

Tuning Novoalign for NextSeq WES Variant Calling

Introduction

Sequencing reads from NextSeq two colour chemistry have a higher error rate than four colour reads and in some situations have over estimated base qualities [1]. This causes some issues with Novoalign, primarily an increase in run time. The mapping of reads with multiple sequencing errors combined with high base qualities leads to reduced precision and recall when compared to reads from Illumina's other platforms.

We look at Novoalign option settings to optimise variant calling and run time for NextSeq WES reads. Precision was increased from 94% to 98% with a slight improvement (0.15% or 40 SNPs) in Recall and a large decrease in run time.

NIST High Confident variants for NA12878 are used as True variants sets with a set of WES Nextera Rapid Capture reads downloaded from Illumina Basespace. Variants were called using Freebayes with default options and then compared to GIAB variants using hap.py.

ID	Command	CAF		Indel		SNP		Prec'n	Recall
		Decoy	SNP	CPU	FN	FP	FN	FP	
1	Default	No	No	100%	285	207	236	1757	93.69%
2	--hlimit 9	No	No	26%	285	207	236	1713	93.83%
3	--hlimit 9 -t 0,2	No	No	12%	289	219	240	1409	94.71%
4	--hlimit 9 -t 0,2 -x 3	No	No	11%	297	180	253	1039	95.99%
5	--hlimit 9 -t 0,2 -x 3 -H 22 --trim3hp	No	No	5%	293	158	256	1067	95.97%
6	--hlimit 9 -t 0,2 -x 3 -H 22 --softclip 35 --trim3hp	No	No	6%	287	156	234	1076	95.95%
7	--hlimit 9 -t 0,2 -x 3 -H 22 --softclip 35 --trim3hp --matchreward 3	No	No	6%	290	151	251	638	97.37%
8	--hlimit 9 -t 0,2 -x 3 -H 22 --softclip 35 --trim3hp --matchreward 3	No	70%	4%	289	150	202	644	97.35%
9	--hlimit 9 -t 0,2 -x 3 -H 22 --softclip 35 --trim3hp --matchreward 3	Yes	70%	5%	289	148	198	535	97.72%
10	--hlimit 9 -t 0,2 -x 3 -H 22 --softclip 35 --trim3hp --matchreward 3 -k	Yes	70%	6%	283	132	197	507	97.86%
	BWA MEM Default	No	NA		321	202	205	2117	92.64%

Table 1: Details of options used for Illustration 1.

Note. This test was done with a single set of reads on NA12878. The final set of options used may reduce the ability to map complex variants and longer indels as they can be soft clipped. GATK Haplotypecaller uses soft clipped bases and its performance should not be affected, other variant callers such as Freebayes may have more false negatives for complex variants.

