‘To sequence or not to sequence’ is not a question anymore. BUT…

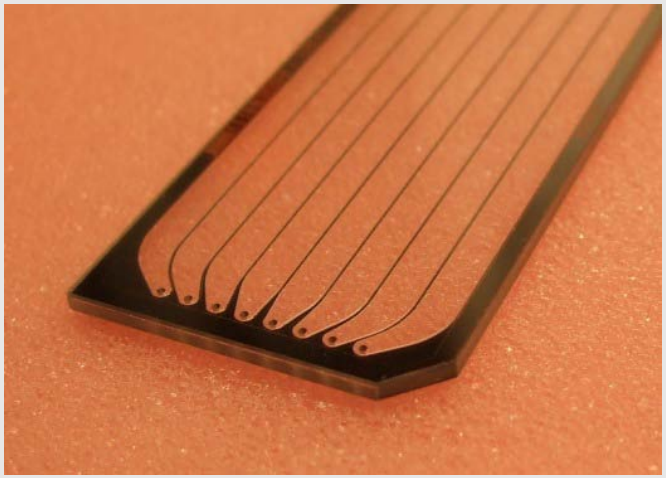
Vladimír Beneš

21 June 2011

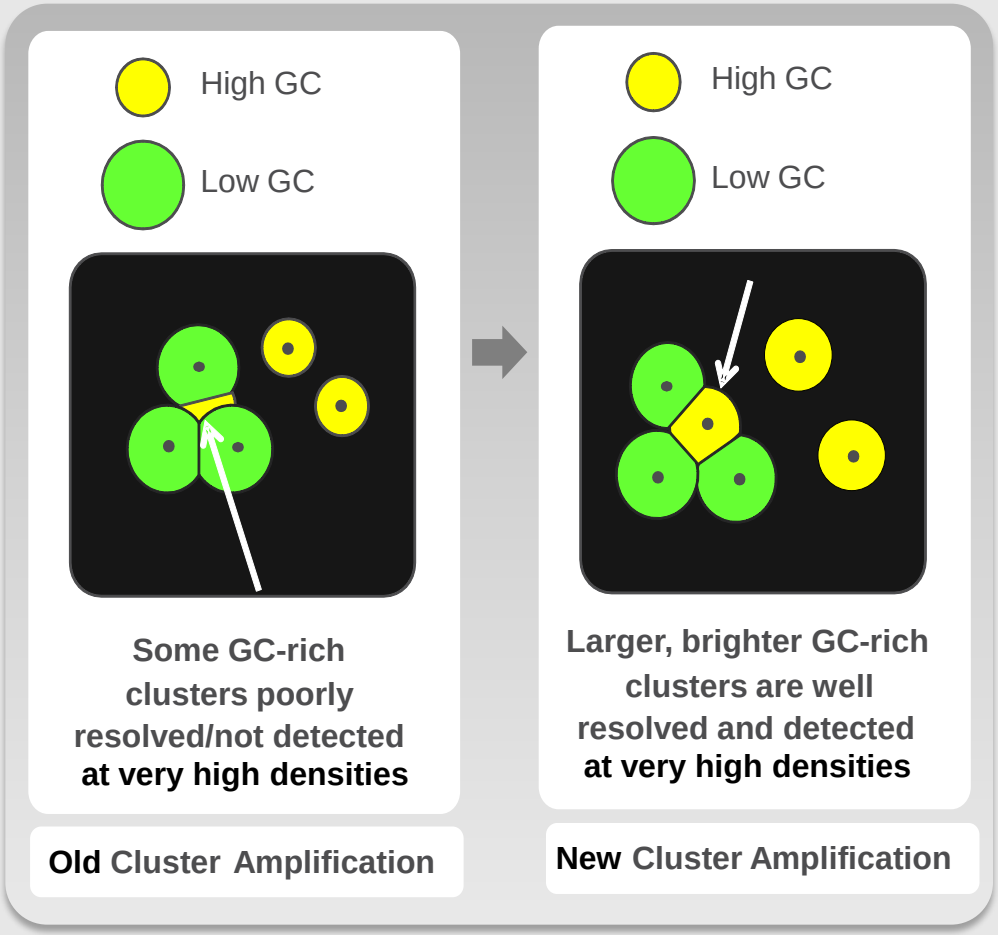
EMBL

http://www.genecore.embl.de

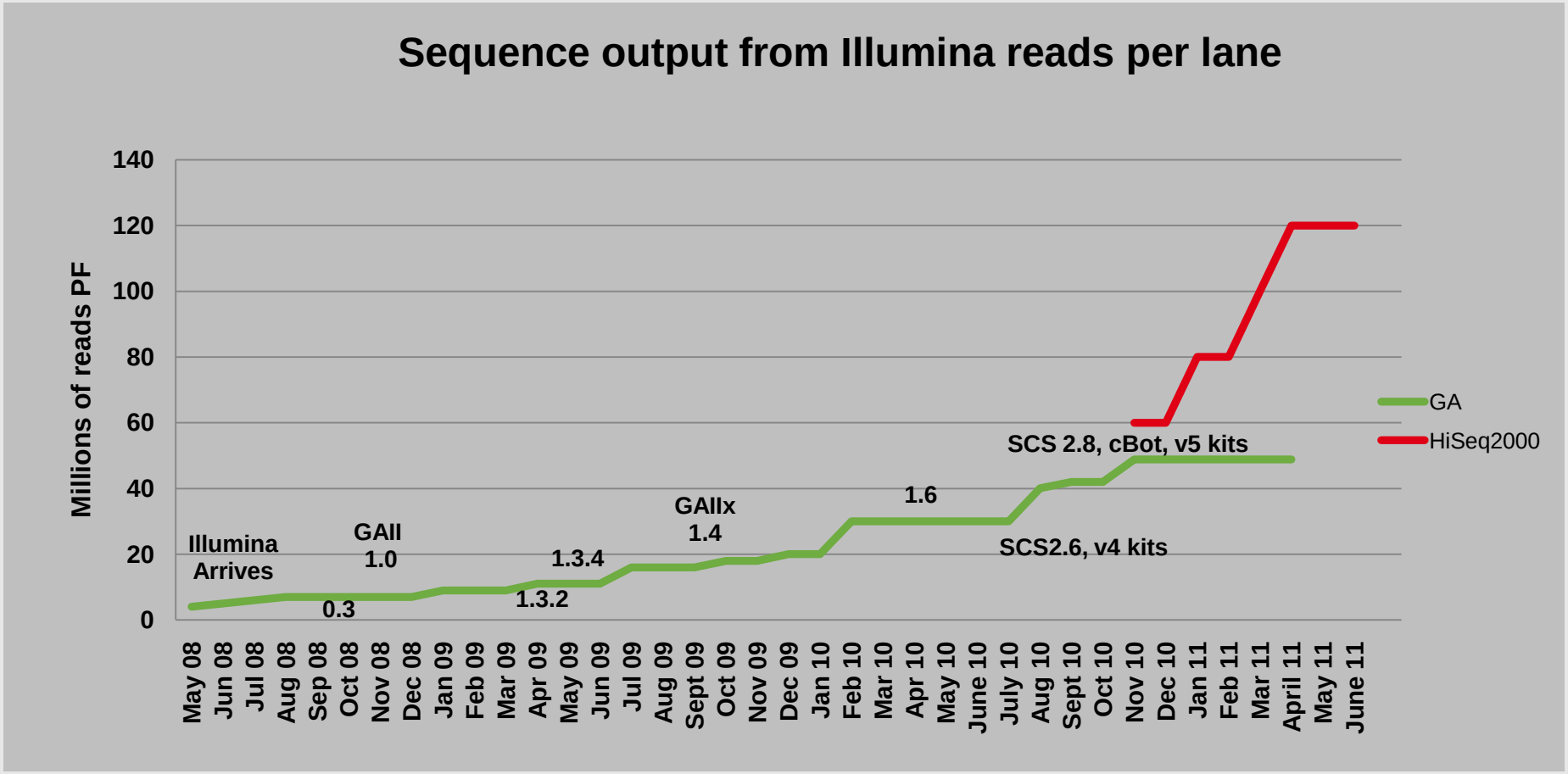
More data on their way to you!



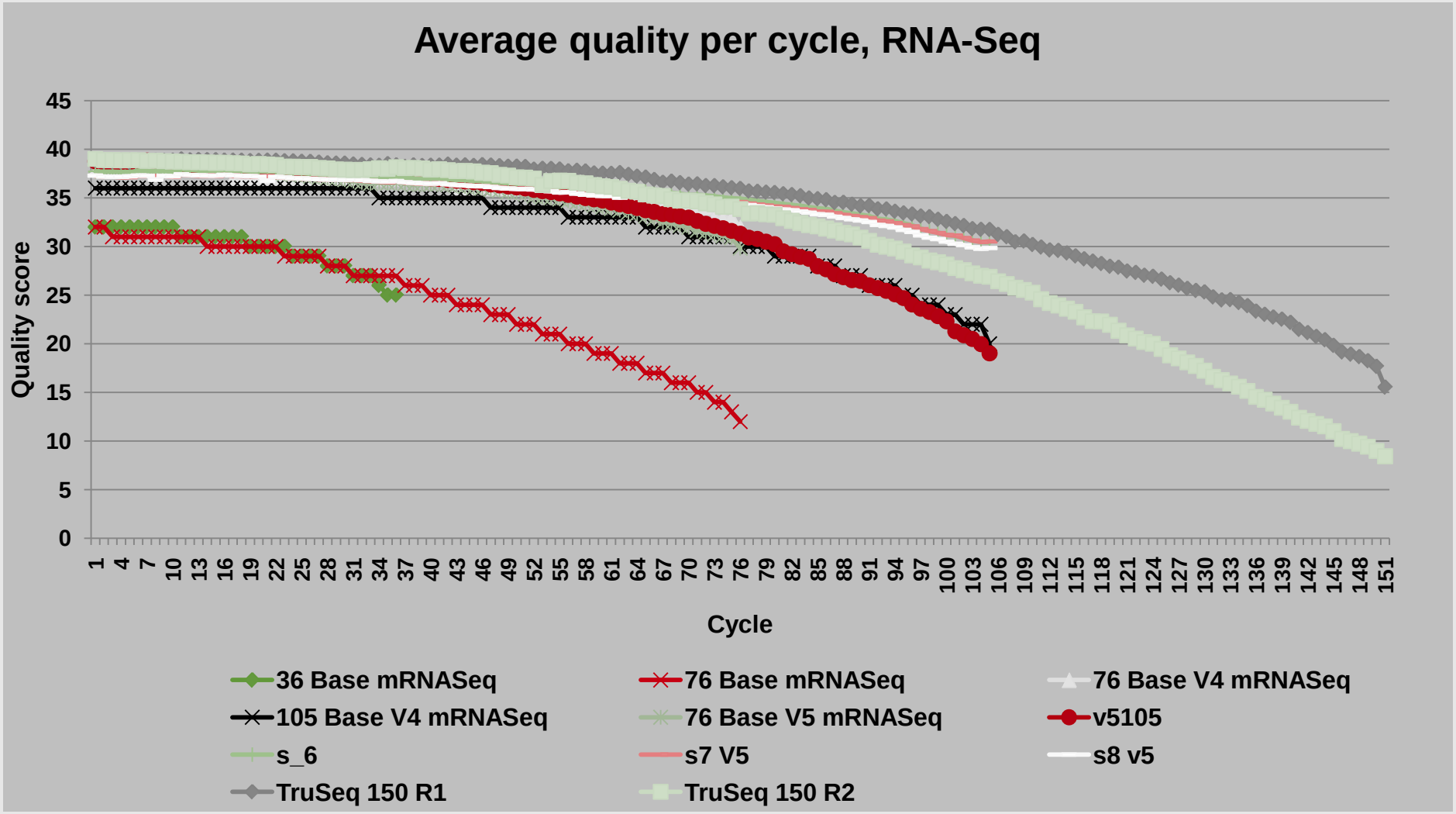
v3 flowcell
imaged area larger by 50 %!
v3 sequencing chemistry



