Clustering

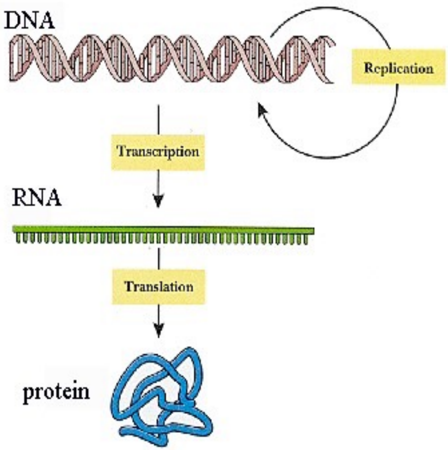
Introduction to omics data

Overview

- Recap of the molecules of life
- High-throughput datasets/omic datasets
- Transcriptomic data
- Computational tasks with transcriptomic data

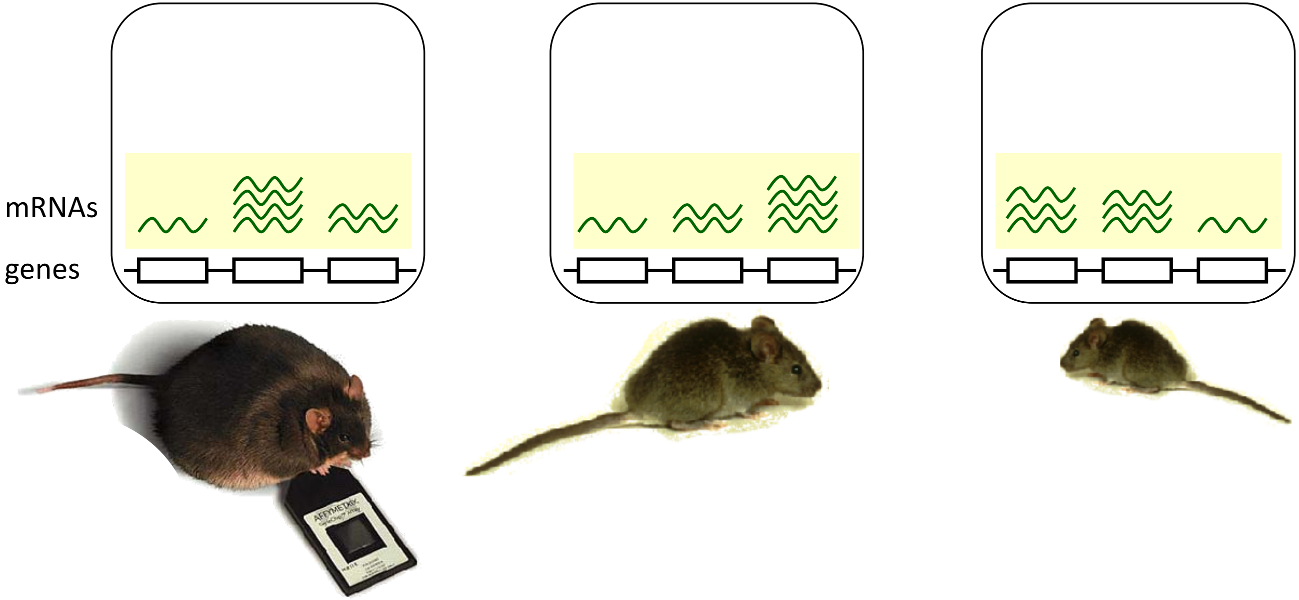
Molecules of life

- DNA
- RNA
  - mRNA
  - ncRNA
- Proteins
- Metabolites
- Whereas DNA is mostly static, RNA, proteins, metabolites *change* between cell types, tissues, environments and conditions



