

# **Bioinformatics**

**CS300**

**Chapter 1:**

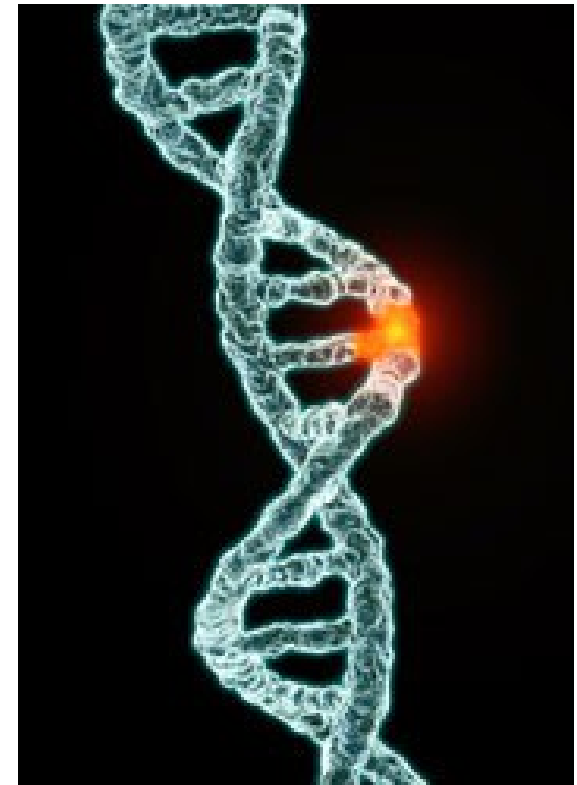
**Using Bioinformatics to  
study genetic disorders**

**Fall 2017**

**Oliver Bonham-Carter**

# Genetic Disorder

- a disorder/disease with a genetic component
- caused by one or more abnormalities in the genome
  - substitutions
  - insertions
  - deletions
  - rearrangements



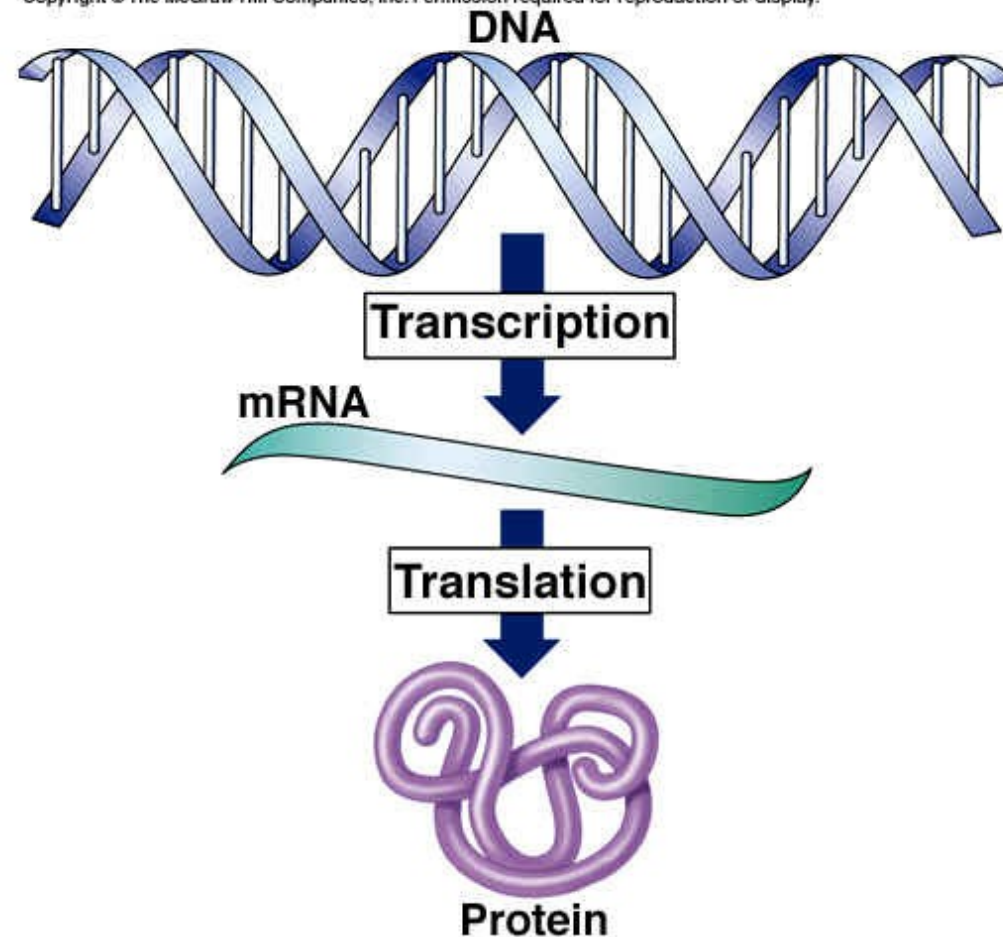


# Genetic Disorder

- a disorder/disease with a genetic component
- caused by one or more abnormalities in the genome
- single gene disorders
  - mutation(s) in the sequence of a single gene
  - alters or eliminates protein product

