Serial Analysis of Gene Expression

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

Serial analysis of gene expression, or SAGE, is a technique designed to take advantage of high-throughput sequencing technology to obtain a quantitative profile of cellular gene expression. Essentially, the SAGE technique measures not the expression level of a gene, but quantifies a "tag" which represents the transcription product of a gene. A tag, for the purposes of SAGE, is a nucleotide sequence of a defined length, directly 3'-adjacent to the 3'-most restriction site for a particular restriction enzyme. As originally described[1], the length of the tag was nine bases, and the restriction enzyme NlaIII. Current SAGE protocols produce a ten to eleven base tag[2], and, although NlaIII remains the most widely used restriction enzyme, enzyme substitutions are possible. The data product of the SAGE technique is a list of tags, with their corresponding count values, and thus is a digital representation of cellular gene expression.

However, to say that SAGE produces a digital output, is not to imply that no loss of fidelity occurs from the conversion of an actual transcript and its expression level to a tag and its count value. Accuracy in both the assignment of tags to genes as well as the ability to quantify a gene's expression level are sacrificed in order to increase throughput, and therefore increase the speed and lower the cost of analysis. A ten base tag is by no means a perfect representation of a gene's entire transcript. There will be instances in which two or more genes share the same tag (i.e., the tag to gene assignment is ambiguous), and instances in which one gene has more than one tag (i.e., through alternate termination in an individual, and polymorphism in a population, the gene to tag assignment is not specific). And, as if this inherent difficulty in making specific and unambiguous tag to gene assignments wasn't enough, an entirely acceptable sequencing error rate from the point of view of most sequencing tasks can have several disturbing effects on SAGE tag data, when dealing with such short sequences.

So there are, really, two problems to be tackled when dealing with SAGE data in the form of tags and counts. The first deals with insuring that the tags and their counts are a valid representation of transcripts and their levels of expression, and the second, with making valid tag to gene assignments.

In consideration of the first problem – the valid data problem – sequencing error has the greatest effect. Assuming that there is an average 1% per base sequencing error rate, for ten bases, the chance of one or more errors occurring is roughly 10%[3]. The error, if it occurs, will, of course, lower the correct tag count by one, but will also either increase the tag count of an already established tag by one, or will establish and count a tag which does not, in reality, exist. The former effect is not of great concern when drawing conclusions from tags with relatively high counts, since raising or lowering a tag count by one or two should, overall, have no great effect. The former and latter effects, on the other hand, do much to increase suspicion of the tags with low counts, particularly those with a count of 1. Currently, the only way this suspicion has been dealt with has been to remove from the data tags counted only once[4]. This may not be an optimal approach, and investigations are currently underway to discover if a better approach might exist[5].

In consideration of the second problem – making valid tag to gene assignments – unspecific and ambiguous tag to gene assignments, as well as sequencing error, both play a role in creating confusion. In making tag to gene assignments, a certain degree of messiness is encountered. It would be preferable if specific and unambiguous gene assignments could be made for every experimental derived tag, but this is definitely not the case. The difficulties are several, and begin with the set of sequences from which tags are derived. The transcriptome of Homo sapiens has yet to be entirely sequenced, let alone characterized. Until it is, there is only an incomplete set of sequences from which to derive tags. Next, considering the nature of the roughly 1.5 million transcript-source sequences stored in GenBank, only about 18,000 are well-characterized cDNA/mRNA sequences, while the vast majority are expressed sequence tag (EST) sequences. The problem with using EST sequences for the derivation of 10 base tags is that they are usually only single-pass sequenced, and therefore, have, roughly, an average 1%

per base error rate. Following the reasoning in the paragraph above, this means a 10% chance that a 10 base tag will include one or more errors. Considering that tag to gene assignments are based upon the sequence from these tags, it stands to reason that roughly 10% of the assignments that are extracted from these sequences will be incorrect. This compounds the "naturally" unspecific and ambiguous tag to gene assignments which are already expected without considering sequencing error.

The naturally unspecific and ambiguous tag to gene assignments can be measured by extracting SAGE tags from the 17,000 or so, nearly error-less, well-characterized cDNA/mRNA sequences in GenBank, and matching those tags to some set of defined gene-units. As noted above, since the transcriptome of Homo sapiens have not yet been sequenced or characterized, an artificial method for defining gene-units must be chosen. This may be as simple as taking the title of the GenBank sequence entry, or as complicated as using a set of gene contigs, or using gene-based sequence clusters, such as the UniGene gene set[6]. Tags extracted from these nearly error-less sequences give an idea of baseline specificity and ambiguity before sequences with higher sequencing error rates are considered.

A heuristic approach has been used to select only 3' EST sequences (using sequence submission label, and polyadenylation signal and tail) before extracting tags and making tag to gene assignments. Using the estimation of the percentage of tags due to error (i.e., 10%), a "correction" of these EST-based tag to gene assignments can be made by removing 10% of the most rarely occurring tags for a particular gene as well as 10% of the most rarely occurring genes for a particular tag. This correction has been performed on a rank-ordering of tag-gene pairings[7], and gene-tag pairings[8]. Once this correction is made, combining the tag to gene assignments of the well-characterized sequences with those of the corrected EST sequences gives us as much reliable information as possible, while greatly reducing the effect of erroneous tag to gene, and gene to tag, assignments[9]. However, an obvious drawback to this approach is that there is no method to tell the difference between tag to gene assignments due to sequencing error, and rare, naturally occurring assignments, which could occur through rare alternate splicing events or polymorphisms, and these would both be removed.

Two tag to gene assignment lists, a.k.a. tag-gene mappings, based upon all human transcript-source sequences in GenBank and the UniGene gene clustering algorithm, have been constructed and are publicly available from this website, in both downloadable and interactive versions. These assignments consist of a reliable mapping and a full mapping. The reliable mapping includes the correction for EST sequencing error; the full mapping does not. In addition, this website provides public access to NCI's Cancer Genetic Anatomy Project (CGAP) SAGE data from human colon and brain tissues, and a statistical test for differential analysis between data sets (or libraries).

Gene expression technologies allow us to confront large amounts of data from which we endeavor to glimpse the inner workings of the cell (transcriptome analyses), or differentiate one type of cell from another (differential expression analyses). Obtaining this data is a difficult task unto itself, but, however it is done, collection of the data is only half the fun. Information must be forcefully extracted from the data – the cacophony quieted – through assumption, knowledge, intuition and fortune. As is true for all analyses, the pump must be primed before information can flow.

Notes and References:

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995 Oct 20;270(5235):484-7.

2. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW. Gene expression profiles in normal and cancer cells. Science. 1997 May 23;276(5316):1268-72.

3. $1.0 - (0.99)^{10} \sim 0.10$

4. This empirical approach has been used in SAGE tag-count sets in which roughly 250,000 total tags have been sequenced. For SAGE data sets, or combinations of sets with, for example, over

one million total tags sequenced, it might be necessary to exclude tags with counts of less than 2, 3 or more.

5. A possible approach is to calculate, and make use of, the expected number (or percentage) of nearest neighbors (i.e., one base substitution, insertion or deletion) with a count of one, two, etc., given the number of total tags sequenced, and the actual tags in the data set.

6. For more information and data about the UniGene project, see http://www.ncbi.nlm.nih.gov/UniGene.

7. The tag to gene rank-ordering is a calculated value consisting of the number of sequences containing a particular tag falling into a particular gene group divided by the total number of sequences containing that tag.

8. The gene to tag rank-ordering is a calculated value consisting of the number of sequences falling into a particular gene group containing a particular tag divided by the total number of sequences falling into that gene group.