Introduction to Trinity RNA-Seq

Berlin, June 2018

Brian Haas
Broad Institute

Nicolas Delhomme
Umeå universitet

Physalia
Courses

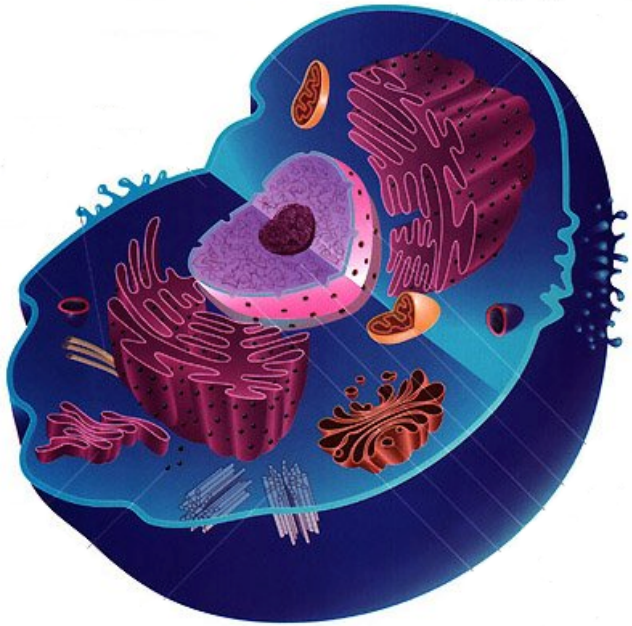
Welcome to the Berlin 2018 Trinity Workshop Wiki!

Day	Time	Activities
Monday, June 11	morning	Workshop Introduction -slides-
		Exploring the Computational Infrastructure -slides-
	afternoon	Unix command-line review
		Data overview and setup
		Using FASTQC and Trimmomatic -slides-
Tuesday, June 12	morning	Trinity de novo transcriptome assembly -slides-
	afternoon	Uploading own data or identifying and downloading SRA studies of interest -slides-
Wednesday, June 13	morning	Expression quantification -slides-
		Quality assessment for assembly -slides-
	afternoon	QC samples and replicates
Thursday, June 14	morning	Statistical methods for differential expression analysis -slides-
	afternoon	Transcript clustering and expression profiling
		Methods for functional annotation -slides-
		Trinotate and TrinotateWeb
Friday, June 15	morning	Functional enrichment analysis
		Review and custom data analyses
		Comments on software installations for later use on different resources

https://github.com/trinityrnaseq/BerlinTrinityWorkshop2018/wiki

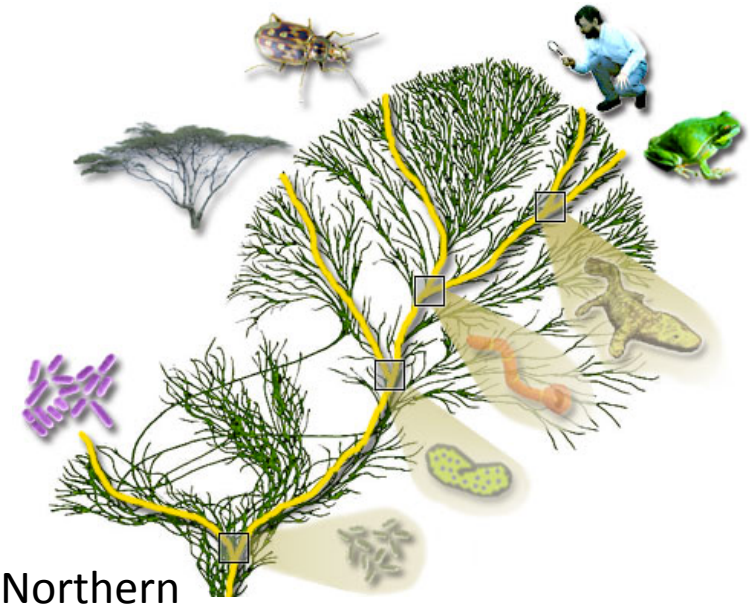
Transcriptomics

A powerful tool for studying molecular biology

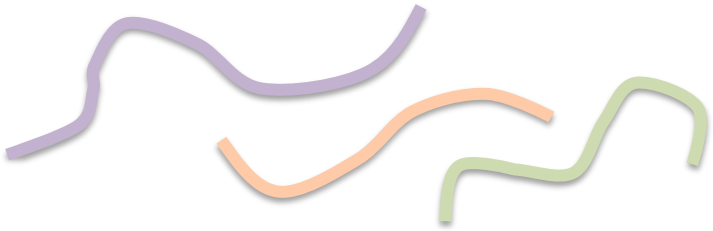


- Yields insights into:
  - Identity of expressed transcripts.
  - Levels of expression.
  - Differences in expression across samples or conditions.

Biological Investigations Empowered by Transcriptomics

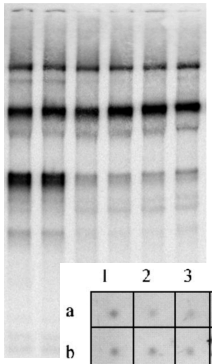


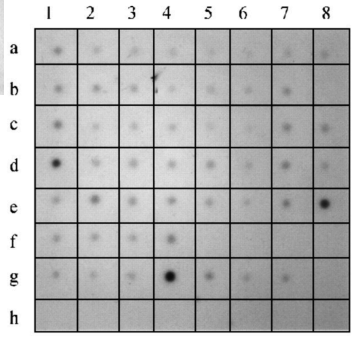
Extract RNA,
… some protocol for processing, …



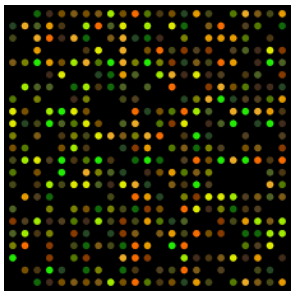
Analysis Method
(pick your favorite)

Northern

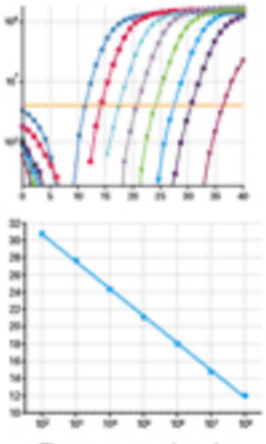




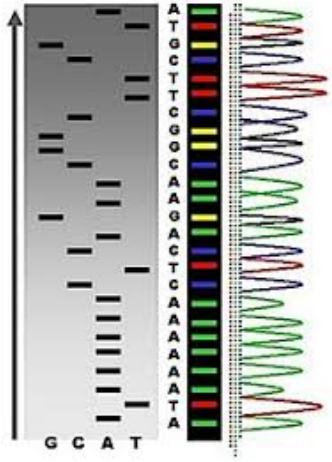
Dot Blot



Microarray



qRT-PCR



Sanger Sequencing

Other…

Minion

MinION

Historical Timeline of Transcriptomics (from 1970)

Reverse Transcription (1970)

Northern Blot
Sanger Sequencing
(1977)

Expressed Sequence Tags (1992)

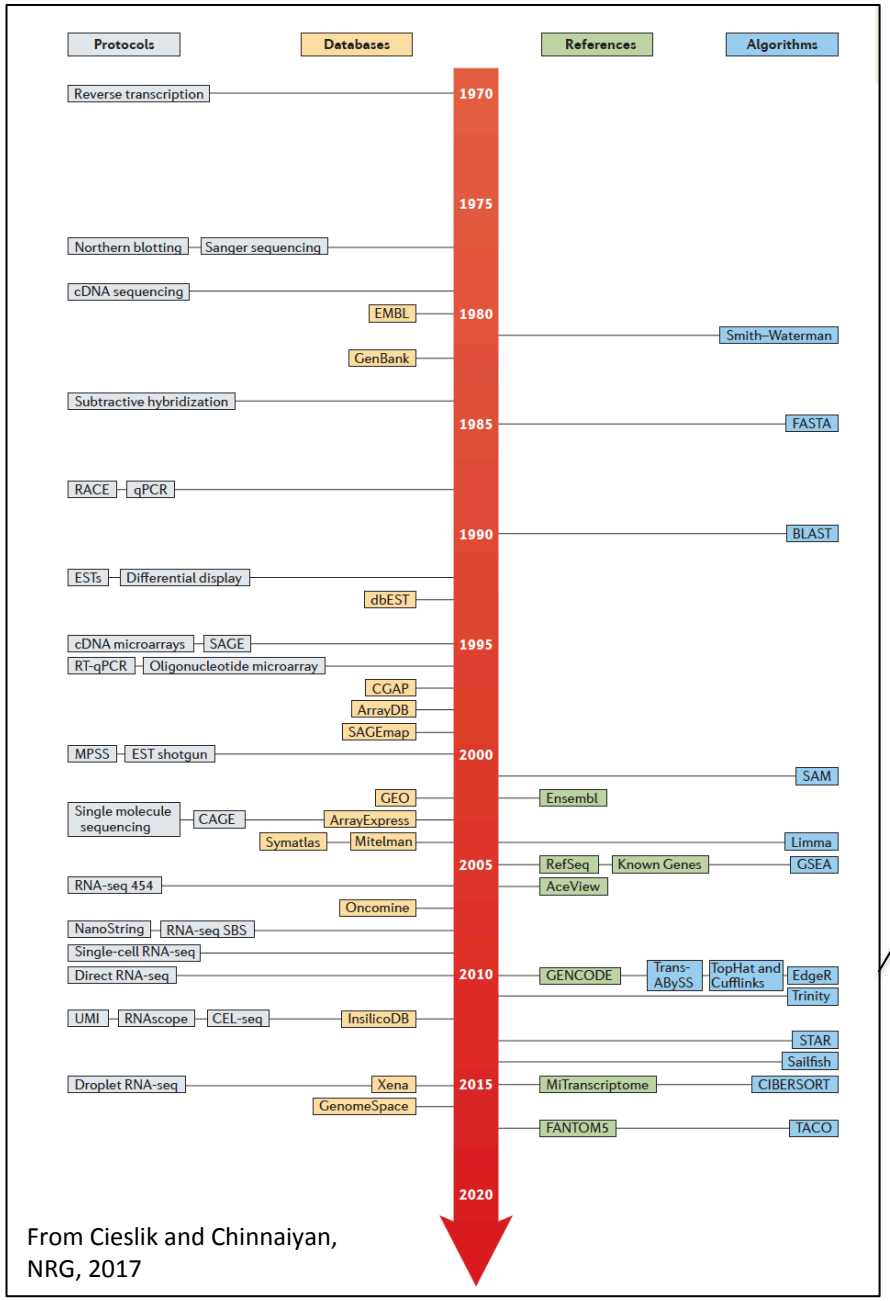
cDNA microarrays (1995)

RNA-Seq (2006-2008)

PacBio IsoSeq (2014)

Droplet single cell RNA-Seq (2015)

Direct RNA Seq Nanopore (2018)



From Cieslik and Chinnaiyan, NRG, 2017

Note: Just a small sampling of what's available.