# Central Dogma of Gene Expression

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

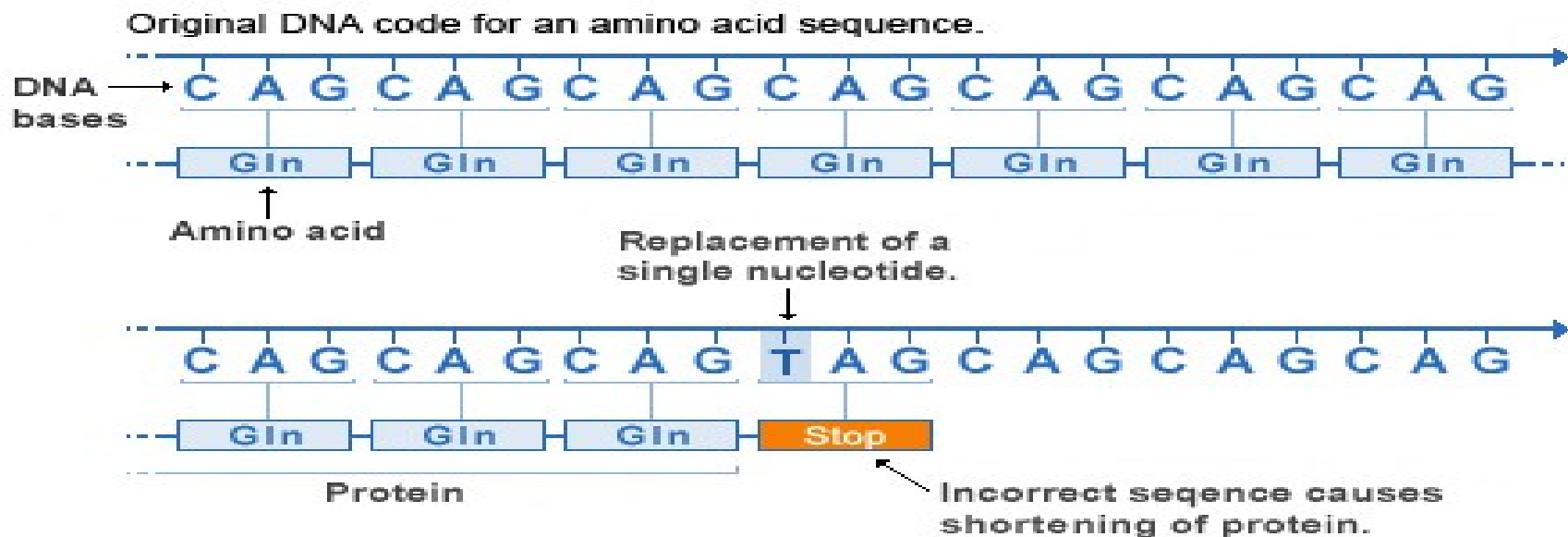


# Mutations and Their Potential Effects

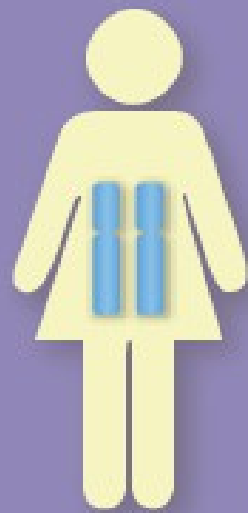
## Mutations that Alter Protein Products and Mutations that Eliminate Protein Products

- Missense substitutions
- Nonsense substitutions
- Insertions/Deletions

### Nonsense mutation

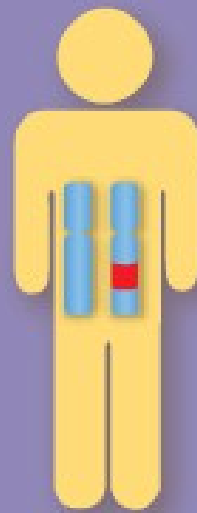


# Autosomal Dominant Inheritance



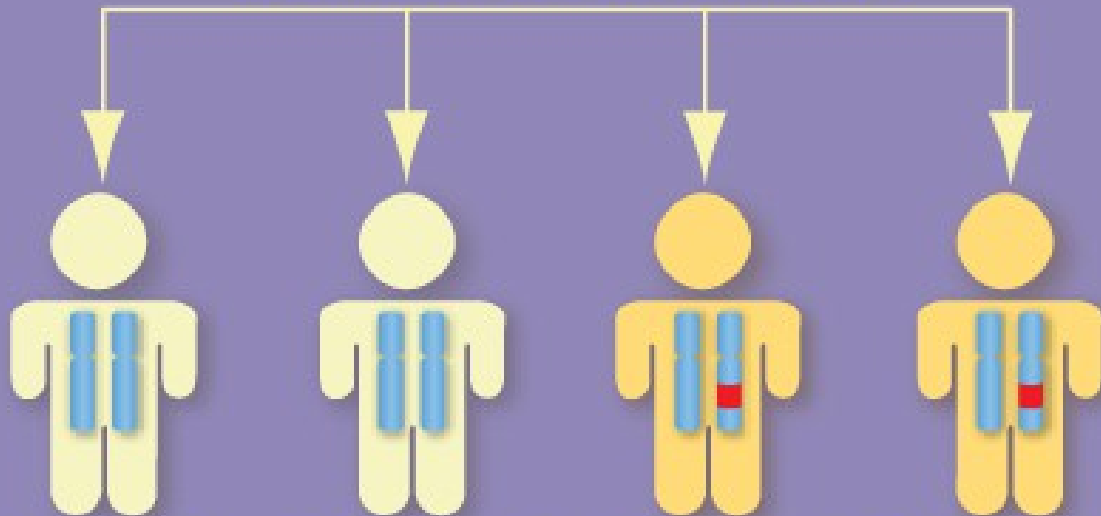
Normal

MOM DAD



Affected

Possible combinations:



