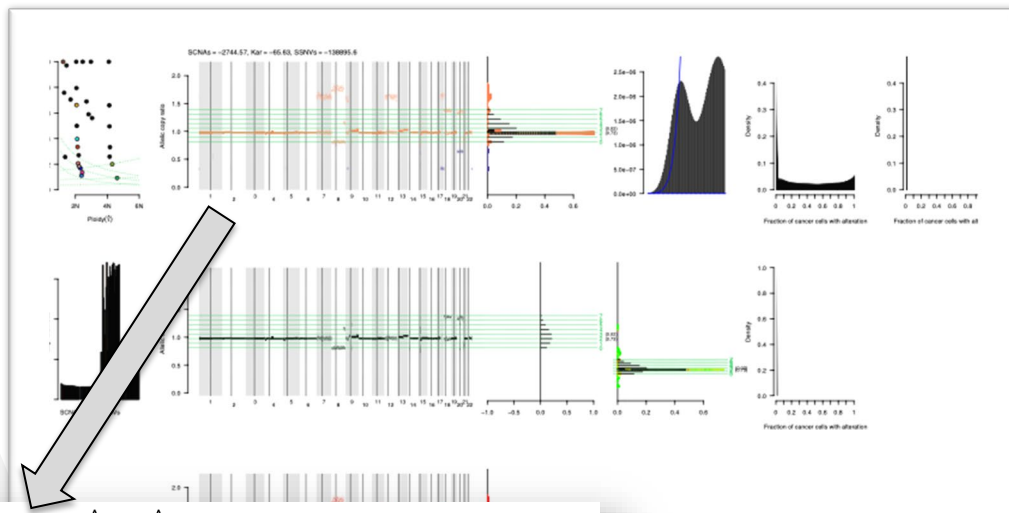
1. *GATK4 CNV caller*
 - *Non-integer copy number segmentation file (segmean).*
 - *Per request by genomics analysis groups, we also provide a controlled-access intermediate analysis archive tarball for expert manual review purpose.*
2. *ABSOLUTE CNV Caller*
 - *Use GATK4 CNV and GATK4 MuTect2 input to derive integer-level copy number segmentation, model PDF, and gene-level copy numbers.*
 - *Not released yet.*



*Output: Segmentation,
gene-level CNV, purity/
ploidy estimation*

- *More on ABSOLUTE*
 - *ABSOLUTE is run in two steps: 1) “model generation” that generates multiple possible CNV models (purity and ploidy combinations); 2) “model extraction” that extracts the segmentations from one selected model.*
 - *We plan to provide a PDF report that contains all models from a run, and also automatic model extraction of the first model called “ABSOLUTE GATK4 CNV Auto”. With curation provided by community experts, we may later release “ABSOLUTE GATK4 CNV Curated” when available.*

ABSOLUTE Model PDF Report



Other WGS Data and Workflows

1. *BAM metrics, including coverage and other information. They are on the AlignedReads node, and available via API query, various facet filtering in GDC portal, and single BAM entity page.*
2. *Microsatellite Instability (MSI) by MSISensor2, at the same place as other BAM metrics, but only available in Tumor BAMs.*
3. *Tumor Purity and Ploidy. They are on the segmentation file node from ascatNGS and ABSOLUTE workflows, available via API.*

Development in Progress

1. *Tumor Mutational Burden (TMB)*
2. *Tumor Mutational Signature (MuSiCal)*

Downloading WGS Data (Demo)

Goal: Download the following:

- Annotated Somatic Mutation VCFs (SNVs + Indels)
- Manta BEDPE files

from cases with lung cancer

GDC 2.0 Workflow

Build Cohort



Cohort Builder

Build and define your custom cohorts using a variety of clinical and biospecimen features.

Download Cohort Data



Repository

Browse and download the files associated with your cohort for more sophisticated analysis.

View Projects



Projects

View the Projects available within the GDC and select them for further exploration and analysis.

