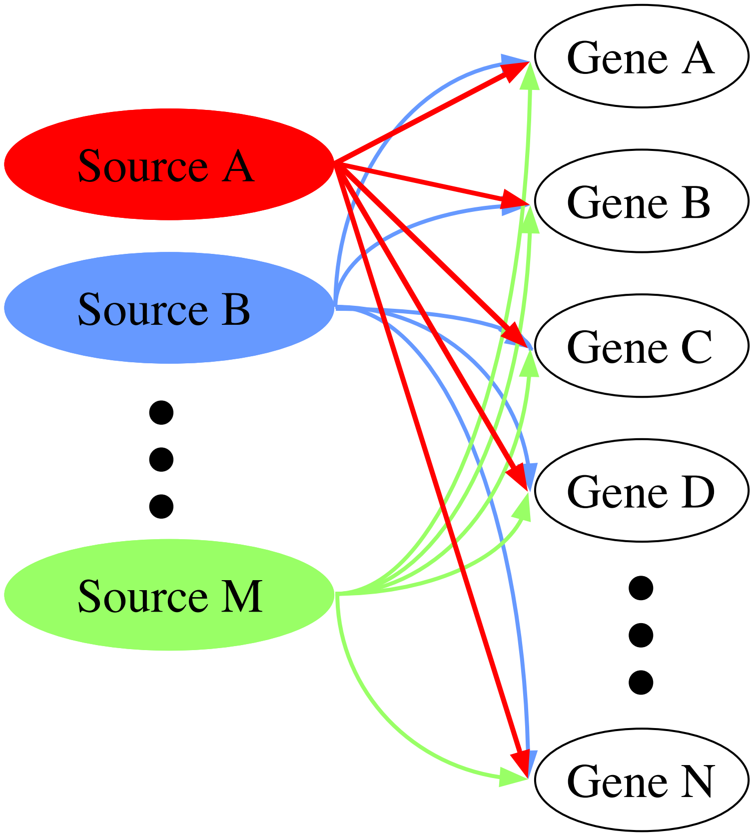
SVD Modeling of Genome-Wide Expression Data

Alter, Brown & Botstein, *PNAS* 97, 10101 (2000);
http://www.alterlab.org/singular_value_decomposition/

Alter, Brown & Botstein, *Proc SPIE* 4266, 171 (2001); http://www.alterlab.org/SVD/

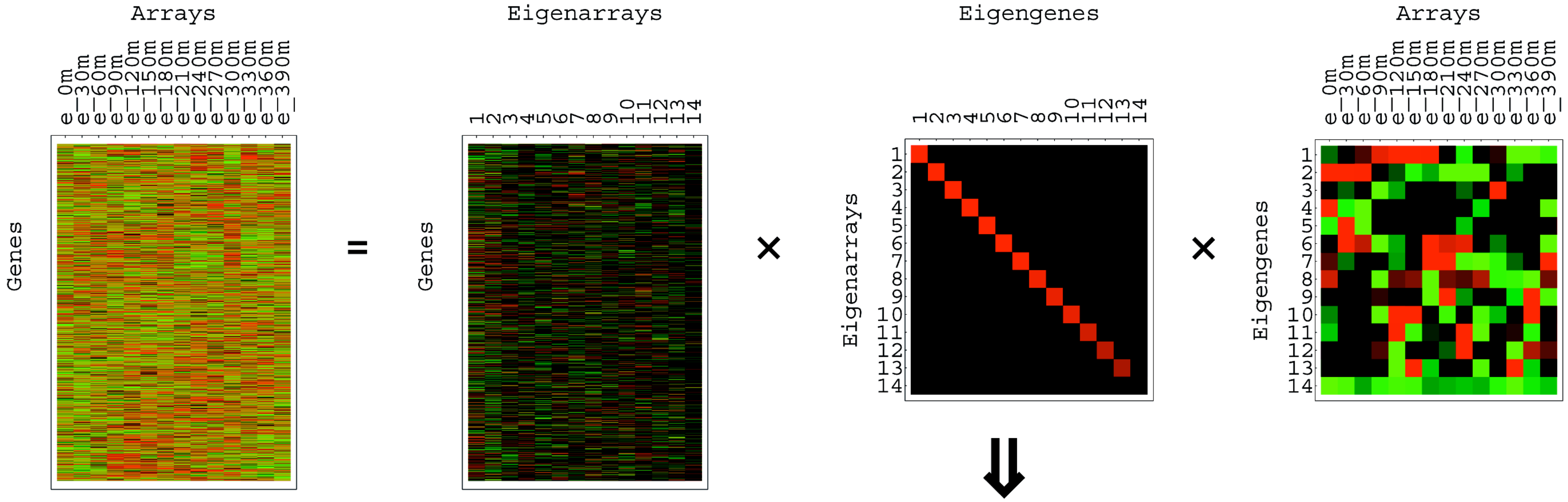
SVD formulates genome-wide expression as a superposition of the genome-wide effects of several independent sources of expression, such as regulatory programs, biological processes and experimental artifacts.

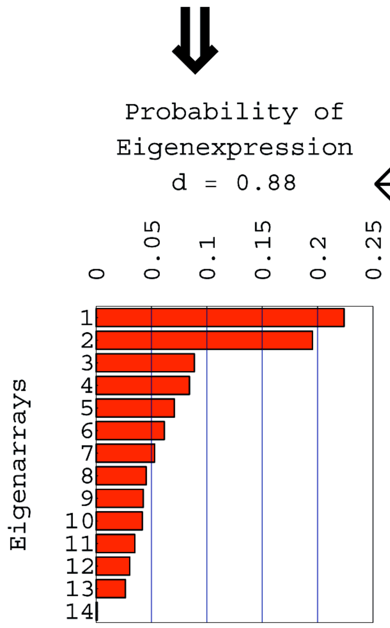
→ Data Normalization
→ Data Classification
→ Additional Data Incorporation



Singular Value Decomposition (I)

Linear transformation of gene expression data from genes × arrays space to reduced diagonalized “eigengenes” × “eigenarrays” space.





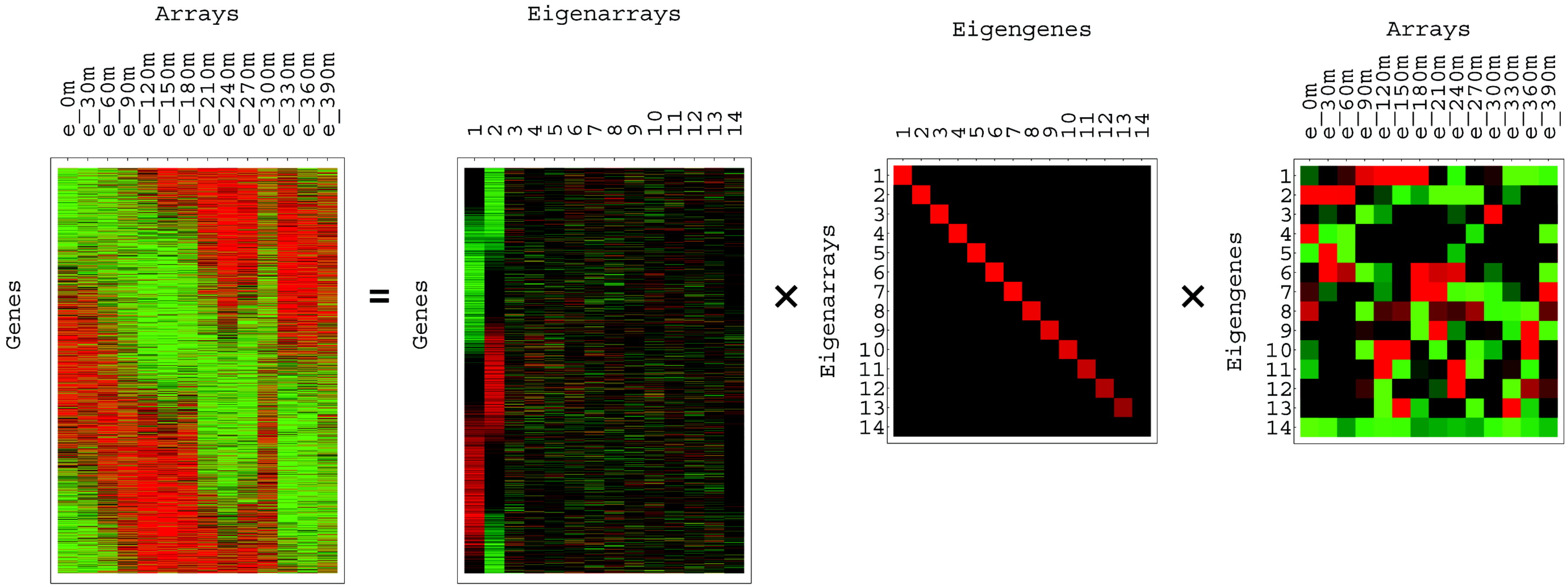
⟵ Normalized Shannon entropy

Yeast Cell Cycle: Elutriation

Spellman et al., *MBC* 9, 3273 (1998).

Singular Value Decomposition (II)

The eigengenes and eigenarrays are data-driven unique (except for a phase of ±1 and in degenerate subspaces), orthonormal (decorrelated and normalized), and decoupled superpositions of genes and arrays.



