**Aim 1: Evaluate the accuracy of the novel method MGMCluster**

*Hypothesis and rationale-* Increasingly available omics data provide an unprecedented opportunity to understand the heterogeneity of cancers in a comprehensive way. We developed a novel method, mixed graphical model based clustering (MGMCluster), to cluster cancer patients into subgroups based on their molecular profiles at different ‘omic’ levels. However, the accuracy of MGMCluster remains unknown. The objective of this aim is to evaluate the accuracy of MGMCluster. *We hypothesize that MGMCluster has high accuracy in discovering subgroups of cancer patients.* MGMCluster incorporates expectation–maximization (EM) algorithm with Mixed Graphical Models (MGM) to cluster cancer patients. EM is an iterative method to find the maximum of latent variable model likelihood, where the latent variables are the indicators of subgroups. Meanwhile, we use MGM to model the joint distribution of multi-omics data. Since the true subtypes of cancers are never known, we will conduct simulation study to objectively evaluate the performance of MGMCluster and two other available method: icluster[1] and PARADIGM[2]. In addition, with the breast invasive carcinoma multi-omics dataset hosted by The Cancer Genome Atlas (TCGA), we will compare the subtypes discovered by the three integrated clustering methods and the ones discovered by well-studied biomarkers. Successful completion of this aim will provide a comprehensive evaluation of the integrated clustering methods. At the completion of this aim, it is our expectation that we will demonstrate the superior accuracy of MGMCluster over other integrated clustering methods.

*Approach*- Mixed Graphical Models were established by Yang et al. [3] to specify parametric multivariate distributions over mixed types of variables (**Fig. 1**). MGM is the first model that provides a joint distribution, P(X,Y,Z|), of different types of measurements. With a fully specified joint distribution, we further assume a latent variable, K, as the indicators of cancer subtypes, which in turn implies the distinct conditional distributions, P(X,Y,Z|), for different cancer subtypes. Expectation–maximization (EM) algorithm is a method to estimate the latent variable, K, for each patient by iteratively updating K and the parameters until convergence. The algorithm is as follow:

1. First, initialize the parameters   to some random values.
2. Compute the best values for K given these parameter values.
3. Then, use the just-computed values of K to compute a better estimate for the parameters . Parameters associated with a particular value of K will use only those data points whose associated latent variable has that value.
4. Iterate steps 2 and 3 until convergence.

MGMCluster incorporates MGM and EM to estimate each patient’s cancer subtype from multi- omics data. Note that in this aim we only focus on copy number variation (CNV), microRNA and RNA-Seq (mRNA) data, but the MGMCluster can take to any type of omics data as input.

Figure 1 Mixed Graphical Model (MGM) Each node represents an entity of on measurement; each edge represents dependency of two entities with a measurement or between two different measurements.

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 copy number variation (CNV), microRNA expression, and mRNA expression by using the data of 1098 cases of breast invasive carcinoma on TCGA. In particular, we assume CNV follows multinomial distribution[4],



Where g represents the gene location and k represents the largest copy number of this gene. For each gene, g, we can estimate the corresponding parameters via the maximum likelihood estimation (MLE) [5] with the copy numbers of 1098 cases. Similarly, we assume microRNA expression and mRNA expression follow Poisson distribution , and can be estimated by MLE for each gene for microRNA.

Secondly, we will investigate the correlations among measuring units at each omics scale, and the correlations of the measuring units across different scales by non-parametric spearman’s [6]. 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 the different possible causes of breast invasive carcinoma including copy number variation, transcriptional dysregulation.

We will first simulate three subtypes of breast invasive carcinoma. Each cancer subtype is caused by the similar mechanism mutation. Then, we will apply MGMCluster, PARADIGM, and iCluster to discover the three breast invasive carcinoma subtypes. Since we know the true subtypes of the simulated samples, we can calculate the misclassification rate for each integrated clustering method. To comprehensively evaluate the three methods, we will simulate different molecular mechanism of carcinogenesis and choose different number of total subtypes.

The primary outcome of the simulation study is an objective evaluation of the accuracy of the MGMCluster and the other two most accepted integrated clustering methods. This will determine the superior accuracy of MGMCluster. In addition, we will know the computational feasibility of MGMCluster in large dataset.

To test the performance of the MGMCluster in a real biological dataset, we will employ the breast invasive carcinoma dataset (1098 samples) hosted by TCGA to compare the subtypes unveiled by MGMCluster, PARADIGM, iCluster and the well-accepted biomarkers.

The outcome of the breast cancer study is a comparison of the subtypes discovered by three integrated clustering methods and the well-accepted biomarkers. We expect to see better overlap between our discovered subtypes and the ones discovered by well-accepted biomarkers. In addition, we also should find novel subtypes that could not be identified by well-accepted biomarkers due to the lower statistical power.

*Problems and alternatives*- The performance of MGMCluster on high dimensional data has not been tested. If it fails as the dimensionality increases, we will introduce L1 Norm in the model to force model sparsity [7]. Also, we do not know if it is computationally too intensive to conduct EM algorithm as the dimensionality increases. If the EM algorithm takes too long to converge, we will write the algorithm with faster programming languages such as C++. Lastly, we need to specify the number of cancer subtypes before conducting EM algorithm. This may not be realistic for some cancers. An alternative way is to conduct EM by specifying different number of cancer subtypes and then test in which case the model fits the data significantly better.

[1] 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.

[2] Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237-i45.

[3] 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.

[4] Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, et al. A robust statistical method for case-control association testing with copy number variation. Nature genetics. 2008;40:1245-52.

[5] Aldrich J. RA Fisher and the making of maximum likelihood 1912-1922. Statistical Science. 1997;12:162-76.

[6] Spearman C. The proof and measurement of association between two things. The American journal of psychology. 1904;15:72-101.

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