



Serial Analysis of Gene Expression

*Cloning of Tissue-Specific Genes Using SAGE
and a Novel Computational Subtraction
Approach. Genomic (2001)*

Hung-Jui Shih



Outline of Presentation

- SAGE
- EST
- Article
 - TPE algorithm
 - Results & Conclusion
- Reference



Serial Analysis of Gene Expression (SAGE)

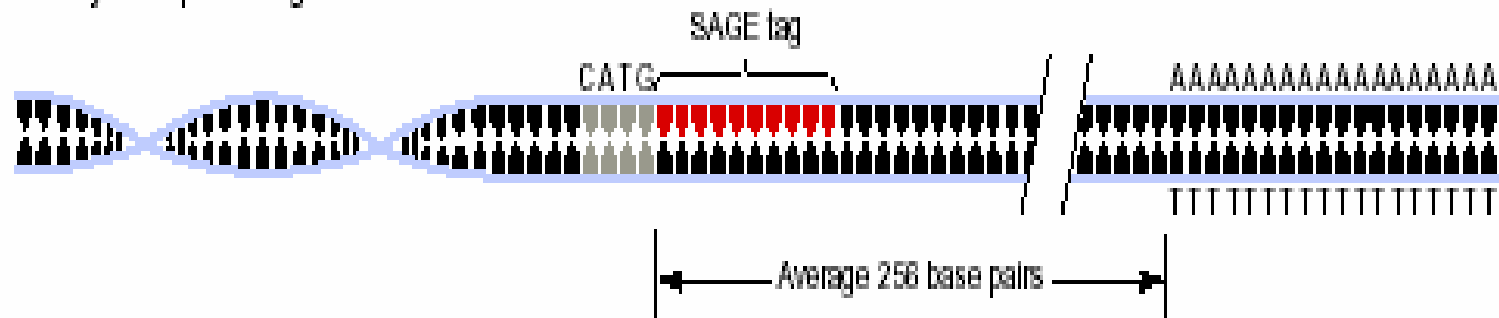
- Developed in 1995 (Velculescu et al. *Science* 270 (5235) : 484-487)
- Allows the quantitative and simultaneous analysis of a large number of transcripts.
- Identify novel expressed genes.

Principles of SAGE

- 1. Short sequence tag contains sufficient information to uniquely identify a transcript

SAGE principle 1

A short oligonucleotide sequence from a defined location within a transcript, a 'tag', encodes sufficient complexity to identify an expressed gene.

