QIAGEN

White Paper

White paper on Probabilistic Variant Caller *1.1*

April 3, 2013

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com · support-clcbio@qiagen.com

Contents

1 Introduction

This is a White Paper on the Probabilistic Variant Detection analysis tool available for CLC Genomics Workbench and Genomics Gateway. The tool can run on the CLC Genomics Workbench as well as on the CLC Genomics Server.

The Probabilistic Variant Caller is designed to call genomic variants from sequencing reads mapped to a reference genome. All kinds of sequencing technologies are supported. Furthermore, the algorithm is able to call variants in haploid (e.g. bacteria), diploid (e.g. human) as well as polyploid genomes (e.g. cancer, higher plants).

The first part of this paper explains in detail the algorithm behind the Probabilistic Variant Caller while the second part presents the results of performance benchmarks for the variant caller. The benchmarks were performed on an *E. Coli* paired-end Illumina dataset to measure specificity when calling heterozygote variants, as well as on two *Homo sapiens* Illumina datasets to give an idea about the sensitivity.

2 The Probabilistic Variant Caller

The purpose of the Probabilistic Variant Caller is to call Single Nucleotide Variants (SNVs), insertions, deletions and MNVs (Multiple Nucleotide Variants) using read mapping data. It can detect variations in datasets from haploid (e.g. bacteria), diploid (e.g. human) and polyploid organisms (e.g. cancer and higher plants).

The algorithm combines a Bayesian model with a Maximum Likelihood approach to calculate prior and error probabilities for the Bayesian model.

After prior and error probabilities have been calculated, the probability for each combination of alleles (e.g. A/G) after observing a certain combination of nucleotides from the reads at every position in the genome is determined. This probability is then used to determine which of the allele combinations is the most likely for each position. If the base or bases identified do not match the base at that position in the reference genome, a variation is called.

2.1 Calculation of the prior and error probabilities

The prior probabilities are estimated using only the mapped reads through four rounds of Expectation Maximization and are calculated for each potential combination of alleles (site types). Thus, the prior probabilities reflect the likelihood of observing each combination of alleles in the genome studied. The reference sequence is not taken into account during the first part of the analysis. More about the Maximum Likelihood estimation (MLE) can be found at `http://en.wikipedia.org/wiki/Maximum_likelihood`.

For a diploid organism, the initial parameters for the priors, which are then updated, are shown in Table 1. The sum of the probabilities for all site types is always 1.

If the expected ploidy level is set to 1, analogous values to table 1 are calculated. Here, only the values for the homozygous site types like A, C, G, T and - would be calculated.

If the expected ploidy is set to 3, the analogous values are calculated, which here would be values for site types like A|A|A, A|C|G, G|G|- and so on.

Error probabilities are calculated alongside the priors for each observed allele and assumed

Site Type	Prior probability
A/A	0.2475
A/C	0.001
A/G	0.001
A/T	0.001
T/C	0.001
T/G	0.001
T/T	0.2475
G/C	0.001
C/C	0.2475
G/G	0.2475
G/-	0.001
A/-	0.001
C/-	0.001
T/-	0.001

Table 1: Site Types for a diploid organism with example probabilities.

reference allele, before the reference sequence is incorporated into the analysis. Table 2 illustrates an example of the values calculated in an error probability matrix.

	A	**C**	**G**	**T**	**-**
A	0.90	0.025	0.025	0.025	0.025
C	0.025	0.90	0.025	0.025	0.025
G	0.025	0.025	0.90	0.025	0.025
T	0.025	0.025	0.025	0.90	0.025
-	0.025	0.025	0.025	0.025	0.90

Table 2: Error probability matrix - observed sequenced nucleotide in read versus actual nucleotide at this position.

If quality values are available, an error matrix is calculated for each quality value.

2.2 Calculation of the likelihood

After the prior and error probabilities have been estimated, the calculation of the likelihood is undertaken. For every combination of reference allele (site types) and nucleotide in every read, the probability of the observed allele being the same as the reference is calculated. These probabilities are then multiplied for all nucleotides in the reads at that position.

