Transcriptomes

All the parts of a genome that are represented in RNA

Biased:

✓expression microarrays (different kinds)
✓exon microarrays
✓splicing juntion microarrays

Unbiased:

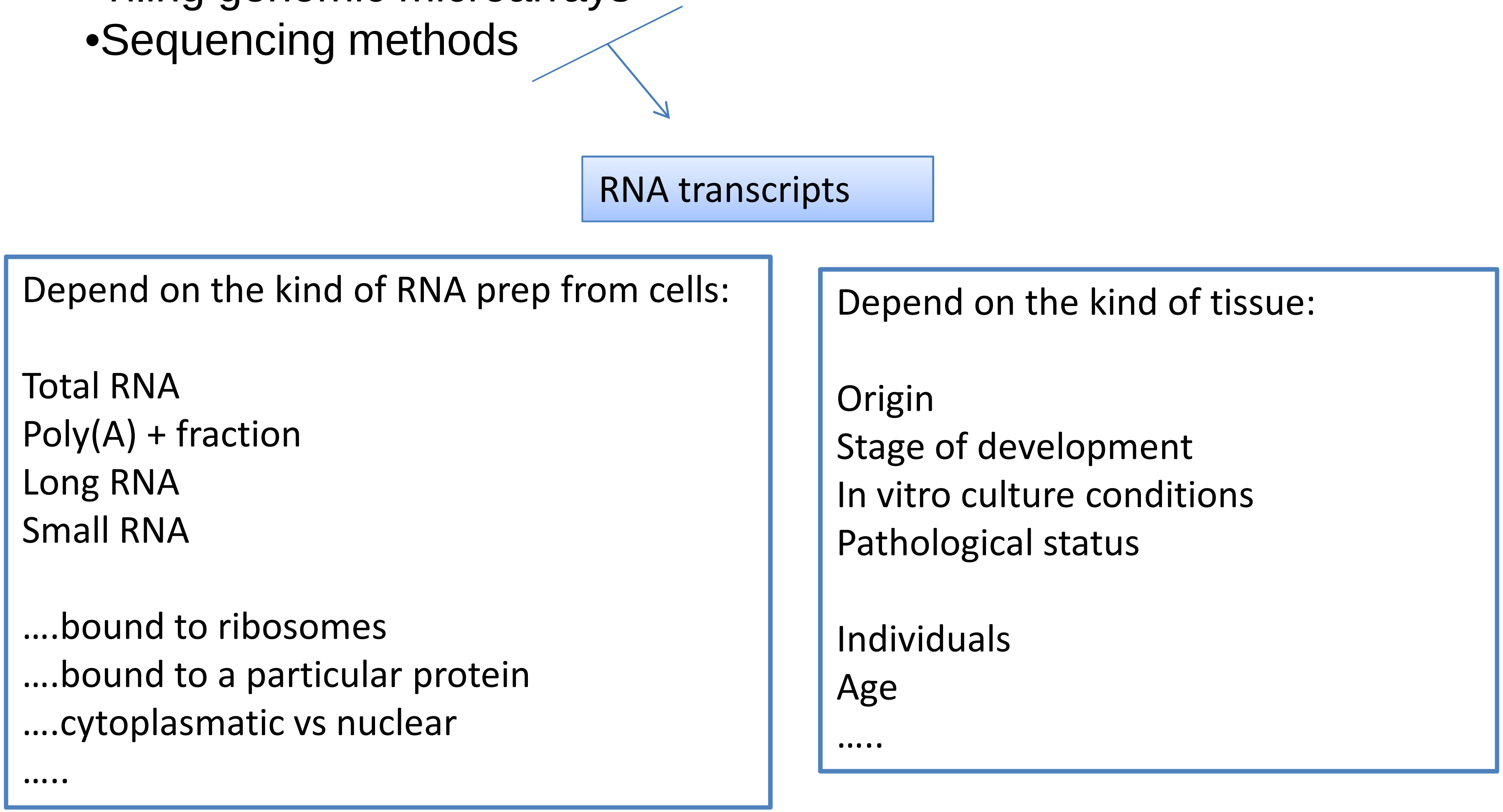
✓tiling microarrays

✓sequencing approaches

- EST
- SAGE
- CAGE
- RNA-Seq

What a transcriptome analysis tell us

- Expression Microarrays
- Tiling genomic microarrays
- Sequencing methods



RNA transcripts

Depend on the kind of RNA prep from cells:

Total RNA
Poly(A) + fraction
Long RNA
Small RNA

….bound to ribosomes
….bound to a particular protein
….cytoplasmatic vs nuclear
…..

Depend on the kind of tissue:

Origin
Stage of development
In vitro culture conditions
Pathological status

Individuals
Age
…..

Tiling arrays

Strong limitations for tiling arrays:

Main problem: unspecific hybridization that can not be controlled (theoretically feasible, not practically)

minor problem: 5’-end and 3’-end indentified not at the nucleotide resolution (unless overlapping probes are used)

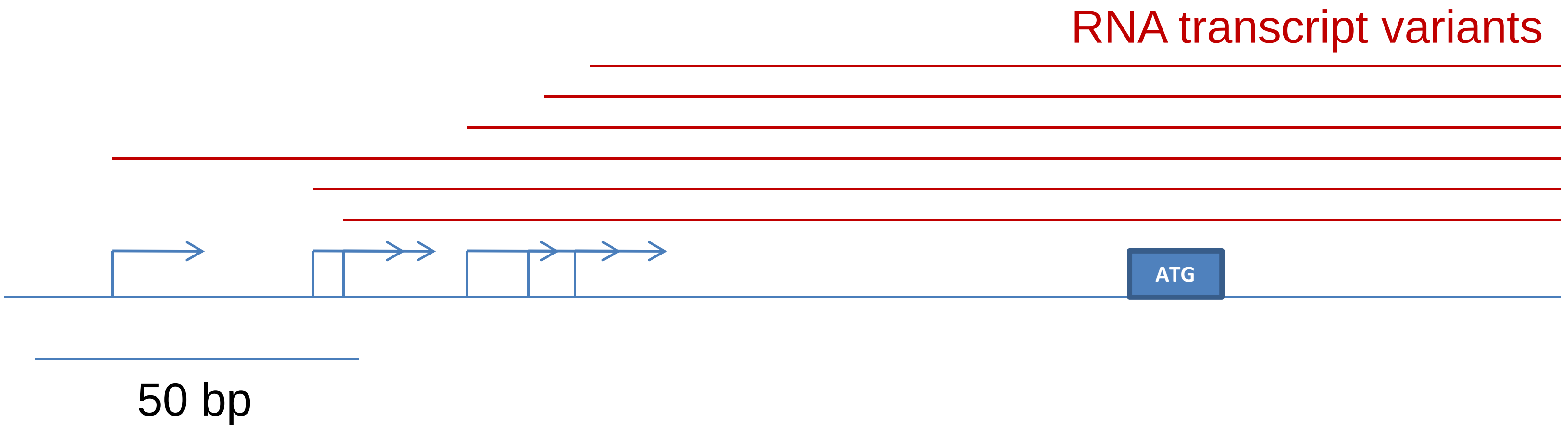
CAGE analysis (5’-end SAGE-like)

CAGE analysis was very important, since it identifies genuine Transcription Start Sites (TSS)

