

Standard Sequencing Service Getting Started Guide

Standard Pipeline 2.0

This document provides an overview of Complete Genomics data and references to additional documentation, tools, and resources that will familiarize you with Complete Genomics data and maximize the value of your sequencing project.

Table of Contents

5 Things You Need to Know	3
Understanding the Data Structure and Content.....	4
Data Structure.....	4
Data Content.....	5
Quality Information.....	5
Small Variations	6
Coverage.....	6
Copy Number Variations (CNVs).....	6
Structural Variations (SVs)	6
Mobile Element Insertions (MEIs).....	7
How to Manage Sequencing Data.....	8
Data Volume	8
Data Integrity.....	8
Step 1: Make Backup Copies of Your Data.....	8
Step 2: Check that the Data Package is Complete.....	8
Resources for Interpreting the Data	10
Analysis Tools and Resources.....	10
Documentation.....	10
Getting Support	11
Complete Genomics Data Analysis Training	11
Call Center	11
Questions about Complete Genomics Data Files and Analysis.....	11
Questions about CGA Tools.....	11
Field Application Science Team	11
Complete Genomics User Community.....	11

5 Things You Need to Know

If you haven't worked with Complete Genomics data before, we want to do everything we can to make sure you are off to a good start. This short checklist will help you get the most from your sequencing data:

- Schedule a local [Workshop Training](#) if you haven't yet done so.
- [Backup your data and perform a data integrity check.](#)
- Familiarize yourself with the [structure and content](#) of the data.
- Prepare yourself for [downstream analysis](#) by becoming acquainted with Complete Genomics open-source software tools package [CGA™ Tools](#) and our [Tool Repository](#).
- Make sure you know where to get help: learn about Complete Genomics [support resources](#).

Understanding the Data Structure and Content

Data Structure

The data delivered by Complete Genomics include sequenced reads, their mappings to a reference human genome, and the annotated variants detected with respect to the reference. The data files are organized according to the directory structure shown in Figure 1 and 2. Most files are provided in tab-delimited text format (.tsv) or in a compressed version of this format (tsv.bz2). Detailed descriptions of Complete Genomics data files can be found in our [Data File Formats](#) document.

