

5-2017

INTEGRATIVE ANALYSIS OF OMICS DATA IN ADULT GLIOMA AND OTHER TCGA CANCERS TO GUIDE PRECISION MEDICINE

Xin hu

xin hu

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Disease Modeling Commons](#), [Health Information Technology Commons](#), and the
[Translational Medical Research Commons](#)

Recommended Citation

hu, Xin and hu, xin, "INTEGRATIVE ANALYSIS OF OMICS DATA IN ADULT GLIOMA AND OTHER TCGA CANCERS TO GUIDE PRECISION MEDICINE" (2017). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 729.
https://digitalcommons.library.tmc.edu/utgsbs_dissertations/729

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

INTEGRATIVE ANALYSIS OF OMICS DATA IN ADULT GLIOMA AND OTHER TCGA CANCERS TO GUIDE PRECISION MEDICINE

by
Xin Hu, M.S.

APPROVED:

Roel G.W. Verhaak, Ph.D., Supervisory advisor

Erik P. Sulman, M.D., Ph.D., On-site advisor

Keith A. Baggerly, Ph.D.

John N. Weinstein, M.D., Ph.D.

John F. de Groot, M.D.

APPROVED:

Dean, The University of Texas

MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

INTEGRATIVE ANALYSIS OF OMICS DATA IN ADULT GLIOMA AND OTHER TCGA CANCERS TO GUIDE PRECISION MEDICINE

A Thesis Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

In Partial Fulfillment of the Requirements for the

Degree of DOCTOR OF PHILOSOPHY in

Bioinformatics, Biostatistics and Systems Biology

by Xin Hu, M.S.

Houston, Texas

May, 2017

DEDICATIONS

To my parents, my advisors, my committee members, my co-workers and friends,
who kindly supported me throughout this critical stage of career transition and evolution, for
sharing your precious time, knowledge and experiences, I couldn't have accomplished my
graduation works without your generous help.

ACKNOWLEDGEMENTS

I would like to extend my gratitude to all of my past and current committee members in MD Anderson Cancer Center. First I would like to recognize Dr. Roel Verhaak, my dissertation committee chairman and supervisory advisor, for his kind encouragement, guidance and patience in the completion of my Ph.D. thesis. I cherished those precious moments when we sat together in front of his PC screen, sharing his opinions about research data or ideas towards scientific career. I appreciate Dr. Verhaak, who served as a role model of rapid success not only in academic research but also in overall life and career. I am so impressed by his open-minded, broad vision in research perspectives and efficient construction of his vivid lab. Surely I wholeheartedly appreciate the intensive research atmosphere and collaborative infrastructures in his lab that sparked my interest in genomic medicine, while expanding my awareness of diverse topics of translational genomics and computational biology.

My genuine appreciations are extended to Dr. Jean-Pierre Issa and Dr. Marcos Estecio, Dr. Jaroslav, my former advisor and instructors, who not only boosted my interest in epigenomes but also offered me with so many generous help, valuable suggestions and kind supports for my graduate study and career development. I really appreciate Dr. Issa for constantly putting his trainees' best interests at heart; he taught me essential elements of critical thinking.

I also express my sincere gratitude to Dr. Keith Baggerly and Dr. Ralf Krahe, for their dedicated efforts and time with rigorous training and constructive guidance, their rigorous style of re-productivity and precision on research data have great impact on my long-term research career. The critical thinking, highly organized research style I learned from Dr. Baggerly greatly benefited my collaborative work in my later Ph.D. stages, and I believe his rigorous philosophy will impact my research in a long run. Many thanks for Dr. Weinstein, for his dedicated contributions to my committee and suggestions of my projects, as well as the

organization of BBSB program, which provided valuable resources for my scientific growth and career development. Many thanks to Dr. Erik Sulman, who served as my on-site advisor and raise very insightful questions in the previous committee meeting, and to the other members of my committee, Dr. John Degroot, plus previous committee members Dr. Xiaobing Shi, Dr. Gigi, Dr. Peng Huang, etc, for both advisory and candidacy exam committees, thank you very much for the constructive feedback and suggestions. I also owe my thanks to Dr. Shoudan Liang, Dr. Yue Lu and Dr. Kenneth Hess, who provided me with important guidance in bioinformatics and statistics training.

I am very grateful for BBSB program co-directors Dr. Wenyi Wang, Dr. Prahlad Ram, Dr. Ken Chen and Dr. Sanjay Shete for their kind advice on my presentation and career development. My sincere appreciation also extended to Dr. William Mattox, Dr. Michelle Barton, Dr. Knutson, Ms. Brenda Gaughan, Ms. Patricia Cruz Bruesch, Ms. Joy Lademora, Ms. Lourdes Perez, Ms. Lily, Mr. Michael Valloadolid and other staff, for your dedicated and friendly support throughout my Ph.D. study. I really appreciate Ms. Brenda for her great administrative efforts to guide me through the final stage of GSBS. Personally I considered GSBS as the best program in US, even in the world, for its flexibility, diverse program structure, and enriched resources, as well as humanized management system and culture, taking care about students' research interests, benefits and career development into full play.

I would like to express my sincere acknowledge to IT system administrators, Jinzhen (Jenny) Chen, Sally Boyd, Daniel Jackson, Benton Karen, Dian (Oscar) Jiao, Rong Yao, Dr. Bradley Broom, other HPC teams, as well as Jun Zhang, Allen Chang, and other DQS staffs, and also Jeff Davis from cancer medicine division, for their efficient and responsible information & technology support. Jenny, Sally, Rong often worked very early or after hours even during the weekend to take care of the HPC system.

I owe faithful gratitude to the Verhaak lab members, Dr. Siyuan Zheng, Dr. Qianghu Wang, Dr. Emmanuel Martinez-Ledesma, Dr. Kosuke Yoshihara, Dr. Floris Barthel, Dr. Amin,

Samirkumar, Dr. Ming Tang, Dr. Hoon Kim, Dr. Zeyan Zhang, Dr. Lee, Soo Hyun, for the precious collaborative activities and discussions within the group, as well as other faculty and research scientists in the department of genomic medicine, Dr. Andy Futreal, Dr. Jianhua Zhang, Dr. Chiachin Wu, etc. Moreover, I appreciate very much for many of my schoolmates from GSBS, for their friendly help and support.

Last but not least, I would deeply appreciate my parents, for their enormous sacrifice, unselfish forgiveness and mentally supporting me to pursue my academic career without coming to bother me, who missed out the duty of taking care of them while struggling on this long career path. Words could never express my profound apologies to them. Finally, I would like to extend my truthful appreciation to those kind souls from those who helped me walk along this special journey of career transition.

INTEGRATIVE ANALYSIS OF OMICS DATA IN ADULT GLIOMA AND OTHER TCGA CANCERS TO GUIDE PRECISION MEDICINE

Xin Hu, M.S.

Supervisory advisor: Roel Verhaak, Ph.D.

Transcriptomic profiling and gene expression signatures have been widely applied as effective approaches for enhancing the molecular classification, diagnosis, prognosis or prediction of therapeutic response towards personalized therapy for cancer patients. Thanks to modern genome-wide profiling technology, scientists are able to build engines leveraging massive genomic variations and integrating with clinical data to identify “at risk” individuals for the sake of prevention, diagnosis and therapeutic interventions. In my graduate work for my Ph.D. thesis, I have investigated genomic sequencing data mining to comprehensively characterize molecular classifications and aberrant genomic events associated with clinical prognosis and treatment response, through applying high dimensional omics genomic data to promote the understanding of gene signatures and somatic molecular alterations contributing to cancer progression and clinical outcomes. Following this motivation, my dissertation has been focused on the following three topics in translational genomics.

1) Characterization of transcriptomic plasticity and its association with the tumor microenvironment in glioblastoma (GBM). I have integrated transcriptomic, genomic, protein and clinical data to increase the accuracy of GBM classification, and identify the association between the GBM mesenchymal subtype and reduced tumor purity, accompanied with increased presence of tumor-associated microglia. Then I have tackled the sole source of microglial as intrinsic tumor bulk but not their corresponding neurosphere cells through both transcriptional and protein level analysis using a panel of sphere-forming glioma cultures and their parent GBM samples. Furthermore I have demonstrated my hypothesis through

longitudinal analysis of paired primary and recurrent GBM samples that the phenotypic alterations of GBM subtypes are not due to intrinsic proneural-to-mesenchymal transition in tumor cells, rather it is intertwined with increased level of microglia upon disease recurrence. Collectively I have elucidated the critical role of tumor microenvironment (Microglia and macrophages from central nervous system) contributing to the intra-tumor heterogeneity and accurate classification of GBM patients based on transcriptomic profiling, which will not only significantly impact on clinical perspective but also pave the way for preclinical cancer research.

2) Identification of prognostic gene signatures that stratify adult diffuse glioma patients harboring 1p/19q co-deletions. I have compared multiple statistical methods and derived a gene signature significantly associated with survival by applying a machine learning algorithm. Then I have identified inflammatory response and acetylation activity that associated with malignant progression of 1p/19q co-deleted glioma. In addition, I showed this signature translates to other types of adult diffuse glioma, suggesting its universality in the pathobiology of other subset gliomas. My efforts on integrative data analysis of this highly curated data set using optimized statistical models will reflect the pending update to WHO classification system of tumors in the central nervous system (CNS).

3) Comprehensive characterization of somatic fusion transcripts in Pan-Cancers. I have identified a panel of novel fusion transcripts across all of TCGA cancer types through transcriptomic profiling. Then I have predicted fusion proteins with kinase activity and hub function of pathway network based on the annotation of genetically mobile domains and functional domain architectures. I have evaluated a panel of in frame gene fusions as potential driver mutations based on network fusion centrality hypothesis. I have also characterized the emerging complexity of genetic architecture in fusion transcripts through

integrating genomic structure and somatic variants and delineating the distinct genomic patterns of fusion events across different cancer types. Overall my exploration of the pathogenetic impact and clinical relevance of candidate gene fusions have provided fundamental insights into the management of a subset of cancer patients by predicting the oncogenic signaling and specific drug targets encoded by these fusion genes.

Taken together, the translational genomic research I have conducted during my Ph.D. study will shed new light on precision medicine and contribute to the cancer research community. The novel classification concept, gene signature and fusion transcripts I have identified will address several hotly debated issues in translational genomics, such as complex interactions between tumor bulks and their adjacent microenvironments, prognostic markers for clinical diagnostics and personalized therapy, distinct patterns of genomic structure alterations and oncogenic events in different cancer types, therefore facilitating our understanding of genomic alterations and moving us towards the development of precision medicine.

KEYWORDS

Cancer genomics, Omics data mining, Transcriptomics profiling, Gene signature, Risk prediction, Tumor microenvironment, Precision medicine, Glioma, Pan-cancer, Fusion transcripts, Genomic rearrangement, Next-generation sequencing, Prediction model

TABLE OF CONTENTS

APPROVAL PAGE	I
TITLE PAGE	II
ACKNOWLEDGEMENTS	IV
ABSTRACT	VII
LIST OF FIGURES	XIV
LIST OF TABLES	XIX
CHAPTERS	
1. Introduction	1
1.1 Promise of next generation sequencing (NGS) data applied in precision medicine	1
1.2 NGS technology promoted the advances of cancer genome	3
1.3 Molecular classification and potential clinical implications in cancer	4
1.4 Link tumor microenvironment (TME) with cancer molecular classification.....	5
1.5 The implications of cancer heterogeneity on molecular classifications.....	9
1.6 Towards personalized therapy for patients with glioma	10
1.7 Supervised learning for personalized medicine.....	12
1.8 Fusion genes as an emerging target for precision medicine	18
1.9 Motivations and Rationale of the study in this thesis.....	20
2. Tumor microenvironment associated with intrinsic transcriptomic subtype during glioma evolution.....	24
2.1 Introduction.....	24
2.2 Methods	26
2.2.1 Data sources for multiplatform classification comparison	26
2.2.2 Transcriptome data processing	27
2.2.3 Identification of gene signatures for refined GBM subtype	28
2.2.4 Molecular classifications based on ssGSEA enrichment scores	29

2.2.5 Evaluate the heterogeneity of GBM subtype	30
2.2.6 Tumor purity assessment	30
2.2.7 Establish glioma neurosphere cultures (GSCs)	31
2.2.8 Western blotting	31
2.2.9 Immunohistochemistry	31
2.3 Results	32
2.3.1 Harnessing glioma sphere-forming cells identifies GBM specific inter-tumoral transcriptional heterogeneity	32
2.3.2 Multi-activation of subtype signatures associated with intra-tumoral heterogeneity...	40
2.3.3 Transcriptional subtypes differentially activate the immune microenvironment.....	45
2.3.4 Phenotypic plasticity upon GBM recurrence	51
2.3.5 Tumor microenvironment transitions upon GBM recurrence	54
2.3.6 Treatment-induced immunological microenvironment changes upon GBM relapse..	60
2.4 Discussion	62
3. Multi-Gene Signature for Predicting Prognosis of Patients with 1p19q Co-deletion	
Diffuse Glioma	65
3.1 Introduction.....	65
3.2 Methods	67
3.2.1 Process of the datasets	67
3.2.2 Predicting 1p/19 status using gene expression	71
3.2.3 Correlation of somatic mutations and clinical outcome	72
3.2.4 Gene signature selection and risk based classification	72
3.2.5 Association of risk classification and clinical outcome	73
3.2.6 Gene Set Variation Analysis of Associated Genes and Top Gene Ontology Terms...	73
3.2.7 Evaluation of Tumor Purity with ESTIMATE gene signatures.....	74
3.3 Results	74

3.3.1 Effects of somatic mutations on patient outcome	74
3.3.2 Constructing a gene expression data set of 1p/19q co-deleted glioma	75
3.3.3 Identification of a 35-gene signature associated with overall survival	78
3.3.4 Multivariable analysis shows prognostic power of 35-gene signature	88
3.3.5 Functional Annotation of 35 Gene Signatures	89
3.3.6 Applying the 35-Gene Signature across Glioma	92
3.4 Discussion	95
4. Prediction of emerging fusion transcripts with oncogenic potential in The Cancer Genome Atlas pan- cancers	99
4.1 Introduction.....	99
4.2 Methods	100
4.2.1 Data resources	101
4.2.2 Identification of fusion transcripts	101
4.2.3 Validation of fusion transcripts through integrating genomic features	102
4.2.4 Pathway and fusion centrality analysis	103
4.2.5 Exon expression analysis	103
4.3 Results	104
4.3.1 Distribution of fusion transcripts in different cancer types	104
4.3.2 Gene fusion annotation and tight association with genomic structure alterations ...	108
4.3.3 Hotspot fusion transcripts are associated with genomic instability	113
4.3.4 Prioritizing functional fusions	116
4.3.5 Recurrent fusion transcripts in normal tissues	123
4.3.6 Recurrent fusion transcripts are mutually exclusive with somatic mutations.....	124
4.3.7 Perturbed pathways associated with fusion formation.....	128
4.3.8 Cancer type–specific fusion gene networks and hubs	129
4.3.9 Altered kinase domains in fusion transcripts deregulate oncogene functionality.....	134

4.3.10 Loss of epigenetic modification domains in fusion transcripts deregulate tumor suppressor genes.....	138
4.3.11 Gain or loss of post-translational modification sites in fusion deregulate oncogenes and tumor suppressor genes	140
4.4 Discussion	143
5. Conclusions and future perspectives	148
5.1 Summary	148
5.2 Significance, pitfalls and perspective explorations.....	148
5.2.1 Classification of glioma integrating transcriptomic profiling from both intrinsic tumor and infiltrated microenvironment	148
5.2.2 Identify gene signatures associated with clinical outcomes in low grade glioma	152
5.2.3 Characterize distinct genomic patterns of fusion transcripts and identify novel fusion events conferring oncogenic potential in pan-cancers	156

LIST OF FIGURES

Figure 2.1 Selection of gene signature for molecular classification of IDH-WT GBMs	34
Figure 2.2 Comparison between GCIMP- GBM specific classification and previously TCGA defined GBM subtypes	34
Figure 2.3 Gene signatures applied for GBM classification	35
Figure 2.4 Concordance of transcriptional classification of GBMs cross multiple platforms.	39
Figure 2.5 Multi activation of transcriptional subtypes associated with intra-tumoral heterogeneity.....	40
Figure 2.6 Genomic alteration patterns for each subtype	41
Figure 2.7 Association between genomic and transcriptomic intratumoral heterogeneity...	42
Figure 2.8 Impact of intratumoral heterogeneity on overall survival between different transcriptional subtypes	43
Figure 2.9 Impact of intra-tumoral heterogeneity on patient survival in each transcriptional subtype	44
Figure 2.10 Transcriptome classification of five bulk tumor samples their derived 502 single GBM cells	45
Figure 2.11 Transcriptional subtypes differentially activate the immune microenvironment	46
Figure 2.12 Comparison of tumor purity and immune cell fraction between GBMs with different NF1 genomic status	47
Figure 2.13 Comparison of immune cell fraction among different subtypes of GBM	49
Figure 2.14 Presence of macrophages/microglia in MES GBM	49
Figure 2.15 Characterization of the source of microglia in GBM	50
Figure 2.16 Integrative views of molecular classification and genomic alterations across molecular subtypes in paired primary and recurrent glioblastoma	52

Figure 2.17 Comparison between transcriptional subtype of primary and paired recurrent tumors	53
Figure 2.18 Microenvironment transition between primary and paired recurrent tumors	55
Figure 2.19 Comparison of M2 Macrophage fractions between MES and none-MES subtypes during tumor evolution.....	56
Figure 2.20 Comparison of M2 macrophage cell fractions in primary and matched recurrent tumors	57
Figure 2.21 Survival analysis of paired IDH wild type GBM	59
Figure 2.22 Survival analysis on MES versus non-MES patients	60
Figure 2.23 Comparison of CD8+ T cell fraction associated with gain of hyper-mutation induced by chemotherapy.....	61
Figure 2.24 Comparison of immune cell fractions in paired samples upon relapse after different period of radiation	62
Figure 3.1 Workflow for identification and validation of prognostic gene signature using the elastic net	69
Figure 3.2 Co-deletion of 1p/19q inferred by gene expression profiling	76
Figure 3.3 Distribution of normalized gene expression datasets from different data sources	77
Figure 3.4 Partial likelihood deviances as function of regularization parameter λ for 3-fold-cross validation in the training dataset.....	80
Figure 3.5 Kaplan-Meier survival analysis of glioma patients harboring 1p/19q co-deletion according to 35-gene signature derived risk scores	81
Figure 3.6 Consistent prediction of prognosis in high- and low-risk groups	82
Figure 3.7 Hazard Ratios (HR) with 95% CIs for gene signatures in high- versus low-risk groups.....	83

Figure 3.8 Kaplan-Meier survival analysis of co-deletion glioma according to 35-gene signature derived risk scores	84
Figure 3.9 Schoenfeld residual plots to evaluate the proportional hazard assumption of signature genes	85
Figure 3.10 Martingale residuals to evaluate the linearity of log hazard on signature genes	86
Figure 3.11 Correlation between gene signatures derived risk scores and overall survival time	88
Figure 3.12 mRNA differential expression patterns of 35 signature genes in two risk groups.....	90
Figure 3.13 Association of risk groups with gene ontology (GO) function.....	91
Figure 3.14 Comparison of ESTIMATE scores in high- and low-risk group in the first validation dataset of 1p/19q code1	92
Figure 3.15 Prediction of outcome in non-code1 IDH-mutant glioma and IDH-wildtype glioma.....	93
Figure 3.16 Comparison of ESTIMATE scores in high- and low-risk group of IDH-mutant-non-code1 and IDH-wildtype glioma	94
Figure 4.1 Spectrum of filtered fusion transcripts across 33 cancer types.....	105
Figure 4.2 Counts of fusion transcripts supported by structure variants at different chromosome arms in 23 cancer types	106
Figure 4.3 Distribution of genomic features among fusion transcripts across 33 cancer types	108
Figure 4.4 Counts of fusion transcripts and structure variants in each tumor sample	109
Figure 4.5 Distribution of fusion transcripts associated with structure variants	110
Figure 4.6A Fusion transcript derived from inversion in BLCA	111
Figure 4.6B Fusion transcript derived from translocation in SKCM	112

Figure 4.6C Fusion transcript derived from large fragment deletion in SARC	113
Figure 4.7 Association between hotspots of recurrent fusion transcripts with chromosome rearrangement and copy number alterations in SARC	115
Figure 4.8 Frequency of recurrent fusion transcripts across 33 cancer types.....	116
Figure 4.9 Most frequent partner genes in recurrent fusion transcripts across 33 cancer types	117
Figure 4.10A Centrality scores of fusion partner genes across 33 cancer types.....	118
Figure 4.10B Proportion of driver fusions among all fusion transcripts across 33 cancer types	119
Figure 4.11 Number of fusion transcripts composed of tumor suppressor genes and oncogenes across 33 cancer types	120
Figure 4.12A Increased transcription levels in recurrent fusion transcripts of C10orf68-CCDC7 in multiple cancer types	122
Figure 4.12B Z-normalized exon expression of C10orf68-CCDC7 in multiple cancer types	123
Figure 4.13 Frequency of recurrent fusion transcripts in normal tissue across 12 tissue types	124
Figure 4.14A Recurrent TYK2-containing fusion transcripts are mutually exclusive with somatic mutation events in multiple cancer types	126
Figure 4.14B Recurrent EIF2AK2-containing fusion transcripts are mutually exclusive with somatic mutation events in multiple cancer types	127
Figure 4.15 Most represented ontological categories in gene ontology analysis of fusion transcripts in 33 cancer types	128
Figure 4.16 Perturbed pathways from fusion transcripts across 33 cancer types	129
Figure 4.17 Network of gene fusions in multiple cancers	130

Figure 4.18 Network of gene fusions in cancer types with high frequency of fusion transcripts.....	131
Figure 4.19 Network of gene fusions in liver cancer (LIHC).....	132
Figure 4.20 Network of gene fusions in stomach adenocarcinoma (STAD).....	133
Figure 4.21 Novel fusion transcript <i>PDGFRA-USP8</i> harboring kinase domains in sarcomas.....	135
Figure 4.22 Novel fusion transcript involved kinase <i>PRKCB</i> in different cancer types.....	137
Figure 4.23 Novel fusion transcript affects chromatin remodeling domain (SET) in BRCA.....	139
Figure 4.24 Novel fusion transcripts <i>ARID1B-EZR</i> in Uterine Corpus Endometrial Carcinoma (UCEC)	140
Figure 4.25 Fusion-induced ubiquitination binding site losses in oncogene	142
Figure 4.26 Fusion-induced ubiquitination binding sites gained in tumor suppressor genes from fusion <i>HKR1-CEACAM7</i> in esophageal carcinoma (ESCA).....	143
Figure 5.1 Clinical relevant transcriptomic subtypes determined by both inflammatory and adaptive immune components	151
Figure 5.2 Subtype-specific molecular alterations and distinct clinical presentations in Lower-Grade Gliomas (LGG)	153
Figure 5.3 Partial plots for gene predictors from codel glioma	154

LIST OF TABLES

Table 2.1 Gene signatures applied for transcriptomic classification of GBM	35
Table 3.1 Clinical characteristics of glioma patients harboring 1p/19q co-deletion	70
Table 3.2 Data sources for gene expression and CNV in codel glioma patients	71
Table 3.3 Annotation of 35 gene signature for prediction	79
Table 3.4 Variance inflation factors (VIF) of 35 gene signature for prediction in the training dataset	87
Table 3.5 Performance of Multivariable Analysis in Validation Dataset	89
Table S4.1 Number of TCGA samples in 33 cancer types	158
Table S4.2 Number of fusions filtered by each step in 33 cancer types	159

CHAPTER 1

INTRODUCTION

1.1 Promise of next generation sequencing (NGS) data applied in precision medicine

1.1.1 Big (NGS) data brings the promise of precision medicine to clinical oncology

The molecular classification of cancer based on next generation sequencing (NGS) has built the foundation for more precision drug development. Taking advantage of genomic sequencing technologies, scientists are able to decipher complex features of an individual's cancer genome, determine the risk factors for prognosis and tailor targeted therapies based on gene signatures and genomic variants. In the genomic era, the paradigm of clinical oncology is largely shifting from empirical, retrospective diagnosis with uniform treatment strategies to mechanism-based diagnosis and treatment driven by molecular diagnostics and risk based individualized regimens. For instance, a meta-analysis of 570 phase II clinical trials involving 32,149 patients demonstrated that patients whose treatment was selected based on the molecular characteristics of their tumor exhibit significantly better outcomes.(Schwaederle, Zhao et al. 2015) Applying advanced genomic sequencing, the Oncotype DX® breast cancer test has efficiently identified patients with higher likelihood of response to chemotherapy in both pre-invasive stage (DCIS) and invasive breast cancer with 600,000 cases examined across more than 90 countries, demonstrating the power of personalized medicine by taking account of genomics as a critical part of clinical decision making for oncologist. The Cancer Genome Atlas (TCGA) network has sequenced more than 11,000 tumors across 34 cancer types, including the identification of germ-line and somatic aberrations (SNVs), DNA copy number alterations, fusion events, differential DNA methylation, and transcriptomic classification. (Martinez-Ledesma, de Groot et al. 2015)

1.1.2 The challenge of leveraging cancer genomics in personalized medicine

Regarding profiling based assignment of cancer therapeutics, while stratifying patients and developing targeted therapies based on genomic biomarkers, scientists and clinicians still face challenges regarding specimen acquisition and heterogeneity. Somatic mutation, clonal selection and genetic drift are the basic processes shaping cancer evolution, the relationships between sample sizes, mutation frequency / rate and predictability create a complex and non-monotonic milieu. Time-to-event predictability is more difficult in tumors with limited mutational burdens due to stochastic mutation generation and drift events. The development of an integrative cancer evolutionary framework using refined computational modeling is of critical importance in order to obtain accurate predictions. The predictive models need to incorporate the spatial constraints in tumor samples with optimal sampling approaches and input parameters dependent on the variability of the tumor, such as growth, metastasis and driver events. Nevertheless, cancer evolution processes influenced by stochastic effects on the entire clonal composition will still influence the predictive power in clinical scenario. (Lipinski, Barber et al. 2016) In addition, passenger mutations are “collateral damage” resulting from genomic instability and are not required for maintaining the transformed phenotype, therefore are “noise” in the predictive system, and since most cancers are rapidly evolving biologic entities, major challenges remain to sort out “drivers” from “passengers”, and these events may change over time. Driver mutations in the signaling pathways affect highly integrated “wiring” comprised of multiple signal transduction flows. Perturbation of a single component of a regulatory network will lead to activation of other components due to feedback activation or loss of feedback repression, which can contribute to drug resistance. It should be noted that many novel epigenetic drivers are not currently targetable with available drugs. Determining trial designs and endpoints are also challenges aspects for precision medicine. Other practical challenges for moving precision oncology into cancer care include, improving sequencing platforms, designing clinical utility, and exploring novel perspectives. There are also ethical and social issues, such as patients’

consent for cancer genetic testing, genomic/genetics test result confidentiality and disclosure, insurance costs for pharmaco-genomic testing, and combined targeted therapeutic regimens for cancer patients. (McGowan, Settersten et al. 2014)

1.2 NGS technology promoted advances in our understanding of cancer genomes

Over the last decade, rapid advances in cancer genomics have shed new lights on cancer biology, promoting comprehensive understanding at a systematic level of the disease and unraveling the multi-dimensional landscapes of aberrant genomic structures that contribute to cancer progression, which has led to novel perspectives for managing the disease (Samantarrai, Dash et al. 2013). NGS has been broadly applied for genome, epigenome and transcriptome sequencing. NGS outperforms traditional genetic test for its accuracy, sensitivity and efficiency, and it has proven mature enough for routine diagnostic use in many laboratories. Next-generation sequencing encompasses worldwide collaborative efforts, such as The Cancer Genome Atlas (TCGA) and International Genome Consortium (ICGC) projects, to characterize the genomic landscape and transcriptomic subtypes of thousands of cancer genomes across various cancer types. These discoveries lead to new fundamental understanding of disease pathogenesis. Molecular profiling has also been established for the identification of unique somatic mutations and gene signatures that accrue in cancer cells. Cancer genomic technology in turn has evolved to facilitate molecular profiling, enabling the assessment of all potential causative or predictive genes in panels using targeted sequencing, especially using circulating tumor ctDNA to noninvasively identify cancer biomarkers for early detection (Cortesi, Palleschi et al. 2015). Since most common clinical oncology phenotypes show diverse responses to the same therapy regimens in patients with equivalent diagnose, it is conceivable to have both their germline and cancer genomes sequenced for each patient to facilitate rationally guided molecular therapies, thanks to the updated NGS technology, allowing several genomes to be

sequenced simultaneously using one affordable instrument run by one scientist within 10 days. Thus, by integrating traditional pathologic diagnosis with other approaches such as molecular imaging, NGS technology allow us to treat each genetic abnormality as an independent variable. Given thousands of variables for each patient, we can comprehensively integrate genomic information and clinical phenotypes with new trial designs and statistical methods (Meldrum, Doyle et al. 2011, Gagan and Van Allen 2015) (Katsios, Ziogas et al. 2012) (Boerno, Grimm et al. 2010) (Barbieri, Demichelis et al. 2012) (Roukos 2010, Roychowdhury and Chinnaiyan 2016) Moreover, with revolutionized NGS technology, emerging advances in cancer genomics may change the strategy for current surgical oncology practice as well (Roukos 2011). Since treatment decisions such as complete tumor resection with or without adjuvant radio/chemo therapy depends on cancer type and stage, genome sequencing in cancer patients could systematically assess the complex nature of cancer initiation and metastasis, with widespread variability of somatic mutations, genomic rearrangements and copy-number changes, and these alterations in turn deregulated signaling pathways. Therefore the genomic features will ultimately determine the molecular classification of patients associated with their clinical-pathologic status.

