**Name:**_____

**Bio393**

**Genetic Analysis**

**Final exam**

**Wednesday June 10, 2015**

Graded exams will be available after 9 AM on Friday June 12th outside Cook 3125. If you have any questions about the grading of the exam, return your exam with a written explanation by Monday June 15th. The grade distribution will be available on the course website.

Thank you for a fun quarter. Enjoy your summer break or your next adventure!

**Question 1 (5 points):**
You are studying a dominant Mendelian disease via linkage analysis and are focusing on a single marker. Two large families have been genotyped at the same marker and scored for the disease.

In Family I, ten offspring are genotyped: eight children inherited the marker locus and disease locus without recombination; two children appear to be recombinants. You test many values of the recombination fraction (theta) and discover that theta = 0.2 gives the maximum odds ratio, which is 6.87 (LOD = 0.837).

In Family II, 20 offspring are genotyped: 17 children inherited the marker locus and disease locus without recombination; three children appear to be recombinants. You test many values of theta and discover that theta = 0.15 gives the maximum odds ratio, which is 223.4 (LOD = 2.34).

To combine data across Family I and Family II, you multiply odds ratios (add LOD scores). The final estimate of the odds of linkage relative to the null as 1534.8 (LOD = 3.18). Explain what is wrong with this calculation.

*You can not combine LOD scores when they are calculated using different values of theta.*

**Question 2 (6 points):**
Imagine you are doing a genome-wide linkage study in Finnish families looking for the genetic determinants of blood pressure in humans. You have five multi-generational families; each individual is genotyped at 1000 markers and his/her blood pressure is measured. A recent, published study in Icelandic families identified a highly significant locus on chromosome 10 responsible for blood pressure variation. You look through your results and see no significant linkage between the genotype and the disease in your data. Your nearest marker to this locus is 30 cM away.

Give three reasons why you might have failed to find linkage to the chromosome 10 locus. Please explain each reason with one or two sentences.

*(1) Locus heterogeneity: the chromosome 10 locus is not polymorphic in the Finnish families and not responsible for disease variation.*
*(2) Complexity: the chromosome 10 locus is polymorphic in your family but there are other causative loci with greater effect.*
*(3) Marker spacing: your closest marker is too far away to give an odds of linkage that can be distinguished from the null hypothesis.*
*(4) Number of patients: your number of individuals in the Finnish families is too small to get significant linkage.*
*(5) Number of markers: because of multiple testing of 1000 markers, your significance threshold is too high to detect linkage*
*(6) Environment: the individuals in your families have very different lifestyle habits and this environmental variation trumps modest genetic effects.*

**Question 3 (6 points):**
Describe three factors that contribute to the ability to detect QTL using genome-wide association studies?

*(1) Number of individuals. The more, the better*

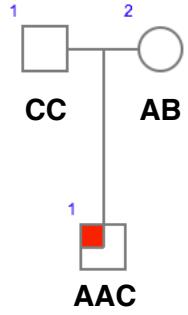
*(2) Effects of underlying genetic variants. The more significant the effect of the variant on the phenotype the better the mapping.*

*(3) Allele frequency. The closer the variant is to 50% in the population the more statistical power to detect a significant difference exists.*

*(4) Linkage disequilibrium. Allows researchers to map using tag-SNPs, giving two advantages: 1. do fewer statistical tests, 2. cheaper to genotype fewer markers*

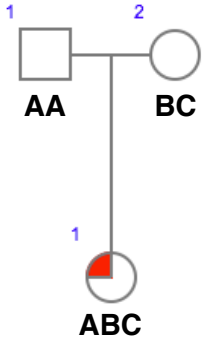
**Question 4 (4 points each):**
Chromosomal abnormalities cause a large fraction of aborted pregnancies and severe developmental disorders. For the following two parts, please describe what caused the inheritance of the extra chromosome in the affected child?

**(a)** Down syndrome is caused by inheritance of an extra copy of chromosome 21. A marker on chromosome 21 was genotyped in both parents and child shown below.




*The extra chromosome was caused by non-disjunction during meiosis II. Sister chromatids harboring the A allele of the chromosome 21 marker failed to go to separate gametes.*

**(b)** Patau syndrome is caused by inheritance of an extra copy of chromosome 13. A marker on chromosome 13 was genotyped in both parents and child shown below.




