What transcripts are found in a human cell?

ArticleInfo		
ArticleID	:	3565
ArticleDOI	:	10.1186/gb-2000-1-1-reports031
ArticleCitationID	:	reports031
ArticleSequenceNumber	:	56
ArticleCategory	:	Paper report
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 2000–2–14 Received : 2000–2–14 OnlineDate : 2000–4–27
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	
ArticleContext	:	130591111

Hamish Scott

Abstract

Serial analysis of gene expression (SAGE) has been used to analyze the complete human 'transcriptome' - the number, identity and level of expression of genes in humans.

Significance and context

Understanding how the spatial and temporal expression patterns of individual genes relate to expression patterns of other genes and to changes in the organism's behavioral state, onset of disease and response to drugs should provide valuable insight into molecular physiology. Hence the great interest in high-throughput differential gene-expression technologies that can quantitatively analyze all mRNAs expressed by a cell or tissue type at any given time. The number, identity and level of expression of the entire set of genes expressed from a eukaryotic genome for a defined population of cells is defined as a 'transcriptome'. An alternative to the much-hyped cDNA microarray technology for this purpose is serial analysis of gene expression (SAGE), which, curiously, was originally described in the same issue of *Science* as the microarray methodology from Stanford. As with expressed sequence tags (ESTs), SAGE relies on sequencing to identify genes and can be considered as a variant of EST analysis. Unlike EST analysis, however, SAGE identifies only a short sequence from a defined position within the transcript. The use of short tags enables approximately 40 times as many transcripts to be identified by SAGE as can be identified in an EST project for the same sequencing effort.Theoretically, the defined position of the tag within the transcript enables unambiguous transcript identification, in contrast to ESTs. Also, SAGE does not depend on prior knowledge of transcript sequence; experimental data are electronically stored and can be reanalyzed as genome projects advance or are completed; and SAGE provides absolute rather than relative expression levels. The analysis by Velculescu *et al.* provides a true snapshot of how many transcripts are needed to make up a human cell's biochemical machinery and the levels of expression of these transcripts, and provides a wealth of data for analysis in other laboratories.

Key results

A total of 3.5 million transcripts was analyzed from 19 tissues (both normal and diseased, mainly cancer cell lines or primary tumor samples). From this, the number of genes in the human genome was estimated at around 84,000. The average number of different but related transcripts corresponding to each gene is 1.6 (134,135 transcripts total, mainly due to differences in polyadenylation). More than

43,000 transcripts were expressed in a single cell type (colorectal cancer cell lines) with expression levels ranging from 0.3 to 9,417 copies of the transcript per cell. Of the transcripts, 83% were present at levels as low as one copy per cell; 55 transcripts present at over 500 copies per cell made up 18% of the cellular mRNA mass (Figure 1a); and the most highly expressed 633 genes accounted for 45% of the cellular mRNA. Most unique transcripts were produced at low levels, with just under 25% of the cellular mRNA mass being made up of 94% of the unique transcripts (Figure 1b). Approximately 9,000 genes of known function and 63,000 genes of unknown function were matched to the transcripts; the remaining transcript tags, mainly from genes expressed at a low level (46%), had no matches in existing (public) databases (Figure 1c). Differences in gene expression between different cell types were greater than the changes in gene expression observed in different physiological states of a given cell type. Expression levels of tissue-specific transcripts present at more than ten copies per cell ranged from 0.05% to 1.76% (as a percentage of total cellular mRNA), and 50% of these transcripts had no database match. Approximately 1,000 'ubiquitously' expressed transcripts were detected and may be viewed as a minimal transcriptome.

Figure 1 **(a)** Representation of the 3,496,829 total 'tags', representing 143,135 unique transcripts by mRNA expression levels (copies per cell) as percentages of total cellular mRNA. **(b)** The number of unique transcripts classified by mRNA expression levels. **(c)** The proportion of unique transcripts with and without matches to GenBank mRNA or EST seqeunces, classified by mRNA expression levels.

[MediaObjects/13059_2000_Article_3565_Fig1_HTML.jpg]

Links

More information about SAGE is available from the Serial analysis of gene expressionhttp://www.sagenet.org homepage and the SAGEmap site at the National Center for Biotechnological Information (NCBI). For information on cDNA microarray technology visit the Gene Chips (DNA Microarrays)http://www.gene-chips.com website.

Reporter's comments

Velculescu *et al.*have provided a database of transcripts that is ripe for the picking. There are known and unknown transcripts and ones that are ubiquitously expressed or specific to certain cell types. By making comparisons with this published human transcriptome, SAGE analyses of human tissue will be able quickly to identify transcripts unique to other biological situations. Perhaps the most troubling aspect of this data for future studies in all mammals is the quantification of the large number of transcripts that are identified as being present at fewer than five copies per cell. These rare transcripts comprise 25% of cellular mRNA by mass but 94% of unique transcripts, and only 50% match transcript sequences (mRNAs and ESTs) in GenBank or EMBL (Figure 1). Many of these transcripts will be at

least partially identified by genomic sequencing. But will microarray technologies be sensitive enough not only to detect but also to quantitate these low-abundance transcripts?

Many researchers are already combining different techniques, including SAGE and cDNA microarrays, in an attempt to get the best of both techniques. Probes representing a substantial group of genes of known different expression levels (from SAGE) and an additional labeled target RNA (from a cell line for which SAGE data exists) could be added to microarray experiments as an additional set of controls to allow conversion of microarray data to more absolute expression levels.

Table of links

Nature Genetics

Serial analysis of gene expression

SAGEmap

Gene Chips (DNA Microarrays)

References

1. Velculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C, Lal A, Wang CJ, Beaudry GA, Ciriello KM, et al: Analysis of human transcriptomes. Nat Genet. 1999, 23: 387-388. 1061-4036

Image Object

1. MediaObjects/13059_2000_Article_3565_Fig1_HTML.jpg