Small Variant Assembler Methods

File format v2.0

Software v2.0.0

March 2012

Table of Contents

Introduction

Complete Genomics small variant assembler is a component of the Analysis Pipeline that is used to detect, call, and score small variants (SNPs, insertions, deletions, and substitutions) that are usually up to a few tens of bases long. The variant calling process begins with a fast mapping of reads to the human reference genome, identifying regions of the genome that likely differ from the reference. Using alignment information of their mates, individual reads that lie in these regions of interest are recruited and local *de novo* assembly is performed by constructing and traversing a De Bruijn graph. The paths discovered during local *de novo* assembly are used as seeds into an optimization process, where hypotheses are scored using a Bayesian framework. Variants are then called—based on the best scoring hypotheses—that have a likelihood exceeding a significance threshold. Variant scores are calculated based on two likelihood models, variable allele fraction (VAF) and equal allele fraction (EAF). Finally, ambiguity of variant localization in cases where variants are found in duplicative regions or regions of large-scale similarity is resolved by rescoring variants based on a correlation analysis. Variants and their scores are reported in Complete Genomics small variant files.

The goal of this document is to describe the process of the small variant assembler and present the details of the algorithms used. This document is written with respect to the Complete Genomics Analysis Pipeline version 2.0. We recommend that you become familiar with the following additional documentation as preparation for reading this document:

1. *Data File Formats* — A description of the organization and content of the format for complete genome sequencing data delivered by Complete Genomics. [www.completegenomics.com/documents/DataFileFormats-100357139.html]
2. *Release Notes* — The Analysis Pipeline Release Notes indicate new features and enhancements by release. [www.completegenomics.com/documents/ReleaseNotes-100358389.html]
3. *Human Genome Sequencing Using Unchained Base Reads on Self-assembling DNA Nanoarrays* — *Science* publication describing Complete Genomics proprietary sequencing technology, including brief methods and performance of our Analysis Pipeline. [www.completegenomics.com/customer-support/publications]
4. *Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads* — A publication in the *Journal of Computational Biology* describing the processes and methods used in Complete Genomics Small Variant Caller in Analysis Pipeline versions prior to version 2.0. [www.completegenomics.com/customer-support/publications]

We recommend that you are familiar with the DNB Generation Modeling and Bayesian Modeling, which are the foundation of the hypothesis scoring used to call variants. This document includes information for both modeling types in appendices.

Additional documentation is available in the Support section of the Complete Genomics website:

www.completegenomics.com/customer-support/documentation

Steps in the Small Variant Assembler

The process of discovering, calling, and scoring small variants within a re-sequenced genome can be described by the following six steps:

1. **Mapping Reads to Human Reference Genome —** initially maps reads in a constrained process that does not allow for insertions and deletions and only allows a small number of mismatches.
2. **Interval Discovery —** identifies hypotheses: regions that are likely to differ from the reference genome where local *de novo* assembly procedure and sequence optimization will be applied in the subsequent step.
3. **Optimization —** gathers individual reads that are likely to lie in the regions of interest using mate alignment information and perform a local *de novo* assembly. Results serve as input into an optimization procedure that uses a Bayesian framework to compute the likelihood of each hypothesis.
4. **Variant Calling —** determines the most likely hypothesis from a set of scored hypotheses, generated during the optimization step, to either call variants or to make no-call.
5. **Hypothesis Rescoring —** computes scores (*varScoreVAF* and *varScoreEAF*) to indicate confidence for each called variant. This step was introduced in Analysis Pipeline version 2.0.
6. **Correlation Filtering —** identifies regions where the probability computations based on a single interval are likely to be unreliable due to sequence similarity with other regions of the genome. These regions are turned into no-calls to reduce the false positive rate of variant detection in repetitive regions.

Step 1: Mapping Reads to the Human Reference Genome

All mate-pair constraint-satisfying paired-end mappings (i.e., DNA Nanoballs (DNBs™) where both mate arms map) are used to detect small variants. For each arm of each DNB, the mapping process is able to guarantee finding all perfect matches and all 1-discordance (k=1) matches. It is able to find a substantial fraction of the one-arm matches up to k=5.

