

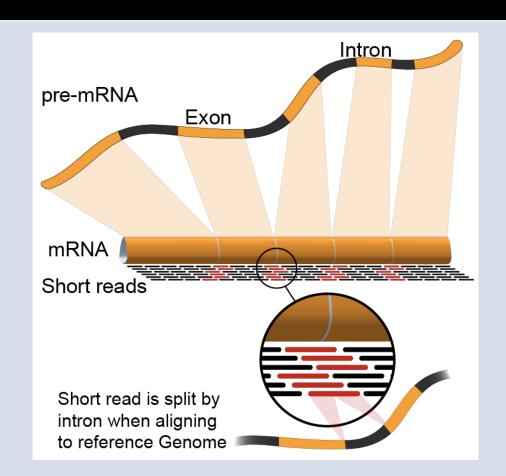
Advanced Sequencing Technologies & Applications

http://meetings.cshl.edu/courses.html



RNA-Seq Module 2 Alignment and Visualization (lecture)

Malachi Griffith, Obi Griffith, Jason Walker, Alex Wagner Advanced Sequencing Technologies & Applications November 7 - 20, 2016





Learning objectives of the course

- Module 0: Introduction to cloud computing
- Module 1: Introduction to RNA Sequencing
- Module 2: Alignment and Visualization
- Module 3: Expression and Differential Expression
- Module 4: 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 2

- RNA-seq alignment challenges and common questions
- Alignment strategies
- HISAT2
- Introduction to the BAM and BED formats
- Basic manipulation of BAMs
- Visualization of RNA-seq alignments in IGV
- Alignment QC Assessment
- 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 HISAT2 the only mapper to consider for RNA-seq data?
 - http://www.biostars.org/p/60478/

Three RNA-seq mapping strategies

De novo assembly

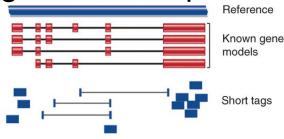


Assemble transcripts from overlapping tags



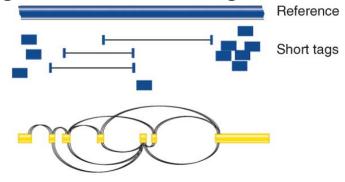
Optional: align to genome to get exon structure

Align to transcriptome



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

Align to reference genome



Infer possible transcripts and abundance

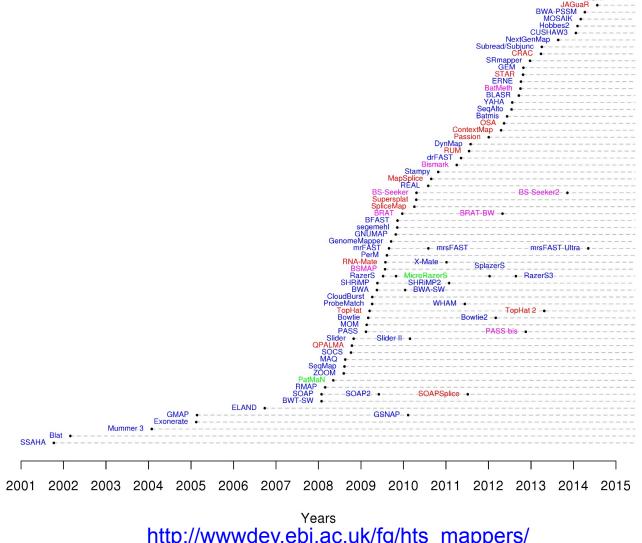
Diagrams from Cloonan & Grimmond, Nature Methods 2010

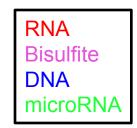
Which alignment strategy is best?

- 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?

