>3 hours drastically reduces the level of gene expression. Maximal reporter expression was seen after a single dose of red light, followed by a 3-hour incubation in the dark. However, as dramatic as the increase in reporter expression was (1000-fold over negative controls), it was still only about one-sixth of the level of endogenous GAL4 expression. Whether this translates to gene expression that reaches physiological levels remains to be seen.

Clearly, Shimizu-Sato and colleagues have developed a regulatable system that has the potential to raise the bar on inducible gene expression systems, almost bridging the gap between science fiction and reality. It is with great anticipation that we await the next step in the journey.

References

1 Lewandoski, M. (2001) Conditional control of gene expression in the mouse. *Nat. Rev. Genet.* 2, 743–755

2 Garcia, E.L. and Mills, A.A. (2002) Getting around lethality with inducible Cre-mediated excision. *Semin. Cell Dev. Biol.* 13, 151–158

3 Weber, W. *et al.* (2002) Macrolide-based transgene control in mammalian cells and mice. *Nat. Biotechnol.* 20, 901–907

4 Blau, H.M. and Rossi, F.M. (1999) Tet B or not tet B: advances in tetracycline-inducible gene expression. *Proc. Natl Acad. Sci. USA* 96, 797–799

5 No, D. *et al.* (1996) Ecdysone-inducible gene expression in mammalian cells and transgenic mice. *Proc. Natl Acad. Sci. USA* 93, 3346–3351

6 Braselmann, S. *et al.* (1993) A selective transcriptional induction system for mammalian cells based on Gal4-estrogen receptor fusion proteins. *Proc. Natl Acad. Sci. USA* 90, 1657–1661

7 Wang, X.J. *et al.* (1999) Development of gene-switch transgenic mice that inducibly express transforming growth factor beta1 in the epidermis. *Proc. Natl Acad. Sci. USA* 96, 8483–8488

8 Kellendonk, C. *et al.* (1999) Inducible site-specific recombination in the brain. *J. Mol. Biol.* 285, 175–182

9 Gossen, M. and Bujard, H. (1992) Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc. Natl Acad. Sci. USA* 89, 5547–5551

10 Negeri, D. *et al.* (2002) Inducible RNA interference uncovers the *Drosophila* protein Bx42 as an essential nuclear cofactor involved in Notch signal transduction. *Mech. Dev.* 117, 151–162

11 Shimizu-Sato, S. *et al.* (2002) A light-switchable gene promoter system. *Nat. Biotechnol.* 20, 1041–1044

12 Quail, P.H. (2002) Phytochrome photosensory signalling networks. *Nat. Rev. Mol. Cell Biol.* 3, 85–93

13 Ni, M. *et al.* (1999) Binding of phytochrome B to its nuclear signalling partner PIF3 is reversibly induced by light. *Nature* 400, 781–784

14 Gambetta, G.A. and Lagarias, J.C. (2001) Genetic engineering of phytochrome biosynthesis in bacteria. *Proc. Natl Acad. Sci. USA* 98, 10566–10571

0167-7799/03/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0167-7799(02)00022-7

The new role of SAGE in gene discovery

Kenneth R. Boheler and Michael D. Stern

Laboratory of Cardiovascular Science, National Institute on Aging, NIH, 5600 Nathan Shock Drive, Baltimore, MD 21224, USA

The sequencing of the human genome has led to *in silico* predictions of far fewer genes than anticipated. Recent studies using serial analysis of gene expression have cast doubt on this finding. One study predicts the presence of many unidentified low abundance transcripts, whereas two others have mapped unique tags to previously unpredicted exons. Genome and transcriptome complexity is thus greater than predicted and many of the missing genes are probably expressed in low copy numbers or only in early embryonic tissues.

With the near completion of the human genome sequence [1,2], the number of predicted human genes (30 000–40 000) has turned out to be much lower than earlier estimates (45 000–140 000). However, recent data have called these predictions into question [3,4] and the number of functional transcriptional units in the mouse and human genomes remains controversial. Given that the identification of all genes in a mammalian genome is crucial to understanding the function and regulation of transcriptional units, a definitive determination of the correct number of genes depends on an accurate estimate and physical map of all of the genes in a genome.

Expressed sequence tags

Identification of all transcripts and RNA splicing variants in the mouse and human transcriptomes is the ultimate complement to the genome project. This has been accomplished primarily through the analysis of expressed sequence tags (ESTs) with subsequent EST mapping to the corresponding genome. Although data collected from EST projects have contributed significantly to the identification of all the genes, the rate of novel sequence discovery using this method has decreased as a function of time. Although it is possible that most of the unique transcripts have been identified, it is more likely that technical or other limitations to the EST approach have been reached. EST projects on early embryonic cells and stem cells, for example, remain a minor component of the mammalian EST collection (~4 million ESTs from human and ~2 million ESTs from mouse: dbEST [5]) because of technical limits in the production of cDNA or EST libraries from small sample sizes or because of ethical problems associated with obtaining tissues from human embryonic tissues and preimplantation embryos (eggs to blastocysts). It therefore remains probable that many unique transcripts, either from novel genes or in the form of splicing variants, remain to be identified.

