



CNV, SV, and MEI Frequently Asked Questions (FAQ)

Updated September 2011

General Information	2
How do I assess the quality of the CNV, SV, or MEI data for my genome?	2
How do I get my delivered genome from Complete Genomics re-assembled on Complete Genomics' latest and greatest assembly software? How can I get updated CNV/SV data for the genomes that Complete Genomics has already sequenced and delivered to me?.....	3
Where can I find more information on Complete Genomics data and results?.....	3
Copy Number Variation (CNV) Analysis	4
What is the difference between the different CNV files, and how do I select which file to use for my samples?	4
How should I use the various coverage values provided in the depthOfCoverage and coverageRefScore files?	4
How do I interpret CNV types "hypervariable" and "invariant"? Should I consider them as candidate CNV regions?.....	5
I am comparing CNV data for two different genomes sequenced by Complete Genomics. How can a region be labeled with cnvType "invariant" in one genome and be labeled '+' in the other if assignment of "hypervariable" and "invariant" regions is based on the coverage profiles of the reference genomes set that is used to process both genomes?	6
What is "calledLevel" in my <i>cnvSegmentsNondiploidBeta</i> file and how does it relate to ploidy?	7
How do I identify segments that are amplified or reduced in my tumor genome?	8
I have matched tumor and normal samples. How do I get "somatic" CNV calling?.....	8
What are the limitations of using a single matched genome as the baseline for the paired analysis used to identify somatic CNVs?	8
How do I evaluate the confidence of a called CNV? What sorts of underlying evidence can I look at?	9
What does it mean when "CNVTypeScore" is much greater than "ploidyScore"?.....	9
How do CNV and SV results relate to one another? Would structural variation events that involve changes in copy number be reported in both CNV and SV files?.....	9
Why did Complete Genomics miss a known event or an event that is obvious from the raw data in my sample?	9
Can I get access to the reference data set used to generate the baseline?	9

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.

What is the Lesser Allele Fraction (LAF), and where can I find it?	10
How do I interpret the Lesser Allele Fraction (LAF)?	10
Structural Variation (SV) Analysis	11
What is a junction?	11
Does Complete Genomics indicate the structural variant type represented by a junction in SV files?	12
Are somatic events identified for my tumor-normal pair?	12
Does Complete Genomics indicate zygosity of the junction?	12
Are the indels reported in the var-[ASM-ID].tsv file also reported in the junctions files?	12
The <i>allJunctionsBeta</i> , <i>highConfidenceJunctionsBeta</i> , <i>somaticAllJunctionsBeta</i> , <i>somaticHighConfidenceJunctionsBeta</i> , and <i>evidenceJunctionClustersBeta</i> files have the same file format. What are the differences between these files?	12
How do I evaluate the confidence of a called junction? What sorts of underlying evidence can I look at?	13
Are there limitations on the classes of junctions Complete Genomics can discover?	13
Can I get access to the reference data used to create the SV baseline?	14
Mobile Element Insertion Detection	14
Are MEIs detected using the same method as SV detection?	14
What is the resolution of insertion site detection? Does Complete Genomics assemble the insertion site?	14
What MEI type does Complete Genomics detect?	14
How should I filter for high-confidence MEIs?	14
Does Complete Genomics identify somatic MEIs?	15
Is the zygosity of events reported?	16
Can I get access to the reference data used to create the MEI baseline?	16

General Information

How do I assess the quality of the CNV, SV, or MEI data for my genome?

The summary file (*summary-[ASM-ID].tsv*) contains a variety of CNV, SV, and MEI metrics that one would expect to be roughly consistent across genomes from individuals of the same ethnicity or disease type. These metrics can be quite useful for quality assessment. They include:

- Total CNV segment count
- Total number of bases in CNV segments
- Fraction of novel CNV (by segment count)
- Fraction of novel CNV (by base count)
- Total junction count
- Mobile element insertion count
- Fraction of novel MEI

Please note that while the application of these and other metrics to normal diploid genomes is relatively clear, correctly interpreting these and similar calculations for a cancer or non-diploid genome can be more difficult.

In addition, the REPORTS directory includes several files reporting various aspects of the sequence data that can be used to assess the quality of the delivered genome. For example:

- ***circos-[ASM-ID].html*** and ***circos-[ASM-ID].png*** (also: ***somaticCircos-[ASM-ID].html*** and ***somaticCircos-[ASM-ID].png***): visual summary of small and large variation data for each genome; includes junctions, called level or called ploidy, and Lesser Allele Fraction (LAF).
- ***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). With this information, you can create a plot of genome-wide coverage distribution. For standard-coverage genomes, you would expect the mean coverage to be at least 40, and for high-coverage genomes the mean coverage would be at least 80.
- ***coverageCoding-[ASM-ID].tsv***: Reports same information as ***coverage-[ASM-ID].tsv*** for only the coding regions of the reference genome.
- ***coverageByGcContent-[ASM-ID].tsv***: Reports normalized coverage for cumulative GC base content percentile, allowing you to assess the level of GC bias across the genome.
- ***coverageByGcContentCoding-[ASM-ID].tsv***: Reports normalized coverage for cumulative GC base content percentile, allowing you to assess the level of GC bias across the exome.

How do I get my delivered genome from Complete Genomics re-assembled on Complete Genomics’ latest and greatest assembly software? How can I get updated CNV/SV data for the genomes that Complete Genomics has already sequenced and delivered to me?

Customers can order reanalysis of a new genome from assembly pipeline version 1.5.0 or later, using part number REA000RNM. Reanalysis is available using assembly pipeline versions 1.10 and later, and customers have the option to indicate whether they prefer a specific version or would prefer reanalysis on the most current assembly pipeline version at the time of processing. Since Complete Genomics does not retain customers’ data, the complete and original data set must be shipped back to Complete Genomics via hard disk drive. For more information, see the [Complete Genomics Reanalysis Service Product Note](#), or contact us at info@completegenomics.com.

Where can I find more information on Complete Genomics data and results?

This document is part of a set that describe aspects of the Complete Genomics human genome data:

- *Complete Genomics Data File Formats* — Describes the organization and content of the format used to deliver Complete Genomics human genome data.
- *Variation Data Frequently Asked Questions* — Provides detailed answers to questions about Complete Genomics variation and evidence files.
- *Getting Started Frequently Asked Questions* — Describes how to prepare for receiving Complete Genomics data files.

