# The Fruits of the Genome Sequences for Society

David Botstein

NIGMS

CENTER FOR QUANTITATIVE BIOLOGY

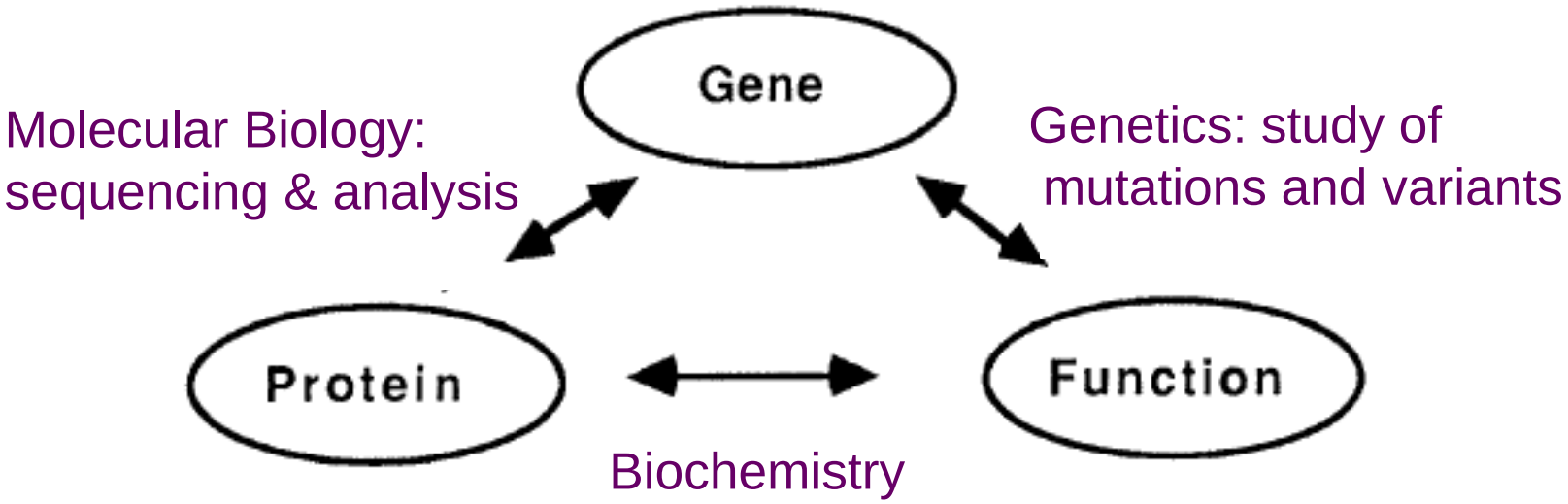
Lewis-Sigler Institute for Integrative Genomics
Princeton University

# Genome Sizes and Gene Numbers

| Organism | Genome Size | Genes (for Proteins) |
|---|---|---|
| Yeast | 12 megabases | 5,800 |
| Worm | 100 megabases | 19,400 |
| Fly | 120 megabases | 13,400 |
| Plant | 115 megabases | 25,500 |
| Human/Mouse | 3300 megabases | 22,000 |

The basic cellular functions of all eukaryotes are carried out by proteins (and RNAs) whose **structure and function** are conserved .

# Associating Biological Information with DNA Sequence




**Most of these associations were made, and likely will continue to be made, by basic scientists working with eukaryotic model systems (yeast, flies, worms, mice)**

# The Intellectual Impact of the Genomic View

- The "grand unification" of biology: all the functional parts of all living things are related by lineage. Despite the diversity, the fundamental biological mechanisms must also ultimately be related.

"*Once we understand the biology of E. coli, we will understand the biology of the elephant*" ---Jacques Monod, ca.1960

- The challenge for the future is to understand not just mechanisms at the individual process level, but also the interactions among all the processes and their mechanisms.
- Genomics makes possible experiments and analysis at the "systems" level. Because of the huge combinatorial possibilites for interactions, this means not just highly parallel experimental methods but also computation-intensive analysis.

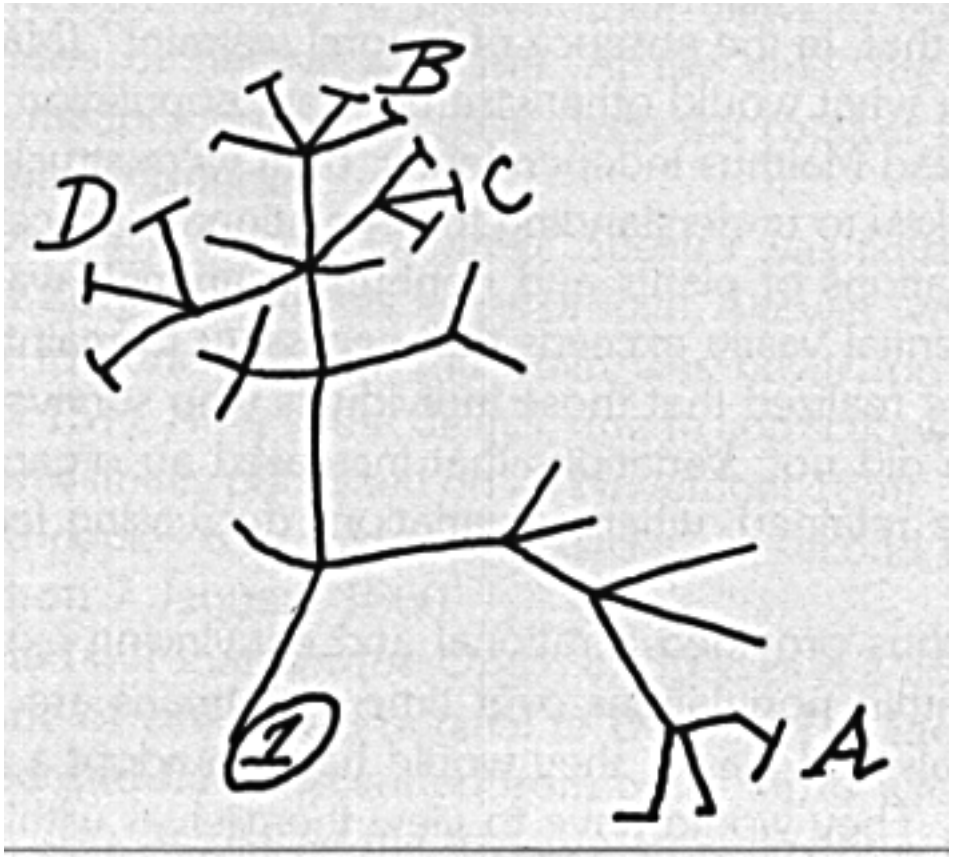
| Yeast/Mammalian Protein | Sequence Identity (%) | Function |
|---|---|---|
| Ubiquitin | 96 | yes |
| Actin | 89 | yes |
| ADP-Ribosylation Factor | 77 | yes |
| Beta-tubulin | 75 | partial |
| Alpha-tubulin | 74 | partial |
| Heat Shock HSP70 | 73 | |
| YPT1/Rab1 | 71 | yes |
| HMG-CoA Reductase | 67 | yes |
| Transcription Initiation Factor IID | 65 | yes |
| Cytochrome C | 63 | |
| KAR2/BiP | 62 | yes |
| Calmodulin | 60 | yes |
| RAS1/N-ras; RAS2/K-ras | 60 | yes |
| CDC28/CDC2 | 59 | yes |
| SEC18/NSF | 46 | yes |
| Cu-metallothionein | 30 | |
| Dihydrofolate Reductase | 32 | yes |
| Profilin | 28 | yes |
| P-glycoprotein/MDR | 26 | yes |
| Glucose Transporter | 25 | yes |

Botstein and Fink, 1988 (updated)

# Fruits of the Genome

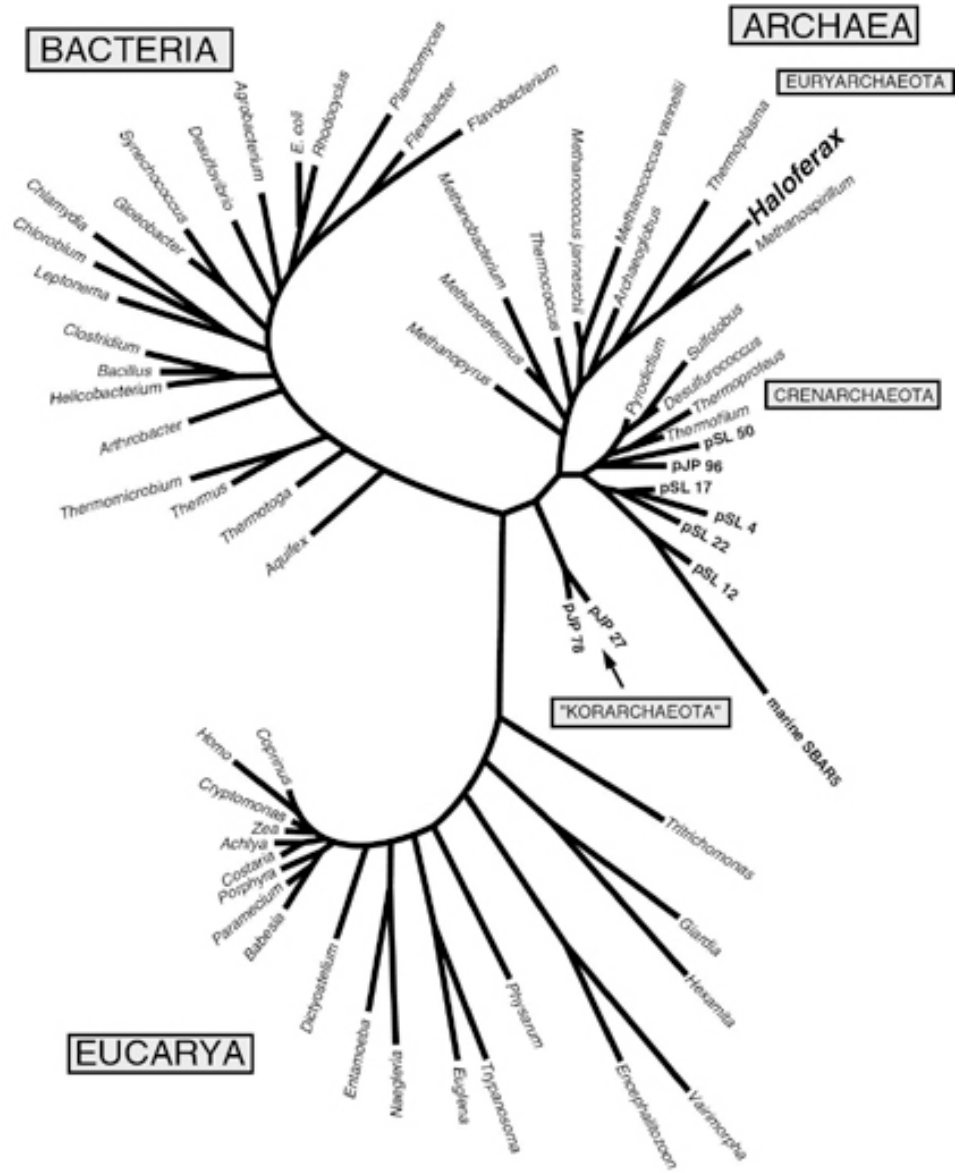
- Quantitative understanding of evolution from sequence.
- Comparative Genomics: the “grand unification” of biology.
- New comprehensive technologies--- metagenomics, metabolomics, etc.
- The many uses of DNA sequence variation: from forensics to disease gene mapping and identification.
- Functional Genomics: defining diseases through gene identities and genome-scale patterns of gene expression.
- DNA Diagnostics: detecting disease, disease progression and predisposition to disease.

