Jessica Williamson

MBB 452a Fall 2001

12/13/01

Analyzing SAGE data using the Unigene Collection

Serial analysis of gene expression, or SAGE, is one of several experimental techniques that are currently used to quantify and compare gene expression levels among various cell types. SAGE takes advantage of simple molecular biology and high-throughput sequencing to generate large amounts of expression data (1,2,3). In essence, SAGE measures the expression of a "tag" rather than the entire transcription product of a gene. The tag is generated by cleaving a mRNA population with an anchoring enzyme, most often with NlaIII. The 3'-end of the transcript is retained, and cleaved again by a tagging enzyme 10 bp after the NlaIII site (CATG). The tag is thus defined as the 10 bp segment of a gene directly 3'-adjacent to the 3'-most NlaIII restriction site of a gene. These 10 base pairs are sufficient to uniquely identify each gene. The entire set of tags is concatenated and sequenced to give a list of all the tags in the sample, and their corresponding count values (1,3). There are two challenges associated with gathering SAGE data. The first is ensuring that the tags and counts are accurate representations of actual expression levels. The second challenge is making valid tag to gene assignments. Both of these are affected by errors in sequencing, which can cause up to 10% inaccuracy in tag sequences (3).

In order to match tags to their gene counterparts successfully, a transcript map of the genome is needed. The Unigene Collection is an initiative of the National Center for Biotechnology Information. It was developed in response to the large number of expressed sequence tags (ESTs) being added to GenBank in the early-mid 90's (4,5). ESTs are short fragments of genes that have been sequenced only once and are therefor relatively inaccurate (6). ESTs in GenBank have caused a high amount of redundancy among the sequences (4,5). Unigene is an experimental system that organizes GenBank sequences into a non-redundant set of gene-oriented clusters. This involves grouping subsets of sequences that belong together into a larger set that represents one gene.

Unigene consists of both the well-characterized gene sequences and ESTs. The Unigene build involves staged clustering in six steps, with each step adding less reliable data than the previous step (5,7).

The first step is a quality control step where all the sequences to be compiled are screened for non-human DNA contaminants, vector, mitochondrial and ribosomal sequences, repeat elements, and low-complexity regions. After screening, only non-ambiguous sequences that are 100 bp or longer are included. In the second step of the process, all the sequences are compared to each other. Sequences that are sufficiently similar are linked together to form initial clusters using MEGABLAST. EST to gene and EST to EST links are included in these clusters. The third step involves forming anchored clusters. An anchored cluster must have a 3' polyadenylation signal or have two ESTs labeled as 3'. Any sequences that join two distinct clusters from a previous step are discarded. In the fourth step, clone-base edges are added. These are clone associated 5' ESTs that allow non-overlapping 5' and 3' ESTs to belong to the same cluster. Here, clone-based edges that link at least two 5' ends to a single cluster that contains at least two 3' ends from the same clones are found and added. The fifth and sixth steps deal with singleton clusters (clusters that contain only one EST or sequence) and non-anchored clusters. These are compared back to the anchored clusters at a lower stringency than the previous passes. This process merges several clusters and reduces the number of singleton and non-anchored clusters. In the end, Unigene is distributed as a set of flatfiles, where the cluster identifiers change with each new build. The Unigene Collection is updated about one week after each GenBank release. Since Unigene can change frequently, it is best to identify the clusters by their GenBank accession number, and not by their clustering ID when analyzing data (5,7).

