UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DISCOVERING MOLECULAR PATTERNS WITH THERAPEUTIC IMPLICATIONS IN LARGE-COHORT HETEROGENEOUS CROSS-CANCER DATA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Yulia Newton

December 2016

The Dissertation of Yulia Newton
is approved:

Professor Joshua Stuart, Ph.D., Chair

Professor David Haussler, Ph.D.

Sofie Salama, Ph.D.

Olena Morozova, Ph.D.

Dean Arthur Ramirez
Vice Provost and Dean of Graduate Studies Tyrus Miller

Table of Contents

List of Figures viii

List of Tables xxi

Abstract xxii

Dedication xxv

Acknowledgments xxvi

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Motivation: Cross-cancer Analysis Reveals Clinically Significant Findings 5
  - 1.2.1 Conclusion . . . 9

2 Novel Tools and Methods to Help Tumor Subtyping 10
- 2.1 UCSC Tumor Map: Exploring the molecular similarities of cancer samples on an interactive portal . . . 11
  - 2.1.0.1 Publication Title and Author List . . . 12
  - 2.1.1 Abstract . . . 12
  - 2.1.2 Background . . . 13
  - 2.1.3 Results . . . 15
    - 2.1.3.1 Supplemental Results . . . 26
  - 2.1.4 Methods . . . 32
    - 2.1.4.1 Datasets . . . 32
    - 2.1.4.2 Map Creation To Reveal Molecularly Similar Sample Groups . . . 35
    - 2.1.4.3 Attributes for Interpreting Biological Relevance of Sample Groups . . . 38
    - 2.1.4.4 Methods for Analysis of the Integrated Pan-Cancer-12 Map . . . 44

2.1.5 Discussion . . . 51
2.1.6 Future Direction: Submaps . . . 54
2.1.7 Conclusion . . . 55
2.2 Data Transformations Aid in Molecular Pattern Discovery . . . 55
2.2.1 Reciprocal Significance of Similarities (RSS) . . . 56
2.2.2 Applications of RSS to Analysis of Cancer Datasets . . . 58
2.2.2.1 Batch Effect Removal with RSS (Proof of Concept) . . 58
2.2.2.2 Batch Effect Removal with RSS In Joint Gliomas Analysis 60
2.2.2.3 Integration Of Platforms As Data Types For Tumor Map 64
2.2.2.4 Integration Of Platforms For Master Regulator Analysis 65
2.2.3 Conclusion . . . 69
2.3 Kernel Space Comparison Helps Contrasting Transformations . . . 70
2.3.1 Conclusion . . . 71

3 Discovering New Biology in Cancer Has Potential to Help Cancer Patients 73
3.1 Analysis of Gliomas of Combined Grades and Histologies . . . 74
3.1.1 Unsupervised Analysis of RNA-Seq and Methylomics Combined Space Using Tumor Map . . . 75
3.1.1.1 Combining Multi-platform Multi-tumor Datasets . . . 75
3.1.1.2 Combined RNASeq and Methylation Space Reveals Important Relationships Between Molecular Subtypes of Gliomas . . . 76
3.1.2 Pathway Analysis Reveals Important Molecular Differences Between GBM and LGG Tumors Within The Same Subtypes . . . 77
3.1.3 Conclusion . . . 84
3.2 Analysis of Cholangiocarcinoma . . . 85
3.2.1 Clustering Based on Tumor Map Layout Positions Provides an Alternative Way to Infer Molecular Subtypes . . . 86
3.2.2 Tri-cancer Multi-platform Analysis of Cholangiocarcinoma, Pancreatic Adenocarcinoma, Liver Hepatocellular Carcinoma Helps Identifying Important Histological Subtypes From Molecular Data 88
3.2.3 Conclusion . . . 90
3.3 Analysis of Testicular Germ Cell Tumor . . . 91
3.3.1 Analysis of Match Primary Tumors Shows Independent Origin of These Malignancies . . . 91
3.3.2 Unsupervised Molecular Subtypes in PARADIGM IPL Space Correlate With Histological Labels . . . 92
3.3.3 Integrated Molecular Space Reveals Important Relationships Between Histological Subtypes . . . 96
3.3.4 Analysis of Pathway Activity Space Helps Characterizing Histological Subtypes . . . 97

3.3.4.1 Characterization of Histological Subtypes Through Unsupervised Analysis . . . 98
3.3.4.2 Epithelial to Mesenchymal Transition Pathway Plays Important Role in Mixed Nonseminoma Tumors . . . 99
3.3.4.3 ERBB Pathway Plays Important Role in Mixed Nonseminoma Tumors . . . 103
3.3.4.4 MYC Pathway Plays Important Role in Embryonal Nonseminoma Tumors . . . 105
3.3.5 Conclusion . . . 107
3.4 Analysis of Mesothelioma . . . 107
3.4.1 Pathway Activity Space Reveals Prognostically Significant Molecular Subtypes of Mesothelioma . . . 108
3.4.2 Molecular Space Shows Important Spacial Separation Of Prognostically Important Subtypes . . . 112
3.4.3 Comparison of Two Most Differentially Surviving Molecular Subtypes Reveals Important Markers of Aggresiveness In Poor Survivors 112
3.4.4 Pan-cancer Analysis Reveals Sarcoma-like Subtype of Mesothelioma Tumors . . . 115
3.4.4.1 Tumor Map Reflection . . . 119
3.4.5 Conclusion . . . 123
3.5 Analysis of Sarcomas . . . 123
3.5.1 Analysis Driven By Molecular Data Aids In Imputing Histology 124
3.5.2 Conclusion . . . 126
3.6 West Coast Dream Team Analysis of Castration Resistant Prostate Cancer 127
3.6.1 Identifying Molecular Subtypes Of CRPC And Defining Small Cell Phenotype Signature . . . 128
3.6.2 Deriving Stemness Signature By Rank Aggregation . . . 131
3.6.2.1 Discovering Differences Between CD49f-high And Small Cell Signatures . . . 133
3.6.3 Conclusion . . . 138
3.7 Identification of Early Metastatic Signature in Prostate Adenocarcinoma 139
3.7.1 Metastatic Biopsies Exhibit More Similarity to the Matched Primary Tumors Than Unrelated Metastatic Tumors . . . 140
3.7.2 Data Preprocessing . . . 142
3.7.3 Subtyping Primary and Metastatic Prostate Adenocarcinoma Identifies Metastatic-like Primary Subtype . . . 145
3.7.4 Metastatic Signature Helps Identifying High Risk Individuals in Primary Cohort . . . 146
3.7.5 Validation Using Matched PrimaryMetastatic Samples Highlights Advantages of Our Method . . . 149
3.7.6 Important Metastatic Markers Are Revealed Through Analysis of Metastatic-like Prostate Adenocarcinoma Subtype . . . 150
3.7.7 Conclusion . . . 153

3.8 Chapter Conclusion 154

4 Bringing Cancer Informatics Into the Clinic To Advance the Field of Personalized Medicine 155

4.1 Comparison with Cancer Genomic Datasets Can Benefit Individual Pediatric Cancer Patients: Clinical Case Report 156
4.1.0.1 Publication Title and Author List 157
4.1.0.2 Abstract 157
4.1.0.3 Introduction 158
4.1.0.4 Overview of the Current Field of Pediatric Cancers and Therapy and Genomic-guided Therapy 160
4.1.0.5 Case Presentation and Clinical History 161
4.1.0.6 Results 162
4.1.0.7 Discussion 167
4.1.0.8 Methods 169
4.1.0.9 Conclusion 172
4.2 California Kids Cancer Comparison Initiative 173
4.2.1 General Approach 174
4.2.2 CKCC Data Preprocessing 177
4.2.3 Methods for Assessing CKCC Map Robustness 178
4.2.3.1 Local Neighborhood Robustness 179
4.2.3.2 Global Layout Robustness 187
4.2.4 N-of-1 Tumor Map Placement 189
4.2.4.1 Placement Based on Nearest Neighbors 190
4.2.4.2 Future Directions for N-of-1 Map Placement 191
4.2.5 Additional Future Directions 194
4.2.5.1 Methods for Assessing N-of-1 Placement Robustness . . 194
4.2.5.2 Neighborhood Analysis 196
4.2.5.3 Derived Feature Spaces 199
4.2.6 Tool Deployment Can Help Others Use Our Tools 200
4.3 Chapter Conclusion 200

5 Additional Work and Future Directions 202

5.1 Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer 202
5.1.1 Metastatic and Primary Prostate Tumors Separate In Transcriptional Space 203
5.1.2 Conclusion and Future Work 204
5.2 Centromere Reference Models for Human Chromosomes X and Y Satellite Arrays 205
5.2.1 Data 206
5.2.2 SatNP Method Successfully Classifies Human Populations Based on DYZ3 Sequences 207

5.2.3 Future Direction 209
5.2.4 Conclusion 210
5.3 The Molecular Taxonomy of Primary Prostate Cancer 210
5.4 Pan-cancer Analysis of Small Cell Tumors Can Shed Light Into the Biology of This Tumor Type 211
5.4.1 Future Directions 212
5.5 Clinical Mutation Tests Are Predictive Of Candidacy For Immunotherapy 213
5.5.1 Introduction 214
5.5.2 Results 214
5.5.2.1 TCGA in the context of FoundationOne gene panel . . 214
5.5.2.2 MMR Pathway in the context of TCGA and FoundationOne gene panel 216
5.5.2.3 Utilizing FoundationOne gene panel to predict hypermutated phenotype 220
5.6 Discussion 222
5.6.0.4 Future Directions 229
5.7 Chapter Conclusion 230

Bibliography **232**

List of Figures

1.1 Tumor Map depicting RNA-Seq space of GBM and LGG tumors. Upper left: tumors are colored by the cancer type. Lower right: tumors are colored by the molecular subtype. 8
1.2 Kaplan Meier showing survival of the three groups of the GBM+LGG joint cohort. The three groups are: GBM samples, LGG samples that cluster with GBM, LGG samples that cluster away from GBM. Survival of the LGG samples that cluster with GBM is closer to the GBM tumors than it is to other LGG tumors. 8
1.3 Genomic aberration space of GBM and LGG tumors. It shows that LGG IDH WT tumors are more similar in that space to GBM tumors than they are to the other LGG tumors. 9

2.1 Outline of Tumor Map construction. Individual molecular platforms (Omics Data) are provided as input from which pairwise similarities between samples are calculated to produce Similarity networks and standardized using the RSS (RSS; see 2.2) to create a coherent space of standardized similarity networks. Map layouts are created with OpenOrd algorithm using coherent sample networks. Integrated multi-platform maps are created from several coherent networks, combined before input to OpenOrd layout procedure. Shown is an mRNA-based map; colors represent tissue of origin. Attributes – clinical, molecular, phenotype, or outcome metadata – annotate samples using colors and color gradients based on groupings that can be defined by the user. 16
2.2 Maps produced from different molecular data types. Samples in the maps are colored by tissue of origin. (A) Maps produced from each of the six molecular data types. (B) Maps produced from inferred gene activities using the PARADIGM and SPIA methods. (C) Maps integrating more than one molecular data type. 17

2.3 Known biology recapitulated by the mRNA expression map. (A) BRCA molecular subtypes and their layout in the map. The map shows that tumors of the same molecular subtype tend to cluster together, with very little mixing of those subtypes, indicating that those subtypes are indeed molecularly different. (B) Estrogen signaling (yellow) and basal signaling (blue) are the top differentiating programs between the basal and non-basal BRCA tumors. Very few tumors exhibit signals of both programs (green). The group of samples labeled as Basal subtype in part A predominantly exhibits the basal signaling program and the group of samples consisting of LumA (for luminal A) and LumB (for luminal B) in part A predominantly exhibits the estrogen signaling program. (C) Mutual exclusivity of PIK3CA (yellow) and TP53 (blue) mutation events in BRCA samples. Most samples carry mutations in only one of those genes. Very few samples show mutations in both of the genes (green), illustrating the documented phenomenon of mutual exclusivity of PIK3CA and TP53 mutations. (D) COAD and READ tumors cluster together (left) and the map separates genomically stable and unstable tumors (right). (E) BLCA tumors separate into three previously discovered molecular subtypes. These subtypes are BLCA-core, BLCA-lung-like, and BLCA-squamous-like. (F) KIRC tumors are deficient in MSH2 (activity level indicated by the intensity of yellow), a component of the DNA mismatch repair pathway. This is a known characterization of KIRC tumors. (G) Co-occurrence of VHL mutations (green) and high HIF1A activity (indicated by the intensity of yellow) in KIRC tumors. Each sample in the map is colored by two colors. Samples colored in yellow indicate an absence of VHL mutation but high HIF1A activity. Samples colored in green indicate both the presence of VHL mutation and high HIF1A activity. No samples show a presence of VHL mutation and low HIF1A activity (such samples would be colored in blue). 19

2.4 Tumor Map reflection analysis shows the overlap in gene signatures distinctive of the Tumor Maps relative positioning of BRCA molecular subtypes. HER2+ tumors share more in common between luminal than basal tumors. A reflection analysis was performed for each of the molecular subtypes; the subtype groupings are defined by BRCA sample annotations. The reflection analysis resulted in 150-gene signatures for each subtype. Venn-diagram of these gene signatures shows the overlap between gene sets for each of the subtype. 20

2.5 Tumor Map rendering of Pan-Cancer-12, an integrated cross-cancer Tumor Map based on six molecular data platforms. (A) Several groups of interest are shown including: (i) BRCA tumors cluster into two major groups, with basal samples grouping with squamous tumors. (ii) LAML tumors separate into two major groups with one group significantly enriched for favorable cytogenetic risk. (iii) Separation of endometrioid UCEC tumors into two major groups, one of which is characterized by a 1q chromosome amplification event. (iv) An integrated pan-cancer cluster, defined by tumors from nine different tissues of origin, exhibits a strong immune signature. (B). Pathway representation of immune signaling-related genes characterizing the integrated pan-cancer cluster showed in A, including markers of both the innate and adaptive immune systems. 22
2.6 Tissue and molecular subtype distribution among the samples in the entire cohort (A-B), and in the pan-cancer cluster (C-D). The pie charts represent the number of samples from each tissue of origin in the entire cohort (A) and the integrated pan-cancer cluster (C). Black and white matrices illustrate the presence of molecular features of each platform (x-axis) across samples (y-axis), in the entire cohort (B) or in the integrated pan-cancer cluster (D). Data available for this sample for a given platform is marked black, otherwise the entry is white. 23
2.7 Enrichment of Immune signaling in the integrated pan-cancer cluster. Different level of evidence support the association of the integrated pan-cancer cluster with an immune phenotype. (A) Enrichment of T- and B-cell signaling shown on the integrated map, yellow gradient. (B) Enrichment of high ESTIMATE scores in the pan-cancer samples, waterfall plot with pan-cancer cluster samples in red. (C) Enrichment of the immune-related pathways identified by differential expression analysis. (D) Unsupervised analysis of master regulator activities inferred by the MARINa method. Gene clusters enriched for T- and B-cell signaling, interferon signaling, and TNFA via NFKB signaling, red font. (E) The top enriched pathways based on the output of master regulator scores derived with MARINa are immune-related. 24
2.8 Purity estimates in the integrated pan-cancer cluster compared to all other samples in the cohort. The pan-cancer cluster (green box) shows lower purity when compared to the whole cohort as a background (yellow box). This finding is consistent with other analyses indicating high immune signaling in the pan-cancer cluster. 25

2.9 Copy number events in the integrated pan-cancer cluster compared to other samples in the full cohort. The pan-cancer cluster shows a lower number of copy number events in both arm-level events and focal events. (A) Arm-level events (pan-cancer group on the right, background cohort on the left). (B) Focal events (pan-cancer group on the right, background cohort on the left). 25
2.10 Mutation frequencies among the genes that are part of the DNA mismatch repair (MMR) pathway across the whole TCGA cohort. The barplot shows all MMR genes sorted (from left to right) by the frequency of mutations in those genes across all the samples in the TCGA cohort. MSH2 gene is ranked 8th among the 23 MMR genes. 32
2.11 Statistical tests computed by different attribute enrichment analysis (AEA) tools available in the Tumor Map. 40
2.12 Method to compute differential expression for samples in the integrated pan-cancer group vs. other samples in the TCGA cohort. Tissue composition imbalance was corrected for by performing t-tests within each tissue. For each gene, t-statistics were computed within each tumor type separately and then summarized per-gene t-statistics were calculated as an arithmetic mean, weighted by the inverse variance of the all the tissue-specific t-statistics values. 45
2.13 Transcription factors inferred by the MARINa method contrasting differential gene expression between the integrated pan-cancer cluster and other samples within each tissue. Specific master regulators vary depending on the tissue of origin but some themes are shared across tumor types. Each matrix shows results for each tissue type. For each transcription factor (rows), the expression levels of each of its targets (tick marks) are colored according to whether the TF is predicted to activate (red) or inactivate (blue) the target. The inferred activity is illustrated to the right of the factor (activated, red; inactivated, blue) along with its expression level and the number of its targets. A P-value is written along the last column. 48
2.14 Venn diagram representing the innate and adaptive immune systems and different levels of evidence supporting higher activity of each of the components of those systems in the pan-cancer cluster when compared to the rest of the TCGA cohort. We found evidence of both innate and adaptive immune system signaling with a number of different analyses. 50
2.15 Application of RSS method for batch effect removal when combining multi-platform mRNA expression datasets for 6 cell lines achieves the best clustering of the same cell lines together. A) Hierarchical clustering of the pre-transformed microarray and RNA sequencing data after combining it. B) Hierarchical clustering of the data after ComBat batch effect removal method was applied to it. C) Hierarchical clustering of the data after RSS method was applied to it. 60

2.16 Distribution of molecular platforms and IDH molecular subtypes in the joint gliomas mRNA expression data. A) Distribution of IDH subtypes in the GBM data across microarray and RNA sequencing platforms. B) Distribution of the IDH subtypes in the GBM and LGG data. 62
2.17 Match samples that are common between the GBM microarray and RNA-Seq datasets. 62
2.18 View of the mRNA expression data for GBM and LGG tumors prior to applying data transformation. A) Hierarchical clustering of the mRNA expression data shows clear clustering by the experimental platform. B) Principle Component Analysis of the mRNA expression data shows separation by the experimental platform. 63
2.19 Summary of how RSS method was applied to the gliomas mRNA expression data. 63
2.20 View of the mRNA expression data for GBM and LGG tumors after applying RSS data transformation. 64
2.21 Outline of the application of RSS method to mRNA expression, CNV, and methylomics CHOL data. 67
2.22 Validation of the SPIA results by showing the top differential pathway between the two groups of interest. As expected, the top differentiating pathway is Oxidative Phosphorylation. On the left: Gene Set Enrichment Analysis of the differential master regulator signature, showing the statistical significance of the Oxidative Phosphorylation pathway genes. On the right: each dot is a sample in the cohort; the samples are separated into two groups, reflecting the two groups of interest in our analysis; the pathway levels were obtained by aggregating the master regulator scores across all the genes in this pathway for each of the samples. 68
2.23 Resulting master regulator network, output by PATHMARK method. Red nodes are high in the "Oxidative Phosphorylation" group of samples. Blue nodes are high in the "Chromatin Remodeling" group of samples. The biggest difference between the two groups is proliferative signaling (high in the "Oxidative Phosphorylation" group). 69
2.24 Application of kernel alignment method to two sets of kernel matrices. A) First set of matrices are patient-to-patient age difference converted to similarity space and the second set of matrices are bladder cancer mutation profile similarities, computed using Humming similarity. B) First set of matrices are a patient-to-patient age difference converted to similarity space and the second set of matrices are similarities of glioblastoma MRI images represented by voxels. 71

3.1 Distribution of the standard deviation of the gene features for the GBM and LGG tumors. The red line shows the SD cutoff for the features used to by the Tumor Map method. Genes on the right of the red line where included into the analysis. 79

3.2 Tumor Map analysis based on individual platforms shows the separation of glioma molecular subtypes differs in the mRNA expression and DNA methylation spaces A) mRNA expression Tumor Map B) DNA methylation Tumor Map. 80
3.3 Tumor Map based on mRNA expression and DNA methylation data. Each data point is a TCGA sample colored coded according to their identified status. 81
3.4 Tumor Map based on mRNA expression and DNA methylation data. Each data point is a TCGA sample colored coded according to their identified status. 82
3.5 Table showing distribution of GBM and LGG samples in various dichotomies. 82
3.6 The method used to extract significant pathways driving the LGr3 and LGr4 groups. 83
3.7 Pathways involved in progression of the two expression subtypes of the GBM and LGG tumors. The figure is displayed as it was included into the published manuscript as a part of the supplemental figure S5. The two pathways are part C and D of figure S5. Part (C) shows the pathways that drive LGr3 (IDH mutant enriched subtype). Part (D) shows the pathways that drive LGr4 (IDH WT enriched subtype). 84
3.8 Tri-cancer RNASeq data clustered using Tumor Map. A) CHOL, LIHC, and PAAD samples laid out in Tumor Map based on RNASeq data. B) Cluster assignments (k=7) based on the sample positions in the Euclidean 2-D space (using $1 - Euclideandistance$ as a measure of similarity). C) Sample-to-sample correlations in the original RNASeq space based on the Tumor Map clustering solution. D) Feature (gene expression) space (from RNASeq data) based on the Tumor Map clustering solution. . . 87
3.9 Tumor Map depicting tri-cancer (CHOL, LIHC, PAAD) integrated genomic space (RNA-Seq, CNV, methylation). The samples in the map are colored based on the tissue of origin. 89
3.10 A view of the CHOL cohort in light of various genomic markers and signatures (exert from main figure in the manuscript). The samples (listed below the heatmap) are ordered based on their molecular subtype along the x-axis and are annotated by various features along the y-axis. All the samples that clustered with PAAD samples robustly correspond to the extrahepatic cholangiocarcinoma (ECC) subtype. IDH mutants are enriched for high scores of the Oxidative Phosphorylation signature. . . 90
3.11 Outline of the PARADIGM analysis pipeline and its various parts. . . 94
3.12 Results of the unsupervised analysis of TGCT cohort in PARADIGM space (based on top 3000 most varying IPLs). The molecular subtypes highly correlate with the histological subtypes. 95

3.13 Comparison of match primaries in PARADIGM space. Each first primary was correlated with its corresponding match primary. Each dot in the plot represents a single PARADIGM IPL. There are 5 plots for the 5 individuals with match primaries. 96
3.14 Tumor Map view of combined mRNA expression, somatic copy number, and methylation spaces of TGCT cohort. The map shows almost perfect separation between the major histological subtypes. 97
3.15 View of TGCT histological subtypes in PARADIGM IPL space. 99
3.16 Epithelial to mesenchymal transition is enriched in the mixed non-seminomas. On the left: GSEA plot for the EMT pathway based on the differential IPL signature. On the right: IPLs for the genes in the EMT pathway were aggregated per-sample and the aggregated levels are plotted based on the histological group. 101
3.17 Epithelial to mesenchymal transition pathway identified by the PATHMARK method as relevant to mixed nonseminoma tumors. 102
3.18 ERBB signaling is enriched in the mixed non-seminomas as compared to other groups. 104
3.19 ERBB pathway identified by the PATHMARK method as relevant to mixed nonseminoma tumors. 104
3.20 MYC pathway identified by the PATHMARK method as relevant to embryonal nonseminoma tumors. 106
3.21 Description of the MESO cohort clusters in PARADIGM IPL space. A) PARADIGM clustering solution (k = 4) and relevant feature spaces (PARADIGM IPL, mRNA expression, and CNV) under that solution. All samples are ordered by the cluster annotations at the top. B) Kaplan Meier plot of survival (in days) of the sample groups from the PARADIGM clustering solution. C) Best (blue) and worst (red) surviving clusters across all the platforms are compared. Jaccard Index is computed separately for the best and the worst surviving groups as a measure of distance between that and the corresponding group in PARADIGM solution. 110
3.22 Best and worst surviving groups for each of the platforms analyzed for MESO cohort as defined by the Kaplan Meier (KM) survival plot for each platform. Each survival plot was produced by the individual collaborators working on that platform. KM plots from each platform are combined into a single figure here and best and worst surviving groups are indicated with the corresponding arrows. 111
3.23 Mesothelioma tumors clustered in PARADIGM space using Tumor Map. Each tumor is a node in the map. The nodes are laid out based on similarities of their PARADIGM IPL profiles. A) The samples are colored by histological labels. B) The samples are colored by PARADIGM subtypes. The good surviving cluster is 3 (blue) and the worst surviving cluster is 1 (red). 112

3.24 Analysis of differential activities in the good survivors (cluster 4/blue) and bad survivors (cluster 1/red) groups. The rows of the heatmap are statistically significant protein-coding PARADIGM IPLs. The pathway enrichment analysis of positive and negative differential IPLs was performed using MSigDB resource. 114
3.25 Analysis of differential aggregated activities within statistically significantly differential pathways between cluster 1 (bad survivors) and cluster 3 (good survivors). 114
3.26 Multi-tumor RNASeq data clustered using Tumor Map. There are 8 different types of tumors in this map (per expert suggestion from the group only basal breast carcinoma tumors were used from that tumor type cohort). A) Tumors are colored by tumor type. B) Tumors are colored by mesothelioma histology (other tumors are gray). C) Tumors are colored by sarcoma histology (other tumors are gray). D) Tumors are colored by MESO PARADIGM subtypes (other tumors are gray). . . . 117
3.27 Mesothelioma samples that do not cluster with either other mesothelioma tumors or sarcoma tumors in the multi-tumor RNASeq data clustered using Tumor Map that incorporates 8 different types of cancer (per expert suggestion from the group only basal breast carcinoma tumors were used from that tumor type cohort). 118
3.28 Distribution of histology labels across the PARADIGM subtypes (cluster 3 are the good survivors; cluster 1 are the bad survivors). 118
3.29 Reflection analysis of sarcoma-like MESO tumors and the UP genes that drive the similarity of those tumors to undifferentiated sarcomas. A) Venn diagram shows overlaps of UP genes from four reflection analyses (sarcoma-like MESO tumors, non-sarcoma-like MESO tumors, epithelioid-only sarcoma-like MESO tumors, and epithelioid non-sarcoma-like MESO tumors). B) MSigDB enrichment analysis of the 52 genes overlapped between epithelioid-only sarcoma-like tumors and all sarcoma-like tumors. 119
3.30 Sarcoma tumors in RNA-Seq space analyzed with Tumor Map show three major groups, driven by histological subtype groupings. 125
3.31 Multi-cancer analysis of ten different tumor types, including sarcomas, in RNA-Seq space using Tumor Map. The samples labeled as "Exclude" during the AWG analysis are marked and identified in the figure. . . . 126
3.32 Outline of the preprocessing pipeline of the 89-sample CRPC mRNA expression data prior to unsupervised clustering. 129

3.33 Chosen solution of k = 8 for the unsupervised census k-means clustering. The solution was chosen based on the silhouette score method. The clusters for the samples are presented by the heatmap. They show how histological labels assigned by a panel of pathologists are distributed among the 8 clusters. The heatmap in the bottom left shows statistical significance of the histological label enrichments in the clusters. The box is colored red if the p-value <= 0.5 and pink if 0.05 <= p-value <= 0.1. The value shown in the boxes is the FDR for each significance. 130
3.34 Results of feature space clustering of the k = 8 sample clustering solution. The gene clusters are annotated by their enrichments in Hallmark and Canonical Pathways sets from MSigDB pathway database. 131
3.35 Comparison of gene ranks in full 20,500-gene CD49f-high and supervised Small Cell Neuroendocrine signatures. Very few ranks actually correlate between the two signatures. 133
3.36 Overview of the background model for rank inconsistencies between the CD49f-high signature (designated as Sig1 in the figure) and Small Cell Neuroendocrine signature (designated as Sig2 in this figure). 135
3.37 Pathways significantly enriched and differential between the CD49f-high and supervised small cell signatures. 136
3.38 Overview of the method used to extract the pathways relevant to the differences of the CD49f-high and supervised small cell signature. . . . 137
3.39 Individual pathways relevant to the CD49f-high (EMT and Wnt) and supervised small cell signatures (Interferon signaling). 138
3.40 Correlation of the nine pairs of the metastatic TCGA samples to their corresponding match primary tumors (red) and to other metastatic tumors in the same tissue (blue). Metastatic and primary tumors within the same patient show higher similarity than if different patients are compared. 141
3.41 Principle component analysis (PCA) of the mRNA expression dataset included into the meta-analysis before (A-B) and after (C-D) ComBat adjustment. A and C show dataset distribution among samples (pre- and post-ComBat). B and D show platform distribution among samples (pre- and post-ComBat). 143
3.42 Variance across 4,895 genes in the combined ComBat-transformed mRNA expression dataset. The red line indicates where the cutoff for variance was placed. All genes on the left of the line were filtered out of the consecutive analysis. 144
3.43 Heatmaps of (a) primary and (b) metastatic cluster solutions with additional clinical covariate annotation bars. 146
3.44 Description of possible primary-to-metastatic progression models. . . . 148
3.45 Confusion matrix (balanced accuracy) for leave-one out cross-validation model trained to predict primary prostate cancer subtypes. 148

3.46 Two different views of mapping between the metastatic and primary subtypes. A) Barplot shows which primary cluster each metastatic sample mapped to. Primary cluster 2 has the most metastatic samples mapped to it. Clusters 3 and 4 has the least number of metastatic samples mapped to it. B) Ribbon plot shows primary and metastatic subtypes and their mapping to the corresponding primary cluster. Multiple metastatic subtypes map into primary cluster 2. 149
3.47 Two top enriched pathways in the metastatic-like signature. 152
3.48 Differential subnetworks, based on mRNA expression in primary prostate cancer. Red color correspond to genes up-regulated in the metastatic-like primaries and blue color corresponds to the genes up-regulated in less aggressive subtypes. Node size corresponds to the connectivity (edge count); large nodes indicate network "hubs". 153
3.49 Overview of the method Graim and I co-authored to detect early metastatic signature in primary prostate cancer. 154

4.1 Patient 1s RNA sequencing profile in the context of the reference cohort of 38 different tumor types, both pediatric and adult. A) A projection of the entire tumor cohort on a 2-D map using Tumor Map method. Each tumor in the map, represented by a hexagon, is colored by the tumor type as described in the legend. Patient 1s tumor is shown in green in the lung cluster. B) LUAD tumors cluster in several subtypes on the Tumor Map visualization. The arrows point out different clusters LUAD tumors belong to. C) In yellow we highlight the tumors that were considered part of the Patient 1 cluster. We ran differential gene expression analysis of those tumors vs. other lung tumors. The intersection of the statistically significant differentials and the druggable up outliers (called Therapeutic Targets) is shown by the Venn diagram. 165
4.2 Patient 1s tumor expresses ALK at a level similar to those in ALK-driven malignancies . A) Expression of ALK and JAK1 as compared to the expression of tumors in the reference cohort, separated by tumor type. B) Expression of ALK in Patient 1 is comparable to ALK-driven lung adenocarcinoma and neuroblastoma . C) Comparison of ALK and JAK1 expression levels in the Patient 1 cluster with other LUAD tumors. . . . 166
4.3 Proposed pathway that may contribute to the disease in Patient 1. This pathway demonstrates how both ALK and JAK1 participate in the activation of tyrosine kinase signaling in Patient 1s tumor. 167
4.4 CKCC version 1 reference dataset visualized using Tumor Map method. Each dot/hexagon in the map is a sample in the reference cohort and the samples are laid out based on the similarity of their gene expression profiles and are colored by the disease. 176

4.5 Results of the CKCC analysis on 19 individuals. The therapeutic leads are color coded by the type of the recommendation. The columns provide information about specific molecular pathways for which the lead was found. 177
4.6 High level outline of how we assess the stability of the reference map and robustness of the sample placements in the map. 179
4.7 Overview of the method for assessing local neighborhood robustness: local neighborhood specificity (bottom left) and local neighborhood variance (top and bottom right). 180
4.8 Assessment of CKCC version 1 local neighborhood specificity across cohort samples under feature space subsampling (C) and under feature space shuffling (B). Most samples retain between 90% to 100% of their true local neighborhood under the condition of feature space subsampling. 183
4.9 Median number of times samples appear in the local neighborhood of a given sample over 1,000 iterations of either subsampling of the feature space (bar plot on the right) or shuffling of gene labels (bar plot on the left). For most samples local neighborhoods do not vary at all over the feature space subsampling iterations. 185
4.10 Total number of samples that appear in the local neighborhood of a given sample over 1,000 iterations of either subsampling of the feature space (B) or shuffling of gene labels (C). For most samples under the subsampling conditions the cardinality of the local neighborhoods across all iterations is very close to the true size of the local neighborhood (N = 6), suggesting that these neighborhoods do not vary much as compared to the gene shuffling conditions where cardinality is in several 1,000s. . 186
4.11 Overview of the method for assessing global neighborhood robustness: global neighborhood sensitivity. 188
4.12 Example of a "landscape pin" indicating an N-of-1 placement into Tumor Map. A) A birds-eye-view of the entire reference cohort map. B) A zoom-in into the area of the map where the pivot sample is placed. In this example the pivot is placed into with sarcoma tumors. Mesothelioma tumors, which are biologically similar to sarcomas (see 3.4), are near by. 190

5.1 Visualization of the transcriptomic space of 13 different types of cancer. A) The map of the entire cohort. Tumors are colored by the tissue of origin. B) metastatic tumors separate from primary and benign tumors in the transcriptomic space of the Grasso *et al.* dataset. 204
5.2 Satellite DNA consists of tandem repeats. These repeats may vary among individuals and populations and sequence kmer frequencies carry information about variation in these repeats. 206
5.3 Overview of SatNP method. A) High-level diagram of the SatNP pipeline. B) Overview of the feature selection pipeline. 208

5.4 Results of our SatNP pipeline to differentiate between western European and east African populations. We identified 1,153 informative kmers and demonstrated our model makes predictions with high accuracy. 208
5.5 Description of pan-cancer small tumor work. A) List of datasets included into this study. B) Overview of the batch/dataset correction pipeline applied to the combined data. C) Consesus k-means clustering of the samples after the batch correction was applied. Small cell tumor appear to mostly cluster together. 212
5.6 Exploration of the mutational landscape of the TCGA cohort in the context of FoundationOne gene set. A) Top 25 cancer hallmarks (MSigDB) represented within the FoundationOne gene panel. B) Top canonical pathways (MSigDB) represented within the FoundationOne gene panel. C) Mutation counts across TCGA cancers within the context of FoundationOne genes only. D) Mutation counts across TCGA cancers (entire genome). 216
5.7 Exploration of MMR pathway and its mutations with TCGA cohort. A) Mutation counts per sample in MMR pathway genes. Most samples have a single mutation within MMR pathway showing that a single mutation is sufficient to disrupt this pathway. B) Mutation frequencies by each MMR gene within TCGA cohort. Different genes exhibit different mutation frequency suggesting difference in importance to MMR functionality. C) Top 4 out of 5 MMR genes in FoundationOne panel exhibit high mutual exclusivity suggesting that a mutation in a single gene is sufficient to disrupt the pathway function. D) Mutation frequencies in the TCGA cohort broken down by MMR mutants vs. not when using all genes in the genome. If there was at least a single mutation in the MMR pathway the sample was assigned to MMR mutants group. Welch t-test p-value was computed between the two groups. E) Mutation frequencies in the TCGA cohort broken down by MMR mutants vs. not when using FoundationOne genes only. If there was at least a single mutation in the MMR pathway the sample was assigned to MMR mutants group. Welch t-test p-value was computed between the two groups. F) Expression profiles of MMR mutants and MMR wt were compared (Pearson Rho). Expression profiles of the samples that have MMR disruptions are more similar than the expression profiles of the samples without MMR disruptions as well as expression profiles between the two groups. 219
5.8 Random Forest model using 313 genes mutation profiles only. 224
5.9 Random Forest model using mutation frequency only. ROC for each of 5 folds of 5-fold cross validation Random Forest model when using mutation frequency as the only input feature. Each color indicates each fold. Average AUC is displayed on the bottom of the plot. 225
5.10 Random Forest model using both 313 genes mutation profiles and mutation frequencies. 226

5.11 Random Forest model using 313 genes mutation profiles only. 461 samples with a mutation in one of the 5 MMR FoundationOne genes were excluded. 227
5.12 Random Forest model using mutation frequency only. 461 samples with a mutation in one of the 5 MMR FoundationOne genes were excluded. ROC for each of 5 folds of 5-fold cross validation Random Forest model when using mutation frequency as the only input feature. Each color indicates each fold. Average AUC is displayed on the bottom of the plot. 228
5.13 Random Forest model using both 313 genes mutation profiles and mutation frequencies. 461 samples with a mutation in one of the 5 MMR FoundationOne genes were excluded. 229

List of Tables

2.1 Summary of samples in the pan-cancer cluster in the integrated Tumor Map. 45

3.1 Datasets used in the meta-analysis. 144
3.2 Predicted primary clusters are enriched in samples that metastasized early. 150

5.1 Results from MMR predictors in TCGA cohort 231

Abstract

Discovering Molecular Patterns with Therapeutic Implications in Large-cohort Heterogeneous Cross-cancer Data

by

Yulia Newton

Recent advances in high-throughput genomic technologies and high-performance computing have propelled the science of computational genomics into a new era and launched the field of precision medicine. Computational genomics is now an integral part of biomedical research and genomic testing is routinely performed in clinical settings. In the field of cancer informatics, the integration of genomics has led to invaluable insights and discoveries. We study cancers in order to better understand tumorigenesis and disease progression. This understanding can, in turn, inform and guide therapeutic decisions and suggest directions for drug development and repositioning. The ultimate goal of cancer precision medicine is to sequence and analyze every patients tumor in order to provide the most effective and least toxic treatment.

Various experimental platforms are available for collection of different perspectives or views of the cell state, which help us characterize and understand molecular signals driving the cell phenotype. We collectively refer to these views as 'omics' data. While vast amounts of 'omics' data are being collected from tumor samples at an accelerating rate, few resources exist to aid biologists and clinicians in identifying trends in these data, finding connections within and between cancer subtypes, and matching

patients to previously studied patient groups to infer therapeutic implications. In our analysis we also utilize bioinformatics methods that manipulate, transform, and integrated these views to derive new views of the cell. In my doctoral thesis, I present my work developing new tools and methods to aid the scientific community in understanding and interpreting cancer biology (Chapter 2). I also present my work applying such methods to contribute to cancer subtype-specific analyses as part of various projects and collaborations during my doctoral work (Chapter 3).

Finally, I describe my work and contributions to the field of personalized medicine in pediatric cancer. While similar in some ways to adult cancers, pediatric cancers differ dramatically from their adult counterparts on a molecular level. For example, pediatric tumors generally have fewer genomic alterations than adult tumors. Further,childhood cancers are rarer than adult cancers and thus more difficult to study due to a lack of sufficiently large patient cohorts. While some clinics now regularly sequence pediatric tumors for bioinformatic analysis, the sequencing of patient genomes in the clinic is only beginning to impact patient care. Most computational methods for detecting differentially expressed genes are designed for analyzing patient cohorts in research settings and are thus unsuitable for interpreting RNA sequencing data from a single patient. Further, analyzing individuals genomic data leads to actionable treatment options in only fifteen percent of all childhood cancer cases. This is because pediatric cancers are often not driven by non-hereditary genomic changes, and any genomic aberrations that do exist may not be targetable by existing drugs. More sophisticated informatics tools and methods are needed in the field of personalized medicine. To

this end, I describe my work developing methods for single-patient analyses in pediatric cancer (Chapter 4). While my methods were developed for pediatric cancers, they may also be used to analyze adult tumors.

To RJ, my son.

You make it all worth it.

Acknowledgments

I want to thank the members of my committee, Josh Stuart, Sofie Salama, Olena Morozova and David Haussler, for committing their time and leadership to my work and for all the input and mentoring they have provided to me and my work. Specifically, my advisor Dr. Joshua Stuart for guiding and leading me in my endeavors. I would also like to thank the members of the Stuart Lab, Haussler Lab, and the Treehouse team for countless hours of discussions and ideas. I would like to thank my PI Josh Stuart and the National Institute of Health (NIH) for their T-32 Training Grant for my funding over the course of my doctoral work.

Chapter 1

Introduction

1.1 Background

Cancer is a multifactorial multigenic disease that arises from changes to the deoxyribonucleic acid (DNA), which is a molecule that contains cellular blue print in every live organism on Earth. While the causes of these changes to the DNA are multi-source and complex, they cause disruptions to important molecular, regulatory, and metabolic pathways in the cell, often gearing it towards a tumorigenic phenotype. This is why cancers often seem to "target" specific pathways that are involved in cell survival and evading natural cell death (unlimited proliferation, DNA replication, cell cycle, and other cancer hallmarks [43]), often accumulating these disruptions until some "tipping point" at which a normal cell becomes a tumor cell. Frequently genes that promote these pathways are amplified or mutated to activate their function. These genes are referred to as "oncogenes". Another class of genes that usually prevent or counteract tumorigenesis

are called "tumor suppressors". These genes are usually deleted or deactivated via mutations in cancer cells.

We should mention that pediatric cancers (cancers that arise in infants, children, and young adults) are an exception to the previous statement that cancers often require an accumulation of genomic aberrations that target vital pathways. Generally, pediatric cancers are characterized by low mutational load (fewer mutations than we often see in adults). Often these tumors contain oncogenic germ line mutations and other aberrations.

Genomic aberrations are not the only way that cancers emerge. In addition to genome structure there are additional mechanisms that determine cell phenotype. Gene expression is a process by which gene products (proteins, protein components, and non-coding RNAs) are produced by the cell from the blue print that is encoded by the DNA structure. Many expressed genes participate in a downstream gene regulation process as promoters, activators, enhancers, or inhibitors. DNA methylation is a way that cells activate or silence gene expression by adding methyl groups to the DNA molecules. Gene silencing can also occur by micro-RNA (miRNA) molecules binding to messenger RNA (mRNA) molecules, which are normally transported from the cell nucleus into cytoplasm for translation into proteins. We can think about all of these as just different views of the cell, or an insight into the cell phenotype from different angles.

In the age of high-throughput experimental technologies and availability of high-computing power we study cancer by analyzing data from these various views of the cell state. To do that we utilize bioinformatics tools and methods as well as various

transformations of these views to new ones that make information extraction easier. Both unsupervised and supervised methods have their own strengths and utilities, depending on the biological question at hand. Both provide an invaluable methodology for cancer informatics. Methods that can integrate multiple views can often provide additional information about the cell, not attainable by examining single views. Additionally, data transformations have been proven useful for signal extraction from the data. The field of cancer genomics is multi-disciplinary and often benefits from collaborations involving expects with different backgrounds (bioinformaticians, biologists, clinicians, statisticians, etc.).

Because there is no single "recipe" for how cancer cells arise and continue to thrive, there are no two tumors that have exactly the same molecular profiles. Therefore, each tumor is different. Even tumors that come from the same tissue or cell of origin. No two patients have the exact same disease. Consequently, cancer can be considered to be a disease umbrella that incorporates many different individual diseases with individual causes and prognostic and therapeutic implications. This is one of the reasons why curing/treating cancer is such a complex problem. The hypothesis space is just too big for brute force approaches. As cancer researches, we approach this task as a data mining problem. We study cancer cohorts to identify common molecular and clinical patterns. This process is called "subtyping". Subtyping allows us to form and later test a hypothesis about groups of individuals with "similar" tumors and how those groups can be characterized from the prognostic and/or therapeutic perspectives. Looking across multiple cancer types allows us to find oncogenic patterns that transcend cancer

type boundaries. This type of research can drive drug repositioning to new indications. Numerous examples in the cancer research literature demonstrate the utility of such cross-cancer, or often referred to as pan-cancer, analysis. Hoadley *et al.* [59] showed that clinically important groupings can be found across different tissues in their analysis of 12 different cancer types. Many mutations targetable by drugs also span different tissue types [142], opening avenues for utilizing those drugs for additional tumor types. In fact, it has been suggested that we reclassify tumors based on their molecular subtypes [59] in lieu of traditional classification by the tissue of origin. Taking it a step further, we can rethink the traditional therapeutic paradigm as a task of treating oncogenic signatures [53] rather than treating a particular tumor type.

Sometimes different tumors cluster together if they arise from the same cell of origin (as is the case with squamous tumors from multiple tissues [59]) or if they exhibit disruptions to similar molecular pathways (e.g. RAS pathway is known to be disrupted in multiple types of cancers). As researchers, we discover and characterize various groups of tumors by analyzing different views of tumor cells. When a new sample clusters with a group of tumors we already know something about, we can make inferences about this sample by association. From the prognostic and therapeutic perspectives, if a tumor clusters with a group that has therapy associated with it, the same therapy might be effective in treating that tumor.

Finally, while it is important to identify clinically and therapeutically important groups of patients, it is only the first step in helping patients in clinic. We do this type of research with a hope that it can help prospective cancer patients by iden-

tifying biomarkers important in understanding the progression of a particular subtype of cancer. Furthermore if we can match a new patient to this subtype of cancer via common biomarkers or similar molecular patterns, the same prognostic and therapeutic implications that apply to the subtype would apply to the new patient. Most tumor donors involved in these types of academic studies have already passed away and can no longer be helped. The goal of such studies is to advance patient care for future patients.

1.2 Motivation: Cross-cancer Analysis Reveals Clinically Significant Findings

As was previously shown by many other studies, performing large-cohort analysis of tumors can lead to therapeutic implication findings. For example, Hoadley *et al.* [59] found that one of the subtypes of bladder urothelial carcinoma (BLCA) exhibits a squamous signature and patients in that group should be treated as if they have a squamous tumor, which differs from the standard BLCA treatment protocol. Yuan *et al.* [142] describe a number of mutational profiles that span multiple cancers. These finding suggest that drugs that target individual mutations can be repositioned for multiple cancer type indications. For example, Vemurafenib is an oral drug that targets BRAF V600 mutation, which occurs in about 50% of cutaneous melanomas and less than 5% of non-melanoma cancers [38]. It was found that this drug, while originally indicated for melanoma tumors, is effective in BRAF V600 positive patients in non-melanoma tumors [38]. In fact, Hoadley *et al.* [59] suggest a new paradigm shift in the cancer

diagnostic field by proposing a new classification scheme that is based on oncogenic phenotype/signature rather than the tissue of origin. This idea was further elaborated on by Ciriello *et al.* [53] who suggested that instead of treating a cancer type we think of cancer therapy as treating one or more oncogenic signature detected in the particular tumor. An extensive scientific literature now supports this new way of thinking about treating individuals with cancer.

Here I will present a snippet of the work I completed as a part of my doctoral work, which highlights the utility of cross-cancer analysis and the importance of molecular subtyping for prognostic and therapeutic considerations. Traditionally, brain tumors are diagnosed by either histology or tumor grade. Therapeutic decision is usually made based on the grade of the tumor. Tumors with a lower grade, called lower grade gliomas (LGG), are generally less aggressive and exhibit slower progression. Therefore, oncologists generally suggest the "wait-and-see" approach. However, this approach does not work for every LGG patient. Some patients progress very quickly and succumb to the disease despite having been diagnosed with a non-aggressive type of LGG. This behavior reminds another type of glioma tumors. Highest grade gliomas, called glioblastomas (GBM), are highly aggressive tumors and need to be treated more aggressively. Therefore, we combined two tumor cohorts (GBM and LGG) to identify any molecular commonalities between them.

I analyzed RNA-Seq data from joint glioblastoma (GBM) and lower grade glioma (LGG) cohorts. I used Tumor Map method to visualize the joint cohort high-dimensional data as a projection on a 2-D plane (Figure 1.1). We found that some of

the LGG tumors cluster with GBM tumors (Figure 1.1 upper left). When looking at the map in the context of the molecular subtype (Figure 1.1 lower tight) we can see the most of the LGG IDH wild type (WT) tumors cluster with GBMs. In fact, all the LGG tumors that cluster with GBMs are IDH WT subtype. Isocitrate dehydrogenase 1 (NADP+), soluble (IDH) gene is one of the markers of molecular subtypes of LGG tumors. Those cells that have no mutations in this gene, referred to as IDH WT, are molecularly different tumors than those that carry a mutation in this gene [93]. In fact, if we look at the survival of the three groups of tumors (GBM samples, LGG samples that cluster with GBM, LGG samples that cluster away from GBM) we find that survival of the LGG samples that cluster with GBM is closer to the GBM tumors than it is to other LGG tumors (Figure 1.2). This finding is significant because it shows that the RNA-Seq space recapitulates the genomic aberration space (Figure 1.3). It shows that RNA-Seq view of the tumors reflects the genomic aberration space and that Tumor Map and other similar unsupervised analyses methods capture these tumor similarities, suggesting that Tumor Map groupings are clinically relevant.

Our work contributed to the World Health Organization (WHO) updating classifications of the adult brain tumors in 2016 [39]. New classification takes molecular subtypes into account and the recommendation for treatment of the lower grade gliomas is now dependent on the mutation status of the IDH gene.

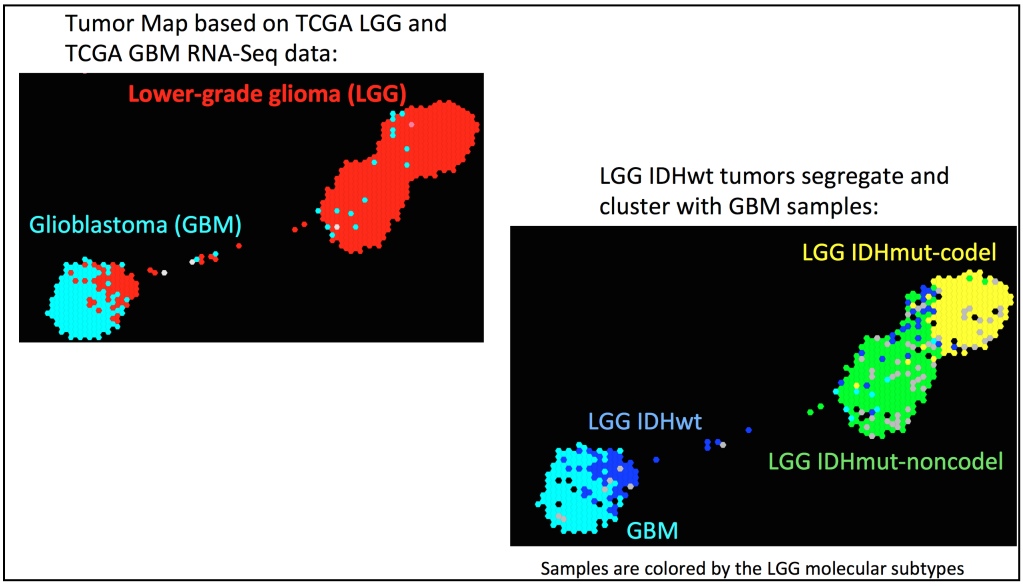


Figure 1.1: Tumor Map depicting RNA-Seq space of GBM and LGG tumors. Upper left: tumors are colored by the cancer type. Lower right: tumors are colored by the molecular subtype.

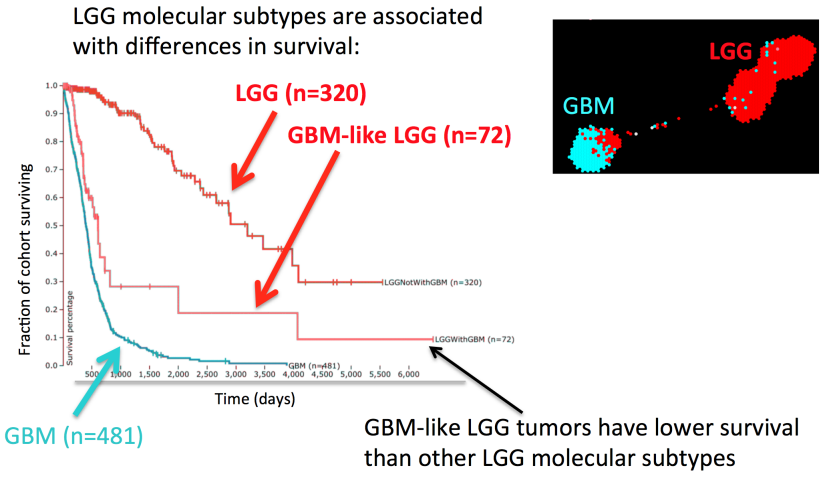


Figure 1.2: Kaplan Meier showing survival of the three groups of the GBM+LGG joint cohort. The three groups are: GBM samples, LGG samples that cluster with GBM, LGG samples that cluster away from GBM. Survival of the LGG samples that cluster with GBM is closer to the GBM tumors than it is to other LGG tumors.

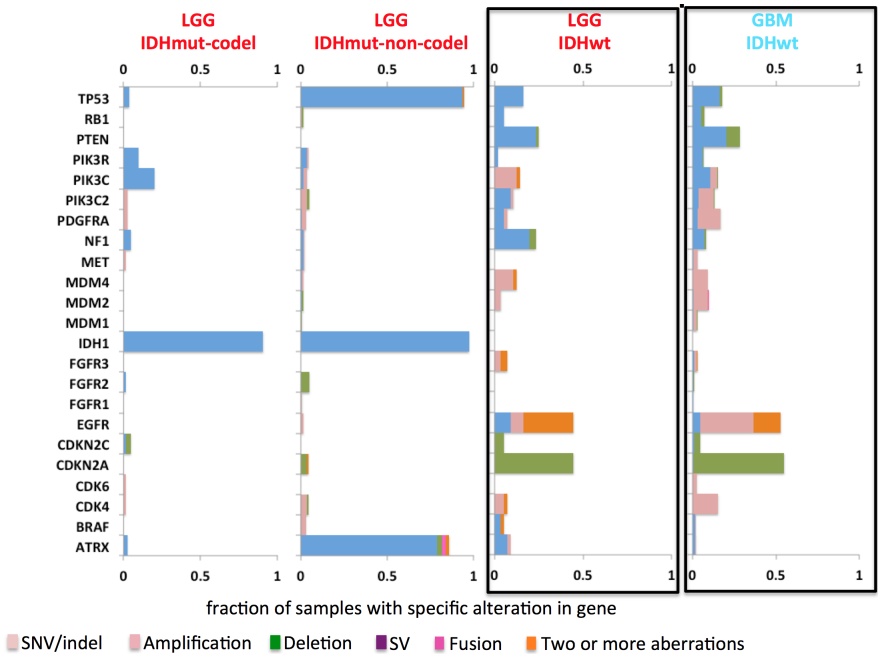


Figure 1.3: Genomic aberration space of GBM and LGG tumors. It shows that LGG IDH WT tumors are more similar in that space to GBM tumors than they are to the other LGG tumors.

1.2.1 Conclusion

We analyzed joint cohorts of the lower grade glioma tumors and glioblastomas. We found that one of the molecular subtypes of lower grade gliomas is very similar to more aggressive glioblastoma tumors. Therefore, individuals that exhibit lower grade but the molecular characterizations of the LGG IDHwt subtype should be treated just as aggressively as high grade tumors. We showed that grade-base and histology-based classifications are not sufficient in patient prognosis. Our work contributed to WHO changing their classification of adult brain tumors in 2016 to account for molecular markers of aggressiveness.

Chapter 2

Novel Tools and Methods to Help Tumor Subtyping

In the field of cancer genomics we heavily rely on the cutting edge tools and methods to perform bioinformatics on cancer data in order to extract useful information from the data and advance the field by learning new cancer biology. However, the field of cancer genomics consists not only of highly trained bioinformaticians, who write and develop these methods, but also cancer biologists, clinical oncologists, pathologists, and other medical professionals. Often, the tasks that are needed to be performed are repetitive and are better off being automated. Novel tools that simplify hypothesis generating and testing are needed in this field, especially if they are straightforward to use for non-bioinformaticians. Often for the analysis tools to be successful, especially in the unsupervised analysis settings, some data transformation or normalization must

be first applied.

In this chapter I describe my work and contributions to the cancer genomics field with new analysis, data transformation, and data integration tools. In the first section (2.1) I describe a novel tool called Tumor Map that allows projection and visualization of the high-dimensional heterogeneous genomic landscape on a 2-D map, similar to navigational Google Maps. This tool comes with a statistical toolbox that allows on-the-fly hypothesis generating and testing for associations of molecular, genomic, phenotypic, and clinical annotations with tumor groupings. In this section I also describe the application of this tool to analyzing Pan-cancer12 dataset, a set of 12 cancers from The Cancer Genome Atlas (TCGA) project [59]. In the second section (2.2) I describe a novel method for data transformation and integration and describe its applications in my doctorate work. Finally, in Section 2.3 I describe my work in kernel space comparison, a part of the Graim *et al.* study currently in editorial review.

2.1 UCSC Tumor Map: Exploring the molecular similarities of cancer samples on an interactive portal

In this section I present the manuscript for genomic data projection and exploration resource called Tumor Map. We are planning to submit it to the Genome Biology journal in the near future.

2.1.0.1 Publication Title and Author List

Title: UCSC Tumor Map: Exploring the molecular similarities of cancer samples on an interactive portal

Authors: Newton, Yulia [1], Novak, Adam M. [1], Swatloski, Teresa [1], McColl, Duncan C. [1], Chopra, Sahil [1,3], Graim, Kiley [1], Weinstein, Alana S. [1], Baertsch, Robert [1], Salama, Sofie R. [1], Ellrott, Kyle [1,2], Chopra, Manu [1,4], Goldstein, Theodore C. [1,5], Haussler, David [1], Morozova, Olena [1] & Stuart, Joshua M. [1]

[1] Biomolecular Engineering and Bioinformatics, University of California, Santa Cruz

[2] Oregon Health & Science University

[3] Stanford University

[4] Pacific Collegiate School

[5] Hematology-oncology Department, University of California, San Francisco

2.1.1 Abstract

While vast amounts of omics data are being collected on tumor samples at an accelerating rate, few resources exist for biologists to readily identify important trends in these data to find connections within and between cancer subtypes. Intuitive browsing interfaces that organize samples based on their molecular similarities to aid pattern discovery are lacking. In response to this demand, we created the Tumor Map portal that provides intuitive global overviews and statistical analyses of sample subtypes.

Samples are arranged on a hexagonal grid based on their similarity to one another and rendered with Googles Map technology. Maps can be made based on any high-throughput platform from which such similarities can be derived. When applied to the TCGA Pan-Cancer-12 [59] dataset, the Tumor Map recapitulates established subtypes as distinct areas of the map that can be spotted by eye including those for breast, endometrial, and bladder cancers. A map created with all of the TCGA data platforms reveals a previously undescribed subtype made of various tumors from diverse tissues that exhibit signatures of immune cell invasion. Thus, exploring cross-cancer specimens using the map metaphor shows promise for generating hypotheses that could inform treatment decisions and generalizes to the application of biospecimens beyond cancer genomics datasets.

2.1.2 Background

Genomic aberrations such as mutations that accumulate in a particular cells DNA, together with the tissue microenvironment, contribute to the initiation and progression of malignancies. The Cancer Genome Atlas (TCGA) and similar projects have catalogued the molecular changes in thousands of tumor samples of various cancer types using different data modalities including genomic, transcriptomic, proteomic, and epigenomic information, collectively referred to as omics data. The availability of these datasets facilitates intra-tumor and cross-tumor (pan-cancer) comparisons. The goal of pan-cancer analysis is to find similarities across cancers originating in different tissues and to reveal clinically and prognostically relevant subtypes that share common

molecular driver events and pathway aberrations.

