Accepted for publication, *Algorithms Special Issue on Biological Knowledge Discovery from Big Data*

Molecular Subtyping and Outlier Detection in Human Disease Using the Paraclique Algorithm*

Ronald D. Hagan and Michael A. Langston
Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville, TN 37996
rhagan@vols.utk.edu, langston@tennessee.edu

Abstract. Recent discoveries of distinct molecular subtypes have led to remarkable advances in treatment for a variety of diseases. While subtyping via unsupervised clustering has received a great deal of interest, most methods rely on basic statistical or machine learning methods. At the same time, techniques based on graph clustering, particularly clique-based strategies, have been successfully used to identify disease biomarkers and gene networks. A graph theoretical approach based on the paraclique algorithm is described that can easily be employed to identify putative disease subtypes and serve as an aid in outlier detection as well. The feasibility and potential effectiveness of this method is demonstrated on publicly-available gene co-expression data derived from patient samples covering twelve different disease families.

Keywords: Molecular subtyping, outlier detection, paraclique algorithm, transcriptomic data

1. Introduction

It has long been established that many disease families exhibit a wide range of heterogeneity. This is especially true in cancer. Lung cancers, for example, fall into two overall types based on histological characteristics: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Although histological classification remains crucial, significant advances in the treatment of NSCLC over the last decade have centered around the development of therapies targeting subtypes at the molecular level, such as those defined by genetic mutations [1]. In particular, therapies targeting alterations in the epidermal growth factor receptor (EGFR) and anaplastic lymphoma kinase (ALK) genes have produced dramatic improvements in outcomes for patients in the underlying subgroups [2, 3]. In addition to providing new paths for treatment, advances in molecular subtyping allow practitioners to avoid needless high-risk therapies. For example, studies have identified transcriptomic signatures for chemo-resistance in both acute myeloid leukemia and breast cancer [4, 5]. Recent research has made positive steps towards targeted care for a variety of diseases, including Asthma, Alzheimer's, and Crohn's disease [6-8]. A key development driving these advances is the successful identification of molecular subtypes.

Given the potential impact of disease subtype identification, it is not surprising therefore that the search for effective clustering methods has become an intense area of interest. Traditional approaches such as *k*-means and hierarchical clustering have long been used to identify sets of genes or samples that exhibit similar expression patterns [9-11]. Machine learning techniques based on neural networks have been investigated as well [12-14]. Latent variable and mixture models have also been used [15-17]. Meanwhile, a graph theoretical approach is to model a set of genes or samples as vertices in a graph, with edges connecting them based on thresholding some similarity metric. A systematic comparison of clustering methods over well-annotated *S. cerevisiae* (baker's yeast) gene co-expression data can be found in [18], where it was shown that clique-centric graph theoretical algorithms generally outperform other approaches. Moreover,

* A preliminary version of a portion of this paper was presented at the International Workshop on Biological Knowledge Discovery from Big Data, held in Linz, Austria, in August, 2019.

the top-down paraclique algorithm introduced in [19] was found to possess considerable computational advantages over other clique-based tools. Maximal clique [20], for example, is output bound, while k-clique communities [21] is hobbled by bottom-up inefficiencies. Paraclique has seen prior application in transcriptomics [22], proteomics [23], epigenetics [24] and the exposome [25] as well as in the study of specific diseases including lung cancer [26], diabetes [27], allergic rhinitis [28] and community-acquired pneumonia [29], and even in investigations of the effects of radiation on living organisms [30]. Nevertheless, to the best of our knowledge, this paper represents the first attempt to gauge paraclique's potential merit in the context of molecular disease subtyping.

To address this gap, we describe an initial study of putative subtypes based on molecular signatures using the paraclique method. Our technique is general and applies easily to other types of data such as protein interaction, metabolite abundance, or DNA methylation profiles, but we focus our experimentation on gene co-expression data thanks to its relative quality and ubiquity. In addition to subtype discovery, we will show how our techniques can be used to help pinpoint potential outliers, providing an automatic means for the identification of suspected data collection errors such as mislabeled samples or misdiagnosed patients. We hasten to observe, however, that biological variation can be inscrutable, inconsistent, and unpredictable. No method is therefore likely to be extraordinarily accurate. We will address this and related issues in the sequel.

This paper is organized as follows. In the next section, we provide a brief review of the paraclique algorithm and discuss details of our workflow for subtyping in gene co-expression data. In Section 3, we outline our testing procedures, provide GO enrichment results that indicate functional biological relevance of the subtypes we identify, and describe additional testing with labeled data over known subtypes that demonstrate the fidelity of further stratification using this approach. In Section 4, we consider outlier detection and discuss how methods such as these can help address this problem. In a final section, we summarize results, place them in context, and consider avenues for future work.

2. Methodology

Clique-centric methods have long been used in a wide variety of applications [31]. On real and noisy data, however, clique finders may be inherently prone to high false negative rates. Indeed, an entire clique may be missed if even a single edge is lost. Thus, the paraclique algorithm is an effort to ameliorate difficulties posed by noise. Its essential strategy is first to isolate a maximum clique, and then expand it by glomming onto any new vertex that is adjacent to all but some predefined number of vertices already in this clique. This number is termed the glom term, g. An illustration of paraclique construction with $g = 2$ is provided in Figure 1. Paraclique details and a thorough discussion of clique selection, edge weights, densities, and other important algorithmic features can be found in [32, 33]. Web-based versions of paraclique and related tools are available to the community via GrAPPA [34].

In this effort, we were mainly concerned with case-control transcriptomic data, for which we applied an initial filtering step to limit the effects of confounding factors. False discovery rate adjusted p-values for the differential expression of genes between case data and control data were calculated using the Benjamini-Hochberg method [35] accessed via the EntropyExplorer R package [36]. Only those genes with p-values less than or equal to 0.1 were retained. The motivation for such a filter was to restrict attention to genes of potential interest in the differential diagnosis of disease. After all, we wanted to concentrate on potential disease subtypes and not be distracted by irrelevant subgroups such as age, ethnicity, or hair color. Once filtering was complete, we focused our attention only on case data, and reversed the roles of variables and

correlations. We therefore calculated pairwise Pearson correlation coefficients between samples (not genes), and across their corresponding lists of expression levels (not patients). We then thresholded the resultant correlation matrix using spectral methods as in [37], and constructed an unweighted graph with vertices representing samples and edges between highly correlated sample pairs. Once this graph had been created, we invoked the paraclique algorithm to extract dense, noise resilient subgraphs. Thus, each such subgraph represented a putative subtype or outlier.

