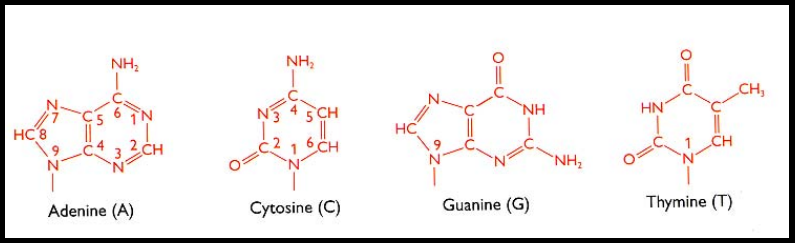
Genomic Medicine: Basic Molecular Biology

Atul Butte, MD
atul_butte@harvard.edu
Children's Hospital Informatics Program
www.chip.org

Children's Hospital • Boston
Harvard Medical School
Massachusetts Institute of Technology

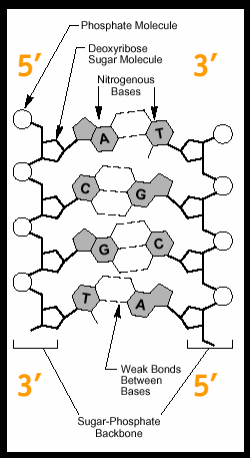
Basic Biology

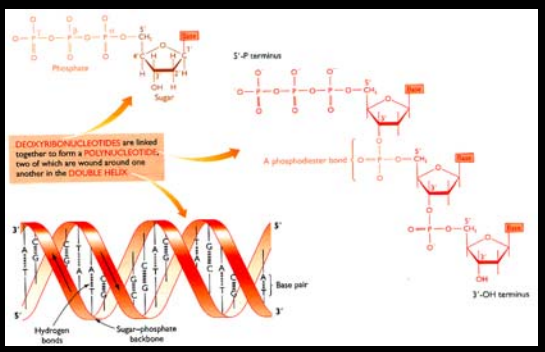
- Organisms need to produce proteins for a variety of functions over a lifetime
  - Enzymes to catalyze reactions
  - Structural support
  - Hormone to signal other parts of the organism
- Problem one: how to encode the instructions for making a specific protein
- Step one: nucleotides



Basic Biology

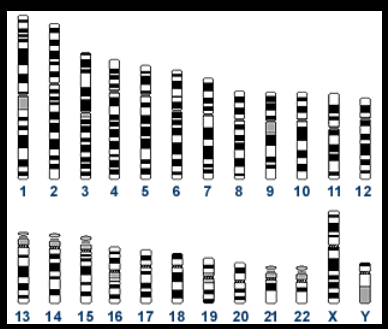
- Complementary nucleotides form base pairs
- Base pairs are put together in chains (strands)
  - Naturally form double helixes
  - Redundant information in each strand

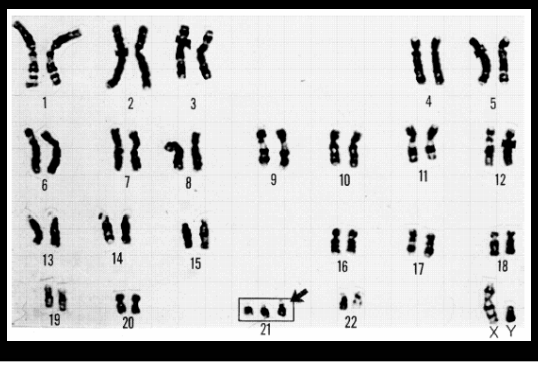




Chromosomes

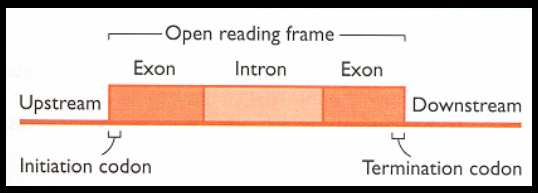
- We do not know exactly how strands of DNA wind up to make a chromosome
- Each chromosome has a single double-strand of DNA
- 22 human chromosomes are paired
- In human females, there are two X chromosomes
- In males, one X and one Y





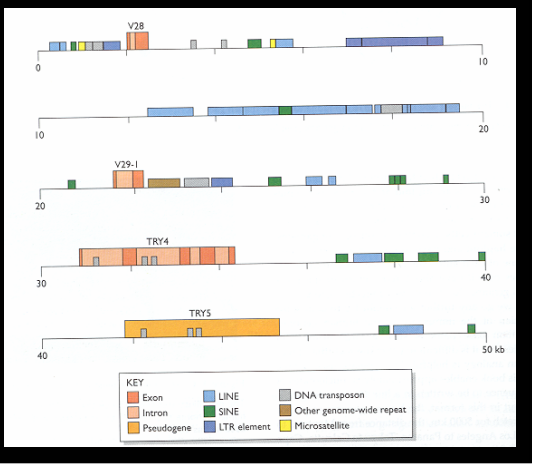
What does a gene look like?