The visualization of genomic datasets greatly aids in the identification of patterns that inspire hypothesis generation [126, 44, 82, 19, 24, 102, 96, 73]. This is especially true in cancer genomics investigations, where the number of measured features (e.g. 20,000 genes) and samples (e.g. a few thousand) can be large. Tools are needed that enable biologists and clinicians to navigate complex datasets and identify putative associations and new tumor biology without additional expertise in computer programming and statistical inference.

A critical step in the analysis of a cancer cohort is the identification of subtypes – groups of patient samples that share common sets of molecular alterations revealed by available omics data. The presence of such subtypes, and any genomic or clinical features that characterize them, may provide insights into therapeutic avenues for treating patients. Clustered heatmaps are often used to identify samples with common profiles but this and other approaches have serious limitations that undermine their usefulness (see Discussion). Alternative approaches that give biologists an intuitive portal into exploring cancer subtypes are urgently needed.

We present the UCSC Tumor Map (https://tumormap.ucsc.edu), an interactive visualization and analysis portal to explore tumor samples, or more generally any observations, arranged relative to one another based on their molecular similarities, or more generally any feature space. The Google Map API is used to visualize the landscape. Other applications of Google Maps to the analysis of oncological specimens have produced visually effective presentations [75]. In Tumor Map sample attributes, such

as disease histological subtypes, can be identified easily by eye as contiguous regions. Even without the aid of additional computer and programming expertise, users can discover trends and perform enrichment statistics analysis in complex genomics datasets for scientific and therapeutic hypothesis generation. We illustrate the advantages of our approach by the analysis of a large collection of cancer specimens from multiple tissue types.

2.1.3 Results

We applied the Tumor Map to the analysis of the TCGA Pan-Cancer-12 cohort [27]. This cohort contains over five thousand tumor samples, spanning twelve different tumor types. Multiple platforms of data on these samples were collected and have revealed clinically relevant subtypes. Maps were created for each individual platform as well as integrated maps that combine two or more types of data (e.g. PARADIGM, or 6-way integrated similarities Figure 2.1). This resulted in six maps for individual platforms based on mRNA-Seq, miRNA-Seq, reverse-phase protein arrays RPPA, DNA methylation, somatic copy number alterations (SCNA), and somatic single nucleotide variants (see Figure 2.2). We also loaded 4,177 attributes that describe phenotypic and outcome related information about the samples and patients (e.g. tissue of origin, tumor stage, histology, etc; see Methods). On most of the single platform maps, as well as the integrated maps, the organization of the tumors mirror their tissue of origin and histological type as previously documented by the TCGA consortium [59]. However, intriguing cross-tumor-type relationships also are revealed by each data modality.

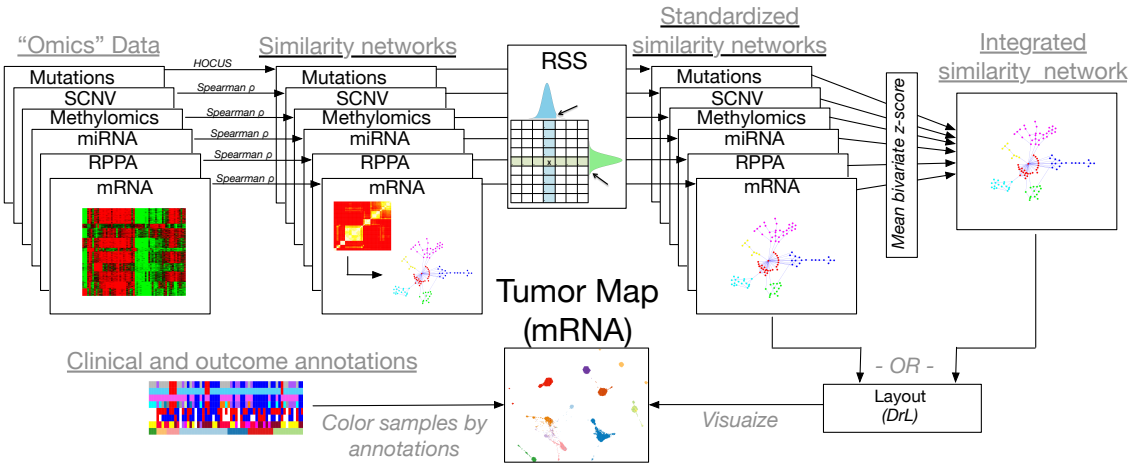


Figure 2.1: Outline of Tumor Map construction. Individual molecular platforms (Omics Data) are provided as input from which pairwise similarities between samples are calculated to produce Similarity networks and standardized using the RSS (RSS; see 2.2) to create a coherent space of standardized similarity networks. Map layouts are created with OpenOrd algorithm using coherent sample networks. Integrated multi-platform maps are created from several coherent networks, combined before input to OpenOrd layout procedure. Shown is an mRNA-based map; colors represent tissue of origin. Attributes – clinical, molecular, phenotype, or outcome metadata – annotate samples using colors and color gradients based on groupings that can be defined by the user.

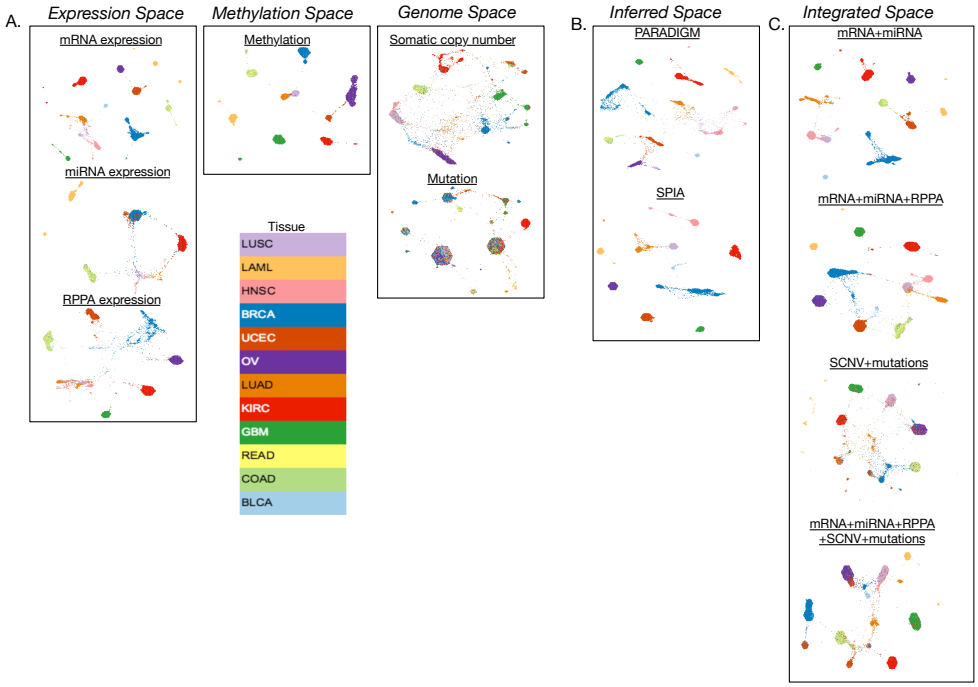


Figure 2.2: Maps produced from different molecular data types. Samples in the maps are colored by tissue of origin. (A) Maps produced from each of the six molecular data types. (B) Maps produced from inferred gene activities using the PARADIGM and SPIA methods. (C) Maps integrating more than one molecular data type.

To illustrate the ease with which the Tumor Map reveals biologically relevant subtypes, we investigated several positive controls and found the visualization clearly depicted expected distinctions in the dataset. To reveal clinical and molecular properties shared among tumors that are placed near one another in the map, attributes for the samples are scored according to their degree of clustering on the map using an Attribute Enrichment Analysis (AEA; see Methods). For example, in the mRNA-Seq-based map, AEA reveals that the separation of basal and luminal [59] breast carcinoma (BRCA) samples is predominantly driven by differences in the estrogen signaling

pathway as expected (Figure 2.3A-B). Within BRCA tumors, HER2-amplified samples (HER2+) cluster closer to luminals compared to basals (see also Figure 2.3A), supported by overlapping distinctive gene sets for each subtype (Figure 2.4), which may support a luminal origin bias for HER2+ tumors. Also for BRCA tumors, mutual exclusivity of samples harboring either PIK3CA or TP53 mutations is readily visible (Figure 2.3C). Carcinomas that arise in the colon (COAD) or rectum (READ) are indistinguishable in the transcriptomic map, consistent with previous reports [59]. Further, among the COAD/READ tumors, the genomically stable tumors, thought to arise from MLH1 hypermethylation, are spatially separated from genomically unstable COAD/READ tumors (Figure 2.3D). AEA of these two regions suggests that unstable tumors have higher HIF1A/ARNT complex activity, indicating hypoxic and/or metabolic differences exist between hypermutated and non-hypermutated colorectal subtypes. The mRNA-Seq space also recapitulates the three major subtypes of bladder carcinomas (BLCA) (Figure 2.3E) identified by Hoadley *et al.* [59]. Moreover, the map shows HIF1A activation in kidney renal clear cell carcinoma (KIRC) tumors with VHL1 mutations (Figure 2.3G) as well as a deficiency of MSH2 (Figure 2.32F), a major component of the DNA mismatch repair pathway whose downregulation has previously been described in KIRC tumors [41].

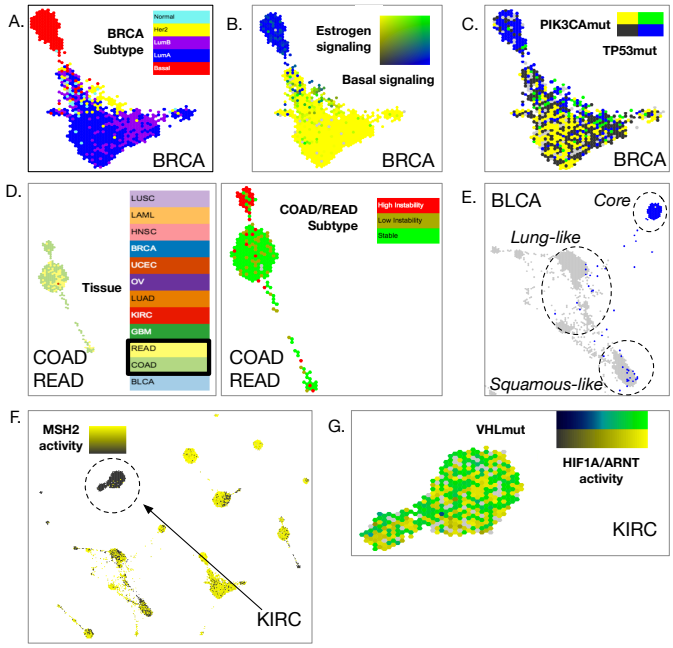


Figure 2.3: Known biology recapitulated by the mRNA expression map. (A) BRCA molecular subtypes and their layout in the map. The map shows that tumors of the same molecular subtype tend to cluster together, with very little mixing of those subtypes, indicating that those subtypes are indeed molecularly different. (B) Estrogen signaling (yellow) and basal signaling (blue) are the top differentiating programs between the basal and non-basal BRCA tumors. Very few tumors exhibit signals of both programs (green). The group of samples labeled as Basal subtype in part A predominantly exhibits the basal signaling program and the group of samples consisting of LumA (for luminal A) and LumB (for luminal B) in part A predominantly exhibits the estrogen signaling program. (C) Mutual exclusivity of PIK3CA (yellow) and TP53 (blue) mutation events in BRCA samples. Most samples carry mutations in only one of those genes. Very few samples show mutations in both of the genes (green), illustrating the documented phenomenon of mutual exclusivity of PIK3CA and TP53 mutations. (D) COAD and READ tumors cluster together (left) and the map separates genomically stable and unstable tumors (right). (E) BLCA tumors separate into three previously discovered molecular subtypes. These subtypes are BLCA-core, BLCA-lung-like, and BLCA-squamous-like. (F) KIRC tumors are deficient in MSH2 (activity level indicated by the intensity of yellow), a component of the DNA mismatch repair pathway. This is a known characterization of KIRC tumors. (G) Co-occurrence of VHL mutations (green) and high HIF1A activity (indicated by the intensity of yellow) in KIRC tumors. Each sample in the map is colored by two colors. Samples colored in yellow indicate an absence of VHL mutation but high HIF1A activity. Samples colored in green indicate both the presence of VHL mutation and high HIF1A activity. No samples show a presence of VHL mutation and low HIF1A activity (such samples would be colored in blue).

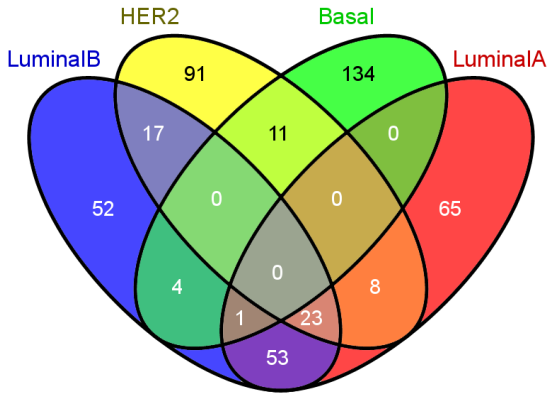


Figure 2.4: Tumor Map reflection analysis shows the overlap in gene signatures distinctive of the Tumor Maps relative positioning of BRCA molecular subtypes. HER2+ tumors share more in common between luminal than basal tumors. A reflection analysis was performed for each of the molecular subtypes; the subtype groupings are defined by BRCA sample annotations. The reflection analysis resulted in 150-gene signatures for each subtype. Venn-diagram of these gene signatures shows the overlap between gene sets for each of the subtype.

Next, we investigated sample groupings revealed by an integration of all of the TCGA omics platforms. To derive a consensus overview, we created an integrated map for the Pan-Cancer-12 dataset using a novel method to standardize similarity spaces (see Methods). We combined six different platforms, each representing a distinct feature type, that included mRNA transcription, miRNA transcription, protein expression, methylation levels, somatic copy number changes, and somatic single nucleotide variants (Figure 2.5A, Figure 2.2C, bottom). All platforms contributions to the integration were treated equally also recapitulated many known connections between the tumor samples. Squamous-like characteristics of basal BRCA tumors, reported by TCGA [105] and oth-

ers, are easily detected in the integrated map (Figure 2.5A-i). The transcriptional data alone fails to reveal this relationship possibly due to the strong tissue-of-origin signal that multiple data modalities tease apart. We also found that the integrated map separates favorable from poor cytogenetic risk groups in acute myeloid leukemia (LAML), which are characterized by differential survival outcomes with statistical significance (Figure 2.5A-ii). The integrated map also revealed that the uterine corpus endometrial carcinoma (UCEC) tumors separate into three major molecular subtypes (Figure 2.5A-iii). While some of the individual UCEC tumors scatter around the map, 126 of them cluster near luminal BRCA tumors, 177 cluster near COAD tumors, and another 171 cluster near ovarian serous cystadenocarcinoma (OV) tumors. The majority of the luminal BRCA-like and COAD-like UCEC samples are endometrioid tumors, while most of the OV-like UCEC tumors are serous. The OV-like tumors are further characterized by mutations in TP53. When comparing these two endometrioid clusters, the top distinguishing feature is a chromosome 1q arm amplification, a known marker in some endometrial tumors [74] and is a known poor prognostic marker in some other cancers [14]. These two major endometrioid subtypes were not found by previous Pan-Cancer-12 analysis.

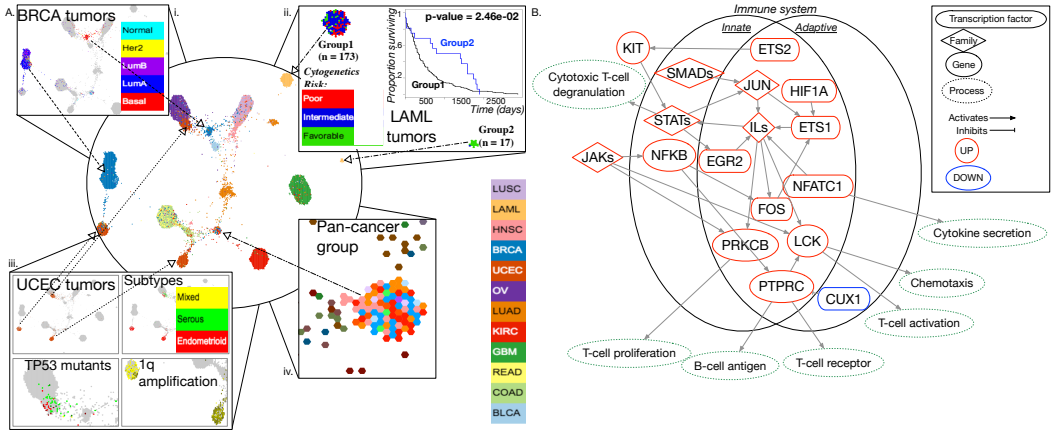


Figure 2.5: Tumor Map rendering of Pan-Cancer-12, an integrated cross-cancer Tumor Map based on six molecular data platforms. (A) Several groups of interest are shown including: (i) BRCA tumors cluster into two major groups, with basal samples grouping with squamous tumors. (ii) LAML tumors separate into two major groups with one group significantly enriched for favorable cytogenetic risk. (iii) Separation of endometrioid UCEC tumors into two major groups, one of which is characterized by a 1q chromosome amplification event. (iv) An integrated pan-cancer cluster, defined by tumors from nine different tissues of origin, exhibits a strong immune signature. (B). Pathway representation of immune signaling-related genes characterizing the integrated pan-cancer cluster showed in A, including markers of both the innate and adaptive immune systems.

Importantly, while the integrated map revealed many of the same connections previously found by the TCGA study of this dataset, it further suggested molecular subtypes that were undetected by single-platform analysis and previous integration strategies. The integrated map revealed an immune-related cross-tumor subtype (n=75), consisting of samples from nine different tissues of origin (Figure 2A-iv, Additional Results). The distribution of tissues among these samples represents different ratios of tumor types compared to the entire cohort (Figure 2.6), suggesting that this grouping of samples did not occur by chance ($P < 1.518e\text{-}10$). Evidence suggests that several

major regulators of both the adaptive and innate immune response exhibit differential activities in these samples as compared to samples outside of this cluster (see Additional Results; Figures 2.7, 2.8, 2.9).

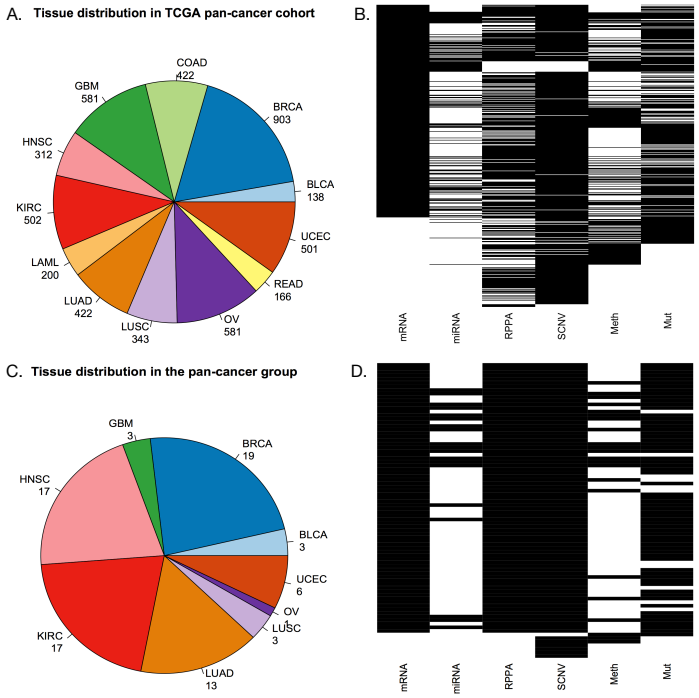


Figure 2.6: Tissue and molecular subtype distribution among the samples in the entire cohort (A-B), and in the pan-cancer cluster (C-D). The pie charts represent the number of samples from each tissue of origin in the entire cohort (A) and the integrated pan-cancer cluster (C). Black and white matrices illustrate the presence of molecular features of each platform (x-axis) across samples (y-axis), in the entire cohort (B) or in the integrated pan-cancer cluster (D). Data available for this sample for a given platform is marked black, otherwise the entry is white.

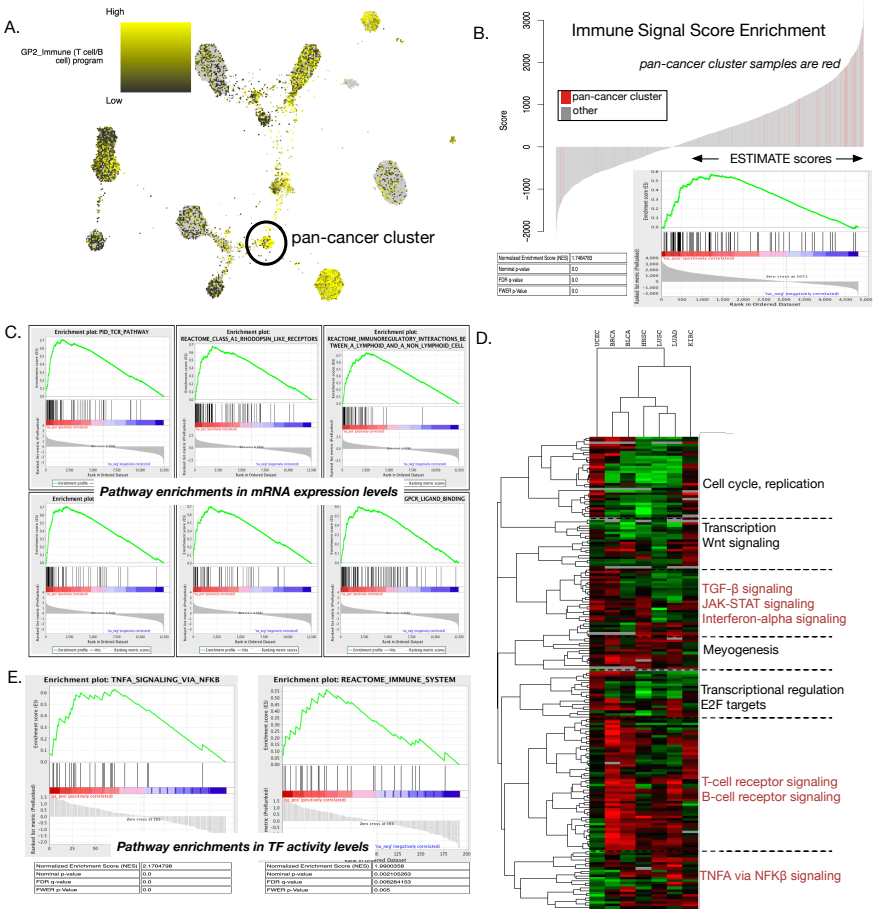


Figure 2.7: Enrichment of Immune signaling in the integrated pan-cancer cluster. Different level of evidence support the association of the integrated pan-cancer cluster with an immune phenotype. (A) Enrichment of T- and B-cell signaling shown on the integrated map, yellow gradient. (B) Enrichment of high ESTIMATE scores in the pan-cancer samples, waterfall plot with pan-cancer cluster samples in red. (C) Enrichment of the immune-related pathways identified by differential expression analysis. (D) Unsupervised analysis of master regulator activities inferred by the MARINa method. Gene clusters enriched for T- and B-cell signaling, interferon signaling, and TNFA via NFKB signaling, red font. (E) The top enriched pathways based on the output of master regulator scores derived with MARINa are immune-related.

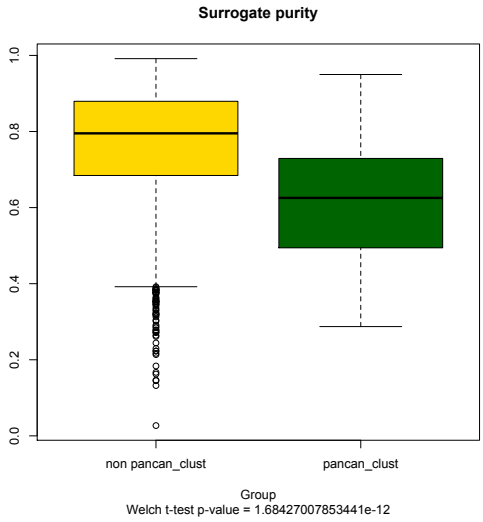


Figure 2.8: Purity estimates in the integrated pan-cancer cluster compared to all other samples in the cohort. The pan-cancer cluster (green box) shows lower purity when compared to the whole cohort as a background (yellow box). This finding is consistent with other analyses indicating high immune signaling in the pan-cancer cluster.

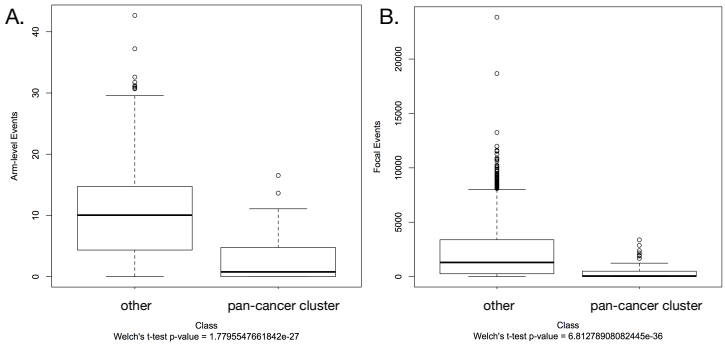


Figure 2.9: Copy number events in the integrated pan-cancer cluster compared to other samples in the full cohort. The pan-cancer cluster shows a lower number of copy number events in both arm-level events and focal events. (A) Arm-level events (pan-cancer group on the right, background cohort on the left). (B) Focal events (pan-cancer group on the right, background cohort on the left).

2.1.3.1 Supplemental Results

Integrated Map

Additional Integrated Mapl Pan-cancer Cluster Results

The analysis of the integrated Pan-Cancer-12 map, which incorporates 6 different data modalities, revealed a tight cluster of samples spanning 9 different types of tissues. We analyzed this group of samples further to identify what distinguishes them from other samples in the Pan-Cancer-12 cohort.

These samples are characterized by elevated T-cell and B-cell immune gene programs (Figure 2.7A) as well as enriched for ESTIMATE [140] immune signaling scores (Figure 2.7B). These samples also show lower tumor purity estimates compared to other samples, supporting the notion of a higher immune infiltrate in the biopsy specimens (Figure 2.8). We investigated this subtype further by computing differential expression between samples inside versus outside of the group while controlling for the tissue composition (see Methods). We found that the most enriched functions were T-cell receptor and interferon pathways (Figure 2.7C). We also used the MARINa [15] algorithm, which implicates transcription factors responsible for observed expression changes, to find that the cluster is enriched in T-cell, B-cell, and interferon signaling (Figure **??**).This group of samples is also characterized by a lower number of somatic copy number alteration events, both arm-level and focal (Figure 2.9), possibly indicating a difficulty in detecting these events due to lower tumor purity. The gene network in Figure 2.5B illustrates regulatory pathways of the immune response in this group of

samples (see Methods).

Additional mRNA Expression Map Tumor-specific Results

Tumor Map reveals that cancer clustering topology differs depending on the omic data type. For example, tumors that are similar in transcriptomic space may form different groupings [59] in methylation space. As was shown in the Pan-Cancer12 analysis, tissue of origin strongly correlates with tumor diagnosis, prognosis, and clinical implications. Tumor Map shows that tissue signal drives transcriptome and proteome expression, as well as autosomal methylation profiles, and is the biggest discriminator of the tumor groupings in these omic spaces (Figure 2.2A, left and middle). Furthermore, expression drives both the inferred (Figure 2.2B) and the integrated molecular subtypes (Supplemental Figure 1C). In contrast, the samples do not separate by tissue as clearly when considering genome structure (CNV and mutations) (Figure 2.2A, right). This suggests that gene and protein expression profiles as well as methylation, which regulates gene expression, are the most influential in defining cell phenotype and for diagnostic classifications.

Breast Invasive Carcinoma

Many previously discovered and known molecular relationships are recapitulated by Tumor Map. For example, samples of breast invasive carcinoma (BRCA) clearly separate into groupings driven by PAM50 [99] molecular subtypes (Figure 2.3A). A basal phenotype discriminates between the two major groups of BRCA samples. However, there are clear clusterings of HER2+, Luminal A and Luminal B subtypes within the non-basal island. These map placements are supported by our previous knowledge about

molecular features of these subtypes. Many basal tumors are triple negative (ER-, PR-, HER2-) and exhibit very different properties compared to other breast cancer subtypes. The HER2+ subtype contains an amplification and hence overexpression of the HER2 receptor, while Luminals express ER and PR. The spatial placing of Luminal B next to HER2+ subtype supports our intuition about the relationships of these tumors [31] (Figure 2.4). Further analysis using the Tumor Map differential statistics tool, shows that the top two differentiating attributes between the two major BRCA clusters are the estrogen signaling program and basal signaling program (Figure 2.2B). These two differential programs also explain why some of the basal samples track with the luminal positioning on the map. It turns out that those samples have high estrogen signaling and resemble luminal samples in expression space.

While TP53 mutations are prevalent in basal BRCA tumors, they appear in some non-basal tumors as well. PIK3CA mutations are also prevalent in BRCA tumors, but they are most commonly seen in luminal tumors. The majority of the BRCA tumors have one of these two genes mutated. However, both of these genes are rarely mutated in the same patient [24]. This is because these two mutations exhibit a phenomena of mutual exclusivity - mutation in only one of them is required to achieve disruption of certain molecular pathways. Such mutual exclusivity is easily seen in the Tumor Map (Figure 2.3C).

Colorectal Cancer

It has been previously shown that rectum adenocarcinoma (READ) and colon adenocarcinoma (COAD) tumors are very similar in gene expression space [59]. Tu-

mor Map shows these tumor types mixing in mRNA expression space (Figure 2.3D, left). Furthermore, the Tumor Map groupings are consistent with two major molecular subtypes of READ/COAD tumors, genomicaly namely stable and unstable tumors [59] (Figure 2.3D, right). Highly unstable tumors group together and are characterized by high activity of the HIF1A/ARNT complex (data not shown), which is known to associate with poor survival [16]. Poor survivors also have high Metallothionein 2A (MT2A) gene activity (data not shown), which has previously been associated with other epithelial cancers, such as breast invasive carcinoma and prostate adenocarcinoma. The low-instability and stable tumors do not separate on the map, indicating that low-instability tumors are molecularly more similar to stable tumors than unstable tumors.

Bladder Carcinoma

Tumor Map reveals three main areas in which bladder carcinomas (BLCA) group together based on patterns of expression in the mRNA-Seq data (Figure 2.3E). These three groupings were initially identified by Hoadley et al. [59]. The largest of these groups is composed almost entirely of samples from bladder tissues. The other two groups are those samples clustering with LUAD tumors (BLCA-adeno-like) and those clustering with HNSC and LUSC tumors (BLCA-squamous-like). These results are consistent with the TCGA analysis that also found three main integrated tumor types [59], much like the results from mRNA-only subtyping. The BLCA-adeno-like and BLCA-squamous-like tumors were shown to be associated with poorer survival outcomes than those in the BLCA-only cluster [59].

The distinction between these three major BLCA subtypes can be investigated quickly using Tumor Map and its associated metadata. For example, it is natural to ask what genomic differences (e.g. mutations, copy number changes, etc.) exist between the three groups. Some previously discovered associations are revealed by Tumor Map. Mutations in the super-enhancer EP300 on chromosome arm 3p and in several chromatin remodelers such as ARID1A and MLL3 are known to be differential within the BLCA molecular subtypes [59]. We hypothesize that these tumors arise from squamous cells making up the lining of the bladder, and are thus clinically and therapeutically distinct from other bladder tumors. Additional associations are easily found by the differential statistic tool in Tumor Map (see section 2.1.4.3). As expected, the squamous cell differentiation expression program (data not shown) is one of the top discriminators between the core BLCA subtype and the squamous subtype. The top differentiator is activity of CDK6 as inferred by PARADIGM; CDK6 is an important regulator of the cell cycle [62]. XBP1 and CDC25B activity inferred by PARADIGM was also significantly lower in the lung-like and squamous-like BLCA tumors as compared to the other BLCA tumors. In combination with a higher TP53 mutated program [59] in the lung-like and squamous-like BLCA tumors as compared to most core BLCA tumors, this suggests that the squamous-like BLCA tumors may be more aggressive than the other core BLCA tumors.

Kidney Renal Clear Cell Carcinoma

It has been previously described that the DNA mismatch repair (MMR) pathway is often targeted by cancer [60]. Often this pathway is turned off either by the

accumulation of mutations within it or de-activation via upstream regulators. Mutations within this pathway often exhibit mutual exclusivity in cancers. MSH2 is one of the most commonly compromised genes within the MMR pathway (Figure 2.10). It has been shown that the MMR pathway is deficient in kidney renal clear cell carcinoma (KIRC) tumors [41]. While MSH2 was not included in the high-confidence mutation set for the Pan-Cancer12 dataset, Tumor Map reveals that MSH2 activity as inferred by PARADIGM in KIRC is very low (Figure 2.3F).

KIRC tumors often carry VHL1 mutation [54]. These tumors often also exhibit high HIF1A activity, a consequence of VHL1 mutation. Tumor Map allows for easy visualization of this relationship in the KIRC tumors (Figure 2.3G).

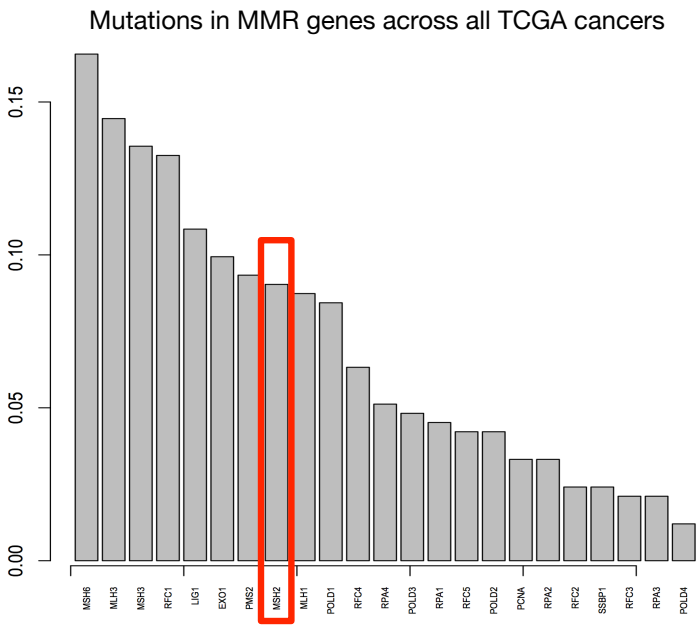


Figure 2.10: Mutation frequencies among the genes that are part of the DNA mismatch repair (MMR) pathway across the whole TCGA cohort. The barplot shows all MMR genes sorted (from left to right) by the frequency of mutations in those genes across all the samples in the TCGA cohort. MSH2 gene is ranked 8th among the 23 MMR genes.

2.1.4 Methods

This section describes the data and the methods used in this paper.

2.1.4.1 Datasets

We obtained a previously published, pre-processed, and normalized dataset made available by the Cancer Genome Atlas (TCGA) consortium that is referred to as the Pan-Cancer-12 [28] dataset. The dataset consists of mRNA expression data for 3,934 samples, miRNA expression data for 1,397 samples, RPPA probe levels for 3,467 samples, copy number variation (CNV) data for 4,692 samples, and Illumina In-

finium HM27 platform DNA methylation data for 2,223 samples. Each of these datasets were treated as independent observations on the samples. A particular samples vector from one of these datasets was considered to be a feature from which similarities to other samples could be computed (see below). For single-platform maps (Figure 2.2A), the vectors of data were used directly as the features in the similarity computation: the mRNA features had 12,471 gene expression levels; the CNV features had 20,287 gene-level summarized copy number estimates from GISTIC2; the miRNA features had 1,070 expression levels for non-coding genes; and the RPPA features had 130 proteins and specific modified protein species. Mutation features were constructed from the 313 high-confidence mutation calls defined by two separate TCGA studies [125, 70]. The methylation features were computed by mapping methylation probes to their nearest genes, aggregating duplicate genes by computing mean methylation levels, and extracting autosomal genes. As a result, the methylation feature vectors had 11,799 autosomal gene features. In this way, all data except for miRNA, was converted to gene-centric estimates and the official HGNC HUGO gene symbol was used as the unique identifier to connect the data across the platforms.

We next describe features that were derived from the individual platform features. These were also used separately to compute sample-sample similarities and from which new maps were created.

PARADIGM Features

PARADIGM [132] is a method that infers gene activities based on gene expression and copy number data that factors in a genes context in a background path-

way. It considers the activities of the network neighbors to make these inferences. The TCGA consortium provided PARADIGM inferences for all of the Pan-Cancer-12 samples. PARADIGM outputs inferred pathway levels (IPLs) as the activities of molecular entities, which include protein coding genes (e.g. TP53), complexes (e.g. pyruvate dehydrogenase complex), protein families (e.g. GCPR family), and abstract processes (e.g. apoptosis). All of the IPLs were used in the calculation of sample-sample correlations.

SPIA Features

Much like PARADIGM, the Signaling Pathway Impact Analysis (SPIA) method [127] creates gene-level activities for each sample based on a given input pathway diagram. SPIA uses the gene expression levels observed for a sample and then uses a method, much like Googles PageRank, to propagate information on a directed network to estimate how much influence/impact each gene has accumulated from the sum of expression changes upstream of the gene. Because we are more interested in the degree to which a gene explains changes downstream, we reversed the network edges in the pathway diagram before running SPIA. In this way, we estimate the responsibility of a particular gene on other gene expression changes that lie logically downstream of the genes influence. Genes with large values have the highest impact and can be thought of as pathway hubs or master regulators. We ran the SPIA algorithm on each sample represented in the Pan-Cancer-12 dataset for which gene expression data was available.

HOCUS Features for Mutation Maps

We used the HOCUS method (Graim et al. 2016, in review) to derive sample-sample similarities based on mutation data. HOCUS uses a community detection scheme

to output 1st, 2nd, and higher-order similarities based on the primary data. It first constructs 1st order features from all pairwise Hamming distances of one sample to all other samples; it then creates new 2nd order features from the similarities of the 1st order features, etc.

2.1.4.2 Map Creation To Reveal Molecularly Similar Sample Groups

Computing Sample Similarities with a Reciprocal Significance of Similarities (RSS)

Maps with different combinations of omics features were created as well as one that combined all six of the omics datasets as mentioned in the main text (Figure 2.2C). We applied RSS (see Section 2.2 for details) method to compute integrated multi-platform Tumor Map maps.

Layout Rendering and Visualization

To allow for easy viewing and exploration of related samples, the Tumor Map renders a 2D hexagonal packing of the samples based on their pairwise similarities. This is accomplished in two steps: first, a preliminary projection of the samples in the X-Y plane is found; then, the X-Y locations are snapped to a hexagonal grid. Each of these steps is described in more detail below.

As in all multidimensional scaling (MDS) approaches, the 2D solution sought by the Tumor Map should preserve the distances from the original features space (e.g. RNA-Seq expression) in the new, lower dimensional projection: i.e. tumors with high similarity should be near one another while dissimilar ones can be further apart. The

quasi-physics based layout engine OpenOrd (formerly known as DrL) [85], implemented in the igraph R package [51], is used to derive an initial set of X-Y positions for the samples [138]. OpenOrd treats the similarities as spring constants and searches for a configuration among the samples that produces an arrangement to relax the spring tension of the system as much as possible. For computational convenience and because the resulting hexagonal lattice will only allow for 6 neighbors in the ultimate solution, we only provide OpenOrd with a sparse matrix made up of the top six neighbors, and their similarities, to each sample.

In the second step, the OpenOrd X-Y coordinates are snapped to their nearest hexagon to arrange all of the samples on a tiling of regular hexagons.

Each sample is placed in a grid cell, as determined by its OpenOrd-determined position. If the predetermined cell is occupied, the sample is snapped to an empty grid cell within a minimal distance from the original cell. Multiple samples that compete for a location will thus spiral around a central hexagon in the neighbors around the central location. Thus, dense clumps are separated so that they can be viewed on approximately the same scale as the distances that separate them. Hexagons were selected as the shape for the grid cell in order to illustrate that there are no inherently preferred axis-aligned directions in the OpenOrd output.

Google Maps API [4] is then used to load the resulting layout into a browsing environment. The API provides the ability to interactively navigate, zoom, and explore various annotations of locations on the map analogous to Google Maps and Google Earth applications. Layouts based on several of the individual platforms are shown in

Figure 2.2.

Additional Layout Rendering Methods

While OpenOrd method described above is the default method of rendering layout of nodes in the 2-D map used by the Tumor Map tool, It also allows creating visualization based on a number of more conventional methods. I implemented the following additional methods of layout rendering for Tumor Map creation:

1. tSNE [102]
2. MDS [20]
3. PCA [109]
4. ICA [63]
5. Isomap [130]
6. Spectral embedding [118, 95, 81]

Pre-computed Euclidian Plane Coordinates

Sometimes the researcher will already have the coordinates for the Euclidian 2-D plane for the samples in their cohort. These coordinates might come from another analysis and/or visualization tools. For example, if the researcher used their own version of a multi-dimensional projection or clustering method, they used one of the more conventional tools but with a different distance/similarity function, or if they utilized a specialized kernel applicable to their particular research question. Even when a researcher already has a visualization of their cohort, Tumor Map can still offer additional

advantages by providing ability to perform dynamic statistical tests and other analysis described in section 2.1.4.3. In such case, Tumor Map allows producing a visualization from (x, y) coordinates of samples in a 2-D plane. Users are still able to import their own attributes and annotations for these samples.

2.1.4.3 Attributes for Interpreting Biological Relevance of Sample Groups

Tumor Map provides the ability to view and explore associations between sample groupings and clinical, molecular, and phenotypic annotations. We refer to these annotations as attributes in this manuscript. Attributes include clinical annotations (e.g. tumor stage), molecular subtypes (e.g. breast cancer PAM50 subtypes), prognostic and survival indicators, and genomic alteration flags (e.g. TP53 mutation).

Attribute Sources

Here we describe the attributes pre-loaded and available to any user in the Pan-Cancer-12 map. The genomic alterations annotations include 313 high-confidence mutations as well as 2,986 gene amplifications and deletions. Attributes additionally include inferred annotations, such as per-sample transcription factor activities summarized from PARADIGM results, for 774 transcriptional regulators and per-sample drug program scores [59]. Finally, the attributes include basic patient information, such as age, height, and weight. We also provide a pre-loaded set of pathway annotations. These attributes are provided for gene features and indicate a genes membership in pathway sets.

User-Defined Sample Groups

One can define custom groups of map entities from the layout-based clusters and groupings by using one of the Select tools available in the Tumor Map application. This operation creates a custom user-defined binary attribute, where the selected samples have a value of 1 and all other entities have a value of 0.

Set Manipulations

Any of the pre-loaded or user-defined attributes can be manipulated into new attributes through the use of set manipulation tools provided in the Tumor Map application. Attributes created via set manipulations can be further subjected to additional set manipulation operations to create increasingly complex attributes.

Attribute Density

One of the powerful features of Tumor Map is its ability to associate attributes with the topology of the map. Here we describe a method for computing associations between every pre-loaded attribute and sample groupings in the map layout based on how densely the attribute is distributed within those groups. This can provide insights into the biological significance of the attributes and their relationships to molecular profiles (i.e. if samples with similar attribute values cluster together in map layout). For example, perhaps samples containing a TP53 mutation cluster in different parts of the map, but each cluster of TP53 mutants can be characterized as tight rather than spread out. The density of the TP53 mutation attribute is indicative of a similar molecular phenotypes these mutants exhibit.

We divide the map into a 25 x 25 grid. We survey each pre-loaded attribute in the map and compute how significant the density of that attribute is in each of the 625

grid partitions. We scan through each partition, from left to right first and then from top to bottom, and compare the density of the attribute in the partition as compared to the density of the same attribute outside the partition.

An appropriate statistical test is performed depending on the type of the attribute variable (summarized in Figure 2.11 Density Statistic section). A p-value is computed for each of the 625 partitions of the 25 x 25 grid. The best p-value out of the 625 p-values is selected to represent the density significance for a given attribute. We perform the Benjamini & Hochberg [18] correction of all p-values.

Density Statistic:

Variable	Binary	Categorical	Continuous
Density test	Binomial	Chi-Square	Mann-Witney U-test

Layout-independent Statistic:

Attribute type	Binary	Categorical	Continuous
Binary	Fisher Exact Test	Chi Square test	Welch's t-test
Categorical		Chi Square test	Kruskal-Wallis
Continuous			Pearson Rho

Differential Statistic:

Attribute type	Binary	Categorical	Continuous
Statistical test	Fisher Exact Test	Chi Square test	Welch's t-test

Figure 2.11: Statistical tests computed by different attribute enrichment analysis (AEA) tools available in the Tumor Map.

Attribute Enrichment Analysis (AEA) to Uncover Statistical Trends of Sample Groups

One of the more powerful features of the Tumor Map is the ability to identify attributes that distinguish one group of samples from another as revealed by the layout. The differential presence or absence of a finding in samples of one group compared to another may reveal important biology that can be leveraged for interpreting subtypes (e.g. over-representation of EGFR amplifications in a subset of lung cancers). By

converting sample groups into attributes themselves (e.g. a binary attribute that defines if a sample is either in or out of a specified group) this kind of differential analysis can be performed by comparing pairs of attributes. Attribute Enrichment Analysis (AEA) performs this comparison using an appropriate statistical test that is chosen based on the type of each attribute (binary, discrete, or continuous). In addition to the usual pairwise enrichment tests, we introduce a new association test included in AEA that uses information about how samples are arranged in the layout. The former approaches we refer to as layout independent while the latter are layout dependent and these are described in more detail below.

User defined attributes, such as those created after defining visual sample groups, are queried against a background database of pre-loaded as well as any previously defined user attributes. Benjamini & Hochberg [18] adjustments are used to correct any of the resulting p-values for multiple testing since the number of tested associations can be quite large.

Layout-Independent Associations

Layout-independent associations can be discovered outside of the Tumor Map tool by surveying pairs of attributes and performing appropriate statistical tests on the groups of samples annotated by those attributes. However, Tumor Map provides an easy on-the-fly way to query these associations.

Given a query attribute, Tumor Map scans through all other pre-loaded attributes and computes an appropriate statistical test, depending on the data types of both the query and the reference attribute (summarized in Figure table2). If both at-

tributes are binary, then a Fishers exact test is performed. If one attribute is binary and the other attribute is categorical, then a Chi Square test is performed. If one attribute is binary and the other attribute is continuous, then a Welchs t-test is performed. If one attribute is categorical and the other attribute is continuous, then a Kruskal-Wallis test is performed. Finally, if both attributes are continuous, then a Pearson Rho test is performed.

Special cases of binary vs. binary, binary vs. categorical, and binary vs. continuous attribute associations can be considered when the dichotomy of the binary query attribute is based not on values of 1 or 0, as seen in the pre-loaded binary attributes, but rather on the value 1 in two different binary attributes, pre-loaded or user-defined (see User-Defined Sample Groups). These associations can be thought of as based on the differential of the first two binary attributes.

Layout-Aware Associations

Layout-aware attribute associations are one of the most powerful features of the Tumor Map tool. The discovery of these associations is driven by the map layout and, therefore, is impossible outside of the Tumor Map tool.

Samples cluster in the map based on their molecular features. Often these groupings contain samples with molecularly equivalent genomic alterations or exhibit patterns of mutual exclusivity. Mutual exclusivity of different genomic alterations may lead to similar gene expression profiles. This is often seen in mutations in the DNA mismatch repair pathway where only a single mutation is required to disrupt the function of the whole pathway. Such associations between mutations can be easily detected with

the Tumor Map tool.

We compute a layout-aware statistic by considering only pairs of binary attributes A and B, and employ an adaptive gridding method to divide the map surface into partitions. The specific statistical test depends on the types of variables that represent attributes A and B (summarized in Figure 2.11). The adaptive gridding method iteratively divides the map in 4 partitions, and any partition with fewer than 2 samples is discarded. Partitions with more than 25 samples are further divided into 4 partitions. We iterate until every non-discarded partition in the map has between 2 and 25 samples. Within each partition i, we count the number of occurrences of the value 1 in the query attribute (A_i) and the number of occurrences of the value 1 in the contrast attribute (B_i), discarding those counts where both attributes A and B have value 1.

We construct the background distribution of the total samples in the map and each cell of the grid by counting the number of samples in each grid partition that was not originally discarded (C_i). We also keep track of total number of samples (N) considered for this vector C. We normalize every element of vector C by quantity N. We then multiply each value of vector C by 5, a value indicating the weight of the vote given to each sample in the background distribution. This is done to give more weight to those partitions in the grid that have many samples.

We use this background distribution in place of pseudocounts when no occurrences of attributes A or B are found in a particular partition but the partition contains a sufficient number of samples to prevent it from being discarded. We add the elements of vector C to the corresponding elements of vectors A and B. We then compute Pear-

sons r correlation between vectors A and B. The p-value is computed as a two-tailed probability from the normal distribution.

In order to avoid uninformative significant findings, i.e. those comparisons in which too few values participate, we compare the total number of samples with value 1 in the query and the contrast attributes. If the number of samples with value 1 in the contrast attribute is fewer than 5% of the number of samples with value 1 in the query attribute, then we skip the statistical test described above and assign a p-value of 1 to the comparison. In this case, a p-value is assigned only for implementation of attribute sorting.

2.1.4.4 Methods for Analysis of the Integrated Pan-Cancer-12 Map

Pan-Cancer Cluster Differential Expression Analysis

The pan-cancer cluster included 75 samples from nine different tissues of origin/diseases (Figure 2.5A-iv, Table 2.1). To account for an imbalance in the number of samples contributed by each tissue, separate t-tests were performed within each tissue type for each gene (n=8 tissue types used here; ovarian serous cystadenocarcinoma tissue was excluded due to insufficient number of samples). For each gene, we obtained 8 separate t-statistics. We summarized these values into a single statistic per gene by calculating the arithmetic mean, weighted by the inverse variance of the all the tissue-specific t-statistics values. Figure 2.12 describes an outline of the method we employed to compute differential expression.

Tissue	N samples	% tissue cohort
BRCA	19	2.1%
HNSC	17	5.4%
KIRC	17	3.4%
LUAD	13	3.1%
UCEC	6	1.2%
BLCA	3	2.2%
GBM	3	0.5%
LUSC	3	0.9%
OV	1	0.2%

Table 2.1: Summary of samples in the pan-cancer cluster in the integrated Tumor Map.

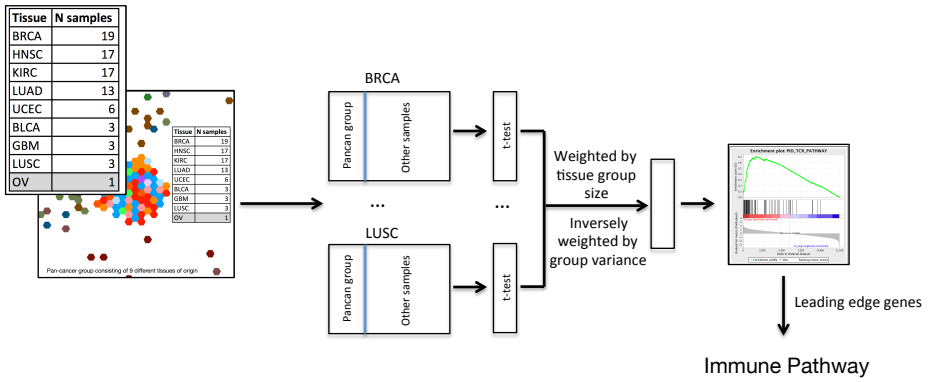


Figure 2.12: Method to compute differential expression for samples in the integrated pan-cancer group vs. other samples in the TCGA cohort. Tissue composition imbalance was corrected for by performing t-tests within each tissue. For each gene, t-statistics were computed within each tumor type separately and then summarized per-gene t-statistics were calculated as an arithmetic mean, weighted by the inverse variance of the all the tissue-specific t-statistics values.

ESTIMATE Analysis of Pan-cancer Cluster

We downloaded ESTIMATE [140] scores for every sample in the TCGA cohort. We ordered these scores from smallest to largest (from left to right) and plotted them as a waterfall plot (sorted barplot) while coloring samples that belong to the pan-cancer cluster in red and all other samples in gray. Figure 2.7B (top) shows the waterfall plot of ESTIMATE scores for the TCGA Pan-Cancer-12 cohort. The plot demonstrates clear enrichment of the pan-cancer cluster samples in the high tail of the plot. To quantify this enrichment we utilized Gene Set Enrichment Analysis (GSEA) [124] method to obtain significance p-value. Instead of using sets of genes, as is usually done with GSEA method, we used sets of samples. We created a custom set that contains only samples that belong to pan-cancer cluster and input ESTIMATE scores into the method. GSEA plot for this analysis and significance estimates are shown in Figure 2.7C (bottom). The pan-cancer cluster is significantly enriched in the high immune scores.

Gene Set Enrichment Analysis of Differential Gene Expression

We applied GSEA to the cross-tissue-summarized differential expression described in the previous section. We utilized the Canonical Pathways set from the MSigDB pathway sets database [78]. Figure 2.7C shows some of the top results from that analysis. The top enriched pathways are immune-related pathways. Class A1 rhodopsin-like receptors are chemokine receptors and compose the largest subfamily of G proteincoupled receptor (GPCR) family. Seeing enrichment of this family of genes along with T-cell receptor and B-cell receptor enrichments supports earlier findings about increased immune signaling in the pan-cancer group.

Pan-Cancer Cluster Master Regulator Analysis

We used the Master Regulator Inference algorithm (MARINa) as described by Aytes et al. [15] to infer candidate master regulators (MRs) driving the phenotype of the pan-cancer cluster based on differential expression of the regulators downstream targets. We ran MARINa separately for seven of the nine cancer types in the pan-cancer cluster, in each case comparing samples in the cluster to those outside the cluster (Figure 2.13) . Ovarian serous cystadenocarcinoma (OV) tumor (n = 1) was excluded from this analysis due to an insufficient number of samples. Glioblastoma multiforme (GBM) tumors were excluded from this analysis due to lack of expression data for the GBM samples in the pan-cancer group (n=3). The following parameters were used to run MARINa: number of random sample/gene permutations for null model computation = 1,000; minimum number of gene targets in the regulon (see below) for a transcriptional regulator to be considered as a potential MR = 25. This gave activity scores scores for putative master regulator transcription factors (TFs) in each tissue type. We computed the overall activity score for each inferred TF by summing the scores across all tissues for that TF. This ranked list of TFs inferred to have increased activity in the pan-cancer cluster was then used for downstream analysis.

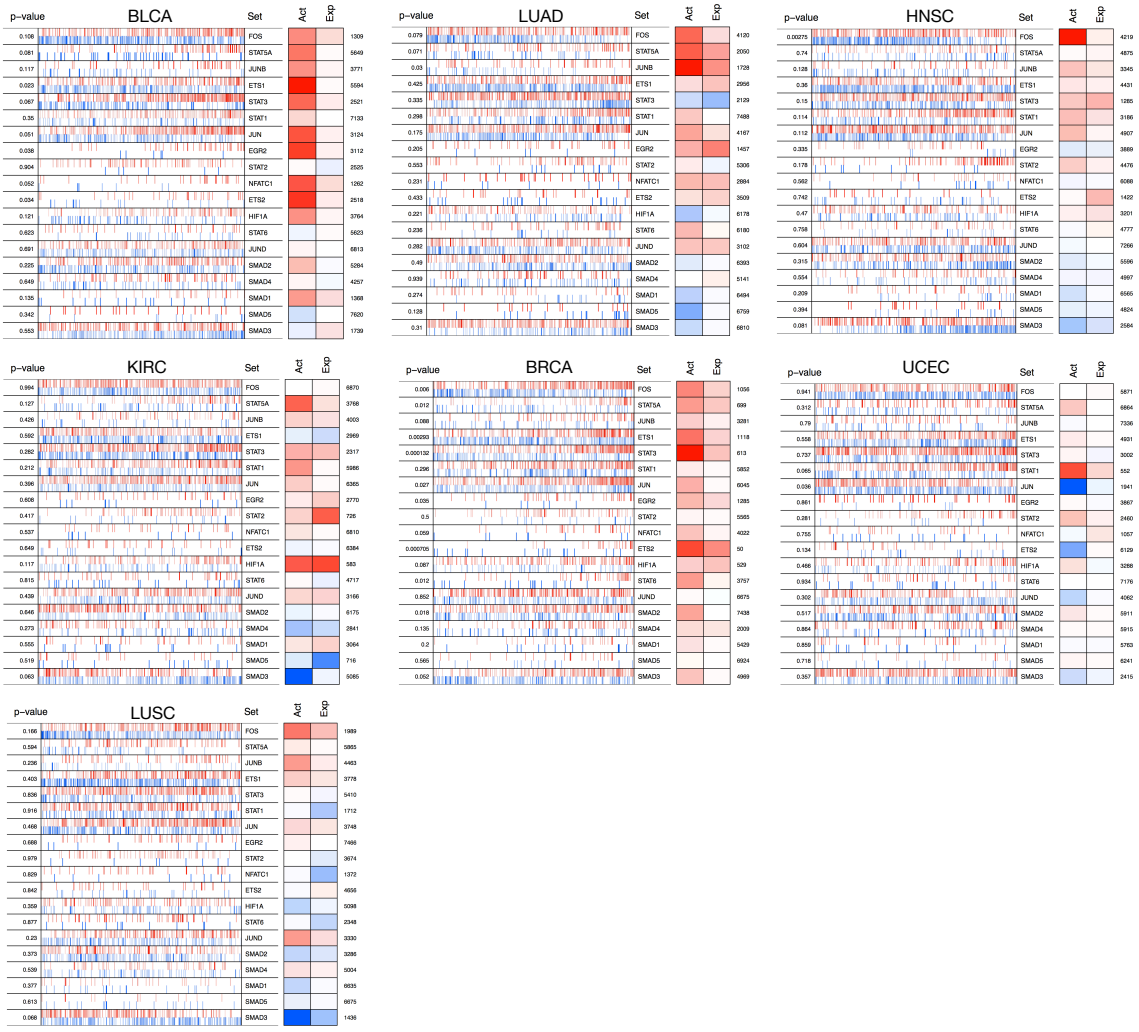


Figure 2.13: Transcription factors inferred by the MARINa method contrasting differential gene expression between the integrated pan-cancer cluster and other samples within each tissue. Specific master regulators vary depending on the tissue of origin but some themes are shared across tumor types. Each matrix shows results for each tissue type. For each transcription factor (rows), the expression levels of each of its targets (tick marks) are colored according to whether the TF is predicted to activate (red) or inactivate (blue) the target. The inferred activity is illustrated to the right of the factor (activated, red; inactivated, blue) along with its expression level and the number of its targets. A P-value is written along the last column.