For consistency, and because this work mainly represents a proof of concept, we set the glom term to g=1 throughout this effort. Depending on the data under study, however, crisper results may naturally be anticipated with fine tuning. In [27], for example, a glom term of g=5 was found to produce superior ontological enrichments when studying non-obese diabetic mice as a model of type 1 diabetes mellitus.

3. Experimental Results

3.1 Discussion

We applied this novel analytical approach to a dozen sets of publicly-available gene co-expression data obtained from the Gene Expression Omnibus (GEO). These data were selected because they provide a wide cross-section of human disease, and because each has both a case and a control group for the aforementioned filtering task. Table 1 provides an overview of the datasets we studied.

Our investigation into the effectiveness of this proposed new methodology was focused on two guiding questions: (1) are these tools capable of reliably and robustly identifying putative subtypes, and (2) are these subtypes appropriate to the associated disease as supported by biological evidence from clinical, published, analytical or other orthogonal information source(s)?

The answer to the first question seems to be an unequivocal yes. As summarized in Table 2, our methods decomposed raw data into putative subtypes in ten of our datasets. In the case of Asthma, for example, every patient sample fell into some paraclique. In other cases, patients were sometimes left unclassified, which is hardly surprisingly given limitations on dataset sizes coupled with possible extremes in disease as well as sample heterogeneity. Only for Parkinson's disease and Type 2 Diabetes were no subtypes identified. It's probably no coincidence then that these two diseases also have by far the smallest datasets, especially in light of clinical subtyping evidence to the contrary [38, 39].

The second question is considerably more difficult to answer because it depends on the availability of alternate, non-transcriptomic data sources. We therefore followed a two-prong approach in putative subtype comparisons. First, we calculated GO enrichments and their associated p-values for the top 100 differentially expressed genes in each paraclique. These results and their corresponding GO categories are summarized in Table 3. In every case, we found statistical evidence for biological significance among the genes separating samples into subgroups, with enrichment p-values ranging from 1.1E-4 for asthma to 4.92E-46 for prostate cancer. Next, we performed a literature search to check the top scoring genes for involvement in known subtypes. As such, this is at best a hit or miss proposition, and one depending for each disease on whether the research community has studied subtyping issues, found results, and published them in venues that we were able to search. Despite these obstacles, however, we found strong evidence in print to support our putative subtype decompositions for four of the diseases we studied. These are asthma, breast cancer, chronic lymphocytic leukemia, and colorectal cancer.

3.2 A Search for Unrecognized Subtypes

Asthma

The incidence of asthma in the U.S has been on the rise for two decades. It is currently estimated that nearly one in ten children under 18 are asthmatic. The risk for some groups is based largely on ethnicity (particularly African American and Puerto Rican), with incidence among those with lower socioeconomic status rising as high as one in six [40].

GEO series GSE4302 data was derived from the Affymetrix Human Genome U133 Plus 2.0 Array, and is designed to identify genes associated with response to corticosteroid treatment in asthmatics [41]. It consists of transcriptomic data taken from the epithelial airway brushings of 42 asthmatics, 28 healthy subjects and 16 smokers. To avoid potential confounds, we discarded data taken from smokers and used only the healthy subjects as controls.

Filtering reduced the number of probes from 54,676 to 2322. Our method produced three paracliques with respective sizes 31, 8 and 3 that were stable until they began to merge as the threshold was lowered below 0.93. The 100 most differentially expressed genes across the two larger putative subtypes included CLCA1, periostin, and ovalbumin, which are all known to serve as markers of a Th2-high endotype of asthma [42].

Breast Cancer

Genetic factors have long been known to play a significant role in breast cancer. Studies have shown that in families with at least four breast cancer cases, most can be linked to mutations in either BRCA1 or BRCA2 genes [43, 44]. Moreover, breast cancer has a variety of known subtypes that significantly impact prognosis and treatment. For example, tumors negative for estrogen receptors, progesterone receptors, and the expression of HER2 are indicative of triple-negative breast cancer, a subtype identified with higher risk of recurrence and a five-year mortality rate [45].

GEO series GSE10810 data was also derived from the Affymetrix Human Genome U133 Plus 2.0 Array, although values for only 18,382 probes were provided. This study was designed to investigate links between gene co-expression and phenotypic breast cancer differences [46], and contains data for 31 tumor samples and 27 healthy tissues.

Filtering reduced the number of probes to 11,531. Our tools produced two paracliques of size 22 and 5 that persisted to a threshold of 0.8, and left four tumor samples unclassified. The 100 most differentially expressed genes between these putative subtypes include SLC39A6, S100a4, AGR3, Cd24, and epcam, all of which have been reported in the literature as biomarkers for distinct breast cancer phenotypes [47-51].

Chronic Lymphocytic Leukemia

Chronic lymphocytic leukemia is one of the most common types of leukemia, with pathogenesis characterized by an overproduction of neoplastic B cells in the bloodstream. The current median age at diagnosis is 65, with males affected more often than females [52]. Chronic lymphocytic leukemia typically presents with a slow progression in which patients are able to enjoy a more or less a normal life expectancy. In some cases, however, chronic lymphocytic leukemia can be aggressive, with death occurring less than five years after the onset of symptoms.

GEO series GSE8835 data was instead derived from the Affymetrix Human Genome U133A Array with 22,283 probes, and was designed to study the effects of chronic lymphocytic leukemia on T cells in peripheral blood [53]. The study comprised 24 CD4 cell samples from chronic lymphocytic leukemia patients and 12 CD4 cell samples from healthy, age-matched donors.

Filtering reduced the number of probes to 1338. At a threshold of 0.8, our tools produced two paracliques of size 4 and 18, leaving 2 samples unclassified. The most differentially expressed genes across these two putative subtypes included ZAP-70, previously identified as the best discriminator of Ig-mutated and Ig-unmutated chronic lymphocytic leukemia [54].

Colorectal Cancer

The incidence of colorectal cancer has been in decline since the mid 1980's [55]. Despite this significant drop in prevalence, it still accounts for both the third highest number of new cases of cancer, and the third highest number of cancer deaths each year [56]. As with breast cancer, there are known hereditary links to this disease. For example, a mutation of the gene APC is responsible for two syndromes, Familial Adenomatous Polyposis and Hereditary Nonpolyposis Colorectal Cancer, that each carry a significant increase in the risk of developing colorectal cancer [57].

