Riassunto puntate precedenti

- Expression Microarrays
- Tiling genomic microarrays
- Sequencing methods

RNA transcripts

Depend on 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 kind of tissue:

Origin
Stage of development
In vitro culture conditions
Pathological status

Individuals
Age
…..

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.

review

VOLUME 10 | JANUARY 2009 | 57

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 transciption 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)

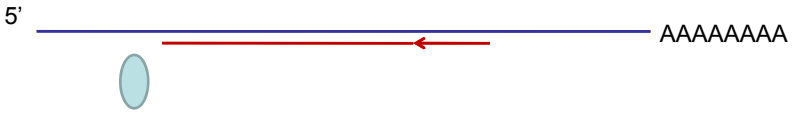
Conclusions: there is persavive transcription and the majority of RNAs do not show protein-coding evidence

Il problema della corretta mappatura delle T.U. Definizione del 5' e 3'.

One problem is that EST, SAGE and other serial methods do not define the TSS (transcription start site)

RNA databases 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.

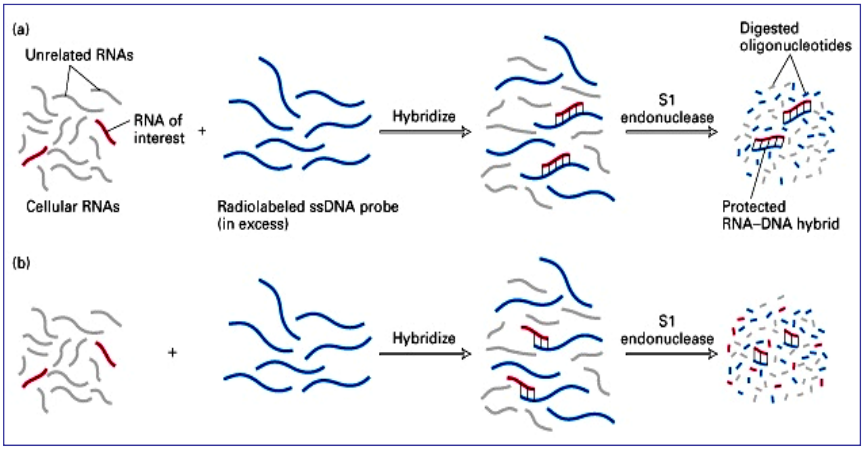


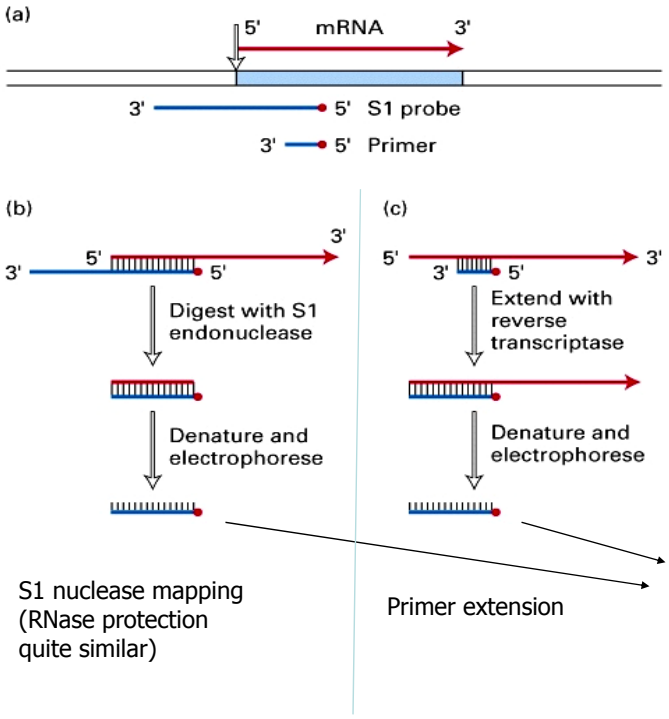
Seq-methods are more precise, but often a pletora of different transcripts are seen at the 5' and 3' ends of a T.U.

Another common problem:
How to distinguish "transcripts" from RNA fragments or processed RNAs ?