You can find these documents on the Complete Genomics web site:

www.completegenomics.com/customer-support/

Copy Number Variation (CNV) Analysis

What is the difference between the different CNV files, and how do I select which file to use for my samples?

Complete Genomics provides two versions of each CNV summary file: one version for samples that are presumably diploid and one version for samples expected to exhibit gross copy number abnormalities.

CNV files	When to use these files
cnvDetailsDiploidBeta cnvSegmentsDiploidBeta	Use this set of files when interested in identifying CNVs (compared to the reference genome) in samples that are primarily diploid. Typically, this case includes the majority of non-tumor samples and samples that are homogeneous with regard to limited copy number changes; including, for example, trisomy 21.
cnvDetailsNondiploidBeta cnvSegmentsNondiploidBeta	Use this set of files when interested in identifying CNVs (compared to the reference) in samples that exhibit widespread or long-range divergence from a copy number of two across all autosomal chromosomes. This case includes the majority of tumor samples and perhaps genomes with mosaic aneuploidy.
somaticCnvDetailsDiploidBeta* somaticCnvSegmentsDiploidBeta*	Use this set of files when interested in identifying somatic copy number events (compared to a matched normal sample) in samples that are primarily diploid. While this is unusual for tumor samples, these files may be useful for some tumor-types that are known to exhibit low levels of genomic copy number alterations, such as leukemias.
somaticCnvDetailsNondiploidBeta* somaticCnvSegmentsNondiploidBeta*	Use this set of files when interested in identifying somatic copy number events (compared to a matched normal sample) in samples exhibiting gross copy number abnormalities.

* Cancer Sequencing Service output, only.

How should I use the various coverage values provided in the *depthOfCoverage* and *coverageRefScore* files?

The *coverageRefScore* and *depthOfCoverage* files come with several different calculations for coverage, including: *uniqueSequenceCoverage*, *weightSumSequenceCoverage*, *gcCorrectedCvg*, and *avgNormalizedCoverage*. Each of these values may be useful in understanding certain aspects of the genome or the analysis process.

- *uniqueSequenceCoverage*: This value counts only full mappings of DNBs whose weight ratio is > 0.99, indicating that the estimated probability of the mapping being correct is > 99%. Its primary value over the other measures described here is that in repeat regions of the genome (especially ubiquitous repeats or segmental duplications), only those DNBs that can be assigned to a specific copy with some confidence are counted. For repeat copies that contain unique differences from other copies, this may allow determination that a specific copy of a repeat has

been duplicated or deleted. It may also be useful in evaluating the quality of a variation call in a repeat region, or in understanding why a given region was no-called despite there being many (nonunique) mappings to the region.

- *weightSumSequenceCoverage*: This value counts all mappings, giving fractional attribution to alternative placements of a single DNB. It can be useful, compared with *uniqueSequenceCoverage*, in evaluating regions of the genome that are near-perfect copies of other regions of the genome, as such regions will receive essentially no *uniqueSequenceCoverage*. The ratio of the two measures (as captured by the *fractionUnique* metric in *cnvDetails* files) may provide insight into whether a CNV called at a particular location is really a duplication or loss of that region per se, or whether it might be better understood as reflecting a change in the copy number to a repeat class rather than a specific instance.

As compared to *gcCorrectedCvg*, this value is closer to the raw data, and might help decide whether a given contrast post-GC correction seems justified. It might be useful as input to an alternative approach to bias correction.

- *gcCorrectedCvg*: This value is based on the *weightSumSequenceCoverage*, and reflects a transformation of that value that adjusts for coverage biases corresponding to local GC content; coverage at each position is adjusted based on the GC content of a 1000-base window whose 501st base is the position of interest. This correction improves the overall comparability of coverage values sample-to-sample in the event of any library-to-library change in the extent and pattern of GC content bias to coverage. As such, this value might be preferred as input to a sample-to-sample comparison or normalization process.
- *avgNormalizedCoverage*: This value is based on the *gcCorrectedCvg*, and reflects a transformation of that value that adjusts for forms of repeatable coverage bias other than local GC content. Coverage in each window is adjusted based on the coverage in a collection of 'baseline' samples. This adjustment attempts to account for the apparent copy number relative to what would be expected of a sample that was diploid in that window. This may be useful in estimating copy number in the current sample using a model that expects its inputs to be proportional to absolute copy number (as opposed to relative change in copy number relative to a reference standard).

How do I interpret CNV types “hypervariable” and “invariant”? Should I consider them as candidate CNV regions?

In the *cnvSegmentsDiploidBeta* file, a segment labeled as “invariant” indicates that coverage does not support normal ploidy but the abnormal ploidy is observed to be invariant in both the sequenced genome and the 52 samples that represent the reference genomes for CNV analysis. See Figure 2. Invariant regions may result from some regions of the reference being wrong (such as missed contig overlaps) or rare. In either case, it may not be useful to include the segment in a list of CNVs. Segments labeled “hypervariable” indicate that coverage does not support normal ploidy but coverage is highly variable across the sequenced genome and the 52 reference genomes and is without appreciable clustering. See Figure 1. Hypervariable regions may result from sequencing artifacts or high copy repeat regions (such as STRs or segdups) with high degree of polymorphism. Depending on your relative tolerance for false positives and false negatives, it may be reasonable to include these as candidate CNV segments to be validated using orthogonal technologies.

Figure 1: Invariant Region

An invariant region is called when coverage consistently indicates same CNV across many genomes, including the sequenced genome itself.