# Darwin's Great Intuitive Insight



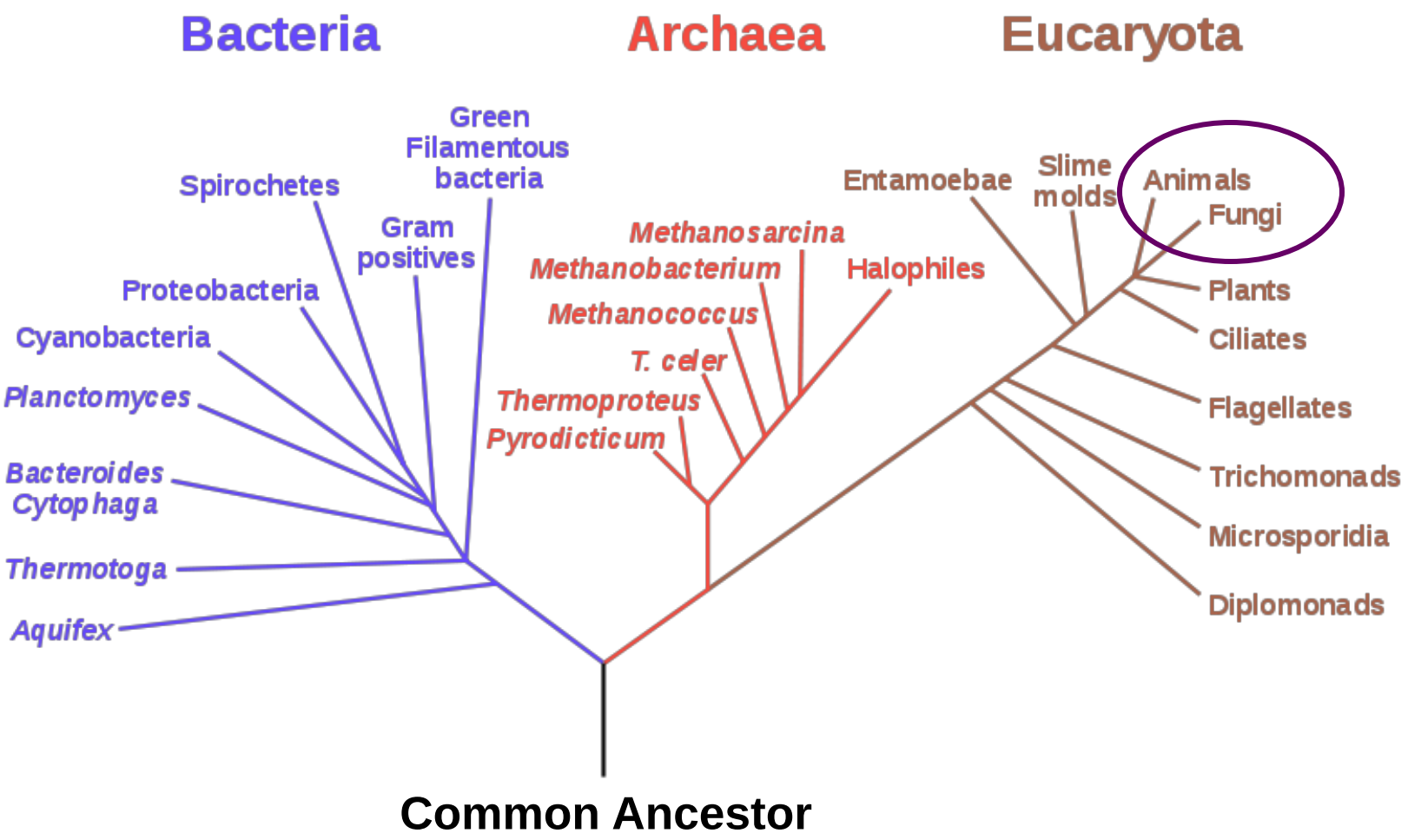

Charles Darwin (1837)

# “Universal” Unrooted Phylogenetic Tree of Life

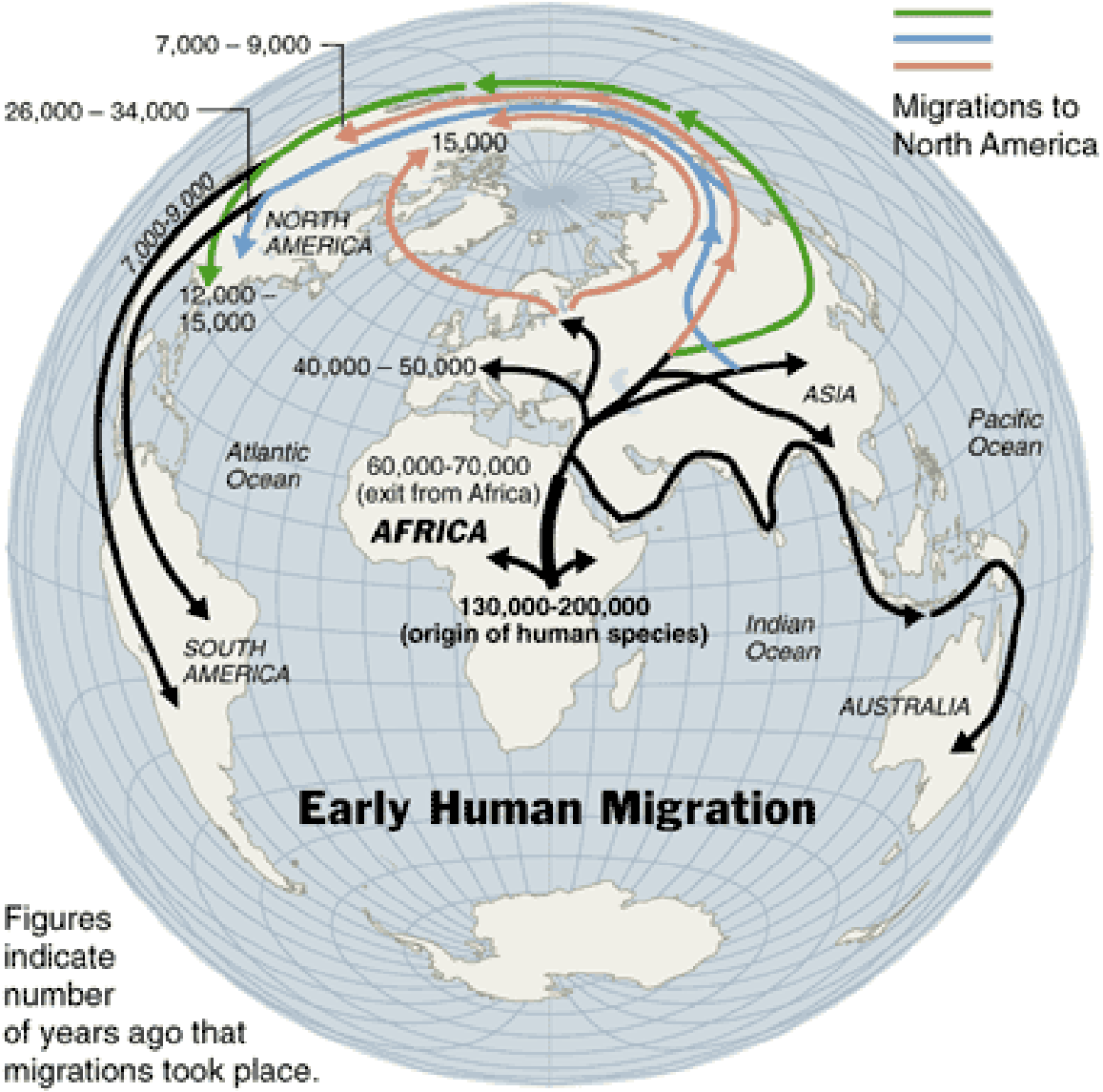



Barnes, S.M. *et al.*, 1996, Proc. Natl. Acad. Sci. USA, **93**: 9188-9193.

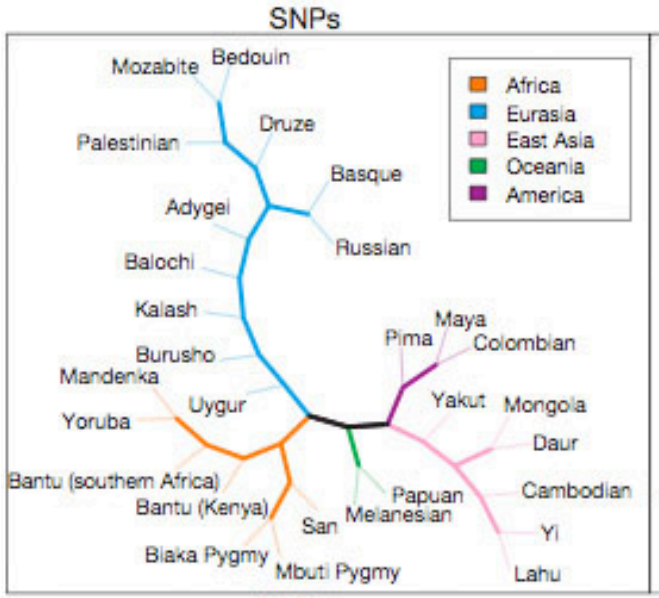
# Rooted Phylogenetic Tree of Life

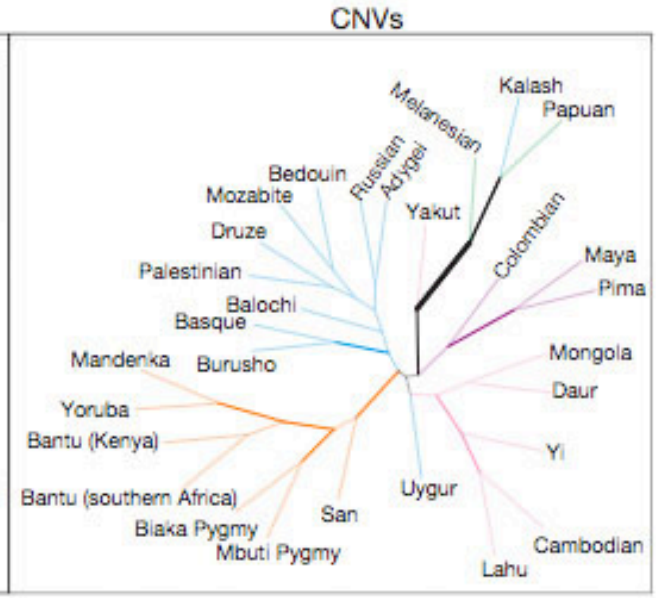


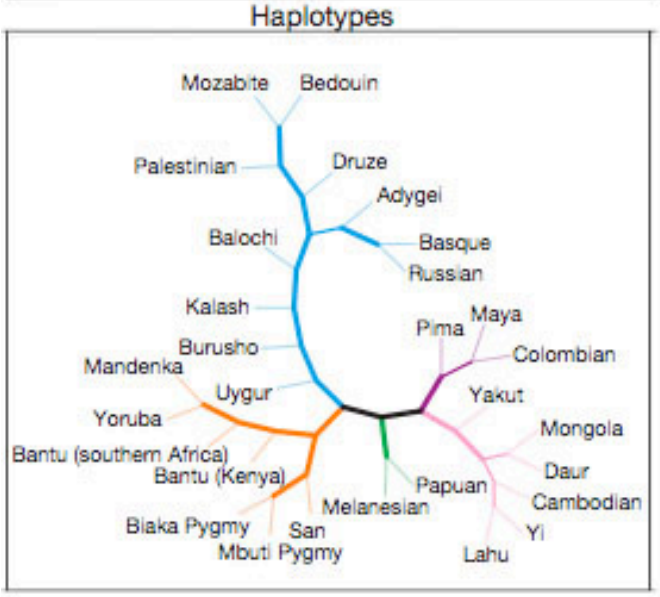
# Out of Africa: The evolutionary path of the human species



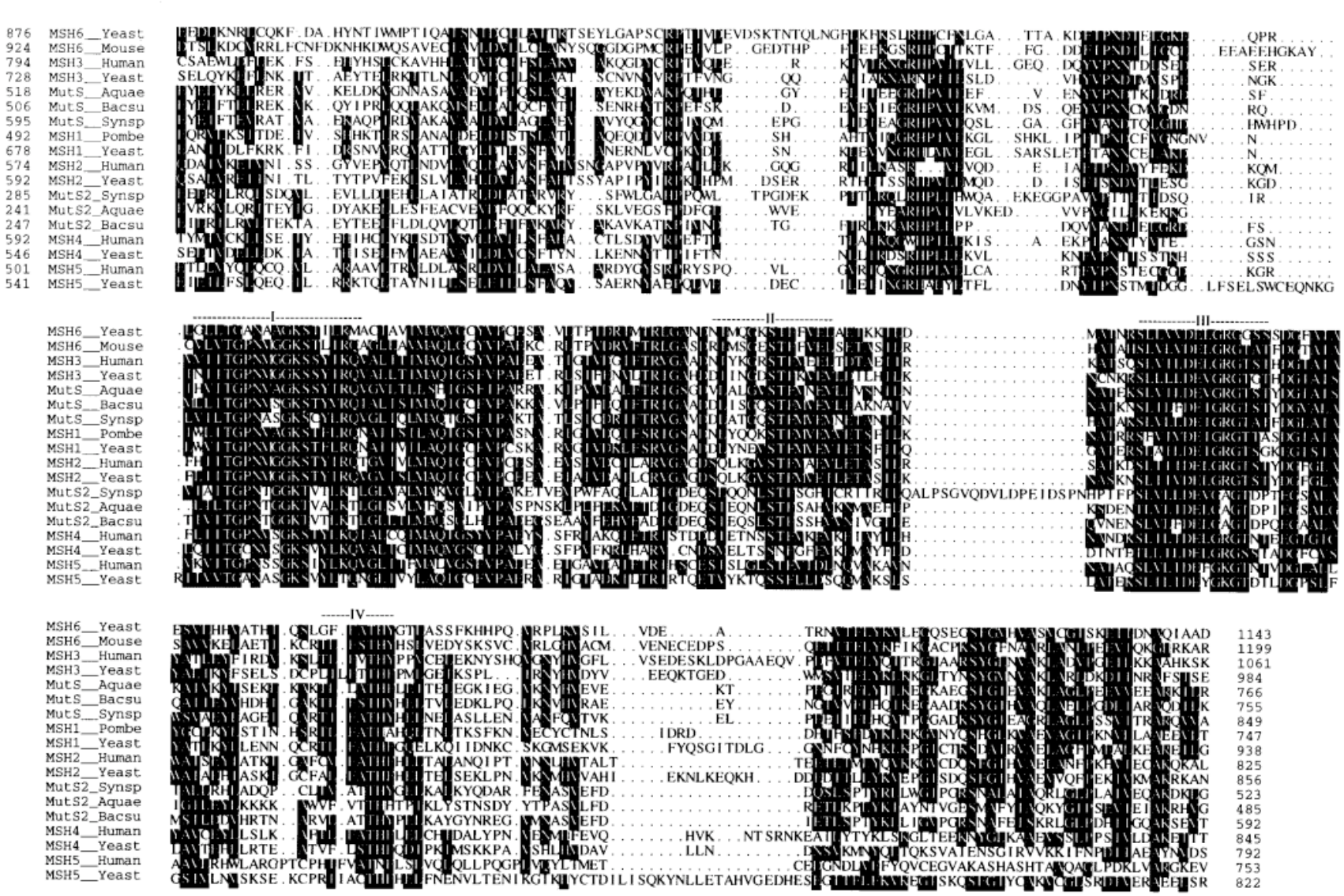
# Age and Diversity of Human Populations



