生物信息学：导论与方法
Bioinformatics: Introduction and Methods

生物信息学：导论与方法

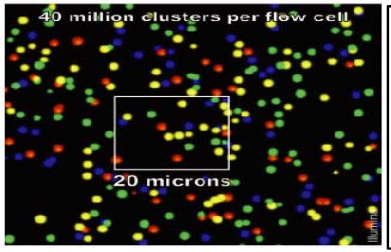
Bioinformatics: Introduction and Methods

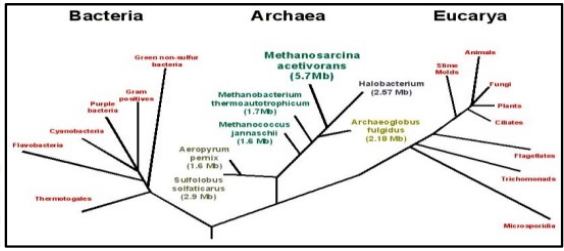
北京大学生物信息学中心 高歌、魏丽萍

Ge Gao & Liping Wei

Center for Bioinformatics, Peking University

TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



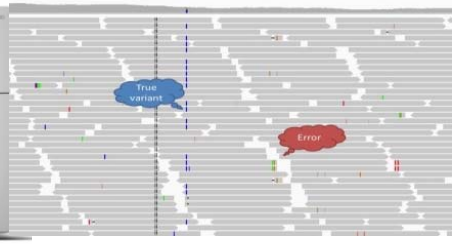


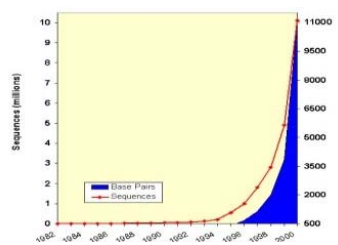
Explore Transcriptome using NGS

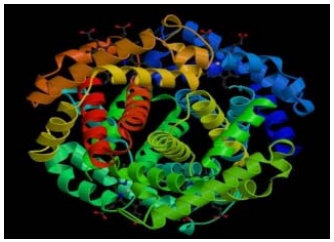
北京大学生物信息学中心 高歌

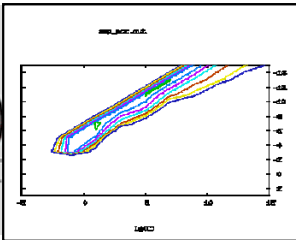
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

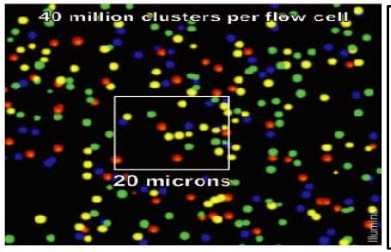


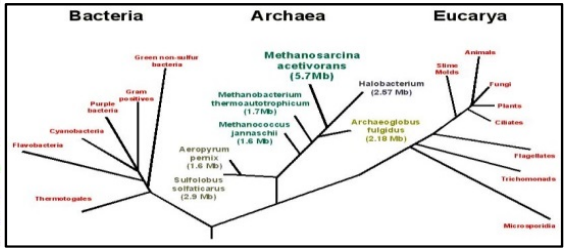






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



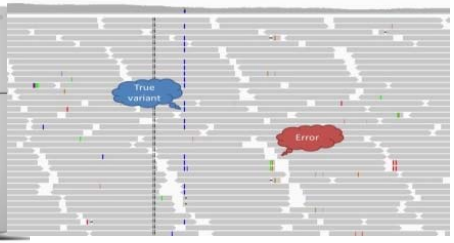


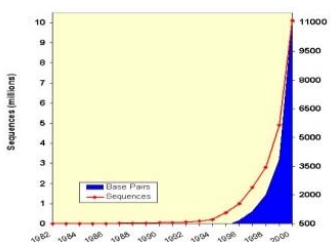
Unit 1: Transcriptome: an Overview

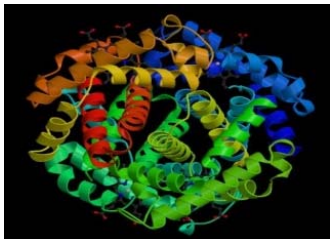
北京大学生物信息学中心 高歌

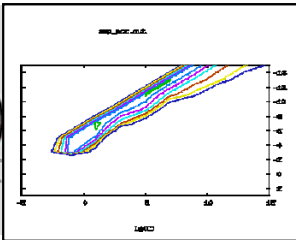
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

