PROBLEM SET 7 – MAPPING, DATABASE ANALYSIS, POPULATION GENETICS
(covering weeks 9 & 10)

1. Use the UCSC genome browser to learn more about the region around the HTT gene associated with Huntington's disease.

1a. Pedigree analysis indicated that the gene was located between markers D4S180 and D4S182. In the most recent assembly of the human genome (the Mar 2006 assembly, also known as NCBI build36 or as hg18), what are the physical locations of these markers?

1b. How far apart are these markers?

1c. Based on the RefSeq gene annotations, how many genes are in this interval?

1d. Based on the RefSeq annotations, which (if any) of these genes have alternatively spliced forms?

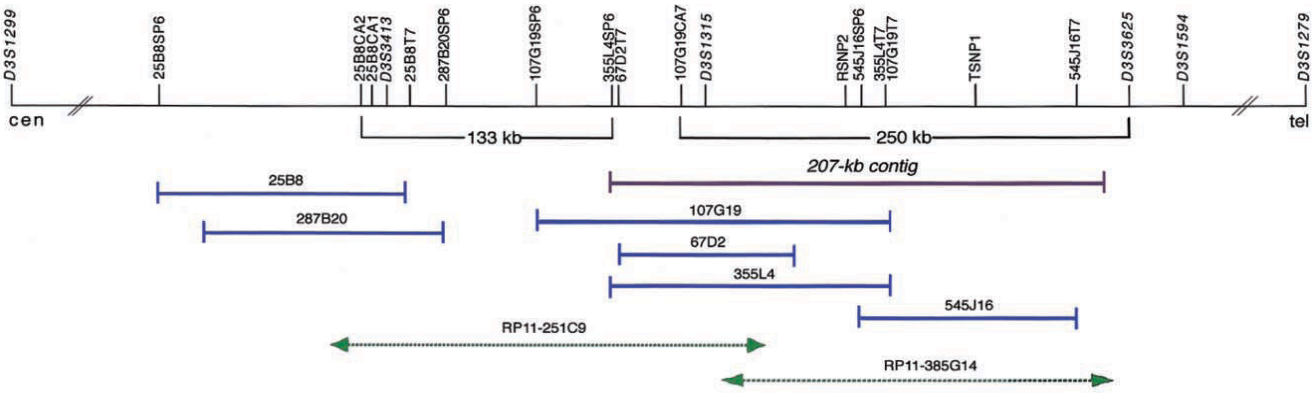
1e. Recall that the International Human Genome Sequencing Consortium used a clone-by- clone approach to sequence and assemble the human genome. The actual tiling path of BAC clones used in the assembly can be displayed by turning the "Assembly" track on in full mode. How many BACs were used to tile across this region?

2. Usher syndrome type 3 (USH3) is an autosomal recessive disorder characterized by progressive hearing loss and severe retinal degeneration. The USH3 gene has been mapped to 3q21-q25. You are researching this rare disease and want to find the causative gene. After fine mapping in this region using affected families, you have refined the location of the USH3 gene to an approximately 400-kb genomic region between markers 25B8CA2 and *D3S3625.*

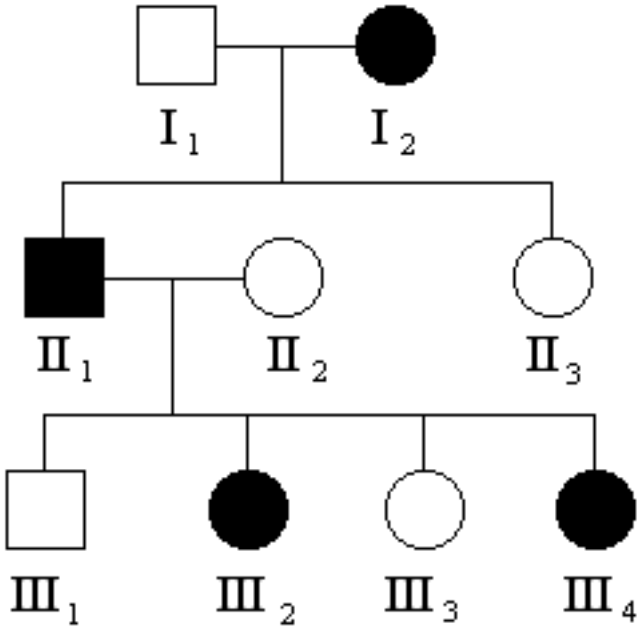
Questions for Thought:
2a. How do you think the flanking markers were previously identified?

2b. How were the fine mapping analyses performed?

2c. Below is a schematic of the genomic region and the contigs you made for positional cloning analysis (make sure you understand the methods for generating a contig). Unfortunately, there are no annotated genes in this region on the UCSC Browser (you checked), but you are sure your gene must be here somewhere. How will you know what genomic sequences could represent genic regions, and more importantly, the USH3 gene? Explain how you would use modern molecular and computational tools to identify the USH3 gene and its association with this disorder.



3. You are studying an autosomal dominant disease trait and have the following family pedigree:



Using 8 polymorphic markers you obtain the following results:

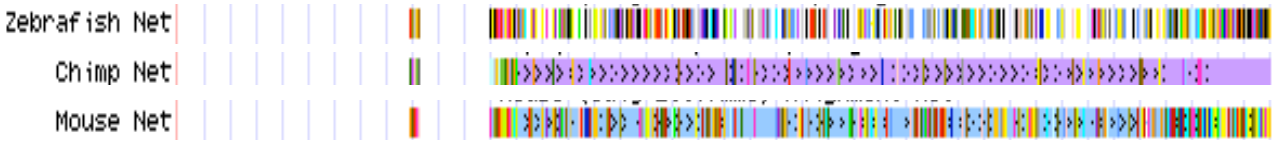
Markers	I1	I2	II1	II2	II3	III1	III2	III3	III4
A	1,3	6,8	3,8	6,7	1,8	3,6	8,7	3,7	8,6
B	5,11	9,4	11,4	10,9	5,4	11,10	4,9	11,9	4,10
C	10,9	10,11	9,10	9,11	10,11	9,9	10,11	9,11	10,9
D	4,5	5,8	5,5	7,7	4,8	5,7	5,7	5,7	5,7
E	3,3	4,5	3,4	5,3	3,4	3,5	4,3	3,5	4,5
F	6,7	6,7	7,6	6,7	6,6	7,6	7,7	6,6	6,6
G	12,10	11,11	10,11	11,12	12,11	10,12	10,12	11,11	11,11
H	9,7	8,6	7,8	5,6	9,8	7,6	7,6	8,5	8,5

3a. How many recombination events can you detect and where do they occur?

3b. Narrow down the region in which you believe the disease trait is located.

