Novoalign Version 3
Release Notice, March 2013

This release is a major upgrade to Novoalign, extending the maximum read length to 950bp and allowing for alignments with longer indels especially in Single end reads. The basic algorithm hasn't changed and sensitivity & specificity have improved especially for longer reads.

Internally the code changes include:

1. Increased Alignment Score Range
2. Increased Indel Length in Single end Reads
3. Changes to default alignment threshold calculation
4. Restructuring for Performance
5. Identical Results on Rerun

Contents

Increased Alignment Score Range

The first stage of alignment involves use of SIMD (Single Instruction Multiple Data) code to quickly establish an alignment score for a read at seeded location. This code is limited to an alignment score of 255 and to around 8 mismatches when using scoring scheme in Novoalign version 2. In version 3 we reduce the scale of the scores by a factor of 2-6 depending on read length. When reduced by a factor of 2 a mismatch at a high quality base scores 15, this means we can have up to 16 mismatches before hitting the score limit of 255. The scoring factor is chosen based on the maximum alignment threshold set for the read. This only affects the scoring in the initial SIMD routines, the best alignments have their score recalculated using conventional code at full scale before reporting.

This change allows us to confidently increase the maximum length reads that can be aligned by allowing for more mismatches and longer indels in an alignment.

Increased Indel Length in Single end Reads

A banded Needleman-Wunsch algorithm is used to align single end reads. The maximum width of the band in version 2 was 32bp allowing alignment of reads with indels up to a maximum of 15bp. The maximum width of the band has been increased to allow alignments with indels up to 60% of read length.

Effects of soft-clipping alignments back to the best local alignment causes detection rates to start dropping for indels longer than about 33% of read length when using pileup type consensus callers. Other indel callers, such as PINDEL and DINDEL, may well be able to call these longer indels from the Novoalign alignments.

2 shows results from testing different aligners using simulated mutations (dwgsim with corrections) and reads on C.Elegans genome, 150bp single reads at 60X read depth. 3 shows similar results using paired end reads. Novoalign V3 is the only aligner that performs equally well on single end reads and paired end reads.