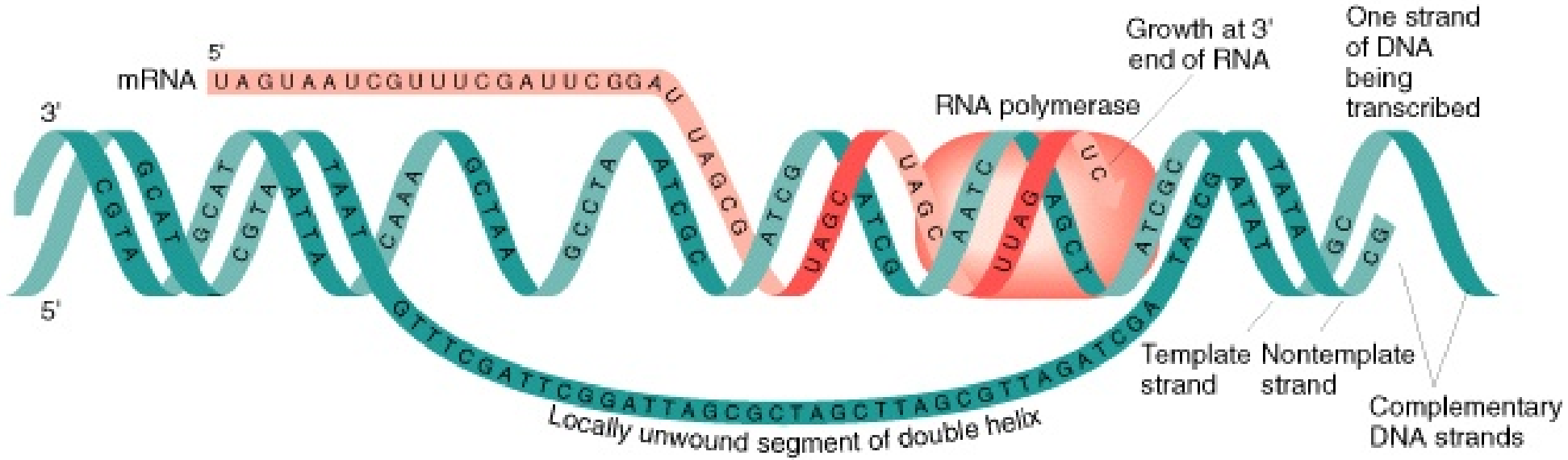






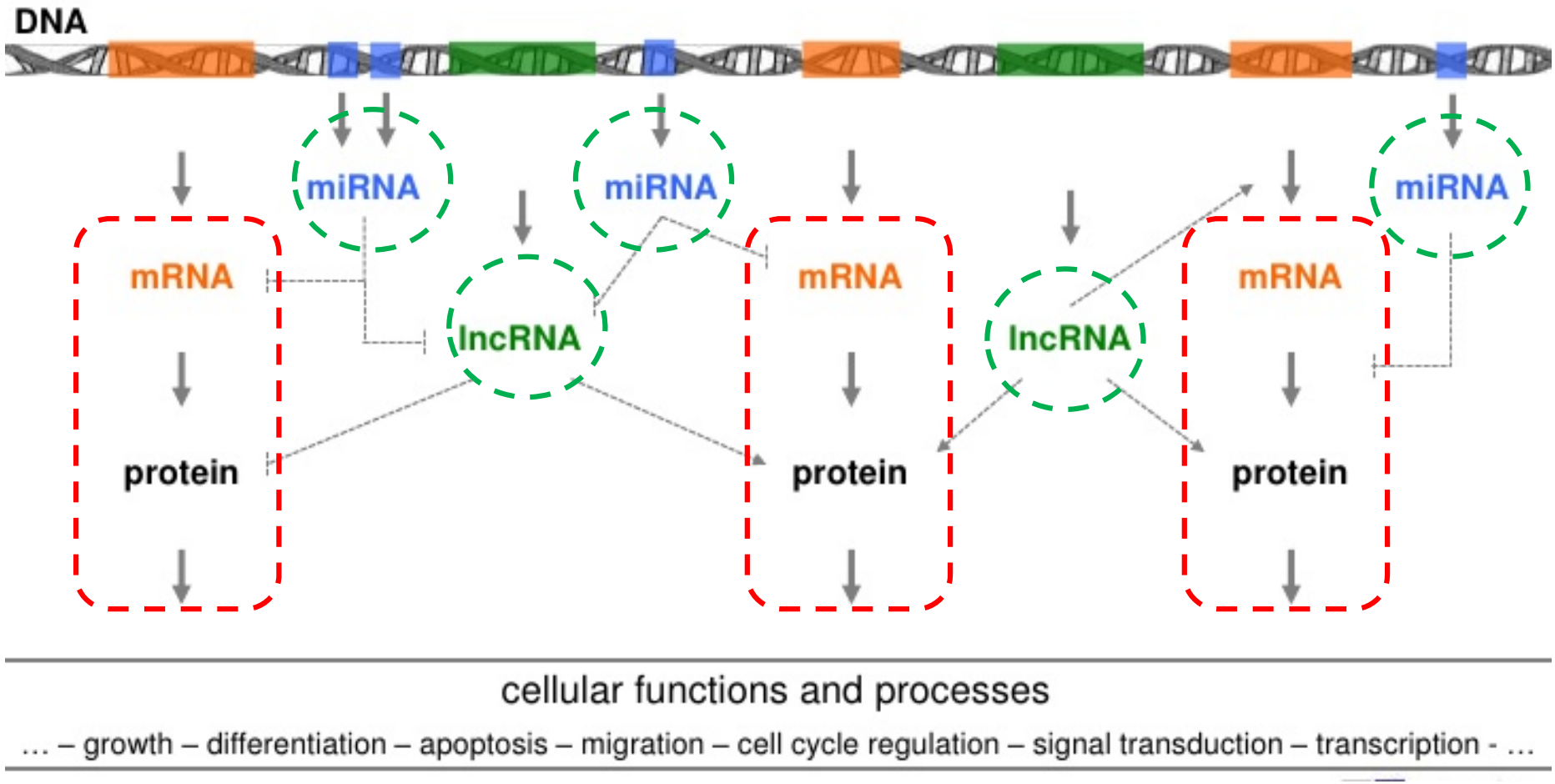


"A transcriptome is a collection of all the transcripts present in a given cell." (NHGRI factsheet, NIH, US)



Source: http://www.mun.ca/biology/scarr/Gr10-11.html

- the transcriptome



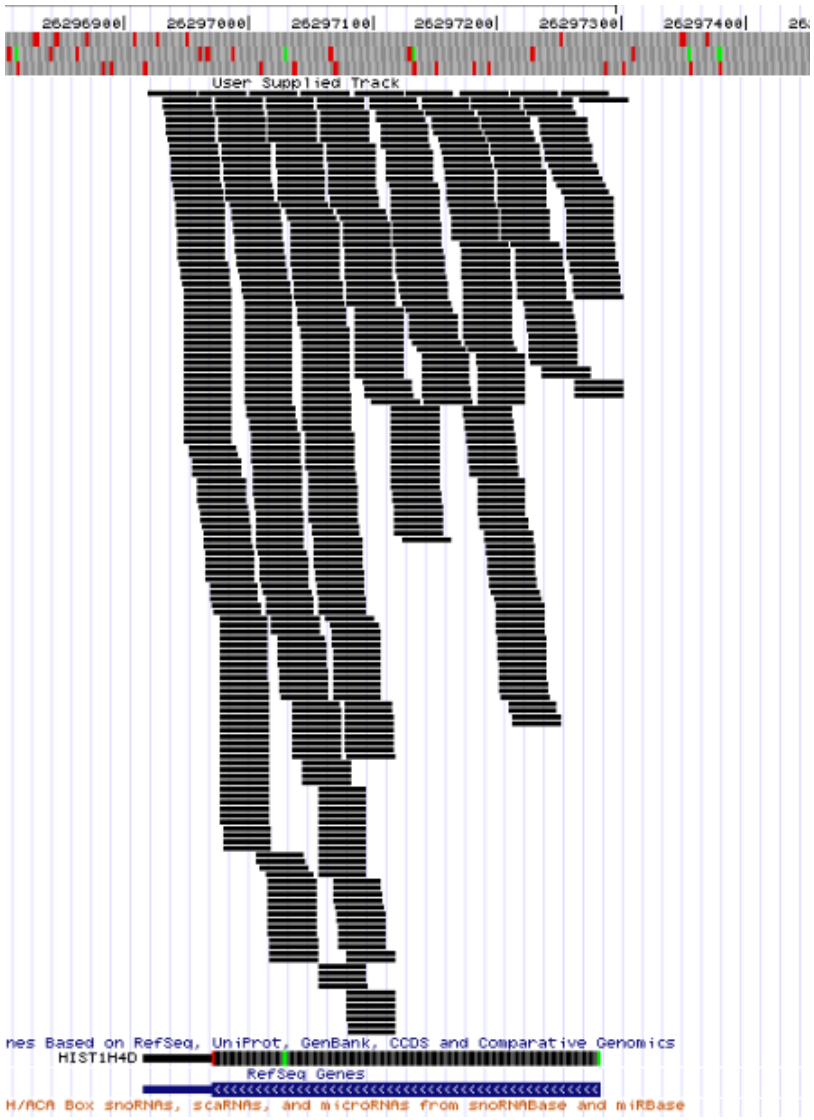
(Modified from http://www.slideshare.net/mateongenaert/05-mestdagh)

