Complete genomics

# Small Variants
# Frequently Asked Questions (FAQ)

Updated September 2011

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.

# Summary Information for each Genome

## How does Complete Genomics map reads and call variations?

Reads are initially mapped to the reference genome using a fast algorithm, and these initial mappings are both expanded and refined by a form of local *de novo* assembly, which is applied to all regions of the genome that appear to contain variation (SNPs, indels, and block substitutions) based on these initial mappings. The *de novo* assembly fully leverages mate-pair information, allowing reads to be recruited into variant calling with higher sensitivity than genome-wide mapping methods alone typically provide. Assemblies are diploid, and we produce two separate result sequences for each locus in diploid regions (exceptions: mitochondria are assembled as haploid and for males the non-pseudoautosomal regions in the sex chromosomes are assembled as haploid). Variants are called by independently comparing each of the diploid assemblies to the reference.

The process is described in more detail in our *Science* paper (Science 327 (5961), 78. [DOI: 10.1126/science.1181498]). We also recommend that you read the Complete Genomics Service FAQ as background for this document. You can access the paper at www.rdrmanac.com at no charge.

Copy number variable (CNV) regions are called based on depth-of-coverage analysis. Sequence coverage is averaged and corrected for GC bias over a fixed window and normalized relative to a set of standard genomes. A hidden Markov model (HMM) is used to classify segments of the genome as having 0, 1, 2, 3 copies…up to a maximum value. Structural variations (SVs) are detected by analyzing DNB mappings found during the standard assembly process described above and identifying clusters of DNBs in which each arm maps uniquely to the reference genome, but with an unexpected mate pair length or anomalous orientation. Local *de novo* assembly is applied to refine junction breakpoints and resolve the transition sequence. The process for CNV and SV detection is described in more detail in *Complete Genomics Data File Formats*.

## How do I assess the quality of a genome produced by Complete Genomics?

In the summary file (***summary-[ASM-ID].tsv***), you will see a variety of metrics that may be helpful in understanding the quality of the delivered genome. For example:

- Gross Mapping Yield (Gb) — total base-pairs of sequence reads mapped to the reference genome (based on the initial mapping process only)
- Fully called genome fraction — percentage of reference genome with full (diploid) calls in the sequenced sample (following assembly)
- Fully called exome fraction — percentage of reference exome with full (diploid) calls in the sequenced sample (following assembly)
- Genome fraction where *weightSumSequenceCoverage* ≥n — Fraction of the reference genome bases where coverage is greater than or equal to n, with n being 5x, 10x, 20x, 30x and 40x.
- Exome fraction where *weightSumSequenceCoverage* ≥n — Fraction of the reference exome bases where coverage is greater than or equal to n, with n being 5x, 10x, 20x, 30x and 40x.

There are additional biological metrics which one would expect to be roughly consistent across genomes from individuals of the same ethnicity (even to genomes sequenced using other methods). These are also quite useful for quality control. They include:

- SNP total count (for genome and exome)
- SNP heterozygous/homozygous ratio (for genome and exome)
- SNP transitions/transversions ratio (for genome and exome)
- SNP novelty fraction (for genome and exome)

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 the REPORTS directory of our data delivery, you will find 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***): Shows a visual summary of small and large variation data for each genome. The image includes density of homozygous SNPs, density of heterozygous SNPs, gene symbols for impacted genes, and, when applicable, density of somatic variants.
- ***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.

## What is the difference between “Gross mapping yield” and “Both arms mapped yield” in the summary file?

“Gross mapping yield” counts aligned bases within DNBs where at least one arm is mapped to the reference genome, excluding reads marked as overflow (large number of mappings to the reference genome indicative of highly repetitive sequence). “Both arms mapped yield” counts aligned bases within DNBs where both arms mapped to the reference genome on the correct strand and orientation and within the expected distance.

## What are the definitions for Fully Called, Partially Called, Half-Called and No-Called?

“Fully called” indicates that the assemblies of both diploid alleles meet the minimum required confidence thresholds, and thus both alleles are considered called. In this case, both alleles may be variant, or one may be reference and the other variant. If both are variant, they may be the same (homozygous) or different (heterozygous).

At a “partially called” or “half-called” site, only one allele meets the threshold to call the site confidently while the other does not. The Complete Genomics software reports this partial information for that locus (rather than no-calling the site entirely). Effectively, this is a statement that “we know this allele is present” but we can say little about what other allele is also present in a diploid region.

A “no-called” allele is one where we cannot determine the sequence of the sample at our minimum thresholds. See “What exactly is a reference call? How is this different from a no-call?”

## In the *summary-[ASM-ID].tsv* file, how is the number of homozygous SNPs calculated?

The number of homozygous SNPs is calculated from the ***var-[ASM-ID].tsv*** file, and is equal to the sum of all diploid loci where the same SNP is present on both alleles.

## In the *summary-[ASM-ID].tsv* file, how is the number of heterozygous SNPs calculated?

The number of heterozygous SNPs is calculated from the ***var-[ASM-ID].tsv*** file, and is equal to the sum of SNPs present in the following types of loci:

- het-ref SNP: A single-base diploid locus where a SNP is present on one allele, and the other allele is reference.
- alt-alt SNP: A single-base diploid locus where each allele contains a different SNP.

## In the *summary-[ASM-ID].tsv* file, how is the total number of SNPs calculated?

The total number of SNPs is calculated from the ***var-[ASM-ID].tsv*** file, and includes SNPs present in all of the following types of loci:

- het-ref SNP: A single-base diploid locus where a SNP is present on one allele, and the other allele is reference.
- hom SNP: A single-base diploid locus where the same SNP is present on both alleles.
- alt-alt SNP: A single-base diploid locus where each allele contains a different SNP.
- hap SNP: A single-base haploid locus where a SNP is present on the single allele.