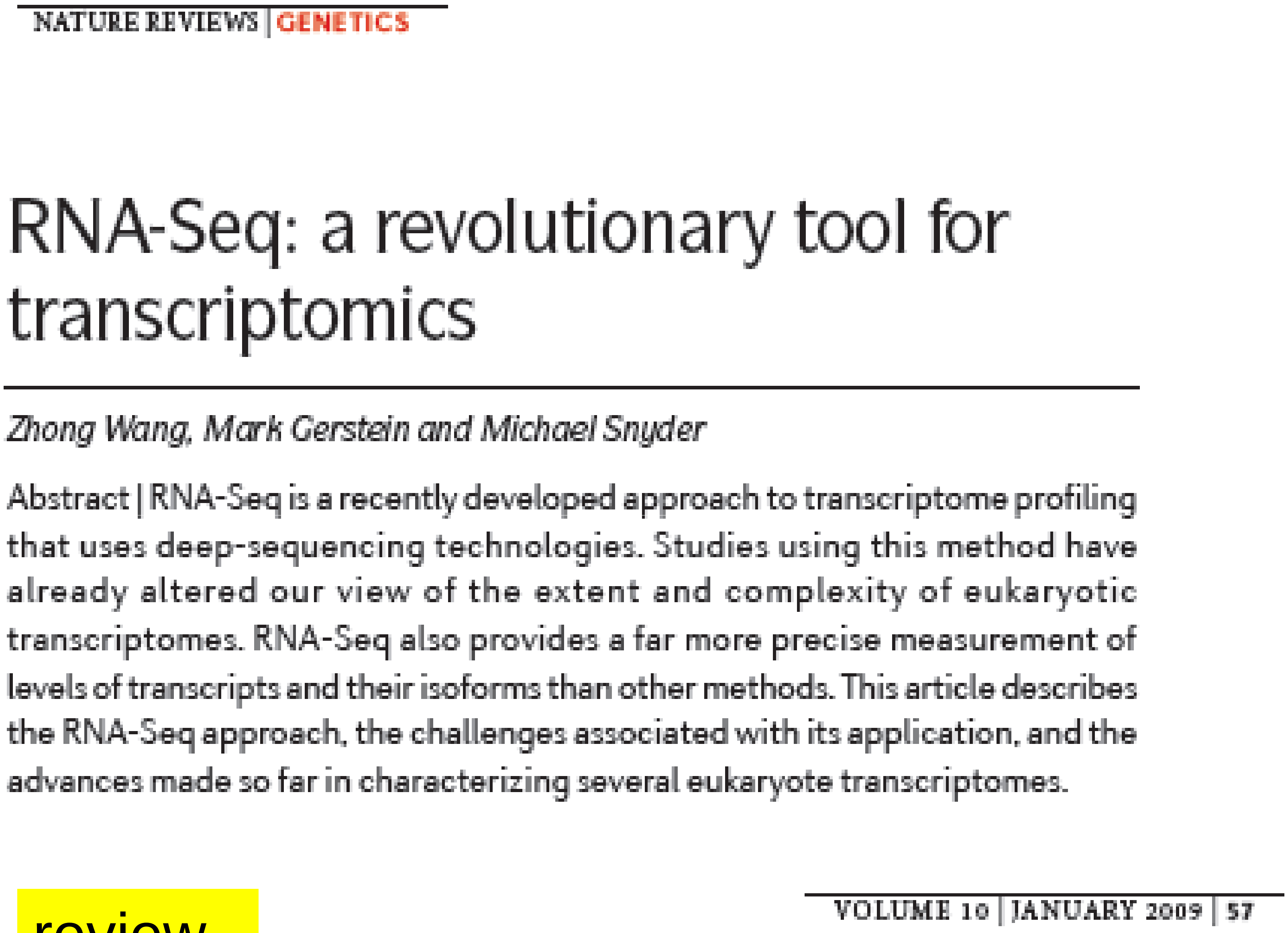
The number of CAGE tags identified was much greater than the number of mapped genes.

Why ?

- New unidentified genes
- Noncoding RNAs (small and large, see next lectures)
- Antisense transcripts within genes
- Multiple TSS in the same gene, with the same CDS



3’ SAGE suffers the same limitation, i.e. it maps correctly the 3’-end but does not identify variant transcripts nor does it link unambiguously to a 5’ tag.



NATURE REVIEWS | GENETICS

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.

VOLUME 10 | JANUARY 2009 | 57

review

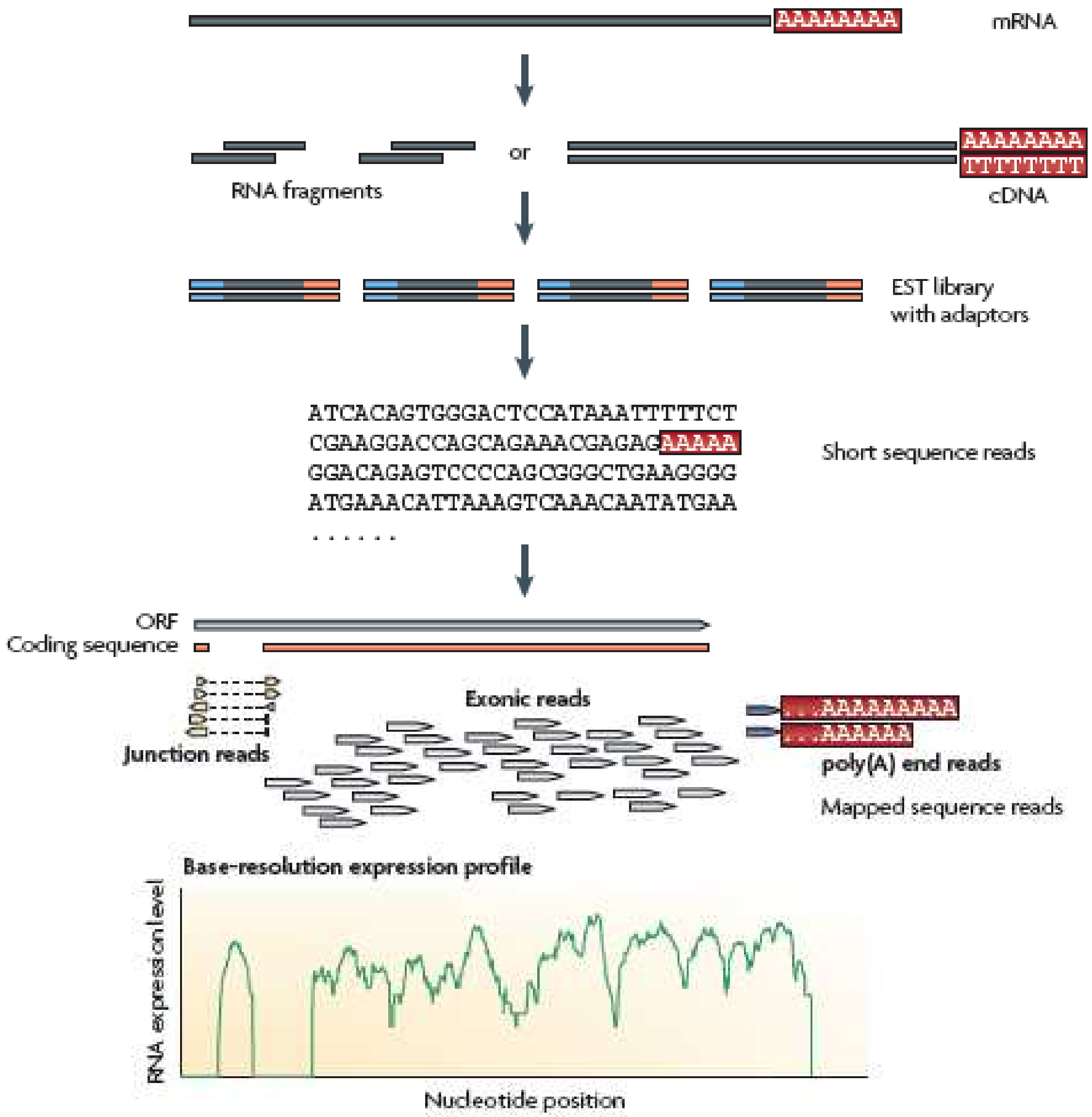


Figure 1 | **A typical RNA-Seq experiment.** Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation (see main text). Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each 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.

