Advanced Sequencing Technologies & Applications

http://meetings.cshl.edu/courses.html
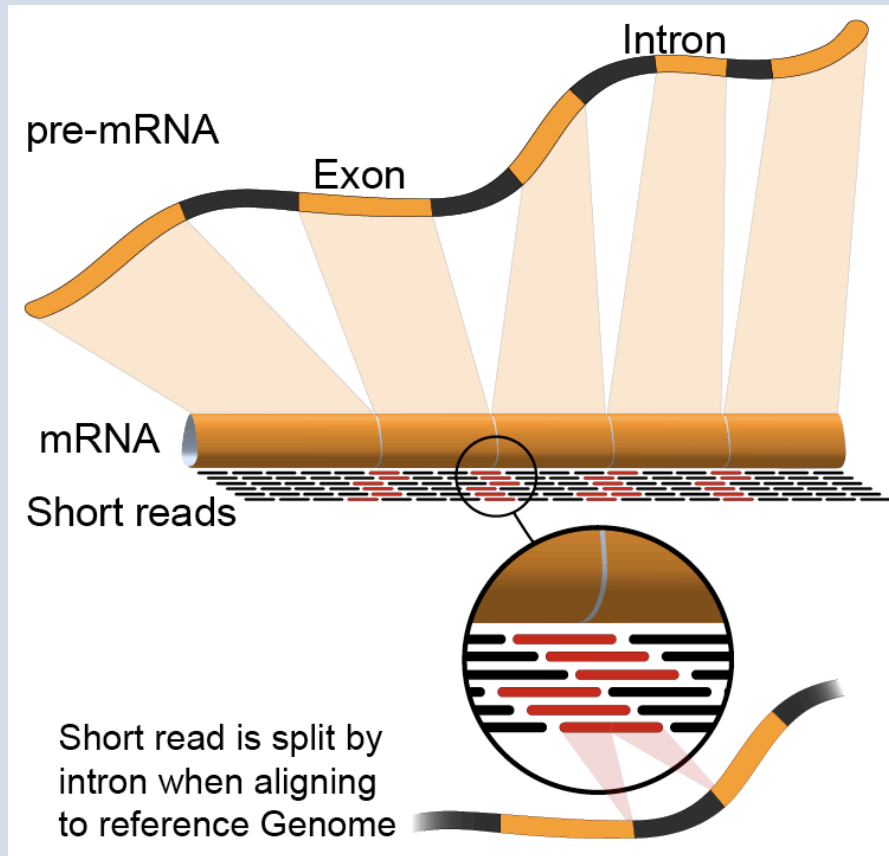
Learning objectives of the course

- Module 1: Introduction to cloud computing
- Module 2: Introduction to RNA sequencing
- **Module 3: RNA-seq alignment and visualization**
- Module 4: Expression and Differential Expression
- Module 5: Isoform discovery and alternative expression

- Tutorials
  - Provide a working example of an RNA-seq analysis pipeline
  - Run in a 'reasonable' amount of time with modest computer resources
  - Self contained, self explanatory, portable

Learning Objectives of Module

- RNA-seq alignment challenges and common questions

- Alignment strategies

- Bowtie/TopHat

- Introduction to the BAM and BED formats

- Basic manipulation of BAMs

- Visualization of RNA-seq alignments in IGV

- BAM read counting and determination of variant allele expression status

RNA-seq alignment challenges

- Computational cost
  - 100's of millions of reads

- Introns!
  - Spliced vs. unspliced alignments

- Can I just align my data once using one approach and be done with it?
  - Unfortunately probably not

- Is TopHat the only mapper to consider for RNA-seq data?
  - http://www.biostars.org/p/60478/

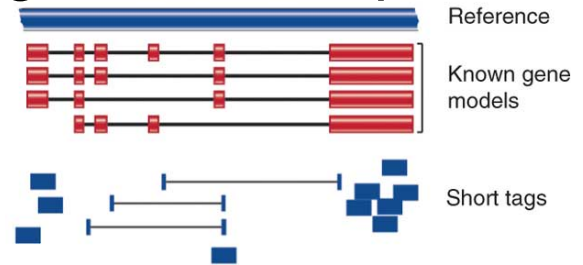# Three RNA-seq mapping strategies

De novo assembly

Assemble transcripts from overlapping tags

Reference

Optional: align to genome to get exon structure

Align to transcriptome

Reference

Known gene models

Short tags

Use known and/or predicted gene models to examine individual features

Align to reference genome

Reference

Short tags

Infer possible transcripts and abundance

Diagrams from Cloonan & Grimmond, Nature Methods 2010

Which alignment strategy is best?

