L3.2 – Transcriptomes
RNA-seq

AGENDA

1. Pre-NGS Transcripts Annotation
2. RNA-seq protocol and basic variations
3. Mapping of RNAseq data for transcripts annotation (Qualitative)
4. Gene expression studies by RNAseq (Quantitative)
5. Extra material on NGS sequencing platforms

UNBIASED Transcriptome Analysis

pre-NGS:

- Tiling microarrays
- SAGE
- CAGE

NGS:

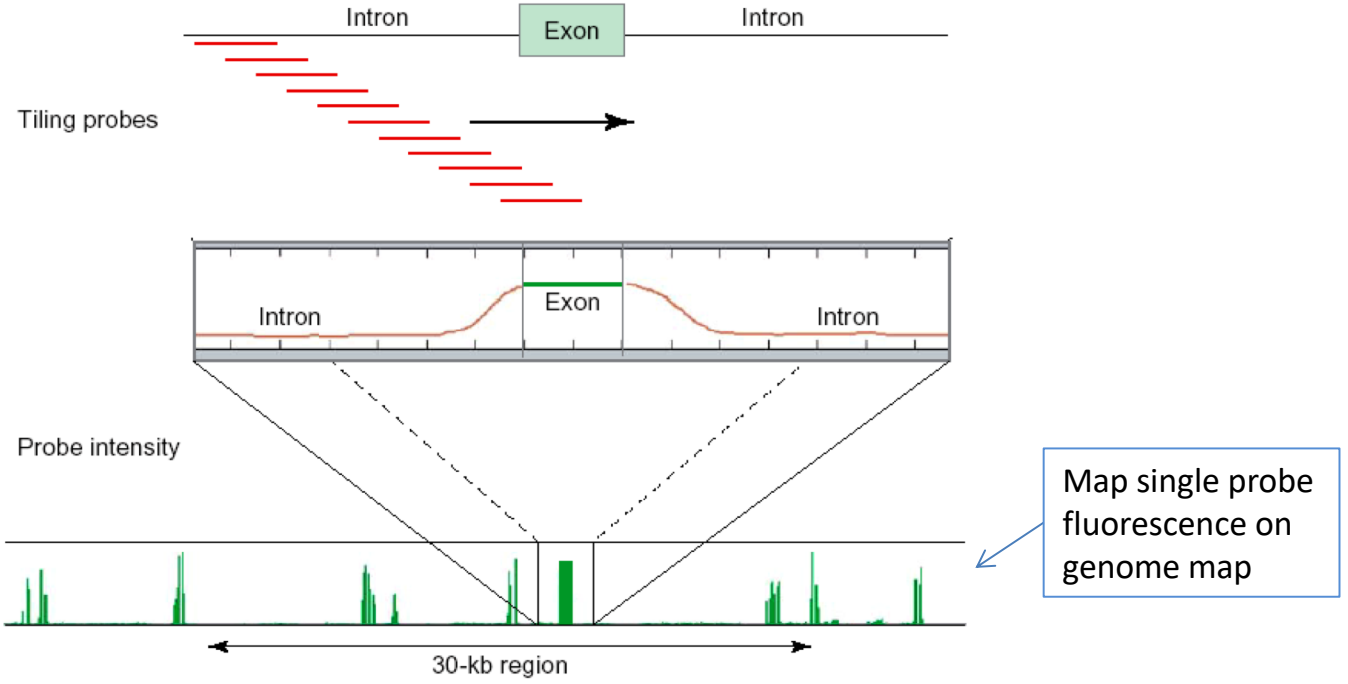
- CAGE
- RNA-Seq (all flavors)
- Strand-specific RNA-seq
- GRO-seq (genomic run-on-seq)
- …

Tiling microarrays (2002-2007)

Qualitative

A tiling microarray is composed of probes that cover (nonrepetitive) genome sequences, irrespective of gene prediction

Millions probes required ! Human Genome: 3.2*10E9 → 1.5 nonrepetitive
You would need 50 millions 30-mer probes !

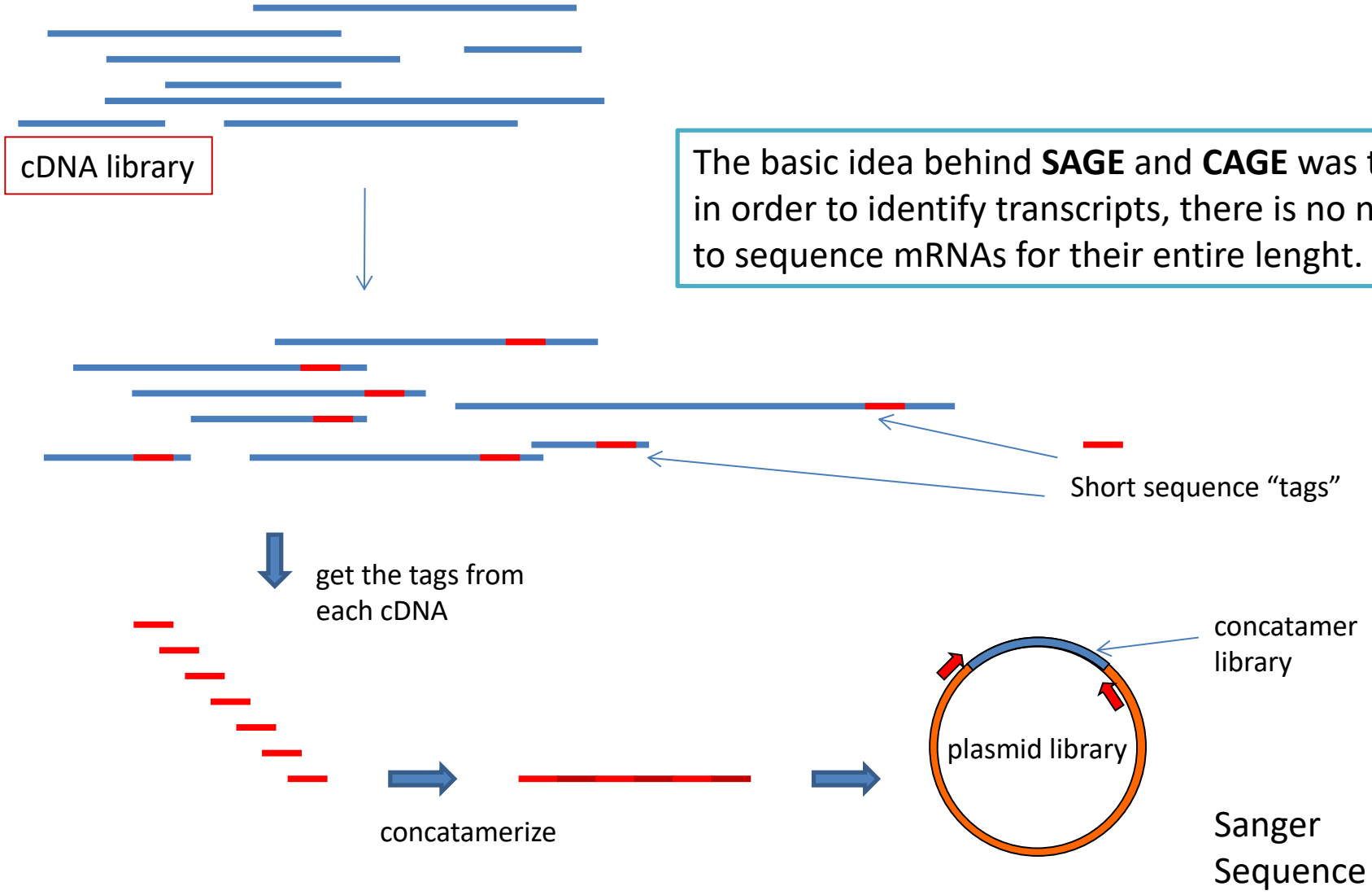


Box 1. Tiling microarray experiments

Tiling microarrays are designed to assay transcription at regular intervals of the genome using regularly spaced probes (horizontal red lines) that can be overlapping (Figure I) or separated. The distance between the centers of successive probes is the 'step' size and probes can be selected to be complementary to one strand (as shown) or both strands. Probes can be synthesized directly onto or spotted onto glass slides, and can be synthesized oligonucleotides or PCR products. They are hybridized with fluorescently labeled cRNA or cDNA prepared from cell samples. Regions of greater fluorescent intensity (green peaks in lower panel) can reveal transcription within a large genomic region. In addition, the correlation of probe intensities in several different tissues (co-expression analysis) can be used to identify probes that are detecting exons of the same transcript. The lower panel shows the extent of a hypothetical transcript within the genome. The middle panel is a schematic, magnified view of the hybridization of a genomic region containing an exon.

SAGE, CAGE 1.0 (1995-2005)

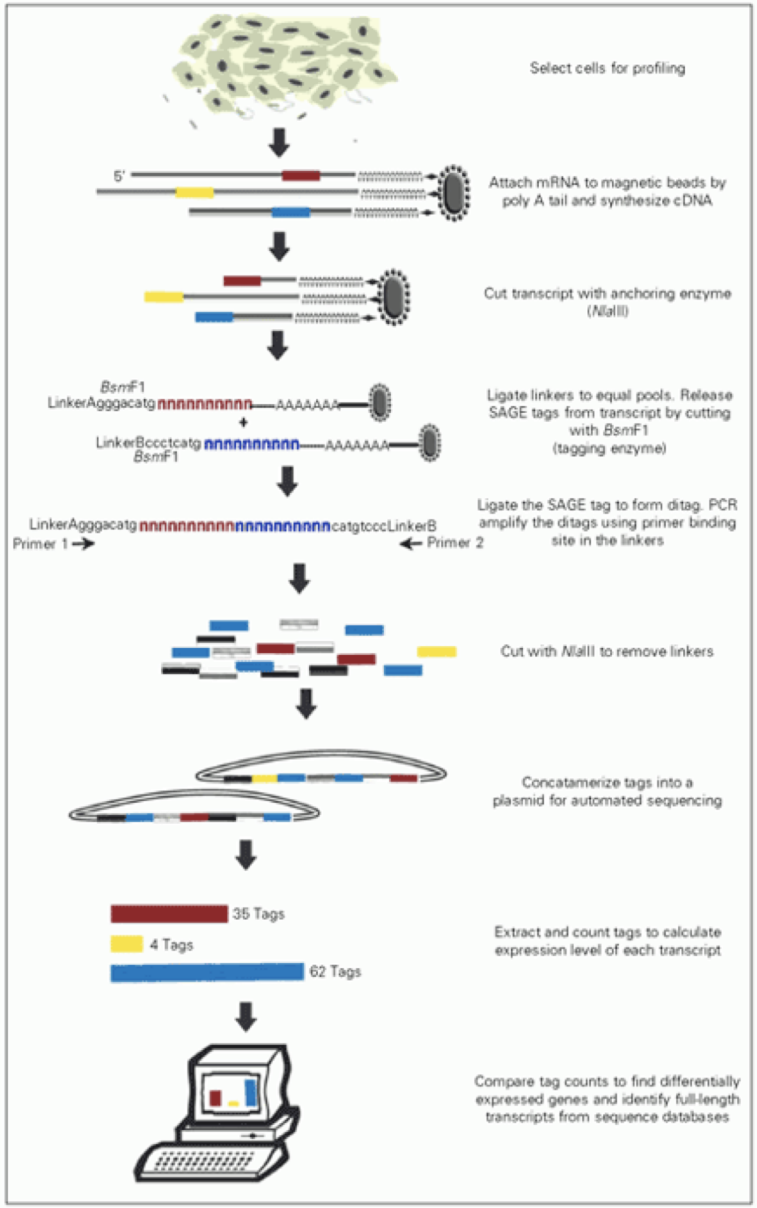
SAGE Serial Analysis of Gene Expression
CAGE Cap Analysis of Gene Expression.



The basic idea behind **SAGE** and **CAGE** was that in order to identify transcripts, there is no need to sequence mRNAs for their entire lenght.

SAGE

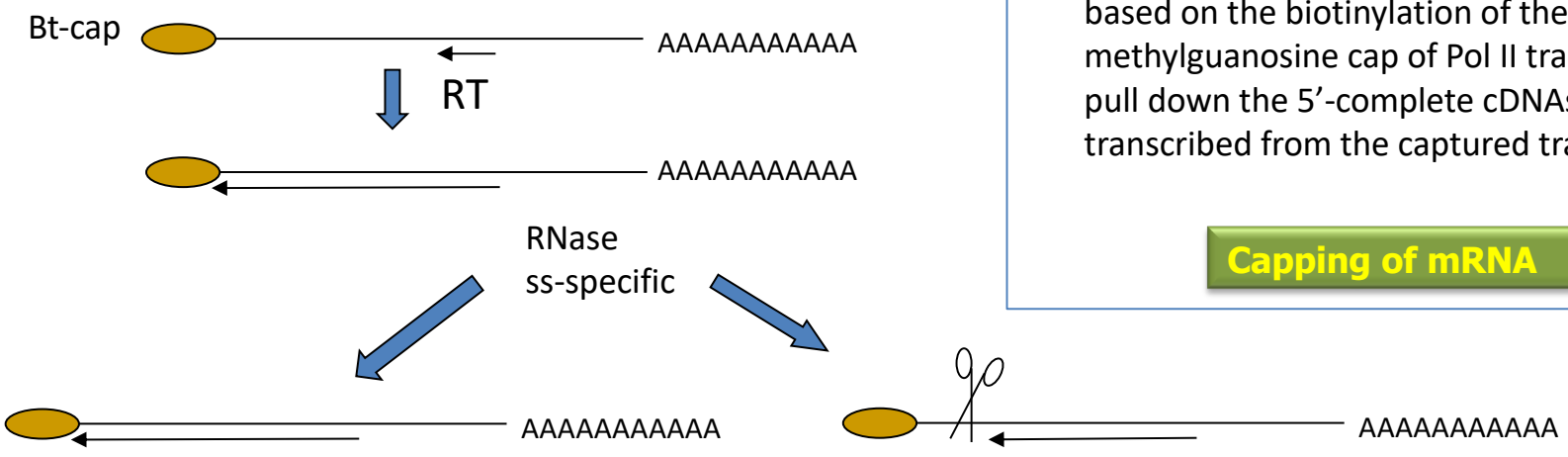
How library preparation for SAGE works



- SAGE tags are short sequences from the 3’ end
- The number of times each tag is present is proportional to the amount of mRNA present (Quantitative Information)
- Lots of SAGE data present in NCBI

CAGE

How library preparation for CAGE works



- CAGE utilizes a “cap-trapping” technology based on the biotinylation of the 7-methylguanosine cap of Pol II transcripts, to pull down the 5’-complete cDNAs reversely transcribed from the captured transcripts.

