Complete Genomics Service Frequently Asked Questions (FAQ)

Updated December 2012

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject. Information, descriptions and specifications in this publication are subject to change without notice.

General Information

Who is Complete Genomics?

Complete Genomics, Inc. is a leading commercial provider of complete human genome sequencing services. Our sequencing service provides high-coverage and high-accuracy results at an affordable price. We do not sell instruments or reagent kits; rather we provide a service that includes sample quality control, library construction, complete genome DNA sequencing, and bioinformatics analysis of human DNA samples. For more information, please contact us at info@completegenomics.com or visit our website at www.completegenomics.com.

Where can I learn more about Complete Genomics sequencing?

There is a high-level Technology White Paper and additional information posted at www.completegenomics.com. A publication in the journal *Science* authored by Complete Genomics scientists and collaborators (Drmanac et al. *Science* 2010; *Science* Express 2009) describes our process and also reviews results from three reference genomes. You can access the paper at www.drmanac.com at no charge. A publication in the Journal of Computational Biology authored by Complete Genomics bioinformaticians (Carnevali et al. J Comp Bio 2011) describes our original variant calling pipeline. Updates since this earlier version include the introduction of allele fraction calling and indel rescoring. Current details on variant calling are captured in the Small Variants Methods document as well as other Methods documents available on our website.

Where can I get sample Complete Genomics data and what is available?

Complete Genomics has made several complete human genome data sets available on its FTP server (ftp2.completegenomics.com). The genomes were sequenced at the Complete Genomics commercial genome sequencing center in Mountain View, California as part of our Complete Genomics Analysis Service (CGA™ Service). These data are largely consistent with the quality and attributes of data provided to Complete Genomics customers.

These data sets include 69 genomes representing the output from the Standard Sequencing Service, including trios, a large pedigree, and a diverse set of samples from nine different populations. Collections were drawn from the Coriell Institute for Medical Research.

There are also four samples representing the output from the Cancer Sequencing Service, including two tumor-normal cell line pairs. These collections were drawn from ATCC.

For additional information about this dataset, and access for free download, please visit www.completegenomics.com/public-data. For additional information, contact support@completegenomics.com.

Where do I get technical support for data sets and tools produced by Complete Genomics?

Please contact our Technical Support Team.

What is the turnaround time for the Complete Genomics sequencing service?

Complete Genomics quotes the turnaround time at 90 to 120 days. In late 2010, we delivered data to customers with an average turnaround time of 83 days. In 2011, the median turnaround time was just 68 days. Complete Genomics continues to focus on driving this number down.

Sequencing

What are the input sample requirements?

Complete Genomics recommends ≥ 5 µg and requires 3.5 µg (based on quantity measurements performed by Complete at the time of sample QC) of high molecular-weight double-stranded DNA (majority over 20 kb). Samples must be at a concentration of 30 to 300ng/µl, with a volume of 50 to 200 µl and in TE, pH 8.0. Note that because there is inherent variability in quantitation between sites and users, targeting 3.5 µg for sample submission could result in a significant number of samples failing to meet sample acceptance criteria, resulting in a delay in sample processing. To ensure that samples are processed efficiently, Complete strongly encourages customers to send ≥ 5 µg when available. Currently, whole genome amplified (WGA) DNA or formalin fixed paraffin embedded (FFPE) samples are not supported. DNA should be quantified by a Pico-green assay (preferably with the Quant-iT™ PicoGreen® dsDNA kit from Invitrogen). Spectrophotometric quantification (by optical density) is not recommended, as contaminating protein and RNA may result in inaccurate estimation of concentration. DNA concentration should be measured after diluting to the range specified above to improve the accuracy of DNA quantification, minimizing the number of sample failures due to insufficient DNA. A detailed protocol for how PicoGreen quantitation is performed at Complete Genomics is available in the Sample Quality Control Protocol. The minimum number of samples the Complete Genomics accepts is eight. Complete Genomics is actively working on additional protocols to support smaller sample DNA amounts.

Refer to the Sample Submission Instructions for details on what types of samples are accepted for sequencing and for further details and guidelines.

What sequencing technology does Complete Genomics use?

