# Bioinformatic Analysis of SAGE Data and Applications to Programmed Cell Death

**Pleasance ED, Chittaranjan S, Freeman JD, Varhol RJ, Zuyderduyn SD, Marra MA, Gorski SM, Jones SJM**

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

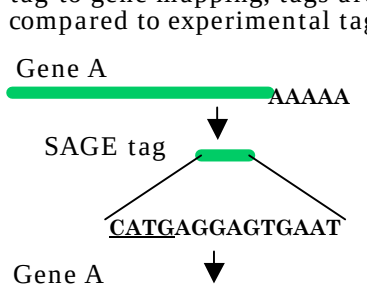
Genome Sciences Centre

www.bcgsc.ca
info@bcgsc.ca

## 1. SAGE overview

### Serial Analysis of Gene Expression

SAGE involves isolating small segments of transcripts ("SAGE tags") for sequencing in such a way that the frequency of each SAGE tag is directly proportional to the expression of the transcript from which it was derived. The sequence and length of the extracted SAGE tag are dependant on the choice of two restriction enzymes used in the SAGE procedure, known as the anchoring enzyme and the tagging enzyme. To determine the gene represented by a SAGE tag, a process called tag-to-gene mapping, tags are extracted from known sequences and compared to experimental tags.




Polyadenylated transcript

SAGE tag extracted from 3'-most anchoring enzyme (*Nla*III) site using tagging enzyme (*Bsm*FI)

Determine gene represented by the tag (tag-to-gene mapping)

...CGAT**CATGAGGAGTGAAT**CCATTTCAATGTGATG...

SAGE is a relatively unbiased method of large-scale gene expression profiling as, unlike microarray methods, it does not require prior knowledge of the genes expressed. Thus, it has the potential to identify novel genes.

### Genes not amenable to SAGE

It is not necessarily possible to determine the expression of every gene using SAGE. There are two primary reasons for this:

- Genes with no anchoring enzyme site will not be present in SAGE libraries, as no tag will be extracted
- Multiple genes which produce the same SAGE tag will not be differentiated when tag-to-gene mapping is done (the tags from multiple genes will be "ambiguous")

There has been little comprehensive study on the importance of these effects in SAGE, despite their potential importance to the use of SAGE for transcript identification.

**SAGE enzymes**

*Anchoring enzyme:* The anchoring enzyme determines the site in a transcript from which the SAGE tag is derived. *Nla*III (CATG) is the most commonly used anchoring enzyme; *Sau*3A (GATC) is also used, and other enzymes are theoretically possible.

*Tagging enzyme:* The tagging enzyme determines the length of the SAGE tag that is extracted. The original and most commonly used tagging enzyme is *Bsm*FI, which extracts a 14 bp SAGE tag. Recently the *Mme*I enzyme has been introduced, which extracts a 21 bp tag. Intermediate tag sizes are not possible with the currently used SAGE procedure and available enzymes.

## 2. Objectives

- Assess the efficacy of SAGE in identifying transcripts by determining the number of genes that cannot be accurately analysed due to lack of an anchoring enzyme site or due to tag ambiguity. Perform this assessment with varying choices of anchoring enzyme, SAGE tag length, and model organism.
- Using tag-to-gene mappings derived as part of the above assessment, identify genes represented in a *D. melanogaster* SAGE library constructed from a tissue undergoing programmed cell death, and demonstrate the utility of SAGE for identification of novel genes involved in this important biological process.

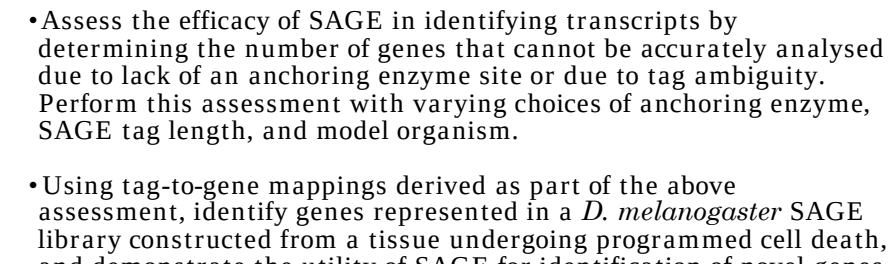
## 3. SAGE assessment: Tag-to-gene mappings

To assess the efficacy of SAGE in identifying transcripts, it is necessary to have complete full-length transcript sets from which to extract tags for tag-to-gene mappings. Otherwise, if two transcripts produce the same tag (making that tag ambiguous) but only one of those transcripts is included in the set, the tag will appear to be unambiguous when this is not the case.

