SAGE2SPLICE: UNMAPPED SAGE TAGS REVEAL NOVEL SPLICE JUNCTIONS

by

BYRON YU-LIN KUO

B.Sc., University of British Columbia, 2002
B.Sc., University of British Columbia, 1999

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA
September 2005

Abstract

Serial analysis of gene expression (SAGE) not only is a method for profiling the global expression of genes, but also offers the opportunity for the discovery of novel transcripts. SAGE tags are mapped to known transcripts to determine the source of tags. We hypothesized that tags that map neither to a known transcript nor to the genome span a splice junction, for which the exon combination or exon(s) are unknown. Splice junctions are typically recognized by the pair of highly conserved dinucleotides at each edge of an intron, GT at the 5′ end and AG at the 3′ end, as well as by other less conserved nucleotides flanking the junctions. In the known transcriptome, between 1.6 to 6.2% of *predicted tags* span a splice junction. We have developed an algorithm, SAGE2Splice, to efficiently map these unmapped SAGE tags to potential splice junctions in a genome. An evaluation scheme was designed based on position weight matrices to assess the quality of candidates. Candidates were classified into three types of *spliced tags*, reflecting the previous annotations of the putative splice junctions. A *Type 1* tag spans a novel junction where the exons are known; a *Type 2* tag spans a previously known and an unknown exon; and a *Type 3* tag spans two previously unknown exons. Analysis of *predicted tags* extracted from EST sequences demonstrated that candidate junctions having the splice junction located closer to the centre of the tags are more reliable. Using high sensitivity and high specificity parameters, 7,757 candidates were predicted from 1,639 of 20,000 unmapped tags by SAGE2Splice. We selected 12 candidates splice junctions and tested them using RT-PCR. Nine of these twelve candidates were validated by RT-PCR and sequencing, and among these, four revealed previously uncharacterized exons. To screen more unmapped SAGE tags, we proposed methods to improve SAGE2Splice in engineering efficiency, program usability, and candidate evaluation methods, as well as to include a high throughput laboratory procedure for testing the predicted candidates. We expect that many more novel transcripts can be discovered using SAGE2Splice. SAGE2Splice is available online at http://www.bcgsc.ca/sage2splice/.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

cDNA	complementary DNA
CGAP	Cancer Genome Anatomy Project
DG	dependency graph
EST	expressed sequence tag
GLGI	generating longer cDNA fragments from SAGE tags for gene identification
HMM	hidden Markov model
MDD	maximal dependence decomposition
MGC	Mammalian Genome Collection
NCBI	National Center for Biotechnology Information
PHP	PHP Hypertext Processor; a widely-used general-purpose scripting language for web development.
PWM	position weight matrix
RefSeq	Reference Sequence Project
ROC	Receiver Operating Characteristics
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SAGE	serial analysis of gene expression
snRNP	small nuclear ribonucleoprotein
TFBS	Transcription Factor Binding Sites
UCSC	University of California Santa Cruz

Acknowledgement

I would like to thank my academic supervisors, Drs. Elizabeth M. Simpson and Wyeth W. Wasserman, for providing the opportunity and guidance to work on this interesting project throughout my Masters. Under their guidance, I have learned how to conduct and present research, and learned to be self-motivated. I would also like to thank Drs. Steven J. Jones, Öjvind Johansson, and Asim Siddiqui for their helpful scientific discussions; Miss Ying Chen and Miss Slavita Bohacec for their assistance in the preparation and the execution of wet laboratory experiments; and Miss Tracey D. Weir and Miss Veronica Yakoleff for helpful comments on the manuscript.

I would like to thank the members of the Simpson Laboratory: Kathleen Banks, Slavita Bohacec, Ying Chen, Rhonda Ellwyn, Candace E. Hofmann, Ravinesh A. Kumar, Lisa Lee, Catherine Van Raamsdonk, Tracey D. Weir, and Bibiana Wong; and the members of the Wasserman Laboratory: David Arenillas, Jochen Brumm, Alice Chou, Debra Fulton, Shannan J. Ho Sui, Carol Huang, Danielle Kemmer, Andrew Kwon, Jonathan Lim, Wynne Locke, Dora Pak, Elodie Portales-Casamar, Peter Sudmant, and Dimas Yusuf, for creating a friendly research environment. I would also like to thank the teams in the Mouse Atlas of Gene Expression Project for their support and encouragement.

I greatly appreciate the financial support that was provided by the Canadian Institutes of Health Research (CIHR) and the Michael Smith Foundation for Health Research (MSFHR) Strategic Training Program in Bioinformatics. The Mouse Atlas of Gene Expression was sponsored by Genome Canada. Without these supports, this work would not have been possible.

Finally, I would especially like to thank my family for their support and encouragement throughout my years of schooling, and church friends for their motivations.

Co-Authorship Statement

Miss Ying Chen performed RT-PCR and sequencing validations, and co-authored *Section 2.4.11 RT-PCR* and helped in the preparation of *Figure 2-7(B)*. Miss Slavita Bohacec assisted in the preparations of tissue samples for RT-PCR, and co-authored *Section 2.4.10 RNA Extraction*. Dr. Öjvind Johansson made intellectual contribution to the algorithm design. Drs. Elizabeth M. Simpson and Wyeth W. Wasserman supervised the project.

Chapter 1 Introduction

1.1 Gene Expression

The focus of this thesis is on the exploration of the transcriptome. Understanding how genes are regulated and how they are expressed is a critical step toward comprehending the transcriptome. In a typical gene expression study, one often compares different tissues or cell types, either between the same tissues under different physiological conditions or time points, or between a diseased tissue and a normal tissue. In such studies, statistical and computational methods are used to extract a set of genes that are differentially expressed or that form interesting patterns for further biological investigations.

1.1.1 Current Technologies of Gene Expression

Several technologies and their variants have been developed for gene expression experiments, including hybridization-based methods, such as microarrays, and sequencing-based methods, such as serial analysis of gene expression (SAGE) [1]. Technology based on hybridization methods, while often low in cost, requires prior knowledge of the genes being studied and the data are presented as relative levels of hybridization. In contrast, though often more expensive, sequencing-based methods, because they do not require prior knowledge of the genes being studied, offer the opportunity for the discovery of unknown transcripts. In addition, the expression levels are presented in absolute quantities. Both technologies have been intensively applied to the field of molecular biology, genomics, and medical studies, and have produced fruitful results in these fields. However, the advantage of SAGE for transcript discovery has made it the focus of this thesis.

1.1.2 Serial Analysis of Gene Expression

SAGE offers high-throughput quantification and analysis of global gene expression patterns of a particular tissue. In this technology, a short nucleotide sequence, called a *tag*, is extracted from the 3′ end of a transcript adjacent to the poly-A tail [1]. Due to modifications in SAGE protocols, a SAGE tag extracted in the original protocol is 14 bp; in LongSAGE [2], 21 bp; and in SuperSAGE [3], 26 bp. The SAGE technology relies on two basic principles. First, a short oligonucleotide sequence extracted from a position defined by a specific restriction endonuclease, the *anchoring enzyme*, typically *Nla*III, uniquely identifies the specific mRNA transcript of origin. Second, the concatenation of each of these oligonucleotide sequences allows the tags to be detected during a sequencing process in an efficient manner [4, 5]. The SAGE tags analyzed in this project were collected by the Mouse Atlas of Gene Expression Project [6] and were extracted using the LongSAGE protocol.

1.2 The Mouse Atlas of Gene Expression Project

Because of the high degree of genetic similarity to human, the mouse has emerged as a model organism for studying development and disease [7]. The Mouse Atlas of Gene Expression Project, funded by Genome Canada, aims to construct a comprehensive atlas of gene expression by using the SAGE method to explore the different stages of mouse development, from the single cell zygote to the adult. In the project, SAGE libraries are constructed for 200 tissues, often those enriched for specific cell types. In addition to these SAGE libraries, the Atlas Project has developed an open source software, DiscoverySpace [8], to provide statistical and annotation tools for manipulating gene expression datasets, especially SAGE. The Mouse Atlas Project is a public resource for basic and clinical researches for the study of genetic pathways controlling development and disease.

1.3 SAGE Tag-to-Gene Mapping

1.3.1 Methods and Problems

For a particular tissue under a specific condition, the collection of SAGE tags and their frequencies is called a *SAGE library*. The frequency of each tag reflects the abundance of its respective transcript. To analyze SAGE data, the transcript from which each tag is derived is identified, a process termed *tag-to-gene mapping* [9]. Technical details of tag-to-gene mapping are described in Chapter 2. As a sequencing-based method, SAGE is prone to sequencing errors and these errors affect the accuracy of tag-to-gene mapping. Furthermore, ambiguities also arise when a tag maps to multiple transcripts and when multiple tags map to the same transcript. Often assumptions have to be made and data cleaning is required to cope with such sequencing errors and ambiguities [10]. Tags are mapped to two types of resources, transcriptome databases and the genome. A tag that does not map to a known transcript but does map to the genome may indicate a potential novel transcript [2, 11]. Chen *et al.* [12] suggested that, in their study, 67% of tags that did not map to a transcript originated from novel transcripts.

1.3.2 Tags Spanning a Splice Junction

As a general rule, a tag that maps to a transcript will find a corresponding match in the genome of the respective organism. However, as will be described in Chapter 2, between 1.6 to 6.2% of tags span a splice junction, hence no match in the genome is observed. While the tags that map neither to the transcriptome nor to the genome may be artifacts, we hypothesize that these tags span previously uncharacterized splice junctions and represent a rich source for the discovery of novel transcripts.

1.4 Splice Junction Properties

1.4.1 Introns and Exons

One of the major differences between eukaryotic and prokaryotic genes is the presence of introns and exons. Discovered by Sambrook in 1977 [13], eukaryotic genes consist of expressed sequences, the *exons*, and intervening sequences, the *introns*. During the transcription process, both the exons and the introns are transcribed to RNA. Through a processed called *RNA splicing*, the intron sequences are removed from the recently transcribed RNA sequence. The consequence of splicing produces a continuous sequence, which is consisted of only the exons and contains information for the translation of proteins. At the junction of exons and introns where the splicing reactions occur, a conservation of sequence pattern is observed. These patterns surrounding the splice junctions, which at the 5′ end of the intron is called the *donor* and at the 3′ end is called the *acceptor*, are conserved across genes and across species. The most invariant bases are the dinucleotides on each end of an intron flanking the splice junction. At the donor end, the bases are GU, and at the acceptor end AG (the *GU-AG rule*). An additional invariant base is an A nucleotide situated in the central region of an intron. Other bases flanking the splice junctions are less conserved, but high frequencies are observed for certain nucleotides [14, 15].

1.4.2 The Splicing Reaction

Small ribonucleoprotein particles (snRNP), which are formed by complexes of protein and small nuclear RNA (snRNA), recognize the regions surrounding these invariant nucleotides. A group of snRNPs form the spliceosome, a functional unit that binds to the intron and subsequently catalyzes the splicing reaction and removes the intron. Through a transesterification reaction, one end of the intron is released from the junction and attaches to the invariant adenine nucleotide to form a lariat-like configuration. Subsequently, the lariat is released from the RNA and another transesterification reaction joins the two exons together. The splicing reactions take place in the nucleus and yield mRNA molecules from the precursor RNA [14, 15].

To ensure the accuracy of splicing, the sequences of the splice sites and the branch point are checked several times before the transesterification reactions are allowed to proceed. Nevertheless, splicing is a complex process. Stochastic events in splicing can result in unexpected forms of mRNA that serve no biological function. Furthermore, splicing errors, such as exon skipping and the use of splice sites that closely resemble true splice junctions, are often observed. [15]. These transcripts are produced sufficiently often to be detected by sensitive gene expression profiling techniques such as SAGE, and cannot be distinguished from functional transcripts based on sequence analysis.

1.5 Computational Gene Prediction

1.5.1 Current Approaches for Gene Prediction

With the ever increasing availability of genomic sequences, computational approaches have been developed for predicting potential genes. Current approaches of *in silico* gene prediction use two methods: *ab initio* and homology-based [16, 17]. *Ab initio* gene predictions rely on DNA sequence signals and nucleotide composition. This is possible because signals such as transcription factor binding sites, promoters, and translation start and stop codons typically show a certain degree of sequence conservation. Moreover, certain base combinations are usually used more frequently in coding regions. A common step in *ab initio* prediction is the use of Hidden Markov Model (HMM) to assesses the probability of the observed nucleotide usage in an exon [16, 18]. Several tools have been developed based on *ab initio* search, including GENSCAN [19], GRAIL [20], GeneID [21, 22], and FGENES [23]. Conversely, homology-based methods use known sequences as a template and make predictions for sequences that are homologous to a known gene in another organism. It is assumed that sequences that are conserved have similarly conserved function, thereby similarity to known sequences may be strong evidence of functional sequences. Programs that are based on the similarity-based method have been developed, including TWINSCAN [24], an extension of GENSCAN; SGP-2 [25], which extends from GeneID; and SLAM [26]. Predictions based solely on signal and pattern recognitions have improved over the last decade, although the accuracy varies among algorithms and organisms. Conversely, although the use of known transcripts to annotate the genomic sequence may provide higher confidence, this technique can be limiting because it is possible that genes may not have a homologous sequence known in other organisms. In this thesis, we developed an algorithm that combines the detection of sequence signals and the evidence offered by SAGE tags for the prediction of novel splice junctions.

