Sequencing and Transcriptomics

Nathan Hall
n.hall@latrobe.edu.au

Senior Research Bioinformatician
La Trobe University
Life Sciences Computation Centre, VLSCI

vlsci
LIFE SCIENCES COMPUTATION

LA TROBE UNIVERSITY

Bioinformatics

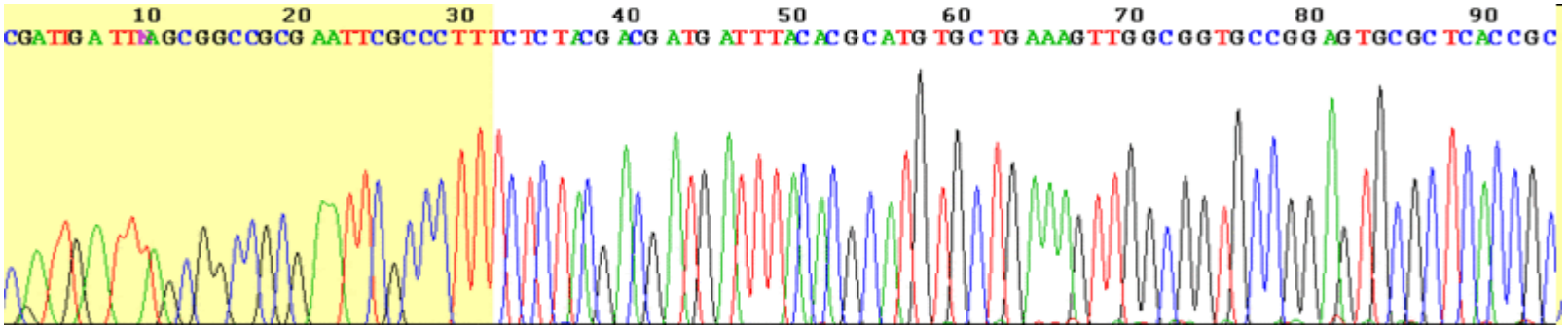
- Quackenbush says … bioinformatics has become "an essential foundation for all molecular and genomic science," that gives him the opportunity to inquire, discover, and do things no one has ever been able to do.

New way of doing science!

- OLD
  - Come up with a hypothesis
  - Do an experiment to try and test it
  - Get your results and see if you were right
- NEW
  - Massive amounts of data
  - Analyse data and see what you find
  - Follow up interesting things

Sequencing Genomes

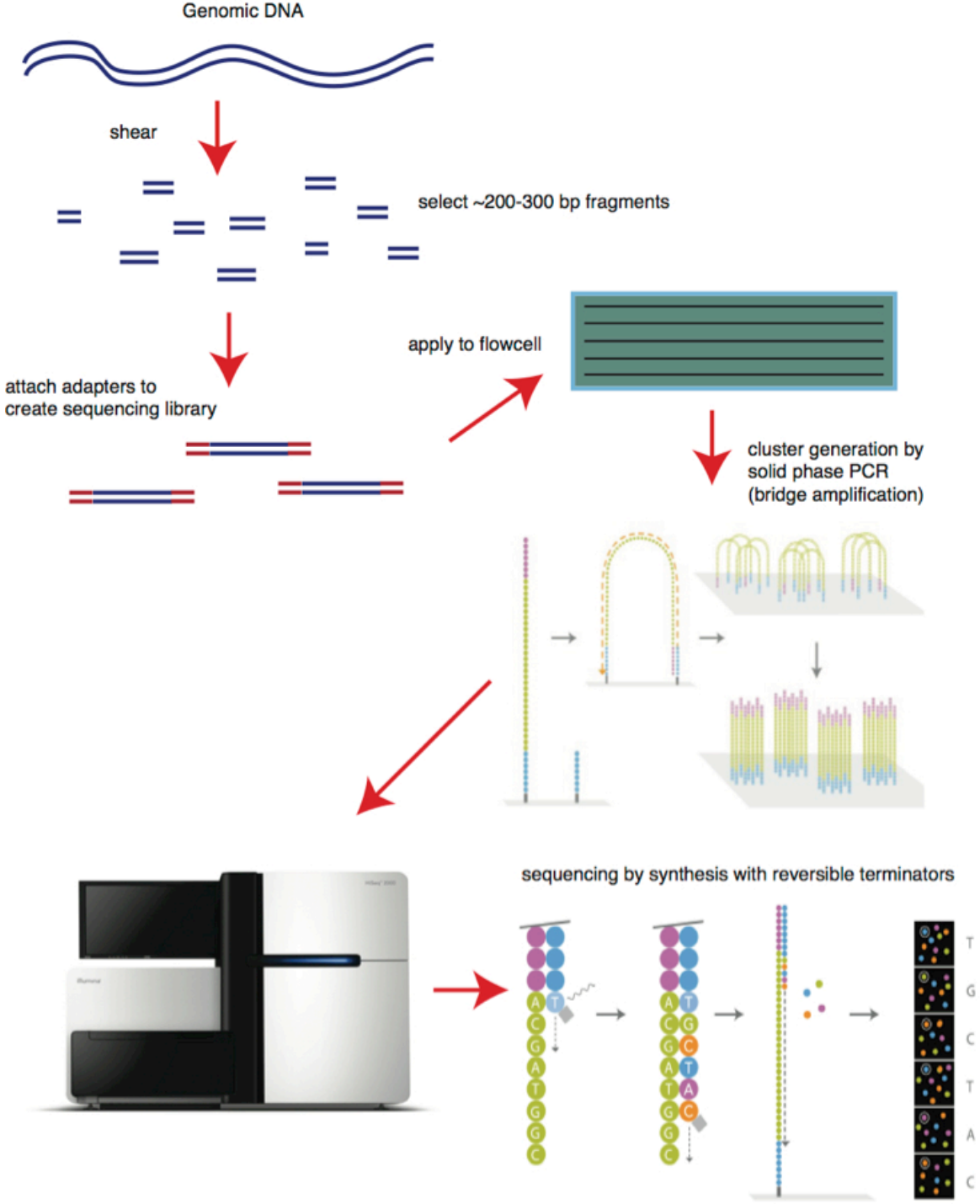
- Historically genomes have been sequenced using *Sanger Sequencing*
  - *Slow*
  - *Expensive*
  - *Dye/Gel-based*
  - *Accurate*
  - *Reads up to 1000bp*
  - *Used for sequencing prior to about 2004*
    - *Human Genome Project*



Illumina

- ~80% of next gen sequencing market
- HiSeq4000 (& MiSeq)
- Reads typically 100bp (up to 350)
- Short reads can make analysis complicated
- http://www.youtube.com/watch?v=77r5p8IBwJk
- http://www.youtube.com/watch?v=HtuUFUnYB9Y

Sequence by synthesis



I.

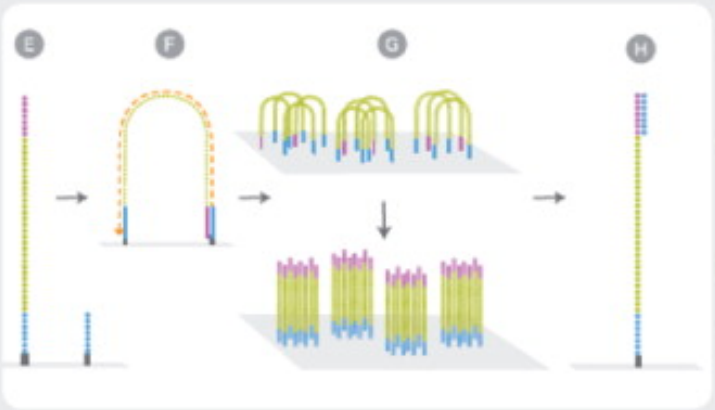
A Fragment DNA

B Repair ends/ Add A overhang

C Ligate adapters

D Select ligated DNA

II.



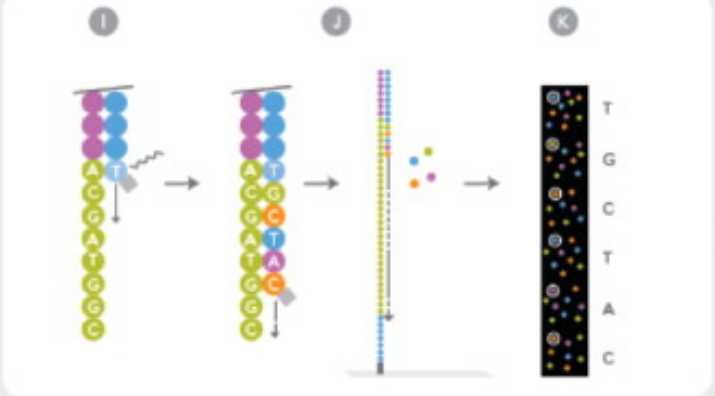
E Attach DNA to flow cell

F Perform bridge amplification

G Generate clusters

H Anneal sequencing primer

III.