Figure 1: Complete Genomics Genome Data Structure

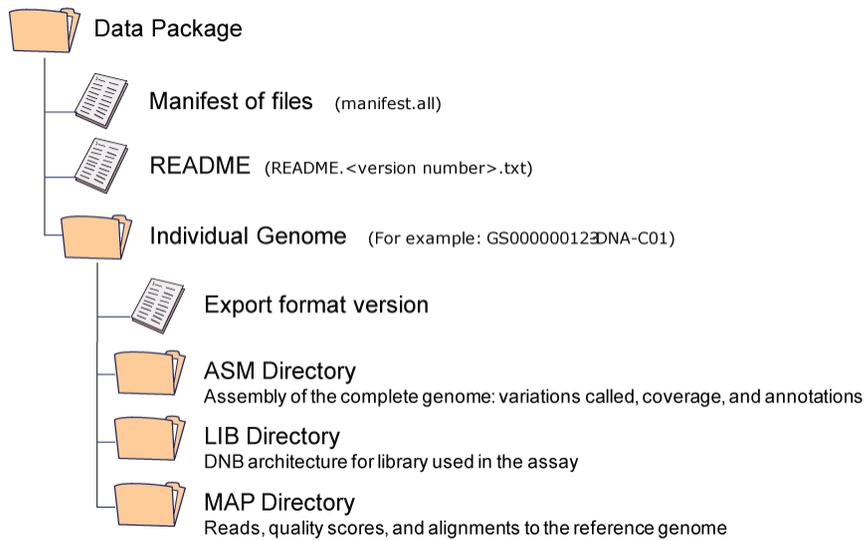
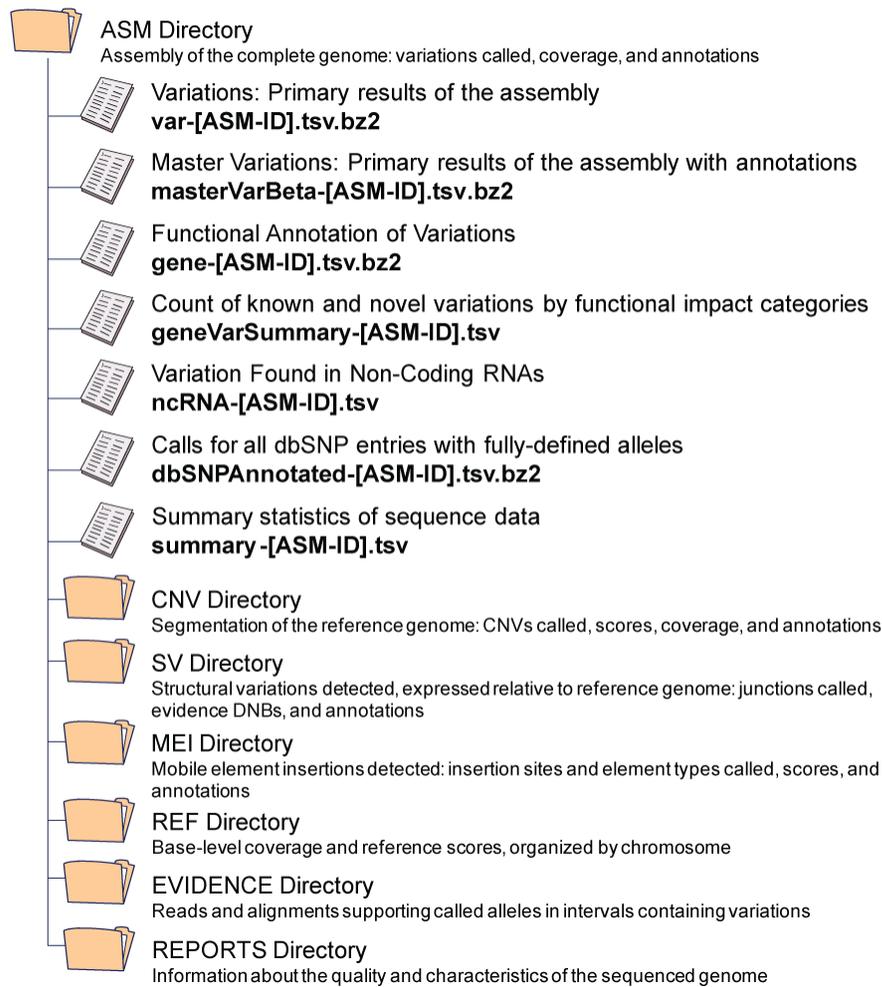


Figure 2: ASM Directory Structure

Data Content

Each individual genome data package contains three sub-directories:

- **ASM** — contains information on coverage, annotated called variants, and the reads and mappings that support the variant calls.
- **LIB** — contains information on the library of clones (DNBS) used for sequencing.
- **MAP** — contains the raw reads, their quality scores, and their initial mappings to the reference genome. The mappings in this directory support homozygous reference calls across the genome.

Quality Information

Information on the quality of the genome sequence can be found in two locations:

- The **summary-[ASM-ID].tsv** file (located in the ASM directory) contains summary statistics of sequencing metrics and variations. This file provides information about coverage, call rate, the number of variants (including small variations, copy number variations, and structural variations), and the transition-to-transversion ratio.
- The **REPORTS** directory contains additional files that support the statistics described in the summary file and also contains files describing the distribution of coverage, GC bias, and the size

of small indels across the genome. These files can be used to plot the genome-wide or exome-wide coverage distribution or to plot coverage as a function of GC content.

