



cgatools Release Notes

Version 1.4.0

Related Documents	1
New Features and Enhancements in Version 1.4	1
Changes to Version 1.3	3
Changes to Version 1.2	4
Changes to Version 1.1.1	4
Changes to Version 1.1.0	4
Changes to Version 1.0.0.15	5
Changes to Version 1.0.0.14	5
Changes to Version 1.0.0.13	5
Changes to Version 1.0.0	5
Changes to Version 0.5.0 (Initial Version)	5

Related Documents

Customers should consult the *cgatools Methods* document for detailed information on specific tools. This document can be downloaded at: <http://cgatools.sourceforge.net>.

New Features and Enhancements in Version 1.4

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.

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.

Copyright © 2011 Complete Genomics Incorporated. All rights reserved.

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 xlxD 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 xlxD 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

2. 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.
3. On rare occasions, Pfam annotation is duplicated for a given locus in the **masterVar** file output from the generatemastervar(beta) tool.
4. 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:
 1. mkdir <SAMPLEDIR>/LIB
 2. 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