- Each gene encodes instructions to make a single protein
- DNA before a gene is called upstream, and can contain regulatory elements
- Introns may be within the code for the protein
- There is a code for the start and end of the protein coding portion
- Theoretically, the biological system can determine promoter regions and intron-exon boundaries using the sequence syntax alone



Area between genes

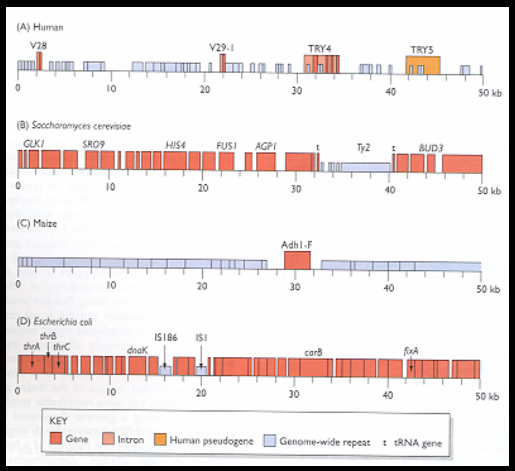
- The human genome contains 3 billion base pairs (3000 Mb) but only 35 thousand genes
- The coding region is 90 Mb (only 3% of the genome)
- Over 50% of the genome is repeated sequences
  - Long interspersed nuclear elements
  - Short interspersed nuclear elements
  - Long terminal repeats
  - Microsatellites
- Many repeated sequences are different between individuals



Genome size

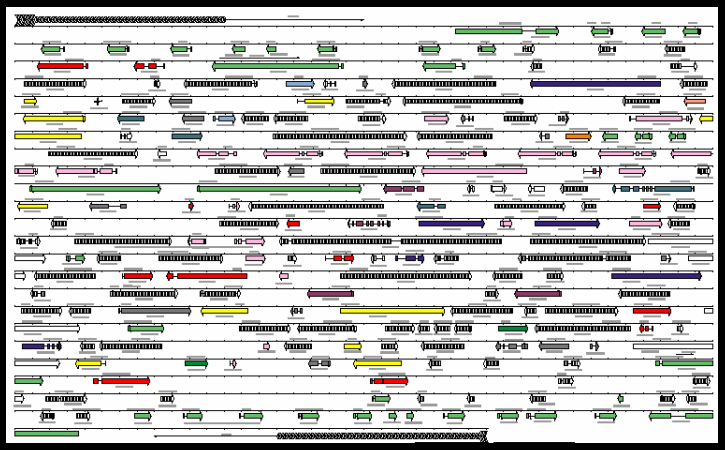
- We're the smartest, so we must have the largest genome, right?
- Not quite
- Our genome contains 3000 Mb (~750 megabytes)

- E. coli has 4 Mb
- Yeast has 12 Mb
- Pea has 4800 Mb
- Maize has 5000 Mb
- Wheat has 17000 Mb



Genomes of other organisms

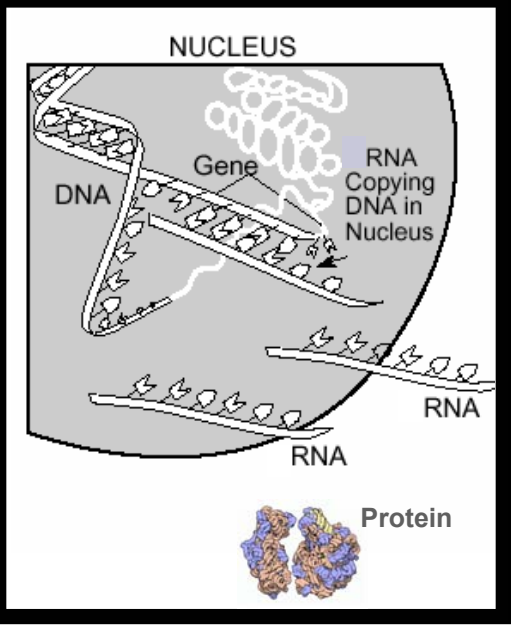
- *Plasmodium falciparum* chromosome 2



Gardner M, et al. Science; 282: 1126 (1998).

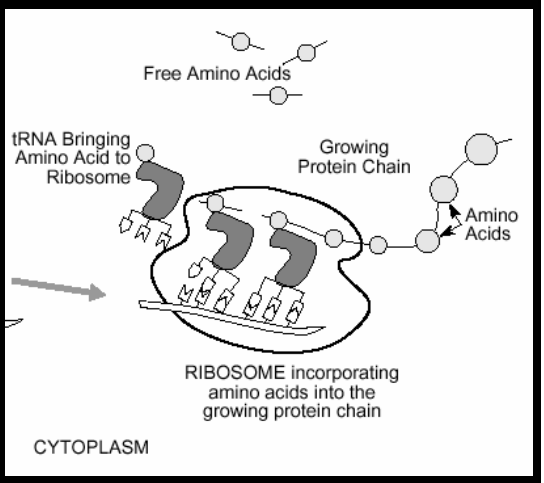
mRNA is made from DNA

- Genes encode instructions to make proteins
- The design of a protein needs to be duplicable
- mRNA is transcribed from DNA within the nucleus
- mRNA moves to the cytoplasm, where the protein is formed

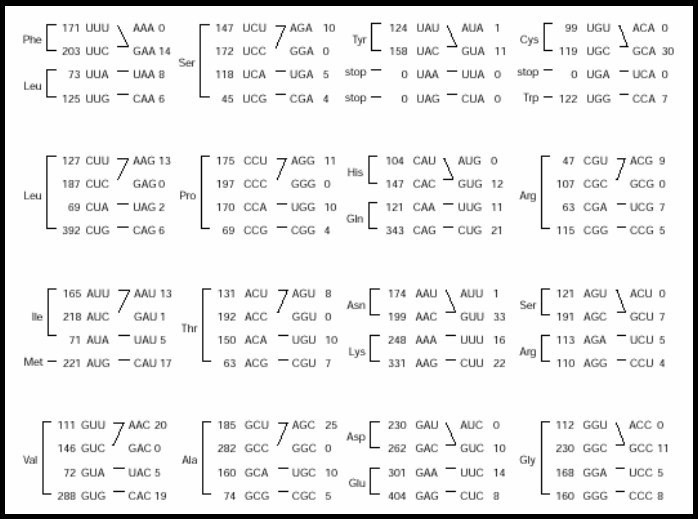


Digitizing amino acid codes

- Proteins are made of 20 (21) amino acids
- Yet each position can only be one of 4 nucleotides
- Nature evolved into using 3 nucleotides to encode a single amino acid
- A chain of amino acids is made from mRNA



Genetic Code



Nature; 409: 860 (2001).