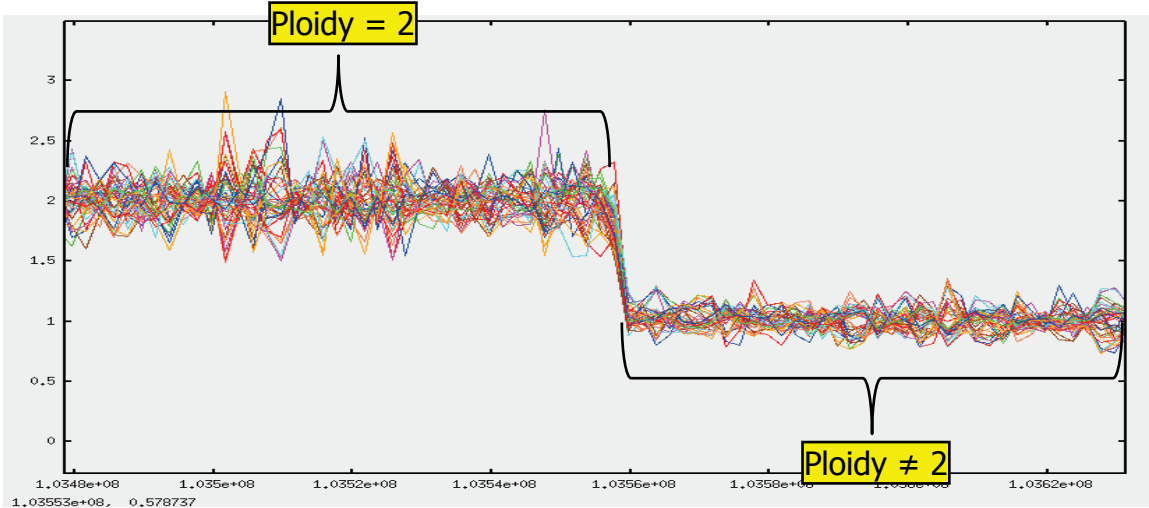
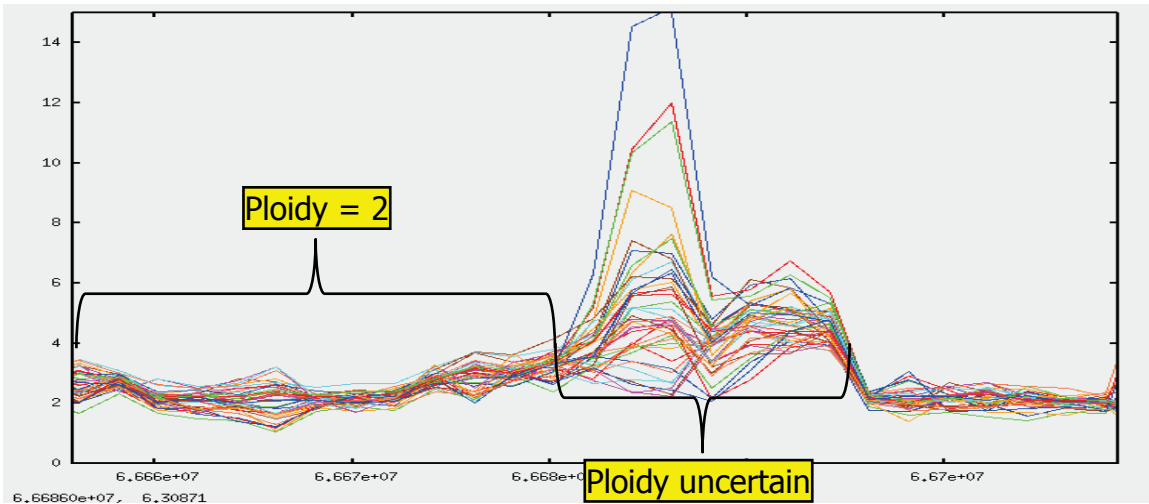


Figure 2: Hypervariable Region

A hypervariable region is called when coverage is highly variable across many genomes, including the sequenced genome, without appreciable clustering.



I am comparing CNV data for two different genomes sequenced by Complete Genomics. How can a region be labeled with `cnvType` “invariant” in one genome and be labeled “+” in the other if assignment of “hypervariable” and “invariant” regions is based on the coverage profiles of the reference genomes set that is used to process both genomes?

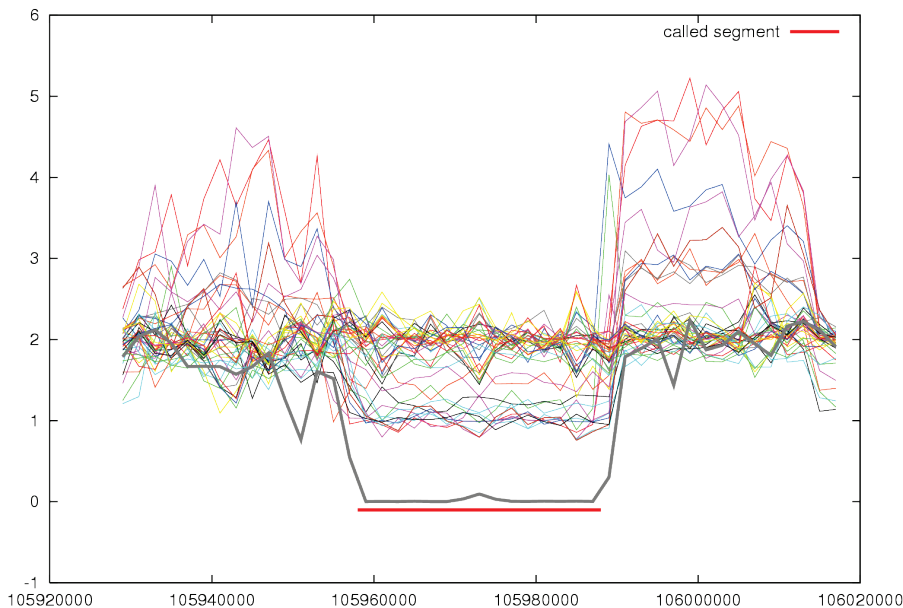
The assignment of “hypervariable” and “invariant” regions depends not only on the reference genome set, but on the specifics of the genome in question. Specifically, there are some heuristic cutoffs for determining whether to label a region “hypervariable” or “invariant”. The cutoffs include

the coverage of the genome of interest. Thus, it is possible for the same genomic region to be assigned different *cnvType* values in different genomes. This occurs when the coverage for the genome in question is far enough outside the range of values seen in the baseline set. Furthermore, regions are labeled “hypervariable” or “invariant” only if they would otherwise be *cnvType* ‘+’ or ‘-’. Thus, a genome can be labeled *cnvType* ‘+’ or ‘-’ in a region that is sometimes called “invariant” or “hypervariable” in other genomes. Additionally, a genome can be labeled *cnvType* ‘=’ in a region that is sometime called “invariant” or “hypervariable” in other genomes.