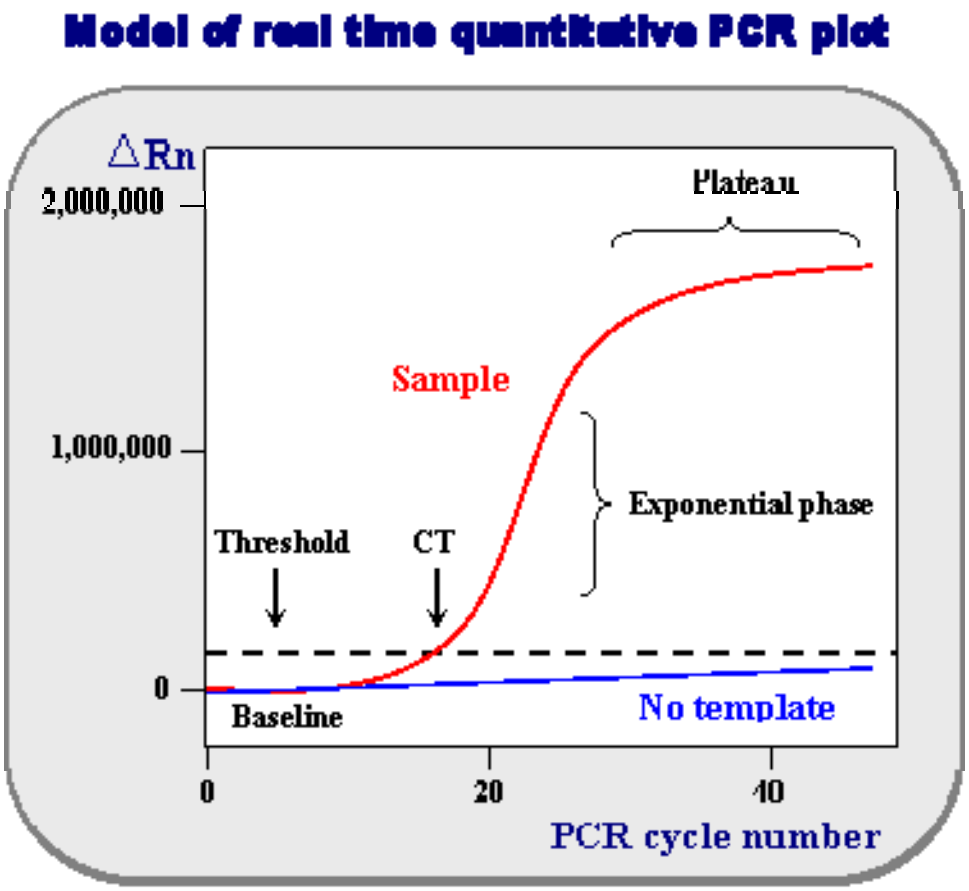
Qualitative

Identify all transcripts, i.e. expressed genes as well as their isoforms

Quantitative

Estimate the expression level of these transcripts, i.e. the transcript abundance of expressed genes/isoforms



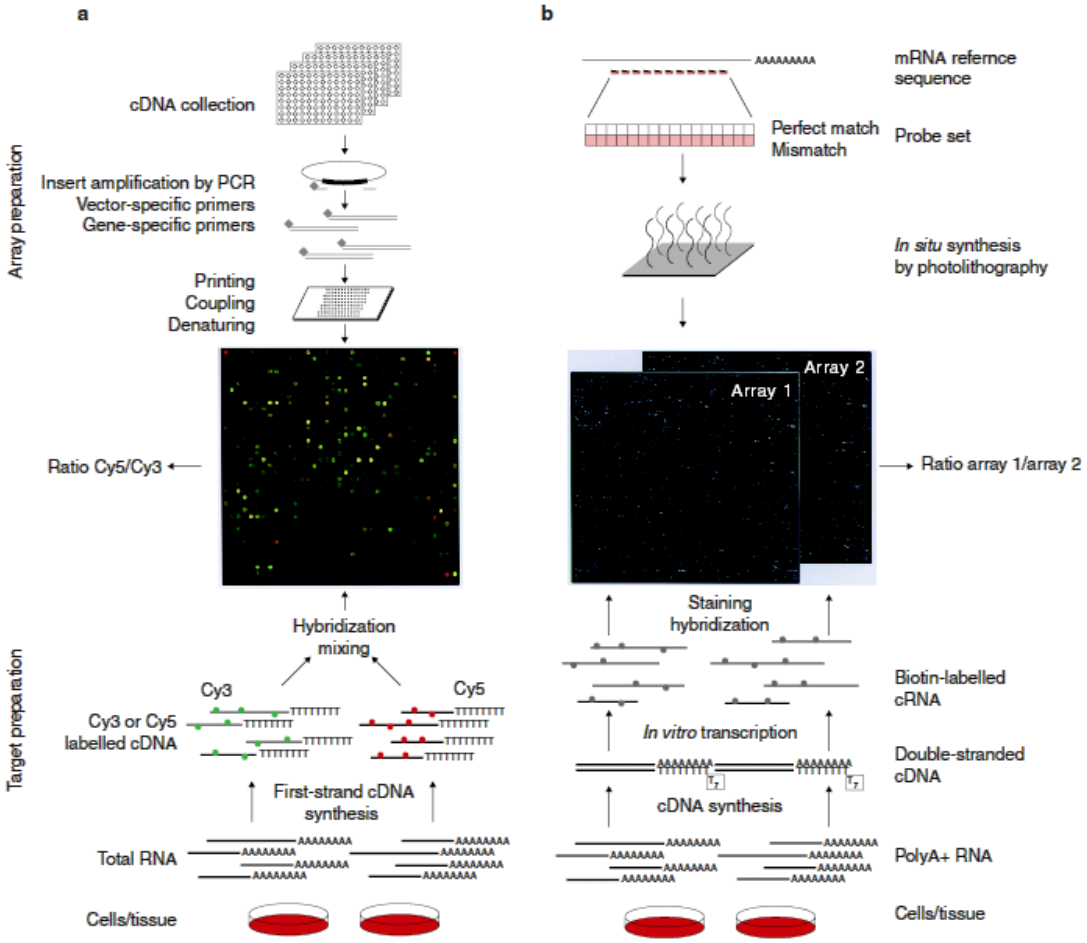


Real-Time qRT-PCR

- Based on complementary hybridization reaction.
  - Development of PCR technology
- Widely accepted as the "Gold Standard"
- Low-throughput
- Prior knowledge of transcript sequences needed!

(Source: http://www.ncbi.nlm.nih.gov/genome/probe/doc/TechQPCR.shtml)

Microarray



(Almut Schulze *et al.*, 2001)

- DNA microarrays are used to analyze gene expression based on complementary hybridization reaction.
- Labeled targets: RNAs derived from biological samples
- Probes: a large number of ordered sets of immobilized nucleotide molecules with known sequences.
- Prior knowledge of transcript sequences needed!

Expresse

Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project

MARK D. ADAMS, JENNY M. KELLEY, JEANNINE D. GOCAYNE, MARK DUBNICK, MIHAEL H. POLYMEROPOULOS, HONG XIAO, CARL R. MERRIL, ANDREW WU, BJORN OLDE, RUBEN F. MORENO, ANTHONY R. KERLAVAGE, W. RICHARD MCCOMBIE, J. CRAIG VENTER*

- Randomly se
  library
  - Short "tag":
  - One-shot: ra
- NO prior kno
  needed!
  - Not only m
- Middle-throu

Automated partial DNA sequencing was conducted on more than 600 randomly selected human brain complementary DNA (cDNA) clones to generate expressed sequence tags (ESTs). ESTs have applications in the discovery of new human genes, mapping of the human genome, and identification of coding regions in genomic sequences. Of the sequences generated, 337 represent new genes, including 48 with significant similarity to genes from other organisms, such as a yeast RNA polymerase II subunit; *Drosophila* kinesin, *Notch*, and *Enhancer of split*; and a murine tyrosine kinase receptor. Forty-six ESTs were mapped to chromosomes after amplification by the polymerase chain reaction. This fast approach to cDNA characterization will facilitate the tagging of most human genes in a few years at a fraction of the cost of complete genomic sequencing, provide new genetic markers, and serve as a resource in diverse biological research fields.