- other: All SNPs that occur in loci that do not fall into the above categories, including SNPs present in loci containing no-calls.

## In the *summary-[ASM-ID].tsv* file, what regions of the genome are included in the “exome”?

The exome is defined as the coding regions (CDS) of protein-coding genes, plus all of the untranslated genes, minus any transcripts (coding or otherwise) that are rejected by the annotation pipeline. A small percentage of transcripts in Build 36 and Build 37 are excluded from the annotation results due to the one or more of the following reasons:

- The transcript contains an unknown (“X”) amino acid.
- The start and/or stop codon positions are unknown.
- The transcript contains unspecified nucleotides.
- The transcript maps to an unknown location/chromosome.

To obtain the list of excluded transcripts, please contact support@completegenomics.com.

## In the *summary-[ASM-ID].tsv* file, how is the number of SNPs in the exome calculated*?*

The number of exonic SNPs is calculated from the ***var-[ASM-ID].tsv*** file and includes SNPs present in all of the following categories:

- SYNONYMOUS: A SNP in the coding region of a transcript that has no affect on the protein sequence.
- NON-SYNONYMOUS: A SNP in the coding region of a transcript that alters protein sequence.
- DISRUPT: A SNP in the splice donor or splice acceptor sites of a transcript, predicted to disrupt splicing.
- NO-CHANGE: A SNP located outside the coding and splice donor/acceptor regions of a transcript, i.e., in the 3’ or 5’ untranslated regions (UTR3 or UTR5).
- UNKNOWN-TR: SNPs located in transcripts that are rejected by the annotation pipeline.

Each variation is only counted once per genome. If a variation has different functional consequences in different transcripts, only the most severe functional effect is counted.

## In the *summary-[ASM-ID].tsv* file, how are variations in potentially redundant regions of the genome counted?

The following rules are used for counting variations:

- Variants in the pseudo-autosomal regions are only counted once.
- If a variant is contained within more than one transcript, we count only the most functionally deleterious mutation (we have a defined hierarchy).

# Variant Calls: SNPs and Small Indels

## When Complete Genomics calls a variation (or a variant allele) what exactly does that mean?

A "variation" or "variant" refers to an allele sequence that is different from the reference at as little as a single base or for a longer (potentially much longer) interval. In general the distinction between "variation" and "polymorphism" is that polymorphisms are by definition variable sites within or between populations. "Variation" makes no assumption about degree of polymorphism except by comparison between a sample and the reference (recall that the reference sequence can be wrong at some sites). Thus, scientists will sometimes use the term Single Nucleotide Variant (SNV) over SNP (single Nucleotide Polymorphism). However, we continue to use the acronym SNP as it is more ubiquitous, if not entirely precise in this case.

## What types of variants are indicated in the variation files?

SNPs, small insertions and deletions, and small block substitutions are indicated as variants in the Complete Genomics variation files, found in the ASM directory. By "small", reported indels may be up to about 50 bases in length, although the precise upper limit varies by region and coverage. In addition to these variants, we also call Copy Number Variants (CNVs), Structural Variants (SVs), and Mobile Element Insertions (MEIs), which are reported in separate folders within the ASM directory. While these variants are all determined in comparison to the human genome reference, genomes submitted for the Cancer Sequencing Service are additionally analyzed for somatic variants called in comparison to the baseline genome within the submitted pair or trio.

## What exactly is a reference call? How is this different from a no-call?

Complete Genomics makes a strong distinction between a no-call and a confident homozygous reference call. Some other pipelines identify variants in sequence but do not make this distinction. Where they fail to call variants, one must rely on rough surrogate measures (such as depth of coverage and mapping scores) to help interpret whether non-variant sites are homozygous reference or are simply not callable. This distinction can be one source of confusion when comparing data across technologies.

Errors in regions falsely called homozygous reference (false-negative variant calls) are included in Complete Genomics overall error rate estimates.

## What is a "sub" or a "delins"?

A "sub" is a block substitution, where a series of nearby reference bases have been replaced with a different series of bases in an allele. The sample's allele and reference may be the same length ("length-conserving") or not ("length-changing"). In data generated by Complete Genomics pipeline versions prior to 1.7 a "sub" was denoted as a "delins".

## What defines a "locus"? Are loci variant or are alleles variant? Explain the asymmetric calls Complete Genomics produces at some loci?

Complete Genomics calls variants on each allele by comparing the assembly of that allele to the reference sequence. This process is repeated independently for each of the two diploid alleles at each autosomal locus. Bases in the genome with variants on either or both alleles in close proximity are grouped together as a single variant locus. For example, the middle three positions at the site below are considered one variant locus:

```
Reference:       TAG TCG CCT
```

```
Allele 1:        TAG TTG CCT     one ref + one SNP + one ref
Allele 2:        TAG CAC CCT     a 3 base block substitution
```

## How does Complete Genomics determine when to call a site with multiple variant bases as one locus or as multiple loci? For example, two neighboring variant bases could be coded as two SNPS or as one two-base block substitution.

Generally, if two or more reference bases on both alleles are called between two variant sequences, then the site is broken into smaller events.

## Each variant allele has been identified as allele "1" or "2". Does that mean that all of the allele 1 variants are located on the same parental chromosome?

No. The allele number is assigned arbitrarily at each locus and does not indicate phase. Where phase is determined, generally because variants are within the same vicinity, the *haplink* field in the variations file will be populated to indicate this. Variant alleles with the same haplink ID are known to be in *cis*-phase, that is, on the same parental chromosome.

Note that prior to pipeline version 1.8, the "allele" column in the variation file was called "haplotype".

## What do "N" and "?" in calls mean? Are they always in alleles marked as no-calls?

An "N" indicates that a specific base could not be resolved on the allele in question; however the flanking (non-N) sequence may have been called. A "?" indicates that the unresolved region may include zero or more unknown bases. For example, "ATGC?" means that the exact number and composition of bases (if any) immediately after ATGC on that allele could not be determined.