GEO series GSE9348 data was again derived from the Affymetrix U133 Plus 2 array, and was intended to search for transcriptomic signatures of early stage colorectal cancer that is prone to metastasis [58]. The study contains gene co-expression data that was taken from 70 colorectal cancer patient tumors as well as tissues from 12 healthy subjects who were matched by age and ethnicity.

Filtering reduced the number of probes from 54675 to 22968. At a threshold of 0.87, our tools produced two paracliques of size 63 and 5, covering all but two of the case samples. The list of 100 genes most differentially expressed between these two putative subtypes include Cd24, identified as a prognostic marker for colorectal cancer [59] as well as OLFM4, indicated in as a marker for tumor differentiation and progression [60, 61].

3.3 Alignment with Previously Known Subtypes

The experimental effort just described suggests that our methods have the potential to identify both known and novel subtypes, as based on biologically relevant genetic signatures. The lack of any widespread established ground truth, however, places a limitation on any in-depth interpretation of these results. In an effort to address this shortcoming, we identified two sets of publicly-available data on GEO that include metadata labeling in the form of known subtyping information. These are based on gastric cancer and non-small cell lung cancer. Because our intent is to identify and contrast novel subtypes in disease, our metric of interest is patient stratification.

Gastric Cancer

GEO series GSE35809 data, from the Affymetrix Human Genome U133 Plus 2.0 Array, was derived from 70 primary gastric tumors intended for use as a validation set for subtype classifier testing [62-64]. The data contains values for 54675 probes. Arrays are subdivided into a collection of 29 identified as coming from proliferative tumors, 26 from invasive, and 15 from metabolic.

Filtering was irrelevant, because no healthy tissues were studied that could be used as controls. At a threshold of 0.955, paraclique produced subsets of size 29 and 16 and performed admirably. All but one of the invasive samples it classified were placed in the first paraclique, while all but one of the proliferative samples it classified were placed in the second. Metabolic samples proved only slightly more challenging, with 75% of those classified placed in the first paraclique. See Table 4.

Non-Small Cell Lung Cancer

GEO series GSE10245 data, also from the Affymetrix Human Genome U133 Plus 2.0 Array with 54675 probes, was derived from 40 adenocarcinoma tumors and 18 squamous cell carcinoma tumors, NSCLC's two most prevalent subtypes. These data were intended to provide a basis for studying co-expression differences between these two cancers [65].

Filtering was again irrelevant. Paraclique performed quite well on this data too. At a threshold of 0.94, it produced three subsets of size 26, 12 and 8. Roughly 74% of the adenocarcinoma samples were placed in the first paraclique while the second contained none, and 80% of the squamous cell carcinoma samples were placed in the second paraclique while the third had none. Again, see Table 4.

Comparison with Other Methods

We sought to compare this basic and untuned version of the paraclique algorithm with well-known strategies such as k-means and hierarchical clustering, as implemented in core-R through the functions *kmeans( )* and *hclust( )*.

Results for the k-means method were mixed. It proved extremely successful on the gastric cancer data. There it divided samples into two subsets of size 26 and 44. All but one of the invasive samples were placed in the first cluster, while all of the proliferative and all but one of the metabolic samples were placed in the second. But k-means failed completely on the non-small cell lung cancer data. Samples were divided into subsets of size 28 and 30, with both the adenocarcinoma and the squamous cell carcinoma samples spread almost evenly across these two clusters. The hierarchical approach was also a rather uneven performer. On the gastric cancer data, it divided samples into two subsets of size 33 and 37. While all of the proliferative samples found their way to the second cluster, the first contained 84% of the invasive and about 73% of the metabolic. On the non-small cell lung cancer data, it divided samples into subsets of size 9 and 49. All the adenocarcinoma samples were admirably grouped in the second cluster, but the squamous cell carcinoma samples were not convincingly stratified at all, with exactly half placed in each cluster. These results are also summarized in Table 4.

As demonstrated by these experiments, the paraclique methodology can provide excellent patient stratification, further motivating the use of graph theoretical methods to differentiate samples based on their underlying genetic signatures. Such stratification is not perfect, of course, nor should we expect it to be given data limitations and biological variability. Moreover, unlike techniques such as k-means and hierarchical clustering, patients are not forced into a cluster under paraclique, as is evidenced by the 25 samples it left unclassified in the gastric cancer data. We suggest therefore that the tools we have described here may be best suited to fast screening tasks, for example, when transcriptomic data is relatively easy to obtain. Once clinical and/or additional forms of data have been collected, histological and other more laborious techniques will likely help provide more comprehensive subtyping of entire patient populations.

4. Outlier Detection

The methodology just described is readily extensible to automating the task of outlier detection. This follows from the observation that an outlier would be expected to appear as its own distinct subtype, and not reside in a paraclique of even modest size. Although detection may be accomplished with our algorithms in several ways, we endorse the use of thresholding, as follows. A normalized threshold of 0.0 will of course produce a single large clique, and a threshold of 1.0 will generally yield an edgeless graph, under the assumption that no two samples are perfectly correlated. As the threshold value is lowered from 1.0, the effect on cliques and paracliques is slightly nuanced. As more and more edges are added, cliques and paracliques will get larger but also begin to merge. If a vertex consistently fails to join any of these dense subgraphs, then the sample it represents is flagged as a potential outlier. The process is illustrated in Figure 2. At this point it may be tempting simply to single out isolated vertices, but at any given threshold a vertex may of course have a variety of neighbors and yet still be a member of no paraclique.

While this approach has intuitive appeal, we conducted a series of six experiments using known misclassifications to test its limitations. We formed test instances by introducing data from one randomly chosen healthy sample into data from the case samples for breast cancer (GSE10810), chronic lymphocytic leukemia (GSE8835), colorectal cancer (GSE9348), lung cancer (GSE7670), pancreatic cancer (GDS4102), and prostate cancer (GSE6919). For the breast, colorectal, lung, and pancreatic cancer sets, we observed that the normal sample was either the last or the next to last vertex to be drawn into the final paraclique. For the chronic lymphocytic leukemia and prostate cancer sets, we found instead that the healthy sample fell into a large paraclique early on and stayed there. From our previous experience with outlier detection [66], these observations suggest to us that although paraclique has a pronounced potential to serve as an automated outlier screening tool, feature selection [67] should probably first be performed to reduce any positive bias that results from whole genome correlations. We will revisit this topic in the next section.