Normal

Normal

Affected

Affected



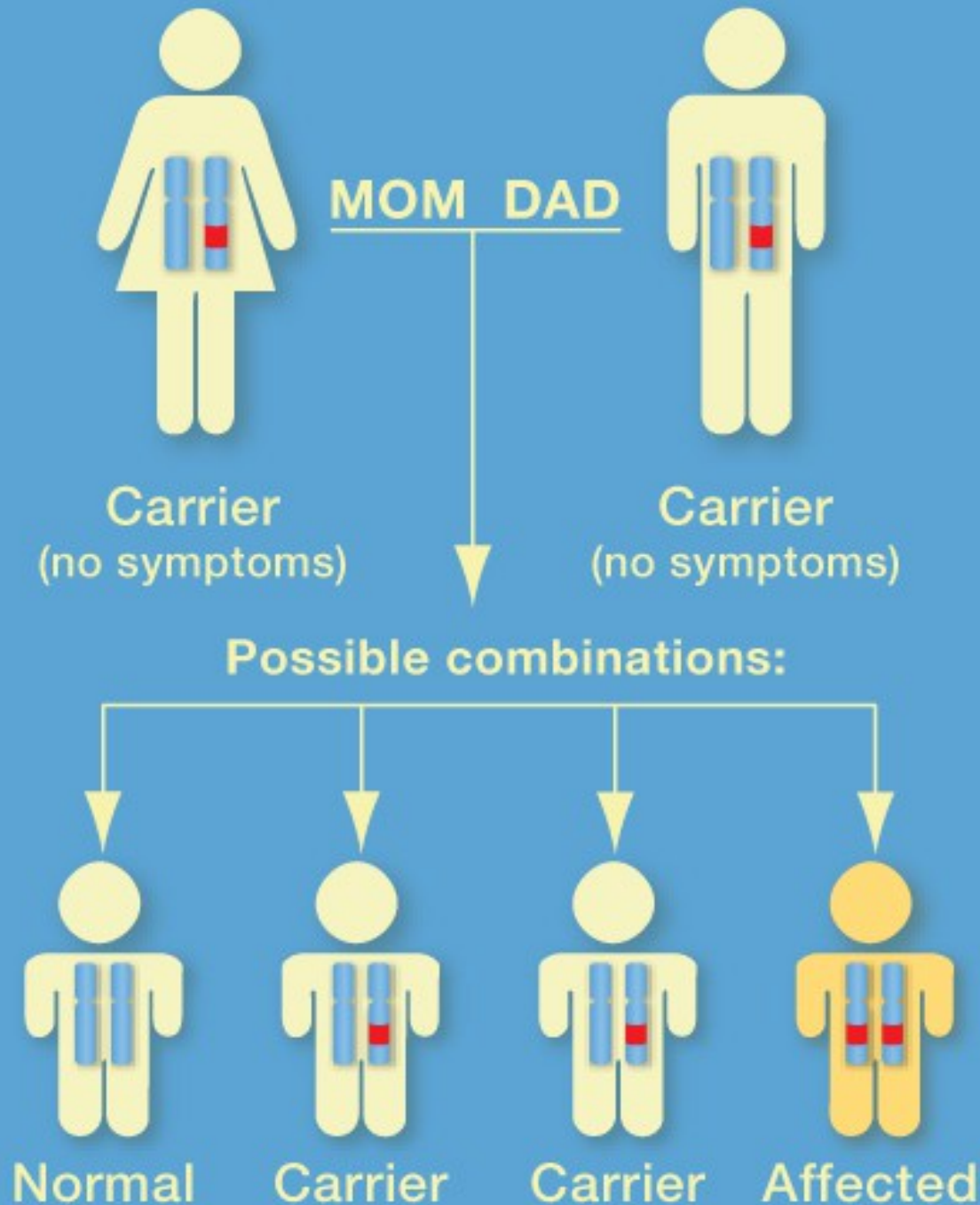
Chromosome with  
normal copy of gene



Chromosome with  
defective copy of gene

Each child inherits a normal copy from Mom and either a normal or a defective copy from Dad.

# Autosomal Recessive Inheritance

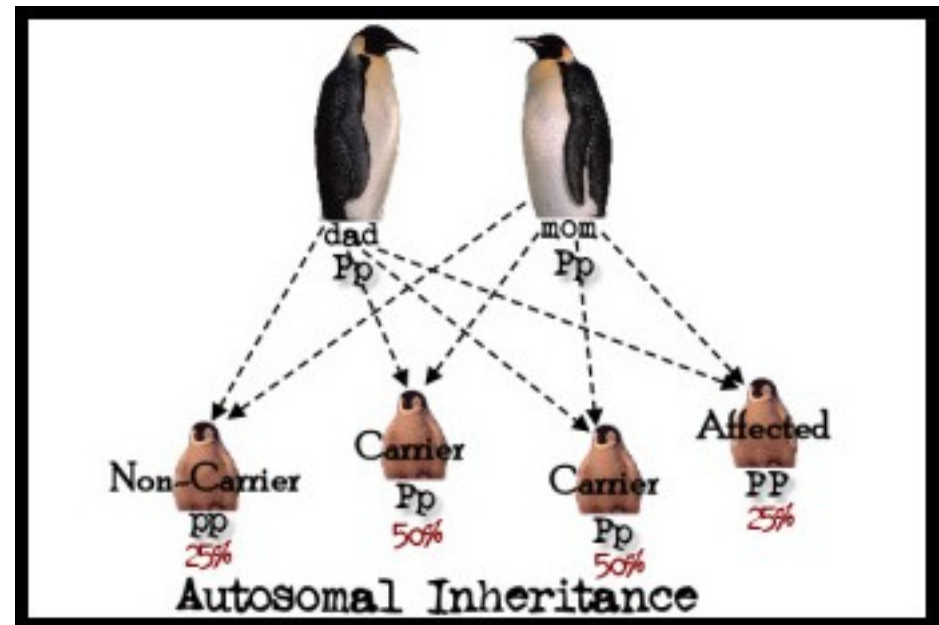


- Chromosome with normal copy of gene
- Chromosome with defective copy of gene

Each child inherits one copy of the gene from each parent.