Increasing yield



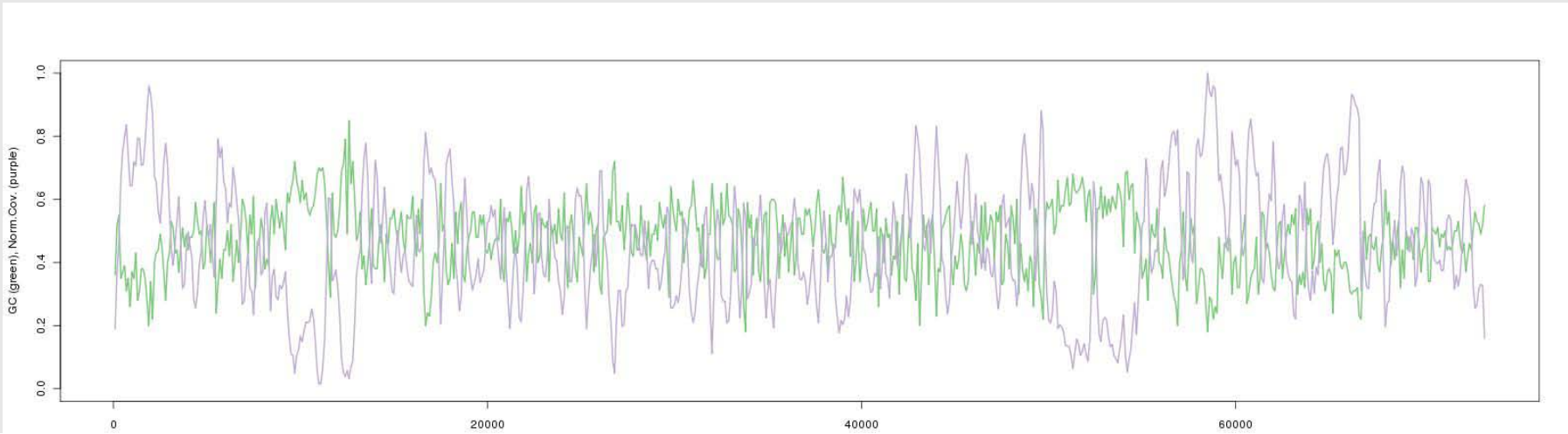
Improving quality of called bases

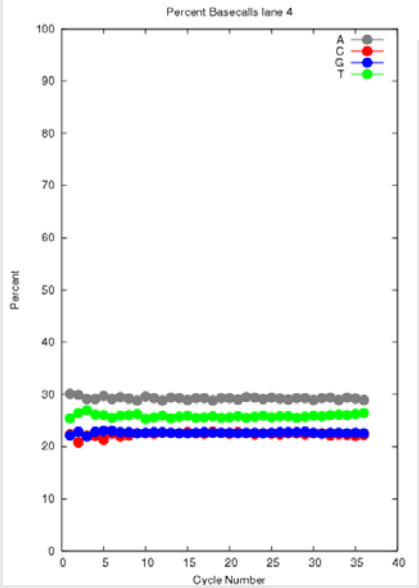


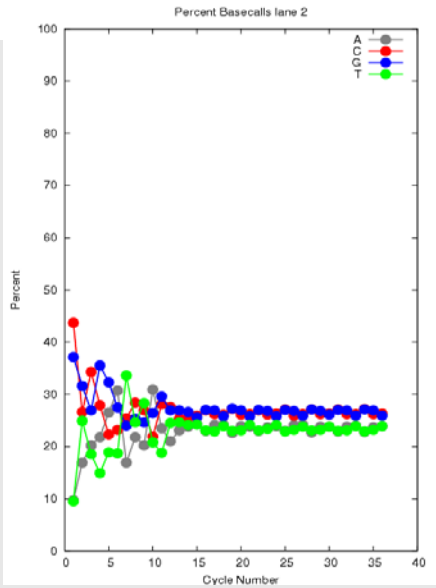
Challenges

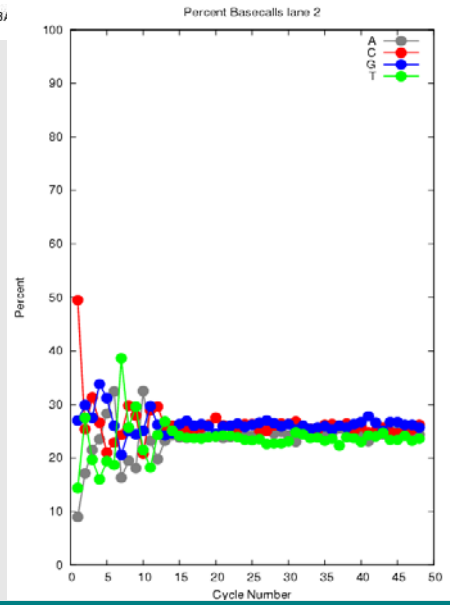
- The only thing constant in life is change…
- Distorted expectations of users
- Data (‘massive’ amounts, formats…)
- Interpretation of results (suboptimal experimental design; is everything relevant?)
- Incomplete understanding of sources of error and bias in MPS data

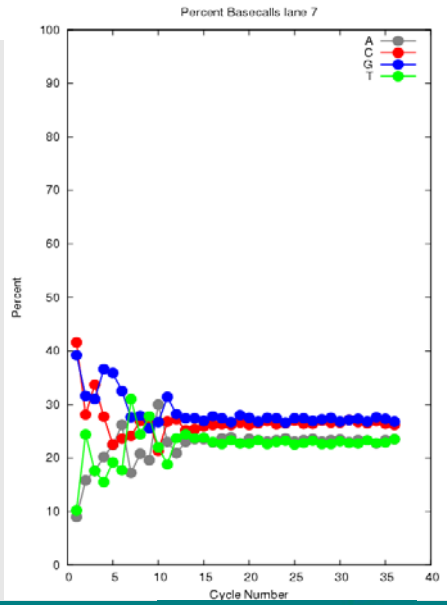
Bias is never good…











OPEN ACCESS Freely available online

PLoS one

Comparison of Sequence Reads Obtained from Three Next-Generation Sequencing Platforms

Shingo Suzuki[1☯], Naoaki Ono[1☯], Chikara Furusawa[1], Bei-Wen Ying[1], Tetsuya Yomo[1,2,3]*

1 Department of Bioinformatics Engineering, Graduate School of Information Science and Technology, Suita, Osaka, Japan, **2** Exploratory Research for Advanced Technology (ERATO), Japan Science and Technology Agency, Suita, Osaka, Japan, **3** Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka, Japan

Published onli

Biase
by

Ka

[1]Di
[2]De
[3]De

BMC

RNA-seq : technical variability and sampling

BMC Genomics 2011, **12**:293 doi:10.1186/1471-2164-12-293

Lauren M McIntyre (mcintyre@ufl.edu)
Kenneth K Lopiano (KLopiano@ufl.edu)

Sequ Geng Tian[1,2]
sequencers

Kensuke Nakamura
Hirofumi Yoshikaw
Hiroki Takahashi[1],

METHOD Open Access

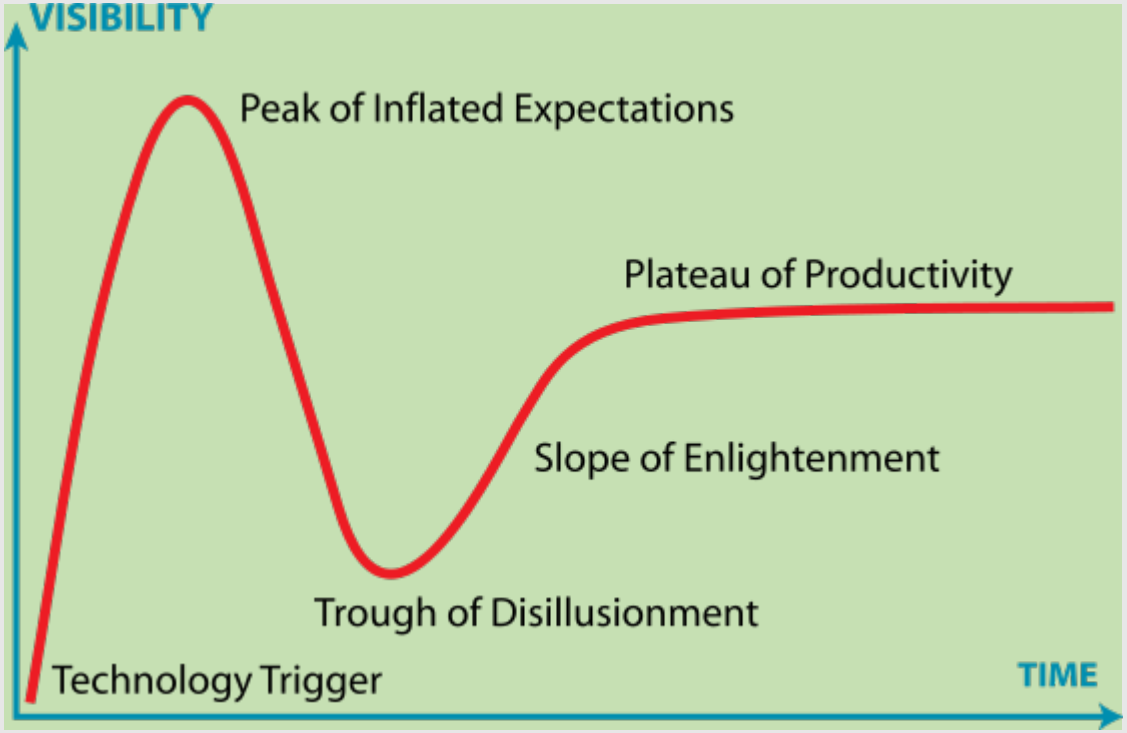
Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries

Daniel Aird[1], Michael G Ross[1], Wei-Sheng Chen[2], Maxwell Danielsson[2], Timothy Fennell[3], Carsten Russ[1], David B Jaffe[1], Chad Nusbaum[1], Andreas Gnirke[1*]

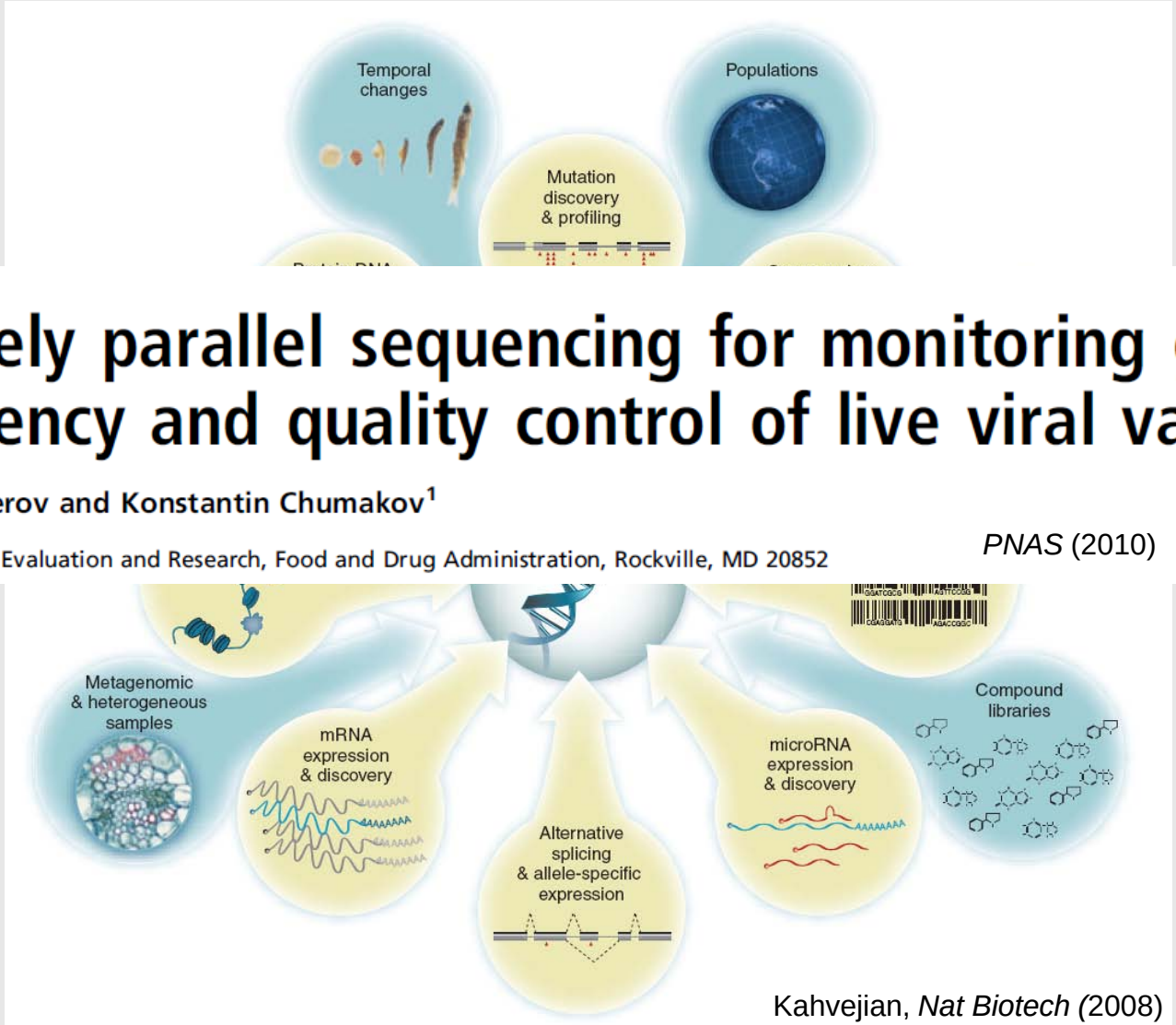
Gene Core

EMBL

Hype/hope curve



MPS space



Kahvejian, *Nat Biotech (*2008)

Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines

Alexander Neverov and Konstantin Chumakov[1]

Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, MD 20852

PNAS (2010)

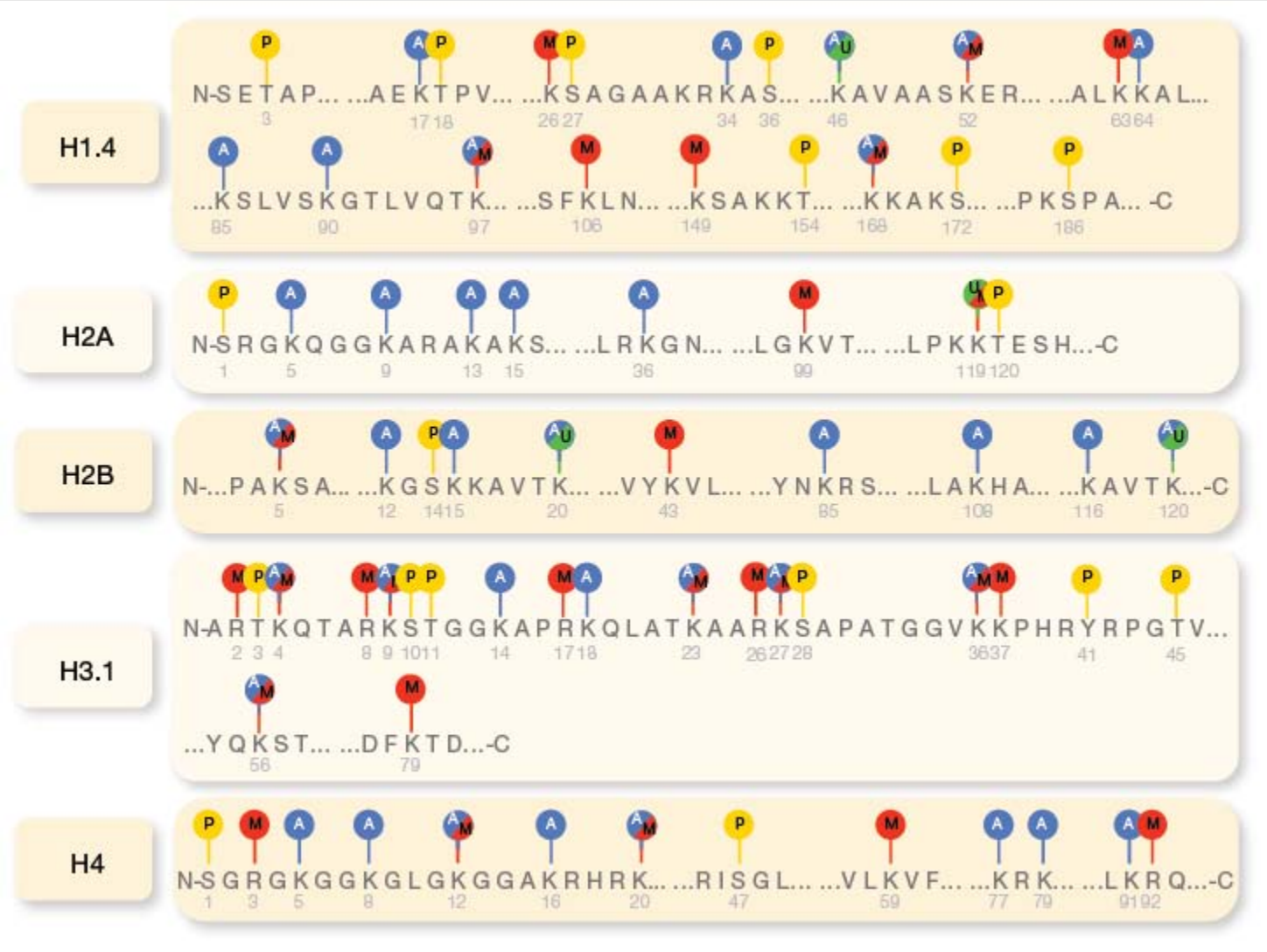
Available MPS applications

transcriptome RNA-Seq, Tag-Seq	**yes**
miRnome smallRNA-Seq	**yes**
protein-NA interactions ChIP-Seq, CLIP-Seq	**yes**
epigenome Methyl-Seq	**yes**
***de novo* & re-sequencing**	**yes**
Metagenomics	**yes**
Genome capture, multiplexing	**yes**