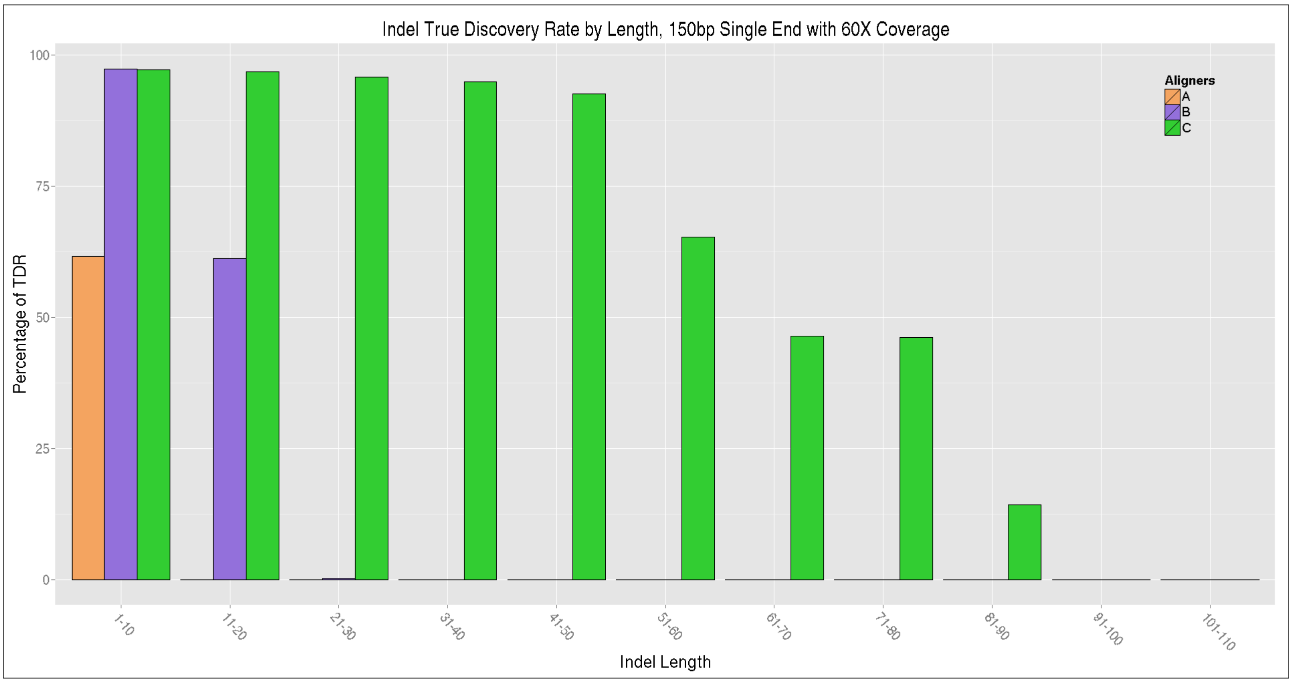

Illustration 1: True discovery rate for indels using different aligners with 60X cover of 150bp single end reads: A BWA; B Novoalign V2 & C Novoalign V3. Pipeline steps: Align, Samtools mpileup -F 0.07 -m 1.

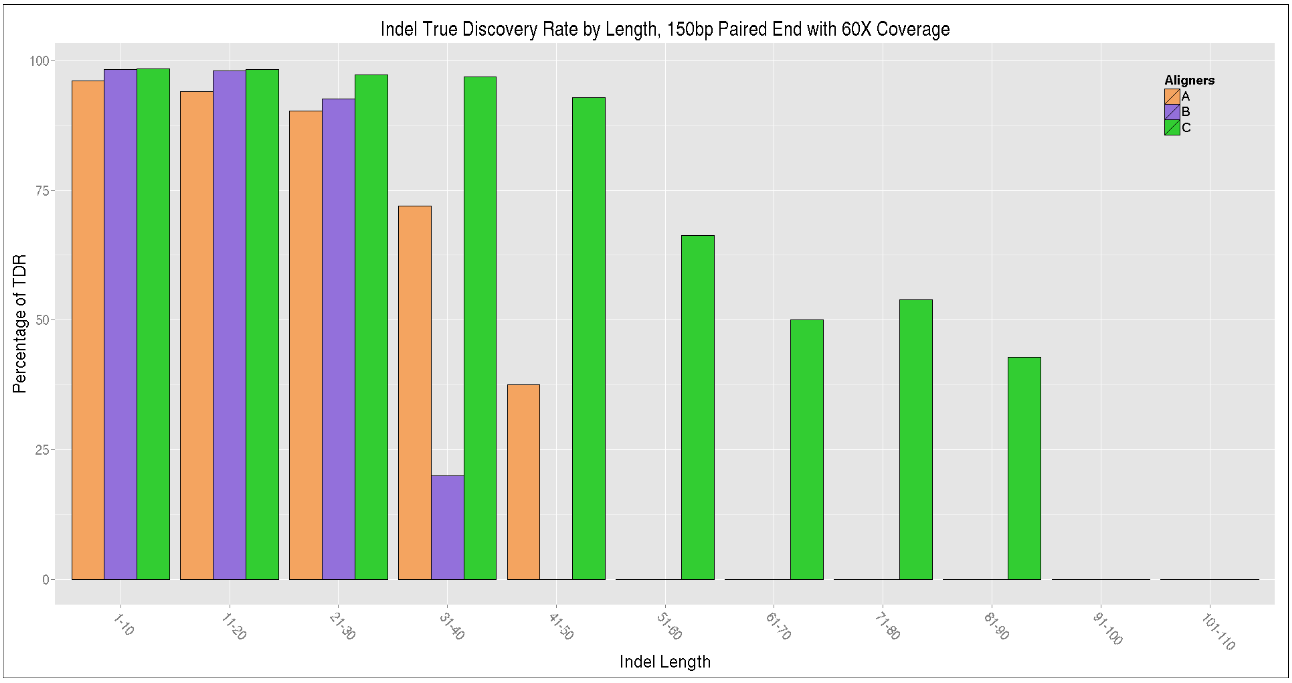

Illustration 2: True discovery rate for indels using different aligners with 60X cover of 150bp paired end reads: A BWA; B Novoalign V2 & C Novoalign V3. Pipeline steps: Align, Samtools mpileup -F 0.07 -m 1.

Aligners and consensus callers were usually run with default options however we needed to adjust samtools mpileup options to improve indel calling. In our testing we found an issue with samtools mpileup that results in wrong indel being called even though most alignments were supporting the correct indel. This happens if a few reads have sequencing errors that result in slightly different alignments and indel positions for some of the reads. Samtools mpileup will emit a cluster of indel calls all with the same quality even though the correct alignments far out number the incorrect alignments, vcfutils then selects one of these indel calls, usually the wrong one. This problem can be avoided using -m and -F options to mpileup. Petr Danecek, Sanger Institute suggested we used -m 3 -F 0.2. Further testing showed that -m1 -F 0.07 produced the best indel TDR rates for this data.

