BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 8 2010, pages 1029–1035
doi:10.1093/bioinformatics/btq092

Sequence analysis

Advance Access publication on February 26, 2010

High quality SNP calling using Illumina data at shallow coverage

Nawar Malhis* and Steven J. M. Jones

Genome Sciences Centre, BC Cancer Agency, Vancouver BC, Canada

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Detection of single nucleotide polymorphisms (SNPs) has been a major application in processing second generation sequencing (SGS) data. In principle, SNPs are called on single base differences between a reference genome and a sequence generated from SGS short reads of a sample genome. However, this exercise is far from trivial; several parameters related to sequencing quality, and/or reference genome properties, play essential effect on the accuracy of called SNPs especially at shallow coverage data. In this work, we present Slider II, an alignment and SNP calling approach that demonstrates improved algorithmic approaches enabling larger number of called SNPs with lower false positive rate. In addition to the regular alignment and SNP calling, as an optional feature, Slider II is capable of utilizing information about known SNPs of a target genome, as priors, in the alignment and SNPs calling to enhance it's capability of detecting these known SNPs and novel SNPs and mutations in their vicinity.

Contact: nmalhis@bcgsc.ca

Supplementary information and availability: Supplementary data are available at *Bioinformatics* online and at http://www.bcgsc.ca/platform/bioinfo/software/SliderII

Received on October 27, 2009; revised on January 14, 2010; accepted on February 23, 2010

1 INTRODUCTION

The ability to perform an unambiguous alignment of sequence reads (Langmead *et al*., 2009; Li *et al*., 2008; Li and Durbin, 2009; Li *et al*., 2009; Lin *et al*., 2008; Malhis *et al*., 2009; Schatz *et al*., 2007; Smith *et al*., 2008) and the detection of sequence variants (Li *et al*., 2008; Li *et al*., 2009) is an essential task in processing the output of second generation sequencing. As this technology evolves, algorithms used in alignment and single nucleotide polymorphism (SNP) calling need to be modified to adapt to rapidly changing parameters such as read length, overall read quality, the number of reads and the sequence coverage being generated. Furthermore, information about reference genomes, in particular known polymorphisms, have the potential to improve the alignment and SNP calling accuracy. For example, the current human reference sequence typically used for alignment reflects, within any region, a single haplotype that will contain private SNPs as well as ones of low frequency within the human population. Such polymorphic regions, through their generation of an alignment mismatch, will impede the ability to determine nearby mutations of interest.

A commonly used application for alignment and SNP calling from Illumina reads is the MAQ aligner (Li *et al*., 2008). MAQ uses hash tables to align the most probable sequence of each read to a reference genome, and then uses the Bayesian theorem to construct consensus sequence from these aligned and calibrated most probable reads, a list of SNPs are called on differences between the reference sequence and the consensus, and filtered by a set of threshold cut-off parameters such as SNP quality score, minimum and maximum coverage, the maximum allowed number of SNPs in a small window (dense SNPs) and the size of this small window. Previously, we reported on the ability of the Slider algorithm (Malhis *et al*., 2009) to use a merge-sort approach for aligning the probability (prb) values at each position of Illumina short reads. By utilizing the probability values to reconstruct likely sequences from each read, Slider was able to achieve considerably higher alignment accuracy, reducing the number of mis-mapped reads arising due to sequence error.

The algorithmic approach of Slider II continues to utilize the probability of all four possible nucleotides generated by Illumina to improve the overall alignment. This information has also been incorporated to improve SNPs calling quality.

In brief, Slider II high SNP calling accuracy is a result of higher alignment accuracy of Slider (Malhis *et al*., 2009), higher SNP calling capability achieved in utilizing information provided in prb data and higher SNPs filtering accuracy:

- High alignment accuracy: while most aligners are designed to align the most probable sequence, higher alignment accuracy is achieved by Slider by aligning the prb sequences for all four bases (Malhis *et al*., 2009). Slider II uses the merge-sort approach of Slider to align seeds of the first 31 bases (higher quality bases) of each read and then extend these seeds to full reads.
- After the alignment step, Slider II utilizes more information provided in prb data for SNP calling:
  - Using all four prb values in a consensus construction yield a more discriminatory probability accumulation compared to only using the probability of the most probable base (mpb).
  - With heterozygous SNPs, the expectation is to have approximately equal coverage representation for the reference nucleotide and the allele nucleotide, however, this coverage, due to the low numbers, is likely to be skewed in favor of one nucleotide or another. Since probability accumulates exponentially to the level of coverage, such coverage unbalance is likely to mask the probability of the under represented nucleotide, therefore, in order to overcome this, Slider II builds a coverage probability consensus that only include data when the reference nucleotide has a lower probability than some threshold value. Again, having the probability values of all four nucleotides enables more accurate filtering of these bases.

*To whom correspondence should be addressed.