3c. How many of the recombinations are helpful (or informative) in identifying the region to which the disease causing allele is located?

4. Shown are synteny maps of three organisms—zebrafish, mouse, and chimp—with the human chromosome 21. Match each of these organisms with their respective synteny map.

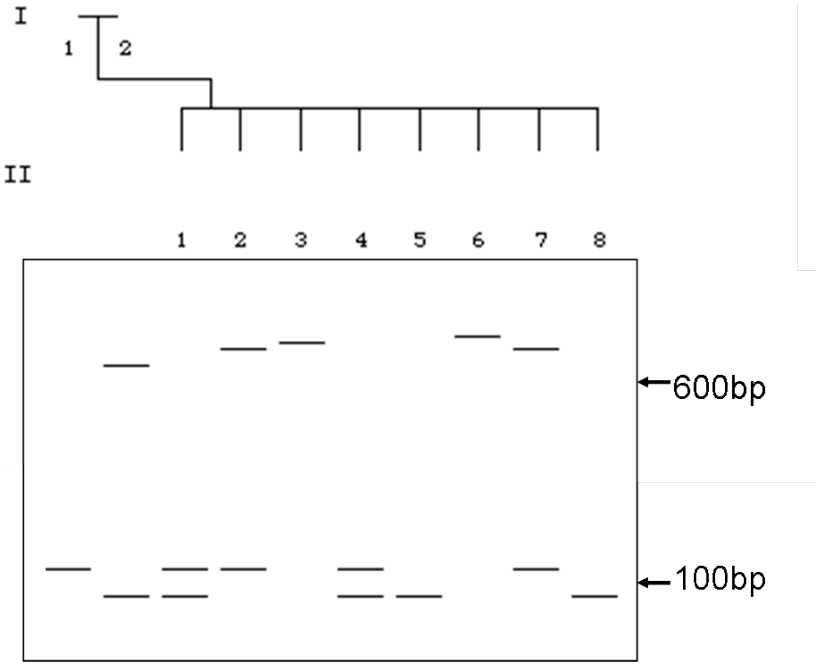


5. Fragile X syndrome in human occurs when there are more than 200 repeats of the CGG trinucleotide in the FMR1 gene. Below is the DNA sequence from the 5' UTR of the FMR1 gene. The unique sequence in the gray boxes can be used to design PCR primers to amplify the trinucleotide repeat region.

5' GGTGACGGAGGCGCCGCTGCCAGG—(CGG)n—CCTCGAGCGCCCGCAGCCCACCTCT 3'
3' CCACTGCCTCCGCGGCGACGGTCC—(GCC)n—GGAGCTCGCGGGCGTCGGGTGGAGA 5'

5a. Show the sequences of a set of PCR primers that could be used to amplify the FMR1 repeat.

5b. Shown below is the pedigree for a family segregating fragile X syndrome and a gel showing PCR amplification fragments of each individual in the pedigree. Each person's DNA sample is shown directly below that person. Based on the result of PCR fill in the pedigree to show the phenotype and gender of each individual.



6a. For the following pairs of aligned sequences, determine the alignment score (log-odds table on last page).

GWTQLPE **KQRAAGLIV**
GFSNEPE **RERAVGVVV**

6b. Given the amino acid frequencies in proteins (amino acid frequency table on the last page) compute how often the indicated pairs of amino acids are found aligned in the known related proteins used to compute the log-odds table (last page):

A aligned with A:
A aligned with E:
F aligned with Y:

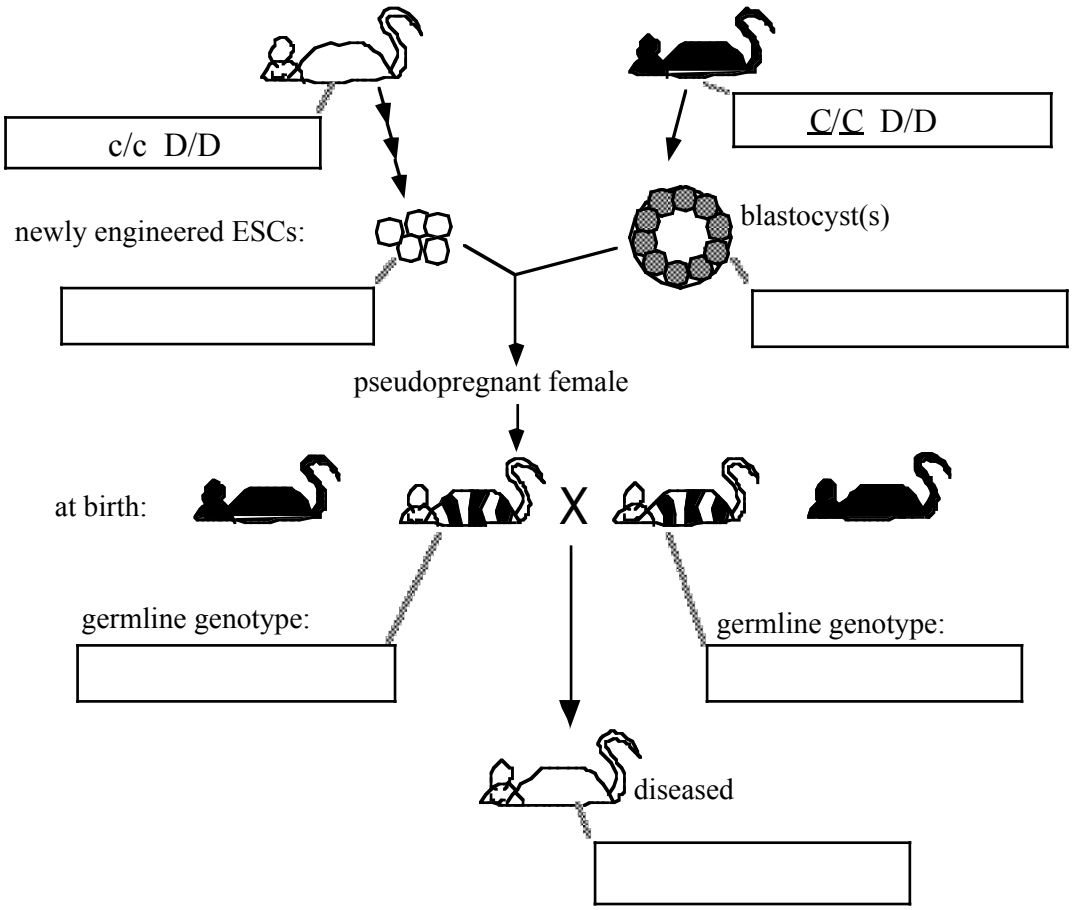
6c. In a large set of random protein sequences each with a length of 5 amino acids, how frequently would the specific sequence **GATLP** appear?

