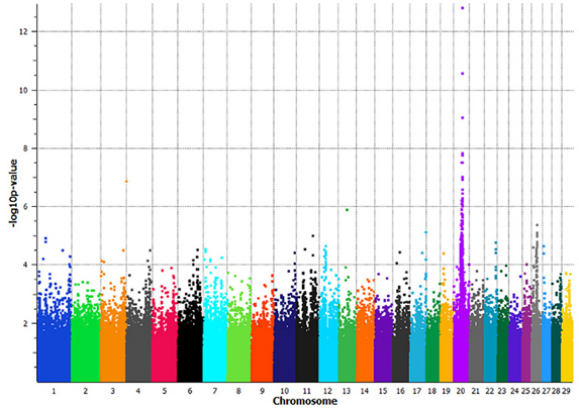
Final Exam 2016 **Name:**
34 points

1. (3) Huson et al. (2014) are interested in the genetic basis of the "slick" hair trait of cows, in which slick haired cows have shorter hair than other cows. The trait is common among cows in the tropics. They genotyped a large population of cows and assayed their hair length. Below is a Manhattan plot of SNPs (x axis) and their P value (y axis) for an association study. **(a)** What does this plot tell you about the genetic control of slick hair? Explain **(b)** Imagine an immunity gene conferring tolerance to a number of tropical diseases is on chromosome 20 (which has the points with highest y values). Explain if/how this would affect your answer to (a).



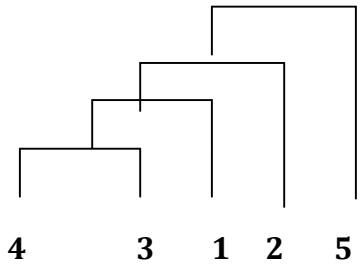
2. (2) You are a veterinarian, and a person comes in with their dog. They are concerned the dog has a predisposition for heart disease and wonder whether to give the animal expensive medication. The dog is genotyped and is homozygous for G at locus X. You read a paper that notes that the relative risk of a dog with GG developing cancer relative to other genotypes is four(4). The paper also gives the odds ratio. **(a)** Explain what "relative risk" and "odds ratio" mean **(b)** Given the relative risk of 4, what would you tell the owner in terms of medication? Do you need more information? If so, what?

3. (3) You want to determine the genetic distance (# substitutions per site) between two sequences. The percent of different nucleotides between the sequences will likely be an incorrect distance estimate. Give and explain three reasons why.

4. (4) Given the sequence data below, where the ... represents additional nucleotides:

```
          12345
seq 1     TACTT...
seq 2     TACAT...
seq 3     TACCT...
seq 4     TCCTT...
seq 5     TCCAT...
```

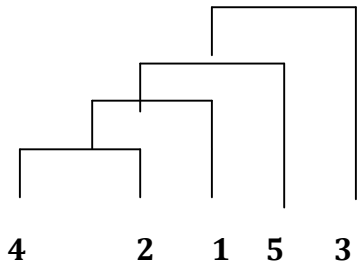
a) Here is a tree with the sequence IDs at the terminal nodes. Calculate the parsimony score <u>for position 4</u> in the alignment above. Show your work.



b) Can you draw a tree for position 4 with an equal or better parsimony score? Do so!

c) Here are the genetic distances calculated from all nucleotide positions (e.g. 1 to the end ...). Given these distances, is the phylogeny above or below a better one given Fitch-Margoliash criteria? Explain.

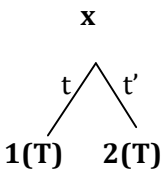
	1	2	3	4	5
1	-	0.2	0.2	0.2	0.4
2		-	0.4	0.4	0.2
3			-	0.1	0.6
4				-	0.6
5					-



5. (4) Two taxa (1 and 2) are joined by a shared node x. These two taxa each have a single homologous nucleotide T. Assume that prior probability of any nucleotide at a node is 0.25. The probabilities of change from one nucleotide to another over the two branches x to 1 and x to 2 are as follows:

Branch	P value
x to 1	P_{AT}=0.05
x to 1	P_{GA}=0.05
x to 1	P_{CA}=0.05
x to 1	P_{TT}=0.85
x to 2	P_{TT}=0.79
x to 2	P_{GA}=0.07
x to 2	P_{CA}=0.07
x to 2	P_{AA}=0.07

(a) Calculate the likelihood score for this tree.