Figure 3 illustrates an example of a region that could potentially be called “hypervariable” in some genomes and *cnvType* ‘+’ or ‘-’ in other genomes. This figure shows coverage profile plots for individual genomes in the reference genome set, along with the coverage profile of the sequenced genome of interest (represented by the thicker grey line). The red called-segment line marks a region where the lack of sharply separated clusters is considered grounds for assignment of “hypervariable” region. However, because the sequenced genome of interest is clearly well separated from the reference genomes, the region is assigned a *cnvType* ‘-’. If, in a different sequenced genome, the coverage is not well separated from the reference genomes, the region would be assigned “hypervariable”.

Figure 3: Coverage Profile Plots

*A segment with *cnvType* ‘-’ is called for the sequenced genome in a region where the reference set genomes display hypervariable characteristics.*



What is “calledLevel” in my *cnvSegmentsNondiploidBeta* file and how does it relate to ploidy?

In tumor genome processing, we identify discrete coverage levels based on the distribution of the observed normalized coverage levels in the tumor genome. Once the levels are determined, a hidden Markov model (HMM) is used to segment the genome into regions assigned to the identified levels. The called levels are identified by their coverage relative to the median of the diploid portion of the genome. Thus, the results describe segments with values > 1 being amplified relative to the genome median and values < 1 being reduced relative to the genome median. Due to either or both tumor heterogeneity and/or normal contamination (presence of DNA from normal cells in the tumor sample), coverage levels may not correspond to integer-valued ploidy.

How do I identify segments that are amplified or reduced in my tumor genome?

Because tumors are usually nondiploid samples (due to heterogeneity and widespread copy number aberrations), the CNV files for 'Nondiploid' samples are generally the most appropriate for studying copy number changes in these samples. In these files, segment ploidy and CNV Type are not reported, but it is possible to filter on the *calledLevel* column of the *cnvSegmentsNondiploidBeta* file. Look at the header of this file and identify #MEAN_LEVEL with value closest to 1. This indicates a coverage level that is most representative of the median genome coverage. Filtering on the *calledLevel* column for value not equal to #MEAN_LEVEL with value closest to 1 will help you obtain segments of interest. However, this approach does not necessarily correspond to identifying non-diploid segments; for instance, in a tumor that is largely tetraploid, it will result in labeling regions that have three copies or fewer as "reduced" and regions that have five copies or more as "amplified." Due to the potential for misinterpretation, we do not explicitly identify levels as amplified or reduced.

I have matched tumor and normal samples. How do I get "somatic" CNV calling?

For samples submitted for the Cancer Sequencing Service, Assembly Pipeline version 2.0 provides paired analysis between tumors and baseline samples submitted as a group. The output is provided in the *somaticCnvDetails-[ASM-ID]-N1* and *somaticCnvSegments-[ASM-ID]-N1* files and reflects the direct comparison of coverage in the tumor sample with coverage in the normal match. For samples submitted for the Standard Sequencing Service, no somatic analysis is performed. There are currently no tools provided for making this comparison. Files that could be useful for comparing copy number between to samples include the *cnvDetails* file, which provides normalized coverage, as well as GC-corrected coverage, and *cnvSegmentsNondiploid*, which identifies called CNV segments and boundaries.

What are the limitations of using a single matched genome as the baseline for the paired analysis used to identify somatic CNVs?

Although the paired analysis approach is most appropriate for identifying somatic CNVs in tumor-normal pairs, there are certain limitations of the paired analysis, where a single sample is used as the baseline genome:

- The CNV calling is based on an HMM model containing a fixed set of states; in the case of the 'diploid' analysis, these correspond to integer-valued copy number, while in the case of the 'nondiploid' analysis, they correspond to strongly exhibited relative coverage levels. In portions of the genome where the paired baseline sample is itself not diploid, a change in copy number in the derived sample, for example a gain of one copy, will result in different changes in relative coverage than in portions of the genome where the baseline sample is diploid.

For instance, if the baseline sample is triploid, an increase in the derived genome to tetraploid will result in a relative coverage of 1.33 whereas in a region where the baseline is diploid, an increase to triploid will result in a relative coverage of 1.5. This may lead to difficulty calling the correct copy number change in the derived sample in the regions where the baseline is nondiploid. In the limit, if a highly aneuploid sample is used as the baseline, it may substantially negatively impact the paired CNV calling, which will adversely affect the analysis of matched samples where the sample designated as the baseline is an aneuploid tumor.
- The paired baseline sample, being a single sample, may by chance be more different from the target sample in some regions than is the standard baseline. That is, the measurement variance on the baseline sample will be higher than that for the multi-genome baseline. This may lead to a modest number of somatic CNV calls even where two paired samples are truly copy number identical.

How do I evaluate the confidence of a called CNV? What sorts of underlying evidence can I look at?

The Phred-like scores reflecting the confidence that the segment has the called ploidy and that the segment has the correct CNV type are indicated in the *ploidyScore* and *CNVTypeScore* fields. The higher the *ploidyScore* and *CNVTypeScore*, the more confidence you have in the CNV. In addition to these statistical measures of confidence, segments represented by three or more window lengths (6 kb) are more likely to be true CNVs, as segments smaller than 6 kb tend to be associated with higher false positive rates.

What does it mean when “CNVTypeScore” is much greater than “ploidyScore”?

This means that we are much more confident that the segment is either amplified (if CNV type = “+”) or deleted (if CNV type = “-”) than we are in the actual magnitude of copy number change.

How do CNV and SV results relate to one another? Would structural variation events that involve changes in copy number be reported in both CNV and SV files?

Results from CNV and SV analysis are generated independently, using different methods. We employ read depth and discordant mate pair methods to detect CNVs and SVs, respectively. Each approach has strengths that allow detection of a class of event that the other does not. For example, read-depth analysis is able to identify CNVs in complex regions of the genome rich in segmental duplications, while discordant mate pair mapping is able to detect events, such as inversions, that do not result in copy number changes. While some events, such as deletions, may be detected by both methods and reported in both CNV and SV files, we do not provide information that would directly link the calls that represent the same event.

