Attribution Key

for more information see: http://open.umich.edu/wiki/AttributionPolicy

Use + Share + Adapt

{ Content the copyright holder, author, or law permits you to use, share and adapt. }

PD-GOV **Public Domain – Government**: Works that are produced by the U.S. Government. (17 USC § 105)

PD-EXP **Public Domain – Expired**: Works that are no longer protected due to an expired copyright term.

PD-SELF **Public Domain – Self Dedicated**: Works that a copyright holder has dedicated to the public domain.

(cc) ZERO **Creative Commons – Zero Waiver**

(cc) BY **Creative Commons – Attribution License**

(cc) BY-SA **Creative Commons – Attribution Share Alike License**

(cc) BY-NC **Creative Commons – Attribution Noncommercial License**

(cc) BY-NC-SA **Creative Commons – Attribution Noncommercial Share Alike License**

GNU-FDL **GNU – Free Documentation License**

Make Your Own Assessment

{ Content Open.Michigan believes can be used, shared, and adapted because it is ineligible for copyright. }

PD-INEL **Public Domain – Ineligible**: Works that are ineligible for copyright protection in the U.S. (17 USC § 102(b)) *laws in your jurisdiction may differ

{ Content Open.Michigan has used under a Fair Use determination. }

© FAIR USE **Fair Use**: Use of works that is determined to be Fair consistent with the U.S. Copyright Act. (17 USC § 107) *laws in your jurisdiction may differ

Our determination **DOES NOT** mean that all uses of this 3rd-party content are Fair Uses and we **DO NOT** guarantee that your use of the content is Fair.

To use this content you should **do your own independent analysis** to determine whether or not your use will be Fair.

The Human Genome I

M1 Patients and Populations

David Ginsburg, MD

Fall 2012

University of Michigan
Medical School

Relationships with Industry

UMMS faculty often interact with pharmaceutical, device, and biotechnology companies to improve patient care, and develop new therapies. UMMS faculty disclose these relationships in order to promote an ethical & transparent culture in research, clinical care, and teaching.

- I am a member of the Board of Directors for Shire plc.
- I am a member of the Scientific Advisory Boards for Portola Pharmaceuticals and Catalyst Biosciences.
- I benefit from license/patent royalty payments to Boston Children’s Hospital (VWF) and the University of Michigan (ADAMTS13).

Disclosure required by the UMMS Policy on Faculty Disclosure of Industry Relationships to Students and Trainees.

Learning Objectives

UNDERSTAND:

- The basic anatomy of the human genome [eg. 3 X10^9 bp (haploid genome); 1-2% coding sequence (~20,000 genes); types and extent of DNA sequence variation].
- Recombination and how it allows genes to be mapped
- Genetic data for a pedigree, assigning phase, defining haplotypes
- Linkage: Distinction between a linked marker and the disease causing mutation itself
- Linkage disequilibrium and haplotype blocks
- Genome wide association studies (GWAS) to identify gene variants contributing to complex diseases/traits
- The implications of GWAS findings for clinical care and "Personalized Medicine"
- The implications of "Next-Gen" sequencing for future clinical medicine

DNA Sequence Variation

- DNA Sequence Variation:
  - Human to human: ~0.1% (1:1000 bp)
    - Human genome = $3X10^9$ bp X 0.1% =~$3X10^6$ DNA common variants
  - Human to chimp: ~1-2%
  - More common in “junk” DNA: introns, intergenic regions

- **poly·mor·phism**
  Pronunciation: "päl-i-'mor-"fiz-&m
  Function: *noun*
  : the quality or state of existing in or assuming different forms: as **a (1) :** existence of a species in several forms independent of the variations of sex **(2) :** existence of a gene in several allelic forms (3) : existence of a molecule (as an enzyme) in several forms in a single species

Polymorphisms and Mutations

- Genetic polymorphism:
  - Common variation in the population:
    - Phenotype (eye color, height, etc)
    - genotype (DNA sequence polymorphism)
  - Frequency of minor allele(s) ≥ 1%

- DNA (and amino acid) sequence variation:
  - Most common allele ≤ 0.99 = polymorphism (minor allele(s) ≥ 1%)
  - Variant alleles < 0.01 = rare variant

- Mutation-- any change in DNA sequence
  - Silent vs. amino acid substitution vs. other
  - neutral vs. disease-causing
  - $1X10^{-8}$/bp/generation (~70 new mutations/individual)

- balanced polymorphism= disease + polymorphism

- Common but incorrect usage:
  - "mutation vs. polymorphism"

All DNA sequence variation arises via mutation of an ancestral sequence

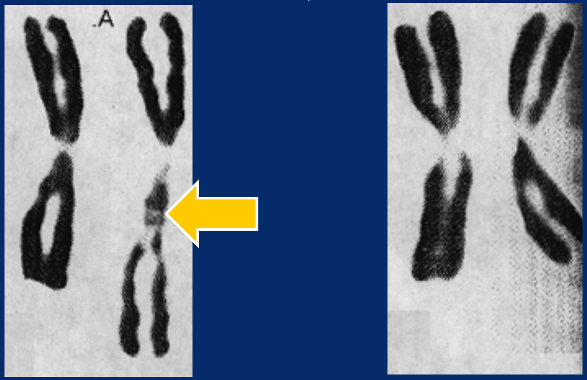
	< 1%	**≥ 1%**
“Normal”	Rare variant or “private” polymorphism	polymorphism
“Disease”	Disease mutation	*Example: Factor V Leiden (thrombosis)* *5% allele frequency*

Common but incorrect usage:

“a disease-causing mutation” **OR** *“a polymorphism”*

Donahue, 1968

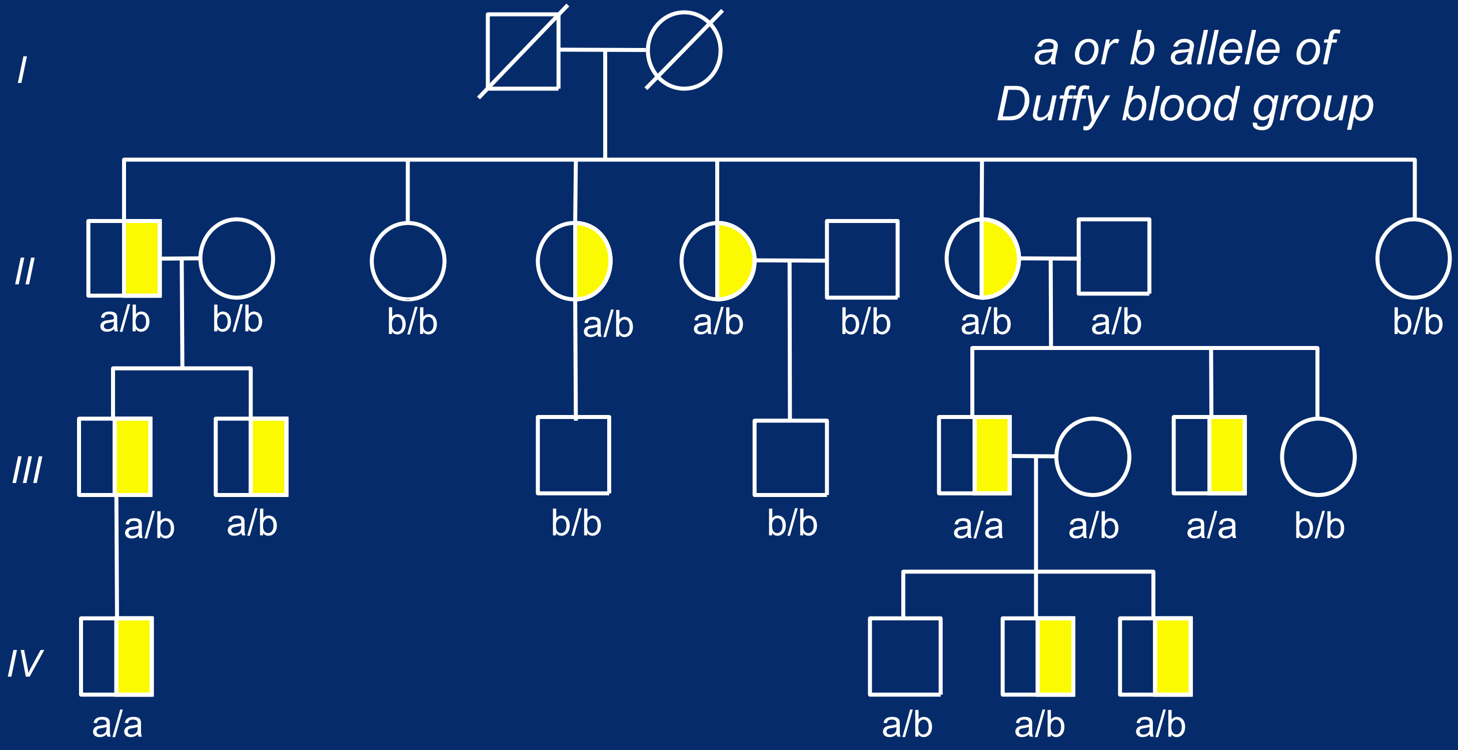
◨ Heteromorphism of chromosome 1 (one copy)



□ 2 normal copies of chromosome 1

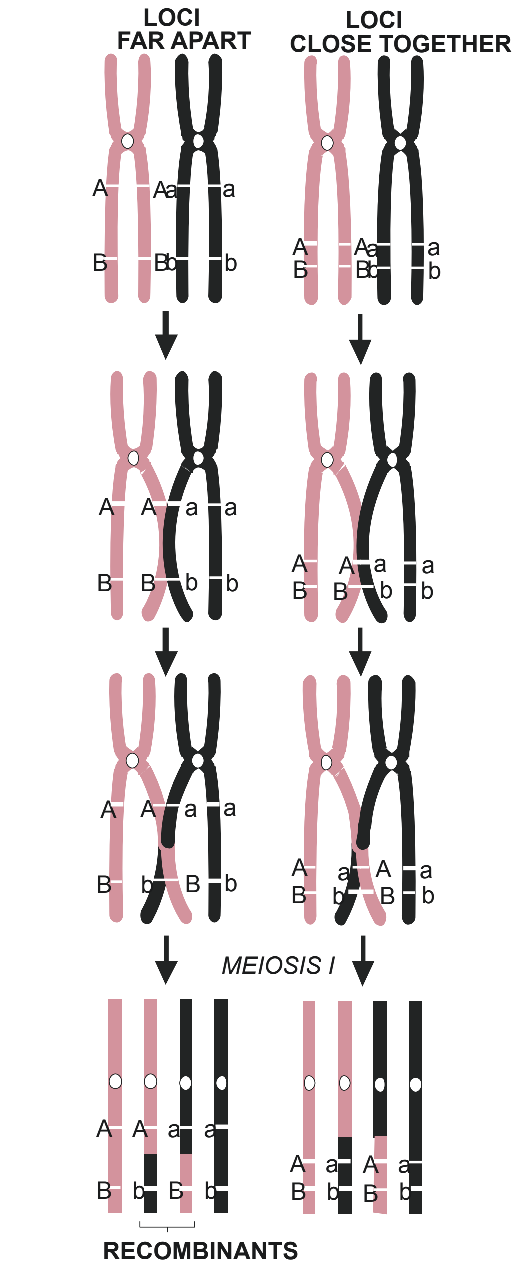
B.