map to genome or to transcript/mRNA database

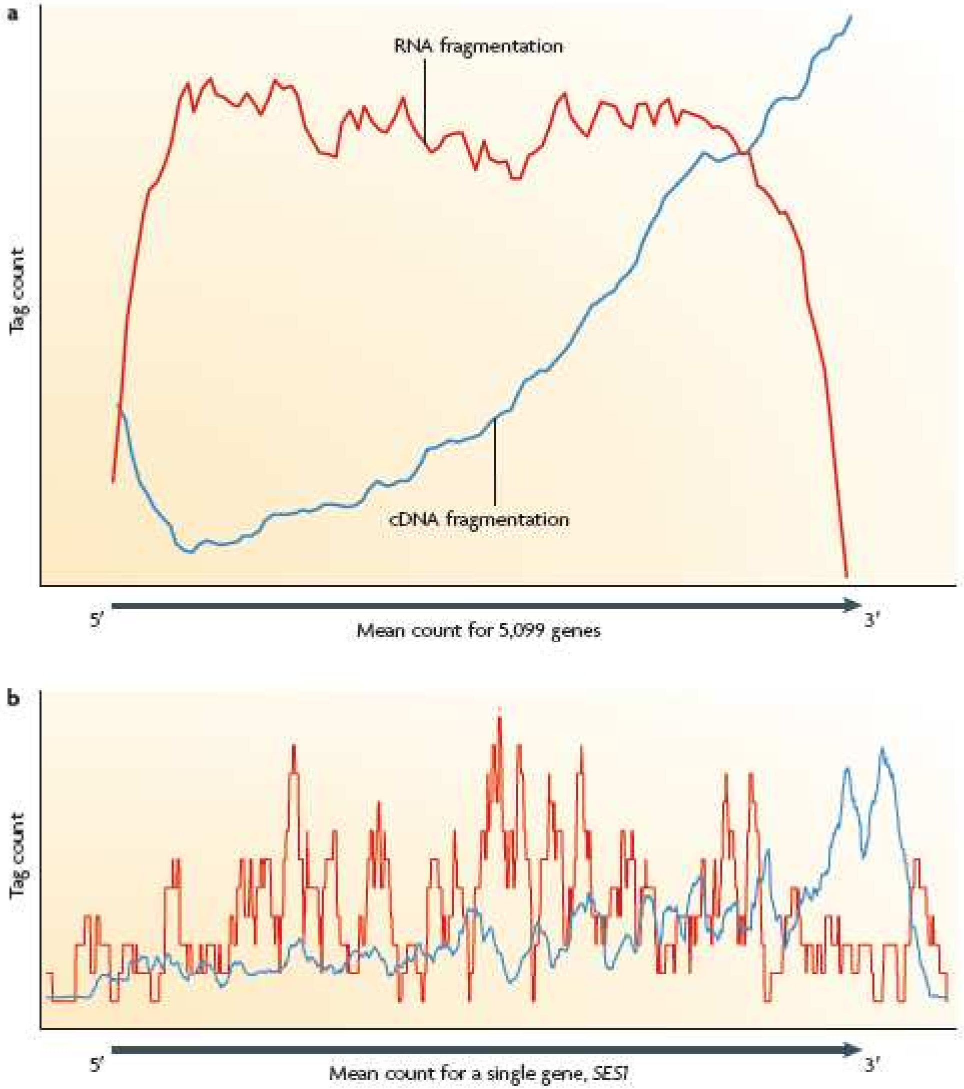


Figure 3 | **DNA library preparation: RNA fragmentation and DNA fragmentation compared.** **a** | Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3′ end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5′ and 3′ ends. Note that the ratio between the maximum and minimum expression level (or the dynamic range) for microarrays is 44, for RNA-Seq it is 9,560. The tag count is the average sequencing coverage for 5,000 yeast ORFs[18]. **b** | A specific yeast gene, *SES1* (seryl-tRNA synthetase), is shown.

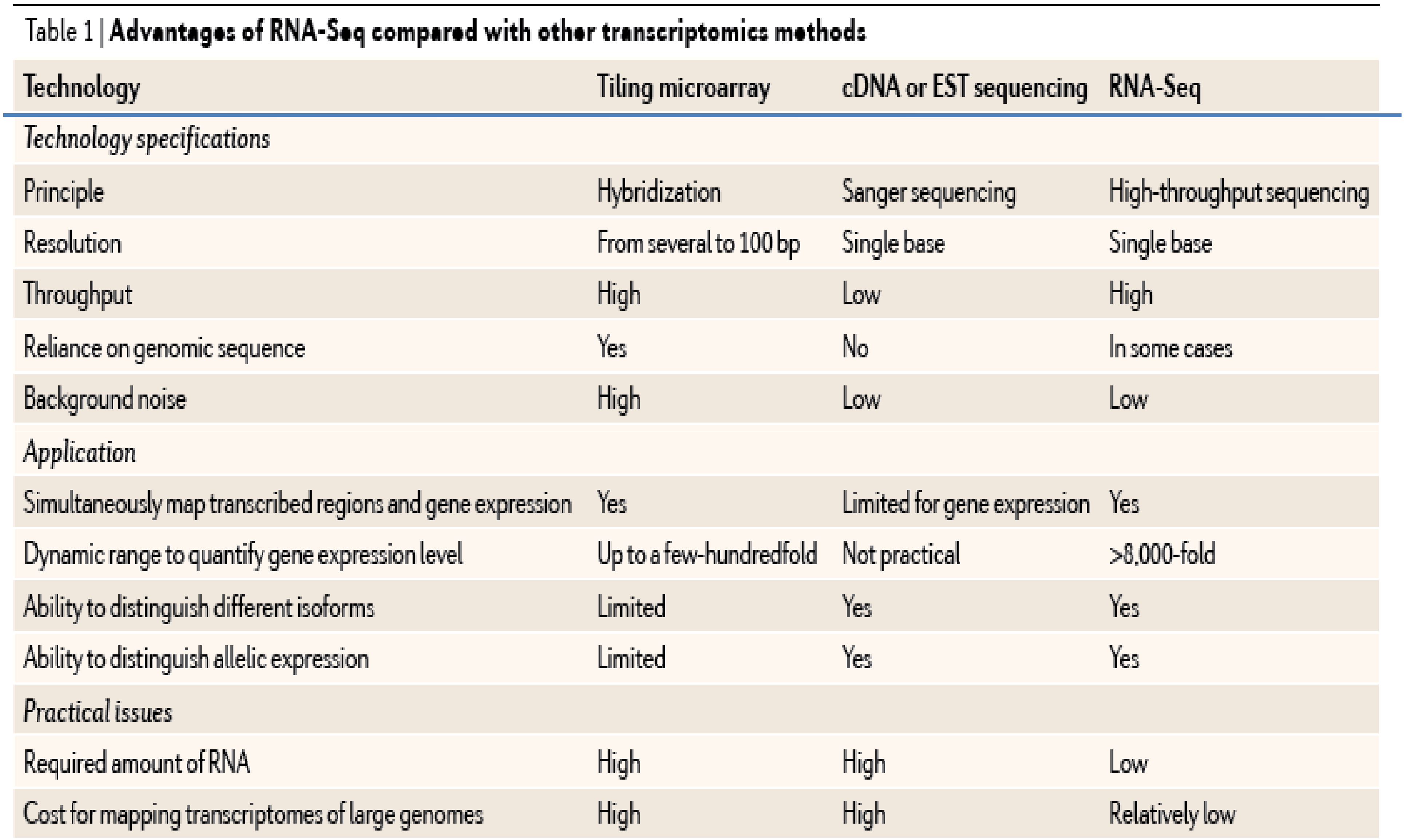


Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

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

Are tiling array experimens and RNA-Seq comparable?

