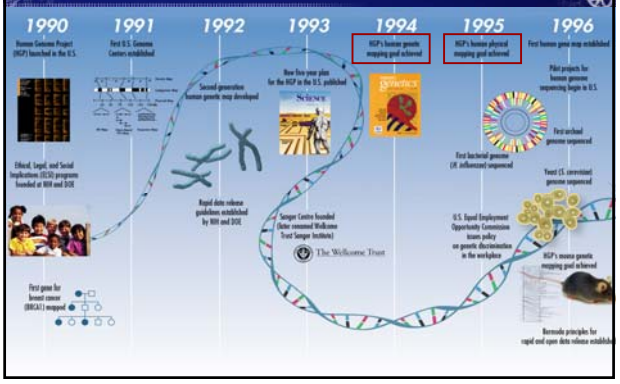
Human Genome Databases

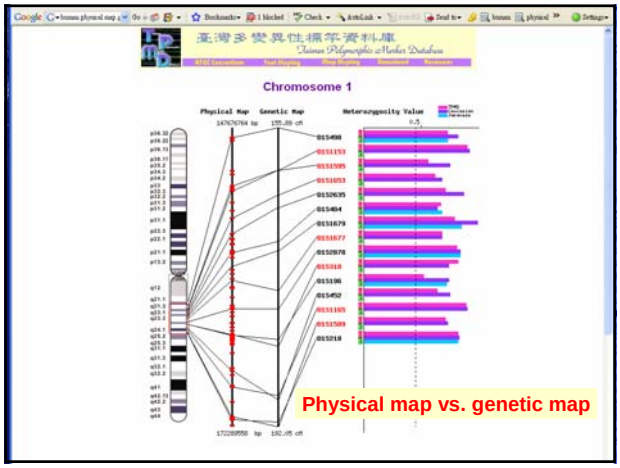
Molecular Medicine Research Center
Chang Gung University
Wen-Hui Su, Ph.D.

What lies inside human genome

- DNA
  - DNA sequence
    - Promoter, introns , exons, CpG islands .....
  - Polymorphism
    - STS/microsatellite markers
    - Single Nucleotide Polymorphism
  - Mutations
- RNA
  - Transcripts
    - Alternative splicing
  - miRNAs, rRNAs
- Protein
  - Proteins sequences
  - Secondary and 3-D structure
  - Domain and motifs
  - Protein families

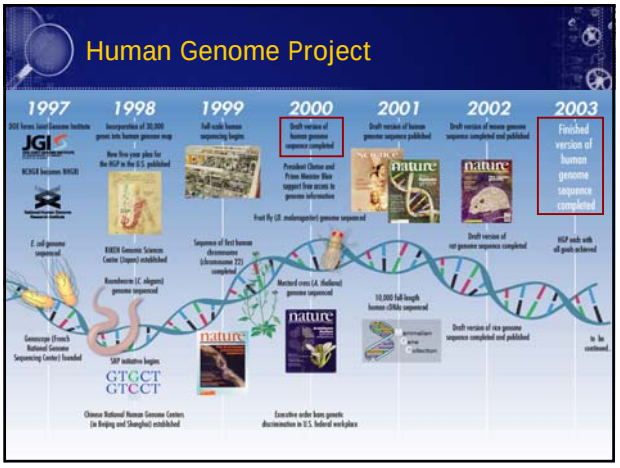
Human Genome Project





Physical map vs. genetic map

Human Genome Project



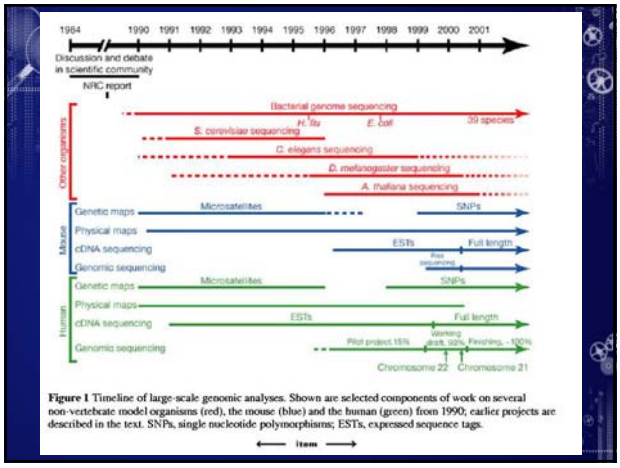
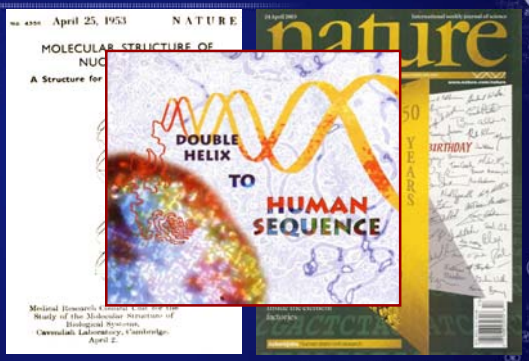


Figure 1 Timeline of large-scale genomic analyses. Shown are selected components of work on several non-vertebrate model organisms (red), the mouse (blue) and the human (green) from 1990; earlier projects are described in the text. SNPs, single nucleotide polymorphisms; ESTs, expressed sequence tags.

