Molecular Biology 101

Data

Selected milestones in genome sequencing

	Year	Common Name	Species	# of Chromosomes	Size (base pairs)
	1995	Bacterium	Haemophilus influenzae	1	1.8×10^6
	1996	Yeast	Saccharomyces cerevisiae	16	1.2×10^7
	1998	Worm	Caenorhabditis elegans	6	1.0×10^8
	1999	Fruit Fly	Drosophila melanogaster	4	1.3×10^8
	2000	Human	Homo sapiens	23	3.1×10^9
	2002	Mouse	Mus musculus	20	2.6×10^9
	2004	Rat	Rattus norvegicus	21	2.8×10^9
	2005	Chimpanzee	Pan troglodytes	24	3.1×10^9

Sequence is freely available

NCBI - http://www.ncbi.nlm.nih.gov

UCSC - http://genome.ucsc.edu

But Wait, There's More…

- > 1000 other publicly available databases pertaining to molecular biology
- GenBank

  > 209 million sequence entries

  > 253 billion bases
- UniProtKB / Swis-Prot

  > 120 million protein sequence entries

  > 40 billion amino acids
- Protein Data Bank

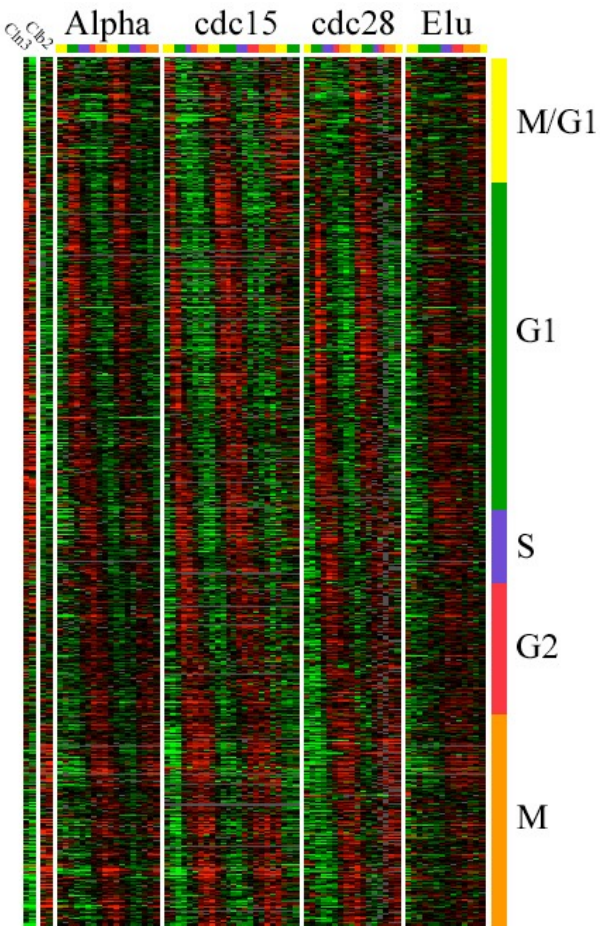
  144,042 protein (and related) structures

* all numbers current as of 9/18

More Data: High-Throughput Experiments

- RNA abundances
- protein abundances
- small molecule abundances
- protein-protein interactions
- protein-DNA interactions
- protein-small molecule interactions
- genetic variants of an individual (e.g. which DNA base does the individual have at a few million selected positions)
- something (e.g. viral replication) measured across thousands of genetic variants
- etc.

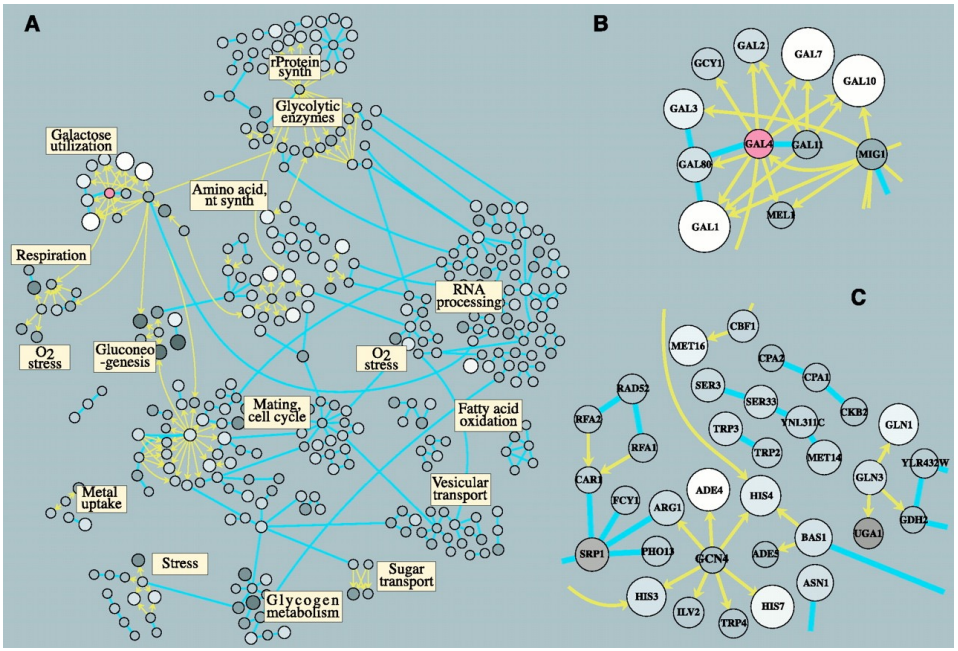
Example HT Experiment



- this figure depicts one yeast gene-expression data set
  - each row represents a gene
  - each column represents a measurement of gene expression (mRNA abundance) at some time point
  - red indicates that a gene is being expressed more than some baseline; green means less

Figure from Spellman et al., Molecular Biology of the Cell, 9:3273-3297, 1998

More Data: Interactions



- each node represents a gene product (protein)
- blue edges show direct protein-protein interactions
- yellow edges show interactions in which one protein binds to DNA and affects the expression of another

Figure from Ideker et al., Science 292(5518):929-934, 2001

Significance of the Genomics Revolution

- data driven biology
  - functional genomics
  - comparative genomics
  - systems biology
- molecular medicine
  - identification of genetic components of various maladies
  - diagnosis/prognosis from sequence/expression
  - gene therapy
- pharmacogenomics
  - developing highly targeted drugs
- toxicogenomics
  - elucidating which genes are affected by various chemicals