- Higher SNP filtering accuracy: in addition to dense SNPs (Marth *et al*., 1999; Li *et al*., 2008) and SNPs with high coverage (Li *et al*., 2008), Slider II also utilizes SNP average location in the set of covering reads to reduce the effect of structural rearrangements on SNPs calling. We have determined that bona fide SNPs typically are randomly distributed along the lengths of the covering reads, whereas false positive SNPs tend to show a biased distribution, see Section 2.2.2.
- And finally, a unified SNP score which correlates with SNP prediction accuracy is generated by penalizing or filtering out some observed SNPs that (based on SNP average location, total coverage and dense SNPs) are likely to be false positive results from structural rearrangements.

We use the phrase 'shallow coverage' for those sections of DNA that are covered by two to four reads only. Since coverage of the reference vary between sections, the ability of identifying SNP at shallow coverage is particularly important for those sections that are less-covered.

In addition to regular alignment and SNP calling, Slider II provides an option of utilizing known SNPs from SNP databases, as priors, in the alignment and the SNP calling process. Utilizing known SNPs improve alignment quality by accurately aligning reads on locations with more variants (known and unknown variants), otherwise such reads would either not be aligned or misaligned to an inaccurate location. This helps in the general identification of SNP, and known SNPs are also used as priors in the SNP calling process, which enable SNP calling of known SNPs at lower coverage.

2 METHODS

2.1 Alignment

The first 31 bases of each prb line are used as seed sequences, which aligned to the reference sequence as explained previously in Malhis *et al*., 2009. Aligned seeds are extended to full reads, RD*s*, allowing up to three base mismatches per RD. When using paired-end sequence data (PET), if only one side can be uniquely matched to the reference, SliderII will attempt:

- If the other side has multiple matches. Slider II will resolve the multiple matches using the PET span distribution information.
- If the other side was not successfully aligned. SliderII will attempt to align it without the restriction of one mismatch in a seed.

We define sequence map-ability at any location as the number of different reads of 36 bases that can be uniquely mapped to cover this location. Considering forward and reverse complement, sequence map-ability for any location is an integer in the range of [0 … 72].

To facilitate identifying unique matches, we also define a sequence commonness factor as the number of bases starting at any chromosomal position that are needed to define a unique location on the reference. Reference sequence commonness and sequence map-ability for every point in the reference are calculated and stored during the reference database construction process. Figure 1 shows the percentage of human genome with an equals or higher sequence commonness.

2.2 SNP calling

After alignment of sequenced reads to a reference genome, SNPs can be observed at any location l as a consistent base difference between reads bases and the reference base at l. In many cases the ability to detect SNPs is obfuscated by many factors, including read misalignments, paralogous mapping, indels and genomic rearrangements.

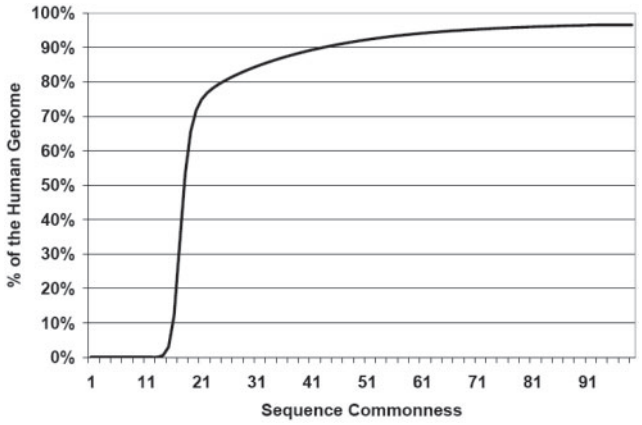

Fig. 1. The percentage of human genome with an equal or higher sequence commonness value.

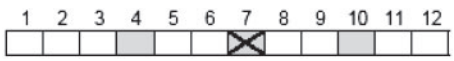

Fig. 2. This 12-base read has a SNP at base 7, plus, some non-crisp bases including bases 4 and 10, where the accurate nucleotide is not the mpb, as a result, the read has 3 points distance of its accurate location on the genome, and if this read is to be matched to the reference with any aligner starting from the most probable sequence allowing up to two base mismatches, it is going to be misaligned.

Read misalignments take place when a read is wrongly aligned to the reference at an erroneous location; this can be due to a combination of factors such as sequencing error(s), SNPs and low map-ability. Regions of low map-ability are particularly problematic, as a single-base call error has a higher probability of allowing the read to be erroneously aligned elsewhere in the genome, a problem that is contained through the more aggressive utilization of the probability data to identify the second most likely base call for poorly called base pairs. Figure 2 demonstrates a scenario resulting from the misalignment of a 12 bp read. Misaligned reads are likely to introduce an inaccurate coverage profile that affects the accuracy of SNP calling, while this inaccurate coverage is less likely to build up with high enough consistency for calling a false SNP; it is likely to add a level of noise capable of masking a SNP at shallow coverage regions. Examples 1 and 2 show how a single misaligned read can easily mask a SNP at shallow coverage. In utilizing the probability values for all four nucleotides, Slider reduces the percentage of misaligned reads significantly compared to aligners that utilize only the most probable sequence or calibrated most probable sequence (Malhis *et al*., 2009).

