Managing Data
Frequently Asked Questions (FAQ)

Updated February 2013

Receiving Data

I just received a shipment of data from Complete Genomics. What should I do first?

After receipt of your data, you should ***immediately*** verify that all data files are present and uncorrupted (see "How do I verify that the data files are present and uncorrupted? When should I do this?"). Complete Genomics strongly recommends that you make backup copies—at least two separate copies on separate devices—of any critical data. Complete Genomics makes no commitments to retain data after delivery. If you delete or lose data that you have not backed up, it is not retrievable.

If I am just using the variant files and other processed output, can I get rid of the reads and initial mappings?

It is up to you to determine which data you need to archive, but keep in mind that Complete Genomics does ***not*** retain customer data, so any data you permanently delete is irretrievable. Also, recall that all disk drives, including those sent by Complete Genomics, have a finite lifetime and a failure rate. Complete Genomics strongly recommends that you make and keep backup copies (at least two separate copies on separate devices) of any critical data.

If you intend to publish your results, then you may be required by the journal or by your funding source to submit the reads to a central database. You may wish to investigate any such requirement before making decisions about data retention.

If you will be focusing on the processed data from Complete Genomics (such as variant calls), but wish to retain the reads and initial mappings, you may want to consider storing them onto slower less expensive storage than the other files. Cloud storage such as Amazon's Web Services (AWS) may also be an option worth considering. AWS is an infrastructure web services platform that provides remote computing power, storage, and other services.

What type of computer do I need to transfer the data from a Complete Genomics disk drive?

Disk drives provided by Complete Genomics are formatted using NTFS, which is readable by most operating systems.

Important: Do not connect the hard disk drive to a computer running the Windows XP operating system. Windows XP is not compatible with 3 TB drives due to a maximum disk volume size of 2 TB. For more information, see: msdn.microsoft.com/en-us/windows/hardware/gg463525.aspx.

What is the fastest method for copying data from drives shipped by Complete Genomics? How long will it take?

Drives are shipped with a USB 3.0 interface. One can expect that the transfer will require 15-30 minutes to transfer data from the drive to the computer for one standard (40x) genome, depending on other aspects of the system.

If you are connecting the drive to one computer and transferring the data through your network to a second computer (such as a file server), then the network speed will also greatly impact the time required. Almost any wireless network will be quite slow, as will either a 10Mbit or 100Mbit wired connection. Furthermore, when using these types of networks, your data copy may effectively monopolize the network and impact other users. A 1-gigabit wired network with good quality switches is strongly preferred. You may even wish to set up a dedicated subnet for this purpose.

We recommend that you contact your computer support staff for help on these issues. Complete Genomics can only be of limited assistance because each computing and network environment is unique.

How do I verify that the data files are present and uncorrupted? When should I do this?

As of Complete Genomics' pipeline version 1.9, we provide a manifest file containing SHA-256 checksums of each file, suitable for use with the sha256sum tool present on most Linux operating systems (and available for many other platforms). The checksums are computed on all the delivered files in our EXP package (except for the ***manifest.all*** and ***manifest.all.sig***). If the files are uncompressed and recompressed then the SHA-256 hashes may not match.

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. Customers should immediately contact support@completegenomics.com if any issues are noticed.

To determine whether a transmission error has occurred:

- Check the integrity of the package:

  On Linux:

  ```
  sha256sum -c manifest.all
  ```

  On Mac OS, from the terminal window:

  ```
  shasum -a 256 -c manifest.all
  ```

  If no errors are reported, the verification was successful.

To determine if data has been intentionally modified:

1. Select and download an appropriate X.509 certificate from Complete Genomics at: ftp://ftp.completegenomics.com/Certs/

   The certificate should match the software version used to analyze the data. The software version can be found in the header of each file in the package.