Capping of mRNA

cDNA not fully extended leave some ssRNA
CAP is cleaved

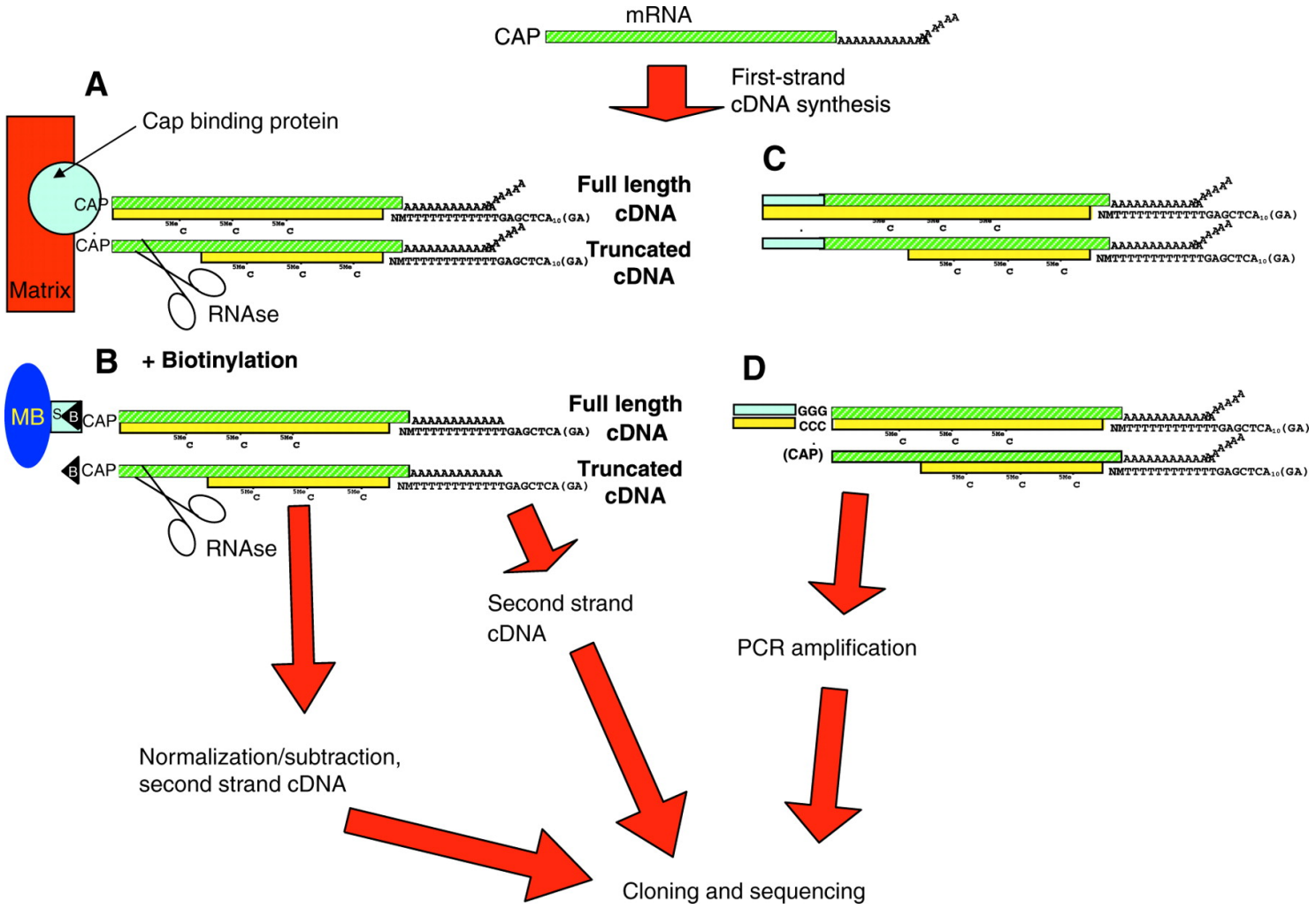
Streptavidin-sepharose selection

RNA removed

cDNA completed by linker ligation - extension

- CAGE ha been instrumental for identifying TSSes and promoters
- CAGE 2.0 – combination with deep sequencing (used in FANTOM and ENCODE projects)

Schematic representation of different methods for preparing full-length cDNA libraries.



Piero Carninci J Exp Biol 2007;210:1497-1506

The Transcriptional Landscape of the Mammalian Genome

The FANTOM Consortium* and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group)*

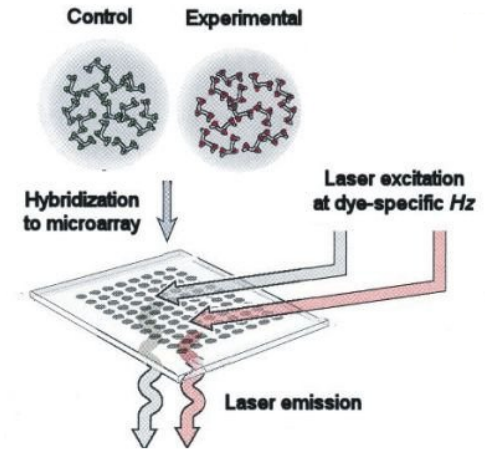
This study describes comprehensive polling of transcription start and termination sites and analysis of previously unidentified full-length complementary DNAs derived from the mouse genome. We identify the 5′ and 3′ boundaries of 181,047 transcripts with extensive variation in transcripts arising from alternative promoter usage, splicing, and polyadenylation. There are 16,247 new mouse protein-coding transcripts, including 5154 encoding previously unidentified proteins. Genomic mapping of the transcriptome reveals transcriptional forests, with overlapping transcription on both strands, separated by deserts in which few transcripts are observed. The data provide a comprehensive platform for the comparative analysis of mammalian transcriptional regulation in differentiation and development.

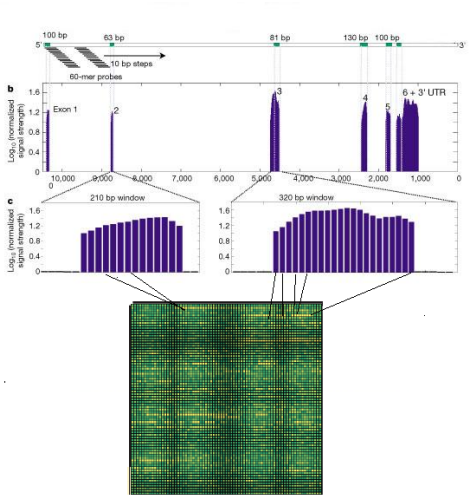
SCIENCE VOL 309 2 SEPTEMBER 2005 **1559**

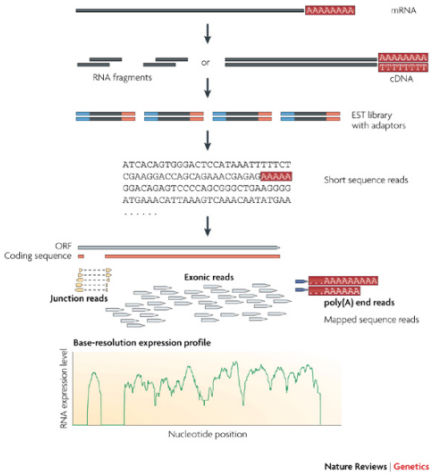
CAGE Website:

The Riken FANTOM project analyzed approximately 1000 kinds of samples using the CAGE method. With those results, the activity of approximately 185,000 promoter sites and 44,000 enhancer sites were identified. Half of the identified promoters were discovered for the first time. This suggest that there are at least 3 alternative promoters for each gene on average.

The evolution of transcriptomics







1995 P. Brown, et. al.
Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

Textbook

INNOVATION

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, Mark Gerstein and Michael Snyder

Abstract | RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods. This article describes the RNA-Seq approach, the challenges associated with its application, and the advances made so far in characterizing several eukaryote transcriptomes.

Wang et al. (2009). Nat. Rev. Genet. 10:57-69

General concept: A population of RNA (total or fractionated) is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing). Reads can be 30–400 bp long depending on the DNA sequencing technology used.

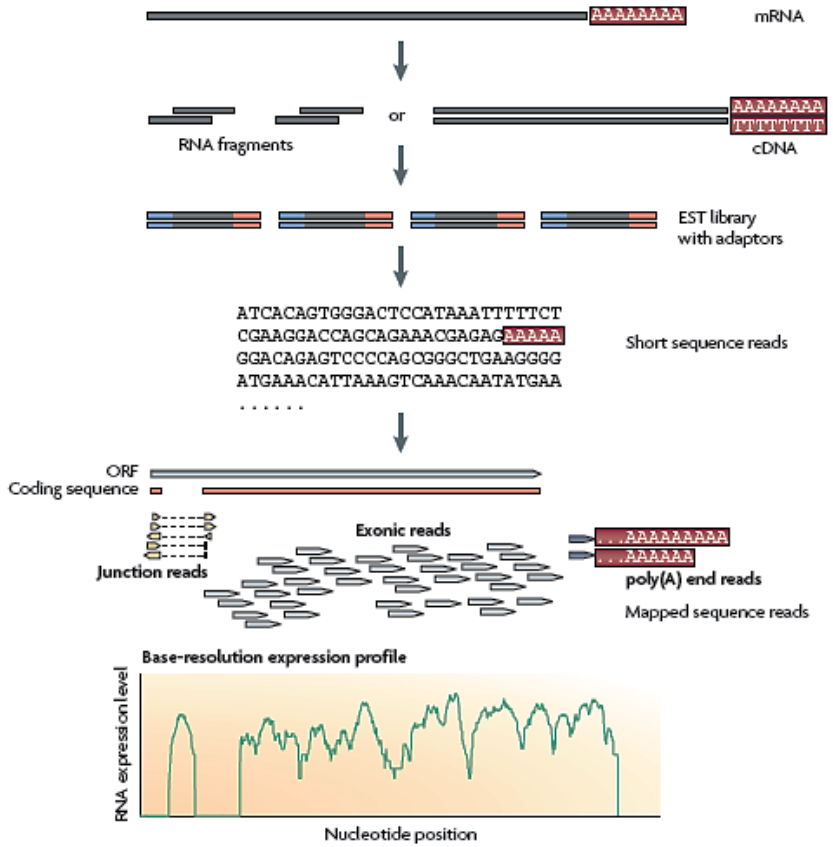


Figure 1 | **A typical RNA-Seq experiment.** Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation (see main text). Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown.

Sample preparation

Next generation sequencing (NGS)

Data analysis

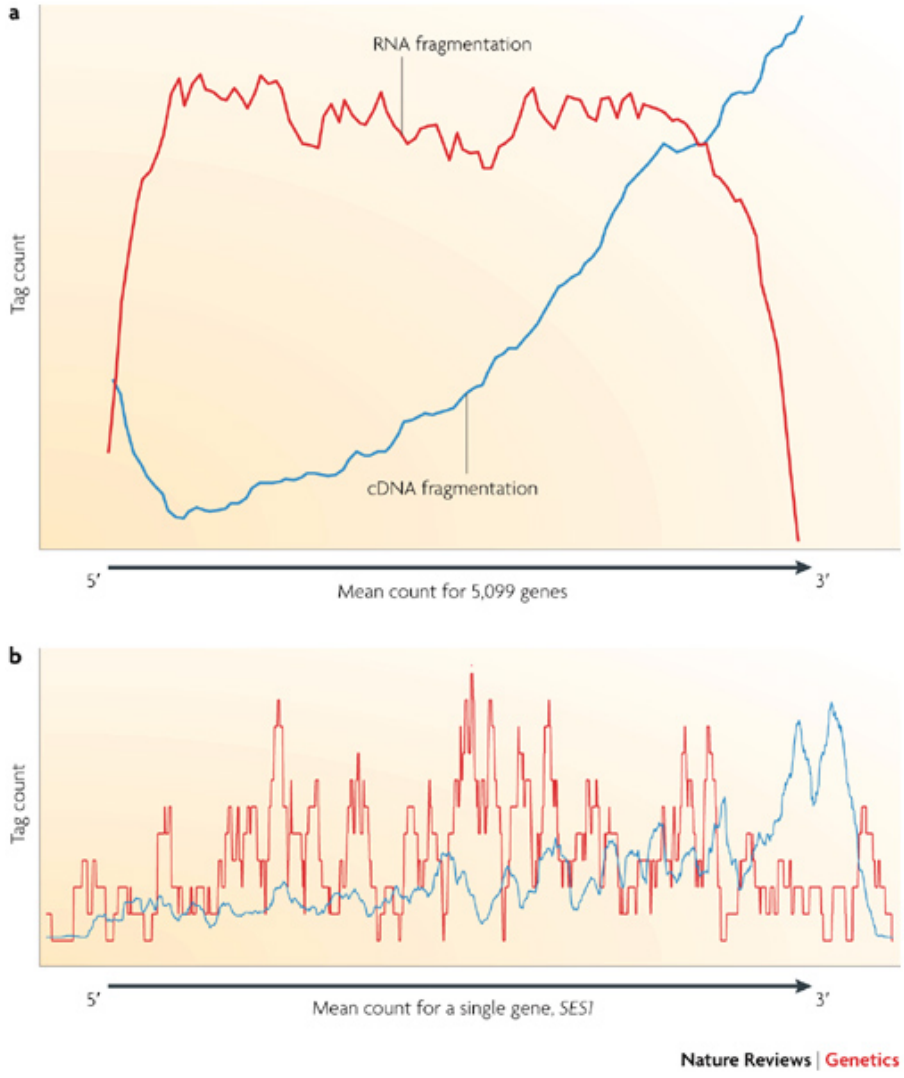
- ✓Mapping reads
- ✓Visualization (Gbrowser)
- ✓De novo assembly
- ✓Quantification

Wang et al. (2009)

Sample Preparation: Challenges for Library Construction

- Unlike small RNAs (microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), short interfering RNAs (siRNAs) and many others), which can be directly sequenced after adaptor ligation, larger RNA molecules must be fragmented into smaller pieces (200–500 bp) to be compatible with most deep-sequencing technologies.
- Common fragmentation methods include RNA fragmentation (RNA hydrolysis or nebulization) and cDNA fragmentation (DNase I treatment or sonication).
- Each of these methods creates a different bias in the outcome.

Sample preparation



Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends.

