Visualizing BayGenomics Genetraps with Genome Browsers

Thomas Ferrin, Ph.D.
Department of Pharmaceutical Chemistry and Biopharmaceutical Sciences
Director, Resource for Biocomputing, Visualization, and Informatics
Overview of BayGenomics
(www.baygenomics.ucsf.edu)

- Our goal is to use gene-trap vectors to inactivate thousands of genes in mouse embryonic stem (ES) cells for the purpose of generating knockout mice
- Many mutants have been transmitted through the germline
- We make our ES cell lines and knockout mice are available to the scientific research community
- Each ES cell line is incorporated into the searchable BayGenomics Database
- *in situ* hybridization images also available for analyzing gene-expression patterns
- We conduct periodic workshops on bioinformatics and make available on-line bioinformatics tutorials
Steady growth in volume of data
Typical cell line annotation

GeneTrap Resource DB Results: RRE094

Cell Line: RRE094 (380 bp)
Genome Mapping: UCSC Ensembl
Vector: pGT2Lxt
Species: Mus musculus
Category: Putative Mouse ID
Identification: GenBank: NM_134079, LocusID: 11534 - 100.0% identity over 100.0% length of the RRE094 sequence tag
Gene Symbol (JAX): Adk
Synonyms: 5033405D03Rik / AK / MGC:6593
Chromosome: 14
Top Mouse dbEST Match: CK787663.1 - matched 100.0% identity over 100.0% length of the RRE094 sequence tag [BLAST]
Sequence: >RRE094 (Adk; adenosine kinase)

AGTGAATAATGTCCTATTTGGAATGATGGAATTCTCTCTTCCTGAGACATCTCTGCTGTACGACAAAGATTTTCCTTTGATAACT
TTCTCTGAAACCCGAAATGACCAAGATCTTGGCTGAGACAAAGCAAGCAAGAACTGTGTTTGATGAACTTTGTTGAAAATAATCAAG
TTGAATATTCATGTGGTGGCTCTACGCAAATTCAATGAAATGGCTGCTGATTCGAGACCAACGGACACAAAGCAAGCAAGCAAG
TTCTCTGAAATGACCAAGATCTTGGCTGAGACAAAGCAAGCAAGCAAGAACTGTGTTTGATGAACTTTGTTGAAAATAATCAAG
CTATGAGGACGAGACGAGCCCAACAGAAGAACTTTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

Download or view ABI trace file used for RRE094.

Post Date: Aug 02, 2002
Updated: Nov 29, 2003
Availability: This sequence was obtained by 5′RACE PCR from a mouse embryonic stem (ES) cell clone with an insertional mutation from a gene trap vector. This cell line is available to the scientific community for the purpose of generating a gene knockout mouse.

Copyright 2002 Regents of the University of California. All rights reserved.
Browsing gene-traps at Ensembl

Select ContigView
Sequence data must be kept current

Problem: How to ensure newly determined cell line sequences are available at Ensembl

Possible Solutions:

- Get Ensembl to download our cell line sequences and do searches against the mouse golden path whenever new BayGenomics data is available
- Use a distributed database approach
The Distributed Annotation System

- A client-server system in which a single client integrates information from multiple servers. It allows a single machine to gather up genome annotation information from multiple distant web sites, collate the information, and display it to the user in a single view. Little coordination is needed among the various information providers.

- The Lightweight DAS Server (LDAS) is designed for use by small to medium sites. It is ``lightweight'' in the sense that once all the software is installed, annotations can be loaded and updated using tab-delimited text files.

http://biodas.org/servers/LDAS.html
What’s behind the curtains...

