BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 4 2010, pages 585–586
doi:10.1093/bioinformatics/btp695

Databases and ontologies

GonadSAGE: a comprehensive SAGE database for transcript discovery on male embryonic gonad development

Tin-Lap Lee[1,*], Yunmin Li[2], Hoi-Hung Cheung[1,3], Janek Claus[4], Sumeeta Singh[4], Chandan Sastry[4], Owen M. Rennert[1], Yun-Fai Chris Lau[2] and Wai-Yee Chan[1,3]

[1]Section on Developmental Genomics, Laboratory of Clinical Genomics, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, [2]Department of Medicine, VA Medical Center, University of California, San Francisco, CA 94121, USA, [3]School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China and [4]Divsion of Information Technology, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA

Received on September 17, 2009; revised on November 24, 2009; accepted on December 15, 2009

Advance Access publication December 21, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Serial analysis of gene expression (SAGE) provides an alternative, with additional advantages, to microarray gene expression studies. GonadSAGE is the first publicly available web-based SAGE database on male gonad development that covers six male mouse embryonic gonad stages, including E10.5, E11.5, E12.5, E13.5, E15.5 and E17.5. The sequence coverage of each SAGE library is beyond 150K, 'which is the most extensive sequence-based male gonadal transcriptome to date'. An interactive web interface with customizable parameters is provided for analyzing male gonad transcriptome information. Furthermore, the data can be visualized and analyzed with the other genomic features in the UCSC genome browser. It represents an integrated platform that leads to a better understanding of male gonad development, and allows discovery of related novel targets and regulatory pathways.

Availability: GonadSAGE is at http://gonadsage.nichd.nih.gov.

Contact: leetl@mail.nih.gov

1 INTRODUCTION

Serial analysis of gene expression (SAGE) is a powerful genomic tool to detect RNA species based on sequencing approaches. It offers a comprehensive and unbiased method for novel transcript discovery not found in microarray platforms. Previously, we successfully applied SAGE to identify novel transcript species during male germ cell development, including stage-specific splicing variants, anti-sense transcripts and transcripts of unknown function (Chan, *et al*., 2006a, b; Lee *et al*., 2006, 2009; Wu *et al*., 2004). Bioinformatic analyses revealed unique transcriptional regulation of various transcription factors and promoter elements, and the involvement of stage-specific gene networks (Chan *et al*., 2006; Lee *et al*., 2006). We recently completed comprehensive SAGE profiling of male embryonic gonad development (Lee *et al*., 2009) and demonstrated global and temporal patterns of gene expression at different developmental time points. We established molecular staging and transcription 'hotspots' based on the gene expression signature. A significant number of novel genes, expressed at specific developmental time points, related to sex determination, meiosis and steroidogenesis were identified. These observations correlated with developmental defects reported in established animal models.

GonadSAGE is the first public interactive database that contains transcriptome information on male embryonic gonad development. Importantly, it permits dynamic analysis and comparison with other genomic features and databases. It provides flexible search parameters, and the data may be visualized and analyzed with other genomic datasets that use the genome browser format. It is an invaluable tool for identification of novel transcripts and regulatory pathways in male gonadal development.

2 DATABASE CONTENT

A total of six male mouse embryonic gonad stages were included in the current version (E10.5, E11.5, E12.5, E13.5, E15.5 and E17.5). To discriminate between female and male gonads at E10.5 and E11.5 prior to sex determination, the gonads were dissected from the embryos using a stereomicroscope, individually placed into PCR tubes, and snap-frozen with liquid nitrogen. Tail tissue from the dissected embryo and used to isolate DNA for PCR genotyping with *Sry* specific primers. It took 100 to 150 fetal gonads for a single RNA pool preparation for each library. The sequence coverage for each SAGE library is above 150K. All duplicate ditags were eliminated when the number of sequenced tags was compiled. Tags derived from linkers were also eliminated. Overall, the six transcriptomes gave a total of 47 255 annotated transcripts, which are comprised of 29 060 singletons and 18 195 tags that mapped to Unigene clusters with multi-hits. The total number of unique tags in singletons and multi-hits was 13 947 and 12 506, whereas the unique Unigene ID was 23 964 and 11 028, respectively. This translates to a total of 36 470 unique tags and 24 975 unique Unigene IDs. Raw dataset can be downloaded directly from the website.

3 DATABASE DESIGN

The GonadSAGE application provides an organized approach for sharing genomic data in a browser extensible display (BED) format.

*To whom correspondence should be addressed.

It utilizes the UCSC Genome Browser (Kuhn *et al*., 2009) to visualize experimental datasets. Genome coordinate information in the BED format was obtained by blasting BLASTN analysis of the SAGE tag sequence generated by the SAGEmap mapping procedure (Lash *et al*., 2000) against the mouse genome (NCBI Build 37 assembly); only perfectly matched tags were retained.

If a tag matched more than one Unigene cluster, the complete Unigene list will be retrieved by a 'Full text or specific field search' in the search page. The core of GonadSAGE is based on a domain model that is comprised of Java objects that describe the business, operations and object relationship. The domain model is established using Hibernate (Bauer and King, 2006). The view layer is rendered primarily using Java Server Pages. The application layers are joined using Spring (Walls and Breidenbach, 2007). The system will check the availability of the host site (http://genome.ucsc.edu) before rendering the BED files.