At this step, DNBs of the following characteristics are filtered out:

- DNBs where both clone ends may map to the same genomic position are de-duplicated.
- DNB arms with too many good reference mappings may be marked as 'overflow', and the mappings of such DNBs are omitted from participating in small variant detection.
- DNB mappings with no consistent mate mapping are omitted.

The result of this step is:

- Initial mappings provided in the ***mapping_[SLIDE-LANE]_00X.tsv.bz2*** file of the MAP directory in the data package.

Step 2: Interval Discovery

The interval discovery process involves identifying hypotheses by scanning the genome for regions that may harbor SNPs or short indels, where a fast version of local *de novo* assembly is used to find indels. This process involves trying each region for:

1. All possible one-base variations for SNPs in homozygous and heterozygous form.
2. All possible one-base insertions and deletions where local *de novo* assembly indicates even slight evidence of an indel existing in homozygous and heterozygous form.

3. All single-copy insertions or deletions in a tandem repeat period up to 10 bases in homozygous and heterozygous form where local *de novo* assembly yields evidence for indel.
4. Known indels and short block substitutions taken from the Complete Genomics Diversity Panel and dbSNP.
5. Short indels (of several nucleotides) discovered by a fast version of local *de novo* assembly.

For each hypothesis, the assembler computes the likelihood of that hypothesis being correct, *L(G)*. At most locations, *L(G)* is computed to be negative, indicating that the reference sequence is more likely than any other variations at that position. Where a one-base variation is present, *L(G)* is computed to be large and positive.

A "reference score" for each hypothesis is then computed as –max(*L(G)*) for every base. This value provides a confidence score for a call of homozygous reference outside of the interval. In regions harboring longer variations, *L(G)* for one-base variations is usually still negative, but to a much lesser degree than in regions where no variation is present. In this event, *L(G)* is used to indicate the presence of a nearby variation; such regions are marked for optimization in the subsequent step. Intervals that are longer than 200 bases are no-called without attempting optimization, as optimization becomes too computationally intensive.

The results of this step are:

- A set of variation intervals—genomic regions that may plausibly contain SNPs or short indels—that are investigated in greater detail in the Optimization stage.
- A reference score computed at each position in the genome, which gives an indication of the likelihood that a variant exists at any given base. A reference score of 10 or less marks the interval for optimization in the next step. Reference scores are provided in the ***coverageRefScore*** files of the ASM directory in a data package.

Step 3: Optimization

For each variation interval discovered in the previous step, the results of local *de novo* assembly and the set of known indels and block substitutions contained by the interval are combined. They then serve as input, or seeds, into a greedy optimization procedure, which searches for the most likely combination of alleles to identify the maximum likelihood hypothesis. Specifically, the optimization procedure is seeded by the most likely hypothesis, out of the following hypotheses:

- The reference hypothesis.
- The set of hypotheses discovered as plausible hypotheses by using a local *de novo* procedure, the input from Interval Discovery.
- The set of hypotheses in the set of indels and block substitutions assembled in the Complete Genomics Diversity Panel, as well as in dbSNP.

The first iteration of the optimization process starts with the top hypothesis out of all the seeds. Each iteration of the optimization evaluates the likelihood of each hypothesis produced by deviating from the starting point by a single-allele variation corresponding to a single SNP, one base indel, or one insertion or deletion that adds or subtracts a single copy of a simple repeat, such as homopolymer and dinucleotide run. Each subsequent iteration gets as input the best hypothesis discovered during the previous iteration. When an iteration of the optimization is unable to find a more likely hypothesis, we have converged at a local minimum and the optimization completes. This approach allows discovery of both isolated variations as well as any combinations of multiple SNPs and indels within an interval, and overlapping distinct variations on opposite haplotypes.

For most regions of the genome (the autosomes and chrX for females), the optimization procedure works on hypotheses with two alleles. When considering a heterozygous hypothesis, the optimization procedure finds the maximum likelihood allele fraction for each allele, such that the sum of allele fractions is 1. In general, as long as a single DNB in support of a non-reference hypothesis exists, the maximum likelihood heterozygous hypothesis will always be more likely than the homozygous reference hypothesis that is not constrained by hypothesis priors. For example, under the assumption that all base calls are equally good, if there are 99 DNBs in support of the reference and one DNB in support of the SNP, the hypothesis where $f_{ref} = .99$ and $f_{SNP} = .01$ is more likely than the homozygous reference hypothesis, by a minute amount.

