

# Expression Profiling

Mark Voorhies

4/4/2011

- Sequence analysis
- hmmbuild (not hmmfit)
- JackHMMer

It's hard work at times, but you have to be realistic. If you have a large database with many variables and your goal is to get a good understanding of the interrelationships, then, unless you get lucky, this complex structure is bound to require some hard work to understand.

Bill Cleveland and Rick Becker

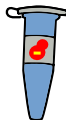
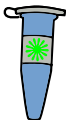
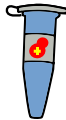
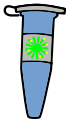
<http://stat.bell-labs.com/project/trellis/interview.html>

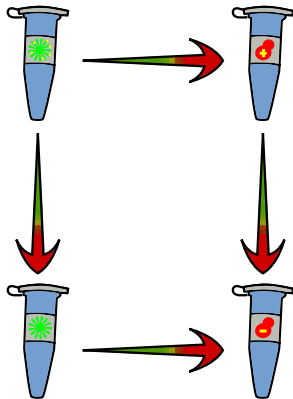
Why profile *transcription*?

Why profile *transcription*?

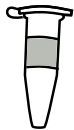
- Major mode of replication
- Due to feedback, “shadows” other modes of regulation
- Thanks to Watson-Crick base pairing, we can assay arbitrary nucleic acids in a uniform way

# Sample Preparation





# Pooled Reference Design





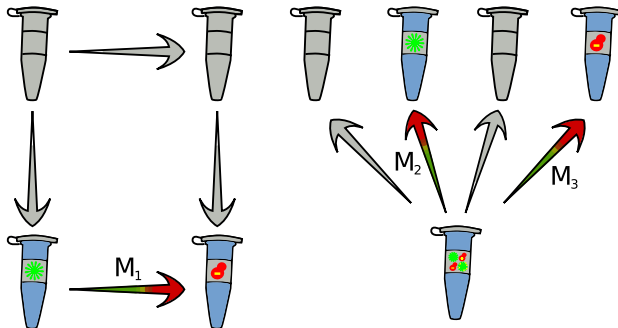
# Transforming Ratios



# Transforming Ratios

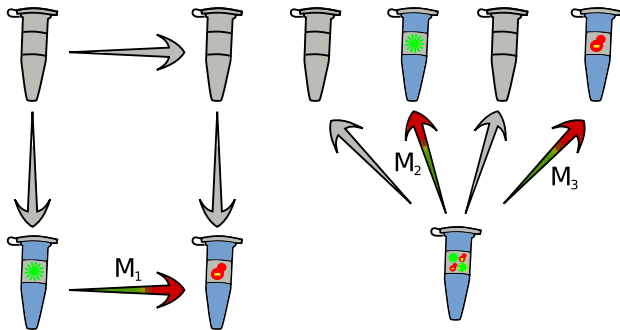


# Transforming Ratios



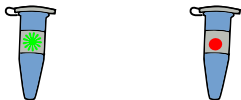
$$M_1 = M_3 / M_2$$

# Transforming Ratios



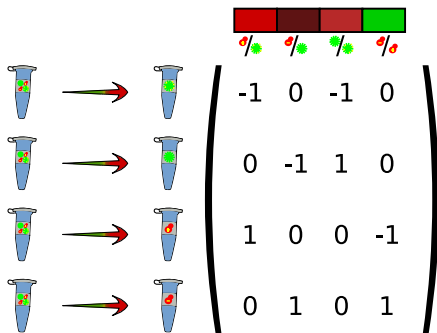
$$\log_2 M_1 = \log_2 M_3 - \log_2 M_2$$