Several methods can be used to map the TSS One gene at a time …

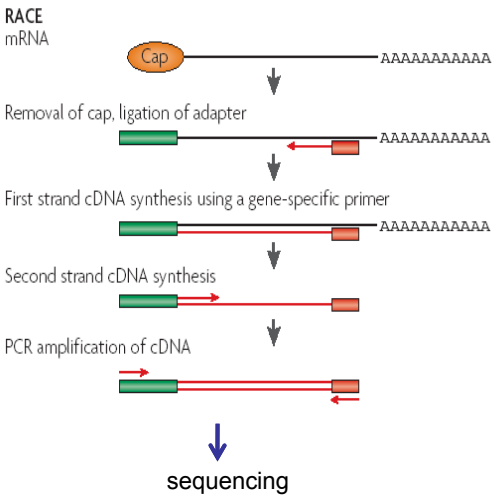
Most common method for mapping, for individual genes, the transcriptional initiation site is S1 Nuclease (a single-strand specific endonuclease) analysis, based on the hybridization of genomic DNA with mRNA:



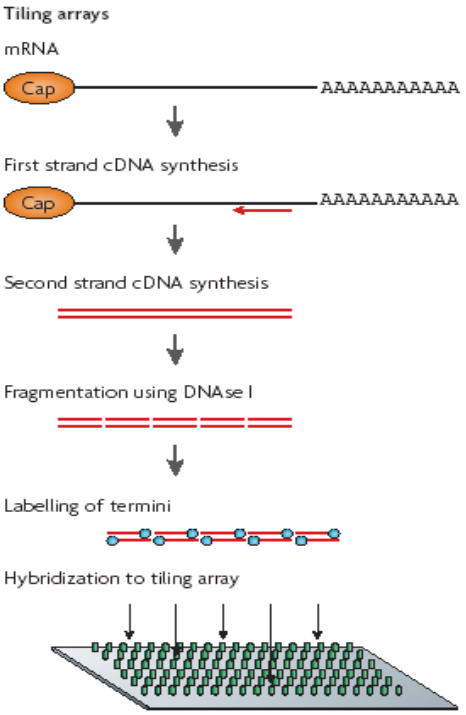


Primer extension is an alternative mapping method

RACE= rapid amplification of cDNA ends

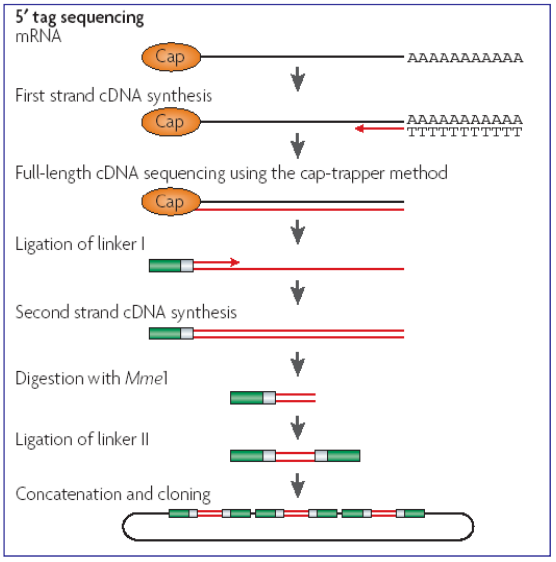


To mapping TSS in high-throughput, tiling arrays have been used



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

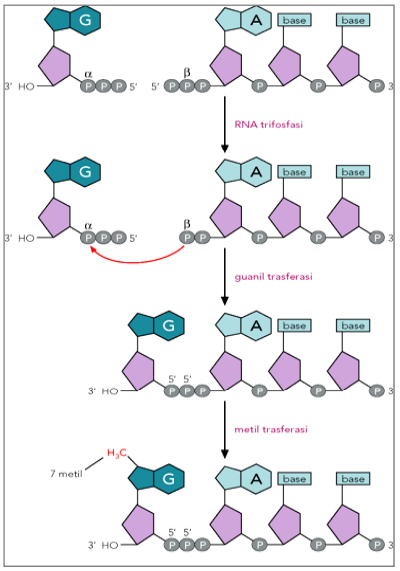
(…. provided a method for CAP selection)

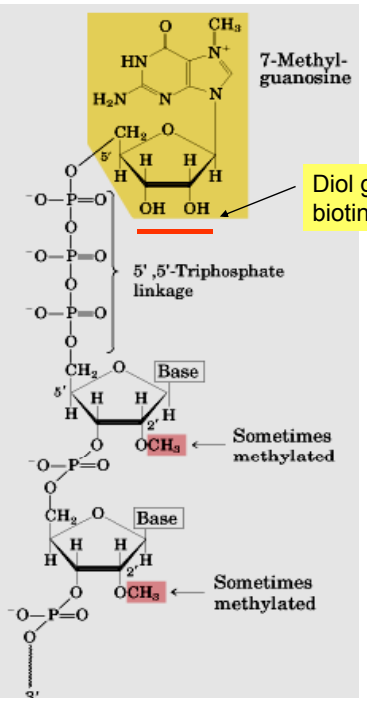


A Sepharose-conjugated CAP-binding protein affinity chromatography has been used to isolate “capped” RNAs.

