**Specific Aims**

Population heterogeneity raises a big challenge to the realization of precision medicine. Understanding the heterogeneity of the disease is essential in overcoming this challenge. Classifying cancer subtypes using high-throughput data made a major contribution to our understanding of the disease heterogeneity1-4. However analyzing disease heterogeneity with a single type of high-throughput (‘omics’) data yields high false positive rate [ref] and overlooks the interactome -- interaction within and between all types of molecules. With the advance of high-throughput technologies, a broad range of ‘omics’ data became available, from whole genome sequencing data to transcriptomic, proteomic and methylomic data. This provides unprecedented opportunity to systematically profile a patient from nearly every molecular aspect and the interaction among them. A challenge faced by scientists lies in interpreting and understanding such large amount of mixed omics data. Yang et al.5 developed Mixed Graphical Models (MGM) to simultaneous model the distribution of mixed types of omics data – binary, categorical, continuous etc. Building on MGM, we developed a novel method, Mixed Graphical Models based clustering (MGMCluster), to identify patients’ cancer subtypes based on the integrated multi-omics data. ***We hypothesize that integration of multiple ‘omics’ data will improve the accuracy to cluster cancer patients into subgroups.***We will comprehensively evaluate MGMCluster by comparing its performance to the competing methods, PARADIGM and iCluster and then apply MGMCluster to find the common driving mutations among different cancers. Successful completion of this proposed study will have a major impact on our ability to use multi-omics data for cancer subtype identification, thus resulting in discovering new common driving mutations among different cancers which in turn sheds light on drug repurposing.

**Aim 1: To determine the accuracy of MGMCluster in identifying cancer subtypes**

Mixed Graphical Models (MGM) was developed to estimate the distribution of multi-scale omics data. Yet, its application in clustering cancer patients relying on their multi-omics data has not been demonstrated. We developed a novel method, MGMCluster, to incorporate expectation–maximization (**EM**) algorithm with MGM to cluster cancer patients into cancer subtypes. EM is an iterative method to find the most likely values of latent variables, where the latent variables in MGMCluster are the indicators of cancer subtypes. We hypothesize that MGMCluster has high accuracy in identifying cancer subtypes.

**(1) Simulation study.** To objectively compare the accuracy of MGMCluster and that of PARAGIGM and iCluster, we will perform a simulation study to compare their true positive rates, false positive rates, and receiver operating characteristics (ROCs). The distributions of different omics data will be estimated from publically available datasets. And we will create different scenarios of mutations, including mutations at different omic levels and/or aberrant interactions between omics.

**(2) Validation study using a breast invasive carcinoma dataset (1098 subjects) available at The Cancer Genome Atlas (TCGA).** To evaluate the performance of MGMCluster in analyzing a real biological data, we will conduct all three integrative methods, MGMCluster, PARADIGM, and iCluster, on a breast invasive carcinoma dataset, and then compare the breast cancer subtypes identified by each method with the ones identified by the well-accepted biomarkers.

**(3) Software:** We will create an R package that allows easy access to all the functions of this method. We will provide examples for implementing the method in a variety of cases. The software will be available through Bioconductor. We will maintain the package for a minimum of three years.

**Aim 2: To identify the common driving mutations among different cancers**

The integrative method, MGMCluster, not only will provide new insights for clustering cancer patients into subtypes but will also help find the commonality between the subtypes of different cancers. We hypothesize that subtypes of different cancers share similar interactome. Using the rich multi-omics data available on TCGA, we aim to identify the interactome similarity among different cancers and thus discover new interactome informed super-subtypes – the subtypes shared by different types of cancers.

