GDC Data User's Guide

NCI Genomic Data Commons (GDC)

Contents

Chapter 1

Introduction

Introduction

The Genomic Data Commons receives, processes, and distributes genomic, clinical, and biospecimen data from cancer research programs. General information about data in the GDC can be found on the GDC website.

This document provides details about data included in the Genomic Data Commons, including information about the GDC data model, data formats, data processing, data security, and data releases.

Chapter 2

GDC Data Model

GDC Data Model

Introduction

The GDC Data Model is the central method of organization of all data artifacts in the GDC. An overview of the data model, including a visual representation of its components, is provided on the GDC website. This section provides technical details about its implementation for data users, submitters, and developers.

Entities, Properties, and Links

Although the GDC Data Model may contain some cyclic elements, it can be helpful to think of it as a Directed Acyclic Graph (DAG) composed of interconnected **entities**. Each entity in the GDC has a set of properties and links.

- **Properties** are key-value pairs associated with an entity. Properties cannot be nested, which means that the value must be numerical, boolean, or a string, and cannot be another key-value set. Properties can be either required or optional. The following properties are of particular importance in constructing the GDC Data Model:
  - **Type** is a required property for all entities. Entity types include `project`, `case`, `demographic`, `sample`, `read_group` and others.
  - **System properties** are properties used in GDC system operation and maintenance. They cannot be modified except under special circumstances.
  - **Unique keys** are properties, or combinations of properties, that can be used to uniquely identify the entity in the GDC. For example, the tuple (combination) of `[ project_id, submitter_id ]` is a unique key for most entities, which means that although `submitter_id` does not need to be unique in GDC, it must be unique within a project. See GDC Identifiers below for details.
- **Links** define relationships between entities, and the multiplicity of those relationships (e.g. one-to-one, one-to-many, many-to-many).

The GDC Data Dictionary determines which properties and links an entity can have according to entity `type`.

Functionally similar entity types are grouped under the same **category**. For example, entity types `slide_image` and `submitted_unaligned_reads` belong to `data_file` category, which comprises entities that represent downloadable files.

GDC Identifiers

UUIDs

When an entity is created, it is assigned a unique identifier in the form of a version 4 universally unique identifier (UUID). The UUID uniquely identifies the entity in the GDC, and is stored in the entity's `id` property.

Program Name, Project Code, and Project ID

Programs are the highest level of organization of GDC datasets. Each program is assigned a unique `program.name` property. Datasets within a program are organized into projects, and each project is assigned a `project.code` property.

The `project_id` property is associated with most entities in the GDC data model and is generated by appending `project.code` to `program.name` as follows:

```
program.name-project.code
(e.g. TCGA-LAML)
```

Note that `program.name` never contains hyphens.

Submitter ID

In addition to UUIDs stored in the `id` property, many entities also have a `submitter_id` property. This property can contain any string that the submitter wishes to use to identify the entity (e.g. a "barcode"). This can be used to identify a corresponding entry in the submitter's records. The GDC requires that `submitter_id` be unique for each entity within a project: the tuple (combination) of `[ project_id, submitter_id ]` is a unique key.

Note: The `submitter_id` of a `case` entity corresponds to the `submitted_subject_id` of the study participant in dbGaP records for the project.

Working with the GDC Data Model

Data Users

Users can access information stored in the GDC Data Model using the GDC Data Portal, the GDC API, and the GDC Data Transfer Tool. For more information see Data Access Processes and Tools.

Data Submitters

Data submitters can create and update submittable entities in the GDC Data Model and upload data files registered in the model using the GDC Data Submission Portal, the GDC API, and the GDC Data Transfer Tool. For more information see Data Submission Processes and Tools.

Chapter 3

Data Security

Data Security

To protect the privacy of research participants and support data integrity, the GDC requires user authorization and authentication for:

- downloading controlled-access data
- submitting data to the GDC

To perform these functions, GDC users must first obtain appropriate authorization via dbGaP and then authenticate via eRA Commons. The GDC sets user permissions at the project level according to dbGaP authorizations.

See Data Access Processes and Tools to learn more about the difference between open-access and controlled-access data.

Authorization via dbGaP

Instructions for obtaining authorization via dbGaP are provided in Obtaining Access to Controlled Data and Obtaining Access to Submit Data.

Authentication via eRA Commons

The following authentication methods are supported by the GDC:

GDC Tool	Authentication Method
GDC Data Portal	Log in using eRA Commons account
GDC Data Submission Portal	Log in using eRA Commons account
GDC Data Transfer Tool	Authentication Token
GDC API	Authentication Token

Authentication Tokens

The GDC Data Transfer Tool and the GDC API use tokens for authentication. GDC authentication tokens are alphanumeric strings of characters like this one:

```
ALPHANUMERICTOKEN-01234567890+AlPhAnUmErIcToKeN=0123456789-ALPHANUMERICTOKEN-01234567890+AlPhAnUmErIcToKeN=01234
```

Obtaining A Token

Users can obtain authentication tokens from the GDC Data Portal and the GDC Data Submission Portal. See the GDC Data Portal User's Guide and the GDC Data Submission Portal User's Guide for instructions.

Token Expiration

Tokens are valid for 30 days from the time of issue. Any request to the GDC API that uses an expired token will result in an error.

Tokens can be replaced at any time by downloading a new token, which will be valid for another 30 days.

Checking User Permissions

Users can view the permissions granted to them by the GDC system as follows:

0. Log into the GDC Data Portal or the GDC Data Submission Portal using your eRA Commons account.
1. Open the URL `https://portal.gdc.cancer.gov/auth/user` to see a JSON object that describes user permissions.

Chapter 4

File Format: MAF

GDC MAF Format v.1.0.0

Introduction

Mutation Annotation Format (MAF) is a tab-delimited text file with aggregated mutation information from VCF Files and are generated on a project-level. MAF files are produced through the Somatic Aggregation Workflow The GDC produces MAF files at two permission levels: **protected** and **somatic** (or open-access). One MAF files is produced per variant calling pipeline per GDC project. MAFs are produced by aggregating the GDC annotated VCF files generated from one pipeline for one project.

Annotated VCF files often have variants reported on multiple transcripts whereas the MAF files generated from the VCFs (*protected.maf) only report the most critically affected one. Somatic MAFs (*somatic.maf), which are also known as Masked Somatic Mutation files, are further processed to remove lower quality and potential germline variants. For tumor samples that contain variants from multiple combinations of tumor-normal aliquot pairs, only one pair is selected in the Somatic MAF based on their sample type. Somatic MAFs are publicly available and can be freely distributed within the boundaries of the GDC Data Access Policies.

The GDC MAF file format is based on the TCGA Mutation Annotation Format specifications, with additional columns included.

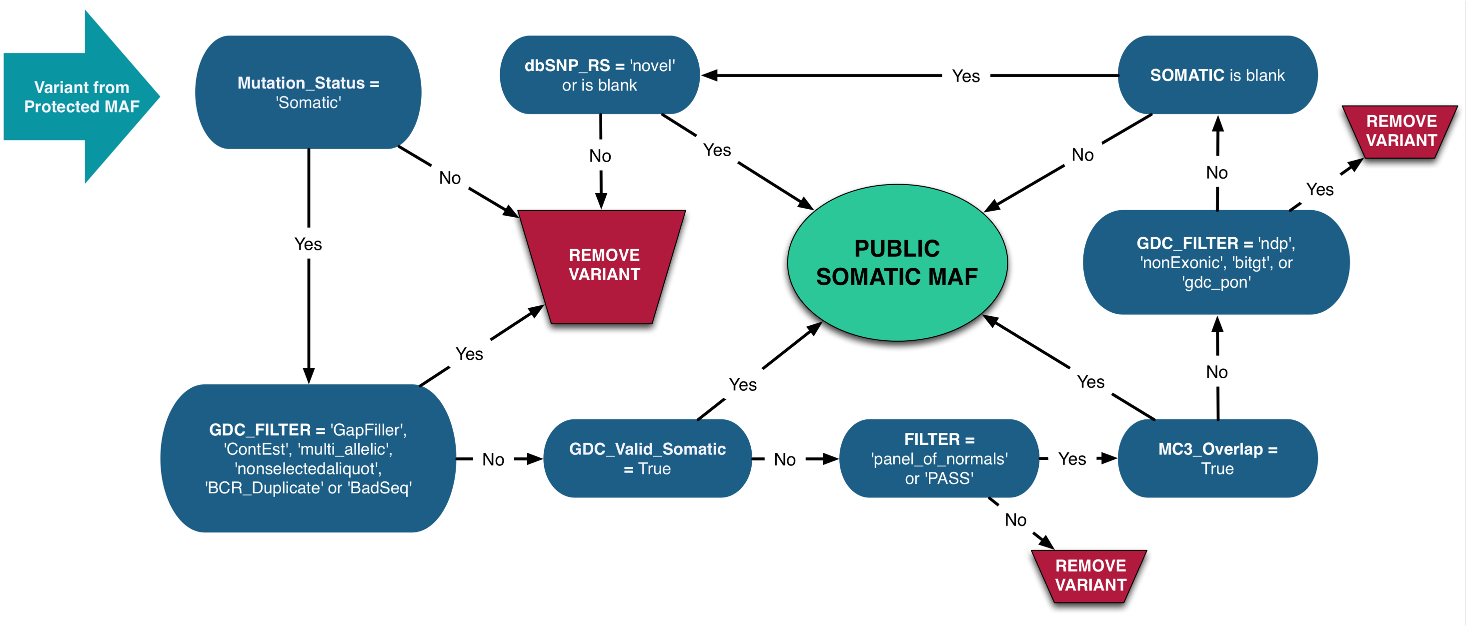
Note: The criteria for allowing mutations into open-access are purposefully implemented to overcompensate and filter out germline variants. If omission of true-positive somatic mutations is a concern, the GDC recommends using protected MAFs.

Somatic MAF File Generation

The process for modifying a protected MAF into a somatic MAF is as follows:

- Aliquot Selection: only one tumor-normal pair are selected for each tumor sample based on the plate number, sample type, analyte type and other features extracted from tumor TCGA aliquot barcode.
- Low quality variant filtering and germline masking:
  1. Variants with **Mutation_Status != 'Somatic'** or **GDC_FILTER = 'Gapfiller', 'ContEst', 'multiallelic', 'nonselectedaliquot', 'BCR_Duplicate' or 'BadSeq'** are **removed**.
  2. Remaining variants with **GDC_Valid_Somatic = True** are **included** in the Somatic MAF.
  3. Remaining variants with **FILTER != 'panel_of_normals' or PASS** are **removed**. Note that the `FILTER != panel_of_normals` value is only relevant for the variants generated from the MuTect2 pipeline.
  4. Remaining variants with **MC3_Overlap = True** are **included** in the Somatic MAF.
  5. Remaining variants with **GDC_FILTER = 'ndp', 'NonExonic', 'bitgt', 'gdc_pon'** are **removed**.
  6. Remaining variants with **SOMATIC != null** are **included** in the Somatic MAF.
  7. Remaining variants with **dbSNP_RS = 'novel' or null** are **included** in the Somatic MAF.
  8. Remaining variants are **removed**.

- Removal of the following columns:
  - vcf_region
  - vcf_info
  - vcf_format
  - vcf_tumor_gt
  - vcf_normal_gt
  - GDC_Valid_Somatic
- Set values to be blank in the following columns that may contain information about germline genotypes:
  - Match_Norm_Seq_Allele1
  - Match_Norm_Seq_Allele2
  - Match_Norm_Validation_Allele1
  - Match_Norm_Validation_Allele2
  - n_ref_count
  - n_alt_count



Variant from Protected MAF
Mutation_Status = 'Somatic'
dbSNP_RS = 'novel' or is blank
Yes
SOMATIC is blank
REMOVE VARIANT
No
No
Yes
No
No
Yes
Yes
REMOVE VARIANT
PUBLIC SOMATIC MAF
GDC_FILTER = 'ndp', 'nonExonic', 'bitgt', or 'gdc_pon'
Yes
Yes
No
Yes
Yes
GDC_FILTER = 'GapFiller', 'ContEst', 'multi_allelic', 'nonselectedaliquot', 'BCR_Duplicate' or 'BadSeq'
No
GDC_Valid_Somatic = True
No
FILTER = 'panel_of_normals' or 'PASS'
Yes
MC3_Overlap = True
No
REMOVE VARIANT

Figure 4.1: Somatic MAF Generation

Protected MAF File Structure

The table below describes the columns in a protected MAF and their definitions. Note that the somatic (open-access) MAF structure is the same except for having the last six columns removed.