Loci with either or both "N's" or "?" will also always be marked as no-call, no-call-rc (no-call, reference-consistent) or no-call-ri (no-call, reference-inconsistent) as appropriate (see "Please explain "no-call-ri", "no-call-rc", "ref-consistent" and "ref-inconsistent". How should I use these?").

## Please explain "no-call-ri", "no-call-rc", "ref-consistent" and "ref-inconsistent" in the *var* file. How should I use these?

All no-call variant types indicate that the sequence could not be fully resolved, either because of limited or no information, or because of contradictory information. When some portions of the allele sequence can be called but others not, we indicate this as "no-call-rc" (no-call, reference-consistent) if those called portions are the same as the reference. We use no-call-ri (no-call, reference-inconsistent) if they are not. Ref-consistent and ref-inconsistent are the names for no-call-rc and no-call-ri, respectively, used by versions of Complete Genomics pipeline versions prior to 1.7. We changed the names to highlight the fact that these alleles contain no-calls.

In some cases, one may wish to be conservative and consider any such region entirely no-called, and thus neither a match nor a mismatch between sample and reference.

## What causes loci to be partially (or half) called?

For a small fraction of assembled loci, there can be support for one allele (reference or variant) but some ambiguity as to whether the other allele is supported by the data. This can happen, for example, when very few reads from one of the two chromosomes are seen. Also, in regions of low coverage, the algorithm may see reads consistent with a single allele (i.e., consistent with a homozygous call), but may judge that too few reads in total were seen to have had a good chance of sampling both chromosomes. In these cases the variation file reports a partial or half-called locus; a fully resolved allele (reference or variant) on one chromosome, but a no-call on the other.

## Does Complete Genomics assume a diploid model when calling small variants?

Diploidy is not assumed when calling small variants. The small variant caller considers heterozygous hypotheses at a wide range of allele frequencies between 20% and 80%, including but not limited to 50%. This is to accommodate small variants that occur at sites of copy number variation as well as in samples that are not pure: for example, due to tumor heterogeneity or sample mosaicism.

Note that two variant scores are provided for each called allele: one derived from the probability of this call assuming variable allele fractions (*allele1VarScoreVAF*, *allele2VarScoreVAF*, or *varScoreVAF*), and one derived from the probability of the given call assuming equal allele fraction, or diploidy (*allele1VarScoreEAF*, *allele2VarScoreEAF*, or *varScoreEAF*). Additionally, triploid hypotheses are considered in the assembly optimization step, and the step of alleles in an ***evidenceInterval*** record may describe a triploid top hypothesis. Regardless of models used to call small variants, the results of variation intervals where the top hypothesis is triploid will still be presented as two alleles at each locus.

Mitochondria and sex chromosomes are handled as special cases: see "How does Complete Genomics handle mitochondrial sequences?" and "How does Complete Genomics handle the sex chromosomes?"

## Which score do I use when filtering my small variant calls for quality?

The *varScoreVAF* and *varScoreEAF* are the best indicators of variant quality (these correspond to *allele1VarScoreVAF*, *allele2VarScoreVAF*, *allele1VarScoreEAF*, and *allele2VarScoreEAF* in the master variations file ***masterVar***). The *varScoreEAF* best reflects the quality of a call for variants at 50% allele fraction, while the *varScoreVAF* is a better score for variants at low allele fraction.

For reference-called positions, Complete Genomics provides scores in the ***coverageRefScore*** files in the REF directory, rather than the ***var*** or ***masterVar*** files. The reference scores within that directory are the best indicator of the quality of reference calls.

For variants, select the score based on the type of sample being studied, as follows:

- The *VarScoreEAF* is based on the assumption that the sample is diploid. It will generally be the best fit for samples that are homogeneous and are not expected to exhibit gross copy number changes. Examples include population studies, samples representing congenital disorders, and the matched normal samples in tumor-normal pairs.
- The *VarScoreVAF* is based on the assumption that the alleles in the sample can vary greatly from being diploid. It will generally be the best fit for samples that are heterogeneous and/or exhibit gross copy number changes. Examples include most tumor samples as well as samples expected to contain mosaicism.

When there is not enough information about the sample to determine the best score approach, Complete recommends using *varScoreVAF* as the general-purpose variant score.

## How does Complete Genomics handle mitochondrial sequences?

Mitochondrial sequences are treated as having a ploidy of 1. The circular nature of mitochondrial DNA is taken into account so that coverage is not suppressed at the start and end of the mitochondrial chromosome.

## How does Complete Genomics handle the sex chromosomes?

In males, the majority of the X chromosome is treated as having a ploidy of 1 while in females the X chromosome has a ploidy of 2. In males, variants in the pseudoautosomal region of the Y chromosome are reported on the corresponding regions of the X chromosome, where ploidy 2 is assumed. The pseudoautosomal region of the Y chromosome itself will be indicated as "PAR-called-in-X" in the variant file.

## How does Complete Genomics handle regions of the genome where multiple divergent references are known, such as MHC?

Areas of the genome that are highly variable are assembled using the default reference sequence at this time. Therefore, the no-call rate may be higher than other locations of the genome. We are looking into improved calling methods for these regions in the future.

## I see loci in the variant file with the same start and end position and a "?" for the sequence. What is a zero length no-call?

This is a locus in the genome where we cannot rule out the possibility that there is an insertion present.

## Does Complete Genomics call variants where multiple nearby bases have changed?

Yes, we find both length-conserving and length-changing block substitutions in our assembly process, at both homozygous and heterozygous loci. In many genomes, we find a number of these substitutions where a portion of the locus is a known variant (such as, SNP), while the remainder of the substitution is novel and called with high confidence.

## What source is used for calling variants?

We presently call variants relative to the NCBI reference in each genome sequenced. This facilitates comparison between any set of samples desired.