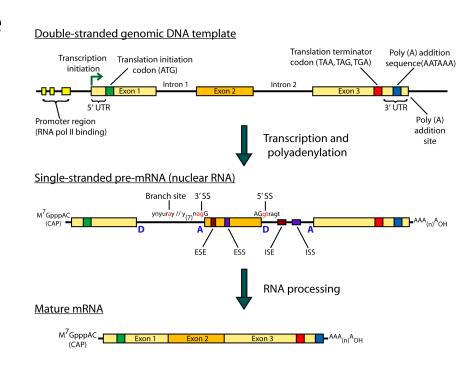




http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Should I use a splice-aware or unspliced mapper

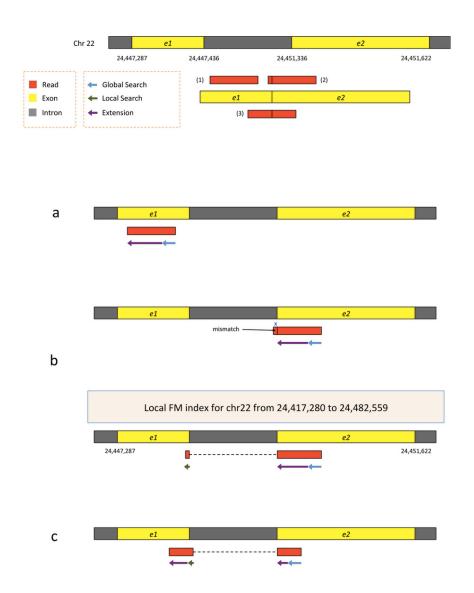
- RNA-seq reads may span large introns
- The fragments being sequenced in RNA-seq represent mRNA and therefore the introns are removed
- But we are usually aligning these reads back to the reference genome
- Unless your reads are short (<50bp) you should use a splice-aware aligner
 - HISAT2, STAR, MapSplice, etc.



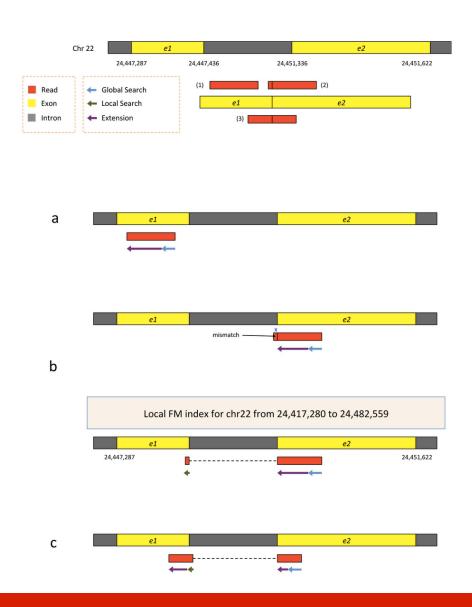
HISAT/HISAT2

- HISAT is a 'splice-aware' RNAseq read aligner
- Requires a reference genome
- Very fast
- Uses an indexing scheme based on the Burrows-Wheeler transform and the Ferragina-Manzini (FM) index
- two types of indexes for alignment
 - a whole-genome FM index to anchor each alignment
 - numerous local FM indexes for very rapid extensions of these alignments.

Kim et al. 2015. Nat Methods 12:357-360



HISAT/HISAT2



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 Cufflinks (and StringTie?)
 - Definitely should allow multi-mapped reads for gene fusion discovery

What is the output of HISAT2?