A specific yeast gene, SES1 (seryl-tRNA synthetase)

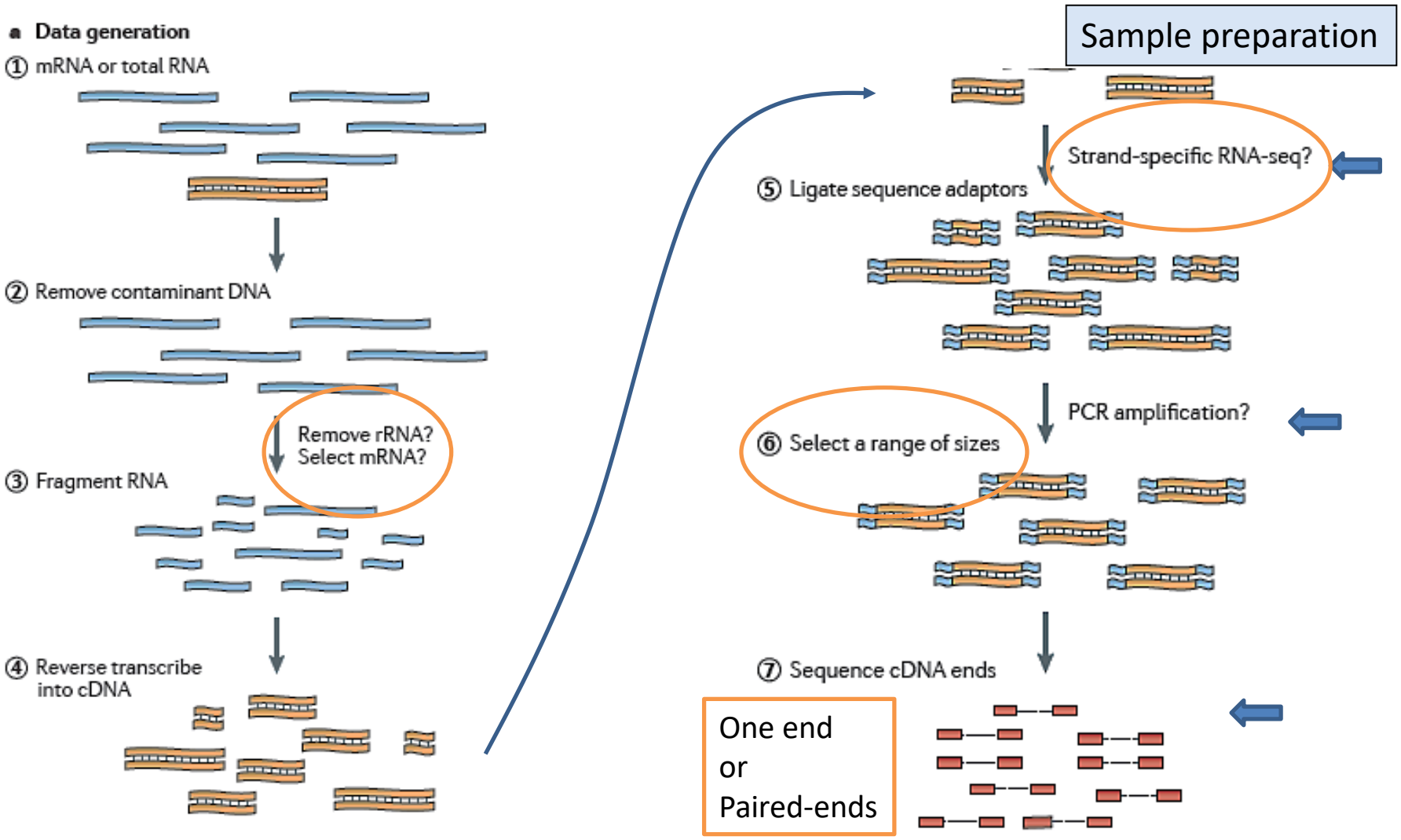


Figure 1 | **The data generation and analysis steps of a typical RNA-seq experiment. a** | Data generation. To generate an RNA sequencing (RNA-seq) data set, RNA (light blue) is first extracted (stage **1**), DNA contamination is removed using DNase (stage **2**), and the remaining RNA is broken up into short fragments (stage **3**). The RNA fragments are then reverse transcribed into cDNA (yellow, stage **4**), sequencing adaptors (blue) are ligated (stage **5**), and fragment size selection is undertaken (stage **6**). Finally, the ends of the cDNAs are sequenced using next-generation sequencing technologies to produce many short reads (red, stage **7**). If both ends of the cDNAs are sequenced, then paired-end reads are generated, as shown here by dashed lines between the pairs. rRNA, ribosomal RNA. *(Martin & Wang, 2011)*

Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study

Sheng Li[1,2,24], Scott W Tighe[3,24], Charles M Nicolet[4], Deborah Grove[5], Shawn Levy[6], William Farmerie[7], Agnes Viale[8], Chris Wright[9], Peter A Schweitzer[10], Yuan Gao[11], Dewey Kim[11], Joe Boland[12], Belynda Hicks[12], Ryan Kim[13,23], Sagar Chhangawala[1,2], Nadereh Jafari[14], Nalini Raghavachari[15], Jorge Gandara[1,2], Natàlia Garcia-Reyero[16], Cynthia Hendrickson[6], David Roberson[12], Jeffrey A Rosenfeld[17], Todd Smith[18], Jason G Underwood[19], May Wang[20], Paul Zumbo[1,2], Don A Baldwin[21], George S Grills[10] & Christopher E Mason[1,2,22]

High-throughput RNA sequencing (RNA-seq) greatly expands the potential for genomics discoveries, but the wide variety of platforms, protocols and performance capabilitites has created the need for comprehensive reference data. Here we describe the Association of Biomolecular Resource Facilities next-generation sequencing (ABRF-NGS) study on RNA-seq. We carried out replicate experiments across 15 laboratory sites using reference RNA standards to test four protocols (poly-A–selected, ribo-depleted, size-selected and degraded) on five sequencing platforms (Illumina HiSeq, Life Technologies PGM and Proton, Pacific Biosciences RS and Roche 454). The results show high intraplatform (Spearman rank $R > 0.86$) and inter-platform ($R > 0.83$) concordance for expression measures across the deep-count platforms, but highly variable efficiency and cost for splice junction and variant detection between all platforms. For intact RNA, gene expression profiles from rRNA-depletion and poly-A enrichment are similar. In addition, rRNA depletion enables effective analysis of degraded RNA samples. This study provides a broad foundation for cross-platform standardization, evaluation and improvement of RNA-seq.

Li et al., Nature Biotech., 2014

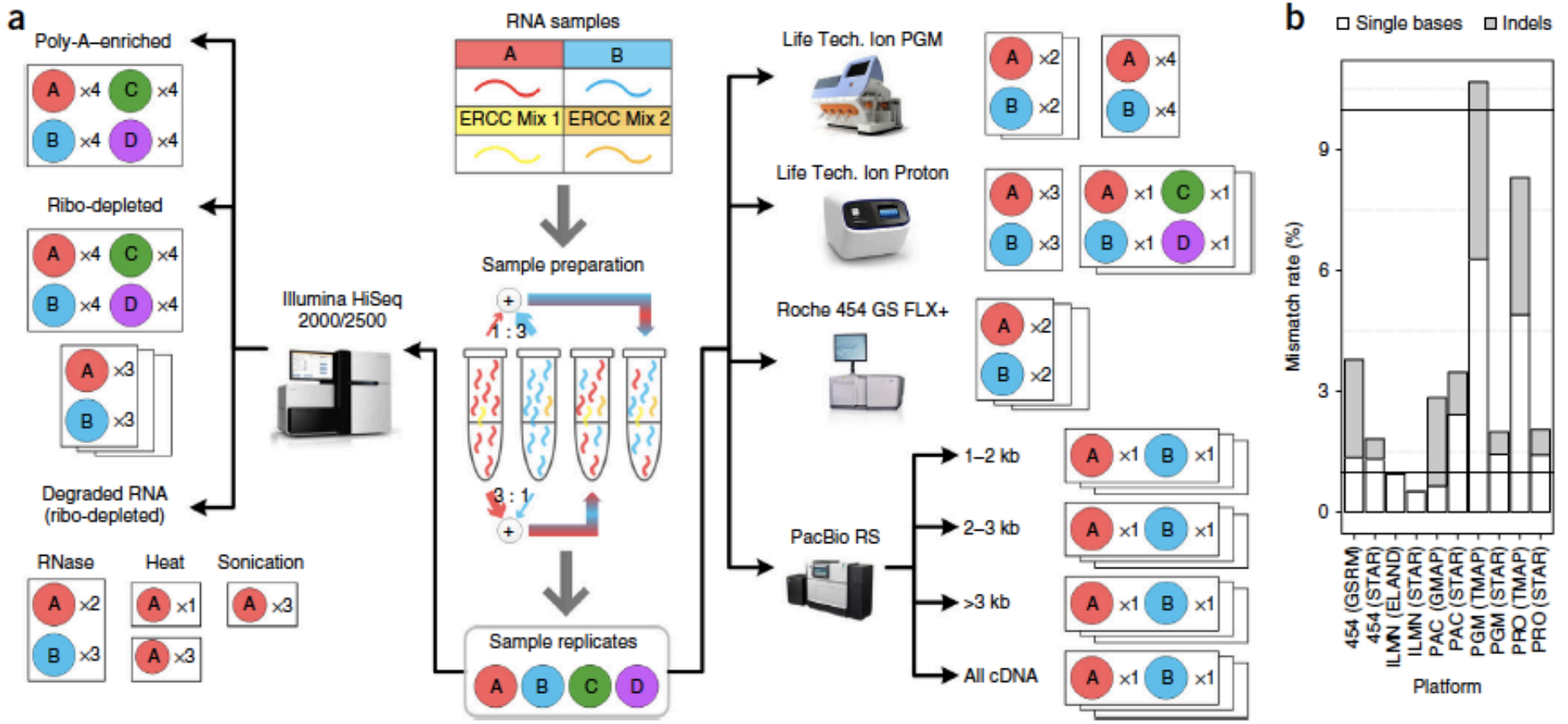
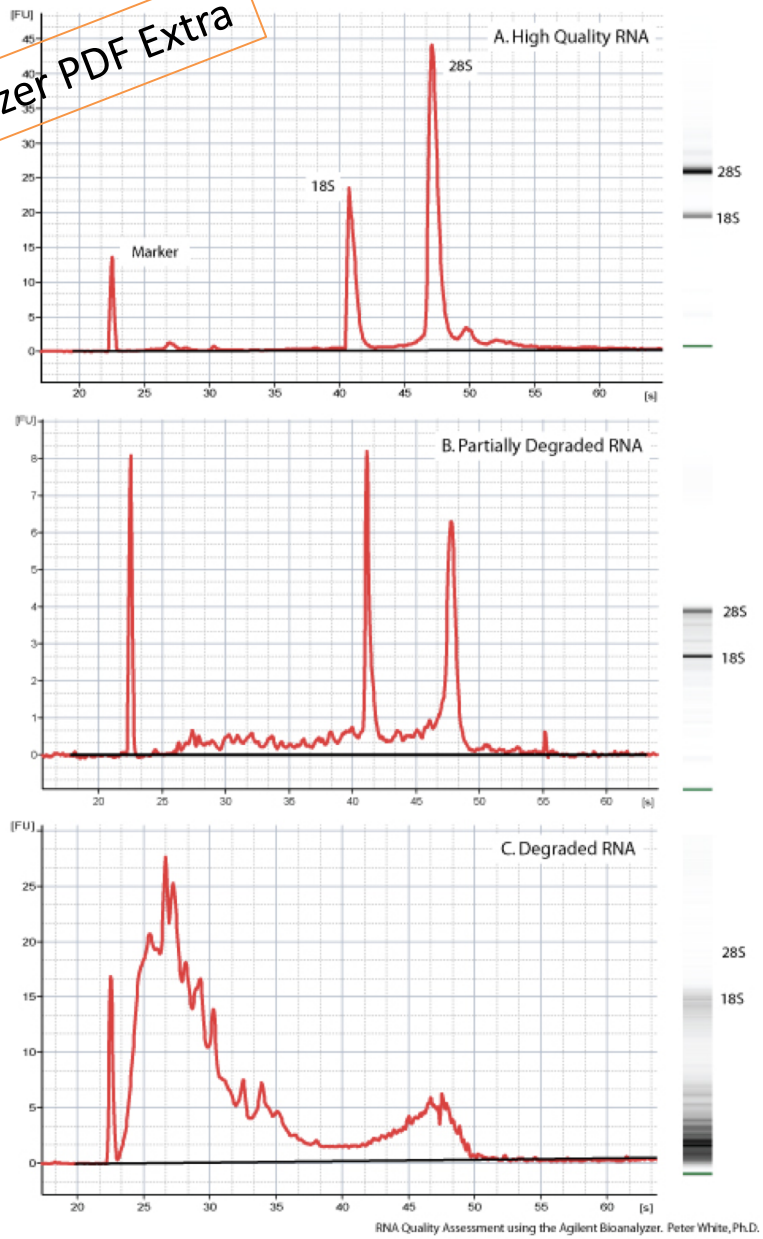


Figure 1 Experimental design and sequencing platforms. (**a**) Two standard RNA samples (A = Universal Human Reference RNA and B = Human Brain Reference RNA) were combined with two sets of synthetic RNAs (ERCCs) to prepare a set of samples to be sequenced on five platforms: Illumina (ILMN) HiSeq 2000/2500, Life Technologies Personal Genome Machine (PGM), Life Technologies Proton (PRO), Pacific Biosciences (PacBio) RS (PAC), and the Roche 454 GS FLX+. Additional RNA samples were also generated: samples C and D were prepared as defined mixtures of A and B, while other aliquots of A and B were degraded by three methods. All these additional samples were ribo-depleted for RNA-seq on the HiSeq platform. The number of technical replicates (×2, ×3 or ×4) of each sample set is indicated for each platform and method. The number of stacked rectangles indicates the number of sites performing the same experiment. (**b**) Stacked bar plots of the sequencing platforms' mismatch rates (*y* axis) for single-base mismatches (white) and insertions/deletions (indels, gray) based on different aligners for each platform (*x* axis). Q10 (90% accuracy) and Q20 (99% accuracy) are shown as the top and bottom line, respectively. *X* axis indicates the platform name, with the aligner name in parentheses.