Complete Genomics develops an open source tools package, Complete Genomics Analysis Tools (**cgatools**), for downstream analysis of Complete Genomics data. Currently, **cgatools** contains tools for comparing variants between two genomes. We are working on additional methods for multiple sample and other comparison tools. For more information on **cgatools**, see the Complete Genomics website: www.completegenomics.com/sequence-data/cgatools.

## Are known variants (such as those in dbSNP) considered when assembling and calling loci?

Yes. Known variants are used as a supplementary source of seeds for local *de novo* assembly when searching for candidate variants. Knowledge of which variants are known and which are novel is not used in variant scoring. The set of known variants used to supplement local *de novo* assembly is comprised of indels and short block substitutions from dbSNP (dbSNP 130 for Build 36, and dbSNP 132 for Build 37) and the Complete Genomics Diversity Panel (69 genomes using assembly pipeline version 1.10).

### Can Complete Genomics use a different reference or directly assemble one genome against another?

No, currently Complete Genomics does not use another reference other than NCBI nor can we directly assemble one genome against another.

### Does Complete Genomics remove duplicate reads?

Any pair of DNBs from the same library that have at least one arm whose initial mappings have a common mapping (based on chromosome, offset, and strand) are considered candidates for de-duplication. Each pair of candidate duplicate DNBs is evaluated for sequence similarity. If the DNBs have at most four discordances for each arm (up to two discordances per read), allowing the gaps to differ by up to two bases (except for the clone end reads), they are considered duplicates, and one of the two DNBs is selected at random for removal. De-duplication is performed such that it does not affect the initial mappings or coverage, but it does apply to all of the following:

- Variations, variation scores, evidence scores, and reference scores
- Contents of evidence files
- Read counts in ***masterVar*** files
- Evidence mappings (BAM)

## Annotations

### What annotations does Complete Genomics provide describing the variants called?

Complete Genomics currently annotates called variants using five external data sources:

- dbSNP: to cross-reference variants
- RefSeq: to identify overlap with and impact to genes (as well as some ncRNAs)
- Database of Genomic Variants (DGV): to cross-reference called CNVs
- COSMIC: to cross-reference variants
- miRBase: to identify variants that overlap with microRNAs

### How does Complete Genomics compute the functional impact of variants in coding regions?

Complete Genomics uses the alignment data from the ***seq_gene*** file contained in a NCBI annotation build (see "What is NCBI build 36.3 (or 37.2)? How does it differ from build 36 (or 37)?") to compute the location of the variant within the RefSeq mRNA sequence. The variant sequence is then substituted into the corresponding location within the mRNA sequence. The resulting nucleotide sequence is translated to protein sequence using the appropriate codon table (standard code or vertebrate mitochondrial code). This permuted protein sequence is then compared to the RefSeq protein sequence, and the impact, if any, is noted. Synonymous changes are noted as such.

An important consideration is that RefSeq mRNA sequences often differ from corresponding sequences in the reference genome. This difference is because the reference genome is derived from a single individual at any given locus (the same individual was not used for the entire genome), meaning that it will contain alleles different than those seen by RefSeq curators, particularly when the reference genome allele is the rare allele. Complete Genomics explicitly annotates calls that are variant with respect to the reference genome but yield protein sequences identical to that reported in RefSeq.

## What version of the reference genome database is used? What versions of the annotation databases are used?

Customers can choose either NCBI build 36 (corresponding to Hg18) or NCBI build 37 (Hg19) as the reference genome. The gene annotations are those provided in each build. For Build 36, the known-variant annotations are from dbSNP 130 and for Build 37 the known-variant annotations are from dbSNP 131 or dbSNP 132, for data generated on assembly pipeline version 1.11 or greater. Prior to software version 1.8, the genome build was NCBI Build 36 and dbSNP 129. The format version was in the #VERSION header.

## What is NCBI build 36.3 (or 37.2)? How does it differ from build 36 (or 37)?

NCBI build 36 refers to the genome build—the FASTA files describing chromosomes 1-22, X, and Y. Build numbers such as 36.1 describe annotation releases provided by the NCBI, where features such as mRNAs from RefSeq are mapped to the genome build. There may be multiple annotation releases, each with a different version of the annotation source data, all mapped to the same reference genome build. For example, builds 36.1, 36.2, and 36.3 all contain annotations mapped to genome build 36, but the RefSeq sequences used in each are from three different points in time. Similarly, build 37.1 contains annotations mapped to genome build 37.

NCBI Build information:

- Release notes for all genome and annotation builds

  www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html
- Information describing annotation build 36.3

  www.ncbi.nlm.nih.gov/projects/mapview/stats/BuildStats.cgi?taxid=9606&build=36&ver=3
- Information describing annotation build 37.1

  www.ncbi.nlm.nih.gov/projects/mapview/stats/BuildStats.cgi?taxid=9606&build=37&ver=2

Complete Genomics uses the RefSeq alignments in NCBI's annotation builds for functional annotation of variants. These alignments can be found in the following locations:

- Build 36.3:

  ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/mapview/seq_gene.md.gz
- Build 37.2:

  ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/mapview/seq_gene.md.gz

The nucleotide and amino acid sequence of the RefSeqs used by NCBI in the production of annotation build can be obtained interactively using the Entrez website (http://www.ncbi.nlm.nih.gov/Entrez/) or in batch using NCBI eUtils (http://eutils.ncbi.nlm.nih.gov/). The accession and version specified in the ***seq_gene*** file must be used in your query to ensure you are using the exact sequence described in the alignments.