Column	Description
1 - Hugo_Symbol	HUGO symbol for the gene (HUGO symbols are always in all caps). "Unknown" is used for regions that do not correspond to a gene
2 - Entrez_Gene_Id	Entrez gene ID (an integer). "0" is used for regions that do not correspond to a gene region or Ensembl ID

Column	Description
3 - Center	One or more genome sequencing center reporting the variant
4 - NCBI_Build	The reference genome used for the alignment (GRCh38)
5 - Chromosome	The affected chromosome (chr1)
6 - Start_Position	Lowest numeric position of the reported variant on the genomic reference sequence. Mutation start coordinate
7 - End_Position	Highest numeric genomic position of the reported variant on the genomic reference sequence. Mutation end coordinate
8 - Strand	Genomic strand of the reported allele. Currently, all variants will report the positive strand: '+'
9 - Variant_Classification	Translational effect of variant allele
10 - Variant_Type	Type of mutation. TNP (tri-nucleotide polymorphism) is analogous to DNP (di-nucleotide polymorphism) but for three consecutive nucleotides. ONP (oligo-nucleotide polymorphism) is analogous to TNP but for consecutive runs of four or more (SNP, DNP, TNP, ONP, INS, DEL, or Consolidated)
11 - Reference_Allele	The plus strand reference allele at this position. Includes the deleted sequence for a deletion or "-" for an insertion
12 - Tumor_Seq_Allele1	Primary data genotype for tumor sequencing (discovery) allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases
13 - Tumor_Seq_Allele2	Tumor sequencing (discovery) allele 2
14 - dbSNP_RS	The rs-IDs from the dbSNP database, "novel" if not found in any database used, or null if there is no dbSNP record, but it is found in other databases
15 - dbSNP_Val_Status	The dbSNP validation status is reported as a semicolon-separated list of statuses. The union of all rs-IDs is taken when there are multiple
16 - Tumor_Sample_Barcode	Aliquot barcode for the tumor sample
17 - Matched_Norm_Sample_Barcode	Aliquot barcode for the matched normal sample
18 - Match_Norm_Seq_Allele1	Primary data genotype. Matched normal sequencing allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases (cleared in somatic MAF)
19 - Match_Norm_Seq_Allele2	Matched normal sequencing allele 2
20 - Tumor_Validation_Allele1	Secondary data from orthogonal technology. Tumor genotyping (validation) for allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases
21 - Tumor_Validation_Allele2	Secondary data from orthogonal technology. Tumor genotyping (validation) for allele 2
22 - Match_Norm_Validation_Allele 1	Secondary data from orthogonal technology. Matched normal genotyping (validation) for allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases (cleared in somatic MAF)
23 - Match_Norm_Validation_Allele 2	Secondary data from orthogonal technology. Matched normal genotyping (validation) for allele 2 (cleared in somatic MAF)

Column	Description
24 - Verification_Status	Second pass results from independent attempt using same methods as primary data source. Generally reserved for 3730 Sanger Sequencing
25 - Validation_Status	Second pass results from orthogonal technology
26 - Mutation_Status	An assessment of the mutation as somatic, germline, LOH, post transcriptional modification, unknown, or none. The values allowed in this field are constrained by the value in the Validation_Status field
27 - Sequencing_Phase	TCGA sequencing phase (if applicable). Phase should change under any circumstance that the targets under consideration change
28 - Sequence_Source	Molecular assay type used to produce the analytes used for sequencing. Allowed values are a subset of the SRA 1.5 library_strategy field values. This subset matches those used at CGHub
29 - Validation_Method	The assay platforms used for the validation call
30 - Score	Not in use
31 - BAM_File	Not in use
32 - Sequencer	Instrument used to produce primary sequence data
33 - Tumor_Sample_UUID	GDC aliquot UUID for tumor sample
34 - Matched_Norm_Sample_UUID	GDC aliquot UUID for matched normal sample
35 - HGVSc	The coding sequence of the variant in HGVS recommended format
36 - HGVSp	The protein sequence of the variant in HGVS recommended format. “p.=” signifies no change in the protein
37 - HGVSp_Short	Same as the HGVSp column, but using 1-letter amino-acid codes
38 - Transcript_ID	Ensembl ID of the transcript affected by the variant
39 - Exon_Number	The exon number (out of total number)
40 - t_depth	Read depth across this locus in tumor BAM
41 - t_ref_count	Read depth supporting the reference allele in tumor BAM
42 - t_alt_count	Read depth supporting the variant allele in tumor BAM
43 - n_depth	Read depth across this locus in normal BAM
44 - n_ref_count	Read depth supporting the reference allele in normal BAM (cleared in somatic MAF)
45 - n_alt_count	Read depth supporting the variant allele in normal BAM (cleared in somatic MAF)
46 - all_effects	A semicolon delimited list of all possible variant effects, sorted by priority ([Symbol,Consequence,HGVSp_Short,Transcript_ID,RefSeq,HGVSc,Impact,Canonical,Sift,PolyPhen,Strand])
47 - Allele	The variant allele used to calculate the consequence
48 - Gene	Stable Ensembl ID of affected gene
49 - Feature	Stable Ensembl ID of feature (transcript, regulatory, motif)
50 - Feature_type	Type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature (or blank)

Column	Description
51 - One_Consequence	The single consequence of the canonical transcript in sequence ontology terms
52 - Consequence	Consequence type of this variant; sequence ontology terms
53 - cDNA_position	Relative position of base pair in the cDNA sequence as a fraction. A “-” symbol is displayed as the numerator if the variant does not appear in cDNA
54 - CDS_position	Relative position of base pair in coding sequence. A “-” symbol is displayed as the numerator if the variant does not appear in coding sequence
55 - Protein_position	Relative position of affected amino acid in protein. A “-” symbol is displayed as the numerator if the variant does not appear in coding sequence
56 - Amino_acids	Only given if the variation affects the protein-coding sequence
57 - Codons	The alternative codons with the variant base in upper case
58 - Existing_variation	Known identifier of existing variation
59 - ALLELE_NUM	Allele number from input; 0 is reference, 1 is first alternate etc.
60 - DISTANCE	Shortest distance from the variant to transcript
61 - TRANSCRIPT_STRAND	The DNA strand (1 or -1) on which the transcript/feature lies
62 - SYMBOL	The gene symbol
63 - SYMBOL_SOURCE	The source of the gene symbol
64 - HGNC_ID	Gene identifier from the HUGO Gene Nomenclature Committee if applicable
65 - BIOTYPE	Biotype of transcript
66 - CANONICAL	A flag (YES) indicating that the VEP-based canonical transcript, the longest translation, was used for this gene. If not, the value is null
67 - CCDS	The CCDS identifier for this transcript, where applicable
68 - ENSP	The Ensembl protein identifier of the affected transcript
69 - SWISSPROT	UniProtKB/Swiss-Prot accession
70 - TREMBL	UniProtKB/TrEMBL identifier of protein product
71 - UNIPARC	UniParc identifier of protein product
72 - RefSeq	RefSeq identifier for this transcript
73 - SIFT	The SIFT prediction and/or score, with both given as prediction (score)
74 - PolyPhen	The PolyPhen prediction and/or score
75 - EXON	The exon number (out of total number)
76 - INTRON	The intron number (out of total number)
77 - DOMAINS	The source and identifier of any overlapping protein domains
78 - GMAF	Non-reference allele and frequency of existing variant in 1000 Genomes
79 - AFR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined African population
80 - AMR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined American population

Column	Description
81 - ASN_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined Asian population
82 - EAS_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined East Asian population
83 - EUR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined European population
84 - SAS_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined South Asian population
85 - AA_MAF	Non-reference allele and frequency of existing variant in NHLBI-ESP African American population
86 - EA_MAF	Non-reference allele and frequency of existing variant in NHLBI-ESP European American population
87 - CLIN_SIG	Clinical significance of variant from dbSNP
88 - SOMATIC	Somatic status of each ID reported under Existing_variation (0, 1, or null)
89 - PUBMED	Pubmed ID(s) of publications that cite existing variant
90 - MOTIF_NAME	The source and identifier of a transcription factor binding profile aligned at this position
91 - MOTIF_POS	The relative position of the variation in the aligned TFBP
92 - HIGH_INF_POS	A flag indicating if the variant falls in a high information position of a transcription factor binding profile (TFBP) (Y, N, or null)
93 - MOTIF_SCORE_CHANGE	The difference in motif score of the reference and variant sequences for the TFBP
94 - IMPACT	The impact modifier for the consequence type
95 - PICK	Indicates if this block of consequence data was picked by VEP's pick feature (1 or null)
96 - VARIANT_CLASS	Sequence Ontology variant class
97 - TSL	Transcript support level, which is based on independent RNA analyses
98 - HGVS_OFFSET	Indicates by how many bases the HGVS notations for this variant have been shifted
99 - PHENO	Indicates if existing variant is associated with a phenotype, disease or trait (0, 1, or null)
100 - MINIMISED	Alleles in this variant have been converted to minimal representation before consequence calculation (1 or null)
101 - ExAC_AF	Global Allele Frequency from ExAC
102 - ExAC_AF_Adj	Adjusted Global Allele Frequency from ExAC
103 - ExAC_AF_AFR	African/African American Allele Frequency from ExAC
104 - ExAC_AF_AMR	American Allele Frequency from ExAC
105 - ExAC_AF_EAS	East Asian Allele Frequency from ExAC
106 - ExAC_AF_FIN	Finnish Allele Frequency from ExAC
107 - ExAC_AF_NFE	Non-Finnish European Allele Frequency from ExAC

Column	Description
108 - ExAC_AF_OTH	Other Allele Frequency from ExAC
109 - ExAC_AF_SAS	South Asian Allele Frequency from ExAC
110 - GENE_PHENO	Indicates if gene that the variant maps to is associated with a phenotype, disease or trait (0, 1, or null)
111 - FILTER	Copied from input VCF. This includes filters implemented directly by the variant caller and other external software used in the DNA-Seq pipeline. See below for additional details.
112 - CONTEXT	The reference allele per VCF specs, and its five flanking base pairs
113 - src_vcf_id	GDC UUID for the input VCF file
114 - tumor *bam*uuid	GDC UUID for the tumor bam file
115 - norma l_bam_uuid	GDC UUID for the normal bam file
116 - case_id	GDC UUID for the case
117 - GDC_FILTER	GDC filters applied universally across all MAFs
118 - COSMIC	Overlapping COSMIC variants
119 - MC3_Overlap	Indicates whether this region overlaps with an MC3 variant for the same sample pair
120 - GDC_V alidation_S tatus	GDC implementation of validation checks. See notes section (#5) below for details
121 - GDC_V alid_Somati c	True or False (not in somatic MAF)
122 - vcf_region	Colon separated string containing the CHROM, POS, ID, REF, and ALT columns from the VCF file (e.g., chrZ:20:rs1234:A:T) (not in somatic MAF)
123 - vcf_info	INFO column from VCF (not in somatic MAF)
124 - vcf_format	FORMAT column from VCF (not in somatic MAF)
125 - vcf_t umor_gt	Tumor sample genotype column from VCF (not in somatic MAF)
126 - vcf_n ormal_gt	Normal sample genotype column from VCF (not in somatic MAF)

Notes About GDC MAF Implementation

1. Column #4 **NCBI_Build** is GRCh38 by default
2. Column #32 **Sequencer** includes the sequencers used. If different sequencers were used to generate normal and tumor data, the normal sequencer is listed first.
3. Column #61 VEP name "STRAND" is changed to **TRANSCRIPT_STRAND** to avoid confusion with Column#8 "Strand"
4. Column #94 **IMPACT** categories are defined by the VEP software and do not necessarily reflect the relative biological influence of each mutation.
5. Column #122-125 **vcf_info, vcf_format, vcf_tumor_gt, and vcf_normal_gt** are the corresponding columns from the VCF files. Including them facilitates parsing specific variant information.
6. Column #120 **GDC_Validation_Status**: GDC also collects TCGA validation sequences. It compares these with

variants derived from Next-Generation Sequencing data from the same sample and populates the comparison result in "GDC_Validation_Status".

- "Valid", if the alternative allele(s) in the tumor validation sequence is(are) the same as GDC variant call
- "Invalid", if none of the alternative allele(s) in the tumor validation sequence is the same as GDC variant call
- "Inconclusive" if two alternative allele exists, and one matches while the other does not
- "Unknown" if no validation sequence exists

7. Column #121 **GDC_Valid_Somatic** is TRUE if GDC_Validation_Status is "Valid" and the variant is "Somatic" in validation calls. It is FALSE if these criteria are not met

FILTER Value Definitions (column 111)

- **oxog :** Signifies that this variant was determined to be an OxoG artifact. This was calculated with D-ToxoG
- **bPcr :** Signifies that this variant was determined to be an artifact of bias on the PCR template strand. This was calculated with the DKFZ Bias Filter.
- **bSeq :** Signifies that this variant was determined to be an artifact of bias on the forward/reverse strand. This was also calculated with the DKFZ Bias Filter.

Impact Categories

VEP

- **HIGH (H)**: The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function, or triggering nonsense mediated decay
- **MODERATE (M)**: A non-disruptive variant that might change protein effectiveness
- **LOW (L)**: Assumed to be mostly harmless or unlikely to change protein behavior
- **MODIFIER (MO)**: Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact

PolyPhen

- **probably damaging (PR)**: It is with high confidence supposed to affect protein function or structure
- **possibly damaging (PO)**: It is supposed to affect protein function or structure
- **benign (BE)**: Most likely lacking any phenotypic effect
- **unknown (UN)**: When in some rare cases, the lack of data does not allow PolyPhen to make a prediction

SIFT

- **tolerated**: Not likely to have a phenotypic effect
- **tolerated_low_confidence**: More likely to have a phenotypic effect than 'tolerated'
- **deleterious**: Likely to have a phenotypic effect
- **deleterious_low_confidence**: Less likely to have a phenotypic effect than 'deleterious'

Chapter 5

File Format: VCF

GDC VCF Format

Introduction

The GDC DNA-Seq somatic variant-calling pipeline compares a set of matched tumor/normal alignments and produces a VCF file. VCF files report the somatic variants that were detected by each of the four variant callers. Four raw VCFs (Data Type: Raw Simple Somatic Mutation) are produced for each tumor/normal pair of BAMs. Four additional annotated VCFs (Data Type: Annotated Somatic Mutation) are produced by adding biologically relevant information about each variant.

The GDC VCF file format follows standards of the Variant Call Format (VCF) Version 4.1 Specification. Raw Simple Somatic Mutation VCF files are unannotated, whereas Annotated Somatic Mutation VCF files include extensive, consistent, and pipeline-agnostic annotation of somatic variants.

VCF file structure

Metadata header

A VCF file starts with lines of metadata that begin with `##`. Some key components of this section include:

- **gdcWorkflow:** Information on the pipelines that were used by the GDC to generate the VCF file. Annotated VCF files contain two *gdcWorkflow* lines, one that reports the variant calling process and one that reports the variant annotation process.
- **INDIVIDUAL:** information about the study participant (`case`), including:
  - *NAME:* Submitter ID (barcode) associated with the participant
  - *ID:* GDC case UUID
- **SAMPLE:** sample information, including:
  - *ID:* NORMAL or TUMOR
  - *NAME:* Submitter ID (barcode) of the aliquot
  - *ALIQUOT_ID:* GDC aliquot UUID
  - *BAM_ID:* The UUID for the BAM file used to produce the VCF
- **INFO:** Format of *additional information* fields
  - NOTE: GDC Annotated VCFs may contain multiple INFO lines. The last INFO line contains information about annotation fields generated by the Somatic Annotation Workflow (see GDC INFO Fields below).
- **FILTER:** Description of filters that have been applied to the variants
- **FORMAT:** Description of genotype fields

- **reference:** The reference genome used to generate the VCF file (GRCh38.d1.vd1.fa)
- **contig:** A list of IDs for the contiguous DNA sequences that appear in the reference genome used to produce VCF files
  - NOTE: Annotated VCFs include contig information for autosomes, sex chromosomes, and mitochondrial DNA. Unplaced, unlocalized, human decoy, and viral genome sequences are not included.
- **VEP:** the VEP command used by the Somatic Annotation Workflow to generate the annotated VCF file.