Because of this fact, the small variant assembler in Complete Genomics Analysis Pipeline version 2.0 constrains the optimization so that the maximum likelihood allele fraction (variable allele fraction model) is used 1) only if the allele fraction is at least 0.2 for each allele, and 2) only if the maximum likelihood is at least 20 dB more likely than the hypothesis where one of the two alleles has allele fraction 0 (i.e., homozygous). If these criteria are not met, the heterozygous hypothesis is constrained so that the allele fraction is equal for all alleles (equal allele fraction model). Essentially, the caller uses a hybrid maximum likelihood allele fraction model. This allows the assembler to:

- Detect alleles present at low allele frequency, as long as there is strong support for them.
- Make homozygous calls where there is strong support that a homozygous hypothesis is more likely than a diploid heterozygous variant.
- No-call where there is little support or substantial conflicting support.

Upon completion of the optimization procedure, the triploid hypothesis considering the alleles of the top two diploid hypotheses is evaluated. If its likelihood is at least 20 dB greater than the likelihood of the most likely diploid hypothesis, the triploid hypothesis is considered to be the top hypothesis. Otherwise, the most likely diploid hypothesis is considered to be the top hypothesis. Note that because the optimization procedure works on a small region (up to 200 reference bases), it is unlikely there will be more than three distinct haplotypes in this region. Where this does occur, the small variant assembler may not be able to detect all the variants present.

The result of this step is:

- For each interval, a list of the most likely hypotheses, which is used as input into the next step where variations are called based on these values.

Step 4: Variant Calling

The variant caller step is responsible for turning the scored hypotheses from the optimization step into scored calls and no-calls. This process entails aligning variants to the reference genome, determining where to make a call and where to no-call, scoring each call, and producing phase information for nearby calls. A Bayesian model, described in "Appendix A: Bayesian Modeling," is used to compute a probability ratio for any two hypotheses from the optimization step; variant calls are then made based on the most likely hypothesis according to this Bayesian probability model.

Determining the initial set of calls

Based on the alignment of the top hypothesis, the initial set of calls is determined, and the variation interval is split into initial variant loci defined by the following rules:

- Calls that overlap by at least one reference base are merged into a single locus.
- Calls that have 0 reference bases (i.e., insertions) are merged with any adjacent locus.

After the variation interval is split into variant loci, the loci are coerced to the appropriate ploidy. For triploid hypotheses, each locus is separately coerced into a diploid hypothesis. Most triploid hypotheses can be coerced into diploid variant loci, as at each locus there are typically only two distinct alleles. Upon coercion of a triploid hypothesis into diploid loci, some phasing information (reported in the *hapLink* field of variation files) is lost. Variant loci with three alleles are no-called.

For each additional hypothesis within 10 dB of the top hypothesis, we align the hypothesis to the reference using the same rules as for the top hypothesis, except gaps may be preferentially placed at the same position as variants in the top hypothesis. For each such hypothesis alignment, we compare the aligned bases to the top hypothesis. At any position of discrepancy, we must no-call.

Initial scores are calculated for each call as the logarithm of the probability ratio (decibel separation, dB) of the most likely hypothesis compared to the next best homozygous hypothesis not containing a given candidate variation. If the score for a given variation exceeds a threshold (currently fixed at 10 dB and 20 dB for homozygous and heterozygous variations, respectively), the variation is called along its score. If the score is below the threshold, a “no-call” is reported for the corresponding portion of the reference.

For heterozygous calls, the call score is the difference between the top hypothesis score and the first hypothesis that is homozygous at the position of the call, but discordant with the call. Thus the score is more indicative of the existence of the call than the correctness of the call. As an example of this definition, consider the following:

Top Hypothesis (score 100):	ACAG-AAAAAAAATGC ACAGAAAAAAAAATGC
Next Hypothesis (score 30):	ACAG--AAAAAAATGC ACAGAAAAAAAAATGC
Reference Hypothesis (score 0):	ACAGAAAAAAAAATGC ACAGAAAAAAAAATGC

In this case, a heterozygous one-base deletion with score 100, rather than 70 would be called. Although there are 70 dB of support for the one-base deletion with respect to the two-base deletion, there are 100 dB of support for a non-reference variant. This way of defining the score yields an improved receiver operating characteristic (ROC) for somatic events where the germline sequence is reference, but the ROC for mismatching events is worse. The poorer ROC for mismatching events is mitigated by setting a threshold of 20 dB on the score to the next best hypothesis. As a result of this threshold, heterozygous calls at a lower score threshold than 20 dB cannot be called.

