Complete genomics

SERVICE NOTE

Service Deliverables

- Complete set of variants reported: SNPs, indels, CNVs, SVs, mobile element insertions, and lesser allele fractions
- Both somatic and germline variants provided
- Annotation of variants with known genes and ncRNAs
- Assembled sequence for each variant allele
- Reads, quality scores, and mappings
- Summary reports
- Analysis tools (CGA™ Tools)

Overview of Data Delivered

Introduction

This document contains an overview of the data delivered for each human genome sequenced by Complete Genomics. Two sequencing services are available: the Standard Sequencing Service for individual samples and the Cancer Sequencing Service for pairs and trios (Table 1). The Standard Sequencing Service delivers a comprehensive data set for each genome sequenced. Similarly, the Cancer Sequencing Service delivers the same comprehensive data set for each genome within a multi-genome comparison. Additionally, though, this service includes direct pairwise comparisons resulting in the identification of somatic events within a tumor as compared to its matched normal.

For all services, the complete data set is provided in convenient, text-based formats. The total size of the data set for a single genome at standard coverage (≥ 40X) is ~ 300 to 350 GB. The size of the data set for genomes sequenced at high coverage (≥ 80X) is twice the size, ~ 600 to 700 GB. Approximately 90% of these data sets are comprised of the initial reads and their mappings. The remaining 10% of the data (approx. 35 to 75 GB

	STANDARD SEQUENCING SERVICE	CANCER SEQUENCING SERVICE		
	Stand-alone Sample	Normal Sample	Tumor Sample	Somatic Events
VARIATIONS				
SNPs and indels	✓	✓	✓	✓
Small variant confidence scores	✓	✓	✓	✓
Copy number variations (CNVs)	✓	✓	✓	✓
Structural variations (SVs)	✓	✓	✓	✓
Mobile element insertions	✓	✓	✓	-
Lesser allele fractions (LAFs) For LOH and UPD events	✓	✓	✓	✓
ANNOTATIONS				
Annotations Includes genes, ncRNAs, dbSNP allele frequency, and more	✓	✓	✓	✓
SV event interpretation	✓	✓	✓	-
SUMMARIES				
Circos plot	✓	✓	✓	✓
VCF files	✓	✓	✓	✓
Genome QC summary report	✓	✓	✓	-
READ-LEVEL DATA				
Evidence folders Compliation of reads used to make variation calls	✓	✓	✓	✓
Raw reads and mappings*	✓	✓	✓	✓

Table 1. *Complete Genomics services and descriptions.*

per genome) is comprised of summary reports that include lists and annotations of all called variations as well as the reads and mappings supporting each variant call (Figure 1).

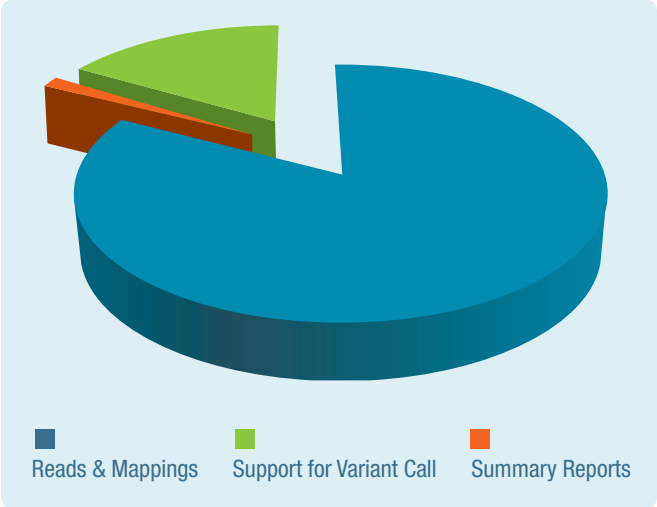

Figure 1. *Breakdown of the data delivered for a single genome.*

Summary Reports

Complete Genomics computes a variety of summary metrics for each sequenced genome (Table 2). These metrics, which include coverage and overall variation call statistics, provide an overview of the results and data quality. We also provide empirical measures of the insert sizes in each library, which is useful in the interpretation of the reads and initial mappings.

Assembly & Identified Variations

These data contain the primary results of the assembly and include the read-depth coverage for each position in the reference genome, a refScore which measures the confidence that the given position is homozygous reference, and all called variants where the assembled genome differs from the reference genome (Table 3). Supporting assemblies for each of the called variants allow for a detailed investigation of the evidence underlying each called allele. Variants are richly annotated with information about known variants in the dbSNP, COSMIC, and DGV databases. Functional genomic annotations for variants present in RefSeq and miRBase transcripts indicate predicted impacts on the protein (for example, frameshift, nonsynonymous, etc.). A summary of the counts of known and novel variants located in each transcript is also provided to enable researchers to rapidly scan for mutated or changed genes (for example, by function or pathway). Finally, comprehensive information describing variants called is provided in ***masterVar***, a tabular file which focuses primarily on small variants, and ***vcfBeta***, a VCF-format file that describes all variants (small variants, copy number variations (CNVs), structural variants (SVs), and mobile element insertions (MEIs)) in VCF format.