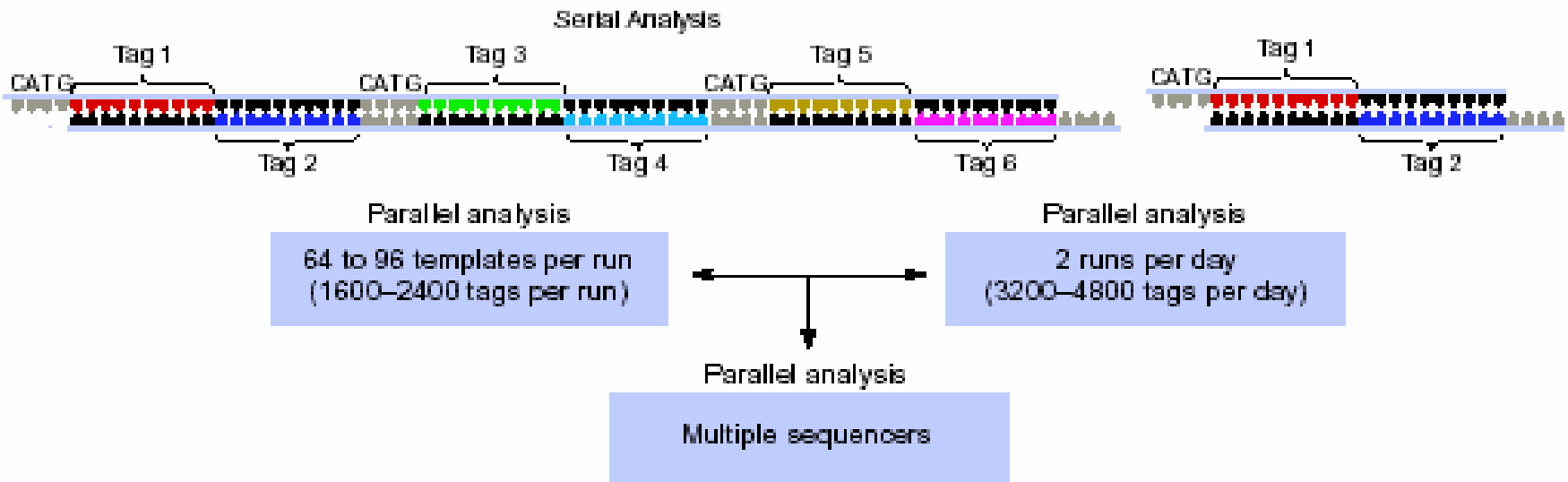


Principles of SAGE

- 2. Sequence tags can be linked together to form long serial molecules that can be cloned and sequenced.

(b) SAGE principle 2

A combination of serial and parallel analysis maximizes throughput.