1.3 Molecular classification and potential clinical implications in cancer

Cancer is a biologically heterogeneous disease, which is reflected by complex etiological pathways that drive tumor evolution, and the network of signaling pathways could be potentially modulated by distinct genetic/ epigenetic and transcriptional / translational alterations. Therefore understanding of molecular heterogeneity could facilitate clinical management and cancer prognosis, and intra-tumor heterogeneity of genetic aberrations could classify discrepancies in the validation of oncologic biomarkers, and treatment resistance (Eder and Kalman 2014). Cancer diagnosis is traditionally based on histological

examination and in the precision medicine era, it is crucial to accurately stratify patients by integrating genomic and histological diagnosis. A series of molecular diagnostic tests have been developed based on intrinsic gene signatures incorporated into the histopathological classification systems for different types of cancer. (Vitucci, Hayes et al. 2011) (Francis, Namlos et al. 2007)

Scientists have made many efforts to characterize the correlation of molecular subtype and phenotypes for cancer, towards a better understanding of cancer development and individual risk prediction, and guiding early detection, clinical decisions and prevention from recurrence in more informative manners than the traditional “wait and see” approach. (Kocarnik, Shiovitz et al. 2015) (Aine, Eriksson et al. 2015, Prat, Pineda et al. 2015, Chen, Xu et al. 2016) The profiling of cancer has yielded a number of genetic, epigenetics (DNA methylation), mRNA, microRNA expression, proteomic, metabolic, and imaging biomarkers for molecular classification, facilitated by next-generation sequencing. As a matter of fact, molecular classification has already been successfully implemented in many types of cancer, such as classification of EGFR / ALK positive non-small-cell lung cancers, subtyping of luminal A/B estrogen receptor-positive breast cancer, classification of WNT pathway and Sonic Hedgehog pathway medulloblastomas. The efficiency of molecular classification facilitating accurate diagnosis and combination chemo-radiation therapy for cancer is apparent, such as molecular classification of pheochromocytomas and gastrointestinal stromal tumors using 92-gene assay in metastatic lesions and primary patients (Greco 2013, Brachtel, Operana et al. 2016) Overall, previous studies suggest that molecular profiling based classification outperforms current pathology-based systems in terms of clinically relevance.(Hoadley, Yau et al. 2014)

1.4 Linking the tumor microenvironment (TME) with cancer molecular classification

1.4.1 The tumor microenvironment (TME) is a pathological driver that modulates tumor evolution

The tumor microenvironment (TME) plays critical roles in multiple spatial and temporal genomic alterations, particularly during immune-invasion and distal metastasis, which commonly lead to treatment resistance (Sun 2016). The essential functional components in the stroma from typical TME include neuroendocrine cells, fibroblasts, immune and inflammatory cells, myofibroblasts, blood /lymphatic vascular networks, adipose cells and extracellular matrix (ECM).(Razavi, Lee et al. 2016) The stroma could suppress tumorigenesis in its native status and is a predictor for favorable longevity (Elkhattouti, Hassan et al. 2015), yet when stroma was transformed into tumor-associated neighbor components by various angiogenic and/or inflammatory stimuli, it exhibit adverse effects and significantly promote cancer progression, thereby the status and functionality of TME have substantial impacts on the clinical decision for frontline therapeutic interventions.(Chen, Zhuang et al. 2015) For example, stroma-mediated drug resistance through regulating innate immune response, tumor infiltration and vascular permeability were reported in animal models of breast cancer. (Nakasone, Askautrud et al. 2012) And cancer-associated fibroblast (CAF) mediating the resistance to EGFR inhibitor erlotinib through inducing epithelial to mesenchymal transition (EMT) was previously reported in lung cancer. (Choe, Shin et al. 2015) In addition, stromal cell-mediated mitochondrial redox adaptation was described to regulate drug resistance in childhood acute lymphoblastic leukemia (ALL). (Liu, Masurekar et al. 2015)

1.4.2 The immune classifications of tumors and precision immunotherapy targeting immune microenvironments (TME)

Associations between the cellular composition of the tumor microenvironment and genomic features of the tumor are emerging. Dissecting the microenvironment of tumor molecular

subgroups could facilitate the prediction of disease progression or response to immunotherapies. Cancer cells are intertwined with a tumor microenvironment comprised of stromal cells (e.g. endothelial cells, fibroblasts and ectodermal stem cells) and immune cells in a complicated interactive manner. This interaction could determine the clinical outcome of cancer progression and response to therapy. For example, the adaptive immune system with high activity of cytotoxic and memory T cells could regulate tumor growth and metastasis, and this immune signaling could lead to favorable patient prognosis in colorectal cancer (CRC) (Markman and Shiao 2015). Other adaptive immune cells such as type I T helper (Th1) lymphocytes could activate both cytotoxic T cells and B-lymphocytes, which secrete antibodies attacking tumor (Fridman, Pages et al. 2012). Lymphocytes forming aggregates surrounding the tumor (Jiang, Mason et al. 2013, Ladanyi 2013) that mediate systemic immune responses have also been reported in non-small-cell lung cancer. These findings demonstrated the potential of immune response to translate into the clinics, such as immune checkpoint inhibitors stimulating cytotoxic T cells activity demonstrated favorable clinical responses in patients with advanced grade of malignancy. Bodies of evidence have demonstrated the association between the constitution of the tumor microenvironment and a patient's prognosis, and more noteworthy, some of these components in the tumor microenvironment can be therapeutically targeted. With distinct correlation between tumor infiltrated cellular compositions and their genomic features, as well as response to immunotherapy, it is of paramount importance to analyze the microenvironment associated with tumor molecular subgroups and develop corresponding multi-dimensional tumor classification systems to select the responders and tailor the treatment regimens. Several types of cancer have been classified into molecularly homogeneous subgroups, such as GBM, LGG, colon cancer, breast cancer, and so on. Overall these subtypes are established through unsupervised classification of 'omics' data, based on the distinct molecular signatures that link genomic features and clinical

characteristics of the patients. This strategy was also applied to dissect the relationship between molecular subtypes and their associated immune microenvironment in diverse cancer types. For example, applying transcriptomic profiles from infiltration of immune and stromal cellular components identified molecular subgroups of clear-cell renal cell carcinoma (ccRCC) (Remark, Alifano et al. 2013). This study revealed that various immune cellular components in the tumor microenvironment are influenced by the metastasis derived malignant cells adjacent to the same surrounding pulmonary tissues and suggested essential roles of metastasis-deriving malignant cells in modulating the tumor immune microenvironment, and indicated significant correlation between the tumor molecular signature and their immunological features. This finding also implied that the originating anti-tumor cells could exhibit suppressed activity accompanied with co-expression of checkpoint molecules in a highly inflammatory microenvironment, suggesting that the potential responders to PD-1 pathway inhibition are enriched in some molecular subgroup of ccRCC.

It was also well documented that pro-tumorigenic inflammation signals could regulate the immune system to restore homeostasis disruption, such as infections or wound healing. Tumor cells acquired the innate capacity to decay inflammatory signals and produce excessive mutagens, angiogenic growth factors and activate extracellular matrix remodeling and collagen turnover pathways (Candido and Hagemann 2013). Inflammation plays a critical role by suppressing acquired anti-tumor immune responses via stimulating MDSC (myeloid-derived suppressor cells), regulatory T cells, and immunosuppressive factors such as transforming growth factor (TGF β). One of the strategies of pharmacogenomic development of tailored immunotherapies is to simultaneously restore the adaptive immune response and dampen the pro-tumor inflammatory response, underscoring the translational value of dissecting the tumor microenvironment to integrate molecular classification of both immune and intrinsic tumors.

Another striking example was revealed by the stromal and immune classification of colorectal cancer, which was significantly correlated with established subtypes of the tumor. (Becht, Giraldo et al. 2015) A novel 'immune-high' subgroup of CRC, is in alignment with poor-prognosis mesenchymal subgroup that expressed the genes triggering adaptive immune response and checkpoint molecules. Tumors in this subtype were also infiltrated with macrophages, increased angiogenesis, high expression of inflammatory genes and fibroblast infiltration, plus abundant immunosuppressive factors such as TGF β . This study suggested that increased inflammation could break the cytotoxic cells activity in the colorectal cancer (CRC) mesenchymal subtype, therefore anti-angiogenic and anti-inflammatory treatments, combined with checkpoint inhibitors, such as anti-PD-L1/2, could simultaneously damp inflammatory signaling and restore cytotoxic T-cell functionality. These treatment strategies will be beneficial for CRC patients in this subgroup.

Noteworthy, another independent study showed that overall mutational load is associated with beneficial response to PD-1 blockage in non-small cell lung cancer (Rizvi, Hellmann et al. 2015, Van Allen, Miao et al. 2015). This further proved the antigenicity to trigger an adaptive immune response in the tumor microenvironment, which maybe a major driver in response to the checkpoint inhibitors, is tightly associated with somatic non-synonymous mutations and cytotoxic orientation of the microenvironment.

Taken together, these previous discoveries have proved the tight association between genomic features and immune classifications of tumors, and the relevance for precision immunotherapy targeting different immune microenvironments.

1.5 The implications of cancer heterogeneity on molecular classifications

A series of studies demonstrated both intra-tumor heterogeneity (cellular heterogeneity within individual tumors) and inter-tumor heterogeneity (cancer subtypes) have significant implications on the prognosis, response to therapy and progression with resistances. For

instance, a number of drugs for targeted therapy are conducted for clinical trials in patients diagnosed with metastatic colorectal cancer (mCRC), through stratifying the patients based on distinct intrinsic tumor subtypes of colorectal cancer while integrating the degree of intra-tumor heterogeneity could provide optimal clinical outcomes. Several lines of evidence revealed that intra-tumor heterogeneity, both hierarchical and stochastic, cause various responses to chemoradio-therapy, yet intra-tumor heterogeneity was not examined by generally known biomarker-based approaches. The fate of a drug combination for an individual patient, responsive or resistant, is largely determined by the genetic and epigenetic background, along with its tumor microenvironment in the clonal populations. The spatial distribution and crosstalk between tumor and stromal cells within the tumor microenvironment could dramatically affect their interactions, which consequently impact differentiation, proliferation, morphology and a variety of biological functions during tumor progression. Therefore in order to accurately predict response to chemotherapy, it is rational to take into account this clinical 'global' heterogeneity in terms of genomic and epigenomic features, tumor microenvironment and cellular architectures. Therefore ideally classification of cancer patients should consider inter-individual heterogeneity with "unique tumor principles" including exogenous stimuli and endogenous factors in each patient, where the endogenous factors refer to patients' epigenomic, genomic and metabolic background, and the exogenous factors refer to lifestyle, dietary intake and environmental exposure, which collectively act as external stimuli in signaling pathways and modulate the innate tumor properties.

1.6 Towards personalized therapy for patients with glioma

Glioma describes tumors that arise from the supportive glial tissue of the brain and spinal cord. Depending on the cell of origin, including oligodendrocytes, astrocytes and ependymal cells, glioma is categorized as astrocytoma, oligoastrocytoma (mixed gliomas) and

ependymoma. WHO grades of gliomas are based on their growth rate, with grade 1 showing slowest growth and grade 4 showing aggressive growth). Low-grade gliomas (WHO grade II) refer to those which are highly-differentiated (neither anaplastic nor benign) but exhibit favorable prognosis in general. In contrast, high-grade (WHO grade III-IV) glioma refer to undifferentiated (anaplastic) glioma with dismal prognosis. Grade IV glioma (or glioblastoma, GBM) is the most fatal glioma and the overall prognosis remains dismal, with a median survival time of only 12-15 months.

Combined therapies with surgical resection, temozolomide (TMZ) and radiotherapy have been established as standard treatment and have improved clinical outcomes for patients diagnosed with glioblastoma. Optimal management requires a multidisciplinary approach with both in depth understanding of tumor progression and the mechanisms of actions in the treatment. Thanks to integrated sequencing and bioinformatics strategies, advances in the investigation of epigenomic and genomic alterations and molecular classifications of glioma have provided improved understanding of the genomic features and biological relevance of the disease. (Verhaak, Hoadley et al. 2010, Cancer Genome Atlas Research, Brat et al. 2015, Ceccarelli, Barthel et al. 2016) Molecular biomarker-driven strategies that modulate the function of “actionable molecular targets” have been used in clinical trials over the past several decades. Nevertheless, single “one size fits all” targeted treatment has yielded limited clinical efficacy in glioblastoma to date, due to lack of accurate tumor imaging, complex tumor heterogeneity, and pharmacodynamics / pharmacokinetic failures. For example, MGMT promoter methylation is significantly correlated with TMZ response and clinical benefit, yet other molecular biomarkers associated with therapeutic response remain to be explored in GBM.

Given the phenotypic, genomic and clinical heterogeneity, it is necessary to conduct a biomarker -based selection of therapeutic regimens for GBM patients, and apply whole-exome, transcriptome or gene expression profiles that collected from a patient's specimen to

build predictive models based upon multidimensional profiles, administrate with multiple agents that show better prognosis in prospective clinical trials. And importantly, it is important to collect multiple biopsies of GBM patients from different disease loci upon surgery dissections, including both infiltrative and non-infiltrative regions of tumor, and perform extensive genome-wide profiling and select individualized drug combinations with multiple agents that are predicted to eliminate actionable targets within the residual, diffused areas of malignant lesion. When feasible, matched samples from individual patients upon relapse should be collected to validate the strategy and assess the drug resistance.(Kim, Zheng et al. 2015) In addition, given the increasing options of clinical targets, a more rapid screening process is needed, along with larger patient population to design small, efficient trials to discern early efficacy signals.

First, we should integrate genomic profiling including copy numbers, somatic mutations and transcriptomic profiling to deduce the molecular signatures of GBM, then create simulation-based model to predict rational combinations of FDA-approved targeted agents through incorporating comprehensive disease pathophysiological information with "actionable" genomics data, to generate personalized regimens and validate these responses in ex-vivo testing, which is proved to be an important step towards clinical translation to design rational, precision drug combinations for GBM. For instance, based on the molecular features of GBM from one patient, in-silico simulation could predict the inhibitors of PI3K and mTOR pathways to be effective against this tumor.

1.7 Supervised learning for personalized medicine

1.7.1 Supervised learning methods predicting output values from high-dimensional datasets

Supervised learning processes perform learning the relation between two variables: observed variable x and predicted variable y , usually uses a 1D array of n samples to fit a (x, y) model and given observations X , $\text{predict}(X)$ method returns the predicted y .

The simplest classification and regression rule is k nearest neighbors (KNN), as non-generalizing machine learning methods, given a new observation X , search for the observations in the training set with the closest feature vectors (e.g., Euclidean metric functions). The principle underlying nearest neighbor methods is to obtain an optimized number of training samples close to the new observation, and predict the new label.

Linear regression is most common method that fits a linear model to the training dataset by optimizing the parameters to minimize the sum of squared residuals from the model. When there are few data points available in each dimension (small sample size), high variance will be induced by noise from the observations. One regularized solution is to shrink the regression coefficients to zero to prevent overfitting. When two sets of randomly selected observations are likely to be uncorrelated, ridge regression is used to decrease the contribution of non-informative features and return corresponding non-sparse coefficients. In contrast, Lasso (least absolute shrinkage and selection operator), as a sparse penalization approach, could set some coefficients to zero in regularization path for feature selection purposes.

Support Vector Machines (SVM) belongs to the discriminant model family. The goal is to build a hyperplane that maximize the margin between the two subgroups by selecting a subset of samples. Regularization is fine-tuned by a parameter C , where small value indicates the margin is calculated using all or most observations around the separating line (heavy regularization); whereas large value C indicates the margin is estimated based on those observations approximate to separating line (weak regularization). When two classes could not be separated linearly in feature space, SVM can build a decision function

using polynomial kernel tricks to create decision energy by positioning kernels on the observations.

1.7.2 Survival analysis of high-dimensional covariates from cancer genomics

International organizations such as International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) have produced multiple dimensional analysis at a variety of genome-scales, aiming at identifying novel cancer biomarkers to enable clinical cancer researchers to tailor the therapies and conduct personalized risk predictions through gathering massive amounts of epi/genomic data of cancer patients collected from multiple treatment centers, so called “bio-profiles”. High-dimensional issues arise as the number of genomic covariates from these datasets often far exceeded the sample size. Another challenge for survival analysis and feature selection is censored data are often observed rather than precisely measured time-to-event information. In the past decades, a series of parametric and semi-parametric models were proposed to overcome these two challenges, including regularized Cox-regression models, regularized accelerated failure time models, supervised principal components, partial least squares, etc. Yet the efficiency of these models largely depends on the underlying assumptions. More recently, non-parametric machine learning algorithms became a focus of growing interest to deal with high-dimensional issues. Ensemble based approaches, including boosting and random forests, are the most widely applied, albeit boosting with high dimensional censored data is not fully investigated. It was reported that a gradient boosting procedure fit smoothing splines to estimate proportional hazard models and identify non-linear effects of important variables (genes) from microarray data that are correlated to the risk of any specific event (Li and Luan 2005).

Since classical Cox regression method is used to select single biomarker, without considering strong correlation between those co-expressed genes, the network-based Cox

regression models (such as fastcox, AdaLnet and Net-Cox) were proposed to overcome such drawbacks, and permutation-based algorithms provided favorable validations on the selection of cancer signature genes involved in regulatory pathway/networks. (Iuliano, Occhipinti et al. 2016) Other regularization methods, including lasso and group lasso for sparse estimation were also designed for the sake of accurate prediction. In addition, these regularization algorithms demonstrated better performance when an additional simple procedure (e.g. Cox model) is initially applied to reduce the number of covariates.

1.7.3 Feature selection for high dimensional regression using LASSO (Least Absolute Shrinkage and Selection Operator) algorithms

High-dimensional data mining and feature selection are amongst the most challenging topics in modern statistics. The field of high-dimensional statistics spans a wide range of models including supervised methods in regression and classification models, as well as unsupervised approaches for clustering, multiple testing or graphical models (Buhlmann and van de Geer, 2011). One of the major statistical challenges with high-dimensional data mining is overfitting in regression where the number of variables far exceeds the number of observations (sample size). In these situations standard estimation methods, such as ordinary least square, failed in accurate prediction. Therefore, a wealth of efforts emerging to handle the high-dimensional regression, typically employing penalty based regularization for dimension reduction and/or feature selection.

LASSO, proposed by Tibshirani in 1996, is the most popular method by far. By construction, the lasso not only fits the regression model, it simultaneously performs variable selection by shrinking regression coefficients of unimportant variables to zero. So it is suitable for prediction and model construction, by producing a sparse solution and consequently extracting most important variables. Shrinkage is performed by placing a constraint on the size of the regression coefficients and adding a penalty term to residual

sum of squares, where penalty $J_\lambda(|\beta_j|)$ depends on tuning parameter λ that regulate the extent of the shrinkage, which take on various forms, typically involving $\lambda |\beta_j|^r$, where r refer to different methods, including ridge regression with a penalty $J_\lambda(|\beta_j|) = \lambda |\beta_j|^2$ with $r = 2$, lasso by assigning $r = 1$. All regularization methods depend on one or more tuning parameters controlling the model complexity including the number of variables preselected in subset selection, the number of derived inputs to use in principal components regression or the amount of shrinkage in shrinkage methods. It is critical to determine tuning parameters for model fitting and obtaining right balance between bias and variance to minimize prediction errors.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|y - X\beta\|_2^2 + \sum_{j=1}^P J_\lambda(|\beta_j|) \right\}$$

However, standard lasso regression model may be less efficient when underlying patterns of dataset are more complex. For instance, the effects of the covariates might deviate from linearity, or they might interact with each other or additional measurable quantities outside the data, confounding the prediction. In many situations, or given assumptions arising from relevant (biological) knowledge, the standard lasso might not be adequate to solve high dimensional problems. Consequently the standard lasso has been extended and modified to deal with more complex data structures that are hidden in high-dimensional data sets. The adaptive LASSO, group LASSO and Elastic Net are the most prevalent methods developed in this regard.

The adaptive LASSO (ALASSO) penalty applied a weighted penalty term that was introduced via the least - angle regression (LARS) algorithm; it yields consistent estimations on the parameters while keeping the convexity property of the lasso. The motivation is to favor predictors with univariate strength on regression coefficients and to avoid spurious selection of noise predictors, it uses univariate regression coefficients in place of full least

squares estimates when the number of features is far beyond the sample size. It recovers the correct model under milder conditions than does the lasso. The major advantage of ALASSO is the penalty terms are assigned to regression coefficients adaptively according to the importance of the corresponding covariates and it relies on neither censoring distribution nor baseline survival function. (Dai, Koutrakis et al. 2016, Raeisi Shahraki, Pourahmad et al. 2016)

In some statistical scenarios of regression with categorical predictors, the predictors belong to different pre-defined groups (e.g. genes functioning in the same biological pathway). The group lasso was designed to solve this problem by shrinking and selecting the elements from a group together. Standard group lasso algorithms use coordinate descent, and assume that the design matrix in each group is orthonormal and use simple soft-thresholding. If the size of each group is 1, it is exactly equivalent to the regular lasso solution. While the group lasso generates a sparse set of groups, when a group is included in the model, all coefficients in that group will be nonzero. For predictors with few levels it is reasonable to use the group lasso - sparsity within group is unnecessary as groups are small. As the number of levels per predictor rises, the sparse-group lasso show best performance by setting the coefficients for many levels equal to 0 even in nonzero groups, to reduce the predictors whose levels are less informative. (Liu, Wu et al. 2014, Wang and Xue 2015)

There are common situations with distinct correlations among the variables (e.g. co-expression of genes belonging to the same molecular networks). The lasso penalty works less sensitively to those strong but correlated variables. Ridge algorithms are proposed to shrink the coefficients of correlated variables towards each other. The elastic net penalty is a compromise of both lasso and ridge solutions, in the form where the front term favors a sparse solution for coefficients of averaged features, the back term favors highly correlated variables to be averaged, consequently more than $\min(N, p)$ coefficients can be nonzero.

$$\sum_{j=1}^p (\alpha |\beta_i| + (1 - \alpha) \beta_j^2)$$

1.8 Fusion genes as an emerging target for precision medicine

Genomic instability, structure rearrangement of the genome, in particular translocations, as well as non-structural rearrangement mechanisms (such as transcription read-through of neighboring genes or mRNA trans-splicing or cis-splicing), could lead to the formation of a large proportion of gene fusions, which collectively played important roles in tumor-initiating process. Therefore fusion genes could serve as potential biomarker for both diagnostic tools and therapeutic targets, due to their inherent expression in a subset of tumors. In general, the frequency of gene fusions varies between different cancer types and appears inversely correlated with the frequencies of other somatic mutations at both per cancer types and tumor samples. Kinase, DNA-binding domain and chromatin modifiers are often involved as one of the partners of chimeric genes, with many fusion genes fusing with only one other partner (Yoshihara, Wang et al. 2015). While the distributions of fusion events vary largely amongst different cancer types, delineation of fusion genes and their genomic features in multiple cancer types will provide more precise perspectives for precision therapies for those cancer patients.

The rapid development of next generation sequencing (NGS) accompanied with novel computational models has exponentially facilitated the discovery of fusion genes in solid tumors, which include sarcoma, carcinoma and tumors in central nervous system, besides hematologic malignancies (Parker and Zhang 2013). Knowledge regarding the prevalence and function of gene fusions has been revolutionized, coinciding with the advances of NGS technology, bioinformatics algorithms and large-scale computational biology, though raising the issues in terms of driver oncogenic or passenger events encoded by these predicted fusions for cancer development. Several lines of studies have highlighted the fusion-driven oncogenesis that enriched with certain combinations of functional domains

encoded by fusion transcripts (Ortiz de Mendibil, Vizmanos et al. 2009) (Frenkel-Morgenstern, Lacroix et al. 2012). The genomic hallmarks of oncogenic fusion genes revealed by public databases include the expression changes in fusion transcripts due to untranslated region (UTR)/ promoter swapping, protein-protein interaction interfacing with novel oncogenic functions, as well as other genomic features associated with replication timing, which could serve as the foundation for predicting oncogenic drivers of novel fusion events by referring to their common features with known fusions (Shugay, Ortiz de Mendibil et al. 2012).

It is also important to explore the mechanisms through which fusion transcripts or proteins are precisely targeted for the sake of potential drug development. Prediction of the function of fusion products is non-trivial, through inferring the secondary structure/ domain architectures and regulatory elements of the parent protein and the interaction of the functional domains in chimeric proteins. Based on the structural features of fusion proteins revealed by a series of studies, those proteins form fusions tend to present fewer domains than other proteins, while fusion transcripts encode more domains than expected by chance; fusion proteins are more enriched at specific domains than randomly permuted arrangement would suggest; and increased intrinsic reorder in fusion proteins promotes viable joining of different constituent domains into flexible proximity for internal interactions.

Identification of both common and rare gene fusions has numerous impacts on personalized clinical care. For example, the *TMPRSS2-ERG* fusion transcript functions were established as a urinary biomarker for prediction of localized prostate cancer recurrence after surgery (Leyten, Hessels et al. 2014) and *KIF5B-RET* gene fusions are discovered as a potential target for existing TKIs (vandetanib) in lung cancer (Kohno, Ichikawa et al. 2012). Fusion genes involving kinase gene fusions served as promising therapeutic targets due to their susceptibility to kinase inhibitors (Stransky, Cerami et al. 2014). And fusion genes

harboring histone methyltransferases were discovered as other attractive drug targets (Helin and Dhanak 2013).

Yet there is still room to increase the specificity and sensitivity of fusion gene prediction, as well as functional characterization of these frequently oncogenic mutations, which will play important roles in elucidating disease processes across diverse tumor types. The pivotal advances in targeting therapy against fusion proteins harboring kinases and chromatin modifiers (Chen and Tseng 2014) (Kannan, Coarfa et al. 2015), pave a way for fusion transcripts as promising targets for future therapeutic avenues in genomic era.

1.9 Motivations and Rationale of the studies in this thesis

The sole purpose for cancer genomics is to advance personalized medicine through NGS sequencing and characterize the patient tumors based on genetic alterations associated with featured biological phenotypes, in order to develop more efficient prevention, diagnosis and treatment strategies corresponding to these genetic variables. Such study for future iterations will fuel more efficient designs of clinical trials in which patients are eligible for novel treatments based on their genomic profiling, molecular classification and predicted outcomes.

However the accuracy of applying NGS data for risk prediction and patient stratification has been largely constrained due to many limitations, such as clinical sample detection, filter and annotation of the variants, computational and statistical algorithms, etc. For example, due to molecular heterogeneity and anatomic diversity of cholangiocarcinoma (CCA), the subtypes of cholangiocarcinoma show various epidemiological behaviors and are correlated with inconsistent prognostic risks, various tissue origins, which render the CCA molecular classifications prone to ambiguity, consequently preventing the precise identification of functional mutations, biomarkers, and target therapies in basic research and clinical setting. Since CCA often develops an inflammatory milieu including cirrhosis and

cholangitis, the tumor microenvironments (TME) are likely to promote the progression of the malignancy, contributing to phenotypical heterogeneity in terms of genomic alterations, cellular morphology and resistance to therapy (Raggi, Invernizzi et al. 2015).

A wealth of studies have been carried out to elucidate the mechanisms for radiation resistance of GBM, some studies reported glioma cells undergoing Epithelial-Mesenchymal Transition (EMT) involving in GBM recurrence (Kubelt, Hattermann et al. 2015), while other studies reported only a subset of proneural patient-derived glioma sphere cultures (GSCs) differentiated to MES state mediated by TNF- α /NF- κ B, accompanied with CD44 subpopulations enrichment and radio resistant phenotypes, while some GBM cases present constitutive MES signatures upon removing from the microenvironment, implying cell intrinsic mechanisms could also sustain MES network (Bhat, Balasubramanian et al. 2013). These observations invoked different hypothesis, whether there is co-existence of proneural and mesenchymal GSC within individual tumor, and Mes-like radiation-resistant cells preferentially survive and emerge as dominant population, or intrinsic transition from proneural to mesenchymal occur upon GBM recurrence. And IHC analysis revealed that proneural and mesenchymal markers co-exist in the same tumor region, indicating transcriptomic plasticity in GBM samples. And MES GBMs exhibit high content of necrosis (Cooper et al., 2012) and macrophages/microglial infiltration (Engler et al., 2012; Li et al., 2012). Given the impact of tumor microenvironment on molecular classification and contradictory interpretation of clinical outcomes in each distinct tumor subtype, it is important to systematically investigate the confounder factors associated with ambiguities of molecular classification established in previous TCGA network study of GBM (Brennan, Verhaak et al. 2013), so I have deciphered the effects of tumor microenvironment on transcriptomic dynamics of GBM and revealed the tumor associated microenvironment mimic mesenchymal phenotype of GBM, which is described in **Chapter 2**.