Column Header Line

Each variant is represented by a row in the VCF file. Below each of the columns are described:

0. **CHROM:** The chromosome on which the variant is located
1. **POS:** The position of the variant on the chromosome. Refers to the first position if the variant includes more than one base
2. **ID:** A unique identifier for the variant; usually a dbSNP rs number if applicable
3. **REF:** The base(s) exhibited by the reference genome at the variant's position
4. **ALT:** The alternate allele(s), comma-separated if there are more than one
5. **QUAL:** Not populated
6. **FILTER:** The names of the filters that have flagged this variant. The types of filters used will depend on the variant caller used.
7. **INFO:** Additional information about the variant. This includes the annotation applied by the VEP.
8. **FORMAT:** The format of the sample genotype data in the next two columns. This includes descriptions of the colon-separated values.
9. **NORMAL:** Colon-separated values that describe the normal sample
10. **TUMOR:** Colon-separated values that describe the tumor sample

See Variant Call Format (VCF) Version 4.1 Specification for details.

GDC INFO fields

The following variant annotation fields are currently included in Annotated Somatic Mutation VCF files. Please refer to the DNA-Seq Analysis Pipeline documentation for details on how this information is generated. VEP Documentation provides additional information about some of these fields.

Field	Description
Allele	The variant allele used to calculate the consequence
Consequence	Consequence type of this variant
IMPACT	The impact modifier for the consequence type
SYMBOL	The HUGO gene symbol
Gene	Ensembl stable ID of the affected gene
Feature_type	Type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature.
Feature	Ensembl stable ID of the feature
BIOTYPE	The type of transcript or regulatory feature (e.g. protein_coding)
EXON	Exon number (out of total exons)
INTRON	Intron number (out of total introns)
HGVSc	The HGVS coding sequence name
HGVSp	The HGVS protein sequence name
cDNA_position	Relative position of base pair in cDNA sequence
CDS_position	Relative position of base pair in coding sequence

Field	Description
Protein_position	Relative position of the affected amino acid in protein
Amino_acids	Change in amino acids (only given if the variant affects the protein-coding sequence)
Codon	The affected codons with the variant base in upper case
Existing_variation	Known identifier of existing variant; usually a dbSNP rs number if applicable
ALLELE_NUM	Allele number from input; 0 is reference, 1 is first alternate, etc.
DISTANCE	Shortest distance from variant to transcript
STRAND	The DNA strand (1 or -1) on which the transcript/feature lies
FLAGS	Transcript quality flags
VARIANT_CLASS	Sequence Ontology variant class
SYMBOL_SOURCE	The source of the gene symbol
HGNC_ID	HGNC gene ID
CANONICAL	A flag indicating if the transcript is denoted as the canonical transcript for this gene
TSL	Transcript support level
APPRIS	APPRIS isoform annotation
CCDS	The CCDS identifer for this transcript, where applicable
ENSP	The Ensembl protein identifier of the affected transcript
SWISSPROT	UniProtKB/Swiss-Prot identifier of protein product
TREMBL	UniProtKB/TrEMBL identifier of protein product
UNIPARC	UniParc identifier of protein product
RefSeq	RefSeq gene ID
GENE_PHENO	Indicates if the gene is associated with a phenotype, disease or trait
SIFT	The SIFT prediction and/or score, with both given as prediction (score)
PolyPhen	The PolyPhen prediction and/or score
DOMAINS	The source and identifier of any overlapping protein domains
HGVS_OFFSET	Indicates by how many bases the HGVS notations for this variant have been shifted
GMAF	Non-reference allele and frequency of existing variant in 1000 Genomes
AFR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined African population
AMR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined American populatio
EAS_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined East Asian populati
EUR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined European populatio
SAS_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined South Asian popula
AA_MAF	Non-reference allele and frequency of existing variant in NHLBI-ESP African American population
EA_MAF	Non-reference allele and frequency of existing variant in NHLBI-ESP European American population
ExAC_MAF	Frequency of existing variant in ExAC combined population
ExAC_Adj_MAF	Adjusted frequency of existing variant in ExAC combined population
ExAC_AFR_MAF	Frequency of existing variant in ExAC African/American population
ExAC_AMR_MAF	Frequency of existing variant in ExAC American population
ExAC_EAS_MAF	Frequency of existing variant in ExAC East Asian population
ExAC_FIN_MAF	Frequency of existing variant in ExAC Finnish population

Field	Description
ExAC_NFE_MAF	Frequency of existing variant in ExAC Non-Finnish European population
ExAC_OTH_MAF	Frequency of existing variant in ExAC combined other combined populations
ExAC_SAS_MAF	Frequency of existing variant in ExAC South Asian population
CLIN_SIG	Clinical significance of variant from dbSNP
SOMATIC	Somatic status of existing variant(s)
PHENO	Indicates if existing variant is associated with a phenotype, disease or trait
PUBMED	Pubmed ID(s) of publications that cite existing variant
MOTIF_NAME	The source and identifier of a transcription factor binding profile aligned at this position
MOTIF_POS	The relative position of the variation in the aligned TFBP
HIGH_INF_POS	A flag indicating if the variant falls in a high information position of a transcription factor binding pr
MOTIF_SCORE_CHANGE	The difference in motif score of the reference and variant sequences for the TFBP
ENTREZ	Entrez ID
EVIDENCE	Evidence that the variant exists

Chapter 6

Bioinformatics Pipeline: DNA-Seq Analysis

DNA-Seq Analysis Pipeline

Introduction

The GDC DNA-Seq analysis pipeline identifies somatic variants within whole exome sequencing (WXS) and whole genome sequencing (WGS) data. Somatic variants are identified by comparing allele frequencies in normal and tumor sample alignments, annotating each mutation, and aggregating mutations from multiple cases into one project file.

The first pipeline starts with a reference alignment step followed by co-cleaning to increase the alignment quality. Four different variant calling pipelines are then implemented separately to identify somatic mutations. Somatic-caller-identified variants are then annotated. An aggregation pipeline incorporates variants from all cases in one project into a MAF file for each pipeline.

DNA-Seq analysis is implemented across six main procedures:

- Genome Alignment
- Alignment Co-Cleaning
- Somatic Variant Calling
- Variant Annotation
- Mutation Aggregation
- Aggregated Mutation Masking

Data Processing Steps

Pre-Alignment

Prior to alignment, BAM files that were submitted to the GDC are split by read groups and converted to FASTQ format. Reads that failed the Illumina chastity test are removed. Note that this filtering step is distinct from trimming reads using base quality scores.

Alignment Workflow

DNA-Seq analysis begins with the Alignment Workflow. Read groups are aligned to the reference genome using one of two BWA algorithms [1]. BWA-MEM is used if mean read length is greater than or equal to 70 bp. Otherwise BWA-aln is used. Each read group is aligned to the reference genome separately and all read group alignments that belong to a single aliquot are merged using Picard Tools SortSam and MergeSamFiles. Duplicate reads, which may persist as PCR artifacts, are then flagged to prevent downstream variant call errors.

Quality Control

Quality control metrics are collected before and after the alignment workflow and reviewed to identify potential low-quality data files. Basic metrics such as GC content and mean read length as well as quality score metrics are collected from unaligned reads using FASTQC. Quality metrics collected by the GDC for aligned reads include samtools idxstat and flagstat. Alignment information is collected using Picard CollectMultipleMetrics for both WGS and WXS. Coverage information is collected using picard CollectWgsMetrics for WGS and picard CollectHsMetrics for WXS.

Quality control metrics for each file endpoint can be accessed through the API using the `expand=analysis.metadata.read_groups,analy` parameter. Click here for an example query.

Reference Genome

All alignments are performed using the human reference genome GRCh38.d1.vd1. Decoy viral sequences are included in the reference genome to prevent reads from aligning erroneously and attract reads from viruses known to be present in human samples. Ten types of human viral genomes are included: human cytomegalovirus (CMV), Epstein-Barr virus (EBV), hepatitis B (HBV), hepatitis C (HCV), human immunodeficiency virus (HIV), human herpes virus 8 (HHV-8), human T-lymphotropic virus 1 (HTLV-1), Merkel cell polyomavirus (MCV), Simian vacuolating virus 40 (SV40), and human papillomavirus (HPV). Reference sequences used by the GDC can be downloaded here.

I/O	Entity	Format
Input	Submitted Unaligned Reads or Submitted Aligned Reads	FASTQ or BAM
Output	Aligned Reads	BAM

DNA-Seq Alignment Command Line Parameters

Step 1: Converting BAMs to FASTQs with Biobambam - biobambam2 2.0.54

```
bamtofastq \
collate=1 \
exclude=QCFAIL,SECONDARY,SUPPLEMENTARY \
filename= <input.bam> \
gz=1 \
inputformat=bam
level=5 \
outputdir= <output_path> \
outputperreadgroup=1 \
outputperreadgroupsuffixF=_1.fq.gz \
outputperreadgroupsuffixF2=_2.fq.gz \
outputperreadgroupsuffixO=_o1.fq.gz \
outputperreadgroupsuffixO2=_o2.fq.gz \
outputperreadgroupsuffixS=_s.fq.gz \
tryoq=1 \
```

Step 2: BWA Alignment - bwa 0.7.15 - samtools 1.3.1

If mean read length is greater than or equal to 70bp:

```
bwa mem \
-t 8 \
-T 0 \
-R <read_group> \
<reference> \
<fastq_1.fq.gz> \
```

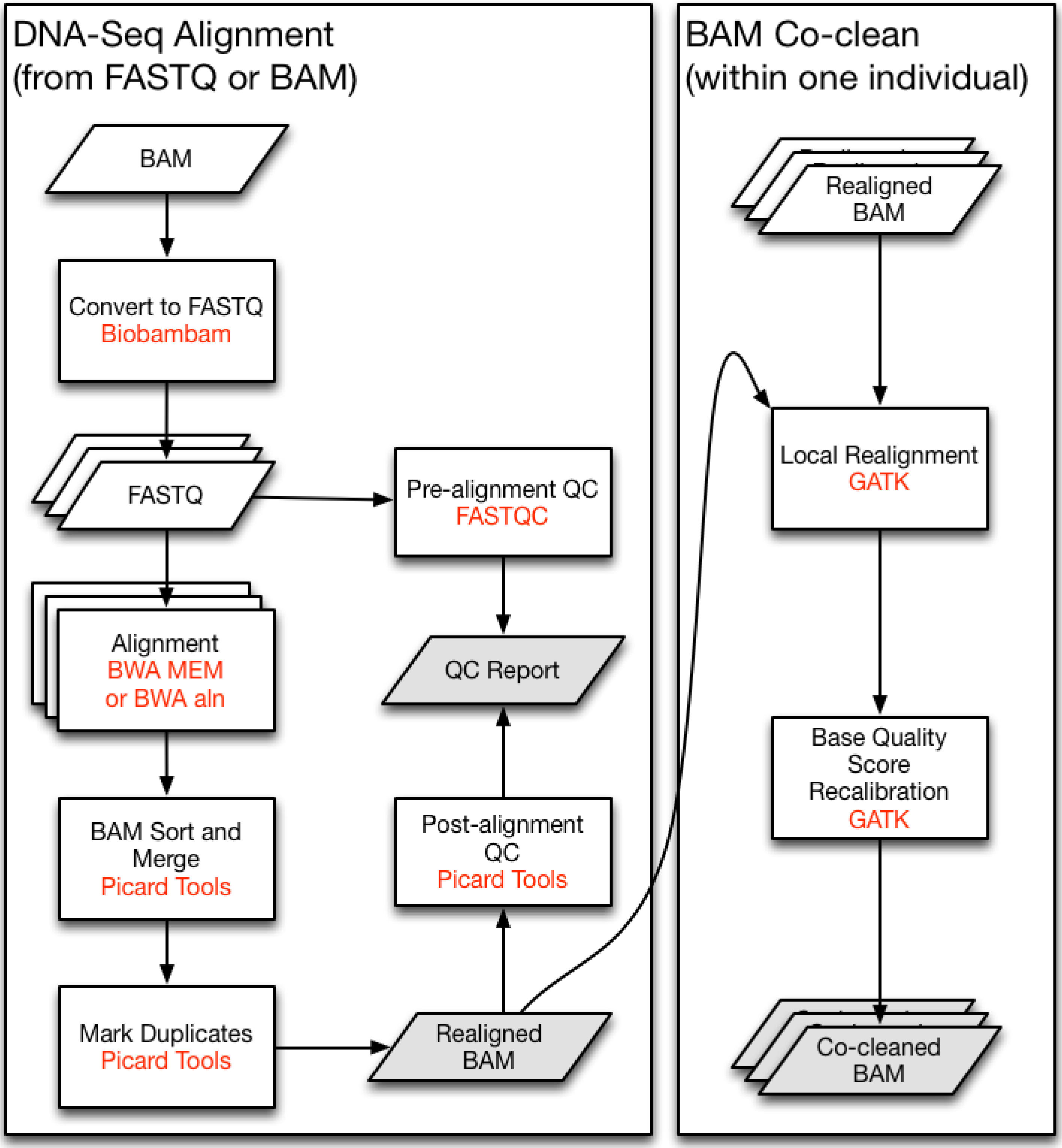


Figure 6.1: DNA-Seq Alignment Pipeline

```
<fastq_2.fq.gz> |
samtools view \
-Shb
-o <output.bam> -
```

If mean read length is less than 70bp:

```
bwa aln -t 8 <reference> <fastq_1.fq.gz> > <sai_1.sai> &&
bwa aln -t 8 <reference> <fastq_2.fq.gz> > <sai_2.sai> &&
bwa sampe -r <read_group> <reference> <sai_1.sai> <sai_2.sai> <fastq_1.fq.gz> <fastq_2.fq.gz> | samtools
    view -Shb -o <output.bam> -
```

If the quality scores are encoded as Illumina 1.3 or 1.5, use BWA aln with the "-l" flag.

Step 3: BAM Sort - picard 2.6.0

```
java -jar picard.jar SortSam \
CREATE_INDEX=true \
INPUT=<input.bam> \
OUTPUT=<output.bam> \
SORT_ORDER=coordinate \
VALIDATION_STRINGENCY=STRICT
```

Step 4: BAM Merge - picard 2.6.0

```
java -jar picard.jar MergeSamFiles \
ASSUME_SORTED=false \
CREATE_INDEX=true \
[INPUT= <input.bam>]  \
MERGE_SEQUENCE_DICTIONARIES=false \
OUTPUT= <output_path> \
SORT_ORDER=coordinate \
USE_THREADING=true \
VALIDATION_STRINGENCY=STRICT
```

Step 5: Mark Duplicates - picard 2.6.0

```
java -jar picard.jar MarkDuplicates \
CREATE_INDEX=true \
INPUT=<input.bam> \
VALIDATION_STRINGENCY=STRICT
```

Co-cleaning Workflow

The alignment quality is further improved by the Co-cleaning workflow. Co-cleaning is performed as a separate pipeline as it uses multiple BAM files (i.e. the tumor BAM and normal tissue BAM) associated with the same patient. Both steps of this process are implemented using GATK.