Smith Waterman (1981)

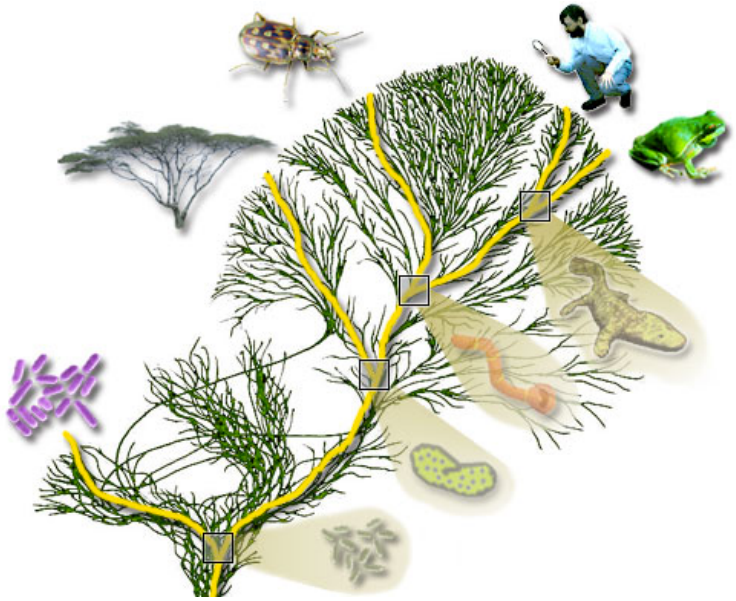
BLAST (1990)

Tophat/Cufflinks (2010)

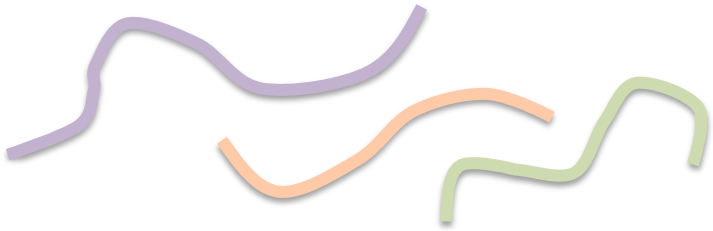
RSEM (2011)

Kallisto (2016)
Salmon (2017)

Modern Transcriptome Studies Empowered by RNA-seq



Extract RNA, convert to cDNA



Next-gen Sequencer
(pick your favorite)

Millions to Billions of Reads

Generating RNA-Seq: *How to Choose?*

Platform	Project Firefly 2018	MiniSeq	MiSeq	Next Seq 550	HiSeq 2500 RR	Hiseq 2500 V3	HiSeq 2500 V4	HiSeq 4000	HiSeq X	Nova Seq S1 2018	Nova Seq S2	Nova Seq S4	5500 XL	318 HiQ 520	Ion 530	Ion Proton P1	PGM HiQ 540	RS P6-C4	Sequel	R&D end 2018	Smidg ION RnD	Mini ION R9.5	Grid ION X5	Prome thION RnD	Prome thION theor etical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
Reads: (M)	4	25	25	400	600	3000	4000	5000	6000	3300	6600	20000	1400	3-5	15-20	165	60-80	5.5	38.5	--	--	--	--	--	--	400	1600	1600	--
Read length: (paired-end*)	150*	150*	300*	150*	100*	100*	125*	150*	150*	150*	150*	150*	60	200 400	200 400	200	200	15K	12K	32K	--	--	--	--	--	--	100*	50	--
Run time: (d)	0.54	1	2	1.2	1.125	11	6	3.5	3	1.66	1.66	1.66	7	0.37	0.16	--	0.16	4.3	--	--	--	2	2	2	--	--	1	0.4	--
Yield: (Gb)	1	7.5	15	120	120	600	1000	1500	1800	1000	2000	6000	180	1.5	7	10	12	12	5	150	4	8	40	2400	11000	80	200	8	--
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	55	166	400	600	600	1200	3600	30	5.5	50	--	93.75	2.8	--	--	--	4	20	1200	5500	--	200	20	--
Reagents: ($K)	0.1	1.75	1	5	6.145	23.47	29.9	--	--	--	--	--	10.5	0.6	--	1	1.2	2.4	--	1	--	0.5	1.5	--	--	0.5	--	--	--
per-Gb: ($)	100	233	66	50	51.2	39.1	31.7	20.5	7.08	18	15	5.8	58.33	--	--	100	--	200	80	6.6	--	62.5	37.5	20	4.3	--	--	--	--
hg-30x: ($)	12000	28000	8000	5000	6144	4692	3804	2460	849.6	1800	1564	700	7000	--	--	12000	--	24000	9600	1000	--	7500	4500	2400	500	--	600	--	
Machine: ($)	30K	49.5K	99K	250K	740K	690K	690K	900K	1M	999K	999K	999K	595K	50K	65K	243K	242K	695K	350K	350K	--	--	125K	75K	75K	--	200K	--	--
#Page maintained by http://twitter.com/albertvilella http://tinyurl.com/ngslytics #Editable version: http://tinyurl.com/ngsspecsshared																													
#curl "https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | grep -v '^"' | column -t -s\, | less -S																													

*Not all shown at scale

Generating RNA-Seq: *How to Choose?*

Platform	Project Firefly 2018	MiniSeq	MiSeq	Next Seq 550	HiSeq 2500 RR	Hiseq 2500 V
Reads: (M)	4	25	25	400	600	300
Read length: (paired-end*)	150*	150*	300*	150*	100*	100
Run time: (d)	0.54	1	2	1.2	1.125	1
Yield: (Gb)	1	7.5	15	120	120	60
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	5
Reagents: ($K)	0.1	1.75	1	5	6.145	23.4
per-Gb: ($)	100	233	66	50	51.2	39.
hg-30x: ($)	12000	28000	8000	5000	6144	469
Machine: ($)	30K	49.5K	99K	250K	740K	690

#Page maintained by http://twitter.com/albertvilella http://t

#curl "https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q

g	Mini ION R9.5	Grid ION X5	Prome thION RnD	Prome thION theor etical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
--	--	--	--	--	400	1600	1600	--
--	--	--	--	--	--	100*	50	--
--	2	2	2	--	--	1	0.4	--
4	8	40	2400	11000	80	200	8	--
--	4	20	1200	5500	--	200	20	--
--	0.5	1.5	--	--	0.5	--	--	--
--	62.5	37.5	20	4.3	--	--	--	--
--	7500	4500	2400	500	--	600	--	
--	--	125K	75K	75K	--	200K	--	--

"What I especially like about this baby is this little drawer where I can keep my lunch."

Thx Joshua Levin, for the cartoon. ☺

Small to Large

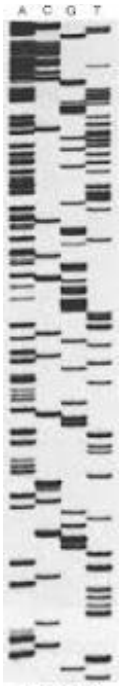
Each has pros/cons

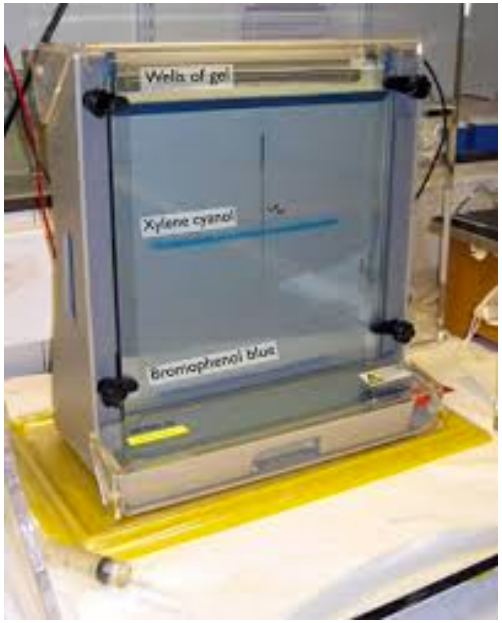
Small to Less Large

Each technology continues to advance

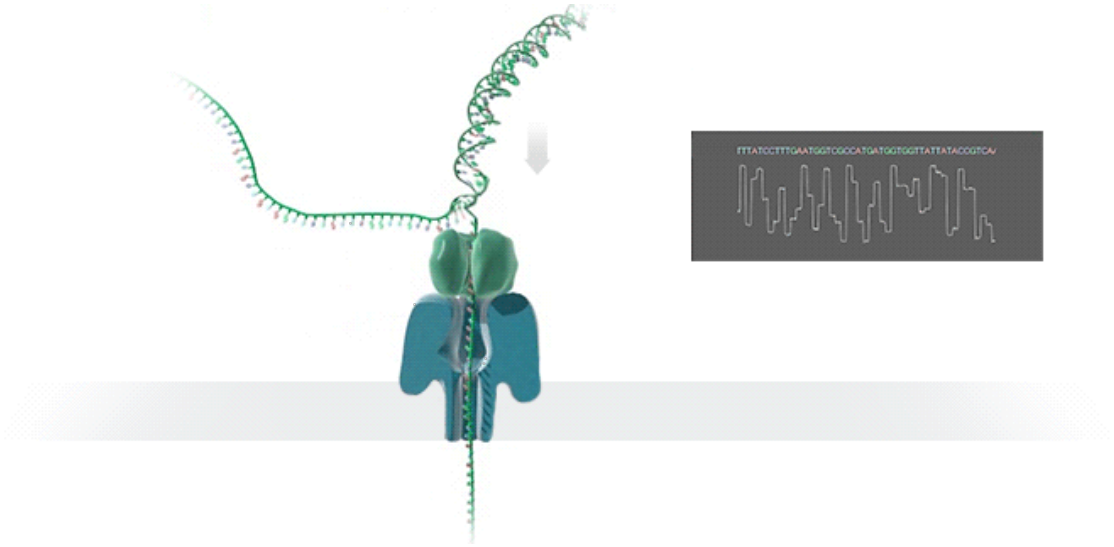
Personal Reflections… (and haunting memories)

Circa 1995

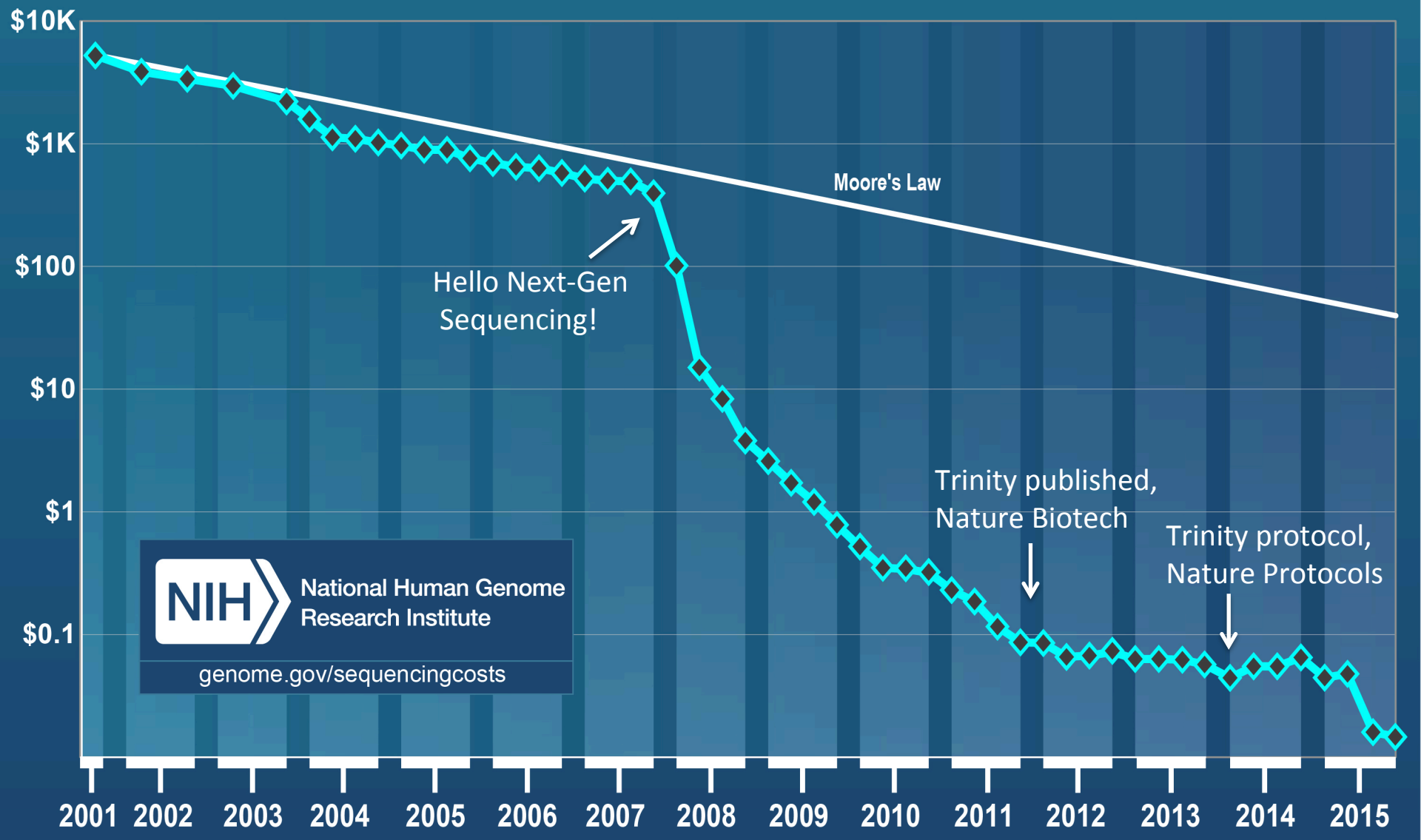




Now (2018)