Small Variations

Small variant calls (SNPs, substitutions, insertions and deletions of ≤ 50 bp) across the entire genome are provided in two formats. The ***masterVarBeta-[ASM-ID].tsv.bz2*** file presents variants in a single line per locus format, whereas the ***var-[ASM-ID].tsv.bz2*** file presents variants in a multiline format. Both are located in the ASM directory. ***masterVarBeta*** contains a rich set of annotations including:

- Read count support for each allele
- Variant scores (*varScoreVAF* and *varScoreEAF*) and confidence category (*varQuality*) that indicate whether the called variation is likely to be true
- Matching dbSNP and COSMIC record IDs
- RefSeq transcript, PFAM, and functional annotations
- Overlapping miRNAs
- Copy number information

Customers interested in the exome can refer to the ***gene-[ASM-ID].tsv.bz2*** file, which lists all variants identified in each RefSeq transcript. It provides the location of the variant within the transcript, COSMIC and dbSNP annotations, PFAM information, and functional annotations, including information on protein coding alterations. The ***geneVarSummary-[ASM-ID].tsv*** file summarizes this information, distilling it down to the number of known and novel variants predicted to affect protein function for each RefSeq transcript.

The EVIDENCE directory contains the reads and mappings that support small variant calls. The mappings in this directory are generated by local *de novo* assembly, which identifies reads that support the variants called at the given locus. Because of the presence of the variant sequence, some of these reads may not have aligned to the reference genome during the initial mapping, but will be present here. Conversely, some reads mapped to this location during initial mapping will be absent from the EVIDENCE directory, because the more precise and stringent local *de novo* assembly will exclude them.

Coverage

The ***coverageRefScore-[chr]-[ASM-ID].tsv.bz2*** file, located in the ASM/REF directory, provides various measurements of coverage for each position in the reference genome. These measurements are based on the initial mappings of the reads to the reference genome and can be used to plot coverage across the genome.

Copy Number Variations (CNVs)

Coverage information is used to segment the genome into distinct regions of ploidy. The coverage, confidence scores, ploidy, and annotations are then reported for each segment in the ***cnvSegmentsDiploidBeta-[ASM-ID].tsv*** file, located in the ASM/CNV directory. It also contains annotations for regions of abnormal ploidy with information about genes that overlap or are contained within the CNV and links to known CNVs and repetitive elements. The related ***cnvDetailsDiploidBeta-[ASM-ID].tsv*** file provides ploidy and coverage information for 2kb windows across the reference genome and can be used to visualize CNVs.

Structural Variations (SVs)

Discordant mate-pair analysis is used to identify junctions, adjacent regions in the genome that are not adjacent in the reference genome. Where possible, reads that straddle the junction are assembled to provide the sequence of the junction. The ***highConfidenceJunctionsBeta-[ASM-ID]-ASM.tsv*** file, located in the ASM/SV directory, lists high-quality junctions in the sample genome that have been

filtered using a variety of quality metrics to enhance specificity. Junction information includes the junction location, the assembled junction sequence, the number of reads supporting a given junction, and the size of the insertion or deletion for intrachromosomal junctions. The file is also annotated with information on repeats and transcripts that overlap the junction as well as known structural variations.

In addition to providing a list of junctions, Complete Genomics rationalizes those junctions into the structural rearrangement events from which they derive, including insertions, deletions, translocations, and inversions. The ***highConfidenceSvEventsBeta-[ASM-ID].tsv*** file lists all SV events predicted by rationalizing high confidence junctions and annotates each putative event with information including event type, genomic location, and any gene fusions or disruptions predicted by the event.

