

RNA-Seq

Lalit Ponnala

CBSU

What is RNA-Seq

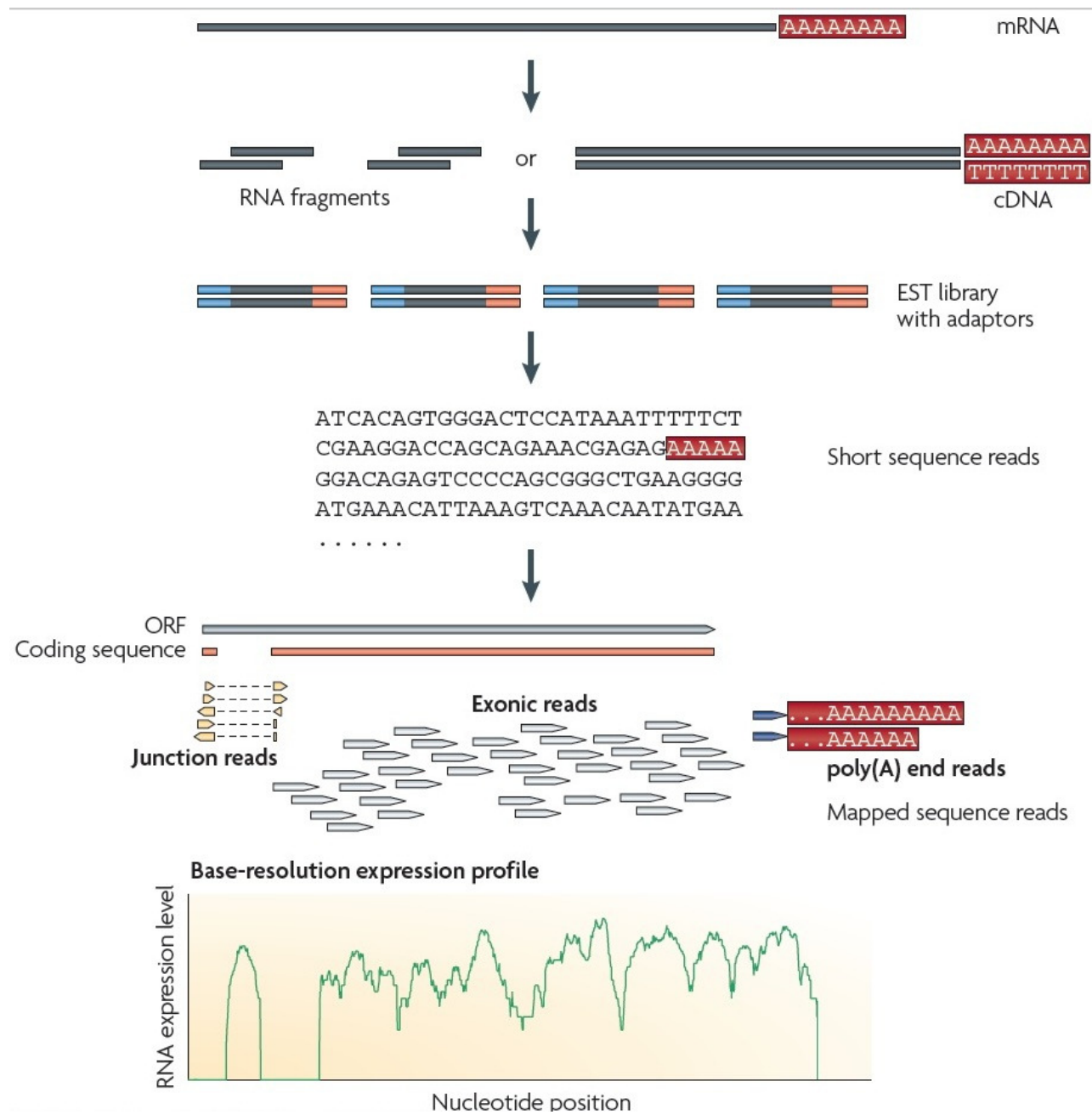
- Massively parallel sequencing method for transcriptome analyses
- Complementary DNA (cDNA) generated from RNA are sequenced using next-generation “short read” technologies
- Reads are aligned to a reference genome and a transcriptome map is constructed

Transcriptome

- The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition
- Understanding the transcriptome is essential for
 - interpreting the functional elements of the genome
 - revealing the molecular constituents of cells, tissues
 - understanding development and disease

Aims of RNA-Seq

- To quantify mRNA abundance
- To determine the transcriptional structure of genes: start sites, 5' and 3' ends, splicing patterns
- To quantify the changing expression levels of each transcript during development and under different conditions



RNA-Seq: a revolutionary tool for transcriptomics Nat Rev Genet. 2009 Jan;10(1):57-63.

