The meGA Computational Pipeline: A Data-Driven Approach to Discover Associations among Genomic Alterations and Identify Drug Treatment Targets across Cancers

By:
Manu Raj Chopra
Pacific Collegiate School
Santa Cruz, CA 95060
And
Biomolecular Engineering Department
University of California at Santa Cruz
Santa, Cruz, CA

Mentors:
Prof. Josh Stuart, Mr. Duncan McColl
Biomolecular Engineering Department
University of California at Santa Cruz
Santa, Cruz, CA

And
Mr. Trung Lai
Pacific Collegiate School
Santa, Cruz, CA

Abstract

In 2015, cancer killed 8.8 million people worldwide. Unlike most illnesses, cancer can manifest itself in many different ways, making it difficult to develop a panacea. Clinical studies typically detail a single cancer subtype, while following small patient cohorts - making it difficult to generalize findings across cancers. In order to help scientists develop treatments across cancers, the NIH has been funding the development of the TCGA Pan-Cancer (PANCAN) Dataset, which currently consists of thousands of tumor samples across dozens of cancers.

With the development of PANCAN, researchers now have the data necessary to find genomic associations across cancers, but prior techniques did not scale well to data of this magnitude and did not provide an intuitive visual interface, by which doctors and researchers could easily inspect potentially related samples and genomic alterations. Working towards the need for a general visualization platform for PANCAN, I helped build the UCSC TumorMap over the last three years [57].

Pre-PANCAN analyses relied on co-occurrence of genomic alterations to identify associations between genomic alterations implicated in specific cancer subtypes. While this is helpful in detailing the interactions at play in singular strains of cancer, it is does not help provide a holistic understanding of pathway disruptions across cancers. Thus, prior analyses are relatively unsuccessful in finding drug targets that can be exploited across cancers.

Herewith, I propose meGA, a novel methodology to help identify these cross cancer associations. Leveraging the massive cross-cancer PANCAN dataset and the UCSC Tumor Map Visualization Platform, meGA finds the associations between mutually exclusive genomic alterations with similar gene expression profiles and visualizes them in 2-dimensional space. These mutually exclusive genomic alterations never co-occur in the same patients but seem to play similar roles in driving tumorigenesis -- indicating that they disrupt regulatory pathways similarly and could potentially be treated similarly.

I demonstrated the efficacy of the meGA computational pipeline by applying it to the PANCAN dataset, which consists of 5074 samples across 12 different cancer types. The initial analysis results in 10+ million event associations, which the meGA pipeline filters to the ~ 10,000 most significant associations. Using the resultant map from meGA, I conducted four case studies. In these studies, I identified potentially novel drivers of AML, a potential link between pediatric brain cancers (CNS-PNET) and lymphocytic leukemias (CLL), novel associations

between genomic alterations that drive hormone signaling pathways, and novel associations between genomic alterations critical to oncogenic PI3K and Ras-MAPK pathways.

As I push to publish meGA, cancer researchers will be able to use the pipeline to identify novel associations across cancers in the hope of reusing existing subtype-specific drugs for other cancer strains, as well as identifying new cross-cancer treatment targets. Furthermore, meGA can be utilized in clinical settings to help doctors develop personalized treatments for their patients.

1 Introduction

In 2015, cancer killed 8.8 million people worldwide [1]. Unlike most illnesses, cancer can manifest in many different ways, making it difficult for there to be a panacea [2]. Typically, this disease is characterized by the onset of malignant tumorous growths, which are caused by perturbations at the genomic and epigenomic levels. These perturbations are better known as Somatic Genomic Alterations, i.e. “events”, which ultimately disrupt downstream signaling pathways, leading to unregulated cell growth and division [4-5].

1.1 Research Background

Scientists currently believe that cancer events can be primarily categorized as Single Nucleotide Variants (SNVs) and Structural Variations (SVs) [8]. SNVs usually take the form of somatic mutations, which come in three forms [9-10]. The first type is a substitution, wherein a single nucleotide of a DNA sequence is modified, thus altering the function of the gene encoded at that location. Second, there are nonsense mutations, where the DNA encoding of a gene is altered to introduce a stop codon. This acts as a kill switch, such that all the gene encodings thereafter are no longer read. Third, there are frameshift mutations, where a single nucleotide is inserted or deleted within an existing DNA sequence, altering the frame utilized by mRNA to read the nucleotide base pairs. This results in the production of malformed proteins. Ultimately, somatic mutations can lead to the development of cancer if they alter cell cycle inhibitors, allowing for unregulated cell growth [10-11]. Meanwhile, SVs are the addition or deletion of chromosomal segments [12]. Repetition is known as Copy Number Gain (Amplification), while removal is known as Copy Number Loss (Deletion). SVs play a critical role in cancer, as they can either amplify the expression of oncogenes or stop the production of tumor suppressors [13].

1.2 Research Problem

Most cancer researchers specialize in studying specific cancer subtypes, which are first characterized by the tissue of origin and further categorized by the specific cellular pathways that have been disrupted to allow for uncontrolled cell growth. Individual SNVs and SVs may not directly lead to cancer, but rather two to eight such events in combination are needed for the onset of tumorigenesis [6]. It is alteration of these regulatory mechanisms that ultimately leads to the misregulation of the so-called cancer hallmarks [49].

Studying how specific events relate to signaling pathways is essential to cancer research as it provides both a better understanding of cancer progression and may reveal potential targets for drug therapies. For example, researchers identified BRAF mutation as an activator of the MEK pathway, which is involved in most melanomas. With this insight, researchers developed a drug known as Trametinib to inhibit the MEK pathway downstream, removing the negative effects of BRAF mutation [14-15].

Analysis of associations amongst events and pathways is critical to cancer research, but it is also quite difficult. While patients diagnosed with the same cancer have similarly corrupted pathways, there can be many different underlying events that drive the same pathway alterations. Oncologists have attempted to identify these associations by finding correlated events that occur in cohorts of patients afflicted by the same cancer. While this strategy has been effective, it is intrinsically limited in scope. Due to small cohort sizes, such studies are often influenced by individual variance amongst patients, making it difficult to develop accurate hypotheses and generalize findings to other, potentially similar, types of cancer [16]. Furthermore, the subtype-oriented focus of these studies does not typically allow for cross-cancer application - where existing treatments could be considered for other, related cancers [17-18].

Bioinformatics has provided alternate methodologies for studying cancer. One such approach is to identify mutually exclusive events across samples, i.e. events that do not occur in the same patients. These events may ultimately play similar roles in their respective pathways such that same treatment may apply to both alterations. Current methodologies, focused on identification of these mutually exclusive events, only seem to successfully establish strong associations amongst such events when several are mutex and are found across a broad set of patients. Leveraging an orthogonal dataset, such as relatedness of gene expression profiles, may help one filter through the list of mutually exclusive events to identify the associations with greatest potential. Gene expression profiles quantify the relative transcription level of a gene in a tumor; they can characterize gene expression patterns of events by measuring the difference in activity of genes among samples with and without an event. By examining mutually exclusive events correlated in their differential gene expression profiles, one may be able identify hitherto unknown events which play similar roles across cancers [19].

1.3 Research Objective

There have been several previous attempts to identify associations between mutually exclusive events, but these methodologies have not proven to be effective on large datasets of thousands of samples and multiple cancer types [20-22]. Furthermore, most of these prior methodologies have not provided an intuitive visual interface, by which doctors and researchers can easily inspect potentially related events.

To address the issues with these prior techniques, I developed the Mutually Exclusive Genomic Association (meGA) Computational Pipeline, a new methodology that couples a data-driven analysis of genomic events with an interactive visualization of these associations. I applied this pipeline to the PANCAN12 dataset, verifying the validity of the results by identifying known event associations and discovering novel associations amongst mutually exclusive events, which may play similar roles in their respective pathways across cancers. In the future, this method will be augmented with a set of clinically actionable events, in hopes of providing evidence for novel markers of approved treatments.