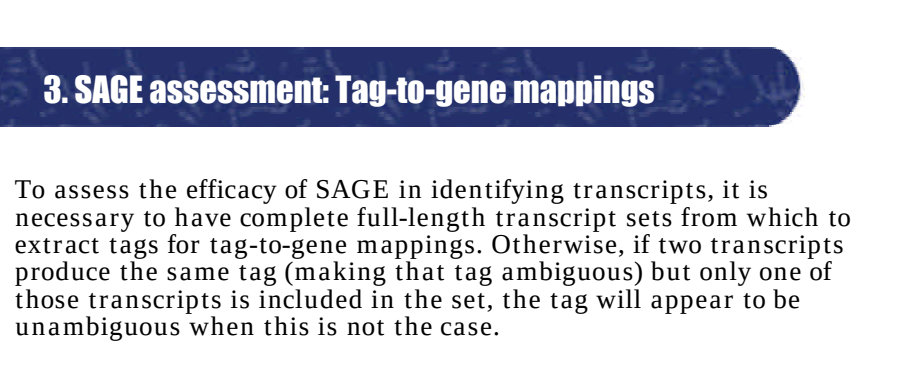
### Constructing full-length transcripts

The model organisms *D. melanogaster* and *C. elegans* both have fully sequenced and annotated genomes, and thus full coding sequences are available for most genes. However, as SAGE tags correlate to the 3'-most anchoring enzyme site in a gene, many SAGE tags are expected to be derived from the 3' untranslated region (UTR) which is not consistently included in the gene predictions. To construct full-length transcripts, UTRs are added as follows, based on sequence information from GadFly and WormBase integrated into ACEDB databases and accessed using Perl scripts:

*Expressed sequences available:*




*Expressed sequences not available:*




### Evaluating tag-to-gene mappings

Tag-to-gene mappings derived from conceptual transcript sets were evaluated for accuracy by comparison to tags extracted from *D. melanogaster* full-length cDNAs.

| Sequences used for mapping | % of tags correctly mapped |
|---|---|
| *Overall accuracy* | |
| cDNAs | 100% |
| Conceptual transcripts | 89% |
| *Represent genes with known expressed sequences (ESTs)* | |
| ESTs | 81% |
| Transcripts constructed without cDNAs | 80% |
| *Represent genes with no known expressed sequences* | |
| Predicted coding sequences | 47% |
| Transcripts constructed without ESTs or cDNAs | 76% |

Mappings are as accurate and less ambiguous (data not shown) as mappings derived from ESTs, and more accurate than those derived from predicted genes alone. Overall accuracy is less than 100% due to genes missing from the genome annotation, and errors in gene predictions.

### Human full-length transcripts

Human full-length transcripts have not yet been constructed as the genome sequence finishing and annotation is still in progress, and so partial full-length transcript sets were derived from the MGC and RefSeq sequence databases.

**Alternative transcripts**

It is important when determining SAGE tag ambiguity to consider the influence of alternative transcripts, as multiple transcripts derived from alternative splicing of the same genomic locus are much more likely to share SAGE tags. Thus, in all the work presented here, "ambiguous" tags that are all derived from the same locus are not considered in the total ambiguity. If alternative transcripts are considered independently, which may be desirable if they have potentially different functions, ambiguity can increase by 50-300% (data not shown).

**Extracting SAGE tags**

SAGE tags were extracted from full-length transcripts at both the 3'-most anchoring enzyme site, as well as upstream enzyme sites. Tags extracted from upstream sites may be relevant if shorter alternative transcripts exist, or if the estimated UTRs are artificially long.

**Data availability**

*D. melanogaster* and *C. elegans* constructed transcripts and tag-to-gene mappings are available from http://sage.bcgsc.bc.ca/tagmapping/.