We also ran tests using several different consensus calling programs, here we chose Bambino, Freebayes, Samtools and both the UnifiedGenotyper and HaplotypeCaller in GATK.4 Shows indel TDR by indel length for the different consensus calling programs. Samtools showed best performance except for indels over 70bp where GATK HaplotypeCaller showed improved results.

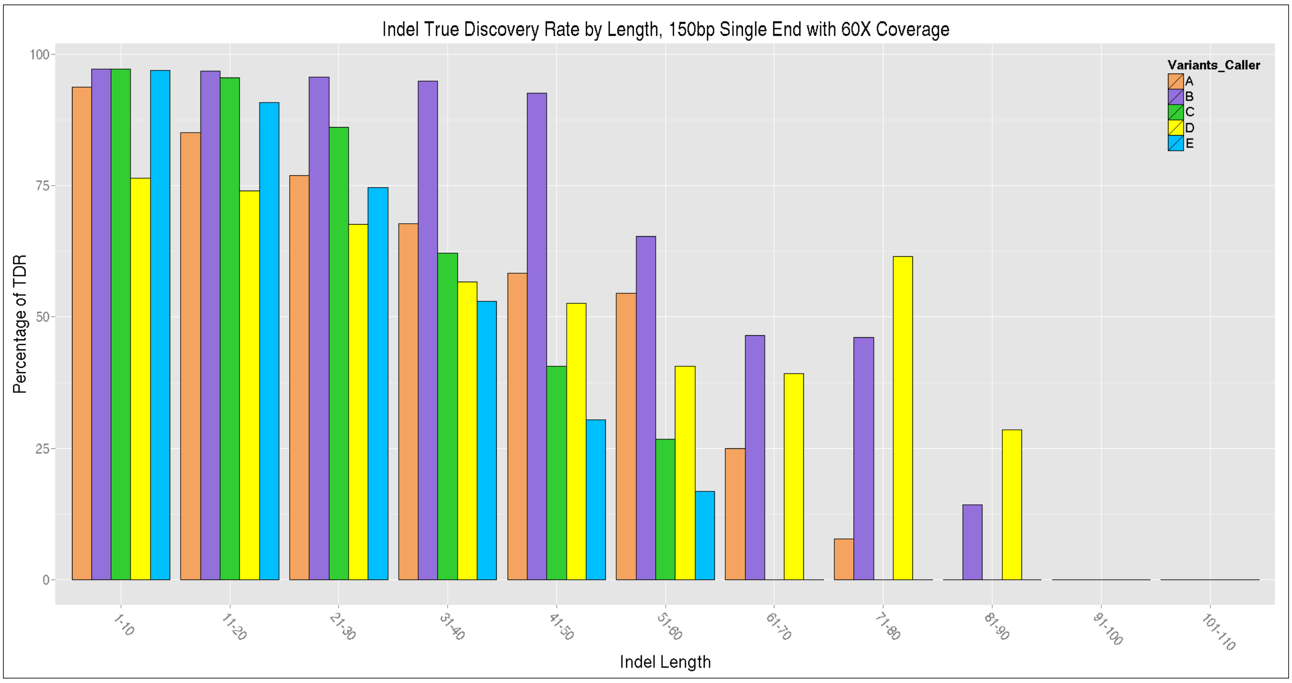

Illustration 3: Indel TDR using different consensus calling programs. A) Bambino; B) samtools mpileup -m 1 –F 0.07; C) GATK UnifiedGenotyper; D) GATK HaplotypeCaller;E) FreeBayes. Pipeline:- Novoalign, Sort, Consensus Calling.

Most "Best Practice" guides suggest using indel realignment before consensus calling. To evaluate whether this was beneficial we tested samtools mpileup with and without realigning. Checking alignments in IGV showed some definite improvements in alignments that had sequencing errors in or near an indel however it made only a small difference to the TDR for indels.

