Variant calling comparison

Denis C. Bauer, Queensland Brain Institute, Australia

July 12, 2011

1 Abstract

This work aims at addressing the question whether the new CASAVA1.8, which boasts improvements such as local realigments of reads, is at par with the well accepted pipeline of BWA mapping, duplicate removal, local realignment, re-calibration and variant calling using GATK. We therefore compare the two methods on chromosome 21 of a Yoruba trio and compare the results to the genotype identified by the 1000 genomes project.

We find that the mapping performance is the same for CASAVA1.8 and the academic pipeline, resulting in a mean coverage of about 22. CASAVA1.8 and GATK both call about 70.000 SNPs per individual of which 80% overlap between CASAVA1.8, GATK and the 1000 genomes project. This stands in contrast to the indel calling performance were CASAVA1.8 calls about 12000 indels while GATK calls 16000. Furthermore, CASAVA1.8 has a higher Mendelian error rate and frequently more than one alternative allele per locus indicating a non-optimal alignment.

We conclude that CASAVA1.8 has come a long way and can be considered a mature SNP calling approach. However, CASAVA1.8 does not deliver the same quality in the indel calling set compared to the newly incorporated Dindel-algorithm of GATK. It hence remains the best practice to use CASAVA1.8 for producing fastq files and switch at this stage to the academic tools for mapping, alignment improvement and variant calling.

2 Results

The following section outlines the results for alignment, SNP and indel calling results.

2.1 Alignment

In this section we investigate whether there are differences between the mapping ability of the two methods based on chromosome 21 of the Yoruba trio. The mapping statistics for CASAVA1.8 is shown in Tab 1 (first three lines).

To re-generate raw read data to be mapped by BWA, the reads uniquely assigned to chromosome 21 by CASAVA1.8 are supplemented with all unmapped, duplicated and low quality reads that could not be aligned to another chromosome by CASAVA1.8 (see section 5.1). This set of reads are then aligned against the whole human reference genome using BWA. Furthermore, the reads undergo a local realignment using GATK, which aims at minimising the overall number of mismatches by taking all reads mapping at a specific location into account when assessing a shift. Finally, the base quality scores are re-calibrated using GATK, which again uses the information of all local reads as well as known variants to give a better estimate of the base call confidence, which becomes important for the variant calling. The mapping statistics of this corrected alignment is shown in Tab 1 (last three lines).

While BWA processes all reads that comes of

Method	*Sample*	*accepted reads*	*duplicates*	*read pairs mapped*	*read pairs mapped in region*	*mean coverage*
CAS	NA18506	12002158 (14%)	402845 (0.4%)	11196106 (13%)	11196106 (13%)	22.90
CAS	NA18507	11867906 (14%)	195484 (0.2%)	11069736 (13%)	11069736 (13%)	23.18
CAS	NA18508	11751535 (15%)	330736 (0.4%)	11175066 (15%)	11175066 (15%)	22.80
BWA	NA18506	84008356 (100%)	13744921 (16%)	52703392 (62%)	11583036 (14%)	22.15
BWA	NA18507	84734652 (100%)	11258750 (13%)	52073470 (62%)	11514002 (14%)	22.35
BWA	NA18508	76348136 (100%)	10271085 (13%)	48414920 (63%)	11429300 (15%)	22.00

Table 1: **Alignment statistics for chromosome 21.** The table shows for CASAVA1.8 (top) and BWA (bottom) and each of the three HapMap samples the number of reads that were found acceptable by the method (third column) and the final yield of reads that mapped and were properly paired on chr21 (sixth column).

the sequencer, the stringent quality filter in CASAVA1.8 deems only 14% usable for further processing. This is reflected in a very low number of duplicates (0.3%) that are found in the CASAVA1.8 mapped reads compared to about 14% in the BWA mapped reads. Subsequently, CASAVA1.8 is only able to map both read pairs for 14% of the total reads, while BWA mapped 62%. However, a close inspection reveals that of the BWA mapped reads, only 14% indeed map to chromosome 21, resulting in a equivalent mean coverage achieved by both methods of 23 for CASAVA1.8 and 22 for BWA, respectively.

