Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

RNA-seq

- "Whole Transcriptome Shotgun Sequencing" ("WTSS" or just "WTS") is a technology that uses the capabilities of Next-Generation Sequencing (NGS) to reveal a snapshot of RNA abundance from a genome and tissue-type at a given moment in time

- Provides the ability to look at:

  - Alternative gene-spliced transcripts

  - Post-transcriptional changes

  - Gene fusions

  - mutations/SNPs

  - exon/intron boundaries

  - Changes in:
    - total RNA
    - mRNAs (gene expression)
    - small RNAs (including miRNAs)
    - tRNAs
    - ribosomal RNAs

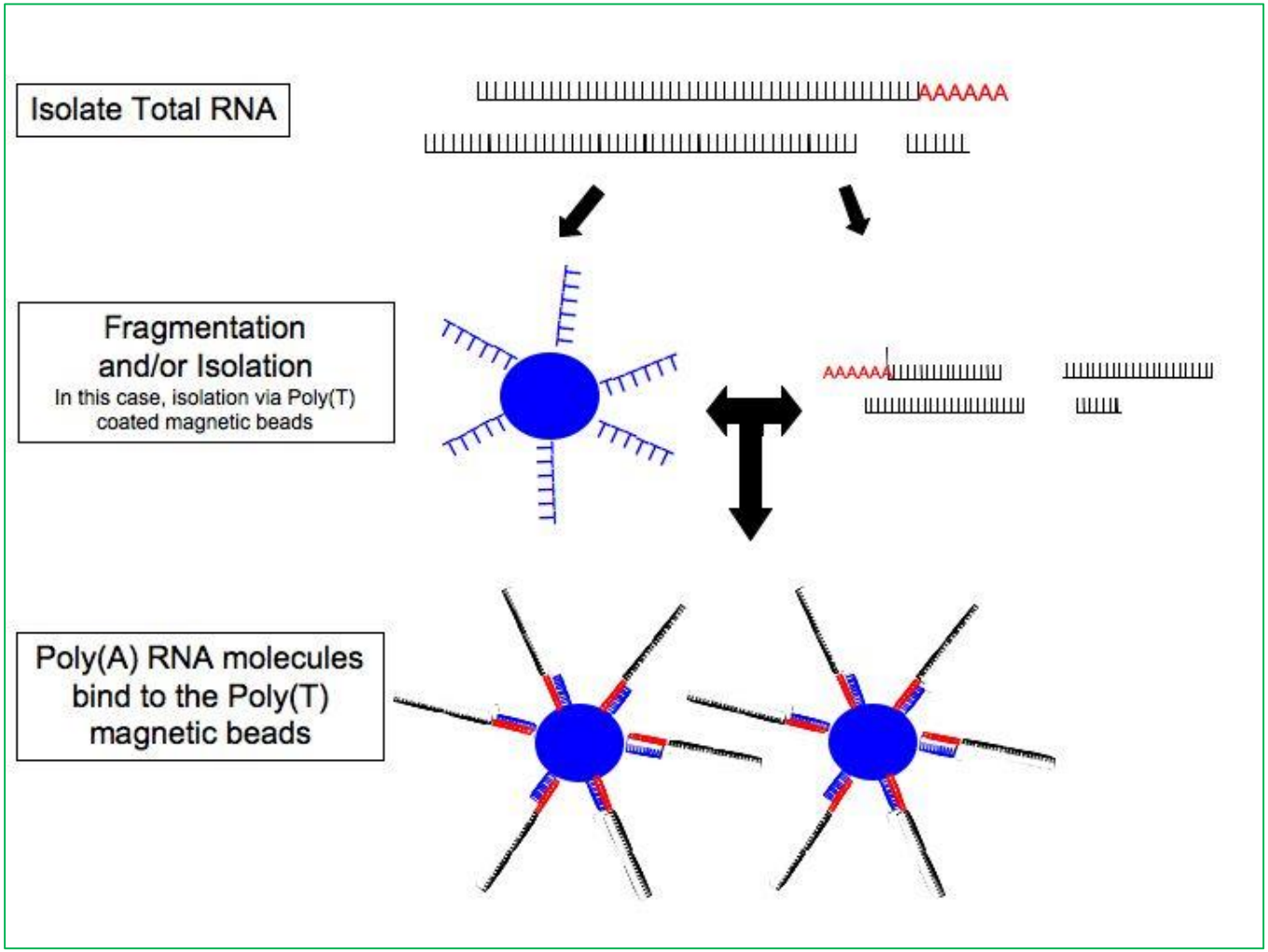
Advantages of RNA-seq

- The main deficiency of microarrays that makes RNA-Seq more attractive has been limited coverage:
  - Arrays target the identification of known common alleles that represent only approximately 500,000 to 2,000,000 SNPs of the more than 10,000,000 in the human genome
  - Microarrays aren't usually available to detect and evaluate rare allele variant transcripts
  - Microarrays are only as good as the SNP databases they're designed from
  - Many cancers are caused by rare <1% mutations and go undetected with microarrays
- The second main deficiency with microarrays is additional noise due to cross-hybridization
  - RNA-seq is more "digital" and has better signal-to-noise ratios
- With RNA-seq, one can sequence to any desired depth in order to get the necessary coverage and no more