1.5.2 Splice Junction Prediction

Gene prediction in eukaryotic organism is more complicated than in prokaryotic organisms because of the presence of introns and exons. Therefore, the identification of signals that indicate candidate splice sites, exons, and introns is, a crucial element in the approach. An *ab initio* approach to predict exons generally attempts to detect four types of signals: the translation start site, the donor splice site, the acceptor splice site, and the translation stop codon [16]. For internal exons, certain nucleotides that code for specific amino acids are also used as a measure of evaluation. One of the earliest method for splice site prediction and evaluation was the use of position weight matrices (PWM), which evaluates splice site signals by detecting nucleotide usage at specific positions [27]. Statistical models that describe the dependencies between base positions have also been studied. The gene prediction software, GENSCAN, uses a decision tree method, maximal dependence decomposition (MDD), to predict splice junctions [28]. Cai *et al.* [29] applied Baysian networks to model splice sites. A recent study predicts splice sites with dependency graphs (DG) and their expanded Baysian networks [30]. The DG model was able to achieve >90% for both sensitivity and specificity. Because nucleotides flanking the splice junctions are conserved for spliceosome recognition, in this project we adopted the PWM method to assess the quality of predicted splice junctions.

1.5.3 Position Weight Matrix

For the detection of regulatory elements, such as transcription factor binding sites (TFBS), along a stretch of DNA sequences, a commonly applied method is the use of a *motif model* [31]. A consensus sequence pattern is often observed for a common family of TFBS. Each category of binding sites often has a fixed length and specific nucleotides are used at every position. In the motif model, by using a list of transcription factor binding sites, a matrix is built to indicate the frequency of nucleotide usage at every position. The frequency matrix is then converted to a PWM for evaluating the DNA sequence of interest. During the evaluation, the weights of nucleotides, according to the weight matrix, at each position of the sequence are summed. A pre-determined threshold value is used to decide whether or not the sequence under evaluation is a consensus binding site. The PWM method for identification of DNA binding sites is generally reliable and is able to detect more binding sites than is sequence alignment methods [31]. This prediction method, however, does suffer from a high number of false positives. The PWM evaluation method has been adopted for the detection of splice site signals [27, 32] because, similar to TFBS, the sequences flanking the donor and the acceptor splice junctions are specifically recognized and bound by spliceosomes that control the splicing reactions. As suggested by Burset *et al.* [32], the PWM method can be used to predict splice junctions and can be incorporated into gene prediction programs. For my project, I have chosen to use the PWM method to evaluate tags that are predicted to span a splice junction because of its sensitivity to predict DNA binding sites. Tag sequences are additional evidence to support the splice junction predictions.

1.6 Overview of the Project

Motivated by the the potential of transcript discovery, in this thesis project, we have mapped SAGE tags that were unassigned to a known transcript or to the genome. The frequency of SAGE tags that span a splice junction in a transcriptome database was investigated. An algorithm, SAGE2Splice, was developed to identify candidate splice junctions covered by SAGE tags. Tags are split into two portions, which we termed the *edges*, and mapped to the genome within a confined distance and satisfying splice junction sequence patterns. A web interface was developed to offer this new functionality to the online community (http://www.bcgsc.ca/sage2splice). We tested the program with *spliced tags*, tags known to span a splice junction, to assess the sensitivity and the specificity, and to choose the parameters and parameter optimums for predicting candidate splice junctions. In addition, by using a different set of spliced tags, we determined that candidate tags having their predicted splice junction closer to the centre of the tag are more likely to be validated in an experiment. Using 20,000 unmapped tags taken from the Mouse Atlas of Gene Expression Project, SAGE2Splice predicted that 6% span a candidate splice junction. Twelve candidate junctions were selected, based on evidence of previously characterized exons (Type 1) and computer predicted exons (Types 2 and 3), for laboratory testing using RT-PCR and sequencing, of which nine revealed novel transcripts. The results demonstrate that SAGE tags that map to neither the transcriptome nor to the genome are a rich source for the identification of novel transcripts.

1.7 References

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484-487.

2. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE: **Using the transcriptome to annotate the genome.** *Nat Biotechnol* 2002, **20**:508-512.

3. Matsumura H, Reich S, Ito A, Saitoh H, Kamoun S, Winter P, Kahl G, Reuter M, Kruger DH, Terauchi R: **Gene expression analysis of plant host-pathogen interactions by SuperSAGE.** *Proc Natl Acad Sci U S A* 2003, **100**:15718-15723.

4. Madden SL, Wang CJ, Landes G: **Serial analysis of gene expression: from gene discovery to target identification.** *Drug Discov Today* 2000, **5**:415-425.

5. Patino WD, Mian OY, Hwang PM: **Serial analysis of gene expression: technical considerations and applications to cardiovascular biology**. *Circ Res* 2002, **91**:565-569.

6. Siddiqui AS, Khattra J, Delaney A, Zhao Y, Astell C, Asano J, Babakaiff R, Barber S, Beland J, Bohacec S, Brown-John M, Chand S, Chaters AM, Cullum R, Dhalla N, Featherstone R, Hirst M, Hoffman B, Holt R, Hou J, Kuo BY-L, Lee LLC, Lee S, Leung D, Ma K, Matsuno C, Mayo M, McDonald M, Prabhu A, Pandoh P, Ruis de Algara T, Rupert JL, Smailus D, Stott J, Tsai M, Varhol R, Vrljicak P, Wong D, Wu MK, Xie Y-Y, Yang G, Zhang I, Hirst M, Jones S, Helgason CD, Simpson EM, Hoodless PA, Marra M: **A Mouse Atlas of Gene Expression: Large-scale, digital gene expression profiles from precisely defined developing C57BL/6J mouse tissues and cells**. Submitted to Proceedings of the National Academy of Sciences.

7. Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, Hutton RD, Mullenger IR, Phillips KJ, Smith J, Stalker J, Threadgold GJ, Birney E, Wylie K, Chinwalla A, Wallis J, Hillier L, Carter J, Gaige T, Jaeger S, Kremitzki C, Layman D, Maas J, McGrane R, Mead K, Walker R, Jones S, Smith M, Asano J, Bosdet I, Chan S, Chittaranjan S, Chiu R, Fjell C, Fuhrmann D, Girn N, Gray C, Guin R, Hsiao L, Krzywinski M, Kutsche R, Lee SS, Mathewson C, McLeavy C, Messervier S, Ness S, Pandoh P, Prabhu AL, Saeedi P, Smailus D, Spence L, Stott J, Taylor S, Terpstra W, Tsai M, Vardy J, Wye N, Yang G, Shatsman S, Ayodeji B, Geer K, Tsegaye G, Shvartsbeyn A, Gebregeorgis E, Krol M, Russell D, Overton L, Malek JA, Holmes M, Heaney M, Shetty J, Feldblyum T, Nierman WC, Catanese JJ, Hubbard T, Waterston RH, Rogers J, de Jong PJ, Fraser CM, Marra M, McPherson JD, Bentley DR: **A physical map of the mouse genome.** *Nature* 2002, **418**:743-750.

8. Varhol R, Robertson N, Oveisi-Fordorei M, Fiell C, Leung D, Siddiqui AS, Marra M, Jone S: **DiscoverySpace: A tool for gene expression analysis and biological discovery**. *Poster* 2005.

9. Pleasance ED, Marra MA, Jones SJ: **Assessment of SAGE in transcript identification**. *Genome Res* 2003, **13**:1203-1215.

10. Ng RT, Sander J, Sleumer MC: **Hierarchical Cluster Analysis of SAGE Data for Cancer Profiling**. *BIOKDD01: Workshop on Data Mining in Bioinformatics (with SIGKDD01 conference)* 2001:65-72.

11. Gorški SM, Chittaranjan S, Pleasance ED, Freeman JD, Anderson CL, Varhol RJ, Coughlin SM, Zuyderduyn SD, Jones SJ, Marra MA: **A SAGE approach to discovery of genes involved in autophagic cell death**. *Curr Biol* 2003, **13**:358-363.

12. Chen J, Sun M, Lee S, Zhou G, Rowley JD, Wang SM: **Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags**. *Proc Natl Acad Sci U S A* 2002, **99**:12257-12262.

13. Sambrook J: **Adenovirus amazes at Cold Spring Harbor**. *Nature* 1977, **268**:101-104.

14. Griffiths AJF: *Modern genetic analysis : integrating genes and genomes*, 2nd edn. New York: W.H. Freeman; 2002.

15. Alberts B: *Molecular biology of the cell*, 4th edn. New York: Garland Science; 2002.

16. Baxevanis AD, Ouellette BFF: *Bioinformatics : a practical guide to the analysis of genes and proteins*, 3rd edn. Hoboken, N.J.: John Wiley; 2005.

17. Guigo R: **Computational gene identification**. *J Mol Med* 1997, **75**:389-393.

18. Durbin R: *Biological sequence analysis : probalistic models of proteins and nucleic acids*. Cambridge, UK New York: Cambridge University Press; 1998.

19. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78-94.

20. Uberbacher EC, Mural RJ: **Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach**. *Proc Natl Acad Sci U S A* 1991, **88**:11261-11265.

21. Parra G, Blanco E, Guigo R: **GeneID in Drosophila**. *Genome Res* 2000, **10**:511-515.

22. Guigo R, Knudsen S, Drake N, Smith T: **Prediction of gene structure.** *J Mol Biol* 1992, **226**:141-157.

23. Solovyev VV, Salamov AA, Lawrence CB: **Identification of human gene structure using linear discriminant functions and dynamic programming.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3**:367-375.

24. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction.** *Bioinformatics* 2001, **17 Suppl 1**:S140-148.

25. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R: **Comparative gene prediction in human and mouse.** *Genome Res* 2003, **13**:108-117.

26. Alexandersson M, Cawley S, Pachter L: **SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model.** *Genome Res* 2003, **13**:496-502.

27. Staden R: **Computer methods to locate signals in nucleic acid sequences.** *Nucleic Acids Res* 1984, **12**:505-519.

28. Brunak S, Engelbrecht J, Knudsen S: **Prediction of human mRNA donor and acceptor sites from the DNA sequence.** *J Mol Biol* 1991, **220**:49-65.

29. Cai D, Delcher A, Kao B, Kasif S: **Modeling splice sites with Bayes networks.** *Bioinformatics* 2000, **16**:152-158.

30. Chen TM, Lu CC, Li WH: **Prediction of splice sites with dependency graphs and their expanded bayesian networks.** *Bioinformatics* 2005, **21**:471-482.

31. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16**:16-23.

32. Burset M, Seledtsov IA, Solovyev VV: **Analysis of canonical and non-canonical splice sites in mammalian genomes.** *Nucleic Acids Res* 2000, **28**:4364-4375.

Chapter 2 SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice Junctions[1]

2.1 Introduction

The complexity of the transcriptome is significantly greater than that of the genome due to alternative splicing. It is estimated that between 35-65% of human genes are alternatively spliced [1, 2]. The *slo* gene, for example, is estimated to produce more than 500 distinct transcripts, which regulate various responses of the hair cells of the inner ear to sound [3]. Identification of the transcripts present within a cell can provide insights into the regulatory processes that control the cell-specific interpretation of the genome [4].

Serial analysis of gene expression (SAGE), in which a representative tag (14 to 26 bp) is excised from each transcript, is a powerful and efficient technology for high-throughput qualitative and quantitative profiling of global transcript expression patterns [5]. SAGE quantitatively measures transcript levels, providing the absolute number of each transcript-specific tag within a library of all tags. That no prior knowledge of the transcripts being studied is required makes SAGE advantageous over array-based methods for the discovery of novel transcripts [6-11].