Fig 1 shows an equivalent coverage over the whole chromosome 21 for both methods. The longer arm of chromosome 21 receives a even coverage, whereas the shorter arm has large coverage spikes at its only gene-rich region and an absent of mapped reads otherwise. This predominantly even coverage is reflected in that about 70% of the bases have a higher coverage than 10, about 68% a coverage of 20 and only about 7% a coverage above 50, for both methods.

Keeping in mind that BWA allows multi-mappers and was given the reads assigned to chromosome 21 by CASAVA1.8 as well as all unmapped reads, the high percentage of reads mapping outside chromosome 21 is not surprising. We can hence summarise that both CASAVA1.8 and BWA return an alignment of similar quality.

2.2 SNPs

This section asks the question whether the SNP call set made on the realigned and re-calibrated BWA mapping using GATK's Bayesian model returns better results than using the new CASAVA1.8, which only uses local realignment to improve the initial read mapping and features a simpler SNP calling model. To answer this question we contrast the SNP calls made by CASAVA1.8 and GATK to the set accepted by the 1000 genomes project for two of the three HapMap samples.

The overview statistics from the SNP call sets are listed in Tab 2. Both method produce a comparable number of SNPs with a comparable statistics for concordance with SNPdb (v132), known to novel SNP rate, and Ti/Tv rate, which is also in agreement with the 1000 genomes project SNP call set. The overlap between the three call set methods is exemplified for NA180507 in Fig 2, which shows that in the majority of their SNP calls the three methods agree, only 12% (8991/74592) of all GATK and 9% (7198/73708) of all CASAVA1.8 calls are not in agreement with of the other methods.

GATK has a tendency to call location to be heterozygote (higher het/hom ratio), which is likely due to its primary application area in cancer genetics, where a true SNP in the cancer tissue is diluted down by the contamination of normal tissue in the sequenced sample. This higher sensitivity might cause erroneous SNP calls in genotyping applications, where the allele imbalance is not expected to be extreme. Indeed the more conservative CASAVA1.8 SNP calls contain only one Mendelian error (where both methods make the same genotype assessment), whereas GATK produces 15 additional errors, which are all due to hom/het disagreements between parents and child. However, a closer inspection reveals that the higher

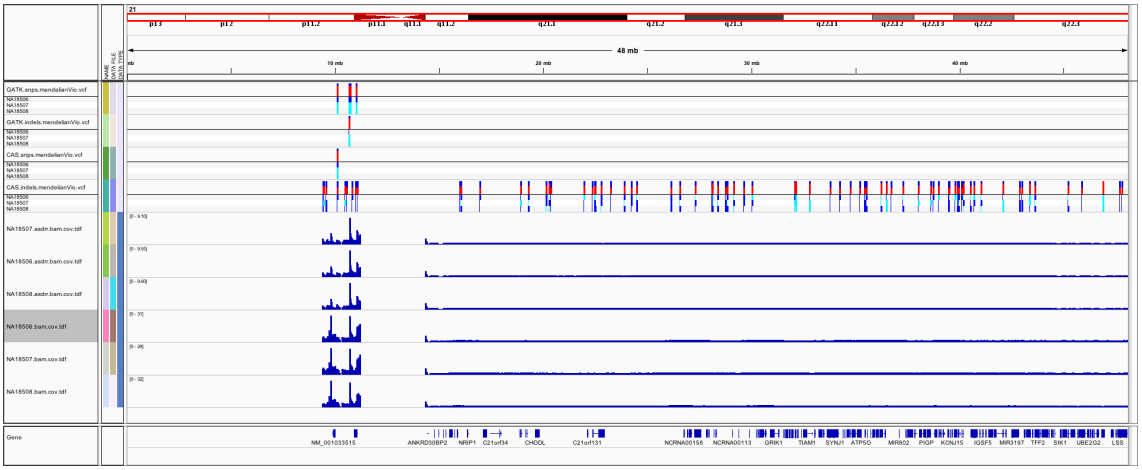

Figure 1: **Coverage and Mendelian error location for both approaches.** The lower half of the figure shows the coverage for the three HapMap individuals for reads mapped using BWA and realigment (upper three lines) and CASAVA1.8 (lower three lines). The top half of the figure shows the location of Mendelian errors for GATK called SNPs (first line), indels (second), CASAVA1.8 called SNPs (third) and indels (fourth).