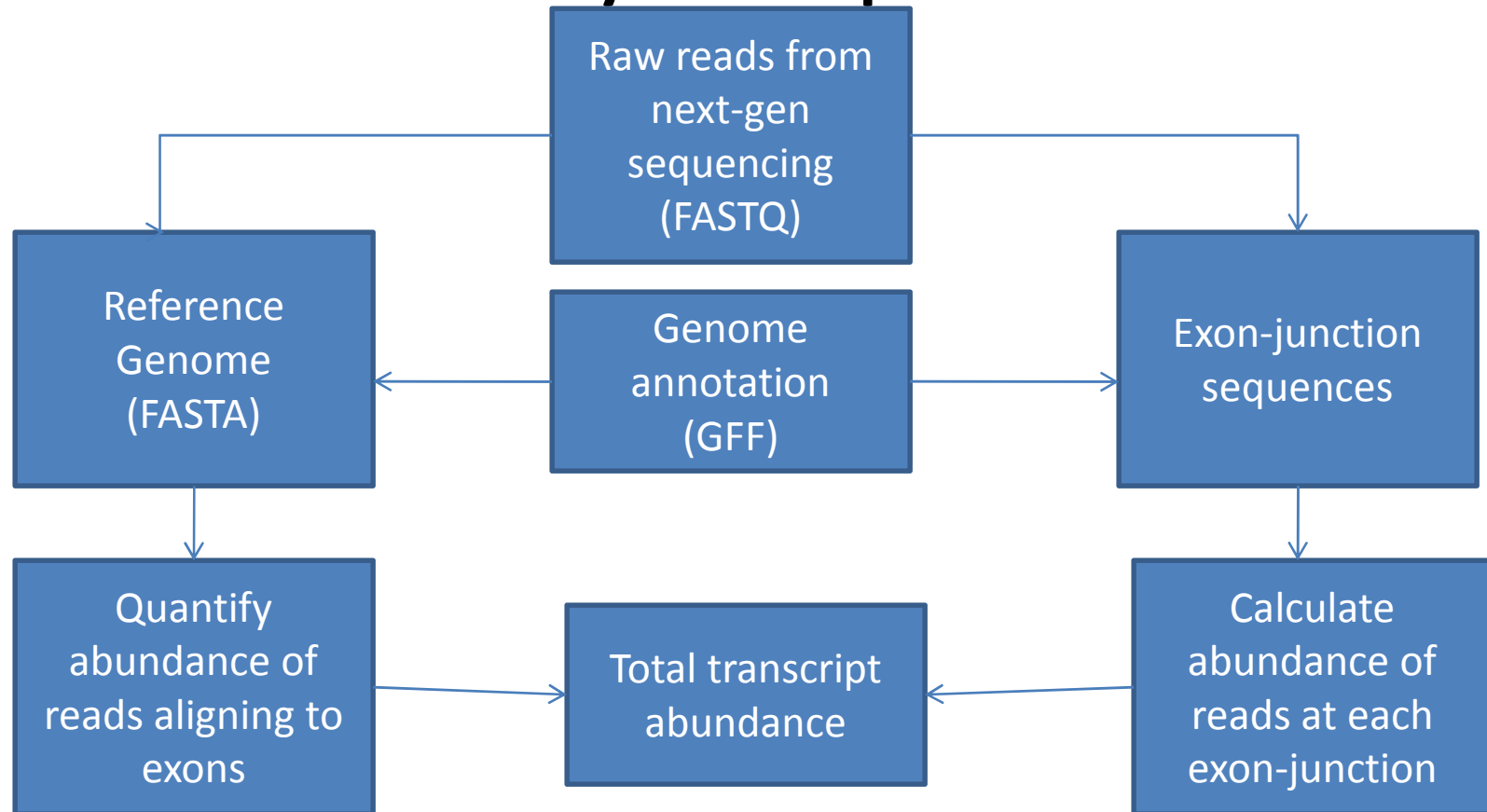
Technology

- Single-end, paired-end
- Typically 30-400bp reads
- Popular platforms: Illumina, 454, SOLID
- >10 million reads in a single “lane”
- Alignment tools: Bowtie, BWA, Eland etc
- Additional step: align to exon-junctions
- Automated pipeline for RNA-Seq:
 - **Tophat** : for alignment
 - **Cufflinks** : for calculating expression levels

