Harnessing the unseen for next generation population genomics and epigenomics

Sequencing of large human populations has the potential to transform disease diagnosis and treatment. In order to harness the full power of this data avalanche, it is crucial to model and leverage the information and covariates that we do not see. I will illustrate this concept with examples from genomics and epigenomics. I will first discuss a close collaboration with the largest exome sequencing consortium (ExAC), where we have aggregated high-quality protein coding sequences (exomes) of 60K healthy individuals. Even with this large dataset, we can identify only a small fraction of the potentially harmful mutations that exist in the human population, because most of these variants are very rare. We developed a new algorithm that leverages the sequenced individuals to accurately infer statistical properties of the unseen genetic variation. This approach has strong mathematical guarantees and provides a unified framework to quantify the natural selection acting on our genome, annotate disease variants, and predict the discovery rate of future sequencing projects.

If time allows, I will describe complementary work to identify changes in the packing and chemical modifications of DNA across individuals—i.e., epigenomic variation—that are associated with diseases. This work requires flexible models of unseen covariates, especially of cell-type composition, to make valid statistical estimation.

Joint work with Greg Valiant, Paul Valiant, Daniel MacArthur and Jennifer Listgarten.

James Zou (Microsoft Research New England and MIT)