Lecture 29 - Polymorphisms in Human DNA Sequences

- SNPs
- SSRs

Eukaryotic Genes and Genomes

genome = DNA content of a complete haploid set of chromosomes
= DNA content of a gamete (sperm or egg)

Species	Chromosomes	cM	DNA content/ haploid(Mb)	year sequence completed	genes/ haploid
E. coli	1	N/A	5	1997	4,200
S. cerevisiae	16	4000	12	1997	5,800
C. elegans	6	300	100	1998	19,000
D. melanogaster	4	280	180	2000	14,000
M. musculus	20	1700	3000	2002 draft 2005 finished?	30,000?
H. sapiens	23	3300	3000	2001 draft 2003 finished	30,000?

Note: cM = centi Morgan = 1% recombination
Mb = megabase = 1 million base-pairs of DNA
Kb = kilobase = 1 thousand base-pairs of DNA

Species	cM	DNA content/ haploid (Mb)	generation time	design crosses?	true breeding strains?
E. coli	N/A	5	30 min	yes	yes
S. cerevisiae	4000	12	90 min	yes	yes
C. elegans	300	100	4 d	yes	yes
D. melanogaster	280	180	2 wk	yes	yes
M. musculus	1700	3000	3 mo	yes	yes
H. sapiens	3300	3000	20 yr	no	no

- Human genetics is retrospective (vs prospective). Human geneticists cannot test hypotheses prospectively. The mouse provides a prospective surrogate.
- Can't do selections
- Meager amounts of data — Human geneticists typically rely upon statistical arguments as opposed to overwhelming amounts of data in drawing connections between genotype and phenotype.
- Highly dependent on DNA-based maps and DNA-based analysis

The unique advantages of human genetics:

- A large population which is self-screening to a considerable degree
- Phenotypic subtlety is not lost on the observer
- The self interest of our species

A locus is said to be polymorphic if two or more alleles are each present at a frequency of at least 1% in a population of animals.

1) SNPs = single nucleotide polymorphisms = single nucleotide substitutions

H_{nuc} = average heterozygosity per nucleotide site

In human populations: = 0.001

2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

SYNONOMOUS CHANGES

TTT GCT GGC CAC

Phe Ala Gly His

TTT GCT GGA CAC

Phe Ala Gly His

NON-SYNONOMOUS CHANGES

TTT GCT GGC CAC

Phe Ala Gly His

TTT GCT TGC CAC

Phe Ala Cys His

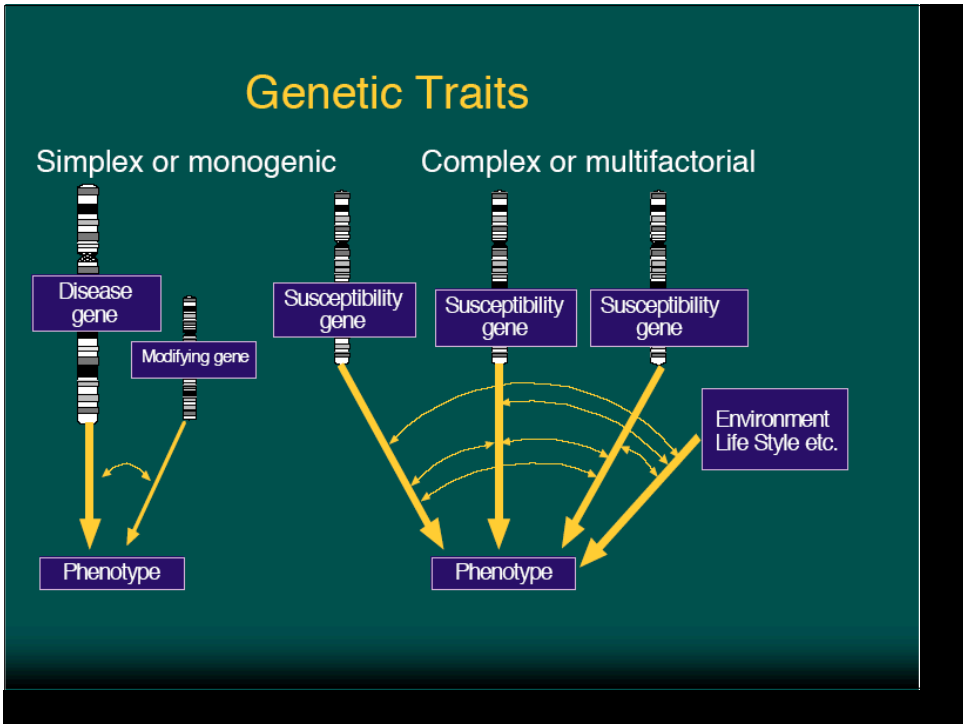
The great majority (probably 99%) of SNPs are selectively "neutral" changes of little or no functional consequence:

