RNA-Seq and Differential Gene Expression

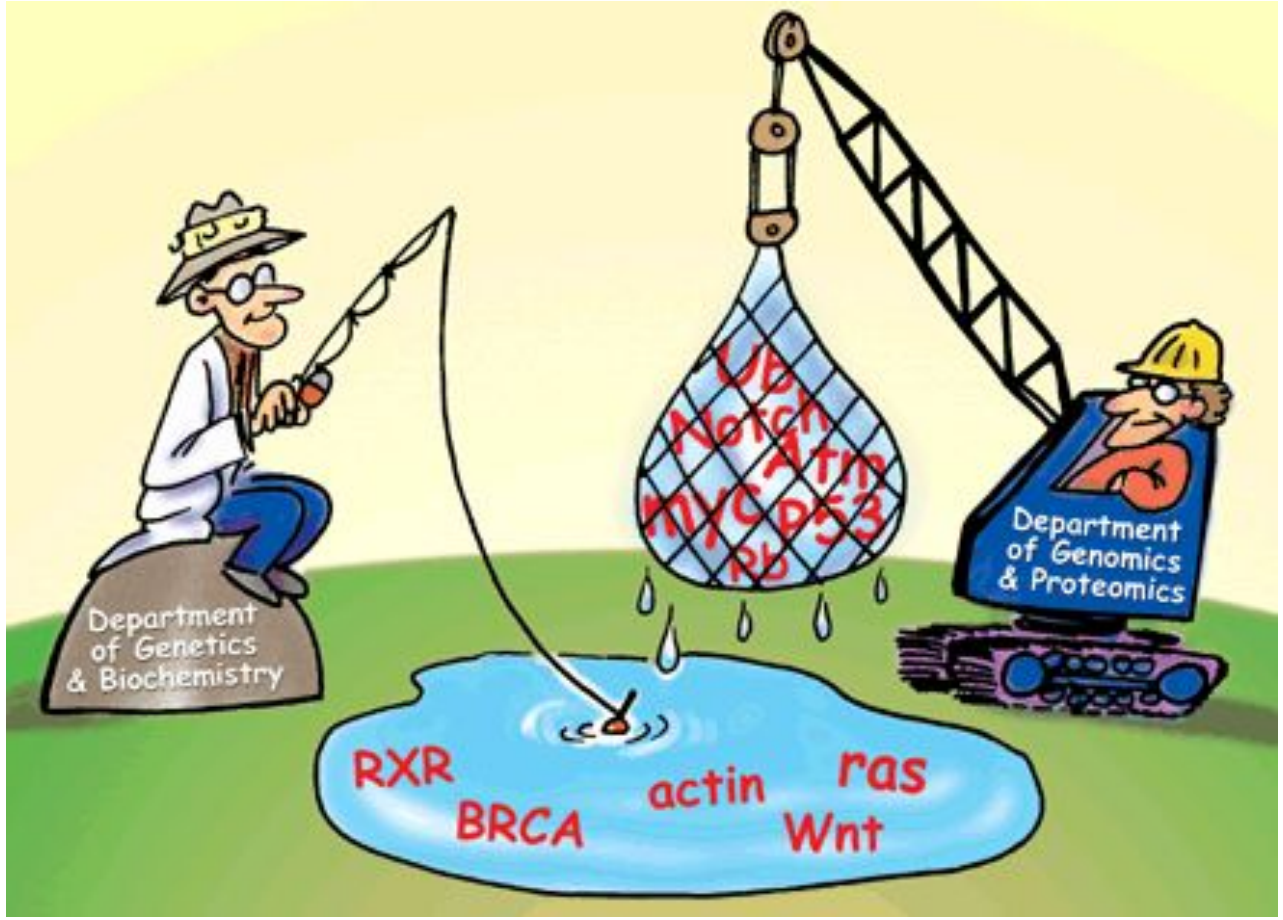
Callum J. Bell, Ph.D.

New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

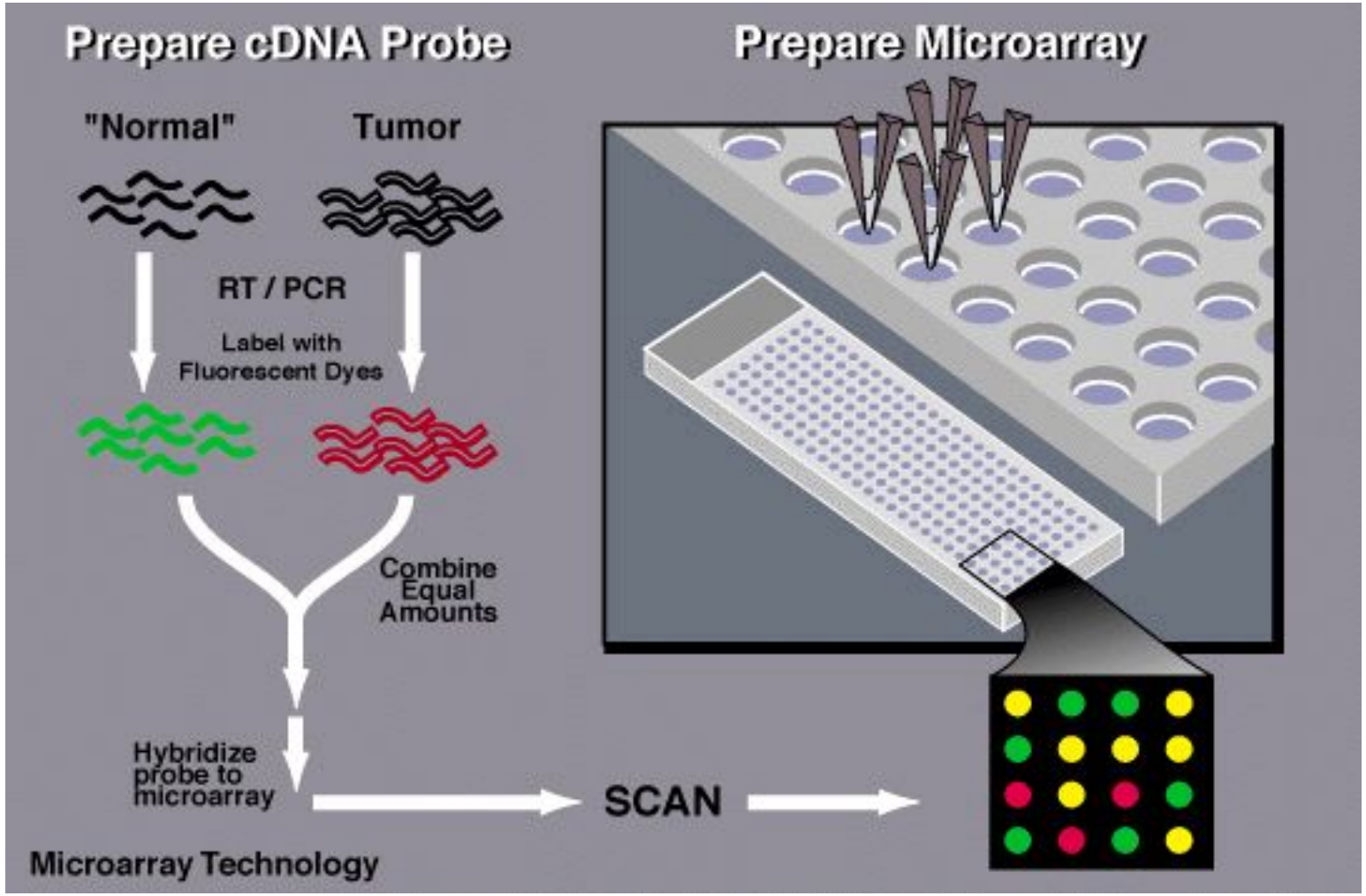
NCGR
National Center for Genome Resources

Methods for studying gene expression

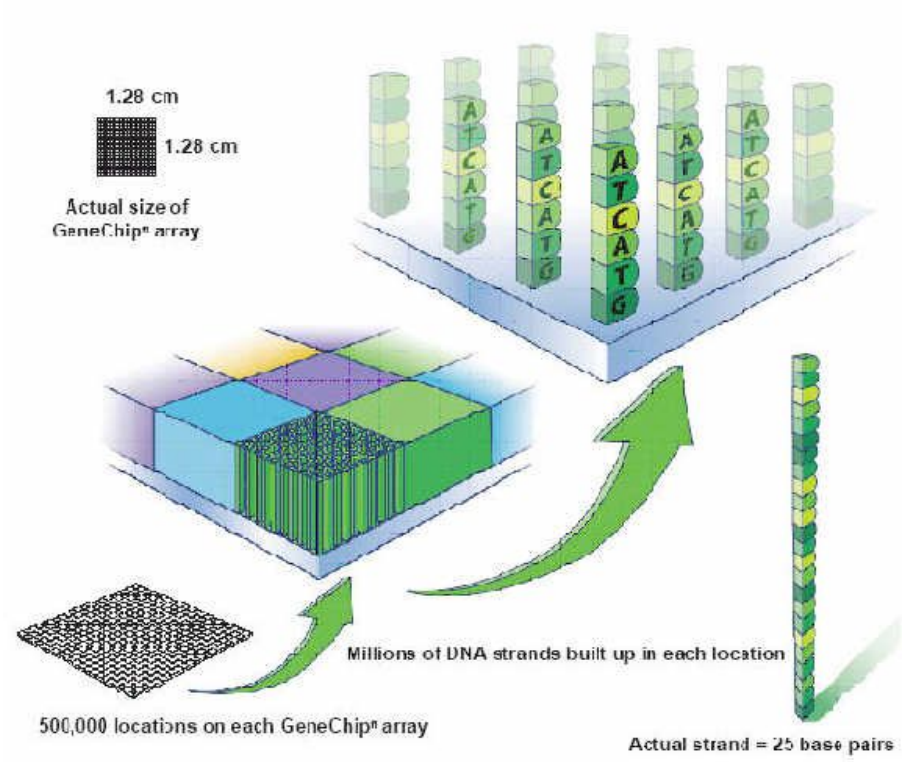
- Northern blot
- Tissue blots
- RT-PCR
- cDNA array blots
- Microarrays
- RNA-Seq