An essential step in the analysis of SAGE data is the assignment of each tag to the transcript from which it was derived [10]. This process, termed *tag-to-gene mapping*, involves comparison of tag sequences to transcript databases. A commonly used

[1] A version of this chapter has been submitted for publication.
Byron Yu-Lin Kuo, Ying Chen, Slavita Bohacec, Öjvind Johansson, Wyeth W. Wasserman, and Elizabeth M. Simpson. (2005) SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice Junctions. Sumitted to PLoS Computational Biology.

technique is to compare SAGE tags to *predicted tags* (also known as *virtual tags*). Based on known transcript sequences, predicted tags are those expected to be generated by a SAGE protocol [12]. Often, the predicted tags closest to the 3′ end of transcripts are emphasized, as SAGE protocols impart a location bias. However, in a SAGE experiment, due to alternative splicing or incomplete enzyme digestion [13, 14], tags can be excised from other positions. The choice of sequence databases impacts the quality of tag-to-gene mapping [10]. A highly curated and more complete transcriptome database not only facilitates mapping of more tags, but also increases confidence in the mappings. Many resources have been developed for mapping SAGE tags to genes, including NCBI's SAGEmap [15], CGAP's SAGE Genie [16], the Mouse SAGE Site [17], Identitag [12], and DiscoverySpace (personal communication, Steven J. Jones, British Columbia Cancer Agency, Vancouver, Canada). Despite these efforts, however, a major problem of tag-to-gene mapping exists as ~1/3 of the tags is unmapped. Inability to map tags limits the information obtained in SAGE studies [6, 7, 10]. The identification of unmapped tags remains an active research topic in SAGE analysis.

Recent studies have attempted to map SAGE tags that did not match the known transcriptome. Chen *et al.* [18] studied 1,000 unmapped SAGE tags from publicly available libraries by generating longer cDNA fragments from SAGE tags for gene identification (GLGI), and concluded that 67% of the unmapped tags originated from novel transcripts. In an analysis of unmapped long SAGE tags (21 bp), Saha *et al.* [19] predicted 60% were from transcripts of novel genes and 40% were from unidentified internal exons of predicted genes. Gorski *et al.* [8] identified 225 cases of genes, that previously had been unidentified by gene prediction programs. Each of these studies affirmed the capacity of SAGE profiling to facilitate identification of novel transcripts.

Tags that do not map to the transcriptome or to the genome may span adjacent exons of which one or both were previously unidentified [8]. We analyzed predicted tags that had been derived from known transcripts and observed between 2 to 6% of these tags span a splice junction. Thus, even tags that do not map to the genome are anticipated to be a resource for the discovery of novel transcripts. To test our hypothesis, we developed an algorithm, SAGE2Splice, for mapping tags to potential splice junctions in a genome. Applying this new method for tag-to-gene mapping, we demonstrated that 6% of unmapped tags span candidate splice junctions. By using high sensitivity and high specificity parameters, we identified 3,458 candidate junctions for 1,212 tags from a collection of 20,000 high quality unmapped SAGE tags. Nine out of the twelve tested tag mappings were validated by RT-PCR.

2.2 Results

2.2.1 Some Predicted SAGE Tags Span a Splice Junction

We defined four distinct types of *spliced tags*, tags that span a splice junction (Figure 2-1). A *Type 0* tag matches portions of two exons at a known splice junction. Type 0 tags were identified by mapping to known transcripts. A *Type 1* tag also spans two known exons, but the junction is not present in the transcriptome databases. A *Type 2* tag spans a previously known exon and a previously unknown exon. Both Type 1 and Type 2 tags indicate a novel transcript of a previously characterized gene. A *Type 3* tag spans two previously unknown exons and indicates either two novel exons of a characterized gene, or two exons of a novel gene.

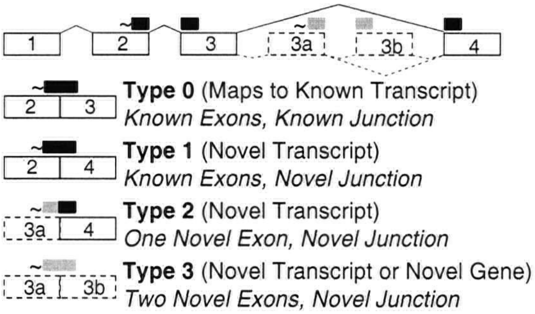
1
2
3
3a
3b
4
~
2
3
Type 0 (Maps to Known Transcript)
Known Exons, Known Junction
~
2
4
Type 1 (Novel Transcript)
Known Exons, Novel Junction
~
3a
4
Type 2 (Novel Transcript)
One Novel Exon, Novel Junction
~
3a
3b
Type 3 (Novel Transcript or Novel Gene)
Two Novel Exons, Novel Junction

Figure 2-1: Tags that span a splice junction may reveal novel genes or novel transcripts. This schematic demonstrates four known exons (1, 2, 3, and 4, boxes in solid lines). The 3′-most *Nla*III enzyme restriction site (represented as ~) lies near the 3′ edge of exon 2 and a known predicted SAGE tag (▬) spans exons 2 and 3 (Type 0 tag). Predicted exons (boxes in dashed line) 3a and 3b are examples of exons predicted by SAGE2Splice. Three other types of tags (Types 1 to 3) have been defined as potential candidates in SAGE2Splice predictions. Tag portions arising from known exons (■), whereas tag portions arising from novel exons (▨). Solid lines connecting exons indicate known combinations, whereas dashed lines indicate unknown combinations.

To determine the portion of predicted tags that span splice junctions of known transcripts, we studied RefSeq sequences. From 17,848 sequences studied, 198,419 predicted tags were extracted based on the identification of all *Nla*III restriction sites. One hundred and Ninety-three RefSeq sequences (approximately 1.08%) did not contain a *Nla*III restriction site, and thus, were unable to give rise to a SAGE tag. Among the predicted tags, 12,297 (6.2%) overlapped a splice junction (Type 0). In addition, 14 predicted tags traversed two splice junctions (Table 2-1). These were due to very small exons [20], between 1 bp to 4 bp in length. Since the SAGE technique excises tags from the *Nla*III restriction site closest to the 3′ end of transcripts, from the RefSeq sequences, 17,655 predicted tags were extracted from the 3′-most position and investigated. Among these predicted tags, only 292 (1.6%) were Type 0. The different Type 0 frequencies between the all-position set and the 3′-most set reflects that exons are generally longer at the 3′ end of a transcript [20]. In the analyzed RefSeq sequences, the average length of all exons was 262 bp, whereas the average for all 3′-most exons was 1,068 bp. Hence, at the 3′-most position, the probability of finding a splice junction within a tag is lower than that from the set of all *Nla*III positions.

Table 2-1: 6.2% of predicted tags from all *Nla*III restriction sites, and 1.6% from 3′-most sites were found to span a known splice junction (Type 0 tags).

Tag Position	Number of Predicted Tags[1]	Number of Type 0 Tags	Number of Tags Spanning Multiple Junctions
All *Nla*III	198,419	12,301 (6.2%)	14
3′-most *Nla*III	17,655	283 (1.6%)	1

[1] Curated RefSeq cDNA collection was analyzed to detect *Nla*III restriction sites and the downstream 17 bp sequences (predicted SAGE tags). Predicted tags were extracted from UCSC Annotation Database (July 16, 2004).

2.2.2 Intron Properties

In our development of SAGE2Splice, an important search criterion was to determine the maximum length the algorithm should allow for candidate introns. Previous studies have shown that, although a typical intron is 40-125 bp in length, the average length is approximately 1,000 bp because the sizes of introns vary over a very wide range [20, 21]. In our studies of the RefGene annotations, we confirmed that within the known splice junctions, introns vary from 6 to 1,195,292 bp in length, with a median of 1,271 bp (Figure 2-2). Ninety percent of introns were smaller than 10,000 bp and 95% were smaller than 20,000 bp. We incorporated 10,000 bp as the default for maximum intron size in the search for candidate splice junctions.

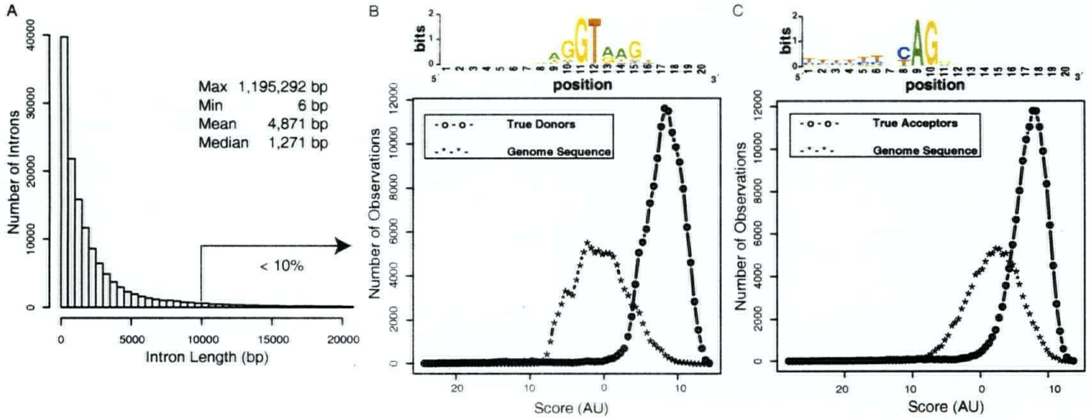

Figure 2-2: Length and boundary nucleotides of introns are important properties for detecting a splice junction. (A) Less than 10% of introns in RefGene annotation were greater than 10,000 bp in length. (B) and (C) a position weight matrix (PWM) for splice junctions was applied to true splice junctions defined by RefGene annotations and to randomly selected genome sequences containing the canonical dinucleotide pair at the appropriate position. The scores, which were computed based on the profile model, for donors and acceptors were plotted and showed that true splice junctions acquired high scores. The information content and the relative frequency of nucleotides at each position are measured in bits (vertical axis of the sequence logo diagrams) to indicate the strength of signals. Two bits of information are required to determine the content of a DNA sequence. AU: arbitrary units.

To gain a more detailed understanding of the sequence patterns of splice junctions, we examined 10 bp flanking each side of the donor junctions and 10 bp flanking each side of the acceptor junctions. For each junction type, we constructed a matrix representing the frequency of each nucleotide at each position. Position weight matrices (PWM) were constructed by converting the frequencies into scores relative to the expected frequency of a randomly selected nucleotide (see Materials and Methods). By using these scoring matrices, we generated genuine score distributions for true splice junctions in RefSeq and empirical score distributions for randomly selected sequences from the genome. By superimposing the genuine distribution on the empirical distribution, it was shown that genuine splice junctions typically had high scores and were located on the far right end of the empirical curve (Figure 2-2). Hence, we incorporated these properties into our SAGE2Splice algorithm for ranking and determining the likelihood of candidates.

2.2.3 The SAGE2Splice Algorithm

2.2.3.1 Pre-processing the Input SAGE Tags

In a 21-bp SAGE tag, if a splice junction exists within the sequence, one of the two portions is no shorter than 11 bp in length. Each 21-bp tag is therefore split into two equal portions of 11 bp (overlapping by one bp), which are used as search strings simultaneously. We term these equal-sized portions as the *halftags*. Prior to a search, complementary sequences for the halftags were constructed because genes can be located on either strand of the genome. The program reads the sequences of each chromosome one segment of 100,000 bp at a time. To perform a complete search, the algorithm holds three such segments in memory at any one time: the previous segment, the current segment, and the next segment. Searching for a candidate splice junction in SAGE2Splice consists of three progressive levels (Figure 2-3). At each level, only if the defined matching criteria are fulfilled will the algorithm proceed to the next level. Otherwise, the algorithm imports a new segment of the genome into memory, and the search starts over from the first level.