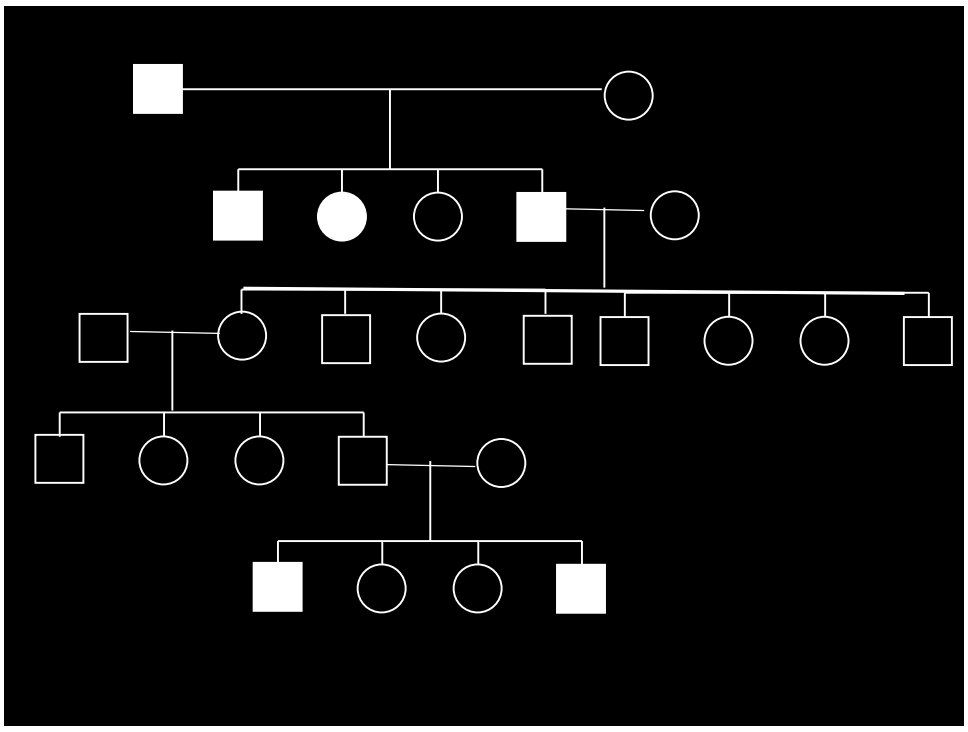
- outside coding or gene regulatory regions (>97% of human genome)
- silent substitutions in coding sequences
- some amino acid substitutions do not affect protein stability or function
- disadvantageous SNPs selected against --> further underrepresentation

A small minority of SNPs are of functional consequence and are selectively advantageous or disadvantageous.

Affymetrix chip

Genetic Traits





TUMORS
C57black
AA

X

NON-TUMORS
C57black
aa

↓

All Tumorous Aa

↓

3 Tumors :: 1 non-tumor

TUMORS NON-TUMORS

C57black X AKR

↓

All NON-TUMORS (normal)

↓

13/16 NON-TUMORS:: 3/16 tumors

AKR HAS A GENE (B) THAT SUPPRESSES TUMORS

TUMORS NON-TUMORS

C57black X AKR

AAbb aaBB

↓

AaBb

All Non-Tumors (normal)

