NGS applications in biomedical sciences

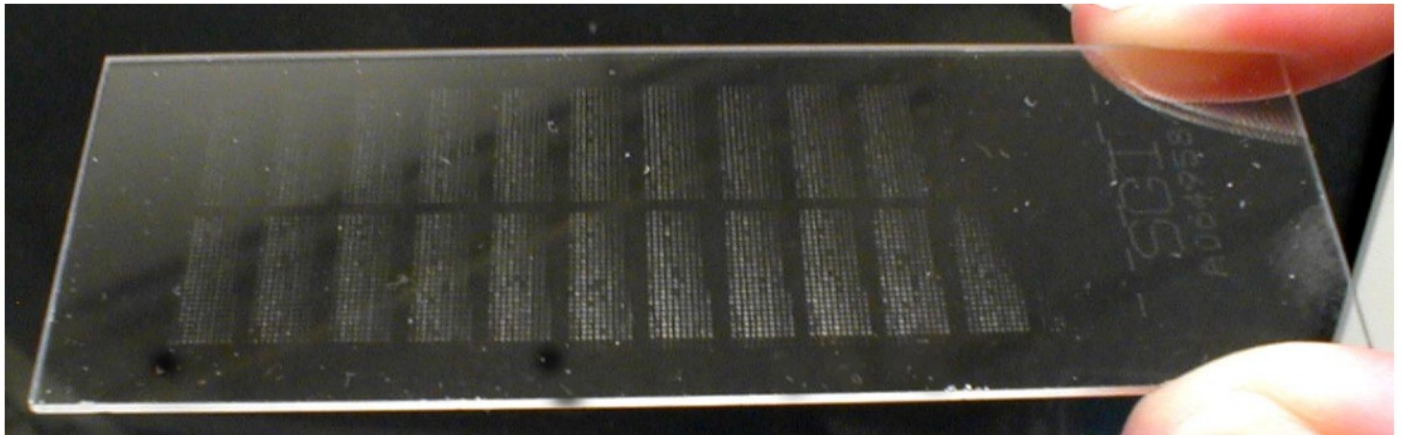
IMGGI,
November 2013

Comparison of Methods for Studying the Transcriptome

Technology	**Tiling microarray**	**cDNA or EST sequencing**	**RNA-Seq**
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

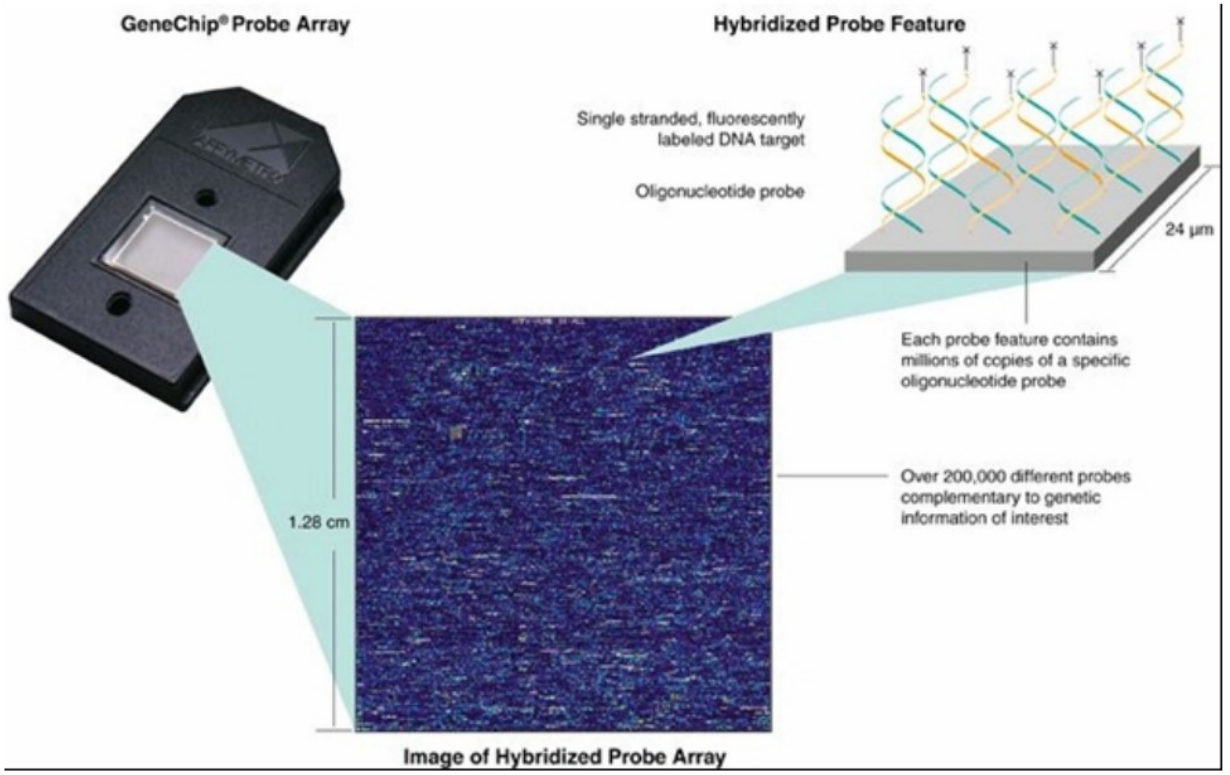
Microarrays

- a microarray is a solid support, on which pieces of DNA are arranged in a grid-like array
- measures RNA abundances by exploiting complementary hybridization



Affymetrix Microarray

GeneChip


GeneChip® Probe Array
Hybridized Probe Feature
Single stranded, fluorescently labeled DNA target
Oligonucleotide probe
24 µm
Each probe feature contains millions of copies of a specific oligonucleotide probe
Over 200,000 different probes complementary to genetic information of interest
1.28 cm
Image of Hybridized Probe Array

Types of microarrays

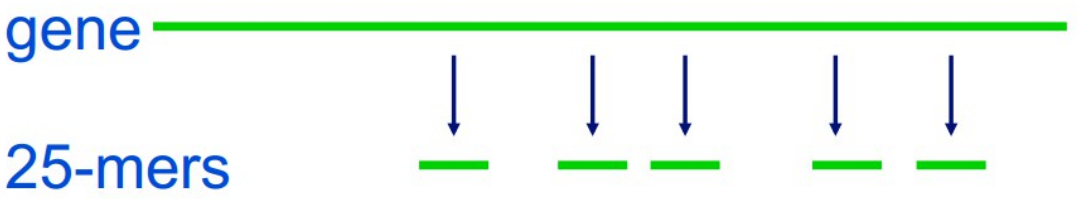
Spotted vs. oligonucleotide arrays

- *spotted arrays*: synthesize samples of cDNA (full-length transcripts or shorter sequences) and then spot them onto array

- *oligonucleotide arrays*: synthesize sets of DNA oligonucleotides (fixed length sequences, typically 25-60 nucleotides in length) on array
  - Affymetrix uses a photolithography process similar to that used to make semiconductor chips
  - Nimblegen (in Madison) uses an array of millions of tiny mirrors + photo deposition chemistry

Oligonucleotide array

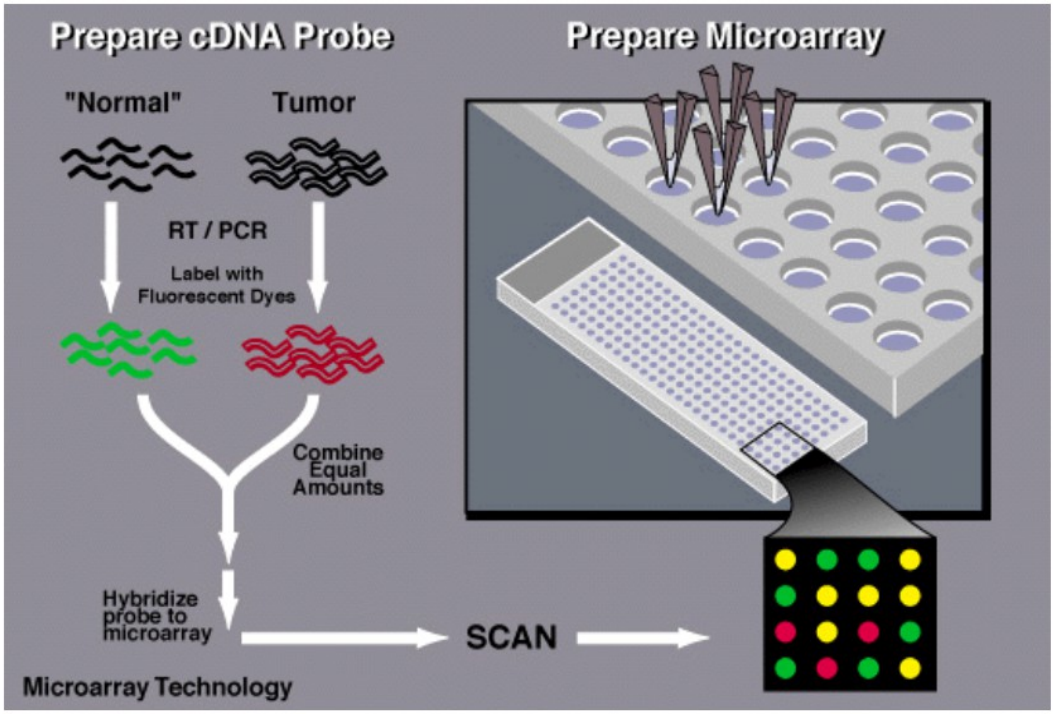
- given a gene to be measured, select different *n*-mers for the gene



- can also select *n*-mers for noncoding regions of the genome
- selection criteria
  - specificity
  - hybridization properties
  - ease of manufacturing

Microarray technology

- RNA is isolated from matched samples of interest, and is typically converted to cDNA. It is labeled with fluorescence and then hybridized to.



Microarray measurements

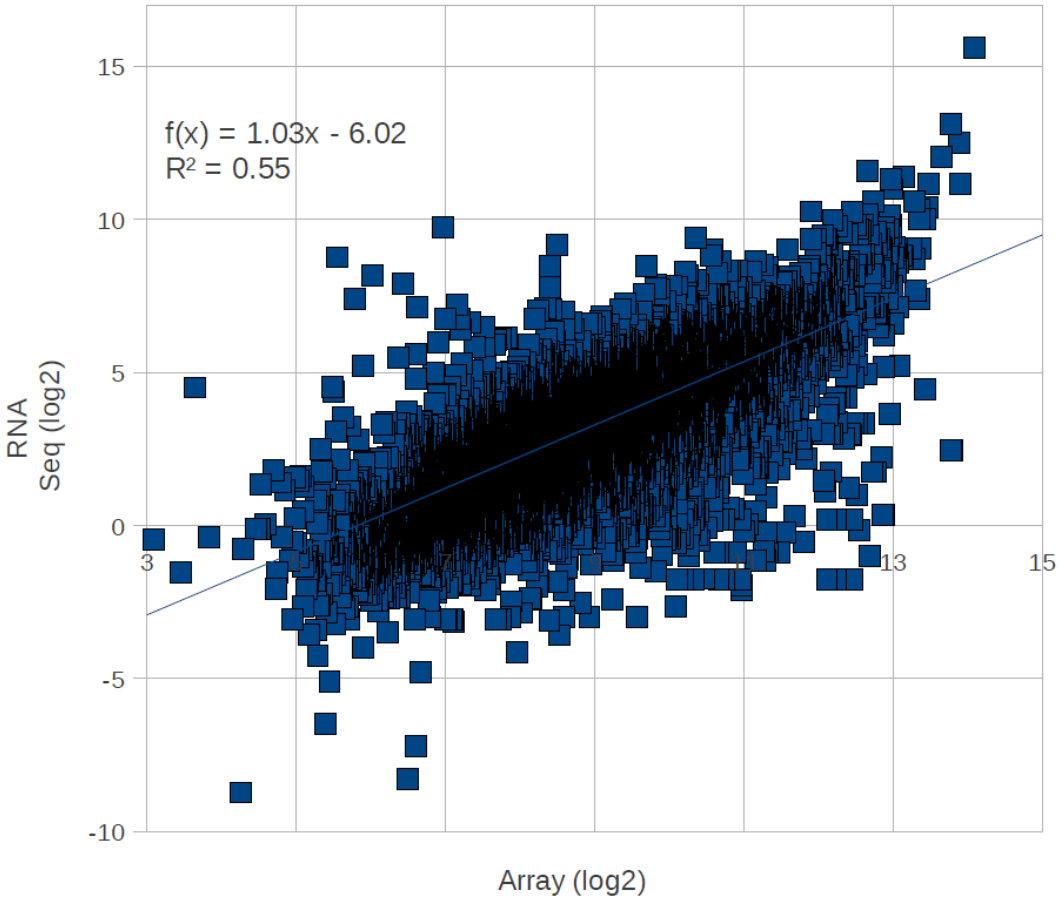
- we can't detect the absolute amount of mRNA present for a given gene, but we can measure a relative quantity
- for two color arrays, the measurements represent

$$G_i = \log \frac{\text{red}_i}{\text{green}_i}$$

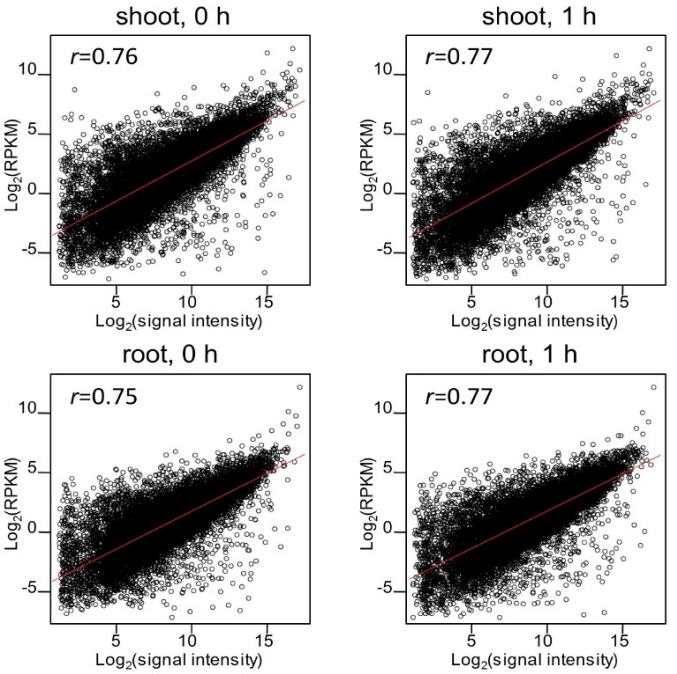
where red is the test expression level, and green is the reference level for gene G in the i th experiment