## 4. SAGE assessment: Tag length and enzyme

### Effect of tag length

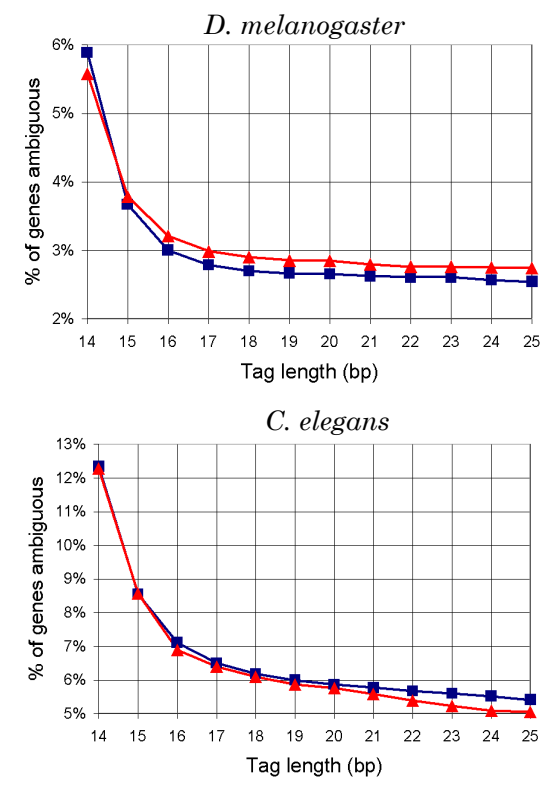



*Figure 1.* Effect of varying tag length on ambiguity.

Tags of varying length were extracted from *D. melanogaster* and *C. elegans* full-length transcripts. Results suggest that increasing tag length decreases ambiguity up to ~16-17 bp. Increasing tag length beyond this point does not have as significant an effect. As currently only 14 bp and 21 bp tags can be extracted, using the longer tag length does not confer a significant advantage.

### Effect of tag length

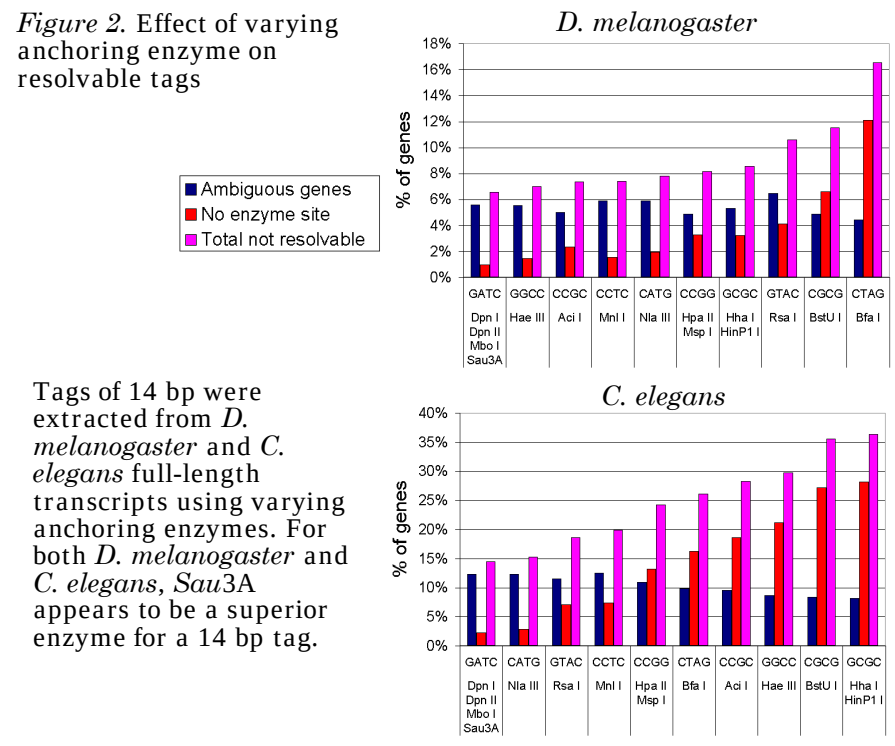



*Figure 2.* Effect of varying anchoring enzyme on resolvable tags

Tags of 14 bp were extracted from *D. melanogaster* and *C. elegans* full-length transcripts using varying anchoring enzymes. For both *D. melanogaster* and *C. elegans*, *Sau*3A appears to be a superior enzyme for a 14 bp tag.

### Human transcriptome

Because the RefSeq sequences do not represent the entire human transcriptome, estimates of likely ambiguity in gene identification by SAGE can be derived from the dependence of ambiguity on transcriptome size.