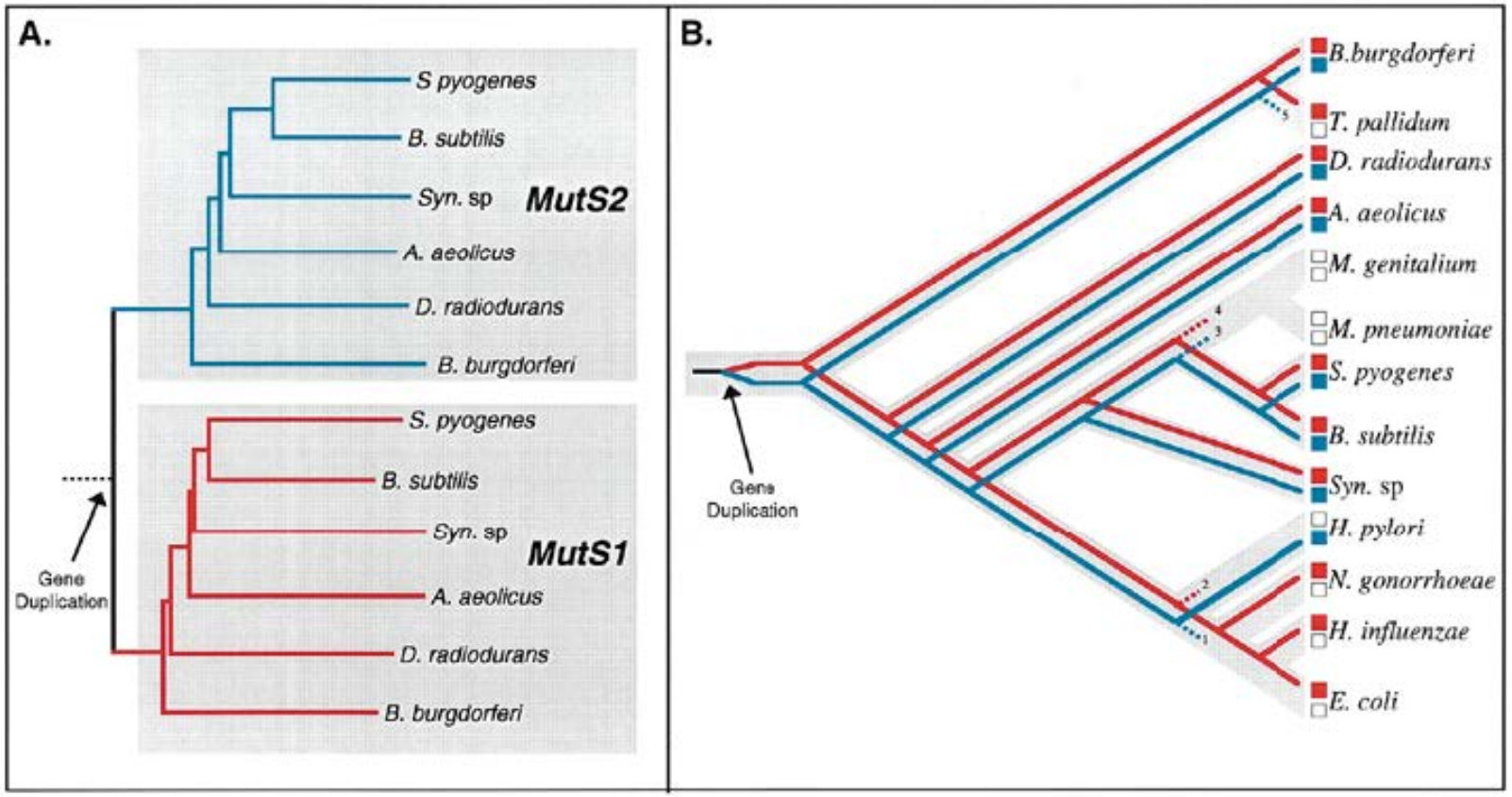



# Multiple Sequence Alignment of *mutS* Homologs



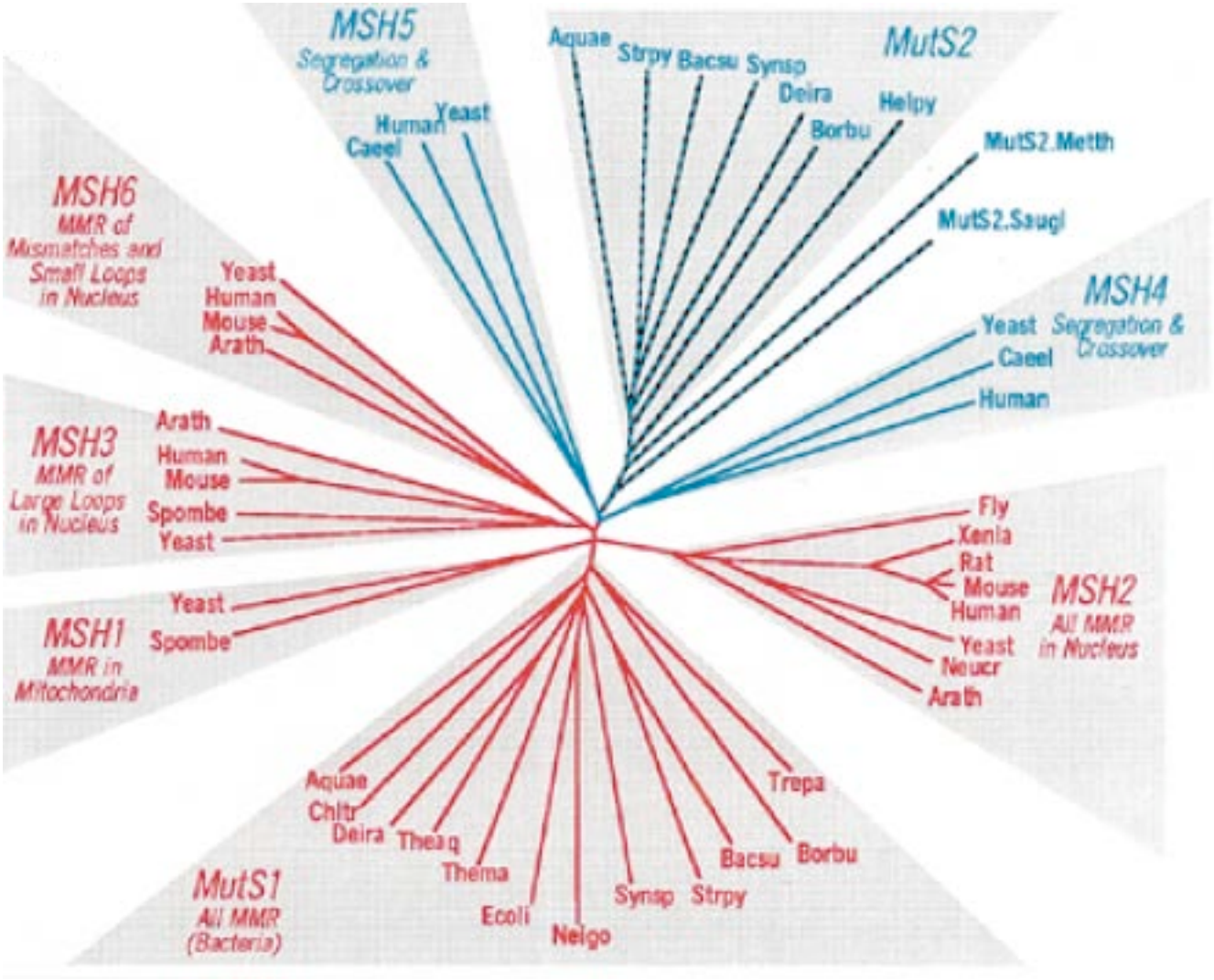

[J.A. Eisen *Nucleic Acids Research, 1998, Vol. 26, No. 18*]

# Distinguishing Orthologs and Paralogs from a Gene Family by Parsimonious Assignment of Gene Duplications and Losses




[J.A. Eisen *Nucleic Acids Research, 1998, Vol. 26, No. 18*]

# MutS Homologs Evolve Diverged Functions




[J.A. Eisen *Nucleic Acids Research, 1998, Vol. 26, No. 18*]

# Extracting Functional Information from the Human Genome Sequence

- Finding and Characterizing Human Disease Genes
  - DNA polymorphisms (SNPs & haplotypes)
  - Simple Mendelian (ca. 5000)
  - Complex (relatively few)
  - Pharmacogenomics (just starting)
- Comparative Genomics: associating human genes with their functional equivalents in experimental model systems
  - Using the evolutionary information: orthologs and paralogs
  - Genetic alterations, RNAi and other gene-based interventions
- Patterns of Gene Expression
  - DNA microarrays & Quantitative PCR
  - Immediately useful for diagnosis (e.g. cancer subtypes)
- Systems Biology: understanding at a different level?
  - Signal transduction, pathways, interactions

# Mapping Human Genes using DNA Polymorphisms

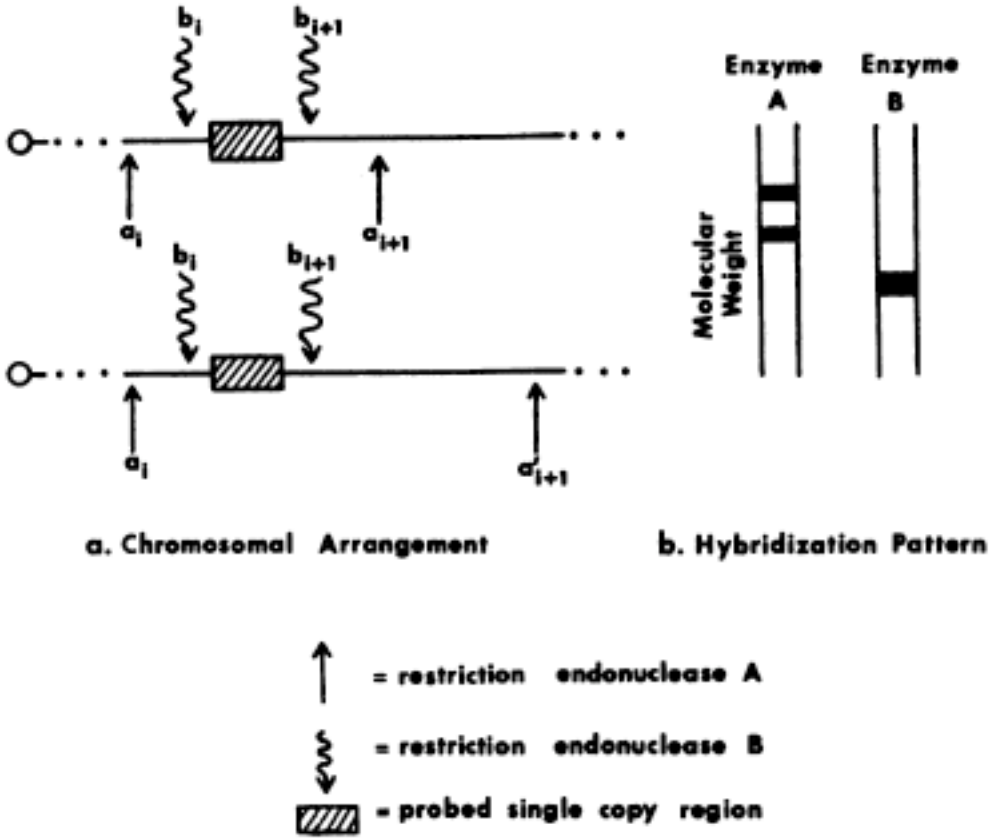
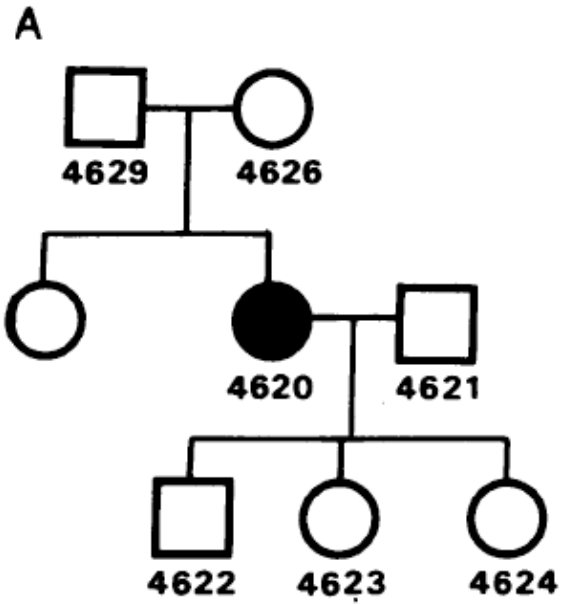