7. By hand, find what you think the optimal alignment of the following pair of protein sequences is. Introduce gaps ('-') as needed.

CGATWQMNPLTSWRALA
CGAWQMNPITSWRRALA

8. To make a mouse "model" for a **recessive** inherited human disease, embryonic stem cells (ESCs) made by mating two **fully homozygous** white (albino) non-disease mice are used. One of the normal alleles (**D**) of the mouse "disease" gene is replaced by an inactivated allele (**d**). The modified ESCs are then injected into mouse blastocysts made by mating two **fully homozygous** black non-disease mice. Black coat color is conferred by an allele (**C**) that is dominant to the allele causing albinism (**c**). The resulting embryos are implanted in a pseudopregnant female (this is a fancy term for a surrogate mother) and allowed to develop. (See the drawings below).

Some of the mice from these embryos are coat color chimeras. Two of these black/white mice are mated to each other. One offspring is found that has a completely white (albino) coat and shows the disease trait! (See the drawings below). In each of the boxes in the drawing below, write in the genotypes **for the coat color gene** and **for the disease gene**.



9. Refer to the simulation graphs below to answer these questions. Each graph shows multiple independent allele frequency simulations, with each in a different color. If you don't have a color figure you should still be able to make out lots of gray-shade lines.

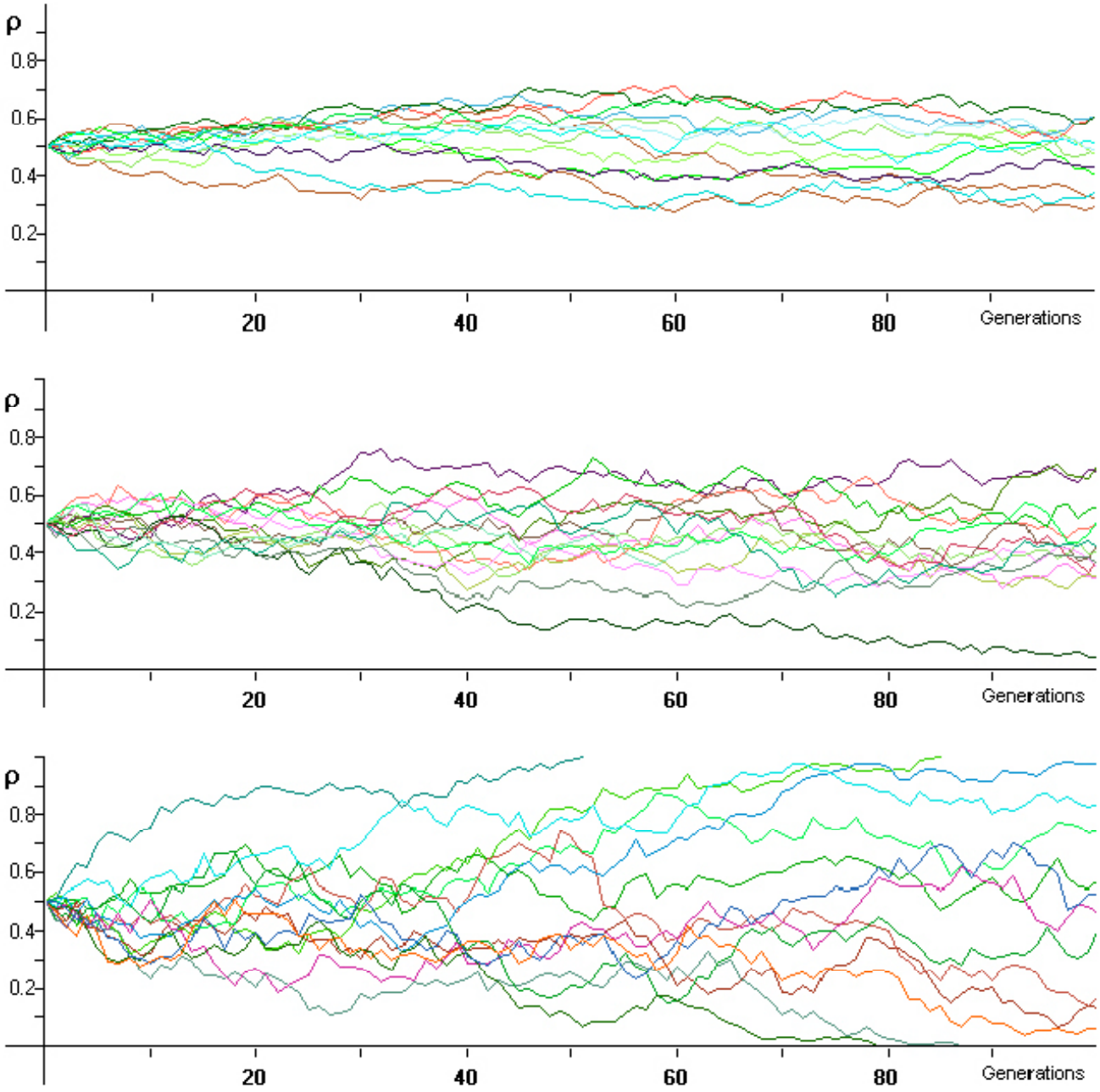
9a. Do you think there is any selection happening in the simulations?

9b. Between the top two simulations, which one was run with a smaller population size?

9c. If the population size in the first simulation were infinitely large, where would the allele frequency

line go?

9d. In the third simulation something special happened with the run that appears as the topmost line (if you can see color it is a sort of dark aqua). What happened? What is this called?



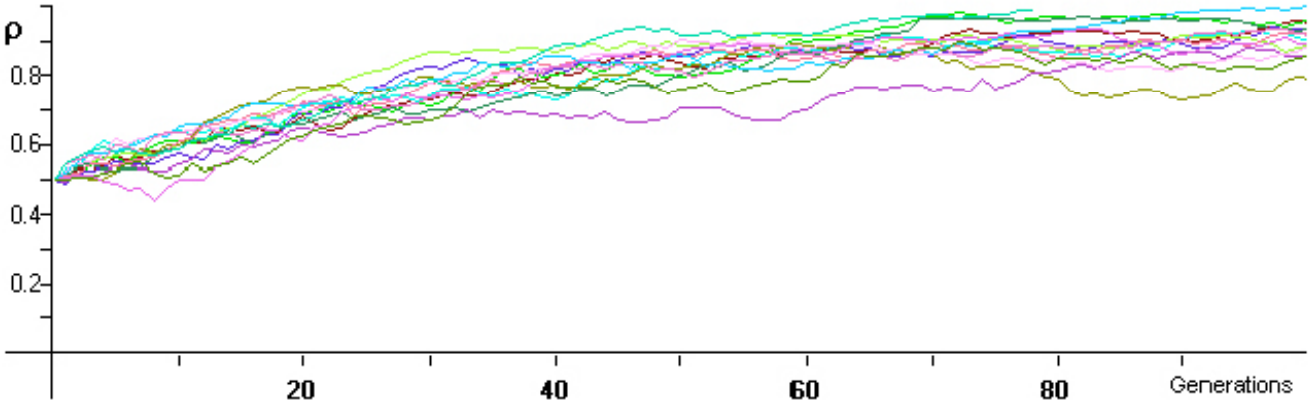
10. For the first two parts to this question refer to the simulation graphs below. Both simulations used the same population size and involved selection. As in class, the value p indicates the frequency of the A allele (the other allele a is always at frequency 1 – p).

