James Diao
MB&B 452
24 February 2017

Section 4: Bioinformatics for Next-Gen Sequencing

The two readings introduce the landscape of research around computing with next-generation sequencing data. The first, Rozowsky et al., is a methods paper describing PeakSeq, a novel approach for scoring Chip-seq experiments. The second, Cooper and Shendure, is a review paper summarizing the present study of disease-causing variants. Both papers demonstrate the utility of standardized computational techniques in the analysis of new data made available through next-generation sequencing.

PeakSeq was developed for analysis of ChIP-seq, or chromatin immunoprecipitation followed by tag sequencing. It implements a two-pass approach to correct for false signal peaks arising from open chromatin. The first pass determines potential binding sites by comparing the signal map (of ChIP mapped reads) to simulated segments. Here, mapped reads are extended to the average fragment length, and control data is normalized to the ChIP-seq sample (based on the linear regression). A threshold may be set based on the desired false discovery rate. The second pass identifies differentially enriched target regions (in experimental compared to control) and significances. This scoring procedure gives a P-value from the binomial distribution and corrects for multiple hypothesis testing. Rozowsky et al.'s method corrects for the genomic variation in mappable sequences, and differences in chromatin accessibility, allowing for more reliable experiments on protein-DNA interactions.

Cooper and Shendure present a review of current and past research efforts to document disease-causal variants. Throughout the paper, Cooper and Shendure emphasize the difficulty of differentiating causal variants from neutral ones – hence the title: needles in stacks of needles. The authors begin with computational approaches. First, comparative sequence analysis allows researchers to estimate deleteriousness from conservation. Second, linkage analysis and genome-wide association studies offer more data to inform variant classification. Third, knowledge of protein and amino acid biochemistry, structure, and experimental function may be coded quantitatively. Many classifiers have been developed to use these features to predict deleteriousness. Alternatively, experimental characterization may demonstrate causality. This often demonstrating *in vitro* molecular changes or *in vivo* phenotypic results. Typically, experimental methods are labor-intensive and difficult to scale. However, recent high-throughput methods have enabled high-resolution mapping of sequence-function relationships. Current major challenges include accuracy assessment (benchmarking), study of non-coding variants, and higher-order interactions between the variant and the genomic environment.

- Rozowsky, J, Euskirchen, G, Auerbach, RK, Zhang, ZD, Gibson, T, Bjornson, R, Carriero, N, Snyder, M, Gerstein, MB (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat. Biotechnol., 27, 1:66-75
- Cooper, GM, Shendure, J (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet., 12, 9:628-40