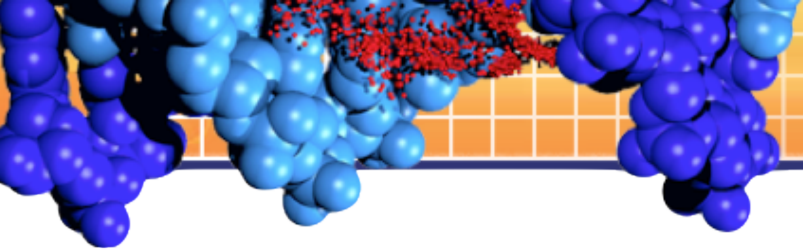


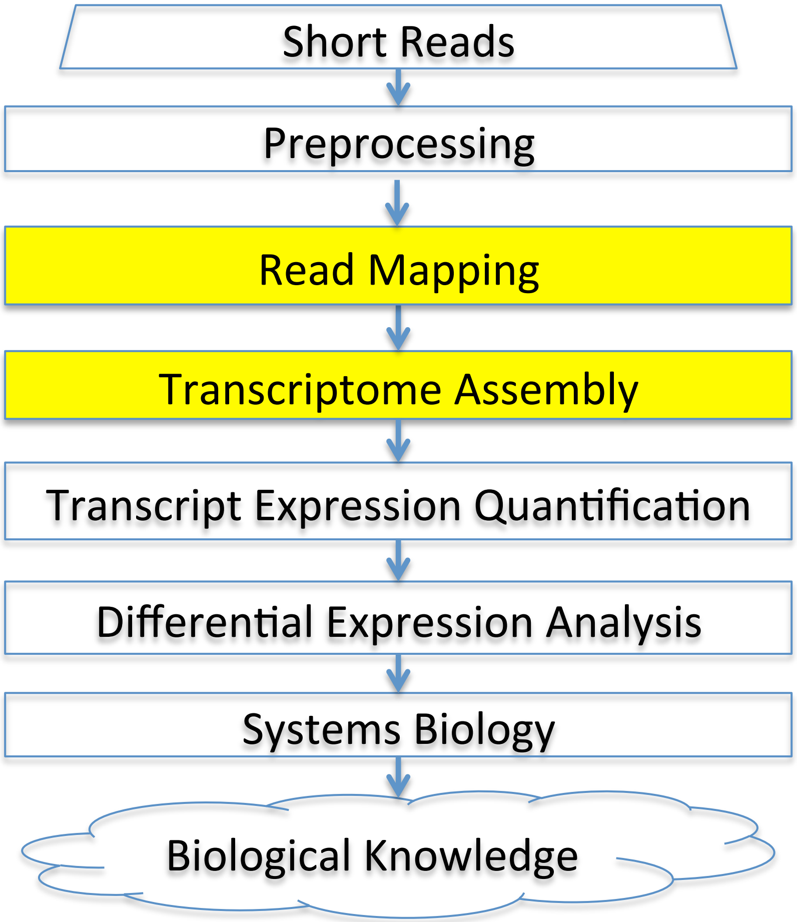
Bioinformatics Short Course: RNA-Seq Data Analysis

Part III: Transcriptome Assembly (Lecture)

Chuming Chen, Ph.D.
University of Delaware

May 22-23, 2012

RNA-Seq Data Analysis Workflow



Short Reads

Preprocessing

Read Mapping

Transcriptome Assembly

Transcript Expression Quantification

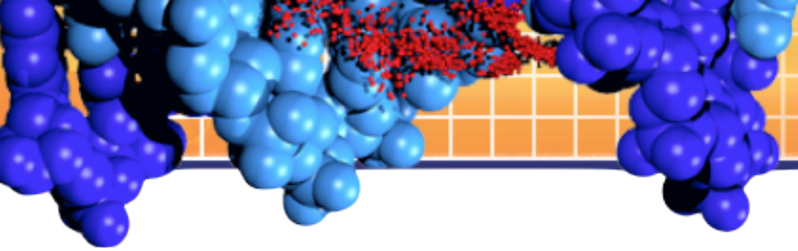
Differential Expression Analysis

Systems Biology

Biological Knowledge

Objective

- Learn the basics of transcriptome assembly using RNA-Seq data.
  - Transcriptome assembly strategies
  - Short read aligners
  - Alignment format and SAMtools
  - Alignment visualization

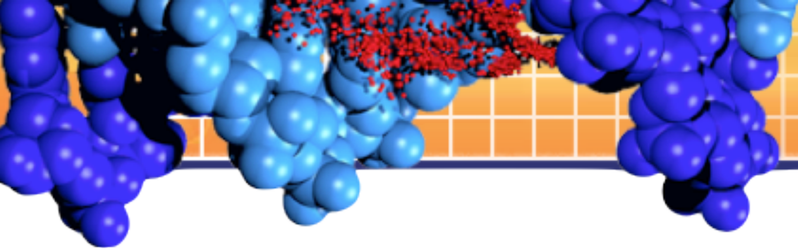


Transcriptome Sequencing (RNA-Seq)

- A very powerful technology for transcriptome studies.
- Uses high-throughput sequencing technologies to sequence the RNA molecules within a biological sample.
- Determines the primary sequence and relative abundance of each RNA molecule.
- Provides a comprehensive picture of the transcriptome including the complete quantification of all genes and their isoforms across samples.

EST Sequencing vs. RNA-Seq

- EST sequencing
  - Sanger sequencing technology.
  - Low-throughput.
  - Good at detecting more abundant transcripts.
- RNA-Seq
  - provide a near-complete snapshot of the expressed transcripts in a cell.



Microarrays vs. RNA-Seq

- Microarrays
  - High-throughput
  - Depends on the prior knowledge to design probes.
  - Cannot detect novel splicing variants, novel genes and transcripts.
- RNA-Seq
  - Can achieve base-pair level resolution.
  - Has higher dynamic range of expression levels.
  - Has low background noise and high sensitivity.
  - Uses less sample and becoming more cost-effective.

RNA-Seq Applications

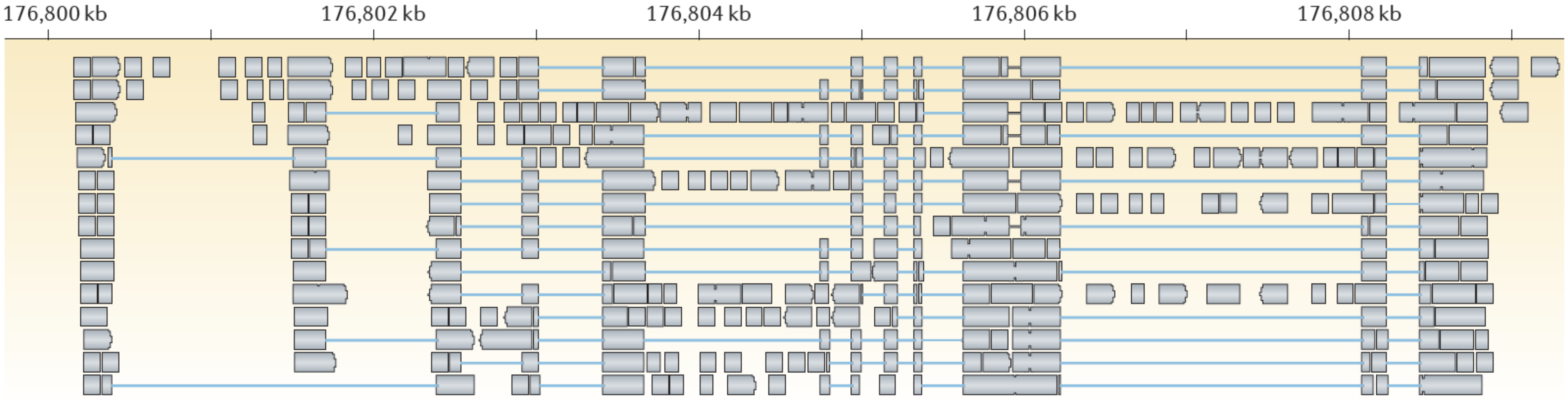
- Identify novel genes, transcripts, exons, alternative splicing events etc.
- Detect RNA editing and exonic SNPs/Indels.
- Transcriptome quantification and differential expression (gene and transcript levels).

Transcriptome Assembly Strategies

- Reference-based or ab initio assembly
  - Require a reference genome for the target transcriptome.
  - RNA-Seq reads are aligned to a reference genome using a splice-ware aligner.
  - Overlapping reads from each locus are clustered to build a graph for all possible isoforms.
  - Example tools: Cufflinks and Scripture etc.
- De Novo assembly
  - No reference genome required.
  - Leverages the redundancy of short-reads to find overlaps between them and assembles them into transcripts.
  - Example tools: Trans-Abyss, Trinity and Oases etc.
- Combined assembly
  - high sensitivity of reference-based assemblers.
  - the ability of De Novo assemblers to detect novel transcripts.

Reference-based Transcriptome Assembly (Step 1)

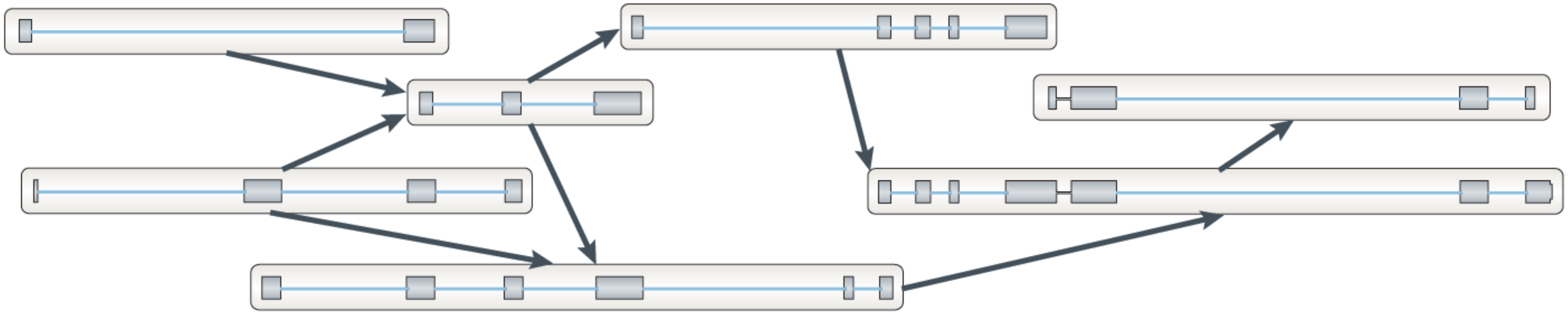
- Align the RNA-Seq reads to a reference genome using a splice-aware aligner such as Blat, TopHat, SpliceMap, MapSplice or GSNAP.



(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

Reference-based Transcriptome Assembly (Step 2)

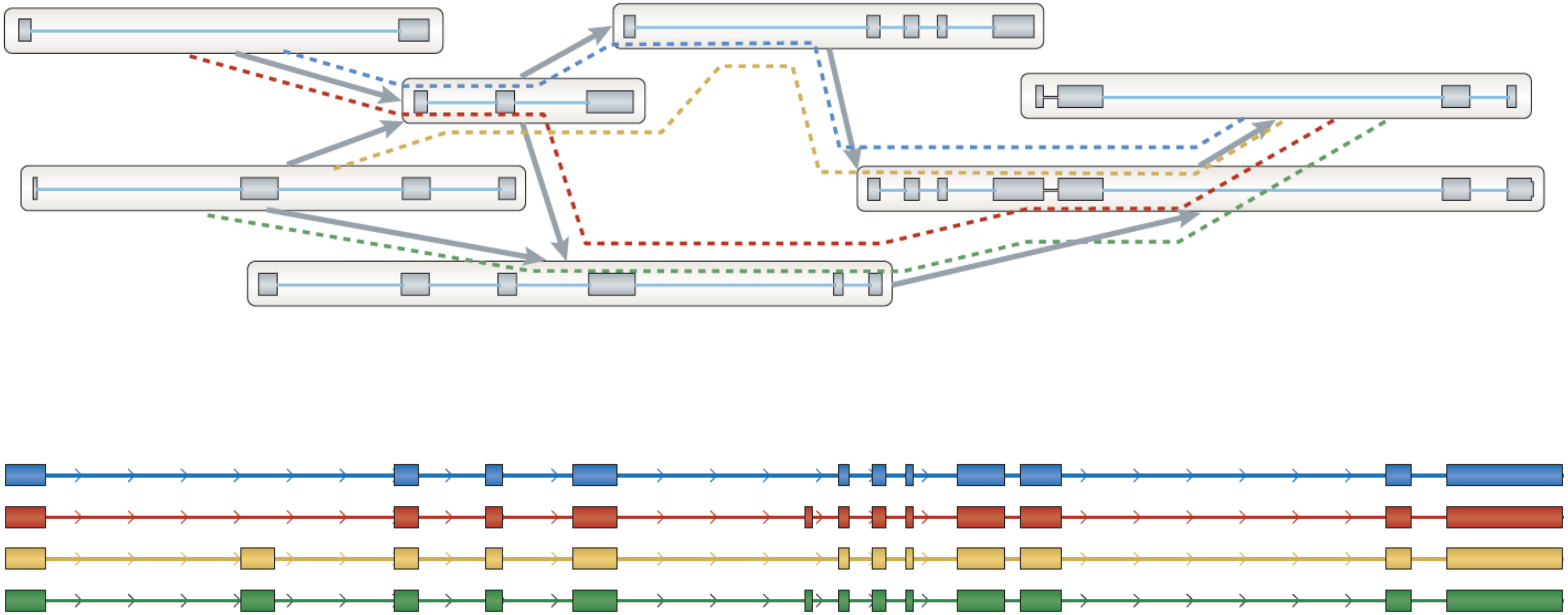
- Build a graph representing alternative splicing events by clustering the overlapping reads from each locus.



(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

Reference-based Transcriptome Assembly (Step 3)

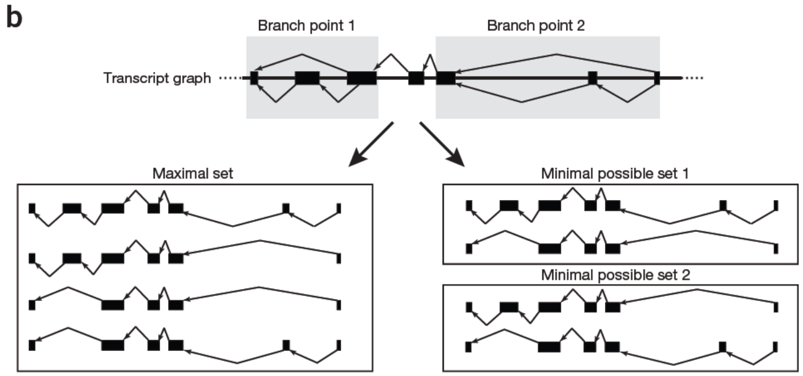
- Traverse the graph to join the compatible reads together into isoforms.



