Sequence Alignment - NGS

Zemin Ning

The Wellcome Trust Sanger Institute

wellcome trust sanger institute

Outline of the Talk:

- ***Global and Local Alignment***
- ***Alignment methods***
- ***Alignment tools: BWA and Smalt***
- ***Comparison of the results***
- ***Data visualisation***

Biological Motivation

Why We Need Sequence Alignment

Inference of Homology

- Two genes are homologous if they share a common evolutionary history.
- Evolutionary history can tell us a lot about properties of a given gene
- Homology can be inferred from similarity between the genes

- Variation Detection – SNP, indel, CNV

Sequence Alignment

Global Alignment:

Goal: How similar are two sequences S_1 and S_2

Input: two sequences S_1, S_2 over the same alphabet

Output: two sequences S'_1, S'_2 of equal length

(S'_1, S'_2 are S_1, S_2 with possibly additional gaps)

Example:

- S_1= GCGCATGGATTGAGCGA
- S_2= TGCGCCATTGATGACC
- A possible alignment:

```
S'1= -GCGC-ATGGATTGAGCGA
S'2=  TGCGCCATTGAT-GACC--
```

Sequence Alignment (cont)

Local Alignment:

Goal: Find the pair of substrings in two input sequences which have the highest similarity

Input: two sequences S_1, S_2 over the same alphabet

Output: two sequences S'_1, S'_2 of equal length

(S'_1, S'_2 are substrings of S_1, S_2 with possibly additional gaps)

Example:

- $S_1 =$ GCGCATGGATTGAGCGA
- $S_2 =$ TGCGCCATTGATGACC
- A possible alignment:

$S'_1 =$ ATTGA-G

$S'_2 =$ ATTGATG

Global vs. Local Alignment

- The Global Alignment Problem tries to find the longest path between vertices *(0,0)* and (*n*,*m*) in the edit graph.

- The Local Alignment Problem tries to find the longest path among paths between **arbitrary vertices** (*i*,*j*) and (*i'*, *j'*) in the edit graph.

Global vs. Local Alignment (cont’d)

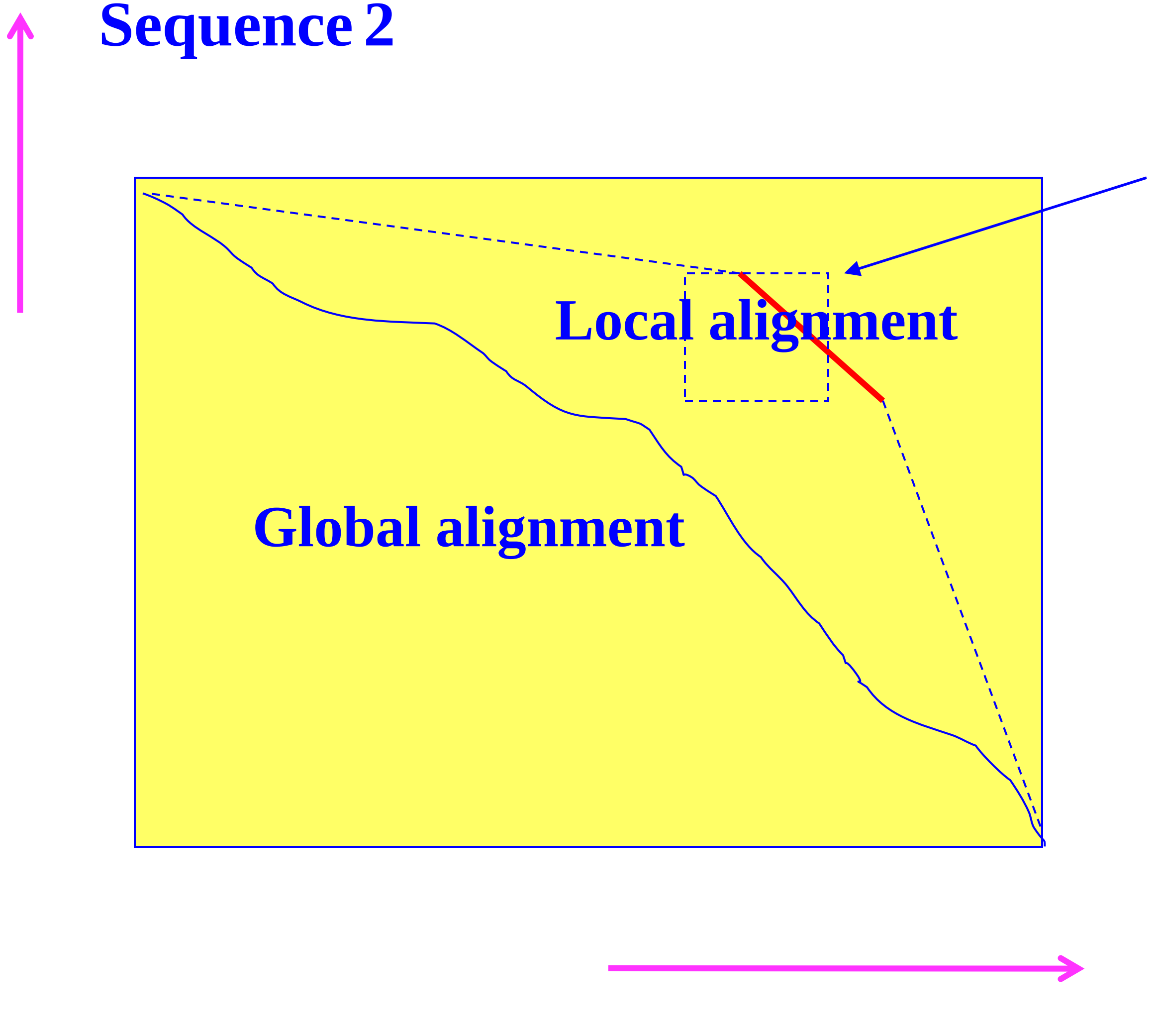
- **Global Alignment**

```
--T—-CC-C-AGT—-TATGT-CAGGGGACACG—A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || | | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG—T-CAGAT--C
```

- **Local Alignment—betten alignment to find conserved segment**

```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                     ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Local Alignment: Example



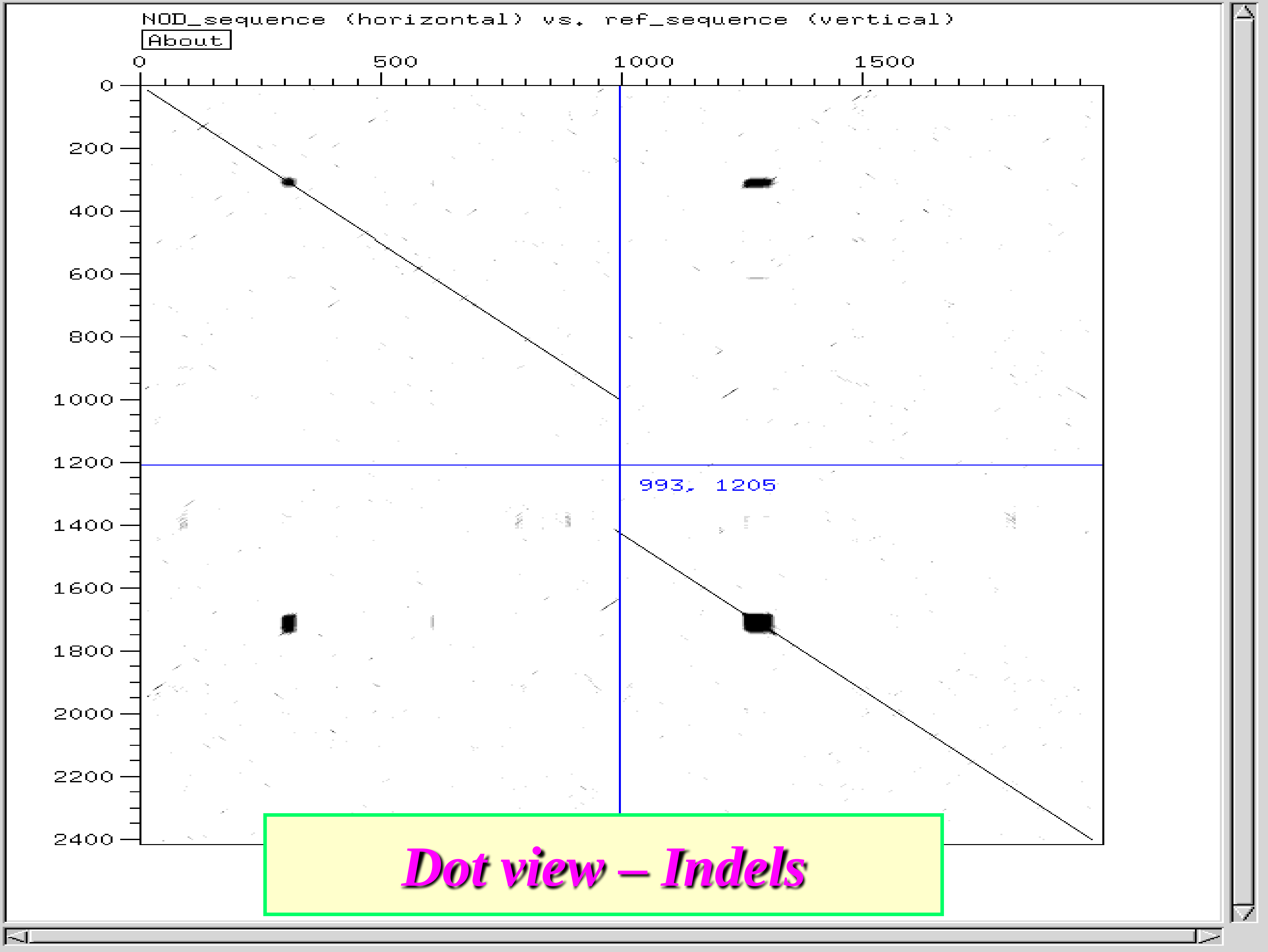
Compute a “mini” Global Alignment to get Local

Methods of DNA Sequence Alignment

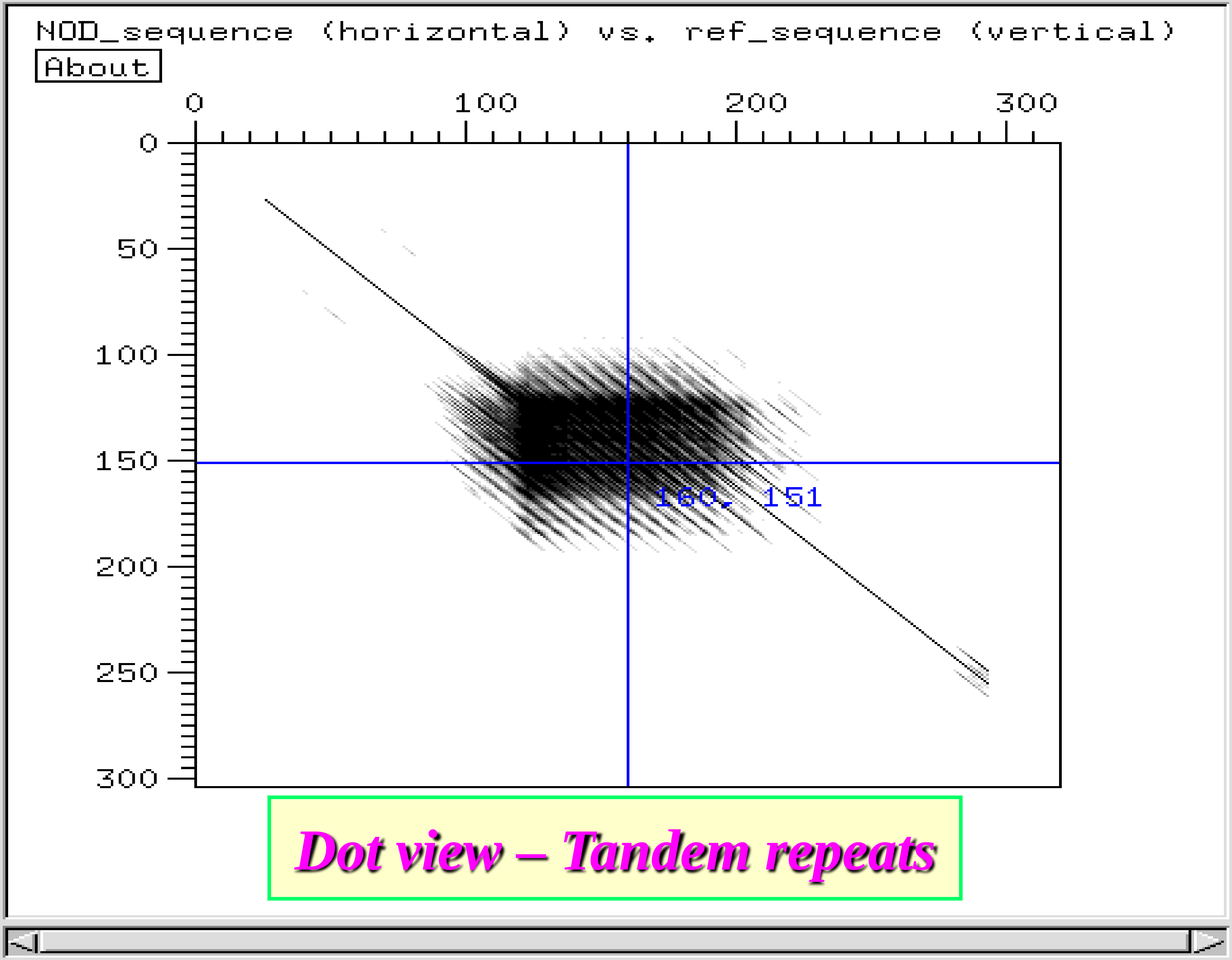
- **Dot matrix analysis**
- **The dynamic programming (DP) algorithm**
    - **Needleman-Wunsch Algorithm**
    - **Smith-Waterman Algorithm**
- **Burrows-Wheeler Index (BWA, Bowtie, SOAP2, etc)**
- **Hash table based algorithm (ssaha2, smalt, novoAlign, etc)**

Dot Matrix Analysis

- A dot matrix analysis is a method for comparing two sequences to look for possible alignment (Gibbs and McIntyre 1970)
- One sequence (A) is listed across the top of the matrix and the other (B) is listed down the left side
- Starting from the first character in B, one moves across the page keeping in the first row and placing a dot in many column where the character in A is the same
- The process is continued until all possible comparisons between A and B are made
- Any region of similarity is revealed by a diagonal row of dots
- Isolated dots not on diagonal represent random matches



Dot view – Indels


NOD_sequence (horizontal) vs. ref_sequence (vertical)

Dot view – Tandem repeats

Smith-Waterman Algorithm

- **Only works effectively when gap penalties are used**
- **In example shown**
  - **match = +1**
  - **mismatch = -1/3**
  - **gap = -1+1/3k (k=extent of gap)**
- **Start with all cell values = 0**
- **Looks in subcolumn and subrow shown and in direct diagonal for a score that is the highest when you take alignment score or gap penalty into account**

	C	A	G	C	C	T	C	G	C	T	T	A	G
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
T	0.0	0.0	0.8	0.3	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	1.0	0.7	0.0	1.0	3.0	1.7	?						
A													
T													
T													
G													
A													
C													
G													
G													

Local alignment score H = 3.0

A--GCC
ATGCC

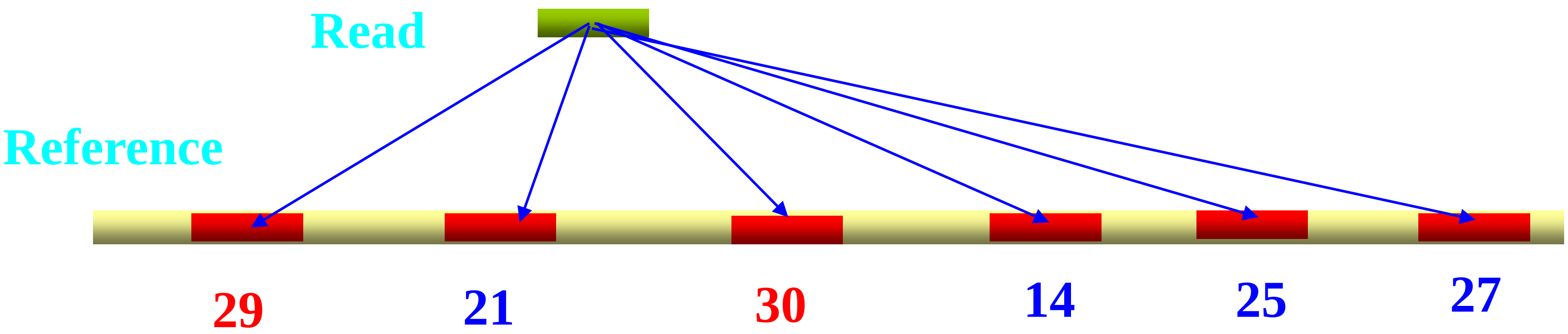
$$H_{ij}=\max\{H_{i-1,\,j-1}+s(a_i,b_j),\ \max\{H_{i-k,j}-W_k\},\ \max\{H_{i,\,j-l}-W_l\},\ 0\}$$

Mapping Score in Short Read Alignment

Read mapping score is used to assess the repetitive feature of the read in the genome. With a maximum mapping score 50, we have:

$$S_{map} = \begin{matrix} 10*(S_{\max} - S_{\max 2}) & if\,(S_{map} <= 50) \\ 50 & if\,(S_{map} > \ 50) \end{matrix}$$

S_{max} - maximum alignment score (smith-waterman) of the hits on genome; S_{max2} - second best alignment score of the hits on genome; Say you have one read of 30 bases which has a few hits on the genome: Best hit: exact match with S_{max} 30; Second best hit: one base mismatch with S_{max2} 29. The mapping score for this read is S_{map} = 10;



Short Read Alignment Tools

Bfast

BioScope

Bowtie

BWA

CLC bio

CloudBurst

Eland/Eland2

GenomeMapper

GnuMap

Karma

MAQ

MOM

Mosaik

MrFAST/MrsFAST

NovoAlign

PASS

PerM

RazerS

RMAP

Smalt

SSAHA2

Segemehl

SeqMap

SHRiMP

Slider/SliderII

SOAP/SOAP2

Srprism

Stampy

vmatch

ZOOM

……

Overview of the BWA algorithm

- **Based on FM-index (Burrows-Wheeler Transform plus auxilliary data structures) which enables fast exact matching.**
- **Short-read algorithm: alter the read sequence such that it matches the reference exactly.**
- **Long-read algorithm (BWA-SW): sample reference subsequences and perform Smith-Waterman alignment between the subsequences and the read.**
- **Work for Illumina and SOLiD single-end (SE) and paired-end (PE) reads; new component BWA-SW for 454/Sanger SE reads.**

Key Features

- □ Fast and moderate memory (<4GB)
- □ SAM output by default
- □ Gapped alignment for both SE and PE reads
- □ Effective pairing to achieve high alignment accuracy; suboptimal hits considered in pairing.
- □ Non-unique read is placed randomly with a mapping quality 0; all hits can be outputted in a concise format.
- □ The default configuration works for most typical input

  -Automatically adjust parameters based on read lengths and error rates.

  -Estimate the insert size distribution

Running BWA

- Input: ref.fa, read1.fq, read2.fq and long-read.fq
- Step 1: Index the genome (3 CPU hours for the human genome):