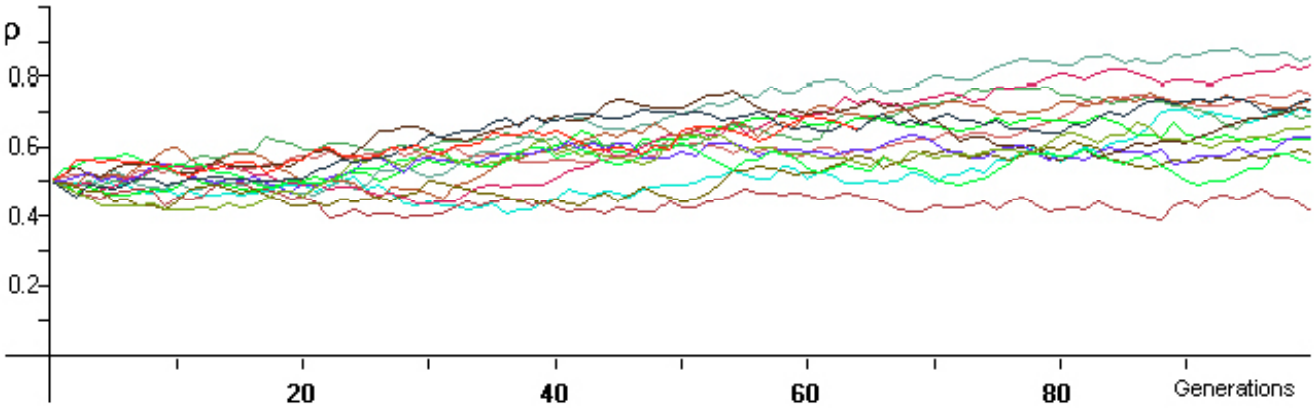
10a. Which allele is at a selective advantage in both simulations? Explain.

10b. In the lower simulation, how can the lowermost line drop below the starting A frequency of 0.5?

10c. (not refering to the two simulation graphs) If the a allele is completely recessive and homozygotes die at birth, what fraction of progeny will fail to reproduce at each generation for each of the following a allele frequencies? What consequence does this pattern have for recessive human disease alleles?

a allele frequency	**Fail to reproduce**
0.1	
0.01	
0.001	
0.0001	





AMINO ACID FREQUENCY TABLE:

amino acid	one-letter	frequency	percent
alanine	A	0.0768	7.68
cysteine	C	0.0162	1.62
aspartate	D	0.0526	5.26
glutamate	E	0.0648	6.48
phenylalanine	F	0.0409	4.09
gylcine	G	0.0689	6.89
histidine	H	0.0225	2.25
isoleucine	I	0.0586	5.86
lysine	K	0.0596	5.96
leucine	L	0.0958	9.58
methionine	M	0.0236	2.36
asparagine	N	0.0435	4.35
proline	P	0.0490	4.90
glutamine	Q	0.0394	3.94
arginine	R	0.0521	5.21
serine	S	0.0700	7.00
threonine	T	0.0558	5.58
valine	V	0.0663	6.63
tryptophan	W	0.0121	1.21
tyrosine	Y	0.0315	3.15
		1.0000	100.00

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	-2	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E	-1	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F	-2	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H	-2	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I	-1	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L	-1	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1
N	-2	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-2
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-3
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-2
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-2
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	2
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7

PROBLEM SET 7 – MAPPING, DATABASE ANALYSIS, POPULATION GENETICS (covering weeks 9 & 10)

1. Use the UCSC genome browser to learn more about the region around the HTT gene associated with Huntington's disease.

1a. Pedigree analysis indicated that the gene was located between markers D4S180 and D4S182. In the most recent assembly of the human genome (the Mar 2006 assembly, also known as NCBI build36 or as hg18), what are the physical locations of these markers?

D4S180 chr4:3270487-3270920 4p16.2
D4S182 chr4:2783123-2783337 4p16.3

1b. How far apart are these markers?

487,798bp

1c. Based on the RefSeq gene annotations, how many genes are in this interval?

7 different Refseq genes

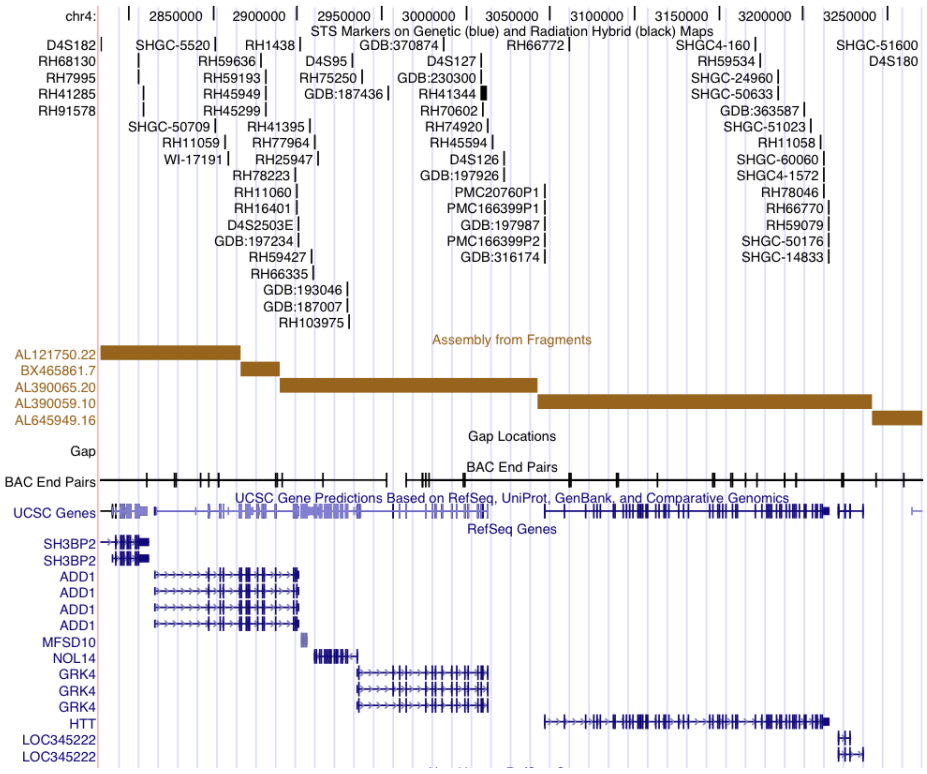
1d. Based on the RefSeq annotations, which (if any) of these genes have alternatively spliced forms?

