Small Variant Score Calibration Methods

October 2012

RM_CAL-02

Table of Contents

Overview

The Complete Genomics Analysis Pipeline identifies most types of genomic alterations in a sequenced genome, including small variants: SNPs, insertions, deletions, and substitutions. To facilitate downstream filtering and interpretation, the analysis process richly annotates variants with biological annotation and scores. Although the scores provide confidence in the calls made and are the best indicator of variant quality, they are currently not calibrated to absolute error rate. Complete Genomics has made available a series of calibration files generated from replicate library datasets to enable researchers to translate provided scores into false positive and false negative rates.

These calibration files can be used in two different contexts: they are required for somatic and calibrated variant call output from the CGA Tools calldiff and mkvcf commands (as described in the *CGA Tools User Guide*) and they may also be used outside of CGA Tools.

This document describes how calibration files are used outside of CGA Tools to calibrate variant scores provided for each genome. In addition, it describes the methods used to generate the calibration files, their content, and file format.

Obtaining the Calibration Files

The calibration files described in this document can be downloaded from this location:

ftp://ftp.completegenomics.com/ScoreCalibrationFiles/var-calibration-v2.tgz

Prerequisite Reading

This document assumes that readers are already familiar with variant score definition and the principals on which they are based. In addition, we recommend that you are familiar with the following Complete Genomics documentation:

- *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]
- *Release Notes* — The Analysis Pipeline Release Notes indicate new features and enhancements by release. [www.completegenomics.com/documents/ReleaseNotes-100358389.html]
- "Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads" — An article describing the original Complete Genomics computational methods for small variant detection. These methods have evolved over the development of further Analysis Pipeline versions. (Journal of Computational Biology, Volume: 19 Issue 3: March 8, 2012) This document is available on the Liebert web site:

  http://online.liebertpub.com/doi/full/10.1089/cmb.2011.0201

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

www.completegenomics.com/customer-support/support/

Calibration Workflow

The replicate calibration workflow performed by Complete Genomics to generate the calibration files includes the following three steps:

1. **Assemble genomes** — Map, assemble, and perform variant calling for a series of replicate libraries, including:
   - Four NA19240 replicate libraries of various coverage depths. These libraries are used in the iterative refinement analysis to calibrate scores under the assumption of 50% allele fraction.
   - Two NA12878 replicate libraries (high and low coverage) and two mixed libraries (high and low coverage) with 60% NA12878 and 40% NA19240. These libraries are used to produce score distributions for the purposes of calibrating for 20% allele fraction.
2. **Compare genomes** — Run CGA™ Tools calldiff to determine loci of concordance and discordance among replicates.
3. **Iteratively refine score calibration** — Calculate the likelihood each discordant site is a false positive or false negative and construct score calibration given the set of true and false calls.

The remainder of this document focuses on step 3. The methods and approaches of genome assembly (step 1) and genome comparison using CGA™ Tools calldiff (step 2) are described in the Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads publication and the *CGA Tools User Guide*, respectively.

Calibrated Score Definition

The replicate calibration data provided and described in this document gives a sense for the empirically observed quality of variant calls and reference calls, given the score and local coverage. A calibrated score for the purposes of this calibration is defined as follows:

$$-10 \log_{10} \frac{\text{P(False call)}}{\text{P(True call)}}$$

Thus, given a calibrated score, we can determine the likelihood the call is correct. This score definition is consistent with Complete Genomics uncalibrated scoring for small variants, and is approximately equal to the Phred score, defined as $-10 \log_{10}$ P(False call), for high scoring calls. It can differ substantially from the Phred score for low scoring calls. Table 1 gives a sense for how the calibrated score corresponds to P(True call):

Table 1: Correlation between Calibrated Scores and Probabilities of a True Call

Score	P(True call)
-10	0.09091
0	0.50000
10	0.90909
20	0.99010
30	0.99900
40	0.99990
50	0.99999

Calibration Files

The results of replicate calibration described in this document are presented as a set of files. Each file provides the calibrated score as a function of two metrics: (1) the uncalibrated score and (2) the coverage.

Using *varScoreVAF*, *varScoreEAF*, or *totalScore* (if data was generated from Analysis Pipeline versions prior to 2.0) with *totalReadCount* allows calibration of scores to a false positive, undercall, or overcall rate. Using *refScore* with *uniqueSequenceCoverage* allows calibration to a false negative rate.

The calibration files are provided in two sub-directories:

- `version2.0.0` for calibration of scores from Analysis Pipeline version 2.0 and greater
- `version0.0.0` for calibration of scores from Analysis Pipeline versions prior to 2.0

> **Note**
>
> The version0.0.0 calibration files are generated using data from Analysis Pipeline 1.12.0, where *totalScore* column was used as both the *varScoreVAF* and *varScoreEAF* for the purposes of calibration. The small variant assembler has not changed substantially since Analysis Pipeline version 1.5.0, and it is believed the 1.12.0 calibration is useful for all pipeline versions between 1.5 and 1.12.0.

Within each calibration data file, each column represents a coverage bin and each row gives the calibration for a different score. As shown in Figure 1, each column header lists the minimum coverage value for the coverage bin. For example, if the file includes columns score, cvg0, cvg20 and cvg30, then the cvg0 column refers to data where the coverage level is between 0 and 19, the cvg20 column refers to data where the coverage level is between 20 and 29, and the cvg30 column refers to data where the coverage level is 30 or higher.