2 Methodology

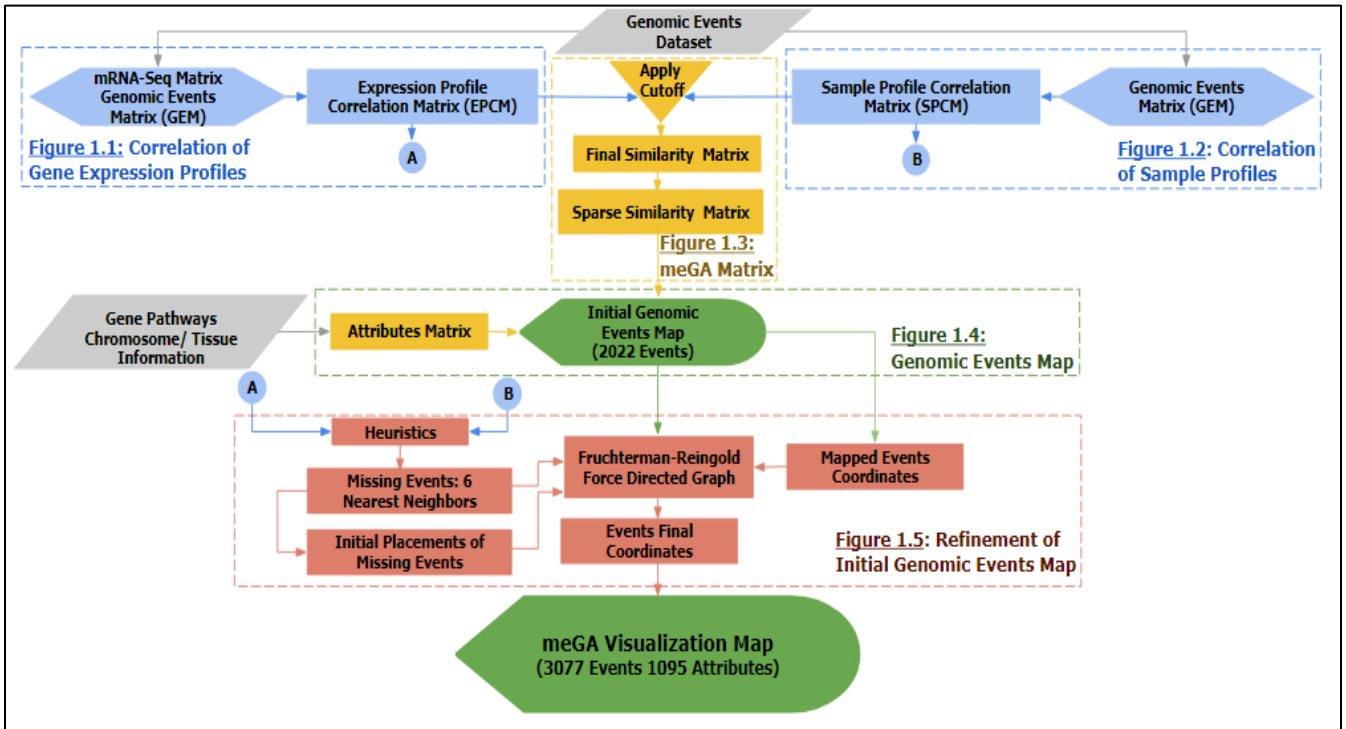


Figure 1 : Mutually Exclusive Genomic Associations (meGA) Visualization Pipeline

The Mutually Exclusive Genomic Association (meGA) Computational Pipeline consists of five stages:

1) Computation of Correlations amongst Gene Expression Profiles
2) Computation of Correlations amongst Sample Profiles
3) Computation of the meGA Matrix
4) Development of the Initial Genomic Events Map
5) Refinement of this map to produce the meGA Visualization Map (Fig 1).

This methodology builds upon the TumorMap, cancer visualization software, which I helped develop over the past two and a half years at UCSC's Bioinformatics lab. I co-authored a paper about the TumorMap, which was published in the November 2017 issue of ***Cancer Research*** [57]. It was previously utilized to cluster tumor samples according to genetic similarities. The meGA Computational Pipeline adds new capabilities to the platform to develop hypotheses regarding event associations within and across cancer types. The meGA Computational Pipeline was built using Python and JavaScript, leveraging packages such as Scipy, Numpy, scikit-learn, and networkX, along with the Meteor Pipeline and Google Maps API [23-26].

2.1 PANCAN12

I applied meGA to the PANCAN12 Dataset, but any similar dataset can be processed by the pipeline. PANCAN12 consists of 5074 patient samples across twelve tissue types (GBM, OV, BRCA, LUSC, LUAD, COAD, READ, KIRC, UCEC, BLCA, HNSC and LAML) collected by the Cancer Genome Atlas (TCGA) project in 2013 [27]. I used the Genomic Event and mRNA-Sequence (mRNA-Seq) matrices from the TumorMap. The Genomic Event Matrix (GEM) is of the form [Sample, Event], consisting of 5074 Samples by 3320 binary events (SMs, SVs), where values indicate the presence of an event in the sample. The mRNA-Seq Matrix contains gene expression values across each sample. It is of the form [Sample, Gene], consisting of 3500 samples by 12471 genes. The 3500 samples primarily consist of a subset of the 5074 from GEM.

2.2 Computing Correlation of Gene Expression Profiles

Given GEM and a respective mRNA-Seq Matrix, I utilized a two-tailed t-test to quantify how much each gene is differentially expressed for each event in the dataset. The output [Gene, Event] matrix consists of t-statistics, which indicate whether genes are upregulated or downregulated for patients with a specific event. I then took every pair of events in the output matrix and computed a pearson correlation score to measure the association between the events on a scale of [-1, 1]. Here, -1 indicates an inverse correlation, 0 indicates no correlation, and 1 indicates a positive correlation. The results are recorded in an Expression Profile Correlation Matrix (EPCM) -- an [Event, Event] matrix of similarity scores (Fig 1.1).

2.3 Computing Correlation of Sample Profiles

Many of the events in EPCM are highly correlated with one another because they occur in the same patients. While this is intuitive, it impedes our research objective of identifying transcriptionally similar events that occur across different patients and cancers. In order to quantify sample overlap, I leveraged cosine distance. For (e1, e2) in GEM, i.e. every pair of 5074-long sample vectors, I compute $\boldsymbol{d_{e1e2} = 1 - \frac{e1 \cdot e2}{\|e1\| \, \|e2\|}}$. Two events that occur in all the same patients will have a distance of 0. Meanwhile, two events that never occur in the same patients will have a distance of 1. Applying this distance metric for every pair of events produces the Sample Profile Correlation Matrix (SPCM), an [Event, Event] matrix that quantifies the sample overlap between the events in PANCAN12 (Fig 1.2).

2.4 Computing Mutually Exclusive Genomic Association (meGA) Matrix

Next, I applied cutoffs to filter the results of EPCM and SPCM before integrating them to compute a Final Similarity Matrix (FSM). I filtered the EPCM to only keep the associations with top 15% of Pearson Correlation scores and filtered the SPCM to only retain mutually exclusive event pairs (Fig 1.3). I then multiplied these matrices to derive the FSM. Thus, the visualization included only the strongest associations between mutually exclusive events. In order to visually represent the events according to their top associations, the FSM was further processed by keeping only the top 6 nearest neighbors for each event, producing the meGA Matrix.

2.5 Computing the Initial Genomic Events Map

The meGA Matrix was fed as input into the TumorMap (Fig 1.4) to produce the Genomic Events Map, providing a visualization of the associations between events across cancer subtypes