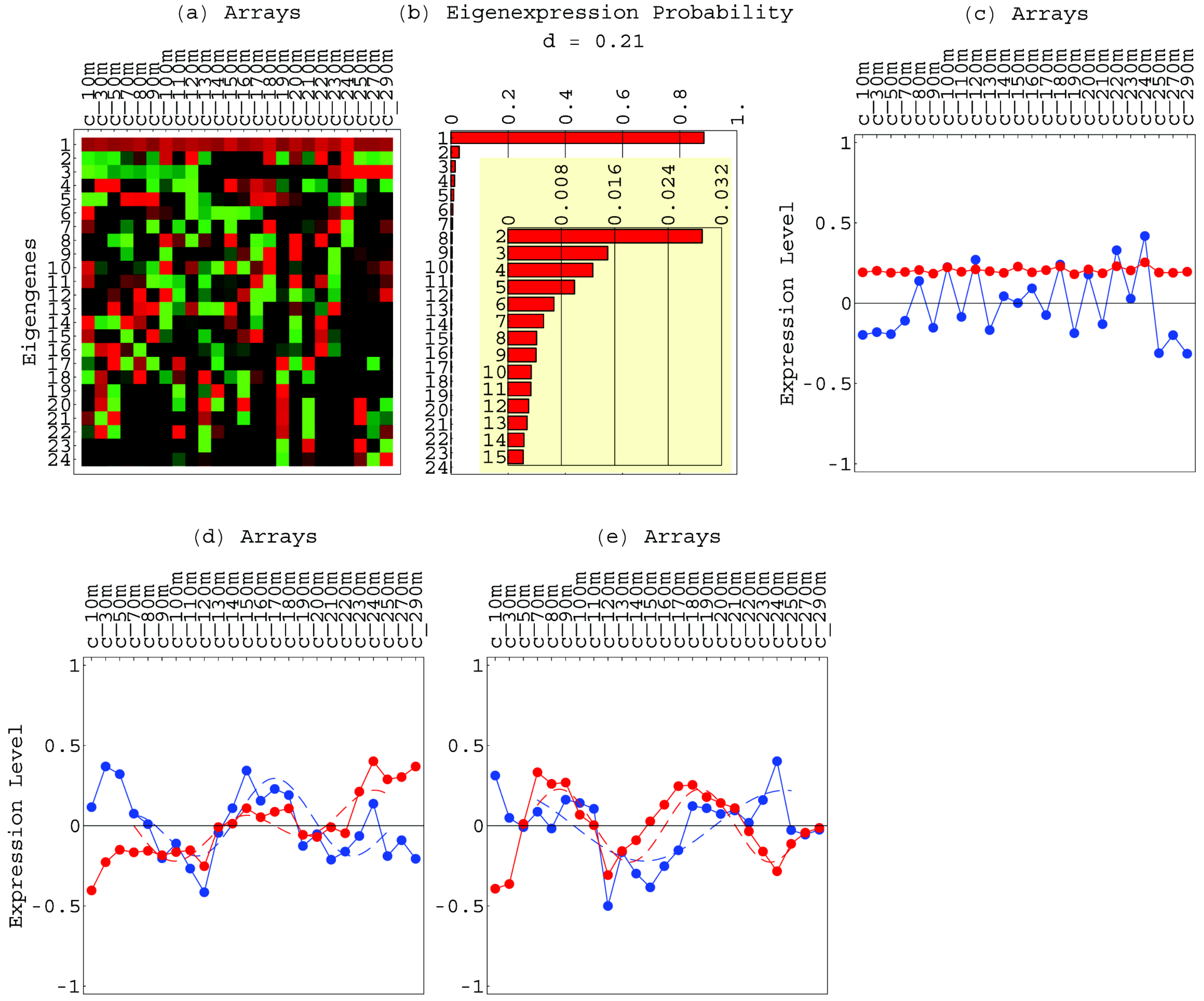
Does the decomposition of the genes (and arrays) expression to eigengenes (and eigenarrays) unravel the biological generation of the expression signal as a superposition of several cellular processes, biological and experimental (and the corresponding cellular states)?

Math Variables → Biology

Significant eigengenes → independent biological processes and experimental artifacts:

90% of expression is steady state,
2.5% is day-of-hybridization artifact,
less than 7.5% is periodic →

Weak Signal Detection



Yeast Cell Cycle: Cdc15 Spellman et al., *MBC* 9, 3273 (1998).

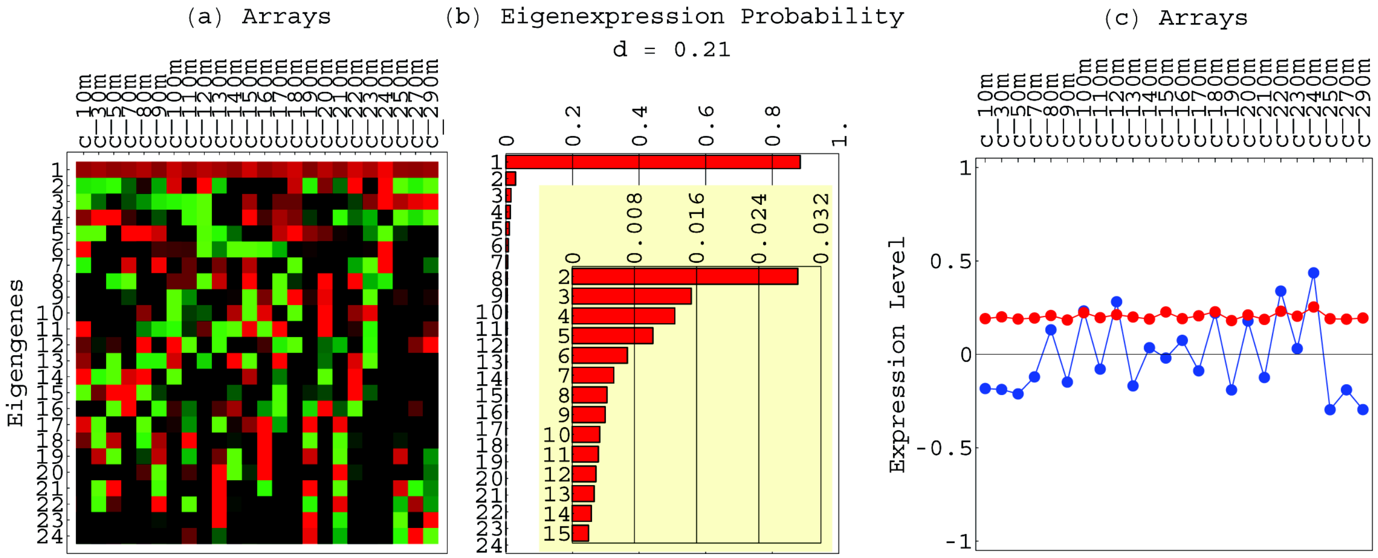
Math Operations → Biology

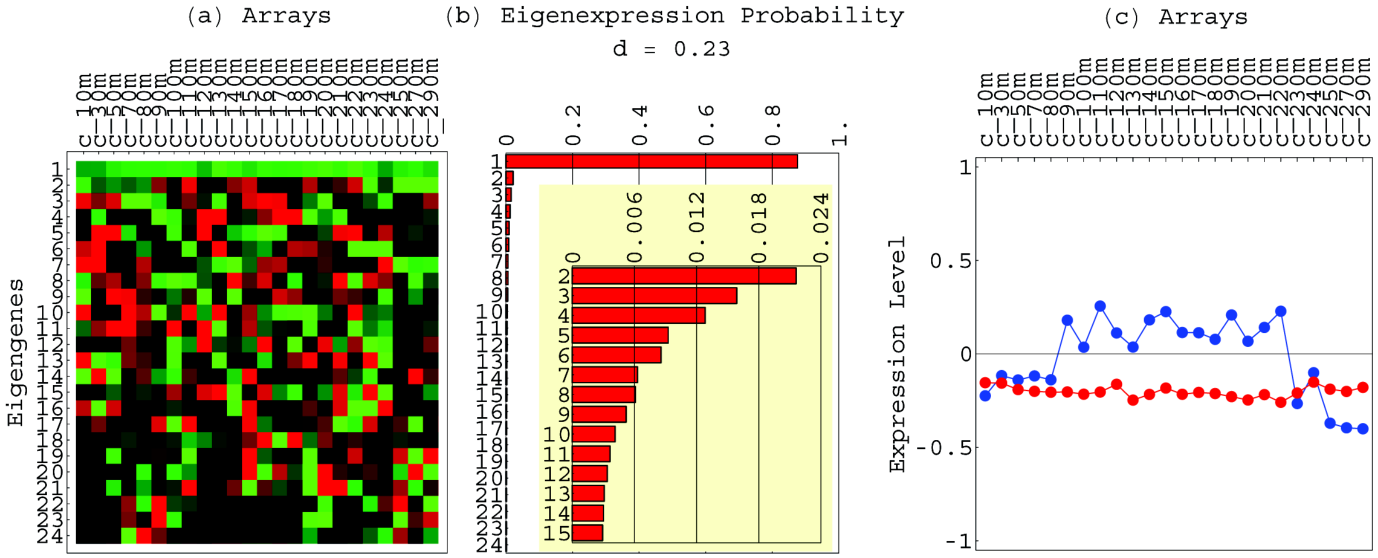
Detection of artifacts →

Filtering data without eliminating genes or arrays:

Normalization

Center data at additive steady state (and filter out additive day-of-hybridization artifact) …