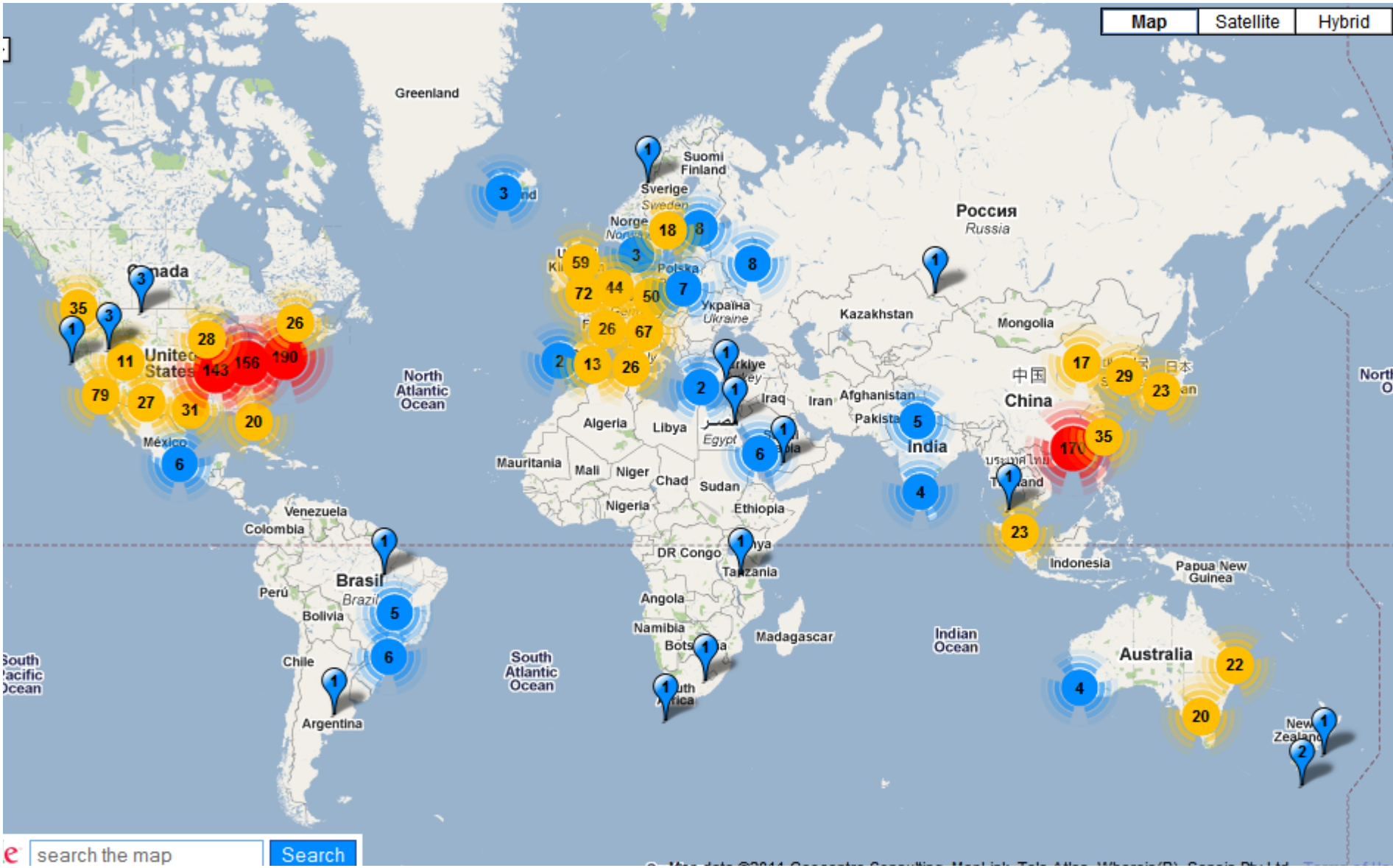


I Extend first base, read, and deblock

J Repeat step above to extend strand

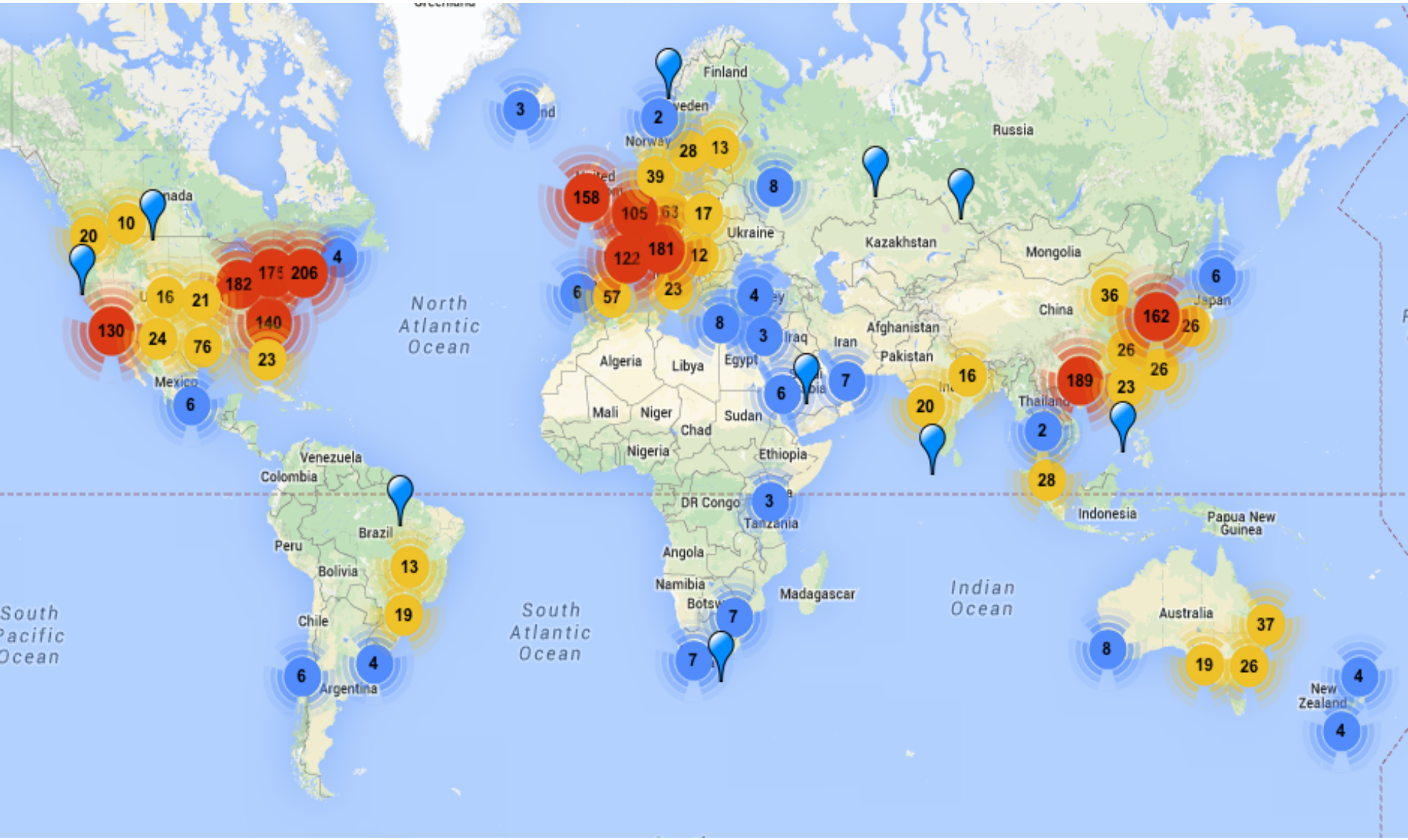
K Generate base calls

Next Generation Genomics: World Map of High-throughput Sequencers 2011



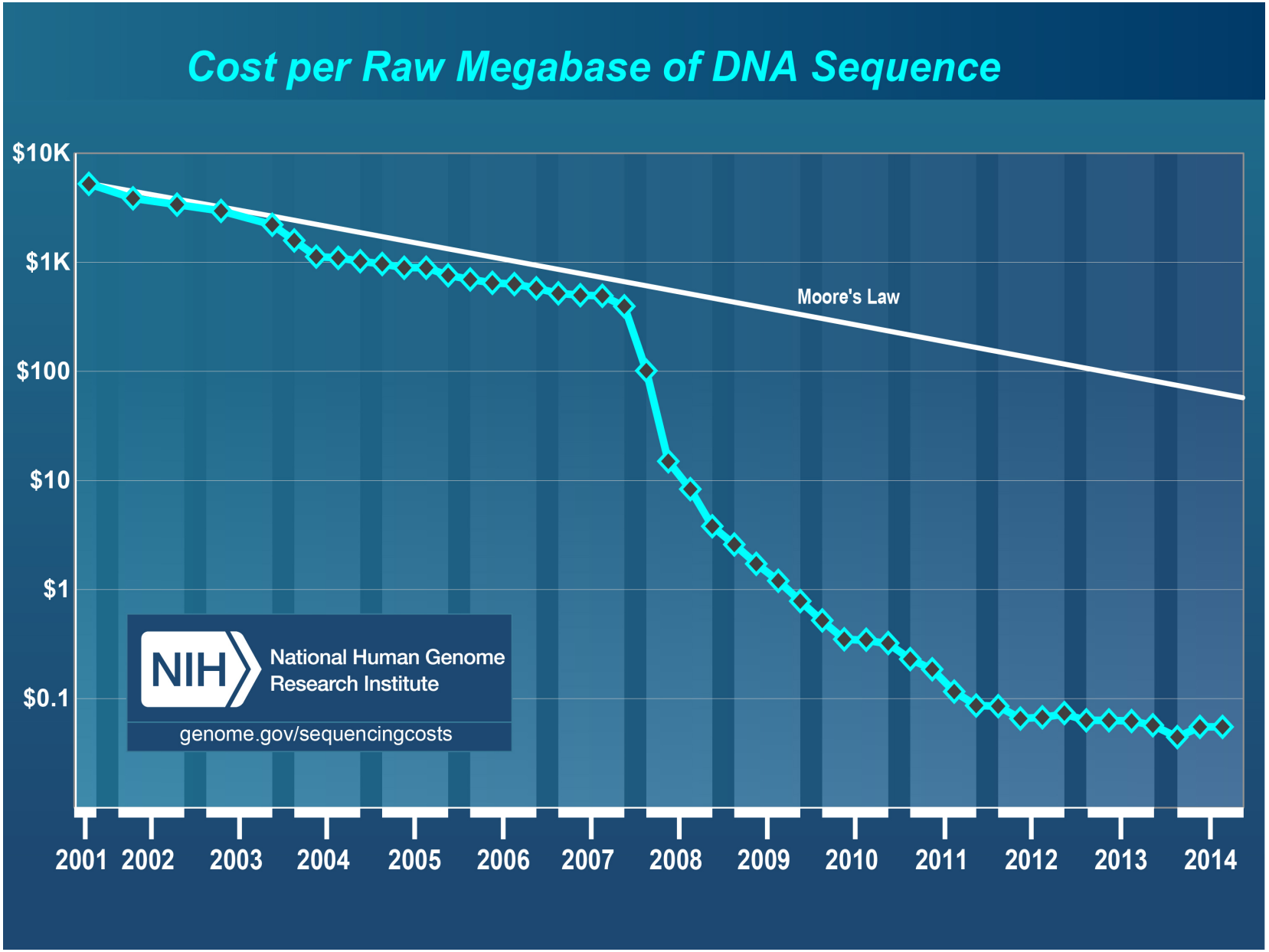
http://omicsmaps.com/

Next Generation Genomics: World Map of High-throughput Sequencers April 2014



http://omicsmaps.com/

Costs Dramatically Decreasing



Oxford Nanopore MinION

Prototypes available
Single Molecule
Potentially very long reads
with no loss of quality

DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

DNA DOUBLE HELIX

❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

TGATATTGCTTTTGATGCCG

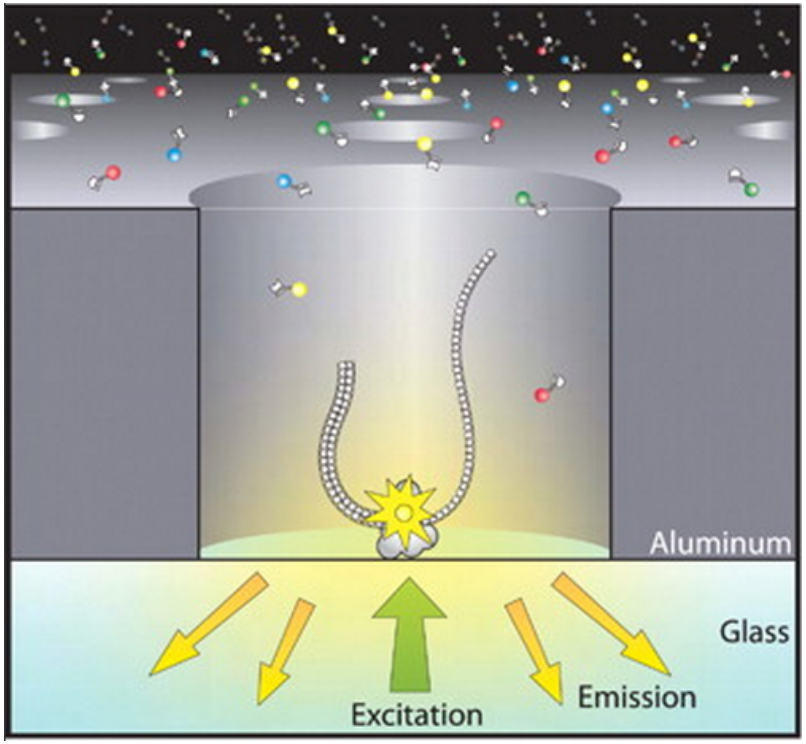
❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

MEMBRANE

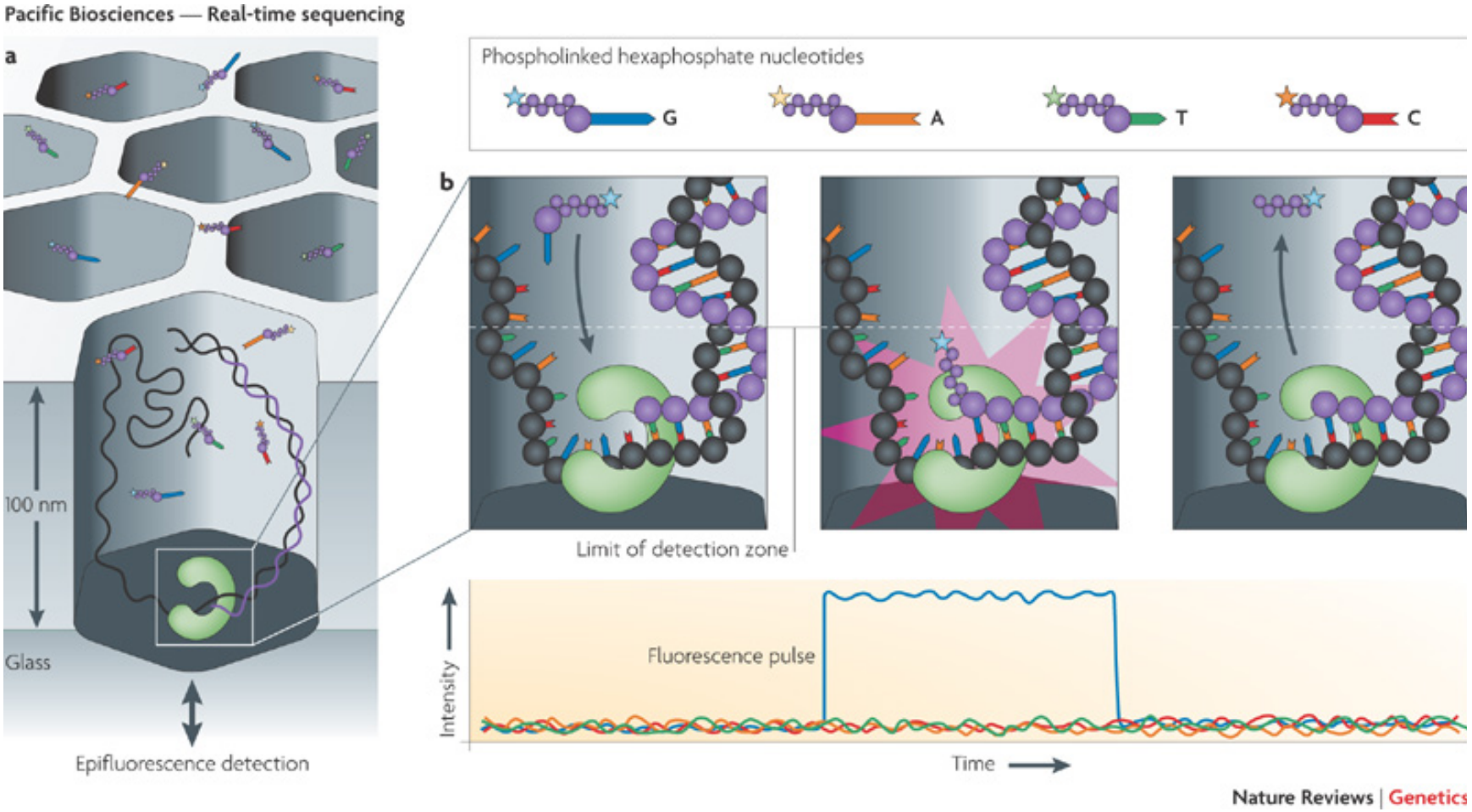
http://www.technologyreview.com/files/84993/nanopore_x616[1].jpg

Pac Bio sequencing

- Single molecule real time sequencing
- Long reads 10kb or more
- Very high error
  - 14 %, but *random*
- More expensive
- Best applications
  - Microbial genome assembly
  - Structural rearrangements
- Human genome would cost $2-500,000

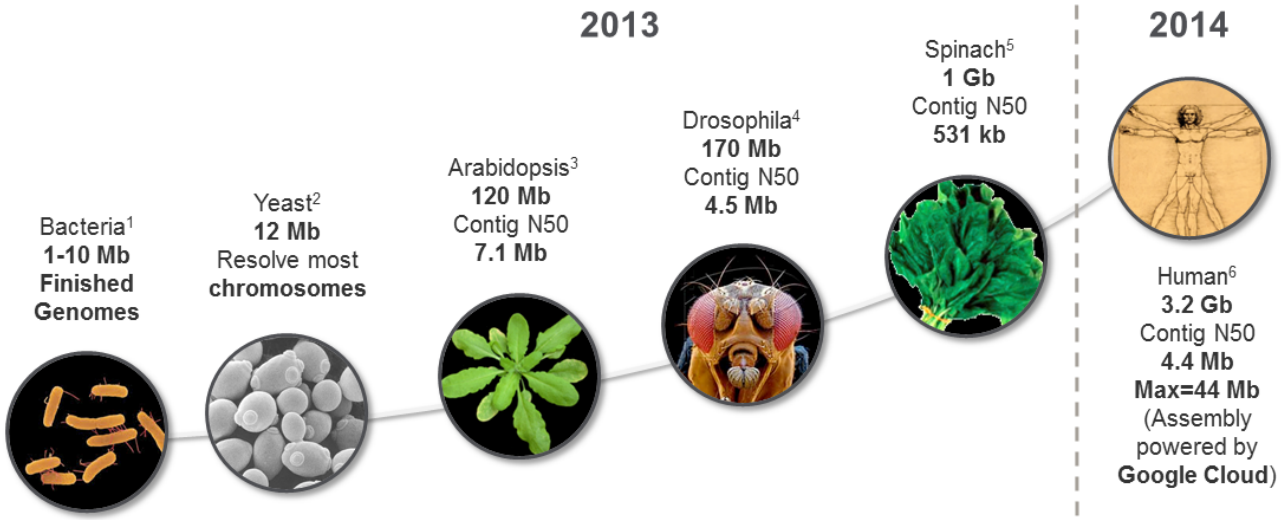


PacBio Technology – (SMRT) Single Molecule Real Time



PacBio

- High error rate (~12%) but ***random***
- Long reads, average 20 kb, up to 60+ kb
- Eliminates high/low GC problems
- Eliminates repetitive sequencing problems
- Very Expensive - $300,000 per genome
- Financially viable for bacterial genomes
- https://www.youtube.com/watch?v=v8p4ph2MAvI



Phred qualities

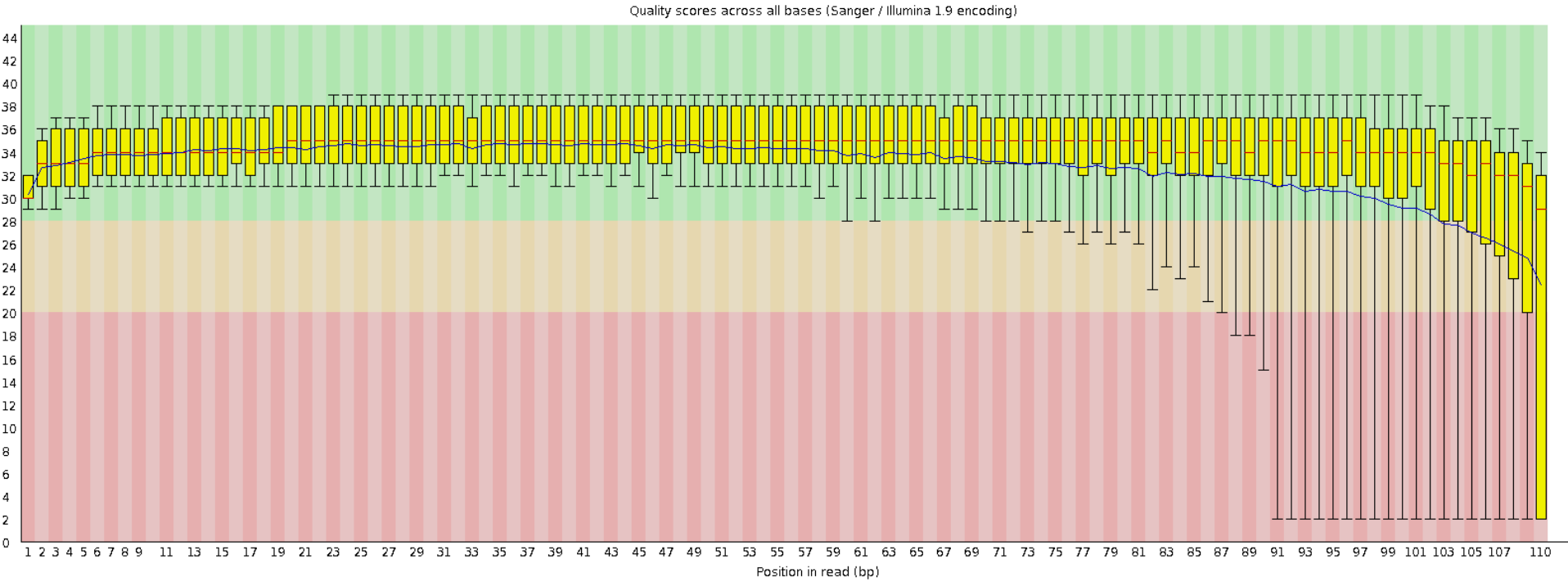
Quality value	Chance it is wrong	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Data Output - Fastq files

