Release Notes for Genomes Processed Using Complete Genomics Software

Software Version 2.5

Related Documents

Consult the *Data File Formats* corresponding to any data set for information specific to those data. In addition, you can check the header information in the files to determine which version of Complete Genomics documentation applies.

Changes to Version 2.5

New Features and Enhancements

The following new features and enhancements are provided in this release:

1. Pipeline 2.5 supports Complete Genomics version 2 library generation process. The reads generated from these libraries have a smaller number of intra-read gaps. Positive intra-read gaps have been eliminated. See the Analysis Pipeline 2.5 *Data File Formats* for more information.
2. To support the changes to library and read structure, the Analysis Pipeline uses new CNV and SV baselines and replicate score calibration data. There are several consequences of these changes:
   - The baseline used to normalize coverage prior to making CNV calls was tuned to reduce coverage bias in the assemblies generated using library version 2. This change will affect the normalization factor calculated for a given region of the genome and can affect estimation of copy number and coverage level. The locations of invariant and hypervariable regions may also differ between Analysis Pipeline 2.5 and earlier pipelines.
   - The presence of interchromosomal junctions and the frequency at which a junction is observed in the baseline differs between Analysis Pipeline 2.5 and previous pipelines.
   - There are very minor changes to the *refScore* and *somaticScore* values as a consequence of updating the variant score calibration data.
3. For insertions, the VQLOW cutoff in Analysis Pipeline 2.5 provides higher sensitivity at the cost of slightly lower specificity. For SNPs, deletions, and substitutions, there is an increase in sensitivity with no change in specificity.

   The sensitivity and specificity of the *somaticScore*, the SQLOW cutoff, and the FET30 cutoff are equivalent between Pipeline 2.5 and Pipelines 2.0-2.4.
4. The SV junction calling algorithm has been tuned for samples sequenced using library version 2. Among other changes, the number of DNBs required to support junctions provided in the ***allJunctionsBeta-[ASM-ID].tsv*** file was increased from 3 to 5.
5. The mobile element insertion (MEI) calling algorithm as been tuned for samples sequenced using library version 2 to reduce the number of false positives.
6. Additional sample and library information is now provided in the header of all text-based output files:
   - LIBRARY_SOURCE indicates whether version 1 or 2 of Complete Genomics' library generation process was used to create the clones for sequencing.
   - LIBRARY_TYPE indicates whether the sequencing library was generated using Complete Genomics' standard process or using Long-Fragment Read (LFR)technology.
   - TUMOR_STATUS indicates whether the sequenced genome is derived from a tumor.
   - CUSTOMER_SAMPLE_ID is the sample ID reported on the sample manifest.
   - SAMPLE_SOURCE is the source of the DNA reported on the sample manifest.
   - REPORTED_GENDER is the gender reported for the genome on the sample manifest.
   - CALLED_GENDER is the gender predicted by the Complete Genomics Analysis Pipeline.
7. Analysis pipeline 2.5 no longer supports Build 36 of the human genome reference.
8. The following annotation sources have been updated:
   - NCBI annotation build 104 is now used for genes, transcripts, and proteins
   - COSMIC version 65

- miRBase version 20
- Pfam annotations were extracted from NCBI annotation build 104 protein records on June 26, 2013
- DGV version 2

Fixed Issues

1. Cancer Sequencing Service: In Pipeline version 2.4, the names track (outermost circle) of the ***somaticCircos-[ASM-ID]-[comparison ID]*** plot shows both germline and somatic variants when it should only show high quality somatic variants. This is now fixed.
2. Deletions that span the splice donor and acceptor sites of an intron were assigned the IMPACT "DONOR" instead of both "DONOR" and "ACCEPTOR". This is now fixed.

Known Issues

1. Pipeline 2.5 has an increased number of insertions relative to the previous pipeline. These are predominantly in homopolymer runs.

Changes to Version 2.4

New Features and Enhancements

The following new features and enhancements are provided in this release:

2. Estimates of Lesser Allele Fraction (LAF) for 100 kb windows of the genome are now provided for each individual sample. This information is captured in the *bestLAFSingle*, *lowLAFSingle*, and *highLAFSingle* columns of the ***cnvDetailsNondiploidBeta-[ASM-ID].tsv.bz2*** file and the corresponding CGA_LAFS, CGA_LLAFS, and CGA_ULAFS tags of the ***vcfBeta-[ASM-ID].vcf.bz2*** and ***somaticVcfBeta-[ASM-ID].vcf.bz2*** files. The Cancer Sequencing Service already contains an estimate of LAF in the tumor based on loci identified as heterozygous in the normal. To reduce confusion between the two types of LAF calculations, column headers containing the somatic paired-sample LAF calculations have been renamed to include the term "paired" (e.g., *bestLAFPaired*, *lowLAFPaired*, and *highLAFPaired*) and the corresponding VCF tags have been renamed CGA_LAFP, CGA_LLAFP, and CGA_ULAFP.
3. The variant quality flagging system has been updated with the following features:
   - The Analysis Pipeline now provides variant quality flags for the purpose of negative filtering. As a consequence, the VQHIGH and SQHIGH flags are no longer provided. Instead, variants that did not meet VQHIGH or SQHIGH criteria are designated VQLOW or SQLOW, respectively. This change aligns the meaning of variant quality filters among the ***var-[ASM-ID].tsv.bz2***, ***masterVar-[ASM-ID].tsv.bz2***, ***vcfBeta***, and ***somaticVcfBeta*** file formats.
   - Variant quality flags are now concatenated in the new *varFilter* column of the ***var*** file and the new *alleleXVarFilter* columns of the ***masterVarBeta*** file. These columns replace the *varQuality* and *somaticQuality* columns.
4. Complete Genomics now reports ambiguous calls where possible in the ***var***, ***masterVarBeta***, ***vcfBeta***, and ***somaticVcfBeta*** files. These are alleles for which there is strong evidence that the allele is not reference, but for which there is insufficient evidence to make a single high-confidence call. For these variants, the Analysis Pipeline now selects the top scoring hypothesis as the primary call, designates the allele "AMBIGUOUS", and provides a concatenated list of up to 200 additional supported variant calls and their relative scores. Variant annotations are made relative to the primary call.

In the ***var*** and ***masterVarBeta*** files, the "AMBIGUOUS" designation is provided in the new *varFilter* columns (*varFilter* in ***var*** and *allele1VarFilter* and *allele2VarFilter* in ***masterVarBeta***). The additional supported variant calls and scores are available in the new alternative calls columns (*alternativeCalls* in ***var*** and *allele1AlternativeCalls* and *allele2AlternativeCalls* in ***masterVarBeta***).

In the ***vcfBeta*** and ***somaticVcfBeta*** files, "AMBIGUOUS" is a new possible value of the FT tag in the FORMAT column, and the additional supported variant calls and scores are provided in the new CGA_ALTCALLS tag in the FORMAT column.

5. Variants reported in the ***var*** and ***masterVarBeta*** files that are present in dbSNP are now annotated with the minor allele frequency of the variant allele as reported by the 1000 Genomes Project in dbSNP, where available. This information is provided in the new *alleleFreq* columns (*alleleFreq* in ***var***, *allele1Freq* and *allele2Freq* in ***masterVarBeta***). The minor allele frequency is also available in ***vcfBeta*** as a new 'AF=<value>' tag-value pair in the INFO column.
6. The mitochondrial chromosome is now considered diploid for the purpose of calling variants. This and other optimizations to the Analysis Pipeline increase its sensitivity to variants present in mitochondrial DNA at low allele fraction (such as due to heteroplasmy) and decrease the number of no-calls in the mitochondrial genome.
7. A variety of improvements were made to the Cancer Sequencing Service, including:
   - Somatic variant calling has been enhanced with the addition of a new score that uses a Fisher's exact test to measure the confidence in the read support for a called somatic variant. The p-value output of this test is provided in the *fisherSomatic* column.
   - For optimal somatic variant specificity, Complete Genomics recommends filtering for somatic variants with a *fisherSomatic* score of >=30 (corresponding to a p-value < 0.001). To enable users to easily filter out somatic variants that do not meet this criterion, a new FET30 flag, which indicates lower confidence in the somatic call, is provided in the corresponding *alleleXVarFilter* column of the ***masterVarBeta*** file. This flag is also provided, where appropriate, as a possible value for the FT tag in the FORMAT column in the ***somaticVcfBeta*** file.
   - The *locusDiffClassification* column in the tumor ***masterVarBeta*** contains comparison classifications for tumor-normal genome comparisons. This information is now provided in the normal ***masterVarBeta*** file within a tumor-normal pair or tumor-tumor-normal trio. The *locusDiffClassification* column in the normal ***masterVarBeta*** file is sample specific (as indicated by the comparison suffix), with separate columns for each non-baseline genome. For example, "*locusDiffClassification-T1*".
   - To facilitate the identification of blocks of loss-of-heterozygosity (LOH) in tumor genomes, heterozygous loci in the normal genome that are homozygous in the matched tumor genome are now identified in the normal ***masterVarBeta*** file. These loci, which may be homozygous reference or homozygous variant in the tumor, are flagged in the new *varFlags* column with an "lohVar-[comparison-suffix]" flag. The comparison suffix corresponds to the tumor sample in which the variant locus is homozygous. For example "lohVar-T1".
   - Refer to the *Data File Formats* document for more details on file content and format.
8. The following annotation sources have been updated:
   - dbSNP v137 for GRCh37
   - COSMIC version 61 for GRCh37
   - miRBase version 19 for GRCh37

Fixed Issues

1. In rare cases, the *allele2ReadCount* for a locus in ***masterVarBeta*** would equal *totalReadCount* when in fact it should not. This is now fixed.

Known Issues

1. There are a small number of loci designated with a loss-of-heterozygosity (LOH) flag in the ***somaticVCFBeta*** file that are not designated with a similar "lohVar" flag in the corresponding normal ***masterVarBeta*** file for the same sample. The opposite is also true: there a small number of loci designated LOH in the ***masterVarBeta*** file that are not designated LOH in the ***somaticVcfBeta*** file. This phenomenon is a consequence of differences in the sizes of loci used by comparison algorithms that output to ***masterVarBeta*** or ***vcfBeta***. The larger superloci used for the comparisons that populate ***masterVarBeta*** can contain adjacent no-calls (i.e., no-calls present in a locus adjacent to the LOH locus in ***vcfBeta***) or nearby variants that result in the comparison algorithm classifying the locus in tumor and normal as consistent, or result in the tumor being considered heterozygous (thus not LOH).

Changes to Version 2.2

New Features and Enhancements

The following new features and enhancements are provided in this release:

2. A new file, ***vcfBeta-[ASM-ID].vcf.bz2***, has been added to the ASM directory. This file contains the results of small variant, CNV, MEI, and SV detection with scores and annotations in VCF 4.1 format. The data is sourced from multiple Complete Genomics files, including ***masterVar***, ***cnvDetailsDiploidBeta***, ***cnvDetailsNondiploidBeta***, ***allJunctionsBeta,*** and ***mobileElementInsertionsBeta***. The ***vcfBeta*** file is generated by running CGA™ Tools mkvcf, a new tool publicly available in CGA Tools version 1.6.
3. A variety of improvements were made to the ***somaticVCFBeta-[ASM-ID].vcf.bz2*** file including:
   - The Somatic Status (SS) tag has been moved from the INFO field to the FORMAT field. This allows SS annotations on a per-allele basis and more specialized filtering.
   - Two new fields have been added to the INFO field:
     i. Allele count (AC)
     ii. Allele number (AN)
   - The FORMAT field has been enriched with several new tags, including:
     i. Genotype likelihood (GL)
     ii. Genotype likelihood, calibrated based on EAF score (CGA_CEGL)
     iii. Haplotype quality, calibrated based on EAF score (CGA_CEHQ)
   - Refer to the *Data File Formats* document for more details on file content and format.
4. The following annotation sources have been updated:
   - dbSNP v135 for GRCh37
   - COSMIC version 59 for both NCBI Build 36 (hg18) and GRCh37 (hg19)
   - miRBase version 18 for GRCh37

5. The ***mapping_[SLIDE-LANE]_00X.tsv.bz2*** file now includes a new *armWeight* column. This column provides a Phred encoding of the probability that this half-DNB mapping is incorrect, independent of the mappings of its mate.