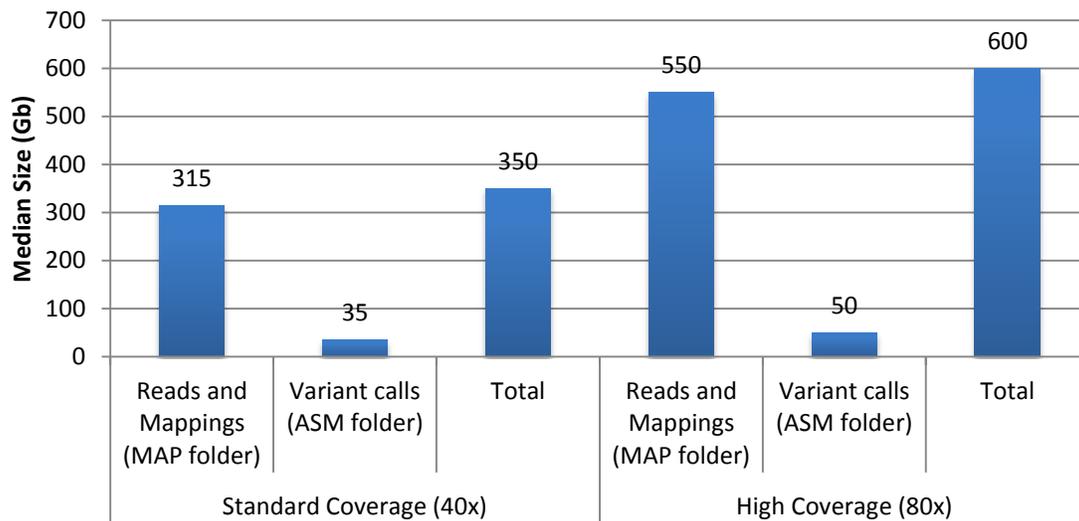
Mobile Element Insertions (MEIs)

Novel MEIs (those not present in the reference genome) are identified by locating clusters of mate pairs where one end maps unexpectedly to a mobile element and the other maps to unique reference sequence. The presence, rather than the precise location or complete sequence of the novel MEI is recorded. The ***mobileElementInsertionsBeta-[ASM-ID].tsv*** file provides information on each insertion, including location, mobile element type, number of supporting reads, and overlapping genes. ***mobileElementInsertionsRefCountsBeta-[ASM-ID].png*** graphs the number of DNBS supporting the reference sequence for each MEI and can be used to predict MEI zygosity. Finally, the ***mobileElementInsertionsROCBeta-[ASM-ID].png*** graph shows the relationship between a specific MEI confidence score and the predicted frequency of false negatives and false positives, and should be used to select a score cutoff to balance the needs of specificity and sensitivity.

How to Manage Sequencing Data

Data Volume

This chart shows an overview of the sizes of the data packages produced using Complete Genomics Standard Sequencing Service:



For more information about data management, including receiving, checking, and working with Complete Genomics data, see the frequently asked questions in [Managing Data FAQ](#).

Data Integrity

IMPORTANT

You **must** ensure the security and completeness of your data within the first 30 days after receipt. As indicated in the Complete Genomics Service contract, Complete Genomics will begin deleting customer data 30 days after shipment to the customer.

Complete Genomics recommends taking the following steps to ensure and maintain data integrity:

- [Step 1: Make Backup Copies of Your Data](#)
- [Step 2: Check that the Data Package is Complete](#)

Step 1: Make Backup Copies of Your Data

Complete Genomics strongly recommends that you make backup copies of the data. If you receive data on a hard disk drive, we do not recommend using the delivered hard drives as the primary backup location. This storage method provides no redundancy in case of drive failure.

Step 2: Check that the Data Package is Complete

After receiving Complete Genomics data, one of your first actions should be to verify that all data files are present and uncorrupted. Data verification should be performed using the **manifest.all** file provided for each genome assembly. The **manifest.all** file contains SHA-256 checksums of each file in

the data package (except for the *manifest.all* and *manifest.all.sig*), and is suitable for use with the sha256sum tool present on Linux operating systems (also available for other platforms).

Assuming the data is copied to another system immediately upon receipt (both to provide working storage and as a backup), customers should check the SHA-256 sums on the copy made, and if any problems arise, check the SHA-256 sums on the original hard disk drive. If any data appears to be missing or corrupted you should contact support@completegenomics.com immediately.

To check the integrity of the data package:

On Linux:

```
sha256sum -c /path/to/manifest.all
```

On Mac OS X, from the terminal window:

```
shasum -a 256 -c /path/to/manifest.all
```

If no errors are reported, the verification was successful.

Complete Genomics also provides a security certificate that can be used to ensure that the data provided to the customer was shipped from Complete Genomics. Please see "[How do I verify that the data files are present and uncorrupted?](#)" in the Managing Data FAQ for more information.