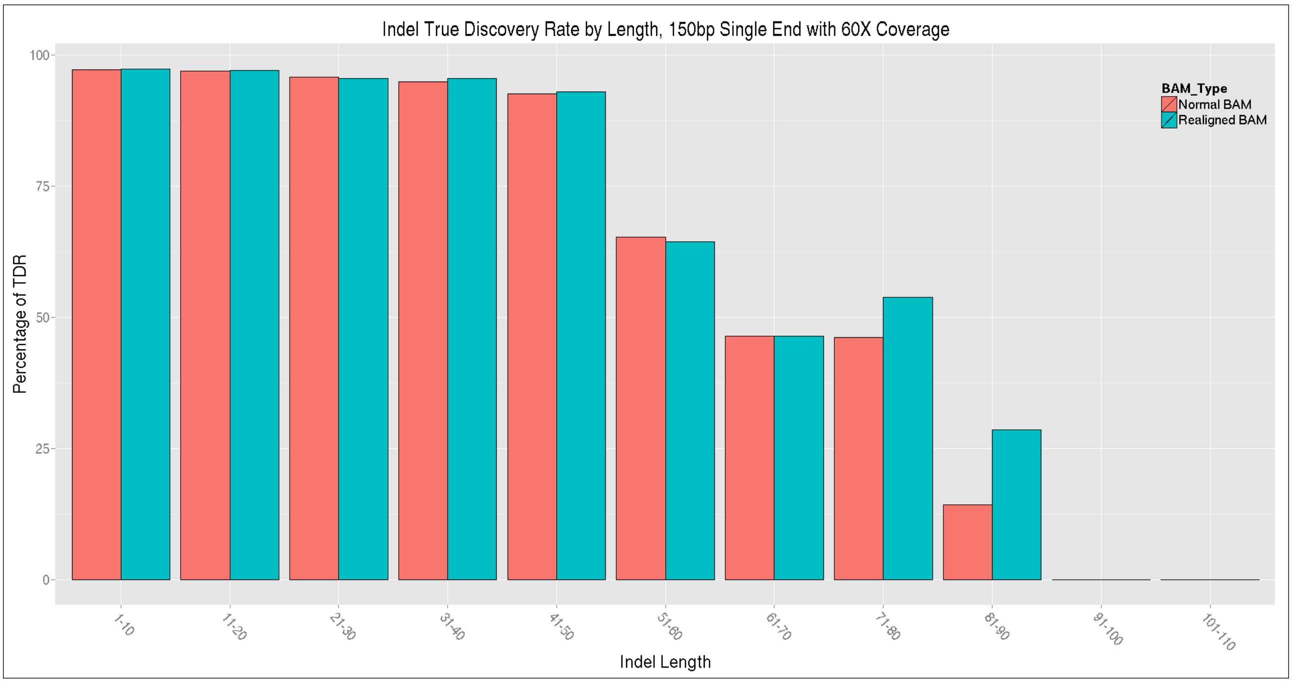

Illustration 4: Effect of GATK realigner on Indel TDR with Novoalign V3 and Samtools mpileup -m 1 -F 0.07. Realigning produced modest improvements.

5 Shows Indel TDR rate for 100bp and 150bp single end reads. Detection rate starts to drop off at 33% of read length. This is mostly due to the fact that indels near the ends of a read are likely to get

soft clipped from the alignment so there are fewer reads with the indel.