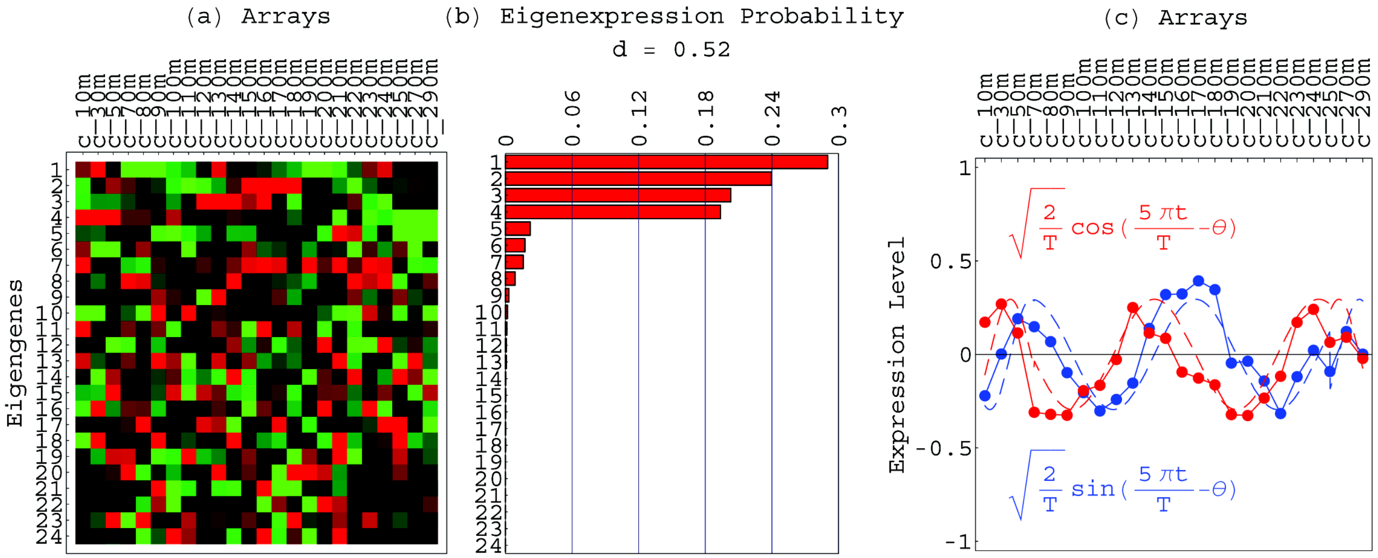


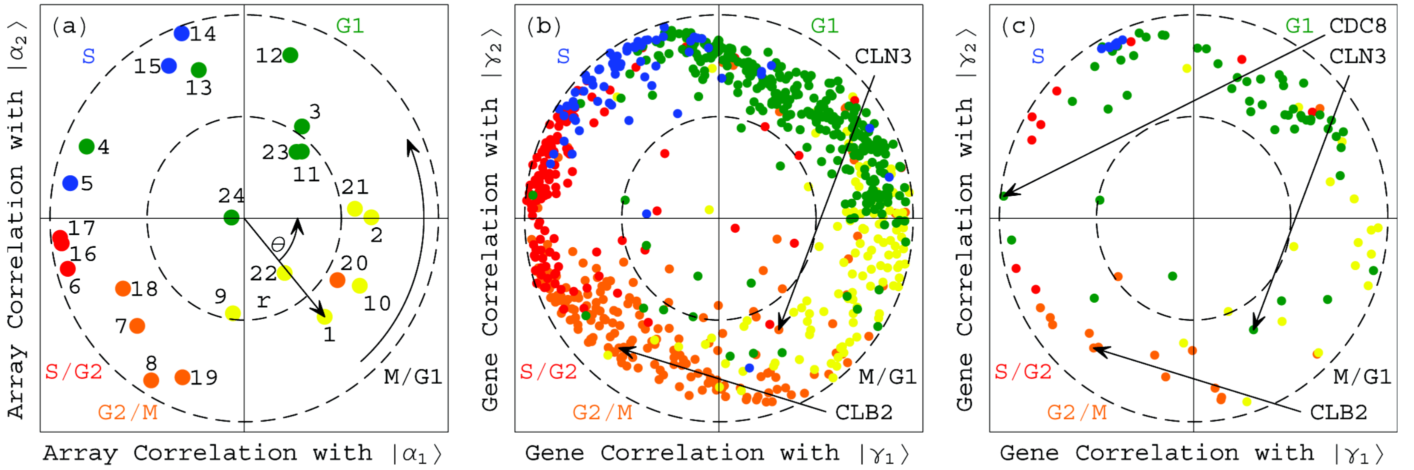
… and normalize by multiplicative steady variance (and filter out multiplicative day-of-hybridization artifact).

Math Variables → Biology

Two dominant and periodic eigengenes of similar significance, and corresponding eigenarrays, span the

Cell Cycle Subspace:



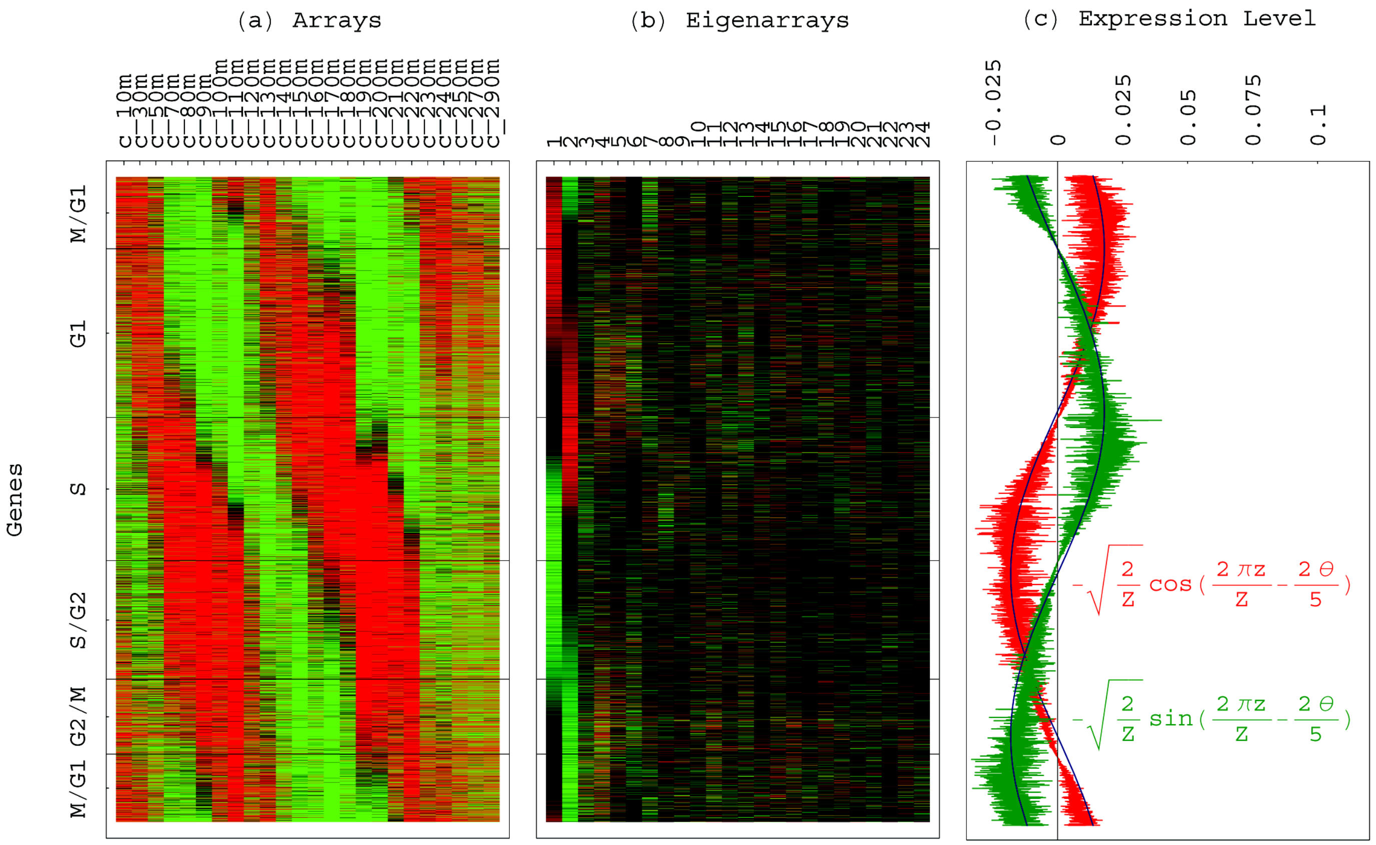


Math Operations → Biology

Detection of biological signals → sorting the data according to the eigengenes and eigenarrays, rather than overall expression:

Classification

Traveling Wave of Expression



Consistent model for the expression of almost the full yeast genome during cell division.

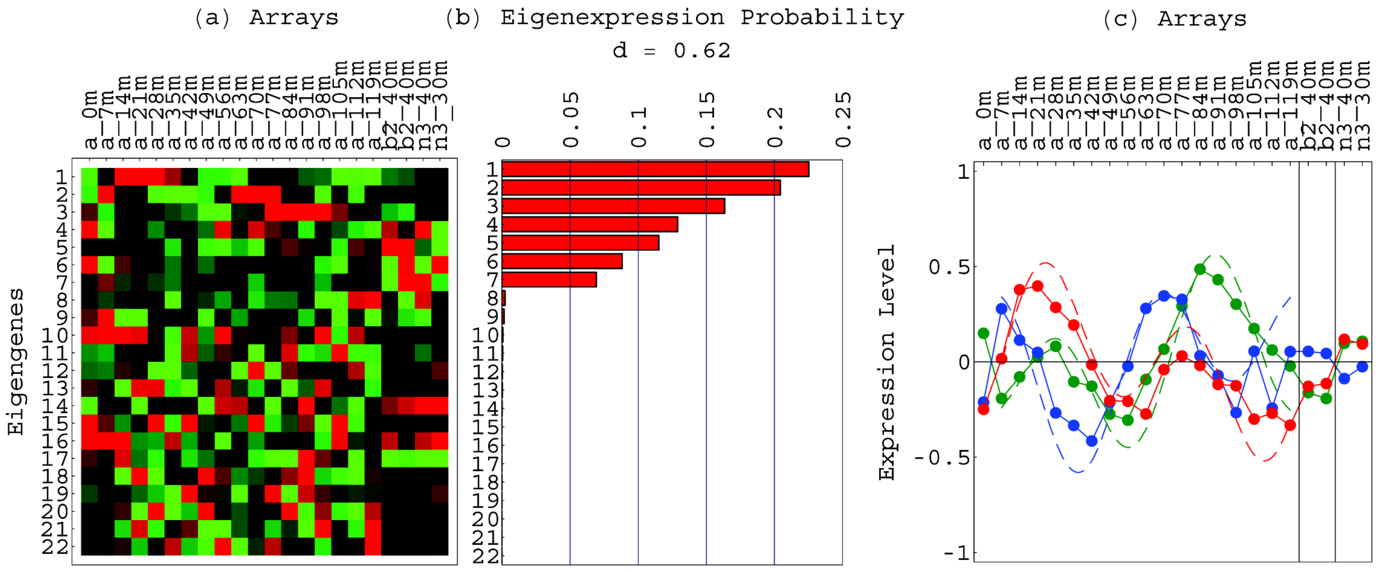
- → Which yeast genes exhibit periodic expression during the cell cycle?
- → Is there a genome-wide relation between DNA replication and RNA transcription?