5. Summary, Discussion and Directions for Future Research

We have developed and described a disease classification strategy based on the paraclique algorithm that can identify putative subtypes, segregating samples based on signatures in their molecular profiles. Although our tools are easily applicable to many types of biological data, we have focused on gene co-expression data largely thanks to its overall quality and availability. We have analyzed high throughput data taken from a dozen different disease samples obtained from the Gene Expression Omnibus, and sought to validate the significance of our findings by reviewing the literature and examining ontological enrichment for the biological relevance of genes differentially expressed across putative subtypes. We also performed testing over data augmented with phenotypic information for known subtypes. Overall, the results of this study indicate a strong utility for this approach in the confirmation of known, and the discovery of novel, disease subtypes. Additionally, we described the extension of our methodology to the task of outlier identification. By iteratively lowering the threshold and re-running the paraclique algorithm, we can detect samples resistant to subtype coalescence. Such a finding can point to critical clinical errors such as tissue misclassifications and/or patient misdiagnoses. Throughout, our aim has been to employ scalable, cutting edge graph theoretical methods that can help automate the disease subtyping process, which can in turn accelerate the pace of discovery and lead to improvements in targeted therapies.

We emphasize that this exploratory effort has focused exclusively on unsupervised techniques and tools that require no prior knowledge. To keep things simple, we even refrained from fine tuning the glom term

for each dataset. This therefore bolsters the argument that the methods we have espoused are really quite effective. In large-scale clinical applications, however, techniques such as feature selection and paraclique anchoring will almost surely prove helpful to narrow the focus on genes or other variables of interest and their disease-associated relationships. In the context of community-acquired pneumonia, for example, we have previously found it advantageous to anchor paraclique analytics at the interleukin genes *IL-6* and *IL-10*. See [68].

To place these results in proper context, we note that any subtyping method based on tissue morphology or molecular signatures is almost certain to be highly imperfect. And this holds true whether it be implemented *in silico* or conducted manually by a human pathologist. A 2015 study [38] underscores this problem. There, an expert panel of pathologists created a baseline diagnosis based on consensus of opinion for 240 breast tissue biopsies with samples that included malignancy, pre-cancerous cells and benign tumors. Pathologists from eight states with at least one year of experience in diagnosing cancer were then invited to examine these samples. 115 of them completed their analysis and provided their best diagnoses. Although findings showed that 96% of the invasive breast cancer samples had been diagnosed in concordance with expert consensus, 13% of the diagnoses underreported the severity of stage I breast cancer, while 48% (17%) underreported (overreported) the severity of precancerous samples. False negatives and false positives such as these can have devastating effects on patients. They may also lead to a wide spectrum of poor outcomes that includes excessive delay, unnecessary treatment, additional expense, needless worry, and even premature morbidity and death.

Finally, we wish to emphasize that this work represents but a first step in determining the utility of paraclique in the molecular subtyping of disease. Although clique-based methods have been used as a basis for tasks such as biomarker detection and gene network elucidation, disease subtyping has received surprisingly little attention. In future work, it would thus be interesting to see systematic comparisons of this and other emergent subtyping technologies. Numerous other research directions beckon. For example, we would like to gain a better understanding of the impacts of improved feature selection, and see extensions of the basic method to multiple heterogeneous data types, an area that has attracted a flurry of recent attention [69-71]. Collaborative opportunities to partner with disease specialists may of course also help in subtyping verification via graph theoretical methods at large. Better thresholding and filtering methods may be studied as well, in hopes of increasing the accuracy of subtyping and in turn reducing the likelihood of confounding factors. In conclusion, we observe that the overall approach we have described can be applied to numerous other sorts of biological data, as well as data from application domains as diverse as cyberattack detection and social network analysis.

Acknowledgements

This research has been supported in part by the National Institute of Alcohol Abuse and Alcoholism and the National Institute on Drug Abuse under grant R01AA018776, by the National Institute of Diabetes and Digestive and Kidney Diseases under grant R01DK125586, and by the Department of Veterans Affairs under grant HX002680. We thank the anonymous reviewers for their careful reading, thoughtful critiques, and helpful comments.

References

[1] P. Savas, B. Hughes, and B. Solomon, "Targeted Therapy in Lung Cancer: IPASS and Beyond, Keeping Abreast of the Explosion of Targeted Therapies for Lung Cancer," *Journal of Thoracic Disease,* vol. 5, no. Suppl 5, pp. S579, 2013.

[2] T. S. Mok, Y.-L. Wu, S. Thongprasert, C.-H. Yang, D.-T. Chu, N. Saijo, P. Sunpaweravong, B. Han, B. Margono, and Y. Ichinose, "Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma," *New England Journal of Medicine,* vol. 361, no. 10, pp. 947-957, 2009.

[3] A. T. Shaw, D.-W. Kim, K. Nakagawa, T. Seto, L. Crinó, M.-J. Ahn, T. De Pas, B. Besse, B. J. Solomon, and F. Blackhall, "Crizotinib versus chemotherapy in advanced ALK-positive lung cancer," *New England Journal of Medicine,* vol. 368, no. 25, pp. 2385-2394, 2013.

[4] C. P. Leith, K. J. Kopecky, J. Godwin, T. McConnell, M. L. Slovak, I.-M. Chen, D. R. Head, F. R. Appelbaum, and C. L. Willman, "Acute myeloid leukemia in the elderly: assessment of multidrug resistance (MDR1) and cytogenetics distinguishes biologic subgroups with remarkably distinct responses to standard chemotherapy. A Southwest Oncology Group study," *Blood, The Journal of the American Society of Hematology,* vol. 89, no. 9, pp. 3323-3329, 1997.

[5] J. M. Balko, R. S. Cook, D. B. Vaught, M. G. Kuba, T. W. Miller, N. E. Bhola, M. E. Sanders, N. M. Granja-Ingram, J. J. Smith, and I. M. J. N. m. Meszoely, "Profiling of residual breast cancers after neoadjuvant chemotherapy identifies DUSP4 deficiency as a mechanism of drug resistance," *Nature Medicine,* vol. 18, no. 7, pp. 1052-1059, 2012.

[6] M. E. Kuruvilla, F. E.-H. Lee, and G. B. Lee, "Understanding asthma phenotypes, endotypes, and mechanisms of disease," *Clinical Reviews in Allergy and Immunology,* vol. 56, no. 2, pp. 219-233, 2019.

[7] G. Di Fede, M. Catania, E. Maderna, R. Ghidoni, L. Benussi, E. Tonoli, G. Giaccone, F. Moda, A. Paterlini, and I. Campagnani, "Molecular subtypes of Alzheimer's disease," *Scientific Reports,* vol. 8, no. 1, pp. 1-14, 2018.

