Novoalign Version 4.0 Release Notes

A major update to Novoalign

- Reduced run time
- Revised default options for alignment scoring
- Removal of obsolete functions
- Some command line options have changed
- BAM output format (not in V4.00.Pre-20190624)

Run time

Run time has been reduced by

1. Major rewrites to SIMD alignment code including increased use of AVX2
2. Added caching to reduce duplicated processes during iterative alignment.
3. New default alignment scoring options help with run time performance.

V3 vs V4 Variant Calling results

WES NextSeq

Reference was hs37 plus decoys from Broad bundle and for Novoalign IUPAC ambiguous codes were added for Common SNPs with reference allele CAF < 0.70. Variants compared to GIAB[1] High confident variants using hap.py[5]. Variant counts were taken at Q10 for freebayes[3] (version: v1.1.0-3-g961e5f3) and Q30 for GATK[4] (gatk-4.1.2.0). This is close to variant quality with maximum F2 score. The truth set comprised 29679 GIAB high confident variants in the target regions.

Freebayes was run with options --min-alternate-fraction 0.05 -m 5

Bwa mem[2] (0.7.12-r1039) was run at default settings.

Reads - NextSeq 500 v2: Nextera Rapid Capture Exome (CEPH 9plex) - H3GJKBGX Rep1

Freebayes

Release	Options	dT[1] (mins)	True Positives	False Positives
V3.09.02	Default	605	28533	852
V4	V3 Defaults	52	28535	839

[1] Elapsed time for alignment using 32 core 64 thread server

V3.09.02	-g 40 -x 1 --matchReward 4 --softclip 50,30 --trim3hp -H 22 -t 0,2.5 --hlimit 8 -v 150 --pechimera	42	28594	827
V4	--tune NextSeq --trim3hp ACGT	14	28596	820
bwa mem	0.7.12-r1039 Default	6	28561	944

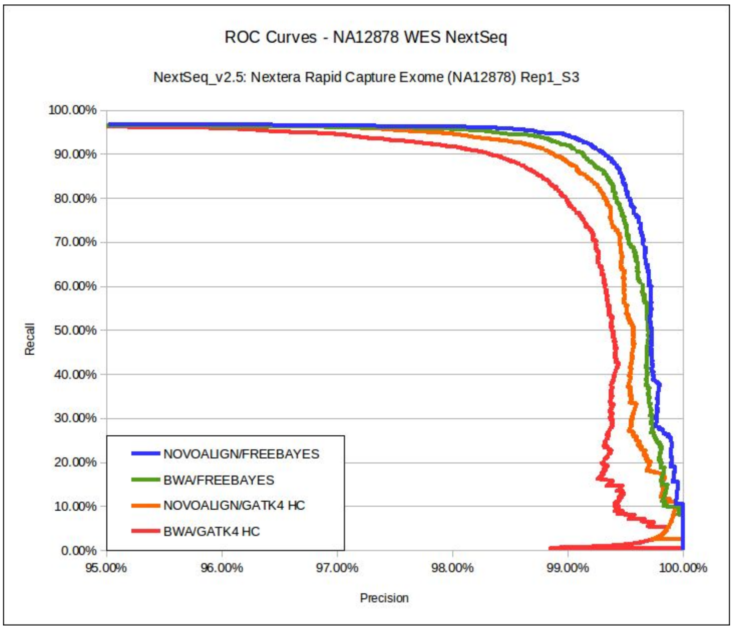
GATK4 HaplotypeCaller

Release	**Options**	**dT (mins)**	**True Positives**	**False Positives**
V3.09.02	Default	605	28614	1514
V4	V3 Defaults	52	28630	1522
V3.09.02	-g 40 -x 1 --matchReward 4 --softclip 50,30 --trim3hp -H 22 -t 0,2.5 --hlimit 8 -v 150 --pechimera	42	28674	1772
V4	--tune NextSeq --trim3hp ACGT	14	28668	1755
bwa mem	0.7.12-r1039 Default	6	28595	1583