a or b allele of Duffy blood group



Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.1

Key Concepts: Linkage and Recombination

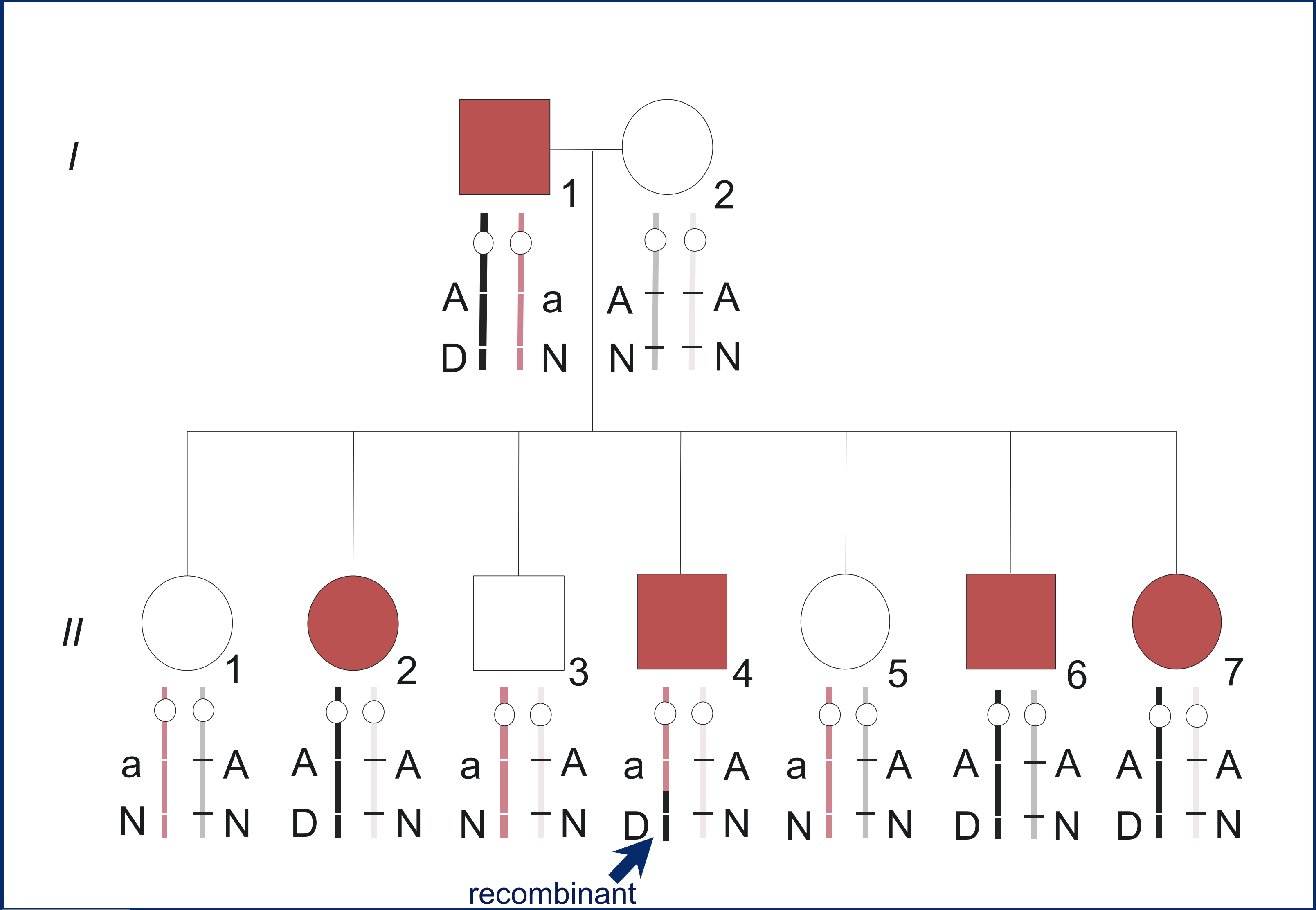


Linkage: A/a and B/b tend to be inherited together

the A and B loci are linked.

Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.2

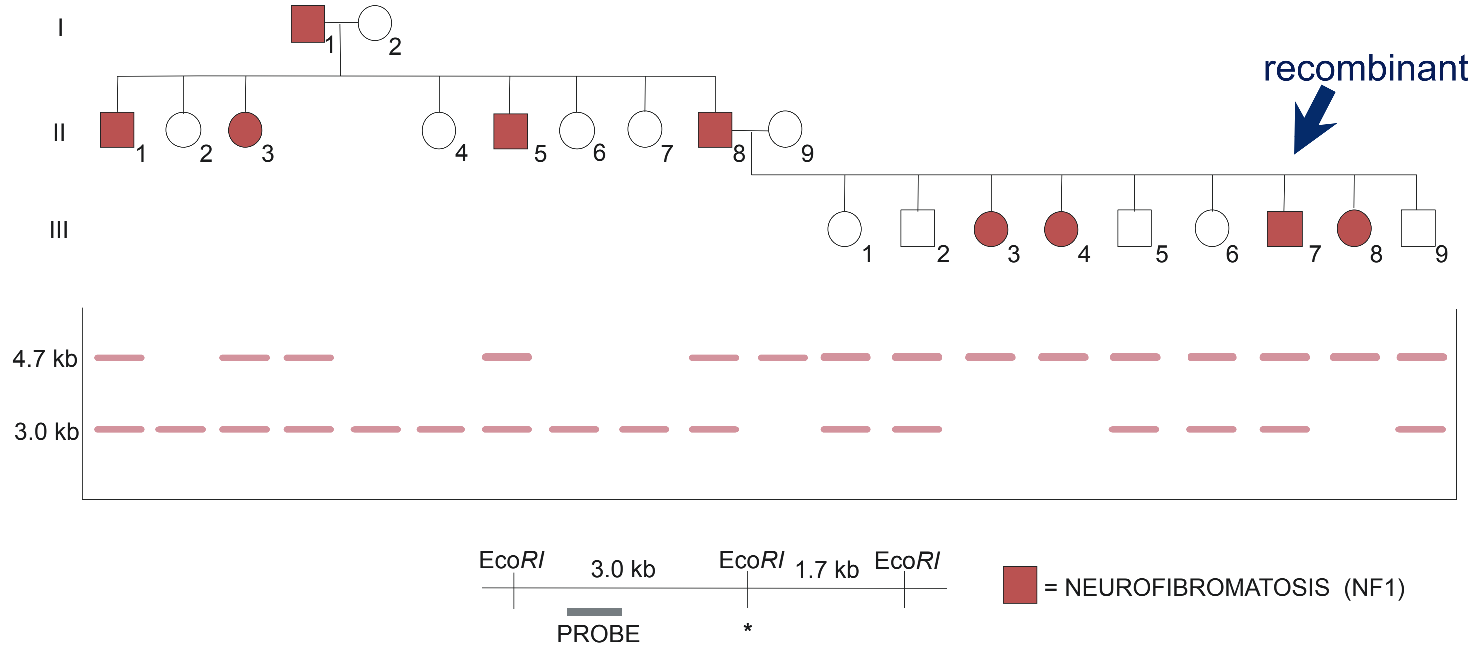
Linkage between Marker A/a and Disease D



Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.3

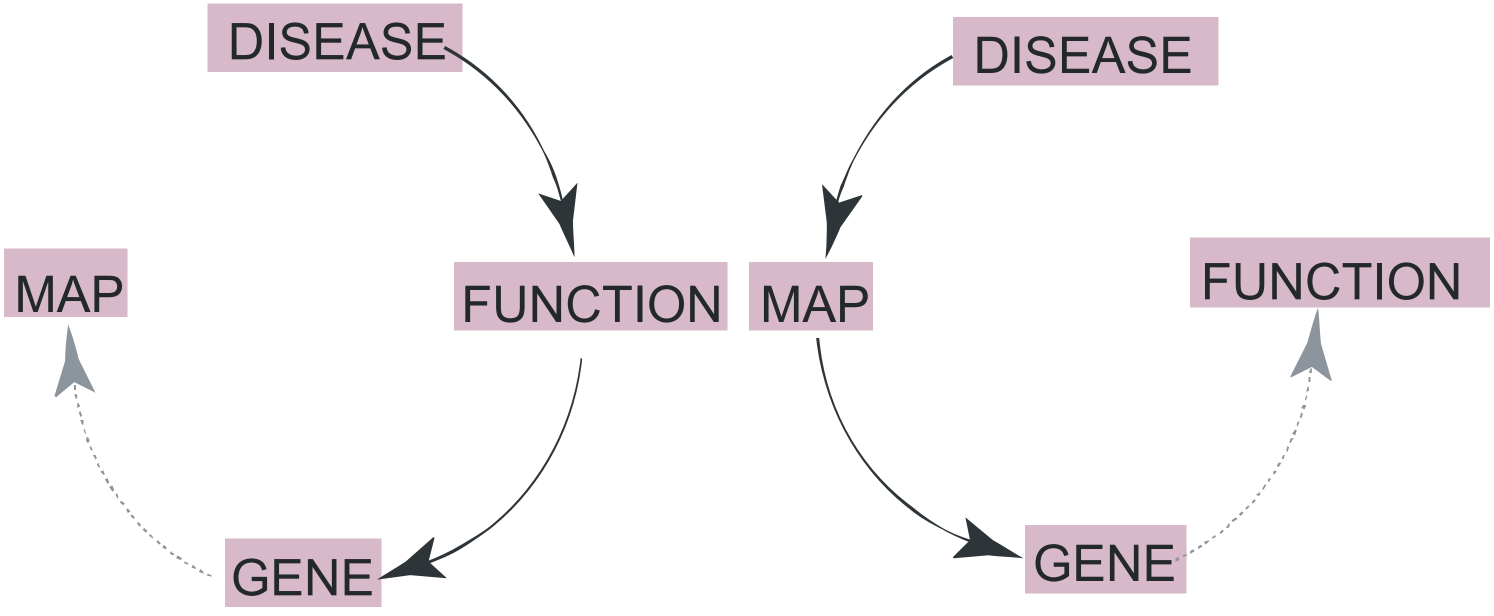
Marker= A or a
Disease allele = D
Normal allele = N

Linkage between NF and RFLP marker



■ = NEUROFIBROMATOSIS (NF1)

PD-INEL Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.5

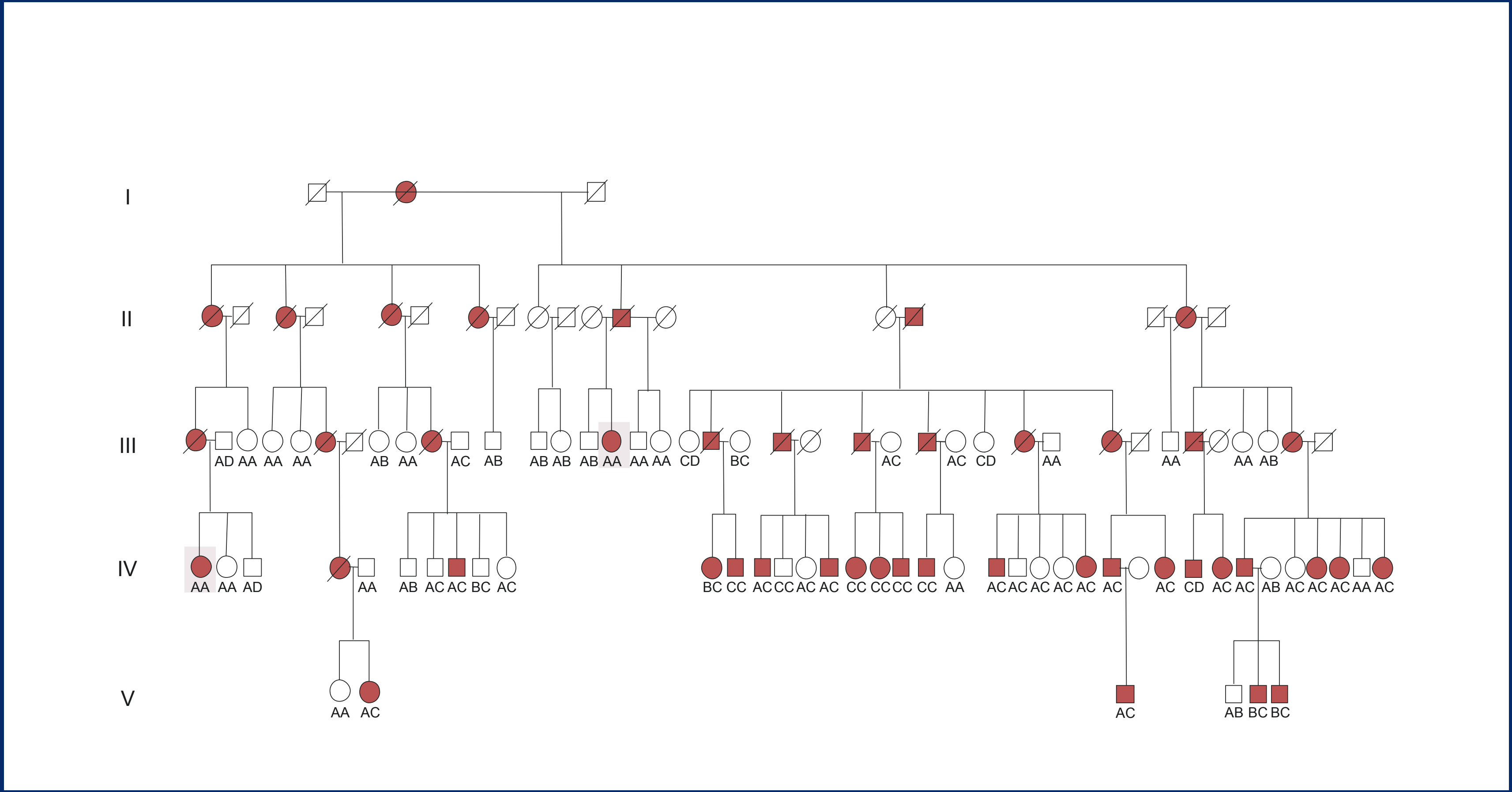


FUNCTIONAL CLONING

POSITIONAL CLONING

Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.15

HD linked to C allele: Two recombinant s (III13, IV1)