MARINa requires as input a known regulatory network consisting of tran-

scription factors (TFs) and their targets (the TFs regulon). We created our own network for this purpose by combining pathway information from four sources: the Superpathway [25], the Literome [104], Multinet [72], and ChEA [76].This collection of networks was filtered to include only links corresponding to regulators known to act at the transcriptional level, and further filtered to include only regulators with at least 15 targets. This gave a network of 419 TFs with 61,504 total targets in their regulons. More detailed information about the assembly of this network can be found at *https://github.com/epaull/UCSC_VIPER/blob/master/pathways/README*. For the specific analysis in this manuscript, we further filtered our network so that it included only those TF regulons for which our Pan-Cancer-12 dataset contained expression data. Additionally, we increased from 15 to 25 the minimum number of targets in a TFs regulon for the regulator to be considered as a MR.

Unsupervised Analysis of Master Regulator Activities

We performed unsupervised hierarchical clustering with average linking on the scores inferred by the MARINa method (Figure 2.7D).

Gene Set Enrichment Analysis of Master Regulators from MARINa

As described in the main text, we performed GSEA on the pan-cancer master regulators resulting from the MARINa analysis described above. We utilized the Canonical Pathways set from the MSigDB pathway sets database [78]. We limited the genes in the gene sets to include only transcription factors included in the MARINa input network (see above).

Construction of The Immune Pathway that Distinguishes The Inte-

grated Pan-can Cluster

We summarized the various levels of evidence we collected to support our hypothesis that the pan-cancer cluster exhibits high immune signaling (Figure 2.14). To explain the specific mechanisms of this signaling we constructed an immune network that depicts the immune activity in the pan-cancer cluster.

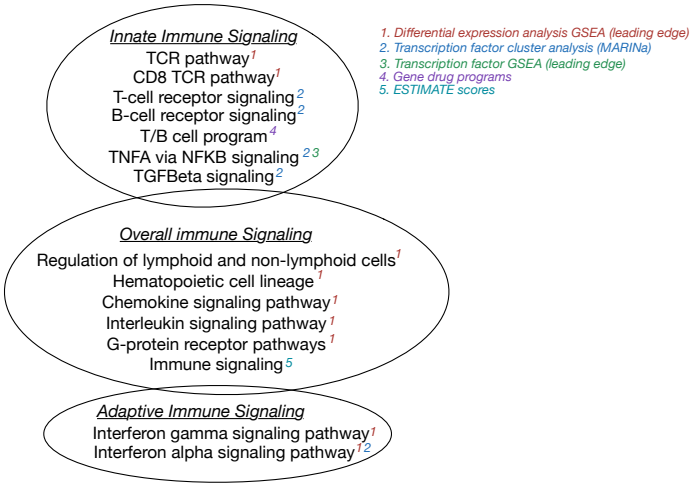


Figure 2.14: Venn diagram representing the innate and adaptive immune systems and different levels of evidence supporting higher activity of each of the components of those systems in the pan-cancer cluster when compared to the rest of the TCGA cohort. We found evidence of both innate and adaptive immune system signaling with a number of different analyses.

We extracted leading edge genes from the immune-related enrichments from GSEA analysis of differential gene expression and MARINa master regulators described above, as well as the transcription factors from the three immune-related clusters described in Figure 2.7D. We combined the extracted genes (n = 151) into a single gene set (Q) and queried the Superpathway [25] for a connected sub-pathway that includes these genes and their nearest neighbors. From this sub-network we excluded any leaves

(terminal vertices) that are complexes, families, or genes that do not belong to set Q. We also removed any non-terminal vertices that are complexes, families, or abstracts by directly connecting the vertices they have an edge to. Figure **??** shows the immune pathways constructed via this procedure. We further simplified this pathway by removing any vertices that did not have extreme high or low differential expression and transcription factor scores as inferred by MARINa. In cases where those vertices connected two or more other nodes, we created a direct connection between those nodes.

2.1.5 Discussion

Finding molecular connections between cancer subtypes will provide a clearer picture of the interplay between cells, tissues, and altered pathways as they contribute to tumorigenesis. Patterns present in many cancer samples may reveal driving genomic aberrations and pathway signatures that aid our understanding of the initiation, progression, and therapeutic options of this disease. The Tumor Map provides a biologist with an interactive browser for exploring molecular commonalities across thousands of samples with an analogous interface to navigating a virtual geographical landscape. Application to the TCGA Pan-Cancer-12 collection of samples suggests this metaphor can successfully survey the major distinctive underpinnings of these samples. Notably, the discovery of a novel, biologically significant subtype, missed by prior analyses, underscores the value of new visualization modalities to further enlighten our understanding of these data. Furthermore, the immune-related pan-cancer subtype suggests that the integration method can aid in identifying samples that are good candidates for im-

munotherapy.

A popular method to identify subtypes is the clustered heatmap, a mainstay of cancer genomic analysis [26, 45, 135]. Asymmetric heatmaps present a matrix of samples (columns) by genes (rows) in which both the columns and rows are clustered to allow an investigator to view patterns of molecular activity. Symmetric heatmaps of samples-by-samples use an organized color gradient to present the pairwise correlations between samples. A heatmap can convey correlations between samples as large bands of columns sharing a similar color trend. While heatmaps are helpful for displaying the molecular patterns shared among sample groups, they are limited for illustrating sample-sample relationships. Although similar samples can be arranged near one another, the placement is restricted to a 1-dimensional axis where distance may not reflect similarity, confounding the eyes ability to identify trends. In addition, samples may be related to multiple groups at different levels of correlation, forcing the eye to scan across non-adjacent columns of the heatmap.

Alternative approaches to the heatmap, such as principal component analysis (PCA) [109] or multidimensional scaling (MDS) [20], have been employed that allow projecting high-dimensional data onto more than a single axis. Several new approaches add to this classic repertoire. One example is GATE [82], which organizes genes into hexagons on a regular lattice according to their mutual co-expression. Overlaying other functional information about the genes (such as their expression levels in a particular sample) enables the identification of co-regulated sets of genes as swathes of similarly-colored hexagons. Another 2-D approach by Kim et al. [73] maps entities (genes or

samples) into a plane where the distance between any these entities is proportional to the dissimilarity between them in the original space. The resulting map creates dense and sparse areas that give the impression of a geographical landscape. Landscapes have been shown to provide an intuitive representation that is much easier to recall than an equivalent 2-D heatmap. Landscape visualizations likely tap aspects of human cognition – our ability to navigate complex terrains without getting lost in the forest.

Other online portals for browsing high-dimensional genomic data exist in an MDS-like solutions. One of such portals is called MEREDITH [128], which provides single-platform and integrated views of 19 TCGA tumors. The integrated view includes mRNA-Seq, miRNA-Seq, DNA methylation, and somatic copy number alterations data. While the portal is publicly available, allowing a user to inspect each sample one at a time, no interactivity is provided, nor the ability to assess the statistical significance with sample groupings of sample attributes such as clinical/demographic (e.g. age, gender, etc.) and genomics (mutations, copy number changes, etc).

Tumor Map is a novel portal for interactive visualization and pattern discovery for the analysis of a variety of genomics datasets, demonstrated here in application to cancer datasets. Tumor samples are depicted on a 2-dimensional grid to combine the strengths of the lattice and landscape approaches. Like MDS, the positioning of samples is guided by feature vector similarities so that nearby samples are like one another. The maps can be constructed from any data type in which molecular features encode observations across a set of samples and from which pairwise similarities between samples can be calculated. In addition, it offers an impressive variety of 4,176

attributes, including clinical, diagnostic, and outcome annotations, genomic aberrations and computationally-derived sample descriptors (see Methods). Using this available metadata on samples, we are able to perform statistical associations to test for the presence of over- or under-represented facts that distinguish a group of samples from the rest or one group from another. Finally, our novel platform integration method enables multiple datasets of different feature types to be combined into a single integrated map, which reveals a number of interesting biologically relevant groupings of samples single platform maps and other integration methods do not find.

2.1.6 Future Direction: Submaps

Often it is insufficient for biologists, clinicians and other researchers to utilize a fixed visualization of the cohort of interest. Once within "a big picture" one may want to examine closer a sub-set of the cohort (e.g. a specific cluster of samples that are laid out together on the map). Tumor Map is a fully dynamic tool that offers such functionality. We call this functionality *Submaps* and it adds an incredible power and ability to mine for informatics in genomic datasets. Once in one of the provided "fixed" maps, the users are able to select a sub-set of nodes in the map (see *User-Defined Sample Groups* in section 2.1.4.3 for details) and build a map that only contains that set of nodes on the fly. This new map has all the attributes for these samples imported into it and users are able to perform dynamic statistical tests and other analysis described in section 2.1.4.3. This functionality sets Tumor Map apart from other visualization tools currently available to the scientific community for scientific research.

2.1.7 Conclusion

The Tumor Map provides an intuitive and interactive map of tumor samples that facilitates the identification of cancer subtypes based on common molecular activities. A toolbox of statistical tests is included that allows researchers to find associations between sample groupings and attributes – clinical, phenotypic, molecular, and outcome annotations (described in Methods). Future versions will make it possible to view user-contributed samples together with publicly available data sets as a backdrop. The portal available at *https://tumormap.ucsc.edu* contributes a new type of integrated genomics browser, which utilizes Google Maps [4] for visualization, that biologists and bioinformaticians can use to richly interrogate cancer genomics data. The approach is easily extendable to applications beyond the comparison of cancers, such as navigation through the landscape of stem and progenitor cells. To facilitate its wider applications and extensions, the code repository is available at *https://github.com/ucscHexmap/hexagram.git*.

2.2 Data Transformations Aid in Molecular Pattern Discovery

As described in the introduction, there are many views that can represent a cell state. Additional data transformations can provide novel views of the data, allow for data views unification, and provide a coherent space for merging data sets (combatting batch effect that arises from combining datasets from multiple sources). This section describes a method I developed called Reciprocal Significance of Similarities

(RSS), which provides such transformation. This method is an adaptation of Context Likelihood Relatedness (CLR) score described by Faith *et al.* [48], with novel ability to make absolutely no assumptions about the pre-transformed space. Our method is very versatile and has many potential uses. Here we describe several applications of RSS method to various experiments.

2.2.1 Reciprocal Significance of Similarities (RSS)

The choice of similarity measure to compare samples can greatly influence their apparent cluster structure and the way they are rendered on a two-dimensional projection. We used a version of the Context Likelihood of Relatedness (CLR) method [30] adapted for sample-sample comparisons. CLR identifies gene-gene interactions by computing the mutual information (MI) between the expression levels of every gene pair. The approach assumes most of the similarities are spurious and thus can be used to form empirical background distributions from the MI values for each gene. Every MI value between genes i and j can then be transformed into a Z-score reflecting the relative significance of the measure on is distribution and on js distribution using the formula $\sqrt{Z_i^2 + Z_j^2}$, where Z_i^2 is the z-score of the MI on is distribution and Z_j^2 is the z-score of the MI on js distribution. Faith et al. [49] showed that CLR can successfully identify validated regulatory interactions in bacteria.

In the same way, we would like to detect pairs of samples of the same cancer subtype from those that are less well-related. We reason that using an approach like CLR would be justified for the case of comparing many samples to each other since one

might safely assume that the vast majority of similarities are spurious, outnumbering those from truly related pairs. We term the approach the Reciprocal Significance of Similarities (RSS) transformation.

In our case, we would like to distinguish positive from negative correlation. Therefore, rather than MI, RSS uses Spearman rank correlation as our non-parametric and signed measure. In addition, if the reciprocal significance measures report conflicting information, such as marginal Z-scores of opposing sign, the resulting RSS Z-scores should reflect this discrepancy by assigning a lower value (near zero) to such pairs.

To this end, let S_d represent the similarity matrix for dataset d, where $S_d(i,j)$ records the similarity of two samples i and j. A reciprocal significance measure $Z_d(i,j)$ is computed between the two samples using the arithmetic mean of the z-scores from each samples marginal distribution:

$$Z_d(i,j) = \frac{1}{2}\left(\frac{S_d(i,j) - m_d(i)}{\sqrt{v_d(i)}} + \frac{S_d(i,j) - m_d(j)}{\sqrt{v_d(j)}}\right) \tag{2.1}$$

Where $m_d(k)$ is the mean and $v_d(k)$ is the variance of the similarities to a particular sample k in dataset d. Of course different metrics can be used in addition to the Spearman rank correlation such as Pearson correlation, or Kendalls Tau (or Jaccard index or Tanimoto similarities [64] for dichotomous features).

An advantage of RSS is that it enables the straightforward integration of different omics platforms. Z-scores computed from each platform separately can be averaged together because the original similarities are transformed into a common scale. One

could consider matching the distributions of Z-scores from each different dataset (e.g. using quantile normalization) before adding separate Z-scores together. However, in this work we chose to use the direct Z-scores as we observed the distributions to be comparable. The individual dataset-specific RSS Z-scores for each pair were combined across different data platforms using the formula:

$$Z^*(i,j) = \frac{\sum_{d=1}^{D} I(d,i)I(d,j)Z_d(i,j)}{\sum_{d=1}^{D} I(d,i)I(d,j)} \tag{2.2}$$

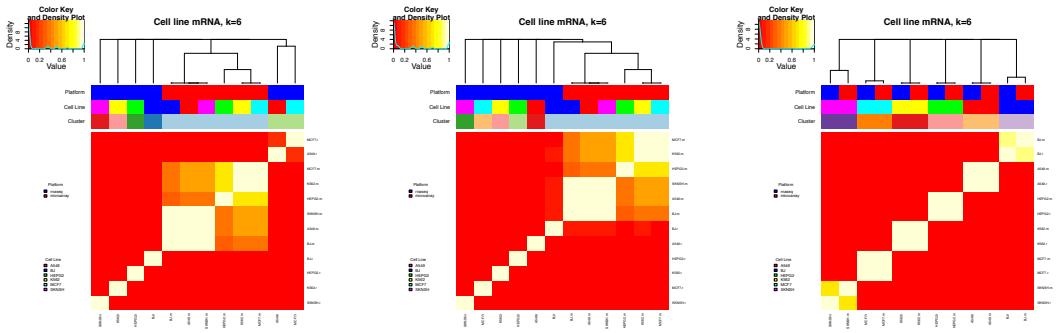
where I(d,k) is an indicator function that records whether sample k had data in dataset d. Thus, Z*(i,j) represents an average of the relative similarities between two samples across the datasets for which both samples have non-missing observations. We note weightings could be incorporated to capture the importance (or non-redundancy) of each of the D platforms being combined, but we have not pursued this direction in the current work.

2.2.2 Applications of RSS to Analysis of Cancer Datasets

2.2.2.1 Batch Effect Removal with RSS (Proof of Concept)

I applied RSS method to mRNA expression data of 6 cell lines from two different datasets. The cell lines represent a variety of different tissues of origin: A549 (lung carcinoma, human), BJ (normal foreskin, human), HEPG2 (normal liver, human), K562 (bone marrow, human), MCF7 (breast adenocarcinoma, human), and SKNSH (neuroblastoma, human). The first dataset contains mRNA expression obtained by RNA Se-

quencing [114]. The second dataset contains mRNA expression obtained by microarray experiments in the Cancer Cell Line Encyclopedia (CCLE) database [17]. The CCLE dataset contains mRNA expression data for many more cell lines than 6, but only 6 were in common between the two datasets. I combined the mRNA expression for the 6 cell lines from the two dataset, resulting in mRNA expression of 12 samples. Figure 2.24 describes the results of our experiment. Figure 2.15(a) shows that simply combining the microarray and RNA sequencing data without transforming it causes the cell line mRNA expression to cluster by the platform. This demonstrates the phenomena called "batch effect", where the data clusters by the source rather than by biological covariates. I applied a commonly utilized batch effect removal method called ComBat [68]. Figure 2.15(b) shows that not all cell lines cluster together and that ComBat does not remove all platform signals. Finally, I applied RSS method to each of the datasets prior to combining them. Figure 2.15(c) shows that all the corresponding cell lines cluster together instead of by the experimental platform. RSS appears to work better than the most popular batch effect removal method. Of course, we should note that this is only possible due to all the molecular subtypes being equally represented in the two datasets. More than two datasets can be combined using RSS. However, the assumption that this method makes is that the datasets have similar representation of biological covariates and patterns (e.g. all molecular subtypes are represented in the data in every dataset).



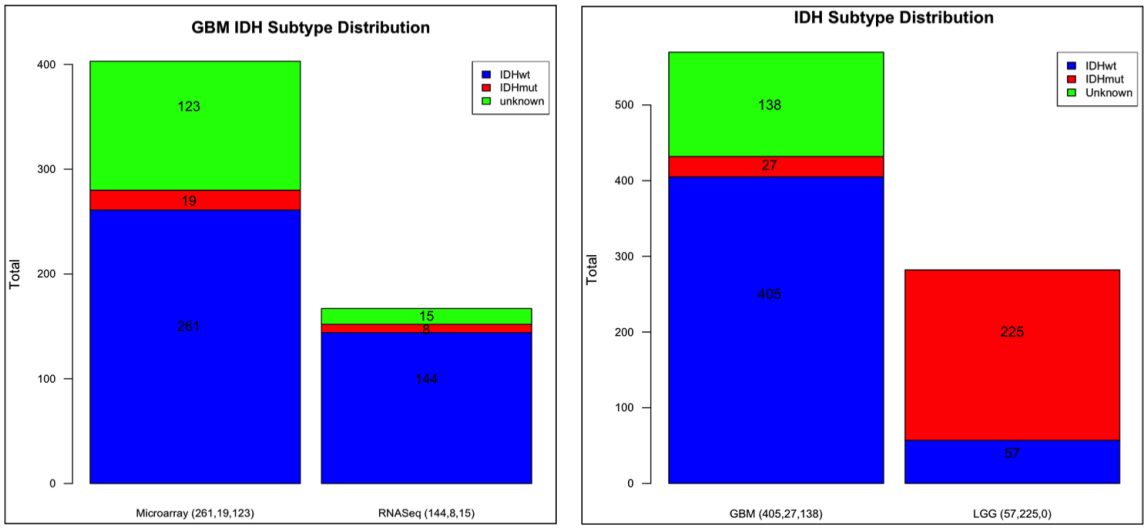
(a) Pre-transformed. (b) ComBat transformed. (c) RSS transformed.

Figure 2.15: Application of RSS method for batch effect removal when combining multi-platform mRNA expression datasets for 6 cell lines achieves the best clustering of the same cell lines together. A) Hierarchical clustering of the pre-transformed microarray and RNA sequencing data after combining it. B) Hierarchical clustering of the data after ComBat batch effect removal method was applied to it. C) Hierarchical clustering of the data after RSS method was applied to it.

2.2.2.2 Batch Effect Removal with RSS In Joint Gliomas Analysis

As a part of the joint gliomas - Glioblastoma (GBM) and Lower Grade Gliomas (LGG) - The Cancer Genome Atlas (TCGA) Analysis Working Group (AWG) we analyzed mRNA expression data from both miroarray and RNA sequencing experimental platforms for GBM (only RNA sequencing data was available for LGG tumors). A number of the samples in the GBM dataset had mRNA expression data available for both the microarray and RNA-Seq platforms (Figure 2.17). As a first step of assessing whether RSS would be an appropriate method to apply to these data, we considered distribution of IDH molecular subtypes (a major molecular diagnostic and prognostic subtype in glioma tumors). Figure 2.16 shows that the IDH mutant and IDH wild type (WT) subtypes are sufficiently represented across both platforms and both tumor types.

I combined data from both platforms, microarray and RNA-Seq, for both tumor types, GMB and LGG. Figure 2.18 shows that cohort samples cluster by the experimental platform prior to any transformation of the data. I applied RSS method separately to LGG, GBM microarray, and GBM RNA-Seq data and then combined all the transformed datasets (Figure 2.19). We found that experimental platforms mix across the data cohort now (Figure 2.20), as well as the tumor types match to some extent. The separation of the tumor types in the cohort can now be explained by the biology of the molecular subtypes. For example, LGG tumors are enriched for IDH mutants and tend to cluster together. We found that two clusters (clusters 3 and 4) are enriched in IDH mutant tumors and most GBM IDH mutants cluster with some LGG mutant in cluster 4. We also found that many of the LGG IDH WT tumors cluster with GBM tumors in cluster 5.



(a) Distribution of IDH subtype in GBM.

(b) Distribution of IDH subtype (GBM, LGG).

Figure 2.16: Distribution of molecular platforms and IDH molecular subtypes in the joint gliomas mRNA expression data. A) Distribution of IDH subtypes in the GBM data across microarray and RNA sequencing platforms. B) Distribution of the IDH subtypes in the GBM and LGG data.

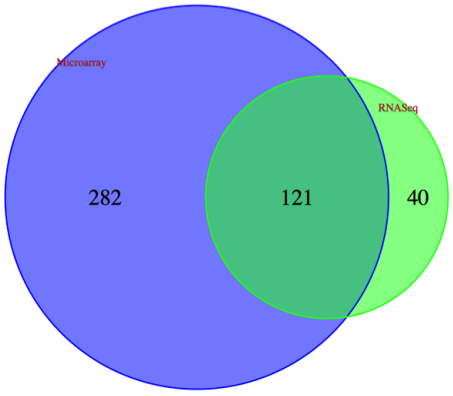
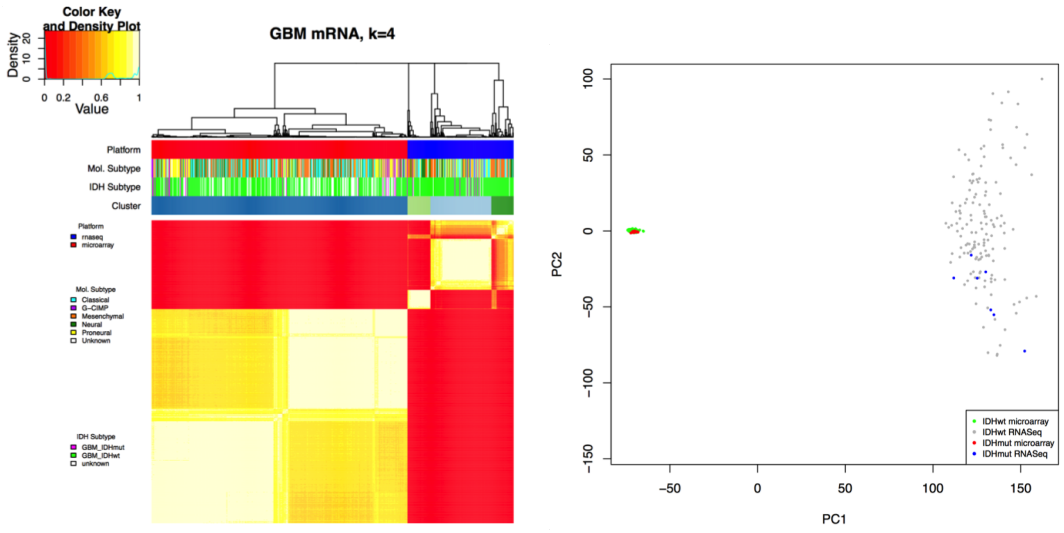


Figure 2.17: Match samples that are common between the GBM microarray and RNA-Seq datasets.



(a) Pre-transformation gliomas clustering.

(b) Pre-transformation gliomas PCA.

Figure 2.18: View of the mRNA expression data for GBM and LGG tumors prior to applying data transformation. A) Hierarchical clustering of the mRNA expression data shows clear clustering by the experimental platform. B) Principle Component Analysis of the mRNA expression data shows separation by the experimental platform.

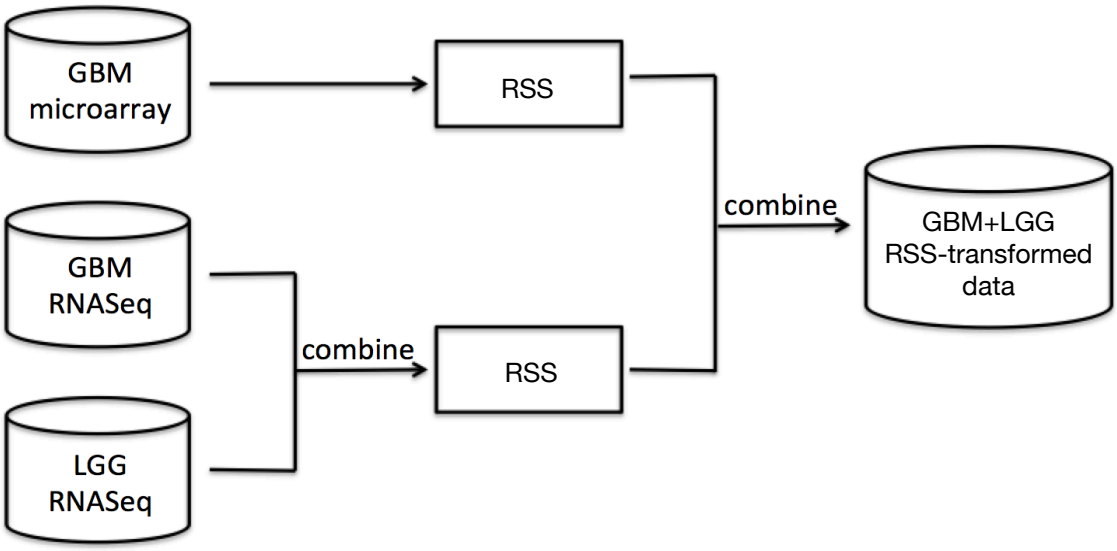


Figure 2.19: Summary of how RSS method was applied to the gliomas mRNA expression data.

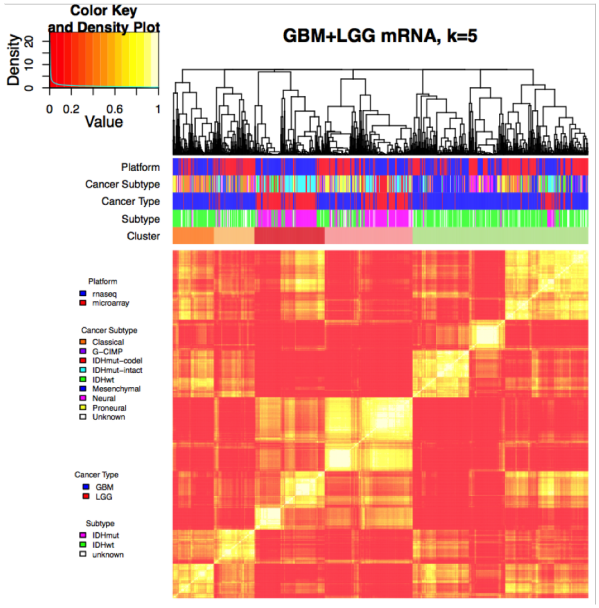


Figure 2.20: View of the mRNA expression data for GBM and LGG tumors after applying RSS data transformation.

2.2.2.3 Integration Of Platforms As Data Types For Tumor Map

Since RSS method makes no assumptions about the nature of the data it is being applied to as long as the data is some type of features by samples, one can apply it to square similarity matrices. In fact, RSS is a perfect approach for combining multiple kernel/similarity spaces. I utilized RSS method to integrate multiple data platforms for the Tumor Map method (described above) on Pan-cancer 12 dataset [59]. I computed individual correlation spaces for 5 different experimental platforms (mRNA expression, miRNA expression, methylation levels, protein expression, and somatic copy number variation) and applied Tumor Map method to build visualization of the integrated genomic space. I describe my findings in the Section 2.1.

2.2.2.4 Integration Of Platforms For Master Regulator Analysis

Finally, I describe the application of RSS to multiple modalities of data for master regulator inference in Cholangiocarcinoma (see Section 3.2 for background information of this tumor type). As a part of the Cholangiocarcinoma (CHOL) Analysis Working Group (AWG), we identified four distinct molecular subtypes. Two of these subtypes captured the interest of the group. One is enriched in oxidative phosphorylation (OxPhos) signaling and the other is enriched in chromatin remodeling signaling. We also found that as one signal increases, the other signal decreases, indicating anti-correlation of the OxPhos and chromatin remodeling signatures across CHOL samples. The question the group posed was what biological markers differentiate between these two groups and what regulatory elements might be playing a role in driving these two subtypes, as well as what possible therapeutic implications are possible to infer from these subtypes. Furthermore, the group was interested in answering these questions from some kind of integrative analysis approach in order to incorporate information presented by multiple platforms.

I performed integrated-platform master regulator analysis to see if we can answer those questions. The objective of this experiment was to identify master regulators and relevant molecular pathways that are differential between the two groups of samples. Master regulators are genes that, if targeted by a drug, have the highest effect in reversing the disease given the topology of the gene regulatory network. These genes are often referred to as network hubs as they usually exhibit high connectivity in the

context of the a gene regulatory network. I utilized a method called SPIA [12], which takes a directed network and a set of scores for the nodes in that network and performs iterative updates of the scores based on the connectivity of the nodes in the network and directionality of the edges. As a result, nodes that have many parents and ancestors are highly scored. Because our problem requires just the opposite - we want nodes that are sitting at the top of the regulatory chains and propagate to a lot of descendants to score highly - I reversed the directionality of the edges in the network. I used the Superpathway network described by Vaske et al. [132] and applied RSS method to the gene-level mRNA expression (RNA-Seq), gene-level somatic copy number variation, and gene-level methylation datasets for CHOL samples in the two groups of interest (Figure 2.21). I applied RSS to each of the data modalities and reversed the sign of the methylation RSS-transformed space because of the inverse relationship between the methylation levels and expression levels. Then I combined the three RSS-transformed spaces. I used the final RSS score as inputs into the SPIA method. This method produced per-sample master regulator scores across the genome. I used LIMMA [116, 120, 87] differential analysis method to create a differential master regulator signature (OxPhos subtype vs. chromatin remodeling subtype), which I used as an input into the PATHMARK [132] method to extract significantly connected subnetwork components describing the differential signature. As the first step in validating the SPIA results, I checked that the top differential pathway in the LIMMA signature is indeed Oxidative Phosphorylation (Figure 2.22). Figure 2.23 shows the PATHMARK result for the master regulator network with a few excerpts from it for notable subnetwork components. The differential

network indicates that OxPhos tumors are more proliferative and aggressive.

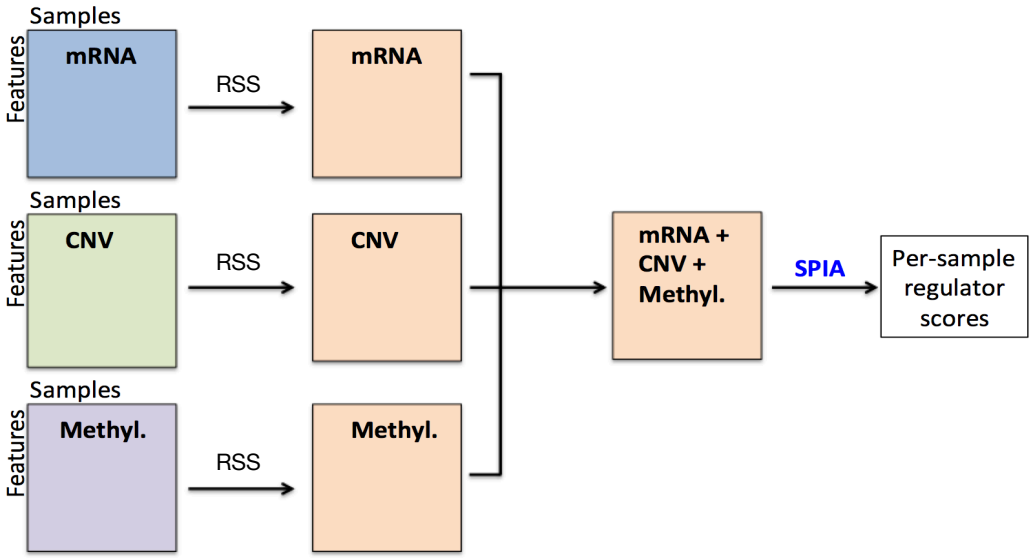


Figure 2.21: Outline of the application of RSS method to mRNA expression, CNV, and methylomics CHOL data.

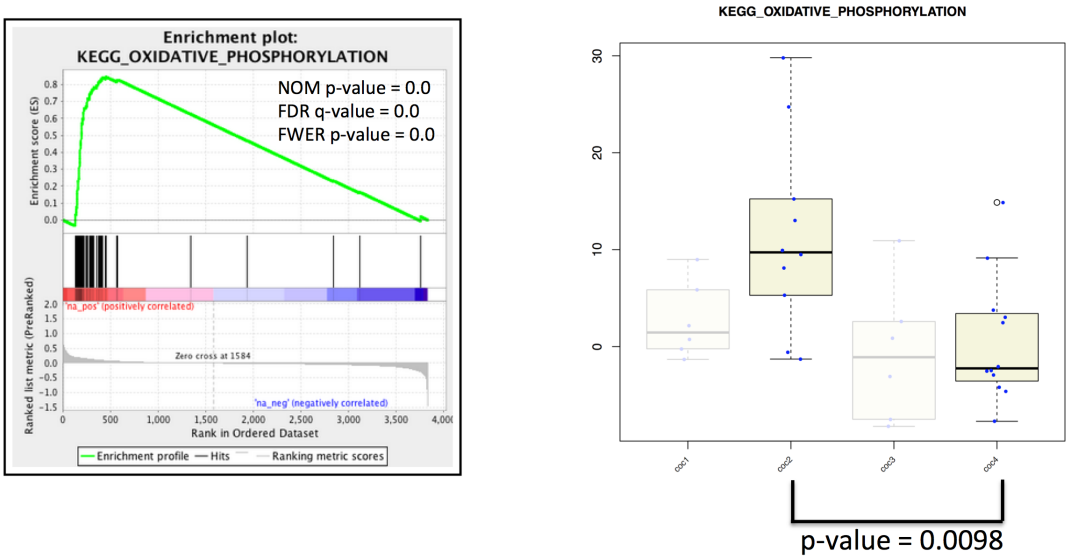


Figure 2.22: Validation of the SPIA results by showing the top differential pathway between the two groups of interest. As expected, the top differentiating pathway is Oxidative Phosphorylation. On the left: Gene Set Enrichment Analysis of the differential master regulator signature, showing the statistical significance of the Oxidative Phosphorylation pathway genes. On the right: each dot is a sample in the cohort; the samples are separated into two groups, reflecting the two groups of interest in our analysis; the pathway levels were obtained by aggregating the master regulator scores across all the genes in this pathway for each of the samples.

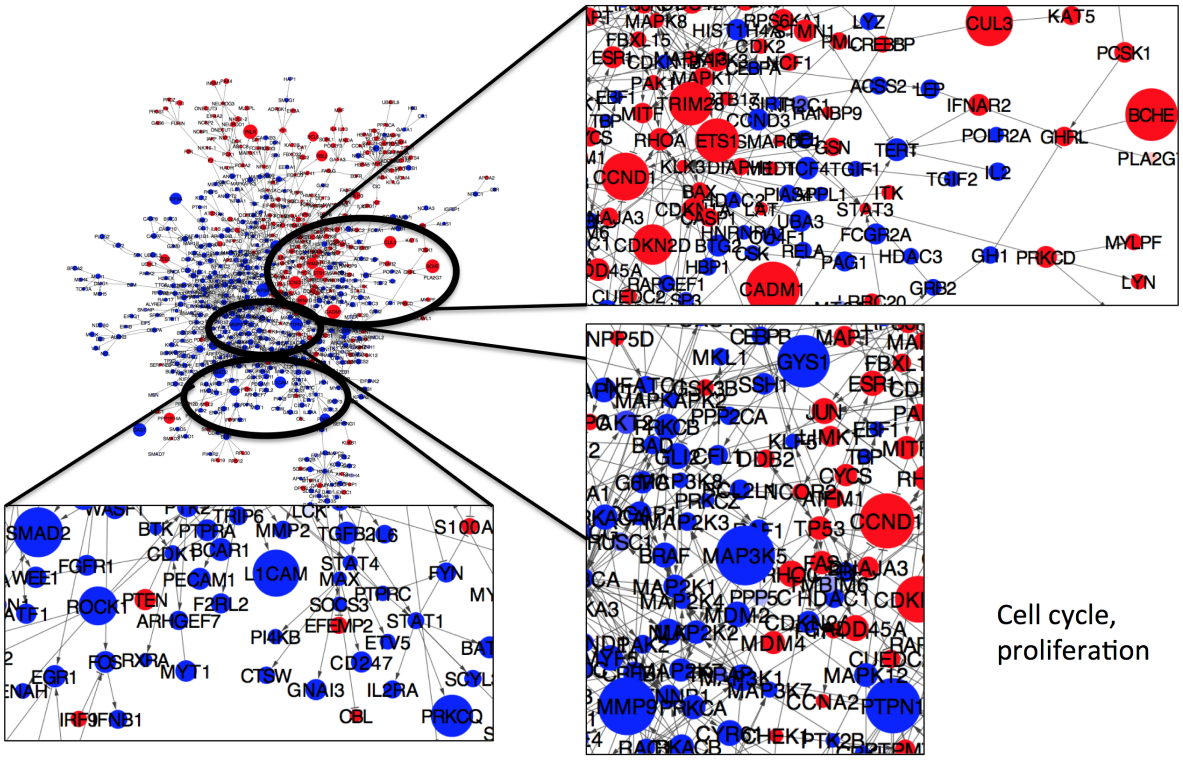


Figure 2.23: Resulting master regulator network, output by PATHMARK method. Red nodes are high in the "Oxidative Phosphorylation" group of samples. Blue nodes are high in the "Chromatin Remodeling" group of samples. The biggest difference between the two groups is proliferative signaling (high in the "Oxidative Phosphorylation" group).

2.2.3 Conclusion

In this section I describe the Reciprocal Significance of Similarities method I developed, an adaptation on a previously published Context Likelihood of Relatedness (CLR) method by Faith *et al.*, for transforming a given space (a feature space or a similarity space) into the one that incorporates sensitivity and specificity in the global context of the reference distribution. I demonstrate the application of this method to a

variety of use cases and biological problems.

2.3 Kernel Space Comparison Helps Contrasting Transformations

As a part of the Graim *et al.* study of community detection in genomic similarity networks, I implemented a pairwise comparison method for kernel matrices. Kernel matrices are square similarity matrices that are symmetric positive-definite and are produced by applying a kernel function to feature vectors. My method computes a metric that can be thought of as correlation of two kernel matrices, in a manner similar to correlating one-dimensional vectors. Our method is an adaptation of a kernel alignment method proposed by Cristianini *et al.* [36]. Their alignment metric can be interpreted as a cosine similarity or cosine of the angle between two kernel spaces. We modified the proposed method by normalizing this measure of similarity, essentially converting it to a Pearson correlation metric.

HOCUS, method proposed by Graim *et al.* (currently in editorial review), exponentiates kernel matrices to define community-based sample groupings. It demonstrates the method on several datasets, including mutation profiles of bladder tumors from The Cancer Genome Atlas (TCGA) project and glioblastoma magnetic resonance imaging (MRI) data. The goal of the kernel alignment experiment was to compare kernel spaces for exponentiated (i^{th} order) similarity matrices of the described two datasets to the exponents of patient-to-patient age difference, converted to similarity space. I

computed age difference between every pair of patients and converted these distances to a similarity space by taking an absolute value of each age difference, normalizing by the maximum age difference in this space, and subtracting the resulting value from 1. I considered this age similarity space, the baseline correlation space, and the second, third, and fourth order spaces and computed pairwise alignments between each pair of kernel spaces. In bladder cancer we found that the age similarity space is most correlated to the second order similarity space while it is weakly correlated with the baseline correlation space. It is also highly correlated with the third order and the fourth order kernel spaces, with slightly lower correlation measures than with the second order kernel space. In the GBM imaging data we found that the age similarity matrix is most correlated with the fourth order kernel space.

BLCA	age	baseline	2nd order	3rd order	4th order
age		2.6005376E-01	9.8090696E-01	9.8090557E-01	9.8090557E-01
baseline			3.1099244E-01	3.1387880E-01	3.1387880E-01
2nd order				9.9607143E-01	9.9607143E-01
3rd order					1.0000000E+00
4th order					

(a) BLCA mutations.

GBMIMG	age	baseline	2nd order	3rd order	4th order
age			8.1426135E-01	9.7716712E-01	9.7719457E-01
baseline					
2nd order				8.4640692E-01	8.4559980E-01
3rd order					9.9999856E-01
4th order					

(b) GBM MRI imaging.

Figure 2.24: Application of kernel alignment method to two sets of kernel matrices. A) First set of matrices are patient-to-patient age difference converted to similarity space and the second set of matrices are bladder cancer mutation profile similarities, computed using Humming similarity. B) First set of matrices are a patient-to-patient age difference converted to similarity space and the second set of matrices are similarities of glioblastoma MRI images represented by voxels.

2.3.1 Conclusion

As a part of the my work on Graim *et al.* paper I implemented and applied an adaptation to an already published kernel alignment method, which compares two

kernel spaces by computing a measure which has a similar interpretation to a Pearson correlation of two vectors performed on matrices.

Chapter 3

Discovering New Biology in Cancer Has Potential to Help Cancer Patients

When we perform experimental procedures to gather together various data that represent views of the cell state for various cell types - whether these are cell lines, patient tumor biopsies, or tumor xenographs - we want to analyze these data to give us insights into biological functions and mechanisms that can describe these cells. We usually know a set of basic information about the cells we are analyzing, e.g. tissue type, cell line family, or histological annotations by a pathologist. Often when analyzing samples that come from cancer patients we have more extensive descriptions of the data, such as treatment and outcome information. These descriptions, or annotations, of each sample

in the cohort help us associate particular molecular descriptors with clinical descriptors (e.g. molecular functions enriched in a particular histological subtype of a given tumor type). Performing unsupervised analysis of the data allows us to identify groupings of samples that are purely driven by the molecular features. Various supervised methods help us identify features relevant to particular biological or clinical labels of interest.

As a part of my doctoral work, I participated in a number of TCGA analysis working group (AWG) collaborations that sought to answer a number of biological questions about individual tumor types as well as find cross-cancer patterns. In this chapter I describe the most notable work I completed. Some of my work uses the tools and methods described in Chapter 2, and some use additional methods.

3.1 Analysis of Gliomas of Combined Grades and Histologies

Lower Grade Glioma (LGG) and Glioblastoma (GBM) are two types of brain cancer that arises from glial cells. These types of tumors make up most of the adult brain tumors. LGG tumors consist of lower grade glioma malignancies (stage 2 and stage 3), while GBM tumors consist of the highest grade glioma malignancies (grade 4).

TCGA GBM/LGG AWG analyzed a number of platforms for a number of glioma samples. The group completed manuscript describing the results of our work and published in Cell journal [29]. I performed two forms of analysis as a part of this publication. First analysis involved unsupervised clustering of combined RNASeq

and methylomics spaces using Tumor Map (see 3.1.1). The second analysis involved identifying significant molecular pathways involved with progression from LGG to GBM (see 3.1.2). My analysis contributed to one main and one supplemental figures in the manuscript.

3.1.1 Unsupervised Analysis of RNA-Seq and Methylomics Combined Space Using Tumor Map

3.1.1.1 Combining Multi-platform Multi-tumor Datasets

I utilized the ComBat [68] batch effect removal method in order to combine mRNA expression data from the GBM RNA-seq (n=154), GBM Agilent (n=525), LGG RNA-seq (n=513), and LGG Agilent (n=27) datasets. We chose to use data generated using Agilent microarray platform over those generated using Affymetrix because such data were available for both tumor types, while Affymetrix data were only available for GBM samples. I combined the 4 datasets and ran ComBat. We flagged 4 batches, one for each dataset, as input into the ComBat method. One hundred and forty nine GBM samples were analyzed using both Agilent and RNA-seq platforms. Twenty seven LGG samples were analyzed using both Agilent and RNA-seq platforms. I utilized these matched samples as biological covariates in the ComBat method, indicating to the method that those samples are similar and should be kept together. Upon completion of the data transformation, I removed all redundant samples analyzed using the Agilent platform whenever the sample was also analyzed using RNA-seq. This combined mRNA

expression dataset (n=1043) was used for Tumor Map analysis.

3.1.1.2 Combined RNASeq and Methylation Space Reveals Important Relationships Between Molecular Subtypes of Gliomas

Prior to the analysis, technical and batch effects in the gene expression data were mitigated as a preprocessing step described above. I computed sample-by-sample pair-wise correlations. From RNA expression data, we selected 6002 genes whose expression was the most variable, based on the variance distribution (see Figure 3.1). The 1301 methylation probes were selected by manual curation of the probe list by the experts in the group. I computed sample-by-sample pair-wise correlations of the methylation profiles, then combined the RNA expression and methylomics spaces using RSS (see 2.2.1) method.

The analysis of the individual platforms (Figure 3.3) shows that the glioma molecular subtypes separate differently in mRNA expression and DNA methylation spaces. The mRNA expression based map recapitulates some of the well-known relationships between the glioma molecular subtypes. For example, some of the IDHwt LGG tumors cluster with Classical GBM tumors, the G-CIMP GBM tumors separate into two groups, one of which clusters with IDHmut-non-codel LGGs and the other clusters with Proneural GBMs. Interestingly, IDHmut-non-codel LGG tumors separate into two groups in the mRNA expression space. Similarly, Neural GBM tumors separate into two groups as well. In the DNA methylation space LGG molecular subtypes appear to have very distinct methylomics signatures, with an exception of the IDHmut-non-codel

tumors separating into two groups. On the other hand, GBM molecular subtypes seem mix quiet a lot, separating into two major methylation groups, each consisting of a mixture of molecular subtypes.

When combining the mRNA expression and methylation spaces we can capture patterns captured by both of those spaces (Figure 3.3 shows the three views of the map included into the published manuscript). Figure 3.4 summarizes the separation of the molecular subtypes in this combined space. This combined space shows groupings of samples that correlate well with mRNA expression clusters and to a lesser extent with the methylation clusters (parts B and C of Figure 3.3).

3.1.2 Pathway Analysis Reveals Important Molecular Differences Between GBM and LGG Tumors Within The Same Subtypes

When combining GBM and LGG data some GBM and LGG tumors cluster together. We wanted to understand the differences between tumors of the highest and lower grades within the same clusters in order to understand the features associated with progression of these tumors.

I used mRNA expression for samples available through RNA-seq platform only and the CNV data to transform the data into inferred pathway activity levels using PARADIGM (Vaske et al., 2010). I then considered a number of dichotomies, such as LGm1 GBM vs. LGG (see Figure 3.5). Some of the dichotomies I considered have significantly different numbers of samples in each class (see Figure 3.5). In order to make statistically strong inferences about pathway activities I only considered those

dichotomies in which both classes are well represented by their members and the variance within the classes is much smaller than the variance between the classes. In other words, I selected those dichotomies where sample scatter is small within the classes and classes are separable in the pathway space. Based on the PARADIGM IPLs (Inferred Pathway Levels) we computed pair-wise Spearman rank correlation for each pair of samples and then computed within-class and between-class variance of the correlations, first for the first class and then for the second class. I then computed the F-statistic for each of the classes in the dichotomy and the p-value based on the F-distribution. I aggregated the p-value for the dichotomy by computing the mean p-value. We selected those dichotomies that had an aggregated p-value $<= 0.05$. Figure 3.5 shows final dichotomies analyzed for the differential pathway activities. For each dichotomy selected, I computed differential activity levels using the linear models for microarrays and RNA-seq data (LIMMA) method [116, 120, 87]. I then applied Gene Set Enrichment Analysis (GSEA) [124] to the HUGO members of the full differential vector. I extracted only those pathways that had FDR-adjusted q-value of $<= 0.1$. At the same time, I extracted statistically significant differentials (multiple hypothesis adjusted p-value $<= 0.05$) and ran PATHMARK [132] on the statistically significant differential activities obtained from LIMMA to extract significantly connected components of the global Superpathway [132] regulatory network. An additional filter of 3 standard deviations was applied to the PATHMARK method. This means only those activities that fall outside 3 standard deviations of the empirical distribution of the statistically significant differentials pass through the filter. A network connection is extracted if both vertices connected by

that connection pass the filter. For each pathway gene set that passed the GSEA q-value of 0.1 I computed the overlap of the pathway genes and those that survived the PATHMARK filter as well as the over-representation hypergeometric p-value. I then extracted those pathways that passed with the p-value of $<= 0.05$. Figure 3.6 shows an overview of the described above process for extracting significantly active pathway from the glioma data. Figure 3.7 shows pathway views of the significant IPLs from Figure 3.5 in which IPLs representing families, complexes, phopho-events and redundant complexes were removed for better visualization.

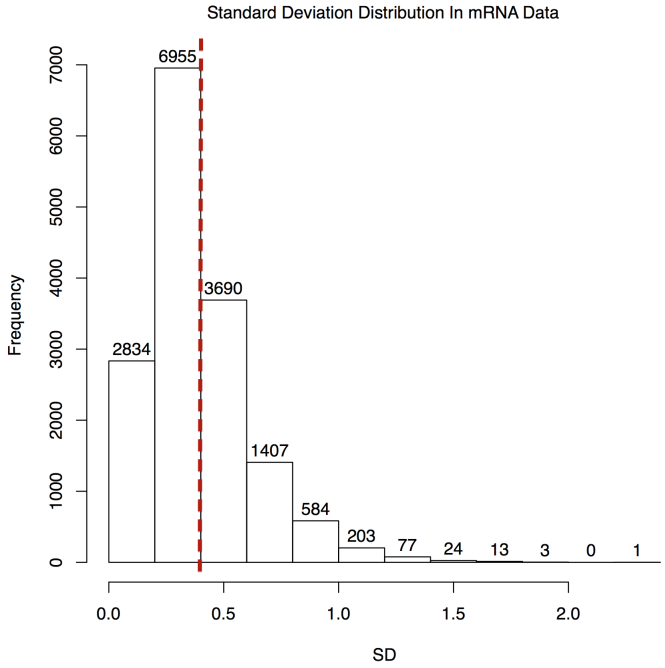


Figure 3.1: Distribution of the standard deviation of the gene features for the GBM and LGG tumors. The red line shows the SD cutoff for the features used to by the Tumor Map method. Genes on the right of the red line where included into the analysis.

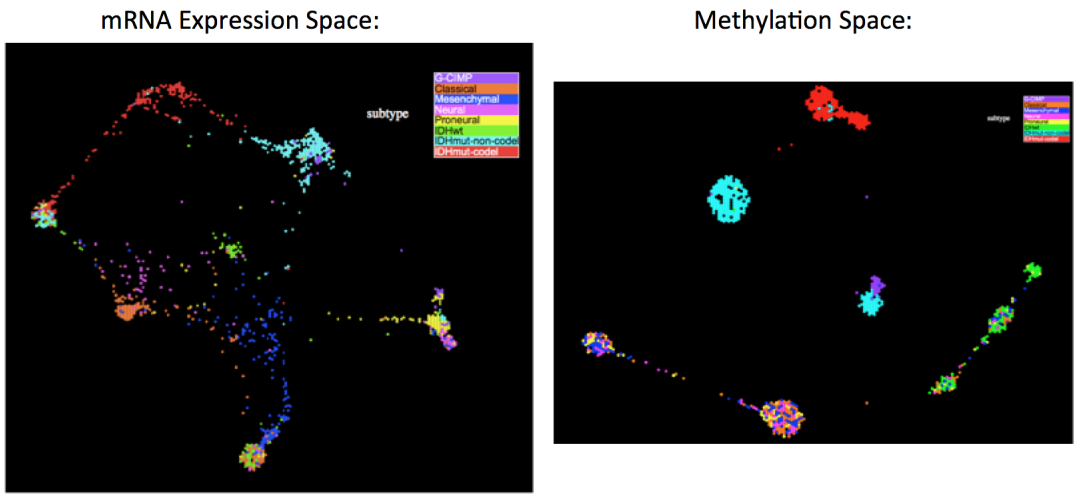
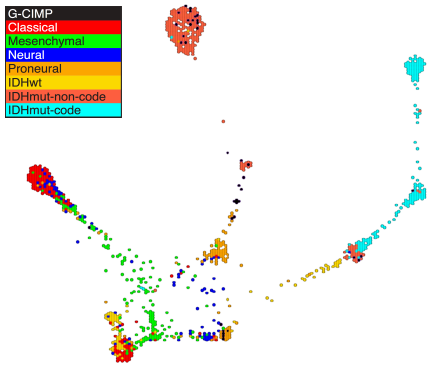
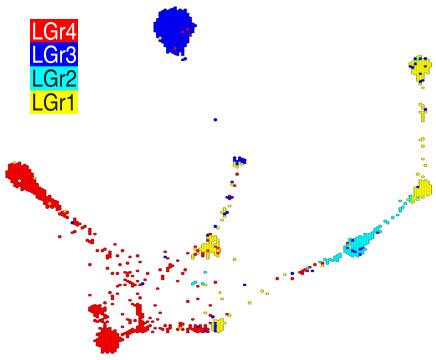


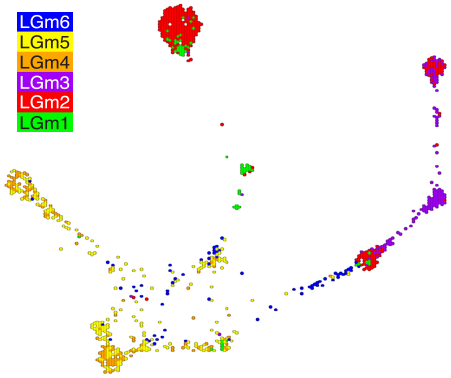
Figure 3.2: Tumor Map analysis based on individual platforms shows the separation of glioma molecular subtypes differs in the mRNA expression and DNA methylation spaces A) mRNA expression Tumor Map B) DNA methylation Tumor Map.



(a) Molecular subtypes.



(b) mRNA joint subtypes.



(c) Methylation joint subtypes.

Figure 3.3: Tumor Map based on mRNA expression and DNA methylation data. Each data point is a TCGA sample colored coded according to their identified status.

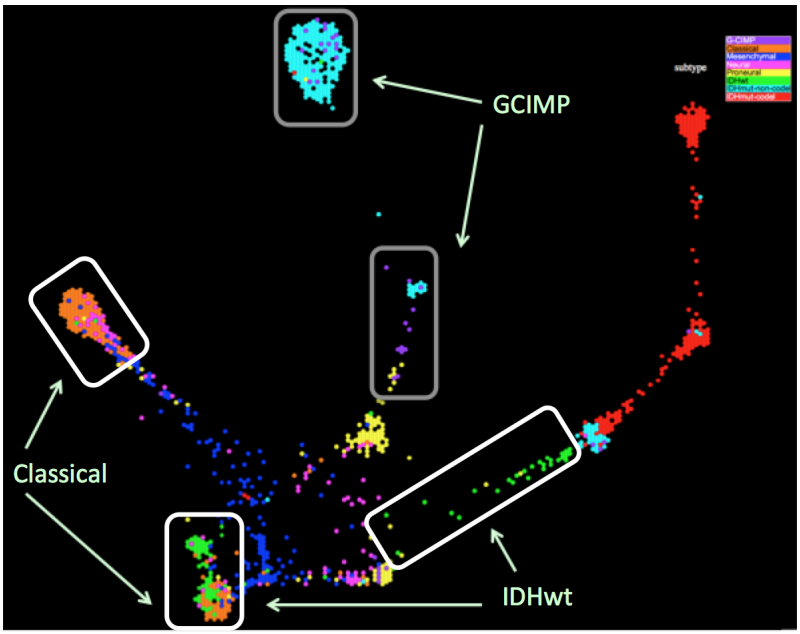


Figure 3.4: Tumor Map based on mRNA expression and DNA methylation data. Each data point is a TCGA sample colored coded according to their identified status.

Dichotomy	Class 1	Class 0	# samples class 1	# samples class 0
LGM1_gbm_vs_lgg	GBM	LGG	24	47
LGM2_gbm_vs_lgg	GBM	LGG	12	266
LGM4_gbm_vs_lgg	GBM	LGG	143	23
LGM5_gbm_vs_lgg	GBM	LGG	235	46
LGM6_gbm_vs_lgg	GBM	LGG	55	27
LGR1_gbm_vs_lgg	GBM	LGG	99	130
LGR3_gbm_vs_lgg	GBM	LGG	37	228
LGR4_gbm_vs_lgg	GBM	LGG	512	78
Hyper_vs_Hypomethylated	Hypermethylated	Hypomethylated	243	25
LGM1_Hyper_vs_Hypomethylated	Hypermethylated	Hypomethylated	35	25

Figure 3.5: Table showing distribution of GBM and LGG samples in various dichotomies.

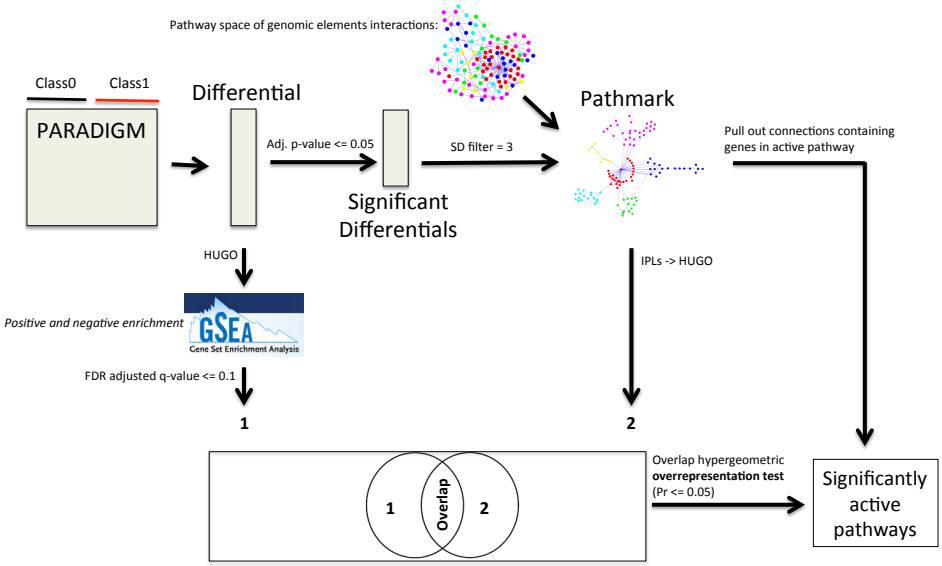


Figure 3.6: The method used to extract significant pathways driving the LGr3 and LGr4 groups.

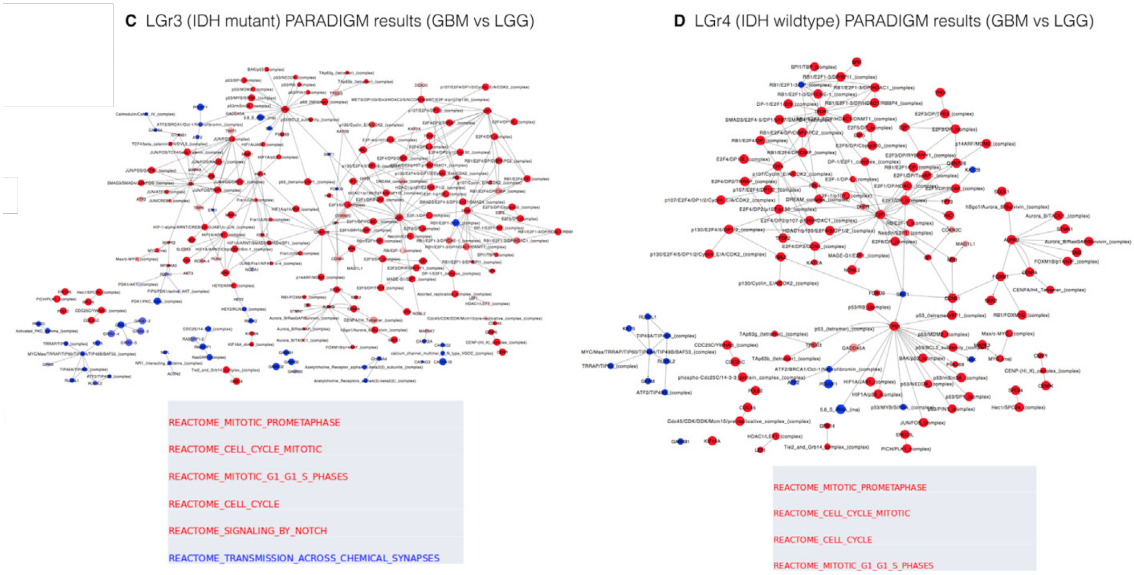


Figure 3.7: Pathways involved in progression of the two expression subtypes of the GBM and LGG tumors. The figure is displayed as it was included into the published manuscript as a part of the supplemental figure S5. The two pathways are part C and D of figure S5. Part (C) shows the pathways that drive LGr3 (IDH mutant enriched subtype). Part (D) shows the pathways that drive LGr4 (IDH WT enriched subtype).