Why did Complete Genomics miss a known event or an event that is obvious from the raw data in my sample?

A known limitation in our current pipeline that we are working to improve is a reduced sensitivity to CNVs less than 5 KB in length and low sensitivity to CNV less than 3 KB in length. Also, regions where there is high copy number polymorphism (such as high copy segmental duplication, short tandem repeats, and heterochromatic regions), are generally no-called. For tumor genome segmentation, sufficient heterogeneity in a tumor may make it difficult to correctly identify all the relevant coverage levels. In addition, excessive normal contamination may make differences in ploidy within the tumor portion of a sample lead to differences in coverage that are too small to be modeled reconstructed, even if the tumor is itself homogeneous.

Can I get access to the reference data set used to generate the baseline?

Yes. For samples analyzed using Assembly Pipeline version 2.0 or later, the baseline genome set is comprised of 52 unrelated genomes from the Complete Genomics Diversity Panel (69 genomes based on Assembly Pipeline version 1.10). A file that summarizes the underlying data and normalization constants for each of the CNV baseline genomes is available from the Complete Genomics FTP site:

ftp://ftp2.completegenomics.com/Baseline_Genome_Set/CNVBaseline

The accompanying *CNV Baseline Genome Dataset: Data Format Description* document provides the identifiers for each genome in the CNV baseline set and describes the data file format for the CNV baseline genome composite file. For more details on how the CNV baseline genomes are used to normalize and ‘no-call’ CNV data for the sequenced genome, please consult the [CNV Methods](#) document. Note that the same genomes are used to construct our CNV, SV, and MEI baseline sets.

What is the Lesser Allele Fraction (LAF), and where can I find it?

The allele fraction for alleles at a heterozygous site in a diploid portion of the genome should be 50% each, but due to heterogeneity copy number variation, the allele fraction of each allele may be greater or less than 50%. The Lesser Allele Fraction (LAF) is the fraction of the sample containing the allele that is present in $\leq 50\%$ of the sample. Therefore, the range of LAF values is 0 to 0.5. The LAF is similar to “B-allele frequency” estimates from microarray genotyping data, but captures the fraction of the less abundant allele/haplotype rather than the fraction of an arbitrary allele at each locus.

The LAF is calculated for samples submitted to the Cancer Sequencing Service and is described in several files in the Complete Genomics output. These include the *masterVarBeta* file for each tumor sample, all CNV segments and details files, the VCF file, and each version of the Circos plot.

There are two calculations for LAF: a multi-sample LAF and a single-sample LAF. These calculations are made for both tumor and normal samples submitted to the Cancer Sequencing Service. Multi-genome LAF calculations apply to 100 kb windows and are based on allele read counts in the tumor at loci that are called heterozygous in the matched baseline sample. Single-sample LAF calculations apply to 100 kb windows and are based on the read counts of each allele wherever a heterozygous call is made. The single-sample LAF is only shown in the Circos plots and is not included in the CNV files.

For more information, see [“How do I interpret the Lesser Allele Fraction \(LAF\)?”](#)

How do I interpret the Lesser Allele Fraction (LAF)?

The Lesser Allele Fraction (LAF), described in [“What is the Lesser Allele Fraction \(LAF\), and where can I find it?”](#), ranges from 0 to 0.5. Samples that are pure and diploid and not affected by loss of heterozygosity should exhibit LAFs at 0.5, indicating that each allele is present in equal amounts.

Single-sample LAF (in Circos plots, only):

- Where LAF is = 0.5, the simplest explanation is that the sample is diploid, with two distinct haplotypes in that region, though it can also reflect any even copy number with equal numbers of two haplotypes.
- Where LAF is < 0.5 indicates copy number variation with unequal numbers of the major and minor haplotypes.
- Regions that have experienced loss of heterozygosity (LOH) will have low but typically non-zero estimated LAF and the ‘noisiness’ of the estimate will be apparent in extended regions. This is because the sites used for LAF estimation will be false heterozygous variations, hence they will be sparse and typically with low reads counts.

Multi-sample LAF:

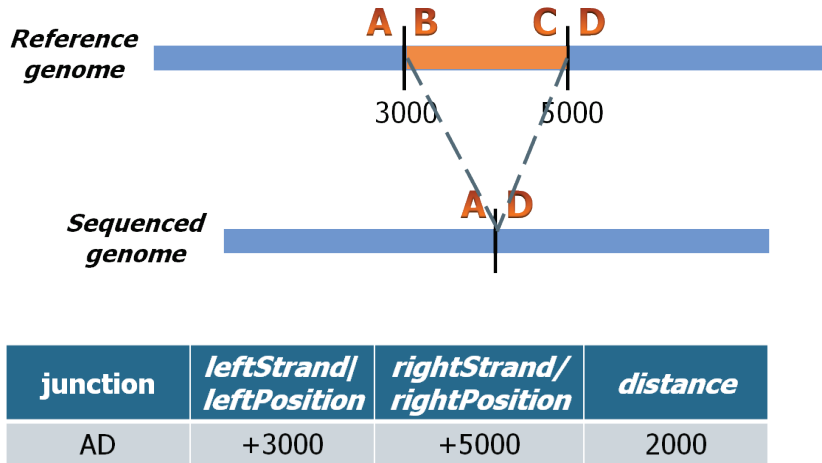
- Where LAF is = 0.5, the simplest explanation is that the sample is diploid, with two distinct haplotypes, in that region, though it can also reflect any even copy number with equal numbers of two haplotypes.
- Where LAF is < 0.5 indicates copy number variation with unequal numbers of the major and minor haplotypes. The LAF in conjunction with depth-of-coverage information may permit improved estimation of major and minor allele copy number, as well as estimation of normal contamination or other heterogeneity.
- Where LAF is = 0 across a region of the genome, this is indicative of Loss of Heterozygosity (LOH).

Structural Variation (SV) Analysis

What is a junction?