Integrative analysis on around 600 glioma patients from TCGA and several lines of other studies revealed significant prolonged survival in glioma patients harboring 1p/19q codeletion, with median survival for ~115 months, while we observed that distinct variations in these patients with relatively favorable outcome also present in terms of clinical behaviors including timing of tumor progression, response to therapy and consequent survival. (Cancer Genome Atlas Research, Brat et al. 2015) Approximately 85% of the gliomas with 1p/19q codeletions in the TCGA cohort are comprised of oligodendroglial components, which could be divided into low grade (grade II) and high grade (grade III) glioma based on WHO classification criteria, yet we also found that the clinical outcome doesn't show significant concordance with grade, suggesting other factors might determine the clinical status. Moreover a series of clinical trials reviewed that combined chemotherapy with procarbazine, lomustine (CCNU), and vincristine (PCV) following standard radiation therapy delayed glioma development and favorable clinical outcomes in patients diagnosed of anaplastic oligodendroglioma, and additional chemotherapy only exhibit benefits to those patients harboring 1p/19q co-deletion, yet the advantage to adjuvant chemotherapy only observed in part but not all amongst these patient cohort (Buckner, Gesme et al. 2003). In addition, an independent study indicated that some patients diagnosed with glioblastoma harboring 1p/19q co-deletion do not show improved survival outcomes (Boots-Sprenger, Sijben et al. 2013). As a matter of fact, contemporary management of low-grade glioma is still controversial, including the necessary components of the diagnostic strategies, the role of “wait-and-see” criteria, the nature of surgical intervention and radio/chemo therapeutic regimens (Zadeh, Khan et al. 2015). Although putative molecular markers, such as IDH mutant, TP53 mutation, MGMT methylation were applied for the prediction of LGG, there are further biological signatures that could distinguish glioma with 1p/19q codeletion into subgroups, which will impact the risk of malignant progression and clinical outcome, in order to discriminate low-risk patients within glioma harboring 1p/19q co-deletion for whom intensive

adjuvant chemotherapy might be ignored, I have developed the gene signature applying machine learning algorithms and evaluated its predictive performance in independent datasets, which is presented in **Chapter 3**.

Though gene fusions have been proved to serve as important drivers of cancer progression and target therapy, our understanding of the prevalence and associated genomic features for gene fusions in different cancer types is still insufficient, since the recurrence rate for fusion events are relatively lower than that for somatic mutations, fusion genes are not fully discovered due to limited sample size, bias due to RNAseq library and diverse computational frameworks, and fusion transcripts in rare cancer types are less appreciated in the past decades. Therefore taking advantage of massive RNA-Seq data generated from TCGA project, I performed fusion transcripts detection across 33 TCGA cancer types using the PRADA pipeline and discovered numerous novel gene fusions as well as known fusion events in never reported cancer types. And I have explored the associated genomic features with these fusion genes and predicted their oncogenic properties based on their context in protein networks, functional domains. My study will allow the clinicians and biologists to further explore these gene fusions datasets with a few mouse clicks and get the world's most comprehensive gene fusions from major cancer types, fusion frequencies across each cancer type, fusion association with genomic rearrangement as well as fusion mRNA expression level and predictive protein domains in the fusion transcripts leading activation or inactivation of the coding products, elucidating possible causal factors of the cancer, decode the pathways that are perturbed by different fusions, and predict of potential targets for drug development and patient stratifications, as discussed in **Chapter 4**.

CHAPTER 2

Tumor microenvironment associated with transcriptomic subtype plasticity during glioma evolution

(The methods and results in this chapter have been published in biorxiv, Qianghu Wang*, Xin Hu*, Baoli Hu, Florian Muller, Hoon Kim, Massimo Squatrito, Tom Mikkelsen, Lisa Scarpace, Floris Barthel, Yu-Hsi Lin, Nikunj Satani, Emmanuel Martinez-Ledesma, Edward Chang, Adriana Olar, Guocan Wang, Ana C. deCarvalho, Eskil Eskilsson, Siyuan Zheng, Amy B. Heimberger, Erik P. Sulman, Do-Hyun Nam, Roel G.W. Verhaak, "Tumor evolution of glioma intrinsic gene expression subtype associates with immunological changes in the microenvironment". According to the journal policy, the author retains the right to include the submitted and published article in full or part in a dissertation.) *Co-first authors

2.1 Introduction

The intrinsic capacity of glioblastoma (GBM) tumor cells to infiltrate normal brain impedes surgical eradication and predictably results in high rates of early recurrence. To better understand determinants of GBM tumor evolution and treatment resistance, The Cancer Genome Atlas Consortium (TCGA) performed high dimensional profiling and molecular classification of nearly 600 GBM tumors (Cancer Genome Atlas Research 2008, Nounshmehr, Weisenberger et al. 2010, Verhaak, Hoadley et al. 2010, Brennan, Verhaak et al. 2013, Ceccarelli, Barthel et al. 2016). In addition to revealing common mutations in genes such as *TP53*, *EGFR*, *IDH1*, and *PTEN*, as well as the frequent and concurrent presence of abnormalities in the RB, p53 and receptor tyrosine kinase pathways. Unsupervised transcriptome analysis identified four clusters, referred to as classical, mesenchymal, neural and proneural, that were tightly associated with genomic abnormalities (Verhaak, Lintsen et

al. 2010). The proneural and the mesenchymal expression subtypes have been most consistently described in literature with proneural relating to a more favorable outcome and mesenchymal to unfavorable survival (Phillips, Kharbanda et al. 2006, Huse, Phillips et al. 2011, Zheng, Chheda et al. 2012), but these findings were affected by the relatively favorable outcome of IDH-mutant glioblastoma which are consistently classified as proneural (Noushmehr, Weisenberger et al. 2010, Verhaak, Lintsen et al. 2010). Proneural to mesenchymal switching upon disease recurrence has been described as a source for treatment resistance in GBM relapse (Bao, Wu et al. 2006, Phillips, Kharbanda et al. 2006, Bhat, Balasubramanian et al. 2013, Ozawa, Riester et al. 2014), but the relevance of this phenomenon in glioma progress remains ambiguous.

GBM tumor cells along with the tumor microenvironment create a complex milieu that ultimately promotes tumor cell plasticity and disease progression (Olar and Aldape 2014). The presence of tumor-associated stroma results in a mesenchymal tumor gene signature and poor prognosis in colon cancers. (Isella, Terrasi et al. 2015) Furthermore, the association between a mesenchymal gene expression signature and reduced tumor purity has been identified as a common theme across cancer (Yoshihara, Shahmoradgoli et al. 2013, Martinez, Yoshihara et al. 2015). Tumor-associated macrophages/microglia in GBM have been proposed as regulators of proneural-to-mesenchymal transition through NF- κ B activation (Bhat, Balasubramanian et al. 2013) and may provide growth factor mediated proliferative signals, which could be therapeutically targeted (Pyonteck, Akkari et al. 2013, Patel, Tirosh et al. 2014, Yan, Kong et al. 2015).

In this study we explored the properties of the microenvironment in different GBM gene expression subtypes and characterized the transition between molecular subtypes before and after therapeutic intervention. In doing so, we improved the robustness of gene expression subtype classification through revised gene signatures and proposed analytical methodology. Our results suggested that the tumor microenvironment interferes with

expression based classification of GBM, both at the primary disease stage as well as at disease recurrence, and suggest a role for the macrophage/microglia in treatment response.

2.2 Methods

2.2.1 Data sources for multiplatform classification comparison

U133A array profiles for 543 primary GBM, and RNA-Seq data for 166 primary and 13 recurrent GBM (part of GBM samples were measured by both U133A array and RNA-Seq) were obtained from TCGA portal <https://tcga-data.nci.nih.gov/tcga/>. Mutation calls and DNA copy number profiles were obtained for all samples, where available. All raw data from non-TCGA resources (either microarray profiles or RNA-Seq) were retrieved from GEO. Processed primary/recurrence expression data was analyzed using GlioVis <http://recur.bioinfo.cnio.es/>.

Tissues from 20 initial GBM and matched recurrent tumors were obtained from Henry Ford Hospital (n = 9) in accordance with institutional policies and all patients provided written consent, with approval from the Institutional Review Boards (Henry Ford Hospital IRB protocol #402). All RNA samples tested were obtained from frozen specimens. All of the recurrent GBMs had been previously treated with chemotherapy and radiation. Three cases were diagnosed with lower grade astrocytoma prior to primary GBM (HF-2869/HF-3081/HF-3162). Tumors were selected solely on the basis of availability. RNA-Seq libraries were generated using RNA Truseq reagents (Illumina, San Diego, CA, USA) and paired-end sequenced using standard Illumina protocols. Read length was 76 base pairs for cases sequenced by TCGA and from Henry Ford hospital. RNA-Seq data on frozen tissue from 44 patients with initial and recurrent GBM that received resection at Samsung Medical Center and Seoul National University Hospital were provided by Dr. Nam's lab. Surgery specimens were obtained in accordance to the Institutional Review Board (IRB) of the Samsung Medical Center (No. 2010-04-004) and Seoul National University Hospital. (No. C-1404-056-

572) Affymetrix CEL files of 39 pairs of initial and recurrent glioma were retrieved from the Gene Expression Omnibus (GEO accession GSE4271, GSE42670, GSE62153) (Phillips, Kharbanda et al. 2006, Joo, Kim et al. 2013, Kwon, Kang et al. 2015). The expression profiles of the 23 pairs from GSE4271 were determined using Affymetrix HG-U133 GeneChips, the 1 pairs from GSE42670 were analyzed using the Affymetrix HuGene-1_0-st platform, the 15 pairs from GSE62153 were analyzed using Illumina Human HT-12 V4.0 expression BeadChip. The RNA sequencing data of 14 and 5 pairs of primary and recurrent low grade glioma were from TCGA LGG cohort and EGAS00001001255 (Mazor, Pankov et al. 2015), respectively. Genome wide DNA copy number profiling and exome sequencing on thirteen TCGA tumor pairs and nine of ten Henry Ford tumor pairs were performed and data was analyzed using standard protocols and pipelines as previously described (Kim, Zheng et al. 2015).

RNA sequencing data was available for 162 primary GBMs (Brennan, Verhaak et al. 2013) for which an Affymetrix HT-U133A gene expression profile was also available. We observed a low Pearson Correlation Coefficient (< 0.15) between RNA sequencing based reads per kilo base of transcript per million reads (RPKM) and Affymetrix HT-U133A profiles in eighteen cases and these were removed from further analysis. In summary, in order to assess the concordance between classification results of the new 70-gene signatures and previously published 210-gene signatures (Verhaak, Hoadley et al. 2010), 144 GBMs which were profiled in both RNA sequencing and Affymetrix U133A platforms were used in our further analyses.

2.2.2 Transcriptome data processing

The latest version custom CDF files (Version19, <http://brainarray.mbni.med.umich.edu>) (Dai, Wang et al. 2005, Sandberg and Larsson 2007) were used to map probes from the Affymetrix HG-U133A and HuGene-1_0-st GeneChip platforms to the Ensemble transcript

database, combined in one probe set per gene and normalized using the AROMA package with default parameters, resulting in RMA normalized and log transformed gene expression values (Bengtsson, Ray et al. 2009). All RNA sequencing data was processed by the PRADA pipeline (Torres-Garcia, Zheng et al. 2014). Briefly, reads were aligned using BWA against the genome and transcriptome. After initial mapping, the aligned reads were filtered out if their best placements are only mapped to unique genomic coordinates. Then quality scores are recalibrated using the Genome Analysis Toolkit (GATK), and duplicate reads are flagged using Picard. Mapped features were quantified and normalized per kilobase of transcript per million reads (RPKM) and were converted to a log2 scale to represent a gene expression level. RPKM values measuring the same gene that mapped to the Ensemble transcript with longest size were selected to obtain one expression value per gene and sample. The statistical environment R was used to perform all the statistical analysis and graph plots.

2.2.3 Identification of gene signatures for refined GBM subtype

A pair-wise gene expression analysis identified 5,334 genes which are significant higher expressed in glioma bulk samples compared to their derivative glioma stem cells (GSCs). These genes were excluded from the gene list for developing tumor-specific molecular subtypes. Consensus non-negative matrix factorization (CNMF) clustering method identified three distinct subgroups among the 369 IDH-wt primary GBMs. A set of 270 GBMs was selected as core samples based on a positive silhouette width. The gene expression values of each subtype were compared with those from the other two subtypes combined (Verhaak, Hoadley et al. 2010). Signature genes per cluster were selected based on the differences in gene expression level and were considered significant if they reached the cut-off value with t-test $p\text{-value} < 1E-3$ for higher expressed in this class, while also showing a significant lower expression with t-test $p\text{-value} < 1E-3$ in the other two classes. In the original gene signatures,

both down and up-regulated genes were included, while only up-regulated genes (n=70 per gene signature) were selected for revised gene signatures. Only genes measured on both RNAseq and U133A platforms were included, and the U133A data from 162 GBM samples measured on both platforms (which included the 144 cases used to compare U133A and RNAseq platform) was used in the final comparative analysis.

2.2.4 Molecular classifications based on ssGSEA enrichment scores

Single sample gene set enrichment analysis was performed as follows. For a given GBM sample, gene expression values were rank-normalized and ordered based on their ranks. The Empirical Cumulative Distribution Functions (ECDF) of the signature genes and the remaining genes were calculated. Then the statistic was deduced by integration of the difference between the ECDFs, the method is similar to GSEA but based on absolute expression value rather than differential expression (Barbie, Tamayo et al. 2009). Since the ssGSEA test is based on the ranking of genes by expression level, the uncentered and log-transformed U133A and RPKM expression levels were used as input for ssGSEA. Since the scores of the three signatures were not directly comparable, we performed a resampling procedure to generate null distributions for each of three subtypes. First we generated a virtual sample matrix $V(s, g)$ (numbers of permutation, $N_s > 1,000,000$) to simulate the gene expression by randomly selecting an expression value of the same gene (g_j) in the remainder of the samples ($s_1, s_2 \dots s_n \dots s_{i-1}$). Then ssGSEA scores of three signatures in each sample (s) were calculated to generate a large number ($> 1,000,000$) of random ssGSEA scores for each subtype, and build the null distribution of ssGSEA scores for simulated samples, from which we derived empirical p-values (the numbers of ssGSEA scores in simulated samples higher than the one in original sample in each subtype adjusted by the number of permutations) corresponding to the raw ssGSEA scores for each sample. By testing on multiple datasets with different sample sizes, we found the resampling

generated distribution could be replaced with student-t distribution (sample size > 30) or normal distribution (sample size > 50) to obtain similar results.

2.2.5 Evaluate the heterogeneity of GBM subtype

For each sample, first we reversely rank the empirical p-values for each subtype to generate ordered statistics as $R_{N-1}, R_{N-2} \dots R_1, R_0$. In particular, R_0 equals to the minimum empirical p-value and corresponding to the dominant subtype (top activated subtype). The accumulative distance to the dominant subtype (ADDS) was defined as:

$$ADDS = \sum_{i=1}^{N-1} (R_i - R_0)$$

Similarly, the accumulative distance between non-dominant subtypes (ADNS) as:

$$ADNS = \sum_{j>i>0} (R_j - R_i)$$

Thus the values of ADDS and ADNS are positively and negatively correlated with single activated subtype, respectively. Hence, we defined the simplicity score by combining ADDS and ADNS together and adjusted with a constant $\frac{(R_{N-1}-R_0)}{N-1}$ as follows:

$$Simplicity\ score = [ADDS - ADNS] \times \frac{(R_{N-1} - R_0)}{N - 1}$$

2.2.6 Tumor purity assessment

The ESTIMATE package was used to evaluate tumor purity on the basis of the expression level of marker genes in stromal and immune cells (Yoshihara, Shahmoradgoli et al. 2013), where the fraction of stromal cells and immune cells in each sample were represented by stromal score and immune score respectively, and the mixed fraction of both stromal and immune cells was represented by estimate scores. The ABSOLUTE package was used to confirm the tumor purity on the basis of chromosome copy number and allele fraction ratios on samples for which SNP array data were available (Carter, Cibulskis et al. 2012).

2.2.7 Establish glioma neurosphere cultures (GSCs)

Upon approval from the institutional review board of The University of Texas M.D. Anderson Cancer Center, glioblastoma tumor tissues were collected and labeled in the order that they were acquired. Each tissue was enzymatically and mechanically dissociated into single cells and grown in DMEM/F12 media supplemented with B27 (Invitrogen), EGF (20 ng/ml), and bFGF (20 ng/ml), resulting in neurosphere growth. All cell lines were tested to ensure mycoplasma infection negative. To minimize any batch effect the downstream molecular analyses were performed on identical cell culture batches. Total RNA from formalin fixed, paraffin embedded tumor tissues and matching neurospheres was prepared using the Masterpure complete DNA and RNA isolation kit (Epicenter) after proteinase K digestion per to the instructions from the manufacturer. Paired-end Illumina HiSeq sequencing assays were performed resulting in a medium number of 50 million 75bp paired end reads per sample. We employed the PRADA pipeline to process the RNA sequencing data (Torres-Garcia, Zheng et al. 2014). Briefly, Samtools, Burroughs-Wheeler alignment and Genome Analysis Toolkit (GATK) were used to align short reads to the human genome (hg19) and transcriptome (Ensembl 64). RPKM gene expression values were generated for 135,994 transcripts consist of 21,165 protein coding genes in Ensembl database.

2.2.8 Western blotting

Lysates were prepared from fresh frozen sections using RPPA lysis buffer (1% Triton X-100 50mM HEPES pH 7.4, 150mM NaCl, 1.5mM MgCl₂, 1mM EGTA, 100mM NaF, 10mM Na pyrophosphate, 1mM Na₃VO₄, 10% glycerol, plus protease and phosphatase inhibitors cocktails from Roche Applied Science #05056489001 and 04906837001), with sonication and clearing by centrifugation at 10,000g. Protein concentration was measured using the BCA kit (Thermo Scientific -Pierce #23225). SDS-PAGE and western blotting was performed using Midi gel system (Life Technologies - #WR0100) and NuPage-Novex 4-12% Bis-Tris

Midi (20-well) Protein Gels (Life Technologies - #WG1402) using the following antibodies: ITGAM (CD11B) (Sigma Aldrich – #HPA002274), IBA1 (AIF1) (Sigma Aldrich – #HPA049234), GFAP (Cell Signalling – #3670), YKL40 (CHI3L1) YKL40 (CHI3L1, Santa Cruz Biotechnology - #sc-30465) a-actinin (Sigma Aldrich A5044) and Tubulin (Sigma Aldrich T9026).

2.2.9 Immunohistochemistry

Formalin-fixed, paraffin-embedded tissue sections (4 µm thick) were produced on superfrost plus slides. Briefly, tissue sections were deparaffinized with xylene and ethanol and rehydrated with 95, 70 and 50% ethanol. Sections were antigen unmasked using citrate buffer (Vector Labs #H-3300) and heating. Peroxidase block was conducted with 3% H₂O₂ and blocking was with 5% goat serum (Vector Labs #S-1000). Primary rabbit polyclonal antibody against IBA1 (AIF1) (WAKO #016-20001) at 1:400 was used overnight. Secondary antibody was done using with the Rabbit-on-Rodent HRP-Polymer (Biocare #RMR622L) for 1 hr at room temperature. The slides were developed with Nova-red (Vector Labs #SK-4800) and counterstained with haematoxylin, mounted and scanned with Panoramic 250 slide scanner (Caliper Life Sciences). Unbiased quantification of microglial (IBA1+) percentage in primary and recurrent GBMs was performed using the Caliper Vectra image system and InForm analysis software. Thirty scan fields were automatically selected on from entire tumor section. Nineteen scan fields were select from the primary tumor of patient #2 due to the small size of tumor section. Percentages of the median and high levels (2+, 3+) of IBA1 were used for the comparison.

2.3 Results

2.3.1 Harnessing glioma sphere-forming cells identifies GBM specific inter-tumoral transcriptional heterogeneity

We set out to elucidate the tumor-intrinsic and tumor associated microenvironment independent transcriptional heterogeneity of GBMs. We performed a pairwise gene expression comparison of independent set of GBMs and the derivative glioma sphere-forming cells (GSCs) (n = 37) (Galli, Binda et al. 2004). In total, 5,334 genes were found to be significantly higher expressed in parental GBMs relative to derived GSCs that could be attributed by the tumor associated GBM microenvironment (**Figure 2.1A**). To focus the analysis on the tumor-intrinsic transcriptome, these genes were filtered from further analysis, given that this step may additionally exclude genes exogenously expressed in activated glioma cells. GBMs with IDH mutations have distinct biological properties and favorable clinical outcomes compared to IDH wild-type GBMs (Noushmehr, Weisenberger et al. 2010, Brennan, Verhaak et al. 2013, Cancer Genome Atlas Research, Brat et al. 2015, Ceccarelli, Barthel et al. 2016). Using the filtered gene set, we performed consensus non-negative matrix factorization clustering (CNMF) to identify three distinct subgroups amongst 369 GBM samples with IDH wild type, then we further selected 270 core samples with positive silhouette width of each cluster (CL: n=97, MES: n=94, PN: n=79) (**Figure 2.1B; Figure 2.1C**). When comparing the clustering result with the previously defined proneural (PN), neural (NE), classical (CL) and mesenchymal (MES) classification (Verhaak, Hoadley et al. 2010, Brennan, Verhaak et al. 2013), three subgroups were distinctly enriched for CL, MES and PN GBMs, respectively (**Figure 2.2**). Consequently, we labeled the groups as CL, MES and PN. None of the three subgroups was enriched for the NE class, suggesting the neural phenotype is non-tumor specific. The NE group has previously been related to the tumor margin where normal neural tissue is more likely to be present (Sturm, Witt et al. 2012, Gill, Pisapia et al. 2014) and such contamination might explain why the neural subtype was the only subtype lack of distinct gene abnormalities (Brennan, Verhaak et al. 2013, Verhaak, Tamayo et al. 2013).

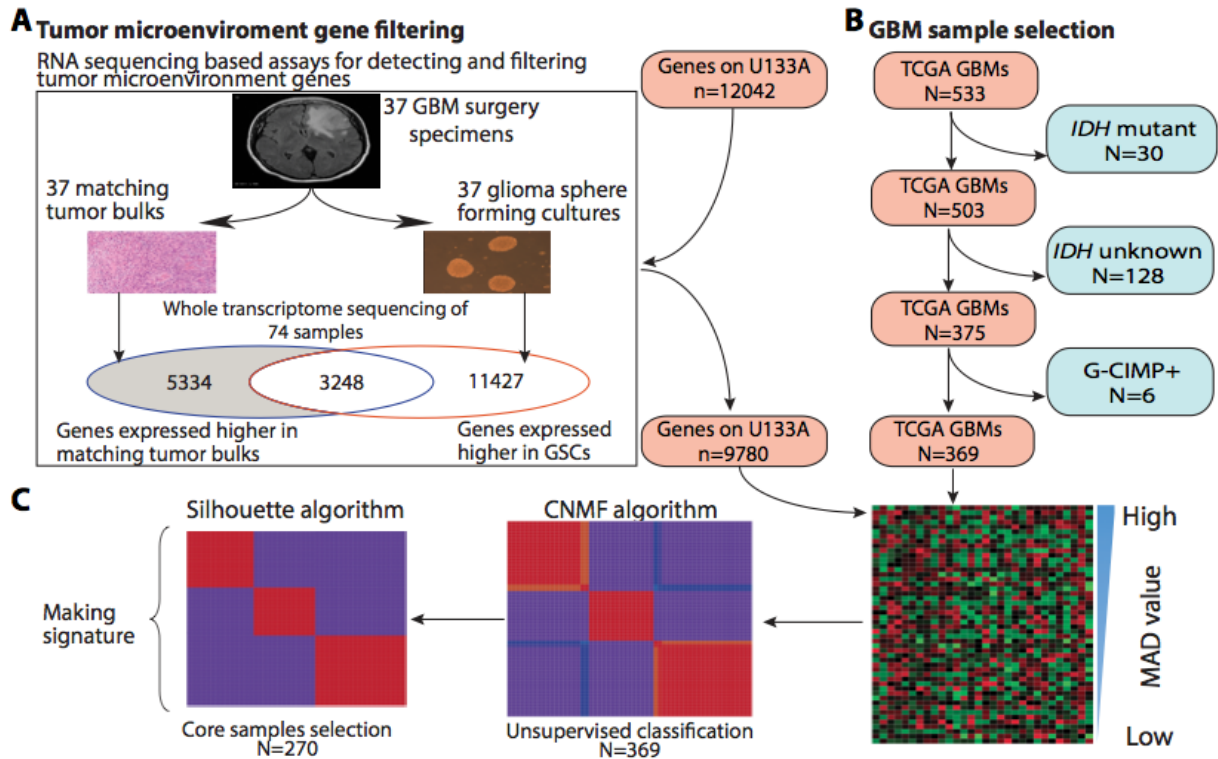


Figure 2.1 Selection of gene signature for molecular classification of IDH-WT GBMs

(A) Exclude genes specifically expressed in tumor associated microenvironment.

(B) Exclude GBM samples associated with IDH mutation/ G-CIMP+.

(C) Selection of signature genes based on NMF clustering.

[* Courtesy of Erik Sulman, Qianghu Wang]

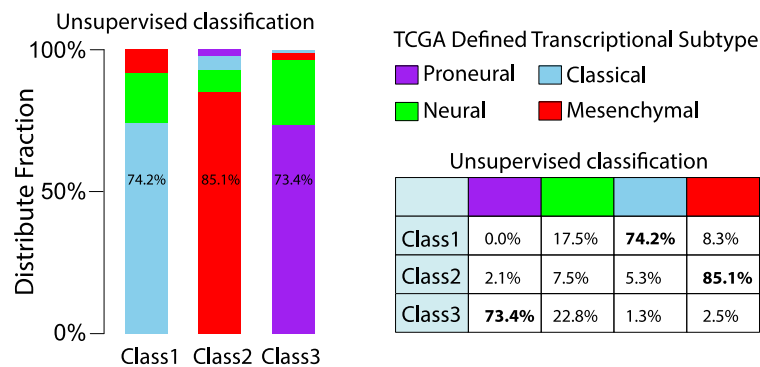


Figure 2.2 Comparison between GCIMP- GBM specific classification and previously TCGA defined GBM subtypes

New assignment Class1: Classical; Class2: Mesenchymal; Class3: Proneural

In order to classify external GBM samples, we implemented a single sample gene set enrichment analysis (ssGSEA) based equivalent distribution resampling classification

strategy using 70-gene signatures with significantly differential gene expression (p-value< 1E-3, t-test) and up-regulated for each subgroup (**Table 2.1**)(**Figure 2.3**), to assign each sample with three empirical classification p-values derived from ssGSEA on which we determined the significantly activated subtype(s) in each GBM sample.

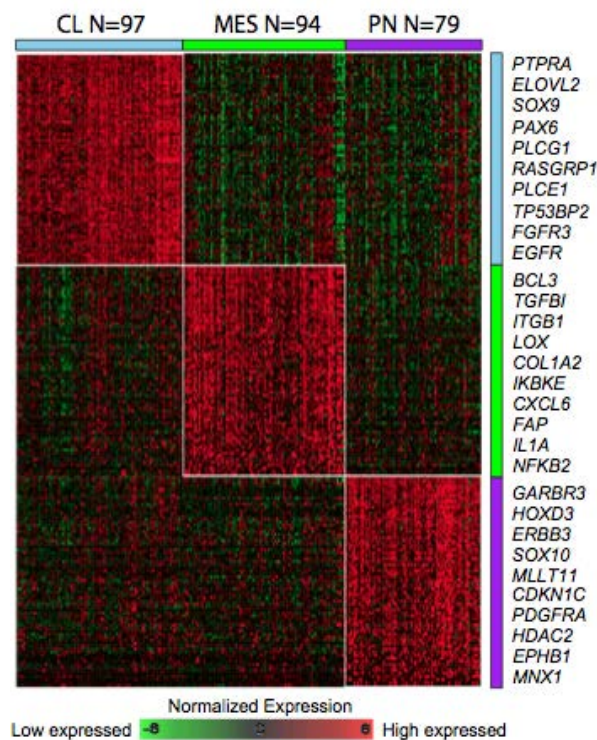


Figure 2.3 Gene signatures applied for GBM classification

Heatmap of 70-gene signatures by gene expression subtype was developed based on 270 core samples of GBMs. Selected ten genes are listed for each subtype.

Table 2.1 Gene signatures applied for transcriptomic classification of GBM

Gene expression value was normalized cross samples.