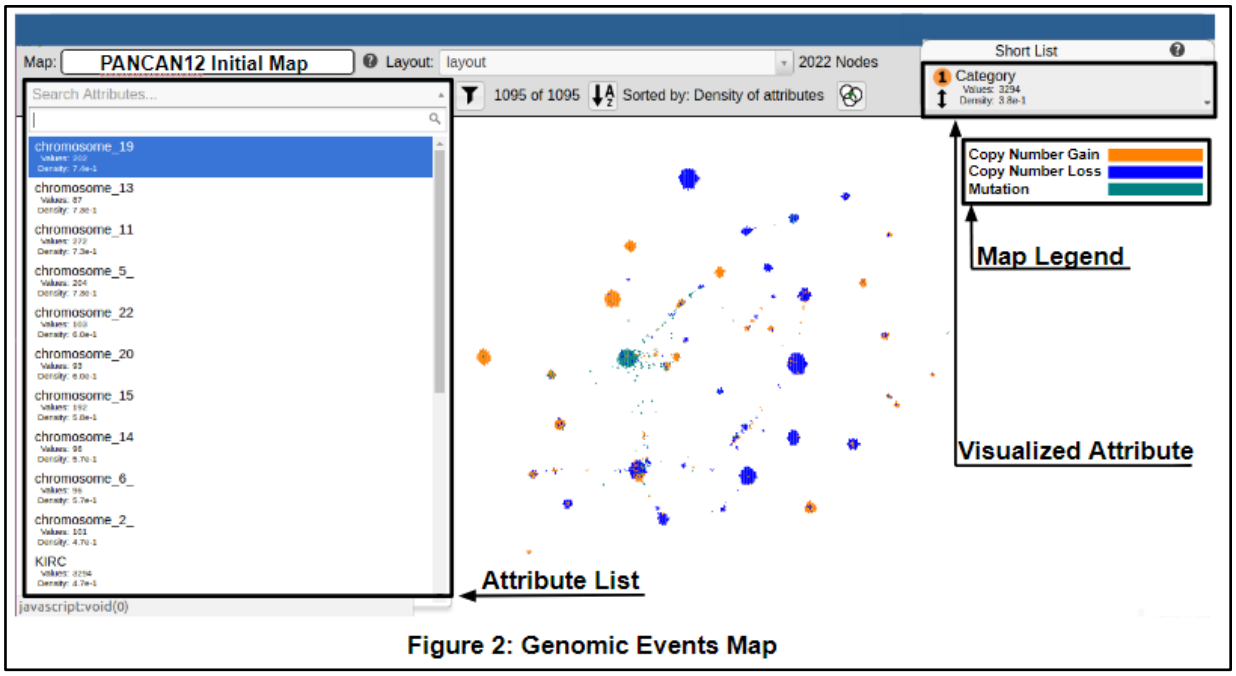


Figure 2: Genomic Events Map

(Fig 2). The platform utilizes a force directed graph algorithm known as OpenOrd to cluster events upon the scores in the meGA Matrix [27]. The algorithm outputs (x, y) coordinates for all events, which are placed as hexagons on a 2D grid for visualization [57].

The event clusters are analyzed in part by adding “attributes”, i.e. clinical data, to the map. Attributes provide the ability to visually segment the events on the map into meaningful categories with coloration (Fig 2).

They allow us to visually inspect the roles of the clusters of genomic events in pathway enrichment and pathway disruption. These attributes were taken from genomic events in pathway enrichment and pathway disruption. These attributes were taken from several public databases including the BROAD Institute’s Molecular Signature Database (msigDb) and then assigned to the events on the map [29-31]. In total, 1095 attributes were included, which can be grouped into the following categories: Gene Pathway Lists, Chromosome Number, Chromosome Position, Type of Event, Known Oncogenes, and Tissue Similarity Attributes (Fig 2). These attributes are correlations computed between the expression profiles of patients with a given event and the expression profiles of patients with cancer originating in a certain tissue.

2.6 Refinement of Initial Genomic Events Map

The meGA Matrix contains values for nearly two thirds (n=2022) of the 3320 events, selected after applying the cutoffs described above. As expected, frequently mutated genes (FMGs), which include many known oncogenes, were absent from the map because they cannot be mutually exclusive with many (if any) events. Thus, we needed a way for connecting the

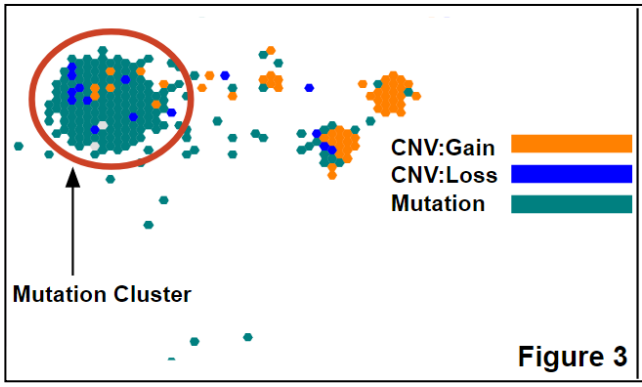


Figure 3

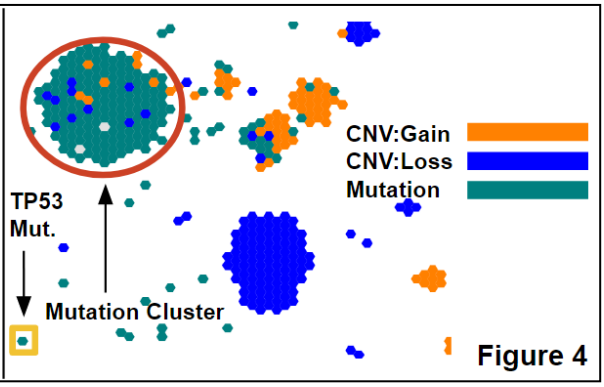


Figure 4

FMGs into the map to help identify associations between known and novel events. Therefore, I expanded the map by linking in the 1298 FMG events using a procedure described herewith.

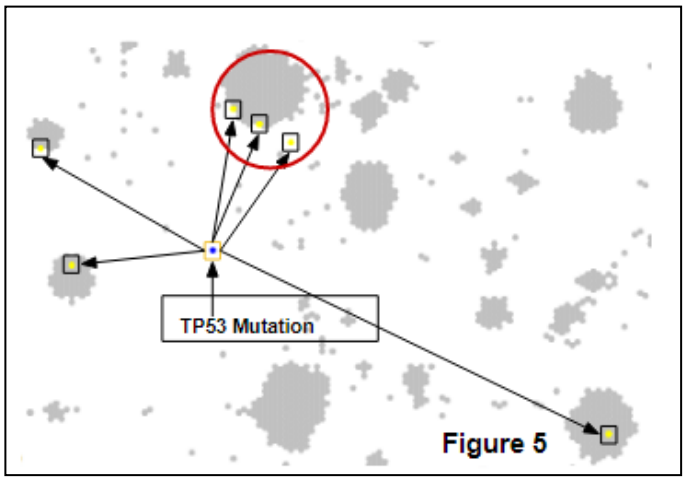


Figure 5

To integrate the FMGs with the map, I developed a heuristic to select the most appropriate event-event similarity scores with which to characterize the FMGs. First, I subsetted the EPCM and SPCM, only examining associations between missing events and those already placed on the map. For each missing event, I selected the associated events with the 100 largest cosine distances in the subsetted version of SPCM, so that I could characterize missing events by their association to those with which they had the least sample overlap (Fig 1.5). Next, I iterated over each missing event and selected the 6 largest event-event Pearson Correlation scores from the subsetted EPCM. Then, I generated initial (x, y) coordinates for the FMGs by taking the average of the (x, y) coordinates

of the 6 previously mapped events to which they are most correlated. Finally, I ran the Fruchterman-Reingold Force Directed Graph Algorithm (FRFDGA), keeping the mutually exclusive events' coordinates fixed from the initial map and using (x, y) coordinates computed in the previous step as seeds for the missing events [32]. The final output coordinates were then used to produce the meGA Visualization Map of 3077 events. 243 events from the original dataset still failed to meet the cutoff criteria, even after this refinement step.

Throughout the PANCAN12 meGA map, there was only one large mutation cluster, circled in red in Figure 3. My hypothesis was that this cluster existed because the events were similarly influenced by TP53 Mutation, a previously missing FMG. Examining the refined map, we see that TP53 Mutation landed right outside the mutation cluster (Fig 4). The placement of TP53 seems to validate my hypothesis, as it seems that the three events most correlated with TP53 are located in the aforementioned cluster (Fig 5). This is just one example where the previously missing FMG may provide context about the pathways affected by mutually exclusive events.

2.7 Mutual Exclusivity Examined

Upon inspection, meGA seems to cluster events which are not mutually exclusive with one another, but this is expected. Two events e1, e2 may not be mutually exclusive but may be correlated and mutually exclusive with the same other events. As a result, clusters on the map consist of substructures that are mutually exclusive from one another - even though the events within the substructures may not be mutually exclusive with one another. For instance, DNMT3A Mutation and NPM1 Mutation are known to be significantly associated, even occurring in the same patients, but they constitute a substructure that is mutually exclusive from substructure of deletions within the same cluster.

3 Results & Discussion

Using the meGA Computational Pipeline, I was able to calculate significant associations among mutually exclusive events in PANCAN12 and visualize these relationships. Now, I demonstrate the efficacy of the meGA Pipeline by presenting four case studies, which comprises of a preliminary investigation of the event clusters within the map. First, I verify the results of my meGA Computational Pipeline by examining a cluster, which consists of events known to be associated with AML progression. Second, I identify novel associations between events and

AML progression. Third, I identify a novel regulatory relationship between chr11 deletions and chr19 amplifications that may link pediatric brain cancers (CNS-PNET) with lymphocytic leukemia (CLL). Fourth, I perform cancer-agnostic analysis to identify novel associations between events that promote the activation of known oncogenic pathways.