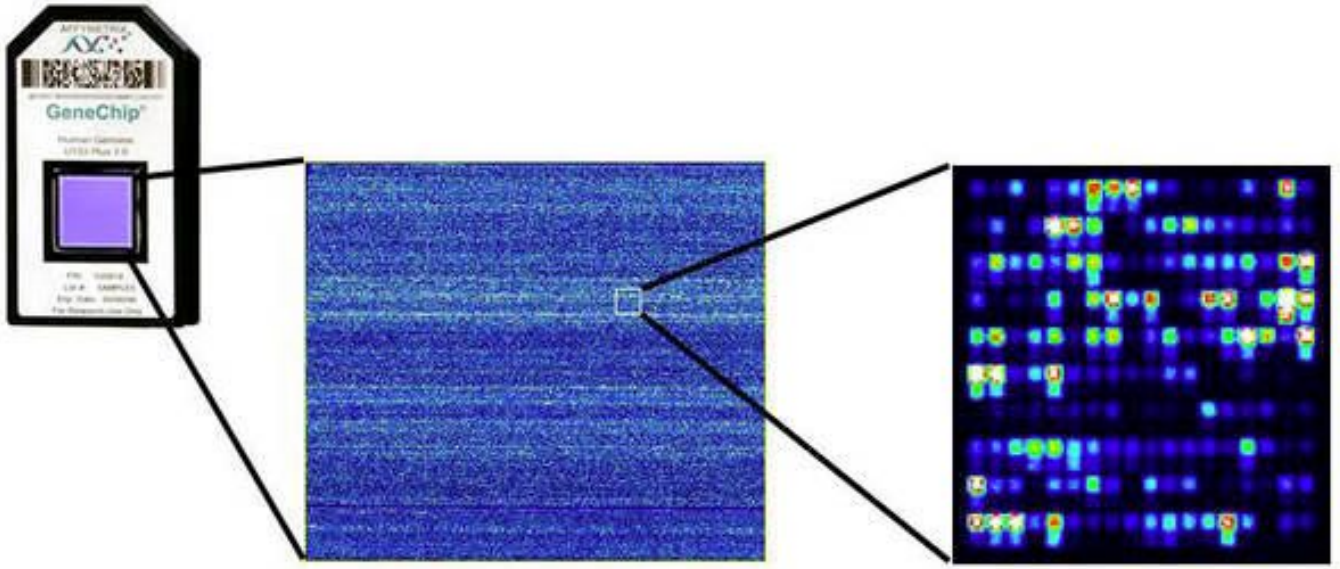


cDNA microarrays

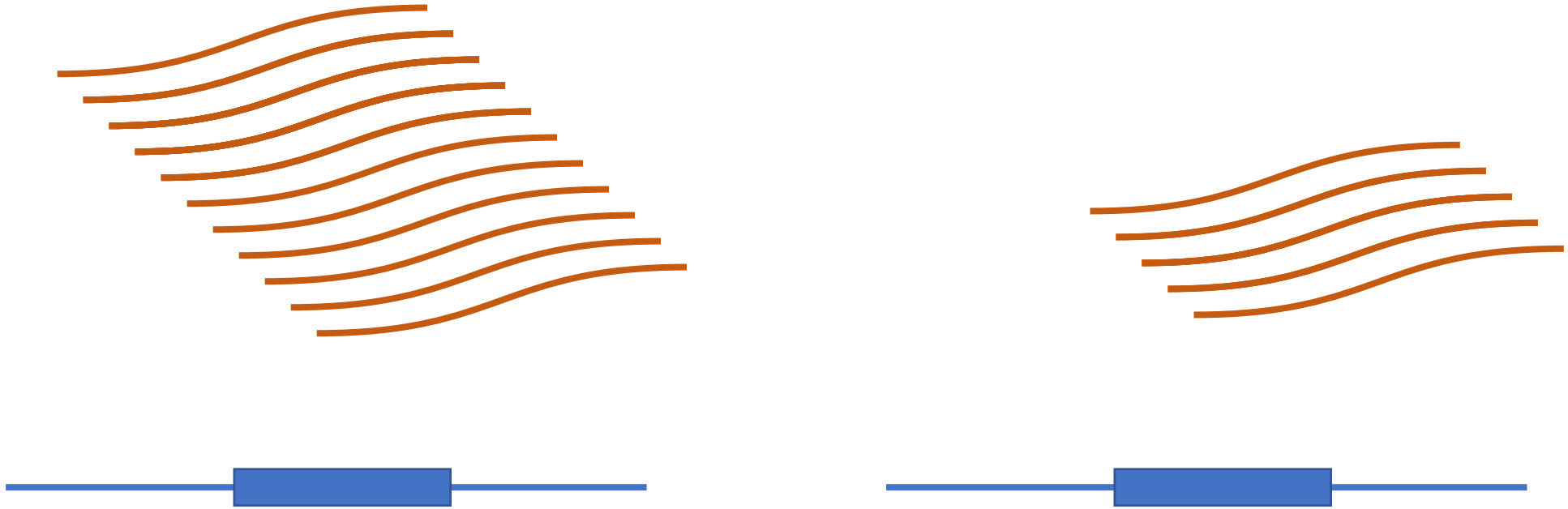


Affymetrix oligo arrays



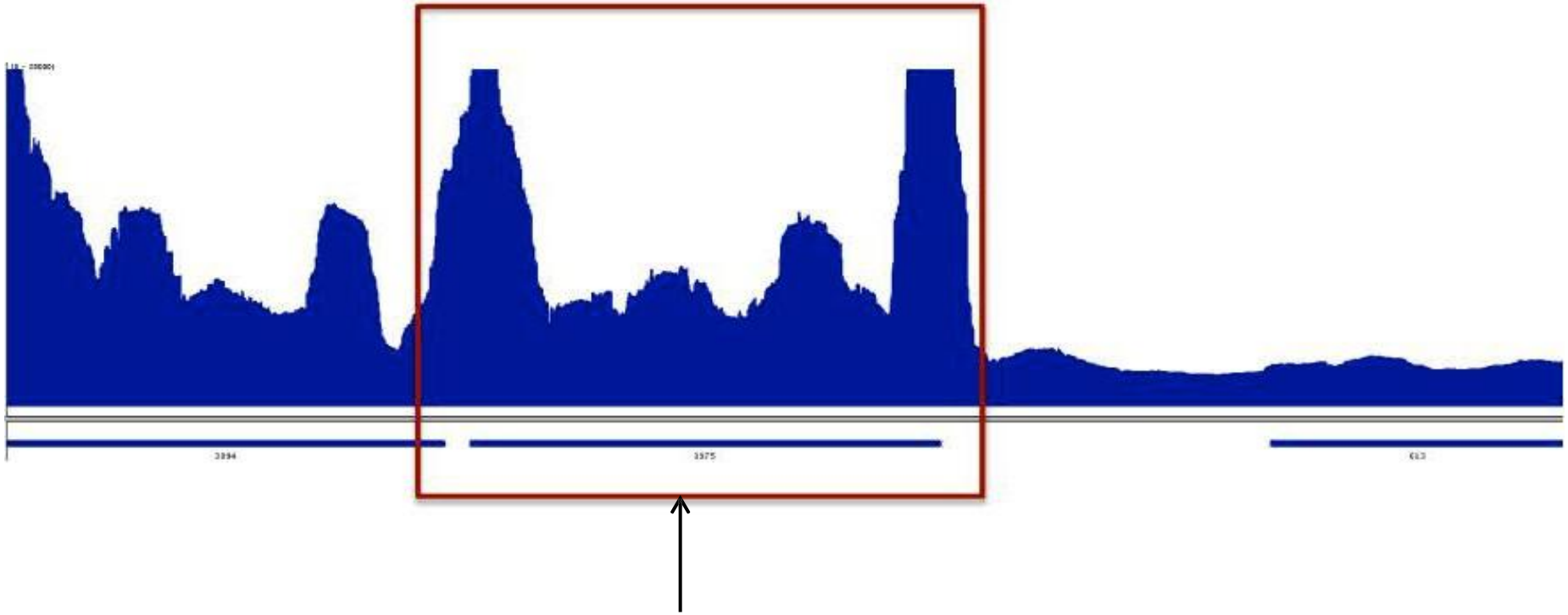


RNA-Seq



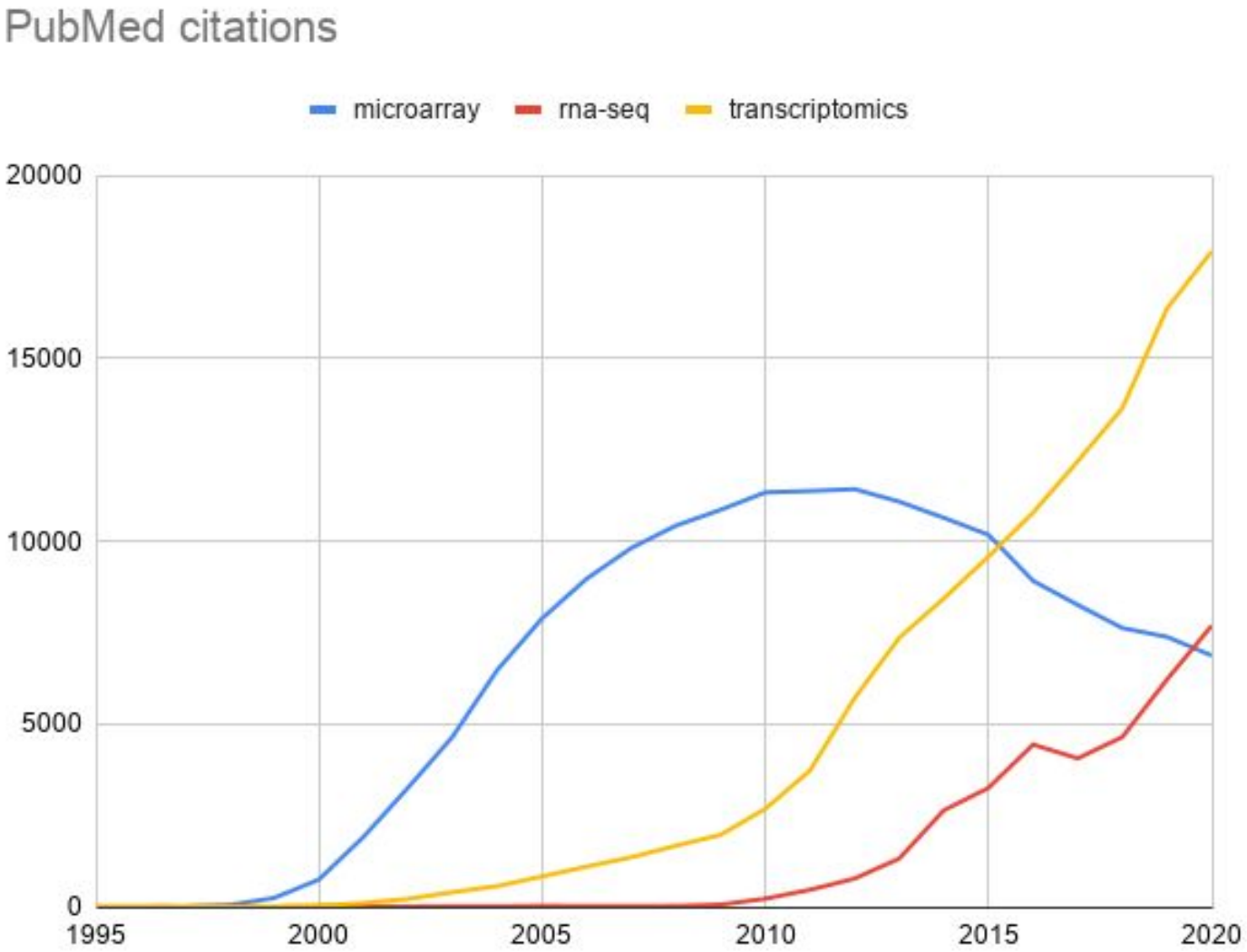
Quantitative gene expression measurement

The number of reads mapped to a gene is a quantification of its expression



algA in *P. aeruginosa*

PubMed citations



New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

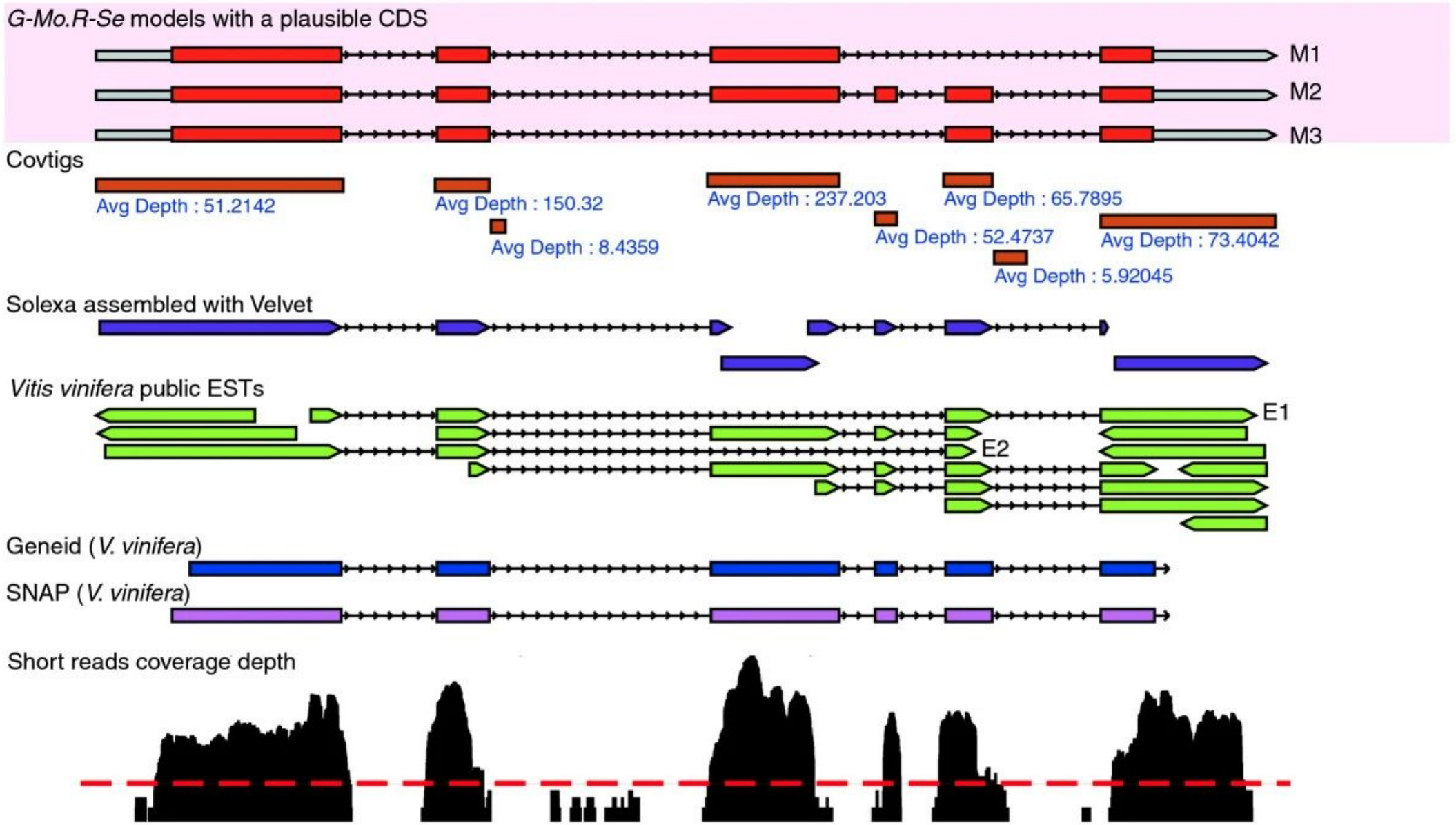
NCGR
National Center for Genome Resources

Early RNA-seq paper

Genome Biol. 2008;9(12):R175. doi: 10.1186/gb-2008-9-12-r175. Epub 2008 Dec 16.

Annotating genomes with massive-scale RNA sequencing

France Denoeud [1], Jean-Marc Aury, Corinne Da Silva, Benjamin Noel, Odile Rogier, Massimo Delledonne, Michele Morgante, Giorgio Valle, Patrick Wincker, Claude Scarpelli, Olivier Jaillon, François Artiguenave