Paralogous mapping of reads happens when aligning reads originate from repeats or duplications in the sample genome (or near repeats, e.g. a repeat with a single base difference) and are represented only once in the reference (if a repeat is represented more than once in the reference, such reads will end up as multiple matches). Such paralogous mapping should result in abnormally high coverage to these regions of the sequence and will likely generate a large number of false positive SNPs. In some cases, such false positive SNPs will be clustered into dense groups.

Coverage depth: typically, most SNP prediction tools call SNPs utilizing both sequence differences and the sequence coverage of the base mismatch. A single sequence read indicating a base mismatch would not usually be considered sufficient to reliably identify a SNP, and redundant observations are needed. A relevant question is what depth of coverage is required

to robustly call a SNP? To address this, Slider II relies on two main parameters:

- The expected ratio of SNPs in our sample reflects our confidence in the reference genome to match the sample. For example, on one extreme, if we know with absolute certainty that our sample is SNP free (i.e. the SNP free PHIX control), no mater how many reads are consistently showing a mismatch, a SNP should not be called; this can be contributed to contamination, paralogous mapping and/or sequencing errors. On the other extreme, if we have very little confidence in the quality of the reference, we might legitimately accept a base mismatch with very low coverage.
- The alignment quality of the covering reads and the mismatch bases quality: Higher quality alignments with high quality bases can provide enough confidence in a SNP with less number of read coverage.

Most tools (i.e. MAQ) used *q*Cal values and alignment quality to score SNPs, but it requires a user input for a minimum coverage (and maximum coverage) for calling SNPs, Slider II does not need such an input.

In principle, SliderII calls SNPs at any location with consensus coverage different from the reference nucleotide, when the confidence accumulated from the aligned reads is higher than our confidence of the base in the reference genome, which is a function of the expected ratio of SNPs in that genome. The consensus sequence constructed from aligned bases that shows a low-probability for the reference nucleotide, taking into account the alignment quality and the base-calibrated quality, *q*Cal, gives the accumulated probability, $Pacc_{\mathrm{Bass},l}$, of each one of the three nucleotides (not including the reference) at each location l. Whereas, reference base probabilities, $Pref_{\mathrm{Bass},l}$, of each one of the four nucleotides at each location is computed using reference sequence and the expected SNPs ratio; this expected polymorphism ratio vary from one population to another, for example, in the human population a polymorphism is believed to occur on average every 1000 bases, so the default probability of the reference nucleotide can be estimated at 99.9%, and 0.033% for each other nucleotide. Other species are likely to have different SNP ratio, in the *Caenorhabditis elegans*, *CE*. While the Hawaiian strain is likely to have about the same SNP ratio as in the human genome, in the *CE* California strain from Pasadena, the ratio of SNPs is about 1 SNP in 5 kb, and in *CE* N2 Bristol strain, our analysis show this ratio is less than 1 SNP in more than 10 kb.

When utilizing a set of known SNPs as priors, SliderII heuristically set the reference base probabilities such that the reference base is 66.6%, and each possible SNP nucleotide is set to 33.3%, and 0.033% for every other nucleotide; then these four values are normalized to the sum of one.

A SNP is called at a location l if:

- The mpb, at l is not the reference base.
- The confidence accumulated from sequence coverage is higher than the confidence in the reference sequence. In practice, Slider II calculates score_1 by dividing the sum of the probabilities of the consensus two lowest bases, $Pacc_{\mathrm{Bass},l}$, by the sum of those of the reference, $Pref_{\mathrm{Bass},l}$. Score_1 must be >1.
- Pass the dense SNPs condition: while Slider II penalizes SNPs that are likely to appear as a result of structural rearrangements or indels by down scoring or filtering them out in the final SNP scores step after the SNP calling step, the dense SNPs condition can be made more effective if it is implemented during the SNPs calling step as described in the following paragraph.

The dense SNPs condition: a commonly used filtering technique that reduces the percentage of false SNPs that are likely to appear clustered in dense groups. Filtering out dense SNPs has been used by Li *et al.* (2008) and Marth *et al.* (1999). However, while dense base mismatches at contig edges and paralogous (false SNPs resulting from contig edges and paralogous will be discussed in Section 2.2.2) are likely to generate dense false SNPs, in many cases, only few of these base mismatches reach high enough coverage for SNPs calling, which will reduce the effectiveness of this technique, therefore, we introduced the low-quality bases concept: *LQB* are consensus bases in which the probability of the reference base in the consensus is <5%. A SNP can be called if the number of LQBs in any window (of size $\mathrm{Size}_{\mathrm{Window}} = \mathrm{read}_{\mathrm{AverageSize}} - \mathrm{seed}_{\mathrm{Size}}$) that cover the SNP is less than $\mathrm{Size}_{\mathrm{Window}}/3$.

2.2.1 Constructing consensus Consensus is constructed by accumulating sequence bases probability from coverage read bases probabilities using Bayes theorem. The prb values provided in prb files first needs to be adjusted for sequencing quality and alignment quality. Calibrated sequencing base qualities, *q*Cal, are derived from prb values by computing the percentage of mapped bases with good mapping quality for every prb value. In practice a high accuracy prediction of read alignment quality (mapping quality) is not possible given that the similarity between the reference and the sample is unknown; many alignment parameters such as contamination, chromosomal rearrangements, etc., are not measurable. We propose a heuristic alignment weight W_a (score) based on three parameters:

- the size of the aligned read, *Size*;
- the calibrated probability of the aligned base at read location x, P_x; and
- the reference commonness factor at the start of the read, k.