# Single Gene Disorders

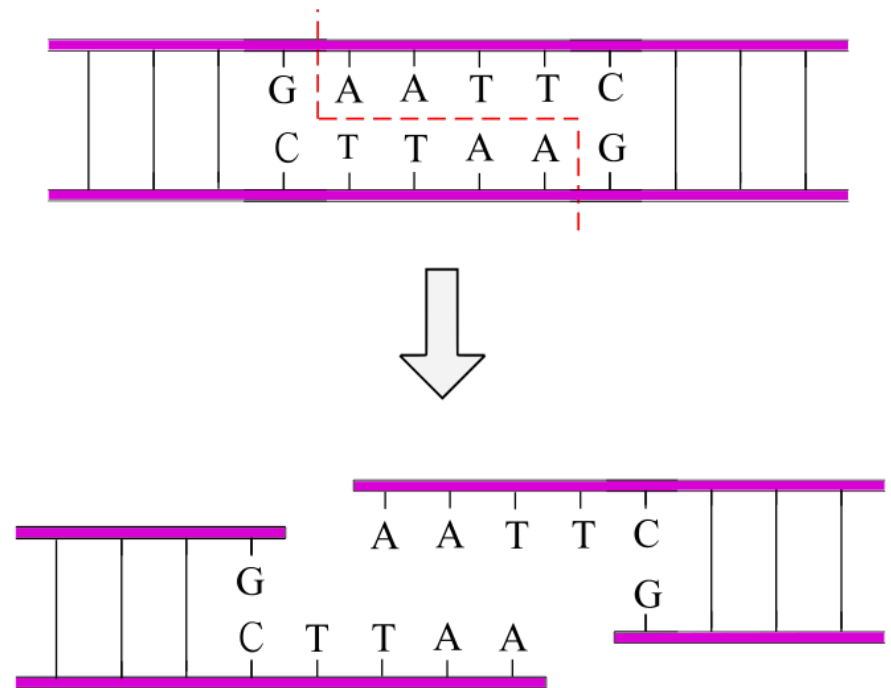
- Inheritance patterns relatively simple
- Chance of inheritance in the text generation can be predicted by studying patterns in past generations.





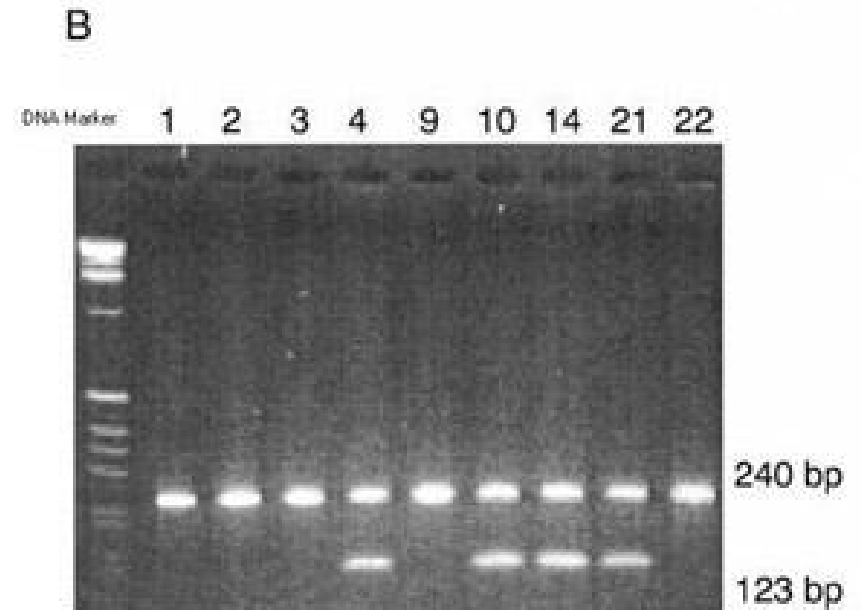
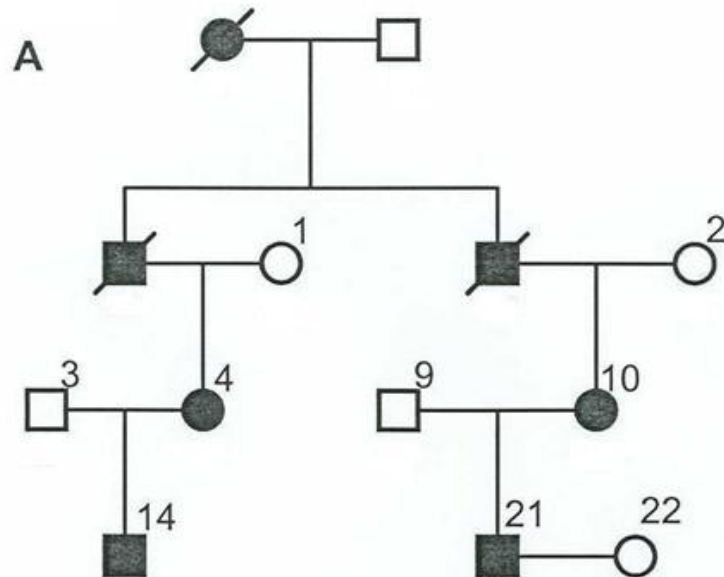
# Single Gene Disorders