The calibration is differentiated by the following attributes, which are indicated in the file name:

- The variant type: snp, ins, del, or sub
- The likelihood model: VAF (variable allele fraction) or EAF (equal allele fraction)
- The error mode: fp (false positive), fn (false negative), uc (undercall), or oc (overcall)
- The allele fraction: all calibrations are based on a diploid assumption (50% allele fraction), unless indicated as “af20” (20% allele fraction)

For example, calibration files describing the deletions (del) for variable allele fraction (vaf) in the error mode for false positives (fp) will be named “del-vaf-fp.tsv” or “del-vaf-fp-af20.tsv”.

It is important to keep in mind that unless files are marked for 20% allele fraction (“af20”), the replicate calibration makes the assumption that the genome is diploid and that all heterozygous variants called have 50% allele fraction. Given that a variant present at a lower allele fraction is likely to have a lower variation score than a variant present at 50% allele fraction, and if a substantial number of true variants are expected to be present at low allele fraction, the calibrated false positive score given in files not named ***-af20.tsv*** may indicate a lower call confidence than is warranted. Thus, to calibrate scores for variants present at a lower allele fraction than 50%, you should use calibration files named ***-af20.tsv***. For concrete plots of calibrations of variants at 50% allele fraction and 20% allele fraction, see “20% Allele Fraction Calibration” in the Validation of the Calibration section.

To calibrate scores, choose a calibration file based on the calibration attributes given in the subdirectory and the file name. Table 2 and Table 3 list the calibration files provided in the version2.0.0 and version0.0.0 subdirectories and summarize their attributes. Additionally, an example of a calibration file, ***snp-vaf-fp.tsv***, which can be used to calibrate *varScoreVAF* for SNPs based on false positive rates assuming 50% allele fraction, is shown in Figure 1.

Note

The replicate calibration approach is not well suited to account for systematic errors in variant calling that occur for every genome. As such, the replicate calibration is most useful in the context of a comparison between one or more genomes sequenced by Complete Genomics. For example, if genome A has a variant where genome B has a reference call, the calibrated scores give a sense for whether the difference between the two genomes can be trusted.

As an example, CGA™ Tools calldiff version 1.5 and Analysis Pipeline version 2.0 use the calibrated scores to compute the somatic score for the reported somatic variants. Because replicate analysis does not account for systematic errors, the actual error rates could be much higher than the error rates indicated by this calibration.

Table 2: Analysis Pipeline Versions 2.0 and Greater Calibration Files (version2.0.0 directory)

File Name	Variant Type to be Calibrated	Likeli-hood Model	Uncalibrated Scores to Use	Coverage to Use	Calibrate Scores To
del-eaf-fn.tsv	deletion	EAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
del-eaf-fp.tsv	deletion	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
del-eaf-uc.tsv	deletion	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
del-eaf-oc.tsv	deletion	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
del-vaf-fn.tsv	deletion	VAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
del-vaf-fp.tsv	deletion	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
del-vaf-uc.tsv	deletion	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
del-vaf-oc.tsv	deletion	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
del-vaf-fp-af20.tsv	deletion	VAF	varScoreVAF	*totalReadCount* (***masterVarBeta*** file)	false positive rate
ins-eaf-fn.tsv	insertion	EAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
ins-eaf-fp.tsv	insertion	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
ins-eaf-uc.tsv	insertion	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
ins-eaf-oc.tsv	insertion	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
ins-vaf-fn.tsv	insertion	VAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate

File Name	Variant Type to be Calibrated	Likeli-hood Model	Uncalibrated Scores to Use	Coverage to Use	Calibrate Scores To
ins-vaf-fp.tsv	insertion	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
ins-vaf-uc.tsv	insertion	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
ins-vaf-oc.tsv	insertion	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
ins-vaf-fp-af20.tsv	insertion	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
snp-eaf-fn.tsv	single nucleotide polymorphism	EAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
snp-eaf-fp.tsv	single nucleotide polymorphism	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
snp-eaf-uc.tsv	single nucleotide polymorphism	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
snp-eaf-oc.tsv	single nucleotide polymorphism	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
snp-vaf-fn.tsv	single nucleotide polymorphism	VAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
snp-vaf-fp.tsv	single nucleotide polymorphism	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
snp-vaf-uc.tsv	single nucleotide polymorphism	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
snp-vaf-oc.tsv	single nucleotide polymorphism	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
snp-vaf-fp-af20.tsv	single nucleotide polymorphism	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
sub-eaf-fn.tsv	substitution	EAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
sub-eaf-fp.tsv	substitution	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
sub-eaf-uc.tsv	substitution	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
sub-eaf-oc.tsv	substitution	EAF	*varScoreEAF*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
sub-vaf-fn.tsv	substitution	VAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
sub-vaf-fp.tsv	substitution	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
sub-vaf-uc.tsv	substitution	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	undercall rate

File Name	Variant Type to be Calibrated	Likeli-hood Model	Uncalibrated Scores to Use	Coverage to Use	Calibrate Scores To
sub-vaf-oc.tsv	substitution	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
sub-vaf-fp-af20.tsv	substitution	VAF	*varScoreVAF*	*totalReadCount* (***masterVarBeta*** file)	false positive rate