G-Mo.R-Se models with a plausible CDS
M1
M2
M3
Covtigs
Avg Depth : 51.2142
Avg Depth : 150.32
Avg Depth : 8.4359
Avg Depth : 237.203
Avg Depth : 65.7895
Avg Depth : 52.4737
Avg Depth : 73.4042
Avg Depth : 5.92045
Solexa assembled with Velvet
Vitis vinifera public ESTs
E1
E2
Geneid (V. vinifera)
SNAP (V. vinifera)
Short reads coverage depth

RNA-Seq compared to microarrays

Microarrays

- Known genes only
- *A priori* knowledge required
- Usually not strand specific
- Hard to interrogate spliceforms
- No new sequence information
- Smaller dynamic range
- Rare transcripts challenging

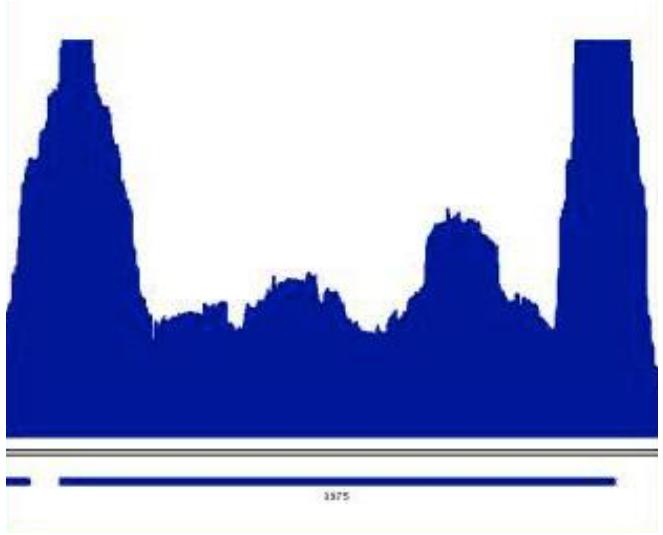
RNA-Seq

- Non-targeted detection
- No reference data required
- Can be strand specific
- Can explore alternative splicing
- Can detect novel variants
- Allele-specific expression
- Large dynamic range
- Detects rare transcripts

Why are microarrays persisting?