[8] M. Weiser, J. M. Simon, B. Kochar, A. Tovar, J. W. Israel, A. Robinson, G. R. Gipson, M. S. Schaner, H. H. Herfarth, and R. B. Sartor, "Molecular classification of Crohn's disease reveals two clinically relevant subtypes," *Gut,* vol. 67, no. 1, pp. 36-42, 2018.

[9] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, and X. Yu, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature,* vol. 403, no. 6769, pp. 503-511, 2000.

[10] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences,* vol. 95, no. 25, pp. 14863-14868, 1998.

[11] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, and L. A. Akslen, "Molecular portraits of human breast tumours," *Nature,* vol. 406, no. 6797, pp. 747-752, 2000.

[12] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE,* vol. 78, no. 9, pp. 1464-1480, 1990.

[13] F. Luo, L. Khan, F. Bastani, I.-L. Yen, and J. Zhou, "A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles," *Bioinformatics,* vol. 20, no. 16, pp. 2605-2617, 2004.

[14] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences,* vol. 96, no. 6, pp. 2907-2912, 1999.

[15] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics,* vol. 25, no. 22, pp. 2906-2912, 2009.

[16] F. Ambrogi, E. Biganzoli, P. Querzoli, S. Ferretti, P. Boracchi, S. Alberti, E. Marubini, and I. Nenci, "Molecular subtyping of breast cancer from traditional tumor marker profiles using parallel clustering methods," *Clinical Cancer Research,* vol. 12, no. 3, pp. 781-790, 2006.

[17] J. Wessman, T. Paunio, A. Tuulio-Henriksson, M. Koivisto, T. Partonen, J. Suvisaari, J. A. Turunen, J. Wedenoja, W. Hennah, and O. P. Pietiläinen, "Mixture model clustering of phenotype features reveals evidence for association of DTNBP1 to a specific subtype of schizophrenia," *Biological psychiatry,* vol. 66, no. 11, pp. 990-996, 2009.

[18] J. J. Jay, J. D. Eblen, Y. Zhang, M. Benson, A. D. Perkins, A. M. Saxton, B. H. Voy, E. J. Chesler, and M. A. Langston, "A systematic comparison of genome-scale clustering algorithms," *BMC Bioinformatics,* vol. 13, no. 10, pp. S7, 2012/06/25, 2012.

[19] E. J. Chesler, and M. A. Langston, "Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data," *Systems Biology and Regulatory Genomics*, E. Eskin, ed., pp. 150–165: Springer, 2006.

[20] C. Bron, and J. Kerbosch, "Algorithm 457: finding all cliques of an undirected graph," *Communications of the ACM,* vol. 16, no. 9, pp. 575-577, 1973.

[21] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature,* vol. 435, no. 7043, pp. 814-818, 2005.

[22] M. A. Langston, A. D. Perkins, A. M. Saxton, J. A. Scharff, and B. H. Voy, "Innovative Computational Methods for Transcriptomic Data Analysis: A Case Study in the Use of FPT for Practical Algorithm Design and Implementation," *The Computer Journal,* vol. 51, no. 1, pp. 26-38, 2008.

[23] A. Schoenrock, B. Samanfar, S. Pitre, M. Hooshyar, K. Jin, C. A. Phillips, H. Wang, S. Phanse, K. Omidi, Y. Gui, M. Alamgir, A. Wong, F. Barrenas, M. Babu, M. Benson, M. A. Langston, J. R. Green, F. Dehne, and A. Golshani, "Efficient prediction of human protein-protein interactions at a global scale," *BMC Bioinformatics,* vol. 15, pp. 383, Dec 10, 2014.

[24] D. Macartney-Coxson, M. C. Benton, R. Blick, R. S. Stubbs, R. D. Hagan, and M. A. Langston, "Genome-wide DNA methylation analysis reveals loci that distinguish different types of adipose tissue in obese individuals," *Clin Epigenetics,* vol. 9, pp. 48, 2017.

[25] M. A. Langston, R. S. Levine, B. J. Kilbourne, G. L. Rogers, A. D. Kershenbaum, S. H. Baktash, S. S. Coughlin, A. M. Saxton, V. K. Agboto, D. B. Hood, M. Y. Litchveld, T. J. Oyana, P. Matthews-Juarez, and P. D. Juarez, "Scalable combinatorial tools for health disparities research," *Int J Environ Res Public Health,* vol. 11, no. 10, pp. 10419-43, Oct 10, 2014.

[26] M. C. Grubb, B. J. Kilbourne, and C. Kilbourne, "Socioeconomic, Environmental and Geographic Factors and United States Lung Cancer Mortality, 1999-2009," *Family Medicine and Community Health,* vol. 5, pp. 3-12, 2017.

[27] J. D. Eblen, I. C. Gerling, A. M. Saxton, J. Wu, J. R. Snoddy, and M. A. Langston, *Graph Algorithms for Integrated Biological Analysis, with Applications to Type 1 Diabetes Data*, p.^pp. 207-22: World Scientific, 2009.

[28] S. Bruhn, F. Barrenas, R. Mobini, B. A. Andersson, S. Chavali, B. S. Egan, E. Hovig, G. K. Sandve, M. A. Langston, G. Rogers, H. Wang, and M. Benson, "Increased expression of IRF4 and ETS1 in CD4+ cells from patients with intermittent allergic rhinitis," *Allergy,* vol. 67, no. 1, pp. 33-40, Jan, 2012.

[29] O. M. Peck Palmer, G. Rogers, S. Yende, D. C. Angus, G. Clermont, and M. A. Langston, "Graph Theoretical Analysis of Genome-Scale Data: Examination of Gene Activation Occurring in the Setting of Community-Acquired Pneumonia," *Shock,* vol. 50, no. 1, pp. 53-59, Jul, 2018.

[30] B. H. Voy, J. A. Scharff, A. D. Perkins, A. M. Saxton, B. Borate, E. J. Chesler, L. K. Branstetter, and M. A. Langston, "Extracting Gene Networks for Low-Dose Radiation using Graph Theoretical Algorithms," *PLoS Comput Biol,* vol. 2, no. 7, pp. e89, Jul 21, 2006.

[31] I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo, "The Maximum Clique Problem," *Handbook of Combinatorial Optimization*, D.-Z. Du and P. M. Pardalos, eds.: Kluwer Academic Publishers, 1999.

