**Aim 1: Determine the applicability of Mixed Graphical Models in single subject analysis**

*Hypothesis and rationale-* Increasingly available omics data provide an unprecedented opportunity to decipher the molecular mechanisms that underlie disease progression in a comprehensive way. While many models, including Mixed Graphical Models (MGM), have been successfully applied to multi-omics integration with a large sample[1-4], we still fall short of an approach to integrate multi-omics data in the single-subject analysis. With the wealth of multi-omics data, single subject analysis circumvents its shortcoming of high false positive rate and offers well-rounded perspectives of a single patient. We hypothesize that Approximate Bayesian Computation (ABC) can facilitate the application of MGM in the single subject analysis. ABC assists MGM over the limitation of small sample size and further equips MGM with the capacity to incorporate the domain knowledge. We will simulate multi-omics data pertaining to molecular pathways involved in cancer progression, with which the accuracy of the proposed method can be tested. In addition, with the breast invasive carcinoma dataset hosted by The Cancer Genome Atlas (TCGA), we will test the proposed method’s performance in analyzing real biological data. We will test the association between patient survival time and the abnormal pathways identified by the proposed method. At the completion of this aim, we will develop the first method for single-subject multi-omics analysis, which may greatly contribute to the realization of personalized medicine.

*Approach*- We will simulate different scenarios of pathway perturbation. To simulate the joint distribution of an unaltered pathway under study, we will first examine the statistical distributions of each measuring unit of SNP, copy number variation (CNV), mRNA expression, and protein expression. Secondly, we will investigate the correlations among measuring units at each omics scale, and the correlations of the measuring units across different scales. With the independent empirical distributions of all measuring units at all scales and the correlations of the measuring units within/between scales, we can jointly simulate any pathway of interest. We have collaborated with a biologist to design different scenarios of pathway perturbation, including genetic mutation, transcriptional and translational dysregulation, and abnormal post-transcriptional regulation. For each pathway, we will simulate 1000 unaltered replicates and will incorporate all types perturbations and their combinations to simulate perturbed pathways with 1000 repetitions for each scenario.

With the comprehensive simulation dataset, we will evaluate the accuracy of the proposed method in each perturbation scenario and the overall accuracy by assessing the false positive rate, true positive rate, and area under the receiver operating characteristic curve (AUC). The Bayesian framework of the proposed method provides a flexible way to incorporate the external understanding of the unaltered pathway in the inference through the prior distributions. In other words, a physician can apply their knowledge of the molecular mechanisms in pathway perturbation for diagnosis. When testing the accuracy of the method, we will distort the prior distributions gradually away from the true prior distributions and test the influence of correctly incorporating expert knowledge.

The primary outcome of the simulation study is an objective evaluation of the accuracy of the proposed method and its accuracy in different pathway-perturbation scenarios. This will determine the applicability of the proposed method in single-subject multi-omics analysis. In addition, we will understand better the impact of utilizing external expert knowledge in the inference. The outcome in turn will provide insights to further improve the proposed method.

To test the performance of the proposed method in a real biological dataset, we will employ the breast invasive carcinoma dataset hosted by TCGA, which includes SNP, CNV, mRNA expression, and protein expression. We will apply the proposed method to identify the perturbed pathways of each patient, and then use the perturbation-level as a summary statistic to covariates in the patient survival analysis. Perturbation level is calculated by converting the p-value of perturbation to Z score using the inverse function of cumulative normal density function. Prior distributions of the model will be estimated from all available breast cancer omics datasets in Gene Expression Omnibus (GEO). We will use cox *lasso* regression [5, 6], with tuning parameter to minimize 10-fold cross validation error to calculate test association between the perturbation Z scores and patient survival times. The *lasso* provides dimension reduction and parameter estimation from a large number of covariates.

The outcome of the breast cancer study is the significant associations between pathways and the survival time. We will compare this result with the pathways identified in published literatures. We expect to see some overlap between our identified pathways and the ones identified by others. In addition, we also should find more pathways significantly associated with survival time, because the single-subject analysis should provide more patient-specific signals and information from multi-omics should capture the molecular change of different levels.

*Problems and alternatives*- This computationally intensive approach will require a huge number of cpu hours to perform. Estimating prior distributions from all available breast cancer datasets from GEO may not be feasible. Alternatively, we can reduce the number of datasets needed to estimate prior distributions but bring in more subjective understanding of the nature of the prior distributions. Also, the decision rule of ABC is that the generated putative data exactly math the observed data. For high-dimensional observations like the multi-scale data of a pathway, exact match rarely exists. We may need to loose the match criterion to allow approximate matches.

[1] Yang E, Ravikumar P, Allen GI, Baker Y, Wan Y-W, Liu Z. A General Framework for Mixed Graphical Models. arXiv preprint arXiv:14110288. 2014.

[2] Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences. 2013;110:4245-50.

[3] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25:2906-12.

[4] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature methods. 2014;11:333-7.

[5] Avalos M, Pouyes H, Grandvalet Y, Orriols L, Lagarde E. Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies: a simple algorithm. BMC Bioinformatics. 2015;16 Suppl 6:S1.

[6] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. 1996;Series B (Methodological) 267-88.