3.1.3 Conclusion

We analyzed a cohort of two tumors: lower grade gliomas and glioblastomas. Both are brain tumor types. The group analyzed these two tumor types to identify their molecular similarities and differences. We identified that one subtype of lower grade glioma, which is a less aggressive type of tumor, clusters with glioblastoma tumors, a more aggressive tumor type. This suggests that even as a generally less aggressive tumor type, this particular subtype should be treated more aggressively in clinic. We found that the integrated RNA-Seq and methylation genomic space accurately reflects what we know about the biology of these tumors. We also derived and described the pathways

that differentiate between GBM and LGG tumors within the same molecular subtypes. I proposed and described a method for the pathway extraction pipeline (Figure 3.6). Generally, we found that GBM tumors exhibit aggressive markers when compared to LGG tumors. For example, cell cycle and related cell activities are enriched in GBM tumors (over LGG tumors) within LGr4 molecular subtype, a subtype derived through clustering of the combined RNA-Seq data. Similarly, cell division functions are enriched in the GBM tumors compared to LGG tumors in the LGr3 molecular subtype.

3.2 Analysis of Cholangiocarcinoma

Cholangiocarcinoma (CHOL) is a type of epithelial cancer that originates in bile duct. This cancer is rare and only 2,000 to 3,000 people in the United States a year develop this type of neoplasm [2] and is generally classified as an adenocarcinoma.

TCGA CHOL AWG analyzed the output of a number of genomic platforms for 36 tumors classified as having bile duct as their tissue of origin. The group completed manuscript describing the results of our work and submitted to Cell journal. I performed two forms of analysis as a part of this publication. First analysis involved unsupervised clustering of RNASeq data using Tumor Map layout (see 3.2.1). The second analysis involved unsupervised analysis of three different types of cancer utilizing multiple platforms (see 3.2.2). My analysis contributed to two main figures in the manuscript.

3.2.1 Clustering Based on Tumor Map Layout Positions Provides an Alternative Way to Infer Molecular Subtypes

I utilized the Tumor Map method to perform unsupervised joint pancreatic adenocarcinoma (PAAD), liver hepatocellular carcinoma (LIHC), and cholangiocarcinoma (CHOL) mRNA cluster analysis. I combined mRNA expression RNA-Seq data for the three cancers and excluded all liver-specific genes from the feature space, then analyzed the remaining 15,269 genes for 588 samples. I computed Tumor Map Euclidean space (x,y) coordinates for each sample, using N=6 closest neighbors for rendering the map and then computed Euclidean distance between each pair of samples based on the (x,y) coordinates. I performed k-means clustering based on those distances. I chose 7 clusters as the best solution because it best recapitulated the expected biological relationships between the tumors (Figure 3.8).

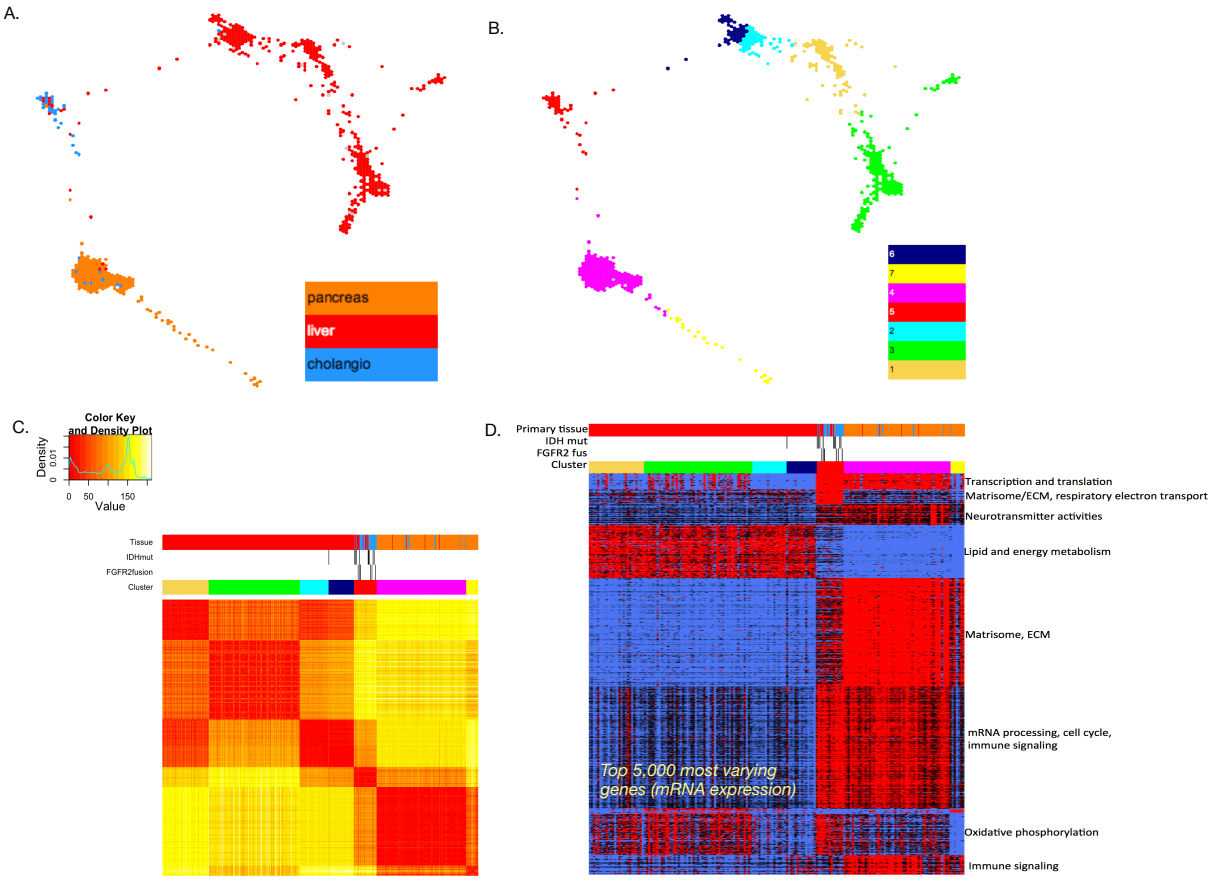


Figure 3.8: Tri-cancer RNASeq data clustered using Tumor Map. A) CHOL, LIHC, and PAAD samples laid out in Tumor Map based on RNASeq data. B) Cluster assignments (k=7) based on the sample positions in the Euclidean 2-D space (using $1 - Euclideandistance$ as a measure of similarity). C) Sample-to-sample correlations in the original RNASeq space based on the Tumor Map clustering solution. D) Feature (gene expression) space (from RNASeq data) based on the Tumor Map clustering solution.

3.2.2 Tri-cancer Multi-platform Analysis of Cholangiocarcinoma, Pancreatic Adenocarcinoma, Liver Hepatocellular Carcinoma Helps Identifying Important Histological Subtypes From Molecular Data

To build tri-cancer multi-platform map, I combined tumor mRNA expression, copy number variation, and methylation profiles for cholangiocarcinoma (CHOL), liver hepatocellular carcinoma (LIHC), and pancreatic adenocarcinoma (PAAD). First, I combined mRNA expression data from RNA-seq for cholangiocarcinoma (n = 36), liver hepatocellular carcinoma (n = 373), and pancreatic adenocarcinoma (n = 179) into a single dataset (n = 588). Second, I combined copy number Gistic calls for cholangiocarcinoma (n = 36), liver hepatocellular carcinoma (n = 370), and pancreatic adenocarcinoma (n = 184) into a single dataset (n = 590). Third, I combined methylation profiles from HumanMethylation450 (HM450) platform for cholangiocarcinoma (n = 36), liver hepatocellular carcinoma (n = 429), and pancreatic adenocarcinoma (n = 186) into a single dataset (n = 651). I computed sample-by-sample pair-wise similarities for each dataset, producing three square similarity matrices, using Spearman rank correlation as a similarity measure on these continuous-valued datasets (mRNA expression, copy number variation, and methylation). Next, I standardized each similarity matrix using the RSS method (see 2.2.1). To build the map layout, the closest neighborhood of 10 samples was selected for each sample from the standardized integrated similarity matrix. I applied the DrL layout method (see 2.1.4.2) to create a map of these tumors.

While tumor clusters in this multi-platform space were mostly driven by tis-

sue type, some samples mixed with other samples from different tissues (Figure 3.9A). Specifically, some CHOL samples clustered with both PAAD and LIHC tumors and some LIHC samples clustered with CHOL tumors. Specifically, two CHOL samples cluster with LIHC samples. Those two samples and two LIHC samples that were originally annotated as LIHC but later included with CHOL analysis are annotated as HCC in the heatmap representing unsupervised analysis of CHOL molecular subtypes and clinical annotations/markers (Figure 3.10). In addition, five CHOL samples clustered with PAAD tumors (Figure 3.10). Each of these five samples robustly correspond to the extrahepatic cholangiocarcinoma (ECC) phenotype and track with somatic copy number alteration, methylation, and mRNA (RNASeq) unsupervised clustering solutions (Figure 3.10).

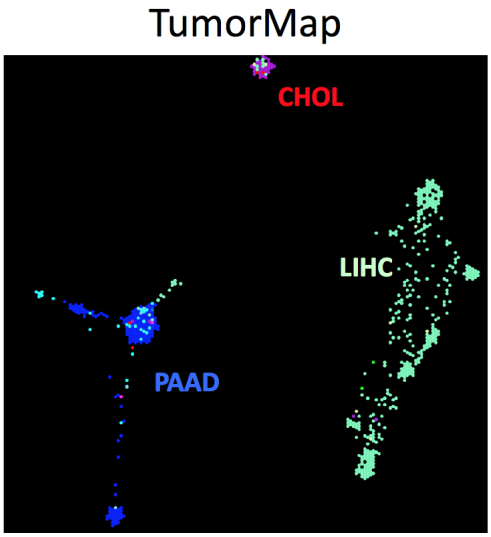


Figure 3.9: Tumor Map depicting tri-cancer (CHOL, LIHC, PAAD) integrated genomic space (RNA-Seq, CNV, methylation). The samples in the map are colored based on the tissue of origin.

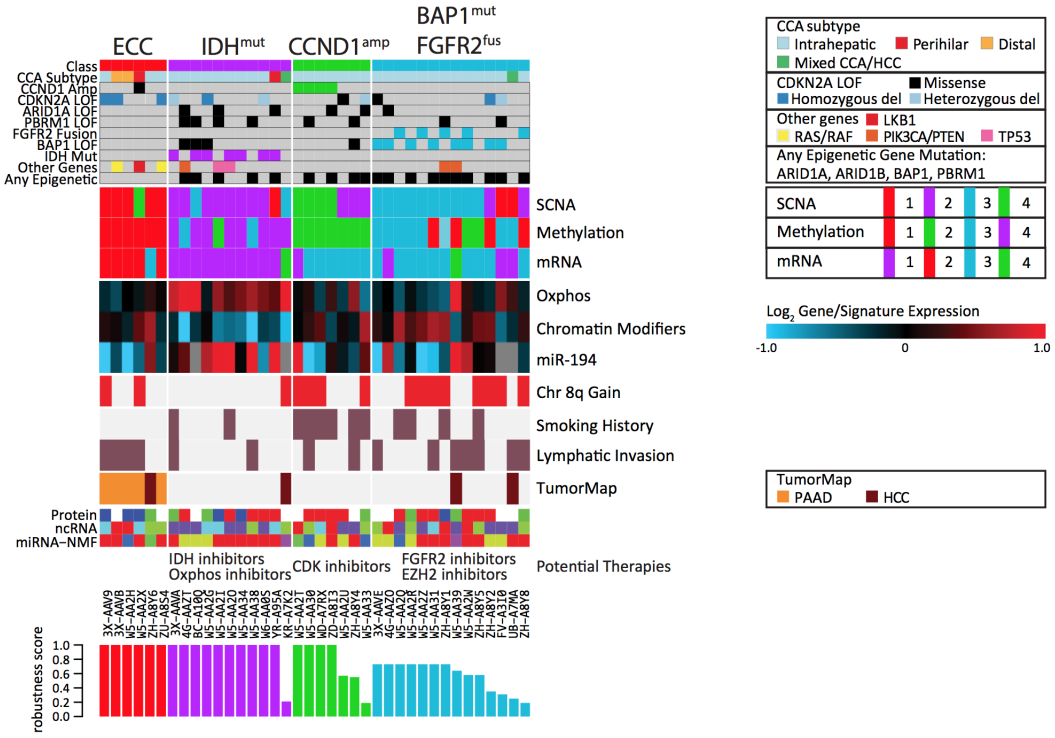


Figure 3.10: A view of the CHOL cohort in light of various genomic markers and signatures (exert from main figure in the manuscript). The samples (listed below the heatmap) are ordered based on their molecular subtype along the x-axis and are annotated by various features along the y-axis. All the samples that clustered with PAAD samples robustly correspond to the extrahepatic cholangiocarcinoma (ECC) subtype. IDH mutants are enriched for high scores of the Oxidative Phosphorylation signature.

3.2.3 Conclusion

We analyzed a cohort of cholangiocarcinoma tumors on their own as well as in the context of other tumors. We utilized Tumor Map method to perform integrated multi-cancer analysis of RNA-Seq, CNV, and methylation genomic platforms and identify several major molecular subtypes of cholangiocarcinoma. We also utilized Tumor Map layout of RNA-Seq tumors from three cancer types to drive the sample clustering

(we clustered based on the layout positions rather than original genomic features) and described these derived molecular subtypes.

3.3 Analysis of Testicular Germ Cell Tumor

Testicular Germ Cell Tumor (TGCT) is a type of testicular malignancy that arises in the germ cells. Seminoma tumors compose about half of all TGCTs and are generally considered less aggressive and slow growing. The rest of the TGCTs consists of several non-seminoma subtypes (embryonal, teratomas, and yolk sack), which are more aggressive and more likely to spread and metastasize. TGCTs are significantly associated with cryptochidism, a phenomena of undescended testicles. However, molecular mechanisms driving this association are not yet clear. The TGCT analysis working group (AWG) set out to answer this and other biological questions about the TGCTs.

3.3.1 Analysis of Match Primary Tumors Shows Independent Origin of These Malignancies

Interestingly, because testicles are paired organs they have a potential for multiple primary tumors arising in different sides. This was the case for 5 individuals in our cohort. In fact, having a primary tumor in one testicle makes an individual more likely to get a primary tumor in the other testicle as compared to an individual with no prior TGCT diagnosis. We were able to analyze 10 primary tumors (5 pairs of tumors, or match tumors) and describe their molecular characteristics. This is the only tumor

type in the TCGA cohort that has match primary tumors. These match tumors arose independently from their counterparts from the same patient and do not share a cell of origin. The group found that different primary tumors in the same patients share very little with each other and in an unsupervised analysis setting behave as if they came from different patients.

In this section I describe my work and contributions to the manuscript currently in preparation.

3.3.2 Unsupervised Molecular Subtypes in PARADIGM IPL Space Correlate With Histological Labels

I analyzed 137 TGCT tumors by applying PARADIGM [133] method to mRNA expression and somatic copy number data (Figure 3.11). I first pre-processed the data to only include the genes present in both data types (n = 19,512) and then ran PARADIGM, which outputs inferred pathway activity level for each feature in the Superpathway [133]. I used the output of PARADIGM method for various analysis described later in this section.

I performed consensus k-means clustering [136] of the PARADIGM inferred pathway levels (IPLs). Based on the silhouette score method, k = 4 solution was the most optimal (Figure 3.12 bottom left). I found that histological labels are the biggest separator of the unsupervised subtypes of the TGCT samples in PARADIGM space (Figure 3.12 heatmap on the right), in fact providing almost perfect separation between seminomas and non-seminomas. Furthermore, within non-seminomas we see

two major molecular subtypes, one is enriched in embryonal and the other is enriched in teratomas and yolk sack. This separation reflects previously known biology of these tumors. Finally, I found that seminoma tumors separate into two major classes, which upon further analysis was determined to not correlate with tumor purity, highlighting a possible difference in molecular drivers within seminoma tumors. One of these seminoma subtypes is enriched in KIT mutants. Separation of KIT mutants from other seminomas was also found to be a major signal in the mRNA expression data by another member of the AWG.

Interestingly, looking at the 5 match primaries in the TGCT cohort confirmed by findings from other genomic platforms that these tumors are no more similar than any two tumors that come from two different TGCT patients (Figure 3.13). The IPL results from PARADIGM were compared between each pair of the match primaries (5 individuals). The scatter plot shows the IPLs, where y-axis is the first primary and the x-axis is the second primary.

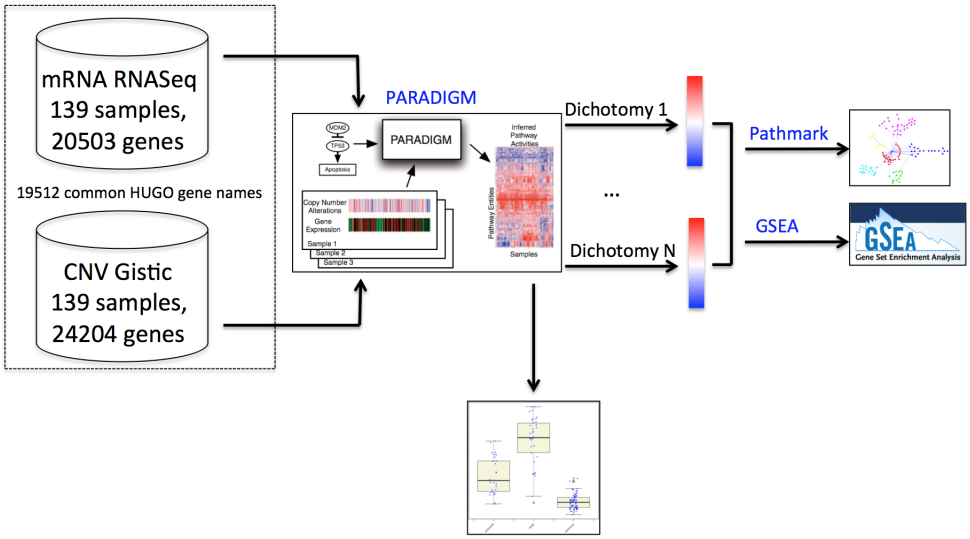


Figure 3.11: Outline of the PARADIGM analysis pipeline and its various parts.

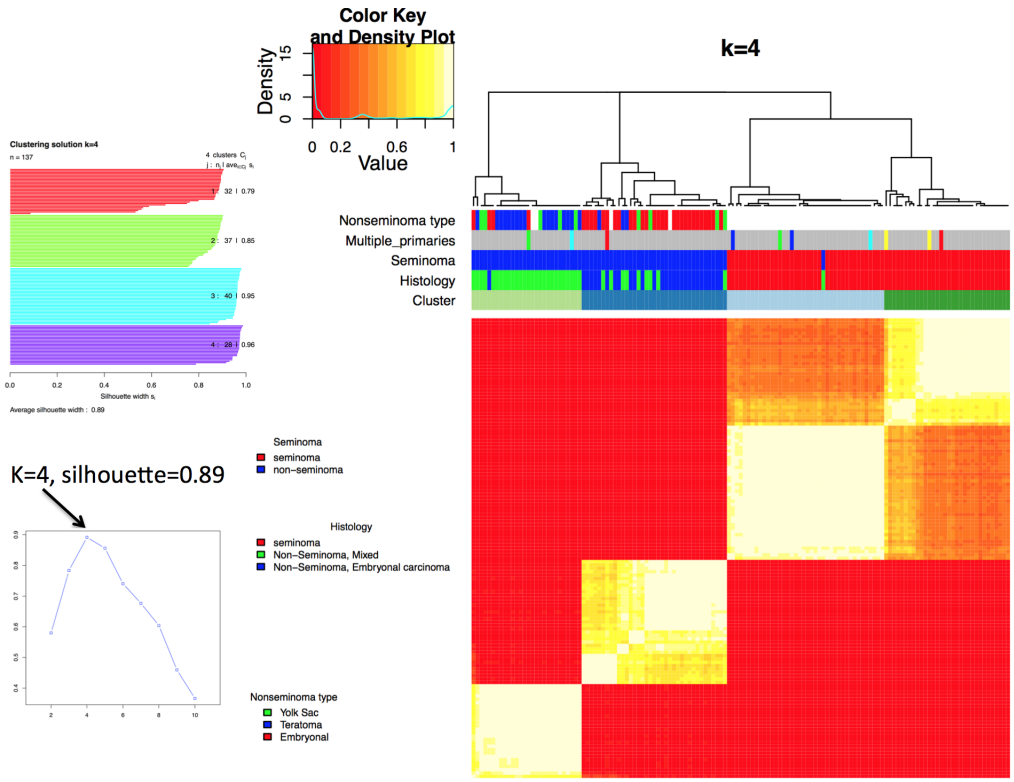


Figure 3.12: Results of the unsupervised analysis of TGCT cohort in PARADIGM space (based on top 3000 most varying IPLs). The molecular subtypes highly correlate with the histological subtypes.

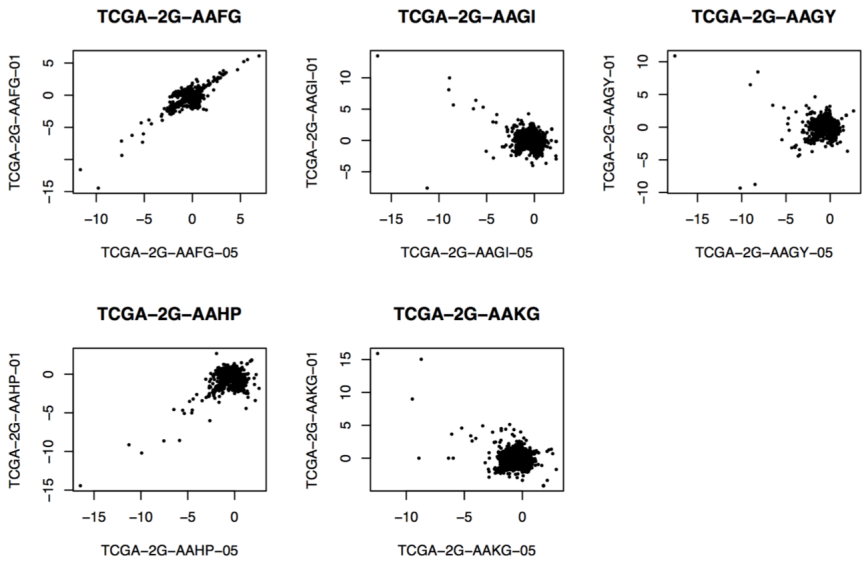


Figure 3.13: Comparison of match primaries in PARADIGM space. Each first primary was correlated with its corresponding match primary. Each dot in the plot represents a single PARADIGM IPL. There are 5 plots for the 5 individuals with match primaries.

3.3.3 Integrated Molecular Space Reveals Important Relationships Between Histological Subtypes

I built visualization of the integrated TGCT molecular space and relationships between the tumors in the cohort using Tumor Map 3.14 method. We integrated RNA-Seq mRNA expression somatic copy number and methylation spaces into a single view and built an integrated map of the TGCT tumors. I used mRNA expression data from RNA-seq (n = 137), copy number Gistic calls (n = 137), and methylation profiles from HumanMethylation450 (HM450) platform (n = 137) to built this multi-platform map. I found that my unsupervised analysis identified groups that closely correlated with histology classification. Using the Tumor Map view, I found that seminoma and non-

seminoma tumors exhibit clear separation in the molecular space and are very distinct from each other. Furthermore, embryonal non-seminoma tumors are very distinct from other non-seminomas. The same clear separation of histological groups was observed during analysis of individual molecular spaces using other unsupervised methods as well. I also see that the KIT mutants within seminomas cluster together. KIT mutation is specific to seminoma tumors and this separation in molecular space is consistent with other unsupervised analysis of seminoma transcriptomic and genomic space.

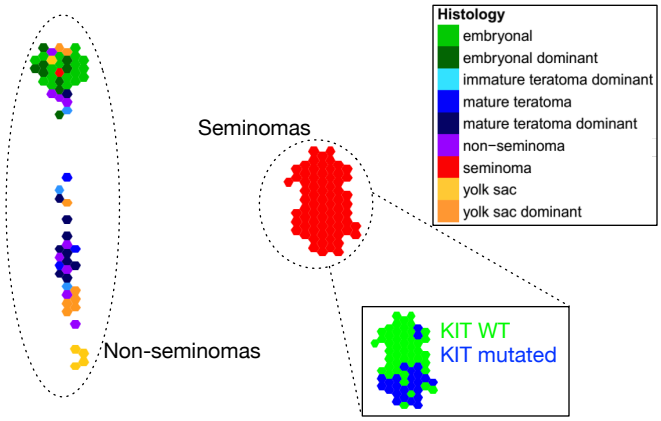


Figure 3.14: Tumor Map view of combined mRNA expression, somatic copy number, and methylation spaces of TGCT cohort. The map shows almost perfect separation between the major histological subtypes.

3.3.4 Analysis of Pathway Activity Space Helps Characterizing Histological Subtypes

As the histological subtypes turned out to be the biggest separators of the cohort unsupervised subtypes in every molecular space the AWG had looked at, the experts in the group made a decision to analyze the TGCT cohort based on the histological

subtypes and not based on unsupervised molecular subtypes.

3.3.4.1 Characterization of Histological Subtypes Through Unsupervised Analysis

I ran pathway activity inference analysis using PARADIGM method and performed analysis of pathways activities differential in each of the tumor grouping. From PARADIGM results I extracted top 3,000 most varying Interred Pathway Levels (IPLs). Within each of the histological groups of samples I performed unsupervised hierarchical clustering of the IPL profiles of those samples. I also performed unsupervised hierarchical clustering of the 3,000 IPLs (Figure 3.15). I identified seven major pathway activity clusters in the histological subtypes. All seminomas exhibit higher proliferation and cell cycle signaling, specifically KRAS and E2F pathway activities. They are also enriched in various components of the immune signaling, such as TNFA via NFKB, TCR pathway, interleukins, and inflammation. In contrast, molecular signaling in non-seminomas is less uniform and more specific to the non-seminoma histological subtypes. All non-seminomas are enriched in such carcinogenic markers as high Wnt and MYC signaling, as well as hypoxia and myogenesis. Embryonal tumors exhibit higher GPCR signaling, while non-embryonal non-seminomas show high FOXA2/3 transcription factor network activity, mTOR signaling, and oxidative phosphorylation. Finally, high AR signaling appears to be specific to teratoma tumors.

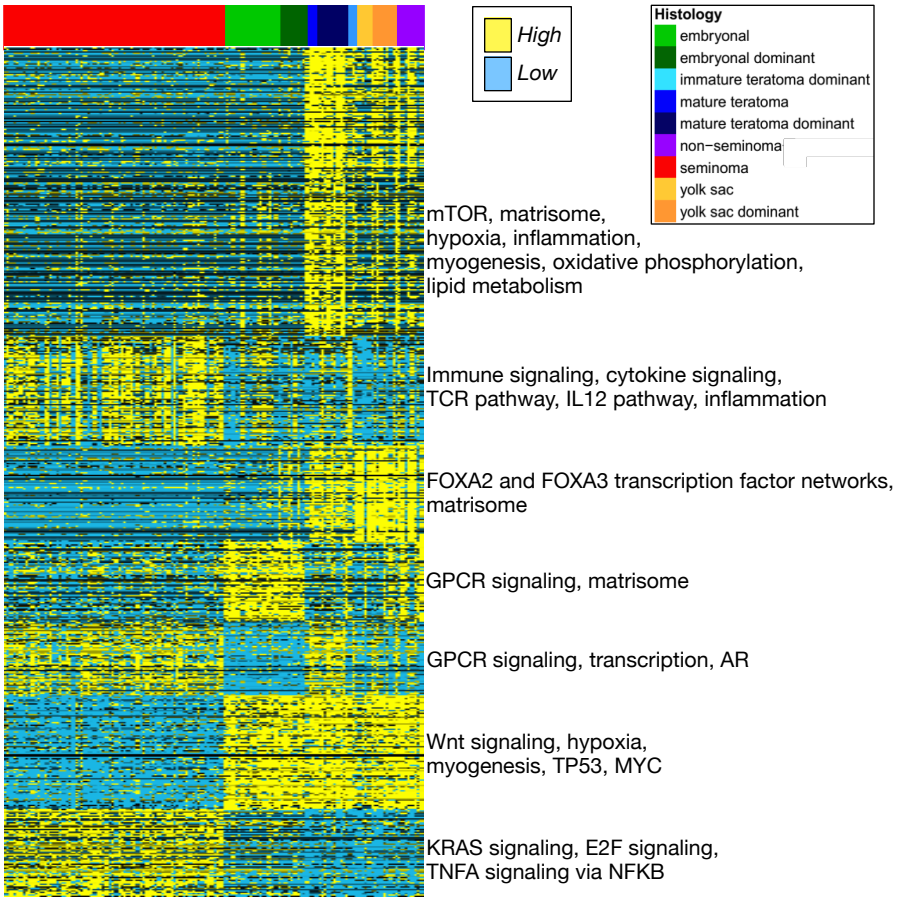


Figure 3.15: View of TGCT histological subtypes in PARADIGM IPL space.

3.3.4.2 Epithelial to Mesenchymal Transition Pathway Plays Important Role in Mixed Nonseminoma Tumors

Mixes non-seminomas were defined as non-seminoma tumors of non-embryonal histology. Because non-seminomas are generally more aggressive tumors and there is a clear molecular separation of the non-seminoma groups, the group had an interest in identifying cellular mechanisms involved in these particular malignancies. I computed

a differential between the IPLs within the mixed nonseminoma group to IPLs in other TGCT samples. I used this differential signature to perform pathway enrichment analysis using Gene Set Enrichment Analysis (GSEA) method [124]. I found that Epithelial To Mesenchymal Transition (EMT) pathway is one of the most enriched pathways in this signature (Figure 3.16 on the left). I also confirmed the statistical significance of this enrichment by aggregating IPLs for the genes that belong to the EMT pathway and plotting the aggregated levels within each histological group (Figure 3.16 on the right). I pursued this finding further to identify the specific elements that might be driving this signaling. I used the differential signature and used a modified PATHMARK [133] method to pull out interesting molecular pathways directly involved in EMT signaling in these particular tumors (Figure 3.17). Some of the things that jump out in this pathway are high activity and "hubbiness" of PIK3CA gene, a known cancer-associated gene. This gene, along with AURKA and together with inhibition by MDM2, activates TP53 activity. It is also notable that NME1, a metastasis suppressor gene, shows lower activity in the mixed nonseminoma tumors. This could explain some of the aggressive behavior in these tumors.

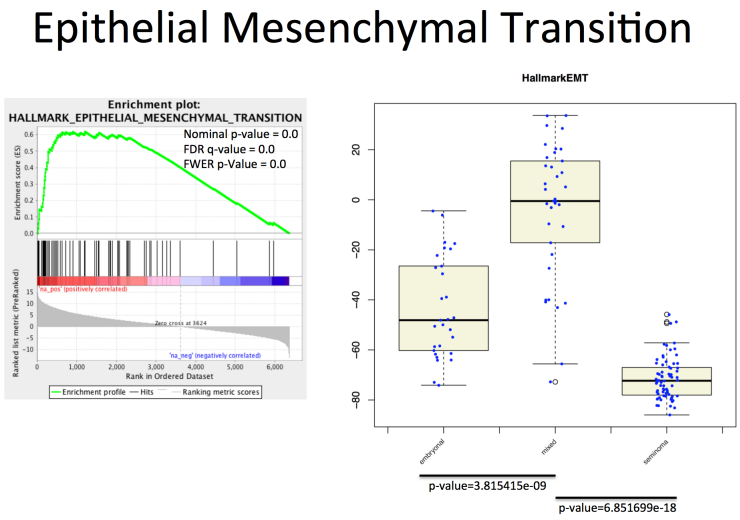
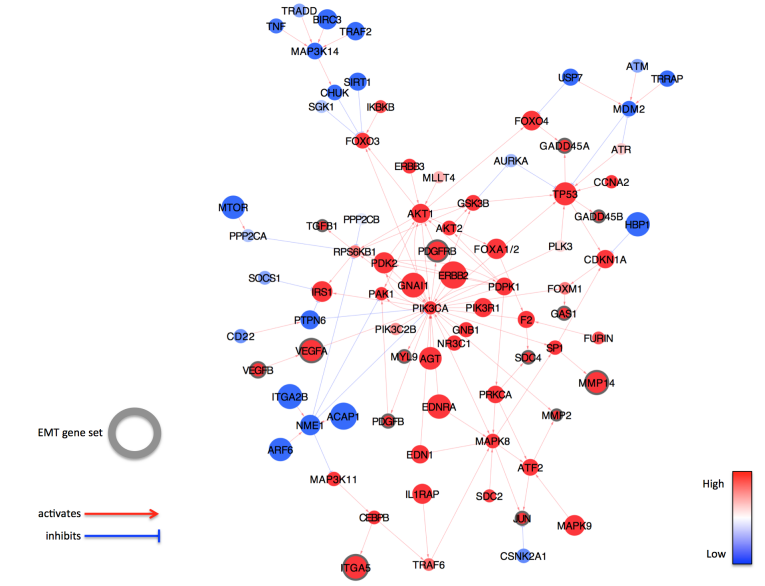
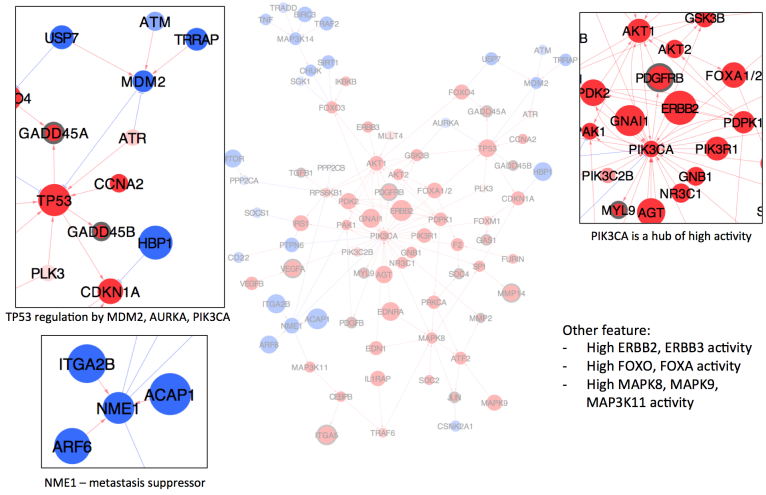


Figure 3.16: Epithelial to mesenchymal transition is enriched in the mixed non-seminomas. On the left: GSEA plot for the EMT pathway based on the differential IPL signature. On the right: IPLs for the genes in the EMT pathway were aggregated per-sample and the aggregated levels are plotted based on the histological group.



(a) EMT pathway.



(b) Closer look.

Figure 3.17: Epithelial to mesenchymal transition pathway identified by the PATHMARK method as relevant to mixed nonseminoma tumors.

3.3.4.3 ERBB Pathway Plays Important Role in Mixed Nonseminoma Tumors

Together with the EMT signaling, I also examined ERBB signaling in mixed nonseminoma tumors. While I saw some of the ERBB pathway genes in the previously described EMT section, I saw that ERBB signaling alone appears to be enriched in the mixed non-seminomas (Figure 3.18). I used the differential IPL signature, described above, and pulled out ERBB-related pathways. I found that quiet a few ERBB-related genes and complexes exhibit high activity in these mixed nonseminoma tumors (Figure 3.19(a)). However, we were unable to identify from this pathway any major activators of the ERBB pathway, except for ADAM17 gene, which is not a known regulator in TGCTs. I hypothesized that there is another mechanism by which ERBB signaling is activated in these cells, possibly miRNA activation. I used the aggregated per-sample ERBB activity levels, described in the above section, as an ERBB activity signature and correlated it to the expression of every miRNA (Figure 3.19(b)). I found a significant anti-correlation, as would be expected in the case of activation by miRNAs, with many known miRNA activators of ERBB signaling. For example, miR-125b, miR-205, miR-7, miR-331-3p, miR-148b, miR-149, miR-326 and miR-520a are all known activators of ERBB2. This supports my hypothesis that ERBB signaling in TGCTs is driven by miRNAs. miRNA nodes are under-represented in the Superpathway [133] used by our modified PATHMARK analysis. Therefore, the pathway we pulled out lacks the true representation of the ERBB activation mechanism.

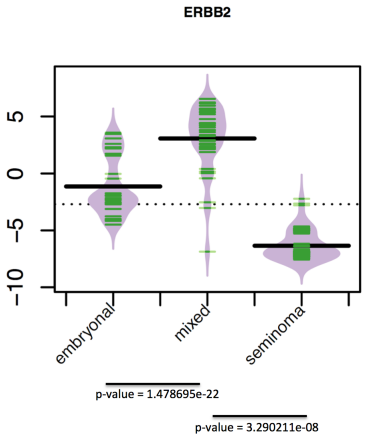
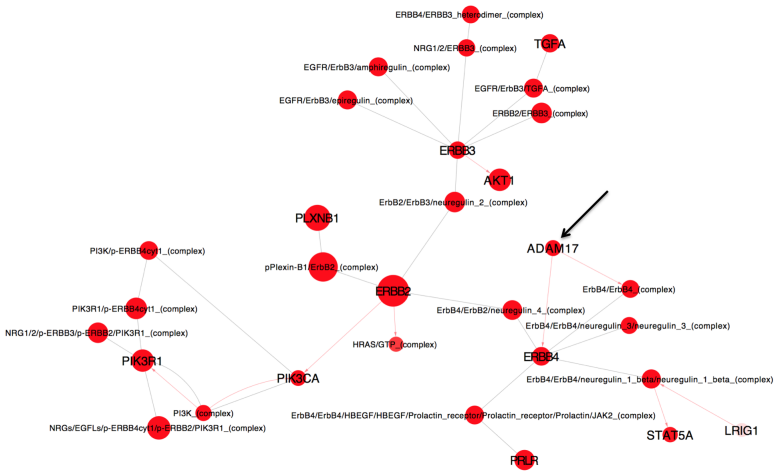


Figure 3.18: ERBB signaling is enriched in the mixed non-seminomas as compared to other groups.



(a) ERBB pathway.

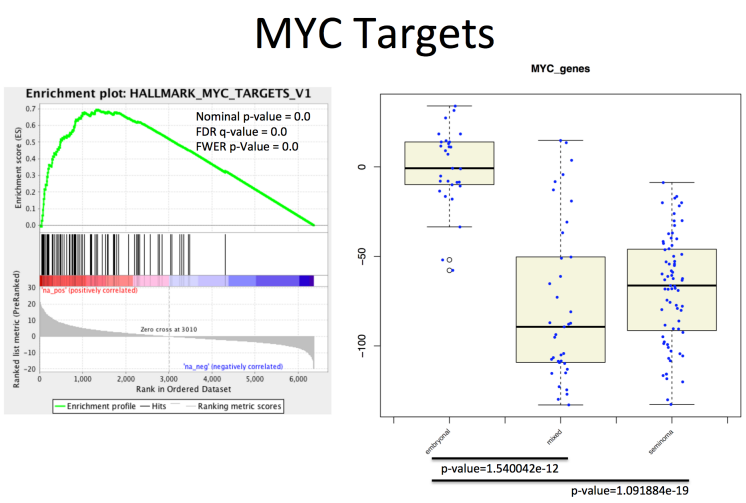
miRNA	correlation	pvalue	FDR
miR-125b-1-3p	0.620482085	2.16E-15	2.36E-14
miR-125b-2-3p	0.591673493	7.98E-14	6.53E-13
miR-205-5p	0.587974099	1.24E-13	9.99E-13
miR-148b-5p	0.556453591	4.28E-12	2.90E-11
miR-125b-5p	0.545104955	1.40E-11	8.57E-11
miR-7-2-3p	0.428543306	8.73E-07	2.77E-06
miR-205-3p	0.347236002	4.78E-05	0.000122007
miR-148b-3p	0.230432083	0.007856993	0.014249122
miR-331-5p	-0.227113585	0.008823109	0.015783527
miR-7-1-3p	-0.553103787	6.10E-12	4.01E-11

(b) miRNAs vs. ERBB signature.

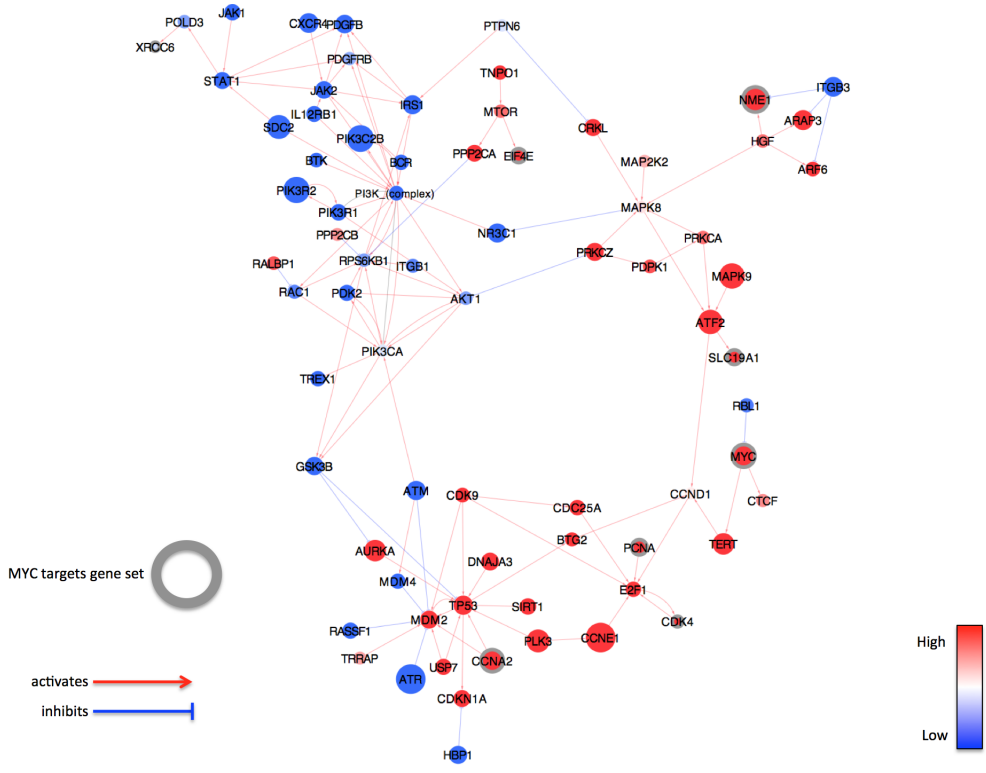
Figure 3.19: ERBB pathway identified by the PATHMARK method as relevant to mixed nonseminoma tumors.

3.3.4.4 MYC Pathway Plays Important Role in Embryonal Nonseminoma Tumors

I also analyzed pathway enrichments in embryonal nonseminoma tumors. I computed a differential IPL signature of embryonal nonseminoas vs. all other tumors. As described above for the mixed nonseminomas, I performed GSEA of this signature as well as the aggregated IPL level analysis. I found that MYC targets is one of the most enriched pathways in this embryonal signature (Figure 3.20(a)). I applied our modified PATHMARK method to pull out MYC-related genes for this signature (Figure 3.20(b)). I found that some of the big cancer players exhibit high activity in these tumors (e.g. PLK3, AURKA, TERT, and many more). Interestingly, NME1 activity appears to be up in embryonal tumors, possibly suggesting a different metastasis potential compared to the mixed tumors. On the other hand, TP53 activity in embryonal malignancies is also high, supporting what was previously known about TP53 activity across all nonseminoma tumors.



(a) MYC targets enrichment.



(b) MYC targets pathway.

Figure 3.20: MYC pathway identified by the PATHMARK method as relevant to embryonal nonseminoma tumors.

3.3.5 Conclusion

We analyzed a cohort of testicular germ cell tumors to derive and describe molecular subtypes of this cancer type. The group found that every platform recapitulated histological subtypes with such high accuracy and robustness that the decision was made to analyze this tumor in the context of historical labels rather than derived major molecular subtypes. We show that integrated Tumor Map view is also driven by histology. We described molecular differences between the histological subtypes. In particular, we identified molecular and pathway markers of the nonseminoma tumors, which is a more aggressive subtype of TGCT. We also found that match primaries are no more similar to each other than they are to another primary TGCT tumor.

3.4 Analysis of Mesothelioma

Mesothelioma (MESO) is a cancer that develops in the lining of the internal organs, most often originating in the lining of lungs and chest wall. However, it also occurs in the lining of abdomen, testis, and surroundings of heart. This cancer is often associated with asbestos exposure.

TCGA MESO AWG analyzed the output of a number of genomic platforms for 75 tumors classified as varying histological subtypes mesothelioma. We performed PARADIGM [133] analysis on these tumors and discovered four PARADIGM-based subtypes. Two of these subtypes exhibited statistically significant survival difference. I followed up with unsupervised analysis of the PARADIGM IPL profiles (see 3.4.2).

I performed further analysis to identify genes and pathways differential between the good survivors and the bad survivors (see 3.4.4). I also performed unsupervised analysis of eight different tumor types, including MESO tumors, to identify which other tumor types these samples are most similar to (see **??**). The group is in the process of completing the manuscript.

3.4.1 Pathway Activity Space Reveals Prognostically Significant Molecular Subtypes of Mesothelioma

I utilized PARADIGM [133] method to infer pathway-level activities of various members of biological and metabolic pathways. One of the collaborators (Dr. Graim) performed unsupervised hierarchical clustering of the MESO samples in the PARADIGM IPL feature space. The best solution (k = 4) significantly separates the cluster groups by survival (Figure 3.21B). The two particular groups of interest are the "good" survivors (best surviving group) and the "bad" survivors (worst surviving group), the two extreme survival groups. One of the interesting biological questions the AWG went after is finding the molecular and clinical differences between these groups. Identifying molecular markers that describe these two groups is important because of possible diagnostic, prognostic, and therapeutic implications. Graim and I found that the poor survivors had a high MYC and TP53 signaling, as well as being more proliferative and aggressive and exhibiting elevated immune activity. Although PARADIGM clustering solution showed the best survival separation out of all the platforms analyzed for this tumor type, we still want to understand how molecular subtypes in other plat-

forms compare to PARADIGM. Specifically, we wanted to see how the best and the worst surviving groups in other platforms compare to the best and the worst surviving groups in PARADIGM in terms of sample memberships. I defined the "best" and the "worst" surviving groups in each of the platform based on the Kaplan Meier plots (Figure 3.22). I compared the sample memberships of the PARADIGM solution's best (blue) and the worst (red) survivors to the memberships of the same samples in other platforms (Figure 3.21C); I combined all the intermediate survivors into a single group "Other". I computed the "distance" between each pair of groups (best vs. best or worst vs. worst in PARADIGM vs. another platform) as Jaccard Index ($\frac{A \cap B}{A \cup B}$), showing in Figure 3.21C on the right of the cluster membership heatmap. I found that another integrative analysis method iCluster [117] closely recapitulates the best and the worst surviving subtypes, which adds to the confidence to our solution and suggests that multiple integrative analysis methods are able to come to similar answers to the problem of molecular subtyping. Some individual platforms are very good at recapitulating the integrative molecular subtypes. For example, lncMRNA platform's best and worst surviving groups closely track with the PARADIGM best and worst surviving groups. We also found that some platforms closely recapitulate one of the PARADIGM extreme surviving groups but not the other. E.g. the best surviving group in the copy number space highly correlates with the best surviving group in PARADIGM space while the worst surviving groups in these spaces have very little overlap. This suggests that the molecular changes driving poor survival are not probably driven by copy number alterations and require other regulatory and possibly more complex explanations.

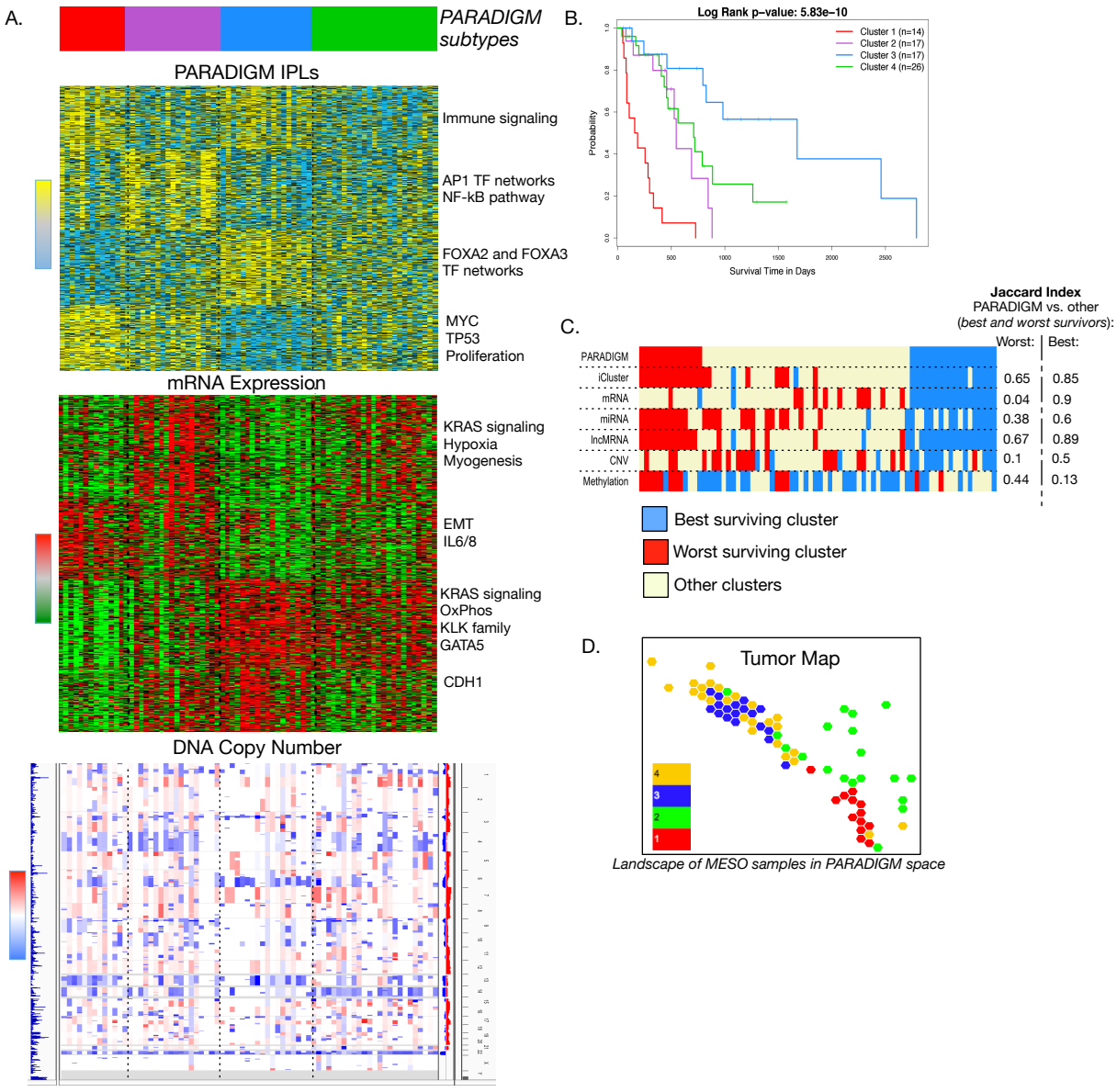


Figure 3.21: Description of the MESO cohort clusters in PARADIGM IPL space. A) PARADIGM clustering solution (k = 4) and relevant feature spaces (PARADIGM IPL, mRNA expression, and CNV) under that solution. All samples are ordered by the cluster annotations at the top. B) Kaplan Meier plot of survival (in days) of the sample groups from the PARADIGM clustering solution. C) Best (blue) and worst (red) surviving clusters across all the platforms are compared. Jaccard Index is computed separately for the best and the worst surviving groups as a measure of distance between that and the corresponding group in PARADIGM solution.

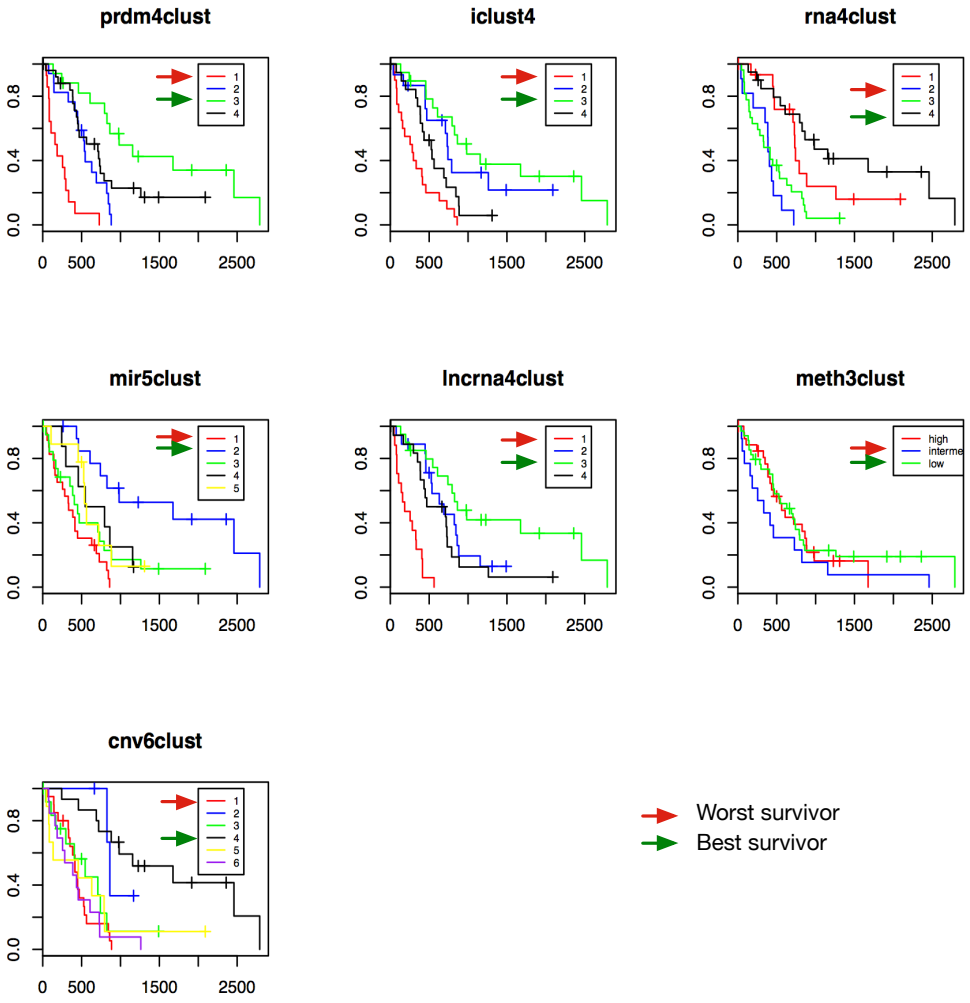


Figure 3.22: Best and worst surviving groups for each of the platforms analyzed for MESO cohort as defined by the Kaplan Meier (KM) survival plot for each platform. Each survival plot was produced by the individual collaborators working on that platform. KM plots from each platform are combined into a single figure here and best and worst surviving groups are indicated with the corresponding arrows.

3.4.2 Molecular Space Shows Important Spacial Separation Of Prognostically Important Subtypes

We used PARADIGM IPLs as features and performed unsupervised analysis of the MESO tumors in PARADIGM space using Tumor Map method. We discovered that histological subtypes are not well separated by their PARADIGM IPL profiles (Figure 3.23). However, the good surviving PARADIGM subtype is clearly enriched in epithelioid histology.

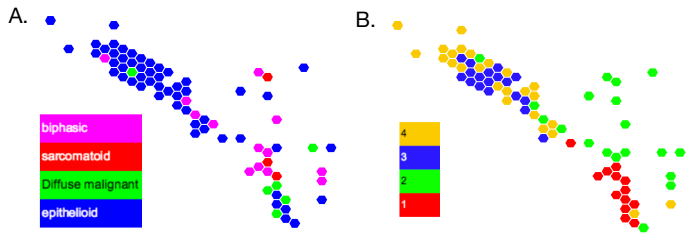


Figure 3.23: Mesothelioma tumors clustered in PARADIGM space using Tumor Map. Each tumor is a node in the map. The nodes are laid out based on similarities of their PARADIGM IPL profiles. A) The samples are colored by histological labels. B) The samples are colored by PARADIGM subtypes. The good surviving cluster is 3 (blue) and the worst surviving cluster is 1 (red).

3.4.3 Comparison of Two Most Differentially Surviving Molecular Subtypes Reveals Important Markers of Aggresiveness In Poor Survivors

I analyzed the difference between the good and bad surviving tumors in the PARADIGM IPL space in order to identify the differences between the good and the bad surviving groups of tumors. I ran differential analysis using LIMMA method and

filtered out all IPLs that did not have Bonferroni adjusted p-value less than or equal to 0.05. I then only considered protein-coding IPLs and separately analyzed positive and negative differential groups of genes with MSigDB. I identified that good survivors have elevated ERBB signaling while the bad survivors appear to exhibit more proliferative (E2F and G2M pathway, mitosis signaling), growth (glycolysis, PLK1 pathway) and aggressive (EMT pathway) signaling (Figure 3.24).

I also scanned across canonical pathways (c2cp set from MSigDB) and identified significantly differential pathways between the good survivors (3) and bad survivors (1) (Figure 3.25). For each pathway and each sample, I extracted only those protein-coding IPLs that belong to the pathway and aggregated all IPLs for that pathway by summing across for this sample. I repeated this for each patient and computed Welch's t-test between the distribution of aggregated IPLs for the samples in cluster 1 and cluster 3 to identify those pathways that exhibited significant difference in aggregated pathway activities between the two groups. The pathways identified with this analysis concurred with the results of the differential analysis of individual IPLs described above.