Note. GATK false positive SNP calls can be reduced by using option --softclip 30,20

ROC Curves

ROC Curves were generated for different versions of test pipeline. The aligner was either Novoalign --tune NextSeq or bwa mem at default. Alignments sorted by novosort with duplicates removed and then variant calling with either Freebayes or GATK4 HaplotypeCaller. Freebayes has significantly higher precision than HaplotypeCaller which is why we chose it as part of our validation process.



```
freebayes  -f hs37d.fa
--min-alternate-fraction 0.05 -m 5 NA12878.bam >NA12878.vcf

gatk HaplotypeCaller -R hs37d5.fa -I NA12878.bam -O NA12878.vcf -stand-call-conf
1 --java-options "-Xmx3G"
```

WGS HISEQ-X

Reference genome as per WES tests. Variant Caller: Freebayes

Projects : HiSeqX: PCR-free v2.5 : NA12878-rep1

The PCR-free library was prepared using Human DNA from Coriell sample NA12878 and sequenced on a HiSeqX instrument using v2.5 chemistry.
Owner - Illumina Public Data
Created - 2016-03-17 14:20
Size - 1.02 TB

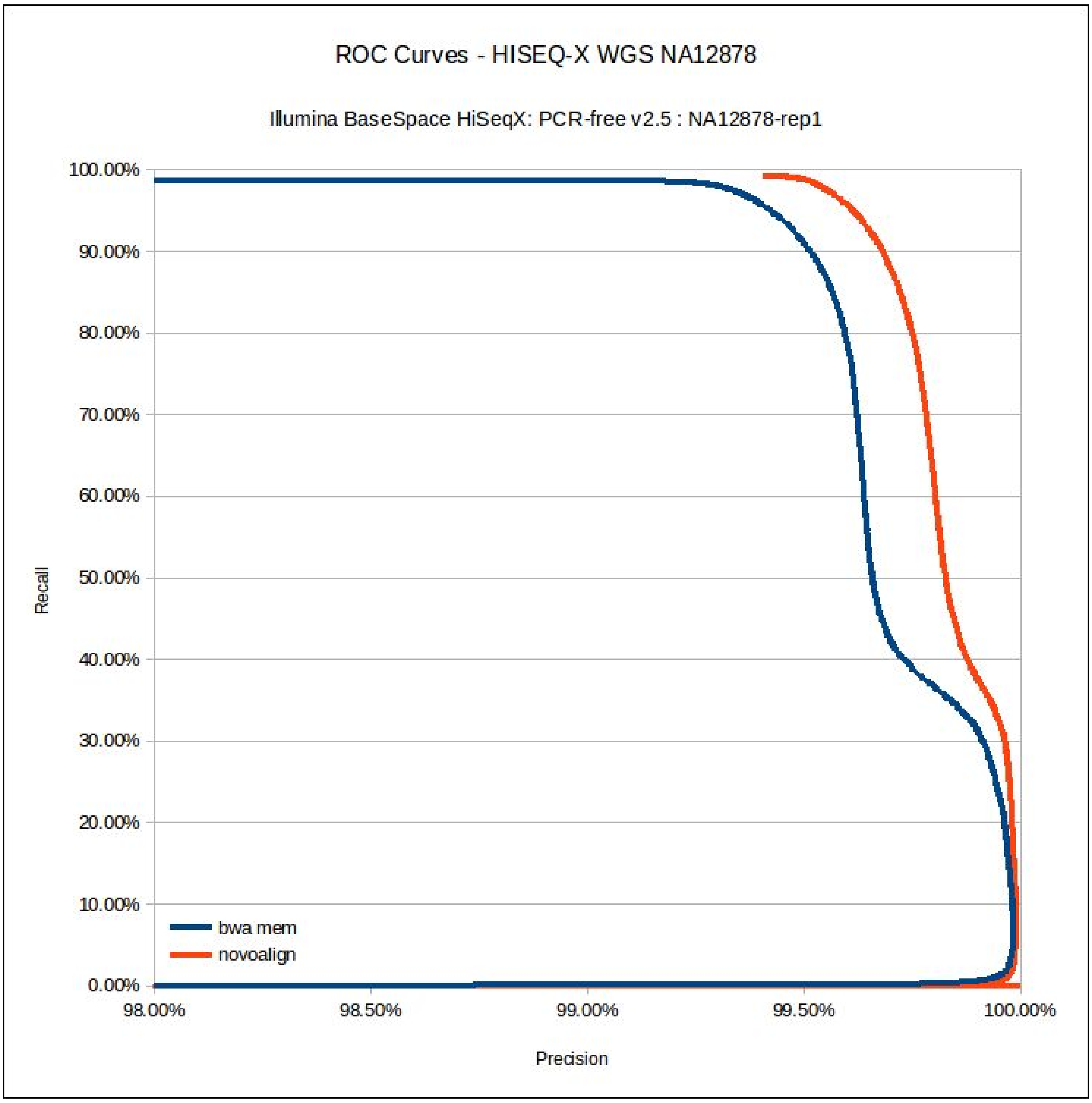
342,593,213 pairs of 150bp

Timings

Aligner	**dT[2] (hrs)**	**SNP Recall %**	**SNP Precision %**	**Indel Recall %**	**Indel Precision %**
Novoalign --tune HISEQX	2.8	99.97	98.91	94.51	95.81
bwa mem	1.8	99.96	97.30	90.98	93.72

ROC Curves

[2] Elapsed time for alignment step using 32 core 64 thread server.



WGS MGISEQ2000

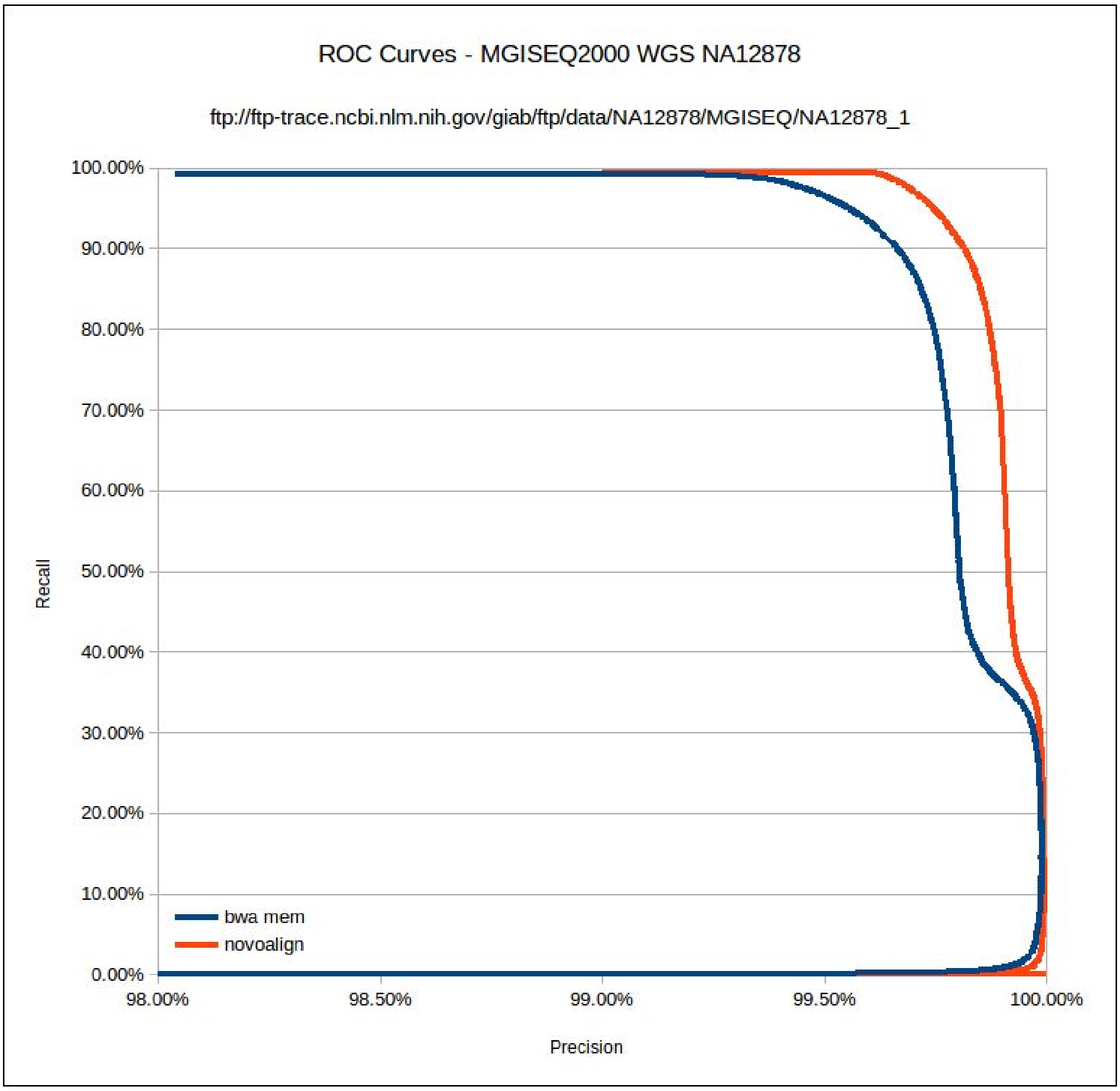
Mapped to GRCH38 with alternate scaffolds, decoys and, for novoalign, with IUPAC codes for SNPs with reference allele CAF < 0.7. Variant caller: Freebayes