Aligned seeds are first extended and an alignment weight is calculated:

$$W_a = \prod_{i=1}^{k} P_i + \frac{\sum_{j=k+1}^{\mathrm{size}} P_j}{k}$$

If W_a is greater than one, W_a is set to be one.

W_a is likely to be correlated with alignment accuracy sine bases quality and read size that are positively correlated with alignment quality and k, which is negatively correlated with alignment quality and all correlate in the same way with W_a. The initiative behind this is that the first part

$$\prod_{i=1}^{k} P_i$$

should calculate the probability of the first k bases of the read to match the reference at the aligned location. If the sequenced sample was identical to the reference, this should be enough to identify a unique match. The (*size*–k) bases at the tail of the read (in the following part) are used to adjust for variations, smaller k value (which reflect higher reference complexity around the alignment area) should result a higher significant for these bases in the second part.

Alignment quality is:

$$Q_a = 0.5 + 0.5 * W_a$$

Base calibrated quality adjusted for alignment quality $q\mathrm{Cal}_a$ is:

$$q\mathrm{Cal}_a = q\mathrm{Cal} * Q_a + (1 - Q_a)/4$$

Finally, sequenced accumulated probability $P_{Acc,l}$ is calculated from aligned bases that show low reference nucleotide probability (<5%) using Bayesian theorem. Starting from uniform priors probabilities, posterior probabilities are computed by updating the priors with one of the coverage base calibrated probabilities adjusted for alignment quality, $q\mathrm{Cal}_a$, as conditional probabilities. The resulting posterior is then used as a prior for the next coverage base. Given a read RD aligned at location l, the posterior probability of nucleotide n at the location $l+i$ given base at location i, RD_i, $P(n_{l+i}|\mathrm{RD}_i)$, is:

$$P(n_{l+i}|\mathrm{RD}_i) = \frac{P(\mathrm{RD}_i|n) * P(n_{l+i})}{\sum_{m=A}^{T} P(\mathrm{RD}_i|m) * P(m_{l+i})} \quad (1)$$

Where: $P(RD_i|n)$ is the calibrated quality adjusted for alignment of nucleotide n in read RD at location i. or $q\mathrm{Cal}_a$ at read location i.

$P(n_{l+i})$ is the priors probability of nucleotide n at location $l+i$.

The following two hypothetical examples are designed to illustrate the advantage of using qCal values of all four bases versus the mpb in building a consensus sequence and to illustrate the effect of misaligned reads in masking SNPs at shallow coverage.

EXAMPLE 1. Lets assume a location in the reference with an A nucleotide is covered by a three non-crisp bases with $q\text{Cal}_a = q\text{Cal}$ values of $[(-27,-27,1,-1),(-8,-27,-27,8)$ and $(-27,10,-27,-10)]$. Starting from uniform priors, a final posterior probability is calculated by updating the priors three times, Equation (1), using $q\text{Cal}_a$ values of all four bases. The resulting posterior quality $(-48,-40,-42,37)$, Table I, represents a T with enough quality for calling a SNP on the human genome (1 SNP in 1000 bases). In using the qCal values of the mpb only, Table II, the posterior quality will be $(-17,1,-11,-3)$, this represents a very low quality C.

Table I	**Q(A)**	**Q(C)**	**Q(G)**	**Q(T)**	**P(A)%**	**P(C)%**	**P(G)%**	**P(T)%**
coverage 1	−27	−27	1	−1	0.2	0.2	55.7	43.8
coverage 2	−8	−27	−27	8	13.68	0.2	0.2	85.92
coverage 3	−27	10	−27	−10	0.2	90.9	0.2	8.69
Posterior	−48	−40	−42	37	0.00	0.01	0.01	99.98
Misaligned base	−14	5	−6	−27	3.75	75.97	20.08	0.20
Posterior (masked)	**−35**	**−14**	**−22**	**13**	**0.03**	**4**	**0.65**	**95.32**

Table II	**Q(A)**	**Q(C)**	**Q(G)**	**Q(T)**	**P(A)%**	**P(C)%**	**P(G)%**	**P(T)%**
coverage 1	–	–	1	–	14.76	14.76	55.73	14.76
coverage 2	–	–	–	8	4.56	4.56	4.56	86.32
coverage 3	–	10	–	–	3.03	90.91	3.03	3.03
Posterior	−17	1	−11	−3	1.86	**55.86**	7.03	35.25

A single misaligned read adding an extra error base coverage might mask such SNP call, for example, an error base of $(-14,5,-6,-27)$ will result in a posterior probability accumulation not sharp enough for calling a SNP, Table I, last two lines.