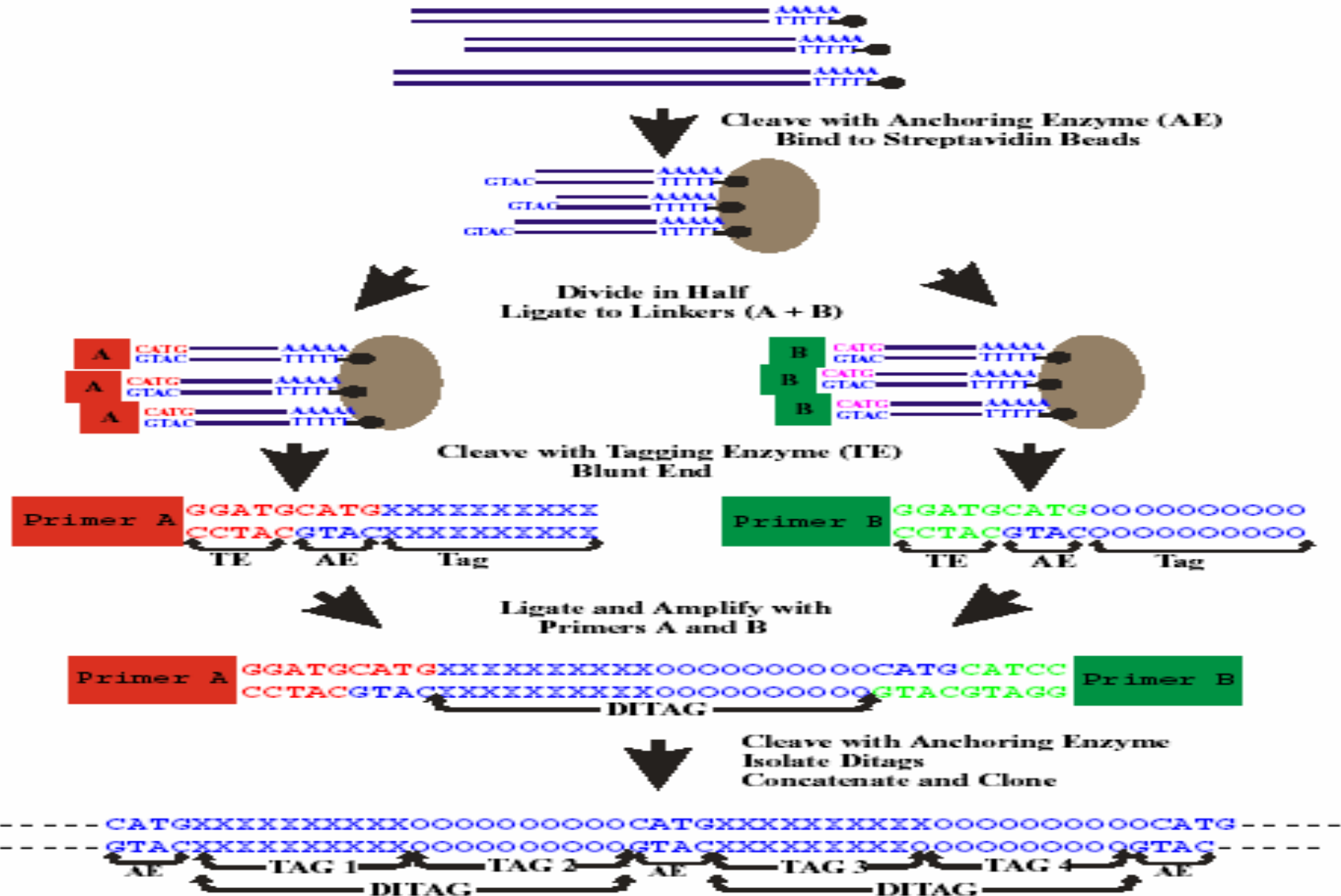
Principles of SAGE

- 3. Quantization of the number of times a particular tag observed provides the expression level of the corresponding transcript.

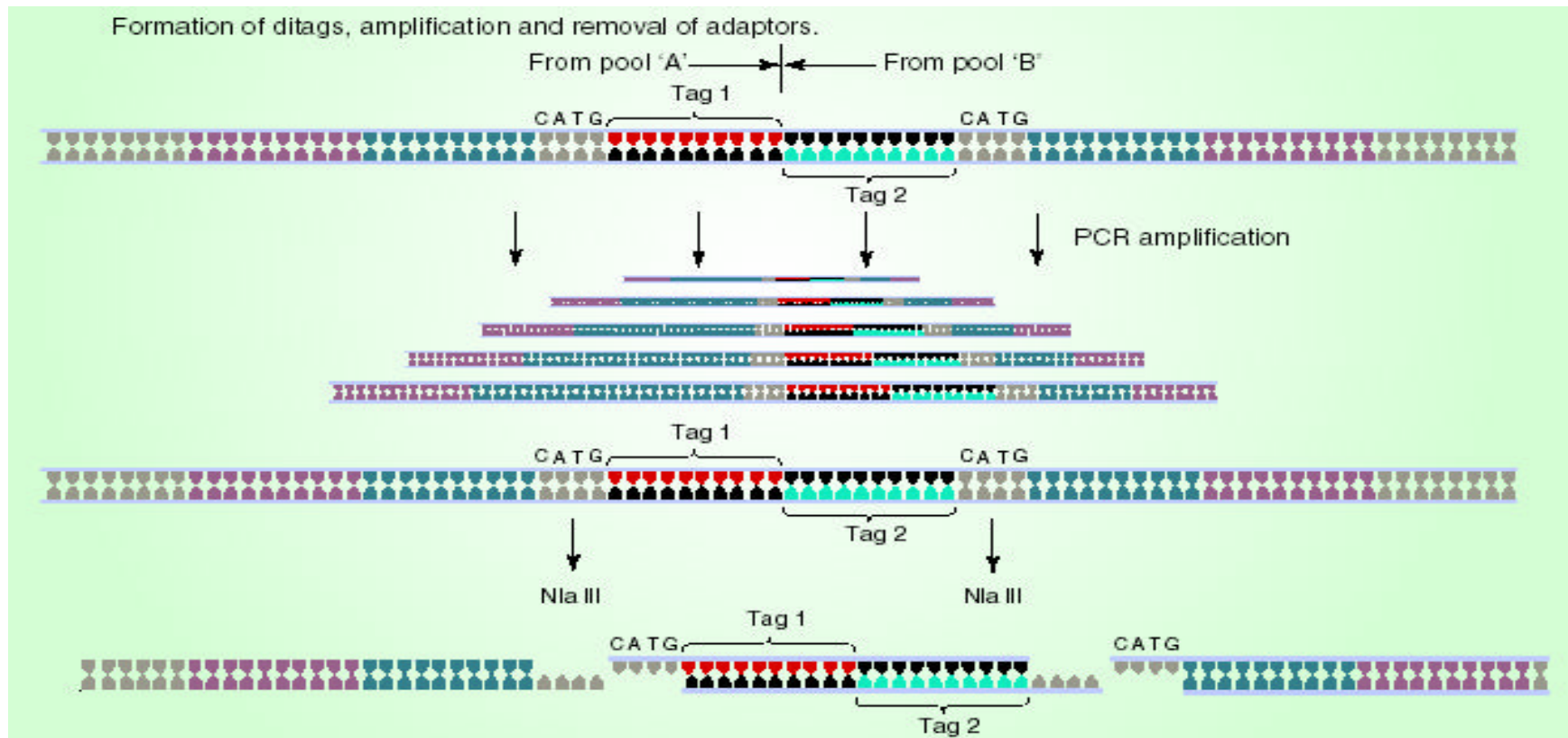
<u>SAGE tag</u>	<u>Tag Count (in 100 000 tags)</u>	<u>Absolute abundance</u>
<u>CATGGACGTCTTAAT</u>	33 TAGS	0.033%
<u>CATGGTGACCTCCTT</u>	63 TAGS	0.063%
<u>CATGTGAAGAGAAGA</u>	22 TAGS	0.022%
<u>CATGAGTGGAGGTGG</u>	9 TAGS	0.009%