[32] R. D. Hagan, M. A. Langston, and K. Wang, "Lower Bounds on Paraclique Density," *Discrete Applied Mathematics,* vol. 204, pp. 208-212, 2016.
[33] Y. Lu, C. A. Phillips, E. J. Chesler, and M. A. Langston, "Clique Selection and its Effect on Paraclique Enrichment: An Experimental Study," in Proceedings, International Conference on Bioinformatics and Computational Biology, San Francisco, California, 2020.
[34] "Graph Algorithms Pipeline for Pathway Analysis," https://grappa.eecs.utk.edu.
[35] Y. Benjamini, and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 57, no. 1, pp. 289-300, 1995.
[36] K. Wang, C. A. Phillips, A. M. Saxton, and M. A. Langston, "EntropyExplorer: an R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression," *BMC Research Notes,* vol. 8, pp. 832, 12/30, 2015.
[37] A. D. Perkins, and M. A. Langston, "Threshold Selection in Gene Co-Expression Networks Using Spectral Graph Theory Techniques," *BMC Bioinformatics,* vol. 10, 2009.
[38] E. D. Pablo-Fernández, A. J. Lees, J. L. Holton, and T. T. Warner, "Prognosis and Neuropathologic Correlation of Clinical Subtypes of Parkinson Disease," *JAMA Neurology,* vol. 76, no. 4, pp. 470-479, 2019.
[39] E. R. Pearson, "Type 2 Diabetes: A Multifaceted Disease," *Diabetologia,* vol. 62, no. 7, pp. 1107-1112, 2019.
[40] E. T. Bope, and R. D. Kellerman, *Conn's Current Therapy 2016*: Elsevier Health Sciences, 2015.
[41] P. G. Woodruff, H. A. Boushey, G. M. Dolganov, C. S. Barker, Y. H. Yang, S. Donnelly, A. Ellwanger, S. S. Sidhu, T. P. Dao-Pick, and C. Pantoja, "Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids," *Proceedings of the National Academy of Sciences,* vol. 104, no. 40, pp. 15858-15863, 2007.
[42] P. G. Woodruff, "Subtypes of asthma defined by epithelial cell expression of messenger RNA and microRNA," *Annals of the American Thoracic Society,* vol. 10, no. Supplement, pp. S186-S189, 2013.
[43] D. Ford, D. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, D. Bishop, B. Weber, G. Lenoir, and J. Chang-Claude, "Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families," *The American Journal of Human Genetics,* vol. 62, no. 3, pp. 676-689, 1998.
[44] D. Easton, D. Bishop, D. Ford, and G. Crockford, "Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium," *American journal of human genetics,* vol. 52, no. 4, pp. 678, 1993.
[45] R. Dent, M. Trudeau, K. I. Pritchard, W. M. Hanna, H. K. Kahn, C. A. Sawka, L. A. Lickley, E. Rawlinson, P. Sun, and S. A. Narod, "Triple-negative breast cancer: clinical features and patterns of recurrence," *Clinical cancer research,* vol. 13, no. 15, pp. 4429-4434, 2007.
[46] V. Pedraza, J. A. Gomez-Capilla, G. Escaramis, C. Gomez, P. Torné, J. M. Rivera, A. Gil, P. Araque, N. Olea, and X. Estivill, "Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness," *Cancer,* vol. 116, no. 2, pp. 486-496, 2010.
[47] N. Srour, M. A. Reymond, and R. Steinert, "Lost in translation? A systematic database of gene expression in breast cancer," *Pathobiology,* vol. 75, no. 2, pp. 112-118, 2008.
[48] S. de Silva Rudland, L. Martin, C. Roshanlall, J. Winstanley, S. Leinster, A. Platt-Higgins, J. Carroll, C. West, R. Barraclough, and P. Rudland, "Association of S100A4 and osteopontin with specific prognostic factors and survival of patients with minimally invasive breast cancer," *Clinical Cancer Research,* vol. 12, no. 4, pp. 1192-1200, 2006.
[49] E. R. King, C. S. Tung, Y. T. Tsang, Z. Zu, G. T. Lok, M. T. Deavers, A. Malpica, J. K. Wolf, K. H. Lu, and M. J. Birrer, "The anterior gradient homolog 3 (AGR3) gene is associated with differentiation and survival in ovarian cancer," *The American journal of surgical pathology,* vol. 35, no. 6, pp. 904, 2011.

[50] S. Ricardo, A. F. Vieira, R. Gerhard, D. Leitão, R. Pinto, J. F. Cameselle-Teijeiro, F. Milanezi, F. Schmitt, and J. Paredes, "Breast cancer stem cell markers CD44, CD24 and ALDH1: expression distribution within intrinsic molecular subtype," *Journal of clinical pathology*, pp. jcp. 2011.090456, 2011.

[51] T. Yamashita, M. Forgues, W. Wang, J. W. Kim, Q. Ye, H. Jia, A. Budhu, K. A. Zanetti, Y. Chen, and L.-X. Qin, "EpCAM and α-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma," *Cancer research*, vol. 68, no. 5, pp. 1451-1461, 2008.

[52] C. Rozman, and E. Montserrat, "Chronic lymphocytic leukemia," *New England Journal of Medicine*, vol. 333, no. 16, pp. 1052-1057, 1995.

[53] G. Görgün, T. A. Holderried, D. Zahrieh, D. Neuberg, and J. G. Gribben, "Chronic lymphocytic leukemia cells induce changes in gene expression of CD4 and CD8 T cells," *The Journal of clinical investigation*, vol. 115, no. 7, pp. 1797-1805, 2005.

[54] A. Wiestner, A. Rosenwald, T. S. Barry, G. Wright, R. E. Davis, S. E. Henrickson, H. Zhao, R. E. Ibbotson, J. A. Orchard, and Z. Davis, "ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile," *Blood*, vol. 101, no. 12, pp. 4944-4951, 2003.

[55] B. K. Edwards, E. Ward, B. A. Kohler, C. Eheman, A. G. Zauber, R. N. Anderson, A. Jemal, M. J. Schymura, I. Lansdorp-Vogelaar, and L. C. Seeff, "Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates," *Cancer*, vol. 116, no. 3, pp. 544-573, 2010.

[56] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, no. 1, pp. 7-30, 2016.

[57] K. W. Kinzler, and B. Vogelstein, "Lessons from hereditary colorectal cancer," *Cell*, vol. 87, no. 2, pp. 159-170, 1996.

[58] Y. Hong, T. Downey, K. W. Eu, P. K. Koh, and P. Y. Cheah, "A 'metastasis-prone'signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics," *Clinical and Experimental Metastasis*, vol. 27, no. 2, pp. 83-90, 2010.

[59] L. Belov, J. Zhou, and R. I. Christopherson, "Cell surface markers in colorectal cancer prognosis," *International journal of molecular sciences*, vol. 12, no. 1, pp. 78-113, 2010.