- Most genes identified in the 1980s-1990s
  - Pre-Bioinformatics:  
biological wet-lab work
- Restriction enzymes to cut sequences
- cut DNA at specific sequence
  - 100s of different patterns
  - Disorder-breeding sequences could be studied



# Single Gene Disorders

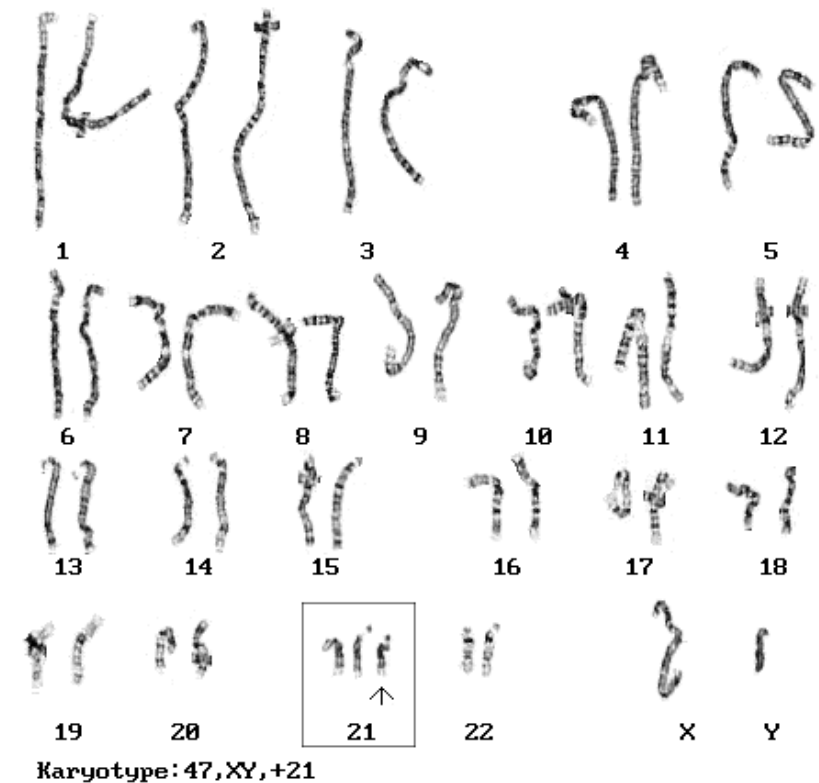
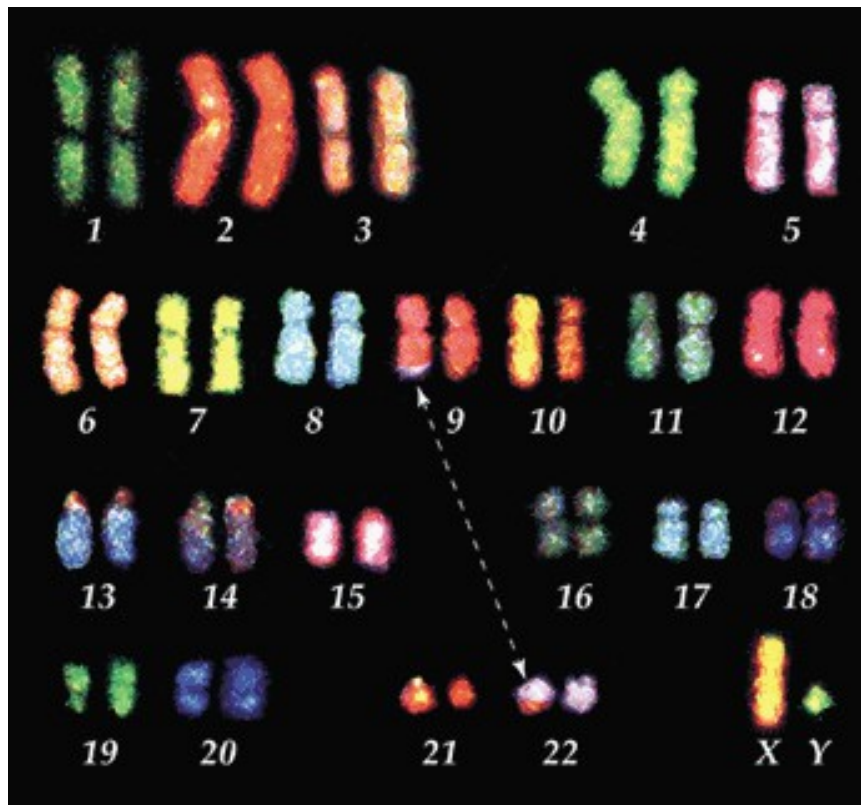
## Pedigree analysis + Restriction Digest Analysis



# Single Gene Disorders

## Cytogenetics

- field of biology concerned with mapping genes to specific locations on chromosomes