3.1 Case Study 1: Verifying Known Associations in AML

In this case study, I validate the efficacy of the meGA Computational Pipeline as a methodology for visual identification of event associations. Looking at the attribute list is a good way to determine clusters that would be interesting to study. Each attribute has a density score, which provides a measure as to how spatially associated the events are on the map for a certain attribute. LAML has the highest density score amongst the tissue type attributes on the meGA Map. This implies that the events with the greatest correlation to the LAML expression profile are more tightly clustered than events most highly correlated with other tissues. Due to the high density scores, I began exploring the events correlated with LAML.

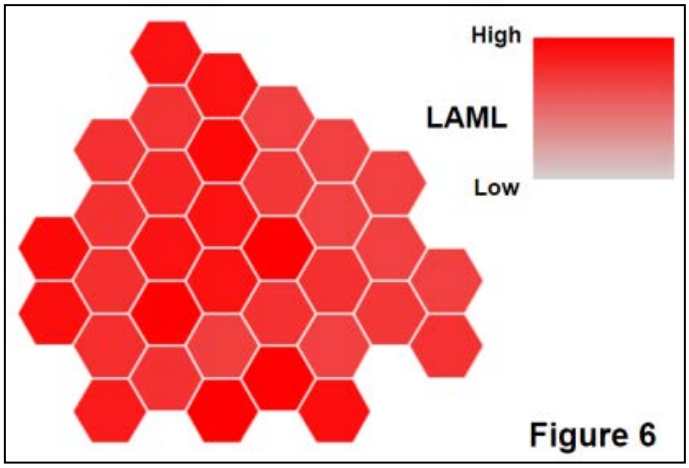


Figure 6

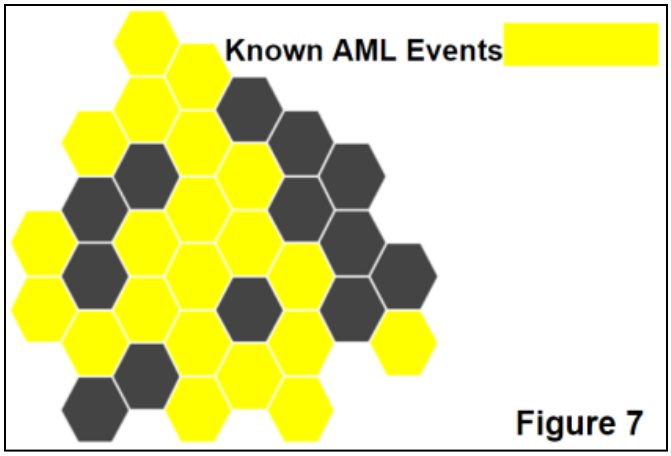


Figure 7

Selecting the LAML attribute colors the meGA Map such that brighter shades of red indicate stronger correlation with LAML and lighter shades of red indicate weaker correlation with LAML. Examining the map, I further investigated the brightest LAML cluster (Fig 6). An investigation of genomic alterations in this cluster finds that 20 of the 33 events grouped together here, have been shown to be highly associated with the development of AML in previous literature [33-36] (Fig 7). The 20 events include the most frequently recurring mutations implicated in the development of AML, i.e.

DNMT3A Mutation, NPM1 Mutation, IDH1 Mutation, CEBPA Mutation, IDH2 Mutation, FLT3 Mutation, NRAS Mutation, TET2 Mutation, and PTPN11 Mutation [48].

3.1.1 Mutual Exclusivity in AML

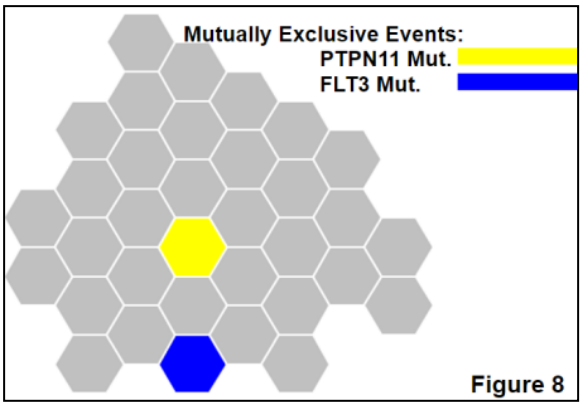


Figure 8

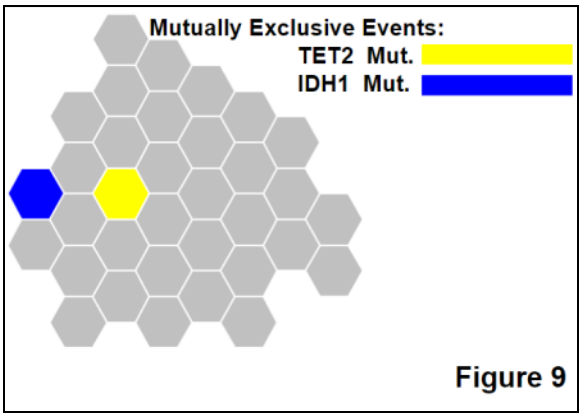


Figure 9

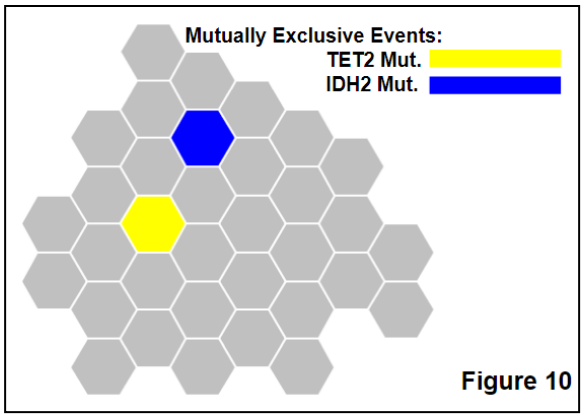


Figure 10

With the meGA Map, we are able to find associations in this cluster between frequently recurring mutations in AML that are known to be mutually exclusive with one another, i.e. might play similar roles in their respective pathways. Examining this cluster, I found PTPN11 Mutation and FLT3 Mutation, two events that are mutually exclusive (Fig 8). Additionally, I found IDH1 Mutation, IDH2 Mutation, and TET2 Mutation in this cluster (Fig 9-10), where the IDH mutations were mutually exclusive with TET2 mutation.

One of the primary theories about AML is that it requires a Class I and Class II mutation to develop. Class I mutations activate signal transduction pathways for cell proliferation. Class II mutations affect transcription factors, impairing pathways associated with hematopoietic differentiation. According to a clinical study done on patients with AML, both PTPN11 Mutation and FLT3 Mutation belong to Class I. [37-38]. Similarly, IDH1, IDH2, and TET2 belong to Class I. Thus, these should be mutually exclusive and produce similar expression states, as hypothesized by the clustering on the map.

3.2 Case Study 2: Discovering Novel Event Associations in AML

After verification of the validity of the LAML cluster, I performed further investigation of this cluster to identify potentially novel associations among these events. My next step was to examine the distribution amongst the types of genomic alterations in this cluster (Fig 11). Here, frequently recurring Mutations (Green) are associated with a group of Copy Number Losses (Blue), aka deletions. Some of the highlighted deletions have been implicated in AML, e.g. GPX3 Del. and CDKN1B Del. [39-40] (Fig 11).

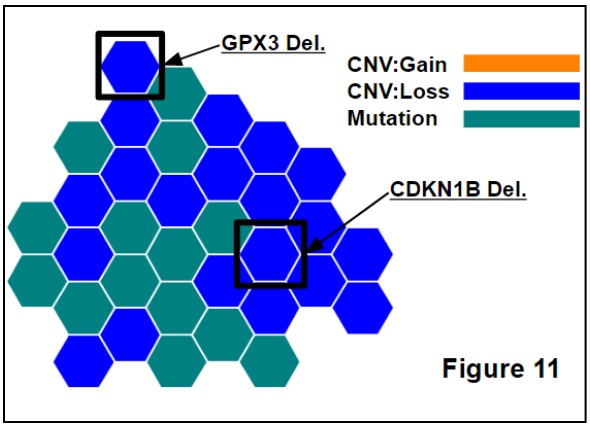


Figure 11

Examining the events in this cluster (Fig 11), I have identified 13 deletion events not previously implicated in AML progression in existing cancer literature (Fig 12). For the sake of brevity, I present further analysis of the 2 most interesting deletions, as potential drivers for the onset of oncogenesis and the development of AML. Further analysis must be performed on the 11 others to gain similarly insightful analyses. The two events I selected to further investigate were GNAQ Deletion and chr12-12262378-12267068 Deletion (Fig 12).