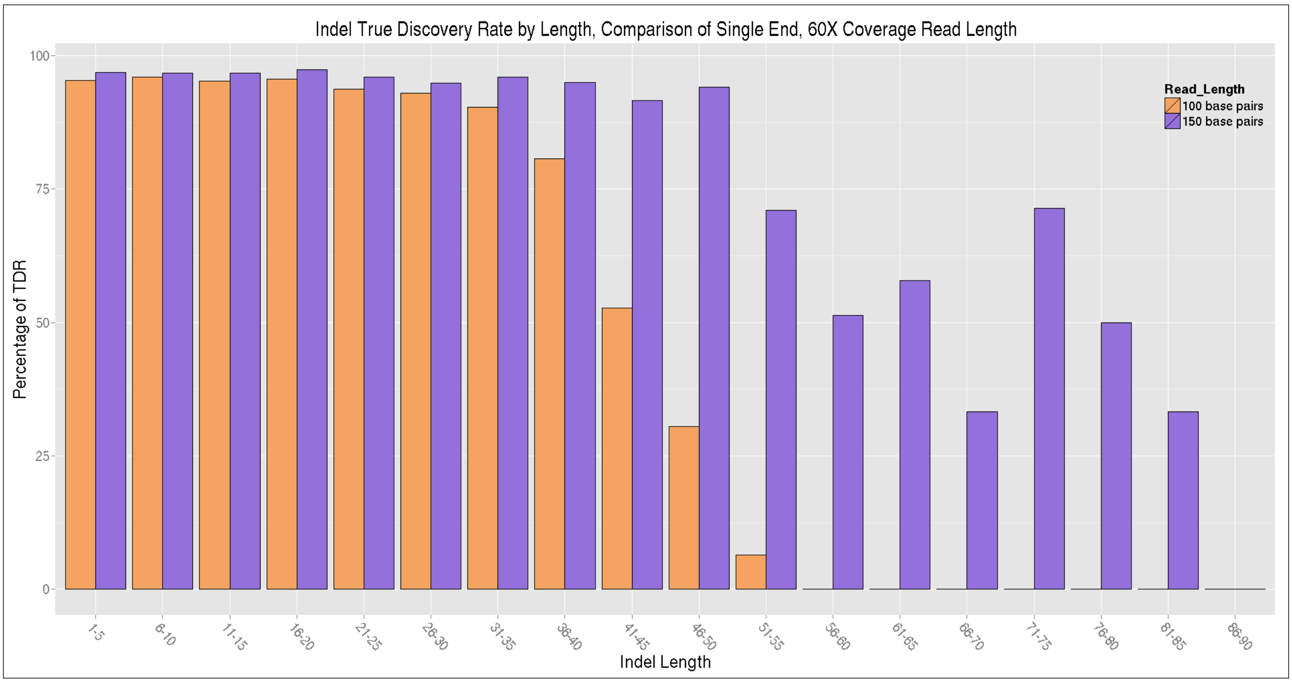

Illustration 5: Comparison of indel TDR for 100bp and 150bp single end reads. TDR rate is high for indels up to 33% of read length.

Default Alignment Threshold

In version 2 the default alignment threshold was calculated as min(254 , 5*(R_l – ($\log_4$(N) + 5))) where R_l is the effective read length (discounting low quality bases) and N is the length of the genome.

The default in version 3 is now min(1524, 4.5(R_l -($\log_4$(N)+1)))

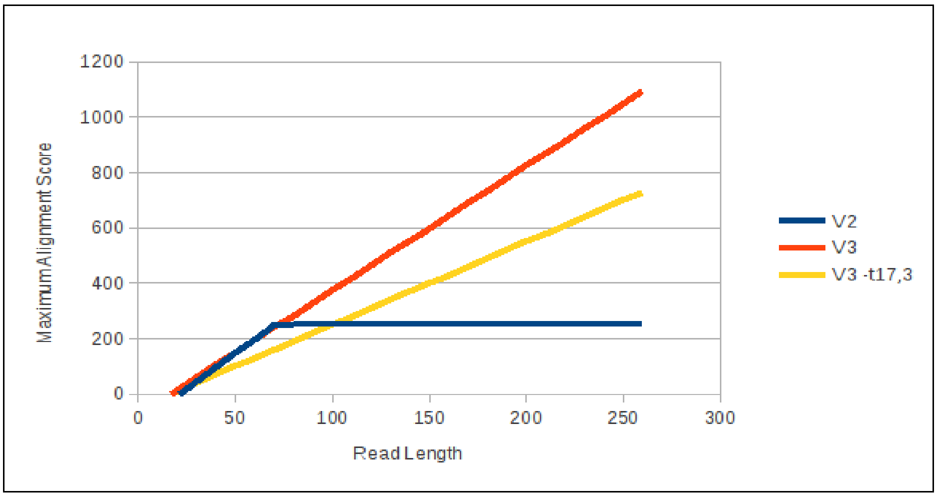

Illustration 6: Maximum alignment scores for different length reads. Version 2 is limited by maximum score of 254. Using -t 17,3 in V3 gives similar limit to V2 on 100bp reads.

The default in version 3 is probably higher than necessary and possibly only necessary for alignment of highly divergent genomes. In a 100bp read it allows 12 mismatches or 88% identity. For really long reads the limit of 1500 is reached which would allow 50 mismatches in say a 950bp read.

We suggest overriding the default with lower setting such as -t 17,3 or similar which should have the same sensitivity on 100bp reads as version 2.

Restructuring for Performance

Allowing longer indels in alignments and higher alignment score threshold had a considerable negative effect on performance, to offset this we went through a major performance tuning step that restructured code and data tables to move computation of detailed alignments to the reporting functions. This means that for any alignment not being reported we have only done the first step of the alignment process. In previous versions when reads aligned to repeats most of the alignment would have had the detailed alignment step done.