EXAMPLE 2. In this example, a location in the reference with an A nucleotide, a total coverage of two, one crisp base and one non-crisp base of $[(-27,-27,-27,25)$ and $(-27,-27,4,-4)]$, Table I, the posterior quality using all qCal values for all four bases is $(-49,-49,-23,23)$, this represents a T with enough quality for calling a SNP on the human genome. However, in using the qCal values of the mpb only, Table II, the posterior quality is $(-30,-30,-20,19)$ that is not enough for calling a SNP.

Table I	**Q(A)**	**Q(C)**	**Q(G)**	**Q(T)**	**P(A)%**	**P(C)%**	**P(G)%**	**P(T)%**
coverage 1	−27	−27	4	−4	0.20	0.20	71.24	28.36
coverage 2	−27	−27	−27	25	0.20	0.20	0.20	99.40
Posterior	−49	−49	−23	23	0.00	0.00	0.50	**99.50**
misaligned base	−14	5	−6	−27	3.75	75.97	20.08	0.20
Posterior (msk)	**−35**	**−22**	**5**	**−5**	**0.03**	**0.61**	**75.66**	**23.70**

Table II	**Q(A)**	**Q(C)**	**Q(G)**	**Q(T)**	**P(A)**	**P(C)**	**P(G)**	**P(T)**
coverage 1	–	–	4	–	9.49	9.49	71.53	9.49
coverage 2	–	–	–	25	0.11	0.11	0.11	99.68
Posterior	−30	−30	−20	19	0.10	0.10	0.99	**98.80**

Again, a single misaligned read might mask such SNP call, the error base of $(-14,5,-6,-27)$ is enough for masking this SNP, Table I, last two lines.

In general, the coverage of one crisp base from a good aligned read provides enough probability for calling a SNP in a genome with one SNP expected up to every 100 bases, and two crisp bases are enough for a SNP call in a genome with expected SNPs ratio up to 1 SNP in every 10 000 bases.

2.2.2 Generating the final SNP scores During the SNPs calling process, a set of SNP evaluation values are generated for each called SNP, these evaluation values are to be used for evaluating the relative confidence in each called SNP with respect to the overall set of predicted SNPs. The evaluation values include the SNP score$_1$ derived from consensus accumulated probability and reference probability as described in Section 2.2, the average SNP location in covering reads, the SNP coverage and the alignment quality of the best aligned read. While these values provide a separate evaluation of each SNP accuracy, a more accurate evaluation is obtained by combining all of these values in a single score in the range [0 … 100], this score correlates with SNPs' prediction accuracy such that SNPs with the highest accuracy have a score of a 100 and the lowest accuracy has a 0 score. Lets first try to understand the main features that are likely to contribute to false SNP calling and what parameters can be used to identify them.

Indels and/or genomic rearrangements between the sequenced sample and the reference are a major source of false positive called SNPs. Genomic rearrangements in the form of different continued sections of the sequenced genome, contigs, are located at different locations on the reference. With respect to our SNPs calling problem, contig edges result in calling false SNPs type C_E. Reads are aligned by first aligning their seeds and then extending these seeds to full reads, lets assume that the contig end (by an indel or genomic rearrangements) just few bases, x, short of some reads. If x is a small number, the last x bases of these reads might, by chance, match the start of the next contig with one base mismatch that will result in calling a false SNP. Ofcourse, if there are several base mismatches, the resulting cluster of dense SNPs will be filtered out by dense SNPs filtering. The average locations of these false SNPs type C_E in covering reads is likely to be either toward the tail or the beginning of reads. Figure 3 shows a sample of true SNP in contig A, and one false SNPs type C_E at the beginning of contig B.

Figure 4 shows the concordance of called SNPs as a function of their average location percentage in the reads at different coverage levels.

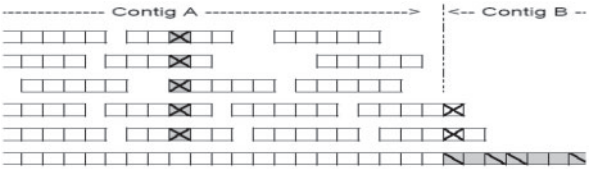

Fig. 3. Coverage of an edge between two contig portions of a reference showing a true SNP marked with a shaded X, and a false C_E SNP marked with X over the first base of a next uncovered (shaded) contig (a deletion in the sample data), the shaded locations in the reference are part of the second contig, those which are crossed with a single line are bases that differ from the sample.

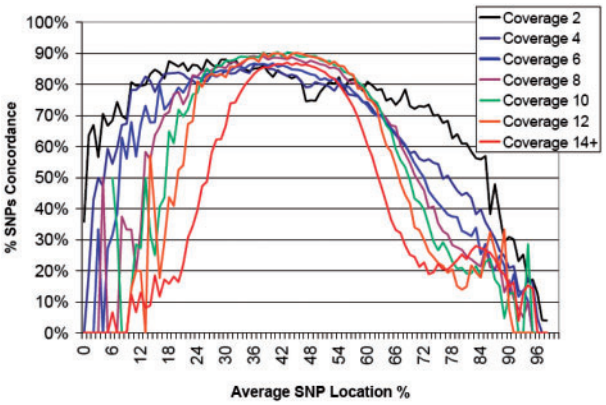

Fig. 4. The concordance of called SNPs as a function of their average location in the reads and coverage. The significance of SNPs average location is higher for higher coverage SNPs.