This approach gives **low yield** of capped RNAs

Biotinylation of “cap” is better to allow efficient 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 Katayama[1], Kazuro Shimokawa[1], Jasmina Ponjavic[1,20], Colin A M Semple[1,4], Martin S Taylor[1,5], Pär G Engström[3], Martin C Frith[1,6], Alistair R R Forrest[6], Wynand B Alkema[3], Sin Lam Tan[7], Charles Plessy[2], Rimantas Kodzius[1,2], Timothy Ravasi[1,6,8], Takeya Kasukawa[1,9], Shiro Fukuda[1], Mutsumi Kanamori-Katayama[1], Yayoi Kitazume[1], Hideya Kawaji[1,9], Chikatoshi Kai[1], Mari Nakamura[1], Hideaki Konno[1], Kenji Nakano[1,9], Salim Mottagui-Tabar[3,20], Peter Arner[10], Alessandra Chesi[11], Stefano Gustincich[11], Francesca Persichetti[12], Harukazu Suzuki[1], Sean M Grimmond[6], Christine A Wells[19], Valerio Orlando[13], Claes Wahlestedt[3,20], Edison T Liu[14], Matthias Harbers[15], Jun Kawai[1,2], Vladimir B Bajic[1,7,16], David A Hume[1,6,21] & Yoshihide Hayashizaki[1,2,17,18]

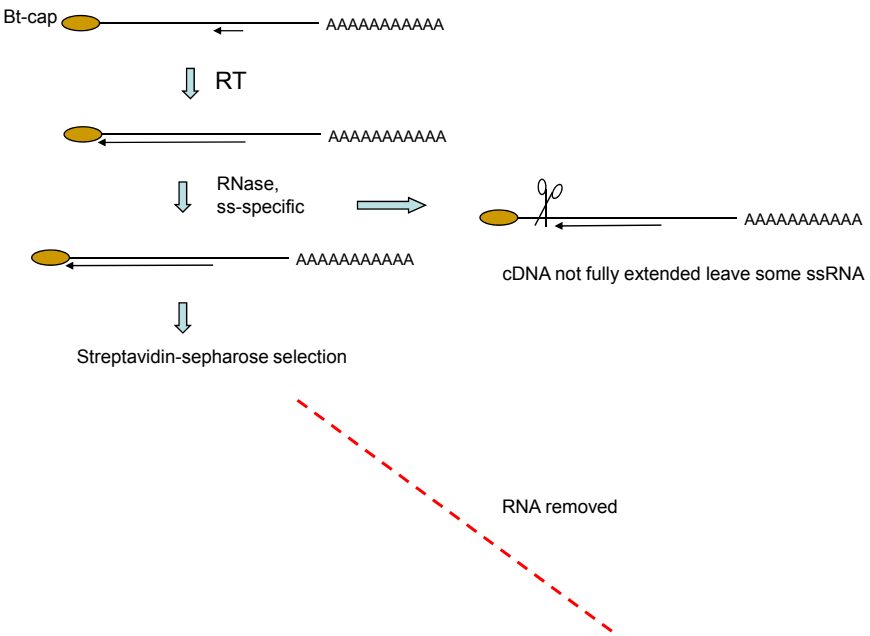
Mammalian promoters can be separated into two classes, conserved TATA box–enriched promoters, which initiate at a well-defined site, and more plastic, broad and evolvable CpG-rich promoters. We have sequenced tags corresponding to several hundred thousand transcription start sites (TSSs) in the mouse 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.

626 VOLUME 38 | NUMBER 6 | JUNE 2006 NATURE GENETICS

146 mouse cDNA libraries

41 human cDNA libraries

!



Shiraki et al. (2003) PNAS 100:15776

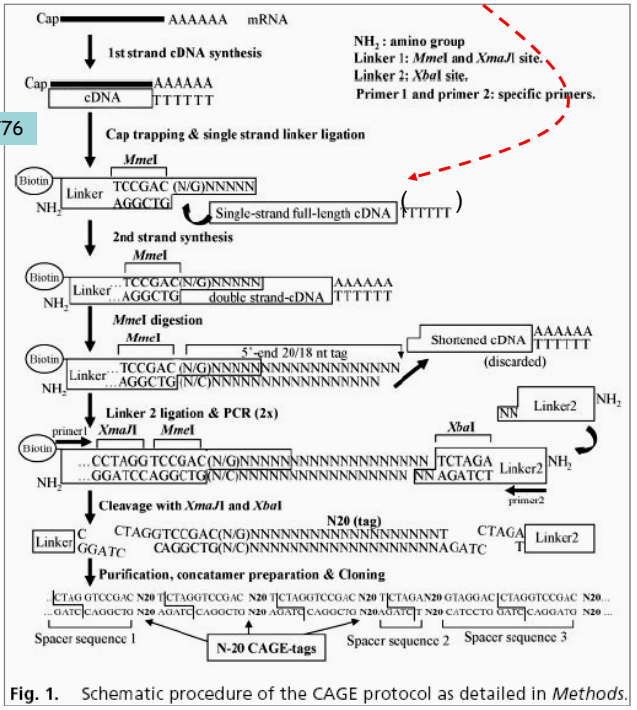


Fig. 1. Schematic procedure of the CAGE protocol as detailed in *Methods*.