Table 3: Analysis Pipeline Versions 1.5 – 1.12 Calibration Files (version0.0.0 directory)

File Name	Variant Type to be Calibrated	Likeli-hood Model	Uncalibrated Scores to Use	Coverage to Use	Calibrate Scores To
del-eaf-fn.tsv	deletion	EAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
del-eaf-fp.tsv	deletion	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
del-eaf-uc.tsv	deletion	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
del-eaf-oc.tsv	deletion	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
del-vaf-fn.tsv	deletion	VAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
del-vaf-fp.tsv	deletion	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
del-vaf-uc.tsv	deletion	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
del-vaf-oc.tsv	deletion	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
del-vaf-fp-af20.tsv	deletion	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
ins-eaf-fn.tsv	insertion	EAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
ins-eaf-fp.tsv	insertion	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
ins-eaf-uc.tsv	insertion	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
ins-eaf-oc.tsv	insertion	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
ins-vaf-fn.tsv	insertion	VAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
ins-vaf-fp.tsv	insertion	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
ins-vaf-uc.tsv	insertion	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
ins-vaf-oc.tsv	insertion	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
ins-vaf-fp-af20.tsv	insertion	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
snp-eaf-fn.tsv	single nucleotide polymorphism	EAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate

File Name	Variant Type to be Calibrated	Likeli-hood Model	Uncalibrated Scores to Use	Coverage to Use	Calibrate Scores To
snp-eaf-fp.tsv	single nucleotide polymorphism	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
snp-eaf-uc.tsv	single nucleotide polymorphism	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
snp-eaf-oc.tsv	single nucleotide polymorphism	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
snp-vaf-fn.tsv	single nucleotide polymorphism	VAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
snp-vaf-fp.tsv	single nucleotide polymorphism	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
snp-vaf-uc.tsv	single nucleotide polymorphism	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
snp-vaf-oc.tsv	single nucleotide polymorphism	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
snp-vaf-fp-af20.tsv	single nucleotide polymorphism	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
sub-eaf-fn.tsv	substitution	EAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
sub-eaf-fp.tsv	substitution	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
sub-eaf-uc.tsv	substitution	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
sub-eaf-oc.tsv	substitution	EAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
sub-vaf-fn.tsv	substitution	VAF	*refScore*	*uniqueSequenceCoverage* (***coverageRefScore*** file)	false negative rate
sub-vaf-fp.tsv	substitution	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate
sub-vaf-uc.tsv	substitution	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	undercall rate
sub-vaf-oc.tsv	substitution	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	overcall rate
sub-vaf-fp-af20.tsv	substitution	VAF	*totalScore*	*totalReadCount* (***masterVarBeta*** file)	false positive rate

Example *snp-vaf-fp.tsv*

Figure 1: Example Calibration File snp-var-fp.tsv (not all rows are shown)

>score	cvg0	cvg20	cvg30	cvg40	cvg50	cvg60	cvg70	cvg80	cvg90	cvg110
…	…	…	…	…	…	…	…	…	…	…
400	47.1	53.4	54.8	51	43	38.7	32.6	27.6	22.8	15.7
401	47.1	53.4	54.8	51	43.1	38.8	32.7	27.7	22.9	15.8
402	47.1	53.4	54.8	51	43.1	38.8	32.8	27.7	23	15.8
403	47.1	53.4	54.8	51	43.2	38.9	32.8	27.8	23	15.9
404	47.1	53.4	54.8	51	43.2	39	32.9	27.9	23.1	15.9
405	47.1	53.4	54.8	51	43.3	39.1	33	27.9	23.1	16
406	47.1	53.4	54.8	51	43.4	39.1	33	28	23.2	16
407	47.1	53.4	54.8	51	43.4	39.2	33.1	28	23.2	16.1
408	47.1	53.4	54.8	51	43.5	39.3	33.2	28.1	23.3	16.1
409	47.1	53.4	54.8	51	43.6	39.4	33.2	28.2	23.4	16.2
410	47.1	53.4	54.8	51	43.6	39.4	33.3	28.2	23.4	16.2
411	47.1	53.4	54.8	51	43.7	39.5	33.4	28.3	23.5	16.3
…	…	…	…	…	…	…	…	…	…	…
1000	47.1	53.4	54.8	51	48.9	51.1	48.6	48	50.6	44

Header Description *snp-vaf-fp.tsv*

Key	Description	Allowed Values
#SOFTWARE_VERSION	Analysis pipeline version.	Two or more digits separated by periods. Either “2.0.0” for Analysis Pipelines 2.0 and greater, or “0.0.0” for Analysis Pipeline versions prior to 2.0.
#CALIB_FORMAT_VERSION	Version number of the file format.	One or more digits separated by periods. For example, “1” or “1.5”
#CALIB_VERSION	Calibration version number.	One or more digits separated by periods.
#GENERATED_AT	Date and time of calibration file generation.	Month-Day-Year- Time. For example “9/16/2011 3:32:50 PM”.
#VARTYPE	Class of variant characterized by the calibration data.	“snp”: single nucleotide polymorphism “del”: deletion “ins”: insertion “sub”: substitution
#LIKELIHOOD_MODEL	Model used for variant score calculation.	“vaf”: variable allele fraction “eaf”: equal allele fraction
#ERROR_MODE	Type of error described by calibrated scores.	“fp”: false positive “fn”: false negative “oc”: overcall “uc”: undercall
#ALLELE_FRACTION	Assumed allele fraction used for score calculation.	“0.5”: 50% allele fraction “af20”: 20% allele fraction
#TYPE	Type of data contained in the file.	“SCORE_CALIBRATION”: calibrated score for each uncalibrated score and coverage level.