- De novo assembly
  - If a reference genome does not exist for the species being studied
  - If complex polymorphisms/mutations/haplotypes might be missed by comparing to the reference genome
- Align to transcriptome
  - If you have short reads (< 50bp)
- Align to reference genome
  - All other cases

- Each strategy involves different alignment/assembly tools

# Which read aligner should I use?
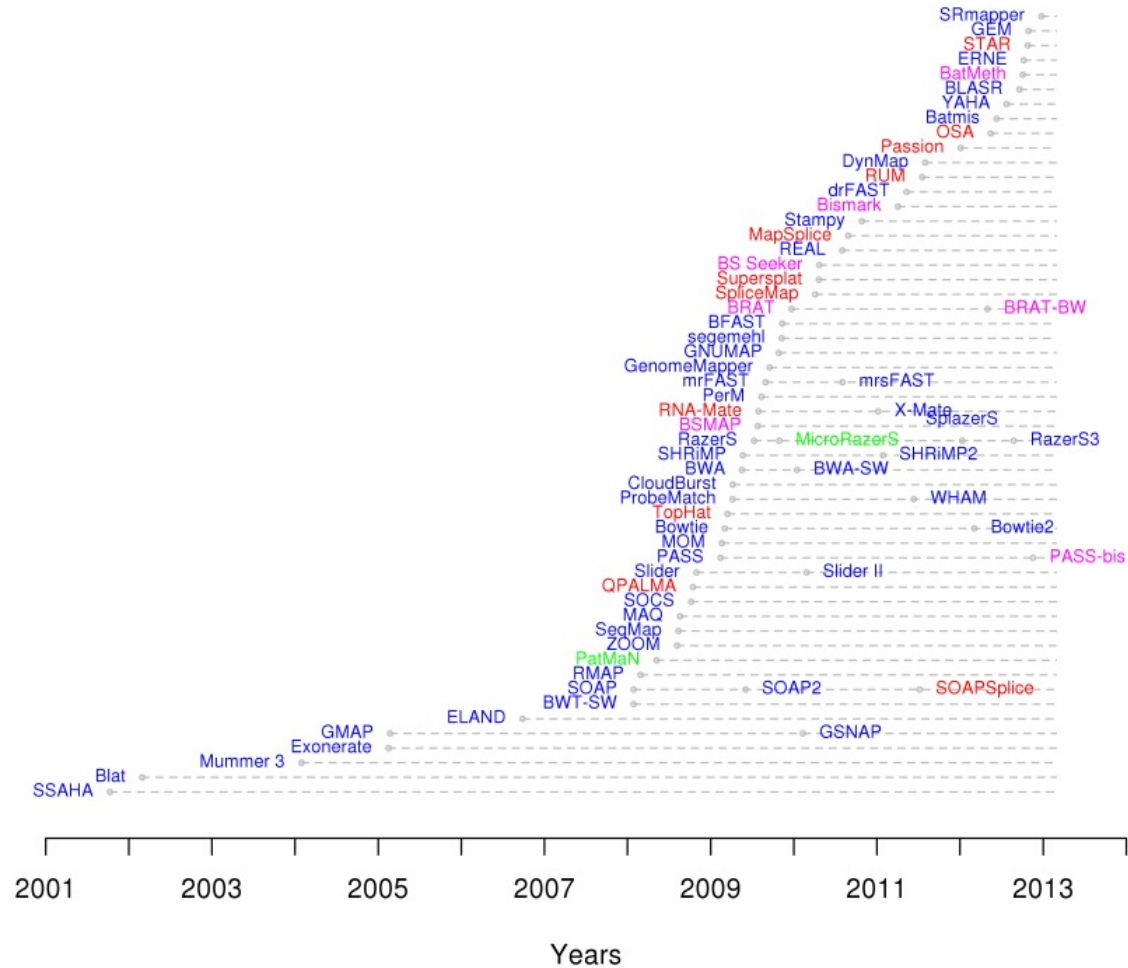
http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Should I use a splice-aware or unspliced mapper

- RNA-seq reads may span large introns
- The fragments being sequenced in RNA-seq represent mRNA and therefore the introns are removed
- But we are usually aligning these reads back to the reference genome
- Unless your reads are short (<50bp) you should use a splice-aware aligner
  - TopHat, STAR, MapSplice, etc.