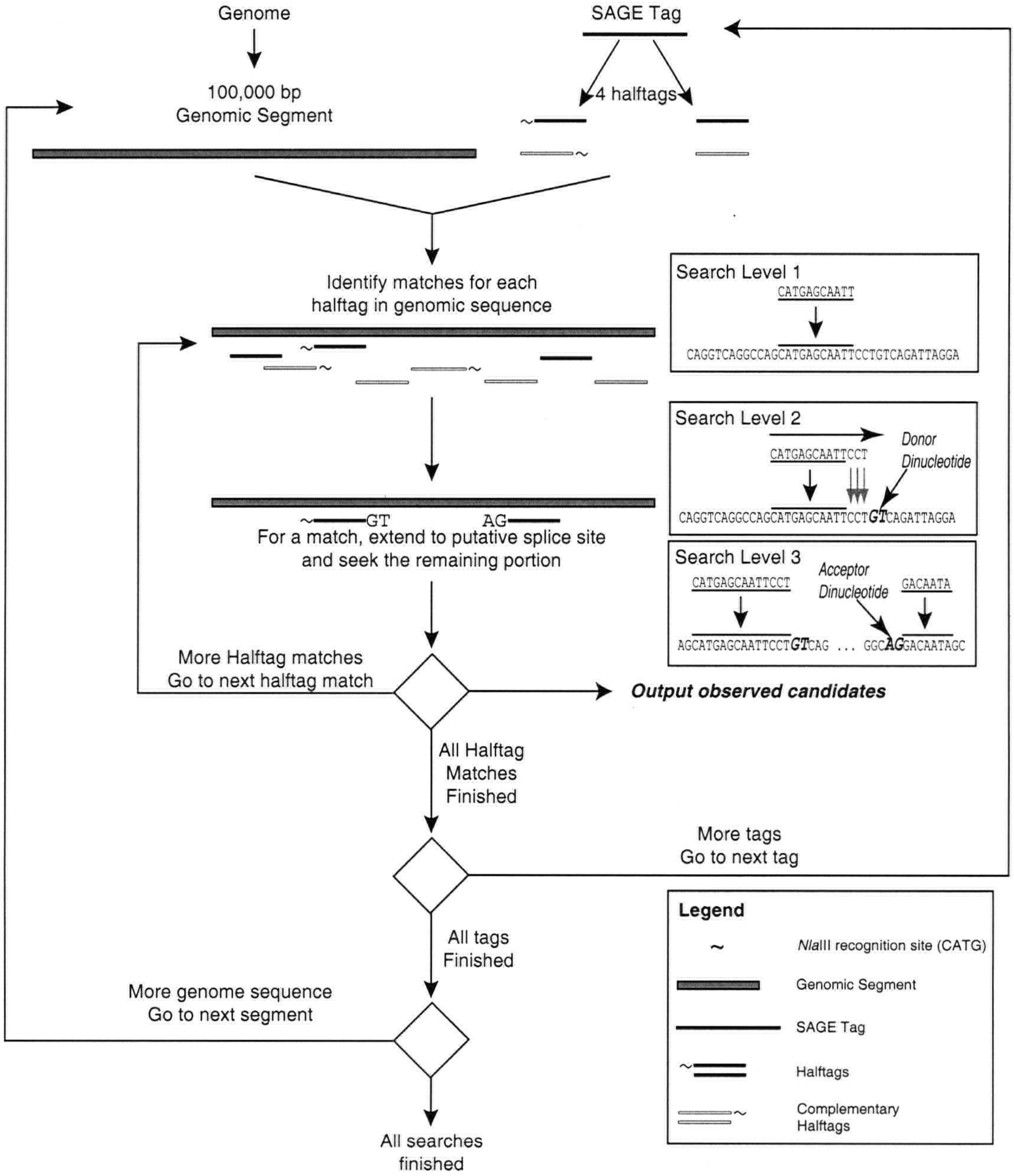

Figure 2-3: SAGE2Splice algorithm searches the genome for novel splice junctions. By splitting each tag into 2 halftags and making complementary copies, the algorithm searches for candidate splice junctions against continuous segments of the genome in three progressive steps. After each level, if the matching criteria were fulfilled, the algorithm would go on to the next level. If criteria were not fulfilled, the algorithm would analyze the next tag. Once all tags have been analyzed, the next genomic segment is read and the algorithm returns to the first level.

2.2.3.2 Search Level 1: Matching Halftags

In Search Level 1, SAGE2Splice searches each halftag against the current segment by using the pattern-matching function built into the Perl programming language (version 5.6). Positions of all matches are stored as a tab-delimited string. A complementary halftag match, indicating a position on the complementary strand, is stored as a negative position. If at least one halftag match is found, the algorithm proceeds to Search Level 2. Otherwise, the next segment of the chromosome is imported and the search for candidate splice junctions returns to Search Level 1.

2.2.3.3 Search Level 2: Extending Halftags

SAGE2Splice searches for one boundary of a potential candidate intron before searching for the other boundary. During Search Level 2, SAGE2Splice attempts to find, for each halftag match, one of the edges of a potential intron. From Search Level 1, a 5′ halftag match to the genomic segment indicates a search of a potential donor intron-exon boundary in Search Level 2. Conversely, a 3′ halftag match suggests a search for the acceptor boundary. Hence, in the second level, the SAGE2Splice algorithm extends the first level halftag match, base-by-base against the original tag. At every base extension, depending on whether or not the halftag match is 5′ or 3′, the respective intron boundary dinucleotide is added and matched to the genome segment. As a result, all potential candidates for one edge of an intron are discovered for every halftag match. For the 5′ halftag match, the extension is toward the 3′ end and the donor dinucleotide is GT, whereas for the 3′ halftag match, the extension is toward the 5′ end and the acceptor dinucleotide is AG. A match of the complementary halftags indicates a potential candidate on the complementary strand of the genome sequence and, thus, the base extension direction is opposite that of the sense strand. If a potential intron-exon boundary is found, the algorithm continues to Search Level 3. Otherwise, SAGE2Splice reads the next genomic segment and returns to Search Level 1.

2.2.3.4 Search Level 3: Searching Remaining Portions

In Search Level 3, the remaining tag portion for the corresponding candidate splice junction is sought within 10,000 bp, or a maximum distance set by the user. If the preceding level found a candidate donor junction, the search looks for candidate acceptor junctions with the conserved dinucleotide, AG, toward the 3′ direction, in accord with the definition of splice junctions [21]. If, on the other hand, the previous search returned a candidate acceptor junction, the search for candidate donors is toward the 5′ direction and the conserved dinucleotide is GT. Searches for the remaining tag portions for the complementary halftag are in the opposite direction. When a candidate splice junction is returned, the algorithm proceeds to scoring and ranking the candidate. Because a match in Search Level 1 could be close to the edges of the current genomic segment, having the previous and the next segments in memory allows for potential matches located beyond the current segment. If, however, Search Level 3 does not return a candidate splice junction, the search returns to Search Level 1 to start on a new segment of the chromosome.

2.2.4 Scoring Candidate Splice Junctions

Once a candidate is discovered and returned by Search Level 3, for both the donor and the acceptor, 10 bp flanking each side of the boundary are extracted and evaluated using the respective PWM. Probability values (p-values) are generated by comparing the observed scores against empirical score distributions. For a tag that matches multiple candidates, SAGE2Splice ranks the candidates according to the composite p-value. After this process, SAGE2Splice returns for each candidate the following information to the user: the chromosome number; the two tag portions with their positions, scores, and p-values; the composite p-value; and the predicted intron length.

2.2.5 Efficiency Tuning of SAGE2Splice

Five parameters affect the performance of SAGE2Splice, including the number of SAGE tags in the search, the length of SAGE tags, the cutoffs for p-values, the cutoff for maximum intron length, and the length of genomic segment in memory. Other than the length of genomic segment in memory, all factors depend on either the input SAGE tags or user-specified parameters. We investigated the use of genomic segments of different lengths to fine-tune SAGE2Splice for best performance (Figure 2-4). The total execution time of SAGE2Splice decreased until it reached a segment size of 100,000 bp, and linearly increased thereafter.

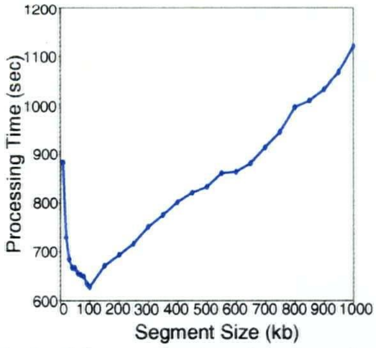

Figure 2-4: SAGE2Splice was optimized for processing time by using different genomic segment lengths (ranging from 10 kb to 1000 kb). For SAGE2Splice performance, 100 kb was determined as the optimal size.

2.2.6 Sensitivity and Specificity

To test the accuracy of SAGE2Splice and determine the optimal parameter settings, we investigated the sensitivity and the specificity for various p-value cutoffs, ranging from 0.00001 to 1. The receiver operating characteristic (ROC) curve demonstrates a tradeoff between sensitivity and specificity (Figure 2-5). As we varied the overall p-value cutoffs, it was observed that to achieve a specificity of close to 95%, sensitivity dropped to 55%. The ROC curve shows that, although SAGE2Splice can achieve high sensitivity, specificity suffers dramatically at such settings. Moreover, the positive predictive value, which indicates the proportion of the candidates that are true positives, decreases as the p-value cutoffs increase (Figure 2-5). Such results correspond to previous studies [22, 23] that showed that true splice junctions acquire high profile scores in the evaluation scheme and, thus, candidates with lower p-values are more likely to be true. In the ROC curve, the point with the minimum number of misclassified candidates (defined by a tangent line for which the slope equals 1) occurs when the composite p-value cutoff is approximately 0.0025, leading to a sensitivity (true positive rate) of 0.9 and a specificity of 0.82 (false positive rate = 0.18) (Figure 2-5). Similarly, separate analyses of the donor junction and the acceptor junction revealed the optimal cutoffs to be 0.06 and 0.15, respectively.

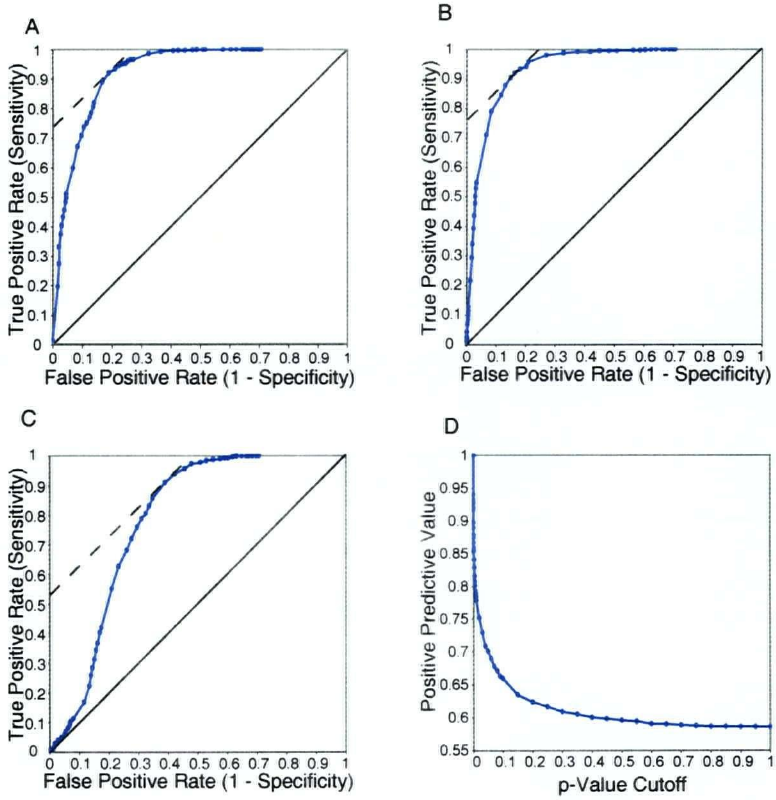

Figure 2-5: SAGE2Splice achieves high sensitivity but relatively low specificity. (A) The area under the receiver operating characteristics (ROC) curve is 0.9232, indicating a candidate found by SAGE2Splice was much better than expected by random chance. Conversely, to achieve high specificity, the sensitivity (true positive rate) was significantly compromised. The tangent point of the dashed line is the optimal point when the costs of misclassifying positive and negative candidates are equal. This point corresponds to a p-value cutoff of 0.0025. (B) Analysis of the ROC curve for the donor splice junctions indicates a cutoff p-value of 0.06 as the optimal point. (C) For the acceptor splice junctions, the optimal cutoff p-value is determined to be 0.15. (D) The positive predictive value indicates that a high probability (greater than 0.9) of correct predictions requires a restrictive p-value (less than 0.0001).

2.2.7 Edge Length and Rank Analysis

To analyze the relationship between search accuracy and the position of a splice junction within a junction-spanning tag, we obtained EST transcript annotations from the UCSC Genome Browser and extracted *Type 0* predicted tags that had GT and AG for the donor and acceptor boundary dinucleotides, respectively, and had introns between 50 bp (minimum imposed to avoid gaps in annotation) and 10,000 bp in length. Among the

200,000 unmapped SAGE tags in the Mouse Atlas of Gene Expression Project (detailed below) [24], 261 such tags, which did not map to RefSeq, Ensembl, MGC, or the mouse genome, were found to match these EST predicted tags. These 261 tags are distinct from the transcript dataset used in initial parameter selection and junction profile model building, thus providing an independent test set. For each splice junction position within the tags, the percentage of tags correctly mapped by using the optimal p-value cutoff values was determined (Figure 2-6). As illustrated, a minimum length of 5 bp for the shorter edge produces reliable predictions. In many cases, a laboratory researcher is prepared to test multiple candidate predictions. Therefore, we investigated, for each length, the number of top ranking candidates required to detect a true junction (Figure 2-6). The closer a splice junction is to the centre of the tag, the fewer candidates are required to find a validated result. For each tag, by testing the candidate with the lowest p-value, investigators can expect 90% of tags to be mapped successfully, if the junction is at least 5 bp from the edge of the tag.