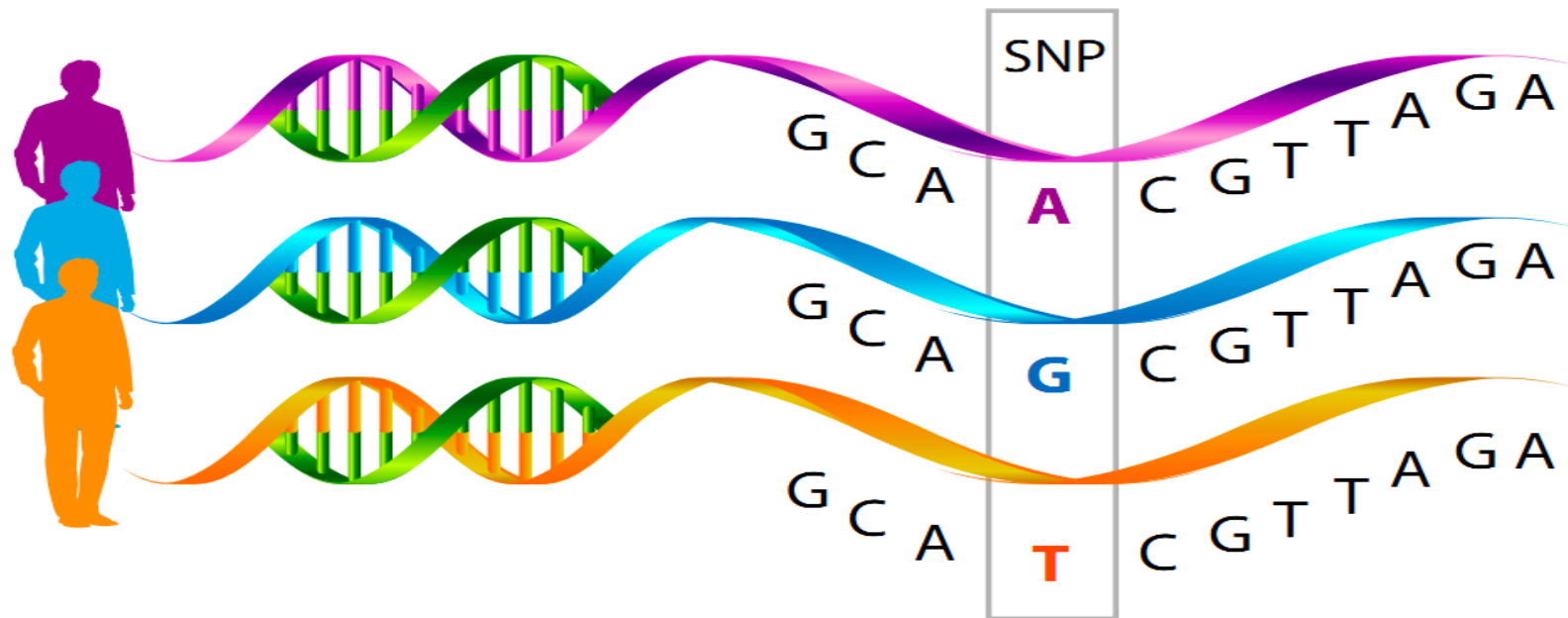


# Genetic Disorder

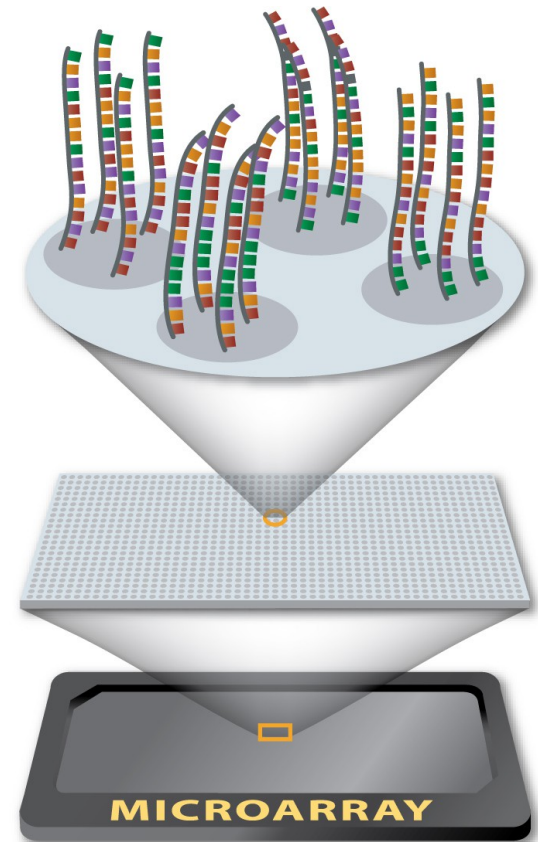
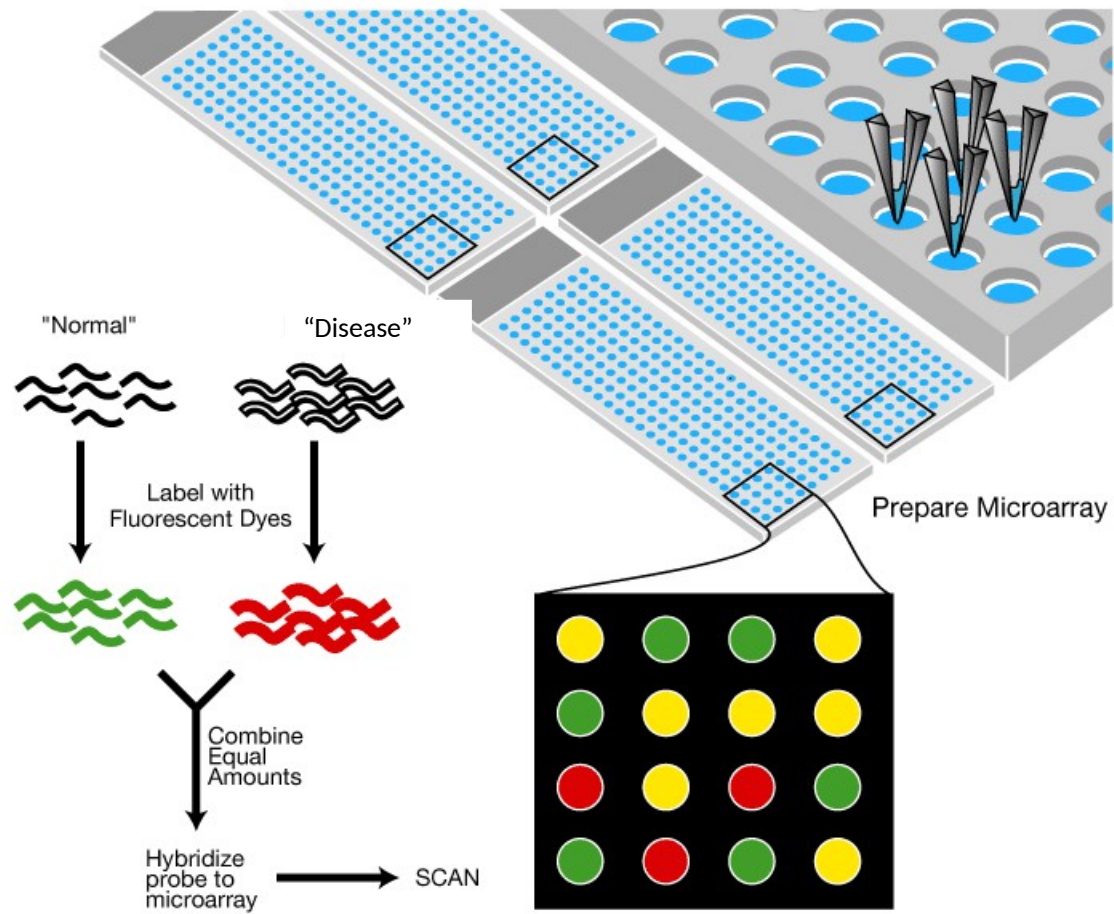
- a disease with a genetic component
- caused by one or more abnormalities in the genome
- complex or multifactorial disorders
  - do not have a single genetic cause
  - likely associated with the effects of multiple genes in combination with lifestyle and environmental factors
  - do not have a clear cut pattern of inheritance