het/hom ratio is not the explanation for the observed Mendelian errors, as shown in Fig 3, where a slightly higher allele frequency of the reference allele in the child (A 17%, C 82%) caused the SNP to be heterozygote (A/C), while parents remain homozygote (5%A,95%C and 5%A,92%C). In contrast, CASAVA1.8 calls all three individuals heterozygote. This observation is also the case for five of the 15 Mendelian error locations. For the remaining locations CASAVA1.8 was not able to make call sets for one or more individuals and hence no error is recorded. This is exemplified in the right image of Fig. 3, where the child is called heterozygote (C 90%, T 10%), while parents remain homozygote for the alternative (94%C,6%T and 94%C,6%T).

Looking back at the coverage chart (Fig. 1) offers the explanation for GATK's erroneous calls and CASAVA1.8 lack of calls: all Mendelian errors made by both methods are located in the area with large fluctuations in coverage and are hence likely to be artefacts.

So in summary, the SNP call set made by CASAVA1.8 is equivalent in quality compared to the set made by GATK from the re-calibrated and realigned BWA mapping and both methods have problems in the coverage fluctuating area.

2.3 Indels

In the final section we want to investigate, whether CASAVA1.8 is also on par with GATK, which now incorporates Dindel, in terms of its indel calling ability. Unfortunately, the 1000 genomes project have not released individual indel calls for the Yoruba trio so we cannot assess the overlap. Tab 3 lists a similar statistics for indels as reported above for SNPs. We can see that CASAVA1.8 is much more conservative in calling indels than GATK is, which is rewarded with a slightly higher concordance rate with SNPdb for the known indels. As shown in Fig 2, 84% of the CASAVA1.8 indel calls are also present in the GATK set.

However, CASAVA1.8 predicts a higher percentage of novel indels (38% compared to 33% by GATK). In doing so CASAVA1.8 makes more Mendelian errors compared to GATK. While GATK's Mendelian error, which is not shared with CASAVA1.8, is again located in the area with coverage spikes, CASAVA1.8's errors are scattered throughout the chromosome (Fig 1). An example of these 77 errors is shown in Fig. 4, where the local realignment made by GATK was more successful than the one made by CASAVA1.8 in finding the true position of the reads. However, the fact that the indel locations with Mendelian vio-

Method	*Sample*	*total SNPs*	*known SNPs*	*conc. dbSNP*	*known Ti/Tv*	*novel SNPs*	*novel Ti/Tv*	*overlap indiv.*	*Mend.* errors	*het/hom* ratio
CAS	all	104967	91081	99.79	2.09	13886	1.69	29792	1	2.03
CAS	NA180506	71986	64348	99.84	2.11	7638	1.65			
CAS	NA180507	72709	65832	99.84	2.08	6877	1.63			
CAS	NA180508	72256	63407	99.84	2.09	8849	1.69			
GATK	all	102795	92393	99.75	2.07	10402	1.61	37228	16	2.36
GATK	NA180506	74437	67916	99.75	2.06	6521	1.49			
GATK	NA180507	74592	68932	99.76	2.04	5660	1.52			
GATK	NA180508	73612	66321	99.74	2.05	7291	1.54			
1kg	NA180507	60892	58794	99.92	2.19	2098	1.99			
1kg	NA180508	58904	56546	99.93	2.21	2358	1.88			

Table 2: **SNP statistics for chromosome 21.** The table shows for CASAVA1.8 (top) and GATK (middle) and each of the three HapMap samples the number of SNPs predicted for chr21 along with their general statistics. This is contrasted to the SNPs that are found by the 1000 genomes project for the samples and the chromosome (bottom).

Method	*Sample*	*total Indels*	*known Indels*	*concordance dbSNP*	*known Ti/Tv*	*novel Indels*	*novel Ti/Tv*	*overlap individuals*	*Mendelian* errors
CAS	all	18095	11202	85.55	0.75	6893	1.76	3864	77
CAS	NA180506	12120	8295	90.12	0.73	3825	1.76		
CAS	NA180507	12094	8463	90.71	0.74	3631	1.76		
CAS	NA180508	12220	8215	90.34	0.72	4005	1.75		
GATK	all	20068	13364	86.03	0.77	6704	1.76	6161	1
GATK	NA180506	16269	11306	85.21	0.76	4963	1.76		
GATK	NA180507	15870	11280	86.00	0.78	4590	1.76		
GATK	NA180508	15771	10950	85.60	0.78	4821	1.76		

Table 3: **Indel statistics for chromosome 21.** The table shows for CASAVA1.8 (top) and GATK (bottom) and each of the three HapMap samples the number of indels predicted for chr21 along with their general statistics.

lations had on average 1.99 alternative alleles (average 1.10 overall in CASAVA1.8 calls and 1.00 in GATK calls) suggests that, the second realignment made by the Dindel-based approach in GATK is the improving factor for indel calls.

