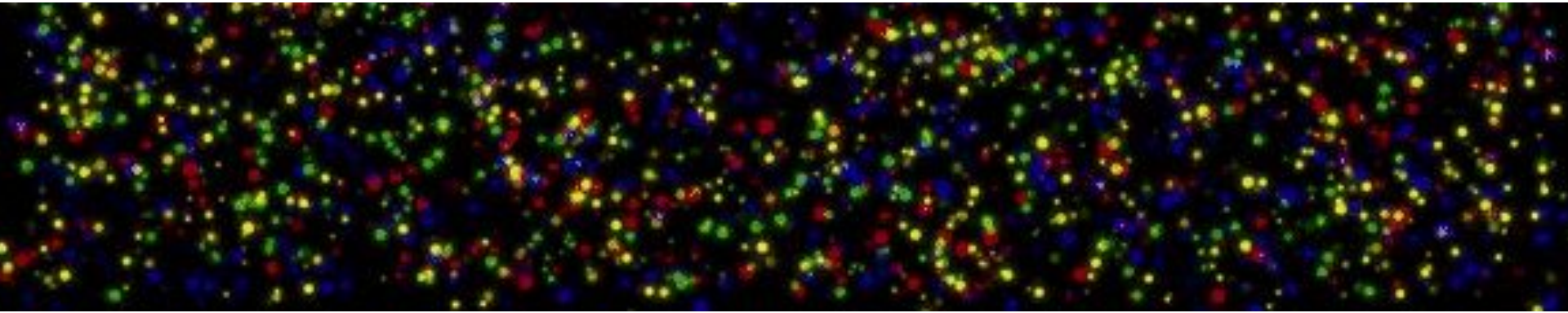
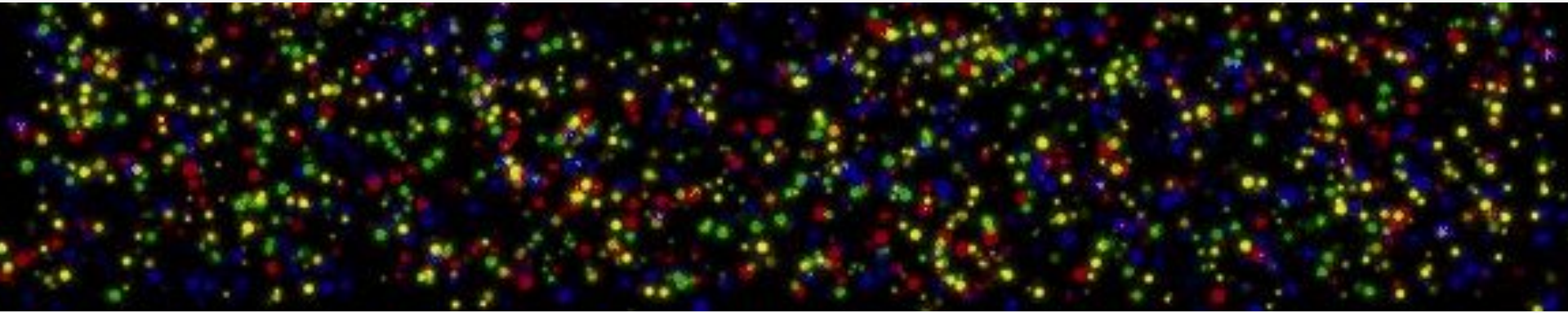
Intro to RNAseq technology and data



Mary-Ann Blätke
JJ Szymanski

Intro to RNAseq technology and data



- Get familiar with jargon of RNAseq/HTS (paired-end, cluster density, fastq, index, etc)
- Understand Illumina's "Sequencing by Synthesis" (SBS) technology
- Set priorities when planning a sequencing experiment
- Understand the basics of library prep for HTS

Paired-end

Clustering density

Flow cell

Q30

fastq

NextSeq

Single-end

fasta

SBS

Library

Read depth

HiSeq

Read length

Indicies

The birthplace of high-throughput sequencing (HTS)

David Klenerman

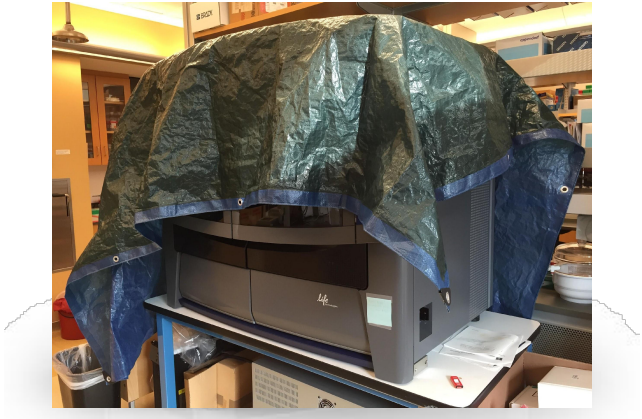
Shankar Balasubramanian

Sequencing circa 2012

Roche 454 ‘pyroseq…’

Life Technologies Ion Torrent

Life Technologies SOLiD



Illumina

The year of sequencing

In 2007, the next-generation sequencing technologies have come into their own with an impressive array of successful applications. Kelly Rae Chi reports.

CREATING THE GENOME ANALYZER

When John West started as CEO of Solexa Ltd. in August of 2004, the longest stretch of DNA that the company could sequence was only six bases long. "That was a little bit intimidating," recalls West, now the vice president and general manager of Illumina's DNA sequencing business unit following the acquisition of Solexa Ltd. by Illumina. "The problem was we never had a commercial platform to be able to sequence it on."

The next-generation sequencing platform they developed and commercialized by June 2006, West says, has been a huge success. Since the commercial release of the platform they have sold 100 instruments and increased the scale of what they have tackled using the technology—from a 5,300-base-pair viral genome to a 150-million-base-pair human X chromosome. But the machine was a challenge to develop. The developers had to bring together key elements of chemistry, people and technology to make it work.

Illumina's (Solexa) Genome Analyzer.

By the time Solexa Ltd. announced its plans to merge with Lynx Therapeutics in August 2004, a lot of the core sequencing-by-synthesis chemistry and molecular biology had already been done, West says. The early chemistry work, done by Solexa Ltd. in the United Kingdom, led to the creation of a reversible terminator nucleotide and a polymerase that would incorporate it. Solexa Ltd. and Lynx Therapeutics had bought cluster technology for solid phase amplification together from the Swiss company Manteia SA, and did some instrumentation design.

Still, the company was in a state of flux in late 2004 and needed to figure out how to combine the chemistry and technology into a complete system. The researchers needed to meet and brainstorm, operating on an eight-hour time difference between the two branches, one in the UK and one in the US. Ligation chemistry developed by Lynx Therapeutics was an option. There were differing opinions about which chemistry, ligation or polymerase would work best, but ultimately they took a gamble on the Solexa polymerase because it could complete reactions faster. "It was risky at the time because it wasn't all proven," West remembers. "We basically put all our eggs in one basket."

And it worked. It became easier, with the cluster technology and polymerase chemistry, to increase the sequence lengths. By February 2005, they increased the read length to 25 bases. They sequenced the 5,300-base-pair virus PHIX174 genome, the same genome that Sanger first sequenced. In October 2005, they sequenced an 180,000-base-pair bacterial artificial chromosome.

Illumina acquired Solexa, Inc. in January of last year, and commercial sales increased, West says. In 2007, the company completed sequencing the human X chromosome and has since moved on to sequencing the human genome using paired-end sequencing. "I think [2007] has been a great year for us," West says. *KRC*

The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Ugrappa Nagalakshmi,[1]* Zhong Wang,[1]* Karl Waern,[1] Chong Shou,[2] Debasish Raha,[1] Mark Gerstein,[2,3] Michael Snyder[1,2,3]†

The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3′-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

6 JUNE 2008 VOL 320 **SCIENCE**

Illumina sequencing

MiniSeq System | MiSeq Series | NextSeq Series | HiSeq Series | HiSeq X Series | NovaSeq Series

Data output

Instrument cost

Per base cost

MiniSeq

- well-suited for a lab
- not sufficient for transcriptomics
- Applications include:
  - ✓amplicon sequencing (16S)
  - ✓bacterial genome seq
  - ✓QC of RNAseq libraries

MiSeq

NextSeq

0.4 - 1.1 billion reads
150bp paired-end
24 hr run time

NovaSeq

- best fit for a large core facility
- requires a dedicated technician
- complicated set-up
- Applications include:
  - ✓Just about anything