I
II
III
IV
V
AD AA AA AA AB AA AC AB AB AB AB AA AA AA CD BC AC AC CD AA AA AA AB
AA AA AD AA AB AC AC BC AC BC CC AC CC AC AC CC CC CC CC AA AC AC AC AC AC AC AC CD AC AC AB AC AC AC AA AC
AA AC AC AB BC BC

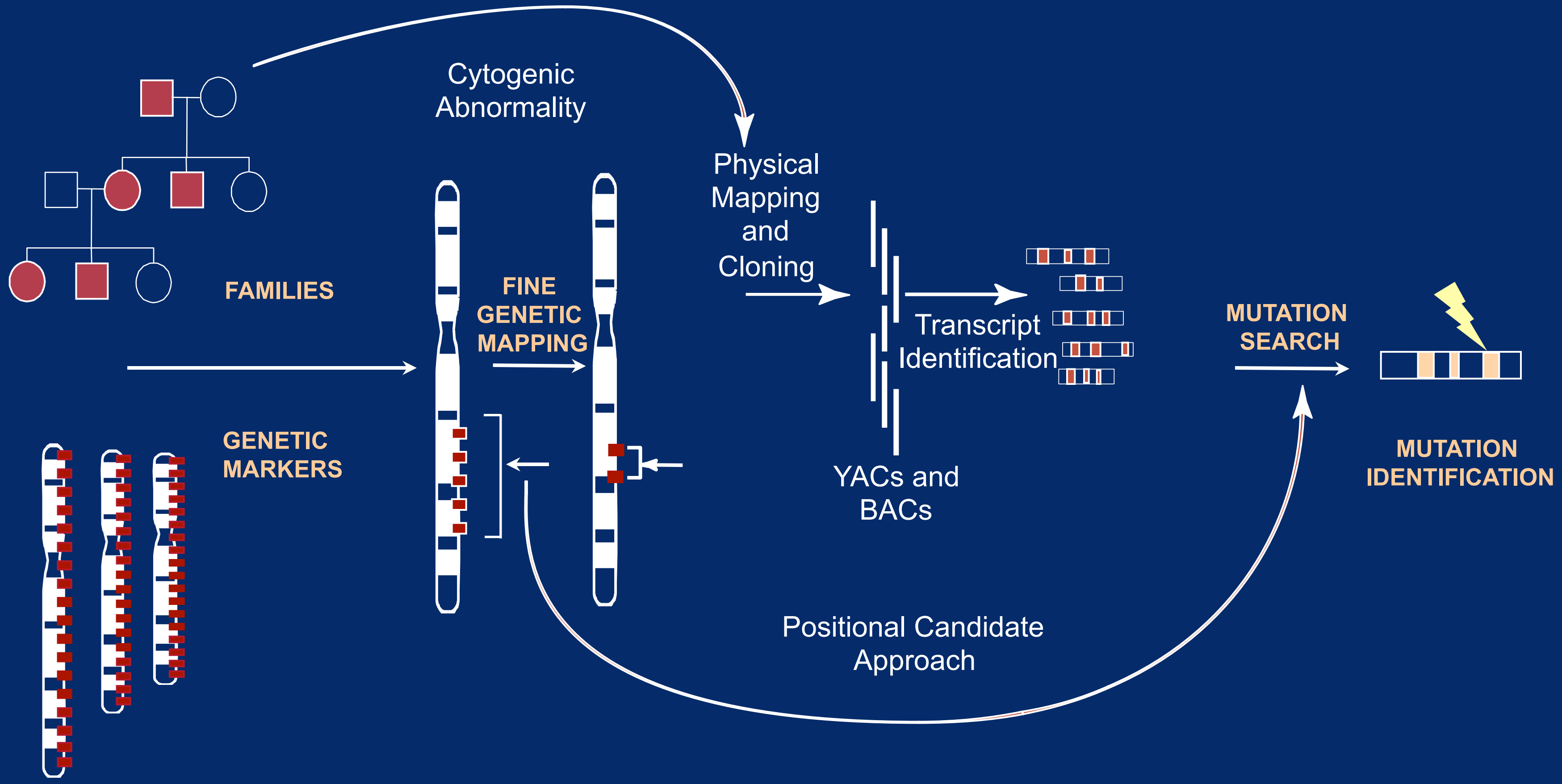
Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.26

Gusella, et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature 306:234-238, 1983.*

The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell 72:971-983, 1993.*

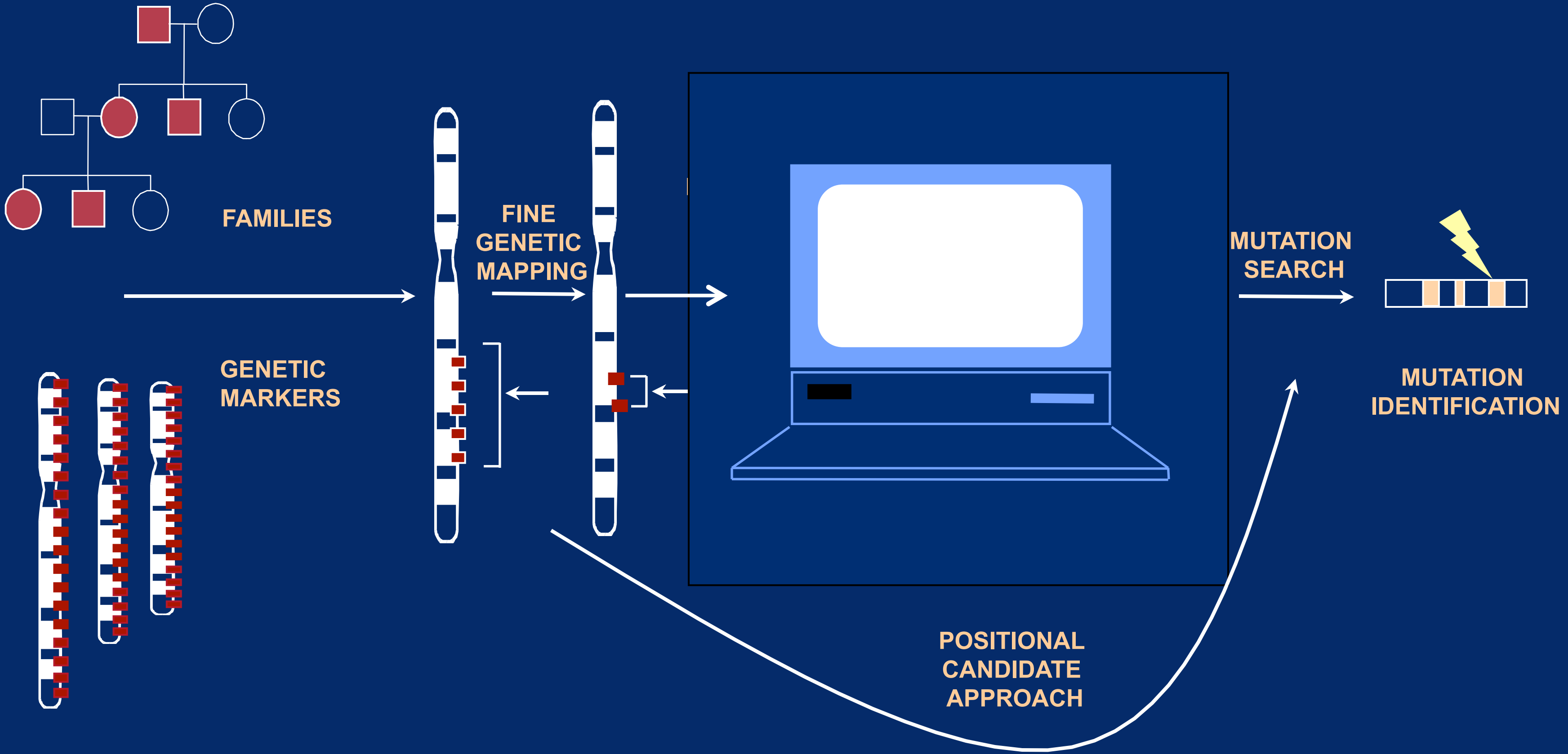
Textbook: Figure 9.26

Positional Cloning

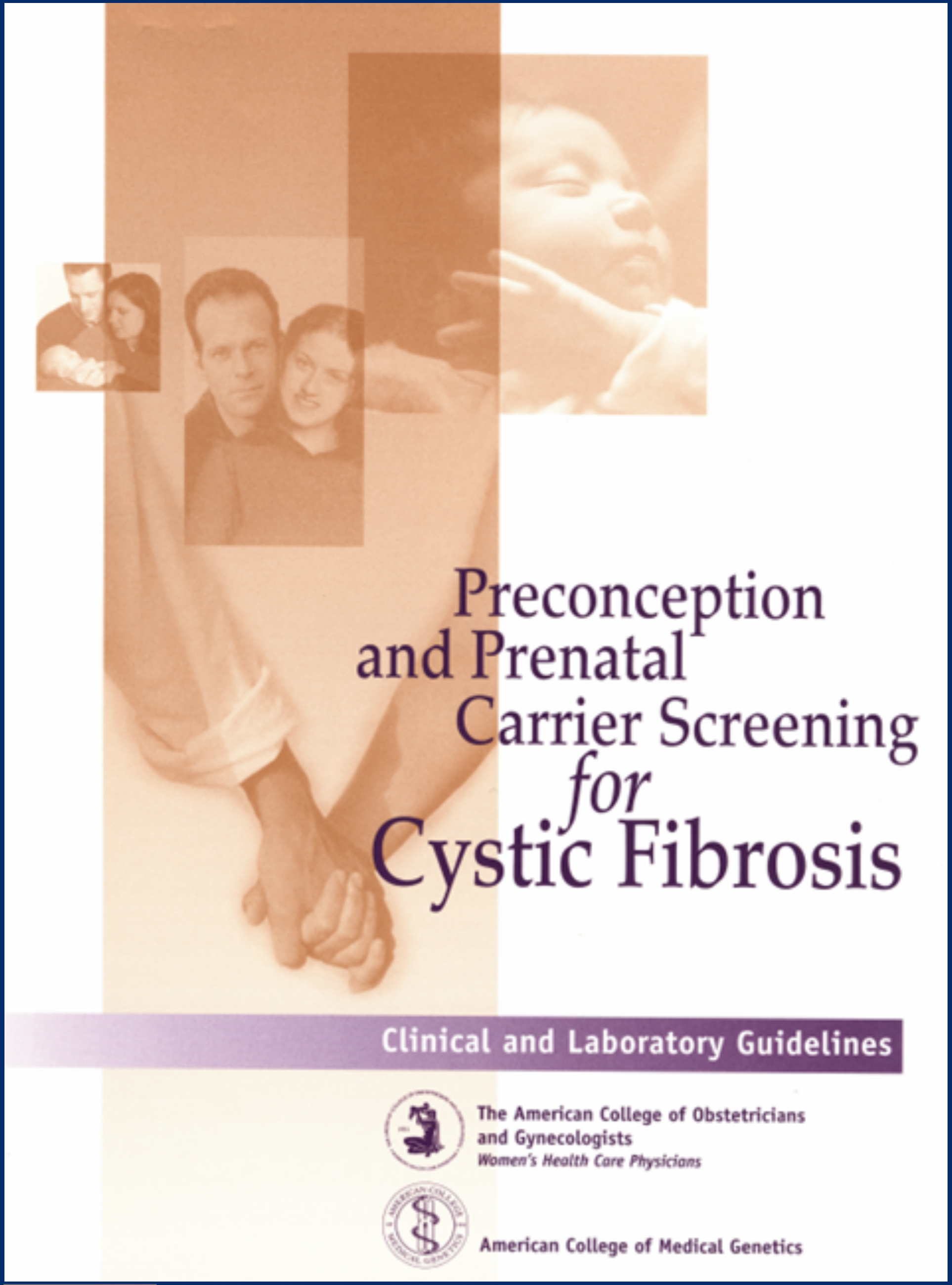


Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.31

Positional Cloning



Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.15



Preconception and Prenatal Carrier Screening *for* Cystic Fibrosis

Clinical and Laboratory Guidelines

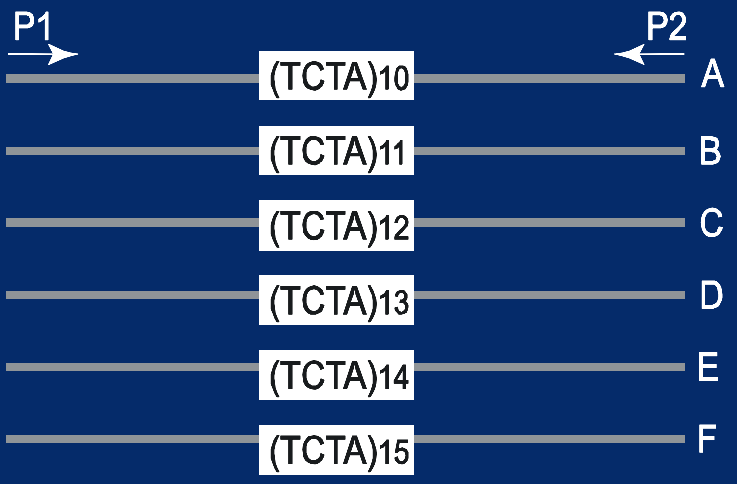
The American College of Obstetricians and Gynecologists
Women's Health Care Physicians