Analyze Cohort

ANALYSIS TOOLS

BAM Slicing Download ▾
1,121 Cases

Clinical Data Analysis ▾
1,310 Cases

Cohort Comparison ▾
1,310 Cases

Gene Expression Clustering ▾
1,039 Cases

Mutation Frequency ▾
1,039 Cases

OncoMatrix ▾
1,039 Cases

ProteinPaint ▾
1,039 Cases

Sequence Reads ▾
1,121 Cases

Set Operations ▾

GDC 2.0 Workflow – Downloading Files

Build Cohort



Cohort Builder

Build and define your custom cohorts using a variety of clinical and biospecimen features.



Download Cohort Data



Repository

Browse and download the files associated with your cohort for more sophisticated analysis.



View Projects



Projects

View the Projects available within the GDC and select them for further exploration and analysis.



Analyze Cohort

ANALYSIS TOOLS

SAM Slicing Download •
1,321 Cases

Clinical Data Analysis •
1,310 Cases

Cohort Comparison •
1,310 Cases

Gene Expression Clustering •
1,039 Cases

Mutation Frequency •
1,039 Cases

OncoMatrix •
1,039 Cases

ProteinPaint •
1,039 Cases

Sequence Reads •
1,021 Cases

Set Operations •

Genomic Data Commons Data Portal

Harmonized Cancer Datasets

A repository and computational platform for cancer researchers who need to understand cancer, its clinical progression, and response to therapy.

Explore Our Cancer Datasets

Data Portal Summary

[Data Release 42.0 - January 30, 2025](#)



86
Projects



69
Primary Sites



44,736
Cases



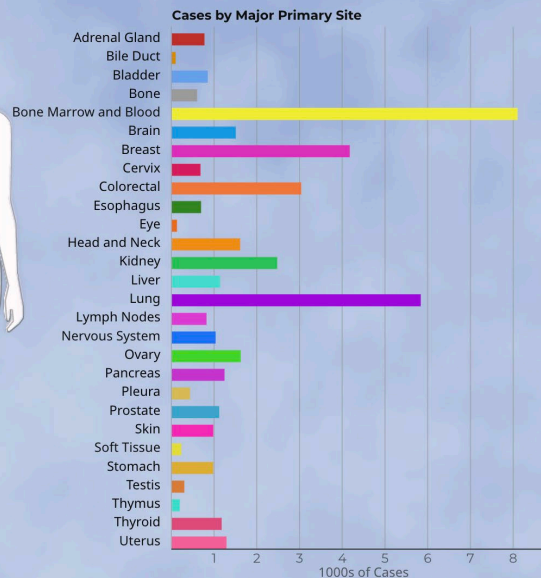
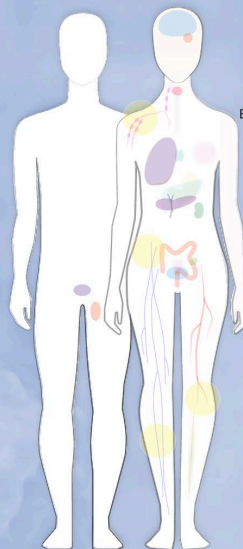
1,121,816
Files



22,534
Genes



2,940,240
Mutations



Unsaved_Cohort
44,736 CASES

Cohort not saved

General Diagnosis

☐ BEATAML1.0	882 (1.97%)
☐ CDDP_EAGLE	50 (0.11%)
☐ CGCI	645 (1.44%)
☐ CMI	299 (0.67%)
☐ CPTAC	1,687 (3.77%)

+ 16 more

Disease Status and History

☐ BEATAML1.0-COHORT	826 (1.85%)
☐ BEATAML1.0-CRENOLANIB	56 (0.13%)
☐ CDDP_EAGLE-1	50 (0.11%)
☐ CGCI-BLGSP	324 (0.72%)
☐ CGCI-HTMCP-CC	212 (0.47%)

