SAGE Tag matching

We developed an algorithm to reliably match SAGE Tags to UniGene clusters that is similar to that developed by Lash et al. (2000) and also incorporates ideas presented by Caron et al. (2001). Briefly, the ehm-Tag-Mapping method extracts a SAGE Tag from each sequence in a UniGene cluster, only if the orientation and 3′ end of the sequence can be confirmed by identifying poly(A) signals and/or tails. To minimize the extraction of SAGE Tags from entries with potential sequencing errors, SAGE Tags not representing at least 20 % of all Tags extracted from a given UniGene cluster are removed from the final ehm-Tag-Mapping flatfile.

In addition to the SAGE Tag sequence, UniGene ID number, and description, three additional parameters are reported for a SAGE Tag that provide additional criteria on which to base the confidence of the Tag-to-UniGene cluster match: 1) the number of times this particular SAGE Tag was extracted from sequence files in the UniGene cluster (NUM), 2) the total number of times any SAGE Tag was extracted from sequence files in the UniGene cluster (SUM), and 3) the number of unique SAGE Tag sequences extracted from the UniGene cluster (COUNT). As stated in the last paragraph, NUM/SUM must be ≥ 0.2 for a SAGE Tag to be listed in the ehm-Tag-Mapping file. For clarity, the ehm-Tag-Mapping log file output for UniGene clusters Mm.1 and Mm.5 (the first 2 clusters in the UniGene database) are reported in Figure 0.1 on page ii. As an example, the NUM, SUM, and COUNT values for UniGene cluster Mm.1 are 59, 77, and 18 respectively.

To keep the false positive rate of this method to a minimum, the algorithm was specifically designed to prevent the extraction of SAGE Tags from sequence entries for which there was no obvious poly(A) signal and/or tail. Since 1) poly(A) tails are sometimes removed from sequences before they are added to the public databases, and 2) a small proportion of genes will contain noncanonical poly(A) signals, there may be a number of 'true' positive Tag-to-UniGene cluster matches

```
(1) Found UniGene cluster #1
Sending UniGene cluster #1 with 168 files to FindTags at Mon Jan 15 12:59:42 2001
Finished FindTags on UniGene cluster #1
calcium binding protein A11 (calgizzarin)
AAGAAGAAGG => 1
AAGCAGAAGG => 59 <-    ### Lines 5 to 22 contain the list of all SAGE Tags
GGATTAATTG => 1    \   ### found in this UniGene cluster, followed by the
GGTTCAATTG => 1     \  ### number of times the SAGE Tag was identified.
CAATGACTAT => 1      \
AGCATACCCC => 1       \### Note that all Tags except for the one noted on
TGCCTCAATA => 1        ### line 6 were only observed once or twice. In fact,
AATCAGAAGG => 1        ### many of them have a similar sequence to the SAGE
AGGAATTTGC => 2        ### Tag observed 59 times, indicating that they are
GCAGAGAAGT => 1        ### likely due to EST sequencing errors.
TAGCAGCAGG => 1
AACCAGAAGA => 1
CCGCGTGTTA => 1        ### The first value on line 23 is the Number of all
AAGAAGAAAG => 1        ### unique Tags. The second value is the Sum of all
AAGCTGGAGG => 1       /### extracted Tags.
TAATGACTAT => 1      /
AAGCAGAAGA => 1     /
AAGCAGAAAG => 1    /   ### The number on line 24 is the fraction of all Tags
18           77 <-/    ### represented by the Tag observed 59 times and sent
0.766233766233766 <----### to the final ehm-Tag-Mapping output file.


(2) Found UniGene cluster #5
Sending UniGene cluster #5 with 44 files to FindTags at Mon Jan 15 12:59:42 2001
Finished FindTags on UniGene cluster #5
homeo box A10
CATTCAAGGC => 1
CATATAAGGG => 1
CATAATAGGG => 1
AAAAATCCCC => 1
TCTATTATAA => 1
CATAAAACGG => 1
CATAAAAGCG => 1
CATAAAAGGG => 9
CATCAAAGGG => 1
CATAAGAGGG => 1
CATCGAAGGG => 1
GAAGCCTAGG => 1
AACAATTGCC => 1
13           21
0.428571428571429
```

The first entry of the log file has been annotated for explanation purposes after the ###.