Fixed Issues

1. Altered the variant caller alignment algorithm to improve VCF concordance with cgatools calldiff.
2. Changed alpha to 10^{-6} from 10^{-12} when computing mapping weight for DNBs with no consistent mate pairs. This value is reported in the weight column of mapping files (***mapping_[SLIDE-LANE]_00X.tsv.bz2***) in the MAP directory. This improves comparability of weight between DNBs with and without consistent mappings.
3. In Analysis Pipeline 2.0, the ***allSvEventsBeta-[ASM-ID].tsv*** and ***highConfidenceSvEvents-[ASM-ID].tsv*** files were missing "inversion" events > 1kb in length. This was due to a problem that has now been corrected in the CGA Tools junctions2events tool, which is used in the pipeline to create these files.

Changes to Version 2.0

New Features and Enhancements

With the release of Assembly Pipeline version 2.0, Complete Genomics now offers two sequencing services:

- Standard Sequencing Service — provides analysis and annotation of an individual genome.
- Cancer Sequencing Service — provides analysis and annotation of a tumor-normal pair or a tumor-tumor-normal trio.

These services rely on the same assembly pipeline algorithms for calling events, but contain differences in the delivered output. Broadly, the Cancer Sequencing Service includes Standard Sequencing results for individual genomes, as well as comparisons between the matched genomes within the pair or trio sample set. The following new features and enhancements are provided in this release, and apply to both services unless otherwise indicated.

Identifier Mapping File (Cancer Sequencing Service)

1. Each genome directory within the multi-genome dataset produced in the Complete Genomics Cancer Sequencing Service includes an ***idMap-[ASM-ID].tsv*** file. This file provides a mapping among various identifiers of a sample in a multi-genome dataset.

Assembly Pipeline Small Variant Caller Component

Several enhancements have been made to the Small Variant (SNPs, insertions, deletions, and block substitutions) Caller component of the assembly pipeline.

2. To improve sensitivity of small variant detection:
   - Two independent scoring methods are now used to quantify confidence in the reported call for each allele:
     - The Variable Allele Fraction (VAF) Scoring method does not constrain the allele fraction to be the same for each allele and reports the call confidence as *varScoreVAF*. The VAF Scoring method is particularly useful for discovering variants in non-diploid regions of a tumor or normal genome.

    - The Equal Allele Fraction (EAF) Scoring method assumes that the allele fraction for each allele is the same and reports the call confidence as *varScoreEAF*.

  The Complete Genomics Assembly Pipeline employs both scoring methods and generates both scores quantified. As a result, the *totalScore* field and its allele-specific equivalents have been replaced with *varScoreVAF* and *varScoreEAF* and their allele-specific equivalents.

- Thresholds required to make homozygous and heterozygous calls have been lowered. Previously, the minimum score (*totalScore*) for calling a homozygous variant and heterozygous variant were 20 dB and 40 dB, respectively. The thresholds required to make homozygous and heterozygous calls have been lowered to *varScoreVAF* of 10 dB and 20 dB, respectively. To facilitate filtering for high confidence variants, a *varQuality* column in ***var*** file and *allele1VarQuality* and *allele2VarQuality* columns in the ***masterVarBeta*** file have been added; possible values are 'VQLOW' and 'VQHIGH', where 'VQHIGH' is assigned for homozygous calls with score of at least 20 dB and other scored calls (i.e., heterozygous loci and half-called loci) with a score of at least 40 dB.
- Triploid hypotheses are considered during local *de novo* assembly optimization. Prior to Assembly Pipeline version 2.0, only diploid or haploid hypotheses were considered. This change entertains the possibility of there being more than two alleles, as may be the case in a heterogeneous tumor or one that is contaminated with normal tissue. If the triploid hypothesis is found to be the top hypothesis, information for all three alleles is recorded in the ***evidenceInterval*** and ***evidenceDnbs*** files. Once the variation interval is split into variant loci, each locus is separately fitted into a diploid hypothesis. Loci with three alleles are no-called.
- To reduce the false negative rate of known indels and short block substitutions, local *de novo* assembly is attempted for all known indels and block substitutions in a database of variants compiled from Complete Genomics Diversity Panel and dbSNP (versions 130 and 132 for genomes assembled using NCBI Build 36 and GRCh37, respectively).

3. To improve specificity of small variant detection:
   - This assembly pipeline version includes reduced spurious calls in segmental duplications and other repetitive regions through improvements made in correlation-based filtering.
   - Mate pairs (DNBs) that are not independently generated (i.e., any pair with at least one arm whose initial mappings have the same start and end position and share high sequence similarity) are de-duplicated; specifically, one mate pair is retained at random while the remaining mate pairs are discarded. Since this filter is applied after the mapping process, coverage values and initial mappings are not affected. However, de-duplication impacts variants, variation scores, evidence scores, reference scores, content of evidence files and evidence mappings, and read counts in ***masterVarBeta*** files.
   - This assembly pipeline version improved modeling of DNB sequence artifacts introduced during library process (for example, single base change or insertion/deletion of a few bases). Specifically, the Complete Genomics Small Variant Caller models the likelihood of a SNP or indel being introduced in the DNB.
4. Previous Assembly Pipeline versions used a right-shifted canonical form of indels. The current assembly pipeline now uses a left-shifted canonical form of indels, which is consistent with the recommended placement of variants for VCF format.
5. Variant alignment is more often consistent with a canonical placement of variants, simplifying genome comparison between Complete Genomics genomes.
6. Variant scores now indicate the quality of the variant with respect to the next best homozygous hypothesis that does not include that variant. Previously, a score indicated quality of the variant with respect to the next conflicting hypothesis. As a result of this change, our score emphasizes the confidence that a variant exists, while our old score that emphasized confidence that our exact call is correct.

7. Alignment of the top hypothesis in evidence files is now a straightforward per-base alignment. Previously, alignment was consistent with the calls, but alignment within a call was not done. The greatest impact of this change is in alignment of long no-called stretches where the top hypothesis differs in length from the reference.

Variations File

The following enhancements have been made to the Variations file (***var-[ASM-ID].tsv.bz2***):

8. As a result of implementing new variation scoring methods, the *totalScore* column has been replaced with two new columns, *varScoreVAF* and *varScoreEAF*.
9. A *varQuality* column has been added to indicate confidence category (low or high) of the call.

Master Variations File (Standard Sequencing Service)

The following new files and enhancements have been made to the Master Variations file (***masterVarBeta-[ASM-ID].tsv.bz2***) produced by the Complete Genomics Standard Sequencing Service:

10. *allele1Score* and *allele2Score* columns have been removed and are replaced by the following six score columns:
    - *allele1VarScoreVAF* reports confidence in the allele 1 call, under a variable allele fraction model
    - *allele2VarScoreVAF* reports confidence in the allele 2 call, under a variable allele fraction model
    - *allele1VarScoreEAF* reports confidence in the allele 1 call, under an equal allele fraction model
    - *allele2VarScoreEAF* reports confidence in the allele 2 call, under an equal allele fraction model
    - *allele1VarQuality* indicates confidence category (low or high) for allele 1 call, based on *allele1VarScoreVAF*
    - *allele2VarQuality* indicates confidence category (low or high) for allele 2 call, based on *allele2VarScoreVAF*
11. *xRef* column has been removed and is replaced by *allele1XRef* and *allele2XRef*. Previously, the *xRef* column combined the *xRef* values from all calls that are represented by a given locus in the ***masterVarBeta*** file. With the current assembly pipeline release, this information is presented in an allele-specific manner.
12. Two columns, *relativeCoverageNondiploid* and *calledLevel*, have been added to the report: for the segment that overlaps a given locus, both the relative coverage and the called coverage level determined using a non-diploid model are reported (see also "Assemble Pipeline CNV Component (Standard Sequencing Service)").

Master Variation File (Cancer Sequencing Service)

The following enhancements have been made to the Master Variation file (***masterVarBeta-[ASM-ID].tsv.bz2***) produced by the Complete Genomics Cancer Sequencing Service:

13. The modifications applied to the Standard Sequencing Service were also applied to the Cancer Sequencing Service.
14. For the Normal Genome: The ***masterVarBeta-[ASM-ID]-T1.tsv.bz2*** file (the "T1" in the file name assumes a comparison to a tumor genome) indicates whether or not the variations called for the sample are also supported in the tumor sample(s) in the dataset. The file includes read counts for the tumor sample(s) in the same multi-genome dataset.

15. For the Tumor Genome: The ***masterVarBeta-[ASM-ID]-N1.tsv.bz2*** file (the "N1" in the file name assumes a comparison to a normal genome) indicates whether or not the variations called for the sample are present or supported in the matched normal sample, including somatic scores to indicate confidence that the variant called in tumor is not present in normal sample and read counts for the matched normal sample. The following information are also included in the ***masterVarBeta*** file for the tumor genome:
    - *somaticCalledLevel* reports the coverage level of the overlapping segment, as called using a non-diploid CNV model where the matched normal is used for normalization
    - *bestLAF* reports the maximum likelihood estimate of the Lesser Allele Fraction (LAF) of the overlapping segment
    - *lowLAF* reports the minimum value within the 99% confidence interval on the Bayesian posterior estimate of LAF
    - *highLAF* reports the maximum value within the 99% confidence interval on the Bayesian posterior estimate of LAF
    - *locusDiffClassification* reports the tumor-normal comparison classification
    - *somaticCategory* reports the category for the somatic variant
    - *somaticRank* reports the estimated rank of the somatic variant amongst all true somatic variants within a given *somaticCategory*
    - *somaticScore* reports the score indicating the quality of the somatic variant
    - *somaticQuality* reports the category for the somatic variant quality, with value SQHIGH indicating somatic variants where *somaticScore* ≥ -10.

Somatic VCF File (Cancer Sequencing Service)

16. A ***somaticVcfBeta-[ASM-ID]-N1.vcf.bz2*** file is provided for each tumor genome. This VCF file reports small variants, CNVs, and junctions detected in both the tumor genome and the matched normal genome in a multi-genome dataset. Refer to the *Data File Formats* document for more details on file content and format.

dbSNP Annotation File

The following enhancements have been made to the ***dbSNPAnnotated-[ASM-ID].tsv.bz2*** file:

17. *scoreA* and *scoreB* fields have been replaced by *varScoreVAFA*, *varScoreEAFA*, *varScoreVAFB*, and *varScoreVAFB* fields representing new scores produced by the Small Variant Caller.
18. The minor allele and minor allele frequency (MAF) for dbSNP entries detected in the 1000 Genomes Project dataset have been added. This information is reported in the *1000GenomesProjectMinorAllele* and *1000GenomesMAF* fields. This information is only provided for genomes assembled against human reference genome GRCh37, where files are annotated with dbSNP version 132. If MAF is not available for the dbSNP entry, *1000GenomesProjectMinorAllele* and *1000GenomesMAF* fields are empty. For genomes assembled against human reference genome NCBI Build 36, where files are annotated with dbSNP version 130, MAF information is not available and *1000GenomesProjectMinorAllele* and *1000GenomesMAF* fields contain 'NA'.

Summary File

19. For metrics that are based on variant count (e.g., total SNP count, missense loci, SNP het/hom ratio, junction count), two values for each metric are now provided in the ***summary-[ASM-ID].tsv*** file; one value calculated using all variants detected and the other value calculated using high confidence variants.

EVIDENCE Directory (Standard Sequencing Service)

The following enhancements have been made to the EVIDENCE directory:

20. ***evidenceInterval-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2***:
    - As a result of the new variation scoring methods, *Score* column has been replaced with *EvidenceVAF* and *EvidenceEAF* columns.
    - As a result of the implementation of triploid hypothesis testing, the following columns were added:
        - *Allele3* reports the sequence of allele 3, if present.
        - *Allele3Alignment* reports the alignment of allele 3, if present.
21. ***evidenceDnbs-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2***:
    - As a result of the implementation of triploid hypothesis testing, *ScoreAllele3* column was added to indicate confidence of the alignment of the DNB to allele 3, if present.