To summarise, while indels called by CASAVA1.8 agree mostly with the set made by GATK, the ones that differ have a high likelihood of being erroneous.

3 Runtime analysis

To be done at a later stage.

4 Conclusion

In conclusion, we aimed as presenting a complete comparison for variant calling between the new CASAVA1.8 and the well accepted pipeline of BWA mapping, duplicate removal, local realignment, re-calibration and variant calling using GATK. Using chromosome 21 of a Yoruba trio, we found that the CASAVA1.8 produces a similar mapping as the academic pipeline. Similarly, CASAVA1.8 and GATK produce a comparable SNP set of about 70.000 SNPs per individual with a concordance rate over 99% and an overlap of 80% between CASAVA1.8, GATK and the 1000 genomes project. This stands in contrast to the indel calling performance were CASAVA1.8 calls

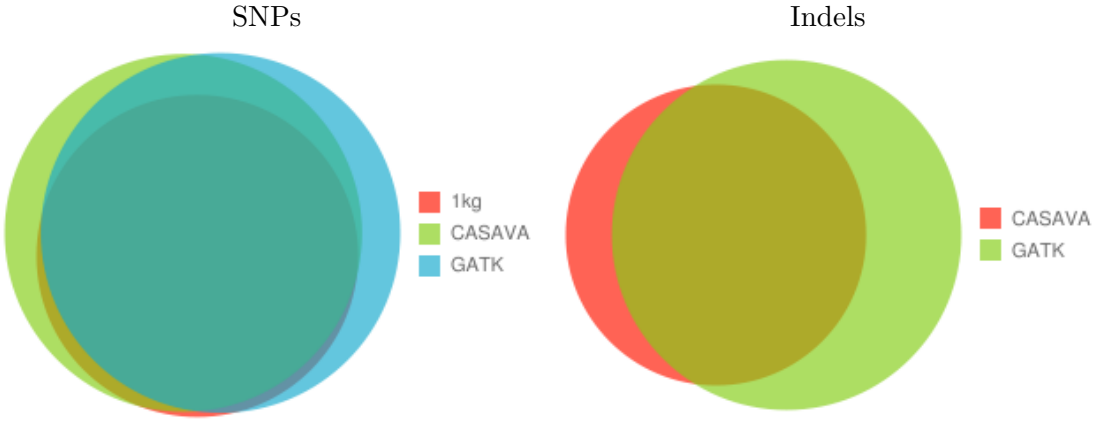

Figure 2: **Overlap between the different call sets** The figure on the left shows the agreement in SNP call sets between the three methods. The right figure shows the agreement in indel calls between GATK and CASAVA1.8.

about 12000 indels of which 38% are novel while GATK calls 16000 with 33% novels. Furthermore, CASAVA1.8 has a higher Mendelian error rate and calls frequently for every individual an alternative allele, which is indicative of an erroneous alignment. This suggests that a Dindel-based approach, performing a local-realignemt at every putative indel locus, is necessary for identifying indels with high confidence.

While CASAVA1.8 has become a mature SNP calling approach it does not deliver the same quality for indel calling. GATK, specifically with the Dindel-algorithm, remains method of choice for indel calling. Given that the CASAVA1.8 alignment output needs to be transformed (see section 5.3) before the variant call can be performed using GATK, the best practice is still to use CASAVA1.8 for producing fastq files and switch at this stage to the academic tools for mapping, alignment improvement and variant calling.