4 DATABASE ACCESS AND WEB INTERFACE

GonadSAGE offers two navigation options. The data can be visualized freely in the genome browser format through 'Genome view of complete data set' or searched by specific criteria using 'Full text or specific field search'. Under genome view format, the users can compare the GonadSAGE data to various UCSC genome annotations by adding tracks from the table browser. In addition, the users can upload processed data from their own or other public resources using UCSC Genome Browser's custom annotation track feature. The custom annotation track is viewable on top of the GonadSAGE dataset. This interactive approach provides a dynamic way to analyze GonadSAGE data. To perform complex queries or analysis, selected dataset can be imported to Galaxy (Blankenberg *et al*., 2007) in the UCSC genome browser.

GonadSAGE also provides a powerful search function option for identification and discovery of transcript species in the development programs of male gonads. Selecting 'Full text or specific field search' or 'Data search' tab will direct one to the search page. GonadSAGE offers a wide variety of search parameters. Users can search the transcripts by gene name/symbol, Unigene ID, sequence, chromosomal location and gene ontology (Fig. 1A). A combination of searching parameters can be applied to answer specific biological questions. The advanced search option allows transcript search by SAGE tag counts at each stage using different operators. For example, to look for genes predominantly expressed before sex determination at E12.5, we put $\geqslant$10 in E10.5 and E11.5 and $\leqslant$5 from E12.5 to E17.5. GonadSAGE will return a number of developmental genes, such as the SRY-Box Containing Gene 11 (*Sox11*) (Fig. 1B). The results also include two unknown transcripts (transcribed locus on chromosome 2 and 10) (data not shown), which might represent novel gene candidates during early developmental programming of the male gonad. The results and the raw data can be downloaded in comma-separated values format at the end of search result page. The genome view of SAGE tag locations can be retrieved by clicking the Enterz ID (Fig. 1C). The developmental stages are indicated on the left hand side of the genome browser. It also provides evidence of potential splicing patterns due to alternative 3′UTR usage in the given transcript. The data can be captured or analyzed in the genome browser application.

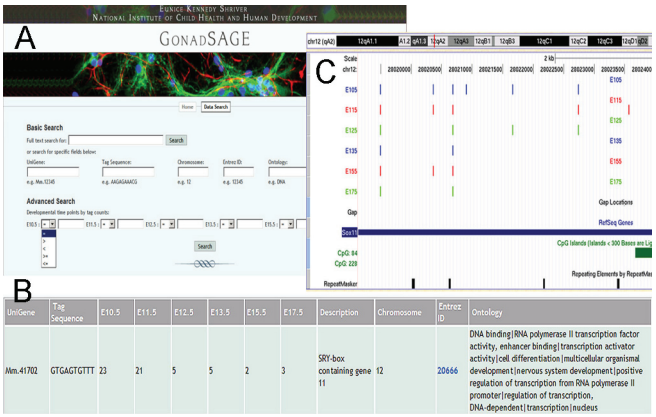

Fig. 1. GonadSAGE layouts. (**A**) Main search page with basic and advanced options. (**B**) Search result page that contains Unigene/Enterz ID, sequence, tag count, chromosomal and gene ontology information. (**C**) Each transcript can in the result page be further visualized and analyzed in the genome browser. The location of SAGE tag is represented in form of vertical bar.

5 FURTHER DEVELOPMENTS

We are developing algorithms to extract the cellular dynamics contained in the GonadSAGE data, which include transcriptional regulation through over-representation or co-expression of promoter sequence elements and gene interaction networks at a particular embryonic stage. We will include the morphological data, and reveal germ cell specific genes in male gonadal development by comparing them with the somatic cells at each stage.

Funding: Intramural Research Program of the National Institutes of Health (NIH); Eunice Kennedy Shriver National Institute of Child Health and Human Development; National Institutes of Health (HD-33728) in part.

Conflict of Interest: none declared

REFERENCES

Bauer,C. and King,G. (2006) *Java Persistence with Hibernate*. Manning Publications, Greenwich, USA.

Blankenberg,D. *et al*. (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res.*, **17**, 960–964.

Chan,W.Y. *et al*. (2006a) Transcriptome analyses of male germ cells with serial analysis of gene expression (SAGE). *Mol. Cell Endocrinol.*, **250**, 8–19.

Chan,W.Y. *et al*. (2006b) The complexity of antisense transcription revealed by the study of developing male germ cells. *Genomics*, **87**, 681–692.

Kuhn,R.M. *et al*. (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.

Lash,A.E. *et al*. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.

Lee,T.L. *et al*. (2006) Application of transcriptional and biological network analyses in mouse germ-cell transcriptomes. *Genomics*, **88**, 18–33.

Lee,T.L. *et al*. (2009a) GermSAGE: a comprehensive SAGE database for transcript discovery on male germ cell development. *Nucleic Acids Res.*, **37**, D891–D897.

Lee,T.L. *et al*. (2009b) Developmental staging of male murine embryonic gonad by SAGE analysis. *J. Genet. Genomics*, **36**, 215–227.

Walls,C. and Breidenbach,R. (2007) *Spring in Action*. Manning Publications, greenwich, USA.

Wu,S.M. *et al*. (2004) Analysis of mouse germ-cell transcriptome at different stages of spermatogenesis by SAGE: biological significance. *Genomics*, **84**, 971–981.