RNA-Seq vs microarray

The correlation of FPKM values for 10874 RefSeq genes found to be expressed in islets (values FPKM=0 excluded). r=0.74

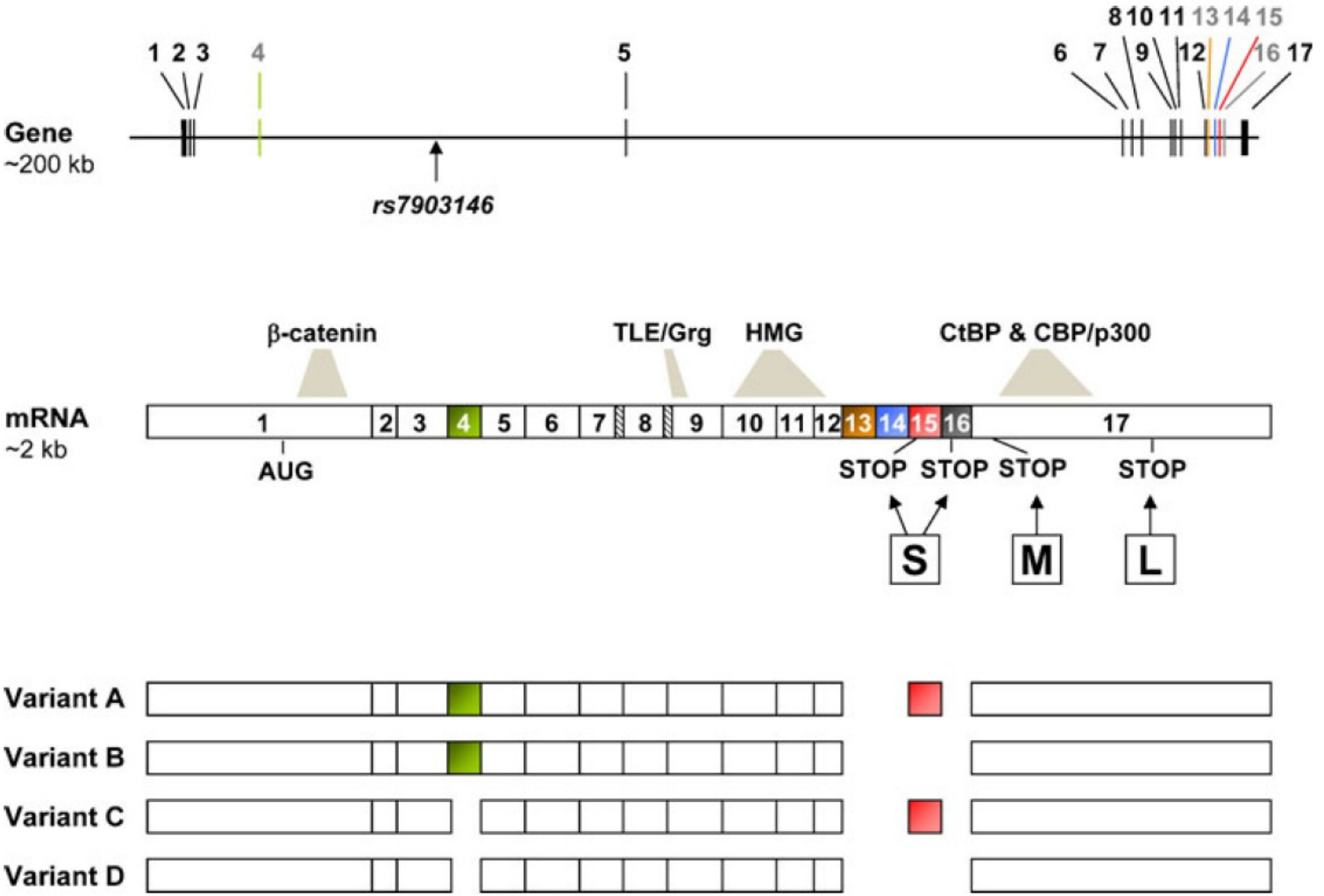


Literature:



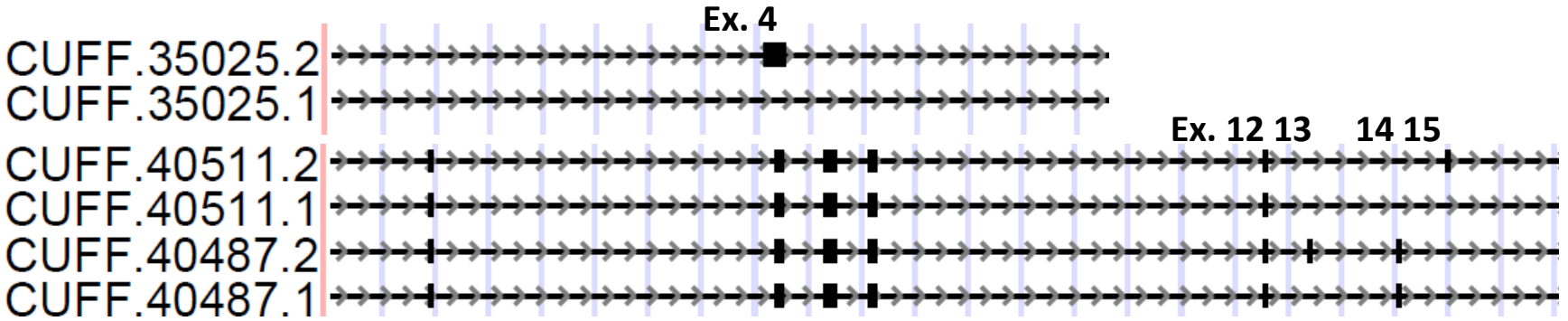
Mizuno et al, BMC Genomics 2010

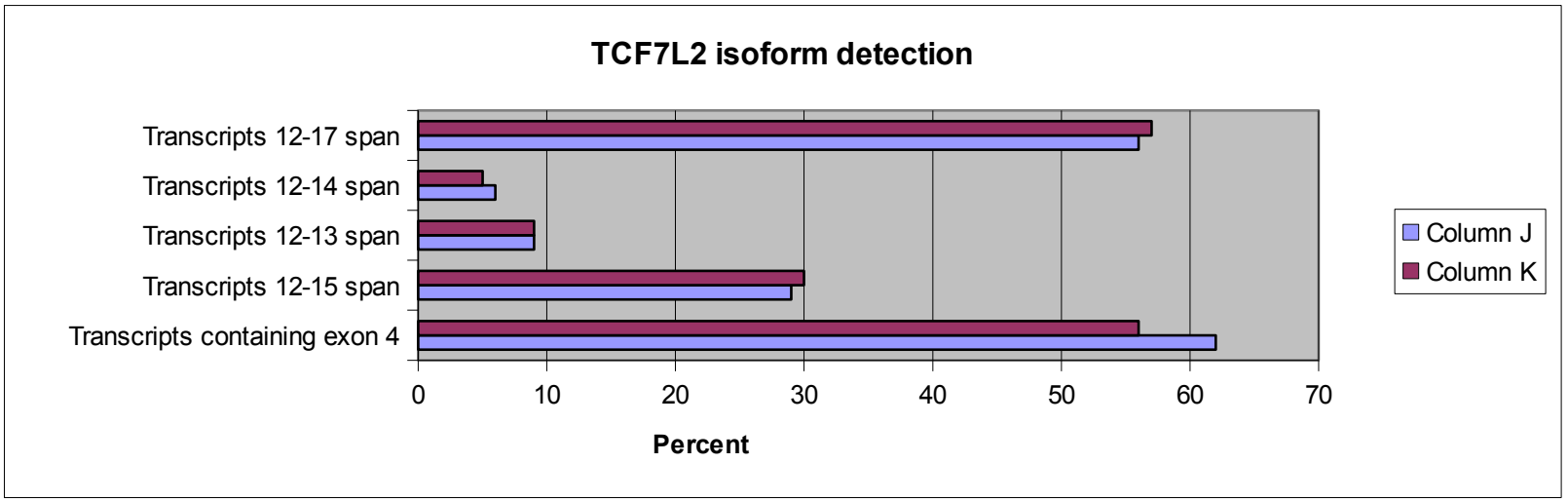
TCF7L2 splicing



Curr Diab Rep. 2010 Dec;10(6):444-51. Molecular function of TCF7L2: Consequences of TCF7L2 splicing for molecular function and risk for type 2 diabetes.

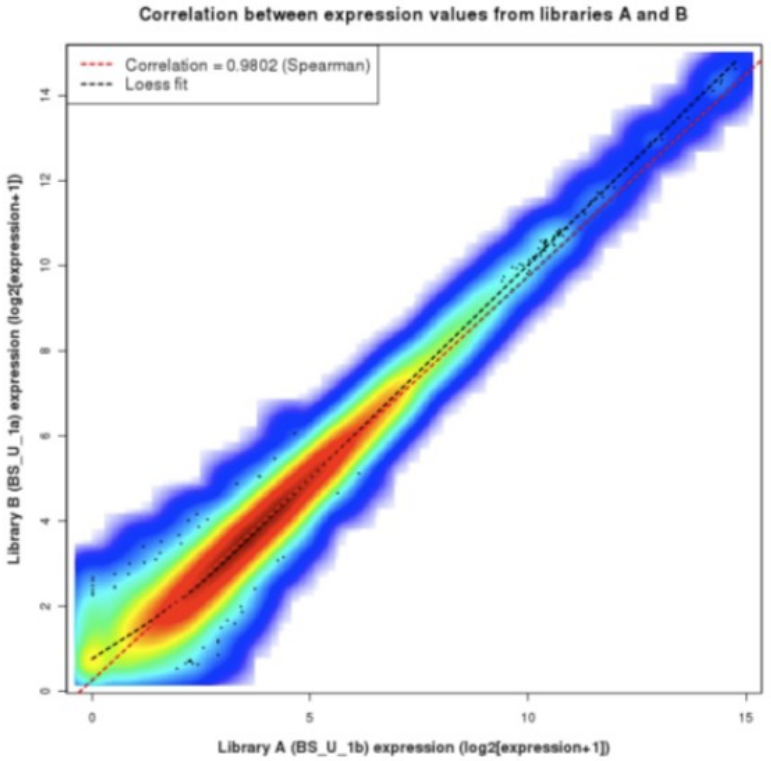
Quantification of annotated transcripts with RNA-Seq





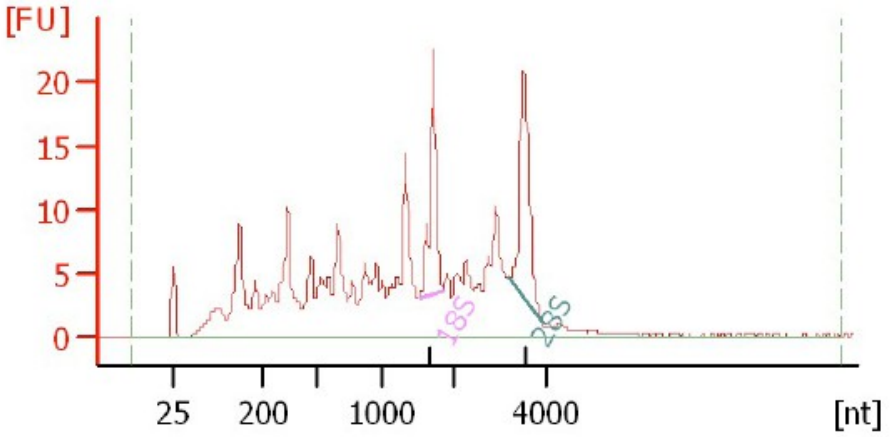
Replicates

- Technical Replicate
  - Multiple instances of sequence generation
    - Flow Cells, Lanes, Indexes
- Biological Replicate
  - Multiple isolations of cells showing the same phenotype, stage or other experimental condition
  - Some example concerns/ challenges:
    - Environmental Factors, Growth Conditions, Time
  - Correlation Coefficient 0.92-0.98