MPS methods used in epigenomics

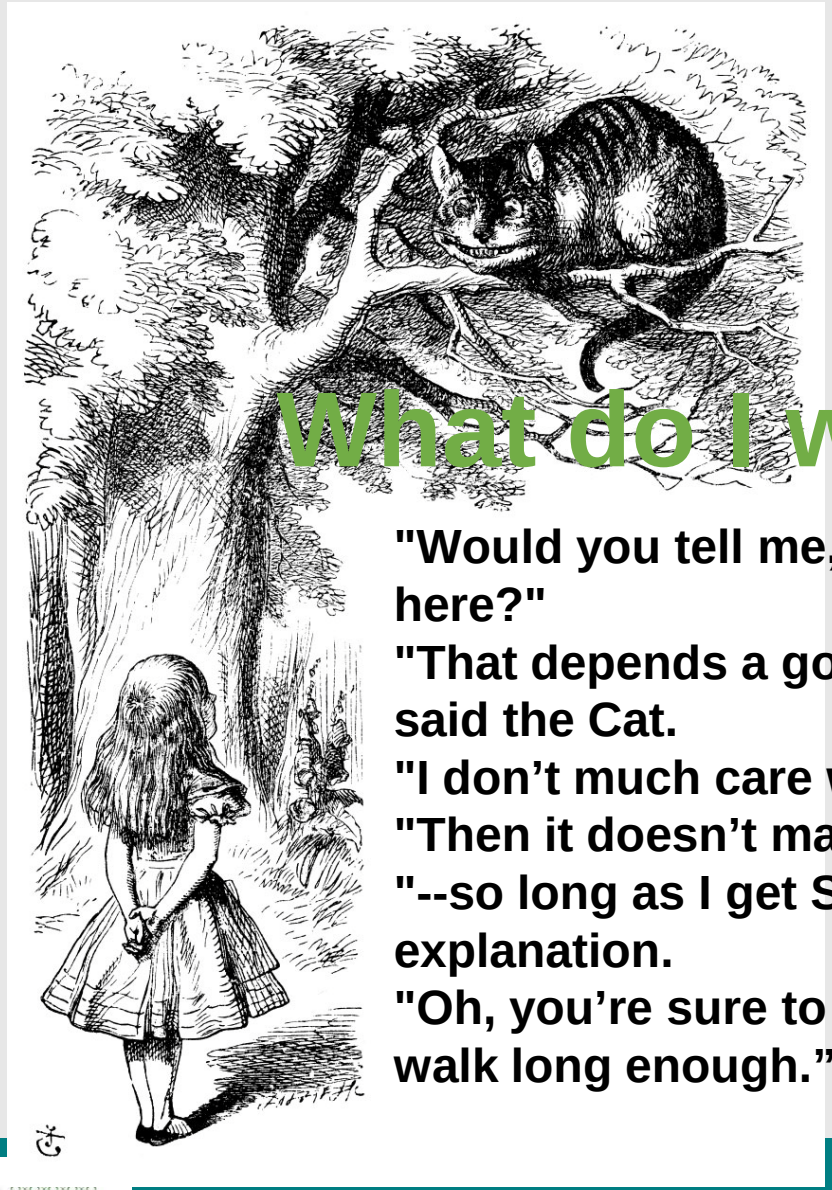
Epigenetic modification	Method
DNA methylation	MethylC-seq
	BS-seq
	MeDIP-seq
	MRE-seq
	MethylCap-seq
	RRBS
Histone post-translational modifications	ChIP-seq
Histone variants	ChIP-seq
Chromatin modifiers and remodelers	ChIP-seq
Chromatin accessibility	DNAseI-seq
	FAIRE-seq
	Sono-seq
Nucleosome positioning and turnover	MNase-seq
	CATCH-IT
Long-range chromatin interactions	Hi-C
	ChIA-PET
Allele-specific chromatin signatures	haploChIP

Rada-Iglesias & Wysocka, *Genome Medicine* (2011)



Portela & Esteller, *Nature Biotechnology* (2011)

Importance of experimental design



What do I want to study?

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where--" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"--so long as I get SOMEWHERE," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Lewis Carroll's Alice in Wonderland

Which sequencing mode to use?

Sequencing type	Recommendation
Exon capture	**50Mb Kit, human: 105b SR – to get sufficient coverage**
Whole genome sequencing	***Large rearrangements*: Mate-pairs large insert** ***Resequencing*: SNPs/indels: Coverage is good 100+ PE.** **If you don't get the coverage at the start you'll regret it ☹.**
RN	**uld** **s.** **E** **depending on a *de novo*/spliced read mapping approach or map pairs to detect also alternative splicing.** ***Strand-specific libraries*: complex insight into transcriptome**
Chip-Seq	**36b SR unless you have real concerns about 'alignability' of your target (i.e. some strange looking enhancer region)**
Multiplexing	**Coverage is the key!**

Coverage is the key!

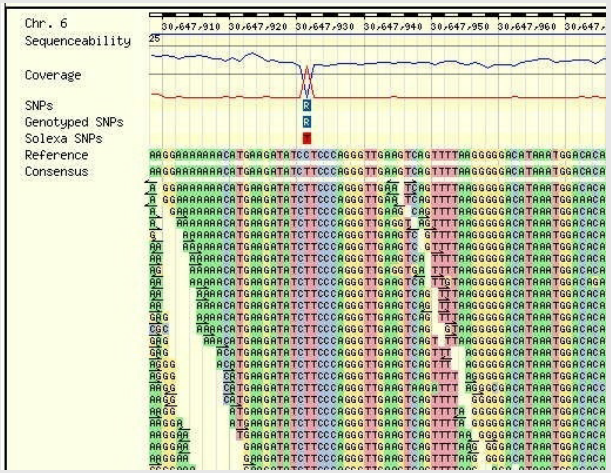
MPS workflow

1) Library preparation

2) Cluster generation on a flow cell

3) Sequencing & imaging

SE, PE reads,
36-150 bases



4) Data processing & analysis

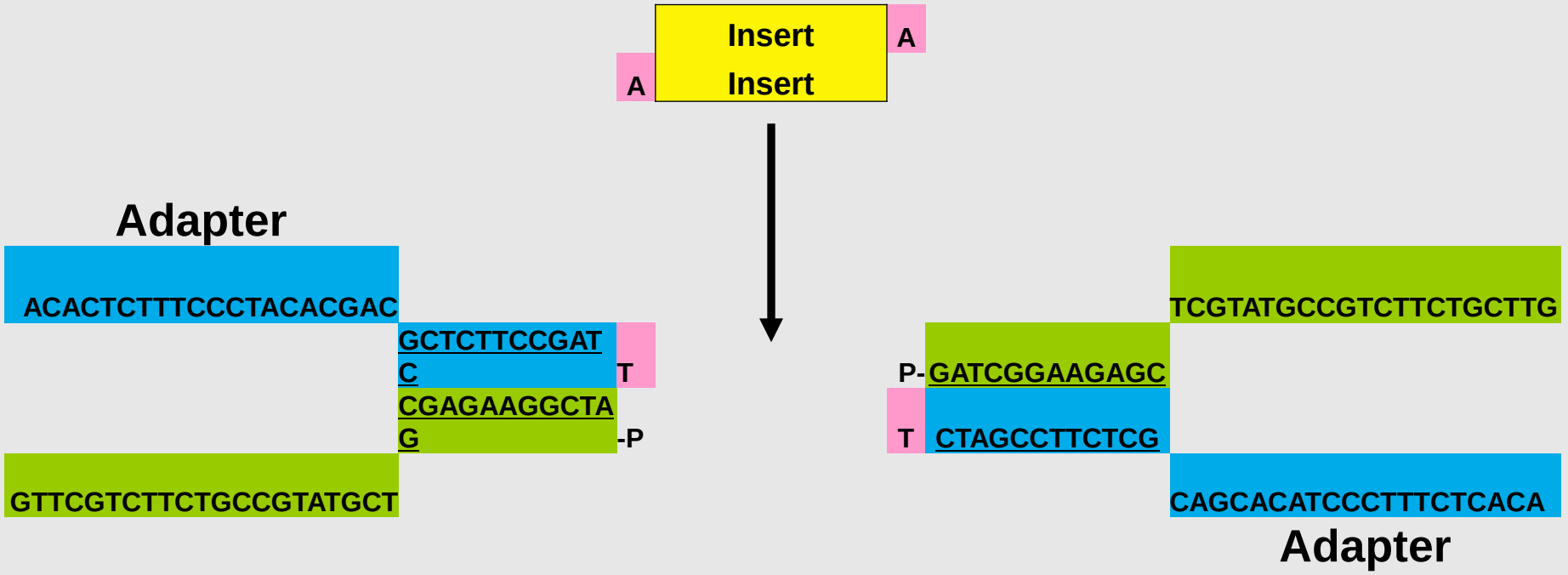
MPS library preparation

```
5′AATGATACGGCGACCACCGA-ACACTCTTTCCCTACACGACGCTCTTCCGATCT--INSERT--TCGTATGCCGTCTTCTGCTTG
  TTACTATGCCGCTGGTGGCT-TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA--INSERT--AGCATACGGCAGAAGACGAAC5′
```

where

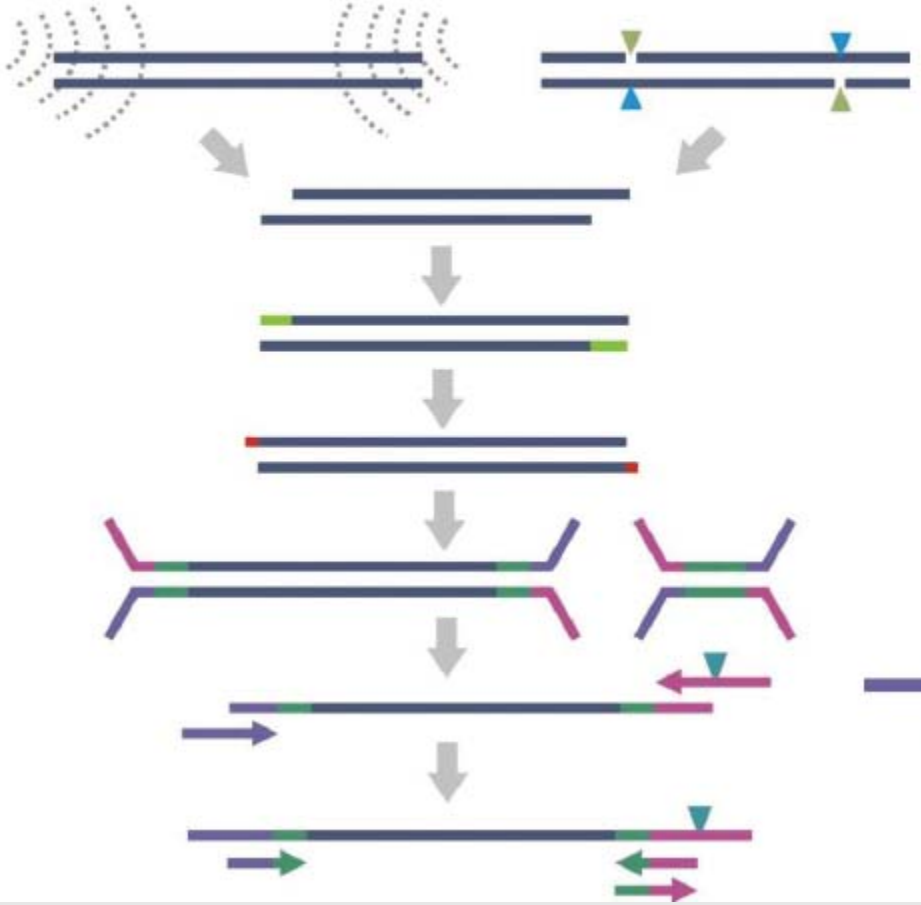
- **5’AATGATACGGCGACCACCGA** **is the P5 attachment/amplification primer sequence**
- **5’CAAGCAGAAGACGGCATACGA** **is the P7 attachment/amplification primer sequence**
- **5’ACACTCTTTCCCTACACGACGCTCTTCCGATCT** **is the SBS3 sequencing primer sequence**
- **INSERT** **is a complex mix of DNA fragments**

