Use + Share + Adapt

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

- **Public Domain – Government**: Works that are produced by the U.S. Government. (17 USC § 105)
- **Public Domain – Expired**: Works that are no longer protected due to an expired copyright term.
- **Public Domain – Self Dedicated**: Works that a copyright holder has dedicated to the public domain.
- **Creative Commons – Zero Waiver**
- **Creative Commons – Attribution License**
- **Creative Commons – Attribution Share Alike License**
- **Creative Commons – Attribution Noncommercial License**
- **Creative Commons – Attribution Noncommercial Share Alike License**
- **GNU – Free Documentation License**

Make Your Own Assessment

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

- **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**: 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
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 [e.g. $3 \times 10^9$ 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
  – $1 \times 10^{-8}$/bp/generation (~70 new mutations/individual)

• balanced polymorphism= disease + polymorphism

• Common but incorrect usage:
  – “mutation vs. polymorphism”
Common but incorrect usage:

“a disease-causing mutation” OR “a polymorphism”
B.

**Heteromorphism of chromosome 1**

(one copy)

2 normal copies of chromosome 1

*a or b allele of Duffy blood group*

Donahue, 1968

Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E*; Figure 9.1
Linkage: A/a and B/b tend to be inherited together

the A and B loci are linked.
Linkage between Marker A/a and Disease D

Marker = A or a
Disease allele = D
Normal allele = N
Linkage between NF and RFLP marker

Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E*; Figure 9.5
Functional Cloning

Positional Cloning

Disease

Map

Function

Gene

Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E*; Figure 9.15
HD linked to C allele: Two recombinants (III13, IV1)


Positional Cloning

Genetic Markers

Families

Cytogenic Abnormality

Physical Mapping and Cloning

Transcript Identification

YACs and BACs

Positional Candidate Approach

Mutation Search

Mutation Identification

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:** Restriction Fragment Length Polymorphism
- **VNTR:** Variable Number of Tandem Repeats  
  - or minisatellite  
  - ~10-100 bp core unit
- **SSR:** Simple Sequence Repeat  
  - or STR (simple tandem repeat)  
  - or microsatellite  
  - ~1-5 bp core unit
- **SNP:** Single Nucleotide Polymorphism  
  - Commonly used to also include rare variants
- **Insertions or deletions**  
  - INDEL – small (few nucleotides) insertion or deletion
- **Rearrangement** (inversion, duplication, complex rearrangement)
- **CNV:** Copy Number Variation
SNP

- 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

1981

3 markers

1991

53 markers

1994

393 markers

1996

791 markers

2010

- 23,653,737 total human entries in dbSNP

- Chromosome 4
  - 4,311,728 SNPs

- ~1M SNP chip commercially available

Gelehrter, Collins and Ginsburg: Principles of Medical Genetics 2E; Figure 10.3
Genes Identified: Monogenic Diseases

Human Genome Project Begins

Online Mendelian Inheritance in Man
Haploid Human Genome $3 \times 10^9$ bp, ~20,000 genes

1 Chromosome
~1300 genes

Single Gene
~1.5 Kb (Globin to 2 $\times 10^6$ bp (Dystrophin)

H. Influenzae
~1700 genes

S. Cerevisiae
~6250 genes

D. Melanogaster
~14000 genes

C. Elegans
~18500 genes
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

- 70%
- 30%

~20,000 genes
Characteristics of the Human Genome Sequence

- 99% of euchromatin is covered, 2.85 Gb
- Error rate: $\ll 1:100,000$ bp
- $<350$ unclonable gaps
- All data is freely accessible without restriction
- Humans have fewer genes than expected
  - $\sim 20,000$ from prev. estimates of 100,000
  - $?$ human genes make more proteins
- $\sim 1-2\%$ of genome = coding sequences
- $\sim 1\%$ = highly conserved noncoding sequences
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 similarity and protein function. Click here to find out more about the Protein Clusters database.

New dbGaP
NCBI's dbGaP Genome Wide Association Database

http://genome.ucsc.edu
Key Concepts: Linkage and Recombination

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

the A and B loci are linked.
The HLA (MHC) Locus

Gelehrter, Collins and Ginsburg: *Principles of Medical Genetics 2E;* Figure 9.12
Assigning Phase

A.

I.

1

A1, A29, B7, B8, DR3, DR4

II.

1 2 3 4 5

A2, A29, B7, B35, DR4, DR13

A24, A29, B7, DR1, DR4

A1, A24, B7, B8, DR1, DR3

A2, A29, B7, B35, DR4, DR13

A2, A29, B7, B35, DR1, DR4

B.

I.

1

A29, B7, DR4

A1, B8, DR3

II.

1 2 3 4 5

A29, B7, DR4

A2, B35, DR13

A24, B7, DR1

A29, B7, DR3

A24, B7, DR1

A29, B7, DR4

A2, B35, DR13

A29, B7, DR4

A2, B35, DR1
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...
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
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

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

*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. $3 \times 10^9$ 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 27: Gelehrter, Collins and Ginsburg: Principles of Medical Genetics 2E
Slide 28: University Of California Santa Cruz, http://genome.ucsc.edu
Slide 30: University Of California Santa Cruz, http://genome.ucsc.edu
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