The central dogma

RNA levels are dynamic

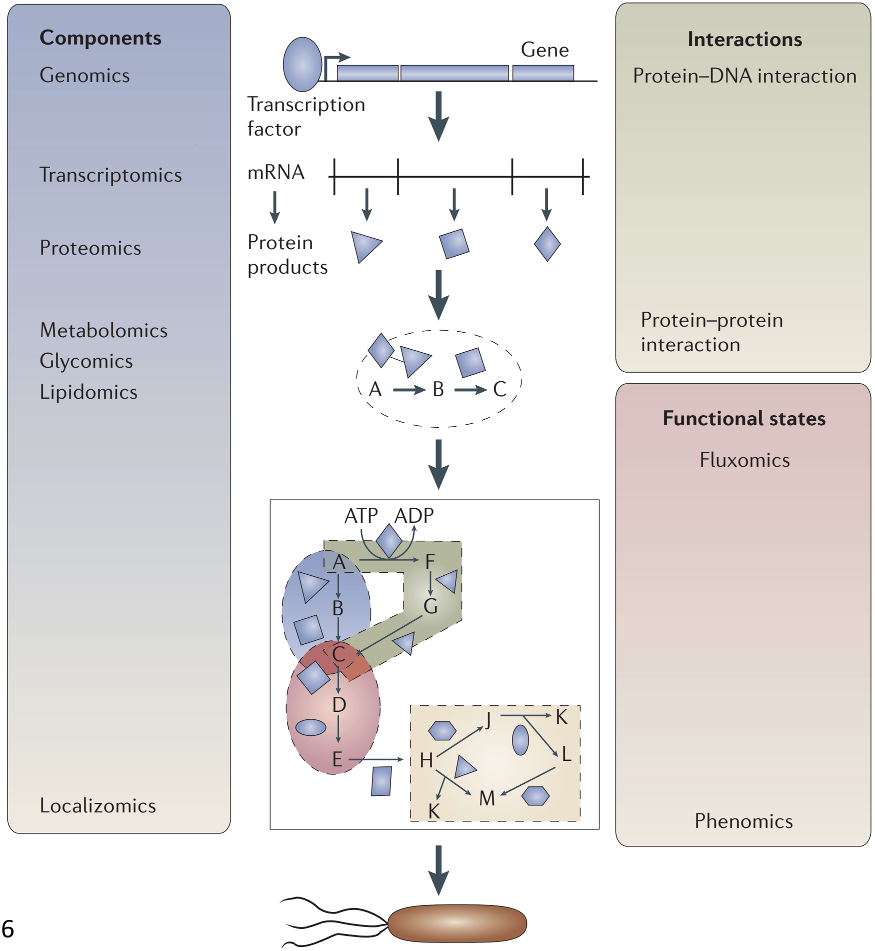


- What is varied: individuals, strains, cell types, environmental conditions, disease states, etc.
- What is measured: RNA quantities for thousands of genes, exons or other transcribed sequences

High-throughput datasets and “omes”

- Aim to measure as many components of a sample of cells simultaneously
- Types of omes
  - Genome: collection of DNA in a cell
  - Epigenome: all of the chemical modifications on the genome
  - Transcriptome: all of the RNA in cell
  - Proteome: all of the proteins in a cell
  - Metabolome: all of the metabolites present in a cell
  - Interactome: all of the interactions within a cell

Omics data provide comprehensive description of nearly all components of the cell