This change has offset the performance costs of the indel changes so that run time of version 2 and version 3 are similar. It has also reduced runtime memory requirements.

Concordant Reruns

To meet requirements for clinical use we have ensured Novoalign(CS) produces identical results when rerun with the same data and options. NovoalignMPI also produces identical results to Novoalign.

Multi-threading can change the order of execution and in version 2 the application of random numbers and some penalties would result in reruns producing slightly different numbers of good alignments and slightly different alignment qualities. The differences were usually very small, < 0.01% and had no noticeable affect on SNP calling or expression analysis.
The factors affecting results in version 2 were...

1. Fragment length penalties were calculated from the empirical fragment length distribution and were updated every few thousand reads. Slight changes in order of processing changed which reads were included in the updates and could affect the results.
2. Random numbers were used to select alignments to report when the -r Random option was used and a read had multiple good alignment locations. Threading meant different random numbers were generated for each read & run.
3. Quality calibration, like fragment length penalties, was recalculated periodically and could result in reads having slightly different calibrated qualities in different runs.

The above factors combined meant that results in version 2 were not 100% reproducible.

Version 3 fixes this with two changes...

1. Recalculation of base quality calibration tables and fragment length penalties is done at specific intervals and we wait for all threads to complete current alignments before recalculating the penalties and restarting alignments. This check pointing adds some small delays to processing.
2. Random numbers are generated per read so that a read always has the same random numbers regardless of run, thread or MPI slave aligning the read.

This can be disabled by including option –nonC on the command line.

ION Torrent Reads

The increased alignment score range and support for longer reads has enhanced the utility of Novoalign with Ion Torrent and Ion Proton reads.

We tested Novoalign V3.00 on 400bp Ion Torrent reads to test variant discovery.

Most aligner evaluations are done using simulated reads with insilco generated mutations. This often fails to give a good measure of the aligners ability to handle real data. Common problems include the setting of simulation parameters to best suit the aligners algorithm, failure to include contaminate reads that have come from typical contaminants (when included, reads which are designed to simulate contaminates are usually randomly generated nucleotide sequences with no similarity to real sequences), and simulated sequencing errors that do not reflect the types of errors generated by current 3rd generation sequencers. This particular problematic for ION Torrent reads due to the homopolymer run length errors.

To avoid this issue we did all testing with real reads rather than simulated reads.

For this experiment, reads from E. *coli* K12 are aligned to a related E. *coli* strain, CFT073, and then we compared the aligners ability to recover the differences between the two strains. Mummer was first used to do a whole genome comparison of K12 against CFT073 and establish a list of "true" SNPs and Indels.
The read dataset used is B7-295 (2Gb 400bp Ion PGM 318 Single reads of E. *coli* K12) from Life Technologies.
Reads were aligned with BWA, TMAP and Novoalign V3.00.e (pre-Release). Variants were called with samtools mpileup, GATK UnifiedGenotyper and FreeBayes.

Read Mapping

	NovoalignV3	TMAP	BWA
Total Reads	6,668,556	6,668,556	6,668,556
Reads Mapped	5,533,476	5,721,160	2,293,297
Percentage	82.98%	85.79%	34.39%

Table 1. Summary of read mapping rates for the short read alignment programs.

Read depth reads are varies in different aligners. BWA read depth mostly affected by the low identity of CFT043 alignment and sequencing errors. The average read depth for BWA, TMAP, and Novoalign are 220, 450, and 446 respectively.