```
bwa index -a is ref.fa
```

- Step 2a: Generate alignments in the suffix array coordinate:

```
bwa aln ref.fa read1.fq > read1.sai
bwa aln ref.fa read2.fq > read2.sai
```

- Step 3a: Generate alignments in the SAM format:

```
bwa sampe ref.fa read1.sai read2.sai read1.fq read2.fq > aln.sam
```

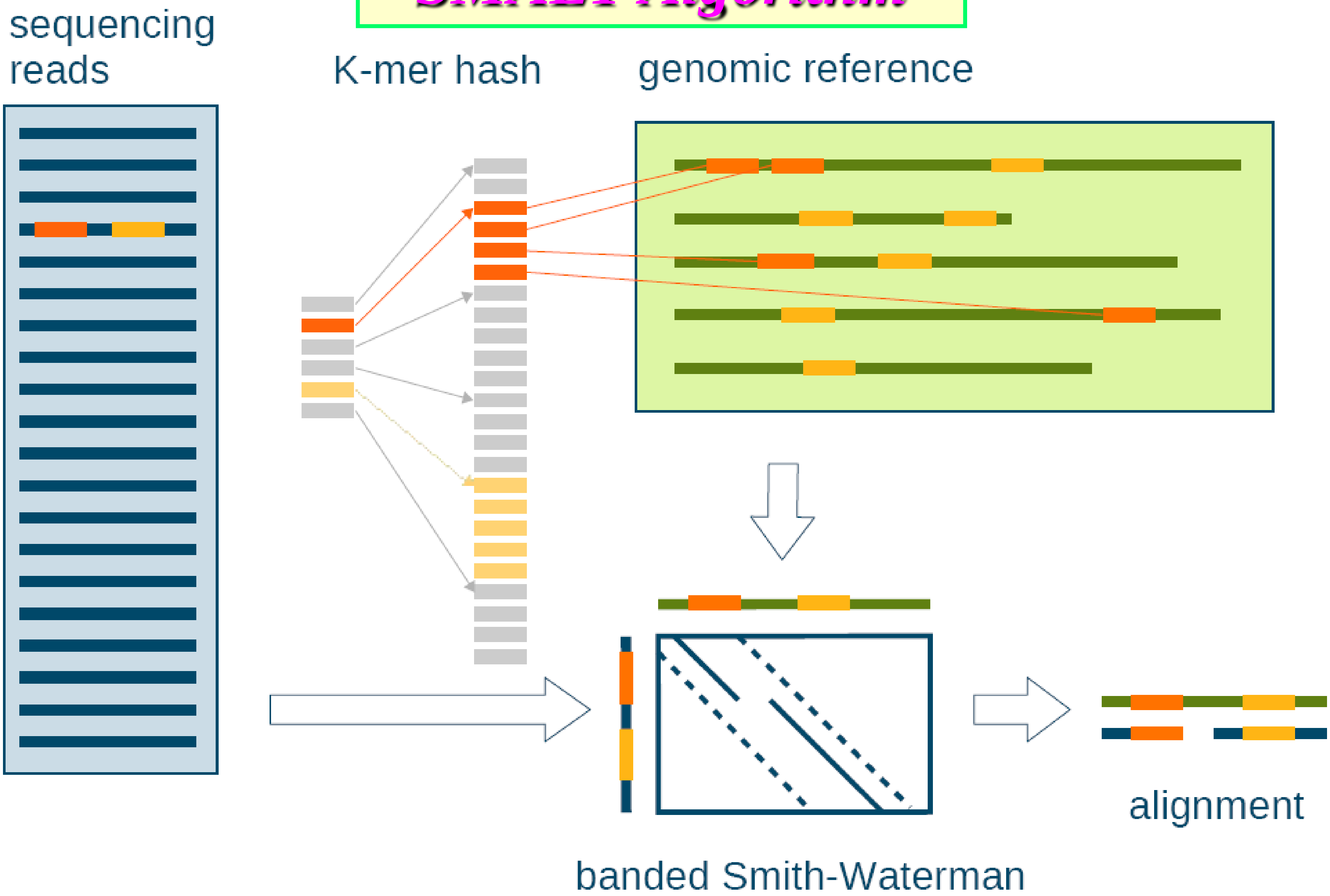
- Step 4a: Get multiple hits:

```
bwa samse -n 100 ref.fa read1.sai read1.fq
```

- Step 2b: Use BWA-SW for long reads:

```
bwa bwasw ref.fa long-read.fq > aln-long.sam
```

SMALT Algorithm



Non-overlap Hashing v Overlap Hashing

ATGGCGTGCAGTCCATGTTCGGATCA

ATGGCGTGCAGT

TGGCGTGCAGTC

GGCGTGCAGTCC

GCGTGCAGTCCA

CGTGCAGTCCAT

Overlap hashing
$W = N-k+1$
($k = 12$)

ATGGCGTGCAGTCCATGTTCGGATCATTACGTAAGC

ATGGGCAGATGT

CCATGTTCGGAT

CATTACGTAAGC

Non-overlap Hashing
$W = N/k$

Sequence Representation

Sequence S: $(s_1 s_2, \ldots, s_i, \ldots, s_m)$ $\quad i = 1, 2, \ldots, m$

K-tuple: $(s_i s_{i+1} \ldots s_{i+k-1})$

Using two binary digits for each base, we may have the following representations:

"A" =00; "C" = 01; "G" = 10; "T" = 11

For any of the m/k no-overlapping k-tuples in the sequence, an integer may be used to represent the k-tuple in a unique way

$$E = \sum_{i=1}^{2k} \beta_i 2^{i-1} \quad \text{with } E_{\max} = 2^{2\mathrm{k}} \text{ -1}$$

where β_i = 0 or 1, depending on the value of the sequence base and E_{max} is the maximum value of the possible E values.

***Hash Table**: A 2-tuple hashing table of S1, S2 and S3*

E	*k*-tuple	N_i	Indices and Offsets						
0	AA	1	2, 19						
1	AC	3	1, 9	2, 5	2, 11				
2	AG	2	1, 15	2, 35					
3	AT	2	2, 13	3, 3					
4	CA	7	2, 3	2, 9	2, 21	2, 27	2, 33	3, 21	3, 23
5	CC	4	1, 21	2, 31	3, 5	3, 7			
6	CG	1	1, 5						
7	CT	6	1, 23	2, 39	2, 43	3, 13	3, 15	3, 17	
8	GA	4	1, 3	1, 17	2, 15	2, 25			
9	GC	0							
10	GG	5	1, 25	1, 31	2, 17	2, 29	3, 1		
11	GT	6	1, 1	1, 27	1, 29	2, 1	2, 37	3, 19	
12	TA	1	3, 25						
13	TC	6	1, 7	1, 11	1, 19	2, 23	2, 41	3, 11	
14	TG	3	1, 13	2, 7	3, 9				
15	TT								

S1=(GTGACGTCACTCTGAGGATCCCCTGGGTGTGG)
S2=(GTCAACTGCAACATGAGGAACATCGACAGGCCCAAGGTCTTCCT)
S3=(GGATCCCCTGTCCTCTCTGTCACATA)

Query sequence: S_q = *(TGCAACAT)*

E	*k*-tuple	N_i	Indices and Offsets						
0	AA	1	2, 19						
1	AC	3	1, 9	2, 5	2, 11				
2	AG	2	1, 15	2, 35					
3	AT	2	2, 13	3, 3					
4	CA	7	2, 3	2, 9	2, 21	2, 27	2, 33	3, 21	3, 23
5	CC	4	1, 21	2, 31	3, 5	3, 7			
6	CG	1	1, 5						
7	CT	6	1, 23	2, 39	2, 43	3, 13	3, 15	3, 17	
8	GA	4	1, 3	1, 17	2, 15	2, 25			
9	GC	0							
10	GG	5	1, 25	1, 31	2, 17	2, 29	3, 1		
11	GT	6	1, 1	1, 27	1, 29	2, 1	2, 37	3, 19	
12	TA	1	3, 25						
13	TC	6	1, 7	1, 11	1, 19	2, 23	2, 41	3, 11	
14	TG	3	1, 13	2, 7	3, 9				
15	TT								

Query sequence:

$S_q = (TGCAACAT)$

Array of index and offset data

k-tuples	*f(t)*	*F(t)*	*-(t-1)*	$F_s(t)$
TG	1, 13	1, 13	0	1, 5
	2, 7	2, 7	0	1, 13
	3, 9	3, 9	0	2, -2
GC			-1	
CA	2, 3	2, 1	-2	2, 1
	2, 9	2, 7	-2	2, 1
	2, 21	2, 19	-2	2, 4
	2, 27	2, 25	-2	2, 7
	2, 33	2, 31	-2	2, 7
	3, 21	3, 19	-2	2, 7
	3, 23	3, 21	-2	2, 7
AA	2, 19	2, 16	-3	2, 16
AC	1, 9	1, 5	-4	2, 16
	2, 5	2, 1	-4	2, 19
	2, 11	2, 7	-4	2, 21
CA	2, 3	2, -2	-5	2, 25
	2, 9	2, 4	-5	2, 28
	2, 21	2, 16	-5	2, 31
	2, 27	2, 22	-5	3, -3
	2, 33	2, 28	-5	3, 9
	3, 21	3, 16	-5	3, 16
	3, 23	3, 18	-5	3, 18
AT	2, 13	2, 7	-6	3, 19
	3, 3	3, -3	-6	3, 21

Running SMALT

- Data files: genome_ref.fa, read1.fastq, read2.fastq
- Hash the reference genome:

  smalt index –k 13 –s 6 hash_ref genome_ref.fa

- Generate alignments in the SAM format:

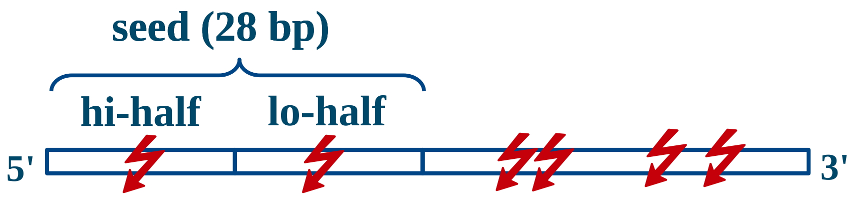
  smalt map -i 800 –j 20 –o aln.sam -f samsoft hash_ref read1.fastq read2.fq

- Where to download:

  http://www.sanger.ac.uk/resources/software/smalt/

Burrows-Wheeler *vs* Hashing

BOWTIE/TOPHAT



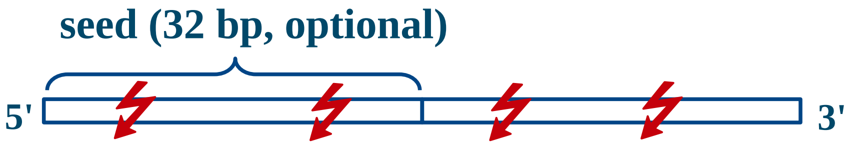
seed (28 bp)

hi-half lo-half

5' 3'

depth-1st by default, breadth-1st slower
no indels

BWA

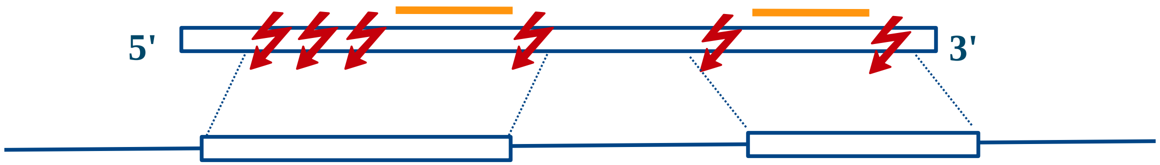


seed (32 bp, optional)

5' 3'

breadth first,
upper bound on edit distance, e.g. max 5 mismatches in 100bp read. Can deal with indels.

SMALT/SSAHA2



5' 3'

Exact matching k-segment (1 kmer) required.
Partial alignments (indels, splice junctions)

Burrows-Wheeler *vs* Hashing

- Strengths and weaknesses (trends)
  - Burrows-Wheeler, e.g. bwa, bowtie
    - Fast, esp. (multiple) exact matches
    - High sensitivity at repetitive regions
    - less robust at high genomic variation
  - Hashing (overlapping k-mer words, e.g SMALT/SSAHA2, Stampy)
    - Slower (more memory hungry)
    - Less sensitivity at repetitive regions
    - tolerate high genomic variation
    - partial alignments (junction reads) easier
    - Flexible (multiple sequencing platforms)

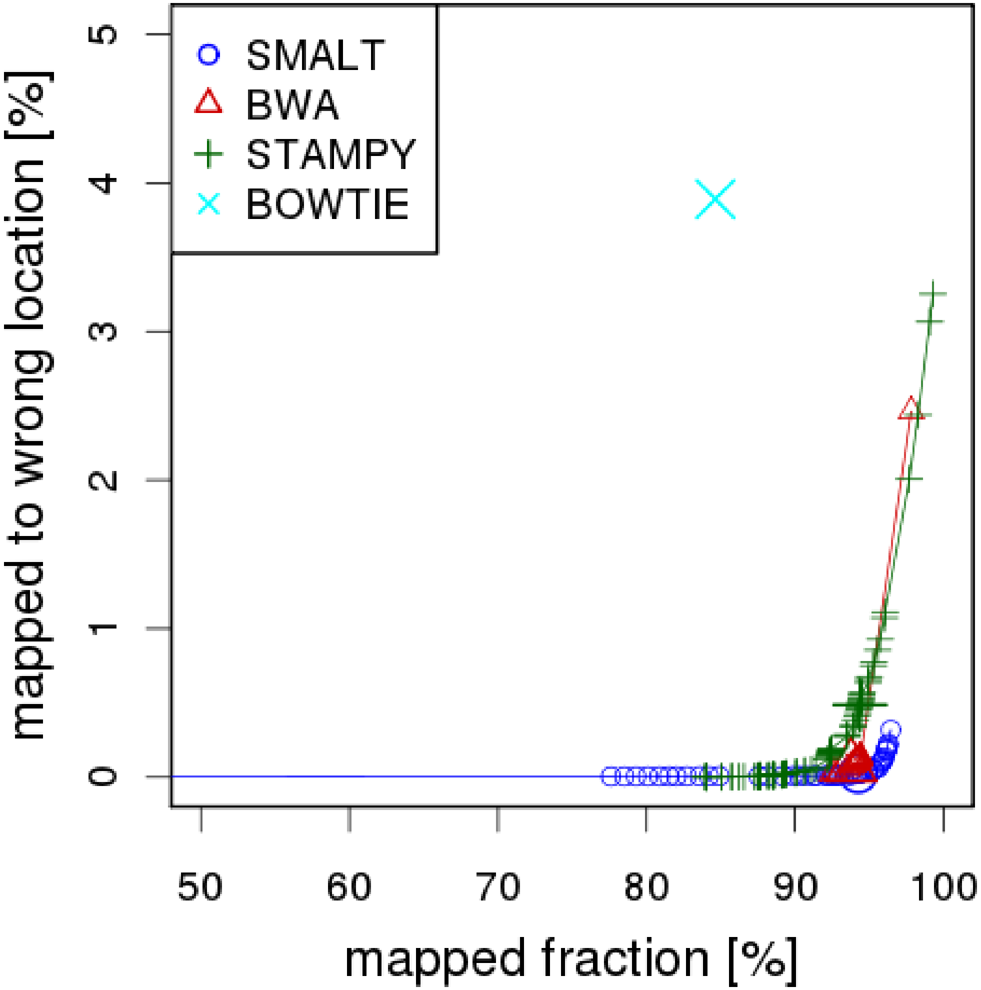
Performance Assessment

on simulated reads

	SSAHA2			SMALT			BWA		
variation	1%	2%	5%	1%	2%	5%	1%	2%	5%
rate [10^6 pairs/h]	0.34	0.22	0.18	0.71	0.60	0.47	1.35	0.74	0.63
memory [GB]	3.8	3.8	3.8	3.3	3.3	3.3	2.3	2.3	2.3
mapped [%]	97.3	97.2	96.1	97.1	97.0	96.5	95.6	89.1	48.1
error [%]	0.09	0.16	0.49	0.08	0.14	0.44	0.09	0.17	0.41

human genome
10^5 read pairs 2 x 100 bp (insert 500)
20% of variations indels (max. 10)

Performance of mappers
(genome re-sequencing)

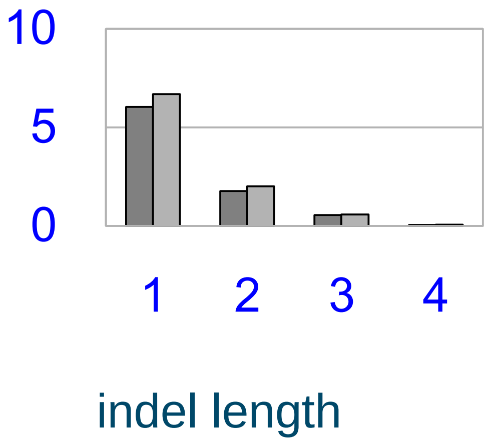


Simulated for human genome:

4×10^6 x 100 bp single reads

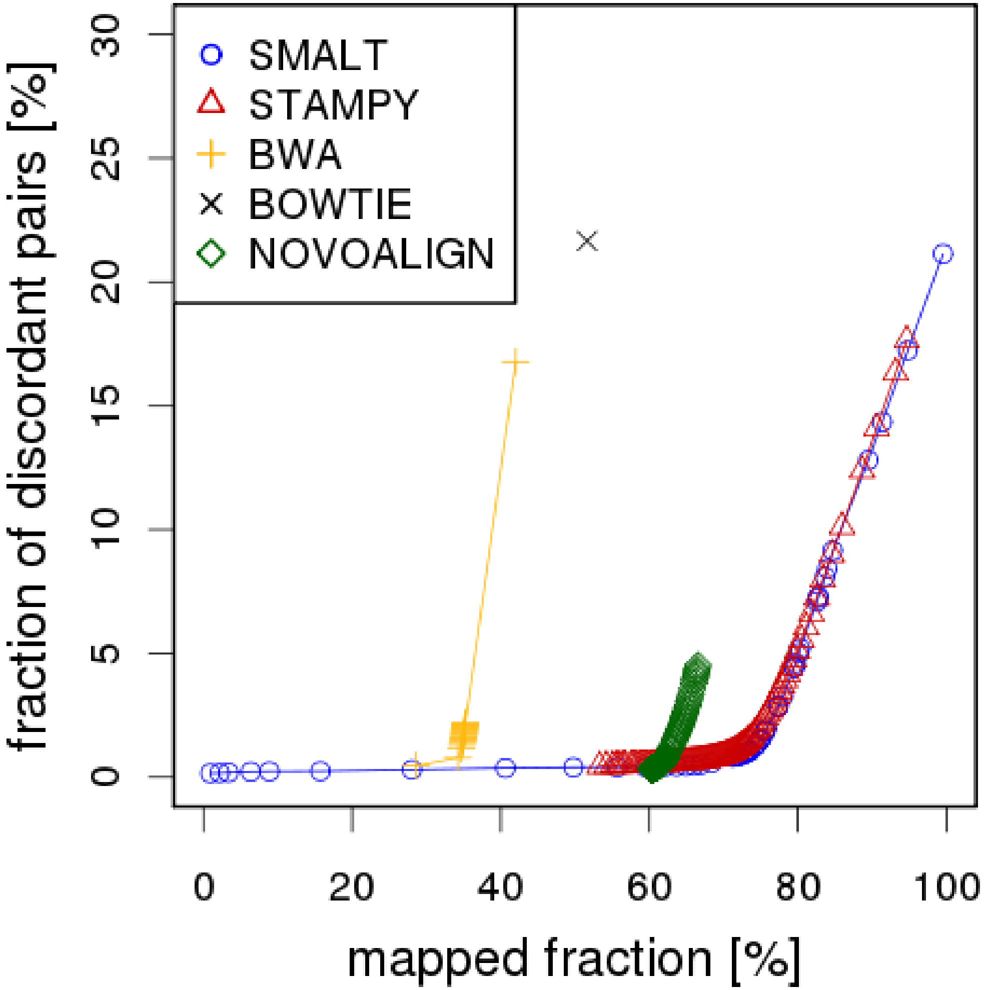
1% variation of which 20% indels

14 bp maximum indel length



Sensitivity Assessment

~ 2% genomic variation

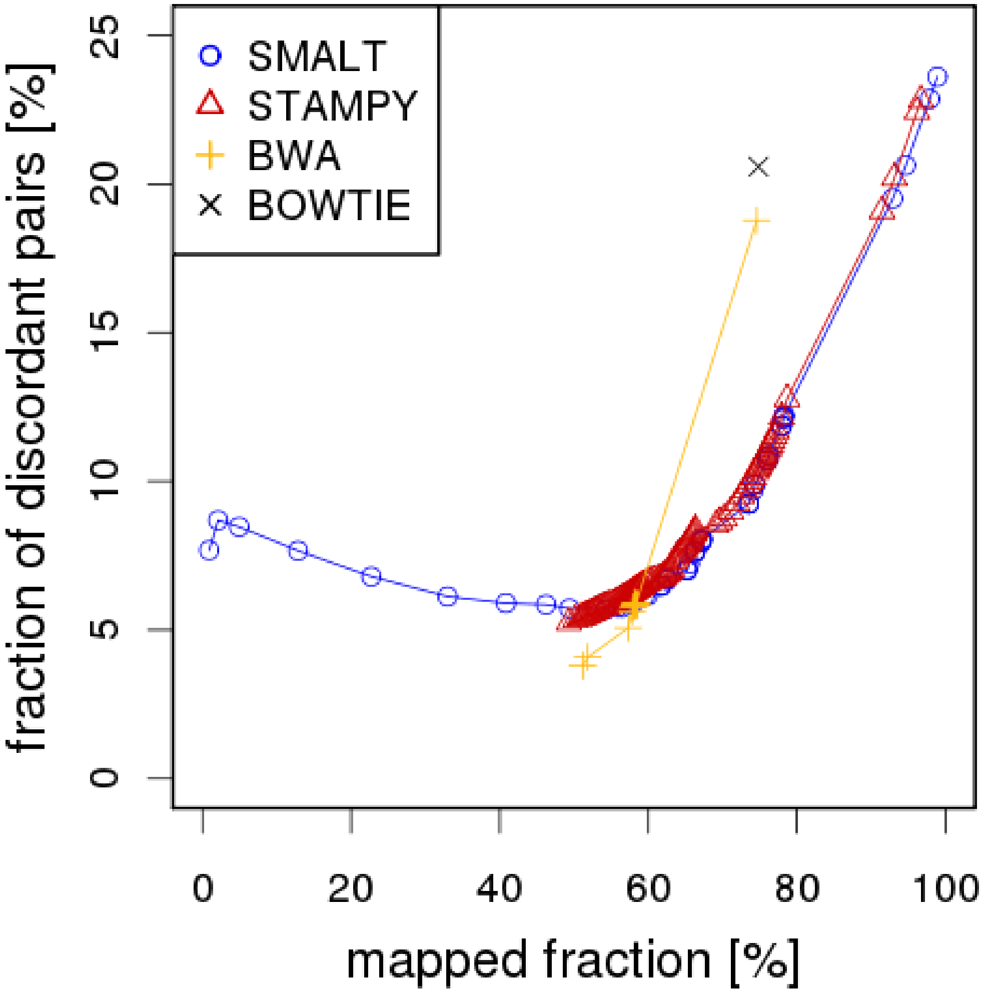


Reads:
M. spretus
whole genome shotgun
2 x 108 bp, insert 250 bp
Reference:
M. musculus
NCBI build 37

- **independently mapped reads**
- **Count discordant pairs as erroneous mappings**

Sensitivity Assessment

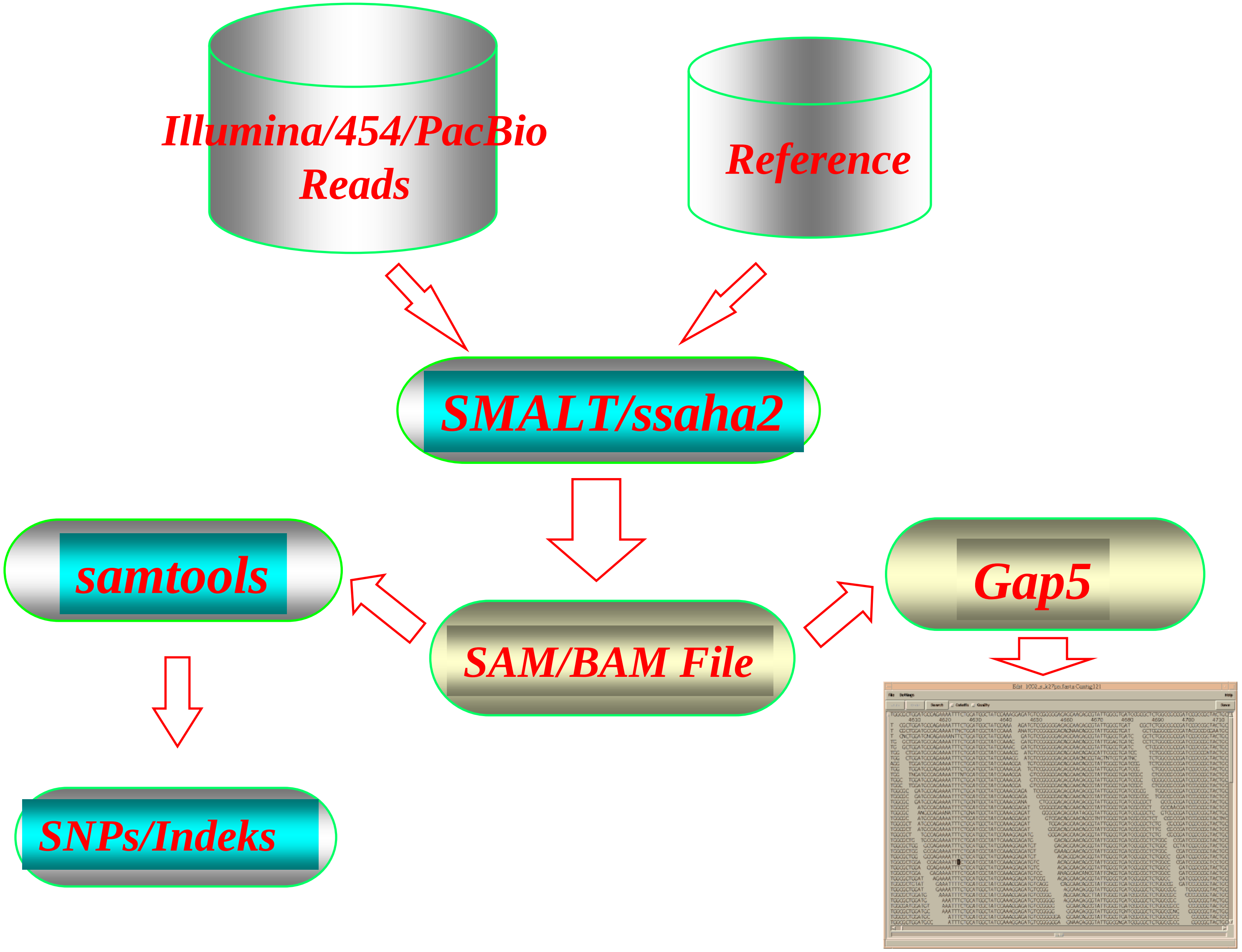
RNA-Seq data



Reads:
M. musculus C57B6
RNA-Seq data
2 x 76 bp, insert 320 bp
Reference:
M. musculus
NCBI build 37

- **independently mapped reads**
- **count discordant pairs (> 10 kbp apart) as erroneous mappings**

SMALT Workflow



Illumina/454/PacBio Reads

Reference

SMALT/ssaha2

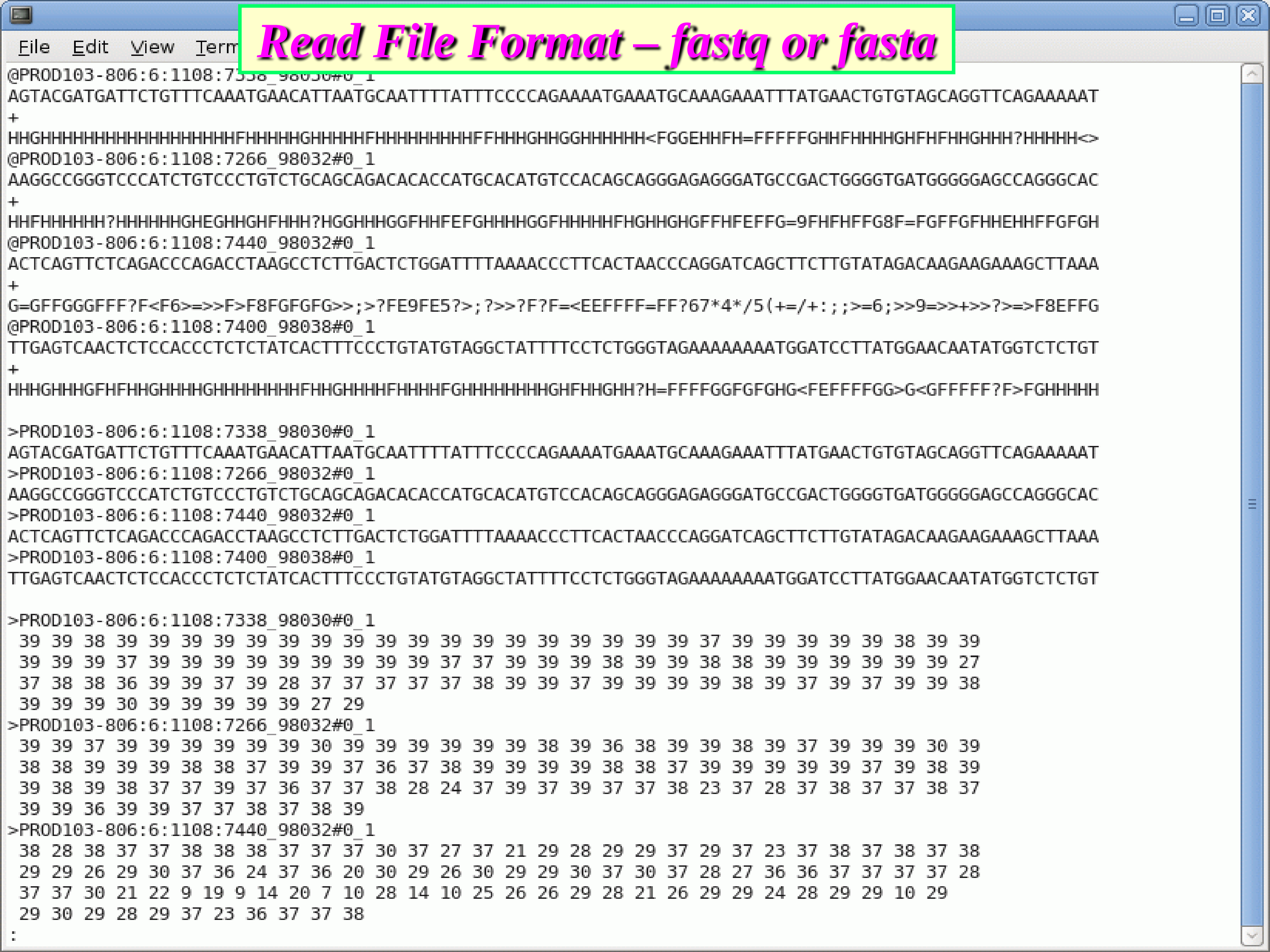
SAM/BAM File

samtools

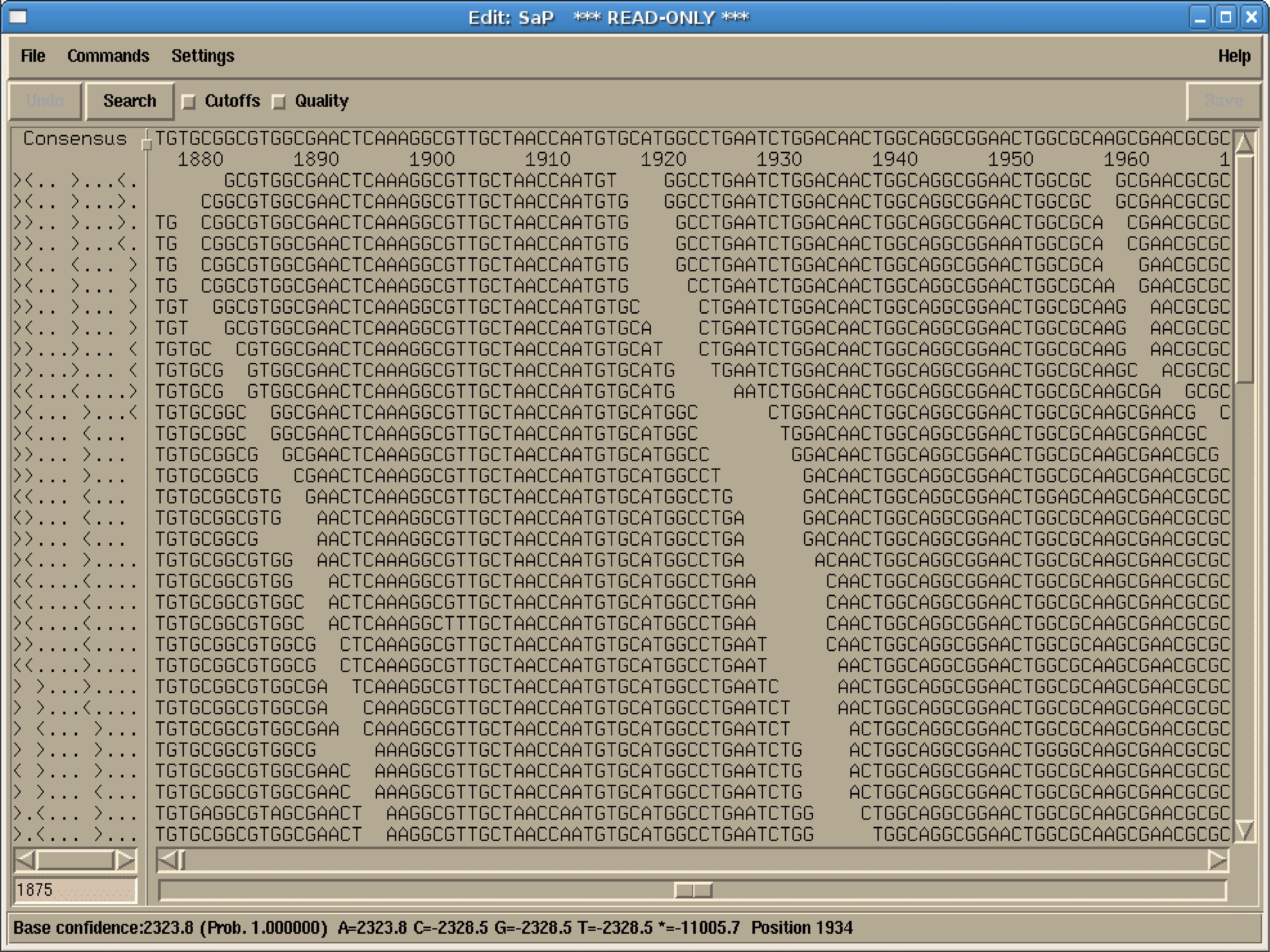
SNPs/Indeks

Gap5

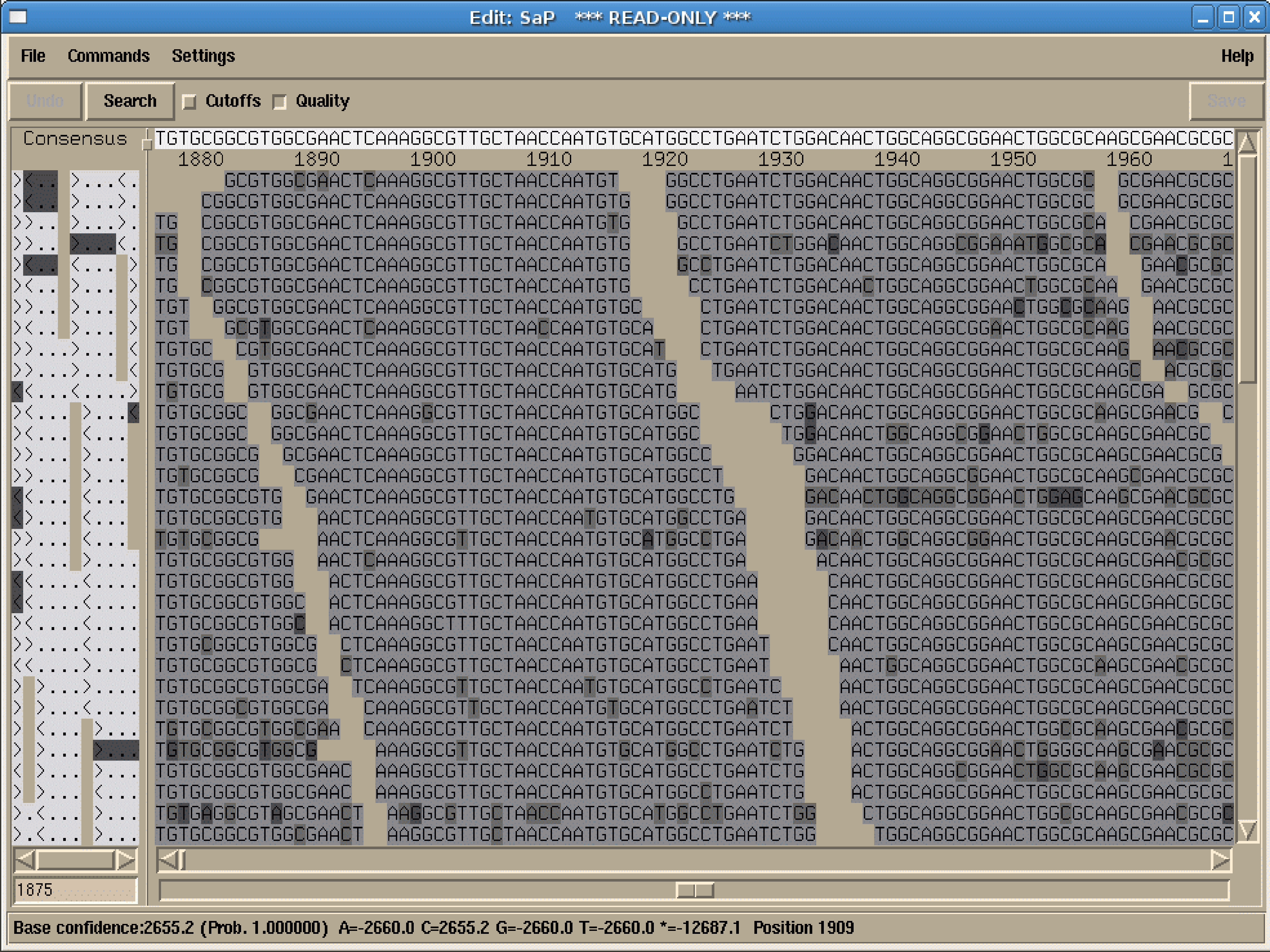
Read File Format – fastq or fasta