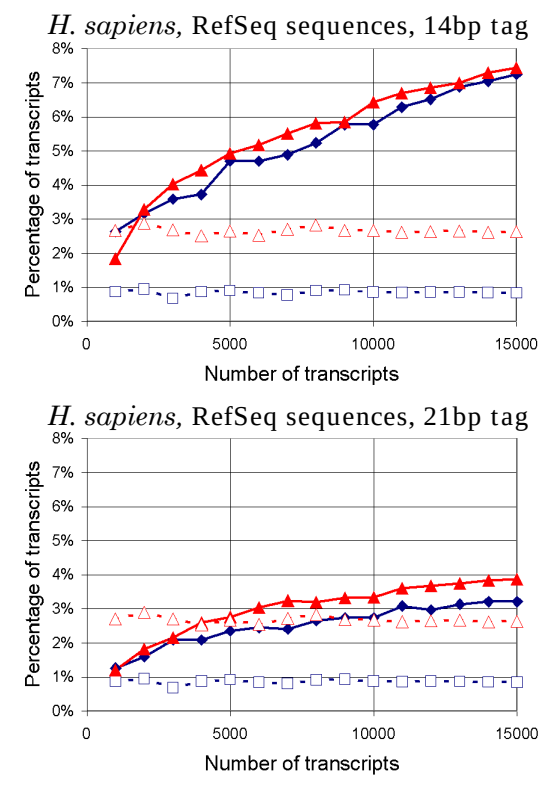
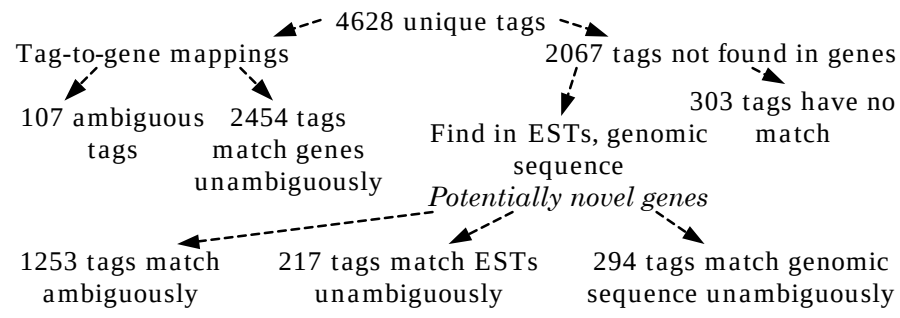



*Figure 3.* Effect of varying transcriptome size, tag length, and anchoring enzyme on resolvable tags

Extrapolation from these graphs suggests that using a 14bp tag, 30,000 transcripts may have up to 9% ambiguity and 50,000 transcripts may have up to 15% ambiguity. With 21bp tags, the ambiguity is reduced by 50%. Note also that *Nla*III is superior to *Sau*3A for human SAGE library construction due to the large number of genes lacking a *Sau*3A recognition site.

## 5. Identifying novel genes

SAGE libraries were constructed from *D. melanogaster* larval salivary glands at three successive time points prior to onset of developmentally-regulated programmed cell death (See posters by S. Chittaranjan and S. Gorski). Using the *D. melanogaster* tag-to-gene mappings described above, SAGE tags were mapped as follows:




If UTRs were not estimated when constructing the transcripts for tag-to-gene mapping, only 2267 instead of 2561 tags matched genes, demonstrating the usefulness of this method of tag-to-gene mapping. Nearly half of the SAGE tags did not match known genes, and over 500 match ESTs or genomic sequence unambiguously, thus pinpointing positions of potentially novel genes.

## 6. Conclusions

- SAGE can identify novel genes with potential roles in *D. melanogaster* programmed cell death
- The preferred enzyme and tag length for SAGE library construction varies and should be considered when designing a SAGE experiment

*Acknowledgements:*
BC Cancer Agency Genome Sciences Centre
Natural Sciences and Engineering Research Council of Canada
Michael Smith Foundation for Health Research
BC Cancer Agency and the BC Cancer Foundation

*References:*
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial Analysis of Gene Expression. *Science* 270:484-487.
BDGP: http://www.bdgo.org/ MGC: http://mgc.nci.nih.gov/
WormBase: http://www.wormbase.org/ RefSeq: http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
GSC SAGE tag-to-gene mappings can be downloaded from http://sage.bcgsc.bc.ca/tagmapping/.