# Genome-Wide Association Studies

- new technology/analysis - early 2000s
  - bioinformatics
- screen 1000s of genomes at once for **SNPs**
  - single nucleotide polymorphisms
  - Some SNPs may indicate disorders



# DNA Microarray







# GWAS - Genome-wide Association Studies

NHGRI FACT SHEETS

genome.gov

Individuals with disease

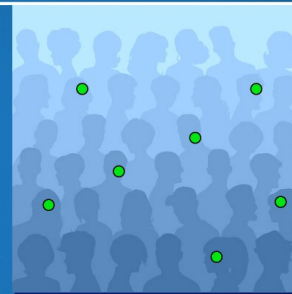
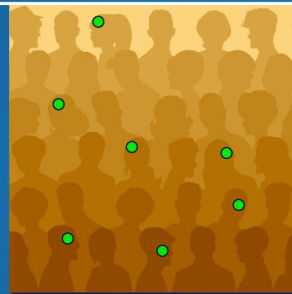


Individuals without disease



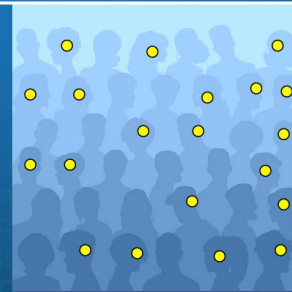
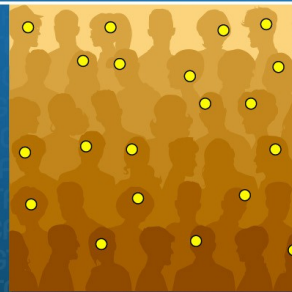
Using a CHIP can genotype  
500,000 - 5 Million SNPs

SNP 1



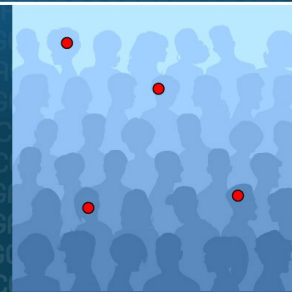
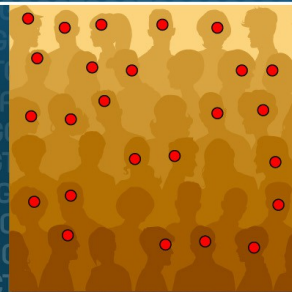
SNP 1  
No association  
to disease