(b) Assume that taxon 1 has an A instead of a T. Will this tree have a higher or lower likelihood than the one above? Explain (or show) why.

(c) We have learned how to calculate the likelihood of one phylogeny for one position. How do we calculate the likelihood of a tree for an entire sequence alignment with N positions? (describe in words or by using an equation).

(d) It is not true here, but why may G<->A changes and C<->T changes have a higher probability than other changes?

6. (2) **(a)** Describe what "bootstrap support" means in the context of measuring confidence for phylogenetic relationships. **(b)** Can you explain how to calculate it?

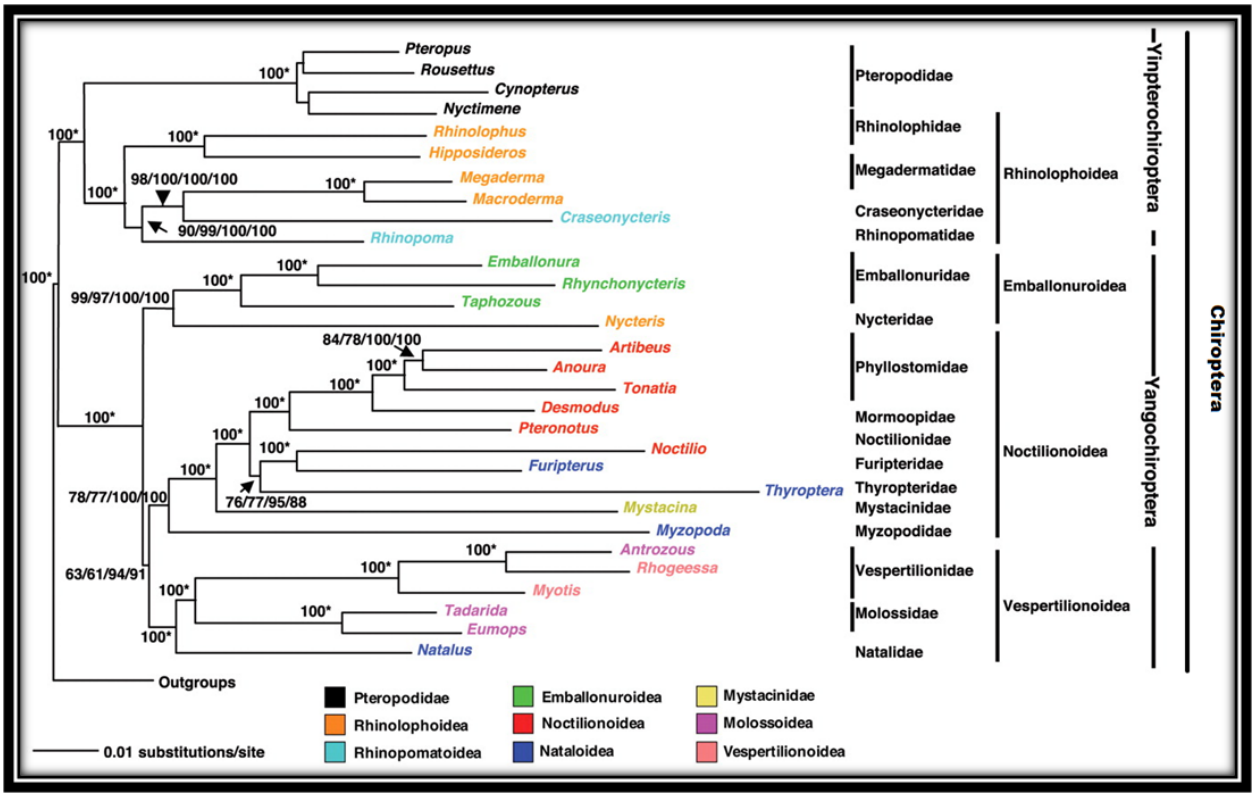
7. (1.0) You randomly generate one nucleotide sequence of 10000nt and align it to a randomly generated second 10000nt sequence without gaps. The frequency of each nucleotide is 0.25. You then record the # of matches as the score of the alignment. You obtain a score of 2750. Do you think this score significantly higher than expected ($P<0.05$)? Explain using equations.

