

RELEASE NOTES FOR GENOMES PROCESSED USING COMPLETE GENOMICS SOFTWARE VERSION 1.9.0.

CONTENTS

New features and enhancements.....	1
Changes to version 1.8.0	6
Changes to version 1.7.4	8
Addendum	8
Changes to version 1.7.3	8
Changes to version 1.7.2	8
Changes to version 1.7.1	8
Addendum	10
Changes to version 1.6	11
Changes to version 1.5	12
Changes to version 1.4	12

Customers should consult the DataFileFormats.pdf file corresponding to any data set for information specific to those data. Customers can check the header information in the files to determine which version of CGI documentation applies.

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).

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject.

support@completegenomics.com

Information, descriptions and specifications in this publication are subject to change without notice.

© Complete Genomics, Inc., 2010

Published in U.S.A., June 2010

- 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 (e.g. number of 1-base insertions, number of 2-base insertions, etc).
 - e. indelLengthCoding-[ASM-ID].tsv: reports number of insertions and deletions seen per length in the coding regions of the genome (e.g. number of 1-base insertions, number of 2-base insertions, etc).
 - 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, and UTR3 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).

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, “nucleotidePosition” field for non-coding transcripts where “impact” values were “UNDEFINED” was incorrect. The first haplotype of the first reported locus always had “nucleotidePosition” value of 0, while the second haplotype had the correct “nucleotidePosition” 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 “nucleotidePosition” value as the second haplotype of the first locus. The second haplotype of the second locus then had “nucleotidePosition” 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 pseudoautosomal regions (PAR) sex chromosomes and “ploidy” = “1” for males on non-PAR region and mitochondrion.
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. In dbSNPAnnotated-[ASM-ID].tsv.bz2 file, if dbSNP entry has more than one called variant overlapping the entry genomic coordinate range, “varType” field = “multiple”. When this is the case, “hap” field sometimes contains the “ERROR” string followed by the called sequence. Correct behavior is to just list the called sequence.
10. 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”.
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 “PARSE-ERROR”
12. 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.
13. 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”.
14. 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.
15. 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

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 (i.e. `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/nucore/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, e.g. 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.

ADDENDUM

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject.

support@completegenomics.com

Information, descriptions and specifications in this publication are subject to change without notice.

© Complete Genomics, Inc., 2010

Published in U.S.A., June 2010

1. We have made an addendum to Version 1.7.1 Release Notes. Please refer to the Addendum section (#2) for details of the change.

CHANGES TO VERSION 1.7.4

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

ADDENDUM

2. We have made an addendum to Version 1.7.1 Release Notes. Please refer to the Addendum section for details of the change.

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 (FileNumLane) 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, the 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

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject.

support@completegenomics.com

Information, descriptions and specifications in this publication are subject to change without notice.

© Complete Genomics, Inc., 2010

Published in U.S.A., June 2010

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 (counts of nonsynonymous SNPs, possible frameshifts, etc.).
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 20kb 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.
 1. 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.
 2. CNV details table: Provides information on estimated ploidy every 2kb 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:
 1. 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.
 2. 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 e.g. dbSNP version, Genome Reference sequence (and RefSeq gene annotations) used, etc.

- 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. This provides customers more convenience if they wish 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 CGI pipeline always use the same threshold set, and only one set (this has been true for some time) so this column was extraneous. 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.
 1. #BUILD to #SOFTWARE_VERSION
 2. #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. In that folder is one directory per library, and one set of files per library directory (normally, 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 Data File Format.PDF are now compressed using bzip2 (.bz2 extensions) rather than gzip (.gz extensions). Customers should 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

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject.

support@completegenomics.com

Information, descriptions and specifications in this publication are subject to change without notice.

© Complete Genomics, Inc., 2010

Published in U.S.A., June 2010

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.

1. Mapping and reads files in the subdirectories of the MAP folder have been broken into “chunks” in order to keep their sizes <5GB per file. This allows compatibility with storage systems (such as certain cloud storage providers) for which 5GB 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 (eg 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, etc.
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 mappability 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 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. Customers can contact support@completegenomics.com for further information.