↓

13/16 non-tumors :: 3/16 tumors

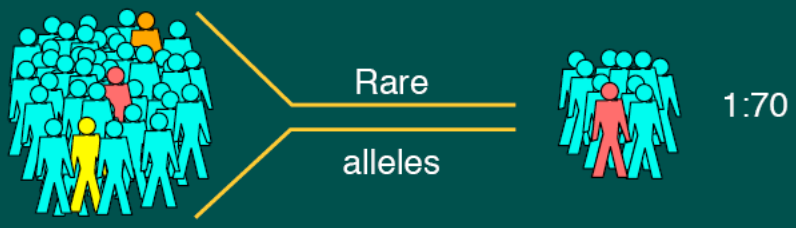
A-B-
aaB-
aabb

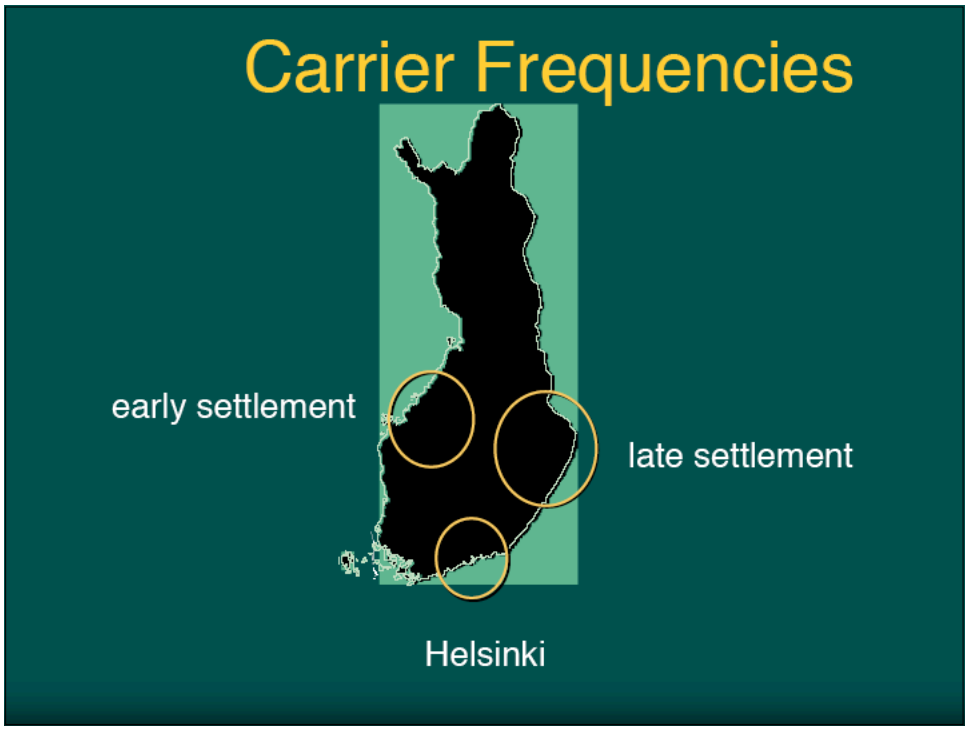
A-bb

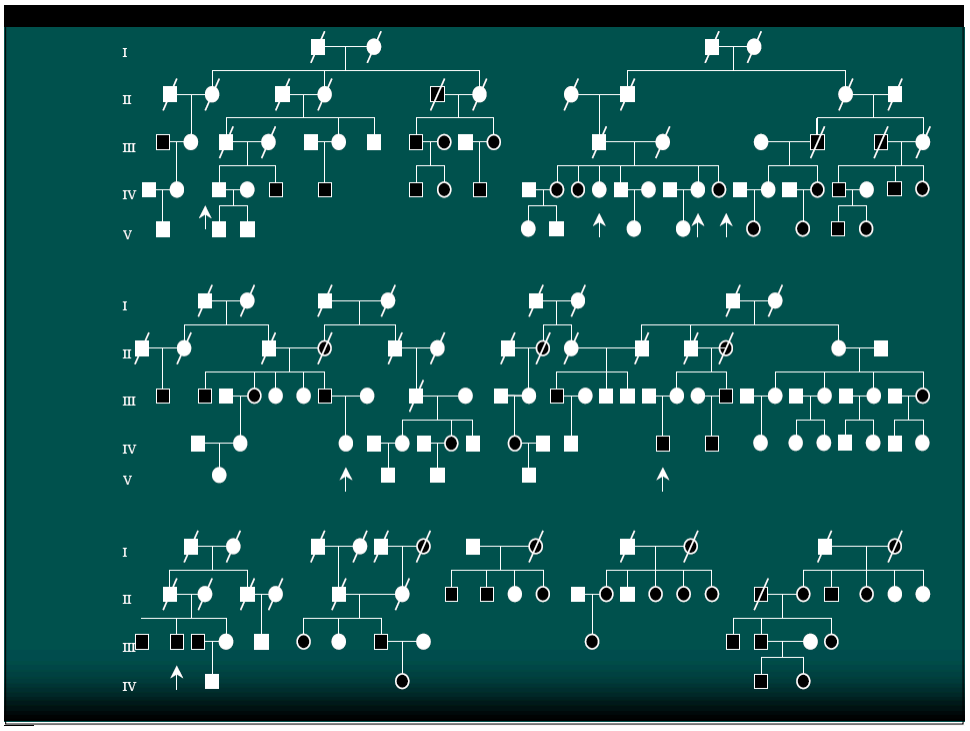
Efforts to Simplify the Complex Genetic Background of Common Diseases

- Familial cases
- Population isolates
- Defined clinical phenotype
- Animal models