Class1 (Classical)						
GeneSymbol	avg_expr_class1	avg_expr_class2	avg_expr_class3	ttest p-val (expressed higher in class1)	ttest p-val (expressed lower in class2)	ttest p-val (expressed lower in class3)
PTPRA	0.9357816	-0.7061414	-0.3800021	3.85E-48	9.88E-17	6.21E-06
ELOVL2	0.9705643	-0.5175024	-0.6306896	3.13E-44	2.14E-11	1.82E-12
SOX9	0.7932283	-0.699071	-0.4061383	2.56E-40	3.83E-14	4.90E-04
MLC1	0.8189504	-0.4340693	-0.669476	2.54E-39	1.37E-06	1.36E-12
CENTD1	0.8029843	-0.3346061	-0.8082492	3.53E-39	4.08E-04	2.19E-13
PAX6	0.7994163	-0.6163521	-0.5406433	1.10E-38	7.82E-12	4.75E-07
ARNTL	0.8436674	-0.4300888	-0.7367687	4.35E-38	1.41E-07	3.18E-12
BBS1	0.7582391	-0.4954904	-0.5619164	1.44E-36	2.26E-07	1.79E-08
DENND2A	0.846962	-0.6578756	-0.44469	1.71E-36	3.66E-14	2.47E-06
SGEF	1.0029159	-0.5439714	-0.6099211	3.23E-35	4.42E-13	3.51E-15
PLCG1	0.917316	-0.5059301	-0.4578873	3.63E-35	6.90E-12	2.10E-07
VAV3	0.9561432	-0.4396791	-0.6426364	3.64E-35	7.71E-10	7.02E-14
ZHX3	0.9076196	-0.5415373	-0.4858609	3.97E-35	3.60E-12	5.59E-08
RASGRP1	0.8597176	-0.4858334	-0.6945396	4.15E-35	5.75E-08	7.46E-13
BBOX1	0.7284083	-0.3182443	-0.7069998	7.72E-35	8.66E-04	1.04E-12
EYA2	0.9021703	-0.5056218	-0.5358631	6.59E-34	9.14E-12	4.46E-10
ZC3H14	0.8685518	-0.5453511	-0.3702896	1.02E-33	3.32E-12	3.16E-05
C14orf159	0.8682383	-0.310451	-0.7534334	2.18E-33	6.34E-05	6.80E-15
ACSL3	0.9163126	-0.3970603	-0.6249772	2.19E-33	1.94E-07	1.24E-12
LHFP	0.7887958	-0.4455733	-0.5433495	2.80E-33	1.25E-07	1.43E-08
MYO6	0.8569401	-0.635871	-0.3659682	3.26E-33	9.54E-13	1.99E-05
NCOA1	0.7781773	-0.5789235	-0.3887896	3.43E-33	3.65E-11	5.25E-05
CDH4	0.943382	-0.6625805	-0.4733951	3.98E-33	1.51E-19	1.03E-07
PLCE1	0.8614252	-0.5801901	-0.4635842	4.45E-33	9.43E-14	1.56E-07
USP8	0.787354	-0.4360011	-0.5225909	5.61E-33	2.03E-07	1.01E-07
METTL8	0.8718112	-0.5698252	-0.3853115	5.87E-33	4.41E-14	1.75E-05
ACSBG1	0.831067	-0.5571798	-0.4996926	6.68E-33	7.85E-11	5.98E-08
TP53BP2	0.8436473	-0.6831884	-0.394815	7.63E-33	1.04E-14	1.14E-04
FGFR3	0.9450108	-0.681269	-0.5723826	2.78E-32	3.46E-17	1.92E-10
SLC20A2	0.8564888	-0.4413945	-0.6730649	9.80E-32	3.52E-07	5.18E-13
CST3	0.7416099	-0.3550387	-0.5378808	1.87E-31	5.24E-06	2.45E-09
ZFXH4	0.8208435	-0.4297797	-0.5015759	2.02E-31	2.15E-08	1.34E-07
ZNFX5	0.8680729	-0.5274126	-0.457103	4.90E-31	1.21E-11	1.57E-07
DTNA	0.7778008	-0.5534839	-0.423241	7.48E-31	1.32E-09	3.76E-06
SEPT11	0.8448154	-0.4932475	-0.488272	7.96E-31	5.59E-10	1.26E-08
TJP1	0.8534586	-0.6236913	-0.5109239	1.00E-30	1.68E-13	2.91E-07
MEOX2	0.7852753	-0.385538	-0.6244527	1.10E-30	5.39E-06	1.00E-08
ZNFX211	0.8111843	-0.593195	-0.3432982	1.69E-30	6.30E-13	1.17E-04
SALL1	0.7407417	-0.4442233	-0.4652971	3.30E-30	4.52E-08	6.44E-06
UPF1	0.876064	-0.4072911	-0.5207814	7.41E-30	2.81E-08	5.36E-10
STXBP3	0.7742647	-0.3384675	-0.506417	7.66E-30	5.21E-06	1.44E-08
MYO5C	0.7732981	-0.4271908	-0.733887	8.39E-30	5.18E-06	5.08E-15
MOSC2	0.8820548	-0.57641	-0.512501	9.11E-30	1.05E-12	1.33E-07
KIAA0329	0.8142222	-0.6286991	-0.3280085	1.17E-29	1.51E-15	3.92E-04
KIAA0355	0.7590722	-0.5633495	-0.3193907	2.01E-29	3.32E-11	2.72E-04
SUOX	0.8149632	-0.5391405	-0.359782	2.44E-29	5.22E-12	2.26E-05
EGFR	0.8131495	-0.5633015	-0.5273916	2.50E-29	7.09E-12	1.68E-07
PPARGC1A	0.9405416	-0.6799903	-0.3912386	2.62E-29	2.16E-20	9.99E-06
SLC4A4	0.8387492	-0.4173314	-0.5811297	2.67E-29	1.20E-06	7.43E-12
POLRMT	0.8255901	-0.545333	-0.3998956	2.75E-29	9.95E-12	1.29E-05
SPRY2	0.7180249	-0.3427663	-0.5881643	3.05E-29	9.69E-05	1.33E-08
GRIK1	0.9074649	-0.5670344	-0.4500018	7.27E-29	4.23E-14	9.07E-08
RBCK1	0.8423863	-0.4240406	-0.5863766	8.94E-29	4.10E-07	1.07E-10
LPIN2	0.8258538	-0.4842192	-0.4697258	9.86E-29	5.10E-09	5.30E-08
C5orf4	0.8371379	-0.4955205	-0.5708525	1.28E-28	3.26E-09	1.89E-10
PNPLA6	0.7089793	-0.3370047	-0.5753727	1.53E-28	1.47E-04	3.86E-11
NPEPL1	0.8436841	-0.3546394	-0.7383391	1.67E-28	1.66E-05	3.15E-17
ST5	0.6797893	-0.3450434	-0.6571912	2.02E-28	3.66E-04	2.68E-12
BCKDHB	0.7946407	-0.645992	-0.3460715	2.34E-28	4.51E-15	2.44E-04
PHKB	0.777893	-0.4398621	-0.3774378	2.63E-28	4.50E-09	4.21E-06
CAMK2B	0.7957804	-0.6046373	-0.3502266	6.52E-28	1.53E-14	1.52E-04
BAG5	0.7576005	-0.3494488	-0.414406	6.90E-28	1.47E-06	6.82E-07
SCAMP4	0.748153	-0.4195556	-0.5457961	9.20E-28	7.91E-07	1.33E-09
SLC3A2	0.7798822	-0.339521	-0.5563424	1.12E-27	1.71E-05	3.43E-10
MAP4	0.6873813	-0.3207317	-0.5186339	1.32E-27	1.17E-04	6.69E-10
SSFA2	0.7708881	-0.320791	-0.6220349	1.67E-27	5.55E-05	1.56E-12
TMEM131	0.7284812	-0.3375683	-0.4463644	2.14E-27	1.47E-05	3.80E-07
PTPN11	0.7767987	-0.5541996	-0.3014475	2.68E-27	5.44E-12	5.00E-04
VAPB	0.846061	-0.5545247	-0.3324442	3.74E-27	3.50E-13	1.47E-04
SLTM	0.715334	-0.4163261	-0.4405037	4.50E-27	5.81E-07	1.31E-05

Class2 (Mesenchymal)						
GeneSymbol	avg_expr_class1	avg_expr_class2	avg_expr_class3	ttest p-val (expressed higher in class2)	ttest p-val (expressed lower in class1)	ttest p-val (expressed lower in class3)
S100A11	-0.2840709	0.8732734	-0.6027328	1.04E-31	1.99E-05	3.29E-12
ARPC1B	-0.262509	0.9546035	-0.7071391	4.41E-31	2.84E-05	2.56E-20
CTSC	-0.5056449	0.9520953	-0.3093346	4.38E-30	1.82E-15	2.14E-06
NPC2	-0.2350836	0.7957376	-0.6517269	1.19E-29	5.45E-04	1.56E-13
GLIPR1	-0.2727147	0.8923129	-0.5656891	2.11E-28	7.37E-06	3.38E-13
VDR	-0.476813	0.9633463	-0.5328591	1.52E-25	1.49E-11	3.39E-13
BCL3	-0.3301806	0.8350919	-0.6683016	5.71E-25	1.27E-05	1.54E-16
PLAUR	-0.3242087	0.9030538	-0.6050933	1.39E-24	9.12E-07	9.55E-15
PRSS23	-0.3130529	0.7505447	-0.3762038	1.51E-24	1.21E-06	6.45E-07
TGFB1	-0.4530202	0.8005453	-0.2208456	2.40E-24	4.55E-11	7.01E-04
LY96	-0.2337267	0.8068546	-0.506927	1.51E-23	7.57E-05	1.06E-09
RAB27A	-0.2511158	0.9282769	-0.6349317	2.48E-22	1.55E-05	8.62E-18
P4HA2	-0.3243781	0.8230918	-0.5169029	3.58E-22	1.51E-06	1.64E-10
TNFAIP8	-0.2727737	0.8002663	-0.6176751	3.64E-22	1.17E-04	4.24E-15
CLEC2B	-0.247904	0.8130403	-0.5733106	4.81E-22	1.71E-04	4.65E-11
IGFBP6	-0.3942758	0.9260064	-0.4439989	9.14E-21	5.89E-10	4.61E-09
S100A4	-0.2161899	0.7794643	-0.6191985	1.40E-20	9.82E-04	1.61E-11
BACE2	-0.2372396	0.743845	-0.5103648	1.73E-20	4.00E-04	3.94E-12
RUNX1	-0.2320887	0.8228707	-0.6410888	3.78E-20	4.17E-04	3.24E-17
CAV1	-0.2960075	0.6907206	-0.3536667	4.60E-19	1.07E-05	6.61E-06
TD02	-0.3284925	0.7889462	-0.4641819	7.54E-19	5.27E-06	2.24E-09
GCNT1	-0.3985057	0.9208544	-0.5366694	1.17E-18	3.57E-09	6.00E-15
IL7R	-0.3733264	0.8217958	-0.5341366	1.64E-18	3.15E-07	1.99E-13
ITGB1	-0.1998173	0.7372965	-0.4428406	1.56E-17	1.84E-04	7.70E-09
FTL	-0.2262859	0.4573017	-0.2301706	2.24E-17	7.67E-05	2.69E-04
DKK1	-0.4969302	0.8742982	-0.2981178	2.30E-17	7.15E-16	3.03E-06
SLPI	-0.3804185	0.6777462	-0.3790458	9.35E-17	2.27E-06	5.63E-06
SOCS3	-0.2926501	0.6366954	-0.438964	1.21E-16	1.88E-04	2.99E-07
ACPP	-0.4394624	0.7743585	-0.367546	2.95E-16	2.27E-12	9.64E-08
LOX	-0.3524578	0.7583321	-0.4568902	3.94E-16	1.20E-06	5.09E-10
CDCP1	-0.2967071	0.6905856	-0.5123239	1.05E-15	7.49E-05	7.05E-13
COL1A2	-0.296882	0.7128535	-0.3870224	2.07E-15	2.90E-06	1.47E-06
IKBKE	-0.2739739	0.7006965	-0.5181077	5.91E-15	5.50E-04	2.38E-10
SLC16A3	-0.3825022	0.6813685	-0.2965056	8.98E-15	2.32E-07	1.30E-04
SYNGR2	-0.2725832	0.5958415	-0.4071258	1.28E-14	1.25E-04	3.97E-06
SDC1	-0.4082852	0.8111449	-0.3517891	2.18E-14	5.90E-11	1.51E-07
CD72	-0.5211012	0.7154377	-0.2690326	2.74E-14	8.91E-13	6.73E-04
CNN2	-0.347207	0.7604034	-0.3483462	3.15E-14	5.99E-08	4.01E-06
LUM	-0.2452876	0.7013139	-0.3866303	3.80E-14	2.83E-05	5.09E-07
PTGS2	-0.3305929	0.703282	-0.3998502	6.00E-14	2.27E-06	4.56E-07
FHL2	-0.323916	0.6982201	-0.2728921	6.36E-14	2.64E-07	2.52E-04
BNC2	-0.23598	0.6563749	-0.4943579	1.12E-13	7.95E-04	6.93E-09
COL5A1	-0.2734181	0.6789417	-0.4534127	1.34E-13	4.01E-05	4.05E-08
PDK3	-0.3153254	0.5736999	-0.4082927	2.15E-13	1.31E-04	2.93E-06
ANPEP	-0.4003642	0.7104048	-0.3826919	4.50E-13	1.10E-09	3.84E-08
COL15A1	-0.375221	0.5974213	-0.3807798	5.85E-13	4.15E-07	1.86E-06
LGALS8	-0.2507261	0.59947	-0.4378582	6.35E-13	2.46E-04	5.15E-07
SFT2D2	-0.2391181	0.6804692	-0.4063897	7.07E-13	1.86E-04	5.43E-07
ECGF1	-0.2920745	0.6721983	-0.4229439	1.22E-12	3.97E-05	5.72E-09
UAP1	-0.2401265	0.6263704	-0.1853797	3.78E-12	1.37E-06	7.15E-04
TGM2	-0.2524213	0.7216703	-0.5211358	4.68E-12	1.16E-04	5.16E-13
CXCL6	-0.336495	0.7262296	-0.3410279	5.95E-12	2.40E-07	3.46E-07
LOXL2	-0.3034555	0.5788239	-0.2949444	6.46E-12	1.32E-05	6.89E-04
FAP	-0.2757887	0.6831894	-0.5098614	6.80E-12	1.54E-04	3.38E-10
PTGES	-0.3594317	0.7605498	-0.3910543	7.41E-12	6.90E-08	2.00E-08
FTH1	-0.2160197	0.5176467	-0.3091485	1.43E-11	4.12E-04	1.18E-04
DSC2	-0.3736628	0.6772477	-0.3616846	1.82E-11	7.54E-09	1.47E-07
BST1	-0.4226815	0.6341181	-0.2578699	2.16E-11	1.33E-08	8.75E-04
FTHP1	-0.2316578	0.5387358	-0.3347232	2.71E-11	5.30E-04	4.01E-05
CTSW	-0.3156596	0.5571091	-0.3580143	3.95E-11	3.12E-05	9.26E-06
LOC57228	-0.2331183	0.6392667	-0.3679127	4.53E-11	2.87E-04	2.93E-07
DYRK3	-0.301936	0.5304544	-0.2728341	4.61E-11	7.59E-05	3.84E-04
PAPPA	-0.3562188	0.7196138	-0.3682087	5.55E-11	9.62E-08	1.68E-07
DCBLD2	-0.4040357	0.717021	-0.2757229	6.89E-11	5.57E-10	8.87E-05
IL1A	-0.3758778	0.5674055	-0.2926806	7.59E-11	1.63E-07	4.05E-04
CMKLR1	-0.2815862	0.5756278	-0.3603891	1.44E-10	5.50E-05	9.56E-07
NFKB2	-0.2753014	0.5079234	-0.3667047	1.85E-10	2.69E-04	9.12E-05
AFP	-0.3121311	0.4902358	-0.2473632	2.84E-10	9.10E-09	7.38E-06
ITGBL1	-0.2830229	0.5872313	-0.3929757	3.08E-10	2.33E-04	8.59E-08
CPZ	-0.2854125	0.6145076	-0.4090309	5.76E-10	7.08E-05	5.54E-08

Class3 (Proneural)						
GeneSymbol	avg_expr_class1	avg_expr_class2	avg_expr_class3	ttest p-val (expressed higher in class3)	ttest p-val (expressed lower in class1)	ttest p-val (expressed lower in class2)
TMSL8	-0.3016644	-0.3935403	1.0409984	1.69E-31	2.02E-06	2.15E-09
MLLT11	-0.300483	-0.4066704	0.9085497	2.43E-31	9.81E-06	1.77E-06
HN1	-0.4049562	-0.2284888	0.9900083	4.84E-28	4.26E-12	8.64E-05
RAB33A	-0.4049605	-0.4549991	1.1470241	1.09E-26	1.10E-09	2.54E-11
MYT1	-0.245166	-0.5912831	1.0938803	3.39E-23	1.50E-04	1.70E-21
FAM77C	-0.4384087	-0.4591366	1.1840729	6.39E-23	1.10E-11	6.29E-13
HOXD3	-0.3548425	-0.4723868	1.0499486	6.66E-22	1.15E-06	1.31E-12
HDAC2	-0.1291983	-0.385803	0.9200472	4.26E-21	5.36E-04	2.24E-10
KLRC3	-0.3356782	-0.5144566	1.1402149	1.64E-20	1.59E-07	2.49E-16
C1QL1	-0.4610427	-0.2934364	0.9228527	9.60E-20	1.09E-10	1.28E-04
LOC81691	-0.2833291	-0.4930101	1.0172924	2.07E-19	2.48E-05	3.72E-13
NPPA	-0.3153572	-0.4686646	0.9409388	4.33E-19	1.40E-06	1.06E-13
MNX1	-0.4180833	-0.3953119	1.0118808	6.23E-19	1.46E-09	2.13E-09
CA10	-0.3157046	-0.4688162	1.0604452	1.44E-18	7.80E-07	1.94E-12
PTTG1	-0.2818446	-0.252283	0.844351	1.47E-18	2.51E-07	3.83E-05
HRASL5	-0.2326171	-0.4952451	0.9821306	3.91E-18	1.34E-04	1.53E-12
UGT8	-0.3180807	-0.2779364	0.8696151	1.00E-17	8.69E-07	1.45E-05
PFN2	-0.2776102	-0.33691	0.7191347	1.94E-17	1.96E-04	2.90E-06
MTSS1	-0.2706171	-0.4257463	0.8904839	2.34E-17	3.33E-05	3.08E-09
TBPL1	-0.2659113	-0.2001825	0.7887958	6.17E-17	5.48E-07	3.97E-04
EPHB1	-0.3091217	-0.4121933	0.8981786	6.48E-17	3.15E-06	1.12E-08
TCF1	-0.1486039	-0.2863144	0.743927	1.48E-16	7.18E-04	3.62E-07
DCTN3	-0.3434929	-0.2588493	0.7933687	5.28E-16	1.54E-07	1.50E-04
PAK7	-0.4231327	-0.3797896	1.0390323	6.65E-16	5.27E-11	2.77E-09
PTTG3	-0.2113549	-0.3517602	0.7656064	5.21E-15	7.16E-04	1.64E-06
ERBB3	-0.2775686	-0.3967001	0.8179164	5.97E-15	6.36E-05	2.56E-07
RASL11B	-0.4051988	-0.3092294	0.9299594	1.25E-14	1.23E-09	5.82E-06
SOX10	-0.3945649	-0.2814716	0.9141436	2.03E-14	7.37E-10	1.23E-05
H2AFZ	-0.1963131	-0.1802585	0.6072791	2.17E-14	6.95E-05	8.21E-04
SMPD3	-0.3332917	-0.4428489	1.0003917	3.44E-14	4.62E-07	4.85E-11
MYB	-0.3896851	-0.2193906	0.6199307	3.53E-14	3.13E-13	3.57E-04
SLC1A1	-0.2338844	-0.3408194	0.9832234	3.94E-14	4.01E-06	2.93E-09
CAMKV	-0.3390277	-0.4534665	0.8591069	4.33E-14	2.21E-06	1.02E-10
NARF	-0.2342995	-0.2648324	0.6863992	1.08E-13	3.58E-05	1.67E-04
C2orf27	-0.234069	-0.382135	0.8408418	2.18E-13	2.94E-04	2.75E-08
CDKN1C	-0.3173029	-0.311338	0.7682126	1.16E-12	1.05E-05	1.27E-05
ZNF804A	-0.2896509	-0.3170174	0.8409686	1.36E-12	3.07E-06	2.17E-07
PDGFRA	-0.2863832	-0.2665715	0.7920235	1.78E-12	1.95E-05	3.24E-05
BCL11A	-0.2938466	-0.3326121	0.7117052	2.40E-12	1.05E-04	6.64E-06
ANKS1B	-0.2864596	-0.2537275	0.7728923	3.19E-12	7.59E-07	1.09E-05
NDUFB11	-0.2122945	-0.2156074	0.5922207	4.77E-12	2.60E-04	2.99E-04
NIMU	-0.3326142	-0.2724806	0.7529392	7.22E-12	8.89E-07	5.22E-05
DYNC1I1	-0.2379179	-0.2891583	0.6570059	1.31E-11	6.53E-04	5.87E-05
JPH3	-0.3590108	-0.330846	0.6349092	1.41E-11	1.21E-05	6.99E-05
GABRA3	-0.3993775	-0.1963359	0.8250111	2.07E-11	1.74E-09	8.56E-04
FA2H	-0.2702598	-0.2703448	0.6537774	2.15E-11	2.50E-04	2.10E-04
MAS1	-0.3046347	-0.3354363	0.6970807	3.38E-11	5.87E-05	6.83E-06
IL1RAPL1	-0.2749149	-0.3020451	0.5898744	5.70E-11	1.77E-04	8.31E-06
B4GALNT1	-0.397308	-0.1787731	0.876193	6.32E-11	6.18E-10	8.78E-04
C20orf42	-0.3240526	-0.2442618	0.7082123	8.20E-11	5.04E-06	7.53E-04
SIM2	-0.2750338	-0.3803632	0.7011917	1.44E-10	3.35E-04	1.20E-07
GPR23	-0.3244403	-0.2835967	0.7246279	1.62E-10	3.73E-06	3.71E-05
TNRC4	-0.2574299	-0.2826808	0.6339999	3.23E-10	2.51E-04	9.58E-05
ACOT7	-0.2205781	-0.2311432	0.6440496	5.07E-10	2.65E-04	5.68E-04
REC8	-0.2056802	-0.2448583	0.7122678	5.84E-10	3.39E-04	1.61E-04
SLC17A6	-0.2944968	-0.3424367	0.8368816	6.19E-10	5.37E-06	8.18E-08
MAGEL2	-0.2990331	-0.2544413	0.7158774	6.21E-10	7.83E-06	2.84E-04
BRSK2	-0.2521112	-0.3088813	0.6295423	2.05E-09	4.61E-04	2.33E-05
PKMYT1	-0.2764801	-0.3067307	0.6672489	2.17E-09	1.23E-04	1.84E-05
KLRK1	-0.3539004	-0.2152245	0.7865578	3.31E-09	1.07E-08	3.72E-04
DCT	-0.3140625	-0.2525283	0.6157531	3.40E-09	1.72E-07	2.27E-04
SUSD5	-0.2582856	-0.3080969	0.8010203	1.87E-08	2.11E-05	7.34E-07
GABRB3	-0.3135961	-0.2774201	0.6789456	2.10E-08	6.69E-06	6.60E-05
GBX2	-0.3601163	-0.2994861	0.522656	3.11E-08	8.38E-06	2.41E-04
CENPJ	-0.2337801	-0.2822015	0.7430913	9.06E-08	8.93E-05	1.82E-05
KLRC4	-0.2695634	-0.2908663	0.7819548	9.10E-08	1.47E-06	3.35E-07
GRID2	-0.2664074	-0.2884373	0.7201703	5.18E-07	6.27E-05	1.25E-05
CENTG1	-0.2585238	-0.3198148	0.5513022	1.32E-06	1.46E-04	1.43E-06
DAZ4	-0.3063219	-0.2130719	0.7405883	2.54E-06	2.41E-07	1.37E-04
DAZ1	-0.2125429	-0.2513081	0.6276327	5.31E-06	5.74E-04	1.13E-04

Applying this method we found that the overall concordance of cluster assignments on 144 TCGA GBM samples profiled using both RNA sequencing and Affymetrix U133A microarrays was 95.14% (**Figure 2.4**). This was an improvement over the 77% subtype concordance determined using previously reported methods (Verhaak, Hoadley et al. 2010).

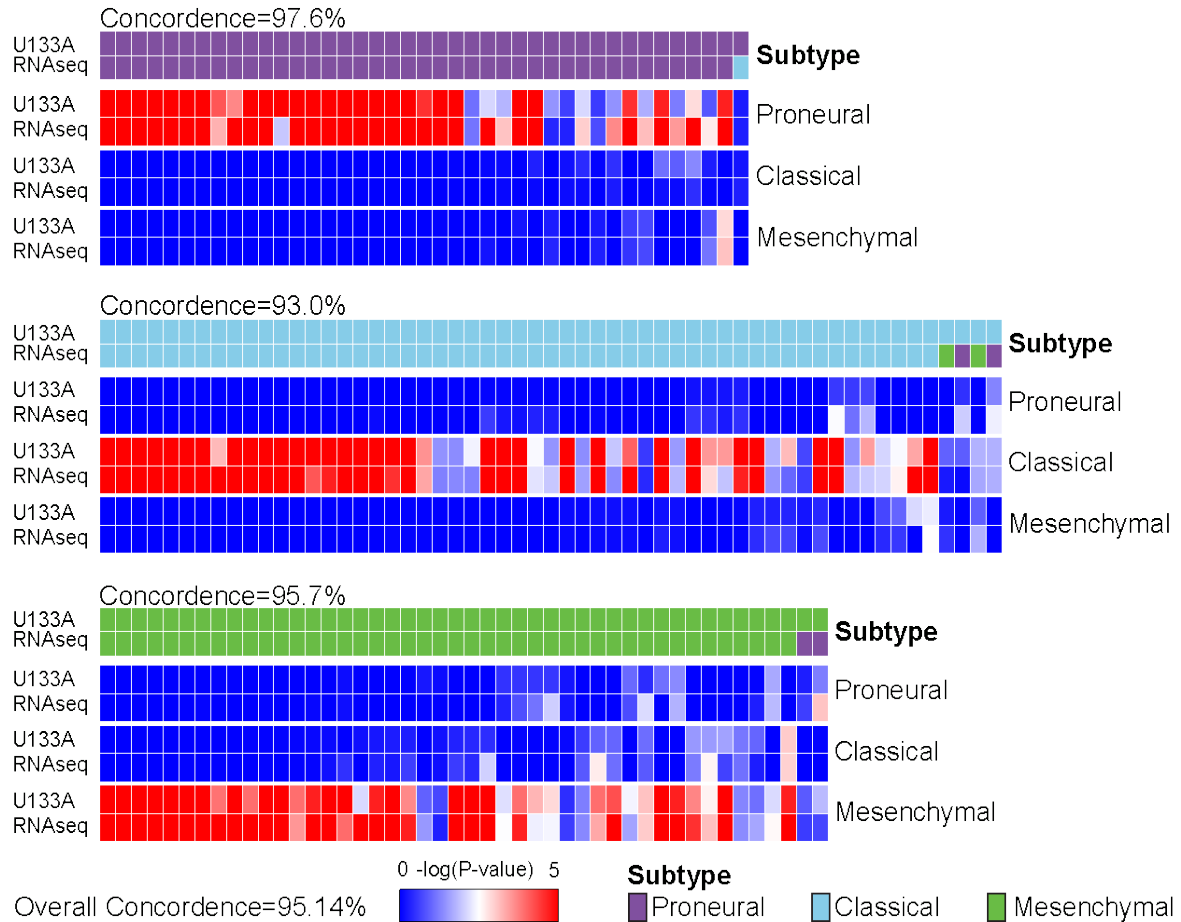


Figure 2.4 Concordance of transcriptional classification of GBMs cross multiple platforms

Through TCGA, the expression profiles of 144 GBM were analyzed using both Affymetrix U133A gene expression arrays and RNA sequencing. The empirical $-\log$ (P-value) of raw ssGSEA enrichment scores at each signature are shown as heatmaps, with dark blue representing no activation and bright red as highly activated. For each panel, the first row shows U133A based classification, and the second row indicates RNA-seq subtype classification.

2.3.2 Multi-activation of subtype signatures associated with intra-tumoral heterogeneity

We observed that 34/369 (9.2%) samples showed significant enrichment of multiple ssGSEA scores (empirical classification p-value<0.05), suggesting these cases activate more than one transcriptional subtype. To quantify this phenomenon, a score ranging from 0 to 1 was defined to quantitatively evaluate the simplicity of subtype activation based on order statistics of ssGSEA score. Samples with high simplicity scores activated a single subtype and those with lowest simplicity scores activated multiple subtypes. All multi-subtype TCGA samples showed simplicity scores of less than 0.1 (**Figure 2.5**).