Double-stranded genomic DNA template
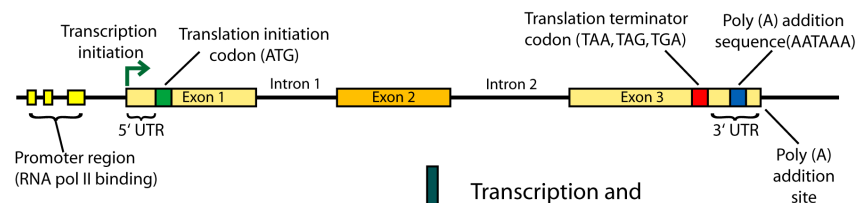
Single-stranded pre-mRNA (nuclear RNA)
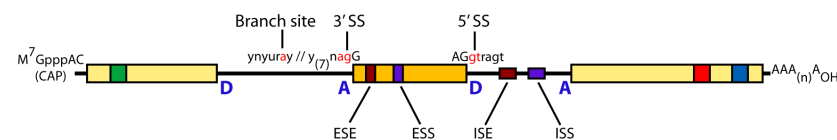
Mature mRNA
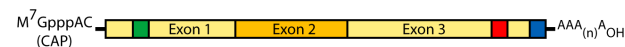
Bowtie/TopHat

- TopHat is a 'splice-aware' RNA-seq read aligner

- Requires a reference genome

- Breaks reads into pieces, uses 'bowtie' aligner to first align these pieces

- Then extends alignments from these seeds and resolves exon edges (splice junctions)
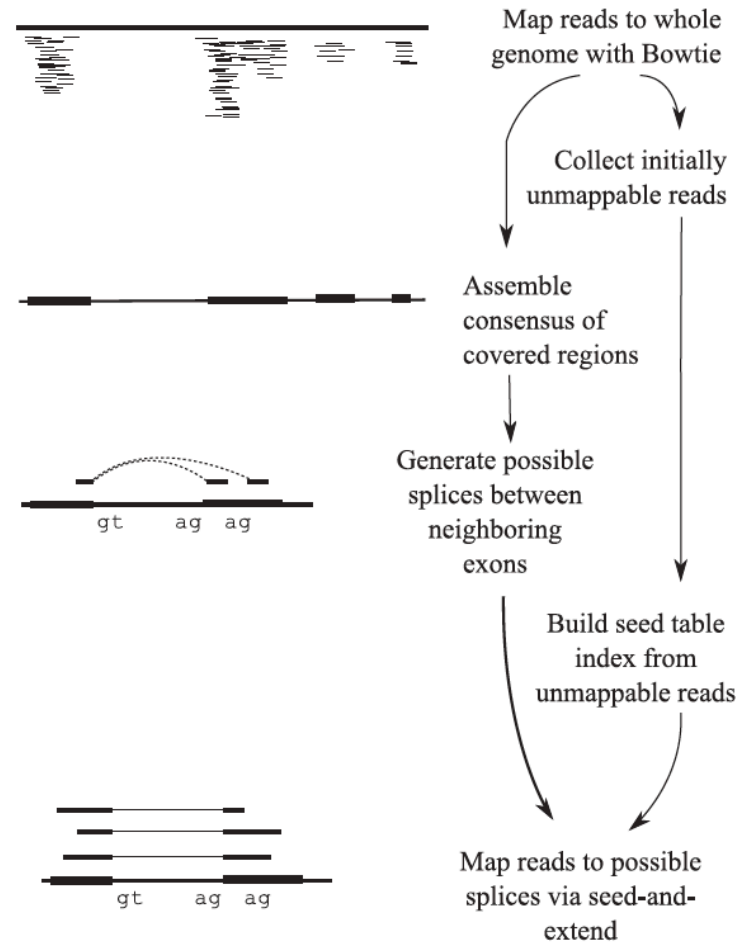
Trapnell et al. 2009

Should I allow 'multi-mapped' reads?