+ 80 more

Disease Specific Classifications

☐ acute lymphoblastic leukemia	1,086 (2.43%)
☐ adenomas and adenocarcinom...	14,344 (32.06%)
☐ adnexal and skin appendage ne...	22 (0.05%)
☐ basal cell neoplasms	22 (0.05%)
☐ blood vessel tumors	1 (0.00%)

+ 39 more

Treatment

Primary Diagnosis	
Name	Cases
☐ acinar adenocarcinoma	187 (0.42%)
☐ acinar cell carcinoma	266 (0.59%)
☐ acinar cell tumor	31 (0.07%)
☐ acute leukemia, nos	11 (0.02%)
☐ acute lymphoblastic leukemia, n...	144 (0.32%)
☐ acute lymphocytic leukemia	1,170 (2.62%)

+ 194 more

Exposure

Primary Site	
Name	Cases
☐ accessory sinuses	1 (0.00%)
☐ adrenal gland	724 (1.62%)
☐ anus and anal canal	82 (0.18%)
☐ base of tongue	27 (0.06%)
☐ bladder	767 (1.71%)
☐ bones, joints and articular cartila...	257 (0.57%)

+ 63 more

Biospecimen

Tissue or Organ of Origin	
Name	Cases
☐ abdomen, nos	236 (0.53%)
☐ adrenal gland, nos	608 (1.36%)
☐ ampulla of vater	4 (0.01%)
☐ anal canal	1 (0.00%)
☐ anterior floor of mouth	2 (0.00%)
☐ anterior mediastinum	40 (0.09%)

+ 194 more

Molecular Filters

Available Data

Custom Filters

Case ID

Upload Cases

Unsaved_Cohort

Cohort not saved

Molecular Filters

Available Data

Custom Filters

Workflow Type

Name ^	Cases ^
☐ Aliquot Ensemble Somatic Varian...	2 (0.00%)
☐ AscatNGS	5,444 (12.17%)
☐ BRASS	5,444 (12.17%)
☐ BWA with Mark Duplicates and ...	15,271 (34.14%)
☐ CaVEMan	5,444 (12.17%)
☐ GATK4 CNV	13,473 (30.12%)
☐ GATK4 MuTect2	10,941 (24.46%)
☐ GATK4 MuTect2 Annotation	10,941 (24.46%)
☐ GATK4 MuTect2 Tumor-Only	2 (0.00%)
☐ GATK4 MuTect2 Tumor-Only Ann...	2 (0.00%)
☐ Manta	5,710 (12.76%)
☐ Pindel	5,444 (12.17%)
☐ SvABA	5,528 (12.36%)
☐ SvABA Indel	1,013 (2.26%)
☐ SvABA Indel Annotation	1,013 (2.26%)
☐ VCF LiftOver	458 (1.02%)

show less

Data Format

Name ^
☐ bam
☐ bedpe
☐ maf
☐ tar
☐ tsv
☐ txt

5,444 (12.17%)
13,728 (30.69%)

+ 1 more

Experimental Strategy

WGS

Name ^

Cases ^

☒ WGS

15,712 (35.12%)

show less

Unsaved_Cohort

ⓘ

+

10,945 CASES

Cohort not saved

Unsaved_Cohort

Clear All

<<

>>

EXPERIMENTAL STRATEGY

←

WGS

×

×

WORKFLOW TYPE

←

GATK4 MuTect2 Anno...

×

SvABA Indel Annotat...

×

×

× REPOSITORY

Filters

Collapse All

+ Add a Custom Filter

JSON

TSV

Download Associated Data

Manifest

View Images

Add All Files to Cart

Remove All From Cart

Experimental Strategy

🔍

🔗

↺

Name	Files
☐ ATAC-Seq	351 (0.03%)
☐ Diagnostic Slide	9,088 (0.81%)
☐ Expression Array	602 (0.05%)
☐ Genotyping Array	109,439 (9.76%)
☐ Methylation Array	33,417 (2.98%)
☐ miRNA-Seq	35,064 (3.13%)