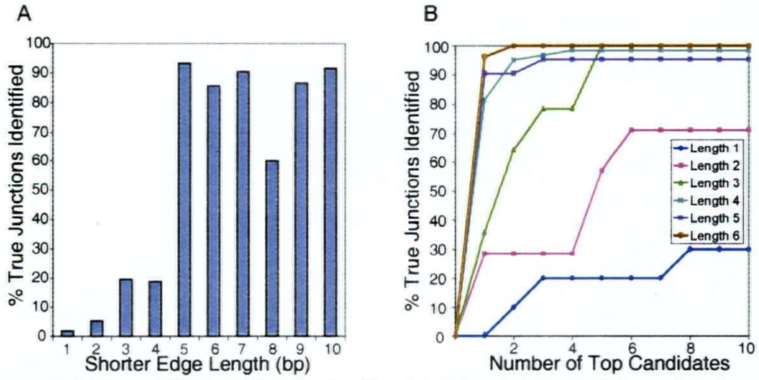

Figure 2-6: The probability of finding the true splice junction is lower if the splice junction is located closer to the edge of a tag. By using the unmapped tags in the Mouse Atlas Project that map to spliced tags predicted from EST transcripts, the percentage of true splice junctions found was analyzed for each short edge length. (A) By using high specificity parameters (cutoffs of 0.06, 0.15, and 0.25 for donor, acceptor, and composite p-values, respectively), 93% of the true splice junctions were found when the shorter edge is ≥5 bp in length. (B) With no p-value cutoffs, 90% of the true splice junctions were found with the top-ranked p-value when the shorter edge is 5 bp in length.

2.2.8 Unmapped Tag Search Results

Exhaustive mapping of the SAGE tags in the Mouse Atlas of Gene Expression project [24] resulted in 200,000 unmapped tags. From these unmapped tags, SAGE2Splice was applied to 20,000 of the highest quality SAGE tags from this set. (see Materials and Methods). There were 7,757 splice junction candidates (0.38785 per tag) found to fulfill the p-value thresholds of 0.06, 0.15, and 0.0025 for the donor, the acceptor, and overall, respectively (maximum intron length was set at 10,000 bp). Among the 1, 639 (8.2%) tags that were found to have candidate junctions, we observed that a few tags mapped to multiple candidate sites. Among the 20,000 SAGE tags in the search, six returned more than 100 candidate junctions, 90 returned between 10 and 100 candidates, 113 returned between 5 and 10 candidates, 271 returned 2 candidates, 939 returned 1 candidate, and 18,361 matched no candidate.

Perl scripts were written to computationally classify the candidates into tag types. Based on matching both donor and acceptor positions to the UCSC annotation databases, 15 candidate junctions corresponded to Type 1 tags. There were 803 junctions, corresponding to Type 2 tags, for which either only the donor position or only the acceptor position matched a known exon. The remaining 6,939 candidate junctions matched no known exons and were associated with Type 3 tags. By mapping candidates corresponding to Type 2 and Type 3 tags to exons predicted by GenScan, TwinScan, or SGP, five candidates that matched Type 2 tags and three candidates that matched Type 3 tags were further categorized as prediction supported. Based on RNA sample availability, we picked eight candidates from the Type 1 category, two candidates from the Type 2 category, and two candidates from the Type 3 category for RT-PCR testing (Table 2-2).

Table 2-2: Twelve candidates were selected for RT-PCR validation.

ID[1]	Chr	Donor Match	Donor Position	Acceptor Match	Acceptor Position	Intron Size	Composite p-Value[2]	Gene Name[3]	RT-PCR Validation.[4]	Accession Number[4]
1-1	1	CATGGTGAAGCTCGCAAAG	86244556	GA	86238632	5924	2.2 E -06	*Ncl*	✗	ND
1-2	1	CATGGTGAAGCTCGCAAAG	86244556	GA	86240496	4060	2.2 E -05	*Ncl*	✗	ND
1-3	4	CATGTAGTGTTTG	117657859	AATGTTCC	117656489	1370	9.2 E -05	*Ppih*	✓	DQ113644
1-4	5	CATGTCCCTCAAG	126140225	GTGTTCTC	126134146	6079	1.6 E -05	*AK081926*	✓	DQ113645[5]
1-5	10	CATGAGAGCGAAG	128675985	GCTGAAGC	128675467	518	5.3 E -06	*Rpl41*	✓	DQ113647
1-6	14	CATG	20780218	CCAAAGGAGTAGATCTG	20785233	5015	4.9 E -05	*Rps24*	✗	ND
1-7	19	CATGCGAGCTG	6710208	GCATTCGTCC	6711938	1730	9.6 E -06	*Tpt1h*	✓	DQ113648
1-8	X	CATG	124592868	GAAAGCGGCGTTACGAC	124593658	790	6.5 E -06	*Rpl136a*	✓	DQ113649
2-1	4	CATG	132062103	GAGGACACTTGTCAGGA	132060011	2092	2.0 E -05	*Ccs*	✓	DQ113650
2-2	11	CATGCAGGGTGATG	75371984	ATTCCTA	75375252	3268	3.7 E -04	*Ywhae*	✓	DQ113651
3-1	4	CATGCCCAG	135998365	GTCCACGGCTCC	135998673	308	3.0 E -04	*s2sEMS1*	✓	DQ113652
3-2	13	CATGGACAT	111936186	ATTCCTTTTGCC	111933949	2237	2.5 E -04	*s2sEMS2*	✓	DQ113653

[1] The first digit of the ID indicates the type of tag. The second digit is a sequential number.
[2] A *Composite p-value* was computed as the product of the donor p-value and the acceptor p-value.
[3] All selected candidates fulfill cutoffs of 0.06, 0.15, and 0.25 for donor, acceptor, and composite p-values. Gene Ontology names were assigned to Types 1 and 2 candidates. Candidate 1-4 did not match to a characterized gene. Accession number of the matched mRNA transcript was assigned. Gene names for candidates 3-1 and 3-2 were assigned by this project.
[4] ✓, as predicted; ✗, not as predicted; ND, not done. For sequences that corresponded to the predicted transcript, a GenBank Accession number is assigned.
[5] Candidate 1-4 generated two strong RT-PCR bands, one an unpredicted novel transcript (DQ113646).

2.2.9 Candidate Validation

For the selected candidates, primers were designed based on the contiguous exons predicted by SAGE2Splice (Table 2-3). RT-PCR results showed that nine of the twelve tested candidates generated products of the predicted length (Figure 2-7). The other three candidates produced bands that were larger than expected. The latter were candidates that had their splice junctions located close to the edges of the SAGE tags. However, 2 of the 9 candidates did have the correct band sizes, even though they had their splice junction located only 4 bp away from the tag edge. Sequencing results of the RT-PCR products matched the expected sequences. Two strong bands were observed for candidate 1-4, one that matched the size of the expected length (221 bp) and the other one larger (361 bp). Sequence of the expected band corresponded to the novel alternative combination predicted; sequence of the larger product revealed an unpredicted, previously unidentified alternative transcript of the same gene. Unpredicted larger bands were also observed for candidates 1-7 and 1-8 (306 bp and 197 bp, respectively) and corresponded to known transcripts.

Table 2-3: RT-PCR primers were designed for the selected candidates based on sequences of the two predicted exons.

ID	Tissue	Forward Primer (name)	Reverse Primer (name)	Product Size bp
ß-actin	All tissues used	GCATGGGTCAGAAGGAT (oEMS1507)	CCAATGGTGATGACCTG (oEMS1508)	615
1-1	P84 Days Visual Cortex	TGAGCTCTTCCGAGCTGCT (oEMS2184)	GTGAAACAGATCGTCCATCAA (oEMS2185)	165
1-2	P84 Days Visual Cortex	TGAGCTCTTCCGAGCTGCT (oEMS2184)	TGCCAAACACTTTTAAACCAG (oEMS2186)	153
1-3	E11.5 Days Whole Head	CAAACAGTGGTCCCAGTACAA (oEMS2156)	GCCTGTGGGAACATTCAAA (oEMS2157)	102
1-4	P27 Days Visual Cortex	AAGGAAGATGGCGAAGACAGT (oEMS2152)	AGGGGAGGCTCATCTTCTGAA (oEMS2153)	215
1-5	E11.5 Days Whole Head	CATGAGAGCGAAGGCTGAA (oEMS1650)	TGAGACTCATTACCGATGGCA (oEMS2149)	157
1-6	P84 Days Visual Cortex	TGCGCGTTGATATGATTGGT (oEMS2176)	GCAGACGTGTAGGAGCTTTTT (oEMS2177)	168
1-7	P84 Days Hypothalamus	CCGAAATGTGCAGCTGTCTAA (oEMS2160)	TAGGGGTCCATCGATGAACA (oEMS2161)	127
1-8	P84 Days Visual Cortex	GCTCCTGCGAACATGGAAA (oEMS2180)	TTGCGGAAAATAGGCTTAGTC (oEMS2181)	79
2-1	P20 Days Visual Cortex	ATCACCAACTGCTGTGCTGTG (oEMS2168)	AGATGGCAAAGTCCTGACAA (oEMS2169)	172
2-2	E17.5 Days Skeletal Muscle	AGCAGCTTTTGATGACGCAA (oEMS2164)	TTAGGAATCATCACCCTGCA (oEMS2165)	136
3-1	P21 Days Uterus	ATAGAATCCTCGTCGCCATC (oEMS2174)	ACAACAATGGAAGCCTCCTT (oEMS2175)	233
3-2	P42 Days Visual Cortex	CCGTGAGAGTGACTTTGGATT (oEMS2172)	AACCACTGTCCGGGTGTTGTA (oEMS2173)	263

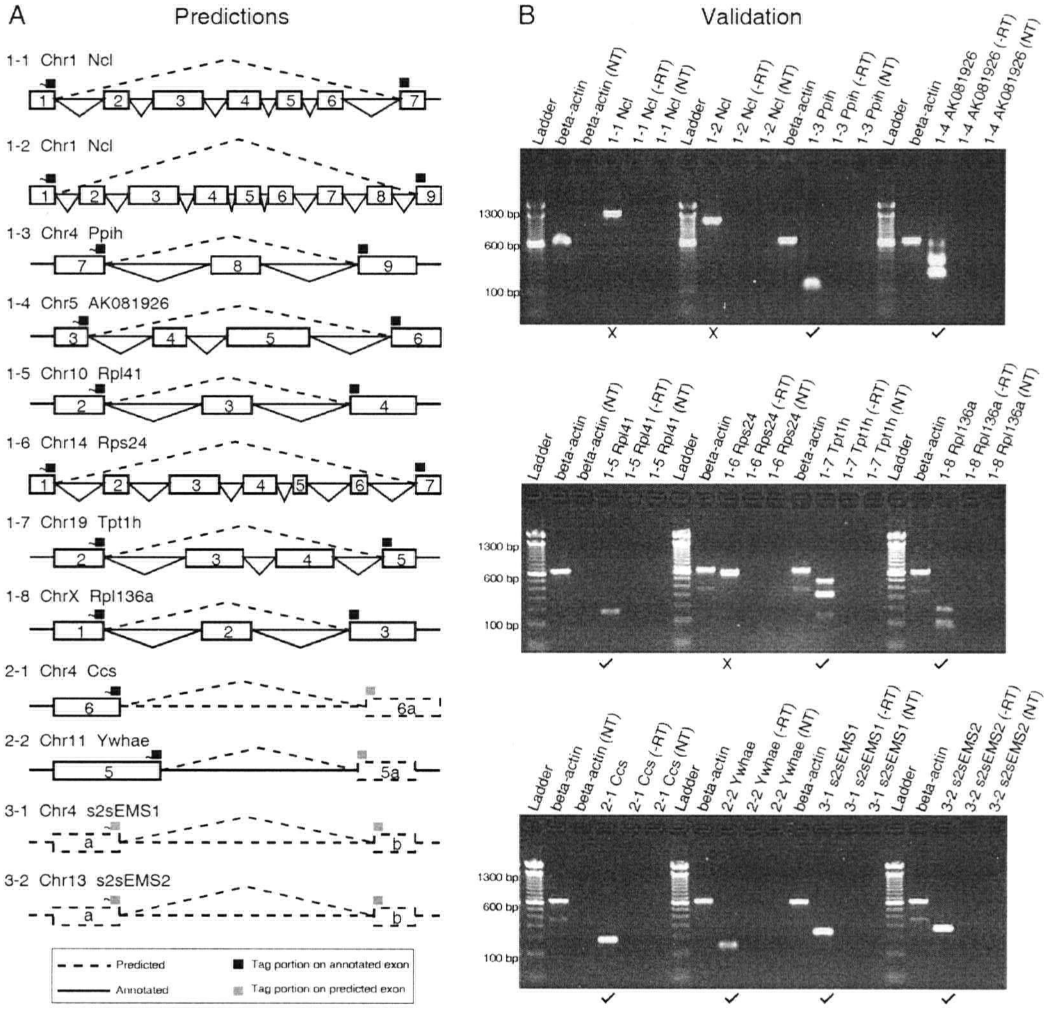

Figure 2-7: Nine of twelve selected candidates revealed novel splice junctions by RT-PCR and sequencing. (A) Predicted splice junctions of the 12 selected candidates. First digit of the candidate ID indicates the tag type; the second digit is arbitrarily assigned. (B) Except for Candidates 1-1, 1-2, and 1-6, all candidates show the correct product size and were sequence validated. A larger band from an unpredicted novel splice junction was also observed for candidate 1-4. Larger bands were also observed for candidates 1-7 and 1-8, but were shown to be known splice variants. Candidates that were validated by RT-PCR and by sequencing are indicated in ✓ under the respective lane; candidates not validated, by X. NT: negative control with no RNA template; -RT: negative control with no reverse transcriptase.