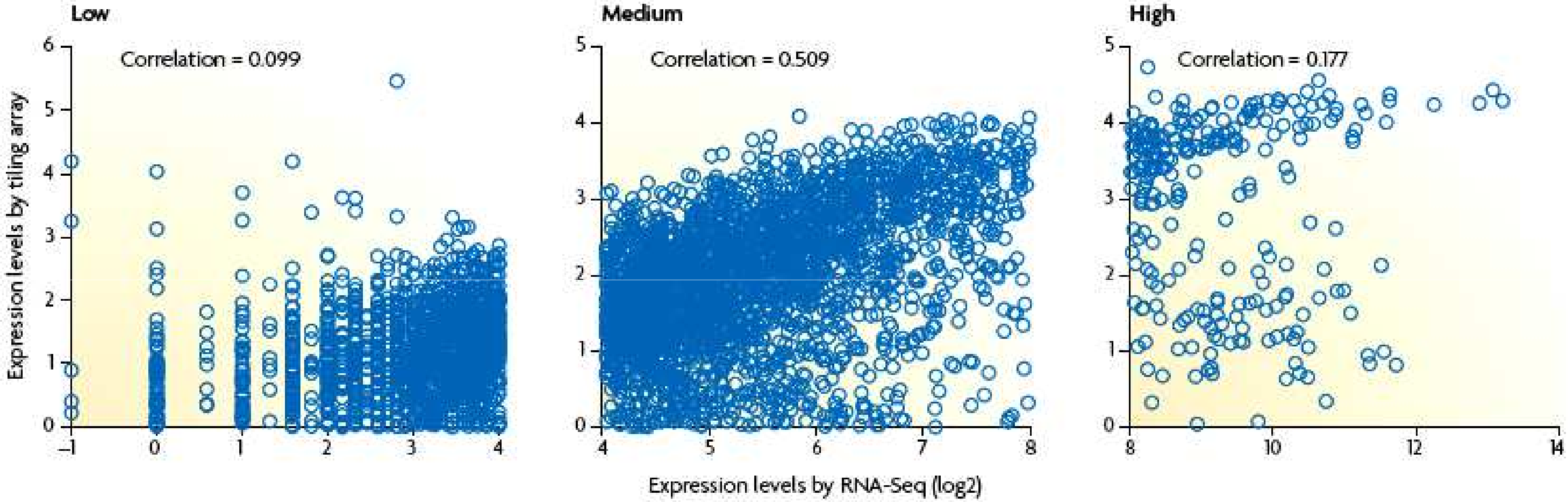


Figure 2 | **Quantifying expression levels: RNA-Seq and microarray compared.** Expression levels are shown, as measured by RNA-Seq and tiling arrays, for *Saccharomyces cerevisiae* cells grown in nutrient-rich media. The two methods agree fairly well for genes with medium levels of expression (middle), but correlation is very low for genes with either low or high expression levels. The tiling array data used in this figure is taken from REF. 2, and the RNA-Seq data is taken from REF. 18.

How many tags are required?
(sequencing depth)

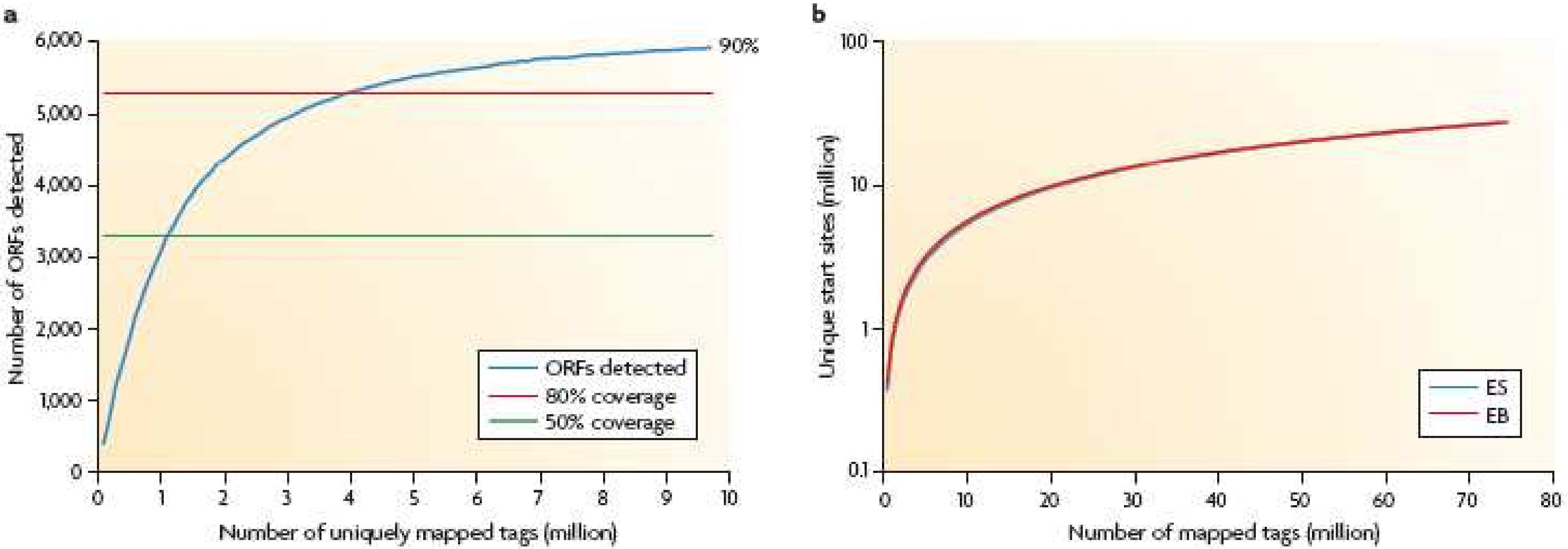


Figure 5 | **Coverage versus depth.** a | 80% of yeast genes were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3′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

quantitative fluorogenic real-time PCR and RT-PCR

Is a method quantitative? How to assess this point?

Data validation

A number of genes are taken at random, with care that they represent all the dynamic range

On the same RNA, they are analyzed by qRT-PCR

Spike-in method

Some synthetic RNA or cDNA, not represented in the analyte, are added to the sample at known concentrations as an internal reference standard.

Identification of several unknown T.U, (transcription units)

→ encoding new protein
→ noncoding small RNA (20-30 nt)
→ noncoding long RNA

From different chromosomal locations:

✓Classical protein encoding genes in regions previously "intergenic"
✓"Within genes" (intragenic) transcripts in Sense and anti-sense orientation
✓Intronic transcripts (S/AS)
✓Small 5' and 3' transcripts

Some of these RNA functionally classified:

rRNA
tRNA
Protein-coding RNA
snRNA
snoRNA
Micro-RNA (miRNA) – siRNA – piRNA

Few noncoding "long" RNA with known or suspected function

A plethora of short and long transcripts with unknown functions

Evidence of pervasive transcription derived from high-throughput studies:

- -EST libraries
- -Tiled microarray analysis
- -SAGE analysis
- -RNA-seq (deep sequencing)
- -CAGE analysis

> **Throughput** **Da Wikipedia, l'enciclopedia libera.**
>
> *Nell'ambito delle telecomunicazioni, si intende per **throughput** di un link (canale) di comunicazione, la sua capacità di trasmissione effettivamente utilizzata. Il “throughput” è la quantità di dati trasmessi in una unità di tempo, il secondo.*

In the mouse, at least 63% of the genome is transcribed. The majority of transcriptional units (TU) do not encode for proteins.

(Carninci et al., 2005, Science 309:1559-63)

In humans, wide transcription seen in 10 chromosomes, 43% of RNA stay in nuclei and are not polyadeylated

(Cheng et al., 2005, Science 308:1149-54)

The ENCODE project results on 1% of the human genome show 93% of the genome transcribed in multiple RNAs.

