Complete genomics

SERVICE NOTE

Public Genome Data Repository

General Information

Complete Genomics offers whole human genome sequence data sets on its FTP server (ftp2.completegenomics.com) 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.

Standard samples include a Yoruban trio, a CEPH/Utah pedigree of 17 family members, a Puerto Rican trio, and a diversity panel representing nine different populations, from the NIGMS and NHGRI Repositories (HapMap and 1000 Genomes Project samples). Collections were drawn from the Coriell Institute for Medical Research. Samples were sequenced to an average genome-wide coverage of about 80X (range of 51X to 89X). The first release of these genomes was generated using an earlier Complete Genomics Analysis Pipeline for calling variations across the genome (version 1.10.0). The current release of the data set was generated with the most recent Analysis Pipeline (version 2.0.0), and some samples were resequenced with the latest library preparation process.

Matched tumor and normal cell line sequence data are available for two patients with breast cancer. Collections were drawn from ATCC. Samples were sequenced to an average genome-wide coverage of 123X for three of the samples, and 92X for HCC2218 BL.

By providing the research community with this public data set, we encourage researchers to validate our sequencing performance and to further improve data analysis and interpretation methods.

Reference Databases

Complete Genomics can use either NCBI build 36 (hg18) or GRCh37 (hg19[1]), as a reference genome during its data analysis process.

- All genomes are analyzed with GRCh37, using dbSNP version 132 for annotating known variations. RefSeq annotations come from NCBI annotation build 37.2.
- Genomes are also analyzed with NCBI build 36, using dbSNP version 130 for annotating known variations. RefSeq annotations come from NCBI annotation build 36.3.

[1]Note that hg19 uses a Yoruba mitochondrion sequence while Complete Genomics uses the Cambridge Reference Sequence.

Available Sample Information

Example sequence data resulting from our Standard Sequencing Service are available for a diversity panel (Table 1), extended CEPH/Utah Pedigree 1463 (Table 2), and the Yoruba and Puerto Rican Trios (Table 3). Example sequencing data resulting from our Cancer Sequencing Service are available for tumor-normal pairs (Table 4).

DIVERSITY SET			
Coriell ID	**Population**	**Coriell ID**	**Population**
NA19700	ASW	NA19020	LWK
NA19701	ASW	NA19025	LWK
NA19703	ASW	NA19026	LWK
NA19704	ASW	NA21732	MKK
NA19834	ASW	NA21733	MKK
NA06985	CEU	NA21737	MKK
NA06994	CEU	NA21767	MKK
NA07357	CEU	NA19735	MXL
NA10851	CEU	NA19648	MXL
NA12004	CEU	NA19649	MXL
NA18526	CHB	NA19669	MXL
NA18537	CHB	NA19670	MXL
NA18555	CHB	NA20502	TSI
NA18558	CHB	NA20509	TSI
NA20845	GIH	NA20510	TSI
NA20846	GIH	NA20511	TSI
NA20847	GIH	NA18501	YRI
NA20850	GIH	NA18502	YRI
NA18940	JPT	NA18504	YRI
NA18942	JPT	NA18505	YRI
NA18947	JPT	NA18508	YRI
NA18956	JPT	NA18517	YRI
NA19017	LWK	NA19129	YRI

Table 1. *Diversity Panel*

Legend, Tables 1-3:

ASW: African ancestry in Southwest USA
CEU: Utah residents with Northern and Western European ancestry from the CEPH collection
CHB: Han Chinese in Beijing, China
GIH: Gujarati Indian in Houston, Texas, USA
JPT: Japanese in Tokyo, Japan
LWK: Luhya in Webuye, Kenya
MKK: Maasai in Kinyawa, Kenya
MXL: Mexican ancestry in Los Angeles, California
TSI: Toscans in Italy
YRI: Yoruba in Ibadan, Nigeria
PUR: Puerto Rican in Puerto Rico
YRI: Yoruba in Ibadan, Nigeria