Figure 0.1: Representative output from the ehm-Tag-Mapping log file for the first 2 analyzed UniGene clusters

that are missed by this method.

A fourth FIT value (Equation 0.1) was calculated to summarize the NUM, SUM, and COUNT values into a single informative number that is correlated with the quality of the Tag-UniGene cluster match.

$$\text{FIT} = \frac{\text{NUM} + \text{COUNT} - 1}{\text{SUM}} \tag{0.1}$$

The FIT value represents the proportion of Tag-UniGene cluster matches that come from the given Tag sequence plus the COUNT of all other Tag matches in a given UniGene cluster. FIT values of 1 indicate that all other Tags matching the UniGene cluster were observed once, and are therefore likely due to sequencing errors. Values less than 1 indicate an increasing proportion of other Tags matching the UniGene cluster. As a representative example, the FIT values for UniGene clusters Mm.1 and Mm.5 are 0.99 and 1, respectively. In our experience, FIT values above 0.9 are generally due to EST sequencing errors and indicative of only 1 SAGE Tag matching a UniGene cluster, while values less than 0.8 may suggest that an additional SAGE Tag matches the UniGene cluster.

The method presented by Caron et al. is more complex. Among other additions, they adjust for multiple types of sequencing errors, which results in matching a slightly higher number of experimental SAGE Tags to UniGene clusters. However, their algorithms are not easily portable (due to several complex interactions with multiple in-house databases and programming modules) and a Tag to gene matching file is presently available for human data only. As with their method, ours also significantly reduces the false positive rate relative to those of earlier reliable Tag Mapping builds available at NCBI (`http://ncbi.nlm.nih.gov/SAGE`).

The ehm-Tag-Mapping method is implemented through the use of several Perl scripts (Appendix **??** on page ??) designed to extract Tag-to-UniGene cluster infor-

mation from the UniGene flatfiles available at NCBI. A current ehm-Tag-Mapping file was generated in approximately 8 hours on an 800 MHz PIII PC, running Red-Hat Linux 6.2. Because of the versatile portability of Perl and relatively stable format of the UniGene databases, in theory, these scripts can be implemented on a variety of platforms and for all species of UniGene clusters, which currently include Human, Mouse, Rat, Zebrafish, and Cow.

To summarize our SAGE data, we have found it useful to classify SAGE Tags as representing known genes, ESTs, or no match. However, the algorithm used to build UniGene clusters sometimes assigns the sequences of one gene to multiple UniGene clusters, resulting in the same SAGE Tag matching multiple UniGene clusters. To adjust for this, an extended set of Perl scripts and Access 97 (Microsoft Corp.) queries were designed to analyze the description line of each UniGene cluster and classify each SAGE Tag as matching a gene or EST. SAGE Tags matching both a gene and an EST were only classified as matching a gene.