SH3BP2, ADD1, GRK4, LOC345222

1e. Recall that the International Human Genome Sequencing Consortium used a clone-by- clone approach to sequence and assemble the human genome. The actual tiling path of BAC clones used in the assembly can be displayed by turning the "Assembly" track on in full mode. How many BACs were used to tile across this region?

5 BACs

Here is a browser snapshot to help.



2. Usher syndrome type 3 (USH3) is an autosomal recessive disorder characterized by progressive hearing loss and severe retinal degeneration. The USH3 gene has been mapped to 3q21-q25. You are researching this rare disease and want to find the causative gene. After fine mapping in this region using affected families, you have refined the location of the USH3 gene to an approximately 400-kb genomic region between markers 25B8CA2 and *D3S3625.*

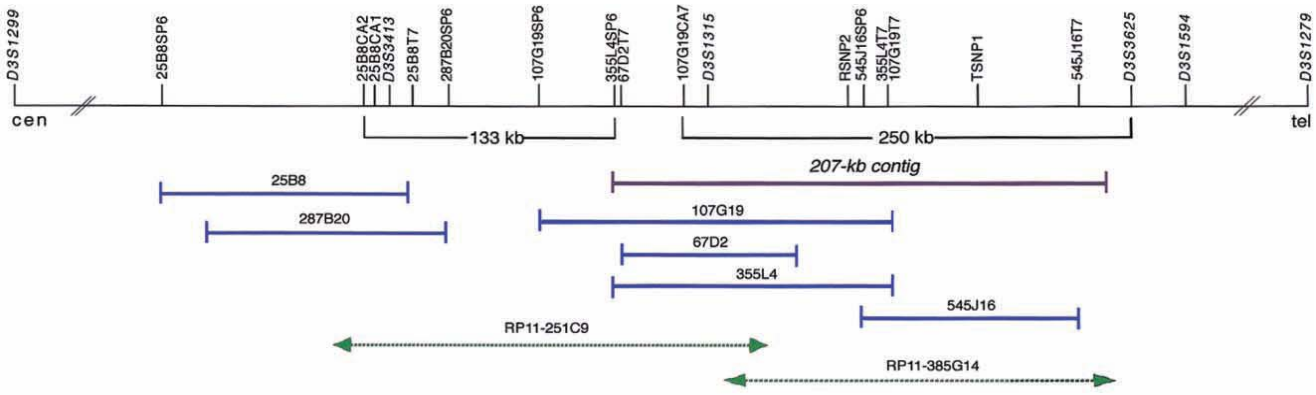
Questions for Thought:
2a. How do you think the flanking markers were previously identified?

These are unique regions in the genome that would have been identified experimentally using a variety of techniques (RFLPs, microsatellites, different probe hybridizations, etc.)

2b. How were the fine mapping analyses performed?

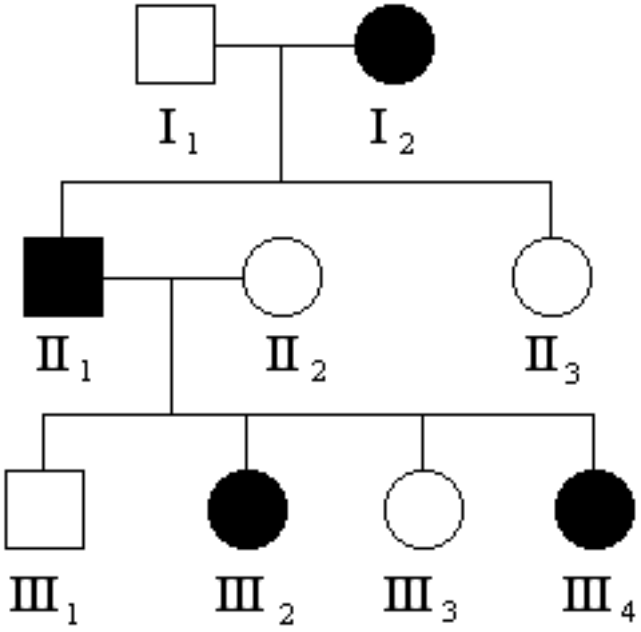
For fine mapping you will narrow down the region using more markers and LOD scores. Using this information you would build haplotype maps. This should allow you to identify a much smaller region. Once this is done you would look for genes in this region to test and see whether or not they are your gene of interest.

2c. Below is a schematic of the genomic region and the contigs you made for positional cloning analysis (make sure you understand the methods for generating a contig). Unfortunately, there are no annotated genes in this region on the UCSC Browser (you checked), but you are sure your gene must be here somewhere. How will you know what genomic sequences could represent genic regions, and more importantly, the USH3 gene? Explain how you would use modern molecular and computational tools to identify the USH3 gene and its association with this disorder.



Look for conserved promoter, enhancer or any other sort of conserved elements.
Identify ORFs
Comparative genomics (compare your region to possible genic regions in other genome sequences.
BLAST search.

3. You are studying an autosomal dominant disease trait and have the following family pedigree:



Using 8 polymorphic markers you obtain the following results:

Markers	I1	I2	II1	II2	II3	III1	III2	III3	III4
A	1,3	6,8	3,8	6,7	1,8	3,6	8,7	3,7	8,6
B	5,11	9,4	11,4	10,9	5,4	11,10	4,9	11,9	4,10
C	10,9	10,11	9,10	9,11	10,11	9,9	10,11	9,11	10,9
D	4,5	5,8	5,5	7,7	4,8	5,7	5,7	5,7	5,7
E	3,3	4,5	3,4	5,3	3,4	3,5	4,3	3,5	4,5
F	6,7	6,7	7,6	6,7	6,6	7,6	7,7	6,6	6,6
G	12,10	11,11	10,11	11,12	12,11	10,12	10,12	11,11	11,11
H	9,7	8,6	7,8	5,6	9,8	7,6	7,6	8,5	8,5

3a. How many recombination events can you detect and where do they occur?