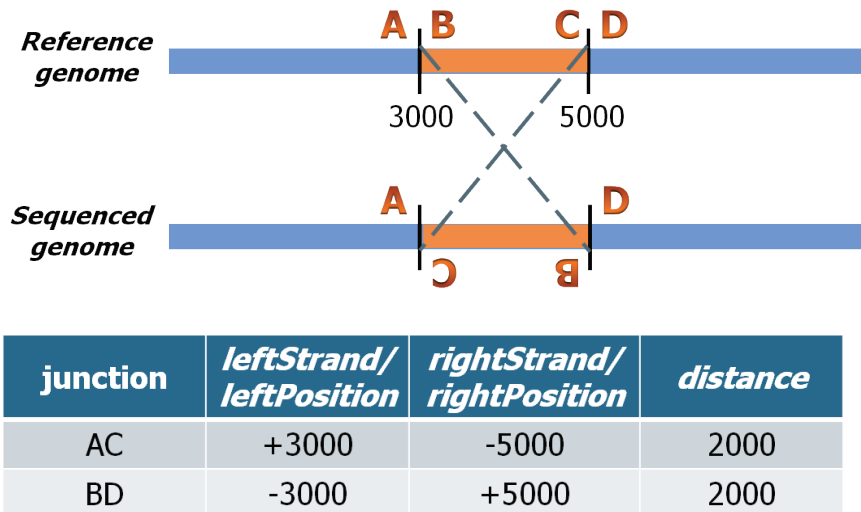
A junction is defined as two separate regions of the reference genome that appear to be near each other in the genome being sequenced. Deletions are represented by a single junction (Figure 4), while other events such as inversions and intrachromosomal translocations can be represented by more than one junction (Figure 5).

Figure 4: Deletion of Segment BC in the Sequenced Genome Represented by Junction AD



As shown in Figure 4, deletion of segment BC in the sequenced genome would be represented by junction AD: a junction that connects sections A and D. *leftStrand*, *leftPosition*, *rightStrand*, *rightPosition*, and *distance* are fields reported in **junction** files. *leftStrand* and *rightStrand* values indicate that left and right side of the junction have the same strand orientation, while the *distance* value of 2,000 indicates that the position on left and right side of the junction closest to the breakpoint is 2,000 bp apart on the reference genome.

Figure 5: Inversion of Segment BC in the Sequenced Genome Represented by Two Different Junctions



As shown in Figure 5, inversion of segment BC in the sequenced genome would be represented by two different junctions: junction AC that connects sections A and C and junction BD that connects sections B and D. Note that unlike what is shown in Figure 5, coordinates for paired junctions are not typically identical for real events. *leftStrand* and *rightStrand* values indicate that, for both junctions, left and right side of the junction have the opposite strand orientation, while the *distance* value of 2,000 indicates that the positions on left and right side of the junction closest to the breakpoint are 2,000 bp apart on the reference genome.

Does Complete Genomics indicate the structural variant type represented by a junction in SV files?

Complete Genomics provides two files—*allSvEventsBeta* and *highConfidenceSvEventsBeta*—that report structural variation events involving identified junctions found in the *allJunctionsBeta* and *highConfidenceJunctionBeta* files, respectively. The *cgatools* junctions2events is used to identify structural variation events such as deletions, inversions, and translocations from lists of junctions. It determines which event type a junction is consistent with by identifying possible relationships among the provided junctions. Single-sample junctions are rationalized into event types using this tool, but somatic junctions are not rationalized into event types at this time.

Are somatic events identified for my tumor-normal pair?

Currently, Complete Genomics does not report somatic events. In other words, we do not attempt to rationalize somatic junctions into somatic events. The *EventId*, *Type*, and *RelatedJunctions* annotations in the *somaticAllJunctionsBeta* file and the *somaticHighConfidenceJunctionBeta* file refer to a description of the event identified in the tumor sample.

Does Complete Genomics indicate zygosity of the junction?

Currently, we do not attempt to call zygosity of the junction. However, zygosity can be inferred, to a certain extent, by interrogating the coverage in junction region. For example, if coverage in a putative deletion junction region is near zero, you can infer that it is likely a homozygous deletion event.

Are the indels reported in the *var-[ASM-ID].tsv* file also reported in the junctions files?

Small insertion and deletion events are detected during the assembly process. They are only reported in the *var* file, and not repeated in the *junctions* files, as they are not detected by discordant mate pair mapping method employed for detection of larger structural variations.

The *allJunctionsBeta*, *highConfidenceJunctionBeta*, *somaticAllJunctionsBeta*, *somaticHighConfidenceJunctionBeta*, and *evidenceJunctionClustersBeta* files have the same file format. What are the differences among these files?

These five files represent outputs at various steps of our SV detection pipeline. Junctions are detected by identifying clusters of DNBs in which each arm maps uniquely to the reference genome, but with an unexpected mate pair distance or anomalous orientation. If a cluster contains three or more DNBs, a junction is output. These junctions, annotations such as the putative junction breakpoints and size of structural variant, and transition length estimated from this initial clustering of DNBs are reported in the *evidenceJunctionClustersBeta* file. Once junctions are detected, local *de novo* assembly is attempted on each junction. These junctions, annotations such as breakpoint, size of SV, transition sequence, and length that have been refined by local *de novo* assembly are reported in the *allJunctionsBeta* file. So, while *evidenceJunctionClustersBeta* and *allJunctionsBeta* files report the same junctions, for junctions in which local *de novo* assembly was successful, junctions annotations differ.

A set of filtering criteria is applied to junctions in the ***allJunctionsBeta*** file to obtain a list of high-confidence junctions, which are then reported in the ***highConfidenceJunctionBeta*** file. So, ***highConfidenceJunctionBeta*** file contains a subset of the ***allJunctionsBeta*** files, but the annotations for junctions found in both files are the same.

For samples submitted for the Cancer Sequencing Service, two additional files are provided. The ***somaticAllJunctionsBeta*** file represents junctions that are identified in the ***allJunctionsBeta*** file for the tumor but not in the ***allJunctionsBeta*** file for the normal sample. The ***somaticHighConfidenceJunctionBeta*** file includes junctions identified in the ***highConfidenceJunctionBeta*** file for the tumor, but not in the ***allJunctionsBeta*** file for the normal match.

How do I evaluate the confidence of a called junction? What sorts of underlying evidence can I look at?

There are several columns of information in the ***allJunctionsBeta*** and ***highConfidenceJunctionBeta*** files that can be used to gauge the confidence level of the called junction. These same metrics are used to filter for high-confidence junctions reported in the ***highConfidenceJunctionBeta*** file.