Indel Local Realignment

Local realignment of insertions and deletions is performed using IndelRealigner. This step locates regions that contain misalignments across BAM files, which can often be caused by insertion-deletion (indel) mutations with respect to the reference genome. Misalignment of indel mutations, which can often be erroneously scored as substitutions, reduces the accuracy of downstream variant calling steps.

Base Quality Score Recalibration

A base quality score recalibration (BQSR) step is then performed using BaseRecalibrator. This step adjusts base quality scores based on detectable and systematic errors. This step also increases the accuracy of downstream variant calling algorithms. Note that the original quality scores are kept in the OQ field of co-cleaned BAM files. These scores should be used if conversion of BAM files to FASTQ format is desired.

I/O	Entity	Format
Input	Aligned Reads	BAM
Output	Harmonized Aligned Reads	BAM

DNA-Seq Co-Cleaning Command Line Parameters

Step 1: RealignTargetCreator

```
java -jar GenomeAnalysisTK.jar \
-T RealignerTargetCreator \
-R <reference>
-known <known_indels.vcf>
[ -I <input.bam> ]
-o <realign_target.intervals>
```

Step 2: IndelRealigner

```
java -jar GenomeAnalysisTK.jar \
-T IndelRealigner \
-R <reference> \
-known <known_indels.vcf> \
-targetIntervals <realign_target.intervals> \
--noOriginalAlignmentTags \
[ -I <input.bam> ] \
-nWayOut <output.map>
```

Step 3: BaseRecalibrator; dbSNP v.144

```
java -jar GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R <reference> \
-I <input.bam> \
-knownSites <dbsnp.vcf>
-o <bqsr.grp>
```

Step 4: PrintReads

```
java -jar GenomeAnalysisTK.jar \
-T PrintReads \
-R <reference> \
-I <input.bam> \
--BQSR <bqsr.grp> \
-o <output.bam>
```

Somatic Variant Calling Workflow

Aligned and co-cleaned BAM files are processed through the Somatic Mutation Calling Workflow as tumor-normal pairs. Variant calling is performed using four separate pipelines:

- MuSE [2]
- MuTect2 [3]
- VarScan2 [4]
- SomaticSniper [5]

Variant calls are reported by each pipeline in a VCF formatted file. See the GDC VCF Format documentation for details on each available field. At this point in the DNA-Seq pipeline, all downstream analyses are branched into four separate paths that correspond to their respective variant calling pipeline.

Pipeline Descriptions

Four separate variant calling pipelines are implemented for GDC data harmonization. There is currently no scientific consensus on the best variant calling pipeline so the investigator is responsible for choosing the pipeline(s) most appropriate for the data. Some details about the pipelines are indicated below.

The MuTect2 pipeline employs a "Panel of Normals" to identify additional germline mutations. This panel is generated using TCGA blood normal genomes from thousands of individuals that were curated and confidently assessed to be cancer-free. This method allows for a higher level of confidence to be assigned to somatic variants that were called by the MuTect2 pipeline.

Basic outlines for the other three pipelines can be found here:

- VarScan2 pipeline
- MuSE pipeline
- SomaticSniper pipeline

Indels

Indel mutations that were generated with the MuTect2 and VarScan pipeline are detected and reported in GDC VCF files.

Germline Variants

At this time, germline variants are deliberately excluded as harmonized data. The GDC does not recommend using germline variants that were previously detected and stored in the Legacy Archive as they do not meet the GDC criteria for high-quality data.

I/O	Entity	Format
Input	Aligned Reads	BAM
Output	Raw Simple Somatic Mutation	VCF

Variant Call Command-Line Parameters

MuSE

MuSEv1.0rc_submission_c039ffa; dbSNP v.144

Step 1: MuSE call

```
MuSE call \
-f <reference> \
-r <region> \
<tumor.bam> \
<normal.bam> \
-O <intermediate_muse_call.txt>
```

Step 2: MuSE sump

```
MuSE sump \
-I <intermediate_muse_call.txt> \
-E \
-D <dbsnp_known_snp_sites.vcf> \
-O <muse_variants.vcf>
```

Note: -E is used for WXS data and -G can be used for WGS data.

MuTect2

GATK nightly-2016-02-25-gf39d340; dbSNP v.144

```
java -jar GenomeAnalysisTK.jar \
-T MuTect2 \
-R <reference> \
-L <region> \
-I:tumor <tumor.bam> \
-I:normal <normal.bam> \
--normal_panel <pon.vcf> \
--cosmic <cosmic.vcf> \
--dbsnp <dbsnp.vcf> \
--contamination_fraction_to_filter 0.02 \
-o <mutect_variants.vcf> \
--output_mode EMIT_VARIANTS_ONLY \
--disable_auto_index_creation_and_locking_when_reading_rods
```

SomaticSniper

Somatic-sniper v1.0.5.0

```
bam-somaticsniper \
-q 1 \
-L \
-G \
-Q 15 \
-s 0.01 \
-T 0.85 \
-N 2 \
-r 0.001 \
-n NORMAL \
-t TUMOR \
-F vcf \
-f ref.fa \
<tumor.bam> \
<normal.bam> \
<somaticsniper_variants.vcf>
```

VarScan

Step 1: Mpileup; Samtools 1.1

```
samtools mpileup \
-f <reference> \
-q 1 \
-B \
<normal.bam> \
<tumor.bam> >
<intermediate_mpileup.pileup>
```

Step 2: Varscan Somatic; Varscan.v2.3.9

```
java -jar VarScan.jar somatic \
<intermediate_mpileup.pileup> \
<output_path> \
--mpileup      1 \
--min-coverage 8 \
--min-coverage-normal 8 \
--min-coverage-tumor 6 \
--min-var-freq 0.10 \
--min-freq-for-hom 0.75 \
--normal-purity 1.0 \
--tumor-purity 1.00 \
--p-value 0.99 \
--somatic-p-value 0.05 \
--strand-filter 0 \
--output-vcf
```

Step 3: Varscan ProcessSomatic; Varscan.v2.3.9

```
java -jar VarScan.jar processSomatic \
<intermediate_varscan_somatic.vcf> \
--min-tumor-freq 0.10 \
--max-normal-freq 0.05 \
--p-value 0.07
```

Variant Call Annotation Workflow

Raw VCF files are then annotated in the Somatic Annotation Workflow with the Variant Effect Predictor (VEP) v84 [6] along with VEP GDC plugins.

The VEP uses the coordinates and alleles in the VCF file to infer biological context for each variant including the location of each mutation, its biological consequence (frameshift/ silent mutation), and the affected genes. See the documentation on the GDC VCF Format for more details. Variants in the VCF files are also matched to known variants from external mutation databases. The following databases are used for VCF annotation:

- GENCODE v.22
- sift v.5.2.2
- ESP v.20141103
- polyphen v.2.2.2
- dbSNP v.146
- Ensembl genebuild v.2014-07
- Ensembl regbuild v.13.0
- HGMD public v.20154
- ClinVar v.201601

Due to licensing constraints COSMIC is not utilized for annotation in the GDC VEP workflow.

In addition to annotation, False Positive Filter is used to label low quality variants in VarScan and SomaticSniper outputs. Variants with SSQ < 25 in SomaticSniper are also removed.

I/O	Entity	Format
Input	Simple Somatic Mutation	VCF

I/O	Entity	Format
Output	Annotated Somatic Mutation	VCF

Somatic Aggregation Workflow

The Somatic Aggregation Workflow generates one MAF file from multiple VCF files; see the GDC MAF Format guide for details on file structure. In this step, one MAF file is generated per variant calling pipeline for each project, and contains all available cases within this project.

I/O	Entity	Format
Input	Multiple Annotated Somatic Mutation	VCF
Output	Aggregated Somatic Mutation	MAF

Masked Somatic Aggregation Workflow

The MAF files generated by Somatic Aggregation Workflow are controlled-access due to the presence of germline mutations. Open-access MAF files are modified for public release by removing columns and variants that could potentially contain germline mutation information. See the GDC MAF Format for details about the criteria used to remove variants.

While these criteria cause the pipeline to over-filter some of the true positive somatic variants in open-access MAF files, they prevent personally identifiable germline mutation information from becoming publicly available. The GDC recommends that investigators explore both controlled and open-access MAF files if omission of certain somatic mutations is a concern.

I/O	Entity	Format
Input	Aggregated Somatic Mutation	Protected MAF
Output	Masked Somatic Mutation	Somatic MAF

File Access and Availability

Files from the GDC DNA-Seq analysis pipeline are available in the GDC Data Portal in BAM, VCF, and MAF formats. Descriptions are listed below for all available data types and their respective file formats.

Data Type	Description	File Format
Aligned Reads	Reads that have been aligned to the GRCh38 reference and co-cleaned. Unaligned reads and reads that map to decoy sequences are also included in the BAM files.	BAM
Raw Simple Somatic Mutation	A tab-delimited file with genotypic information related to genomic positions. Genomic variants are first identified here.	VCF
Annotated Somatic Mutation	An annotated version of a raw simple somatic mutation file. Annotated files include biological context about each observed mutation.	VCF
Aggregated Somatic Mutation	A tab-delimited file derived from multiple VCF files. Contains information from all available cases in a project.	MAF
Masked Somatic Mutation	A modified version of the Aggregated Somatic Mutation MAF file with sensitive or potentially erroneous data removed.	MAF

Data Type	Description	File Format

[1]. Li, Heng, and Richard Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform.” Bioinformatics 25, no. 14 (2009): 1754-1760.

[2]. Fan, Yu, Liu Xi, Daniel ST Hughes, Jianjun Zhang, Jianhua Zhang, P. Andrew Futreal, David A. Wheeler, and Wenyi Wang. “Accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling for sequencing data.” bioRxiv (2016): 055467.

[3]. Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.” Nature biotechnology 31, no. 3 (2013): 213-219.

[4]. Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. “VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.” Genome research 22, no. 3 (2012): 568-576.

[5]. Larson, David E., Christopher C. Harris, Ken Chen, Daniel C. Koboldt, Travis E. Abbott, David J. Dooling, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. “SomaticSniper: identification of somatic point mutations in whole genome sequencing data.” Bioinformatics 28, no. 3 (2012): 311-317.

[6] McLaren, William, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. “Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.” Bioinformatics 26, no. 16 (2010): 2069-2070.

Chapter 7

Bioinformatics Pipeline: mRNA Analysis

mRNA Analysis Pipeline

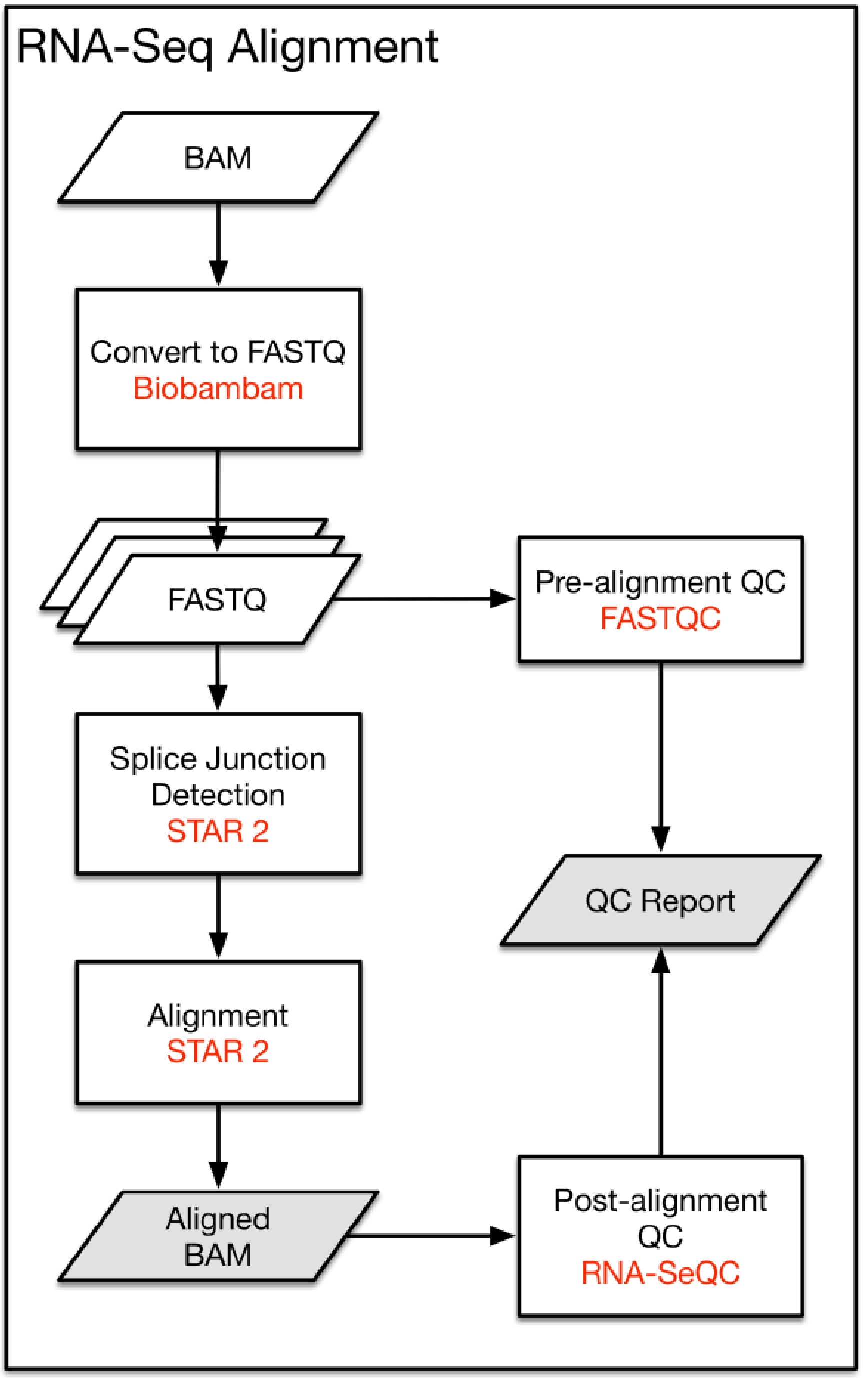
Introduction

The GDC mRNA quantification analysis pipeline measures gene level expression in HT-Seq raw read count, Fragments per Kilobase of transcript per Million mapped reads (FPKM), and FPKM-UQ (upper quartile normalization). These values are generated through this pipeline by first aligning reads to the GRCh38 reference genome and then by quantifying the mapped reads. To facilitate harmonization across samples, all RNA-Seq reads are treated as unstranded during analyses.

