1. Being a really cool scientist you decide to shotgun sequencing the genome of your beloved dog Dakota, you find that 155 of the 1,630 (9.5%) protein coding genes are essential to immunity and yet do not have a known function. One particular integral gene, F470, has become of interest to you and you would like to know of any other proteins in Dakota’s genome that may interact with the F470 protein. Through the use of diagrams and explanations design two experimental approaches that could be used to solve this inquiry. Are any one of these approaches better than the others, if so, why?

2. BRCA1 is a human gene belonging to a class known as tumor suppressors. It has been found that certain mutations in the BRCA1 gene can greatly increase a woman’s risk in developing breast and/or ovarian cancer. Being a future Nobel Laureate, you want to find a new gene mutation that causes tumor growth. Using functional genomics techniques, from the “Emerging Technologies in Yeast Genomics” paper or other sources, explain how you would discover 1) the gene of interest and 2) the deleterious mutation.

3. You are a student working in Craig Venter’s lab doing genome assembly research. Your colleague’s 21 run was last night and today he’s saying that he can construct whole species sequences from the water samples from the Sorcerer II. Being a sober Sally and diligent lab worker, you believe this to be untrue. Explain to him why he cannot isolate an entire final sequence (virus or other) in its entirety.

4. You love beer and you would like to employ mutagenesis on yeast to identify all or at least the majority of the genes that are involved in the tasty beer cycle. When genes coding for the tasty beer cycle are knocked out the yeast will have a lower fitness when grown at 40° C. Why is random transposon induced mutagenesis with engineered bacterial transposons Tn3 or Tn7 not adequate to cover all genes involved in this cycle? What technique would you suggest & why? Describe & outline this technique below.

5. An exceedingly wealthy and powerful scientist has decided to attempt to recreate the endosymbiotic event between mitochondria and non-mitochondrial-eukarya by inserting a minimal mitochondrion into a bacterial cell.

   A) Design an experimental overview whereby this scientist can use his phenomenal resources to reproduce this definitive divergent event. Include an explanation of what organism could be used to study the host-mitochondrial relationship, how the involved genes could be identified, as well as a suitable vector (or vector system) for the transformation of a bacterial cell into an “endosymbiotic” bacteria.

   B) Due to the enormous complexity of this experiment, it is critical that all genomic sequences involved be correct. If there is doubt about the veracity of the sequence as a whole, how could it be partially checked using sequence assembly tools and without re-sequencing the genome?

6. Describe the automated gene annotation program(s) (Otto), as well as the methods used to validate annotated genes?

7. Craig Venter’s human genome project produced two separate but similar computational assemblies. What were they? Why were two assemblies done? Which one was chosen for further downstream analysis? Why? How was it determined which of the two were more accurate?

8. A new protein, XYZ, was recently discovered from annotating the newly sequenced genome of the eukaryotic organism *Horribulis expensiveness*. This protein contains what is thought to be a DNA binding domain, but no other information is known.
(a) Identify and explain a genomic technique that you could use to gather information on this gene.
(b) What information about XYZ would this generate? What potential caveats/misinformation could it produce that you should be aware of?

9. A microbiologist has just read *The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples*, and found it to be an interesting look at bacterial evolution. His own project has been studying human gut flora, and wonders if viruses would have a similar effect on gut flora as they did on ocean flora. His gene of interest is thiamine-phosphate synthase (TPS), an enzyme involved in the synthesis of thiamine. Propose an experiment that would measure the differences in this gene between *E.coli* in a normal environment, and those aided by viruses.

10. You are this funded researcher and have been given the opportunity to sequence a newly discovered insect in the Amazon. What method of sequencing would you choose to use between the Hierarchical Clone-by-Clone and the Whole Genome Assembly, and why? Make sure to give specifics on each type of strategy for sequencing.

11. As an undergraduate research assistant at Western Washington University, you have recently identified *Candida bellinghamis*, a yeast that has been colonizing in the empty kegs at Boundary Bay. Being the diligent researcher that you are, your research advisor has asked you to design an experimental approach to better characterize the gene *hope1* (pronounced hop-ee-one) in *C. bellinghamis*. Funding is not an issue for you, but efficiency is.
   A) In theory, what experimental tool can you engineer to do this (what must it contain) and what actual steps are involved to use this tool? (Feel free to use pictures and diagrams in your explanations)
   B) Based on the design of your tool, what information do you hope to obtain?
   C) What characteristic of yeast makes this technique possible?
   D) What are some potential problems inherent to the tool you chose?