```
@PROD103-806:6:1108:7338_98030#0_1
AGTACGATGATTCTGTTTCAAATGAACATTAATGCAATTTTATTTCCCCAGAAAATGAAATGCAAAGAAATTTATGAACTGTGTAGCAGGTTCAGAAAAAT
+
HHGHHHHHHHHHHHHHHHHHHFHHHHHGHHHHHFHHHHHHHHHHFFHHHGHHGGHHHHHH<FGGEHHFH=FFFFFGHHFHHHHGHFHFHHGHHH?HHHHH<>
@PROD103-806:6:1108:7266_98032#0_1
AAGGCCGGGTCCCATCTGTCCCTGTCTGCAGCAGACACACCATGCACATGTCCACAGCAGGGAGAGGGATGCCGACTGGGGTGATGGGGGAGCCAGGGCAC
+
HHFHHHHHH?HHHHHHGHEGHHGHFHHH?HGGHHHGGFHHFEFGHHHHGGFHHHHHFHGHHGHGFFHFEFFG=9FHFHFFG8F=FGFFGFHHEHHFFGFGH
@PROD103-806:6:1108:7440_98032#0_1
ACTCAGTTCTCAGACCCAGACCTAAGCCTCTTGACTCTGGATTTTAAAACCCTTCACTAACCCAGGATCAGCTTCTTGTATAGACAAGAAGAAAGCTTAAA
+
G=GFFGGGFFF?F<F6>=>>F>F8FGFGFG>>;>?FE9FE5?>;?>>?F?F=<EEFFFF=FF?67*4*/5(+=/+:;;>=6;>>9=>>+>>?>=>F8EFFG
@PROD103-806:6:1108:7400_98038#0_1
TTGAGTCAACTCTCCACCCTCTCTATCACTTTCCCTGTATGTAGGCTATTTTCCTCTGGGTAGAAAAAAAAATGGATCCTTATGGAACAATATGGTCTCTGT
+
HHHGHHHGFHFHHGHHHHGHHHHHHHHHFHHGHHHHFHHHHFGHHHHHHHHHGHFHHGHH?H=FFFFGGFGFGHG<FEFFFFGG>G<GFFFFF?F>FGHHHHH