Forked adapters



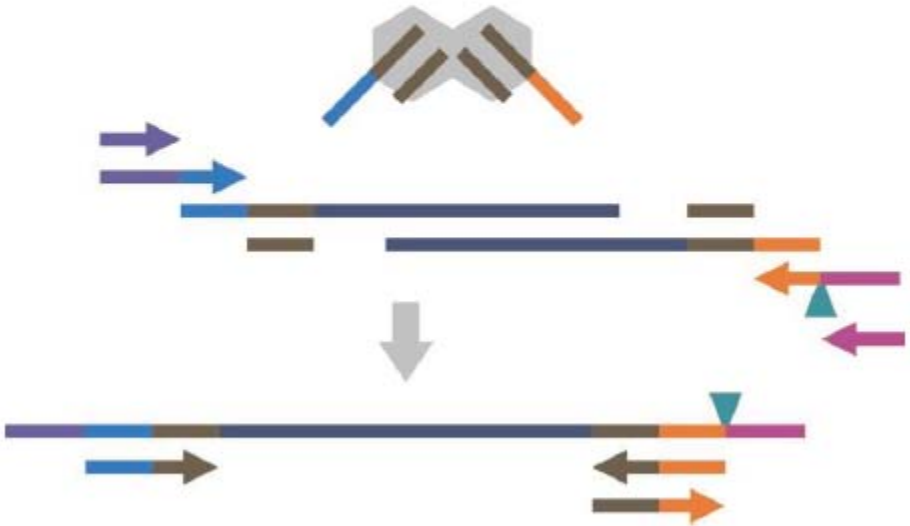
Library preparation II.

(a) **A conventional protocol**



at least 500 ng of gDNA

(b) **A "nextera" protocol**

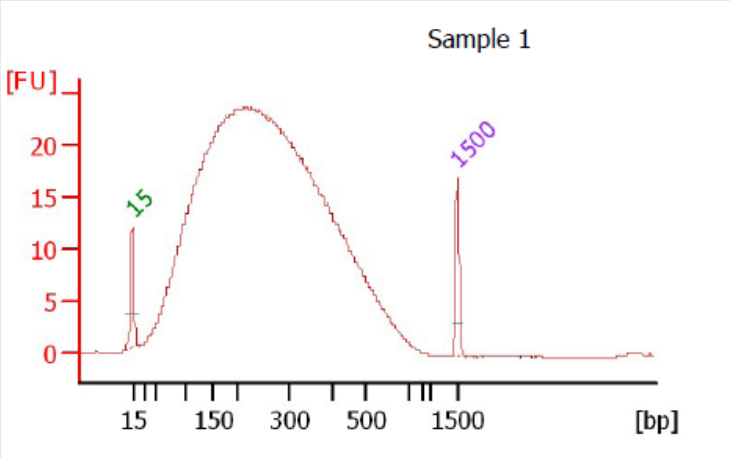


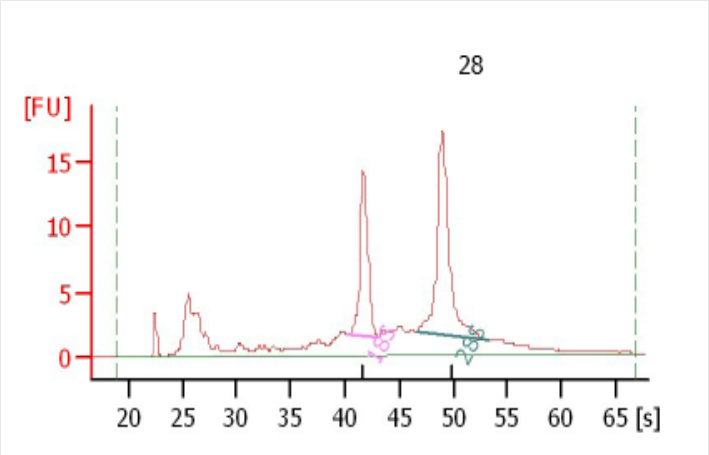
<50 ng of gDNA!

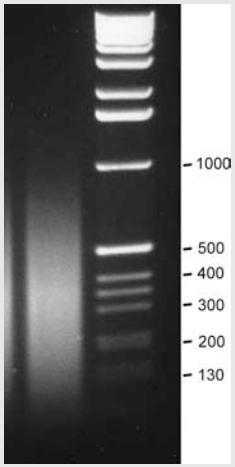
Adey *et al*., *Genome Biology* (2010)

Library preparation III.

- **Strict QC of starting material (GiGo)**
  - Qubit quantification
  - gel images, bioanalyzer/experion traces

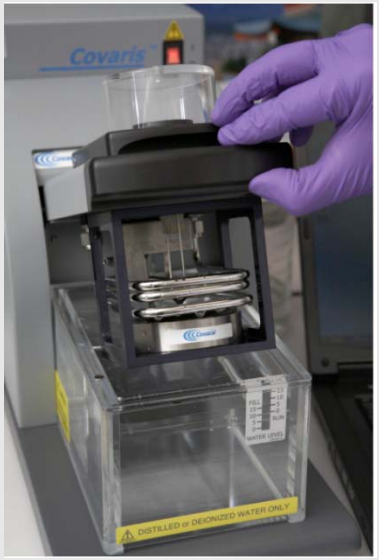






Library preparation IV.

- Bioruptor, probe (ChIP-Seq)
- Covaris *vs* nebulization
- Kits (proprietary, home-brewed, NEB!)
- Size selection using gel extractor, E-gel, Pippin prep, SPRIworks…,
- Lo-bind tubes!



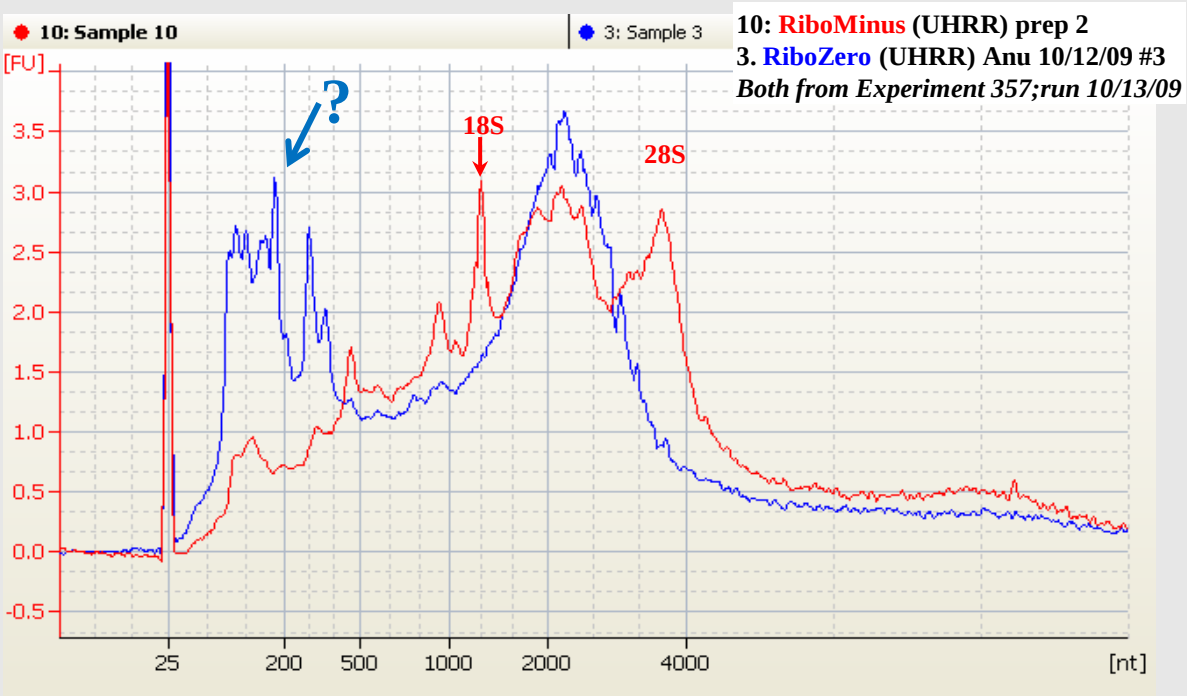
Covaris

Hydroshear

RNA-Seq libraries

- rRNA depletion (oligo-dT beads, Ribo-Minus, Ribo-Zero…)

BUT mitochondria-derived rRNA mostly ignored!!



- strand-specific library, Levine *et al*., *Nat Meth* (2010)

The rocks and shallows of deep RNA sequencing: Examples in the *Vibrio cholerae* RNome

CARSTEN A. RAABE,[1] CHEE HOCK HOE,[2] GERRIT RANDAU,[1] JUERGEN BROSIUS,[1,3] THEAN HOCK TANG,[2] and TIMOFEY S. ROZHDESTVENSKY[1,3]

[1]Institute of Experimental Pathology, University of Muenster, 48149 Muenster, Germany
[2]Infectious Diseases Cluster, Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia, 13200 Penang, Malaysia

ABSTRACT

New deep RNA sequencing methodologies in transcriptome analyses identified a wealth of novel nonprotein-coding RNAs (npcRNAs). Recently, deep sequencing was used to delineate the small npcRNA transcriptome of the human pathogen *Vibrio cholerae* and 627 novel npcRNA candidates were identified. Here, we report the detection of 223 npcRNA candidates in *V. cholerae* by different cDNA library construction and conventional sequencing methods. Remarkably, only 39 of the candidates were common to both surveys. We therefore examined possible biasing influences in the transcriptome analyses. Key steps, including tailing and adapter ligations for generating cDNA, contribute qualitatively and quantitatively to the discrepancies between data sets. In addition, the state of 5′-end phosphorylation influences the efficiency of adapter ligation and C-tailing at the 3′-end of the RNA. Finally, our data indicate that the inclusion of sample-specific molecular identifier sequences during ligation steps also leads to biases in cDNA representation. In summary, even deep sequencing is unlikely to identify all RNA species, and caution should be used for meta-analyses among alternatively generated data sets.