- 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)

```
mgriffit@linus270 -> samtools view -H /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam | grep -P "SN\:22|HD|RG|PG"
@HD
        VN:1.4 SO:coordinate
        SN:22 LN:51304566
                               UR:ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa.gz AS:GRCh37-lite M5:a718acaa6135fdca8357d5bfe9
4211dd SP:Homo sapiens
       ID:2888721359 PL:illumina
                                       PU:D1BA4ACXX.3 LB:H_KA-452198-0817007-cDNA-3-lib1
                                                                                               PI:365 DS:paired end DT:2012-10-03T19:00:00-0500
                                                                                                                                                       SM:H KA-452198-0817007 CN:WUGSC
       ID:2888721359
                       VN:2.0.8
                                       CL:tophat --library-type fr-secondstrand --bowtie-version=2.1.0
                                                                                               {\tt CL:net.sf.picard.sam.MarkDuplicates} \ \ INPUT=[/gscmnt/gc13001/info/build\_merged\_alignments/merged-alignment-blader] \\
       ID:MarkDuplicates
                               PN:MarkDuplicates
                                                       PP:2888721359 VN:1.85(exported)
e10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300.bam] OUTPUT=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jw
alker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300-post_dup.bam METRICS_FILE=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-1543
4-136080019/staging-liuJS/H_KA-452198-0817007-cDNA-3-lib1-2888360300.metrics REMOVE_DUPLICATES=false ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=9500 TMP DIR=[/qscmnt/qc13001/info/build merged al
ignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y] VALIDATION_STRINGENCY=SILENT_MAX_RECORDS_IN_RAM=5000000 PROGRAM_RECORD_ID=MarkDuplicates_PROGRAM_GROUP_NAME=Mark
DUPLICATE MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000 SORTING_COLLECTION_SIZE_RATIO=0.25 READ_NAME_REGEX=[a-za-Z0-9]+:[0-9]+):([0-9]+).* OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 VERBOSITY=INFO
QUIET=false COMPRESSION LEVEL=5 CREATE INDEX=false CREATE MD5 FILE=false
mgriffit@linus270 ~>
```

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

```
mgriffit@linus270 -> samtools view -f 3 -F 1804 /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam | head
HWI-ST495 129147882:3:2114:15769:38646 99
                                             11306 3
                                                          100M
                                                                                    ACTGCGGGGCCCTCTTGCTTACTGTATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCAGCACGC
                                                                       11508
                                      1
                                                                            302
CC:Z:15 MD:Z:5A94
                                                                                                       PG:Z:MarkDuplicates
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:
      XN:i:0 X0:i:0 CP:i:102519765 AS:i:-5 XS:A:+
                                                                                    HWI-ST495 129147882:3:2114:15769:38646 147
;5:CDCDCDECEFCD@9E=?7EEIIIIHCEGGIJJJJIIJJIHF@?00IHHFFGG?*JJJJJJJJJJJJJJJJJJHHCIEJJJHFHHGHFFEDFCCB
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:
                                                                                    CC:Z:15 MD:Z:34A65
                                                                                                       PG:Z:MarkDuplicates
      XN:i:0 X0:i:0 CP:i:102519563 AS:i:-6 XS:A:+ YT:Z:UU
HWI-ST495 129147882:3:1210:1257:16203 163
                                             11810 3
                                                          100M
                                                                       12055
                                                                            345
                                                                                    CCFFFFFHFHAFGGIIIJJJEEHGIGGGIJIJJGI?@EHIGIJDGHIHIGGIJJJJJJJJJJJGHHHGHFFFCDDDDDDCDCCCCCA;>@>@AA@:AA>AA
                                                                                    CC:Z:15 MD:Z:100
                                                                                                       PG:Z:MarkDuplicates
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
      XN:i:0 X0:i:0 CP:i:102519261 AS:i:0 XS:A:- YT:Z:UU
HWI-ST495 129147882:3:1210:1257:16203 83
                                                                                    GAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATAC
                                      1
                                             12055 3
CC:Z:15 MD:Z:100
                                                                                                       PG:Z:MarkDuplicates
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
      XN:i:0 X0:i:0 CP:i:102519016 AS:i:0 XS:A:+ YT:Z:UU
HWI-ST495 129147882:3:2111:3117:78828 163
                                                                                    @@FFFFFDHHHH9FHGIIFGAFDHEGII>GHIIIIIIIIIIIIIIIIIIFHDDFFEEECEECCCACCCCC:AADCCBCC>CAC<CCCCC:@CB@@BAB##
                                                                                    CC:Z:15 MD:Z:85G14
                                                                                                       PG:Z:MarkDuplicates
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:
      XN:i:0 X0:i:0 CP:i:102518437 AS:i:-5 XS:A:- YT:Z:UU
HWI-ST495 129147882:3:2111:3117:78828 83
                                             12746 3
                                                          100M
                                                                                    GGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCATCTCCTGTCTCCTGGAGAGGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTD
                                                                       12634
                                                                            -212
DCABDBDDDDDDDDDDDDDDBDB@BDDDB@;CCCCCDEFD@;.?<HIGGEIGEHIGJJJIIGIGIIHEGFEHFJIIIIIGJJJJHHHHHFFFFFC@
                                                                                    CC:Z:15 MD:Z:37G62
                                                                                                       PG:Z:MarkDuplicates
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:1 XM:i:
      XN:i:0 X0:i:0 CP:i:102518325 AS:i:-5 XS:A:-
HWI-ST495_129147882:3:1102:4242:26638
                               99
                                             13503 3
                                                                       13779
                                                                                    CGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAC
CCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJGIIIIJJFHGGIJGIJJJEGIJIJJHHIHHGHFFEFDEEEECCCAACDDACDCDDDDDB?8?<B>A@CDC
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
                                                                                    CC:Z:2 MD:Z:100
                                                                                                       PG:Z:MarkDuplicates
      XN:i:0 X0:i:0 CP:i:114357414 AS:i:0 XS:A:+ YT:Z:UU
HWI-ST495 129147882:3:1309:15328:74082 99
                                             13534 3
                                                          100M
                                                                       13780
                                                                                    AGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACAGG@
CC:Z:2 MD:Z:100
                                                                                                       PG:Z:MarkDuplicates
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
      XN:i:0 XO:i:0 CP:i:114357383 AS:i:0 XS:A:+ YT:Z:UU
HWI-ST495 129147882;3;1308;10126;19636 99
                                      1
                                             13779 3
                                                          100M
                                                                       14027 348
                                                                                    CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCCGATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTGC
CC:Z:2 MD:Z:100
                                                                                                       PG:Z:MarkDuplicates
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
      XN:i:0 XO:i:0 CP:i:114357140 AS:i:0 XS:A:+ YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638 147
                                      1
                                                                                    CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCCGATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTG#
##DCCDDDCCBBBABCCDDDCBDDBBDHC?=GIIJIIIIJIGIIIIJJHJJIJJIGCIIJJJJJJIGHGJJIJJJJJJJIJIIIIGGFGHHHHFFFFFCCC
                                                                                                                           RG:Z:2888721359 XG:i:0 NH:i:2 HI:i:0 NM:i:0 XM:i:
                                                                                    CC:Z:2 MD:Z:100
                                                                                                       PG:Z:MarkDuplicates
      XN:i:0 XO:i:0 CP:i:114357140 AS:i:0 XS:A:+ YT:Z:UU
mgriffit@linus270 <>
```