EVIDENCE Directory (Cancer Sequencing Service)

The following enhancements have been made to the EVIDENCE directory:

22. The modifications applied to the Standard Sequencing Service were also applied to the Cancer Sequencing Service.
23. In addition to providing evidence supporting called variations for each genome in the EVIDENCE directory, Complete Genomics provides evidence for the presence or absence of the same variation in matched genomes within a multi-genome analysis group (this is done regardless of whether the same variation was called in the matched genome). This information is provided in an EVIDENCE-<comparison_ASM-ID> directory, where <comparison_ASM-ID> identifies another sample from the same analysis group. Specifically:
    - For the Normal Genome, an EVIDENCE-<comparison_ASM-ID> directory for each matching tumor genome has been added (where <comparison_ASM-ID> identifies the ASM-ID of the tumor genome). This directory contains files, ***evidenceIntervals-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2*** and ***evidenceDnbs-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2,*** representing reads and alignments in the tumor genome at called alleles against variations in the normal genome.
    - For the Tumor Genome, an EVIDENCE-<comparison_ASM-ID> directory for the matching Normal Genome has been added (where <comparison_ASM-ID> identifies the ASM-ID of the normal genome). This directory contains files, ***evidenceIntervals-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2*** and ***evidenceDnbs-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2,*** representing reads and alignments in the normal genome at called alleles against variations in the tumor genome.

Assemble Pipeline CNV Component (Standard Sequencing Service)

Several enhancements have been made to the CNV Pipeline:

24. The set of reference baseline samples used for normalization and assignment of no-call categories in single-sample CNV analysis has been updated. The baseline genome set now consists of 52 genomes from the Complete Genomics Diversity Panel. A file that summarizes the underlying data and normalization constants for each of the CNV baseline genomes can be downloaded from the Complete Genomics FTP site. For more information on these genomes and how they were processed to construct the baseline, refer to *CNV Baseline Genome Dataset: Data Format and Description*, which can be found at the same location as the composite file.
25. Window definitions have changed so that window boundaries are round numbers. For example, for 2K windows, window boundaries will end with "x000", where x is an even digit. Exceptions to this are windows at the ends of contigs. Windows will never span bases taken from more than

one contig, even if the gap between contigs is small enough to permit this. Bases outside the outermost full default windows for each contig will either be added to the first full window towards the center of the contig or be placed in their own window, depending on whether the number of bases is larger than ½ the window width or not.

26. Each sequenced genome is now processed with both the diploid model (model applied to "Normal Genome" in previous assembly pipeline versions) and non-diploid model (model applied to "Tumor Genome" in previous assembly pipeline versions) of the CNV pipeline. Results from both the diploid and non-diploid models are provided for each sequenced genome. Specifically, files provided for the following Complete Genomics Sequencing Services are outlined below:

    - ***cnvSegmentsDiploidBeta-[ASM-ID].tsv*** file provides the diploid-model segmentation of the complete reference genome into regions of distinct ploidy levels, giving the estimated ploidy, the average and relative adjusted coverage for each segment, and measures of confidence in the called segments.
    - ***cnvSegmentsNondiploidBeta-[ASM-ID].tsv*** file provides the non-diploid-model segmentation of the complete reference genome into regions of distinct coverage levels, giving the estimated level, the average and relative adjusted coverage for each segment, and measures of confidence in the called segments.
    - ***cnvDetailsDiploidBeta-[ASM-ID].tsv.bz2*** file reports estimated ploidy, average and relative adjusted coverage, and confidence scores for every 2 kb along the genome.
    - ***cnvDetailsNondiploidBeta-[ASM-ID].tsv.bz2*** file reports estimated level, average and relative adjusted coverage, and confidence scores for every 100 kb along the genome.
    - ***depthOfCoverage_100000-[ASM-ID].tsv*** reports unique and weight-sum sequence coverage, along with GC bias-corrected weight-sum coverage for every 100 kb non-overlapping window along the sequenced genome.

Assemble Pipeline CNV Component (Cancer Sequencing Service)

27. For the Normal Genome, all files described above for the Standard Sequencing Services are provided.

28. For each Tumor Genome, all files described above for the Standard Sequencing Service are provided.

29. In addition, the segmentation and details files for somatic comparisons are provided in four new files. For somatic comparisons, coverage in the tumor sample is normalized using GC-corrected coverage in the matched Normal Genome instead of the set of unrelated baseline samples used in the non-somatic CNV analysis. The "N1" in each file name indicates the comparison to an assumed normal genome.

    - ***somaticCnvSegmentsDiploidBeta-[ASM-ID]-N1.tsv*** file provides the diploid segmentation of the complete reference genome into regions of distinct ploidy levels, giving the estimated ploidy, the average and relative adjusted coverage for each segment, measures of confidence in the called segments, and estimates of Lesser Allele Frequency (LAF) for each segment.
    - ***somaticCnvSegmentsNondiploidBeta-[ASM-ID]-N1.tsv*** file provides the non-diploid segmentation of the complete reference genome into regions of distinct coverage levels, giving the estimated level, the average and relative adjusted coverage for each segment, measures of confidence in the called segments, and estimates of LAF for each segment.
    - ***somaticCnvDetailsDiploidBeta-[ASM-ID]-N1.tsv.bz2*** file reports estimated ploidy, average and relative adjusted coverage, confidence scores for every 2 kb along the genome, and estimates of LAF for each window.

- ***somaticCnvDetailsNondiploidBeta-[ASM-ID]-N1.tsv.bz2*** file reports estimated level, average and relative adjusted coverage, confidence scores for every 100 kb along the genome, and estimates of LAF for each window.

Assembly Pipeline SV Component (Standard Sequencing Service)

30. Identified junctions in the sequenced genome are now composed into structural variation events and reported in two files. Data reported in these files are generated by running CGA Tools junctions2events.
    - ***allSvEventsBeta-[ASM-ID].tsv*** file reports events identified by rationalizing all junctions found in the ***allJunctionsBeta*** file, along with annotations of impact on genes.
    - ***highConfidenceSvEvents-[ASM-ID].tsv*** file reports events identified by rationalizing high-confidence junctions found in the ***highConfidenceBeta*** file, while considering junctions in the ***allJunctionsBeta*** file as potential event partners. Annotations of impact on genes are also provided.
31. The following fields have been added to the ***allJunctionsBeta-[ASM-ID].tsv*** and ***highConfidenceJunctionsBeta-[ASM-ID].tsv*** files. Information in these fields helps relate junctions to events they represent.
    - *EventID* provides the identifier of the event that this junction is a part of.
    - *Type* identifies the structural rearrangement that this junction represents or of which it is a part.
    - *RelatedJunctions* provides the identifiers of other junctions that, together with this junction, make up the event indicated in *Type* field.
32. The set of reference baseline samples used for *FrequencyInBaseline* field in various junction files has been updated. The baseline genome set now consists of 52 genomes from the Complete Genomics Diversity Panel. A file that summarizes the junction data across SV baseline genomes can be downloaded from the Complete Genomics FTP site. For more information on these genomes and how they were processed to construct the baseline, refer to the *SV Baseline Genome Dataset: Data Format and Description* document at the same location as the composite file.
33. Previously, *LeftGenes* and *RightGenes* fields of various junction files reported the transcript name and strand of gene(s) that overlap the left and/or right section of a junction. We now report the gene symbol of these gene(s).

Assembly Pipeline SV Component (Cancer Sequencing Service)

34. The modifications applied to the Standard Sequencing Service were also applied to the Cancer Sequencing Service.
35. For the Tumor Genome, somatic junctions, defined as junctions that are present in the tumor sample but absent in the matched normal sample, are reported in two new files: ***somaticAllJunctionsBeta-[ASM-ID].tsv*** and ***somaticHighConfidenceJunctionsBeta-[ASM-ID].tsv.*** Data reported in these files are generated by running CGA Tools junctiondiff.

REPORTS Directory (Standard Sequencing Service)

The following files have been added to the REPORTS directory within the ASM directory for each genome:

36. ***circos-[ASM-ID].html*** and ***circos-[ASM-ID].png*** files represent a Circos visualization of small variant, CNV, and structural variation data, along with other associated data such as Lesser Allele Fraction and homozygous and heterozygous SNP density. The ***circosLegend.png*** file provides the legend that defines the data being plotted.

REPORTS Directory (Cancer Sequencing Service)

37. The modifications applied to the Standard Sequencing Service were also applied to the Cancer Sequencing Service.
38. Additionally, for each Tumor Genome, ***somaticCircos-[ASM-ID]-N1.html*** and ***circos-[ASM-ID]-N1.png*** files represent a Circos visualization of somatic small variant, CNV, and structural variation data, along with other associated data such as Loss of Heterozygosity (LOH), LAF, and homozygous and heterozygous SNP density. The ***somaticCircosLegend.png*** file provides the legend that defines the data being plotted.

Fixed Issues

1. On rare occasions, the ***var-[ASM-ID].tsv.bz2*** file reported a heterozygous locus where the top hypothesis was homozygous. This occurred when an allele could be aligned more than one way against the reference. This has been fixed.
2. In some cases, due to no-called alleles of the top hypothesis split into separate loci, the top hypothesis contained a ref allele while the ***var-[ASM-ID].tsv.bz2*** file contained a locus where the alt call corresponded to the ref allele of the top hypothesis. This has been fixed.
3. In rare cases, the mate gap specified in ***evidenceDnbs-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2*** file was not within the empirical mate gap distribution. This has been fixed.
4. The source file used for gene annotations of ***allJunctionsBeta-[ASM-ID].tsv*** and ***highConfidenceJunctionsBeta-[ASM-ID].tsv*** files was different from the source file used for other Complete Genomics files. Differences corresponded to mRNA accessions found in one source file but not the other. This has been fixed such that gene annotations of the junctions file are consistent with other Complete Genomics files.
5. In Assembly Pipeline versions 1.11 and 1.12, the *xRef* columns in the ***allJunctionsBeta***, ***highConfidenceJunctionsBeta***, and ***evidenceJunctionsClusterBeta*** files of genomes assembled using human reference GRCh37 were always empty. This has been fixed such that the field now reports variation in dbSNP that overlaps region between *leftPosition* and *rightPosition* of a junction. The field also indicates a deletion not reported in dbSNP.
6. In the ***var-[ASM-ID].tsv.bz2*** file, two SNPs that are on opposite haplotypes but adjacent on the reference are reported as a single locus. This has been fixed.

Known Issues

1. The *uniqueSequenceCoverage* column in the ***coverage-[ASM-ID].tsv*** and ***coverageCoding-[ASM-ID].tsv*** files reports the number of unique, fully and half mapping reads at a given coverage depth instead of the number of unique, fully mapping reads as described in the *Data File Formats* documentation. Additionally, the *cumulativeUniqueSequenceCoverage* column in the same files reports the cumulative number of unique, fully and half mapping reads at a given coverage depth instead of the cumulative number of unique, fully mapping reads.
2. On rare occasions, for variants in the PAR, the same variant is being annotated with different dbSNP entries across different genomes.
3. For all files that report #COSMIC value in the header, the COSMIC version used for annotation is incorrect. It currently lists version 48 but the actual version being used is v53.
4. The following lines in the header of the ***allSvEventsBeta*** and ***highConfidenceSvEventsBeta*** files are missing: #ASSEMBLY_ID, #FORMAT_VERSION, #GENOME_REFERENCE, #SAMPLE, and #GENE_ANNOTATIONS. The values in these missing fields are the same as the values for these header rows in other files within the same directory for the genome.