Specific to the Cancer Sequencing Service is the identification of variants present in the tumor but absent from its matched normal sample. These somatic events are supported by a statistical score that indicates the likelihood that the variant is present in the tumor only. The evidence reads underlying each somatic call are provided for both the tumor and the normal samples. To simplify comparison of the related genomes, small variant, CNV and SV calls for the paired samples are summarized in a single somaticVcf file.

Copy Number, Structural Variations, Mobile Element Insertions, and Lesser Allele Fractions

In addition to small variants (SNPs, substitutions, and small indels), Complete Genomics identifies larger events across every genome sequenced. CNVs are summarized in discrete and regular windows tiled across the genome, or alternately in segments according to ploidy breakpoints. LAF estimates, which can be used to identify regions of loss of heterozygosity (with or without copy number change, including uniparental disomy), are also summarized within the same windows used for CNV calculation. Structural variation reports list detected SV breakpoints and include event interpretation such as gene fusions, inversions, and deletions. Finally, transposable element insertions that are novel with respect to the reference genome are identified and reported in the MEI files (Table 4).

The Cancer Sequencing Service provides all of this data for each genome, and also includes somatic CNVs, somatic SVs, and somatic LAF estimations. Somatic CNVs and somatic LAFs are derived from a direct comparison of the tumor with its matched normal. Identification of somatic SVs is based on the selection of those events that are detected in the tumor and not in the matched normal. Further, when samples are submitted as pairs or

FILENAME	DESCRIPTION
summary-XXX.tsv	Summary statistics for genome sequence
coverage-XXX.tsv coverageCoding-XXX.tsv	Reports number of bases represented at a give unique and weight-sum sequence coverage depth for the whole genome and coding regions of the genome, respectively.
coverageByGcContent-XXX.tsv coverageByGcContentCoding-XXX.tsv	Reports normalized coverage for cumulative GC base content percentile for the whole genome and coding regions of the genome, respectively.
depthOfCoverage_100000-XXX.tsv	Reports unique and weight-sum sequence coverage, along with GC bias-corrected weight-sum coverage for every 100 kb window along the sequenced genome
indelLength-XXX.tsv indelLengthCoding-XXX.tsv	Reports length of each small insertion and deletion called in the whole genome and coding regions of the genome, respectively.
substitutionLength-XXX.tsv substitutionLengthCoding-XXX.tsv	Reports length of each substitution called in the whole genome and coding regions of the genome, respectively.
lib_* _XXXX.tsv	Library file
circos-XXX.html circos-XXX.png somaticCircos-XXX.html somaticCircos-XXX.png	Circos plot including visualization of SNP calls, interchromosomal junctions, called level or called ploidy, and Low Allele Frequency (LAF). Somatic version is for tumors in paired analyses only, filtered for somatic events only, plus loss of heterozygosity (LOH) and more.

Table 2. *Reports of summary metrics provided.*

FILENAME	DESCRIPTION
var-XXX.tsv.bz2	Called sequence with respect to the reference genome, indicating variant and non-variant regions
masterVarBeta-XXX.tsv.bz2	Called sequence indicating variant and non-variant regions along with annotations in a one line per locus format. For the Cancer Sequencing Service only, read counts for the related genomes are provided for each called sequence.
gene-XXX.tsv.bz2	Annotations of variations in known protein coding gene sequences
geneVarSummary-XXX.tsv	Summary of variations in known protein coding gene sequences
dbSNPAnnotated-XXX.tsv.bz2	Calls at dbSNP loci
ncRNA-XXX.tsv.bz2	Variants that fall within mature microRNAs and pre-microRNAs identified in the miRBase Sequence database
coverageRefScore-XXX.tsv	Base-level coverage and scores
evidenceIntervals-XXX.tsv.bz2	Contains the assembled sequence for each variant allele, including alleles unique to the alternate sample in the case of tumor-normal comparisons
evidenceDnbs-XXX.tsv.bz2	Contains the supporting reads for each assembled sequence
correlation-XXX.tsv.bz2	Correlations between assemblies that share supporting reads, for example duplicated regions
reads-XXX.tsv.bz2	Reads and base-level quality scores
mapping-XXX.tsv.bz2	Initial (pre-assembly) mappings of reads
vcfBeta-XXX.vcf.bz2 somaticVcfBeta-XXX-N1.VCF.bz2	Comprehensive information describing all variant calls (small variants, CNVs, SVs, MEIs, LAF) for a single genome assembly in VCF format. Somatic version summarizes both tumor and normal within a pair for small variants, CNVs, SVs, and LAF.

