

Transcriptome Analysis Tools: Visualization and Management of Ultra-High Volume of DNA Sequence Data

Irina Khrebtukova¹, Christian D. Haudenschield, Daixing Zhou, William Nelson, Selene M. Virk, Maria Johnson, Keith Moon, Thomas Vasicek

Keywords: short read sequences, annotation, genome browser, reference transcriptome

1 Introduction.

It is generally believed that the first step toward systems biology is to construct a comprehensive expression database that catalogs all the mRNA and regulatory RNA in a sample and documents their expression levels. When a database consisting of expression data from many samples is constructed, it can serve as a reference transcriptome database. Massively Parallel Signature Sequencing (MPSSTM) is an extremely efficient method for generating short DNA sequences. It has been routinely used for the identification of transcribed regions including those that are processed into mRNAs and small regulatory RNAs (miRNA and siRNA). Because MPSS measures absolute copy number of transcripts, the reference transcriptome databases established by MPSS are truly expandable and exchangeable.

MPSS, in its current implementation, routinely analyses up to 1 million DNA molecules simultaneously from a single sample preparation, each yielding a 20 bp signature. When analyzing transcriptome data, we find that 97% of the MPSS cDNA signatures map to a unique mRNA transcript. We have generated more than 600 million ESTs from a wide variety of tissues, isolated cells, and cell lines of many organisms. The enormous depth of these data has revealed transcripts from many previously uncharacterized loci.

This level of throughput and ultra high volume of short sequence reads requires special tools for the data management, annotation and visualization. This presentation will describe Lynx pipeline for data processing and annotation, and the genome browser for viewing short sequence data in the genome and transcriptome context.

2 Signature Genome Browser and MPSS data integration.

We developed a data integration and visualization tool called “Signature Genome Browser” (SGB) for rapid and reliable display of MPSS data in the genome context [1]. This browser displays MPSS signatures and transcripts mapped to any chosen region of the genome. To build the SGB, we first extract all possible GATC-17mers and GATC-20mers (“virtual signatures”) from the genome and the mRNA sequences while recording their coordinates in the genome and their positions among the transcripts. We then built a relational database to include the signatures, their positions in the genome and the transcripts, the annotation tables downloaded from UCSC Golden Path [2], and the expression levels detected by MPSS. Finally, we constructed a searchable graphic interface to allow the end-users to query the database and display the signatures, along with their expression levels and their associated genomic information in the genome context (see Figure 1).

¹ Lynx Therapeutics, Inc., Hayward, CA. E-mail: irina@lynxgen.com

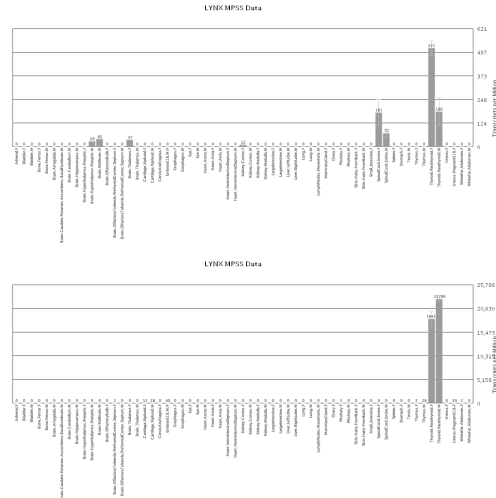
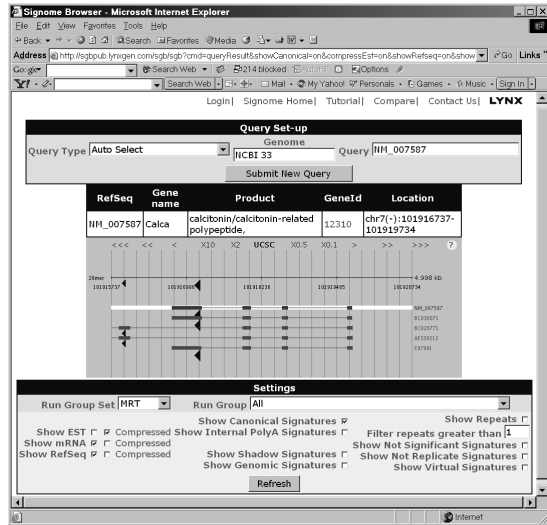


Figure 1: An example of SGB display showing expression profile of two alternative transcripts of the calcitonin in 61 mouse tissues. Data from mouse reference transcriptome project [3].

3 Reference Transcriptomes.

MPSS™ is the ideal technology for establishing a reference transcriptome database. It provides superior sensitivity and dynamic range. It is the only technology that can routinely provide the sampling depth needed for accurate and quantitative determination of the expression level of every gene in a particular sample. Unlike most other gene expression technologies that provide analog expression data, MPSS provides digital expression information, which is crucial for data exchange, comparison and expansion.

Powered by MPSS and jointly funded by many NIH Institutes, we have recently established the mouse reference transcriptome (MRT) for normal mouse tissues. The MRT database is hosted in NCBI and publicly available [3]. Establishing reference transcriptomes of other species using the same technology is underway. In addition, Lynx has adapted MPSS technology to catalog nearly all the small regulatory RNAs (siRNA and microRNA) in a sample, ready for establishing reference databases for small regulatory RNAs.

References

- [2] <http://genome.ucsc.edu/>
- [1] <http://sgbpub.lynxgen.com/sgb/sgb>
- [3] <http://www.ncbi.nlm.nih.gov/genome/guide/mouse/MouseTranscriptome.html>