Research Statement for Jeffrey T. Leek

I am interested in finding simple, robust statistical solutions to important high-dimensional problems in biology. My research approach consists of three steps. First, in collaboration with biologists and clinicians I identify concrete problems that require statistical solutions. Next, I try to develop methods that focus on the concrete problem. Then, after solving the specific problem, I look for generalizations of my methods that may be useful for other problems in high-dimensional statistics. My research has mostly been applied in the context of gene expression microarrays, but the solutions I have developed can be applied across a variety of fields, including genetic epidemiology, brain imaging, spatial epidemiology, and computer science.

My doctoral research focused on dependence in high-dimensional multiple testing. My work was motivated by data analysis problems in a study of the molecular response to massive trauma, the Inflammation and the Host Response to Injury Glue Grant. We noticed that heterogeneity among patients resulted in a loss of power and confounding when looking for associations between expression and clinical outcomes like multiple organ failure. This heterogeneity had not been identified or addressed by any method for gene expression analysis. We solved this problem by recognizing that the gene expression data for each patient is a multivariate observation subject to the same conditions. By borrowing strength across genes, it is possible to estimate latent factors driving the heterogeneity among patients. We showed that including carefully estimated latent factors in the model relating gene expression to clinical outcomes eliminated confounding due to heterogeneity, improved the ranking of genes for differential expression, and improved power.

Based on the idea of borrowing strength across genes to estimate latent structure, we showed that any continuous high-dimensional data can be decomposed into a component that is dependent across features and a component that is independent. We also showed that a low-dimensional set of vectors is often sufficient to parametrize the dependent component of high-dimensional data. The parameters in this "dependence kernel" can easily be estimated by borrowing strength across the large number of features in high-dimensional data. The dependence kernel represents an interesting reversal of the "curse of dimensionality" often encountered in high-dimensional data analysis. In my dissertation I developed asymptotic theory designed for this new situation, where the number of samples is fixed, but the number of features grows large. This asymptotic approach has been fruitful in the study of P-values from multiple tests, but has yet to be developed in detail for the underlying high-dimensional data.

Following my doctorate I spent seven months in a molecular biology laboratory learning about the experiments I plan to analyze as a statistician. I learned about cell culture, sources of variation in molecular biology experiments, and specific techniques like PCR under the supervision of molecular biology graduate students and postdocs. I helped develop a pipeline for analyzing high-dimensional methylation, gene expression, and protein expression data produced by the lab and contributed statistical support to several other lab projects. Throughout this time, I continued my statistical research in multiple testing dependence. My laboratory experience will be useful both in developing my biological intuition and when working with collaborators.

After working in the lab, I returned to full time statistical research for my second postdoctoral position. Like my doctoral research, my current postdoctoral work began with a concrete statistical problem arising when classifying patients with microarray data. In breast cancer studies, physicians classify patients into low, intermediate, or high risk of recurrence based on

phenotypic variables like estrogen receptor status, tumor size, and age of the patient. Unfortunately, many classifiers based on gene expression give the same classification as a physician would based on standard phenotypic data. An important problem is to develop classifiers that build on the expertise of physicians, and use microarray data to improve classification of intermediate, difficult to assign patients. I developed a way to model covariates – such as a physician's classification or phenotypic information – when building top scoring pair (TSP) classifiers.

To understand the TSP idea, imagine a pair of genes where Gene 1 has higher expression than Gene 2 in low risk patients, but Gene 2 has higher expression than Gene 1 in high risk patients. The pair that shows the greatest difference in relative ranking between the groups is the TSP. I showed that the TSP selection criteria is equivalent to a model selection on a logic regression model. The logical outcomes in the model are the indicator functions giving the relative ranking of the genes. By placing the TSP in the regression framework, it is straightforward to incorporate covariates when selecting the best pair of genes. The resulting covariate adjusted TSP classifiers show substantial improvement over the physician's judgement alone. Another advantage of the model framework is that TSP classifiers can now also be based on more complicated data types, such as survival data.

In the future I would like to build on the work I have done in multiple testing dependence and classification, to apply these techniques to data from my current collaborations, and to form new collaborations in areas where my work can be useful. Specifically I would like to work on the following projects.

Short-Term

- Continue a collaboration with the Glue Grant showing that modeling the dependence kernel improves reproducibility in clinical gene expression studies
- Continue my work on the dependence kernel and asymptotic theory for continuous high-dimensional data
- Continue a collaboration with Giovanni Parmigiani, Don Geman, and Leslie Cope to create methods based on the model framework I developed for top scoring pairs
- Continue a collaboration with Giovanni Parmigiani, Luigi Marchionni, and Antonio Wolff to improve physicians' classification of breast cancer patients with microarrays

Midterm

- Develop methods for estimating the dependence kernel for fMRI data, where dependence between voxels is a key analysis obstacle
- Extend the multiple testing dependence and asymptotic theory for high-dimensional data to discrete distributions
- Apply these new ideas to analyze high-throughput sequencing and whole genome association data

Biological experiments and technology are evolving at a rapid pace. Since the goal of my research program is to address important and relevant problems in biology with statistics, my longer term plans are flexible. I hope to take advantage of new technologies and a good research environment to tackle interesting high-dimensional statistical problems in a simple, robust way.