Exactly as in the case of SAGE, CAGE produce a table of frequencies for all the 21-mers, fro 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

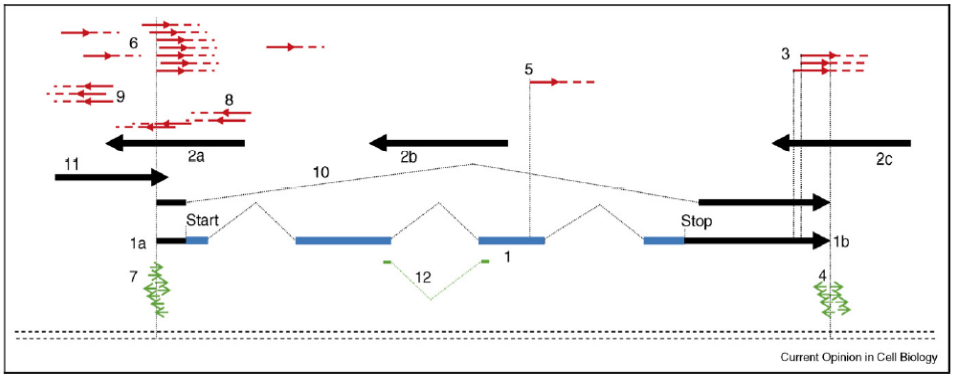
ETC....

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.

SCIENCE VOL 309 2 SEPTEMBER 2005



Multiple RNAs from a single locus. Image of a hypothetical gene with a noncomprehensive map of noncoding RNAs surrounding the main mRNA. CAGE tags (short stretches encompassing the first 20 nt of a capped RNA) are depicted in red and do not identify the end of the transcript; see text for details. (1) Main protein-coding mRNA transcript with start (1a) and termination sites (1b) and CDS highlighted in blue; (2a–c) antisense RNAs in head-to-head (2a), full overlap (2b), and tail-to-tail (2c) overlap; (3) CAGE tags identify transcript in the 3′ UTRs; (4) TASRs; (5) transcript from esonic promoters; (6) CAGE tags identifying TSS (exact location can vary) and may overlap PALRs; (7) PASRs; (8) antisense transcription events detected by CAGE from the first introns or first exon; (9) RNA from bidirectional promoters; (10) ncRNA overlapping starting and termination sites of the coding mRNA; (11) PALRs. Not included are other phenomena, like miRNA produced from introns and acting elsewhere *in trans*, and miRNA produced from other loci and acting *in trans* on the 3′ UTRs of the mRNAs.

Diversi lavori pubblicati nel 2009 con RNA-Seq. Su lievito e mammiferi (anche molti procarioti)

MOLECULAR BIOLOGY

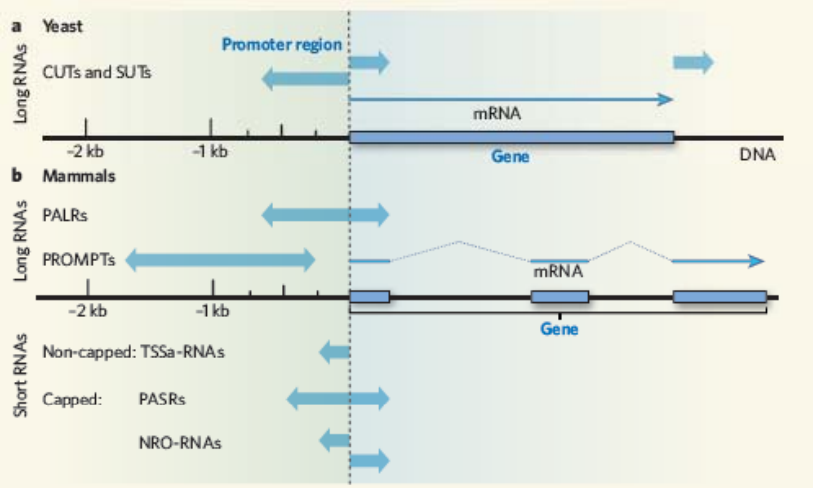
The long and short of RNAs

Piero Carninci

The known world of RNA is expanding faster than that of any other cellular building block. The latest additions are types of long and short non-coding RNAs formed by bidirectional transcription and unusual processing.

NATURE|Vol 457|19 February 2009

Comments on 6 papers published at the beginning of 2009



Un lavoro pubblicato (il primo) su cellule umane, con RNA-Seq (Illumina-Solexa)

Limitato però a poly(A)+ RNA

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.

Supporto 1