TOTAL OF 658,307 FILES
10,945 CASES
6.91 PB

☰

Cart	Access	File Name	Cases	Project	Data Category	Data For
	Controlled	11c9e0b9-c94c-466f-877a-65325a76dcac.rna_seq.chimeric.qdc_realn.bam	1	TCGA-KIRC	Sequencing Reads	BAM
	Open	TCGA-KIRC.7181379b-1d8a-4d22-818d-b3f7c20a38be.ascat3.allelic_specific.seq.txt	1	TCGA-KIRC	Copy Number Variation	TXT
	Controlled	TCGA-KIRC.997e4ed0-8d01-4901-acb3-50a8f7d15a3c.arriba.rna_fusion.bedpe	1	TCGA-KIRC	Structural Variation	BEDPE
	Controlled	TCGA-KIRC.fc76ae2d-e4d0-4ef0-b01e-e4c8e649c74b.arriba.rna_fusion.bedpe	1	TCGA-KIRC	Structural Variation	BEDPE

Documentation and Repos

Documentation – <https://docs.gdc.cancer.gov>

The screenshot shows the GDC Documentation website. At the top, the NIH National Cancer Institute logo is on the left, and a search bar and 'GDC Apps' link are on the right. Below the header is a navigation bar with links: Home, API, Data Portal, Data Submission, Data Transfer Tool, Data Dictionary, **Data**, and Encyclopedia. The 'Data' link is highlighted with a red box. A red arrow points from this box to a link in the left sidebar: 'Bioinformatics Pipeline: DNA-Seq Analysis: Whole Genome Sequencing Variant Calling'. This link is also highlighted with a red box. The main content area is titled 'DNA-Seq: Whole Genome Sequencing Variant Calling' and 'Introduction'. It contains a paragraph about WGS variant calling pipelines and a list of software packages: Strelka, GATK4 MuTect2, SvABA, and Manta. A right sidebar contains a 'Table of contents' with links to various topics like 'Simple Nucleotide Variation', 'Structural Variant Pipelines', and 'Copy Number Variation'.

NIH NATIONAL CANCER INSTITUTE GDC Documentation

Search GDC Apps

Home API Data Portal Data Submission Data Transfer Tool Data Dictionary **Data** Encyclopedia

Data
Introduction
GDC Data Model
Data Security
File Format: MAF
File Format: VCF
Bioinformatics Pipeline: DNA-Seq Analysis: Whole Genome Sequencing Variant Calling
Bioinformatics Pipeline: DNA-Seq Analysis: Whole Exome and Targeted Sequencing Analysis Pipeline
Bioinformatics Pipeline: mRNA Analysis
Bioinformatics Pipeline: miRNA Analysis
Bioinformatics Pipeline: Copy Number Variation Analysis
Bioinformatics Pipeline: Methylation Analysis Pipeline
Bioinformatics Pipeline: Protein Expression
Aligned Reads Summary Metrics
Release Notes
Download PDF

DNA-Seq: Whole Genome Sequencing Variant Calling

Introduction

Variant calls from Whole Genome Sequencing (WGS) data are produced using pipelines distinct from those used for WXS and Targeted Sequencing samples. The GDC WGS variant calling workflows currently generate multiple downstream data types, including simple somatic mutations (SSMs), structural variants (SVs), and copy number variations (CNVs), leveraging the following software packages:

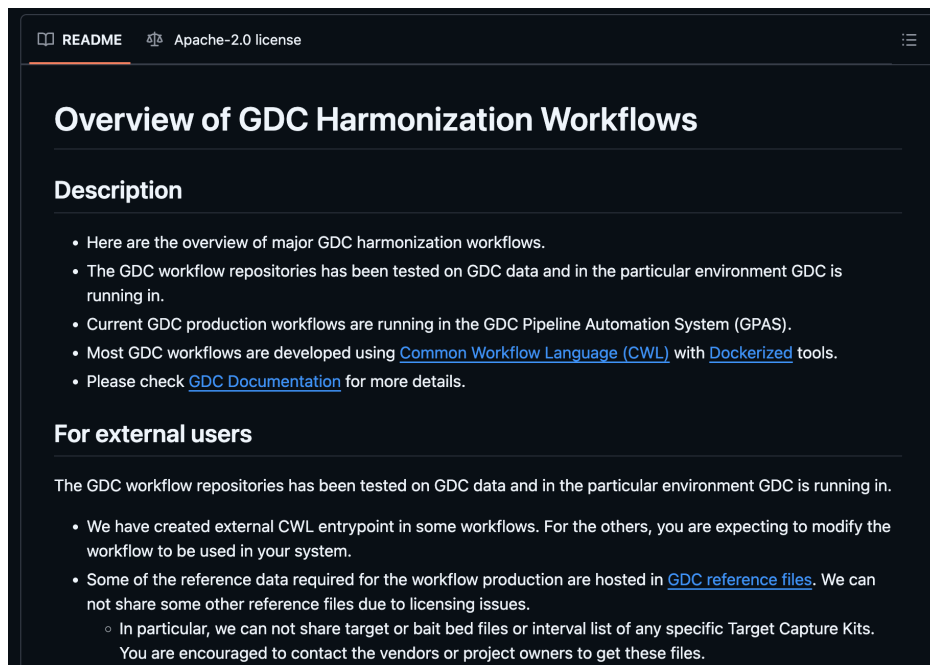
- **Strelka**: Simple nucleotide variants, including both point mutations and Indels, which are available in VCF format.
- **GATK4 MuTect2**: Simple nucleotide variants, including both point mutations and Indels, which are available in VCF format.
- **SvABA**: Indel variants only, which are available in VCF format, and structural variants, which are available in both VCF and BEDPE format.
- **Manta**: Structural variants, which are available in both VCF and BEDPE format.

• Additionally, Manta generates a set of candidate indels, which are subsequently used as input for the **Strelka** SSM calling pipeline to enhance Strelka's coverage across all indel sizes. Note that the Manta candidate indel file is not released on the GDC data portal.

Table of contents
Introduction
Simple Nucleotide Variation
GATK4 MuTect2 CLI
Step 1: Calculate Contamination
Step 2: MuTect2 Calling
Step 3: Filter MuTect2
Step 4: Filter Alignment Artifacts
Structural Variant Pipelines
SvABA CLI
Strelka-Manta CLI:
BEDPE File Format
Copy Number Variation
GATK4 CNV CLI
CNV from WGS File Format

Bioinformatics Pipeline Documentation and Resources:

<https://github.com/NCI-GDC/gdc-workflow-overview>



The screenshot shows the GitHub repository page for 'Overview of GDC Harmonization Workflows'. The page has a dark theme. At the top, there are tabs for 'README' and 'Apache-2.0 license'. The main heading is 'Overview of GDC Harmonization Workflows'. Below it is a section titled 'Description' with a bulleted list of information about GDC harmonization workflows. Another section titled 'For external users' follows, containing a paragraph and a bulleted list of instructions for users.

README Apache-2.0 license

Overview of GDC Harmonization Workflows

Description

- Here are the overview of major GDC harmonization workflows.
- The GDC workflow repositories has been tested on GDC data and in the particular environment GDC is running in.
- Current GDC production workflows are running in the GDC Pipeline Automation System (GPAS).
- Most GDC workflows are developed using [Common Workflow Language \(CWL\)](#) with [Dockerized](#) tools.
- Please check [GDC Documentation](#) for more details.

For external users

The GDC workflow repositories has been tested on GDC data and in the particular environment GDC is running in.

- We have created external CWL entrypoint in some workflows. For the others, you are expecting to modify the workflow to be used in your system.
- Some of the reference data required for the workflow production are hosted in [GDC reference files](#). We can not share some other reference files due to licensing issues.
 - In particular, we can not share target or bait bed files or interval list of any specific Target Capture Kits. You are encouraged to contact the vendors or project owners to get these files.

Questions?

U.S. Department of Health & Human Services
National Institutes of Health | National Cancer Institute

<https://www.cancer.gov/>

1-800-4-CANCER

Produced April 2024