- Paired End data
  - FILENAME_1.fastq (or _R1)
  - FILENAME_2.fastq (or _R2)
- Header, sequence, redundant + line & quality line
- Sometime 200 million reads or more per file

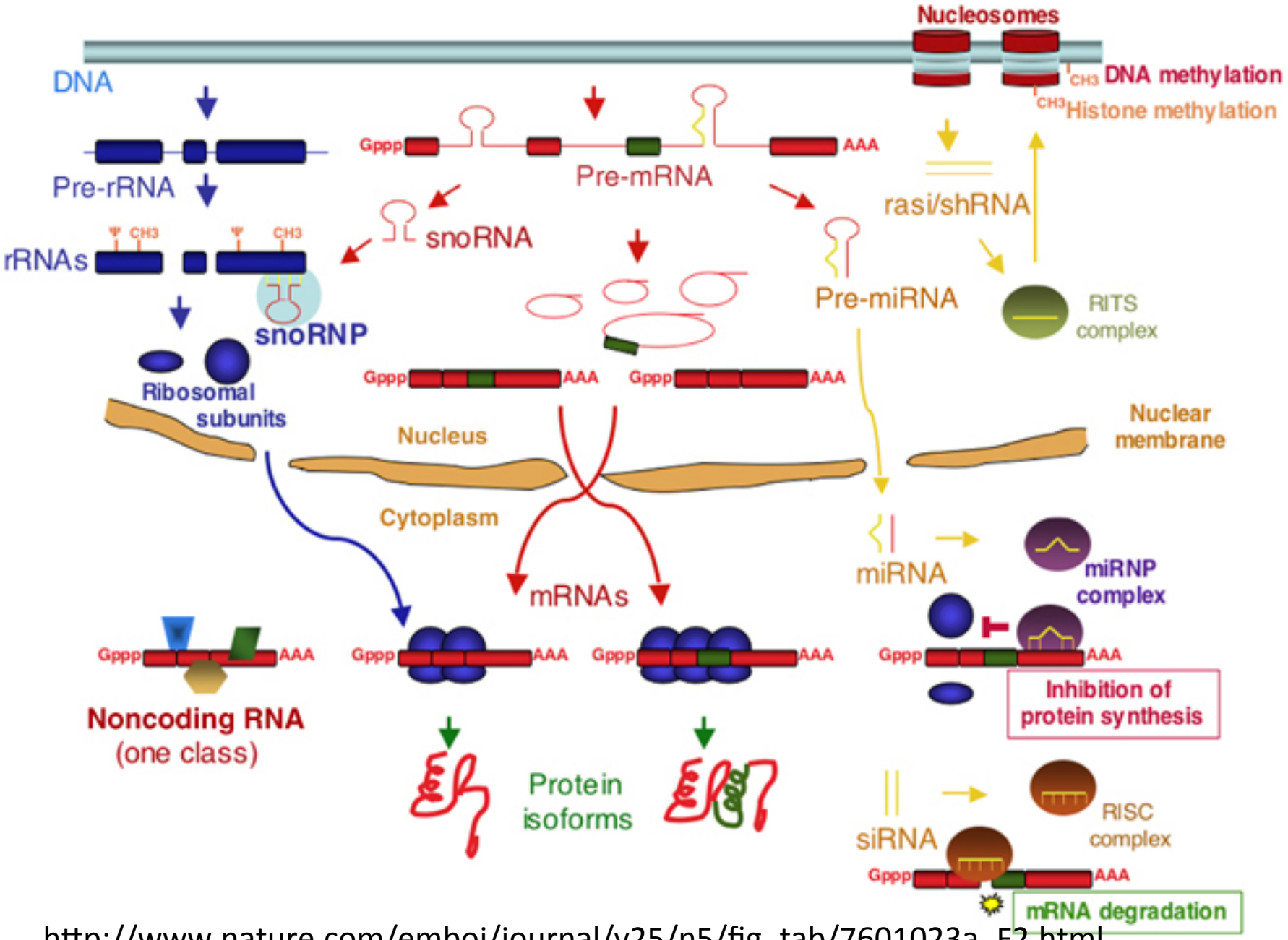
```
@HWI-ST945:139:C0M5FACXX:6:1101:5454:3483 2:N:0:ATTCCT
TGTATCAAGGCGAGGGCGGAGAGGCAGACGACACGCGAAAGGACCAGGCGATAGACGAGTACT
+
CCBFFFFFHHHHHJJJJJJIJIIJJIJIIHHFFDDDDDDDDDDDDDDDDDDDDDDDB8?CD
@HWI-ST945:139:C0M5FACXX:6:1101:5256:3485 2:N:0:ATTCCT
GTGGAACTTGTCCTGGCACACTGAACATTCCCAGCAGATGCCATGGTAACTGGCACATTCACT
+
C@BFFFFFHHHHHJJJJIJJJJJJJJJJJJJJJJJJJIJJJJJJJJIJJJIJJJJJJIJJJJ
```

fastqc - quality analysis



RNA-seq -- Transcriptomics

- Massively parallel sequencing of RNA
- Discovery – what is being expressed
- Differential Expression
  - Comparison of RNA from different conditions
  - Compare transcriptome of different populations



http://www.nature.com/emboj/journal/v25/n5/fig_tab/7601023a_F2.html

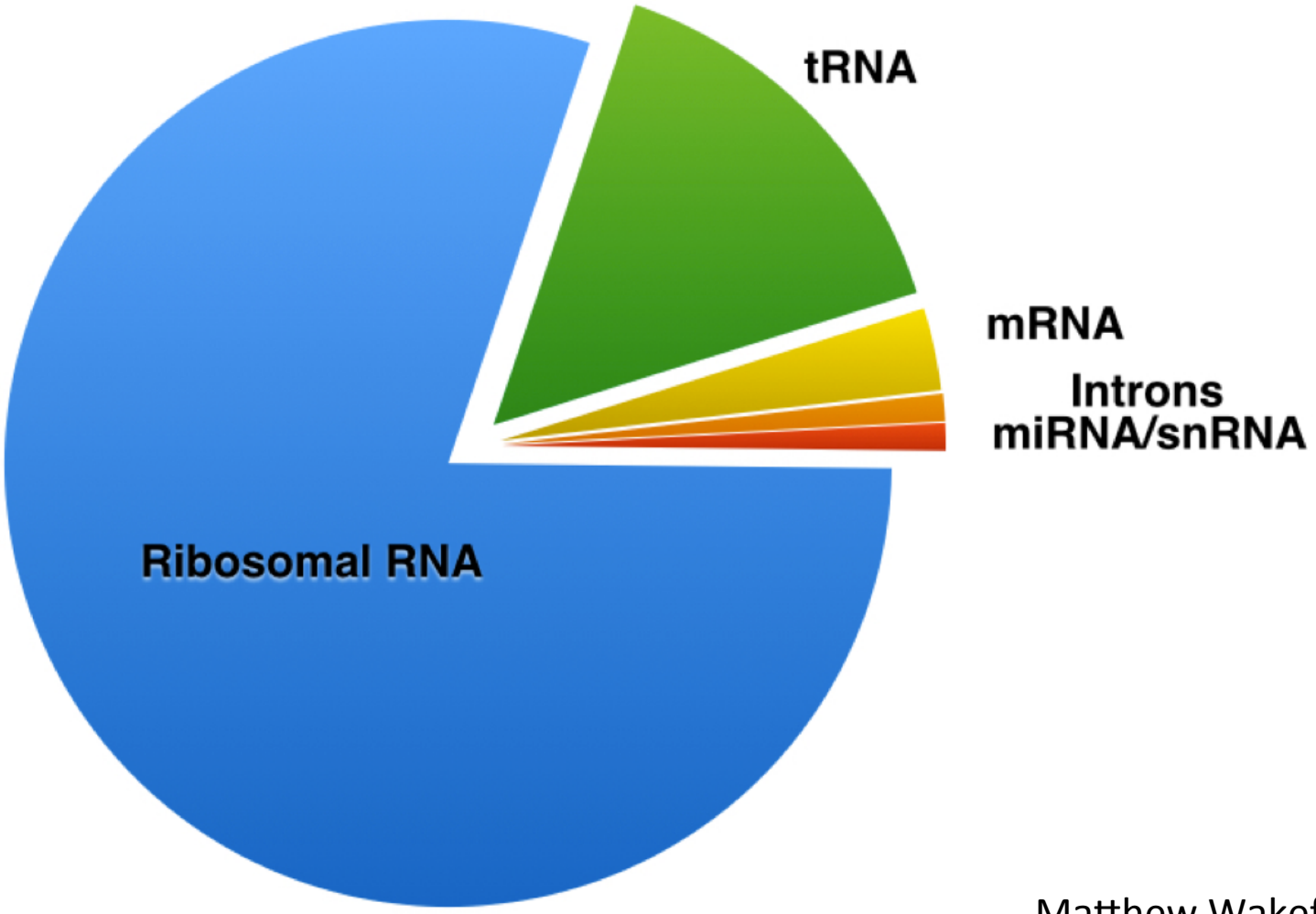
Gene expression

- Can measure what and how much RNA
- Which genes are over/under expressed in healthy v's normal patients/environments
- Which genes are correlated to a phenotype
- What genes are differentially expressed in different cell types

Gene expression 2

- What genes are expressed by our gut microbiota
- How are gene expression and early disease development correlated
- What cellular pathways are activated during heat stress
- What genes are differentially spliced in different cell/tissue types/environments
- What genes are found in novel organism

RNA in the Cell



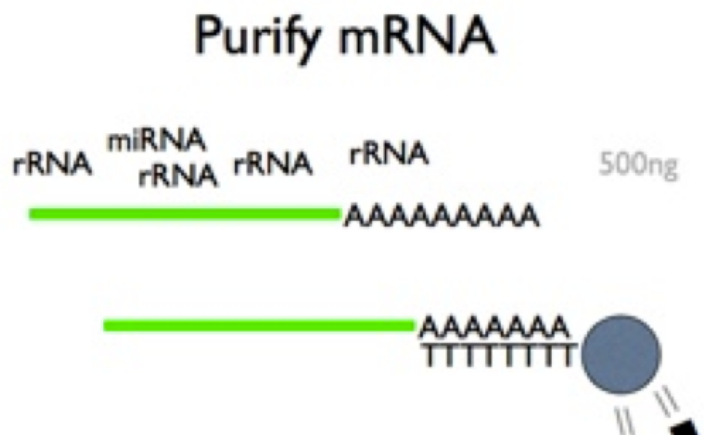
Matthew Wakefield

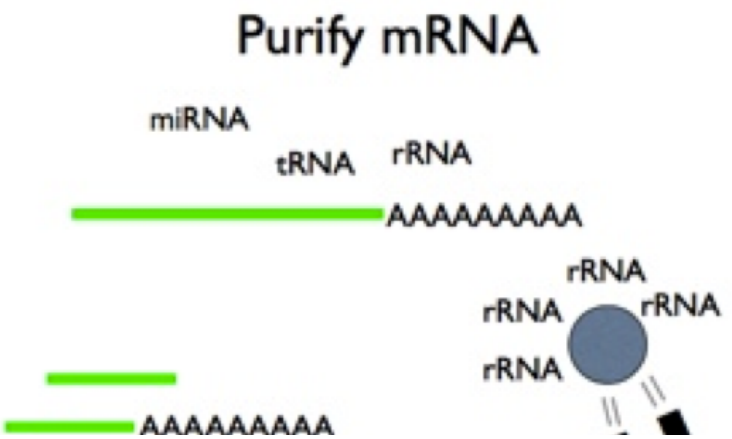
mRNA purification

We usually don't want to sequence rRNA and tRNA. We can:

Keep only mRNA:
Poly(A) pull down

Remove unwanted RNA:
ribosomal depletion

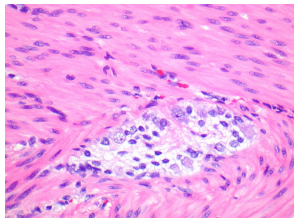




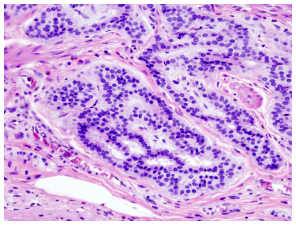
Matthew Wakefield

RNAseq - technique

Samples of interest

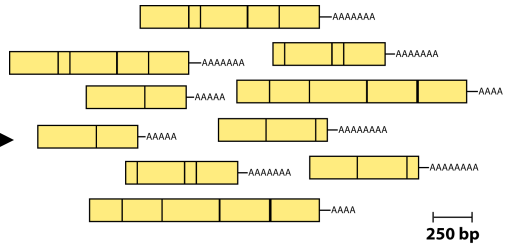


Condition 1
(normal colon)

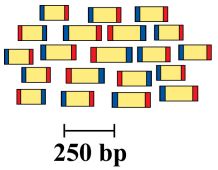


Condition 2
(colon tumor)

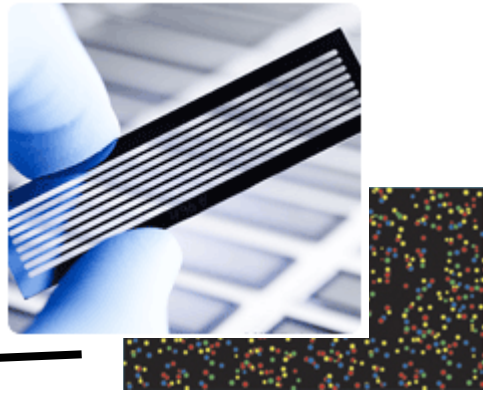
Isolate RNAs



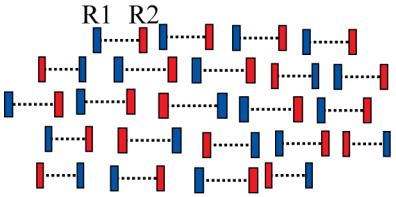
Generate cDNA, fragment, size select, add linkers