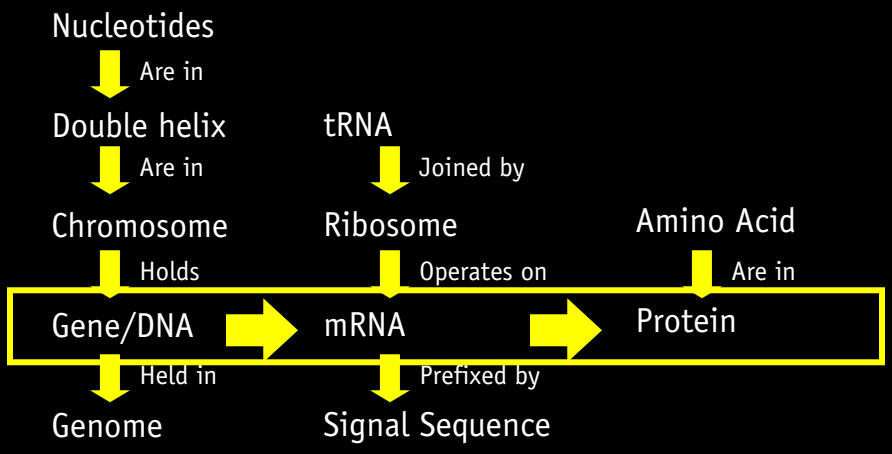
Molecular Biology

Nucleotides
↓ Are in
Double helix
↓ Are in
Chromosome
↓ Holds
Gene/DNA
↓ Held in
Genome

tRNA
↓ Joined by
Ribosome
↓ Operates on
mRNA
↓ Prefixed by
Signal Sequence

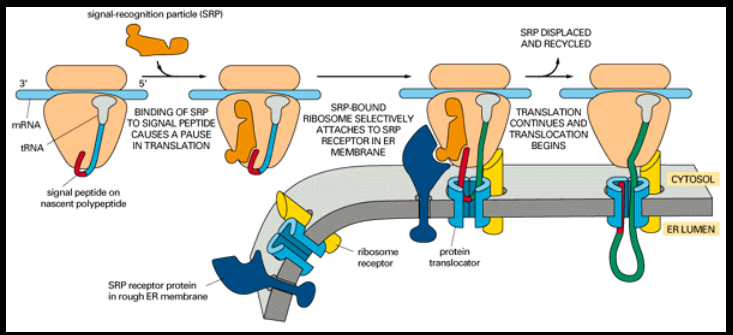
Amino Acid
↓ Are in
Protein

Central Dogma



Protein targeting

- The first few amino acids may serve as a signal peptide
- Works in conjunction with other cellular machinery to direct protein to the right place



Transcriptional Regulation

- Amount of protein is roughly governed by RNA level
- Transcription into RNA can be activated or repressed by transcription factors

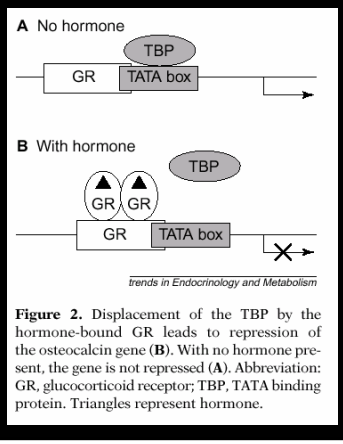


Figure 2. Displacement of the TBP by the hormone-bound GR leads to repression of the osteocalcin gene (**B**). With no hormone present, the gene is not repressed (**A**). Abbreviation: GR, glucocorticoid receptor; TBP, TATA binding protein. Triangles represent hormone.

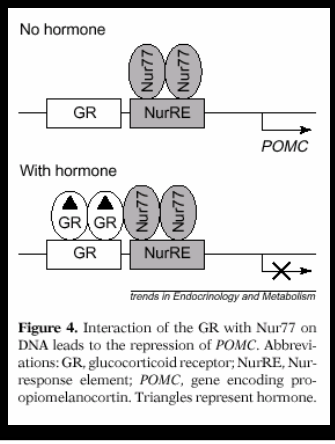
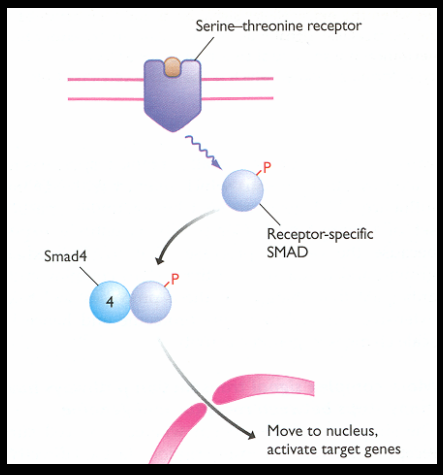


Figure 4. Interaction of the GR with Nur77 on DNA leads to the repression of *POMC*. Abbreviations: GR, glucocorticoid receptor; NurRE, Nur-response element; *POMC*, gene encoding pro-opiomelanocortin. Triangles represent hormone.

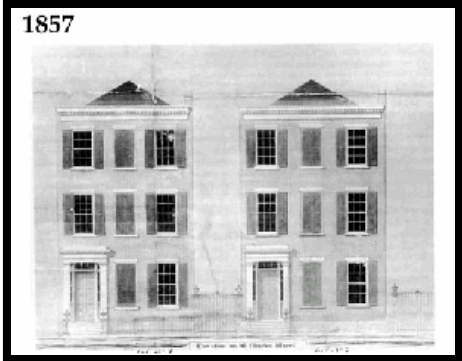
What starts the process?