- Very good for targeted applications
- Particular disease focus area
- E.g. evaluate a panel of genes in a tumor
- Drug response
- Toxicology
- Regulated environments
- Sunk cost
- Reuse

Differential gene expression



Gene	Counts
FOXP2	10
PARK7	20
TNF	45
EGFR	2
IL6	50
AKT1	0

?

Gene	Counts
FOXP2	6
PARK7	14
TNF	65
EGFR	2
IL6	50
AKT1	6

Differential Gene Expression Overview

- Experimental design
- Sample preparation
- Sequencing
- Alignments
- Read counting
- (Normalization)
- DE analysis

Experimental Design

- Scientific question
- Sample limitations (can you get enough RNA)
- Variables (cartesian product grows quickly)
- (Controls, 6 timepoints, three treatments, three replicates...)
- Replicates (more is better)
- Missing data
- Optimize cost and scope
- http://scotty.genetics.utah.edu/

Influence of the organism

- Novel: little or no data
- Some data available: ESTs
- Non-model with a draft genome
- Model
  - Good reference genome available
  - Well annotated gene models
  - Transcriptome resources
  - Multiple genomes
  - Mutant lines

Influence of the organism

- Novel: little or no data
- Some data available: ESTs
- Non-model with a draft genome
- Model
  - Good reference genome available
  - Well annotated gene models
  - Transcriptome resources
  - Multiple genomes
  - Mutant lines
- These influence sequencing configuration and depth (and $)
- A well characterized organism requires fewer and shorter reads

Controlling cost

- Separate transcriptome assembly and DE runs
- Highest throughput platform
- Multiplex as much as safely possible (skew)
- Avoid unnecessary failure: QC before and after library prep
- Consider a pilot study, e.g. to optimize timepoints

Summary

Feature	Discovery/Assembly	DGE
Biological replicates	Helpful	Essential
Coverage/transcript	Important	Less important
Sequencing Depth	Maximize for rare transcripts	Sufficient for robust statistics
Duplex Specific Nuclease	Can help with rare transcripts	Never
Stranded library	Ideally	Not necessary with annotated reference
Long reads	Ideally	Not necessary with annotated reference
Paired end reads	Ideally	Less important

Sequencing (really part of design)

- Library protocol
- rRNA depletion method
- Strand preservation
- Paired or single end
- Read length
- Depth
  - “normal” RNA-Seq: 20 M reads/sample
  - Alternative splicing: 40-60 M reads per sample
  - Allele-specific expression 60-100 M
  - Low abundance transcripts: 60-100 M

Platform and multiplexing

	platform	read config	output
Production scale	HiSeq 2500	2 x 250	180 Gb - 1 Tb
	HiSeq 4000	2 x 150	1.5 Tb
	HiSeq X	2 x 150	1.8 Tb
	NovaSeq	2 x 250	6 Tb
benchtop	NextSeq	2 x 150	120 Gb
	MiSeq	2 x 300	15 Gb
	Iseq	2 x 150	1.2 Gb
	MiniSeq	2 x 150	7.5 Gb

Library preparation steps (1)

Target enrichment

- mRNA (and lncRNA)
  - Poly-A \+ fraction selection
  - Ribosomal RNA depletion
- Non-coding RNA (miRNA, etc.)
  - Size selection

Library preparation steps (2)

RNA fragmentation

- Fragmentation improves coverage on short read platforms
- Enzymatic, heat, divalent cations, ultrasound
- Aim for a population centered around 200 bp

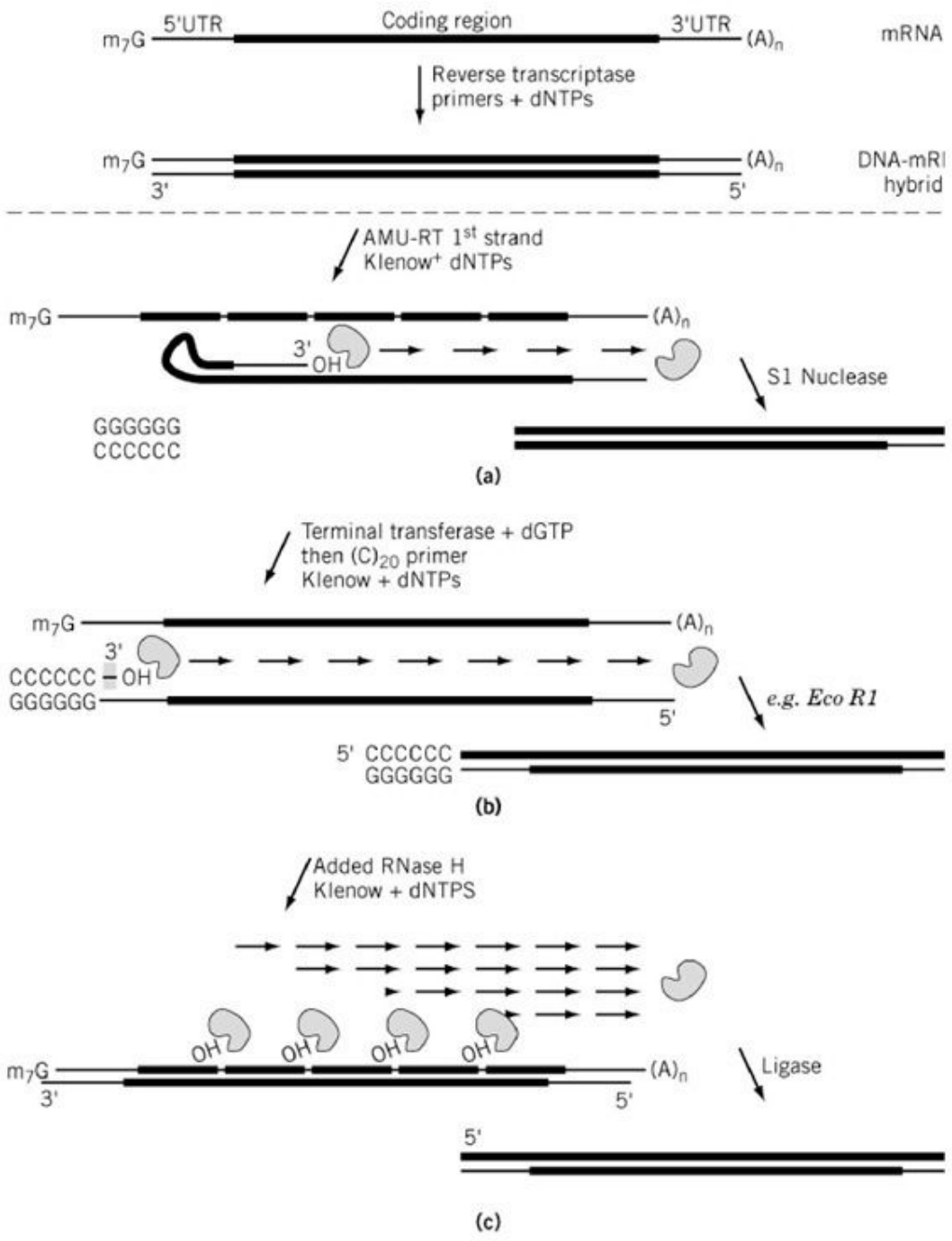
Library preparation steps (3)

