

In the previous lecture, we have discussed:

- classical sequencing methods
- newer automatic sequencing methods
- solid-phase parallel sequencing
- Next Generation mass-sequencing methods

Then, we went on to discuss “genome expression” and described:

- Microarrays

- | | |
|--|-------------------------|
| 1. Spotted – Probes: PCR products or long oligos | 2-colors relative meas. |
| 2. In situ synthesized: | |
| 1. short – Affymetrix (photolithography, probeset) | 1 color absolute meas. |
| 2. longer (ink-jet technology) | 1 or 2 – colors |
| 3. arrayed beads (Illumina Bead-arrays) | 1 color absolute meas. |

Today's lecture

- 1° part of the lecture: more microarrays
 - One of the first studies published with genome-wide microarray analysis as a paradigm of what microarrays can tell us.
 - Sensitivity issues
 - Gene mapping using microarrays.
- 2° part → Sequencing approaches for gene expression studies
 1. EST (cloning and serial sequencing of cDNA libraries)
 2. SAGE (short 3' tags concatamers from cDNA)
 3. CAGE (Cap-driven 5' tags concatamers from cDNA)
 4. RNA analysis by mass-sequencing (RNA-Seq).

<http://www.bio.davidson.edu/Courses/genomics/chip/chip.html>

(generale sull'uso dei microarrays)

Specifico per uso AffyChip®

http://www.affymetrix.com/corporate/media/genechip_essentials/gene_expression/Gene_Expression_Analysis.affx

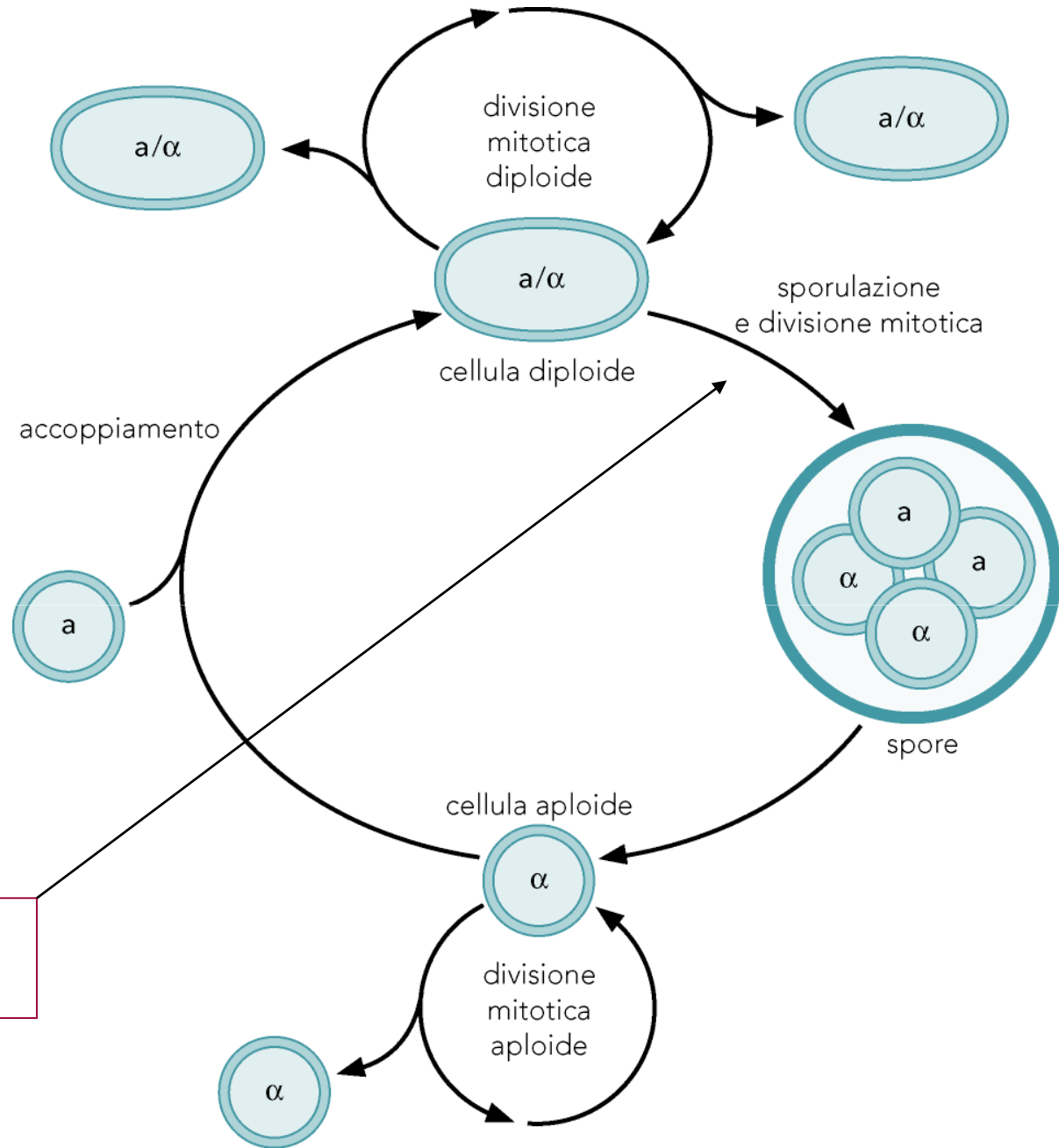
Assigned paper

The Transcriptional Program of Sporulation in Budding Yeast

S. Chu,* J. DeRisi,* M. Eisen, J. Mulholland, D. Botstein,
P. O. Brown,† I. Herskowitz†

Diploid cells of budding yeast produce haploid cells through the developmental program of sporulation, which consists of meiosis and spore morphogenesis. DNA microarrays containing nearly every yeast gene were used to assay changes in gene expression during sporulation. At least seven distinct temporal patterns of induction were observed. The transcription factor Ndt80 appeared to be important for induction of a large group of genes at the end of meiotic prophase. Consensus sequences known or proposed to be responsible for temporal regulation could be identified solely from analysis of sequences of coordinately expressed genes. The temporal expression pattern provided clues to potential functions of hundreds of previously uncharacterized genes, some of which have vertebrate homologs that may function during gametogenesis.

Saccaromyces cerevisiae



nitrogen-deficient medium induces sporulation

DNA microarrays

Sensitivity

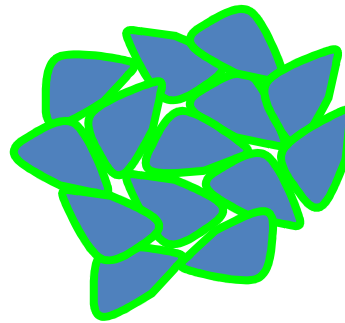
Issues:

Sensitivity

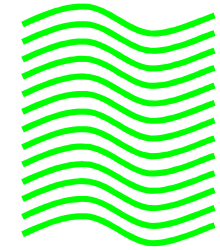
What an array does measure

What an array does not measure

Direct labelling, no amplification



RNA extraction



Primers: oligo(dT) or random oligomers

Reverse transcriptase

dNTPs

1 labelled NTP (e.g. Cy3-CTP)



Sensitivity:

mRNA is 2-4 % of total RNA

1 μ g Tot RNA \rightarrow 20-40 ng mRNA

Assuming that 10,000 genes are expressed, on average each mRNA species is 2-4 pg

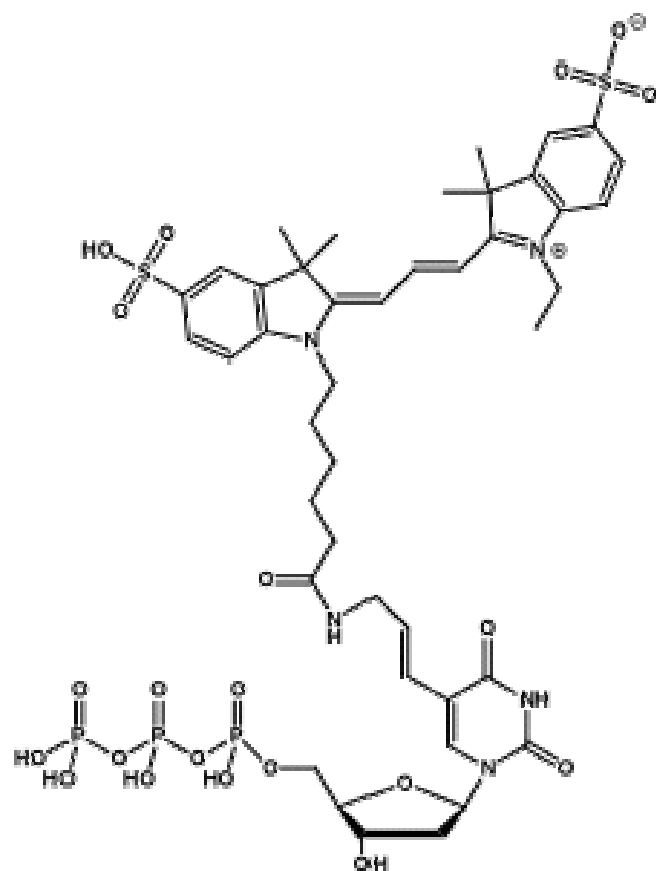
The number of mRNA molecules/cell of individual genes ranges from 0 to some thousands.

i.e. , for some genes, we are measuring a **very low** number of mRNA molecules.

Microarrays:

Sample preparation

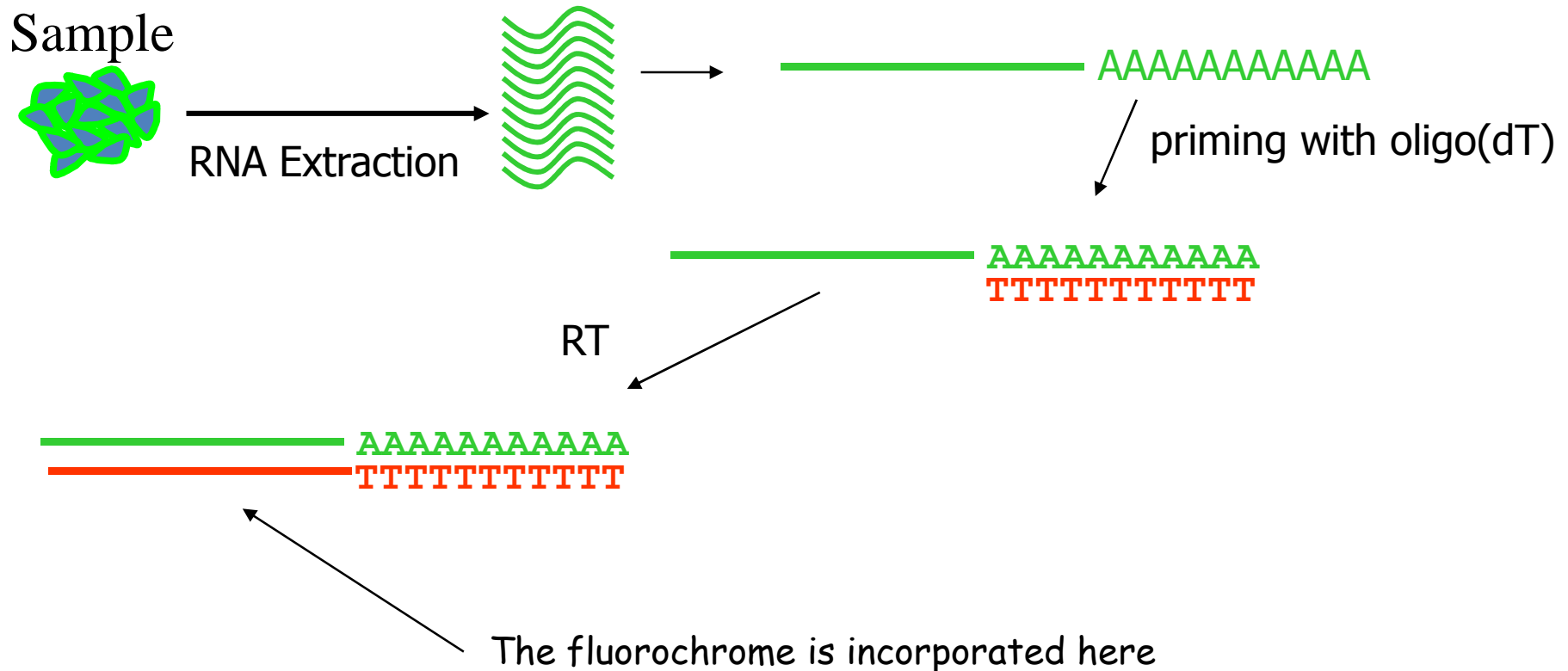
(sometimes the RNA sample is also called “complex probe”....not to be confused with the “probes” that are oligos or cDNA fixed to the chip surface...)