- Depends on the application

- In *DNA* analysis it is common to use a mapper to randomly select alignments from a series of equally good alignments

- In *RNA* analysis this is less common
  - Perhaps disallow multi-mapped reads if you are variant calling
  - Definitely should allow multi-mapped reads for expression analysis with TopHat/Cufflinks
  - Definitely should allow multi-mapped reads for gene fusion discovery

What is the output of bowtie/tophat?

- A SAM/BAM file
  - SAM stands for Sequence Alignment/Map format
  - BAM is the binary version of a SAM file

- Remember, compressed files require special handling compared to plain text files

- How can I convert BAM to SAM?
  - http://www.biostars.org/p/1701/

Example of SAM/BAM file format

Example SAM/BAM header section (abbreviated)

Example SAM/BAM alignment section (only 10 alignments shown)

Introduction to the SAM/BAM format

- The specification
  - http://samtools.sourceforge.net/SAM1.pdf
- The SAM format consists of two sections:
  - Header section
    - Used to describe source of data, reference sequence, method of alignment, etc.
  - Alignment section
    - Used to describe the read, quality of the read, and nature alignment of the read to a region of the genome
- BAM is a compressed version of SAM
  - Compressed using lossless BGZF format
  - Other BAM compression strategies are a subject of research. See 'CRAM' format for example
- BAM files are usually 'indexed'
  - A '.bai' file will be found beside the '.bam' file
  - Indexing aims to achieve fast retrieval of alignments overlapping a specified region without going through the whole alignments. BAM must be sorted by the reference ID and then the leftmost coordinate before indexing

SAM/BAM header section

- Used to describe source of data, reference sequence, method of alignment, etc.
- Each section begins with character '@' followed by a two-letter record type code. These are followed by two-letter tags and values
  - @HD The header line
    - VN: format version
    - SO: Sorting order of alignments
  - @SQ Reference sequence dictionary
    - SN: reference sequence name
    - LN: reference sequence length
    - SP: species
  - @RG Read group
    - ID: read group identifier
    - CN: name of sequencing center
    - SM: sample name
  - @PG Program
    - PN: program name
    - VN: program version

SAM/BAM alignment section

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

Example values

```
1     QNAME   e.g.   HWI-ST495_129147882:1:2302:10269:12362 (QNAME)
2     FLAG    e.g.   99
3     RNAME   e.g.   1
4     POS     e.g.   11623
5     MAPQ    e.g.   3
6     CIGAR   e.g.   100M
7     RNEXT   e.g.   =
8     PNEXT   e.g.   11740
9     TLEN    e.g.   217
10    SEQ     e.g.   CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGT
11    QUAL    e.g.   CCCFFFFFHHHHHJJIJFIJJJJJJJJJJJHIJJJJJJJJIJJJJJGGHIJHIJJJJJJJJJGHGGIJJJJJJIJEEHHHHFFFFCDCDDDDDDDB@ACDD
```

SAM/BAM flags explained

- http://broadinstitute.github.io/picard/explain-flags.html
- 11 bitwise flags describing the alignment
- These flags are stored as a binary string of length 11 instead of 11 columns of data
- Value of '1' indicates the flag is set. e.g. 00100000000
- All combinations can be represented as a number from 0 to 2047 (i.e. $2^{11}-1$). This number is used in the BAM/SAM file. You can specify 'required' or 'filter' flags in samtools view using the '-f' and '-F' options respectively

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

Note that to maximize confusion, each bit is described in the SAM specification using its hexadecimal representation (i.e., '0x10' = 16 and '0x40' = 64).

CIGAR strings explained

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

- The CIGAR string is a sequence of base lengths and associated 'operations' that are used to indicate which bases align to the reference (either a match or mismatch), are deleted, are inserted, represent introns, etc.

- e.g. 81M859N19M
  - A 100 bp read consists of: 81 bases of alignment to reference, 859 bases skipped (an intron), 19 bases of alignment