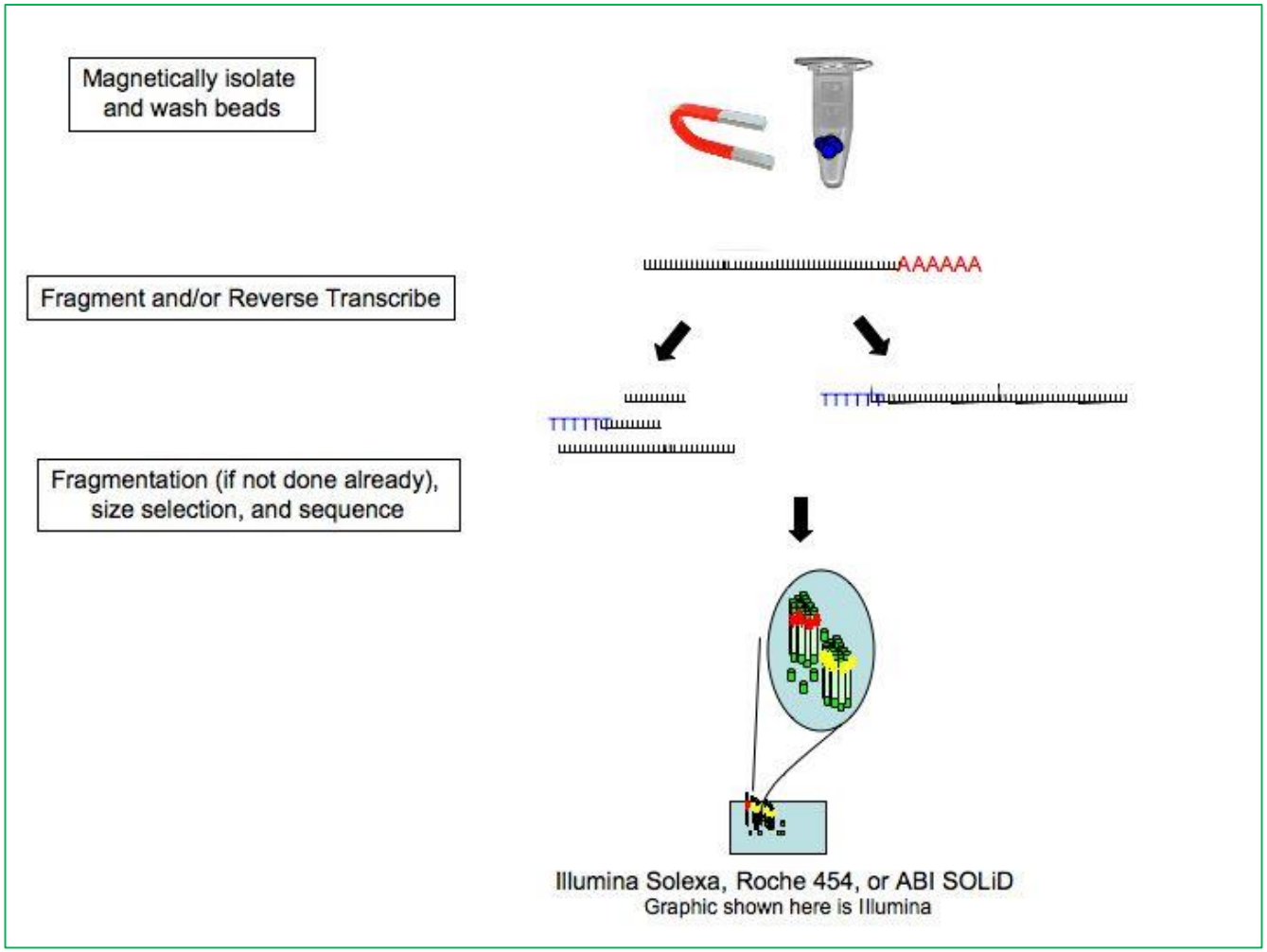
Selecting for mRNA, poly-A RNA, Ribosomal

- Frequently, in mRNA analysis the 3' polyadenylated (poly-A) tail is targeted in order to ensure that coding RNA is separated from noncoding RNA.
  - Accomplished simply with poly-T oligos covalently attached to a given substrate. Presently many studies utilize magnetic beads for this step.
  - The flow-through RNA (non-poly-A RNA) contains noncoding RNA
- Probe hybridization with microarrays can separate out Ribosomal RNA
  - Ribosomal RNA represents over 90% of the RNA within a given cell
  - Removing ribosomal RNA before sequencing greatly increases the percentage of the reads that are from the remaining portion of the transcriptome (saves $).
- When sequencing RNA other than mRNA, such as miRNA or other small RNAs, selection is based on the desired size range
  - size exclusion gel
  - size selection magnetic beads

Selecting for mRNA via poly(A) RNA



Selecting for mRNA via poly(A) RNA II



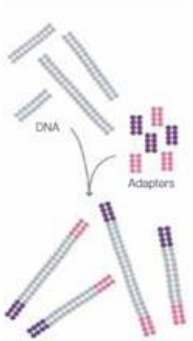


Figure 1
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

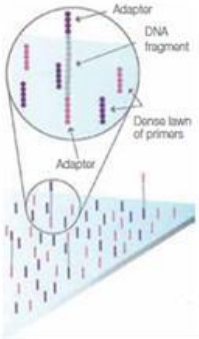


Figure 2
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

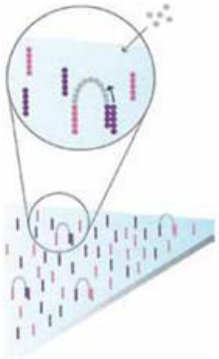


Figure 3
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

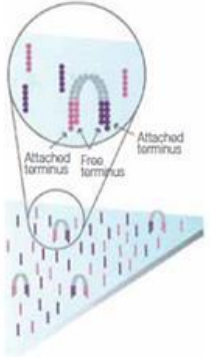


Figure 4
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

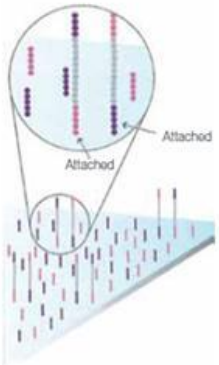


Figure 5
Denaturation leaves single-stranded templates anchored to the substrate.