Serial analysis of gene expression (SAGE)

Serial analysis of gene expression (SAGE) has the

Corresponding author: Kenneth R. Boheler (bohelerk@grc.nia.nih.gov).

capability of detecting and quantifying the expression of large numbers of known and unknown transcripts [6]. Two major principles underlie this method. First, short DNA sequences are sufficient to identify individual gene products, and second, concatenation (linking together) of the short DNA sequences (i.e., SAGE tags) increases the efficiency of identifying unique transcripts in a sequence-based assay. The technique generates large numbers of short (e.g., 10 base pair) tags, originating from the last (most 3′) unique location of an enzyme recognition site in a single transcript, that when sequenced, can potentially identify 4^{10} (1 048 576) unique transcripts. The newer LongSAGE method [7] can distinguish 4^{17} different tags, a number sufficient to be virtually unique even within the whole genome. There are now >2 380 000 (464 989 unique) and 274 000 (141 665 unique) SAGE tags of human and mouse origin, respectively in the SAGE database (http://www.ncbi.nlm.nih.gov/SAGE/index.cgi?cmd = printstats). Within these datasets, many SAGE tags do not correspond to sequences located within currently predicted genes. Although some of the mismatches might be owing to sequencing errors that are inherent to the SAGE method, recent breakthrough studies using SAGE catalogs have altered the previous estimates of biological complexity and the number of novel genes in mouse and human genomes.

SAGE estimates the number of transcriptional units in early development

Pilot surveys of early mouse embryonic tissues and preimplantation embryos suggest that a large number of developmental regulated genes are only expressed in mouse embryonic tissues and might be rare in most EST collections [8]. We used SAGE to examine mouse R1 embryonic stem (ES) cell lines to identify molecular traits of pluripotentiality and to predict the number of unique transcripts present in cells derived from the inner cell mass of preimplantation embryos [9]. From this SAGE catalog, we identified 44 569 unique SAGE tags, of which 35.2% did not match any sequence previously deposited in any EST dataset. Similar findings were found with three SAGE libraries (47 608 unique tags, ~33% with no EST sequence matches) from pluripotent P19 embryonic carcinoma cells induced to differentiate *in vitro* to cardiomyocytes [10]. These findings support the notion that many genes, known and unknown, might be uniquely expressed in preimplantation embryos.

In the ES cell SAGE catalog, the number of transcripts increases rapidly (as an inverse-square power law) as the expression level detected is reduced [11]. This makes it difficult, for statistical reasons, to estimate accurately the total number of unique transcripts but a naïve correction for sampling and sequencing error indicates that there must be >54 000 unique transcripts in this cell population. Monte-Carlo simulations based on the inverse-square model indicate that as many as 130 000 unique transcripts are compatible with our observations. Given that a wide variation in the number of unique transcripts is possible, it is likely that many very low abundance transcripts have been missed. Because up to a third of the tags in the ES cell library do not map to any previously described EST dataset, the number of unique transcripts (splice variants or novel gene transcripts) that have not yet been identified might be very high (~16 000 to 39 000). Only by sequencing and analyzing sufficient numbers of tags from early developmental stages will it be possible to identify all the transcripts that might be expressed uniquely in embryonic tissues.

Identification of novel transcripts and genes through use of SAGE tags

In the first study of its kind, Chen *et al.* [12] identified many unique transcripts and potentially novel genes though the use of SAGE tags. These authors used publicly available SAGE datasets from diverse tissue and cell sources to identify 375 856 unique SAGE tags of human origin. Of these, 62.3% did not match any UniGene cluster. The authors verified >1000 SAGE tags, after generating longer cDNA fragments from SAGE tags for gene identification (GLGI) [13] and showed that most of the unmatched SAGE tags (67%) originated from novel transcripts that did not match existing ESTs. The authors then generated 17 full-length cDNAs to confirm that these tags were derived from authentic transcripts and not from genomic contamination or sequencing error. Each of the full-length cDNAs had a putative open reading frame, and the sequences could be matched to exons present in the human genome. Importantly, 13 out of 17 sequences did not match any EST, and nine of these did not match any predicted exon. Some of the latter might have been transcribed from the antisense strand of known genes, and might have a regulatory role in RNA stability, whereas, the others might be indicative of novel genes. As pointed out by the authors, converting novel SAGE tags to longer sequences using GLGI should accelerate the discovery of novel transcripts or novel genes in the human genome.

In a similar study, Saha *et al.* [7] used long SAGE tags (21 base pairs) to annotate the genome. Because of their uniqueness, these long SAGE tags can be reliably mapped back to the genome. In a library of ~28 000 tags from a single colorectal cancer line, they found 14 020 distinct tags, of which 8 570 could be matched to the genome databases of Celera or the Human Genome Project (the remainder presumably being owing to sequencing errors or gaps in the database). Within this small collection, there were ~900 novel genes and another ~600 previously unidentified internal exons of predicted genes. The authors estimate that, conservatively, there must be >15 000 previously unpredicted genes in the human genome. LongSAGE should thus accelerate the discovery of novel genes in the human genome.

Outlook of SAGE to identify novel genes

A complete catalog of all the RNA transcripts and splicing variants is the ultimate complement to the genome project. The challenge is to complete the catalog and identify all the genes in the mouse and human genomes. As Chen *et al.* [12] and Saha *et al.* [7] have shown, SAGE tags can efficiently identify novel transcripts or novel genes in the genome. These findings furthermore confirm that earlier predictions of the number of genes in the human genome might be too conservative. A systematic large-scale analysis of the genome by LongSAGE or by SAGE coupled to GLGI is thus complementary to other EST approaches for gene identification and might be crucial to overcome the current computational inadequacies of gene prediction. If we are ever to generate a complete catalog of gene transcripts, these 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 © 2002 Elsevier Science Ltd. All rights reserved.
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).