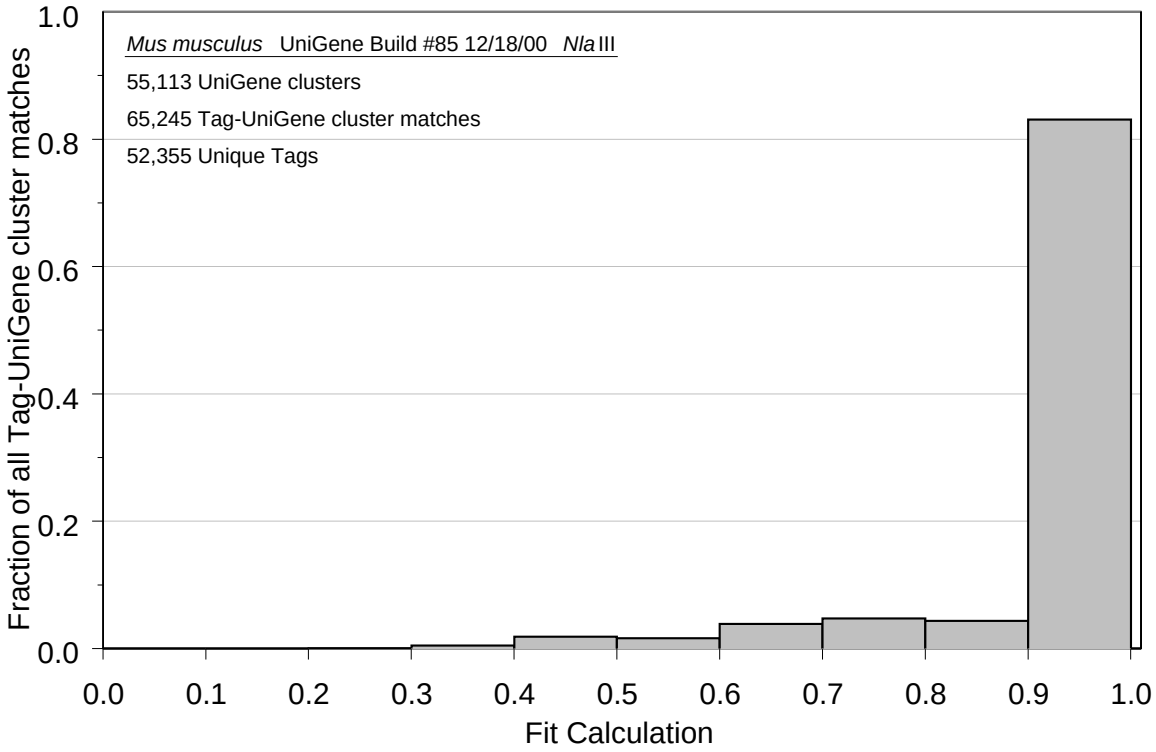
Quality of the ehm-Tag-Mapping method

From the mouse UniGene build #85 on December 18, 2000 (containing 83,862 different UniGene clusters), the ehm-Tag-Mapping method identified 65,245 Tag-UniGene cluster matches, representing 55,113 UniGene clusters and 52,355 unique Tags. 86.4 % of the Tags matched only one UniGene cluster, while 85 % of the UniGene clusters had only one matching SAGE Tag. These percentages increased to 96.3 % and 97.1 % by including Tags matching 2 UniGene clusters, and UniGene clusters with 2 matching SAGE Tags, respectively. There were several SAGE Tags with low sequence complexity, or matching a B2 repetitive element (data not shown) that made up the bulk of the remaining Tag-to-UniGene cluster matches.

Manual analysis on a limited number of Tag-UniGene cluster matches suggests the ehm-Tag-Mapping method produced a correct Tag-to-UniGene cluster

match approximately 90 % of the time (32 out of 35 were correct). Ambiguous Tag matches were the result of minimal and low quality sequence data in the UniGene cluster, suggesting that the limiting factor for this Tag matching method, as well as others relying on EST data, is the quality of sequences submitted to public databases and the ability to correctly group them into different gene clusters.

Since the scope of manual Tag confirmations was only performed on a relatively small subset of Tag-UniGene cluster matches, we also analyzed the distribution of FIT values for all Tag-UniGene cluster matches produced by the ehm-Tag-Mapping method. FIT values of 1 indicate that all other SAGE Tags matching the UniGene cluster were each observed once (see Equation 0.1 on page iii for details). Over 80 % of the FIT values for all Tag-UniGene cluster matches fell between 0.9 and 1 (Figure 0.2 on page vi), indicating that other Tag matches present in the UniGene cluster, and subsequently removed from the final ehm-Tag-Mapping file, were usually observed once and likely due to sequencing errors. Since this is the first Tag-Mapping method to supply these quantitative measurements, it is not possible to compare these values across the different methods.



FIT values were calculated as described in Equation 0.1 on page iii and represent the proportion of Tag-UniGene cluster matches that come from the given Tag sequence plus the COUNT of all other Tag matches in the given UniGene cluster. In our experience, FIT values above 0.9 generally result from UniGene clusters with numerous EST sequencing errors and usually are indicative of only 1 SAGE Tag matching a UniGene cluster.

Figure 0.2: Distribution of FIT values for all Tag-UniGene cluster matches

BIBLIOGRAPHY

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S., van Kampen, A., and Versteeg, R. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, 291(5507):1289–1292.

Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J., and Altschul, S. F. (2000). SAGEmap: a public gene expression resource. *Genome Research*, 10(7):1051–1060.