We computationally predicted the longest open reading frames (ORFs) within the RT-PCR and sequencing-validated candidates based on the sequence information of the two exons . Candidates 1-3 and 2-2 encoded short alternative C-terminal sequences (Table 2-4). Candidates 1-5, 1-7, and 1-8 contain alternative ORFs. ORFs were predicted for the novel sequences in candidates 1-4, 2-1, 3-1, and 3-2. Protein-protein BLAST (BLASTP) to all NCBI all organism non-redundant database showed no significant matches for candidates 1-3, 1-4, 1-5, 1-7, 1-8, and 2-2. However, candidate 2-1 matched a dog zinc finger DHHC domain containing protein. Candidates 3-1 and 3-2 showed significant similarities to rat proteins. Significant matches to known proteins in a different organism are strong evidence that these three predicted transcripts are functional.

Table 2-4: Open reading frame and BLASTP analyses of the RT-PCR and sequencing validated candidates.

ID	ORF[1] Impact	BLASTP[2,3] Results for New Sequence
1-3	Alternative C-terminus Pre-mature stop	No significant match
1-4	ORF predicted with stop codon ≥ 95 amino acids	No significant match
1-5	Alternative ORF 32 amino acids	No significant match
1-7	Alternative ORF without stop codon ≥ 22 amino acids	No significant match
1-8	Alternative ORF without stop codon ≥ 23 amino acids	No significant match
2-1	ORF predicted with stop codon ≥ 46 amino acids	Match to *Canis familiaris* zinc finger DHHC domain containing protein (XP_854957.1)
2-2	Alternative C-terminus Pre-mature stop	No significant match
3-1	ORF predicted with stop codon ≥ 79 amino acids	Match to *Rattus norvegicus* heparin sulfate proteoglycan 2 (XP_233606.3)
3-2	ORF predicted without stop codon ≥ 88 amino acids	Match to *Rattus norvegicus* integrin alpha 1 (NP_112256.1)

[1] ORF, open reading frame.
[2] BLASTP, protein-protein BLAST versus NCBI nr (all organisms) database (September 7, 2005)
[3] Similarity on both predicted exons is required for a significant match.

2.3 Discussion

We have developed a tool, SAGE2Splice, for efficient mapping of SAGE tags to potential splice junctions in a genome. By using a scoring system that generates a probability value for each candidate splice junction, SAGE2Splice allows users to assess the quality of the candidates. Furthermore, the *in silico* validation pipeline automatically classifies the candidates into three categories, based on overlaps with annotated and predicted exons. We identified candidate junctions for 1,639 unmapped tags, using parameters designed for high specificity. This is the first attempt to investigate systematically SAGE tags that span splice junctions and to use this characteristic for transcript identification. The online version of SAGE2Splice allows users to search the genome sequences for human, mouse, rat, and worm, the four most common organisms in NCBI's SAGE database. All source code and data are available for download from the SAGE2Splice website.

Scanning a genome for potential splice junctions is computationally challenging. The mouse genome, roughly 3 Gb, takes on the order of several minutes to scan. Disk access dominates the running time when the number of input tags is low. As the number of input tags increases, the search time becomes dominant. Due to the increased probability of observing halftag matches that trigger more computationally intensive searches, longer maximum intron length settings increase run time. The time efficiency of SAGE2Splice is *O(nm)*, where *n* is the number of input tags and *m* is the size of the genome. Since SAGE2Splice reads and keeps only a fixed length of genomic segment in memory at any time, memory usage is minimal. Memory space is dependent on the number of input tags, and, thus, is *O(n)*, where *n* is the number of input tags.

The portion of tags corresponding to splice junctions in a SAGE library is unclear. Incomplete enzyme digestion or alternative splicing at the 3′ end of a transcript could give rise to multiple tag types from the same gene [13]. Thus, we expect the portion of spliced tags in a SAGE experiment to be higher than 1.6%, which was based on predictions from the 3′-most tags in RefSeq transcripts, but lower than 6.2%, which was based on predicted tags from all positions. Among the high sequence quality and highly expressed unmapped tags, the portion of spliced tags is expected to be higher. In our analysis of such unmapped SAGE tags, 8.2% matched a candidate splice junction when high specificity parameters were used.

As in other studies [22, 23], we adopted PWM profiles for splice site detection. In addition, SAGE2Splice uses tag sequence as support and includes a criterion for the presence of the canonical dinucleotide prior to scoring the candidates. This heuristic requirement for the canonical dinucleotide pair limited our searches to about 96.27% of potential splice junctions (according to known splice junctions in RefSeq annotation). In the future, we would like to incorporate methods such as decision trees into our splice junction evaluation scheme and, thus, allow SAGE2Splice to detect non-canonical candidate junctions.

Nine of the twelve laboratory tested candidates confirmed predicted novel splice junctions. Based on these results, we showed that SAGE2Splice is a potent tool for computational prediction of novel splice junctions using unmapped tags. Furthermore, the results indicate that unmapped SAGE tags represent a rich resource for the discovery of novel transcripts.

A minimum edge length from the splice position to the closest edge of a tag is required for reliable predictions. Of the tested candidates, the two with a shorter edge of 2 bp and the one with a shorter edge of 4 bp were not detected by RT-PCR. Conversely, all candidates with a splice junction closer to the centre of the tag were confirmed by RT-PCR. For candidates that have their predicted splice junction closer to the tag boundary, less support from the tag sequence was given, and thus less confidence can be assessed from the p-value. Based on our data, we recommend eliminating candidates with edges less than 5 bp. Applying this recommendation to our data will result in a prediction of 1,212 (6%) spliced tags (3,458 candidate junctions). Based on the evidence in edge length and rank analyses, we think this 6% of tags represents a list of reliable predictions.

In conclusion, we have developed an algorithm that uses unmapped SAGE tags to search for candidate splice junctions. We validated nine of the twelve tested candidates by RT-PCR. As the annotation of genomes and the characterization of genes and transcripts continue, systematic exploration of candidate novel splice junctions through the use of SAGE2Splice will help elucidate the transcriptome.

2.4 *Materials and Methods*

2.4.1 Source of Transcripts and Known Splice Junctions

The genomic sequences of C57BL/6J mouse (mm5, May 2004) and the RefGene annotation database of RefSeq transcripts (July 16, 2004) were obtained from the University of California Santa Cruz (UCSC) Genome Browser [25]. Sequences in RefSeq are considered to be high quality because they have been examined and curated by experts [26]. The UCSC genome annotation pipeline maps the transcript sequences to the mouse genome and identifies the exon coordinates.

For each transcript, the RefGene annotations include the chromosome, the orientation, the exon coordinates, and the translated region coordinates. Based on this information, we developed programming scripts in the Perl language (version 5.6) to re-construct the RefSeq sequences from the mouse genome sequence. These re-constructed RefSeq sequences enabled us to examine the boundary patterns of each splice junction, as well as to analyze the predicted SAGE tags and the number of *Type 0* tags.

2.4.2 Extraction of Predicted SAGE Tags

We computationally extracted, from the RefSeq transcript sequences, all *predicted SAGE tags*, by obtaining 21 bp (LongSAGE) downstream of each *Nla*III anchoring enzyme restriction site. Each predicted tag was annotated with its distance from the 3′ end, which was given the position 0.

2.4.3 Scoring Splice Junctions

For each observed splice junction, we examined the window of 10 bp on either side. By counting the occurrences of each nucleotide at every position, frequency matrices were constructed for donor and for acceptor patterns. Assuming that in a random sequence all four nucleotides have equal probability, we converted these matrices, for every nucleotide at every position, to position weight matrices (PWM) [27] by using the formula $S_{pos} = \log_2\left(\frac{frequency}{0.25}\right)$. For each donor and acceptor junction, 10 bp from each side of the boundary were extracted and, by using their respective PWM, a score was computed as $score = \sum_{pos=-10}^{pos=10} S_{pos}$. To generate empirical score distributions for p-value assignments, 100,000 sequences, which were 20 bp in length and had G and T at the 11th and 12th positions, were randomly selected from the genome and each were scored by the donor PWM. Similarly, 100,000 sequences of 20 bp containing A and G at the 9th and 10th positions, were selected and scored by the acceptor PWM. For each candidate intron, the proposed donor and acceptor junctions were scored separately, according to their respective matrices. A p-value was assigned based on the relative position of the observed score on the junction's empirical distribution. Assuming independence, a composite p-value was computed as $p(Donor, Acceptor) = p(Donor)p(Acceptor)$.

2.4.4 SAGE2Splice Implementation and Features

The core program of SAGE2Splice was written in the Perl programming language (version 5.6), and executed by using a compiled version to increase performance. An Internet interface was created by using the PHP scripting language. In addition to providing a list of SAGE tags as inputs, the user has the options of specifying the following: the anchoring enzyme recognition sequence (default is *Nla*III, CATG), the maximum intron size (default is 10,000 bp), and the cut-off p-values for the donor candidate, the acceptor candidate, and the composite candidate (defaults are 0.06, 0.15, and 0.0025, respectively). The implementation of SAGE2Splice allows the user to adapt to different organisms simply by modifying the configuration file. The SAGE2Splice program and the web interface PHP script, are available for download (http://www.bcgsc.ca/sage2splice/).

2.4.5 Efficiency Tuning of SAGE2Splice

We tested a series of different genomic segment size settings to find an optimal size for computational efficiency. Tested sizes include: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000 kbp. For each size, we performed five iterations of the SAGE2Splice algorithm to search for 10 randomly selected SAGE tags, and recorded the average execution time in seconds. Efficiency analysis was performed on a 14 node cluster, in which each node had two Intel© Xeon processors at 2.4 GHz with 1.5 GB random access memory running RedHat© Linux version 7.3. Perl version 5.6 was used to compile the core SAGE2Splice program.

2.4.6 Sensitivity and Specificity

We randomly chose from the list of predicted tags 1,000 tags that were known to span a splice junction to have GT and AG as the junction dinucleotide pairs, and to have the introns within 10,000 bp of each other, as our positive controls for testing SAGE2Splice. By searching against the corresponding genome using SAGE2Splice, *true* *positives* (TP) were identified if the original splice junctions were found, and *false negatives* (FN) were identified if no known splice junction was found. For negative controls, we chose from the same predicted tag lists, 1,000 tags that were known not to traverse a splice junction. A *true negative* (TN) evaluation is when no candidate was output by SAGE2Splice, whereas a *false positive* (FP) identifies a candidate junction for a negative tag. *Sensitivity* of SAGE2Splice was computed as $\frac{TP}{TP+FN}$, whereas *specificity* was computed as $\frac{TN}{TN+FP}$.

2.4.7 Source of SAGE Tags

In searching for novel transcripts, we utilized the SAGE data generated from the Mouse Atlas of Gene Expression Project [24]. The Atlas project aims to examine comprehensively and quantitatively the expression of genes of various organ and tissue types throughout the development of mouse, from a single cell zygote to the adult. For genetic homogeneity, throughout the project only the C57BL/6J strain of mouse was used for library construction. At the end of the project, 200 SAGE libraries will have been generated. The LongSAGE protocol [19], which is similar to the original SAGE [5] in preparation but generates 21 bp tags, is being used in the majority of the SAGE libraries constructed. In this study, only the 21 bp tags were used. All SAGE data and analysis tools are public and can be downloaded from the web (http://www.mouseatlas.org).

2.4.8 Searching Unmapped SAGE Tags

SAGE tags from the available libraries in the Mouse Atlas of Gene Expression Project [24] were pooled to generate a meta-library. As described by the authors, each tag sequence was assigned a quality factor, which was computed by using PHRED scores [28], and a tag sequence probability value (p-value) was assigned based on the quality factor and the rate of errors in library construction. For tags observed more than once, individual p-values were multiplied to obtain a composite p-value. The more frequent the observations, the more confidence in the existence of the tag, thus resulting in a lower p-value. We exhaustively mapped the tags in this meta-library to all predicted tags extracted from RefSeq [26], Ensembl transcripts , Mammalian Gene Collection (MGC) [29], mRNA sequences, EST collections and the C57BL/6J mouse genome (NCBI Build 33), as well as to the full mouse UniGene mapping of SAGEmap (Build 145) [15], and then we selected 20,000 SAGE tags with the lowest p-value for further study. These tags were searched by using SAGE2Splice on the latest release of the C57BL/6J mouse genome sequence (NCBI Build 33). We used the default 10,000 bp maximum intron length and p-value cut-offs of 0.06, 0.15, and 0.0025 for the donor, the acceptor, and the overall score, respectively.

