What's special about stem cells?

Using molecular profiling
to look at gene activity

George Bell and Fran Lewitter
Bioinformatics and Research Computing

2011 Spring Lecture Series for High School Students

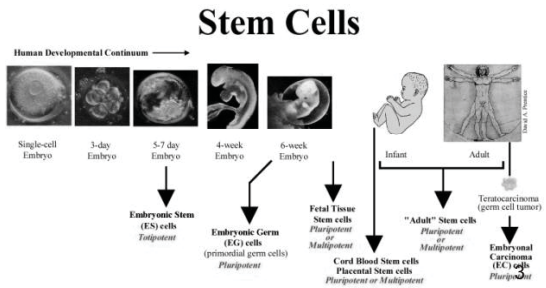
What we'll do today

- Research questions in stem cell biology
- Measuring gene expression levels
- Starting with gene levels in different stem cells and other cells
  - Identify most variable genes
  - Get relative gene levels
  - Cluster to group most similar genes and most similar cell types
- Compared to differentiated cells
  - What genes are changed in all stem cells?
  - What genes are changed in some types of stem cells?

Types of stem cells

- What is a stem cell, anyway?
  - ability to self-renew (and produce more stem cells)
  - ability to differentiate into different/any cell types
- Embryonic stem cells
- Adult stem cells
- Induced pluripotent stem cells

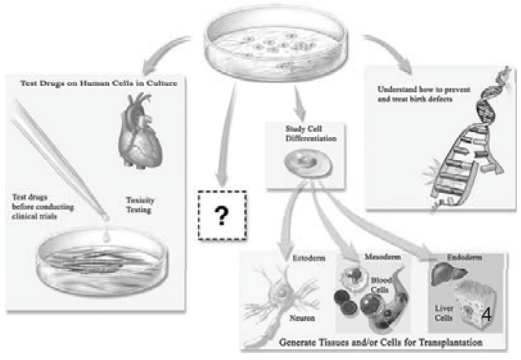
www.stemcellresearch.org



What genes are special in stem cells?

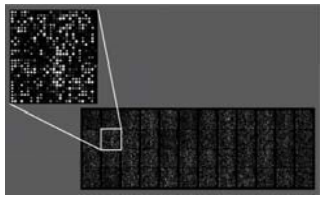
- Given that stem cells can self-renew and differentiate into many or all types of cells,
  - What genes are responsible for this behavior?
- Can these genes teach us about
  - Human development?
  - Cell division?
  - Differentiation?
  - Regenerating damaged tissue?

stemcells.nih.gov



Measuring levels of each gene

- DNA microarrays
  - Glass slides with up to millions of spots of short DNA sequences
  - When a solution of DNA (often converted from RNA) is added, genes stick to spots which are found in their sequence



- High-throughput sequencing
  - Convert RNA to DNA and break into small pieces
  - Read short DNA sequence from one or both ends

Bioinformatics

- Bioinformatics = the application of computational methods to the field of molecular biology
  - Also called Computational Biology
- More and more biology experiments include lots and lots of measurements so many biologists need to
  - Use computers to analyze data
  - Use statistics to help determine the confidence of any conclusions

Using math to understand biology

- Log transformations (bases) $\log_2 8 =$
- Median median {5,8,10,12} =
- Standard deviation (to measure variability)
- Mean/median centering
- Log (A/B) = log A – log B

values	32	256	1024	4096
log2	5	8	10	12
Median centered	5 - 9 = -4	8 - 9 = -1	10 - 9 = 1	12 – 210 =

Sources of expression data

- The final step of an experiment is usually publishing the project in a journal
- When a project is published, all of the data may be made public so
  - Others can verify the findings
  - Others can use the data to help with their research
- The National Center for Biotechnology Information (NCBI) hosts much of this data on their web sites
- Today's data is from:
  - Whitehead Institute (Guenther et al., 2010)
  - UCLA (Lowry et al., 2008; Chin et al., 2009)

Matrix of gene "expression levels"

- Each column represents a cell type
- Each row represents a gene
  - The levels of some genes is measuring from more than one spot (probe) on the microarray

Probe	Gene symbol	Fibroblasts	ES cells	iPS cells
220184_at	NANOG	5	11	12
208286_x_at	POU5F1	7	13	12
228038_at	SOX2	4	12	13

Matrix of gene "expression levels" (details)

- Open Expression_log2_values_HS_2011.xlsx and look at
  - The first row
  - The first two columns
  - What information is shown?
- The numbers represent
  - Levels of mRNA
  - Measured by the amount of dye-bound DNA that binds to a DNA probe (a spot containing a gene tag)
  - Log2-transformed
    Since $2^{10} = 1024$,
    A RNA level of 1024 has been converted to 10