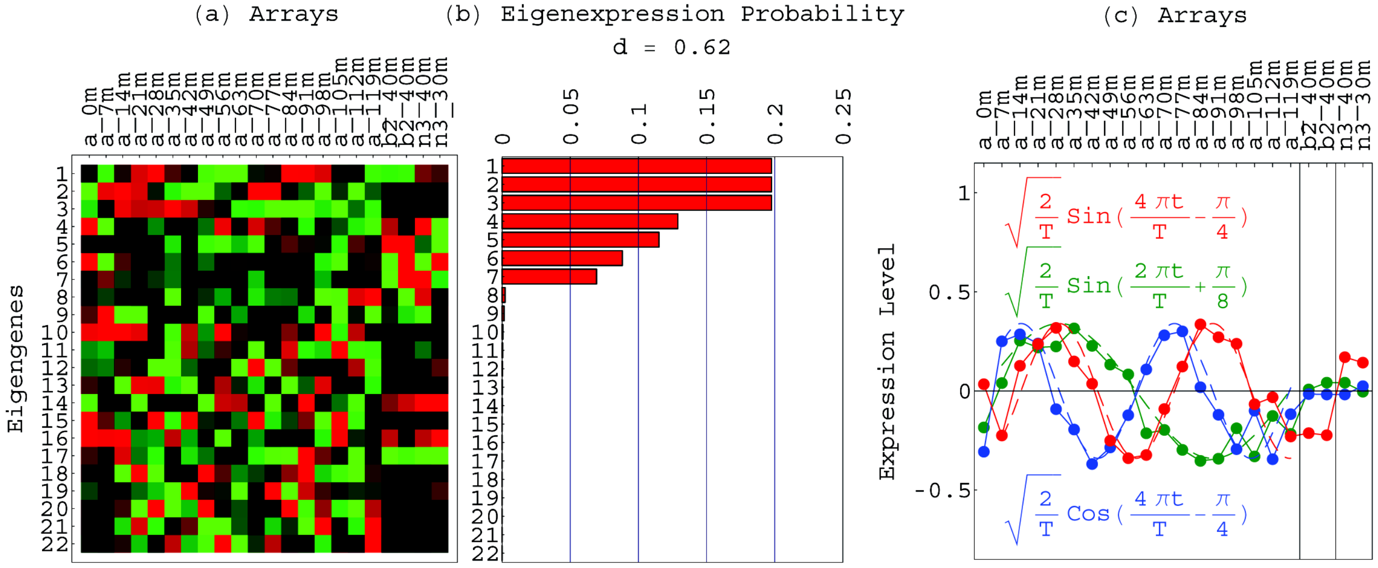
Math Operations → Biology

Degeneracy of eigengenes (eigenarrays) subspace →
Unique rotation of eigengenes (eigenarrays) for better data interpretation and presentation:

Data Comparison and Integration

Cln3, Clb2 genome–wide effects = ± first eigengene





Yeast Cell Cycle: Alpha Factor, Clb2 & Cln3

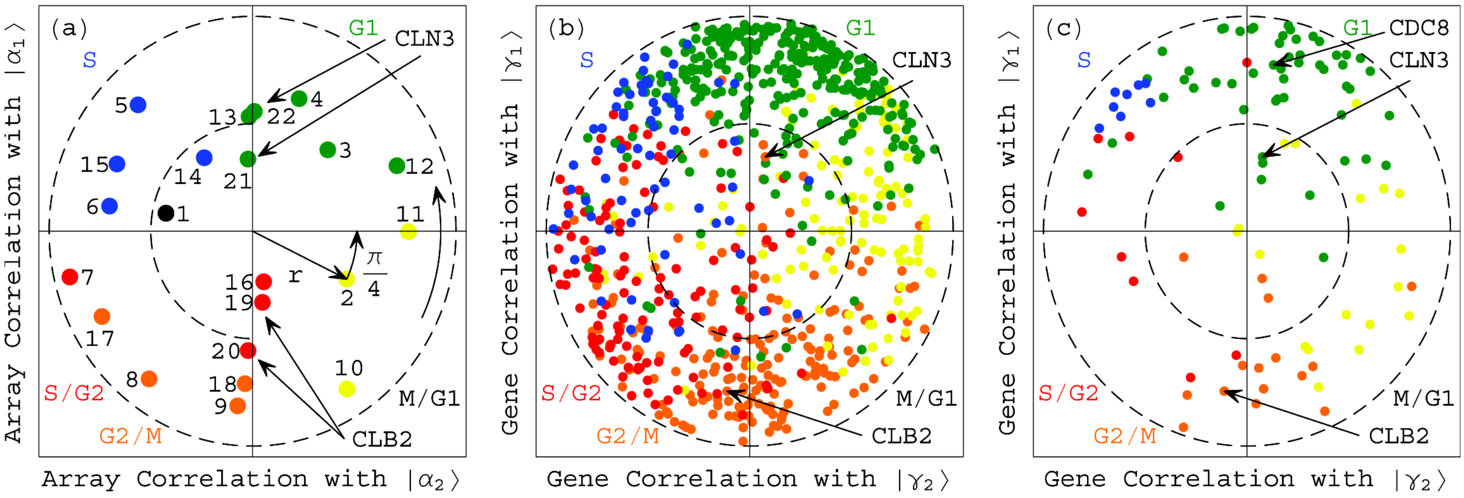
Spellman et al., *MBC* 9, 3273 (1998).

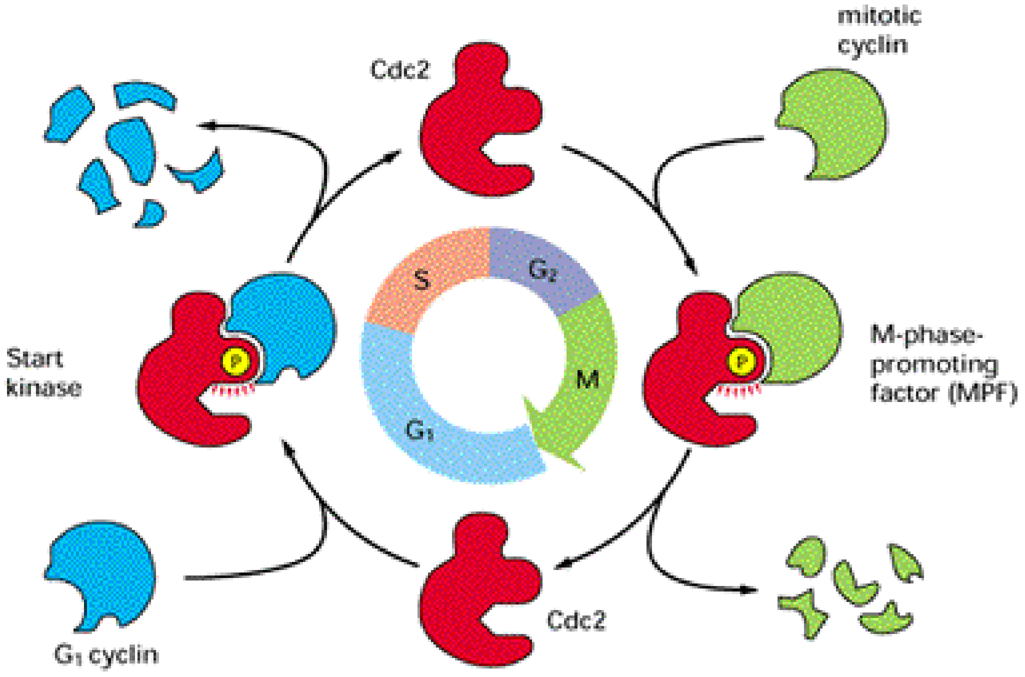
Math Variables → Biology

Significant eigengenes and eigenarrays → genome-wide effects of regulators, and samples in which these regulators are overactive, respectively:

Cln3, Clb2 genome–wide effects = ± first eigengene

Cln3, Clb2 overactive samples = ± first eigenarray



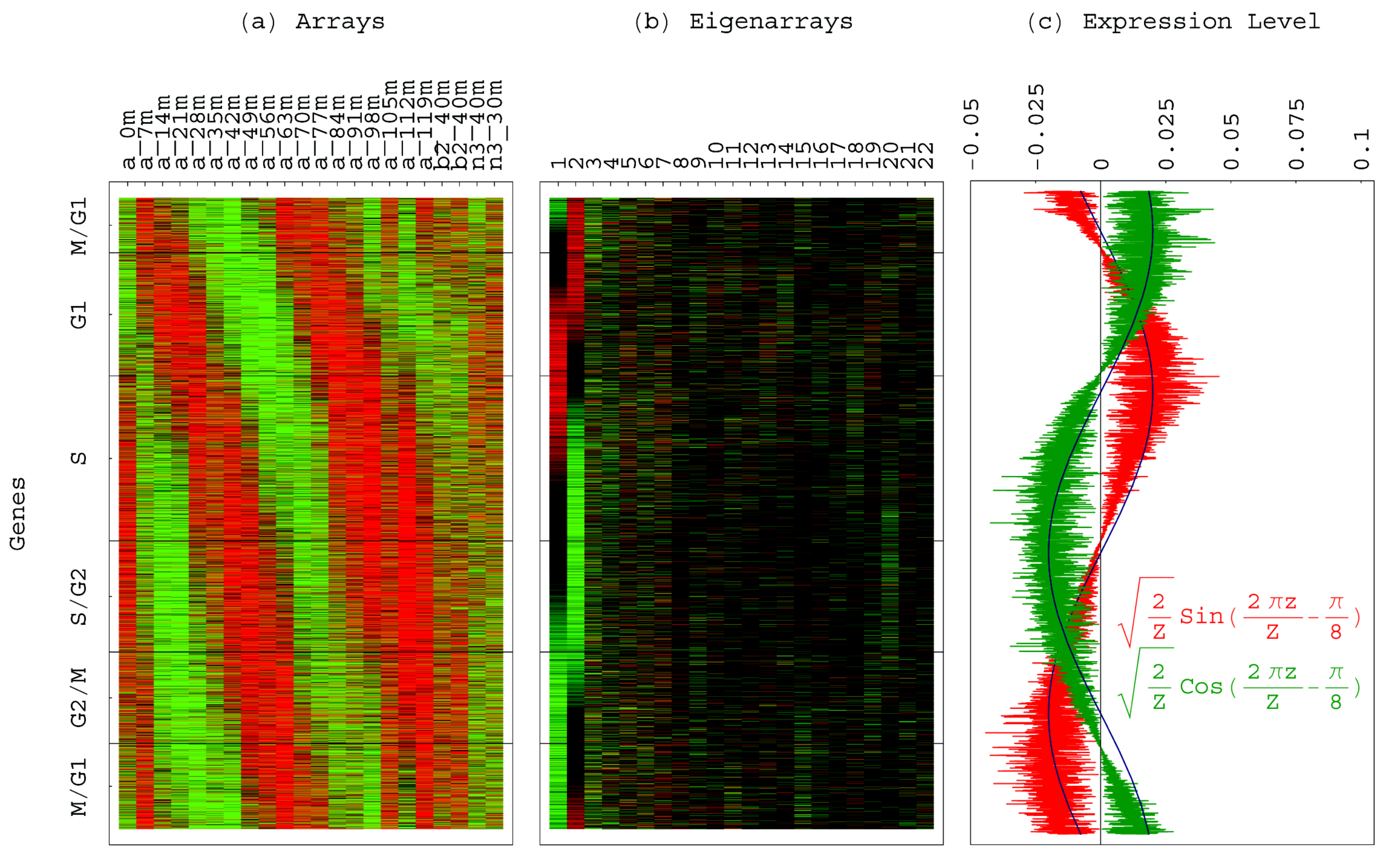


Alberts et al., *Molecular Biology of the Cell* (1994).

Math Operations → Biology

Traveling Wave of Expression

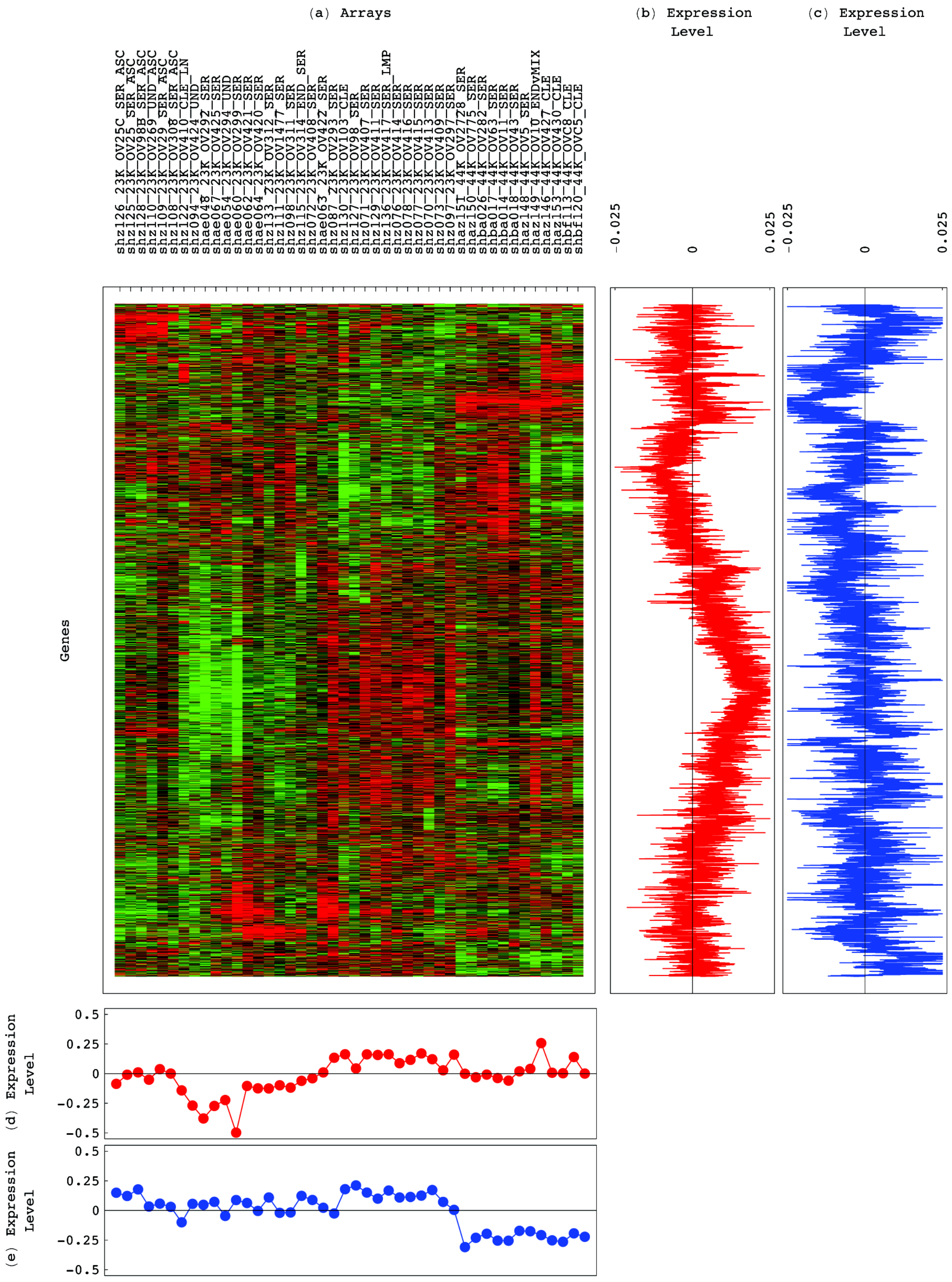
Cln3, Clb2 overactive samples = ± first eigenarray



Consistent model for the expression of almost the full yeast genome during cell cycle, in a subspace spanned by only two eigengenes and corresponding eigenarrays.

- → Are there only two cellular elements or modules that drive the yeast cell cycle?
- → Can we design a synthetic genetic network analogous to the analog harmonic oscillator, which would simulate the yeast cell cycle?