Complete Genomics' sequencing platform employs high-density DNA nanoarrays that are populated with DNA nanoballs (DNBs™). Base identification is performed using an unchained ligation-based read technology known as combinatorial probe-anchor ligation (cPAL™). The sequencing instrumentation is custom-developed to support this process. Details are described in our Technology White Paper and in our *Science* publication (Drmanac et al., 2010). See "Where can I learn more about Complete Genomics sequencing?"

Does Complete Genomics perform paired-end or mate-pair sequencing? What is the gap size between the reads? What are the implications of this?

DNBs are a mate-pair construct. We target ~400-500 bp inserts to maximize the power of the data (a) for assembly through many duplications and repeats (most particularly, Alu elements, which are numerous), and (b) for identification of structural variants and larger indels. The actual gap insert size in any specific library can be empirically measured from the mapping results, and (as of version 1.7 of the Complete Genomics Analysis Pipeline) we provide such a distribution with each genome. Useful metrics, such as mean mate gap estimated for the library and the 95% confidence interval for the mate gap distribution range, are reported in the ***summary-ASM-ID].tsv*** file in the ASM directory.

Some mate-pair library protocols have been known to generate biased or low-complexity libraries. Complete Genomics' recent laboratory protocols have been extensively tuned to reduce bias (for example, as a function of AT/GC ratio), and we achieve very high complexity (as measured by low duplication rates) using these methods. As a result, we have been able to provide thousands of a genomes with a median genome call rate > 96% and exome call rate > 98% in the first three quarters of 2012. Additional sequencing metrics that quantify library performance are present in the summary files provided with each genome.

Can I order a phased genome, using Long Fragment Read (LFR) technology?

Complete Genomics is currently working on commercializing the LFR technology, and plans to offer whole genome sequencing with phasing based on LFR in 2013.

Assembly, Mapping, and Variant Calling

How does Complete Genomics map reads and call variants?

Reads are initially mapped to the reference genome using a fast algorithm, and these initial mappings are both expanded and refined by a form of local *de novo* assembly, which is applied to all regions of the genome that appear to contain variants (SNPs, indels, and block substitutions) based on the initial mappings. The *de novo* assembly leverages mate-pair information, which allows reads to be recruited into variant calling with higher sensitivity than genome-wide mapping methods provide. Assemblies are diploid, and thus we produce two separate result sequences for each locus in diploid regions. Variants are called by independently comparing each of the diploid assemblies to the reference. The process is described in our *Science* paper (Drmanac et al. *Science*, Jan 2010, www.drmanac.com), and our assembly algorithms are described in detail in our publication in the Journal of Computational Biology (Carnevali et al, J Comp Bio 2011).

Copy number variable (CNV) regions are called based on depth-of-coverage analysis. Sequence coverage is averaged and corrected for GC bias over a fixed window and normalized relative to a set of standard genomes. In the case of a tumor-normal comparative analysis provided through our Cancer Sequencing Service, coverage in the tumor genome is normalized to coverage for the same region in the matched normal genome. A hidden Markov model (HMM) is used to classify segments of the genome as having 0, 1, 2, 3 copies…up to a maximum value.

Structural variations (SVs) are detected by analyzing DNB mappings found during the standard assembly process described above and identifying clusters of DNBs in which each arm maps uniquely to the reference genome, but with an unexpected mate pair length or anomalous orientation. Local *de novo* assembly is applied to refine junction breakpoints and resolve the transition sequence. Novel insertions of mobile elements into the sequenced genome are identified as clusters of reads that uniquely map to the reference genome with one arm and to ubiquitous sequence with the other arm. The location, type, and orientation of the inserted elements are identified using mate pairs that map in the vicinity of the insertion site, aligning each unmapped arm to sequences of a defined set of mobile elements. The process for CNV and SV detection is described in more detail in *Complete Genomics Data File Formats*.

What type of events does Complete Genomics call?

Complete Genomics identifies small variants, including SNPs, indels, and block substitutions, as well as copy number variations (CNVs), structural variations (SVs), and mobile element insertions (MEIs). For the Standard Sequencing Service, all of these variation types are identified in comparison to the human genome reference. For the Cancer Sequencing Service, somatic small variants, CNVs, and SVs are each also identified in comparison to the baseline sample within a pair or trio.

Please explain the gaps within the reads. What are the implications of these gaps on mapping, assembly, and variant calling?

The Complete Genomics read structure is described in our *Science* paper (Drmanac et al. 2010) and in the Complete Genomics technology whitepaper (available at www.completegenomics.com).