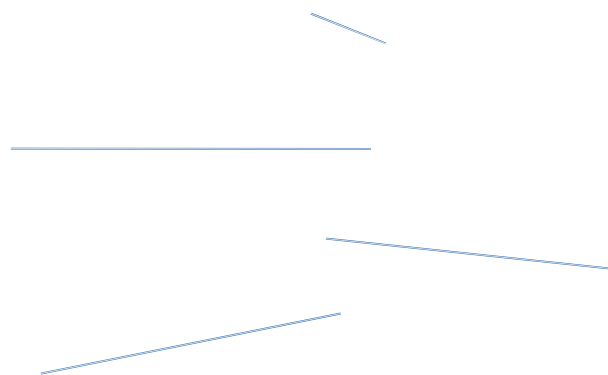
Cy3-dUTP

Nonamplified complex probe preparation:

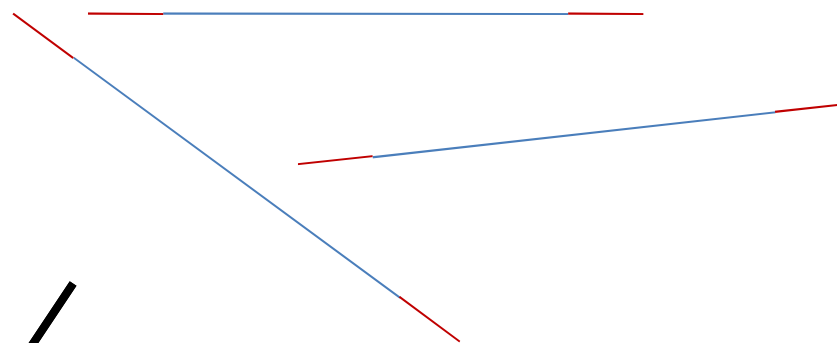
RNAs are labelled using fluorochrome-conjugated ribonucleotides (NTP) directly by reverse transcription (RT), priming the synthesis either with oligo-dT or with random primers.



PCR amplification ?

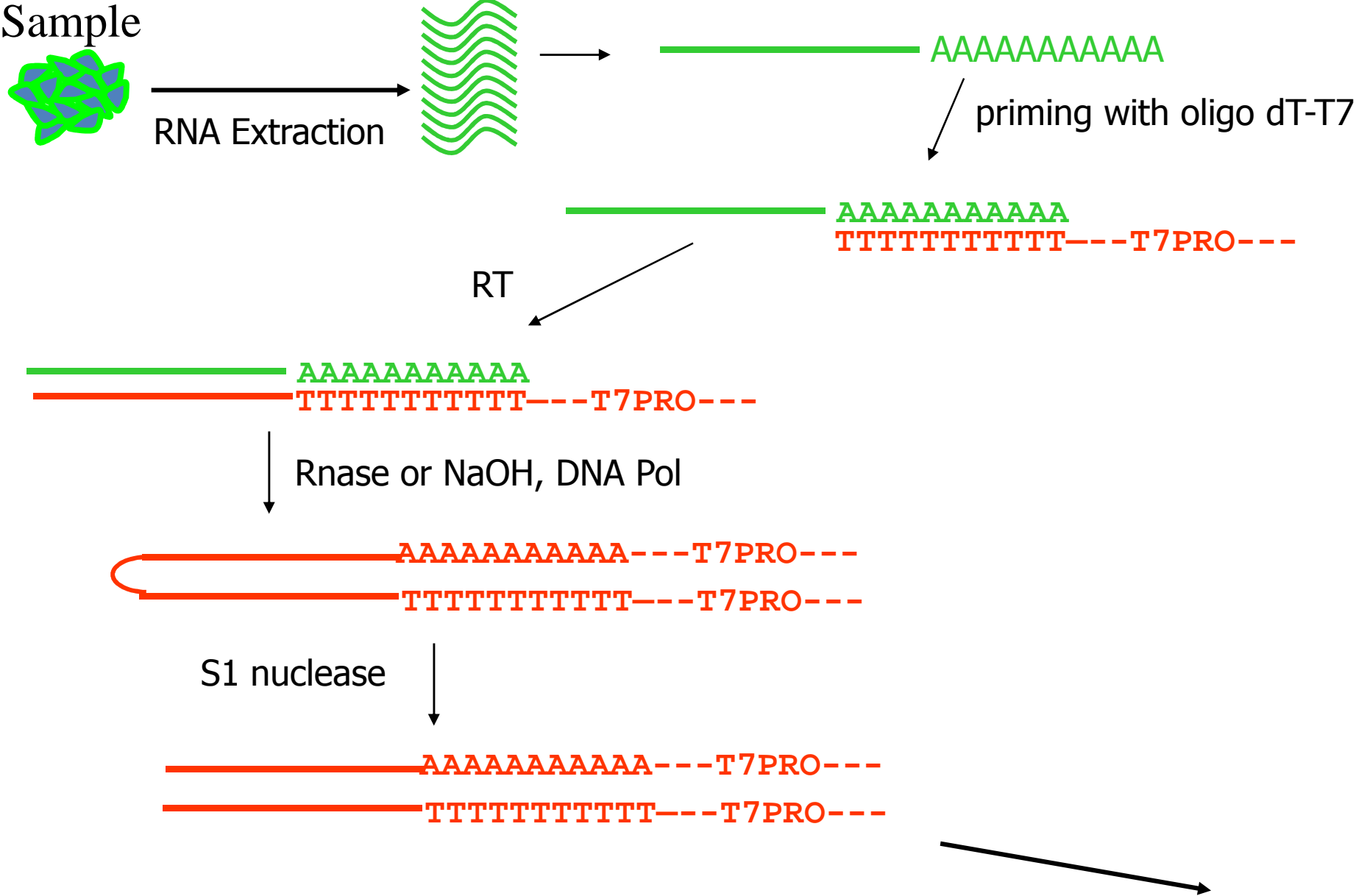


Linker ligation



PCR amplification

Amplified complex probe preparation

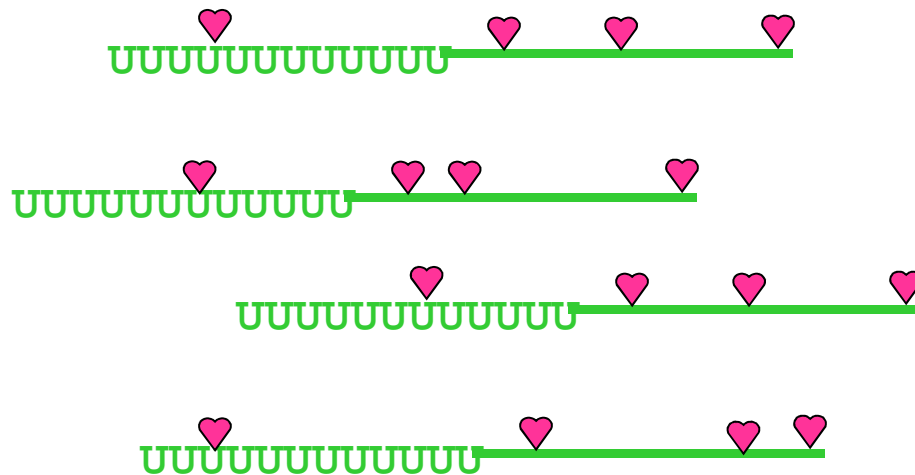


---T7PRO---TTTTTTTTTTT
---T7PRO---AAAAAAAAAAA



T7 RNA Polymerase, NTPs, + labelled UTP

♥ = label



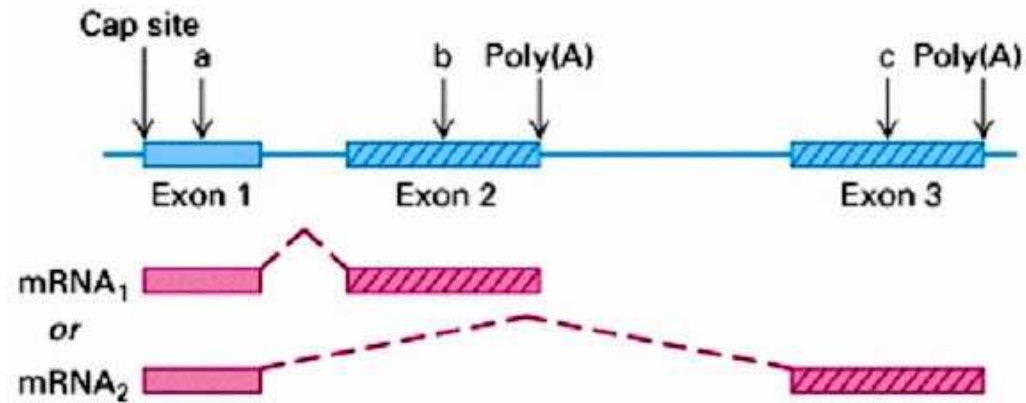
Label may be a fluorochrome or a detectable modification, like biotin or digoxigenin, or a chemical group that can be conjugated with fluorochrome after transcription (e.g. allyl-UTP)

TRANSCRIPTION
Linear Amplification of each
sequence that was originally
present in starting RNA, but
complementary
= "cRNA"

