Study Questions for "Addressing biases in gene-set enrichment analysis: a case study of Alzheimer's Disease" by Bakulin et al.

1. What are the main similarities and differences between over-representation analysis (ORA) and functional class scoring (FCS)?
2. Is pyPAGE an ORA or an FCS enrichment testing method?
3. What distribution does the count of the number of annotations per gene follow?
4. What specific type of bias does pyPAGE aim to address? How does it use conditional mutual information to do this (i.e., what is conditioned on and what effect does that have)? You might first start by thinking about what mutual information measures and how it is implemented.
5. Which data modalities does pyPAGE use to define gene sets for testing? Which of these are demonstrated in the Alzheimer's Disease analyses?
6. What do the authors mean by "non-monotonic complex relationships are captured (e.g. dual regulators)" on p.6?
7. **Challenge question:** In the simulations, area under the precision-recall curve is used rather than other performance metrics (e.g., area under the ROC curve). The justification given is that "the number of unperturbed gene sets is much greater than the number of perturbed gene sets". Can you explain this logic? Do you agree with it?
8. **Challenge question:** Equation (2) on p. 26 is proposed as a way to quantify differential gene expression. What do the two terms represent and why do you think the authors chose this equation? Would you use it?