SVD Detection of Array-Production Artifacts in Ovarian Tumor Data

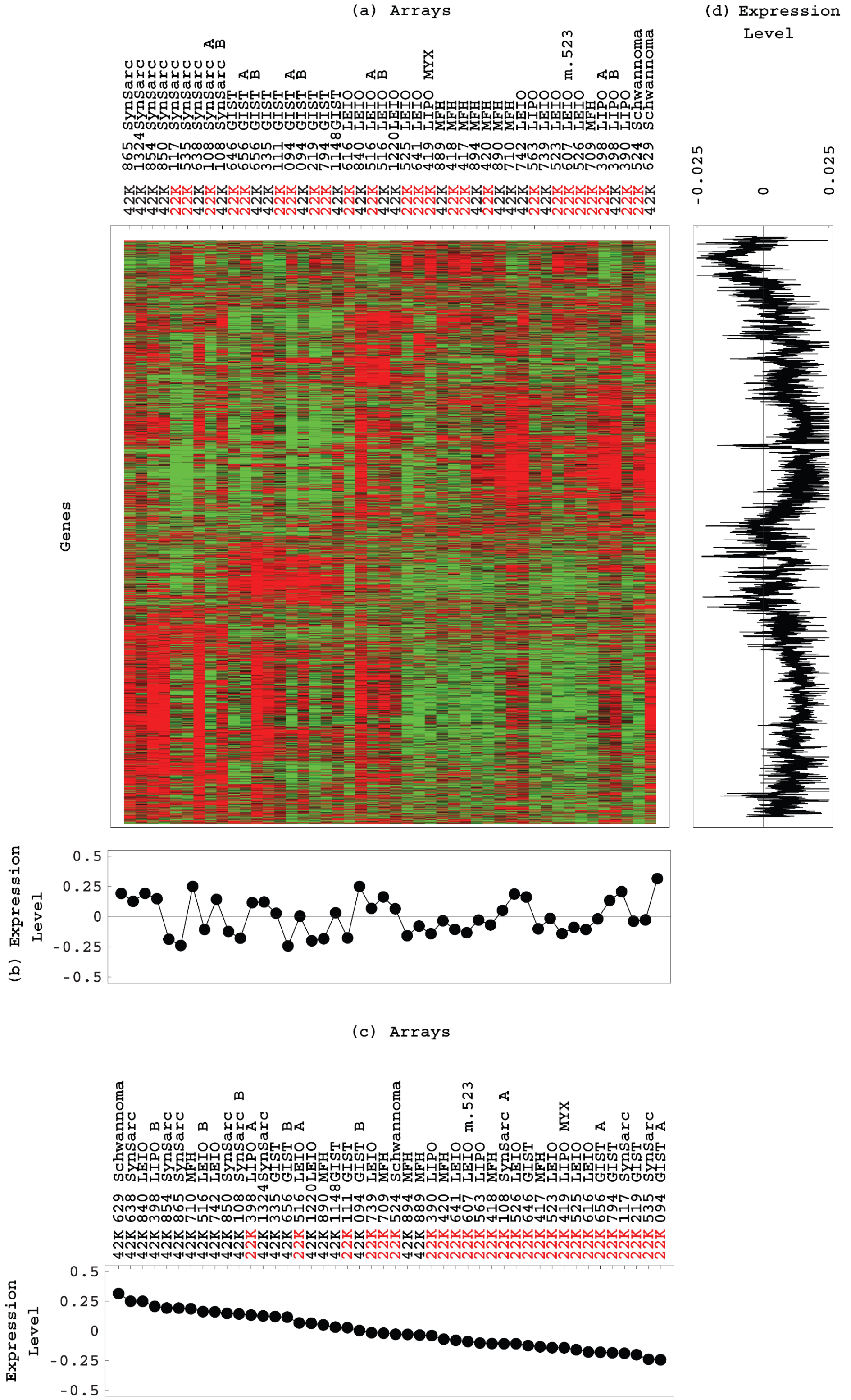


A: ↓ one array batch

B: ↑ 23K array prints ↓ 44K array prints

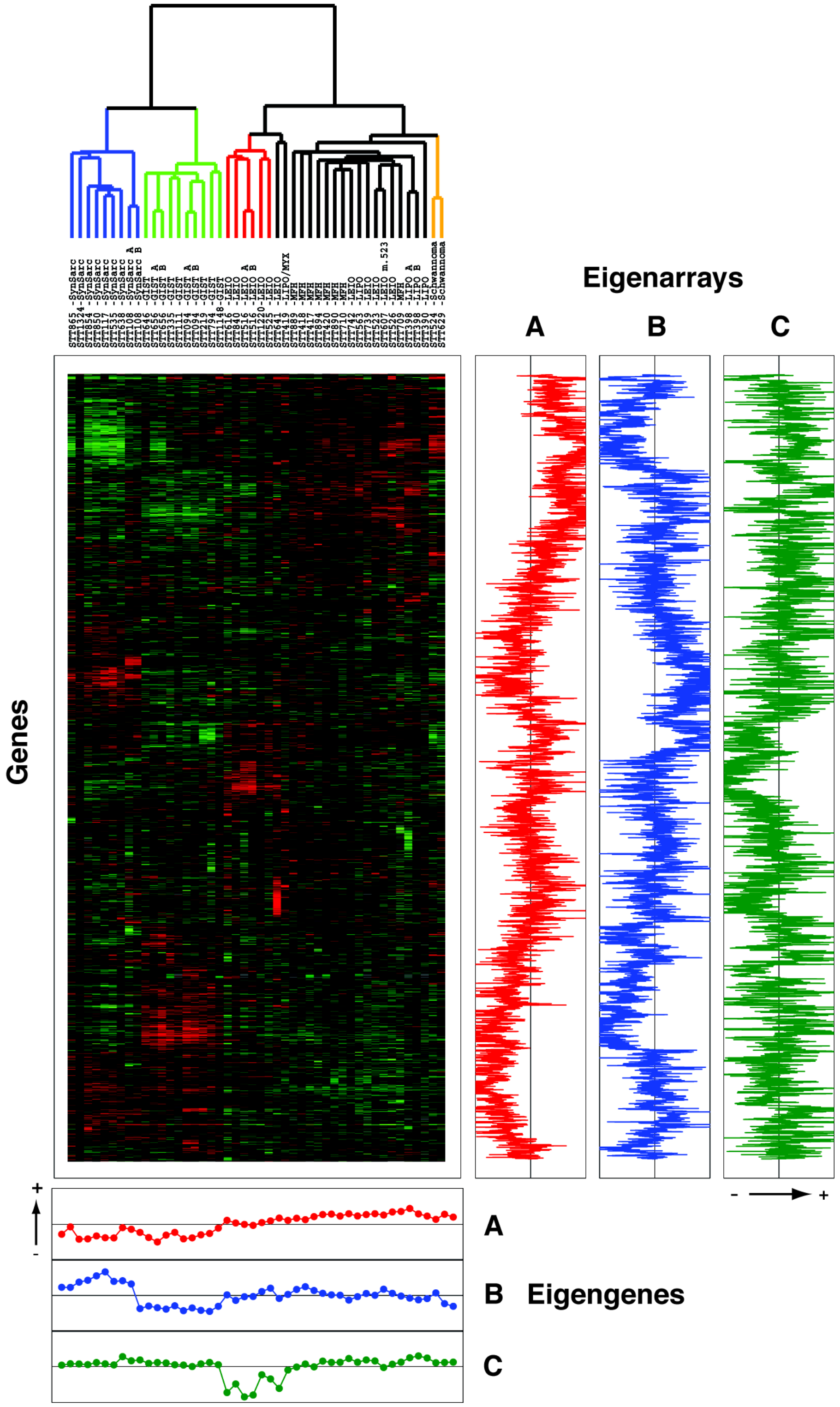
SVD Identification of Array-Print Artifact in Soft-Tissue Tumor Data

Nielsen, West, Linn, Alter et al., *Lancet* 359, 1301 (2002).



SVD Classification of Tumor Data

Nielsen, West, Linn, Alter et al., *Lancet* 359, 1301 (2002).

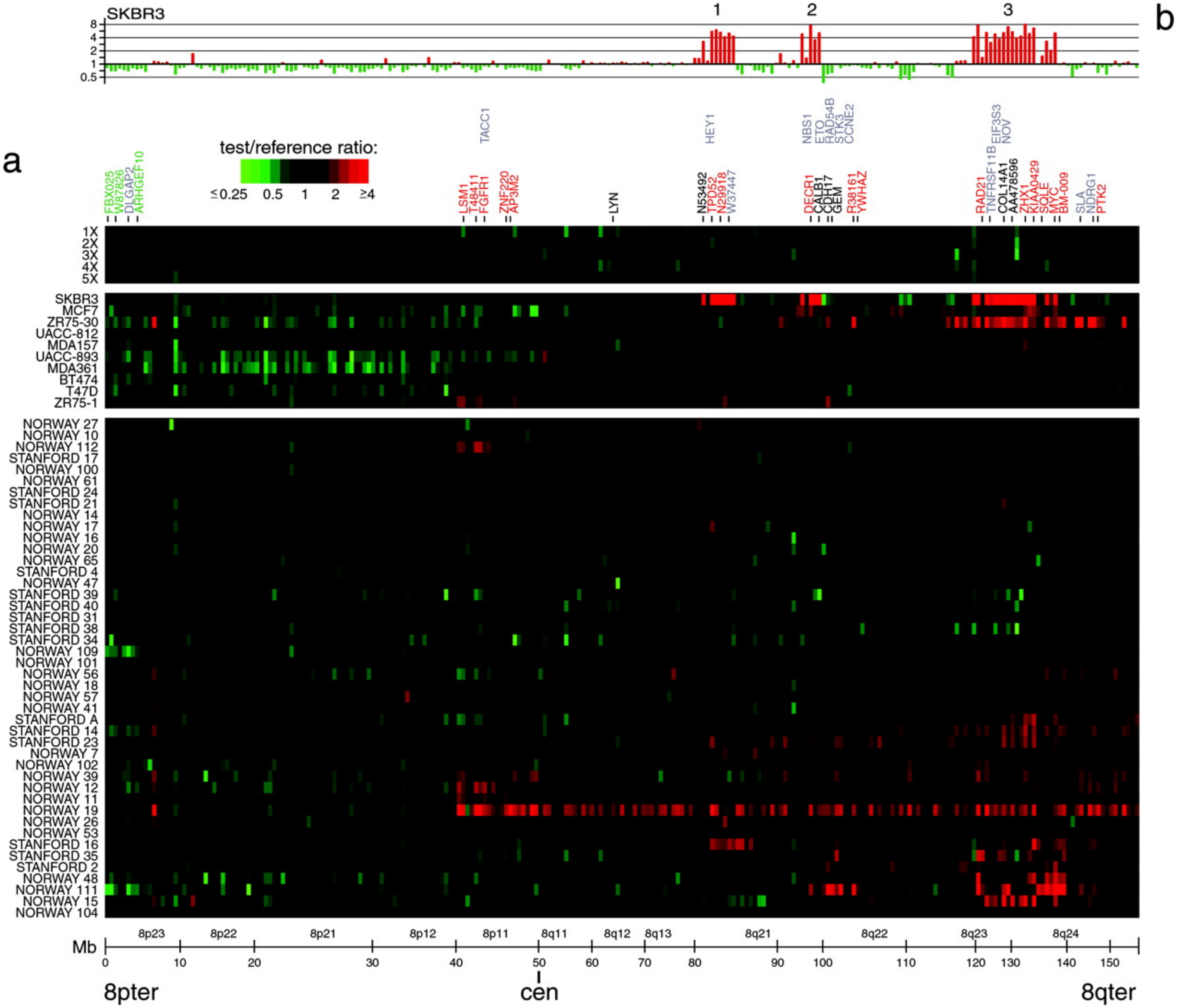


A: ↓ synovial sarcomas and gastrointestinal stromal tumors (GISTs)

B: ↑ synovial sarcomas ↓ GISTs

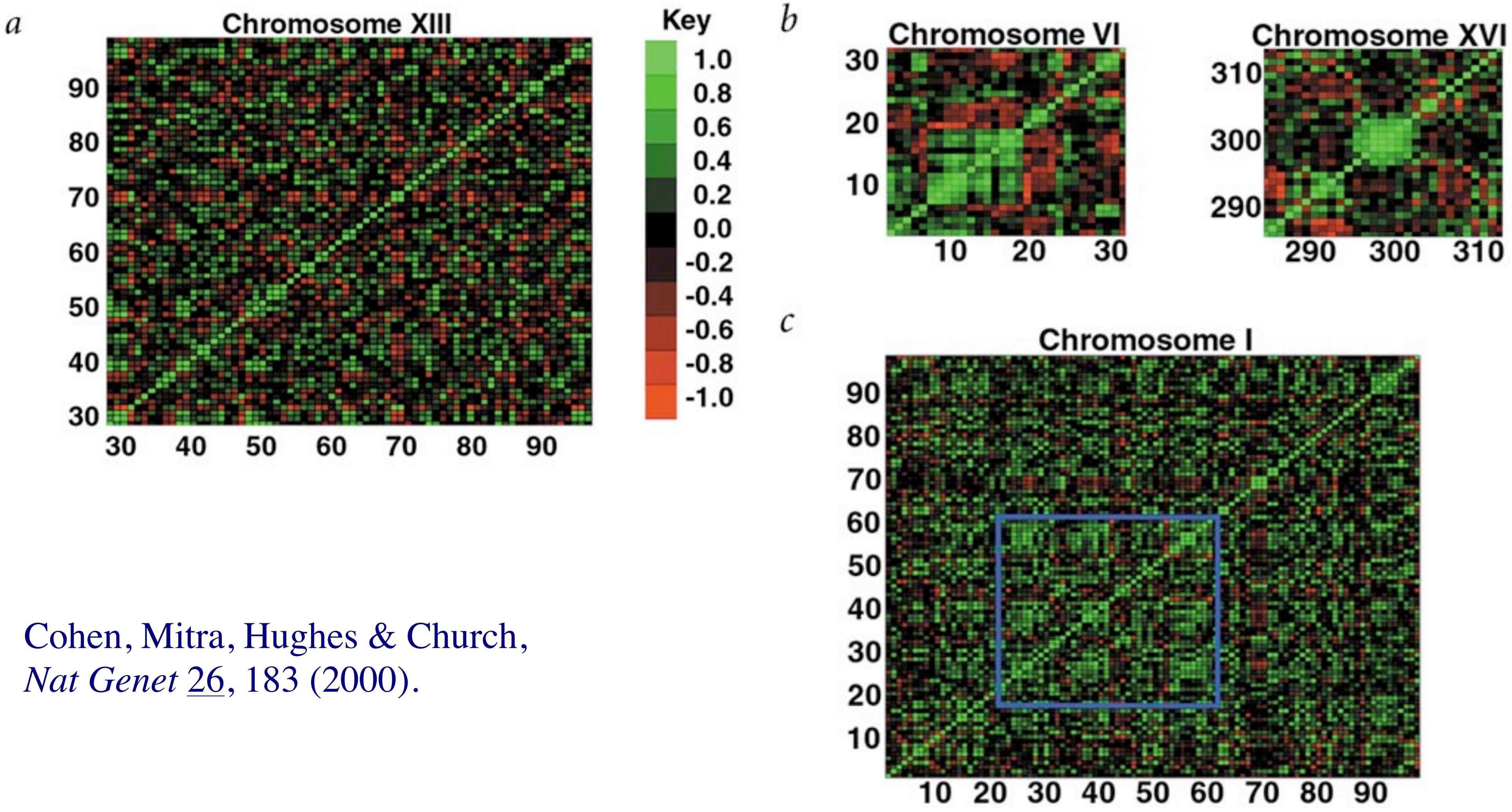
C: ↓ leiomyosarcomas that express a group of muscle genes