April, 1953 → April, 2003





Major Findings of the Human Genome Project

- Number of Genes
  - Number of genes only ~ 30,000
    - <2% of genome encode genes
  - Proteome is complex
    - 1 gene codes up to ~ 1000 proteins
- Genetic Variation
  - Comparisons between humans indicate that we are all 99.9% genetically identical
  - More than 1.4 million single nucleotide polymorphisms (SNPs)
    - Occur ~ 2000 base pairs apart, but density varies
    - Used to create haplotypes to differential between maternal / paternal chromosomes
    - May relate to disease susceptibility
    - Used for genome-wide association studies

How do we differ? – Let me count the ways

- Single nucleotide polymorphisms
  - 1 every few hundred bp, mutation rate* $\approx 10^{-9}$
- Short indels (=insertion/deletion)
  - 1 every few kb, mutation rate v. variable
- Microsatellite (STR) repeat number
  - 1 every few kb, mutation rate $\leq 10^{-3}$
- Minisatellites
  - 1 every few kb, mutation rate $\leq 10^{-1}$
- Repeated genes
  - rRNA, histones
- Large inversions, deletions
  - Rare, e.g. Y chromosome

TGCATTGCGTAGGC
TGCATTCCGTAGGC

TGCATT---TAGGC
TGCATTCCGTAGGC

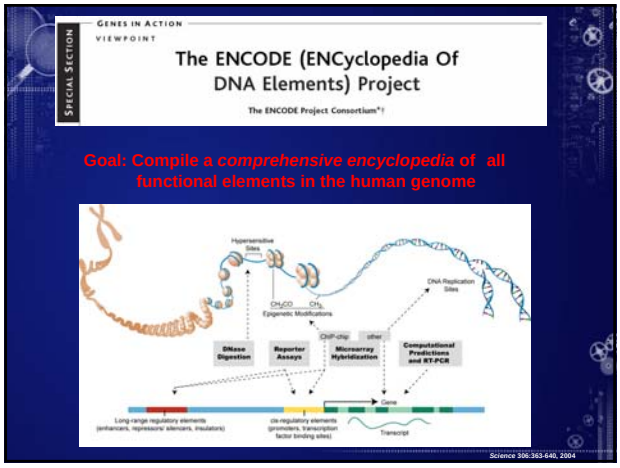
TGCTCATCATCATCAGC
TGCTCATCA------GC



*per generation

Beyond Human Genome Project

- Gene annotation and expression
  - Microarray, Proteomic...
  - EST, SAGE...
- Functional elements
  - ENCODE project
  - Comparative genomic approach
- Polymorphism
  - SNP (SNP Consortium)
  - HapMap (HapMap Project)
- Mutation
  - Cancer Genome Project (resequencing)



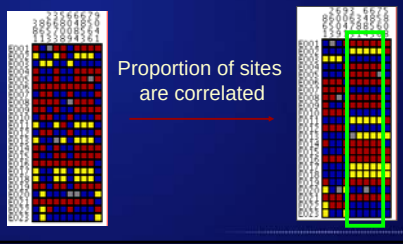
Beyond Human Genome Project

- Gene annotation and expression
  - Microarray, Proteomic...
  - EST, SAGE...
- Intergenic region
  - ENCODE project
  - Comparative genomic approach
- Polymorphism
  - SNP (SNP Consortium)
  - HapMap (HapMap Project)
- Mutation
  - Cancer Genome Project (resequencing)

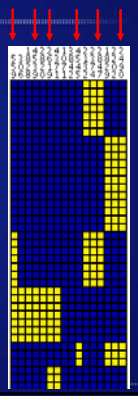
SNPs Are Correlated (aka linkage disequilibrium)

"the nonindependence of alleles at different sites." Pritchard and Przeworski 2001

Genotype at one site can predict genotype at another site

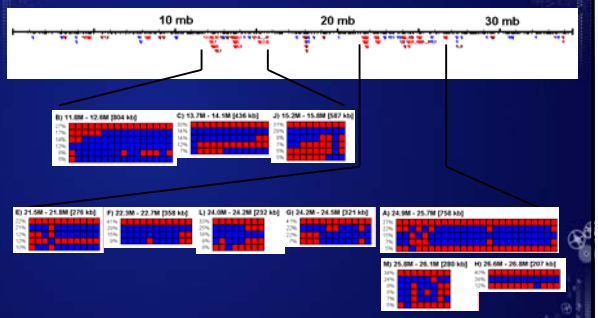


Haplotype Blocks and tagSNPs



LTA:
16 SNPs (MAF >10%)
6 "common" haplotypes

Haplotype blocks on chromosome 22



Haplotyp



Goals:
- Define patterns
- Guide selectior
- Public release
- The DNA samp

70 people:

30 t
45 t
45 t
30 (

y).

Beyond Human Genome Project

- Gene annotation and expression
  - Microarray, Proteomic...
  - EST, SAGE...
- Intergenic region
  - ENCODE project
  - Comparative genomic approach
- Polymorphism
  - SNP (SNP Consortium)
  - HapMap (HapMap Project)
- Mutation
  - Cancer Genome Project (resequencing)

What's the difference between a mutation and a polymorphism? Frequency!

A frequency value of 1% of the polymorphic allele is usually taken as a threshold between mutation and polymorphism.

Major Human Genome Databases

- Integrated databases
  - NCBI
    - Entrez Gene
    - Map Viewer
    - PubMed, OMIM
    - UniGene, UniSTS, dbSNP
  - Ensembl
    - Comparative genomic result
    - BioMart data extraction system
  - UCSC Genome Browser
    - Display different genome assembly versions
    - Golden Path
    - User defined tracks

Homework

- List chromosome position, gene symbol and description of human genes located in chromosome (**?**) with transmembrane domain and involve in cell communication (GO:0007154) in a CSV file.
- Mail to whsu@mail.cgu.edu.tw before **June 30 noon** with title "Bioinformatic homework".
- Please put your name and "學號" in the mail content.