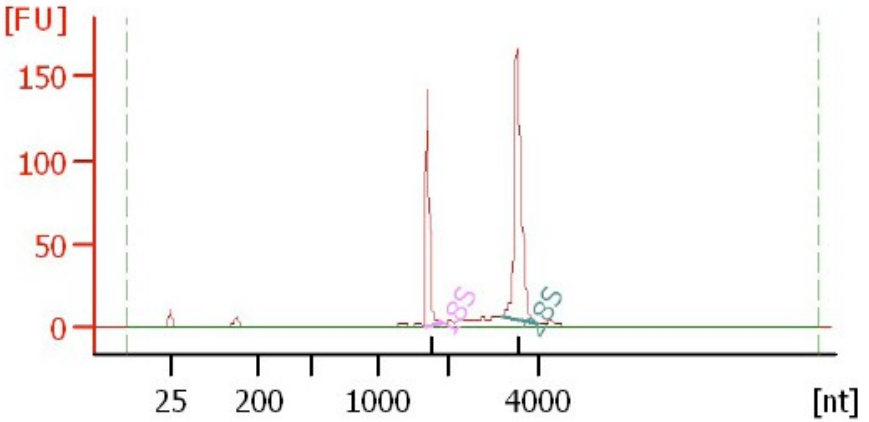


RNA quality – Agilent bioanalyzer

- 'RIN' = RNA integrity number
  - 0 (bad) to 10 (good)

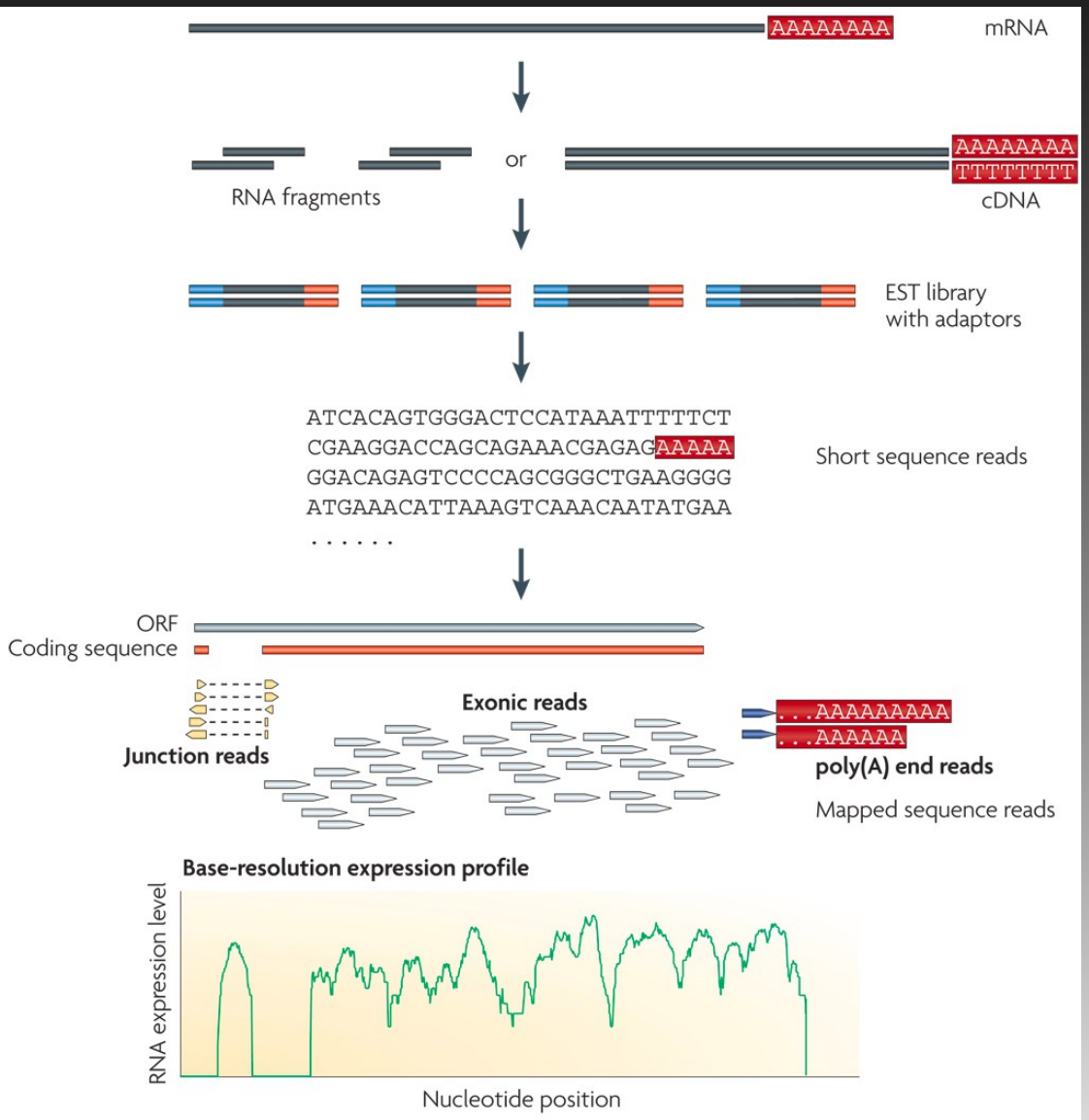


RIN = 6.0



RIN = 10

RNA Sequencing



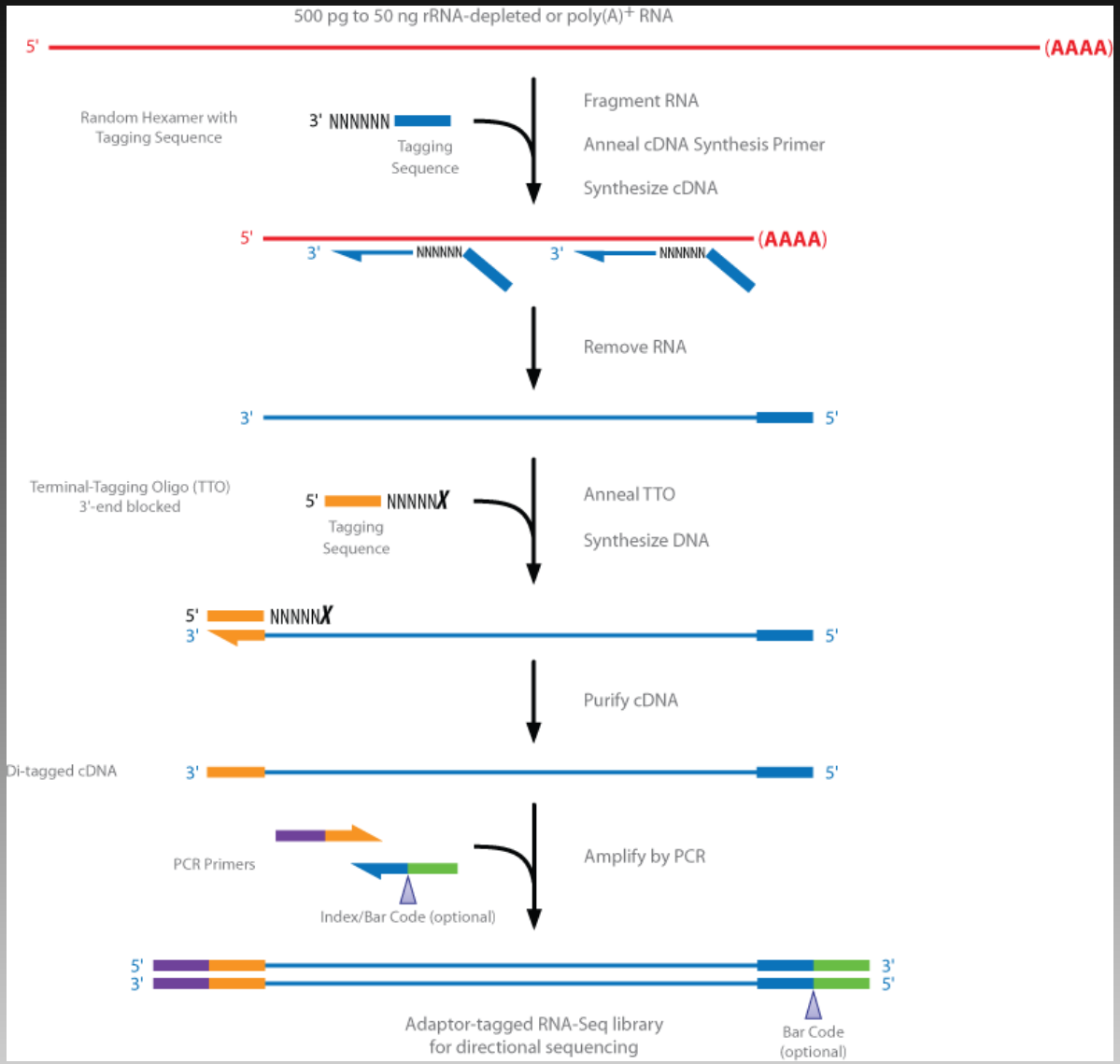
Population of RNA (poly A+) converted to a library of cDNA fragments with adaptors attached to one or both ends

Solid Phase Amplification performed

Molecules sequenced from one end (Single End) or both ends (Pair End)

Reads are typically 30-400bp depending on sequence technology used

TRUSEQ Library Preparation



Library Construction
Effective elimination of ribosomal RNA (negative selection) followed by polyA selection (for mRNA)

High Quality Strand Information

Can be used with low quality/low abundance RNA (10-100ng)

48 barcodes allows for multiplexing

Small RNAs can be directly sequenced

Large RNAs must be fragmented

Experimental Design: Single End (SR) vs Paired End (PE)

Single Read: one read sequenced from one end of each sample cDNA insert (Rd1 SP: Read 1 Sequencing Primer)

Paired End: two reads (one from each end) sequenced from each sample cDNA insert (Rd1 and Rd2 sequencing primer)

SR: often used for expression studies or SNP detection; NOT good for splice isoforms
PE: used for discovery of novel transcripts, splice isoforms and for de novo transcriptome assembly

Experimental Design: How many reads do I need

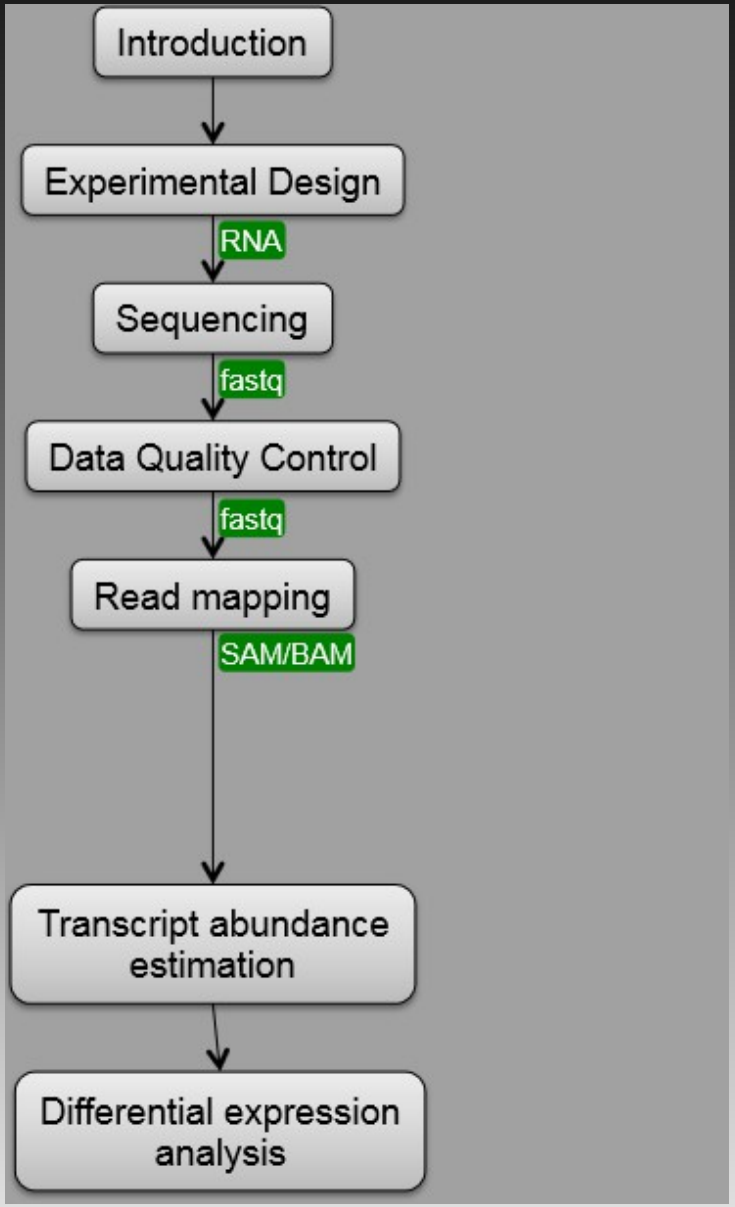
Greater Sequencing Depth correlates with better genomic coverage and more robust differential gene expression analysis

