Readings for Module 6: Clustering.

Applications:

- Spellman P.T., Sherlock G., Zhang M.Q., Vishwanath R.I., Anders K., Eisen M.B., Brown P.O., Botstein D. (1998), Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization, *Molecular Biology of the Cell* **9**:3273-3297.
- Chu S., DeRisi J., Eisen M.B., Mulholland J., Botstein D., Brown P.O., Herskowitz I. (1998). The Transcriptional Program of Sporulation in Budding Yeast. *Science* **282**:699-705.
- Iyer V.R., Eisen M.B., Ross D.T., Schuler G., Moore T., Lee J.F.C., Trent J.M., Staudt L.M., Hudson J., Boguski M.S., Lashkari D., Shalon D., Botstein D., Brown P.O. (1999). The Transcriptional Program in the Response of Human Fibroblast to Serum. *Science* **283**:83-87.
- Gasch A.P., Spellman P.T., Kao C.M., Carmel-Harel O., Eisen M.B., Storz G., Botstein D., Brown P.O. (2001). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell* **11**:4241-4257.
- Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998). Cluster Analysis and Display of Genome-Wide Expression Patterns. *PNAS* **95**:14863-14868.
- Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., Golub T.R. (1999). Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and an Application to Hematopoietic Differentiation, *PNAS* **96**:2907-2912.
- Tavazoie S., Hughes J.D., Campbell M.J., Cho R.J., Church G.M. (1999). Systematic Determination of Genetic Network Architecture. *Nature Genetics* **22**:281-285

Principal components and clustering:

- Yeung K.Y., Ruzzo W.L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics* **17** (9) 762-744.

Computational methods and choice of number of clusters:

- McShane L.M., Radmacher M.D., Freidlin B., Yu R., Li M.C., Simon R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* **18**:1462-1469.
- Ben-Hur A. Elisseeff A. and Guyon E. (2002). A stability based method for discovering structure in clustered data. *Pac. Symp. Biocomputing 02*.
- Ben-Hur A., Guyon I. (2003). Detecting stable clusters using principal component analysis. *Functional Genomics: Methods and Protocols*, Brownstein and Khodursky eds (Methods in Molecular Biology Series, Vol 224) (Humana Press),159-182.
- Dudoit S., Fridlyand J. (2002). A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biology* **3**(7):0036.
- Dudoit S., Jane Fridlyand (2003), "Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* **19**:1090-1099.
- Tibshirani R., Walther G. and Hastie T. (2001). Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society B* **63**:411-423.
- Kerr M.K. and Churchill G. (2000). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *PNAS* **98**:8961-8965.

Mixture-based clustering:

- Yeung K.Y., Fraley C., Murua A., Raftery A.E. and Ruzzo W. L. (2001). Model-Based Clustering and Data Transformation for Gene Expression Data, *Bioinformatics* **17** (10) 977-987.
- McLachlan G.J., RW Bean, D Peel (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**:413-422.

Other:

- Lazzeroni L., Owen A. (2002). Plaid models for gene expression data. *Statistica Sinica* **12**(1):61-86.
- Gasch A.P., Eisen M.B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology* **3**(11):research0059.
- Bar-Joseph Z., Demaine E.D., Gifford D.K., Srebro N., Hamel A.M., Jaakkola T.S. (2003). K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics* **19**:1070-1078.
- Peddada S., Lobenhofer E.K., Li L., Afshari C.A., Weinberg C.R. and Umbach D.M. (2003). Gene selection and clustering for time-course and dose–response microarray experiments using order-restricted inference. *Bioinformatics* **19**: 834-841.

General review:

- Goldstein D.R., Conlon E., Ghosh D. (2002). Statistical issues in the clustering of gene expression data. *Statistica Sinica* **12**(1):219-240.