5. On rare occasions, Pfam annotation is duplicated for a given locus in the ***masterVarBeta-[ASM-ID].tsv.bz2*** file.
6. No-call loci within blocks of 10 or longer no-calls are excluded from the ***gene-[ASM-ID].tsv.bz2*** file.
7. In a few cases, large differences between the genomic reference and RefSeq sequence within a Pfam domain lead to coordinate conversion difficulties and subsequent failure to annotate any part of the Pfam domain.
8. A small percentage of transcripts in NCBI Build 36 and GRCh37 are excluded from the annotation results due to the one or more of the following reasons: (1) the transcript contains an unknown ("X") amino acid; (2) start and/or stop codon positions are unknown; (3) the transcript contains unspecified nucleotides; and (4) the transcript maps to an unknown location/chromosome. To obtain the list of transcripts, please contact support@completegenomics.com.
9. For genes that partially map to the reference genome, the 5' transcriptional start site is misidentified for a small set of genes (about 144 transcripts) in the ***gene-[ASM-ID].tsv.bz2*** file. As a result, variants may be incorrectly annotated as falling within the TSS-UPSTREAM region (7.5 kb upstream of 5' transcriptional start site). To obtain a list of affected transcripts, please contact support@completegenomics.com.
10. For a few transcripts in which alignment information cannot be parsed, the *impact* field in ***gene-[ASM-ID].tsv.bz2*** file will be annotated with 'UNKNOWN-TR'.
11. For predicted genes without stop codons, the *impact* field in ***gene-[ASM-ID].tsv.bz2*** file will be annotated with 'UNKNOWN-TR'.
12. Because COSMIC does not provide a transcript version number, COSMIC annotation in the gene file is copied over from the *xRef* field of the variation file. Thus, the transcript described in the ***gene-[ASM-ID].tsv.bz2*** file may not be the transcript that is associated with the COSMIC record.

Changes to Version 1.12.0

New Features and Enhancements

The following new features and enhancements are provided in this release:

1. Data File Format version has been changed to v1.7.
2. In the ***gene-[ASM-ID].tsv.bz2*** file, the value "COMPATIBLE" in the *impact* field was used to indicate a variation where the DNA sequence for the transcript that the variation overlaps has changed but there is no corresponding change in the protein sequence. The value "COMPATIBLE" has been changed to "SYNONYMOUS" to be more consistent with standard terminology.
3. Mobile element insertion (MEI) detection has been added to our Assembly Pipeline as a Beta feature. Incorporation of transposable elements that are novel with respect to the reference genome are identified, and the insertion element type is determined by alignment of mate pair arms to the sequences of various possible mobile elements in our sequence database (please see *Data File Formats* for descriptions of the mobile element sequence database). Key information for MEI analysis is provided in the MEI subdirectory of the ASM directory of the data. New features and outputs for MEI include:
   - Summary statistics from MEI detections have been added to the ***summary-[ASM-ID].tsv*** file. See *Data File Formats* for description of these statistics.
     i. Mobile element insertion count
     ii. Fraction of MEIs that are novel relative to MEIs detected by 1000 Genomes Project

- Results of the MEI analysis are reported in three files. They can be found in the MEI sub-directory of the ASM directory.
  i. ***mobileElementInsertionsBeta.tsv*** reports detected mobile element insertion events, and associated information including reference coordinates where the mobile element was inserted, element type, Phred-like score to indicate confidence in the insertion, overlapping genes, and presence or absence of the detected event in the 1000 Genome Projects MEI dataset.
  ii. ***mobileElementInsertionsROCBeta-[ASM-ID].png*** provides a graph showing the number of known and novel MEIs detected for various score cutoffs. This graph can be used to choose a score threshold to apply to the dataset to achieve the desired balance between specificity and sensitivity.
  iii. ***mobileElementInsertionsRefCountsBeta-[ASM-ID].png*** provides a graph showing the distribution of mate pair counts that support the reference allele for MEI events. This graph can be used interpret the zygosity of the called insertion event.

4. A new file, ***masterVarBeta-[ASM-ID].tsv.bz2*** has been added to the ASM directory. This is a simple, integrated master variation file that reports the variant calls and annotation information produced by the Complete Genomics assembly process. The file format is derived heavily from the existing variation file (***var-[ASM-ID].tsv.bz2***), but integrates annotation information from data in other Complete Genomics data files. This aggregation of information facilitates filtering for variations of interest using information that was previously distributed across multiple files. Another benefit of this format is that it can more easily be converted into other standard variation file formats. See *Data File Formats* for description of the ***masterVarBeta-[ASM-ID].tsv.bz2*** file.
5. The following annotation sources have been updated:
   - NCBI build 37.2 annotation release
   - COSMIC version 52 for both NCBI Build 36 and GRCh37
   - miRBase version 16 for GRCh37
   - Pfam annotations were taken from NCBI's Conserved Domain Database on April 21st, 2011 for both NCBI Build 36 and GRCh37

Fixed Issues

1. In the ***gene-[ASM-ID].tsv.bz2*** file, allele 2 was sometimes reported before allele 1. This has been fixed such that allele 1 is always reported before allele 2.
2. Previously, a hierarchical system was used to determine which alternate start codons are permitted for the genes that use them. In some cases, this allowed more alternatives than are actually known to be relevant for a given gene. This has been fixed. The data now specify exactly which alternative start codon is permitted for genes that use them.
3. If there was a frameshift in the reference genome with respect to RefSeq, the reference amino acid sequence was reported in the *genomeRefSequence* field of the ***gene-[ASM-ID].tsv.bz2*** as if the frameshift had not occurred. This has been fixed.
4. CNV calls spanning windows of extremely high coverage were sometimes assigned to an arbitrary *calledPloidy* or *calledLevel*. Additionally, regions that are not good fits to any *calledPloidy* or *calledLevel* were sometimes assigned to *calledPloidy*=0 or assigned to the lowest *calledLevel*. These issues have been fixed.
5. In Assembly Pipeline version 1.11, non-coding regions were annotated with Pfam domains, vastly increasing the number of Pfam annotated variations. This has been fixed.

Known Issues

1. On rare occasions, Pfam annotation is duplicated for a given locus in the ***masterVarBeta-[ASM-ID].tsv.bz2*** file.
2. In rare cases, the mate gap specified in ***evidenceDnbs-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2*** file is not within the empirical mate gap distribution.
3. No-call loci within blocks of 10 or longer no-calls are excluded from the ***gene-[ASM-ID].tsv.bz2*** file.
4. On rare occasions, the ***var-[ASM-ID].tsv.bz2*** file reports a heterozygous locus where the top hypothesis is homozygous. This occurs when an allele can be aligned more than one way against the reference.
5. In a few cases, large differences between the genomic reference and RefSeq sequence within a Pfam domain lead to coordinate conversion difficulties and subsequent failure to annotate any part of the Pfam domain.
6. A small percentage of transcripts in Build 36 and GRCh37 are excluded from the annotation results due to the one or more of the following reasons: (1) the transcript contains an unknown ("X") amino acid; (2) start and/or stop codon positions are unknown; (3) the transcript contains unspecified nucleotides; and (4) the transcript maps to an unknown location/chromosome. To obtain the list of transcripts, please contact support@completegenomics.com.
7. For genes that partially map to the reference genome, the 5' transcriptional start site is misidentified for a small set of genes (about 144 transcripts) in the ***gene-[ASM-ID].tsv.bz2*** file. As a result, variants may be incorrectly annotated as falling within the TSS-UPSTREAM region (7.5 kb upstream of 5' transcriptional start site). To obtain a list of affected transcripts, please contact support@completegenomics.com.
8. For a few transcripts in which alignment information cannot be parsed, the *impact* field in ***gene-[ASM-ID].tsv.bz2*** file will be annotated with "UNKNOWN-TR".
9. For predicted genes without stop codons, the *impact* field in ***gene-[ASM-ID].tsv.bz2*** file will be annotated with "UNKNOWN-TR".
10. Because COSMIC does not provide a transcript version number, COSMIC annotation in the gene file is copied over from the *xRef* field of the variation file. Thus, the transcript described in the ***gene-[ASM-ID].tsv.bz2*** file may not be the transcript that is associated with the COSMIC record.

Addendum

The following issues were discovered after the release of Assembly Pipeline version 1.12:

1. On rare occasions, the ***var-[ASM-ID].tsv.bz2*** file reports a heterozygous locus where the top hypothesis is homozygous. This occurs when an allele can be aligned more than one way against the reference. This problem was fixed in Assembly Pipeline version 2.0.
2. In some cases, due to no-called alleles of the top hypothesis split into separate loci, the top hypothesis contains a ref allele and, the ***var-[ASM-ID].tsv.bz2*** file contains a locus where the alt call corresponds to the ref allele of the top hypothesis. This problem was fixed in Assembly Pipeline version 2.0.

Changes to Version 1.11.0

New Features and Enhancements

The following new features and enhancements are provided in this release by comparison with previous data shipped or released by Complete Genomics:

1. Data File Format version has been changed to v1.6
2. The following changes have been made to the ***summary-[ASM-ID].tsv*** file:
   - The following new summary statistics and metrics have been added. See *Data File Formats* for description of statistics and metrics.
     - Both mates mapped yield (Gb)
     - Genome fraction where weightSumSequenceCoverage >=5x
     - Genome fraction where weightSumSequenceCoverage >=10x
     - Genome fraction where weightSumSequenceCoverage >=20x
     - Genome fraction where weightSumSequenceCoverage >=30x
     - Genome fraction where weightSumSequenceCoverage >=40x
     - Fully called exome fraction
     - Partially called exome fraction
     - No-called exome fraction
     - Exome fraction where weightSumSequenceCoverage >=5x
     - Exome fraction where weightSumSequenceCoverage >=10x
     - Exome fraction where weightSumSequenceCoverage >=20x
     - Exome fraction where weightSumSequenceCoverage >=30x
     - Exome fraction where weightSumSequenceCoverage >=40x
     - Homozygous SNP count
     - Heterozygous SNP count
     - SNP novel fraction
     - Het SNP novel rate
     - For exome variations: SNP total count
     - For exome variations: Hom SNP count
     - For exome variations: Het SNP count
     - For exome variations: SNP novel rate
     - For exome variations: Hom SNP novel rate
     - For exome variations: Het SNP novel rate
     - For exome variations: SNP het/hom ratio
     - For exome variations: SNP Transitions/transversions
     - For exome variations: INS total count
     - For exome variations: INS novel rate
     - For exome variations: INS het/hom ratio
     - For exome variations: DEL total count
     - For exome variations: DEL novel rate
     - For exome variations: DEL het/hom ratio
     - For exome variations: SUB total count
     - For exome variations: SUB novel rate
     - For exome variations: SUB het/hom ratio
     - Non-synonymous SNP loci
     - Misstart SNP loci
     - Disrupt SNP loci
     - Total number of CNV segments
     - Total bases in CNV segments

    - Fraction novel CNV (by segment)
    - Fraction novel CNV (by bases)
    - Total junction count
    - High confidence junction count
  - A *Category* field has been added to the file to indicate the type of metric reported. Possible category values include: Genome coverage, Exome coverage, Library, Genome variations, Exome variations, Functional impact, CNV, and SV.
  - For the mate distribution range (95% "CI") metric used to report range of mate gap that captures 95% of the data, the Min and Max of this range are now reported on two separate lines.

3. In previous releases, the ***coverage-[ASM-ID].tsv*** and ***coverageByGcContent-[ASM-ID].tsv*** files provided in the REPORTS directory reported coverage data for the whole genome.
   In this release, two additional files, ***coverageCoding-[ASM-ID].tsv*** and ***coverageByGcContentCoding-[ASM-ID].tsv***, are provided for exome coverage and exome coverage as a function of GC content, respectively.
4. The following changes have been made to the ***dbSNPAnnotated-[ASM-ID].tsv.bz2*** file:
   - The *found* and *exactMatch* fields have been removed.
   - New fields, *alleleAGenotype* and *alleleBGenotype* have been added to report the dbSNP allele matched to allele 1 and allele 2 of the variant file. The special value "NO-CALL" is used to denote a no-call in the variant file. "NO-MATCH" is given if the locus was called but did not match any of the dbSNP alleles.
   - In previous releases, possible *zygosity* field values were: "hom", "het", or empty for homozygous, heterozygous, and unknown, respectively. In this software version, possible values to indicate zygosity of *alleleAGenotype* and *alleleBGenotype* are: "unknown", "no-call", "hap", "half", "het-ref", "het-alt", and "hom".
   - In *varTypeA* and *varTypeB* fields, a prefix "multiple:" is added to the semi-colon separated list of *varTypes* from the ***var-[ASM-ID].tsv.bz2*** file when more than one call is required to recapitulate dbSNP entry.
5. A change was made to the ***coverage-[ASM-ID].tsv*** file. Previously, we reported a sum for the number of reads at coverage depth of 1000 or greater and indicated coverage as "1000+". In the current release, we are now reporting number of reads at every coverage depth observed in the sequenced genome.
6. Enhancements were made to the ***gene-[ASM-ID].tsv.bz2*** file:
   - For variation where *component* = "DONOR" or "ACCEPTOR", *nucleotidePos* and *proteinPos* values represented mRNA or protein positions adjacent to the exon/intron boundary. Values for these fields could be negative or very large for splice sites that occur before the beginning of the protein or mRNA. This has been changed such that values reported for *nucleotidePos* and *proteinPos* are "0".
7. Enhancements were made to the ***geneVarSummary-[ASM-ID].tsv*** file:
   - Functional impact categories "DISRUPT" and "MISSTART" have been added to the ***geneVarSummary-[ASM-ID].tsv*** file. It now reports total and novel count of "DISTRUPT" and "MISSTART" variations that fall within a RefSeq transcript.
8. Provided dbSNP annotations are from build 132 when chosen human reference genome is GRCh37. Previously dbSNP 131 was used for GRCh37.
9. Strand information has been added for transcript annotations in the ***allJunctionsBeta-[ASM-ID].tsv*** and ***highConfidenceJunctionsBeta-[ASM-ID].tsv*** files. For example, "NM_173508:+".