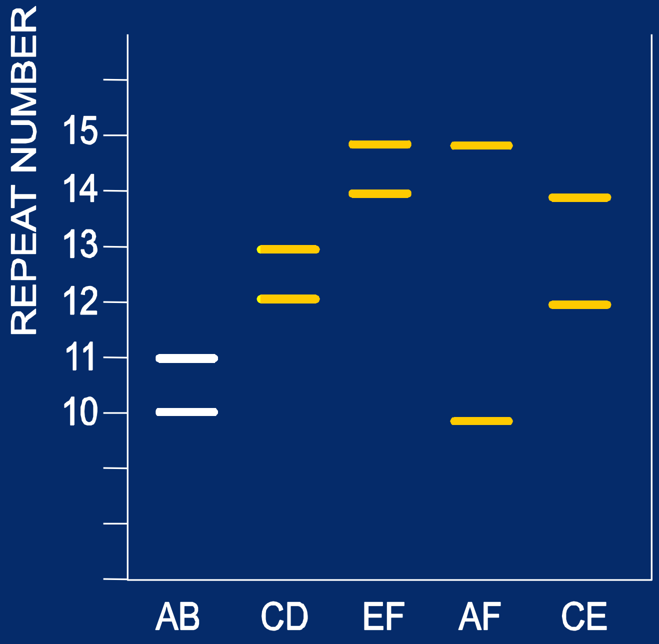
American College of Medical Genetics

Types of DNA Sequence Variation

- RFLP: **R**estriction **F**ragment **L**ength **P**olymorphism
- VNTR: **V**ariable **N**umber of **T**andem **R**epeats
  - or minisatellite
  - ~10-100 bp core unit
- SSR : **S**imple **S**equence **R**epeat
  - or STR (simple tandem repeat)
  - or microsatellite
  - ~1-5 bp core unit
- SNP: **S**ingle **N**ucleotide **P**olymorphism
  - Commonly used to also include rare variants
- Insertions or deletions
  - INDEL – small (few nucleotides) insertion or deletion
- Rearrangement (inversion, duplication, complex rearrangement)
- CNV: **C**opy **N**umber **V**ariation

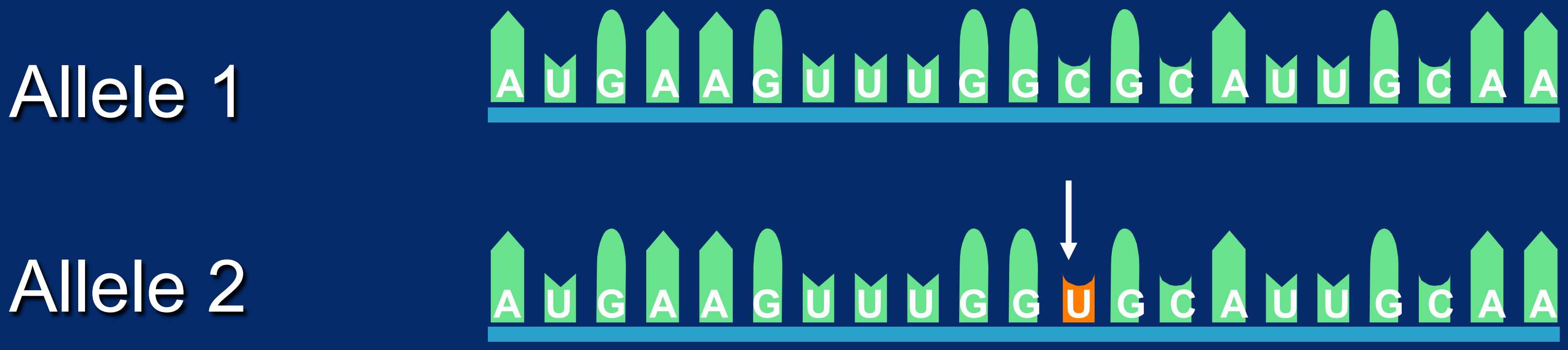
STR





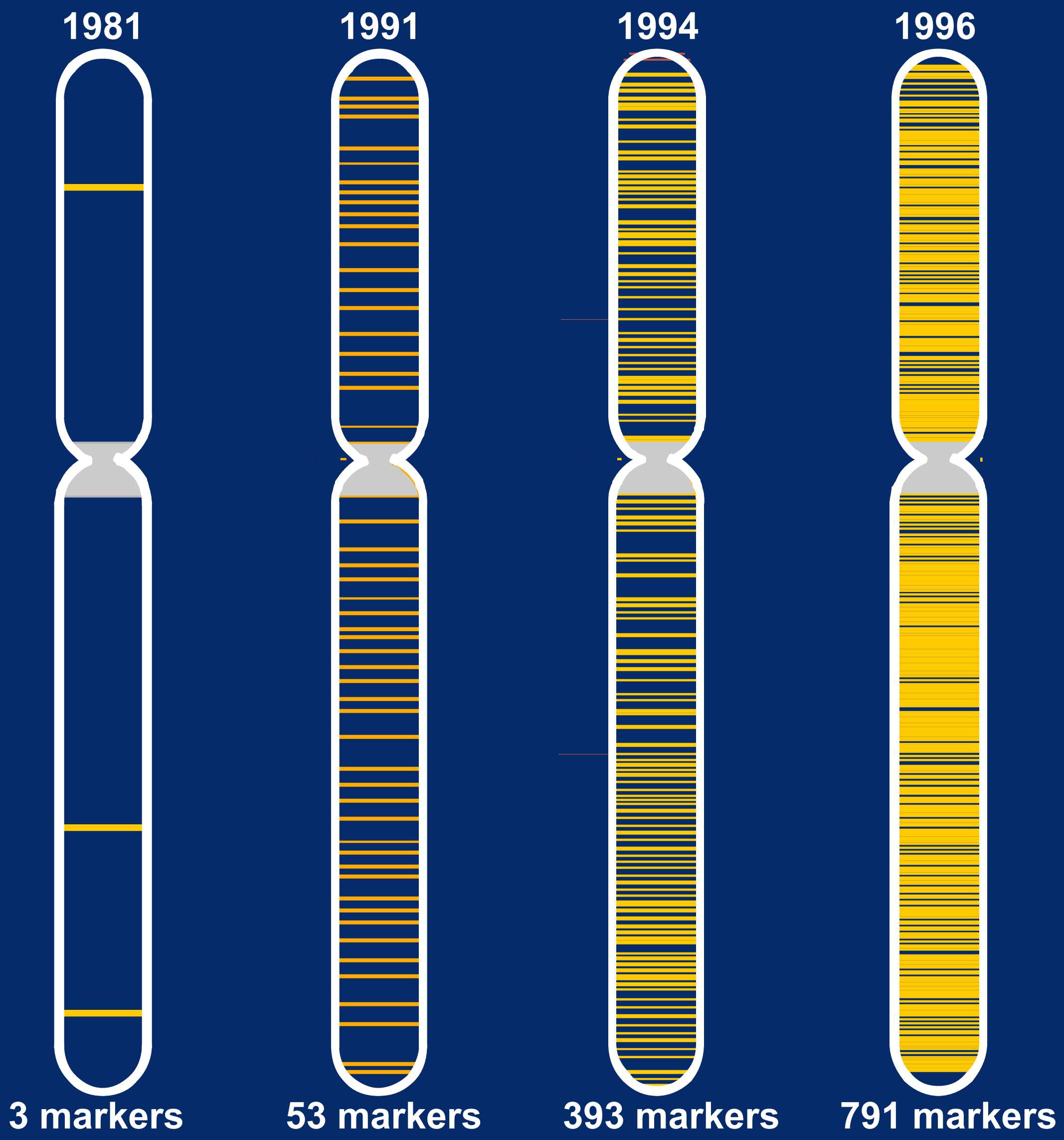
PD-INEL Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 5.22

SNP


Allele 1
A U G A A G U U U G G C G C A U U G C A A
Allele 2
A U G A A G U U U G G U G C A U U G C A A

- Most are "silent"
- Intragenic
- Promoters and other regulatory sequences
- Introns
- Exons
  - 5' and 3' untranslated regions
  - Coding sequence (~1-2% of genome)

Human Chromosome 4



2010

- 23,653,737 total human entries in dbSNP

http://www.ncbi.nlm.nih.gov/projects/SNP/

- Chromosome 4
  - 4,311,728 SNPs
- ~1M SNP chip commercially available

PD-INEL Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 10.3

15 February 2001

nature

www.nature.com

the human genome

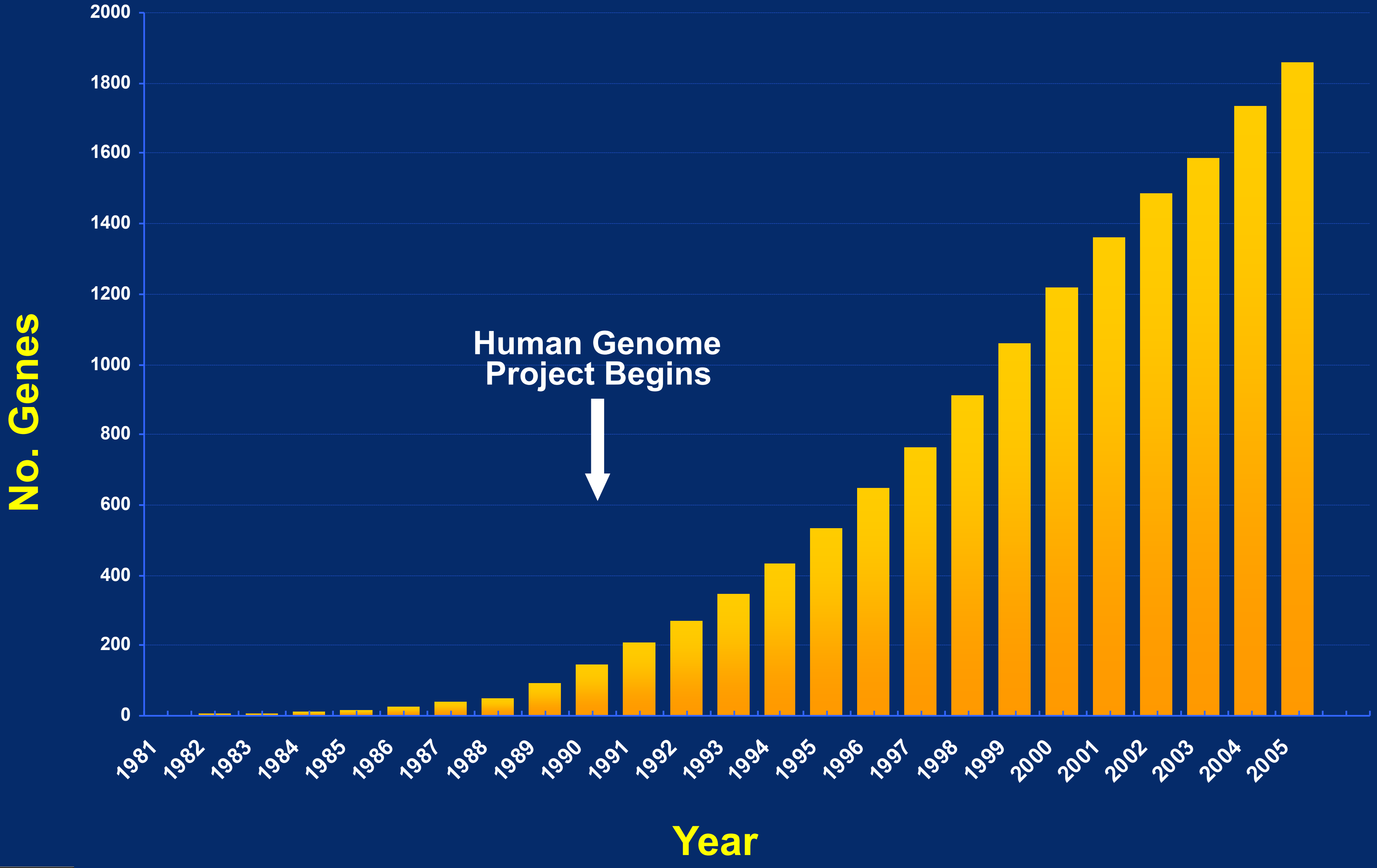
Nuclear fission
Five-dimensional energy landscapes