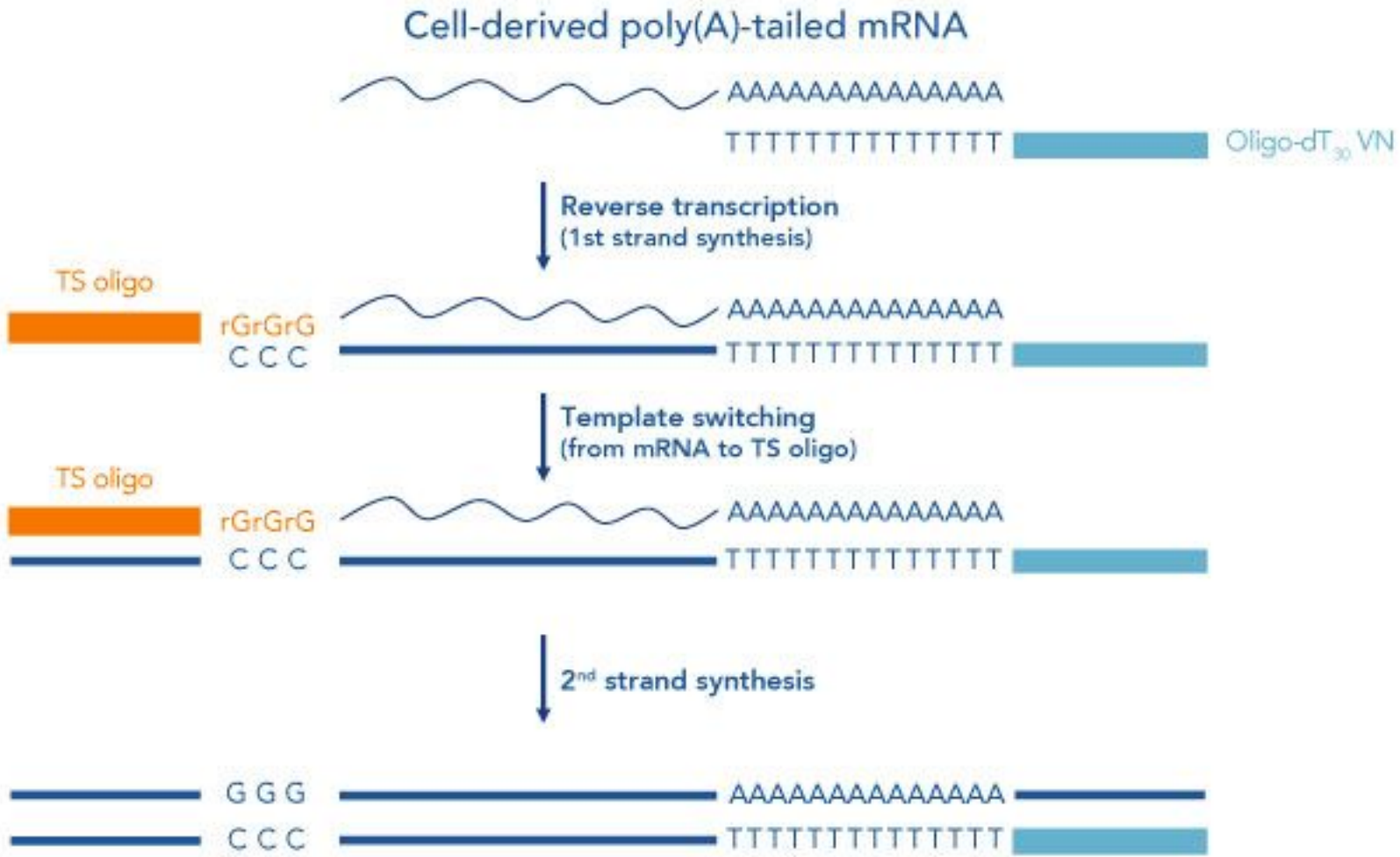
First strand cDNA synthesis

- RNA-directed DNA synthesis
- Oligo-dT priming
- Random 6-mers
  - Can recover non polyadenylated RNA
  - No 3’ coverage bias

Library preparation steps (4)

- Second strand cDNA synthesis
  - Hairpin formation
  - Terminal transferase
  - Okayama-Berg
  - Template switching

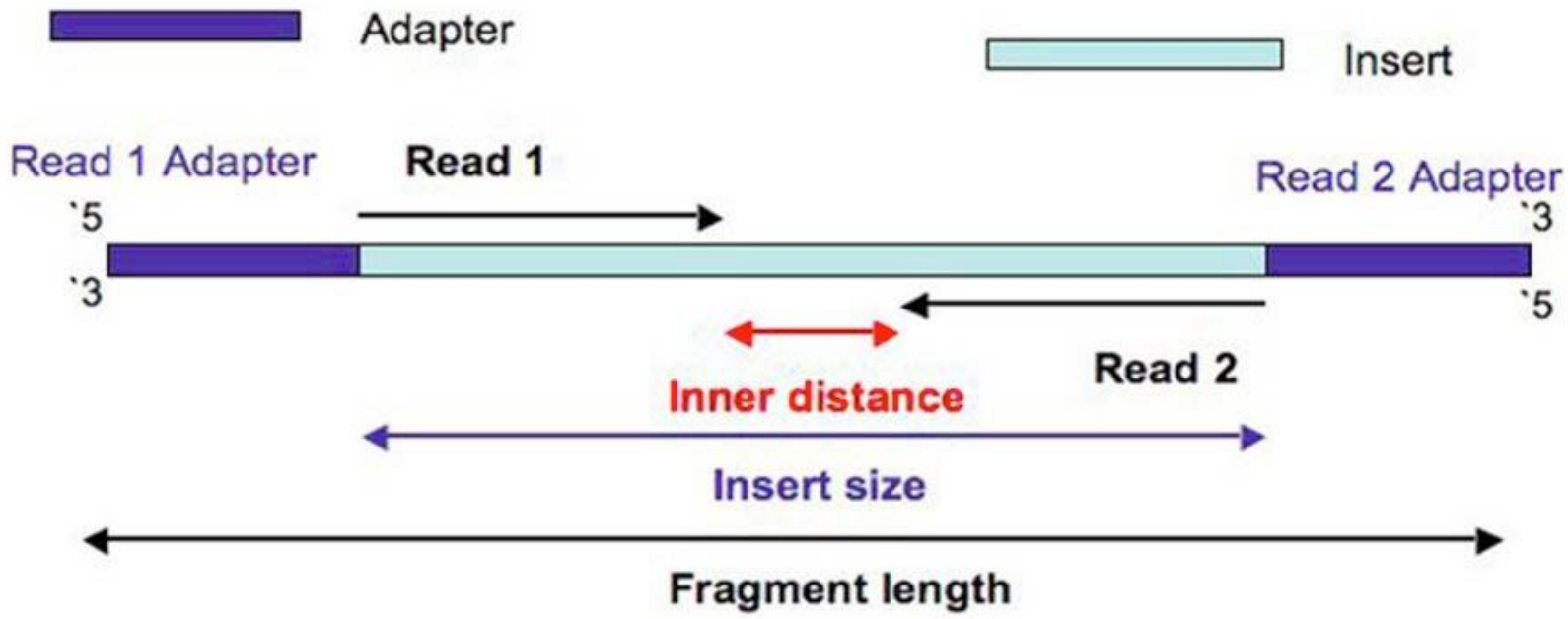



Cell-derived poly(A)-tailed mRNA
AAAAAAAAAAAAAAA
TTTTTTTTTTTTTTT
Oligo-dT30 VN
Reverse transcription
(1st strand synthesis)
TS oligo
rGrGrG
C C C
AAAAAAAAAAAAAAA
TTTTTTTTTTTTTTT
Template switching
(from mRNA to TS oligo)
TS oligo
rGrGrG
C C C
AAAAAAAAAAAAAAA
TTTTTTTTTTTTTTT
2nd strand synthesis
G G G
C C C
AAAAAAAAAAAAAAA
TTTTTTTTTTTTTTT

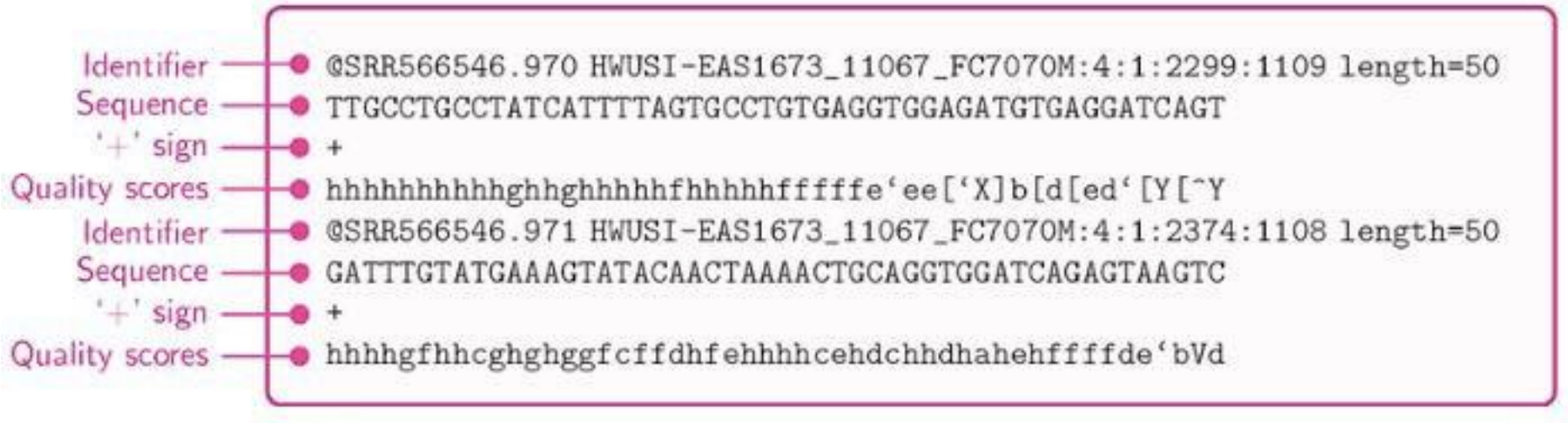
Library preparation steps (5&6)

- End repair
  - Remove 3’ overhangs
  - Fill in 5’ overhangs
- Adenylate 3’ ends
  - Add a single A to the blunt fragments to prevent self-ligation