>PROD103-806:6:1108:7338_98030#0_1
AGTACGATGATTCTGTTTCAAATGAACATTAATGCAATTTTATTTCCCCAGAAAATGAAATGCAAAGAAATTTATGAACTGTGTAGCAGGTTCAGAAAAAT
>PROD103-806:6:1108:7266_98032#0_1
AAGGCCGGGTCCCATCTGTCCCTGTCTGCAGCAGACACACCATGCACATGTCCACAGCAGGGAGAGGGATGCCGACTGGGGTGATGGGGGAGCCAGGGCAC
>PROD103-806:6:1108:7440_98032#0_1
ACTCAGTTCTCAGACCCAGACCTAAGCCTCTTGACTCTGGATTTTAAAACCCTTCACTAACCCAGGATCAGCTTCTTGTATAGACAAGAAGAAAGCTTAAA
>PROD103-806:6:1108:7400_98038#0_1
TTGAGTCAACTCTCCACCCTCTCTATCACTTTCCCTGTATGTAGGCTATTTTCCTCTGGGTAGAAAAAAAAATGGATCCTTATGGAACAATATGGTCTCTGT

>PROD103-806:6:1108:7338_98030#0_1
 39 39 38 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 37 39 39 39 39 39 38 39 39
 39 39 39 37 39 39 39 39 39 39 39 39 39 37 37 39 39 39 38 39 39 38 38 39 39 39 39 39 39 27
 37 38 38 36 39 39 37 39 28 37 37 37 37 37 38 39 39 37 39 39 39 39 38 39 37 39 37 39 39 38
 39 39 39 30 39 39 39 39 39 27 29
>PROD103-806:6:1108:7266_98032#0_1
 39 39 37 39 39 39 39 39 39 30 39 39 39 39 39 39 38 39 36 38 39 39 38 39 37 39 39 39 30 39
 38 38 39 39 39 38 38 37 39 39 37 36 37 38 39 39 39 39 38 38 37 39 39 39 39 39 37 39 38 39
 39 38 39 38 37 37 39 37 36 37 37 38 28 24 37 39 37 39 37 37 38 23 37 28 37 38 37 37 38 37
 39 39 36 39 39 37 37 38 37 38 39
>PROD103-806:6:1108:7440_98032#0_1
 38 28 38 37 37 38 38 38 37 37 37 30 37 27 37 21 29 28 29 29 37 29 37 23 37 38 37 38 37 38
 29 29 26 29 30 37 36 24 37 36 20 30 29 26 30 29 29 30 37 30 37 28 27 36 36 37 37 37 37 28
 37 37 30 21 22 9 19 9 14 20 7 10 28 14 10 25 26 26 29 28 21 26 29 29 24 28 29 29 10 29
 29 30 29 28 29 37 23 36 37 37 38
:
```

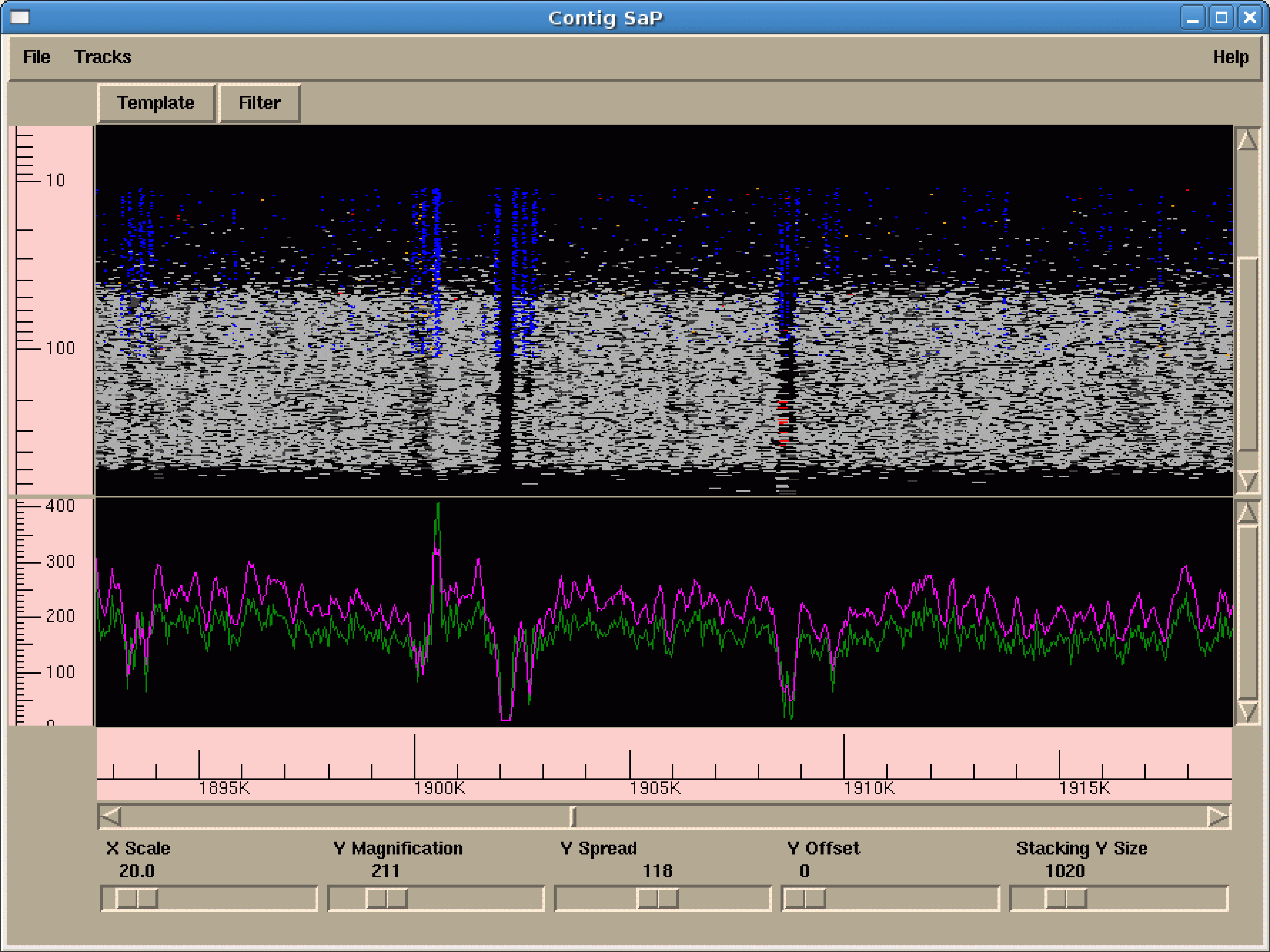