Seafloor spreading
The view from under the Arctic icepack

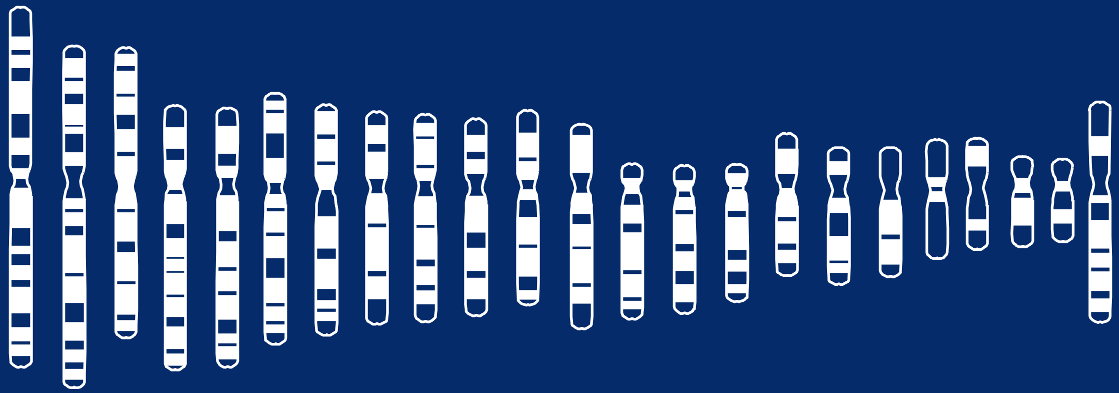
Career prospects
Sequence creates new opportunities

naturejobs
genomics special

Genes Identified: Monogenic Diseases

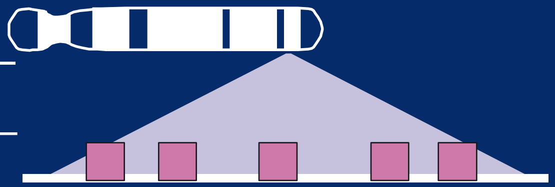


Online Mendelian Inheritance in Man



Haploid Human Genome 3 X 10^9 bp, ~20,000 genes

1 Chromosome
~1300 genes



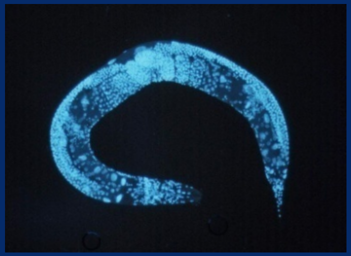
Single Gene
~1.5 Kb (Globin to
2 X 10^6 bp (Dystrophin)

H. Influenzae
~1700 genes

S. Cerevisiae
~6250 genes

D. Melanogaster
~14000 genes

CC BY-SA Andre Karwath (wikipedia)



C. Elegans
~18500 genes

PD-GOV U.S. Federal Government (wikimedia)

Genomes

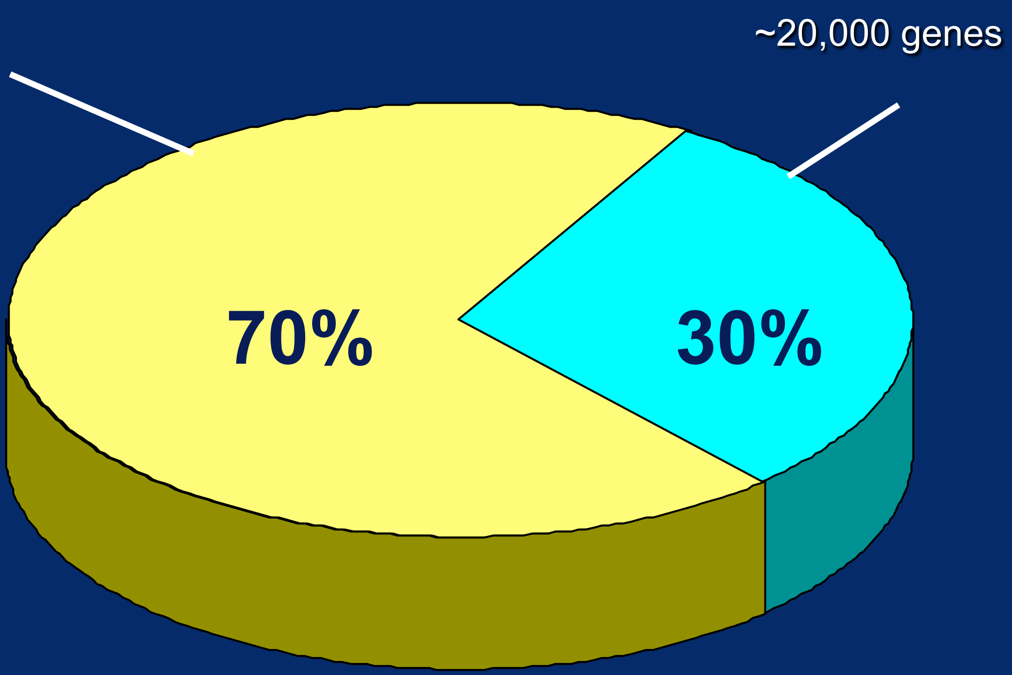
- *Complete human genome (~100 individual genomes, 1000 genomes in progress)*
- *Complete genomes of >6500 other species*
- Plants (arabidopsis, oat, soybean, barley, wheat, rice, tomato, corn) …
- Yeast, fly, worm, human, mouse, rat, zebrafish, mosquito, malaria, ciona …
- Cow, pig, frog, chimp, gorilla, dog, chicken, cat, bee …

The Human Genome

23 pairs of chromosomes made of 3 billion base pairs

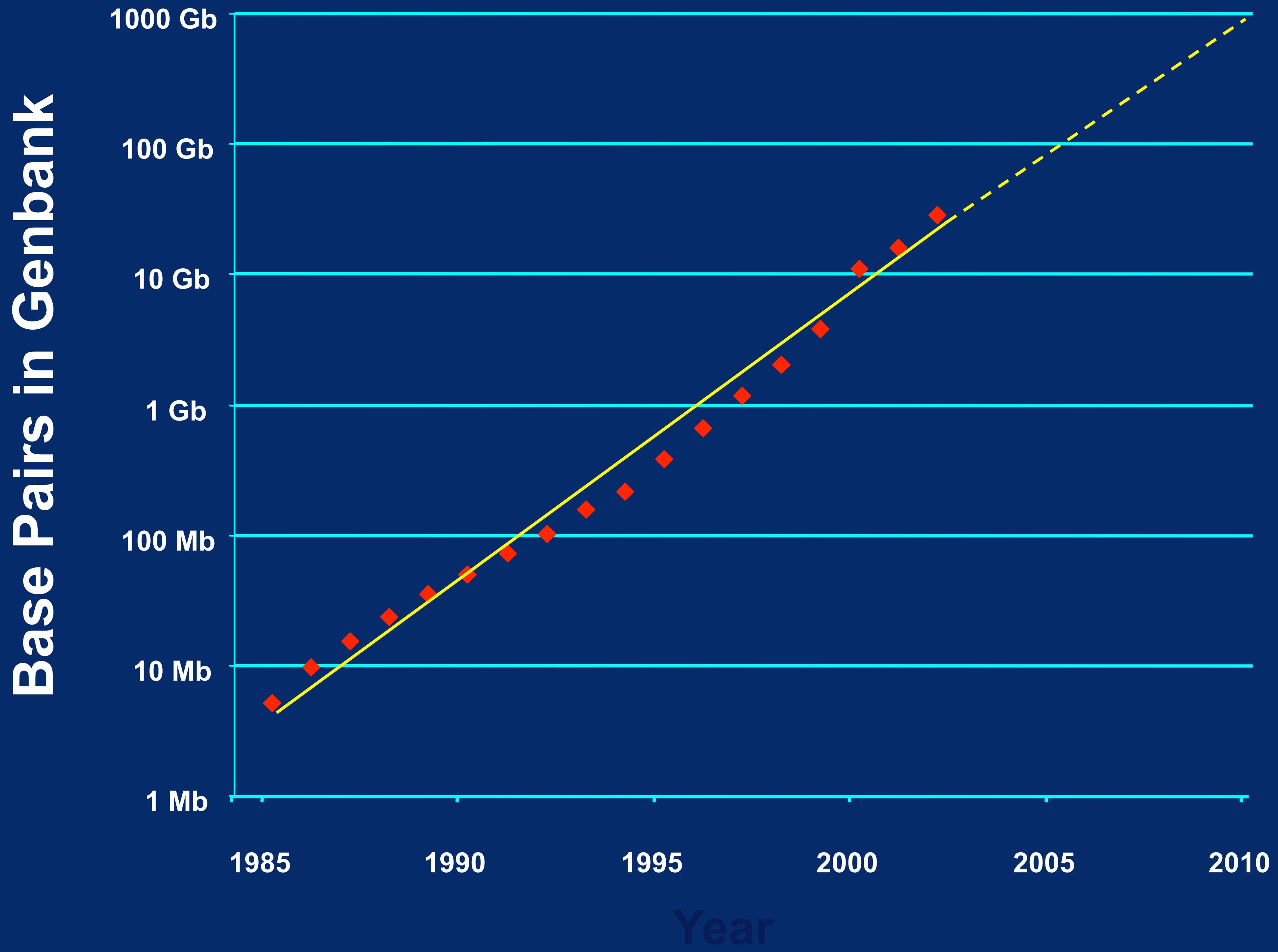
Extragenic DNA

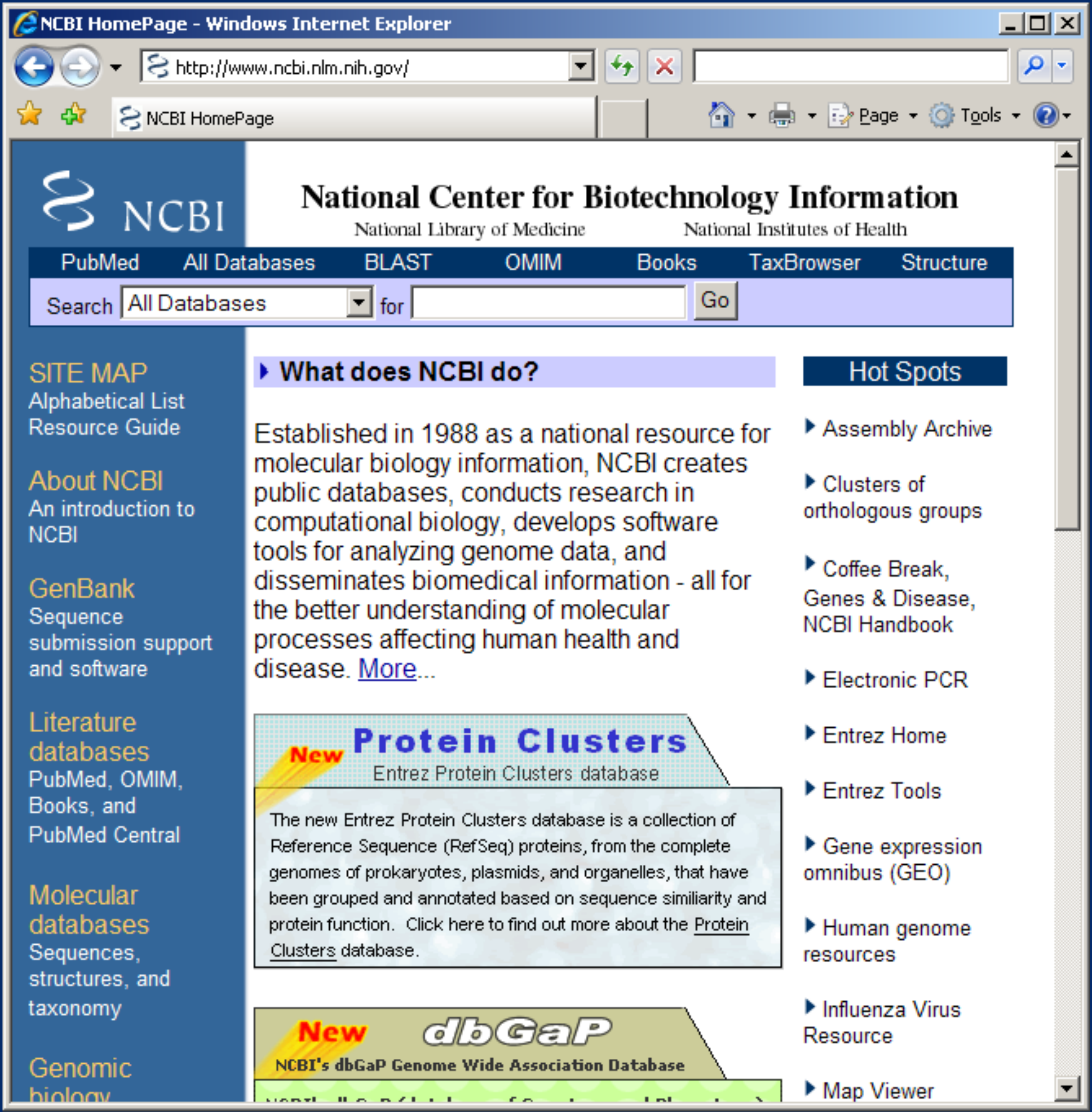
- Repetitive sequences
- Control regions
- Spacer DNA between genes
- Function mostly unknown