10. The following changes were made to *gcCorrectedCvg* and *avgNormalizedCvg* fields reported in the ***cnvDetailsBeta-[ASM-ID].tsv.bz2***, ***cnvTumorDetailsBeta-[ASM-ID].tsv.bz2***, ***cnvSegmentsBeta-[ASM-ID].tsv***, and ***cnvTumorSegments-[ASM-ID].tsv*** files:
    - In previous version, *gcCorrectedCvg* can be negative in regions of the genome with very high or very low GC content. This has been changed such that *gcCorrectedCvg* is now no-called ('N') in regions of the genome with very high or very low GC content.
    - In previous version, *avgNormalizedCvg* can result in a negative value from explicit no-calling of the normalization when the average coverage in the baseline samples is extremely low. *avgNormalizedCvg* can also be negative when *gcCorrectedCvg* is negative. This has been changed such that *avgNormalizedCvg* is no-called ('N') in these scenarios.
11. Starting with this pipeline release, the following file name convention is enforced for files in the ASM directory:

    ```
    [^-]+-[a-zA-Z0-9_-]+-ASM
    ```

    where `[^-]+-` denotes file name convention and `[a-zA-Z0-9_-]+-ASM` denotes ASM ID convention. For example:

    ```
    geneVarSummary-GS000000474-ASM.tsv
    ```

    Renaming of Complete Genomics ASM files or writing code to handle these files should take this convention into consideration. CGA Tools also considers this convention when handling ASM files.

Fixed Issues

1. In Assembly Pipeline release version 1.10, the ***geneVarSummary-[ASM-ID].tsv*** file for tumor sample was supposed to have the same file format as file for non-tumor sample. However, *calledPloidy* and *relativeCvg* fields were missing, and a *calledLevel* field was reported. This has been changed: the *calledLevel* field has been removed and *calledPloidy* and *relativeCvg* fields have been added to the tumor sample file.
2. In some cases, *genomeRefSequence* field of the ***gene-[ASM-ID].tsv.bz2*** file reported incorrect amino acid sequence for no-calls. This has been fixed.
3. A small number of computationally predicted genes do not map fully to the reference genome and hence lack a start codon, a stop codon, or both. In previous releases, annotations reported in the ***gene-[ASM-ID].tsv.bz2*** file of variants found throughout these genes were incorrect in that many are called *impact* = "MISSTART" or "NONSTOP" even though the gene lacked a start codon, a stop codon, or both. This has been fixed.
4. In rare cases where we are annotating a call that is close to a position where the RefSeq and genomic reference sequence differed, the called amino acid sequence, reference amino acid sequence, and the functional impact reported for the call in the ***gene-[ASM-ID].tsv.bz2*** file may have been incorrect. This has been fixed.
5. In previous releases, when an indel was found in the RefSeq transcript with respect to the reference, that indel was applied to the reference sequence when determining the amino acid sequence reported in the *genomeRefSequence* field of the ***gene-[ASM-ID].tsv.bz2*** file. Thus, this field contained non-reference sequence. This has been fixed.
6. The ***coverage-[ASM-ID].tsv*** file in the REPORTS folder skipped reporting coverage 999. This has been fixed.
7. Some values in *nucleotidePos* and *proteinPos* fields in the ***gene-[ASM-ID].tsv.bz2*** file had "-1" appended. This has been fixed.
8. In Assembly Pipeline version 1.10, the *DeletedTransposableElement* field in the ***allJunctionsBeta-[ASM-ID].tsv***, ***highConfidenceJunctionsBeta-[ASM-ID].tsv***, and ***evidenceJunctionClustersBeta-[ASM-ID].tsv*** files only considered L1 and AluY subtypes with

divergence at or below 2% when annotating junctions. This has been changed such that all transposable elements up to 20% divergence are considered when annotating junctions.

9. In previous release, some calls equivalent to RefSeq mRNA sequence were incorrectly classified as "FRAMESHIFT", in the case when NCBI alignments of RefSeq to the reference genome contains nearby indel alignments.
10. In previous release, some "INSERT" calls in the ***gene-[ASM-ID].tsv.bz2*** file incorrectly classified as "INSERT+" and some "DELETE" calls incorrectly classified as "DELETE+".
11. In previous releases, some lines in the ***gene-[ASM-ID].tsv.bz2*** file have an extra tab at the end. This has been removed.
12. In previous releases, *nucleotidePos* and *proteinPos* fields of the ***gene-[ASM-ID].tsv.bz2*** file contained negative values if the position of the call relative to mRNA start sequence was negative. This has been fixed such that negative values are now 0.
13. Insertions, deletions, and length changing substitutions were annotated as "NONSENSE" in the ***gene-[ASM-ID].tsv.bz2*** file only if the length of mRNA changed by a multiple of three. In addition, these variants were called "NONSENSE" if a stop codon was introduced anywhere along length of mRNA. This has been fixed such that these length-changing variants are annotated as "NONSENSE" if the first affected codon of the new sequence is a stop codon. If the variant occurs in-frame, it is annotated as "NONSENSE" if a stop codon is introduced anywhere along the length of the variant.
14. In Assembly Pipeline version 1.10, *LeftPosition* or *RightPosition* in the ***allJunctionsBeta-[ASM-ID].tsv*** file and ***highConfidenceJunctionsBeta-[ASM-ID].tsv*** file were incorrect when transition sequence alignment included gaps. This has been fixed.
15. In previous releases, the number of variants with *impac*t = "FRAMESHIFT" in the ***gene-[ASM-ID].tsv.bz2*** file was not consistent with total count of "FRAMESHIFT" in the ***geneVarSummary-[ASM-ID].tsv*** file, due to incorrect counting in the ***geneVarSummary-[ASM-ID].tsv*** file. This has been fixed.
16. The count for variants with *impac*t = "MISSENSE" was incorrect because it included variants with *impact* = "MISSTART" in the ***geneVarSummary-[ASM-ID].tsv*** file. This has been fixed: the count for *impact* = "MISSTART" is now separated out into its own field.
17. In previous versions, the name of file ***depthOfCoverage-100000-[ASM-ID].tsv*** breaks the convention of Complete Genomics file name format. This has been fixed such that name of file is: ***depthOfCoverage_100000-[ASM-ID].tsv***.
18. In previous versions, scores were not present for some deletions in ***dbSNPAnnotated-[ASM-ID].tsv.bz2*** file. This has been fixed.

Known Issues

1. No-call loci within block of 10 or longer no-calls are excluded from the ***gene-[ASM-ID].tsv.bz2*** file.
2. On rare occasion, the ***var-[ASM-ID].tsv.bz2*** file reports a heterozygous loci where top hypothesis is homozygous. This occurs when an allele can be aligned more than one way against the reference. If because of no-calling, one allele is aligned differently than the other allele, a fully called heterozygous locus could result. Ideally, variant should always be called homozygous whenever the top hypothesis is homozygous.
3. In a few cases, large differences between the genomic reference and RefSeq sequence within the Pfam domain lead to coordinate conversion difficulties and subsequent failure to annotate any part of the Pfam domain. We should annotate the portion of the Pfam domain that is consistent with the reference sequence.

For GRCh37, the few cases are:

- Sulfotransfer_1 (pfam00685) hit on NP_006031.2 (corresponding to NM_006040.2)
- 7tm_1 (pfam00001) hit on NP_001002905.1 (corresponding to NM_001002905.1)

For Build 36:

- Sulfotransfer_1 (pfam00685) hit on NP_006031.2 (NM_006040.2)
- DUF1193 (pfam06702) hit on NP_064608.2 (NM_020223.2)
- PARG_cat (pfam05028) hit on NP_003622.2 (NM_003631.2)

4. A small percentage of transcripts in Build 36 and GRCh37 are excluded from the annotation results due to the one or more of the following reasons: (1) the transcript contains unknown ("X") amino acid; (2) start and/or stop codon positions are unknown; (3) the transcript contains unspecified nucleotides; and (4) the transcript maps to unknown location/chromosome. To obtain the list of transcripts, please contact support@completegenomics.com.
5. For genes that partially map to the reference genome, 5' transcriptional start site is misidentified for a small set of genes (25 transcripts in Build 36 and 26 transcripts for GRCh37) in the ***gene-[ASM-ID].tsv.bz2*** file. As a result, variants are incorrectly annotated as falling within the TSS-UPSTREAM region (7.5 kb upstream of 5' transcriptional start site). To obtain a list of affected transcripts, please contact support@completegenomics.com.
6. If there is a frameshift in the reference genome with respect to RefSeq, the reference amino acid sequence is reported in the *genomeRefSequence* field of the ***gene-[ASM-ID].tsv.bz2*** as if the frameshift had not occurred.
7. For a few transcripts in which alignment information cannot be parsed, the *impact* field in ***gene-[ASM-ID].tsv.bz2*** file will be annotated with "UNKNOWN-TR".
8. For predicted genes without stop codons, which are not parsed correctly, the *impact* field in ***gene-[ASM-ID].tsv.bz2*** file will be annotated with "UNKNOWN-TR".
9. Because COSMIC does not provide a transcript version number, COSMIC annotation in the gene file is copied over from the *xRef* field of the variation file that is based on genomic coordinate. Thus, the transcript described in the ***gene-[ASM-ID].tsv.bz2*** file may not be the transcript that is associated with the COSMIC record.

Addendum

The following issues were discovered after the release of Assembly Pipeline version 1.11:

1. Some non-coding regions were annotated with Pfam domains, vastly increasing the number of Pfam annotated variations.
2. On rare occasions, the ***var-[ASM-ID].tsv.bz2*** file reports a heterozygous locus where the top hypothesis is homozygous. This occurs when an allele can be aligned more than one way against the reference. This problem was fixed in Assembly Pipeline version 2.0.
3. In some cases, due to no-called alleles of the top hypothesis split into separate loci, the top hypothesis contains a ref allele and, the ***var-[ASM-ID].tsv.bz2*** file contains a locus where the alt call corresponds to the ref allele of the top hypothesis. This problem was fixed in Assembly Pipeline version 2.0.