Here is an example:

Assumed reference allele: A/C

Read 1: C [$\frac{1}{2}$ (P(C|A)) + $\frac{1}{2}$(P(C|C))] *

Read 2: C [$\frac{1}{2}$(P(C|A)) + $\frac{1}{2}$(P(C|C))] *

Read 3: A [$\frac{1}{2}$(P(A|A)) + $\frac{1}{2}$(P(A|C))] *

Read 4: A [$\frac{1}{2}$(P(A|A)) + $\frac{1}{2}$(P(A|C))] *

Read 5: T [$\frac{1}{2}$(P(T|A)) + $\frac{1}{2}$(P(T|C))]

Here, P(X|Y) is the probability that we will observe nucleotide X in a read when the true reference sequence is Y.

2.3 Calculation of the posterior probability for each site type at each position in the genome

Based on the probabilities calculated, one can determine which of the site types is the best fit at each position in the genome. The site type determined to be the most likely at each position can then be compared with the allele in the reference sequence at the same position. If it is likely to be different, it suggests the presence of a variation.

Therefore the posterior probability is formed as follows:

$$P(site\ type|Obs) = \frac{P(Obs|site\ type) * P(site\ type)}{P(Obs)}$$

where

$$P(Obs) = \sum_{Site\ types} P(Obs|site\ type) * P(site\ type)$$

2.4 Comparison with the reference sequence and identification of candidate variants

Once we have all of the probabilities for each combination of alleles for all positions in the reference sequence, the next step is to determine which of them have the highest probability of existing in the sample. These are the candidate variations. Nucleotide combinations that are the same as the reference sequence are not reported. At this point in the algorithm, a probability threshold is taken into consideration, utilizing a threshold provided by the user.

The threshold provided by the user indicates how sure one would like to be that the candidate variant differs from the reference type. The threshold is applied by the Probabilistic Variant Caller by considering the inverse situation: is the probability of the candidate variant being the same as the reference position lower than 1 minus the threshold. So, for a user-provided threshold of 90%, the Probabilistic Variant Caller requires that any given site type has a probability of less than or equal to 0.1 (i.e. 1 - 0.9) of being the same as the reference type. For example, if a user gave a threshold of 90%, and a particular position was found to have a probability of 15%, or 0.15, of being the same as the reference (equivalently, having a probability of 85% of being different than the reference), then this position would not be called as a variant. If the threshold had been set to 80%, then this position would have been called as a variant, as 0.15 is less then 0.20, or in other words, the position has a high enough probability of being different than the reference according to the user-defined threshold, to be reported as a variant.

If a variant is called at a given position, the second step performed by the algorithm is to determines the allele combination (type site) with the highest probability. This type site, together with the corresponding probability, will be reported as the candidate variant.

2.5 Posterior filtering and reporting of variants

The algorithm includes several filters to reduce the rate of false positive variants. These filters can be activated or deactivated by the user.

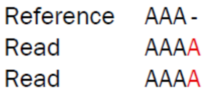
2.5.1 Filtering of variants in homopolymeric regions

Different sequencing platforms generate different types of sequencing errors, which can cause incorrectly called variants. The most common source of sequencing errors across platforms is the determination of nucleotides in so-called homopolymeric regions. These are regions that include stretches of the same nucleotide (e.g. AAAAA or TTTTTTTT). As a result of the internal chemistry used on platforms such as 454 and Ion Torrent, the number of identical nucleotides in such regions is often not accurately reported. This causes variant-callers to identify within homopolymer regions, insertions and deletions not actually present in the sample. The Illumina platform has a similar problem in which one nucleotide is surrounded by other nucleotides of the same type (e.g. AAAAGAAAA). Such cases are sometimes misread, with the different base identified as being the same as the surrounding nucleotides. This can lead to incorrect SNV calls. For example, a region of AAAAGAAAA in the sample may appear as AAAAAAAAA in the read. This could lead to a variant allele, A, being called where the G appears in the reference, when in fact the sample itself did contain a G at that position.

The Probabilistic Variant Caller includes an internal filter to recognize and prevent variants being reported in homopolymeric regions.