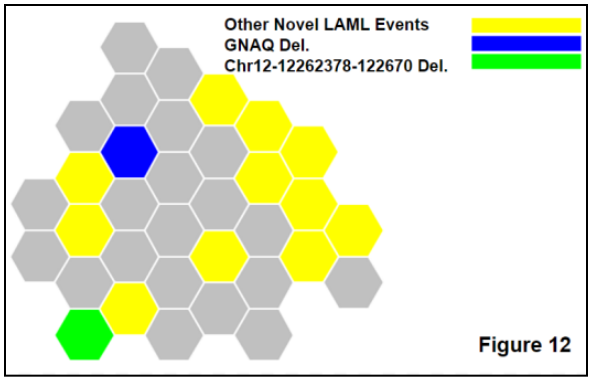


Figure 12

GNAQ Deletion (Blue) is highly correlated with the LAML attribute, yielding a Pearson Correlation score of r = 0.72. This indicates a strong similarity in expression phenotypes amongst patients with GNAQ Deletion and those with AML. Table 1 lists the top 6 events associated with the GNAQ Deletion from the meGA Matrix. GNAQ Deletion is highly correlated with the expression signatures of many of the most frequently occurring events in AML (Table 1).

Most Similar Events	Pearson Correlation Score
RUNX1 MUTATION	0.739242
FLT3 MUTATION	0.728182
CEBPA MUTATION	0.715067
NPM1 MUTATION	0.689799
DNMT3A MUTATION	0.673952
MIR142 MUTATION	0.653096

Table 1: GNAQ Del. Event Assoc.

GNAQ is a gene that produces a

nucleotide binding protein, which attaches to a receptor to activate a molecule known as GTP [41]. In turn, this molecule activates signaling pathways that help a cell regulate the development and function of blood vessels. Thus, GNAQ deletion will result in less production of GTP and lead to unregulated cell development amongst blood vessels. Furthermore, it is mutually exclusive with the known events listed in Table 1, indicating that it induces a similar expression state to commonly recurring mutations in AML, while never occurring in the same patients.

Most Similar Events	Pearson Correlation Score
FLT3 MUTATION	0.795949
DNMT3A MUTATION	0.794400
IDH2 MUTATION	0.793131
CEBPA MUTATION	0.788816
WT1 MUTATION	0.766585
NPM1 MUTATION	0.764020

Table 2: chr12-12262378-12267068 Del. Event Assoc.

Thus, I have identified a novel association between GNAQ Deletion and AML progression, as it seems to play a similar role to known AML oncogenes in pathways usually disrupted during AML tumorigenesis.

Chr12-12262378-12267068 Deletion (Green) is the other event in this cluster, whose relationship with AML I examine herewith (Fig 12). This event is an Open Reading Frame meaning that it is region that has the potential to be translated but the region has otherwise not been characterized by any genetic or biochemical study. It has a 0.78 similarity score to LAML, which is approximately the same as the correlation scores between known AML drivers and the LAML attribute. Currently, researchers do not know what this chromosomal region encodes, but here meGA identifies its deletion as a novel driver of AML, due to its high correlation with other known drivers of AML (Table 2). Further research must be performed on Chr12-12262378-12267068 to determine its role in genetic pathways.

3.3 Case Study 3: Discovery of a Novel Association between Chromosome 19 Amplifications & Chromosome 11 Deletions

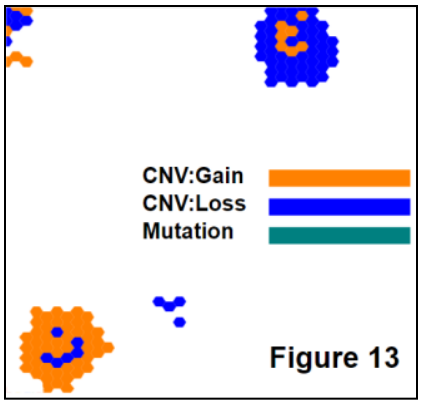
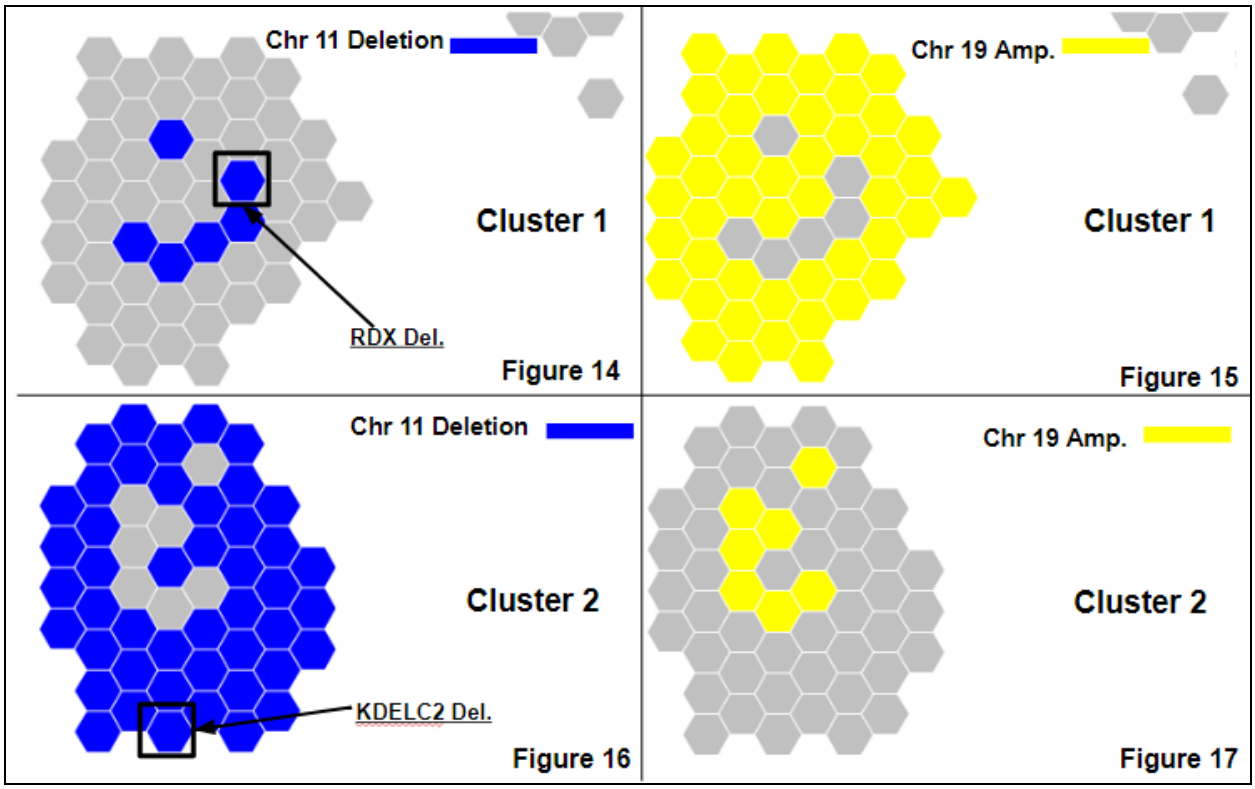


Figure 13

Using meGA I identified a potentially novel regulatory relationship between events in pathways involved in pediatric brain cancers (CNS-PNET) and lymphocytic leukemias (CLL). Upon inspection of the distribution in genomic event types, I found two clusters that seem to be mirror images of one another

(Fig 13). Cluster 1, bottom left corner of Fig 13, consists of CNV Gains (Orange) with a few CNV Losses (Blue). Cluster 2, top right corner of Fig 13, consists of CNV Losses (Blue) and a



few CNV Gains (Orange).