To do – Select most variable genes

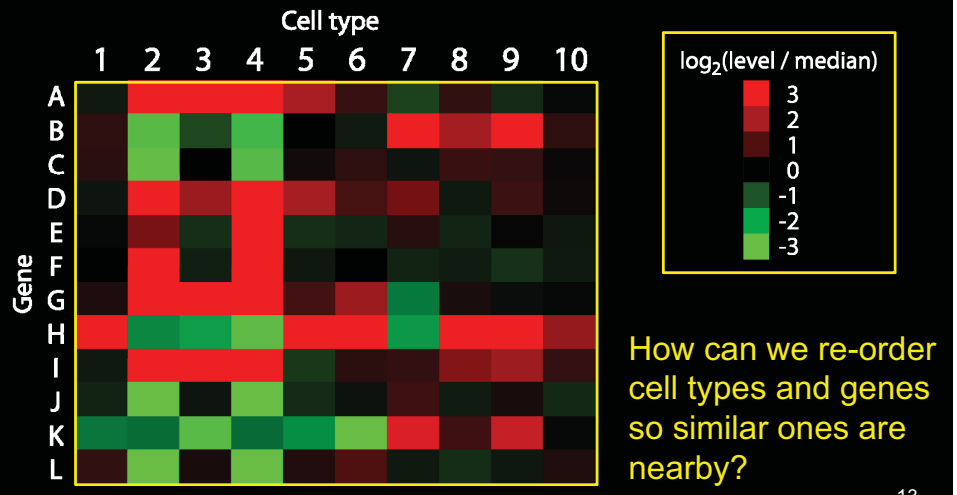
1. Open the matrix of log2 expression values (Expression_log2_values_HS_2011.txt) in Cluster.
   - File => Open data file
   - How many genes are you starting with?
2. Filter data (to remove genes with relatively constant levels
   - Check "SD (Gene Vector)"
   - Enter 1.5 in following box (to filter out genes with a standard deviation < 1.5)
   - Click on "Apply Filter"
   - How many genes remain after filtering?
   - When complete, click on "Accept [Filter]"

To do – get relative gene levels

While in Cluster

- Click on the "Adjust Data" tab
- Check "Center genes" and select "Median" which will
  - Find the median gene level of each gene in all cell types
  - Subtract the median from each gene level which will set it
    - to 0 if its level is the median level
    - to a positive value x (if 2^x -fold above the median)
    - to a negative value y (if 2^{-y} -fold below the median)
  - Help us identify genes that increase or decrease their levels
- Click on "Apply" to center your expression matrix
- Look at the bottom of the program to make sure it says "Done adjusting data"

Displaying a matrix as a heatmap

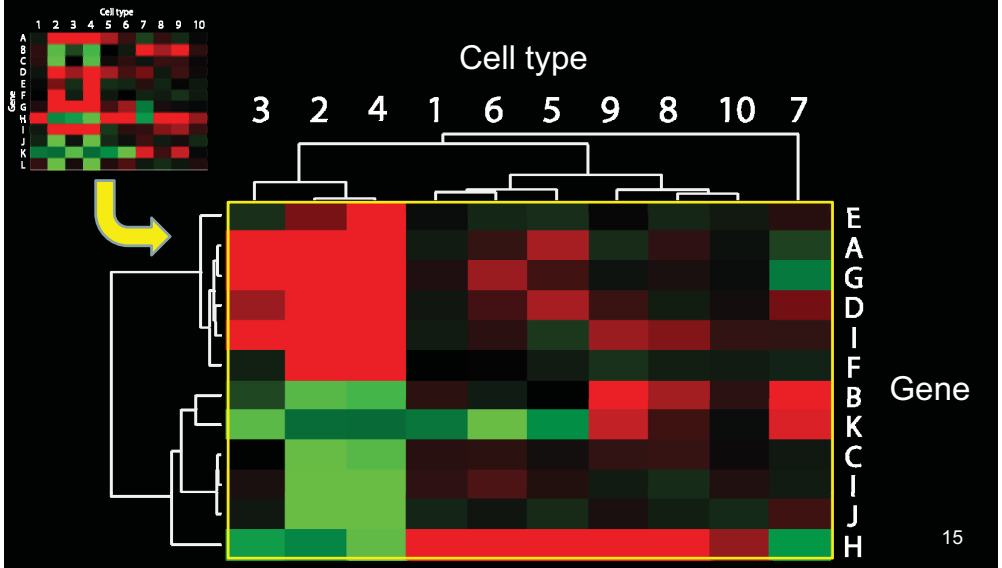


How can we re-order cell types and genes so similar ones are nearby?