Content Description *snp-vaf-fp.tsv*

All calibration files include content similar to the following example, except for these differences:

- The *Score* may be based on *varScoreEAF*, *totalScore*, or *refScore*
- If *refScore* is being calibrated, Coverage is based on *uniqueSequenceCoverage*

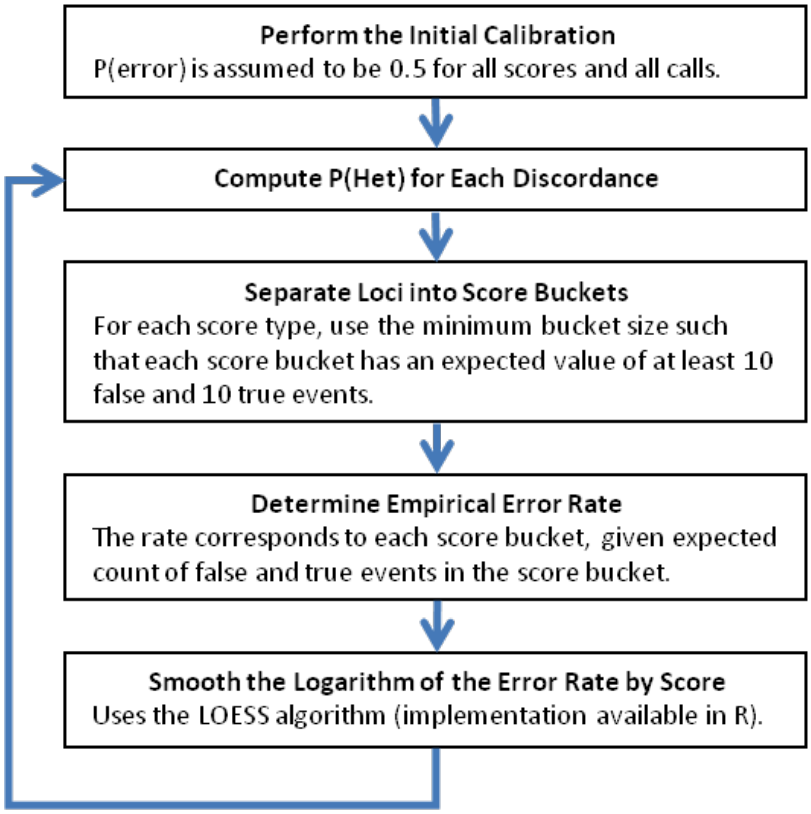
	Column	Description
1	Score	*varScoreVAF* for the variant to be calibrated.
2	cvg0	Minimum coverage value for the coverage bin between 0 and 19. This bin describes the coverage level (*totalReadCount* from the ***masterVarBeta*** file) for the variant to be calibrated.
3	cvg20	Minimum coverage value for the coverage bin between 20 and 29. This bin describes the coverage level (*totalReadCount* from the ***masterVarBeta*** file) for the variant to be calibrated.
4	cvg30	Minimum coverage value for the coverage bin between 30 and 39. This bin describes the coverage level (*totalReadCount* from the ***masterVarBeta*** file) for the variant to be calibrated.
5	cvg40	Minimum coverage value for the coverage bin between 40 and 49. This bin describes the coverage level (*totalReadCount* from the ***masterVarBeta*** file) for the variant to be calibrated.
…	…	…
11	cvg110	Minimum coverage value for the coverage bin ≥ 110. This bin describes the coverage level (*totalReadCount* from the ***masterVarBeta*** file) for the variant to be calibrated.

Score Calibration

Iterative Refinement of Score Calibration

The replicate calibration analysis tests the likelihood that a replicate discordance results from a calling error (false positive, false negative, overcall, or undercall), given calibrated scores for the calls in both replicate genomes. Concordant loci are assumed to be true. The calibration process is an iterative analysis represented by a feedback loop that starts with initial estimates for calibrated scores. It determines the likelihood each discordant site is an error based on those calibration estimates and then constructs improved estimates of the score calibration, given the set of true and false calls. The process iterates through this loop three times and writes results to the calibration files. Figure 2 describes this feedback loop.

Figure 2: Iterative Calibration by Replicates Method



The score calibration is run separately for each variant type (snp, ins, del, or sub), and separately for each likelihood model (VAF or EAF). For each combination of variant type and likelihood model, the method is run once to calibrate the false positive rate (FP) given a variation score(*varScoreVAF*, *varScoreEAF*, or *totalScore*) and false negative rate (FN) given a *refScore*. These results are referred to as FP-FN calibration. The method is run again to calibrate the undercall rate (UC) given the *varScore* of the reference call in a ref-het locus and the overcall rate (OC) given the minimum *varScore* of a homozygous alt locus. These results are referred to as UC-OC calibration.

To calibrate the scores for a particular variant type or score type, we first categorize the loci in a replicate comparison as “homozygous concordant,” “heterozygous concordant,” or “discordant.”