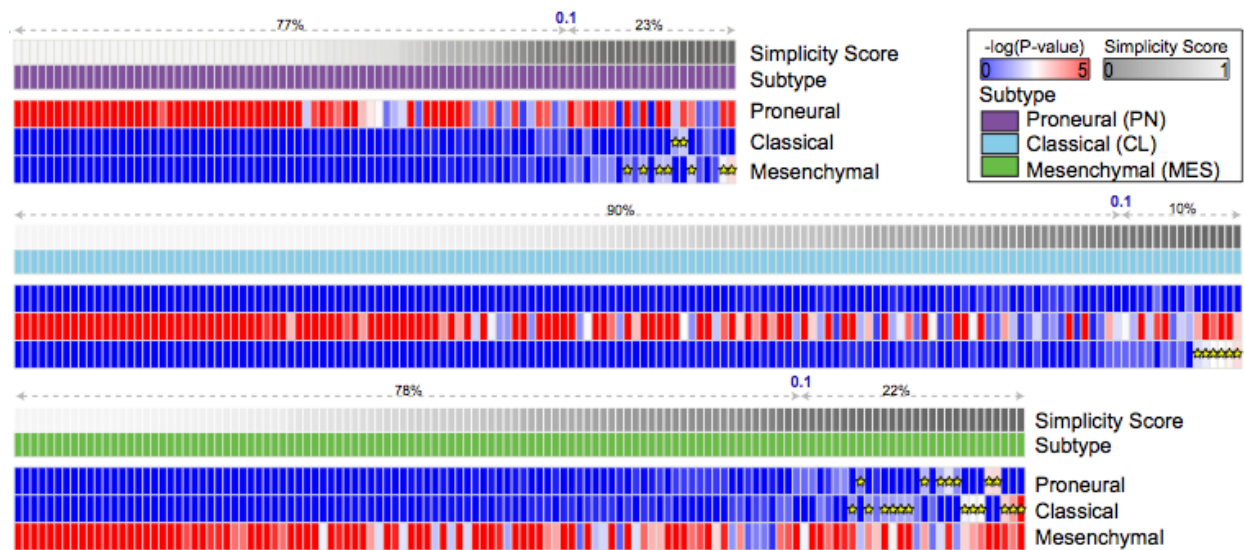


Figure 2.5 Multi activation of transcriptional subtypes associated with intra-tumoral heterogeneity

The expression profiles of 369 IDHwt GBMs were analyzed using Affymetrix U133A. The empirical $-\log(P\text{-value})$ of raw ssGSEA enrichment scores at each signature are shown as heatmaps, with dark blue representing no activation and bright red as highly activated. Yellow star indicates the secondary activated subtype (empirical p-value<0.05). For each panel, the first row shows simplicity score, and the second row indicates transcriptional subtype.

Then we evaluated the distribution of somatic variants across these three molecular subtypes (**Figure 2.6**) and confirmed the strong associations between transcriptomic subtypes and genomic abnormalities in previously reported driver genes (Fisher's exact test) (Verhaak, Hoadley et al. 2010, Brennan, Verhaak et al. 2013).

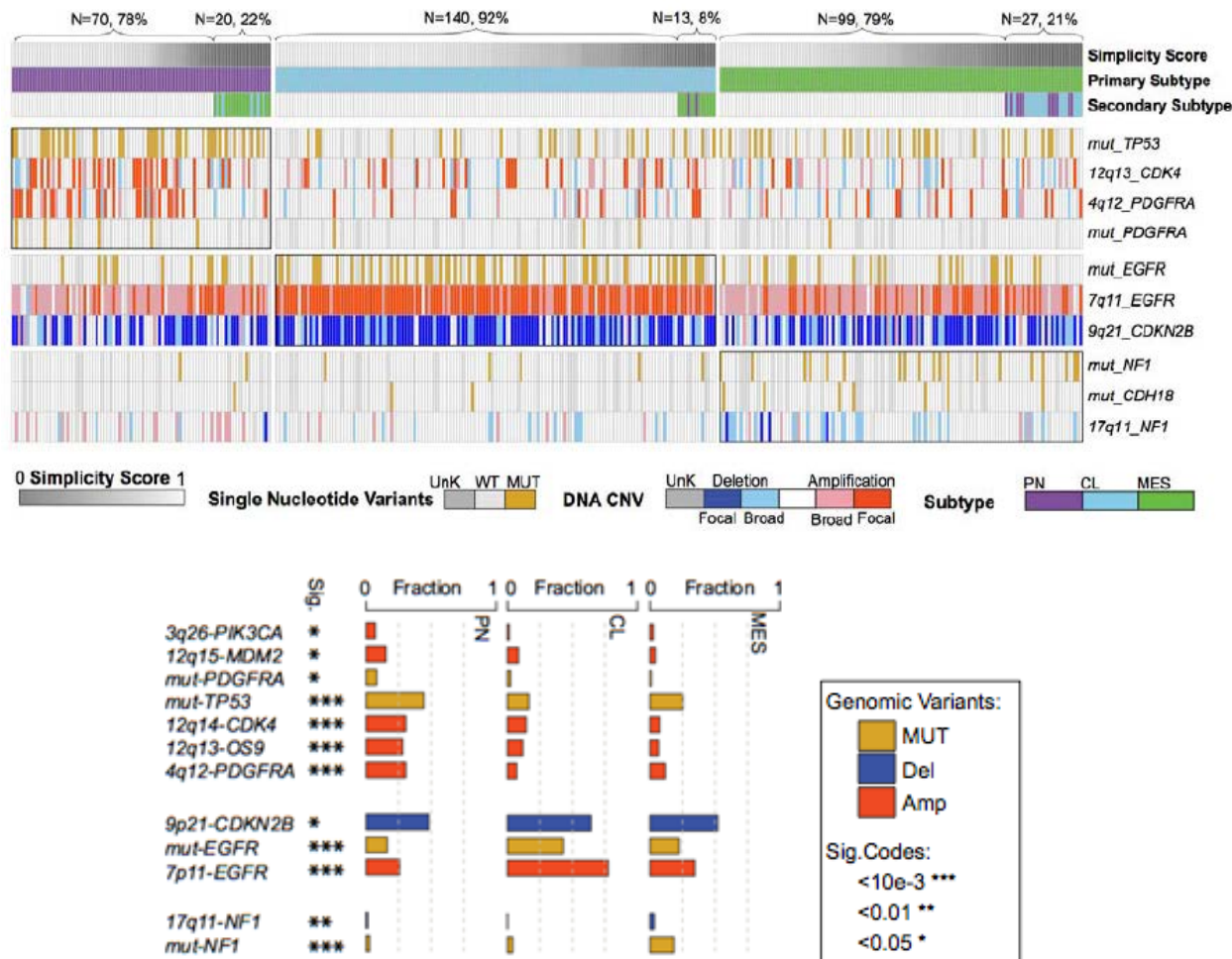


Figure 2.6 Genomic alteration patterns for each subtype

(A) The most prominent GBM somatic events. In the upper panel, the first row shows the simplicity score for each sample, the second row shows primary subtypes of each samples determined by ssGSEA using U133A array expression data. The third row shows the secondary subtype of each sample. In the bottom panel, mutations and copy number alterations in key GBM genes are shown. Missing values i.e. when exome data or SNP array data was not available are labeled in gray.

(B) Frequency of subtype related somatic genomic alterations.

To determine whether transcriptional heterogeneity associated with genomic intra-tumoral heterogeneity, we correlated simplicity scores, total mutation rates and subclonal mutation rates. Included in the analysis were 224 TCGA GBMs with available whole exome sequencing data (Kim, Zheng et al. 2015) and ABSOLUTE (Carter, Cibulskis et al. 2012) determined high tumor purity (> 0.8) to equalize the mutation detection sensitivity (Aran, Sirota et al. 2015). Although not significant (Wilcoxon rank test p -value=0.143), the total mutation rate was less in the bottom 30% with lowest simplicity scores versus the top 30% samples with highest simplicity scores. The subclonal mutation rate was significantly higher (p -value=0.024) in samples with lowest simplicity scores (Figure 2.7), suggesting that increased intra-tumoral heterogeneity associates with increased transcriptional heterogeneity.

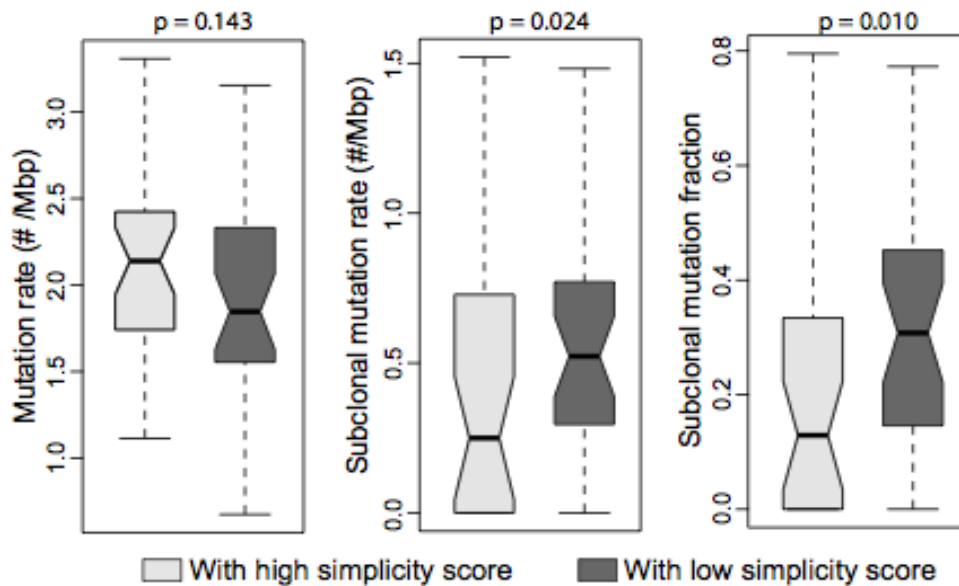
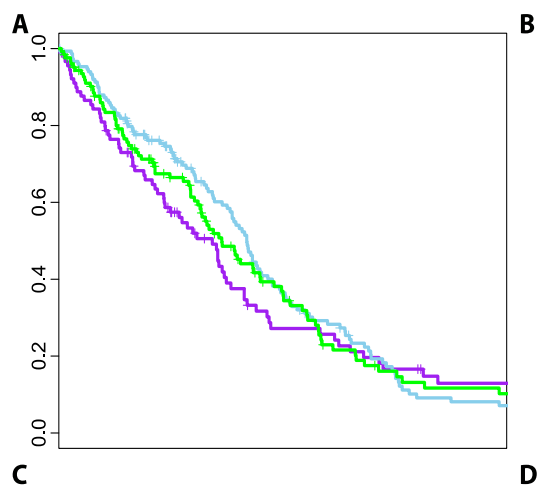


Figure 2.7 Association between genomic and transcriptomic intratumoral heterogeneity

Comparison of mutation rate, subclonal mutation rate and subclonal mutation fraction between IDHwt GBMs with high and low simplicity scores, P-values were calculated using Wilcoxon rank test and shown at the top of each panel.

We compared outcomes amongst the three transcriptional groups and observed no significant differences. However, when restricting the analysis to samples with high simplicity scores, a distinct trend of MES showing worst survival and PN the most favorable outcome became visible. For example, Kaplan-Meier analysis of 88 samples with simplicity scores >0.99 showed a median survival of 11.4, 14.7 and 16.7 months were detected in MES, CL and PN, respectively, which was significantly different (log rank test, $p=0.048$) (Figure 2.8).



(N=7)

Figure 2.8 Impact of intratumoral heterogeneity on overall survival between different transcriptional subtypes

The Kaplan-Meier survival curves for each transcriptional subtype are shown. The patients were selected based on increased simplicity score as threshold from panel A, B (>0.9), C (>0.99), D (0.999).

In addition, higher simplicity scores correlated with relative favorable outcome within the PN set, non-significant in the CL subtype, and correlated with relatively unfavorable survival in the MES class (**Figure 2.9**).

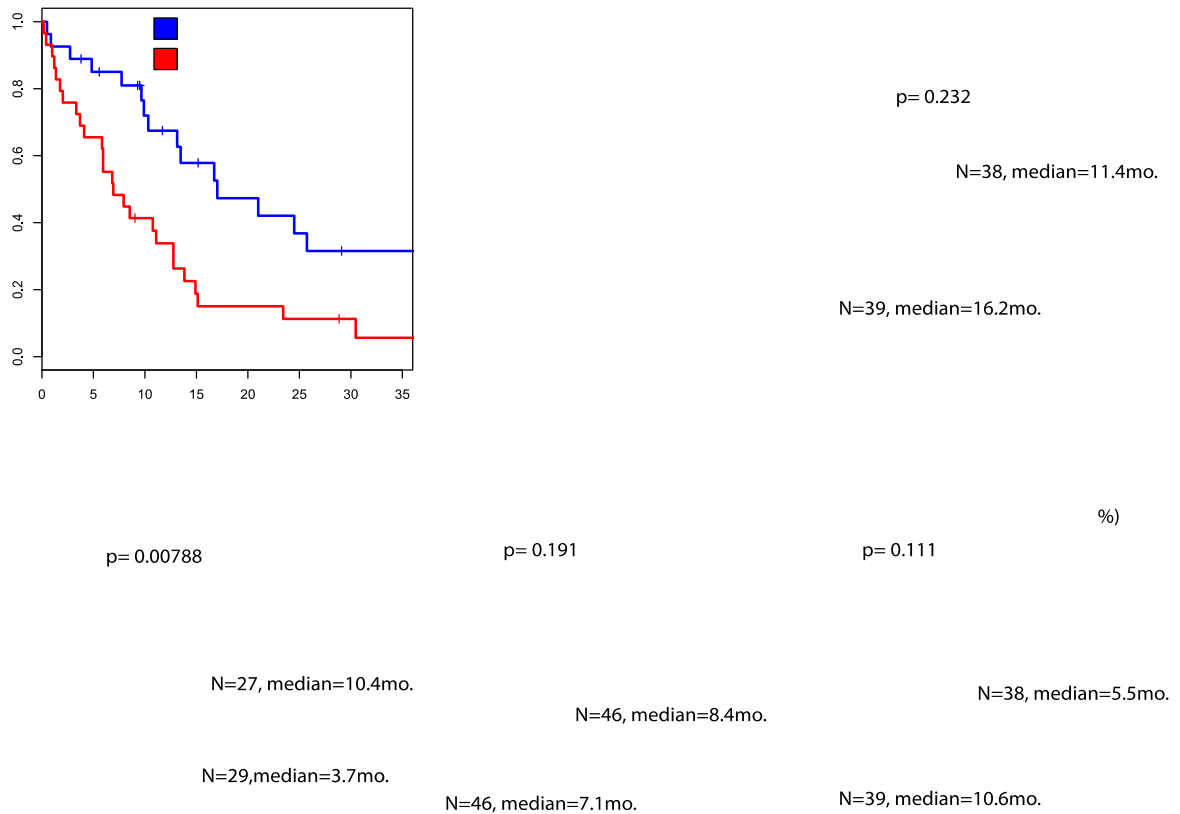


Figure 2.9 Impact of intra-tumoral heterogeneity on patient survival in each transcriptional subtype

The Kaplan-Meier survival curves for each transcriptional subtype are shown. The patients were grouped based on their simplicity scores. The blue curve denotes the patients with 30% high simplicity scores, the red curve denotes the patients with 30% low simplicity scores.

Single GBM cell RNA sequencing recently suggested that GBMs are comprised of a mixture of tumor cells with variable GBM subtype footprints (Patel, Tirosh et al. 2014). We used this data to classify 502 single GBM cells in addition to the bulk tumor derived from five primary glioblastomas. All bulk tumor samples showed simplicity scores less than 0.05 suggesting high transcriptional heterogeneity compared to 45 of 369 TCGA GBM samples with simplicity scores below 0.05. In four of five cases (MGH26, MGH28, MGH29 and MGH30), the bulk tumor samples were classified in the same primary subtype as the majority of their single cells (**Figure 2.10**). Our analysis suggests that the heterogeneity observed at the single cell level is captured in the expression profile of the bulk tumor, and that the five GBM samples studied at the single cell level represented samples with relatively high transcriptional heterogeneity.

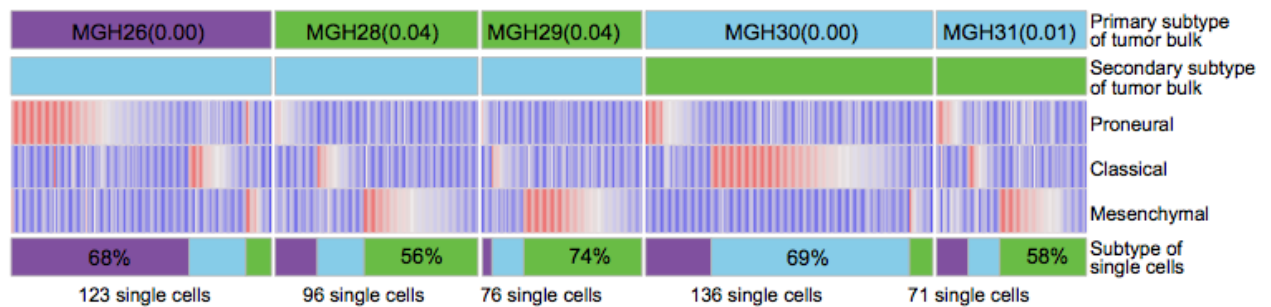


Figure 2.10 Transcriptome classification of five bulk tumor samples their derived 502 single GBM cells

The top two row of each panel show the dominant and secondary subtype of the GBM tumor bulk. The heatmap of each panel shows the empirical $-\log(P\text{-value})$ of the ssGSEA scores of the derived single GBM cells on each of the three subtype signatures. The bottom row shows the subtype distribution of derived single GBM cells within the same GBM tumor of origin.

2.3.3 Transcriptional subtypes differentially activate the immune microenvironment

Despite restricting the cluster analysis to genes exclusively expressed by GBM cells, we found that tumor purity predictions based on ABSOLUTE were significantly reduced in GBM classified as MES (Student T-test $p\text{-value} < 10e-14$; **Figure 2.11A**). This was concordant by

gene expression based predictions of tumor purity using the ESTIMATE method (Student T-test p-value < 10e-32; **Figure 2.11B**) (Yoshihara, Shahmoradgoli et al. 2013) . The ESTIMATE method has been optimized to quantify tumor-associated fibroblasts and immune cells (Yoshihara, Shahmoradgoli et al. 2013) and the convergence of a decreased ABSOLUTE and decreased ESTIMATE tumor purity confirmed previous suggestions on the increased presence of microglial and neuroglial cells mesenchymal GBM (Bao, Wu et al. 2006, Engler, Robinson et al. 2012, Ye, Xu et al. 2012, Gabrusiewicz, Rodriguez et al. 2016). The mean simplicity score of samples classified as MES was 0.53 which was significantly lower than in PN (Wilcoxon rank test p-value < 0.019) and CL subtypes (Wilcoxon rank test p-value < 0.0001), confirming increased transcriptional heterogeneity.

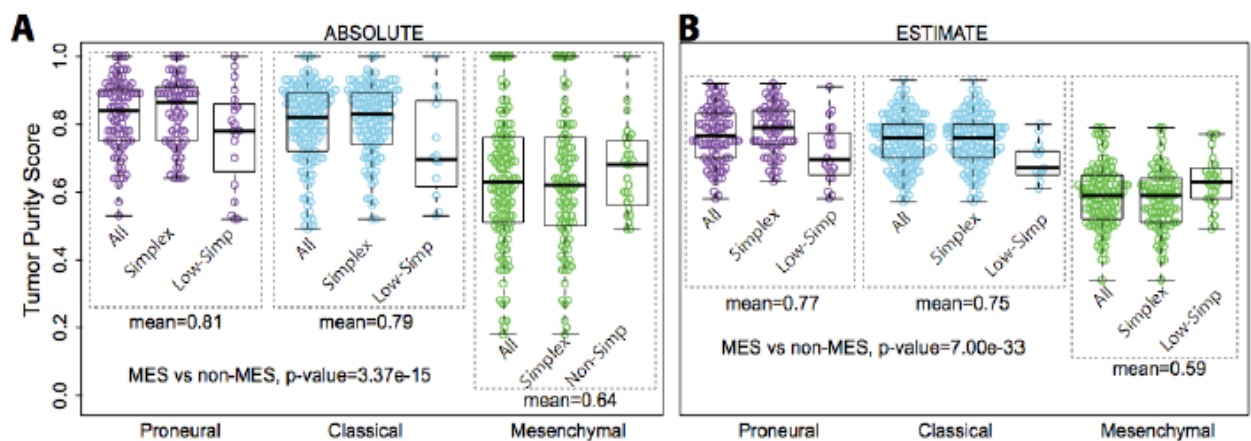


Figure 2.11 Transcriptional subtypes differentially activate the immune microenvironment

Tumor purity of 364 respectively 369 TCGA IDHwt GBM samples was determined by ABSOLUTE (A) and ESTIMATE (B) The difference in tumor purity between subtypes was evaluated using two-sample heteroscedastic t-test.

In order to identify genomic determinants of macrophage/microglia chemo-attraction, we compared genomic alterations between mesenchymal class samples with high (n=51) and low (n=51) ABSOLUTE based tumor purity. GBM carrying heterozygous loss of NF1 or somatic mutations in NF1 showed reduced tumor purity compared to GBM with wild type NF1 (Wilcoxon rank test p-value=0.0007) and this association was similarly detected when

limiting the analysis to MES samples (Wilcoxon rank test p -value=0.017) (**Figure 2.12**). Formation of dermal neurofibromas in the context of *Nf1* loss of heterozygosity has been reported to be context and microenvironment dependent (Le 207 et al., 2009). Further functional studies may clarify whether *NF1* deficient GBM are able to recruit cells that provide them with a proliferative advantage, or whether *NF1* loss provides that proliferative advantage in a specific tumor-associated microenvironment context.

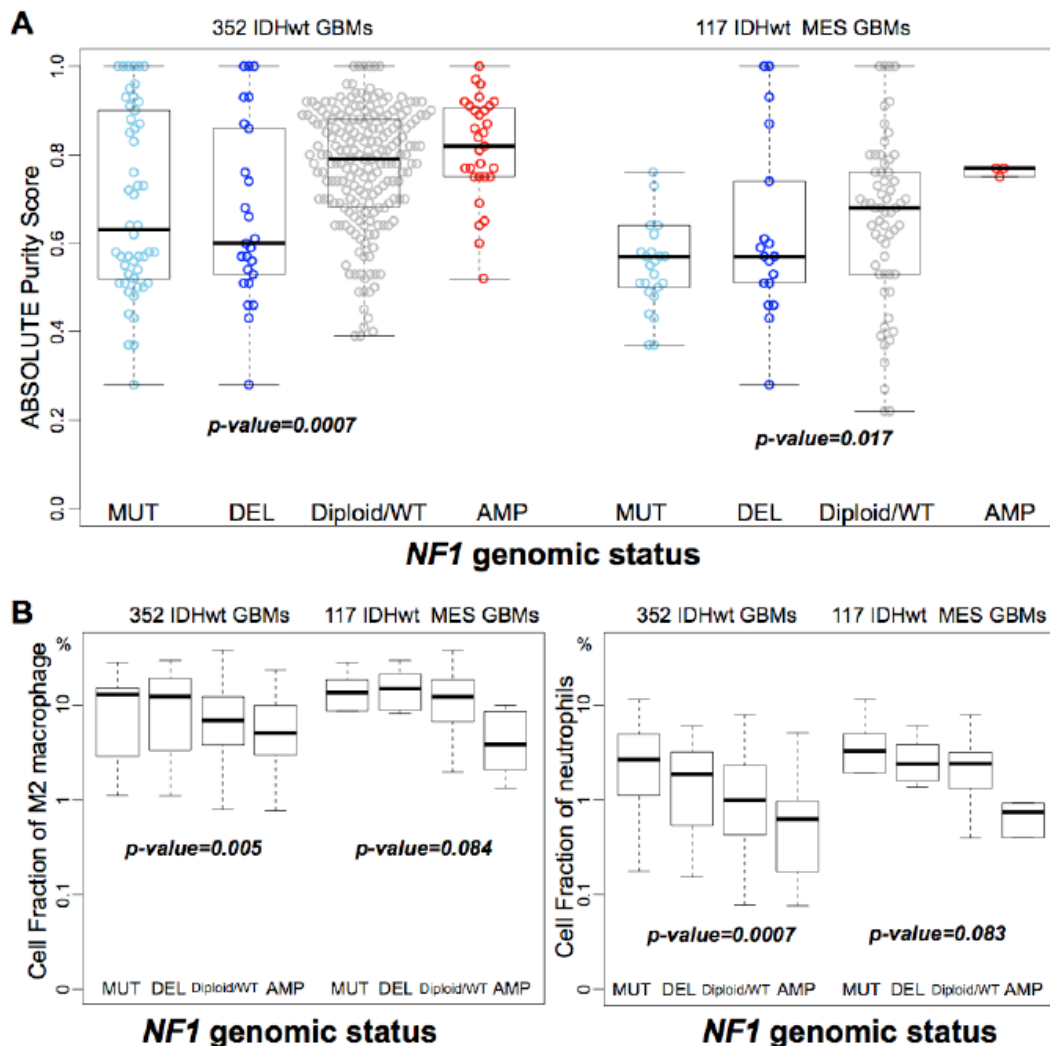


Figure 2.12 Comparison of tumor purity and immune cell fraction between GBMs with different *NF1* genomic status

P -values were calculated using Wilcoxon rank test between samples carrying *NF1* deletion/mutation versus *NF1* WT and carrying amplification. [* Courtesy of Qianghu Wang]

To determine the cellular components of the tumor microenvironment across different transcriptional subtypes, we used the CIBERSORT in silico cytometry (Newman, Liu et al. 2015) method to evaluate absolute immune cell fractions. We evaluated 22 different immune cell types in 69 PN, 137 CL and 96 MES samples, after filtering samples with classification simplicity scores less than 0.1. Microglia is the resident macrophages in the central nervous system. Peripheral blood monocytes also give rise to tumor associated macrophages. These innate immune cells can be broadly classified as the proinflammatory M1 type and the alternative tumor promoting M2 type (Hambardzumyan, Gutmann et al. 2015). The M2 macrophage gene signature showed a greater association with the MES subtype (13.4%) relative to the PN (4.6%) and CL (6.0%) (**Figure 2.13**), consistent with previous analysis of the TCGA database (Doucette, Rao et al. 2013, Gabrusiewicz, Rodriguez et al. 2016). In addition to the M2 macrophage gene signature, there was also a significantly higher fraction of MES samples that expressed M1 macrophage (Student T-test p-value $3.20\text{E-}5$) and neutrophil (Student T-test p-value $1.30\text{E-}9$) gene signatures. In contrast, the activated natural killer T-cell gene signature (Student T-test p-value $4.91\text{E-}2$) was significantly reduced in the MES subtype and resting memory CD4+ T cells (Student T-test p-value $5.40\text{E-}7$) were less frequently expressed in the PN subtype.

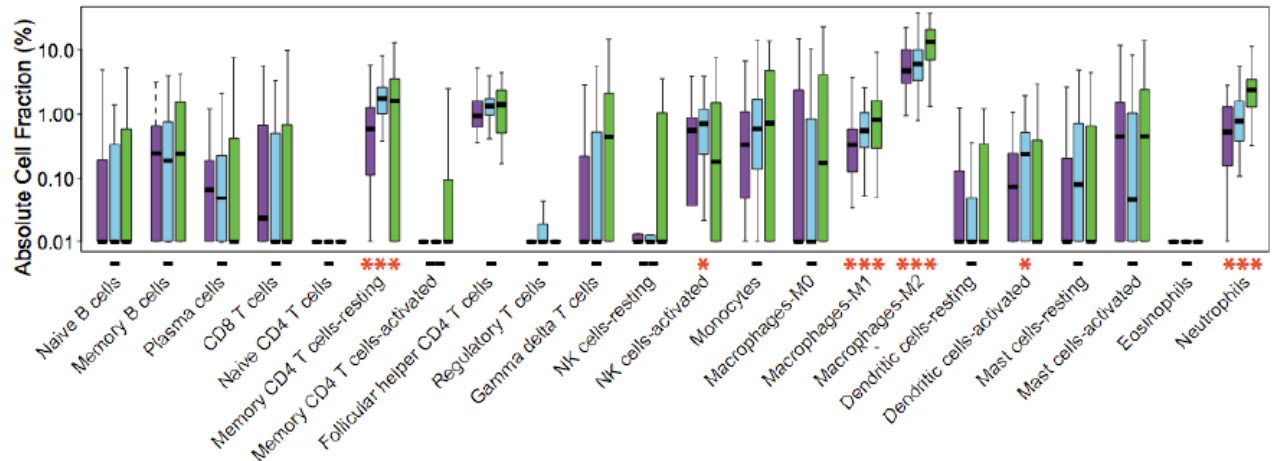


Figure 2.13 Comparison of immune cell fraction among different subtypes of GBM

Purple, blue and green boxplots indicate PN, CL and MES subtype, respectively. Immune cell fraction was estimated using CIBERSORT and corrected using ABSOLUTE purity scores. Difference of cell fraction between subtypes was evaluated using Mood's test.

To confirm the association of macrophages/microglia with the MES GBM subtype, we assessed protein expression levels of the ITGAM (alternatively known as CD11B) and IBA1 (also known as AIF1) macrophage/microglial markers in a set of 18 GBM for which we also characterized the expression subtype (**Figure 2.14**).