# Call Thresholds and Scoring

## What scores are produced for variant calls? What score thresholds are used?

In all putatively variant regions, the assembler considers many hypotheses (essentially, possible consensus sequences) and computes probabilities of the observed read data under each these hypotheses. We perform a likelihood ratio test between the most likely hypothesis and the next most likely, and we express this score in decibels (dB). Bioinformaticists will recognize dB as the basis of the Phred scale: 10 dB means the likelihood ratio is 10:1, 20 dB means 100:1, 30 dB is 1000:1, etc. The variant scores factor in quantity of evidence (read depth), quality of evidence (base call quality values), and mapping probabilities. The column header for the variation score is "total score" in the variations file.

Scores for variants are not calibrated on an absolute scale to error rate. A score of 30 dB does not necessarily indicate that the P(error)=0.001.

10 dB is the minimum score for calling a homozygous variant and 20 dB is the minimum for a heterozygous variant. These thresholds are chosen to maximize sensitivity to low quality variants, but also mean that filtering by variant quality is of paramount importance. The **cgatools** varfilter can be used to filter by quality score. For more information on filtering by quality score, see "Complete Genomics produces separate scores for the two alleles at a locus. How do I interpret this information?"

## Can I compare scores across variants?

Scores for variants can be compared, but only robustly within a specific class of variant, as the probability model for each class of variant is slightly different. For example, a SNP call with a score of 60 is more likely correct than a SNP call with a score of 50, and a deletion with a score of 60 is more likely correct than one at 50. However these numbers do not precisely indicate the strength of evidence for either of the SNPs relative to that for either of the deletions.

The score calibration produced by Complete Genomics may provide insight on actual quality of scores, using *varType*, *zygosity*, and local coverage. For more information on score calibration, see *Calibration Methods*, available on the Complete Genomics website.

## Complete Genomics produces separate scores for the two alleles at a locus. How do I interpret this information?

Roughly, the higher of the two scores can be considered the strength of evidence that this allele is present at the locus. The lesser score similarly indicates the strength that we have fully called the complete diploid genotype correctly at the locus.

## What criteria are used to call a region homozygous reference?

We report the strength of evidence for homozygous reference calls in the ***coverageRefScore*** files. See "What is the "Reference Score" and what is it used for?" Reference score is one metric used to flag regions of the genome for *de novo* assembly. (A second metric based on De Bruijn graphs is designed to also flag regions for *de novo* assembly containing indels and block substitutions, where mapping individual reads is more difficult.)

To call an unassembled region homozygous reference, a reference score of 10 dB or higher must be achieved. To achieve this score, typically at least four reads need to map to the position and be consistent with the reference sequence, although the precise number needed depends on mapping probabilities, base-call quality scores, and the number of concordant and discordant calls. In addition, a small number of homozygous reference regions are determined by the assembler.

## Can I change calling thresholds to detect more true variants (at a higher false-positive rate) or detect fewer (at a higher false-negative rate)?

The variant file includes variant calls made at a minimum threshold designed to allow for high sensitivity (see What scores are produced for variant calls? What score thresholds are used?). To limit the number of false positives in the results, scores are provided to filter out variants called with lower confidence. For more information, see Complete Genomics produces separate scores for the two alleles at a locus. How do I interpret this information?

The variants shipped by Complete Genomics all meet our minimum thresholds and we do not report possible variants below that level. Thus it is not possible to drop thresholds below this level.

## What depth of coverage is required to call a variant allele?

Various parameters are taken into account when calling and scoring variant alleles, such as coverage depth, mapping probabilities of reads, and base-call qualities. Heterozygous variants marked as `VQHIGH` generally require at least two high quality, well mapped reads per allele. Homozygous variants marked as `VQHIGH` generally require at least seven reads. Variants marked as `VQLOW` may have fewer reads supporting the call, and are accompanied by a lower score indicating the lower confidence in the call.

Note that most variants are called with much greater read count support (depth of coverage) than these minimums, and we find that the scores (reference scores and variant scores) are excellent indicators of the relative quantity, quality, and consistency of evidence.

## What criteria are used to no-call a region on one or both alleles?

Sites that do not meet the criteria to be called either homozygous reference or variant are considered no-calls. Loci can be partially (or half) called, where one allele sequence is determined but the other is not.

Alleles can also be "incompletely" called in some cases, which is a different behavior than a partial call. In this case, some of the bases of an allele are determined at the minimum threshold but others are not. These alleles will have "N" or "?" in their sequence, and will be marked as "no-call-rc" or "no-call-ri". Some loci are considered partially called because one of the two alleles is incompletely called.

## How are duplications and highly conserved repeats in the reference handled?

Mate-pair information is used both in mapping and to recruit reads for assembly. Even when the initial scan indicates that one or both ends of a clone may have multiple possible locations in the genome, the pair may have a single location consistent with the library's known orientation and distance of the read ends. We are thus able to assign many reads to assemblies even in non-unique regions and accurately call variants in those regions.

Even when a single location for a mate-pair is not indicated by the initial mappings, we can allow these reads to participate in more than one assembly, weighted by their mapping probabilities. Recall that assembly is both more sensitive than initial mapping (as it can allow greater degrees of mismatch and yet still align) and more stringent (as it demands that the accepted set of reads consistently explain the final consensus sequence). We have found that this approach allows us to accurately call variants in many (certainly not all) duplications. When a read contributes to variation calls at multiple loci due to sequence similarity, the scores for all affected calls are adjusted down to reflect the correlated evidence. This is not only reflected in the variant scores but also is reported in the correlations file contained in the EVIDENCE folder. If the correlations are too high, which happens when the duplicate regions cannot be well discriminated, then scores will be below threshold and some or all variations within the correlated regions are no-called.

# Data Interpretation and Sample-Sample Comparisons

## How do I identify somatic small variants?

For samples submitted through the Cancer Sequencing Service, each tumor sample is compared to the baseline sample within the pair or trio. The small variants that are unique to the tumor genome and not present in the normal can be isolated by looking at either the *somaticRank* column or the *somaticScore* column in the master variations file. The *somaticRank* column provides the estimated rank of the mutation amongst all true variants for the specific classification of variant. The value is a number between 0 and 1, and is empty for mutations that are not somatic. The *somaticScore* column indicates the quality for all somatic mutation calls, and is empty for mutations that are not somatic.