Democratization of sequencing

MiniSeq System

MiSeq Series

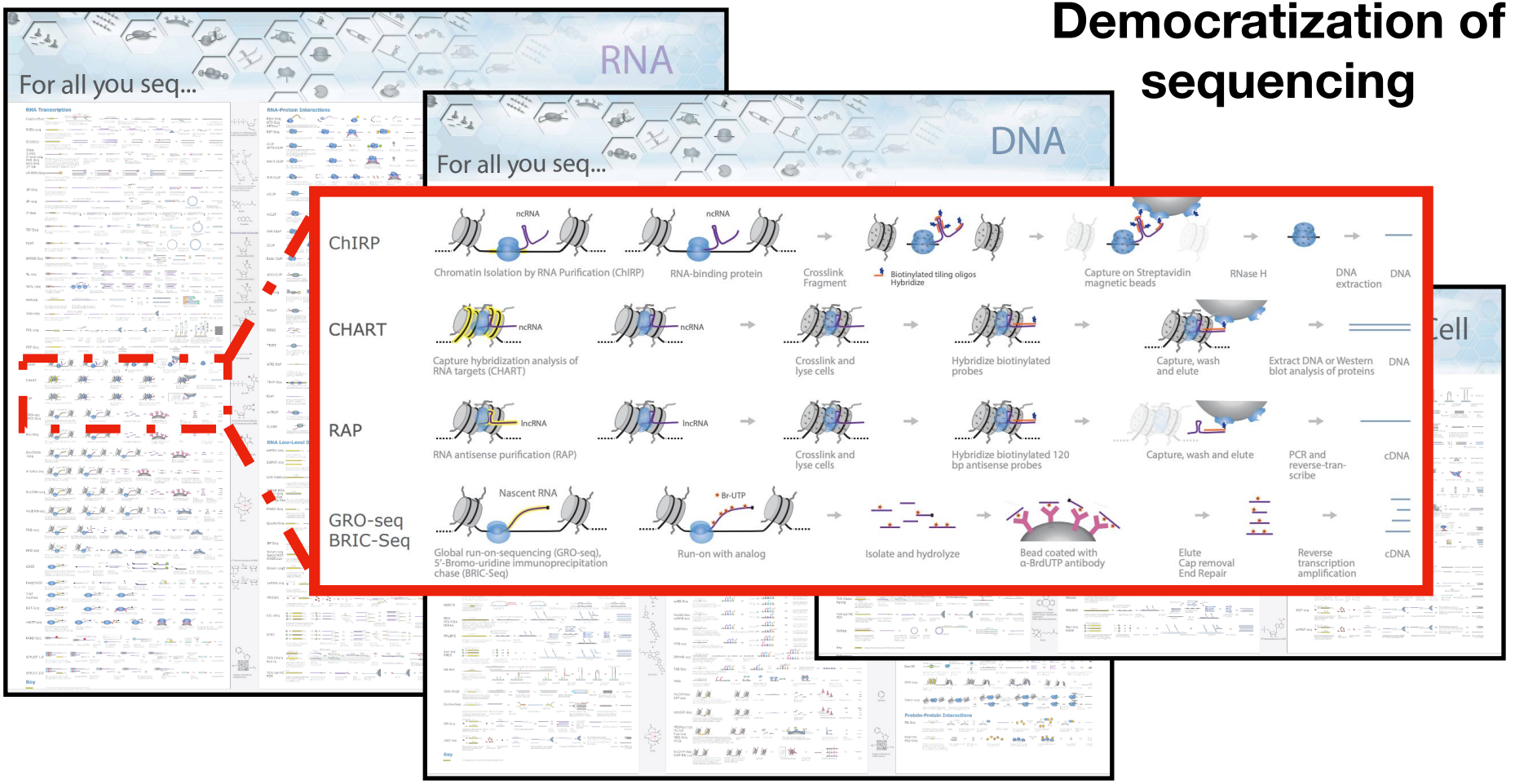
NextSeq Series

HiSeq Series

HiSeq X Series

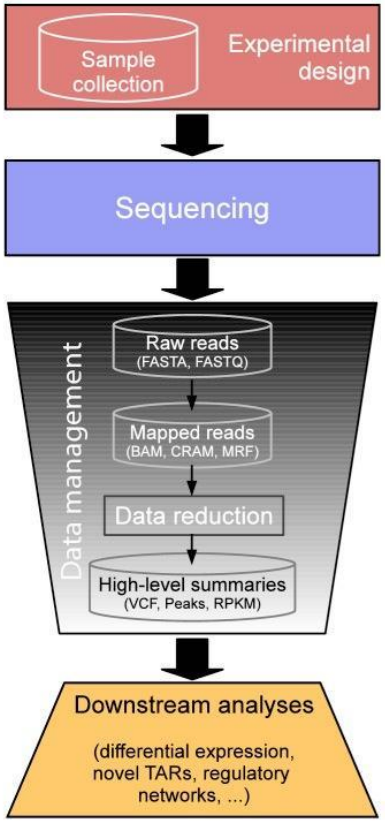
NovaSeq Series

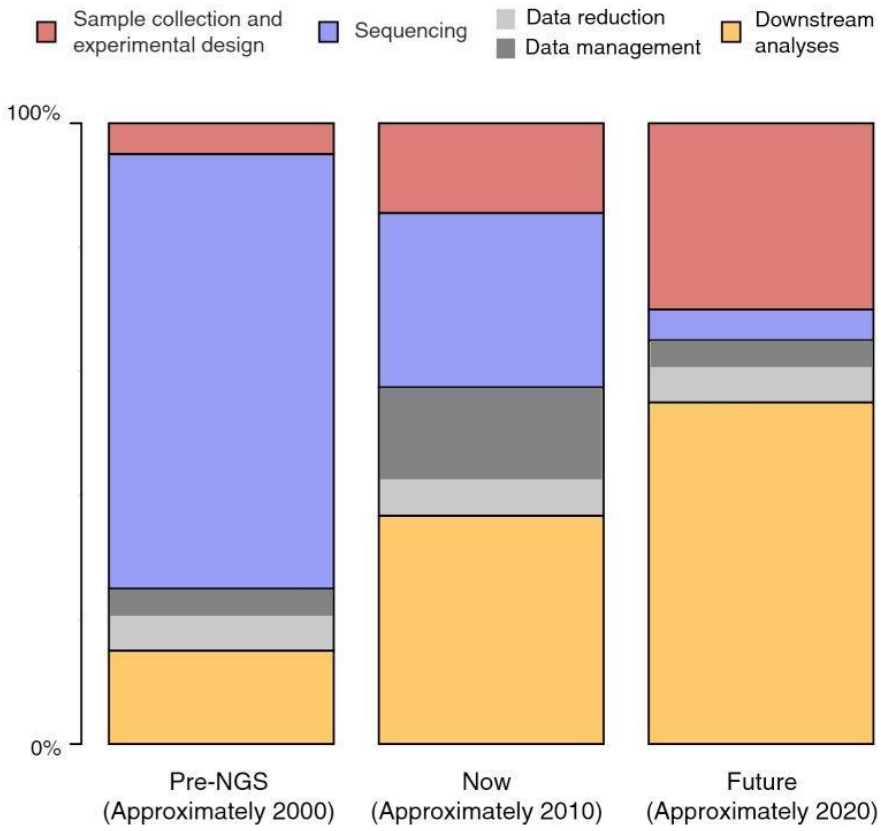
Democratization of sequencing



bit.ly/4ALLUSEQ

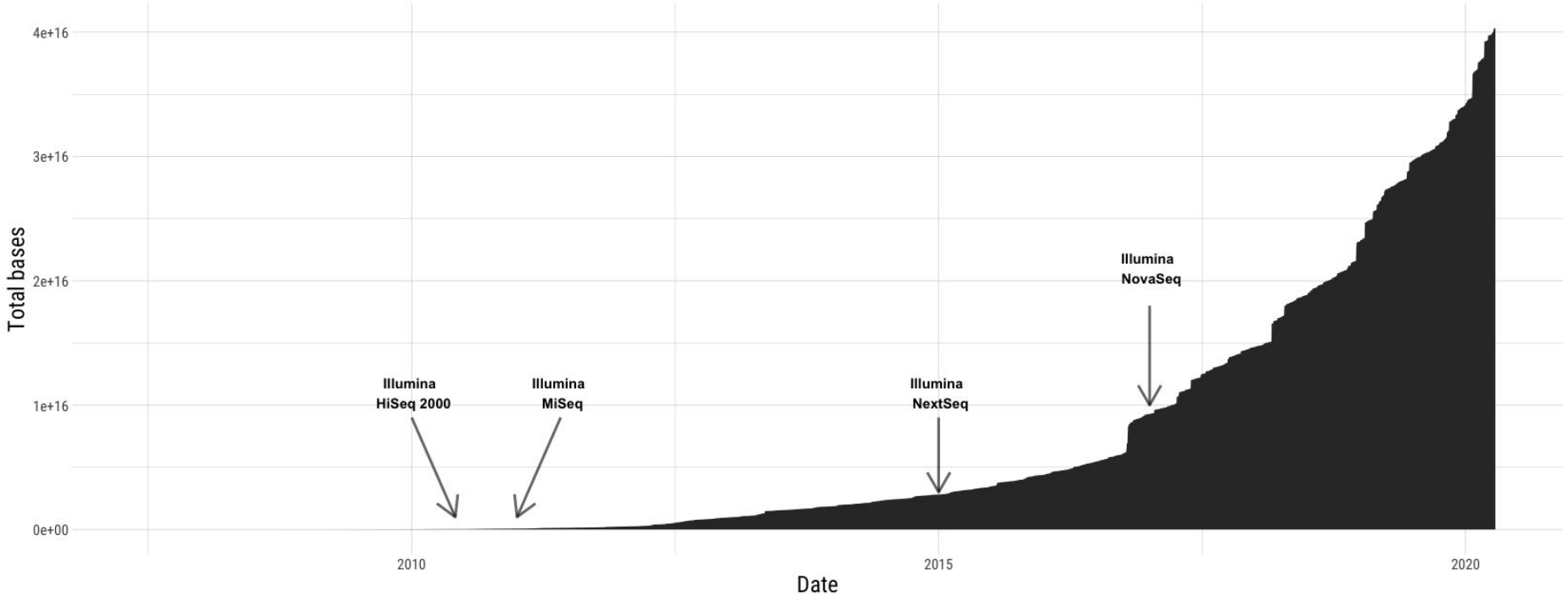
A shift in how time and money are spent on genomics experiments





Sboner A., Genome Biology, 2011

NCBI Sequence Read Archive (SRA)