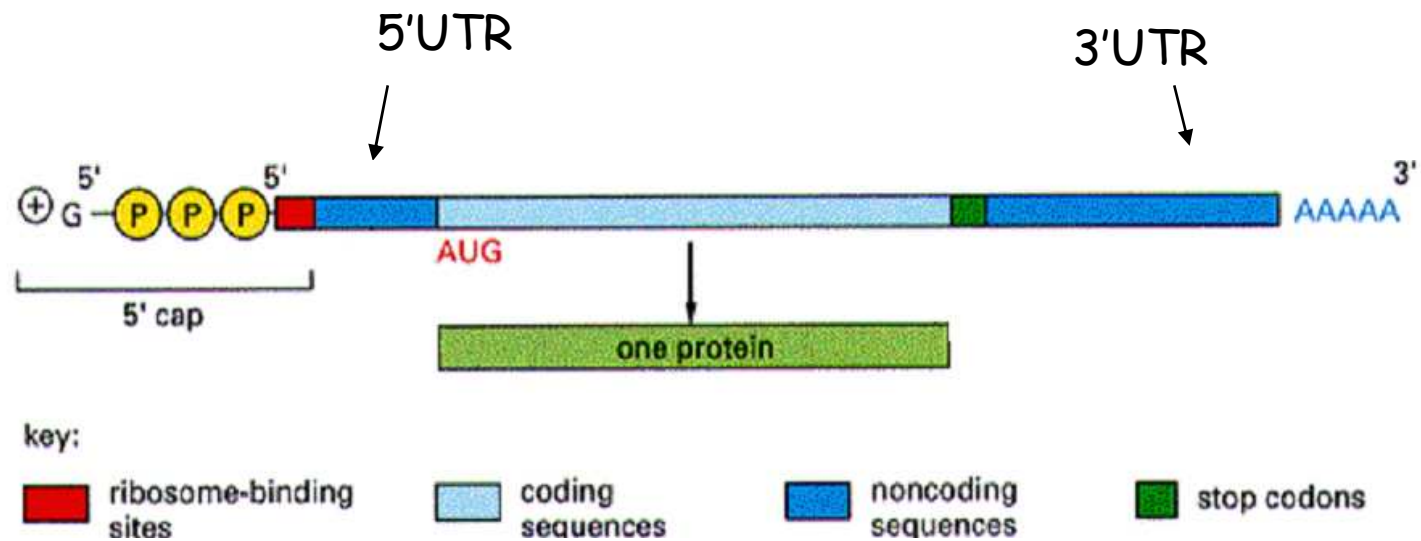
Problem 2.

Which probe? Which target?

More than 60% of the human gene transcripts are thought to undergo alternative splicing



Probe libraries (cDNA, EST collections, etc) vary widely depending on the methods used for preparation



Issues:

Sensitivity

What an array does measure

What an array does not measure

Like Northern Blot, RT-PCR and qRT-PCR, and RNase protection assay (RPA)
traditional microarrays measure steady-state mRNA quantity, **NOT transcription !!!**

Like Northern Blot, RT-PCR and qRT-PCR, and RNase protection assay (RPA)
traditional microarrays measure mRNA, **NOT protein levels !!!**

Microarrays have been and still are extensively used to:

1. monitor changes in gene expression in experimental situations

- development
- regulatory biology
- pathology
- pharmacology
- environment
- nutrition

2. diagnostic and classification

- oncology
- prognosis
- prediction

A second use....

How many genes are in mammalian genomes?

How many sequences are transcribed?

How many and how many types of noncoding RNA (ncRNA) are transcribed?

Do ncRNA play functional roles?

This paper accompanied the publication of the first HG draft in the 15 February 2001 issue of Nature. It concerned the problem of <finding genes> in the genome.

articles

Experimental annotation of the human genome using microarray technology

D. D. Shoemaker*, E. E. Schadt*, C. D. Armour, Y. D. He, P. Garrett-Engle, P. D. McDonagh, P. M. Loerch, A. Leonardson, P. Y. Lum, G. Cavet, L. F. Wu, S. J. Altschuler, S. Edwards, J. King, J. S. Tsang, G. Schimmack, J. M. Schelter, J. Koch, M. Ziman, M. J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M. R. Meyer, M. Mao, J. Burchard, M. J. Kidd, H. Dai, J. W. Phillips, P. S. Linsley, R. Stoughton, S. Scherer & M. S. Boguski

Rosetta Inpharmatics, Inc., 12040 115th Avenue N.E., Kirkland, Washington 98034, USA

** These authors contributed equally to this work*

The most important product of the sequencing of a genome is a complete, accurate catalogue of genes and their products, primarily messenger RNA transcripts and their cognate proteins. Such a catalogue cannot be constructed by computational annotation alone; it requires experimental validation on a genome scale. Using 'exon' and 'tiling' arrays fabricated by ink-jet oligonucleotide synthesis, we devised an experimental approach to validate and refine computational gene predictions and define full-length transcripts on the basis of co-regulated expression of their exons. These methods can provide more accurate gene numbers and allow the detection of mRNA splice variants and identification of the tissue- and disease-specific conditions under which genes are expressed. We apply our technique to chromosome 22q under 69 experimental condition pairs, and to the entire human genome under two experimental conditions. We discuss implications for more comprehensive, consistent and reliable genome annotation, more efficient, full-length complementary DNA cloning strategies and application to complex diseases.

Exonic arrays

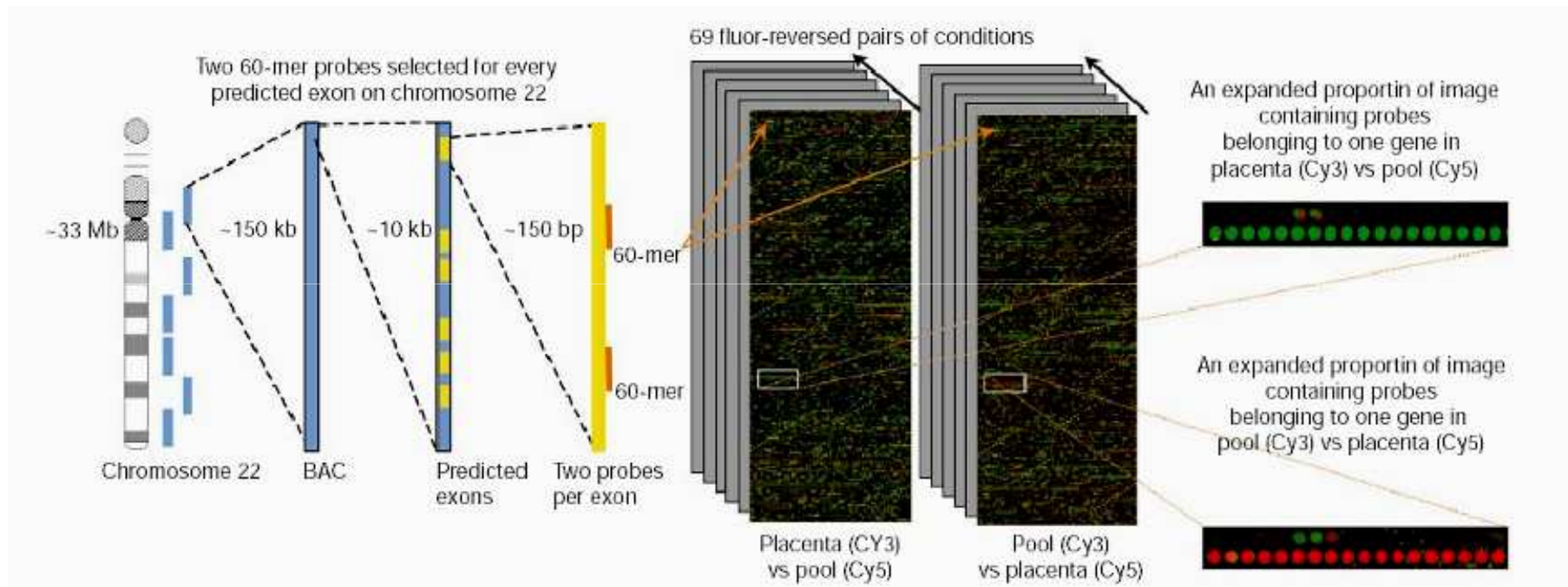


Figure 1 Design and fabrication of exon arrays for the predicted exons on human chromosome 22. Two 60-mers were selected from each of 8,183 predicted exons on human chromosome 22q and printed on a single 1 x 3 inch array (~25,000 60-mers). This array was hybridized with 69 pairs of RNA samples using a two-colour hybridization technique. Each experiment was performed in duplicate with a fluor reversal to minimize

possible bias caused by the molecular structure of the Cy3 and Cy5 dyes (138 arrays in total). Red and green spots, as shown in the expanded panels on the right, are probes representing experimentally verified genes (groups of differentially expressed exons that are located next to each other in the genome).

This is conceptually strongly biased:

“I have to predict genes to make probes” but what about genes that I can not predict ?

One solution (before deep-sequencing) was < tiling microarrays >

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

Of course it makes a lot of probes!

HG: 3.2×10^9 → 1.5 nonrepetitive

I need 50 millions 30-mer probes !



Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments

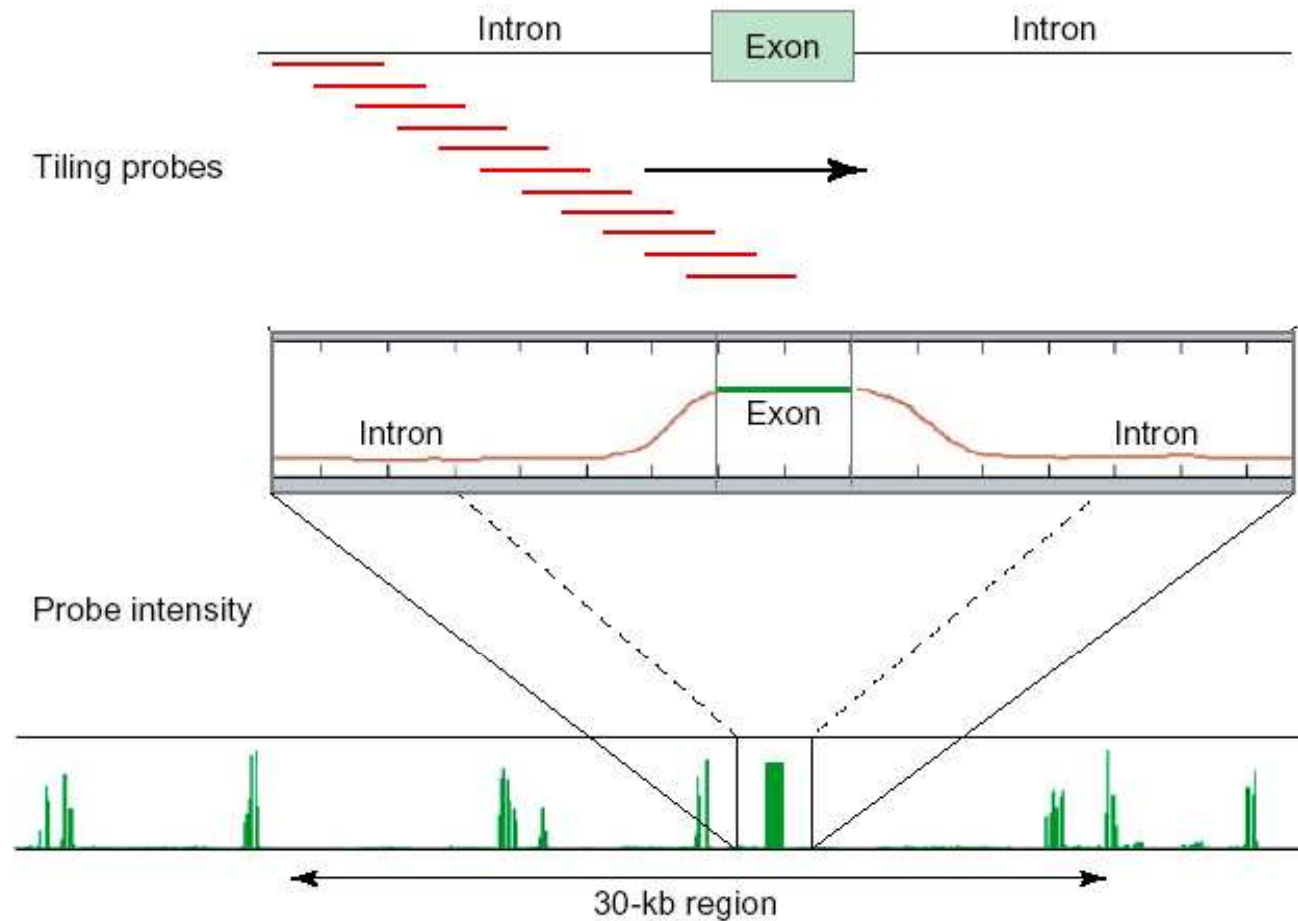
Jason M. Johnson¹, Stephen Edwards¹, Daniel Shoemaker² and Eric E. Schadt¹