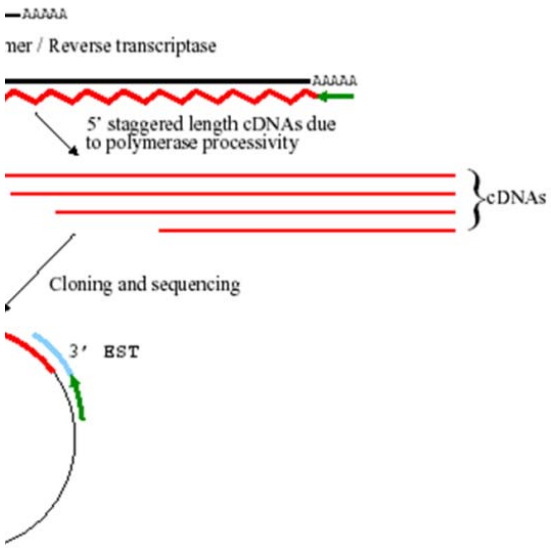
THE HUMAN GENOME IS ESTIMATED TO CONSIST OF 50,000 to 100,000 genes, up to 30,000 of which may be expressed in the brain (*1*). However, GenBank lists the sequence of only a few thousand human genes and <200 human brain messenger RNAs (mRNAs) (*2*). Once dedicated human chromosome sequencing begins in 5 years, it is expected that 12 to 15 years will be required to complete the sequence of the genome (*3*). It is therefore likely that the majority of human genes will remain unknown for at least the next decade. The merits of sequencing cDNA, reverse transcribed from mRNA, as a part of the human genome project have been vigorously debated since the idea of determining the complete nucleotide sequence of humans first surfaced. Proponents of cDNA sequencing have argued that because the coding sequences of genes represent the vast majority of the information content of the genome, but only 3% of the DNA, cDNA sequencing should take precedence over genomic sequencing (*4*). Proponents of genomic sequencing have argued the difficulty of finding every mRNA expressed in all tissues, cell types, and developmental stages and have pointed out that much valuable information from intronic and intergenic regions, including control and regulatory sequences, will be missed by cDNA sequencing (*5*). However, many genome enthusiasts have incorrectly stated that gene coding regions, and therefore mRNA sequences, are readily predictable from genomic sequences and have concluded that there is no need for large-scale cDNA sequencing. In fact, prediction of transcribed regions of human genomic sequence is currently feasible only for relatively large exons (*6*).

On the basis of our high output with automated DNA sequence analysis of 96 templates per day and consideration of the above issues, we initiated a pilot project to test the use of partial cDNA sequences (ESTs) in a comprehensive survey of expressed genes.

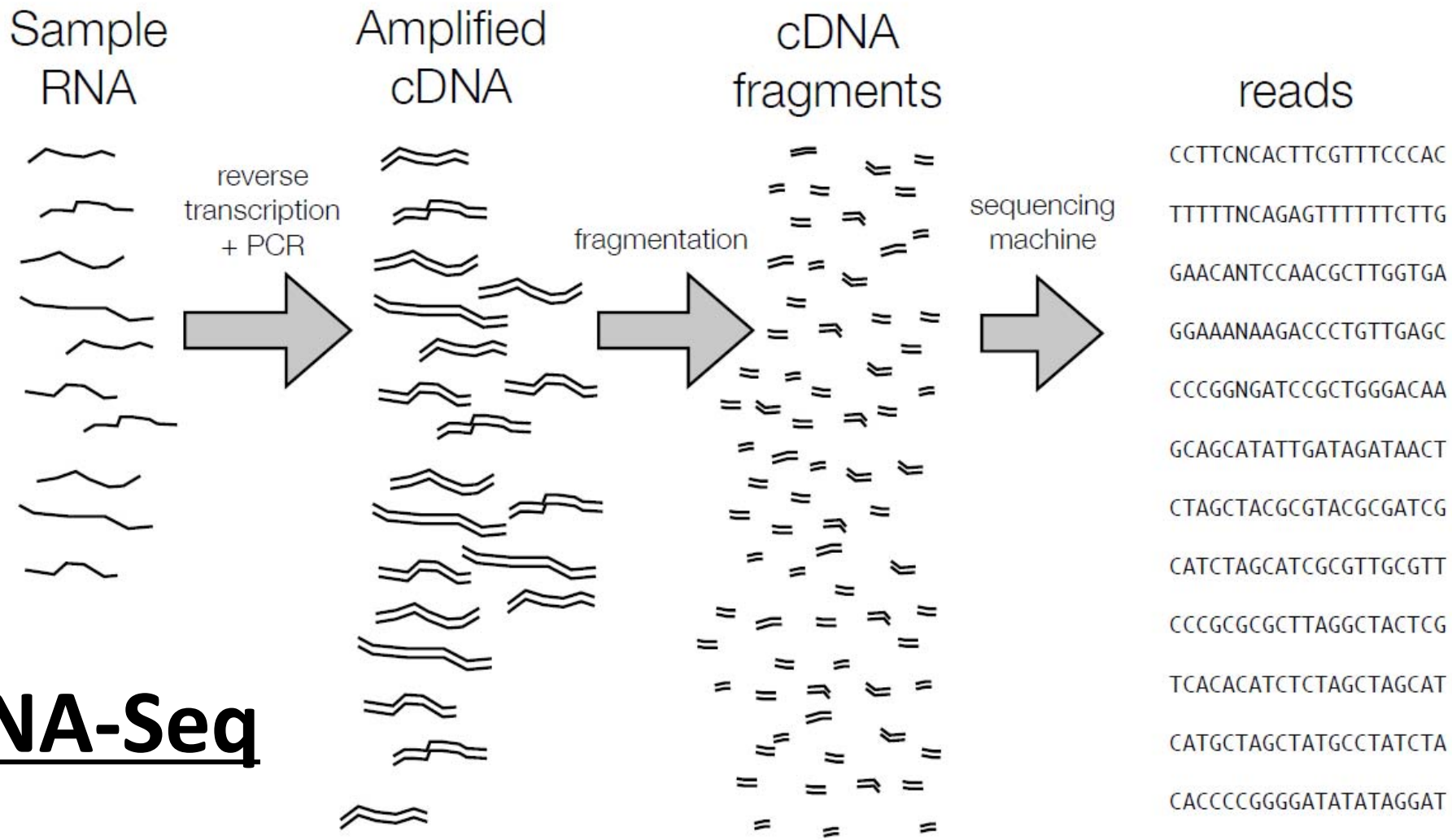
Sequence-tagged sites (STSs) are becoming standard markers for the physical mapping of the human genome (*7*). These short sequences from physically mapped clones represent uniquely identified map positions. ESTs can serve the same purpose as the random genomic DNA STSs and provide the additional feature of pointing directly to an expressed gene. An EST is simply a segment of a sequence from a cDNA clone that corresponds to an mRNA. ESTs longer than 150 bp were found to be the most useful for similarity searches and mapping.