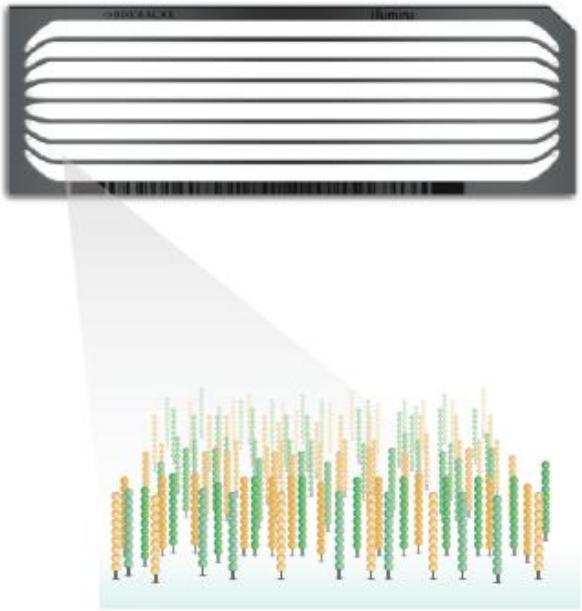
Illumina sequencing

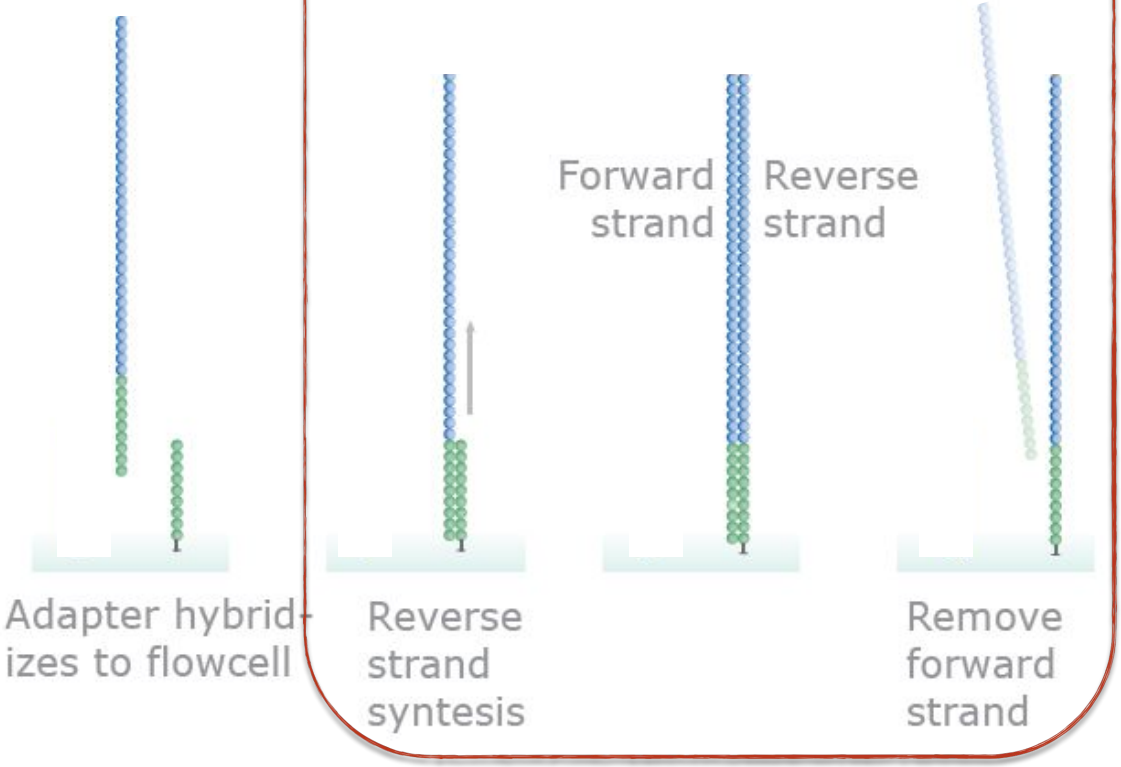
MiniSeq System | MiSeq Series | NextSeq Series | HiSeq Series | HiSeq X Series | NovaSeq Series

‘Sequencing By Synthesis’ (SBS) technology

Sequencing by Synthesis

major technological hurdle

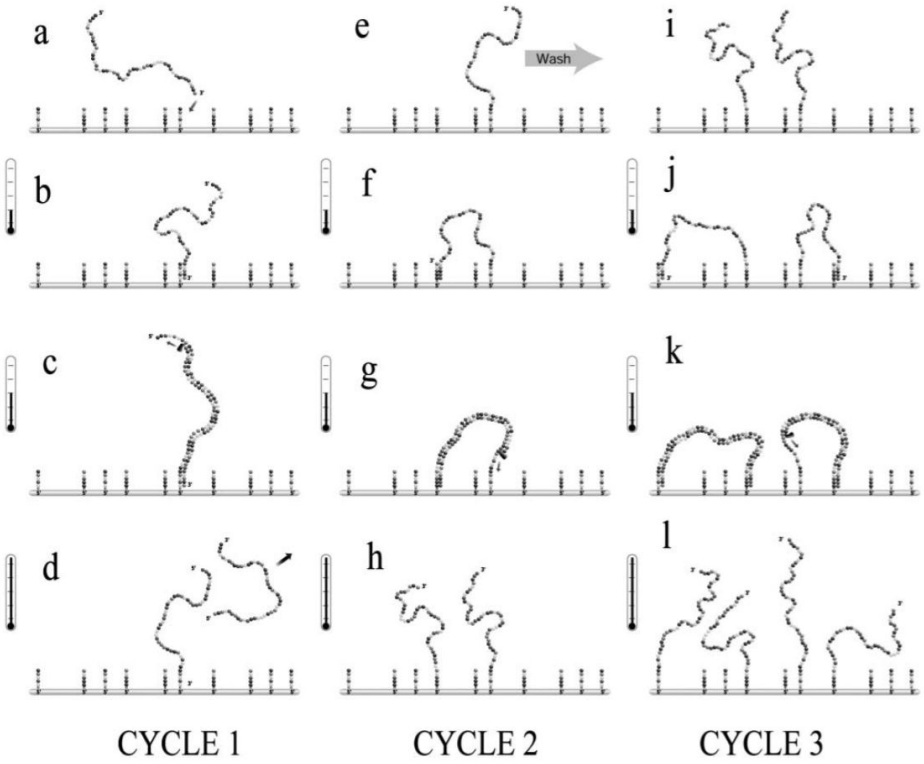




Biophysical Journal Volume 85 October 2003 2075–2086

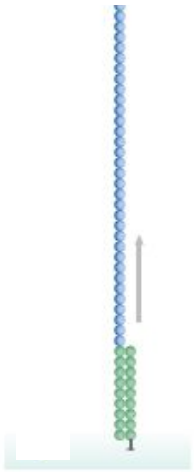
Solid Phase DNA Amplification: A Simple Monte Carlo Lattice Model

Jean-Francois Mercier,* Gary W. Slater,* and Pascal Mayer†
*Department of Physics, University of Ottawa, Ottawa, Ontario, Canada; and †Manteia Predictive Medicine S.A., Coinsins, Switzerland

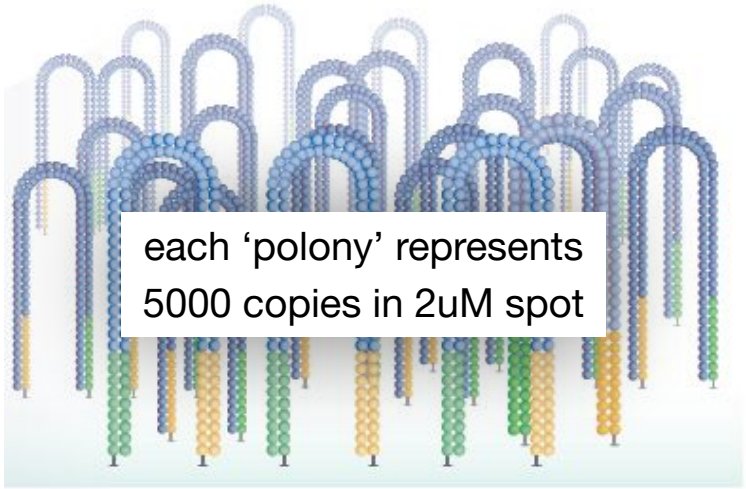


"**solid phase DNA amplification leads to the growth of a colony of molecules attached to the surface and located in the same region**. This characteristic could easily be exploited in the design of DNA microarrays."

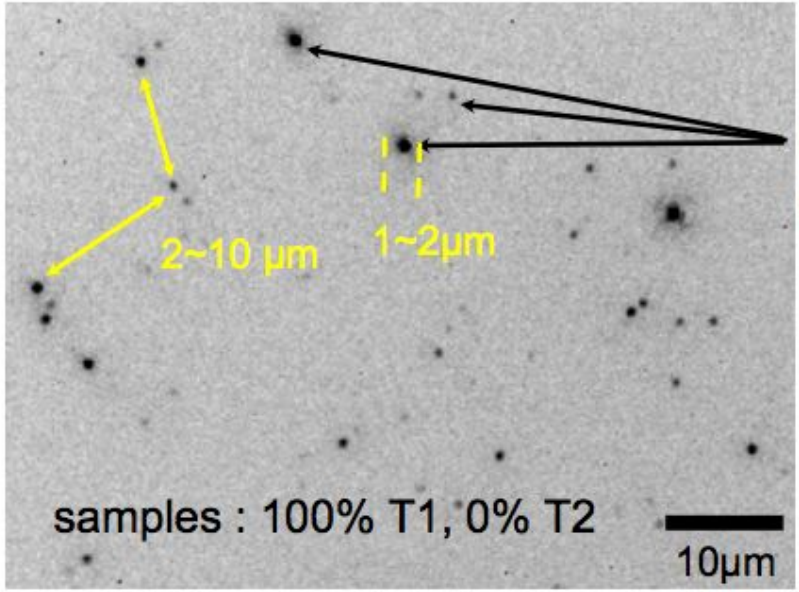
Sequencing by Synthesis (SBS)