¹Rosetta Inpharmatics LLC^{*}, 401 Terry Avenue North, Seattle, WA 98109, USA

²GHC Technologies, 505 Coast Boulevard South, Suite 309, La Jolla, CA 92037, USA

Microarrays provide the opportunity to measure transcription from regions of the genome without bias towards the location of known genes. This technology thus offers an important source of genomic sequence annotation that is complementary to cDNA sequencing and computational gene-finding methods. Recent 'tiling' microarray experiments that assay transcription at regular intervals throughout the genome have shown evidence of large amounts of transcription outside the boundaries of known genes. This transcription is observed in polyadenylated RNA samples and appears to be derived from intergenic regions, from introns of known genes and from sequences antisense to known transcripts. In this article, we discuss different explanations for this phenomenon.

review



TRENDS in Genetics

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 1) 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.

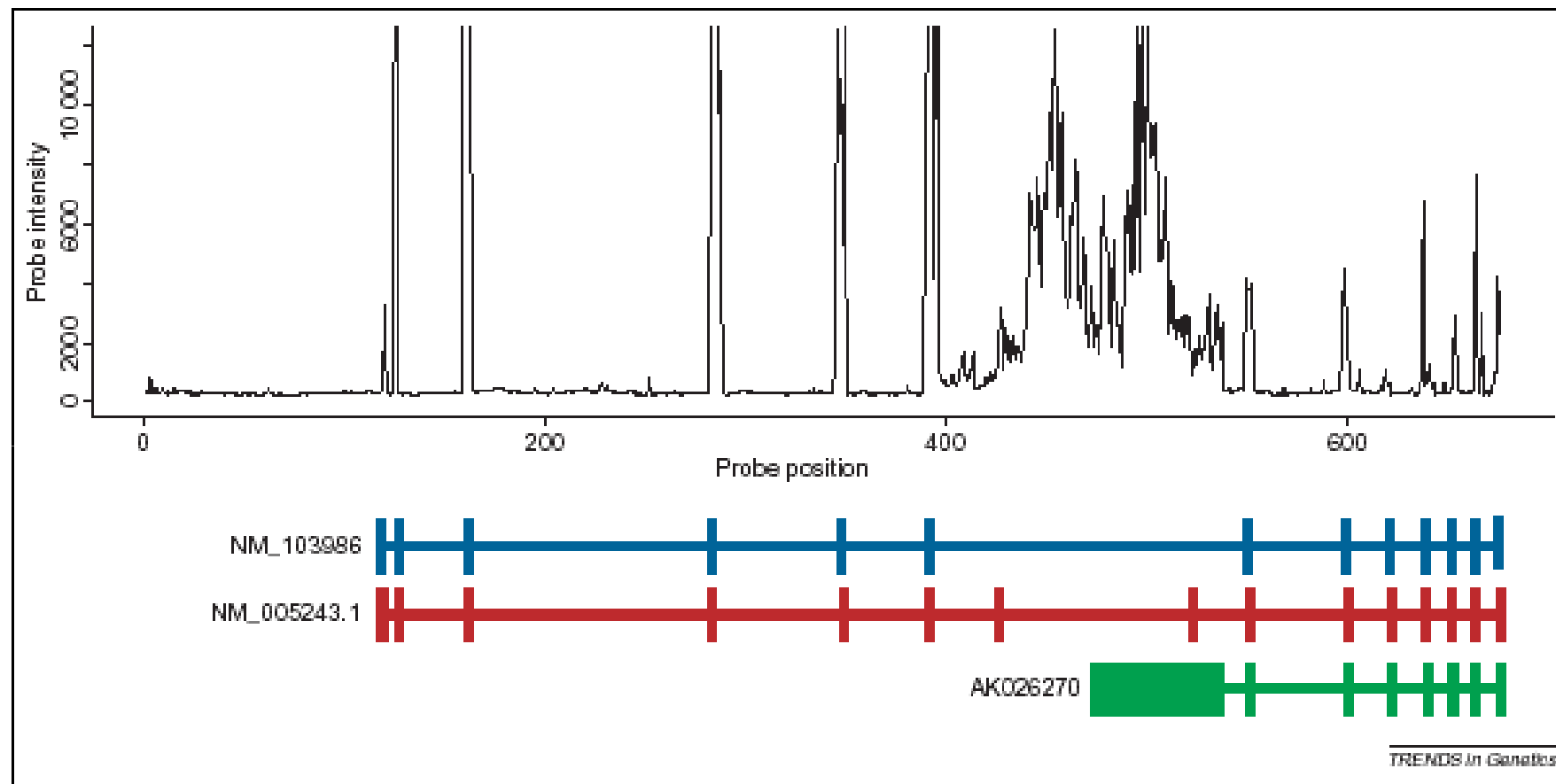


Figure 1. Microarray intensity profile for human thymus poly(A)⁺ cDNA profiled on a 60mer ink-jet tiling array representing the genomic locus of the Ewing sarcoma breakpoint region 1 gene (EWSR1) in 30-nt steps, with probe index as the x-axis. No probes are shown for repeat-masked regions. The tiling data for this locus are shown in relation to the exon positions (below the plot) of three EWSR1 cDNAs (Genbank accession numbers: NM_013986, NM_005243.1 and AK026270). The 5'-most exon lies in a repeat-masked region and is not shown. A few peaks with the highest intensity have been truncated in Figures 1–3.

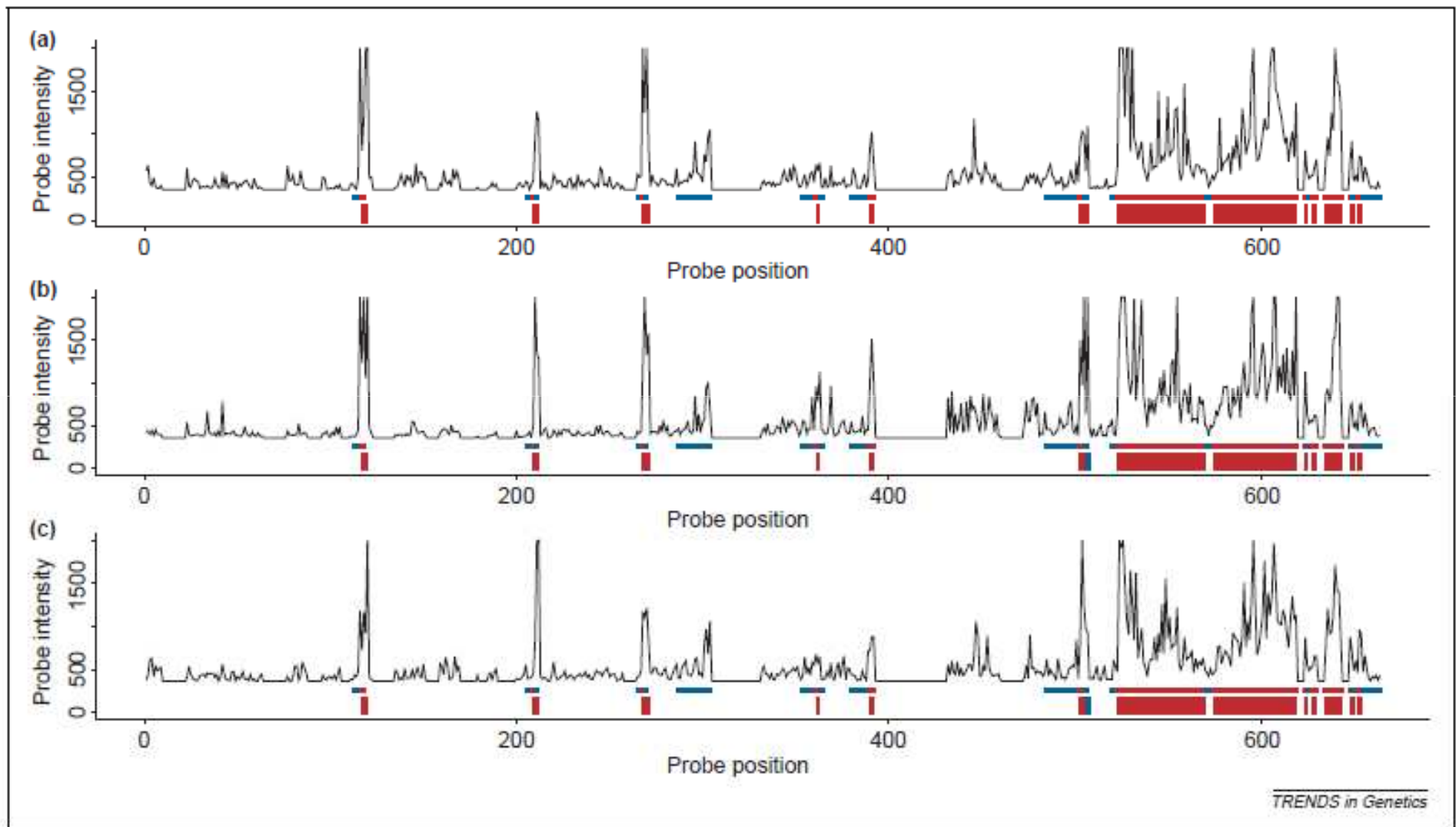


Figure 2. Microarray tiling confirmation of a predicted gene. Microarray tiling probe intensities for a region of human chromosome 20 that contains an *ab initio* gene prediction (made by the program GENSCAN [48]). Predicted exons for this gene are shown with blue lines. The transcription detected by microarrays has been grouped into a single transcriptional unit (dark red) by the correlated behavior of these probes across different human samples [11]. The conditions displayed are (a) thalamus, (b) testes and (c) uterus.

How to measure the activity of all genes (genome-wide) in cells/tissues (mRNA).

Sequencing methods

(EST)

SAGE (LongSAGE, CAGE)

direct re-sequencing (deep sequencing)

Hybridization methods - DNA microarrays, oligonucleotide microarrays.