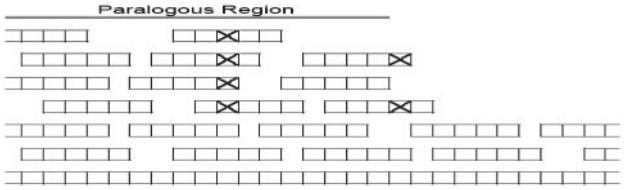

Fig. 5. A paralogous region on the reference with higher coverage than average. One false SNP type P_I is observed inside the paralogous region, and one false SNP type P_E is observed at the edge of this region.

We can see that the significance of this average location is higher as the coverage is higher. Since seed reads are aligned first with up to one mismatch and then extended allowing up to three mismatches in total, the extension portion of aligned reads are, on average, more likely to include false SNP type C_E and those SNPs averaging at the tail of reads are more likely to be inaccurate. A second accuracy score, score_2, is computed from SNPs average percentage location, *bidx*:

```
Seed_Percentage = seed_Size / read_AverageSize
If ( bidx < ((Seed_Percentage / 2) - 10))
   Score2 = bidx
Else If ( bidx > ((Seed_Percentage / 2) + 10))
   Score_2 = 100 * ((Seed_Percentage / 2 + 10) - bidx)
Else
   Score_2 = 100
```

Paralogous regions are regions of the reference, where paralogous reads align resulting in significantly higher coverage. Paralogous regions can generate two types of false SNPs: first, false paralogous SNPs type, P_I, are observed if paralogous reads include some base mismatches which will be observed as SNPs (heterozygous SNPs or if the paralogous reads largely outnumbers the accurately aligned reads, these SNPs might look like homozygous); second type of paralogous false SNPs is observed on the edges of these paralogous regions, false SNPs type P_E. Figure 5 shows a sample paralogous region with a paralogous false SNP type P_I, and one false SNP type P_E at the edge of that region.

While P_I SNPs are likely to have significantly higher than average coverage, P_E SNPs coverage is not as high as P_I's; but they are likely to appear at the edges of reads, and, therefore, the average locations of these SNPs in the covering reads is likely to be toward the edges. The PCR artifact introduced during library preparation can also fabricate false SNPs, some might be observed similar to P_I SNPs with significantly high coverage.

Even the reads alignment quality is incorporated into the consensus base probabilities through $q\text{Cal}_a$ values, when the x coverage increases, some areas of the genome that happen to have low sequence commonness factor are candidate to accumulate enough misaligned reads (perhaps low-quality reads) for calling false SNPs. To improve called SNPs accuracy, the highest alignment quality of each location is used in assessing called SNPs quality.

While the above information provides multiple parameters for measuring the accuracy of called SNPs, it is desirable to integrate this information into a single SNP score that correlate with SNPs accuracy. Slider II uses a heuristic approach that divides the SNP final score range into three different regions based initially on the highest read alignment quality, AQ_{Max}, region three for SNPs with the highest AQ_{Max}, above some threshold value, and region one for those SNPs with the lowest, below some other threshold value. The final SNPs score is computed by combining score_1 and score_2 and normalize the outcome to fit in the appropriate SNP region score range, where region one for scores in the range [67 … 100], region two for scores in [34 … 66] and region three for [0 … 33].

SNPs that are likely to be a result of paralogous alignments (P_I and P_E), or contig edges (C_E) and some PCR artifacts false SNPs are down scored by moving them to a lower region.

A third score that rewords SNPs with higher ratio of reads that are not showing the reference base is computed $\text{score}_3 = \text{cnt}_s/\text{cnt}_t$, where:

cnt_s: is the number of cover bases with the probability of the reference nucleotide less than 5%.

cnt_t: the total number of cover reads.

Ranks are first derived from score_1 and score_2 where each SNP has two ranks, rank_1 from score_1 and rank_2 from score_2. For example, if a SNP score_1 is 32, and this put this SNP at the 78% rank among all other SNPs with respect to score_1, rank_1 for this SNP is 78.

The final SNP score is:

$$\text{Score} = (\text{rank}_2 * \text{lg} + \text{rank}_1 * (2 - \text{lg}) + \text{score}_3)/3$$

where:

```
lg = log10(coverage)
If (lg > 2)
  lg = 2
```

Slider II execution time is about 1.2 to 1.4 times the execution time of Slider (Malhis *et al.*, 2009).