Cost per Raw Megabase of DNA Sequence



From https://www.genome.gov/sequencingcostsdata/

A Plethora of Biological Sequence Analyses Enabled by RNA-Seq

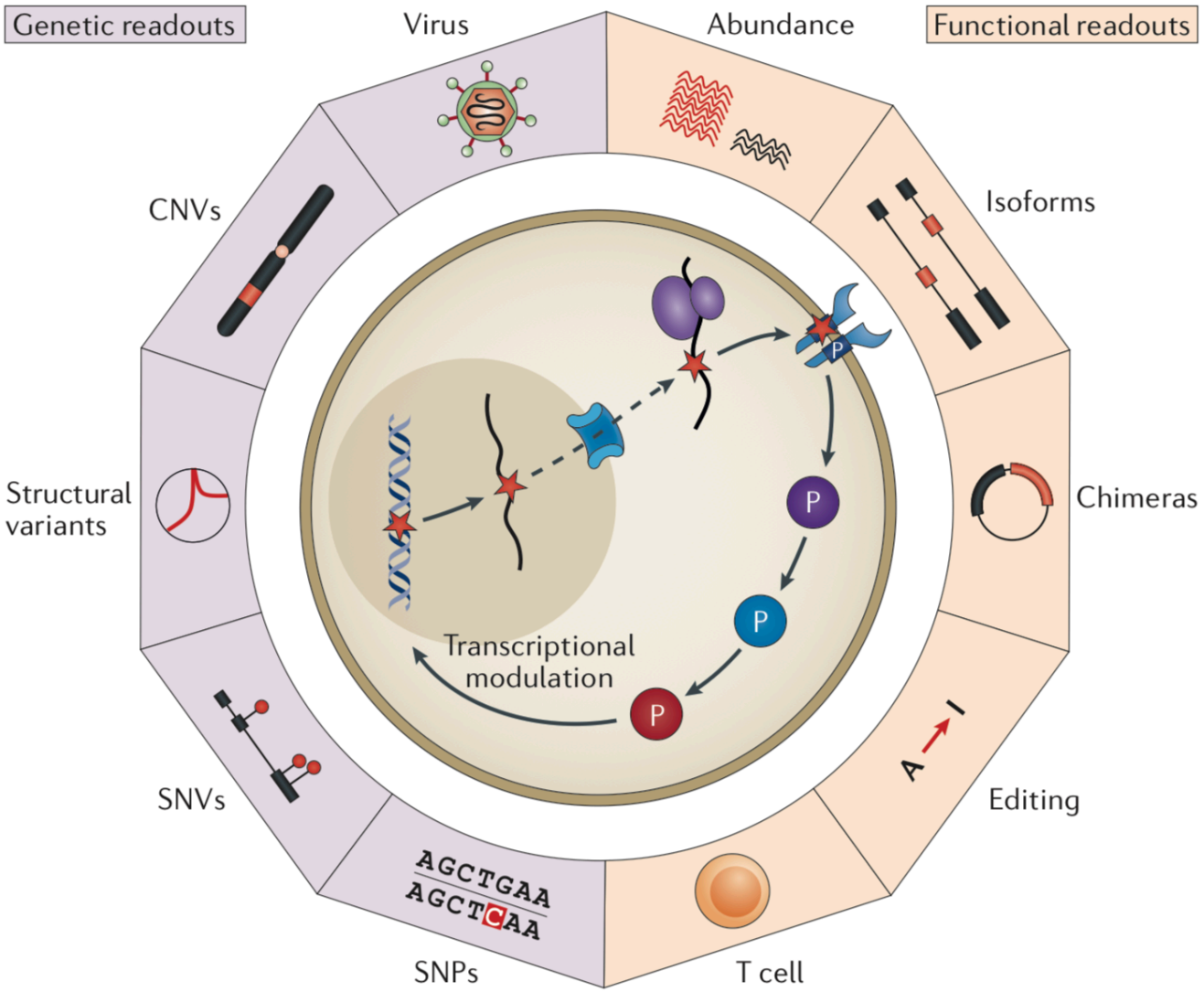
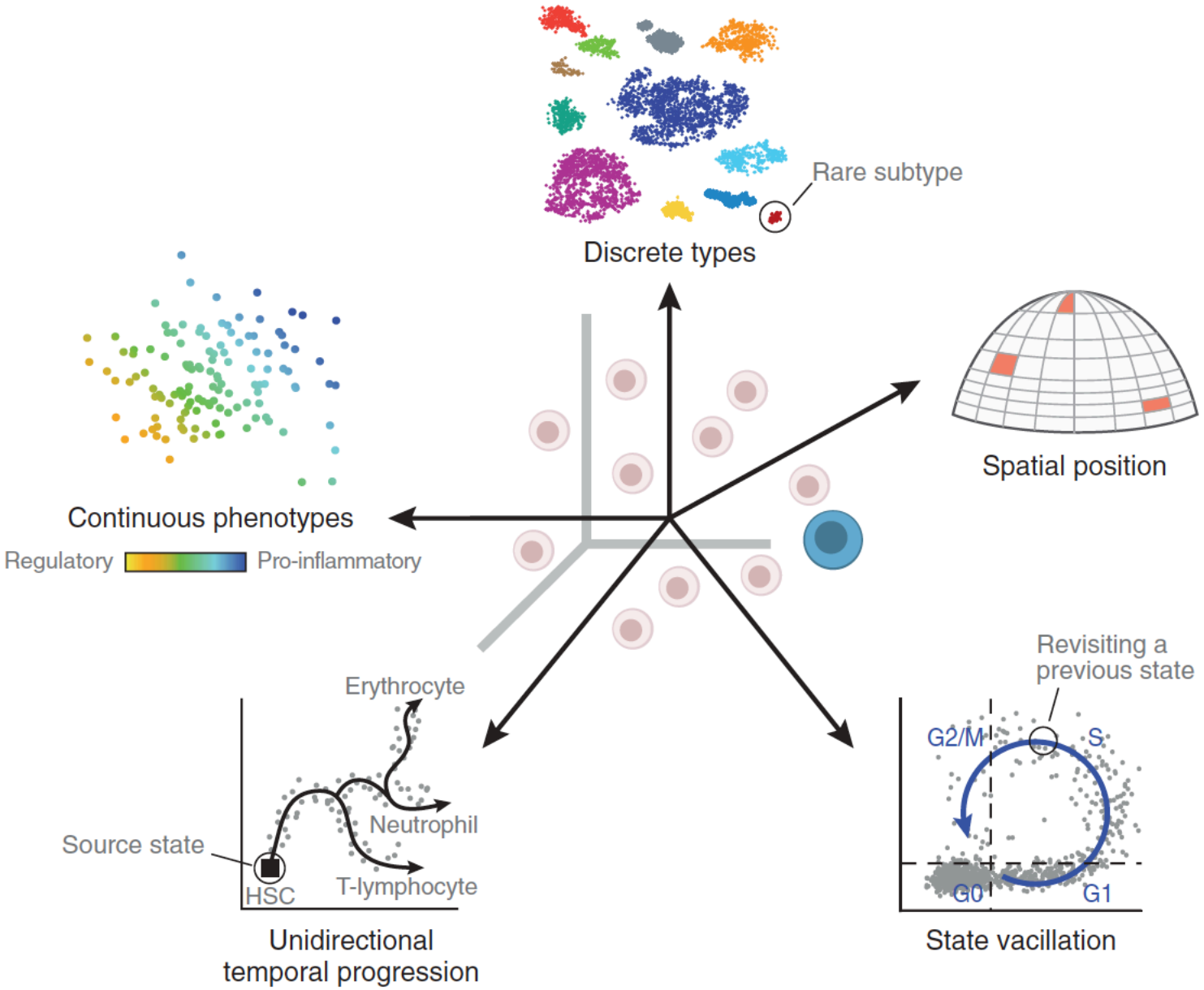


Figure 2 | **Transcriptome profiling for genetic causes and functional phenotypic readouts.**