For samples submitted through the Standard Sequencing Service, somatic events can be identified using **cgatools** calldiff, which includes somatic scoring.

## How do I find the variants which might be disrupting or changing a known protein coding gene?

The information provided by Complete Genomics in the ***geneVarSummary*** and ***gene*** files may help. See *Complete Genomics Data File Formats* for details. Alternatively, you can use the coordinates in the variation file to compare against any database of annotations you wish.

You can find this documents on the Complete Genomics web site: http://www.completegenomics.com/sequence-data/download-data/

## Are SNPs or indels called more likely to be true-positives?

In our *Science* paper (Drmanac et al, *Science* 2010) we showed that the false positive rate for indel variants was somewhat higher than that for SNPs at the thresholds used. This is consistent with additional data, including family studies, which we have analyzed since. We expect these methods to continue to improve over time.

You can access the paper at www.rdrmanac.com at no charge.

## Are novel vs. known variants more likely to be true-positives?

Random sequencing errors are most likely to appear as novel variants. Depending on the goals of the analysis, one's statistical prior on the P(error) of a novel variant call might be greater than that for a known variant call.

## How do I find variants that might be disrupting or changing a known locus that is not a protein coding gene?

The ***gene*** file provided by Complete Genomics summarizes changes in the coding portion of transcripts annotated on the genome in the NCBI build. For other annotated loci in the genome, one would need to look in the variations file by chromosome and position.

## How do I separate novel vs. known SNPs and indels?

The dbSNP file provided compares the results of a Complete Genomics assembly to known variants in dbSNP, and reports a genotype of each site based on the Complete Genomics sequence data. The versions of dbSNP that we are presently using (for Assembly Pipeline version 1.11) are 130 for NCBI Build 36 and 132 for Build 37 (note: dbSNP Build 130 incorporated the 1000 genomes project data). As of pipeline version 1.8 release, we have added the dbSNP version number for when each SNP was added to the database. This can be helpful for filtering novel SNPs from different dbSNP database releases. When looking at these data, please keep in mind that most estimates of the error rate in dbSNP are relatively high, as are most estimates of the error rate in publicly available dbSNP genotypes of reference samples. As these rates are generally thought to be higher than the error rate of Complete Genomics sequence data, a number of discrepancies are to be expected.

## How do I compare variants between two or more samples?

Complete Genomics has an open source tools package, Complete Genomics Analysis Tools (**cgatools**), for downstream analysis of Complete Genomics data. For more information on **cgatools**, see the Complete Genomics website: http://www.completegenomics.com/sequence-data/cgatools/.

## When is comparing SNPs between samples problematic?

As discussed in the section above, we believe Complete Genomics has excellent sensitivity to detect not only SNPs but also insertions, deletions and block substitutions. However as we look at many genomes with this method, we discover that the rate of non-SNP alleles we call is high, and in some cases they occur at loci with simple SNP variants on the other allele. This leads to some complexity when comparing samples.

To illustrate, we'll re-use an example shown previously and refer to it as individual A:

```
Position:         123 456 789 012
Reference:        TAG TCG CCT ACG    locus includes bases 4 to 6
Allele A1:        TAG CAC CCT ACG    3 base block substitution
Allele A2:        TAG TTG CCT ACG    one ref + one SNP + one ref
```

This same location in another genome (individual B) might contain the following called as two separate loci because the changes are further apart:

```
Position:         123 456 789 012
Reference:        TAG TCG CCT ACG
Allele B2:        TAG CCG CCT ACG    locus #1 at base 4, one het SNP
Allele B1:        TAG TCG CCA TGG    3 base het sub, locus #2
```

Methods to handle such situations vary depending on the scientific goal. Questions for choosing a method would include whether you wish to consider, for example, Allele B2 above (a single SNP) the same variant as the corresponding part of Allele A1 (the block sub, which includes the same SNP). If so, one would consider A1 as having a second variant locus at position 6, distinct from the other SNP. Alternatively one might wish to consider all four of these alleles shown as distinct versions of this locus.

We see a few modes of comparing data at this level of detail that have different uses. One method is to maximally break loci into SNPs, which may be required when comparing Complete Genomics data against (array or sequence based) data sets focused on SNP calls. Another would be to group these loci together into blocks when looking across samples to check for conservation of the entire region.

Complete Genomics also has an open source tools package, Complete Genomics Analysis Tools (**cgatools**), for downstream analysis of Complete Genomics data such as performing comparisons of SNP calls between two samples. For more information please contact support@completegenomics.com.

## How do I compare indel variants between two or more samples?

First, note that the discussion in "When is comparing SNPs between samples problematic?" applies to non-SNP variants as well.

A further complication arises in indels and in non-length-conserving block substitutions. The exact point where any algorithm will define the start and stop of the change is based on rules, but even small differences can move the start and end coordinates of a variation, making comparison based on coordinates more difficult. For example:

```
Reference:         ATAATTTTTTTTTGTGTGTGT
Allele 1:          ATAATTTTTTTT-GTGTGTGT
Allele 2:          ATAATTTTTTTTT-TGTGTGT
```

Homopolymers and simple sequence repeats (SSRs) (such as AAAAA, CACACA…, or TAGTAGTAG…) present most obvious examples of this problem, where the choice of indel point is essentially arbitrary. While a fixed rule for defining that point does work well, it fails to provide consistency when a handful of other sites in the SSR sequence have changed (errors or real variation), which can greatly influence the alignment.