CEPH/UTAH PEDIGREE 1463	
Coriell ID	**Population**
NA12877	CEPH/Utah Pedigree 1463
NA12878	CEPH/Utah Pedigree 1463
NA12879	CEPH/Utah Pedigree 1463
NA12880	CEPH/Utah Pedigree 1463
NA12881	CEPH/Utah Pedigree 1463
NA12882	CEPH/Utah Pedigree 1463
NA12883	CEPH/Utah Pedigree 1463
NA12884	CEPH/Utah Pedigree 1463
NA12885	CEPH/Utah Pedigree 1463
NA12886	CEPH/Utah Pedigree 1463
NA12887	CEPH/Utah Pedigree 1463
NA12888	CEPH/Utah Pedigree 1463
NA12889	CEPH/Utah Pedigree 1463
NA12890	CEPH/Utah Pedigree 1463
NA12891	CEPH/Utah Pedigree 1463
NA12892	CEPH/Utah Pedigree 1463
NA12893	CEPH/Utah Pedigree 1463

Table 2. *Extended CEPH/Utah Pedigree 1463*

TRIOS	
Coriell ID	**Population**
HG00731	PUR
HG00732	PUR
HG00733	PUR
NA19238	YRI
NA19239	YRI
NA19240	YRI

Table 3. *Yoruban and Puerto Rican Trios*

Sample	ATCC Number	Details
HCC1187	CRL-2322	Breast cancer (primary ductal carcinoma): TNM stage IIA, grade 3
HCC1187 BL	CRL-2323	Derived from peripheral blood and immortalized with EBV transformation. Normal match for HCC1187.
HCC2218	CRL-2343	Breast cancer (primary ductal carcinoma): TNM stage IIIA, grade 3
NA12880	CRL-2363	Derived from peripheral blood and immortalized with EBV transformation. Normal match for HCC2218.

Table 4. *Tumor-normal Pairs*

FTP Server Organization

The data is organized on Complete Genomics' ftp server (ftp2.completegenomics.com) in multiple directories: YRI_trio, PUR_trio, Pedigree_1463, Diversity, Cancer_pairs, VCF_files, and Multigenome_summaries. Figure 1 is a partial representation of the entire FTP directory. In the example, the filenames use sample NA19240 and build 37. For other samples and/or for build 36, replace NA19240 with the appropriate sample name from above and substitute 37 for 36.

Below these directories are three subdirectories: ASM_Build37_2.0.0, MAP_Build37_2.0.0, and ASM_Build36_2.0.0. Within each of these directories are sample sub-directories as designated by the sample name. Within each sample sub-directory are tar archives containing the assembly results (ASM indicates "assembly").

- NA19240-200-37-ASM-VAR-files.tar
- NA19240-200-37-ASM-EVIDENCE-files-pt1.tar
- NA19240-200-37-ASM-EVIDENCE-files-pt2.tar
- NA19240-200-36-ASM-EVIDENCE-files-pt3.tar
- NA19240-200-37-ASM-EVIDENCE-files-pt4.tar
- NA19240-200-37-ASM-REF-files-pt1.tar
- NA19240-200-37-ASM-REF-files-pt2.tar
- NA19240-200-37-ASM-REF-files-pt3.tar
- NA19240-200-37-ASM-REF-files-pt4.tar
- NA19240-200-37-ASM-REF-files-pt5.tar