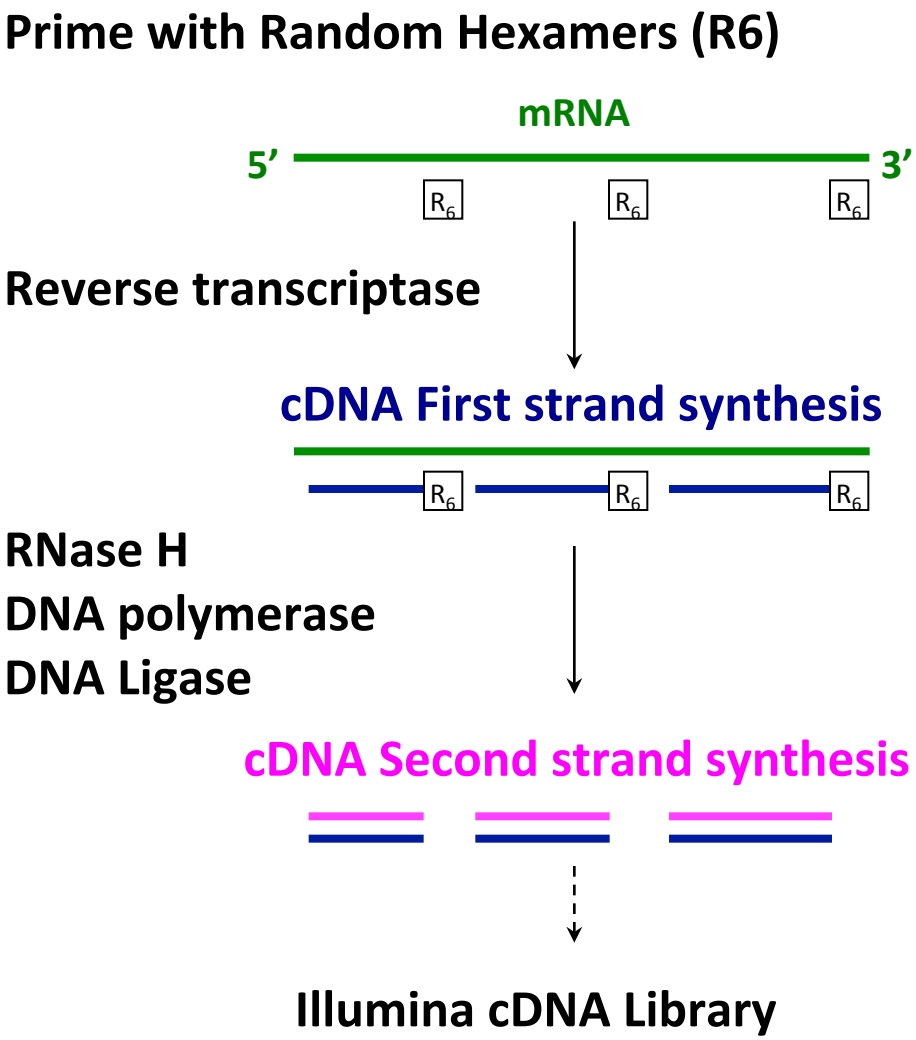
From Cieslik and Chinnaiyan, NRG, 2017

RNA-Seq is Empowering Discovery at Single Cell Resolution



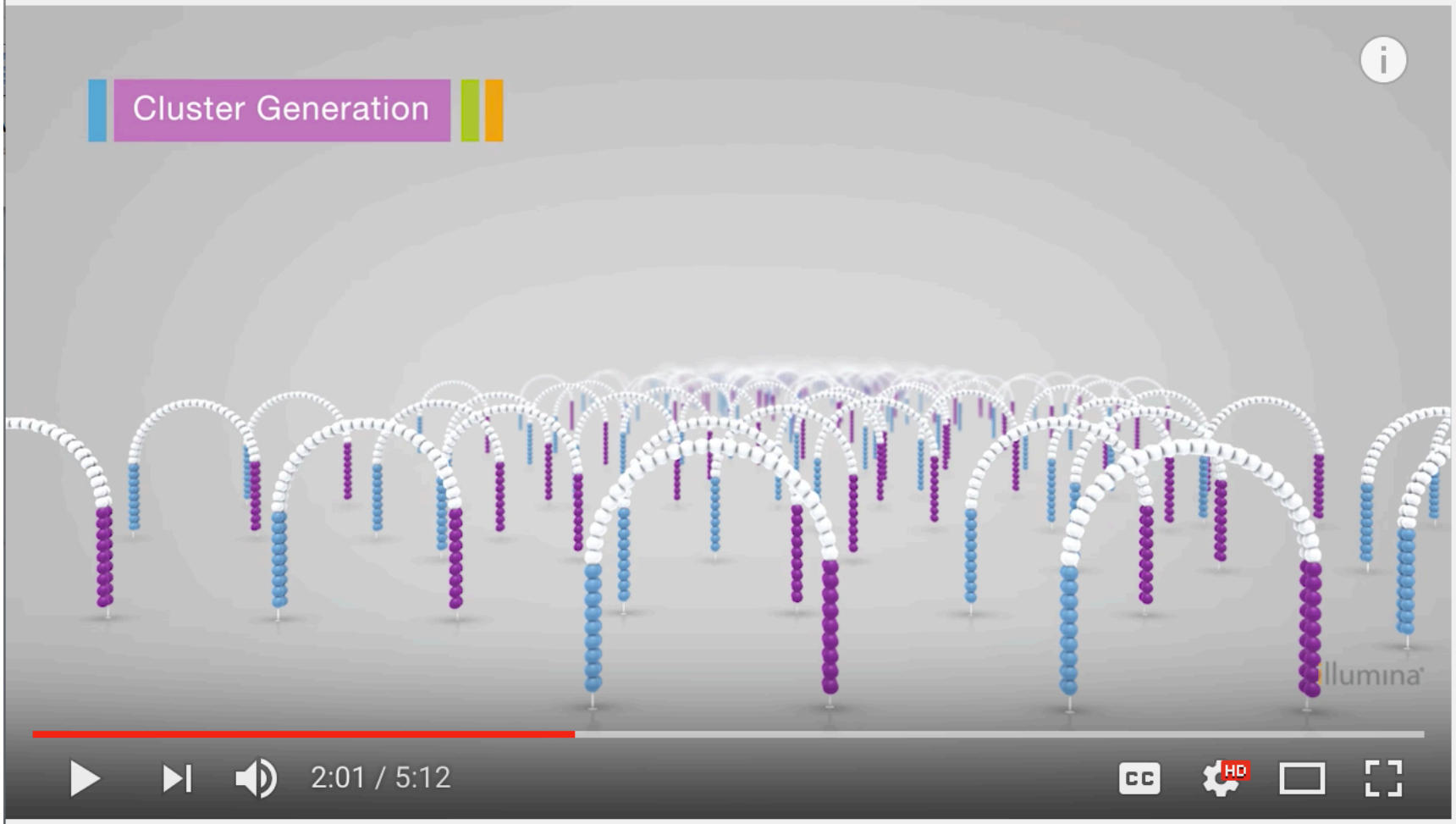
Wagner, Regev, and Yosef. NBT 2016

RNA-Seq: How do we make cDNA?



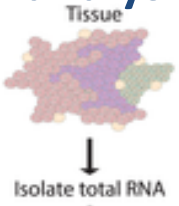
Slide courtesy of Joshua Levin, Broad Institute.

Illumina Sequencing by Synthesis



Illumina Inc https://www.youtube.com/watch?v=fCd6B5HRaZ8

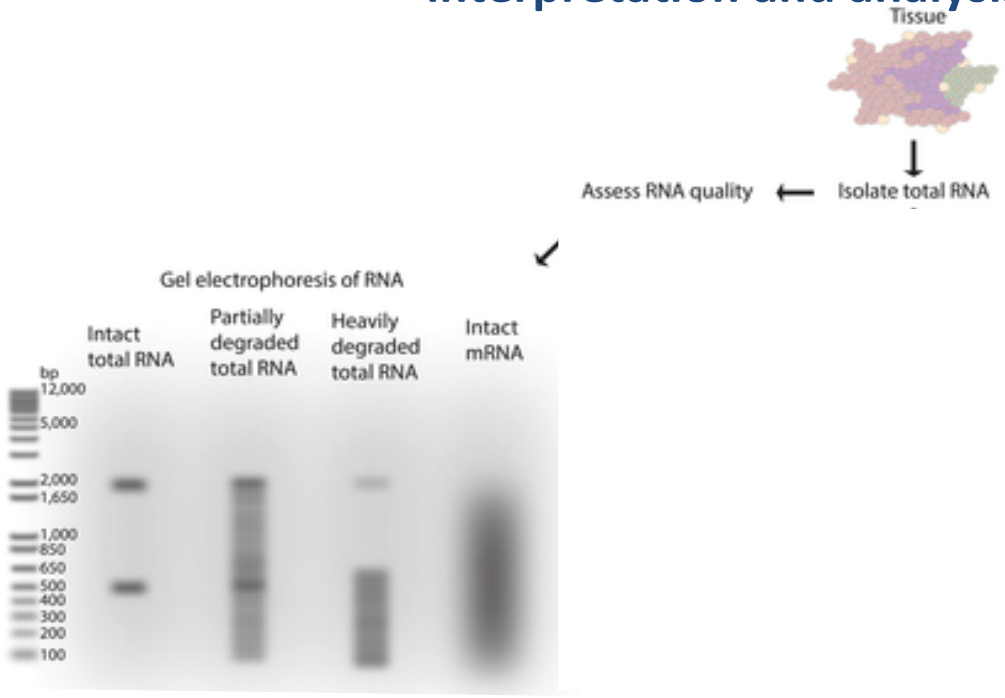
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

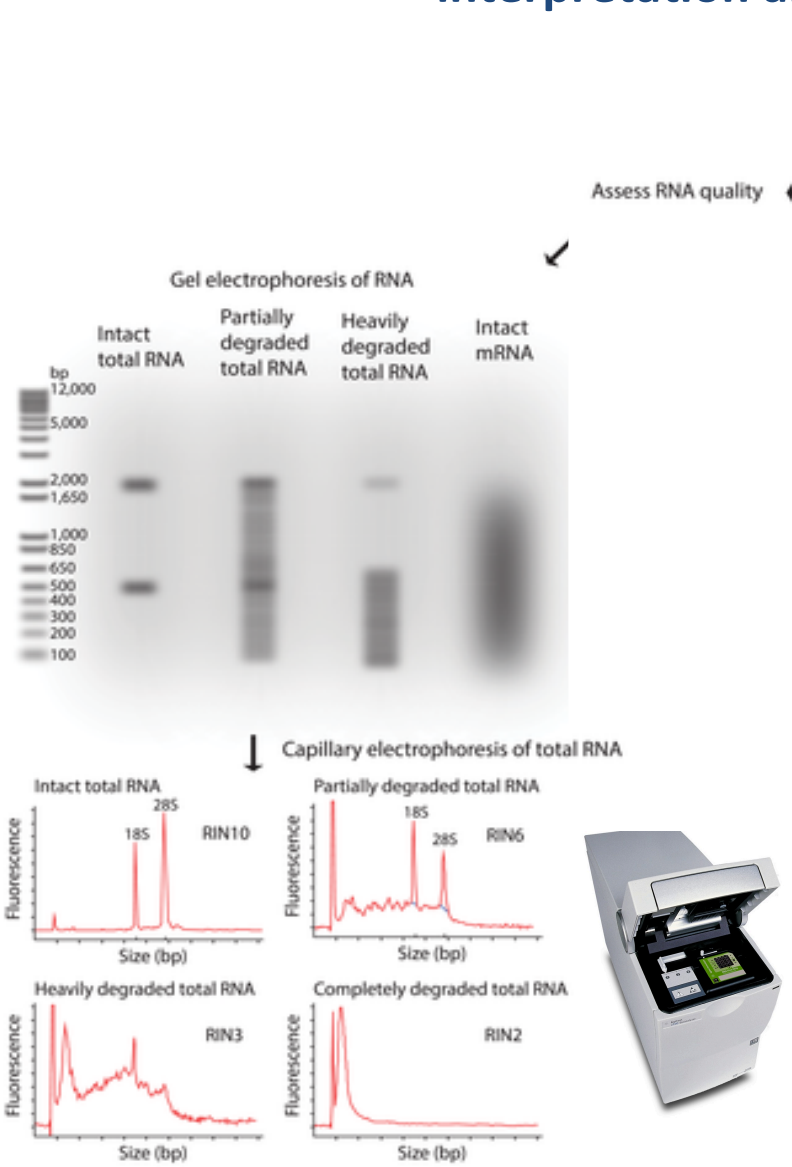
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

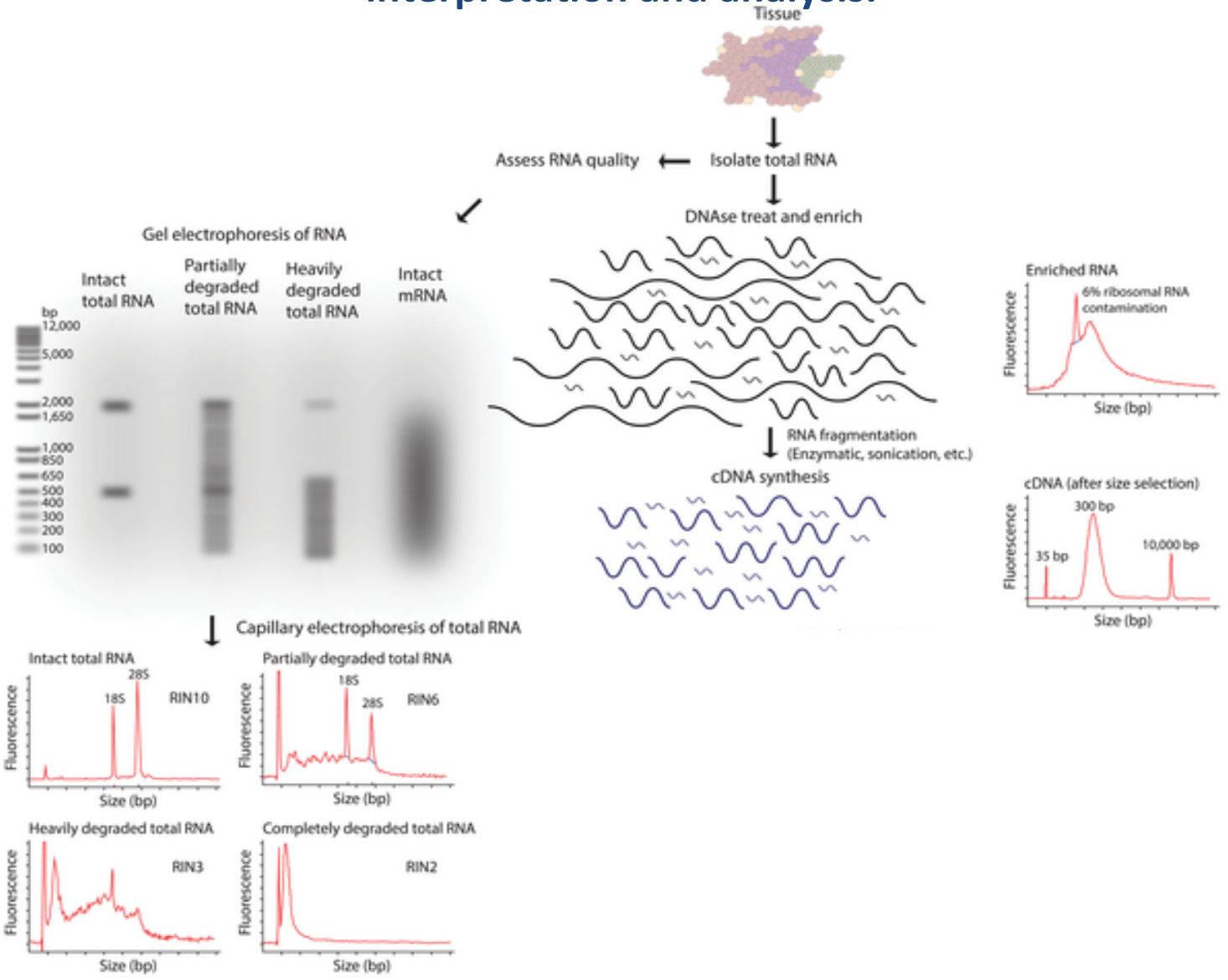
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

