Release Notes for Genomes Processed Using Complete Genomics Software

Version 1.10.0

Related Documents

Consult the *Data File Formats* (DataFileFormats.pdf) 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 CGI documentation applies.

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:
    a. 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.
    b. 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.
    c. 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.
    d. 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:
    a. 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.
    b. 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:
    a. "SPAN" has been added to the *component* field to indicate that variant overlaps an entire exon.
    b. In previous software version, "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:
    a. Added GC bias corrected weight sum sequence coverage values in the *gcCorrectedCvg* field.
    b. Added gross weight sum sequence coverage values in the *grossWeightSumSequenceCoverage* field.

6. Enhancements to ***Summary.tsv*** file:
   a. Added mean mate gap estimated for the library in the *Mate distribution mean* field.
   b. Added range of mate gap that captures 95% of the data in the *Mate distribution range (95% "CI")* field.
7. Enhancements to the assembly process slightly increased call accuracy by reducing the number of half-calls.
8. Improvements 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 GRCh Build 37 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 GRCh Build 37 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 Build 37, 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 Build 37 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 Build 37) 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.

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:
    a. ***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).
    b. ***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.
    c. ***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.
    d. ***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.
    e. ***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.
    f. ***substitutionLength-[ASM-ID].tsv***: reports number of substitutions seen per length.
    g. ***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***:
    a. 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.
    b. 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***:
    a. *hasCodingRegion* field was changed from *codingRegionKnown* to more accurately reflect the information contained in the field.
    b. 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".

c. Variants found within the 7.5 kb upstream region of the 5' transcriptional start site are annotated as "TSS-UPSTREAM" in the *component* field.

d. 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 .

e. 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 build 37 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 NCBI Build 37 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 NCBI Build 37, 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 NCBI Build 37 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 Genome Reference Consortium build 37 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 build 37.
    a. The format is: dbsnp.[build first seen]:[rsID], with multiple entries separated by the semicolon (;). For example, "dbsnp.129:rs12345".
    b. 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:
    a. Renamed *haplotype* column to *allele* in variant file header.
    b. 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:
    a. We have renamed ***gene-var-summary.tsv*** file to ***geneVarSummary.tsv*** for consistency with other files.
    b. 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*
    c. 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.

d. 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.

e. 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.

f. 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.

g. 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".

h. 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 build 37 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.
    a. 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.
   a. 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.
   b. We renamed homozygous reference calls to "ref" rather than "=", although "=" continues to indicate the reference allele in these ref-called regions.
   c. 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.
   d. 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.
   e. The "chunk" numbers appended onto the mapping and reads files now have leading zeros ("_001" for example).
   f. We removed a column in the ***Summary*** file describing the score threshold set used. Genomes produced using a specific version of the CGI 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.
   g. We renamed some of fields in the file headers for clarity:
      i. #BUILD to #SOFTWARE_VERSION
      ii. #VERSION to #FORMAT_VERSION
   h. 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".
    a. 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.

3. 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.
    a. 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.
    b. 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.
4. 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.
5. 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.

6. Updates to documentation accordingly.

Changes to version 1.5

Changes in 1.5 by comparison with earlier data releases.

7. 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.
8. The assemblies and read alignments underlying all called variant regions are now provided in the EVIDENCE folder.
9. 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 CGI 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.