1. *DiscordantMatePairAlignments* indicates the number of discordant mate pairs in the cluster that supports the called junction. A higher number of discordant mate pairs supporting a junction indicate higher confidence that this junction is present in the sequenced genome.
2. *junctionSequenceResolved* indicates whether assembly of sequence across the two sides of the junction was successful. A value of "Y" indicates success and lends strong support to the inference that there is a physical connection between the left and right side of the junction (that is, a higher confidence that this junction corresponds to a real event).
3. *KnownUnderrepresentedRepeat* indicates whether left or right sections of the junction overlaps with repetitive genomic elements that are known to be underrepresented in the human reference genomes. A value of "Y" indicates overlap and indicates less confidence that the junction is real.
4. *LeftLength* and *RightLength* indicate the lengths of the left and right sections of the junction. Longer lengths indicate higher confidence that junction is real.
5. If a junction implies an interchromosomal translocation event (left and right sections of junction map to different chromosomes), the value in *FrequencyInBaseline* field can be used to gauge confidence in the junction, along with metrics described above. *FrequencyInBaseline* indicates the frequency at which the junction is detected in the 52 normal genomes used as the baseline reference set. Given that an interchromosomal event in a normal genome is rare, higher frequency indicates less confidence that the junction is real and that the junction is more likely to have resulted from a processing artifact or sequence similarity of one mate pair to another region of the genome than a true physical connection between left and right sections of the junction.

Are there limitations on the classes of junctions Complete Genomics can discover?

Our pipeline has known limitations that we are working to improve. These limitations are:

1. Sensitivity for deletion events < 1 KB in size.
2. Due to the fact that we only use unique mappings for junction detection, we do not detect junctions involving high-identity repeats such as segmental and tandem duplications.
3. Insertions of transposable elements such as common LINEs and AluY or insertions of sequences not found in the reference genomes are not detected.

Can I get access to the reference data used to create the SV baseline?

Yes. For samples analyzed using Assembly Pipeline version 2.0 or later, the SV baseline genome set is comprised of 52 genomes from the Complete Genomics Diversity Panel (69 genomes based on Assembly Pipeline version 1.10). You can download a file that summarizes the detected junctions and their frequencies across the SV baseline set from the Complete Genomics FTP site:

ftp://ftp2.completegenomics.com/Baseline_Genome_Set/SVBaseline

The accompanying *SV Baseline Genome Dataset: Data Format Description* document provides the identifiers for each genome in the SV baseline set and describes the data file format for the SV baseline genome composite file. Note that the same genomes are used to construct our CNV, SV, and MEI baseline sets.

Mobile Element Insertion Detection

Are MEIs detected using the same method as SV detection?

No. The discordant mate pair mapping method used to detect SVs only considers mate pairs in which both ends map uniquely to the reference genome. To detect MEI events, we identify mate pairs in which one end maps outside the event and one end maps within the events. Thus, we look for clusters of mate pairs in which one end maps uniquely to the reference genome while the other end maps to a sequence that is ubiquitous in the reference, that is, to a sequence for which reads are marked as “overflow”.

What is the resolution of insertion site detection? Does Complete Genomics assemble the insertion site?

The range of likely insertion sites is reported for each event in the *InsertionRangeBegin* and *InsertionRangeEnd* fields within the ***mobileElementInsertionsBeta*** file. This range is determined from the initial mapping of mate pairs that map in the vicinity of the insertion site with one arm and map to ubiquitous sequence with the other arm. Currently, we do not attempt to perform local *de novo* assembly of the mate pairs that map across the insertion point to refine the position to a single base pair resolution.

What MEI type does Complete Genomics detect?

We align each unmapped end of mate pair clusters that support MEI events to sequences of a defined set of mobile element sequences. This set contains the most active mobile element sequences such as L1s, Alus, and SVAs, which are more likely to contribute to the structural variations in the human genome. Please refer to the [Data File Format](#) document for information on this list. As more information becomes available regarding the contribution of other mobile element sequences to variations in the human genome, Complete Genomics will add these relevant sequences to its database.

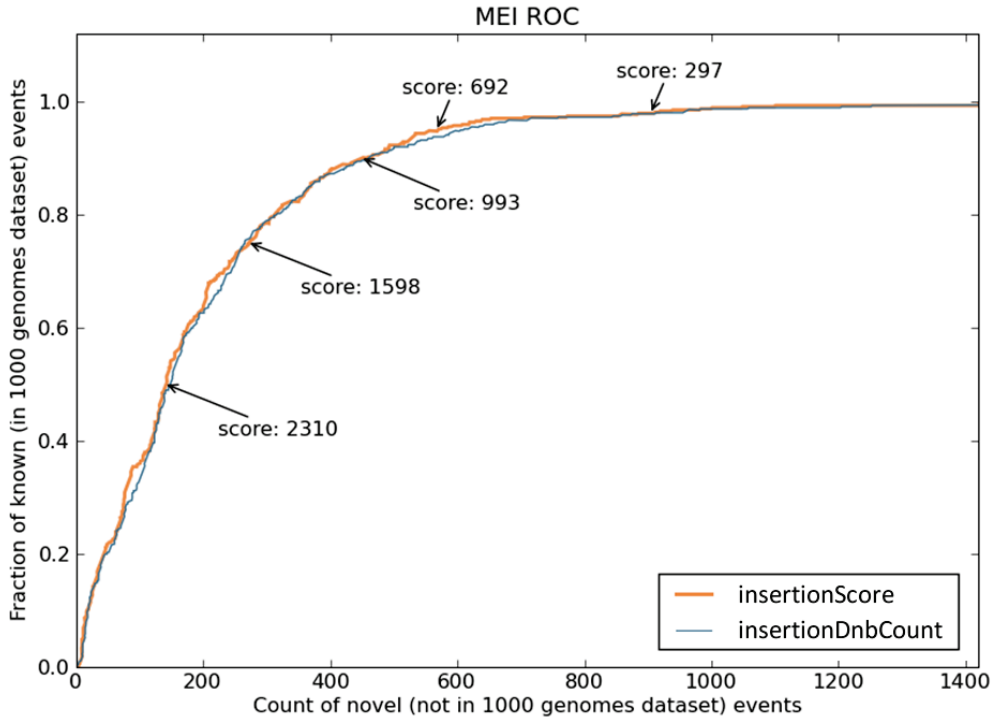
How should I filter for high-confidence MEIs?