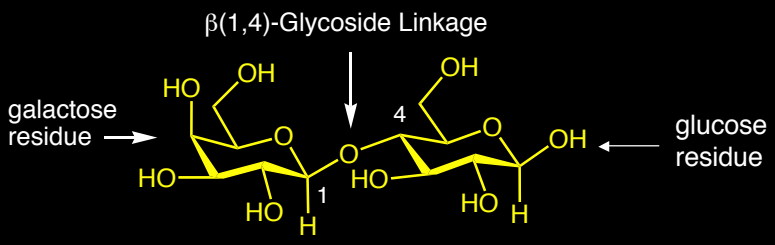
The Effect of Population Bottlenecks to Disease Alleles




Carrier Frequencies
early settlement
late settlement
Helsinki

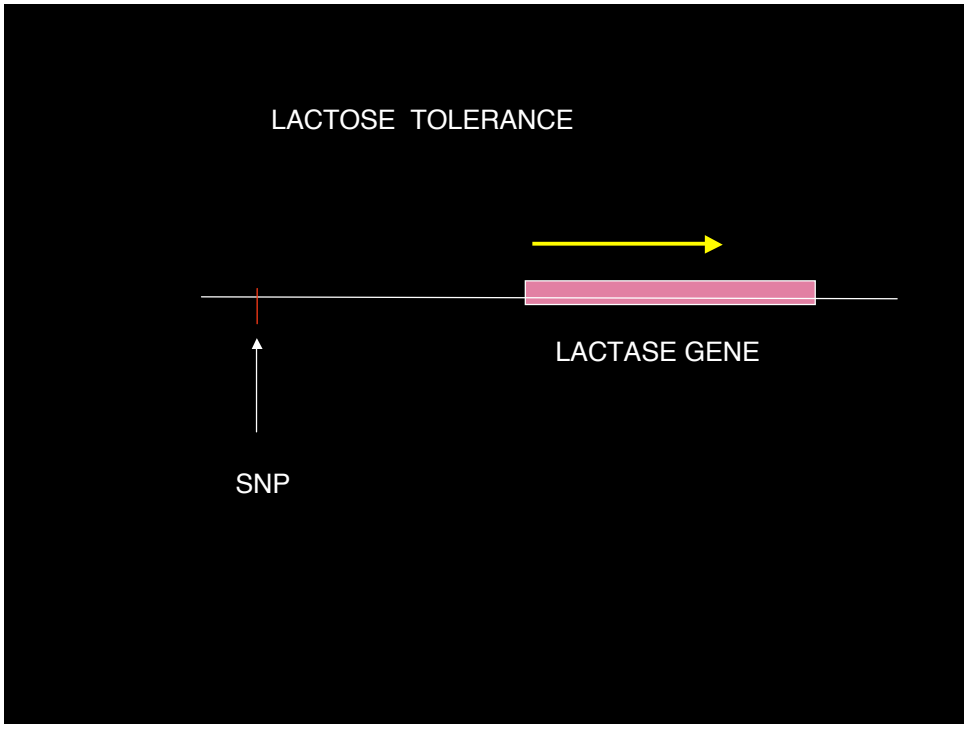

I
II
III
IV
V
I
II
III
IV
V
I
II
III
IV

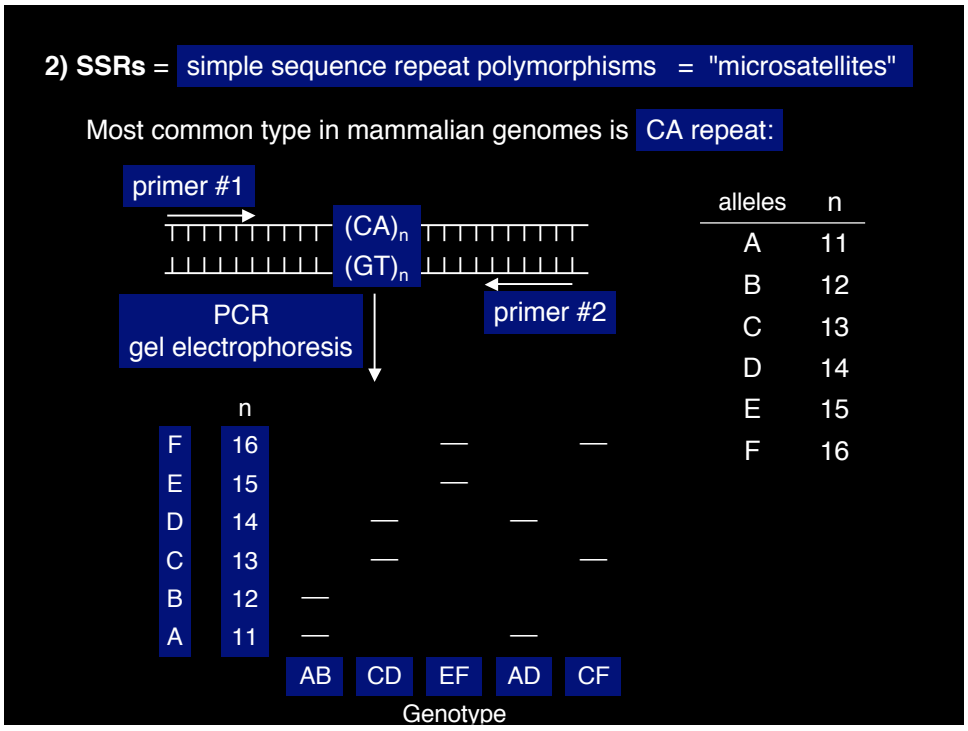
LACTOSE



CANDIDATE GENE

The enzyme lactase that is located in the villus enterocytes of the small intestine is responsible for digestion of lactose in milk. Lactase activity is high and vital during infancy, but in most mammals, including most humans, lactase activity declines after the weaning phase. In other healthy humans, lactase activity persists at a high level throughout adult life, enabling them to digest lactose as adults. This dominantly inherited genetic trait is known as lactase persistence. The distribution of these different lactase phenotypes in human populations is highly variable and is controlled by a polymorphic element *cis*-acting to the lactase gene. A putative causal nucleotide change has been identified and occurs on the background of a very extended haplotype that is frequent in Northern Europeans, where lactase persistence is frequent. This single nucleotide polymorphism is located 14 kb upstream from the start of transcription of lactase in an intron of the adjacent gene *MCM6*. This change does not, however, explain all the variation in lactase expression.


LACTOSE TOLERANCE
LACTASE GENE
SNP