(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

Transcriptome Reconstruction

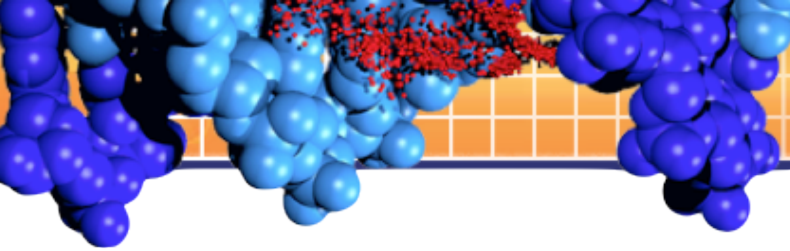
- Minimum path coverage of the graph (Maximum precision)
  - Traverse the graph to assemble isoforms by finding the minimum set of transcripts that “explain” the intron junctions with the reads.
  - Example: Cufflinks
- Maximum path coverage of the graph (Maximum sensitivity)
  - Find all paths through the graph that have a statistically significant read coverage.
  - Example: Scripture



(Garber et. al., **Nature methods**, VOL.8 NO.6, JUNE 2011)

Reference-based Assembly (Pros and Cons)

- Advantages:
  - Locus independent and can be assembled in parallel.
  - Contamination or sequencing artifacts are not expected to be aligned to the reference genome.
  - Small gaps within the transcripts can be filled by reference sequence.
  - Very sensitive and can assemble transcripts of low abundance.
  - Detect novel transcripts (lower expression levels) not present in current annotations.
- Disadvantages:
  - Depends on the quality of the reference genome.
  - Errors from short-read aligners are also carried over into the assembled transcripts.
  - Spliced reads spanning large introns may be missed due to aligners usually only search for introns of smaller lengths.
  - Non-specific reads (reads aligned to the reference in different locations) are hard to deal with by the aligners.
    - Ignore them may introduce gaps in the the assembled transcripts.
    - Random assignment may lead to a transcript from a region of genome that has no transcription.

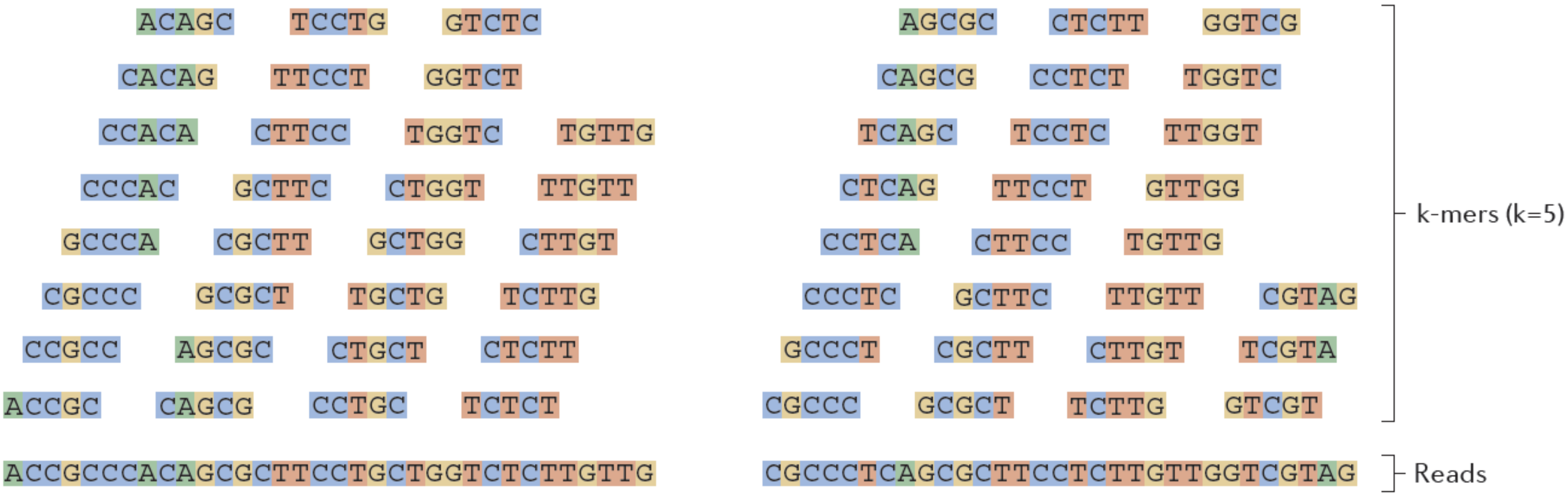


Reference-based Assembly (Usage)

- Simple transcriptomes of bacterial, archaeal and lower eukaryotic organisms (>10X)
- However, overlapping genes that are transcribed from the same strand and have comparable expression levels cannot be easily separated.
- Plant and mammalian transcriptomes are hard to assemble accurately.

De Novo Transcriptome Assembly (Step 1)

- All subsequences of length K (K-mers) are generated from each read.



ACAGC TCCTG GTCTC

CACAG TTCCT GGTCT

CCACA CTTCC TGGTC TGTTG

CCCAC GCTTC CTGGT TTGTT

GCCCA CGCTT GCTGG CTTGT

CGCCC GCGCT TGCTG TCTTG

CCGCC AGCGC CTGCT CTCTT

ACCGC CAGCG CCTGC TCTCT

ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTG

AGCGC CTCTT GGTCG

CAGCG CCTCT TGGTC

TCAGC TCCTC TTGGT

CTCAG TTCCT GTTGG

CCTCA CTTCC TGTTG

CCCTC GCTTC TTGTT CGTAG

GCCCT CGCTT CTTGT TCGTA

CGCCC GCGCT TCTTG GTCGT

CGCCCTCAGCGCTTCCTCTTGTTGGTCGTAG

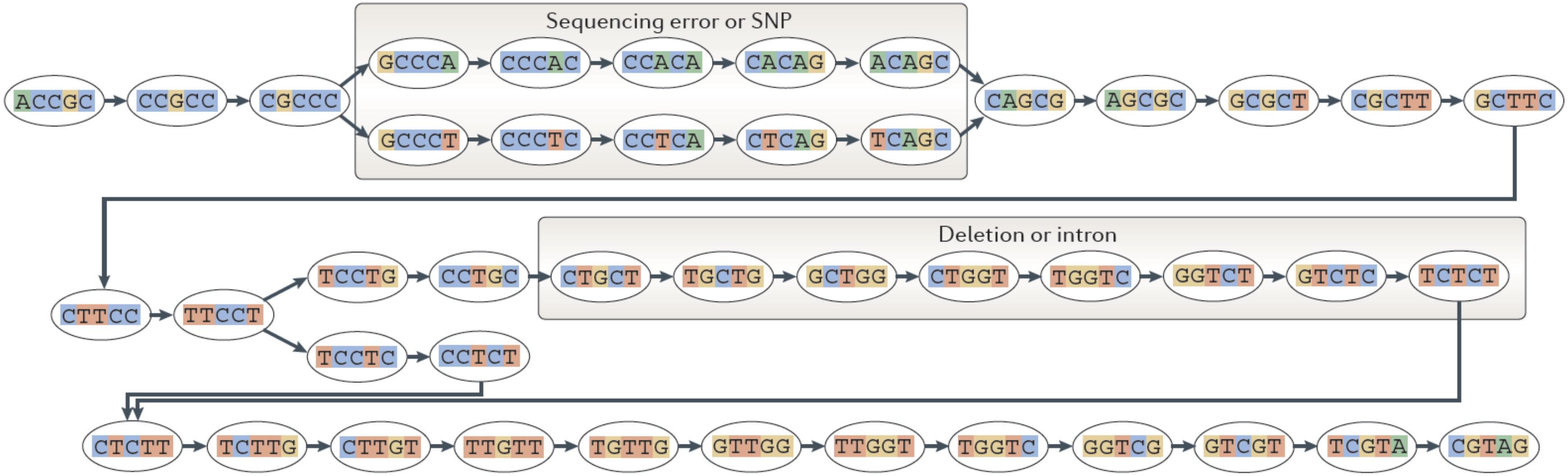
k-mers (k=5)

Reads

(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

De Novo Transcriptome Assembly (Step 2)

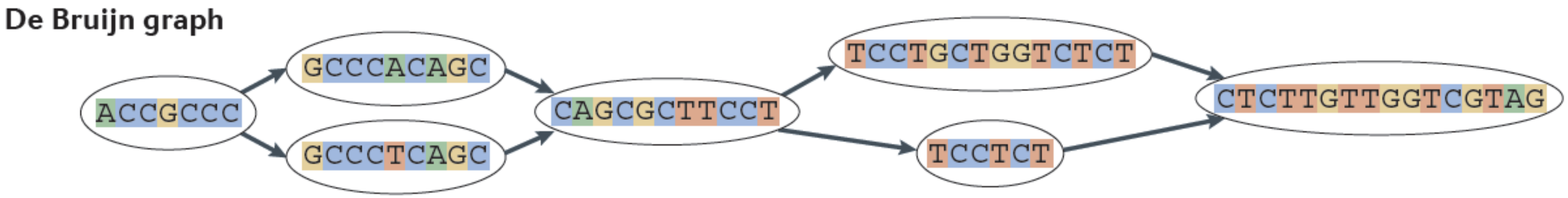
- Each node (or vertex) in the De Bruijn graph is represented by a unique K-mer.
- Pairs of nodes are connected if shifting a K-mer by one character creates an exact K-1 overlap between the two K-mers.
- SNPs cause 'bubbles' of length K in the De Bruijn graph.
- Introns or deletions introduce a shorter path in the graph.



(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

De Novo Transcriptome Assembly (Step 3)

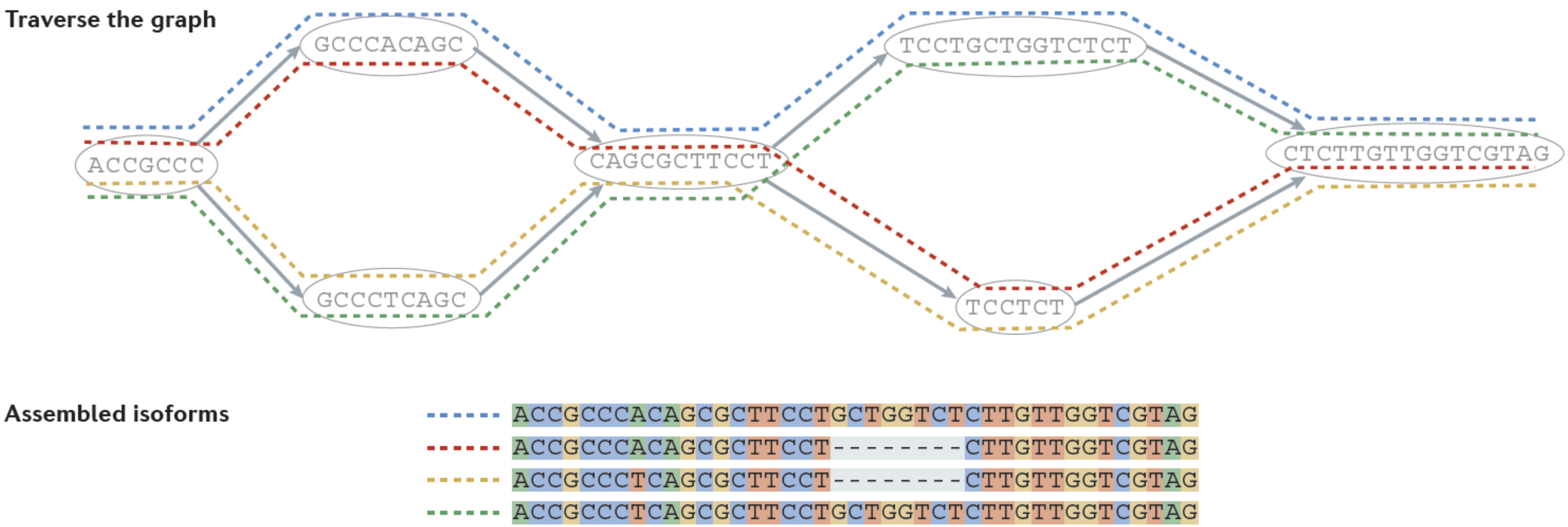
- Chains of adjacent nodes in the graph are collapsed into a single node.



(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

De Novo Transcriptome Assembly (Step 4)

- Traverse the graph to join the compatible reads together into isoforms.