- Transcriptional programs can start from
  - Hormone action on receptors
  - Shock or stress to the cell
  - New source of, or lack of nutrients
  - Internal derangement of cell or genome
  - Many, many other internal and external stimuli



Temporal Programs

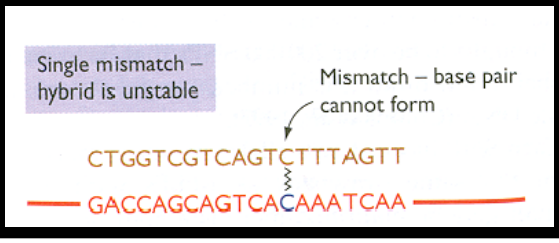
- Segmentation versus Homeosis: same two houses at different times



Scott M. Cell; 100: 27 (2000).

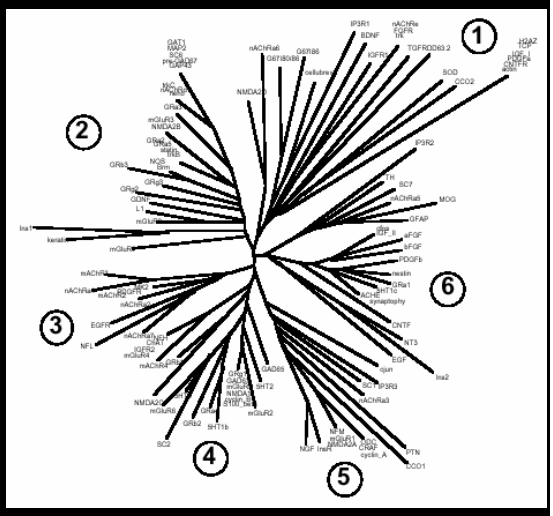
mRNA

- mRNA can be transcribed at up to several hundred nucleotides per minute
- Some eukaryotic genes can take many hours to transcribe
  - Dystrophin takes 20 hours to transcribe
- Most mRNA ends with poly-A, so it is easy to pick out
- Can look for the presence of specific mRNA using the complementary sequence

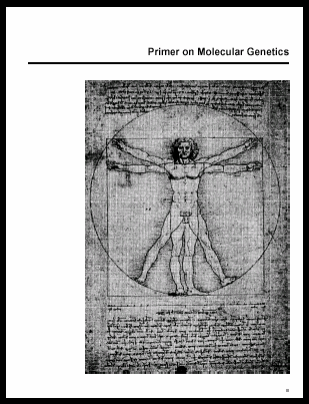


Periodic Table for Biology

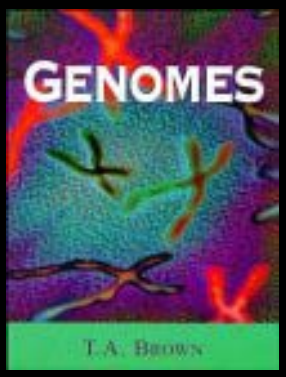
- Knowing all the genes is the equivalent of knowing the periodic table of the elements
- Instead of a table, our periodic table may read like a tree



More Information



- Department of Energy Primer on Molecular Genetics http://www.ornl.gov/hgmis/publicat/primer/primer.pdf



- T. A. Brown, Genomes, John Wiley and Sons, 1999.

Gene Measurement Techniques

DNA

- Sequencing
- Polymorphisms

RNA

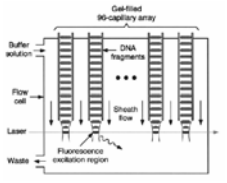
- Serial analysis of gene expression
- DNA Microarrays
- Wafers

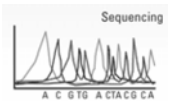
Protein

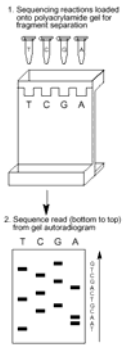
- 2D-PAGE
- Mass spectrometry
- Protein arrays

Sequencing Reactions

- Sanger Reactions
- Four color fluorescence-base sequence detection
- Laser detector
- Automated process

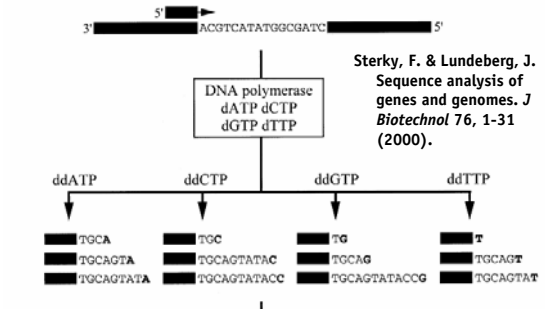






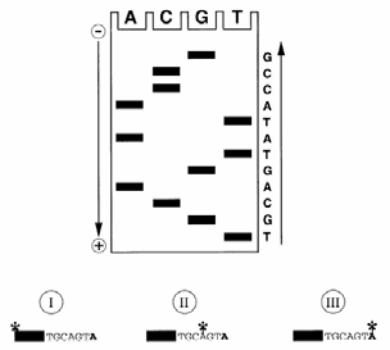
Jaklevic JM, et al. Annu Rev Biomed Eng 1:649 (1999).

Sanger Chain Termination



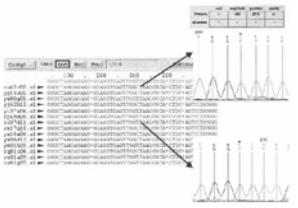
Sterky, F. & Lundeberg, J. Sequence analysis of genes and genomes. *J Biotechnol* 76, 1-31 (2000).

Sanger Method



Sequencing Reactions

- PHRED: base-quality score for each base, based on probability of erroneous call
- PHRED quality score of X means error probability of $10^{-x/10}$
- PHRED score of 30 means 99.9% accuracy for base call

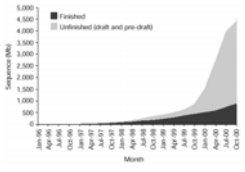


Buetow KH, et al. Nature Genetics 21:323 (1999).