*The extra chromosome was caused by non-disjunction during meiosis I. Homologous chromosomes failed to go to separate gametes, leading to both the A and B alleles of the chromosome 13 marker being inherited.*

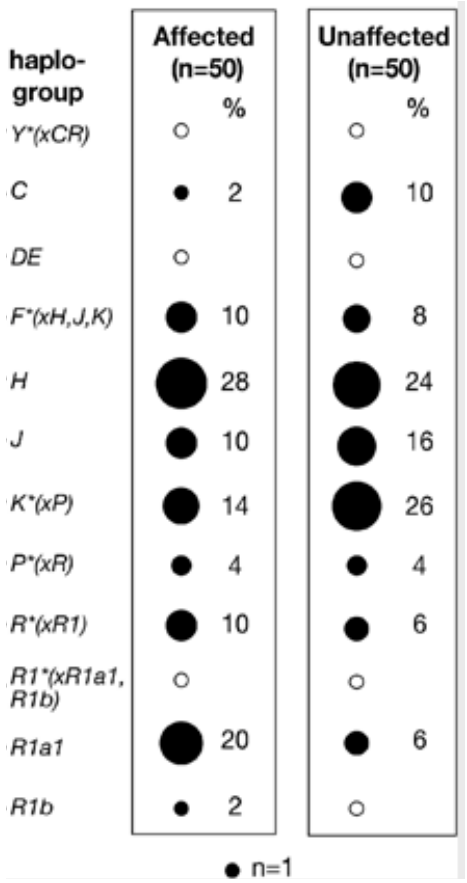
**Question 5 (4 points each):**
**(a)** What properties of the human genome contribute to the presence of haplotypes?

*Linkage disequilibrium and its opposing force recombination shape haplotypes.*

**(b)** What properties of human history contribute to the presence of haplotypes?

*Migration out of Africa, inbreeding in local regions, and bottlenecks all influence and shape haplotypes.*

**Question 7 (6 points):**
The hairy ears trait has long fascinated geneticists, from using evidence from the earliest Italian pedigree in 1907 showing male transmittance to the molecular investigation of southern Indian males in 2004. In the most recent study, 50 affected and 50 unaffected (by visual inspection) southern Indian males were genotyped for markers on the Y chromosome. Twelve distinct Y chromosome haplotypes (haplogroups) were observed in both populations. Please describe whether these data argue for Y linkage and how you arrived at your decision.




*In order for hairy ears to be linked to the Y chromosome, affected males would have to have Y chromosomes not often found in the rest of the population. In other words, affected males would likely share a common Y chromosome haplotype or have a strong correlation between Y chromosome genotype and hairy ears phenotype.*

*Comparison of Y chromosome haplotypes between affected and unaffected males shows little difference in haplotype frequencies. Therefore, unless hairy ears arose independently in many unique Y chromosome haplotypes, it is likely not Y-linked.*

**Question 7 (17 points):**
A psychologist friend of yours heard you studying association analyses for the Bio393 final. He suggested that you help to genetically map whether a student will attend Dillo Day for his senior thesis. He collected DNA and genotyped 1000 Northwestern students who went to Dillo Day all four years (cases) and 1000 students who never went during their four years (controls). One marker is nearby the widely purported "Dillo gene" with the following distribution in genotypes among the students: Cases = 345 AA, 575 AG, 80 GG; Controls = 275 AA, 500 AG, 225 GG.

**(a, 4 points)** Fill out the 2x2 contingency table for this association test for him.

| | | Cases | Controls |
|---|---|---|---|
| Genotype | A | *1265* | *1050* |
| | G | *735* | *950* |

**(b, 4 points)** This chi-squared test returned a *p*-value of 7.2E-12, which is significant given multiple testing correction. Why do we have to do a correction for multiple statistical tests?