Study Type	Reads Needed
Expression Profiling	5-10 Million
Alternative splicing, quantifying cSNPs	50-100 M
De Novo Transcriptome Assembly	100-1000 M

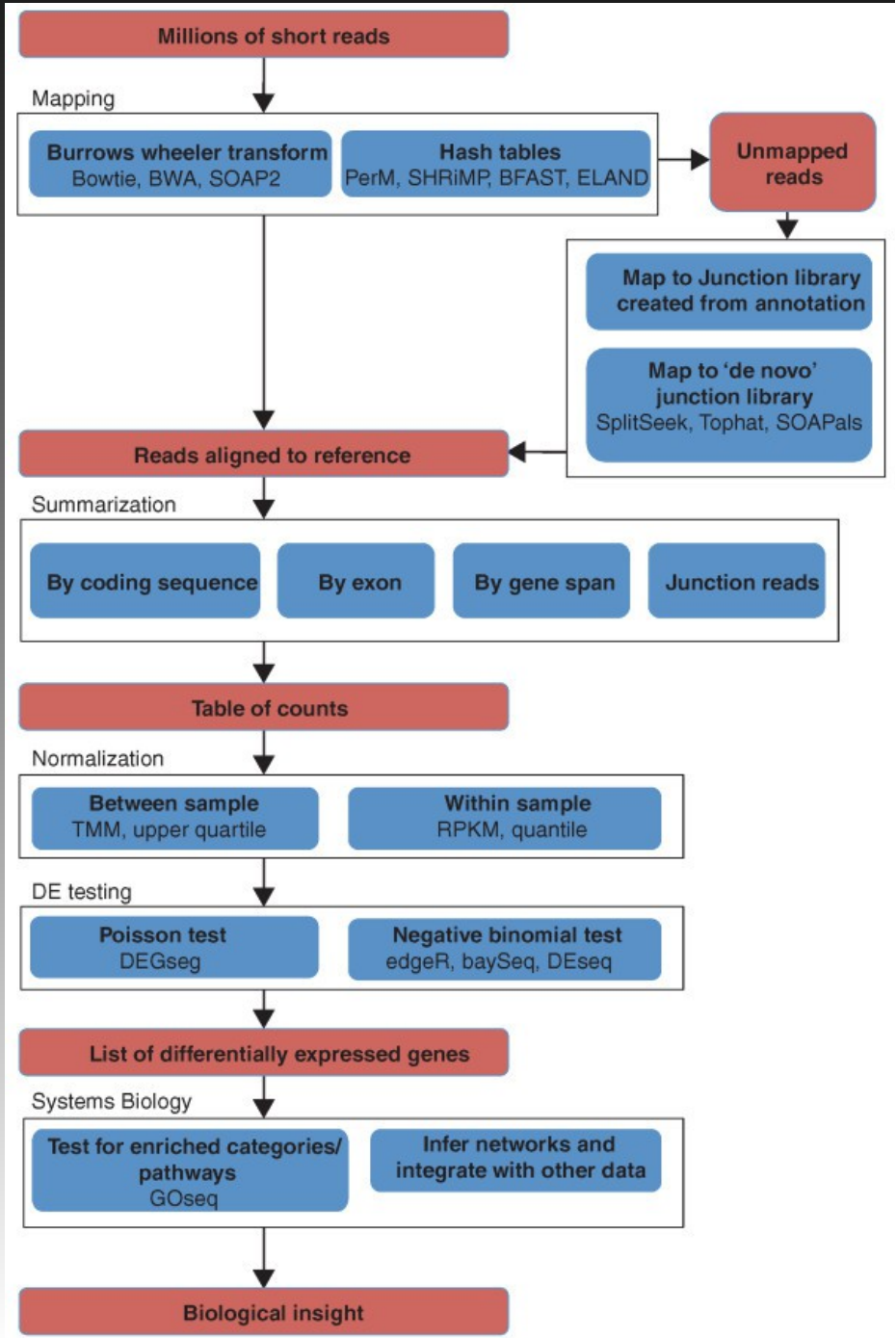
Sequencing Instrument	Reads per Lane (SR:PE)	Reads per Flow Cell
HiSEQ 2500	185:375M	1.5:3 Billion

Sequence Analysis

Theory

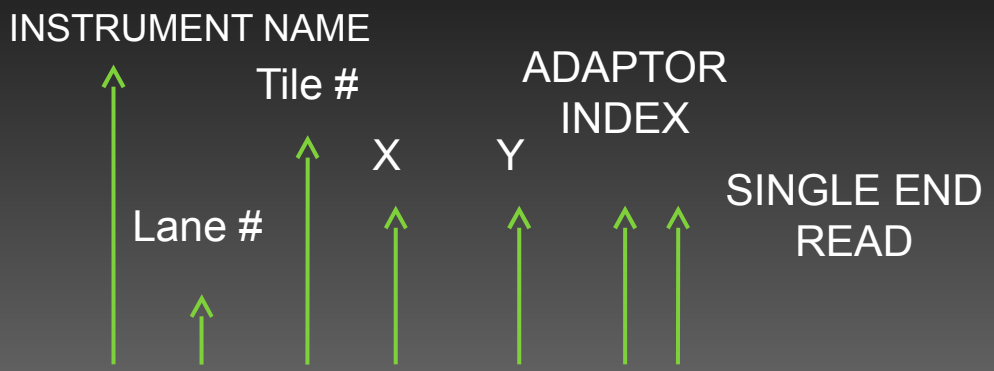


Practice



Converting RAW data to FASTQ

FASTQ File



@SN971:3:2304:20.80:100.00#0/1
NAAATTTCACATTGCGTTGGGAACAGTTGGCCCAAACTCAGGTTGCAGTAACTGTCACAATACCATTCTCCATCAACTTCA
AGAAATGTTCAACAAAACAC
+
@P\ccceeeggggiihhiiiiiiihighiiiiiiiiiiiiifghhhhgfghiifihihfhhiiiihiggggggeeeeeeddcdddccbcdddccccccc

Line 1: begins with '@' followed by sequence identifier
Line 2: raw sequence
Line 3: +
Line 4: base quality values for sequence in Line 2

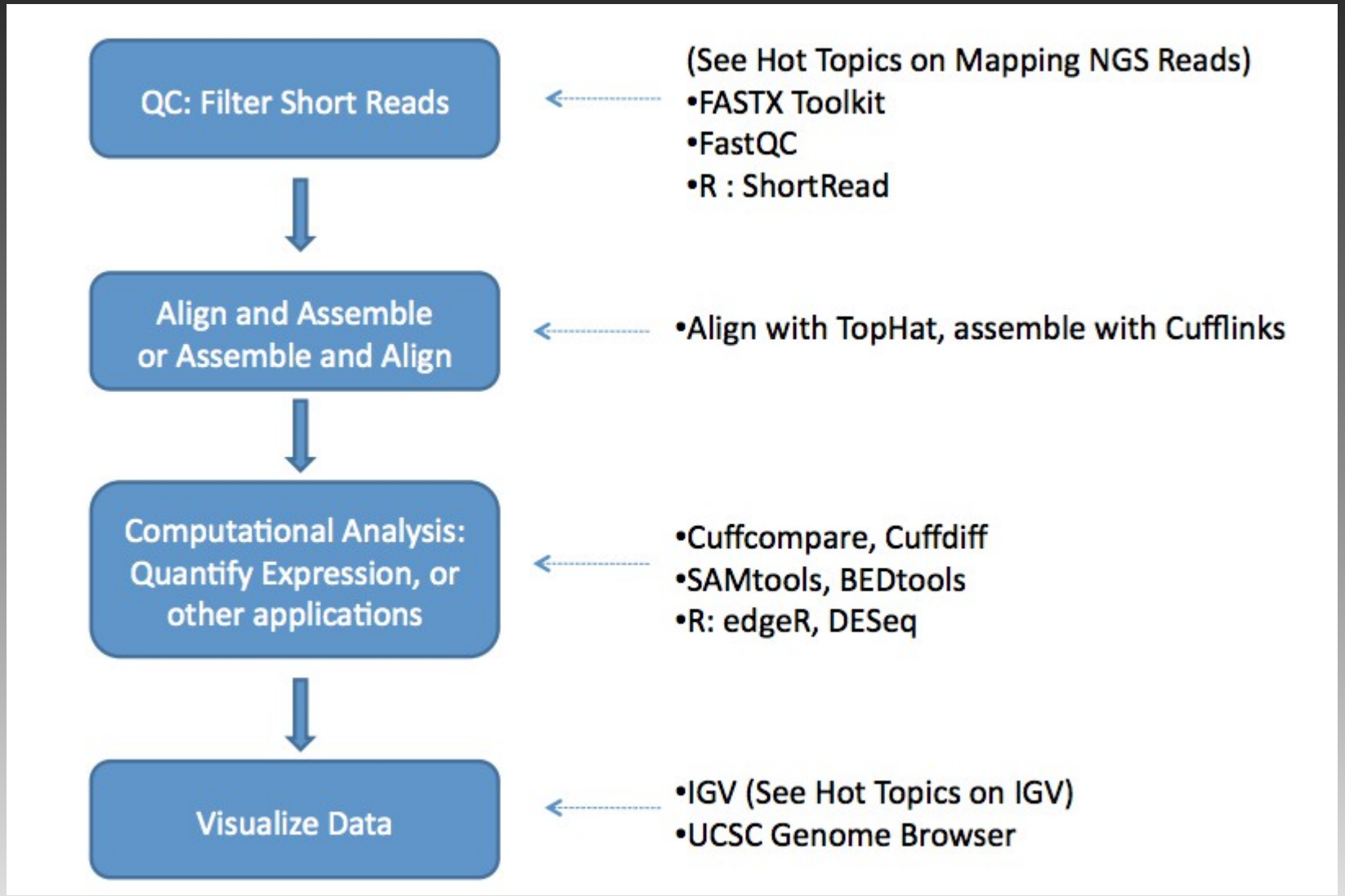
Tool recommendations

- Alignment
  - BWA (PMID: 20080505)
    - Align to genome + junction database
  - Tophat (PMID: 19289445), MapSplice (PMID: 20802226), hmmSplicer (PMID: 21079731)
    - Spliced alignment to genome
- Expression, differential expression alternative expression
  - Cufflinks/Cuffdiff (PMID: 20436464), ALEXA-seq (PMID: 20835245), RUM (PMID: 21775302)
- Fusion detection
  - ChimeraScan (PMID: 21840877), Defuse (PMID: 21625565), Comrad (PMID: 21478487)
- Transcript assembly
  - Trinity (PMID: 21572440), Oases (PMID: 22368243), Trans-ABySS (PMID: 20935650)
- Mutation calling
  - SNVMix (PMID: 20130035)
- Visit the 'SeqAnswers' or 'BioStar' forums for more recommendations and discussion
  - http://seqanswers.com/
  - http://www.biostars.org/

Online Community Forum and Discussion

- http://seqanswers.com/

My RNA Seq Workflow

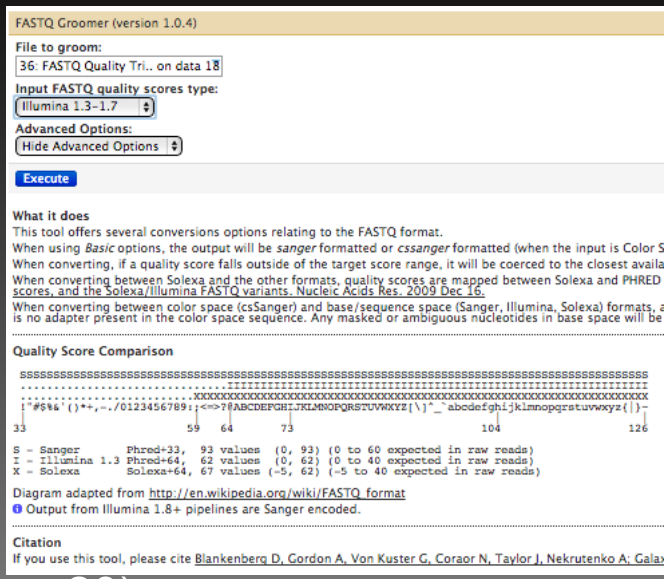


Quality Control

FASTQ Groomer: converts FASTQ data from different sources (ie Illumina, 454 Sequence etc) to a consensus FASTQ file