When performing the FP-FN calibration, the homozygous concordant sites are shared reference calls in the replicates; the heterozygous concordant sites are shared ref-het calls; the discordant sites are the sites where one genome has a ref-het call but the other genome has a homozygous ref call; and all other loci are discarded. When performing the UC-OC calibration, the homozygous concordant sites are shared homozygous alt calls in the replicates; the heterozygous concordant sites are shared ref-het calls; the discordant sites are the sites where one genome has a ref-het call but the other genome has a homozygous alt call and where the alt call is shared; and all other loci are discarded.

Considering a locus within the replicate analysis, we define:

Symbol	Description
hetScoreA	The condition that the locus is called heterozygous with a given score in genome A.
homScoreB	The condition that the locus is called homozygous with a given score in genome B.
d	The condition that the locus is a replicate discordant locus.
c_r	The condition that the locus is called homozygous in both genomes.
c_h	The condition that the locus is called heterozygous in both genomes.
n_{TP}	The number of true heterozygous loci in the genome.
n_{FP}	The number of false called heterozygous loci.
n_{TN}	The number of true homozygous loci.
n_{FN}	The number of false called homozygous loci.
HetA	The condition that genome A is truly heterozygous at the given locus.
HomA	The condition that genome A is truly homozygous at the given locus.
HetB	The condition that genome B is truly heterozygous at the given locus.
HomI	The condition that genome B is truly homozygous at the given locus.
Het	The condition that both genomes are heterozygous at this locus.
Hom	The condition that both genomes are homozygous at this locus.

We note that under the assumption that the two genomes are the same genome (replicates):

$$P(\text{not HetA}) = P(\text{HomA}) = P(\text{HomB}) = P(\text{not Het}) = P(\text{Hom})$$

We wish to determine the likelihood ratio

$$\frac{P(\text{Het}|\text{hetScoreA}, \text{homScoreB}, \text{d})}{P(\text{Hom}|\text{hetScoreA}, \text{homScoreB}, \text{d})}$$

or equivalently,

$$\frac{P(\text{HetA}|\text{hetScoreA}, \text{homScoreB}, \text{d})}{P(\text{HomB}|\text{hetScoreA}, \text{homScoreB}, \text{d})}$$

given the calibration of

hetScoreA ($L_A = \frac{P(\text{HomA}|\text{hetScoreA},c_h)}{P(\text{HetA}|\text{hetScoreA},c_h)}$) and homScoreB ($L_B = \frac{P(\text{HetB}|\text{homScoreB},c_r)}{P(\text{HomB}|\text{homScoreB},c_r)}$).

We begin by applying Bayes' theorem as follows:

$$\frac{P(\text{HetA}|\text{hetScoreA}, \text{homScoreB}, \text{d})}{P(\text{HomB}|\text{hetScoreA}, \text{homScoreB}, \text{d})} = \frac{P(\text{hetScoreA}, \text{homScoreB}|\text{HetA}, \text{d})P(\text{HetA}|\text{d})}{P(\text{hetScoreA}, \text{homScoreB}|\text{HomB}, \text{d})P(\text{HomB}|\text{d})}$$

Under independence assumption we have:

$$= \frac{P(\text{hetScoreA}|\text{HetA}, d)P(\text{homScoreB}|\text{HetA}, d)P(\text{HetA}, d)}{P(\text{hetScoreA}|\text{HomB}, d)P(\text{homScoreB}|\text{HomB}, d)P(\text{HomB}, d)}$$

$$= \frac{P(\text{hetScoreA}|\text{HetA}, d)P(\text{homScoreB}|\text{HetA}, d)n_{FN}}{P(\text{hetScoreA}|\text{HomB}, d)P(\text{homScoreB}|\text{HomB}, d)n_{FP}} \qquad (1)$$

Again using Bayes' theorem we have:

$$L_A = \frac{P(\text{HomB}|\text{hetScoreA}, c_h)}{P(\text{HetA}|\text{hetScoreA}, c_h)} = \frac{P(\text{hetScoreA}|\text{HomB}, c_h)P(\text{HomB}, c_h)}{P(\text{hetScoreA}|\text{HetA}, c_h)P(\text{HetA}, c_h)}$$

$$\frac{P(\text{hetScoreA}|\text{HomB}, c_h)}{P(\text{hetScoreA}|\text{HetA}, c_h)} = L_A \frac{P(\text{HetA}, c_h)}{P(\text{HomB}, c_h)} = L_A \frac{n_{TP}}{n_{FP}} \qquad (2)$$

$$L_B = \frac{P(\text{HetA}|\text{homScoreB}, c_r)}{P(\text{HomB}|\text{homScoreB}, c_r)} = \frac{P(\text{homScoreB}|\text{HetA}, c_r)P(\text{HetA}, c_r)}{P(\text{homScoreB}|\text{HomB}, c_r)P(\text{HomB}, c_r)}$$

$$\frac{P(\text{homScoreB}|\text{HetA}, c_r)}{P(\text{homScoreB}|\text{HomB}, c_r)} = L_B \frac{P(\text{HomB}, c_r)}{P(\text{HetA}, c_r)} = L_B \frac{n_{TN}}{n_{FN}} \qquad (3)$$