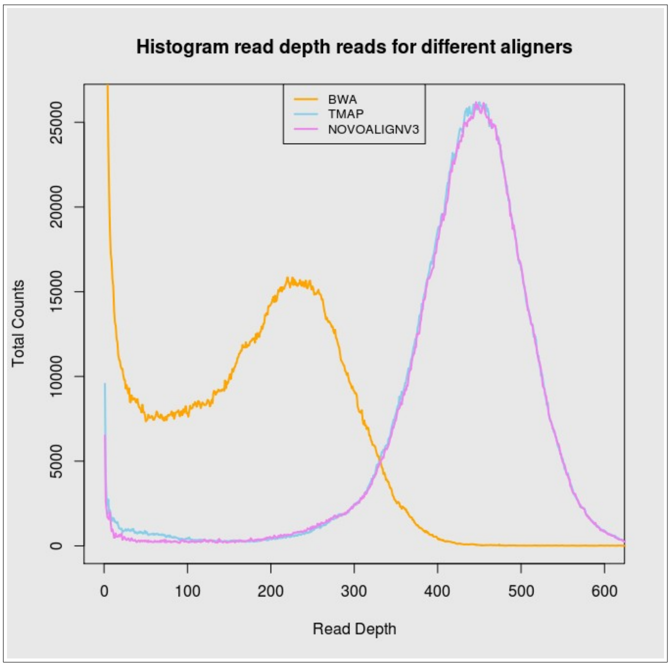

Illustration 7: Read depth histogram for the different aligners. BWA had trouble aligning reads in divergent regions of the genome and reads with high numbers of indel sequencing errors. Novoalign showed slightly more uniform cover than tmap with less low coverage regions.

Variant Recovery Comparison

The SNPs and Indels reported by the Mummer used as the “true” polymorphism for comparison of the three aligners. All Mummer SNPs and Indels location were extracted and compared against the SNPs and Indels calls from the Novoalign V3.00, BWA and TMAP. The comparison was done by using VariantEval from GATK suite of programs.

For SNP recovery, 9, using default alignment parameters, TMAP performed best with a True Discovery Rate (TDR) of 97.58%, NovoalignV3 at 96.66%, and BWA last at 75.04%. Novoalign and TMAP performed very well compared to BWA. For False Discovery Rate (FDR), BWA was best at 0.76%, Novoalign next best at 2.81%, and TMAP at 5.13%.

Next, the Indel variants detected by Mummer were extracted and compared against the indel calls from the three aligners.

True Discovery Rate (TDR) of Indels, 9, was disappointing as none of the aligners performed very well, BWA was best at 5.47%, Novoalign at 1.99% and TMAP at 1.25%. False Discovery Rates (FDR) were also very high for all aligners with NovoalignV3 at 85.23%, TMAP at 92.25%, , and BWA at 91.75%.

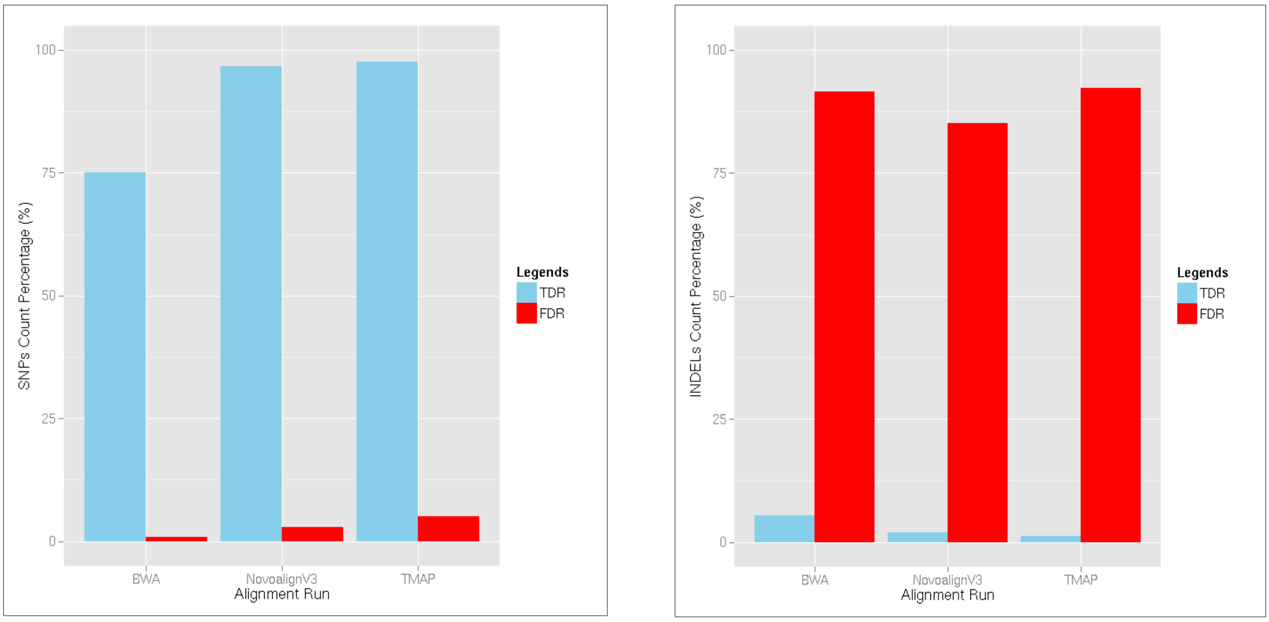

Illustration 8: Comparison of SNP (left) and Indel (Right) True Discovery Rate(TDR) and False Discovery Rate (FDR) using different aligners. Variant calling using samtools mpileup.