Introduction to the BED format

- When working with BAM files, it is very common to want to examine a focused subset of the reference genome
  - e.g. the exons of a gene
- These subsets are commonly specified in 'BED' files
  - https://genome.ucsc.edu/FAQ/FAQformat.html#format1
- Many BAM manipulation tools accept regions of interest in BED format
- Basic BED format (tab separated):
  - Chromosome name, start position, end position
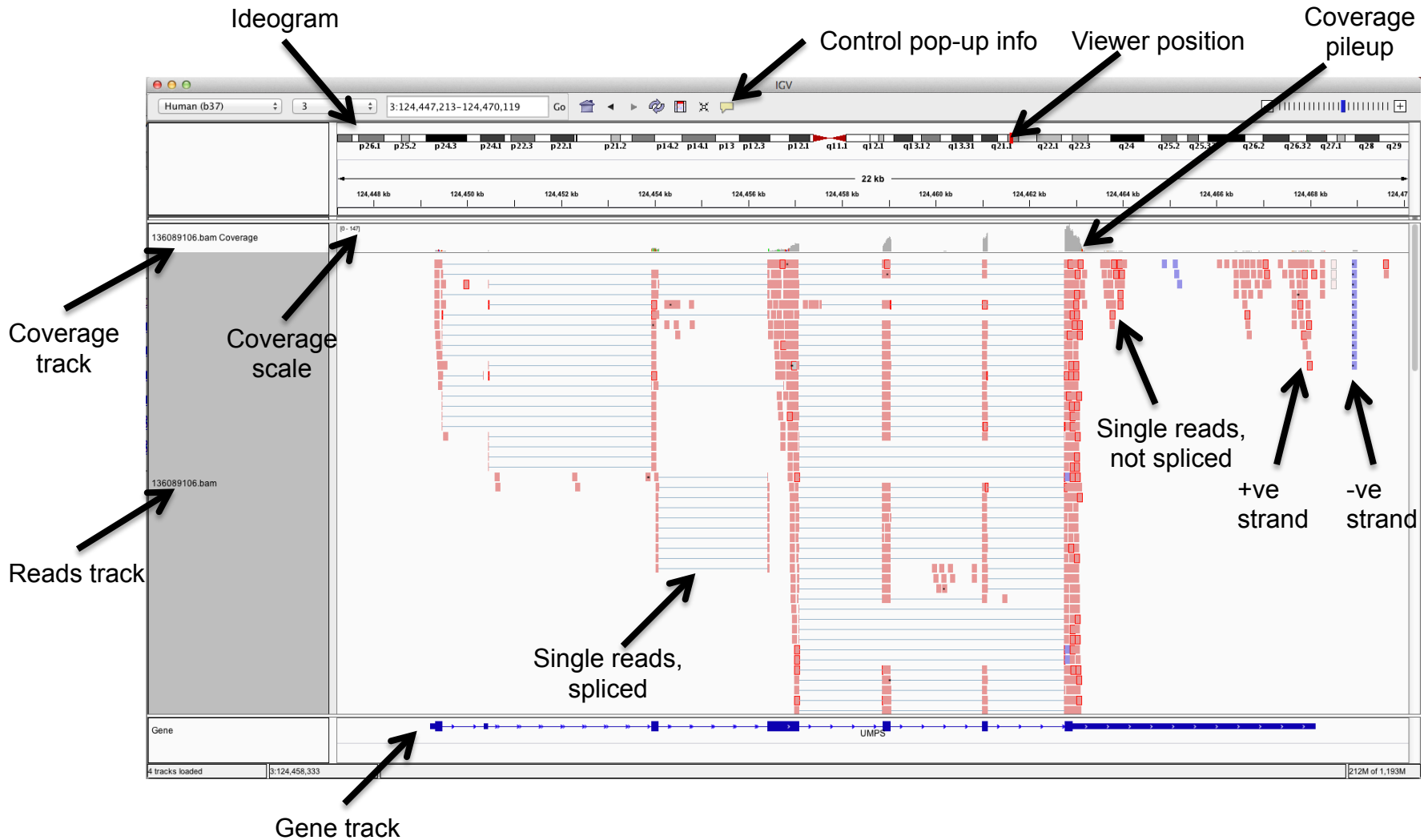 - Coordinates in BED format are 0 based

Manipulation of SAM/BAM and BED files

- Several tools are used ubiquitously in sequence analysis to manipulate these files

- SAM/BAM files
  - samtools
  - bamtools
  - picard

- BED files
  - bedtools
  - bedops

How should I sort my SAM/BAM file?

- Generally BAM files are sorted by <u>position</u>
  - This is for performance reasons
    - When sorted and indexed, arbitrary positions in a massive BAM file can be accessed rapidly

- Certain tools require a BAM sorted by <u>read name</u>
  - Usually this is when we need to easily identify both reads of a pair
    - The insert size between two reads may be large
    - In fusion detection we are interested in read pairs that map to different chromosomes...