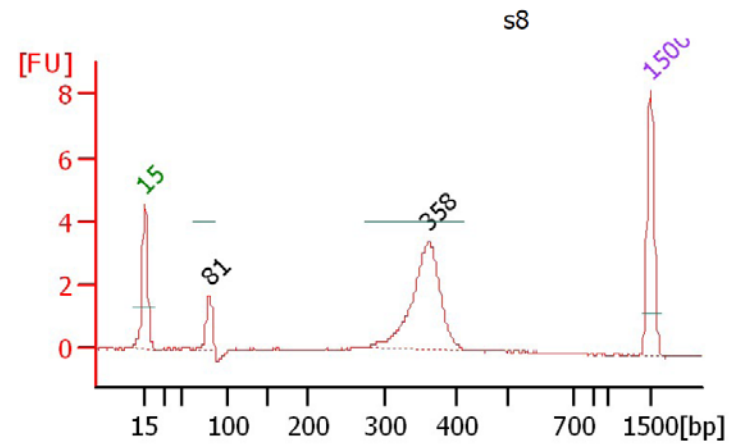
TABLE 1. Sources of potential bias in RNA sequencing

No.	Step	Substeps leading to possible bias
1	Preparation and (counter) selection of RNA starting material	Selection of subcellular organelles or fractions Selection as polysomal mRNAs or other ribonucleoprotein complexes (RNPs) Method of RNA preparation (e.g., loss of small RNAs after LiCl precipitation) Size selection on columns or gels Selection of polyadenyated or capped RNAs Counter-selection of undesired rRNAs, tRNAs, etc. by affinity methods
2	Removal or addition of RNA modifications	Mostly at the termini, such as decapping, dephosphorylation, or phosphorylation
3	Extension of RNA 3′-ends	Tailing with oligo(A) or oligo(C) using poly(A) polymerase Ligation of oligonucleotide adapters
4	Extension of RNA 5′-ends	Ligation of oligonucleotide adapters
5	Reverse transcription	RNA modifications and secondary structures can lead to premature stop of extension
6	Adapter ligation to 3′-ends of (first) single-stranded cDNA (in applicable protocol variants)	
7	PCR amplification	Can lead to bias in amplicon representation due to template size, base composition, repeat content, hairpin structures, etc.
8	Cloning efficiency	In protocols where cDNA is cloned in, e.g., plasmid vectors prior to sequencing Adapter restriction sites for cloning might also be present on cDNA
9	Computational analyses	Different filters and stringency

Shading indicates steps examined in this study.

Library quantification & QC

- Qubit
- Bioanalyzer
  - HS DNA Chip
  - DNA 1000 Chip



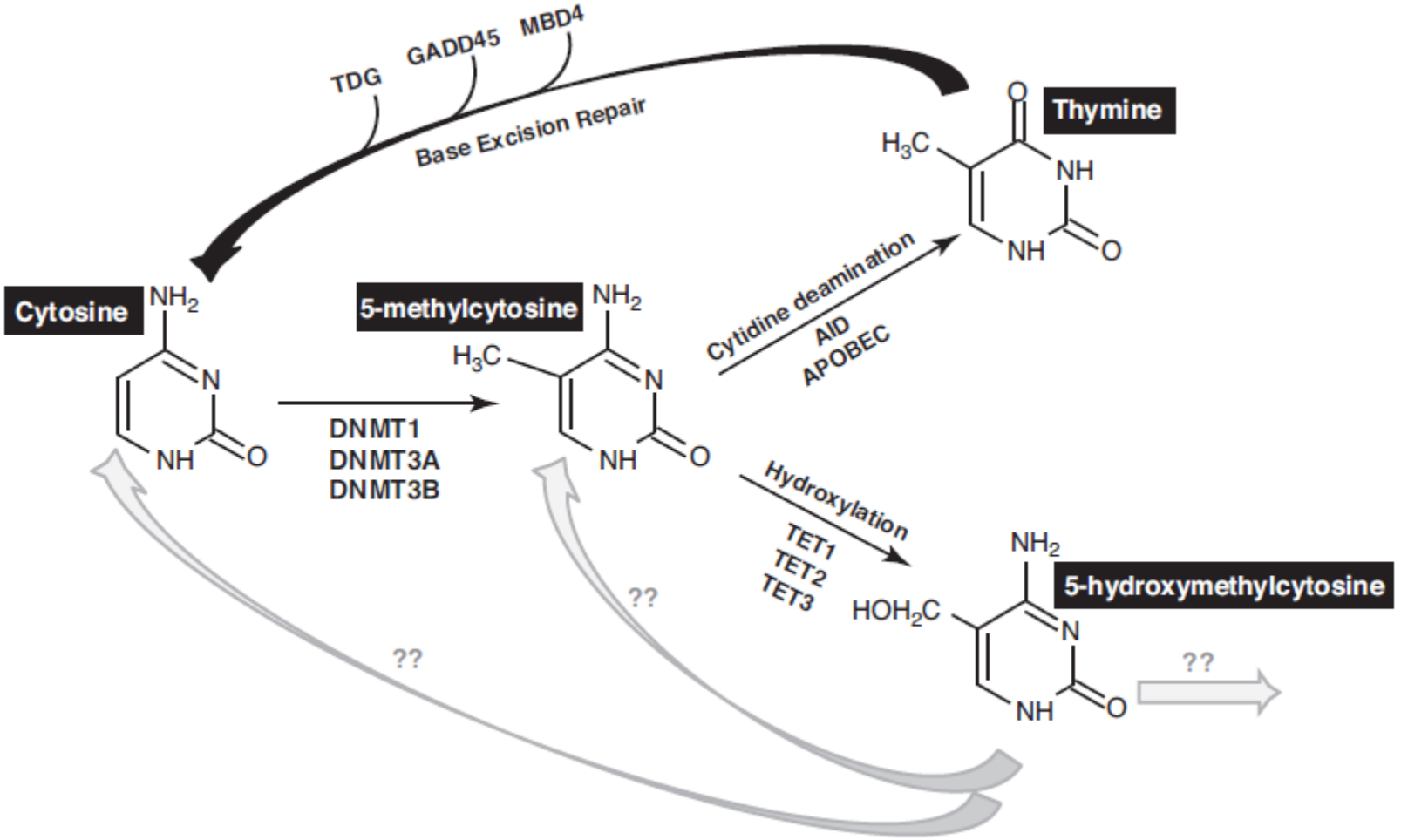
Overall Results for sample 8 : s8

Number of peaks found: 2

Peak table for sample 8 : s8

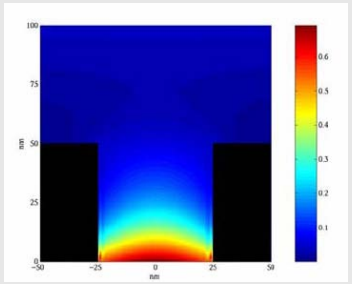
Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	81	0.93	17.4	
3	358	4.57	19.3	
4	1,500	2.10	2.1	Upper Marker

Methyl-Seq

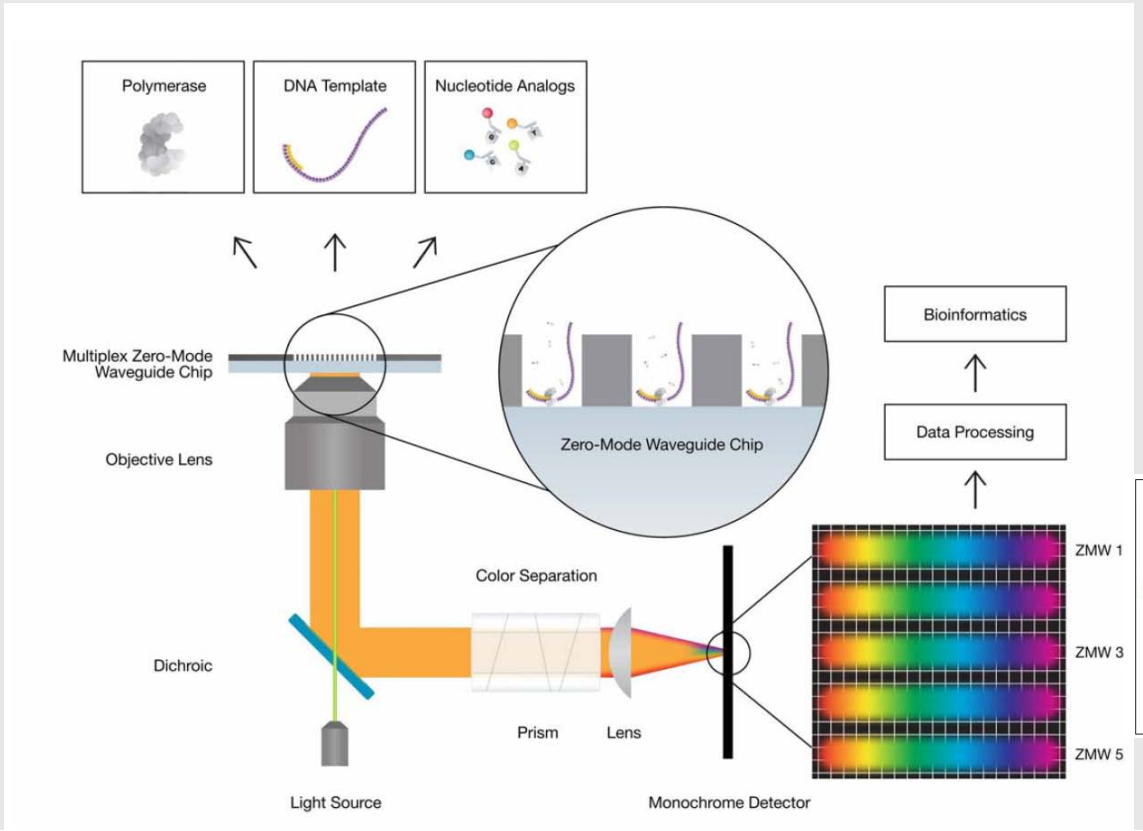


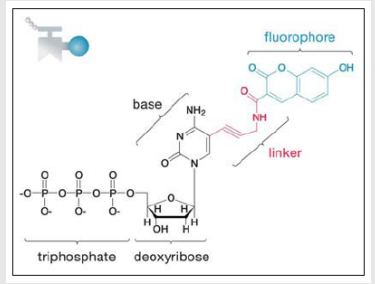
Carey *et al*., *Drug Discovery Today* (2011) Zilbermann & Henikoff, *Development* (2007)