Sequence data

```
ponnala@cbsuss04:~/data/tophat  
[ponnala@cbsuss04 tophat]$ more -10 s_1_sequence.txt  
@HWI-EAS83_20ECVAAXX:1:1:750:288  
TGAAGAAATTGAGTCTTCTAAGATGAATGTGAAAAG  
+HWI-EAS83_20ECVAAXX:1:1:750:288  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]XJOX]]C]]W[[[[[[C  
@HWI-EAS83_20ECVAAXX:1:1:851:310  
AGGATTCAACCCAGTTGTGCTAGAGCATCGACTCTT  
+HWI-EAS83_20ECVAAXX:1:1:851:310  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]T]]]]]MVC[V[N  
@HWI-EAS83_20ECVAAXX:1:1:1000:549  
TGCCCACTTGGTATATCCCTCAGAGGAGTGCCCTT  
+HWI-EAS83_20ECVAAXX:1:1:1000:549  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]C]]XX[V[[[N[[[[C  
@HWI-EAS83_20ECVAAXX:1:1:989:463  
ATTCTTCCAAAAACTTCCCTGATGTACCAGTCCCTTT  
+HWI-EAS83_20ECVAAXX:1:1:989:463  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]T]]K[[[D[N  
@HWI-EAS83_20ECVAAXX:1:1:1001:547  
TGGGTCTTGTAGGACTCAGCAGGAACATCAGCAAAG  
+HWI-EAS83_20ECVAAXX:1:1:1001:547  
[[[[X[[[[[[[[[[[[[[GX[[[[X[[YQ[[XGYYZYJ  
@HWI-EAS83_20ECVAAXX:1:1:765:512  
AAAGAAATATATTTTTCTAAGATCACAAATAACTGAA  
+HWI-EAS83_20ECVAAXX:1:1:765:512  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]G[[Q[T  
@HWI-EAS83_20ECVAAXX:1:1:979:558  
TAGAAGTCGATGGAATAAAAAGTTGCATCTTGACTT  
+HWI-EAS83_20ECVAAXX:1:1:979:558  
[[[[[[[[[[X[[[[X[[[[[[T[[VZ[[[[[[GTTTTJ  
@HWI-EAS83_20ECVAAXX:1:1:829:561  
AACACGGACACGCCTCGGCACACTGCGGATACCACT  
+HWI-EAS83_20ECVAAXX:1:1:829:561  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]XV]]RW[X[[  
@HWI-EAS83_20ECVAAXX:1:1:564:219  
CCCCCCCCCCCCCCCCACCCCCCCCCACCATCAAAG  
+HWI-EAS83_20ECVAAXX:1:1:564:219  
[[[[[[[[[[[[[[TXXQ[[CMZZRGPRQ[[CSJJCLGPCC  
@HWI-EAS83_20ECVAAXX:1:1:917:419  
ACCAGCTTCAGTTCAGCATCAAGACGCTCCCTCTCT  
+HWI-EAS83_20ECVAAXX:1:1:917:419  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]X]]Y[[[[[[[[XC[[[[H  
@HWI-EAS83_20ECVAAXX:1:1:1001:566  
TAGCAATCCATGTTTTATTCCCATTTGTTTTCCCT  
+HWI-EAS83_20ECVAAXX:1:1:1001:566  
[[[[[[[[[[[[[[[[[[[[[[N[[[[[[[[[[C[[T[[V  
@HWI-EAS83_20ECVAAXX:1:1:913:446  
AAACTTTCATCGAGTTGGATTTGGATATTTGCCTCT  
+HWI-EAS83_20ECVAAXX:1:1:913:446  
[[[[[[[[[[[[[[[[[[[[[[X[[[[[[[[[[[[[[[[[[[[[[F  
@HWI-EAS83_20ECVAAXX:1:1:820:517  
ATTCCATAGAAGATTACATTGTTTGTCTGCATTTTGT  
+HWI-EAS83_20ECVAAXX:1:1:820:517  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]X]]C]]F[[[[[[Z  
@HWI-EAS83_20ECVAAXX:1:1:677:257  
TGAGAAAATGGTTTGTGGCGTTGTTCATCCCTCCAT  
+HWI-EAS83_20ECVAAXX:1:1:677:257  
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]S]]C]]]]W[[V[N  
@HWI-EAS83_20ECVAAXX:1:1:976:629  
AATAAACATCTGTGGTTTGTGTGGCCTGAATCGTGT  
+HWI-EAS83_20ECVAAXX:1:1:976:629  
--More-- (0%)
```