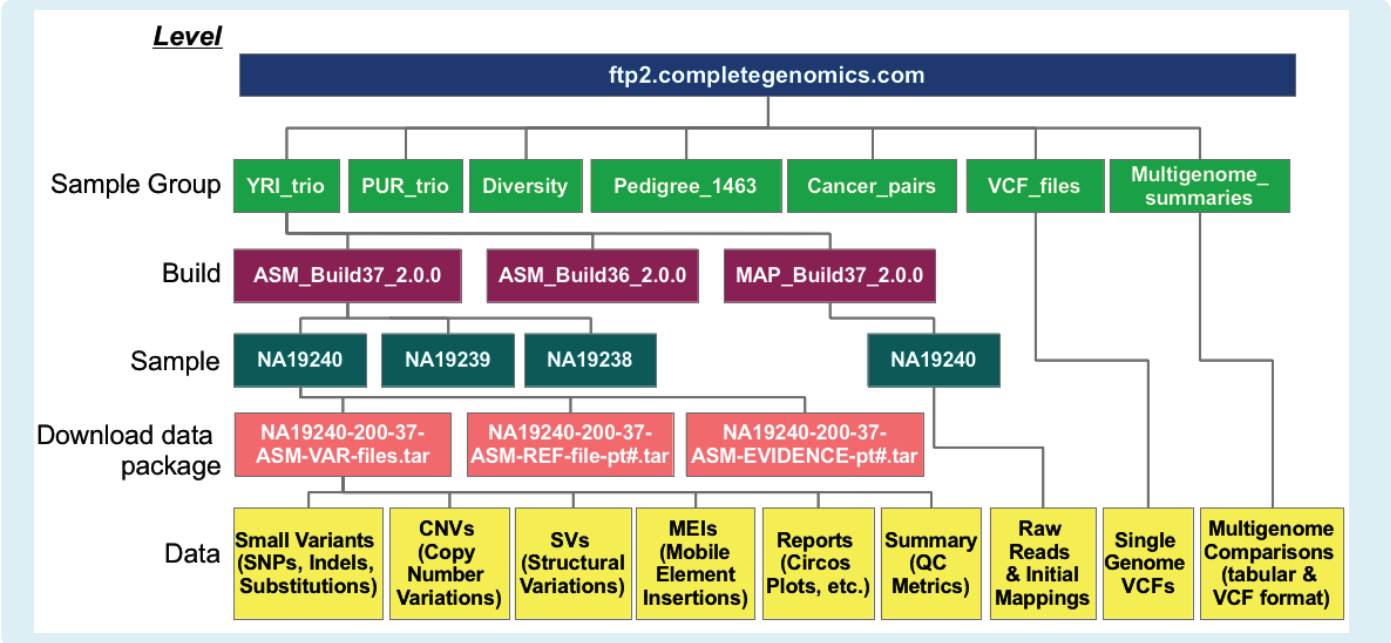

Figure 1. *FTP Server Directory Structure*

The VAR-files archives are approximately 1 GB each and contain the variation calls (including SNPs, indels, copy number variations, structural variations, and mobile element insertions), scores, and annotations. For many purposes this might be the only data needed. The EVIDENCE files contain the reads and mappings that support the variant calls. The REF files contain the refScores (confidence that a position is homozygous reference) and coverage information for every basepair in the reference genome. These files can be used to evaluate mapping results from the assembly process. They are also used in CGA™ Tools for comparing genomes, visualizing read support for variations, and converting reference and evidence mappings to SAM format. These files are broken up into multiple tarballs (pt1, pt2, etc.) approximately 3 to 3.5 GB, to reduce the file sizes for ease in downloading. Collectively these files add up to 25 to 45 GB per genome.

The MAP_Build37_2.0.0 directories (Figure 1) contain all of the raw reads and initial (prior to *de novo* assembly) genome-wide mappings of those reads. These files are approximately 300 GB per genome, or larger, which is why only Build 37 mapping data has been made available.

Contents of the Data Set

Each data set includes several directories. The most relevant directory is the ASM directory which contains the overall results of the sequencing project, including identified variants, scores, and annotations. For those interested in the raw sequencing data, the MAP directory contains individual reads, quality scores, and initial mapping results.

For each publicly available genome, Complete Genomics provides the following whole genome information:

- Variation Reporting, including:
  - SNPs, small insertions, deletions, substitutions, and complex small variants
  - Copy number variations (CNVs)
  - Structural variations (SVs)
  - Mobile element insertions
- Annotations, including:
  - Genes and functional impact
  - dbSNP and minor allele frequency (MAF) from the 1000 Genomes Project
  - Non-coding RNA
  - Known segmental duplications
  - Type of SV
- Extensive scoring
- Evidence files, including read support information for the variation
- Per base-pair coverage for each position in the reference genome
- For cancer pairs only: somatic small variants (eg, SNPs and indels), somatic CNVs, and somatic SVs

Additional output includes:

- VCF files generated using the CGA Tools mkvcf command are provided for each genome in the VCF_files folder. These files summarize all variants (small variants, CNVs, SVs, MEIs) and no-called regions for each sample (excluding the cancer pairs). Note that these VCF files were created from the data generated using Analysis Pipeline 2.0, and so the annotations are not consistent with the current, updated Analysis Pipeline 2.2.
- Multigenome comparisons generated using the CGA Tools testvariants and mkvcf commands. These tools compare the presence of variants identified in the 69 genomes or a 54 genome subset of unrelated individuals. Small variant comparisons are provided in tabular and VCF format, while CNV comparisons are only provided in VCF format. These files are located in the Multigenome_summaries directory and are described in greater detail in the Multigenome_summaries_README.pdf document, located in the same directory.