Characteristics of the Human Genome Sequence

- 99% of euchromatin is covered, 2.85 Gb
- Error rate: <<1:100,000 bp
- <350 unclonable gaps
- All data is freely accessible without restriction
- Humans have fewer genes than expected
  - ~20,000 from prev. estimates of 100,000)
  - ? human genes make more proteins
- ~1-2% of genome = coding sequences
- ~1% = highly conserved noncoding sequences


Base Pairs in Genbank
1000 Gb
100 Gb
10 Gb
1 Gb
100 Mb
10 Mb
1 Mb
1985
1990
1995
2000
2005
2010
Year


NCBI HomePage - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/
NCBI HomePage
Page
Tools

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

New **Protein Clusters**

Entrez Protein Clusters database

The new Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins, from the complete genomes of prokaryotes, plasmids, and organelles, that have been grouped and annotated based on sequence similiarity and protein function. Click here to find out more about the Protein Clusters database.

New **dbGaP**

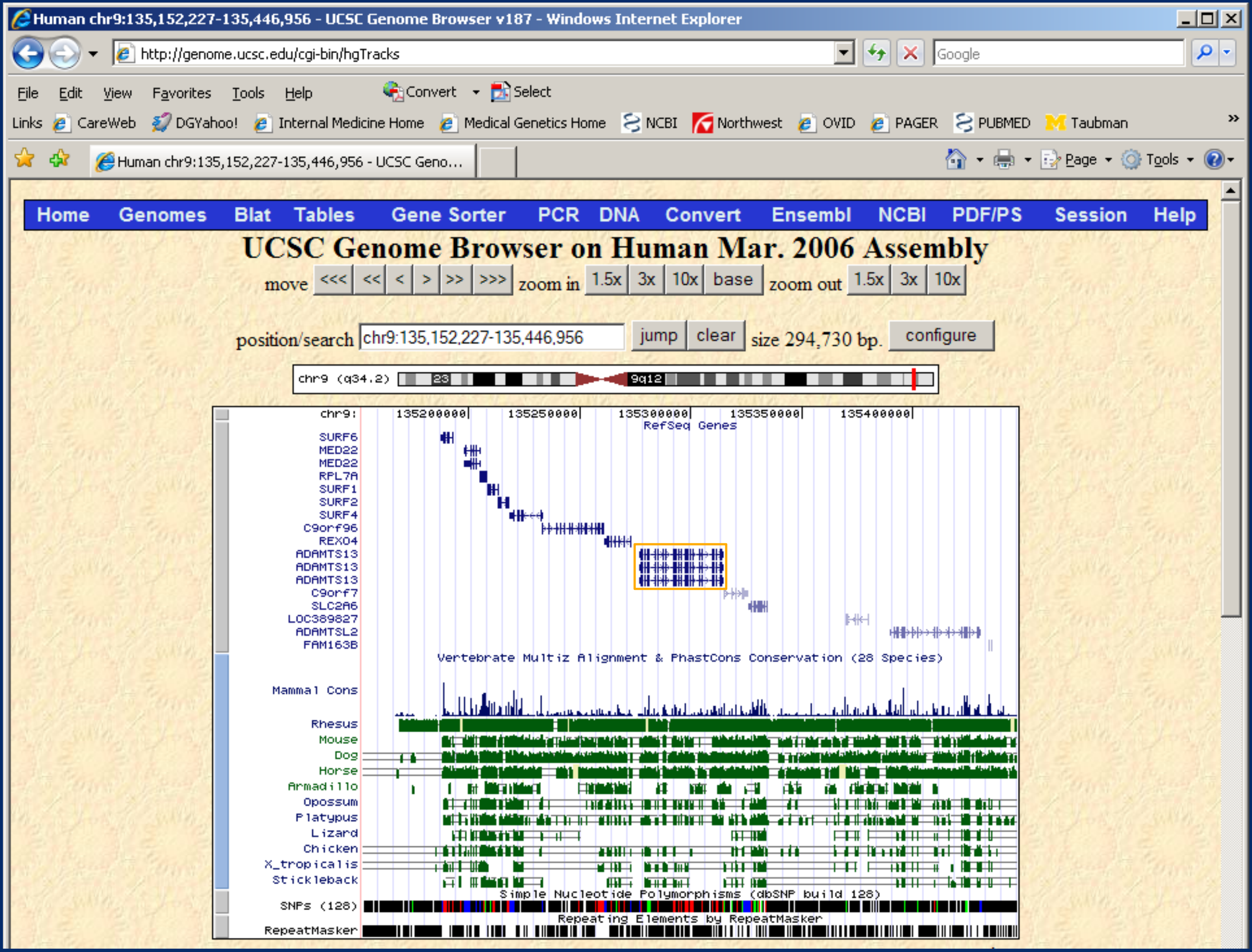
NCBI's dbGaP Genome Wide Association Database

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- Influenza Virus Resource
- Map Viewer

PD-GOV http://www.ncbi.nlm.nih.gov/

http://genome.ucsc.edu



Human chr9:135,152,227-135,446,956 - UCSC Genome Browser v187 - Windows Internet Explorer

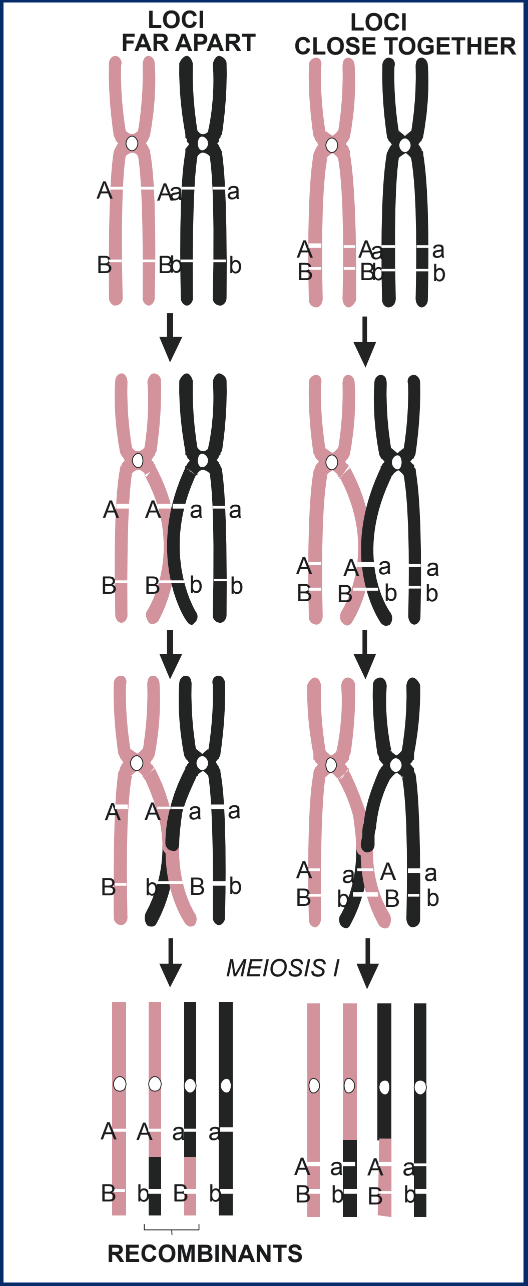
http://genome.ucsc.edu/cgi-bin/hgTracks

UCSC Genome Browser on Human Mar. 2006 Assembly

position/search chr9:135,152,227-135,446,956

size 294,730 bp.

Key Concepts: Linkage and Recombination

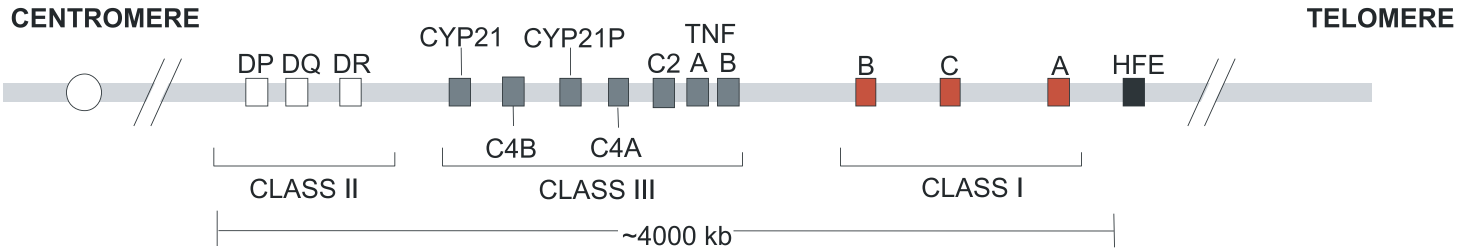


<u>Linkage</u>: A/a and B/b tend to be inherited together

the A and B loci are linked.

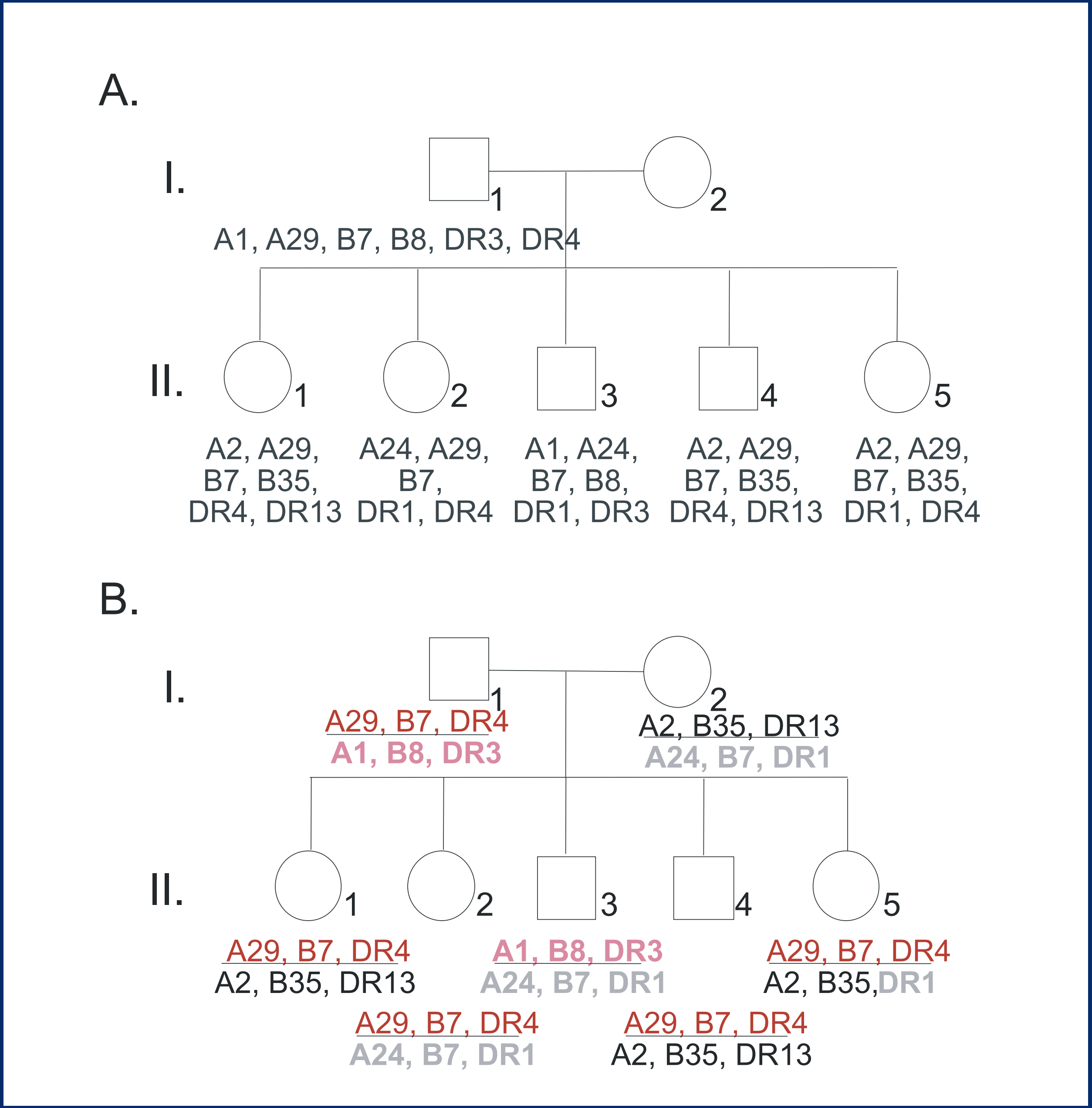
Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.2