Changes to Version 1.10.0

New Features and Enhancements

The following new features and enhancements are provided in this release by comparison with previous data shipped or released by Complete Genomics:

1. Data File Format version has been changed to v1.5.
2. Genomic copy number analysis has been added to our Assembly Pipeline as a Beta feature. The CNV analysis pipeline for non-tumor genomes uses depth of coverage to segment the genome into regions of distinct ploidy. Each segment is reported with the estimated ploidy, CNV type, statistical scores to indicate confidence in called ploidy and CNV type, coverage information, and annotations of genes, repeats, and known CNVs found in Database of Genomic Variants (DGV) overlapping called segment. For CNV analysis of tumor genomes, this approach is modified such that the genome is segmented into region of distinct coverage level. Each segment is reported with the estimated coverage level, statistical score, and coverage information. In addition to these results, other key information for CNV analysis is provided in the CNV directory. Specifically, new features and outputs for CNV include:
   - Metric quantifying coverage variability across the genome has been added to the ***summary-[ASM-ID].tsv*** file. This can be found in the *100k Normalized Coverage Variability* field. If this value is above a defined cutoff of 0.05, the ploidy (for non-tumor genomes) or level (for tumor genomes) will be no-called for the entire genome. Segmentation results will still be provided.
   - Enhancements to the ***depthOfCoverage-100000-[ASM-ID].tsv*** file:
     i. Moved the ***depthOfCoverage-100000-[ASM-ID].tsv*** from the REPORTS directory to the CNV directory.
     ii. Added Baseline normalized coverage values in the *avgNormalizedCvg* field.
   - Results of CNV analysis for non-tumor and tumor genomes are reported in two files. They can be found in the CNV sub-directory of the ASM directory.
     i. ***cnvSegmentsBeta-[ASM-ID].tsv*** file reports segmentation of the complete reference genome into regions of distinct ploidy levels, giving the estimated ploidy, the average and relative adjusted coverage, confidence scores, and annotations for each segment.
     ii. ***cnvDetailsBeta-[ASM-ID].tsv.bz2*** file reports estimated ploidy, average and relative adjusted coverage, and confidence scores for every 2 KB along the genome.
     iii. ***cnvTumorSegmentsBeta-[ASM-ID].tsv*** file reports segmentation of the complete reference genome into regions of distinct coverage levels, giving the estimated coverage level, the average and relative adjusted coverage, and confidence scores for each segment.
     iv. ***cnvTumorDetailsBeta-[ASM-ID].tsv.bz2*** file reports estimated coverage level, average and relative adjusted coverage, and confidence scores for every 100 KB along the genome.
   - Enhancements to the ***geneVarSummary-[ASM-ID].tsv*** file:
     i. For each transcript reported in the ***geneVarSummary-[ASM-ID].tsv*** file, relative coverage of the CNV segments spanned by the transcript is reported in the *relativeCvg* field. If transcript spans more than a single CNV segment, relative coverage for all segments will be listed, separated by ";".

ii. For each transcript reported in the ***geneVarSummary-[ASM-ID].tsv*** file, ploidy of the CNV segments spanned by the transcript is reported in the *calledPloidy* field. If transcript spans more than a single CNV segment, ploidy for all segments will be listed, separated by ";". For tumor genomes, this column will be empty, as ploidy is not called for the identified segments.

3. Structural variation analysis has been added to our Assembly Pipeline as a Beta feature. The SV detection pipeline identifies regions of the genome that show evidence for structural alterations (characterized in Complete Genomics SV data as "junctions"). Junctions are identified by finding clusters of mate pairs that map to the reference genome at unexpected distance or orientation. Once a junction is detected, local de novo assembly is attempted on the junction to refine breakpoints to a single base pair resolution and to resolve the transition sequence, if one exists. Results from SV analysis are provided in the SV directory. Specifically, new features and outputs for SV analysis are:

   - Mean mate gap and the 95% Confidence Interval of mate gap distribution estimates for the sequenced genome have been added to the ***summary-[ASM-ID].tsv*** file. This can be found in the *Mate distribution mean* and *Mate distribution range (95% "CI")* fields, respectively.
   - Results from SV analysis are reported in the following four files:

     i. ***allJunctionsBeta-[ASM-ID].tsv*** file reports all junctions detected in the sequence genome, with associated information including genomic coordinates of breakpoints, number of discordant mate pairs supporting each junction, assembled transition sequence, and annotation of overlapping repeats elements, genes, and known indels in dbSNP.

     ii. ***highConfidenceJunctionsBeta-[ASM-ID].tsv*** file reports high-confidence junctions that are a subset of junctions found in the ***allJunctionsBeta-[ASM-ID].tsv***. Filtering criteria is applied to junctions in the ***allJunctionsBeta-[ASM-ID].tsv*** file. For a description of the filtering criteria, please refer to *Data File Formats*. Junctions that pass the filter are reported, with associated information, in the ***highConfidenceJunctionsBeta-[ASM-ID].tsv*** file.

     iii. ***evidenceJunctionDnbBeta-[ASM-ID].tsv.bz2*** file reports alignment of individual DNBs supporting each called junction.

     iv. ***evidenceJunctionClustersBeta-[ASM-ID].tsv*** file reports all junctions detected in the sequence genome, with associated information such as junction breakpoints and transition sequence length estimated from the initial clustering of discordant mate pairs (before these values are optimized by local de novo assembly).

4. New features and enhancements to ***gene-[ASM-ID].tsv.bz2*** file:

   - "SPAN" has been added to the *component* field to indicate that variant overlaps an entire exon.
   - In previous software versions, "UNKNOWN", "NO-CALL", and (Empty) in the *impact* field were used to indicate that functional impact of the variant cannot be determined for a variety of reasons. These values have been reassigned as "UNKNOWN-VNC", "UNKNOWN-INC", and "UNKNOWN-TR", respectively, to give more information to underlying reason for why functional impact is unknown. "UNKNOWN-VNC" indicates that impact is unknown due to the fact that one or more alleles have no-calls. "UNKNOWN-INC" indicates that impact is unknown due to lack of biological information. "UNKNOWN-TR" indicates that impact is unknown due to the transcript being rejected by our annotation pipeline.

5. Enhancements to ***coverageRefScore-[CHROMOSOME-ID]-[ASM-ID].tsv.bz2*** file:

   - Added GC bias corrected weight sum sequence coverage values in the *gcCorrectedCvg* field.

- Added gross weight sum sequence coverage values in the *grossWeightSumSequenceCoverage* field.

6. Enhancements to ***Summary.tsv*** file:
   - Added mean mate gap estimated for the library in the *Mate distribution mean* field.
   - Added range of mate gap that captures 95% of the data in the *Mate distribution range (95% "CI")* field.
7. Enhancement to the assembly process slightly increased call accuracy by reducing the number of half-calls.
8. Improvement to the mapping process reduced erroneous mappings to repeats regions. This change led to a reduction of coverage spikes in these repeats regions.

Fixed Issues

1. In previous software versions, if GRCh37 is used as the reference genome, variants found within PAR of Chr Y had an incorrect *varType* of "no-ref". This has been fixed such that *varType* is "PAR-called-in-X" for variants found in PAR of Chr Y.
2. Loci in the ***var-[ASM-ID].tsv.bz2*** file where the reference sequence is unspecified (such as at the beginnings and endings of chromosomes) are normally reported with a *varType* field value of "no-ref". In this software version, 11038 bases at the beginning of chromosome 1 where reference sequence is unspecified are not reported in the ***var-[ASM-ID].tsv.bz2*** file if GRCh37 was used as the reference genome. This has been fixed such that these loci are now reported with a *varType* field value of "no-ref".
3. In previous software versions, intronic variations in genes with coding region gave non-empty protein position. This has been fixed.
4. We only annotate non-reference alleles with dbSNP identifiers. In rare cases where RefSeq and reference genome sequences differ, annotation of only non-reference alleles with dbSNP identifiers can lead to counting reference calls as novel for the purpose of tabulation in ***geneVarSummary-[ASM-ID].tsv*** file. This has been fixed.
5. If there were two dbSNP entries that intersect a variant, *zygosity* fields for the entries in the ***dbSNPAnnotated-[ASM-ID].tsv.bz2*** file were incorrectly being reported as homozygous when they are supposed to be heterozygous. This has been fixed.
6. If a variant is found within *component* = "TSS-UPSTREAM" in the ***gene-[ASM-ID].tsv.bz2*** file, *impact* field is empty when it should be "UNKNOWN-INC". This has been fixed.
7. In previous software versions, coverage information for Chr 10 and Chr 20 were missing from the ***depthOfCoverage-[ASM-ID].tsv*** file. This has been fixed.
8. A few *hapA* and *hapB* fields of the ***dbSNPAnnotated-[ASM-ID].tsv.bz2*** file contained "ERROR" followed by the sequence of the A or B allele. We have eliminated the error condition that was causing this to happen and thus, the "ERROR" preceding allele sequence has been removed.

Known Issues

1. The *DeletedTransposableElement* field in the ***allJunctionsBeta-[ASM-ID].tsv, highConfidenceJunctionsBeta-[ASM-ID].tsv,*** and ***evidenceJunctionClustersBeta-[ASM-ID].tsv*** files only considers L1 and AluY subtypes with divergence at or below 2% when annotating junctions. However, all transposable elements, regardless of divergence level, should be considered when annotating junctions.
2. A small number of computationally predicted genes do not map fully to the reference genome and hence lack a start codon, a stop codon, or both. Annotations of variants found throughout

these genes are incorrect in that many are called *impact* = "MISSTART" or "NONSTOP" even though the gene lacks a start codon, a stop codon, or both.

3. No-call loci within block of 10 or longer no-calls are excluded from the ***gene-[ASM-ID].tsv.bz2*** file.
4. In rare cases where we are annotating a call that is close to a position where RefSeq and genomic reference sequence differs, the called amino acid sequence, reference amino acid sequence, and the functional impact reported for the call in the ***gene-[ASM-ID].tsv.bz2*** file may be incorrect.
5. In some cases, if variant matches variant in dbSNP, but not at coordinates listed in dbSNP, *found* field in ***dbSNPAnnotated-[ASM-ID].tsv.bz2*** file lists "N" when it should be "Y".
6. In a few cases, large differences between the genomic reference and RefSeq sequence within the Pfam domain lead to coordinate conversion difficulties and subsequent failure to annotate any part of the Pfam domain. We should annotate the portion of the Pfam domain that is consistent with the reference sequence. For GRCh37, the few cases are
   - Sulfotransfer_1 (pfam00685) hit on NP_006031.2 (corresponding to NM_006040.2)
   - 7tm_1 (pfam00001) hit on NP_001002905.1 (corresponding to NM_001002905.1)

   For Build 36:
   - Sulfotransfer_1 (pfam00685) hit on NP_006031.2 (NM_006040.2)
   - DUF1193 (pfam06702) hit on NP_064608.2 (NM_020223.2)
   - PARG_cat (pfam05028) hit on NP_003622.2 (NM_003631.2)
7. A small percentage of transcripts in Build 36 and GRCh37 are excluded from the annotation results due to the one or more of the following reasons: (1) contains unknown ("X") amino acid; (2) start and/or stop codon positions are unknown; (3) contains unspecified nucleotides; and (4) maps to unknown location/chromosome. To obtain the list of transcripts, please contact support@completegenomics.com.
8. For genes that partially map to the reference genome, 5' transcriptional start site is misidentified for a small set of genes (25 transcripts in Build 36 and 26 transcripts for GRCh37) in the ***gene-[ASM-ID].tsv.bz2*** file. As a result, variants are incorrectly annotated as falling within the TSS-UPSTREAM region (7.5 kb upstream of 5' transcriptional start site). To obtain a list of affected transcripts, please contact support@completegenomics.com.
9. Rarely, when an indel is found in RefSeq transcript with respect to the reference, that indel is applied to the reference sequence when determining the amino acid sequence reported in the *genomeRefSequence* field of the ***gene-[ASM-ID].tsv.bz2*** file. Thus, this field may contain non-reference sequence.
10. If there is a frameshift in the reference genome with respect to RefSeq, the reference amino acid sequence is reported in the *genomeRefSequence* field of the ***gene-[ASM-ID].tsv.bz2*** as if the frameshift had not occurred.
11. For a few transcripts in which alignment information cannot be parsed, *impact* field in ***gene-[ASM-ID].tsv.bz2*** file will be annotated with "UNKNOWN-TR".
12. Predicted genes without stop codon are not parsed correctly, leading to annotation of the variant with "UNKNOWN-TR" in the *impact* field in ***gene-[ASM-ID].tsv.bz2*** file.
13. Because COSMIC does not provide a transcript version number, COSMIC annotation in the gene file is copied over from the *xRef* field of the variation file that is based on genomic coordinate. Thus, the transcript described in the gene file may not be the transcript that is associated with the COSMIC record.

Addendum

The following issues were discovered after the release of Assembly Pipeline version 1.11:

1. Calls that overlap the stop codon for genes on the reverse strand and extend beyond the end of the stop codon were incorrectly called as "NONSTOP". This has been fixed.
2. On rare occasions, the ***var-[ASM-ID].tsv.bz2*** file reports a heterozygous locus where the top hypothesis is homozygous. This occurs when an allele can be aligned more than one way against the reference. This problem was fixed in Assembly Pipeline version 2.0.
3. In some cases, due to no-called alleles of the top hypothesis split into separate loci, the top hypothesis contains a ref allele and, the ***var-[ASM-ID].tsv.bz2*** file contains a locus where the alt call corresponds to the ref allele of the top hypothesis. This problem was fixed in Assembly Pipeline version 2.0.