Li et al., Nature Biotech., 2014

Bioanalyzer PDF Extra



Quality Control

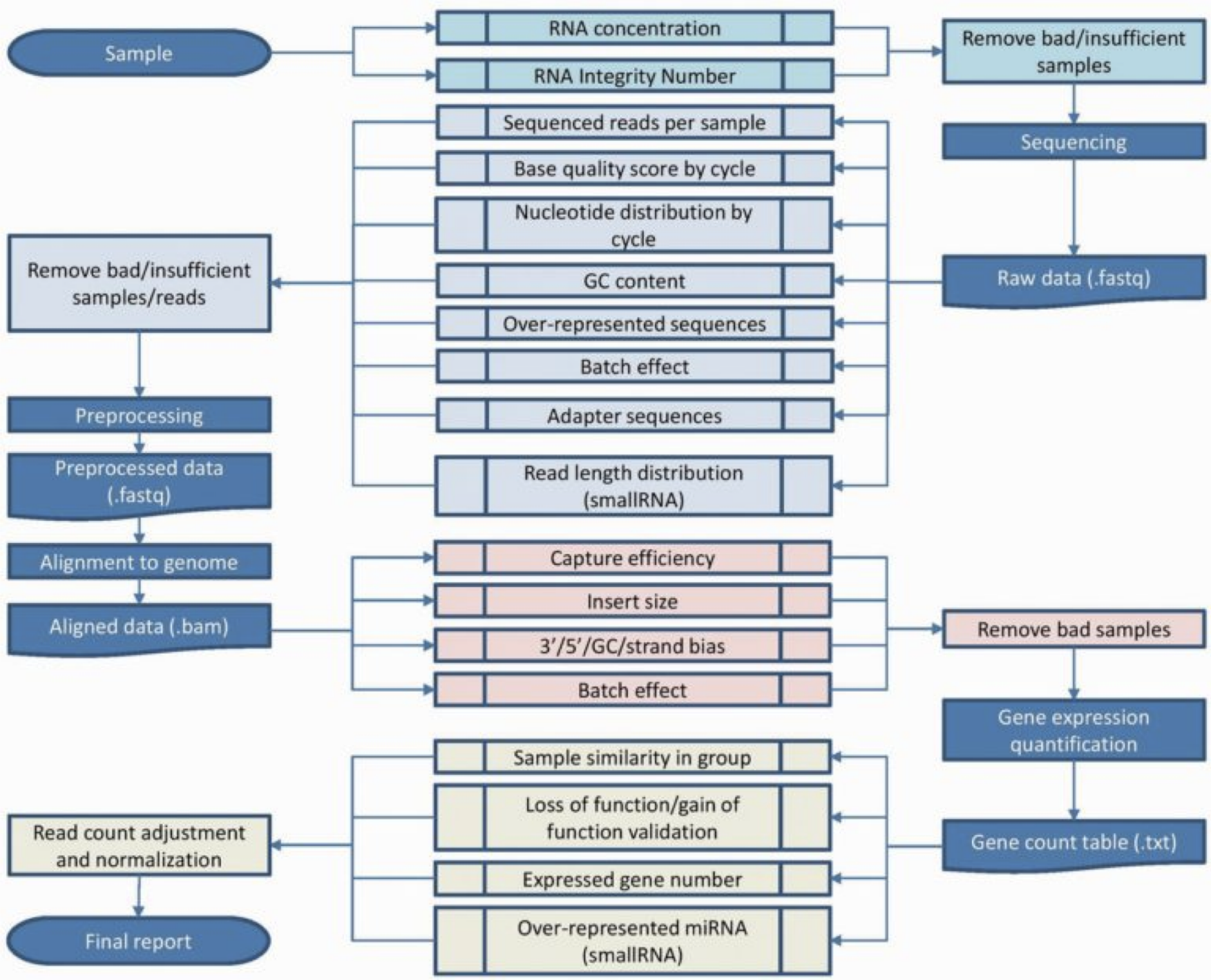
Garbage in = Garbage out !!

1- RNA quality control

2-Pre-processed raw reads

3- Aligned reads

Example of RNA quality control using Agilent Bionanalyzer; RIN: RNA Integrity Number

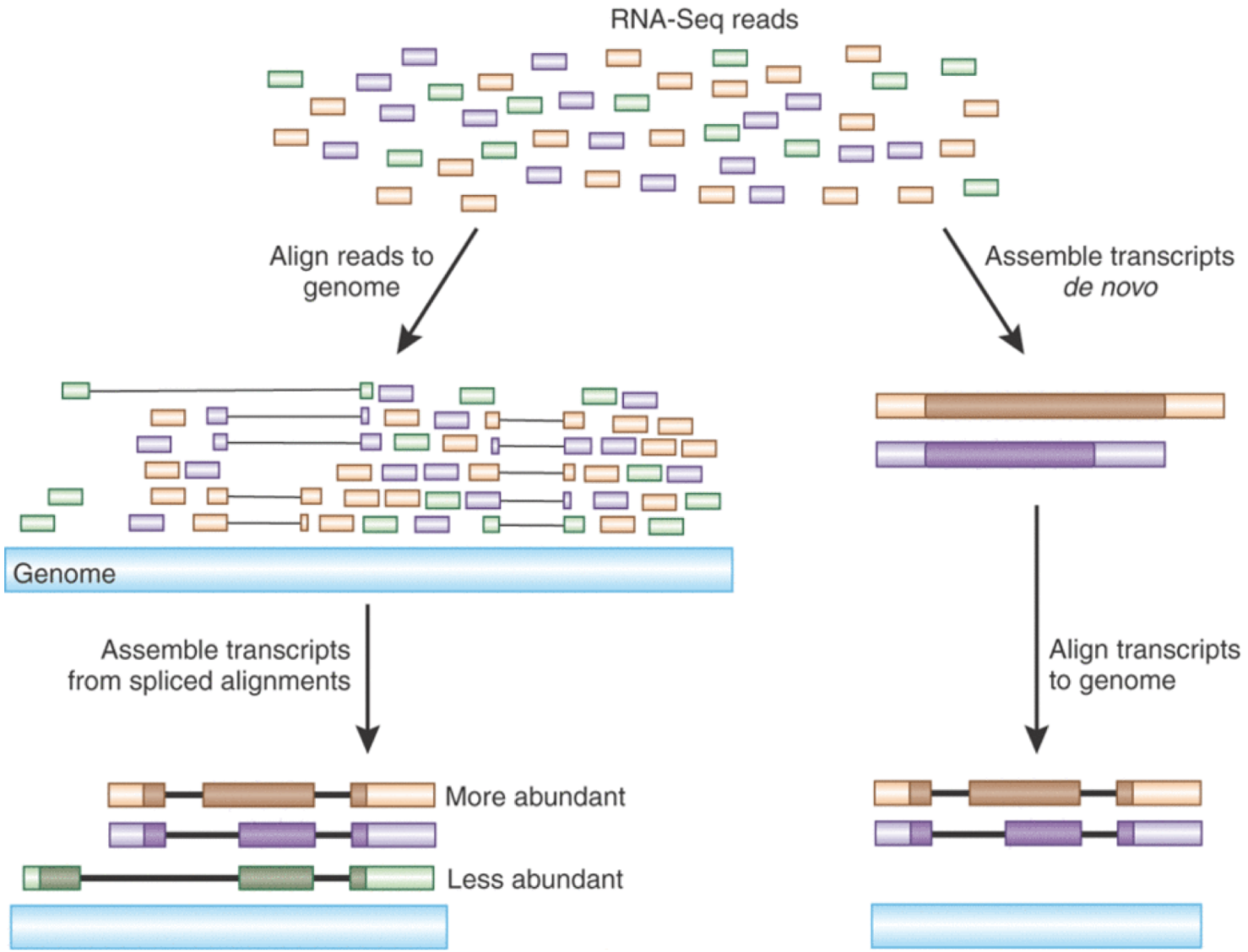

Sample
RNA concentration
RNA Integrity Number
Remove bad/insufficient samples
Sequencing
Sequenced reads per sample
Base quality score by cycle
Nucleotide distribution by cycle
GC content
Over-represented sequences
Batch effect
Adapter sequences
Read length distribution (smallRNA)
Raw data (.fastq)
Remove bad/insufficient samples/reads
Preprocessing
Preprocessed data (.fastq)
Alignment to genome
Aligned data (.bam)
Capture efficiency
Insert size
3'/5'/GC/strand bias
Batch effect
Remove bad samples
Gene expression quantification
Gene count table (.txt)
Sample similarity in group
Loss of function/gain of function validation
Expressed gene number
Over-represented miRNA (smallRNA)
Read count adjustment and normalization
Final report

BREAK then RNAseq DATA ANALYSIS

Data analysis for mRNA-seq: key steps

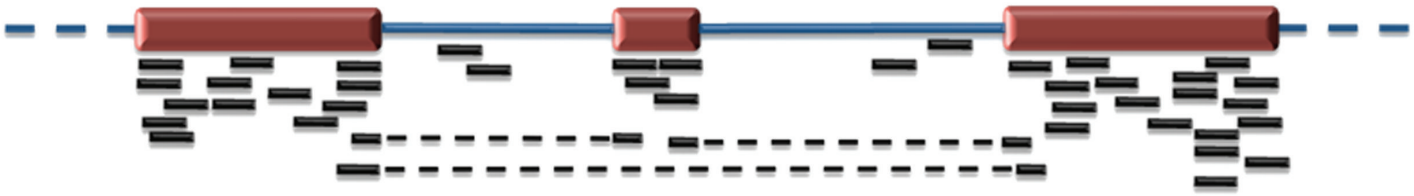
- Mapping reads to the *reference* genome
  - Read mapping of 454 sequencers can be done by conventional sequence aligners (BLAST, BLAT, etc)
  - Short read aligner needed for Illumina or SOLiD reads
- Prediction of novel transcripts
  - Assembly of short reads: comparative vs. *de novo*
- Quantifying the known genes
- Quantifying splicing variants

Reference-based versus *de novo* assembly



Haas and Zody 2 , Nature Biotechnology 28, 421–423 (2010)

Mapping



Reads alignment to the genome

- Easy(ish) for genomic sequence
- Difficult for transcripts with splice junctions

Use of specific alignment tools

(i.e. Bowtie, Tophat, MapSplice…)

Qualitative

Prediction of novel transcripts

Reads are aligned to the reference genome, or to more limited reference of your choice:

- known exons of protein-coding genes (exome)
- Spliced reads
- Genes (sense and antisense)

Comparison to reference libraries of known coding and noncoding RNAs
All nonmapped reads → may define new trascripts

Limitation: Sequencing depth
i.e. many long noncoding RNAs are expressed at very low level
→ very low number of reads....

Sequencing depth *versus* sensitivity

Always remember that the molecules you have sequenced are a «Sample» of the total possible reads from your biological sample.

How representative this sample is will depend on the number of molecules you have sequenced (i.e. the sequencing depth).

Saturation is reached when an increment in the number of reads does not result in additional transcripts being detected or in more differentially expressed gene being identified when two or more conditions are compared.

Increasing sequencing depth (higher coverage) helps identifying new transcripts

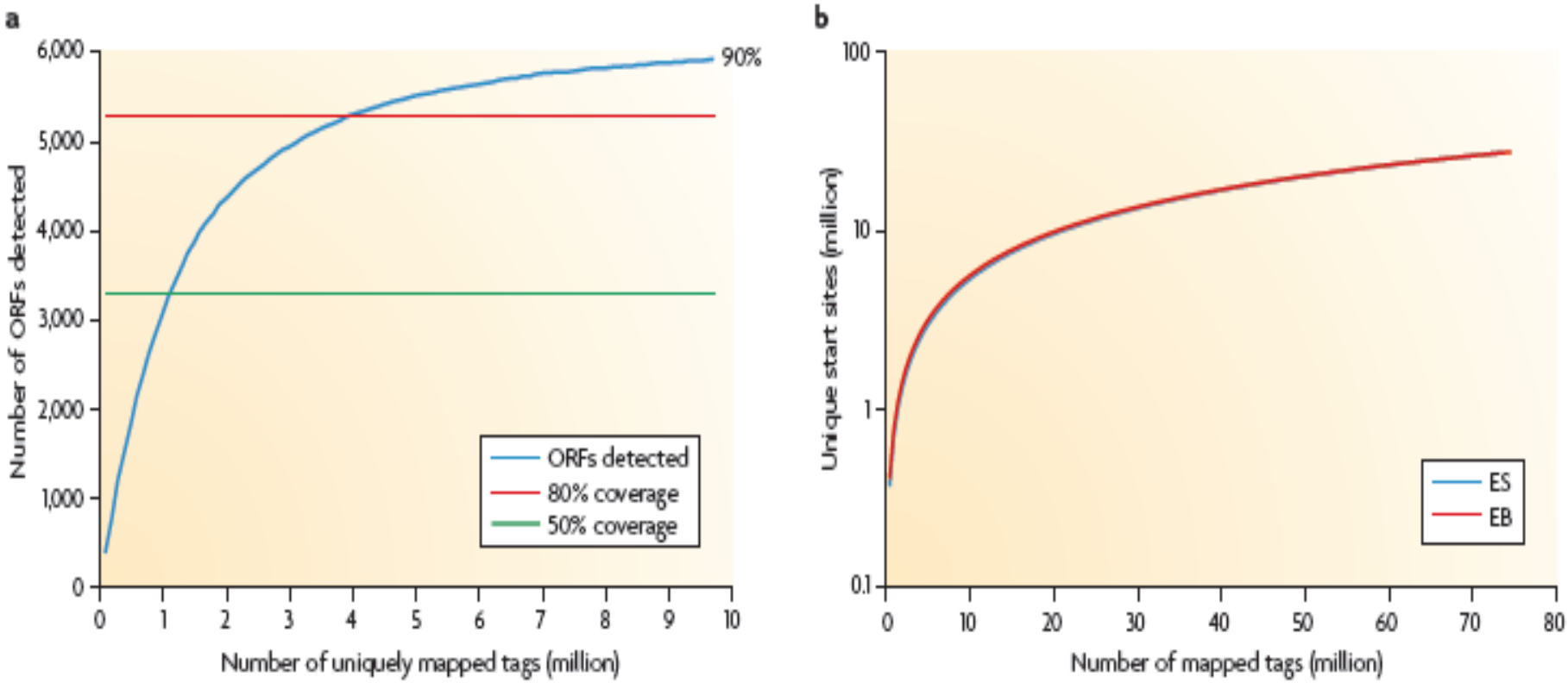
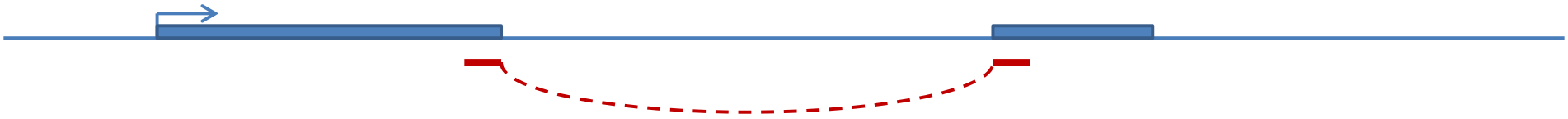


Figure 5 | **Coverage versus depth.** a | 80% of yeast genes were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3′ end. Data is taken from REF. 18.

b | The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes. ES, embryonic stem cells; EB, embryonic body. Figure is modified, with permission, from REF. 22 © (2008) Macmillan Publishers Ltd. All rights reserved.