M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter are in the Section of Receptor Biochemistry and Molecular Biology, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892. M. H. Polymeropoulos, H. Xiao, and C. R. Merril are in the Laboratory of Biochemical Genetics, National Institute of Mental Health, Neuroscience Center at St. Elizabeth's Hospital, Washington, DC 20032.

*To whom correspondence should be addressed.

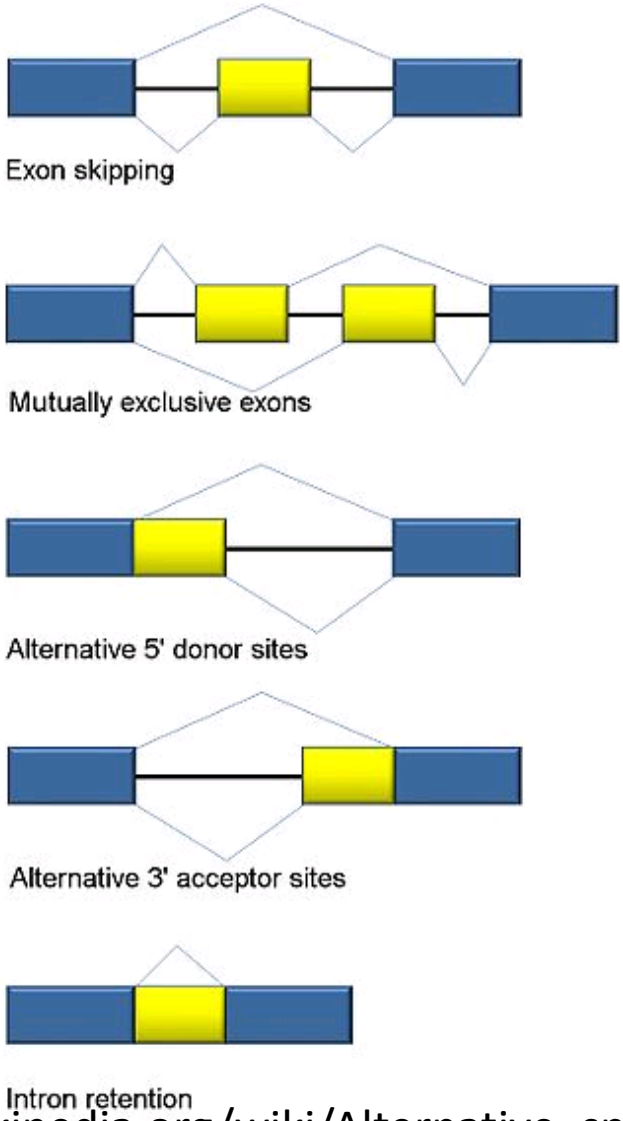


RNA-Seq

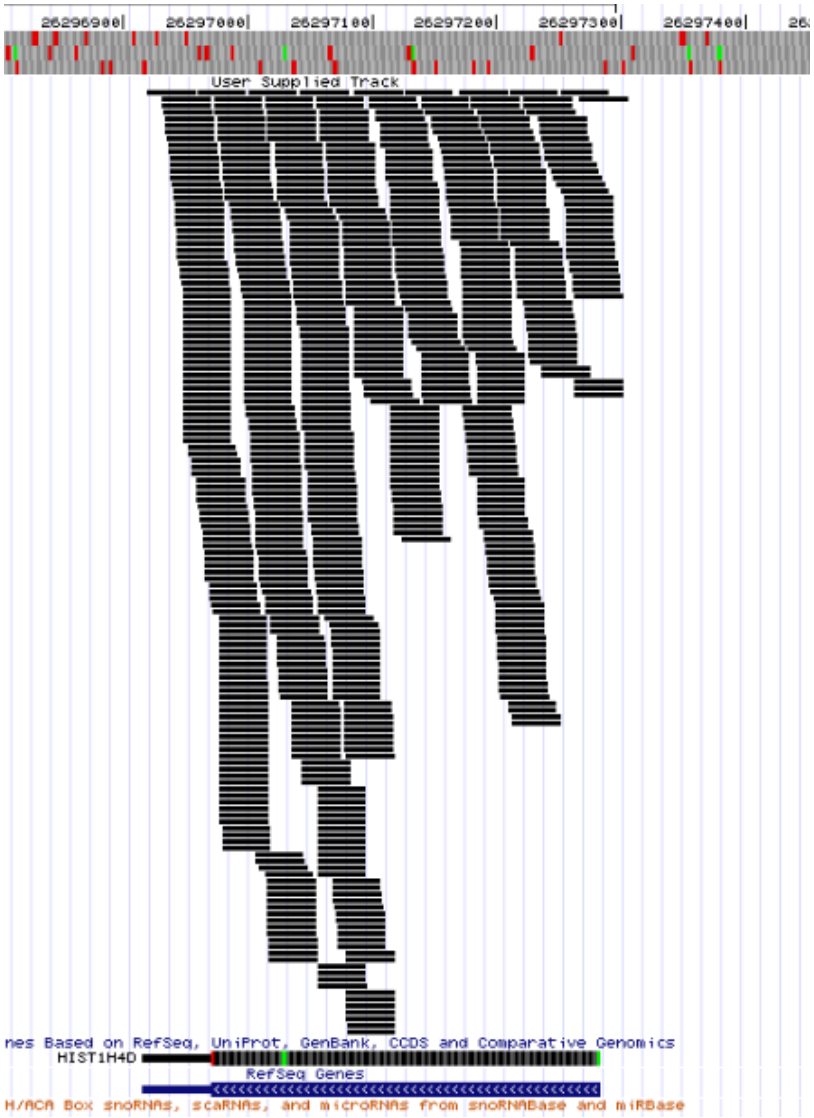


(Modified from Colin Dewey slides at www.biostat.wisc.edu/bmi776/)