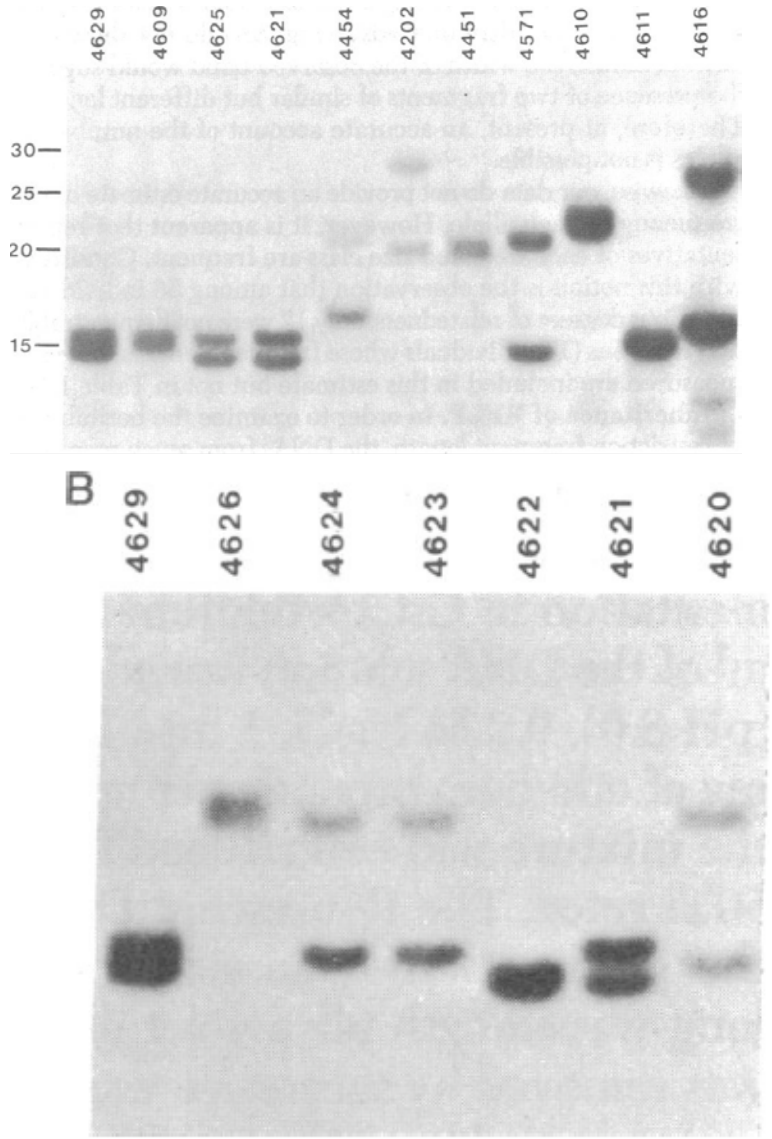

FIG. 1.—*a*, Cuts made in pair of homologous chromosomes by enzyme A and enzyme B; *b*, hybridization pattern of enzymes A and B given cuts of *a*.

[Botstein, White, Skolnick & Davis, 1980]

# DNA Polymorphisms can map human disease genes by linkage

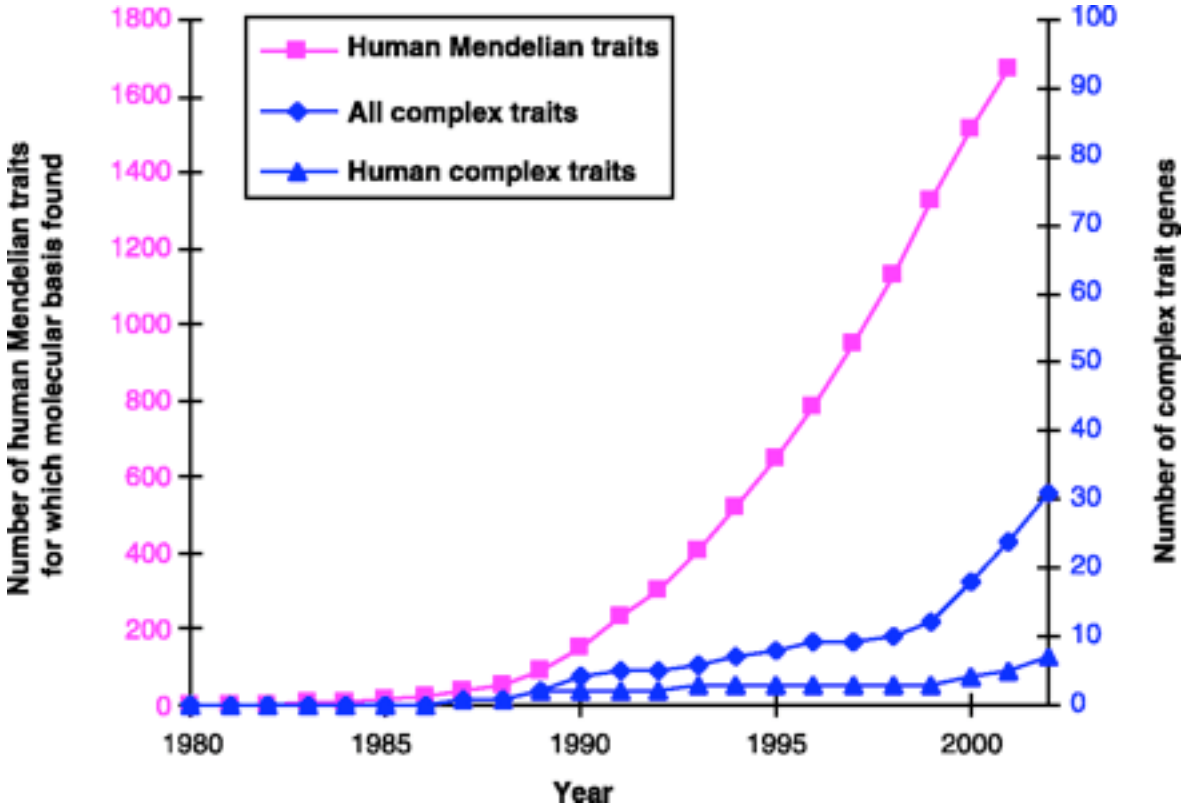
The original RFLP




[Wyman and White, 1980]

# Thousands of Inherited Disease Genes have been Found




[Glazier Nadeau & Aikman, 2006]

In 2006, OMIM had 2,799 of a total of 4,466 Mendelian phenotypes (mostly inherited diseases) as having been associated with specific genes. Today it is nearer 4,000.

# Gene Identification through Linkage Mapping Provides Basic Mechanistic Information for Inherited Diseases

Huntington's Disease ----> class of amplification of trinucleotide repeat diseases (myotonic dystrophy, fragile X, spinocerebellar ataxia, etc.

Amyotrophic Lateral Sclerosis ----> understanding of the critical issues around reactive oxygen species in the brain.

Ataxia-telangiectasia and BRCA1---> implication of cell cycle checkpoints and DNA repair in the etiology of cancer.

Retinoblastoma: Realization that cancer can be caused by loss of function as easily as by inappropriate gain of function

# DNA Evidence is Ubiquitous in Crime Fiction

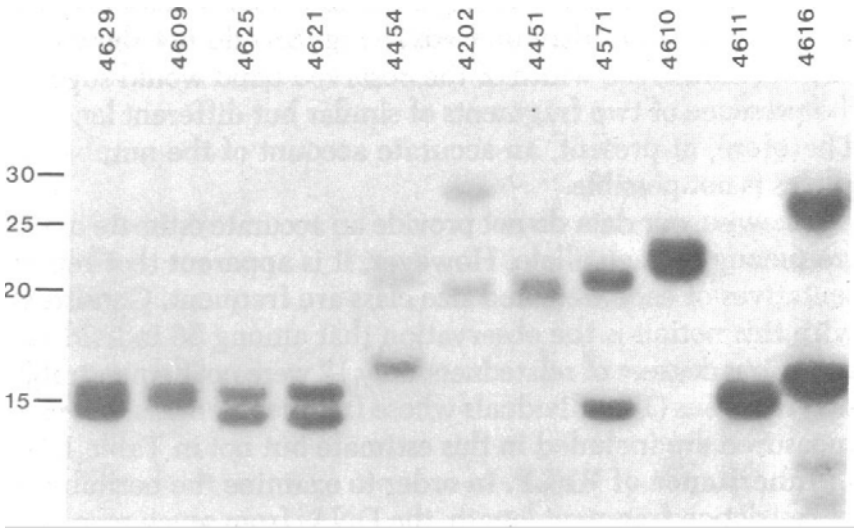


Watching these shows, it becomes clear that most (if not quite all) plots involve DNA evidence.

# DNA Polymorphisms are Abundant in the Human Genome

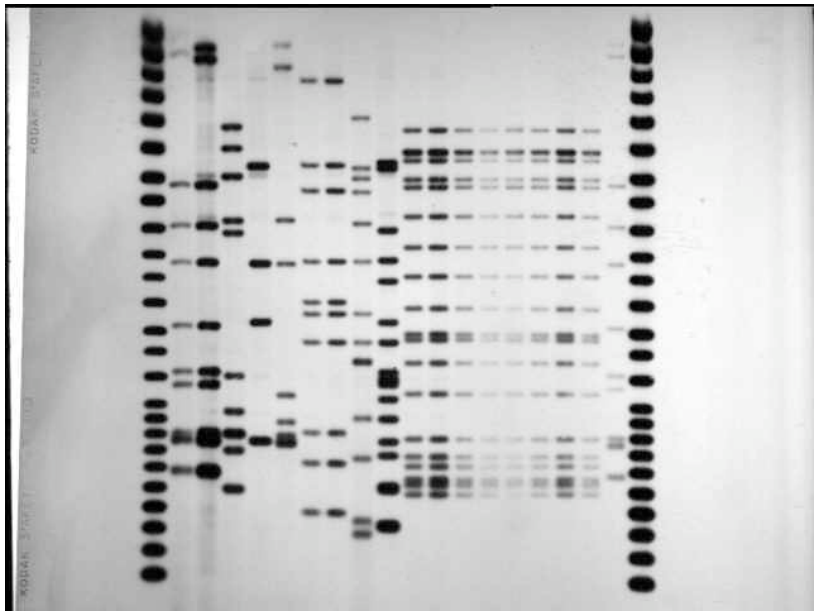
The original RFLP

[Wyman and White, 1980]

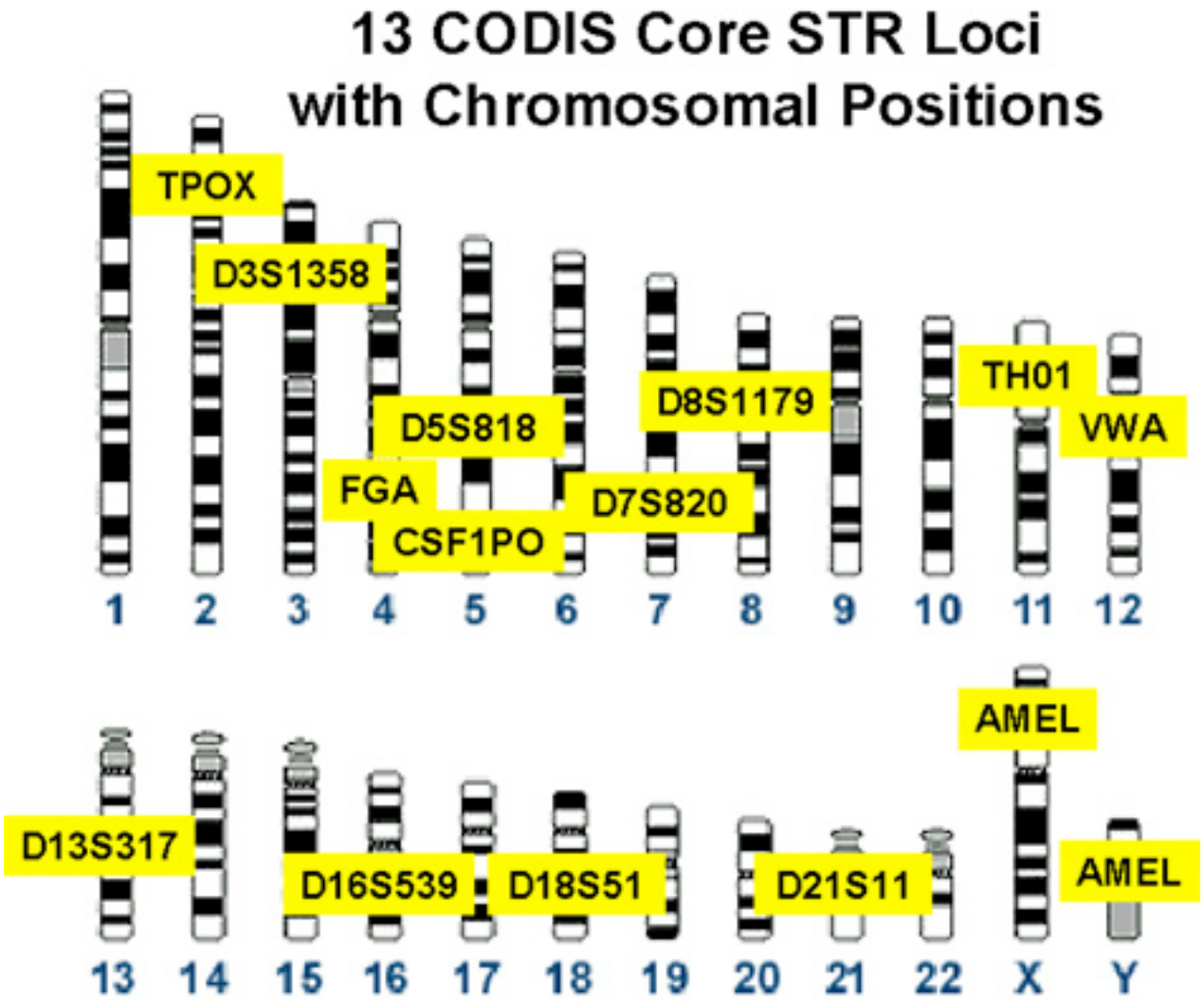



Markers from a commercial DNA Forensics laboratory

[Ryan Forensic website]