In order to process SAGE data, maps connecting Unigene clusters to SAGE tags have been constructed. Lash *et al* (8) explain how tag-Unigene assignments are made based on the NlaIII anchoring enzyme for human. First, individual human sequences are separated out from GenBank submission records that are represented in Unigene. Next, sequence orientation is assigned through a combination of identification poly(A) signal,

poly(A) tail, and orientation (3' or 5') annotations. Then, the 10 bp tag (as defined above) is extracted and each human sequence with a SAGE tag is assigned a Unigene identifier. Finally, for each tag-Unigene pair, two frequencies are calculated. The first is the number of times one tag-Unigene pair has been seen divided by the number of sequences with the one tag. The second is the number of times the pair has been seen divided by the number of sequences with tags in this Unigene cluster. The process results in "full" tag to gene mapping (5,8). This full map includes many types of ESTs, many of which have a high amount of sequencing error. The most infrequent tag-Unigene assignments are most likely due to sequencing error. "Reliable" tag to gene maps also have been generated that leave out the 10% most infrequent tag-Unigene assignments (8). Both full and reliable maps for human, mouse, and rat are available for downloading on the SAGEmap website (5).

Once the SAGE tags have been identified, it is up to the researcher to interpret the expression levels. An example of some typical results of a SAGE comparison is shown in Figure 1. Other tools that are useful for manipulating SAGE data can be found on both the NCBI SAGEmap site (4) and the SAGEnet site (9). The Unigene database and other bioinformatics tools have proven invaluable for SAGE analysis, and will most likely continue to do so as the field grows.

Figure 1. The following table was taken from Zhang *et al* (2). Listed here are the top 20 transcripts that showed a decrease in expression in colorectal cancers versus normal colon cells. NC is normal colon, TU is colon tumors, CL is colon cells.

Tag sequence	SAGE UID	NC/ TU	TU	CL	NC	GenBank match (accession number)
GACCAGTGGC	H545514	45	1	0	45	No match
ATTTCAAGAT	H259108	37	1	0	37	Carbonic anhydrase II (M36532)
GTCATCACCA	H740629	34	0	0	34	Uroguanylin (U34279)
CTTATGGTCC	H511670	34	1	0	34	No match
TGGAAAGTGA	H950457	34	1	1	34	Human cellular oncogene c-fos (V01512)
CCTTCAAATC	H390158	31	1	0	31	Carbonic anhydrase I (M33987)
TCGGAGCTGT	H893564	30	1	4	30	EST 261490 (H98618)
GTCTGGGGGA	H752297	29	1	3	29	EST 81394 (T60135)
GATCCCAACT	H578824	27	1	1	27	Metallothionein from cadmium-treated cells (V00594)
CTTAGAGGGG	H510123	27	1	5	27	No match
ATGATGGCAC	H233106	26	0	2	26	No match
CCTGTCTGCC	H388582	24	1	2	24	EST 122594 5 (T99568)
CTGGCAAAGG	H500747	23	0	0	23	No match
CTTGACATAC	H516402	22	0	0	22	Homo sapiens CL100 mRNA for protein tyrosine phosphatase (X68277)
GGAAGAGCAC	H657554	21	1	1	21	Gal- (1-3/1-4)GlcNAc -2.3-sialyltransferase (X74570)
TCTGAATTAT	H909556	21	1	1	21	Transmembrane carcinoembryonic antigen BGPb (X14831)
TAAATTGCAA	H790417	19	6	1	113	Cytokeratin 20 (X73502)
GTGGGGGCGC	H764570	18	1	1	18	EST 153570 5 (R48529)
ATGGTGGGGG	H241323	18	2	6	36	Homo sapiens zinc finger transcriptional regulator mRNA (M92843)
TCACCGGTCA	H857781	17	7	7	122	Human mRNA for plasma gelsolin (X04412)

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. NCBI's SAGEmap website: http://www.ncbi.nlm.nih.gov/sage
4. NCBI News. August 1996. http://www.ncbi.nlm.nih.gov/Web/Newsltr/aug96.html
5. NCBI's UniGene website: http://www.ncbi.nlm.nih.gov/UniGene
6. Schuler, GD. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J Mol Med. 1997;75(10):694-698.
7. Yuan J, Liu Y, Wang Y, Xie G, Blevins R. Genome analysis with gene-indexing databases. Phamacol Ther. 2001 Aug;91(2):115-32.
8. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF. SAGEmap: A public gene expression resource. Genome Research. 2000 Jul;10(7):1051-60.
9. SAGEnet http://www.sagenet.org