Sequence ends



100s of millions of paired reads
10s of billions bases of sequence



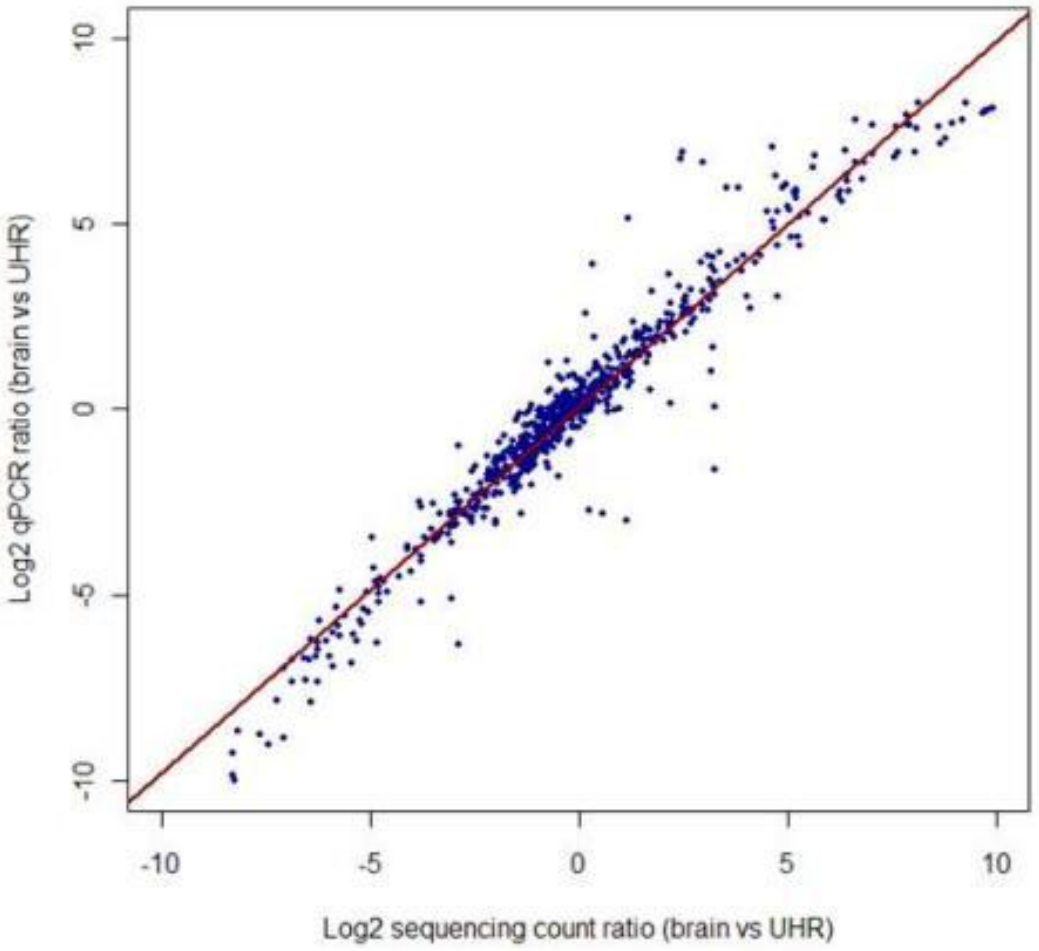
Map to genome, transcriptome, and predicted exon junctions

Downstream analysis

RNA-Seq

- Digital expression
- No *a.priori* information about the transcriptome required
- Sensitive to splicing
- High dynamic range – increase by more sequencing
- Signal normalised across the transcript gives much better gene v's gene quantitation
- Determine bp changes in RNA (Editing events, expressed SNPs/Indels)

RNA-seq vs. qPCR



Accuracy and Sensitivity

illumina®

Alternative Splicing

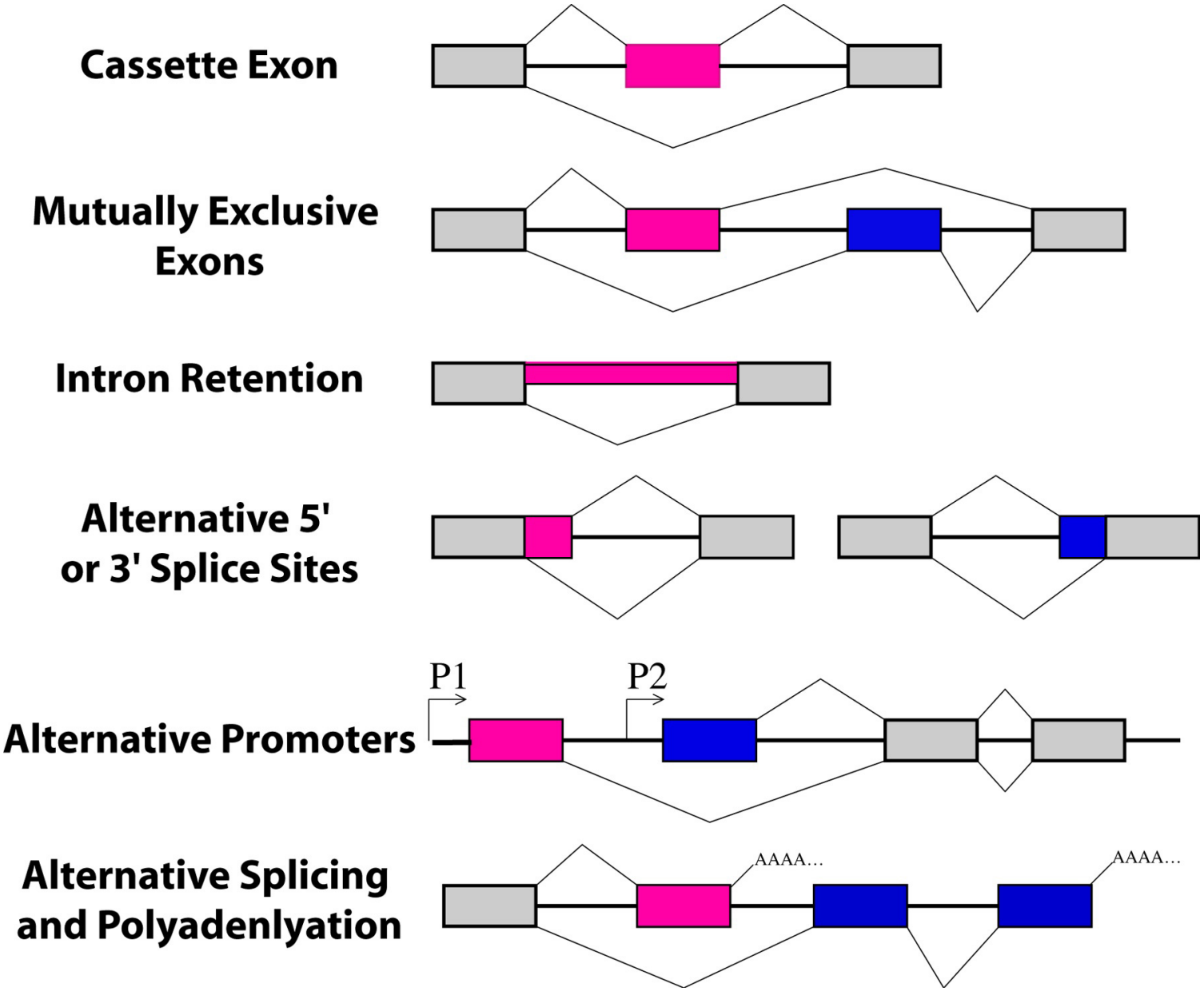
- It is well understood that many genes are alternatively spliced to give different protein or non-coding products
- The old *paradigm* of one gene

  = one transcript

  = one protein

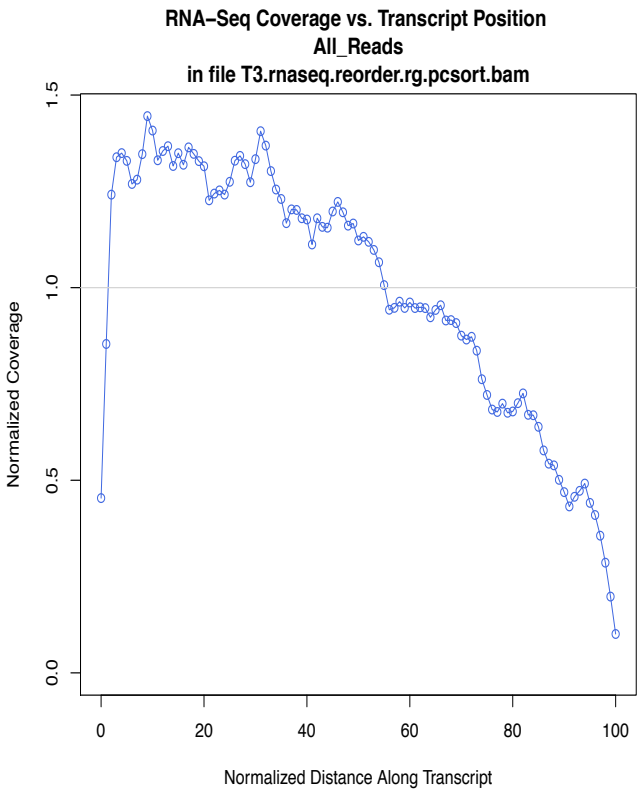
  is long gone

Alternative Splicing

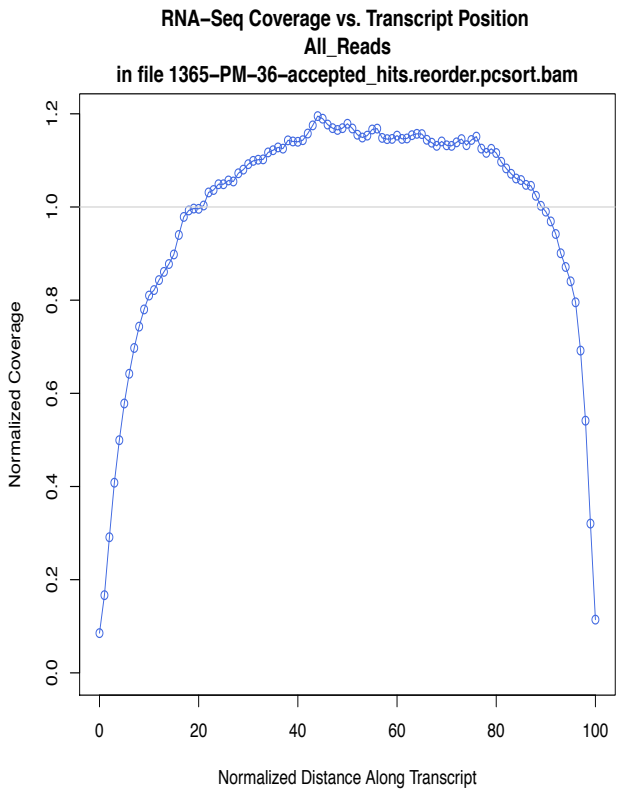


Sample prep can create 3’ or 5’ bias

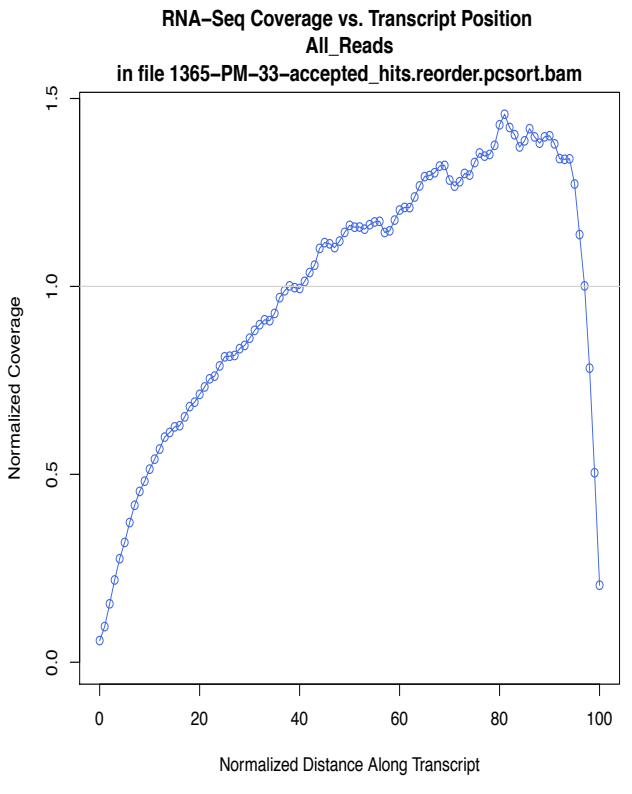
5’ bias
(strand oriented protocol)



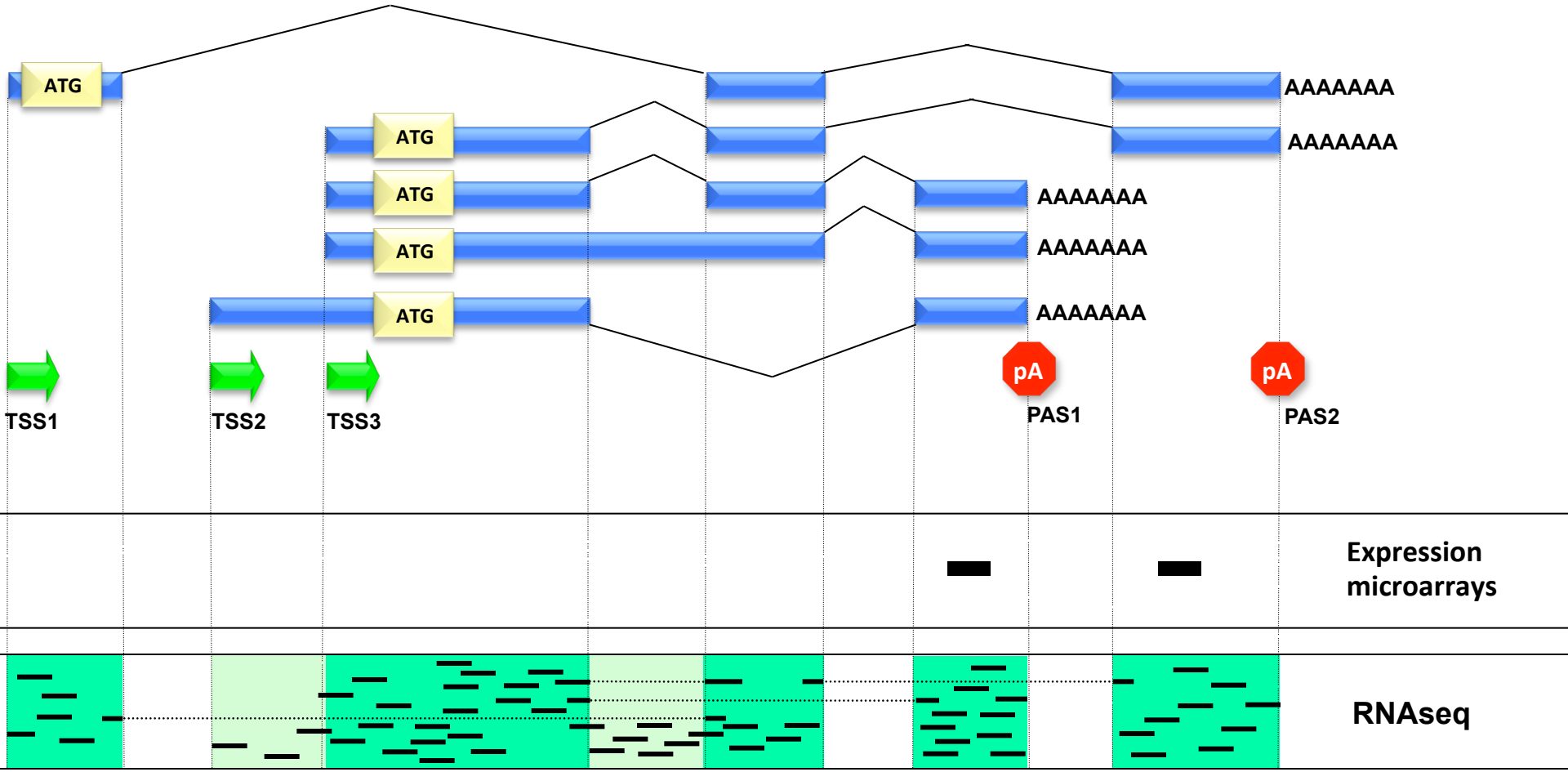
no bias
(low coverage at ends of transcript)



3’ bias
(poly-A selection)