Software Version	Certificate
1.10 and older	cgiexpcert-1_10-and-older.pem
1.11 – 2.0.2.21	cgiexpcert-1_11-2_0_2_21.pem
2.0.2.22 and newer	cgiexpcert.pem

2. Verify the Complete Genomics certificate. For example, use the command:

   ```
   openssl verify cgiexpcert.pem
   ```

   If the verification failed with a message "error 20 at 0 depth lookup: unable to get local issuer certificate" do the following:

   - Get the local issuer certificate. For example, if the certificate authority (CA) is GoDaddy.com:

     ```
     wget https://certs.godaddy.com/repository/gd_bundle.crt -O gd.pem
     ```

   - Verify it:

     ```
     openssl verify gd.pem
     ```

   - If the result is "OK", verify the Complete Genomics certificate:

     ```
     openssl verify -CAfile gd.pem cgiexpcert.pem
     ```

If no errors were displayed, the verification of the certificate was successful.

Note: For the older versions of the certificates, the verification process may produce a "certificate has expired" message. If the "Verified OK" message also appears, the integrity of your data has been verified.

3. Extract the Complete Genomics public key from the certificate:

```
openssl x509 -inform pem -in cgiexpcert.pem -pubkey -noout >cgipubkey.pem
```

4. Use the public key to check the digest of ***manifest.all*** from a delivered package:

```
openssl dgst -sha256 -verify cgipubkey.pem -signature manifest.all.sig
manifest.all
```

If no errors were reported on each step, the verification was successful.

Can I put the Complete Genomics data files on a Windows PC and work with them?

Yes, however there are important caveats to keep in mind. Refer to "What type of computer do I need to transfer the data from a Complete Genomics disk drive?" for more information on compatibility of specific Windows versions. Further, while in general, text files are the same between Windows and other operating systems (particularly Mac OS X, Unix, and Linux), Windows text files use different conventions for marking the ends of lines (a CR and LF character is used, while a LF only is used on the other non-Windows systems). Because most Complete Genomics users have requested to work with the data on Unix, Linux, and Mac OS, Complete Genomics files are provided with LF-only line breaks. Some Windows software works well on these files, in spite of a lack of a CR character, while other Windows programs will require that the files be converted to work properly. Contact your Windows support staff for utilities that can help with this conversion.

Some Complete Genomics users prefer to use a Unix-like environment on their Windows machines to handle this data, most commonly the free Cygwin package. Cygwin can be installed in a mode where LF-only files are read and written (Cygwin can also be installed in a CR-LF mode, which is not recommended for working with our data). Note that using such an environment requires command-line Unix skills. In addition, the FAT32 file system on Windows allows a maximum file size of only 2 GB, which is too small for many genomic data sets—including many of the Complete Genomics data files. Contact your Windows support staff for help, particularly if you have an older computer or are using external drives (which often come pre-formatted with FAT32).

Can I put the Complete Genomics data files on a Linux, Unix, or Mac OS X computer and work with them without having to do data format conversion?

Yes; this is what most of our customers do.

Should I uncompress the data?

Uncompressing all of the data files will increase the required storage for a single genome approximately 3 to 4 fold, typically to over 1.5 TB per standard genome and over 3 TB per high-coverage genome. Approximately ninety percent of this volume is used by the reads and mappings files, and most Complete Genomics customers leave these files in their compressed format for this reason. Further, the utilities in CGA Tools work with compressed data.

In general, there are a number of methods for analyzing data in compressed files. For example, for those familiar with Unix and Linux commands, to use streaming decompression to count all genotypes of known dbSNP variants on chromosome 21 in a Complete Genomics genome (here, GS0000084-ASM) you could run the command:

```
bzcat dbSNPAnnotated-GS0000084-ASM.tsv.bz2 | grep chr21 | wc -l
```

Something seems wrong with a file or I deleted some data. Can Complete Genomics get me a replacement copy?