Clustering an expression matrix

- To be able to better see changes in the levels of genes, we want to
  - Re-order genes so similar ones are closer
  - Re-order cell types so similar ones are closer
- For genes, our program is going to
  - compare every gene (row of numbers) to every other gene
  - Draw a tree showing how close each gene's "profile" is to every other gene's profile
  - Use the tree to order the genes in a new matrix
- The same thing will be done with cell types (columns)

Hierarchical clustering output



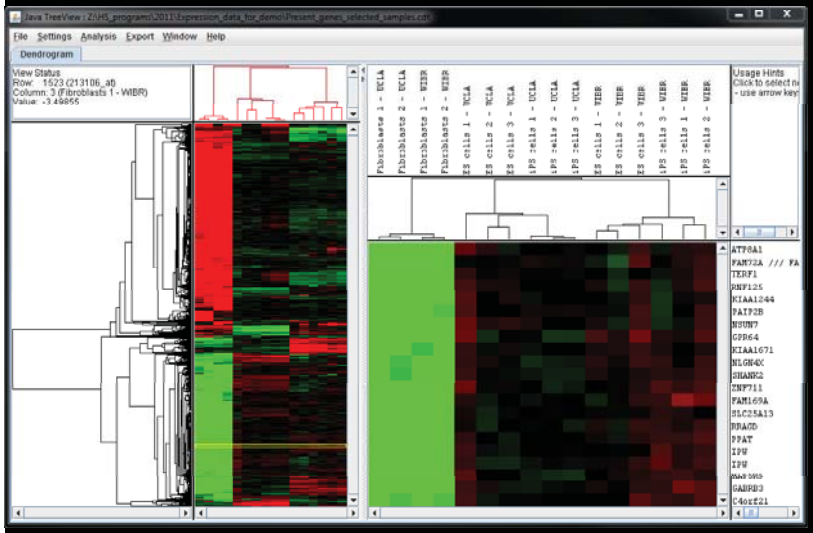
To do – cluster by gene and cell type

- While in Cluster
  - Click on the "Hierarchical" tab
  - In the Genes square, check "Cluster"
  - In the Arrays (cell types) square, check "Cluster"
  - To perform the actual clustering to re-order rows and columns,
    - Click on the "Average linkage" box
  - Look at the bottom of the program to make sure it says "Done Clustering"

To do – open your clustered expression matrix

1. Open the program Java Treeview by double-clicking on it
2. Open your clustered expression matrix
   - File => Open
   - Select the cdt file that you created with Gene Cluster
3. [Click on “Dismiss” if necessary]
4. With your mouse select a region of the colored panel at left.
5. What are you looking at?

To do – examine your heatmap



By convention:

Red = Higher than the median for this gene

Green = Lower than the median for this gene

What genes are changed in stem cells?

- Compared to fibroblasts, how do the levels of these genes change?
- Go to Analysis > Find Genes
  - Pou5f1
  - Sox2 (multiple probes)
  - Nanog
  - Xist (multiple probes)
  - Runx1
  - Fgf7 (multiple probes)
  - Twist2 (multiple probes)

How do ES and iPS cells compare?

- Compared to embryonic stem cells, are any genes
  - Higher in iPS cells?
  - Lower in iPS cells?
- Which genes seem to be laboratory-specific?
- Any other interesting expression patterns?

Getting information about genes

Using the gene symbol

- Method 1: Search the Web with the gene symbol
- Method 2: NCBI Gene
  - http://www.ncbi.nlm.nih.gov/gene/
- Method 3: GeneCards
  - http://www.genecards.org/

- Access these resources via Java Treeview
  - Go to Settings > URL Settings
  - In the box, type
    - http://www.ncbi.nlm.nih.gov/gene?term=HEADER
    - http://www.genecards.org/cgi-bin/carddisp.pl?gene=HEADER
  - Select NAME
  - Check “Enable” and click on “Close”
  - Back on the detailed heatmap, click on a gene symbol

Summary

- Gene expression profiles can be used to examine gene activity
  - Microarrays
  - High-throughput sequencing
- Many genes are expressed at a different level in stem cells compared to differentiated cells
  - Some genes are consistent in all stem cells
  - Some genes seem to be different in some types of stem cells
- Current research addresses:
  - Which of these “stem cell genes” are biologically important?
  - Exactly what do these important genes do?