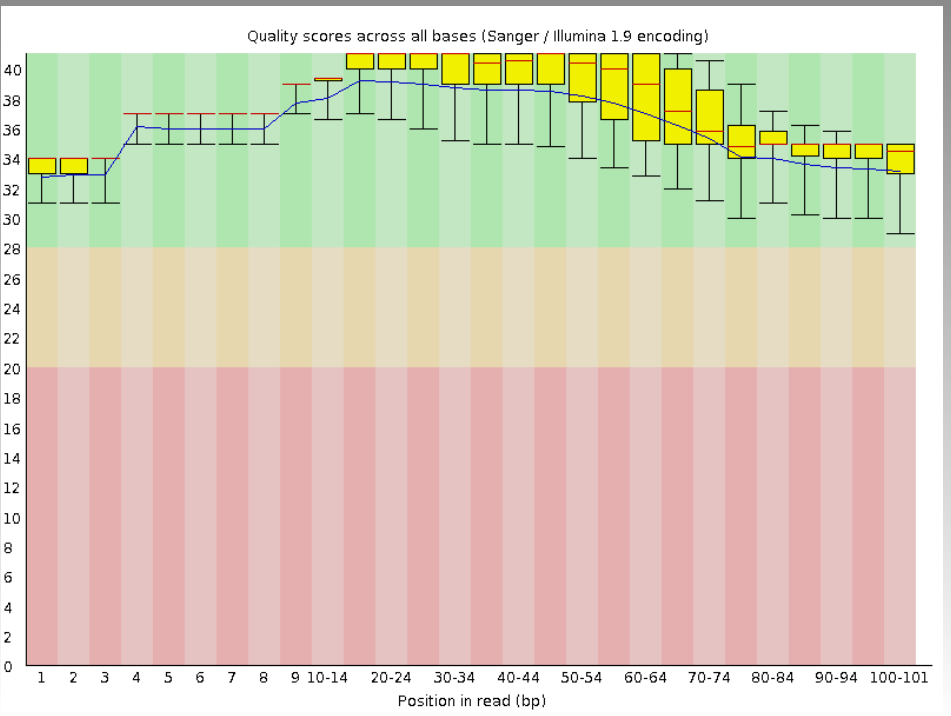
FASTQ QC: assesses base quality of sequence reads

- Per base sequence quality
- per sequence quality scores
- GC content
- Sequence Length
- Sequence Duplication
- Overrepresented sequences
- Kmer content



FASTQ TRIMMER: eliminate sequences below phRed score (usually <20)

Remember to check how many reads are lost from original input after processing

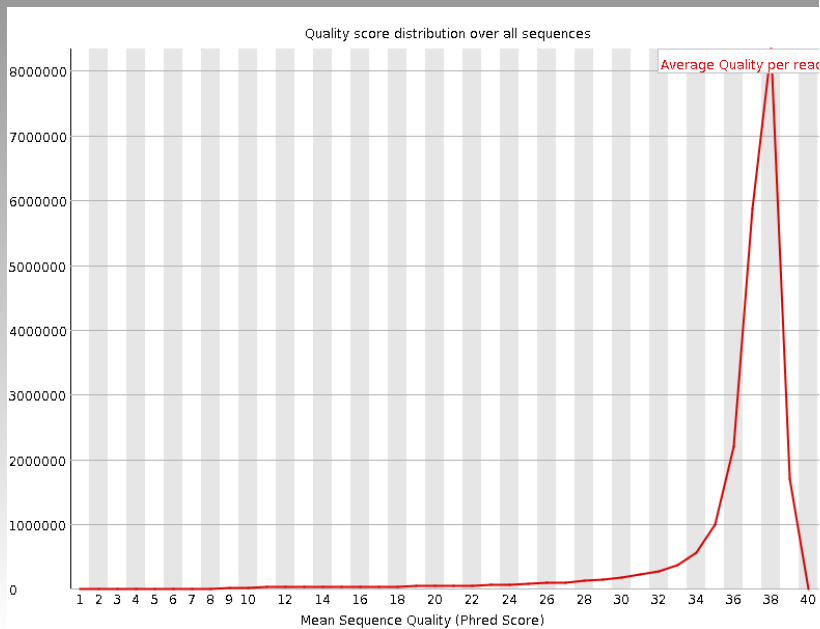


Quality

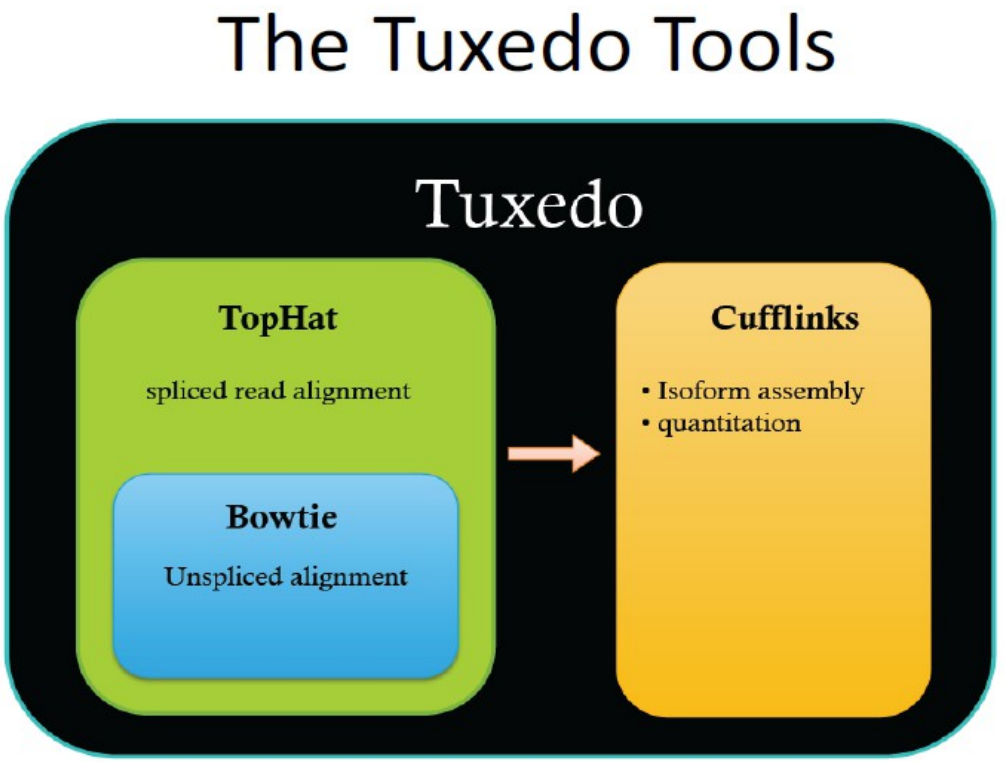
Genhong

Shankar

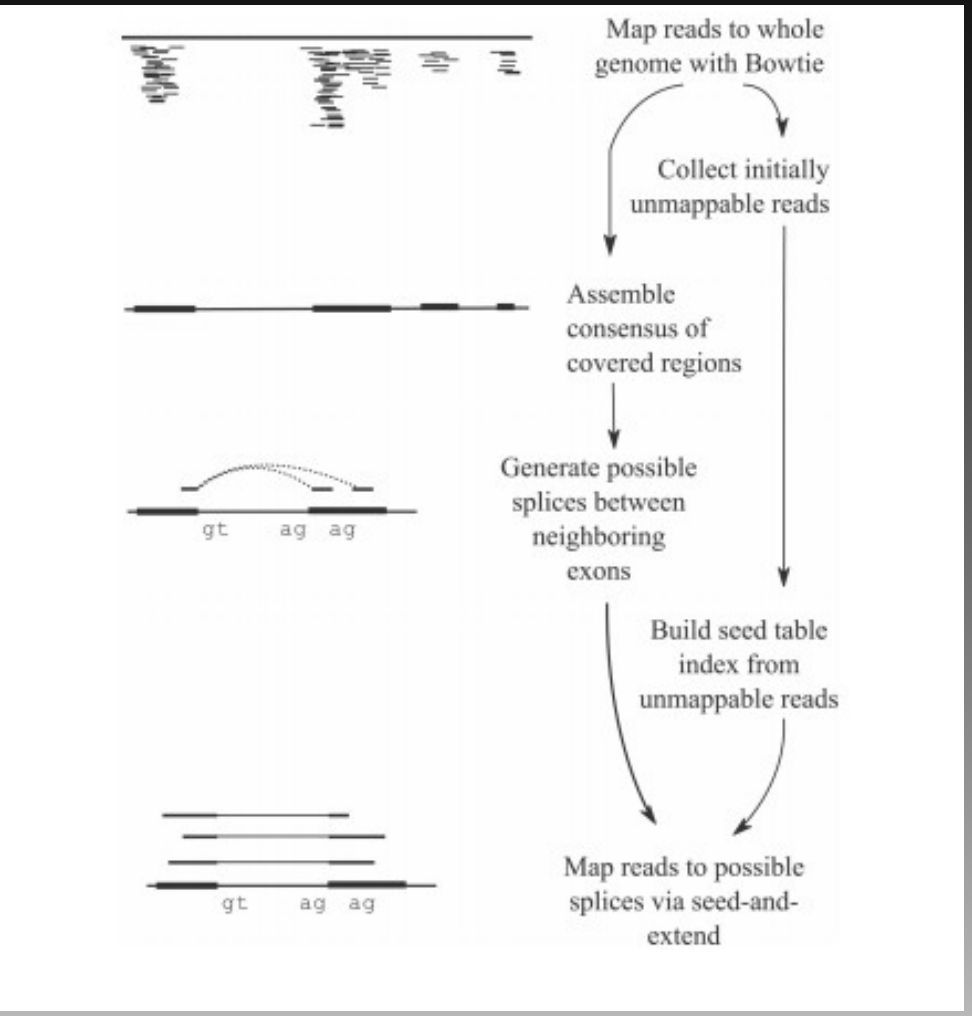
Kislay



RNA-Seq tools



Reference Mapping - TOPHAT



INPUT

FASTQ (processed)

Output (4 files)

Insertions (.bed)

Deletions (.bed)

Junctions (.bed)

Accepted Hits (.bam)

TOPHAT provides both identifying and quantifying information

.bed files can be downloaded to excel

-sam (Sequence Aligment/Map) or bam (binary compressed version of sam) – can be used to visualize reads using UCSC Genome Browser or Integrative Genomics Viewer

TopHat

TopHat Manual: http://tophat.cbcb.umd.edu/manual.html

- Running TopHat

Usage:
tophat [options] <bowtie_index> <reads1[,reads2,...,readsN]>
[reads1[,reads2,...,readsN]]
eg.
bsub “tophat -p 2 --solexa1.3-quals --max-multihits 5 -o s_1_TopHat_Out /nfs/genomes/mouse_gp_jul_07_no_random/bowtie/mm9 s_1_sequence.txt”
Options (See Manual for all available options):
-o/--output-dir Sets the name of the directory in which TopHat will write all of its output.
--solexa-quals Use the Solexa scale for quality values in FASTQ files.
--solexa1.3-quals As of the Illumina GA pipeline version 1.3, quality scores are encoded in Phred-scaled base-64. Use this option for FASTQ files from pipeline 1.3 or later.
-p/--num-threads Use this many threads to align reads. The default is 1.
-g/--max-multihits Instructs TopHat to allow up to this many alignments to the reference for a given read, and suppresses all alignments for reads with more than this many alignments. The default is 40.