8. (1.0) You do a two-tailed t.test to test the difference between the expression value of a gene in 10 individuals treated with drugs compared to 10 individuals treated with a placebo. You get a P value of 0.01. What is this P value the probability of?

9. (2) The mouse genome is sequenced. Some mice have no pigment; others have brown pigment. You have genetically mapped a gene (gene C) that controls coat colour in mice between markers 1 and 2. Explain how you could use this genetic map information to identify the nucleotide sequence difference(s) that cause the coat colour differences. Explain why the genetic distance between markers 1 and 2 is relevant given this task.

10. (3) You know that nucleotides occur at the following frequencies within frogs (A- 0.2, G-0.3, T-0.2, C- 0.3). Nucleotides occur at the following frequencies within toads (A- 0.25, G-0.25, T-0.25, C-0.25). **(a)** Given this information, explain how to calculate the probability of the sequence AATC given that it came from a frog and from a toad? Which is more likely? **(b)** To calculate the probabilities for both sequences, what is one assumption that we make? Explain.

11. (3) The figure below is from a bat phylogeny paper. You may write on the phylogeny to answer the question if it is useful.
The megabats (Pteropodidae, at top) do not echolocate, nor do the outgroups (bottom). The other bats do. **(a)** Using the phylogeny explain and/or show how echolocation may have arisen <u>more than once</u>. **(b)** Using the phylogeny explain and/or show how echolocation may <u>arisen once only once</u>.



(b) About how many substitutions would you expect to have occurred in a 100nt sequence between the Natalus and Myzopoda groups? (these are found at the bottom 1/3 of the tree)

(c) According to this phylogeny, Tadarida and Eumops (at bottom) share a common ancestor. Why are the branch lengths leading to these lineages not equal?

12. (6) Short answer. Define and explain each term.

FASTQ format

"coverage" and "highthroughput sequencing"

optimization approach and algorithmic approach for phylogeny estimations

BLOSUM matrix

"linkage disequilibrium" and "selective sweep"

Chi Square test

EQUATIONS- Feel Free to remove this page if convenient.

$$M = Ratio = \log_{10}(\frac{RED}{GREEN}) \quad A = (\log_{10} Cy5 + \log_{10} Cy3)/2$$

$$D = 1 - (a + f + k + p)$$

$$d_{xy} = -\frac{3}{4}\ln(1 - \frac{4}{3}D) \quad S = \frac{P(Seq \mid Model1)}{P(Seq \mid Model2)}$$

$$L_{(j)} = \sum_m \pi_m \bullet (\sum_{j=1}^{4} P_{mj}^t L_j^{(1)}) \bullet (\sum_{k=1}^{4} P_{mk}^t L_k^{(2)})$$

$$Uncorrected\_cM = \frac{\#\ recombinants}{\#\ offspring} x100 \quad D_{AiBj}=P_{AiBj}-p_{Ai}p_{Bj}$$

$$P(Seq \mid Model1) = \prod_{i=1}^{n} q_{x_i}$$

$$P(Seq \mid Model2) = \prod_{i=1}^{n} q_{x_i}$$

$$F(i,j) = \max\begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

$$P(x) = q_{x_i} \prod_{i=2}^{N} r_{x_{i-1}x_i} \quad t = \frac{\bar{x} - 0}{s/\sqrt{n}} \quad Var = n * a * b \quad sd = \sqrt{Var}$$

$$L_i^{(3)} = (\sum_{j=1}^{4} P_{ij}^t L_j^{(1)})(\sum_{k=1}^{4} P_{ik}^{t'} L_k^{(2)})$$

$$E = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \left| d_{ij} - p_{ij} \right|^{\alpha} \quad X^2 = \sum_i \frac{(n_i^{obs} - n_i^{\exp})^2}{n_i^{\exp}} \quad E = Kmne^{-\lambda x}$$

$$F(i,j) = \max\begin{cases} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

$$Var(x) = s^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n-1}$$