Traverse the graph
ACCGCCC
GCCCACAGC
GCCCTCAGC
CAGCGCTTCCT
TCCTGCTGGTCTCT
TCCTCT
CTCTTGTTGGTCGTAG
Assembled isoforms
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG
ACCGCCCACAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG

(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

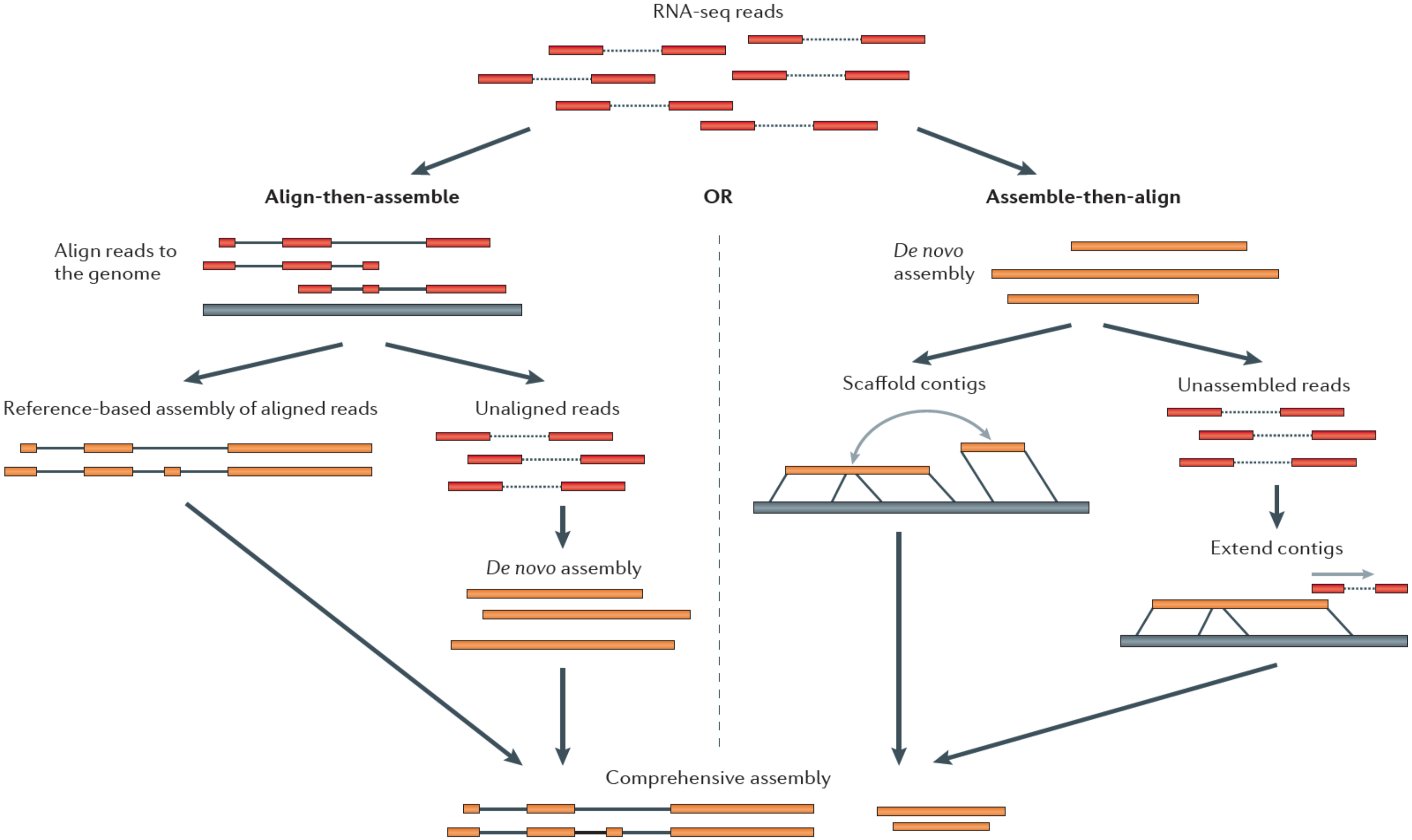
De Novo Assembly (Pros and Cons)

- Advantages:
  - Doesn't depend on the reference genome.
    - Provides an initial set of transcripts for expression analysis.
    - Detects transcripts that are transcribed from the segments of the genome missing in the genome assembly.
  - Doesn't depend on the correct alignment of reads to known splice sites or the prediction of novel splicing sites.
  - Long introns are no longer a concern.
  - Trans-spliced transcripts and transcripts originating from chromosomal rearrangements can also be assembled.
- Disadvantages:
  - Large computing resources are needed.
  - Higher sequencing depth for full length transcript assembly.
  - Sensitive to sequencing errors and chimeric molecules.
  - Highly similar transcripts (different alleles or paralogues) are likely to be assembled together.
  - Need annotation after assembly.

De Novo Assembly (Usage)

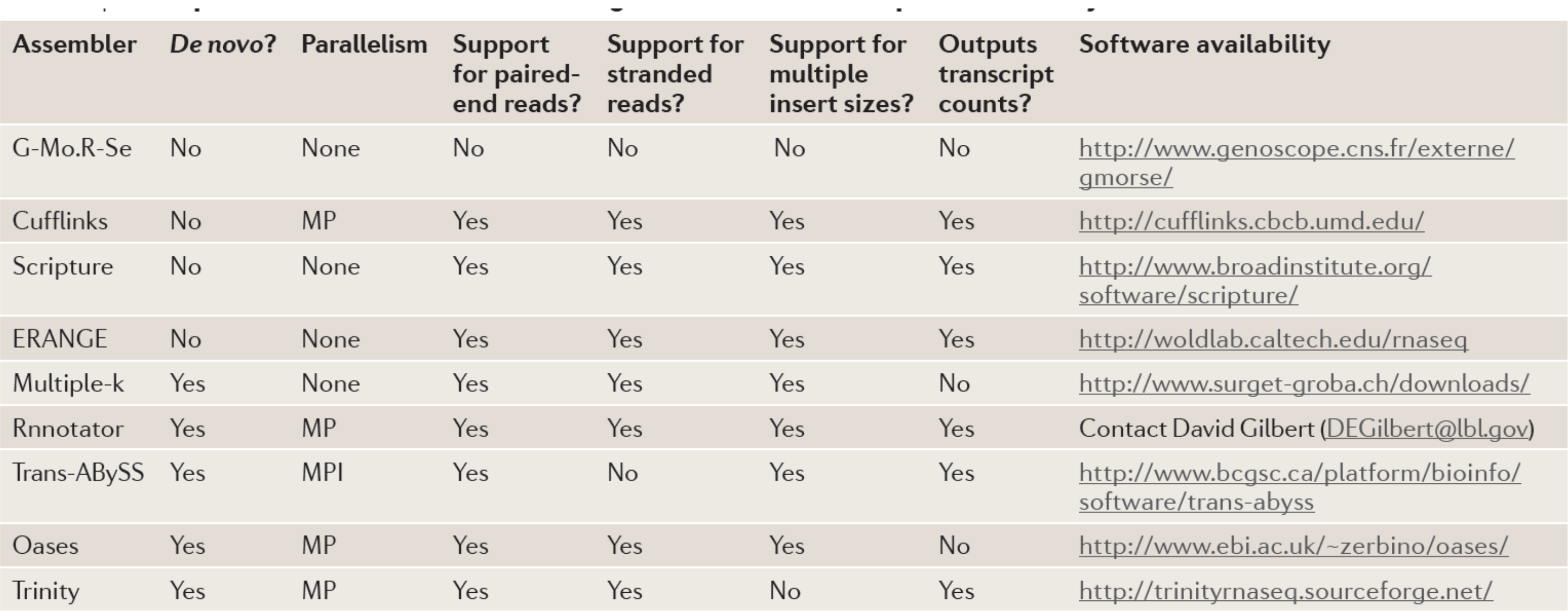
- Assembly of bacterial, archaeal and lower eukaryotic transcriptomes is straightforward (>30X).
- Overlapping genes transcribed from opposite strands can be resolved.
  - Building reverse compliment k-mers in the De Bruijin graph (not losing strand specific info)
  - Aligning the strand-specific reads to contigs after assembly.
- Assembly of higher eukaryotic transcriptomes is challenging.
  - Large genome with complicated alternatively spliced variants.
  - Large data set requires large computational resources.

Combined Assembly



(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

Tools for Transcriptome Assemby



Assembler	*De novo*?	Parallelism	Support for paired-end reads?	Support for stranded reads?	Support for multiple insert sizes?	Outputs transcript counts?	Software availability
G-Mo.R-Se	No	None	No	No	No	No	http://www.genoscope.cns.fr/externe/gmorse/
Cufflinks	No	MP	Yes	Yes	Yes	Yes	http://cufflinks.cbcb.umd.edu/
Scripture	No	None	Yes	Yes	Yes	Yes	http://www.broadinstitute.org/software/scripture/
ERANGE	No	None	Yes	Yes	Yes	Yes	http://woldlab.caltech.edu/rnaseq
Multiple-k	Yes	None	Yes	Yes	Yes	No	http://www.surget-groba.ch/downloads/
Rnnotator	Yes	MP	Yes	Yes	Yes	Yes	Contact David Gilbert (DEGilbert@lbl.gov)
Trans-ABySS	Yes	MPI	Yes	No	Yes	Yes	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
Oases	Yes	MP	Yes	Yes	Yes	No	http://www.ebi.ac.uk/~zerbino/oases/
Trinity	Yes	MP	Yes	Yes	No	Yes	http://trinityrnaseq.sourceforge.net/

(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

Which One to Choose?

- Assembly strategy
  - Existence or completeness of a reference genome.
  - Availability of sequencing and computing resources.
  - Most importantly, the goal of sequencing project.
    - Comprehensive annotation of the transcriptome with a reference genome
      - Multiple paired-end libraries.
      - Sequence the transcriptome at a great depth.
      - Use a combined strategy of reference-based and de novo assembly.
- Assembly program
  - Organism and sequencing platform specific.

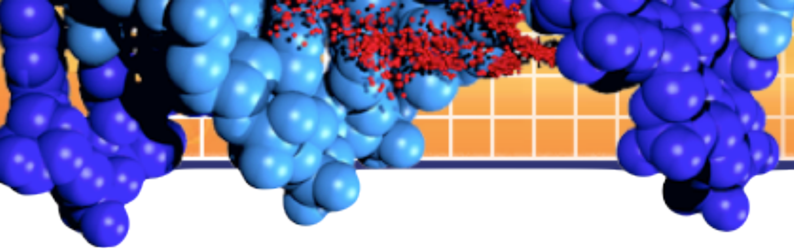
Assembly Quality Assessment

- Given a set of reference transcripts that are expressed in the sample and are derived from the same transcriptome, we can use the following metrics for evaluating the quality of an assembled transcriptome.
  - Accuracy
    - % of the correctly assembled bases estimated using the set of expressed reference transcripts.
  - Completeness
    - % of expressed reference transcripts covered by all the assembled transcripts.
  - Contiguity
    - % of expressed reference transcripts covered by a single, longest-assembled transcript.
  - Chimerism
    - % of chimaeras (contains non-repetitive parts from two or more different reference genes) that occur owing to mis-assemblies among all of the assembled transcripts.
  - Variant resolution
    - % transcript variants assembled and can be calculated as the average of the % of assembled variants with the reference set.

(Martin and Wang, **Nature reviews genetics**, Volume 12, October 2011)

Mapping Short RNA-Seq Reads

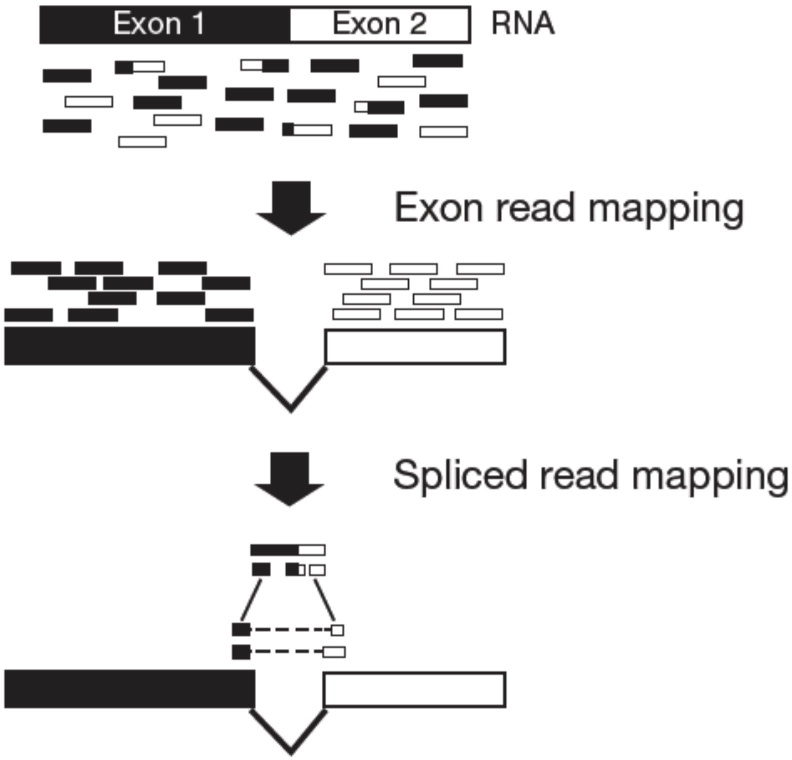
- Read alignment is a classic problem in bioinformatics.
  - Challenges
    - The number of reads per experiment is also increasing dramatically with new sequencing technology.
    - Short, high error rates and many reads span exon-exon junctions.
- Two major approaches:
  - Unspliced read aligners
  - Spliced read aligners



Unspliced Read Aligners

- Align reads to a reference without any large gaps.
- Seed methods (i. e. MAQ and Stampy)
  - Find matches for short subsequences ("seeds") in a read to the reference.
  - Narrow candidate regions using more sensitive methods (Smith-Waterman).
- Burrows-Wheeler transform methods (i. e. BWA, Bowtie, SOAP2)
  - Compact the genome to allow for very efficient search for perfect matches.
  - Performance decreases exponentially with the number of mismatches increase.
- Ideal for mapping reads against a reference cDNA databases for quantification purposes.
- Limited to identifying know exons and junctions.
- Do not allow for the identification of splicing events involving new exons.