PCR colony (aka, ’polony’) size and distribution can be precisely controlled

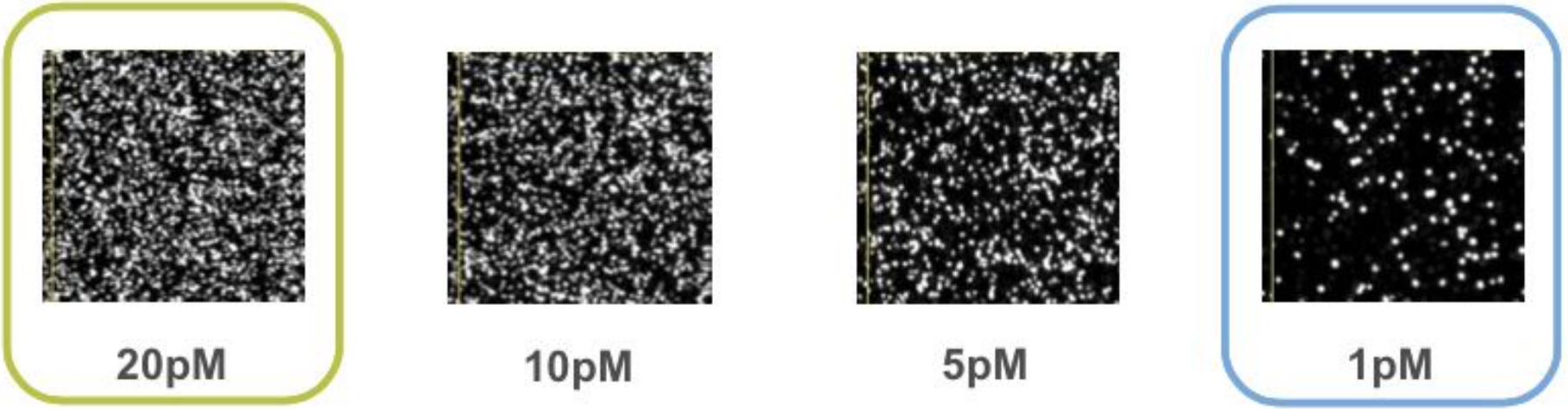


Thousands of molecules are amplified in parallel

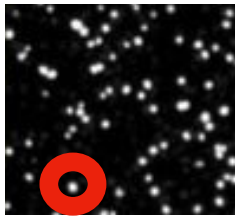


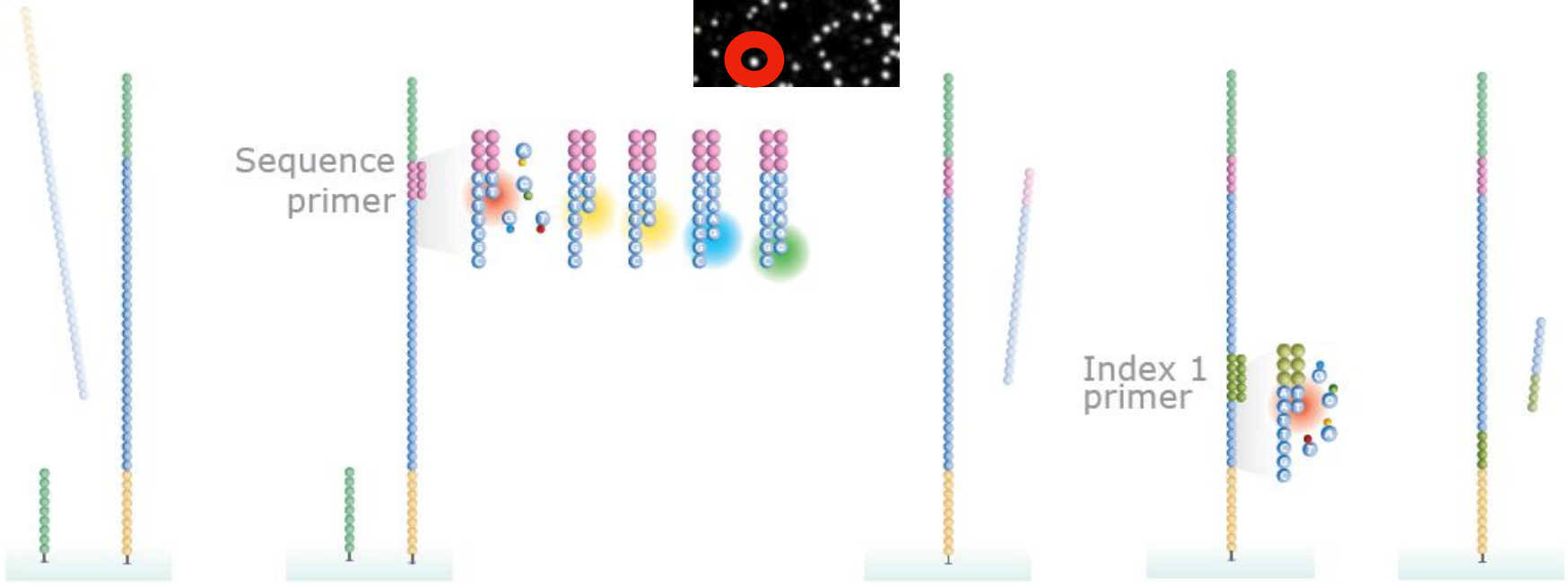
Mayer et al., Presentation 1998

Clustering density is key to data output and quality



HiSeq: 950-1050K clusters/mm^2
NextSeq: 170-220K clusters/mm^2





The reverse strand is cleaved and washed away

With each cycle, four fluores-cently tagged nucleotides compete for addition to the growing chain. Only one is incorporated based on the sequence of the template.

The read product is washed away

Sequence Index1

The read product is washed away

bit.ly/Illumina_SBS

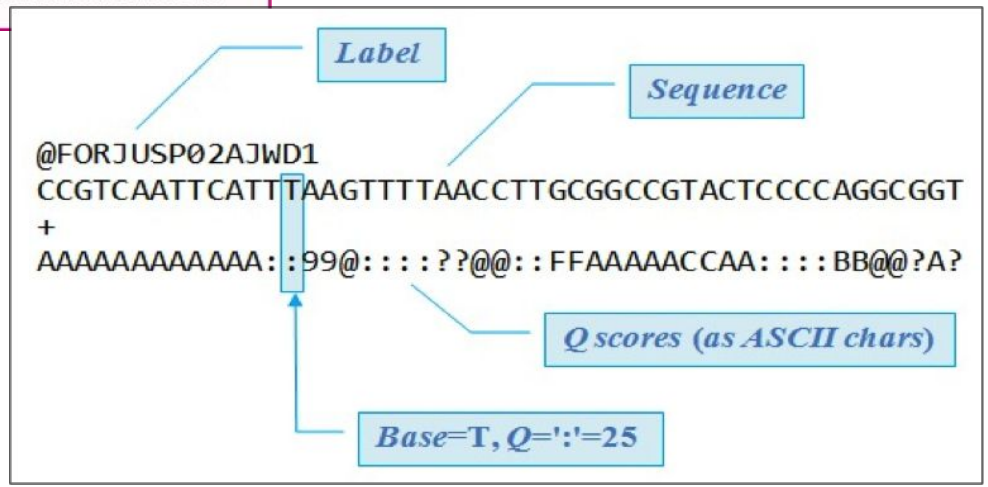
High-throughput sequencing data is stored in .fastq format