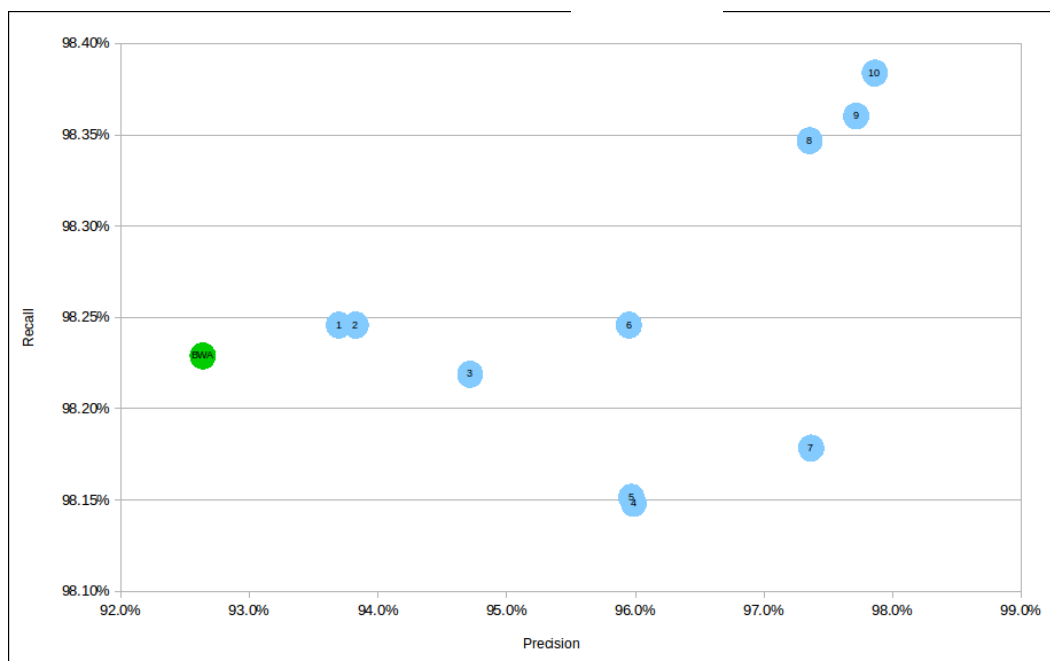


Illustration 1: Path to improved Precision and Recall. In blue is Novoalign

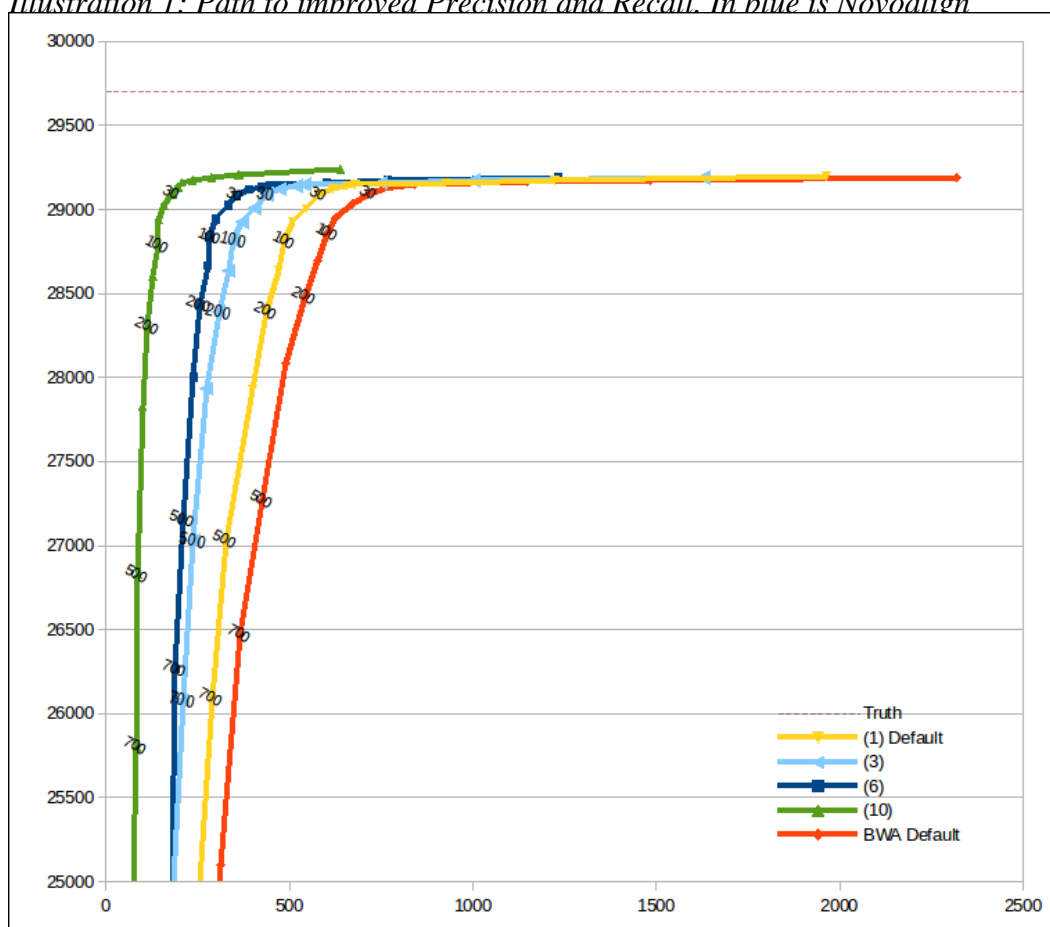


Illustration 2: ROC Curves for selected option combinations with variant quality marks.