Data Processing Steps

RNA-Seq Alignment Workflow

The mRNA Analysis pipeline begins with the Alignment Workflow, which is performed using a two-pass method with STAR. STAR aligns each read group separately and then merges the resulting alignments into one. Following the methods used by the International Cancer Genome Consortium ICGC (github), the two-pass method includes a splice junction detection step, which is used to generate the final alignment. This workflow outputs a BAM file, which contains both aligned and unaligned reads. Quality assessment is performed pre-alignment with FASTQC and post-alignment with RNA-SeQC and Picard Tools.



I/O	Entity	Format
Inpu t	Submitted Unaligned Reads or Submitted Aligned Reads	FASTQ or BAM
Outp ut	Aligned Reads	BAM

RNA-Seq Alignment Command Line Parameters

STAR-2.4.2a

ICGC STAR alignment pipeline **For users with access to the ICGC pipeline:**

```
python star_align.py \
--genomeDir <star_index_path> \
--FastqFileIn <input_fastq_path> \
--workDir <work_dir> \
--out <output_bam> \
--genomeFastaFiles <reference> \
--runThreadN 8 \
--outFilterMultimapScoreRange 1 \
--outFilterMultimapNmax 20 \
--outFilterMismatchNmax 10 \
--alignIntronMax 500000 \
--alignMatesGapMax 1000000 \
--sjdbScore 2 \
--limitBAMsortRAM 0 \
--alignSJDBoverhangMin 1 \
--genomeLoad NoSharedMemory \
--outFilterMatchNminOverLread 0.33 \
--outFilterScoreMinOverLread 0.33 \
--twopass1readsN -1 \
--sjdbOverhang 100 \
--outSAMstrandField intronMotif \
--outSAMunmapped Within
```

For users without access to the ICGC pipeline:

Step 1: Building the STAR index.*

```
STAR
--runMode genomeGenerate
--genomeDir <star_index_path>
--genomeFastaFiles <reference>
--sjdbOverhang 100
--sjdbGTFfile <gencode.v22.annotation.gtf>
--runThreadN 8
```

*These indices are available for download at the GDC Website and do not need to be built again.

Step 2: Alignment 1st Pass.

```
STAR
--genomeDir <star_index_path>
--readFilesIn <fastq_left_1>,<fastq_left2>,... <fastq_right_1>,<fastq_right_2>,...
--runThreadN <runThreadN>
--outFilterMultimapScoreRange 1
--outFilterMultimapNmax 20
--outFilterMismatchNmax 10
--alignIntronMax 500000
--alignMatesGapMax 1000000
--sjdbScore 2
--alignSJDBoverhangMin 1
--genomeLoad NoSharedMemory
--readFilesCommand <bzcat|cat|zcat>
--outFilterMatchNminOverLread 0.33
--outFilterScoreMinOverLread 0.33
```

```
--sjdbOverhang 100
--outSAMstrandField intronMotif
--outSAMtype None
--outSAMmode None
```

Step 3: Intermediate Index Generation.

```
STAR
--runMode genomeGenerate
--genomeDir <output_path>
--genomeFastaFiles <reference>
--sjdbOverhang 100
--runThreadN <runThreadN>
--sjdbFileChrStartEnd <SJ.out.tab from previous step>
```

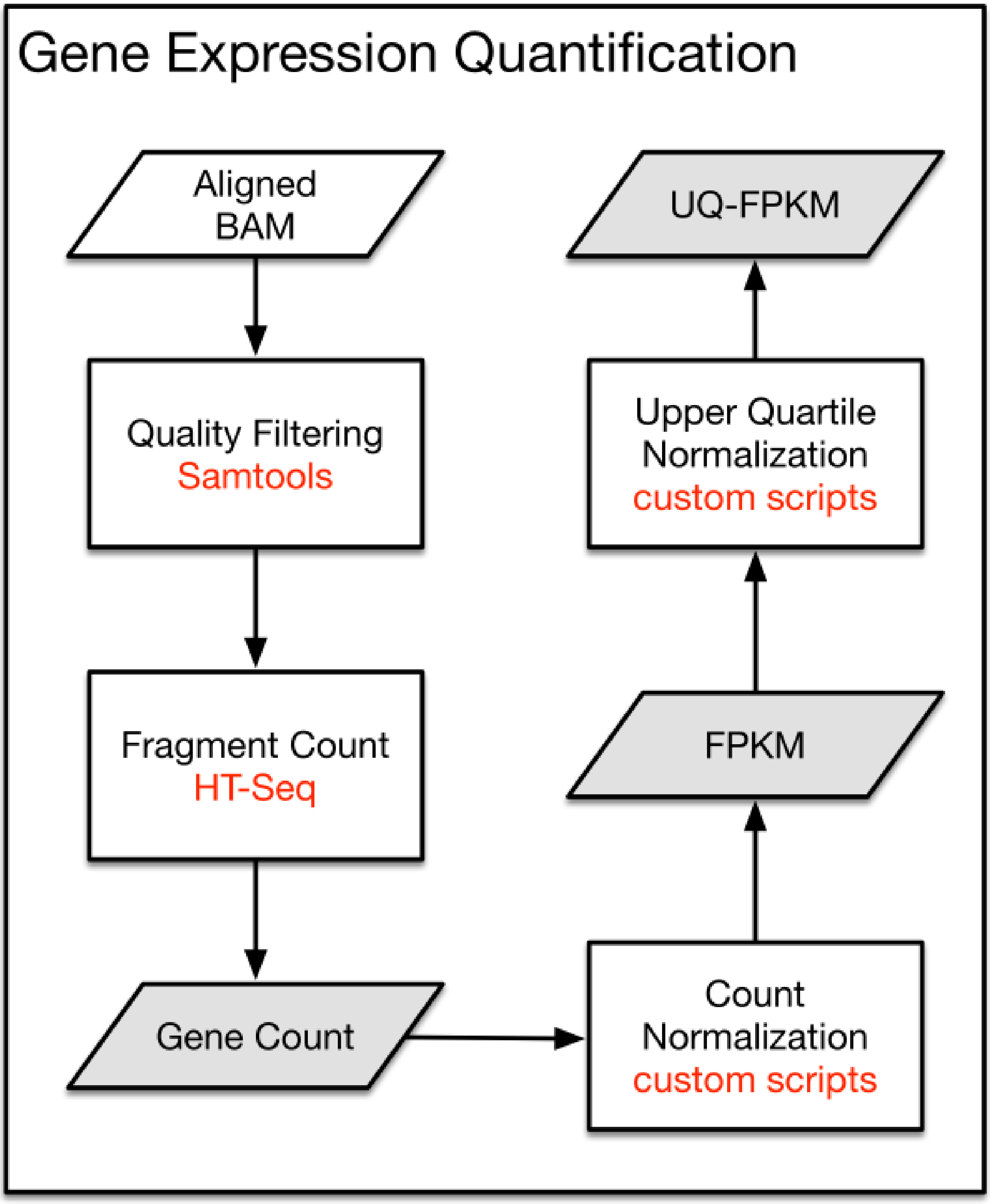
Step 4: Alignment 2nd Pass.

```
STAR
--genomeDir <output_path from previous step>
--readFilesIn <fastq_left_1>,<fastq_left2>,... <fastq_right_1>,<fastq_right_2>,...
--runThreadN <runThreadN>
--outFilterMultimapScoreRange 1
--outFilterMultimapNmax 20
--outFilterMismatchNmax 10
--alignIntronMax 500000
--alignMatesGapMax 1000000
--sjdbScore 2
--alignSJDBoverhangMin 1
--genomeLoad NoSharedMemory
--limitBAMsortRAM 0
--readFilesCommand <bzcat|cat|zcat>
--outFilterMatchNminOverLread 0.33
--outFilterScoreMinOverLread 0.33
--sjdbOverhang 100
--outSAMstrandField intronMotif
--outSAMattributes NH HI NM MD AS XS
--outSAMunmapped Within
--outSAMtype BAM SortedByCoordinate
--outSAMheaderHD @HD VN:1.4
--outSAMattrRGline <formatted RG line provided by wrapper>
```

mRNA Expression Workflow

Following alignment, BAM files are processed through the RNA Expression Workflow.

First the BAM files are filtered for aligned reads using the samtools view function. The reads mapped to each gene are enumerated using HT-Seq count. Expression values are provided in a tab-delimited format. GENCODE v22 was used for gene annotation.



I/O	Entity	Format
Input	Aligned Reads	BAM
Output	Gene Expression (HTSeq count/ FPKM/ FPKM-UQ)	TXT

mRNA Quantification Command Line Parameters

Samtools v1.1; HTSeq-0.6.1p1

```
samtools view -F 4 <input.bam> |
htseq-count \
-m intersection-nonempty \
-i gene_id \
-r pos \
-s no \
- gencode.v22.annotation.gtf
```

mRNA Expression Normalization

RNA-Seq expression level read counts are normalized using two related methods: FPKM and FPKM-UQ. Normalized values should be used only within the context of the entire gene set. Users are encouraged to normalize raw read count values if a subset of genes is investigated.

FPKM

The Fragments per Kilobase of transcript per Million mapped reads (FPKM) calculation normalizes read count by dividing it by the gene length and the total number of reads mapped to protein-coding genes.

Upper Quartile FPKM

The upper quartile FPKM (FPKM-UQ) is a modified FPKM calculation in which the total protein-coding read count is replaced by the 75th percentile read count value for the sample.

Calculations

$$FPKM = \frac{RC_g * 10^9}{RC_{pc} * L} \qquad FPKM - UQ = \frac{RC_g * 10^9}{RC_{g75} * L}$$

- **RCg:** Number of reads mapped to the gene
- **RCpc:** Number of reads mapped to all protein-coding genes
- **RCg75:** The 75th percentile read count value for genes in the sample
- **L:** Length of the gene in base pairs; Calculated as the sum of all exons in a gene

Note: The read count is multiplied by a scalar (109) during normalization to account for the kilobase and 'million mapped reads' units.

Examples

Sample 1: Gene A

- Gene length: 3,000 bp
- 1,000 reads mapped to Gene A
- 1,000,000 reads mapped to all protein-coding regions
- Read count in Sample 1 for 75th percentile gene: 2,000

FPKM for Gene A = (1,000)*(10ˆ9)/[(3,000)*(1,000,000)] = **333.33**

FPKM-UQ for Gene A = (1,000)*(10ˆ9)/[(3,000)*(2,000)] = **166,666.67**

File Access and Availability

To facilitate the use of harmonized data in user-created pipelines, RNA-Seq gene expression is accessible in the GDC Data Portal at several intermediate steps in the pipeline. Below is a description of each type of file available for download in the GDC Data Portal.

Type	Description	Format
RNA-Seq Alignment	RNA-Seq reads that have been aligned to the GRCh38 build. Reads that were not aligned are included to facilitate the availability of raw read sets	BAM
Raw Read Counts	The number of reads aligned to each gene, calculated by HT-Seq	TXT
FPKM	A normalized expression value that takes into account each gene length and the number of reads mapped to all protein-coding genes	TXT
FPKM-UQ	A modified version of the FPKM formula in which the 75th percentile read count is used as the denominator in place of the total number of protein-coding reads	TXT

Chapter 8

Bioinformatics Pipeline: miRNA Analysis

miRNA Analysis Pipeline

Introduction

The GDC miRNA quantification analysis makes use of a modified version of the profiling pipeline that the British Columbia Genome Sciences Centre developed. The pipeline generates TCGA-formatted miRNAseq data. The first step is read alignment. The tool then compares the individual reads to sequence feature annotations in miRBase v21 and UCSC. Of note, however, the tool only annotates those reads that have an exact match with known miRNAs in miRBase and should therefore not be considered for novel miRNA identification or mismatched alignments.

For more information see BCGSC's GitHub or the original publication.

Data Processing Steps

Alignment Workflow

The miRNA pipeline begins with the Alignment Workflow, which in the case of miRNA uses BWA-aln. This outputs one BAM file for each read group in the input.

I/O	Entity	Format
Inpu t	Submitted Unaligned Reads or Submitted Aligned Reads	FASTQ or BAM
Outp ut	Aligned Reads	BAM

miRNA Expression Workflow

Following alignment, BAM files are processed through the miRNA Expression Workflow.

The outputs of the miRNA profiling pipeline report raw read counts and counts normalized to reads per million mapped reads (RPM) in two separate files mirnas.quantification.txt and isoforms.quantification.txt. The former contains summed expression for all reads aligned to known miRNAs in the miRBase reference. If there are multiple alignments to different miRNAs or different regions of the same miRNA, the read is flagged as cross-mapped and every miRNA annotation is preserved. The latter contains observed isoforms.

I/O	Entity	Format
Input	Aligned Reads	BAM
Output	miRNA Expression	TXT

File Access and Availability

Type	Description	Format
Aligned Reads	miRNA-Seq reads that have been aligned to the GRCh38 build. Reads that were not aligned are included to facilitate the availability of raw read sets.	BAM
miRNA Expression Quantification	A table that associates miRNA IDs with read count and a normalized count in reads-per-million-miRNA-mapped.	TXT
Isoform Expression Quantification	A table with the same information as the miRNA Expression Quantification files with the addition of isoform information such as the coordinates of the isoform and the type of region it constitutes within the full miRNA transcript.	TXT

Chapter 9

Bioinformatics Pipeline: Copy Number Variation

Copy Number Variation Analysis Pipeline

Introduction

The copy number variation (CNV) pipeline uses Affymetrix SNP 6.0 array data to identify genomic regions that are repeated and infer the copy number of these repeats. This pipeline is built onto the existing TCGA level 2 data generated by Birdsuite and uses the DNAcopy R-package to perform a circular binary segmentation (CBS) analysis [1]. CBS translates noisy intensity measurements into chromosomal regions of equal copy number. The final output files are segmented into genomic regions with the estimated copy number for each region. The GDC further transforms these copy number values into segment mean values, which are equal to log2(copy-number/ 2). Diploid regions will have a segment mean of zero, amplified regions will have positive values, and deletions will have negative values.

Data Processing Steps

The GRCh38 probe-set was produced by mapping probe sequences to the GRCh38 reference genome and can be downloaded at the GDC Reference File Website.

Copy Number Segmentation

The Copy Number Liftover Workflow uses the TCGA level 2 tangent.copynumber files described above. These files were generated by first normalizing array intensity values, estimating raw copy number, and performing tangent normalization, which subtracts variation that is found in a set of normal samples. Original array intensity values (TCGA level 1) are available in the GDC Legacy Archive under the "Data Format: CEL" and "Platform: Affymetrix SNP 6.0" filters.