```
 Consensus  TGTGCGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
              1880      1890      1900      1910      1920      1930      1940      1950      1960      1
><.. >...<.       GCGTGGCGAACTCAAAGGCGTTGCTAACCAATGT    GGCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGC  GCGAACGCGC
><.. >...>.     CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG   GGCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGC  GCGAACGCGC
>>.. >...>. TG  CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG    GCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGCA  CGAACGCGC
>>.. >...<. TG  CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG    GCCTGAATCTGGACAACTGGCAGGCGGAAATGGCGCA  CGAACGCGC
><.. <... > TG  CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG    GCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGCA   GAACGCGC
><.. >... > TG  CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG     CCTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAA  GAACGCGC
>>.. >... > TGT  GGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTGC     CTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAG  AACGCGC
><.. >... > TGT   GCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCA    CTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAG  AACGCGC
>>...>... < TGTGC  CGTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCAT   CTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAG  AACGCGC
>>...>... < TGTGCG  GTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCATG   TGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAGC  ACGCGC
<<...<....> TGTGCG  GTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCATG     AATCTGGACAACTGGCAGGCGGAACTGGCGCAAGCGA  GCGC
><... >...< TGTGCGGC  GGCGAACTCAAAGGCGTTGCTAACCAATGTGCATGGC      CTGGACAACTGGCAGGCGGAACTGGCGCAAGCGAACG  C
><... <...  TGTGCGGC  GGCGAACTCAAAGGCGTTGCTAACCAATGTGCATGGC       TGGACAACTGGCAGGCGGAACTGGCGCAAGCGAACGC
>>... >...  TGTGCGGCG  GCGAACTCAAAGGCGTTGCTAACCAATGTGCATGGCC       GGACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCG
>>... >...  TGTGCGGCG   CGAACTCAAAGGCGTTGCTAACCAATGTGCATGGCCT       GACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
<<... <...  TGTGCGGCGTG  GAACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTG      GACAACTGGCAGGCGGAACTGGAGCAAGCGAACGCGC
<>... <...  TGTGCGGCGTG   AACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGA     GACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
>>... <...  TGTGCGGCG     AACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGA     GACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
><... >.... TGTGCGGCGTGG  AACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGA      ACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
<<....<.... TGTGCGGCGTGG   ACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAA      CAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
<<....<.... TGTGCGGCGTGGC  ACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAA      CAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
><....<.... TGTGCGGCGTGGC  ACTCAAAGGCTTTGCTAACCAATGTGCATGGCCTGAA      CAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
>>....<.... TGTGCGGCGTGGCG  CTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAAT     CAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
<<....>.... TGTGCGGCGTGGCG  CTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAAT      AACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
> >...>.... TGTGCGGCGTGGCGA  TCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAATC     AACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
> >...<.... TGTGCGGCGTGGCGA   CAAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCT    AACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
> <... >... TGTGCGGCGTGGCGAA  CAAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCT     ACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
> >... >... TGTGCGGCGTGGCG     AAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTG    ACTGGCAGGCGGAACTGGGGCAAGCGAACGCGC
< >... >... TGTGCGGCGTGGCGAAC  AAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTG    ACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
> >... <... TGTGCGGCGTGGCGAAC  AAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTG    ACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
>.<... >... TGTGAGGCGTAGCGAACT  AAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTGG    CTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
>.<... >... TGTGCGGCGTGGCGAACT  AAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTGG     TGGCAGGCGGAACTGGCGCAAGCGAACGCGC

1875

Base confidence:2323.8 (Prob. 1.000000)  A=2323.8 C=-2328.5 G=-2328.5 T=-2328.5 *=-11005.7  Position 1934
```



```
Consensus  TGTGCGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
             1880      1890      1900      1910      1920      1930      1940      1950      1960      1
                 GCGTGGCGAACTCAAAGGCGTTGCTAACCAATGT    GGCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGC  GCGAACGCGC
               CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG   GGCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGC  GCGAACGCGC
           TG  CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG    GCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGCA  CGAACGCGC
           TG  CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG    GCCTGAATCTGGACAACTGGCAGGCGGAAATGGCGCA  CGAACGCGC
           TG  CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG    GCCTGAATCTGGACAACTGGCAGGCGGAACTGGCGCA   GAACGCGC
           TG  CGGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTG     CCTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAA  GAACGCGC
           TGT  GGCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTGC     CTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAG  AACGCGC
           TGT   GCGTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCA    CTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAG  AACGCGC
           TGTGC  CGTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCAT   CTGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAG  AACGCGC
           TGTGCG  GTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCATG   TGAATCTGGACAACTGGCAGGCGGAACTGGCGCAAGC  ACGCGC
           TGTGCG  GTGGCGAACTCAAAGGCGTTGCTAACCAATGTGCATG     AATCTGGACAACTGGCAGGCGGAACTGGCGCAAGCGA  GCGC
           TGTGCGGC  GGCGAACTCAAAGGCGTTGCTAACCAATGTGCATGGC      CTGGACAACTGGCAGGCGGAACTGGCGCAAGCGAACG  C
           TGTGCGGC  GGCGAACTCAAAGGCGTTGCTAACCAATGTGCATGGC       TGGACAACTGGCAGGCGGAACTGGCGCAAGCGAACGC
           TGTGCGGCG  GCGAACTCAAAGGCGTTGCTAACCAATGTGCATGGCC       GGACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCG
           TGTGCGGCG   CGAACTCAAAGGCGTTGCTAACCAATGTGCATGGCCT       GACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTG  GAACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTG      GACAACTGGCAGGCGGAACTGGAGCAAGCGAACGCGC
           TGTGCGGCGTG   AACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGA     GACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCG     AACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGA     GACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGG  AACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGA      ACAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGG   ACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAA      CAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGC  ACTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAA      CAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGC  ACTCAAAGGCTTTGCTAACCAATGTGCATGGCCTGAA      CAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGCG  CTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAAT     CAACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGCG  CTCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAAT      AACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGCGA  TCAAAGGCGTTGCTAACCAATGTGCATGGCCTGAATC     AACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGCGA   CAAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCT    AACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGCGAA  CAAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCT     ACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGCG     AAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTG    ACTGGCAGGCGGAACTGGGGCAAGCGAACGCGC
           TGTGCGGCGTGGCGAAC  AAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTG    ACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGCGAAC  AAAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTG    ACTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGAGGCGTAGCGAACT  AAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTGG    CTGGCAGGCGGAACTGGCGCAAGCGAACGCGC
           TGTGCGGCGTGGCGAACT  AAGGCGTTGCTAACCAATGTGCATGGCCTGAATCTGG     TGGCAGGCGGAACTGGCGCAAGCGAACGCGC
1875
```

Base confidence:2655.2 (Prob. 1.000000) A=-2660.0 C=2655.2 G=-2660.0 T=-2660.0 *=-12687.1 Position 1909



X Scale 20.0

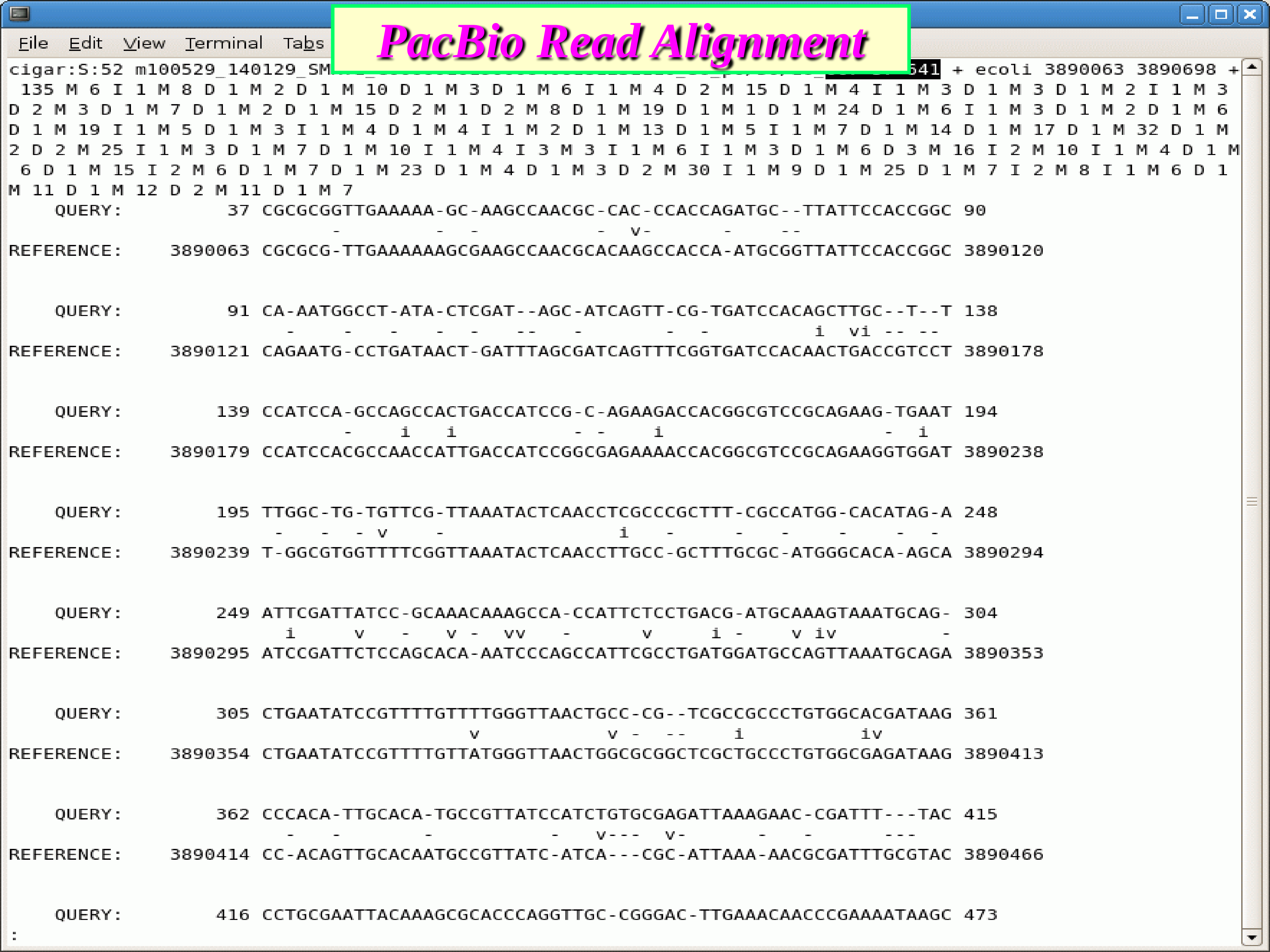
Y Magnification 211

Y Spread 118

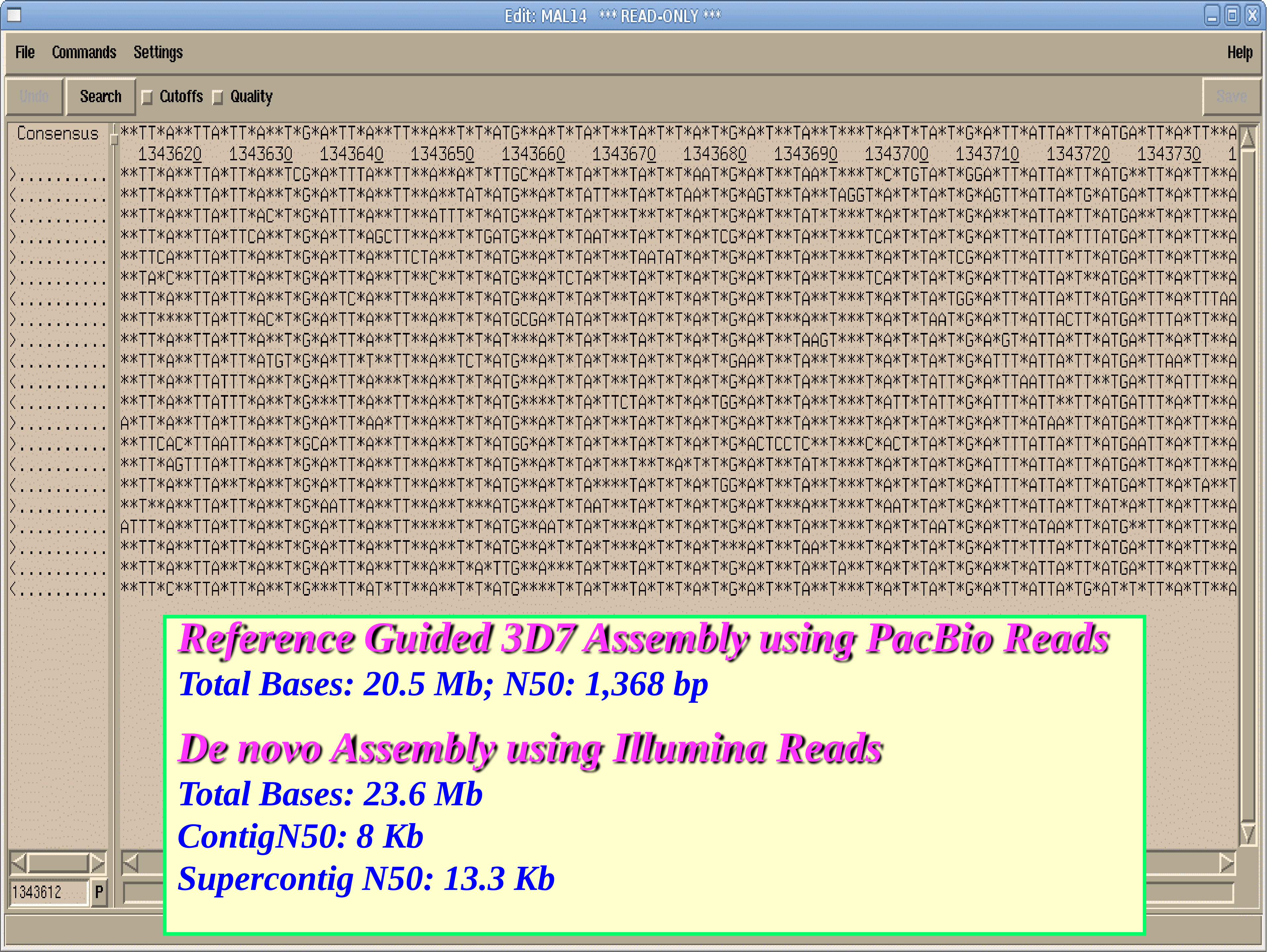
Y Offset 0

Stacking Y Size 1020

PacBio Read Alignment