A ROC curve is generated for each sequenced genome to demonstrate the tradeoff in sensitivity and specificity as a function of two metrics that are well correlated with quality of the event detection: *insertionScore* and *insertionDnbCount*. This curve is provided in the ***mobileElementInsertionsROCBeta*** file; Figure 6 shows an example. The Y-axis is computed with regards to the fraction of 1000 Genomes Projects MEI events that is also called in the sequenced genome to estimate sensitivity at a given *insertionScore* or *insertionDnbCount*. The X-axis uses the count of MEI events that are called in the sequenced genomes that were not called by the 1000 Genomes Project to estimate specificity at a given *insertionScore* or *insertionDnbCount*. These will not

all be false positives, as suggested by the nearly linear relationship below score 1598, which could indicate that the set of true high scoring events are a mix of events found by the 1000 Genomes Project and events missed by the 1000 Genomes Project. The curve enables you to determine the *insertionScore* or *insertionDnbCount* that would achieve a desirable balance between sensitivity and specificity. Once this threshold is determined, events can be easily filtered by this threshold value using the *KnownEventSensitivityForInsertionScore* or *NovelEventCountForInsertionScore* columns in the ***mobileElementInsertionsBeta*** file.

Figure 6: ROC Curve

This image demonstrates the trade-off in sensitivity and specificity as a function of *insertionScore* and *insertionDnbCount*.



The ROC curve graph in the ***mobileElementInsertionsROCBeta-[ASM-ID].png*** file is provided to facilitate selection of a threshold that would best meet your requirements on sensitivity and specificity of the MEI detection.

For each candidate mobile element insertion site, the data also includes the count of the number of DNBs that map across the insertion site—DNBs where one arm map upstream and one arm map downstream of the reference range where the insertion is likely to be located—with mate gap distance that would be unlikely had the DNBs come from the allele where the insertion was present. The count is reported in the *referenceDnbCount* field of the ***mobileElementInsertionsBeta.tsv*** file and allows determination of the zygosity of the MEI events. A distribution graph of these counts for the sequenced genome is provided in ***mobileElementInsertionsRefCountsBeta-[ASM-ID].png*** to help with the selection of the appropriate threshold to separate heterozygous and homozygous events.

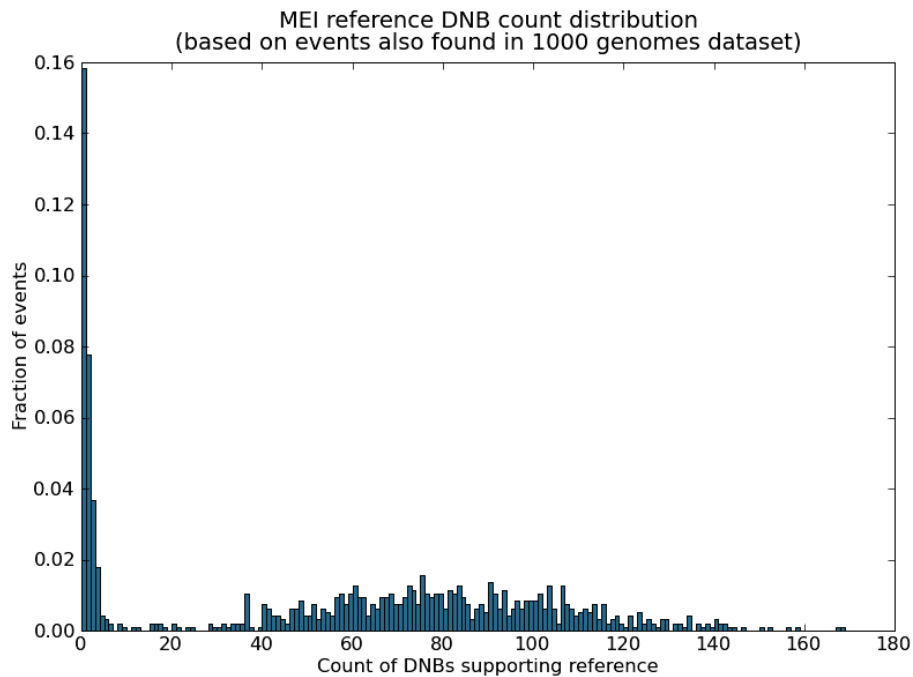
Does Complete Genomics identify somatic MEIs?

At this time, Complete Genomics does not provide somatic analysis for MEI detection.

Is the zygosity of events reported?

The zygosity of an insertion event is not reported in the *mobileElementInsertionsBeta.tsv* file. However, the plot provided in the *mobileElementInsertionsRefCountsBeta-[ASM-ID].png* file can be used to filter for homozygous and heterozygous events. This plot shows the distribution of mate pairs that support reference allele for MEI events detected by the 1000 Genomes Project that were also detected in the sequenced genome. As shown in Figure 7, this distribution is usually bi-modal, corresponding to the homozygous insertions (peaking at zero DNBs) and heterozygous insertions (centered at approximately 80 DNBs for this genome). Thus, for this genome, any threshold between 10 to 30 DNBs would be reasonable to apply when filtering on the *ReferenceDnbCount* column to separate homozygous and heterozygous MEI events.

Figure 7: Bi-Modal Distribution of Mate Pairs



Can I get access to the reference data used to create the MEI baseline?

A composite file profiling MEIs and their frequencies across the baseline set that is used to annotate the *mobileElementInsertionsBeta* file is currently not provided. Because the same genomes were used to construct the CNV, SV, and MEI baseline sets, see either the *CNV Baseline Genome Dataset: Data Format Description* or *SV Baseline Genome Dataset: Data Format Description* documents for the list of genomes. These documents are packaged with the Baseline Genome Set summaries available from the Complete Genomics FTP site:

- ftp://ftp2.completegenomics.com/Baseline_Genome_Set/CNVBaseline
- ftp://ftp2.completegenomics.com/Baseline_Genome_Set/SVBaseline