(Birney et al., 2007, Nature 447: 799-816)

Conclusions: there is pervasive transcription and the majority of RNAs do not show protein-coding evidence

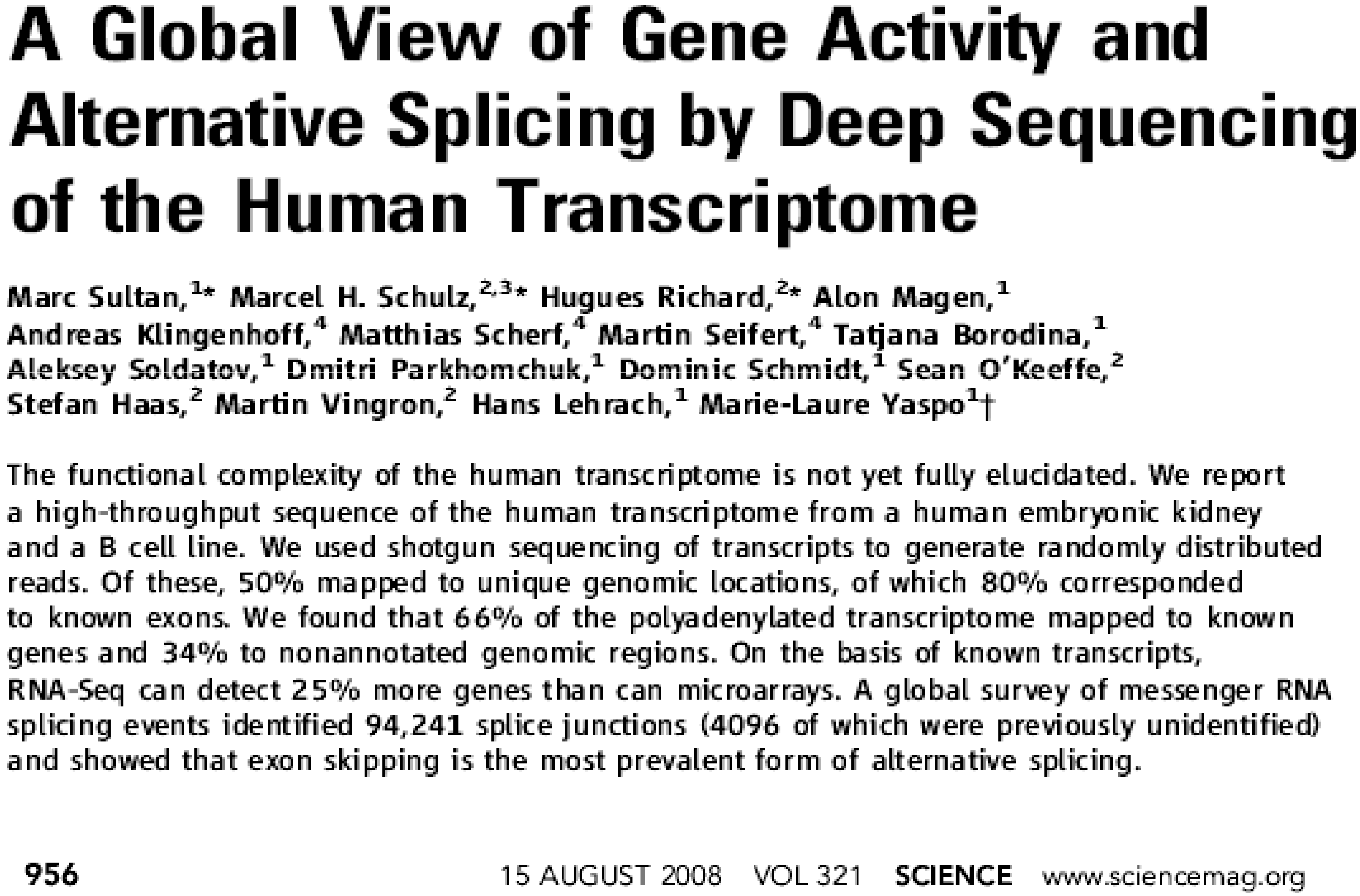
The notion of pervasive transcription, at the level some studies has suggested (see previous slide) is today widely discussed and questioned, since many of the unknown trascripts are very low level (for example, many intergenic transcripts) and may represent “leaky” random transcription.

IMPORTANT:

Sequencing methods, such as SAGE or Deep-Seq, are apparently more sensitive than microarrays for genes showing very low levels of expression.

However, contamination of the RNA sample by thiny amounts of DNA may lead to false positive.

Assigned paper



A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome

Marc Sultan,[1]* Marcel H. Schulz,[2,3]* Hugues Richard,[2]* Alon Magen,[1]
Andreas Klingenhoff,[4] Matthias Scherf,[4] Martin Seifert,[4] Tatjana Borodina,[1]
Aleksey Soldatov,[1] Dmitri Parkhomchuk,[1] Dominic Schmidt,[1] Sean O'Keeffe,[2]
Stefan Haas,[2] Martin Vingron,[2] Hans Lehrach,[1] Marie-Laure Yaspo[1]†

The functional complexity of the human transcriptome is not yet fully elucidated. We report a high-throughput sequence of the human transcriptome from a human embryonic kidney and a B cell line. We used shotgun sequencing of transcripts to generate randomly distributed reads. Of these, 50% mapped to unique genomic locations, of which 80% corresponded to known exons. We found that 66% of the polyadenylated transcriptome mapped to known genes and 34% to nonannotated genomic regions. On the basis of known transcripts, RNA-Seq can detect 25% more genes than can microarrays. A global survey of messenger RNA splicing events identified 94,241 splice junctions (4096 of which were previously unidentified) and showed that exon skipping is the most prevalent form of alternative splicing.

Gene expression by deep-sequencing

All genome can be re-sequenced starting from known primers by Solexa® - Illumina technology or other competitor technology.

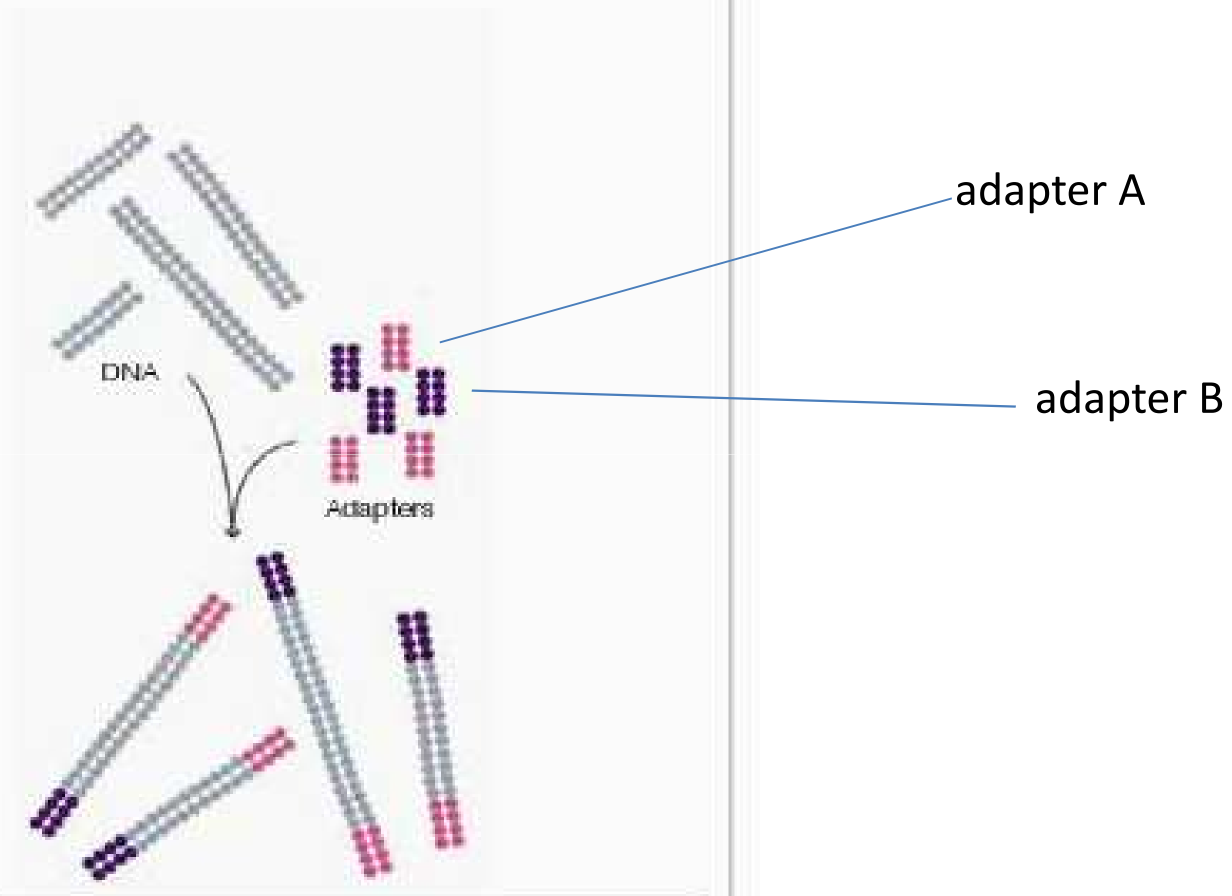
RNA application (RNA-Seq)