Analysis Pipeline



Alignment Issues

- Exon Boundaries
 - What if exon is shorter than read?
- Multiple matches
 - Simple weighting
 - Evidence-based weighting

Units of measurement

- RPKM : Reads per kilobase per million mapped reads

1kb transcript with 1000 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped) will have

$$\text{RPKM} = 1000 / (1 * 8) = 125$$

- FPKM : for paired-end sequencing
 - A pair of reads constitute one fragment

Tophat

- Aligns sequences to the whole genome AND to exon-junctions
- Uses Bowtie, an ultrafast, memory-efficient short read aligner
- Output reported in SAM format
- Independently aligns segments of each read (default 25bp) allowing up to 2 mismatches
- Does not support indels / gapped alignments

<http://tophat.cbcb.umd.edu/index.html>

Tophat : junctions

- From supplied annotation file (GFF) or list of junction coordinates
- Without reference annotation
 - Sets of coverage islands : high coverage regions
 - Paired end reads: using genomic distance between mates
 - Segments of same read mapped far apart: “GT-AG” introns

Running Tophat

- Index the genome:

```
bowtie-build maize_pseudo.fa maize_pseudo
```

- Run tophat:

```
tophat -o zero -G annot.gff --no-novel-juncs maize_pseudo s_1_sequence.txt
```

- Output files:

- accepted_hits.sam
- annot.juncs
- junctions.bed

Viewing the alignments (IGV)

```
samtools faidx maize_pseudo.fa
```

```
samtools view -bt maize_pseudo.fa.fai -o accepted_hits.bam accepted_hits.sam
```

```
samtools index accepted_hits.bam
```



<http://samtools.sourceforge.net/>

<http://www.broadinstitute.org/igv/>

Cufflinks

- can estimate the abundances of the isoforms present in the sample, using either:
 - a known "reference" annotation
 - an ab-initio assembly of the transcripts
- constructs a set of transcripts that "explain" the reads observed in an RNA-Seq experiment
- Input: alignments in SAM format, annotation in GTF (optional)
- Output: assembled transfrags, genes

<http://cufflinks.cbcb.umd.edu/index.html>

Cufflinks

- Command line:

```
cufflinks -G annot.gtf accepted_hits.sam
```

- Output files:

- transcripts.gtf
- transcripts.expr
- genes.expr

Cuffdiff

- Differential expression at the transcript “isoform” level and at the gene level