3 RESULTS

We utilized 68 Illumina lanes of human breast cancer whole genome shotgun sequencing paired-end sequence data, representing a total of 906 million reads ranging in size from 36 to 42 bases (Shah *et al.*, 2009). These data were processed using both Slider II Version 1.1 and MAQ Version 0.7.1. Reads were first aligned to the human genome *hg18* resulting in an average coverage of ∼7.5×. For each aligner, SNPs were called and sorted in descending order, using the SNP score for Slider II and the Phred-like consensus quality score provided by MAQ. We used the concordance of the called SNPs with the Ensembl Variation database (version 50) SNPs to compare SNP calling accuracy. We postulate that a higher level of concordance with the variation database, given its large size, and yet relative rarity of SNPs in the reference can only reflect a higher degree of accuracy in SNP calling. Although SNP concordance does not give an exact estimate of SNP calling accuracy, it provides a good relative measure of performance. Concordance provides a lower boundary of this accuracy (and an upper boundary to the novel SNPs in the sample), for example, in a set of n called SNPs, if we have a concordance of $x\%$, this means that:

- Our SNPs calling accuracy is at least $x\%$. It is exactly $x\%$ when the sample data contain only known SNPs.
- The SNPs in our sample data has at most $(1-x)\%$ novel SNPs. The percentage of novel SNPs is exactly $(1-x)\%$ when we have no calling errors.

Therefore, if the sample data contain $y\%$ novel SNPs and $(0 \leq y \leq 1-x)$, novel SNPs accuracy is $z\%$, and $z = y/(1-x)$.

Figure 6 shows the concordance of top-scored SNPs for Slider II without using known SNPs as priors (Slider II A), with known SNPs as priors (Slider II B) and for MAQ at minimum coverage depths of 3, 2 and 1. Whereas for Slider II, the called SNPs concordance is correlated with their scores,; the MAQ concordance dropped significantly for SNPs reported with high quality scores. This phenomenon appears to be contributed by SNPs generated as a result of alignment to paralogous or duplicated regions since MAQ does not have the capability of filtering out such SNPs.

To provide a more comparable analysis, we filtered out SNPs indicating sequence coverage higher than 15 (Figure 7) within the

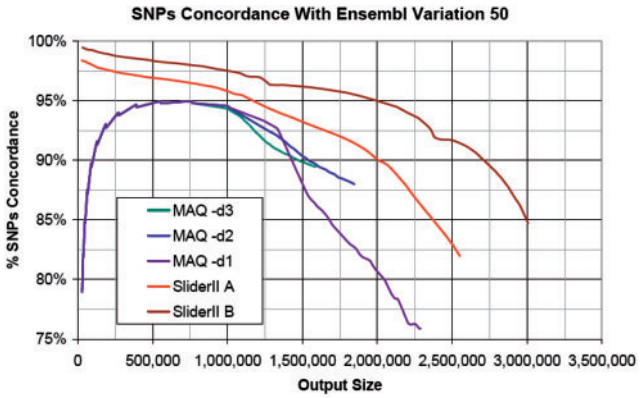

Fig. 6. SNP concordance with the Ensembl Variation database (version 50). SNPs as a function of output size ranked by Phred-like consensus quality for MAQ and SNP score for Slider II. MAQ 1d, MAQ 2d and MAQ 3d indicate a minimum coverage depth 1, 2 and 3, respectively. Slider II A indicated SNPs called by Slider II without using any prior knowledge of known SNPs. Slider II B indicates SNPs called using priors knowledge of known SNPs.

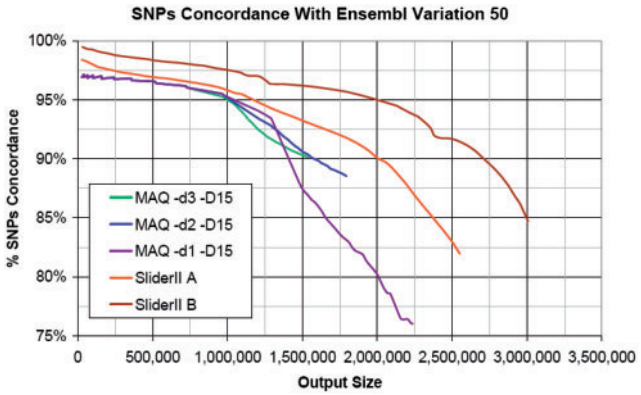

Fig. 7. Labels are as in Figure 6, but MAQ SNPs with sequence coverage higher than 15× are filtered out to reduce the effect of paralogous SNPs.

MAQ output in order to reduce those inaccurately called SNPs at high quality scores. Figure 7 shows that Slider II using the probabilities of all four bases is able to identify more SNPs with higher accuracy than MAQ.

The results in Figures 6 and 7 also show that the usage of polymorphism priors (Slider II B) enables the detection of a larger number of SNPs with substantially higher calling accuracy.

Figure 8 shows the output overlap for Slider II SNPs (not including prior knowledge of known SNPs) with score greater than 25, counting 1 911 839 (or about 75% top scoring SNPs) and all of the MAQ called SNPs with coverage in the range [2 … 15], counting 1 793 659. Given that Slider II and MAQ apply different approaches in alignment, consensus generation, SNPs scoring and filtering, only 70.6% of the SNPs called by Slider II are also called by MAQ, and 75.3% of MAQ SNPs are called by Slider II (using all of the 3 007 401 SNPs called by Slider II utilizing the prior knowledge of known SNPs; this last ratio will become 89.7%).