For example, consider a complex variant, spanning bases 12 through 18 in the alignment shown below.

```
                    1234567890123456789012
Reference:          GGAACTGAACA-----GCTAGC
Allele A1:          GGAACTGATAAGAAATGCTAGC
Allele A2:          GGAACTGAAGA-------TAGC
```

With a single base change, Allele #2 (and the entire diploid locus) could be assigned a different start and end position (10 and 16):

```
                    1234567890123456789012
Reference:          GGAACTGAACG-----GCTAGC
Allele B1:          GGAACTGATAAGAAATGCTAGC
Allele B2:          GGAACTGAA-------GCTAGC
```

Complete Genomics also has an open source tools package, Complete Genomics Analysis Tools (**cgatools**). Use **cgatools** for downstream analysis of Complete Genomics data, such as performing comparisons of indel calls between two or more samples. For more information please contact support@completegenomics.com.

## How should I handle alleles with N's or ?'s when comparing variants between two samples?

In many cases, one may wish to be conservative and consider any such region entirely no-called, and thus neither a match nor a mismatch between the samples.

For more precision, one can use the notion of compatible vs. incompatible alleles. Alleles are "compatible" if one can align the two and in doing so does not reject the hypothesis that they are the same. Incompatible alleles are those that must be different according to this type of analysis. For the reasons described above, an alignment based analysis is required to avoid falsely calling sites different that are in fact compatible.

## I am doing a family study. Can I use the constraints of Mendelian inheritance to further reduce errors? How?

Yes. We note a few things to take into account. First, simple Mendelian constraints on a variant by variant basis will detect a number of errors, but certainly not all, particularly in smaller families. A more powerful method has recently been presented by the Institute for Systems Biology using Complete Genomics data (Roach et al, *Science* 2010). This method involves first using the variant data to build a complete high-resolution recombination map of the family. This map allows greater power to detect errors than the simpler method. With families larger than a trio the power of this method to detect errors can be quite high.

# Assemblies and Evidence

## How big are local assemblies and how much of the genome is typically assembled this way?

Complete Genomics local *de novo* assemblies are typically 30-40 bp although they can be smaller or much larger (100s of bp). Approximately 5-10% of genome is typically assembled.

## Do variations reported correspond one-to-one with individual assembled regions? What are the implications of this on phasing closely neighboring variants?

No, one assembly can contain multiple variant loci. It's possible to phase closely neighboring variants if they are inside the same local assembly. The variant file has a *haplink* column that indicates phasing, if it can be determined.

## How do I see the assembly around a variant call?

By chromosome and position, you can find the appropriate row in the corresponding ***evidenceIntervals*** file. This data provide you with the assembler results (essentially, consensus sequences) for each allele, with a gapped alignment of that result against the reference. The underlying reads in each assembly can be found in the corresponding ***evidenceDnbs*** file, by looking up records (rows) using the evidence interval ID found in the evidenceIntervals file.

## Why is the reference allele always included in an evidence interval even when it is not called (that is, when neither allele appears to be reference sequence)?

The possibility of one or both alleles at a site being reference is always considered as a possible hypothesis in the likelihood ratio tests. Essentially, we demand that the data disprove this hypothesis, which is an appropriate null in that most sites in the genome of any sample are indeed reference.

As mentioned in "What scores are produced for variant calls? What score thresholds are used?", the score reflects separation between the top hypothesis and a hypothesis of homozygous reference. Inclusion of the reference allele in the evidenceIntervals file can help one understand how strongly the reference was rejected.

## Why are there reads in the evidenceDnbs file assigned to a particular locus that are not mapped to that location in the initial mappings files? Similarly why are there reads mapped to a location that are not in the evidenceDnbs file?

The assembly process has both more sensitivity and specificity than the initial mapping process. Reads which were sufficiently different from the reference (for example, those containing many indels or groups of SNPs) or which had multiple possible initial mapping locations may not be initially mapped to the variant site but can be brought into assembly using the mapping of the corresponding mate-pair.

Conversely, reads initially mapping to a region but which prove inconsistent with the preponderance of evidence in an assembly can be down weighted or excluded. Moreover, reads must have their mate-pair mapped nearby to a variant region to participate in an assembly, and thus reads with only one end mapping during the de novo assembly process are presently always excluded from assembly.

### How do I find the evidence underlying a site called homozygous reference? How do I know if another interpretation (such as variant) might be possible?

Complete Genomics does not produce assemblies (evidence intervals) for regions of the genome where the mapped reads are highly consistent with the reference sequence. Furthermore, assemblies are not reported for regions where variants are not found (this is an unusual case, however). Thus, variation scores are also not produced for these regions.

Instead, Complete Genomics computes a "reference score" for every base in the reference genome that is reported in the coverageRefScore file. This score indicates whether the corresponding mapped reads are consistent with the reference sequence (positive values) or not (negative values). This score is an excellent predictor for the strength of evidence for homozygous reference calls. See "What is the "Reference Score" and what is it used for?"

### How do I see evidence for any possibility other than the called variants at a particular locus?

Complete Genomics provides an assembly for the reference allele for all sites, including those called diploid variant.

Determining evidence for other alternate hypotheses is not easily supported by our assembler output at this time. Users would need to perform read level analysis, essentially recapitulating the details of the recruitment and assembly processes at a site (as described in Drmanac et al., *Science* 2010).

You can access the paper at www.rdrmanac.com at no charge.

### Why are reads present in multiple evidence intervals?

In the case of repetitive regions (such as highly conserved segmental duplications or interspersed repeats) the algorithm can allow reads to participate in more than one assembly, weighted by their mapping probabilities. See "How are duplications and highly conserved repeats in the reference handled?"

## Coverage and reference score

### How should I interpret the depth of coverage in Complete Genomics' coverageRefScore files?

As of Complete Genomics pipeline version 1.7.1, our calculation of depth only included bases of reads which, based strictly on the initial mapping results, are highly likely to be uniquely placed. Thus, it generally undercounts actual coverage in areas of duplication or repeats. The initial mapping algorithm is also not resilient to indels nor high degrees of divergence between the sample and reference, so those regions are also undercounted.