Visualization of RNA-seq alignments in IGV browser

Ideogram

Control pop-up info

Viewer position

Coverage pileup

Coverage track

Coverage scale

Reads track

Single reads, not spliced

+ve strand

-ve strand

Single reads, spliced

Gene track

Alternative viewers to IGV

- Alternative viewers to IGV
  - http://www.biostars.org/p/12752/
  - http://www.biostars.org/p/71300/
- Artemis, BamView, Chipster, gbrowse2, GenoViewer, MagicViewer, **Savant**, Tablet, tview
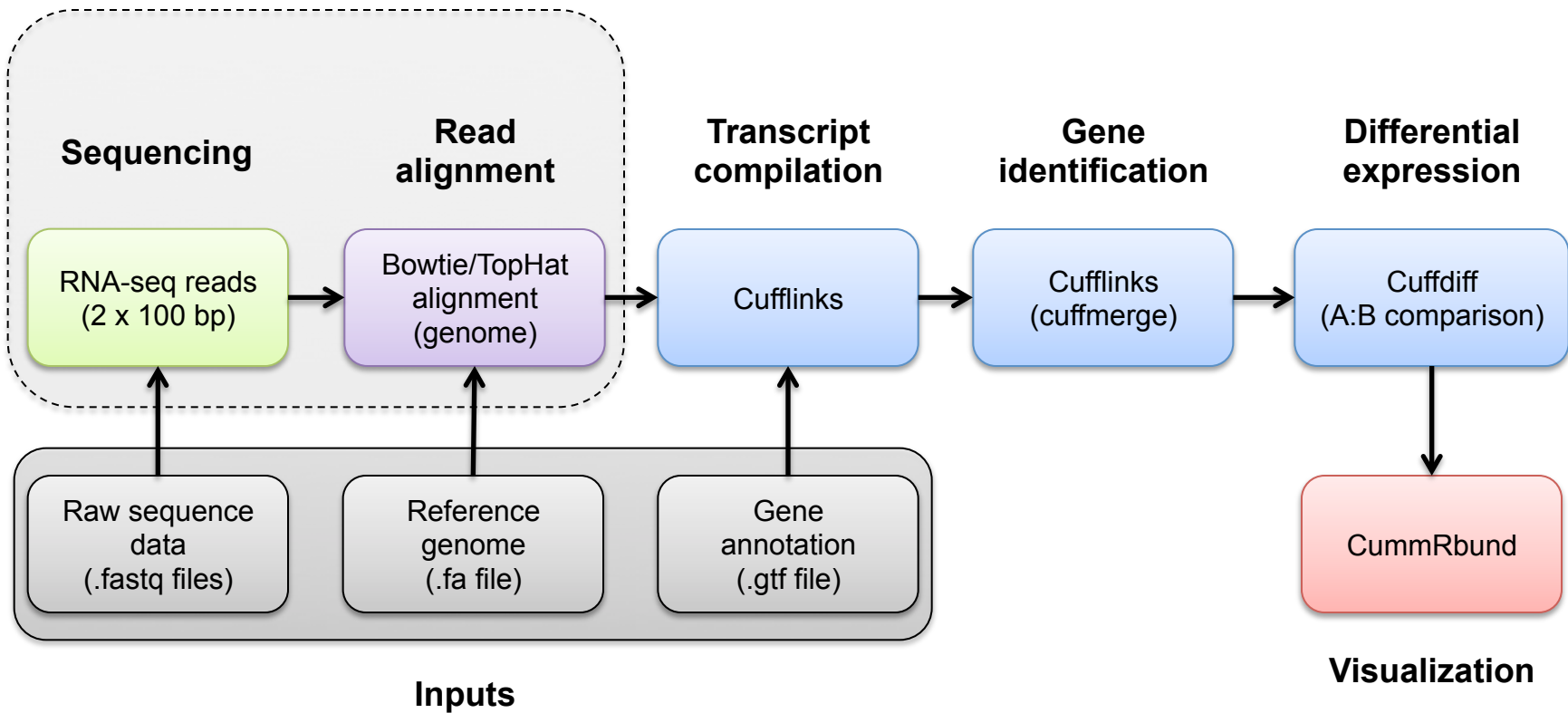