By contrast with some other sequencing technologies, which have a high rate of within-read indels (most single molecule sequencing and pyrosequencing-based methods have this attribute), the intra-read gaps in Complete Genomics data are relatively easy to handle. First, they always occur at, and only at, precisely known locations in each paired end. Secondly, the gaps sizes are highly predictable and generally only +/- 1 base pair from the known mid-value. Thus, algorithms can readily be designed to map, assemble, and call variants in these reads. Complete Genomics' analysis methods for these gapped reads have been shown to produce high quality results for both SNPs and indels variant calls.

Because of the gaps, coverage for comparable detection power does need to be modestly higher than if the reads had no gaps. However, this coverage requirement is balanced by the improved base-call accuracy (and consistent accuracy over the length of the read) in Complete Genomics sequences, improving the power of the data on a per base-call basis. This accuracy is enabled by the gapped construct that provides multiple sequencing reaction priming sites in each arm of each DNB. Because the sequencing of DNBs is highly cost-effective, Complete Genomics can also generate very deep coverage of these reads and thus produce high-quality variant calls over a large fraction of the genome.

What do you mean by a “called” base or locus?

We use stringent thresholds in our variant-calling algorithms that take into account base-call accuracy, mis-mapping probability, and both quantity and consistency of evidence. A fully called position is one where we have determined the full diploid sequence (that is, we have assembled both alleles) at these thresholds. By contrast with some other pipelines, Complete Genomics’ data analysis methods are careful to distinguish regions of the genome that are confidently called homozygous reference from those which are no-called. This greatly facilitates comparison between genomes by reducing false negatives.

For clarification, when we measure a percentage of the genome called, we are referring to a percentage of the bases corresponding to the complete NCBI reference genome sequence. We are not referring to a fraction of the non-repetitive or non-degenerate genome, or to a fraction of the genome within a certain AT/GC range.

If Complete Genomics adds a new feature to its pipeline and I wish to have my data reprocessed, can I?

Customers with genomes processed by Assembly Pipeline version 1.5.0 or later can order re-analysis of these genomes using Assembly Pipeline version 1.10 and later. Customers have the option to indicate whether they prefer a specific version or would prefer reanalysis on the most current Assembly Pipeline version at the time of processing. Since Complete Genomics does not retain customer data, the complete and original data set must be shipped back to Complete Genomics via hard disk drive. For more information, see the Reanalysis Flyer, or contact us at info@completegenomics.com.

Can I get a copy of Complete Genomics’ data processing pipeline to run on my computers?

Complete Genomics’ data processing software is not distributed at this time.

Data Results

What data/results does Complete Genomics provide with each genome?

Complete Genomics provides a complete set of raw and processed results with each genome. As of Complete Genomics Assembly Pipeline version 2.4 each genome includes the following:

Short name	File name(s)	Content	Location
Summary file	summary-*	Summary statistics	ASM
VCF	vcfBeta-*	All called small variants, CNVs, SVs, and MEIs in VCF format	ASM
Variant file	var-*	All called small variants	ASM
Master Variation file	masterVarBeta-*	All called small variants and associated annotations	ASM
Gene annotations	gene-*	Small variants in genes	ASM
dbSNP annotations	dbSNPAnnotated-*	Sequence for all dbSNP sites	ASM
Gene summary	geneVarSummary-*	Summary of small variants in genes	ASM
Non-coding RNA annotations	ncRNA-*	All small variants in known microRNAs	ASM
CNV files	cnvSegmentsDiploidBeta-*, cnvSegmentsNondiploidBeta-*, cnvDetailsDiploidBeta-*, cnvDetailsNondiploidBeta-*, and depthOfCoverage_100000-*	Copy number segmentation and Lesser Allele Fraction estimates	ASM/CNV
SV files	allJunctionsBeta-*, highConfidenceJunctionsBeta-*, evidenceJunctionDnbBeta-*, and evidenceJunctionClustersBeta-*, allSvEventsBeta-*, highConfidenceSvEventsBeta-*	All called junctions, structural variation events, and alignments of evidence DNBs	ASM/SV
MEI files	mobileElementInsertionsBeta-*, mobileElementInsertionsROCBeta-*, mobileElementInsertionsRefCountsBeta-*	All called mobile element insertion events detected	ASM/MEI
Evidence files	evidenceIntervals-*, evidenceDnbs-*, and correlation-*	Assemblies of small variant loci and aligned reads by chromosome	ASM/ EVIDENCE
Coverage and Reference Score	coverageRefScore-*	Values for each position in the reference genome by chromosome	ASM/REF
Report files	circos-*, circosLegend-* coverage-*, coverageByGcContent-*, coverageCoding-*, coverageByGcContentCoding-*, indelLength-*, and substitutionLength-*	Coverage quality characteristics for whole genome and exome, and length distribution of called indels and substitutions	ASM/ REPORTS
Library info	lib-*	Distribution of mate-pair gap sizes	LIB
Reads	reads-*	Sequences and base quality scores	MAP/*
Mapping	mapping-*	Results of initial mapping of reads to reference	MAP/*