Changes to Version 1.9.0

New Features and Enhancements

The following new features and enhancements are provided in this release by comparison with previous data shipped or released by Complete Genomics:

1. Seven files reporting various aspects of the sequence data have been added in the REPORTS folder within the ASM directory. Specifically, these seven files are:
   - ***coverage-[ASM-ID].tsv***: Reports number of bases in the reference genome covered (overlapped) by no reads, by one read, by two reads, etc. Two forms of coverage are computed and reported: uniquely mapping mated reads, and multiply mapping mated reads, appropriately weighted by a mapping confidence factor between 0 and 1 ("weight-sum" coverage).
   - ***coverageByGcContent-[ASM-ID].tsv***: reports normalized coverage across the spectrum of GC content seen in the genome. GC content is computed in 501-bp windows. A GC bin at the 1st percentile indicates that 1% of genomic bases have this or lower %GC. A GC bin at the 99th percentile indicates that only 1% of genomic bases have higher GC content. Normalized coverage over a large span of percentiles (a large proportion of the space 0..100, not lines in the file) indicate a relatively GC-unbiased library.
   - ***depthOfCoverage-[ASM-ID].tsv***: reports unique and weight-sum sequence coverage, along with GC bias-corrected weight-sum coverage for every 100 kb non-overlapping window along the sequenced genome.
   - ***indelLength-[ASM-ID].tsv***: reports number of insertions and deletions seen per length, such as the number of 1-base insertions and number of 2-base insertions.
   - ***indelLengthCoding-[ASM-ID].tsv***: reports number of insertions and deletions seen per length in the coding regions of the genome, such as the number of 1-base insertions and number of 2-base insertions.
   - ***substitutionLength-[ASM-ID].tsv***: reports number of substitutions seen per length.
   - ***substitutionLengthCoding-[ASM-ID].tsv***: reports number of substitutions seen per length in the coding regions of the genome.
2. A new file, ***ncRNA-[ASM-ID].tsv.bz2***, has been added to the ASM directory. This file reports variants that fall within mature microRNAs and pre-microRNAs identified in the miRBase sequence database.

3. New features and enhancements to ***var-[ASM-ID].tsv.bz2***:
   - Phasing information in *hapLink* field are available for many more variants as a result of using mate-pair information to deduce phase between neighboring variants.
   - Variants found in Catalogue of Somatic Mutation in Cancer (COSMIC) are annotated with COSMIC identifiers in the *xRef* column of the variation file. Format: COSMIC:<type>_<identifier>, where type indicates COSMIC classification of somatic variants. For example, "COSMIC:ncv_id:139111", where type indicates non-coding variant.
4. New features and enhancements to ***gene-[ASM-ID].tsv.bz2***:
   - *hasCodingRegion* field was changed from *codingRegionKnown* to more accurately reflect the information contained in the field.
   - Variants that fall within Pfam domains are annotated with Pfam identifier and domain name in a newly added *Pfam* field. Format: PFAM:<identifier>:<domain name>. For example, "PFAM:00069:Pkinase".
   - Variants found within the 7.5 kb upstream region of the 5' transcriptional start site are annotated as "TSS-UPSTREAM" in the *component* field.
   - Variants found in UTR, UTR and CDS, or CDS used to be annotated as EXON in the *component* field. EXON has been replaced by several new *component* values to be consistent with NCBI notation, and to give more precise and accurate information on where variants are found. New values include CDS for variants found in coding regions, UTR for variants found in non-coding genes, UTR5 for variants found in 5' untranslated region of coding genes, andUTR3 for variants found in 3' untranslated region of coding genes .
   - Variants that span exon boundaries are annotated with SPAN5 or SPAN3 in the *component* field, depending on whether they occur immediately before or after an exon, respectively. For example, insertions just before the first base or just after the last base would be annotated as SPAN5 and SPAN3, respectively. This is done to capture the uncertain impact of the variation (affecting coding sequence primarily, splicing primarily, or both).
5. The manifest file in the export package root directory provides sha256sum for all files written to the disk for each genome. Previous software releases (versions 1.6- 1.8) provided md5sum.

Fixed Issues

1. In the ***gene-[ASM-ID].tsv.bz2*** file, insertion of DNA sequence in multiple of 3 was being called "INSERT+" in the *impact* field without regards to the identity of the inserted codon. Thus, insertion of stop codon was incorrectly being called "INSERT+" instead of "NONSENSE". This has been fixed such that the codon represented by insertion or deletion of DNA sequence in multiple of 3 is being considered when assigning *impact* value.
2. In certain cases, assignment of *impact* field in ***gene-[ASM-ID].tsv.bz2*** file was based on amino acid changes relative to the reference genome sequence rather than the RefSeq sequence. This has been fixed such that assignment of "impact" is always based on amino acid changes relative to the RefSeq sequence.
3. In the ***dbSNPAnnotated-[ASM-ID].tsv.bz2*** file, the genome coordinates reported for the second allele of variants in haploid regions of genome (e.g., chrM, male non-PAR chrX) listed dummy value of "chr1, 0,0". The respective genomic coordinate fields for the second allele of variants in haploid regions are now left empty.
4. In previous software releases, it was indicated in our FAQs that gene symbols reported in ***gene-[ASM-ID].tsv.bz2*** and ***geneVarSummary-[ASM-ID].tsv*** files were taken from the ***seq_gene.md*** file that can be downloaded from NCBI. However, gene symbols were actually taken from a different XML file that is downloaded using the NCBI toolkit. We are now taking gene symbol information from the seq_gene.md file.

5. In the ***gene-[ASM-ID].tsv.bz2*** file, *nucleotidePos* field for non-coding transcripts where *impact* values were "UNDEFINED" was incorrect. The first haplotype of the first reported locus always had *nucleotidePos* value of 0, while the second haplotype had the correct *nucleotidePos* value. This initiated an off-by-one error, where the first haplotype of the second reported locus for the same non-coding transcript had the same *nucleotidePos* value as the second haplotype of the first locus. The second haplotype of the second locus then had *nucleotidePos* value of 0. This has been fixed.
6. Counting of introns for negative strand genes in the *componentIndex* field of the ***gene-[ASM-ID].tsv.bz2*** file was not zero-based. Thus, obtaining the correct count of the intron required a -1 adjustment. This has been fixed.
7. In ***var-[ASM-ID].tsv.bz2*** file, for variants where *varType* = "no-ref", *ploidy* value was reported as "?" in software versions. This has been changed such that *ploidy* = "2" for autosomal locus and pseudoautomsomal regions (PAR) sex chromosomes and *ploidy* = "1" for males on non-PAR region and mitochrondrion.
8. In previous software releases, variants found in non-coding transcripts were annotated in the ***gene-[ASM-ID].tsv.bz2*** file with *impact* field of "UNDEFINED" while *impact* of variants found in DONOR and ACCEPTOR components was left empty. Variants where *impact* was either left empty or annotated as "UNDEFINED" are now annotated as "NO-CALL" to be consistent with other situations where biological consequences of change cannot be determined.

Known Issues

1. We only annotate non-reference alleles with dbSNP identifiers. In rare cases where RefSeq and reference genome sequences differ, annotation of only non-reference alleles with dbSNP identifiers can lead to incorrect count of novel mutations in the ***geneVarSummary-[ASM-ID].tsv*** file. For example, consider a heterozygous A/G SNP at a give position within the sequenced genome where there is a dbSNP entry. Reference genome Build 36 has an A in this position, which results in a residue change in the protein T > M (with respect to the RefSeq sequence). Thus, this variant is called a novel missense mutation in the ***geneVarSummary-[ASM-ID].tsv*** file when in fact, the mutation is known.
2. If there are two dbSNP entries that intersects a variant, *zygosity* fields for the entries in the ***dbSNPAnnotated-[ASM-ID].tsv.bz2*** file are incorrect such that if both entries are supposed to be heterozygous, they will be reported as homozygous.
3. Indels affecting the start or stop codon are categorized as "FRAMSHIFT" in the *impact* field of the ***gene-[ASM-ID].tsv.bz2*** file rather than "MISSTART" or "NONSTOP".
4. Approximately 100 transcripts in build 36 and ~150 transcripts in GRCh37 are excluded from the annotation results due to the one or more of the following reasons: (1) contains unknown ("X") amino acid; (2) start and/or stop codon positions are unknown; (3) contains unspecified nucleotides; and (4) maps to unknown location/chromosome. To obtain the list of transcripts, please contact support@completegenomics.com.
5. For genes that partially map to the reference genome, 5' transcriptional start site is misidentified for a small set of genes in the gene-[ASM-ID].tsv.bz2 file. As a result, variants are incorrectly annotated as falling within the TSS-UPSTREAM region (7.5 kb upstream of 5' transcriptional start site). To obtain a list of affected transcripts, please contact support@completegenomics.com.
6. Loci in the ***var-[ASM-ID].tsv.bz2*** file where reference sequence is unspecified (e.g. at the beginnings and endings of chromosomes) are normally reported with a *varType* field value of "no-ref". In this software version, 11038 bases at the beginning of chromosome 1 where reference sequence is unspecified are not reported in the ***var-[ASM-ID].tsv.bz2*** file if GRCh37 was used as reference genome.

7. Rarely, when an indel is found in RefSeq transcript with respect to the reference, that indel is applied to the reference sequence when determining the amino acid sequence reported in the *genomeRefSequence* field of the ***gene-[ASM-ID].tsv.bz2*** file.
8. If there is a frameshift in the reference genome with respect to RefSeq, the reference amino acid sequence is reported in the *genomeRefSequence* field of the ***gene-[ASM-ID].tsv.bz2*** as if the frameshift had not occurred.
9. For NCBI Build 36, variants in PAR in ChrY are annotated with *varType* = "PAR-called-in-X". For GRCh37, variants in PAR in ChrY are annotated with *varType* = "no-ref".
10. For a few transcripts in which alignment information cannot be parsed, *impact* field in ***gene-[ASM-ID].tsv.bz2*** file will be annotated with "PARSE-ERROR".
11. Predicted genes without stop codon are not parsed correctly, leading to annotation of the variant with "PARSE-ERROR" in the *impact* field in ***gene-[ASM-ID].tsv.bz2*** file.
12. If variant is found within *component* = "TSS-UPSTREAM" in the ***gene-[ASM-ID].tsv.bz2*** file, *impact* field is empty when it should be "NO-CALL".
13. If GRCh37 is used as the reference genome, variants found within PAR of Chr Y have incorrect *varType* of "no-ref". The *varType* should be "PAR-called-in-X", as reported if NCBI Build 36 was used as the reference genome.
14. Because COSMIC does not provide a transcript version number, COSMIC annotation in the gene file is copied over from the *xRef* field of the variation file that is based on genomic coordinate. Thus, the transcript described in the gene file may not be the transcript that is associated with the COSMIC record.

Changes to Version 1.8.0

New Features and Enhancements

The following new features and enhancements are provided in this release by comparison with previous data shipped or released by Complete Genomics:

1. Customers can choose either NCBI build 36 or GRCh37 as the reference genome. The most recent RefSeq annotations for each build (NCBI annotation builds 36.3 and 37.1 respectively) were used for annotation.
2. dbSNP annotations are from build 130 for genome build 36 and from build 131 for genome GRCh37.
   - The format is: dbsnp.[build first seen]:[rsID], with multiple entries separated by the semicolon (;). For example, "dbsnp.129:rs12345".
   - Prior to version 1.8, we provided dbSNP 129 annotations for Build 36.
3. We have moved the version file from top-level directory to the individual genome results directory (for example "GS00001-DNA-A01").
4. Several improvements were made to the variations file:
   - Renamed *haplotype* column to *allele* in variant file header.
   - Every dbSNP annotation has been amended to contain the dbSNP version number for when that SNP was added to the database. This can be helpful for filtering novel SNPs from different dbSNP database releases.