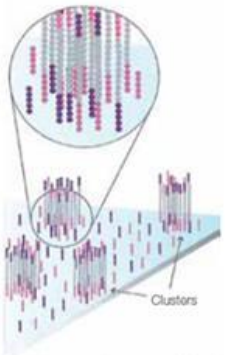


Figure 6
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

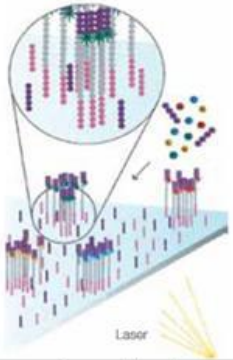


Figure 7
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

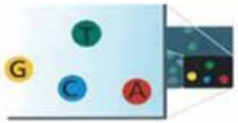


Figure 8
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

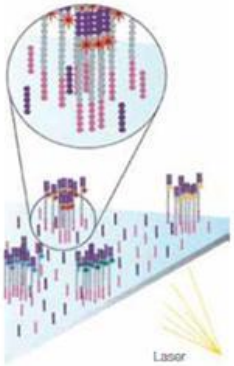


Figure 9
The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

Figure 10
After laser excitation, the image is captured as before, and the identity of the second base is recorded.

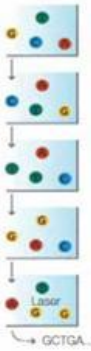


Figure 11
The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

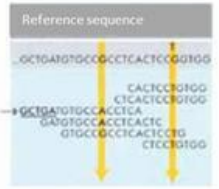


Figure 12
The data are aligned and compared to reference, and sequencing differences are identified.