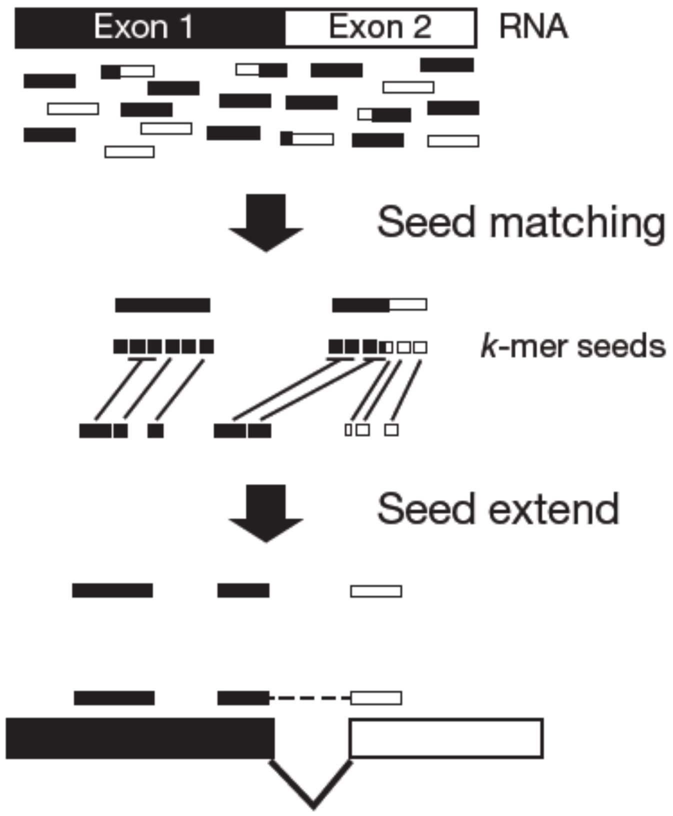
Spliced Read Aligners (Exon-first)



- First, map reads continuously to the genome using the un-spliced read aligners.
- Second, unmapped reads are split into shorter segments and aligned independently. Regions surrounding the mapped read segments are then searched for possible spliced connections.
- Fast and require fewer computational resources.
- Example tools: TopHat, MapSplice, SpliceMap,

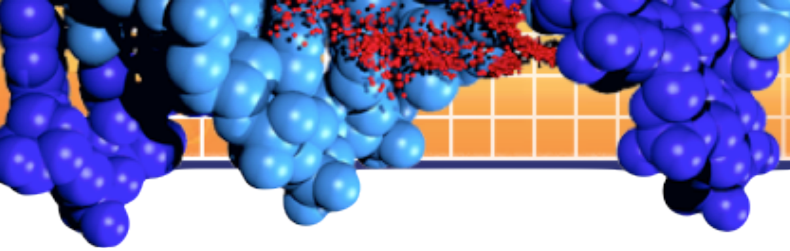
(Garber et. al., **Nature methods**, VOL.8 NO.6, JUNE 2011)

Spliced Read Aligners (Seed-extend)



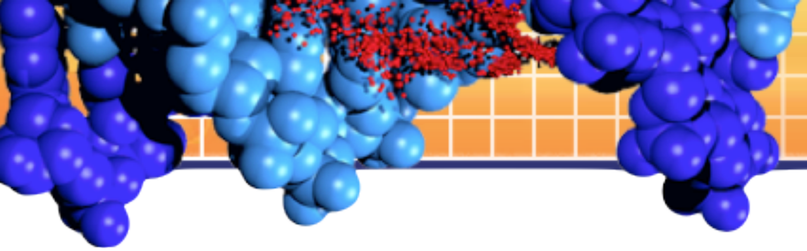
(Garber et. al., **Nature methods**, VOL.8 NO.6, JUNE 2011)

- Break reads into short seeds and placed onto the genome to localize the alignment.
- Candidate regions are then examined to determine the exact spliced alignment by using more sensitive methods or iterative extension and merging of initial seeds.
- Paired-end read mapping can be used to increase alignment specificity.
- Example tools: genomic short-read nucleotide program (GSNAP) and Optimal Spliced Alignments of Short Sequence Reads (QPALMA)



Types of Alignments

- Soft clipped alignment
- Hard clipped alignment
- Spliced alignment
- Padded alignment



Soft Clipped Alignment

- In Smith-Waterman alignment, a sequence may not be aligned from the beginning to the end. Subsequences at the ends may be clipped off.
- In the example alignment record below, on the read sequence, bases in uppercase are matches and bases in lowercase are clipped off.

```
 REF: AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGACTAGTCGATCGATGTG
READ:        gggGTGTAACC-GACTAGgggg
```

Hard Clipped Alignment

- Similar to soft clipped alignment. The only difference is that the hard clipped subsequence is not present in the alignment record.
- In the example alignment record belwo, the sequence stored is “GTGTAACC-GACTAG”, instead of “gggGTGTAACC-GACTAGgggg” as in soft clipped alignment.

```
 REF: AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGACTAGTCGATCGATGTG
READ:        ~~ggg~~GTGTAACC-GACTAG~~gggg~~
```

Spliced Alignment

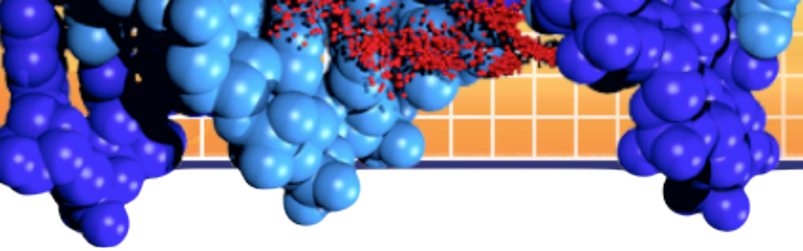
- In cDNA-to-genome alignment, we may need to distinguish introns from deletions in exons.
- In the example alignment record below, '…' on the read sequence indicates the intron.

```
 REF: AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGACTAGTCGATCGATGTG
READ:           GTGTAACCC................................TCAGAATA
```

Padded Alignment

- Most aligners only give the sequences inserted to the reference genome, but do not present how these inserted sequences are aligned against each other.
- Alignment with inserted sequences fully aligned is called padded alignment.
- In the example alignment records below, GA on READ1 and A on READ2 are inserted to the reference.

```
  REF: CACGATCA**GACCGATACGTCCGA           REF: CACGATCA**GACCGATACGTCCGA
READ1:   CGATCAGAGACCGATA                READ1:   CGATCAGAGACCGATA
READ2:     ATCA*AGACCGATAC               READ2:     ATCAA*GACCGATAC
```



Alignment Format (SAM/BAM)

- To store the read alignments against reference sequences.
- SAM stands for **S**equence **A**lignment/**M**ap format
  - **NOT** Significance Analysis of Microarrays.
- It is a Tab-delimited text format
  - Head section (optional, but recommended).
  - Alignment section.
- BAM is the binary version of SAM file
  - Indexed.
  - Compressed by the BGZF library.

The Alignment Record and SAM format (Example)

Alignment record:

```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                  ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                       CAGCGCCAT
```

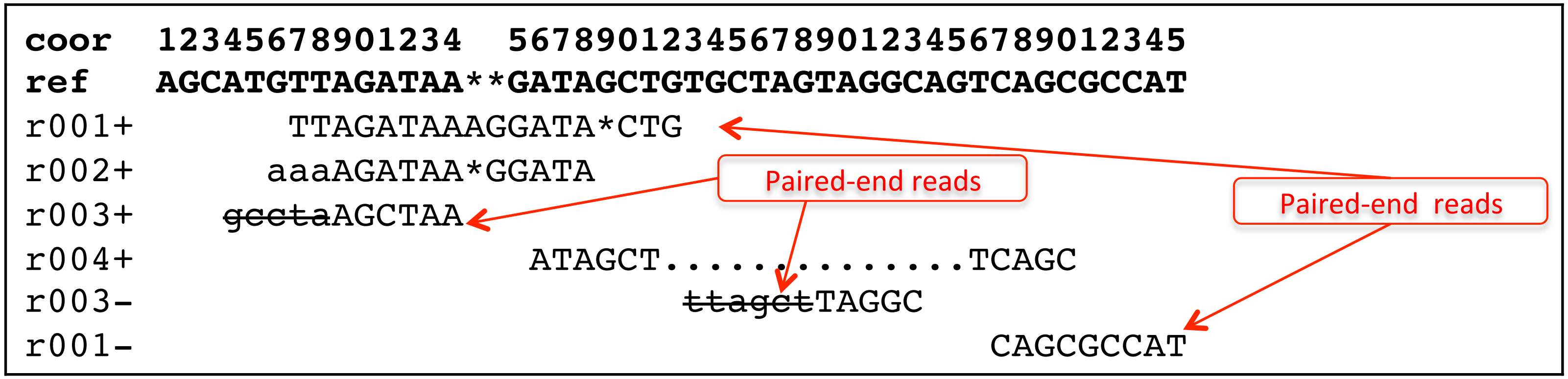
Corresponding SAM format:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

The Alignment Record (Paired-end reads)

Alignment record:



```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                187ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                       CAGCGCCAT
```

Paired-end reads

Paired-end reads

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

The Alignment Record (Soft clipped alignment)

Alignment record:

```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                  ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                       CAGCGCCAT
```

Soft clipped alignment

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

The Alignment Record (Padded alignment)

Alignment record:

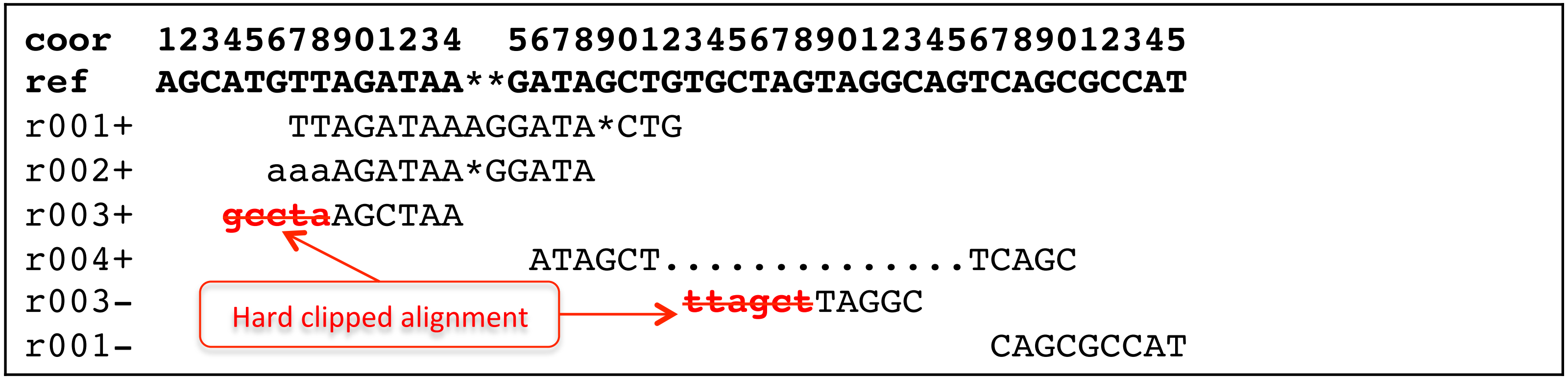
```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                  ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                     CAGCGCCAT
```

Padded alignment

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

The Alignment Record (Hard clipped alignment)

Alignment record:



```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                  ATAGCT..............TCAGC
r003-                   ~~ttagct~~TAGGC
r001-                                 CAGCGCCAT
```

Hard clipped alignment

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

The Alignment Record (Spliced alignment)

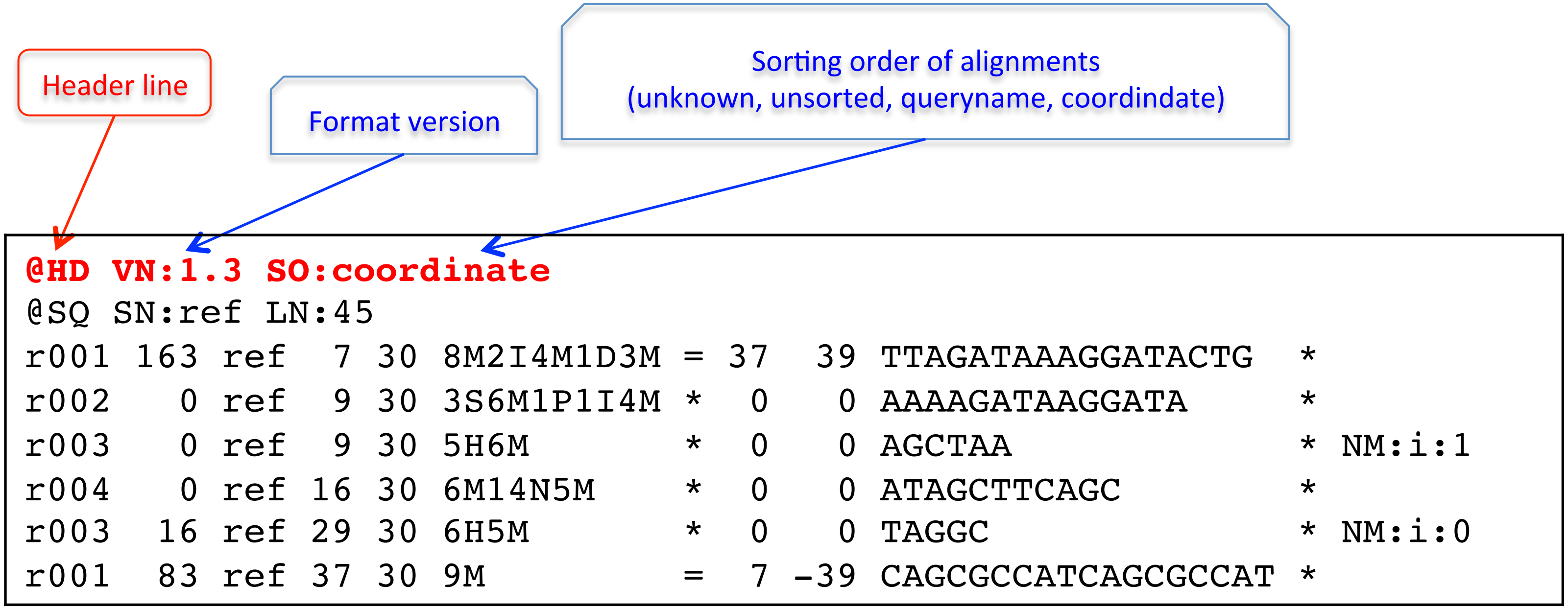
Alignment record:

```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                  ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                       CAGCGCCAT
```

Spliced alignment

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

Head Section (Header line)



Header line

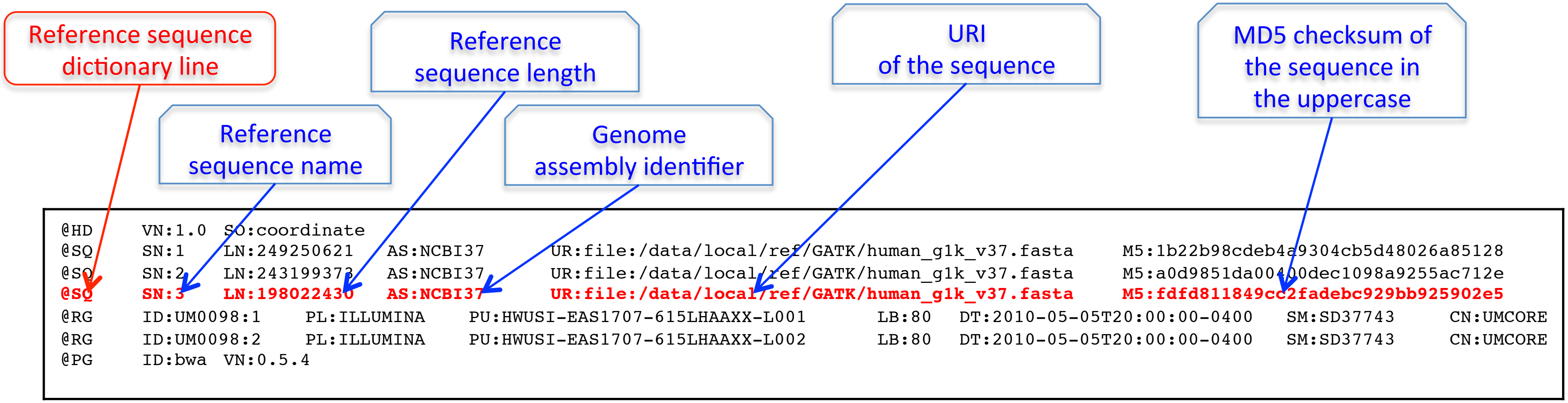
Format version

Sorting order of alignments
(unknown, unsorted, queryname, coordindate)