Fasta

Header
Sequence
Header
Sequence
Header
Sequence

```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

Fastq



These are just text files!

ASCII characters

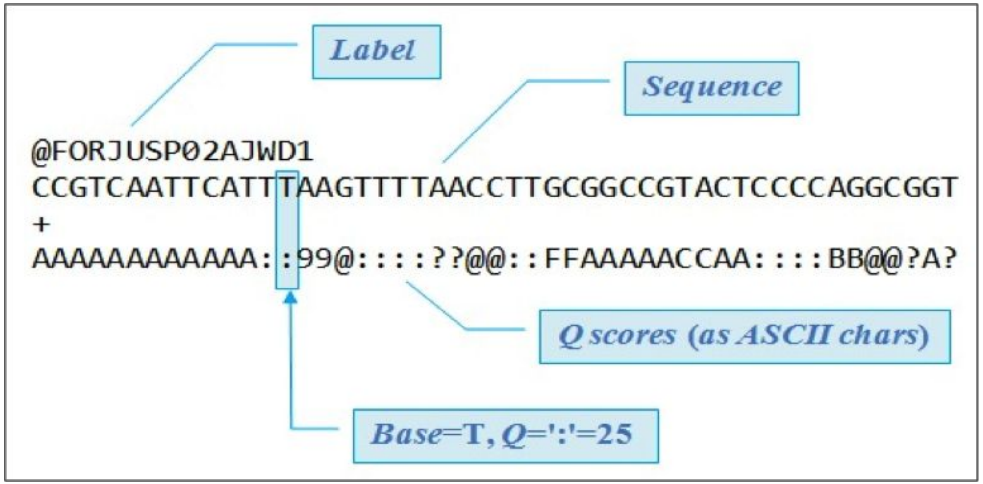
!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	@	A	B	C	D	E	F	G	H	I
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

Quality score

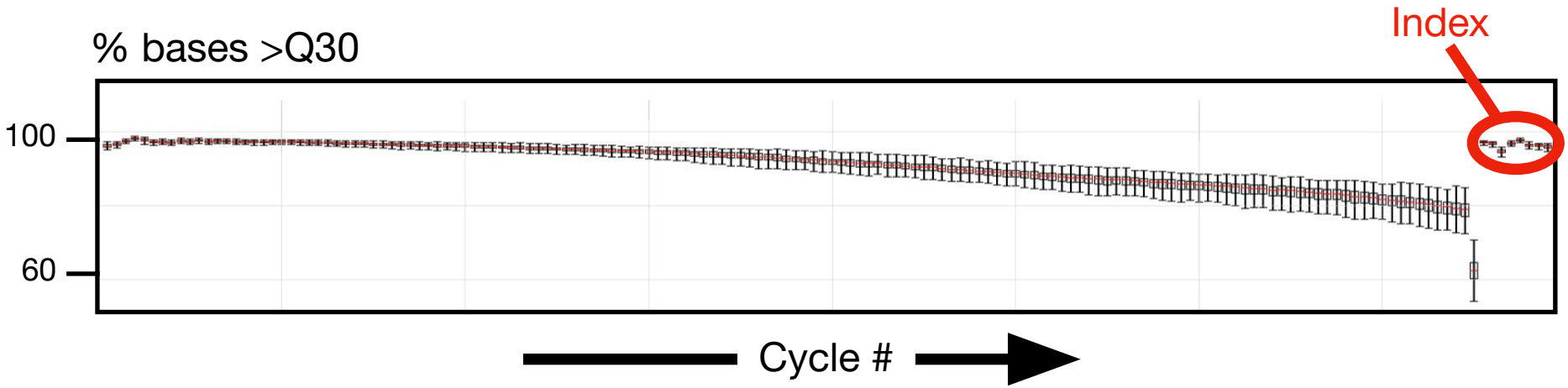
1/1000

1/10000

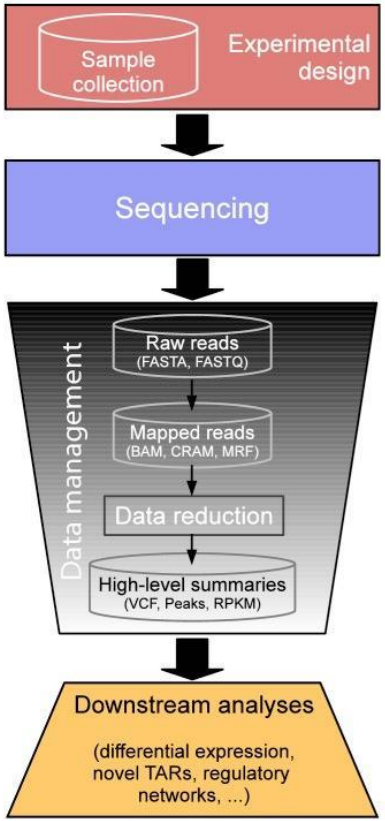
Fastq

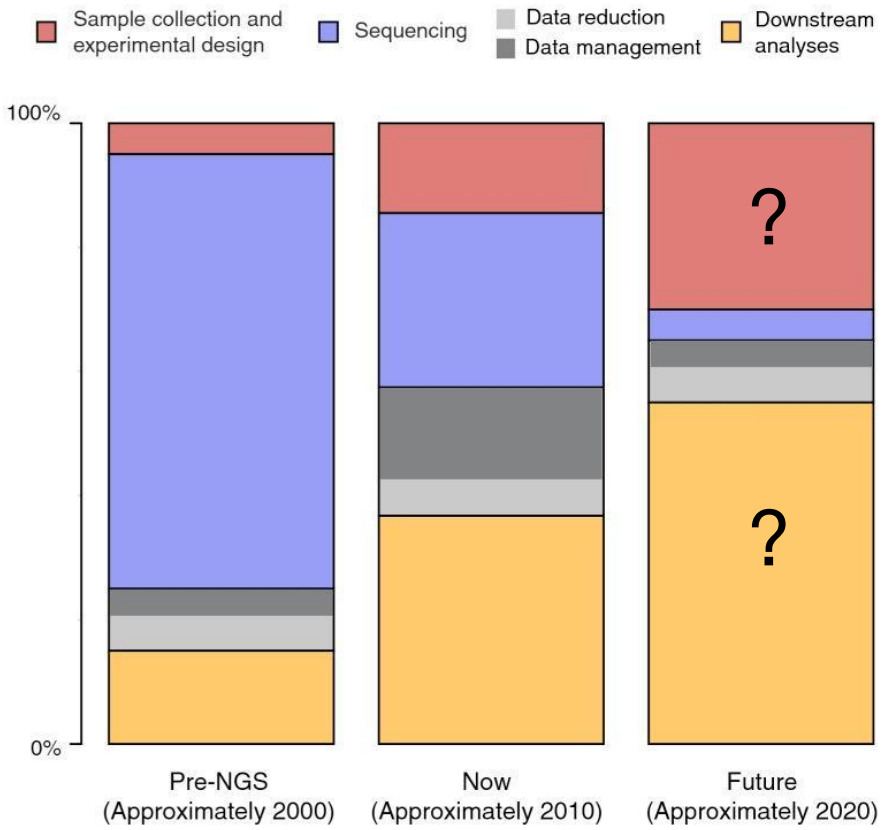


Illumina sequences are short for a reason



A shift in how time and money are spent on genomics experiments





Sboner A., Genome Biology, 2011

Costs associated with bulk* sequencing

Library Preparation

TruSeq kit	cat#	cost
stranded total RNA LT (w/ RiboZero)	RS-122-2201	$5,067
stranded mRNA LT	RS-122-2101	$2,430

all kits process 48 samples

Sequencing

mid output
130M clusters

high output
400M clusters

cycles	cat#	cost
300	FC-404-2004	$4,222
150	FC-404-2002	$2,635
75	FC-404-2005	$1,374

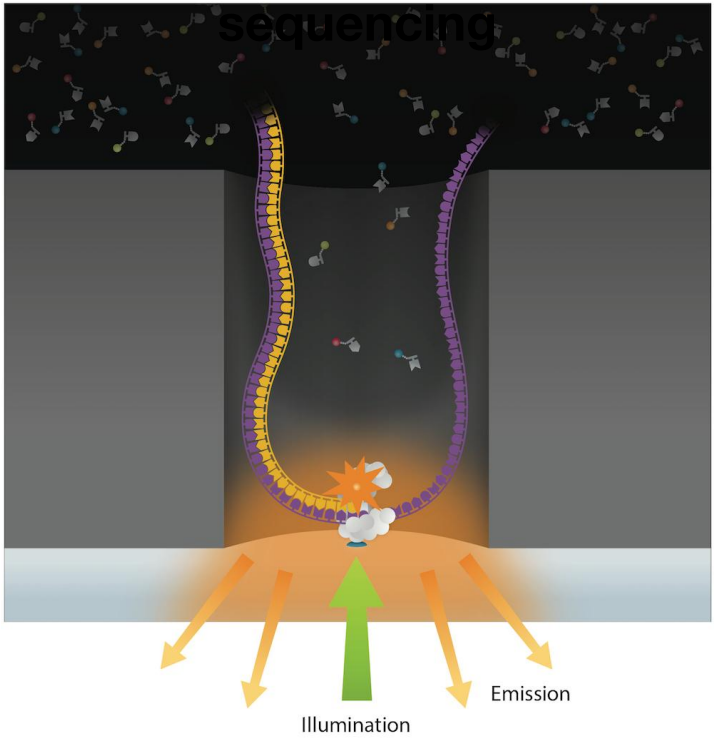
24 samples
library prep = $51/sample
sequencing = $58/sample
data output = 15M reads/sample

**Costs associated with scRNAseq are much more*

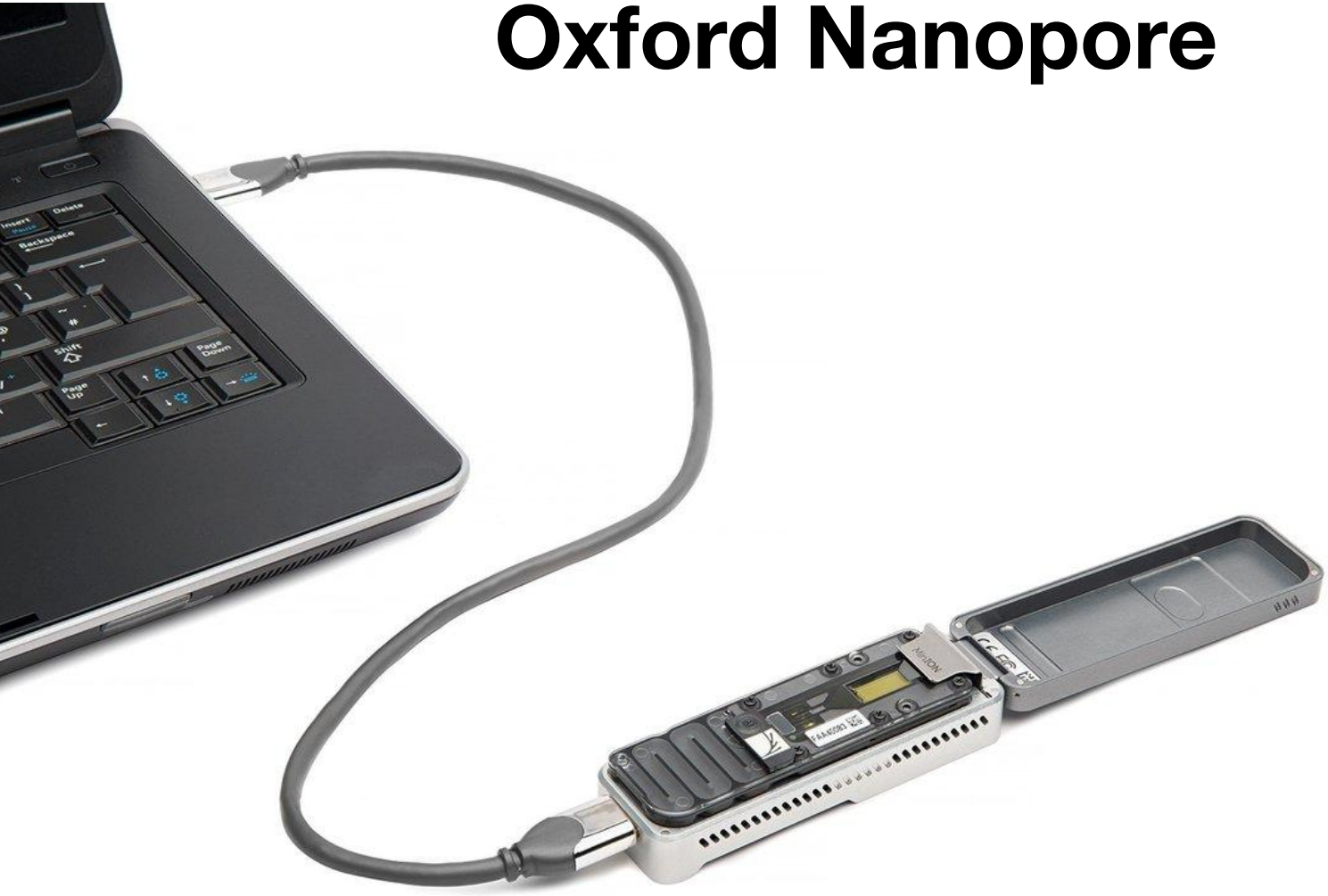
Long read sequencing

PacBio Sequel 2

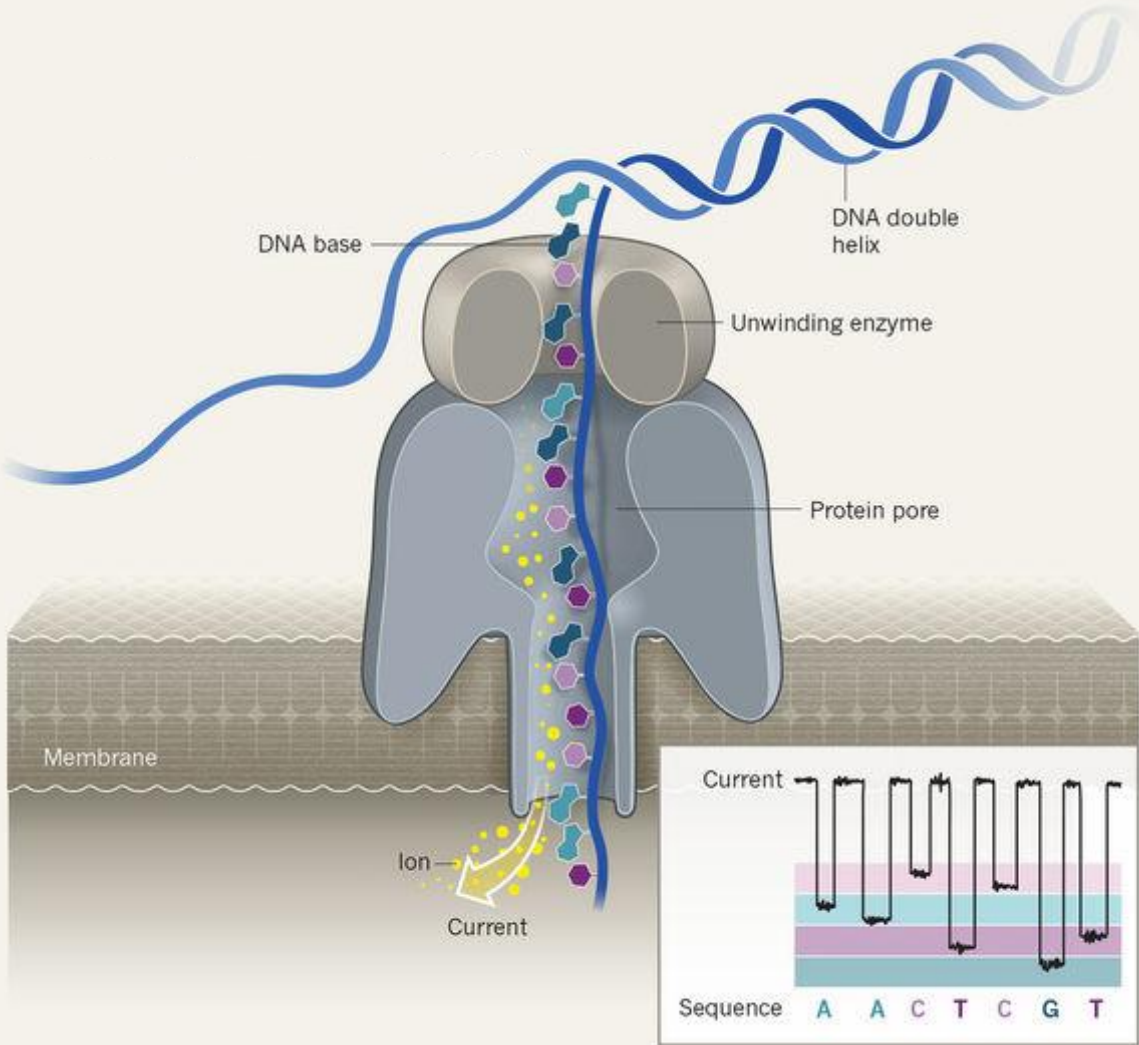
SMRT sequencing



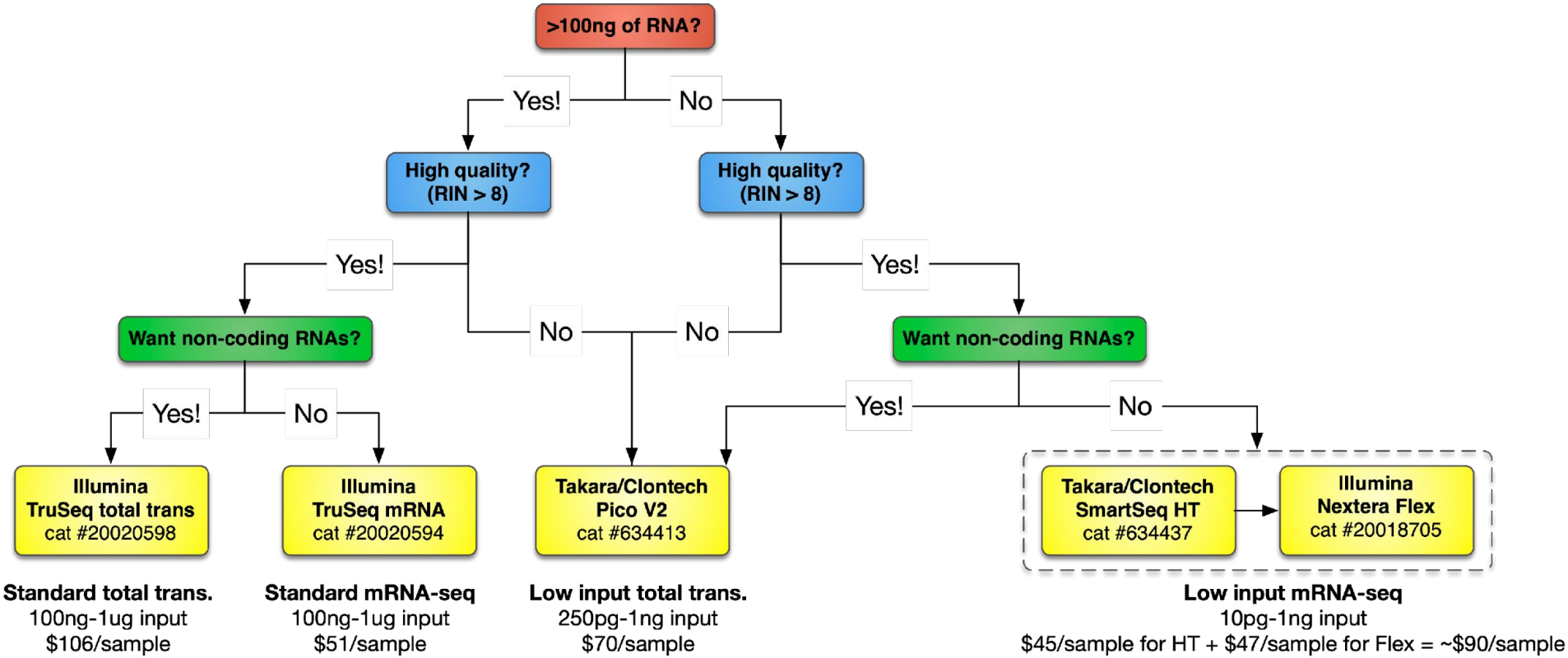
Oxford Nanopore



Nanopore Sequencing



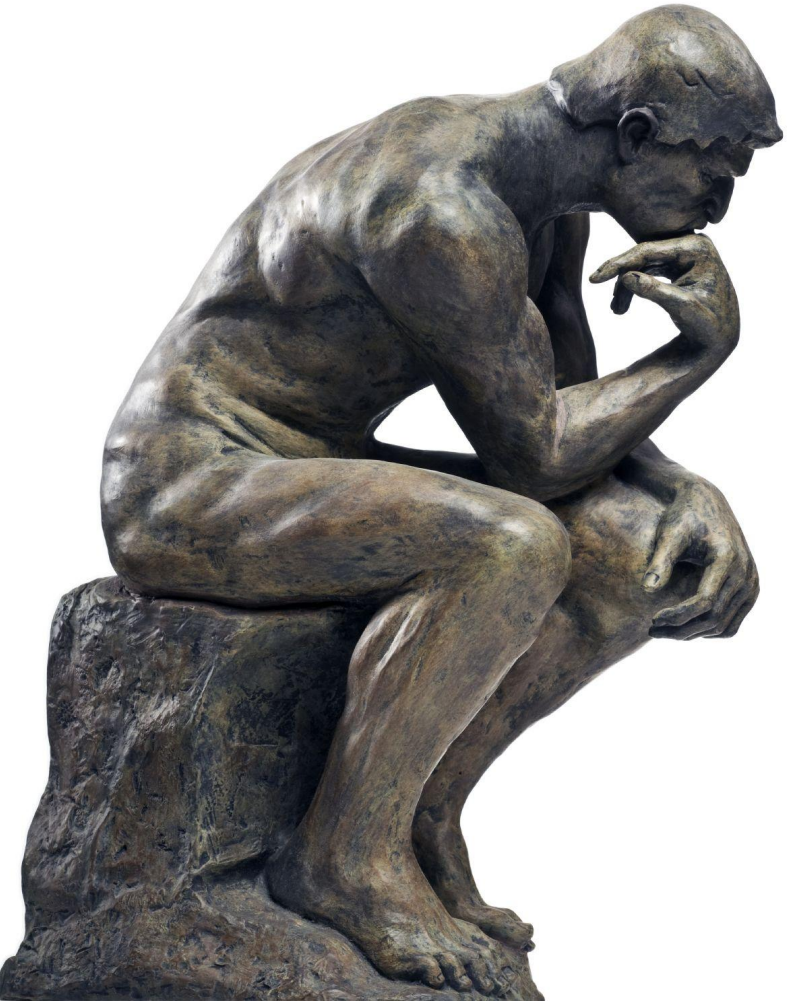
Decision tree for RNAseq library prep



The End