RNA-seq Preparation and Paired-end Sequencing

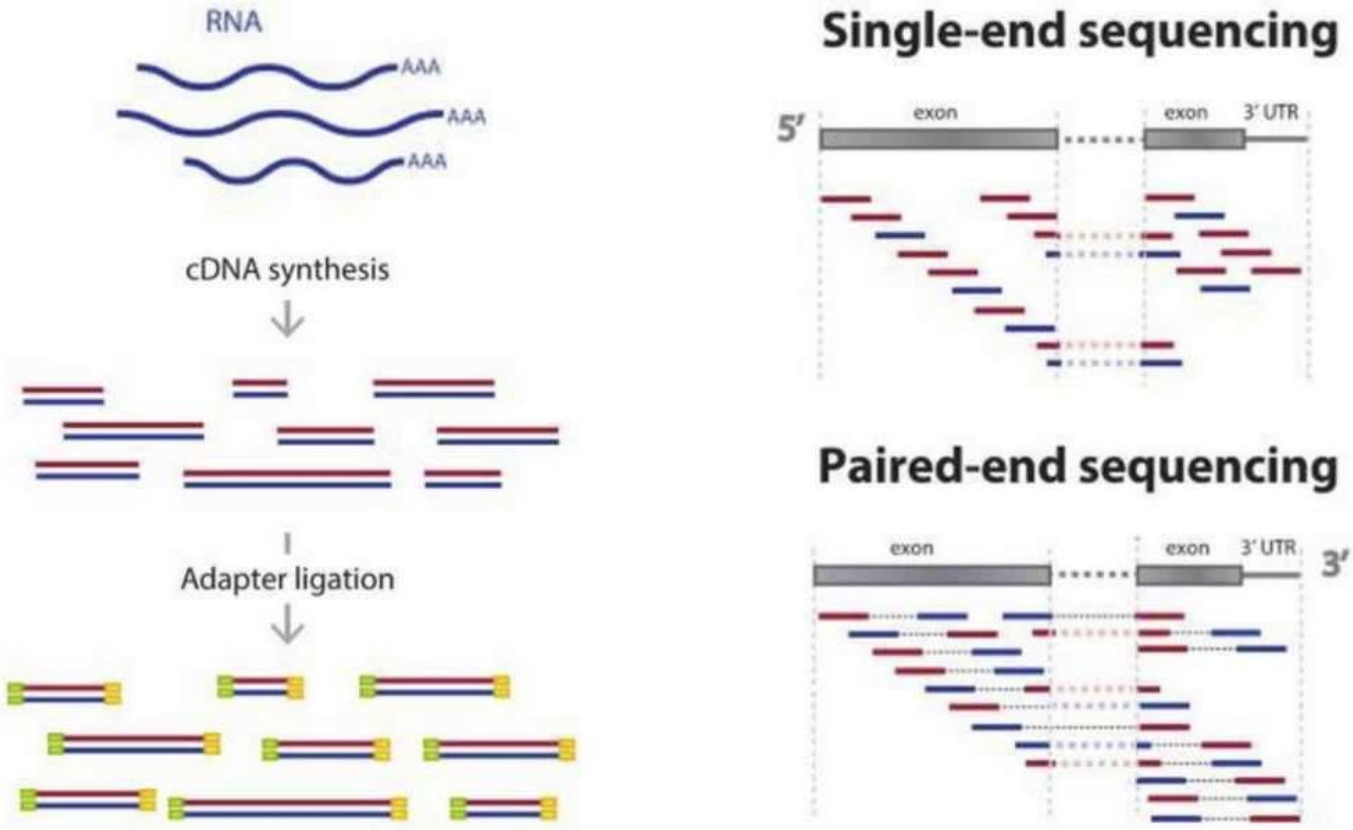
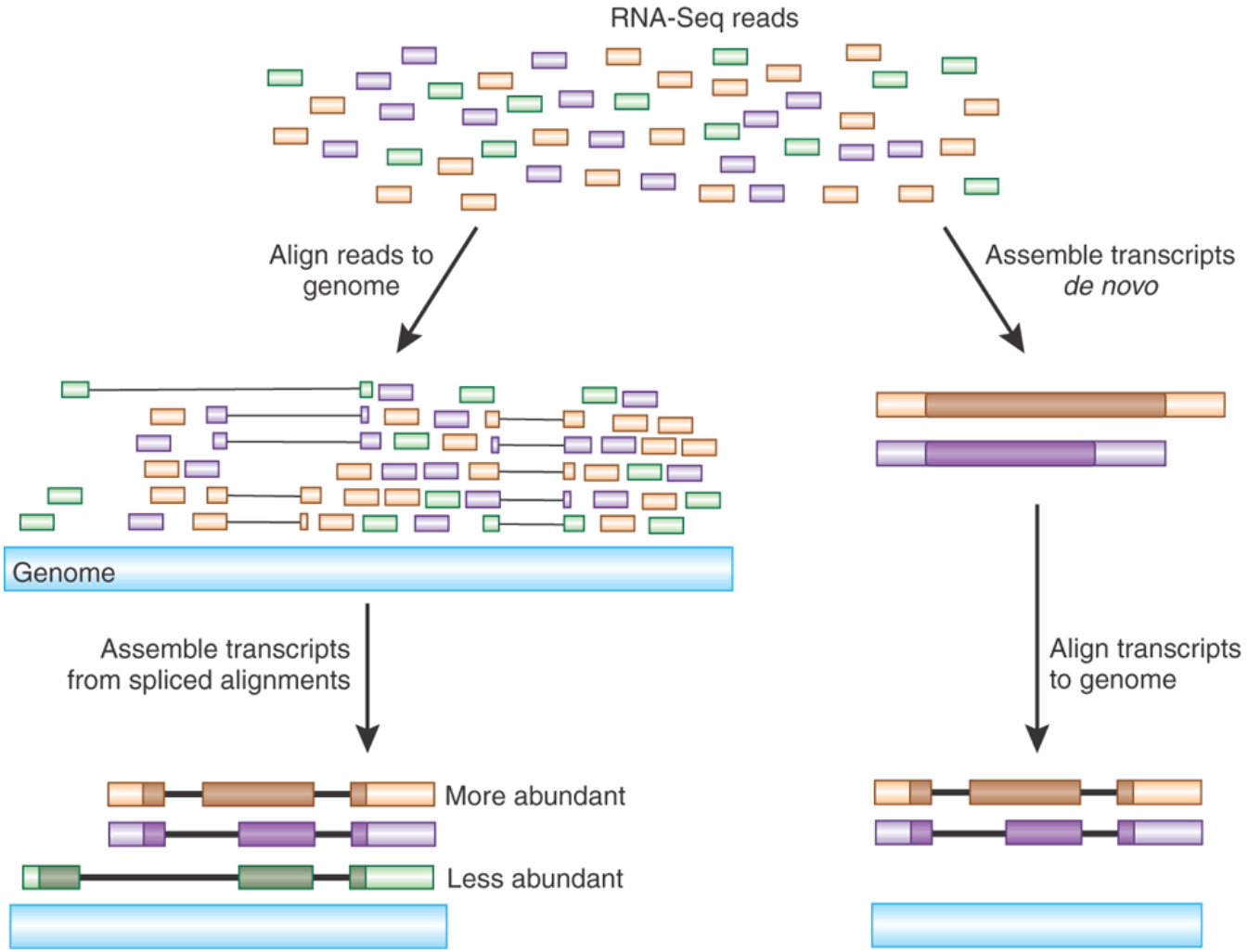


Image adapted from Zhernakova, et al., PLoS Genet. 2013 June; 9(6): e1003594.

RNA-seq "Align & Assemble" Vs. "De Novo"