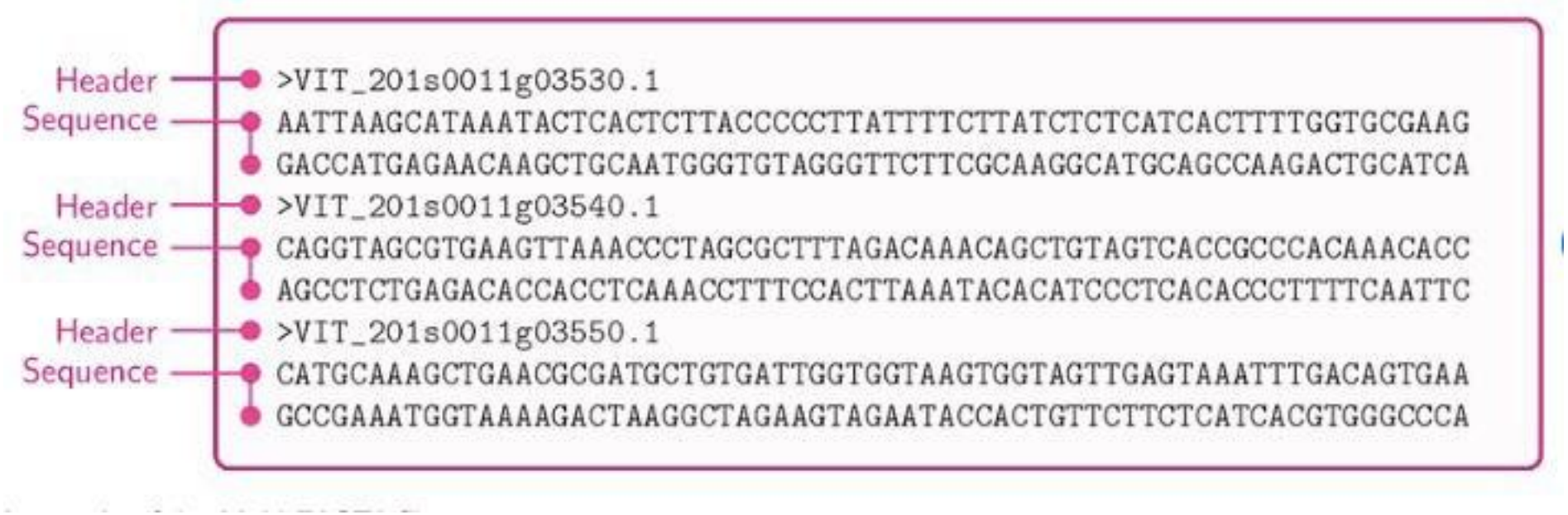
Recap of library structure



FASTQ format



FASTA format



Paired end sequence files

R1

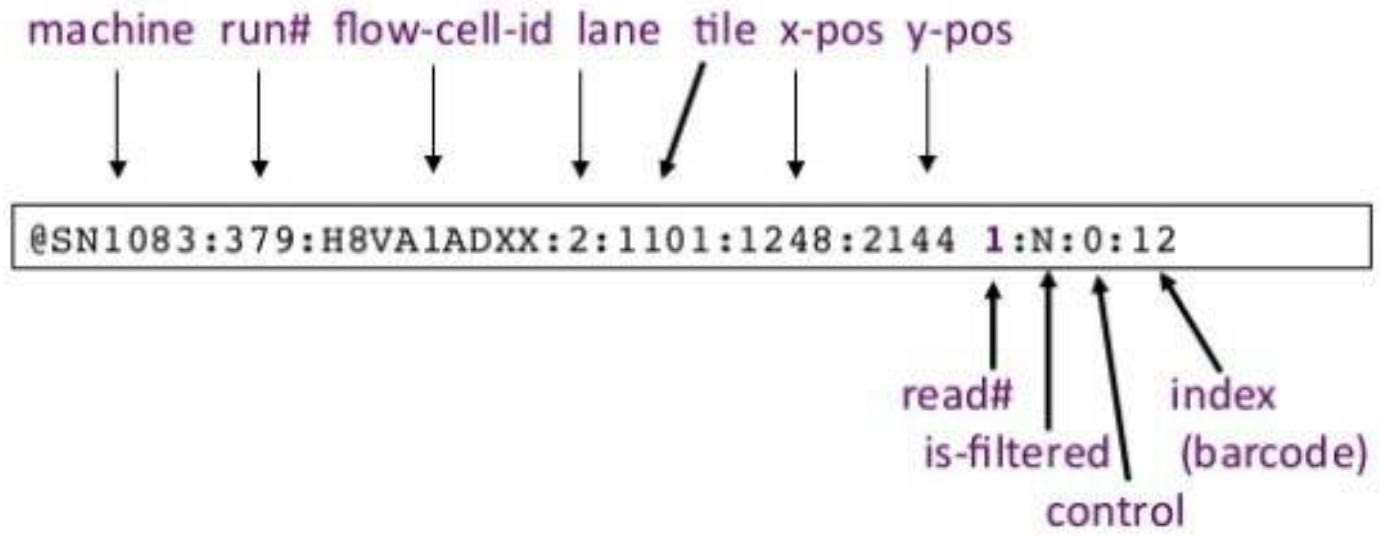
```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 1:N:0:12
CCTAAATGGTGCCATGCTAGGAGGCCGTGCCCTTCTTGAAAAGTTGTATGTGAA
+
BBBFFFFFFBFFFIIIIFI<FFIIIIIFIIIIFBFIIIIIIIIFFFIIIIFIII
```