[60] D. Besson, A.-H. Pavageau, I. Valo, A. Bourreau, A. Bélanger, C. Eymerit-Morin, A. Moulière, A. Chassevent, M. Boisdron-Celle, and A. Morel, "A quantitative proteomic approach of the different stages of colorectal cancer establishes OLFM4 as a new nonmetastatic tumor marker," *Molecular & Cellular Proteomics*, vol. 10, no. 12, pp. M111. 009712, 2011.

[61] M.-Y. Huang, H.-M. Wang, H.-J. Chang, C.-P. Hsiao, J.-Y. Wang, and S.-R. Lin, "Overexpression of S100B, TM4SF4, and OLFM4 genes is correlated with liver metastasis in Taiwanese colorectal cancer patients," *DNA and cell biology*, vol. 31, no. 1, pp. 43-49, 2012.

[62] N.-Y. Chia, N. Deng, K. Das, D. Huang, L. Hu, Y. Zhu, K. H. Lim, M.-H. Lee, J. Wu, and X. X. Sam, "Regulatory crosstalk between lineage-survival oncogenes KLF5, GATA4 and GATA6 cooperatively promotes gastric cancer development," *Gut*, vol. 64, no. 5, pp. 707-719, 2015.

[63] Z. Lei, I. B. Tan, K. Das, N. Deng, H. Zouridis, S. Pattison, C. Chua, Z. Feng, Y. K. Guan, and C. H. Ooi, "Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil," *Gastroenterology*, vol. 145, no. 3, pp. 554-565, 2013.

[64] Y. Wu, H. Grabsch, T. Ivanova, I. B. Tan, J. Murray, C. H. Ooi, A. I. Wright, N. P. West, G. G. Hutchins, and J. Wu, "Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer," *Gut*, vol. 62, no. 8, pp. 1100-1111, 2013.

[65] R. Kuner, T. Muley, M. Meister, M. Ruschhaupt, A. Buness, E. C. Xu, P. Schnabel, A. Warth, A. Poustka, and H. Sültmann, "Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes," *Lung cancer*, vol. 63, no. 1, pp. 32-38, 2009.

[66] D. Macartney-Coxson, M. C. Benton, R. Blick, R. S. Stubbs, R. D. Hagan, and M. A. Langston, "Genome-Wide DNAMethylation Analysis Reveals Loci that Distinguish Different Types of

Adipose Tissue in Obese Individuals," *Clinical Epigenetics,* vol. 9, no. 48, pp. DOI 10.1186/s13148-017-0344-4, 2017.

[67] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. d. Schaetzen, R. Duque, H. Bersini, and A. Nowé, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 30, no. 9, pp. 1106-1119, 2012.

[68] O. M. Peck-Palmer, G. Clermont, G. L. Rogers, S. Yende, D. C. Angus, and M. A. Langston, "Graph Theoretical Analysis of Genome-Scale Data: Examination of Gene Activation Occurring in the Setting of Community-Acquired Pneumonia," *Shock: Injury, Inflammation, and Sepsis: Laboratory and Clinical Approaches,* vol. 50, pp. 53-59, 2018.

[69] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici, "A novel approach for data integration and disease subtyping," *Genome Research,* vol. 27, no. 12, pp. 2025-2039, 2017.

[70] S. Krishnagopal, R. v. Coelln, L. M. Shulman, and M. Girvan, "Identifying and predicting Parkinson's disease subtypes through trajectory clustering via bipartite networks," *PLoS One,* vol. 15, no. 6, pp. e0233296, 2020.

[71] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods,* vol. 11, no. 3, pp. 333, 2014.

Figures and Tables

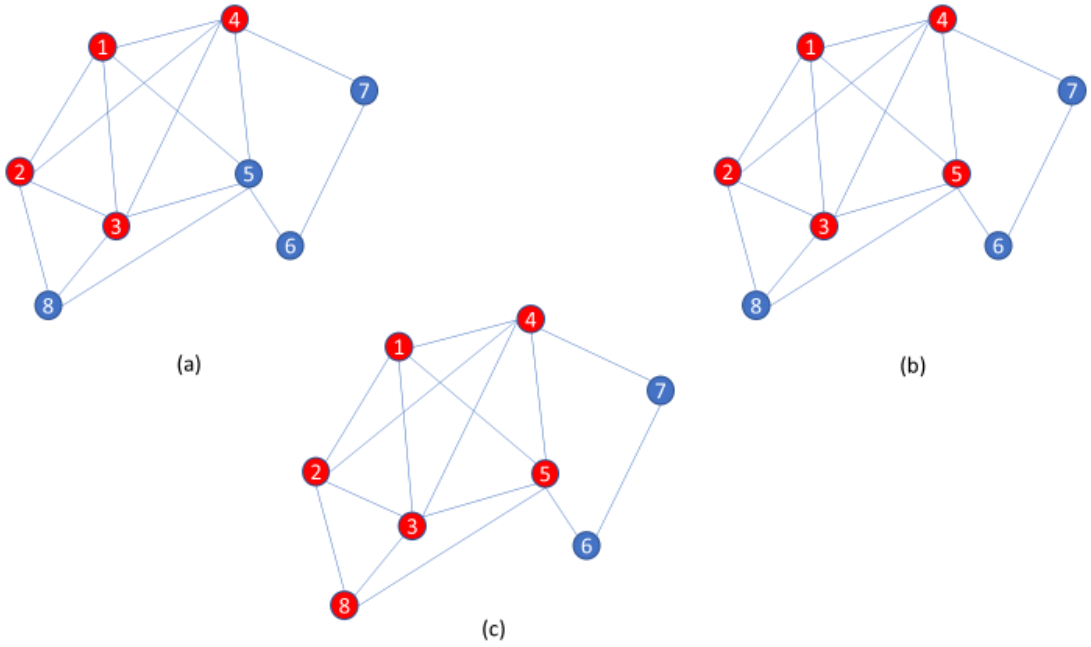

Figure 1. An illustration of the paraclique algorithm with glom term $g = 2$. (a) Starting with a maximum clique of size 4 as shown by red vertices, (b) paraclique first gloms onto vertex 5, (c) and then it gloms onto vertex 8 to form a paraclique of size 6.