Qualitative

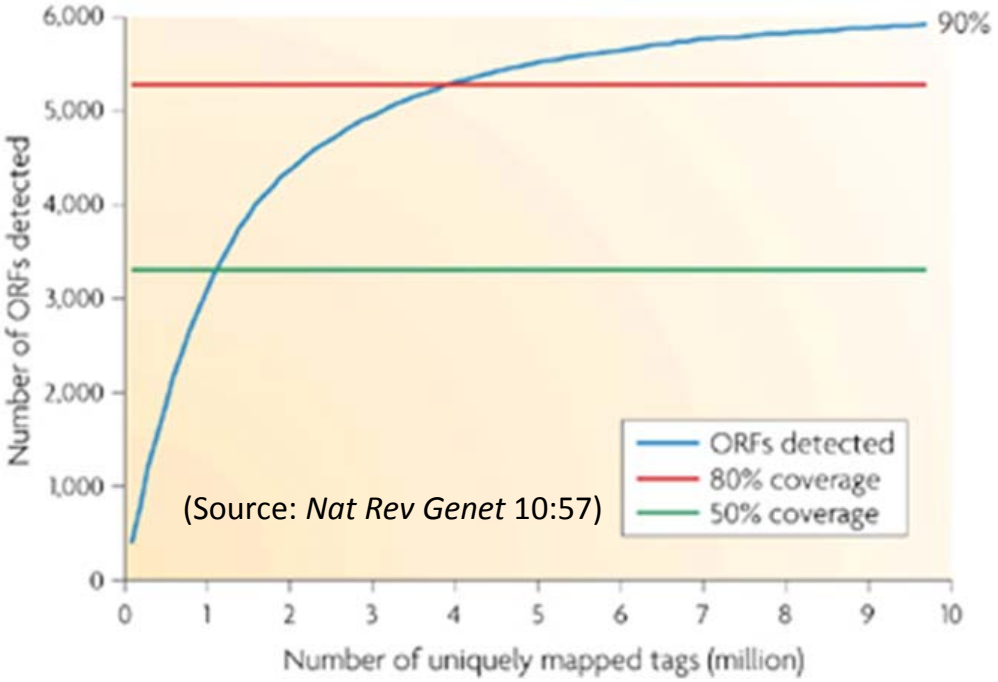


Quantitative



http://en.wikipedia.org/wiki/Alternative_splicing

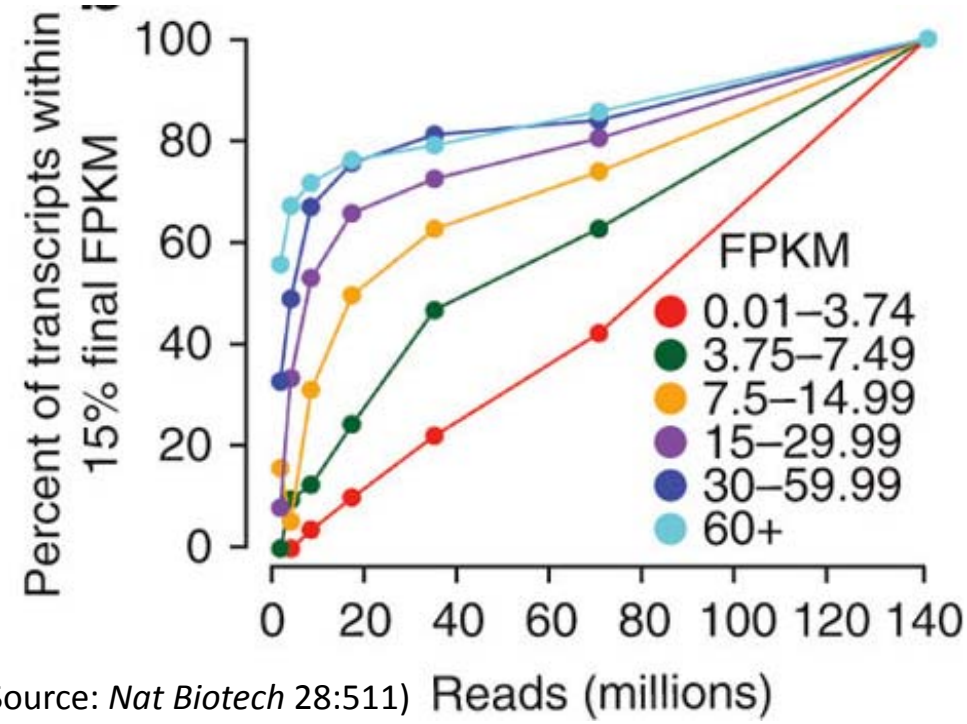
Random sampling the transcriptome



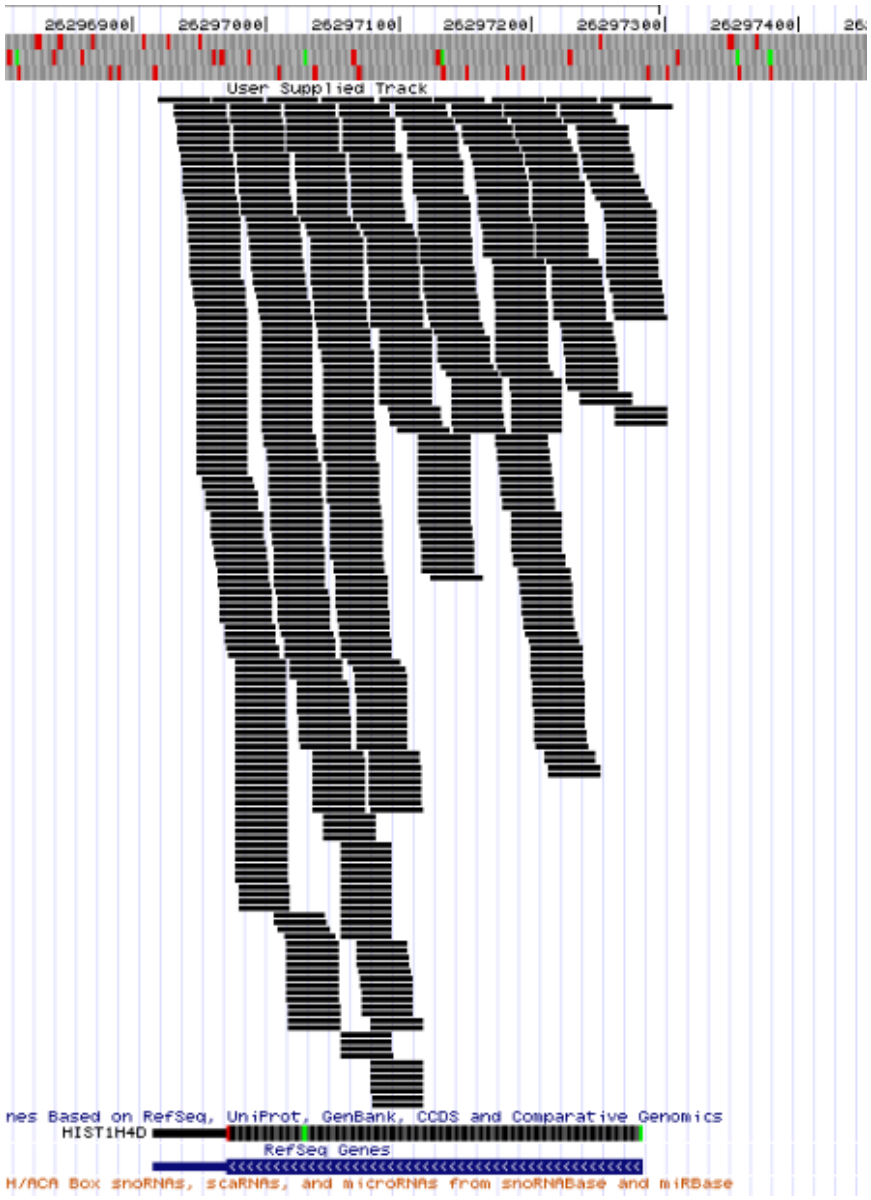
(Source: *Nat Rev Genet* 10:57)