Building alternative transcript models

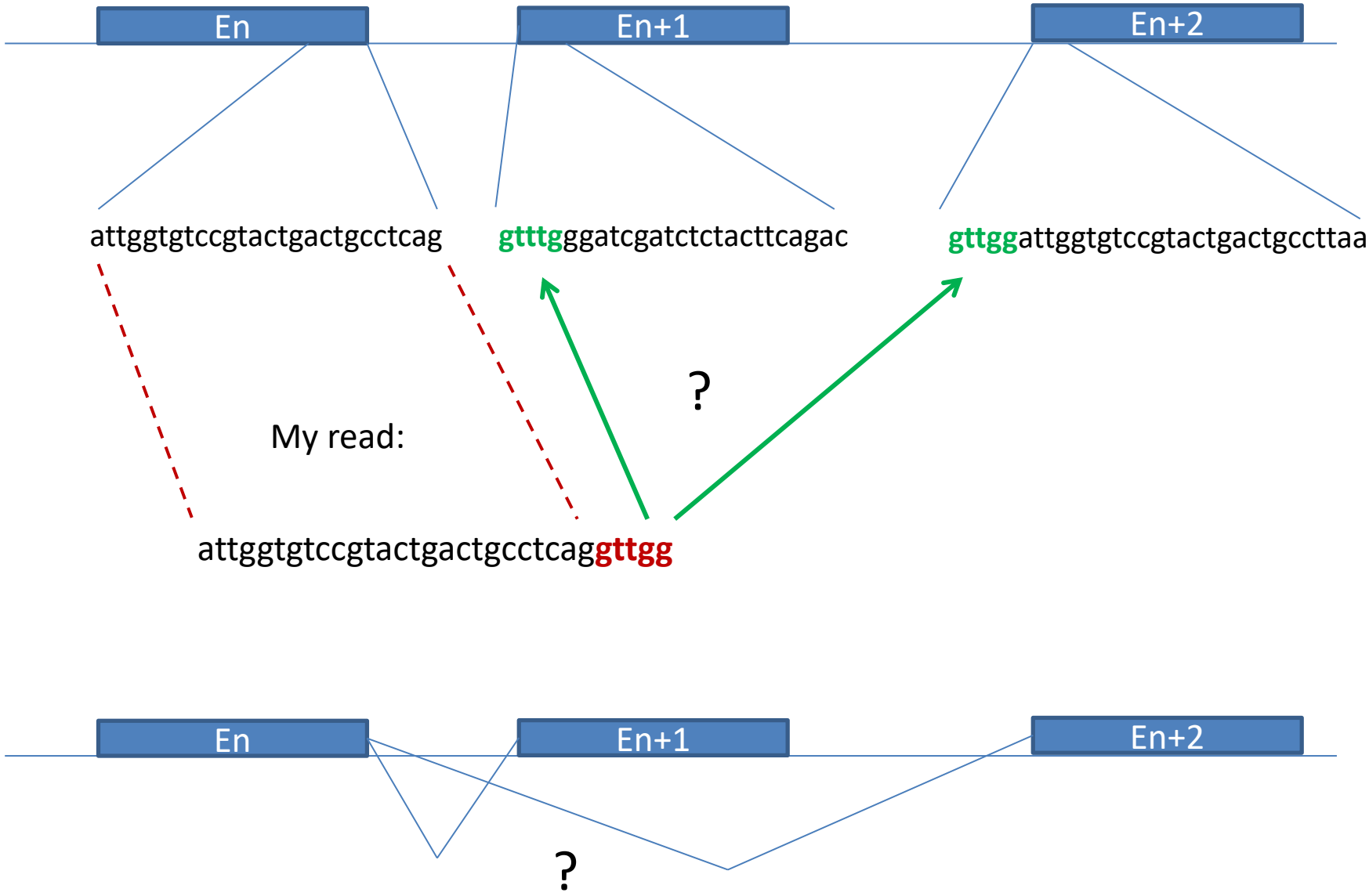
Problem: How can we deal with splicing ?



Junction read

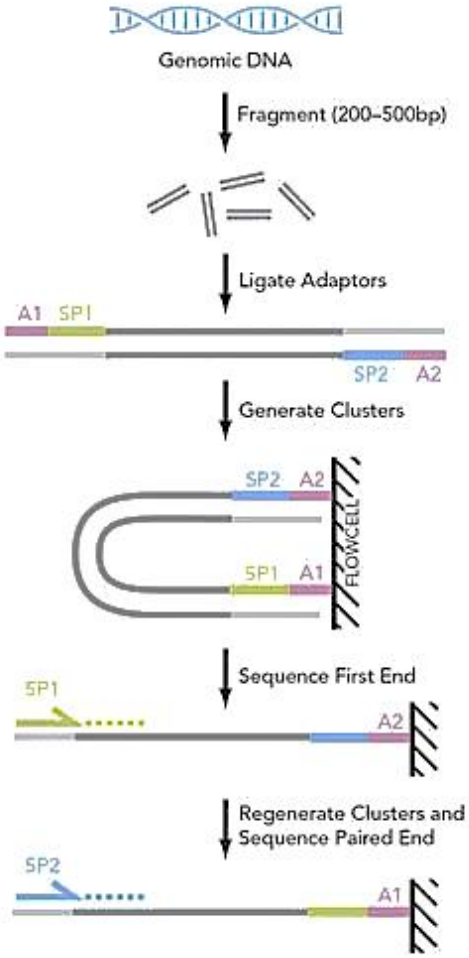
Mechanims of mRNA Splicing

Uncertainty in mapping



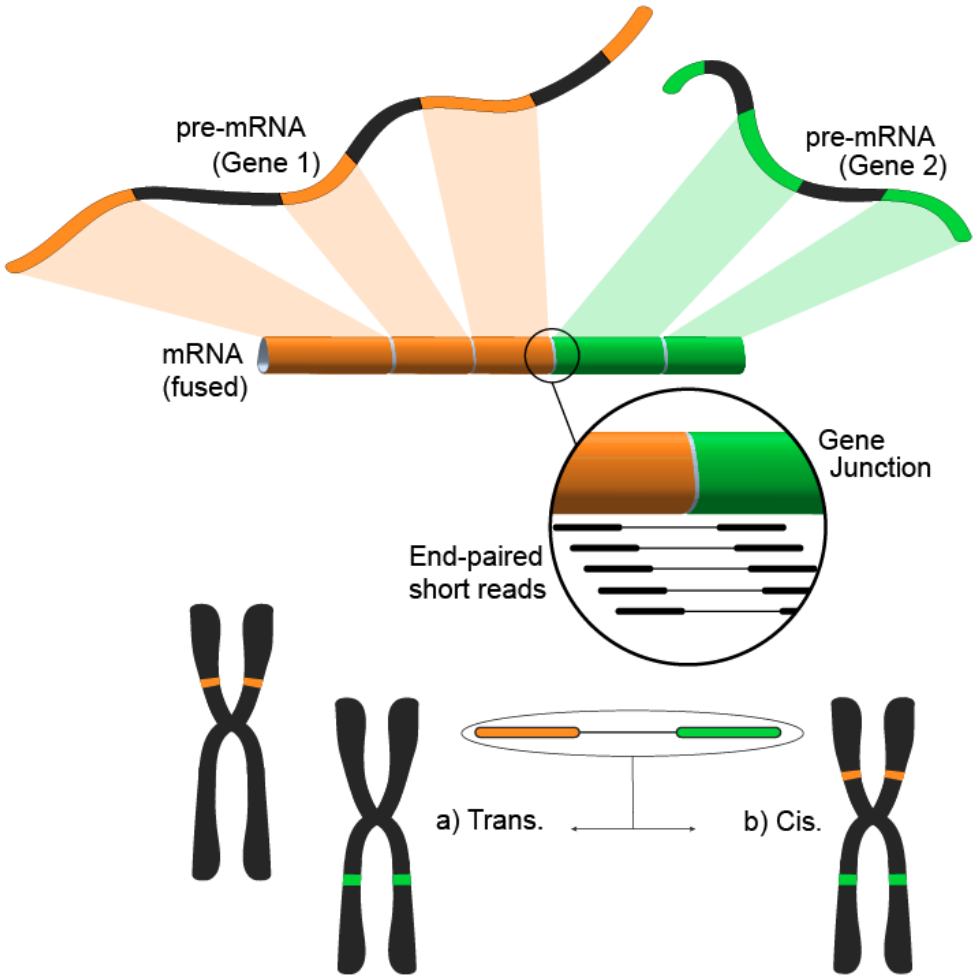
In addition to increasing coverage, difficulties in mapping reads can be helped by technical improvements:

- Longer reads
- Paired-end sequencing
- Strand-specific RNA-seq



Paired-end sequencing

Other annotations from mRNA-seq data: gene fusion events



Following the alignment of the short m-RNA reads to a reference genome, most reads will fall within a single exon, and a smaller but still large set would be expected to map to known exon-exon junctions. The remaining unmapped short reads can then be further analyzed to determine whether they match an exon-exon junction where the exons come from different genes.

Acknowledgement: Wiki – mRNA-seq

Quantitative

Once transcript database defined, a common method to evaluate expression levels is to count the reads that fall within a gene. Since the process of fragmenting and sequencing is stochastic, when the number of reads mapping to one gene is above a treshold, we can assume that the number of reads falling within the gene is proportional to the amount of that specific RNA present in the sample.

- Count number of reads for each transcript in all experimental conditions (samples)
- Normalize
- Statistics → find DE transcripts (differentially expressed genes)

Can you use absolute reads number? What rpkm means?

Quantitative

rpkm = reads per kilobase per million

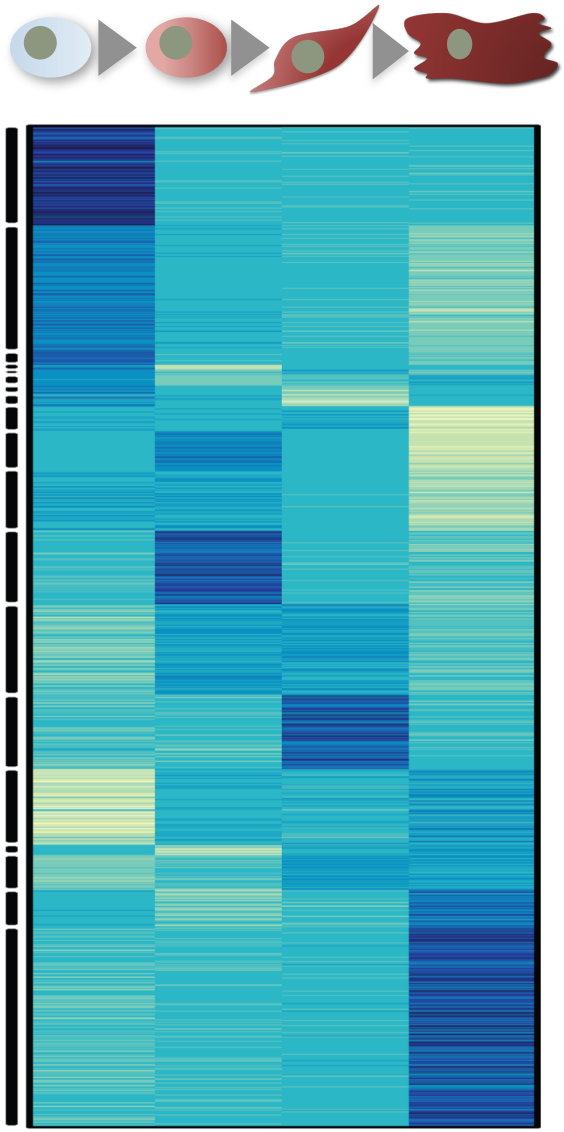
Double normalization for sequencing depth and gene lenght:

1- Divide the read counts by the “per million” scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)

2- Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

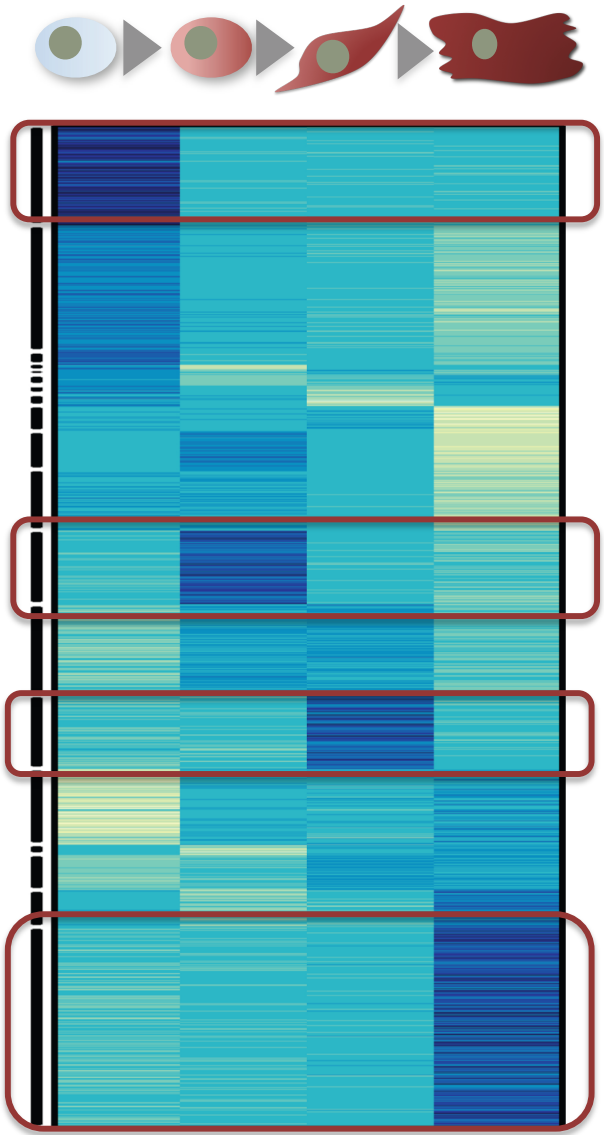
fpkm = fragments per kilobase per million
similar concept adapted for paired-end sequencing where two reads can map to one fragment

Clusters of co-expressed genes



- Use unsupervised clustering to group genes by expression pattern
- Use gene ontology information to determine which kinds of genes are in each group
- Reveal novel associations and gene types

Clusters of co-expressed genes



Pluripotency/stem cell: Nanog, Oct4

Mesoderm/cell fate commitment: Mesp1, Eomes

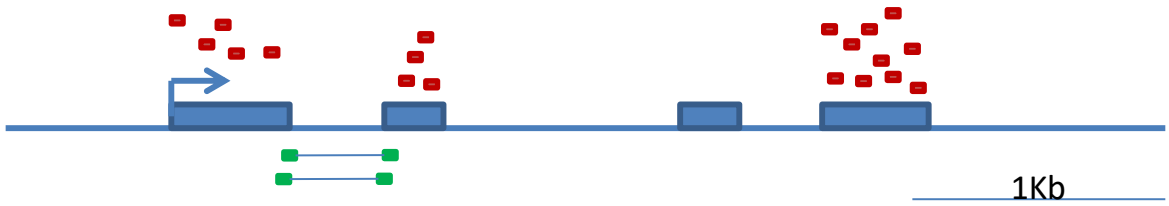
Cardiac precursors: Isl1, Mef2c, Wnt2

Cardiac structure/function: Actc1, Ryr2, Tnni3

Quantitative

Caution: it may be more appropriate to talk about «Transcript» levels rather than RNA or gene levels .