For homozygous calls, the call score of one of the calls is the difference between the top hypothesis score and the first hypothesis that is discordant with the call. The score of the other call is determined using the same rules as for a heterozygous call. In this way, the call with the lower score indicates the certainty that no other allele exists at this locus and the call with the higher score indicates the certainty that this allele exists at this locus.

When we apply the score to the call, we record the next best hypothesis that was used to determine the call score, as this hypothesis is rescored in the hypothesis rescoring stage.

The result of this step is:

- Initial set of calls, along with scores that are further considered in subsequent hypothesis rescoring and correlation filtering.

Step 5: Hypothesis Rescoring

The hypothesis rescoring stage is responsible for separately computing the variation scores determined under VAF and EAF models (*varScoreVAF* and *varScoreEAF* respectively), given the top hypothesis and the hypotheses identified by the variant calling stage. Additionally, this step allows us to achieve a reduction in the false positive rate by ensuring that individual DNBs cannot provide overwhelming support for a hypothesis.

Given the limited number of hypotheses this stage must score, it becomes computationally tractable to generally correct for the limitation within the Complete Genomics Small Variant Assembler that individual DNBs can provide overwhelming support for a hypothesis. For example, in the Optimization stage (Step 3), a DNB may map to the top hypothesis with no discordances, but have no mappings to the reference hypothesis. In this case, based on the approach presented in "Appendix A: Bayesian Modeling," the likelihood of the DNB given the reference hypothesis is α, which we currently set to 10^{-9}. However, the likelihood of the DNB given the top hypothesis where there is a good mapping may be higher than 10^{-3}. Thus, this DNB may support the top hypothesis by over 60 dB. This overwhelming support for the top hypothesis becomes a problem when these DNBs arise from unmodeled errors in the process of DNB generation from the sample (e.g., polymerase stutter in PCR amplification or formation of DNB chimerism). Thus, a goal of this rescoring step is to limit the influence of even the best single DNB, in accordance with a model that assumes even the best DNBs could arise due to artifacts in the process of DNB construction. The details of this approach are presented in the "Appendix B: DNB Generation Modeling".

The result of this step is:

- A set of variants, along with rescored *varScoreVAF* and *varScoreEAF* that are further considered in the correlation filtering step.

Step 6: Correlation Filtering

In all previous steps of the small variant assembler, each region of the genome is considered in isolation, and it is assumed that the rest of the genome is equal to the reference. In the majority of cases, this approach is sufficient, as most DNBs map uniquely to the genome. However, within segmental duplications and other regions of large-scale similarity where DNBs cannot be uniquely mapped, this approach leads to false positive variant calls. Because the DNBs cannot discern in which region of the genome these variants truly exist, no-call must be assigned to such regions. The correlation filtering step is responsible for no-calling the variants called in these duplicative regions, thereby substantially reducing false positive calls in repetitive regions.

In some cases two intervals at a time need to be considered due to DNBs having good mappings to both regions. Consider the two regions, 1 and 2, in the following genomes:

- G_1 differs from the reference in region 1 but is identical to the reference in region 2.
- G_2 differs from the reference in region 2 but is identical to the reference in region 1.
- G_{12} differs from the reference in both regions. It is identical to G_1 in region 1 and identical to G_2 in region 2.

In most cases the equality $L(G_{12}) = L(G_1) + L(G_2)$ will hold (the two intervals are uncorrelated), as DNBs supporting $G_1$are disjoint from the set of DNBs supporting G_2. However, there are three situations where the two sets of supporting DNBs are not disjoint:

- The two active regions are less than ≈ 40 bases, such that a single DNB arm can overlap both.
- The two active regions are at a distance approximately equal to a mate gap length such that a single DNB can overlap both.
- The two active regions are at any distance from each other in the genome, but are similar in sequence, exactly or approximately and a DNB can have good mappings to both regions.