Sequencing Reactions

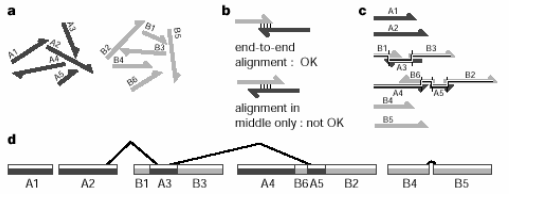
- PHRAP: assembles sequence data using base-quality scores into sequence contigs
- Assembly-quality scores

- Most of the genome was sequenced over 12 months
- Highest throughput center at Whitehead: 100,000 sequencing reactions per 12 hours
- Robots pick 100,000 colonies, sequence 60 million nucleotides per day



Assembly

- Contamination from non-human sequences removed
- Clones overlaid on physical map
- High-quality semiautomatic sequencing from both ends of very large numbers of numbers of human genome fragments
- Overlaps take memory: Drosophila 600 GB RAM
- Human 10 4-processor 4 GB and 16-processor 64 GB, 10K CPU hrs



Genome Browsers

- Genome browsers: University of California at Santa Cruz and EnsEMBL
- Overlap sequence, cytogenetic, SNP, genetic maps
- Overlap annotations, disease genes

Figure 10 Screen shot from UCSC Draft Human Genome Browser. See http://genome.ucsc.edu/.

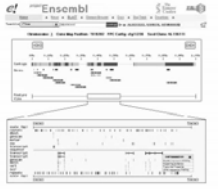
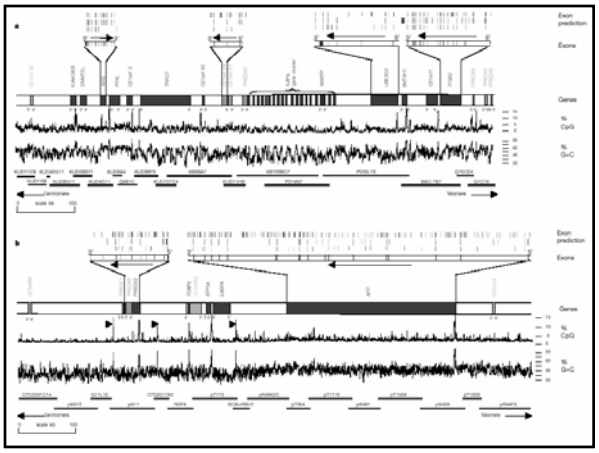
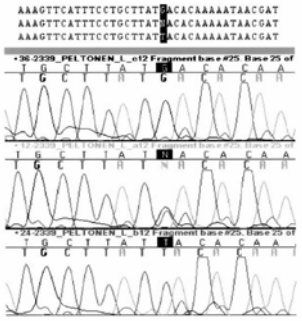


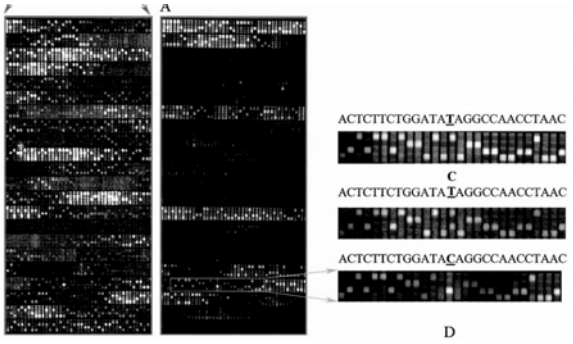
Figure 11 Screen shot from the Genome Browser of Project Ensembl. See http://www.ensembl.org.



Single Nucleotide Polymorphisms

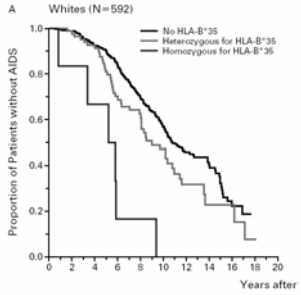
- Three step approach
- First, find the genes you are interested in
- Second, catalog all the polymorphisms in a gene (by sequencing)
- Third, measure those polymorphisms in a larger population





Clinical use of SNPs

- New publication with association of SNP with disease is almost a daily occurrence



Gao, X. et al. Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N Engl J Med* **344**, 1668-75 (2001).

SNPs and pharmacogenomics

A common polymorphism associated with antibiotic-induced cardiac arrhythmia

Federico Sesti*, Geoffrey W. Abbott*†, Jian Wei‡, Katherine T. Murray§, Sanjeev Saksena¶, Peter J. Schwartz||, Silvia G. Priori||, Dan M. Roden‡, Alfred L. George, Jr.‡, and Steve A. N. Goldstein*†

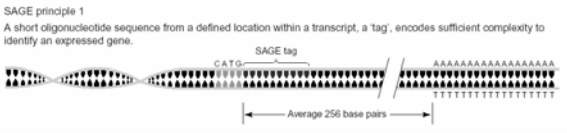
*Departments of Pediatrics and Cellular and Molecular Physiology, Boyer Center for Molecular Medicine, Yale University School of Medicine, New Haven, CT 06536; †Departments of Medicine and Pharmacology, Vanderbilt University, Nashville, TN 37235; ‡Robert Wood Johnson Medical School, Pamair, NJ 07065; and ||Department of Cardiology, University of Pavia and Policlinico San Matteo IRCCS, Pavia, Italy 27100

Edited by Vincent T. Marchesi, Yale University School of Medicine, New Haven, CT, and approved July 6, 2000 (received for review May 16, 2000)

- Genes will help us determine which drugs to use in particular disease subtypes
- Genes will help us predict those who get side-effects