Genome analysis for samples submitted to the Cancer Sequencing Service includes the following additional files:

Short name	File name(s)	Content	Location
idMap files	idMap-*	A mapping among various identifiers of a sample in a multi-genome dataset	Individual Genome Directory
Somatic VCF	somaticVcfBeta-*	Small variants, CNVs, and SVs detected in both samples within a tumor-normal pair in VCF format	ASM
CNV files	somaticCnvSegmentsDiploidBeta-*, somaticCnvSegmentsNondiploidBeta-*, somaticCnvDetailsDiploidBeta-*, and somaticCnvDetailsNondiploidBeta-*	Copy number segmentation and Lesser Allele Fraction estimates unique to the tumor in a tumor-normal comparison	ASM/CNV
SV files	somaticAllJunctionsBeta-*, somaticHighConfidenceJunctionsBeta-*	All called junctions and alignments of evidence DNBs unique to the tumor in a tumor-normal comparison	ASM/SV
Evidence files	evidenceIntervals-[related sample]-*, and evidenceDnbs-[related sample]-*	Assemblies of small variant loci and aligned reads by chromosome for the related sample in a tumor-normal pair	ASM/ EVIDENCE-<comparison _ASM-ID>
Report files	somaticCircos-*, and somaticCircosLegend-*	Circos plot and legend for somatic events in the tumor	ASM/ REPORTS

For more information on the delivered data set, see the *Complete Genomics Overview of Data Delivered*.

Does Complete Genomics retain customer data after it has been delivered to a customer?

Complete Genomics retains data for not less than thirty days after delivery to a customer, but ***deletes*** the data thereafter. After receiving data from Complete Genomics, customers are strongly advised to confirm immediately that the files are valid and to create a backup copy.

How big is the data sent by Complete Genomics for each genome? Are the data compressed?

A single genome at standard coverage (40x) is approximately 300 to 350 GB, although the data set may be larger. Genomes at higher coverage (80x) are approximately double (600 to 700 GB). A tumor-normal pair is therefore approximately 1.2 to 1.4 TB, and a trio is approximately 1.8 to 2.1 TB. About ninety percent of this volume is used by the reads and initial mappings, while the processed data comprises the remaining ten percent. For part numbers providing variations only (no reads and mappings), the data set is approximately 35 GB to 60 GB per genome, depending on coverage. Samples submitted to the Cancer Sequencing Service will have additional data in the Evidence directories, resulting in data sets of approximately 75 GB per genome for variations only.

For more information on the delivered data set, see *Complete Genomics Overview of Data Delivered*.

Most of the files are shipped compressed. Uncompressing all of the data files will increase the required storage for a single genome approximately 3- to 4-fold (for example, to over 1.5 TB). Decompression is not required for compatibility with Complete Genomics downstream analysis tool package, CGA Tools. For these reasons, most of Complete Genomics' customers leave many of these files in their compressed format.

What data formats does Complete Genomics use?

All Complete Genomics data are provided as text files that can be examined and further analyzed using many different tools on all standard computing systems.

Many of Complete Genomics' text data file formats are specific to our platform and provide rich descriptions of the data we generate. These files are also optimized for information density and to keep file sizes as manageable as possible. In addition to platform-specific files, both Standard and Cancer Sequencing Services provide variant calls in VCF 4.1 format.

Complete Genomics has an open source tools package, Complete Genomics Analysis Tools (CGA Tools), for downstream analysis of Complete Genomics data. Currently, CGA Tools contains file format converters to transform the Complete Genomics data to other data formats, such as SAM/BAM and VCF.