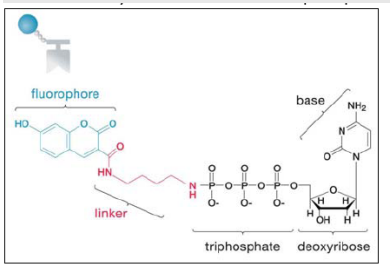
Pacific Biosciences



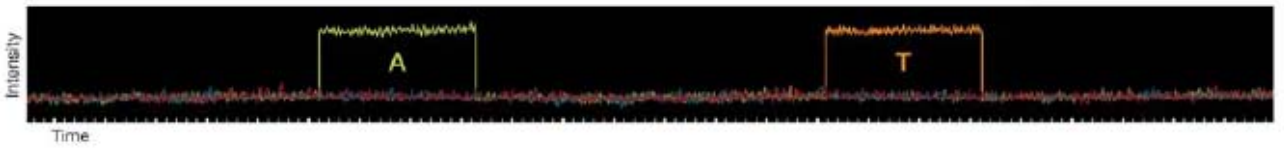
Detection volume





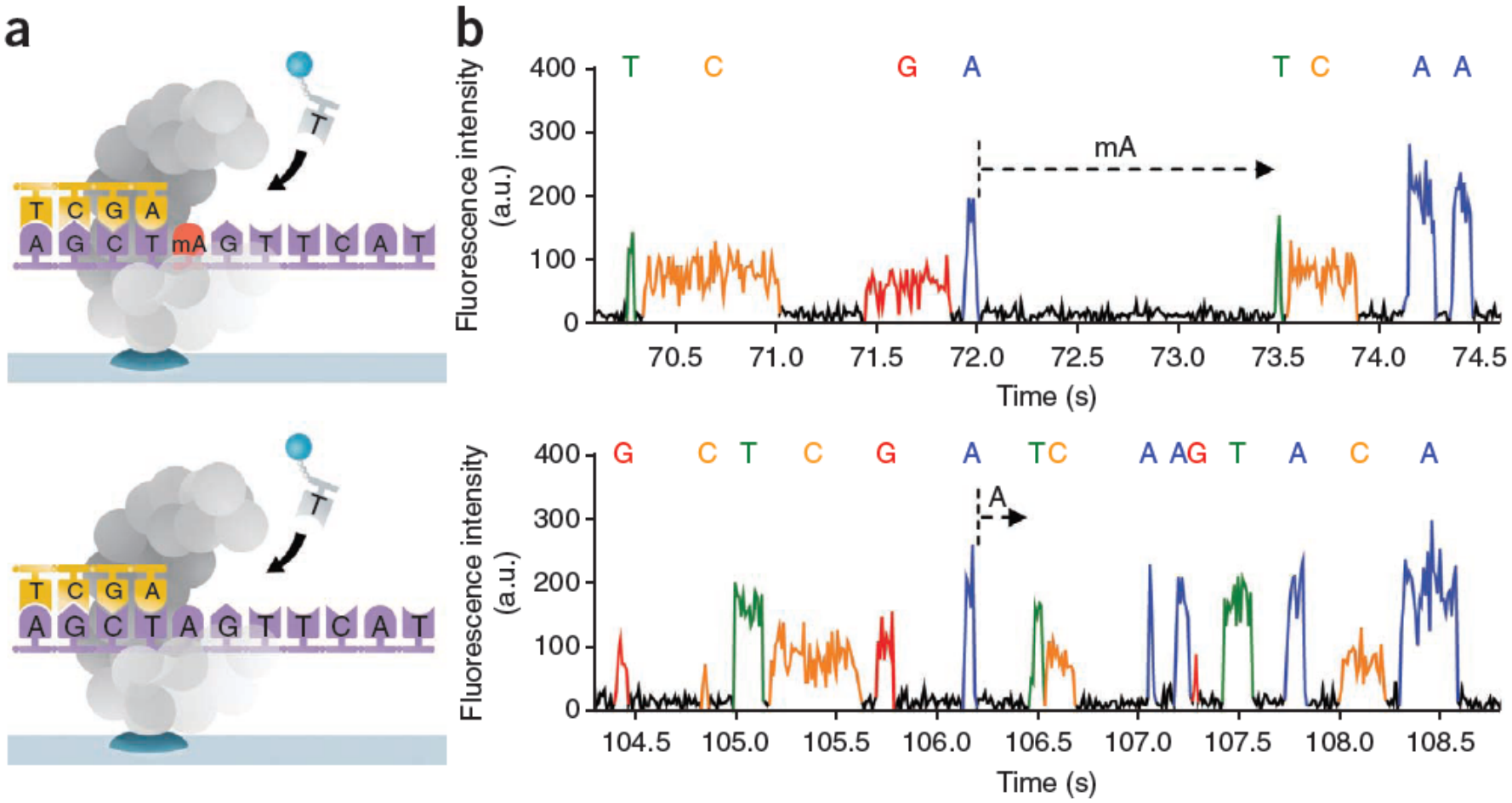


Modified bases



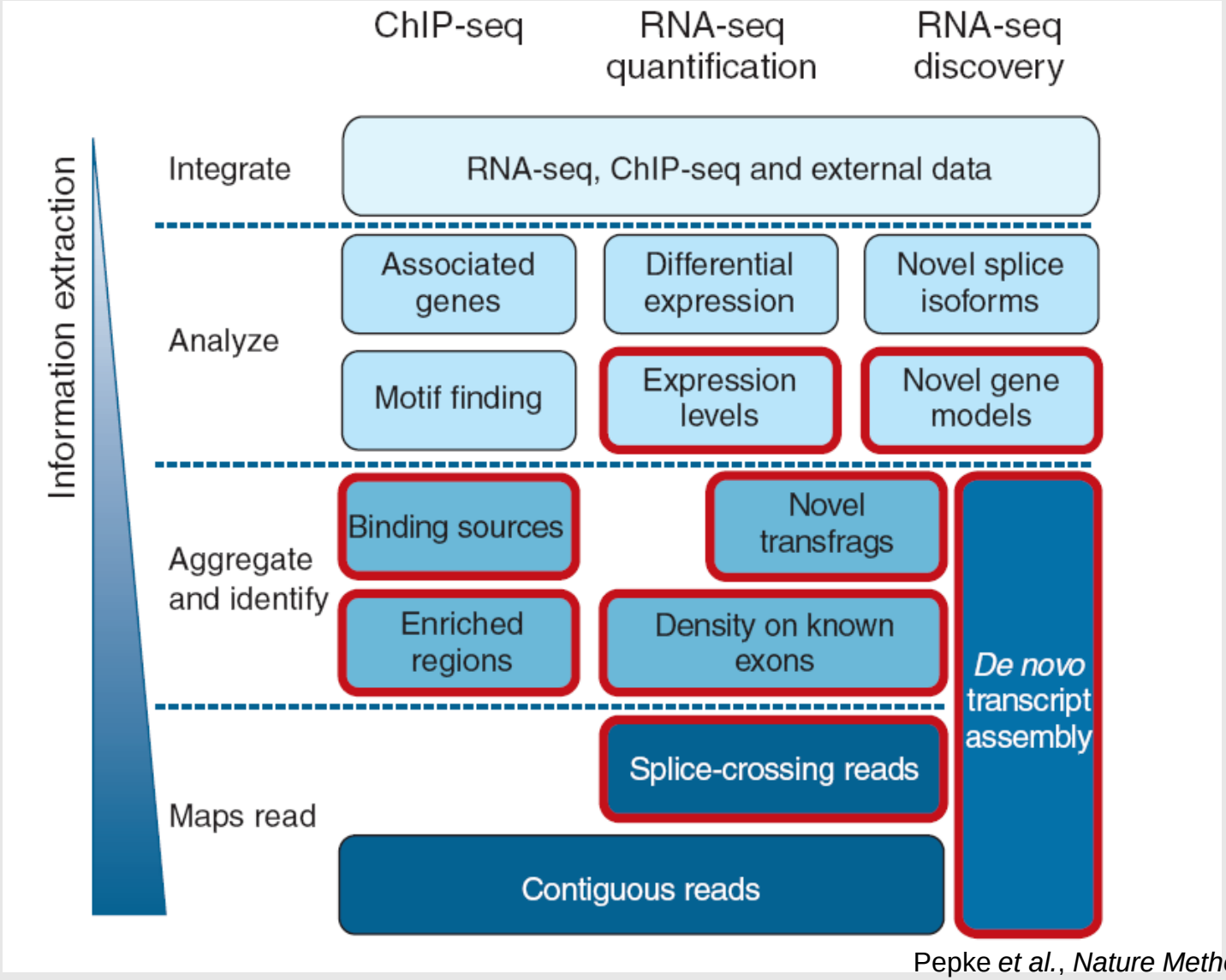
By courtesy of Pacific Bioscience

Direct detection of methylated bases



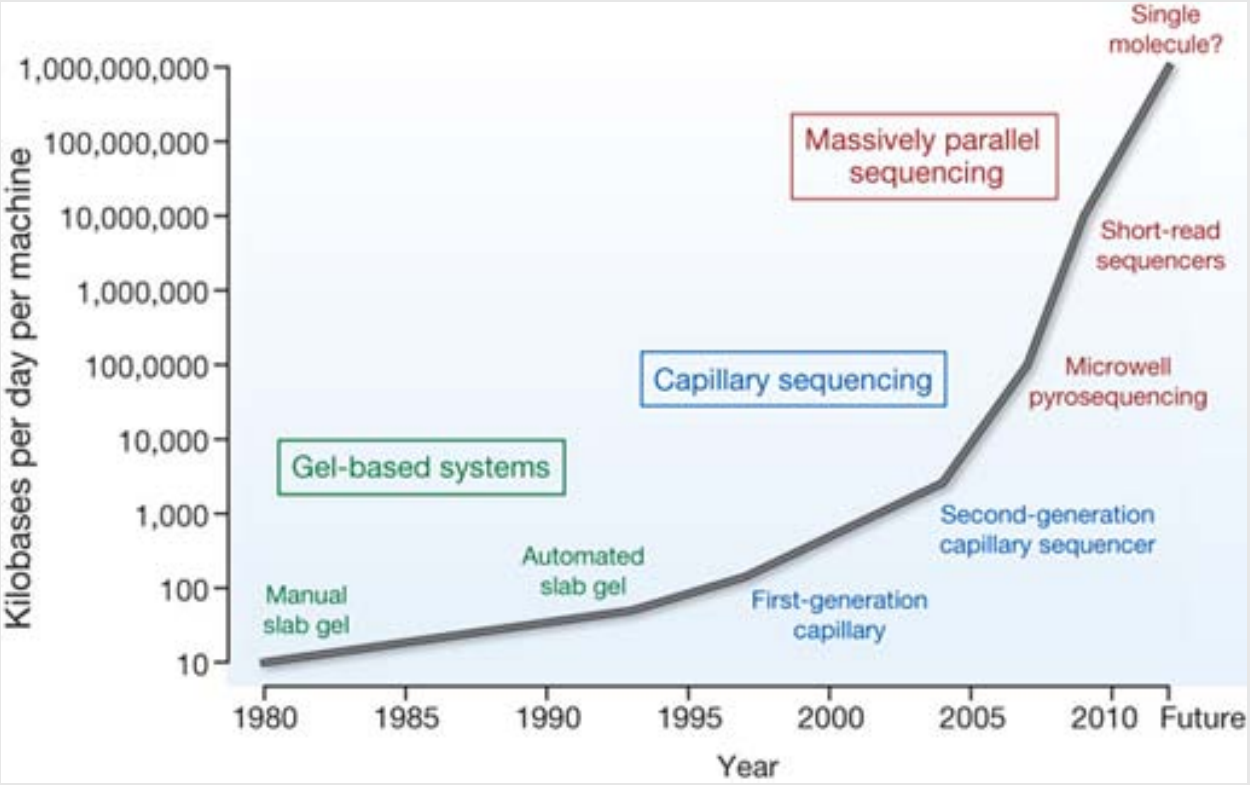
Flusberg *et al*. *Nature Methods* (2010)