SNP 2



SNP 2  
No association  
to disease

SNP 3

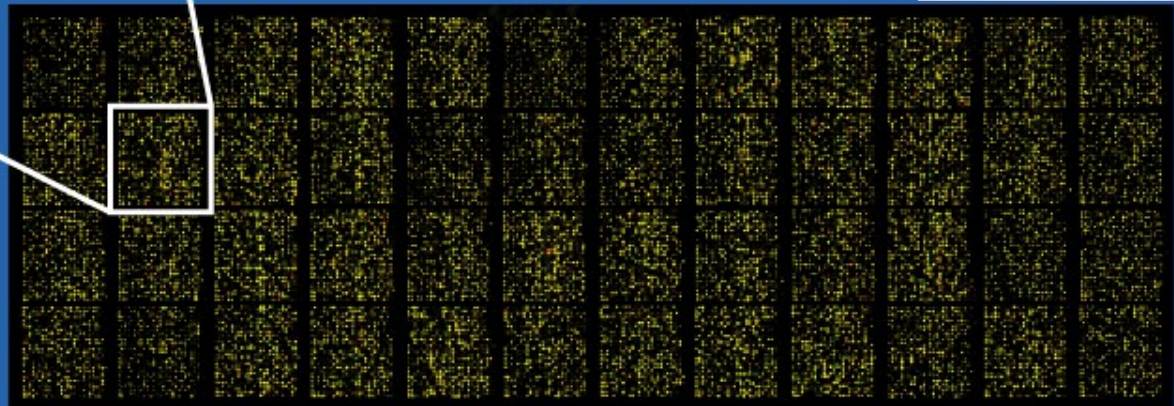
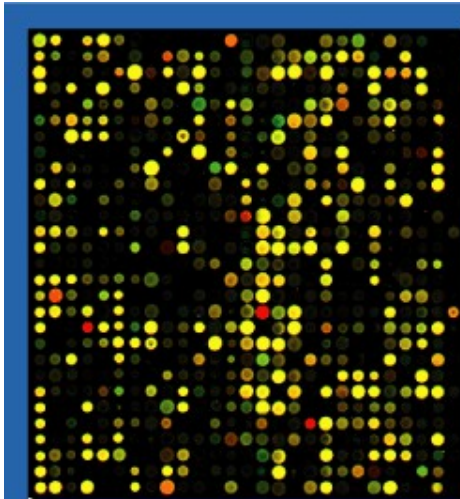


SNP 3  
Associated  
to disease



ALLEGHENY  
COLLEGE

# DNA Microarray







# DNA Microarray

			Human Brain Microarray
Gene	Probe	Fold Change	
SAG	A_23_P5853	97.974	
	CUST_14866_PI416261804	20.417	
PDYN	CUST_653_PI417557136	90.414	
	A_24_P279870	52.069	
	CUST_635_PI417557136	47.493	
	CUST_643_PI417557136	43.811	
	A_23_P40262	41.904	
	CUST_645_PI417557136	41.007	
	CUST_649_PI417557136	38.397	



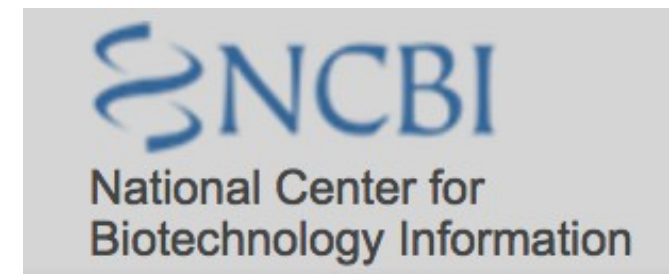
# Candidate SNPs that May Be Correlated with Disorder Using *Genome-Wide Association Studies* (GWAS)

**Table 2.** GWAS results for all SNPs with  $p < 10^{-6}$  in the 23andMe cohort.

SNP	Chr	Position	Region	Alleles	MAF	Cohort	OR	<i>p</i>
rs34637584	12	39020469	LRRK2	G/A	0.002	23andMe	9.615 (6.43–14.37)	$1.82 \times 10^{-28}$
						IPDGC	–	–
i4000416	1	153472258	GBA	T/C	0.005	23andMe	4.048 (3.08–5.32)	$5.17 \times 10^{-21}$
						IPDGC	–	–
rs356220	4	90860363	SNCA	C/T	0.375	23andMe	1.285 (1.22–1.36)	$2.29 \times 10^{-19}$
						IPDGC	–	–
rs12185268	17	41279463	MAPT	A/G	0.211	23andMe	0.769 (0.72–0.82)	$2.72 \times 10^{-14}$
						IPDGC	–	–
rs10513789	3	184242767	MCCC1/LAMP3	T/G	0.201	23andMe	0.803 (0.75–0.86)	$2.67 \times 10^{-10}$
						IPDGC	0.873 (0.83–0.92)	$1.7 \times 10^{-6}$
rs6812193	4	77418010	SCARB2	C/T	0.365	23andMe	0.839 (0.79–0.89)	$7.55 \times 10^{-10}$
						IPDGC	0.90 (0.86–0.94)	$3.29 \times 10^{-6}$
rs6599389	4	929113	GAK	G/A	0.075	23andMe	1.311 (1.19–1.44)	$3.87 \times 10^{-8}$
						IPDGC	–	–
rs11868035	17	17655826	SREBF1/RAI1	G/A	0.309	23andMe	0.851 (0.80–0.90)	$5.61 \times 10^{-8}$
						IPDGC	0.95 (0.91–0.996)	0.033
rs823156	1	204031263	SLC41A1	A/G	0.183	23andMe	0.827 (0.77–0.89)	$1.27 \times 10^{-7}$
						IPDGC	–	–
rs4130047	18	38932233	RIT2/SYT4	T/C	0.313	23andMe	1.161 (1.10–1.23)	$2.44 \times 10^{-7}$
						IPDGC	1.077 (1.03–1.13)	0.0014
rs2823357	21	15836776	USP25	G/A	0.376	23andMe	1.149 (1.09–1.21)	$6.32 \times 10^{-7}$
						IPDGC	0.971 (0.93–1.02)	0.187

# Data for Research

- Free data in public databases
- Typically Protein: **Uniprot**
- <http://www.uniprot.org/>
- Search: Pink1 (protein)
- Typically DNA and Genes: **National Center for Biotechnology Informatics (NCBI)**
- <https://www.ncbi.nlm.nih.gov/>
- Search: “orchid” (*nucleotide*)





# Python Programming

- Loading, Writing files
- Biopython
- Fasta files

Follow along in  
class and save  
your notes in  
a text file!!



python