GIAB ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/MGISEQ/NA12878_1
701,636,066 pairs of 150bp

Timings

Aligner	dT (hrs)	SNP Recall %	SNP Precision %	Indel Recall %	Indel Precision %
Novoalign --tune MGISEQ	6.7	99.96	99.23	97.08	97.76
bwa mem	5.1	99.95	98.28	95.31	96.61

ROC Curves



Obsolete Functions

1. Support for ABI SOLiD colour space reads has been removed. This is still available in V3 with continuing support.
2. Removed support for some read formats such as prb and qseq formats.
3. Removed Pairwise output format

Changes to Default Settings

The new defaults were chosen to maximise F-score on NA12878 whole exome and genome datasets. We used data from different sequencing systems including HiSeqX, HiSeq2500, Novaseq, NextSeq & BGI_500, comparing variants against the GIAB High Confident variants.

Option		V4	V3
-g	Gap open penalty	40	40
-x	Gap extend penalty	2	6
--matchreward	Match Reward	4	6
--softclip	Reward for not soft clipping	50,30	0,0
-H	3’ low quality base trimming	5	off
-t	Alignment score threshold	0,3	20,4
--hlimit	Alignment score threshold for low complexity reads	8	off
-v	Structural variation penalty	150	70
--pechimera	Paired end chimera support	on	off
-r	Reporting for multi-mapped reads	Random	None
-o	Output report format	SAM	Native
--tag LB	SAM LIbrary tag. Redundant as read group identifies the library.	Off	On
-u	Bi-Seq unconverted cytosine penalty	8 iff -b 4	0

Changes to Option Formats & Behaviour

V3	V4	Notes
--alt	--alt [on|off]	Enables alt-scaffold mode. This is now automatically turned on if the reference contains alternate scaffolds. If you want to disable it use **--alt off**
-H [t [m]]	-H [off| t [m]]	Hard clipping low quality 3’ bases now defaults to a quality of 5. Use **-H off** to disable.
--trim3HP	--trim3HP [bases|Off]	This now takes a list of bases to trim. Using **--trim3hp AG** is suitable for 2 colour sequencers such as NextSEQ
--pechimera	--pechimera [on|off]	Now defaults to on.
-5 <r1>[,l1] [r2[,l2]]		If read 2 sequence and length are not specified then no trimming will be done on the 5’ of read 2. In V3 the Read 1 settings were applied to both reads.