I2 between B and C
I2 between D and E
II1 between E and F (2 cases)
II2 between C (or possibly D) and E
II2 between F and G

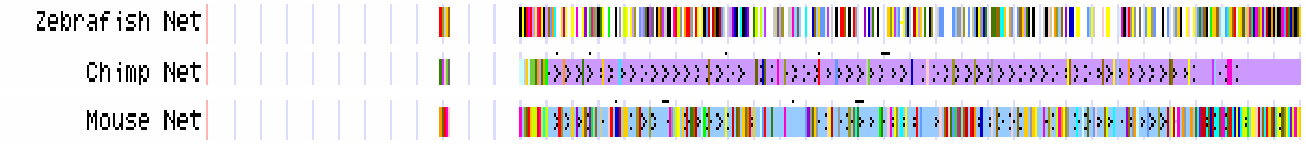
3b. Narrow down the region in which you believe the disease trait is located.

Between markers B and E

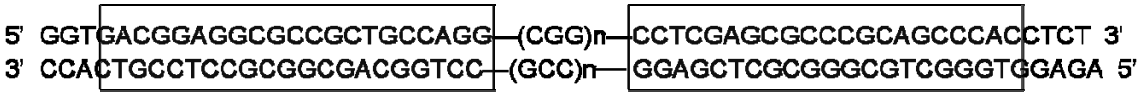
3c. How many of the recombinations are helpful (or informative) in identifying the region to which the disease causing allele is located?

4 (I2 between B and C, both II1 between E and F, and I2 D and E)

4. Shown are synteny maps of three organisms—zebrafish, mouse, and chimp—with the human chromosome 21. Match each of these organisms with their respective synteny map.



5. Fragile X syndrome in human occurs when there are more than 200 repeats of the CGG trinucleotide in the FMR1 gene. Below is the DNA sequence from the 5' UTR of the FMR1 gene. The unique sequence in the gray boxes can be used to design PCR primers to amplify the trinucleotide repeat region.

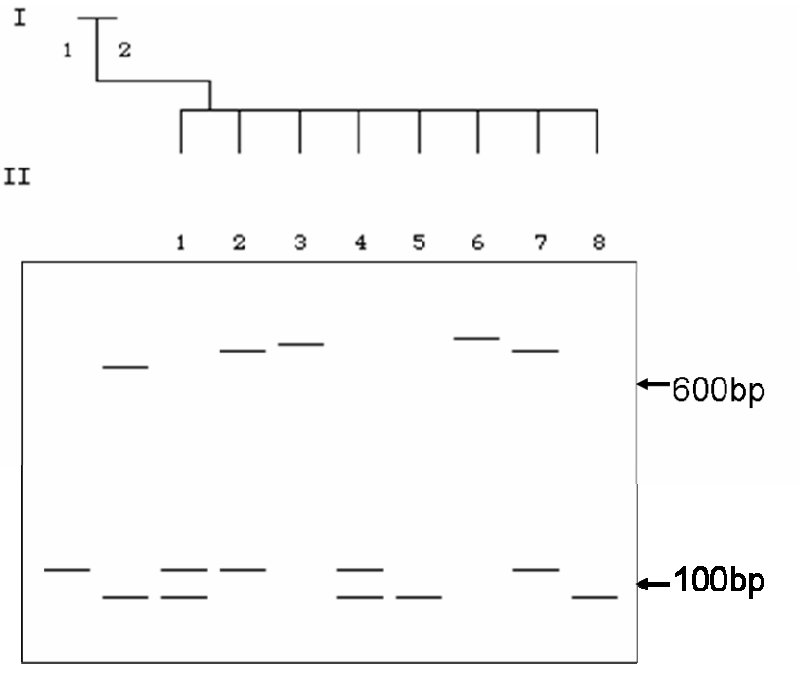


5a. Show the sequences of a set of PCR primers that could be used to amplify the FMR1 repeat.

For the 5' or forward or left primer you would pick ~20bp from this sequence:
5'-GACGGAGGCGCCGCTGCCAGG-3'

For the 3' or reverse or right primer you would pick ~20bp from this sequence:
5'-GTGGGCTGCGGGCGCTCGAGG-3'

5b. Shown below is the pedigree for a family segregating fragile X syndrome and a gel showing PCR amplification fragments of each individual in the pedigree. Each person's DNA sample is shown directly below that person. Based on the result of PCR fill in the pedigree to show the phenotype and gender of each individual.



I1 male wildtype
I2 female fragile X
II1 female wildtype
II2 female fragile X
II3 male fragile X
II4 female wildtype

II5 male wildtype
II6 male fragile X
II7 female fragile X
II8 male wildtype

The females phenotypic symptoms can vary between mild and more severe.

6a. For the following pairs of aligned sequences, determine the alignment score (log odds table on last page).

GWTQLPE	**KQRAAGLIV**
GFSNEPE	**RERAVGVVV**
6+1+1+0-3+7+5=17	2+2+5+4+0+6+1+3+4=27

6b. Given the amino acid frequencies in proteins (amino acid frequency table on the last page) compute how often the indicated pairs of amino acids are found aligned in the known related proteins used to compute the log odds table (last page):

A aligned with A: 0.02359
A aligned with E: 0.00704
F aligned with Y: 0.00364

6c. In a large set of random protein sequences each with a length of 5 amino acids, how frequently would the specific sequence **GATLP** appear?

$0.0689*0.0768*0.0558*0.0958*0.0490=1.39x10^{-6}$

7. By hand, find what you think the optimal alignment of the following pair of protein sequences is. Introduce gaps (' ') as needed.