5. Several improvements were made to the gene annotation files:
   - We have renamed ***gene-var-summary.tsv*** file to ***geneVarSummary.tsv*** for consistency with other files.
   - Renamed several columns in the ***gene-[ASM-ID].tsv.bz2*** file:
     i. *exonCategory(category)* to *component*
     ii. *exon* to *componentIndex*
     iii. *aaCategory* to *impact*
     iv. *aaAnnot* to *annotationRefSequence*
     v. *aaCall* to *sampleSequence*
     vi. *aaRef* to *genomeRefSequence*
   - In Version 1.7.1, we stopped annotating effects of variations for 476 genes in the ***gene-[ASM-ID].tsv.bz2*** and ***gene-var-summary-[ASM-ID].tsv*** files. These genes were affected by exonic indels in build 36 with respect to RefSeq sequence, a situation that led to incorrect frameshift calls in earlier versions of our software. Rather than report these erroneous frameshifts, annotations for these genes were suppressed. This situation is now properly handled by our annotation software, and therefore annotations for these 476 genes have been reintroduced.
   - For genes with standard initiation codons (per RefSeq curation), we have modified the annotation to ensure non-standard initiations are not recognized. Previous releases recognized the following non-standard start codons for all genes: TTG & CTG. For genes with non-standard initiations, (per RefSeq curation; for example, TEF-5 http://www.ncbi.nlm.nih.gov/nuccore/148277074), we do allow alternative start codons.
   - Previously splice sites were annotated only by intron/exon boundaries. We now annotate splice sites as DONOR and ACCEPTOR sites, as well as potential impacts when the variation overlaps the 2 conserved intronic bases immediately adjacent to the intron/exon boundary. If conserved GT/AG, or rare AT/AC becomes something incompatible, variation is annotated as "DISRUPT" in the *impact* column of the ***gene-[ASM-ID].tsv.bz2*** file. The *impact* column is left empty if the variation in donor and acceptor sites does not overlap the 2 conserved intronic bases immediately adjacent to the intron/exon boundary.
   - For *component* = "DONOR" or "ACCEPTOR", the following interpretations are applicable:
     i. *nucleotidePos* represents boundary between exons where the splice site is mapped to nucleotide sequence.
     ii. *proteinPos* represents boundary between exons where the splice site is mapped to protein sequence.
     iii. *sampleSequence* represents the sequence of splice site donor or splice site acceptor region for this allele after modification.
     iv. *genomeRefSequence* represents sequence of splice site donor or acceptor regions for this allele before modification.
   - The numbering of exons is now adjusted for strand, using 0-base numbering. In addition, exon numbering of UTR regions has been fixed; previously all UTRs were labeled "0".
   - In the ***gene-[ASM-ID].tsv.bz2*** file, we have added a *symbol* column indicating the NCBI Gene Symbol, for example "GAPDH".
6. The documentation has been updated and the data file format version number has been incremented to 1.3 to reflect the changes above.

Known issues

1. Approximately 100 transcripts in build 36 and ~150 transcripts in GRCh37 are excluded from the annotation results due to the one or more of the following reasons: (1) contains unknown ("X") amino acid; (2) start and/or stop codon positions are unknown; (3) contains unspecified nucleotides; and (4) maps to unknown location/chromosome. To obtain the list of transcripts, please contact support@completegenomics.com.

Changes to Version 1.7.4

1. We improved the base calling algorithm which resulted in more high quality calls.

Changes to Version 1.7.3

1. We are no longer including output from our beta-CNV algorithm (introduced in 1.7.1) as we continue development, validation and performance tuning of those methods. We expect to release an updated version in the near future.

Changes to Version 1.7.2

1. We have added a new field to the ***evidenceDnb*** file (*FileNumInLane*) to make it easier for customers to link reads and mappings to records in the evidence files. This does mean that any programs written to parse the ***evidenceDnb*** file will need to be changed.
2. We have added a new calculation to the ***coverageRefScore*** file (*weightSumSequenceCoverage*) that reflects coverage of each base in the reference genome factoring in reads which may not map uniquely. This is by contrast with the unique sequence coverage previously and still included in the ***coverageRefScore*** file. We find that the weighted-sum metric is best used in quantitative copy number calculations, for example. It also reflects many reads which can be recruited into de novo assembly. However, please do recall that both of these metrics are computed prior to assembly, and reflect the initial rather than final determination of read placements.
3. We have removed the DOC (documentation) directory from the customer deliverable. In previous versions, *Data File Format* (DataFileFormat.pdf) was included in the DOC directory. This document is now available from support@completegenomics.com.
4. The documentation has been updated and the data format version number has been incremented to 1.1 to reflect the changes above.

Changes to Version 1.7.1

These changes appeared in version 1.7.1 data releases by comparison with earlier versions.

1. We added a gene variation summary report (for example: ***gene-var-summary-GS19240-ASM.tsv.gz***) in the ASM folder. For each protein-coding gene, this file summarizes the numbers of variations with certain functional impacts, such as counts of nonsynonymous SNPs, and possible frameshifts.

2. We added a **BETA** quantitative copy number (CNV) computation. The quality of these CNV calls has not yet been extensively validated, and the exact performance in terms of sensitivity and specificity remains under study. However we believe these initial results may be of use to customers. We will likely modify the CNV calculation over time as we continually improve it, and we request customer feedback on these beta results.
   - Currently, CNV calls are based solely on increases or decreases in mapping rates (mapped coverage) normalized by various factors. Due to fluctuation in coverage owing to phenomena other than CNVs, results are currently limited to putative copy number changes affecting regions of approximately 20 KB or more. CNV reporting is provided in two tab-separated tables: the CNV segmentation table and the CNV details table. Supporting evidence for these CNV calls is provided via mappings and coverage data provided in other files.
     i. CNV segmentation table: Provides a segmentation of the complete reference genome into regions of various ploidy levels, giving the estimated ploidy, the average adjusted coverage for each segment, and measures of confidence in the called segments.
     ii. CNV details table: Provides information on estimated ploidy every 2 KB along the genome, giving average coverage and details regarding the estimated likelihood of each of various possible ploidy levels.
3. Several file format improvements were made from previous versions.
   - We renamed the variation types “ref-consistent” and “ref-inconsistent”. There is no change to semantics of each variation type, although by changing the name we wish to highlight the fact that these represent cases where the assembler was not able to fully resolve the allele sequence:
     i. “Ref-consistent” was renamed to “no-call-rc” (no-call reference consistent) – where one or more bases are ambiguous, but the allele is potentially consistent with the reference.
     ii. “Ref-inconsistent” was renamed to “no-call-ri” (no-call reference inconsistent), where one or more bases are ambiguous, but the allele is definitely inconsistent with the reference.
   - We renamed homozygous reference calls to “ref” rather than “=”, although “=” continues to indicate the reference allele in these ref-called regions.
   - We updated the headers of the variation and annotation files to include a reference to the specific version of the external reference data source used, such as the dbSNP version, Genome Reference sequence, or RefSeq gene annotations used.
   - We changed a number of file names to ensure that all files get unique names within an assembly and between samples, so they remain unique even if files are moved. The name change makes it easier to reorganize the data hierarchy or gather various data subsets.
   - The “chunk” numbers appended onto the mapping and reads files now have leading zeros (“_001” for example).
   - We removed a column in the ***Summary*** file describing the score threshold set used. Genomes produced using a specific version of the assembly pipeline always use the same threshold set, and only one set (this has been true for some time) so this column was extraneous. The technical documentation we are preparing on the analysis process will describe these thresholds in a more user-friendly way.
   - We renamed some of fields in the file headers for clarity:
     i. #BUILD to #SOFTWARE_VERSION
     ii. #VERSION to #FORMAT_VERSION
   - We fixed a minor bug where the “=” allele was not always output in the corresponding column of the variations file in haploid regions. This bug did not affect the results, only the exact syntax of how such calls were reported.

4. We added empirically measured gap information, per library, in a set of new files included in the LIB folder. The LIB folder includes one directory per library and one set of files per library directory (a genome is sequenced from a single library in Complete Genomics current process). Gap distribution information is useful for mapping, assembly and variant calling of the read data. It is also useful in discordant paired-end analyses to look for putative structural variants. Note that these new data files replace the ***lib_**** files previously included in the MAP folder subdirectories.
5. All data files except ***readme.txt*** and ***DataFileFormat.PDF*** are now compressed using bzib2 (.bz2 extensions) rather than gzip (.gz extensions). Be aware that bzip2 can be slower at decompressing than gzip, however the space savings and improved file transfer times were considered helpful by many.
6. The file format version number has been incremented to 1.0 to reflect these changes. We recommend that any code customers or partners write should check this number on any data file(s) read to ensure that the program is compatible with the data file(s).
7. The README.txt and DataFileFormats.PDF document have been updated to reflect the changes above. We have also added this release notes file.

Addendum

1. In Version 1.7.1, we stopped annotating effects of variations for 476 genes. This annotation is provided in the "aaCategory" column of the ***gene-[ASM-ID].tsv.bz2*** file. For these genes, mismatch between RefSeq sequence that we used for the annotation and the reference genome lead to incorrect annotation of the variation (e.g. declaring variation incorrectly as frameshift). Therefore, we stopped annotating variations found in these genes and will address this issue in future software version release. These 476 genes are also excluded from the ***gene-var-summary-[ASM-ID].tsv*** file. To obtain a list of the excluded 476 genes, please contact support@completegenomics.com.
2. An additional impact was added to the ***gene-[ASM-ID].tsv.bz2*** file in release 1.7.0. The additional impact added was "MISSTART".
    - MISSTART: The DNA sequence for this transcript has changed and has resulted in the change of a START codon into a codon that codes for an incompatible start codon resulting in a non-functional gene.

Changes to Version 1.6

These changes appeared in version 1.6 data releases by comparison with earlier versions.

1. Mapping and reads files in the subdirectories of the MAP folder have been broken into "chunks" in order to keep their sizes <5 GB per file. This allows compatibility with storage systems (such as certain cloud storage providers) for which 5 GB is an upper file size limit.
    - To accomplish this, we limit the number of reads described in any one mapping+reads file pair to 30 million. Mapping and reads files remain paired 1:1, and numbers are appended on the end of the mappings and reads files (such as "_1", "_2", "_3") to indicate the files which should be processed together.
    - The offset indexes of reads provided in the ***evidenceDNB*** files now have a particular interpretation. When this number is less than 30,000,000 the reads are in the first chunk (the mapping and reads files with "_1") at this 0-based position (data line) in that file. When the number is 30,000,000 to 59,999,999, the reads are in the second chunk ("_2"), with an offset position in that file of 30,000,000 less than the index provided. When the number is

greater than 60 million, the reads are in the third chunk and 60M should be subtracted from the index to get the position.

2. Subdirectories of the MAP folder are now named by slide and lane, rather than by an arbitrary mapping job number. This makes it easier to find reads and mappings based on knowing the slide and lane information, such as is used in the ***evidenceDNB*** files.
3. Previously, reads were filtered (not stringently) before inclusion into a customer data set. The stringency of these filters has been further reduced as we find doing so provides additional information that can improve accuracy of some variant calls without a significant impact on false positives. Furthermore, providing a more complete set of reads can facilitate reanalysis of the read-level data using various methods. As a consequence, customers should expect to see somewhat lower rates of map-ability as these additional reads are included—the rate has not actually gone down just the number of non-mapping reads included has increased.
4. Updates to documentation accordingly.

Changes to Version 1.5

Changes in 1.5 by comparison with earlier data releases.

1. Improvements were made in the variant calling algorithm that provide better accuracy of calls in duplicated regions and low copy number repeats. The variant scores now factor in uniqueness of evidence to further reduce false positives in such regions. Customers can consult the correlation file in the EVIDENCE folder to the underlying scores used in this calculation.
2. The assemblies and read alignments underlying all called variant regions are now provided in the EVIDENCE folder.
3. Updates to documentation accordingly.

Changes to Version 1.4

Changes in 1.4 by comparison with earlier data releases.

Numerous changes were made between version 1.3.x of the assembly pipeline software, as used in our data submitted to SRA as part of the Drmanac et al. publication (*Science*, Jan 2010 print edition). Among other changes, the C++ API is no longer required nor usable. Contact support@completegenomics.com for further information.