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

Head Section (Reference sequence dictionary line)



Reference sequence dictionary line

Reference sequence name

Reference sequence length

Genome assembly identifier

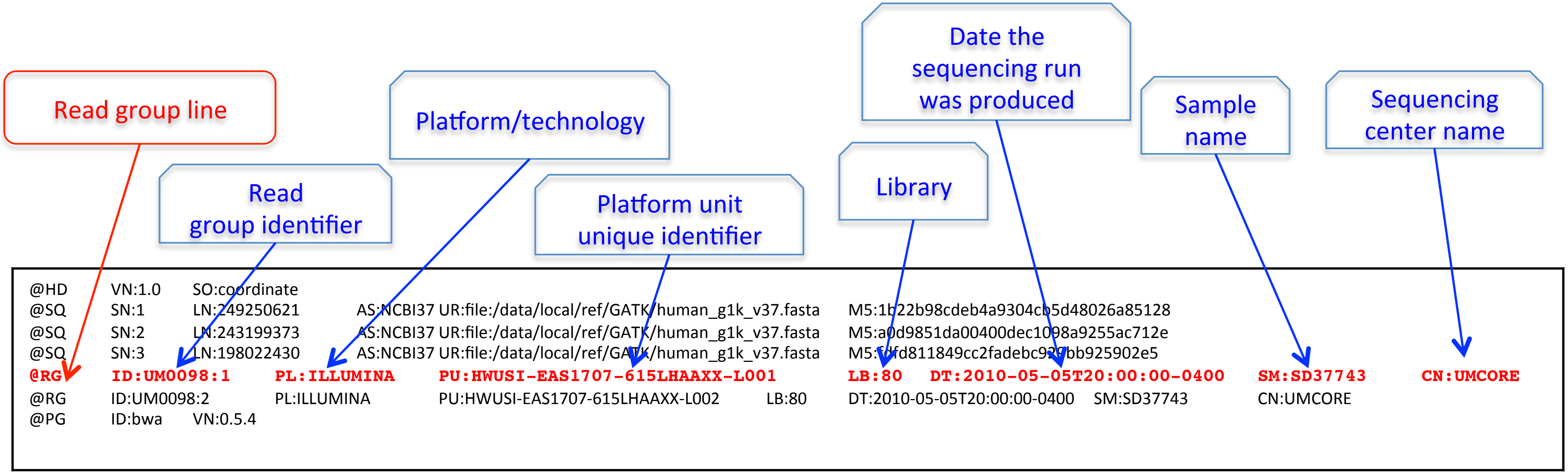
URI of the sequence

MD5 checksum of the sequence in the uppercase

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:fdfd811849cc2fadebc929bb925902e5
@RG     ID:UM0098:1     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001      LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@RG     ID:UM0098:2     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002      LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@PG     ID:bwa  VN:0.5.4
```

(http://genome.sph.umich.edu/wiki/SAM)

Head Section (Read group line)



Read group line

Read group identifier

Platform/technology

Platform unit unique identifier

Library

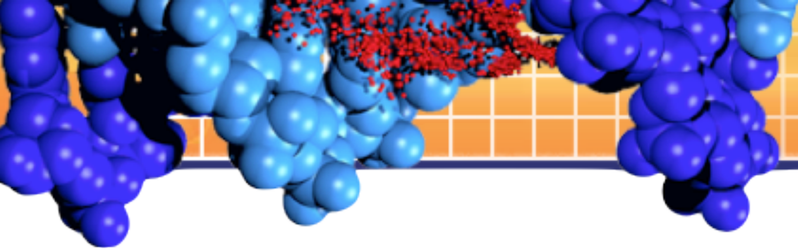
Date the sequencing run was produced

Sample name

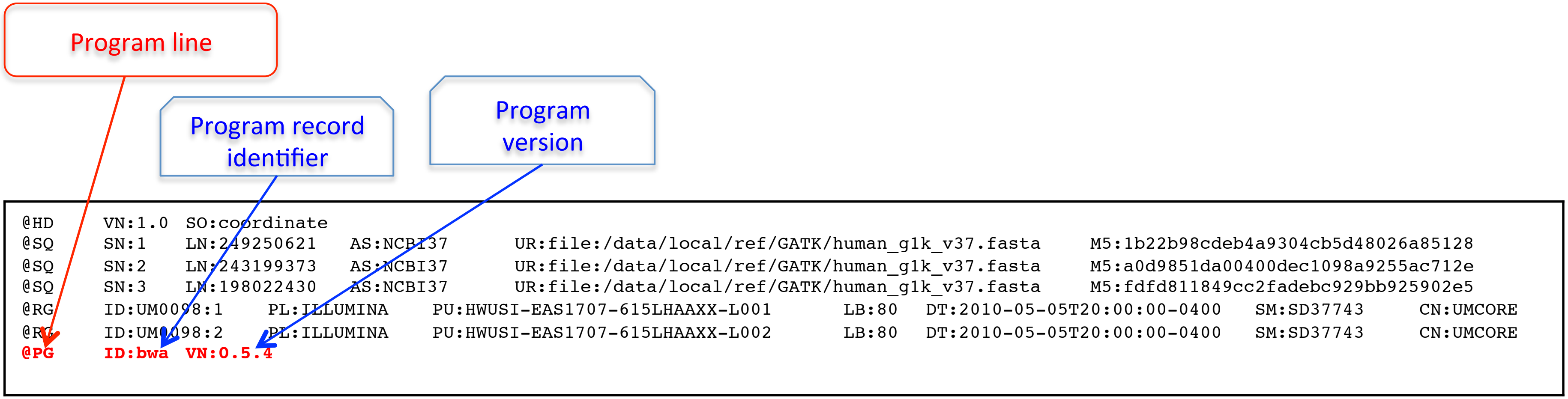
Sequencing center name

```
@HD   VN:1.0       SO:coordinate
@SQ   SN:1         LN:249250621    AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta    M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ   SN:2         LN:243199373    AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta    M5:a0d9851da00400dec1098a9255ac712e
@SQ   SN:3         LN:198022430    AS:NCBI37 UR:file:/data/local/ref/GATK/human_g1k_v37.fasta    M5:fdfd811849cc2fadebc929bb925902e5
@RG   ID:UM0098:1  PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001                                LB:80   DT:2010-05-05T20:00:00-0400    SM:SD37743    CN:UMCORE
@RG   ID:UM0098:2  PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002        LB:80                  DT:2010-05-05T20:00:00-0400   SM:SD37743     CN:UMCORE
@PG   ID:bwa       VN:0.5.4
```

(http://genome.sph.umich.edu/wiki/SAM)



Head Section (Program line)

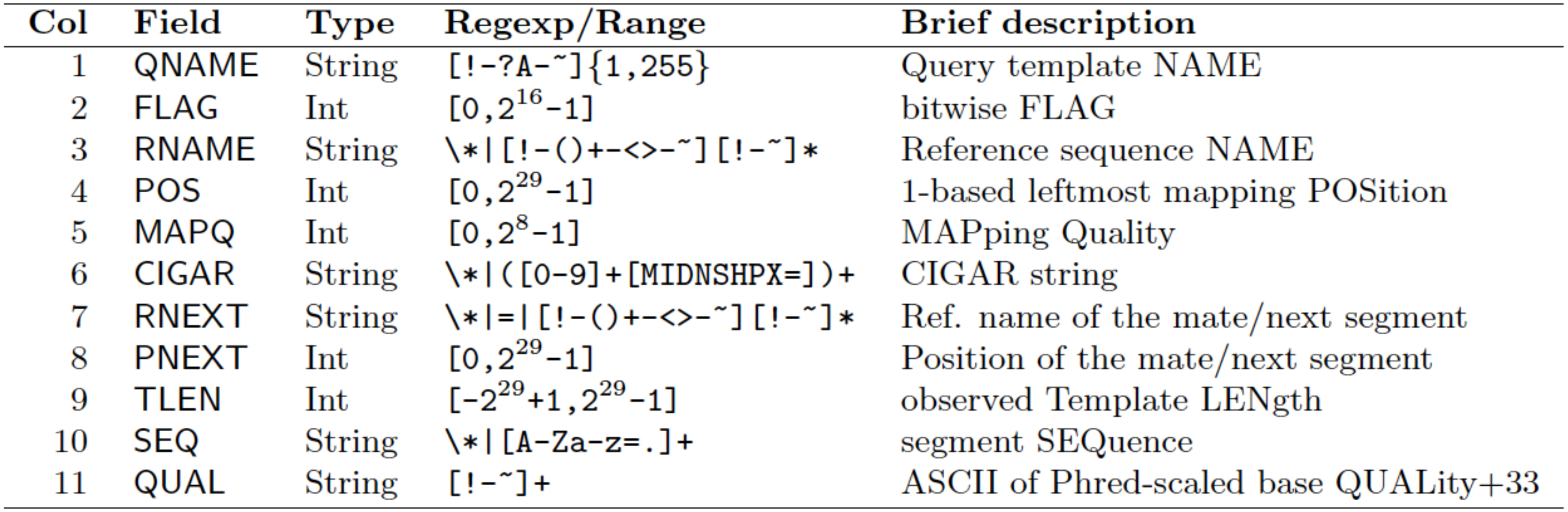