5 Methods

5.1 DNA sequencing data

HapMap data of a Yoruba trio of son (NA180506), father (NA180507), and mother (NA180508) was sequenced mapped and variant called by illumina using CASAVA. The sequencing was done on a HiSeq using the third generation flowcell and chemistry and 100bp paired end reads. To prepare a fastq file for the BWA and GATK analysis all reads mapping to chromosome 21 as well as all reads that were labelled "noMatch", "nonUnique", "qcFail" and "mixed" were combined in a bam file using MergeSamFiles and then converted to fastq with SamToFastq, both from Picard tools.

5.2 CASAVA1.8

The initial reads were mapped and variants were called by illumina. To make the resulting data conform to bam-standards, as accepted by analysis tools such as GATK, some adjustments had to be made.

5.3 Adjust bam file

To make the bam indexing conform to a b37 reference fasta file the following steps are necessary 1) create header with desired name in the same order the old header was in 2) reheader with Samtools (0.1.11 r851) to re-names the chromosomes 3) use Picard tools to force a re-sort with the order of the reference file (this sorts the reads and rewrites the header) 4) use Picard to also add a read-group to each read After this the mean coverage could be calculated with GATK (1.0.5917) and a sliding

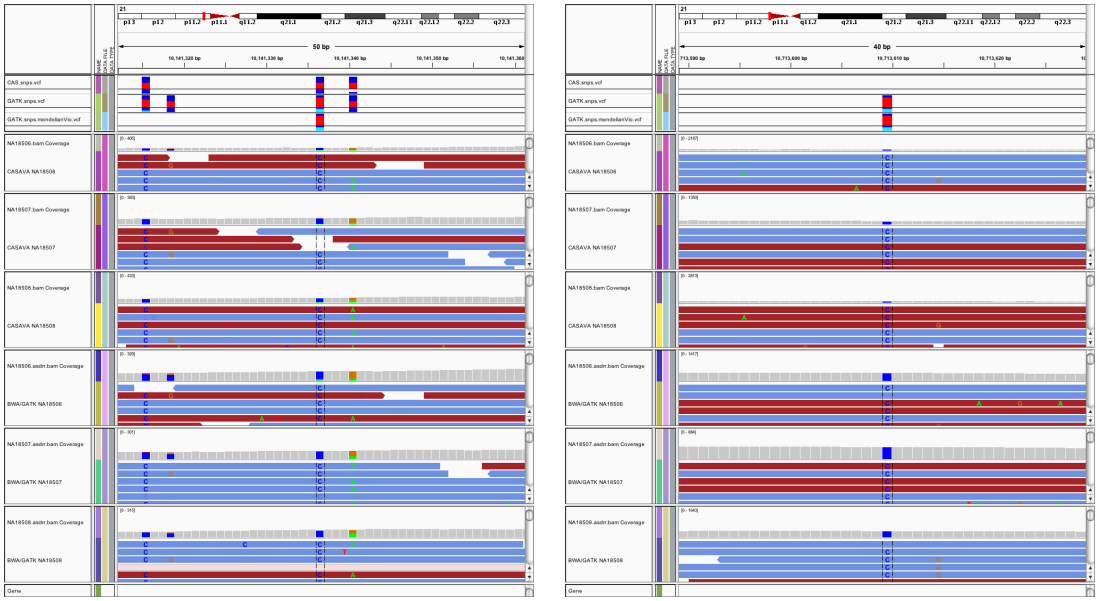

Figure 3: **Mendelian errors made by GATK** The figure shows IGV screenshots from two different regions on chromosome 21, where GATK has a Mendelian error. The figures show SNPs called by CASAVA1.8 (first line), SNPs called by GATK (second), Mendelian errors (third), as well as the reads mapped to this location from the three individuals by CASAVA1.8 (fourth-six) and realigned BWA hits (seventh-nineth).

window coverage track generated using IGVTOOLS (1.5.11).

5.3.1 Adjust vcf file

The vcf format calls for the reference allele to start before the variant. Since CASAVA1.8 calls indels even though the start and end position of this genomic variation might be uncertain, the resulting format can not be conform with the vcf specifications. Here all variants that do not match the specifications are filtered out, these were all "—BKPT=RIGHT" and some "BKPT=LEFT" variants, in total about 1% (200/12593) per sample.

