Serial Analysis of Gene Expression, II

(taken from http://www.ncbi.nlm.nih.gov/SAGE/)

The construction of a SAGE tag to gene mapping is a multistep, automated process. These steps include the following:

1. Separating out individual human sequences from GenBank submission records.
2. Assigning a SAGE tag to each sequence, by:
   (a) assigning sequence orientation through a combination of identification poly-adenylation signal (`ATTAAA` or `AATAAA`), poly-adenylation tail, and sequence label, and
   (b) extracting a10 base tag 3'-adjacent to the 3'-most NlaIII site (`CATG`).
3. Using information from NCBI's UniGene project, assigning an UniGene identifier to each human sequence with a SAGE tag.

The result of this process is a SAGE tag to UniGene identifier mapping, with forward and reverse sequence frequency weights (w) given to the connections. In the tag to UniGene cluster mapping, any one tag may be paired to one or many clusters, and any one cluster may pair to one or many tags. For tag to cluster pairs there is a forward and reverse mapping weight. The forward weight is equal to the number of sequences with tag t in cluster c divided by the total number of sequences with tag t. The reverse weight is equal to the number of sequences with tag t in cluster c divided by the total number of sequences in cluster c (with tags).

Sequencing errors

Any error occurring in the sequencing of these DNA sequences can lead to serious "background noise" in the tag to gene mapping. Different types of sequences have different rates of sequencing error. For instance, single-pass EST sequencing error rates have been estimated to be 1-3%/base. In 10 bases, a 1%/base error rate means roughly a 10% chance $(1.00 - (0.99)^{10})$ of one or more errors occurring, and a 3%/base error rate means roughly a 26% chance $(1.00 - (0.97)^{10})$ of one or more errors occurring. Since single-pass EST sequences make up the bulk of the sequences in GenBank, this error rate becomes extremely important in order to quantify and remove this background noise as completely as possible, without harming the overall utility of the mapping. On the other hand, submissions to GenBank of well-characterized sequences, have a very low sequencing error rate, which has been considered to be negligible for our purposes.

Background noise removal

Currently, perceived sequence errors are not removed from GenBank, UniGene or this SAGE tag mapping, for the very reason that is difficult (or, perhaps, impossible) to separate these errors from sequence polymorphisms. However, in this mapping, we can use certain assumptions about the error rates given above to "recommend" certain connections between tags and genes, and not others. We can construct this "recommeded" mapping simply by accepting all connections above a certain strength, or by accepting a certain fraction of the strongest connections.

We know that, based on our assumed single-base error rate, we can calculate a resultant 10 base error rate. For instance, given a 1%/base error rate, there is a 10% chance of one or more errors occurring in 10 bases. Therefore, 10% of the "weakest" tag to gene connections (and vice versa) are most likely to be due to error. These could be removed from the "reliable" tag to gene mapping, and in this implementation, this is what we have done.

We have taken this "strength of connection" approach for the tag to UniGene mapping, but have changed (10/26/00) our approach to presenting "reliable" UniGene to tag assignments. In the past, we had accepted a certain fraction of the stongest UniGene cluster to tag assignments. We now accept all tags derived from well-characterized mRNA or CDS sequence, as well as the most frequently occurring tag derived from EST data as being "reliable".

Resultant mappings

Two tag to gene mappings result from this entire process. One is a "full" mapping, and the other a "reliable" mapping. Both of these are provided on the SAGEmap FTP site as downloadable files (very large), as well as integrated with the SAGE library data on this site through a searchable interface. If one wishes to search for a small number of tags, we suggest using the searchable interface on lists the Tag to Gene and Gene to Tag pages (see below). Downloadable mappings are available under the "Downloads" section links on the left sidebar. The reliable mapping is also used in the display of tabulated SAGE tag data from the various SAGE libraries.

Distributions for these mappings are available on the Mapping Statistics page.

Tag to Gene and Gene to Tag pages

Ten base tags can be entered and both full and reliable mappings searched from the Tag to Gene webpage. Gene to tag full and reliable mappings can be queried via the Gene to Tag webpage, using the gene's UniGene identifier (ie, Hs.#), which can be procured from the UniGene website by entering the gene symbol in the search box. Gene symbols for various genes of interest can be procured from the NCBI OMIM website.

For more information on current SAGE data analyses, summaries and online analysis tools, please click on the About tools link on the left sidebar.

Caveats

It is possible that some tags map to sequences containing repeat sequences (Alu) or mitochondrial sequences. These sequences have been, in large part, masked in or removed from UniGene, but not from this mapping. Mitochondrial sequences have not been removed from the SAGE library data, but have been flagged in the tag to gene mapping with the following phrase: "WARNING: Tag matches mitochondrial sequence." A future attempt will be made to make a more explicit warning about tags matching to repeat containing sequences.

If the warning "!!This cluster has likely been retired!!" appears under the gene description column, please try back to the site in a few days. This is most likely due to synchronization problems between this mapping and UniGene.