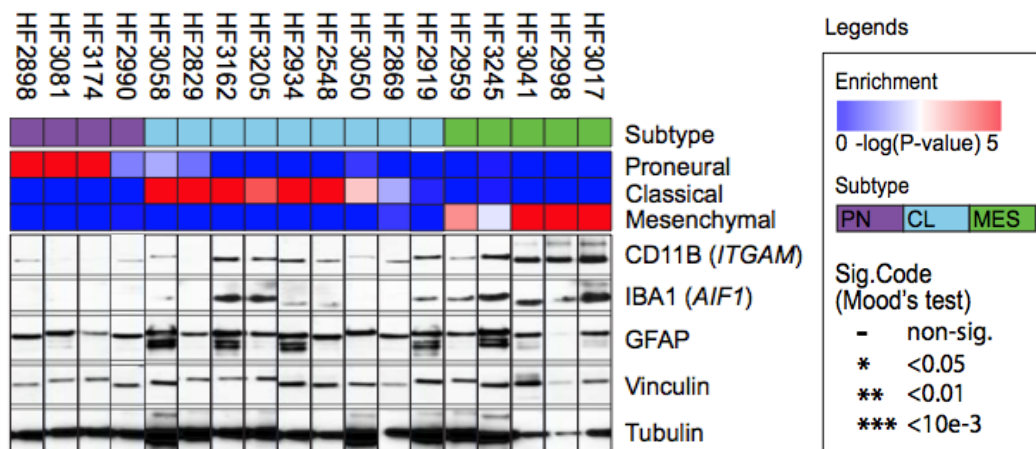


Figure 2.14 Presence of macrophages/microglia in MES GBM

The upper panel shows ssGSEA enrichment scores and associated expression subtype classifications. Bottom panels display protein expression of the microglial markers integrin alpha M (ITGAM) and allograft inflammatory factor 1 (IBA1), astrocyte marker glial fibrillary acidic protein (GFAP) and the loading control vinculin and tubulin. [* Courtesy of Florian Muller]

We confirmed the tumor associated microenvironment as the main source for *ITGAM/IBA1* transcription by comparing transcriptional levels in 37 GBM-neurosphere pairs used for gene filtering, which showed that neurospheres do not express *ITGAM/IBA1* (**Figure 2.15**). The association of the MES GBM subtype with increased level of M2 microglia/macrophages may suggest that in particular MES GBM are candidates for therapies directed against tumor-associated macrophages (Pyonteck, Akkari et al. 2013). Activated dendritic cell signatures (Student T-test p-value 7.36E-3) (**Figure 2.13**) were significantly higher in the CL subtype, suggesting this subtype may benefit from dendritic cell vaccines (Palucka and Banchereau 2012). Dendritic cells may require an activated phenotype in order to direct the immune system. A previous study suggested that MES GBM patients treated with dendritic cells were more likely to benefit (Prins, Soto et al. 2011).

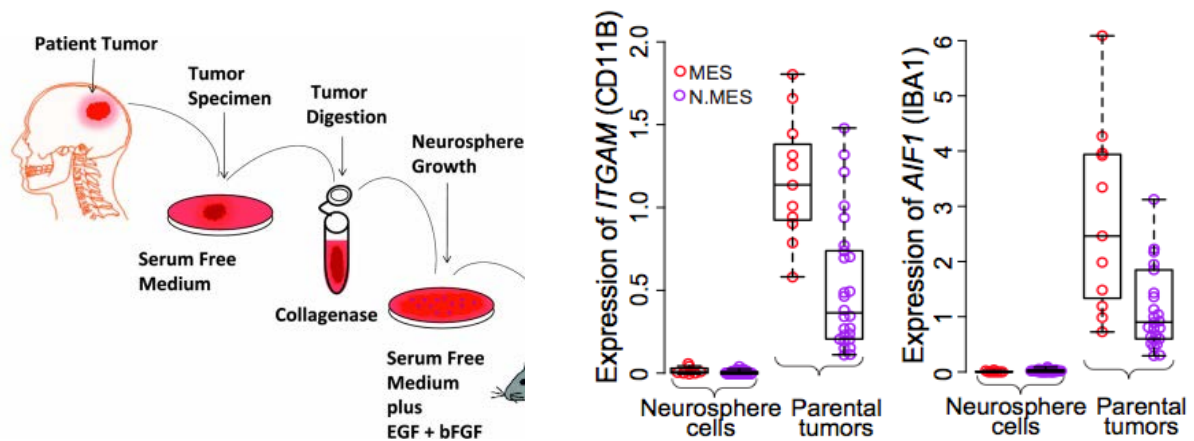


Figure 2.15 Characterization of the source of microglia in GBM

(A) Generation of neurosphere cells that deprived of microglia from tumor associated microenvironment

(B) Comparison of *ITGAM* and *IBA1* gene expression levels between GBM patients and their derived neurosphere models.

[* Courtesy of Erik Sulman, Qianghu Wang]

2.3.4 Phenotypic plasticity upon GBM recurrence

Glioblastoma has long been hypothesized to progress along a proneural to mesenchymal axis (Phillips, Kharbanda et al. 2006). First we evaluated the subtype classification and tumor contents of 22 primary GBM and their matching recurrence case. When considering subtype classifications based on the maximum ssGSEA score, the molecular subtypes remained consistent after disease recurrence for 19 of 22 patients. 3 of 22 tumor samples that switched primary subtypes showed a relative decrease in tumor purity after recurrence. In contrast, of the remaining 19 pairs with different primary-recurrent subtype classifications only 5 showed a decrease in tumor purity at time of recurrence. A remarkable shift away from the EGFR associated classical subtype was observed, with only a single case retaining this classical status. Then we evaluated whether the previously reported genomic associations between tumor subtype and genomic abnormalities persisted after disease recurrence, using 22 pairs for which genome wide DNA copy number levels and exome sequencing data were available in The Cancer Genome Atlas (TCGA) or generated by ourselves and reported elsewhere (Kim, Zheng et al. 2014). Although limited by sample size, we observed an overall trend in genomic associations that was consistent with what we found previously in primary GBM, such as two recurrences that retained the *IDH1* mutation status and the proneural phenotype, three of three recurrent classical GBM with focal *EGFR* amplification and/or *EGFR* mutations, and three of nine recurrent mesenchymal GBM with a non-synonymous mutation in *NF1* (Brennan, Verhaak et al. 2013) (Figure 2.16).

To further determine the relevance of this transition process in IDH wildtype glioma evolution, we performed a longitudinal analysis of the subtype classification and tumor-associated microenvironment in sample pairs obtained at diagnosis and first disease recurrence from 124 glioma patients. The cohort included 96 initial GBM and first recurrence, eight pairs of primary low grade glioma and matching secondary GBM, and 20 pairs of primary and recurrent low grade glioma. Gene expression profiles of 78 tumor pairs were

analyzed through transcriptome sequencing, and remaining pairs were generated using Affymetrix (n = 31) and Illumina (n = 15) microarray, respectively. To facilitate exploration of this dataset we have made it available through the GlioVis portal <http://recur.bioinfo.cnio.es/>.

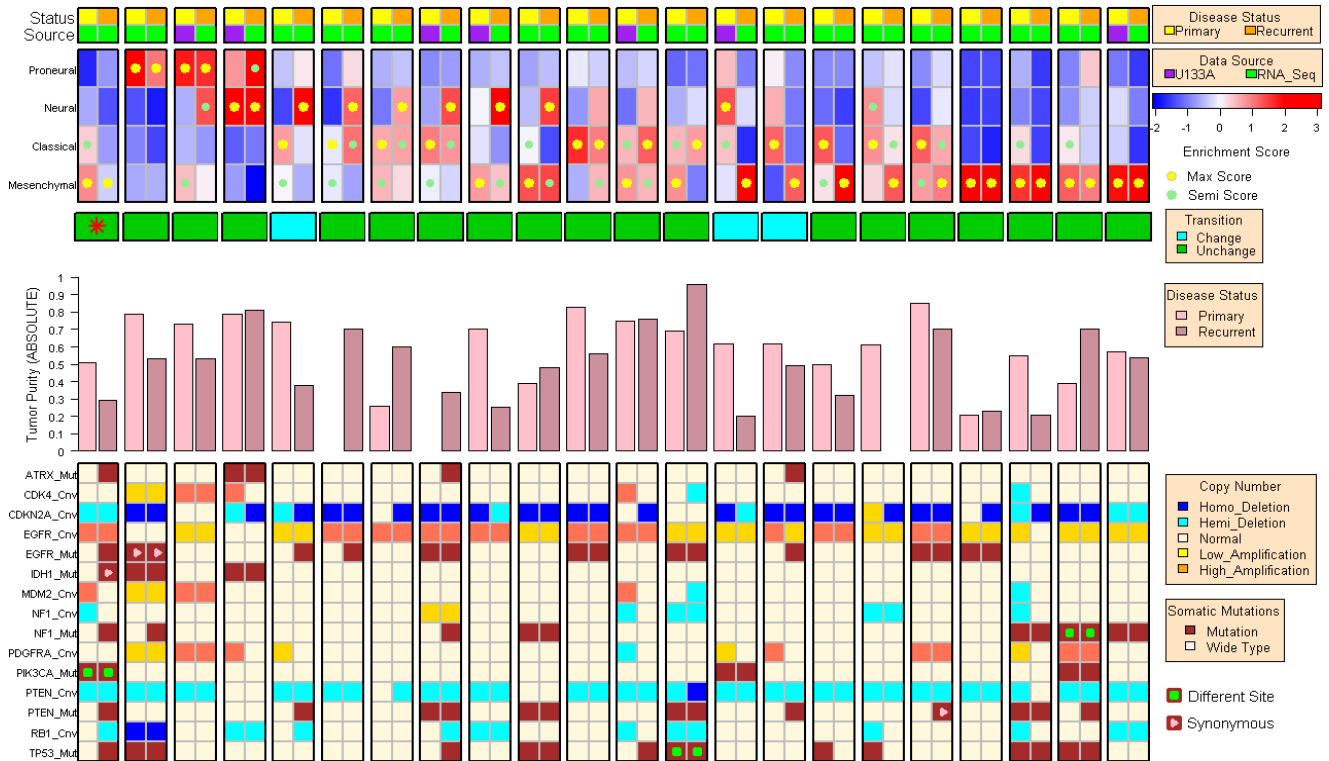


Figure 2.16 Integrative views of molecular classification and genomic alterations across molecular subtypes in paired primary and recurrent glioblastoma

The molecular classifications, tumor purity and genomic alteration landscape of 22 pairs of primary and recurrent GBM. In the upper panel, ssGSEA enrichment scores are shown and the yellow dots indicate subclass, green dots represent the secondary activation scores. In the middle panel, tumor purity based on ABSOLUTE scores is shown, light pink represent primary samples, dark pink represent recurrent samples. Bottom panel displays molecular alterations; mutations are indicated by a brown cell, with synonymous substitution labelled with pink triangle, a different base mutated labelled with green squares. copy number events are illustrated in blue for homozygous deletions, light blue for hemizygous deletions, beige for copy number neutral, orange for low level amplification, and dark red for high level amplifications. A single case with a history of low grade glioma is denoted with a red asterisk (*). The samples are sorted first by primary tumor grade, then by recurrent tumor subtype, and then by primary tumor subtype.

We used a gene expression signature (Baysan, Bozdag et al. 2012) to determine that 33 of 124 cases were IDH-mutant/GCIMP at presentation and recurrence. We used the renewed gene signatures and classification method to determine molecular subtype of the 91 pairs of IDH wild type cases and found that expression class remained consistent after disease recurrence for 48 of 91 IDH-wildtype cases (52%). The MES subtype was most stable (64%) while the CL (47%) and PN (43%) phenotypes were less frequently retained. 9, 16 and 18 post-treatment tumors switched subtypes to become CL, MES and PN at disease recurrence, respectively, indicating that PN and MES increased in higher frequency after recurrence while the CL subtype was least frequently found (**Figure 2.17A**). The CL expression class was previously found to be most sensitive to intensive therapy and it is possible that therapy provides a competitive advantage for non-CL cells, which could explain the reduced post-treatment incidence (Verhaak, Hoadley et al. 2010). Our results did not identify enrichment for proneural to mesenchymal transitions.

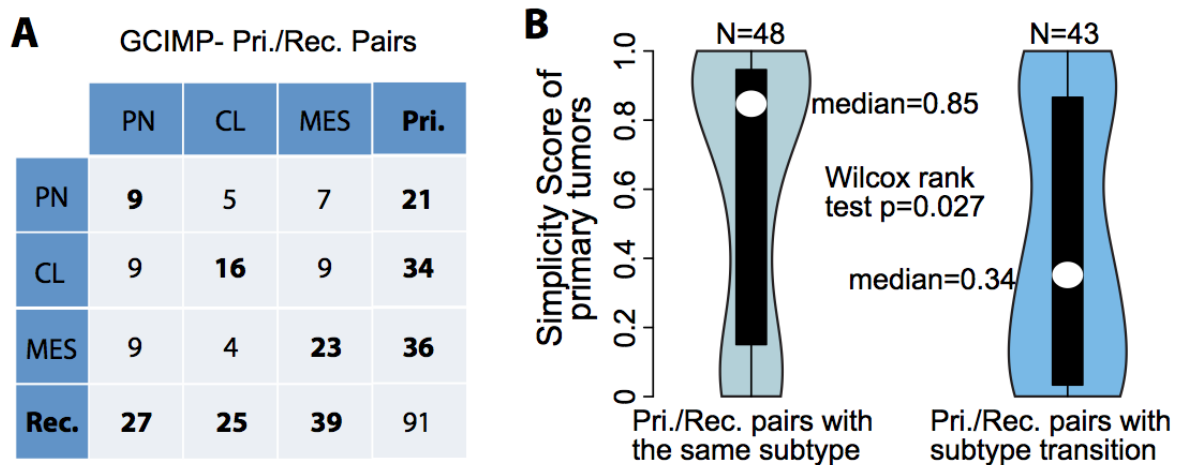


Figure 2.17 Comparison between transcriptional subtype of primary and paired recurrent tumors

(A) Rows and columns of the cross table represents subtype distribution frequency of primary and paired recurrent tumors, respectively.

(B) Violin plots show the distribution of simplicity scores of pairs with (left) and without (right) subtype transition.

We observed a significant difference in transcriptional simplicity between primary GBM retaining their expression class, versus those that switched to a different phenotype upon their recurrence (**Figure 2.17B**). GBMs with a primary tumor simplicity score greater than 0.5, with lower transcriptional heterogeneity, were classified as the same subtype in 31 of 48 (64.5%) cases, compared to 15 cases classified as the same subtype out of 41 (36.6%) cases with primary tumor simplicity scores less than 0.5. (Fisher exact test p-value=0.01)

2.3.5 Tumor microenvironment transitions upon GBM recurrence

Debulking surgery, radiotherapy and chemotherapy provide therapeutic intervention but nonetheless induce tumor evolution, including impact on their tumor infiltrated microenvironment. We explored this hypothesis by comparing the tumor associated microenvironment in primary and recurrent GBMs using CIBERSORT (Newman, Liu et al. 2015). A comparison between 91 primary and recurrent IDH-wild type tumors revealed a decrease in monocyte gene signature expression at recurrence, suggesting relative depletion of circulation derived monocytes (**Figure 2.18A**). Next, we dissected microenvironment fluctuations between primary and recurrent tumors across different subtype combinations. Primary non-MES (CL or PN) tumors showed relatively high tumor purity and consequently, recurrent tumors classified as non-MES demonstrated a relatively global decrease of immune cells while cases transitioning to MES at recurrence represented a trend towards increased immune cell fractions. Gene signatures of immunosuppressive regulatory T cells showed an increase in gene expression at recurrence across several primary-recurrence subtype combinations although the inferred cellular fractions are relatively small (**Figure 2.18B**).

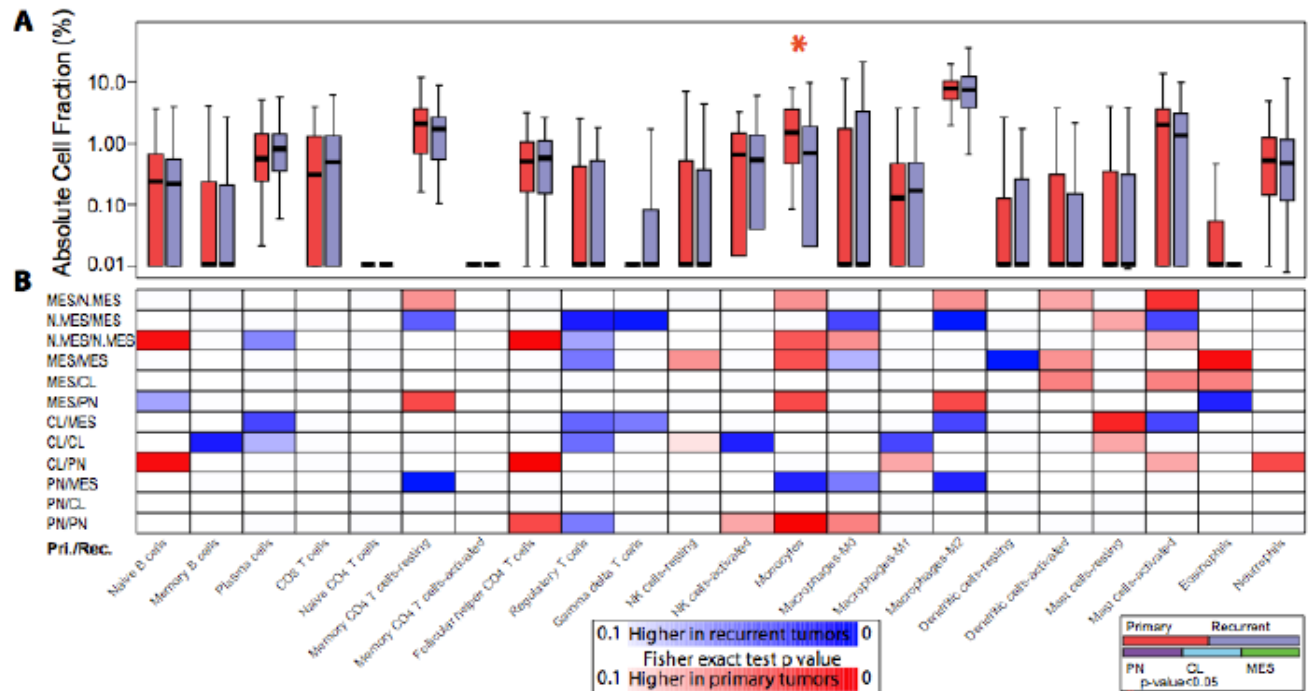


Figure 2.18 Microenvironment transition between primary and paired recurrent tumors

(A) Red and blue boxplots represent the immune cell fraction distribution of each immune cell type. Immune cell fraction was calculated using CIBERSORT and adjusted using ESTIMATE purity scores. Difference between cell fraction of primary and paired recurrent tumors was calculated using Wilcoxon rank test.

(B) The blue-to-red heat-map represents changes of immune cell fraction upon tumor recurrence per subtype transitions labeled on the left of the heat-map.

In contrast to the trend of monocyte depletion, the imputed M2 macrophage frequency was significantly higher at recurrence in cases transitioning to MES (**Figure 2.19**). This observation converges with the higher predicted fraction of M2 macrophages in primary MES GBM relative to primary non-MES GBM. M1 macrophages and neutrophils also correlated with primary MES GBM, but these associations were not confirmed for recurrent GBM.

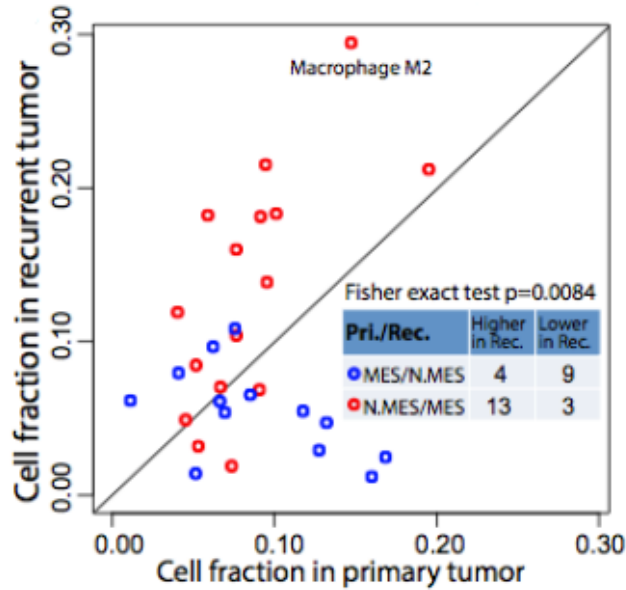


Figure 2.19 Comparison of M2 Macrophage fractions between MES and non-MES subtypes during tumor evolution

Each dot represents M2 macrophage cell fraction in a pair of primary and recurrent GBM. Red dots denote samples transit from none-MES to MES upon recurrence; blue dots denote samples transit from MES to non-MES upon relapse.

We validated the increase in macrophages using immunostaining of IBA1 expression in two primary-recurrent GBM pairs which were classified as CL to MES. IBA1 immune staining signal restricted to macrophages/microglia, cells exhibiting either globular or filamentous/spidery morphology, with no expression in glioma tumor cells. Quantitative analysis of microglia frequency using inform software for automated pathology imaging processing confirmed a significantly higher presence (p value = $2.25e-11$ and $2.12e-13$ for patient #1 and #2, respectively) of at MES recurrence (**Figure 2.20**). These findings further solidify the association between MES GBM and macrophage/microglia and extend this mutual relationship to disease recurrence. MES tumors at recurrence compared to primary MES tumors showed an increase in transcriptional activity associated with non-polarized M0 macrophages, which has been previously described (Gabrusiewicz, Rodriguez et al. 2016), but also dendritic cells which is potentially motivated by the increased levels of neoantigens

at disease recurrence (Kim, Zheng et al. 2015). In contrast, primary PN GBM were found to contain significantly higher fractions of five immune cell categories compared to recurrent PN GBM, indicating a relative absence of immune infiltration in PN GBM upon recurrence.

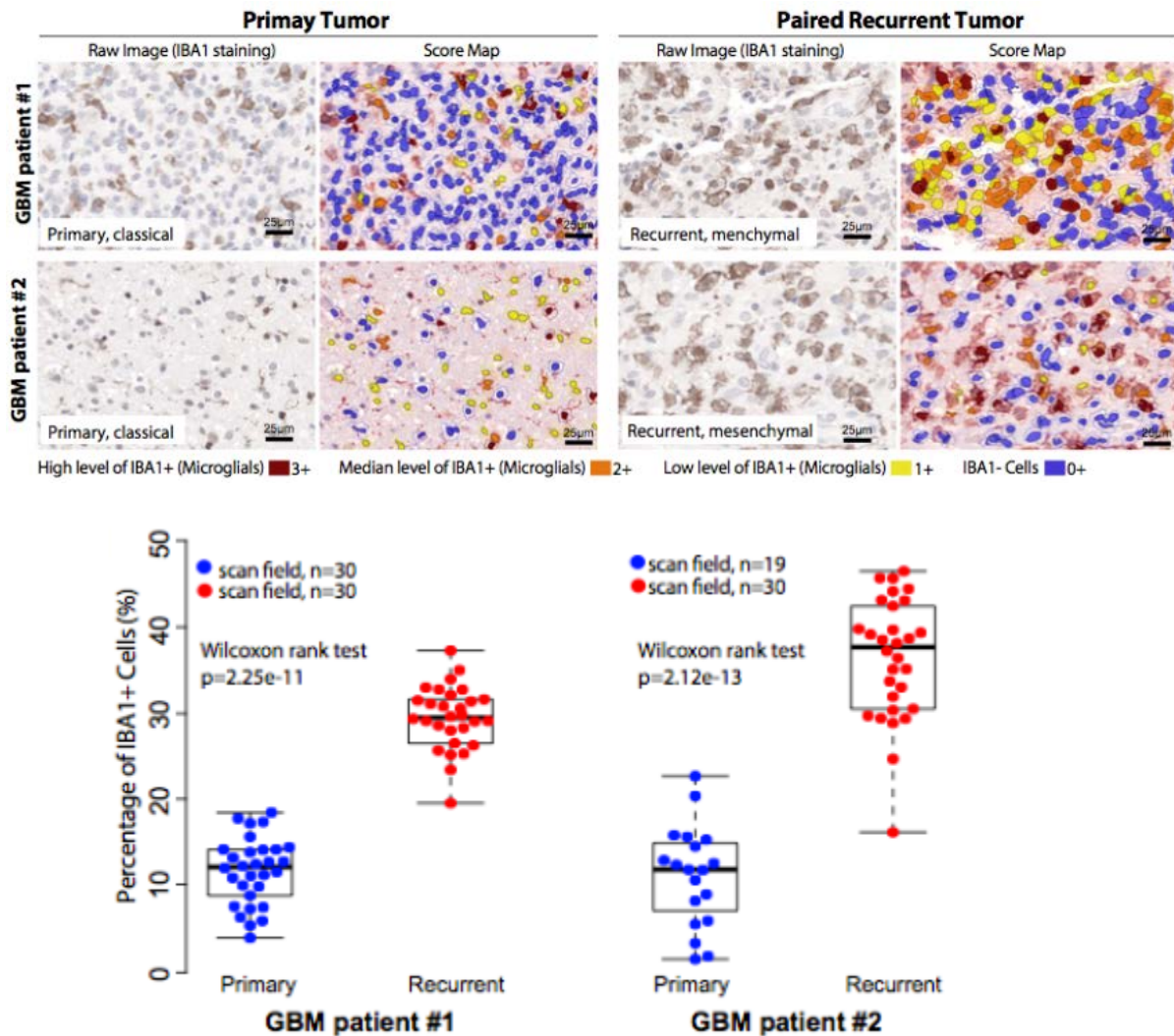


Figure 2.20 Comparison of M2 macrophage cell fractions in primary and matched recurrent tumors

(A) Representative images of IBA1 immunohistochemistry staining and corresponding score map obtained by Inform image analysis in two matched pairs of primary and recurrent GBM. Scale bar, 25 μm.

(B) Unbiased quantification of IBA1+ percentage in primary and recurrent GBMs.