**SIGNIFICANCE.**

There is heterogeneity in the molecular mechanisms of carcinogenesis in patients diagnosed with the same type of cancer. Therefore, precision medicine holds great promise for classifying cancer patients according to their molecular profile at the genomic, proteomic, transcriptomic, and epigenomic level. There is a body of literature on classifying cancer patients by high throughput omics data1-4, but methodology is still lacking for clustering cancer patients using multi-omics data, while taking into account the interactions among omic measures. *Our research will address this gap by developing the first methodology to identify cancer subtypes in the light of cancer molecular profile including mutation status, mRNA expression, methylation level, microRNA expression level, protein expression level, and more importantly their interactions.*

*This contribution will be significant because:*

* Successful classification of cancer patients is the corner stone of precision medicine.
* Increasingly available genomic, proteomic, transcriptomic, and epigenomic data provide an unprecedented opportunity for identifying cancer subtypes. The proposed method, MGMCluster, provides an opportunity to integrate multiple omics data to compressively comprehensively classify cancer subtypes.
* In comparison to other methods, MGMCluster identifies cancer subtypes more accurately since it is informed by the interactome -- the mutation status of the molecular entities at all omic levels, and the interactions among the entities of the same and different omic measures.
* Discovering interactome similarities between subtypes of different cancers could result in existing cancer drugs being retargeted to new cancers.
* Using MGMCluster to integrate multiple omics data will reveal more about the heterogeneity and molecular causes of cancer.

At the completion of this study, we will provide open access software to the research community. This software will allow researchers to identify the subtypes of their patients by integrating any omics data they have.

**INNOV ATION.**

Most of the current cancer subtype clustering approaches only utilize a single type of omics data. Patients with a similar profile at some omics levels, but distinct profiles at other omic levels, may be mistakenly clustered together. Only a few methods are available for integrating multiple omics data to cluster cancer subtypes. However, none of these integrative methods take the interaction between omic measures into consideration.

*The proposed research is innovative because:*

* To our knowledge, this is the first approach to study cancer heterogeneity based on cancer interactome.
* Our method incorporates the first graphical model that captures the joint distribution of multiple omics data. This state of the art mixed graphical model allows us to model the status of each molecular entity as well as the interactions among them.

The novel application of Mixed Graphical Models in clustering will empower interactome study through integrating multi-omics data. Interactome based clustering opens up new opportunities to understand the heterogeneity within cancers and the similarity among different cancers.

**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 subtypes based on their interactome -- molecular profiles at several ‘omic’ levels and the interactions within and between the 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 subtypes of cancer patients.* Accurate identification of cancer subtypes is the cornerstone of cancer precision medicine. MGMCluster uses expectation–maximization (EM) algorithm with Mixed Graphical Models (MGM) to classify cancer patients into subtypes. We assume a mixture of MGM for the multi-omics distribution of patients from different cancer subtypes, where the cancer subtypes are modeled as latent variables. EM iteratively use the current model to estimate the latent cancer subtypes, and then uses the latent value estimates to refine the model. To objectively compare the performance of MGMCluster with other two competing integrative clustering method, icluster6 and PARADIGM7, we will conduct a simulation study in which patients’ true cancer subtypes are known. In addition, with the breast invasive carcinoma multi-omics dataset hosted at 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, we hope to find the superior accuracy of MGMCluster over other integrated clustering methods.

*Approach*- Mixed Graphical Models (MGM) 5 is the first model that fully specifies a parametric multivariate distribution, P(X,Y,Z|), over a mixture of data types ,including binary, categorical and continuous

|  |
| --- |
|  |
| Figure 1 Mixed Graphical Model Each node represents a measuring entity (a copy number of a gene, a expression value of microRNA, or a express value of a mRNA) and each edge represents a dependency between two entities (within or across omics scales) |

data (**Fig. 1**). In MGMCluster, we extend MGM by including a latent variable, K, as the indicators of cancer subtypes. In consequence,

each cancer subtype follows distinct conditional distribution, P(X,Y,Z|). 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 follows:

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.

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.