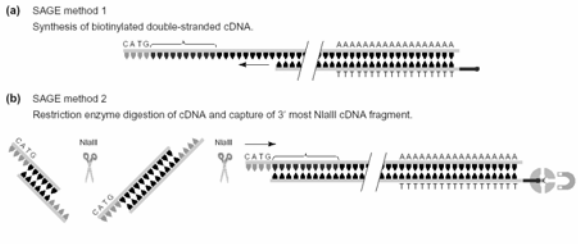
Sesti F. PNAS 97:10613, 2000

Serial Analysis of Gene Expression

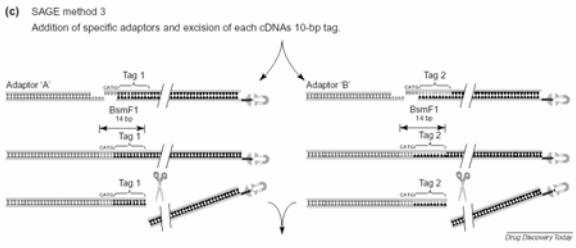


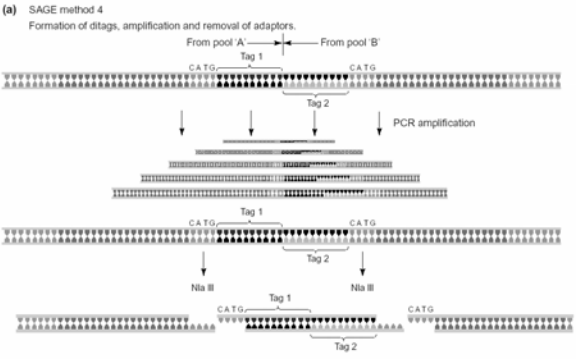
Madden, S. L., Wang, C. J. & Landes, G. Serial analysis of gene expression: from gene discovery to target identification. *Drug Discov Today* **5**, 415-425 (2000).

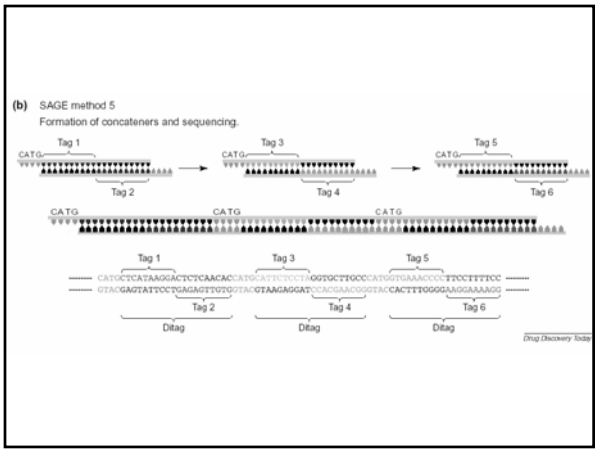
Serial Analysis of Gene Expression



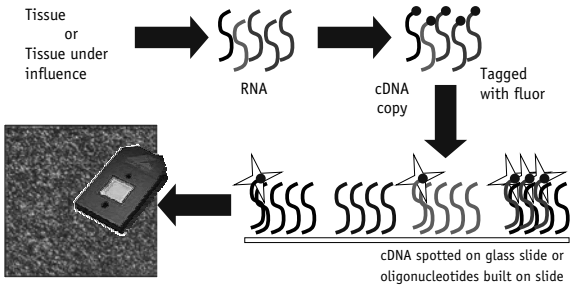
Serial Analysis of Gene Expression





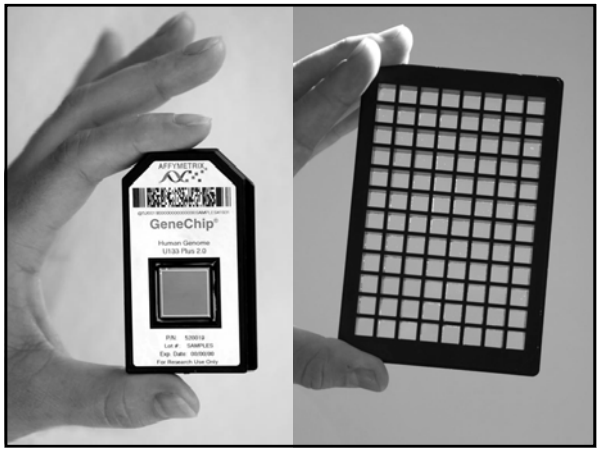


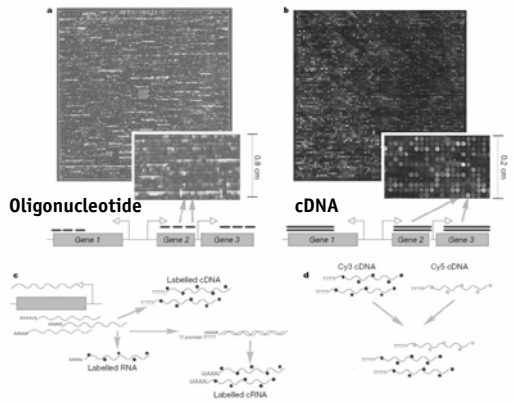
RNA expression detection chips



- Quantitative, absolute or relative
- Genes chosen arbitrarily
- Needs functional tissue

Schena M, et al. PNAS 93:10614 (1996).
Nature Genetics, 21: supplement (Jan 1999).

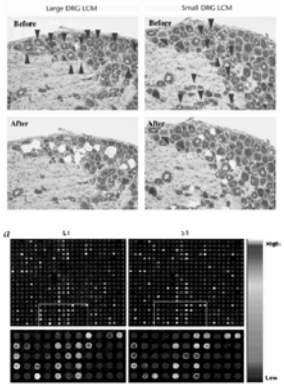




Lockhart, DJ. Winzeler, EA. Nature 405, 827-36 (2000).