*A p-value is the probability that you will reject the null hypothesis when it should otherwise be accepted. In other words, you falsely conclude that your experimental condition is true instead of false that fraction of the time. For a large number of tests, you would reject the null hypothesis many times falsely. A multiple testing correction prevents us from falsely concluding that a marker is correlated with a disease-causing allele.*

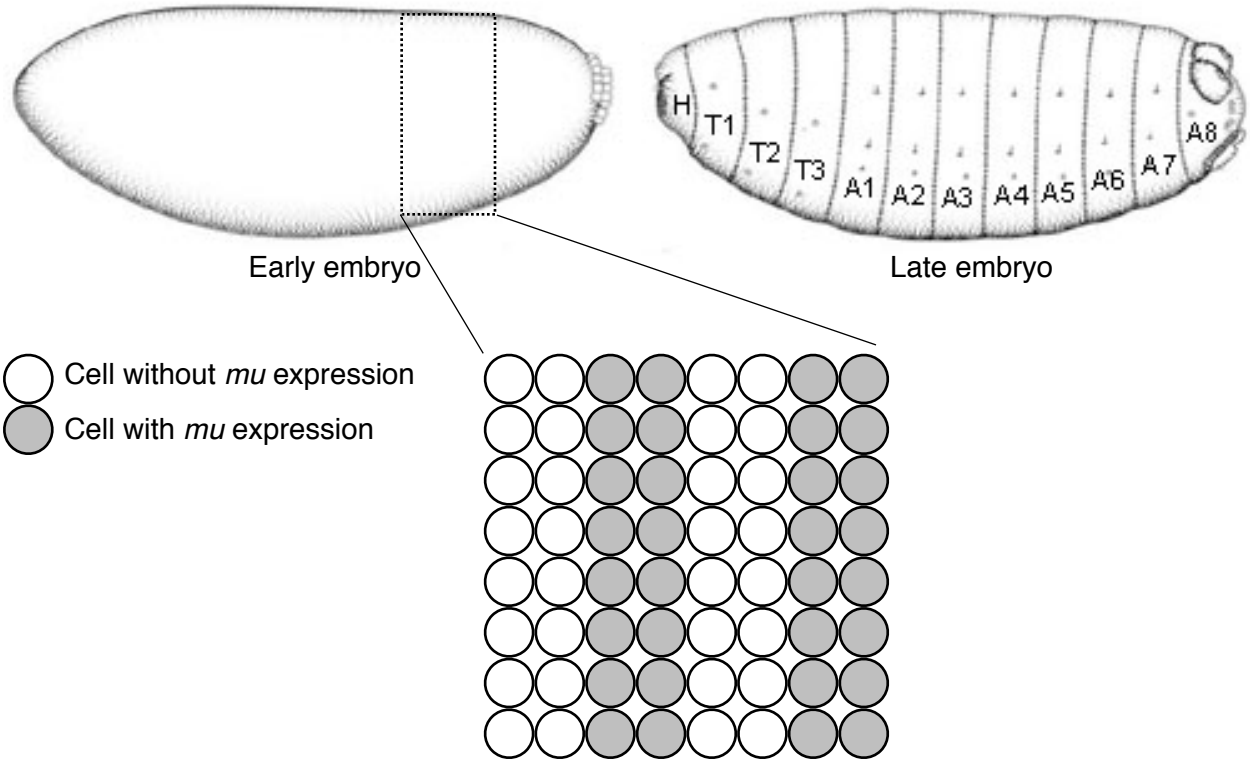
**(c, 4 points)** The group of students genotyped and phenotyped for Dillo Day attendance represented a diverse collection of undergraduates. Why might this statistically significant association be a spurious result of experimental design?

*This association signal could result from population structure. If the students in the case group are enriched for one particular ethnic group not found in the controls, then signal could be correlated with shared alleles from that ethnic group rather than correlated with causal genetic variation for Dillo Day attendance.*

