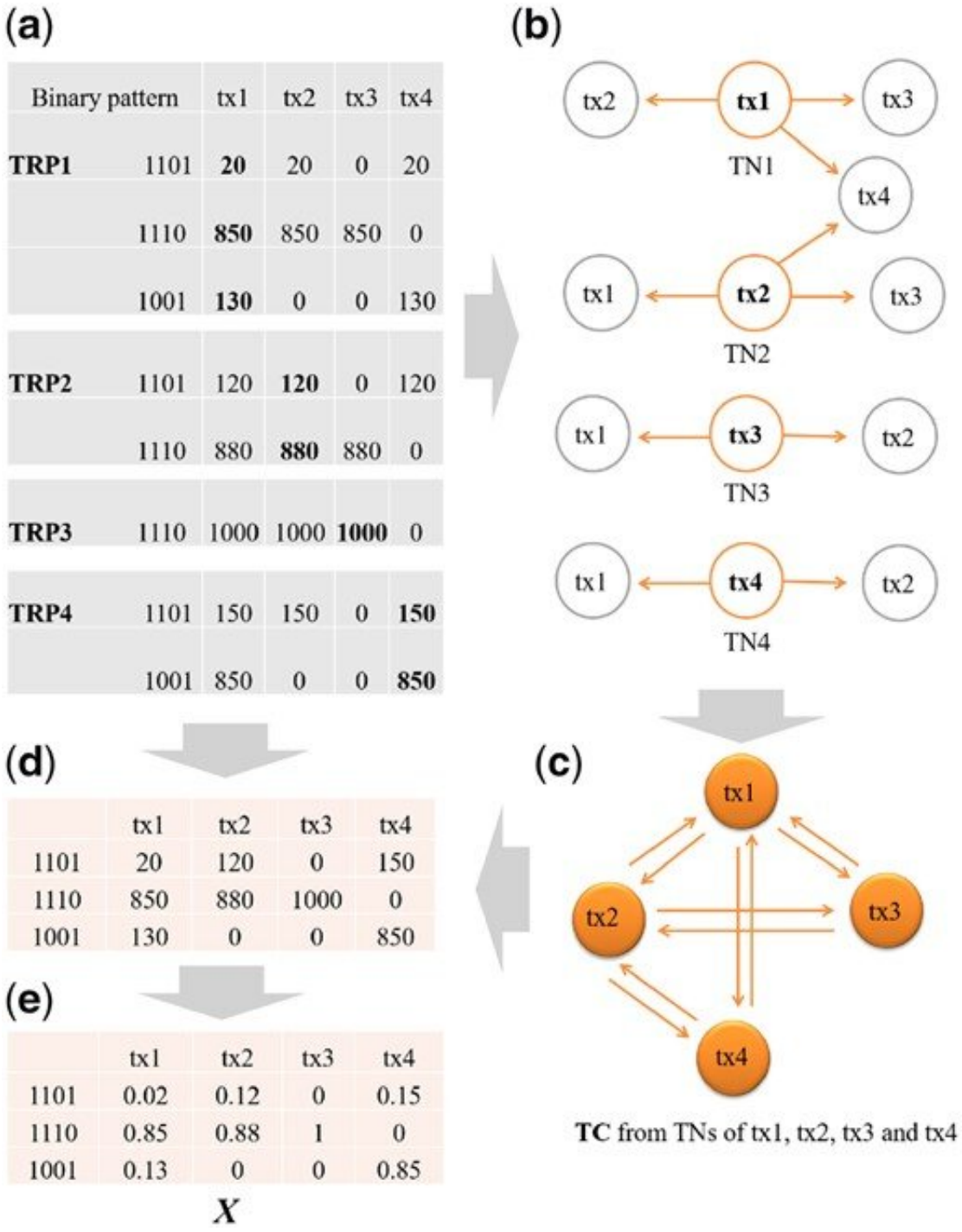
Improved statistical methods for high-throughput omics data analysis

August 25 2021, by Gunilla Sonnebring



Steps to construct the starting design matrix X. (a) TRPs of tx1, tx2, tx3 and tx4, and the summary of binary occupancy patterns from the TRPs. Transcript tx5 does not pass the filtering (H = 2.5%) and is filtered out from TRP1. In each binary pattern, digit 1 means there are reads originating from an eqclass, and 0 otherwise. For example, there are three eqclasses in TRP1: eqclass1, eqclass2 and eqclass3. For eq1 the binary pattern is 1101, which means three transcripts,

i.e. tx1, tx2 and tx4 have reads from eq1. (b) Transcript neighbors (TNs) for tx1 to tx4. (c) Illustration of construction of transcription cluster (TC) from the TNs. We first collect the TNs of tx1, tx2, tx3 and tx4, and then add the connections between transcripts into the TC. For example, from TN1, we add the connection of tx1-tx2, tx1-tx3 and tx1-tx4. In the end, a TC would contain all connections between transcripts sharing exons. (d) The unique set of binary patterns are kept, so three unique patterns remain: 1101, 1001, 1110. We then fill in the read counts from each source TRP. For example, for pattern 1101, in TRP1 the read count is 20 for tx1, in TRP2 the read count is 120 for tx2 and in TRP4 the read count is 150 for tx4. (e) The total reads of each transcript in (d) are standardized to sum to 1 to create the starting design matrix X. Credit: DOI: 10.1093/bioinformatics/btz640

High-throughput omics technology has revolutionized biological and biomedical research and large volumes of omics data have been produced. For this, computational tools to manage and analyze the omics data have been developed and there are big challenges in how to process and interpret the omics data in the best way. Wenjiang Deng has worked to develop novel statistical methodologies and algorithms for omics data analysis, using both simulated and real cancer data to test the methods.

Could you describe some of the results in your thesis?

Yes, in my first study, we identify several genes associated with the survival of high-risk neuroblastoma patients, says Wenjiang Deng, Ph.D. student at the Department of medical epidemiology and biostatistics, MEB. Neuroblastoma is the most common and deadliest cancer in young kids under the age of five. We believe that our findings will provide significant evidence for the treatment and management of patients. Our results can also be meaningful to understand the physiological mechanisms of the disease.

How come you chose to study this particular area?

We are living in the era of "big data," and the high-throughput sequencing data is the predominant "big data" in life science. When I first heard the concept of omics data, I was amazed by its huge volume and the big potential in medical research. Nowadays it is quite easy to produce sequencing data, but we still need efficient and accurate tools to analyze them, so I decided to study the development of algorithms during my time as a Ph.D. student.

What will you do next?

After my defense, I will stay in MEB for a while to wrap up my manuscripts. I will then go to Shenzhen, China, and start working in a biotechnology company which aims to develop new methods for early diagnosis of cancers. I hope that our work there will contribute to the overall health of human beings.

More information: Wenjiang Deng et al, Alternating EM algorithm for a bilinear model in isoform quantification from RNA-seq data, *Bioinformatics* (2019). DOI: 10.1093/bioinformatics/btz640

Provided by Karolinska Institutet