NlaIII site

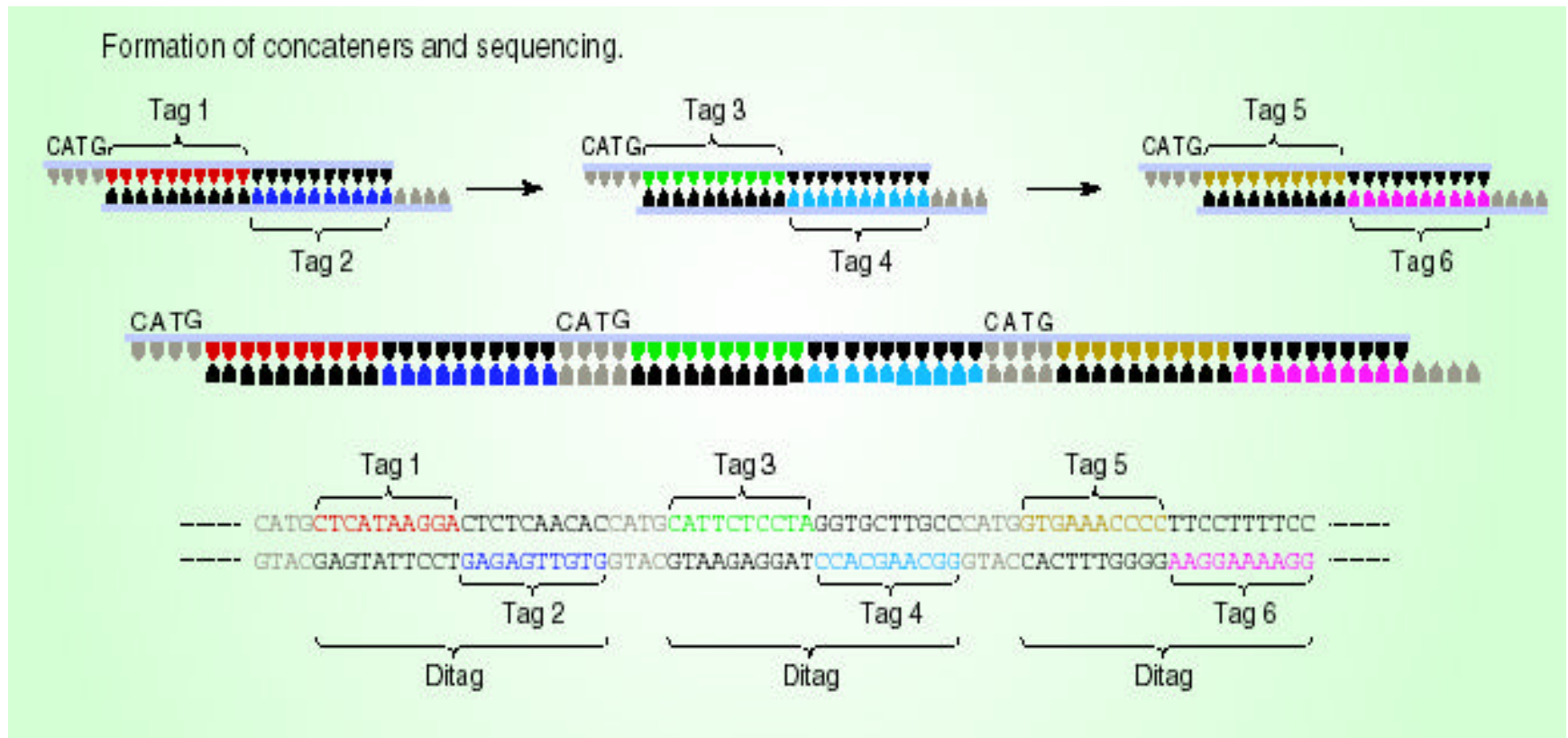
Serial Analysis of Gene Expression (SAGE)