```
@HD     VN:1.0  SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta     M5:fdfd811849cc2fadebc929bb925902e5
@RG     ID:UM0098:1     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001       LB:80   DT:2010-05-05T20:00:00-0400    SM:SD37743      CN:UMCORE
@RG     ID:UM0098:2     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002       LB:80   DT:2010-05-05T20:00:00-0400    SM:SD37743      CN:UMCORE
@PG     ID:bwa  VN:0.5.4
```

(http://genome.sph.umich.edu/wiki/SAM)

Alignment Section (Mandatory Fields)

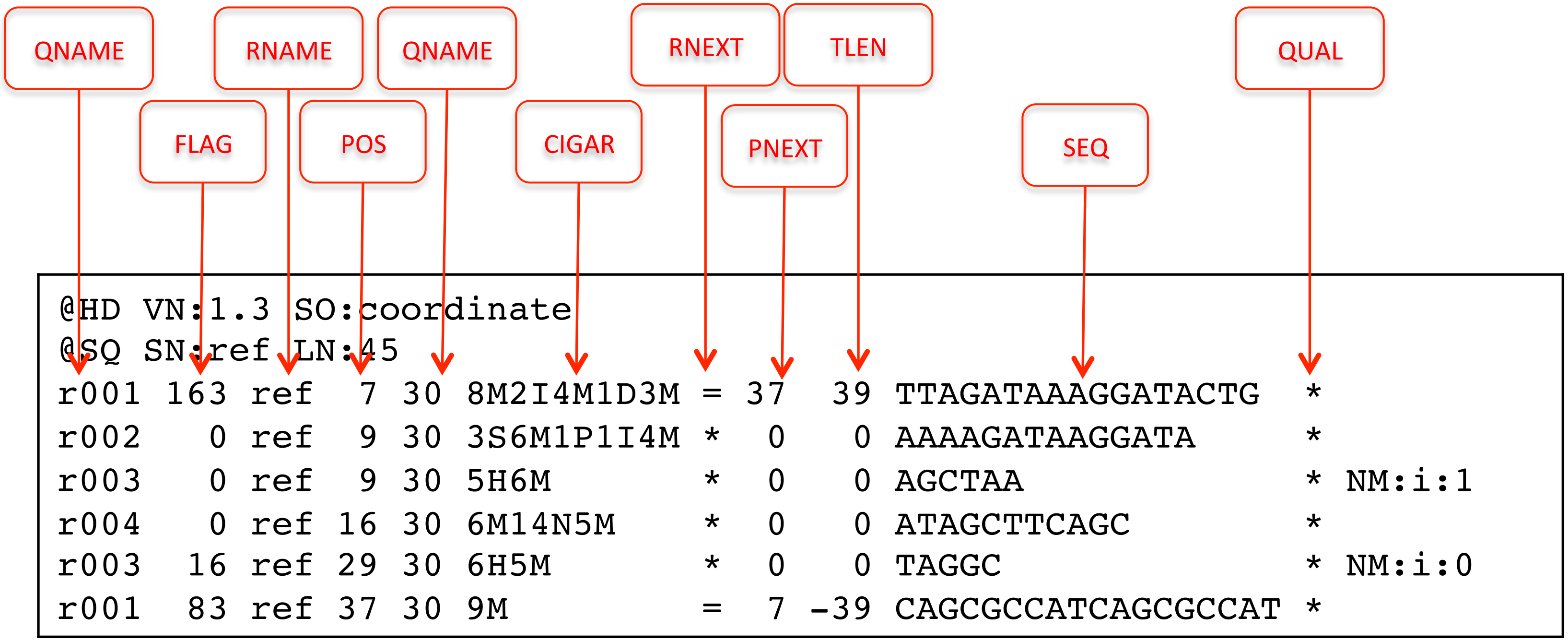
- Each alignment line has 11 mandatory fields for essential alignment information
- Column 12 and anything follows it is optional



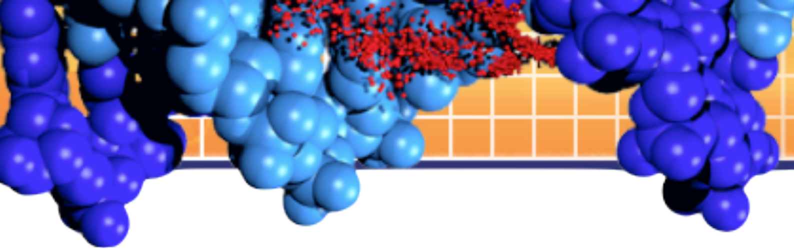
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{29}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	[0,2^{29}-1]	Position of the mate/next segment
9	TLEN	Int	[-2^{29}+1,2^{29}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

(http://samtools.sourceforge.net/SAM1.pdf)

Alignment Section (Example)



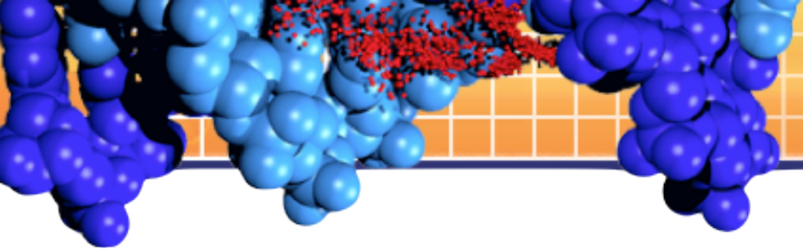
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```



Bitwise Flag – information describing the alignment

Base 10	Base 16	Description	Meaning
1	0x1	Template having multiple segments in sequencing	The read originated from a paired sequencing molecule
2	0x2	Each segment properly aligned according to the aligner	The read is mapped in a proper pair
4	0x4	Segment unmapped	The query sequence itself is unmapped
8	0x8	Next segment in the template unmapped	The query's mate is unmapped
16	0x10	SEQ being reverse complemented	The query is in the reverse strand
32	0X20	SEQ of the next segment in the template being reversed	The query's mate is in the reverse strand
64	0x40	The first segment in the template	The query is the first read in the pair
128	0x80	The last segment in the template	The query is the second read in the pair
256	0x100	Secondary alignment	The alignment is not primary
512	0x200	Not passing quality controls	The read fails paltform/vendor quality checks
1024	0x400	PCR or optical duplicate	The read is either a PCR duplicate or an optical duplicate

(http://samtools.sourceforge.net/SAM1.pdf)



Quiz (Bitwise Flag)

Base 10	Base 16	Description	Meaning
1	0x1	Template having multiple segments in sequencing	The read originated from a paired sequencing molecule
2	0x2	Each segment properly aligned according to the aligner	The read is mapped in a proper pair
4	0x4	Segment unmapped	The query sequence itself is unmapped
8	0x8	Next segment in the template unmapped	The query's mate is unmapped
16	0x10	SEQ being reverse complemented	The query is in the reverse strand
32	0X20	SEQ of the next segment in the template being reversed	The query's mate is in the reverse strand
64	0x40	The first segment in the template	The query is the first read in the pair
128	0x80	The last segment in the template	The query is the second read in the pair
256	0x100	Secondary alignment	The alignment is not primary
512	0x200	Not passing quality controls	The read fails paltform/vendor quality checks
1024	0x400	PCR or optical duplicate	The read is either a PCR duplicate or an optical duplicate

(http://samtools.sourceforge.net/SAM1.pdf)

What does 163 stand for?

Quiz (Bitwise Flag)

Base 10	Base 16	Description	Meaning
1	0x1	Template having multiple segments in sequencing	The read originated from a paired sequencing molecule
2	0x2	Each segment properly aligned according to the aligner	The read is mapped in a proper pair
4	0x4	Segment unmapped	The query sequence itself is unmapped
8	0x8	Next segment in the template unmapped	The query's mate is unmapped
16	0x10	SEQ being reverse complemented	The query is in the reverse strand
32	0X20	SEQ of the next segment in the template being reversed	The query's mate is in the reverse strand
64	0x40	The first segment in the template	The query is the first read in the pair
128	**0x80**	**The last segment in the template**	**The query is the second read in the pair**
256	0x100	Secondary alignment	The alignment is not primary
512	0x200	Not passing quality controls	The read fails paltform/vendor quality checks
1024	0x400	PCR or optical duplicate	The read is either a PCR duplicate or an optical duplicate

(http://samtools.sourceforge.net/SAM1.pdf)

What does 163 stand for?

163 = 128 + …

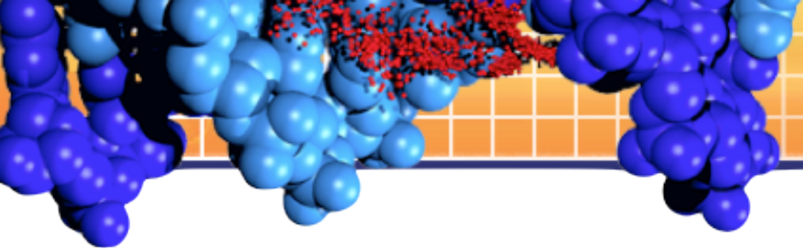
Quiz (Bitwise Flag)

Base 10	Base 16	Description	Meaning
1	0x1	Template having multiple segments in sequencing	The read originated from a paired sequencing molecule
2	0x2	Each segment properly aligned according to the aligner	The read is mapped in a proper pair
4	0x4	Segment unmapped	The query sequence itself is unmapped
8	0x8	Next segment in the template unmapped	The query's mate is unmapped
16	0x10	SEQ being reverse complemented	The query is in the reverse strand
32	**0X20**	**SEQ of the next segment in the template being reversed**	**The query's mate is in the reverse strand**
64	0x40	The first segment in the template	The query is the first read in the pair
128	0x80	The last segment in the template	The query is the second read in the pair
256	0x100	Secondary alignment	The alignment is not primary
512	0x200	Not passing quality controls	The read fails paltform/vendor quality checks
1024	0x400	PCR or optical duplicate	The read is either a PCR duplicate or an optical duplicate

(http://samtools.sourceforge.net/SAM1.pdf)

What does 163 stand for?

163 = 128 + 32 + …



Quiz (Bitwise Flag)

Base 10	Base 16	Description	Meaning
1	0x1	Template having multiple segments in sequencing	The read originated from a paired sequencing molecule
2	**0x2**	**Each segment properly aligned according to the aligner**	**The read is mapped in a proper pair**
4	0x4	Segment unmapped	The query sequence itself is unmapped
8	0x8	Next segment in the template unmapped	The query's mate is unmapped
16	0x10	SEQ being reverse complemented	The query is in the reverse strand
32	0X20	SEQ of the next segment in the template being reversed	The query's mate is in the reverse strand
64	0x40	The first segment in the template	The query is the first read in the pair
128	0x80	The last segment in the template	The query is the second read in the pair
256	0x100	Secondary alignment	The alignment is not primary
512	0x200	Not passing quality controls	The read fails paltform/vendor quality checks
1024	0x400	PCR or optical duplicate	The read is either a PCR duplicate or an optical duplicate

(http://samtools.sourceforge.net/SAM1.pdf)

What does 163 stand for?

163 = 128 + 32 + 2 + …

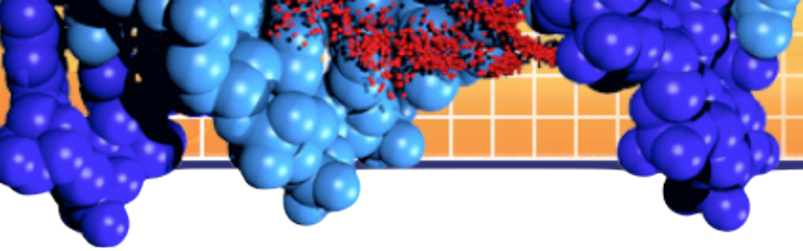
Quiz (Bitwise Flag)

Base 10	Base 16	Description	Meaning
1	0x1	Template having multiple segments in sequencing	The read originated from a paired sequencing molecule
2	0x2	Each segment properly aligned according to the aligner	The read is mapped in a proper pair
4	0x4	Segment unmapped	The query sequence itself is unmapped
8	0x8	Next segment in the template unmapped	The query's mate is unmapped
16	0x10	SEQ being reverse complemented	The query is in the reverse strand
32	0X20	SEQ of the next segment in the template being reversed	The query's mate is in the reverse strand
64	0x40	The first segment in the template	The query is the first read in the pair
128	0x80	The last segment in the template	The query is the second read in the pair
256	0x100	Secondary alignment	The alignment is not primary
512	0x200	Not passing quality controls	The read fails paltform/vendor quality checks
1024	0x400	PCR or optical duplicate	The read is either a PCR duplicate or an optical duplicate

(http://samtools.sourceforge.net/SAM1.pdf)

What does 163 stand for?

163 = 128 + 32 + 2 + 1



Quiz (Bitwise Flag)

Base 10	Base 16	Description	Meaning
1	0x1	Template having multiple segments in sequencing	The read originated from a paired sequencing molecule
2	0x2	Each segment properly aligned according to the aligner	The read is mapped in a proper pair
4	0x4	Segment unmapped	The query sequence itself is unmapped
8	0x8	Next segment in the template unmapped	The query's mate is unmapped
16	0x10	SEQ being reverse complemented	The query is in the reverse strand
32	0X20	SEQ of the next segment in the template being reversed	The query's mate is in the reverse strand
64	0x40	The first segment in the template	The query is the first read in the pair
128	0x80	The last segment in the template	The query is the second read in the pair
256	0x100	Secondary alignment	The alignment is not primary
512	0x200	Not passing quality controls	The read fails paltform/vendor quality checks
1024	0x400	PCR or optical duplicate	The read is either a PCR duplicate or an optical duplicate

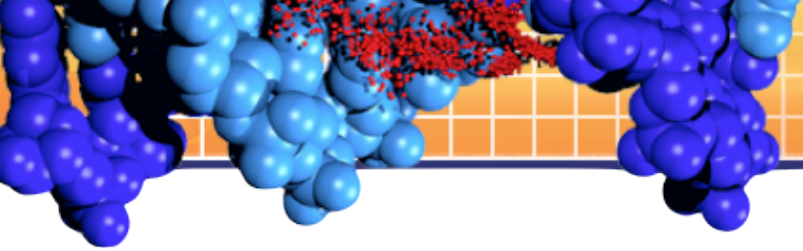
(http://samtools.sourceforge.net/SAM1.pdf)

What does 163 stand for?

163 = 128 + 32 + 2 + 1

Answer:

- It is properly paired (1+2)
- Its mate is mapped on the reverse strand (32)
- It is the second read in the pair (128)



Quiz (Bitwise Flag)

Base 10	Base 16	Description	Meaning
1	0x1	Template having multiple segments in sequencing	The read originated from a paired sequencing molecule
2	0x2	Each segment properly aligned according to the aligner	The read is mapped in a proper pair
4	0x4	Segment unmapped	The query sequence itself is unmapped
8	0x8	Next segment in the template unmapped	The query's mate is unmapped
16	0x10	SEQ being reverse complemented	The query is in the reverse strand
32	0X20	SEQ of the next segment in the template being reversed	The query's mate is in the reverse strand
64	0x40	The first segment in the template	The query is the first read in the pair
128	0x80	The last segment in the template	The query is the second read in the pair
256	0x100	Secondary alignment	The alignment is not primary
512	0x200	Not passing quality controls	The read fails paltform/vendor quality checks
1024	0x400	PCR or optical duplicate	The read is either a PCR duplicate or an optical duplicate

(http://samtools.sourceforge.net/SAM1.pdf)

What does 163 stand for?

163 = 128 + 32 + 2 + 1

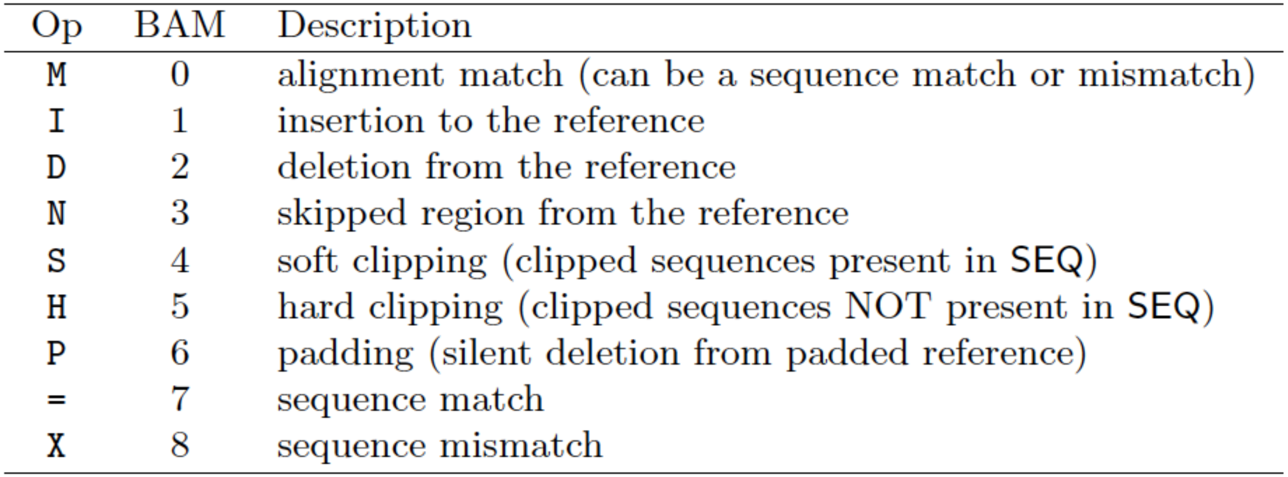
http://picard.sourceforge.net/explain-flags.html

Answer:

- It is properly paired (1+2)
- Its mate is mapped on the reverse strand (32)
- It is the second read in the pair (128)

Extended CIGAR Strings

- A sequence of base lengths and associated operations describing pairwise alignment.
- They are used to indicated things like:
  - Which bases align (either match or mismatch) with the reference?
  - Which bases are deleted from the reference?
  - Which bases are insertions that are not in the reference?
  - Which bases are soft/hard clipped?
  - Which bases are padded?
  - Which bases are spliced alignment?



Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

(http://samtools.sourceforge.net/SAM1.pdf, http://genome.sph.umich.edu/wiki/SAM)

Quiz (Extended CIGAR Strings)

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

```
   REF: CACGATCA**GACCGATACGTCCGA
READ-A:     ATCA*AGACCGATAC
```

What is the CIGAR string for READ-A?

