STATISTICS SEMINAR

Seth Raithel

Master's Defense

Thursday, April 30, 2015
K-State Student Union, Room 226, 8:30-9:30 am

INFERENTIAL CONSIDERATIONS FOR LOW-COUNT RNA-SEQ TRANSCRIPTS: A CASE STUDY ON AN EDAPHIC SUBSPECIES OF DOMINANT PRAIRIE GRASS *ANDROPOGON GERARDII*

Big bluestem (*Andropogon gerardii*) is a wide-ranging dominant prairie grass of ecological and agricultural importance to the US Midwest while edaphic subspecies sand bluestem (*A. gerardii ssp. Hallii*) grows exclusively on sand dunes. Sand bluestem exhibits phenotypic divergence related to epicuticular properties and enhanced drought tolerance relative to big bluestem. Understanding the mechanisms underlying differential drought tolerance is relevant in the face of climate change. For bluestem subspecies, presence or absence of these phenotypes seems to be associated with RNA transcripts characterized by low number of read counts. So called low-count transcripts pose particular inferential challenges and are thus usually filtered out at early steps of data management protocols and ignored for analyses. In this study, we use a plasmode-based approach to assess the relative performance of alternative inferential strategies on RNA-seq transcripts, with special emphasis on low-count transcripts as motivated by differential bluestem phenotypes. Our dataset consists of RNA-seq read counts for 25,582 transcripts (60% of which are classified as low-count) collected from leaf tissue of 4 individual plants of big bluestem and 4 of sand bluestem. We compare alternative ad-hoc data filtering techniques commonly used in RNA-seq pipelines and assess the performance of recently developed statistical methods for differential expression (DE) analysis, namely DESeq2 and edgeR robust. These methods attempt to overcome the inherently noisy behavior of low-count transcripts by either shrinkage or differential weighting of observations, respectively.

Our results indicate that, when DE methods are properly specified, the need for ad-hoc data filtering is circumvented, thus allowing for inference on low-count transcripts. Practical recommendations for inference are provided when low-count RNA-seq transcripts are of interest, as is the case in the comparison of subspecies of bluestem grasses.