Data integration

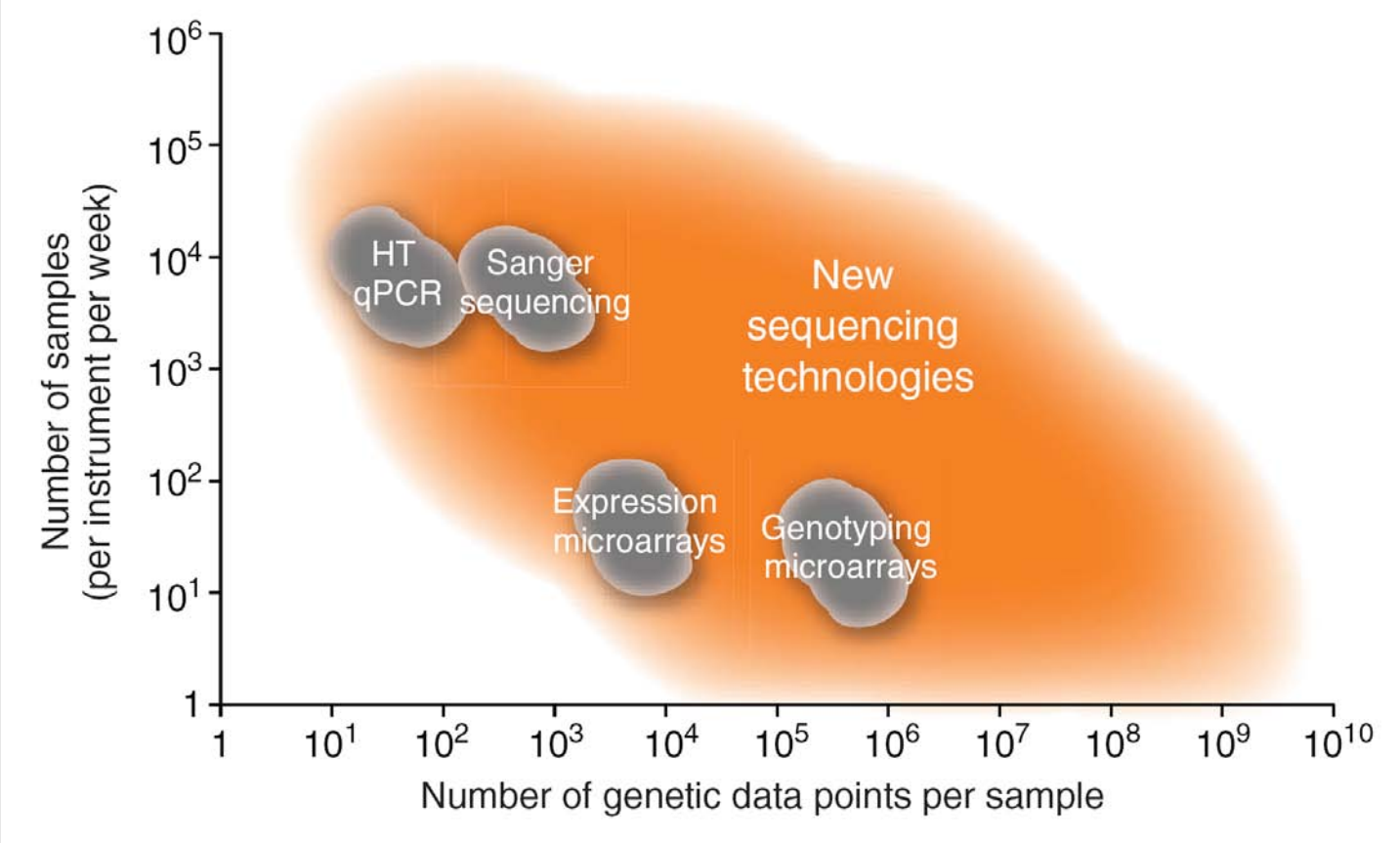


Pepke *et al.*, *Nature Methods* (2009)

Where are we heading?

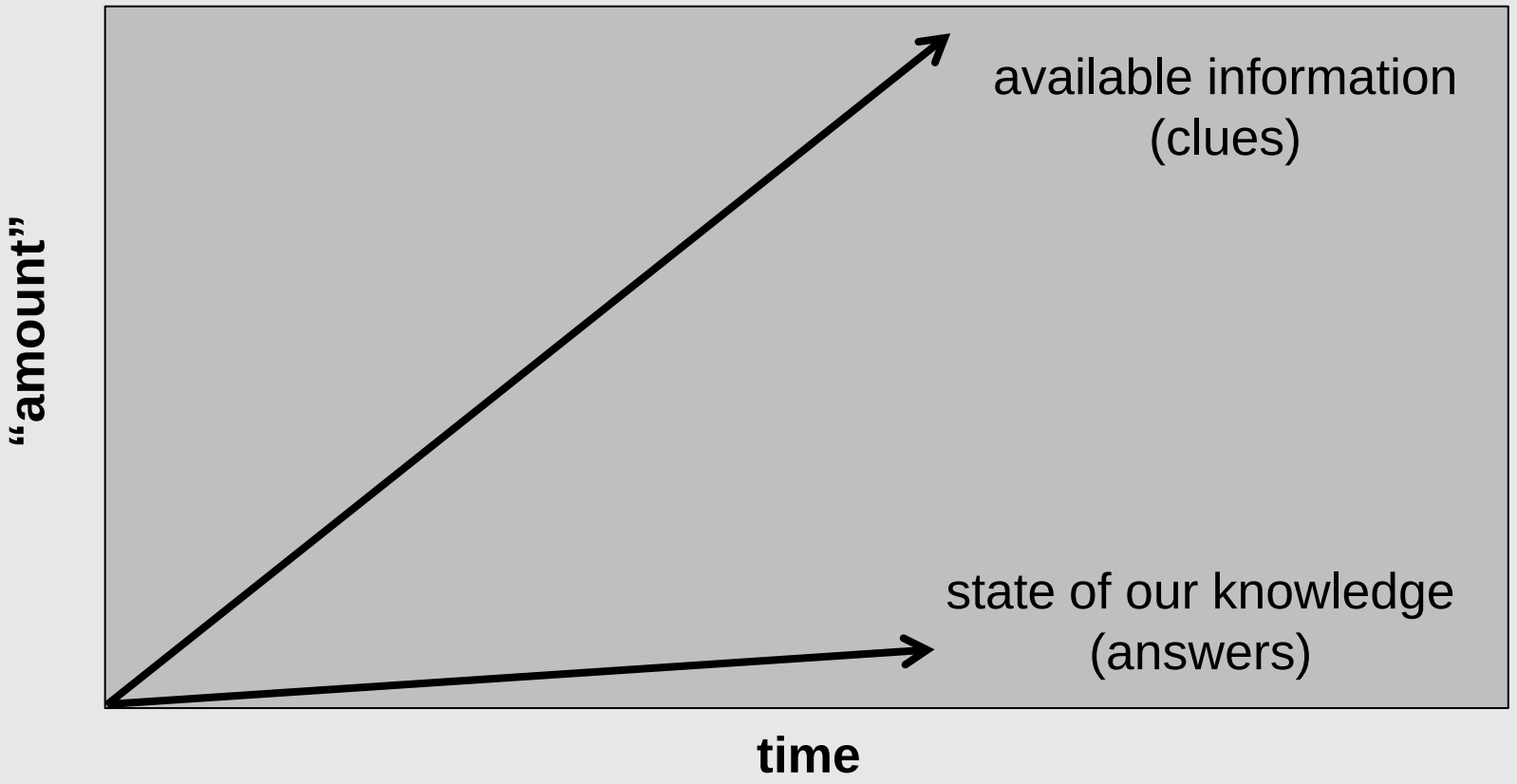


Nucleic acids detection and sequencing techniques



Kahvejian et al., *Nat Biotech* (2008)

Caution required



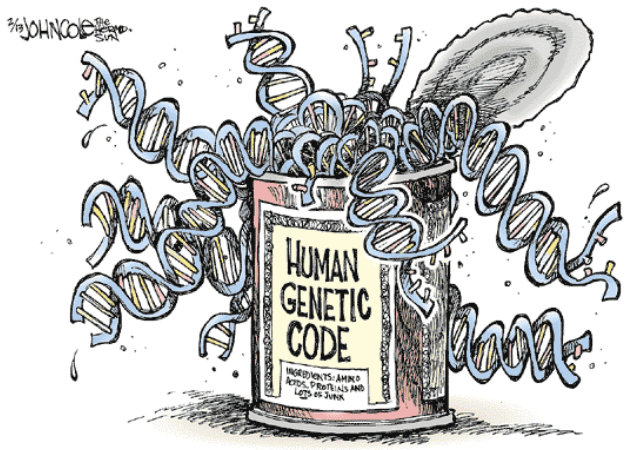
We are drowning in information and starving for knowledge.
Rutherford D. Roger

MPS features

- Unprecedented discovery power
- Hypothesis-free
- Almost unbiased results
- Sensitivity & specificity
- For tag-counting applications truly whole-genome, -transcriptome, -methylome… view
- Only one source of ‘technology’ noise

Acknowledgments

Bettina
Dinko
Jens S
Jonath
Tobias
Jürgen
All our users and former colleagues



Have a nice day!

Science is built with facts as a house is with stones, but a collection of facts is no more a science than a heap of stones is a house.
Jules Henri Poincare