# Linear Representation

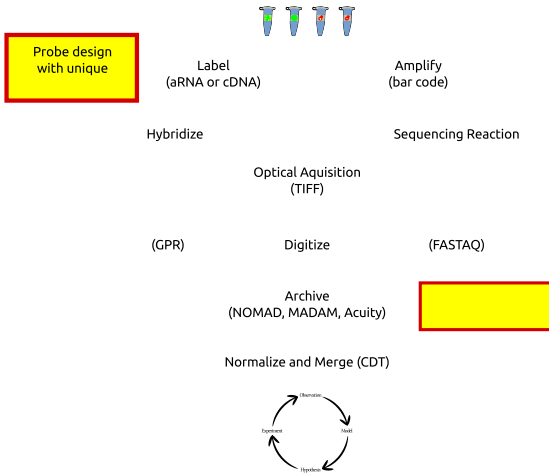


$$(1)\log_2 M_1 = (-1)\log_2 M_2 + (1)\log_2 M_3$$

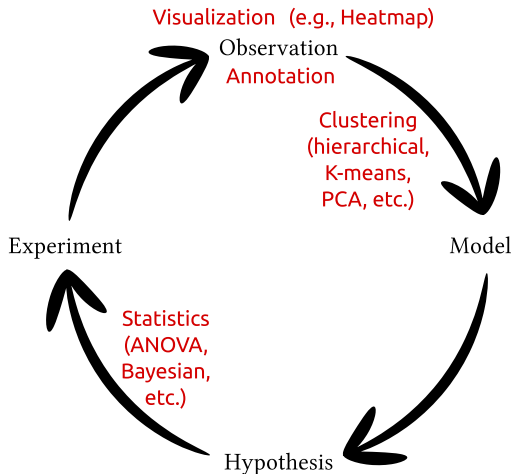
# Linear Representation



# Expression Profiling Workflow



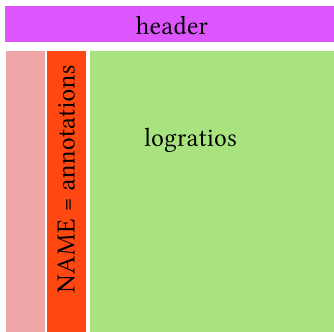
# Expression Profiling Analysis





# The CDT file format

## Minimal CLUSTER input

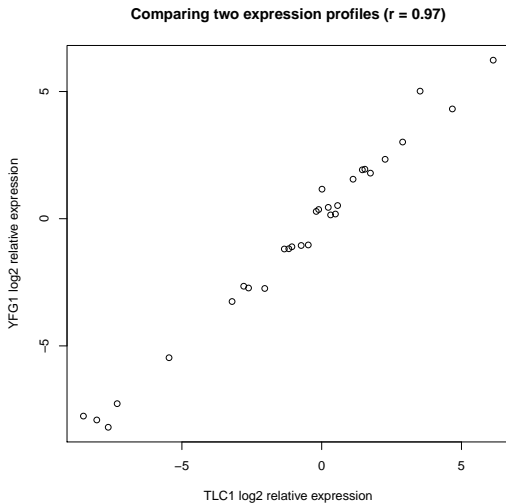


- Tab delimited (`\t`)
- UNIX newlines (`\n`)
- Missing values  $\rightarrow$  empty cells

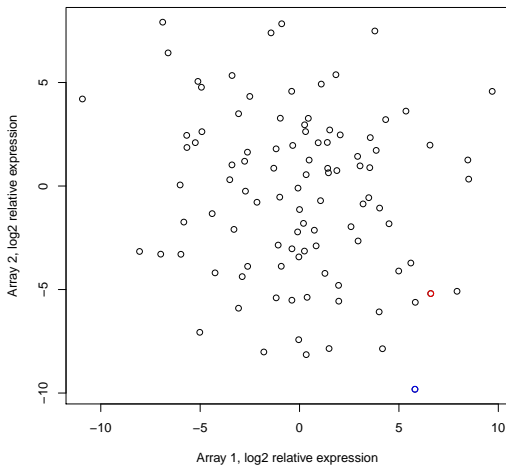
## Cluster3 CDT output



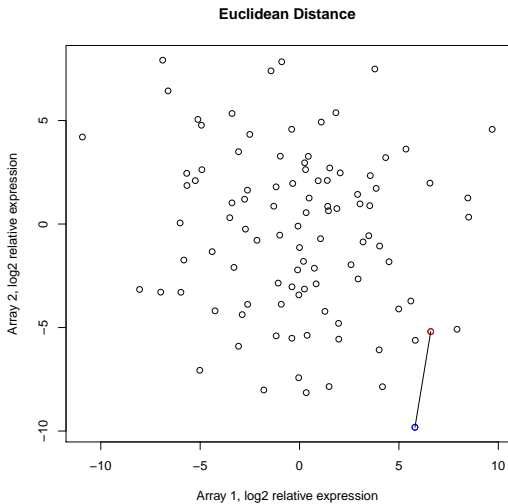
# Comparing all measurements for two genes