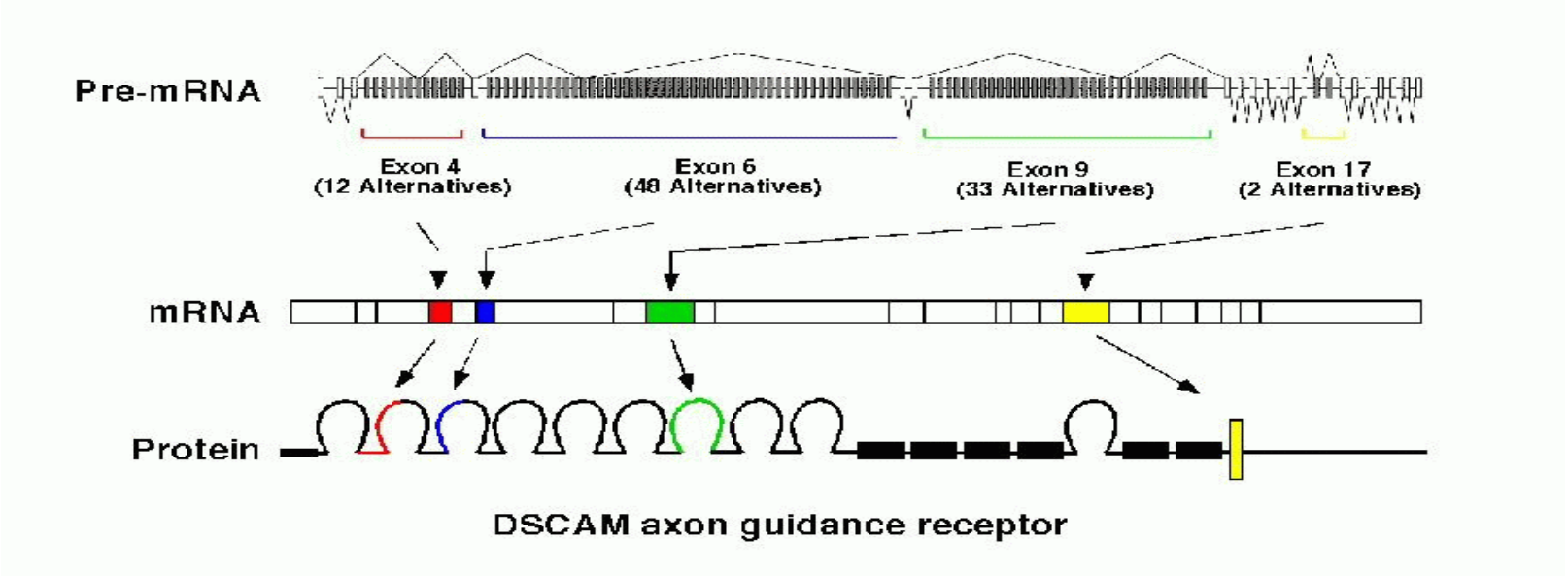
In this example, it is much likely that a splicing isoform exists that incorporates exons 1-2-4 (skipping Exon 3).



Quantification of alternative transcript usage

Quantitative

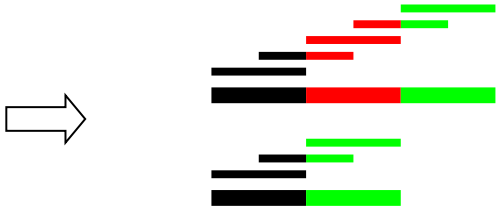
Most human genes show extensive AS and some genes present a huge number of isoforms (*slo* >500, *neurexin* >1000, *DSCAM* > 38000)



Taken from **[Graveley, 2001]**

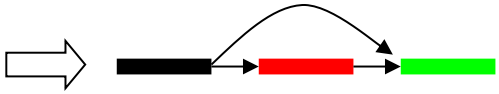
Splicing Graph Approach

Replace the problem of finding a list of consensus sequences



with ***Graph Reconstruction Problem***:

Given an set of expressed sequence, find a minimal graph (*splicing graph*) representing **all** transcripts as paths.



Heber, et. al. ISMB 2002

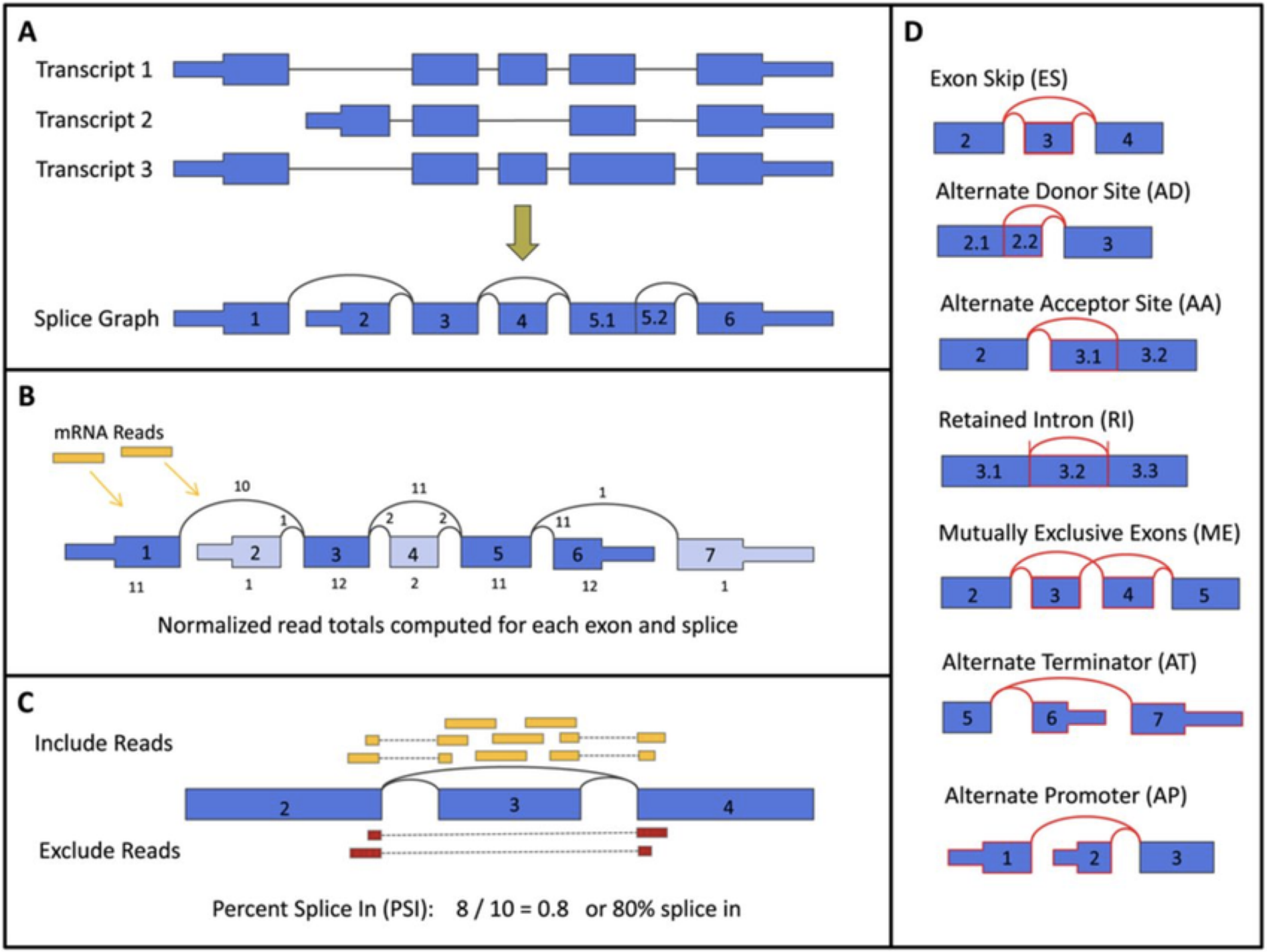
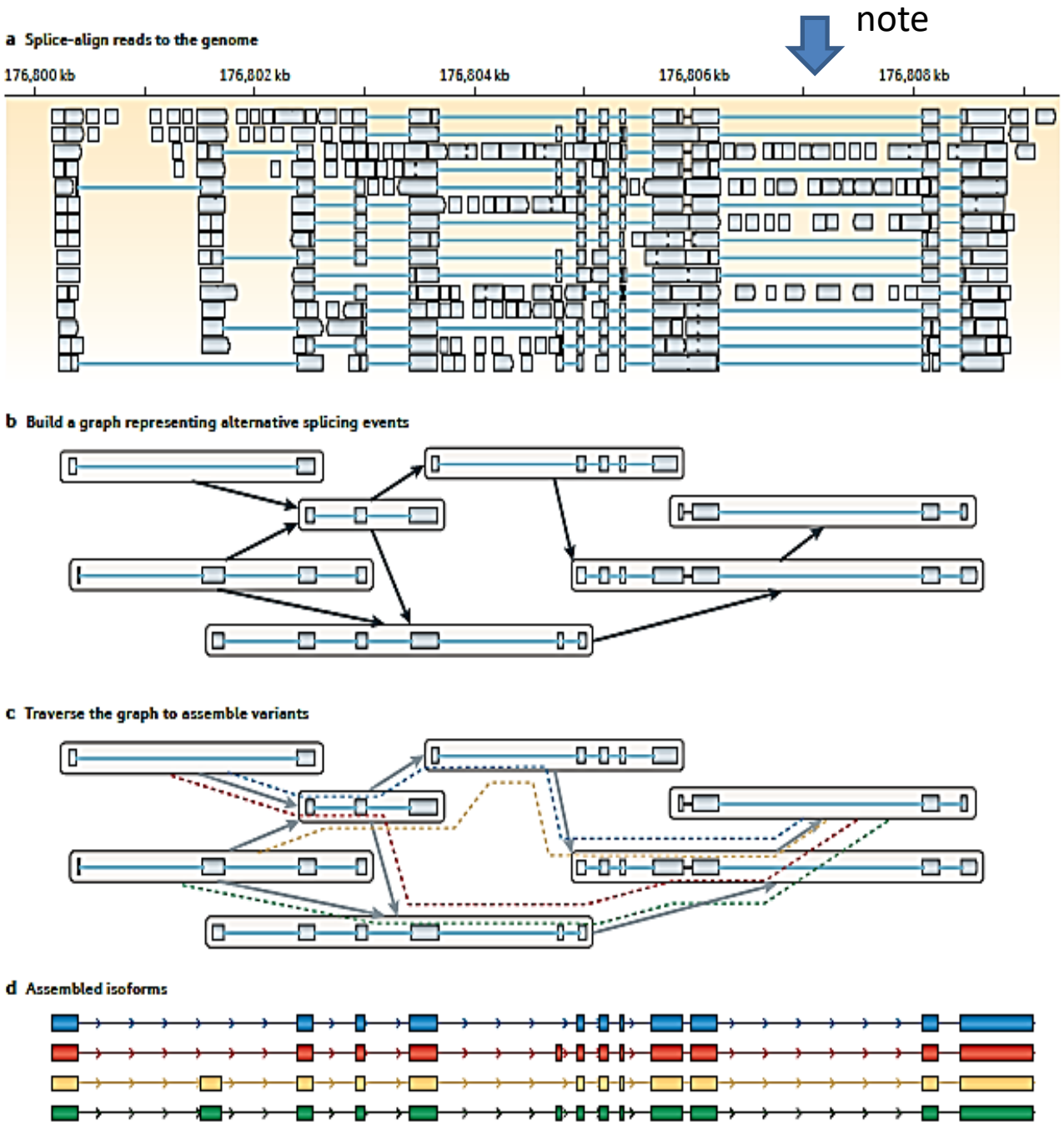

A
Transcript 1
Transcript 2
Transcript 3
Splice Graph
1
2
3
4
5.1
5.2
6
B
mRNA Reads
10
11
1
1
2
2
11
1
2
3
4
5
6
7
11
1
12
2
11
12
1
Normalized read totals computed for each exon and splice
C
Include Reads
Exclude Reads
2
3
4
Percent Splice In (PSI): 8 / 10 = 0.8 or 80% splice in
D
Exon Skip (ES)
2
3
4
Alternate Donor Site (AD)
2.1
2.2
3
Alternate Acceptor Site (AA)
2
3.1
3.2
Retained Intron (RI)
3.1
3.2
3.3
Mutually Exclusive Exons (ME)
2
3
4
5
Alternate Terminator (AT)
5
6
7
Alternate Promoter (AP)
1
2
3

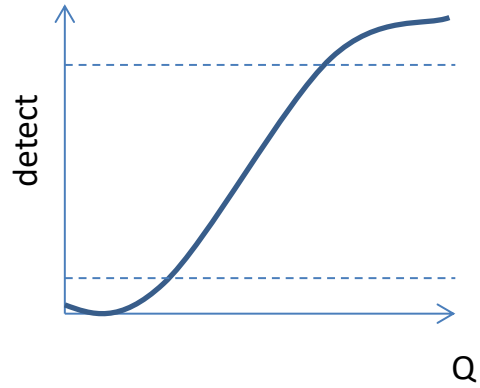
Figure 2 | **Overview of the reference-based transcriptome assembly strategy.** The steps of the reference-based transcriptome strategy are shown using an example of a maize gene (GRMZM2G060216).
a | Reads (grey) are first splice-aligned to a reference genome.
b | A connectivity or splice graph is then constructed to represent all possible isoforms at a locus.
c,**d** | Finally, alternative paths through the graph (blue, red, yellow and green) are followed to join compatible reads together into isoforms.