2) SSRs = simple sequence repeat polymorphisms = "microsatellites"
Most common type in mammalian genomes is CA repeat:
primer #1
(CA)n
(GT)n
primer #2
PCR
gel electrophoresis
alleles n
A 11
B 12
C 13
D 14
E 15
F 16
n
F 16
E 15
D 14
C 13
B 12
A 11
AB CD EF AD CF
Genotype

SSRs are extremely useful as genetic markers in human studies because:

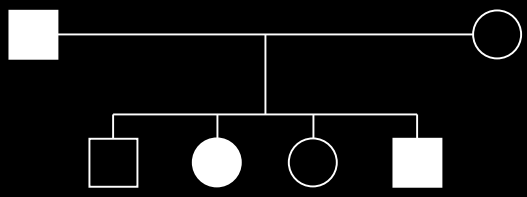
- they are easily scored (by PCR)
- they are codominant
- many SSRs exhibit very high average heterozygosities: H_{SSR} = 0.7 to 0.9

  A randomly selected person is likely to be heterozygous.

- SSRs are abundant

  SSRs occur, on average, about once every 30 kb in the human (or mouse) genomes. > 20,000 SSRs have been identified and mapped within the human genome.

Huntington's disease (HD)

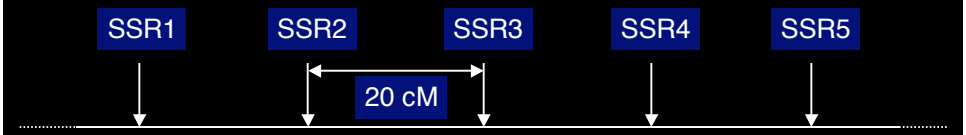


HD: autosomal dominant affecting 1/20,000 individuals

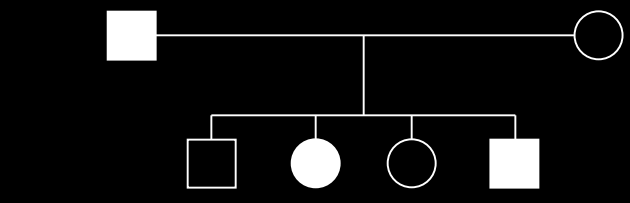
Phenotype: Loss of neurons → personality change, memory loss, motor problem

genetic linkage mapping

We genotype the six members of the family for SSRs scattered throughout the genome (which spans 3300 cM)—perhaps 165 different SSRs distributed at 20 cM intervals so that one SSR must be within 10 cM of the Huntington's gene:



We obtain potentially exciting results with SSR37, on chromosome 4:



Genotypes:	HD	HD/+	+/+	HD/+	+/+	HD/+	+/+
	SSR37	AB	BD	AC	BC	AD	CD
Paternal alleles:	HD		+	HD	+	HD	
	SSR37		B	A	B	A	