The 454/Ion Torrent homopolymer filter does not report insertion or deletion variants found at the ends of regions of two or more nucleotides of the same kind (e.g. AA, TT, GGG).

An example is given in figure 1:



Reference	AAA -
Read	AAAA
Read	AAAA

Figure 1: *Example of insertions filtered out using the 454/Ion Torrent homopolymer filter.*

The red A will not be reported as a variant when the 454/Ion Torrent filter is applied, as it is characteristic of sequencing errors frequently observed on those platforms.

2.5.2 Forward/reverse reads support

This filter is recommended in all cases where an even distribution of forward and reverse reads at every position is expected. However, it should not be used for data sets such as large amplicons, where the ends of an amplicon are likely to be covered by only forward or reverse reads.

Due to sequencing or PCR artifacts and mapping issues, there can be some positions in the reference genome where only forward or only reverse reads are aligned. This can lead to certain alleles being present on one strand only.

If there is a strand bias from sequencing visible in the quality output check after sequencing, these should be regarded as suspicious regions that should be ignored during variant calling. If the user has selected the forward/reverse read support option, only variants that have a forward/reverse read balance of at least 0.05 are reported.

The forward/reverse balance is calculated as:

$$Min((\#forward/\#total)(\#reverse/\#total))$$

where

#forward = number of forward reads supporting the variant
#reverse = number of reverse reads supporting the variant
#total = all reads supporting the variant

3 Investigating the false positive rate for calling heterozygote variations

To investigate the false positive rate for calling heterozygote variations, we run the Probabilistic Variant Caller on aligned reads mapped to the haploid genome of *E. coli* with the ploidy parameter set to 2. This allows us to identify the rate at which variants would be called on a diploid organism, if there were no heterozygous variations. In this example, to ensure that there are no false positive variants, both the reads and the reference genome are derivied from the same *E. coli* strain (DH10B), so under ideal circumstances no variants should be identified. Thus, all of the variant calls obtained should represent false positives, which can then be counted and investigated further.

3.1 Material and Parameters

Data was obtained from E. coli strain DH10B, sequenced using paired-end MiSeq. Illumina reads were downloaded from the Illumina website (http://www.illumina.com/systems/miseq/ecoli.ilmn), imported into CLC Genomics Workbench and mapped against the reference sequence from the same *E. coli* strain using the CLC read mapper (Parameters used: Deletion cost = 3, Insertion cost = 3, Length fraction = 0.8, Mismatch cost = 2, Similarity fraction = 0.8, Min paired distance = 1, Max paired distance = 1000).

The Probabilistic Variant Caller was then run and the results used for this analysis. (Parameters used: Minimum coverage = 10, Maximum expected variations (ploidy) = 2, Ignore non-specific matches = Yes, Require presents in both forward and reverse reads = No, Variant probability = 90.0, Ignore broken pairs = Yes).

3.2 Mapping and Variant calling results

14,078,150 (97.73%) of 14,405,136 sequencing reads were mapped to the reference genome, of which 13,901,326 (96.50%) reads were mapped in pairs. The average read length was 151bp.

The Probabilistic Variant Caller called 61 variants on a reference sequence length of 4,686,137bp. Most of the variants called were observed to be heterozygous and thus, for this analysis, can be considered false positives. 57 variations are Single Nucleotide Variants (SNVs) and 4 are Multiple Nucleotide Variants (MNVs). With the exception of one MNV (a complex insertion),the reference allele was called alongside the variant, and the reference allele frequency was always observed to be higher than the frequency of the genotype allele.

3.3 Filtering of variant calls

The "Filter Marginal Variants" tool can be used to filter against specific thresholds for allele frequency, forward-reverse balance and average base coverage subsequent to calling variations. Using this tool and filtering for a minimum average base quality of 20 leaves only one variant. This variant is the complex insertion mentioned above.

3.4 Running the Variant Caller in haploid mode

Running the variant caller with the correct haploid setting yielded no variations, indicating that no false positive variations would have been called on the haploid genome with the appropriate settings provided to the variant caller.

3.5 Discussion and Conclusions

Running the CLC variant caller on the paired-end Illumina MiSeq data from the *E.coli* DH10B strain test dataset and with a diploid setting resulted in 61 potential false positives. This corresponds to a specificity of greater than 99.99%.