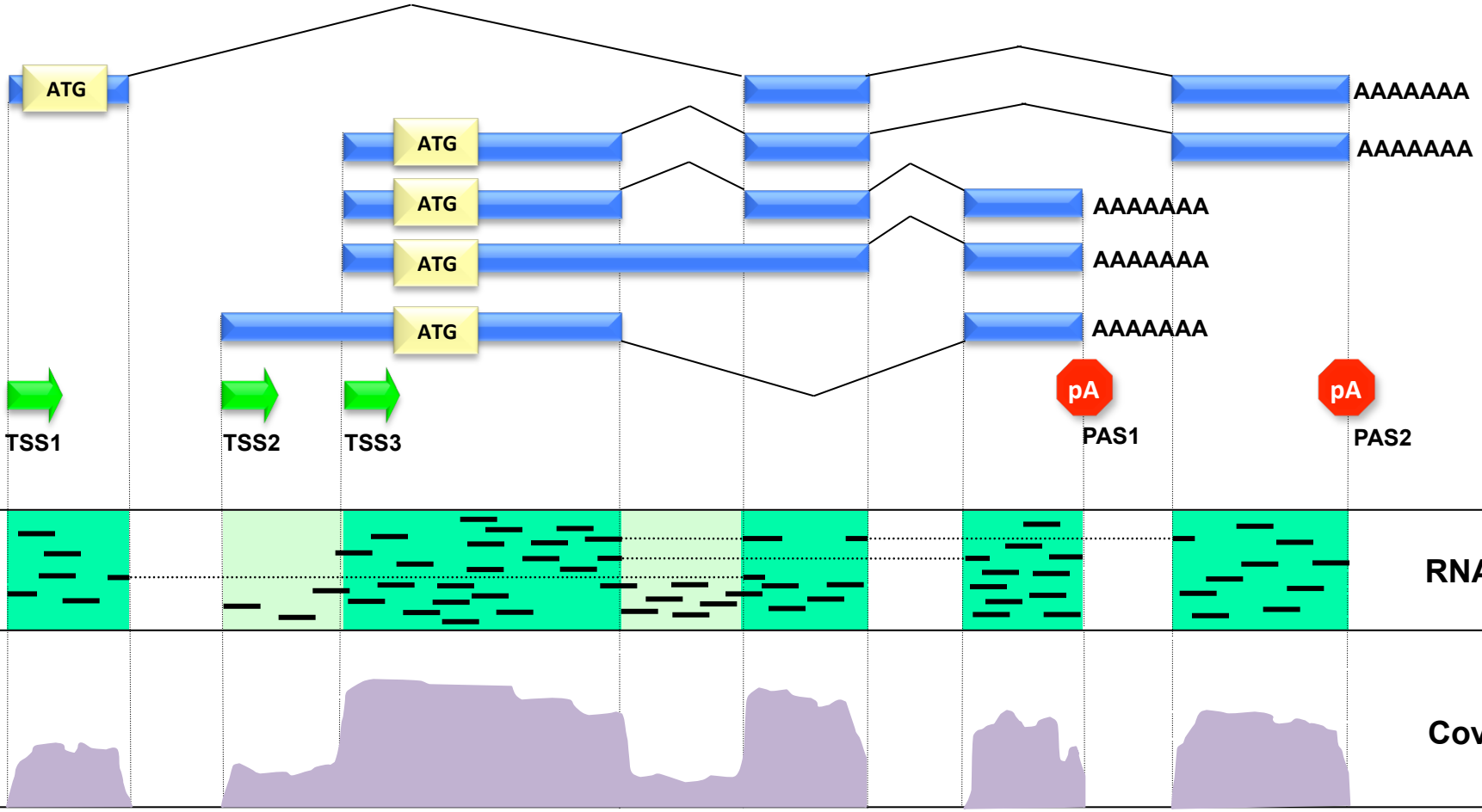


Measuring RNA expression



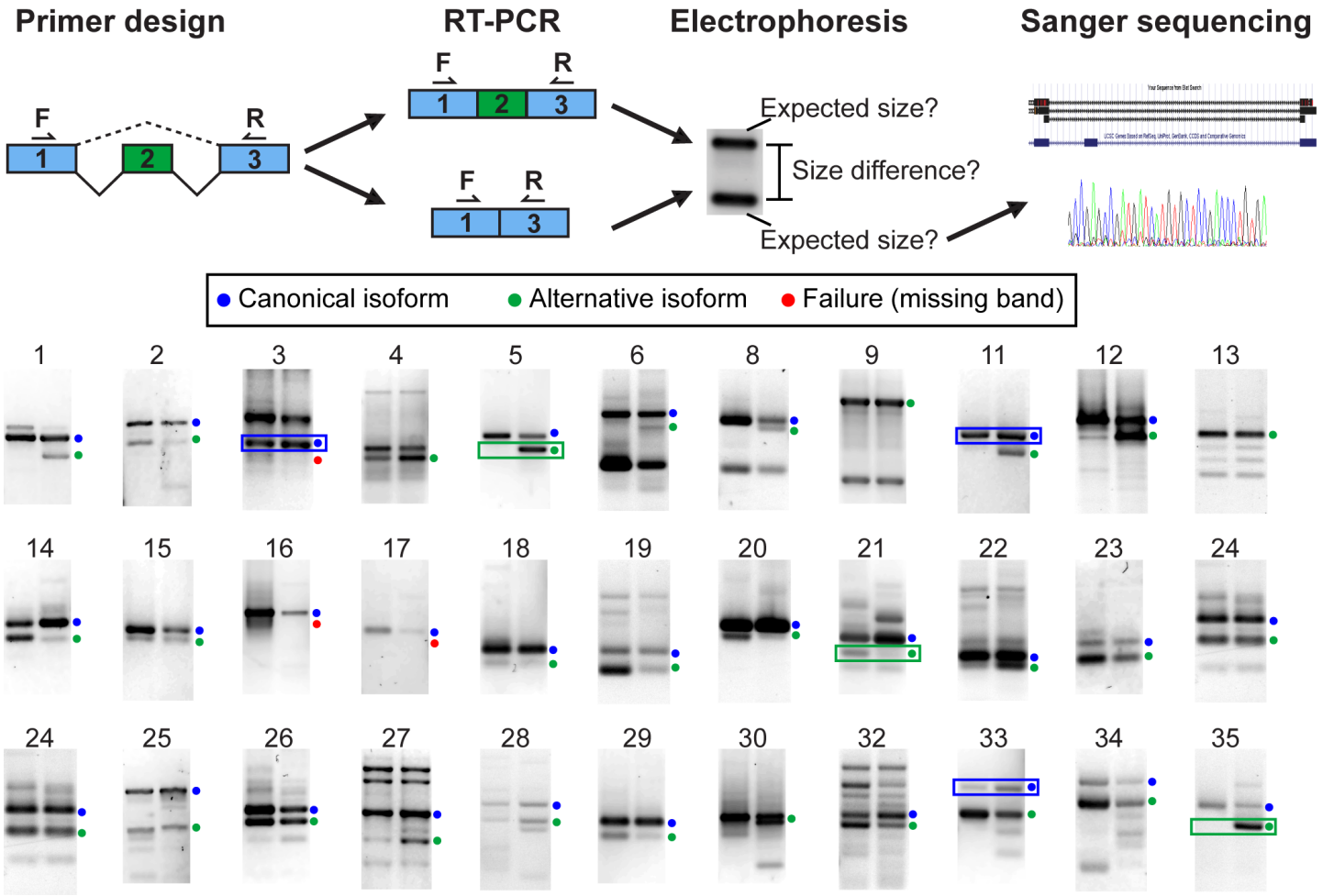
Gabe Kolle

Measuring RNA expression



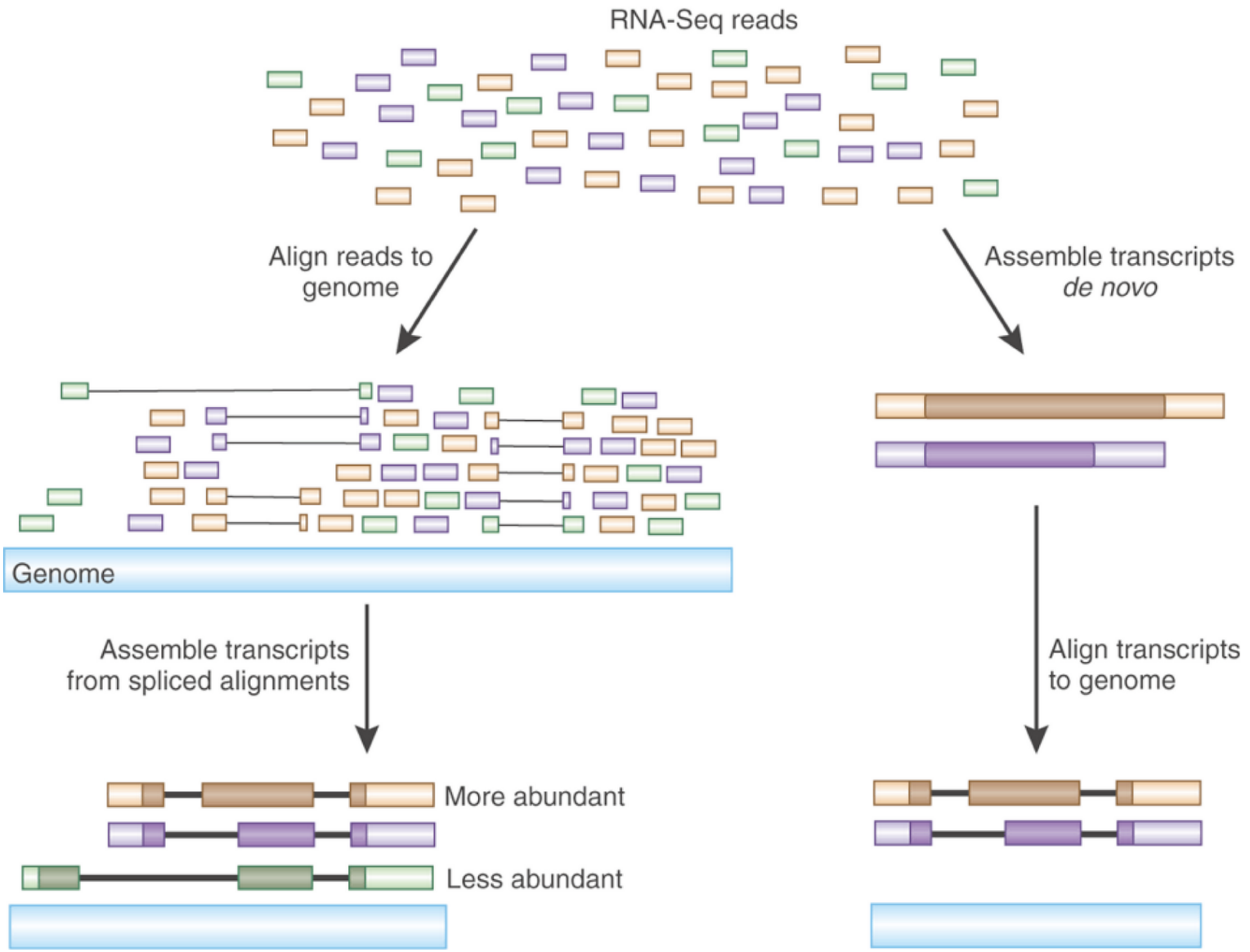
Gabe Kolle

Validation



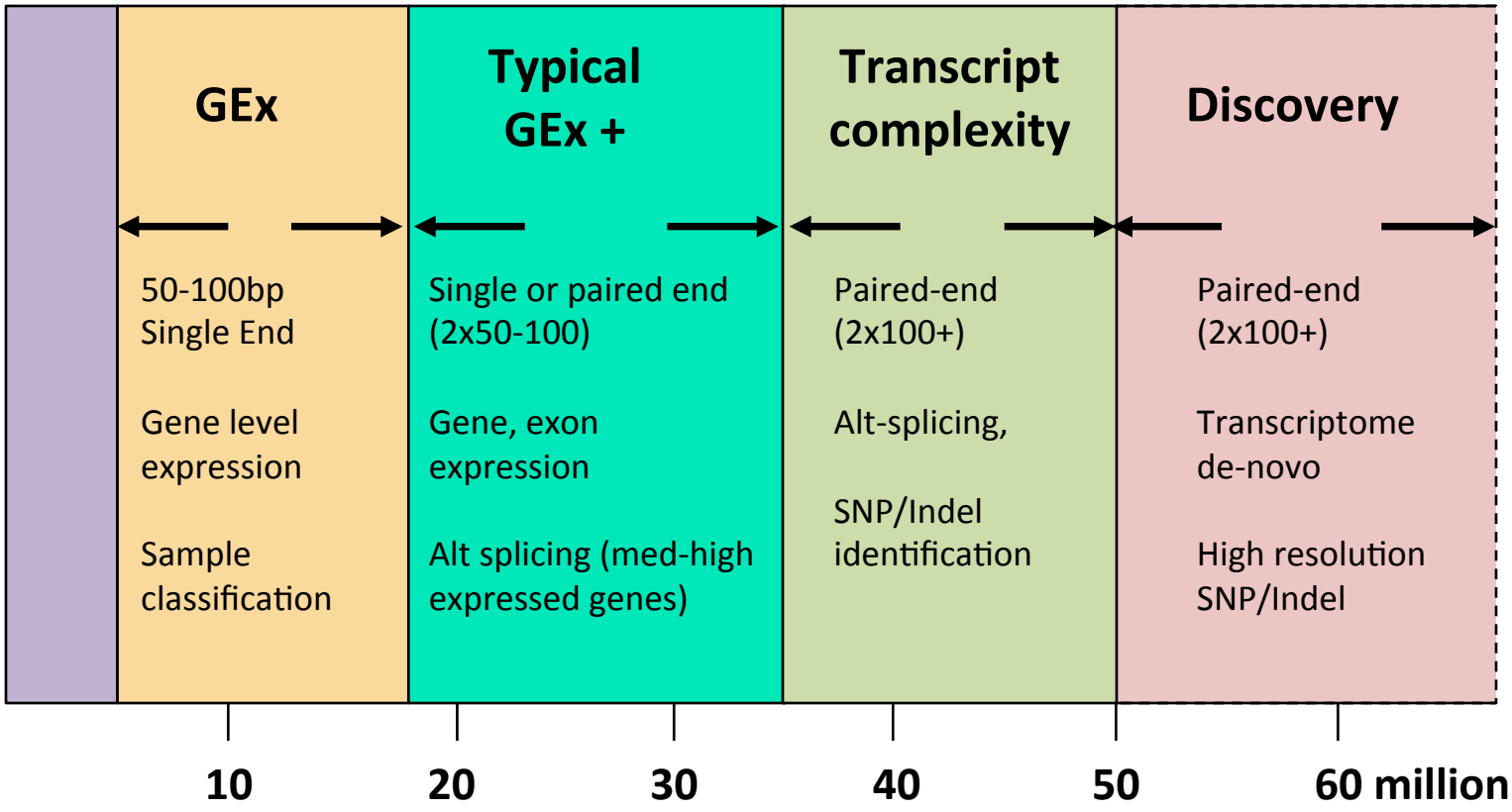
33 of 192 assays shown. Overall validation rate = 85%

Map or assemble?



