CGA™ Tools Release Notes

Software Version 1.7.1

Related Documents

Customers should consult the *CGA™ Tools User Guide* for detailed information on specific tools. This document can be downloaded at: http://cgatools.sourceforge.net.

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

Changes to Version 1.7

New Features and Enhancements in Version 1.7

The following new features and enhancements are provided in this release by comparison with previous CGA Tools released by Complete Genomics:

1. This release is compatible with the new information and file formats produced by Complete Genomics Analysis Pipeline 2.4. The new information is described in *Data File Formats*.
2. Tools that output ***var*** and ***masterVarBeta*** file formats now output those files using the file formats introduced in Analysis Pipeline 2.4. For example, variant quality filters are provided in *varFilter* column of the ***var** file* instead of the *varQuality* column. Similarly, in ***masterVarBeta*** the *alleleXVarFilter* columns replace the *alleleXVarQuality* and *alleleXSomaticQuality* columns. The output files also contain new columns introduced in Analysis Pipeline 2.4, such as *alleleFreq* and *alternativeCalls* (for example in the ***var*** file), though these columns will be empty if the input files were analyzed on Analysis Pipelines before version 2.4. These changes affect output from the following tools: snpdiff, calldiff, generatemasterVar, and varfilter.
3. The following changes have been made to generatemasterVar (beta):
   - generatemasterVar can output the new *fisherSomatic* score (a score introduced in Analysis Pipeline 2.4 that measures confidence in called somatic variants, complementing *somaticScore*) and the new FET30 flag (which indicates lower confidence in the somatic call). These cancer-related features require the source genome to have been sequenced using Complete Genomics Cancer Sequencing Service.
   - generatemasterVar now sources nonsomatic relative coverage (both *relativeCoverageDiploid* and *relativeCoverageNondiploid*) from the overlapping CNV window, not the overlapping CNV segment.
4. The following changes have been made to mkvcf:
   - mkvcf output supports the enhancements to Analysis Pipeline 2.4. These include:
     a. Single-sample LAF calculations using the CGA_LAFS, CGA_LLAFS, and CGA_ULAFS tags in the FORMAT column.

     b. The *fisherSomatic* score via the CGA_SOMF tag in the FORMAT column. The new FET30 flag is provided via the FT tag of the FORMAT column.

     c. The *alleleFreq* minor allele frequency information via the AF tag of the INFO column.

     d. The AMBIGUOUS call designation via the FT tag of the FORMAT column. Alternate calls and their relative scores are provided via the new CGA_ALTCALLS tag in the FORMAT column.
   - To eliminate confusion between single sample LAF measurements and paired-sample measurements provided in the Cancer Service, the paired-sample tag names have been appended with a “P” suffix for “paired”, as follows:

     a. CGA_LAF is now CGA_LAFP

     b. CGA_LLAF is now CGA_LLAFP

     c. CGA_ULAF is now CGA_ULAFP

     These new column names are fixed, thus paired-sample LAF calculations will always be produced using these tag names, regardless of the version of the Analysis Pipeline used to generate the input files.

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. generatemasterVar can output a ***masterVarBeta*** file containing the new *locusDiffClassification* and *varFlags* columns (for a baseline sample in a cancer pair or trio), but this feature is not supported for genomes sequenced on Analysis Pipelines before version 2.4.

Changes to Version 1.6

New Features and Enhancements in Version 1.6

The following new features and enhancements are provided in this release by comparison with previous CGA Tools released by Complete Genomics:

1. Added mkvcf tool that translates Complete Genomics small variant, CNV, SV, and MEI calls to VCF 4.1 format. This tool can be used to convert all variant calls for a single genome assembly to VCF format, and to compare variant calls across multiple genome assemblies.

   VCF conversion of certain variant calls is only available for more recent pipeline versions, and certain variant calls are currently not supported when generating multi-genome VCFs. For more information see the “Compatibility” section for the mkvcf tool in the *CGA™ Tools User Guide*.

Fixed Issues

1. map2sam has been removed from CGA Tools to reduce user frustration with downstream analysis of map2sam-converted data.
2. Fixed a bug where using evidence2sam with the `--keep-duplicates` option produced no output.
3. calldiff now checks for a valid `--calibration-root` directory before performing calculations instead of providing an error when outputting the results.
4. Changes to improve testvariants memory usage.
5. Fixed error where junctions2events does not identify “inversion” events > 1kb.
6. Fixed bug converting alignment to SAM CIGAR strings for DNBs where overlapping base calls overlap an insertion.
7. Removed padding (P) from the ends of SAM CIGAR strings. Previous versions of evidence2sam used the CIGAR string “P” symbol to mean mate gap bases or bases beyond the clone end that correspond to an insertion against the reference. This use of padding enables calculation of the length of the insertion that a read supports. The use of padding in this context has now been removed to improve compatibility with SAM-processing tools.

