Study Questions for "Navigating the Pitfalls of Applying Machine Learning in Genomics" by Whalen et al.

1. Name several examples of data sets in bioinformatics research where the examples are dependent. Try to think of some examples that are not mentioned in the paper. What is the source of the dependence? Can you think of any data sets where the examples are independent?
2. Name several examples of problems in bioinformatics research where classes are unbalanced. Try to think of some examples that are not mentioned in the paper. How would you mitigate the unbalance if you were analyzing this data? Why? Can you think of any problems in bioinformatics with balanced classes?
3. Distributional differences (Pitfall 1) and Confounding (Pitfall 3) are related. Discuss the crux of each pitfall. What makes them similar, and what is different about them?
4. When in the process of cross-validation should batch correction be performed? Why?
5. **Challenge question:** How do the ideas in this paper apply to unsupervised learning and exploratory data analysis, such as clustering and visualizing cell types in single-cell experiments?