Nature Biotechnology **28**, 421–423 (2010)

Sequence depth and type

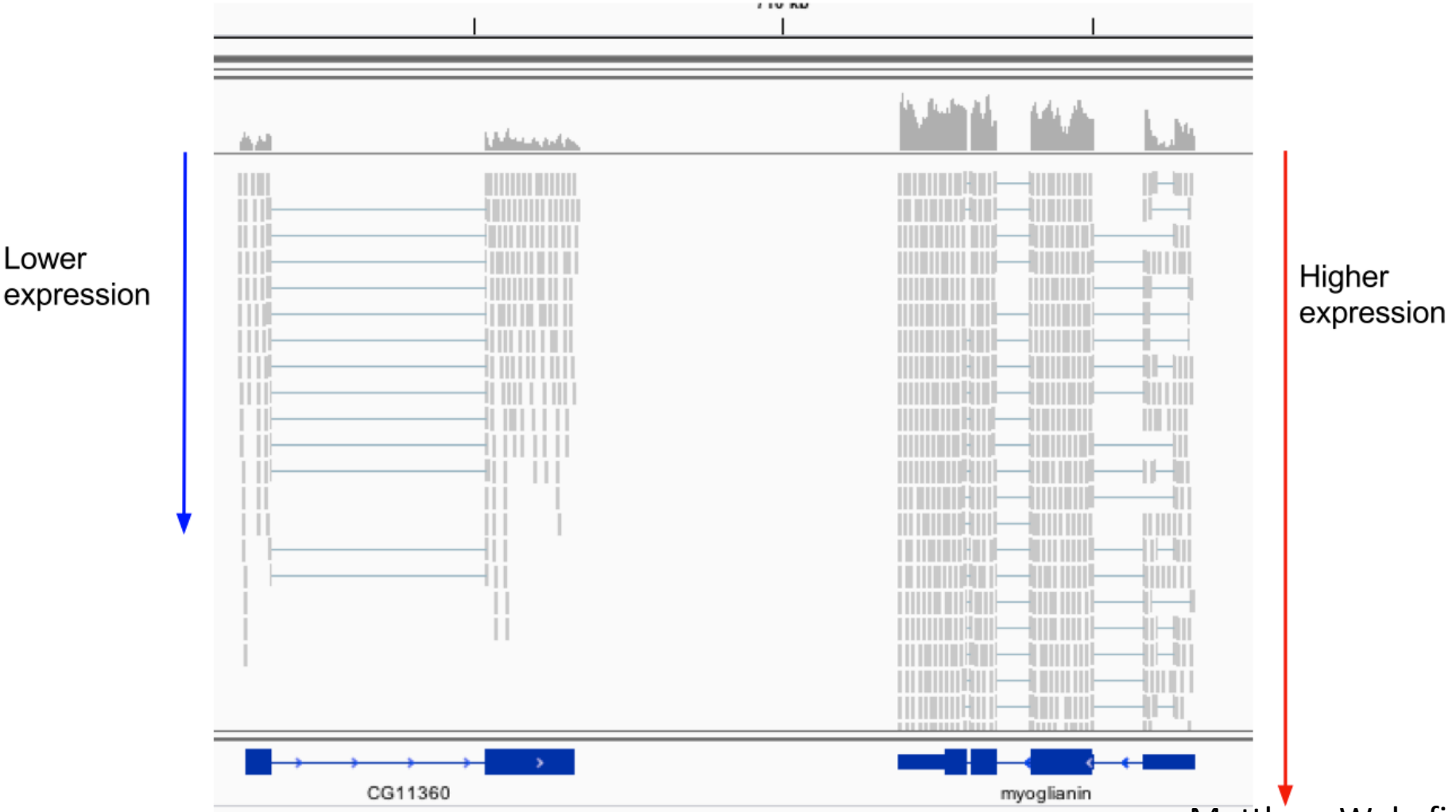


- Level of sensitivity is proportional to the number of reads
- Longer reads/paired end reads result in better mapping and resolution of splicing

Gabe Kolle

Gene expression from mRNA

RNA-seq measures transcripts, so reads should align to exons.



Matthew Wakefield

Differential Expression Visualization (using IGV)

Quantifying gene expression
RPKM, FPKM

- **R**eads (**F**ragments) **P**er **K**ilobase of exon per **M**illion fragments mapped

- Standard way or comparing relative abundance of transcripts
- Normalises for:
  - Longer/shorter transcripts getting more/less reads
  - different experiments with varying data volume

- Major draw back is counting alternative isoforms
  - Often ambiguous which isoform a read belongs to

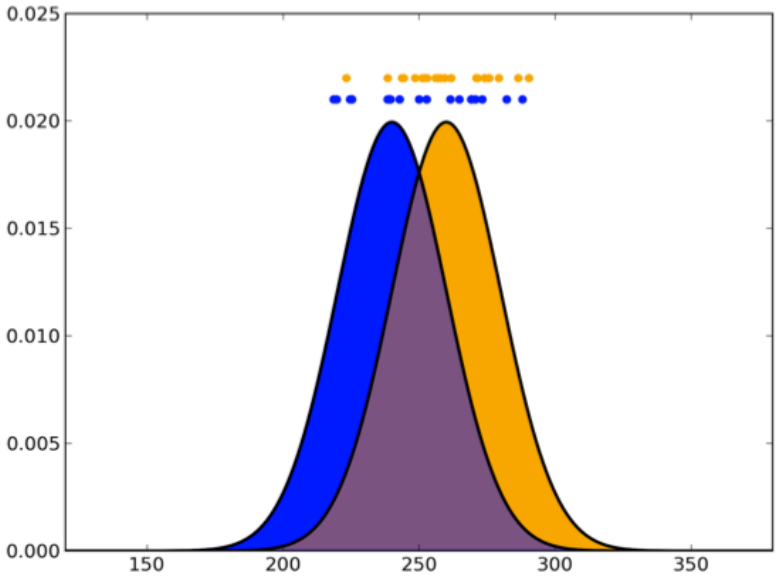
Difference of expression means

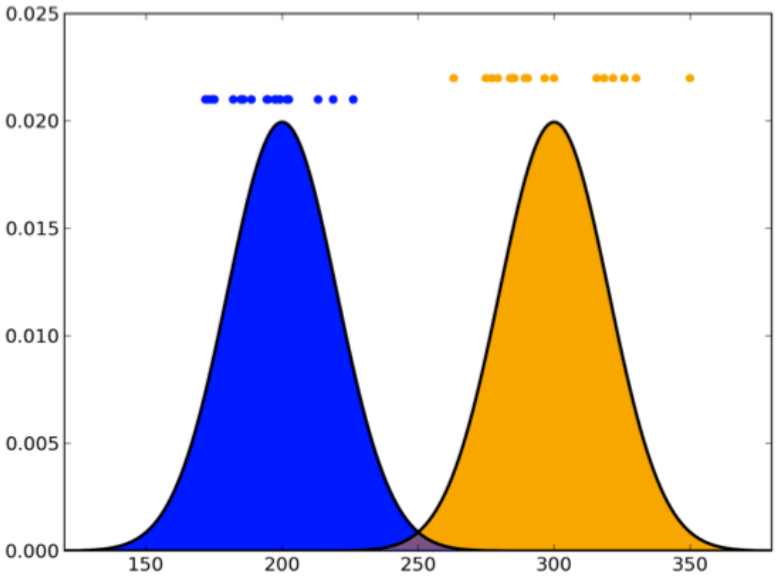
Let's assume a normal distribution.

We can be more confident there are two separate groups if:

- the means of the groups are far apart
- the variances of the groups are small

Variance is smaller if we have many samples!





Matthew Wakefield

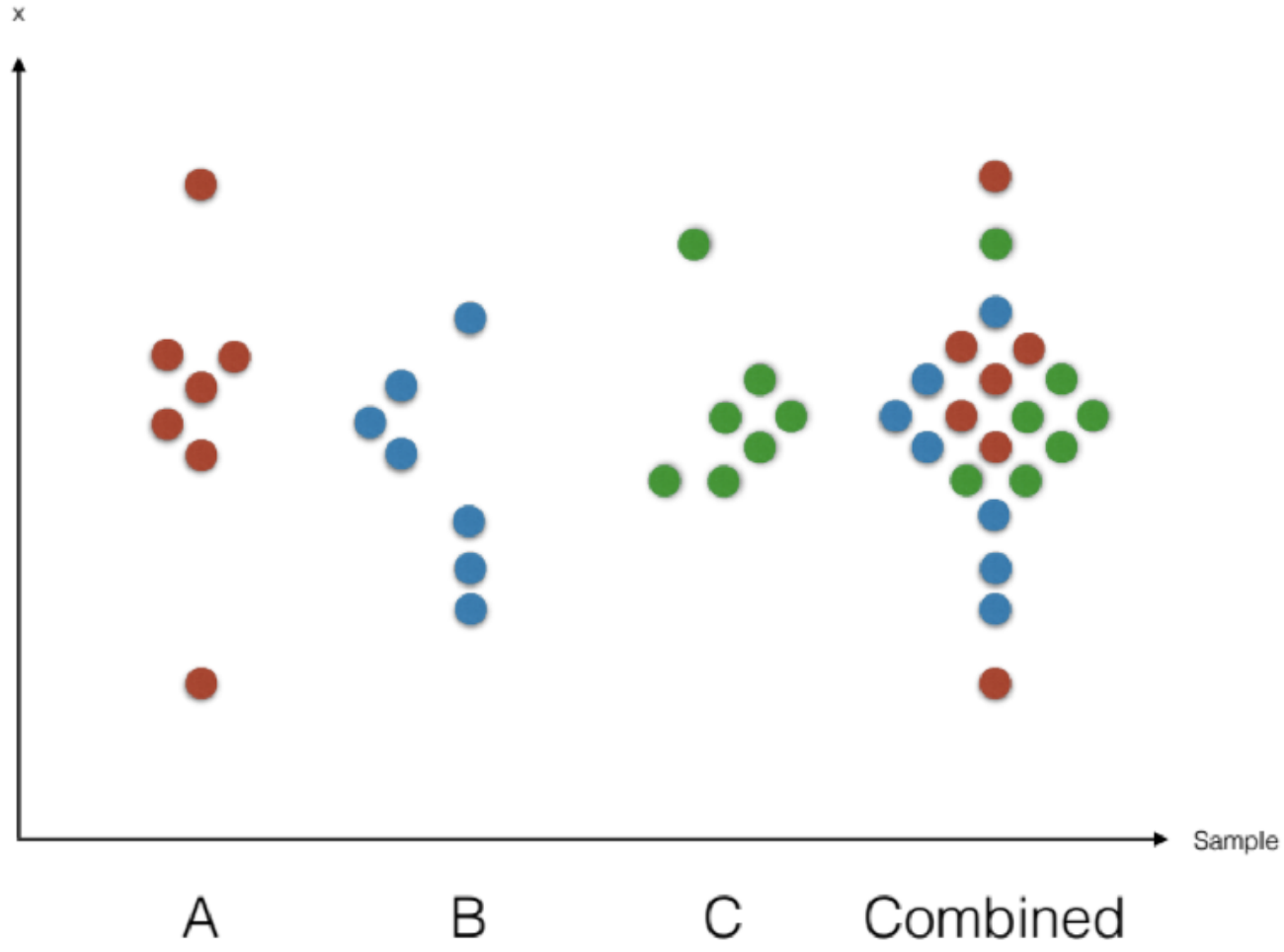
A simple t-test?

- How about a students t-test for each gene?
  - Often have few replicates
  - Distribution is not normal
  - Multiple testing issues

edgeR

- Uses a negative binomial distribution
- modified statistics
- corrects for multiple testing
- Variance estimates “borrow information” from other samples and genes
- *“Implements a range of statistical methodologies based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests.”*

Borrowing Information



Matthew Wakefield

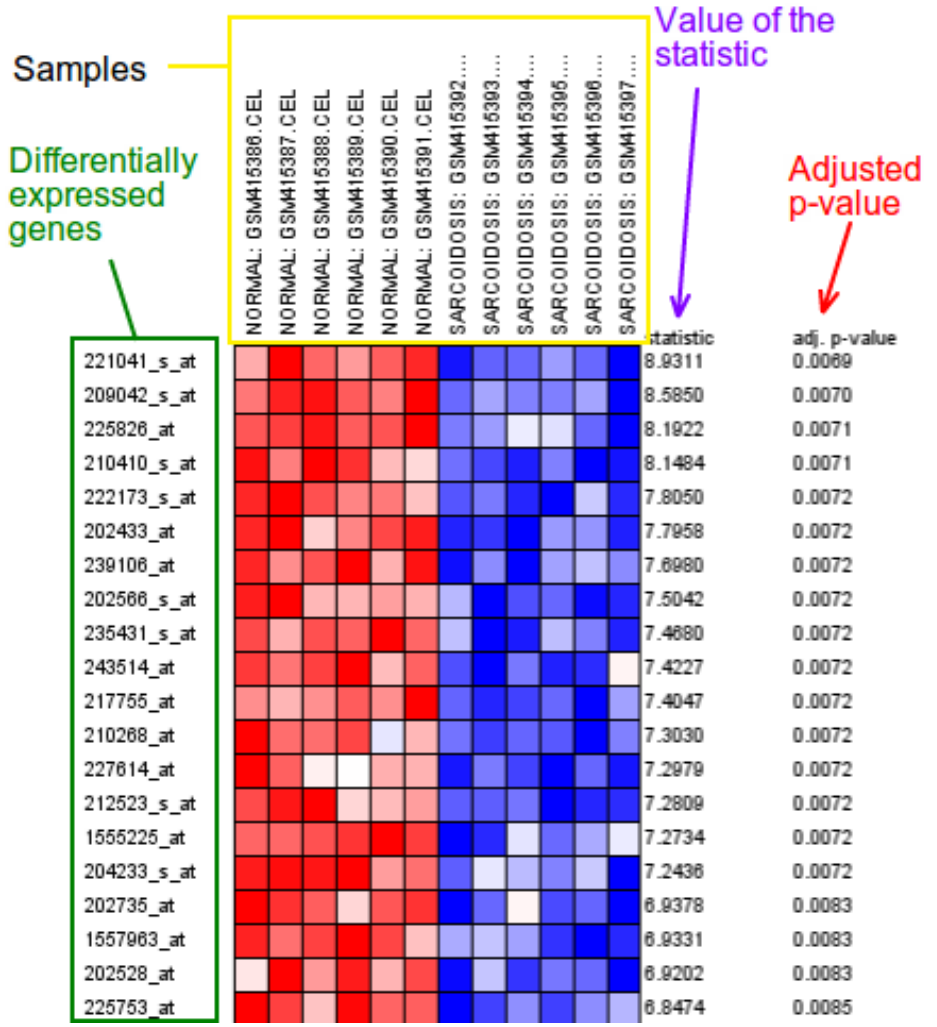
Voom/Limma

- Limma
  - Originally designed for microarray analysis
  - uses linear models
  - Voom function deals with the read counts