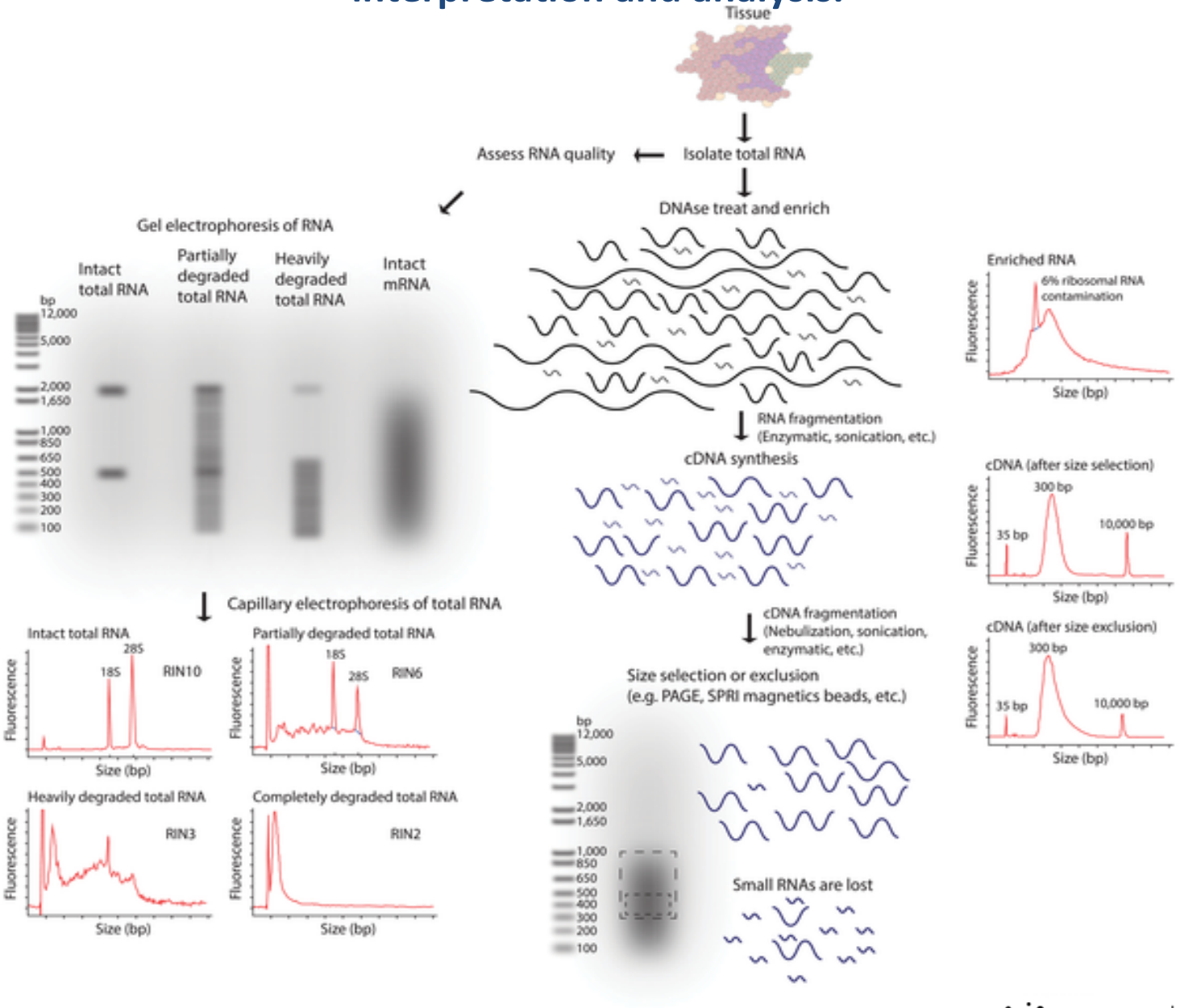
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

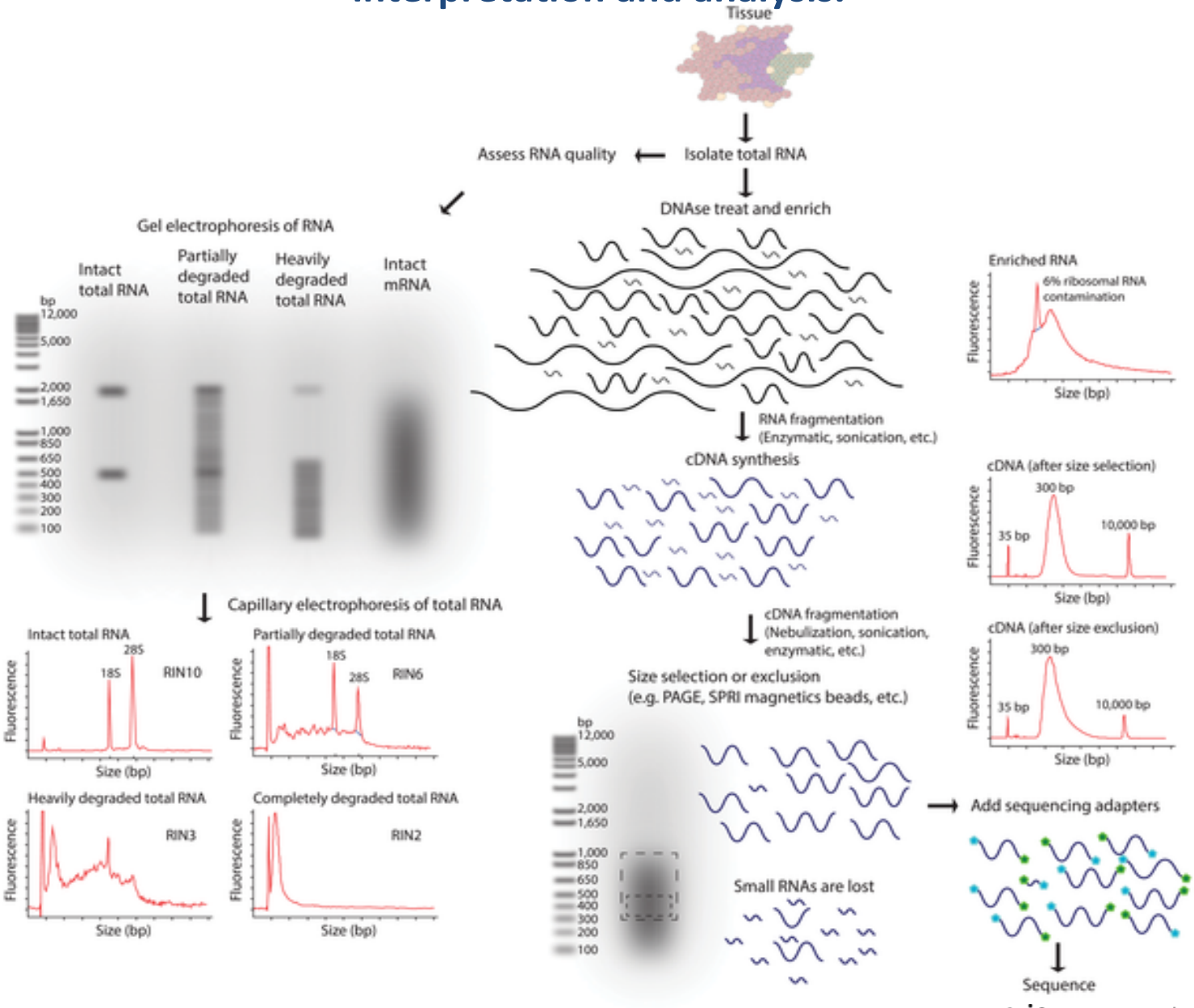
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

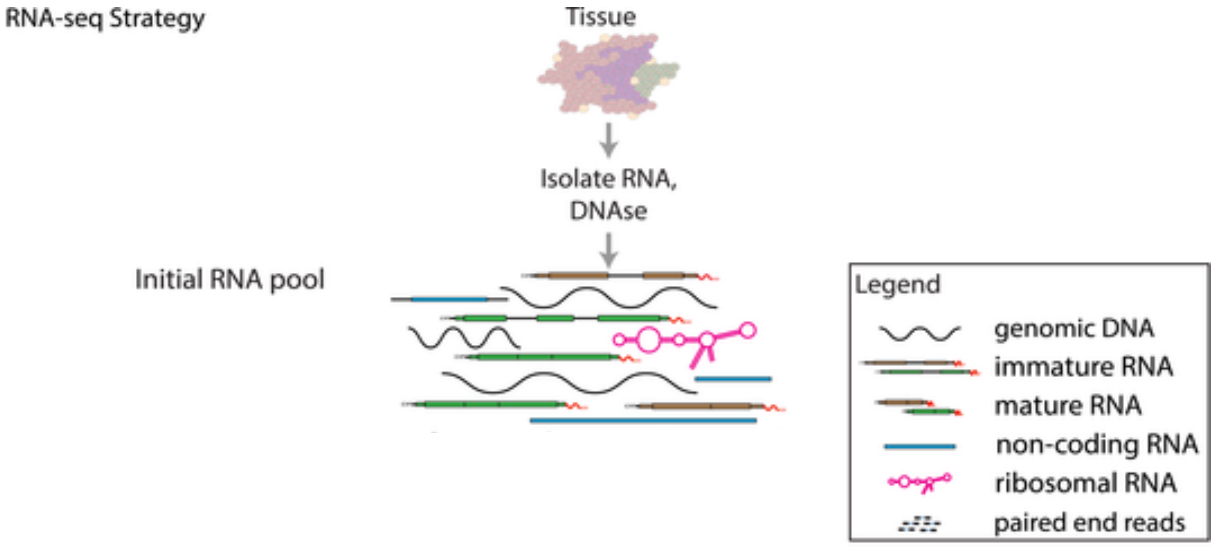
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