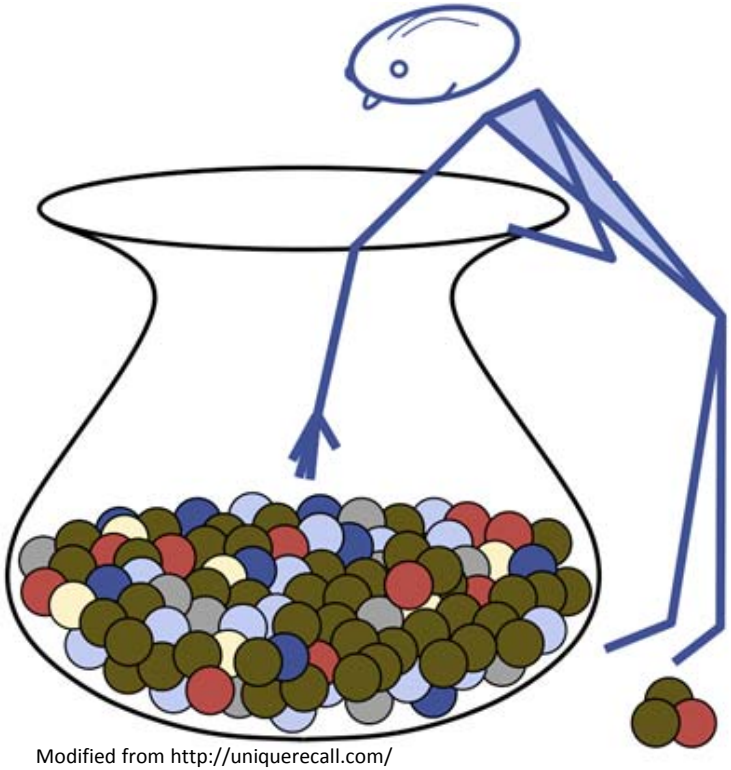
The detection power as well as sensitivity of RNA-Seq is highly dependent on the sequencing depth.

- 100~150x as a decent start for a typical mammalian transcriptome RNA-Seq.



(Source: *Nat Biotech* 28:511)



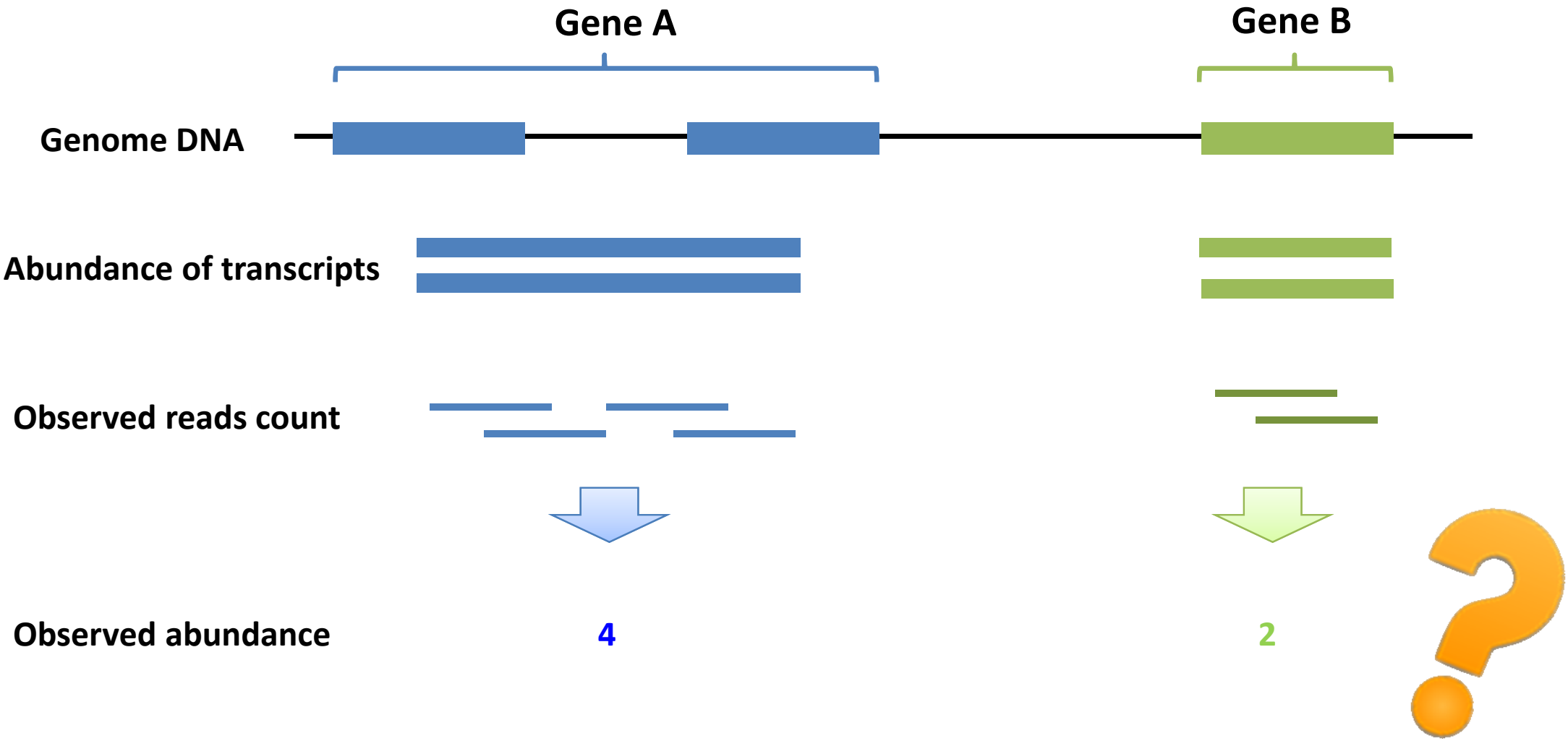


Modified from http://uniquerecall.com/

# of mapped reads ∝ transcript abundance

# of mapped reads ∝ transcript length

# of mapped reads ∝ library depth

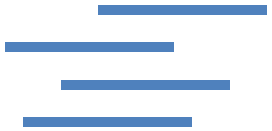


Experiment 1

Gene B

Abundance of transcripts

Total reads: 2n



Experiment 2

Gene B

Total reads: n

From raw count to expression level

RPKM: the number of mapped **R**eads ***p****er* **K**B *per* **m**illion reads.

$$RPKM = 10^9 \frac{\mathrm{C}}{\mathrm{NL}}$$

- C: the number of mapped reads for specified transcript.
- N: the number of total mapped reads.
- L: the length of the specified transcript.