Library preparation

- All sequencing experiments begin by using molecular techniques at the bench to construct a catalog or library of oligonucleotides present in a sample.
- Reagents and steps of library prep are organized in kits, and choice of kit will depend on what you want to do/ask.
- Steps may differ depending on end goal, but many features of library prep are shared (***e.g.***, DNA as output, fragmentation, adapter ligation, barcoding of samples)
- QC is critical before, during and after library prep.
- Major technological advances in library prep protocols/kits for special applications (***e.g.***, FFPE-seq, low-input, single cell sequencing)
- The length of library prep (1-3 days), low volume pipetting and chemistry, make this step particularly prone to error and batch effects



Experiment planning

- Study design
- Library prep and sequencing
- Analysis

RNAseq FAQs

Analysis

1. How do I check the quality of my reads?
2. Which aligner should I use?
3. Am I using the proper statistical methods?
4. What are the best ways to visualize results?

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

RStudio®

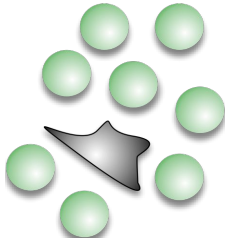
RNAseq FAQs

study design

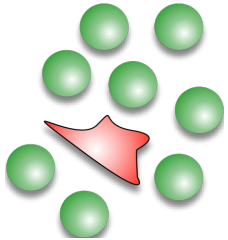
1. **Independent evaluation that your experiment worked?**
2. **Controls will depend on the Q's**
3. **What insight is to be gained from each treatment/condition?**
4. **Signal to noise ratio for your cell type**

Signal to noise

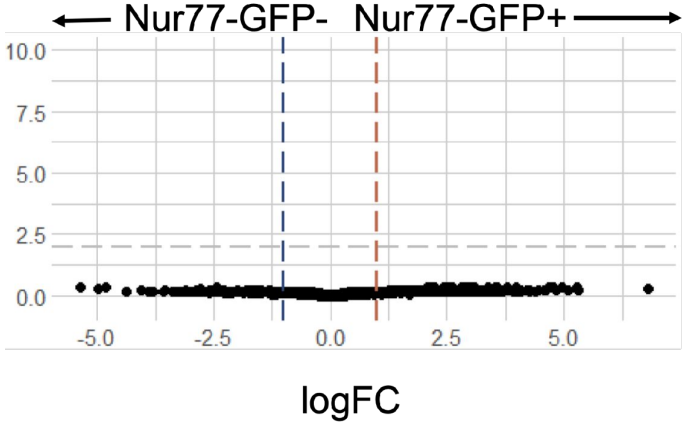