Resources for Interpreting the Data

Analysis Tools and Resources

Several tools are available for downstream analyses of Complete Genomics data:

- [CGA Tools](#) — A repertoire of file conversion and analysis modules developed by Complete Genomics.
- [Complete Genomics Tool Repository](#) — The Tool Repository on the Complete Genomics User Community website contains a collection of Complete Genomics authored scripts for downstream analysis.
- Complete Genomics Public Genome Data Repository — Complete Genomics offers whole human genome sequence data sets on its FTP server for free download and general use. These data result from the sequencing of 69 standard, non-diseased samples as well as two matched tumor and normal sample pairs. This large data set may be used for many things including filtering out non-disease causing variants. Resources available include:
 - [Overview of the Public Genomes](#)
 - [Public Genome Data Repository Service Note](#) — detailed description of the public genome sequence data.
 - [Summary Analysis Readme](#) — Overview of additional files containing summary statistics across the standard genomes and lists of the variants found in each genome.
- [Software Partners](#) — Complete Genomics has partnered with some of the leading developers of software for downstream analysis. The tools offered by our partners have been tested to ensure compatibility with Complete Genomics data.
- [Third-Party Tools](#) — A list of tools that can be used to analyze Complete Genomics data.

Documentation

Complete Genomics provides an abundance of documentation on our website:

<http://www.completegenomics.com/customer-support/documentation/>

The following documents may be particularly helpful to new users of our services.

- [Complete Genomics Technology Whitepaper](#) — A concise description of the Complete Genomics sequencing technology, including the library construction process and the ligation-based sequencing.
- [Complete Genomics Data File Formats Standard Pipeline 2.0](#) — A description of the organization and content of the format for complete genome sequencing data delivered by Complete Genomics.
- [Complete Genomics Analysis Pipeline Release Notes](#) — Indicates new features and enhancements by release.
- [Complete Genomics FAQs](#) — Important questions and answers related to data management, variant calling, and other topics.

Getting Support

Complete Genomics Data Analysis Training

Complete Genomics Field Application Science team provides bioinformatics training to assist its customers in the analysis of Complete Genomics data. These free training programs are organized regionally and scheduled based on customers' needs. Complete Genomics highly recommends you and your team attend a workshop to maximize your knowledge of Complete Genomics sequencing platform, data files, and best practices for downstream analysis. For more information, or to arrange training, contact your Complete Genomics Account Representative or email Complete Genomics support at support@completegenomics.com.

Call Center

Complete Genomics Call Center is available to answer technical questions during regular business hours (8am-6pm Pacific Time). The call center aims to respond to all questions within 24 business hours.

- support@completegenomics.com
- Toll-free in the US or Canada: 1-855-267-5383
- International: 1-650-943-2600

Questions about Complete Genomics Data Files and Analysis

When contacting the call center for questions about data and downstream analysis, provide the following information from the header of the *summary-[ASM-ID].tsv* file located in the ASM directory of the data package:

Header Field	Description
#SAMPLE	Sample ID for the genome(s) in question. For example, "GS11111-DNA_A01".
#SOFTWARE_VERSION	Version of the Complete Genomics Analysis Pipeline software used to generate the data.
#GENOME_REFERENCE	Version of the haploid human reference used for genome assembly.

Questions about CGA Tools

When contacting the call center about CGA Tools, provide the following information:

- Version of CGA Tools (run `cgatools` from the command line with no arguments)
- Operating system (run `uname -a` on the command line)
- Was CGA Tools installed using the provided binaries or recompiled from source?

Field Application Science Team

Complete Genomics Field Application Scientists (FAS) are experienced bioinformaticians who can help you maximize utility of your Complete Genomics data. Contact support@completegenomics.com to obtain the contact information for your local FAS.

Complete Genomics User Community

[Complete Genomics User Community](#) provides a forum where customers can interact, collaborate, and share knowledge with each other. The Tools Repository is also located in the User Community.