PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

SAGE is better than dbEST

ArticleInfo		
ArticleID	:	4311
ArticleDOI	:	10.1186/gb-2002-3-12-reports0062
ArticleCitationID	:	reports0062
ArticleSequenceNumber	:	34
ArticleCategory	:	Paper report
ArticleFirstPage	:	1
ArticleLastPage	:	3
ArticleHistory	:	RegistrationDate : 2002–10–3 Received : 2002–10–3 OnlineDate : 2002–11–22
ArticleCopyright	:	BioMed Central Ltd2002
ArticleGrants	:	

ArticleContext	:	13059331212

Cathy Holding

Summary

The more data compiled from serial analysis of gene expression experiments, the more novel genes are likely to be found, in contrast to the situation with expressed sequence tags.

Significance and context

The Human Genome Mapping Project has published its draft of the entire sequence of the human genome, but the number of functional genes contained in the sequence is still a matter of controversy. The generation of expressed sequence tags (ESTs) has until now been the method of choice for the discovery of novel genes and splice variants. ESTs are nucleotide sequences generated from the ends of randomly selected cDNA clones. A few genes expressed at high level, however, represent a large proportion of the total transcripts and are thus more frequently represented in the EST database. Methods such as subtraction hybridization are often used to try to redress this imbalance. A more recent method of generating fragments of sequence data is by serial analysis of gene expression, or SAGE. Although this technique also reflects relative abundance, the difference is in the method of data generation. SAGE produces tags that are usually 10 bp long, which remarkably is sufficient to identify the transcript from which each tag comes. The process, though, concatenates many unrelated tags together, which are cloned and sequenced. The amount of information generated from SAGE is thus greater than that from ESTs. Some SAGE tags do not match EST data, however, and so SAGE data have previously been dismissed as resulting from sequencing errors. Chen *et al.*have carried out a detailed examination of SAGE data and find that this is not the case after all.

Methodological innovations

Systematic analysis of the probabilities of sequencing errors was carried out, and predictions based on these statistics were compared with the observed numbers of single-copy SAGE tags actually found in and between SAGE libraries.

Conclusions

The number of unique SAGE tags is, for the most part, not due to sequencing errors, and is rising with the total number of SAGE tags. Unmatched SAGE tags therefore represent as-yet unidentified genes in the human genome, and so, the authors conclude, more SAGE data will yield more novel genes. Thus, there are more genes in the human genome than suggested by EST data, and the authors propose that the number may be as much as an order of magnitude greater than first thought. The number of novel sequences found in EST data is falling with the increasing volume of EST data, and the authors conclude that this may be due to the reflection of high copy-number sequences by ESTs and also argue that including a subtraction process may actually result in novel transcripts being lost.

One of the results to come out of checking SAGE data was that the authors found that a gene transcribed in the forward direction of a sequence may have a different function to that transcribed in the reverse direction, and conclude that these must qualify as separate genes. They therefore challenge the practise of collecting together forward and reverse EST transcripts and ascribing them to the same gene, as in the UniGene Database, and conclude that this is also affecting the estimated number of genes in the genome.

Reporter's comments

It happens sometimes that scientists arrive at a tentative conclusion, based perhaps on an opinion, that over the course of time becomes accepted as a true explanation due to its repeated use, but which has never been formally examined. This paper has examined such a misconception and thus potentially prevented another one.

Table of links

Proceedings%20of%20the%20National%20Academy%20of%20Sciences%20of%20the%20United%20States%2

References

1. Chen J, Sun M, Lee S, Zhou G, Rowley J, Wang S: Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. Proc Natl Acad Sci USA. 2002, 99: 12257-12262.