TopHat Output

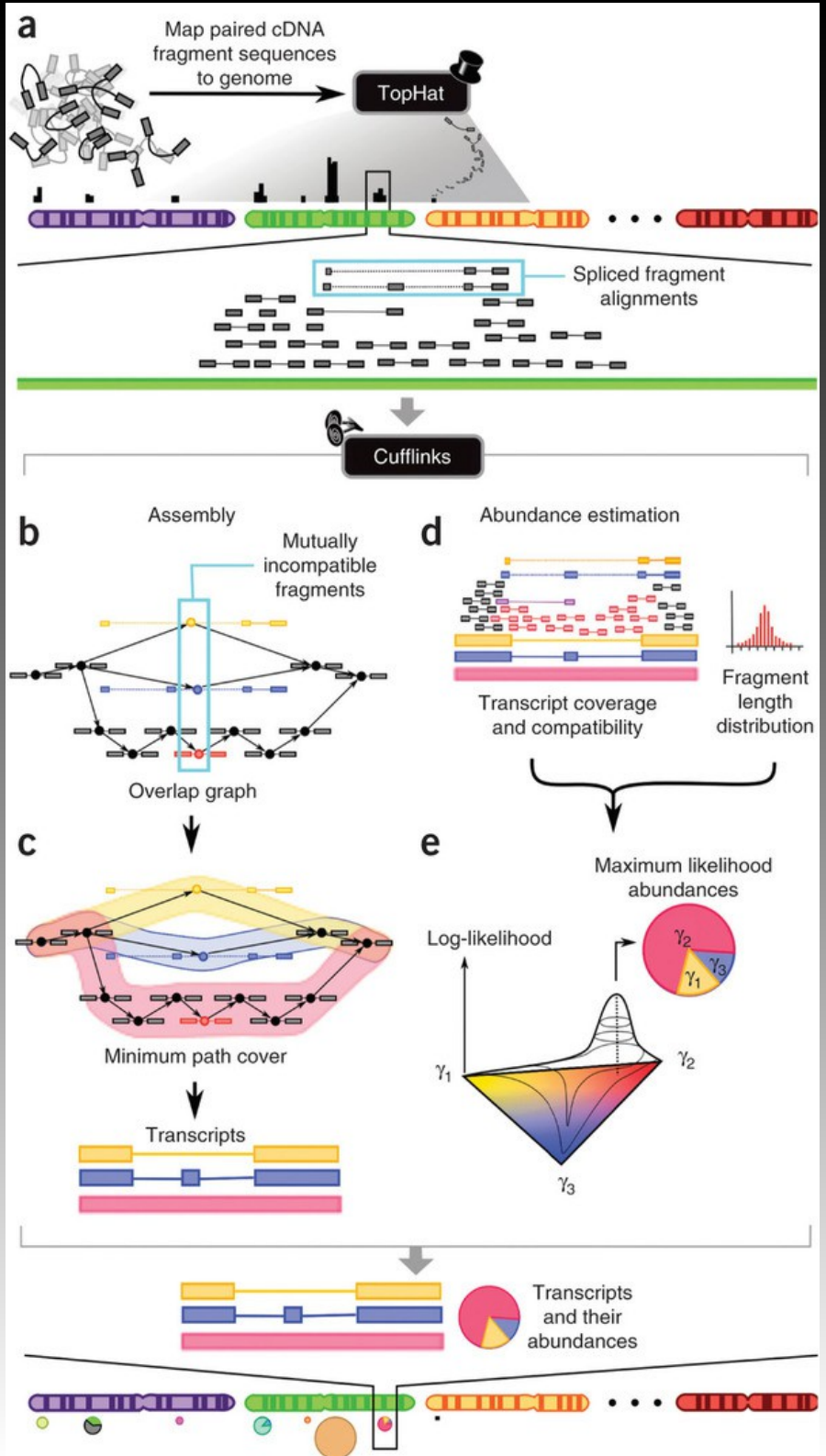
- Output of TopHat is a bam file. Binary version of Sequence Alignment/Map (SAM) file
- Use Integrative Genomics Viewer (IGV) to view bam

file or use SAMtools to analyze bam file

eg. SAM File

26

WICMT-SOLEXA:1:20:670:1533# 137 chr1 3240920 3 30M * 0 0 CTGGATCTGGACCTGGACCTGGATCTATAT ::::::::::::::-:::::::::::: NM:i:1 NH:i:2 CC:Z:chr6 CP:i:83893005
WICMT-SOLEXA:1:69:135:1285# 89 chr1 3269437 1 30M * 0 0 TGCCTAAACTTATTAAGGCAGGCCATGGGC :(/+:::(+:+':/:+++&+//':++::: NM:i:2 NH:i:4 CC:Z:chr7 CP:i:20934843
WICMT-SOLEXA:1:84:584:747# 153 chr1 3270083 0 30M * 0 0 AGCAAGTTTTTTNTTAGCCCTAGATTCCAG ::::::::::::%:::::::::::::::: NM:i:1 NH:i:5 CC:Z:= CP:i:136301734
WICMT-SOLEXA:1:75:1357:1675# 163 chr1 3522128 255 30M = 3522287 0 GTGGCTTTGTGGTCTTCACCAACCTTTCTC :::::::::::::::::::::::::::::: NM:i:1 NH:i:1
WICMT-SOLEXA:1:75:1357:1675# 83 chr1 3522287 255 30M = 3522128 0 CTGTAGGTGTAATCCTAAATTCTTATTACG :::::::::::::::::::::::::::::: NM:i:0 NH:i:1
WICMT-SOLEXA:1:8:59:283# 153 chr1 3522536 3 30M * 0 0 TTTCTGCTTTGATTATGGTACTGATGTCTG ::::::::::4:::::::::::::::::: NM:i:2 NH:i:2 CC:Z:chr5 CP:i:134317691
WICMT-SOLEXA:1:12:1161:945# 89 chr1 3523371 1 30M * 0 0 TCTACATAGCCCAAACTGGCTTTGGACTCT :::::::::::::::::::::::::::::: NM:i:0 NH:i:3 CC:Z:chr10 CP:i:117172515
WICMT-SOLEXA:1:45:1469:1826# 73 chr1 3620888 3 30M * 0 0 CAAGTATTTAATGTTTTCATTAAATTGTTT ::::::::::::::::::::::::::4::: NM:i:0 NH:i:2 CC:Z:chr11 CP:i:22903295
WICMT-SOLEXA:1:14:536:150# 73 chr1 3620943 3 30M * 0 0 CTGGAAGACAATGTCCAAAAACTCTGAATC ::::::::::::::::::::::::::%::&: NM:i:1 NH:i:2 CC:Z:chr11 CP:i:22903240
WICMT-SOLEXA:1:66:646:1188# 137 chr1 3662923 0 30M * 0 0 AAAAAAAAAACACCACCCCCAACAAAAAAA +00++0+0+''0++++:00::.&:::,:,: NM:i:2 NH:i:5 CC:Z:chr10 CP:i:94881279



Estimating Transcript Abundance - Cufflinks

INPUT

.bam file (Accepted Hits)

Reference (.gtf)

Refseq, Ensembl, etc

Output (tabular form, excel)

FPKM quantifiable

Cufflinks:
Assemble and Quantify Reads

- Cufflinks Manual:

http://cufflinks.cbcb.umd.edu/manual.html

- Running Cufflinks
- Optional: Supply annotation in GTF format with

"-G" option

Usage:

cufflinks [options] <hits.bam>

eg.

bsub "cufflinks -p 2 -o s_1_Cufflinks_Out s_1_TopHat_Out/accepted_hits.bam"

eg. cufflinks will assemble and quantify using known transcripts using gtf file supplied

bsub "cufflinks -p 2 -G transcripts.gtf accepted_hits.bam"

Cufflinks Output

- Output of Cufflinks is a GTF file with assembled isoforms

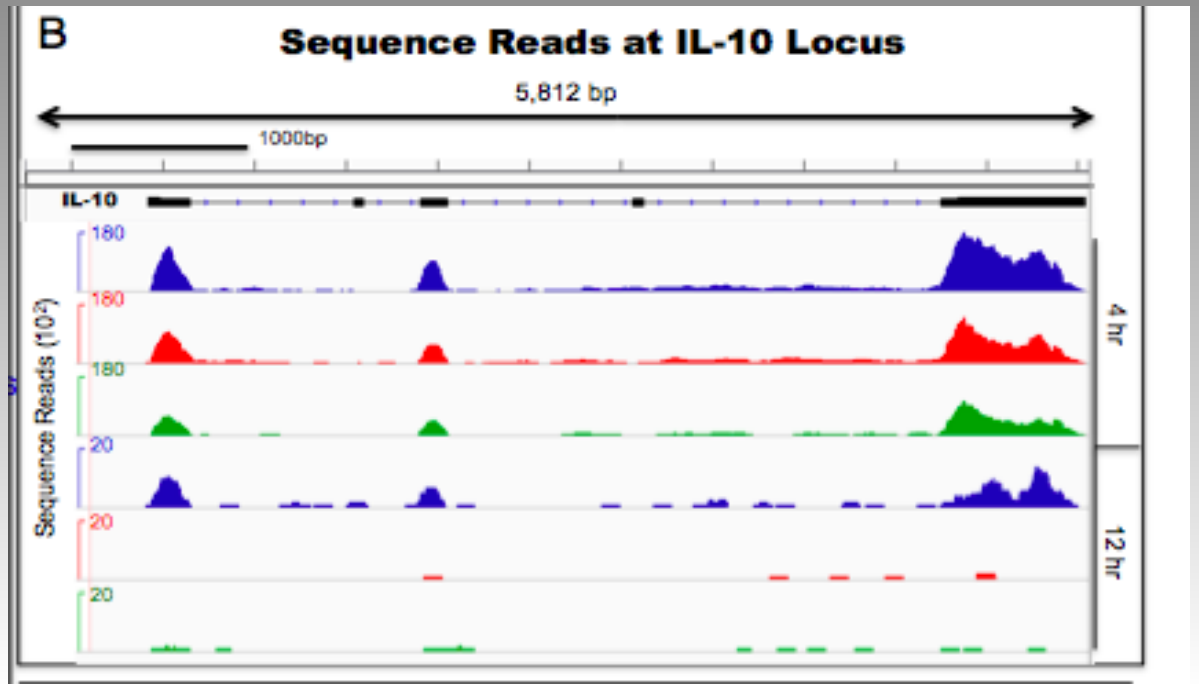
eg.

chr1 Cufflinks transcript 36321447 36330270 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks exon 36321447 36323398 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; exon_number "1"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks exon 36325501 36325554 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; exon_number "2"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks exon 36326058 36326546 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; exon_number "3"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks exon 36330183 36330270 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; exon_number "4"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks transcript 36364578 36380874 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36364578 36364681 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "1"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36373054 36373172 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "2"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36374929 36375026 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "3"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36375333 36375498 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "4"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36375837 36380874 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "5"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";

Visualizing Reads Across the Genome

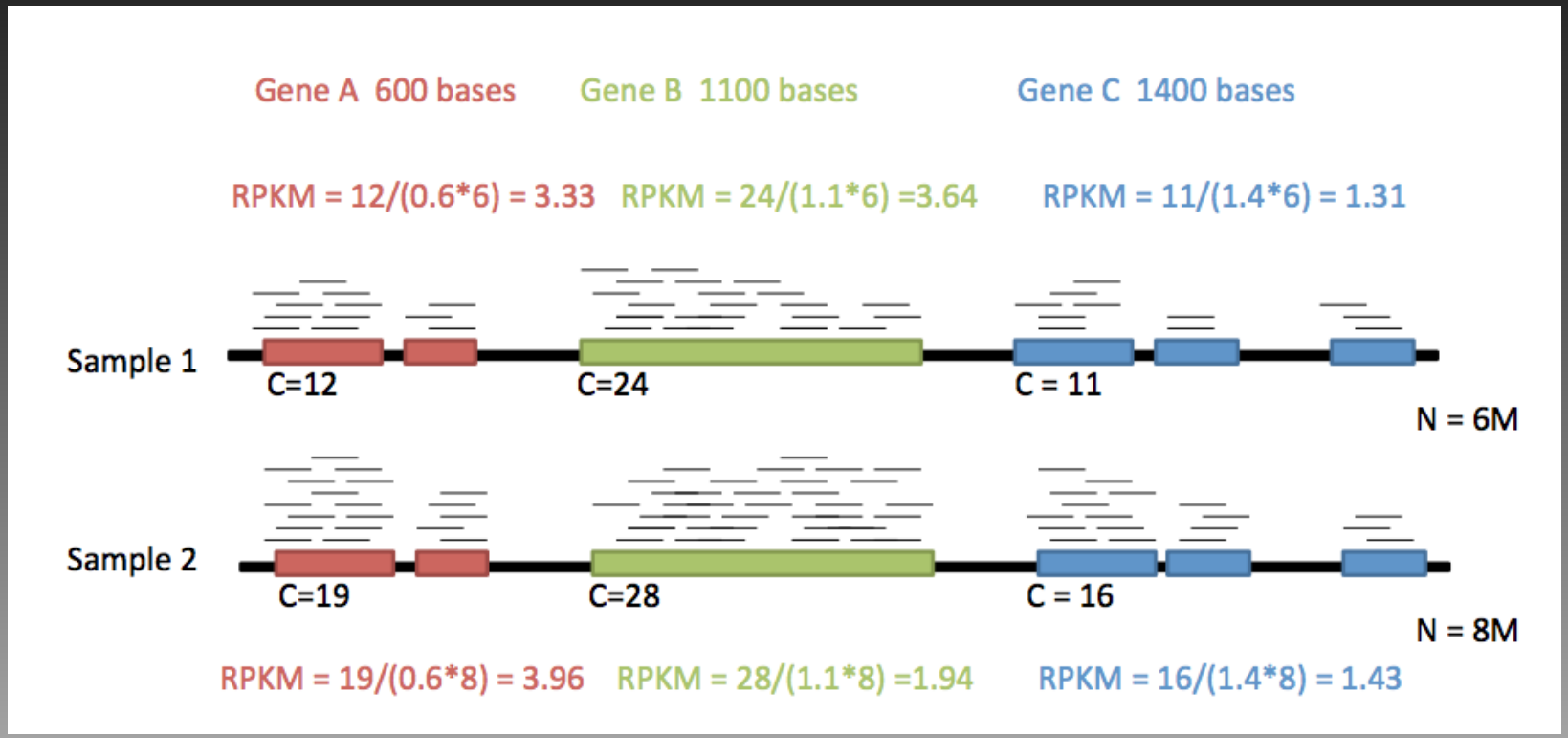
Upload Files to UCSC Genome Browser
Convert .bam file to .bedgraph (using Galaxy)
Requires some coding
Size Limitations

Upload Files to Integrative Genome Viewer
Convert .bam file to .bedgraph (using Galaxy)
Upload directly



How do I quantify expression from RNA-seq?