Having examined several possible attributes, I propose a novel association between Chromosome 19 and 11, which may be driving the formation of these event clusters. The Copy Number Gains that we observed in the clusters (Fig 13) all are Chr19 amplifications (Fig 15, 17) while the Copy Number Losses we observed in the clusters all are Chr11 deletions (Fig 14, 16). All the Chr19 amplifications occur on q13 and all the Chr11 deletions occur on q22 and q23. The fact that all the genes altered by these events belong to such a small subset of loci seems to lend credibility to the hypothesis of an association between these events. Furthermore, the results of the meGA pipeline seem to suggest that these two groups of chromosome aberrations are mutually exclusive of each other; and furthermore amongst all events in the dataset Chr19 amplifications are most correlated with the Chr11 deletions, and vice-versa - with Pearson Correlation scores r = 0.62. Altogether, this seems to indicate that Chr19 amplifications and Chr11 deletions induce similar expression states, i.e. these alterations do not co-occur but might invoke the same pathways in cancer development. Chr19q13 is home to one of the largest

miRNA clusters, known as C19MC. miRNA is a type of RNA molecule involved in post-transcriptional gene regulation. Once RNA has been transcribed from DNA, the next step in the process is for it to be translated into a protein. miRNA regulates other RNAs either through translational repression or through mRNA degradation by binding to RNA strands. miRNA regulation plays an important role in the expression of many genes. The amplification of genes involved in the production of miRNA may significantly alter transcriptional pathways, by eliminating the translation of certain genes.

I postulate that the amplification of genes on Chromosome 19q13, associated with a large microRNA cluster known as C19MC, induces a regulatory effect upon the genes on Chromosome 11q22 and 11q23. These amplifications may lead to a stoppage in the production of necessary proteins, similar to a stoppage that would be induced by the deletion of genes on Chromosome 11q22 and 11q23.

Previous cancer literature has implicated amplifications of C19MC, the miRNA cluster associated with Chromosome 19q13, in the development of neuro-ectodermal brain tumors, e.g. pediatric brain cancers (CNS-PNET)[42-44]. Similarly, Chromosome 11q22 Deletions have been previously implicated in the progression of lymphocytic leukemia (CLL) [45-47]. 10-20% of CLL patients have Chromosome 11q22 and 11q23 Deletions. Two of the genes known to be deleted in most CLL patients with Chromosome 11q22 and 11q23 are present in these clusters: RDX Deletion and KDELC2 Deletion (Fig 14, Fig 16). Here, I link these groups of events together, ultimately hypothesizing that these two types of cancers may invoke the same genetic pathways. There is little prior research as to the signaling pathways with which these chromosomal regions interact, so further research must be performed by biologists to examine the downstream effects of these SVs. Also, oncologists studying these cancer subtypes should work together, as drug treatments for one might be effective for the other.

3.4 Case Study 4: Discovery of Associations in Hormone Signaling Pathways and PI3K/Ras-MAPK Pathways

Lastly, I leveraged meGA to perform cancer-agnostic analysis, identifying relationships between events that may drive pathways across a spectrum of cancers. I examined the cluster shown in Figure 18, which consists mostly of amplifications and deletions, along with two mutations. Ultimately, I found that there are actually two clusters grouped together.

Examining the scores from the meGA matrix, I found that the two mutations are mutually exclusive and correlated with one another (Cluster 1). The amplifications and deletions constitute Cluster 2 as they are not correlated with the mutations, while being mutually exclusive and correlated to one another. The two clusters seem to clump together because the deletions in Cluster 2 and the mutations in Cluster 1 are mutually exclusive and correlated to Chromosome 17 Amplification, highlighted in gray (Fig 19).

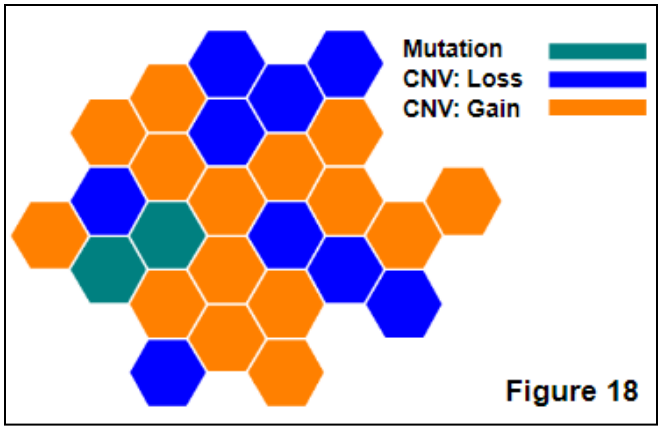


Figure 18

Cluster 1 consists of two mutations: FOXA1 mutation and NR1H2 mutation (Fig 19). FOXA1 mutation is a well-established cancer event, known to drive hormone-based cancers, e.g. breast and prostate cancers. FOXA1 is a transcription factor (TF) that can bind directly to chromatin, allowing other TFs to find their target genes. FOXA1 mutation influences well-known nuclear hormone receptors associated with cancer development, e.g. estrogen receptors (ER) in breast cancer and androgen receptors (AR) in prostate cancer, by facilitating the binding of these nuclear hormone receptors to induce the expression of cell proliferation-related genes [50]. The meGA map shows that FOXA1 mutation is mutually exclusive and correlated with NR1H2 mutation with r = 0.59. NR1H2 is part of a family of nuclear hormone receptors that are involved in many metabolic processes [51]. Biologically, it is feasible that NR1H2 could be involved in hormone-stimulated binding of TFs, as hypothesized by the map clustering. This association has not been reported previously and NR1H2 mutation has rarely been implicated in cancer. Since FOXA1 and NR1H2 mutation induce similar expression states, they may be involved in similar signaling pathways and thus be treated similarly.

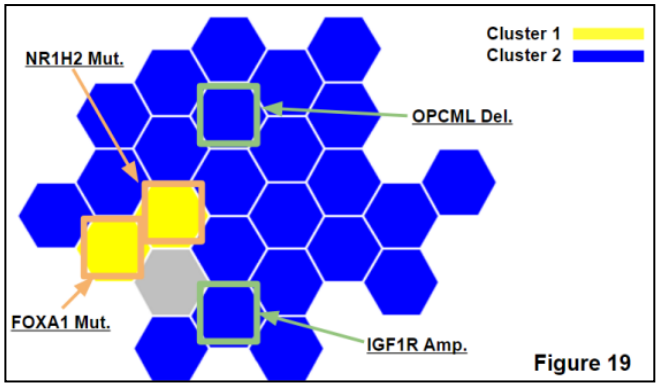


Figure 19

Cluster 2 consists of a group of deletions and amplifications (Fig 19). The group of amplifications and deletions are mutually exclusive with one another and have pearson correlation scores of r = 0.635 with one another. One of these amplifications is known to be an oncogenic event: IGF1R Amplification (Fig 19). IGF1R is a tyrosine kinase receptor, which mediates the actions of insulin-like growth factors. The phosphorylation of insulin receptor substrates (IRSs) leads to the activation of two well-known oncogenic signaling pathways: PI3K-AKT/PKB and Ras-MAPK. The activation of the PI3K pathway leads to the inhibition of apoptosis while the activation of the Ras-MAPK pathway leads to the cell proliferation. Together, these allow for unregulated cell growth and reduction in cell death, leading to tumorigenesis. The amplification of IGF1R has been shown to be a key driver in the development of malignant growths across cancers [52-53]. One of the deletions in this cluster is known to be an oncogenic event: OPCML Deletion (Fig 19). OPCML encodes an opioid binding and cell adhesion protein. More importantly, it has been found to be a tumor suppressor that negatively regulates tyrosine kinase receptors and is either silenced or deleted in many cancers, e.g. epithelial ovarian cancer [54-56]. The deletion of OPCML should lead to unregulated production of tyrosine kinases, which may induce a similar expression state to IGF1R Amplification, which also leads to the overproduction of these kinases. The biological narrative, pearson correlation scores, and mutual exclusivity of these events seem to indicate that IGF1R and OPCML may have similar roles in the PI3K and Ras-MAPK pathways. This is a new connection between two major drivers of cancer. It has not previously been reported and may indicate that similar drug treatments could be used for patients having these alterations, across a large number of cancers.

4 Conclusion

My research focused on uncovering associations between genomic alterations across cancers so that scientists can develop cross-cancer drug treatments and repurpose existing drugs for previously untreatable cancers.

I developed the meGA Computational Pipeline to achieve this goal. This Pipeline provides a data-driven approach to identify mutually exclusive events with correlated gene expression profiles, visually investigate clusters of these events in the context of pathways, and propose novel relationships between events and pathways. These mutually exclusive event associations are points of interest for biologists developing cross-cancer drug therapies.