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
$\bigstar 2$	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

```
QNAME e.g. HWI-ST495 129147882:1:2302:10269:12362 (QNAME)
     FLAG
          e.g. 99
     RNAME e.g. 1
     POS
          e.g. 11623
     MAPQ
     CIGAR e.g. 100M
     RNEXT
     PNEXT e.g. 11740
     TLEN
10
     SEQ
          e.q. CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTTGGGCTGGGGCCTGGCCATGT
11
```

SAM/BAM flags explained

- http://broadinstitute.github.io/picard/explain-flags.html
- 12 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 1 to 2048 (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	
1	0x1	template having multiple segments in sequencing	
2	0x2	each segment properly aligned according to the aligner	
4	0x4	segment unmapped	
8	0x8	next segment in the template unmapped	
16	0x10	SEQ being reverse complemented	
32	0x20	SEQ of the next segment in the template being reverse complemented	
64	0x40	the first segment in the template	
128	0x80	the last segment in the template	
256	0x100	secondary alignment	
512	0x200	not passing filters, such as platform/vendor quality controls	
1024	0x400	PCR or optical duplicate	
2048	0x800	supplementary alignment	

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

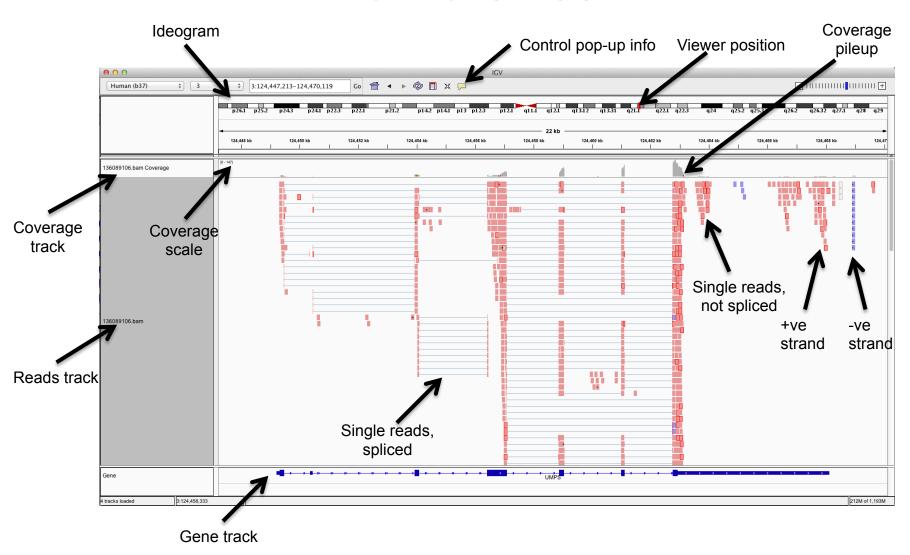
Manipulation of SAM/BAM and BED files

- 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



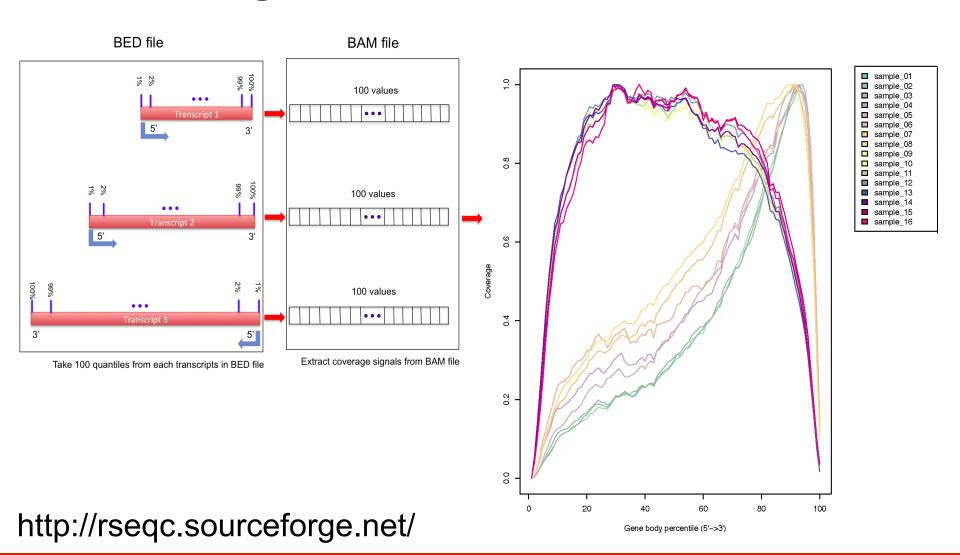
Alternative viewers to IGV

- 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

Alignment QC Assessment

- 3' and 5' Bias
- Nucleotide Content
- Base/Read Quality
- PCR Artifact
- Sequencing Depth
- Base Distribution
- Insert Size Distribution

Alignment QC: 3' & 5' Bias

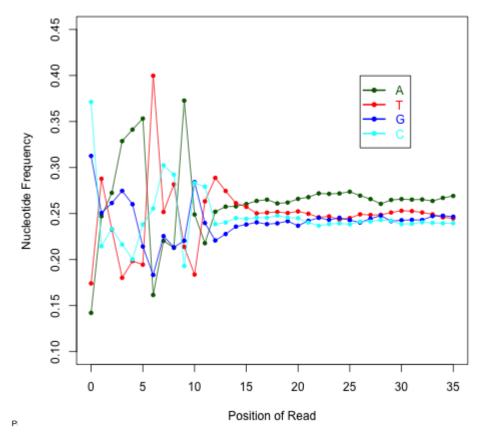


Alignment QC: Nucleotide Content

- Random primers are used to reverse transcribe RNA fragments into double-stranded complementary DNA (dscDNA)
- Causes certain patterns to be over represented at the beginning (5'end) of reads
- Deviation from expected A%=C %=G%=T%=25%