Data

GIAB High Confident Regions & Variants

/giab/ftp/release/NA12878_HG001/NISTv3.3.2/GRCh37

HG001_GRCh37_GIAB_highconf_CG-IIIIB-IIIIGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_nosomaticdel.bed

HG001_GRCh37_GIAB_highconf_CG-IIIIB-IIIIGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_PGandRTGphasetransfer.vcf.gz

NA12878 Reads Illumina Basespace

NextSeq 500 v2: Nextera Rapid Capture Exome (CEPH 9plex) – H3GJKBGXX

NA12878 Rep1-3, Lane 1-4, Total 99,140,245 151bp paired end reads.

Programs

Program	Version	Reference
Hap.py	v0.3.7	https://github.com/Illumina/hap.py , Peter Krusche pkrusche@illumina.com
Freebayes	v1.1.0-3-g961e5f3	Erik Garrison, Gabor Marth "Haplotype-based variant detection from short-read sequencing" arXiv:1207.3907 (http://arxiv.org/abs/1207.3907)
BWA	0.7.12-r1039	Heng Li < lh3@sanger.ac.uk >
Novoalign	V3.08.01	Novocraft Technologies Sdn Bhd. All tests included options -r R -a CTGTCTCTTATACACATCT CTGTCTCTTATACACATCT
Novosort	V1.04.06	
Bedtools	v2.20.1-4-gb877b35	

Pipeline

1. NovoalignMPI (V3.08.01) with options ‘-r R -a CTGTCTCTTATACACATCT CTGTCTCTTATACACATCT’ and othe options as detailed below.
2. Novosort –markduplicates
3. Freebayes variant calling. Default options.
4. Filter variants to Nextera Rapid Capture Exome targets (bedtools intersect) -
nexterarapidcapture_exome_targetedregions_v1.2.bed
5. Filter variants to GIAB Confident regions (bedtools intersect)

6. hap.py variant comparison to GIAB Confident variants with filtering as per steps 4&5.

System

Non-homologous mix of 3 X86-64 servers with NovoalignMPI, non-exclusive use. Due to mix of CPU processing speeds and non-exclusive access the CPU times between runs may not be comparable and is include for illustration of possible benefit.

Novoalign Options

--hlimit <F>

This option limits the alignment threshold so that the maximum number of mismatches allowed is a function of the number of mismatches required for the read to align to a perfect homopolymer.

Where:

F is typically in range 5-15 and limits alignment threshold to $F \cdot N_h$

N_h is number of mismatches required to align to a perfect homopolymer.

As mismatches typically score 30 a value of 10 would allow 1/3 the number of mismatches as there are bases differing from a homopolymer.

We suggest using --hlimit 8 for Bi-sulphite alignments.

Read mappers often avoid using over represented k-mers as seeds as in bwa mem which is a more general way of dealing with low complexity reads.

Usage: bwa mem [options] <idxbase> <in1.fq> [in2.fq]

Algorithm options:

...

-c INT skip seeds with more than INT occurrences [500]

Novoaligns --hlimit option avoids using homomeric seeds. Tests (Table 2) show an increase in precision and a 5 fold reduction in CPU time.

Options	CPU Time (mins)	Precision	Recall
Default	10732	93.69%	98.25%
--hlimit 9	2781	93.83%	98.25%

Table 2: Novoalign test for --hlimit 9 option. The reference is standard grch37, chr1-X&M.

Alignment Threshold (-t)

-t A,B

Sets the alignment score threshold as a function of read length.

$$\text{threshold} = (L - A) * B$$

Where:

L is read length (sum of pairs)

A is usually set $\geq \log_4(\text{Reference genome length})$.

B can be fractional and should always be \leq the gap extend penalty.

Default is **-t log₄(N),4.5** where N is the reference genome length.