```
cigar:S:52 m100529_140129_SM                                         641 + ecoli 3890063 3890698 +
 135 M 6 I 1 M 8 D 1 M 2 D 1 M 10 D 1 M 3 D 1 M 6 I 1 M 4 D 2 M 15 D 1 M 4 I 1 M 3 D 1 M 3 D 1 M 2 I 1 M 3
D 2 M 3 D 1 M 7 D 1 M 2 D 1 M 15 D 2 M 1 D 2 M 8 D 1 M 19 D 1 M 1 D 1 M 24 D 1 M 6 I 1 M 3 D 1 M 2 D 1 M 6
D 1 M 19 I 1 M 5 D 1 M 3 I 1 M 4 D 1 M 4 I 1 M 2 D 1 M 13 D 1 M 5 I 1 M 7 D 1 M 14 D 1 M 17 D 1 M 32 D 1 M
2 D 2 M 25 I 1 M 3 D 1 M 7 D 1 M 10 I 1 M 4 I 3 M 3 I 1 M 6 I 1 M 3 D 1 M 6 D 3 M 16 I 2 M 10 I 1 M 4 D 1 M
 6 D 1 M 15 I 2 M 6 D 1 M 7 D 1 M 23 D 1 M 4 D 1 M 3 D 2 M 30 I 1 M 9 D 1 M 25 D 1 M 7 I 2 M 8 I 1 M 6 D 1
M 11 D 1 M 12 D 2 M 11 D 1 M 7
    QUERY:         37 CGCGCGGTTGAAAAA-GC-AAGCCAACGC-CAC-CCACCAGATGC--TTATTCCACCGGC 90
                            -        -  -          -  v-      -    --
REFERENCE:    3890063 CGCGCG-TTGAAAAAAGCGAAGCCAACGCACAAGCCACCA-ATGCGGTTATTCCACCGGC 3890120


    QUERY:         91 CA-AATGGCCT-ATA-CTCGAT--AGC-ATCAGTT-CG-TGATCCACAGCTTGC--T--T 138
                        -    -   -   -  -   --   -       -  -         i  vi -- --
REFERENCE:    3890121 CAGAATG-CCTGATAACT-GATTTAGCGATCAGTTTCGGTGATCCACAACTGACCGTCCT 3890178


    QUERY:        139 CCATCCA-GCCAGCCACTGACCATCCG-C-AGAAGACCACGGCGTCCGCAGAAG-TGAAT 194
                             -    i   i          - -    i                   -  i
REFERENCE:    3890179 CCATCCACGCCAACCATTGACCATCCGGCGAGAAAACCACGGCGTCCGCAGAAGGTGGAT 3890238


    QUERY:        195 TTGGC-TG-TGTTCG-TTAAATACTCAACCTCGCCCGCTTT-CGCCATGG-CACATAG-A 248
                       -   -  - v    -               i   -     -   -    -    -  -
REFERENCE:    3890239 T-GGCGTGGTTTTCGGTTAAATACTCAACCTTGCC-GCTTTGCGC-ATGGGCACA-AGCA 3890294


    QUERY:        249 ATTCGATTATCC-GCAAACAAAGCCA-CCATTCTCCTGACG-ATGCAAAGTAAATGCAG- 304
                        i     v   -   v -  vv   -      v     i -    v iv         -
REFERENCE:    3890295 ATCCGATTCTCCAGCACA-AATCCCAGCCATTCGCCTGATGGATGCCAGTTAAATGCAGA 3890353


    QUERY:        305 CTGAATATCCGTTTTGTTTTGGGTTAACTGCC-CG--TCGCCGCCCTGTGGCACGATAAG 361
                                        v           v -  --    i          iv
REFERENCE:    3890354 CTGAATATCCGTTTTGTTATGGGTTAACTGGCGCGGCTCGCTGCCCTGTGGCGAGATAAG 3890413


    QUERY:        362 CCCACA-TTGCACA-TGCCGTTATCCATCTGTGCGAGATTAAAGAAC-CGATTT---TAC 415
                        -   -       -          -   v---  v-      -   -      ---
REFERENCE:    3890414 CC-ACAGTTGCACAATGCCGTTATC-ATCA---CGC-ATTAAA-AACGCGATTTGCGTAC 3890466