With the exception of one insertion, all called variations were heterozygous, with the reference allele displaying a higher allele frequency. This single exception, a one base (T) insertion, is putatively missing in the reference genome for *E. coli* strain DH10B and therefore may in fact be a correctly called variant. All other variants observed had a low average base quality (below 20) and are thus likely to be false positive variant calls. Therefore, we recommend using the "Filter Marginal Variants" tool with a threshold for the average base quality of 19 or 20.

Running the variant caller with the correct setting, as described above (Maximum expected variations (ploidy) = 1) resulted in no variations being called, or in other worse, there is a specificity of 100%. Therefore, the correct setting of the ploidy level apparently has a substantial impact on the overall specificity of the variant calling.

4 Variant detection on chromosome 21 of human individual NA19240

Acurately determining the sensitivity and specificity of the Probabilistic Variant Caller in homozygotes as well as heterozygotes for higher organisms is difficult because there is no test dataset available for which all variations have been validated. However, for humans, there exists several good datasets for which some of the variations have been experimentally verified (i.e. by Sanger sequencing). We use one of these datasets from Illumina to measure the sensitivity of our variant caller and to compare it with the CASAVA 1.7. variant caller.

4.1 Material and Parameters

BAM formatted read mappings, CASAVA variant calls and validated variants for chromosome 21 of human individual NA19240 were downloaded from the Illumina website (http://www.illumina.com/truseq/tru_resources/datasets.ilmn).

We downloaded the following datasets and imported them as tracks into CLC Genomics Workbench:

- Read mappings in BAM format

- CASAVA variant calls and validated variants for chromosome 21 of human individual NA19240

File name	**Description**	**Number of variant alleles**
NA19240_,HiSeq_100_chr21.bam	Read mapping file (reference: b36hg18) created by Illumina	
NA19240_GAIIx_100_NCBIb36_CASAVA-1.7.0_snps.txt	Variations called with the CASAVA 1.7.0 software	66,107
dbSNP130_NA1924_snps.txt	Validated variations. Consists of those dbSNP 130 variants that were also observed in a capillary sequencing study of NA19240 ([Kidd et al., 2008])	11,515

Table 3: Files downloaded and imported into CLC Genomics Workbench for determination of sensitivity and specificity of the Probabilistic Variant Caller *1.1*.

In total 13,528,800 reads of 100bp length, were present in the BAM file. 12,726,401 reads were mapped to chromosome 21 of the human b36/hg18 genome using the CASAVA software, of which 11,805,385 reads were found in aligned pairs. 76,200 paired reads were marked as broken due to the distances between pairs being outside the distance range expected, or because incorrect orientations were observed. For 844,817 reads the mate could not be mapped. The mean distance between the paired reads was 312.49bp with a standard deviation of 10.84bp. The average read coverage observed on chromosome 21 was 27.11x.

Using the above data, the Probabilistic Variant Caller was run to generate a list of variations. (Parameters: Minimum coverage = 10, Maximum expected variations (ploidy) = 2, Ignore non-specific matches = Yes, Require presence in both forward and reverse reads = No, Variant probability = 90.0, Ignore broken pairs = Yes). Variant calls were subsequently filtered for forward/reverse balance (> 0.05) and average base quality (> 19).

To aid in comparing the Probabilistic Variant Caller with Casava, the tool was run a second time with all reads taken into account (Parameters as above except: Ignore broken pairs = No). As before, these results were filtered for forward/reverse balance (> 0.05) and average base quality (> 19).

SNVs called by CASAVA 1.7 and the validated SNVs from the study of [Kidd et al., 2008] were transferred to GVF format and were imported as tracks in CLC Genomics Workbench. SNVs were ignored, as these seem to result from unspecific read mappings or wrongly transferred coordinates from hg19.

CASAVA 1.7 and validated SNVs were then compared with the called variants from the Probabilistic Variant Caller.