Typical value is -t 20,3

Example

A 100bp paired-end read on Human Genome has a default threshold of $(200-16)*4.5 = 828$ which allows $828/30 = 27$ mismatches $(200-27)/200 = 86\%$ identity or a deletion of $(828-40)/6 = 131\text{bp}$ which is longer than a single read of the pair.

Allowing mappings with such a low identity can produce false positive mappings by mapping contaminants and reads with high levels of sequencing errors. Lowering the threshold improves precision and reduces run time.

Options	CPU Time (mins)	Precision	Recall
--hlimit 9	2781	93.83%	98.25%
--hlimit 9 -t 20,4	2277	93.95%	98.24%
--hlimit 9 -t 0,2	1260	94.71%	98.22%

Table 3: Tests showing effect of decreasing the alignment threshold. The reference is standard grch37, chr1-X&M.

Gap Penalties (-o & -x)

Lowering gap extend penalty can help preserve alignment of long gaps after we've reduce the alignment threshold(-t) and as it reduces the penalty for longer gaps it means less 'match rewards' are needed to stop the indel being soft clipped. It may also align more indels rather than a series of mismatches.

The original default of gap open of 40 and extend of 6 was based on log odds ratio between a 1bp and a 2bp indels

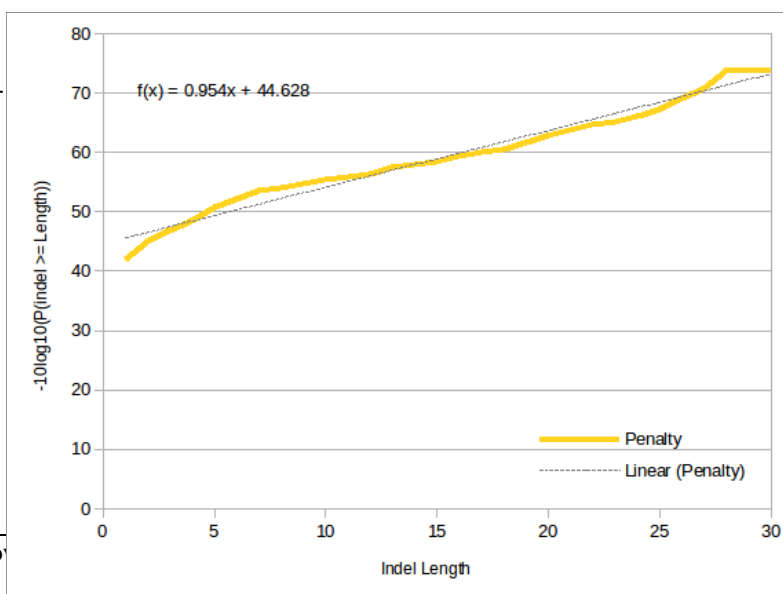


Illustration 3: Phred scaled probability of indel longer than length based on NA12878 GIAB High Confidence variants in TruSeq WES target regions. Linear regression formulae and line shown.

from dbSNP in 2008 to suit use with 36bp Solexa reads.

Now we have better data and longer reads we should revise the penalties. Using Indel distribution from GIAB High Confident VCF file for whole genome and Nextera Exome regions we get phred scaled gap open of 44 and gap extend 1.

Gap penalties also interact with match reward and mismatch penalties (fixed at 30 for high quality bases) and what matters is the ratio between the penalties, match rewards and the alignment threshold.