[* Courtesy of Baoli Hu, Qianghu Wang]

We evaluated the effect of transcriptional class on patient survival. The analysis was performed in 50 cases for whom annotation on overall survival (OS) time and time to disease progression (PFS) were available and with high simplicity scores, indicating low transcriptional heterogeneity. We confirmed the worse prognosis for patients whose primary tumor was classified as MES on overall survival (logrank test $p=0.029$ with $HR=1.97$) (**Figure 2.21 AB**). This pattern was retained in patients whose secondary glioma was classified as MES (logrank test $p=0.09$ with $HR=1.71$) (**Figure 2.21 CD**). Consequently, cases for whom both primary and recurrent tumor were classified as MES subtype showed the least favorable outcome, suggesting an additive effect of transcriptional class at different time points (**Figure 2.21 EF** and **Figure 2.22**)

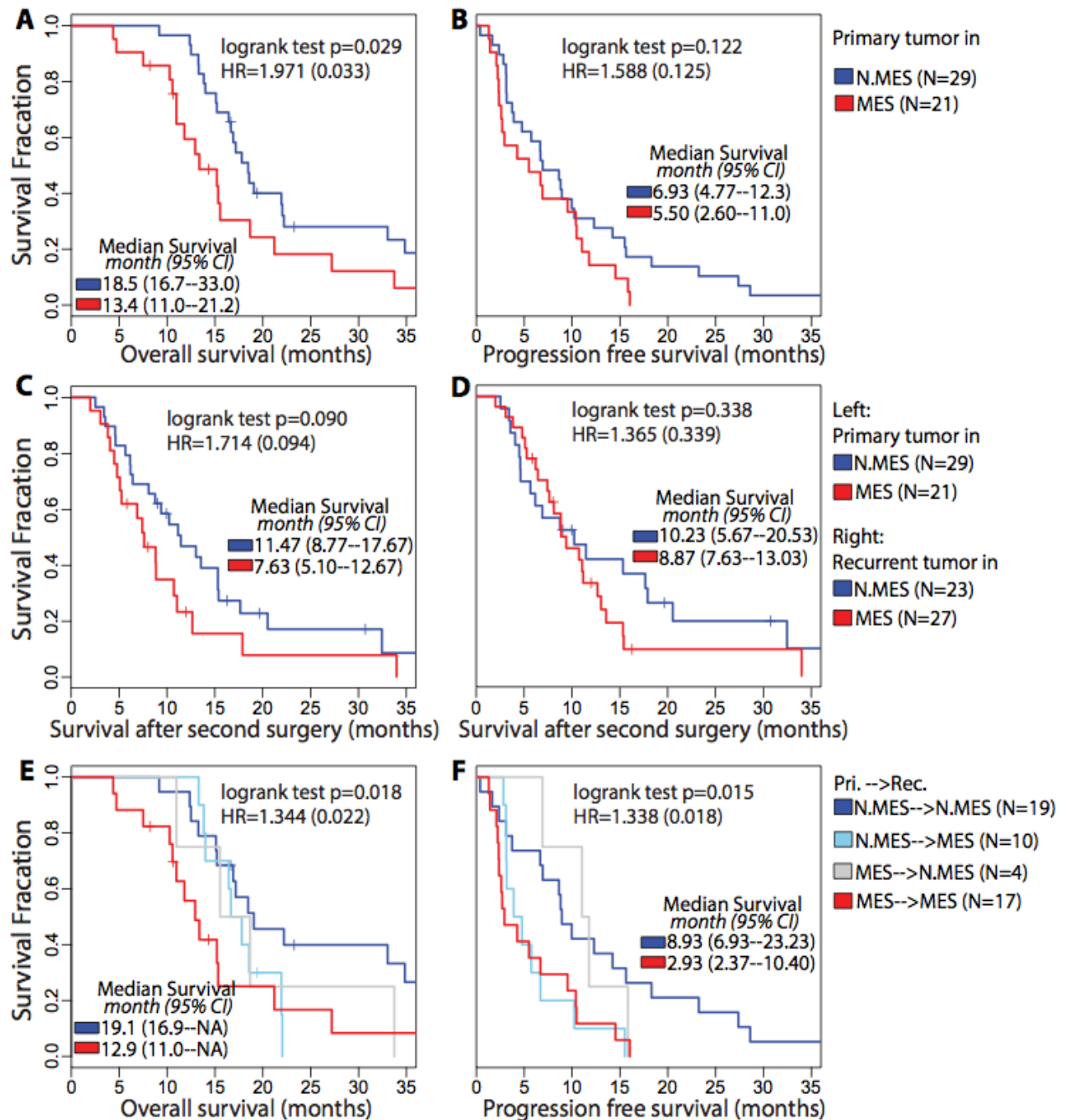


Figure 2.21 Survival analysis of paired IDH wild type GBM

(A, B) OS and PFS analyses between samples with difference primary subtype (C) Difference of survival time after secondary surgery between patients with non-MES and MES in primary tumors. (D) Survival analysis of time after secondary surgery between patients with non-MES and MES in recurrent tumors. (E, F) OS and PFS analyses between samples with different recurrent subtypes

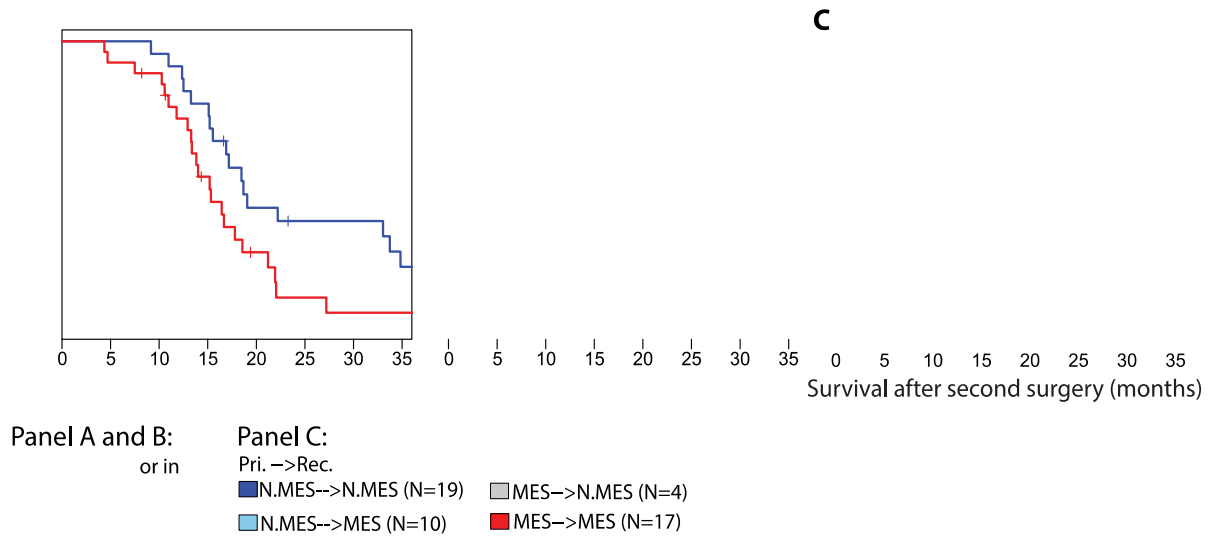


Figure 2.22 Survival analysis on MES versus non-MES patients

(A, B) Overall and progression free survival analysis between samples in different recurrent subtypes of MES versus non-MES. (C) Survival after secondary surgery comparison between different transition types.

2.3.6 Treatment-induced immunological microenvironment changes upon GBM recurrence

Temozolomide treatment of gliomas can induce hyper-mutation (Hunter, Smith et al. 2006, Kim, Zheng et al. 2015). Missense mutations may generate neoantigens that can be recognized by CD8+ T lymphocytes (Schumacher and Schreiber 2015). Using matching exome data we classified five recurrent gliomas underwent hypermutation at (≥ 400 SNVs). The predicted frequency of CD8+ T cells was significantly increased at recurrence in comparison to their primary tumors (median 7.7% vs 1.9%; Wilcoxon rank test p-value=0.008) (**Figure 2.23A**). This observation was further validated by comparing 7 hypermutated primary GBMs to 238 non-hypermutated GBMs (median 7.0% vs 0%; Wilcoxon rank test p-value=0.031) (**Figure 2.23B**). The majority (61%) of non-hypermutated primary GBMs showed predicted CD8+ T cell fractions equal to zero. This observation

suggests that patients with hyper-mutated tumors are more likely to benefit from CD8+ T cell antitumor immunity.

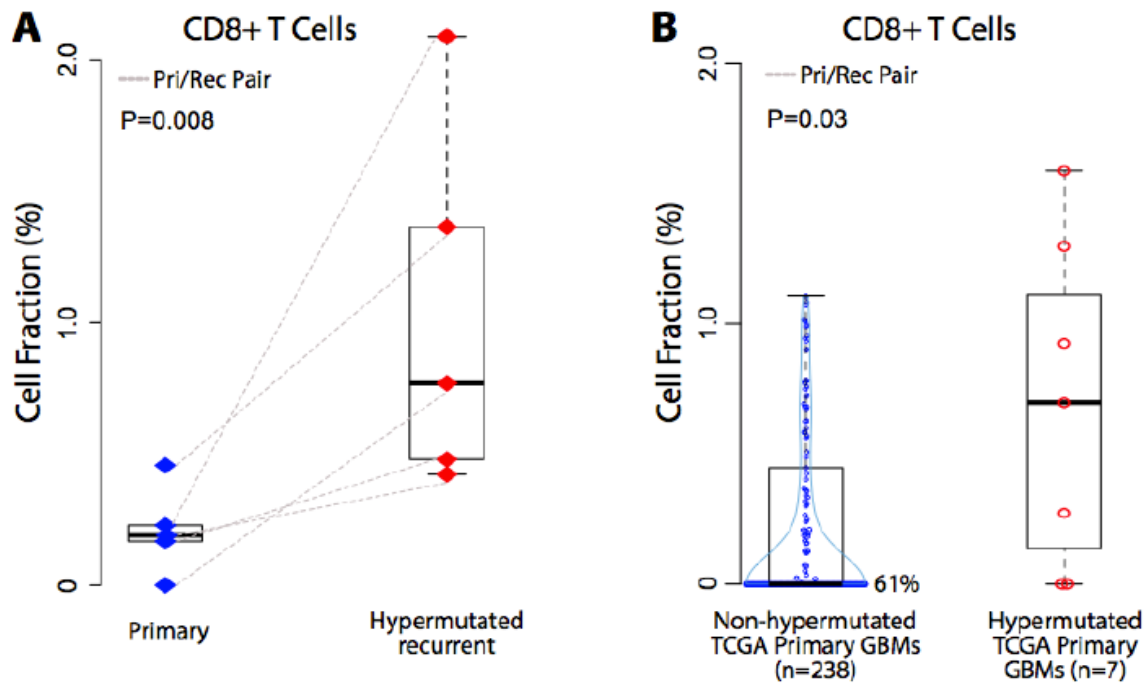


Figure 2.23 Comparison of CD8+ T cell fraction associated with gain of hyper-mutation induced by chemotherapy

(A) Blue and red diamond indicates individual primary and recurrent tumors. Dash line connects paired primary and recurrent tumors.

(B) Blue and red cycle indicates non-hyper-mutated and hyper-mutated primary samples.

Preclinical studies suggested radiation may increase the recruitment of T cells in the tumor microenvironment (Zeng, See et al. 2013, Deng, Liang et al. 2014). We compared the microenvironment of primary GBM treated with radiation therapy and separated short term relapse (PFS > 6 months, n = 27) from late relapse (PFS > 12 months, n = 21). Evaluating the presence of M2 macrophages and CD4+ T cells (CD4+ T memory resting and CD4+ follicular helper cells) based on gene signatures in 75 IDH-WT GBM patients who received radiotherapy, we observed no significant difference between their primary tumors with short-term and long-term relapse but found a significant increase at recurrence post radiation therapy (**Figure 2.24**). M2 macrophages have been speculated to play a role in resistance

to radiotherapy(Meng, Beckett et al. 2010, Ruffell and Coussens 2015) and macrophage targeting immunotherapy (Pyonteck, Akkari et al. 2013, Ries, Cannarile et al. 2014) may boost the radio-sensitivity. The increased level of CD4+ T cells at recurrence for short term relapse tumors implying the blockage of CTLA-4 as adjuvant therapy with radiation.

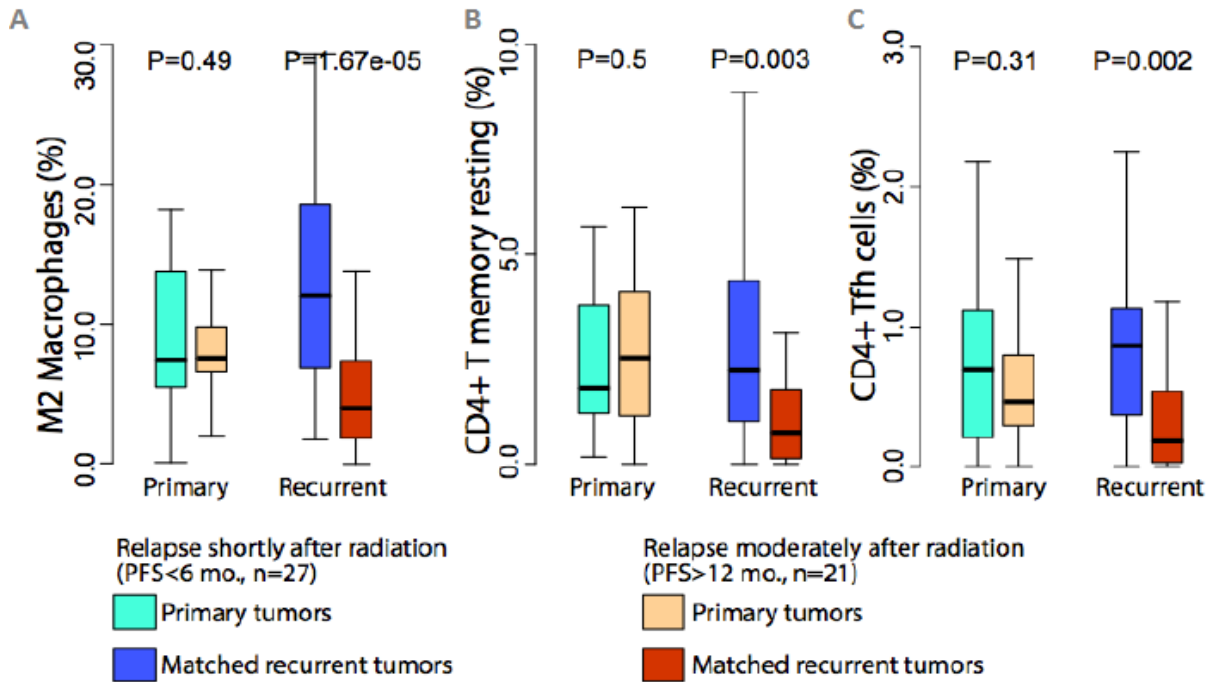


Figure 2.24 Comparison of immune cell fractions in paired samples upon relapse after different period of radiation

Sky blue/dark blue and orange/red boxplots indicate short- and long- term relapsed patients. The cell fraction of M2 macrophage(A), CD4+ T memory resting cells(B), CD4+ follicular helper cells (C) are calculated based on gene signatures implemented in CIBERSORT.

2.4 Discussion

Transcriptome profiling of tumor samples is a commonly used modality for interrogating pathway functionality and phenotype based patient classification. The transcriptional footprint left by the tumor microenvironment, which may constitute 10-80% of cells in a tumor biopsy (Yoshihara, Shahmoradgoli et al. 2013), can obscure the true activity of the signaling network (Isella, Terrasi et al. 2015, Kim and Verhaak 2015). Here, we employed in

silico methods to integrate mRNA expression profiles from glioma samples and glioma cell culture models to provide insights into glioma-intrinsic pathway activities and classification, and to deconvolute the glioma associated stroma into its immunological cellular components.

GBM expression subtype classification has emerged as an important concept to better understand the biology of this devastating disease (Huse, Phillips et al. 2011, Dunn, Rinne et al. 2012, Sturm, Bender et al. 2014). Robust classification of new GBM tumors is therefore critical to ensure consistency in reporting between different studies. The transcriptional glioma subtypes we discovered using tumor-intrinsic gene expression values strongly overlapped with the proneural, classical and mesenchymal subtypes but identified the neural class as normal neural lineage contamination. Our updated methods, released through a R-library, were found to be highly robust and provide the community with a standardized strategy for classification of gliomas.

Through re-classification of primary GBM samples from TCGA and despite using tumor-only transcripts, we observed that the mesenchymal GBM subtype associated with the presence of tumor-associated glial and microglial cells. Mesenchymal glioma cell differentiation status has been found to correlate with enrichment of macrophages/microglia (Kreutzberg 1996, Bhat, Balasubramaniyan et al. 2013). Through in silico cell type identification we additionally detected enrichment of various adaptive immunity cell types, including CD4+ T lymphocytes.

Longitudinal analysis of tumor samples is complicated by the lack of tissue collections including such pairs. Through aggregation of existing and novel datasets we compiled a cohort of 124 glioma pairs, including 91 pairs of IDH wild type tumors. Comparison of pairs of initial gliomas and first disease recurrence did not identify the trend of proneural GBM transitioning to a mesenchymal phenotype that has often been suspected (Phillips, Kharbanda et al. 2006). Mesenchymal subtype at diagnosis and at disease recurrent correlated with relatively poor outcome. The recurrent IDH wild type GBM immune

system showed fewer blood derived monocytes which may reflect lower penetration through the blood brain barrier. While the frequency of M2 macrophage/microglia was increased in recurrent mesenchymal GBM compared to primary non-mesenchymal GBM, the overall fraction of M2 macrophage/microglia remained stable. This possibly suggests that the majority of these cells are derived from resident CNS macrophages than through active recruitment from the circulation.

In summary, our study defines a new strategy to determine transcriptional subtype, and associated expression classes to the tumor-associated immuno-environment. Our findings may aid in the implementation of immunotherapy approaches (Blank, Haanen et al. 2016) in a disease type with very limited treatment options. Collectively, our results have improved our understanding of determinants of GBM subtype classification, the critical impact of the tumor microenvironment, and provide new handles on the interpretation of transcriptional profiling of glioma.

CHAPTER 3

Multi-Gene Signature for Predicting Prognosis of Patients with 1p19q Co-deletion Diffuse Glioma

(The methods and results in this chapter have been accepted by Neuro-Oncology, November, 23, 2016: Xin Hu, Emmanuel Martinez-Ledesma, Siyuan Zheng, Kim Hoon, Floris Barthel, Tao Jiang, Kenneth R. Hess, Roel G.W. Verhaak, “Multi-Gene Signature for Predicting Prognosis of Patients with 1p19q Co-deletion Diffuse Glioma”. According to the journal policy, the first author retains the right to include the published article in full or part in a dissertation.)

3.1 Introduction

According to current guidelines for brain tumors, the diagnosis of grade II-III adult diffuse glioma is assessed primarily by histopathological examination (Weller, van den Bent et al. 2014), while molecular abnormalities have been evolving as supportive markers to facilitate diagnostics and management of these patients. Diffuse gliomas with mutations in *IDH1/IDH2* may represent an entirely different type of disease than those with wild type *IDH1/IDH2*, known as IDH-wildtype glioma (Cancer Genome Atlas Research, Brat et al. 2015, Eckel-Passow, Lachance et al. 2015, Ceccarelli, Barthel et al. 2016). Within the group of IDH-mutant glioma, presence of 1p/19q co-deletion (IDH-mutant-codel glioma) may present an additional prognostic marker, separate from IDH-mutant glioma with intact 1p/19q chromosome arms (IDH-mutant-non-codel glioma). The unique characteristics of IDH-mutant-codel glioma led to recognition of this subtype in the 2016 World Health Organization Classification System of Tumors in Central Nervous System (Louis, Perry et al. 2016). A series of clinical trials revealed that standard radiation therapy followed by adjuvant chemotherapy with procarbazine, lomustine, and vincristine (PCV) delayed disease

progression and increased survival in patients diagnosed with anaplastic oligodendroglioma. Interestingly, patients harboring the 1p/19q co-deletion demonstrated better response to additional chemotherapy than patients whose tumor was 1p and 19q wildtype (Cairncross, Wang et al. 2013, van den Bent, Brandes et al. 2013, Dubbink, Atmodimedjo et al. 2015, Buckner, Shaw et al. 2016). Approximately 85% of diffuse gliomas with the 1p/19q co-deletion in the TCGA cohort have an oligodendroglial component and could be classified as either grade II (low-grade) or grade III (high-grade) glioma according to the World Health Organization (WHO) system. Patients with histologically and molecularly similar glioma may show heterogeneous clinical characteristics and response to treatment, which suggest that additional factors may determine clinical behavior and prognosis. The management of low-grade diffuse glioma, including the components necessary for diagnosis, the role of surveillance, and the nature of surgical intervention, radiation therapy, and chemotherapy, lacking conclusive evidence to support best practices, remains controversial (Zadeh, Khan et al. 2015). Phenotypic and genomic inter-tumor heterogeneity of 1p/19q co-deleted gliomas may account for inconsistency between clinical observations.(Figarella-Branger, Mokhtari et al. 2014, Alentorn, Dehais et al. 2015, Zadeh, Khan et al. 2015) Understanding of the biological components associated with clinical and phenotypic heterogeneity will aid improved disease staging before treatment and tailoring of appropriate therapeutic regimens.

Molecular markers such as *IDH1/2* mutation, promoter methylation of *MGMT*, *ATRX* and *EGFR* gene mutations, and *BRAF* fusion transcripts or point mutations are increasingly recognized as an integral aspect of the clinical management of adult diffuse glioma patients (Siegal 2015). There may be a role for molecular markers in risk classification of 1p/19q co-deletion glioma patients. High-risk patients could receive aggressive treatment with adjuvant chemotherapy, whereas low-risk patients might forgo intensive adjuvant chemotherapy. Several independent studies have demonstrated that gene expression profiling can be applied to identify biomarkers and molecular subtypes of glioma associated with certain

clinical outcomes.(Freije, Castro-Vargas et al. 2004, Yan, Zhang et al. 2012) However, the prognostic profiles these studies identified have few genes in common, and the reported gene signatures are based on survival information and gene expression patterns from histopathological glioma classes. Such gene signatures might not accurately predict survival for patients whose gliomas harbor the 1p/19q co-deletion, as the mRNA expression patterns and underlying biological characteristics of this subgroup may be intrinsically different from those gliomas without the 1p/19q co-deletion, as is implied by its distinct favorable clinical outcomes (Huang, Hsu et al. 2015).

In an integrative analysis of newly diagnosed diffuse glioma patients, we observed that glioma patients harboring the 1p/19q co-deletion exhibit heterogeneous clinical outcomes (Cancer Genome Atlas Research, Brat et al. 2015),(Ceccarelli, Barthel et al. 2016). In the present study, we sought to identify molecular markers associated with the diverse clinical outcomes in this subset of glioma patients.

3.2 Methods

3.2.1 Process of the datasets

Our approach to perform gene signature selection and validation for classification using normalized gene expression datasets is summarized in **Figure 3.1**. We curated gene expression and sample information from five publicly available glioma datasets whose tumors were assessed with microarrays (Gravendeel, Kouwenhoven et al. 2009, Madhavan, Zenklusen et al. 2009, Yan, Zhang et al. 2012) or RNA-Seq (Bao, Chen et al. 2014, Cancer Genome Atlas Research, Brat et al. 2015), summarized in Table S1. Normalized RSEM values for TCGA glioma samples were retrieved from the LGG-GBM project data portal (https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015). RPKM values for CGGA1 RNA-seq data (Bao, Chen et al. 2014) were calculated using in house software (PRADA)(Torres-Garcia, Zheng et al. 2014). We used Affymetrix Human Genome U133 Set

annotation data provided by the Bioconductor library hgu133plus2.db hgu133a.db, and Illumina HumanHT-12 WG-DASL to convert microarray probe signals to gene expression levels. Multiple probe sets were mapped to a single gene by averaging the signals. In addition, we curated and combined two gene expression datasets measured by microarrays and used this as a second validation dataset (Guan, Vengoechea et al. 2014, Weller, Weber et al. 2015) (**Table 3.1**). Affymetrix CEL files in training- and first validation dataset but not the second validation dataset were normalized together. Overall survival time (OS) was defined as time from diagnosis to death; the patients who were alive by the end of each study period were censored at the time of last follow-up. Tumor grade was also established at primary diagnosis.

The microarray expression values as retrieved from GEO and the RPKM/RSEM values resulting from RNA-sequencing were log2 transformed, respectively. Since each expression dataset contained a slightly different set of genes, we merged the expression datasets and retained the genes commonly present in all datasets, then fit elastic net to perform feature (i.e., gene) selection. Each dataset was globally scaled across all the samples and genes and to obtain a mean of 0 with standard deviation of 1 and an empirical Bayes framework (combat) was applied to adjust for batch effects on the merged dataset (Johnson, Li et al. 2007). Using random sampling, we then assigned glioma samples with the 1p/19q co-deletion to the training dataset or the validation dataset.

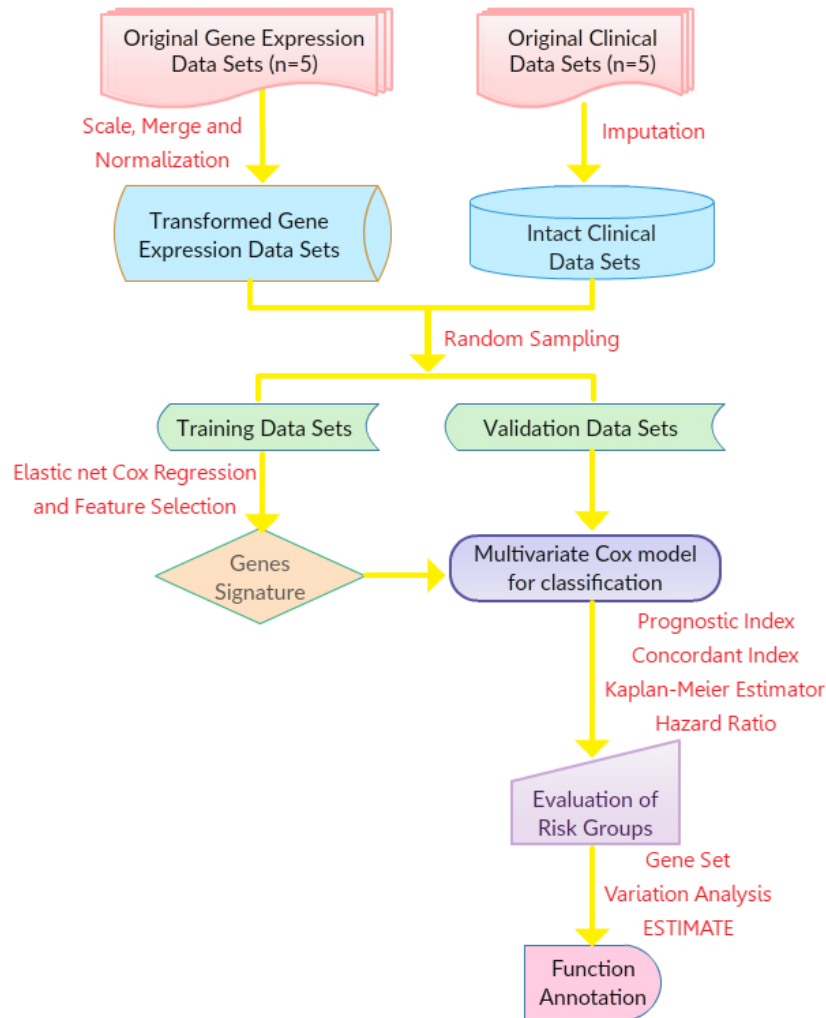


Figure 3.1 Workflow for identification and validation of prognostic gene signatures using the elastic net

Each gene expression data set was scaled separately and batch effects were calibrated using ComBat after merging of the five data sets. The gene signature was selected based on the optimal value of the regularization parameter λ determined from cross-validation in the training model. The training model was applied to the validation dataset to derive the risk scores, which classified the patients in the validation dataset into high- and low- risk groups for assessment of prognostic power.

Table 3.1 Clinical characteristics of glioma patients harboring 1p/19q co-deletion

Characteristic	No. of patients
Age, years (range =17, 87; median = 43)	
< 43	175
> 43	179
43	13
Not available	7
Gender	
Male	217
Female	148
Not available	9
Time to death/last follow-up for event-free subjects: median 22.7months, range 0-182.3 months	
Kaplan-Meier estimated median survival time : 75.7 months (range 1-248 months)	
WHO grade	
II	179
III	159
IV	16
Not available	20
Histologic subtype	
Oligodendroglioma	217
Astrocytoma	32
Oligoastrocytoma	96
GBM	9
Not available	13

Table 3.1 Clinical characteristics of glioma patients harboring the 1p/19q co-deletion

Abbreviations: WHO, World Health Organization; GBM, glioblastoma multiforme.

Note: The dataset of glioma patients with the 1p/19q co-deletion was curated and selected based on copy number variation information from The Cancer Genome Atlas (15 death events / 151 patients), the Chinese Glioma Genome Atlas (22 death events / 81 patients), Gravendeel et al. (35 death events / 40 patients), and Rembrandt et al. (49 death events/61 patients), Weller et al. (4 death events / 31 patients) (Weller, Weber et al. 2015) and Guan et al (1 death event / 10 patients)(Guan, Vengoechea et al. 2014).

3.2.2 Predicting 1p/19 status using gene expression

DNA copy number profiles determining 1p/19q status were available for a subset of the cohort (**Table 3.2**). For cases described below where copy number profiles were absent, we applied a Gaussian window smoothing algorithm on the expression dataset to infer the pattern of chromosome arm sized copy number variations (CNVs). Using the expression values of the genes located at Chr-1p and Chr-19q sorted by their genomic locations from start to end, we used a sliding 100 gene window to determine chromosome arm wide 1p/19q expression levels. We applied the following equation to the resulting gene-specific expression patterns to determine 1p/19q status: where (i) is the estimated copy number (relative value) of sample k at i th gene in genomic-ordered gene list, g_j is the j th gene in the genomic-ordered gene list, and (g_j) is the relative gene expression value of that gene in sample k . Note that the estimated 1p/19q status is often consistent with the chromosome centromere borders, with increased or decreased values within specific chromosomes, suggesting that it accurately represents chromosomal changes. We then applied hierarchical cluster analysis to $CNV_k(i)$ values to assign all the samples from each dataset into a group reflecting 1p/19q co-deletion, and another group with 1p/19q wild type copy number.

Data_resource	Platform	Normalization	CNV_available	Sample_size	Events
TCGA (The Cancer Genome Atlas)	RNA-Seq	Downloaded RPKM from TCGA portal	Yes	151	15
CGGA_1 (PMID:25135958)	RNA-Seq	Calculated RPKM using PRADA	No	63	15
CGGA_2 (PMID: 23090983)	Microarray	Obtained from CGGA database, Probe intensities normalized by GeneSpring GX 11.0	Yes	18	7
Rembrandt Brain database (PMID: 19208739)	Microarray	Quantile normalization using "affy" and annotation using "hgu133plus2.db"	No	61	49
Gravendeel (PMID: 19920198)	Microarray	Quantile normalization using "affy" and annotation using "hgu133plus2.db"	Yes (Partial)	35	40
Weller German Glioma Network (PMID: 25783747)	Microarray	Quantile normalization using "affy" and annotation using "hgu133plus2.db"	Yes	61	49
DASL (PMID: 24614622)	Microarray	Normalized and annotated using "illumina HumanHT-12 WG-DASL"	No	35	40

Table 3.2 Data sources for gene expression and CNV in glioma patients harboring 1p/19q co-deletion

3.2.3 Correlation of somatic mutations and clinical outcome

We applied the Kaplan-Meier estimator to assess the prognostic value of the most prevalent mutations in glioma, *CIC*, *FUBP1*, *NOTCH1*, and *PIK3CA* mutations, on overall survival (OS). Two-sided log-rank tests were applied to examine the differences of overall survival (OS) between the patients with and without any of these mutations. *P* values < 0.05 were considered statistically significant. We conducted this analysis using the TCGA dataset, which included mutation information. Although OS might be affected by treatment bias at the time of tumor progression, OS data are generally more accurate than PFS data; therefore, we used OS to represent clinical outcomes that more accurately reflect disease aggressiveness in each glioma patient.