# The FBI has Settled on a Standard Set of Multiallelic Markers




CODIS: Combined DNA Index System: Federal Bureau of Investigation

# Non-Inherited Dinucleotide Repeat Polymorphisms Appear in Colon Tumor Cells

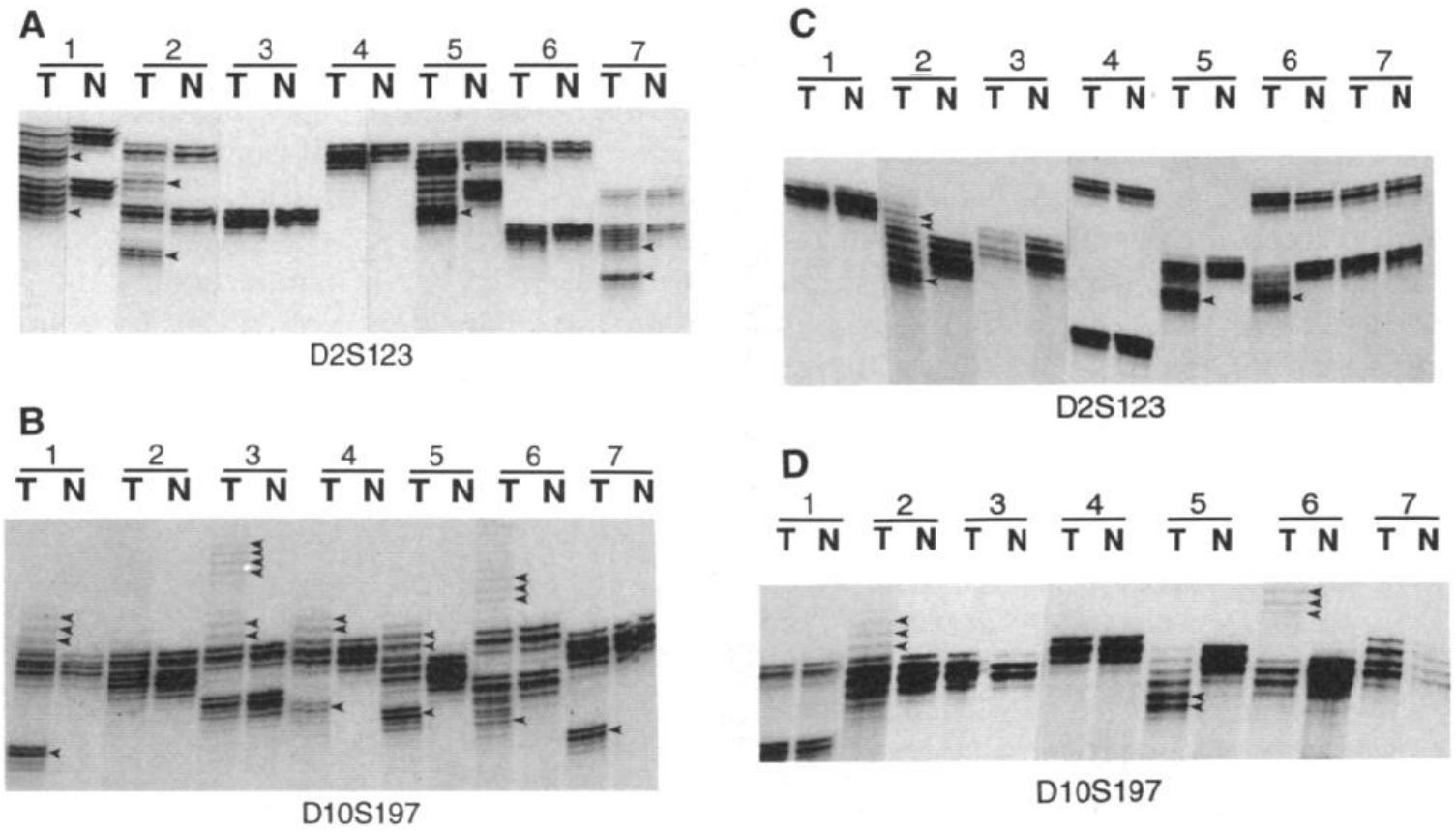



**Fig. 2.** (**A** and **B**) Dinucleotide repeat polymorphisms in normal and tumor tissue from HNPCC patients. The microsatellite markers D2S123 and D10S197 were used in PCR analysis (*5*, *23*), and

[Aaltonenen et al., 1993]

# Isolation of Yeast *msh2* and *mlh1* Mutations, with a Hypothesis, September 1993

## Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair

Micheline Strand*, Tomas A. Prolla†§, R. Michael Liskay‡§ & Thomas D. Petes*

Finally, we note that the phenotype of the mutation involved in one type of familial colorectal cancer (decreased stability of simple repeats)[2–4] is that predicted for a mutation affecting DNA mismatch correction. Such a mutation could represent a functional homologue of *PMS1*, *MLH1* or *MSH2* or another component of the mismatch repair system (for example, a DNA helicase or single-strand binding protein). □

Nature 365:274 (September 16, 1993)

# The Human *MSH2* Ortholog Predisposes to HNPCC (Human Non-Polyposis Colon Cancer)

Cell, Vol. 75, 1027–1038, December 3, 1993, Copyright © 1993 by Cell Press

## The Human Mutator Gene Homolog *MSH2* and Its Association with Hereditary Nonpolyposis Colon Cancer

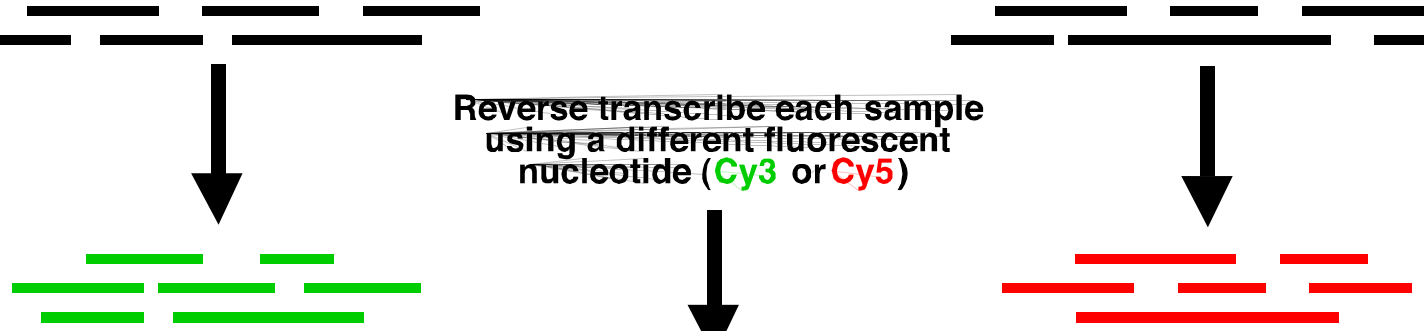
Richard Fishel,* Mary Kay Lescoe,* M. R. S. Rao,§
Neal G. Copeland,† Nancy A. Jenkins,†
Judy Garber,‡ Michael Kane,§
and Richard Kolodner§

Today, it is known that ca. 90% of all familial HNPCC families have mutations in either the human MSH2 or MLH1 homologs

# Genome-Wide Gene Expression Patterns Determined Using Hybridization to DNA Microarrays

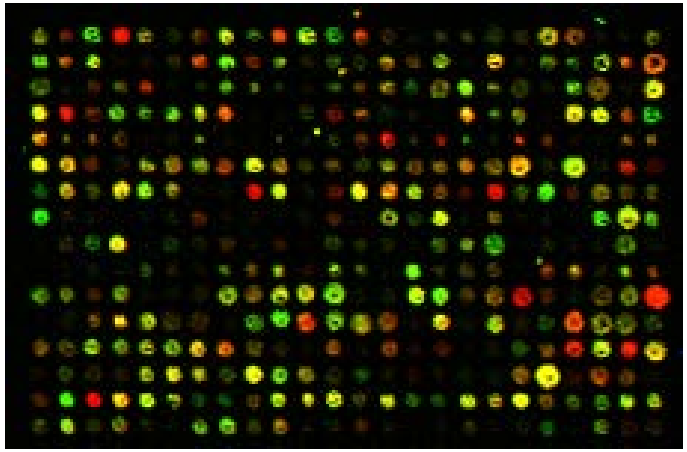
mRNA from Sample 1
(Reference or Experimental Sample)

mRNA from Sample 2
(Experimental Sample)



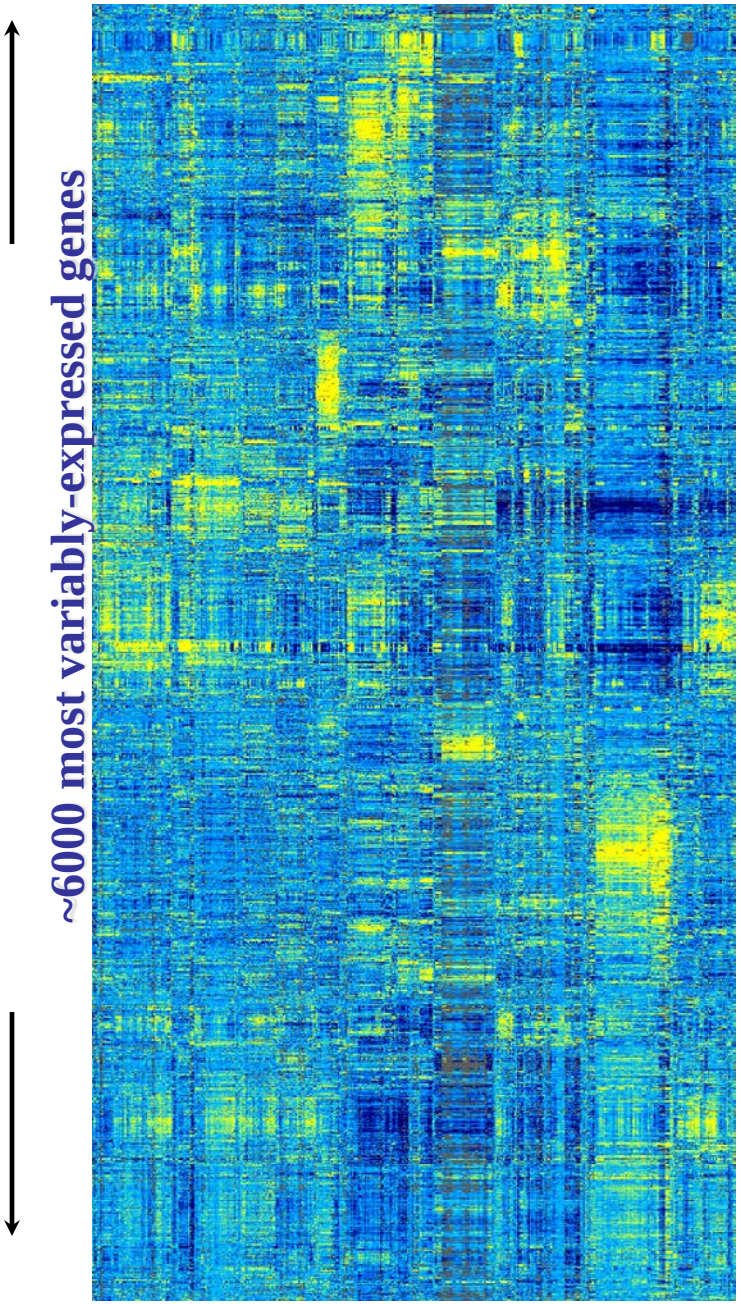
Reverse transcribe each sample using a different fluorescent nucleotide (Cy3 or Cy5)