4.2 Results

With broken pairs discarded, 123,441 variations were called by the Probabilistic Variant Caller. 106,272 of them were SNVs, 7659 were insertions or deletions, and 9510 were classified as MNVs. After filtering for forward/reverse balance (> 0.05) and average base quality (> 19), this was reduced to 76,592 variants, of which 65,637 were SNVs, 7064 are insertions or deletions, and 3891 were MNVs.

With broken pairs included, the Probabilistic Variant Caller predicted 91,031 variants (78,686 SNVs and 4506 MNVs). Filtering for forward/reverse balance (> 0.05) and average base quality (> 19) reduced this to 82,474 variants (70,579 SNVs and 4114 MNVs).

4.2.1 Comparison of validated SNVs with called variants by the Probabilistic Variant Caller and Casava

Of the 11,515 validated variations, 8946 SNVs were called by both Casava and the Probabilistic Variant Caller. An additional 190 variants were called by Casava that were included in MNVs called by the Probabilistic Variant Caller. Thus, a total of 9136 variants (79.3%) were called by both variant callers.

A further 525 validated SNVs were called by the Probabilistic Variant Caller but not Casava 1.7, whereas 695 validated SNVs were called by Casava 1.7 but not the Probabilistic Variant Caller.

1158 variations in the validation dataset were not called by either variant caller, with issues such as insufficient read coverage or low allele frequency in mapped reads accounting for those investigated.

When all reads are used (broken pairs are included) by the CLC Probabilistic Variant Caller, as is the default for Casava, the number of variants called by both variant callers increases to 9385 (81.5%), while the number of variants called only by the Probabilistic Variant Caller falls to 566. Furthermore, the number of validated variants called only by Casava drops to 447. This still leaves 1117 validated variants undetected by either variant caller.

4.2.2 Comparison of non-validated SNVs and MNVs called by the Probabilistic Variant Caller

A substantial number (56,226 SNVs and 3719 MNVs giving a total of 59,945 variants, or 87.5%) of the 67,009 variants called by the Probabilistic Variant Caller are not present in the set of validated variants from [Kidd et al., 2008]. This number becomes even greater if unpaired reads are included in the analysis, increasing to 72,613 (88%) of all called variants.

Casava 1.7 does not provide information on insertions and deletions, thus we only consider the SNVs and MNVs for this analysis. Of the 59,945 variants detected by the Probabilistic Variant Caller, 49,384 matched SNVs found by Casava (82.39%). A further 390 variants represent partial matches, that is, as MNVs mapping to the same location as an SNV called by Casava). However, 37 SNVs called only by the Probabilistic Variant Caller are also present in HapMap 3.3, suggesting that the number of true positives is actually higher than that predicted by simply looking at the concordance between Casava and the Probabilistic Variant Caller.

4.3 Discussion and Conclusions

Using publicly available datasets composed of validated and called variants as well as mapping files from Illumina, we are able to show that the Probabilistic Variant Caller calls variants with a high sensitivity. Over 80% of the validated SNVs from [Kidd et al., 2008] were detected. Variants not detected and which were investigated further were in regions with insufficient coverage (below 10x) or where a low number of reads were observed containing the variants. Thus, when all factors are taken into consideration, and all reads are used to call variations, the Probabilistic Variant Caller called more validated SNVs than Casava 1.7.

5 Variant detection on chromosome 20 of human individual NA12878 and comparison with GATK

GATK is currently one of the most commonly used variant callers. We compare our variant caller with the UnifiedGenotyper tool from GATK using the data from the GATK resource bundle provided on the GATK website (`http://www.broadinstitute.org/gsa/wiki/index.php/GATK_resource_bundle`).

We also compare insertions and deletions called by the Probabilistic Variant Caller with the gold standard dataset from [Mills et al., 2011] and with insertions and deletions from the 1000 Genomes project.

5.1 Material and Parameters

We downloaded the following datasets and imported them as tracks into the CLC Genomics Workbench:

- Reads mapped to human reference genome 19 (b37/hg19) done by BWA in BAM format
- Results from the UnifiedGenotyper run (default parameters) on this dataset and the gold standard indel dataset from [Mills et al., 2011] in vcf format

In total, 50,663,069 mapped reads to chromosome 20 of the human b37/hg19 genome are present in the BAM file with an average read length of 101. 1,101,277 reads were found in "broken pairs", of which 10,604 reads were discarded because of an inverted read or incorrect mapping distance. For the remaining 1,090,673 read pairs, the mate could not be mapped. The resulting average coverage of chromosome 20 was 74.65x.