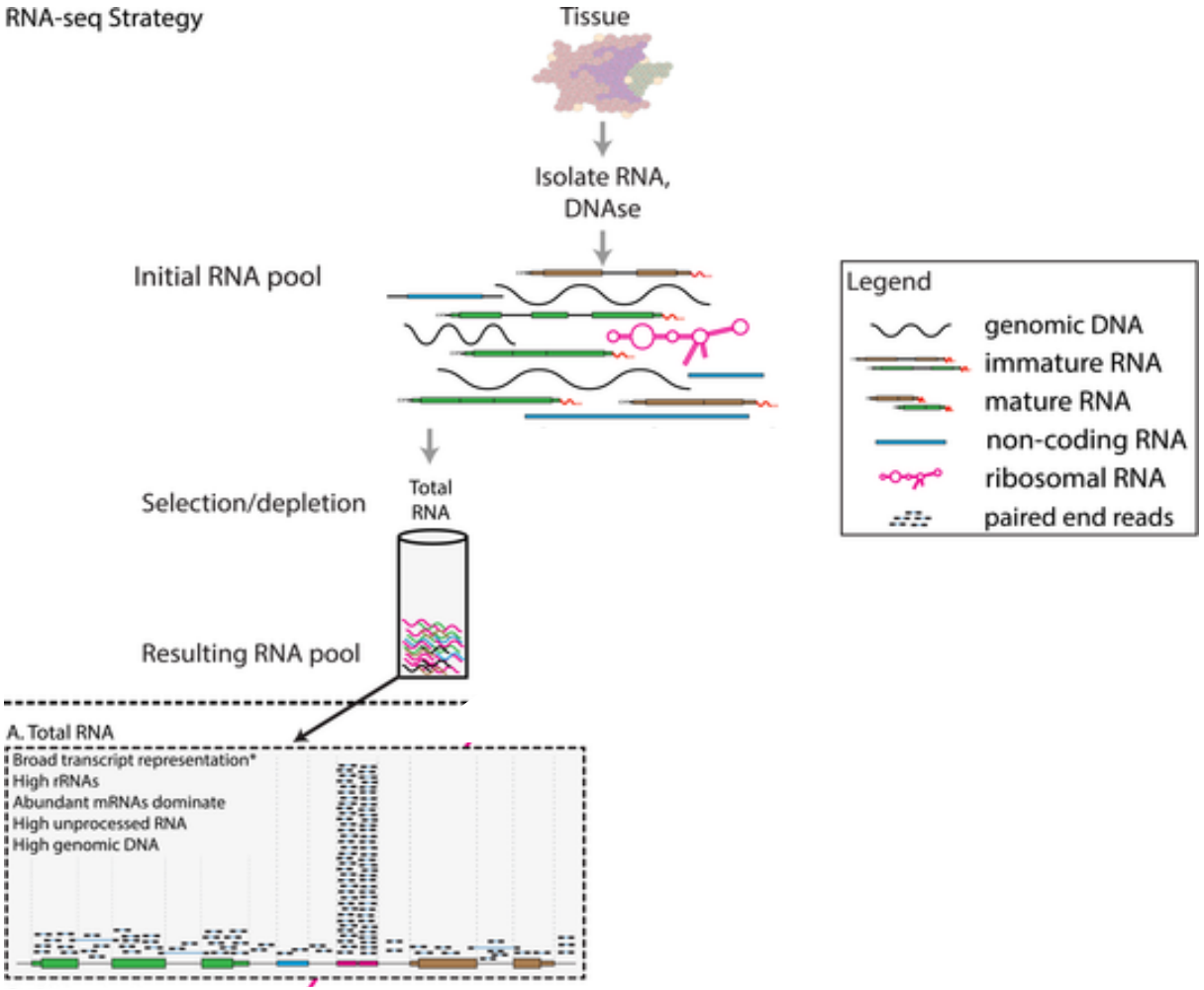
RNA-seq library enrichment strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

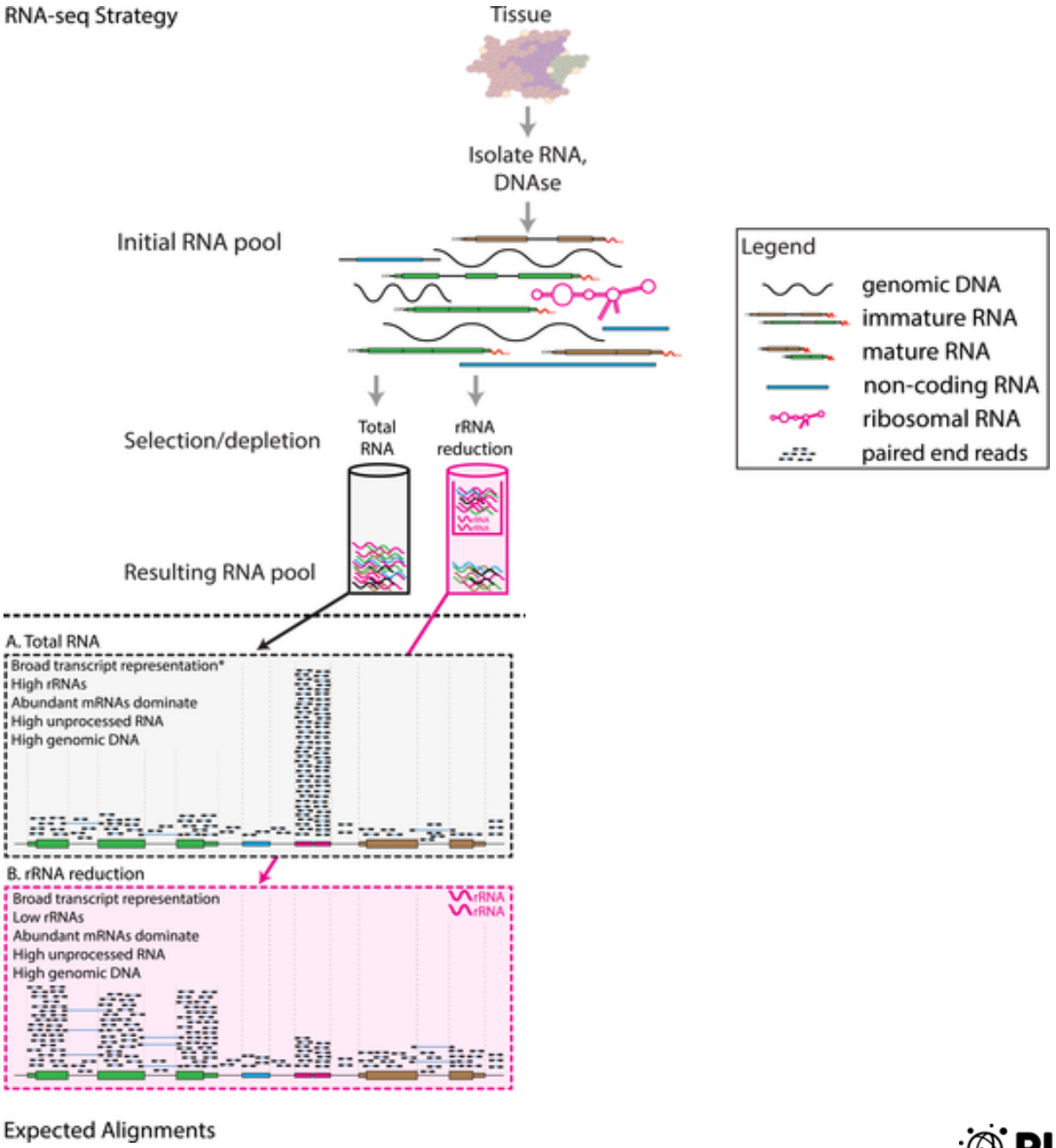
RNA-seq library enrichment strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

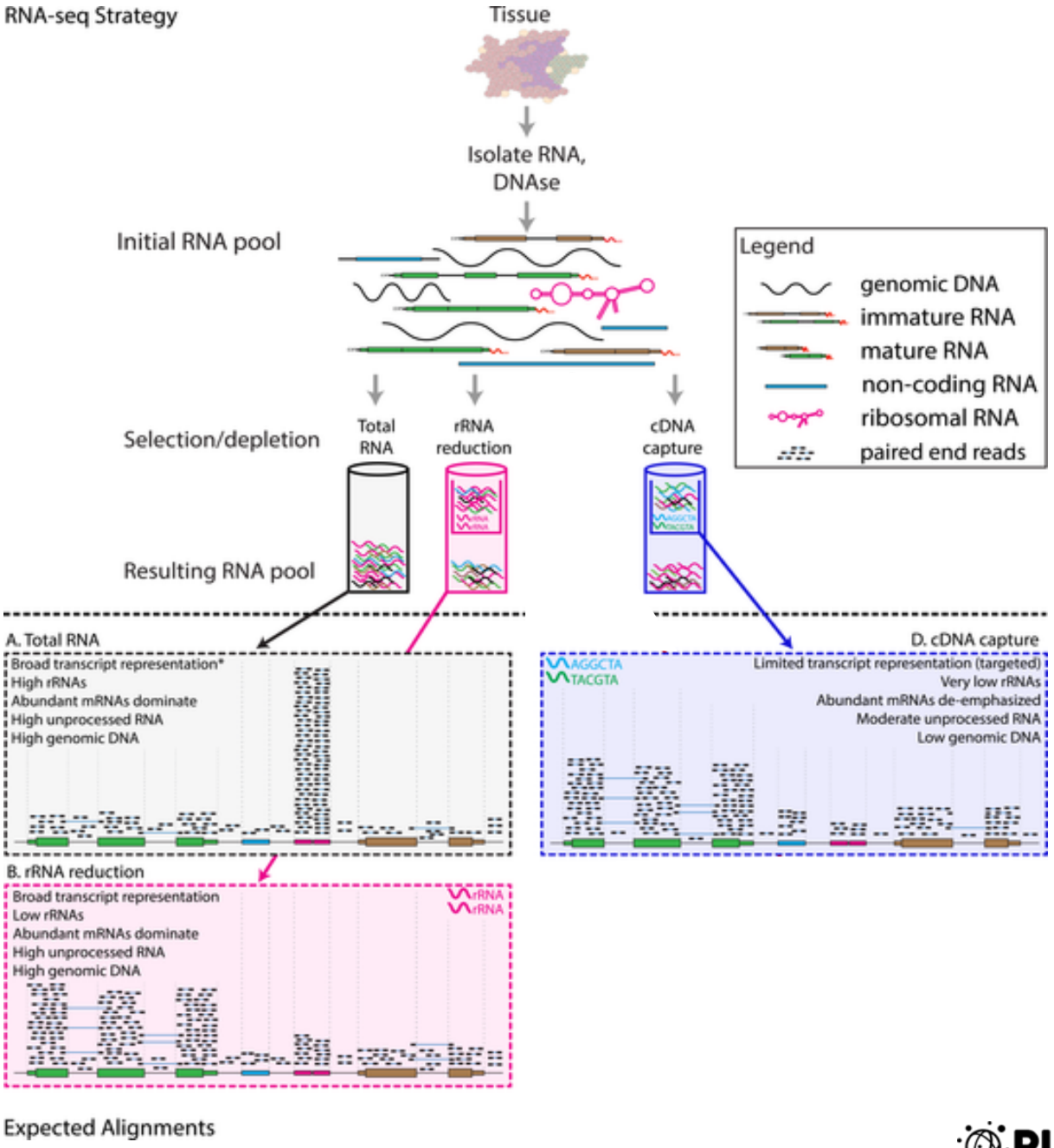
RNA-seq library enrichment strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

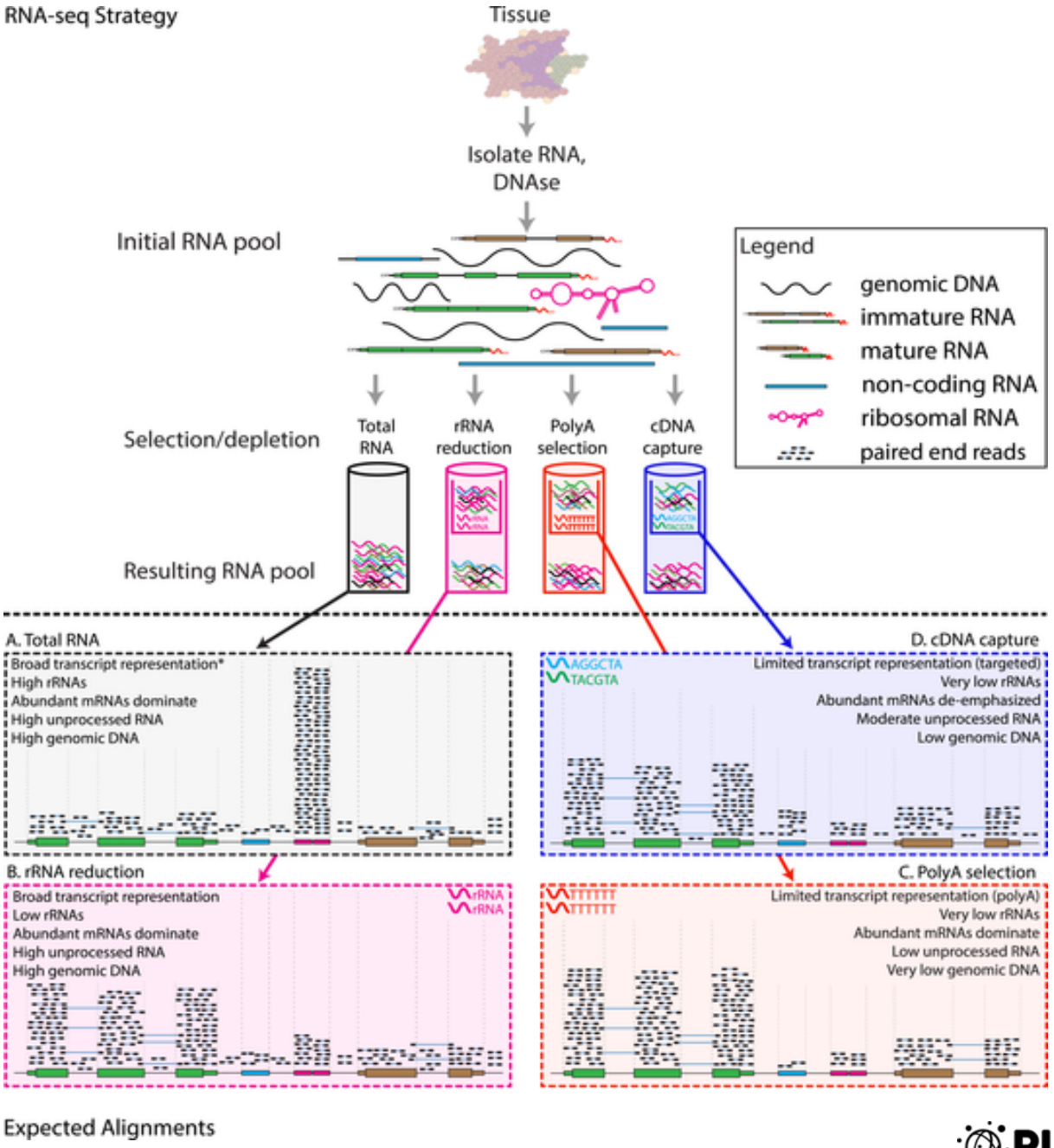
RNA-seq library enrichment strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