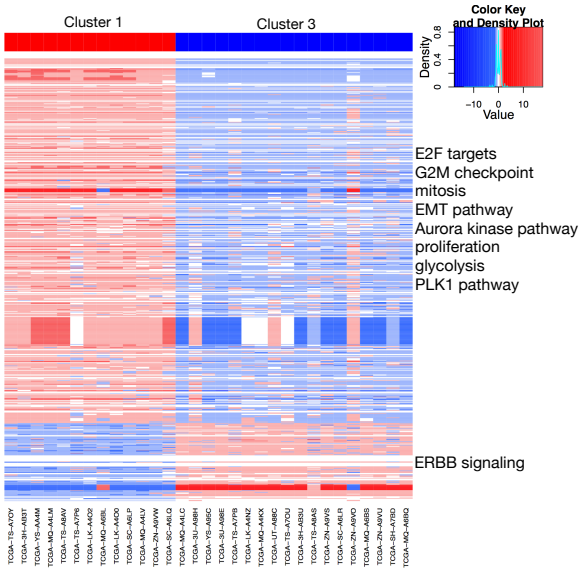


Figure 3.24: Analysis of differential activities in the good survivors (cluster 4/blue) and bad survivors (cluster 1/red) groups. The rows of the heatmap are statistically significant protein-coding PARADIGM IPLs. The pathway enrichment analysis of positive and negative differential IPLs was performed using MSigDB resource.

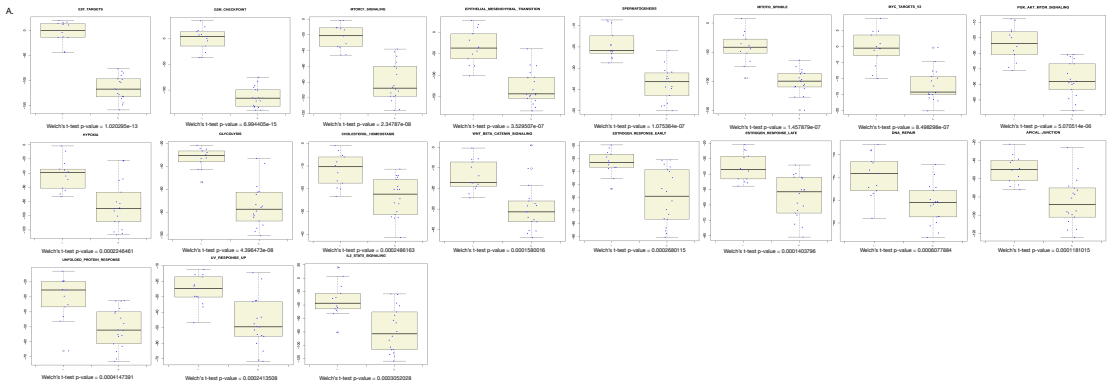


Figure 3.25: Analysis of differential aggregated activities within statistically significantly differential pathways between cluster 1 (bad survivors) and cluster 3 (good survivors).

3.4.4 Pan-cancer Analysis Reveals Sarcoma-like Subtype of Mesothelioma Tumors

I utilized the Tumor Map method to perform unsupervised joint analysis of RNA-Seq mRNA expression data fro 8 different cancers (per expert suggestion from the group only basal breast carcinoma tumors were used from that tumor type cohort). I combined mRNA expression RNA-Seq data for eight different cancer types, including MESO tumors (Figure 3.26). I found that most MESO tumors cluster near a group of sarcoma tumors (Figure 3.26A). Furthermore, some MESO tumors mix with the sarcoma tumors in this map and those sarcoma-like tumors are enriched for biphasic and sarcomatoid histology (Figure 3.26B). At a closer look, the sarcoma group near and with which MESO tumors cluster predominantly exhibits undifferentiated phenotype, or enriched for dedifferentiated liposarcoma, undifferentiated pleomorphic sarcoma, and myxofibrosarcoma histological subtypes (Figure 3.26C). I also found that most MESO tumors clustering with undifferentiated sarcomas belong to the bad surviving PARADIGM subtype (Figure 3.26D). Finally, I found a handful of MESO samples that cluster with tissues other than sarcoma or mesothelioma (Figure 3.27).

In order to understand what is different about sarcoma-like tumors from other MESO tumors, I applied Tumor Map reflection method to identify genes driving the placement of the sarcoma-like MESO tumor and compared them to genes driving the placement of the non-sarcoma-like MESO tumors. The reflection method identifiers top 150 up-regulated and 150 down-regulated genes contribute the most to the placement of

a selected group of samples. Out of concern that non-sarcoma-like tumors are enriched for epitheliod histology (Figure 3.28) and the differential signaling we detect will be overwhelmed by the epithelial signaling, I also applied the analysis within epithelioid tumors only. I performed Tumor Map reflection on 4 groups: sarcoma-like MESO tumors, non-sarcoma-like MESO tumors, epithelioid-only sarcoma-like MESO tumors, and finally epithelioid non-sarcoma-like MESO tumors. I extracted two lists of genes for each of the reflections (UP and DOWN genes). I then compared UP-only and DOWN-only gene lists separately. For each of these two comparisons I overlapped the four gene lists from each of the group (Figure 3.29A) and concentrated on those genes that were common to sarcoma-like MESO tumors and epithelioid-only sarcoma-like MESO tumors (n=52). I performed MSigDB enrichment analysis on those genes and found that sarcoma-like MESO tumors exhibit such elevated aggressiveness and de-differentiation signaling as EMT pathway, hypoxia, and myogenesis (Figure 3.29B). Analysis of DOWN genes obtained through reflection method did not produce any significant enrichment.

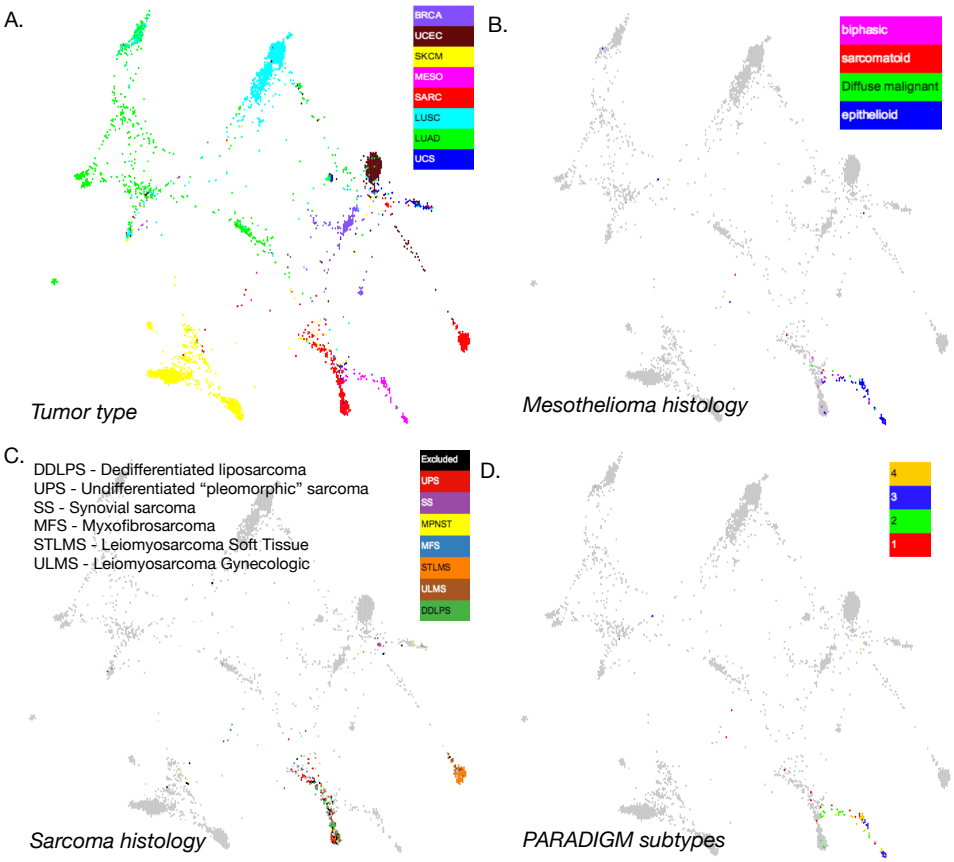


Figure 3.26: Multi-tumor RNASeq data clustered using Tumor Map. There are 8 different types of tumors in this map (per expert suggestion from the group only basal breast carcinoma tumors were used from that tumor type cohort). A) Tumors are colored by tumor type. B) Tumors are colored by mesothelioma histology (other tumors are gray). C) Tumors are colored by sarcoma histology (other tumors are gray). D) Tumors are colored by MESO PARADIGM subtypes (other tumors are gray).

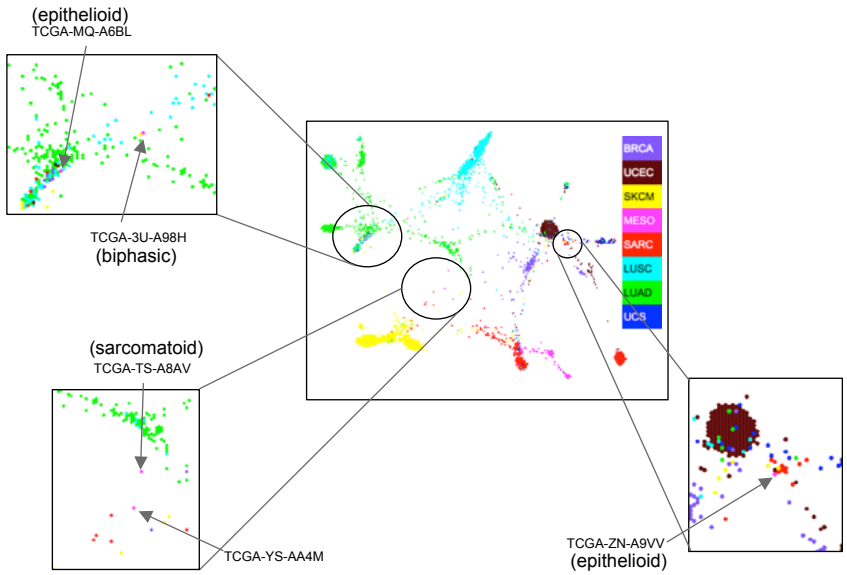


Figure 3.27: Mesothelioma samples that do not cluster with either other mesothelioma tumors or sarcoma tumors in the multi-tumor RNASeq data clustered using Tumor Map that incorporates 8 different types of cancer (per expert suggestion from the group only basal breast carcinoma tumors were used from that tumor type cohort).

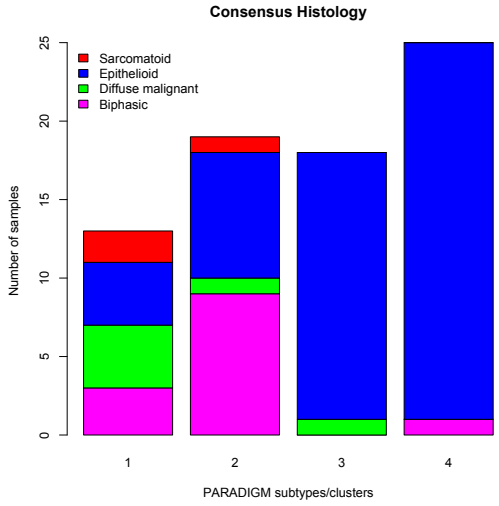
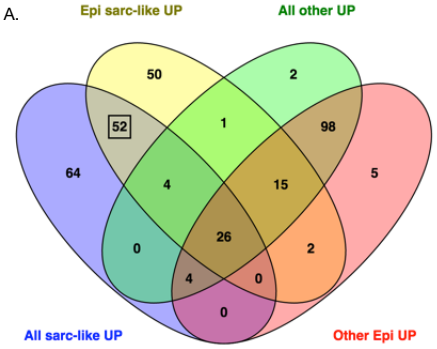


Figure 3.28: Distribution of histology labels across the PARADIGM subtypes (cluster 3 are the good survivors; cluster 1 are the bad survivors).

A.



B.

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION [200]	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.	14		$4.5\ e^{-22}$	$2.25\ e^{-20}$
HALLMARK_APICAL_JUNCTION [200]	Genes encoding components of apical junction complex.	4		$6.85\ e^{-5}$	$1.14\ e^{-3}$
HALLMARK_HYPOXIA [200]	Genes up-regulated in response to low oxygen levels (hypoxia).	4		$6.85\ e^{-5}$	$1.14\ e^{-3}$
HALLMARK_UV_RESPONSE_DN [144]	Genes down-regulated in response to ultraviolet (UV) radiation.	3		$5.3\ e^{-4}$	$6.63\ e^{-3}$
HALLMARK_MYOGENESIS [200]	Genes involved in development of skeletal muscle (myogenesis).	3		$1.37\ e^{-3}$	$1.37\ e^{-2}$

Figure 3.29: Reflection analysis of sarcoma-like MESO tumors and the UP genes that drive the similarity of those tumors to undifferentiated sarcomas. A) Venn diagram shows overlaps of UP genes from four reflection analyses (sarcoma-like MESO tumors, non-sarcoma-like MESO tumors, epithelioid-only sarcoma-like MESO tumors, and epithelioid non-sarcoma-like MESO tumors). B) MSigDB enrichment analysis of the 52 genes overlapped between epithelioid-only sarcoma-like tumors and all sarcoma-like tumors.

3.4.4.1 Tumor Map Reflection

Motivation

Tumor Map computes spatial relationships between entities in the map based on the feature space that defines genomic profiles of those entities. When constructing a sample map, the feature space consists of molecular features from a particular data type. One can think of the transpose of such a map, in which the map entities are molecular features and the feature space consists of the samples in the cohort. Molecular features in this transposed map are positioned based on how similarly they behave across the cohort of samples. In other words, features grouping together may tend to be high or low in the same samples. This concept also applies to an integrated map containing molecular features from a variety of omics platforms constructed as described above.

In the transpose of an integrated map, all the single-platform molecular features are combined to enable identification of clusters of features that behave similarly regardless of their platform of origin.

Given such a transpose map of molecular features, it is possible to flip between the sample map and the molecular features map to visualize the link between entities in the two maps. We call this map flip a Reflection and we describe our method below.

Reflection Method

The Tumor Map Reflection functionality provides a transition between two maps, namely a map of samples (Sample Map), referred to in this manuscript as Tumor Map, and a map of genes, or Gene Map, which plots genes instead of tumor samples according to the DrL layout method described above. The goal of map reflection is to link groups of samples with groups of genes. A link between groups of samples that cluster together on the Sample Map and a group of genes in the Gene Map suggests that those are the genes driving that particular grouping of samples. Viewing the reflection of a group of samples to a Gene Map highlights genes representing the top and bottom extremes of expression summarized across the selected samples. If starting on a Gene Map, viewing the reflection of a group of genes to a Sample Map highlights samples in which the chosen genes are at their extreme values (up and down) across the sample cohort.

The Reflection functionality operates on a modified genomic matrix (e.g. gene expression matrix), R. To produce R, consider a typical genomic matrix E from which two additional matrices E^R and E^C will be created,. E^R and E^C are produced by

z-score normalizing the rows and columns of matrix E, respectively, where rows are features (e.g. genes) and columns are observations (e.g. tumor samples). Cells of E^R and E^C are then combined using the Euclidean distance function to produce matrix R, where each cell R_{ij} is computed as following:

$$R_{ij} = \sqrt[2]{(E_{ij}^R)^2 + (E_{ij}^C)^2} \tag{3.1}$$

Application of the Euclidean distance function results in loss of the notion of the sign for each feature (e.g. up vs. down regulation of gene expression) in R. Euclidean distance is always a positive value. To restore the sign information for the features of R, we compute two standardized z-scores (E_{ij}^R and E_{ij}^C) for each E_{ij} cell in R, using the row and the column as a background distribution accordingly. The agreement of the signs of the z-scores from each corresponding cell of E^R and E^C is examined. If the signs of E_{ij}^R and E_{ij}^Cagree, then the value in the corresponding cell of R_{ij} is given the sign of E_{ij}. If the signs in E_{ij}^R and E_{ij}^Cdo not agree, then the value in R_{ij} is set to 0. The signs may not agree in cases where the feature has an extreme (high or low) value across all the observations but not across all the features in that particular observation or vice versa.

For our continued discussion of the Reflection functionality below, we refer to features (e.g. genes) and observations (e.g. samples) by the more general term nodes, and we generalize Sample Maps and Gene Maps to source maps and target maps, respectively.

The operation applied to the modified genomic matrix R is a t-test-like function. A group of nodes S is selected on the source map. Each node t_i in the target map is scored by contrasting values in group S with the background distribution of the union of S and S (nodes in the source map not in S). For each t_i $\forall i, i \in [1, |T|]$, where $|T|$ is the cardinality of the target map, we compute the following score:

$$t_i = \frac{mean(t_i^S) - mean(t_i^{S \cup S'})}{std(S \cup S')} \tag{3.2}$$

This results in a ranked list of target map nodes. One can think of these rankings as scoring each node of the target map based on how extreme they are on average among nodes in S as compared to all of the nodes in the source map. We then reduce this list to the 150 highest- and lowest-scoring nodes, and these 300 nodes are highlighted on the target map.

The concept of a map reflection is not restricted to mRNA expression data as some of our examples might suggest. For instance, instead of a t-test-like operation, a frequency operation could have been used on a matrix representing binary mutation data. In that case, the reflection from a Sample Map to a Gene Map would highlight genes with the highest mutation frequency in the selected samples. Similarly, more complex reflection functions can be considered, such as activity-based or pathway-based summaries.

3.4.5 Conclusion

We analyzed a cohort of mesothelioma tumors to identify and describe molecular subtypes of this cancer. We found that pathway activity view (PARADIGM IPLs) subtypes provide the best separation of molecular subtypes in survival space. We found four molecular subtypes in the mesothelioma cohort. We also described these subtypes and found, as might be expected, that the worst surviving mesothelioma group exhibits many of the known markers of proliferation and aggressiveness. Interestingly, the good surviving group is significantly associated with pneumonectomy, a surgical removal of a whole or a partial lung. This suggests that surgical intervention has a significant effect on long-term survival and progression of individuals with mesothelioma.

We performed a cross-cancer analysis of RNA-Seq data of MESO and 7 additional tumor types in order to identify tumor groups and signatures that cross cancer type boundaries. We describe a group of MESO samples that is very similar to undifferentiated sarcoma tumors. This suggests that these particular MESO tumors exhibit undifferentiated phenotype in their gene expression profiles. It also suggest stem-like phenotype of these tumors (stemness anti-correlates with differentiation). We used Tumor Map Reflection method 3.4.4.1 to obtain the genes driving these similarities.

3.5 Analysis of Sarcomas

Sarcoma (SARC) is a cancer that arises in the cells of mesenchymal origin and can originate in many different tissues and are generally rare malignancies compared to

carcinomas.

TCGA SARC AWG analyzed the output of a number of genomic platforms for 237 tumors classified as varying histological subtypes sarcomas. The AWG organized their analysis of this type of tumor by histological subtypes (DDLPS - Dedifferentiated liposarcoma, UPS - Undifferentiated pleomorphic sarcoma SS - Synovial sarcoma, MFS - Myxofibrosarcoma, STLMS - Leiomyosarcoma Soft Tissue, ULMS - Leiomyosarcoma Gynecologic, MPNST - Malignant peripheral nerve sheath tumor). Tumor Map analysis of the RNA-Seq data confirmed previous analyses of RNA-Seq data by the AWG and showed that sarcoma tumors separate into three major groups (Figure 3.30). DDLPS, UPS, MPNST, and MFS tumors group together. All these tumors exhibit an undifferentiated cell phenotype. SS tumors appear to cluster separately from other sarcomas. Finally, STLMS and ULMS tumors cluster together, demonstrating that all leiomyosarcoma exhibit similar cell state.

3.5.1 Analysis Driven By Molecular Data Aids In Imputing Histology

A number of samples among these 237 tumors were excluded from the main manuscript analysis because confident histology calls were not possible for those samples. The group expressed interest in being able to speculate about the histology of those samples based on the RNA-Seq data. I performed unsupervised analysis of the RNA-Seq data of ten different tumor types using Tumor Map method (Figure 3.31). We identified that many samples labeled as "Exclude" cluster with sarcomas. I found that some exclude samples cluster with other tissue types. From these placements driven

by molecular profiles we can hypothesize about the cell of origin and histology of these ambiguous samples. For example, those sarcoma samples clustering with melanoma (SKCM) tumors are most likely melanoma tumors mis-labeled as sarcomas. The results presented to the group were received well by the lead pathologist in the AWG, with the intention of reviewing the biopsy slides for some of these samples. Some of these results confirmed what was already suspected by the experts on the team.

The group is in the process of completing the manuscript.

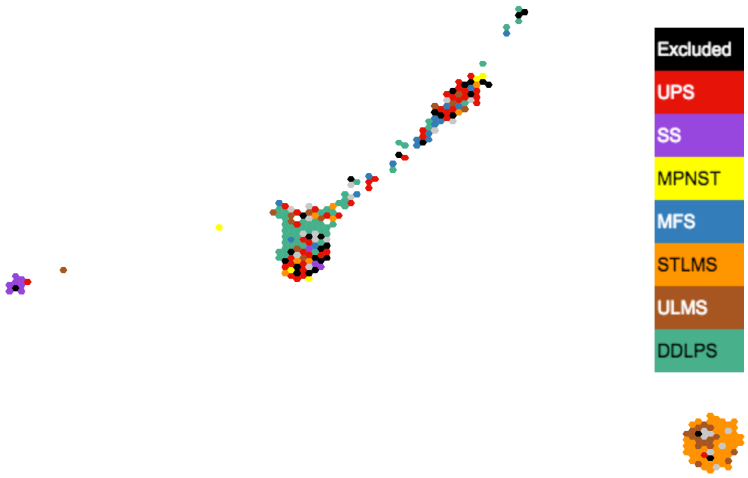


Figure 3.30: Sarcoma tumors in RNA-Seq space analyzed with Tumor Map show three major groups, driven by histological subtype groupings.

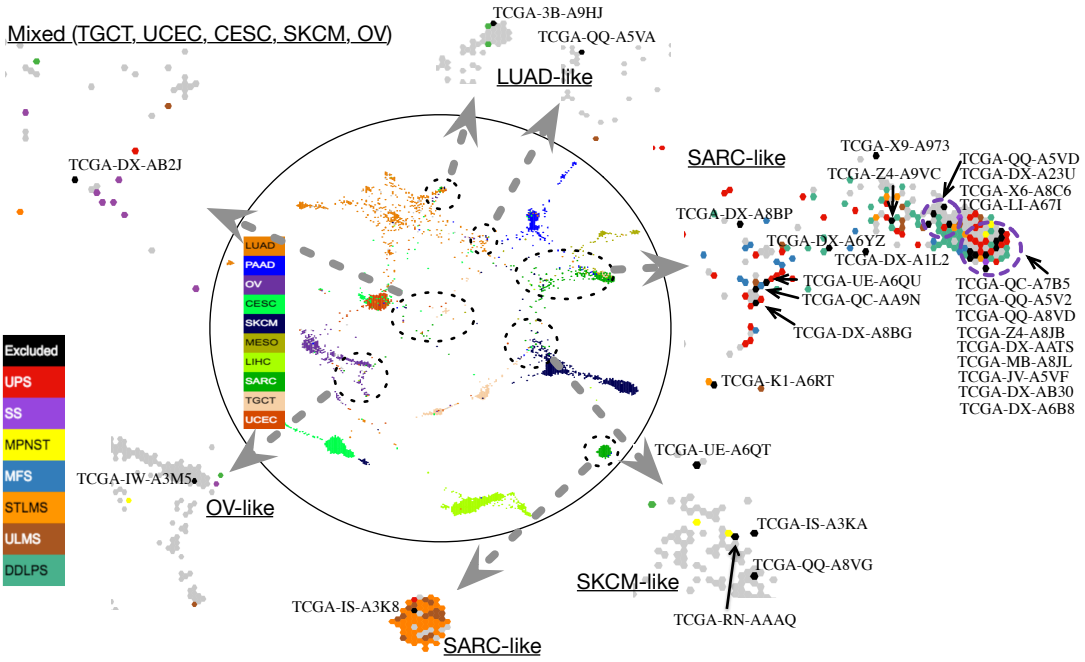


Figure 3.31: Multi-cancer analysis of ten different tumor types, including sarcomas, in RNA-Seq space using Tumor Map. The samples labeled as "Exclude" during the AWG analysis are marked and identified in the figure.

3.5.2 Conclusion

We analyzed a cohort of sarcoma tumors in the context of other cancers based on the RNA-Seq data in order to attempt to describe a number of samples excluded from analysis due to ambiguous histology. The experts suspected that some of the sarcoma-labeled tumors were actually from another cancer type (e.g. melanoma). We performed unsupervised analysis of 10 tumor types, including sarcomas, and were able to describe these Exclude samples. This analysis confirms that we can use Tumor Map method and other unsupervised analysis tools to let the molecular data tell us more

about the tumor samples than clinical and other annotations often provide.

3.6 West Coast Dream Team Analysis of Castration Resistant Prostate Cancer

As a part of the Stand Up To Cancer (SU2C) initiative, we became a part of the West Coast Dream Team working on castration resistant prostate cancer (CRPC). Some of these tumors exhibit small cell phenotype, a particularly distinct histology. CRPC are not the only tumors that exhibit this phenotype, which has been previously identified in lung, ovarian, and some others. In fact, small cell tumors are a part of the small-blue-round-cell tumor family. Adenocarcinoma is the most common type of primary prostate cancer and many CRPC tumors are also of the same histology type. It is yet unclear as to whether the small cell CRPC tumors arise from adenocarcinoma primary tumors or whether they have their own cell of origin. It is also not clear how the mixed histology tumors arise. As a part of my work with WCDT, I participated in several projects that aimed at answering some of those questions about the cell of origin and molecular mechanisms involved in development of hormonal therapy resistance. This section describes my work and contributions to those projects.

3.6.1 Identifying Molecular Subtypes Of CRPC And Defining Small Cell Phenotype Signature

I performed unsupervised analysis of mRNA expression data from RNA sequencing for 89 CRPC samples. Prior to the analysis I performed feature selection on the original 20,500-gene space (Figure 3.32). I first filtered out all the genes that were not expressed in more than 50% of the samples. Second, I computed per-gene variance and selected top 3,000 varying genes. I performed consensus k-means clustering [136], scanning from k = 2 to k = 10. I chose k = 8 based on the silhouette score method (Figure 3.33 top left). The heatmap in top right of Figure 3.33 shows how the samples are distributed among these 8 clusters. It also shows the distribution of the histological labels assigned to samples by a panel of pathologists among these 8 clusters. I computed the significance by isolating each histological label and performing Fisher Exact Test for that label across all the clusters. I computed FDR by performing Benjamini & Hochberg [139] multiple hypothesis testing correction on the Fisher Exact Test p-values. I found that cluster 5 is significantly enriched for small cell label. Clusters 3 and 7 are enriched in adenocarcinoma label, although cluster 3 to a lesser extent than cluster 7. Cluster 3 is significantly enriched for IAC histology, which is an intermediate histological subtype. And finally, cluster 1 is enriched for a mixed histology.

These findings confirmed that there is a signal in the expression data that can help identify molecular markers and drivers for different histological subtypes. I looked for those markers by performing unsupervised clustering of the feature space,

keeping the sample cluster dendrogram in place (Figure 3.34). I analyzed each gene cluster for their enrichments in various pathway sets from Hallmark (on the left of the heatmap) and Canonical Pathways (on the right of the heatmap) sets from MSigDB [78] pathway database. I found that the cluster enriched in small cell histology exhibits strong neural signaling, which has been previously associated with some small cell tumors. The fact that not all small cell tumors cluster together could be explained by a number of hypotheses. For example, small cell tumors may not all be uniform in their molecular signaling and there could be multiple subtypes of small cell tumors. It is also possible that the tumor biopsy had mixed histology and the part of the tumor that the pathologists analyzed had different cell composition than the one that went out for RNA sequencing analysis. Finally, it is possible that there was a mis-labeling by either the pathologists or post-processing team. Whatever the cause is, the molecular data shows us that not all small cell tumors are the same. Through this analysis I also found that many CRPC tumors exhibit immune signaling and aggressive proliferative signaling.

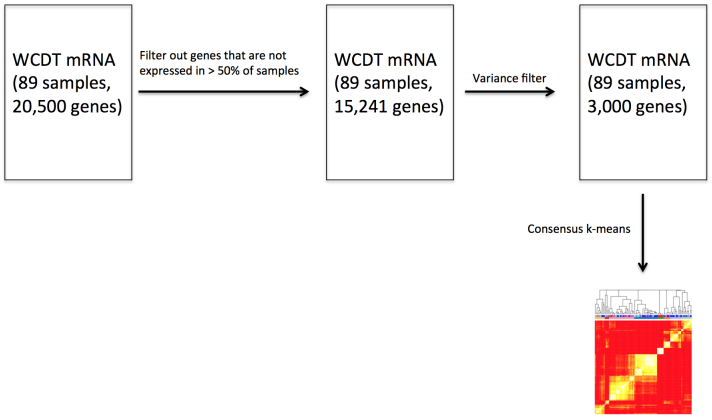


Figure 3.32: Outline of the preprocessing pipeline of the 89-sample CRPC mRNA expression data prior to unsupervised clustering.

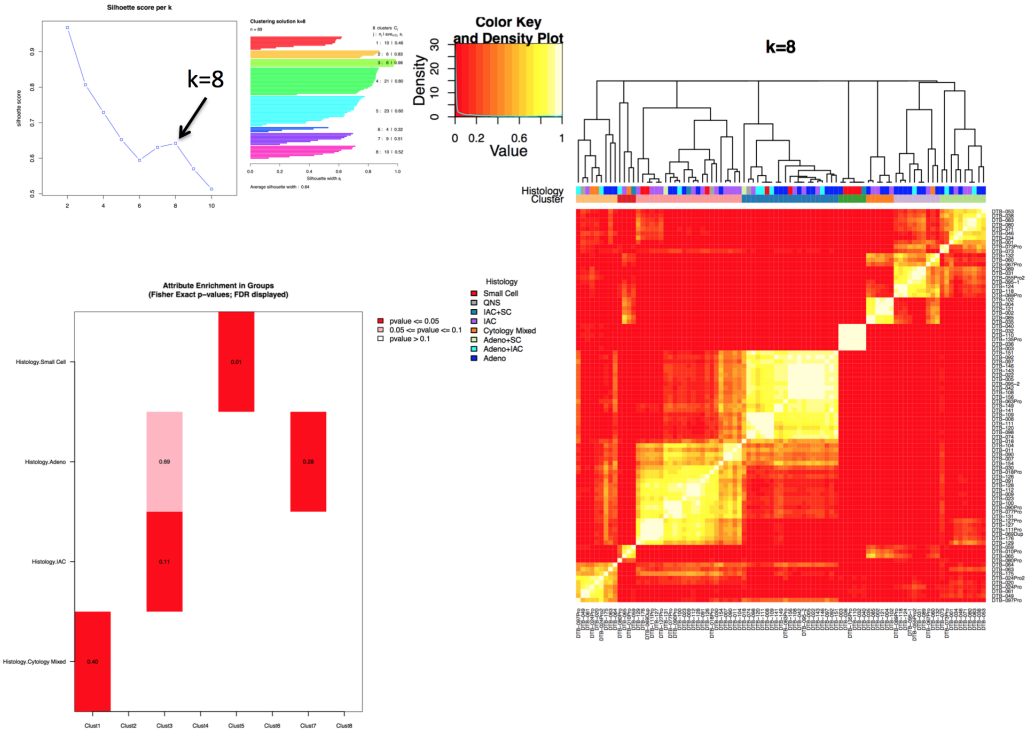


Figure 3.33: Chosen solution of k = 8 for the unsupervised census k-means clustering. The solution was chosen based on the silhouette score method. The clusters for the samples are presented by the heatmap. They show how histological labels assigned by a panel of pathologists are distributed among the 8 clusters. The heatmap in the bottom left shows statistical significance of the histological label enrichments in the clusters. The box is colored red if the p-value <= 0.5 and pink if 0.05 <= p-value <= 0.1. The value shown in the boxes is the FDR for each significance.

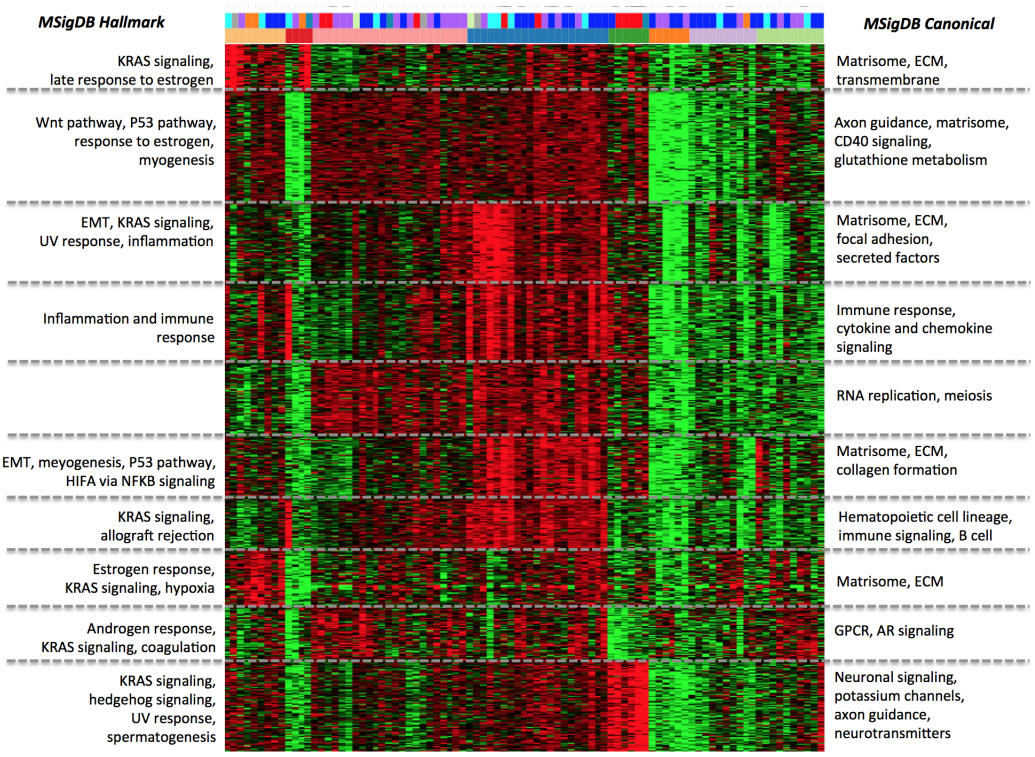


Figure 3.34: Results of feature space clustering of the k = 8 sample clustering solution. The gene clusters are annotated by their enrichments in Hallmark and Canonical Pathways sets from MSigDB pathway database.

3.6.2 Deriving Stemness Signature By Rank Aggregation

As a part of the Smith *et al.* study [22] of aggresiveness in CRPC, I worked on developing a single stemness signature from multiple signatures through rank aggregation methods. This particular study transcriptionally profiled epithelial populations from the human prostate and showed that aggressive prostate cancer is enriched for a prostate basal stem cell signature. High activity of CD49f (alias ITGA6) gene is a known marker of aggressiveness in human basal cells and carries a known stemness characteristics. Transcriptional profiling of tumors from eight patients that undergone

radical prostatectomy was completed and two subpopulations were identified (CD49f low and CD49f high). A CD49f-high phenotype signature was developed by computing differential transcription between the two groups using LIMMA [87] method. An additional 91-gene small cell neuroendocrine phenotype signature was developed by Artem Sokolov, one of the team members, through supervised analysis of the WCDT CRPC cohort. Top 91 differential genes were selected from the transcriptional CD49f-high signature in order to match the size of Sokolov's supervised signature.

Signature aggregation methods provide a way to extract intersecting molecular signals between two or more signatures. In this case, we combined CD49f high vs. low and supervised small cell carcinoma signatures to identify common signals between these two phenotypes. Because these signatures were derived by different methods, I used rank-based aggregation to normalize the weights and scale of each signature before finding the intersection. I used the Kolde *et al.* [108] rank aggregation method developed specifically for aggregating noisy gene lists (Figure 3.35). This method builds a statistical model for informative ranks in each of the lists being combined, then builds a final list of genes ordered by the minimum of the p-value of the order statistics across individual lists.

The results of our work were published in the PNAS journal [22].

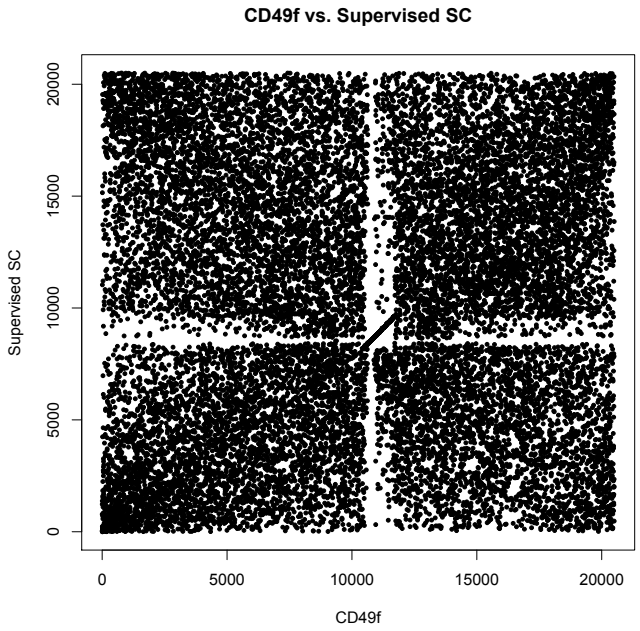


Figure 3.35: Comparison of gene ranks in full 20,500-gene CD49f-high and supervised Small Cell Neuroendocrine signatures. Very few ranks actually correlate between the two signatures.

3.6.2.1 Discovering Differences Between CD49f-high And Small Cell Signatures

While we found that the CD49f-high and the supervised small sell signatures had a lot in common and combining them revealed significant insights into the biology of aggressiveness of prostate cancer, we wanted to understand what the differences between the two are as well. One can look at the difference in ranks between corresponding gene features. However, how does one determine whether a particular gene rank difference is significant? I developed a statistical background model to test for such significance. Figure 3.36 describes our approach to developing this background model. In the figure

I designate CD49f-high signature as Sig1 and Small Cell Neuroendocrine signature as Sig2. The background model is developed by performing k-means clustering on the small cell signature and shuffling the gene modules a fixed number of times (n = 1,000). The true rank differences are compared then to the rank differences between CD49f-high signature and each of the signature in the background model. Then I ask how often each rank difference with the background signatures is at least as big (or exceeds) the rank difference between the two true signatures. I use this method to compute the empirical p-value for each gene. As a result, I identified that 2,168 genes had an empirical p-value <= 0.5. I analyzed those genes for their enrichment in various MSigDB pathway sets. I found that notably CD49f-high signature exhibits high Epithelial To Mesenchymal Transition (EMT) and Wnt pathway signaling while the supervised small cell signature exhibits high interferon signaling (suggesting innate immune response).

I used the PATHMARK [133] method to pull out molecular pathways of interest from Superpathway [133], based on the significantly differential genes (Figure 3.38). Figure 3.39 summarizes the individual pathways significantly differential in the two signatures (red is high in CD49f-high and blue is high in supervised small cell signature). These pathways show many well-known cancer markers associated with aggressiveness and immune signaling.

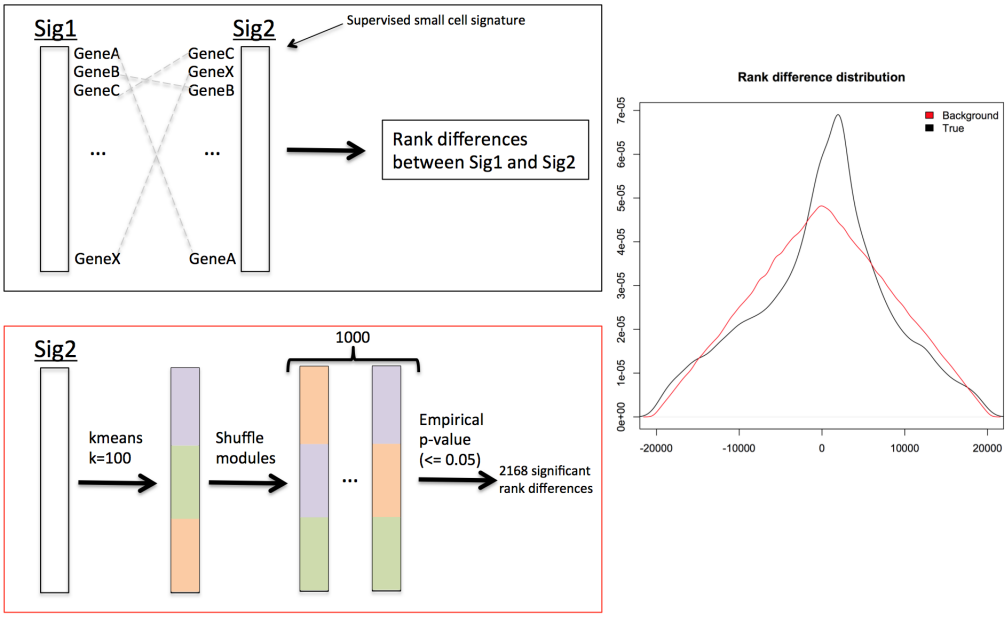


Figure 3.36: Overview of the background model for rank inconsistencies between the CD49f-high signature (designated as Sig1 in the figure) and Small Cell Neuroendocrine signature (designated as Sig2 in this figure).

MSigDB (Hallmark) Results
Genes High In CD49f And Low In SupervisedSC

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION [200]	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.	38		1.51 e-26	7.53 e-25
HALLMARK_P53_PATHWAY [200]	Genes involved in p53 pathways and networks.	25		1.51 e-13	3.78 e-12
HALLMARK_INFLAMMATORY_RESPONSE [200]	Genes defining inflammatory response.	22		5.55 e-11	6.94 e-10
HALLMARK_KRAS_SIGNALING_UP [200]	Genes up-regulated by KRAS activation.	22		5.55 e-11	6.94 e-10
HALLMARK_COAGULATION [138]	Genes encoding components of blood coagulation system; also up-regulated in platelets.	18		1.87 e-10	1.87 e-9
HALLMARK_APICAL_JUNCTION [200]	Genes encoding components of apical junction complex.	20		2.22 e-9	1.85 e-8
HALLMARK_ESTROGEN_RESPONSE_EARLY [200]	Genes defining early response to estrogen.	19		1.29 e-8	8.09 e-8
HALLMARK_TNFA_SIGNALING_VIA_NFKB [200]	Genes regulated by NF-kB in response to TNF [GeneID=7124].	19		1.29 e-8	8.09 e-8
HALLMARK_MYOGENESIS [200]	Genes involved in development of skeletal muscle (myogenesis).	18		7.13 e-8	3.96 e-7
HALLMARK_ALLOGRAFT_REJECTION [200]	Genes up-regulated during transplant rejection.	17		3.7 e-7	1.68 e-6

(a) High in CD49f.

MSigDB (Hallmark) Results
Genes High In SupervisedSC And Low In CD49f

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_SPERMATOGENESIS [135]	Genes up-regulated during production of male gametes (sperm), as in spermatogenesis.	23		9.44 e-13	4.72 e-11
HALLMARK_ESTROGEN_RESPONSE_LATE [200]	Genes defining late response to estrogen.	25		1.1 e-10	2.76 e-9
HALLMARK_INTERFERON_GAMMA_RESPONSE [200]	Genes up-regulated in response to IFNG [GeneID=3458].	24		6.03 e-10	1 e-8
HALLMARK_ESTROGEN_RESPONSE_EARLY [200]	Genes defining early response to estrogen.	23		3.14 e-9	3.14 e-8
HALLMARK_KRAS_SIGNALING_UP [200]	Genes up-regulated by KRAS activation.	23		3.14 e-9	3.14 e-8
HALLMARK_INTERFERON_ALPHA_RESPONSE [97]	Genes up-regulated in response to alpha interferon proteins.	15		3.26 e-8	2.71 e-7
HALLMARK_GLYCOLYSIS [200]	Genes encoding proteins involved in glycolysis and gluconeogenesis.	21		7.38 e-8	5.27 e-7
HALLMARK_UV_RESPONSE_UP [158]	Genes up-regulated in response to ultraviolet (UV) radiation.	18		1.82 e-7	1.14 e-6
HALLMARK_G2M_CHECKPOINT [200]	Genes involved in the G2/M checkpoint, as in progression through the cell division cycle.	20		3.31 e-7	1.66 e-6
HALLMARK_HYPOXIA [200]	Genes up-regulated in response to low oxygen levels (hypoxia).	20		3.31 e-7	1.66 e-6

(b) High in SC.

Figure 3.37: Pathways significantly enriched and differential between the CD49f-high and supervised small cell signatures.

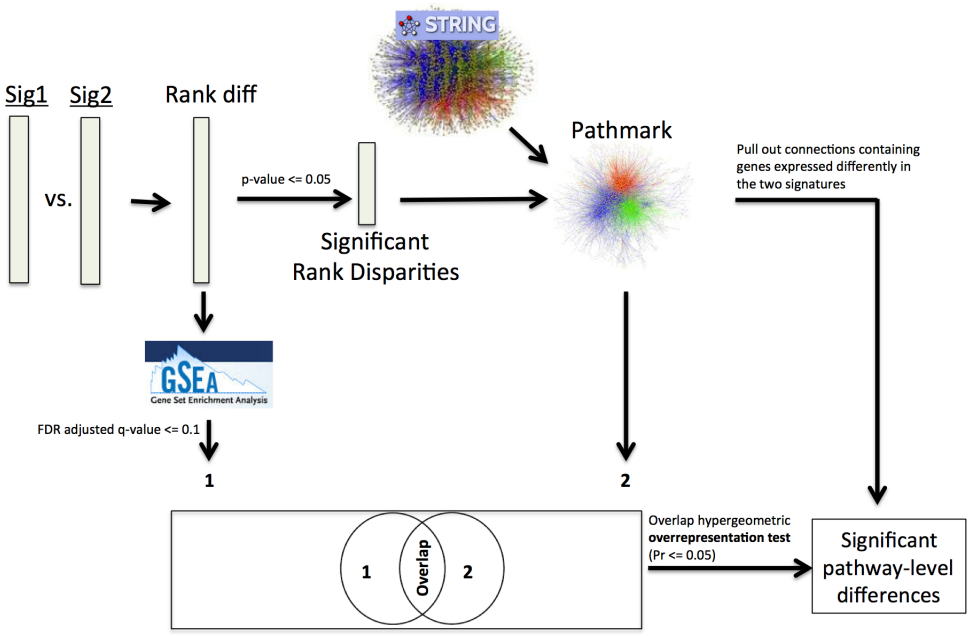
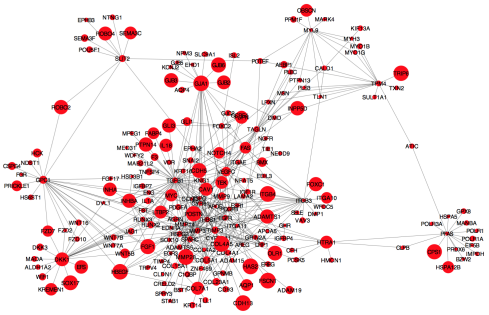
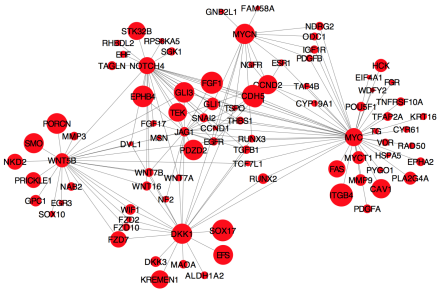


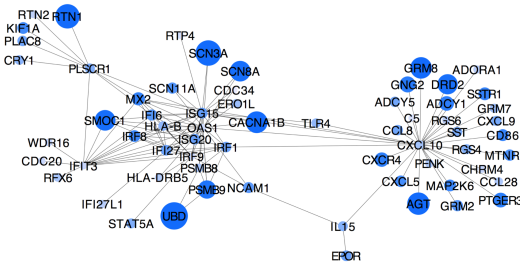
Figure 3.38: Overview of the method used to extract the pathways relevant to the differences of the CD49f-high and supervised small cell signature.



(a) EMT high in CD49f.



(b) Wnt high in CD49f.



(c) Interferon high in SC.

Figure 3.39: Individual pathways relevant to the CD49f-high (EMT and Wnt) and supervised small cell signatures (Interferon signaling).

3.6.3 Conclusion

As a part of the West Coast Dream Team initiative I participated in several projects. First, I worked to characterize a cohort of castration resistant prostate cancer tumors. I was able to identify several biologically and therapeutically important molecular subtypes based on RNA-Seq data. I found that some of these subtypes are significantly associated with distinct histological subtypes. I was able to characterize each of these subtypes by molecular markers and pathways. Second, I participated in a study that derived a stemness signature in CRPC tumors. I used rank aggrega-

tion methods to combine multiple stemness signatures derived by different analyses. In addition to deriving a single stemness signature, I characterized commonalities and differences between the aggregated signatures.

3.7 Identification of Early Metastatic Signature in Prostate Adenocarcinoma

Prostate cancer is the type of malignancy that develops in the prostate gland, a part of a male reproductive system. Prostate adenocarcinoma (PRAD) is the type of prostate cancer that develops in the gland cells and makes up nearly all diagnosed cases of prostate cancer. It generally occurs in older men and is most often detected using prostate-specific antigen (PSA) test. Many individuals diagnosed with this type of malignancy undergo active and ongoing surveillance by the treating oncologist, who can follow up with a number of treatments (surgical resection, radiation, hormone treatment, chemotherapy, or combination). Some of the individuals treated with hormonal androgen inhibition therapy, a therapy that blocks androgen receptor (AR) signaling, develop resistance to this therapy and no longer respond to it. This type of development often leads to castration resistant prostate cancer (CRPC), described above in section 3.6, which turns into metastatic disease. Metastasis is often the cause of death in cancer patients.

Tumor cells undergo very specific molecular changes during invasion and metastasis that allow cells to be detached from the tumor, travel to another location and

successfully start a new colony. Some of such changes include epithelial-mesenchymal transition (EMT) in preparation for invasion or intra-luminal growth and micro-colony formation by circulating tumor cells (CTC). As a result, metastatic tumors have a distinct molecular signatures that are responsible for activating signaling pathways in the cell that are specific to metastasis. For example, NOTCH signaling is known to be activated in metastatic tumors. In this paper we ask whether an early metastatic signature can be detected in primary tumors to predict future development of metastasis. Dr. Kiley Graim and I co-authored a project where we set out to perform a meta analysis of a number of PRAD datasets to see if we can identify an early metastatic signature in primary prostate cancer. This type of analysis could aid in identifying high risk patients early on. Similar approaches have been attempted with copy number [13, 56] and methylation [89] signatures of metastatic development. However, mRNA expression signatures of aggressiveness in prostate cancer have only been described based on their correlation with Gleason [61] score and clinical outcomes [84, 52].

This manuscript is currently in progress.

3.7.1 Metastatic Biopsies Exhibit More Similarity to the Matched Primary Tumors Than Unrelated Metastatic Tumors

First, we wanted to know how the similarities between metastatic and primary tumors from the same patient compare with similarities to other metastatic tumors within the same tumor type. To analyze how similar match metastatic and primary samples from the same patients are I utilized TCGA data across multiple tumor types.

In the TCGA cohort there are nine pairs of matched samples of primary and metastatic tumors: 2 pairs in Skin Cutaneous Melanoma (SKCM), 1 pair in Colon Adenocarcinoma (COAD), 1 pair in Head and Neck Squamous Cell Carcinoma (HNSC), and 7 pairs in Breast invasive carcinoma (BRCA). When I compared metastatic samples (Pearson correlation of transcriptional profiles) to their corresponding match primary samples the similarity was generally higher than when those samples were compared to other metastatic tumors within the same tissue (Figure 3.40). This similarity is even more pronounced when looking with specific metastasis-related pathways (e.g. NOTCH signaling pathway). This observation suggest that some of the primary samples carry an early metastatic signal that can be detected.

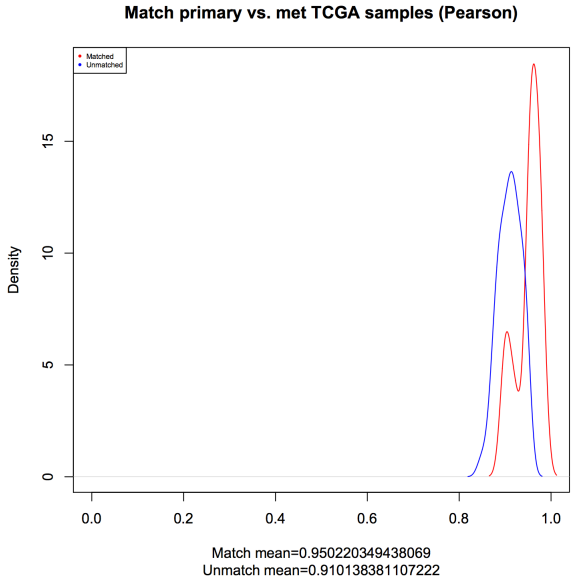


Figure 3.40: Correlation of the nine pairs of the metastatic TCGA samples to their corresponding match primary tumors (red) and to other metastatic tumors in the same tissue (blue). Metastatic and primary tumors within the same patient show higher similarity than if different patients are compared.

3.7.2 Data Preprocessing

I combined prostate cancer data collected from various studies (Table 3.1). These data contain normal, primary, and metastatic data. The collected data also came from multiple platforms (RNA sequencing and microarray). For this analysis I only utilized primary and metastatic prostate cancer data. After combining all the mRNA expression sets into a single matrix I identified 4,895 genes common to all datasets. Those 4,895 genes formed the feature space used in our analysis. At first the combined datasets exhibited a strong batch effect (Figure 3.41A). I utilized a commonly used an Empirical Bayes batch effect removal method called ComBat [68] and treated each dataset as a separate batch. After applying this method the samples from different datasets mixed together (Figure 3.41B) and did not cluster by the dataset they came from. Additionally, the platform is not the driving signal in the transformed data. Furthermore, I applied variance filter to the gene features of the ComBat-transformed data in order to filter out noisy genes. Based on the variance distribution (Figure 3.42) I identified 1,313 genes that had sufficient level of variance across the cohort. We used this ComBat transformed variance-filtered dataset for our further analysis.

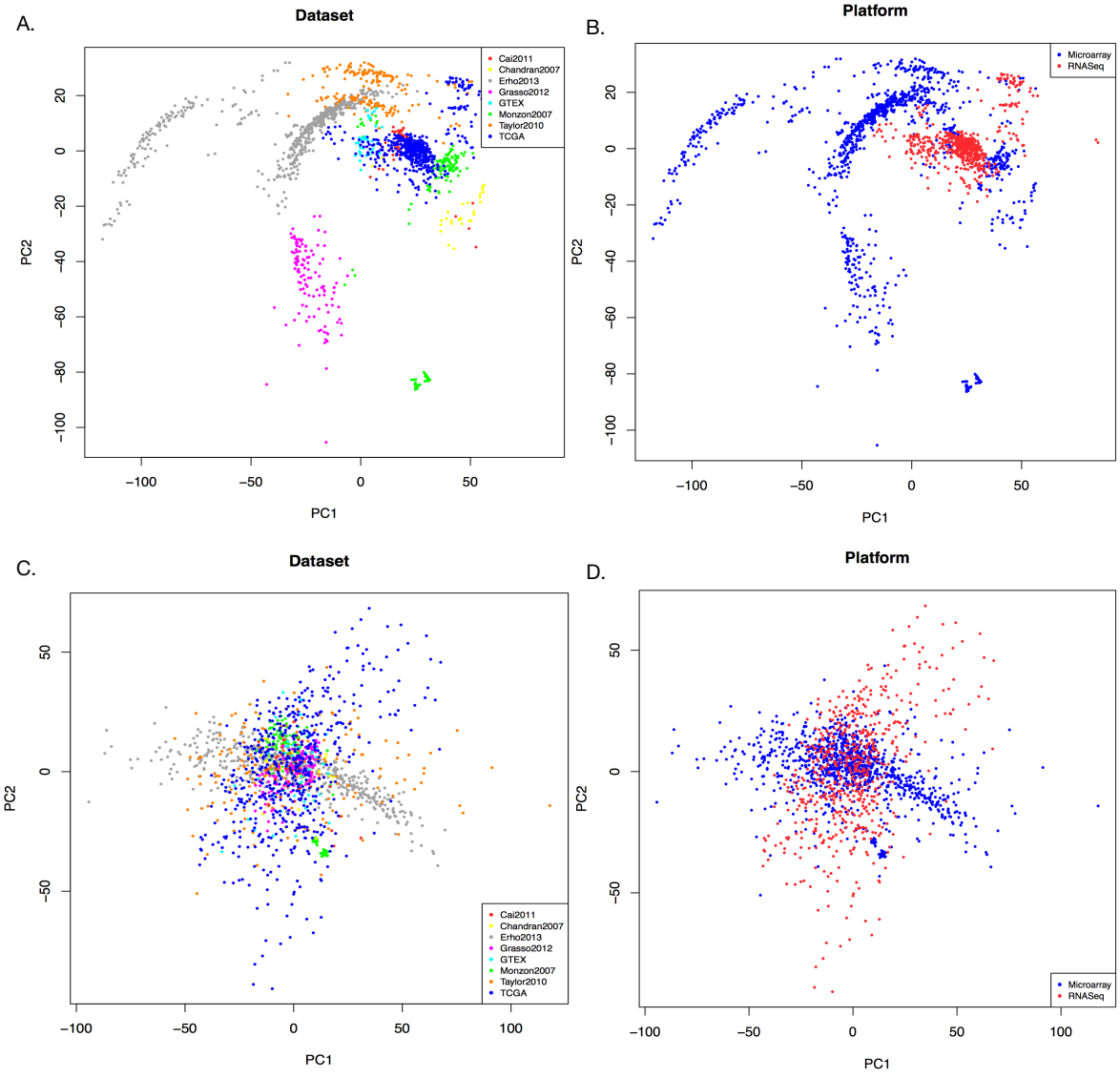


Figure 3.41: Principle component analysis (PCA) of the mRNA expression dataset included into the meta-analysis before (A-B) and after (C-D) ComBat adjustment. A and C show dataset distribution among samples (pre- and post-ComBat). B and D show platform distribution among samples (pre- and post-ComBat).

Dataset	# Normals	# Primaries	# Metastatic	# Genes	Platform
Cai [23]	0	22	29	10,523	Microarray
Chandran [30]	0	10	21	14,997	Microarray
Grasso [55]	28	59	32	15,830	Microarray
GTEx [80]	42	0	0	13,256	Microarray
Monzon [121]	52	65	25	9,383	Microarray
Taylor [129]	29	131	19	19,923	Microarray
TCGA [11]	21	246	0	20,500	RNASeq
Erho [47]	0	545	0	20,500	AffyHumanExon
Joint	172	1078	126	4,895	

Table 3.1: Datasets used in the meta-analysis.