The Probabilistic Variant Caller was then run on the mappings (Parameters: Minimum coverage = 10, Maximum expected variations (ploidy) = 2, Ignore non-specific matches = Yes, Require presence in both forward and reverse reads = No, Variant probability = 90.0, Paired mode = only paired reads are considered). A second run was performed using less strict filtering. (Parameters: Minimum coverage = 10, Maximum expected variations (ploidy) = 2, Ignore non-specific matches = Yes, Require presence in both forward and reverse reads = No, Variant probability = 50.0, Ignore broken pairs = Yes).

Variant calls were then filtered using the "Filter Marginal Variant Calls" tool, with an average base quality threshold of 19.

File name	Description	Size and Type of Variants
NA12878.HiSeq.WGS.bwa.cleaned.recal.hg19.20.bam	Read mapping done by BWA	
NA12878.HiSeq.WGS.bwa.cleaned.recal.hg19.20.vcf	Variant calls from UnifiedGenotyper tool run with default parameters	98,243 SNVs
Mills_and_1000G_gold_standard.indels.b37.sites.vcf	Gold standard indel dataset from [Mills et al., 2011]	26,247 indels 1,372 MNVs
hapmap_3.3.hg19.vcf	HapMap 3.3 variants for chr20	37,150 SNVs 2 indels
1000 Genomes phase 1 data from Ensembl in gvf format	Variants from the 1000 Genomes project pilot 1 and 3	480,933 SNVs 51,5836 indels 133 MNVs
dbsnp_135.hg19.vcf	dbSNP 135 variants for chr20	1,170,634 variants (SNVs, MNVs and indels)

Table 4: Files downloaded and imported into the CLC Genomics Workbench for comparison of the Probabilistic Variant Caller *1.1* with the UnifiedGenotyper tool from GATK.

5.2 Results

In total, 106,569 variants were called by the Probabilistic Variant Caller. This was composed of 92,026 SNVs, 11,732 indels and 2811 MNVs.

After filtering out low average base quality variants, 95,436 were left, composed of 87,915 SNVs, 4776 indels and 2745 MNVs.

5.2.1 Comparison of SNVs and MNVs

Of 87,915 SNVs identified by the Probabilistic Variant Caller, 85,304 SNVs (97.0%) were also called by the UnifiedGenotyper tool and a further 497 MNVs reported by the Probabilistic Variant Caller were also detected in the results from GATK by linking adjacent SNVs.

However, 9964 variants (10%) called by the UnifiedGenotyper tool were not found by the Probabilistic Variant Caller. Of these, only 5 are present in HapMap. Two of these have a read coverage of less than 10 in the dataset analyzed, thus, this region would not have been considered by the variant caller. One more of these 5 variants was detected if the posterior probability threshold was lowered to 50%. Overall, using the relaxed probability cutoff of 50%

reduced the number of variants called by the UnifiedGenotyper tool but not the Probabilistic Variant Caller to 9318. This number can be reduced to 8111 by allowing low quality variations to remain in the CLC-generated dataset. In other words, 1207 variants have an average base quality of less than 19. For 430 of the 8111 non-called variants, there was a read coverage of less than 10 in those regions.

In contrast, 2611 variants called by the Probabilistic Variant Caller were not called by the UnifiedGenotyper tool. Of these, 3 of them are present in HapMap 3.3, with many of the remaining variants found in low complexity regions.

5.2.2 Comparison of Insertions and Deletions

Of the 4776 small insertions and deletions identified by the Probabilistic Variant Caller, 3298 (69%) were also identified in both the dataset of [Mills et al., 2011] and the in the 1000 Genomes data. Of these, 2748 variants are direct matches and the rest are partial matches.

5.3 Discussion and Conclusions

More than 90% of the variants called by the CLC Probabilistic Variant Caller and the UnifiedGenotyper tool are identical for the GATK resource bundle dataset provided by the Broad institute.