Components

Genomics

Transcriptomics

Proteomics

Metabolomics
Glycomics
Lipidomics

Localizomics

Interactions

Protein–DNA interaction

Protein–protein interaction

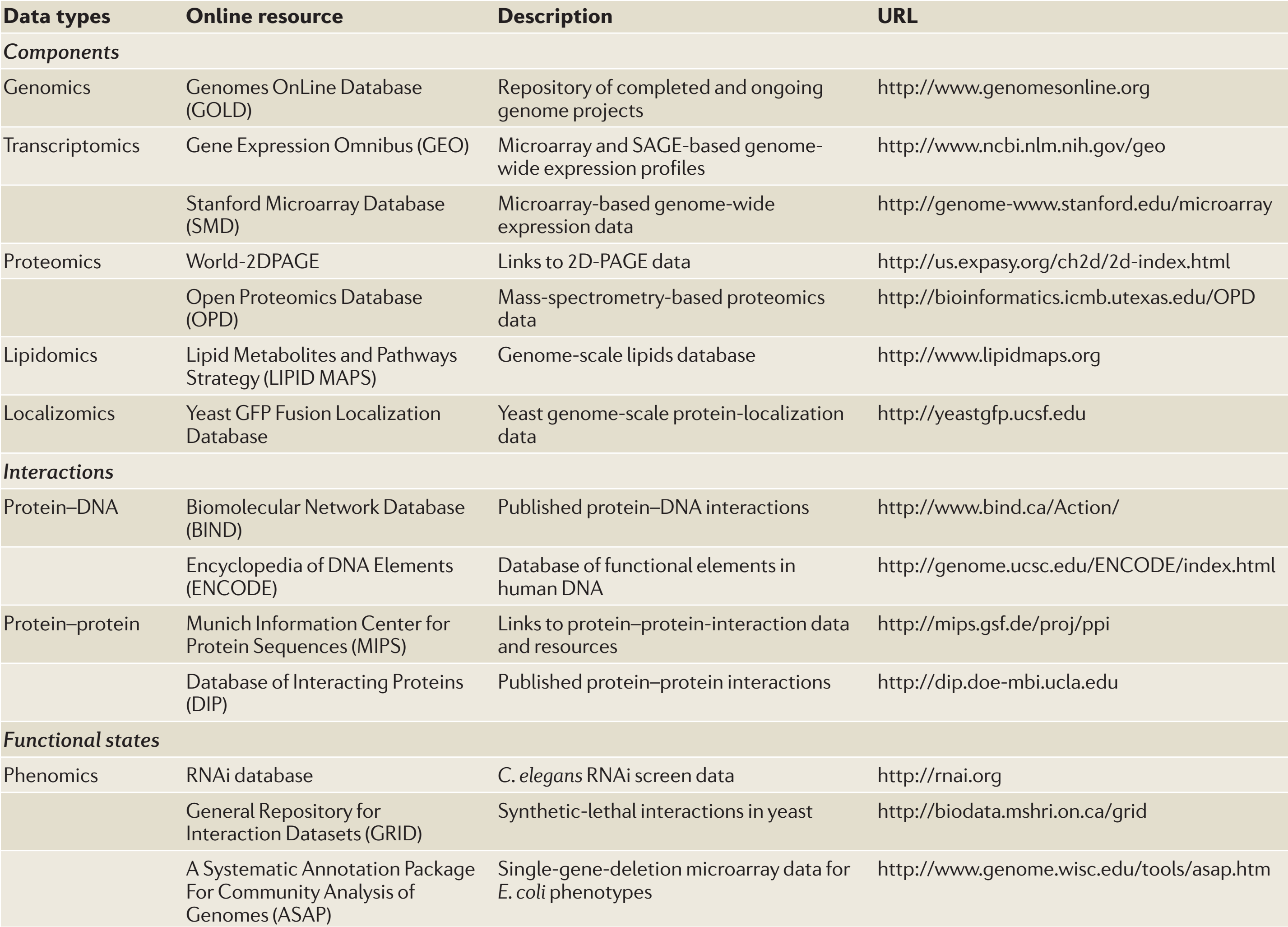
Functional states

Fluxomics

Phenomics

Joyce & Palsson, Nature Mol cell biol. 2006

Databases with omic data



Data types	Online resource	Description	URL
Components			
Genomics	Genomes OnLine Database (GOLD)	Repository of completed and ongoing genome projects	http://www.genomesonline.org
Transcriptomics	Gene Expression Omnibus (GEO)	Microarray and SAGE-based genome-wide expression profiles	http://www.ncbi.nlm.nih.gov/geo
	Stanford Microarray Database (SMD)	Microarray-based genome-wide expression data	http://genome-www.stanford.edu/microarray
Proteomics	World-2DPAGE	Links to 2D-PAGE data	http://us.expasy.org/ch2d/2d-index.html
	Open Proteomics Database (OPD)	Mass-spectrometry-based proteomics data	http://bioinformatics.icmb.utexas.edu/OPD
Lipidomics	Lipid Metabolites and Pathways Strategy (LIPID MAPS)	Genome-scale lipids database	http://www.lipidmaps.org
Localizomics	Yeast GFP Fusion Localization Database	Yeast genome-scale protein-localization data	http://yeastgfp.ucsf.edu
Interactions			
Protein–DNA	Biomolecular Network Database (BIND)	Published protein–DNA interactions	http://www.bind.ca/Action/
	Encyclopedia of DNA Elements (ENCODE)	Database of functional elements in human DNA	http://genome.ucsc.edu/ENCODE/index.html
Protein–protein	Munich Information Center for Protein Sequences (MIPS)	Links to protein–protein-interaction data and resources	http://mips.gsf.de/proj/ppi
	Database of Interacting Proteins (DIP)	Published protein–protein interactions	http://dip.doe-mbi.ucla.edu
Functional states			
Phenomics	RNAi database	*C. elegans* RNAi screen data	http://rnai.org
	General Repository for Interaction Datasets (GRID)	Synthetic-lethal interactions in yeast	http://biodata.mshri.on.ca/grid
	A Systematic Annotation Package For Community Analysis of Genomes (ASAP)	Single-gene-deletion microarray data for *E. coli* phenotypes	http://www.genome.wisc.edu/tools/asap.htm