R2

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 2:N:0:12
CATTTTCGACGTTGTTAATAAGCTCTGCGTACTTGCAAGCTATCTGCGCGAACG
+
BBBFFFFFFFFFIIIIIIIIIIIIIIIIFIIIIIIIIIIIIIIIIIIIFFF
```

Read 1 and read 2 have the same identifier

Header format



Base quality (Phred) scores

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

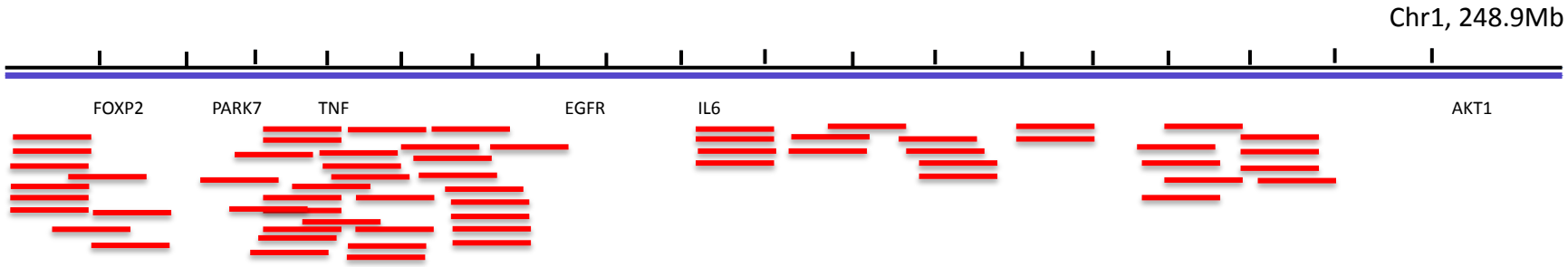
$$P = 10^{\frac{-Q}{10}}$$

Read processing

- Demultiplexing
- Adapter removal
- Quality trimming
- Quality assessment (FASTQC, MULTIQC)

Alignment

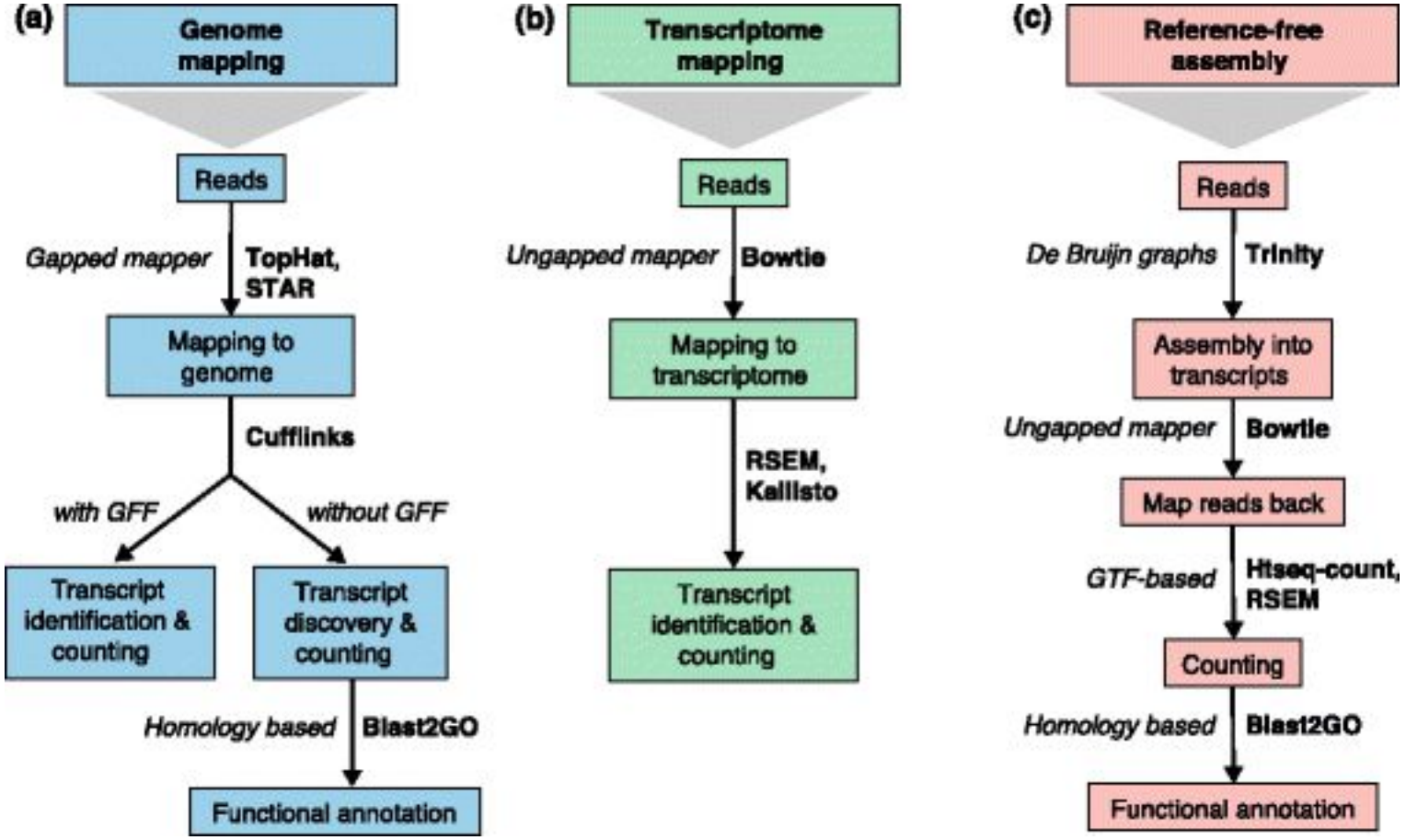
- Place RNA-Seq reads on the reference
- Count the reads in each gene



STEP3: A read counting tool then counts the number of reads aligning to a feature (gene, exon, non coding RNA)

Gene	Counts
FOXP2	10
PARK7	20
TNF	45
EGFR	2
IL6	50
AKT1	0

Alignment strategies



A survey of best practices for RNA-seq data analysis, Genome Biology, 2016. 17:13

New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

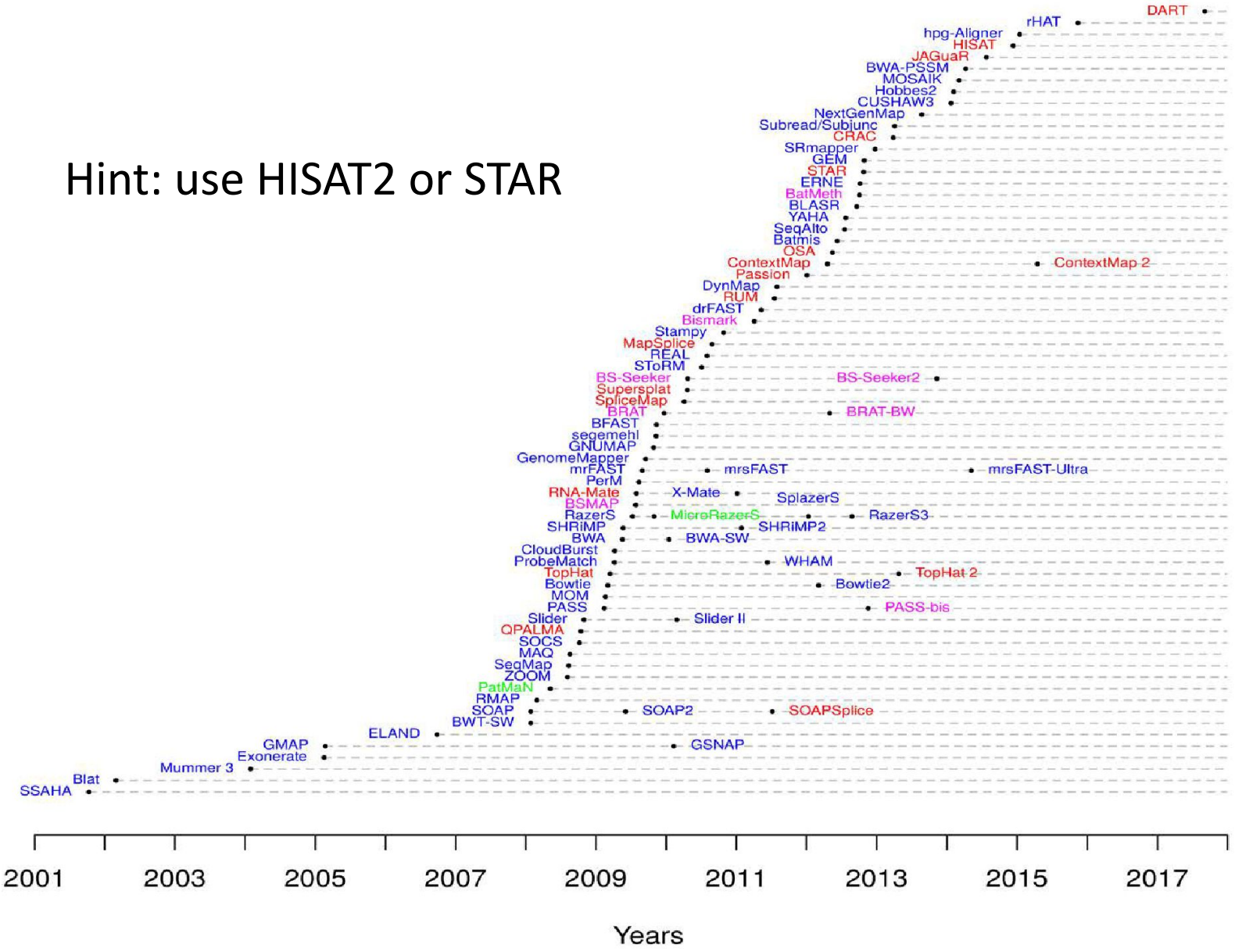
NCGR
National Center for Genome Resources

Aligner needs to

- Allow imperfect matches
  - Sequencing errors
  - SNPs
- Map the read to its origin on the genome or reference transcript
- Catalog reads mapping to multiple locations

- It's never perfect
  - Misassembly
  - Missing regions
  - Repetitive regions
  - Missing transcripts

Hint: use HISAT2 or STAR



New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

NCGR
National Center for Genome Resources

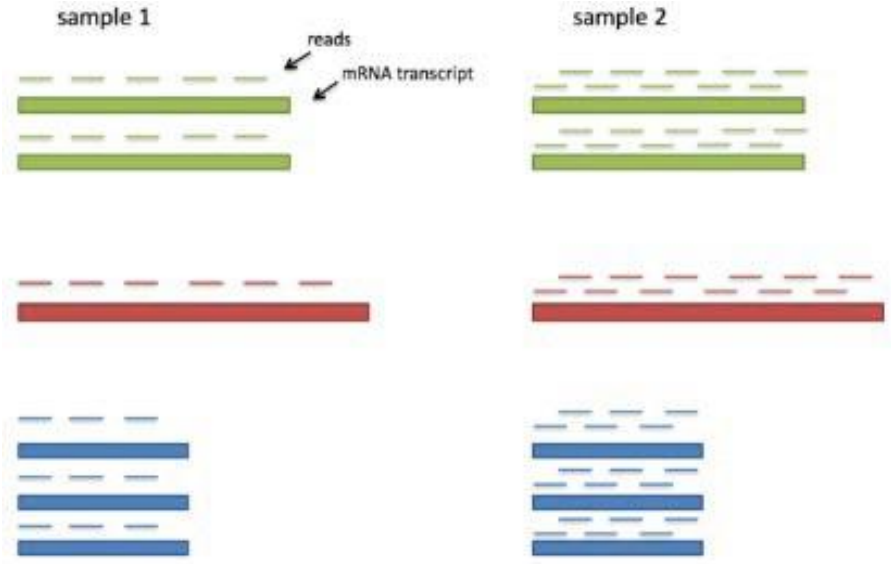
Read counting

- BAM/SAM file from alignment step
- GFF3 file
- Locate feature of interest (list of genes and their coordinates)
- Count the singly mapping genes between the coordinates

Normalization

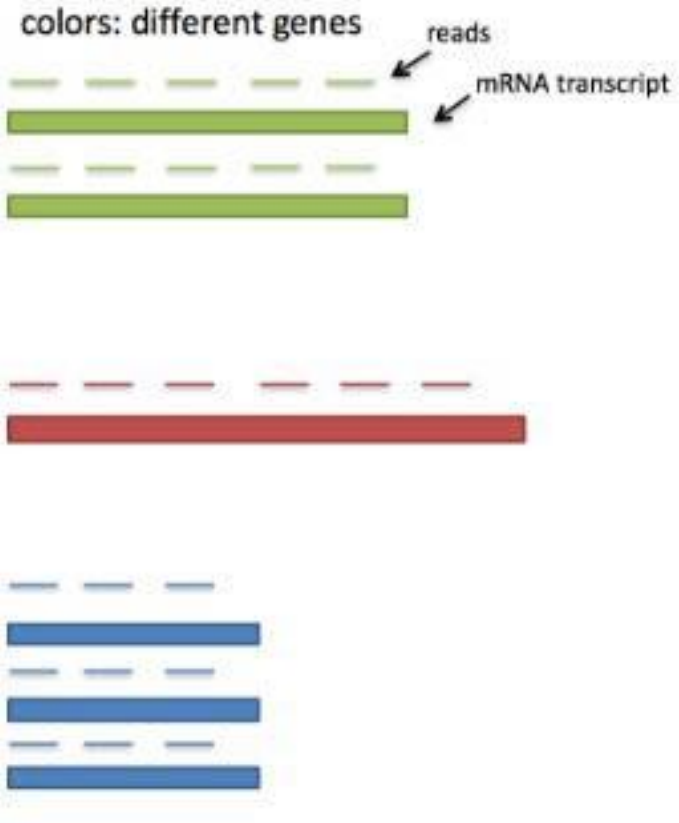
- The goal is to compare expression levels within and between samples
- We first have to consider methods not generally used by DE algorithms

Correction for sequencing coverage



Sequencing depth: Accounting for sequencing depth is necessary for comparison of gene expression between samples. Each gene appears to have doubled in expression in sample 2. This is a consequence of sequencing depth!!!

Correction for gene length



Gene length: Accounting for gene length is necessary for comparing expression between different genes within the same sample. The number of reads mapped to a longer gene can appear to have equal count/expression as a shorter gene that is more highly expressed.

Common methods

Read versus fragment

Two Reads mapping = one Fragment

One Read mapping = one Fragment

Gene

Common methods

- CPM/RPM/FPM (counts per million): counts scaled by total number of reads. This method accounts for sequencing depth only
- RPKM/FPKM (reads/fragments per kilobase million): CPM divided by gene length
- TPM (transcripts per million): FPKM divided by the sum of FPFMs in a sample. Sum of TPM in each sample is the same

FPKM/TPM: beware

- Counts are assigned to longest theoretical isoform
  - It may not exist
  - Relative proportions of isoforms are unknown
- Gene families