The Copy Number Liftover Workflow performs CBS analysis using the DNACopy R-package to process tangent normalized data into Copy Number Segment files, which associate contiguous chromosome regions with log2 ratio segment means in a tab-delimited format. The number of probes with intensity values associated with each chromosome region is also reported (probes with no intensity values are not included in this count). During copy number segmentation probe sets from Pseudo-Autosomal Regions (PARs) were removed from males and Y chromosome segments were removed from females.

Masked copy number segments are generated using the same method except that a filtering step is performed that removes the Y chromosome and probe sets that were previously indicated to be associated with frequent germline copy-number variation.

I/O	Entity	Forma t
Input	Submitted Tangent Copy Number	TXT

I/O	Entity	Forma t
Output	Copy Number Segment or Masked Copy Number Segment	TXT

Copy Number Estimation

Numeric focal-level Copy Number Alteration (CNA) values were generated with “Masked Copy Number Segment” files from tumor aliquots using GISTIC2 [2], [3] on a project level. Only protein-coding genes were kept, and their numeric CNA values were further thresholded by a noise cutoff of 0.3:

- Genes with focal CNA values smaller than -0.3 are categorized as a “loss” (-1)
- Genes with focal CNA values larger than 0.3 are categorized as a “gain” (+1)
- Genes with focal CNA values between and including -0.3 and 0.3 are categorized as “neutral” (0).

Values are reported in a project-level TSV file. Each row represents a gene, which is reported as an Ensembl ID and associated cytoband. The columns represent aliquots, which are associated with CNA value categorizations (0/1/-1) for each gene.

I/O	Entity	Format
Input	Masked Copy Number Segment	TXT
Output	Copy Number Estimate	TXT

GISTIC2 Command Line Parameters

```
gistic2 \
-b <base_directory> \
-seg <segmentation_file> \
-mk <marker_file> \
-refgene <reference_gene_file> \
-ta 0.1 \
-armpeel 1 \
-brlen 0.7 \
-cap 1.5 \
-conf 0.99 \
-td 0.1 \
-genegistic 1 \
-gcm extreme \
-js 4 \
-maxseg 2000 \
-qvt 0.25 \
-rx 0 \
-savegene 1 \
(-broad 1)
```

File Access and Availability

Type	Description	Format
Copy Number Segment	A table that associates contiguous chromosomal segments with genomic coordinates, mean array intensity, and the number of probes that bind to each segment.	TXT

Type	Description	Format
Masked Copy Number Segment	A table with the same information as the Copy Number Segment except that segments with probes known to contain germline mutations are removed.	TXT

[1] Olshen, Adam B., E. S. Venkatraman, Robert Lucito, and Michael Wigler. “Circular binary segmentation for the analysis of array-based DNA copy number data.” Biostatistics 5, no. 4 (2004): 557-572.

[2] Mermel, Craig H., Steven E. Schumacher, Barbara Hill, Matthew L. Meyerson, Rameen Beroukhim, and Gad Getz. “GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.” Genome biology 12, no. 4 (2011): R41.

[3] Beroukhim, Rameen, Craig H. Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina et al. “The landscape of somatic copy-number alteration across human cancers.” Nature 463, no. 7283 (2010): 899.

Chapter 10

Bioinformatics Pipeline: Methylation Liftover

Methylation Liftover Pipeline

Introduction

The DNA Methylation Liftover Pipeline uses data from the Illumina Infinium Human Methylation 27 (HM27) and HumanMethylation450 (HM450) arrays to measure the level of methylation at known CpG sites as beta values, calculated from array intensities (Level 2 data) as Beta = M/(M+U).

Using probe sequence information provided in the manufacturer's manifest, HM27 and HM450 probes were remapped to the GRCh38 reference genome [1]. Type II probes with a mapping quality of <10, or Type I probes for which the methylated and unmethylated probes map to different locations in the genome, and/or had a mapping quality of <10, had an entry of '*' for the 'chr' field, and '-1' for coordinates. These coordinates were then used to identify the associated transcripts from GENCODE v22, the associated CpG island (CGI), and the CpG sites' distance from each of these features. Multiple transcripts overlapping the target CpG were separated with semicolons. Beta values were inherited from existing TCGA Level 3 DNA methylation data (hg19-based) based on Probe IDs.

Methylation Beta Values Table Format

Descriptions for fields present in GDC Harmonized Methylation Beta Values Table are detailed below:

Field	Definition
Composite Element	A unique ID for the array probe associated with a CpG site
Beta Value	Represents the ratio between the methylated array intensity and total array intensity, falls between 0 (lower levels of methylation) and 1 (higher levels of methylation)
Chromosome	The chromosome in which the probe binding site is located
Start	The start of the CpG site on the chromosome
End	The end of the CpG site on the chromosome
Gene Symbol	The symbol for genes associated with the CpG site. Genes that fall within 1,500 bp upstream of the transcription start site (TSS) to the end of the gene body are used.
Gene Type	A general classification for each gene (e.g. protein coding, miRNA, pseudogene)
Transcript ID	Ensembl transcript IDs for each transcript associated with the genes detailed above

Field	Definition
Position to TSS	Distance in base pairs from the CpG site to each associated transcript's start site
CGI Coordinate	The start and end coordinates of the CpG island associated with the CpG site
Feature Type	The position of the CpG site in reference to the island: Island, N_Shore or S_Shore (0-2 kb upstream or downstream from CGI), or N_Shelf or S_Shelf (2-4 kbp upstream or downstream from CGI)

I/O	Entity	Format
Input	Submitted Methylation Beta Values	TXT
Output	Methylation Beta Values or Masked Copy Number Segment	TXT

File Access and Availability

Type	Description	Format
Methylation Beta Value	A table that associates array probes with CpG sites and associated metadata.	TXT

[1]. Zhou, Wanding, Laird Peter L., and Hui Shen. "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes." Nucleic Acids Research. (2016): doi: 10.1093/nar/gkw967

Chapter 11

Release Notes

Data Release Notes

Version	Date
v13.0	September 27, 2018
v12.0	June 13, 2018
v11.0	May 21, 2018
v10.1	February 15, 2018
v10.0	December 21, 2017
v9.0	October 24, 2017
v8.0	August 22, 2017
v7.0	June 29, 2017
v6.0	May 9, 2017
v5.0	March 16, 2017
v4.0	October 31, 2016
v3.0	September 16, 2016
v2.0	August 9, 2016
v1.0	June 6, 2016

Data Release 13.0

- **GDC Product**: Data
- **Release Date**: September 27, 2018

New updates

1. Three new projects are released to the GDC (VAREPOP-APOLLO (phs001374), CTSP-DLBCL1 (phs001184), NCICCR-DLBCL (phs001444)
2. TARGET WGS alignments are released. VCFs will be provided in a later release
3. Clinical data was harmonized with ICD-O-3 terminology for TCGA properties case.primary_site, case.disease_type, diagnosis.primary_diagnosis, diagnosis.site_of_resection_or_biopsy, diagnosis.tissue_or_organ_of_origin
4. Redaction annotations applied to 11 aliquots in TCGA-DLBC

5. Redaction annotations applied to incorrectly trimmed miRNA file in the Legacy Achive

A complete list of files for DR13.0 are listed for the GDC Data Portal and the GDC Legacy Archive are found below:

- gdc_manifest_20180927_data_release_13.0_active.txt.gz
- gdc_manifest_20180927_data_release_13.0_legacy.txt.gz.

Bugs Fixed Since Last Release

- 253 files Copy Number Segment and Masked Copy Number Segment files were released. These were skipped in DR 12.0
- 36 Diagnostic TCGA slides were released. They were skipped in DR 12.0

Known Issues and Workarounds

- 506 Copy Number Segment and 36 Slide Image files are designated as controlled-access on the GDC Data Portal. These files are actually open-access and will be downloadable without a token using this manifest.
- 2 Copy Number Segment files from TCGA-TGCT do not appear on the GDC Portal. They can be downloaded using the Data Transfer Tool using the following UUIDs.
  - 6cd4ef5e-324a-4ace-8779-7a33bd559c83 - RAMPS_p_TCGA_Batch_430_NSP_GenomeWideSNP_6_E07_1538238.nocnv_
  - dfa89ee9-6ee5-460b-bd58-b5ca0e9cb7ac - RAMPS_p_TCGA_Batch_430_NSP_GenomeWideSNP_6_E07_1538238.grch38.s
- TARGET CGI BAMs in the Legacy Archive for the following aliquots should not be used because they were not repaired and concatenated into their original composite BAM files by CGHub.
  - TARGET-20-PASJGZ-04A-02D
  - TARGET-30-PAPTLY-01A-01D
  - TARGET-20-PAEIKD-09A-01D
  - TARGET-20-PASMYS-14A-02D
  - TARGET-20-PAMYAS-14A-02D
  - TARGET-10-PAPZST-09A-01D
- Some miRNA files with QC failed reads were not swapped in DR11.0. 361 aliquots remain to be swapped in a later release
- 74 Diagnostic TCGA slides are attached to a portion rather than a sample like the rest of the diagnostic slides. The reflects how these original samples were handled.
- 11 bam files for TARGET-NBL RNA-Seq are not available in the GDC Data portal
- Two tissue slide images are unavailable for download from GDC Data Portal
- The raw and annotated VarScan VCF files for aliquot `TCGA-VR-A8ET-01A-11D-A403-09` are not available. These VCFs files will be replaced in a later release.
- There are 5051 TARGET files for which `experimental_strategy`, `data_format`, `platform`, and `data_subtype` are blank
- There are two cases with identical submitter_id `TARGET-10-PARUYU`
- TARGET-MDLS cases do not have disease_type or primary_site populated
- Some TARGET cases are missing `days_to_last_follow_up`
- Some TARGET cases are missing `age_at_diagnosis`
- Some TARGET files are not connected to all related aliquots
- Samples of TARGET sample_type `Recurrent Blood Derived Cancer - Bone Marrow` are mislabeled as `Recurrent Blood Derived Cancer - Peripheral Blood`. A workaround is to look at the sample barcode, which is -04 for `Recurrent Blood Derived Cancer - Bone Marrow`. (e.g. `TARGET-20-PAMYAS-04A-03R`)
- FM-AD clinical and biospecimen supplement files have incorrect data format. They are listed as XLSX, but are in fact TSV files.
- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- Most intronic mutations are removed for MAF generation. However, validated variants may rescue these in some cases. Therefore intronic mutations in MAF files are not representative of those called by mutation callers.
- The latest TARGET data is not yet available at the GDC. For the complete and latest data, please see the TARGET Data Matrix. Data that is not present or is not the most up to date includes:

  - All microarray data and metadata
  - All sequencing analyzed data and metadata
  - 1180 of 12063 sequencing runs of raw data
- Demographic information for some TARGET patients is incorrect. The correct information can be found in the associated clinical supplement file. Impacted patients are TARGET-50-PAJNUS.
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MDLS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Data Release 12.0

- **GDC Product**: Data
- **Release Date**: June 13, 2018

New updates

1. Updated clinical and biospecimen XML files for TCGA cases are available in the GDC Data Portal. Equivalent Legacy Archive files may no longer be up to date.
2. All biospecimen and clinical supplement files for TCGA projects formerly only found in the Legacy Archive have been updated and transferred to the GDC Data Portal. Equivalent Legacy Archive files and metadata retrieved from the API may no longer be up to date.
3. Diagnostic slides from TCGA are now available in the GDC Data Portal and Slide Image Viewer. They were formerly only available in the Legacy Archive.
4. Updated Copy Number Segment and Masked Copy Number Segment files are now available. These were generated using an improved mapping of hg38 coordinates for the Affymetrix SNP6.0 probe set.
5. VCF files containing SNVs produced from TARGET WGS CGI data are available. The variant calls were initially produced by CGI and lifted over to hg38.

Updated files for this release are listed here. A complete list of files for DR12.0 are listed for the GDC Data Portal here and the GDC Legacy Archive here.

Bugs Fixed Since Last Release

- TARGET NBL RNA-Seq data is now associated with the correct aliquot.

Known Issues and Workarounds

- Some Copy Number Segment and Masked Copy Number Segment were not replaced in DR 12.0. 253 files remain to be swapped in a later release

- Some miRNA files with QC failed reads were not swapped in DR11.0. 361 aliquots remain to be swapped in a later release
- 74 Diagnostic TCGA slides are attached to a portion rather than a sample like the rest of the diagnostic slides. The reflects how these original samples were handled.
- 36 Diagnostic TCGA slides are not yet available in the active GDC Portal. They are still available in the GDC Legacy Archive.
- 11 bam files for TARGET-NBL RNA-Seq are not available in the GDC Data portal
- Two tissue slide images are unavailable for download from GDC Data Portal
- The raw and annotated VarScan VCF files for aliquot `TCGA-VR-A8ET-01A-11D-A403-09` are not available. These VCFs files will be replaced in a later release.
- There are 5051 TARGET files for which `experimental_strategy`, `data_format`, `platform`, and `data_subtype` are blank
- There are two cases with identical submitter_id `TARGET-10-PARUYU`
- TARGET-MDLS cases do not have disease_type or primary_site populated
- Some TARGET cases are missing `days_to_last_follow_up`
- Some TARGET cases are missing `age_at_diagnosis`
- Some TARGET files are not connected to all related aliquots
- Samples of TARGET sample_type `Recurrent Blood Derived Cancer - Bone Marrow` are mislabeled as `Recurrent Blood Derived Cancer - Peripheral Blood`. A workaround is to look at the sample barcode, which is -04 for `Recurrent Blood Derived Cancer - Bone Marrow`. (e.g. `TARGET-20-PAMYAS-04A-03R`)
- FM-AD clinical and biospecimen supplement files have incorrect data format. They are listed as XLSX, but are in fact TSV files.
- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- Most intronic mutations are removed for MAF generation. However, validated variants may rescue these in some cases. Therefore intronic mutations in MAF files are not representative of those called by mutation callers.
- The latest TARGET data is not yet available at the GDC. For the complete and latest data, please see the TARGET Data Matrix. Data that is not present or is not the most up to date includes:
  - All microarray data and metadata
  - All sequencing analyzed data and metadata
  - 1180 of 12063 sequencing runs of raw data
- Demographic information for some TARGET patients is incorrect. The correct information can be found in the associated clinical supplement file. Impacted patients are TARGET-50-PAJNUS.
- There are 11 cases in project TCGA-DLBC that are known to have incorrect WXS data in the GDC Data Portal. Impacted cases are listed below. This affects the BAMs and VCFs associated with these cases in the GDC Data Portal. Corrected BAMs can be found in the GDC Legacy Archive. Variants from affected aliquots appear in the protected MAFs with GDC_FILTER=ContEst to indicate a sample contamination problem, but are removed during the generation of the Somatic MAF file. In a later release we will supply corrected BAM, VCF, and MAF files for these cases. In the mean time, we advise you not to use any of the WXS files associated with these cases in the GDC Data Portal. A list of these files can be found here. Download list of affected files.
  - TCGA-FF-8062
  - TCGA-FM-8000
  - TCGA-G8-6324
  - TCGA-G8-6325
  - TCGA-G8-6326
  - TCGA-G8-6906
  - TCGA-G8-6907
  - TCGA-G8-6909
  - TCGA-G8-6914
  - TCGA-GR-7351
  - TCGA-GR-7353
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our

BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MDLS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Data Release 11.0

- **GDC Product**: Data
- **Release Date**: May 21, 2018

New updates

1. Updated miRNA files to remove QCFail reads. This included all BAM and downstream count files.
2. TCGA Tissue slide images now available in GDC Data Portal. Previously these were found only in the Legacy Archive

Updated files for this release are listed here. A complete list of files for DR11.0 are listed for the GDC Data Portal here and the GDC Legacy Archive here.