Do I get the individual reads? Can I re-map or re-assemble them using some other software?

Customers receive a complete read data set unless they have ordered variant-only services. Read-level data includes all reads and mappings, as well as Phred-scale base quality scores and other useful related information such as library gap size distributions.

Complete Genomics is not presently aware of any broadly released programs optimized to handle the unique aspects of Complete Genomics read data, such as the intra-read gap structure. We have found that mapping and assembly programs such as MAQ or Velvet, which are well optimized for other data types, will not produce satisfactory results with Complete Genomics data.

Can I get the reads in FASTA or FASTQ format?

Complete Genomics does not provide a translator, but customers could write one quite easily. However, please thoroughly consider the response to "Do I get the individual reads? Can I re-map or re-assemble them using some other software?" and "Can I call variants from mapped Complete Genomics reads using some other program?" before doing so.

Can I get the mappings in some other format, like SAM/BAM?

Complete Genomics has an open source tools package, Complete Genomics Analysis Tools (CGA Tools), for downstream analysis of Complete Genomics data. Currently, CGA Tools contains file format converters to transform the Complete Genomics data to other data formats (such as SAM/BAM and VCF). However, please thoroughly consider the response to "Can I call variants from mapped Complete Genomics reads using some other program?" before doing so.

How accurate are the individual reads? How does accuracy change over the length of a read?

We have examined a number of our data sets in detail and found that the highest scoring 85% of all raw base-calls in uniquely mapped reads are >99.5% concordant with reference (corresponding to a Phred score >23). We also find that our calibrated Phred-scale quality scores are excellent predictors of base-call accuracy. It is important to note that this low discordance rate is achieved with **no** additional filtering of raw reads. Note also that the small number of discordances at these higher quality scores include not only sequencing errors but also real polymorphisms.

Because of the unique aspects of our sequencing chemistry, our read accuracy does not degrade over the length of a read, and the error profile is relatively flat. There are modest fluctuations in accuracy of some positions over others owing to the different oligonucleotides used in each ligation. Our algorithms measure this position-specific discordance rate and use it as a prior on error rate in variant calling.

What is the coverage provided?

Complete Genomics offers two coverage levels. For the standard-coverage products, Complete Genomics generates ≥ 120 GB of reads mappable to the reference genome, providing an average coverage of ≥ 40X across the reference genome. Furthermore, Complete Genomics provides ≥ 90% completeness defined as making a diploid call (i.e., both alleles) at unique loci in over 90% of the reference genome. We believe this high level of coverage provides excellent accuracy for calls over the vast majority of the genome of any sample. On genomes to date we have typically well exceeded these metrics. These metrics are reported for each sequenced genome in an output file (***summary-[ASM-ID].tsv***) that is provided to customers.

For the high-coverage products, the coverage level is doubled, generating ≥ 240 GB of reads mappable to the reference genome, providing an average coverage of ≥ 80X across the reference genome. The additional coverage is useful for increased sensitivity, particularly for heterogeneous samples such as tumors. In the case of saliva samples, the amount of sequencing performed is equivalent to the high coverage products, but because of the bacterial DNA also present and also sequenced, the mapping rate is lower. For saliva samples, Complete Genomics guarantees an average coverage of ≥ 50X across the reference genome. Samples with low bacterial load will generally receive significantly higher coverage. Samples with high bacterial load will receive additional free sequencing to ensure that ≥ 50X coverage is attained for the samples.

What is the read length? What coverage of the genome does this allow?

We sequence 70 bases per DNB, 35 from each end. At the high level and uniformity of base-call accuracy we achieve, a 35-base read has equivalent mapping power of somewhat longer reads from other methods. Perhaps more importantly, the vast majority of mapped base calls contribute significantly to variant detection (such as SNP calling), by comparison with other technologies where accuracy drops off significantly along the length of the read.

We have a variety of data, both from actual assemblies and simulation studies, which show that about 96% of the reference human genome is addressable using this library and sequencing strategy, including a significant fraction of the high-copy repeat sequences in the genome. The remaining 4% includes degenerate regions and larger, highly conserved sequences which are difficult to access using most sequencing methods.

What is the stated accuracy in a Complete Genomics data set?