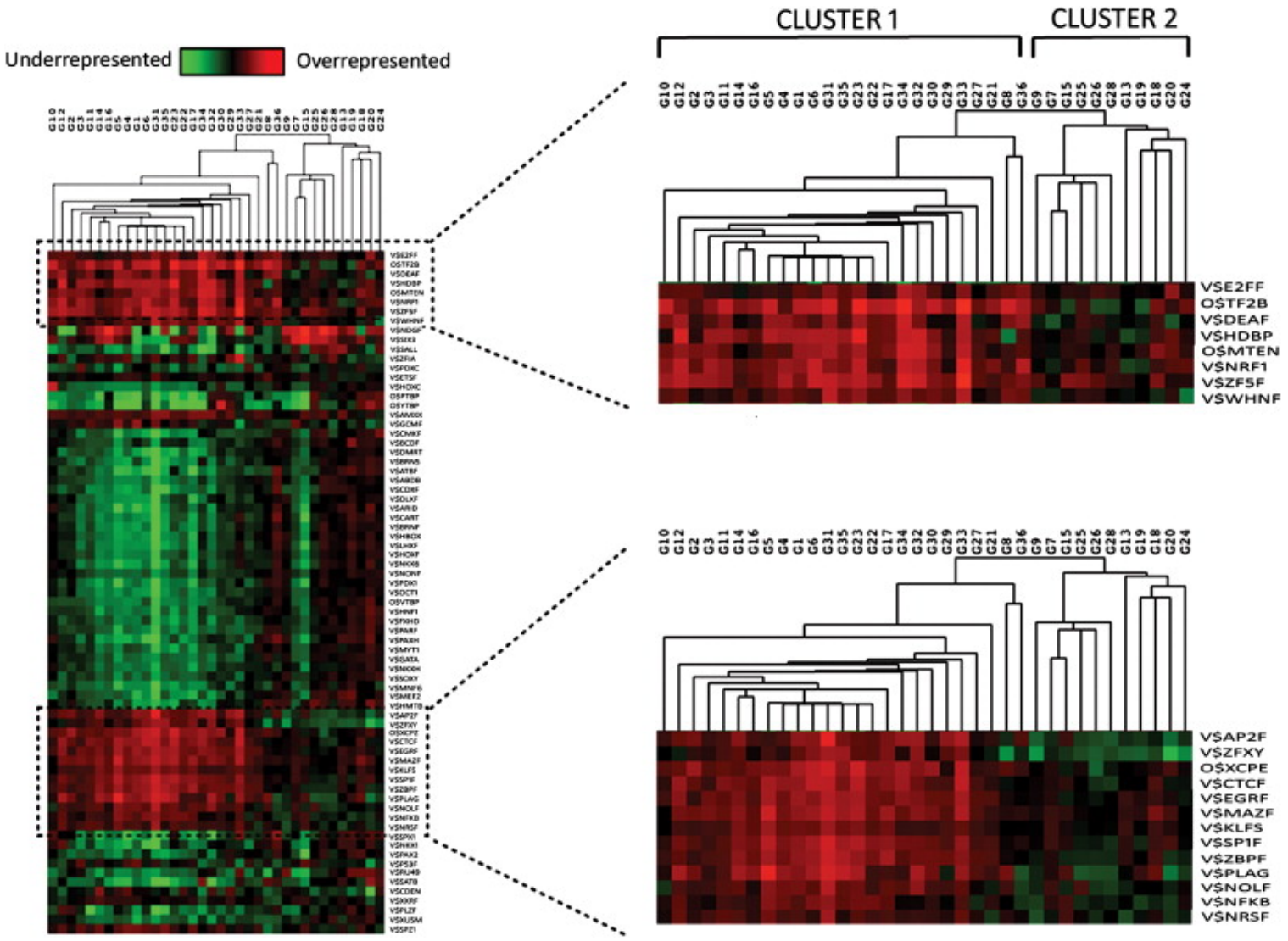
- Generally conservative method

Differential Expression Analysis

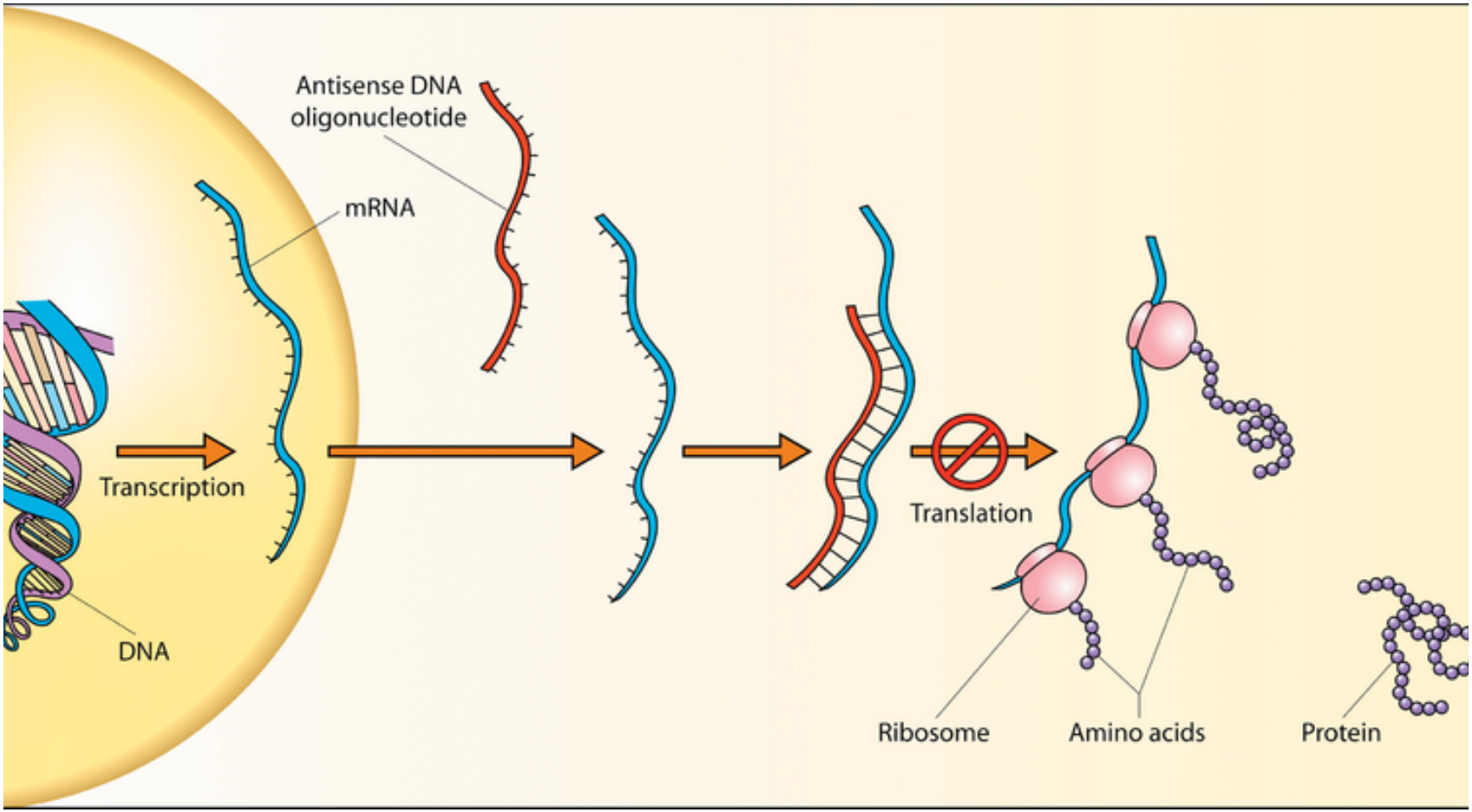
- Requires complex statistics to take into account variance of values
- Often done using the *R* statistics package



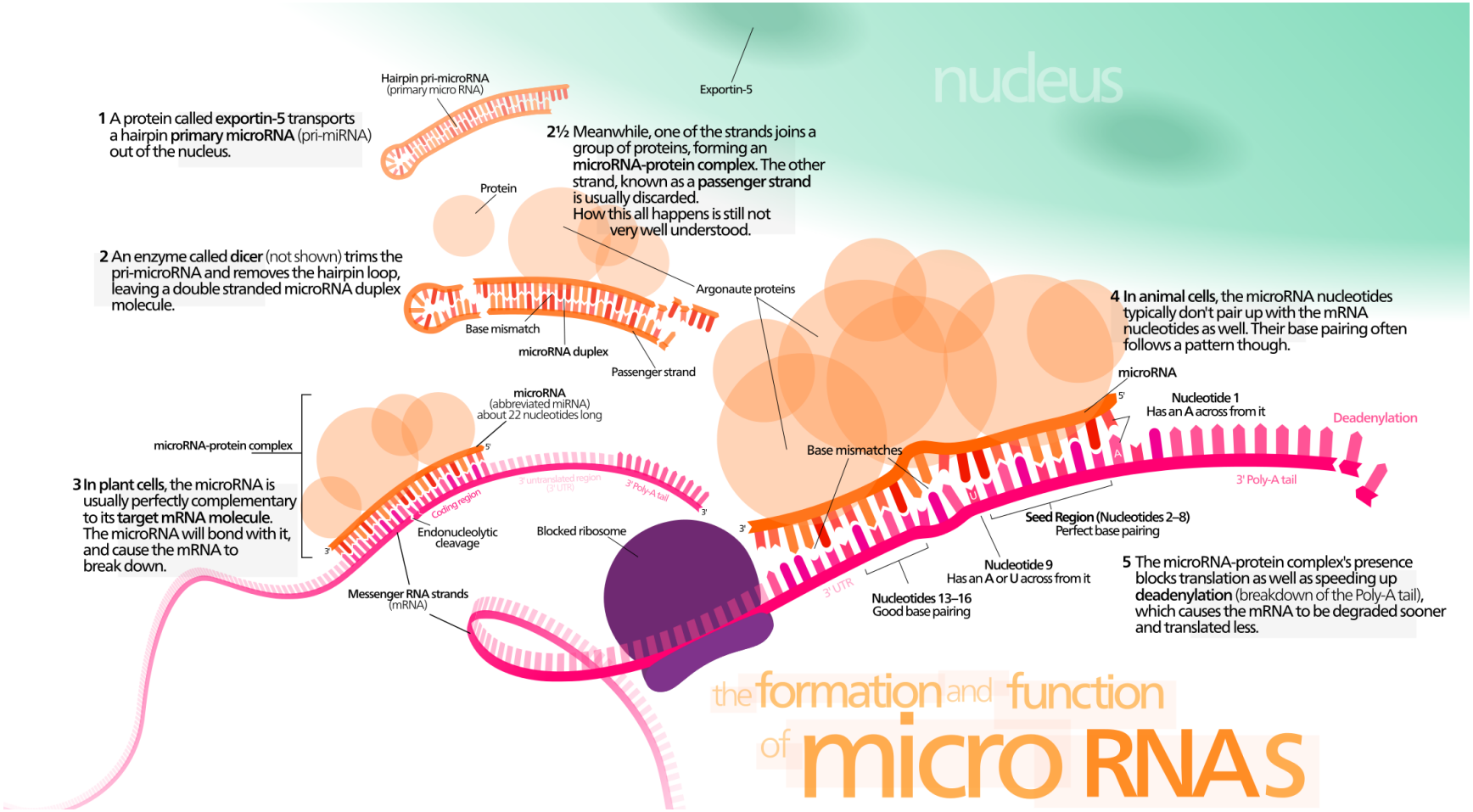
Differential expression heatmap



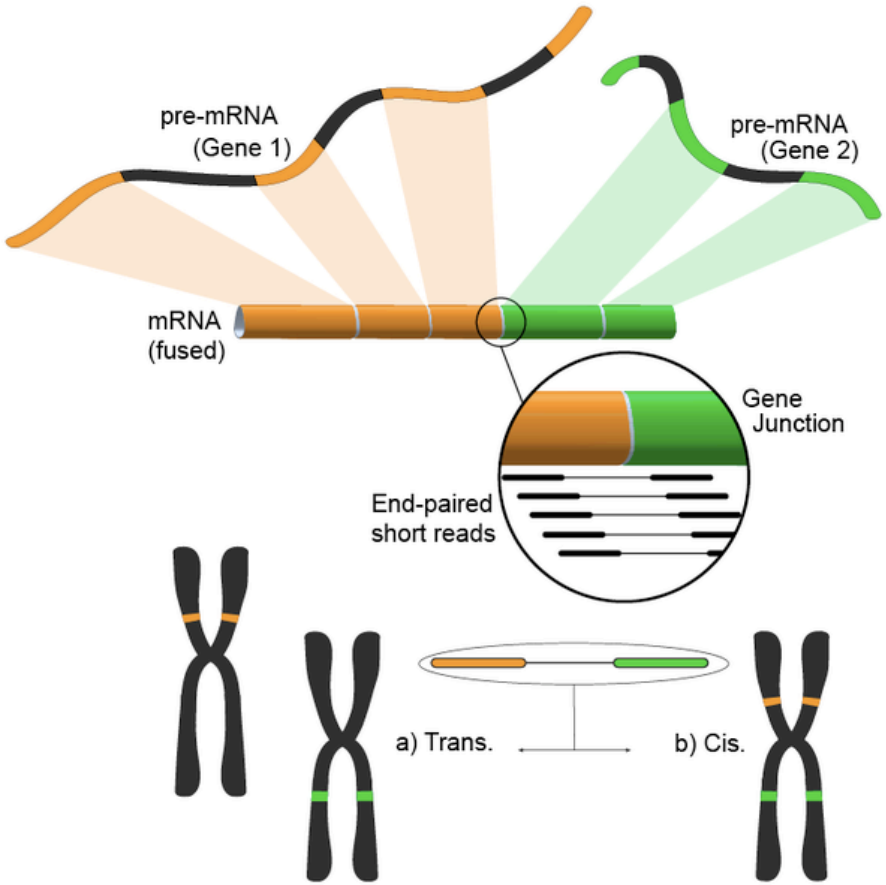
Antisense transcripts



miRNAs – RNA regulation



Fusion Gene detection



Variant (SNP) Detection

- One of the promising aspects of RNA-seq is that it produces actual sequences of mRNA molecules
- Can search for mutations in coding regions
- **However**: coverage is uneven, high background error rate, high level of PCR duplication, possible RNA editing
- Unknown allele frequency
- Possible allele-specific expression

miRNAseq and RNAseq integration

- miRNA affect RNA levels
- Can correlate miRNA and RNA levels in related experiments to decode interactions

Experimental Design

- What is your biological question?
- How many reads/depth of coverage do you need?
- Do you need biological / technical replicates?
- What is the best platform / technology for this application?

RNAseq Replicates

- Biological Replicates essential for Differential Expression (DE)
- Biological Replicates essential for DE
- Biological Replicates essential for DE

- Technical replicates not nearly as important

Why triplicate?

	A			B	
3500	4000	4500	3000	5000	17000
Ave	4000		Ave	8333	
3500	4000	4500	8400	8000	8600
Ave	4000		Ave	8333	

- What are your replicates?
- Cell lines, tissues, same animal?

Essential questions to ask:

- What is the biological question?
- Can RNAseq answer it?
- If so, what type(s) of sequencing are required?
- Is it a discovery and/or quantitative DE study
- Do I have the replicates required?
- What library prep is needed?
- What read length is required?
- Is paired end sequencing required?
- What sequencing depth is required?
- How many samples per lane & how many lanes?