Spotted arrays

In situ synthesized oligo arrays

Bead-arrays®

EST = expressed sequence tags

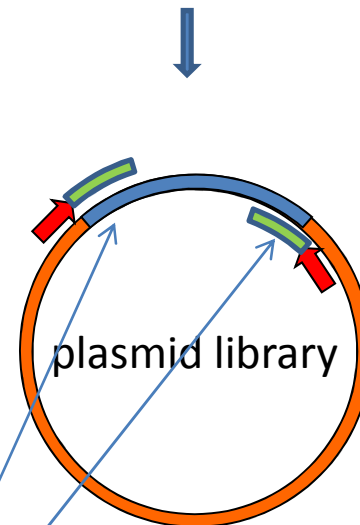
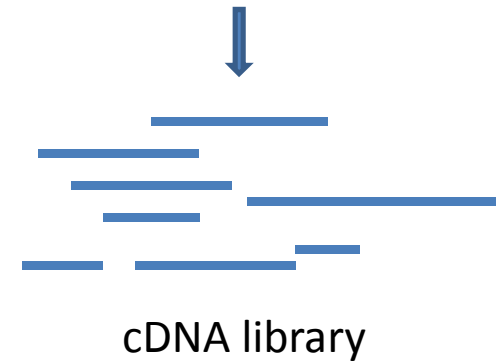
1. mRNA extraction from cells or tissues
2. cDNA synthesis (oligo-dT or random-primed)
3. cloning into plasmid vectors
4. sequencing from vector primers (200-300nt)
5. Estimate expression from frequency

[Current EST Databases contain millions of EST]

how to create a cDNA library

on line sequence databases, BLAST

mRNA from cells

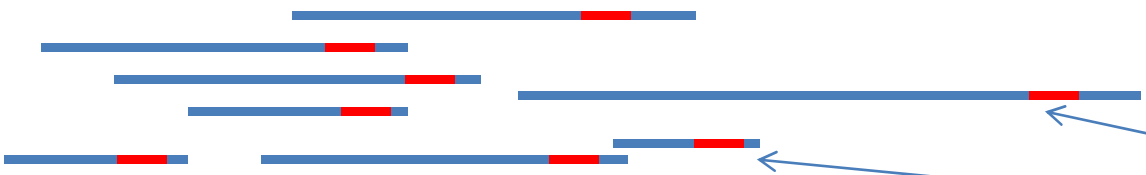


Sequence 200-300 bp for each clone and send to database

cDNA library



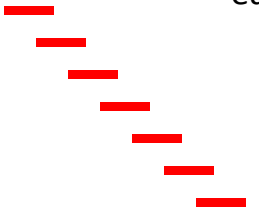
As a matter of fact, to measure gene expression level, we do not need sequencing mRNAs for their entire length, so we can optimize the sequencing job:



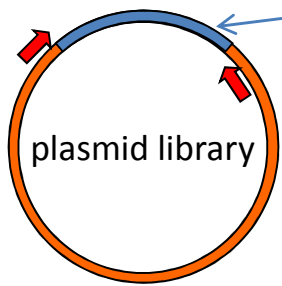
Short sequence "tags" will make the business



get the tags from each cDNA



concatamerize



concatamer library

plasmid library

Sequence

SAGE



SCIENCE • VOL. 173 • 20 OCTOBER 1993

Serial Analysis of Gene Expression

Victor E. Velculescu, Lin Zhang, Bert Vogelstein,
Kenneth W. Kinzler*

The characteristics of an organism are determined by the genes expressed within it. A method was developed, called serial analysis of gene expression (SAGE), that allows the quantitative and simultaneous analysis of a large number of transcripts. To demonstrate this strategy, short diagnostic sequence tags were isolated from pancreas, concatenated, and cloned. Manual sequencing of 1000 tags revealed a gene expression pattern characteristic of pancreatic function. New pancreatic transcripts corresponding to novel tags were identified. SAGE should provide a broadly applicable means for the quantitative cataloging and comparison of expressed genes in a variety of normal, developmental, and disease states.

What you obtain is:

Tag1-Sp-Tag2-Sp-Tag3-Sp-Tag4-Sp.....Sp- Tag_i

(where Sp is a spacer sequence use to make concatamers, e.g. a restriction site).

Each Tag_i can be mapped to the genome, of course (hopefully!)

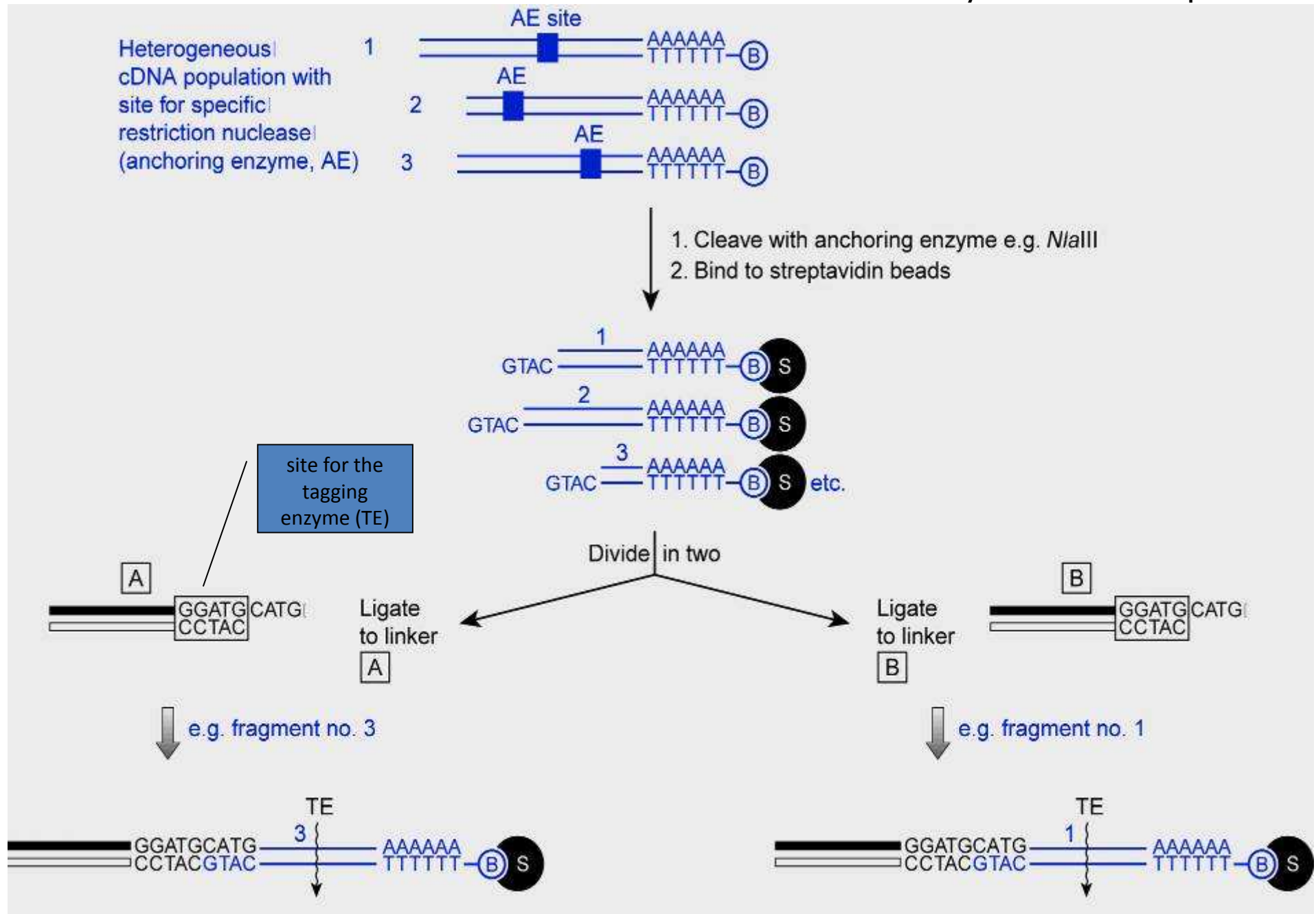
It is very likely that one Tag is represented mor than one time in your library....
so that you get a Table of frequencies:

<u>Tag sequence</u>	<u>times found</u>
CATGGCTACGGGCATTACTGC	125
GTTGGAAGTCATCGTCAGTCG	21
TTAATTAGGGATGACCCCTGAC	1
AAATGTACTCAGTCGTCGTCGT	458
.....
.....

(thousands)

This is proportional
to the relative
abundance in your
sample

SAGE= Serial Analysis of Gene Expression



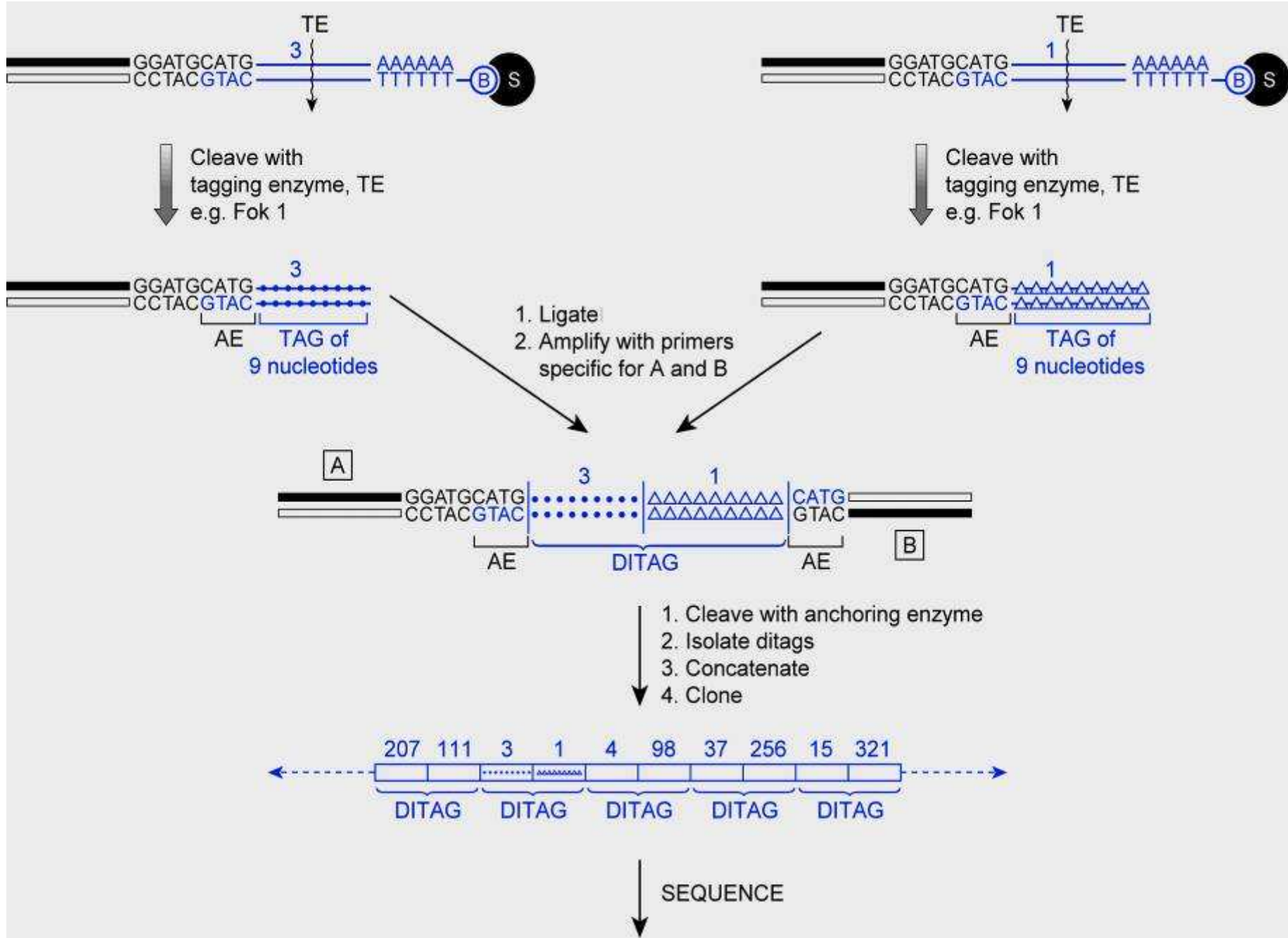
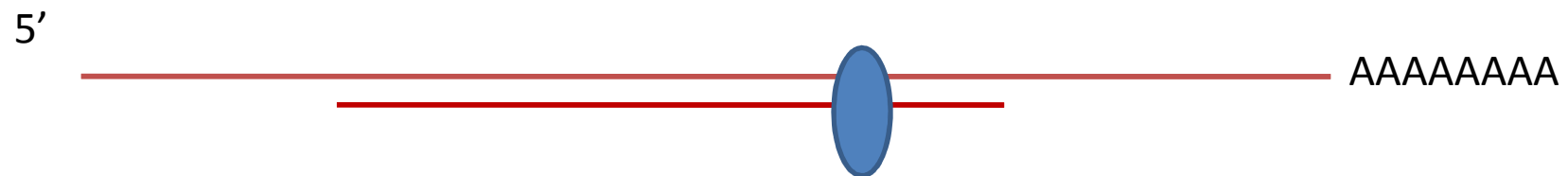