Changes to Version 1.5

New Features and Enhancements in Version 1.5

The following new features and enhancements are provided in this release by comparison with previous CGA Tools released by Complete Genomics:

1. For CGA Tools outputs with header information, data file format version ('FORMAT_VERSION') has been changed to 2.0 and the metadata header SOFTWARE_VERSION will now always be the version of the Assembly Pipeline.
2. With CGA Tools 1.5, support has been added for processing data created with the Complete Genomics Assembly Pipeline version 2.0. This includes support for the new columns of the ***var*** file and ***masterVarBeta*** file, such as *varScoreVAF* and *varScoreEAF*. In order to support Assembly Pipeline version 2.0 and previous Assembly Pipeline versions, CGA Tools translates any older ***var*** file or ***masterVarBeta*** file. Specifically, the *varScoreVAF* and *varScoreEAF* columns are populated by the *totalScore* from the older input file. Additionally the *varQuality* field added in Assembly Pipeline 2.0 remains empty. The corresponding translation is also done for previous versions of the ***masterVarBeta*** file.
3. CGA Tools listvariants has been changed so that indels are reported at the left-shifted canonical form. This is in accordance with the VCF preferred canonicalization, as well as the canonicalization in Complete Genomics Assembly Pipeline 2.0. To avoid re-canonicalization of a large number of variants when running listvariants, we recommend that comparison of ***var*** or ***masterVarBeta*** files from Pipeline versions earlier than 2.0 be performed using listvariants in CGA Tools 1.4.
4. Added var file filtering syntax wherever a ***var*** file or ***masterVarBeta*** file is taken as input. This syntax allows users to specify a comma-separated list of filters, turning calls that pass a defined set of filters into no-call.
5. Added varfilter command. This allows users to create a filtered list of variations that can be further analyzed.
6. The following changes have been made to the generatemasterVar (beta) output:
    - Two new annotation types are available through the -- annotations argument for inclusion into the ***masterVar*** output: *cnvDiploid* and *cnvNondiploid*. If *cnvDiploid* is specified, the diploid CNV annotations (*calledPloidy* and *relativeCoverage*) are added. If *cnvNondiploid* is specified, the non-diploid CNV annotations (*calledLevel* and *relativeCoverage*) are added. If cnv is specified, either the *cnvDiploid* annotation is applied, or the *cnvNondiploid* annotation is applied, or both are applied if your genome has both CNV outputs. Note that prior to Assembly Pipeline version 2.0, only one of the two CNV outputs is present.
7. The following changes have been made to map2sam and evidence2sam tools:
    - DNBs where only one arm maps are now assigned as primary alignments.
    - For reads with an unmapped mate, we are now assigning a mate chromosome of "*" and template length of 0. Previously, we were assigning a mate position of 0 on this chromosome.
    - The mate chromosome, position and template length for secondary mappings, are now provided.
    - The `--pack-cigar` option is no longer available, as `--pack-cigar` is the default and only supported behavior.
    - Each DNB has at least one primary mapping reported but not more than one per mate.

- Records generated for unmapped mates get chromosome and position of the mapped mate.

8. The following changes have been made to calldiff (beta) output:

- SomaticOutput option

  a. Somatic variation scoring has changed in CGA Tools 1.5. For detailed information, refer to the “calldiff for Scoring Somatic Variations (beta)” Appendix section of the *CGA Tools User Guide*. Briefly, the somatic score from CGA Tools 1.4 is comparable to the *somaticRank* from CGA Tools 1.5. The new *somaticScore* is a Phred-like score indicating the log likelihood ratio of likelihood of truth over likelihood of error.

  b. The *VarCvgA* and *RefCvgB* columns have been added to the Somatic Output file.

  c. The *VarScoreARank* and *RefScoreBRank* columns have been replaced by the *SomaticRank* column in the Somatic Output file.

Fixed Issues

1. Gaps in the reference and no-calls are now distinguished in the output file of generatemasterVar (beta).
2. In CGA Tools 1.4 junctions2events output file, *DestinationRegionLength* and *DestinationRegionEnd* columns were swapped. This is now fixed.