Mortazavi A. *et al*. *Nature Methods* 2008

BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 6. 671–683
Advance Access published on 17 September 2012
doi:10.1093/bib/bbs046

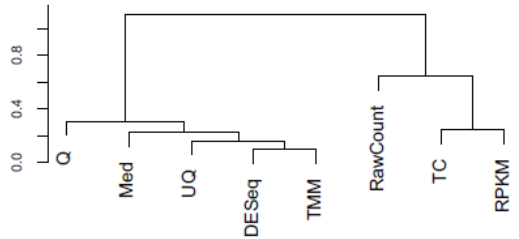
A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

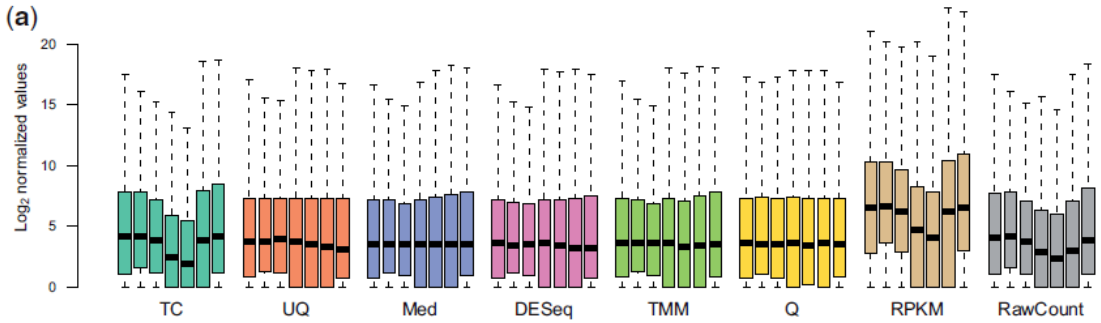
Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

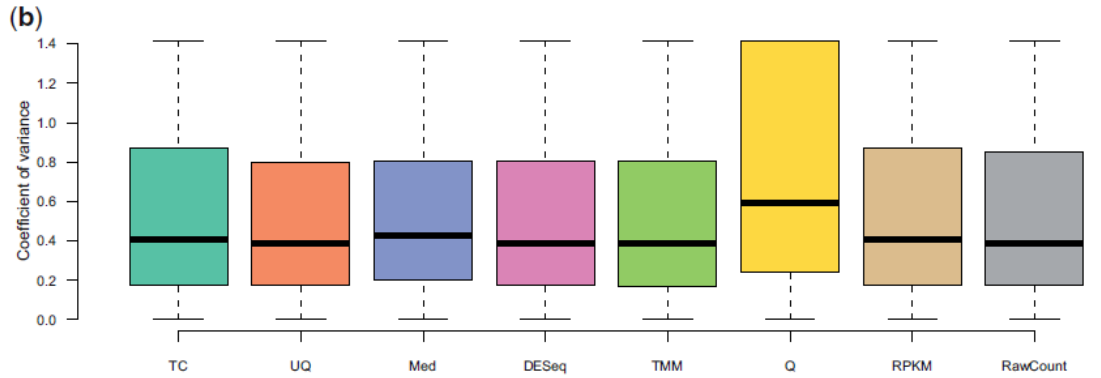
Submitted: 12th April 2012; Received (in revised form): 29th June 2012

Abstract

During the last 3 years, a number of approaches for the normalization of RNA sequencing data have emerged in the literature, differing both in the type of bias adjustment and in the statistical strategy adopted. However, as data continue to accumulate, there has been no clear consensus on the appropriate normalization method to be used or the impact of a chosen method on the downstream analysis. In this work, we focus on a comprehensive comparison of seven recently proposed normalization methods for the differential analysis of RNA-seq data, with an emphasis on the use of varied real and simulated datasets involving different species and experimental designs to represent data characteristics commonly observed in practice. Based on this comparison study, we propose practical recommendations on the appropriate normalization method to be used and its impact on the differential analysis of RNA-seq data.



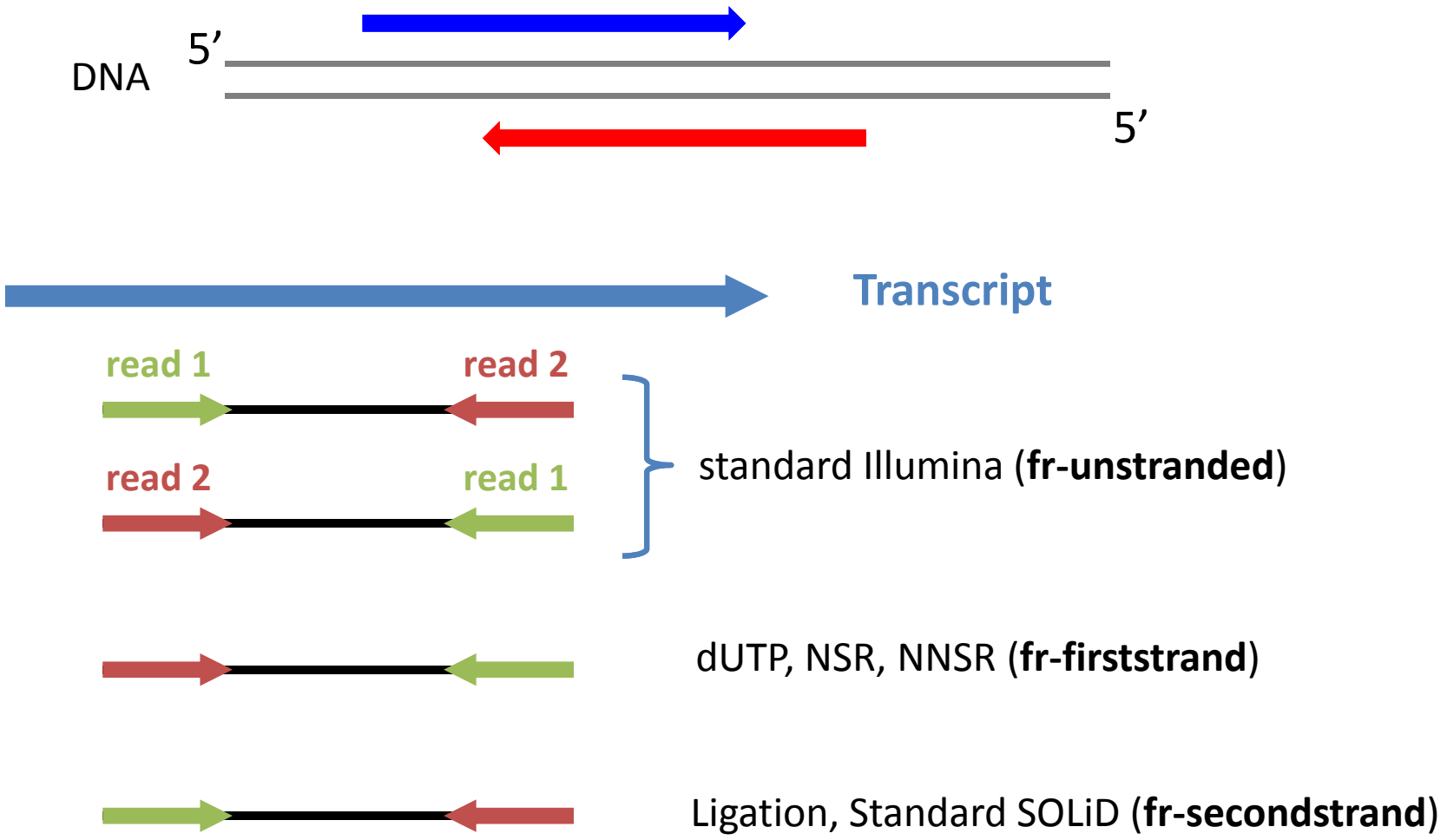




Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	–	+	++	–
RPKM	–	+	+	–	–

A '–' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

Briefings in Bioinformatics. 14(6):671



Summary Questions

Please read the paper “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis” by Marie-Agnes Dillies *et al*. (*Briefings in Bioinformatics*. 14(6):671) first, and

- Re-phrase the (biological) assumptions for each normalization algorithms mentioned in the paper, then
- Explain the Table 3

生物信息学：导论与方法
Bioinformatics: Introduction and Methods