cuffdiff annot.gtf ./zero/accepted_hits.sam ./one/accepted_hits.sam

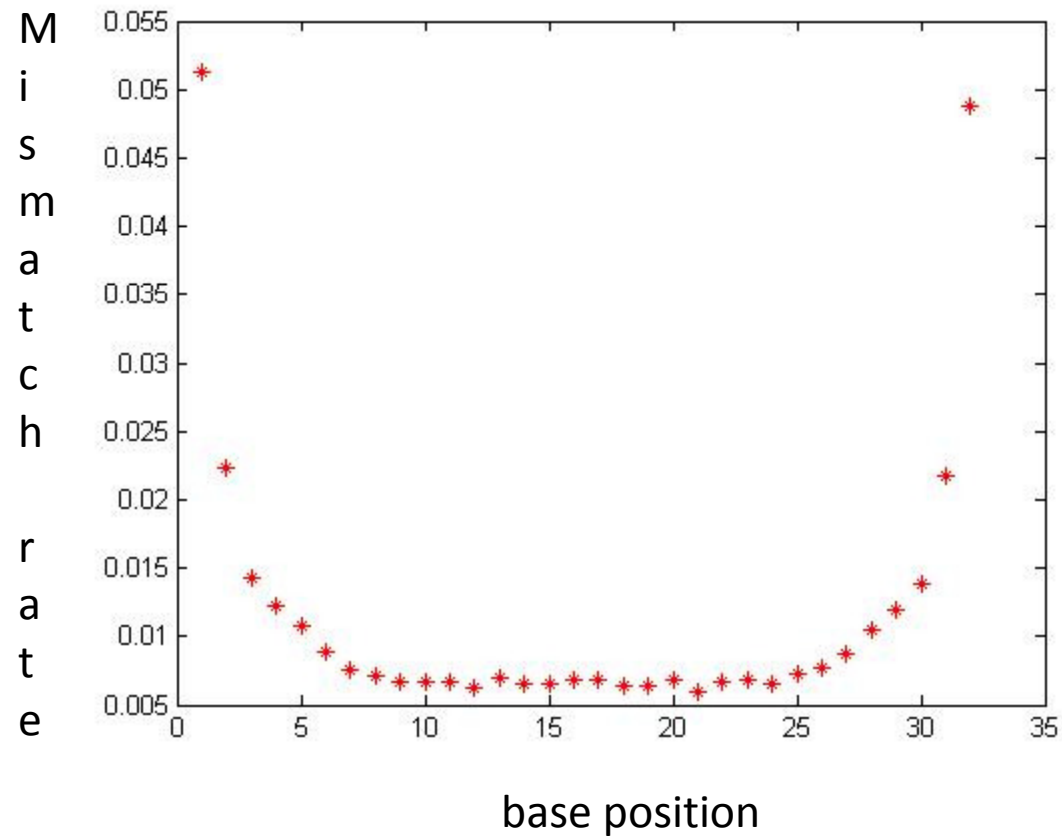
Examine output file: 0_1_gene_exp.diff

```
ponnala@localhost:~/3cpg
AC147602.5_FG002 - chr3:177051908-177053910 OK 10.9779 0 5.3613e-315 2.22507e-308 0 yes
s
AC147602.5_FG003 - chr3:177062944-177064394 OK 27.6166 830.416 3.40351 25.7555 0 yes
AC147602.5_FG004 - chr3:177072768-177074234 OK 16.499 563.468 3.53081 17.4231 0 yes
AC147602.5_FG005 - chr3:177118776-177126033 OK 3.75594 2.55511 -0.385243 -1.42538 0.154047
no
AC148152.3_FG001 - chr2:228345647-228347484 OK 0 0.544941 5.3613e-315 1.79769e+308 0
yes
AC148152.3_FG002 - chr2:228343615-228345268 OK 0 0 0 0 1 no
AC148152.3_FG005 - chr2:228269850-228271219 OK 66.1803 3.00847 -3.09095 -10.005 0 yes
AC148152.3_FG006 - chr2:228224832-228228214 OK 0 0 0 0 1 no
AC148152.3_FG008 - chr2:228196231-228200539 OK 2.09605 68.3161 3.48409 12.3901 0 yes
AC148167.6_FG001 - chr7:11551708-11555289 OK 21.4479 47.8814 0.803101 5.84087 5.19282e-09 yes
AC149475.2_FG002 - chr9:148163756-148166629 OK 118.037 3.92869 -3.4027 -8.46267 0 yes
AC149475.2_FG003 - chr9:148168982-148173108 OK 107.649 10.0885 -2.36748 -11.7115 0 yes
s
AC149475.2_FG004 - chr9:148174039-148176942 OK 0 0 0 0 1 no
AC149475.2_FG005 - chr9:148203125-148213830 OK 5.58129 5.46706 -0.0206777 -0.13312 0.894099
no
AC149475.2_FG007 - chr9:148234408-148235158 OK 32.6681 1.28519 -3.23549 -5.48684 4.0919e-08
yes
AC149475.2_FG008 - chr9:148237619-148240046 OK 0 0 0 0 1 no
AC149633.4_FG001 - chr9:150971905-150982549 OK 2.3421 1.77743 -0.275876 -0.84018 0.400808
no
AC149633.4_FG002 - chr9:150955778-150961584 OK 0.316124 0.840214 0.977522 1.57789 0.
114591 no
AC149633.4_FG003 - chr9:150945485-150947555 OK 0.0768126 0 5.3613e-315 2.22507e-308 0
yes
AC149633.4_FG005 - chr9:150899533-150901308 OK 41.8233 64.3736 0.43125 2.94018 0.00328023 yes
AC149810.2_FG003 - chr9:147237423-147245277 OK 7.42772 3.93657 -0.63491 -3.49688 0.00047072
s
AC149810.2_FG004 - chr9:147249015-147249715 OK 20.366 14.9801 -0.307146 -1.32082 0.18656 no
"./orig_gtf/0_1_gene_exp.diff" 53756L, 4519787C 23,1 0%
```