Known Issues

1. Performing multi-genome comparison with listvariants or testvariants may result in a few cases where variants are included in the multi-genome report but none of the samples are reported as containing that variant. This is caused by the fact that listvariants generates the right-most canonical representation of the variant. This representation of the variant is then subsequently used to compare between genomes. As a result, one of the following two situations may arise which causes testvariants to not report the variant call for a given sample:

   a. If the right-most canonical representation of an indel overlaps with another variant or a no-call then testvariants does not consider that indel as equivalent to the original call and, therefore, does not report the presence of that variant for a given sample.

   b. Currently there is a maximum limit of 50 bp that a superlocus can be extended by prefix and suffix matching. As a result, it is possible that in the right-most canonical representation of the variant generated by listvariants, the call itself will now be outside of the maximum superlocus size and, therefore, will not be detected and reported by testvariants.

2. In rare cases, a no-call next to a position of interest results in testvariants reporting a no-call instead of discordance for the position of interest, when discordance exists.

Changes to Version 1.4

Note: CGA Tools was referred to as "**cgatools** " in this and previous releases.

The following new features and enhancements are provided in this release by comparison with previous **cgatools** released by Complete Genomics:

1. Added junctions2events (beta) tool that identifies structural variation events from lists of junctions. Structural variation events are represented by a single junction (such as for deletions and tandem duplications) or multiple junctions (such as for inversions and translocation). junctions2events considers possible relationships among junctions in the input file and determines which event a junction or multiple junctions is consistent with.

   The tool produces two files:

   - ***Events*** file: Reports structural variation events deduced from input junctions file, along with annotations of genomic location, number of discordant mate pairs supporting the event, genes overlapping event breakpoints or wholly contained within event, and putative gene fusion.
   - ***AnnotatedJunctions*** file: Contains the original junctions of interest, annotated with the event type, list of related junctions, and the unique ID of the event.
2. Updated generatemasterVar (beta) as follows:
   - Changed *neitherAlleleReadCount* to *referenceAlleleReadCount*.
   - Included additional documentation about read count calculation.
   - Allowed *calledCNVType* to be “N” when genome is no-called.
   - Added two new values for *varType*: “no-ref” and “PAR-called-in-X”.
   - Created a new column called *pfam* which reports Pfam domain information. The *allele1Gene* and *allele2Gene* columns no longer report Pfam domain information.
3. Changed the *haplotype* column header in calldiff to *allele*.
4. Fixed the junctiondiff tool such that, when using the `--minlength` parameter, the ***report.txt*** file now accurately summarizes the number of length-filtered unique junctions (column: *filteredIncompatible*).
5. The following changes have been made to the evidence2sam output:
   - *CIGAR*: ‘N’ sections of 0 length were removed; neighbor commands of the same type were merged; overlapping part of neighbor combinations xIxD were replaced with ‘M’, in the case of xPxD with ‘N’.
   - *TAG*: Added a read group tag ‘RG’.
6. The following changes have been made to the map2sam output:
   - *CIGAR*: ‘N’ sections of 0 length were removed; neighbor commands of the same type were merged; overlapping part of neighbor combinations xIxD were replaced with ‘M’, in the case of xPxD with ‘N’.
   - *TAG*: Added a read group tag ‘RG’.
   - For unmapped reads, reference name and position are reported based on the mate information.
   - For unmapped alignments, mate reference name, position and strand are reported.
   - *FLAG*: The value for FLAG (0x2 each fragment properly aligned according to the aligner) is set only if the mate and the alignment are both mapped

7. Modified the help page for all tools (where relevant) as follows:
   - `--export-root` is now `--genome-root`
   - "export package" is now "genome directory"
   - `--export-region` is now `--extract-genomic-region`

Fixed Issues

1. In **cgatools** version 1.3, incorrect values were reported in the *readCounts* field for het-ref insertion calls in the master variation file created by generatemasterVar (beta). This has been fixed.

Known Issues

1. Performing multi-genome comparison with listvariants or testvariants may result in a few cases where variants are included in the multi-genome report but none of the samples are reported as containing that variant. This is caused by the fact that listvariants generates the right-most canonical representation of the variant. This representation of the variant is then subsequently used to compare between genomes. As a result, one of the following two situations may arise which causes testvariants to not report the variant call for a given sample:

   c. If the right-most canonical representation of an indel overlaps with another variant or a no-call then testvariants does not consider that indel as equivalent to the original call and, therefore, does not report the presence of that variant for a given sample.

   d. Currently there is a maximum limit of 50 bp that a superlocus can be extended by prefix and suffix matching. As a result, it is possible that in the right-most canonical representation of the variant generated by listvariants, the call itself will now be outside of the maximum superlocus size and, therefore, will not be detected and reported by testvariants.