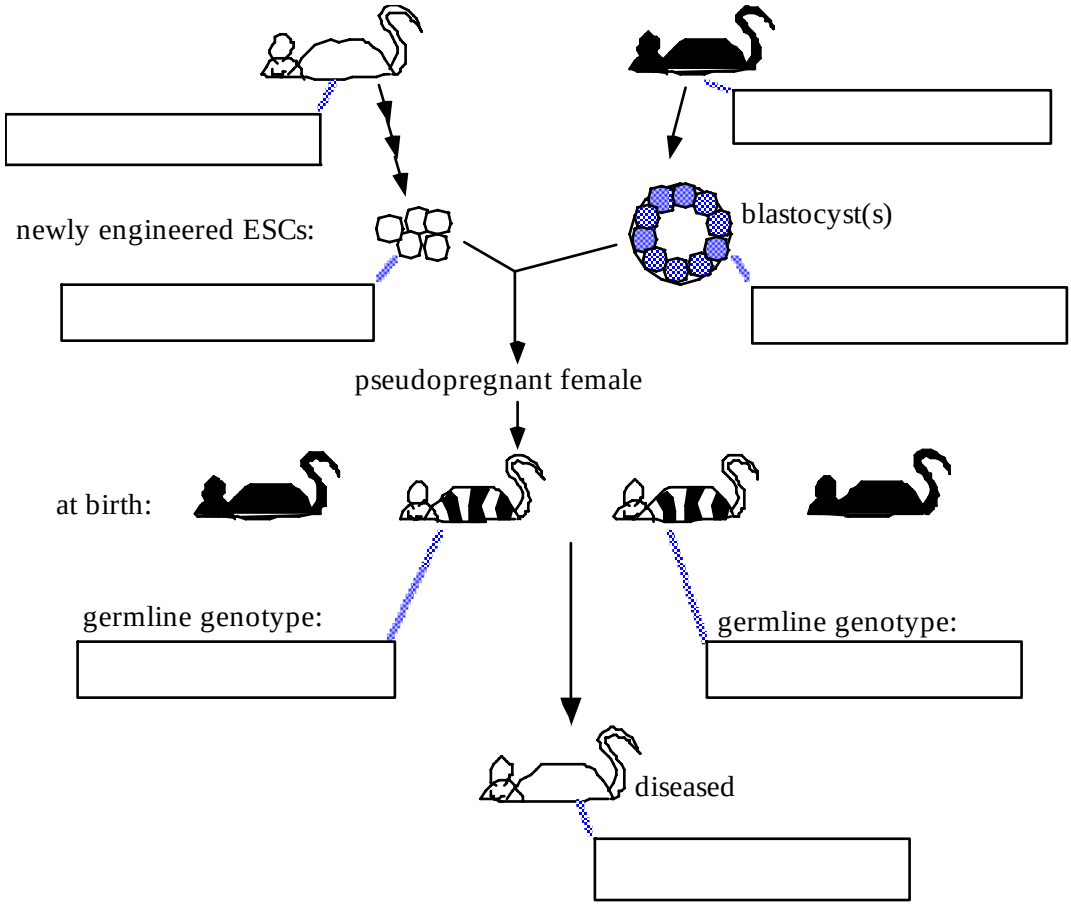
CGATWQMNPLTSWRALA
CGAWQMNPITSWRRALA

CGATWQMNPLTSWR-ALA
CGA-WQMNPITSWRRALA

9+6+4+0+11+5+5+6+7+2+5+4+11+5+0+4+4+4=92

8. To make a mouse "model" for a **recessive** inherited human disease, embryonic stem cells (ESCs) made by mating two **fully homozygous** white (albino) non-disease mice are used. One of the normal alleles (**D**) of the mouse "disease" gene is replaced by an inactivated allele (**d**). The modified ESCs are then injected into mouse blastocysts made by mating two **fully homozygous** black non-disease mice. Black coat color is conferred by an allele (**C**) that is dominant to the allele causing albinism (**c**). The resulting embryos are implanted in a pseudopregnant female (this is a fancy term for a surrogate mother) and allowed to develop. (See the drawings below).

Some of the mice from these embryos are coat color chimeras. Two of these black/white mice are mated to each other. One offspring is found that has a completely white (albino) coat and shows the disease trait! (See the drawings below)**.** In each of the boxes in the drawing below, write in the genotypes **for the coat color gene** and **for the disease gene**.



Newly engineered ESCs – c/c d/D
Blastocyst(s) – C/C D/D
Both germline genotypes – c/c d/D
Diseased – c/c d/d

9. Refer to the simulation graphs below to answer these questions. Each graph shows multiple independent allele frequency simulations, with each in a different color. If you don't have a color figure you should still be able to make out lots of gray shade lines.

9a. Do you think there is any selection happening in the simulations?

There is no observable selection in any of the simulations (perhaps very weak selection, but it could be explained by genetic drift).

9b. Between the top two simulations, which one was run with a smaller population size?

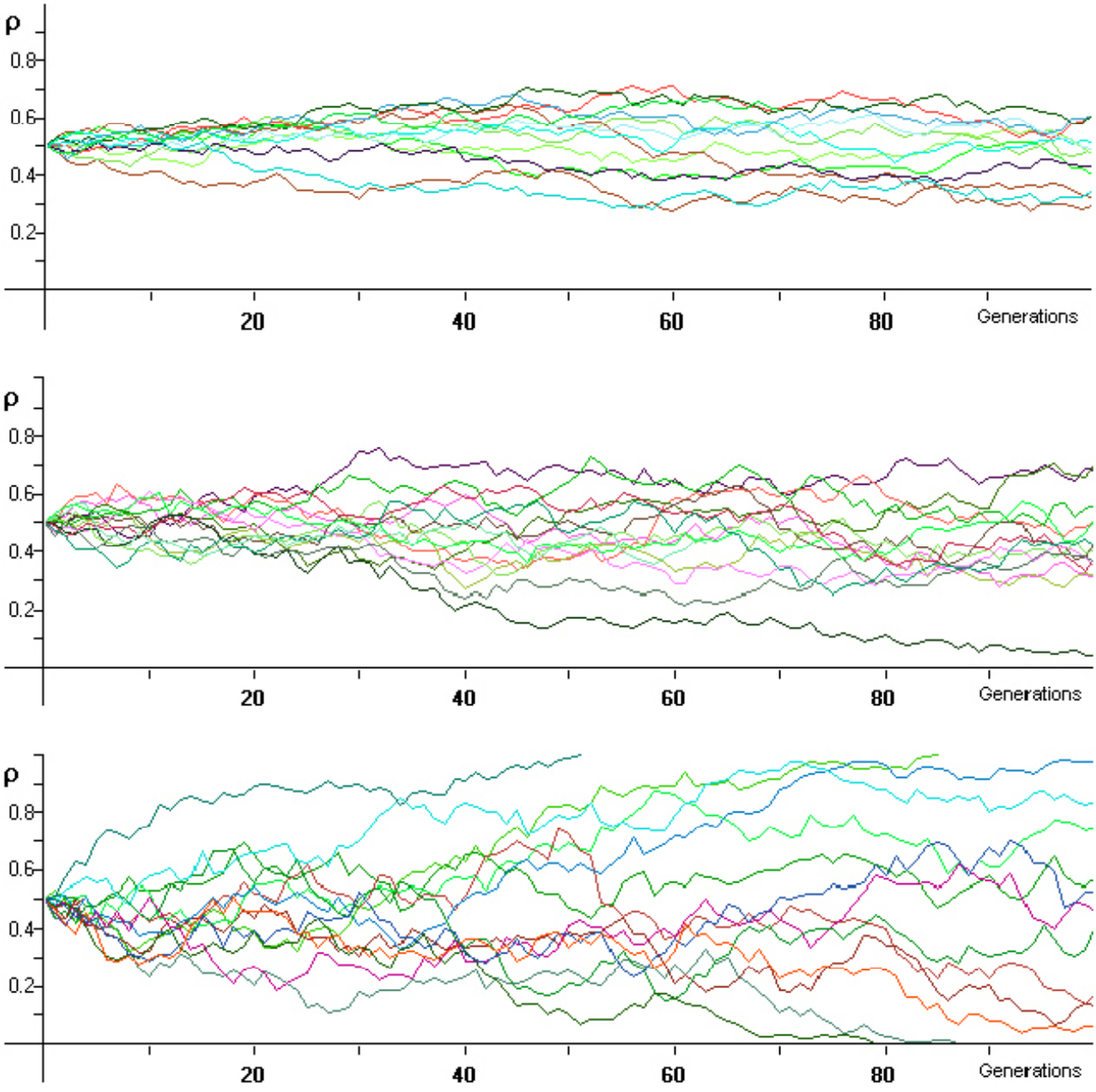
The second simulation.