We now assume the score distribution for true variants is the same for discordant loci as for concordant loci, so that $P(\text{hetScoreA}|\text{HetA}, d) = P(\text{hetScoreA}|\text{HetA}, c_h)$, and likewise for the other score distributions of equations (1), (2), and (3). We can then plug equations (2) and (3) into (1), and we get:

$$\frac{P(\text{HetA}|\text{hetScoreA}, \text{homScoreB}, d)}{P(\text{HomB}|\text{hetScoreA}, \text{homScoreB}, d)} = \frac{n_{FP} L_B n_{TN} n_{FN}}{L_A n_{TP} n_{FN} n_{FP}}$$

$$\frac{P(\text{Het}|\text{hetScoreA}, \text{homScoreB}, d)}{P(\text{Hom}|\text{hetScoreA}, \text{homScoreB}, d)} = \frac{L_B n_{TN}}{L_A n_{TP}}$$

We are given L_A and L_B as inputs to the iteration, and we can estimate n_{TN} and n_{TP} empirically.

In practice, the likelihood of a false call for a given score is strongly dependent on local coverage. To model this behavior, we separate loci into bins by both coverage and score, and smooth the logarithm of the error rate by score for each coverage bin. The calibration is then reported for each coverage bin.

20% Allele Fraction Calibration

The iterative calibration process described above is performed with the assumption that the genome is diploid and all heterozygous variants have 50% allele fraction. However, this assumption is violated in several cases, such as in most tumor samples where heterogeneity and copy number aberrations are common, and non-tumor samples with regions of copy number variation and/or mosaicism. In these cases, heterozygous variants may not have 50% allele fraction.

To enable accurate calibration of scores in these non-diploid genomes or regions, a 20% allele fraction calibration is also provided for false positives using the *varScoreVAF*. This calibration indicates the error rate of variant calls, under the assumption that all heterozygous mutations in a genome are present at 20% allele fraction. We adopt the notation described in "Iterative Refinement of Score Calibration", except we define $\text{Het}_{20\%AF}$ to be the condition that a given

locus is heterozygous, assuming all heterozygous loci are present at 20% allele fraction and $\text{Het}_{50\%\text{AF}}$ to be the condition that a locus is heterozygous, assuming all heterozygous loci are present at 50% allele fraction. Previously, we used Het (without any other notation) to indicate the latter condition.

We can apply Bayes' theorem to the likelihood ratio:

$$\frac{\text{P}(\text{Het}_{20\%\text{AF}}|\text{hetScoreA})}{\text{P}(\text{Het}_{50\%\text{AF}}|\text{hetScoreA})} = \frac{\text{P}(\text{hetScoreA}|\text{Het}_{20\%\text{AF}})\text{P}(\text{Het}_{20\%\text{AF}})}{\text{P}(\text{hetScoreA}|\text{Het}_{50\%\text{AF}})\text{P}(\text{Het}_{50\%\text{AF}})}$$

Given that we are considering a scenario where the set of heterozygous calls is the same as you would expect in a typical genome — they are simply present at 20%AF — we have:

$$\text{P}(\text{Het}_{20\%\text{AF}}) = \text{P}(\text{Het}_{50\%\text{AF}})$$

Therefore:

$$\text{P}(\text{Het}_{20\%\text{AF}}|\text{hetScoreA}) = \text{P}(\text{Het}_{50\%\text{AF}}|\text{hetScoreA})\frac{\text{P}(\text{hetScoreA}|\text{Het}_{20\%\text{AF}})}{\text{P}(\text{hetScoreA}|\text{Het}_{50\%\text{AF}})}$$

That is, we can simply scale $\text{P}(\text{Het}_{50\%\text{AF}}|\text{hetScoreA})$ by the score distribution ratio for true variants to get $\text{P}(\text{Het}_{20\%\text{AF}}|\text{hetScoreA})$.

During score calibration, the score distribution of true variants is assessed for both allele fractions using the sample mixture assemblies. The true variants for the 20%AF and 50%AF case are binned and their ratio is smoothed in the same manner as false and true variants are binned and their ratio is smoothed for calibration of 50%AF variants. This smoothed result is then applied to the 50%AF calibration to obtain the 20%AF calibration. Because of the use of this mixture model, the 20% allele fraction calibration files can be used to calibrate scores for variants present at allele fractions lower than 50%.

Validation of the Calibration Method

This section presents data to show that the calibration method produce high-quality results.