Serial Analysis of Gene Expression (SAGE)



Serial Analysis of Gene Expression (SAGE)





Function of SAGE Software

- Identifies enzyme sites with proper spacing
- Extracts tags
- Record tags in database
- Match tags to genome sequence



Advantages of SAGE

- Highly sensitive - scaleable.
- Detects all genes including unknowns - quantitative data.
- Avoids amplification bias.
- Immortalized data allows for multiple comparisons.
- Circumvented unwanted cross-hybridization

New in SAGE

- LongSAGE

- Characterizes a 21-base pair segment
- Saha S. et al. *Using the transcriptome to annotate the genome (2002). 20(5), 508-512*

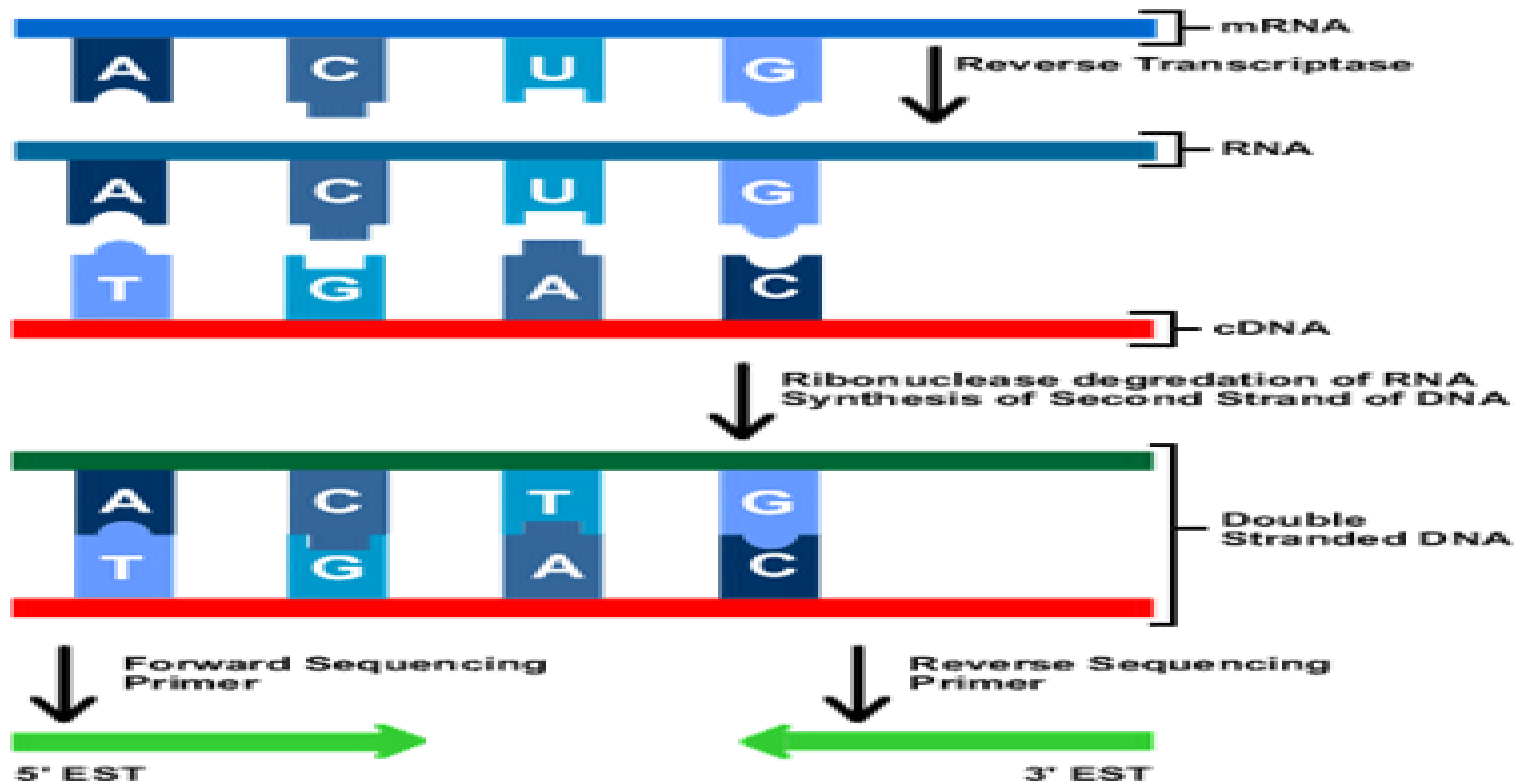
- SAGE database

- I-SAGE™ kit (Invitrogen)

Expressed Sequence Tag (EST)

- Developed in 1991. (Adams M. et al. Science 252:1651-1656)
- A small part of the active part of a gene, made from cDNA, 150-400 bps, a kind of STS (sequence tagged site).
- Can be used to fish the rest of the gene out of the chromosome, by matching base pairs with part of the gene.
- Can be radioactively labeled in order to locate it in a larger segment of DNA.

Expressed Sequence Tag (EST)





Practical Advantages of EST

- Sequences can be generated rapidly and inexpensively
- Only one sequencing experiment is needed per each cDNA generated
- Do not have to be checked for sequencing errors as mistakes.
- Do not prevent identification of the gene from which the EST was derived.



Challenge of EST

- Sequence is with errors including base insertion and deletion.
- Genes with lower expression are not easier to be picked which may have significant functions. It can be solved by...
 - Normalization
 - With large amount of EST

Summary of Characteristics

■ SAGE


- Highly sensitive, efficient , comparability, technically demanding and needs a sequenced genome

■ EST

- Sequence rapidly, inexpensive and business valuable (ex: Merck-EST project, (1998) *Bioinformatics* 14(1):2-13) .

■ Microarray

- Large-Scale, sensitive, technically challenging, limited to known genes and expensive



Cloning of Tissue-Specific Genes

■ Rational

- Defect in expressed genes results in clinical phenotypes.