12. You and your lab team, at the disaster prone Kelera labs, are assigned the task of sequencing the genome of an aquatic microbe using the whole genome assembly approach. You obtain a fresh sample of the aquatic microbe and grow it on a plate. You then select one colony and grow in a large vat of LB broth. A careless scientist contaminates the vat with sea water. Your microbes are then shot gun sequenced and the DNA is stored in a yeast artificial chromosome (YAC) based library. The library is stored in a malfunctioning freezer for an extended period of time at room temperature. After the sequencing process, you run your genetic reads through the WGA computers. For an unknown reason The Overlapper is having difficulty performing its function. Describe two reasons why The Overlapper is having difficulty. Use pictures if necessary.

13. In reference to the paper “The Sequence of the Human Genome,” describe the two approaches they used to assemble the genome by whole-genome assembly (WGA)? Also, how did they manage to reach 8X coverage and resolve any bias from the sequences to reach 8X coverage for WGA?

14. Congratulations, you are a cell biologist who is working on a particular metabolic pathway in *Saccharomyces cerevisiae* in particular you are looking for a gene that function to converts a novel sugar into glucose. You shell out $1,500 for your *S. cerevisiae* library which has up and down tags on each of the knocked out genes. How would you start to identify genes in the process of the metabolism of this novel sugar? Explain how your experiment would narrow down your search for genes involved in the specific pathway. How certain would you be that you know the specific function of the gene?
15. You **are** a smart scientist. You are interested in the viral genes now incorporated in bacterial genomes. You sequence the genome of a particular bacterium, then use Otto program on your data. Your results show a large number of genes annotated, and many suspected genes of viral descent. When you share your results, your lab rival accuses you of "baseless heuristics." Briefly explain your experiment's methods and process, and then justify and explain in detail how you used the Otto program and why it has given better results than other similar software. Also, explain to him why it isn't unlikely for viral genes to occur commonly in bacteria.

16. When Craig Venter and Celera were sequencing the human genome they had to take precautions to introduce as little error in the beginning steps as possible. Two of these measures included “shredding” all of the PFP data and using compartmentalized shotgun assembly (CSA) as a complement to whole genome shotgun assembly. Describe both techniques and how they increased the accuracy of sequencing.

17. You are trying to study an unknown regulatory sequence. A known translation factor is thought to interact with your sequence. Explain the steps you would take to investigate this DNA sequence.

18. In the Venter et al in the *Mycoplasma* paper note: Examination of the gene disruption data, organized by functional role, reveals that all functional classes of genes are not equally mutable under the selective growth conditions used in this study, which suggests that the genes are closer to a minimal set for some cellular functions than for others (12). But also that: It is clear that genes that are individually dispensable may not be simultaneously dispensable.

Is there overlap in these sets? Which partially disrupted proteins turned out to be indispensable? How could you test if these gene families are actually not required in the minimal genome?

19. When working on the human genome, Venter used data from the public project to help complete the sequencing. Using only the techniques that Celera used (ie unitigs, rocks, stones, tiler, shredder etc) would his team have been able to determine the human genome correctly - Why / Why not?

20. You are Craig Venter. (see figure) A mindless PFP researcher got lost on the way to work and ended up in your lab. Before you were able to correct his/her error, they managed to damage your nebulizer so that when you set it to make fragment of size 10 kB, it now produces 5 kB fragments. These fragments are ligated into the mix with your other 10kB vectors. Where specifically in your sequencing project might this cause problems? Try to think of at least two. Be sure to use the word unitig in your answer, in a non-derogatory context. Do not use the word boulder. Take your pick on the figure.

“Bite my shiny, metal #$%.”

21. In the paper, “Strategies for the systematic sequencing of complex genomes”, two of the approaches discussed were the clone-by-clone approach, the whole genome assembly approach.
Compare and contrast these strategies as well as any advantages or disadvantages that may arise with their use.

22. As a budding genomic botanist, you decide that you’re interested in sequencing the genome of the plant, *Dactoris jufyungous* for your first project in graduate school. This organism is known to have a relatively large genome, so you know that you have quite the task ahead of you. Your research advisor tells you that the easiest way to undertake this project would be to do whole-genome shotgun sequencing for the entire *jufyungous* genome while using yeast artificial chromosomes as your vector. Is this a wise decision? Defend your answer by explaining the benefits to approaching this project in the manner your research advisor suggests, or propose a new method that also explains why this is not a good idea.

23. The metagenome paper characterizing viruses in ocean water samples addresses identifying the microbial diversity of a locale. The minimal genome paper looks at pinpointing the bare set of genes mandatory for cellular survival. Both issues required sorting. How did each project sort its data, i.e. how did the metagenome paper identify its sequences and how did the minimal genome project distinguish essential gene from non-essential genes? How did both projects use existing sequences to achieve their respective goals? How is the question of “what is the minimal genome” related to the question of “what is a species”?

24. Why didn't YACs work for the public's approach? Why did they choose them in the first place? How would you be able to test that this was happening in either the Celera or the public approach? What was the solution to the problem for each, and compare and contrast the approaches of Celera and PFP to identify and select sequences with the least amount of overlap.