techniques will need to be used in the study of embryogenesis, a probable source for many unidentified genes.

References

1 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
3 Hogenesch, J.B. *et al.* (2001) A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* 106, 413–415
4 Kapranov, P. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919
5 Boguski, M.S. *et al.* (1993) dbEST – database for 'expressed sequence tags'. *Nat. Genet.* 4, 332–333
6 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487
7 Saha, S. *et al.* (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* 20, 508–512
8 Ko, M.S. (2001) Embryogenomics: developmental biology meets genomics. *Trends Biotechnol.* 19, 511–518
9 Anisimov, S.V. *et al.* (2002) SAGE identification of gene transcripts with profiles unique to pluripotent mouse R1 embryonic stem cells. *Genomics* 79, 169–176
10 Anisimov, S.V. *et al.* (2002) SAGE identification of differentiation responsive genes in P19 embryonic cells induced to form cardiomyocytes *in vitro*. *Mech. Dev.* 202, 25–74
11 Stern, M.D. *et al*. Can transcriptome size be estimated from SAGE catalogs? Bioinformatics (in press)
12 Chen, J. *et al.* (2002) Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl Acad. Sci. U.S.A.* 99, 12257–12262
13 Chen, J.J. *et al.* (2000) Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *Proc. Natl Acad. Sci. U.S.A.* 97, 349–353

0167-7799/03/$ - see front matter
PII: S0167-7799(02)00031-8

Research Focus Response

Response: The new role of SAGE in gene discovery

San Ming Wang

Dept of Medicine, University of Chicago Medical Center, 5841 S. Maryland Ave., MC2115, Chicago, IL 60637, USA

It is interesting to note that the concept of 30 000 to 40 000 genes in the human genome is questioned by several groups of experimental scientists, because these lower estimates are not compatible with their experimental data [1–3] (and see previous article in this issue by Boheler and Stern).

The final proof for the correct number of genes in the human genome will rely on the physical identification of all genes. The current art of physical gene identification is through the analysis of the gene products, transcripts, using the expressed sequence tags (EST) approach [4,5]. Although the decade's performance of EST projects has identified >4.8 million human ESTs, the novel ESTs identified have decreased progressively. For example, the rate of novel sequence identification was 10.6% in 1996, 2.7% in 1998 and only 1.5% in 2002 [1]. It seems that most, if not all, transcripts expressed from the human genome might have been identified.

The serial analysis of gene expression (SAGE) technique provides an alternative tool for gene expression [6]. SAGE collects a short sequence tag (10 bases) from a transcript and concatemerizes multiple tags for a single sequencing reaction. This overcomes the problem of redundancy of transcripts and allows SAGE to be performed without subtraction or normalization, a precondition used by the EST approach [6]. More than 6.8 million SAGE tags have been generated from 171 libraries [7] (see http://www.ncbi.nlm.nih.gov/SAGE/). An interesting observation is that many SAGE tags have no match to known expressed sequences and the majority of these unmatched novel SAGE tags tend to be low copies [1]. What do these novel SAGE tags represent?

A widespread concept, even within the community that favours the use of SAGE, is that many tags are not reliable owing to experimental errors [7,8]. This concept is largely based on the assumption that a SAGE tag has only 10 bases. As the error rate of single-pass sequencing, including substitution, deletion and insertion, can reach nearly 10% per tag, the possible number of mutated tags derived from a wild-type tag can be enormous, particularly for the SAGE tags with low copy numbers. Thus, the low-copy tags should be excluded from the analysis.

However, our analyses show that the number of erroneous tags is far less than presently assumed. Our analysis of the SAGE database shows that the novel SAGE tags do not have a higher rate of mutation than the matched SAGE tags: our experimental data show that the error rate for SAGE tags is <2% per tag, much lower than the assumed error rates, and we confirmed that 67% of novel SAGE tags are derived from novel transcripts. Our data indicate that many novel transcripts remain to be identified considering the presence of millions of novel SAGE tags. These novel transcripts are not only present in some rare cell types but also in many common tissues that have been analyzed in the EST projects as revealed by our analysis of the SAGE data collected from these tissues. Our observations are consistent with those of others using different approaches. For example, an analysis of transcriptional activity in human chromosomes 21 and 22 using oligo-microarrays indicates that the number of transcribed sequences in the human genome could be an order of magnitude greater than current estimate [2]. The study of gene expression in human colorectal cancer cells using the longer SAGE approach also identified a significant number of potential novel genes [3].

Corresponding author: San Ming Wang (swang1@midway.uchicago.edu).

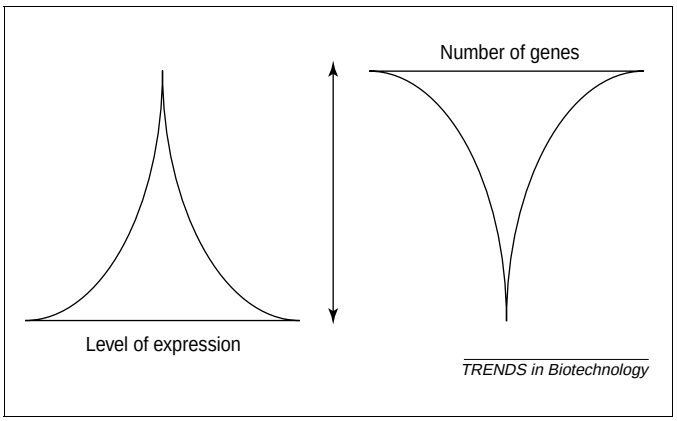

Fig. 1. Reverse relationship between the number of genes and the levels of expression. In eukaryotic genomes, a small number of genes are expressed at high levels and contribute to most of the transcripts whereas the majority of the genes are expressed at low levels and contribute to a small portion of the transcripts.

Owing to the low level of expression, the novel transcripts are difficult to identify using the EST approach but they have been detected as novel SAGE tags using the sensitive SAGE technique. The issue now is to isolate the full-length transcripts represented by the novel SAGE tags. The controversy for my argument, which I am sure will remain, will cease only when these full-length sequences are identified.

In the 'post-genomic' era, I propose that we switch the strategy of physical gene identification from the random sequencing-based EST approach to the novel SAGE tag-based approach to identify the human genes expressed at lower levels. It is in this area that SAGE is superior to the EST approach in providing high efficiency to identify these genes. It is important to remember that the majority of genes in the human genome are expressed at low levels [9] (Fig. 1).

References

1 Chen, J. *et al.* (2002) Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12257–12262
2 Kapranov, P. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919
3 Saha, S. *et al.* (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* 20, 508–512
4 Adams, M.D. *et al.* (1992) Sequence identification of 2,375 human brain genes. *Nature* 355, 632–634
5 Bonaldo, M.F. *et al.* (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* 6, 791–806
6 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487
7 Velculescu, V.E. *et al.* (1999) Analysis of human transcriptomes. *Nat. Genetics* 23, 387–388
8 Boon, K. *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. U. S. A.* 99, 11287–11292
9 Bishop, J.O. *et al.* (1974) Three abundance classes in HeLa cell messenger RNA. *Nature* 250, 199–204

0167-7799/03/$ - see front matter
PII: S0167-7799(02)00035-5