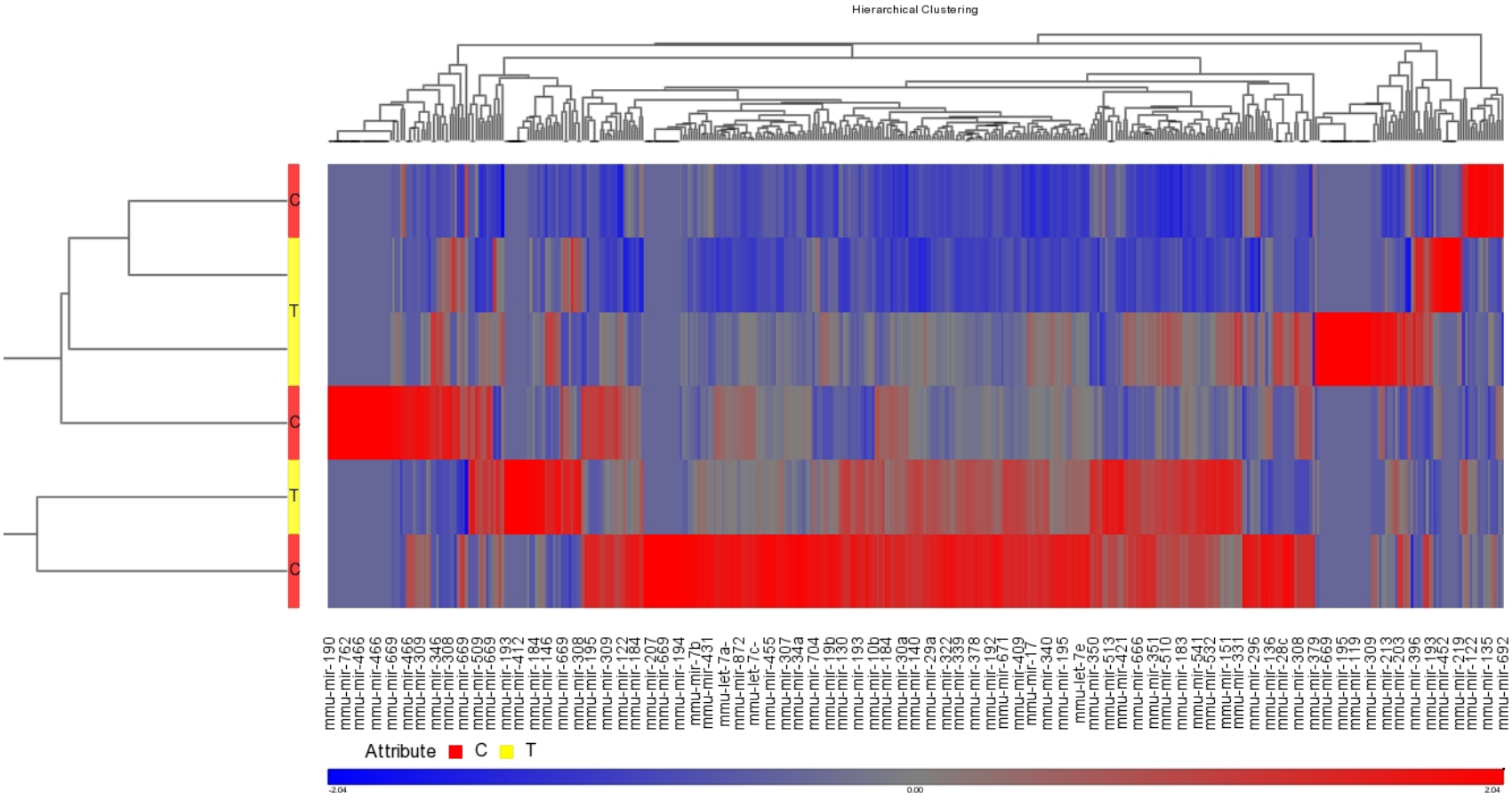
Example 1 – VERY simple

- What pathways are upregulated when mouse is treated with XXX
  - What tissues, will it give you the appropriate readout, what timepoint(s)?
  - Simple gene count required
  - Can do SE 100 bp PolyA
  - Triplicate in (matched) mice
    - Need to minimise biological variation
  - 8-9 samples per lane (1 lane=200 million reads)
    - Over 20 million reads per sample

Example 2 – VERY complicated

- Novel organism (no genome), what genes are upregulated in treatment XYZ.
- It is know that in mouse, microRNA, noncoding transcripts, including antisense and alternative splicing are all relevant
- Solution
  - 100-150 bp PE library for transcriptome assembly
    - Annotate assembly
  - (up to 300 PE if on MiSeq)
  - Stranded to see antisense transcripts
  - Total RNA with ribodepletion to sequence non-PolyA transcripts
  - Separate microRNA library prep
  - Then triplicate 100 bp PE for mapping counts back to assembled transcriptome

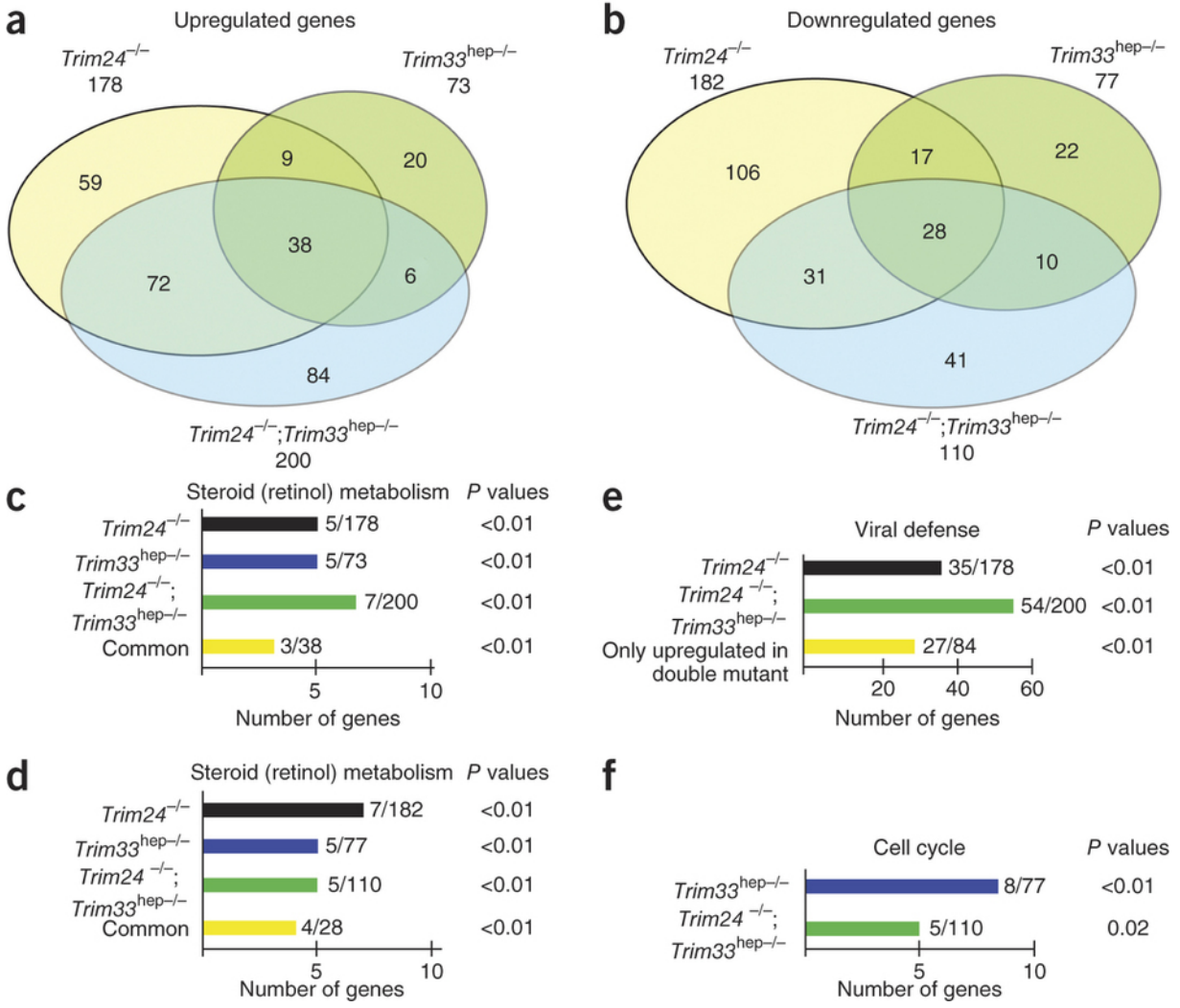
miRNA seq differential expression



http://www.modelingimmunity.org/data/2555-2/

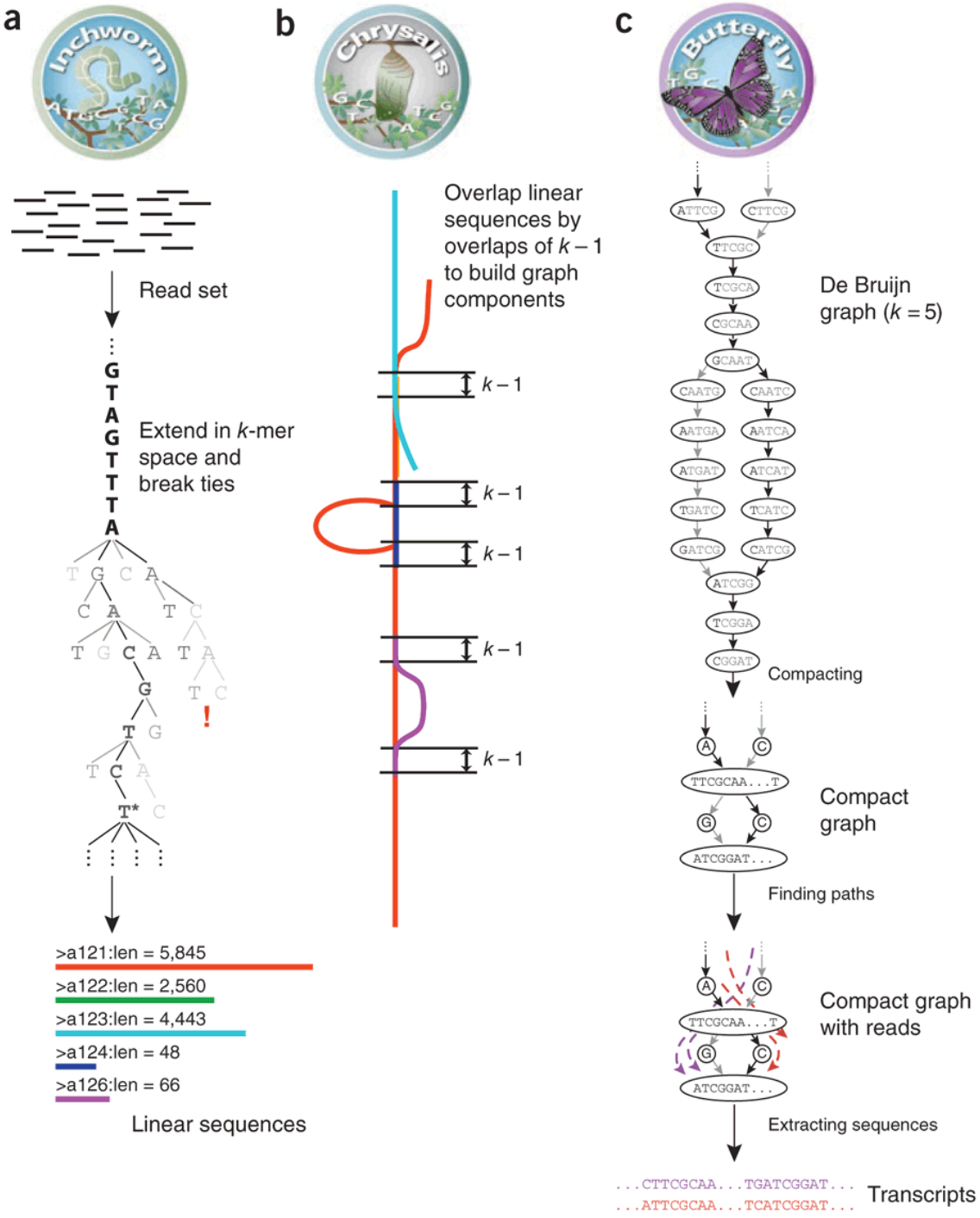
Real Example - 2013 Nature Paper

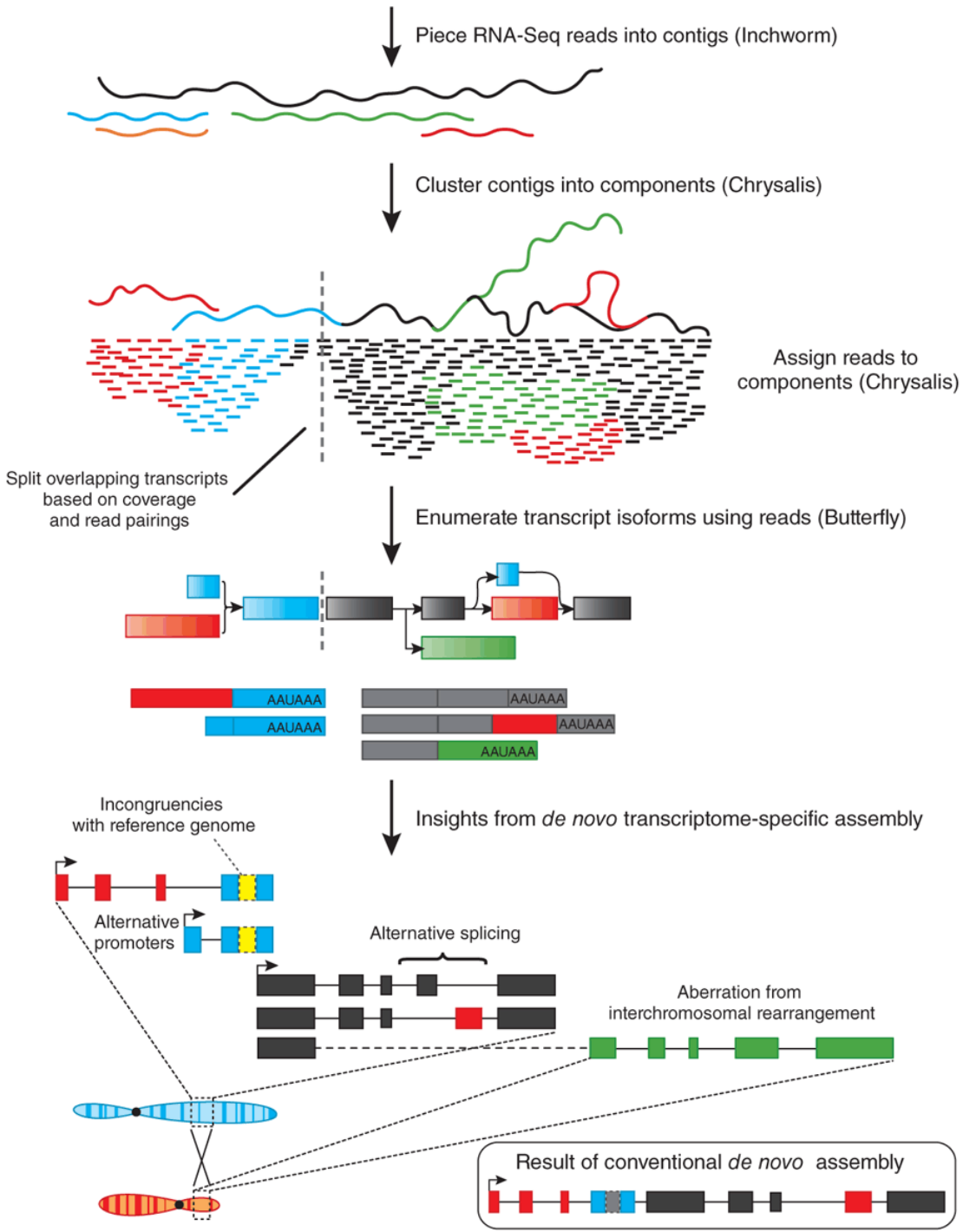
Trim24-repressed VL30 retrotransposons regulate gene expression by producing noncoding RNA



Nature Structural & Molecular Biology **20**, 339–346 (2013)

Trinity RNAseq assembly


a
Inchworm
Read set
G
T
A
G
T
T
T
A
Extend in k-mer space and break ties
>a121:len = 5,845
>a122:len = 2,560
>a123:len = 4,443
>a124:len = 48
>a126:len = 66
Linear sequences
b
Chrysalis
Overlap linear sequences by overlaps of k – 1 to build graph components
k – 1
k – 1
k – 1
k – 1
k – 1
c
Butterfly
ATTCG
CTTCG
TTCGC
TCGCA
CGCAA
GCAAT
CAATG
CAATC
AATGA
AATCA
ATGAT
ATCAT
TGATC
TCATC
GATCG
CATCG
ATCGG
TCGGA
CGGAT
De Bruijn graph (k = 5)
Compacting
TTCGCAA...T
ATCGGAT...
Compact graph
Finding paths
Compact graph with reads
Extracting sequences
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...
Transcripts


Piece RNA-Seq reads into contigs (Inchworm)
Cluster contigs into components (Chrysalis)
Assign reads to components (Chrysalis)
Split overlapping transcripts based on coverage and read pairings
Enumerate transcript isoforms using reads (Butterfly)
AAUAAA
AAUAAA
AAUAAA
AAUAAA
AAUAAA
Insights from de novo transcriptome-specific assembly
Incongruencies with reference genome
Alternative promoters
Alternative splicing
Aberration from interchromosomal rearrangement
Result of conventional de novo assembly

Conclusions

- RNAseq
  - Revolutionising what is understood of transcription
  - Experimental Design very important
    - Understanding the limitations of your experiment
    - Understanding the context of your experiment
    - Leads to better down stream analysis and interpretation