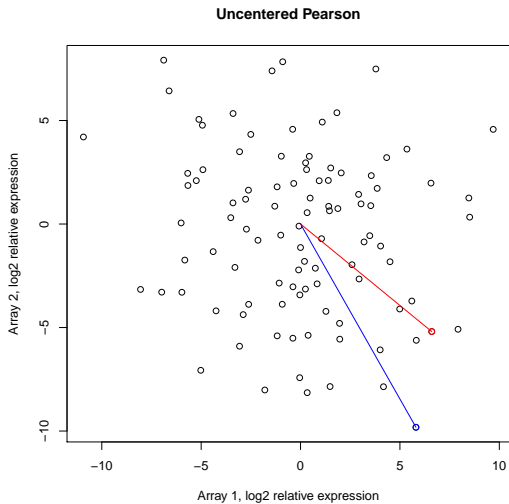
# Comparing all genes for two measurements



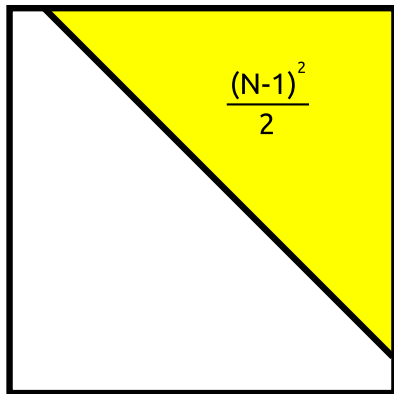
# Comparing all genes for two measurements



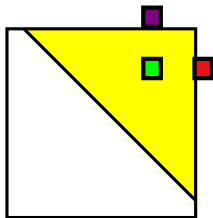
# Comparing all genes for two measurements



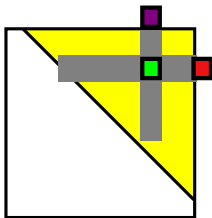
# Measure all pairwise distances under distance metric



# Hierarchical Clustering

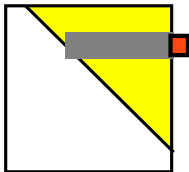


# Hierarchical Clustering





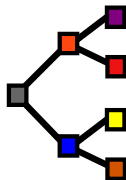
# Hierarchical Clustering



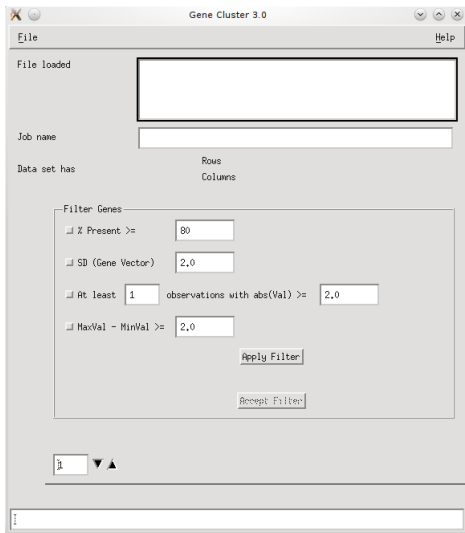
# Hierarchical Clustering



# Hierarchical Clustering



# Using the Cluster3 GUI



Gene Cluster 3.0

File Help

File loaded

Job name

Data set has Rows Columns

Filter Genes

- % Present  $\geq$  80
- SD (Gene Vector) 2,0
- At least 1 observations with  $\text{abs}(\text{Val}) \geq$  2,0
- $\text{MaxVal} - \text{MinVal} \geq$  2,0