I demonstrated the efficacy of the meGA computational pipeline by applying it to the PANCAN12 dataset, which consists of 5074 samples across 12 different cancer types. The initial analysis results in 10+ million event associations, which the meGA pipeline filters to the ~ 10,000 most significant associations. Using the resultant map from meGA, I conducted four case studies. In these studies, I identified potentially novel drivers of AML, a potential link between pediatric brain cancers (CNS-PNET) and lymphocytic leukemias (CLL), novel associations between events that drive hormone signaling pathways, and novel associations between events critical to oncogenic PI3K and Ras-MAPK pathways.

These case studies constitute a preliminary investigation of the meGA map, which researchers can further study to identify many more associations that have not previously been found.

The meGA Computational Pipeline can also be used by oncologists in clinical settings. Doctors can upload their patients' data to the meGA map so that they can gain insight as to the pathways that have been disrupted in similar patients and leverage this information to craft personalized treatments.

Altogether, the meGA Computational Pipeline provides researchers a novel perspective with which to analyze large cross-cancer datasets. By mapping cancer events in a space that removes implicit-sample biases and provide pathway context, meGA will help researchers identify novel associations amongst events and pathways across cancers, allowing for the development of new hypotheses to drive the development of new drug treatments.

5 Future Work

I plan on refining the meGA Computational Pipeline by integrating the results of my lab's drug sensitivity models to directly identify associations between potential treatments and these cancer events. Additionally, I plan on applying the meGA Computational Pipeline to PANCAN33, a much larger dataset of 33 cancer subtypes, whose analysis should yield an even greater number of interesting associations.

Appendix

1. Mathematical Underpinnings of the meGA Computational Pipeline

1.1. Pairwise Distance Matrix with Pearson Correlation Metric

The Pearson Correlation coefficient is a measure of linear correlation between two variables, in this case X and Y. The Linear Correlation coefficient is calculated by dividing the covariance of the two variables by the product of the standard deviations of the variables. Covariance is defined by the probability of the joint variability of two variables, meaning that two variables that track with one another have a positive covariance.

$$\boldsymbol{\rho} x, y = \frac{\mathbf{E}[\mathbf{XY}] - \mathbf{E}[\mathrm{X}]\mathbf{E}[\mathrm{Y}]}{\sqrt{\mathbf{E}[\boldsymbol{X^2}] - [\mathbf{E}[\mathbf{X}]]^2}\sqrt{\mathbf{E}[\mathbf{Y^2}] - [\mathbf{E}[\mathbf{Y}]]^2}}$$

$$\mu_x = \mathbf{E}[\mathbf{X}]\,,\ \mu_y = \mathbf{E}[\mathbf{Y}],\ \ \sigma^2 x = \mathbf{E}[\boldsymbol{X}^2] - [\mathbf{E}[\mathbf{X}]]^{\mathbf{2}},\ \ \sigma^2 y = \mathbf{E}[\boldsymbol{Y}^2] - [\mathbf{E}[\mathbf{Y}]]^{\mathbf{2}}$$

Here μ_x represents the mean of x, μ_y represents the mean of y, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

A pairwise distance matrix is calculated with Pearson Correlation as the metric for which the matrix is to calculate distance. This calculation takes two one dimensional vectors as input, and creates as a distance matrix with pearson correlation as the metric. It does this for each combination of one dimensional vector in the matrix.

$$\boldsymbol{\rho}(\mathbf{u}, \mathbf{v}) = \ \mathbf{1} - \frac{(\boldsymbol{u} - \overline{\boldsymbol{u}}) \cdot (\boldsymbol{v} - \overline{\boldsymbol{v}})}{\left|\left|\boldsymbol{u} - \overline{\boldsymbol{u}}\right|\right|_{\mathbf{2}} \left|\left|(\boldsymbol{v} - \overline{\boldsymbol{v}})\right|\right|_{\mathbf{2}}}$$

Where u and v are the two one dimensional vectors.

1.2. Pairwise Distance Matrix with Cosine Distance Metric

Cosine Distance is the measurement of the distance between two vectors based off of a measurement of the cosine of the angle between them. For two column vectors $\boldsymbol{x}, \boldsymbol{y}$ in our matrix, we compute the L-2 normalized dot product, where $\boldsymbol{k}$ is the cosine distance.

$$\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}\boldsymbol{y}^{\boldsymbol{T}}}{\left|\left|\boldsymbol{x}\right|\right| \left|\left|\boldsymbol{y}\right|\right|}$$

1.3. Fruchterman-Reingold Force Directed Graph Algorithm

Force Directed Algorithms are incredibly useful in calculating the layouts, i.e. the positioning, for undirected graphs. They build graphs based solely on the information within the graph structure. Force Directed Graph Algorithms work place vertices on the graph structure by assigning forces to the edges which are to be placed.

The Fruchterman-Reingold Force Directed Graph Algorithm defines two forces to act on the vertices: an attractive force and a repulsive force.

$$F_{\text{attractive}}(d) = d^2/k$$

$$F_{\text{repulsive}}(u, v) = -k^2/d$$

The forces applied to each vertex in the graph are guided by these criteria. First, each edge in the graph is treated as if it were in a mechanical system where spring forces, guided by Hooke's Law, move all the vertices to a minimal energy state. The springs have logarithmic strength to make sure forces on the edges do not get too strong. Secondly, vertices in the graph are treated as if they were celestial bodies, meaning that the vertices in the graph exert attractive and repulsive forces on one another. *K* in the formulas for force is the optimal distances between vertices on the graph where

$$k = C\sqrt{\frac{area}{number\ of\ vertices}}$$

and C is a certain weight. Finally, the Fruchterman-Reingold Graph Algorithm utilizes the notion of temperature which controls the average amount of displacement of the vertices on the graph. As more and more vertices get placed the temperature decreases, allowing for smaller and smaller adjustments of the positioning of the vertices on the graph.

2. TumorMap

The TumorMap is cancer visualization software developed at UCSC's Biomolecular Engineering Lab, which I have worked on for the last 2 years. The TumorMap provides a visualization of tumor samples on a two dimensional grid and is used for studying sample relationships.

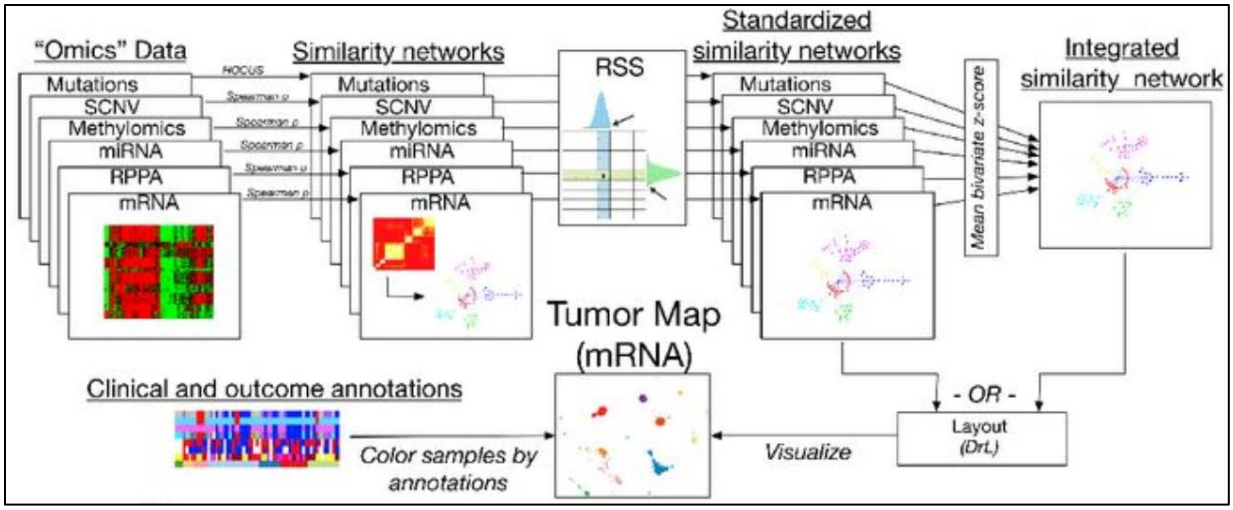


Figure 20: An illustration of TumorMap construction workflow as applied on TCGA Pan-Cancer-12 dataset [57].

My project utilizes and refines the software to visualize the associations between genomic alterations, which comprise of the results from the meGA Computational Pipeline. Figure 20 documents the TumorMap construction workflow.

Bibliography

[1] Cancer. World Health Organization Available at: http://www.who.int/mediacentre/factsheets/fs297/en/. (Accessed: 2nd September 2017)

[2] Blanpain, C. C. A. Tracing the cellular origin of cancer. Nature Cell Biology 15,126–134 (2013).