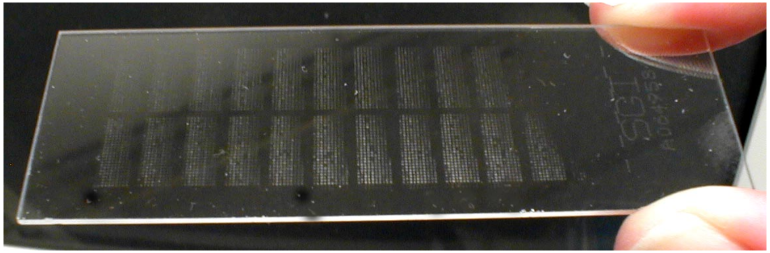
Joyce & Palsson, Nature Mol cell biol. 2006

Understand a cell as a system

- *Measure: identify the parts of a system*
  - Parts: different types of bio-molecules
    - genes, proteins, metabolites
  - High-throughput assays to measure these molecules
- *Model: how these parts are put together*
  - Clustering ✓
  - Network inference and analysis ✓

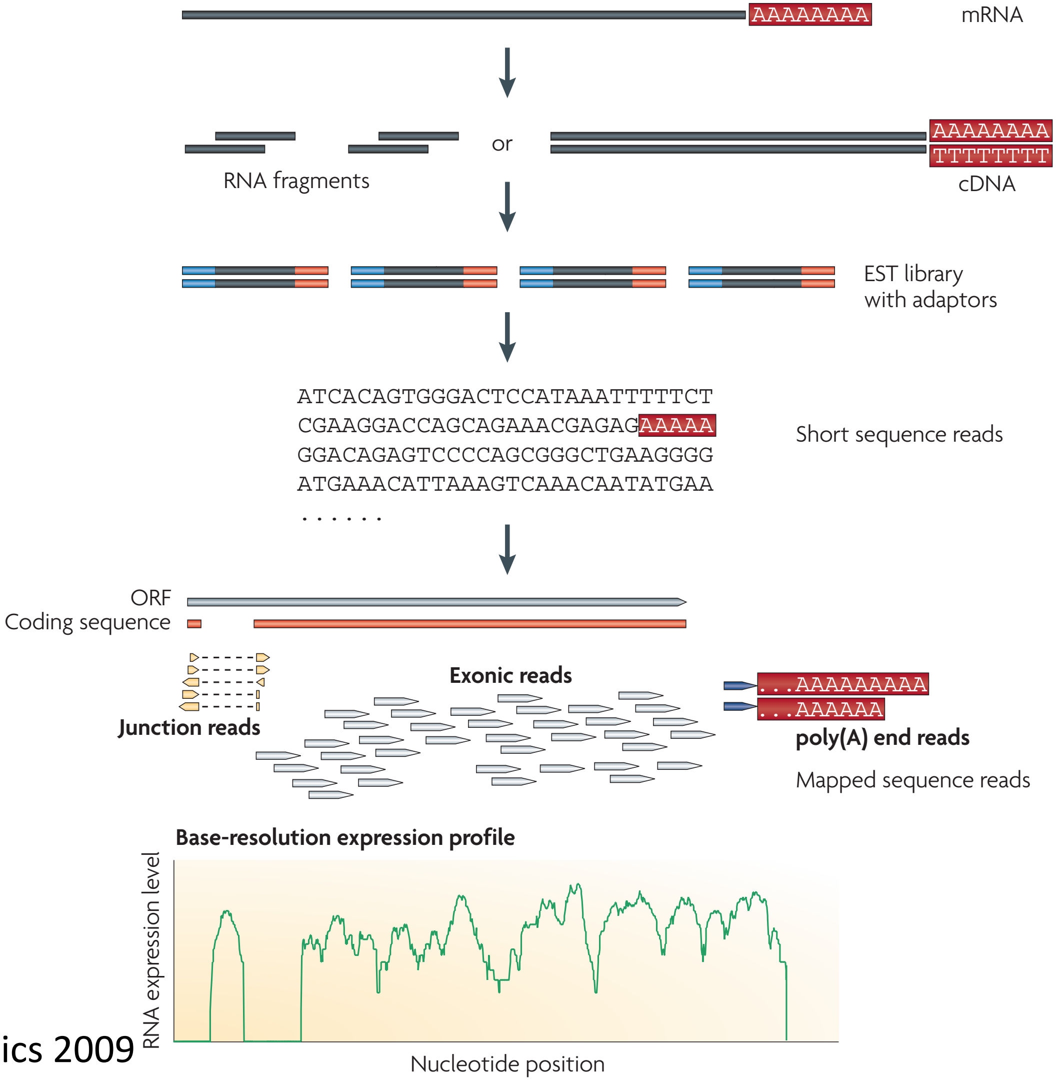
Bio-techniques to measure transcriptomes