2.4.9 Categorization of Splice Junction Candidates

Three pipelines were created to classify the candidates into their respective categories. We obtained, from the UCSC Genome Browser, transcript annotations, including RefSeq, Ensembl transcripts, MGC, mRNA sequences, and EST collections, and gene predictions annotations, including TWINSCAN [30], GENSCAN [31], and SGP [32]. Candidates returned by SAGE2Splice were categorized by matching candidate junction positions to those in known transcripts. Candidates associated with Type 2 and Type 3 tags were further categorized by mapping the candidate junction positions to gene prediction annotations. Candidates that mapped to predicted junctions were classified as high priority in the validation list.

2.4.10 RNA Extraction

All samples were manually dissected and stored at -80 °C until RNA extraction. Frozen tissue was disrupted and homogenized for 30 seconds with a Polytron® PT 1200CL homogenizer (Kinematica AG, through Brinkmann™ Instruments Inc, Mississauga, Canada) at a setting of 3 (~13,000 RPM), equipped with a 7 mm generator (PT-DA 1207/2EC). RNA from each sample was extracted by using either RNeasy Mini Kit or RNeasy Lipid Tissue Mini Kit (Qiagen Inc., Mississauga, Canada), with an on-column DNaseI treatment. Quality assessment and quantification of each RNA sample was done by using RNA 6000 Nano LabChip® Kit on an Agilent 2100 Bioanalyzer (Agilent Technologies Canada Inc., Mississauga, Canada). Tissue samples of embryonic (E) 11.5 whole head (rEMS315), post natal day (P) 84 hypothalamus (rEMS340), P21 uterus (rEMS341.01), and E17.5 skeletal muscle (rEMS344) were processed by using the RNeasy Mini Kit protocol. Samples of visual cortex P20 (rEMS300), P27 (rEMS301), P42 (rEMS304), and P84 (rEMS305) were processed by using the RNeasy Lipid Tissue Mini Kit following manufacturer's directions with the modification of using 1.5 ml Phase Lock Gel-Heavy tube (Eppendorf Scientific, through Fisher Scientific, Canada) for more robust phase separation. All tissues were extracted from male C57BL/6J mice, except for the uterine tissue (rEMS341).

2.4.11 RT-PCR

Primers for each candidate (Table 2-3) were designed by using Web Primers provided by the *Saccharomyces* Genome Database (http://www.yeastgenome.org). RT-PCR amplification was performed with the QIAGEN OneStep RT-PCR Kit (Qiagen Inc. Mississauga, Ontario) as per the manufacturer. Reverse transcription was performed at 50 °C for 30 minutes. Amplification reactions included 0.4 mM of each dNTP, 1× QIAGEN OneStep RT-PCR buffer, 1× Q-Solution 2.0 μl QIAGEN OneStep RT-PCR Enzyme Mix per 50 μl reaction, and 5 U RNase inhibitor (Invitrogen Canada Inc. Burlington, Canada) per reaction. Reverse transcriptase inactivation and PCR activation were performed at 95 °C for 15 minutes, followed by 40 cycles of 94 °C for 30 seconds, 58°C for 30 seconds, and 72 °C for 1 minute, and a final extension step at 72 °C for 10 minutes. Candidates 1-3, 1-5, and 1-8 were performed at 55 °C, 30 seconds for annealing. For the –RT negative controls, the RNA was not added until after the reverse transcriptase inactivation step.

2.5 References

1. Modrek B, Lee C: **A genomic view of alternative splicing.** *Nat Genet* 2002, **30**:13-19.

2. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res* 1999, **9**:1288-1293.

3. Yajima I, Sato S, Kimura T, Yasumoto K, Shibahara S, Goding CR, Yamamoto H: **An L1 element intronic insertion in the black-eyed white (Mitf[mi-bw]) gene: the loss of a single Mitf isoform responsible for the pigmentary defect and inner ear deafness.** *Hum Mol Genet* 1999, **8**:1431-1441.

4. Qiu P, Benbow L, Liu S, Greene JR, Wang L: **Analysis of a human brain transcriptome map.** *BMC Genomics* 2002, **3**:10.

5. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484-487.

6. Boheler KR, Stern MD: **The new role of SAGE in gene discovery.** *Trends Biotechnol* 2003, **21**:55-57; discussion 57-58.

7. Chen JJ, Lee S, Zhou G, Rowley JD, Wang SM: **Generation of longer cDNA fragments from SAGE tags for gene identification.** *Methods Mol Biol* 2003, **221**:207-222.

8. Gorski SM, Chittaranjan S, Pleasance ED, Freeman JD, Anderson CL, Varhol RJ, Coughlin SM, Zuyderduyn SD, Jones SJ, Marra MA: **A SAGE approach to discovery of genes involved in autophagic cell death.** *Curr Biol* 2003, **13**:358-363.

9. Velculescu VE, Vogelstein B, Kinzler KW: **Analysing uncharted transcriptomes with SAGE.** *Trends Genet* 2000, **16**:423-425.

10. Pleasance ED, Marra MA, Jones SJ: **Assessment of SAGE in transcript identification.** *Genome Res* 2003, **13**:1203-1215.

11. Madden SL, Wang CJ, Landes G: **Serial analysis of gene expression: from gene discovery to target identification.** *Drug Discov Today* 2000, **5**:415-425.

12. Keime C, Damiola F, Mouchiroud D, Duret L, Gandrillon O: **Identitag, a relational database for SAGE tag identification and interspecies comparison of SAGE libraries.** *BMC Bioinformatics* 2004, **5**:143.

13. Jones SJ, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R, Marra MA: **Changes in gene expression**

associated with developmental arrest and longevity in Caenorhabditis elegans. *Genome Res* 2001, **11**:1346-1352.

14. Welle S, Bhatt K, Thornton CA: **Inventory of high-abundance mRNAs in skeletal muscle of normal men**. *Genome Res* 1999, **9**:506-513.

15. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: **SAGEmap: a public gene expression resource**. *Genome Res* 2000, **10**:1051-1060.

16. Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, De Souza SJ, Riggins GJ: **An anatomy of normal and malignant gene expression**. *Proc Natl Acad Sci U S A* 2002, **99**:11287-11292.

17. Divina P, Forejt J: **The Mouse SAGE Site: database of public mouse SAGE libraries**. *Nucleic Acids Res* 2004, **32**:D482-483.

18. Chen J, Sun M, Lee S, Zhou G, Rowley JD, Wang SM: **Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags**. *Proc Natl Acad Sci U S A* 2002, **99**:12257-12262.

19. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE: **Using the transcriptome to annotate the genome**. *Nat Biotechnol* 2002, **20**:508-512.

20. Deutsch M, Long M: **Intron-exon structures of eukaryotic model organisms**. *Nucleic Acids Res* 1999, **27**:3219-3228.

21. Alberts B: *Molecular biology of the cell*, 4th edn. New York: Garland Science; 2002.

22. Breathnach R, Chambon P: **Organization and expression of eucaryotic split genes coding for proteins**. *Annu Rev Biochem* 1981, **50**:349-383.

23. Burset M, Seledtsov IA, Solovyev VV: **Analysis of canonical and non-canonical splice sites in mammalian genomes**. *Nucleic Acids Res* 2000, **28**:4364-4375.

24. Siddiqui AS, Khattra J, Delaney A, Zhao Y, Astell C, Asano J, Babakaiff R, Barber S, Beland J, Bohacec S, Brown-John M, Chand S, Chaters AM, Cullum R, Dhalla N, Featherstone R, Hirst M, Hoffman B, Holt R, Hou J, Kuo BY-L, Lee LLC, Lee S, Leung D, Ma K, Matsuno C, Mayo M, McDonald M, Prabhu A, Pandoh P, Ruis de Algara T, Rupert JL, Smailus D, Stott J, Tsai M, Varhol R, Vrljicak P, Wong D, Wu MK, Xie Y-Y, Yang G, Zhang I, Hirst M, Jones S, Helgason CD, Simpson EM, Hoodless PA, Marra M: **A Mouse Atlas of Gene Expression: Large-scale, digital gene expression profiles from precisely**

defined developing C57BL/6J mouse tissues and cells. Submitted to Proceedings of the National Academy of Sciences.

25. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ: **The UCSC Genome Browser Database**. *Nucleic Acids Res* 2003, **31**:51-54.

26. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2005, **33 Database Issue**:D501-504.

27. Stormo GD: **Consensus patterns in DNA**. *Methods Enzymol* 1990, **183**:211-221.

28. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**:186-194.

29. Strausberg RL, Feingold EA, Klausner RD, Collins FS: **The mammalian gene collection**. *Science* 1999, **286**:455-457.

30. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction**. *Bioinformatics* 2001, **17 Suppl 1**:S140-148.

31. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78-94.

32. Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R: **SGP-1: prediction and validation of homologous genes based on sequence alignments**. *Genome Res* 2001, **11**:1574-1583.

Chapter 3 Conclusion and Future Directions

3.1 *Expansion of Important Observations*

3.1.1 Unmapped SAGE Tags are a Source for Novel Transcript Discovery

Not all tags can be mapped to the known transcriptome or to the genome. One explanation for this is that not all transcripts have been discovered and represented in the transcriptome databases. Although polymorphism in nucleotides can also result in the failure of mapping, for our study care was taken to work within an inbred mouse strain both for computational and bench studies, since it has been shown that variation within an inbred strain is negligible [1]. The majority of unmapped SAGE tags we used in the search had low tag counts in each individual library. Although tags that are low in copy are more likely due to sequencing or experimental errors [2], it was also suggested that they could be a source for previously uncharacterized transcripts or genes [3]. In this project, we considered the presence of a SAGE tag as evidence for the presence of a transcript. Our SAGE2Splice search indicated that with high-sensitivity and high-specificity parameters for the donor, the acceptor, and the composite junction p-values, and a 5 bp shorter edge length cutoff, 6% of the 20,000 unmapped tags were predicted to span a candidate splice junction in the genome. RT-PCR and sequencing results confirmed nine of the twelve candidate junctions, including novel alternative variants (Type 1 tags) and novel exon(s) (Types 2 and 3 tags). Thus, SAGE tags that mapped neither to known transcripts nor to the genome represent a rich resource for the discovery of novel transcripts.

3.1.2 Edge Length is an Important Factor for Reliability

In the analysis of position of splice junctions within a tag, we observed that the length of the shorter edge plays an important role in assessing the reliability of predicted splice junctions. Computational analysis showed that more than 90% of the shorter length portions that are ≥ 5 bp from the splice junction revealed true junctions. In laboratory testing, all selected candidate junctions having their shorter edge greater than 4 bp matched true novel transcripts. Therefore, in transcript discovery, candidates that have the splice junction closer to the centre of the tag should be given higher priority for laboratory testing. In addition, for each input tag that was predicted to have candidate junctions, the closer the predicted junction is to the centre of the tag, the fewer candidates required for testing in the laboratory. With a length restriction of 5 bp for the shorter edge, testing the top candidate with the lowest p-value would reveal greater than 90% of true splice junctions.

3.1.3 Quality of Input Tags and Exhaustive Mapping are Required Prior to Performing SAGE2Splice Searches

We assumed the input SAGE tags from the Mouse Atlas of Gene Expression Project did not contain base errors because care was taken to ensure sequences were of high quality, so searches were based on exact sequence matches. The quality of search results is however highly dependent on the quality of tags. It is expected the informed user will assess the quality and exhaustively map tags to all known transcriptome databases and to the genome before performing SAGE2Splice searches.

3.1.4 Functions of the Predicted Candidates are Unknown

Nine of the twelve candidates selected for laboratory testing were confirmed as novel transcripts. We do not assume these novel transcripts will necessarily be functional. As indicated in section 1.4.2, stochastic events in splicing can result in transcripts which serve no specific biological function. It is interesting to note that several of the novel mouse transcripts encode proteins with significant similarity to proteins observed in other species (e.g. dog). While the evolutionary conservation of the splice junction and ORF are suggestive of function, further studies are required to determine the biological significance of the transcripts identified by SAGE2Splice.

3.2 Future Directions

3.2.1 Improvements in Computational Efficiency

Additional optimization of SAGE2Splice can be achieved by improving the computational efficiency through the use of better hardware. A considerable portion of processing is spent on reading the genome from the hard-disk; therefore, having sufficient random access memory to constantly store the genome will minimize the reading time. Moreover, if a computer cluster is available, distribution of the unmapped tags across all nodes will maximize the computing power because searches can be processed in parallel.

3.2.2 Improvements in Usability

Further optimization and improvements can be made to increase the usefulness of SAGE2Splice. Among these is its incorporation into the DiscoverySpace software [4]. The incorporation of SAGE2Splice into DiscoverySpace will add an additional tool to the package for the annotation and discovery of transcripts using the tags that do not map to known transcripts. In addition, for those who wish to use SAGE2Splice as an independent program, the software is also accessible through a web-interface. Continuous monitoring of user feedback will aid the improvement of the program. Currently, the online version of SAGE2Splice supports genomes of four organisms and uses the PWM generated from the mouse annotations for splice junction evaluation. In the future, I would like to include more organisms. Although nucleotides flanking splice junctions generally follow a common pattern, further experiments could be performed to assess the possibility of increasing prediction accuracy by using organism-specific PWM. Should an increase in accuracy be observed, the organism-specific PWM would be incorporated into SAGE2Splice. Organism-specific PWM can be computed according to the method described in Chapter 2.