We will simulate different scenarios of genomic perturbation. To simulate a virtual patient with unaltered number variation (CNV), microRNA and mRNA profile, we will first estimate the statistical distributions of each measuring unit of copy number variation (CNV), microRNA expression, and mRNA expression by using the data of 1098 matching normal samples in the breast invasive carcinoma dataset from TCGA. In particular, we assume CNV follows multinomial distribution8,



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) 9 from the 1098 matching normal samples. 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 10. With the independent empirical distributions of all measuring units at all omic levels and the correlations of the measuring units within and between the omic levels, we can jointly simulate a virtual patient with unaltered multi-omics profile.

We have collaborated with a biologist to design the different possible causes of breast invasive carcinoma including copy number variation, abnormal transcriptomics profile, and transcriptional dysregulation. We will first simulate three subtypes of breast invasive carcinoma. Each cancer subtype is caused by the similar mutation mechanism. Then, we will apply MGMCluster, PARADIGM, and iCluster to discover the three breast invasive carcinoma subtypes. Since we know the true subtypes existing in simulated samples, we can calculate the misclassification rate for each integrated clustering method. Furthermore, in order 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 integrative clustering methods. This will help determine the superior accuracy of MGMCluster. In addition, we will know the computational feasibility of MGMCluster in a large dataset.

We will also test the performance of the MGMCluster in a real biological dataset. We will employ the 1098 tumor samples of the breast invasive carcinoma dataset hosted at TCGA to compare the subtypes unveiled by MGMCluster, PARADIGM, iCluster and the well-accepted biomarkers3.

The outcome of the breast cancer study is a comparison of the subtypes discovered by the three integrative clustering methods and the one discovered by the well-accepted biomarkers. We expect to see a 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 discovered by well-accepted biomarkers due to the narrow view of a single type of molecules.

*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 the model sparsity 11. 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 a faster programming language 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.

**Aim 2: Identify the common driving mutations between breast cancer and ovarian cancer.**

*Hypothesis and rationale*- It has been shown that defining cancers as a disease not by its tissue of origin but by the genetic characteristics provides the possibility of repurposing the existing cancer drugs to treat other genetically alike cancer subtypes [ref]. However, finding the similarity of different cancers based on single omics profiles narrows our understanding of the similarity shared by cancers. The Cancer Genome Atlas (TCGA) project has created a comprehensive molecular characterization of 33 cancers at different molecular levels: single nucleotide polymorphism (SNP), copynumber variation (CNV), microRNA expression, mRNA expression and DNA methylation. This provides unprecedented opportunity to discover the interactome similarity of different cancers. The objective of this aim is to identify similar subtypes between breast cancer and ovarian cancer based on their interactomes. *We hypothesize that breast cancer and ovarian cancer share subtypes with similar interactome.*

We will apply MGMCluster on two possibly similar cancer types together: breast cancer and ovarian cancer. And determine if the samples in the same subgroup but from two different cancer types share similar molecular mechanism of carcinogenesis. Successful completion of this aim will discover similar subgroups of two different cancers and the shared causal molecular mechanism. This will be an evidence that cancers should be defined by the molecular features but not its tissue of origin. In addition, if we discover one subtype with available treatment is similar to another subtype from the other cancer type, the existing treatment can likely be repurposed. At the completion of this aim, we expect to discover the molecular similarity between different cancer types.

*Approach-* Multi-omics data for breast invasive carcinoma and Ovarian serous cystadenocarcinoma will be obtained from TCGA. The following table shows the types of available omics data types and number of available samples from TCGA:

| **Breast invasive carcinoma [BRCA]** | **Total** | **Exome** | **SNP** | **Methylation** | **mRNA** | **miRNA** |
| --- | --- | --- | --- | --- | --- | --- |
| **Cases** | [1098](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=BRCA&tumorNormal=TN&tumorNormal=T&tumorNormal=NT) | [1081](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=BRCA&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=12&platformType=7&platformType=41) | [1095](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=BRCA&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=1&platformType=4&platformType=40) | [1096](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=BRCA&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=2&platformType=42) | [1094](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=BRCA&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=3&platformType=5&platformType=27&platformType=38&platformType=43) | [1077](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=BRCA&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=6&platformType=28) |
| **Ovarian serous cystadenocarcinoma [OV]** | **Total** | **Exome** | **SNP** | **Methylation** | **mRNA** | **miRNA** |
| **Cases** | [586](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=OV&tumorNormal=TN&tumorNormal=T&tumorNormal=NT) | [536](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=OV&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=12&platformType=7&platformType=41) | [579](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=OV&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=1&platformType=4&platformType=40) | [584](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=OV&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=2&platformType=42) | [583](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=OV&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=3&platformType=5&platformType=27&platformType=38&platformType=43) | [582](http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&showMatrix=true&diseaseType=OV&tumorNormal=TN&tumorNormal=T&tumorNormal=NT&platformType=6&platformType=28) |  |

We estimate the size of this dataset is likely to be several terabytes. It will be stored and preprocessed by high-performance computing (HPC) at University of Arizona. We have already registered an account for bulk download, and tested the download speed. This dataset can be downloaded in approximately one week. We will follow the standard process recommended by TCGA for data preprocessing. Preprocess data will be stored in our workstation, which has Intel® Xeon® Processor E5-2650 v3 and 64 GB memory.

After obtaining all needed data from TCGA, we will apply MGMCluster on these two types of cancer using CNV, SNP, microRNA, mRNA, and methylation data. This computationally expensive analysis will be conducted on our local high-end workstation. Figure 1 was generated using simulated hypothetical dataset to illustrate the possible outcome.

Macintosh HD:Users:qikeli:Dropbox:Courses:IMB521GrantWriting:Aim2:clustering.pdf

Figure 3 Clustering of patients of two cancers One subtype of breast cancer and a subtype of ovarian cancer clustered together. This is possibly due to the similar molecular characteristics shared by the two subtypes.

To our knowledge, MGMCluster is the first method to jointly model different levels of omics data while taking into account the interactions between different measurements. It is our expectation that we will uncover new cancer subtypes and likely similar subtypes between different cancers. The hypothetical example in figure 1 shows a subtype of breast cancer has higher molecular similarity to a subtype of ovarian cancer than the other subtype of breast caner does. In this case we will further investigate the mixed graphical model (MGM) of these two similar subtypes. Since MGM captures the state of each node (each measuring unit) and each edge (interactions between units). Discovery of distinct alteration of the nodes and edges suggest the common molecular mechanism of carcinogenesis of the samples belonging to the subgroup. The discoveries, thus, provides the understandings of molecular commonality between two cancers, which can possibly inform the drug repurposing based on the common interactome.

*Problems and alternatives-* Integrating five types of omics data will be computationally intensive. If the computation is not feasible, we will further optimize the computation by parallel computing. If either of the molecularly similar subtypes has an existing treatment, the molecular similarity will be still valuable in understanding the carcinogenesis of the two cancers.

*Future directions-* With the novel methodology, MGMCluster, we can conduct a systematic survey of the heterogeneity of every 33 cancers profiled by TCGA and the interactome similarity shared among all 33 cancer types.

1 Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353-360 (2012).

2 Finak, G. *et al.* Stromal gene expression predicts clinical outcome in breast cancer. *Nature medicine* **14**, 518-527 (2008).

3 Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752 (2000).

4 Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics* **41**, 178-186 (2009).

5 Yang, E. *et al.* A General Framework for Mixed Graphical Models. *arXiv preprint arXiv:1411.0288* (2014).

6 Shen, R., Olshen, A. B. & 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* **25**, 2906-2912 (2009).

7 Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-i245 (2010).

8 Barnes, C. *et al.* A robust statistical method for case-control association testing with copy number variation. *Nature genetics* **40**, 1245-1252 (2008).

9 Aldrich, J. RA Fisher and the making of maximum likelihood 1912-1922. *Statistical Science* **12**, 162-176 (1997).

10 Spearman, C. The proof and measurement of association between two things. *The American journal of psychology* **15**, 72-101 (1904).

11 Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288 (1996).