- Microarrays



- Sequencing
  - RNA-seq

A typical RNA-seq pipeline



Wang et al, Nature Genetics 2009

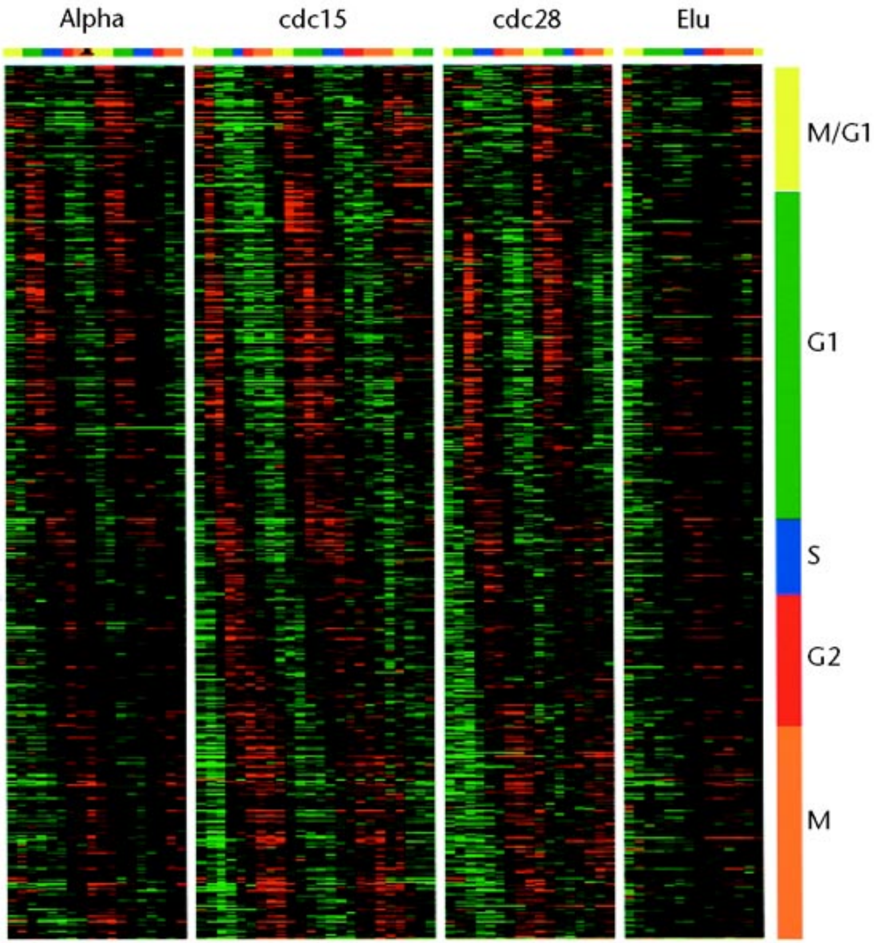
Gene expression profiles

- We will assume we have a 2D matrix of gene expression measurements
  - rows represent genes
  - columns represent different experiments, time points, individuals etc.