```
   REF: CACGATCA**GACCGATACGTCCGA
READ-B:     ATCAA*GACCGATAC
```

What is the CIGAR string for READ-B?

Quiz (Extended CIGAR Strings)

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

```
   REF: CACGATCA**GACCGATACGTCCGA
READ-A:     ATCA*AGACCGATAC
```

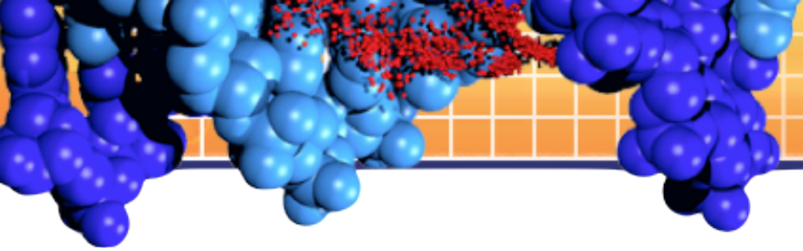
What is the CIGAR for READ-A?

Answer:

4M1P1I9M

```
   REF: CACGATCA**GACCGATACGTCCGA
READ-B:     ATCAA*GACCGATAC
```

What is the CIGAR for READ-B?



Quiz (Extended CIGAR Strings)

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

```
   REF: CACGATCA**GACCGATACGTCCGA
READ-A:     ATCA*AGACCGATAC
```

What is the CIGAR for READ-A?

Answer:

4M1P1I9M

```
   REF: CACGATCA**GACCGATACGTCCGA
READ-B:     ATCAA*GACCGATAC
```

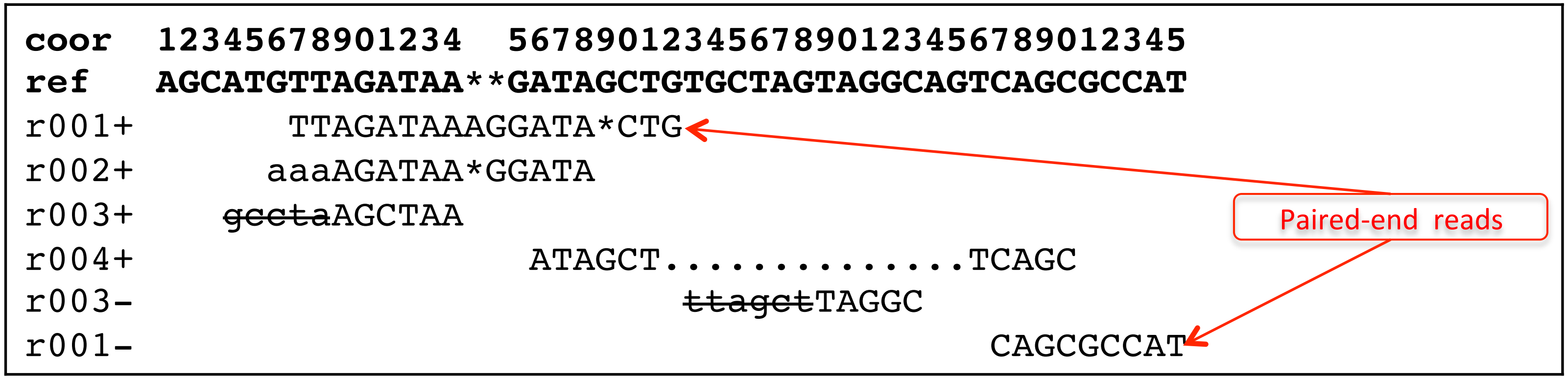
What is the CIGAR for READ-B?

Answer:

4M1I1P9M

Alignment Section (Example - r001)

Alignment record:



```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     ~~gccta~~AGCTAA
r004+                   ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                       CAGCGCCAT
```

Corresponding SAM format:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003    0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003   16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001   83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

Alignment Section (Example - r001)

Alignment record:

```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+   ~~gccta~~AGCTAA
r004+                   ATAGCT..............TCAGC
r003-                        ~~ttagct~~TAGGC
r001-                                        CAGCGCCAT
```

Paired-end reads

Corresponding SAM format:

163=128+32+2+1

- It is properly paired (1+2)
- Its mate is mapped on the reverse strand (32)
- It is the second read in the pair (128)

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003    0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003   16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001   83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

Alignment Section (Example - r001)

Alignment record:

```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     ~~gccta~~AGCTAA
r004+                  ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                        CAGCGCCAT
```

- 8 matches (TTAGATTA -- TTAGATAA)

Corresponding SAM format:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

Alignment Section (Example - r001)

Alignment record:

```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                  ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                     CAGCGCCAT
```

- 8 matches (TTAGATTA -- TTAGATAA)
- **2 insertions (**->AG)**

Corresponding SAM format:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

Alignment Section (Example - r001)

Alignment record:

```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                 ATAGCT..............TCAGC
r003-                        ~~ttagct~~TAGGC
r001-                                      CAGCGCCAT
```

Corresponding SAM format:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

- 8 matches (TTAGATTA -- TTAGATAA)
- 2 insertions (**->AG)
- **4 matches (GATA -- GATA)**

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

Alignment Section (Example - r001)

Alignment record:

```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     ~~gccta~~AGCTAA
r004+                   ATAGCT..............TCAGC
r003-                          ~~ttagct~~TAGGC
r001-                                        CAGCGCCAT
```

- 8 matches (TTAGATTA -- TTAGATAA)
- 2 insertions (**->AG)
- 4 matches (GATA -- GATA)
- **1 deletion (G->*)**

Corresponding SAM format:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

Alignment Section (Example - r001)

Alignment record:

```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                 ATAGCT..............TCAGC
r003-                        ~~ttagct~~TAGGC
r001-
```

- 8 matches (TTAGATTA --TTAGATAA)
- 2 insertions (**->AG)
- 4 matches (GATA -- GATA)
- 1 deletion (G->*)
- **3 matches (CTG -- CTG)**

Corresponding SAM format:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG  *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA             * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC              * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCATCAGCGCCAT *
```

(Li H et al. **Bioinformatics**. 25(16), Aug. 2009)

SAMtools

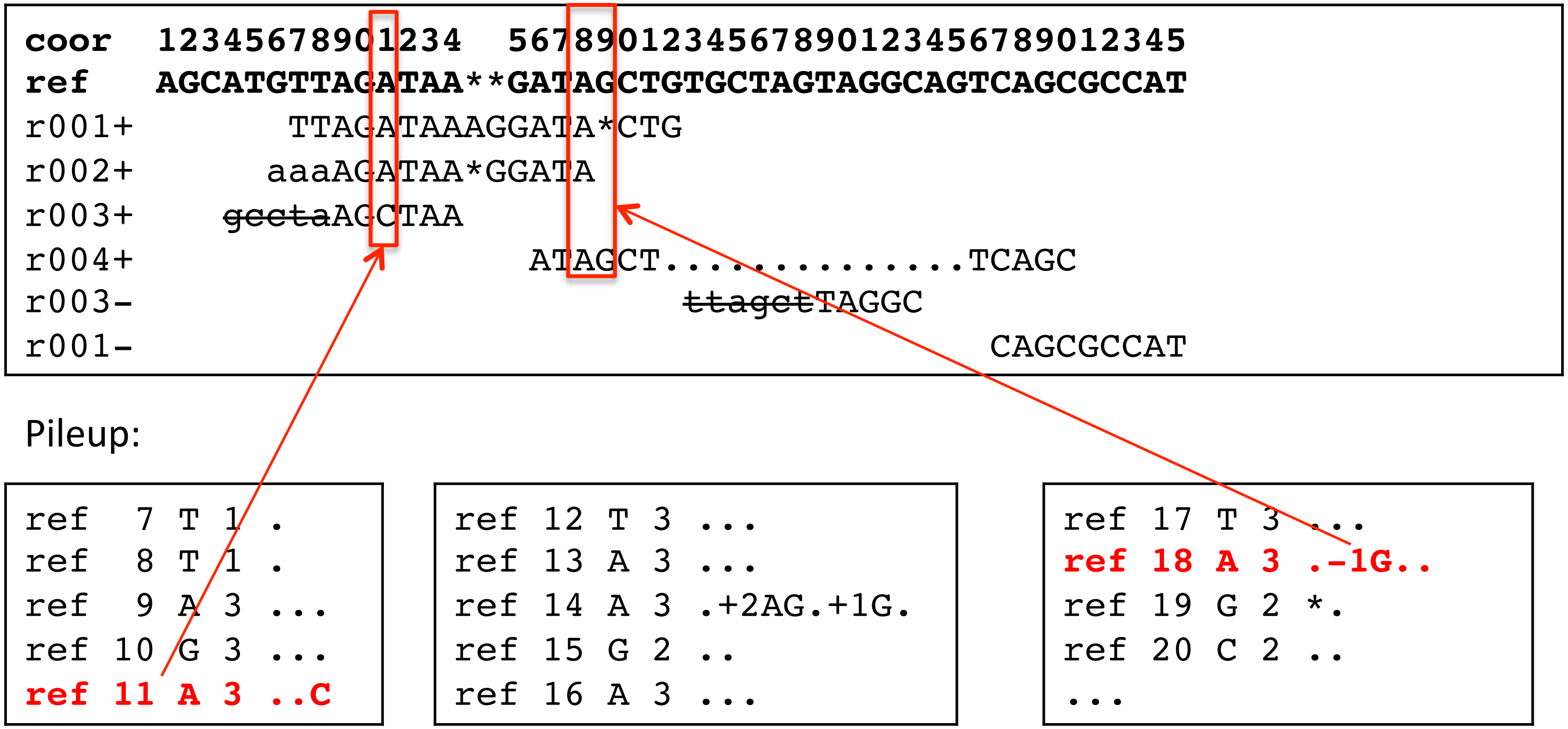
- SAMtools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format etc.

```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.16 (r963:234)

Usage:   samtools <command> [options]

Command:        view            SAM<->BAM conversion
                sort            sort alignment file
                pileup    generate pileup output
                mpileup         multi-way pileup
                depth           compute the depth
                faidx           index/extract FASTA
                tview           text alignment viewer
                index     index alignment
                idxstats        BAM index stats (r595 or later)
                fixmate         fix mate information
                glfview         print GLFv3 file
                flagstat        simple stats
                calmd           recalculate MD/NM tags and '=' bases
                merge           merge sorted alignments
                rmdup           remove PCR duplicates
                reheader replace BAM header
                cat             concatenate BAMs
                targetcut cut fosmid regions (for fosmid pool only)
                phase           phase heterozygotes
```

Pileup Format

- Describe the base-pair information at each chromosomal position.
- Good for SNP/indel calling and brief alignment viewing by eyes.

Alignment record:



```
coor  12345678901234  5678901234567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001+       TTAGATAAAGGATA*CTG
r002+      aaaAGATAA*GGATA
r003+    ~~gccta~~AGCTAA
r004+                  ATAGCT..............TCAGC
r003-                        ~~ttagct~~TAGGC
r001-                                       CAGCGCCAT
```

Pileup:

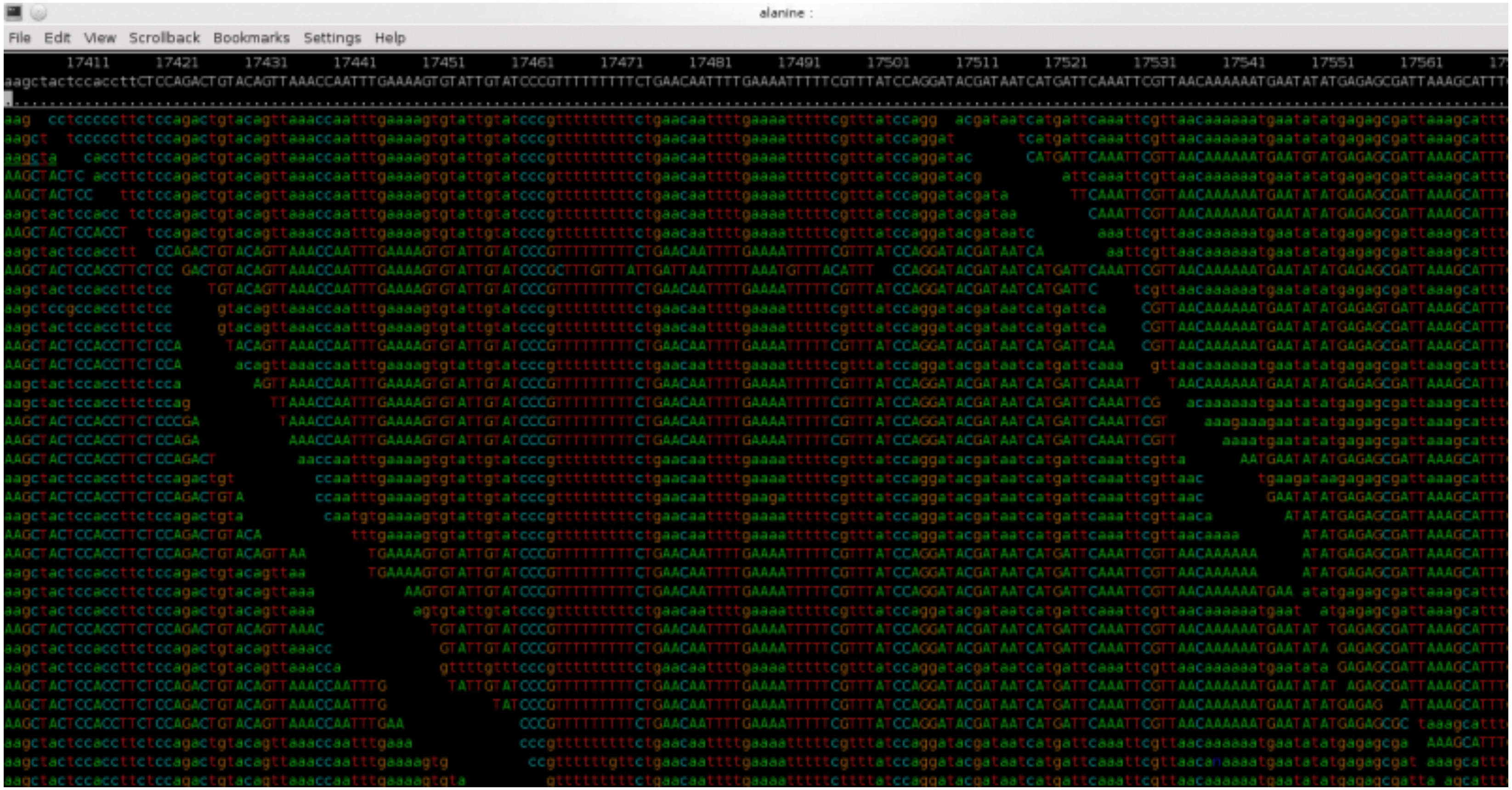
```
ref   7 T 1 .
ref   8 T 1 .
ref   9 A 3 ...
ref  10 G 3 ...
ref  11 A 3 ..C
```

```
ref 12 T 3 ...
ref 13 A 3 ...
ref 14 A 3 .+2AG.+1G.
ref 15 G 2 ..
ref 16 A 3 ...
```

```
ref 17 T 3 ...
ref 18 A 3 .-1G..
ref 19 G 2 *.
ref 20 C 2 ..
...
```