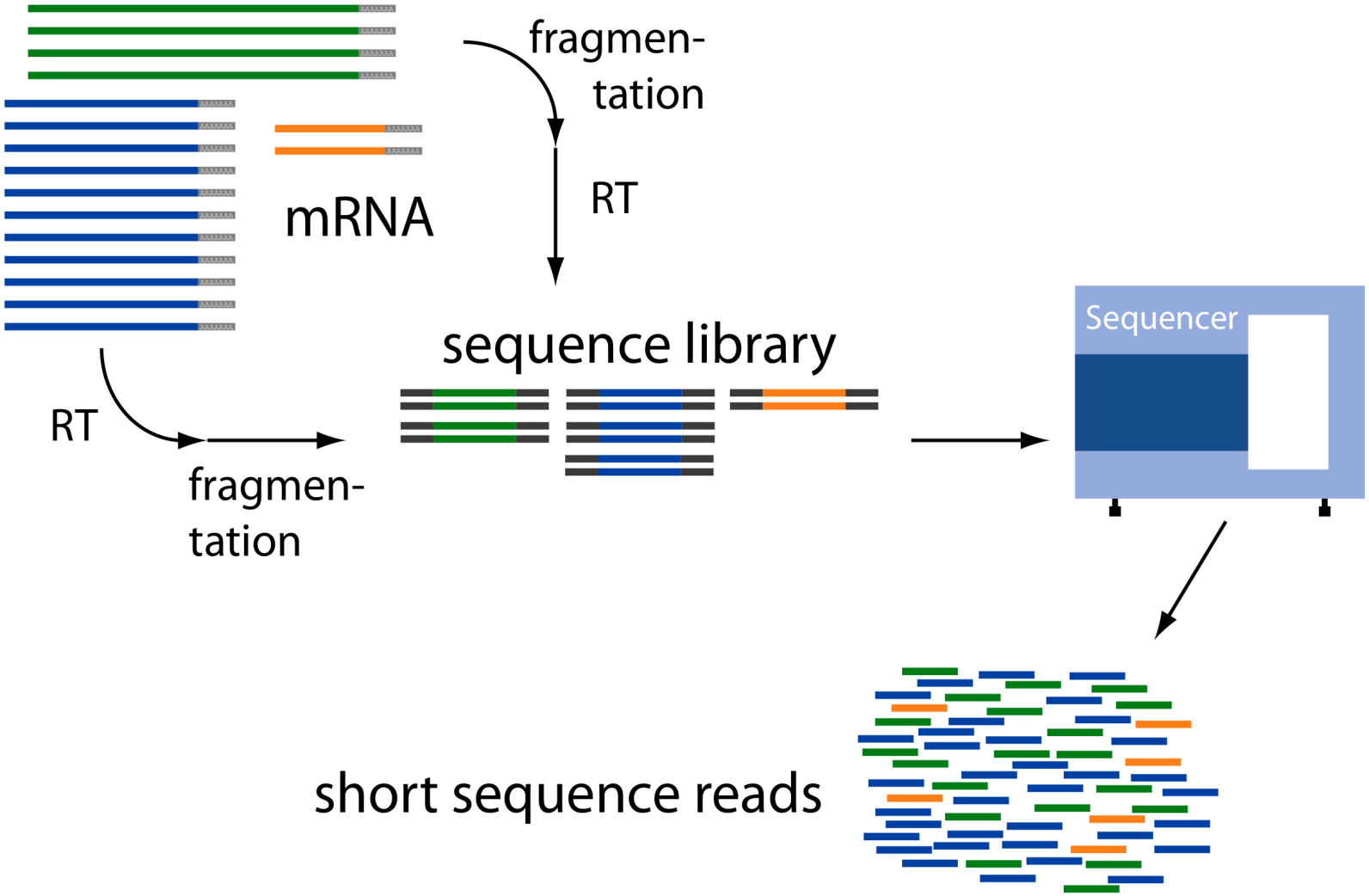
RNA-seq library enrichment strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

PLOS | COMPUTATIONAL BIOLOGY

Overview of RNA-Seq



From: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

Common Data Formats for RNA-Seq

FASTA format:

```
>61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Read

Quality values

Interpreting Base Quality Values

Phred_Quality_Value = AsciiEncodedQual('B') – 33 = 30

Phred_Quality_Value = -10 * $\log_{10}$(Pwrong('T'))

Pwrong('T') = 10^(30/-10) = 10^-3 = 0.001

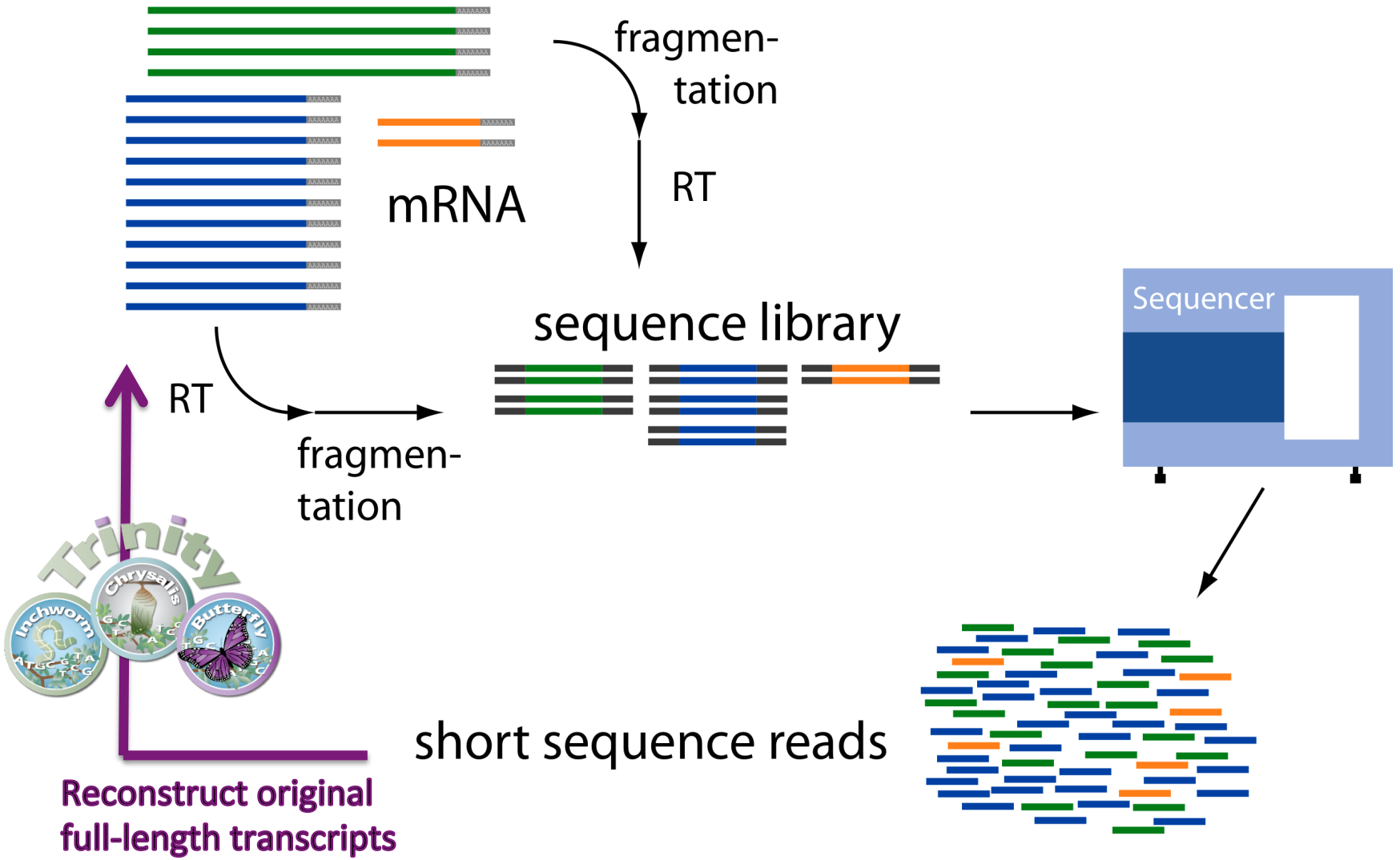
Paired-end Sequences

Two FastQ files, read name indicates left (/1) or right (/2) read of paired-end

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTCAGGATGGAAGAACA
+
C<CCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCC
```

Overview of RNA-Seq



From: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

The Ever-Growing Trinity User Community

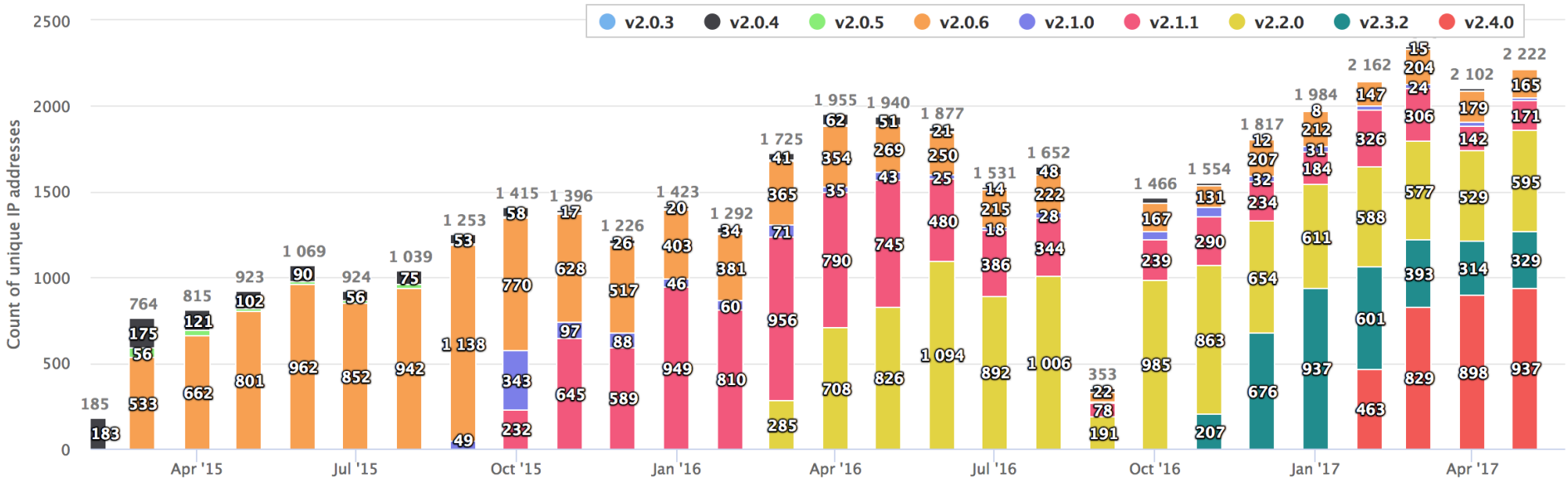
Nature Biotechnology, 2011
Nature Protocols, 2013

- ~3k unique users per month
- ~8k literature citations
- Open Source software development contributions from the Trinity community.