While MAQ requires a minimum coverage cut off value to identify SNPs and a maximum coverage parameter to filter out paralogous

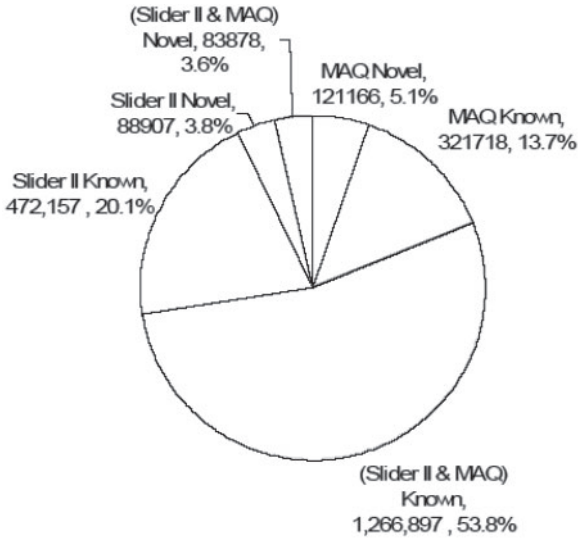

Fig. 8. The called SNPs overlap between Slider II SNPs without using any prior knowledge of known SNPs with score higher than 25 counting 1 911 839 SNPs and MAQ SNPs with coverage of at least 2 and less than 15 counting 1 793 659 SNPs. Known SNPs in the Ensembl Variation database (version 50) are separated from potential novel SNPs that are not in the variation database.

SNPs, Slider II does not rely on such heuristic values, instead utilizes the expected ratio of polymorphism in the reference genome for identifying SNPs at each position. In addition, an automated approach filters out paralogous SNPs without a user instantiated cut off value for maximum coverage.

4 DISCUSSION

In this article, we presented a comprehensive SNPs calling tool from prb values of Illumina short reads, which aligns Illumina short prb reads to a reference genome, utilizes alignment information in calibrating all four prb values and accumulates bases probability using Bayes approach, and finally SNPs are called on nucleotide differences between the accumulated bases probabilities and the reference genome, when our confidence in a consensus accumulated from the reads is higher than our confidence of the used reference. Each called SNP is associated with a single score that ranks its accuracy among the set of called SNPs. SNPs that are likely to be a result of genomic rearrangements are either down scored or removed. As an optional feature, known SNPs of a target genome can be utilized as priors in alignment and SNPs calling, which increases the capability of detecting known SNPs. Results show that Slider II calls more SNPs with higher accuracy than the leading competitive tools; from Figure 7 we see that the concordance with Ensembl known SNPs for Slider II is higher than that of MAQ at every output size, for example, at concordance of 90%, Slider II was able to call more than 2 million SNPs whereas MAQ called about 1.58 million SNPs. We contribute this higher SNPs calling accuracy of Slider II to two factors:

(1) The use of prb values: prb values result in higher alignment quality and more discriminatory probability accumulation outcome, compared to *q*Cal value of the mpb, which enabled

Slider II to call more SNPs at lower X coverage with higher quality.

(2) Higher filtering capabilities: in filtering out SNPs that are likely been called as a result of indels and genomic rearrangements, Slider II implemented two new concepts that resulted in a lower percentage of false positive called SNPs: first, when the average SNP location in covering reads is toward the edges, the SNP is more likely to be a false positive that is called as a result of indels and genomic rearrangements (as described in Section 2.2.2); second, the concept of LQB is used to improve dense SNPs filtering (as described in Section 2.2).

While higher SNPs calling capabilities resulted from utilizing prb values in alignment and probability, accumulation can be more essential at lower coverage areas and is noticeable at the right side of charts in Figures 6 and 7, higher filtering capability of falsely called SNPs utilizing the average SNP location gets more effective as the coverage increases (Figure 4), which can be detected at the left side of charts in Figures 6 and 7.

This algorithmic approach demonstrates that utilizing the probability of all four possible nucleotides provide an important factor in improving SNPs calling outcome. While many researchers and software developers choose to ignore prb data due to its larger files size, Slider II process prb from '.gz' files which are compressed with high ratio (7–10× smaller than the original prb source) due to the high repeat in prb values.

ACKNOWLEDGEMENTS

S.J.M.J. is a senior scholar of the Michael Smith Foundation for Health Research. We thank Stephen Montgomery of the Sanger Institute, and Sohrab Shah and Yaron Butterfield of the BC Cancer Agency for their helpful comment on the script.

Funding: IBM Canada Ltd. (in part).

Conflict of Interest: none declared.

REFERENCES

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*., **10**, R25.

Li,H. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*., **18**, 1851–1858.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,R. *et al*. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Lin,H. *et al*. (2008) ZOOM! Zillions of oligos mapped. *Bioinformatics*, **24**, 2431–2437.

Malhis,N. *et al*. (2009) Slider - maximum use of probability information for alignment of short sequence reads and SNP detection. *Bioinformatics*, **25**, 6–13.

Marth,G.T. *et al*. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet*., **23**, 452–456.

Schatz,M.C. *et al*. (2007) High-throughput sequence alignment using Graphics Processing Units. *BMC Bioinformatics*, **8**, 474.

Shah,S. *et al*. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.

Smith,A.D. *et al*. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, **9**, 128.