RPKM: Reads per Kb million (Mortazavi et al. Nature Methods 2008)



Longer and more highly expressed transcripts are more likely be represented among RNA-seq reads

RPKM normalizes by transcript length and the total number of reads captured and mapped in the experiment

Sequencing depth can alter RPKM values

Differential Gene Expression Analysis

RPKM

- -Can calculate Fold change
- -Input sequence reads must be similar
- -replicates not needed
- -provides NO statistical test for differential gene expression
- -useful for Cluster based classification of genes

http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/Help/4%20Quantitation/4.3%20Pipelines/4.3.1%20RNA-Seq%20quantitation%20pipeline.html

CuffDiff (available on GALAXY)

- -Input .bam file
- -Can set statistical threshold ($p<0.05$)
- -replicates encouraged but not needed
- -Input sequence reads can be somewhat dissimilar
- -can provide differential splicing and promoter usage

DESeq

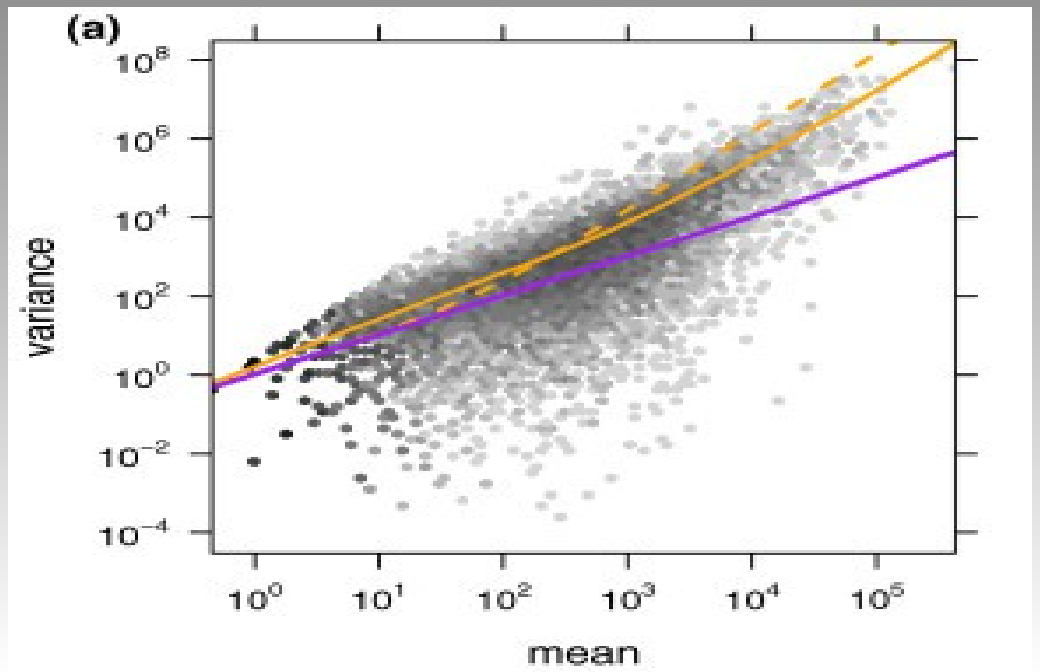
- -Input .bam file
- -Can set statistical threshold
- -Input sequence reads can be somewhat dissimilar
- -Must have replicates
- -Not currently on Galaxy (must use Edge R)

Differential Gene Expression Analysis: Sampling Variance

Consider a bag of balls with K number of red balls where K is much less than the total number of balls. You can sample n number of balls. P represents the proportion of red balls in your sample.

Estimate of the number of balls (u) = pn
K (the actual number of balls) follows a Poisson distribution and hence K varies around the expected value (u) with a standard deviation of 1/ sqroot (u)

Microarray data follows a Poisson distribution. However RNA seq does not.
In RNA Seq genes with high mean counts (either because they're long or highly expressed) tend to show more variance (between samples) than genes with low mean counts. Thus this data fits a Negative Binomial Distribution



Poisson
Negative Binomial

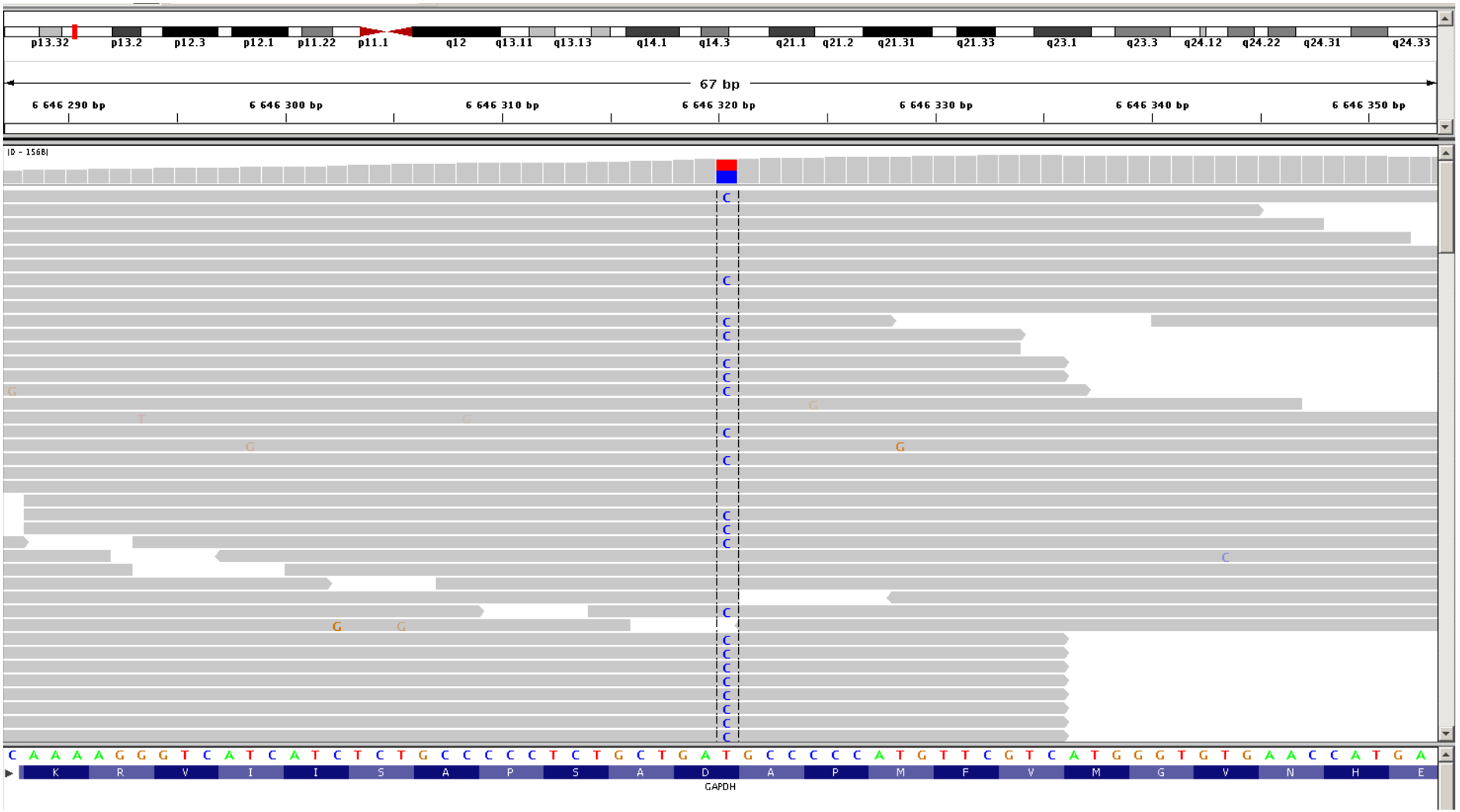
Differential Gene Expression Analysis

CuffDiff: If you have two samples, cuffdiff tests, for each transcript whether there is evidence that the concentration of this transcript is not the same in the two samples

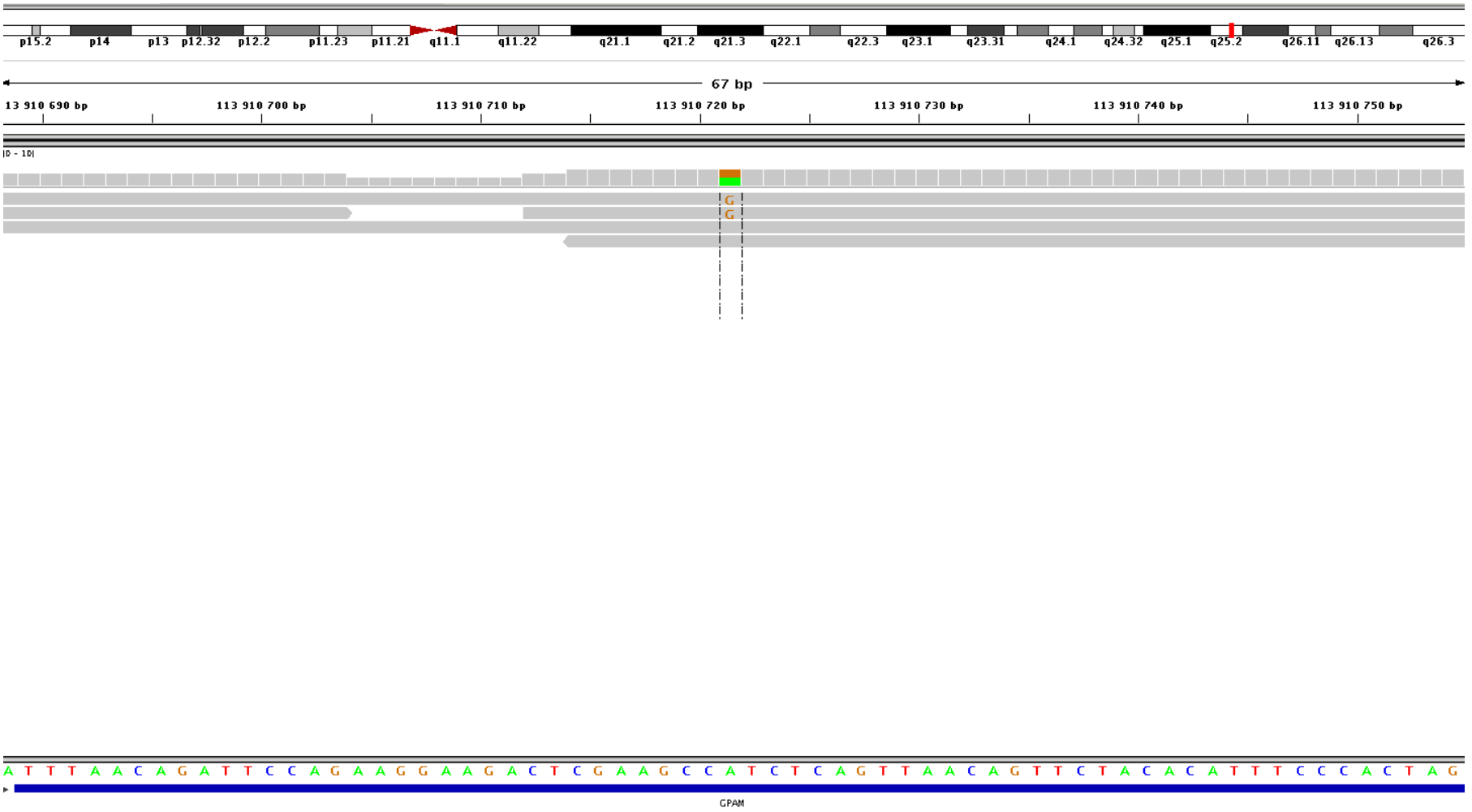
DESeq/EdgeR: If you have two different experimental conditions, with replicates for each condition, DESeq tests whether, for a given gene, the change in the expression strength between the two conditions is large as compared to the variation within each group.