[3] What Is Cancer? *National Cancer Institute* Available at: https://www.cancer.gov/about-cancer/understanding/what-is-cancer. (Accessed: 2nd September 2017)

[4] Rodríguez-Paredes, M. & Esteller, M. The Fundamental Role of Epigenetic Regulation in Normal and Disturbed Cell Growth, Differentiation, and Stemness. *Epigenetic Therapy of Cancer* 1–41 (2013). doi:10.1007/978-3-642-38404-2_1

[5] Paull, E. O. *et al.* Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 29,2757–2764 (2013).

[6] Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* 339,1546–1558 (2013).

[7] Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. Cell 144,646–674 (2011).

[8] Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Reviews Genetics* 10,551–564 (2009).

[9] Chakravarthi, B. V., Nepal, S. & Varambally, S. Genomic and Epigenomic Alterations in Cancer. *The American Journal of Pathology* 186,1724–1735 (2016).

[10] Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. Science 349, 1483–1489 (2015).

[11] Olivier, M., Hollstein, M. & Hainaut, P. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology* 2,(2009).

[12] Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genetics* 45,1134–1140 (2013).

[13] Shlien, A. & Malkin, D. Copy number variations and cancer. *Genome Medicine* 1,62 (2009).

[14] MEK Inhibition in BRAF-Mutated Melanoma. New England Journal of Medicine 367,1364–1365 (2012).

[15] Solit, D. B. *et al.* BRAF mutation predicts sensitivity to MEK inhibition. Nature 439,358–362 (2005).

[16] Nielsen, T. O. Immunohistochemical and Clinical Characterization of the Basal-Like Subtype of Invasive Breast Carcinoma. *Clinical Cancer Research* 10, 5367–5374 (2004).

[17] Kim, Y.-A. & Przytycka, T. M. Bridging the Gap between Genotype and Phenotype via Network Approaches. *Frontiers in Genetics* 3,(2013).

[18] Beroukhim, R. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *PNAS* 104,20007–20012 (2007).

[19] Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Research* 22,375–385 (2011).

[20] Ciriello, G., Cerami, E., Sander, C. and Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Research 22(2), 398-406 (2011).

[21] Lu, S. *et al.* Identifying driver genomic alterations in cancers by searching minimum-weight, mutually exclusive sets. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*(2015). doi:10.1109/bibm.2015.7359942

[22] Wu, H. *et al.* Identifying overlapping mutated driver pathways by constructing gene networks in cancer. *BMC Bioinformatics* 16,(2015).

[23] Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011)

[25] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA),11–15 (2008)

[26] Surhone, L. M., Tennoe, M. T. & Henssonow, S. F. *Node.js*. (Betascript Publishing, 2010).

[27] Weinstein, J. N. *et al*. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature Genetics* 45, 1113–1120 (2013)

[28] Shawn Martin. OpenOrd, Sandia National Laboratories (2012)

[29] Subramanian, A. *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102,15545–15550 (2005).

[30] Peruzzi, L. & Bedini, G. Online resources for chromosome number databases. *Caryologia* 67,292–295 (2014).

[31] Cancer Gene List. *Bushman Lab: Genelists* University of Pennsylvania Available at: http://www.bushmanlab.org/links/genelists.

[32] Fruchterman, T. M. J. & Reingold, E. M. Graph drawing by force-directed placement. *Software: Practice and Experience* 21,1129–1164 (1991).

[33] Tim, L. J., *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456,(2008).

[34] Takahashi, S. Current findings for recurring mutations in acute myeloid leukemia. *Journal of Hematology & Oncology* 4,36 (2011).

[35] Kwok, C.-T., Marshall, A. D., Rasko, J. E. J. & Wong, J. J. L. Genetic alterations of m6A regulators predict poorer survival in acute myeloid leukemia. *Journal of Hematology & Oncology* 10,(2017).

[36] Dombret, H. Gene mutation and AML pathogenesis. *Blood* 118,5366–5367 (2011).

[37] Hou, H.-A. *et al.* Characterization of acute myeloid leukemia with PTPN11 mutation: the mutation is closely associated with NPM1 mutation but inversely related to FLT3/ITD. *Leukemia* 22,1075–1078 (2007).

[38] Inoue, S., Lemonnier, F. & Mak, T. W. Roles of IDH1/2 and TET2 mutations in myeloid disorders. *International Journal of Hematology* 103,627–633 (2016).

[39] Zhou, J.-D. *et al.* Down-regulation of GPX3 is associated with favorable/intermediate karyotypes in de novo acute myeloid leukemia. *International Journal of Clinical and Experimental Pathology* 8, 2384–2391 (2015).

[40] Haferlach, C. *et al.*CDKN1B, encoding the cyclin-dependent kinase inhibitor 1B (p27), is located in the minimally deleted region of 12p abnormalities in myeloid malignancies and its low expression is a favorable prognostic marker in acute myeloid leukemia. *Haematologica* 96, 829–836 (2011).

[41] GNAQ gene - Genetics Home Reference. *U.S. National Library of Medicine* Available at: https://ghr.nlm.nih.gov/gene/GNAQ. (Accessed: 10th September 2017)

[42] Nguyen, P. N. N., Huang, C.-J., Sugii, S., Cheong, S. K. & Choo, K. B. Selective activation of miRNAs of the primate-specific chromosome 19 miRNA cluster (C19MC) in cancer and stem cells and possible contribution to regulation of apoptosis. *Journal of Biomedical Science* 24,(2017).

[43] Li, M. *et al.* Frequent Amplification of a chr19q13.41 MicroRNA Polycistron in Aggressive Primitive Neuroectodermal Brain Tumors. *Cancer Cell* 17,413 (2010).

[44] Spence, T. *et al.* CNS-PNETs with C19MC amplification and/or LIN28 expression comprise a distinct histogenetic diagnostic and therapeutic entity. *Acta Neuropathologica* 128**,**291–303 (2014).

[45] Bullerdiek, J., Kiefer & Tiemann. Chronic lymphocytic leukemia-associated chromosomal abnormalities and miRNA deregulation. *The Application of Clinical Genetics* 21 (2012). doi:10.2147/tacg.s18669

[46] Jiang, Y. *et al.* ATM function and its relationship with ATM gene mutations in chronic lymphocytic leukemia with the recurrent deletion (11q22.3-23.2). *Blood Cancer Journal* 6,(2016).

[47] Puiggros, A., Blanco, G. & Espinet, B. Genetic Abnormalities in Chronic Lymphocytic Leukemia: Where We Are and Where We Go. *BioMed Research International* 2014**,**1–13 (2014).

[48] TCGA. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. New England Journal of Medicine 368, 2059–2074 (2013).

[49] Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* 144**,**646–674

[50] Robinson, J. L. L., Holmes, K. A. & Carroll, J. S. FOXA1 mutations in hormone-dependent cancers. *Frontiers in Oncology* **3,**(2013).

[51] Nr1h2 nuclear receptor subfamily 1, group H, member 2 - Gene - NCBI. National Center for Biotechnology Information

[52] Cao, Y. *et al.* Insulin-Like Growth Factor 1 Receptor and Response to Anti-IGF1R Antibody Therapy in Osteosarcoma. *PLOS One*(2014).

[53] Ribeiro, T. C. *et al.* Amplification of the Insulin-Like Growth Factor 1 Receptor Gene Is a Rare Event in Adrenocortical Adenocarcinomas: Searching for Potential Mechanisms of Overexpression. *BioMed Research International*(2014).

[54] McKie, A. B., *et al.* The OPCML Tumor Suppressor Functions as a Cell Surface Repressor-Adaptor, Negatively Regulating Receptor Tyrosine Kinases in Epithelial Ovarian Cancer. Cancer Discovery 2,156–171 (2012).

[55] Sellar, G. C. *et al.* OPCML at 11q25 is epigenetically inactivated and has tumor-suppressor function in epithelial ovarian cancer. *Nature Genetics* 34**,** 337–343 (2003).

[56] Cui, Y. *et al.* OPCML Is a Broad Tumor Suppressor for Multiple Carcinomas and Lymphomas with Frequently Epigenetic Inactivation. *PLOS One***3,**(2008).

[57] Newton, Y., Novak, A., Swatloski, T., McColl, D., Chopra, S., Graim, K., Weinstein, A., Baertsch, R., Salama, S., Ellrott, K., Chopra, M., Goldstein, T., Haussler, D., Morozova, O., Stuart, J., TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. Cancer Research 77,(2017).