The HLA (MHC) Locus



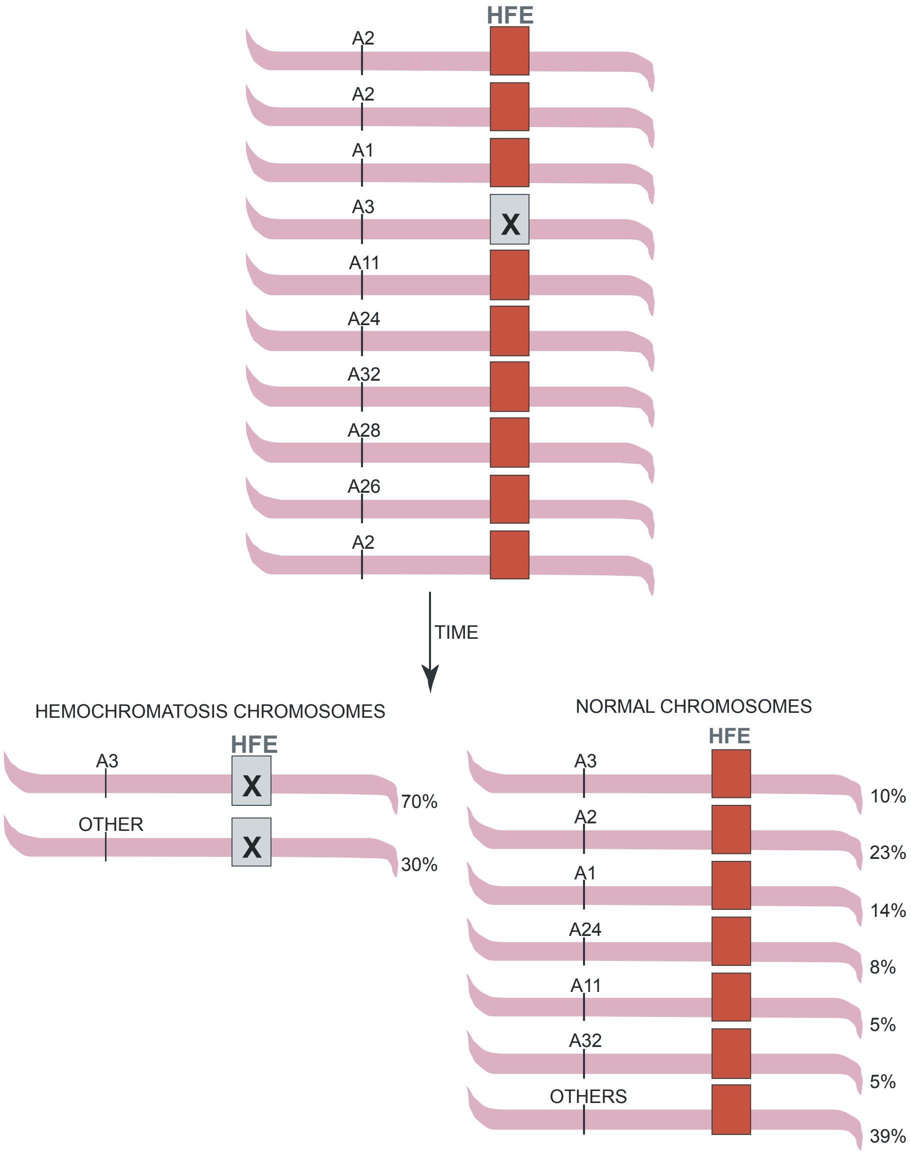
PD-INEL Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.12

Assigning Phase



Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.13

Linkage Disequilibrium



Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.14

These three SNPs could theoretically occur in 8 different haplotypes

…C…A…A…

…C…A…G…

…C…C…A…

…C…C…G…

…T…A…A…

…T…A…G…

…T…C…A…

…T…C…G…

But in practice, only two are observed

...C...A...A...

...C...A...G...

...C...C...A...

...C...C...G...

...T...A...A...

...T...A...G...

...T...C...A...

...T...C...G...

These three variants are said to be in linkage disequilibrium

…C…A…A…

…C…A…G…

…C…C…A…

…C…C…G…

…T…A…A…

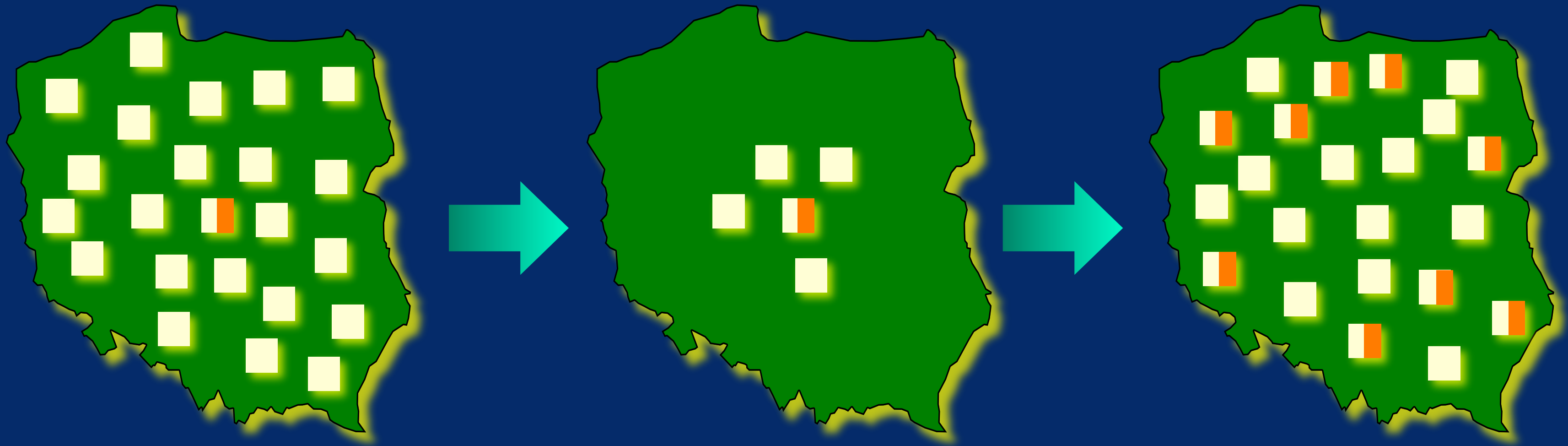
…T…A…G…

…T…C…A…

…T…C…G…

Founder Effect

A high frequency of a specific gene mutation in a population founded by a small ancestral group