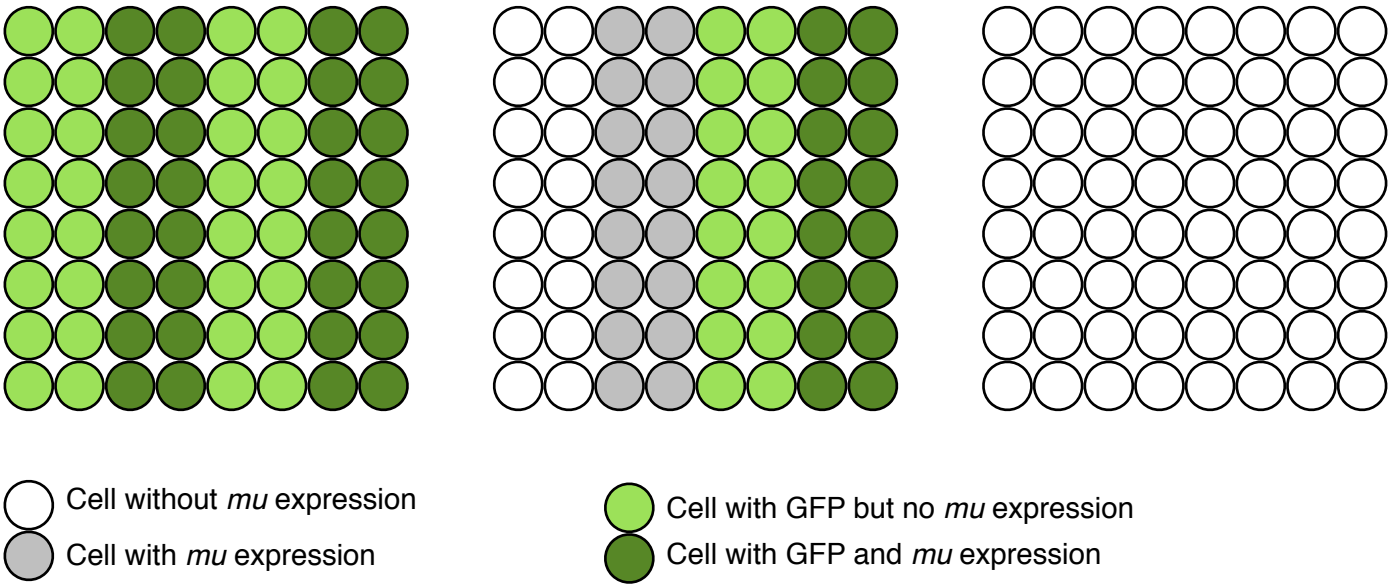
**(d, 5 points)** The HapMap Consortium found that the G allele at this marker is most often found in the African population. Do you think that this allele could have been introduced into our species from interbreeding with Neanderthals? Why or why not?

*No. Neanderthals and modern humans shared the same regions in Europe. The African population did not interbreed with Neanderthals.*

**Question 8 (20 points):**
You are studying pattern formation of the *Drosophila* embryo. Previous studies established that certain genes are expressed in pattern-specific ways, even before the time that patterns can be distinguished morphologically. One such gene is called *mixed up* (*mu*).




Mutations in the gene *disorganized* (*dis*) cause abnormal morphological patterning in the embryo. To determine whether *dis* affects the pattern of *mu* expression, you create genetic mosaics by transplanting nuclei between wild-type and *dis* embryos. Wild-type cells are labeled using the *GFP* gene with a constitutive promoter driving gene expression. In three representative mosaic embryos, you see the patterns of *mu* and *GFP* expression indicated below.



**(a, 8 points)** What two conclusions can you draw about the role of *disorganized* in controlling *mixed up* gene expression?

*(1) disorganized promotes mixed up expression*
*(2) disorganized acts cell non-autonomously*

**(b, 6 points)** Other data suggest that the *confused* (*cf*) family of genes (*cf1* and *cf2*) regulate *mu* gene expression. Previous genetic screens for embryonic patterning mutants failed to identify any *cf* mutants. Assuming the EMS screen was saturated, please suggest three other explanations for why *cf* mutants were not found.

*(1) cf1 and cf2 could act redundantly. One would see the mutant phenotype only if both genes were mutated*

*(2) Both genes could be lethal when mutated (or any other form of pleiotropy preventing the ability to see the mutant embryonic patterning phenotype)*