There are many locations in the genome where initially unmapped DNBs can be and are correctly included in assembly as described in "Why are there reads in the evidenceDnbs file assigned to a particular locus that  are not mapped to that location in the initial mappings files? Similarly why are there reads mapped to a location that are not in the evidenceDnbs file?" These are not reflected in the

***coverageRefScore*** depth. In some cases, regions of 0 depth in the ***coverageRefScore*** files do get assembled and indeed can be accurately called. This is reflected in the evidence files.

In pipeline version 1.7.2 we added an additional "weighted depth" metric to the ***coverageRefScore*** files. This score may be better for some purposes (such as CNV detection) as it also counts non-uniquely mapped reads, giving them fractional counts corresponding to the confidence of each mapping. Like the unique depth metric, it does not consider post-assembly depth. This value is used as input to our CNV detection pipeline.

In pipeline version 1.10 we added a "GC-bias corrected depth" metric to the ***coverageRefScore*** files. The depth-of-coverage approach we take to call CNVs assumes that the number of reads mapping to a genomic region is proportional to the genomic copy number of the region. This assumption is violated when distribution of mappings is biased (for example, in high/low GC content in the DNA sequences). To correct for this bias, coverage is adjusted for GC content calculated over a fixed window. In pipeline version 1.10 a "gross weighted depth" metric was also added to the ***coverageRefScore*** files. This value represents the number of half-DNBs which may map to this location, each weighted by their mapping weight ratio.

## What is the "Reference Score" and what is it used for?

Complete Genomics computes a value called the reference score reported in the ***coverageRefScore*** file. This score indicates whether the corresponding mapped reads are consistent with the reference sequence (positive values) or not (negative values). This score is an excellent predictor for the strength of evidence for homozygous reference calls.

Similarly to the method by which variant scores are computed, the reference score is the log-likelihood ratio of P(ref) over P(non-ref), expressed in dB, where the P(non-ref) involves examining only a limited number of alternate hypotheses. These include all possible SNPs at every position in homozygous and heterozygous form, plus, at selected positions, one-base insertions and deletions, as well as some changes in homopolymer or tandem repeat length. This computation is performed based on the initial mapping results and, like the variation scores, is not precisely calibrated to P(error). Reference scores are also not precisely calibrated to variation scores.

In spite of the lack of calibration, a reference score in one sample can be considered against the variation score of another sample to assist in sample-sample comparison, particularly when asking whether a variant seen in one sample might be a false negative in another.

## I see very high levels of depth in some locations. Why?

This happens at some repeats. The centromeric regions are the most extreme example. Many repetitive genomic regions have the "overflow" flag set on all reads in the mappings file as the read maps to too many sites to be computationally tractable, and thus have no mappings reported: they can appear as coverage zero.

## Why does the coverage vary from one base to the very next one?

This is a consequence of mapping reads with intra-read gaps and is consistent with our knowledge of the biochemistry and alignment properties of these reads.

# Glossary

## Allele (as used in variations file)

An arbitrary designation of one diploid allele over another in a variations file. See "Each variant allele has been identified as allele "1" or "2". Does that mean that all of the allele 1 variants are located on the same parental chromosome?"

## dB (decibel)

A log scale used by Complete Genomics for expressing probabilities and likelihood-ratios. dB are well known to bioinformaticians as the basis of the "Phred scale". See "What scores are produced for variant calls? What score thresholds are used?" Formally, the value of an likelihood-ratio $R= P_1/P_2$ expressed in dB is $10 \times \log_{10} R$. In cases where dB is used to encode an error probability P (as in a basecall quality score or a mis-mapping probability) the score is expressed as $-10 \times \log_{10} P$. In both cases bigger scores in dB are "better".

## DNB

DNA Nano Ball, an individual library construct. Referred to as a "clone" on many other platforms.

## DNB Arm

One end of a DNB insert sequence, from either side of the mate-pair gap. Called an "end" or "read end" or "paired end" on other platforms.

## Evidence

The assembly underlying a small variant call. It includes the final allele sequences called, and for each the alignments of the supporting DNB to that sequence.

## Evidence Interval

The coordinates on the reference genome corresponding to an assembled region.

## Indel

Short for "Insertion or Deletion"

## Initial Mapping

By comparison with some other pipelines used with other types of data, the Complete Genomics bioinformatics process involves an initial mapping followed by a refinement of these mappings by local *de novo* assembly. The assemblies, and not the initial mappings, represent the final determination of the location of a DNB. See "Complete Genomics Service FAQ" for more information.

## Locus (as used in variations file)

A region of the genome containing variations on either or both alleles. An arbitrary threshold is used to determine when nearby variation are part of the same loci or separate loci.

## No-call-rc, No-call-ri, ref-consistent, ref-inconsistent

See "Please explain "no-call-ri", "no-call-rc", "ref-consistent" and "ref-inconsistent". How should I use these?"

## Read Gap, Mate Gap

Complete Genomics reads have two kinds of gaps. There are three specific positions in each DNB arm where the bases do not neighbor in the source DNA: these are intra-read gaps. Also, there is a larger mate-gap region (300-400bp+) in between the two reads from one DNB, as is the case for other paired-end and mate-pair sequencing methods. See the Complete Genomics technology whitepaper at www.completegenomics.com.

## RefScore, Reference Score

See "What is the "Reference Score" and what is it used for?" and "How do I find the evidence underlying a site called homozygous reference? How do I know if another interpretation (such as variant) might be possible?".

## Sub, Delins

A "sub" is a block substitution where a series of reference bases are replaced with another series of bases. This event may or may not be length conserving. In data generated with Complete Genomics pipeline versions prior to 1.7 "sub" was denoted as a "delins".