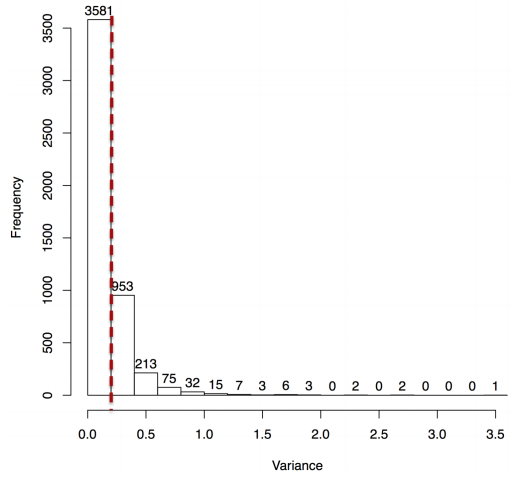
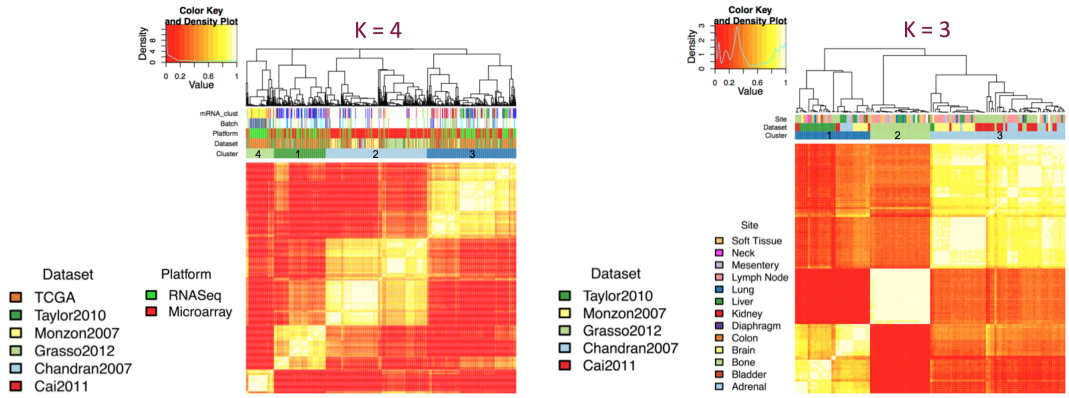


Figure 3.42: Variance across 4,895 genes in the combined ComBat-transformed mRNA expression dataset. The red line indicates where the cutoff for variance was placed. All genes on the left of the line were filtered out of the consecutive analysis.

3.7.3 Subtyping Primary and Metastatic Prostate Adenocarcinoma Identifies Metastatic-like Primary Subtype

In order to identify molecular subtypes of the primary and metastatic prostate adenocarcinoma I performed consensus k-means clustering [136] separately on primary and metastatic samples, using the ComBat-transformed variance filtered mRNA expression data. I extracted primary samples from this dataset and clustered them based on transcriptional profiles (785 samples, 1,313 genes). Similarly, metastatic samples were extracted and clustered (126 samples, 1,313 genes). Consensus k-means clustering was performed using complete linkage and 100 iterations. Based on the silhouette score metric, I identified four primary and three metastatic subtypes (Figure 3.43). From the annotated heatmap of sample clusters for primary samples (Figure 3.43(a)), it is clear that no single subtype is confined to a single dataset or a single platform. Two of the metastatic subtypes (Figure 3.43(b)) contain samples from multiple datasets, while one subtype is entirely composed of samples from the samples from the Grasso *et al.* study. This dataset was obtained from autopsy biopsies while the rest of the metastatic samples were live biopsies. My collaborator and I hypothesized that this Grasso cluster reflects true biology of samples from a dead tissue exhibiting definitively different molecular signal than sampled from live patients.



(a) Primary samples

(b) Metastatic samples.

Figure 3.43: Heatmaps of (a) primary and (b) metastatic cluster solutions with additional clinical covariate annotation bars.

3.7.4 Metastatic Signature Helps Identifying High Risk Individuals in Primary Cohort

The question with we started this project was "which primaries will metastasize?" Now, there are several possible models of primary-to-metastatic progression (Figure 3.44) and we currently do not know which model is at work or if multiple models of progression are possible. We hypothesized that if we can train a classifier model that predicts a primary subtype and apply it to metastatic samples we might be able to associate metastatic tumors with a possible primary subtype(s) that is enriched for the metastatic progression signature. Several consideration went into our decision of setting up our experiment this way as opposed to training a classifier on metastatic samples and predicting which metastatic subtype primary tumors are associated with. First, we know that every metastatic tumor originated from some primary tumor. Forcing

primary samples to associate with metastatic subtypes would constrain the experiment with assumption that every primary turns into metastatic tumor, which is not true. Second, we want to identify primary subtype(s) from which metastatic tumors originate. Therefore, it makes sense to build a classifier based on primary subtypes.

We trained a linear multiclass elastic net model to recognize four the primary subtypes from mRNA expression (1,313 gene features) using glmnet R package [50]. We used leave-one-out cross-validation with balanced success rate ($\sum_{i=1}^{n} \frac{\text{tpr}_i}{\text{pos}_i}$) to validate our model. We found that our model achieved high accuracy (Figure 3.45). My collaborator performed additional robustness and cluster stability tests she already described in her doctoral thesis.

We applied this predictor to the metastatic samples to find associations between primary and metastatic subtypes. We identified that majority of metastatic samples mapped to a single primary subtype (Figure 3.46). We should note that another primary subtype had a fair number of metastatic samples mapped to it. However, we chose to concentrate on the primary subtype to which the most metastatic samples, regardless of their metastatic subtype, mapped. We call this primary subtype "metastatic-like". This is an exciting discovery as it has important therapeutic implications. This suggests that the one-to-many model is the most likely scenario in which metastasis develops in prostate cancer, and possibly other types of cancer. This finding concurs with previously described similar types of analyses based on prostate cancer copy number [79]. One-to-many model is an easier model to approach from the perspective of cancer treatment. If there is an aggressive subtype of the disease and we

can identify those individuals who belong to this aggressive subtype then we can treat them differently and in a more aggressive manner than normally similar tumors would be treated.

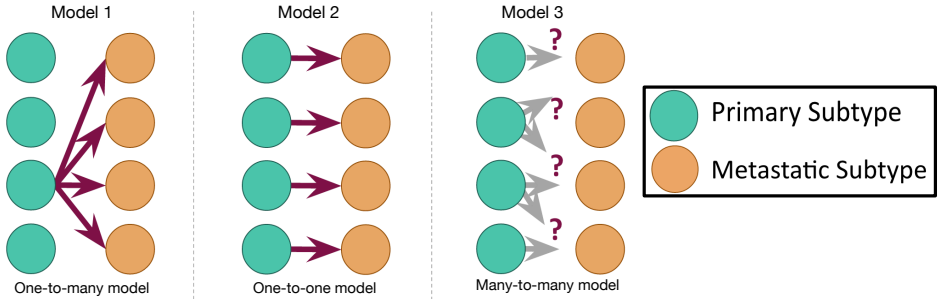


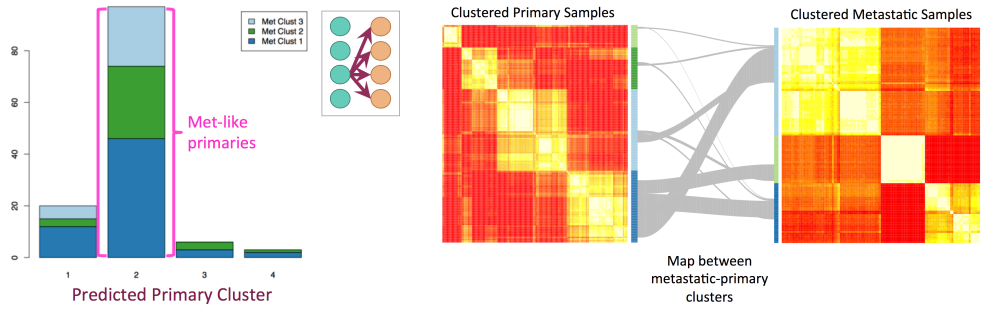
Figure 3.44: Description of possible primary-to-metastatic progression models.

Predicted

True cluster	1	2	3	4
4	0	0	0	1
3	0.010	0	0.990	0
2	0.017	0.983	0	0
1	0.990	0.010	0	0

Balanced Success Rate = 0.991

Figure 3.45: Confusion matrix (balanced accuracy) for leave-one out cross-validation model trained to predict primary prostate cancer subtypes.



(a) Metastatic to primary summary.

(b) Sample mappings.

Figure 3.46: Two different views of mapping between the metastatic and primary subtypes. A) Barplot shows which primary cluster each metastatic sample mapped to. Primary cluster 2 has the most metastatic samples mapped to it. Clusters 3 and 4 has the least number of metastatic samples mapped to it. B) Ribbon plot shows primary and metastatic subtypes and their mapping to the corresponding primary cluster. Multiple metastatic subtypes map into primary cluster 2.

3.7.5 Validation Using Matched PrimaryMetastatic Samples Highlights Advantages of Our Method

We validated our results on a held-out validation set. Erho *et. al* [47] analyzed 545 prostate cancer patients from the Mayo Clinic Registry, from 19872001. Median followup for these patients was 17 years and 212 patients (out of 545) were identified as early metastasis. Metastatic patients were grouped into "no recurrence" and "recurrence within 5 years" groups. Gene expression data for the cohort were collected with microarray assays. Erho *et al.* used a random forest model to classify patients into metastatic vs. not, and randomly split the data into training and test sets. In the control group 21 patients had clinical metastasis. This dataset contains matched primary and metastatic tumors (tumors that come from the same patient). Therefore,

this dataset is ideal to test early metastatic signature on. We applied our trained linear classifier to Erho *et. al* gene expression data. We also applied our classifier to the metastatic samples to see if those samples are still classified as metastatic-like primaries. When applying the classifier to the primary samples we compared the predicted result (metastatic-like vs. not) to the actual metastatic event (Table 3.2). Our approach improves over the original by using several types of data, introducing cross-validation into the model, and by using an independent dataset to validate the results. We also include many more samples, increasing the statistical power of this type of analysis.

		Metastatic Event	
		No	Yes
Predicted Primary Cluster	1	223	133
	2	86	66
	3	24	13

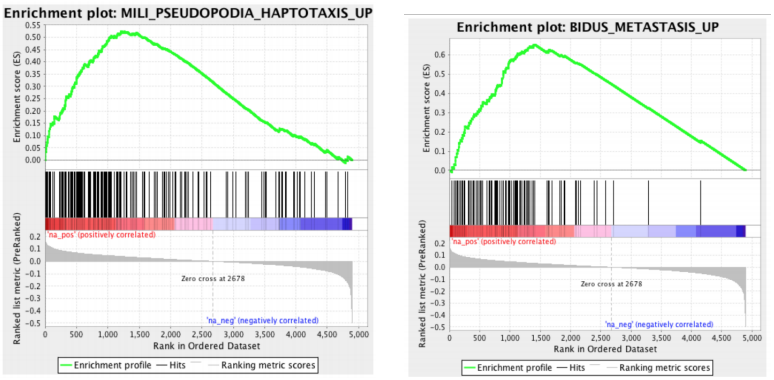
Table 3.2: Predicted primary clusters are enriched in samples that metastasized early.

3.7.6 Important Metastatic Markers Are Revealed Through Analysis of Metastatic-like Prostate Adenocarcinoma Subtype

Based on the ComBat-transformed and variance filtered mRNA expression, I computed metastatic-like differential signature in primary prostate cancer by juxtaposing metastatic-like primary cluster 2 and all other clusters. I identified that the top enriched pathways in that signature are cell mobility and metastatic progression

pathways. It means that genes involved in the tumor cell movement and ability to metastasize are already turned on in the samples of this primary subtype. This observation supports the one-to-many model proposed earlier (Figure 3.44) as it suggests that this particular subtype has differentially higher metastatic potential when compared to the entire cohort of primary tumors as a background.

I used PATHMARK [132] method with Superpathway [132] to extract significantly connected components of the network using the differential signature as an input (Figure 3.48). This network shows some of the well known proliferative signals (PLK1 and FOXM1) in the metastatic-like tumors. We also found that the metastatic-like samples exhibit higher cellularity, which has been previously associated with aggressiveness in cancers. MYB/MYC subnetwork recapitulates the same finding as the GSEA analysis found. These two genes have had a long history of being associated with aggressiveness and proliferation in cancers.



(a) Increased cell mobility.

(b) Metastasic progression signature.

Figure 3.47: Two top enriched pathways in the metastatic-like signature.

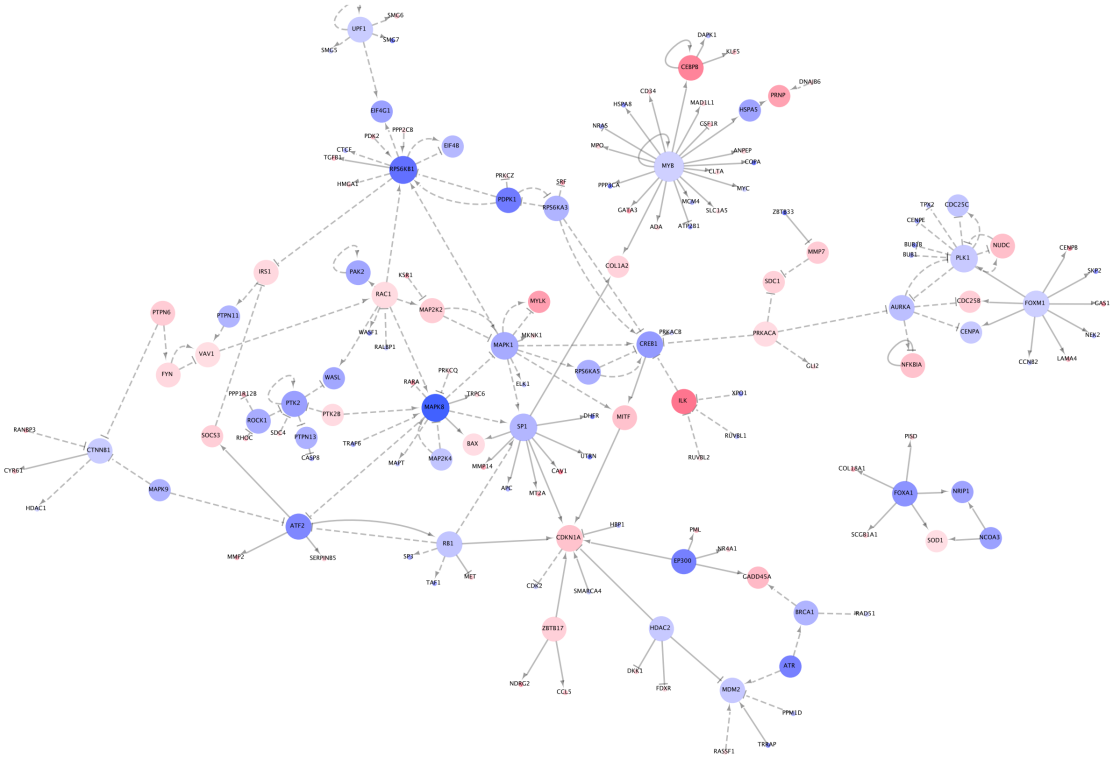


Figure 3.48: Differential subnetworks, based on mRNA expression in primary prostate cancer. Red color correspond to genes up-regulated in the metastatic-like primaries and blue color corresponds to the genes up-regulated in less aggressive subtypes. Node size corresponds to the connectivity (edge count); large nodes indicate network "hubs".

3.7.7 Conclusion

We analyzed a multi-platform multi-study cohort of primary and metastatic prostate tumors. We developed a methodology (Figure 3.49) for detecting early metastatic signature in primary prostate cancer based on mRNA expression data. This methodology aims at helping to identify high-risk patients with primary tumors that have the highest chance of metastasizing. We show that we are indeed able to isolate and identify an aggressive subtype of primary prostate tumors. We also show that we improved on a

previously developed similar approaches. Finally, we identify bio-markers that describe the aggressive primary subtype.

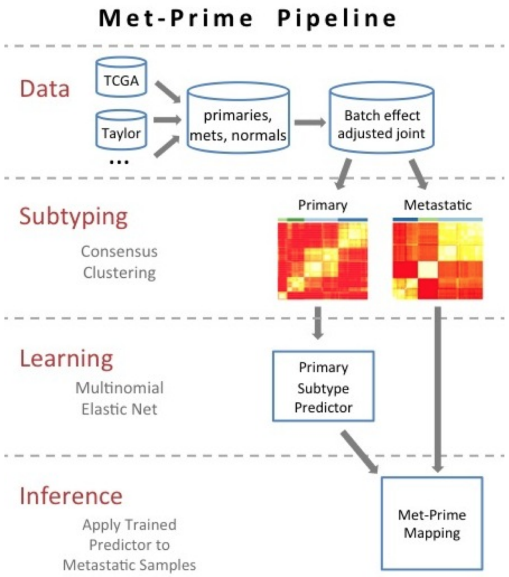


Figure 3.49: Overview of the method Graim and I co-authored to detect early metastatic signature in primary prostate cancer.

3.8 Chapter Conclusion

I described a number of collaborations and projects in which I participated with an aim of better understanding of cancer biology. In each one I was able to produce a significant contribution with specific new discoveries in our understanding of malignant tumorigenesis. I utilized a number of unsupervised and supervised methods and other bioinformatics tools, which aided me in arriving at my results. My work demonstrated that platform integration and multi-view analysis of cancer cells improves our ability to understand cell state and cell biology over single-view methods.

Chapter 4

Bringing Cancer Informatics Into the Clinic To Advance the Field of Personalized Medicine

Despite the encouraging statistics that more children with cancer survive now than ever, curing childhood and rare cancers is not a solved problem. Childhood cancers still kill more children than every other disease combined. On a good note, progress is being made in the field of childhood and rare cancer research. Genomic testing and consequent analysis are becoming more and more common in clinical settings [91]. However, most such testing and analysis are based on limited panels of genes and concentrate on the genomic aberration space (mutations, fusions) [57, 100, 91, 137]. Analyzing individual's mutation and genome structure data leads to actionable treatment leads in

only 15% of all childhood cancer cases [91]. This is because these cancers are often not driven by non-hereditary genome changes and frequently the genomic aberrations that are found do not have drugs that can target them. RNA sequencing has been used but with a focus on fusions. Gene expression analysis has been largely ignored because it has not been very clear how to utilize this information for a single patient. RNA sequencing platform showed vast promise in tumor subtyping in research settings. RNA expression carries important signals about the cell phenotype and can help group the new sample with the most molecularly similar tumors, given a group of reference tumors to compare to. This is especially important with rare tumors because diagnosis not always clear and even when it is we usually do not have many tumors with the same diagnosis to utilize for cohort-based analysis. Therefore, RNA-Seq data can help subtyping single tumors based on expression profiles. Here I present my work that contributes to the field of personalized medicine in pediatric cancers with the aim of improving individualized patient care in clinic. The approaches and methods are not specific to pediatric cancers and can be applied to adult cancers as well.

4.1 Comparison with Cancer Genomic Datasets Can Benefit Individual Pediatric Cancer Patients: Clinical Case Report

In this section I describe the work in collaboration with Olena Morozova, Joshua Stuart, Sofie Salama, Jing Zhu and David Haussler I completed towards a paper

we intend to submit to New England Journal of Medicine journal in the near future.

4.1.0.1 Publication Title and Author List

Title: Comparison with cancer genomic datasets can benefit individual pediatric cancer patients: a clinical case report

Authors: Newton, Yulia [1], NameTBD, NameTBD [2], Swatloski, Teresa [1], Jing, Zhu [1], McColl, Duncan [1], Salama, Sofie [1], Haussler, David [1], Stuart, Joshua [1], NameTBD, NameTBD [2], Morozova, Olena [1]

[1] Biomolecular Engineering and Bioinformatics, University of California Santa Cruz

[2] British Columbia Childrens Hospital

4.1.0.2 Abstract

Despite the recent advances in pediatric oncology, cancers kill more children than all other diseases combined. Genomic analysis of tumors has proven its clinical utility for adult cancer patients and has begun to enter the clinic for pediatric cancer patients as well. While several studies have highlighted the promise of genomic analysis of individual childhood tumors, these investigations typically focus on the analysis of the DNA sequence of a selected set of genes. Even when genome-wide approaches, such as Whole Exome Sequencing (WES), Whole Genome Sequencing (WGS) or RNA sequencing (RNA-Seq) are utilized, the interpretation of these datasets is often limited to mutations, copy number alterations and gene fusions affecting known cancer genes.

Here we introduce a novel unsupervised framework for the analysis of individual tumors gene expression profile in the context of large datasets of previously generated cancer gene expression data. We describe how to make use of these public genomic datasets in order to bring statistical power and context for the interpretation of RNA sequencing information from single patients. Our approach compares RNA sequencing profiles of an individual tumor to those of a reference cohort and determines if the individual's RNA sequencing profile is similar to that of another tumor type that has a treatment associated with it. We also show how to use a collection of public cancer RNA sequencing datasets as a reference for the identification of transcripts significantly upregulated in the individual patient. Our technique complements and often strengthens the information obtained from the analysis of genomic variants from the individual tumor.

4.1.0.3 Introduction

Cancer genomics has entered clinical practice for both pediatric and adult cancer patients. However, current methods for clinical interpretation of genomic data are largely limited to the detection of somatic mutations in well-characterized cancer genes. Unfortunately, in cancers with low mutational loads, such as many childhood cancers, this interpretation approach only produces viable treatment leads for up to 15% of patients [91]. This means that bringing gene panel or even whole exome tests into the pediatric cancer clinic would not help to change treatment outcomes for the majority of patients.

In order to identify and understand molecular drivers in pediatric cancer pa-

tients, genomic alterations other than gene fusions or hotspot mutations in oncogenes need to be considered. In particular, many pediatric cancers are driven by epigenetic changes that can manifest in specific gene expression profiles [143]. However, most computational methods for the detection of differentially expressed genes have been designed for the analysis of patient cohorts in a research setting, and need to be completely reconsidered for interpreting RNA sequencing data from single patients in the clinic. In particular, the identification of differentially expressed genes relies on the presence of multiple tumor samples, which is not available in the single patient setting. Here we propose a computational framework for the clinical interpretation of RNA sequencing-based gene expression profiles from single pediatric cancer patients. Our framework relies on the analysis of individual tumors in the context of large public collection of RNA sequencing data from a diverse pool of cancers. This comparative analysis allows to increase statistical power by identifying tumors with RNA sequencing profiles similar to the patients, effectively placing the single patient data into a cohort of similar tumors. In addition, our approach can identify oncogenic signatures that are not restricted to one tumor type, thereby highlighting opportunities for drug repositioning. In this way, our framework is designed to reveal patient-specific pathway alterations that are not restricted to single mutations and could eventually broaden the scope of individualized therapeutic avenues. Here we provide a proof-of-concept of this framework, as applied to a pediatric cancer patient with a rare tumor, dural-based central nervous system (CNS) sarcoma. We compared the RNA sequencing profile from this patient to a collection of public RNA sequencing profiles from 10,668 tumors and identified JAK/STAT pathway

as a potential therapeutic target. The patient, refractory to previous multiple lines of treatment, had a dramatic response to FDA-approved JAK inhibitor ruxolitinib. While the tumor ended up recurring after treatment, the patients medical team estimated that his life was extended by 2 years from what would be expected for his disease.

4.1.0.4 Overview of the Current Field of Pediatric Cancers and Therapy and Genomic-guided Therapy

To date four studies have investigated the clinical utility of genomic analysis of pediatric cancers [57, 100, 91, 137]. These studies revealed that the majority of pediatric cancer patients evaluated using genomic analysis do not receive treatment recommendations. The iCAT study used a DNA panel of 275 cancer genes and 91 introns covering 30 gene rearrangements to evaluate 100 patients. Thirty-one percent of patients had a recommendation for therapy made based on the genomic analysis of their tumor [57]. The BASIC3 study used a combination of germline and somatic WES to evaluate 150 patients [100]. Twenty-seven percent of patients had either a somatic alteration associated with either established or potential clinical utility. The Peds-MiOncoSeq study took a more comprehensive approach including a combination of WES and RNA sequencing to evaluate 102 patients [91]. Forty-six percent of cases had a potentially actionable finding suggesting that including RNAseq can increase the clinical utility of genomics. Of note, RNAseq in this study was used only for identification of fusions and not for gene expression analysis. Finally, the European INFORM study analyzed 57 patients using WES, methylation and expression arrays

and reported potentially actionable findings in fifty percent of patients.

4.1.0.5 Case Presentation and Clinical History

The patient (referred to as Patient 1 hereafter) was diagnosed at eight years of age with a dural-based central nervous system (CNS) sarcoma. The histology was ambiguous and had features of both desmoplastic small round cell tumor (DSRC) and clear cell sarcoma (CCS). At the time of diagnosis the patient was treated with six cycles of induction chemotherapy: ifosfamide, carboplatin and etoposide (ICE), followed by high-dose chemotherapy with carboplatin, thiotepa and etoposide, followed by autologous stem cell transplant, and local radiation. After two years of remission, the tumor recurred in the lungs, at which point the patient was enrolled on a Personalized OncoGenomics (POG) clinical trial at British Columbia Childrens Hospital. Biopsy material from a lung metastasis was characterized using whole genome sequencing and RNA sequencing, and peripheral blood was characterized using whole genome sequencing as previously described [65]. The analysis of the sequencing data revealed an EWSR1-ATF1 gene fusion, consistent with a sarcoma diagnosis [131, 103]. The sequencing also revealed three somatic variants of unclear therapeutic significance, PDGFRA p.V299F, PRKCB p.D341N and SVIL p.L1374R. No germline SNVs with established cancer relevance were detected in the patient.

4.1.0.6 Results

We compared the RNA sequencing profile of Patient 1 with the RNA sequencing profiles of the reference cohort consisting of 10,668 tumor samples from 38 different tumor types, derived from pediatric and adult cancer patients [28, 10]. The Outlier Analysis (Methods) identified 2,704 genes significantly upregulated in Patient 1 (up outlier genes) and 504 genes significantly downregulated in Patient 1 (down outlier genes). We concentrated on the up outliers as these genes represent most likely therapeutic targets. We used the Drug Gene Interaction Database (DGIdb) [134] (Methods) to narrow this list to known l therapeutic targets. The DGIdb search produced 78 drug-gene interactions, involving 20 unique genes, including many members of the tyrosine kinase signaling pathway.

We used the Tumor Map method (manuscript in submission) to visualize the similarity of Patient 1s RNA sequencing profile to those of other tumors in our reference cohort (Figure 4.1A). This visualization method lays out tumor samples on a two-dimensional space based on similarities of their RNA sequencing profiles, as defined by the expression of 18,357 genes. Although diagnosed as a sarcoma, Patient 1s tumor clustered with lung adenocarcinoma (LUAD) tumors, and a few lung squamous tumors (LUSC) that evidently exhibit similar gene expression profiles. Since the tumor material used for this analysis was taken from a lung metastasis, this placing is not fully surprising. However, not all LUAD samples cluster together (Figure 4.1B), suggesting that there are several molecular subtypes of LUAD tumors and Patient 1 is most similar

to a particular subtype. We performed differential gene expression analysis comparing the cluster in which Patient 1s tumor was placed (n = 350) with all other LUAD tumors (n = 362) and selected only statistically significant differences (adj. p-value ¡= 0.05). Fourteen genes identified as significantly differentially expressed by this analysis were are also in the list of known therapeutic targets identified by DGIdb (Figure 4.1C).

EWSR1-ATF1 fusion found in Patient 1 is known to be associated with CCS tumors [131] and has been shown to activate tyrosine kinase signaling [40]. Among the 14 druggable significantly differentially expressed genes we found anaplastic lymphoma kinase (ALK). Previous reports indicated that patients with EWSR1 fusions may respond to ALK inhibitor treatment [123]. Therefore, given the high expression of ALK in this patients tumor, we hypothesized that tyrosine kinase signaling pathway may be contributing to Patient 1s tumorigenesis. We used fourteen up outliers that were mapped to drugs using DGiDB to conduct functional enrichment analysis using MSigDB [78]. This analysis revealed tyrosine kinase signaling as one of the top signaling pathways (p-value = 4.11e-13). Using signaling interactions mined from the literature as well as the MSigDB [78], we reconstructed Patient 1-specific tyrosine kinase signaling pathway containing highly expressed transcripts, as defined by our comparative RNA sequencing analysis (Figure 4.2). This pathway contains two receptor tyrosine kinases, ALK and FGFR1, both of which were identified as up outliers. The functional enrichment analysis also revealed over expression of two downstream pathways, JAK/STAT and PI3K/AKT/mTOR which converge on the activation of the cell cycle machinery, including CCND1 and CDKN1A. TGFB1, which has been reported to aid in activation of

JAK/STAT signaling [92, 66], was also found to be an up outlier in Patient 1. Both ALK and JAK1 are highly expressed in Patient 1, as compared to the reference cohort (Figure 4.3A), and were both identified as up outliers. The ALK has been implicated as a driver in multiple tumors, including lung adenocarcinoma, anaplastic large-cell lymphoma and neuroblastoma, and ALK inhibitor Crizotinib has been in clinical development for pediatric ALK-driven malignancies [115, 141]. Therefore, we sought additional evidence of ALK as a potential driver in the Patient 1 tumor. We compared ALK expression in Patient 1 with that in other sarcomas, as well as ALK-driven and non-ALK-driven lung and neuroblastoma tumors (Figure 4.3B). We discovered that the level of ALK expression in Patient 1 was comparable to those of ALK-driven lung cancers and ALK-driven neuroblastomas.

Since ALK was not the only receptor tyrosine kinase overexpressed in Patient 1s tumor, we also investigated the expression of JAK1, a key downstream component of the signaling network. We compared both ALK and JAK1 expression levels in the Patient 1 cluster in Tumor Map to other lung tumors in the reference cohort and found that both of these genes are expressed at significantly higher levels in the Patient 1 cluster (Figure 4.3C). These observations suggest that ALK and JAK/STAT pathway could represent therapeutic opportunities for this patient. Consistent with our observations, the treating oncologist prescribed an ALK inhibitor Crizotinib and then JAK inhibitor Ruxolitinib. While eventually the patient succumbed to the disease, these treatments prolonged the patients life by an estimated two years.

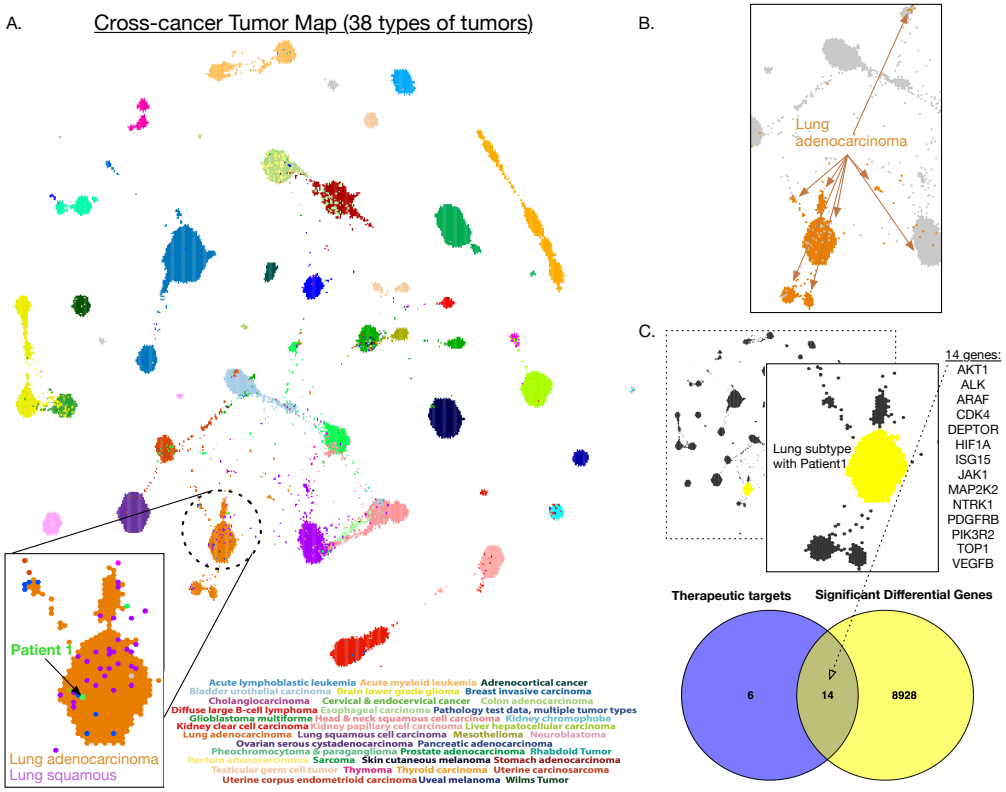


Figure 4.1: Patient 1s RNA sequencing profile in the context of the reference cohort of 38 different tumor types, both pediatric and adult. A) A projection of the entire tumor cohort on a 2-D map using Tumor Map method. Each tumor in the map, represented by a hexagon, is colored by the tumor type as described in the legend. Patient 1s tumor is shown in green in the lung cluster. B) LUAD tumors cluster in several subtypes on the Tumor Map visualization. The arrows point out different clusters LUAD tumors belong to. C) In yellow we highlight the tumors that were considered part of the Patient 1 cluster. We ran differential gene expression analysis of those tumors vs. other lung tumors. The intersection of the statistically significant differentials and the druggable up outliers (called Therapeutic Targets) is shown by the Venn diagram.

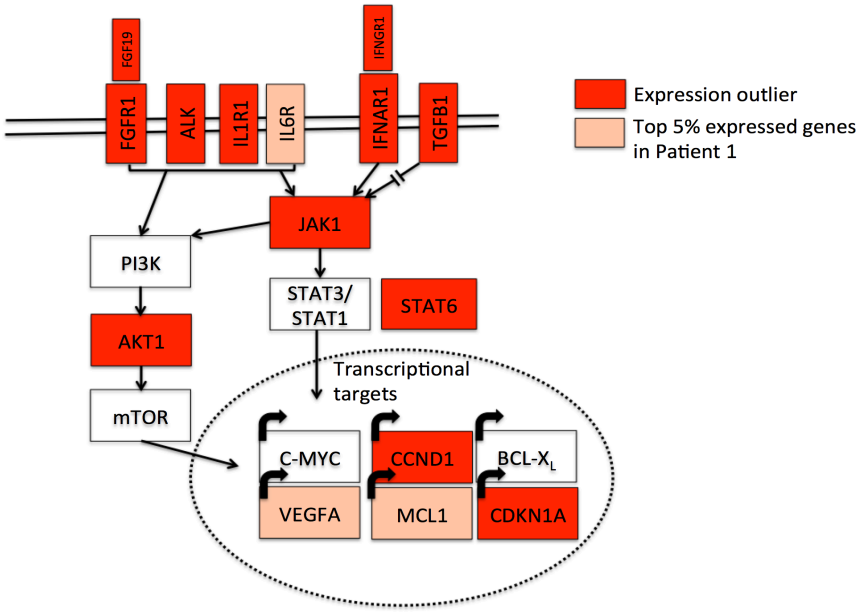


Figure 4.2: Patient 1s tumor expresses ALK at a level similar to those in ALK-driven malignancies . A) Expression of ALK and JAK1 as compared to the expression of tumors in the reference cohort, separated by tumor type. B) Expression of ALK in Patient 1 is comparable to ALK-driven lung adenocarcinoma and neuroblastoma . C) Comparison of ALK and JAK1 expression levels in the Patient 1 cluster with other LUAD tumors.

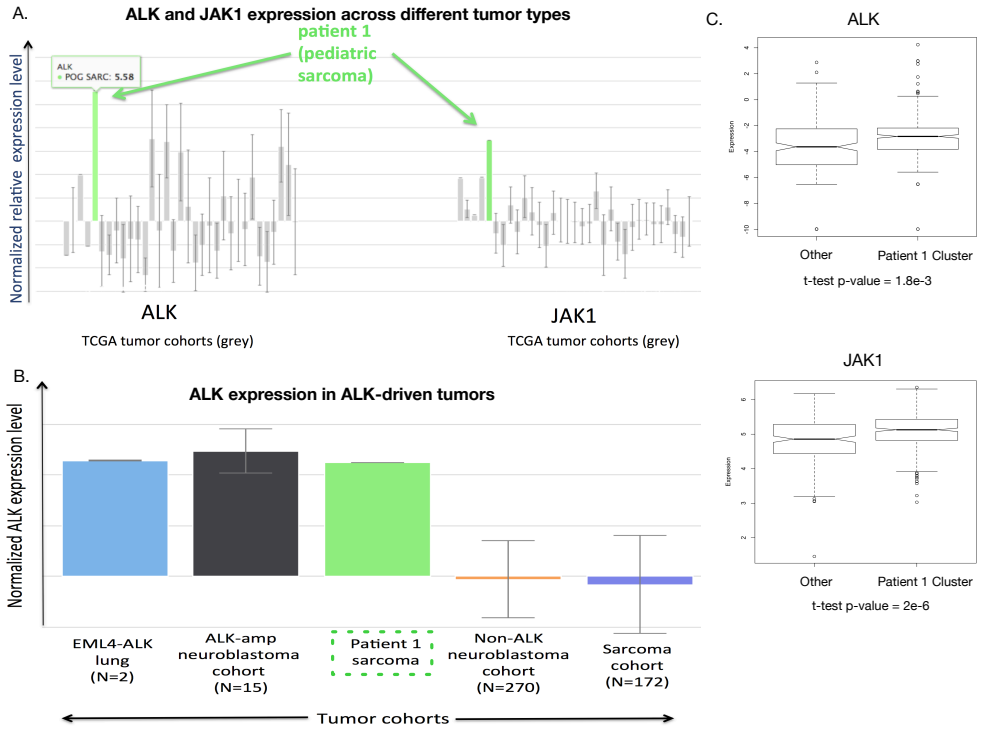


Figure 4.3: Proposed pathway that may contribute to the disease in Patient 1. This pathway demonstrates how both ALK and JAK1 participate in the activation of tyrosine kinase signaling in Patient 1s tumor.

4.1.0.7 Discussion

Simultaneous analysis of multiple tumor types has been previously utilized in research settings to discover molecular patterns that traverse tissue-based cancer diagnosis [53, 59, 38, 142]. Similar molecular patterns in different tumor types can occur when the tumor arising from the same cell of origin can originate in multiple tissue types. For example, some tumors originating as a bladder urothelial carcinoma exhibit squamous cell phenotype [59]. Squamous cell tumors can also originate in other tissues, such as lungs, digestive tract, and prostate. When analyzing RNA sequencing

profiles across these multiple cancer types we usually see squamous tumors from multiple tissues cluster together based on the similarity of their gene expression profiles [59, 128].

Comparative analysis of tumors can benefit single patients [57, 100, 91, 137]. However, the analysis of the genomic data from patients in the clinic is often limited to considering mutation profiles of a selected set of genes. Furthermore, not every genomic aberration that is found in the tumor genome of a patient is targetable by currently available drugs. RNA sequencing approaches have been gaining more traction in a clinical testing, but the application of these technologies is often limited to the discovery of fusion genes and expressed mutations [90]. Gene expression information that could be derived from RNA sequencing data is typically not considered in a clinical setting. This is because the analysis of single samples lacks the statistical power to identify genomic drivers and other molecular patterns important for clinical decision-making. Supervised methods have been utilized to derive diagnostic and prognostic markers from gene expression in adult tumors [98, 99, 106]. However, the application of these methods in clinical settings is limited by the need of prior knowledge about the tumor in order to test for the appropriate signature. With rare cancers, such as many pediatric tumors, we often to do not know which particular diagnostic or prognostic phenotype to test for.

Here we provide a demonstration of how analyzing single tumors in the context of a large cohort of tumors representing multiple types of cancer allows us to make use of gene expression information derived from RNA sequencing. In addition, this approach allows us to simultaneously query the patients gene expression profile for

multiple oncogenic signatures. Comparing the individual patients genomic information to a cohort of reference tumors can help determine if this patients genomic profile is similar to another type of cancer, which could provide insight into the clinical behavior of the individual tumor, thereby providing additional information to the oncologist.

Encouraged by the promising results of applying our framework to Patient 1, we launched f the California Kids Cancer Comparison (CKCC) collaborative effort [1] to advance the field of personalized pediatric oncology, led by the University of California Santa Cruz Genomics Institute. We are currently pursuing further and systematic evaluation and improvement of the single patient comparative analysis framework described here.

4.1.0.8 Methods

Reference Data

We obtained The Cancer Genome Atlas (TCGA) [28] and Therapeutically Applicable Research To Generate Effective Treatments (TARGET) [10] RNA sequencing Fragments Per Kilobase of transcript per Million mapped reads (FPKM) [32] gene expression data from the public repository in Xena [86, 67]. Both datasets were processed using STAR2 [42] for sequence alignment and RSEM [77] for gene expression estimation by the UCSC Genomics Institute [67]. We extracted tumor-only samples from the TCGA dataset, then combined these gene expression data into a single cohort (n = 10,668). From this dataset we extracted only 18,814 gene features that are present in Patient 1s gene expression set obtained using the POG pipeline [65].

Gene Expression Outlier Analysis

Gene-level Reads Per Kilobase of transcript per Million mapped reads (RPKM) data were generated according to the previously published method [65]. We used these data to compute Fragments Per Kilobase of transcript per Million mapped reads (FPKM) by dividing each value by two. The final feature space contains 18,357 individual unique gene features.

We then performed gene expression outlier analysis to identify transcripts significantly enriched in the patient's tumor as compared to 10,668 other cancer samples in the reference compendium. The outlier analysis has been used in an adult cancer setting and led to a clinical benefit in a case report (Jones et al. 2010); however, it has not been evaluated in the pediatric cancer setting. Gene expression outliers were identified as described with the exception of using a more stringent IQR of 2.0 (Jones et al. 2010). We analyzed the outlier genes for enrichment of specific pathways and signaling networks that afford potential druggability using MSigDB [78].

Identification of Druggable Targets

We used the DGIdb [78, 134] to search for drug targets among the genes found to be upregulated in Patient 1 by gene expression outlier analysis. In our DGIdb search we specified all 41 gene categories, and narrowed down the database and the interaction type parameters. We chose the following four databases: MyCancerGenome, MyCancerGenome CLinical Trial, CIVIC and Cancer Commons to focus on drug targets with known cancer relevance. We chose the following six interaction types: antagonist, antibody, blocker, inhibitor, inhibitory allosteric modulator, and suppressor.

Tumor Map Analysis

We first computed pairwise Spearman [101] correlations of RNA sequencing profiles of the tumors in our reference cohort (n = 10,669). This produced a square correlation matrix with 10,669 columns and 10,669 rows.

The Tumor Map method seeks to project high-dimensional genomic observations onto a 2-D plane, while preserving original sample-to-sample distances. Tumors cluster together according to the similarity of their RNA sequencing profiles. Then, we use the quasi-physics based layout engine OpenOrd (formerly known as DrL) [85], implemented in the igraph R package [51], to derive an initial set of (x, y) positions for the samples, based on the correlation matrix [138]. The similarity space is represented as a graph and used as an input into OpenOrd. OpenOrd treats the similarities as spring constants and searches for a configuration among the samples that produces an arrangement to minimize the spring tension of the system as much as possible. We utilize hexagonal packing for space conservation in the projected 2-D plane. For each sample in the full correlation matrix,we extracted samples with top 6 correlation values to compose a sparse matrix of top 6 nearest neighbors. We use this sparse matrix to construct a sparse similarity graph for the samples in the cohort and apply the OpenOrd method to derive the initial (x, y) positions in the map.

Furthermore, to avoid overlapping and crowding samples in the dense graph components, OpenOrd (x, y) coordinates are snapped to their nearest hexagon to arrange all of the samples on a tiling of regular hexagons. Using OpenOrd (x, y) coordinates, each sample is placed in a grid cell. If the predetermined cell is occupied, the

sample is snapped to an empty grid cell within a minimal distance from the original cell. Multiple samples that compete for a location will thus spiral around a central hexagon in the neighbors around the central location. Thus, dense clumps are separated so that they can be viewed on approximately the same scale as the distances that separate them. Hexagons were selected as the shape for the grid cell in order to illustrate that there are no inherently preferred axis-aligned directions in the OpenOrd output.

Google Maps API [4] is then used to load and visualize the resulting layout into a browsing environment. The API provides the ability to interactively navigate, zoom, and explore various annotations of locations on the map analogous to Google Maps and Google Earth applications.

4.1.0.9 Conclusion

We describe a framework designed to reveal patient-specific pathway alterations that are not restricted to single mutations and incorporate tumor-specific changes in gene expression profiles. We provide preliminary evidence that the identification of such events could broaden the scope of individualized therapeutic avenues available to pediatric cancer patients. Our case study demonstrates the utility and benefits of the RNA sequencing data for single pediatric cancer patients treated on clinical trials today. The case study also demonstrates that our informatics approach is capable of producing new therapeutic options not found through genome variant analysis. While the presented framework shows promising results, we suggest that this framework be further developed and evaluated in larger patient cohorts. We highlight the need and

hope to inspire the development of new informatics tools and principled methods in the field of personalized medicine for pediatric cancers.

4.2 California Kids Cancer Comparison Initiative

As a part of our work in advancing the field of precision medicine in pediatric cancers we launched the Treehouse group, which spans several labs in the Biomolecular Engineering department at the University of California Santa Cruz. I had an honor and a pleasure of being one of the founding members of the Treehouse team.

The Treehouse group participated in and won a demo competition run by the governor of California Jerry Brown to fund innovations in the field of precision medicine. UCSC Treehouse group is leading the California Kids Cancer Comparison (CKCC) initiative (http://www.ciapm.org/project/california-kids-cancer-comparison), a collaborative and multi-institutional effort to advance the field of pediatric oncology. In fact, we partner with several pediatric oncology hospitals in California. We take individuals' genomics and analyze it in the context of a large library of previously researched data. We determine if this individual's genomic profile is similar to another tumor type that has a treatment associated with it. This type of approach allows for more targeted and timely treatment decisions that have the most chance to be effective in killing the tumor. This also gives hope to often last-resort individuals with no treatment options available to them. This is a very powerful approach given the large number and diversity of available data in our reference library. In it we currently have genomic data

from over 10,000 tumors representing 38 types of cancers, both pediatric and adult. To our knowledge it is by far the largest such compendium available anywhere in the world. This is really a paradigm shift in the field of personalized medicine. And pediatric cancers are a perfect fit with the personalized medicine initiative because every tumor is so different.

In this section I describe the work I completed as a part of the Treehouse group's participation in the CKCC initiative. I also describe the ongoing and future efforts as we are still continuing our work in advancements in the pediatric cancer precision medicine.

4.2.1 General Approach

Inspired by the initial success of the Patient 1 story and recognizing the need for a more systematic evaluation and the opportunity for improvement of N-of-1 analysis, we formulated our approach in the following way: we analyze RNA sequencing profiles of new patients in the context of a large reference cohort of previously researched cancers. We obtain RNA sequencing data for individual patients from our partner hospitals. Currently, we only utilize gene expression data. However, our future plans include utilizing additional information extracted from RNA sequencing data as well as the ability to obtain and analyze genomic variant data. The initial reference data freeze (CKCCv1; Figure 4.4) included the same data as described in the Patient 1 analysis, consisting of TCGA and TARGET tumors. Going forward, we hope to incorporate previously analyzed de-identified RNA sequencing data into the new data freezes. We

apply a set of N-of-1 analysis tools to the individual's data, then collect, summarize and interpret the results produced by those tools to extract potential new therapeutic leads. As a part of our analysis, we ask several questions that help us arrive at our conclusions:

- Which tumors in the reference cohort is this individual's RNA sequencing profile is most similar to?
- What makes this individual's RNA sequencing profile similar to those tumors?
- What makes this individual's RNA sequencing profile different from tumors in the reference cohort?

As a part of our analysis we utilize a set of currently available to us tools:

- Tumor Map (not yet published) - to place the individual's tumor within the reference cohort. In order to maintain a static reference map across all new individuals we maintain a single "frozen" map derived from the reference cohort. New N-of-1 tumors are being place into the map using N-of-1 placement (see 4.2.4).
- Outlier Analysis [65] - to identify which genes make this tumor different from the tumors in the reference cohort. Generally, two versions of the Outlier Analysis are performed. The first one is done against the entire reference cohort. The second is guided by the Tumor Map placement and is performed within a subcohort restricted to tumors in the cluster where the individual's genomic profile was placed (unless, no clear placement was obtained).

We are hopeful that we can expand our toolbox with additional tools, which will aid in identifying druggable drivers and pathways in the individual patient. So far however, we have been able to provide additional therapeutic leads to 19 Treehouse N-of-1 samples (Figure 4.5).

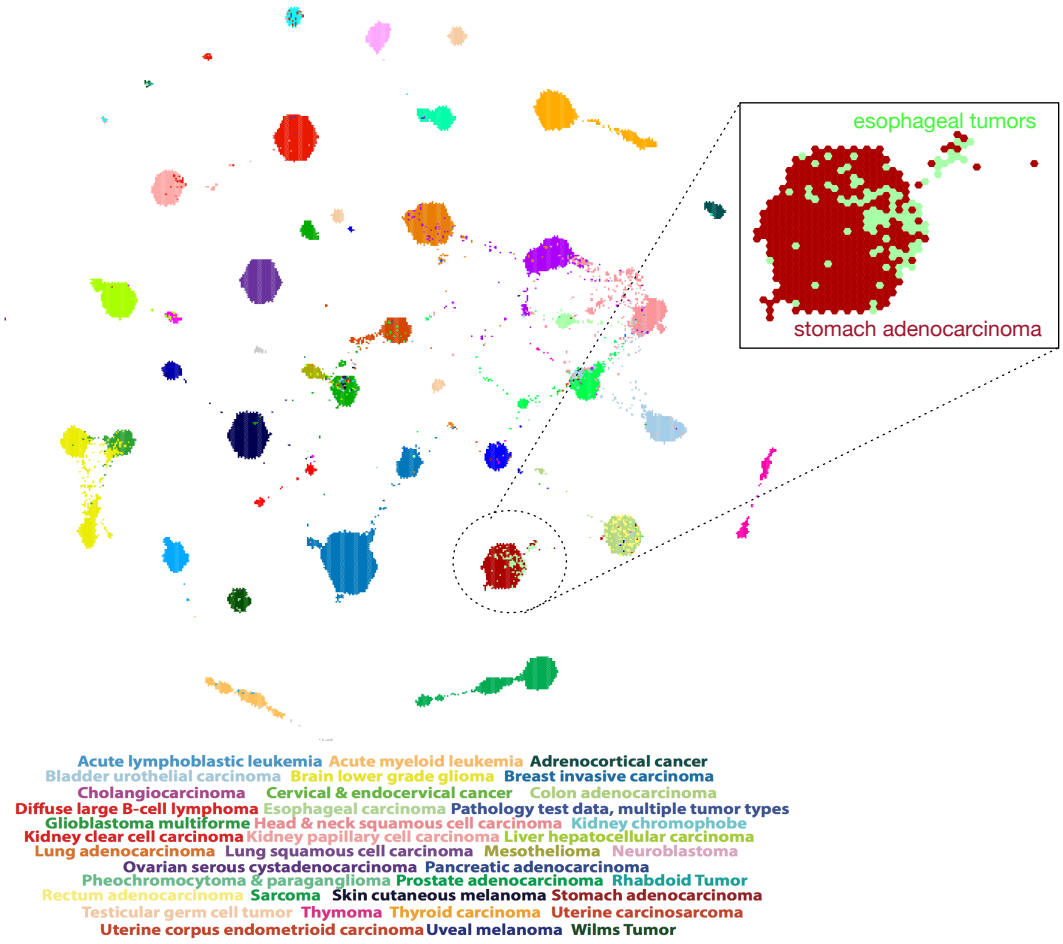


Figure 4.4: CKCC version 1 reference dataset visualized using Tumor Map method. Each dot/hexagon in the map is a sample in the reference cohort and the samples are laid out based on the similarity of their gene expression profiles and are colored by the disease.

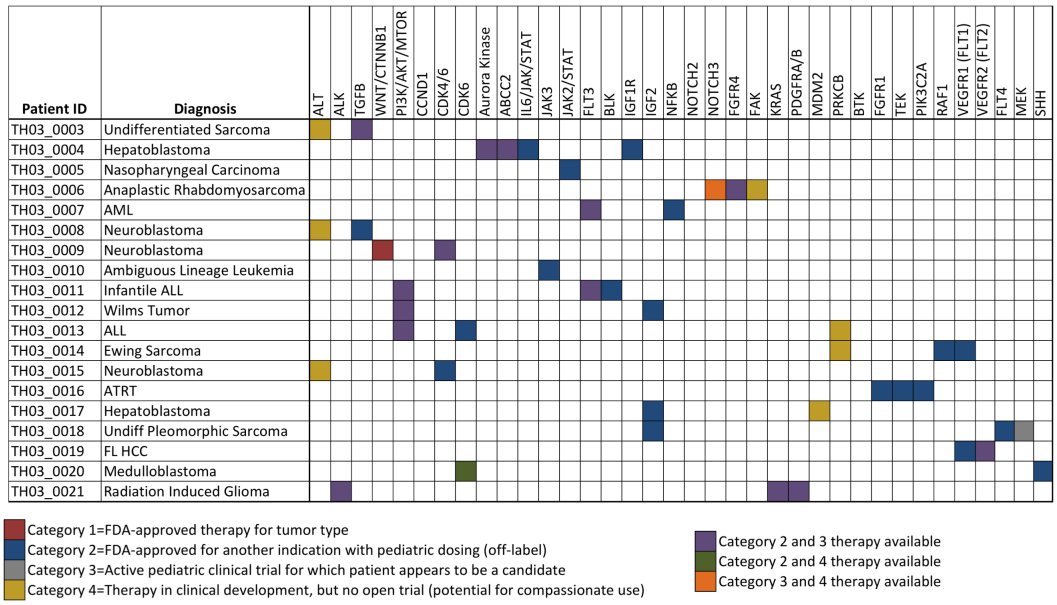


Figure 4.5: Results of the CKCC analysis on 19 individuals. The therapeutic leads are color coded by the type of the recommendation. The columns provide information about specific molecular pathways for which the lead was found.

4.2.2 CKCC Data Preprocessing

We collected The Cancer Genome Atlas (TCGA) [28] and Therapeutically Applicable Research To Generate Effective Treatments (TARGET) [10] RNA sequencing gene expression data (normalized read counts) from the public repository in Xena [86, 67]. In order to minimize batch effects introduced by the source datasets the RNA sequencing data from each project was re-processed in a uniform way across the entire cohort. Both datasets were processed using STAR2 [42] for sequence alignment and RSEM [77] for gene expression estimation by the UCSC Genomics Institute [67]. We extracted tumor-only samples from the TCGA dataset, then combined these gene expression data into a single cohort (n = 10,369, 58,582 gene features). I hereafter refer

to this combined TCGA and TARGET dataset as *CKCC reference* cohort.

Identifying and eliminating those features that obstruct signals in the data is an established technique in the field of machine learning. For instance, features that do not vary enough across all the observations are not informative for cohort-wide analysis. Similarly, features that are only present in a small number of observations can make the cohort-wide patterns fuzzy and make it difficult to extract them. In order to eliminate "noisy" gene features in the CKCC reference dataset I performed two levels of gene filters. First, I filtered out all the genes that were expressed in fewer than 20% of all samples. Second, from the remaining genes I filtered out 20% of least varying genes. These filters left 26,969 gene features.

4.2.3 Methods for Assessing CKCC Map Robustness

One of the important considerations of a reference map is the robustness of the sample placements in the reference map. In other words, we want to measure the stability of the map under small perturbations to the cohort space and feature space from which the maps are built. We make inference about the relationships of the tumor samples in the original high-dimensional space based on how they relate to each other in the projected 2-D map space. Therefore, it is important to asses how robust those relationships in the 2-D projection are. In this section I describe the proposed framework for assessing the map robustness (Figure 4.6). In short, I propose to measure:

1. Local neighborhood robustness - *how stable are the nearest neighbor memberships?*

2. Global layout robustness - *how stable is the global structure of the map?*

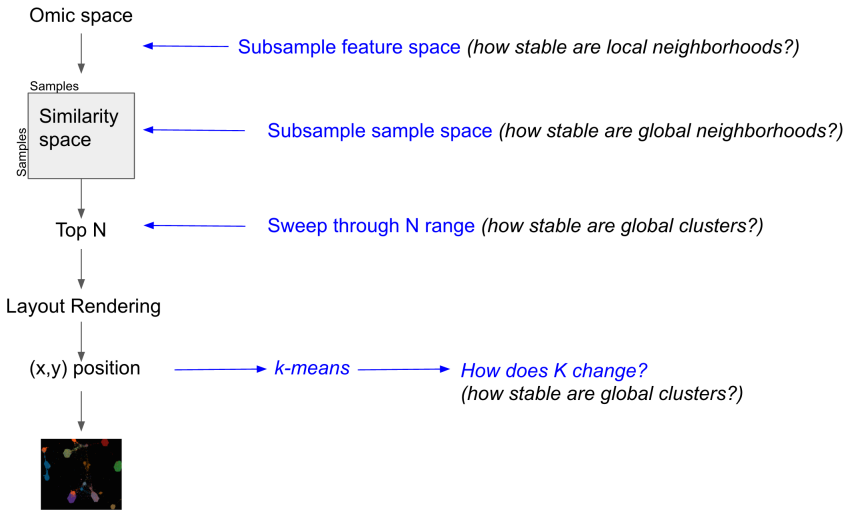


Figure 4.6: High level outline of how we assess the stability of the reference map and robustness of the sample placements in the map.

4.2.3.1 Local Neighborhood Robustness

Any time we make an inference based on the concept of a "local neighborhood" we want to understand how stable that neighborhood is under reasonably small perturbations to the data. A "local neighborhood" for a given observation is the top N observations (excluding self) that are most similar to this observation. If the local neighborhood drastically changes due to small changes to the dataset then such local neighborhood is not robust. Tumor Map belongs to the k-nearest neighbor family of methods. Therefore, assessing the stability of the local neighborhoods (top N neighbors) is important in order to understand how robust the CKCC map is. Here I describe the approach of measuring both sensitivity and variance of the local neighborhoods under

feature space perturbation (Figure 4.7).

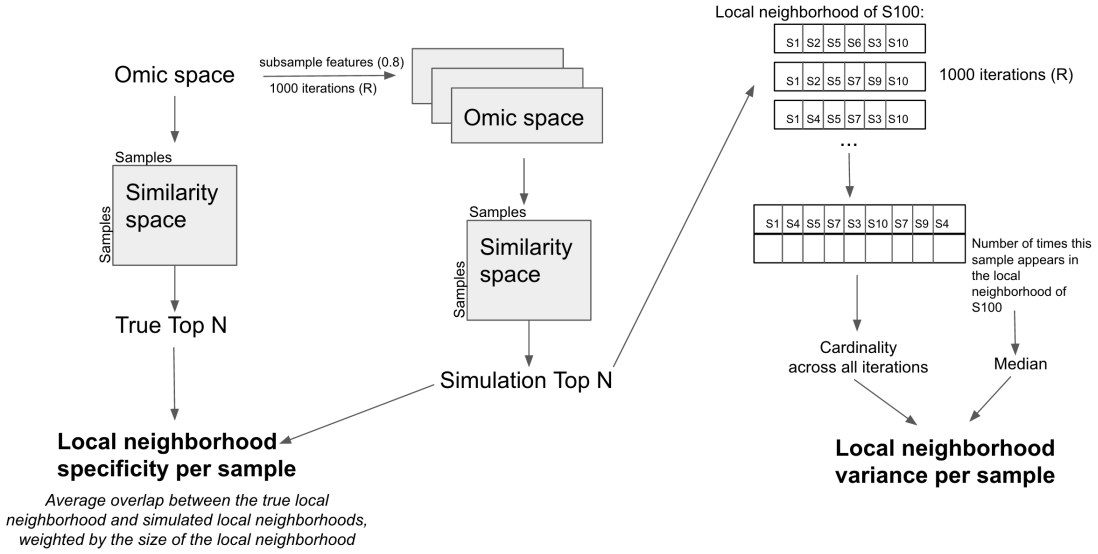


Figure 4.7: Overview of the method for assessing local neighborhood robustness: local neighborhood specificity (bottom left) and local neighborhood variance (top and bottom right).