■ Objective

- Identify functionally specialized genes.

■ Material & Method

- No-match tags
- SAGE
- EST
- TPE algorithm



Tissue Preferential Expression (TPE)

- Based on the number of tissues in which a tag is present (The range of expression) and its expression level in the tissue of interest compared with other tissues (The preferential abundance).
- Calculate and plot scores
- Achieved as the Euclidean distance.

The preferential abundance

$$\text{Ratio}_j(\text{tag}_i) = \log[(0.001 + M(\text{tag}_i)) / (0.001 + N_j(\text{tag}_i))]$$

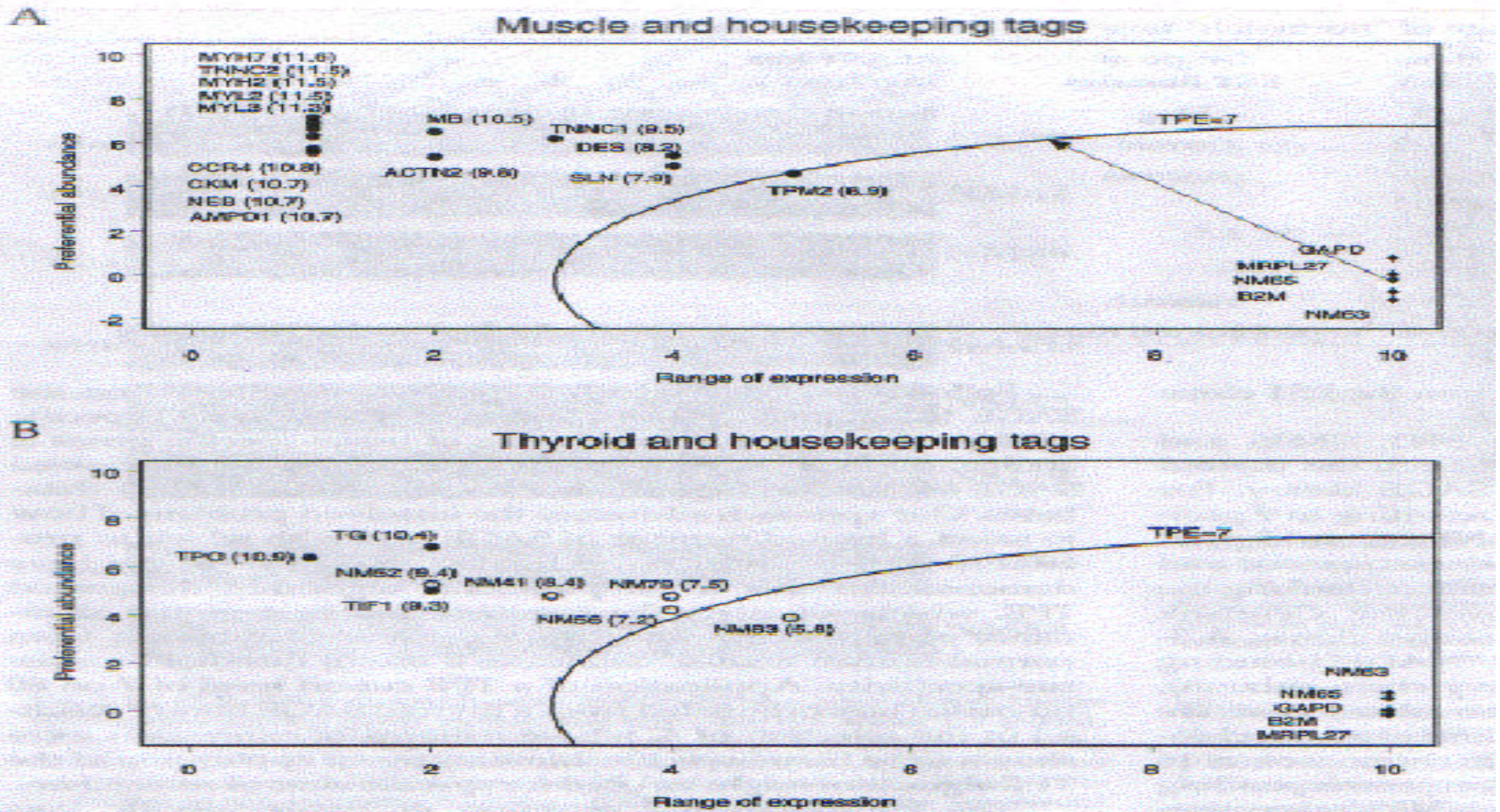
- M: tag count in the tissue of interest
- N_j: tag count in another tissue j
- Add 0.001 to each tag count: prevent division by 0 or taking the logarithm of 0