Convergence of Iterative Refinement Process

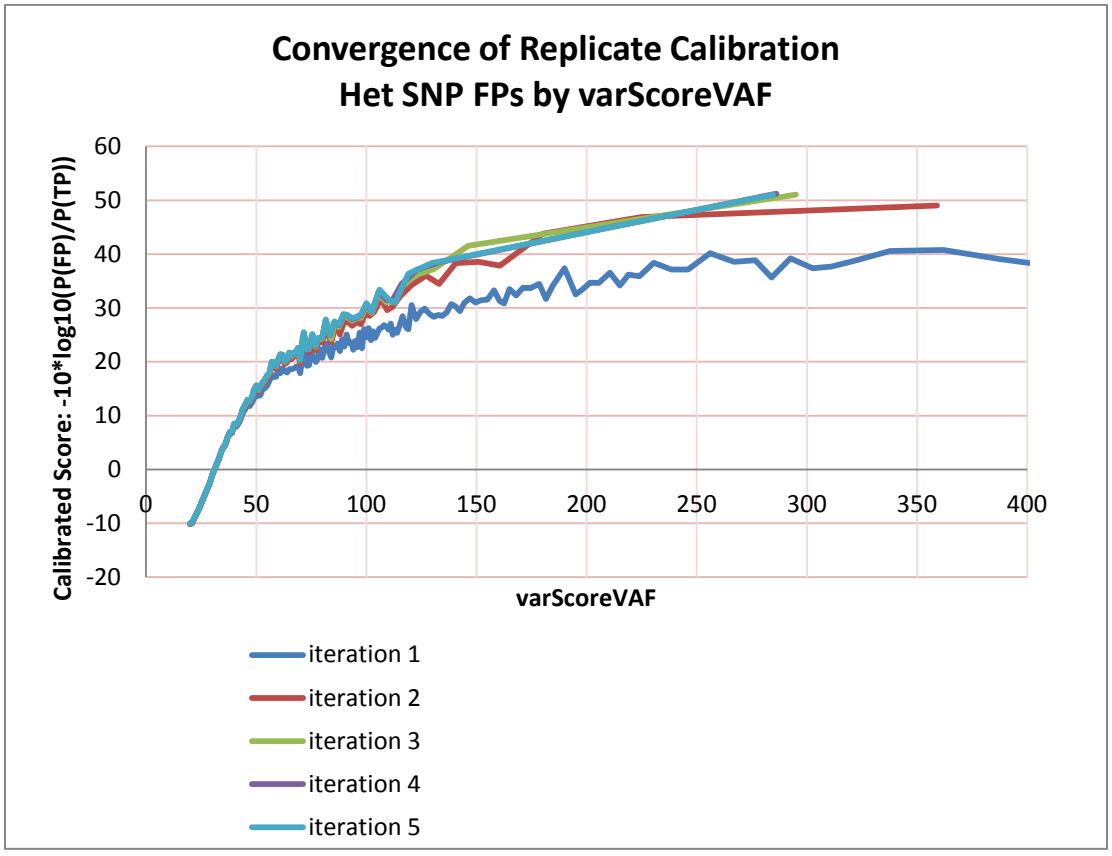
The algorithm described previously in "Iterative Refinement of Score Calibration" rapidly converges to a solution. Figure 3 shows an example of the pre-smoothed convergence for the case of a coverage bin that includes all coverage levels.

For the lowest *varScoreVAF* displayed on the plot, the replicate calibration reports a calibration score of -10. This calibrated score indicates that false calls are ten times as likely as true calls. This is not unexpected for *varScoreVAF* value of 20, since heterozygous SNPs occur approximately every 1000 bases and this is not accounted for in the assembler's probability model.

The best variation scores achieve a calibrated score over 50, indicating there is one error every 100,000 true SNPs at this score, or about one error every 100 million base positions (assuming a true SNP with this good a score every 1 kb).

The deviation of the calibration curve from the expected linear shape indicates that our modeling of DNA generation is imperfect and our variant caller does not exactly predict the likelihood any event is true. Instead, the downward curvature of the calibrated score line indicates that there are some DNBs that occur outside the model of DNB generation and so the shape of the curve can be considered a manifestation of systematic artifacts not explained by the model.

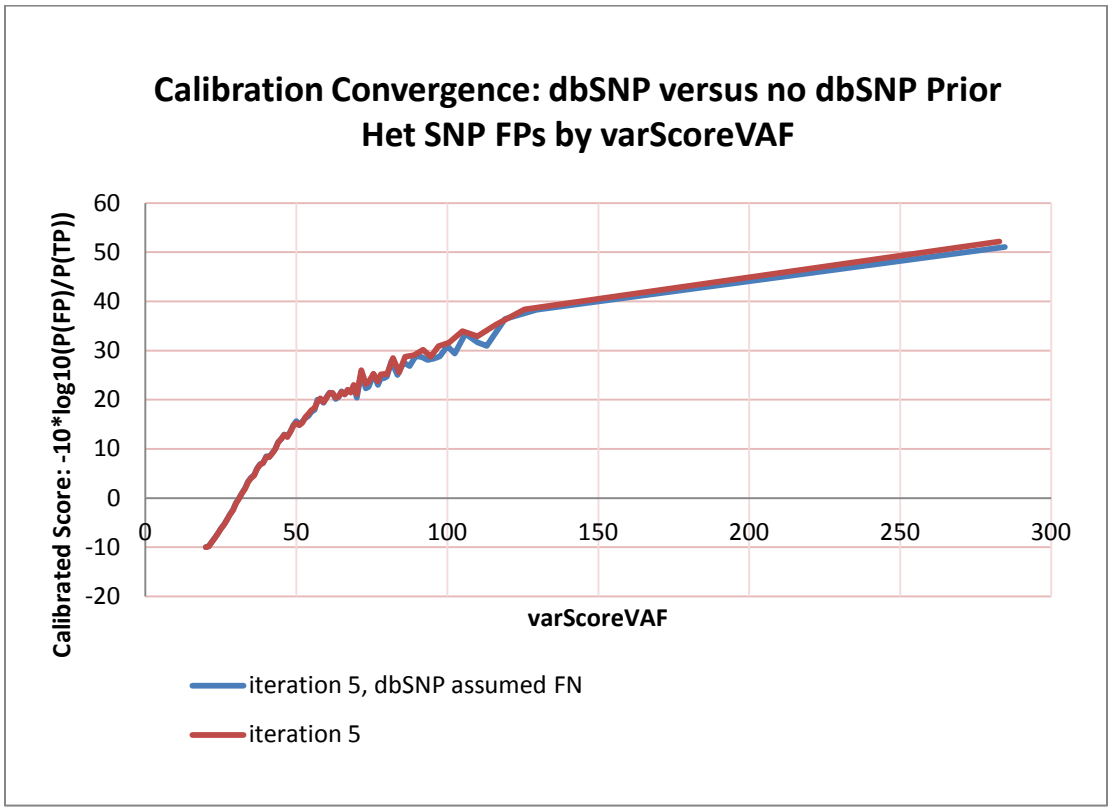
Figure 3: Pre-Smoothed Convergence for All Coverage Levels