Table 1. Pancreatic SAGE tags. Tag indicates the 9-bp sequence identifying each tag, adjacent to the 4-bp anchoring Nla III site. *n* and Percent indicate the number of times the tag was identified and its frequency, respectively. Gene indicates the description and accession number of the GenBank release 87 entry found to exactly match the indicated tag when the SAGE software group was used, with the following exceptions. When multiple entries were identified because of duplicated entries (7), only one entry is listed. For chymotrypsinogen and trypsinogen 1, other genes (adenosine triphosphatase and myosin alkali light chain, respectively) were identified that were predicted to contain the same tags, but subsequent hybridization and sequence analysis identified the listed genes as the source of the tags. Alu entry indicates a match with a GenBank entry for a transcript that contained at least one copy of the Alu consensus sequence (15).

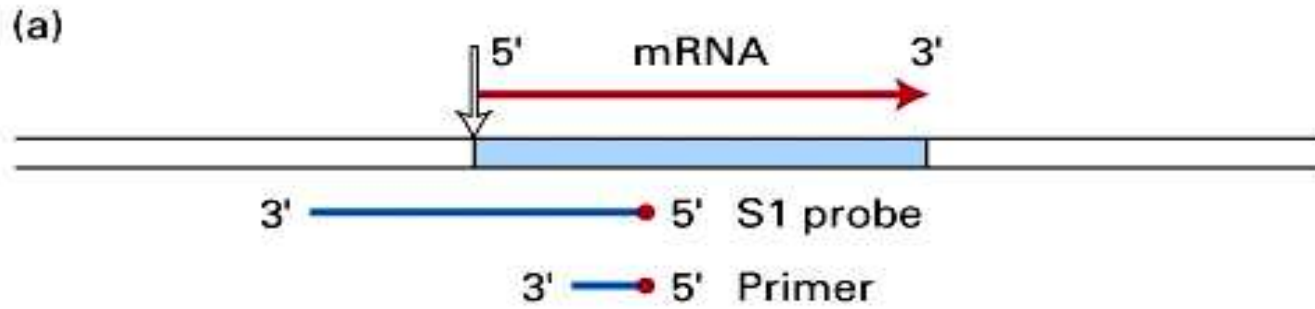
Tag	Gene	<i>n</i>	Percent
GAGCACACC	Procarboxypeptidase A1 (X67318)	64	7.6
TTCTGTGTG	Pancreatic trypsinogen 2 (M27602)	46	5.5
GAACACAAA	Chymotrypsinogen (M24400)	37	4.4
TCAGGGTGA	Pancreatic trypsin 1 (M22612)	31	3.7
GCGTGACCA	Elastase IIIB (M18692)	20	2.4
GTGTGTGCT	Protease E (D00306)	16	1.9
TCATTGGCC	Pancreatic lipase (M93285)	16	1.9
CCAGAGAGT	Procarboxypeptidase B (M81057)	14	1.7
TCCTCAAAA	No match (see Table 2, P1)	14	1.7
AGCCTTGGT	Bile salt stimulated lipase (X54457)	12	1.4
GTGTGCGCT	No match	11	1.3
TGCCGAGACC	No match (see Table 2 P2)	9	1.1
GTGAAACCC	21 Alu entries	8	1.0
GGTGACTCT	No match	8	1.0
AAGGTAACA	Secretory trypsin inhibitor (M11949)	6	0.7
TCCCCTGTG	No match	5	0.6
GTGACCACG	No match	5	0.6
CCTGTAATC	M91159, M29366, 11 Alu entries	5	0.6
CACGTTGGA	No match	5	0.6
AGCCCTACA	No match	5	0.6
AGCACCTCC	Elongation factor 2 (Z11692)	5	0.6
ACGCAGGGA	No match (see Table 2, P3)	5	0.6
AATTGAAGA	No match (see Table 2, P4)	5	0.6
TTCTGTGGG	No match	4	0.5
TTCATACAC	No match	4	0.5
GTGGCAGGC	NF- κ B (X61499), Alu entry (S94541)	4	0.5
GTAAAACCC	TNF receptor II (M55994), Alu entry (X01448)	4	0.5
GAACACACA	No match	4	0.5
CCTGGGAAG	Pancreatic mucin (J05582)	4	0.5
CCCATCGTC	Mitochondrial CytC oxidase (X15759)	4	0.5
SAGE tags occurring:	Greater than three times	380	45.2
	Three times (15 \times 3 =)	45	5.4
	Two times (32 \times 2 =)	64	7.6
	One time	351	41.8
	Total SAGE tags	840	100.0

One common problem in RNA databases is that they are 3' – biased, for practical and historical reasons

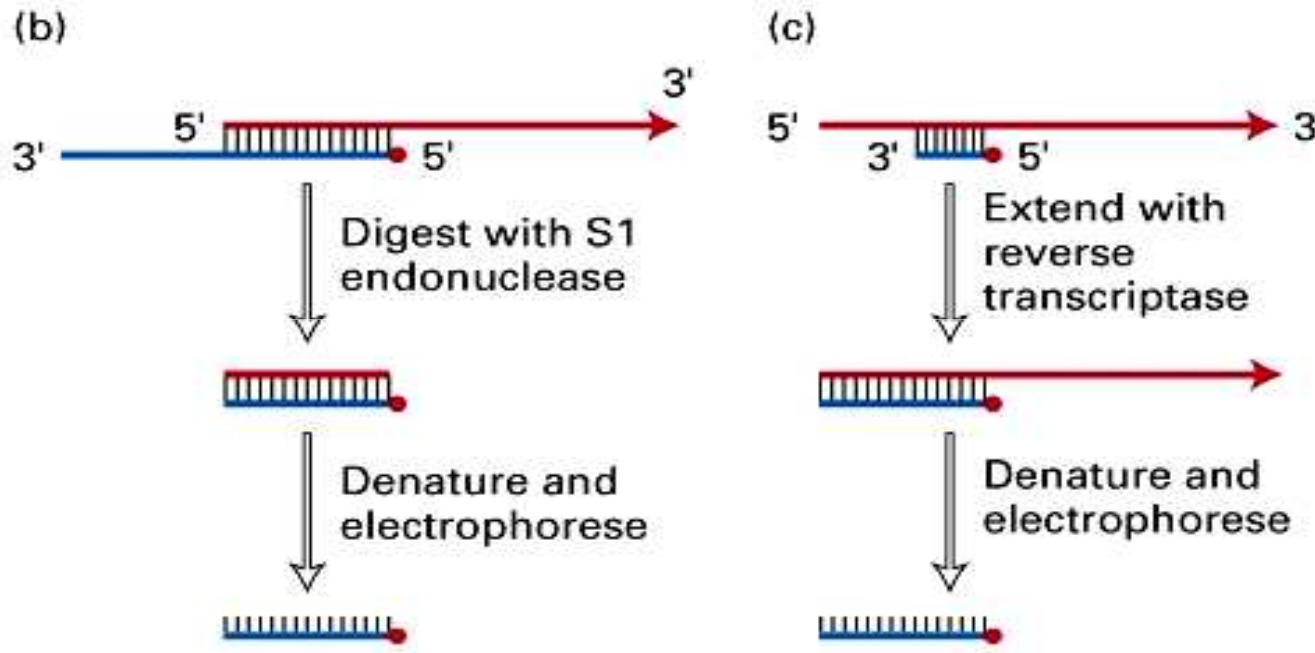
Number 1 problem is that Reverse Transcriptase is not very “processive” and often terminates before reaching the 5' end of RNA.



