Title: Tag-Based Approaches for the Detection of Cis-Encoded Antisense Transcription

Authors: Petrescu A., Delaney A., Marra M.

Abstract

Elucidating the mechanisms by which gene expression is regulated is one of the main challenges in understanding mammalian development. Antisense transcription has recently been recognized as one such mechanism. Prior to the availability of genome sequence data, fewer than 40 sense-antisense (S-AS) transcripts were known to play roles in development (Vanhee-Brossollet and Vaquero, 1998), but with the availability of extensive EST and fully-sequenced cDNA libraries, the number of S-AS genes has grown to encompass nearly half of the transcriptome (Hayashizaki and Carninci, 2006). Although bioinformatic mining of cDNA libraries has enhanced our ability to detect S-AS transcripts, there remains the main challenge of creating a genome-wide catalog of accurate gene expression profiles for S-AS transcripts. Such a resource can then be used to elucidate the role these transcripts play in regulating gene expression. To address this challenge, a technology such as LongSAGE, capable of detecting digital gene expression profiles, can be employed to both enumerate known and novel S-AS transcripts and also provide accurate measurements of expression levels.

To date, SAGE has been used in two studies (Chen et al., TiG, 2005; Chen et al., PNAS, 2002), to show that AS transcription can indeed be detected using SAGE in multiple human tissues. However, these studies suffer from the limitations of the short length of SAGE tags, which do not allow unique assignment of tags to genomic sequences, and thus reduce the number of informative tags. In contrast, the LongSAGE technique generates longer tags with a 75% unique mapping rate to the genome. Only one preliminary study has thus far applied LongSAGE to the discovery of AS transcription (Wahl et al, Bioinformatics, 2005), but was critically limited in scope (one tissue: embryonic mouse tail) and both scale and depth (one library of 200,000 tags). We are now poised to address the issues of scope, scale and depth with the recent completion of a large collection of LongSAGE datasets. We thus analyzed 196 publicly available SAGE libraries generated as part of the Mouse Atlas of Gene Expression (MA) project, 15 libraries from the Mammalian Organogenesis – Regulation by Gene Expression Networks (MORGEN) project, and one Solexa-SAGE library sequenced at the BCCA Genome Sciences Centre. By analyzing libraries sequenced at depths of 100,000 tags (MA), 200,000 and 300,000 tags (MORGEN), and 5 million tags (Solexa-SAGE), we found that the ability to detect antisense transcription is dependent on sequencing depth. We present evidence that half of the genes in EnsEMBL show AS transcription. We also show that one Solexa-SAGE library has a comparable transcript discovery rate to 15 MORGEN libraries and ~20 MA libraries. The sensitivity to transcriptional signal is 2 to 8-fold improved in Solexa-SAGE over classical LongSAGE techniques, and hence the ability to detect infrequently expressed transcripts, such as antisense transcripts, has surpassed existing methods by almost an order of magnitude.

Table 1. Mouse Atlas, Morgen, and Solexa-SAGE library statistics. Tags and tag types from the MA and MORGEN metalibraries, as well as from the Solexa-SAGE library, were enumerated and mapped to the genome (Uniquely Mapping Tag Types). These were further mapped as either sense (EnsEMBL Genes (S)) or antisense (EnsEMBL Genes (AS)) to 27,967 EnsEMBL genes.

	MA	MORGEN	Solexa-SAGE
Libraries	196	15	1
Target Depth	100,000	2-300,000	5 mil
Tags	22,102,568	3,951,294	5,069,100
Tag Types	2,045,306	558,506	1,414,215
Uniquely Mapping Tag Types	1,984,693	174,144	340,190
EnsEMBL Genes (S)	18,755	14,904	14,202
EnsEMBL Genes (AS)	16,935	11,490	13,739

Figure 1. Coverage of the Fantom3 mouse database by antisense tags from Mouse Atlas, Morgen, and Solexa-SAGE libraries and metalibraries. Coverage of the Fantom3 database was assessed for AS tags from various single libraries (ex. MA_100k_sm147_AS is a Mouse Atlas library sequenced to a depth of 100,000 tags), as well as metalibraries composed of all or a random subset of libraries (ex. MA_random_50_AS is a metalibrary of 50 random Mouse Atlas libraries). Computationally re-sampling these libraries at various depths allows an estimate of the sequencing depth required to saturate transcript discovery.