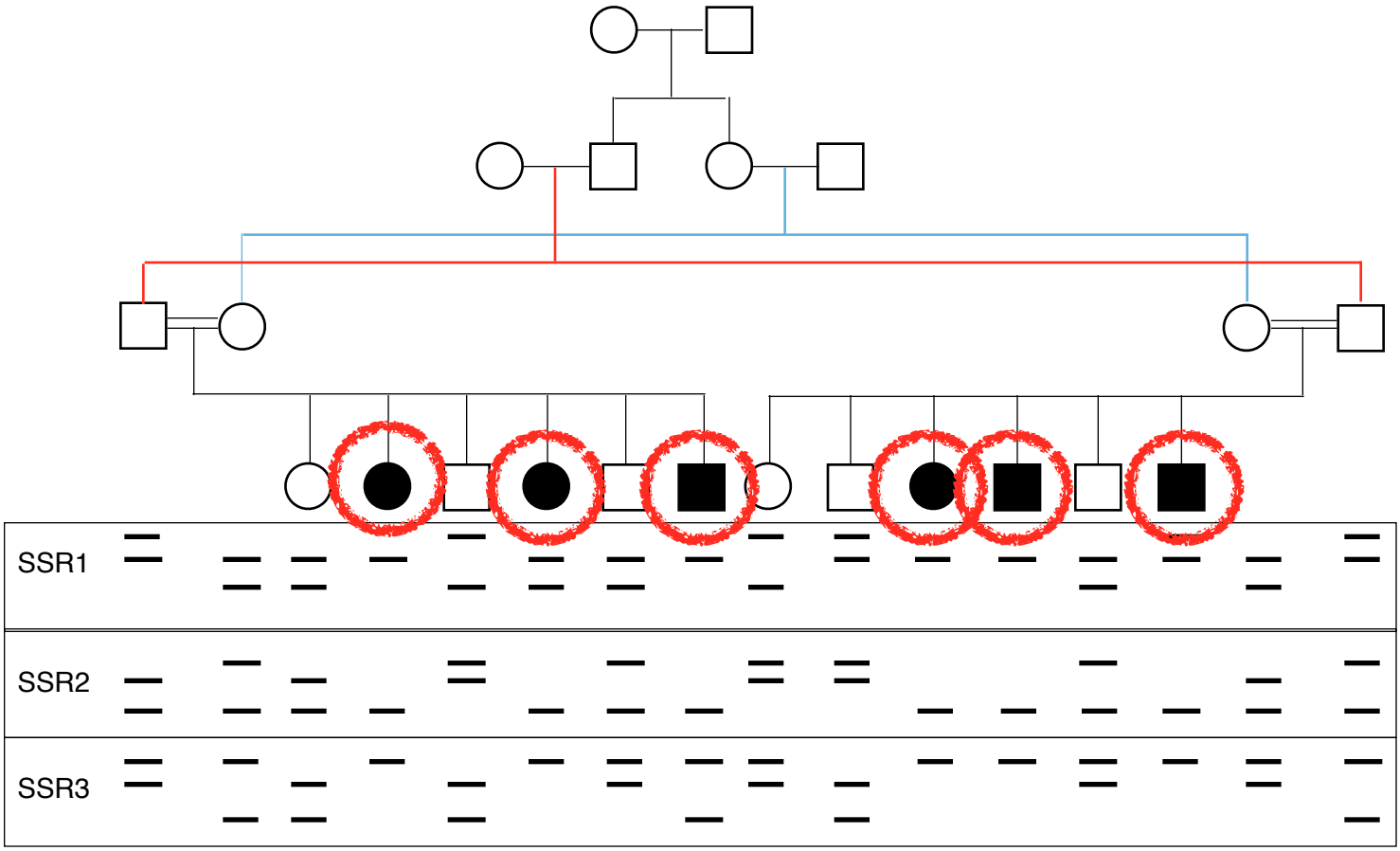
*(3) EMS might not mutate sites in the confused family genes that alter function*

**(c, 6 points)** Using a loss-of-function allele of *cf1*, you perform a genetic screen and identify a new mutation that causes abnormal patterning in the embryo indistinguishable from that caused by a lack of *mu* expression. Without describing specific crosses, describe how you would determine if the effects of this new mutation depended upon the presence of a *cf1* mutation.

*The assumption here is that the abnormal patterning phenotype is dependent on loss of cf1 and another gene. If you outcross this new mutant strain to the wild type and then intercross these heterozygous offspring, you would expect 1/16 of the offspring to have the mutant embryonic patterning phenotype if both alleles confer recessive phenotypes.*

*Alternatively, you could rescue the cf1 mutant in the isolated double mutant and see if the abnormal embryonic patterning is suppressed.*

**Question 9 (32 points):**
You are studying a family with a rare genetic disease with autosomal recessive inheritance. The pedigree is given below along with the segregation of alleles for three linked simple sequence repeats. The order of the three markers is SSR1, SSR2, and SSR3. The double horizontal lines denote consanguineous marriages. The children of generation II are connected with red or blue lines to clarify relationships.