In situations described above, a correlation term appears and $L(G_{12})$ no longer equals the sum of $L(G_1)$ and $L(G_2)$. The value of the correlation term can reveal information that contradicts the conclusions one would have reached by considering $L(G_1)$ and $L(G_2)$ in isolation. For example, in a pair of regions with high sequence similarity, one can have large, approximately equal values $L(G_1) = L(G_2) = L(G_{12})$. In this example, all three of the hypotheses are equally likely: the variant exists in region 1 and not region 2, the variant exists in region 2 and not region 1, or the variant exists in both regions. Thus, for each of the two possible variants, we have conflicting hypotheses with equal likelihood, one hypothesis indicating the variant exists and the other indicating the variant does not exist. Thus, we must no-call.

The result of this step is:

- A set of variants and their scores, *varScoreVAF* and *varScoreEAF* , that are reported in the ***var*, masterVarBeta*,*** and ***somaticVcfBeta**** files.

Appendix A: Bayesian Modeling

A Bayesian model is used to compute a probability ratio for any two hypotheses for a given re-sequenced genome. Variant calls made by the Complete Genomics small variant assembler always represent the most likely hypothesis according to this Bayesian probability model. Scores for called variants are computed from this Bayesian model, which takes into account:

- Quantity of evidence (read depth)
- Quality of evidence (base call quality scores)
- Mapping/alignment probabilities (selection of evidence)
- Empirical priors on gap sizes and discordance rates

The Bayesian probability model used by Complete Genomics indicates the likelihood of a hypothesis (H) given the set of reads, or DNBs (DNA Nano balls), present in the raw data relative to that of the reference genome (j):

$$\frac{P(H_i|DNBs)}{P(H_j|DNBs)} = \frac{P(DNBs|H_i)P_{prior}(H_i)}{P(DNBs|H_j)P_{prior}(H_j)}$$

For a typical human genome, substantial priors exist. For example, for any given base position, the hypothesis that a heterozygous SNP exists is only about 1/1000 as likely as the reference sequence. However, the Complete Genomics small variant assembler uses no prior information about hypothesis likelihood in computing the likelihood ratio. Thus we have:

$$\frac{P(H_i|DNBs)}{P(H_j|DNBs)} = \frac{P(DNBs|H_i)}{P(DNBs|H_j)}$$

$$\frac{P(H_i|DNBs)}{P(H_j|DNBs)} = \frac{\prod_{DNB \in DNBs} P(DNB|H_i)}{\prod_{DNB \in DNBs} P(DNB|H_j)} \qquad (1)$$

In the latter equation, we make the assumption that all DNBs are independent. However, this assumption is sometimes violated, such as in cases where the DNBs may split apart and be sequenced on several spots on our patterned DNB arrays, or a single fragment of DNA may be duplicated in the library construction process and result in multiple DNBs. Thus, DNBs are de-duplicated by sequence similarity before using them in the small variant assembly. Once arriving at the latter equation above, evaluating the likelihood ratio of any two hypotheses becomes a matter of determining $P(DNB|H_i)$ for each DNB and hypothesis.

A hypothesis H_i consists of the alleles $S_{i,k}$, and for each allele $S_{i,k}$ a corresponding allele fraction $f_{i,k}$. The allele fraction represents the fraction of haplotypes in the DNA sample containing the allele. For a typical diploid hypothesis there are two alleles, and the allele fraction is 0.5 for both alleles. Under the assumption that each DNB is equally likely to originate from any of the haplotypes in the sample, we can compute $P(DNB|H_i)$ as follows:

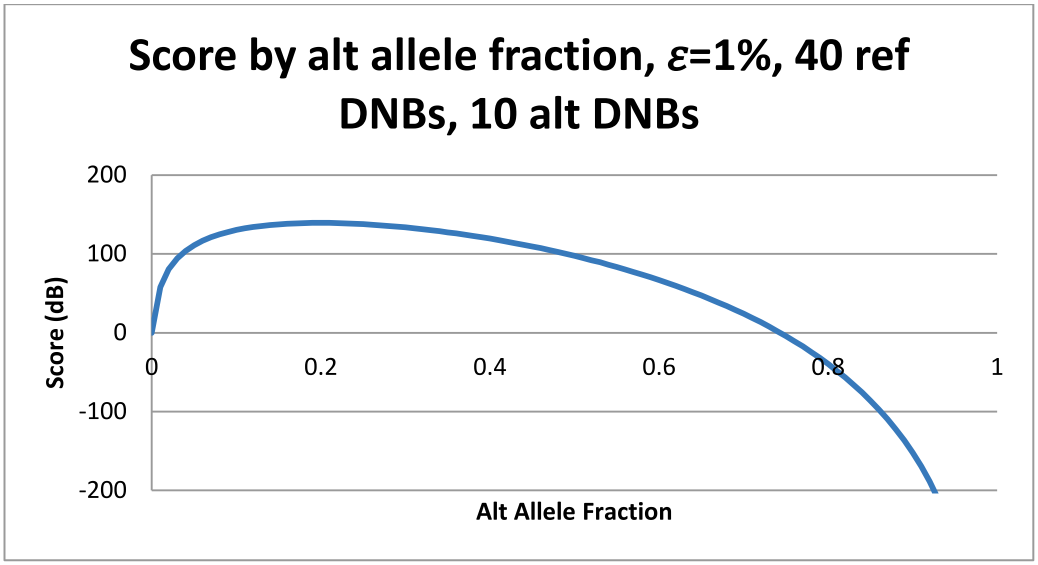
$$P(DNB|H_i) = \sum_k f_{i,k} P(DNB|\, S_{i,k}) \qquad (2)$$

Prior to the Analysis Pipeline version 2.0, hypotheses were constrained to be either diploid hypotheses where $f_{i,k} = 0.5$ or haploid hypotheses where $f_{i,k} = 1$. The small variant assembler in Analysis Pipeline version 2.0 attempts to solve the more general problem where there may be more than two alleles, and where the allele fraction for each allele is not constrained to be the same for each allele. This is particularly important for discovering somatic variants in cancer, as well as variants in non-diploid regions of a normal genome. For example, under the assumption that a cancer tissue is completely diploid, but the sample of that cancer is contaminated by normal tissue, there may still exist four distinct haplotypes with varying allele fraction. Additionally, the cancer tissue itself may be heterogeneous, composed of many different cells, each with its own heredity and somatic mutations.

Allele Fraction Example

We define the score to be the likelihood ratio in equation 1, such that H_j is the reference hypothesis, expressed in decibels. Under the model described above, we can determine the score for a heterozygous hypothesis for any allele fraction, under the assumption that $\varepsilon = 1\%$ for all base calls (this is slightly higher than the geometric mean of ε for a typical sequencing run), and in a scenario where we have 40 DNBs in favor of the reference and 10 DNBs in favor of an alternative SNP:

Figure 1: Score by ALT Allele Fraction, 10 DNBs in Favor of ALT SNP

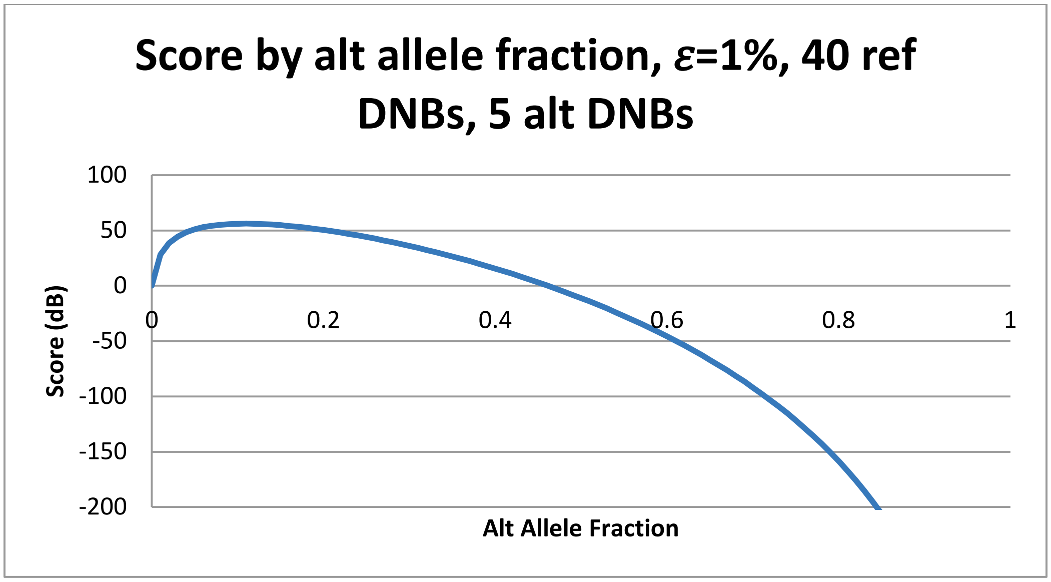


The plot above highlights a few things about the data and the model:

- A strong signal exists for the alt allele, as the maximum likelihood heterozygous allele fraction is ~140 dB more likely than homozygous ref.
- The alternate allele fraction cannot be "called" with any great certainty. For example, for any allele fraction between .04 and .5, the score is within 40 dB of the most likely allele fraction.

Figure 2 shows the same plot for a scenario where we have 40 DNBs in favor of the reference and 5 DNBs in favor of an alt SNP:

Figure 2: Score by Alternate Allele Fraction, 5 DNBs in Favor of Alternate SNP



This plot demonstrates the power of using a model that allows for a non-diploid allele fraction. Under the diploid assumption and either homozygous or heterozygous with 0.5 allele fraction, the homozygous hypothesis is more likely than heterozygous in this case. If the model allowed heterozygous with allele fractions that are not 0.5, it is shown that at allele fraction ~0.1, we would get a score that is substantially higher than the homozygous reference hypothesis. In practice, the small variant assembler in Complete Genomics Analysis Pipeline version 2.0 employs both models, and gives the score for the maximum likelihood allele fraction as *varScoreVAF* and the score for the equal allele fraction as *varScoreEAF*.

Appendix B: DNB Generation Modeling

In modeling the DNB generation process, we assume that each DNB randomly originates at a position of one of the two strands of the genome being sequenced, and from one of two alleles at that position, with all strands, positions, and alleles having equal probability to originate a DNB.

To finally evaluate hypothesis likelihood, we must estimate $\mathrm{P}(\mathrm{DNB}|\ \mathrm{S}_{\mathrm{i,k}})$ as in equation 2. A DNB may be generated at any position in the genome, with reads and the gaps between the reads. We call the position where a DNB arises and the gaps between the reads a "mapping". We can thus determine the likelihood of generating any given DNB by summing the likelihood of generating the DNBs over all possible mappings (M).

$$\mathrm{P}(\mathrm{DNB}|\ \mathrm{S}_{\mathrm{i,k}}) = \sum_{\mathrm{M}} \mathrm{P}(\mathrm{DNB},\mathrm{M}|\mathrm{S}_{\mathrm{i,k}})$$

$$\mathrm{P}(\mathrm{DNB}|\ \mathrm{S}_{\mathrm{i,k}}) = \sum_{\mathrm{M}} \mathrm{P}(\mathrm{DNB}|\mathrm{M},\mathrm{S}_{\mathrm{i,k}})\ \mathrm{P}(\mathrm{M}|\mathrm{S}_{\mathrm{i,k}}) \qquad (3)$$

Here, we assume the likelihood of a DNB arising is the same at any position within the genome, and the change in DNB likelihood due to change in length of the genome for any two hypotheses is roughly equal. In that case, the likelihood ratio in equation 1 is unaffected by the likelihood of DNB arising at any given position, and in the factor $\mathrm{P}(\mathrm{M}|\mathrm{S}_{\mathrm{i,k}})$ we must simply account for the likelihood of the gaps. The gaps likelihoods are determined empirically during the mapping stage of assembly. The gaps within each arm are modeled as dependent on the sequence near the enzyme cut sites, and the small gaps of the left arm, the mate gap, and the small gaps of the right arm are modeled as independent of each other.