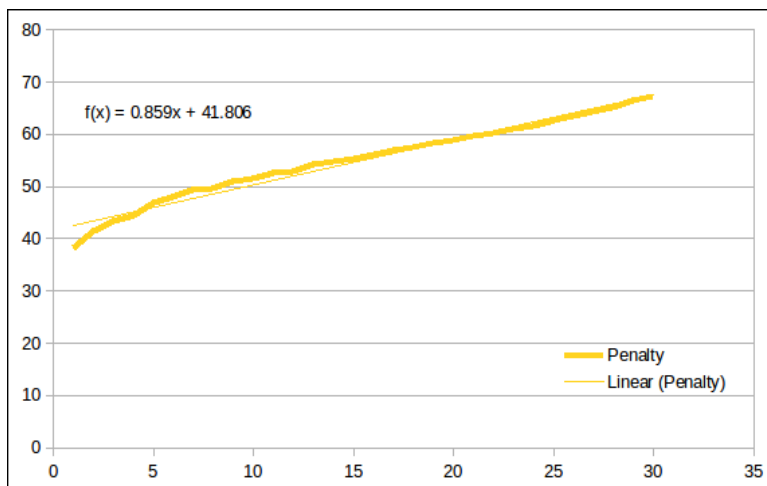


Illustration 4: Similar to above but for Whole Genome. Linear regression formulae and line shown.

Decreasing gap extend to 3 gave a 1% increase in precision that was primarily from a drop in false positive SNVs rather than a change in indel calls.

Options	CPU Time (mins)	Precision	Recall
--hlimit 9 -t 0,2	1260	94.71%	98.22%
--hlimit 9 -t 0,2 -x 3	1223	95.99%	98.15%

Hard Clipping 3' bases (-H & --trim3hp options)

-H [limit [margin]]

Hard clips 3' bases with average quality \leq limit from reads before aligning them. This uses a modified Mott algorithm similar to that used in BWA.

Starting at 3' most base we calculate the base (quality – limit), keeping a running sum of this value. If the running sum exceeds the margin, or we reach the 5' end of the read, then bases from the minimum value to the 3' end of the read are trimmed.

Hard clipping is applied before the polyclonal filter.

Any N's in read are treated as quality 2.

Specifying -H alone sets the quality limit at 2. The margin value defaults to 30.

--trim3HP

Hard clip 3' homopolymers regardless of base quality. Min length 15bp and 88% pure. Useful for reads that degrade to high base quality homopolymer sequences as seen in some 2-dye runs. Applied after -H if used.

Note. Reads were quality binned with values 2,14,21,27,32 & 36

Tests show using -H alone had minimal effect while --trim3hp reduces CPU time dramatically. When used together CPU time is further reduced.

Options	CPU Time (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3			
default	1223	95.99%	98.15%
-H 10	1238	95.99%	98.15%
-H 20	1321	95.82%	98.16%
-H 22	1288	95.86%	98.14%
-H 22 --trim3hp	562	95.97%	98.15%
--trim3hp	663	96.05%	98.16%

A second test using different base options and a reference that included decoy sequences and IUPAC codes for SNPs with CAF < 70% showed a similar drop in CPU time when both options were used and a small increase in precision.

Options	CPU Time (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3 --softclip 35 --matchreward 3 -p 5,20 -k			
default	1401	97.58%	98.40%
-H 22 --trim3hp	674	97.83%	98.39%
-H 20 --trim3hp	754	97.72%	98.40%
-H 10 --trim3hp	999	97.59%	98.40%
--trim3hp	998	97.59%	98.40%
-H 22	1177	97.78%	98.38%

Polyclonal Filtering

-p 99,99 [0.9,99]

Sets thresholds for polyclonal filter. This filter is designed to remove reads that may come from polyclonal clusters or beads. Please refer to paper: *Filtering error from SOLiD Output, Ariella Sasson and Todd P. Michael.*

The first pair of values (n,t) sets the number of bases and threshold for the first 20 base pairs of each read. If there are n or more bases with phred quality below t then the read is flagged as polyclonal and will not be aligned. The alignment status is 'QC'. The second pair applies to the entire read rather than just the first 20bp and is specified as fraction of bases in the read below the given quality. Setting **-p -1** disables the filter. Default for Novoalign is **off**.

This filter was implemented for ABI SOLiD color space reads as their base caller didn't filter polyclonal reads. Solexa & Illumina reads didn't require this filter until Patterned flow cells and 2 color chemistry were introduced. In early tests on these reads we found the polyclonal filter helped reduce CPU time and improve variant calling. However, the tests on this data set have shown almost no affect on variant calls.