Convergence Using dbSNP Variants as Priors

Figure 4 shows that if we modify the step to determine P(Het) for the FP-FN calibration so that all replicate discordances of known variants in dbSNP are assumed to be false negative, the algorithm converges on roughly the same false positive calibration.

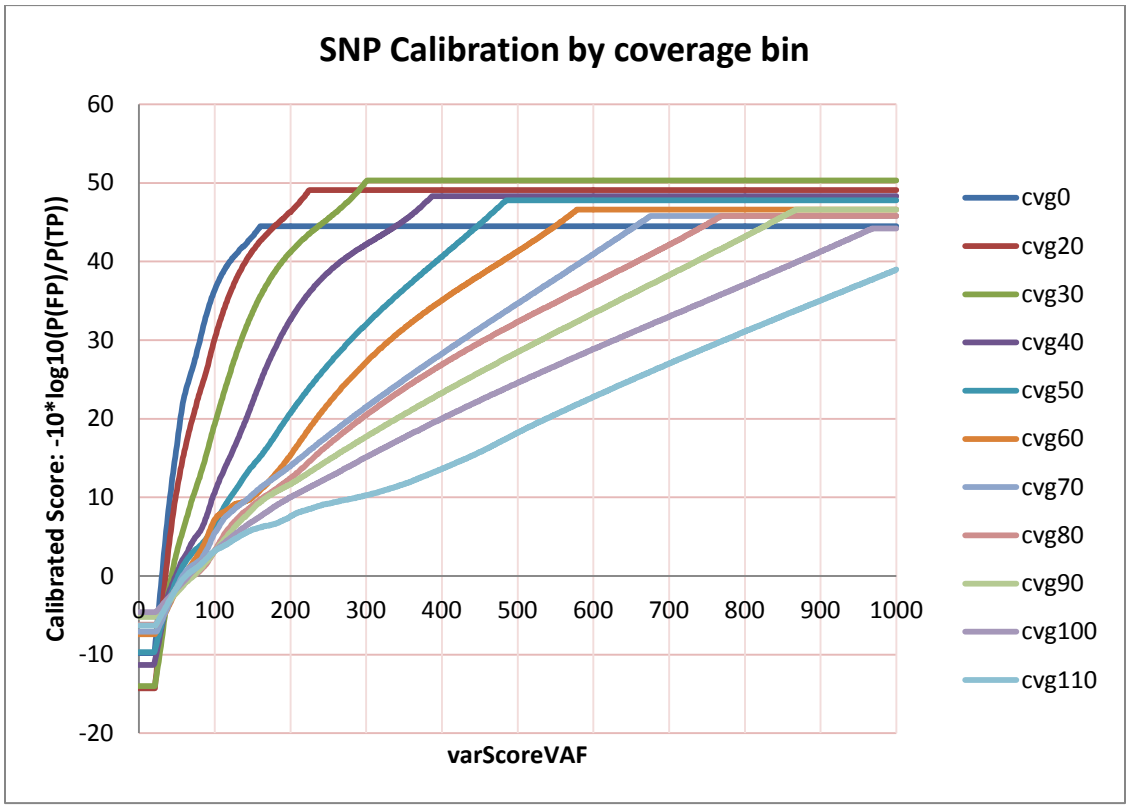
Figure 4: Convergence Using dbSNP Variants as Priors



Calibration by Coverage Bin

Figure 5 shows the per-coverage-bin calibration of SNPs after smoothing. The bins are labeled by the minimum coverage present in that bin. Due to the lack of false events, the calibration cannot be estimated higher than some maximum calibrated score. Also, higher coverage bins have worse calibration curves, indicating that the primary cause for variant errors at high score is DNBs that are not generated in accordance with our model of DNB generation.

Figure 5: Per-Coverage-Bin Calibration of SNPs After Smoothing



20% Allele Fraction Calibration

Figure 6 compares the 20%AF calibration to the 50%AF calibration, for the coverage 40 bin. The calibration curve reveals that we can be much more confident in variants at low *varScoreVAF*, if we assume all variants are present at 20% allele fraction. This behavior is leveraged in the calculation of scores for somatic variants identified in CGA™ Tools calldiff. Specifically, calldiff uses this fact to improve the ROC for low allele fraction variants by using a mixture model where 50% of variants are present at 50%AF and 50% of variants are present at 20%AF.

Figure 6: Comparison Between 20%AF Calibration and 50%AF Calibration