The problem is analogous to mapping the promoter



Primer extension is an alternative mapping method



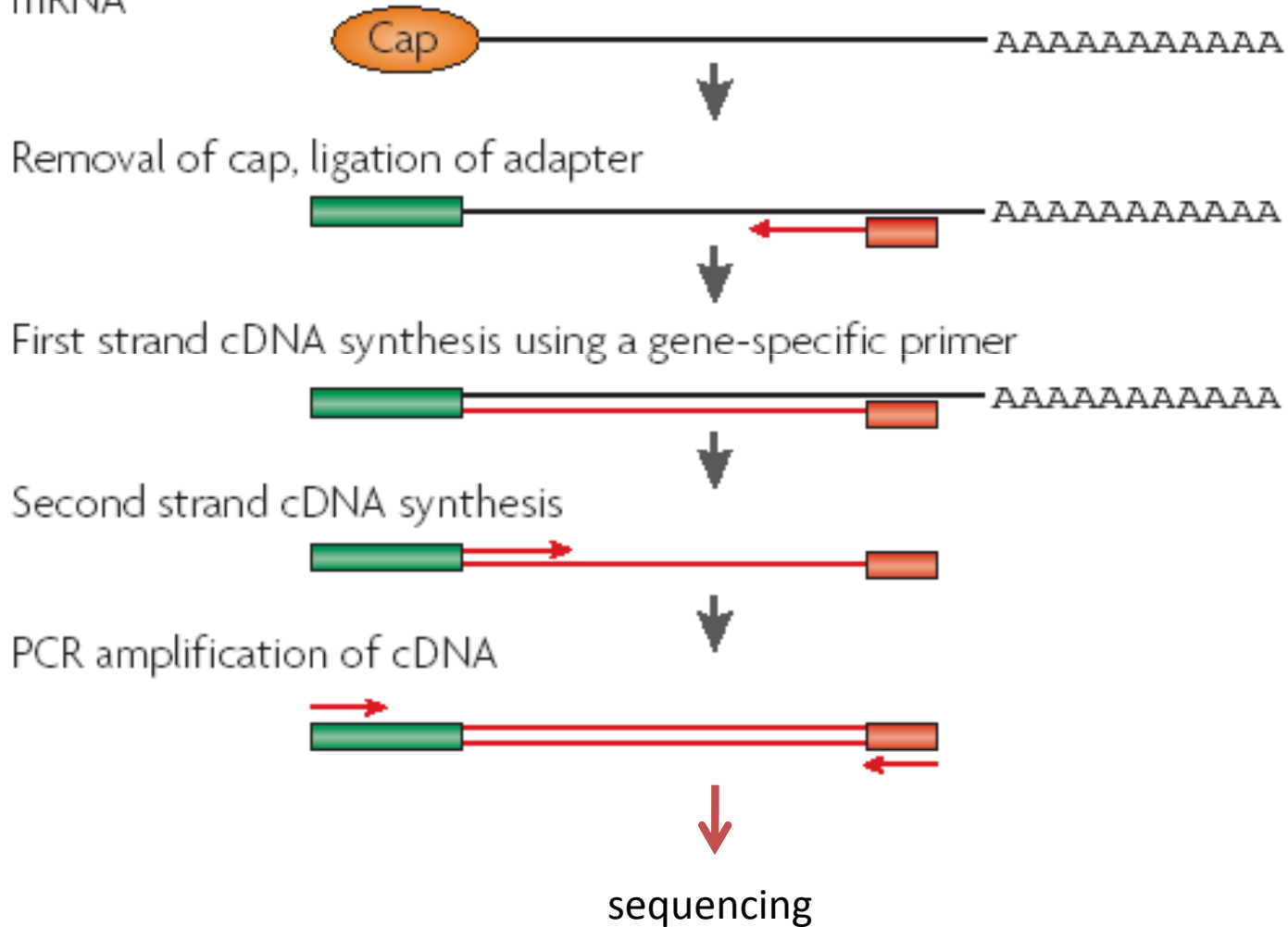
S1 nuclease mapping
(RNase protection quite similar)

Primer extension

Single-stranded DNA is evaluated by polyacrylamide gel electrophoresis

RACE= rapid amplification of cDNA ends

RACE
mRNA



To mapping TSS in high-throughput, tiling arrays have been used

Tiling arrays

mRNA



First strand cDNA synthesis



Second strand cDNA synthesis



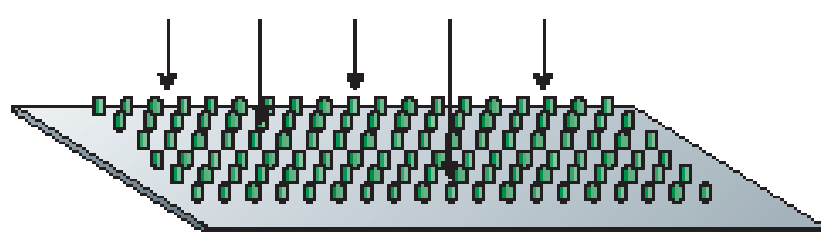
Fragmentation using DNase I



Labelling of termini



Hybridization to tiling array



But in general, **sequencing methods** are much better for **precise** definition of TSS

(... provided a method for CAP selection)

5' tag sequencing

mRNA



First strand cDNA synthesis



Full-length cDNA sequencing using the cap-trapper method



Ligation of linker I



Second strand cDNA synthesis



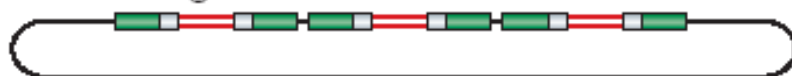
Digestion with *MmeI*



Ligation of linker II



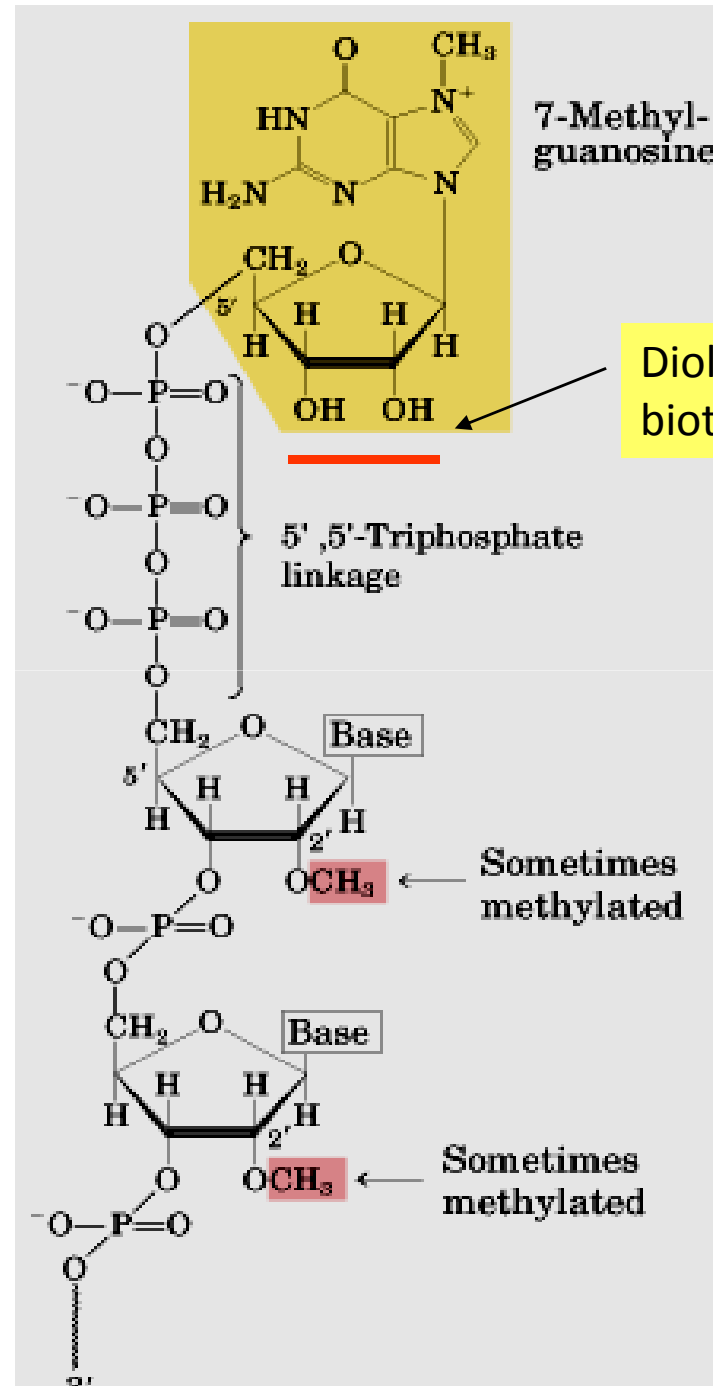
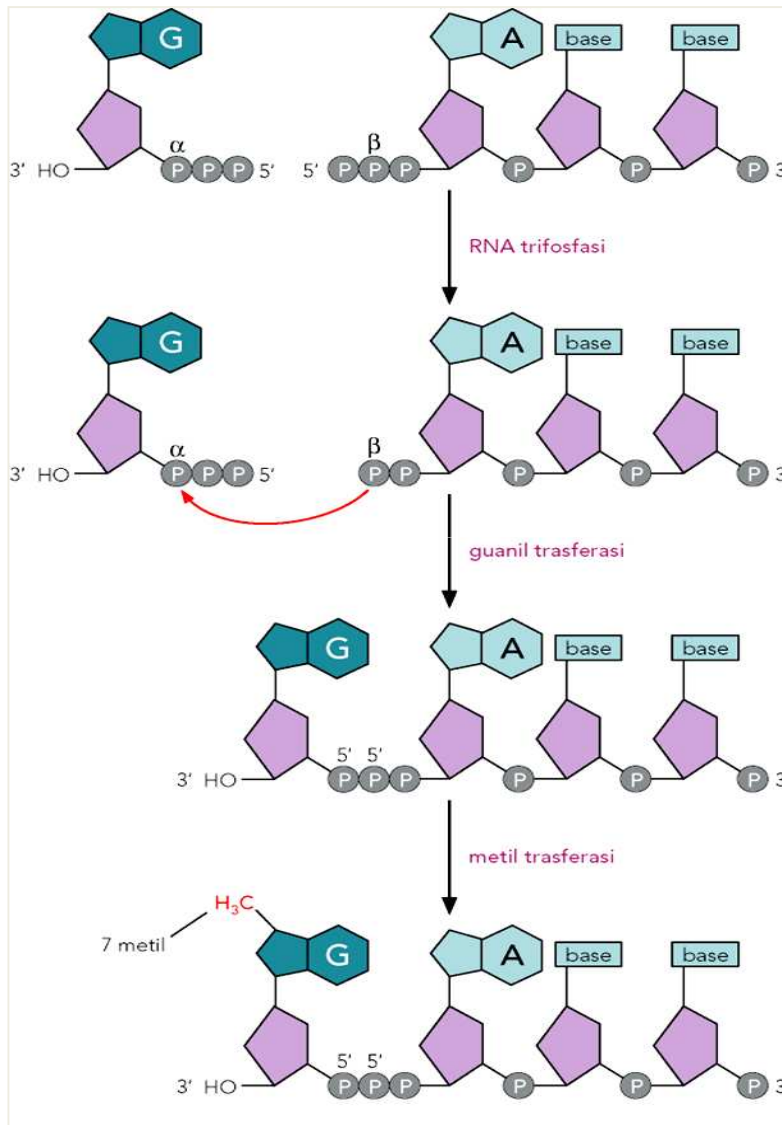
Concatenation and cloning



A Sepharose-conjugated CAP-binding protein affinity chromatography has been used to isolate "capped" RNAs.