1. RNA extraction from cells or tissues
2. cDNA synthesis from oligo(dT) primers
3. ds cDNA tags ligated to A and B primers
4. all cDNA re-sequenced on a chip (microarray) containing millions of spot with complementary aA and bB primers
5. sequence is read → table of frequency of each match

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

How the Solexa-Illumina works

Figure 2: Prepare Genomic DNA Sample



Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

denature and link

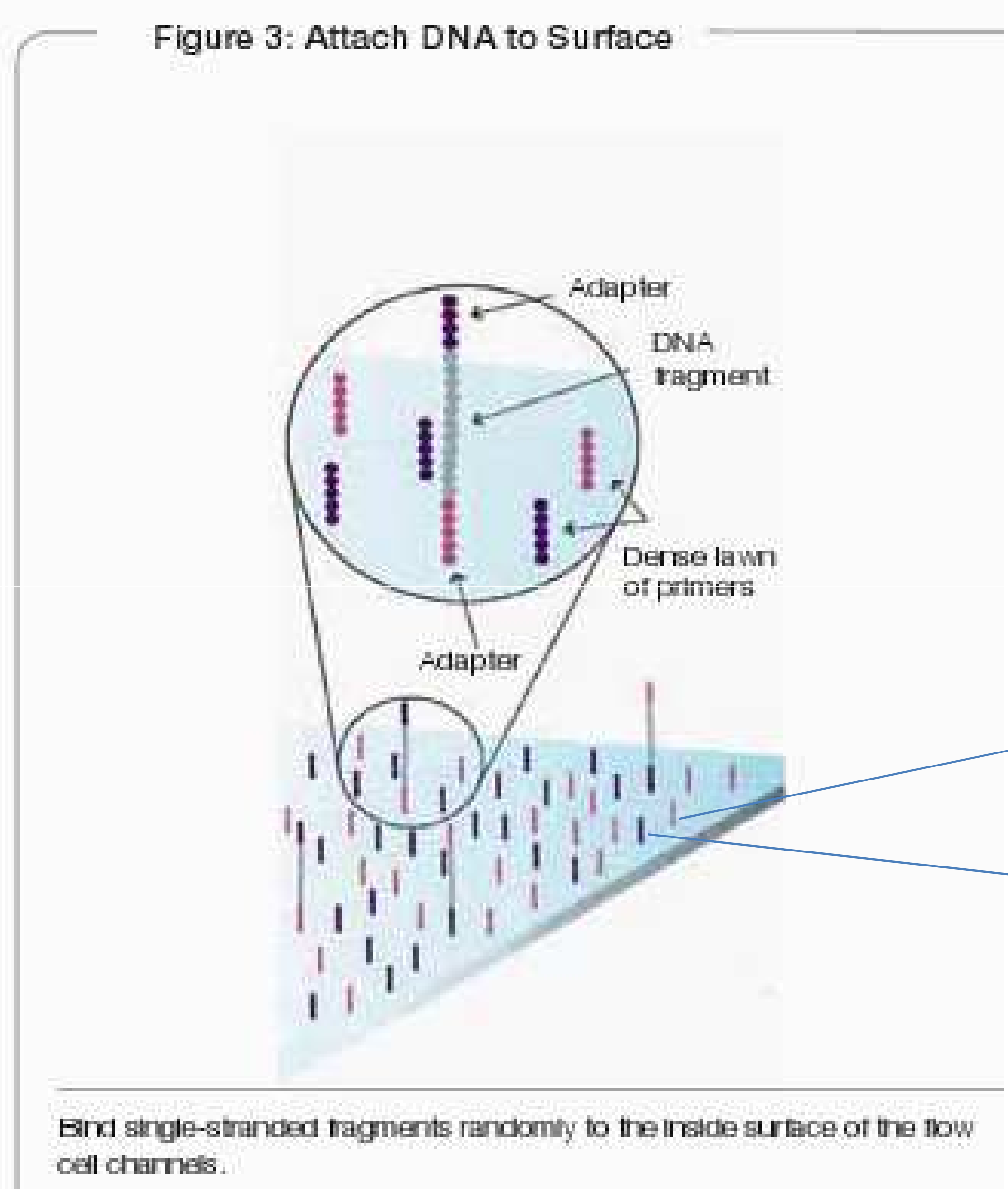


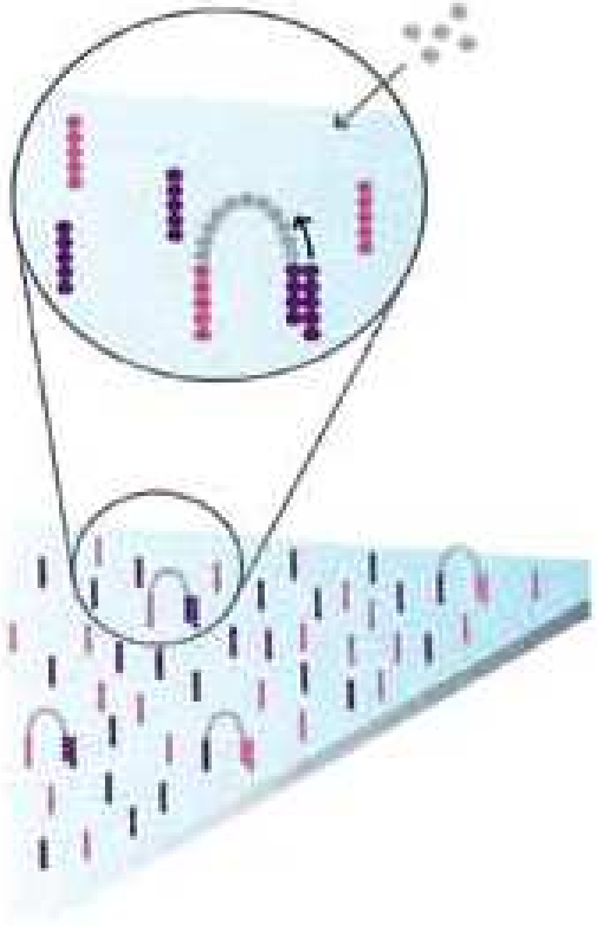
Figure 3: Attach DNA to Surface

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

compl. adapter A

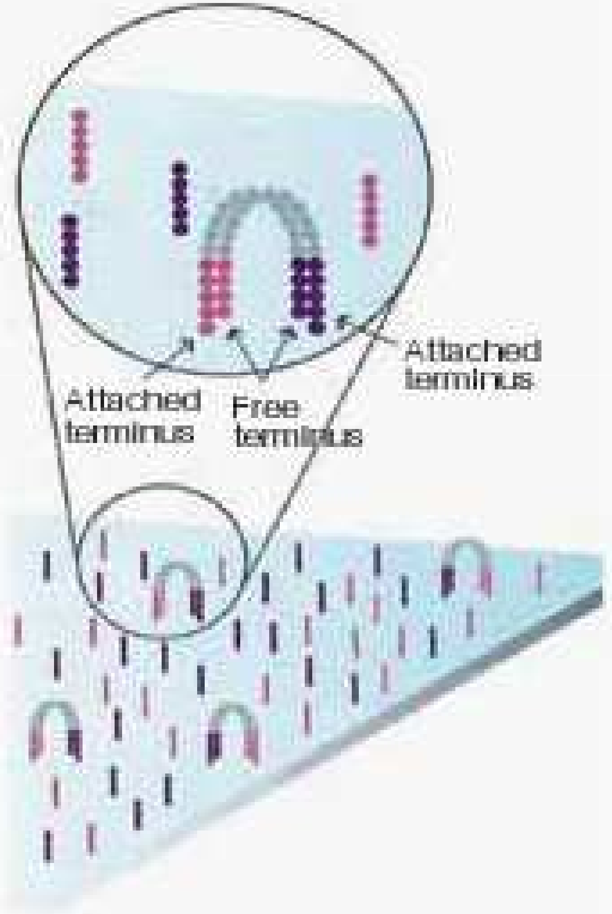
compl. adapter B

Figure 4: Bridge Amplification



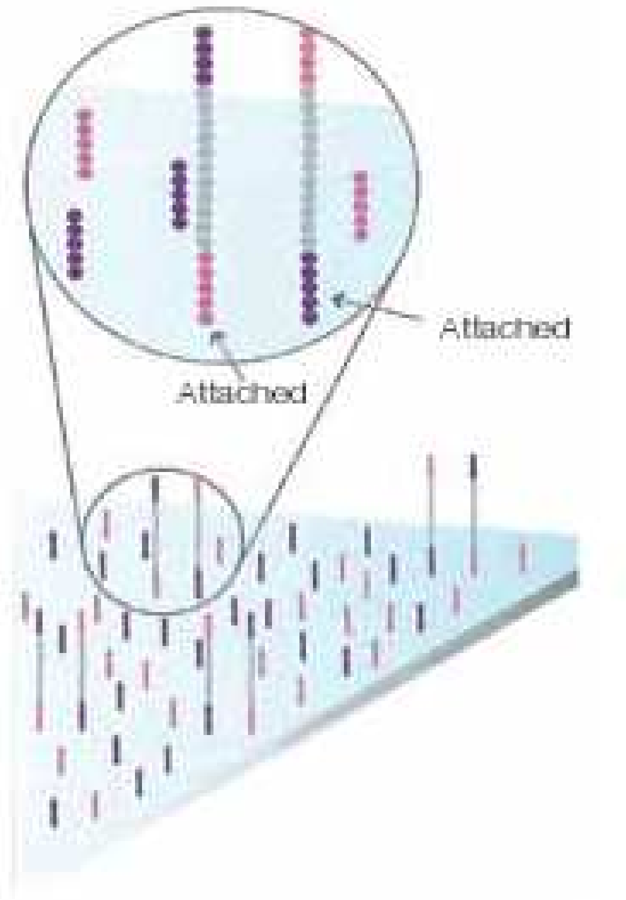