We tried different consensus callers to see if indel TDR could be improved. For this test we tried GATK UnifiedGenotyper using a Realigned BAM and Freebayes.

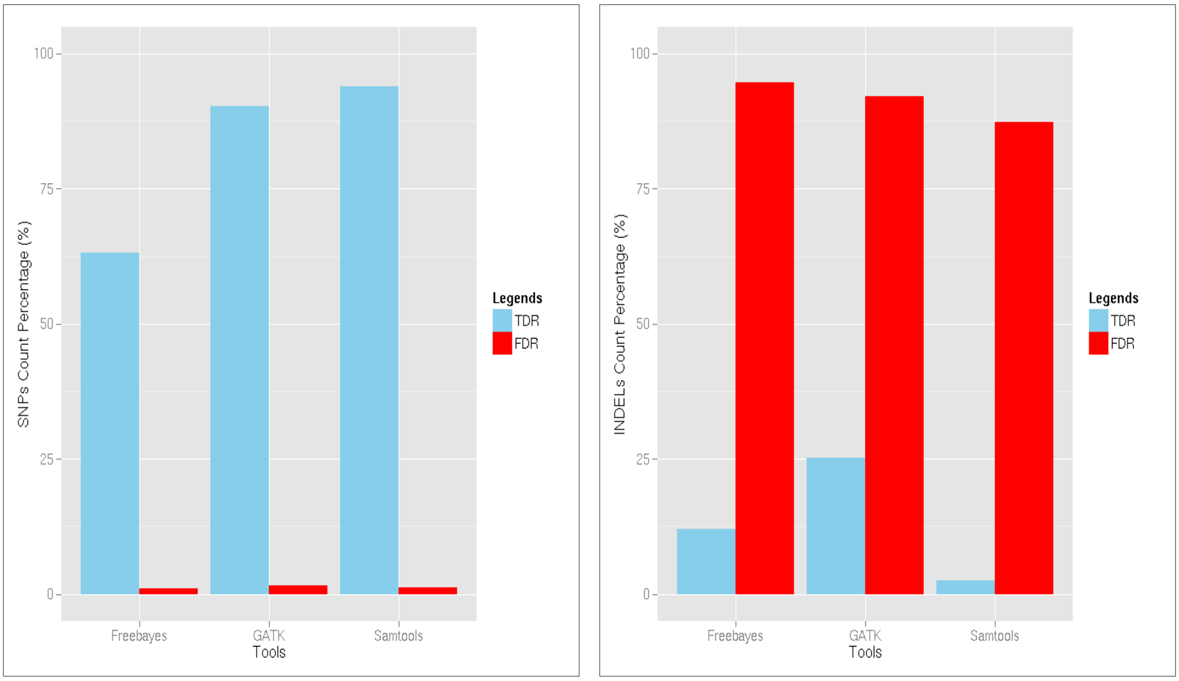

Illustration 9: SNPs (left) and INDELs (right) TDR&FDR rates from different tools. Samtools reported highest SNP TDR. All three tools reported low FDR for SNPs. For Indels GATK reported the highest TDR while Samtools reported the lowest FDR. Alignments by Novoalign V3.00

Tests with different gap penalties and threshold settings showed that default values were close to optimum. Increasing gap open penalty to 50 made a modest improvement in SNP TDR. Indel TDR (samtools mpileup) stayed below 2.5% for all settings tested.

Other Changes

Deleted Command Line Options

-Q 99	Sets lower limit on alignment quality for reporting. Default 0.
-r [0.99]	Sets lower limit on posterior alignment probability for reporting.

Deleted SAM fields

1. Remove AS tag as it has same value as UQ tag.
2. Remove custom tag ZN as it had the same value as NH tag

Soft Clipping

This is the local alignment step that trims alignments back to the best local alignment. The match reward has been increased from 6 to 8 for good quality bases.

NovoalignCS

NovoalignCS benefits from the code restructuring and is around 50% faster than version 2. It also has the new default threshold calculation but as read lengths are <70bp the changes are aren't significant.

Current restriction of maximum of 7bp indel in single end reads remains. Paired end reads support indels up to 33% of read length in one read of the pair.