Condition A



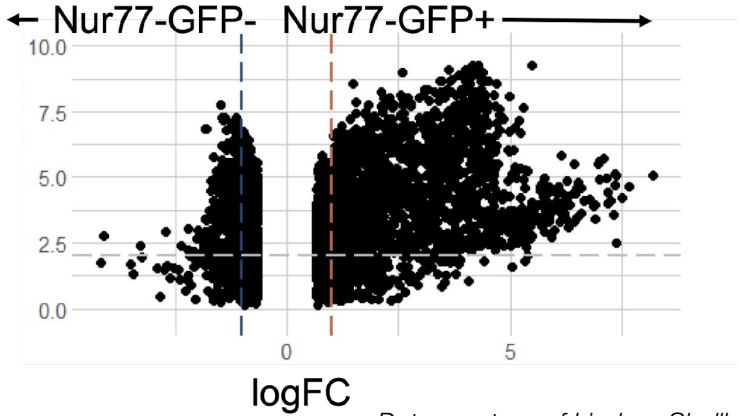
Condition B



Single sorted
DEGs: 0

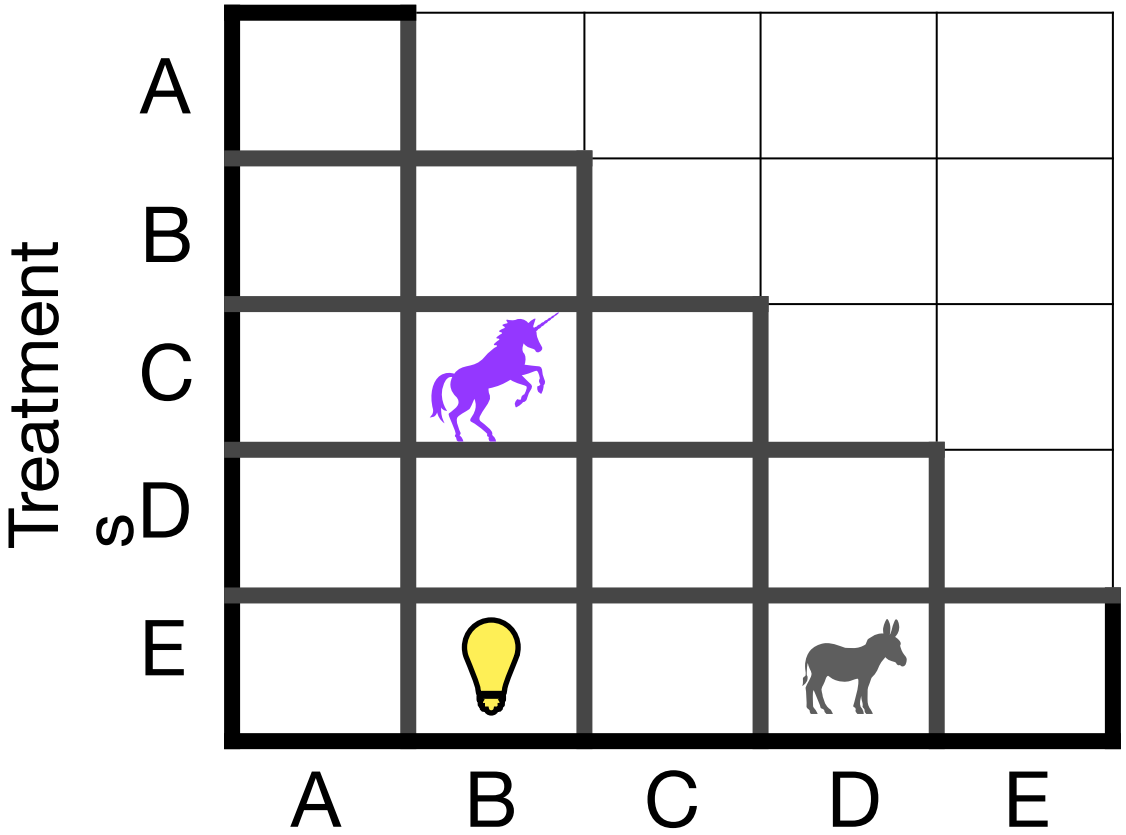


Double sorted
DEGs: 1096



Data courtesy of Lindsey Shallberg (Hunter lab)

What insight is to be gained from each treatment or condition?



Total number of pairwise comparisons

(a)(a-1)/2

(5)(5-1)/2 = 10

RNAseq FAQs

sequencing

1. **How many replicates are needed?**
2. **How deep should I sequence (reads)?**
3. **How much will it cost?**

Deeper sequencing or more replicates?

BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 3 2014, pages 301–304
doi:10.1093/bioinformatics/btt688

Gene expression

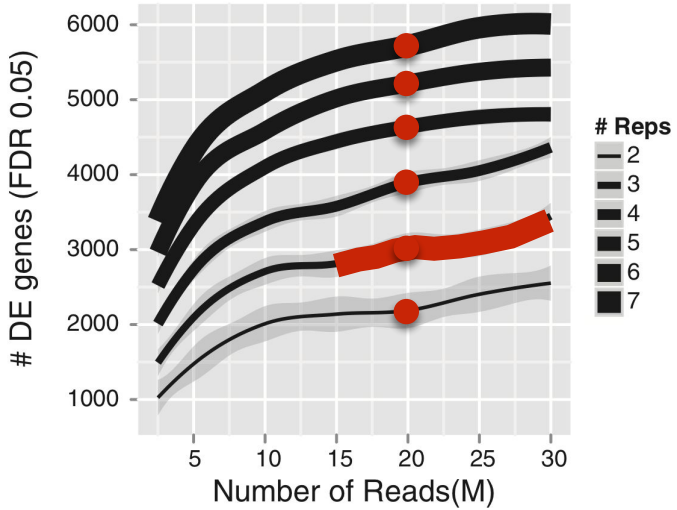
Advance Access publication December 6, 2013

RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Associate Editor: Janet Kelso



Deeper sequencing or more replicates?

BIOINFORMATICS **DISCOVERY NOTE**

Vol. 30 no. 3 2014, pages 301–304
doi:10.1093/bioinformatics/btt688

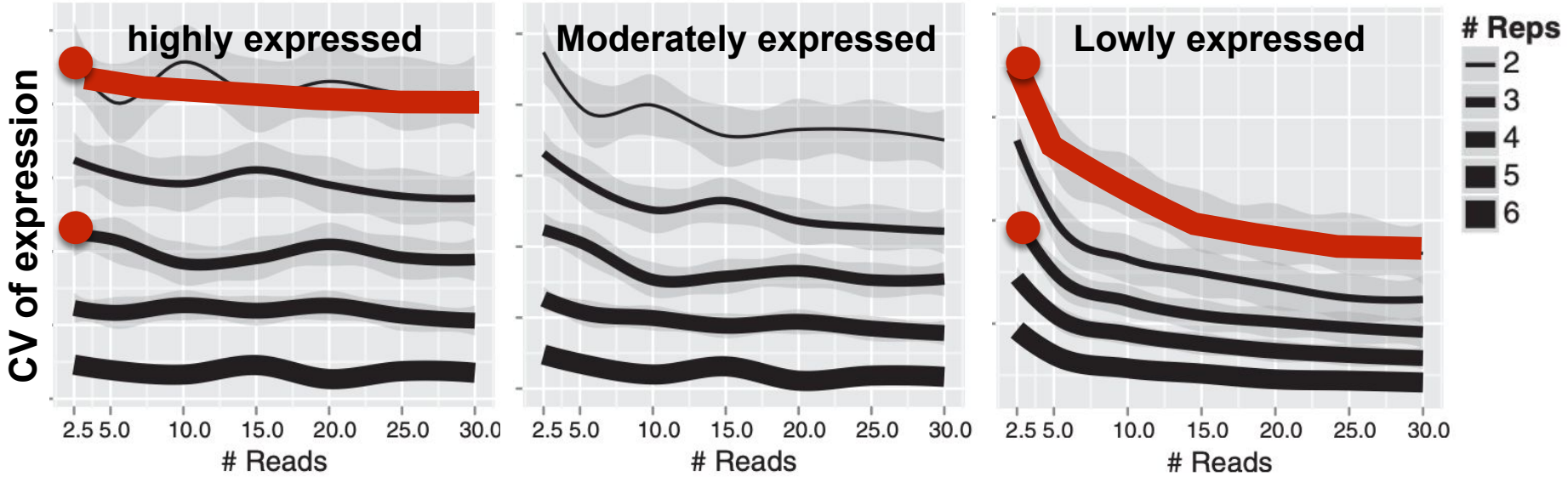
Gene expression

Advance Access publication December 6, 2013

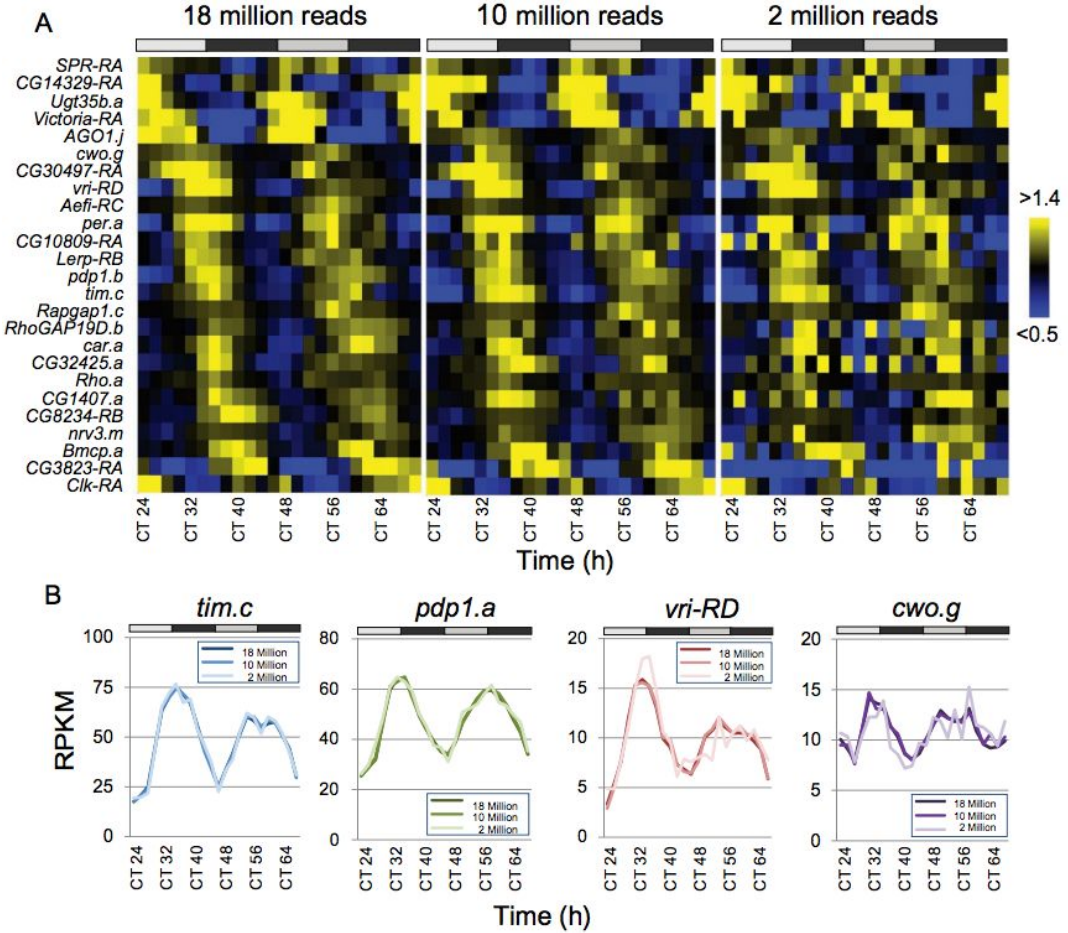
RNA-seq differential expression studies: more sequence or more replication?

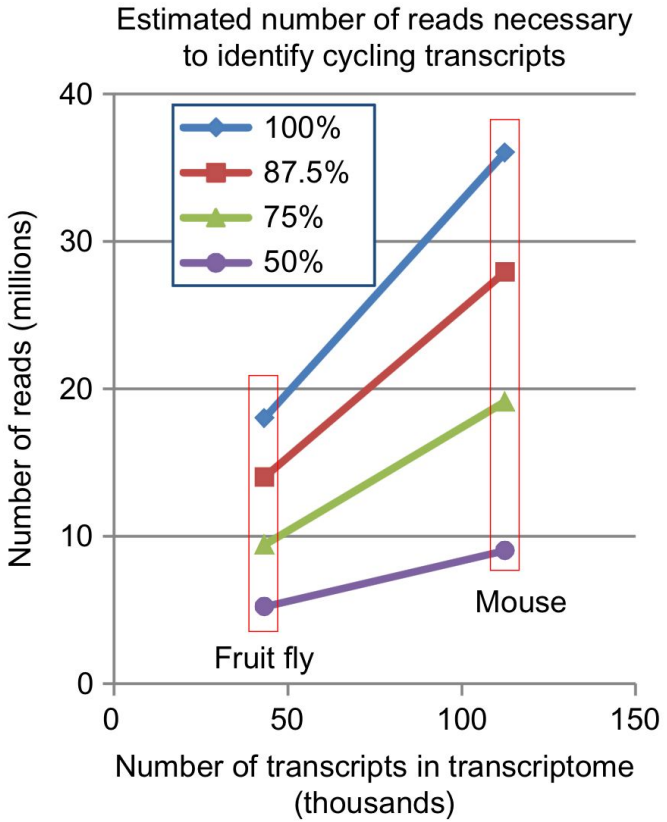
Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA



low sequencing depth can still reveal biology





Li et al., Methods Enzymology, 2015

How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

TABLE 2. A summary of the recommendations of this paper