  - Identical regions will have multimapping reads, which are ignored
  - True expression values will be underestimated

Normalization for DE analysis

- Don’t particularly need to account for gene length
- Same gene, different sample. We want the difference
- (Unless you care about absolute levels)
- DO need to account for RNA composition

Gene	Sample A counts	Sample B counts
FOXP2	10	110
PARK7	20	120
TNF	45	145
EGFR	500	0
IL6	50	150
AKT1	5	105

DE normalization purpose

- Normalization aims to determine a scaling factor to be applied to all genes in each sample.
- Scaling factor will be different for each sample.
- The aim is to make the samples comparable with one another
- The method removes genes at the extremes of dispersion first across, and then within samples

Step 1

- Compute the geometric mean for each gene, across samples
- The mean is called the pseudo-reference sample

Gene	Sample A counts	Sample B counts	Geometric mean
EF2A	1489	906	1161.5
ABCD1	22	13	16.9
MEFV	793	410	570.2
BAG1	76	42	56.5
MOV10	521	1196	883.7

Step 2

- Compute the ratio of the sample counts to the pseudo reference

Gene	Sample A counts	Sample B counts	Geometric mean	Sample A/ Pseudo-ref	Sample b/ Pseudo-ref
EF2A	1489	906	1161.5	1.28	0.78
ABCD1	22	13	16.9	1.30	0.77
MEFV	793	410	570.2	1.39	0.72
BAG1	76	42	56.5	1.35	0.74
MOV10	521	1196	883.7	0.59	1.35

Step 3

- Compute the median value of the count ratios

Gene	Sample A counts	Sample B counts	Geometric mean	Sample A/ Pseudo-ref	Sample b/ Pseudo-ref
EF2A	1489	906	1161.5	1.28	0.78
ABCD1	22	13	16.9	1.30	0.77
MEFV	793	410	570.2	1.39	0.72
BAG1	76	42	56.5	1.35	0.74
MOV10	521	1196	883.7	0.59	1.35
				1.30	**0.77**

Step 4

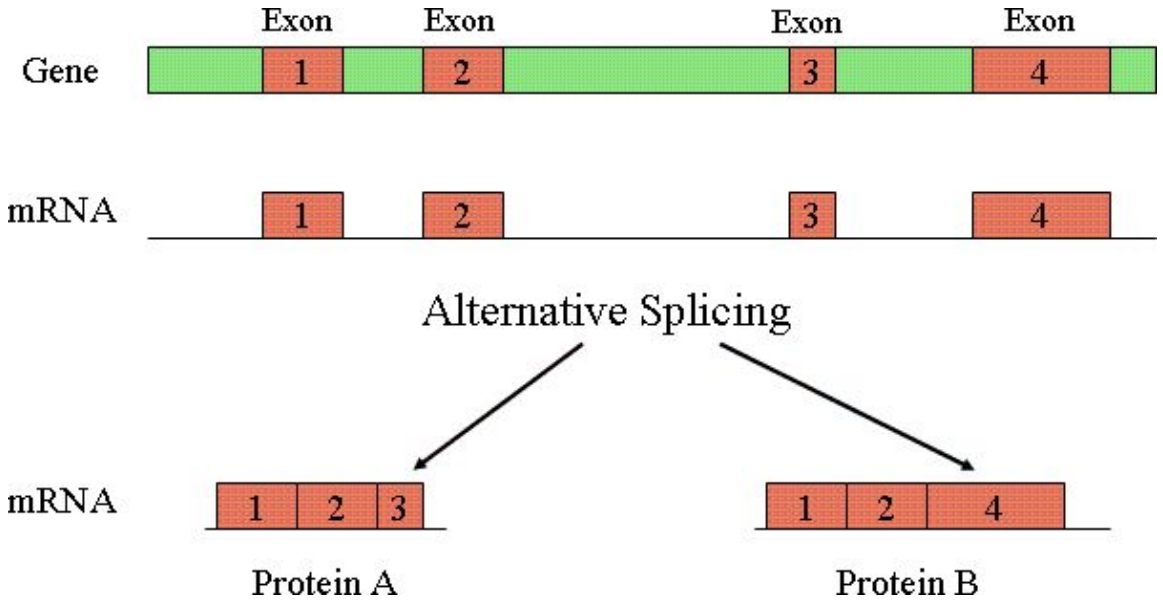
- Divide the counts for each sample by it's own scaling factor

Gene	Sample A counts	Sample B counts	Geometric mean	Sample A/ Pseudo-ref	Sample b/ Pseudo-ref	Sample A Norm'd	Sample B Norm'd
EF2A	1489	906	1161.5	1.28	0.78	**1145**	**1177**
ABCD1	22	13	16.9	1.30	0.77	**17**	**17**
MEFV	793	410	570.2	1.39	0.72	**610**	**532**
BAG1	76	42	56.5	1.35	0.74	**58**	**55**
MOV10	521	1196	883.7	0.59	1.35	**401**	**1553**
				1.30	0.77		

Other applications of RNA-Seq

- Whole transcriptome sequencing
- Gene discovery
- Variant discovery
- Allele-specific expression
- Alternative spliceforms
- Small RNA
- ncRNA

Alternative splicing



Splice junctions with short read sequencing

Individual splice junctions are not difficult

Complete isoform structures are difficult

PacBio Iso-Seq

- Attempt to sequence full length transcripts
- Characterize isoforms across entire transcriptome
- Can discover novel genes and isoforms simultaneously in un characterized samples
- No reference knowledge required
- Sample quality is paramount
- Quantitative?

UMI

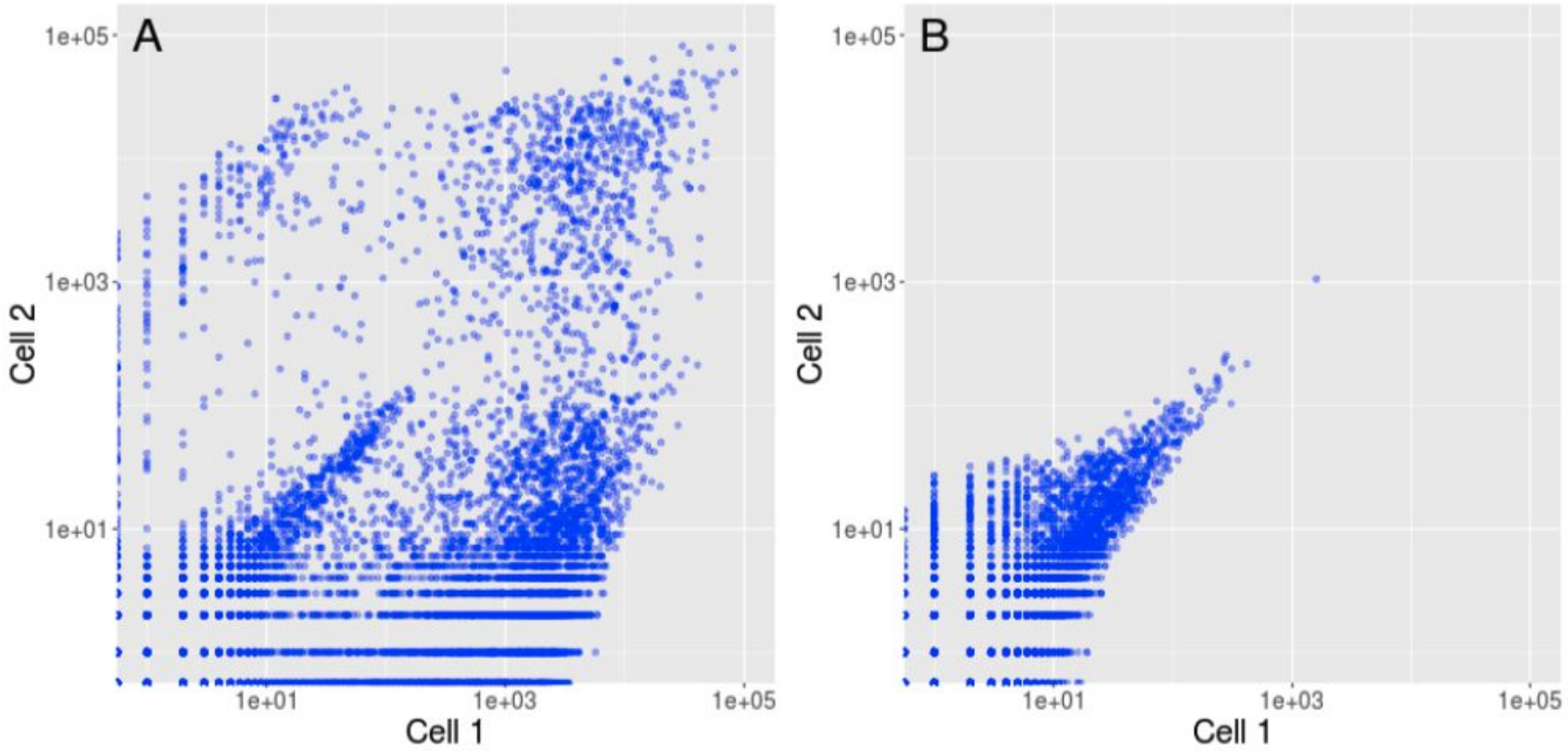


Figure S 1. Comparison of RNA-Seq UMI and read counts from sequencing data derived from two *P. patens* protoplasts. A: read counts per gene. B: UMI counts per gene. UMI and read counts are shown for the same data. Each dot plots a gene having two counts of reads or UMI as indicated by the X and Y axes.