Complete Genomics begins to delete data 30 days after delivery unless special arrangements have been made. Upon receipt of data from Complete Genomics, if any data appears to be missing or corrupted you should contact support@completegenomics.com immediately.

Checking Data Details

Where can I find the format version number for the data files? What is the difference between the format version and the software version numbers?

For Analysis Pipeline 2.0 and later, the versions of the data file formats and the software are synchronized to the same number. The version is provided in the version file located in the root directory (e.g., "`/DID-ID/ASM-ID/GSXXXXX_DNA-YYY`") and also in the header of all Complete Genomics data files under the keys "FORMAT_VERSION" and "SOFTWARE_VERSION".

For Analysis Pipelines prior to 2.0, data included a separate data format version number and software version number. All Complete Genomics data files contained a header (#FORMAT_VERSION) indicating the data format version number. In addition, a small text file named "version" was included in the individual genome results directory (for example, "`GS00001-DNA-A01`") of the data file hierarchy for each genome containing the same value. We recommend that your processing programs check this version number to ensure compatibility.

Prior to software version 1.7, the format version was in the #VERSION header. We changed the name for clarity. Prior to software version 1.8, this file was located in the top-level directory.

How do I determine which version of the reference human genome was used during mapping and assembly?

The header in Complete Genomics data files produced with pipeline version 1.7 or greater specifies the reference genome build used (#GENOME_REFERENCE). Data generated with pipeline versions prior to 1.7 (before this header was included) used NCBI build 36. Starting in software version 1.8, you can choose either NCBI build 36 (corresponding to Hg18) or GRCh37 (Hg19) as the reference genome.

How do I determine which version of the annotation databases (such as RefSeq or dbSNP) was used?

The header in Complete Genomics data files produced with pipeline version 1.7 or greater contains annotation version information for the relevant data sources (e.g. #GENE_ANNOTATIONS and #DBSNP_BUILD). Data generated with pipeline versions prior to 1.7 (before this header was included) always used dbSNP 129, NCBI build 36.3 and RefSeq alignments.

Working with the Data

What do I need to understand about Complete Genomics' methods and output in order to work with the data? Where can I get that information?

The *Complete Genomics Standard and Cancer Sequencing Service Getting Started Guides* provide an overview of Complete Genomics documentation, training, and support resources. These files can be downloaded at the following URLs:

http://www.completegenomics.com/customer-support/documentation/162097975.html

http://www.completegenomics.com/customer-support/documentation/162121635.html

Complete Genomics sent me disk drives. Can I just work directly with the data on those devices?

With the exception of eSATA or USB 3.0 connections, working directly with data on external drives will be much slower than working with the files on either a local disk drive or on a good quality file server over a fast network. It is also strongly recommended that you create a back up copy of your data before working with it (see "I just received a shipment of data from Complete Genomics. What should I do first?"). Apple® Mac® users mounting the NTFS hard drive will have Read-Only access to the files.

What is a .tsv file? Why is this extension not recognized by my OS or software?

TSV stands for "tab-separated values." This generic format is also called "tab-delimited text." Unfortunately, there is no consistent naming convention across all software and systems for this format. Some software defaults to looking for such data in .tab, .tdt, .text, or .txt files. Microsoft applications will default to .csv (generally meaning "comma-separated values") when importing tab-delimited files. To find the file, you can select "all files" in the file type filter and import a .tsv file into your software program of choice.

I have a file derived from Complete Genomics-produced data that is not mentioned in the Data File Formats document. Who can help me with it?

Complete Genomics may be of limited assistance in understanding data that has been transformed by tools not supplied by Complete Genomics, or that has been further analyzed by non-Complete Genomics software.

What is Complete Genomics C++ library and API? Do I need to know how to program in C++ to use Complete Genomics data?

While the variants were provided in text files, Complete Genomics data prior to pipeline version 1.4 required use of a C++ library and API to access the individual reads and mappings. The C++ API does not work with data produced after this time (after version 1.4 or later than September 2009).