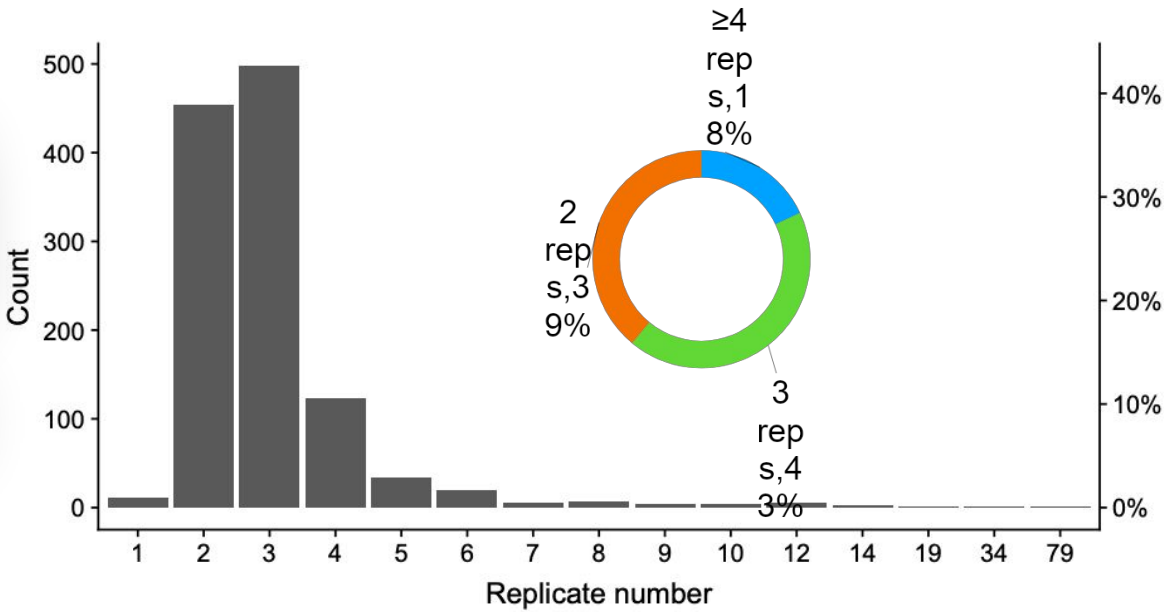
	Agreement with other tools[a]	WT vs. WT FPR[b]	Fold-change threshold (T)[c]	Tool recommended for: (# good replicates per condition)[d] ≤3	≤12	>12
DESeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
EBSeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
edgeR (exact)	Consistent	Pass	0	-	-	Yes
			0.5	Yes	Yes	Yes
			2.0	Yes	Yes	Yes
Limma	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
cuffdiff	Consistent	Fail				
DESeq2	Consistent	Fail				
BaySeq	Inconsistent	Pass				
edgeR (GLM)	Inconsistent	Pass				
DEGSeq	Inconsistent	Fail				
NOISeq	Inconsistent	Fail				
PoissonSeq	Inconsistent	Fail				
SAMSeq	Inconsistent	Fail				

How many replicates are ‘standard’?

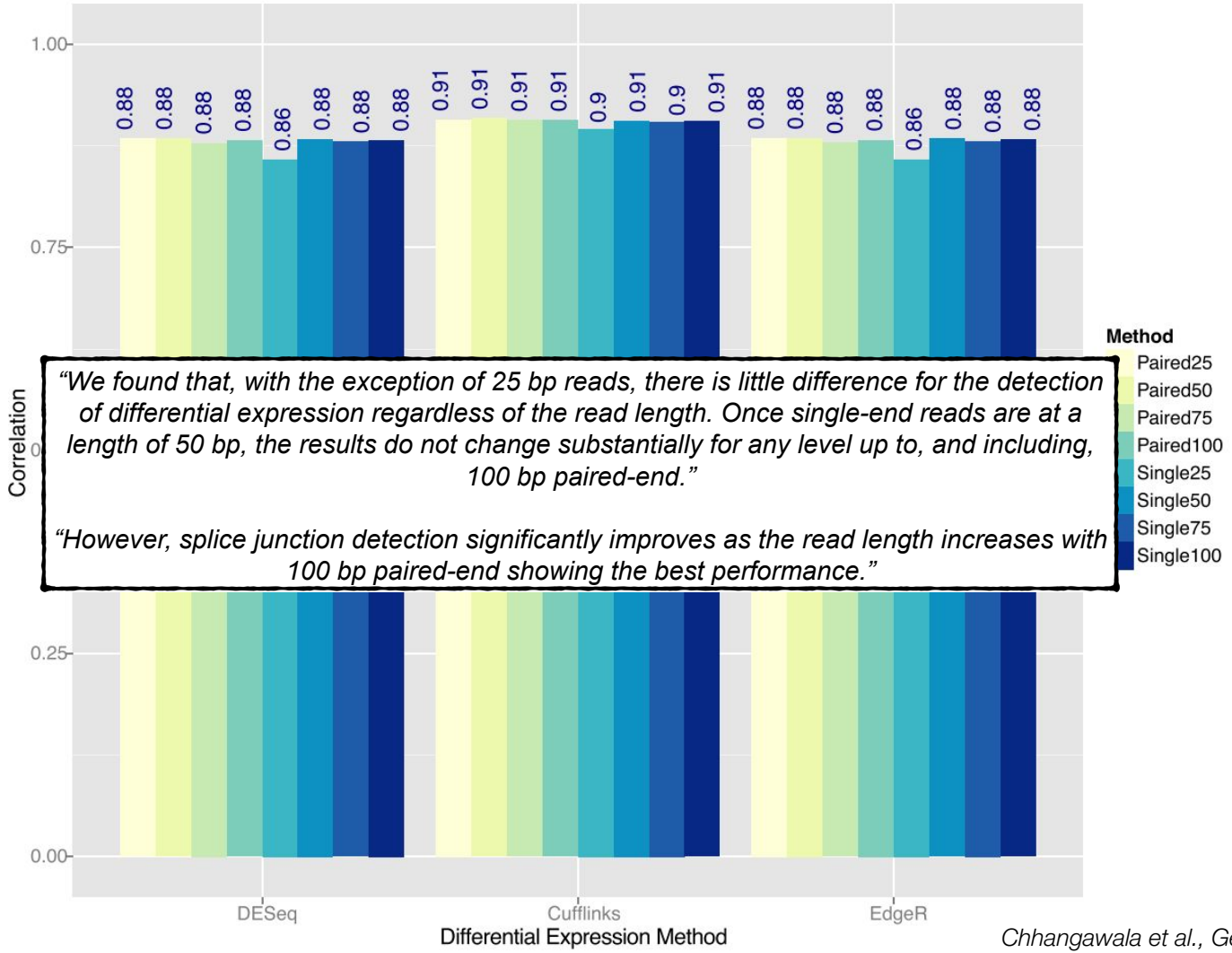
Mined public data from1167 studies were investigators named samples ‘rep1’, ‘replicate 1’, ‘rep_1’, etc,

Your budget is better spent sequencing more replicates rather than fewer reps more deeply

How ‘bout read length?



Li et al., Genome Biology, 2020



"We found that, with the exception of 25 bp reads, there is little difference for the detection of differential expression regardless of the read length. Once single-end reads are at a length of 50 bp, the results do not change substantially for any level up to, and including, 100 bp paired-end."

"However, splice junction detection significantly improves as the read length increases with 100 bp paired-end showing the best performance."

Chhangawala et al., Genome Biology, 2015