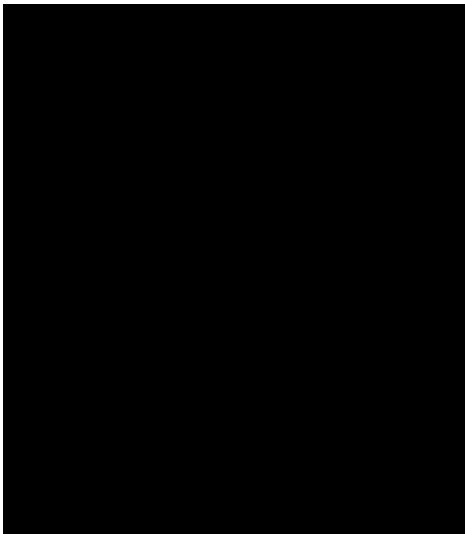
Apply Filter

Accept Filter

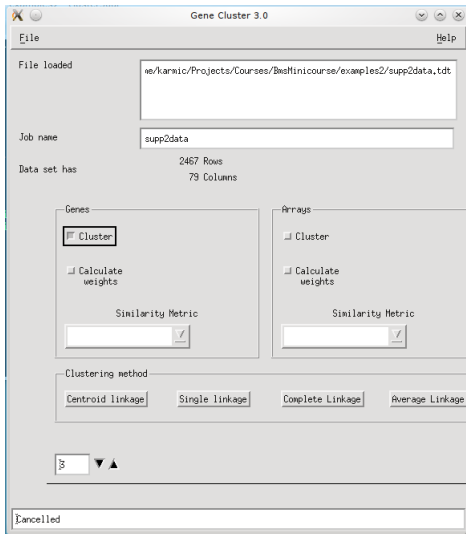
1

1

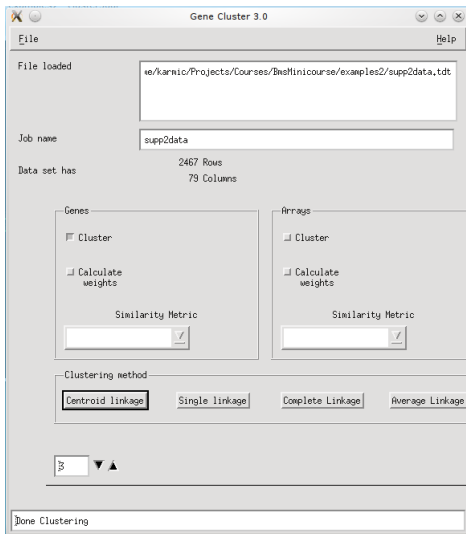
# Load your data



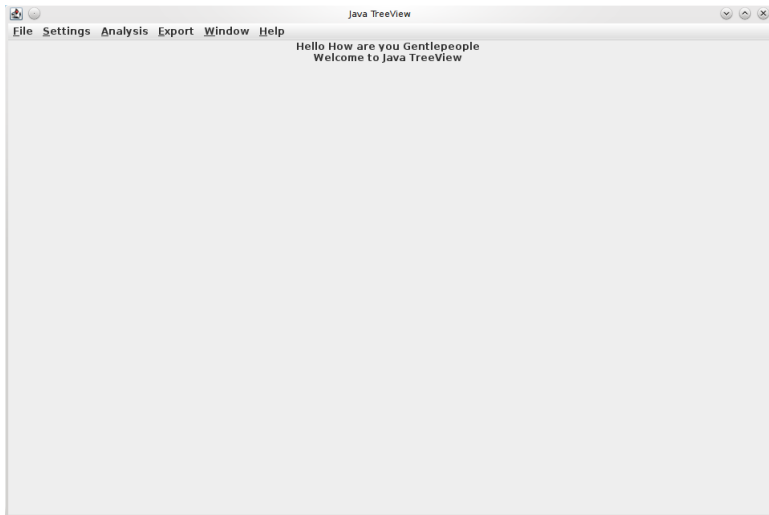
# Choose distance function



# Choose linking method

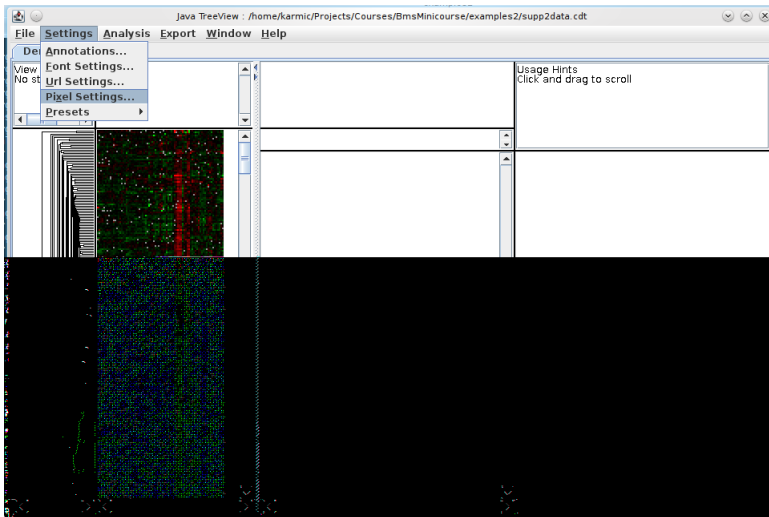


# Using JavaTreeView





# Adjust pixel settings for global view



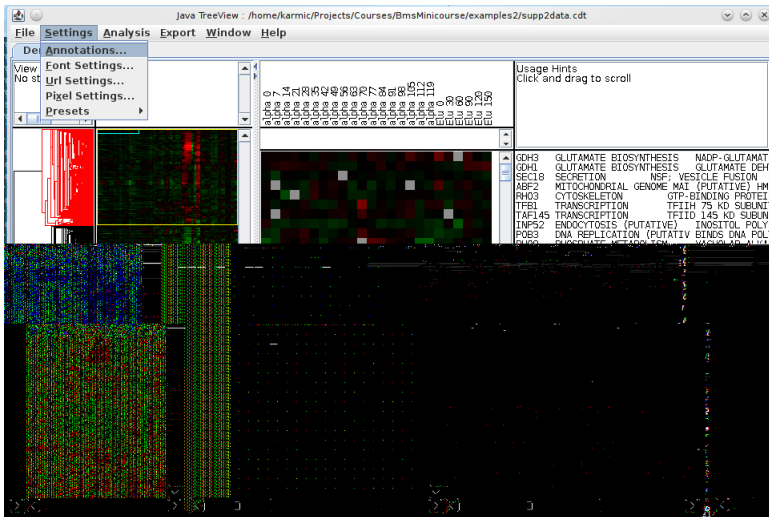
# Adjust pixel settings for global view

The screenshot shows the Java TreeView application window. The title bar indicates