(Martin & Wang, 2011)



Quantitative

Comparing microarrays to RNA-Seq for quantitative purposes:

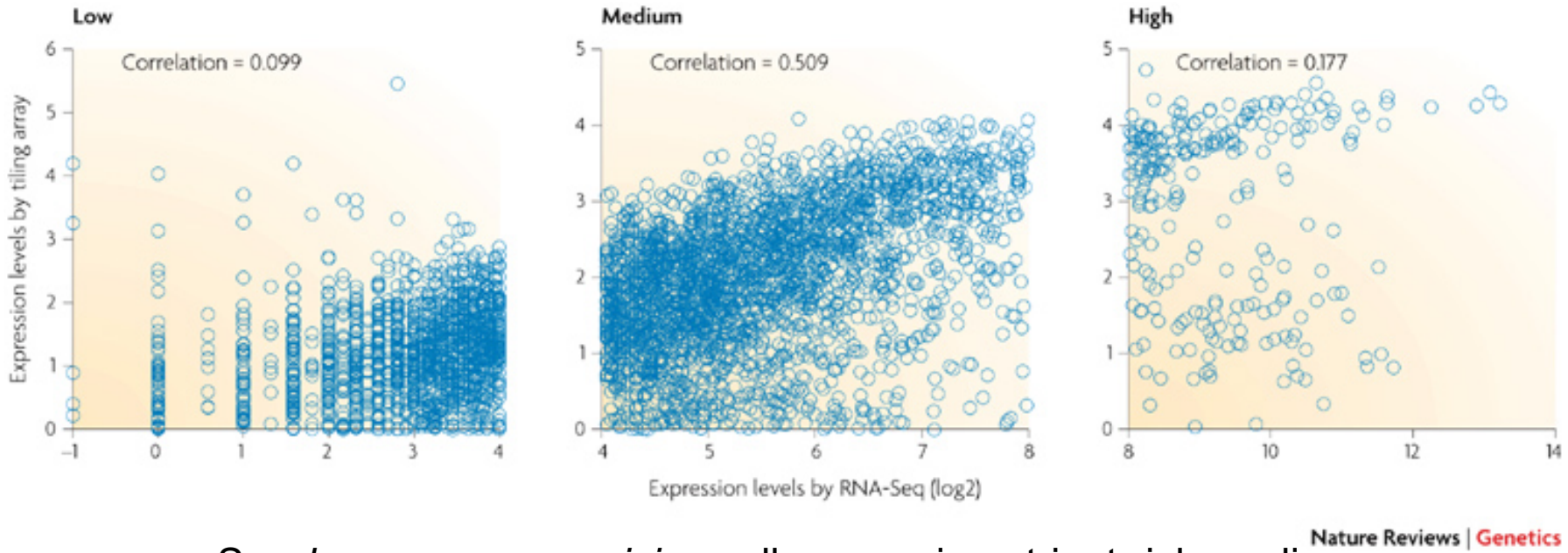


- Microarrays have narrow dynamic range

- RNA-Seq: no upper limit, lower limit depends on sequencing depth

- Often results difficult to compare since RNA-Seq refers to all transcripts, whereas microarray refers to the probed segment only.

RNA-seq and microarray agree fairly well only for genes with medium levels of expression



Saccharomyces cerevisiae cells grown in nutrient-rich media. Correlation is very low for genes with either low or high expression levels.

RNA-seq	Microarray
ID novel genes, transcripts, & exons	Well vetted QC and analysis methods
Greater dynamic range	Well characterized biases
Less bias due to genetic variation	Quick turnaround from established core facilities
Repeatable	Currently less expensive
No species-specific primer/probe design	
More accurate relative to qPCR	
Many more applications	

Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	**Tiling microarray**	**cDNA or EST sequencing**	**RNA-Seq**
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Tomorrow RNASEQ PAPERs

Actual Sequencing Platforms

- Roche/454 (GS FLX+/GS Junior)
- Illumina Genome Analyzer (HiSeq/MiSeq/NextSeq)
- Life Technologies (3500 Genetic Analyzer, Ion Torrent Proton/PGM)
- Pacific Biosciences (PACBIO RSII)
- Applied Biosystems (SOLiD, 3730*xl* DNA Analyzer)

The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"

Sequencing Principles

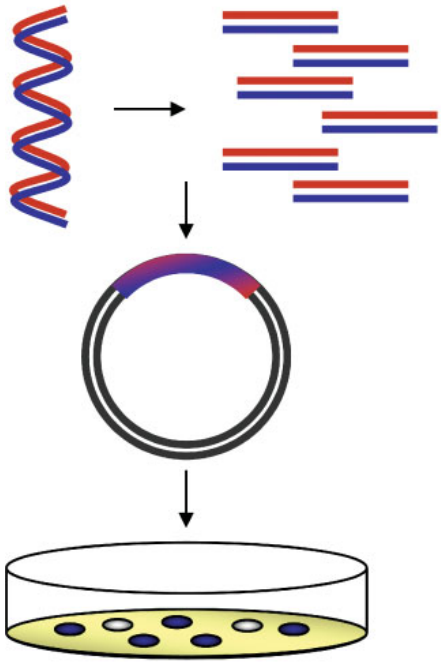
- Sequencing by Synthesis
  - Sanger/Dideoxy chain termination (Life Technologies, Applied Biosystems)
  - Pyrosequencing (Roche/454)
  - Reversible terminator (Illumina)
  - Ion proton semiconductor (Life Technologies)
  - Zero Mode Waveguide (Pacific Biosciences)
- Sequencing by Oligo Ligation Detection
  - SOLiD (Applied Biosystems)
- Other
  - Asynchronous virtual terminator chemistry - HeliScope (Helios)

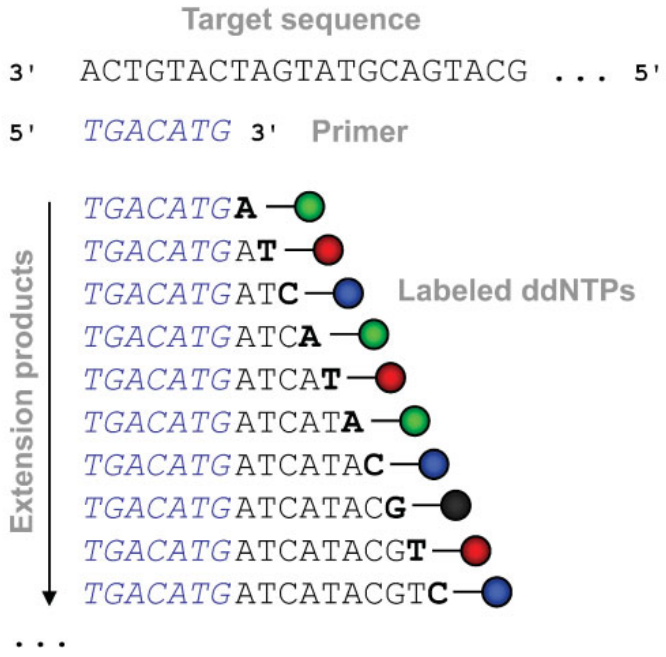
The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"

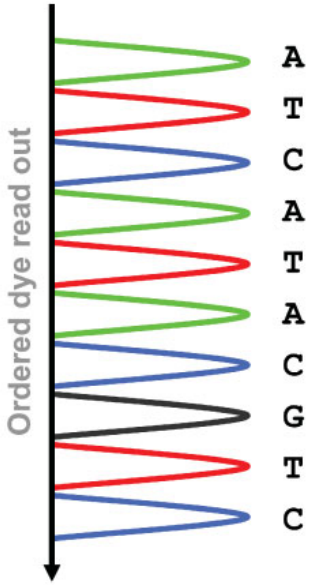
Sanger (3500 GA, 3730*xl* DNA Analyzer)

Sequencing by synthesis

1. Amplification *(by cloning)* → 2. Primer extension in presence of blocked and labeled nucleotides → 3. Separation by electrophoreses & read out of labels



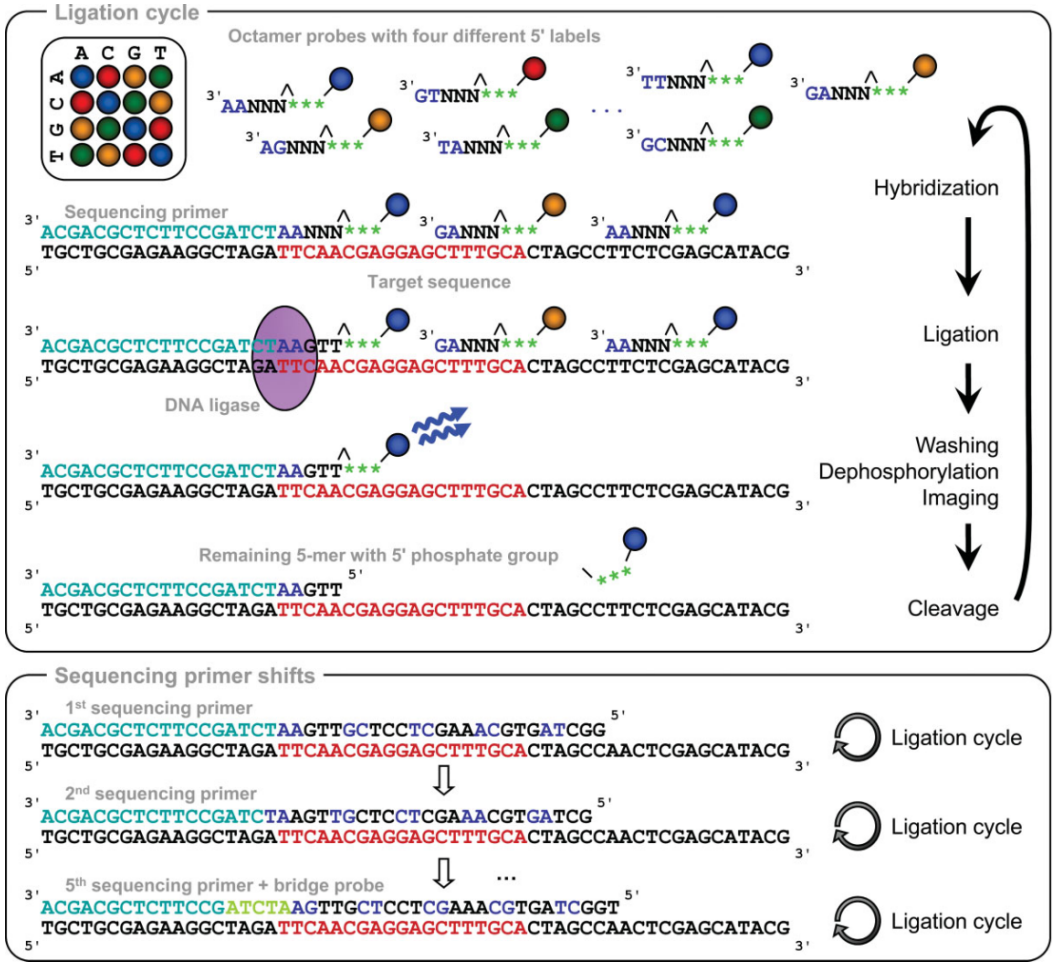




The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"

Oligo Ligation Detection (SOLiD)

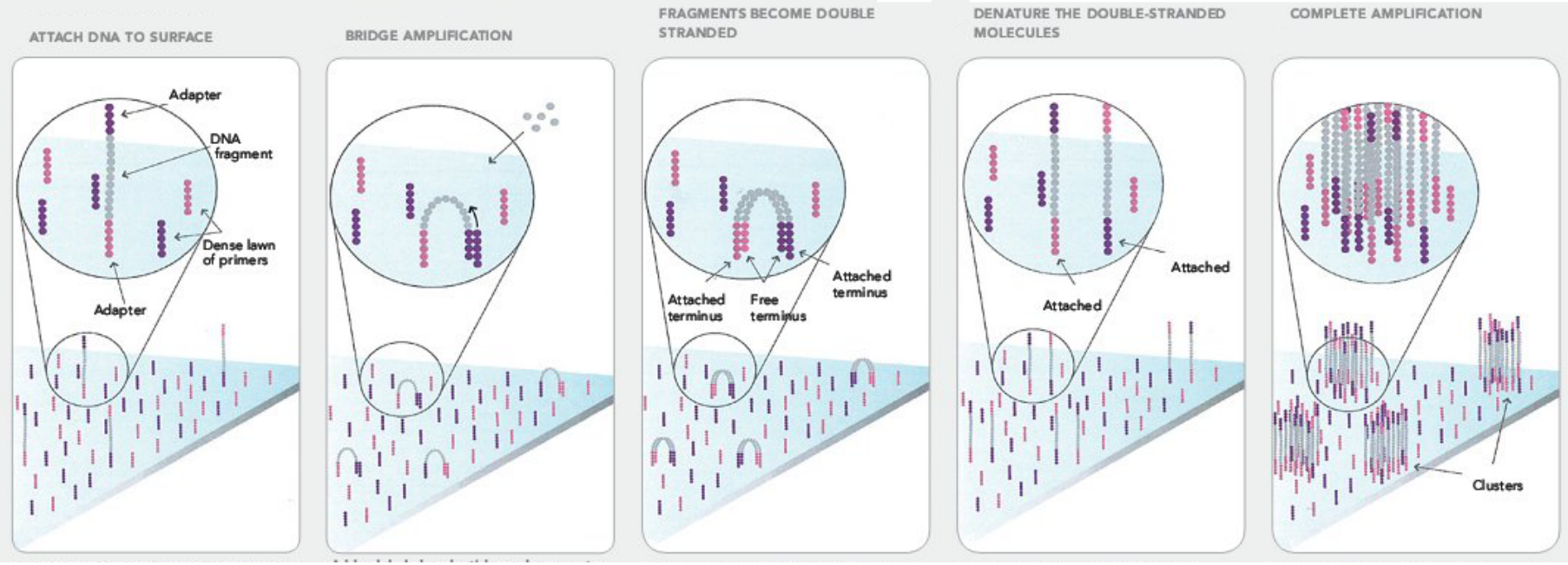
Sequencing by ligation



The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"

Reversible Terminator (HiSeq, MiSeq, NextSeq)