$$\mathrm{P}\big(\mathrm{M}\big|\mathrm{S}_{\mathrm{i,k}}\big) = \mathrm{P}(\mathrm{g_L g_M g_R}|\mathrm{S}_{\mathrm{i,k}})$$

$$\mathrm{P}\big(\mathrm{M}\big|\mathrm{S}_{\mathrm{i,k}}\big) = \mathrm{P}(\mathrm{g_L}|\mathrm{S}_{\mathrm{i,k}})\mathrm{P}(\mathrm{g_M}|\mathrm{S}_{\mathrm{i,k}})\mathrm{P}(\mathrm{g_R}|\mathrm{S}_{\mathrm{i,k}})$$

The remaining factor to resolve in equation 3 is $\mathrm{P}(\mathrm{DNB}|\mathrm{M},\mathrm{S}_{\mathrm{i,k}})$. We assume each base call is independent, so that $\mathrm{P}(\mathrm{DNB}|\mathrm{M},\mathrm{S}_{\mathrm{i,k}})$ is simply the product of base call likelihoods *b* in DNB:

$$\mathrm{P}(\mathrm{DNB}|\mathrm{M},\mathrm{S}_{\mathrm{i,k}}) = \prod_{\mathrm{b\ in\ DNB}} \mathrm{P}(\mathrm{b}|\mathrm{M},\mathrm{S}_{\mathrm{i,k}})$$

Where the base call matches the hypothesis allele for the mapping, $\mathrm{P}\big(\mathrm{b}\big|\mathrm{M},\mathrm{S}_{\mathrm{i,k}}\big)$ is the likelihood the base call is correct. We determine the likelihood a base call is correct (ε) empirically during the mapping stage for correct base calls; it is dependent on the base call score, read cycle, and field within the slide. For bases that do not match the hypothesis, the Analysis Pipeline assumes the likelihood of any given base call is $\varepsilon/3$.

The DNA sequence of a DNB may be modified during the library process. For example, a SNP or indel may be introduced into the DNB. This process of empirically estimating base call likelihood also accounts for SNPs within the sequence of the DNB, except where the SNP occurs at the same position as an overlapping base call (i.e., negative wobble gap). In the case that a SNP is introduced, we are likely to get two correct base calls, each discordant with the correct value of the hypothesis. If we say that θ is the likelihood a SNP is introduced at any given position within the DNB, then the likelihood of two base calls that are concordant with each other, but discordant with each other, turns out to be roughly $\theta/3$, under simple assumptions about the likelihood of any given transition or transversion. In Analysis Pipeline version 2.0, the Complete Genomics assembler models this possibility, with $\theta = .0015$.

The small variant assembler also models the likelihood of an indel being introduced in the DNB, but this model refinement is only employed during hypothesis rescoring stage, as described in "Step 5: Hypothesis Rescoring," due to practical considerations.

Suppose in our example where the DNB maps with no discordances to the top hypothesis but has no mappings to the reference hypothesis, the DNB supports a heterozygous insertion. It is possible this DNB actually originated from the reference sequence, but the base was inserted during PCR or some other process before sequencing that is not otherwise described by our DNB generation model.

To account for this possibility, we assume the likelihood of any given variant arising within a DNB during the DNB construction process is β, so that $P(DNB|reference) \geq \beta\, P(DNB|\ variant)$. The hypothesis rescoring stage uses this knowledge to ensure that no DNB provides overwhelming support for a hypothesis. It does so by increasing $P(DNB|\ S)$ for every allele S to ensure that, for each pair of alleles S_i and S_j within this universe, $P(DNB|\ S_i) \geq \beta\, P(DNB|\ S_j)$. When S_i and S_j differ by only a one-base indel, $\beta = .001$ is used. Otherwise, $\beta = .0001$ is used.

In the model described above, we must technically sum the likelihood $P(DNB, M|S_{i,k})$ for all possible mappings. As there are billions of such mappings for every DNB, in practice we only sum the likelihoods for all "good" mappings (the mappings with few discordant base calls with respect to the hypothesis), and add a term, α, representing the likelihood of generating the DNB from a "bad" mapping. The α term is set to 10^{-9}, and serves as a catch-all factor to decrease the signal of "bad" DNBs—those which arise by means which are not well modeled by the Bayesian math.