the file path: `/home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt`. The menu bar includes **File**, **Settings**, **Analysis**, **Export**, **Window**, and **Help**. The **Dendrogram** tab is active, showing a hierarchical tree structure on the left and a corresponding heatmap on the right. A **Pixel Settings** dialog box is open, displaying the following options:

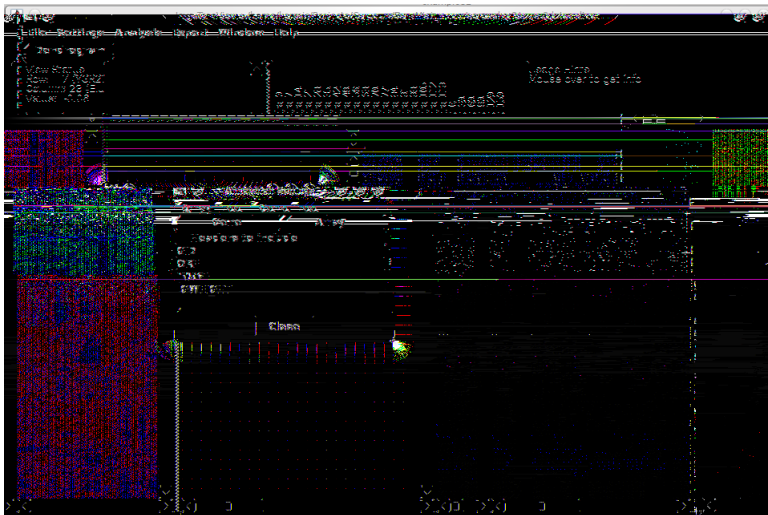
- Global:**
  - Fixed Scale: 481012658227
  - Fixed Scale: 663964329145
  - Fill
- X:** [input field]
- Y:** [input field]