Subsampling Feature Space

It has long been known that the results obtained from unsupervised data modeling (e.g. Tumor Map, hierarchical clustering, etc.) are highly dependent on the original feature space [111]. This is because how similar any two samples are may depend on the particular features that are being considered. Since most unsupervised methods utilize some measure of similarity or distance, feature space is an important factor that affects the results of those methods. If observations drastically change their cluster memberships under small changes to the feature space then we are not confident in such solution. On the other hand, if a clustering solution is relatively stable under varied feature spaces then we are confident in the solution. It is a common practice in the

field of bioinformatics to apply a technique called "subsampling" [119, 122]. We use 26,969 HUGO gene features (see 4.2.2) as our feature space and subsample it at 80%. We repeat feature space subsampling 1,000 times.

Alternatively to subsampling, bootstrapping procedure can also be used in all of our robustness discussions here. Similar to subsampling, bootstrapping is a process of drawing from the original observation space but with replacement. Therefore, we produce a new set of observations of the same cardinality as the original space we drew from but the observations in that space can be repeated. In the case of the local neighborhood robustness, we would produce a matrix of the same dimensions as the original genomic matrix with some gene features possibly duplicated in that space.

Measuring Specificity of the Local Neighborhoods

A specificity of a local neighborhood of a given sample can be thought of as "how often under feature space subsampling the sample recapitulates the same local neighborhood?". I define local neighborhood specificity for a given cohort sample j as:

$$SP_j = \frac{1}{N} \sum_{i=1}^{N} \frac{S_j^{true} \cap S_j^i}{|S_j|} \tag{4.1}$$

, where N is the number of times the feature space was subsampled, S_j^i is the neighborhood of sample j under subsampled feature space i, S_j^{true} is the true local neighborhood for sample j in CKCC reference cohort, and $|S_x|$ is the cardinality of the local neighborhood of sample x.

A similar approach was taken by Taskesen *et al.* [128] in their comparison of

two different solutions based on a method from a k-nearest neighbor family. The metric SP_j is scaled on the interval $[0, 1]$, where 1 indicates high specificity and 0 indicates low specificity. This local neighborhood specificity metric can be interpreted as a measure of how specific local neighborhood (as a set of observations) to a particular sample in the given genomic space.

I computed local neighborhood specificity metric for the CKCC reference cohort (N = 6) under both feature subsampling at 80% and feature label shuffling (Figure 4.8). Both subsampling and shuffling were repeated 1,000 times. I found that majority of samples retain between 90% to 100% of their true local neighborhood under the condition of feature space subsampling while almost no local neighborhoods were preserved under the gene label shuffling. This finding suggests that even under incomplete gene feature space we recapitulate local neighborhood structures for the samples in the CKCC reference cohort.

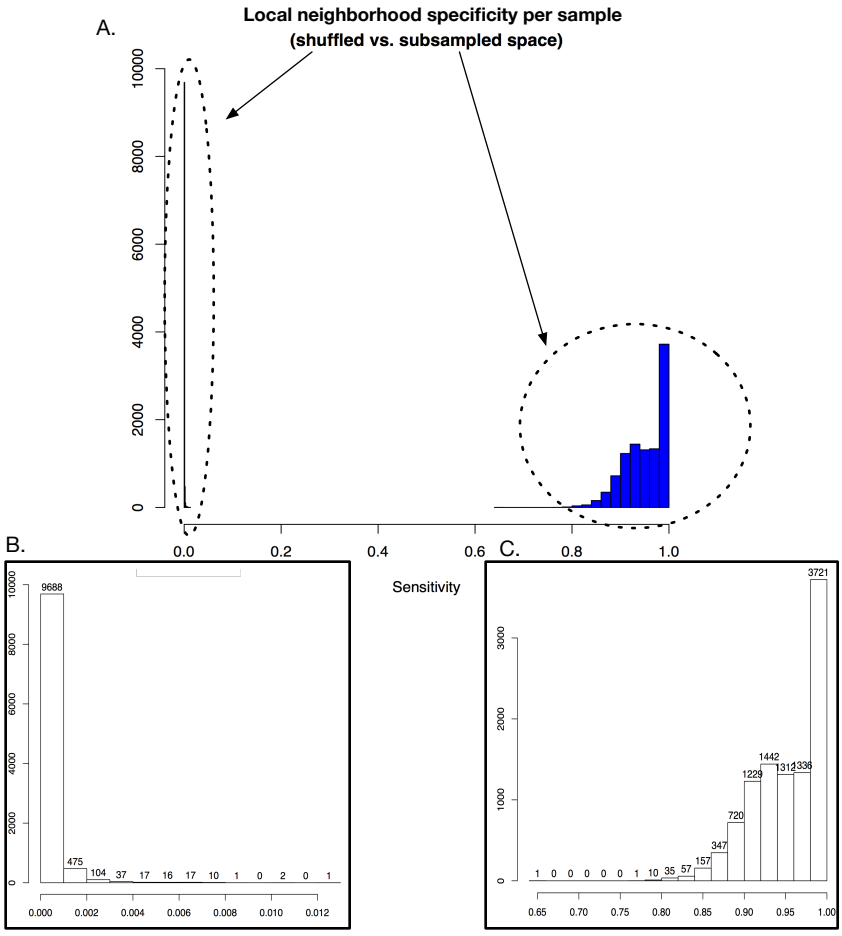


Figure 4.8: Assessment of CKCC version 1 local neighborhood specificity across cohort samples under feature space subsampling (C) and under feature space shuffling (B). Most samples retain between 90% to 100% of their true local neighborhood under the condition of feature space subsampling.

Measuring Variance of the Local Neighborhoods

In addition to measuring how often the local neighborhood remains the same, we can also measure how much the local neighborhoods vary by comparing them under each feature space subsampling for a given sample j. Figure 4.7 describes how we compute this metric. For each feature space subsampling we look at the local neighborhood

and count how many times each sample appeared in the sample j's local neighborhood. This allows us to assess how consistently each sample becomes a local neighbor of sample j. Even with imperfect local neighborhood specificity it is possible to have low local neighborhood variance, suggesting that even if the local neighborhoods under feature space subsampling do not always remain the same as the true local neighborhood they are still pretty consistent across these feature space perturbations. On the other hand, if the local neighborhood variance is high we conclude that these neighborhoods are highly dependent on the feature selection and we cannot be confident in them.

I computed local neighborhood variance for the CKCC reference cohort (N = 6) under both feature subsampling at 80% and feature label shuffling (Figures 4.9 and 4.10). When looking across 1,000 iterations of feature space subsampling we see that for most samples local neighborhoods do not vary at all (Figure 4.9). Most samples appear in a local neighborhood of a given sample 1,000 or close to 1,000 times, suggesting consistency across iterations. In the future, as an alternative we can weight each sample by the number of times it occurs in the local neighborhood of the focus sample. This will help provide easier interpretation to this statistic.

When considering the cardinality of the set of all samples that appear in a local neighborhood of a given sample across 1,000 iterations, we see that it is very close to the cardinality of the true local neighborhood (N = 6). This suggests that the same samples tend to appear in the local neighborhood across all iterations and is consistent with the local neighborhood specificity metric (see 4.2.3.1 *Measuring Specificity of the Local Neighborhoods*). Under the gene label shuffling this cardinality is measured in

1,000's. As a side note, in the future we could normalize the local neighborhood variance metrics by the number of subsampling iterations (1,000 in the described above case) for an ease of interpretation.

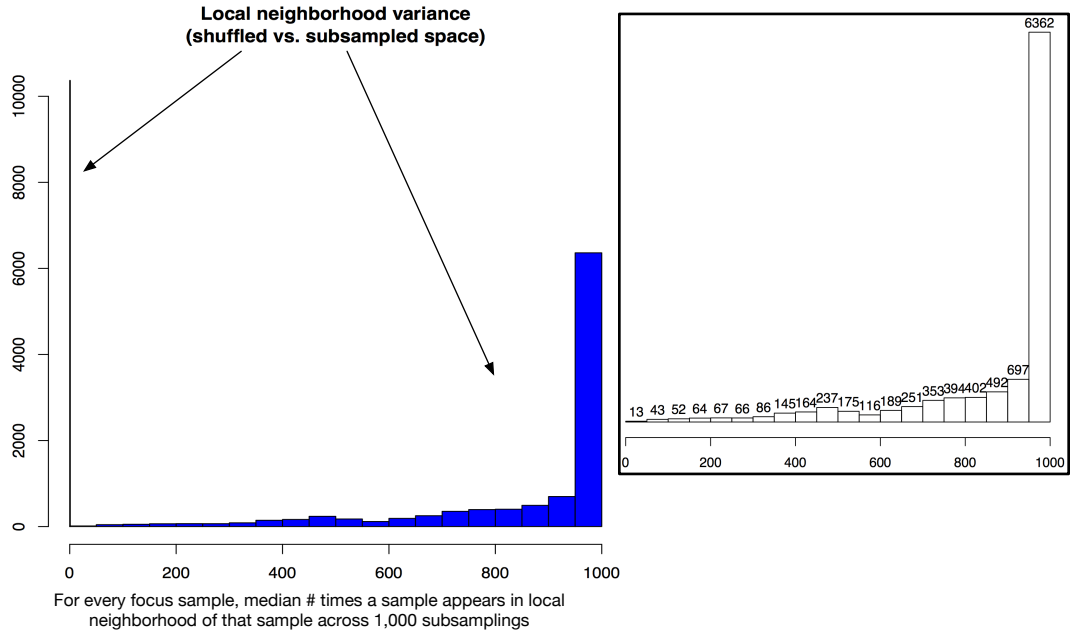


Figure 4.9: Median number of times samples appear in the local neighborhood of a given sample over 1,000 iterations of either subsampling of the feature space (bar plot on the right) or shuffling of gene labels (bar plot on the left). For most samples local neighborhoods do not vary at all over the feature space subsampling iterations.

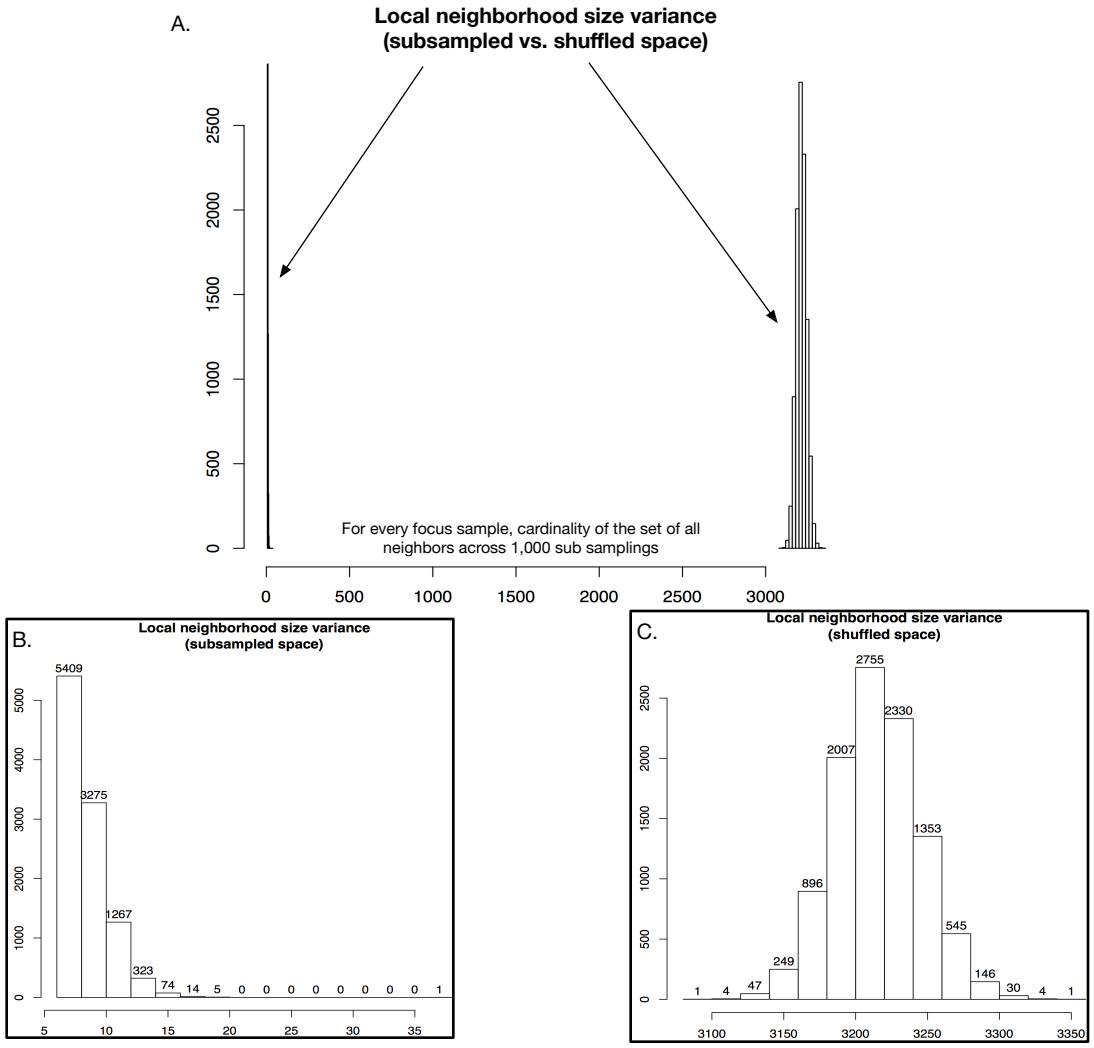


Figure 4.10: Total number of samples that appear in the local neighborhood of a given sample over 1,000 iterations of either subsampling of the feature space (B) or shuffling of gene labels (C). For most samples under the subsampling conditions the cardinality of the local neighborhoods across all iterations is very close to the true size of the local neighborhood (N = 6), suggesting that these neighborhoods do not vary much as compared to the gene shuffling conditions where cardinality is in several 1,000s.

Conclusion: Local Neighborhood Robustness

Tumor Map method belongs to k-nearest neighbors family of methods, there-

fore it is important to asses the robustness and stability of the local neighborhoods produced from the CKCC reference cohort. I propose two ways to measure this stability: through local neighborhood specificity and local neighborhood variance. I propose to use feature space subsampling technique in order to perturb the original high-dimensional genomic space. For each metric, I contrast the results of feature space subsampling to feature label shuffling (fully random model). I showed that CKCC reference cohort produces stable local neighborhoods, which robustly retain the sample memberships across iterations of perturbing the dataset.

4.2.3.2 Global Layout Robustness

In addition to assessing how stable the cohort's local neighborhoods are it is important to assess the robustness of the global structure of the map (Figure 4.11). This assessment is important because it evaluates whether the tumor sample clusters in the map are consistent across similar cohorts and do not arise simply by chance.

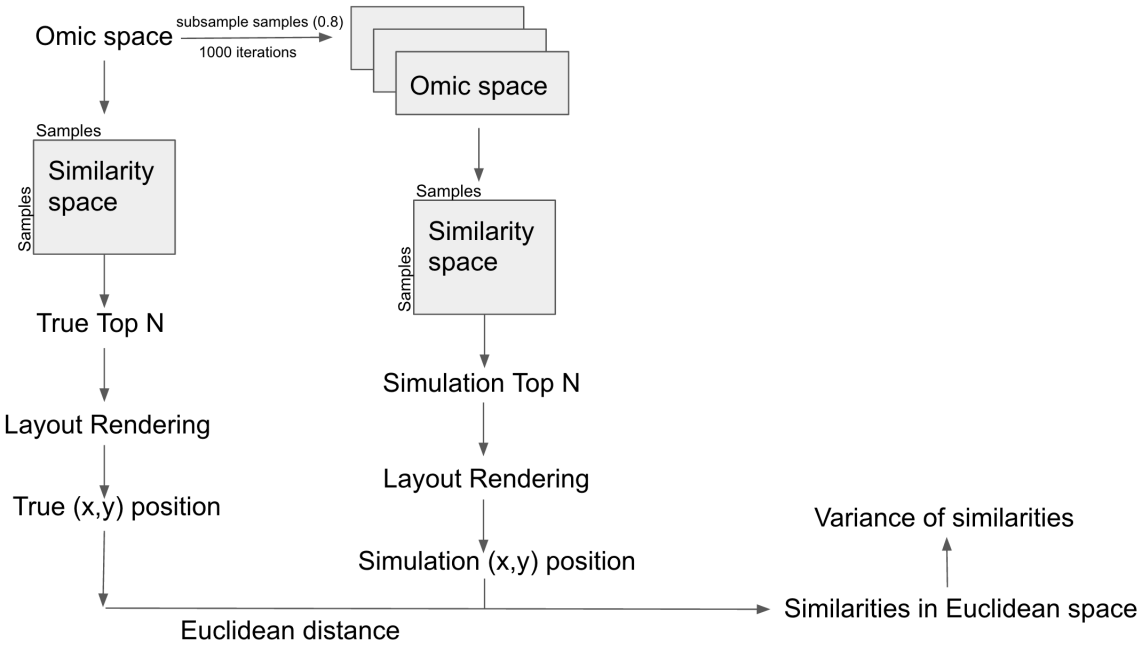


Figure 4.11: Overview of the method for assessing global neighborhood robustness: global neighborhood sensitivity.

Subsampling Sample Space

Similarly to subsampling feature space described above (see 4.2.3.1 *Subsampling Feature Space*), the space of observations (or tumor samples) can be subsampled. While leaving out samples from the cohort would alter the relationships between samples in the map, we contend that leaving out a small number of observations should not drastically change the map topology. Additionally, we utilize the observation that RNA sequencing-based maps appear to be driven by the tissue of origin as a dominant signal. Armed with that observation, I subsample 1,000 times at 80% of each tissue type. I originally explored subsampling at 80% of the entire cohort. However, this has potential of removing entire tumor types from the map (e.g. cholangiocarcinoma tumors

are only represented by 36 samples. Avoiding introduction of new uncertainty into the map structure I made a decision to subsample within each tissue type. After subsampling the cohort I recompute local neighborhoods and render a separate map layout for each cohort subsampling. The rendering of the layout produces (x,y) coordinates for the samples in the map. For each rendering I perform unsupervised clustering based on the Euclidean distances computed from (x,y) positions of the samples in the map. I explored several clustering methods. One of them was k-means clustering, with which I use the Silhouette score [112] method to compute the optimal number of clusters. Affinity propagation clustering [21] automatically selects k. Other clustering methods can be explored in the future as needed. Once the clustering is performed, I compare the optimal number of clusters across all iterations to the true number of clusters computed for the CKCC cohort. We postulate that the number of clusters should not vary greatly across subsampling of the cohort samples.

Again, as discussed above for the local neighborhood robustness , we could use the bootstrap procedure here where we could produce a proxy to our original space of tumor samples by sampling with replacement and producing a genomic matrix with some tumor samples repeated in that space.

4.2.4 N-of-1 Tumor Map Placement

The first step in analyzing a single tumor in the context of a larger reference cohort of cancer samples is to understand where it fits among those samples. Understanding which cancer samples this individual's genomic profile is most similar to can

guide our molecular investigation and direct us to new potential therapeutic avenues. Tumor Map method (see 2.1) projects high-dimensional genomic landscape of a tumor cohort into an easily visualized 2-D map to aid further investigation and interpretation of tumor relationships. We utilize Tumor Map of the "frozen" cohort (Figure 4.4) to place the placement of N-of-1 sample into that map without re-generating it with the new sample. We visualize this placement using a "landmark pin" used by navigational maps in Google Maps application[4] (Figure 4.12).

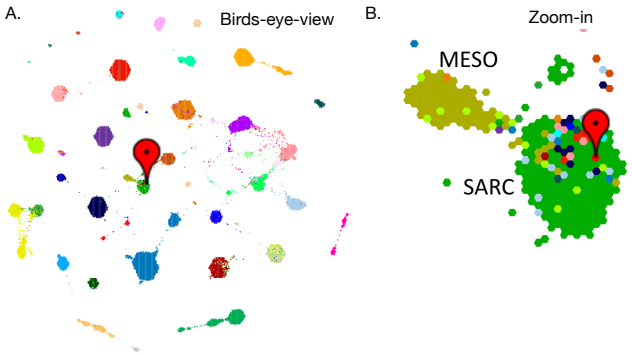


Figure 4.12: Example of a "landscape pin" indicating an N-of-1 placement into Tumor Map. A) A birds-eye-view of the entire reference cohort map. B) A zoom-in into the area of the map where the pivot sample is placed. In this example the pivot is placed into with sarcoma tumors. Mesothelioma tumors, which are biologically similar to sarcomas (see 3.4), are near by.

4.2.4.1 Placement Based on Nearest Neighbors

As a part of rendering the map, Tumor Map method generates (x, y) coordinates in the Euclidean plane for every sample in the cohort. For the N-of-1 sample (also referred to in this text as *pivot*) I compute similarity of the RNA sequencing profile of this sample with every sample in the reference cohort. From this similarity space, I

select 6 nearest neighbors of the pivot and compute the centroid of the map positions of these neighbors, based on the Euclidean coordinates generated by Tumor Map. I utilize 6 nearest neighbors because that is the number of neighbors used to build the "frozen" map of the reference cohort. To compute the centroid I use the median value of the x and the median value of the y coordinates of the 6 nearest neighbors.

4.2.4.2 Future Directions for N-of-1 Map Placement

While it has led to some promising discoveries for the CKCC individuals, the current method of N-of-1 placement into the map is not ideal. First, blankly selecting 6 nearest neighbors ignores any information the 7^{th}, 8^{th}, and so on neighbors might carry that are useful for the pilot sample's molecular investigations. Second, the distances between the samples in the map are a proxy for the relationships between the samples in the original space. However, the map positions are not always directly reflective of the nearest neighbors. Because the map is built by taking into account *all* the neighborhoods and aims to achieve the optimal energy configuration for the whole system rather than any particular neighborhood or a subset of nodes (see 2.1 Methods), the final map positions sometimes do not correlate with every nearest neighbor configuration (in other words, sometimes a sample X that appears in the local neighborhood of sample Y is not near sample Y in the map). Additionally, the local neighborhood relationships are not symmetric (just because sample X appears in the local neighborhood of sample Y does not mean that Y will appear in the local neighborhood of X).

Therefore, the method for placement of N-of-1 sample into Tumor Map de-

mands a closer consideration and further investigation. Improving and developing this method is a part of the ongoing work being done as a part of the CKCC initiative. Here I outline two possible avenues to explore as our future work.

Model-based Placement

The discipline of machine learning has been invaluable to the field of bioinformatics. Many successes in the field of cancer genomics have been due to ability to build and apply supervised models or obtaining results through unsupervised analysis. Exploring a model-based approach in this case makes sense since our Tumor Map placement model perfectly fits a classical regression formulation. We begin with a high dimensional genomic space and end up with a low 2-D space. We want to be able to predict the map's (x, y) coordinates from a high-dimensional genomic space (gene expression in this case). Therefore, there are two continuous outcome variables we want to predict/model from 26,969 (see 4.2.2 for details on how this number was obtained) continuous gene expression values. This is a typical multi-variate multiple regression problem. However, it is still a complex problem to find the appropriate model that will accurately make such predictions.

Our preliminary results indicate that a simple linear regression model might not be appropriate in this case. High-dimensional genomic space might be too noisy to accurately model it in just 2 dimensions. Deep learning is a an emerging area in the field of machine learning and has shown much promise in ability to model complex non-linear relationships as well as has shown promise in applications in bioinformatics [88]. Currently, several software libraries [8, 6, 9] provide implementations of the deep learning

framework that makes it easy to design, build, and run your own models. My initial investigation showed that programming and engineering efforts involved are reasonably small. However, the problem of model selection here is very complex. There are a lot of parameters one can vary with deep learning models (input space, number of hidden layers, configuration of hidden layers, introducing convolution, introducing connection pruning in the hidden layers, and many more). The problem of hyperparameter space search is complex and should be approached with care. This effort is a part of the proposed future work for this project.

As with any machine learning model, the question of evaluation is of importance. Generally, regression-type models are evaluated using mean squared error (MSE) - a measure of deviation of the predicted values from the true values. It is also a common approach in machine learning to leave out a validation set (including leave-X-out) or perform some form of cross-validation. In our case we could employ leave-X-out strategy to leave a set of samples out of training the model and make predictions for those samples. To evaluate the model we can look at the MSE. We can also compute a correlation coefficient between the predicted values and the true values. This process can be repeated a fixed number of times to derive a distribution of the evaluation metric.

Similarity Matrix Clustering

The biggest drawback of the nearest neighbor based Tumor Map placement approach is that we have no principled way of knowing how many neighbors to look at in order to define a local neighborhood. I propose to investigate clustering of the similarity matrix as a way of defining the pivot's local neighborhood. The "frozen"

reference map is built from a similarity space of genomic profiles of the samples in the map. We also compute similarities between the pivot and every sample in the reference dataset. I propose to use methods often utilized to cluster graph structures, as a similarity space can be thought of as a graph connectivity space. Methods like affinity propagation (AP) clustering [21] and spectral clustering [118] have long been used to cluster similarity/adjacency space. Other clustering methods can be investigated here as well. Some methods (e.g. AP clustering) automatically find the number of clusters in the data, while others need help from a Silhouette [112] or other similar k-selection methods. Once the clustering is performed and the number of clusters is finalized, all the samples in the cluster where the pivot sample is are the pivot's neighbors.

4.2.5 Additional Future Directions

While already achieving promising results, CKCC initiative is still in its infancy stages and is an ongoing effort to improve and develop current approaches. Above I proposed some new directions for the N-of-1 Tumor Map placement method. Below I propose several additional future directions that will greatly contribute to already put forward efforts.

4.2.5.1 Methods for Assessing N-of-1 Placement Robustness

It is important to be able to access how good the placement of the pivot sample into the frozen map of the reference cohort is. This is a different problem than assessing the robustness of the map itself. This is a problem of assessing the robustness

of the placement for a particular individual's genomic sample. Several things could be driving the pivot placement and we have to understand if these things are relevant to the potential treatment of the tumor. For example, if the placement is being driven purely by the tissue of origin then the placement does not provide any useful information and, furthermore, can confuse the molecular investigation and point to therapeutic leads that should not be considered. For example, we noticed that diffuse intrinsic pontine gliomas (DIPG) often cluster with adult glioma and glioblastoma tumors when by many experts they are considered biologically and molecularly different tumors. This placement is driven by the tissue of origin of these tumors. Gliomas and glioblastomas just happen to be the most similar tumors in the reference cohort to DIPGs but they should not be placed together. This is one of the big drawbacks of the local-neighborhood-based placement methods. There are always "top" neighbors for a pivot sample. How good those neighbors are is another question. So, we need to assess if the similarities we are seeing between the reference cohort nearest neighbors of the pivot have "good enough" similarity with the pivot. To check for the strength of the tissue signal we are going to incorporate computing similarities of the pivot sample with GTEx [5, 35] data, which is a cohort of RNA sequencing profiles from normal tissue samples. As a part of the CKCC initiative, we re-processed GTEx RNA sequencing data with the same pipeline as the TCGA and TARGET data but have not utilized it yet. If the similarities of the pivot with GTEx samples, especially of the same tissue as pivot, are dominating the top neighbors of the pivot then we can conclude that the signal we are observing is driven by the tissue. If, on the other hand, we find that CKCC reference cohort samples are

still the highest neighbors of the pivot then we are probably observing a tumor-driven molecular signal.

Additionally, I propose to utilize local neighborhood robustness metrics (see *Local Neighborhood Robustness*) in relation to the pivot's neighborhood. These metrics utilize subsampling of the feature space and can measure how consistent (specificity and variance) the local neighborhood of the pivot is. In fact, the same 1,000 subsamplings computed to assess the robustness of local neighborhoods in the map can be utilized to compute local neighborhood robustness for the pivot sample. Some additional avenues for measuring N-of-1 placement robustness will be explored as my work for CKCC will continue.

4.2.5.2 Neighborhood Analysis

The advantage of using a large library of tumor genomics data as a reference cohort is not only that we can look at the most similar tumors but we can also test for enrichment of genomic, clinical, and phenotypic markers in the samples similar to the pivot sample. Similar to Gene Set Enrichment Analysis (GSEA) [124] we can test for occurrence of certain events in the samples that are most similar or least similar to the pivot. These events could be anything from genomic events (e.g. mutation) to diagnostic events (e.g. histology label) to cell phenotype indicators (e.g. pathway activity flag). These events may not occur in the closest neighbors of the pivot but are concentrated at a non-random rate in one tail of the similarity distribution. These enrichments indicate statistically significant deviation from a uniform distribution of the

event we would expect by random. We can employ systematic ways of scanning for and identifying these enrichments. Furthermore, we can incorporate prior knowledge about how common the particular event is by developing a Bayesian approach (see below) to this screening.

Formulation of Bayesian Framework for Neighborhood Analysis

Given input query sample q (represented by feature vector X_q), a background/reference matrix of features by samples X^n for n samples, and a database of binary attributes A for the samples in the reference matrix, we describe a method to predict an attribute A_k in sample q based on the similarity of it to other samples $[1, ..., n]$ in genomic space.

$$A_{kj} = \begin{cases} 1 & \text{if sample } j \text{ has attribute } k \\ 0 & \text{if sample } j \text{ lacks attribute } k \\ 0 & \text{if sample } j \text{ has unknown status of attribute } k \end{cases} \tag{4.2}$$

We compute:

$$S_{kq} = P(A_{kq} = 1 | X_{(+q)}, A_{k,(-q)}) = P(A_{kq} = 1 | X_1, ..., X_n, X_q, A_{k1}, ..., A_{kn}) \tag{4.3}$$

, where $P(A_{kq} = 1)$ is the probability of the query sample having attribute A_k. We use similarity space of all samples in X to query sample X_q and test for attributes enriched in the neighbors of X_q. Let R_{ij} be a measure of relatedness or similarity between two given feature vectors. We compute R_{ij} kernel space as:

$$R_{ij} = f(\bar{X}_i, \bar{X}_j) \tag{4.4}$$

Now S_{kq} can be defined as a function of R_{ij}:

$$S_{kq} = P(A_{kq} = 1|X_{(+q)}, A_{k,(-q)}) = P(A_{kq}|\bar{R}_q, A_{k,(-q)}) \tag{4.5}$$

, where $\bar{R}_q = [R_{q1}, ..., R_{qn}]$. We associate attribute A_k with query sample q if $A_k = 1$ in the neighbors of q (the attribute is present in samples similar to q). We use a statistical test L_{kq} to record association of attribute A_k with sample q based on $\bar{R}_q$. An example of such a statistic is GSEA or a KS test. Given this new measure of association an attribute with the query sample, we again redefine S_{kq} as a function of L_{kq}:

$$S_{kq} = P(A_{kq} = 1|X_{(+q)}, A_{k,(-q)}) = P(A_{kq}|L_{kq}) \tag{4.6}$$

Now, let a^+_{kq} be a set of attributes where $A_{kq} = 1$ and let a^-_{kq} be a set of attributes where $A_{kq} = 0$. Using the Bayes rule, we redefine S_{kq}:

$$S_{kq} = \frac{P(L_{kq}|a^+_{kq})P(a^+_{kq})}{P(L_{kq}|a^+_{kq})P(a^+_{kq}) + P(L_{kq}|a^-_{kq})(1 - P(a^+_{kq}))} \tag{4.7}$$

$P(L_{kq}|a^+_{kq})$ can be estimated by computing $P(L_{kj}|a^+_{kj})$ for $j \in [1, n]$ from the reference data.

4.2.5.3 Derived Feature Spaces

Currently we utilize only RNA sequencing data in our work in CKCC. While providing important view of the cell state, RNA sequencing profiles are only one view of it. The more views of the cell state we can incorporate into our analysis the more information we can extract from the data. If we see the same result coming up across different views then we have even more confidence in that result. Different views of the cell state can provide alternative and complimenting signals about the it. While it is nice to collect data from multiple experimental platforms as it was done in TCGA project, it is not always possible. While obtaining mutation data is one of our goals, currently RNA-Seq data is all we have access to. However, there are various transformations and derived views we can still obtain from RNA-Seq. One of such transformations is a method called VIPER [83], which takes gene expression data for an individual sample and transforms it into transcription factor (TF) activity scores, based on expression levels of the downstream targets of that TFs. We successfully employed this method in other projects. TF activities provide an important view of the cell state and represent promising therapeutic targets as they tend to sit at the top of regulatory networks and simultaneously regulate many pathways and cell functions. Another computational transformation method is called PARADIGM [133] and, while usually integrates gene expression as well as copy number data, can be run only gene expression data only. This method outputs inferred pathway level (IPL) activities, proxies for how active each gene is, given the regulatory network relationships between the gene products and

cell functions. Although copy number data provides an additional view of the cell state, PARADIGM results can still be very useful when based on gene expression data only as the method incorporates knowledge about gene interactions in a molecular signaling network. The method aggregates evidence for how active a particular gene is based on its own expression level as well as the evidence from the neighboring expression levels.

4.2.6 Tool Deployment Can Help Others Use Our Tools

We not only want to automate our methods and tools and make them easy to run, we also want to make them deployable to third parties. Docker is a platform that allows configure, build, and deploy entire systems and/or applications. Tools available through Docker [3] platform do not require special installation or setup. They are deployed as self-contained packages and can be run within a virtual environment independent of the user's operating system, hardware, or system setup. As a part of my work with CKCC I dockerized a number of Tumor Map related functionalities and made them available through Quay [7] docker repository (under hexmap_ucsc user at `https://quay.io/repository/`).

4.3 Chapter Conclusion

In this chapter I describe my current and future work towards bridging the gap between the research community and the clinical practitioners. Currently, much of the academic cancer research does not make it into the clinic, even when it can help

affect patient care. While great strides have been made and much success has been achieved in the field of personalized medicine in general, and personalized genomics more specifically, in both cancer and non-cancer applications, there is still a dire need fore more bioinformatics tools and methods in a single patient setting. My work not only demonstrates that pan-cancer analysis has promise in clinical applications but, as a part of the CKCC initiative, I am already contributing to an ongoing collaboration with treating clinicians here in California. In this chapter I also describe possible future directions and possibly inspire others to contribute to this field.

Chapter 5

Additional Work and Future Directions

In addition to the work described in chapters 2, 3, and 4, I participated in some other projects I, for various reasons, did not describe as my main work. In this chapter I describe those projects and suggest future directions for projects that are not completed.

5.1 Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer

I participated in a multi-institutional multi-team collaboration to analyze transcriptomic and phosphoproteomic data for metastatic castration-resistant prostate cancer (CRPC) patients. Drake *et al.* were able to show that adding phosphoproteomic

data helps in deriving patient-specific aberration networks and brings in information not available in the transcriptomics data. We published the results of our work in the journal Cell [69].

5.1.1 Metastatic and Primary Prostate Tumors Separate In Transcriptional Space

The genomic data for this project came from the work by Grasso *et al.* [55]. These are rapid autopsy tumor samples from prostate patient. This dataset includes both metastatic and primary samples, as well as some benign tumors. Unfortunately, phosphoproteomic data in this study is only available for 27 samples (16 of which are CRPC tumors), we could not integrate both data types for all the samples. I built a visualization of the transcriptomic space of Grasso *et al.* tumors in the context of The Cancer Genome Atlas (TCGA) cancers (Figure 5.1). I used 13 different cancer types (12 previously published cancer types in a pan-cancer study [59] and TCGA primary prostate adenocarcinoma (PRAD) [94]). We noticed that in the gene expression space Grasso tumors do not cluster with TCGA PRAD tumors. This could be a possible batch effect or it could reveal real biology, since Grasso *et al.* data are autopsy samples while TCGA PRAD data is from live patients. Since a number of genes get turned off and on after the organism dies, we cannot be sure that this separation is not driven by the biology of the cells. The interesting finding, however, is that within Grasso *et al.* tumors there is a clear separation between metastatic and primary/benign tumors. While benign tumors cluster with primary tumors, they locate towards the side of the

cluster that is the furthest from the metastatic tumors. This is an interesting find because it suggests that:

1. Metastatic tumors have a distinctly different gene expression profile than primary tumors.

2. Benign tumors exhibit some characteristics of the primary tumors at the molecular level.

3. Benign tumors are least similar to metastatic tumors among all non-metastatic tumors.

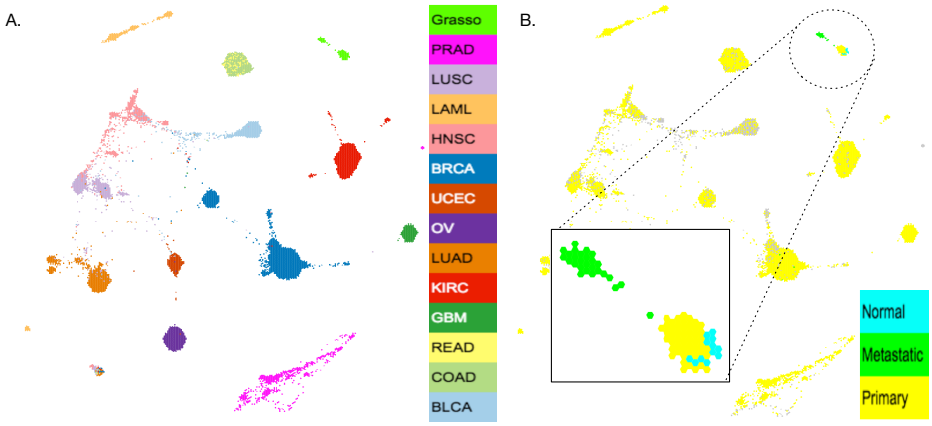


Figure 5.1: Visualization of the transcriptomic space of 13 different types of cancer. A) The map of the entire cohort. Tumors are colored by the tissue of origin. B) metastatic tumors separate from primary and benign tumors in the transcriptomic space of the Grasso *et al.* dataset.

5.1.2 Conclusion and Future Work

I participated in a collaboration in which we investigated if phosphoproteomic data can help derive patient-specific aberration networks for CRPC patients. I built a

visualization of transcriptomic space for both metastatic and primary prostate tumors in the context of a number of other tumor types. This analysis revealed several interesting findings. However, due to a lack of phosphoproteomic data for most samples in the cohort we were unable to build an integrated visualization of the entire cohort. If more phosphoproteomic data becomes available for this project or other projects, it will be interesting to build a visualization based on the phosphoproteomic feature space. It might lead to interesting findings if data for multiple cancer types are available for such analysis. Integrating transcriptomic and phosphoproteomic spaces into a single map might also lead to interesting findings.

5.2 Centromere Reference Models for Human Chromosomes X and Y Satellite Arrays

In collaboration with several colleagues I worked on the task of developing a classifier to predict human populations from the centromeric DNA sequences. Centromeric DNA belongs to a class called "satellite DNA". These DNA sequences are large arrays of tandem repeats (Figure 5.2) and vastly occur in centromeric and telomeric regions. These are regions of DNA that do not contain protein coding sequences but many siRNA and miRNA genes are found in satellite DNA regions. Satellite DNA is also not well characterized as for a long time it was considered "junk" DNA. Several recent studies implicate satellite DNA in a number of diseases [110, 107] as well as the aging process [46, 97]. In the scope of this project we wanted to characterize

centromeric DNA for DYZ3, Y ?-satellite family of centromeric sequences, and investigate how much the centromeric sequences differ between individuals/populations. We published our findings in the journal Genome Research [71].

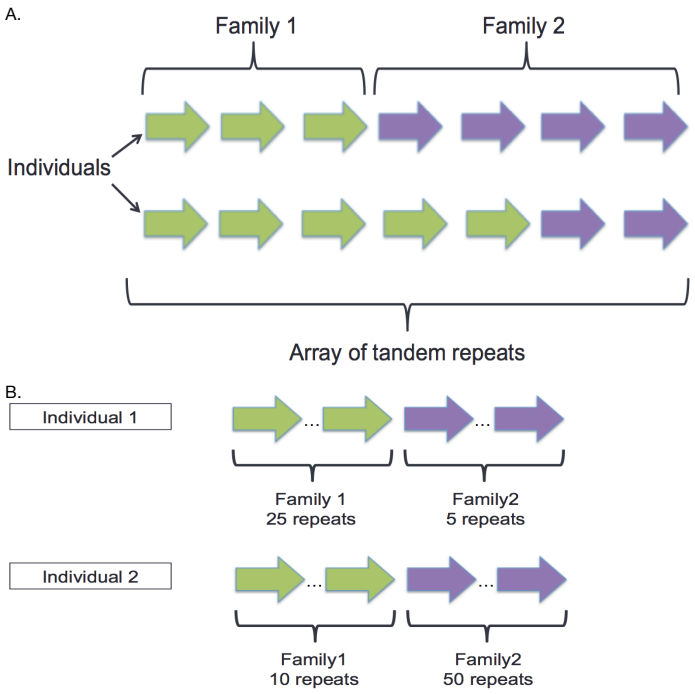


Figure 5.2: Satellite DNA consists of tandem repeats. These repeats may vary among individuals and populations and sequence kmer frequencies carry information about variation in these repeats.

5.2.1 Data

We obtained centromeric sequence data for 373 male individuals [113] from 1000 Genomes project [34] (http://www.internationalgenome.org/). As in the method described by Hayden *et al.* [58], we reformatted each linearized centromeric array into a k-mer library (k=24). We used a sliding window of size 24 and scanned the sequences

of every individual, and computed kmer frequencies for 9,808 kmers specific to DYZ3 satellite family. These 373 individuals represent 12 haplotypes and 72 clades of human populations.

5.2.2 SatNP Method Successfully Classifies Human Populations Based on DYZ3 Sequences

In collaboration with Miten Jain, a colleague, I developed SatNP method for classification human populations based on the DNA sequences characterized by kmer frequencies (Figure 5.3A). Our method incorporates feature selection to identify informative kmers for the model (Figure 5.3B). The classification model is a support vector machine (SVM). We also permute population labels and re-train the model on randomly labeled individuals. The random model is expected to perform poorly while the true model is desired to perform with a much higher accuracy.

We applied our method to build models to differentiate between different clades and different haplotypes. Figure 5.4 describes our results for one specific model, which differentiates between "R1b1b2" western European and "E1b1*" east African clades. We were able to identify 1,153 (out of 9,808) informative kmers for this classification task. Using 2-fold cross validation SVM model trained on the 1,153 informative kmers shows over 91% prediction accuracy while the model trained on permuted labels performed worse (42.5%) than expected at random. For a more detailed description of our methods and results see full publication [71].

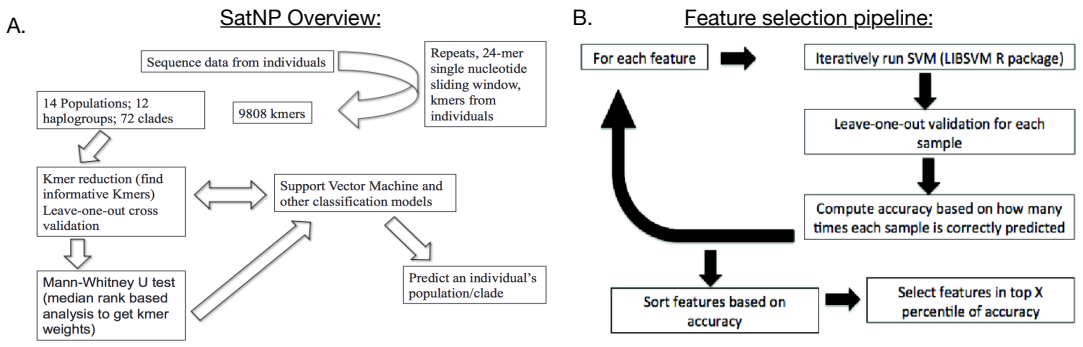


Figure 5.3: Overview of SatNP method. A) High-level diagram of the SatNP pipeline. B) Overview of the feature selection pipeline.

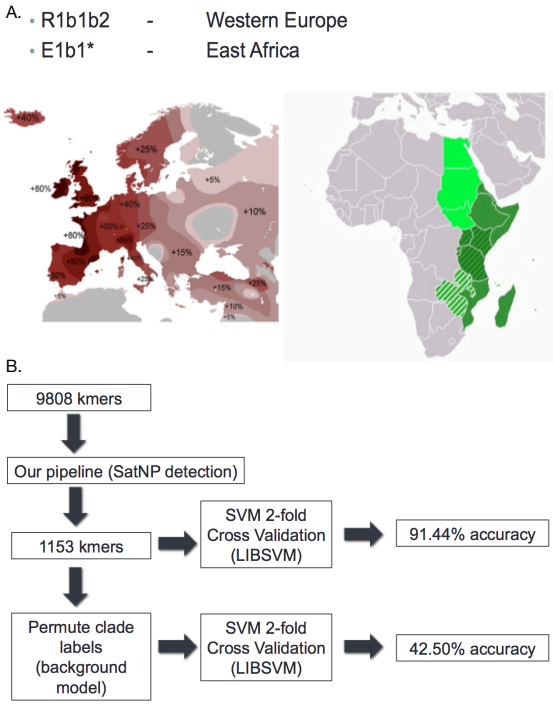


Figure 5.4: Results of our SatNP pipeline to differentiate between western European and east African populations. We identified 1,153 informative kmers and demonstrated our model makes predictions with high accuracy.

5.2.3 Future Direction

While we were able to obtain nice results that show satellite DNA sequences carry important signals in the composition of the tandem repeats they are made up from, we only concentrated on a small region of the satellite DNA. Systematically going through and characterizing every satellite DNA family would be a great contribution to the field. Other high-copy DNA regions, such as mitochondrial DNA, can also be characterized using our approach. SfatNP method can be utilized in a large number of applications, such as disease modeling. Increasingly the research and medical community recognizes that satellite DNA plays an important role in originating and driving of a number of diseases [110, 107], including cancer. Specifically, centromeric DNA is a key player in cell division. Disruptions in the centromeric sequences can affect kinetochore attachment. As we showed above, kmer frequencies can carry an important signal about these sequence disruptions. SatNP method can be applied in the cancer research settings. Additionally, SatNP method is not dependent on the particular feature space, so it can be applied in problem formulations where features are not kmer frequencies. In our preliminary work, we demonstrated (not shown here) that SatNP method can be applied to microarray gene expression data to diagnose chronic obstructive pulmonary disease (COPD) from gene expression profiles. We even identified 524 (out of 54,675) probes that are sufficient to diagnose COPD with 77.7% balanced accuracy, a higher accuracy than 72.5% when all probes are used. This increase in accuracy can be explained by the the fact that there is more noise in the data when all probes are used.

5.2.4 Conclusion

I took part in a study that aimed to characterize a specific region of satellite DNA. As a part of that study we developed an approach for this characterization process that can be applied to other regions of high-copy DNA (other satellite DNA regions and mitochondrial DNA). My contribution to the study was development of a novel method that uses the features characterizing satellite DNA sequences and extracts most informative features to model various human populations and subpopulations. We published our findings and methods in the journal Genome Research [71].

5.3 The Molecular Taxonomy of Primary Prostate Cancer

As a part of one of the analysis groups within The Cancer Genome Atlas (TCGA) initiative, I worked on analysis of primary prostate adenocarcinoma (PRAD) tumors. While my analysis did not directly contribute to any of the figures in the manuscript, I participated in several lines of analysis and presented on two weekly calls. Specifically, I analyzed gene expression (RNA sequencing platform) data for possible batch effect coming from different sequencing centers. I also performed and presented pan-cancer and pan-PRAD-study analyses. The group published our results in the journal Cell [11].

5.4 Pan-cancer Analysis of Small Cell Tumors Can Shed Light Into the Biology of This Tumor Type

Small cell phenotype is a type of histology that is observed in some tumors in various tissues. Generally, small cell tumors are considered to be more aggressive and often develop into a metastatic disease. This type of tumor arises in several tissues of epithelial origin (prostate, lung, ovarian, etc.). More generally, small cell tumors belong to a broader family of small-blue-round-cell tumors, which also incorporate such tumor types as Ewing's sarcoma/primary neuroendocrine tumor (PNET), neuroblastoma, Wilms' tumor, and some others.

I collected a number of datasets that have gene expression data for small cell histology (Figure 5.5A). Since we generally observe batch/dataset effect when we combine data from different studies we applied ComBat [68] method to the combined gene expression data (Figure 5.5B). Once the batch effect was removed, we proceeded to cluster the samples in the combined cohort using consensus k-means clustering [136] (Figure 5.5C). We observed that most of the castration resistant prostate cancer (CRPC) tumors cluster into a single cluster. Zooming in on that cluster (highlighted in the figure) reveals that tumors from some other tissue types cluster with the CRPC samples. This is an exciting finding as it suggests that these tumors also exhibit small cell phenotype and molecular data shows it. It also suggest that our batch effect removal pipeline removed enough tissue and study signal that we are now able to extract additional signals from the molecular data. Our batch effect removal pipeline can also be applied

to any pan-cancer and/or multi-study analysis.

A.

Dataset	# samples	Platform	Tissue	Description
GSE32967	22	Microarray	Prostate	CRPC xenograph; 8 adenos, 5 large cell neuroendocrine, 9 small cell
GSE33277	12	Microarray	Prostate	2 small cell, 9 adenos and 1 intermediate/mixed prostate xenograph
GSE43346	23	Microarray	Lung	small cell lung cancer tumor
GSE49887	4	Microarray	Ovary	small cell ovarian tumor
WCDT prostate	52	RNASeq	Prostate	WCDT CRPC
TCGA PRAD	498	RNASeq	Prostate	TCGA prostate adenocarcinoma, tumor only
TCGA LUAD	512	RNASeq	Lung	TCGA lung adenocarcinoma, tumor only
Pfeiffer et al.	27	RNASeq	Prostate	Small cell prostate
GSE63459	34	Microarray	Lung	Stage I adenocarcinoma lung
GSE54807	25	Microarray	Ovary	Clear cell and serous ovary
GSE5132	32	Microarray	Prostate	Prostate adenos with various gleason scores

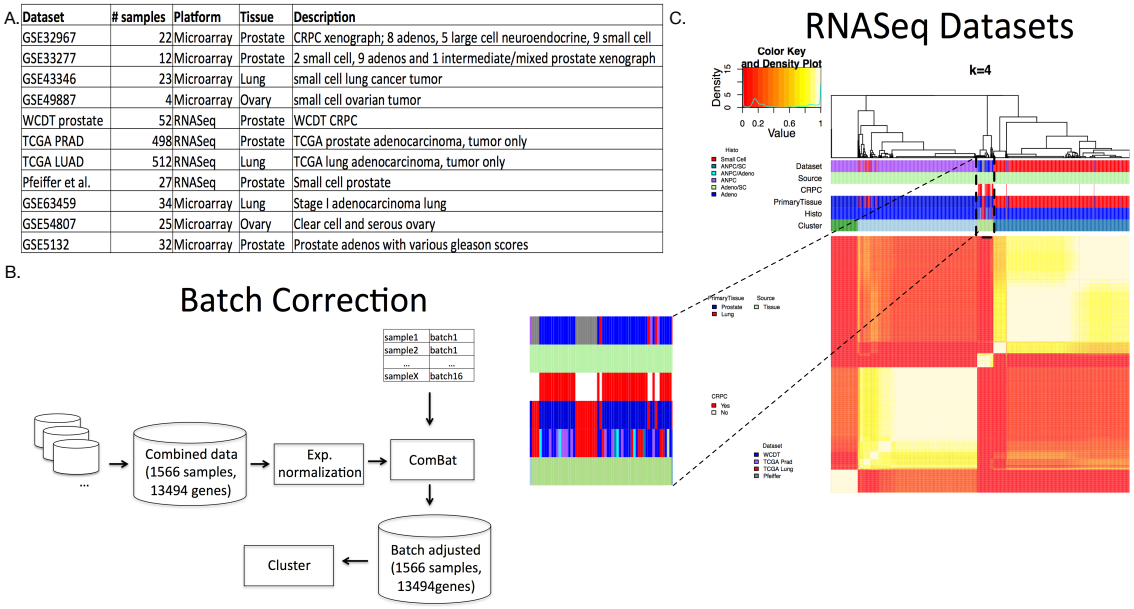


Figure 5.5: Description of pan-cancer small tumor work. A) List of datasets included into this study. B) Overview of the batch/dataset correction pipeline applied to the combined data. C) Consesus k-means clustering of the samples after the batch correction was applied. Small cell tumor appear to mostly cluster together.

5.4.1 Future Directions

This initial investigation showed that it is possible to perform pan-cancer multi-study analysis to better understand the behavior, molecular drivers and mechanisms of small cell tumors across various cancers. The preliminary results are not enough to make any assertions or inferences about small cell phenotype though. More analysis, possibly with additional data, is needed. We could also investigate at a family level by looking at all small-blue-round-cell tumors rather than just small cell.

5.5 Clinical Mutation Tests Are Predictive Of Candidacy For Immunotherapy

Immunotherapy is cancer therapeutic method that gain much popularity in the recent years. It has been previously shown that good candidates of immunotherapy often exhibit a hypermutated phenotype. However, extensive genomic testing is not always available to patients in clinic and clinicians are limited by the genomic tests available in the lab. One of such tests is FoundationOne gene panel that tests for mutations in 315 genes. We analyzed a cross-cancer cross-tissue cohort of 26 cancers and found that FoundationOne gene panel is representative of the mutation rate within these cancers. Furthermore, we analyzed Mismatch Repair pathway within this cohort and found that tumors with a disruption in this pathway have higher mutation rate than tumors that do not. We explored predictive power of the FoundationOne gene panel to predict mutations in this pathway and found that FoundationOne mutation profiles have high prediction accuracy of mutations in the Mismatch Repair pathway, even after we exclude those samples that have mutations in the Mismatch Repair pathway genes present in the FoundationOne gene panel. While we did not achieve success predicting MMR pathway aberrations from just mutation frequency alone, more work is required here.

5.5.1 Introduction

Good candidates of immunotherapy often have a hypermutated phenotype. In fact, it was shown that mutation frequency can be a good predictor of response to checkpoint inhibitors [33]. Often hyper mutated phenotype is associated with disruptions of the Mismatch Repair (MMR) molecular pathway. This pathway is responsible for repairing "mistakes" in the cell DNA. If disrupted, this pathway does not function properly and mutations tend to accumulate, over time leading to hypermutated phenotype. These patients are usually more responsive to immunotherapy. In fact, Colli *et al.* [33] show that 192 non-synonymous mutations is a threshold for ability to predict response to checkpoint inhibitors. In the clinical setting however, often extensive genomic testing is not available and oncologists are limited to standard panel tests. One of such tests is FoundationOne panel [37] that tests for mutations in 315 genes often mutated in cancer.

5.5.2 Results

5.5.2.1 TCGA in the context of FoundationOne gene panel

FoundationOne gene panel is a clinical test clinicians often order for their patients. This panel of genes consists of 315 genes often mutated in cancer and tests for a presence of absence of mutations within those genes. These genes represent many of the cancer hallmarks, such as E2F targets, G2M checkpoint, apoptosis and many signaling pathways generally found disrupted in cancer (Figure 5.6A & B).

We analyzed 26 of the TCGA cancers (6001 samples) in the context of the FoundationOne 315 genes. Out of 315 genes 313 were found in the TCGA mutation data. Mutation rates vary across cancers. They also vary across cancers within just the 313 FoundationOne genes (Figure 5.6C) and are, in fact, comparable when considering the entire genome (Figure 5.6D). While some cancers have clear outliers not representative of other cancers originating in the same tissue, most hypermutated cancers are uterine corpus endometrioid carcinoma (UCEC), skin Cutaneous Melanoma (SKCM), stomach adenocarcinoma (STAD), bladder urothelial Carcinoma (BLCA), and lung adenocarcinoma (LUAD). It is also clear that the mutation rate represented by the FoundationOne gene panel is representative of the overall mutation rate when all genes in the genome are considered.

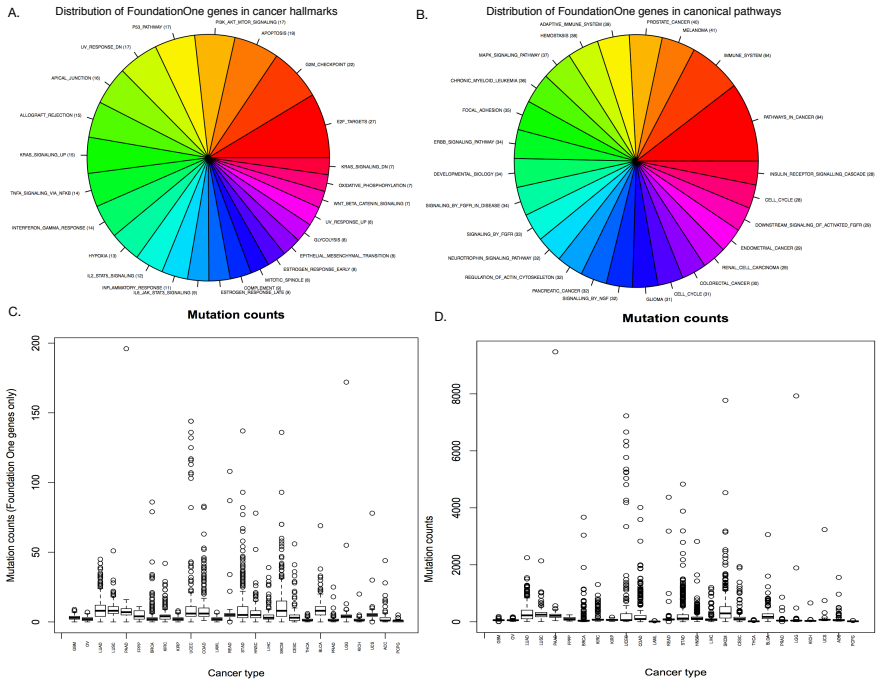


Figure 5.6: Exploration of the mutational landscape of the TCGA cohort in the context of FoundationOne gene set. A) Top 25 cancer hallmarks (MSigDB) represented within the FoundationOne gene panel. B) Top canonical pathways (MSigDB) represented within the FoundationOne gene panel. C) Mutation counts across TCGA cancers within the context of FoundationOne genes only. D) Mutation counts across TCGA cancers (entire genome).

5.5.2.2 MMR Pathway in the context of TCGA and FoundationOne gene panel

Mismatch Repair (MMR) pathway is responsible for repairing DNA replication mistakes and is often disrupted in cancer cells as a way to produce hypermutated phenotype, by which many tumors are characterized. We set out to explore MMR pathway and mutations within that pathway in the TCGA data.