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Figure 5: Fragments Become Double Stranded



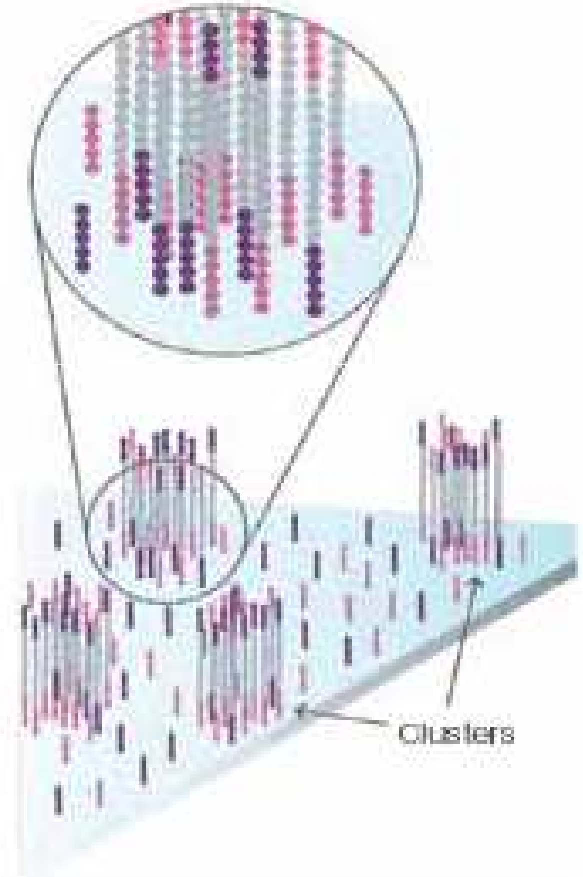
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Figure 6: Denature the Double-Standed Molecules



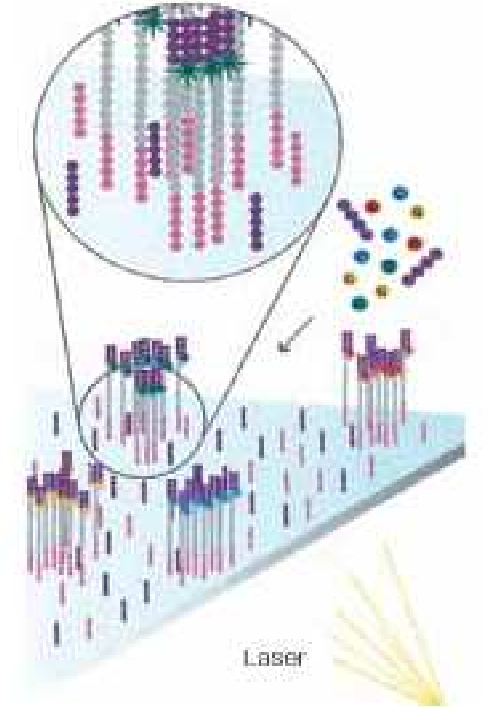
Denaturation leaves single-stranded templates anchored to the substrate.

Figure 7: Complete Amplification



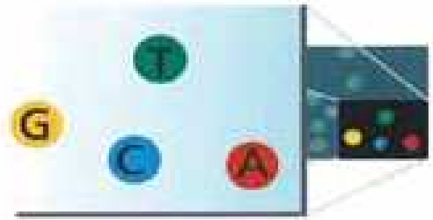
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Figure 8: Determine First Base



The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

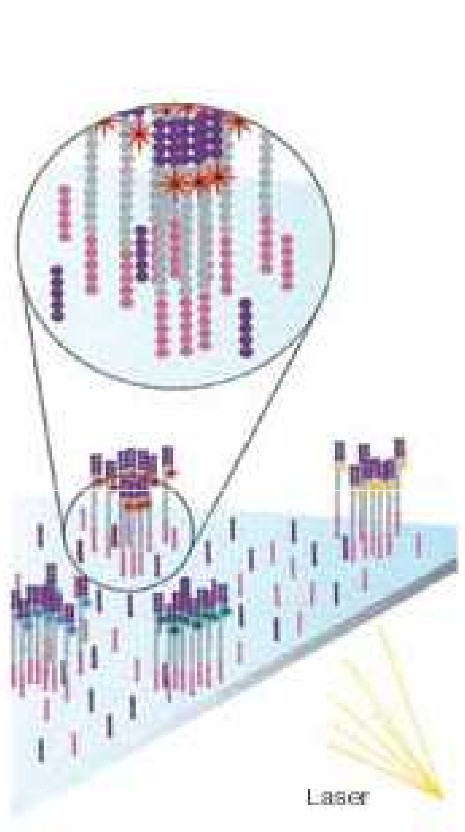
Figure 9: Image First Base



After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

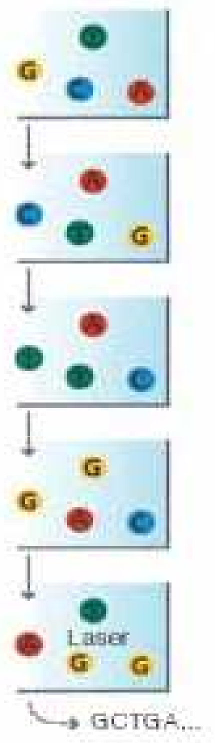
Sequencing is obtained by using flurochrome-labelled reversible terminators