Options	CPU Time (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3 -H 22 --softclip 35 --trim3hp --matchreward 3 No decoy or SNVs			
No -p option	605	97.37%	98.18%
-p 4,20	622	97.38%	98.18%
-p 5,20	618	97.39%	98.18%

Options	CPU Time (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3 -H 22 --softclip 35 --trim3hp --matchreward 3 Decoys and SNVs with CAF<70%			
No -p option	693	97.86%	98.38%
-p 5,20	674	97.83%	98.39%
-p 6,20	670	97.83%	98.39%
-p 7,20	670	97.83%	98.38%

Soft clipping options (--softclip & --matchreward)

--matchreward 9	<p>Sets a match reward. If not set then a match reward of 6 is only used during soft clipping and the initial Needleman-Wunsch alignment does not use a match reward.</p> <p>When match reward is set it is used in soft clipping and acts as an extra penalty for inserted bases during the initial Needleman-Wunsch glocal alignment.</p> <p>Setting a low match reward such as 3 helps soft clip ragged ends from NextSeq alignments.</p>
--softclip 99[,99]	<p>Turns on soft clipping and sets a reward for alignments extending to the start or end of a read. Typical value of 40. The first value is for 5' of read and second for 3' of read. Default 0,0.</p> <p>The reward is only used in the soft clipping routine and is not added to the reported alignment score.</p>

Note. The --softclip option was changed from V3.08.00 to have separate values for 5' & 3' ends of the read. In previous releases the reward was used for both ends of the alignment.

Novoalign uses a glocal Needleman-Wunsch alignment to determine the best mapping location for a read. Mismatches and indels near the ends of an alignment can be incorrectly mapped or called due to either,

1. Indels near the ends of a read may get mapped as mismatches as they have a lower alignment score, or
2. The 3' ends of the read typically have more sequencing errors so mismatches may not be true variants.

Novoalign soft clips alignments by using a modified Smith-Waterman¹ alignment prior to reporting. Adjusting these options can make significant differences to precision and recall.

Using reference with decoy sequences and IUPAC codes for dbSNP SNVs with CAF < 70%

Options	CPU (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3 -p 5,20 -k -H 22 --trim3hp			
--softclip 35 --matchreward 2	672	98.36%	98.18%

¹ . Novoalign Smith-Waterman soft clipping left aligns indels which retains some indels that would not be part of usual Smith-Waterman alignment. Our tests have shown this improves indel calling.

--softclip 35 --matchreward 3	674	97.83%	98.39%
--softclip 35 --matchreward 4	658	97.31%	98.43%
--softclip 35 --matchreward 5	616	96.59%	98.43%

Options	CPU Time (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3 -p 5,20 -k -H 22 --trim3hp			
--softclip 10 --matchreward 3	625	97.85%	98.27%
--softclip 20 --matchreward 3	670	97.86%	98.34%
--softclip 35 --matchreward 3	674	97.83%	98.39%
--softclip 52 --matchreward 3	597	97.83%	98.38%

IUPAC Ambiguous codes for common SNVs

Novoalign fully supports all 15 possible IUPAC nucleotide codes in the reference sequence. This allows common SNVs to be encoded into the reference as IUPAC ambiguous codes and then for novoalign to map reads to the SNVs without incurring a mismatch penalty hence reducing likelihood of the SNV being soft clipped from an alignment.

It's typically used in RNA alignments to reduce allelic bias and can also help with SNV recall.

In this test we add common SNVs to the reference at several allele frequency limits and look at the affect on precision and recall.

We use 00-common_all.vcf from NIST3.3.2 as a source of SNVs and filter these based on allele frequency of the reference allele.

```
##INFO=<ID=CAF,Number=.,Type=String,Description="An ordered, comma delimited list of
allele frequencies based on 1000Genomes, starting with the reference allele followed by alternate
alleles as ordered in the ALT column. Where a 1000Genomes alternate allele is not in the dbSNPs
alternate allele set, the allele is added to the ALT column. The minor allele is the second largest
value in the list, and was previously reported in VCF as the GMAF. This is the GMAF reported on
the RefSNP and EntrezSNP pages and VariationReporter">
```