Table 3. *Variants, annotation, evidence, reads, and mapping files provided.*

trios as part of the Cancer Sequencing Service, all CNVs and SVs for each tumor-normal pair are captured in a single VCF file that also contains the small variant calls for those samples.

Reads & Mapping Data

(Only provided with part numbers containing RM or RNM) Complete Genomics provides the individual reads, quality scores, and initial mappings to the reference genome (currently NCBI Build 36 or GRCh37). Reads and mappings are organized by lane. Each base call is from a

FILENAME	DESCRIPTION
cnvSegmentsDiploidBeta-XXX.tsv somaticCnvSegmentsDiploidBeta-XXX.tsv	Segmentation of the complete reference genome into regions of distinct ploidy levels. Somatic version is for tumors in paired analyses only.
cnvDetailsDiploidBeta-XXX.tsv.bz2 somaticCnvDetailsDiploidBeta-XXX.tsv.bz2	Information on the estimated ploidy and average coverage for every 2 kb along the genome. Somatic version is for tumors in paired analyses only.
depthOfCoverage_100000-XXX.tsv	Reports unique and weight-sum sequence coverage, along with GC bias-corrected weightsum coverage and baseline normalized coverage for every non-overlapping 100 kb window along the genome
cnvSegmentsNondiploidBeta-XXX.tsv somaticCnvSegmentsNondiploidBeta-XXX.tsv	Segmentation of the complete reference genome into regions of discrete coverage levels for nondiploid genomes. Somatic version is for tumors in paired analyses only.
cnvDetailsNondiploidBeta-XXX.tsv somaticCnvDetailsNondiploidBeta-XXX.tsv	Information on the estimated ploidy in nondiploid samples and average coverage for every 100 kb along the genome. Somatic version is for tumors in paired analyses only.
allJunctionsBeta-XXX.tsv somaticAllJunctionsBeta-XXX.tsv	Information for detected junctions represented by 3 or more discordant mate pairs. Somatic version is for tumors in paired analyses only.
highConfidenceJunctionsBeta-XXX.tsv somaticHighConfidenceJunctionsBeta-XXX.tsv	Filtered subset of the junctions reported in the allJunctions file. Somatic version is for tumors in paired analyses only.
evidenceJunctionDnbsBeta-XXX.tsv.bz2	Alignments of the individual discordant mate pairs supporting each junction
evidenceJunctionClustersBeta- XXX.tsv	Information and annotation of detected junctions
mobileElementInsertionsBeta-XXX.tsv	Information on detected mobile element insertion events
mobileElementInsertionsROCBeta-XXX.png	Graph of the trade-off between sensitivity and specificity of MEI detection for various score cutoffs
mobileElementInsertionsRefCountsBeta-XXX.png	Graph that shows the distribution of DNB counts that support reference allele for MEI events

Table 4. *CNV, SV, and MEI files provided.*

single DNB, and its quality score is a Phred-like transformation of the error probability associated with it. These data comprise the bulk of the data deliverable.

Analysis Tools

Complete Genomics provides an open source tools package, Complete Genomics Analysis Tools (CGA™ Tools), for downstream analysis of Complete Genomics data (http://www.completegenomics.com/sequence-data/cgatools/). Currently, the general areas of functionality include genome comparison tools, format conversion tools, annotation, and reference tools.

Documentation

Complete Genomics provides detailed documentation that describes the data file formats and structure along with examples and interpretation. Detailed documentation is available on our website at www.completegenomics.com/customer-support.

Summary

Complete Genomics provides the most comprehensive, high-quality data sets for whole human genome sequencing; including reports on summary statistics and variant calls in addition to the underlying reads, quality scores and mappings. Complete Genomics brings an entirely new level of efficiency and affordability to researchers conducting large-scale human genome studies focused on the elucidation of the genetic underpinnings of complex diseases.

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

Complete genomics

Copyright© 2012 Complete Genomics, Inc. All rights reserved. Complete Genomics and the Complete Genomics logo are trademarks of Complete Genomics, Inc. All other brands and product names are trademarks or registered trademarks of their respective holders.

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject.
support@completegenomics.com Toll-free: 1-855-CMPLETE (1-855-267-5383) or 1-650-943-2600
Information, descriptions and specifications in this publication are subject to change without notice.

Published in U.S.A., November 2012, SN_DD-04