Journal List > Nucleic Acids Res > v.38(12); 2010 Jul > PMC2896536

Nucleic Acids Research



Nucleic Acids Res. 2010 Jul; 38(12): e131. Published online 2010 Apr 14. doi: 10.1093/nar/qkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen, 1,* Steven E. Brenner, 2 and Sandrine Dudoit 1,3

Author information ▶ Article notes ▶ Copyright and License information ▶

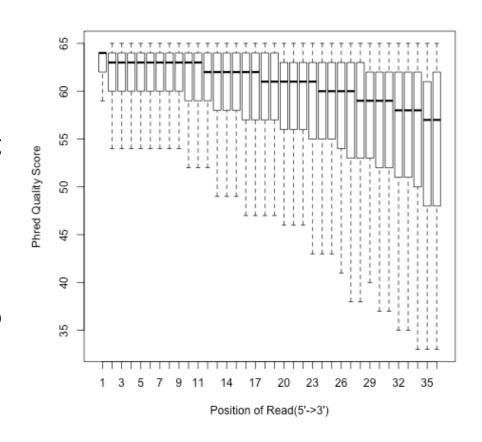
This article has been cited by other articles in PMC.

http://rseqc.sourceforge.net/

- Nucleic Acids Res

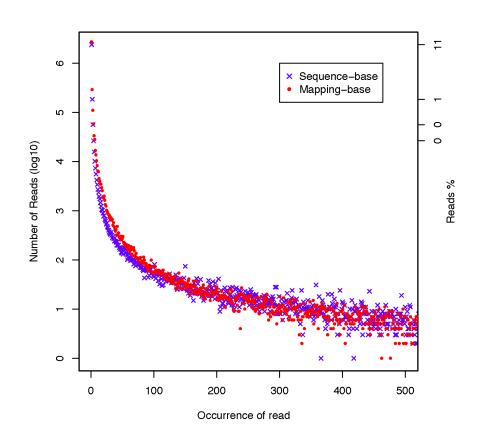
Alignment QC: Quality Distribution

- Phred quality score is widely used to characterize the quality of basecalling
- Phred quality score = -10xlog(10)P, here P is probability that base-calling is wrong
- Phred score of 30 means there is 1/1000 chance that the base-calling is wrong
- The quality of the bases tend to drop at the end of the read, a pattern observed in sequencing by synthesis techniques



Alignment QC: PCR Duplication

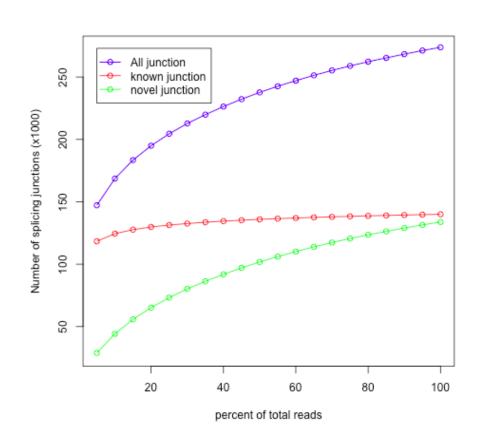
- Duplicate reads are reads that have the same start/end positions and same exact sequence
- In DNA-seq, reads/start point is used as a metric to assess PCR duplication rate
- In DNA-seq, duplicate reads are collapsed using tools such as picard
- How is RNA-seq different from DNAseq?