Experiment Design

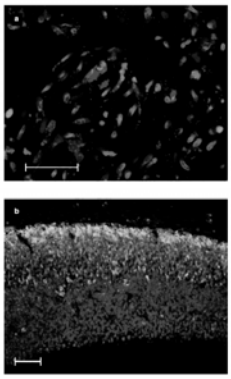
- Quantitate specific RNA expression before and after an intervention
- Compare expression between two tissue types
- Compare expression between different strains or constructed organisms
- Compare expression between neighboring cells



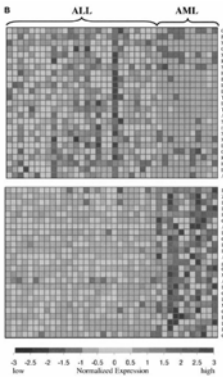
Luo L, et al. Nature Medicine; 5: 117 (1999).

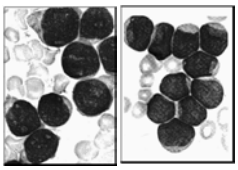
Validation

- In situ hybridization
- Real-time Polymerase Chain Reaction



Microarrays in Diagnosis

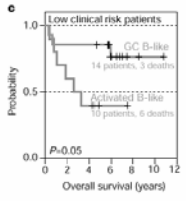




- Difficulty distinguishing between leukemias
- Microarrays can find genes that help make the diagnosis easier

Golub TR. Science 286:531, 1999.

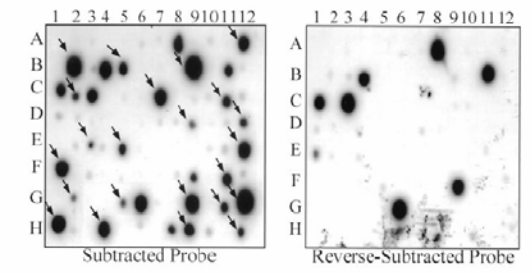
Microarrays in Prognosis



Alizadeh AA. Nature 403:503, 2000

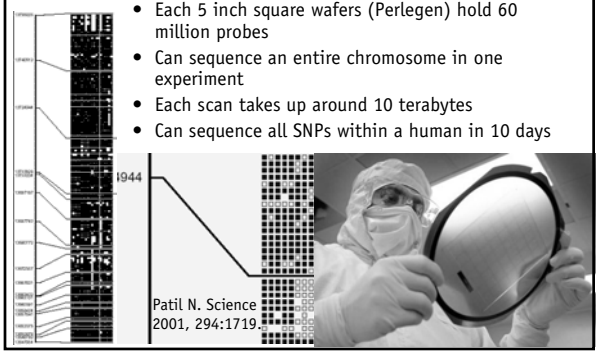
- Patients with seemingly the same B-cell lymphoma
- Looking at pattern of activated genes helped discover two subsets of lymphoma
- Big differences in survival

RNA Subtraction



After microarrays comes wafers...

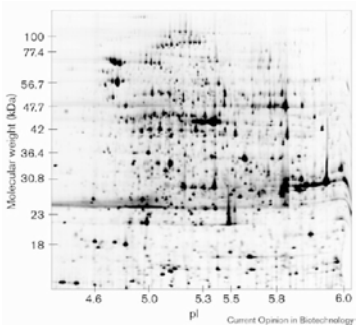
- Chromosome 21 has 21 million base-pairs
- Each 5 inch square wafers (Perlegen) hold 60 million probes
- Can sequence an entire chromosome in one experiment
- Each scan takes up around 10 terabytes
- Can sequence all SNPs within a human in 10 days

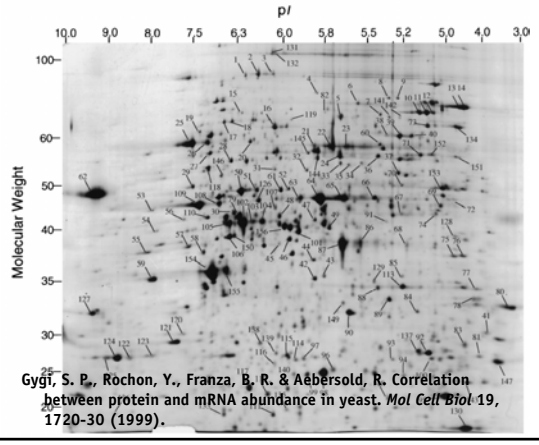


Patil N. Science 2001, 294:1719.

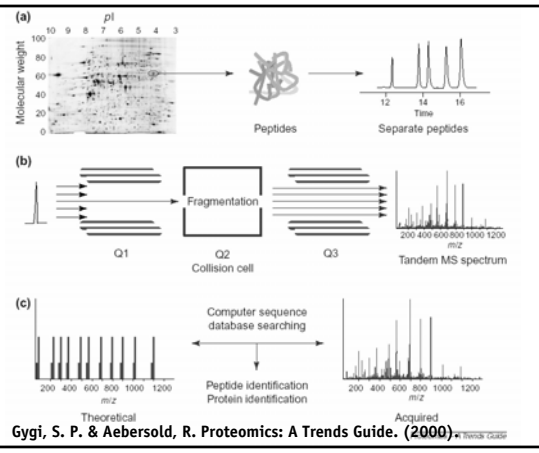
2D-PAGE

- Two axis = two properties of proteins: pH versus mass
- Global view of proteins
- Patterns can be scanned, saved and searched
- Spots need to be picked for identification
- Unfortunately, not very quantitative

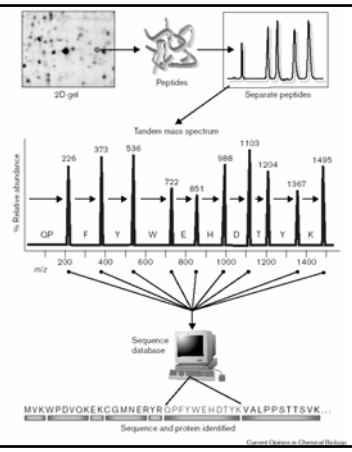




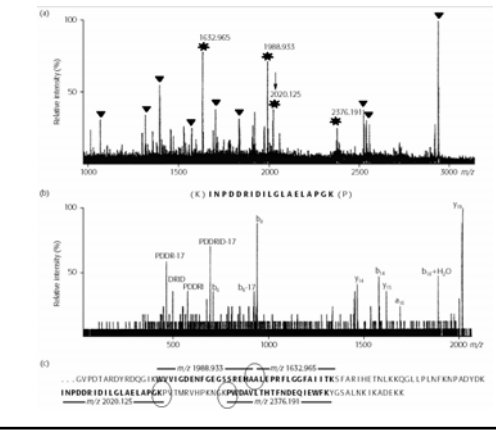
Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 19, 1720-30 (1999).