Advantages of RNA-Seq

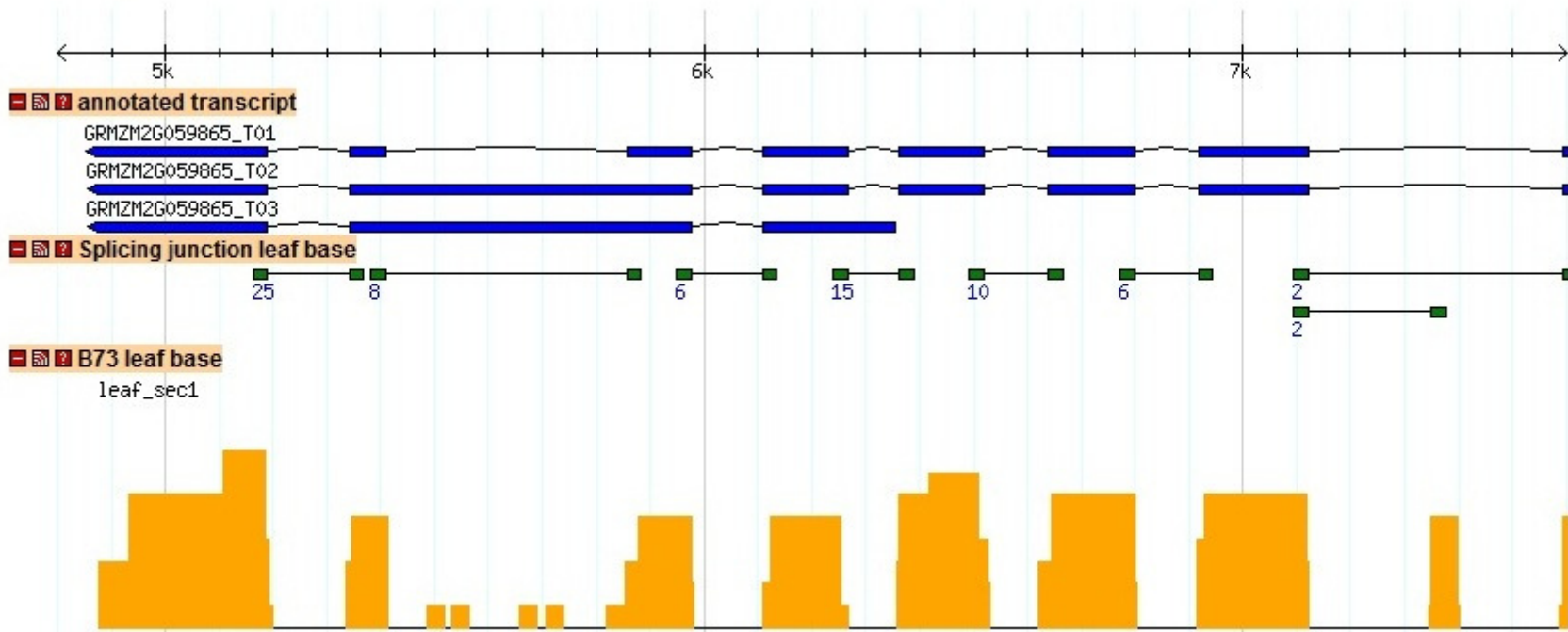
- Does not require existing genomic sequence
 - Unlike hybridization approaches
- Very low background noise
 - Reads can be unambiguously mapped
- Resolution
 - Up to 1 bp
- High-throughput
 - Better than Sanger sequencing of cDNA or EST libraries
- Cost
 - Lower than traditional sequencing
- Can reveal sequence variations (SNPs)

Issues



Issues

- Depth of coverage depends on “sequenceability” of the genomic region



Conclusion

RNA-Seq

- Offers high-throughput quantitative measurement of transcript abundance
- Expression levels correlate well with qPCR
- Costs continue to fall due to multiplexing
- Expected to replace microarrays for transcriptomic studies
- Automated pipeline (Tophat/Cufflinks)

References

- Wang Z, Gerstein M, Snyder M. *RNA-Seq: a revolutionary tool for transcriptomics*. Nat Rev Genet. 2009 Jan;10(1):57-63.
- Nagalakshmi U, Waern K, Snyder M. *RNA-Seq: A Method for Comprehensive Transcriptome Analysis*. Curr Protoc Mol Biol. 2010 Jan;Chapter 4:Unit 4.11.1-13