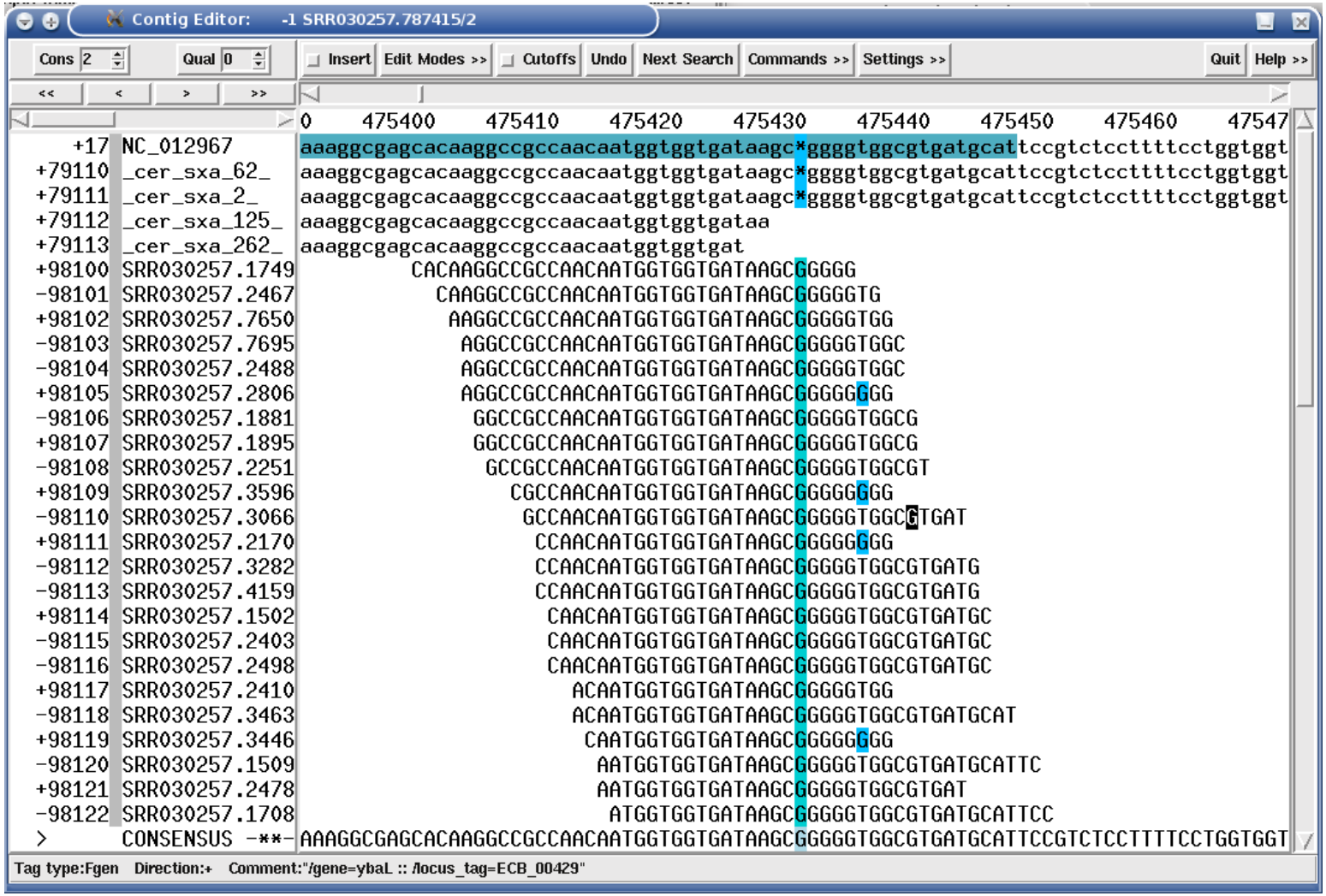


Transcriptome Assembly

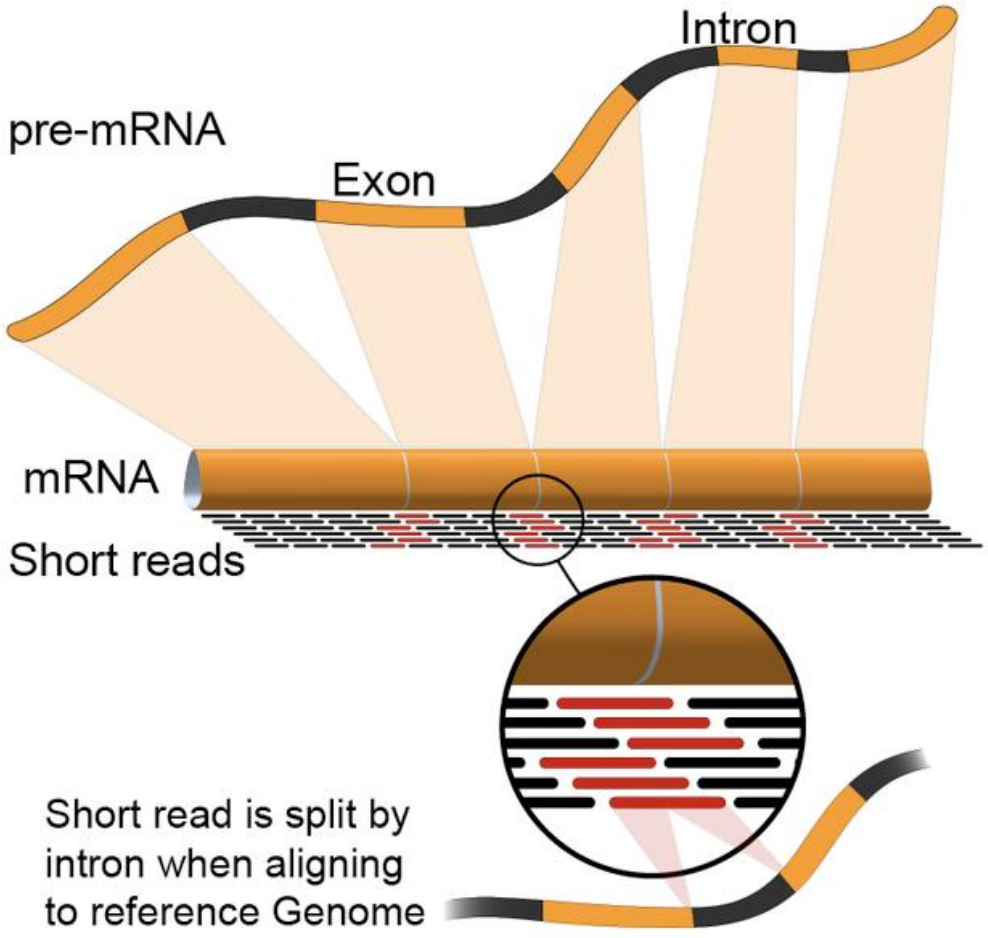
Two different assembly methods are used for producing a transcriptome from raw sequence reads