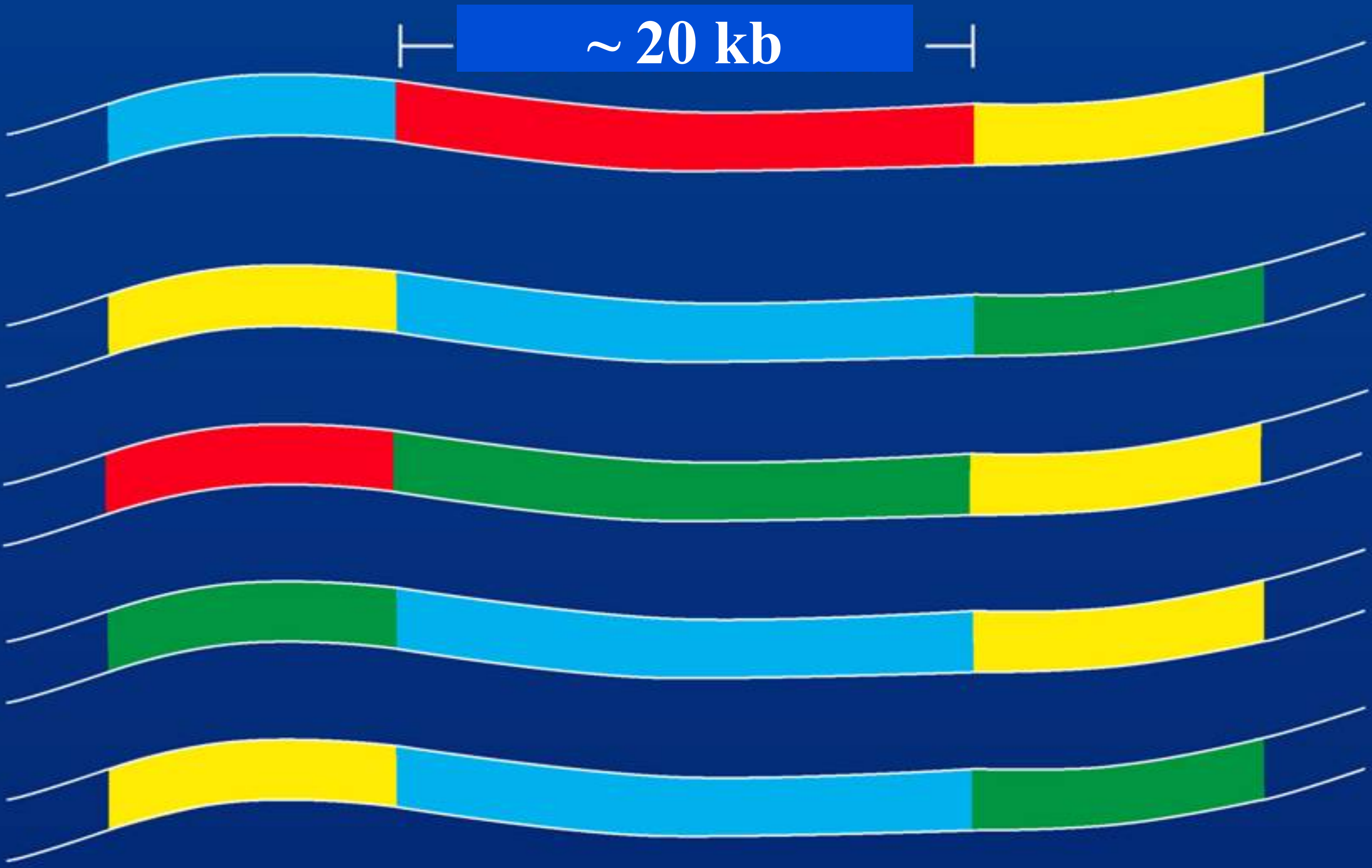


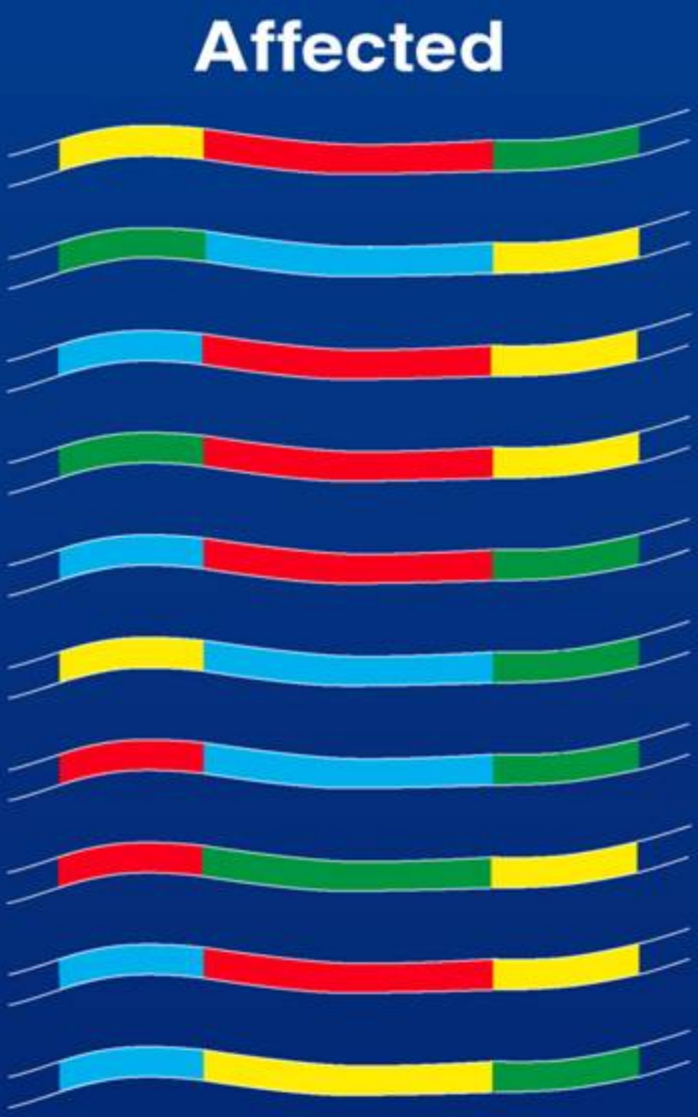
Original population

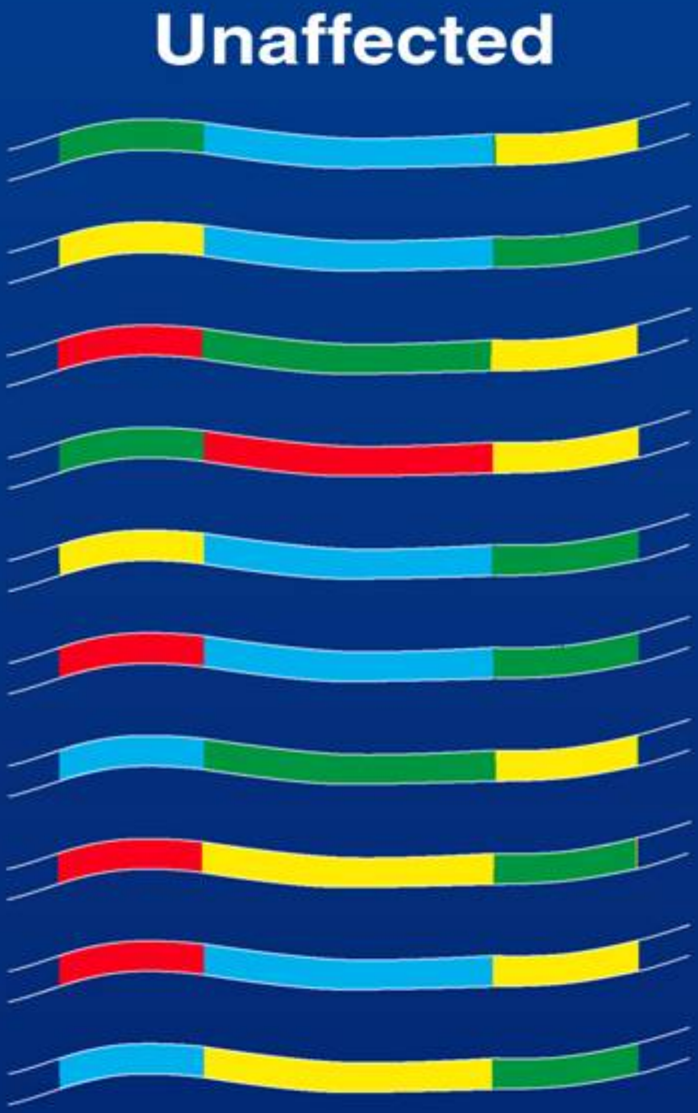
Marked population decrease, migration, or isolation

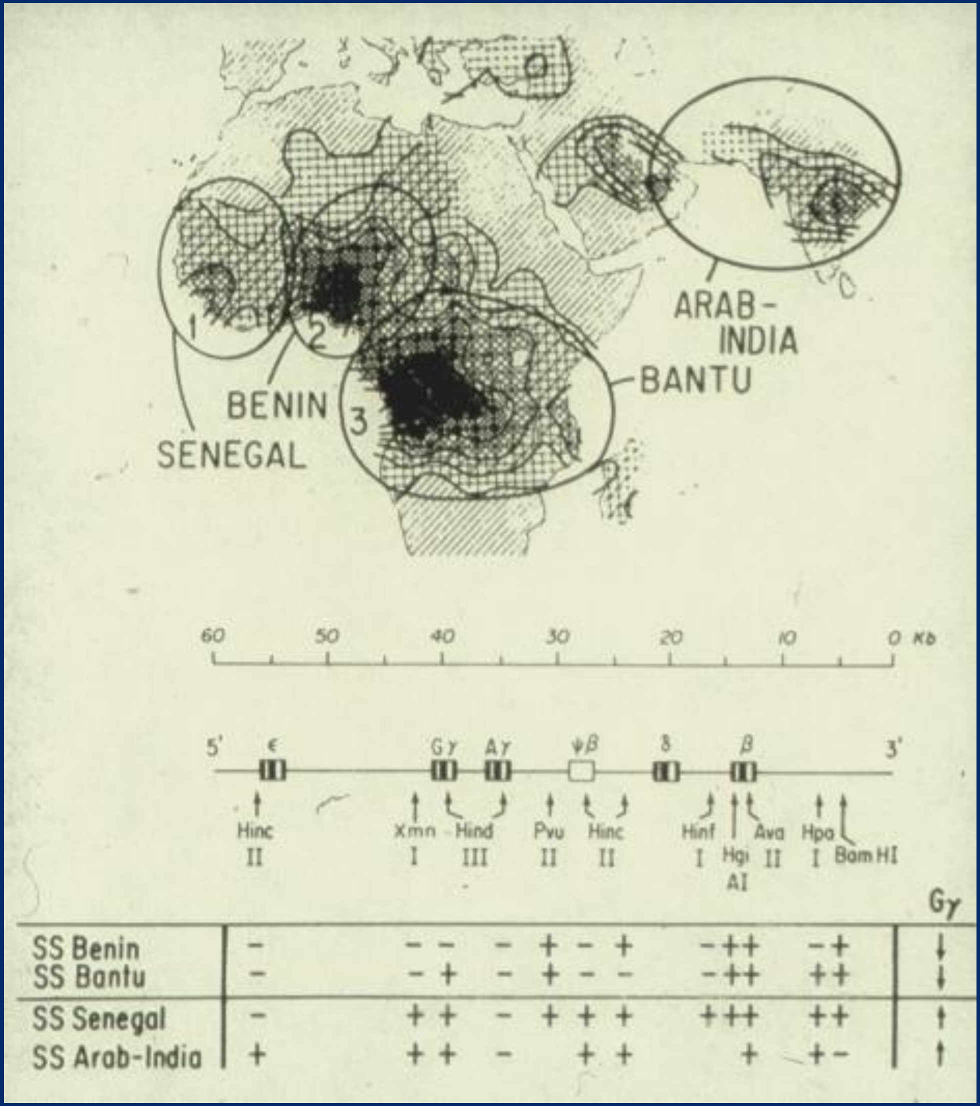
Generations later

ASCO


~ 20 kb


Affected


Unaffected

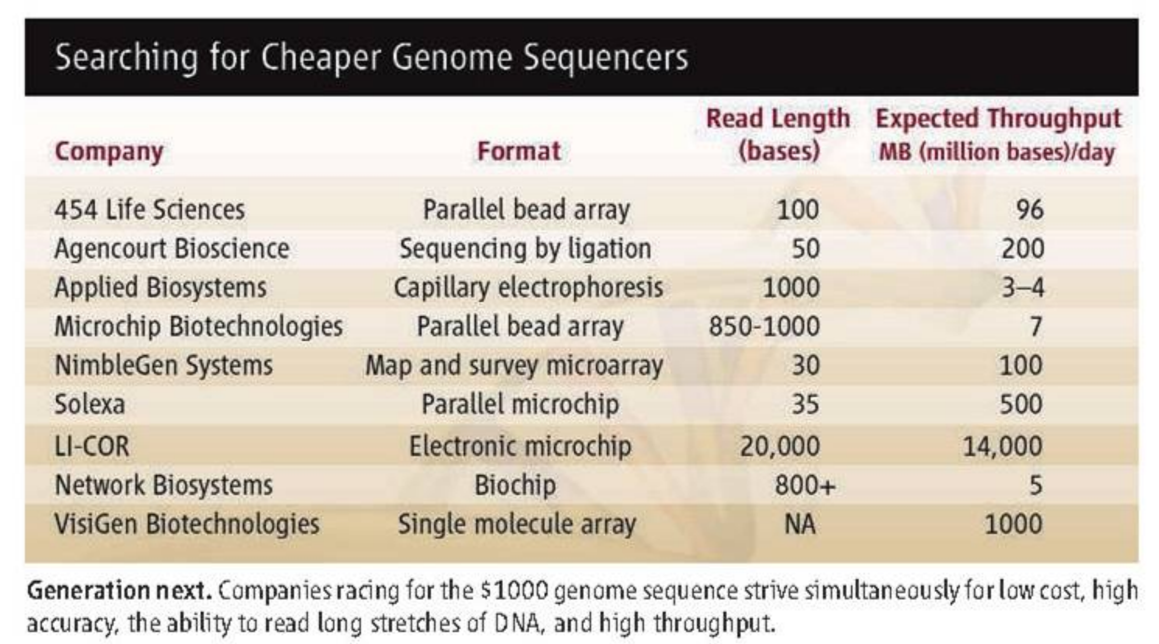


	Hinc II	Xmn I	Hind III	Hind III	Pvu II	Hinc II	Hinc II	Hinf I	Hgi AI	Ava II	Hpa I	Bam HI	Gγ
SS Benin	−	−	−	−	+	−	+	−	+	+	−	+	↓
SS Bantu	−	−	+	−	+	−	−	−	+	+	+	+	↓
SS Senegal	−	+	+	−	+	+	+	+	+	+	+	+	↑
SS Arab-India	+	+	+	−		+	+			+	+	−	↑

Hb S only occurs on 4 haplotypes…only occurred 4 times in history

Could we use this approach to find human disease genes (identify specific haplotypes present more often in patients than in controls)?

Next Generation (NexGen) Sequencing Technologies



Searching for Cheaper Genome Sequencers

Company	Format	Read Length (bases)	Expected Throughput MB (million bases)/day
454 Life Sciences	Parallel bead array	100	96
Agencourt Bioscience	Sequencing by ligation	50	200
Applied Biosystems	Capillary electrophoresis	1000	3–4
Microchip Biotechnologies	Parallel bead array	850-1000	7
NimbleGen Systems	Map and survey microarray	30	100
Solexa	Parallel microchip	35	500
LI-COR	Electronic microchip	20,000	14,000
Network Biosystems	Biochip	800+	5
VisiGen Biotechnologies	Single molecule array	NA	1000

Generation next. Companies racing for the $1000 genome sequence strive simultaneously for low cost, high accuracy, the ability to read long stretches of DNA, and high throughput.

Learning Objectives

UNDERSTAND:

- The basic anatomy of the human genome [eg. 3 X10^9 bp (haploid genome); 1-2% coding sequence (~20,000 genes); types and extent of DNA sequence variation].
- Recombination and how it allows genes to be mapped
- Genetic data for a pedigree, assigning phase, defining haplotypes
- Linkage: Distinction between a linked marker and the disease causing mutation itself
- Linkage disequilibrium and haplotype blocks
- Genome wide association studies (GWAS) to identify gene variants contributing to complex diseases/traits
- The implications of GWAS findings for clinical care and "Personalized Medicine"
- The implications of "Next-Gen" sequencing for future clinical medicine

Additional Source Information
for more information see: http://open.umich.edu/wiki/AttributionPolicy

Slide 22: Source Undetermined; Andre Karwath (wikipedia); U.S. Federal Government (wikimedia)

Slide 24: Source Undetermined

Slide 26: National Center for Biotechnology, http://www.ncbi.nlm.nih.gov/

Slide 27: Gelehrter, Collins and Ginsburg: Principles of Medical Genetics 2E

Slide 28: University Of California Santa Cruz, http://genome.ucsc.edu

Slide 29: Levy, et al. Mutations in a member of the ADAMTS gene family cause thrombotic thrombocytopenic purpura. Nature 413:488-494, 2001.

Slide 30: University Of California Santa Cruz, http://genome.ucsc.edu

Slide 31: National Center for Biotechnology, http://www.ncbi.nlm.nih.gov/Omim/mimstats.html

Slide 41: Regents of The University of Michigan

Slide 42: Regents of The University of Michigan

Slide 46: Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E,* Figure 10.3

Slide 47: Ricardipus, flickr, http://creativecommons.org/licenses/by-sa/2.0/deed.en