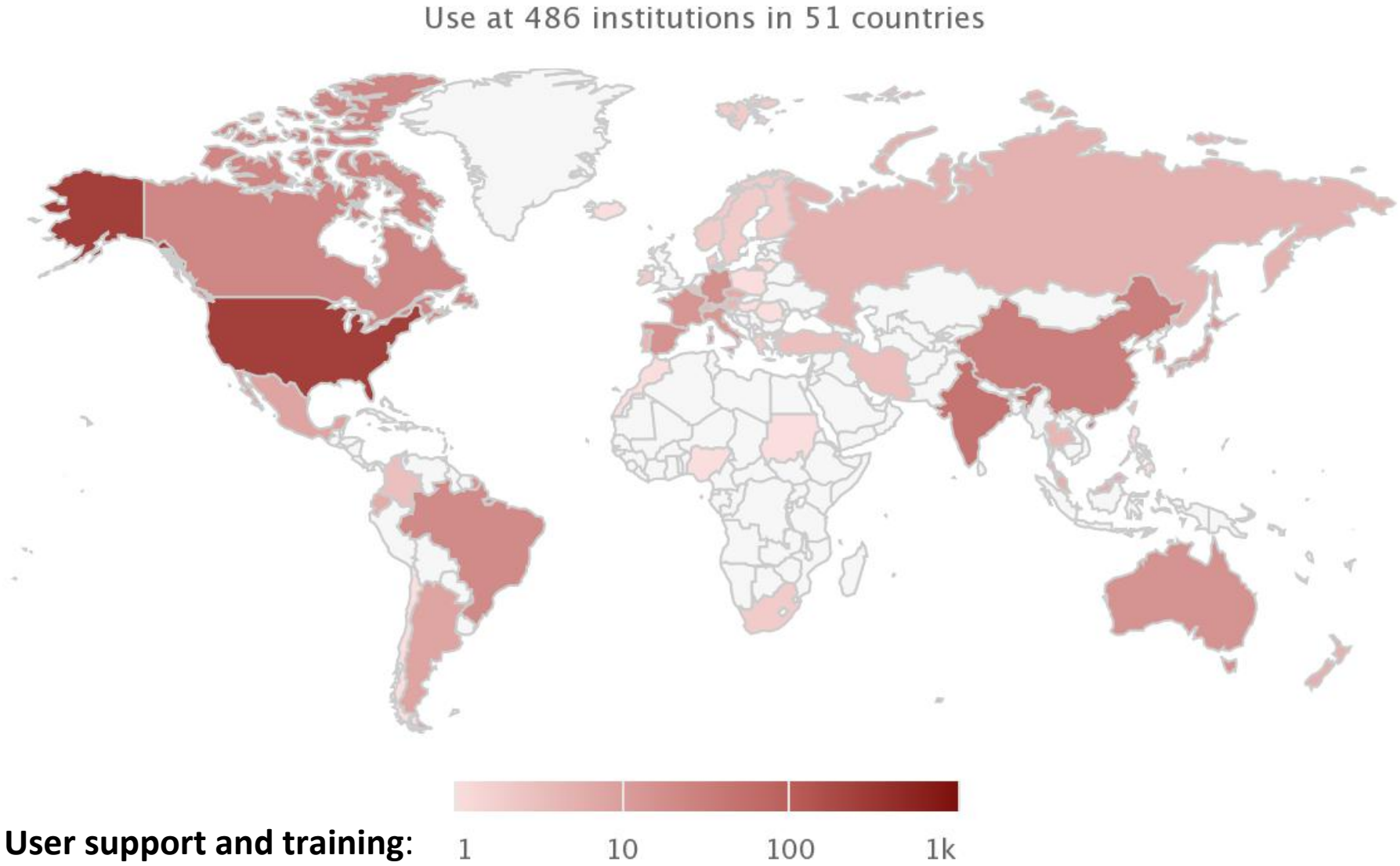
GitHub

http://trinityrnaseq.github.io

Trinity Usage Tracked by Unique IP Address



Trinity Usage is Global



User support and training:

- Google group and Twitter feed for community interaction and support.
- Extensive documentation, user guides, tutorials and protocols
- Demo and training videos
- On-site training workshops

nature protocols

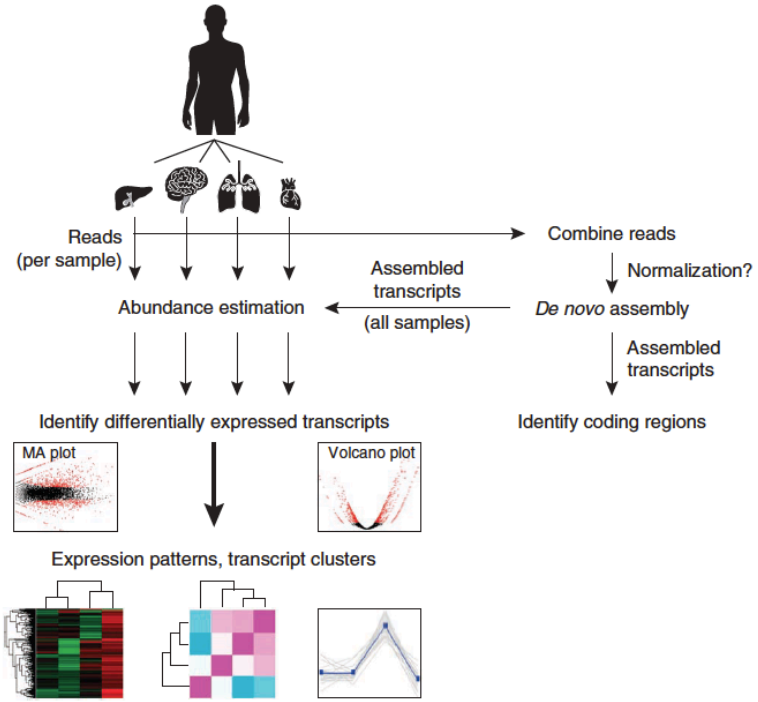
NATURE PROTOCOLS | PROTOCOL

De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

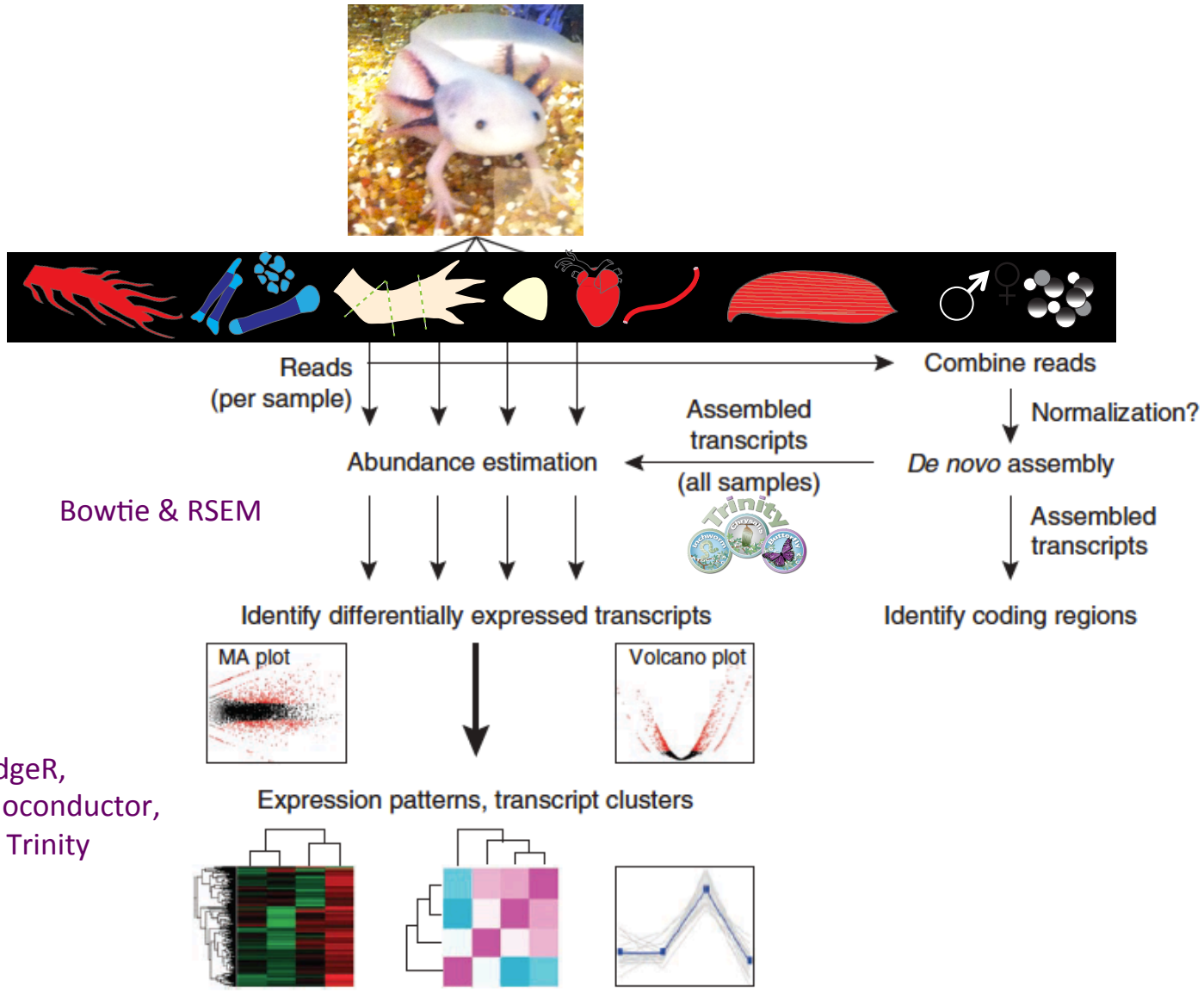
Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

Affiliations | Contributions | Corresponding authors

Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084
Published online 11 July 2013



Framework for De novo Transcriptome Assembly and Analysis



1.3 Billion
Total Reads

86 Million
Normalized Reads

Bowtie & RSEM

EdgeR,
Bioconductor,
& Trinity

RNA-Seq De novo Assembly Using Trinity

Quick Guide for the Impatient

Trinity assembles transcript sequences from Illumina RNA-Seq data.

Download Trinity here.

Build Trinity by typing 'make' in the base installation directory.

Assemble RNA-Seq data like so:

```
Trinity --seqType fq --left reads_1.fq --right reads_2.fq --CPU 6 --max_memory 20G
```

Find assembled transcripts as: 'trinity_out_dir/Trinity.fasta'

Use the documentation links in the right-sidebar to navigate this documentation, and contact our Google group for technical support.

Example Applications of the Trinity RNA-Seq Protocol

Resource

A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors

Donald M. Bryant[1,6], Kimberly Johnson[1,6], Tia DiTommaso[1], Timothy Tickle[2], Matthew Brian Couger[3], Duygu Payzin-Dogru[1], Tae J. Lee[1], Nicholas D. Leigh[1], Tzu-Hsing Kuo[1], Francis G. Davis[1], Joel Bateman[1], Sevara Bryant[1], Anna R. Guzikowski[1], Stephanie L. Tsai[4], Steven Coyne[1], William W. Ye[1], Robert M. Freeman Jr.[5], Leonid Peshkin[5], Clifford J. Tabin[4], Aviv Regev[2], Brian J. Haas[2], , , Jessica L. Whited[1,7],

Original Article

Loggerhead sea turtle embryos (*Caretta caretta*) regulate expression of stress response and developmental genes when exposed to a biologically realistic heat stress

Blair P. Bentley, Brian J. Haas, Jamie N. Tedeschi, Oliver Berry

Got RNA-Seq?

Run Trinity