- We will refer to individual rows or columns as *profiles*
  - a row is a profile for a gene
  - a column is a profile for an experiment, time point, etc.

Gene-expression profiles for yeast cell cycle

- Rows represent yeast genes
- Columns represent time points as yeast goes through cell cycle
- Color represents expression level relative to baseline (red=high, green=low, black=baseline)

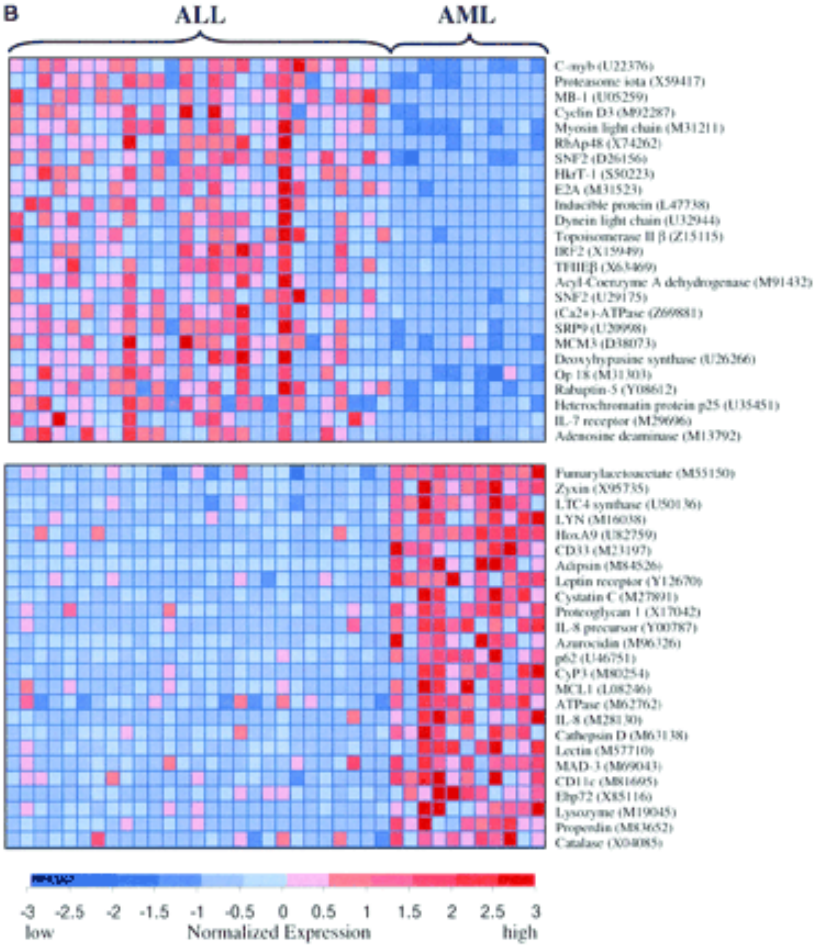


Spellman 1998

Gene-expression profiles for leukemia patients

- rows represent genes
- columns represent people with 2 subtypes of leukemia: ALL and AML

Each column corresponds to a microarray measurement



Commonly asked questions from expression datasets

- If we measure gene expression in a normal versus disease cell type, which genes have different expression levels across two groups?
  - Differential expression
- Which genes seem to be changing together?
  - Clustering genes based on expression profiles of genes across all conditions
- Which treatments/individuals have similar profiles?
  - Clustering samples based on gene expression profiles of all genes
- What does a gene do?
  - To which functional classes does a given gene belong
- What class is a sample from?
  - e.g., does this patient have ALL or AML