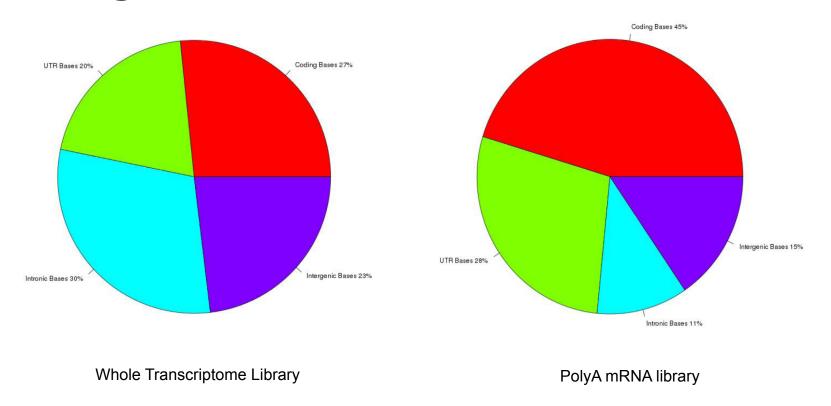
http://rseqc.sourceforge.net/

Alignment QC: Sequencing Depth

- Have we sequenced deep enough?
- In DNA-seq, we can determine this by looking at the average coverage over the sequenced region. Is it above a certain threshold?
- In RNA-seq, this is a challenge due to the variability in gene abundance
- Use splice junctions detection rate as a way to identify desired sequencing depth
- Check for saturation by resampling 5%, 10%, 15%, ..., 95% of total alignments from aligned file, and then detect splice junctions from each subset and compare to reference gene model.
- This method ensures that you have sufficient coverage to perform alternative splicing analyses



Alignment QC: Base Distribution

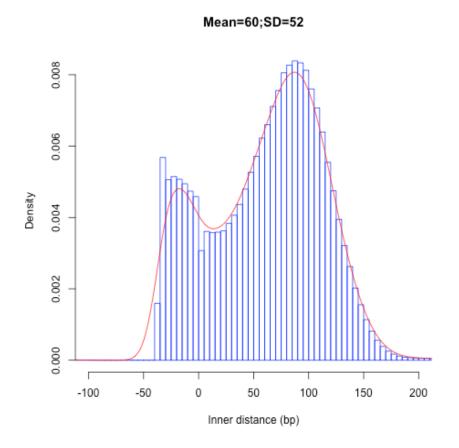


Your sequenced bases distribution will depend on the library preparation protocol selected

Alignment QC: Insert Size

http://thegenomefactory.blogspot.ca/2013/08/paired-end-read-confusion-library.html

Alignment QC: Insert Size



Consistent with library size selection?

http://rseqc.sourceforge.net/

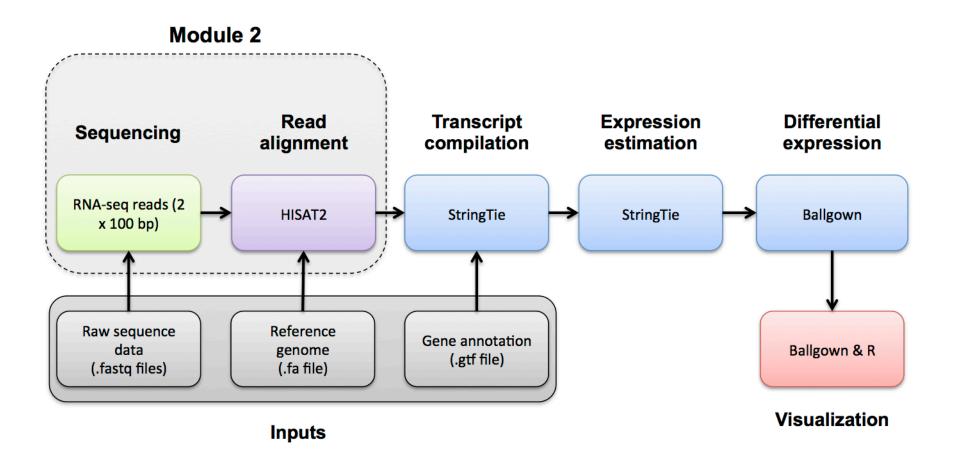
BAM read counting and variant allele expression status



- 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 2)

Bowtie/Tophat/Cufflinks/Cuffdiff RNA-seq Pipeline



Break