BAM read counting and variant allele expression status
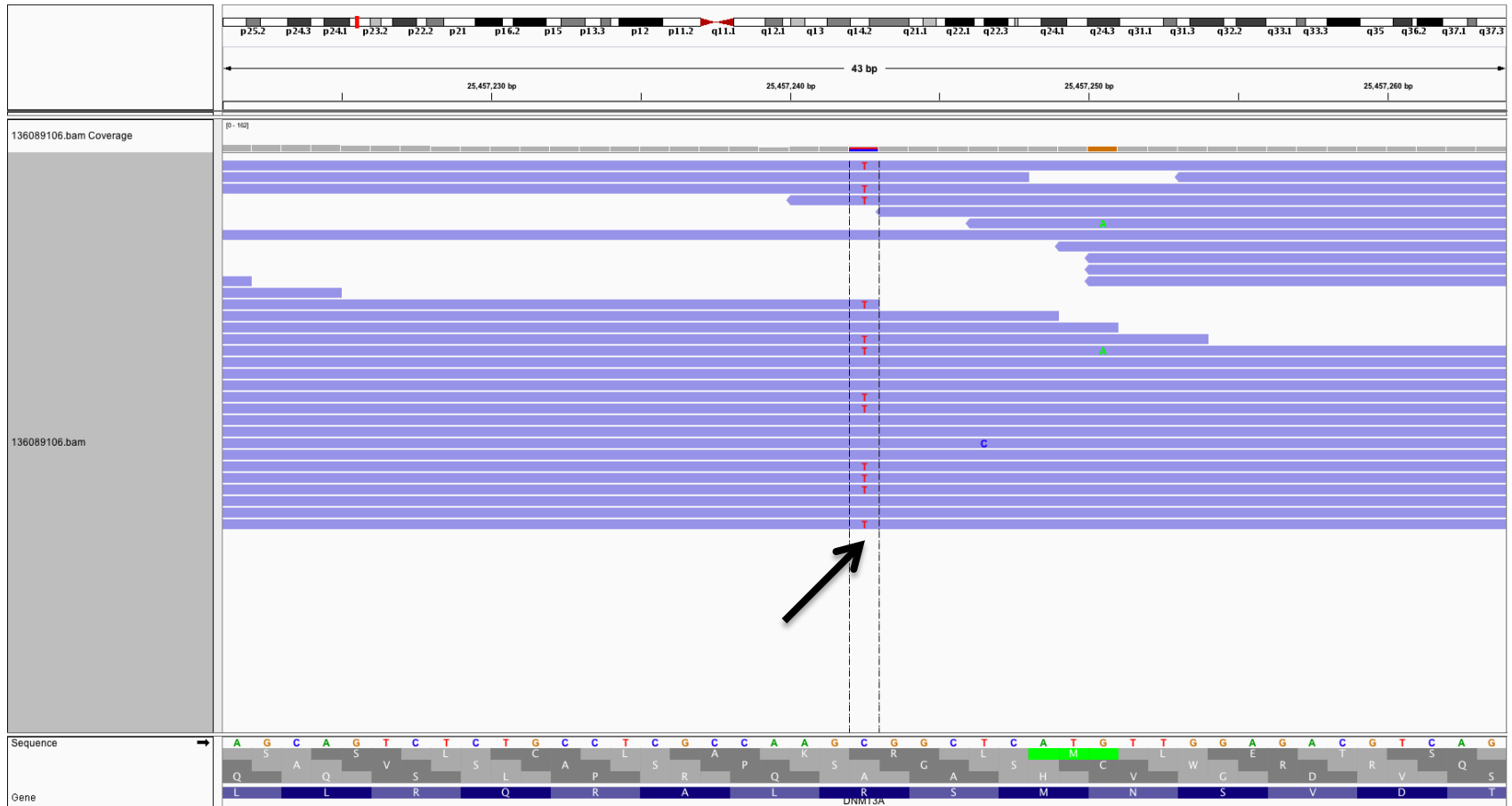
- A variant C->T is observed in 12 of 25 reads covering this position. Variant allele frequency (VAF) 12/25 = 48%.
- Both alleles appear to be expressed equally (not always the case) -> heterozygous, no allele specific expression
- How can we determine variant read counts, depth of coverage, and VAF without manually viewing in IGV?

Introduction to tutorial (Module 3)

Bowtie/Tophat/Cufflinks/Cuffdiff RNA-seq Pipeline

Break