(low yield)

Biotinylation of "cap" is better to allow selection




Genome-wide analysis of mammalian promoter architecture and evolution

Piero Carninci^{1,2,21}, Albin Sandelin^{1,3,21}, Boris Lenhard^{1,3,20,21}, Shintaro Katayama¹, Kazuro Shimokawa¹, Jasmina Ponjavic^{1,20}, Colin A M Semple^{1,4}, Martin S Taylor^{1,5}, Pär G Engström³, Martin C Frith^{1,6}, Alistair R R Forrest⁶, Wynand B Alkema³, Sin Lam Tan⁷, Charles Plessy², Rimantas Kodzius^{1,2}, Timothy Ravasi^{1,6,8}, Takeya Kasukawa^{1,9}, Shiro Fukuda¹, Mutsumi Kanamori-Katayama¹, Yayoi Kitazume¹, Hideya Kawaji^{1,9}, Chikatoshi Kai¹, Mari Nakamura¹, Hideaki Konno¹, Kenji Nakano^{1,9}, Salim Mottagui-Tabar^{3,20}, Peter Arner¹⁰, Alessandra Chesi¹¹, Stefano Gustincich¹¹, Francesca Persichetti¹², Harukazu Suzuki¹, Sean M Grimmond⁶, Christine A Wells¹⁹, Valerio Orlando¹³, Claes Wahlestedt^{3,20}, Edison T Liu¹⁴, Matthias Harbers¹⁵, Jun Kawai^{1,2}, Vladimir B Bajic^{1,7,16}, David A Hume^{1,6,21} & Yoshihide Hayashizaki^{1,2,17,18}


Mammalian promoters can be separated into two classes, conserved TATA box–enriched promoters, which initiate at a well-defined site, and more plastic, broad and evolvable CpG-rich promoters. We have sequenced tags corresponding to several hundred thousand transcription start sites (TSSs) in the mouse and human genomes, allowing precise analysis of the sequence architecture and evolution of distinct promoter classes. Different tissues and families of genes differentially use distinct types of promoters. Our tagging methods allow quantitative analysis of promoter usage in different tissues and show that differentially regulated alternative TSSs are a common feature in protein-coding genes and commonly generate alternative N termini. Among the TSSs, we identified new start sites associated with the majority of exons and with 3' UTRs. These data permit genome-scale identification of tissue-specific promoters and analysis of the *cis*-acting elements associated with them.

146 mouse cDNA libraries
41 human cDNA libraries


!



Bt-cap  —————< AAAAAAAAAA

↓ RT

 —————< AAAAAAAAAA

↓ RNase, ss-specific →

 —————< AAAAAAAAAA

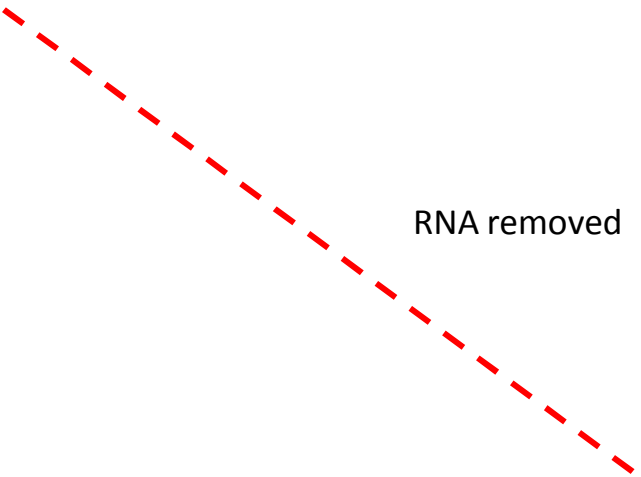
  —————< AAAAAAAAAA

cDNA not fully extended leave some ssRNA

↓

Streptavidin-sepharose selection

RNA removed



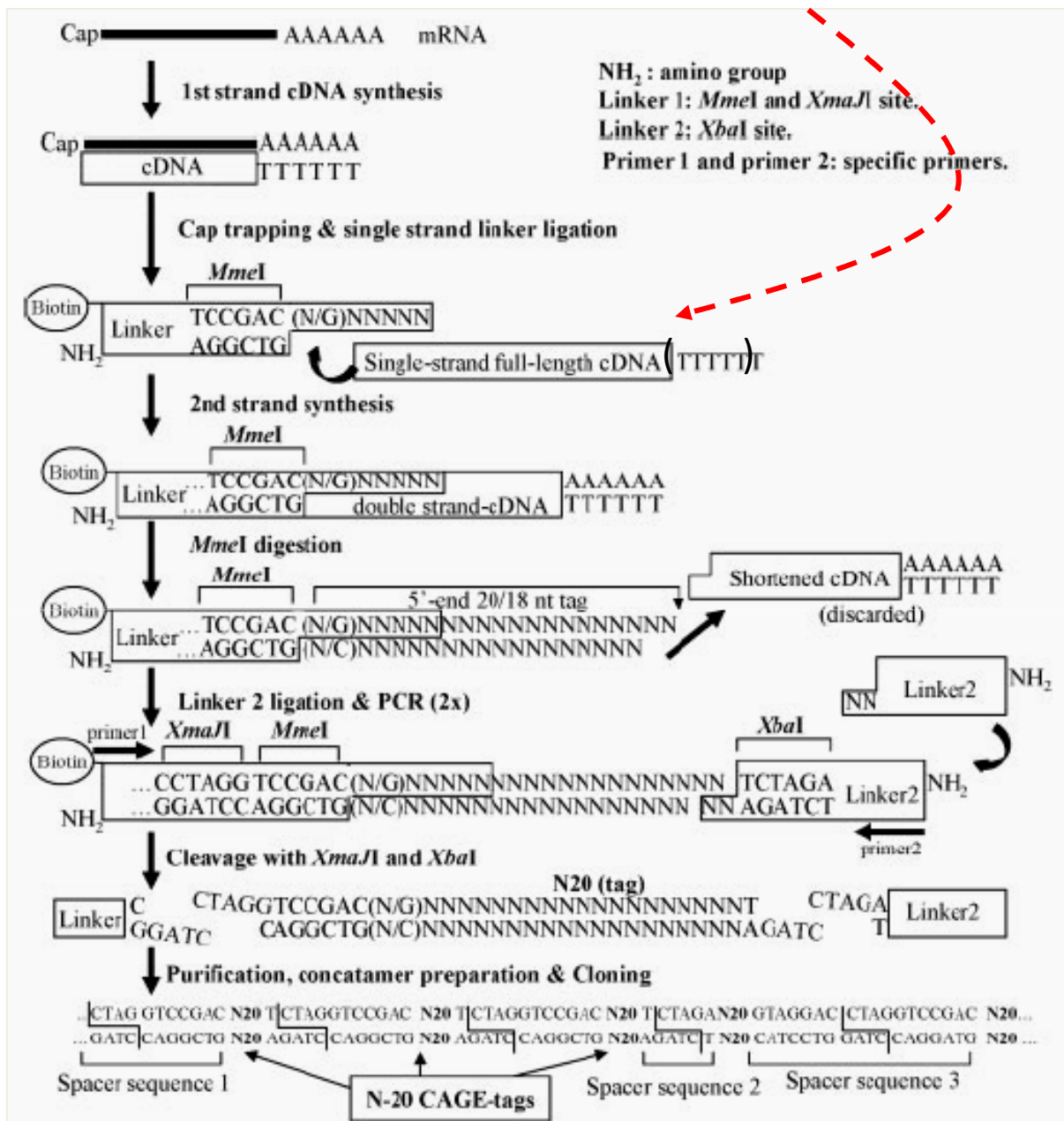


Fig. 1. Schematic procedure of the CAGE protocol as detailed in *Methods*.
from: Shiraki et al. (2003) PNAS 100:15776

Exactly as in the case of SAGE, CAGE produce a table of frequencies for all the 21-mers, from each library sequenced, that are subsequently mapped to the genome:

CAGE Tag	No./total	chr	position
ATTCGTCCAATCCAATCTCGG	123	chr6	23456444-23456465
TTAGGGCATGCTTGCGGCGA	3	chr21	10111578-10111556
ATCAACTCCTCTTCGTCATCG	987	chr8	9876101-9876122
etc.....			

And a table is generated correlating for each CAGE TAG its frequency in different cDNA libraries from different tissues:

	lung	gut	eye	breast	liver	muscle	brain
Tag 1	123	111	2	234	12	14	987
Tag 2	244	12	213	749	22	79	45
Tag 2	1	76	199	32	7	95	265
ETC....							

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.

The most extensive core promoter identification study undertaken so far used CAGE tags to identify 184,379 human and 177,349 mouse core promoters, many of which might contain a cluster of individual TSSs²⁴. A previous analysis that involved full-length cDNA sequencing identified 30,964 human and 19,023 mouse promoters⁴². But even the most recent figures are likely to be a substantial underestimate. First, sequencing 50–100,000 tags in each library can reliably detect only those transcripts that are expressed at a level of at least 10 copies in each cell (as there are at least 400,000 mRNAs in an average mammalian cell⁴³). Many transcripts are not present at this level, either because they are of low abundance in individual cells or are expressed in only a small subset of cells in the tissues that have been studied.

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 tiny amounts of DNA may lead to false positive.

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: see pdf attached to the lecture

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.

What is a gene, post-ENCODE? History and updated definition

Mark B. Gerstein,^{1,2,3,9} Can Bruce,^{2,4} Joel S. Rozowsky,² Deyou Zheng,² Jiang Du,³ Jan O. Korbel,^{2,5} Olof Emanuelsson,⁶ Zhengdong D. Zhang,² Sherman Weissman,⁷ and Michael Snyder^{2,8}

¹Program in Computational Biology & Bioinformatics, Yale University, New Haven, Connecticut 06511, USA; ²Molecular Biophysics & Biochemistry Department, Yale University, New Haven, Connecticut 06511, USA; ³Computer Science Department, Yale University, New Haven, Connecticut 06511, USA; ⁴Center for Medical Informatics, Yale University, New Haven, Connecticut 06511, USA; ⁵European Molecular Biology Laboratory, 69117 Heidelberg, Germany; ⁶Stockholm Bioinformatics Center, Albanova University Center, Stockholm University, SE-10691 Stockholm, Sweden; ⁷Genetics Department, Yale University, New Haven, Connecticut 06511, USA; ⁸Molecular, Cellular, & Developmental Biology Department, Yale University, New Haven, Connecticut 06511, USA

While sequencing of the human genome surprised us with how many protein-coding genes there are, it did not fundamentally change our perspective on what a gene is. In contrast, the complex patterns of dispersed regulation and pervasive transcription uncovered by the ENCODE project, together with non-genic conservation and the abundance of noncoding RNA genes, have challenged the notion of the gene. To illustrate this, we review the evolution of operational definitions of a gene over the past century—from the abstract elements of heredity of Mendel and Morgan to the present-day ORFs enumerated in the sequence databanks. We then summarize the current ENCODE findings and provide a computational metaphor for the complexity. Finally, we propose a tentative update to the definition of a gene: A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products. Our definition sidesteps the complexities of regulation and transcription by removing the former altogether from the definition and arguing that final, functional gene products (rather than intermediate transcripts) should be used to group together entities associated with a single gene. It also manifests how integral the concept of biological function is in defining genes.

Figura 5.10 Risultato di un tipico sistema di annotazione del genoma.

L'esempio mostrato si riferisce all'annotazione di un segmento di 15 kb del genoma umano contenente il gene per il fattore tissutale utilizzando il browser Genotator. Dall'alto le analisi sono: localizzazioni di possibili promotori (elementi regolativi a monte); sequenze corrispondenti a proteine nella banca dati GenPept; sequenze corrispondenti a EST note (si veda la Sezione 3.3.3); localizzazione delle sequenze ripetitive umane note; predizione degli esoni in base ai programmi Genscan, Genefinder, GRAIL e Genie; sequenze corrispondenti a geni noti nella banca dati GenBank; ORF in ciascuno dei tre schemi di lettura. Sotto alla linea nera le analisi sono ripetute per il filamento complementare nella direzione inversa. L'annotazione rivela la posizione di cinque possibili esoni, come indicato dalle frecce nere in alto. Immagine fornita cortesemente da Nomi Harris.

