Curate Patient-Centric Multi-Omics Data for Precision Medicine

Jun Zhu, Icahn School of Medicine at Mount Sinai

A biological system is complex. Multiple levels of regulation enable it to respond to genetic, epigenetic, genomic, and environmental perturbations. To understand whether a patient has a risk for a specific disease or whether a tumor responds to a therapy, we need to examine the patient from genetics, epigenetics, genomics, proteomics, and metabolomics angles. Many predictive models have been generated to assess individual's risk or response based on each type of these omics data. Public databases, such as TCGA (The Cancer Genome Atlas), have been created for depositing diverse types of omics data for public dissemination. However, sample errors, such as sample-swapping or mis-labeling, sample cross-contamination, meta-data errors, are inevitable during the process of data generation and management. If a gene expression profile from a different patient was assign to our patient of interest, we may end up with complete wrong prediction and put out patient at a great risk for wrong medicine, which is totally opposite from the goal of precision medicine. Thus, to accumulate omics and clinical data for precision medicine, it is critical to properly curate patient-centric omics data so that different types of omics data pertaining to the same individual truly match to each other.

We developed multiple methods for finding intrinsic regulations between different omics data types, and developed barcodes to match different omics profiles to different people. We also developed methods for matching omics and meta-data, and methods for checking sample cross contamination based on intrinsic barcodes. We applied our methods into TCGA datasets and detected multiple sample errors in many cancer datasets in TCGA (such as glioblastoma, kidney, lung, prostate, stomach). These results suggest that sample errors are not a dataset specific problem but more global problem in public databases.