Bugs Fixed Since Last Release

- N/A

Known Issues and Workarounds

- Two tissue slide images are unavailable for download from GDC Data Portal
- RNA-Seq files for TARGET-NBL are attached to the incorrect aliquot. The BAM files contain the correct information in their header but the connection in the GDC to read groups and aliquots is incorrect. The linked file below contains a mapping between aliquots where file are currently associated and the aliquot where they should instead be associated (mapping file).
- The raw and annotated VarScan VCF files for aliquot `TCGA-VR-A8ET-01A-11D-A403-09` were not replaced in DR10.0 and thus do not contain indels. However, the indels from this aliquot can be found in the MAF files and are displayed in the Exploration section in the Data Portal. These VCFs files will be replaced in a later release.
- There are 5051 TARGET files for which `experimental_strategy`, `data_format`, `platform`, and `data_subtype` are blank
- There are two cases with identical submitter_id `TARGET-10-PARUYU`
- TARGET-MDLS cases do not have disease_type or primary_site populated
- Some TARGET cases are missing `days_to_last_follow_up`
- Some TARGET cases are missing `age_at_diagnosis`
- Some TARGET files are not connected to all related aliquots
- miRNA alignments include QC failed reads.
- Samples of TARGET sample_type `Recurrent Blood Derived Cancer - Bone Marrow` are mislabeled as `Recurrent Blood Derived Cancer - Peripheral Blood`. A workaround is to look at the sample barcode, which is -04 for `Recurrent Blood Derived Cancer - Bone Marrow`. (e.g. `TARGET-20-PAMYAS-04A-03R`)
- FM-AD clinical and biospecimen supplement files have incorrect data format. They are listed as XLSX, but are in fact TSV files.

- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- Most intronic mutations are removed for MAF generation. However, validated variants may rescue these in some cases. Therefore intronic mutations in MAF files are not representative of those called by mutation callers.
- The latest TARGET data is not yet available at the GDC. For the complete and latest data, please see the TARGET Data Matrix. Data that is not present or is not the most up to date includes:
  - All microarray data and metadata
  - All sequencing analyzed data and metadata
  - 1180 of 12063 sequencing runs of raw data
- Demographic information for some TARGET patients is incorrect. The correct information can be found in the associated clinical supplement file. Impacted patients are TARGET-50-PAJNUS.
- There are 11 cases in project TCGA-DLBC that are known to have incorrect WXS data in the GDC Data Portal. Impacted cases are listed below. This affects the BAMs and VCFs associated with these cases in the GDC Data Portal. Corrected BAMs can be found in the GDC Legacy Archive. Variants from affected aliquots appear in the protected MAFs with GDC_FILTER=ContEst to indicate a sample contamination problem, but are removed during the generation of the Somatic MAF file. In a later release we will supply corrected BAM, VCF, and MAF files for these cases. In the mean time, we advise you not to use any of the WXS files associated with these cases in the GDC Data Portal. A list of these files can be found here. Download list of affected files.
  - TCGA-FF-8062
  - TCGA-FM-8000
  - TCGA-G8-6324
  - TCGA-G8-6325
  - TCGA-G8-6326
  - TCGA-G8-6906
  - TCGA-G8-6907
  - TCGA-G8-6909
  - TCGA-G8-6914
  - TCGA-GR-7351
  - TCGA-GR-7353
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MDLS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Data Release 10.1

- **GDC Product**: Data
- **Release Date**: February 15, 2018

New updates

1. Updated FM-AD clinical data to conform with Data Dictionary release v1.11

Bugs Fixed Since Last Release

None

Known Issues and Workarounds

- RNA-Seq files for TARGET-NBL are attached to the incorrect aliquot. The BAM files contain the correct information in their header but the connection in the GDC to read groups and aliquots is incorrect. The linked file below contains a mapping between aliquots where file are currently associated and the aliquot where they should instead be associated (mapping file).
- The raw and annotated VarScan VCF files for aliquot `TCGA-VR-A8ET-01A-11D-A403-09` were not replaced in DR10.0 and thus do not contain indels. However, the indels from this aliquot can be found in the MAF files and are displayed in the Exploration section in the Data Portal. These VCFs files will be replaced in a later release.
- There are 5051 TARGET files for which `experimental_strategy`, `data_format`, `platform`, and `data_subtype` are blank
- There are two cases with identical submitter_id `TARGET-10-PARUYU`
- TARGET-MDLS cases do not have disease_type or primary_site populated
- Some TARGET cases are missing `days_to_last_follow_up`
- Some TARGET cases are missing `age_at_diagnosis`
- Some TARGET files are not connected to all related aliquots
- miRNA alignments include QC failed reads.
- Samples of TARGET sample_type `Recurrent Blood Derived Cancer - Bone Marrow` are mislabeled as `Recurrent Blood Derived Cancer - Peripheral Blood`. A workaround is to look at the sample barcode, which is -04 for `Recurrent Blood Derived Cancer - Bone Marrow`. (e.g. `TARGET-20-PAMYAS-04A-03R`)
- FM-AD clinical and biospecimen supplement files have incorrect data format. They are listed as XLSX, but are in fact TSV files.
- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- Most intronic mutations are removed for MAF generation. However, validated variants may rescue these in some cases. Therefore intronic mutations in MAF files are not representative of those called by mutation callers.
- The latest TARGET data is not yet available at the GDC. For the complete and latest data, please see the TARGET Data Matrix. Data that is not present or is not the most up to date includes:
  - All microarray data and metadata
  - All sequencing analyzed data and metadata
  - 1180 of 12063 sequencing runs of raw data
- Demographic information for some TARGET patients is incorrect. The correct information can be found in the associated clinical supplement file. Impacted patients are TARGET-50-PAJNUS.
- There are 11 cases in project TCGA-DLBC that are known to have incorrect WXS data in the GDC Data Portal. Impacted cases are listed below. This affects the BAMs and VCFs associated with these cases in the GDC Data Portal. Corrected BAMs can be found in the GDC Legacy Archive. Variants from affected aliquots appear in the protected MAFs with GDC_FILTER=ContEst to indicate a sample contamination problem, but are removed during the generation of the Somatic MAF file. In a later release we will supply corrected BAM, VCF, and MAF files for these cases. In the mean time, we advise you not to use any of the WXS files associated with these cases in the GDC Data Portal. A list of these files can be found here. Download list of affected files.
  - TCGA-FF-8062
  - TCGA-FM-8000
  - TCGA-G8-6324
  - TCGA-G8-6325
  - TCGA-G8-6326
  - TCGA-G8-6906

  - TCGA-G8-6907
  - TCGA-G8-6909
  - TCGA-G8-6914
  - TCGA-GR-7351
  - TCGA-GR-7353
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MDLS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Data Release 10.0

- **GDC Product**: Data
- **Release Date**: December 21, 2017

New updates

1. New TARGET files for all projects
2. TARGET updates for clinical and biospecimen data
3. Replace corrupted .bai files
4. Update TCGA and TARGET MAF files to include VarScan2 indels and more information in all_effects column
5. Update VarScan VCF files

Updated files for this release are listed here. A complete list of files for DR10.0 are listed for the GDC Data Portal here and the GDC Legacy Archive here.

Bugs Fixed Since Last Release

None

Known Issues and Workarounds

- The raw and annotated VarScan VCF files for aliquot `TCGA-VR-A8ET-01A-11D-A403-09` were not replaced in DR10.0 and thus do not contain indels. However, the indels from this aliquot can be found in the MAF files and are displayed in the Exploration section in the Data Portal. These VCFs files will be replaced in a later release.
- There are 5051 TARGET files for which `experimental_strategy`, `data_format`, `platform`, and `data_subtype` are blank
- There are two cases with identical submitter_id `TARGET-10-PARUYU`
- TARGET-MDLS cases do not have disease_type or primary_site populated
- Some TARGET cases are missing `days_to_last_follow_up`

- Some TARGET cases are missing `age_at_diagnosis`
- Some TARGET files are not connected to all related aliquots
- miRNA alignments include QC failed reads.
- Samples of TARGET sample_type `Recurrent Blood Derived Cancer - Bone Marrow` are mislabeled as `Recurrent Blood Derived Cancer - Peripheral Blood`. A workaround is to look at the sample barcode, which is -04 for `Recurrent Blood Derived Cancer - Bone Marrow`. (e.g. `TARGET-20-PAMYAS-04A-03R`)
- FM-AD clinical and biospecimen supplement files have incorrect data format. They are listed as XLSX, but are in fact TSV files.
- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- Most intronic mutations are removed for MAF generation. However, validated variants may rescue these in some cases. Therefore intronic mutations in MAF files are not representative of those called by mutation callers.
- The latest TARGET data is not yet available at the GDC. For the complete and latest data, please see the TARGET Data Matrix. Data that is not present or is not the most up to date includes:
  - All microarray data and metadata
  - All sequencing analyzed data and metadata
  - 1180 of 12063 sequencing runs of raw data
- Demographic information for some TARGET patients is incorrect. The correct information can be found in the associated clinical supplement file. Impacted patients are TARGET-50-PAJNUS.
- There are 11 cases in project TCGA-DLBC that are known to have incorrect WXS data in the GDC Data Portal. Impacted cases are listed below. This affects the BAMs and VCFs associated with these cases in the GDC Data Portal. Corrected BAMs can be found in the GDC Legacy Archive. Variants from affected aliquots appear in the protected MAFs with GDC_FILTER=ContEst to indicate a sample contamination problem, but are removed during the generation of the Somatic MAF file. In a later release we will supply corrected BAM, VCF, and MAF files for these cases. In the mean time, we advise you not to use any of the WXS files associated with these cases in the GDC Data Portal. A list of these files can be found here. Download list of affected files.
  - TCGA-FF-8062
  - TCGA-FM-8000
  - TCGA-G8-6324
  - TCGA-G8-6325
  - TCGA-G8-6326
  - TCGA-G8-6906
  - TCGA-G8-6907
  - TCGA-G8-6909
  - TCGA-G8-6914
  - TCGA-GR-7351
  - TCGA-GR-7353
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MDLS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Data Release 9.0

- **GDC Product**: Data
- **Release Date**: October 24, 2017

New updates

1. Foundation Medicine Data Release

- This includes controlled-access VCF and MAF files as well as clinical and biospecimen supplements and metadata.
- Original Foundation Medicine supplied data can be found on the Foundation Medicine Project Page.

2. Updated RNA-Seq data for TARGET NBL

- Includes new BAM and count files

Updated files for this release are listed here. A complete list of files for DR9.0 are listed here.

Bugs Fixed Since Last Release

None

Known Issues and Workarounds

- miRNA alignments include QC failed reads.
- Samples of TARGET sample_type `Recurrent Blood Derived Cancer - Bone Marrow` are mislabeled as `Recurrent Blood Derived Cancer - Peripheral Blood`. A workaround is to look at the sample barcode, which is -04 for `Recurrent Blood Derived Cancer - Bone Marrow`. (e.g. `TARGET-20-PAMYAS-04A-03R`)
- FM-AD clinical and biospecimen supplement files have incorrect data format. They are listed as XLSX, but are in fact TSV files.
- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- Most intronic mutations are removed for MAF generation. However, validated variants may rescue these in some cases. Therefore intronic mutations in MAF files are not representative of those called by mutation callers.
- The latest TARGET data is not yet available at the GDC. For the complete and latest data, please see the TARGET Data Matrix. Data that is not present or is not the most up to date includes:
  - All microarray data and metadata
  - All sequencing analyzed data and metadata
  - 1180 of 12063 sequencing runs of raw data
- Demographic information for some TARGET patients is incorrect. The correct information can be found in the associated clinical supplement file. Impacted patients are TARGET-50-PAJNUS.
- There are 11 cases in project TCGA-DLBC that are known to have incorrect WXS data in the GDC Data Portal. Impacted cases are listed below. This affects the BAMs and VCFs associated with these cases in the GDC Data Portal. Corrected BAMs can be found in the GDC Legacy Archive. Variants from affected aliquots appear in the protected MAFs with GDC_FILTER=ContEst to indicate a sample contamination problem, but are removed during the generation of the Somatic MAF file. In a later release we will supply corrected BAM, VCF, and MAF files for these cases. In the mean time, we advise you not to use any of the WXS files associated with these cases in the GDC Data Portal. A list of these files can be found here. Download list of affected files.
  - TCGA-FF-8062
  - TCGA-FM-8000
  - TCGA-G8-6324

  - TCGA-G8-6325
  - TCGA-G8-6326
  - TCGA-G8-6906
  - TCGA-G8-6907
  - TCGA-G8-6909
  - TCGA-G8-6914
  - TCGA-GR-7351
  - TCGA-GR-7353
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MDLS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Data Release 8.0

- **GDC Product**: Data
- **Release Date**: August 22, 2017

New updates

1. Released updated miRNA quantification files to address double counting of some normalized counts described in DR7.0 release notes.

Updated files for this release are listed here. A Complete list of files for DR8.0 are listed here.

Bugs Fixed Since Last Release

None

Known Issues and Workarounds

- TARGET-NBL RNA-Seq files were run as single ended even though they are derived from paired-end data. These files will be rerun through the GDC RNA-Seq pipelines in a later release. Impacted files can be found here. Downstream count files are also affected. Users may access original FASTQ files in the GDC Legacy Archive, which are not impacted by this issue.
- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- Most intronic mutations are removed for MAF generation. However, validated variants may rescue these in some cases. Therefore intronic mutations in MAF files are not representative of those called by mutation callers.