We have multiple data points regarding accuracy from validation studies (comparing Complete Genomics data to other laboratory methods), technical replicates, and family studies (using Mendelian constraints to measure errors). These data suggest that Complete Genomics fully calls approximately 97% of the reference genome (and 98% of the exome) with SNP false positive and false negative rates of 1.56×10^{-6} and 1.67×10^{-6}, respectively. The net Mendelian Inheritance Error concordance of all small variant call types (SNPs, indels, and substitutions) was observed to be 99.99971% in called non-repetitive bases and 99.99947% genome-wide. For more information, contact your sales representative or customer support for a copy of Complete Genomics Accuracy White Paper.

Using the Results

What bioinformatics skills and IT infrastructure do I need to work with Complete Genomics data?

Many current Complete Genomics users have had excellent scientific success by studying only the processed results provided by Complete Genomics, in particular, the called variants and their annotations. These customers find they do not require highly specialized bioinformatics skills (such as in genome assembly) nor expensive high-end compute clusters to work with the data. Many sophisticated analyses of these data can be done on high-end desktop and mid-range servers with access to enough disk storage required to keep the data. However even the processed data in the variant and annotation files are large, and these files can be difficult to work with using many desktop software programs. Most notably, this applies to Microsoft Excel, where even the most recent versions of Excel have a 1 million-row limit. Also, visually inspecting even a fraction of the variant calls in any genome can be daunting.

Since there are important bioinformatics considerations as well as logistical issues when interpreting any large genomic data set (including those produced by Complete Genomics) we often recommend that customers have access to at least one individual with good bioinformatics skills, including basic programming (PERL or Python scripting is common), and access to a Unix/Linux environment. This person should have technical experience manipulating large data sets and have a good scientific understanding of genetics, genomic sequence, and genome-annotation databases.

Do I need a data processing pipeline for mapping, assembly, or variant detection to work with Complete Genomics data?

No. Mapping, assembly, variant detection, and annotation are performed by Complete Genomics and are included in the data set provided to customers.

If I am just using the variant files and other processed output, can I get rid of the reads and initial mappings? At least, can I keep them off my computer?

It is up to you to determine which data you need to archive, but keep in mind that Complete Genomics ***deletes*** customer data a short time after delivery to you, 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 on 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.

The ST001V and SHC001VAR part number options offer delivery of all variant files and other processed output, without including the reads and initial mappings, for customers that do not intend to use or store the raw data at all.

Can I call variants from mapped Complete Genomics reads using some other program?

Yes, however the results will differ from those that the Complete Genomics pipeline generates, and customers should be cautious, as the results may be far less accurate. We are not aware of any broadly released variant-calling tools optimized for Complete Genomics data.

Complete Genomics' assembly and variation calling methods have been tuned to various aspects of Complete Genomics' data, such as the flat error profile, the presence of specific length intra-read gaps, and the properties of the analytical process we have chosen. Because of the division of labor between our mapping and assembly processes, our initial mappings have a somewhat different character than mappings often produced for other platforms. For example, traditional SNP calling directly from these initial alignments tends to produce far less satisfactory results than our local *de novo* assembly approach.

Where can I get tools for further processing or visualizing Complete Genomics data?

Complete Genomics has an open source tools package, Complete Genomics Analysis Tools (CGA Tools), for downstream analysis of Complete Genomics data. CGA Tools focuses primarily on variant comparison and file format conversion tools and is available for Linux, Mac OS X, and the Galaxy platform. Complete Genomics software partners provide additional tools and solutions to complement CGA Tools. For more information about our software partnerships, see: www.completegenomics.com/services/partners/.

Additional tools are also available at the Complete Genomics User Community: community.completegenomics.com. The Tool Repository at this site includes scripts and programs that have been written by Complete Genomics members, but are not formal product offerings and, as such, are not fully supported by Complete Genomics.

Cancer Samples

How does the Cancer Sequencing Service differ from the Standard Sequencing Service?

The Standard Sequencing Service supports whole genome sequencing and data delivery for individual genomes. The Cancer Sequencing Service supports whole genome sequencing and data delivery for genome pairs and genome trios. Genome pairs consist of a tumor genome with a matched normal, while genome trios consist of two tumor genomes from the same patient along with the matched normal genome for those two tumors. Pairs and trios (also referred to as Sample Groups) are tracked and coordinated throughout QC, processing, sequencing, assembly, and delivery. Data delivered includes the full data set for each genome within the Sample Group as well as results from paired analysis between the tumor and the matched normal genome samples.