5.4 Mapping with BWA

BWA version 0.5.8c (r1536) was used with the standard settings to map. The resulting sam file was sorted and converted to bam with SAMTOOLS (0.1.11 r851). Duplicates are identified with PICARD and the mean coverage is calculated with GATK (1.0.5917) and a sliding window coverage track generated using IGVTOOLS (1.5.11). The final summary is generated with SAMTOOLS flagstat.

Local realignment was done with GATK RealignerTargetCreator and IndelRealigner and the base quality score was re-calibrated with CountCovariates and TableRecalibration after both steps the completeness of the file was checked with SAMTOOLS flagstat.

5.5 Variant calling

Variants were called with GATK and the parameter setting suggested by the best practice page for a coverage above 10 (last modified on 27 June 2011, at 16:51). UnifiedGenotyper was applied to the realigned and re-calibrated bam files of all three individuals simultaneously with no change to the Bayesian model parameter, but "-glm BOTH", a

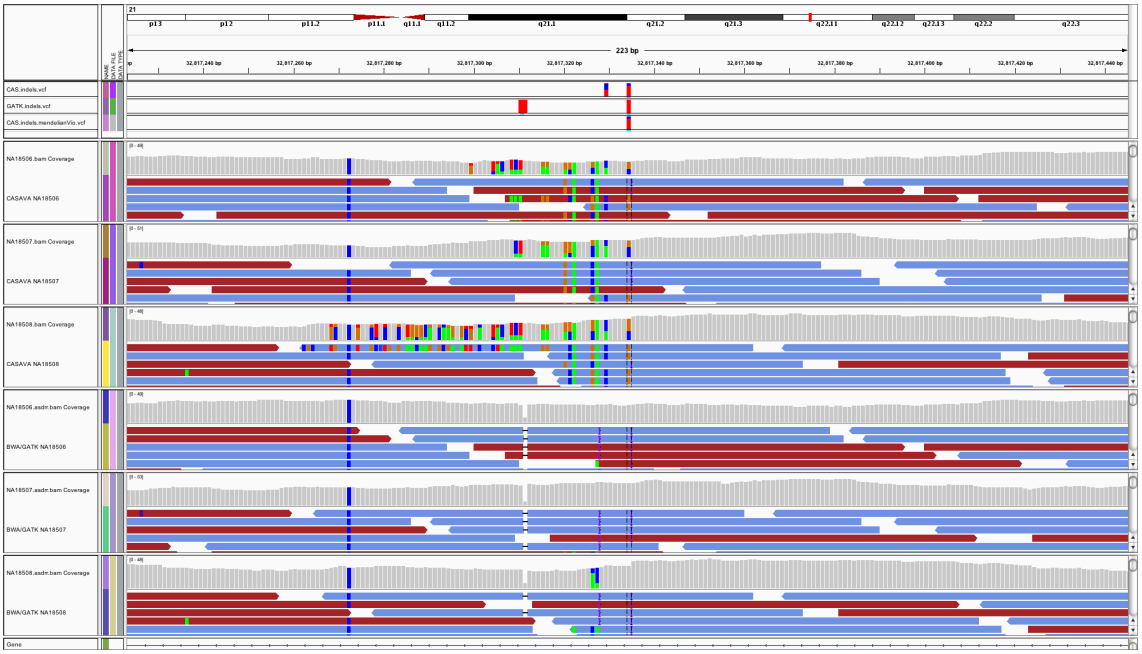

Figure 4: **Mendelian errors made by CASAVA1.8** The figure shows an IGV screenshot from a location where CASAVA1.8 has a Mendelian error. The figures show indels called by CASAVA1.8 (first line), indels called by GATK (second), Mendelian errors (third), as well as the reads mapped to this location from the three individuals by CASAVA1.8 (fourth-six) and realigned BWA hits (seventh-nineth).

standard call confidence of 30 and a standard emit confidence of 10.

The resulting SNPs were extracted using SelectVariants and subjected hard filtering using VariantFiltration such that SNPs around indels are excluded, SNPs cluster within a window of 10 are removed and fulfilling $MQ0 >= 4\&\&((MQ0/(1.0 * DP)) > 0.1)$.

Similarly, indels were extracted from the variant call set and subjected to VariantFiltration and candidates with $MQ0 >= 4\&\&((MQ0/(1.0 * DP)) > 0.1)$, a strand bias ≥ -1.0, and a quality lower than 10 were excluded.