1. De-novo
   - Does not rely on the presence of a reference genome in order to reconstruct the nucleotide sequence.
   - Requires deep coverage and increased computing power to track all the possible alignments
2. Genome-guided
   - Easier and computationally cheaper approach is aligning the millions of reads to a "reference genome".
   - Several software packages exist for short read alignment, and recently specialized algorithms for transcriptome alignment have been developed
     - Bowtie for RNA-seq short read alignment
     - TopHat for aligning reads to a reference genome to discover splice sites
     - Cufflinks to assemble the transcripts and compare/merge them with others

SNP Detection Using RNA-seq

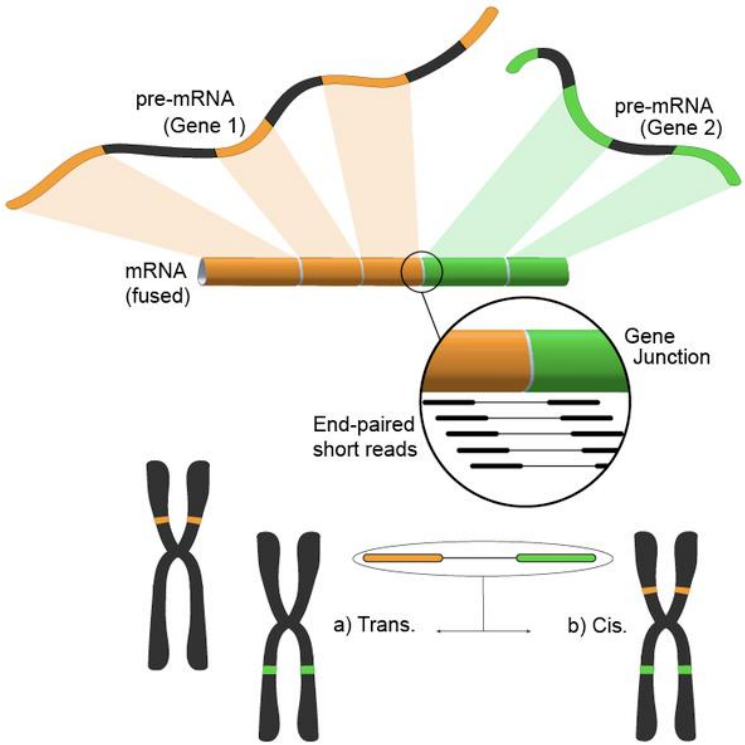


Exon-Intron Boundaries and Isoform Abundance via RNA-seq



Gene Fusion Detection

- Caused by different structural modifications and abnormalities in the genome:
  - Deletions
  - Duplications
  - Translocations
- Fusion genes have gained attention because of their relationship with cancer
- The ability of RNA-seq to analyze a sample's whole transcriptome in an unbiased fashion makes it an attractive tool to find these kinds of common events in cancer



RNA-seq Analysis Workflow

1. Read mapping
2. Counting reads overlapping with genes
3. Analysis of Differentially Expressed Genes (DEGs)
4. Clustering of co-expressed genes
5. Gene set/GO term enrichment analysis

Fastq Raw Reads File

- A FASTQ file normally uses 4 lines per sequence.
  - Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
  - Line 2 is the raw sequence calls.
  - Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
  - Line 4 encodes the phred quality scores for the sequence in Line 2, and must contain the same number of symbols as nucleotide calls in the sequence.

- Example:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- The character '!' represents the lowest quality while '~' is the highest.

- Here are the phred quality value characters in left-to-right increasing order of quality (ASCII):