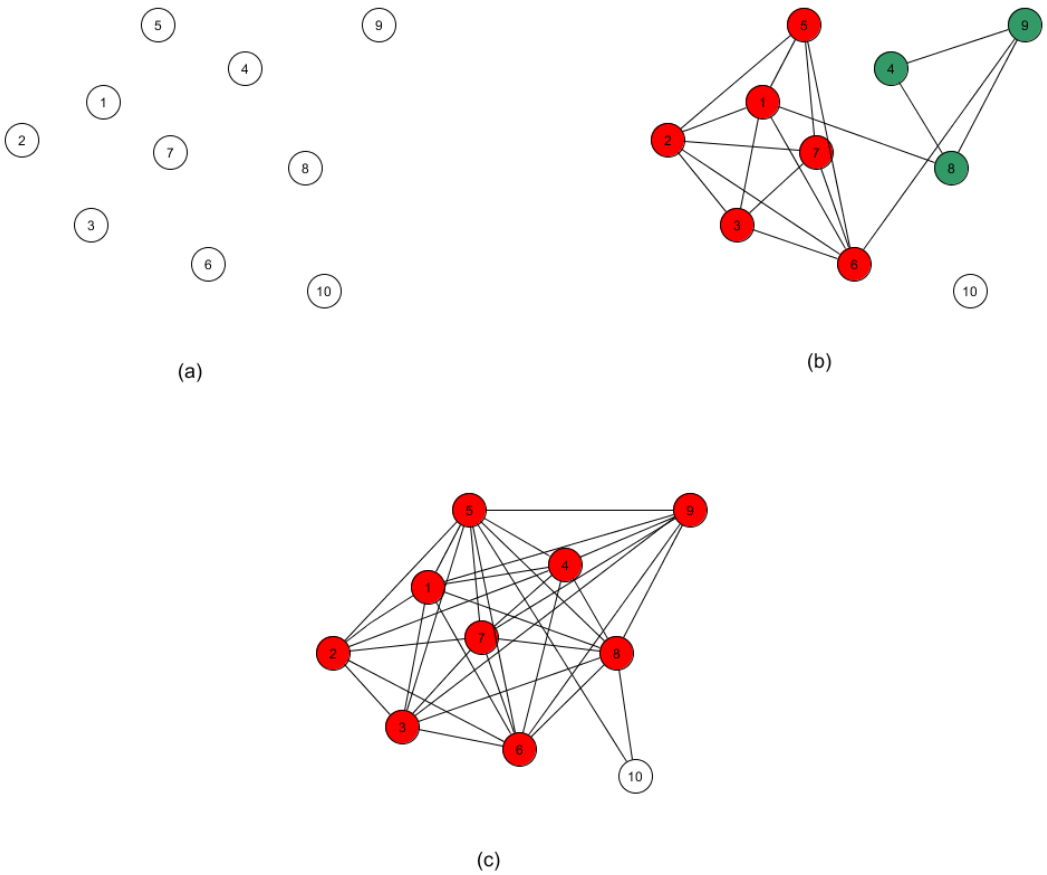

Figure 2. Outlier detection using paracliques. (a) A normalized threshold of 1.0 usually produces an empty graph. (b) As the threshold is lowered, more edges are added and paracliques begin to form and merge. (c) If a vertex consistently joins no paraclique, then it is flagged as a potential outlier.

Disease	GEO Accession	Patients		Probes	
		Case	Control	Initial	Filtered
Asthma	GSE4302	42	28	54675	2322
Breast Cancer	GSE10810	31	27	18382	11531
Chronic Lymphocytic Leukemia	GSE8835	24	12	22283	1338
Colorectal Cancer	GSE9348	70	12	54675	22968
Lung Cancer	GSE7670	27	27	22283	7458
Multiple Sclerosis	GDS3920	14	15	54674	9844
Pancreatic Cancer	GDS4102	36	16	54613	23711
Parkinson's Disease	GSE20141	10	8	54674	6625
Prostate Cancer	GSE6919	61	63	12625	1531
Psoriasis	GSE13355	58	58	54675	29407
Schizophrenia	GSE17612	28	23	54675	4250
Type 2 Diabetes	GSE20966	10	10	61294	93

Table 1. Subtyping datasets. A profile of the datasets used in this study.

Disease	Subgroups Identified	Subgroup Sizes
Asthma	3	31,8,3
Breast Cancer	2	22,5
Chronic Lymphocytic Leukemia	2	4,18
Colorectal Cancer	2	63,5
Lung Cancer	2	21,5
Multiple Sclerosis	2	11,3
Pancreatic Cancer	2	31,5
Parkinson's Disease	1	8
Prostate Cancer	2	56,3
Psoriasis	2	49,5
Schizophrenia	2	19,6
Type 2 Diabetes	1	9

Table 2. Subgroups identified. Summary of the numbers and sizes of putative subgroups identified by our methods in testing data.

Dataset	GO Category	p-value
Asthma GSE4302	Oxireductase	1.1E-4
Breast Cancer GSE10810	Secreted	1.0E-13
Chronic Lymphocytic Leukemia GSE8835	Mhc ii	2.4E-15
Colorectal Cancer GSE9348	Translational elongation	2.8E-28
Lung Cancer GSE7670	Secreted	7.7E-10
Multiple Sclerosis GDS3920	Translational elongation	1.9E-34
Pancreatic Cancer GDS4102	Signal	4.59E-15
Prostate Cancer GSE6919	Translational elongation	4.92E-46
Psoriasis GSE13355	Immune response	3.5E-15
Schizophrenia GSE17612	Organelle membrane	5.24E-4

Table 3. GO enrichments. Listed is the GO term category with the lowest enrichment p-value of the 100 most differentially expressed genes for each disease in this study.

Paraclique Results						
Gastric Cancer			**NSCLC**			
	Paraclique Sizes			**Paraclique Sizes**		
	29	16		26	12	8
Subtype			**Subtype**			
proliferative	1	12	AC	23	0	8
invasive	19	1	SCC	3	12	0
metabolic	9	3				
k-Means Results						
Gastric Cancer			**NSCLC**			
	Cluster Sizes			**Cluster Sizes**		
	26	44		28		30
Subtype			**Subtype**			
proliferative	0	29	AC	18		22
invasive	25	1	SCC	10		8
metabolic	1	14				
Hierarchical Clustering Results						
Gastric Cancer			**NSCLC**			
	Cluster Sizes			**Cluster Sizes**		
	33	37		9		49
Subtype			**Subtype**			
proliferative	0	29	AC	0		40
invasive	22	4	SCC	9		9
metabolic	11	4				

Table 4. Cluster compositions based on known subtypes. Shown is a breakdown of the subtypes obtained from datasets with best available ground truth for paraclique, k-means, and hierarchical clustering.