Alignment Visualization (1)

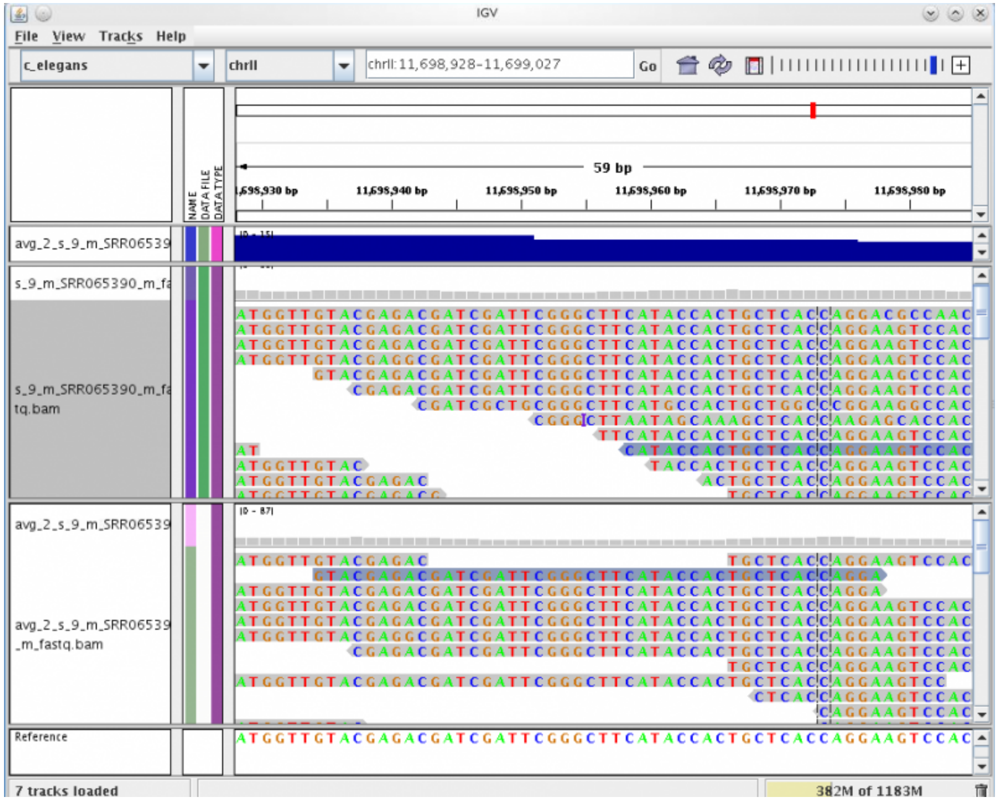
- SAMtools Text Alignment Viewer.



```
         17411      17421     17431     17441     17451     17461     17471     17481     17491     17501     17511     17521     17531     17541     17551     17561     17
aagctactccaccttCTCCAGACTGTACAGTTAAACCAATTTGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAAATGAATATATGAGAGCGATTAAAGCATTT
..................................................................................................................................................................................
aag  cctcccccttctccagactgtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccagg  acgataatcatgattcaaattcgttaacaaaaaaatgaatatatgagagcgattaaagcattt
aagct  tccccccttctccagactgtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggat       tcatgattcaaattcgttaacaaaaaaatgaatatatgagagcgattaaagcattt
aagcta   caccttctccagactgtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatac      CATGATTCAAATTCGTTAACAAAAAAATGAATGTATGAGAGCGATTAAAGCATTT
AAGCTACTC acccttctccagactgtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacg      attcaaattcgttaacaaaaaaatgaatatatgagagcgattaaagcattt
AAGCTACTCC   ttctccagactgtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgata      TTCAAATTCGTTAACAAAAAAATGAATATATGAGAGCGATTAAAGCATTT
aagctactccacc tctccagactgtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataa      CAAATTCGTTAACAAAAAAATGAATATATGAGAGCGATTAAAGCATTT
AAGCTACTCCACCT  tccagactgtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatc      aaattcgttaacaaaaaaatgaatatatgagagcgattaaagcattt
aagctactccacctt  CCAGACTGTACAGTTAAACCAATTTGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCA      aattcgttaacaaaaaaatgaatatatgagagcgattaaagcattt
AAGCTACTCCACCTTCTCC GACTGTACAGTTAAACCAATTTGAAAAGTGTATTGTATCCCGCTTTGTTTATTGATTAATTTTTAAATGTTTACATTT   CCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAAATGAATATATGAGAGCGATTAAAGCATTT
aagctactccaccttctcc     TGTACAGTTAAACCAATTTGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTC     tcgttaacaaaaaaatgaatatatgagagcgattaaagcattt
aagctccgccaccttctcc     gtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattca   CGTTAACAAAAAAATGAATATATGAGAGTGATTAAAGCATTT
aagctactccaccttctcc     gtacagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattca   CGTTAACAAAAAAATGAATATATGAGAGCGATTAAAGCATTT
AAGCTACTCCACCTTCTCCA     TACAGTTAAACCAATTTGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAA   CGTTAACAAAAAAATGAATATATGAGAGCGATTAAAGCATTT
AAGCTACTCCACCTTCTCCA      acagttaaaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaa   gttaacaaaaaaatgaatatatgagagcgattaaagcattt
aagctactccaccttctcca        AGTTAAACCAATTTGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATT   TAACAAAAAAATGAATATATGAGAGCGATTAAAGCATTT
aagctactccaccttctccag        TTAAACCAATTTGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCG   acaaaaaaatgaatatatgagagcgattaaagcattt
AAGCTACTCCACCTTCTCCCGA       TAAACCAATTTGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGT   aaagaaagaatatatgagagcgattaaagcattt
AAGCTACTCCACCTTCTCCAGA       AAACCAATTTGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTT   aaaatgaatatatgagagcgattaaagcattt
AAGCTACTCCACCTTCTCCAGACT     aaccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaattcgtta   AATGAATATATGAGAGCGATTAAAGCATTT
aagctactccaccttctccagactgt     ccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaattcgttaac   tgaagataagagagcgattaaagcattt
AAGCTACTCCACCTTCTCCAGACTGTA    ccaatttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaagatttttcgtttatccaggatacgataatcatgattcaaattcgttaac   GAATATATGAGAGCGATTAAAGCATTT
aagctactccaccttctccagactgta    caatgtgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaattcgttaaca   ATATATGAGAGCGATTAAAGCATTT
AAGCTACTCCACCTTCTCCAGACTGTACA    tttgaaaagtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaattcgttaacaaaa   ATATGAGAGCGATTAAAGCATTT
AAGCTACTCCACCTTCTCCAGACTGTACAGTTAA    TGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAA   ATATGAGAGCGATTAAAGCATTT
aagctactccaccttctccagactgtacagttaa    TGAAAAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAA   ATATGAGAGCGATTAAAGCATTT
aagctactccaccttctccagactgtacagttaaa    AAGTGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAAATGAA atatgagagcgattaaagcattt
aagctactccaccttctccagactgtacagttaaa    agtgtattgtatcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaattcgttaacaaaaaaatgaat  atgagagcgattaaagcattt
AAGCTACTCCACCTTCTCCAGACTGTACAGTTAAAC    TGTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAAATGAATAT  TGAGAGCGATTAAAGCATTT
aagctactccaccttctccagactgtacagttaaacc    GTATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAAATGAATATA GAGAGCGATTAAAGCATTT
aagctactccaccttctccagactgtacagttaaacca   gttttgtttcccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaattcgttaacaaaaaaatgaatata GAGAGCGATTAAAGCATTT
AAGCTACTCCACCTTCTCCAGACTGTACAGTTAAACCAATTTG   TATTGTATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAAATGAATATAT AGAGCGATTAAAGCATTT
AAGCTACTCCACCTTCTCCAGACTGTACAGTTAAACCAATTTG     TATCCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAAATGAATATATGAGAG  ATTAAAGCATTT
AAGCTACTCCACCTTCTCCAGACTGTACAGTTAAACCAATTTGAA     CCCGTTTTTTTTTCTGAACAATTTTGAAAATTTTTCGTTTATCCAGGATACGATAATCATGATTCAAATTCGTTAACAAAAAAATGAATATATGAGAGCGC taaagcattt
aagctactccaccttctccagactgtacagttaaaccaatttgaaa     cccgtttttttttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaattcgttaacaaaaaaatgaatatatgagagcga  AAAGCATTT
aagctactccaccttctccagactgtacagttaaaccaatttgaaaagtg     ccgtttttttgttctgaacaattttgaaaatttttcgtttatccaggatacgataatcatgattcaaattcgttaacaaaaaaatgaatatatgagagcgat aaagcattt
aagctactccaccttctccagactgtacagttaaaccaatttgaaaagtgta     gtttttttttctgaacaattttgaaaatttttctttttatccaggatacgataatcatgattcaaattcgttaacaaaaaaatgaatatatgagagcgatta agcatt
```

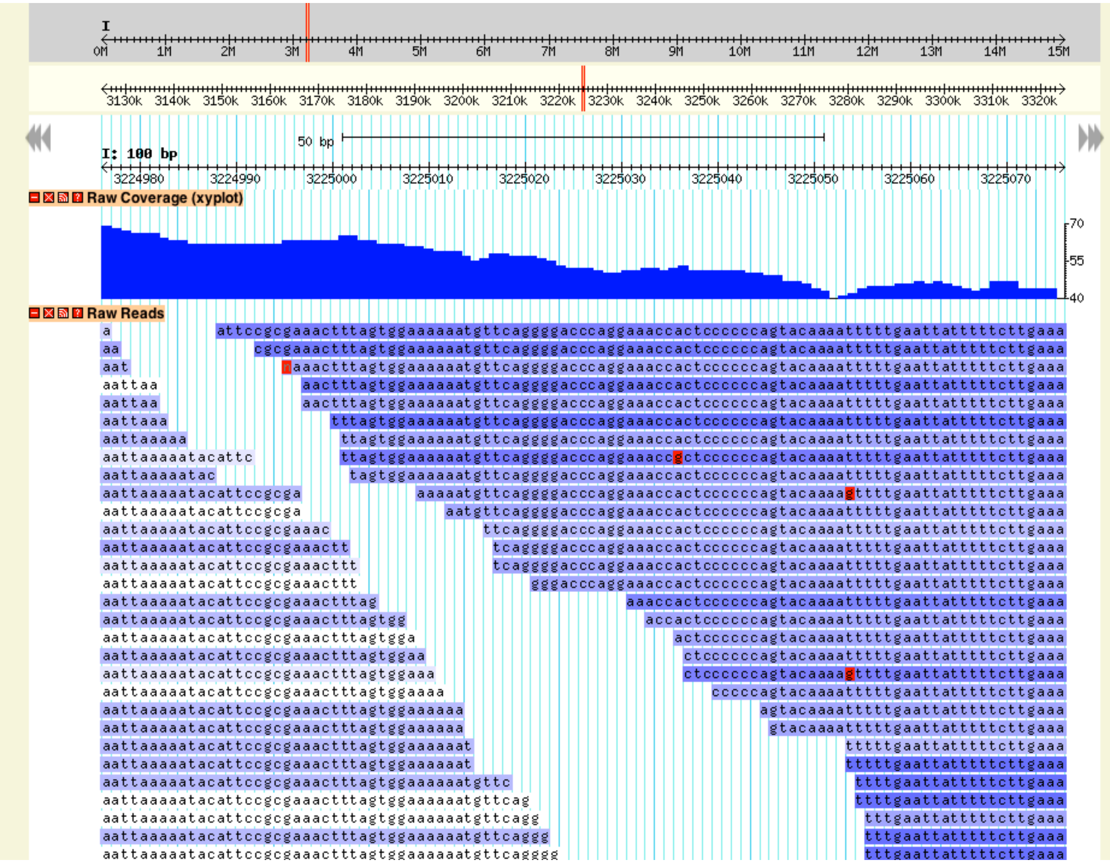
Alignment Visualization (2)

- Integrative Genomics Viewer (http://www.broadinstitute.org/igv/)
- High-performance visualization tool for interactive exploration of large, integrated genomic datasets.
- Supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotation.



Alignment Visualization (3)

- GBrowse (Generic Genome Browse, http://gmod.org/wiki/Gbrowse)
- Combination of database and interactive web pages for manipulating and displaying annotations on genomes.



Summary

- Transcriptome assembly strategies
- Short read aligners
- Alignment format and SAMtools
- Alignment visualization

References

- Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
- Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. Genetics. 2010 Jun;185(2):405-16.
- Au KF et al. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. 2010 Aug;38(14):4570-8.
- Bolstad, B. M., et al. (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2) ,pp 185-193
- Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010 Feb 18;11:94.
- Chen G, Wang C, Shi T. Overview of available methods for diverse RNA-Seq data analyses. Sci China Life Sci. 2011 Dec;54(12):1121-8.
- De Bona et al. Optimal spliced alignments of short sequence reads. Bioinformatics. 2008 Aug 15;24(16):i174-80.
- Garber M et al. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011 Jun;8(6):469-77.
- Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011).
- Guttman M. et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010 May;28(5):503-10.
- Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010 Aug 10;11:422.
- Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform first published online April 19, 2012 doi:10.1093/bib/bbs017.
- Langmead B et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
- Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008 Nov;18(11): 1851-8.
- Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. 2009. Bioinformatics, 25:1754-60.
- Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.
- Li R et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009 Aug 1;25(15):1966-7.

References (Cont.)

- Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011 Sep 7;12(10):671-82. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009 Aug 1;25(15):1966-7.
- Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011 Jun;21(6): 936-9.
- Martin J et al. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics. 2010 Nov 24;11:663.
- Oshlack A et al. From RNA-seq reads to differential expression results. Genome Biol. 2010;11(12):220.
- Roberts A et al. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011 Sep 1;27(17):2325-9.
- Robertson G. et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010 Nov;7(11):909-12.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40.
- Schulz MH et al. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012 Apr 15;28(8): 1086-92. Epub 2012 Feb 24.
- Tarazona S et al. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011 Dec;21(12):2213-23.
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11.
- Trapnell C et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 Mar 1;7(3):562-78.
- Trapnell C et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.
- Wang L et al. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010 Jan 1;26(1):136-8.
- Wang K et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010 Oct;38(18):e178.
- Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010 Apr 1;26(7):873-81.