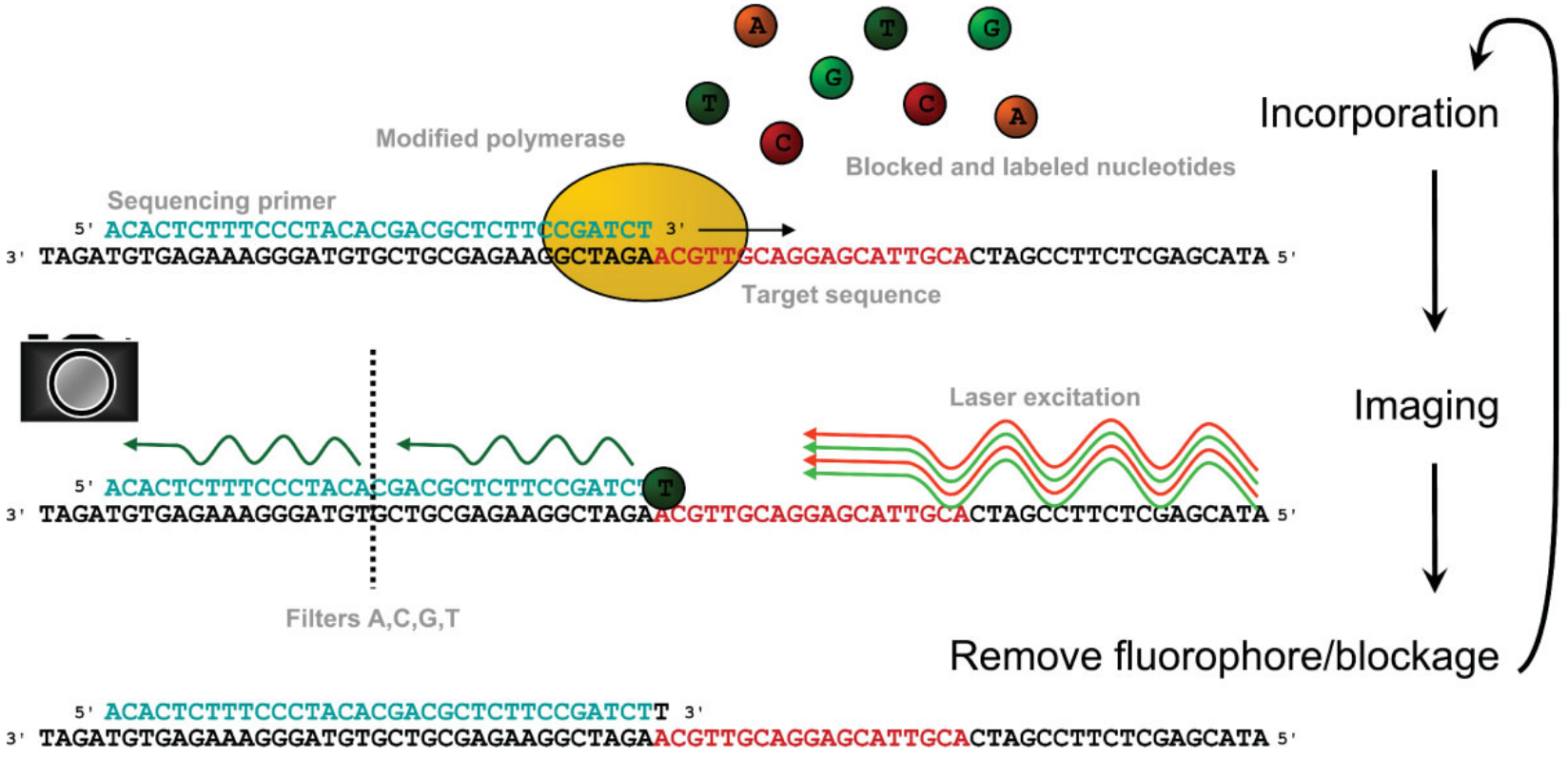
Cluster generation on a flow-cell surface



The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"

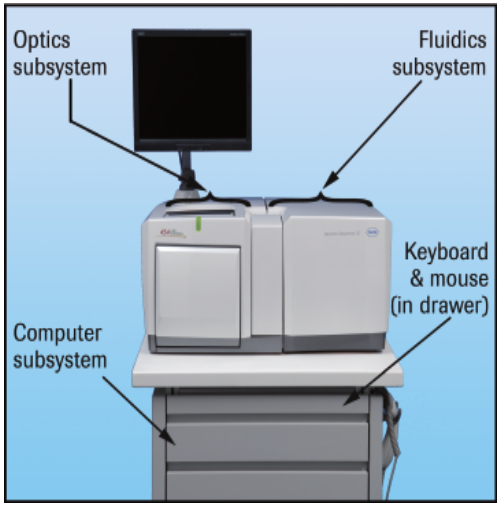
Reversible Terminator (HiSeq, MiSeq, NextSeq)

Sequencing by synthesis

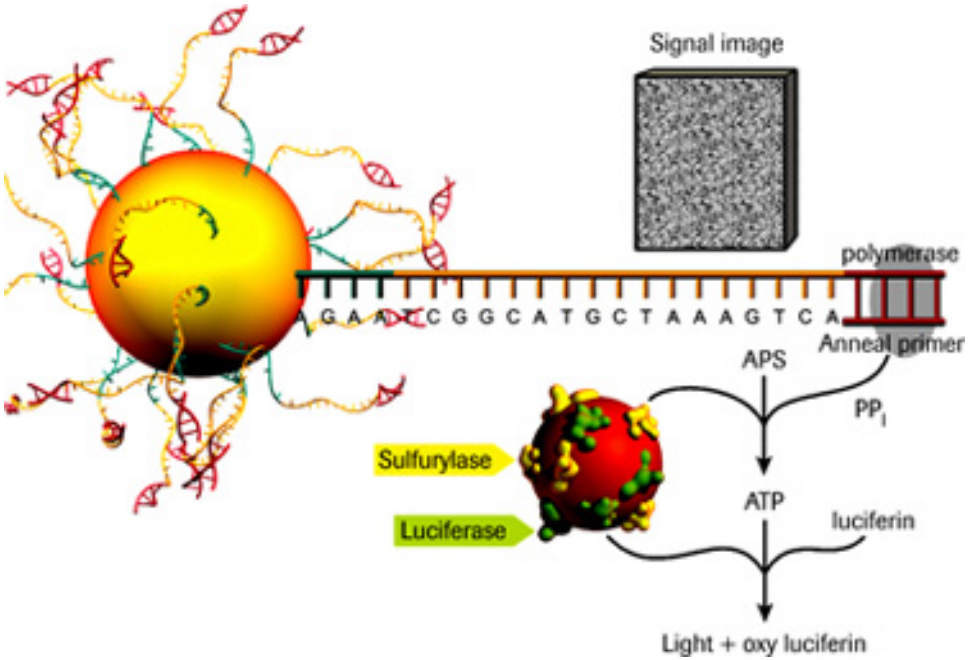


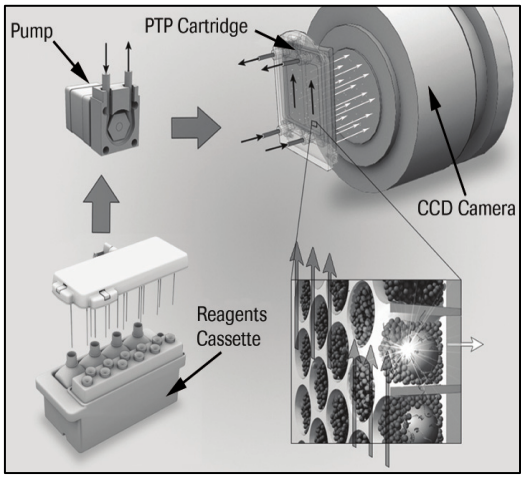
The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"

Pyrosequencing (GS FLX, GS Junior)



Sequencing by synthesis

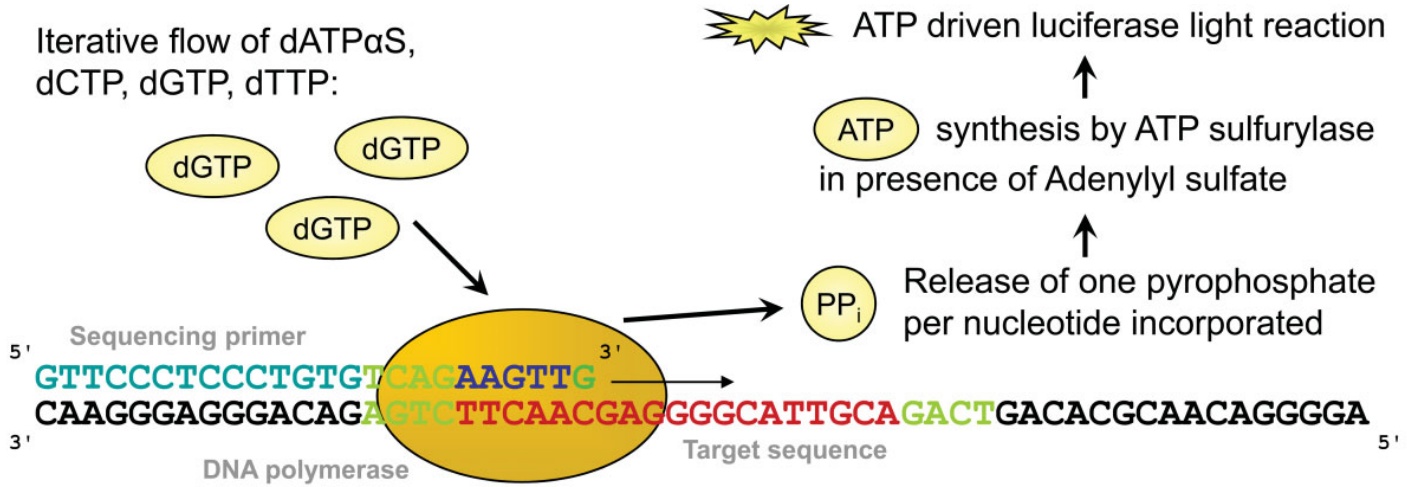


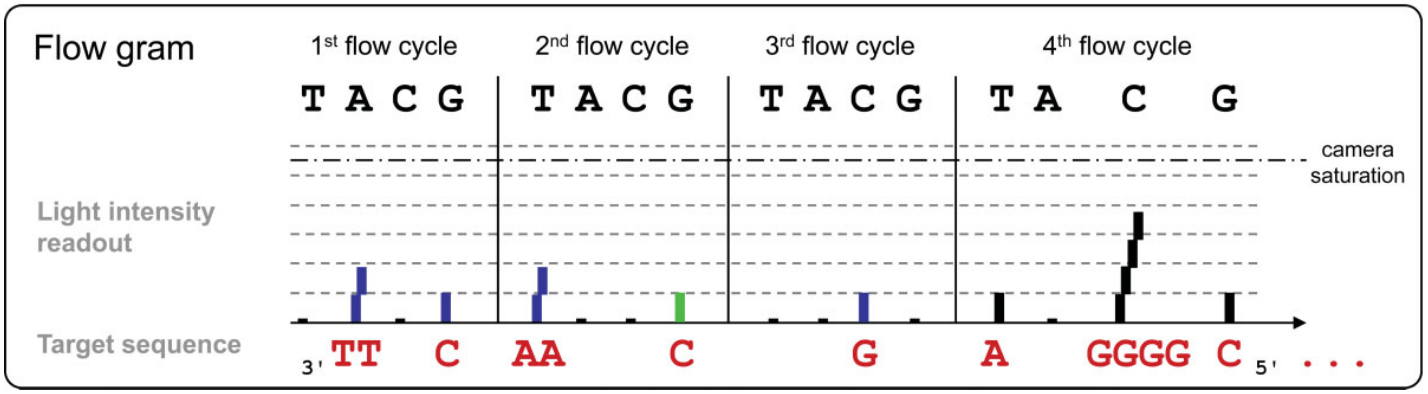


The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"

Pyrosequencing (GS FLX, GS Junior)

Sequencing by synthesis





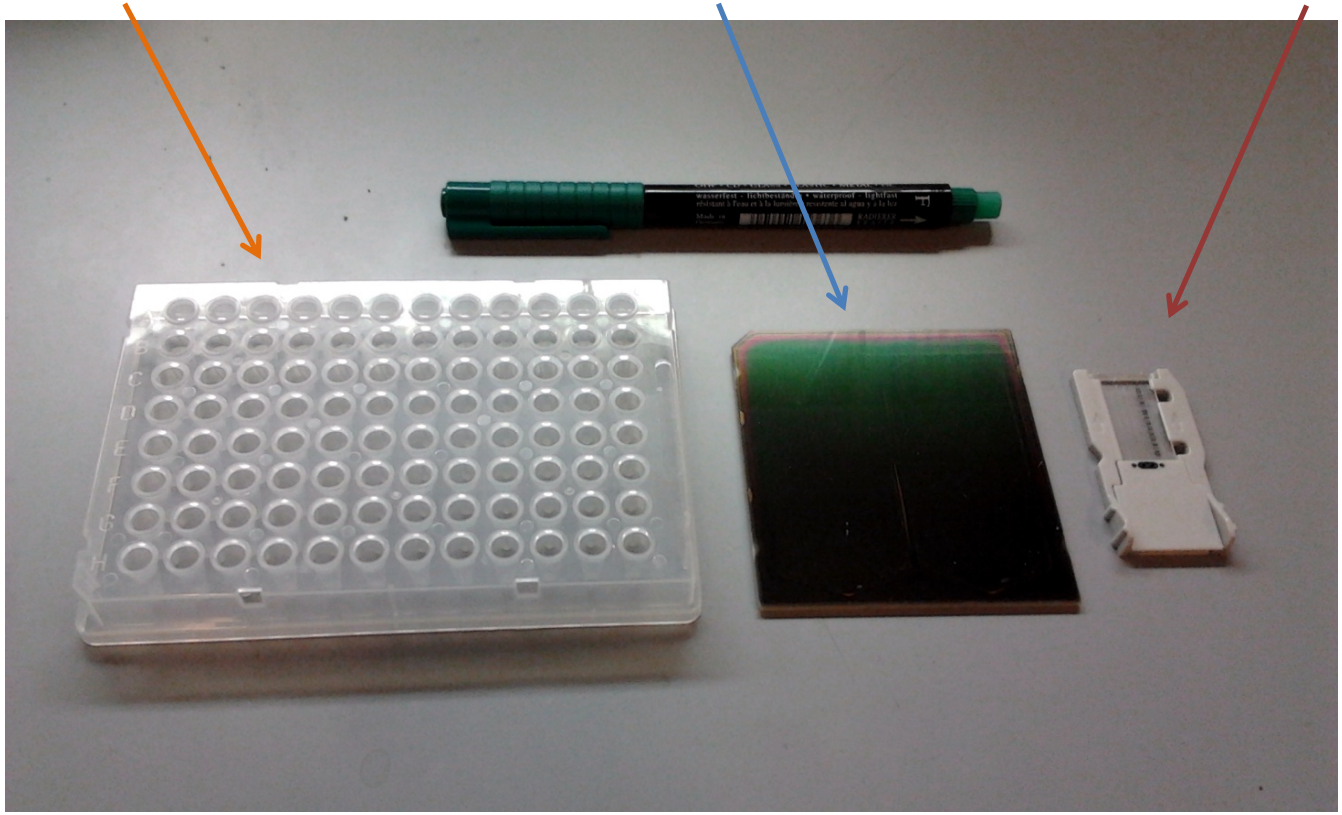
The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"

Sequencing Matrices

Sanger, 96-well, 8 capillaries
96 x 600 bp / 24 h
1400 €

Pyrosequencing, 2 regions
1,000,000 x 600 bp / 20 h
5500 €

Revers. terminator, MiSeq
10,000,000 x 250 bp / 40 h
1150 €



The presentation is supported from the project OP EC CZ.1.07/2.3.00/30.0027
"Founding the Centre of Transgenic Technologies"