3.2.3 Improvements in Candidate Evaluation

The high false positive rates of the PWM in splice junction prediction were addressed by matching junctions to experimental evidence supported by SAGE tags, using conserved boundary dinucleotides, choosing high specificity p-value cutoffs and maximum intron size, and placing a restriction on the minimum length of the shorter edge. In the future, I would like to evaluate other proposed methods, such as decision tree [5] and HMM [6], to determine the optimal assessment for splice junction predictions. In addition, because the length of the shorter edge plays an important role in determining the reliability of a predicted candidate, I would also like to incorporate this into the generation of candidate p-values.

3.2.4 Improvement to Search Non-canonical Splice Junctions

Some of the remaining 91.8% of the unmapped tags are likely to span a splice junction with intron boundary dinucleotides that do not follow the GU-AG rule, the *non-canonical* splice junctions, or have introns longer than 10,000 bp. To detect the less conserved and larger introns, the p-value cutoffs and maximum intron length can be relaxed. Also, improvements to the algorithm could be made to include searches of the non-canonical splice junctions.

3.2.5 Screen More Unmapped Tags Can Lead to the Discovery of More Transcripts

The Mouse Atlas of Gene Expression Project has produced more than 200,000 unmapped tags. Of which, less than 10% were analyzed using SAGE2Splice. To search all 200,000 unmapped tags, allocation of time and of hardware resources is required. By using a computer cluster with several nodes, this task can be completed in several days.

3.2.6 High Throughput Laboratory Processes Will be Required

To test the predicted candidate splice junctions for all 200,000 unmapped tags (an estimate of ~24,000), a high throughput laboratory process is required. First, because of the large number of candidates, the laboratory will require a large amount of RNA for testing. To ensure the resemblance of the RNA samples, the dissection and RNA extraction procedures for the new RNA should be identical to that used for the sample preparation of the SAGE libraries. Second, we demonstrated that candidates with their shorter edge lengths ≥ 5 bp are more likely to reveal true splice junctions (**Error! Reference source not found.**). Our RT-PCR experiments also confirmed this observation. Thus, the high throughput procedure should focus only on candidates with their shorter edges ≥ 5 bp. Third, with a 5 bp length restriction on the shorter edge, > 90% of the true junctions were revealed. For each tag, only the candidate with the lowest composite p-value should be tested. Our data suggests there will be little advantage to test more than one candidate for each tag. Furthermore, candidates with additional computer prediction support should be given higher priority for testing because these are more likely to be validated. Fourth, an automated primer design pipeline should be created. Finally, should a pair of primers yield more than one product, only the product that matches the expected size should be sequenced. We expect that a high throughput laboratory procedure should ensure that candidate junctions predicted by SAGE2Splice are tested efficiently and would assist in the discovery of novel transcripts.

3.3 Conclusion

Continuous discovery of novel transcripts has demonstrated that the transcriptome is more complex than the genome. In addition to the current techniques for novel transcript discovery, we have developed SAGE2Splice that combines experimental evidence of SAGE tags and a computational prediction method to identify candidate splice junctions in a genome. We demonstrated that nine candidates predicted by SAGE2Splice revealed both previously unknown alternative variants and previously unknown exons. With further improvements to the SAGE2Splice candidate evaluation algorithms and the use of better computer hardware, we expect that many novel transcripts and genes can be discovered.

3.4 References

1. Wade CM, Kulbokas EJ, 3rd, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ: **The mosaic structure of variation in the laboratory mouse genome**. *Nature* 2002, **420**:574-578.

2. Akmaev VR, Wang CJ: **Correction of sequence-based artifacts in serial analysis of gene expression**. *Bioinformatics* 2004, **20**:1254-1263.

3. Boheler KR, Stern MD: **The new role of SAGE in gene discovery**. *Trends Biotechnol* 2003, **21**:55-57; discussion 57-58.

4. Varhol R, Robertson N, Oveisi-Fordorei M, Fiell C, Leung D, Siddiqui AS, Marra M, Jone S: **DiscoverySpace: A tool for gene expression analysis and biological discovery**. *Poster* 2005.

5. Brunak S, Engelbrecht J, Knudsen S: **Prediction of human mRNA donor and acceptor sites from the DNA sequence**. *J Mol Biol* 1991, **220**:49-65.

6. Durbin R: *Biological sequence analysis : probalistic models of proteins and nucleic acids*. Cambridge, UK New York: Cambridge University Press; 1998.

Appendix: Novel Transcript Sequence Information

```
LOCUS       DQ113644                 184 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus Ppih-like mRNA, partial sequence.
ACCESSION   DQ113644
VERSION     DQ113644
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 184)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 184)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..184
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="4"
                     /tissue_type="whole head"
                     /dev_stage="embryonic day 11.5"
     misc_feature    <1..>184
                     /note="similar to Ppih"
ORIGIN
  1 ttcaaacagt ggtcccagta caaatggctg ccagttcttt tcaaacagtg gtcccagtac
 61 aaatgggtgc gcagttcttt atcacgtgtt ctaagtgtga ttggctggat ggaaagcatg
121 tagtgtttga atgttcccac aggcaacttc tagtgatgag gaagatttga atgttcccac
181 aggc
```

```
LOCUS       DQ113645                 221 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus unknown mRNA, partial sequence.
ACCESSION   DQ113645
VERSION     DQ113645
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 221)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 221)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..221
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="5"
                     /tissue_type="visual cortex"
                     /dev_stage="post natal day 27"
     misc_feature    <1..>221
                     /note="unknown; transcript variant 2; similar to
                           mRNA in GenBank Accession Number AK081926"
ORIGIN
  1 tctaaggaag atggcgaaga cagtgaggga agagagcaga cgtctgactc cggggtgctt
 61 atctgtgtgg aagagaccgg ttcttcctga ctggcttcat gtccctcaag gtgttctcct
121 ggctcttcaa gtatttaccc gtctgtgtgt gactcatatc caggaaccca ggcagcttcc
181 ctcaatagat tgctggcttc agaagatgag cctcccctaa g
```

```
LOCUS       DQ113646                 361 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus unknown mRNA, partial sequence.
ACCESSION   DQ113646
VERSION     DQ113646
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 361)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 361)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..361
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="5"
                     /tissue_type="visual cortex"
                     /dev_stage="post natal day 27"
     misc_feature    <1..>361
                     /note="unknown; transcript variant 3; similar to
                     mRNA in GenBank Accession Number AK081926"
ORIGIN
  1 ttaaggaaga tggcgaagac agtgaggga gagagcagac gtctgactcc ggggtgctta
 61 tctgtgtgga agagaccggt tcttcctgac tggcttcatg tccctcaagg atcccaaacc
121 aaggctntgg actatttcaa agccantagt aatangggtc agtagtactc agcagccctg
181 ctcctgggtg caaaganacn aggncaggtg cagactgtgc tcncatactt ggaagcttgg
241 tggtggtgga ggtgttctcc tggctcttca agtatttacc cgtctgtgtg tgactcatat
301 ccaggaaccc aggcagcttc cctcaataga ttgctggctt cagaagatga gcctccccta
361 a
```

```
LOCUS       DQ113647                 202 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus Rpl41-like mRNA, partial sequence.
ACCESSION   DQ113647
VERSION     DQ113647
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 202)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 202)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..202
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="10"
                     /tissue_type="whole head"
                     /dev_stage="embryonic day 11.5"
     misc_feature    <1..>202
                     /note="similar to Rpl41"
ORIGIN
  1 tcatgagagc gaaggctgaa ttcatgagag cgaaggctga agcgcaagag aagaaagatg
 61 aggcagaggt ccaagtaagc cagcccgtgc acctacgacg cctgcaggag cagaagtgag
121 ggatgctgag ggccgggaca agctatcgga ctgtgtgctg ccatcggtaa tgagtctcaa
181 tgccatcggt aatgagtctc aa
```

```
LOCUS       DQ113648                 131 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus Tpt1h-like mRNA, partial sequence.
ACCESSION   DQ113648
VERSION     DQ113648
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 131)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 131)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..131
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="19"
                     /tissue_type="hypothalamus"
                     /dev_stage="12 weeks"
     misc_feature    <1..>131
                     /note="similar to Tpt1h"
ORIGIN
  1 ttccgaaatg tgcagctgtc taaggctctg tgcctatgcc cttcgccacg gggccttgaa
 61 gctgggactt cccatgcgag ctggcattcg tccaaattgt gaggtggcgg tgttcatcga
121 tggaccccta a
```

```
LOCUS       DQ113649                  79 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus Rpl136a-like mRNA, partial sequence.
ACCESSION   DQ113649
VERSION     DQ113649
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 79)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 79)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..79
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="X"
                     /tissue_type="visual cortex"
                     /dev_stage="12 weeks"
     misc_feature    <1..>79
                     /note="similar to Rpl136a"
ORIGIN
 1 gctcctgcga acatggaaag cggcgttacg acaggaaaca gagtggctat ggtgggcaga
61 ctaagcctat tttccgcaa
```

```
LOCUS       DQ113650                  178 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus Ccs-like mRNA, partial sequence.
ACCESSION   DQ113650
VERSION     DQ113650
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 178)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 178)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..178
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="4"
                     /tissue_type="visual cortex"
                     /dev_stage="post natal day 20"
     misc_feature    <1..>178
                     /note="similar to Ccs"
ORIGIN
  1 ctatcaccaa ctgctgtgct gtgctctgtg gcccactgcc tcccagcctg attgaccggc
 61 gaggattcgt gcagtctgat accgcgttgc cctcgcccat cagaagtgat gacccggcct
121 gtggagccaa gccagacgcc agcatggagg acacttgtca ggactttgcc atctagaa
```

```
LOCUS       DQ113651                 143 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus Ywhae-like mRNA, partial sequence.
ACCESSION   DQ113651
VERSION     DQ113651
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 143)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 143)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..143
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="11"
                     /tissue_type="leg skeletal muscle"
                     /dev_stage="embryonic day 17.5"
     misc_feature    <1..>143
                     /note="similar to Ywhae"
ORIGIN
  1 cctagcagct tttgatgacg caattgcaga actggaccgc tgaagtgaag aaagctataa
 61 ggactctacg gctcattcat gcagctgcta cgtgataacc ctgacgctgt ggacctcaga
121 catgcagggt gatgattcct aaa
```

```
LOCUS       DQ113652                 243 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus isolate s2sEMS1 mRNA sequence.
ACCESSION   DQ113652
VERSION     DQ113652
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 243)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 243)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..243
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /isolate="s2sEMS1"
                     /db_xref="taxon:10090"
                     /chromosome="4"
                     /tissue_type="uterus"
                     /dev_stage="post natal day 21"
ORIGIN
    1 ctatagaatc ctcgtcgcca tccgtgactg aaggacagac gcttgacctt aactgtgcgg
   61 tgatggggtt gacctacacc caggtcacat ggtacaagcg agggggcagc ctgcctcccc
  121 atgcccaggt ccacggctcc cggctgcggc tcccgcaggt ctcaccggca gactccggag
  181 actatgtgtg ccgagtggag agganngtga cgtgggccct aaggaggctt ccattgttgt
  241 aga
```

```
LOCUS       DQ113653                 266 bp    mRNA    linear   ROD
26-AUG-2005
DEFINITION  Mus musculus isolate s2sEMS2 mRNA sequence.
ACCESSION   DQ113653
VERSION     DQ113653
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 266)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     SAGE2Splice: Unmapped SAGE Tags Reveal Novel Splice
            Junctions
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 266)
  AUTHORS   Kuo,B.Y.L., Chen,Y., Bohacec,S., Wasserman,W.W. and
            Simpson,E.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUN-2005) Medical Genetics, University of
            British Columbia, 950 West 28th Avenue, Vancouver, British
            Columbia V5Z 4H4, Canada
FEATURES             Location/Qualifiers
     source          1..266
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6J"
                     /isolate="s2sEMS2"
                     /db_xref="taxon:10090"
                     /chromosome="13"
                     /tissue_type="visual cortex"
                     /dev_stage="post natal day 42"
ORIGIN
  1 tccgtgagag tgactttgga ttttaacctc actgatccag aaaatgggcc cgtgctcgat
 61 gacgctctgc caaactcagt ccatggacat attccttttg ccaaagactg tgggaacaag
121 gaaagatgcg tttcagacct caccctggat gtgtccacaa cagaaaagaa cctgctgatt
181 gtcagatccc agaatgacaa gttcaatgtc agcctcaccg tcaaaaacaa gggagacagt
241 gcgtacaaca cccggacagt ggttaa
```