Mix the complex probes together and hybridize overnight



Scan and determine fluorescence intensities at each spot

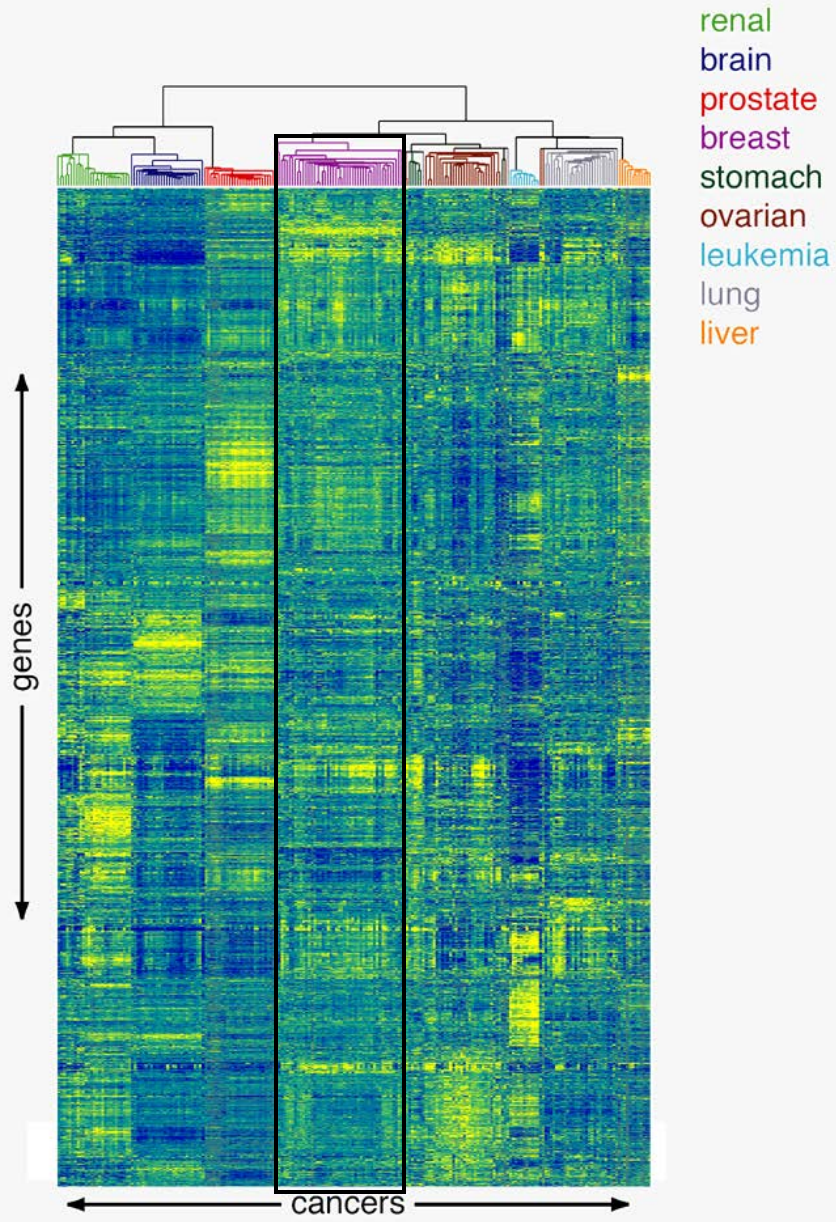
440 human cell and tissue samples (out of more than 20,000)



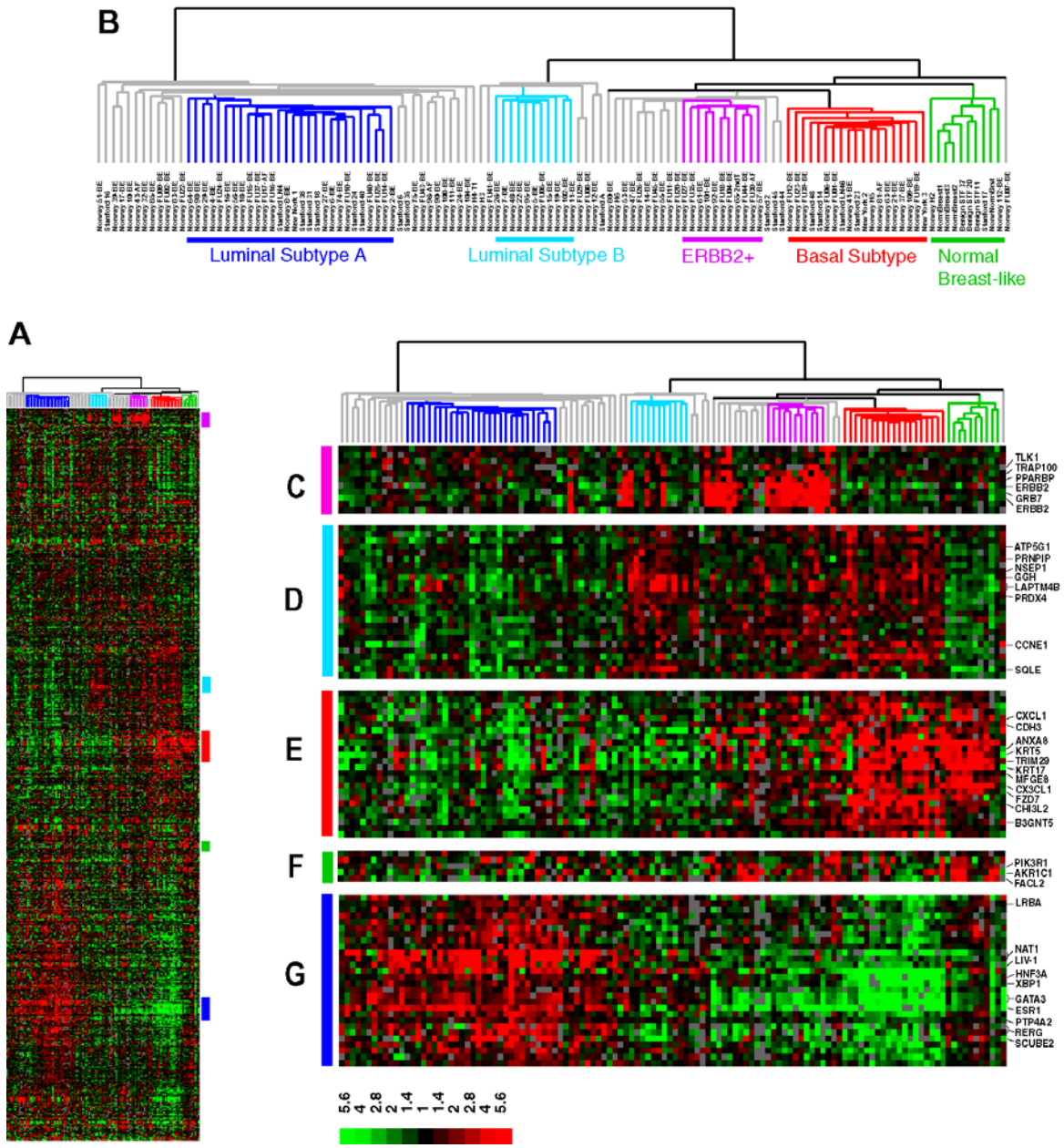

A new kind of map of the human genome…

**Pat Brown**
**Mike Eisen**
Max Diehn
Xin Chen
Jon Pollack
Chuck Perou
Therese Sorlie
Mitch Garber
Marci Schaner
Matt van de Rijn
Gavin Sherlock
Mike Fero

# Molecular portraits of cancer

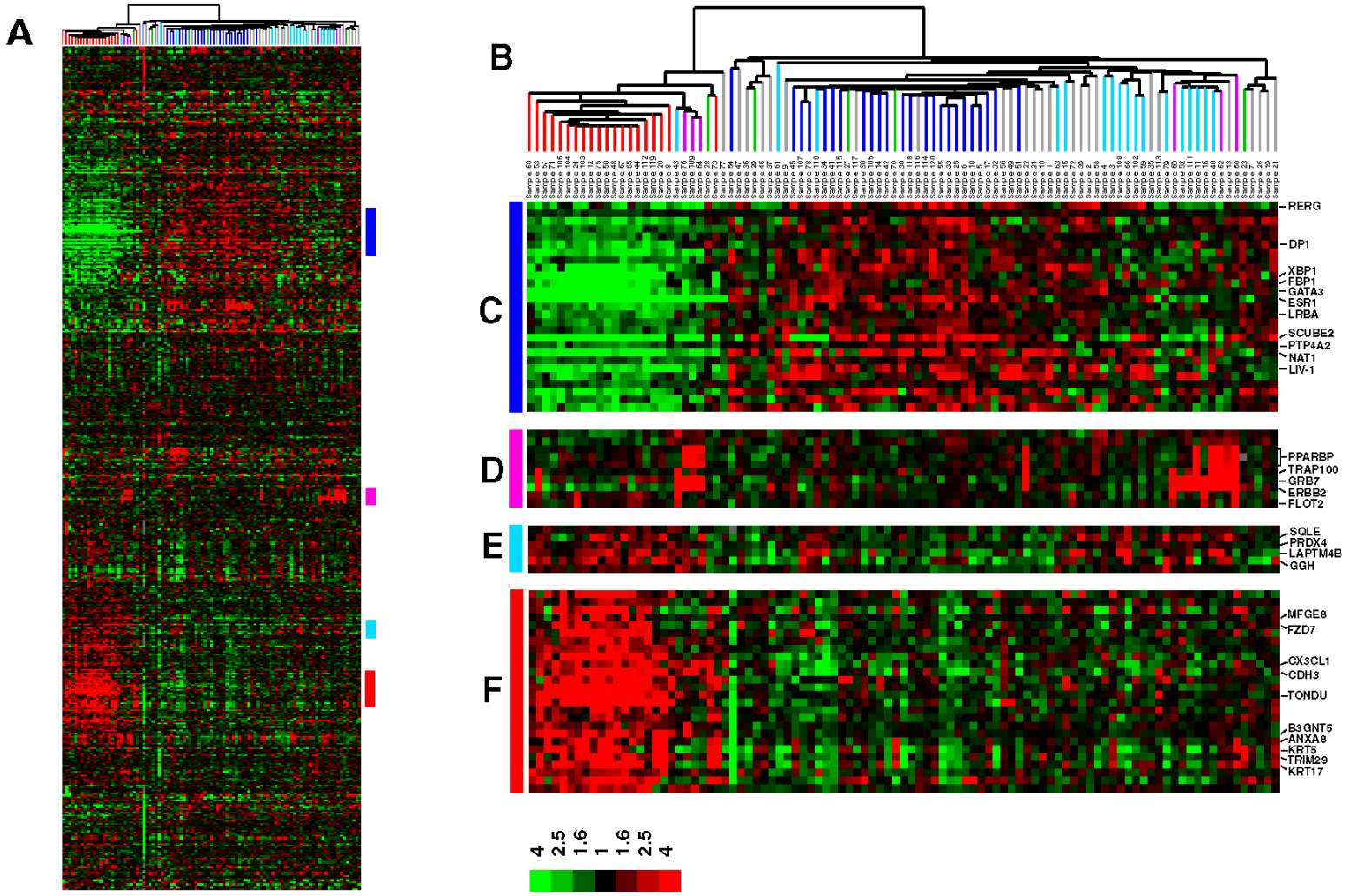


# Molecular Portraits of Breast Tumors: Norway/Stanford Cohort

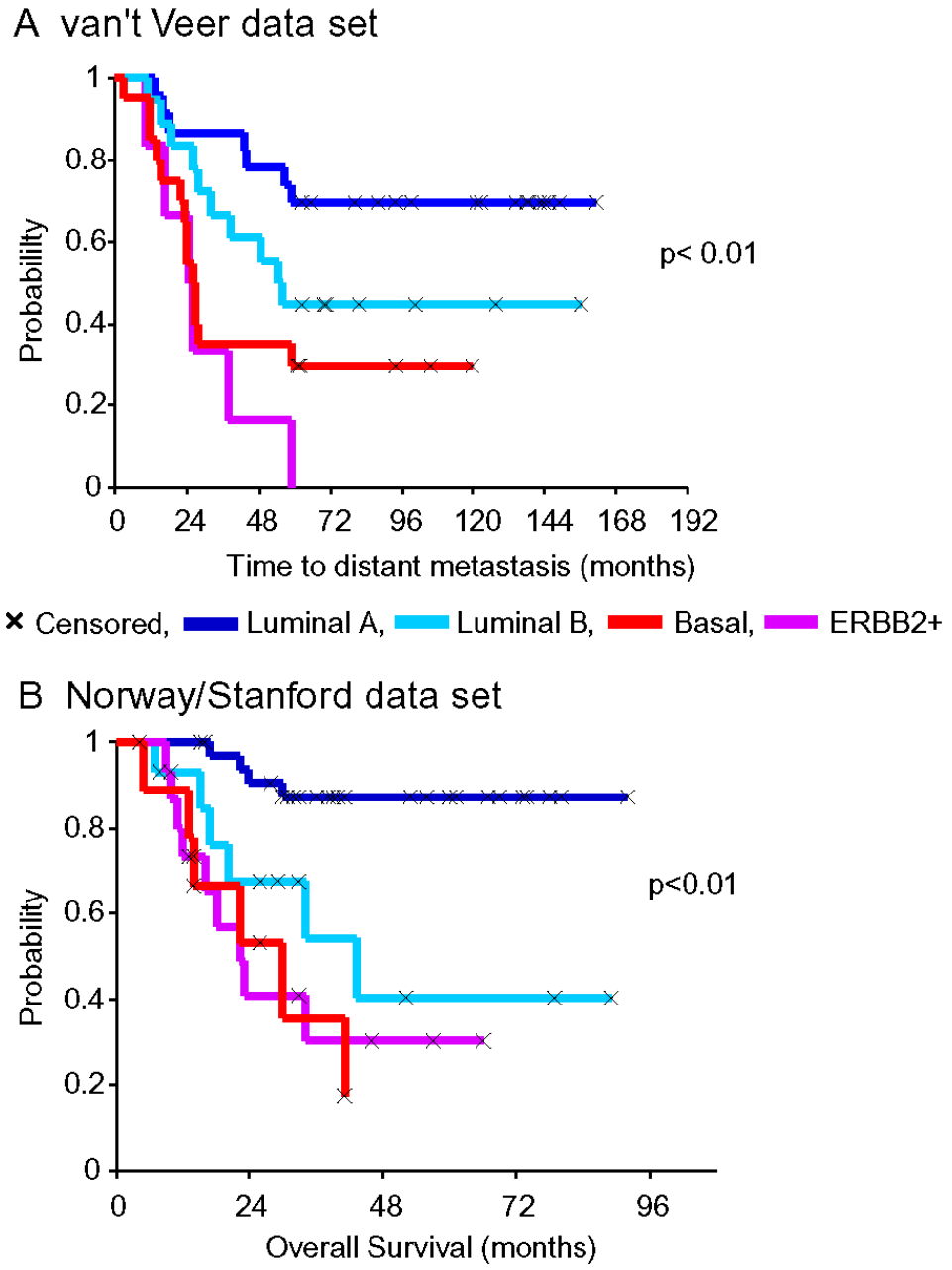


# Molecular Portraits of Breast Tumors: Dutch Cohort

(Data from van t' Veer et al, 2002)