- Ensembl web server
- Mouse genome data
- Mouse goldenpath
- BayGenomics cell line sequences
- SSAHA
- MySQL RDBS
- BayGenomics LDAS server
- Ensembl web server
- Mouse genome data
- BayGenomics cell line sequences
- SSAHA
- MySQL RDBS
- BayGenomics LDAS server
SSAHA

- **Sequence Search and Alignment by Hashing Algorithm**

  SSAHA is a very fast tool for matching and alignment of DNA sequences. It is most useful when you are looking for exact or 'almost exact' matches between two sequences. SSAHA is much faster than BLAST, but has fewer capabilities and less flexibility.

  http://www.sanger.ac.uk/Software/analysis/SSAHA/
Must specify DAS source

- DAS Sources->Manage Sources...
Caveat

- Beware that only a limited number of hits may be displayed!
What about the UCSC Genome Browser?

BayGenomics Cell Line Annotation – Mozilla (Build ID: 2004031613)

GeneTrap Resource DB Results: RRE094

Cell Line: RRE094 (380 bp)
Genome Mapping: Ensembl
Vector: JPT2LaM
Species: Mus musculus
Category: Putative Mouse ID
Identification: adenosine kinase  MGI:187930
   GenBank: NM_134079, LocusID: 11534 - 100.0 % identity over 100.0 % length of the RRE094 sequence tag
Gene Symbol (JAX): Adk
Synonyms: 5033405D03Rik / AK / MGC:6593
Chromosome: 14
Top Mouse dbEST Match: CK787663.1 - matched 100.0% identity over 100.0% length of the RRE094 sequence tag [BLAST]
Sequence:
> RRE094 (Adk; adenosine kinase)
AGTGAAGATGTGATTTGGGATCCTCTGCTGTACTCTGTAGACACACACGATTTTCTCTGATAGTA
TCTCGAACAATGACACAGATCTTGGCGAGACAACAGCAACAGGAACATGATGAATTTGCTGGAAATTTCAAG
TTGAATATCATGCTGTGCCCTACTACGAGATTCAATGAAATGTCATAGTGGTATCAGGAGCCACACAAGCCAACGA
ACGCTCTTTTGATCATGGAGATATACTTTTTGAGATCTTCTGAGCTACTAGCTGAGATCTGCTCTGACATGGATGTCTCA
TPACTATAGCAGAAGAGCGACCACCAACCGGAACCTGTCCTGCTGTCATCACTGCTGCCAA

Download or view ABI trace file used for RRE094.

Post Date: Aug 02, 2002
Updated: Nov 29, 2003
Availability: This sequence was obtained by 5’ RACE PCR from a mouse embryonic stem (ES) cell clone with an insertional mutation from a gene trap vector. This cell line is available to the scientific community for the purpose of generating a gene knockout mouse.

Copyright 2002 Regents of the University of California. All rights reserved.
Results similar to Ensembl browser
Customized “BLAT Track” provides the same kind of functionality as DAS

- Allows someone to browse mouse genome and display custom annotations like BayGenomics cell line sequences
- Requires same type of golden path search as done with Ensembl’s SSAHA
  - Results stored in large text file which then gets sent to UCSC via http
  - http://genome.ucsc.edu/cgi-bin/hgTracks?db=mm4&position=chr14:159...
  - ...&hgt.customText=http://baygenomics.ucsf.edu/blattrack/baygenomics.psl

*de novo* browsing for gene tags requires explicit specification of BLAT track
Data distribution very similar

Researcher’s web browser

Mouse genome data

UCSC web server

Mouse goldenpath

BLAT

BLAT Track text file

BayGenomics cell line sequences

BayGenomics web server

UCSF
Adding a BLAT track

This URL gets stored as a cookie in your browser
Benefits to BayGenomics and the PGA

- Improves visibility and impact of our research activities
- Improves impact of NHLBI funding
- Provides better accessibility for those scientists who prefer genome-centric approach to browsing gene-trap data
Acknowledgements

Collaborators & Staff
- Prof. Patricia Babbitt, Dr. Conrad Huang, Dr. Elaine Meng, Susan Johns, Doug Stryke, Michiko Kawamoto, Patty Chang, Courtney Harper, Dr. Roy Lee, Dr. Stephen Young

NIH NHLBI
- U01-HL66621

NIH National Center for Research Resources
- P41-RR01081
QUESTIONS?!?