Data Download Instructions

The Complete Genomics data files are provided as 'tar' format archives that can be un-archived using a Linux/Unix/MacOSX system's tar command line tool, or by using many other graphical utilities on various platforms (for example, WinRAR on Microsoft Windows). Each tar file associated with the same sample should be extracted starting from the same working directory, as when extracted they will each create and populate various contents of a subdirectory hierarchy starting at a folder named NA19240-200-37-ASM. Many of the contents of the tar files are compressed using bzip2 (.bz2 files), and thus the tar file itself is not compressed. Also available on the ftp server is a file containing the checksum results which can be helpful in ensuring the integrity of the download files (md5sum.txt).

Complete Genomics recommends using one of the following, tested, FTP clients for download.

1. Ncftp is a proven, well documented command-line ftp client. It can be downloaded from: http://www.ncftp.com/download. NcFTP supports the use of scripts to download very large files. NcFTP also provides precompiled binary distributions for several Operating Systems (e.g. Windows, Mac, Unix/Linux).
2. FileZilla is an FTP solution available as a client and server option. It can be downloaded from: http://filezilla-project.org/. FileZilla is a fast and reliable cross-platform FTP, FTPS and SFTP client with an intuitive graphical user interface.

These files should be downloaded in binary mode. If given the option select the binary option (web-based downloads from a compatible browser should automatically use binary mode).

Note that all Complete Genomics data files (when decompressed) use text formats formatted for use on Linux, Unix and MacOSX. Working with these files on a Windows machine may require format conversion because of the differences in line break characters used in standard Windows text files (CR-LF) vs. those on Unix (LF only). Whether conversion is needed will depend on the Windows application you are using: some Windows programs read Unix format files without conversion while other Windows software (example: Notepad) does not.

Getting Started with These Data

Detailed descriptions of the data file formats can be found in the Standard Sequencing Service Data File Formats for the 69 standard genomes and the Cancer Sequencing Service Data File Formats for the tumor-normal pairs. Output and file types are summarized in the Data Deliverable Service Note. We strongly encourage users to download these documents, along with the FAQs, which can be found at http://www.completegenomics.com/customer-support/support/.

Various analysis and visualization tools are useful for performing common analyses on Complete Genomics data sets. Complete Genomics recommends that researchers utilize CGA Tools for analyzing these data. For more information about CGA Tools, see http://www.completegenomics.com/sequence-data/cgatools/.

As an excellent starting point, the masterVarBeta files are designed to help researchers easily search for variations of interest and aid in the conversion of Complete Genomics data into other standard file formats. This file provides an overview of the types of information that Complete Genomics provides.

Citing the Public Genome Data

The data are freely available for use in a publication with the following stipulations:

1. The Coriell and ATCC Repository number(s) of the cell line(s) or the DNA sample(s) must be cited in publication or presentations that are based on the use of these materials.
2. Our Science paper (R. Drmanac, et. al. Science 327(5961), 78. [DOI: 10.1126/science.1181498]) must be referenced.
3. The version number of the Complete Genomics assembly software with which the data was generated must be cited. This can be found in the header of the summary.tsv file (# Software_Version).

Support

Please direct any questions to Complete Genomics Technical Support
support@completegenomics.com
Toll-free: 1-855-267-5383
International: 1-650-943-2600

For additional documentation please see http://www.completegenomics.com/customer-support/support/.

Legal Notice

www.completegenomics.com info@completegenomics.com
2071 Stierlin Court, Mountain View, CA 94043 USA Tel 650.943.2800

Published in U.S.A., August 2012, SNPG-03