- The latest TARGET data is not yet available at the GDC. For the complete and latest data, please see the TARGET Data Matrix. Data that is not present or is not the most up to date includes:
  - All microarray data and metadata
  - All sequencing analyzed data and metadata
  - 1180 of 12063 sequencing runs of raw data
- Demographic information for some TARGET patients is incorrect. The correct information can be found in the associated clinical supplement file. Impacted patients are TARGET-50-PAJNUS.
- There are 11 cases in project TCGA-DLBC that are known to have incorrect WXS data in the GDC Data Portal. Impacted cases are listed below. This affects the BAMs and VCFs associated with these cases in the GDC Data Portal. Corrected BAMs can be found in the GDC Legacy Archive. Variants from affected aliquots appear in the protected MAFs with GDC_FILTER=ContEst to indicate a sample contamination problem, but are removed during the generation of the Somatic MAF file. In a later release we will supply corrected BAM, VCF, and MAF files for these cases. In the mean time, we advise you not to use any of the WXS files associated with these cases in the GDC Data Portal. A list of these files can be found here. Download list of affected files.
  - TCGA-FF-8062
  - TCGA-FM-8000
  - TCGA-G8-6324
  - TCGA-G8-6325
  - TCGA-G8-6326
  - TCGA-G8-6906
  - TCGA-G8-6907
  - TCGA-G8-6909
  - TCGA-G8-6914
  - TCGA-GR-7351
  - TCGA-GR-7353
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MDLS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Data Release 7.0

- **GDC Product**: Data
- **Release Date**: June 29, 2017

New updates

1. Updated public Mutation Annotation Format (MAF) files are now available. Updates include filtering to remove variants impacted by OxoG artifacts and those impacted by strand bias.

2. Protected MAF files are updated to include flags for OxoG and strand bias.
3. Annotated VCFs are updated to include flags for OxoG artifacts and strand bias.

Updated files for this release are listed here. A Complete list of files for DR7.0 are listed here

Bugs Fixed Since Last Release

None

Known Issues and Workarounds

- TARGET-NBL RNA-Seq files were run as single ended even though they are derived from paired-end data. These files will be rerun through the GDC RNA-Seq pipelines in a later release. Impacted files can be found here. Downstream count files are also affected. Users may access original FASTQ files in the GDC Legacy Archive, which are not impacted by this issue.
- Reads that are mapped to multiple genomic locations are double counted in some of the GDC miRNA results. The GDC will release updated files correcting the issue in an upcoming release. The specific impacts are described further below:
  - Isoform Expression Quantification files
    - Raw reads counts are accurate
    - Normalized counts are proportionally skewed (r^2=1.0)
  - miRNA Expression Quantification files
    - A small proportion of miRNA counts are overestimated (mean r^2=0.9999)
    - Normalized counts are proportionally skewed (mean r^2=0.9999)
  - miRNA BAM files
    - no impact
- Mutation frequency may be underestimated when using MAF files for genes that overlap other genes. This is because MAF files only record one gene per variant.
- Most intronic mutations are removed for MAF generation. However, validated variants may rescue these in some cases. Therefore intronic mutations in MAF files are not representative of those called by mutation callers.
- The latest TARGET data is not yet available at the GDC. For the complete and latest data, please see the TARGET Data Matrix. Data that is not present or is not the most up to date includes:
  - All microarray data and metadata
  - All sequencing analyzed data and metadata
  - 1180 of 12063 sequencing runs of raw data
- Demographic information for some TARGET patients is incorrect. The correct information can be found in the associated clinical supplement file. Impacted patients are TARGET-50-PAJNUS.
- There are 11 cases in project TCGA-DLBC that are known to have incorrect WXS data in the GDC Data Portal. Impacted cases are listed below. This affects the BAMs and VCFs associated with these cases in the GDC Data Portal. Corrected BAMs can be found in the GDC Legacy Archive. Variants from affected aliquots appear in the protected MAFs with GDC_FILTER=ContEst to indicate a sample contamination problem, but are removed during the generation of the Somatic MAF file. In a later release we will supply corrected BAM, VCF, and MAF files for these cases. In the mean time, we advise you not to use any of the WXS files associated with these cases in the GDC Data Portal. A list of these files can be found here. Download list of affected files.
  - TCGA-FF-8062
  - TCGA-FM-8000
  - TCGA-G8-6324
  - TCGA-G8-6325
  - TCGA-G8-6326
  - TCGA-G8-6906
  - TCGA-G8-6907
  - TCGA-G8-6909
  - TCGA-G8-6914

  - TCGA-GR-7351
  - TCGA-GR-7353
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MLDS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Data Release 6.0

- **GDC Product**: Data
- **Release Date**: May 9, 2017

New updates

1. GDC updated public Mutation Annotation Format (MAF) files are now available. Updates include leveraging the MC3 variant filtering strategy, which results in more variants being recovered relative to the previous version. A detailed description of the new format can be found here.
2. Protected MAFs are updated to include additional variant annotation information
3. Some MuTect2 VCFs updated to include dbSNP and COSMIC annotations found in other VCFs

Updated files for this release are listed here.

Bugs Fixed Since Last Release

None

Known Issues and Workarounds

- There are 11 cases in project TCGA-DLBC that are known to have incorrect WXS data in the GDC Data Portal. Impacted cases are listed below. This affects the BAMs and VCFs associated with these cases in the GDC Data Portal. Corrected BAMs can be found in the GDC Legacy Archive. Variants from affected aliquots appear in the protected MAFs with GDC_FILTER=ContEst to indicate a sample contamination problem, but are removed during the generation of the Somatic MAF file. In a later release we will supply corrected BAM, VCF, and MAF files for these cases. In the mean time, we advise you not to use any of the WXS files associated with these cases in the GDC Data Portal. A list of these files can be found here. Download list of affected files.
  - TCGA-FF-8062
  - TCGA-FM-8000
  - TCGA-G8-6324
  - TCGA-G8-6325

  - TCGA-G8-6326
  - TCGA-G8-6906
  - TCGA-G8-6907
  - TCGA-G8-6909
  - TCGA-G8-6914
  - TCGA-GR-7351
  - TCGA-GR-7353
- Variants found in VCF and MAF files may contain OxoG artifacts, which are produced during library preparation and may result in the apparent substitutions of C to A or G to T in certain sequence contexts. In the future we will plan to label potential oxoG artifacts in the MAF files.
- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Some validated somatic mutations may not be present in open-access MAF files. Please review the protected MAF files in the GDC Data Portal if you are unable to find your mutation in the open-access files.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- No data from TARGET-MLDS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Details are provided in Data Release Manifest

Data Release 5.0

- **GDC Product**: Data
- **Release Date**: March 16, 2017

New updates

1. Additional annotations from TCGA DCC are available
    - Complete list of updated TCGA files is found here
2. Clinical data added for TARGET ALL P1 and P2
3. Pathology reports now have submitter IDs as assigned by the BCR
4. TARGET Data refresh
    - Most recent biospecimen and clinical information from the TARGET DCC. New imported files are listed here
    - Updated indexed biospecimen and clinical metadata
    - Updated SRA XMLs files
    - Does not include updates to TARGET NBL

Bugs Fixed Since Last Release

1. Missing cases from TCGA-LAML were added to Legacy Archive
2. Biotab files are now linked to Projects and Cases in Legacy Archive

Known Issues and Workarounds

- Some TCGA annotations are unavailable in the Legacy Archive or Data Portal. These annotations can be found here.
- Some validated somatic mutations may not be present in open-access MAF files. When creating open-access MAF files from the protected versions we are extremely conservative in removing potential germline variants. Our approach is to remove all mutations that are present in dbSNP. In a subsequent release we will provide updated open-access MAF files, which preserve variants found in MC3 or a TCGA validation study. Please review the protected MAF files in the GDC Data Portal if you are unable to find your mutation in the open-access files.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- MAF Column #109 "FILTER" entries are separated by both commas and semi-colons.
- TARGET-AML is undergoing reorganization. Pending reorganization, cases from this projects may not contain many clinical, biospecimen, or genomic data files.
- No data from TARGET-MLDS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- Two biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg
- Tumor grade property is not populated
- Progression_or_recurrence property is not populated

Details are provided in Data Release Manifest

Data Release 4.0

- **GDC Product**: Data
- **Release Date**: October 31, 2016

New updates

1. TARGET ALL P1 and P2 biospecimen and molecular data are now available in the Legacy Archive. Clinical data will be available in a later release.
2. Methylation data from 27k/450k Arrays has been lifted over to hg38 and is now available in the GDC Data Portal
3. Public MAF files are now available for VarScan2, MuSE, and SomaticSniper. MuTect2 MAFs were made available in a previous release.
4. Updated VCFs and MAF files are available for MuTect2 pipeline to compensate for WGA-related false positive indels. See additional information on that change here. A listing of replaced files is provided here.
5. Added submitter_id for Pathology Reports in Legacy Archive

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- Some validated somatic mutations may not be present in open-access MAF files. When creating open-access MAF files from the protected versions we are extremely conservative in removing potential germline variants. Our approach is to

remove all mutations that are present in dbSNP. In a subsequent release we will provide updated open-access MAF files, which preserve variants found in COSMIC or a TCGA validation study. Please review the protected MAF files in the GDC Data Portal if you are unable to find your mutation in the open-access files.
- Public MAF files for different variant calling pipelines but the same project may contain different numbers of samples. Samples are omitted from the public MAF files if they have no PASS variants, which can lead to this apparent discrepancy.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- MAF Column #109 "FILTER" entries are separated by both commas and semi-colons.
- TARGET-AML is undergoing reorganization. Pending reorganization, cases from this projects may not contain many clinical, biospecimen, or genomic data files.
- No data from TARGET-MLDS is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- There are 200 cases from TCGA-LAML that do not appear in the Legacy Archive
- Biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg

Details are provided in Data Release Manifest

Data Release 3.0

- **GDC Product**: Data
- **Release Date**: September 16, 2016

New updates

1. CCLE data now available (in the Legacy Archive only)
2. BMI calculation is corrected
3. Slide is now categorized as a Biospecimen entity

Bugs Fixed Since Last Release

- BMI calculation is corrected

Known Issues and Workarounds

- Insertions called for tumor samples that underwent whole genome amplification may be of lower quality. Whether a sample underwent this process can be found in the analyte_type property within analyte and aliquot. TCGA analyte type can be also identified in the 20th character of TCGA barcode, at which "W" corresponds to WGA.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- Public MAFs (those with germline variants removed) are only available for MuTect2 pipeline. MAFs for other pipelines are forthcoming.
- MAF Column #109 "FILTER" entries are separated by both commas and semi-colons.
- TARGET-AML and TARGET-ALL projects are undergoing reorganization. Pending reorganization, cases from these projects may not contain many clinical, biospecimen, or genomic data files.

- No data from TARGET-PPTP is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- There are 200 cases from TCGA-LAML that do not appear in the Legacy Archive
- Biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg

Details are provided in Data Release Manifest

Data Release 2.0

- **GDC Product**: Data
- **Release Date**: August 9, 2016

New updates

1. Additional data, previously available via CGHub and the TCGA DCC, is now available in the GDC
2. Better linking between files and their associated projects and cases in the Legacy Archive
3. MAF files are now available in the GDC Data Portal

Known Issues and Workarounds

- Insertions called for tumor samples that underwent whole genome amplification may be of lower quality. These are present in VCF and MAF files produced by the MuTect2 variant calling pipeline. This information can be found in the analyte_type property within analyte and aliquot. TCGA analyte type can be also identified in the 20th character of TCGA barcode, at which "W" corresponds to WGA.
- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- Public MAFs (those with germline variants removed) are only available for MuTect2 pipeline. MAFs for other pipelines are forthcoming.
- MAF Column #109 "FILTER" entries are separated by both commas and semi-colons.
- TARGET-AML and TARGET-ALL projects are undergoing reorganization. Pending reorganization, cases from these projects may not contain many clinical, biospecimen, or genomic data files.
- No data from TARGET-PPTP is available.
- Slide barcodes (`submitter_id` values for Slide entities in the Legacy Archive) are not available
- SDF Files are not linked to Project or Case in the Legacy Archive
- There are 200 cases from TCGA-LAML that do not appear in the Legacy Archive
- Biotab files are not linked to Project or Case in the Legacy Archive
- SDRF files are not linked to Project or Case in the Legacy Archive
- Portion "weight" property is incorrectly described in the Data Dictionary as the weight of the patient in kg, should be described as the weight of the portion in mg

Details are provided in Data Release Manifest

Initial Data Release (1.0)

- **GDC Product**: Data
- **Release Date**: June 6, 2016

Available Program Data

- The Cancer Genome Atlas (TCGA)
- Therapeutically Applicable Research To Generate Effective Treatments (TARGET)

Available Harmonized Data

- WXS
  - Co-cleaned BAM files aligned to GRCh38 using BWA
- mRNA-Seq
  - BAM files aligned to GRCh38 using STAR 2-pass strategy
  - Expression quantification using HTSeq
- miRNA-Seq
  - BAM files aligned to GRCh38 using BWA aln
  - Expression quantification using BCCA miRNA Profiling Pipeline*
- Genotyping Array
  - CNV segmentation data

Known Issues and Workarounds

- BAM files produced by the GDC RNA-Seq Alignment workflow will currently fail validation using the Picard ValidateSamFiles tool. This is caused by STAR2 not recording mate mapping information for unmapped reads, which are retained in our BAM files. Importantly, all affected BAM files are known to behave normally in downstream workflows including expression quantification.
- All legacy files for TCGA are available in the GDC Legacy Archive, but not always linked back to cases depending on available metadata.
- Public MAFs (those with germline variants removed) are only available for MuTect2 pipeline. MAFs for other pipelines are forthcoming.
- TARGET-AML and TARGET-ALL projects are undergoing reorganization. Pending reorganization, cases from these projects may not contain many clinical, biospecimen, or genomic data files.
- No data from TARGET-PPTP is available.
- Legacy data not available in harmonized form:
  - Annotated VCF files from TARGET, anticipated in future data release
  - TCGA data that failed harmonization or QC or have been newly updated in CGHub: ~1.0% of WXS aliquots, ~1.6% of RNA-Seq aliquots
  - TARGET data that failed harmonization or QC, have been newly updated in CGHub, or whose project names are undergoing reorganization: ~76% of WXS aliquots, ~49% of RNA-Seq aliquots, ~57% of miRNA-Seq.
- MAF Column #109 "FILTER" entries are separated by both commas and semi-colons.
- MAFs are not yet available for query or search in the GDC Data Portal or API. You may download these files using the following manifests, which can be passed directly to the Data Transfer Tool. Links for the open-access TCGA MAFs are provided below for downloading individual files.
  - Open-access MAFs manifest
  - Controlled-access MAFs manifest

Details are provided in Data Release Manifest

Download Open-access MAF files