...

```
#CHROM POS ID REF ALT QUAL FILTER INFO
```

```
1 10642 rs558604819 G A . . CAF=0.9958,0.004193
```

SNVs are added to the reference using 'novoutil iupac'

Example

To build a reference with all common SNVs with reference CAF < 0.70

```
novoutil iupac <( grep -v CAF=0.[789] 00-common_all.vcf) grch37.fa > grch37.070.fa
novoindeix grch37.070.nix grch37.070.fa
```

Options common to all runs were --hlimit 9 -t 0,2 -x 3 -H 22 --softclip 35 --trim3hp --matchreward 3 -p 5,20 without decoy sequences.

Results were a slight drop in CPU time and increase in recall of 0.2% or about 50 SNPs at CAF 70%. Higher CAF limits made no further improvement.

Reference CAF limit	CPU(mins)	Precision	Recall
No SNVs added	618	97.39%	98.18%
<60%	597	97.38%	98.28%
<70%	604	97.37%	98.35%
<90%	597	97.36%	98.36%
<95%	602	97.38%	98.37%
<99%	604	97.35%	98.37%

Using Decoy Sequences

Decoy sequences produced by Heng Li and part of the Broad Bundle are useful for reducing false positive mappings by drawing some mappings away from the main reference sequence to the decoys.

This results in a slight increase in precision.

This test used a reference with IUPAC ambiguous codes for SNVs with CAF <70%

Options	CPU (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3 -p 5,20 -k -H 22 --trim3hp --softclip 35 --matchreward 3			
No decoys	626	97.47%	98.38%
With decoys	674	97.83%	98.39%

Novoaligns Quality Calibration

-k [infile]

Enables quality calibration. The quality calibration data (mismatch counts) are either read from the named file or accumulated from actual alignments. Default is no calibration.

Note. Quality calibration does not work with reads in prb format.

Novoalign is a quality aware align with mismatch penalties adjusted to reflect the base qualities. Novoaligns quality calibration dynamically adjusts base qualities and mismatch penalties according to mismatches seen in the reads. This can have several affects such as improved mapping accuracy, improved MAPQ scores and, as recalibrated qualities are output to the SAM QUAL attribute, it can improve accuracy of variant calling in a similar fashion as GATKs BQSR.

On this dataset quality calibration produced a small improvement in both precision and recall.

The first test had a standard reference with neither decoy sequences nor IUPAC ambiguous codes for common alleles.

Options	CPU (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3 -p 5,20 -k -H 22 --trim3hp --softclip 35 --matchreward 3			
No calibration	604	97.37%	98.35%
With calibration	626	97.47%	98.38%

The second test included both decoy sequences and IUPAC ambiguous codes for common alleles CAF < 70%. Using IUPAC ambiguous codes for common SNPs is recommended with quality calibration as means these SNPs will not be counted as mismatches which helps maintain higher base qualities.

Options	CPU (mins)	Precision	Recall
--hlimit 9 -t 0,2 -x 3 -p 5,20 -k -H 22 --trim3hp --softclip 35 --matchreward 3			
No calibration	645	97.74%	98.36%
With calibration	674	97.83%	98.39%

References

Do you have two colors or four colors in Illumina?

“The 2-channel System is a further development of the 4-channel system resulting in a faster sequencing process. Unfortunately it has also disadvantages: The base quality is determined, among other factors, by the purity of the emitted light signal of each cluster per run. Furthermore, incorrect base calls because of [phasing](#) lead to a rising pollution of the light signals over time, making it more difficult to differentiate the bases and to interpret the base quality. For example, the 2-channel system interprets a mix out of two light signals (red and green) as signal for adenine. Such a mixed red-green signal should always be adenine meaning that a red-green mixed signal due to [phasing](#) (and not by adenine) could be overlooked. This can distort base qualities and even the accuracy of the sequences.”

https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina

Illumina 2 colour chemistry can overcall high confidence G bases

<https://sequencing.qcfail.com/articles/illumina-2-colour-chemistry-can-overcall-high-confidence-g-bases/>