Defaults Options

V4 revises the default settings for options. This is something we've planned to do for a long time but felt it should be part of a major release.

These options have been chosen by maximising variant F2-score using a hill climbing algorithm on NA12878 WES and WGS reads with comparison against GIAB High confident variants using hap.py. Multiple data sets form a broad range of sequencing systems were used.

These settings can be applied using the --tune option.
    --tune [Tuning Option]

Tuning Option	Settings applied
Default	-g 40 -x 2 --matchReward 4 --softclip 50,30 -H 5 -t 0,3.0 --hlimit 8 -v 150 -r Random --pechimera on -u 8[3]
HiSeqX	-g 40 -x 2 --matchReward 4 --softclip 50,30 -H 7 -t 0,**2.0** --hlimit **9** -v 150 -r Random --pechimera on -u 8
HiSeq	-g 40 -x 1 --matchReward 4 --softclip **45**,30 -H 17 -t 0,**2.5** --hlimit **9** -v 150 -r Random --pechimera on -u 8
NextSeq	-g 40 -x 1 --matchReward 4 --softclip 50,30 **--trim3hp AG** -H **22** -t 0,**2.0** --hlimit 8 -v 150 -r Random --pechimera on -u 8
NOVASEQ	-g 40 -x 1 --matchReward 4 --softclip **45**,30 -H **17** -t 0,**2.5** --hlimit 8 -v 150 -r Random --pechimera on -u 8
BGISEQ500	-g 40 -x 1 --matchReward 4 --softclip 50,30 -H **12** -t 0,**3.5** --hlimit 8 -v 150 -r Random --pechimera on -u 8
MGISEQ2000	-g 40 -x 1 --matchReward 4 --softclip 50,30 -H **12** -t 0,**1.5** --hlimit 8 -v 150 -r Random --pechimera on -u 8
IONTorrent-1	-g 95 -x 2 --matchReward **2** --softclip **100,50** -H **15** -t 0,**4.0** --hlimit **9** -v 150 -r Random **-k** -u 8
IONTorrent-2	-g 100 -x **3** --matchReward **2** --softclip **100,25** -H **5** -t 0,3.0 --hlimit **7** -v 150 -r Random **-k** -u 8
V3-Defaults	-g 40 -x **6** --softclip **0,0** -H **2** -t **16,4.5** --hlimit **9** -v **70** -r **None -u 0**

The two ION Torrent settings had similar F2 scores with IONTorrent-1 being considerably better for SNPs while the IONTorrent-2 settings had significantly lower false positives on INDELs and slightly higher false positive SNPs.

[3] -u option only applies in Bi-Seq alignment mode

Testing on these settings is continuing and they are subject to change during pre-release of V4. Use **novoalign --help** to see actual settings.

BAM Output Format

Novoalign V4 can now write directly to BAM format without the need of a conversion tool like samtools view.

To write BAM simply use option -o BAM rather than -o SAM.

The BAM option can also be followed by a compression level and an @RG record. Compression level is a single digit in range 0-9. Default level is 4. If piping directly to Novosort set compression to 0 (zero).

```
novoalign -d …    -o BAM 6 “@RG….” >results.bam
```

This has a performance advantage over piping into samtools view for servers with high core counts.

RNA Alignments

When using novoalign with USEQ package for RNA alignment you need to set --softclip 0,0 to prevent CIGARS with leading & trailing inserts that can cause USEQ to crash.

References

1. An open resource for accurately benchmarking small variant and reference calls.(**Zook, et al., Nature Biotechnology 2019**)
2. Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].
3. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]* 2012
4. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010 *GENOME RESEARCH 20:1297-303*
5. Haplotype Comparison Tools, Peter Krusche pkrusche@illumina.com, https://github.com/Illumina/hap.py