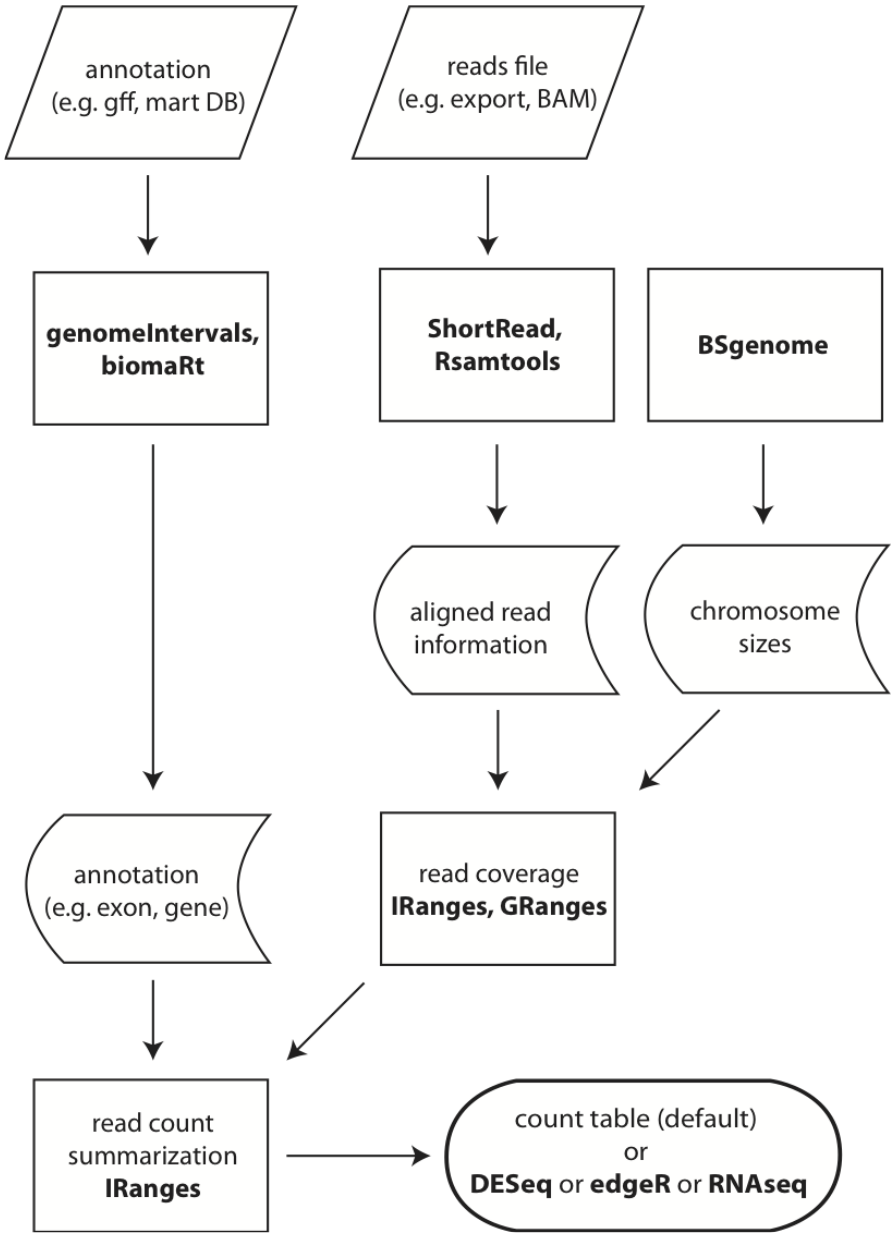
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

SAMTools

- Set of utilities for interacting with and post-processing of short, DNA-sequence read alignments in the SAM/BAM format
- SAM/BAM files are generated as output by short read aligners like BWA, Bowtie, etc.
- Supports complex tasks like variant calling and alignment viewing as well as sorting, indexing, data extraction and format conversion.
- SAM files can be very large (10s of Gigabytes is common), so compression into BAM is used to save space.
- SAM files are human-readable text files, while BAM files are simply the binary equivalent.
- BAM files more efficient for software to work with because the files are much smaller
- SAMtools makes it possible to work directly with a compressed BAM file, without having to uncompress to a SAM file.
- SAM/BAM files are complex - containing reads, references, alignments, quality information, and user-specified annotations

easyRNASeq

- package to ease the processing of RNA-seq data in R/Bioconductor.
- The main function of the easyRNASeq package is easyRNASeq:
  - should be the only processing method you need to know about when using the package.
  - It is essentially a wrapper around other functions performing the different tasks
  - The lower-level functions which are all exported too, if you feel you need to have a look at them



Package 'edgeR'

- "Empirical analysis of Digital Gene Expression data in R"
- A package for the analysis of digital gene expression data arising from RNA sequencing technologies such as SAGE (single-end), CAGE (5'-Cap), Tag-seq (single-end)or RNA-seq (paired-end), with emphasis on testing for differential expression.
- Particular strengths of the package include the ability to estimate biological variation between replicate libraries, and to conduct exact tests of significance which are suitable for small counts.
- The package is able to make maximal use of replicates
- Differential expression analysis of RNA-seq and digital gene expression profiles with biological replication.
- Uses empirical Bayes estimation and exact tests based on the negative binomial distribution.
- Also useful for differential signal analysis with other types of genome-scale count data.

Package 'DESeq'

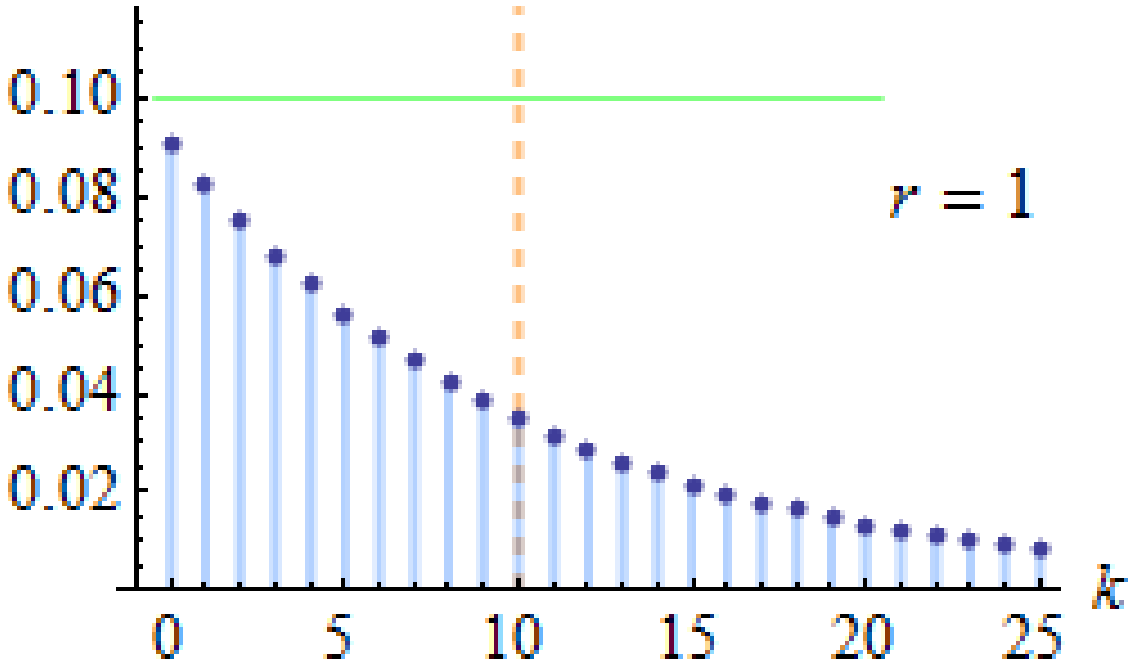
- "Differential gene Expression analysis of Sequencing data based on the negative binomial distribution"