Individuals III-1 and III-4 inherit the middle allele of SSR1, the bottom allele of SSR2, and the top allele of SSR3 from their father individual II-2. Individuals III-2 and III-3 inherit the middle allele of SSR1, the bottom allele of SSR2, and the top allele of SSR3 from their mother individual II-3.

**(a, 6 points)** Circle the progeny in generation IV that are informative for the markers and disease allele.

**(b, 20 points)** For each SSR, calculate the LOD score for the best theta given the data above. Remember that the LOD equation is the log of the odds ratio that you see linkage between a marker and the disease-causing allele. For a recessive disorder, each chromosome must be evaluated. Think of it as every affected individual as the likelihood of two chromosomes coming together. Use the form of the odds ratio (below) to calculate LOD score and show your work on the following page. P denotes the number of informative parental chromosomes, and R denotes the number of informative recombinant chromosomes.

$$\text{Odds ratio} = \frac{\text{Likelihood(theta)}}{\text{Likelihood(0.5)}} = \frac{0.5*((1\text{-theta})^{P} * (\text{theta})^{R} + (1\text{-theta})^{R} * (\text{theta})^{P})}{(0.5)^{P+R}}$$

SSR1 *The only informative individuals are the ones with the disease, because unaffected individuals could be carriers (heterozygotes for the disease-causing allele) or homozygotes for the unaffected allele. We can never know the status of the marker and disease-causing allele definitively.*

*For a recessive disease, the likelihood of inheriting the disease must take into account the likelihood of inheriting two affected chromosomes. In this way, the probability of each chromosome must be calculated.*

*Phase is known.*

$$LOD = \log \frac{(1-0.083)^{11} * (0.083)^{1}}{(0.5)^{12}} = 2.18$$

theta = 1/12 = 0.083
or 1 recombinant chromosome out of 12, in six individuals
P=11, R=1

SSR2

$$LOD = \log \frac{(1-0)^{12} * (0)^{0}}{(0.5)^{12}} = 3.61$$

theta = 0/12 = 0
or 0 recombinant chromosomes out of 12, in six individuals
P=12, R=0

SSR3

$$LOD = \log \frac{(1-0.083)^{11} * (0.083)^{1}}{(0.5)^{12}} = 2.18$$

theta = 1/12 = 0.083
or 1 recombinant chromosome out of 12, in six individuals
P=11, R=1

**(c, 6 points)** Have we established linkage between the disease gene and any of the three SSR markers? Why or Why not?

*Yes, SSR2 has a LOD score of greater than 3. Therefore, we believe that the disease-causing allele is linked to this marker.*

**Question 10 (12 points):**
A patient comes in to your medical office presenting his 23andme results and an extreme sense of worry. Both his father and his grandfather died of prostate cancer, and he is worried that his days are numbered given his genome results.

**(a, 6 points)** Please calculate his risk of prostate cancer given the results below.

| NAME | AVG. RISK | COMPARED TO AVERAGE |
|---|---|---|
| Atrial Fibrillation | 27.2% | 1.25x |
| Prostate Cancer ♂ | 17.8% | 1.33x |
| Gallstones | 7.0% | 1.58x |
| Exfoliation Glaucoma | 0.7% | 2.90x |
| Ulcerative Colitis | 0.77% | 1.30x |

*1.33 * 17.8% = 23.7%*

**(b, 6 points)** Further down on the form you see that he has reduced risk for Alzheimer's disease. He is completely confused how 23andme determined that result. Please briefly explain in words how his risk can be less than 1x and how they calculated it.

| NAME | YOUR RISK | AVG. RISK | COMPARED TO AVERAGE |
|---|---|---|---|
| Gout | 17.1% | 22.8% | 0.75x |
| Venous Thromboembolism | 9.0% | 12.3% | 0.73x |
| Alzheimer's Disease | 4.3% | 7.2% | 0.60x |
| Age-related Macular Degeneration | 3.1% | 6.5% | 0.48x |
| Melanoma | 2.2% | 2.9% | 0.75x |

*23andme calculated his Alzheimer's disease risk by looking at his genome-wide genotype. This individual shares haplotypes (markers) with individuals that have Alzheimer's disease less often than the average person.*

*The alleles present in an individual might provide some protective benefit for certain diseases.*