You will get different answers with different tests

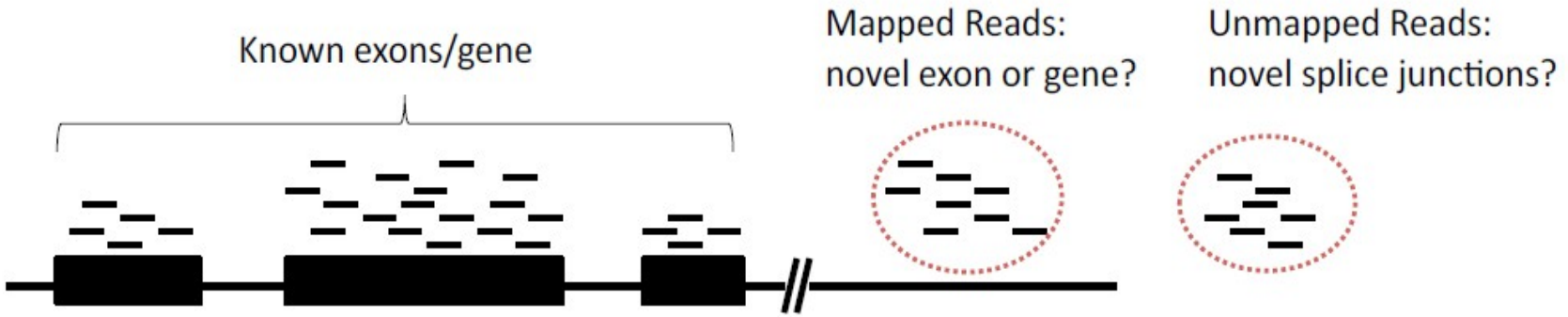
SNP discovery and quantification of allele expression

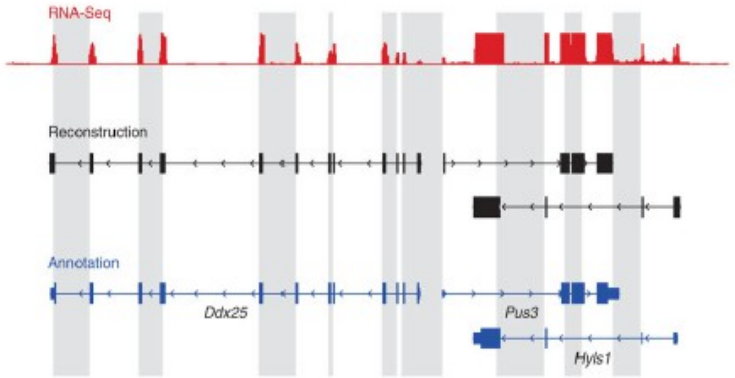


SNP can be detected only with 4 reads

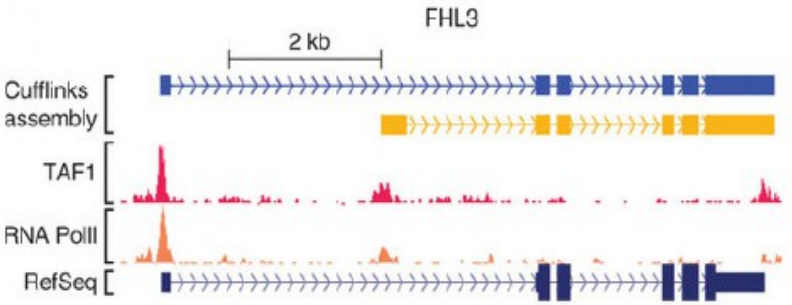


RNA-Seq Applications – Annotation: Identify Known and Novel Transcripts



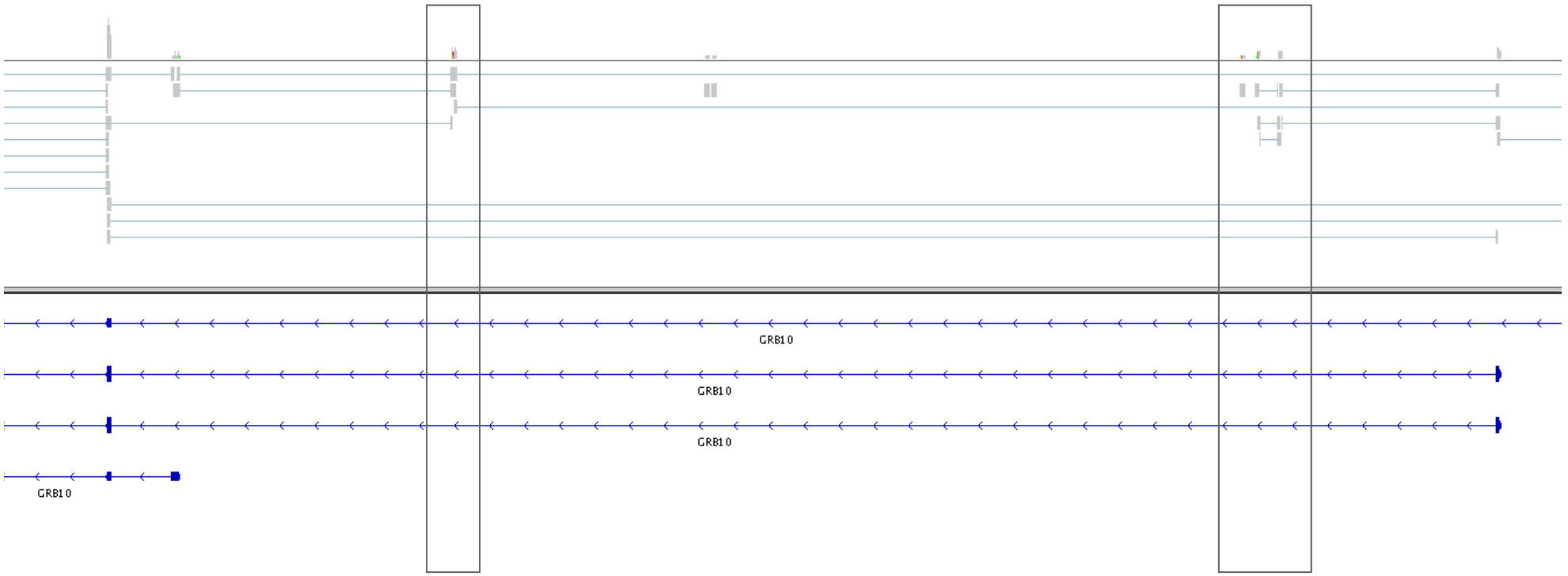


Guttman, M. et al *Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs* Nature Biotechnology (2010)

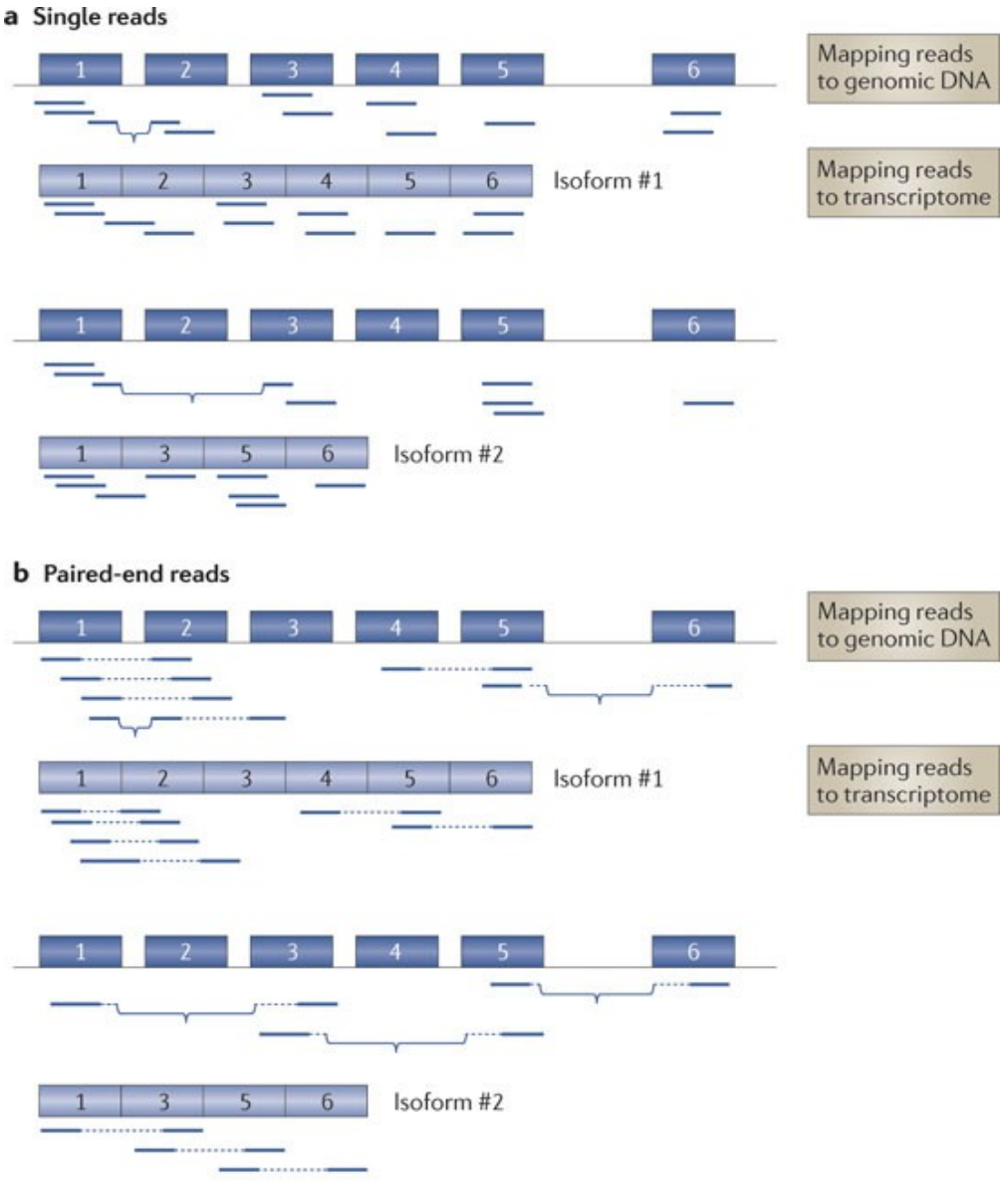


Trapnell, C. et al *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation* Nature Biotechnology (2010)

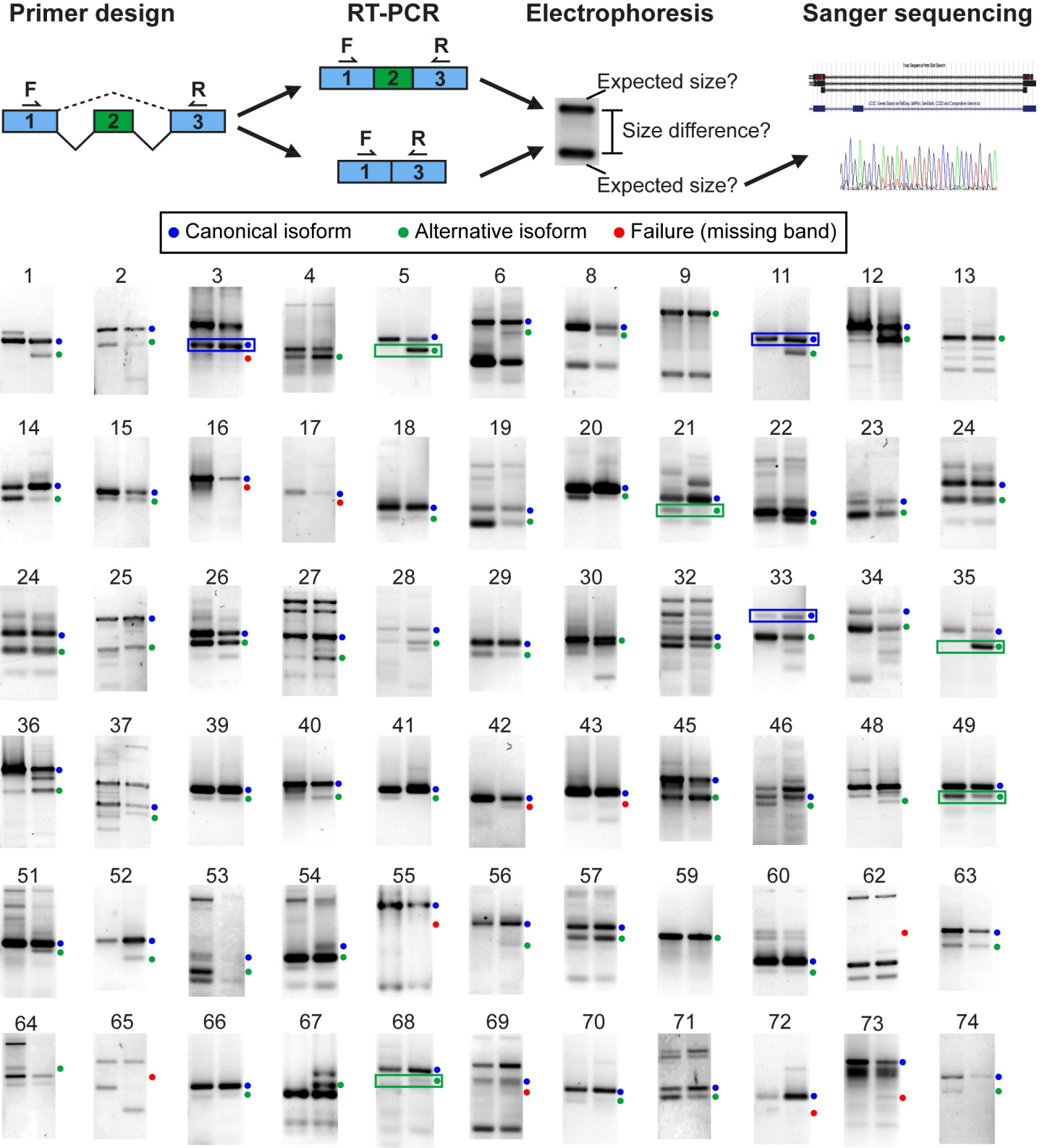
Discovery of novel exons



Alternative splicing events



Nature Reviews | Genetics



Validation

Overall validation rate - 85%