For variants detected by only one variant caller, the presence of the variant in HapMap supports the likelihood that the called allele is a true positive. However, this accounts only for a small percentage of the non-concordant variations.

Furthermore, a greater number (1207) of the variations called only by the UnifiedGenotyper were called where there was an average read base quality of less than 19, which may indicate that these are sequencing errors, rather than real variations. In comparison, the CLC Probabilistic Variant Caller called a greater number of variants in low complexity regions (e.g. homopolymer regions), where it is known that the rate of sequencing errors is higher. Thus, a low complexity filter could help to decrease the number of potential false positives.

In conclusion we cannot make any assumptions about the sensitivity or specificity of either tool as applied to this specific dataset, but we have shown that the UnifiedGenotyper tool from GATK and the Probabilistic Variant Caller behave similarly in calling variants in a diploid organism.

Finally, approximately 70% of the insertions and deletions called are present in the validated datasets, showing the high accuracy of the Probabilistic Variant Caller in this regard. However, it should be mentioned that, especially for detecting insertions and deletions, an accurate mapping or realignment of reads to the reference genome is required. Specific sequencing errors, which occur mainly in data generated by the 454 and Ion Torrent platforms, can significantly alter the performance in calling insertions and deletions.

6 Conclusion

In this White Paper we have demonstrated the performance of the Probabilistic Variant Caller *1.1* on three publicly available datasets.

We highlighted the Probabilistic Variant Caller's very high specificity of over 99.99% when tested on an *E. coli* dataset, and illustrated that when changing the ploidy parameter to a heterozygote model for variant calling, the false positive rate is increased only slightly. In combination with the "Filter Marginal Variants" tool for filtering variants that have a low average base quality and by choosing the right ploidy setting of 1 for a haploid organism, the specificity was increased to 100%.

Using two human datasets available from Illumina and GATK, we show the sensitivity of the Probabilistic Variant Caller and compare it to the Casava variant caller provided by Illumina and the UnifiedGenotyper tool provided by GATK. In both cases, the test data bundles came with a bam file of reads mapped to the human reference genome, making it possible to test the CLC variant caller against other variant callers independently of the mapping.

Using a validated dataset, we also show that the Probabilistic Variant Caller is able to call SNVs with an overall sensitivity of over 80% and that it called more validated SNVs than Casava when unpaired reads were included in the analysis. When calling insertions and deletions, the CLC Probabilistic Variant Caller achieved a sensitivity of over 70%, which is comparable with other tools. A local realignment of some reads as well as a better overall coverage would substantially increase this.

Over 90% of the variants called by the Probabilistic Variant Caller and the UnifiedGenotyper tool from GATK overlap, showing that both tools performed similarly. However, the number of variants called uniquely by one tool or the other raises the question of whether there are true positives that are missed by each variant caller, or whether these variants represent false positives, which one tool then properly dismisses. Unfortunately only experimental validation of these variants would give a proper answer to thise question. Nevertheless, are there some indications that at least a small minority of these variants may be false positives. For example, a small number of variants called only by the UnifiedGenotyper tool show an average base quality of less than 19. Similarly, a very few variants called only by the Probabilistic Variant Caller are in low complexity regions, like homopolymeric regions, where the number of wrongly called bases is typically higher than average. In such cases, increased filtering on quality, or for homopolymer regions, could help to decrease the false positive rate.

In conclusion, the performance of the Probabilistic Variant Caller *1.1* for CLC Genomics Workbench and CLC Genomics Server is excellent and comes out well in comparisons with other commonly used variant callers for haploid and diploid genomes. Unfortunately, there are no test datasets with validated variants available for higher ploidy organisms at the moment, which severely restricts our ability to demonstrate that the Probabilistic Variant Caller's strong performance in calling heterozygous variants for polypoidy organisms.

References

[Kidd et al., 2008] Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64.

[Mills et al., 2011] Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., Kemeza, D. M., Strassler, D. S., Ponting, C. P., Webber, C., and Devine, S. E. (2011). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*, 21(6):830–839.