Organization (I): Chromosomal Order



Pollack et al., *PNAS* 99, 12963 (2002).

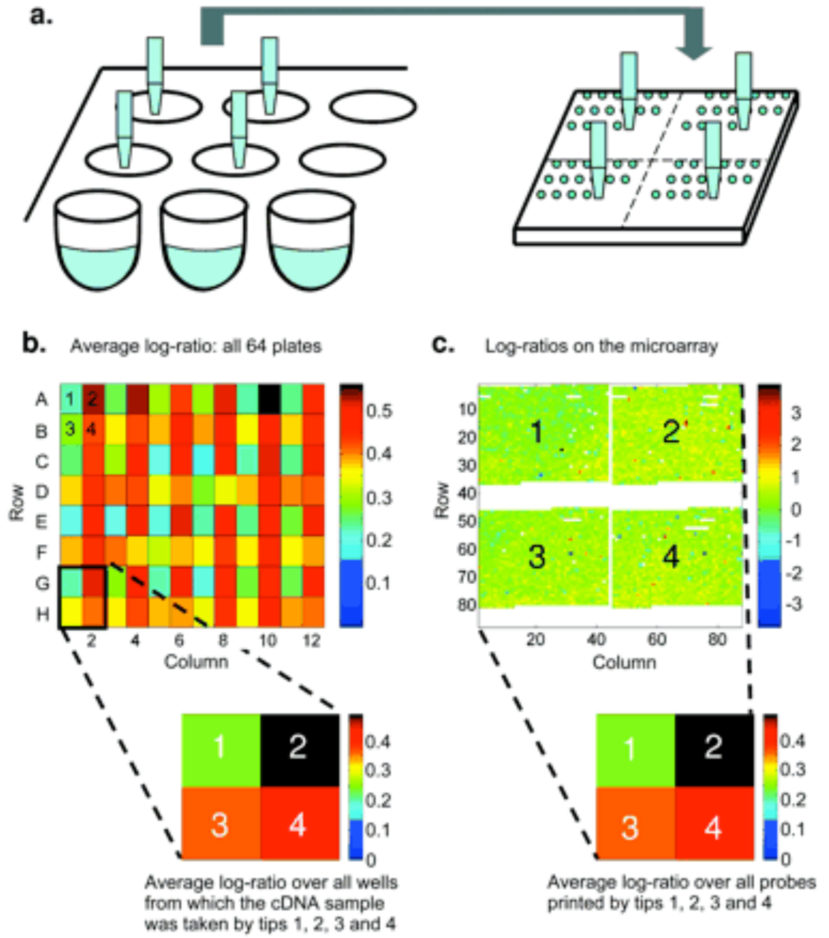
Computational Analysis of Whole-Genome Expression Data Reveals Chromosomal Domains of Gene Expression

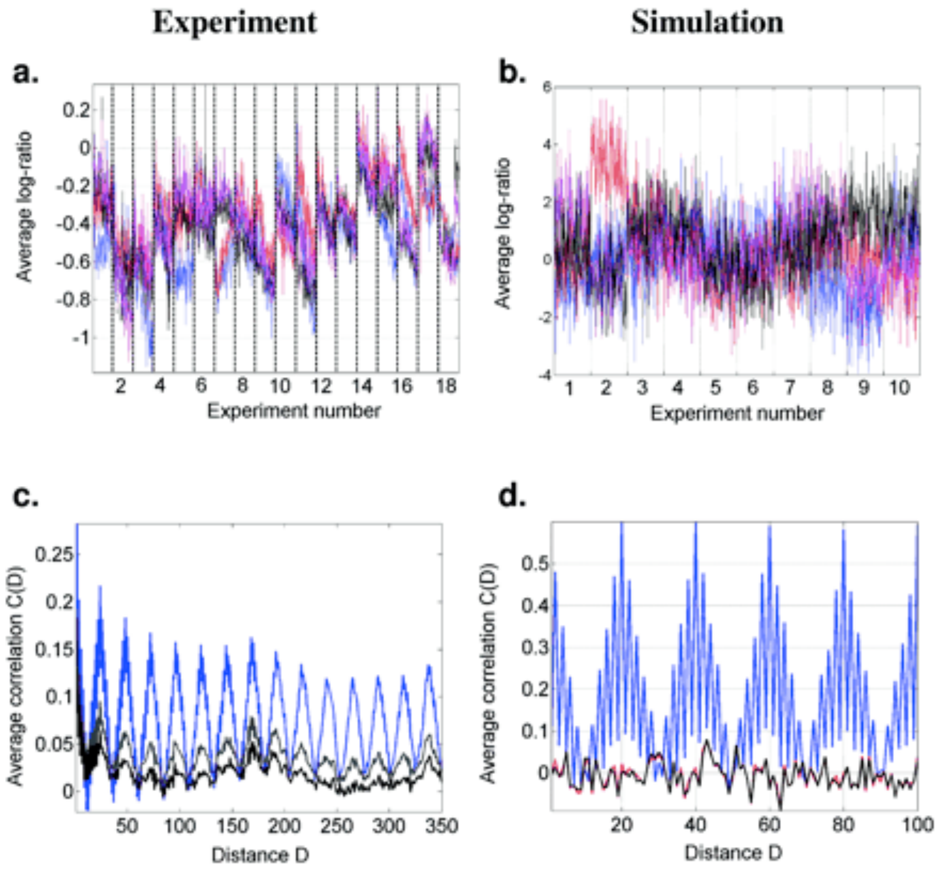


Cohen, Mitra, Hughes & Church,
Nat Genet 26, 183 (2000).

Spurious Spatial Periodicity of Co-Expression in Microarray Data Due to Printing Design

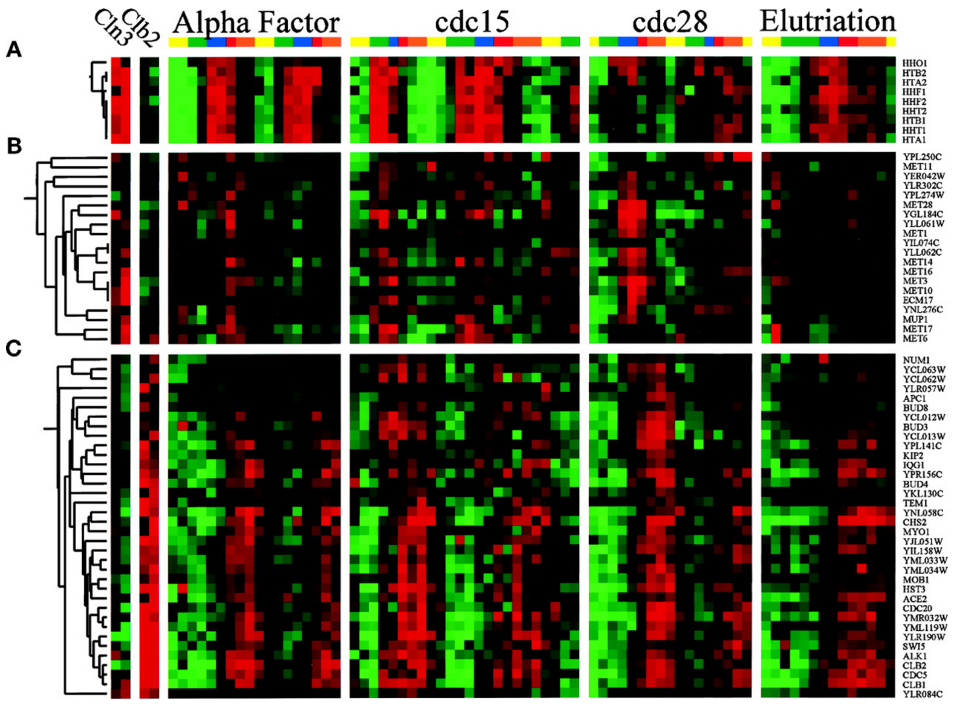
Chromosomal order might be correlated with array sector.