Results

TABLE 1. Tags corresponding to thyroid-specific genes, "no-match" tags, and "housekeeping" genes

Tissue		Thyroid	Breast	Ovary	Vascular	Prostate	Cerebellum	Brain	Colon	Fibroblast	Muscle	TPE
Total no. of tags	762188	10994	49762 ^a	48925	111588 ^a	66687	30774	212699 ^a	100730 ^a	22301	107728 ^a	
Thyroid-specific genes	TPO	2400	-	-	-	-	-	-	-	-	-	10.89
	TG	21000	-	-	-	-	32	-	-	-	-	10.36
	TTF1	300	-	-	-	-	-	12	-	-	-	9.32
No-match tags TPE ≥ 7	NM52	454	-	-	-	-	-	3	-	-	-	9.45
	NM41	454	40	-	-	-	-	24	-	-	-	8.36
	NM79	454	-	-	-	-	-	7	-	-	9	7.53
	NM56	454	-	-	-	14	-	-	89	-	9	7.19
No-match tags TPE ≤ 7	NM83	888	-	20	-	-	194	47	-	-	19	5.80
	NM63	454	60	61	74	134	97	509	89	89	9	0.51
	NM65	454	180	122	103	119	389	272	29	134	92	0.30
Housekeeping genes	GAPD	2070	542	6295	2737	1799	1722	3020	694	313	9161	0.17
	B2M	1182	2753	8032	1825	1829	194	314	2563	1883	241	0.25
	MRP17	454	1085	1941	1430	1229	162	211	108	2152	658	0.30

Results



- TPE levels ≥ 7 are considered indicative for tissue specificity.

Results

TABLE 2. Linkage of "no-match" tags with EST and GenBank sequences

No-match tag	Tag sequence	EST clone (Acc. no.)	Origin of EST libraries	BLAST hits (Acc. no.)
41	ccagetgct	AI37514	lung	human chromosome 16 clone 165E7 (AC007011)
52	ttgggagta	AA632629	thyroid	
56	ctgttggtg	W60005	pancreas	mouse NADPH-dependent oxidase (MMU43384) pig NADPH-dependent oxidase (SSU02476) human NADPH-dependent oxidase (AF127763) human chromosome 15 clone (AC009700)
79	ggaatgctc	A.446209	stomach	
83	cagtgaaaa	A.023948	parathyroid tumor	human chromosome 1p35 clone 462023 (HS462023)



Conclusion

- Computational subtraction of SAGE tags by the proposed TPE algorithm is a rapid and reliable way to expedite the cloning of tissue-specific genes.

Reference

- Patino W. et al (2002) *Serial Analysis of Gene Expression: Technical Consideration and Application to Cardiovascular Biology*. *Circ Res* 91:565-569
- Velculescu, Zhang, Vogelstein & Kinzler. (1995) *Serial Analysis of Gene Expression*. *Science* 270 (5235) : 484-487
- Adams M. et al (1991) *Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project*. *Science* 252:1651-1656
- www.sagenet.org