    QUERY:        416 CCTGCGAATTACAAAGCGCACCCAGGTTGC-CGGGAC-TTGAAACAACCCGAAAATAAGC 473
:
```



```
Consensus   **TT*A**TTA*TT*A**T*G*A*TT*A**TT**A**T*T*ATG**A*T*TA*T**TA*T*T*A*T*G*A*T**TA**T***T*A*T*TA*T*G*A*TT*ATTA*TT*ATGA*TT*A*TT**A
              1343620   1343630   1343640   1343650   1343660   1343670   1343680   1343690   1343700   1343710   1343720   1343730   1
>.........  **TT*A**TTA*TT*A**TCG*A*TTTA**TT**A**A*T*TTGC*A*T*TA*T**TA*T*T*AAT*G*A*T**TAA*T***T*C*TGTA*T*GGA*TT*ATTA*TT*ATG**TT*A*TT**A
<.........  **TT*A**TTA*TT*A**T*G*A*TT*A**TT**A**TAT*ATG**A*T*TATT**TA*T*TAA*T*G*AGT**TA**TAGGT*A*T*TA*T*G*AGTT*ATTA*TG*ATGA*TT*A*TT**A
<.........  **TT*A**TTA*TT*AC*T*G*ATTT*A**TT**ATTT*T*ATG**A*T*TA*T**T**T*T*A*T*G*A*T**TAT*T***T*A*T*TA*T*G*A**T*ATTA*TT*ATGA**T*A*TT**A
>.........  **TT*A**TTA*TTCA**T*G*A*TT*AGCTT**A**T*TGATG**A*T*TAAT**TA*T*T*A*TCG*A*T**TA**T***TCA*T*TA*T*G*A*TT*ATTA*TTTATGA*TT*A*TT**A
>.........  **TTCA**TTA*TT*A**T*G*A*TT*A**TTCTA**T*T*ATG**A*T*TA*T**TAATAT*A*T*G*A*T**TA**T***T*A*T*TA*TCG*A*TT*ATTT*TT*ATGA*TT*A*TT**A
>.........  **TA*C**TTA*TT*A**T*G*A*TT*A**TT**C**T*T*ATG**A*TCTA*T**TA*T*T*A*T*G*A*T**TA**T***TCA*T*TA*T*G*A*TT*ATTA*T**ATGA*TT*A*TT**A
<.........  **TT*A**TTA*TT*A**T*G*A*TC*A**TT**A**T*T*ATG**A*T*TA*T**TA*T*T*A*T*G*A*T**TA**T***T*A*T*TA*TGG*A*TT*ATTA*TT*ATGA*TT*A*TTTAA
>.........  **TT****TTA*TT*AC*T*G*A*TT*A**TT**A**T*T*ATGCGA*TATA*T**TA*T*T*A*T*G*A*T***A**T***T*A*T*TAAT*G*A*TT*ATTACTT*ATGA*TTTA*TT**A
>.........  **TT*A**TTA*TT*A**T*G*A*TT*A**TT**A**T*T*AT***A*T*TA*T**TA*T*T*A*T*G*A*T**TAAGT***T*A*T*TA*T*G*A*GT*ATTA*TT*ATGA*TT*A*TT**A
<.........  **TT*A**TTA*TT*ATGT*G*A*TT*T**TT**A**TCT*ATG**A*T*TA*T**TA*T*T*A*T*GAA*T**TA**T***T*A*T*TA*T*G*ATTT*ATTA*TT*ATGA*TTAA*TT**A
<.........  **TT*A**TTATTT*A**T*G*A*TT*A***T**A**T*T*ATG**A*T*TA*T**TA*T*T*A*T*G*A*T**TA**T***T*A*T*TATT*G*A*TTAATTA*TT**TGA*TT*ATTT**A
<.........  **TT*A**TTATTT*A**T*G***TT*A**TT**A**T*T*ATG****T*TA*TTCTA*T*T*A*TGG*A*T**TA**T***T*ATT*TATT*G*ATTT*ATT**TT*ATGATTT*A*TT**A
>.........  A*TT*A**TTA*TT*A**T*G*A*TT*AA*TT**A**T*T*ATG**A*T*TA*T**TA*T*T*A*T*G*A*T**TA**T***T*A*T*TA*T*G*A*TT*ATAA*TT*ATGA*TT*A*TT**A
>.........  **TTCAC*TTAATT*A**T*GCA*TT*A**TT**A**T*T*ATGG*A*T*TA*T**TA*T*T*A*T*G*ACTCCTC**T***C*ACT*TA*T*G*A*TTTATTA*TT*ATGAATT*A*TT**A
<.........  **TT*AGTTTA*TT*A**T*G*A*TT*A**TT**A**T*T*ATG**A*T*TA*T**T**T*A*T*T*G*A*T**TAT*T***T*A*T*TA*T*G*ATTT*ATTA*TT*ATGA*TT*A*TT**A
<.........  **TT*A**TTA**T*A**T*G*A*TT*A**TT**A**T*T*ATG**A*T*TA****TA*T*T*A*TGG*A*T**TA**T***T*A*T*TA*T*G*ATTT*ATTA*TT*ATGA*TT*A*TA**T
>.........  **T**A**TTA*TT*A**T*G*AATT*A**TT**A**T***ATG**A*T*TAAT**TA*T*T*A*T*G*A*T***A**T***T*AAT*TA*T*G*A*TT*ATTA*TT*AT*A*TT*A*TT**A
>.........  ATTT*A**TTA*TT*A**T*G*A*TT*A**TT*****T*T*ATG**AAT*TA*T***A*T*T*A*T*G*A*T**TA**T***T*A*T*TAAT*G*A*TT*ATAA*TT*ATG**TT*A*TT**A
>.........  **TT*A**TTA*TT*A**T*G*A*TT*A**TT**A**T*T*ATG**A*T*TA*T***A*T*T*A*T***A*T**TAA*T***T*A*T*TA*T*G*A*TT*TTTA*TT*ATGA*TT*A*TT**A
<.........  **TT*A**TTA**T*A**T*G*A*TT*A**TT**A**T*A*TTG**A***TA*T**TA*T*T*A*T*G*A*T**TA**TA**T*A*T*TA*T*G*A**T*ATTA*TT*ATGA*TT*A*TT**A
<.........  **TT*C**TTA*TT*A**T*G***TT*AT*TT**A**T*T*ATG****T*TA*T**TA*T*T*A*T*G*A*T**TA**T***T*A*T*TA*T*G*A*TT*ATTA*TG*AT*T*TT*A*TT**A
```

Reference Guided 3D7 Assembly using PacBio Reads

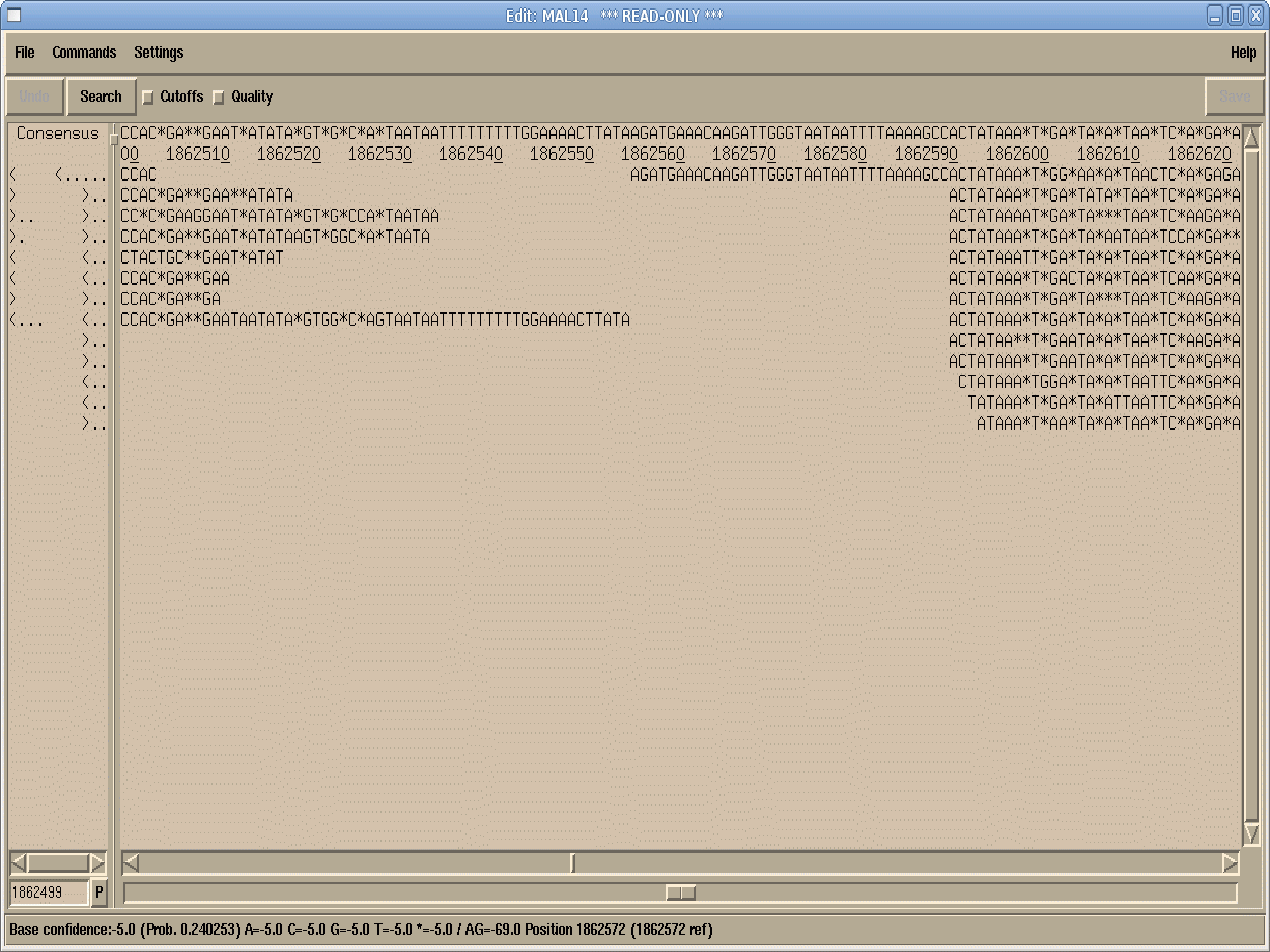
Total Bases: 20.5 Mb; N50: 1,368 bp

De novo Assembly using Illumina Reads

Total Bases: 23.6 Mb

ContigN50: 8 Kb

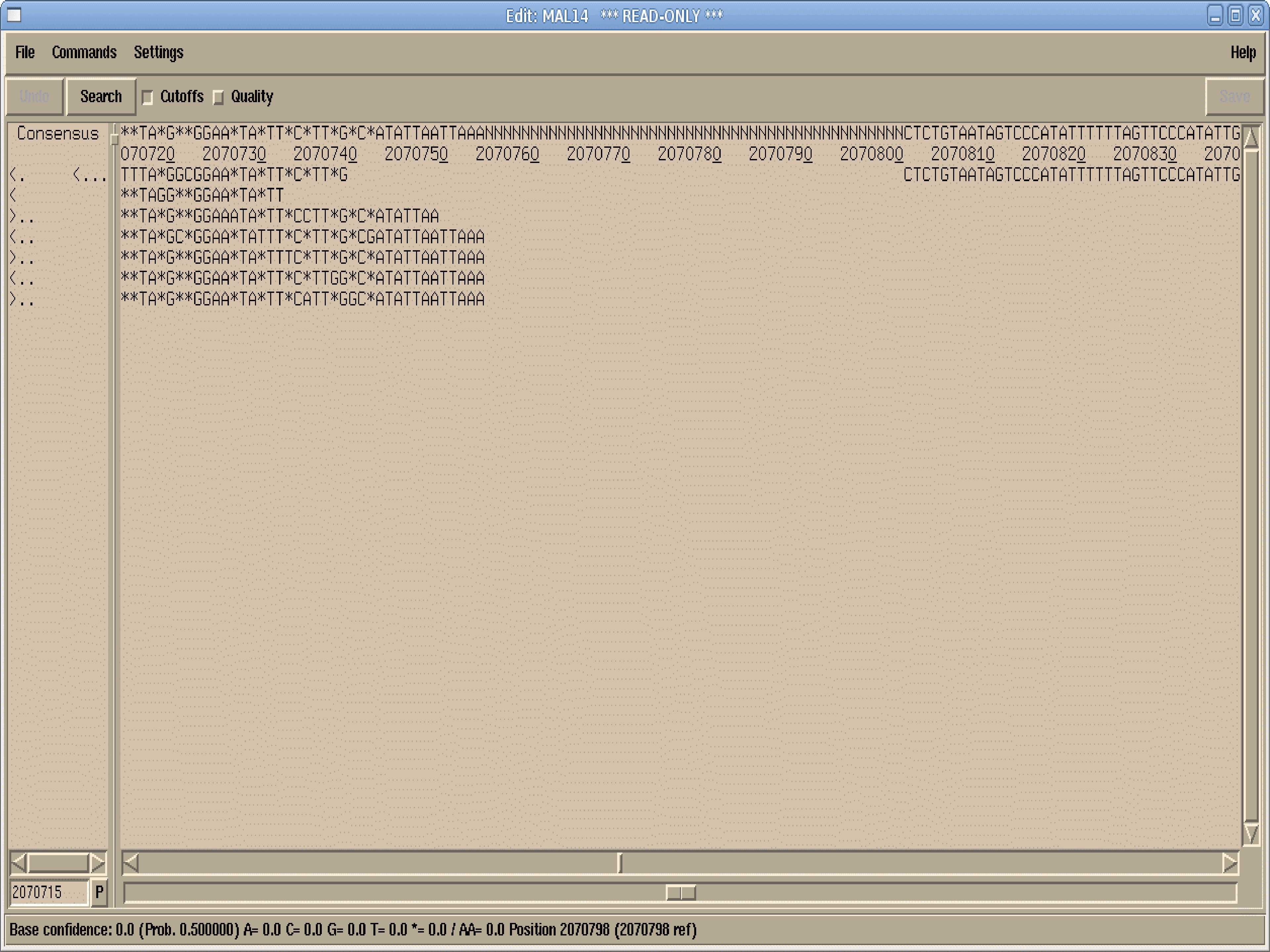
Supercontig N50: 13.3 Kb



```
  Consensus  CCAC*GA**GAAT*ATATA*GT*G*C*A*TAATAATTTTTTTTGGAAAACTTATAAGATGAAACAAGATTGGGTAATAATTTTAAAAGCCACTATAAA*T*GA*TA*A*TAA*TC*A*GA*A
             00   1862510   1862520   1862530   1862540   1862550   1862560   1862570   1862580   1862590   1862600   1862610   1862620
<    <.....  CCAC                                                   AGATGAAACAAGATTGGGTAATAATTTTAAAAGCCACTATAAA*T*GG*AA*A*TAACTC*A*GAGA
>        >.. CCAC*GA**GAA**ATATA                                                                       ACTATAAA*T*GA*TATA*TAA*TC*A*GA*A
>..      >.. CC*C*GAAGGAAT*ATATA*GT*G*CCA*TAATAA                                                       ACTATAAAAT*GA*TA***TAA*TC*AAGA*A
>.       >.. CCAC*GA**GAAT*ATATAAGT*GGC*A*TAATA                                                        ACTATAAA*T*GA*TA*AATAA*TCCA*GA**
<        <.. CTACTGC**GAAT*ATAT                                                                        ACTATAAATT*GA*TA*A*TAA*TC*A*GA*A
<        <.. CCAC*GA**GAA                                                                              ACTATAAA*T*GACTA*A*TAA*TCAA*GA*A
>        >.. CCAC*GA**GA                                                                               ACTATAAA*T*GA*TA***TAA*TC*AAGA*A
<...     <.. CCAC*GA**GAATAATATA*GTGG*C*AGTAATAATTTTTTTTGGAAAACTTATA                                   ACTATAAA*T*GA*TA*A*TAA*TC*A*GA*A
         >..                                                                                           ACTATAA**T*GAATA*A*TAA*TC*AAGA*A
         >..                                                                                           ACTATAAA*T*GAATA*A*TAA*TC*A*GA*A
         <..                                                                                            CTATAAA*TGGA*TA*A*TAATTC*A*GA*A
         <..                                                                                             TATAAA*T*GA*TA*ATTAATTC*A*GA*A
         >..                                                                                              ATAAA*T*AA*TA*A*TAA*TC*A*GA*A
```

1862499

Base confidence:-5.0 (Prob. 0.240253) A=-5.0 C=-5.0 G=-5.0 T=-5.0 *=-5.0 / AG=-69.0 Position 1862572 (1862572 ref)



```
Consensus   **TA*G**GGAA*TA*TT*C*TT*G*C*ATATTAATTAAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCTCTGTAATAGTCCCATATTTTTAGTTCCCATATTG
            070720   2070730   2070740   2070750   2070760   2070770   2070780   2070790   2070800   2070810   2070820   2070830   2070
<.    <...  TTTA*GGCGGAA*TA*TT*C*TT*G                                                             CTCTGTAATAGTCCCATATTTTTAGTTCCCATATTG
<           **TAGG**GGAA*TA*TT
>..         **TA*G**GGAAATA*TT*CCTT*G*C*ATATTAA
<..         **TA*GC*GGAA*TATTT*C*TT*G*CGATATTAATTAAA
>..         **TA*G**GGAA*TA*TTTC*TT*G*C*ATATTAATTAAA
<..         **TA*G**GGAA*TA*TT*C*TTGG*C*ATATTAATTAAA
>..         **TA*G**GGAA*TA*TT*CATT*GGC*ATATTAATTAAA
```

```
2070715
```

```
Base confidence: 0.0 (Prob. 0.500000) A= 0.0 C= 0.0 G= 0.0 T= 0.0 *= 0.0 / AA= 0.0 Position 2070798 (2070798 ref)
```

Acknowledgements:

- ***Jim Mullikin***
- ***Hannes Ponstingl***
- ***Adam Spargo***
- ***Tony Cox (Illumina)***
- ***Tony Cox (Sanger)***
- ***James Bonfield***
- ***Heng Li***

Solexa™

PACIFIC
BIOSCIENCES™