2. On rare occasions, Pfam annotation is duplicated for a given locus in the ***masterVar*** file output from the generatemasterVar(beta) tool.

3. In this version of **cgatools**, map2sam is no longer compatible to the export of Assembly Pipeline versions 1.5 and 1.6. A workaround is available for customers with Assembly Pipeline versions 1.5 and 1.6 data who want to run map2sam. The following modifications to the datasets are required:
   - mkdir <SAMPLEDIR>/LIB
   - For each library, create mkdir <SAMPLEDIR>/LIB/<LIBNAME> and copy a dnbstructure file from a <SAMPLEDIR>/MAP/<LANE> into the corresponding <SAMPLEDIR>/LIB/<LIBNAME>. The <LIBNAME> is a part of the dnb structure file name: lib_DNB_<LIBNAME>.tsv.

Changes to Version 1.3

The following new features and enhancements are provided in this release by comparison with previous **cgatools** released by Complete Genomics:

1. Added generatemasterVar (beta) tool to create a simple, integrated master variation (***masterVar***) file that contains one line per locus and combines variation, annotation, and coverage information. The ***masterVar*** file serves as an aggregated source of information that

is formatted to be easily expanded to include custom annotations and processed with simple command-line tools, including searching and filtering. The ***masterVar*** file provides a structured content that can more easily be converted into other standard variation file formats.

2. Modified the calculation of the *SomaticScore* column in the calldiff somatic output file to use two newly added score ranking columns, *VarScoreARank* and *RefScoreBRank*. The new *SomaticScore* better handles somatic categories which only have a few variations. It is also able to capture potential somatic variations where the normal sample is no-called. Finally, with this new score, a selected cutoff will represent the same sensitivity when applied to different samples.
3. Changed the behavior of listvariants such that for longer 'subs', the range of the loci is trimmed when edge sequence matches the reference. This, potentially, results in the assignment of a different *varType* for variations that can be more explicitly defined.

Addendum

1. After the release, it was discovered that incorrect values were reported in the *readCounts* field for het-ref insertion calls in the master variation file created by generatemasterVar (beta).

Changes to Version 1.2

The following new features and enhancements are provided in this release by comparison with previous **cgatools** released by Complete Genomics:

1. Added junctiondiff (beta) tool.
2. Changed LocusDiffClassification for calldiff LocusOutput from alt-consistent to ref-consistent and from alt-identical to ref-identical, wherever the allele is consistent with the reference. In previous versions, the alt- classification was used whenever reference-consistent calls were aligned to calls of the other genome that were reference-inconsistent.

Changes to Version 1.1.1

1. Fixed crash in join tool when reading ***geneVarSummary*** file.
2. Added Mac OS X binary tarball.

Changes to Version 1.1.0

1. Enhanced the calldiff tool to allow users to identify somatic variants from a tumor/normal pair (beta). This tool takes two variant files—genome A (tumor) and genome B (normal)—as inputs and produces:
   - a report ("SomaticOutput") that lists variants found only in genome A
   - a score that indicates the likelihood that each variant is truly somatic
2. Added listvariants (beta) and testvariants (beta) tools for comparing variants across multiple genomes, allowing users to determine whether a variant was found in a given genome and the frequency of the variant across the set of tested genomes. listvariants generates a list of all fully called variants found in at least one genome within the tested set.

testvariants uses this list as input and reports for each variant whether the allele called in each genome is 1) inconsistent with the variant, 2) is fully called and is consistent with the variant, or 3) has no-calls and allele is consistent with the variant.

3. Added the join (beta) tool, allowing users to combine information from two tab-delimited files by specifying column(s) within the files to be used for determining overlap and column(s) from each file to be included in the merged file.
4. Changed parameters for snpdiff and calldiff tools. In previous releases of **cgatools**, separate parameters were required to output each report (such as Stats, Output, and SuperlocusOutput) to a specific location. These parameters were replaced with two new ones: a parameter that allows you to specify multiple reports to be output and a parameter that allows you to specify a path to the directory to which all output reports will be saved.

Changes to Version 1.0.0.15

1. Fixed snpdiff and calldiff failure processing male build 37 genomes.

Changes to Version 1.0.0.14

1. Fixed evidence2sam to be able to support genomes from assembly format version 1.0.

Changes to Version 1.0.0.13

1. Changed to dynamic linkage on Mac OS X.

Changes to Version 1.0.0

1. Renamed cgi2sam to map2sam.
2. Added evidence2sam (beta).

Changes to Version 0.5.0 (Initial Version)

1. The initial version of **cgatools** included:
   - Reference tools
   - snpdiff
   - calldiff
   - cgi2sam