Figure 10: Determine Second Base



The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

Figure 12: Sequencing Over Multiple Chemistry Cycles



The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

...and so on

A table of reads is generated (with associated frequency)

then reads are aligned to the genome sequence

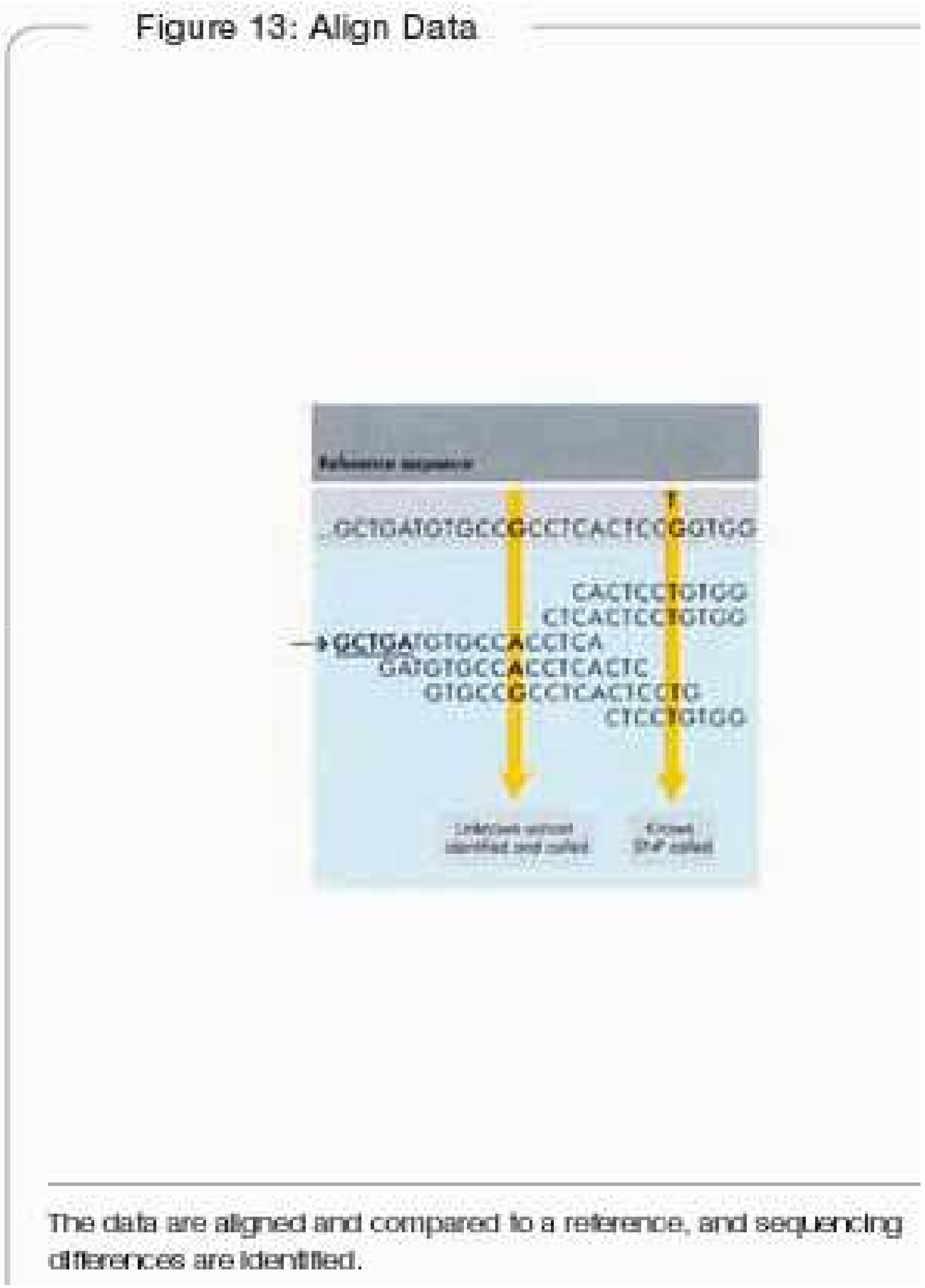
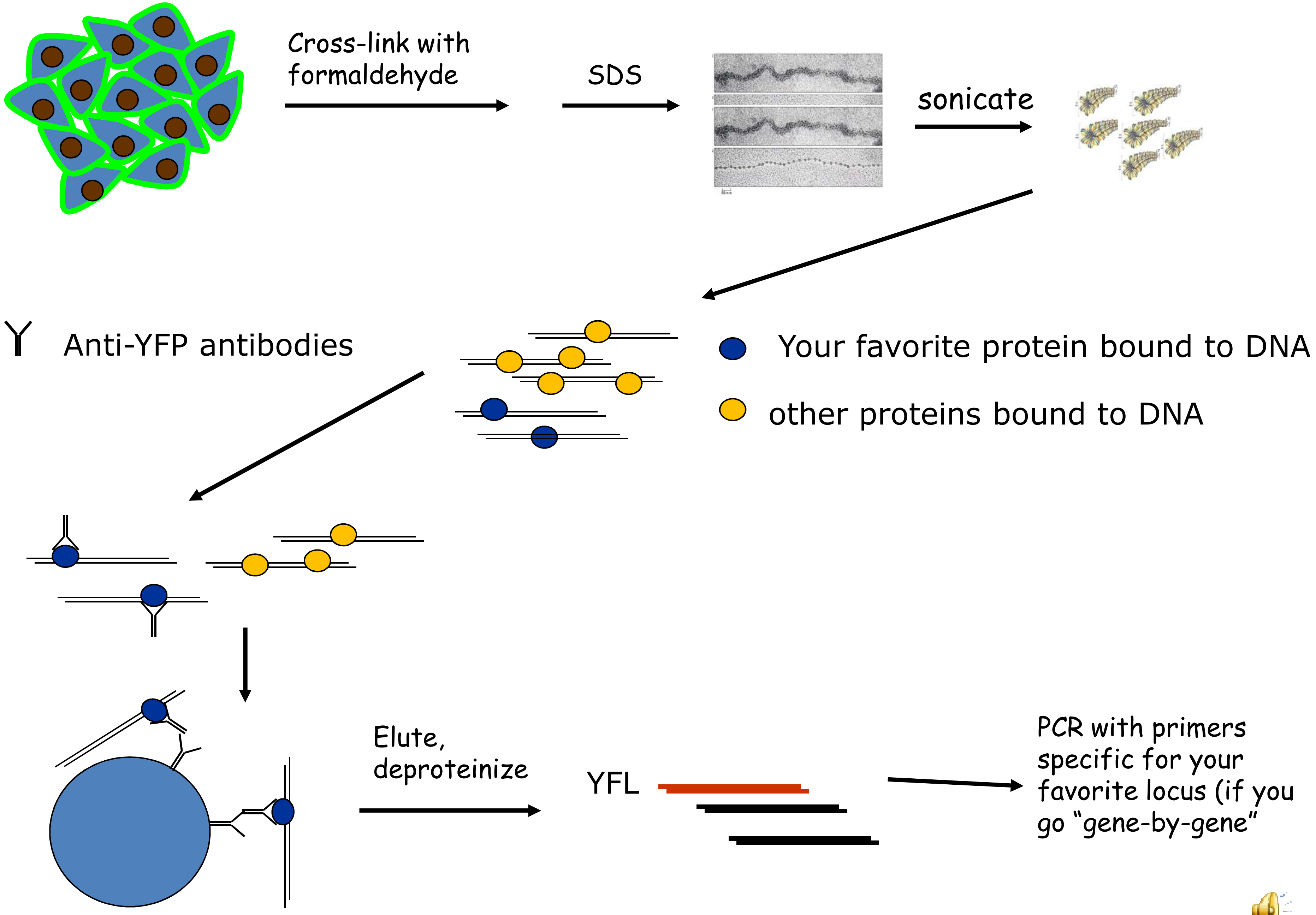


Figure 13: Align Data

The data are aligned and compared to a reference, and sequencing differences are identified.

Chromatin immunoprecipitation (ChIP)



Cross-link with formaldehyde

SDS

sonicate

Anti-YFP antibodies

🔵 Your favorite protein bound to DNA

🟡 other proteins bound to DNA

Elute, deproteinize

YFL

PCR with primers specific for your favorite locus (if you go "gene-by-gene"