# Correlation of Subtype with Outcome in Different Cohorts



# A genomic hypothesis test

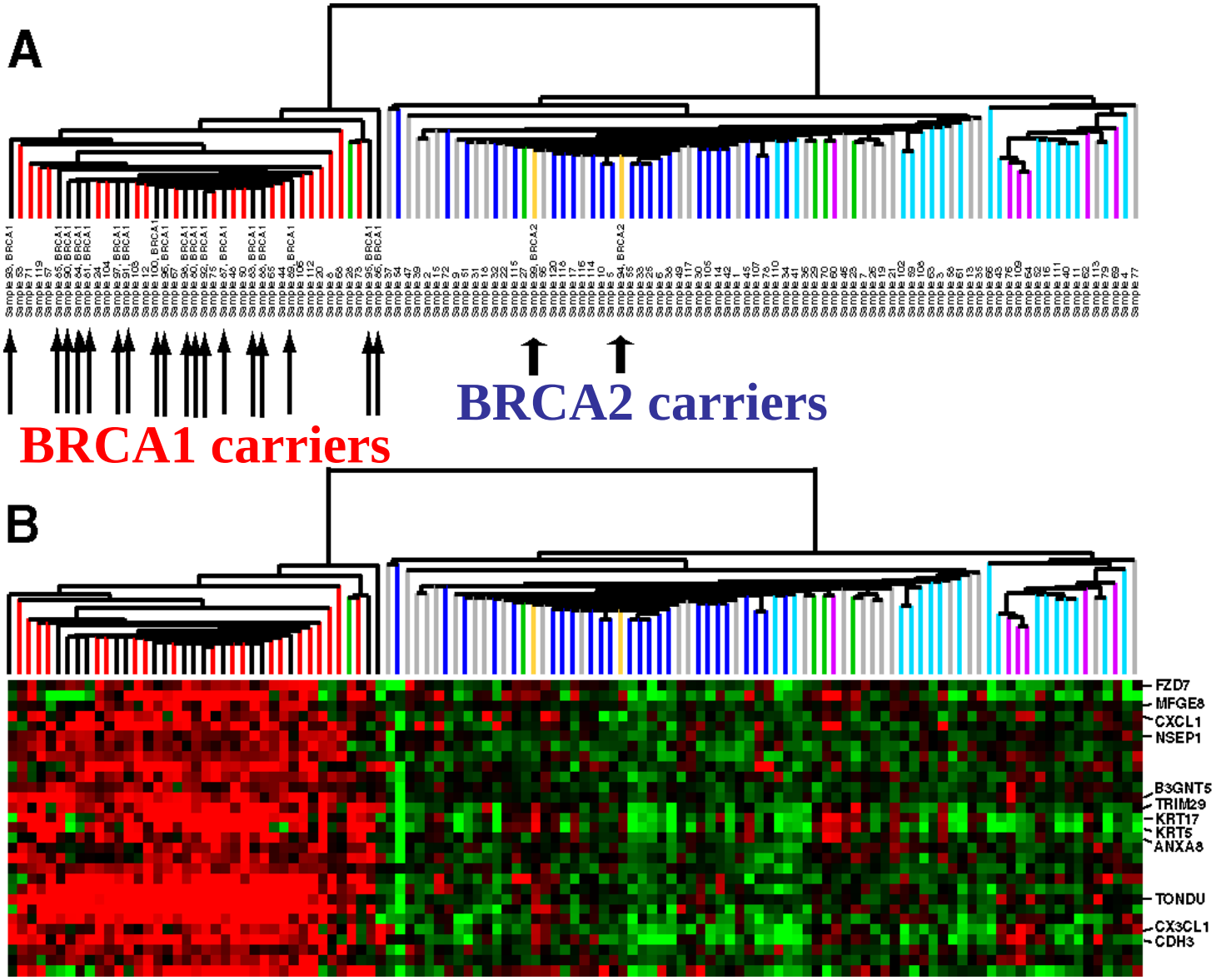
Hypothesis: the four breast cancer subtypes represent fundamentally different diseases arising from different cell types and/or by different pathways of oncogenesis.

If so, then women who inherit genes predisposing to breast cancer, and who thereby have a many-fold increased risk, might all be expected to have the same tumor subtype.

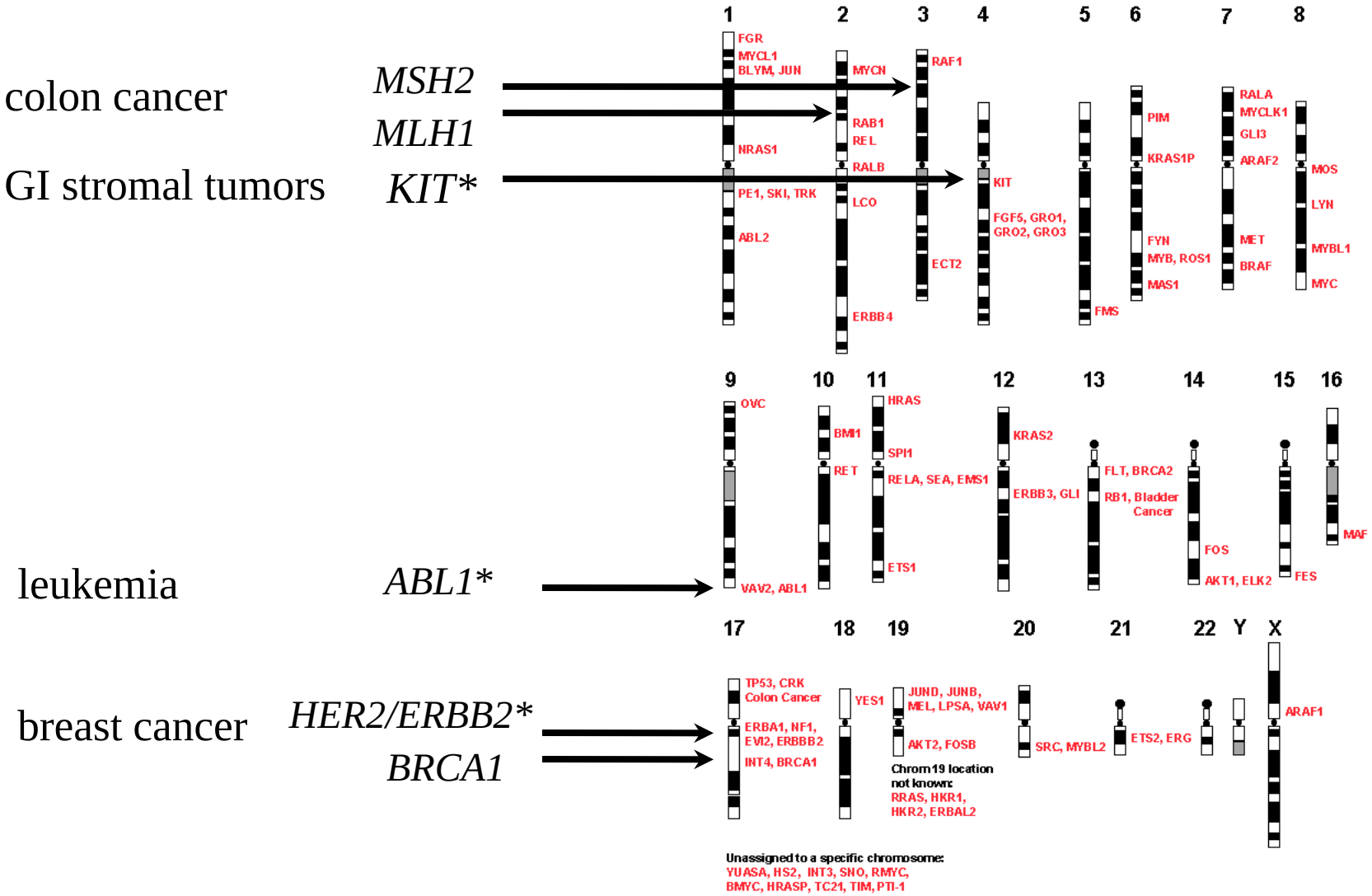
Test: Assess the patterns of gene expression of breast tumors in BRCA1 or BRCA2 carriers.

# BRCA1 mutations predispose to tumors of the "Basal" subtype

(Data from van t' Veer et al, 2002)



# Examples of Human Cancer-Causing Genes

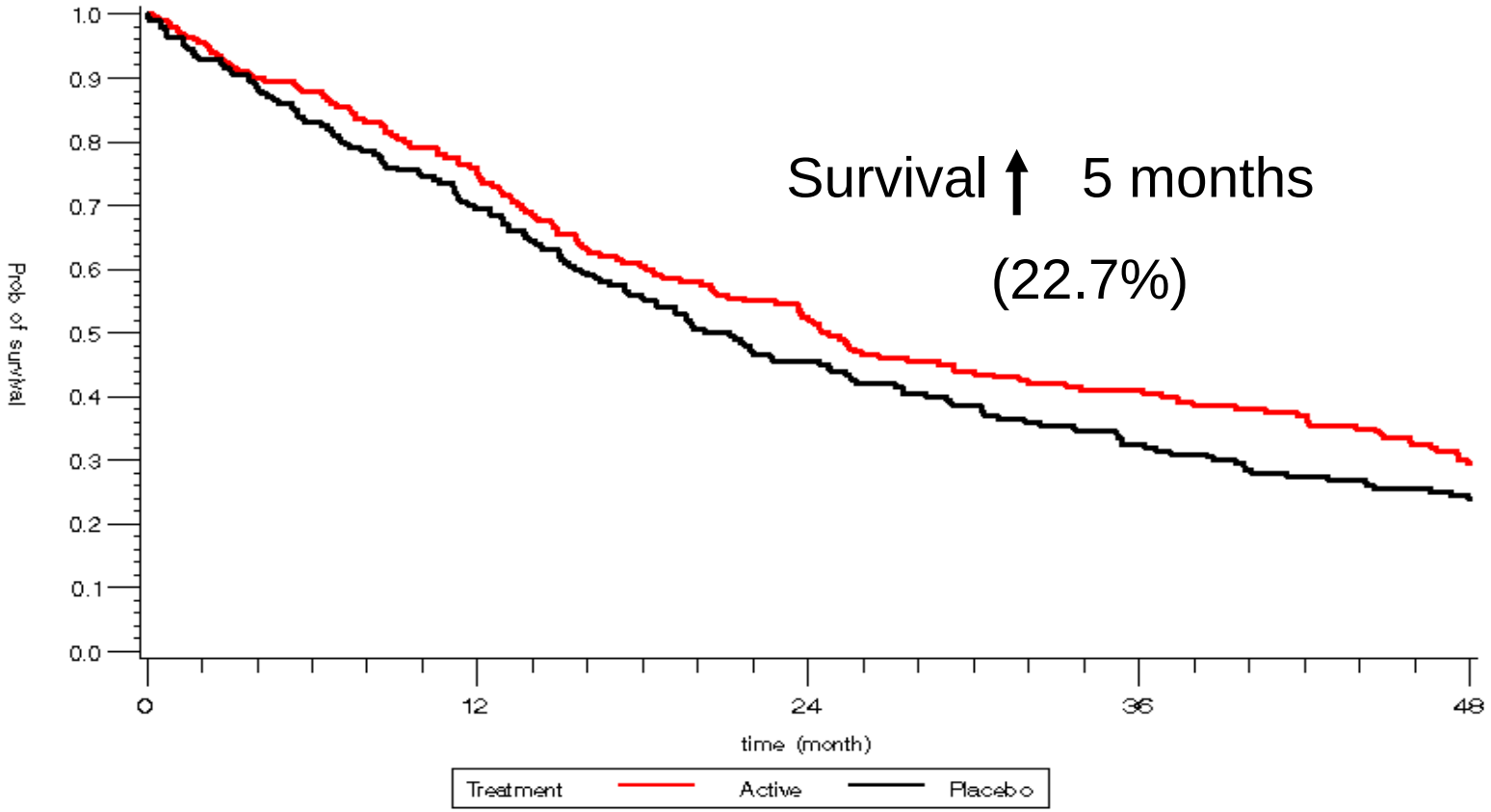



These genes have been implicated in cancer as inherited predispositions and/or as genes functionally altered in cancer cells. (*) targets of successful new drugs.