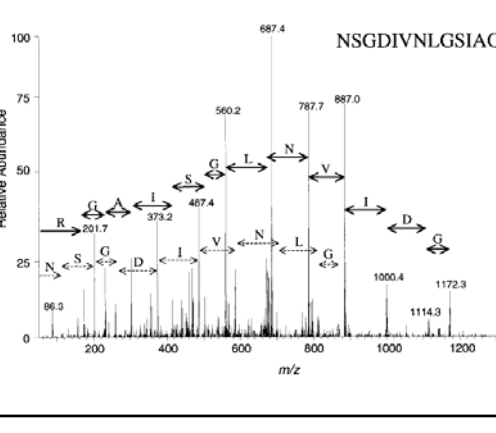


Gygi, S. P. & Aebersold, R. Proteomics: A Trends Guide. (2000).



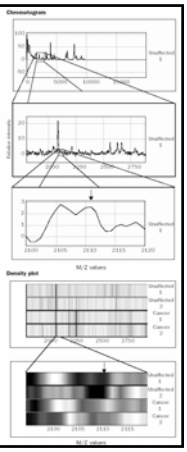
Gygi, S. P. & Aebersold, R. Mass spectrometry and proteomics. *Curr Opin Chem Biol* **4**, 489-94 (2000).





Clinical uses for proteomics

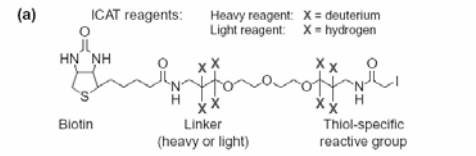
- Petricoin, et al., used this technique on serum
- Finding markers distinguishing ovarian cancer versus non-neoplasia
- Quest for biomarkers

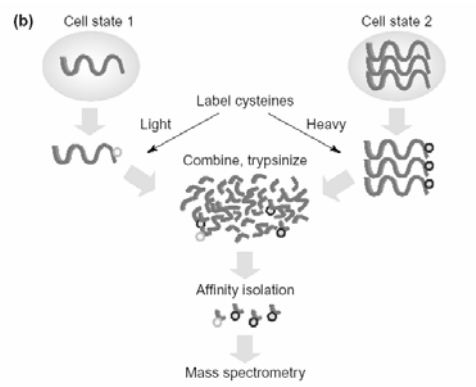


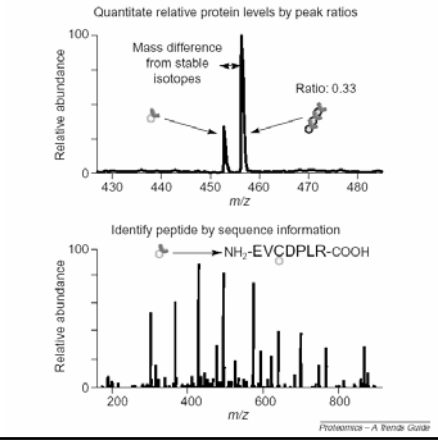
Petricoin, E. F. et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359, 572-7. (2002).

Quantitative proteomics

- The examples so far demonstrate identification, not quantification
- One can take advantage of the extreme sensitivity of detection of mass spectrometry
- Add to the proteins a known amount of label

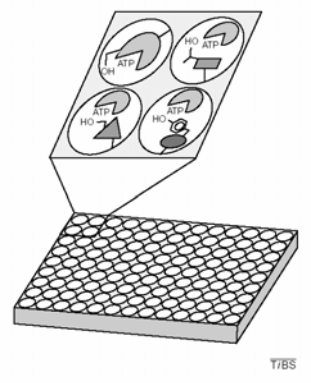






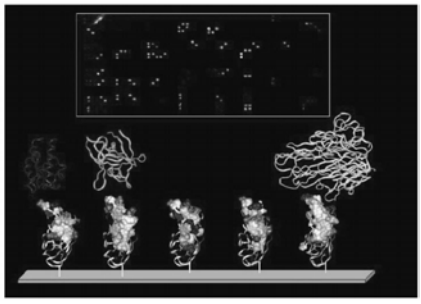
Protein chips

- Detection vs. Function
- Kinase chips



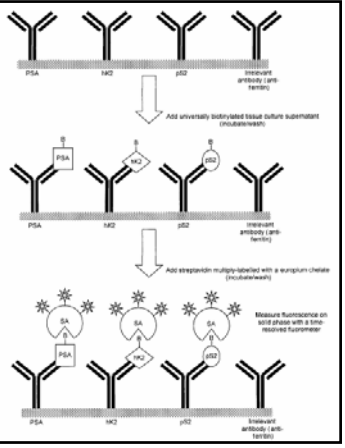
Williams, D. M. & Cole, P. A. Kinase chips hit the proteomics era. *Trends Biochem Sci* **26**, 271-3 (2001).

Functional binding



Protein Detection

- Specific antibodies
- Antibodies need to be available



Gene Measurement Techniques

DNA

- Sequencing
- Polymorphisms

RNA

- Serial analysis of gene expression
- DNA Microarrays
- Wafers

Protein

- 2D-PAGE
- Mass spectrometry
- Protein arrays