The main heatmap displays a complex pattern of colored pixels (red, green, blue, yellow) against a black background, representing gene expression data. The x-axis is labeled **Class** and the y-axis is labeled **Gene**. The heatmap is overlaid with a dendrogram structure, and various control elements like zoom and pan are visible.

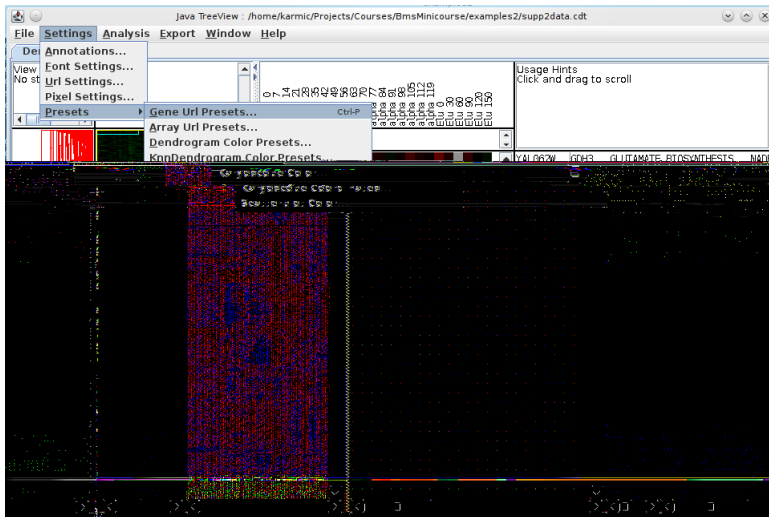
# Select annotation columns



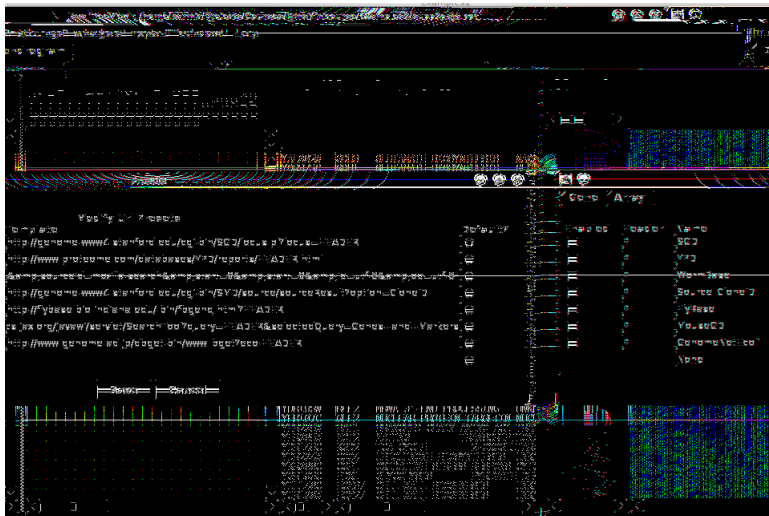
# Select annotation columns



# Select URL for gene annotations



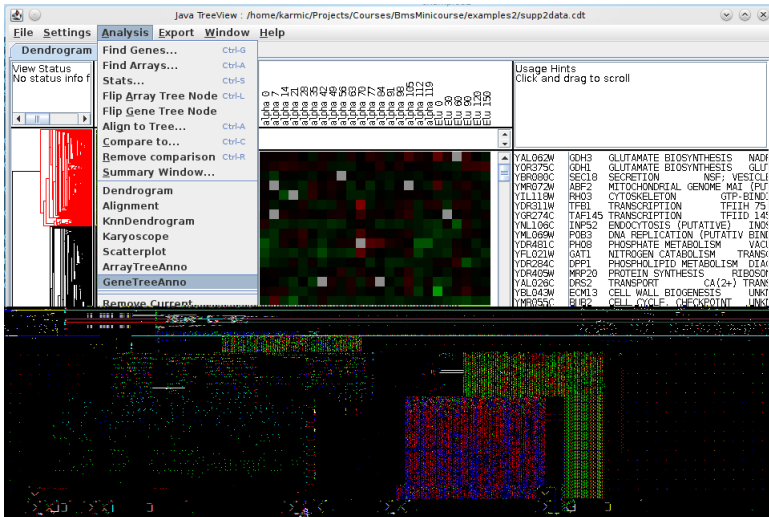
# Select URL for gene annotations



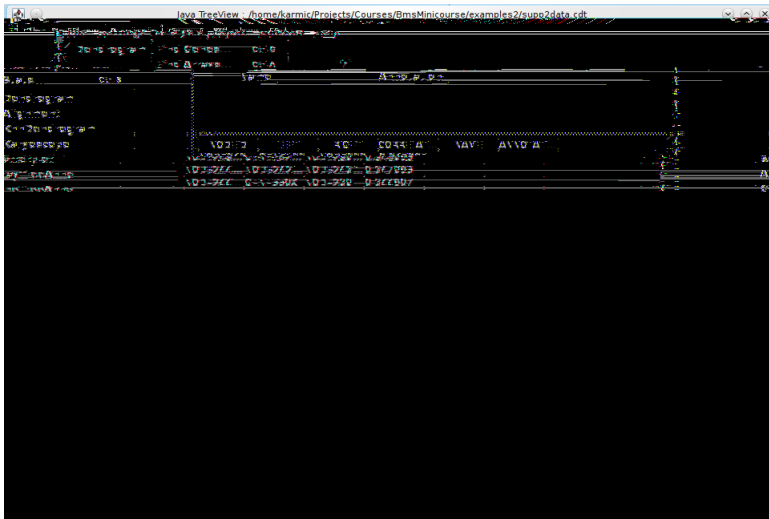
The screenshot displays a genome browser interface with a dark background. At the top, there are navigation icons and a search bar. Below the search bar, a track shows gene models with colored bars representing exons and introns. A vertical line indicates the current genomic position. Below the track, a table lists URLs for gene annotations, including Ensembl, RefSeq, and UniProt. The table has columns for 'Gene', 'Description', and 'URL'. The 'Gene' column lists genes like 'BRCA1', 'BRCA2', and 'TP53'. The 'Description' column provides brief descriptions of these genes. The 'URL' column contains links to the respective databases. At the bottom of the screenshot, there are navigation controls and a search bar.

Gene	Description	URL
BRCA1	BRCA1 protein	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene&amp;term=BRCA1">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene&amp;term=BRCA1</a>
BRCA2	BRCA2 protein	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene&amp;term=BRCA2">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene&amp;term=BRCA2</a>
TP53	TP53 protein	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene&amp;term=TP53">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Gene&amp;term=TP53</a>

# Activate and detach annotation window



# Activate and detach annotation window



The screenshot shows a Java TreeView window titled "java TreeView /home/karmic/Projects/Courses/BmsMinicourse/examples2/supo2date.cdt". The window displays a tree view of a project structure on the left and a table of data on the right. The table has several columns and rows of data, including numerical values and strings.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
102:7	102:7	102:7	102:7	102:7	102:7
102:7	102:7	102:7	102:7	102:7	102:7
102:7	102:7	102:7	102:7	102:7	102:7
102:7	102:7	102:7	102:7	102:7	102:7
102:7	102:7	102:7	102:7	102:7	102:7
102:7	102:7	102:7	102:7	102:7	102:7
102:7	102:7	102:7	102:7	102:7	102:7
102:7	102:7	102:7	102:7	102:7	102:7
102:7	102:7	102:7	102:7	102:7	102:7



# Activate and detach annotation window