# Lessons from Herceptin

## Power of Patient Selection

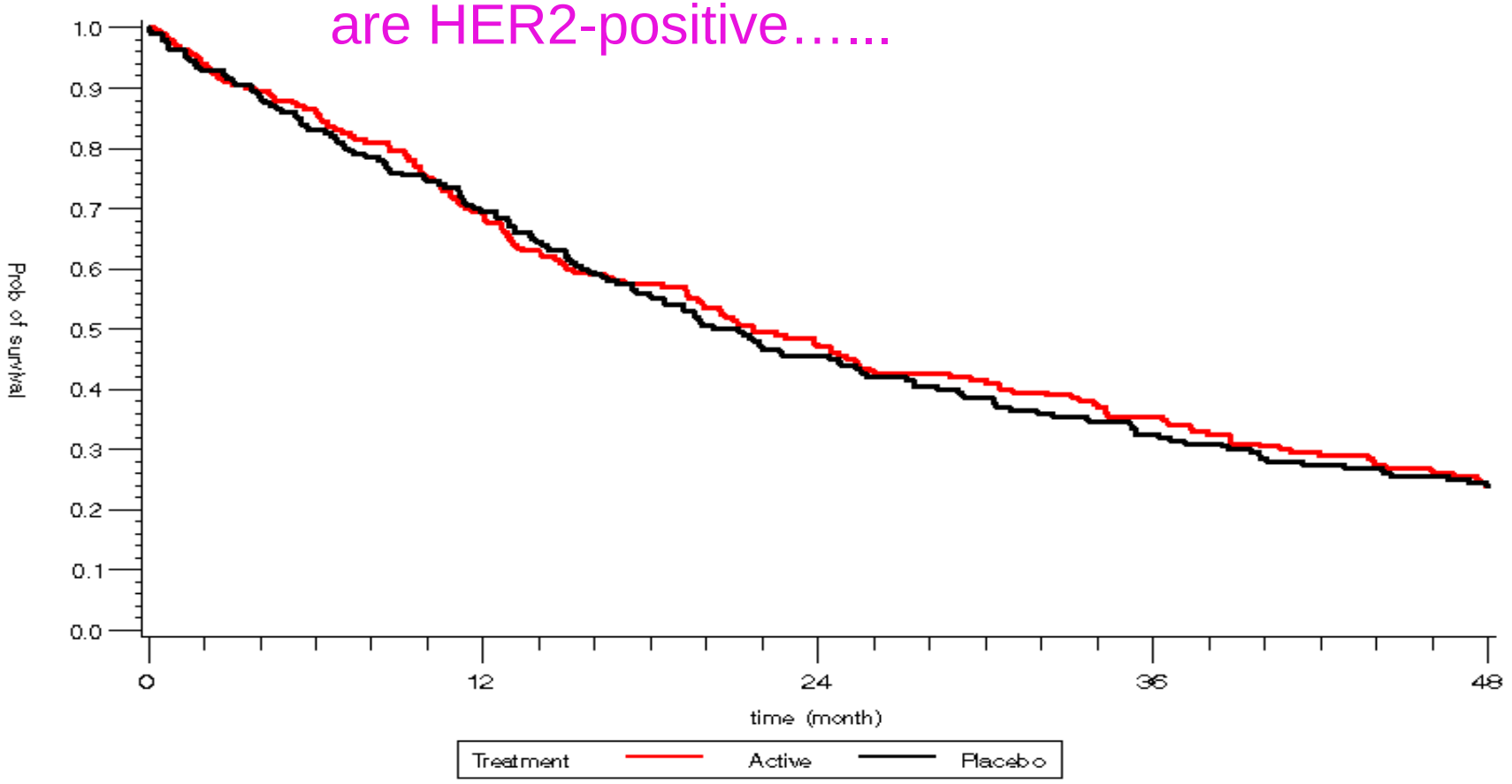
Randomized Phase III: HER2-positive patients selected before randomization



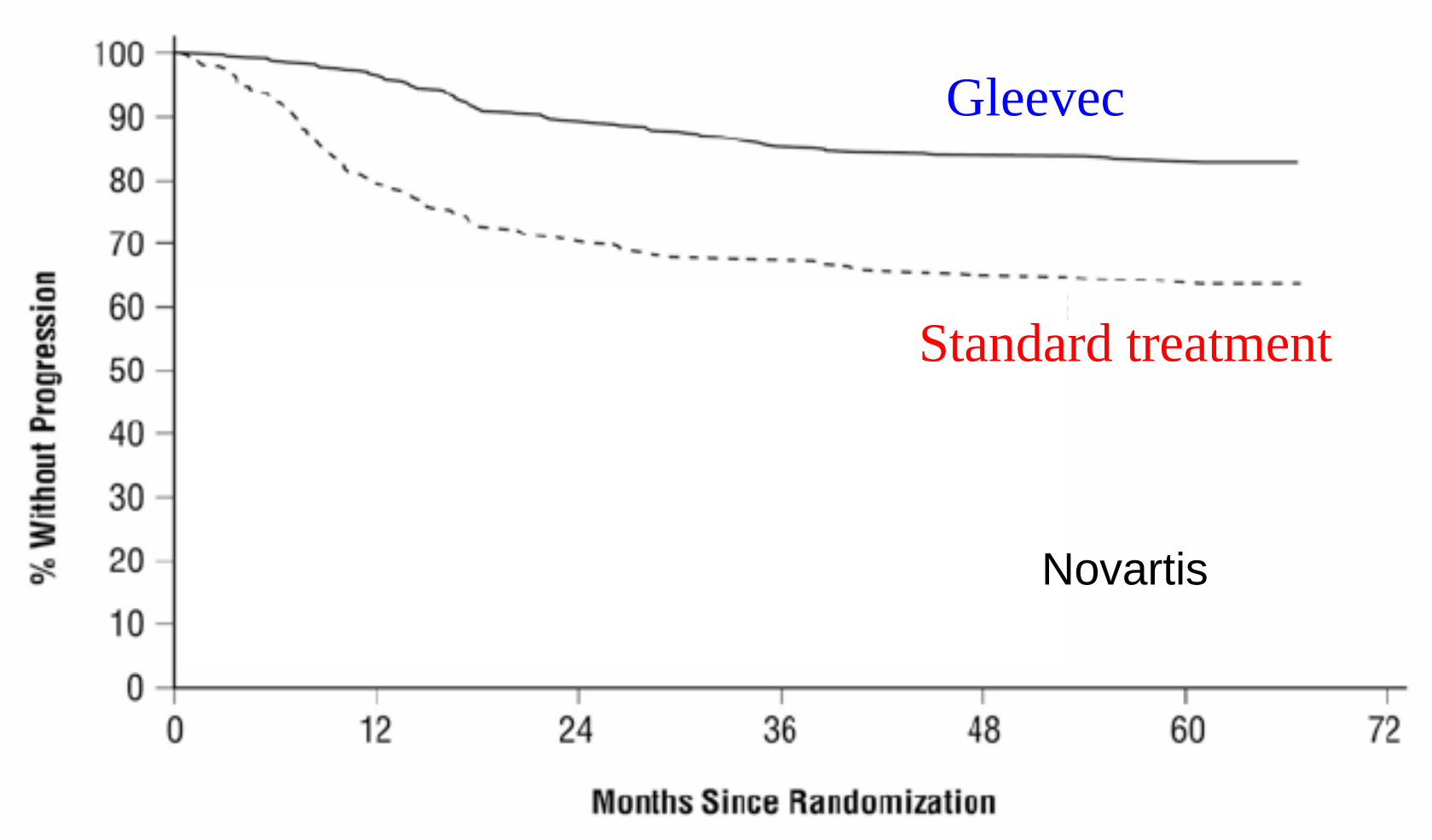
# Lessons from Herceptin

## Power of Patient Selection

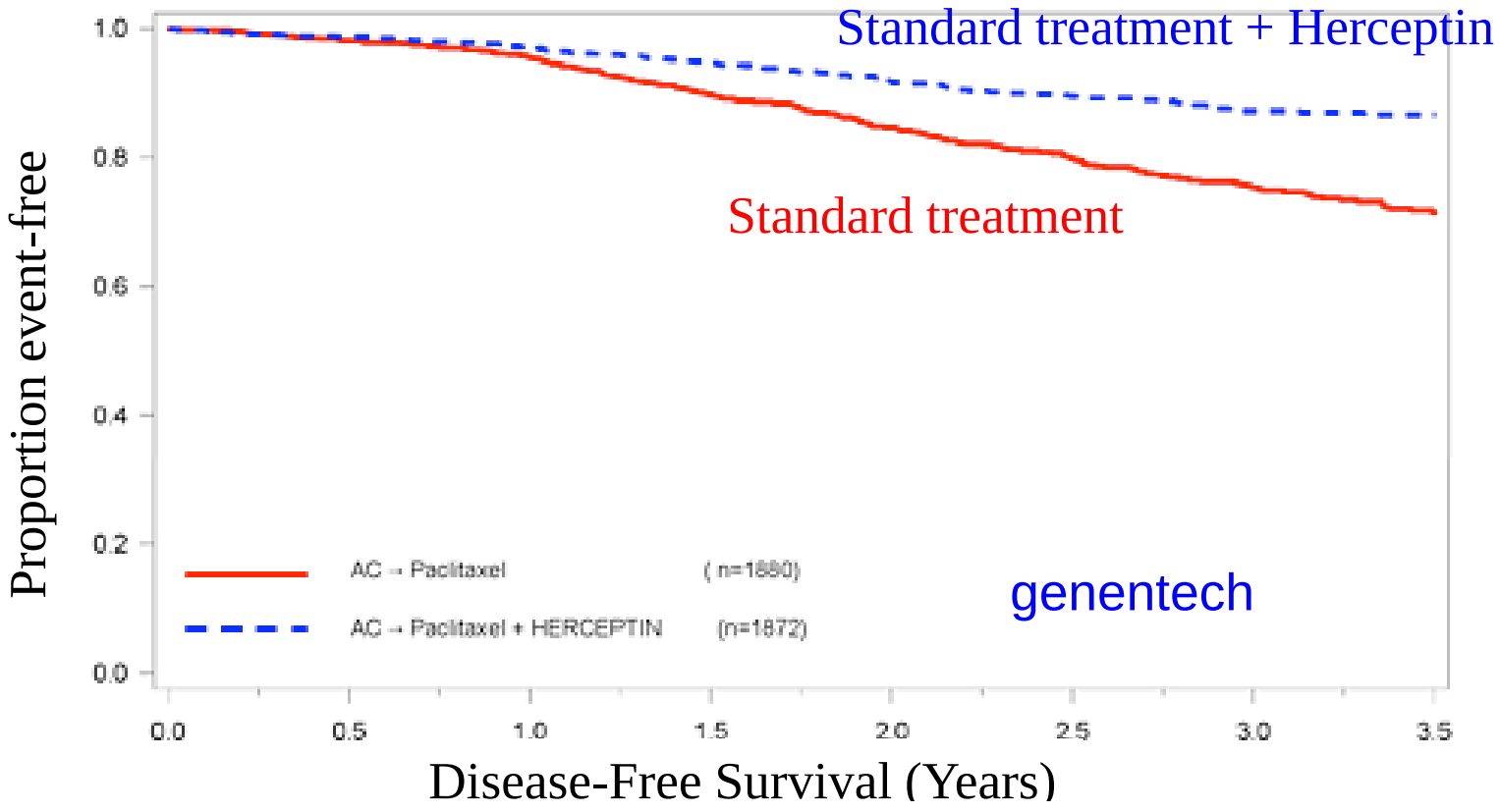
Randomized Phase III Trial: unselected patients [simulation] in which 25% of patients are HER2-positive…...



# Chronic Myelogenous Leukemia Patients Treated with Specific Antagonist (Gleevec) Directed Against the Product of the *ABL* Gene



# Breast Cancer Patients Treated with an Antibody Drug (Herceptin) Directed Against the Product of the HER2 Gene




Results of a randomized trial in which women were treated after removal of the primary tumor: the effect is about 2-fold improvement in survival, and highly significant statistically

# Clinical Applications of Genomic Information to Cancer

- *Better diagnosis*: definition of more biologically and clinically homogeneous cancer subtypes. Greater power to test efficacy in trials.
- *Earlier detection*: detection of secreted molecules, or even mutant DNA, in blood tests
- *New therapeutic targets*: identification of molecules expressed in tumors that can be aimed at.
  - membrane proteins as antibody therapy targets
    e.g. Her2/ERBB2 (Herceptin
  - receptor tyrosine kinases as small molecule targets
    e.g. specific antagonists of Abl or Kit (Gleevec)
- *Monitoring and predicting response*: finding the appropriate therapy, old or new, for each individual tumor

# Issues for the Future

- Personal genome as predictor of health: confronting the reality that we have no robust theory or understanding of the relationship between genotype and complex diseases (as opposed to single-gene Mendelian ones).

- How to reconcile interpretation of DNA sequence by doctors and patients (or somebody else– a statistical geneticist?) with the probabilistic nature of the connections between sequence and disease:

  -- The case of Huntington's (no therapeutic options today)
  -- The case of HNPCC (heightened surveillance, by colonoscopy, of obvious survival value)
  -- The case of HER2 amplification in breast tumors (an effective drug, trastuzumab (Herceptin) available)

# Issues for the Future

• Biology and medicine are being transformed into information sciences. It is increasingly difficult even to understand (let alone make) new discoveries (or diagnoses based on them) without a working command of the underlying mathematical, computational and statistical ideas that made them possible. But even today, most biologists and physicians are finish their education with no more than elementary calculus and no computer science at all.

• The great majority of human genes are not well understood. What we know is largely based on research on their orthologs in model systems (yeast, worms, mice). Yet basic science, the only proven path to understanding, is coming under severe funding pressure by “translational” work that seeks to apply what we don’t yet know.