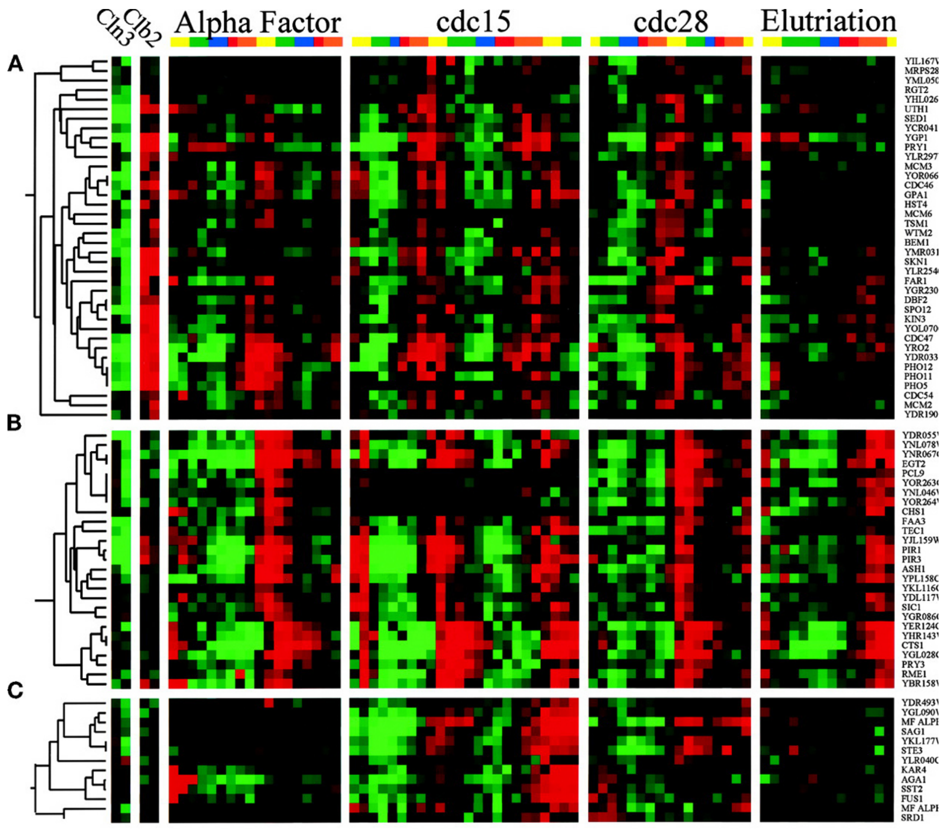




Balazsi, Kay, Barabasi & Oltvai, *Nucleic Acids Res* 31, 4425 (2003).

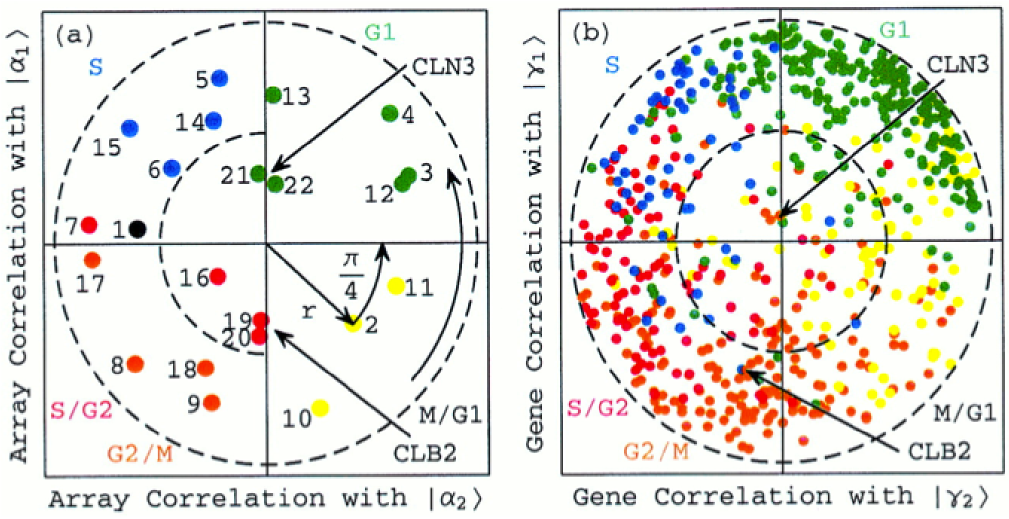
Organization (II): Hierarchical Clustering Groups Genes of Similar Function

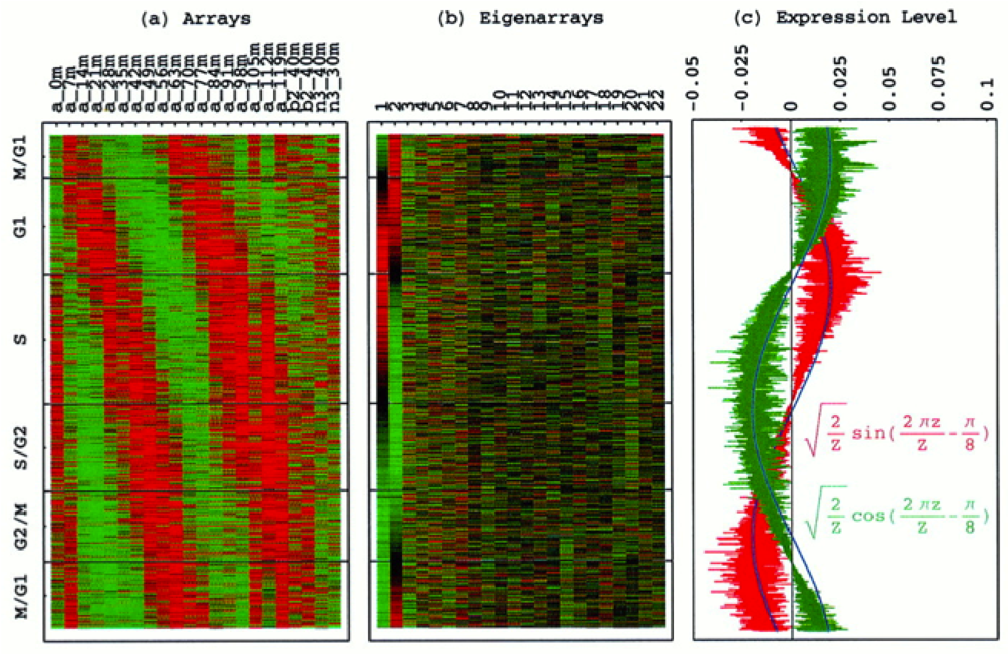




Spellman et al., *MBC* 9, 3273 (1998);
Eisen, Spellman, Brown & Botstein, *PNAS* 95, 14863 (1998).

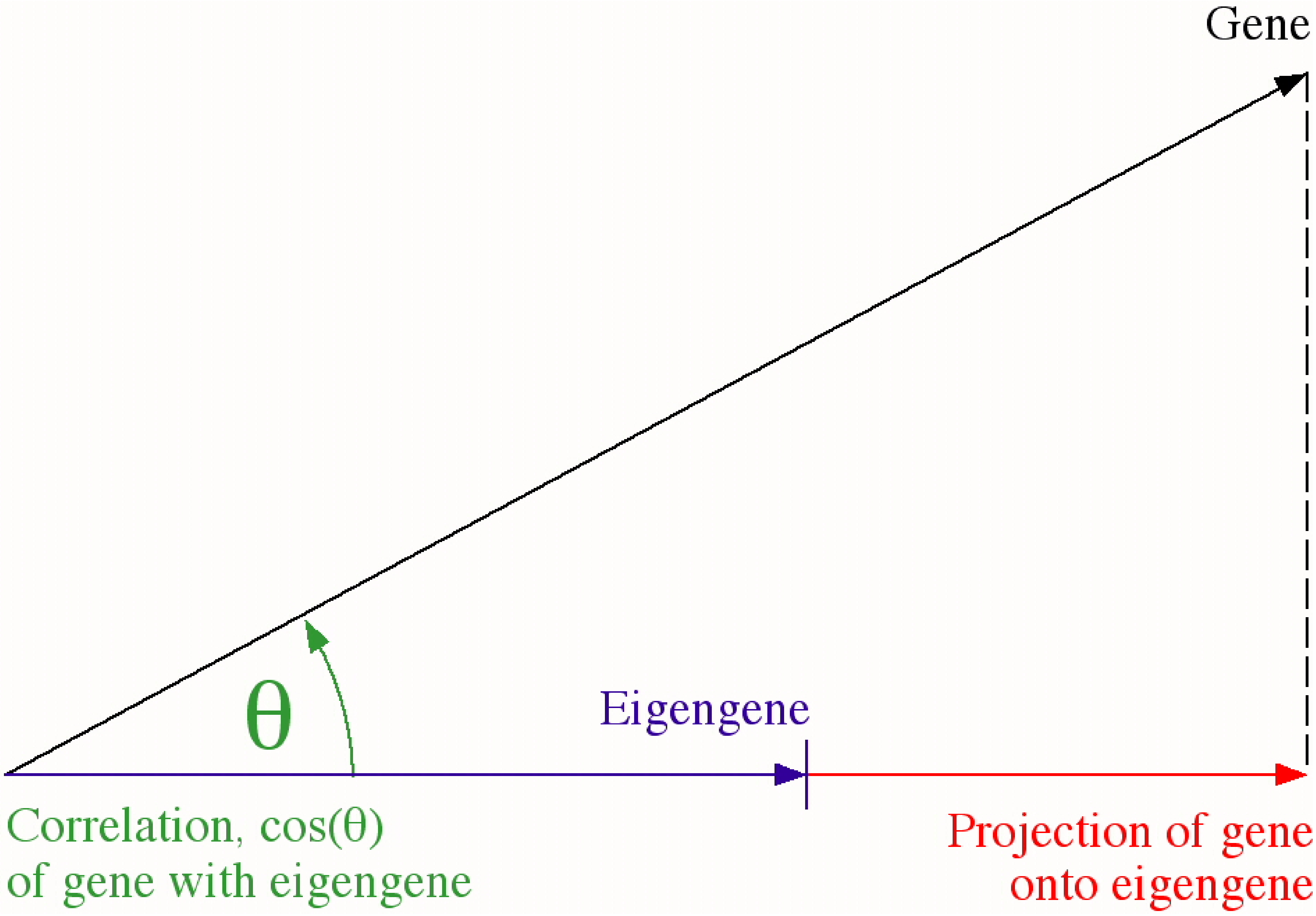
Organization (III): Sorting by Eigengenes and Eigenarrays





Alter, Brown & Botstein, *PNAS* 97, 10101 (2000).

Similarity by Projection and Correlation



Nielsen, West, Linn, Alter et al., *Lancet* 359, 1301 (2002).