Can I get deeper coverage for my tumor samples?

Yes! Complete Genomics offers complete human genome sequencing at two coverage levels. For standard coverage, Complete Genomics guarantees a minimum average of 40x coverage across the complete genome. For high-coverage genomes, the number of sequencing lanes applied to sequencing each genome is doubled, with a guarantee of a minimum average of 80x coverage across the complete genome. It is common for cancer researchers to apply high coverage to tumor samples to mitigate some of the challenges introduced by heterogeneity (the presence of contaminating DNA from multiple genomes within a sample) and gross aneuploidy (widespread copy-number changes). The high coverage option is recommended for researchers working with samples known or expected to exhibit heterogeneity or gross aneuploidy.

Can a tumor sample be submitted using the Standard Sequencing Service rather than the Cancer Sequencing Service?

Yes. Unpaired tumors should be submitted using the Standard Sequencing Service, as they represent individual genomes. Tumor-normal genome pairs could be submitted separately using the Standard Sequencing Service, but there is no benefit to doing this, and there will be no paired analysis provided (i.e., no identification of somatic events). It is recommended that tumor-normal pairs (or trios) be submitted using the Cancer Sequencing Service to enable the detection of somatic events specific to the tumor samples.

Can a sample set containing only tumors or only non-tumors be submitted using the Cancer Sequencing Service rather than the Standard Sequencing Service?

There is no restriction on the types of samples that are submitted using either product type, but it is important to understand some caveats to submitting tumor samples without matched normal samples using the Cancer Sequencing Service. These include the following:

- Somatic output, summarizing small variants, copy number variation, and structural variations, is unidirectional. It is produced comparing the non-baseline sample to baseline sample only. A comparison in the reverse direction is not performed.
- CNV calling will work best when the baseline genomes are diploid/euploid.

Can I get the same paired analysis provided by the Cancer Sequencing Service using CGA Tools?

CGA Tools supports some the paired analyses provided in the Cancer Genome Sequencing Service. For example, calldiff and junctiondiff enable the identification of somatic small variants and structural variations. Not all of the analyses provided in the Cancer Genome Sequencing Service are reproduced by CGA Tools. Refer to the Genome Comparison Tools listed on the Complete Genomics website for more information: www.completegenomics.com/sequence-data/cgatools.

How are Sample Groups treated differently than individual samples throughout the complete workflow?

Samples submitted as a group under the Cancer Sequencing Service are treated as a Sample Group. The workflow and output for samples submitted as a Sample Group differs slightly from samples submitted as individual genomes under the Standard Sequencing Service as follows:

- Sample processing is coordinated at sample acceptance, sequencing, assembly, and delivery.
- Sample QC involves a confirmation that the samples within a group are in fact related, using a panel of 96 SNPs.
- Sequencing results for samples within a group are included in and influence the output for samples within a Sample Group. Specifically:
    - Paired output is provided for each pair, in which one sample is compared to the paired baseline to identify the somatic events specific to the sample. Somatic output includes somatic small variants, somatic CNVs, somatic structural variations, and a somatic Circos plot. A VCF file is provided including variants from both samples within a pair.
    - Allele read counts for paired samples are included in the ***masterVar*** file for each given sample.
    - Evidence at a given locus (mapped reads) is provided not only wherever a variant is found but also wherever a variant is found in the paired sample, even if no variant was found in the specific sample.

What if my Sample Group contains more than three samples, or if I want additional comparisons within my group?

For Sample Groups that contain greater than three samples or that require additional sample comparisons, submit all samples within the Sample Group as a combination of pairs, trios, and/or individuals, depending on sample numbers. For the additional comparisons desired that are not accomplished through the assignment of pairs and trios in the first submission, new sample comparisons can be performed by choosing one of the following options:

- CGA Tools comparisons: Complete Genomics provides analysis tools (CGA Tools) that identify and score somatic small variants and identify somatic SVs. Identifying somatic CNVs by directly comparing the CNV output for each sample is also often performed by customers. Please contact support for more information.
- Reanalysis: Sample groups can be submitted to our Professional Services group for automatic reanalysis immediately after sequencing so that the additional desired paired analyses are delivered shortly after the primary assemblies. If this is the desired route, please inform your Complete Genomics sales representative and Project Manager.