There are 23 genes in the MMR pathway. We examined each tumor that

had mutation in at least one gene in those 23 genes. We discovered that 970 TCGA tumors have at least one mutation within MMR pathway (Figure 5.7A). This finding eludes to the fact that a single mutation within this pathway is sufficient to disrupt its function. The top mutated gene within this pathway is MSH6 (Figure 5.7B), whose product heterodimerizes with the product of MSH2 gene and initiates DNA repair.

Of the 23 MMR genes 5 are present in the FoundationOne panel (Figure 5.7B(i)). Within the top 4 mutated of those genes exhibit high mutual exclusivity (Figure 5.7C). This, again, supports our previous assertion that a single mutation in this pathway is sufficient to disrupt the DNA repair function.

We considered the overall mutation frequency in samples that have at least one MMR mutation and compared it to the mutation frequency of samples that do not have MMR mutations. We found that overall mutation frequency in samples that have an MMR disruption is significantly higher than in samples that do not (Figure 5.7D & E). This finding supports the hypothesis that once MMR function is impaired cancer cells go into uncontrolled mutation frenzy and acquire hypermutated phenotype. If we compare mutation frequencies in MMR mutants vs. not when considering all genes in the genome to the same groupings when only considering FoundationOne gene panel, we see similar distributions and similar p-values between the groups. This suggests that FoundationOne gene panel might be sufficient in recapturing hypermutated phenotype in the absence of the whole genome mutation data.

We compared expression profiles of the MMR mutants group, MMR WT group, as well as between the two groups (Figure 5.7F). We found that expression profiles of

samples within the MMR mutants group are statistically significantly more similar than expression profiles within the non mutants group (Welch t-test p-value = 0) as well as expression profiles between MMR mutants and MMR WT groups (Welch t-test p-value = 0). This suggests that hypermutated tumors are not only similar in mutation profiles but also in expression profiles and expression profiles can carry potential predictive power for hypermutated phenotype.

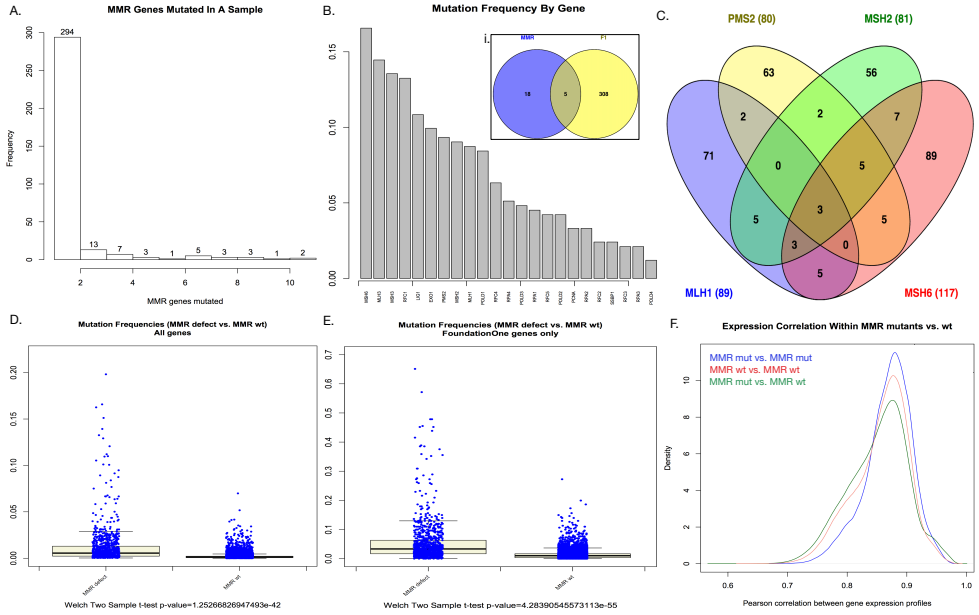


Figure 5.7: Exploration of MMR pathway and its mutations with TCGA cohort. A) Mutation counts per sample in MMR pathway genes. Most samples have a single mutation within MMR pathway showing that a single mutation is sufficient to disrupt this pathway. B) Mutation frequencies by each MMR gene within TCGA cohort. Different genes exhibit different mutation frequency suggesting difference in importance to MMR functionality. C) Top 4 out of 5 MMR genes in FoundationOne panel exhibit high mutual exclusivity suggesting that a mutation in a single gene is sufficient to disrupt the pathway function. D) Mutation frequencies in the TCGA cohort broken down by MMR mutants vs. not when using all genes in the genome. If there was at least a single mutation in the MMR pathway the sample was assigned to MMR mutants group. Welch t-test p-value was computed between the two groups. E) Mutation frequencies in the TCGA cohort broken down by MMR mutants vs. not when using FoundationOne genes only. If there was at least a single mutation in the MMR pathway the sample was assigned to MMR mutants group. Welch t-test p-value was computed between the two groups. F) Expression profiles of MMR mutants and MMR wt were compared (Pearson Rho). Expression profiles of the samples that have MMR disruptions are more similar than the expression profiles of the samples without MMR disruptions as well as expression profiles between the two groups.

5.5.2.3 Utilizing FoundationOne gene panel to predict hypermutated phenotype

We set out to explore how well the mutation profiles of the FoundationOne gene panel can predict MMR defects. As was shown above, MMR defects are indicative of the hypermutated phenotype. Therefore, being able to predict MMR defects would allow to make an assertion about whether the tumor exhibits overall hypermutated state. We explored FoundationOne gene panel feature space to extract markers predictive of the MMR mutant phenotype.

We used FoundationOne gene panel as the feature space and built a Random Forest model based on the TCGA mutation profiles consisting of 313 genes (one bit per gene). We evaluated the model by performing 5-fold cross validation. We found that this model produced average AUC = 0.85235, indicating that mutation profiles are highly predictive of MMR disruptions. We considered mutation frequencies only to train the model. This model produced average AUC = 0.654, indicating that using just the mutation frequency does not carry the same predictive power as using mutation profiles. We then built a model that used both the 313 gene mutation profiles and mutation frequency. We found that this model did not perform better than the model in which only the 313 gene mutation profiles were used. These two models have the same average AUC.

As was shown above, the 5 MMR genes present in the FoundationOne gene panel are the most informative markers of the prediction accuracy of MMR disruption

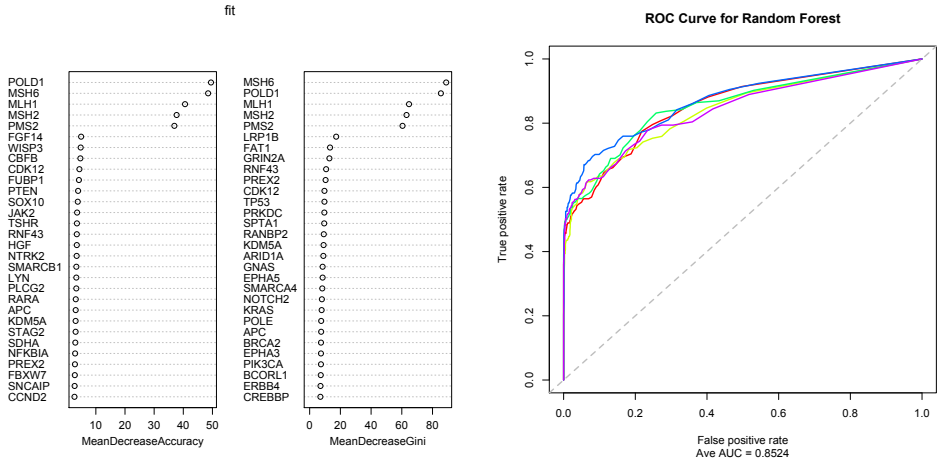
from the FoundationOne mutation profiles (Figures 5.8, 5.10). We wanted to explore what other predictive markers can be used in the absence of a mutation in one of those 5 MMR genes. There are 509 samples in the TCGA cohort that have a mutation in MMR pathway but not in one of the 5 FoundationOne MMR genes. We excluded the 461 samples that have at least one mutation in one of the 5 genes from our next analysis. We repeated the 3 models presented above, while excluding the 461 samples. We found that mutation frequency did not increases average AUC of the model and by itself is not predictive of the MMR disruptions (Figures 5.11-5.13). Some of the predictive markers of the MMR disruptions, in the absence of the 5 FoundationOne MMR marker mutations, are mutations in some of the main cancer players like FUBP1, IDH1, BRAF, AKT, and many others. In fact, out of the top 30 predictive markers, 9 are members of the KEGG_PATHWAYS_IN_CANCER MsigDB [78] gene set (FDR q-value = 3.51e-10). Furthermore, some of these genes are member of pathways involved in specific tissue cancers. For example, 5 of the genes are members of the KEGG_ENDOMETRIAL_CANCER pathway (FDR q-value = 6.5e-8) and 5 are members of the KEGG_MELANOMA pathway (FDR q-value = 2.42e-7). Other genes belong to generic pathways responsible for tumorigenesis and proliferation (e.g. REACTOME_SIGNALLING_BY_NGF with FDR q-value = 9.5e-7). We hypothesize that once the MMR pathway is disrupted genes involved in cancer hallmark pathways are mutated first, activating oncogenes and inactivating tumor suppressors. Table 5.1 summarizes the results of all the above described experiments.

5.6 Discussion

Immunotherapy is cancer therapeutic method that gain much popularity in the recent years. It allows utilizing the patient's immune system to target tumor cells based on specific markers presented on the cell surface of those cells. However, not every patient is a good candidate for immunotherapy and is important to be able to screen, in the clinical settings, for how good of a candidate a particular patient is for immunotherapy. Extensive genomic sequencing assays are not always available within the clinic settings and clinicians are often limited to tests available in the standard clinical lab. One of such tests is FoundationOne gene panel mutation test, which tests for mutations in 315 genes often mutated in cancer. Additionally, it has been previously suggested that hypermutated phenotype correlates with good response to immunotherapy. It has also been previously suggested that hypermutated phenotype is associated with mutations and disruptions in MMR pathway.

We set out to explore, within the TCGA cohort (26 cancers, 6001 samples), how predictive mutations in the MMR pathway are of hypermutated phenotype and how representative FoundationOne gene panel is of that predictive power. It turns out that mutation frequency of the FoundationOne gene panel alone is not predictive of the MMR mutants. It also does not add to predictive power of the model based on both the mutation profiles and mutation frequency. We found that even when we exclude the tumors with mutations in at least one of the 5 MMR genes present in the FoundationOne gene panel we observe good predictability from the mutation profiles and

mutation frequency does not add to prediction power of the model. This suggests that it is not the mutation rate that is indicative of the MMR disruptions but the composition of the mutations in the mutation profile. We found that even after excluding those samples the most predictive markers are simply genes associated with cancer pathways. We hypothesize that MMR disruption gives a green light to unchecked mutations in cancer drivers, activating oncogenes and inactivating tumor suppressors. Another thing to consider is that we performed our experiments on a cross-cancer cross-tissue TCGA cohort, which possibly dampens the mutation marker signals specific to tissues. We suspect that if we were to build predictive models per each tissue we could see much stronger predictive markers of the MMR disruption when no MMR gene mutations are detected by the FoundationOne test.



(a) Importance of various features in the model performance when using mutation profiles.

(b) ROC for each of 5 folds of 5-fold cross validation Random Forest model when using 313-gene TCGA mutation profiles (one bit per gene). Each color indicates each fold. Average AUC is displayed on the bottom of the plot.

Figure 5.8: Random Forest model using 313 genes mutation profiles only.

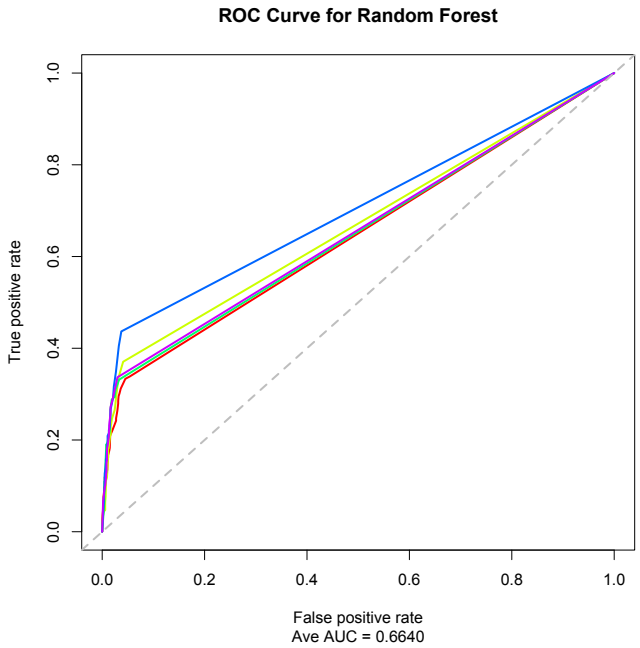
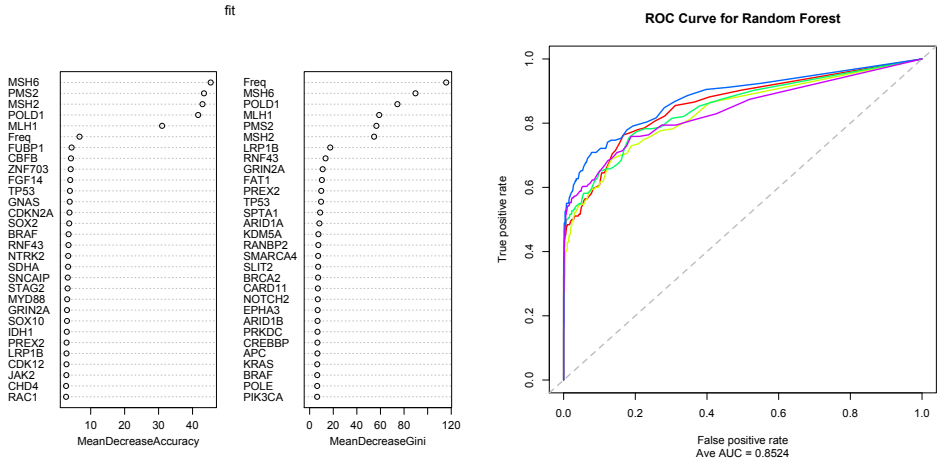


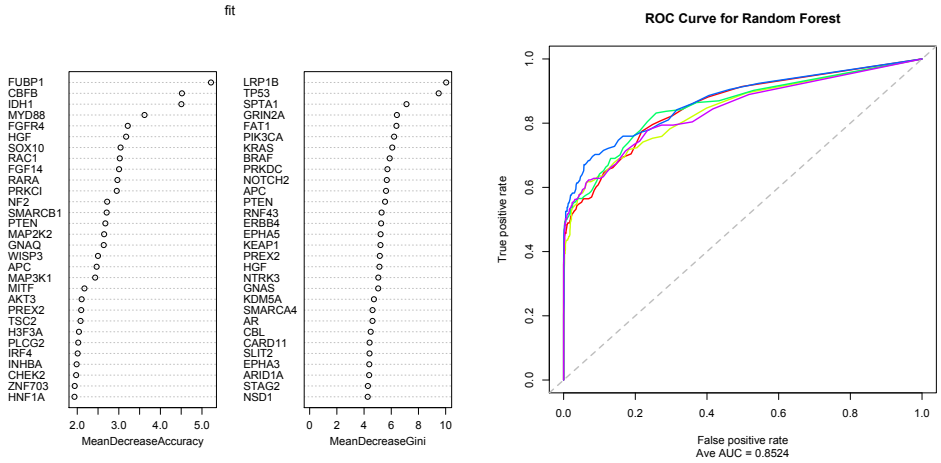
Figure 5.9: Random Forest model using mutation frequency only. ROC for each of 5 folds of 5-fold cross validation Random Forest model when using mutation frequency as the only input feature. Each color indicates each fold. Average AUC is displayed on the bottom of the plot.



(a) Importance of various features in the model performance when using both mutation profiles and mutation frequencies.

(b) ROC for each of 5 folds of 5-fold cross validation Random Forest model when using both 313-gene TCGA mutation profiles (one bit per gene) and mutation frequencies. Each color indicates each fold. Average AUC is displayed on the bottom of the plot.

Figure 5.10: Random Forest model using both 313 genes mutation profiles and mutation frequencies.



(a) Importance of various features in the model performance when using mutation profiles.

(b) ROC for each of 5 folds of 5-fold cross validation Random Forest model when using 313-gene TCGA mutation profiles (one bit per gene). Each color indicates each fold. Average AUC is displayed on the bottom of the plot.

Figure 5.11: Random Forest model using 313 genes mutation profiles only. 461 samples with a mutation in one of the 5 MMR FoundationOne genes were excluded.

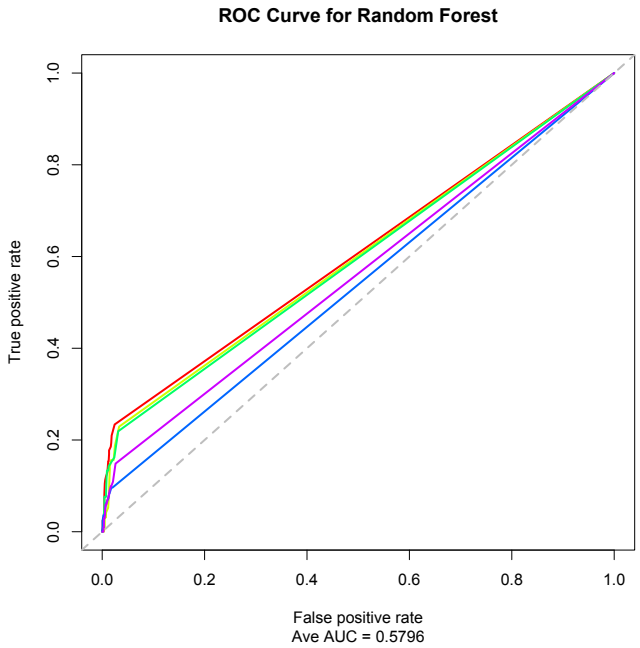
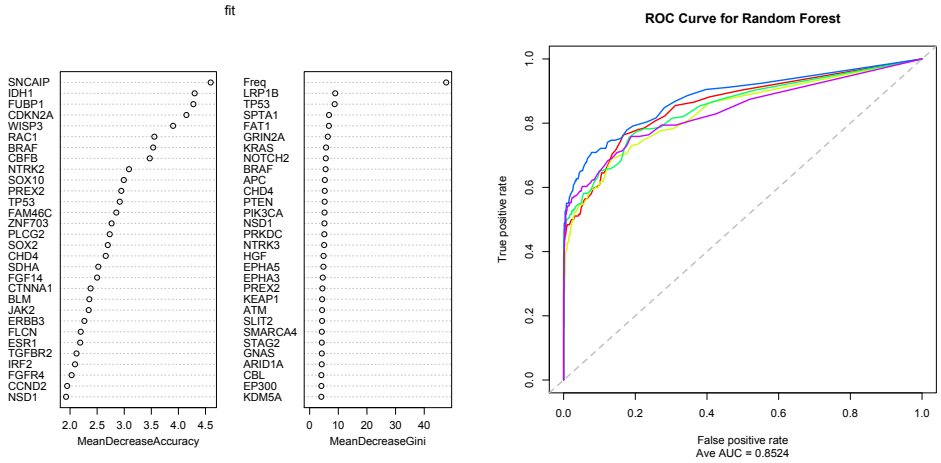


Figure 5.12: Random Forest model using mutation frequency only. 461 samples with a mutation in one of the 5 MMR FoundationOne genes were excluded. ROC for each of 5 folds of 5-fold cross validation Random Forest model when using mutation frequency as the only input feature. Each color indicates each fold. Average AUC is displayed on the bottom of the plot.



(a) Importance of various features in the model performance when using both mutation profiles and mutation frequencies.

(b) ROC for each of 5 folds of 5-fold cross validation Random Forest model when using both 313-gene TCGA mutation profiles (one bit per gene) and mutation frequencies. Each color indicates each fold. Average AUC is displayed on the bottom of the plot.

Figure 5.13: Random Forest model using both 313 genes mutation profiles and mutation frequencies. 461 samples with a mutation in one of the 5 MMR FoundationOne genes were excluded.

5.6.0.4 Future Directions

While initial investigation performed here showed the need to investigate further, time constraints and other priorities prevented me from finishing this project. We must investigate additional predictors to use with mutation frequencies as Random Forest models are not the most appropriate here. We also need to further investigate associations of the aberrations in the MMR pathway and mutational load. We should look at additional datasets as well. Can we make a tighter connection between MMR aberra-

tions and response to the checkpoint inhibitors? We showed that mutation frequencies across cancer within the FoundationOne panel genes is representable of the mutations frequencies across cancers in the entire genome. This is the first step in demonstrating that gene panel tests can be utilized for predicting genome-wise mutation patterns. However, the project warrants more investigation and analysis.

5.7 Chapter Conclusion

In this chapter I describe some additional work I completed during the course of my doctoral research. Some of this was has been published in peer reviewed journals and some just presented internally and is not ready for a publication. In those cases I provide some guidelines for possible avenues of future work that can lead to the project completion.

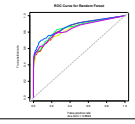
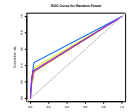
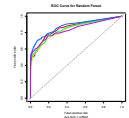
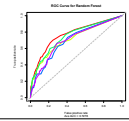
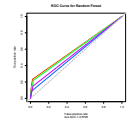
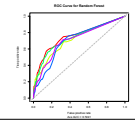
	Experiment	Ave AUC	N features	N samples	5-fold CV ROC
Including F1 MMR marker mutants	Using mutation profiles	**0.852**	313	6001	
	Using mutation frequency	0.654	1	6001	
	Using mutation profiles and frequency	**0.852**	314	6001	
Excluding F1 MMR marker mutants	Using mutation profiles	**0.7278**	313	5540	
	Using mutation frequency	0.5796	1	5540	
	Using mutation profiles and frequency	0.7241	314	5540	

Table 5.1: Results from MMR predictors in TCGA cohort

Bibliography

[1] California kids cancer comparison. http://www.ciapm.org/project/california-kids-cancer-comparison. Accessed: 2016-11-05.

[2] Cholangiocarcinoma Key Statistics cholangiocarcinoma key statistics. http://www.cancer.org/. Accessed: 2016-09-02.

[3] Docker. https://www.docker.com/. Accessed: 2016-11-05.

[4] Google maps JavaScript API V3 reference. https://developers.google.com/maps/documentation/javascript/reference. Accessed: 2016-7-24.

[5] Gtex. http://www.gtexportal.org/home/. Accessed: 2016-11-05.

[6] Keras. https://keras.io/. Accessed: 2016-11-05.

[7] Quay. https://quay.io/. Accessed: 2016-11-05.

[8] Tensorflow. https://www.tensorflow.org/. Accessed: 2016-11-05.

[9] Theano. http://deeplearning.net/software/theano/. Accessed: 2016-11-05.

[10] Therapeutically Applicable Research To Generate Effective Treatments: Data Matrix, howpublished = https://ocg.cancer.gov/programs/target/data-matrix, note = Accessed: 2016-11-04.

[11] Rehan Akbani Adrian Ally Samirkumar Amin Christopher D Andry Matti Annala Armen Aprikian Joshua Armenia Arshi Arora J Todd Auman Miruna Balasundaram Saianand Balu Christopher E Barbieri Thomas Bauer Christopher C Benz Alain Bergeron Rameen Beroukhim Mario Berrios Adrian Bivol Tom Bodenheimer Lori Boice Moiz S Bootwalla Rodolfo Borges dos Reis Paul C Boutros Jay Bowen Reanne Bowlby Jeffrey Boyd Robert K Bradley Anne Breggia Fadi Brimo Christopher A Bristow Denise Brooks Bradley M Broom Alan H Bryce-Glenn Bubley Eric Burks Yaron SN Butterfield Michael Button David Canes Carlos G Carlotti Rebecca Carlsen Michel Carmel Peter R Carroll Scott L Carter Richard Cartun Brett S Carver June M Chan Matthew T Chang Yu Chen Andrew D Cherniack Simone Chevalier Lynda Chin Juok Cho Andy Chu Eric Chuah Sudha Chudamani Kristian Cibulskis Giovanni Ciriello Amanda Clarke Matthew R Cooperberg Niall

M Corcoran Anthony J Costello Janet Cowan Daniel Crain Erin Curley Kerstin David John A Demchok Francesca Demichelis Noreen Dhalla Rajiv Dhir Alexandre Doueik Bettina Drake Heidi Dvinge Natalya Dyakova Ina Felau Martin L Ferguson Scott Frazer Stephen Freedland Yao Fu Stacey B Gabriel Jianjiong Gao Johanna Gardner Julie M Gastier-Foster Nils Gehlenborg Mark Gerken Mark B Gerstein Gad Getz Andrew K Godwin Anuradha Gopalan Markus Graefen Kiley Graim Thomas Gribbin Ranabir Guin Manaswi Gupta Angela Hadjipanayis Syed Haider Lucie Hamel D Neil Hayes David I Heiman Julian Hess Katherine A Hoadley Andrea H Holbrook Robert A Holt Antonia Holway Christopher M Hovens Alan P Hoyle Mei Huang Carolyn M Hutter Michael Ittmann Lisa Iype Stuart R Jefferys Corbin D Jones Steven JM Jones Hartmut Juhl Andre Kahles Christopher J Kane Katayoon Kasaian Michael Kerger Ekta Khurana Jaegil Kim Robert J Klein Raju Kucherlapati Louis Lacombe Marc Ladanyi Phillip H Lai Peter W Laird Eric S Lander Mathieu Latour Michael S Lawrence Kevin Lau Tucker LeBien Darlene Lee Semin Lee Kjong-Van Lehmann Kristen M Leraas Ignaty Leshchiner Robert Leung John A Libertino Tara M Lichtenberg Pei Lin W Marston Linehan Shiyun Ling Scott M Lippman Jia Liu Wenbin Liu Lucas Lochovsky Massimo Loda Christopher Logothetis Laxmi Lolla Adam Abeshouse, Jaeil Ahn. The molecular taxonomy of primary prostate cancer. *Cell*, 4(163):1011–1025, 2015.

[12] Purvesh Khatri Sonia S. Hassan Pooja Mittal Jung-sun Kim Chong Jai Kim Juan Pedro Kusanovic Adi Laurentiu Tarca, Sorin Draghici and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2008.

[13] Eric Brooks Alex Genshaft Shahin Shajahan Michael Ittman G. Steven Bova Jonathan Melamed Ilona Holcomb Robert J. Schneider Alexander Pearlman, Christopher Campbell and Harry Ostrer. Clustering-based method for developing a genomic copy number alteration signature for predicting the metastatic potential of prostate cancer. *J Probab Stat.*, page 873570, 2012.

[14] Shi L Shizhen Z Deng S Xie Z et al. An G, Xu Y. Chromosome 1q21 gains confer inferior outcomes in multiple myeloma treated with bortezomib but copy number variation and percentage of plasma cells involved have no additional prognostic value. *Haematologica*, 99:353359, 2014.

[15] Alvaro Aytes, Antonina Mitrofanova, Celine Lefebvre, Mariano J Alvarez, Mireia Castillo-Martin, Tian Zheng, James A Eastham, Anuradha Gopalan, Kenneth J Pienta, Michael M Shen, Andrea Califano, and Cory Abate-Shen. Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell*, 25(5):638–651, 12 May 2014.

[16] Yoshifumi Baba, Baba Yoshifumi, Nosho Katsuhiko, Shima Kaori, Irahara Natsumi, Andrew T Chan, Jeffrey A Meyerhardt, Daniel C Chung, Edward L Giovan-

nucci, Charles S Fuchs, and Ogino Shuji. HIF1A overexpression is associated with poor prognosis in a cohort of 731 colorectal cancers. *Am. J. Pathol.*, 176(5):2292–2301, 2010.

[17] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam a Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F Berger, John E Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa a Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H Engels, Jill Cheng, Guoying K Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P Mesirov, Stacey B Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E Myer, Barbara L Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L Harris, Matthew Meyerson, Todd R Golub, Michael P Morrissey, William R Sellers, Robert Schlegel, and Levi a Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–7, March 2012.

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, 1995.

[19] Hamid Bolouri, Lue Ping Zhao, and Eric C Holland. Big data visualization identifies the multidimensional molecular landscape of human gliomas. *Proc. Natl. Acad. Sci. U. S. A.*, 113(19):5394–5399, 10 May 2016.

[20] Mair P. Borg I, Groenen PJF. Applied multidimensional scaling. *Springer Science and Business Media*, 2012.

[21] Delbert Dueck Brendan J. Frey. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, 2007.

[22] Vladislav Uzunangelov Robert Baertsch Yulia Newton Kiley Graim Colleen Mathis Donghui Cheng Joshua M. Stuart Bryan A. Smith, Artem Sokolov and Owen N. Witte. A basal stem cell signature identifies aggressive prostate cancer phenotypes. *PNAS*, 112(47):E6544E6552, 2015.

[23] Chen S Coleman I Wang H Fang Z Chen S Nelson PS Liu XS Brown M Balk SP. Cai C, He HH. Androgen receptor gene expression in prostate cancer is directly suppressed by the androgen receptor through recruitment of lysine-specific demethylase. *Cancer Cell*, 4(20):457–71, 2011.

[24] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 4 October 2012.

[25] Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–690, 23 October 2014.

[26] Cancer Genome Atlas Research Network, Cyriac Kandoth, Nikolaus Schultz, Andrew D Cherniack, Rehan Akbani, Yuexin Liu, Hui Shen, A Gordon Robertson, Itai Pashtan, Ronglai Shen, Christopher C Benz, Christina Yau, Peter W Laird, Li Ding, Wei Zhang, Gordon B Mills, Raju Kucherlapati, Elaine R Mardis, and Douglas A Levine. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2 May 2013.

[27] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, October 2013.

[28] Collisson EA Mills GB Shaw KRM Ozenberger BA et al. Cancer Genome Atlas Research Network, Weinstein JN. The cancer genome atlas pan-cancer analysis project. *Nat Genet.*, 45:11131120, 2013.

[29] Malta TM Sabedot TS Salama SR Murray BA Morozova O Newton Y Radenbaugh A Pagnotta SM Anjum S Wang J Manyam G Zoppoli P Ling S Rao AA Grifford M Cherniack AD Zhang H Poisson L Carlotti CG Jr Tirapelli DP Rao A Mikkelsen T Lau CC Yung WK Rabadan R Huse J Brat DJ Lehman NL Barnholtz-Sloan JS Zheng S Hess K Rao G Meyerson M Beroukhim R Cooper L Akbani R Wrensch M Haussler D-Aldape KD Laird PW Gutmann DH; TCGA Research Network Noushmehr H Iavarone A Verhaak RG. Ceccarelli M, Barthel FP. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, 164(3):550–63, 2016.

[30] Uma R Chandran, Changqing Ma, Rajiv Dhir, Michelle Bisceglia, Maureen Lyons-Weiler, Wenjing Liang, George Michalopoulos, Michael Becich, and Federico A Monzon. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC cancer*, 7(1):64, 2007.

[31] Maggie C U Cheang, Stephen K Chia, David Voduc, Dongxia Gao, Samuel Leung, Jacqueline Snider, Mark Watson, Sherri Davies, Philip S Bernard, Joel S Parker, Charles M Perou, Matthew J Ellis, and Torsten O Nielsen. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J. Natl. Cancer Inst.*, 101(10):736–750, 20 May 2009.

[32] Geo Pertea Ali Mortazavi Gordon Kwan Marijke J van Baren Steven L Salzberg Barbara J Wold & Lior Pachter Cole Trapnell, Brian A Williams. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511515, 2010.

[33] Myers TA Jessop L Yu K Chanock SJ Colli LM, Machiela MJ. Burden of Nonsynonymous Mutations among TCGA Cancers and Candidate Immune Checkpoint Inhibitor Responses. *Cancer Res.*, 76(13):3767–72, 2016.

[34] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:5665, 2012.

[35] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.*, 45(6):580585, 2013.

[36] Nello Cristianini, Jaz Kandola, Andre Elisseeff, and John Shawe-Taylor. On kernel target alignment. 194:205–256, 2006.

[37] M. Cronin and JS. Ross. Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology. *Biomark Med.*, 5(3):293–305, 2011.

[38] Igor Puzanov M.D. Vivek Subbiah M.D. Jason E. Faris M.D. Ian Chau M.D. Jean-Yves Blay M.D. Ph.D. Jrgen Wolf M.D. Ph.D. Noopur S. Raje M.D. Eli L. Diamond M.D. Antoine Hollebecque M.D. Radj Gervais M.D. Maria Elena Elez-Fernandez M.D. Antoine Italiano M.D. Ph.D. Ralf-Dieter Hofheinz M.D. Manuel Hidalgo M.D. Ph.D. Emily Chan M.D. Ph.D. Martin Schuler M.D. Susan Frances Lasserre M.Sc. Martina Makrutzki M.D. Florin Sirzen M.D. Ph.D. Maria Luisa Veronese M.D. Josep Tabernero M.D.-Ph.D. David M. Hyman, M.D. and Ph.D. Jos Baselga, M.D. Vemurafenib in multiple nonmelanoma cancers with braf v600 mutations. *N Engl J Med*, 373:726–736, 2015.

[39] Andreas von DeimlingDominique Figarella-BrangerWebster K. CaveneeHiroko OhgakiOtmar D. WiestlerPaul KleihuesDavid W. Ellison David N. Louis, Arie PerryGuido Reifenberger. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica*, 131(6):803820, 2016.

[40] Zhang Y Coxon A Burgess TL-Wagner AJ Fisher DE Davis IJ, McFadden AW. Identification of the receptor tyrosine kinase c-Met and its ligand, hepatocyte growth factor, as therapeutic targets in clear cell sarcoma. *Cancer Res.*, 70(2):639–45, 2010.

[41] Masao Deguchi, Hiroaki Shiina, Mikio Igawa, Masanori Kaneuchi, Koichi Nakajima, and Rajvir Dahiya. DNA mismatch repair genes in renal cell carcinoma. *J. Urol.*, 169(6):2365–2371, June 2003.

[42] Schlesinger F Drenkow J Zaleski C-Jha S Batut P Chaisson M Gingeras TR Dobin A, Davis CA. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[43] Robert A. Weinberg Douglas Hanahan. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646674, 2011.

[44] Denis Dupuy, Dupuy Denis, Bertin Nicolas, César A Hidalgo, Venkatesan Kavitha, Tu Domena, Lee David, Rosenberg Jennifer, Svrzikapa Nenad, Blanc Aurélie, Carnec Alain, Carvunis Anne-Ruxandra, Pulak Rock, Shingles Jane, Reece-Hoyes John, Hunt-Newbury Rebecca, Viveiros Ryan, William A Mohler, Tasan Murat, Frederick P Roth, Christian Le Peuch, Ian A Hope, Johnsen Robert, Donald G Moerman, Barabási Albert-László, Baillie David, and Vidal Marc. Genome-scale analysis of in vivo spatiotemporal promoter activity in caenorhabditis elegans. *Nat. Biotechnol.*, 25(6):663–668, 2007.

[45] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.*, 95(25):14863–14868, 8 December 1998.

[46] Jue Lin Firdaus S. Dhabhar Nancy E. Adler Jason D. Morrow Elissa S. Epel, Elizabeth H. Blackburn and Richard M. Cawthon. Accelerated telomere shortening in response to life stress. *PNAS*, 101(49):1731217315, 2004.

[47] Nicholas Erho, Anamaria Crisan, Ismael a Vergara, Anirban P Mitra, Mercedeh Ghadessi, Christine Buerki, Eric J Bergstralh, Thomas Kollmeyer, Stephanie Fink, Zaid Haddad, Benedikt Zimmermann, Thomas Sierocinski, Karla V Ballman, Timothy J Triche, Peter C Black, R Jeffrey Karnes, George Klee, Elai Davicioni, and Robert B Jenkins. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PloS one*, 8(6):e66855, January 2013.

[48] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, 5(1):e8, January 2007.

[49] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, January 2007.

[50] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[51] Csardi Gabor and Nepusz Tamas. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695, 2006.

[52] Andrew J. Stephenson Robert M. Hoffman Gennadi V. Glinsky, Anna B. Glinskii and William L. Gerald. Gene expression profiling predicts clinical outcome of prostate cancer. *Genomics*, 113(6):913923, 2004.

[53] Bulent Arman Aksoy Yasin Senbabaoglu Nikolaus Schultz & Chris Sander Giovanni Ciriello, Martin L Miller. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45:11271133, 2013.

[54] J R Gnarra, K Tony, and Y Weng. Mutations of the VHL tumour suppressor gene in renal carcinoma (1994). *Nat. Genet.*

[55] Catherine S Grasso, Yi-Mi Wu, Dan R Robinson, Xuhong Cao, Saravana M Dhanasekaran, Amjad P Khan, Michael J Quist, Xiaojun Jing, Robert J Lonigro, J Chad Brenner, et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, 487(7406):239–243, 2012.

[56] Anuradha Gopalan Brett S. Carver Matthew T. Chang Yonghong Xiao Adriana Heguy Kety Huberman Melanie Bernstein Melissa Assel Rajmohan Murali Andrew Vickers Peter T. Scardino Chris Sander Victor Reuter Barry S. Taylor Haley Hieronymus, Nikolaus Schultz and Charles L. Sawyers. Copy number alteration burden predicts prostate cancer relapse. *Proc Natl Acad Sci U S A*, 30(111):1113911144, 2014.

[57] Glade Bender JL Kim A Crompton BD Parker E Dumont IP Hong AL Guo D Church A Stegmaier K Roberts CW Shusterman S London WB MacConaill LE Lindeman NI Diller L Rodriguez-Galindo C Janeway KA Harris MH, DuBois SG. Multicenter Feasibility Study of Tumor Molecular Profiling to Inform Therapeutic Decisions in Advanced Pediatric Solid Tumors: The Individualized Cancer Therapy (iCat) Study. *JAMA Oncol*, 2016.

[58] Merrett SL Lee HR Rudd MK Willard HF Hayden KE, Strome ED. Sequences associated with centromere competency in the human genome. *Mol Cell Biol*, 33(4), pages = "763-772", year = 2013).

[59] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max D M Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A Margolin, Laura J Van't Veer, Nuria Lopez-Bigas, Peter W Laird, Benjamin J Raphael, Li Ding, A Gordon Robertson, Lauren A Byers, Gordon B Mills, John N Weinstein, Carter Van Waes, Zhong Chen, Eric A Collisson, Cancer Genome Atlas Research Network, Christopher C Benz, Charles M Perou, and Joshua M Stuart. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 14 August 2014.

[60] Peggy Hsieh and Kazuhiko Yamane. DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mech. Ageing Dev.*, 129(7-8):391–407, July 2008.

[61] Peter A. Humphrey. Gleason grading and prognostic factors in carcinoma of the prostate. *Modern Pathology*, 3(17):292306, 2004.

[62] Per Hydbring, Marcos Malumbres, and Piotr Sicinski. Non-canonical functions of cell cycle cyclins and cyclin-dependent kinases. *Nat. Rev. Mol. Cell Biol.*, 17(5):280–292, May 2016.

[63] Aapo Hyvarinen and Erkki Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4-5):411–430, 2000.

[64] Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11:3750, 1912.

[65] Samuel Aparicio Stephen Chia Carolyn Ch'ng Rebecca Deyell Peter Eirew Alexandra Fok Karen Gelmon Cheryl Ho David Huntsman Martin Jones Katayoon Kasaian Aly Karsan Sreeja Leelakumari Yvonne Li Howard Lim Yussanne Ma-Colin Mar Monty Martin Richard Moore Andrew Mungall Karen Mungall Erin Pleasance S. Rod Rassekh Daniel Renouf Yaoqing Shen Jacqueline Schein Kasmintan Schrader Sophie Sun Anna Tinker Eric Zhao Stephen Yip Janessa Laskin, Steven Jones and Marco A. Marra. Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. *Cold Spring Harb Mol Case Stud.*, 1:a000570, 2015.

[66] Douglas A. Harrison Jason S. Rawlings, Kristin M. Rosler. The JAK/STAT signaling pathway. *Journal of Cell Science*, 117:1281–1283, 2004.

[67] Austin Nothaft Christopher Ketchum Joel Armstrong Adam Novak Jacob Pfeil Jake Narkizian Alden D. Deran Audrey Musselman-Brown Hannes Schmidt Peter Amstutz Brian Craft Mary Goldman Kate Rosenbloom Melissa Cline Brian O'Connor-Megan Hanna Chet Birger W. James Kent David A. Patterson Anthony D. Joseph Jingchun Zhu Sasha Zaranek Gad Getz David Haussler Benedict Paten John Vivian, Arjun Rao. Rapid and efficient analysis of 20,000 RNA-seq samples with Toil. *Cold Spring Harbor Laboratory*.

[68] Rabinovic A Johnson, WE and C Li. Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[69] Nicholas A. Graham John K. Lee Bryan A. Smith 1 Bjoern Titz Tanya Stoyanova Claire M. Faltermeier Vladislav Uzunangelov Daniel E. Carlin Daniel Teo Fleming Christopher K. Wong Yulia Newton-Sud Sudha Ajay A. Vashisht Jiaoti Huang James A. Wohlschlegel Thomas G. Graeber Owen N. Witte Justin M. Drake, Evan O. Paull and Joshua M. Stuart. Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer. *Cell*, 166:10411054, 2016.

[70] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, Mark D M Leiserson, Christopher A Miller, John S Welch, Matthew J Walter, Michael C Wendl, Timothy J Ley, Richard K Wilson, Benjamin J Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 17 October 2013.

[71] Miten Jain Nicolas Altemose Huntington F. Willard Karen H. Miga, Yulia Newton and W. James Kent. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome research*, 24(4):697–707, 2014.

[72] Ekta Khurana, Yao Fu, Jieming Chen, and Mark Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput. Biol.*, 9(3):e1002886, 7 March 2013.

[73] S K Kim. A gene expression map for caenorhabditis elegans. *Science*, 293(5537):2087–2092, 2001.

[74] S Knuutila, A M Björkqvist, K Autio, M Tarkkanen, M Wolf, O Monni, J Szymanska, M L Larramendy, J Tapper, H Pere, W El-Rifai, S Hemmer, V M Wasenius, V Vidgren, and Y Zhu. DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. *Am. J. Pathol.*, 152(5):1107–1123, May 1998.

[75] Nguyen H-A Cohen D Viara E Grieco L et al. Kuperstein I, Bonnet E. Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. *Oncogenesis*, 4:e160, 2015.

[76] Alexander Lachmann, Huilei Xu, Jayanth Krishnan, Seth I Berger, Amin R Mazloom, and Avi Ma'ayan. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, 26(19):2438–2444, 1 October 2010.

[77] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, 2011.

[78] Pinchback R-Thorvaldsdttir H Tamayo P Mesirov JP. Liberzon A, Subramanian A. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27:17391740, 2011.

[79] Khan S-Vihinen M Kowalski J Yu G Chen L Ewing CM Eisenberger MA Carducci MA Nelson WG Yegnasubramanian S Luo J Wang Y Xu J Isaacs WB Visakorpi T Bova GS. Liu W, Laitinen S. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med.*, 7(15):819, 2009.

[80] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al.

The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[81] Ulrike Von Luxburg. A tutorial on spectral clustering, 2007.

[82] Ben D MacArthur, Alexander Lachmann, Ihor R Lemischka, and Avi Ma'ayan. GATE: software for the analysis and visualization of high-dimensional time series expression data. *Bioinformatics*, 26(1):143–144, 1 January 2010.

[83] Federico M Giorgi Alexander Lachmann B Belinda Ding B Hilda Ye & Andrea Califano Mariano J Alvarez, Yao Shen. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, 48:838847, 2016.

[84] Amir Arsanjani Lixin Zhou Eliza Wickham Garcia Joshua Modder Monica Kostelec David Barker Tracy Downs Jian-Bing Fan Jessica Wang-Rodriguez Marina Bibikova, Eugene Chudin. Expression signatures that correlated with gleason score and relapse in prostate cancer. *Genomics*, 89(6):666672, 2007.

[85] Shawn Martin, Martin Shawn, W Michael Brown, Klavans Richard, and Kevin W Boyack. OpenOrd: an open-source toolbox for large graph layout. In *Visualization and Data Analysis 2011*, 2011.

[86] Teresa Swatloski Melissa Cline Olena Morozova Mark Diekhans David Haussler Mary Goldman, Brian Craft and Jingchun Zhu. The UCSC Cancer Genomics Browser: update 2015. *Nucleic Acids Res.*, 28:43, 2015.

[87] Di Wu Yifang Hu Charity W. Law Wei Shi Matthew E. Ritchie, Belinda Phipson and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 2015.

[88] Leo J. Lee Michael K. K. Leung, Hui Yuan Xiong and Brendan J. Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30:i121i129, 2014.

[89] Marina Bibikova Brandy Klotzle Jian-Bing Fan Shanshan Zhao Ziding Feng Elaine A. Ostrander-Daniel W. Lin Peter S. Nelson Milan S. Geybels, Jonathan L. Wright and Janet L. Stanford. Epigenetic signature of gleason score and prostate cancer recurrence after radical prostatectomy. *Clin Epigenetics*, 8(97), 2016.

[90] Everett J Parsons DW Chinnaiyan AM Mody RJ, Prensner JR. Precision medicine in pediatric oncology: Lessons learned and next steps. *Pediatr Blood Cancer*, 2016.

[91] Lonigro RJ Cao X Roychowdhury S Vats P Frank KM Prensner JR Asangani I Palanisamy N-Dillman JR Rabah RM Kunju LP Everett J Raymond VM Ning Y Su F Wang R-Stoffel EM Innis JW Roberts JS Robertson PL Yanik G Chamdin A Connelly JA Choi S Harris AC Kitko C Rao RJ Levine JE Castle VP Hutchinson

RJ Talpaz M Robinson DR Chinnaiyan AM Mody RJ, Wu YM. Integrative Clinical Sequencing in the Management of Refractory or Relapsed Cancer in Youth. *JAMA Oncol*, 314(9):913–25, 2015.

[92] Aristidis Moustakas. Smad signalling network. *Journal of Cell Science*, 115:3355–3356, 2002.

[93] The Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med*, 372:2481–2498, 2015.

[94] The Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, 163(4):10111025, 2015.

[95] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

[96] Cydney B Nielsen, Michael Cantor, Inna Dubchak, David Gordon, and Ting Wang. Visualizing genomes: techniques and challenges. *Nat. Methods*, 7(3 Suppl):S5–S15, March 2010.

[97] Blackburn EH Newman AB Wu SH Li R Simonsick EM Harris TM Cummings SR Cawthon RM; Health ABC study Njajou OT, Hsueh WC. Association between telomere length, specific causes of death, and years of healthy life in health, aging, and body composition, a population-based cohort study. *J Gerontol A Biol Sci Med Sci.*, 64(8):860–4, 2009.

[98] Mete M Herbolsheimer P Smith KL Bijelic L Boisvert ME Swain SM Nunes RA, Wray L. Genomic profiling of breast cancer in African-American women using MammaPrint. *Breast Cancer Res Treat.*, 159(3):481–8, 2016.

[99] Cheang MC Leung S Voduc D Vickery T Davies S Fauron C He X Hu Z Quackenbush JF Stijleman IJ Palazzo J Marron JS Nobel AB Mardis E Nielsen TO Ellis MJ-Perou CM Bernard PS Parker JS, Mullins M. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.*, 27(8):1160–7, 2009.

[100] Yang Y Wang T Scollon S Bergstrom K Kerstein RA Gutierrez S Petersen AK Bavle A Lin FY Lpez-Terrada DH Monzon FA Hicks MJ Eldin KW Quintanilla NM Adesina AM Mohila CA Whitehead W Jea A Vasudevan SA Nuchtern JG Ramamurthy U McGuire AL Hilsenbeck SG Reid JG Muzny DM Wheeler DA Berg SL Chintagumpala MM Eng CM Gibbs RA Plon SE Parsons DW, Roy A. Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors. *JAMA Oncol*, 2016.

[101] R.L. Graham Persi Diaconis. Spearman's Footrule as a Measure of Disarray, 1977.

[102] Alexander Platzer. Visualization of SNPs with t-SNE. *PLoS One*, 8(2):e56883, 15 February 2013.

[103] Thomas P. Plesec. Gastrointestinal Mesenchymal Neoplasms other than Gastrointestinal Stromal Tumors: Focusing on Their Molecular Aspects. *Patholog Res Int.*, 2016.

[104] Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics*, 30(19):2840–2842, October 2014.

[105] Aleix Prat, Barbara Adamo, Cheng Fan, Vicente Peg, Maria Vidal, Patricia Galván, Ana Vivancos, Paolo Nuciforo, Héctor G Palmer, Shaheenah Dawood, Jordi Rodón, Santiago Ramon y Cajal, Santiago Ramony Cajal, Josep Maria Del Campo, Enriqueta Felip, Josep Tabernero, and Javier Cortés. Genomic analyses across six cancer types identify basal-like breast cancer as a unique molecular entity. *Sci. Rep.*, 3:3544, 18 December 2013.

[106] Christoph Muller Patrik Eden Josefin Fernebro Jeanne-Marie Berner Bodil Bjerkehagen Mans Akerman Par-Ola Bendahl Anna Isinger-Anders Rydholm Ola Myklebost Princy Francis, Heidi Maria Namlos and Mef Nilbert. Diagnostic and prognostic gene expression signatures in 177 soft tissue sarcomas: hypoxia-induced transcription profile signifies metastatic potential. *BMC Genomics*, 8:73, 2007.

[107] Smita M. Purandare and Pragna I. Patel. Recombination Hot Spots and Human-Disease. *Biopolym. Cell*, 7:773–786, 1997.

[108] Priit Adler Raivo Kolde, Sven Laur and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.

[109] Altman RB. Raychaudhuri S, Stuart JM. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput.*, page 455466, 2000.

[110] Pirozhkova I. V. Rich J., Ogryzko V. V. Satellite DNA and related diseases. *Biopolym. Cell*, 24(4):249–259, 2014.

[111] Antai Wang Jianhua Xuan Minetta C. Liu-Edmund A. Gehan Robert Clarke, Habtom W. Ressom and Yue Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8:37–49, 2008.

[112] Peter J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, 20:5365, 1987.

[113] Pauline C Ng Lars Feuk Aaron L Halpern-Brian P Walenz Nelson Axelrod Jiaqi Huang-Ewen F Kirkness Gennady Denisov Yuan Lin Jeffrey R MacDonald Andy Wing Chun Pang J. Craig Venter Samuel Levy, Granger Sutton. The Diploid Genome Sequence of an Individual Human. *Plos Biology*, 2007.

[114] Angelika Merkel Alex Dobin Timo Lassmann Ali Mortazavi Andrea Tanzer Julien Lagarde Wei Lin-Felix Schlesinger Chenghai Xue Georgi K. Marinov Jainab Khatun Brian A. Williams Chris Zaleski Joel Rozowsky Maik Rder Felix Kokocinski Rehab F. Abdelhamid Tyler Alioto Igor Antoshechkin Michael T. Baer Nadav S. Bar Philippe Batut Kimberly Bell Ian Bell Sudipto Chakrabortty Xian Chen Jacqueline Chrast Joao Curado Thomas Derrien Jorg Drenkow Erica Dumais Jacqueline Dumais Radha Duttagupta Emilie Falconnet Meagan Fastuca Kata Fejes-Toth Pedro Ferreira Sylvain Foissac Melissa J. Fullwood Hui Gao David Gonzalez Assaf Gordon Harsha Gunawardena Cedric Howald Sonali Jha Rory Johnson Philipp Kapranov Brandon King Colin Kingswood Oscar J. Luo Eddie Park Kimberly Persaud Jonathan B. Preall Paolo Ribeca Brian Risk Daniel Robyr Michael Sammeth Lorian Schaffer Lei-Hoon See Atif Shahab Jorgen Skancke Ana Maria Suzuki Hazuki Takahashi Hagen Tilgner Diane Trout Nathalie Walters Huaien Wang John Wrobel Yanbao Yu Xiaoan Ruan Yoshihide Hayashizaki Jennifer Harrow Mark Gerstein Tim Hubbard Alexandre Reymond Stylianos E. Antonarakis Gregory Hannon Morgan C. Giddings Yijun Ruan Barbara Wold Piero Carninci Roderic Guigo & Thomas R. Gingeras Sarah Djebali, Carrie A. Davis. Landscape of transcription in human cells. *Nature*, 489(5):101108, May 2012.

[115] Ou SH. Crizotinib: a novel and first-in-class multitargeted tyrosine kinase inhibitor for the treatment of anaplastic lymphoma kinase rearranged non-small cell lung cancer and beyond. *Drug Des Devel Ther.*, 5:471–85, 2011.

[116] Olshen Shen and Ladanyi. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.*, 3(3), 2004.

[117] Olshen Shen and Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 22(25):2906–12, 2009.

[118] Jianbo Shi and Jitendra Malik. Normalized Cuts and Image Segmentation. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2(8):888–905, August 2000.

[119] Mark Smolkin and Debashis Ghosh. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 4:36, 2003.

[120] Carey V Dudoit S R Irizarry WH Smyth GK, Gentleman R. Limma: linear models for microarray data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. *New York: Springer*, pages 397–420, 2005.

[121] Bharathi Laxman Daniel R Rhodes Ro-hit Mehra Scott A Tomlins Rajal B Shah Uma Chandran Federico A Monzon Michael J Becich et al. Sooryanarayana Varambally, Jianjun Yu. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer cell*, 8(5):393406, 2005.

[122] Lawrence O. Hall Steven Eschrich, Nitesh V. Chawla. Generalization Methods in Bioinformatics., 2002.

[123] Gowen K Spritz D Amini B-Wang WL Schrock AB Meric-Bernstam F Zinner R Piha-Paul S Zarzour M Elvin JA Erlich RL Stockman DL Vergilio JA Suh JH Stephens PJ Miller V Ross JS Ali SM Subbiah V, Holmes O. Activity of c-Met/ALK Inhibitor Crizotinib and Multi-Kinase VEGF Inhibitor Pazopanib in Metastatic Gastrointestinal Neuroectodermal Tumor Harboring EWSR1-CREB1 Fusion. *Oncology*, 2016.

[124] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, and Benjamin L Ebert. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *PNAS*, 102(43):15545–15550, 2005.

[125] David Tamborero, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S Lawrence, Gad Getz, Gary D Bader, Li Ding, and Nuria Lopez-Bigas. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, 3:2650, 2 October 2013.

[126] Christopher M Tan, Edward Y Chen, Ruth Dannenfelser, Neil R Clark, and Avi Ma'ayan. Network2Canvas: network visualization on a canvas with enrichment analysis. *Bioinformatics*, 29(15):1872–1878, 1 August 2013.

[127] A L Tarca, S Draghici, P Khatri, S S Hassan, P Mittal, Kim J.-s., C J Kim, J P Kusanovic, and R Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2008.

[128] Huisman SMH Ahmed M Krijthe JH-de Ridder J et al. Taskesen E, Erdogan T. Pan-cancer subtyping in a 2d-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci Rep.*, 6:24949, 2016.

[129] Barry S Taylor, Nikolaus Schultz, Haley Hieronymus, Anuradha Gopalan, Yonghong Xiao, Brett S Carver, Vivek K Arora, Poorvi Kaushik, Ethan Cerami, Boris Reva, et al. Integrative genomic profiling of human prostate cancer. *Cancer cell*, 18(1):11–22, 2010.

[130] De Silva V. Langford J.C. Tenenbaum, J.B. Normalized Cuts and Image Segmentation. *Science*, 290:2319–2323, 2000.

[131] Fisher C. Thway K. Tumors with EWSR1-CREB1 and EWSR1-ATF1 fusions: the current status. *Am J Surg Pathol.*, 36(7):e1–e11, 2012.

[132] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–45, 15 June 2010.

[133] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics (Oxford, England)*, 26(12):i237–45, June 2010.

[134] Ainscough BJ Spies NC Skidmore ZL-Campbell KM Krysiak K Pan D McMichael JF Eldred JM-Walker JR Wilson RK Mardis ER Griffith M Griffith OL Wagner AH, Coffman AC. DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.*, 44(D1):D1036–44, 2016.

[135] J N Weinstein, T G Myers, P M O'Connor, S H Friend, A J Fornace, Jr, K W Kohn, T Fojo, S E Bates, L V Rubinstein, N L Anderson, J K Buolamwini, W W van Osdol, A P Monks, D A Scudiero, E A Sausville, D W Zaharevitz, B Bunow, V N Viswanadhan, G S Johnson, R E Wittes, and K D Paull. An information-intensive approach to the molecular pharmacology of cancer. *Science*, 275(5298):343–349, 17 January 1997.

[136] Matthew D. Wilkerson and D. Neil Hayes. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 12(26):15721573, 2010.

[137] Balasubramanian GP Fiesel P Witt R Freitag A Boudalil M Previti C Wolf S Schmidt S Chotewutmontri S Bewerunge-Hudler M Schick M Schlesner M Hutter B Taylor L Borst T Sutter C Bartram CR Milde T Pfaff E Kulozik AE von Stackelberg A Meisel R Borkhardt A Reinhardt D Klusmann JH Fleischhack G Tippelt S Dirksen U Jrgens H Kramm CM von Bueren AO Westermann F Fischer M Burkhardt B Wmann W Nathrath M Bielack SS-Frhwald MC Fulda S Klingebiel T Koscielniak E Schwab M Tremmel R Driever PH Schulte JH Brors B von Deimling A Lichter P Eggert A Capper D Pfister SM Jones DT Witt O Worst BC, van Tilburg CM. Next-generation personalised medicine for high-risk paediatric cancer patients - The INFORM pilot study. *Eur J Cancer*, 65:91–101, 2016.

[138] B N Wylie, K W Boyack, G S Davidson, and D K Johnson. Visualization of information spaces with VxInsight. Technical Report SAND2000-3100, Sandia National Labs., Albuquerque, NM (US); Sandia National Labs., Livermore, CA (US), 1 December 2000.

[139] Yosef Hochberg Yoav Benjamini. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[140] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, Scott L Carter, Gad Getz, Katherine Stemke-Hale, Gordon B Mills, and Roel G W Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, 4:2612, 2013.

[141] Moss YP. Anaplastic Lymphoma Kinase as a Cancer Target in Pediatric Malignancies. *Clin Cancer Res.*, 22(3):546–52, 2016.

[142] Larsson Omberg Nikhil Wagle Ali Amin-Mansour Artem Sokolov Lauren A Byers Yanxun Xu Kenneth R Hess Lixia Diao Leng Han Xuelin Huang Michael S Lawrence John N Weinstein Josh M Stuart Gordon B Mills Levi A Garraway Adam A Margolin Gad Getz & Han Liang Yuan Yuan, Eliezer M Van Allen. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature Biotechnology*, 32:644652, 2014.

[143] McEvoy J Flores-Otero J Ding L Chen X Ulyanov A Wu G Wilson M Wang J Brennan R Rusch M Manning AL Ma J Easton J Shurtleff S Mullighan C Pounds S Mukatira S Gupta P Neale G Zhao D Lu C Fulton RS Fulton LL Hong X Dooling DJ Ochoa K Naeve C Dyson NJ Mardis ER Bahrami A Ellison D Wilson RK Downing JR Dyer MA. Zhang J, Benavente CA. A novel retinoblastoma therapy from genomic and epigenetic analyses. *Nature*, 481:329–34, January 2012.