9c. If the population size in the first simulation were infinitely large, where would the allele frequency line go?

You would expect it to remain at the p=0.5 line.

9d. In the third simulation something special happened with the run that appears as the topmost line (if you can see color it is a sort of dark aqua). What happened? What is this called?

The A allele became fixed (reached p=1.0) and therefore the a allele became extinct.



10. For the first two parts to this question refer to the simulation graphs below. Both simulations used the same population size and involved selection. As in class, the value p indicates the frequency of the A allele (the other allele a is always at frequency 1 p).

10a. Which allele is at a selective advantage in both simulations? Explain.

The A allele, though it appears to be undergoing much stronger selection in the first simulation (or you could look at it as weak selection in the second simulation).

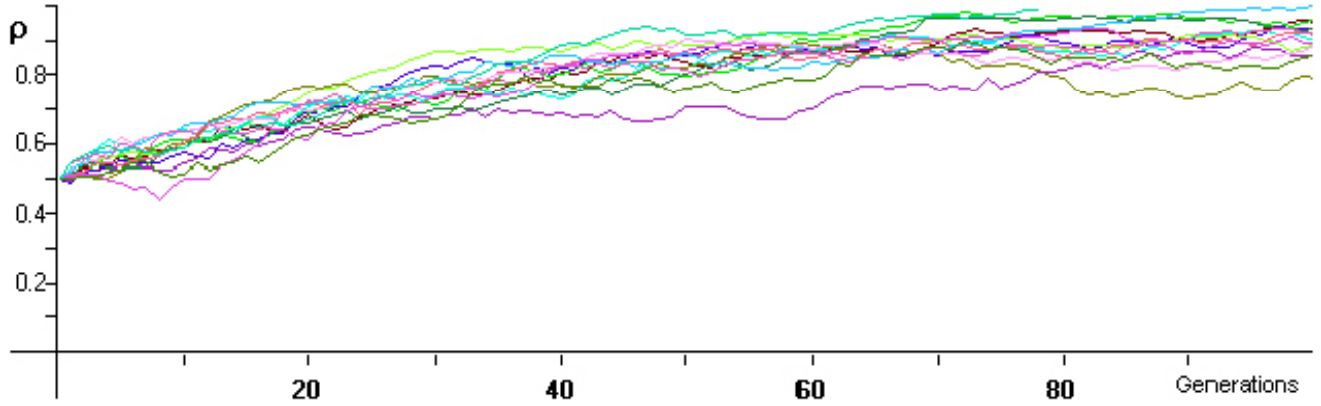
10b. In the lower simulation, how can the lowermost line drop below the starting A frequency of 0.5?

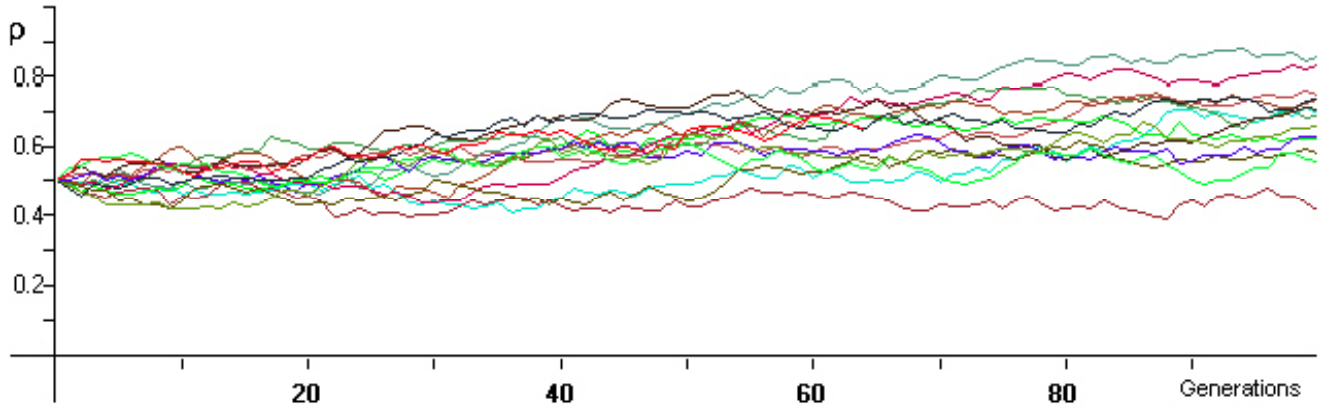
Random drift was able to favor the a allele more than the weak selection acting on the A allele.

10c. (not refering to the two simulation graphs) If the a allele is completely recessive and homozygotes die at birth, what fraction of progeny will fail to reproduce at each generation for each of the following a allele frequencies? What consequence does this pattern have for recessive human disease alleles?

a allele frequency	**Fail to reproduce**
0.1	0.01
0.01	0.0001
0.001	0.000001
0.0001	0.00000001

Since all aa homozygoud individuals will die the a allele can only exist in the population as a heterozygote. Unless there is some strong selection for the a allele it will likely exist at a lower frequency than the A.





AMINO ACID FREQUENCY TABLE:

amino acid	one-letter	frequency	percent
alanine	A	0.0768	7.68
cysteine	C	0.0162	1.62
aspartate	D	0.0526	5.26
glutamate	E	0.0648	6.48
phenylalanine	F	0.0409	4.09
gylcine	G	0.0689	6.89
histidine	H	0.0225	2.25
isoleucine	I	0.0586	5.86
lysine	K	0.0596	5.96
leucine	L	0.0958	9.58
methionine	M	0.0236	2.36
asparagine	N	0.0435	4.35
proline	P	0.0490	4.90
glutamine	Q	0.0394	3.94
arginine	R	0.0521	5.21
serine	S	0.0700	7.00
threonine	T	0.0558	5.58
valine	V	0.0663	6.63
tryptophan	W	0.0121	1.21
tyrosine	Y	0.0315	3.15
		1.0000	100.00

OG ODDS TABLE:

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	-2	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E	-1	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F	-2	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H	-2	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I	-1	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L	-1	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1
N	-2	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-2
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-3
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-2
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-2
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	2
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7