- Estimates variance-mean dependence in count data from high-throughput sequencing assays and tests for differential expression based on a model using the negative binomial distribution

- Negative binomial distribution is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of r failures

    - 6-sided die, define a "1" as failure, and all "non-1"s as successes
    - Roll a die repeatedly until the 3rd time "1" appears (r = 3 failures)
    - The probability distribution of the number of "non-1"s that had appeared will be negative binomially distributed

Negative Binomial Probability Mass Function



- The orange line represents the mean, which is equal to 10 in each of these plots
- The green line shows the standard deviation.

Wig and bigWig

- Wig (Wiggle)
  - Older format for display of dense, continuous data
    - GC percent, probability scores, and transcriptome data
  - Data elements must be equally sized
  - wiggle data is compressed and stored internally in 128 unique bins
  - This compression causes a minor loss of precision when data is exported from a wiggle track
  - If your data is sparse or contains elements of varying size, use the bedGraph or bigBed format instead of the wiggle format

- bigWig
  - Recommended format for almost all graphing track needs data elements must be equally sized
  - Indexed binary format
  - Main advantage of the bigWig files is that only the portions of the files needed to display a particular region are transferred and loaded into browser
  - For large data sets bigWig is considerably faster than regular wiggle files
  - Only the portion that is needed for the chromosomal position you are currently viewing is locally cached as a "sparse file"

Wiggle Format

- **Example 1**

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

- **Example 2 (same result)**

```
variableStep chrom=chr2 span=5
300701 12.5
```

- **Example 3**

```
fixedStep chrom=chr3 start=400601 step=100 span=5
11
22
33
```

  - **causes the values 11, 22, and 33 to be displayed as 5-base regions on chromosome 3 at positions 400601-400605, 400701-400705, and 400801-400805, respectively.**

BED Files

- Provides a flexible way to define the data lines that are displayed in an annotation track
- 3 required fields and 9 additional optional fields
- The number of fields per line must be consistent throughout any single set of data in an annotation track
- The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used

- The first 3 required BED fields are:
  1. chrom
  2. chromStart
  3. chromEnd

- The 9 additional optional BED fields are:
  4. name - Defines the name of the BED line
  5. score - A score between 0 and 1000
  6. strand - Defines the strand - either '+' or '-'
  7. thickStart - The starting position at which the feature is drawn thickly
  8. thickEnd - The ending position at which the feature is drawn thickly
  9. itemRgb - An RGB value of the form R,G,B (e.g. 255,0,0)
  10. blockCount - The number of blocks (exons) in the BED line
  11. blockSizes - A comma-separated list of the block sizes
  12. blockStarts - A comma-separated list of block starts

BedGraph and BigBed

- BedGraph
  - Allows display of continuous-valued data in track format
  - Useful for probability scores and transcriptome data
  - Similar to the wiggle (WIG) format, but unlike the wiggle format, data exported in the bedGraph format are preserved in their original state (no rounding)

- BigBed
  - Stores annotation items that can either be simple, or a linked collection of exons, much as BED files do
  - Indexed binary format
  - The main advantage of the bigBed files is that only the portions of the files needed to display a particular region are transferred and loaded into browser
  - For large data sets bigBed is considerably faster than regular BED files
  - The bigBed file remains on the web accessible server (http, https, or ftp)
    - not on the UCSC server
  - Only the portion that is needed for the chromosomal position you are currently viewing is locally cached as a "sparse file"

GFF (General Feature Format) Files

- Based on the Sanger GFF2 specification.
- 9 required fields that must be tab-separated.
  - If the fields are separated by spaces instead of tabs, the track will not display correctly

- GFF fields:
  1. seqname - The name of the sequence. Must be a chromosome or scaffold.
  2. source - The program that generated this feature.
  3. feature - The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
  4. start - The starting position of the feature in the sequence. The first base is numbered 1.
  5. end - The ending position of the feature (inclusive)
  6. score - A score between 0 and 1000. If the track line useScore attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (higher numbers = darker gray).
  7. strand - Valid entries include '+', '-', or '.' (for don't know/don't care).
  8. frame - If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
  9. group - All lines with the same group are linked together into a single item.

GTF (Gene Transfer Format) Files

- Refinement to GFF that tightens the specification.
- First 8 GTF fields are the same as GFF
- The 9th (Group) field has been expanded into a list of attributes
  - Each attribute consists of a type/value pair
  - Attributes must end in a semi-colon, and be separated from any following attribute by exactly one space
- The attribute list must begin with the two mandatory attributes:
  - gene_id - A globally unique identifier for the genomic source of the sequence
  - transcript_id - A globally unique identifier for the predicted transcript
  - Attribute list Example:
    - gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1
  - The UCSC Genome Browser groups together GTF lines that have the same transcript_id value
    - It only looks at features of type exon and CDS.