3.2.4 Gene signature selection and risk based classification

Using the training set, we first pre-filtered the genes based on Wald p-values generated from univariate Cox models and selected the 1,000 most significant genes, then applied the Cox proportional hazards model with elastic net penalty for variable selection. The univariate and multivariable Cox models were built using the R package “survival”; the elastic net regression (i.e., the combination of L1 regularization and L2 regularization) was performed using R package “glmnet.”(Friedman, Hastie et al. 2010, Hughey and Butte 2015) The penalty parameter λ was chosen based on 3-fold cross validation within the training set, which produced the minimum mean cross validated error for the Cox model. Thus, we used shrinkage-based regularization combined with a univariate Cox model to obtain the gene signature.

Using the training dataset, we fit a multivariable Cox proportional hazards model with the genes identified using the above penalty-based method. We then computed a prognostic index for each individual patient in the validation set through multiplying their gene expression values by the corresponding regression coefficients estimated from the training

data. This resulted in a risk score for each patient in the validation dataset according to a linear combination of the mRNA expression level from the validation data weighted by the multivariable Cox model–derived regression coefficients from the training data. We calculated the concordance index (C-index) and its standard error for the gene signature, age, grade, and gene signature combined with age and grade, using the R package “survcomp” (Schroder, Culhane et al. 2011). We also calculated the hazard ratios (HR) and their 95% confidence intervals between two groups of patients with risk scores above and below the median risk score computed from the training dataset.

3.2.5 Association of risk classification and clinical outcome

We divided the patients from the validation dataset into high- risk and low-risk groups based on their risk scores derived from the linear prediction, using the median risk score in the training set as cut-off value. We used the Kaplan–Meier estimator and the two-sided log-rank test to evaluate the differences in overall survival (OS) between the high- and low-risk patients. To examine the robustness of the risk-based classification using selected genes, we divided the patients into subgroups using a series of different risk scores as cut-off values and evaluated the difference of OS between high- and low-risk groups using the Kaplan–Meier estimator and hazard ratio (HR). To further investigate the trend of the OS pattern to align with the predicted risk scores, we fit a smoothing spline to ascertain the association of risk scores with the OS of the patients in the validation dataset.

3.2.6 Gene Set Variation Analysis of Associated Genes and Top Gene Ontology Terms

We first used a Student’s t-test to identify the genes differentially expressed between the high- and low-risk groups, only including genes with a p-value of less than 0.05 and an absolute difference in median gene expression of 0.4. We then mapped the gene ontology (GO) terms of the corresponding 260 genes that present the most variance in expression

between the two risk groups. We applied gene set variation analysis (GSVA)(Hanzelmann, Castelo et al. 2013) to obtain the enrichment scores for each gene set that corresponding to the GO terms containing those genes in all the patients.

3.2.7 Evaluation of Tumor Purity with ESTIMATE gene signatures

We inferred tumor purity of each sample using ESTIMATE (Yoshihara, Shahmoradgoli et al. 2013), which reflects the enrichment of stromal and immune cell gene signatures in a transcriptional profile.

3.3 Results

3.3.1 Effects of somatic mutations on patient outcome

Recent studies by TCGA and others have revealed genes frequently mutated in IDH-mutant and 1p/19q co-deleted glioma, including the 1p gene *FUBP1* and the 19q gene *CIC*. Mutations in these genes fulfill the classic Knudson tumor suppressor two-hit model in which one allele is lost and the second is inactivated through somatic mutation. Thus, gliomas carrying *CIC/FUBP1* mutations are candidates for further progressing disease. To test this hypothesis, we performed univariate survival analyses of 151 diffuse IDH-mutant-codel gliomas from TCGA. We found significant correlation between overall survival (OS) and the presence of the *FUBP1* mutation (n = 40 of 151; log-rank test p-value = 0.05) but not the *CIC* mutation (n = 71 of 151; log-rank test p-value = 0.71). No associations were observed for other gene mutations frequently detected in IDH-mutant-codel gliomas such as in *NOTCH1* (23/151; log-rank test p-value = 0.46), or *PIK3CA* (20/151; log-rank test p-value = 0.06). Despite the relatively small numbers of death events in the TCGA dataset (15 of 151), our results suggest that these mutations do not significantly affect disease progression or clinical outcomes in IDH-mutant-codel glioma patients.

3.3.2 Constructing a gene expression data set of 1p/19q co-deleted glioma

Since only *FUBP1* somatic point mutations showed a significant association with patient outcome, we set out to identify a gene expression signature with the potential to identify high-risk IDH-mutant-codel patients. We curated gene expression and clinical information from seven publicly available datasets of adult diffuse glioma patients whose tumors were assessed by microarray or RNA-Sequencing. Where available, we used annotation on 1p/19q co-deletion available per the respective publications or data from genome-wide DNA copy number profiling to identify IDH-mutant-codel cases. As this data was unavailable on some data sets, we applied a Gaussian window smoothing algorithm to infer the signal of large scale CNVs for each sample. By suppressing individual gene-specific expression patterns, and averaging relative expression values over large genomic regions, we selected the samples that harbor the 1p/19q co-deletion based on the hierarchical clustering of the CNVs estimated from each dataset (**Figure 3.2**). We found that our method could predict 1p/19q codel status with high accuracy in the TCGA data set (sensitivity=0.97, specificity=0.97, Mathews correlation coefficient (MCC) = 0.94). Using the window sliding method in the Weller et al dataset, which included CGH-based 1p/19q status, we obtained a specificity of 0.91. We found 411 of 2,231 gliomas to contain the 1p/19q co-deletion and retained those samples with survival data, resulting in two combined datasets consisting of 374 (n = 333 for first dataset, n = 41 for second dataset) clinically annotated 1p/19q codel IDH mutant glioma gene expression profiles (**Table 3.1**). The patient characteristics of the 1p/19q co-deletion glioma cohort are summarized in **Table 3.1**. The patients' median age at diagnosis was 43 years (range, 17-87 years) with Kaplan-Meier estimated median survival time of 75.7 months (range, 1-248 months), including 121 events. Amongst censored cases, the median follow up was 22.7 months.

After performing scale normalization using the 13,345 genes common to all data sets, we found no distinct clustering in any of the 5 gene expression datasets in the training- and

first validation set, suggesting that any platform or batch variance across the different datasets had been mostly eliminated (**Figure 3.3**).

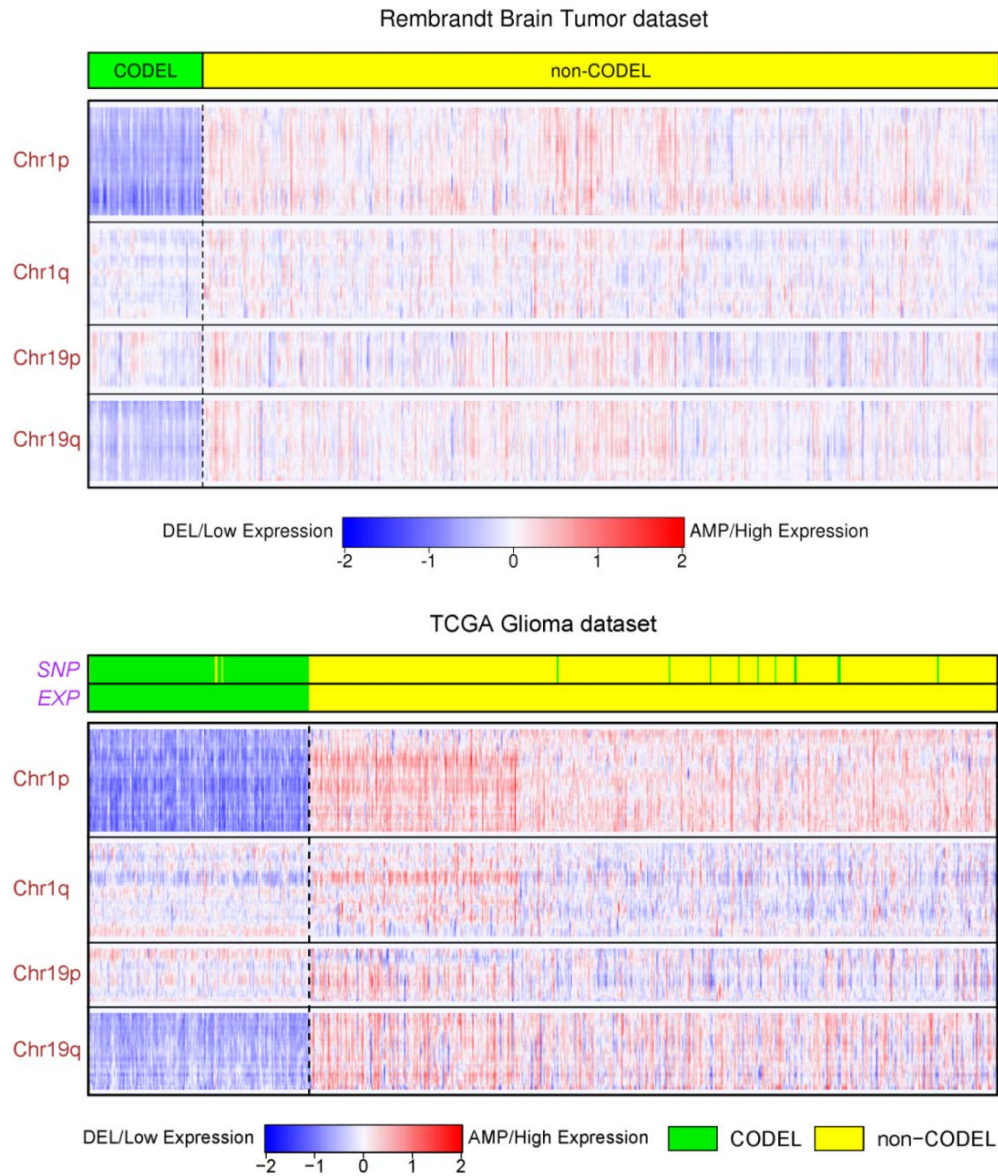


Figure 3.2 Co-deletion of 1p/19q inferred by gene expression profiling

Normalized gene expression levels of chromosome 1 and chromosome 19. The top panel shows the Rembrandt glioma dataset (n = 69 with co-deletion, n = 550 total). The bottom panel shows the TCGA dataset. The top bar denotes the CODEL status based on SNP6 DNA copy number arrays, the bottom bar denotes the CODEL status inferred by our method using gene expression data (n = 162 with co-deletion, n = 667 total). The green bar denotes the samples classified as CODEL, the yellow bar denotes the samples classified as non-CODEL. The averaged expression level is shown in red-white-blue scale.

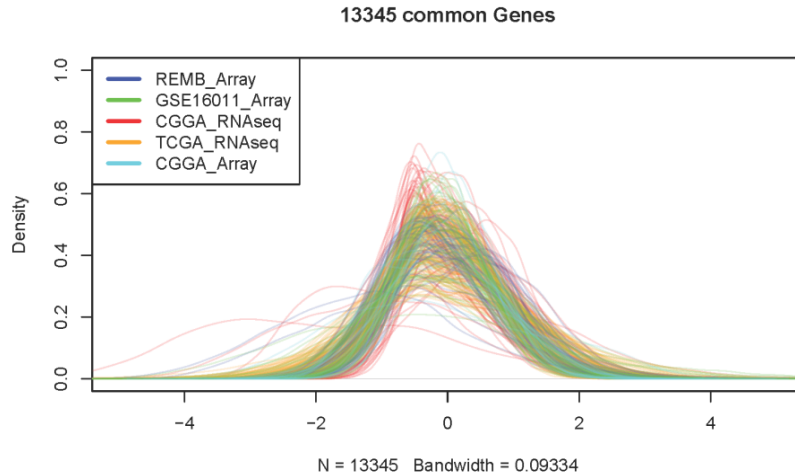


Figure 3.3A Distribution of normalized gene expression datasets from different data sources

The density plot represents the transformed and merged datasets of gliomas with the 1p/19q co-deletion used for featured gene selection and prediction modeling. Datasets were curated from TCGA_RNAseq (151, events = 15), Gravendeel et al (GSE16011, NETH) (40, events = 35), Rembrandt et al (REMB) (61, events = 49), CGGA1_RNAseq (63, events = 15), CGGA2_Microarray (18, events = 7).

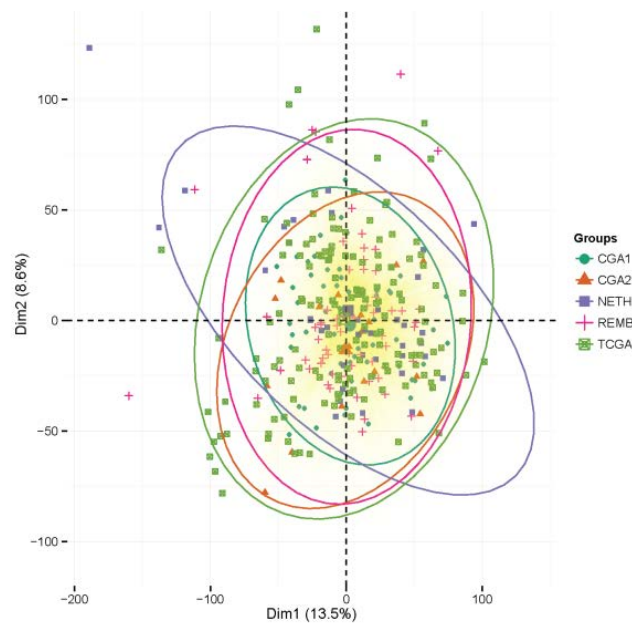


Figure 3.3B The Principal Component Analysis (PCA) of transformed datasets

All samples in the transformed gene expression dataset were plotted using the top principal components (PCs) which captured the majority of the variance in the input dataset. Each sample is denoted by a single dot. And each colored symbol represents a data source as labelled on the right. The inertia ellipses represent 95% of the inertia of the corresponding dataset.

3.3.3 Identification of a 35-gene signature associated with overall survival

We divided 333 patients from the first combined dataset including five original datasets (**Table 3.1**) into training and validation datasets by event-frequency matched random sampling, so that each consisted of comparable numbers of astrocytic, oligodendrocytic, and mixed histological tumor subtypes. We also balanced for chemotherapy and radiotherapy treatment. Through the controlled randomized sampling process, the training dataset (n=170, death events = 64) included 105 samples with treatment annotation of whom 26 patients received radiotherapy and 26 patients received chemotherapy, with eleven of those cases undergoing both radio- and chemo-therapy. The remainder of the 105 patients (n = 64) were surgically debulked without further treatment. The first validation dataset (n=163, death events = 57) consisted of 84 samples with treatment information of whom 22 patients received radiotherapy and 18 patients received chemotherapy. To build the training model, we selected the 1,000 genes with the most significant linear correlation with overall survival (OS). We then applied a linear regression function that fits the Cox model regularized by elastic net penalty (**Figure 3.4**), to select 35 genes as active covariates of the Cox model to assess the prognostic index in the validation sets (**Table 3.3**).

hgnc_symbol	entrezgene	band	start_position	end_position	P.value_Diff_Expression
ADIG	149685	q11.23	38581195	38588463	8.97E-01
ARHGEF15	22899	p13.1	8310241	8322516	1.28E-01
CCDC58	131076	q21.1	122359591	122383231	1.36E-03
CCKAR	886	p15.2	26481400	26490462	2.28E-05
CDC37	11140	p13.2	10391134	10420121	1.25E-03
CRYBA2	1412	q35	218990189	218993421	4.51E-01
DOCK4	9732	q31.1	111726110	112206411	1.83E-03
EGR1	1958	q31.2	138465490	138469315	2.91E-03
EGR3	1960	p21.3	22687659	22693302	2.49E-06
EXOC8	149371	q42.2	231332753	231337852	5.86E-05
FBXO36	130888	q36.3	229922302	230013109	2.28E-02
FOXD2	2306	p33	47436017	47440691	1.40E-05
FRZB	2487	q32.1	182833275	182867162	4.48E-05
GAS2	2620	p14.3	22625642	22813055	1.29E-08
GATA4	2626	p23.1	11676959	11760002	3.24E-02
GRK5	2869	q26.11	119207589	119459742	1.54E-03
GSTO2	119391	q25.1	104268873	104304945	1.83E-02
HSD17B2	3294	q23.3	82035004	82098534	9.55E-04
IARS2	55699	q41	220094102	220148041	1.25E-02
IFT88	8100	q12.11	20567069	20691437	4.13E-05
IL32	9235	p13.3	3065297	3082192	1.11E-06
ITIH3	3699	p21.1	52794768	52809009	3.41E-02
LHCGR	3973	p16.3	48686775	48755730	1.15E-09
MAP9	79884	q32.1	155342658	155376970	2.25E-03
MIPEP	4285	q12.12	23730189	23889419	2.40E-02
MTMR6	9107	q12.13	25246201	25288009	5.34E-04
PF4V1	5197	q13.3	73853189	73854155	1.10E-02
POU4F2	5458	q31.22	146638893	146642474	3.91E-06
PTGDR	5729	q22.1	52267713	52276724	1.75E-03
RPS8	6202	p34.1	44775251	44778779	2.48E-06
SDHC	6391	q23.3	161314257	161375340	3.66E-02
TFAP2B	7021	p12.3	50818723	50847613	5.33E-03
TRAT1	50852	q13.13	108822698	108855005	8.93E-04
TRH	7200	q22.1	129974305	129977938	8.91E-04
TRPA1	8989	q21.11	72019917	72075617	1.17E-06

Table 3.3 Annotation of 35 gene signature for prediction

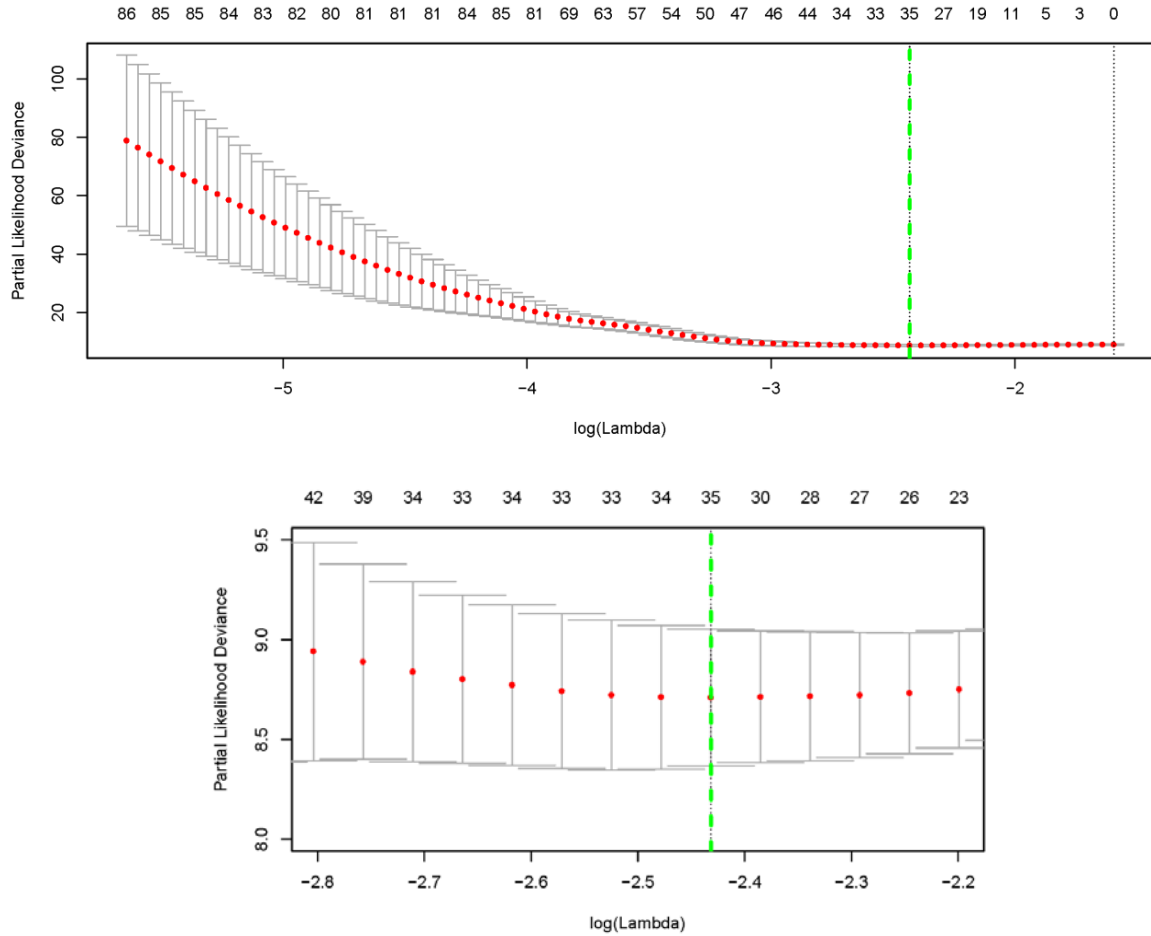


Figure 3.4 Partial likelihood deviances as function of regularization parameter λ for 3-fold-cross validation in the training dataset

The penalization criteria cross-validated error rates in the training dataset were plotted. Each red point represents a λ value along regularization paths for Cox model; with error bars in gray lines representing confidence intervals for the cross-validated error rate (Partial Likelihood Deviance). The left vertical line in green marks the minimum error while the right vertical line denotes the largest value of λ so that the error is in one standard deviation from the minimum. The top number of the plot labeled the size of each model upon shrinkage and selection based on linear regression. The bottom panel illustrates the zoom in plot.

To assess the performance of the signature genes as classifiers, we computed a linear combination of the 35 genes using the coefficients of multivariable Cox regression derived from the training dataset to calculate the risk scores for the patients in the first validation dataset. Using median risk score amongst samples from the training dataset as

the cutoff value to divide the first validation dataset into high risk and low risk groups, we found a significant difference in OS time between the two groups (log-rank test p-value = 0.014) (**Figure 3.5A**). The high risk group associated with HR of 2.03 (95% confidence interval, 1.14-3.60). The median OS duration was 75.2 months for the patients (n=84) with high-risk prognostic indices and 118.2 months for those (n=79) with low-risk indices. We also found a significant difference in OS when dividing the first validation set using the upper and bottom quartile risk scores (log-rank test p-value = 0.0085; HR = 2.9 (95% confidence interval, 1.3- 6.6), **Figure 3.5B**).

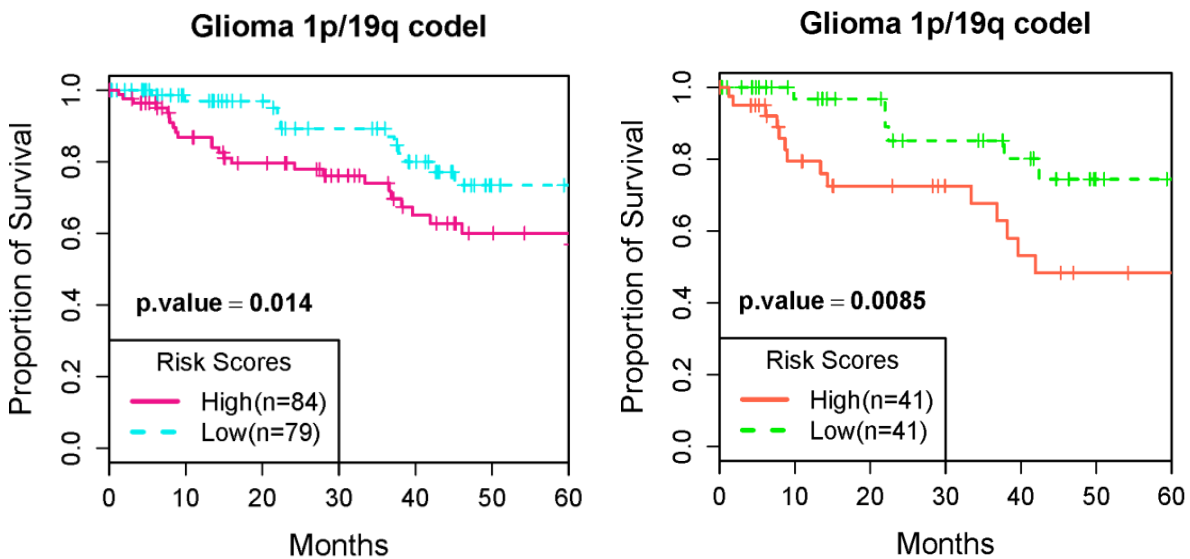


Figure 3.5 Kaplan-Meier survival analysis of glioma patients harboring 1p/19q co-deletion according to 35-gene signature derived risk scores

(A) Kaplan-Meier cumulative survival curves for first validation set glioma patients with 1p/19q co-deletion tumors, classified in two groups based on 35-gene signature derived risk scores. The survival of the high-risk patients (solid line) was significantly worse than that of the low-risk patients (dashed line; p-value = 0.014, log-rank test; hazard ratio = 2.02).

(B) Kaplan-Meier survival curves for the top and bottom quartile of glioma patients with the 1p/19q co-deletion in the first validation dataset based on 35-gene signature derived risk scores are shown. Patients in the first validation dataset were divided into high- and low-risk subgroups according to the risk scores derived from the multivariable Cox model. The survival of the high-risk patients (upper quartile, red line) was significantly worse than that of the low-risk patients (bottom quartile, green line; P = 0.0085, log-rank test).

We evaluated several different arbitrary risk score cutoffs to define high- and low-risk patient groups and found that regardless of cutoff value chosen, the overall survival of the low-risk group show significantly better than that in the high-risk group (**Figure 3.6** and **Figure 3.7**).

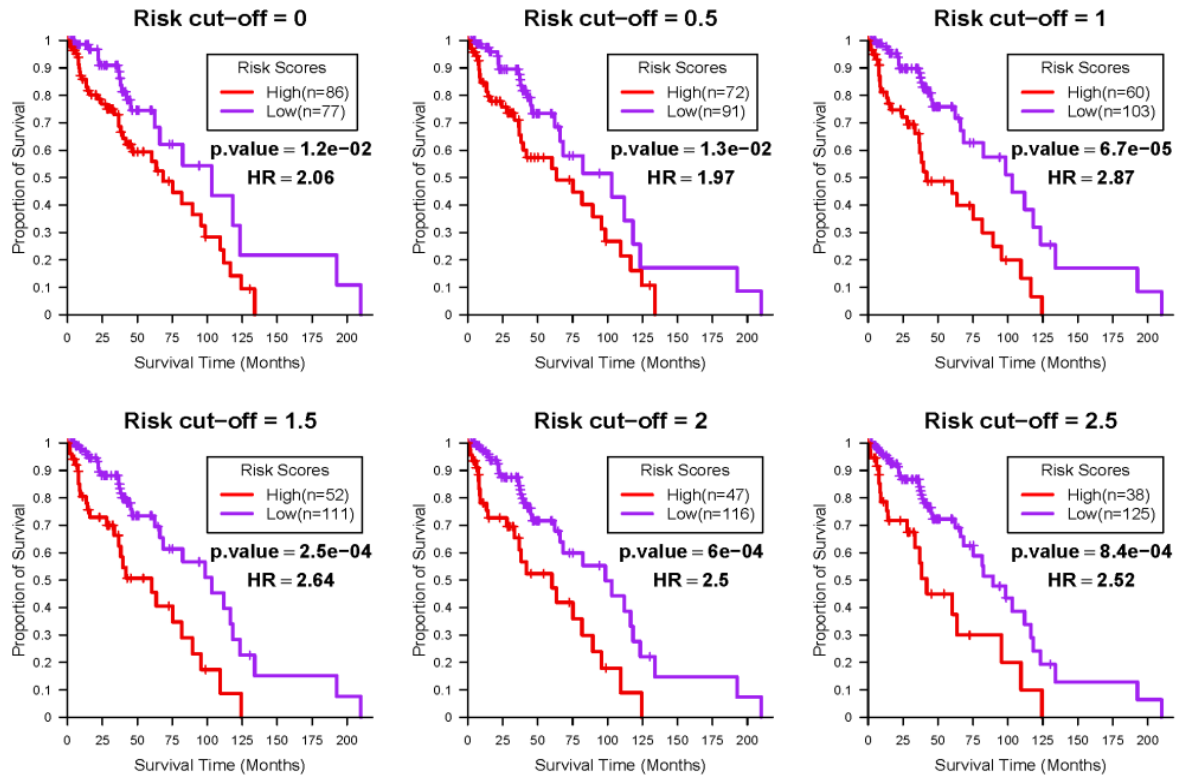


Figure 3.6 Consistent prediction of prognosis in high- and low-risk groups

Patients were divided into high- / low-risk groups using different risk scores (prognostic index) after adjustment for age and stratification by grade. The cut-off value (0, 0.5, 1, 1.5, 2, and 2.5) was assigned to each test respectively. Survival curves for the high-risk patients (red) and low-risk patients (purple) are shown. The log-rank P values and HRs for each subgroup are labeled in each plot.

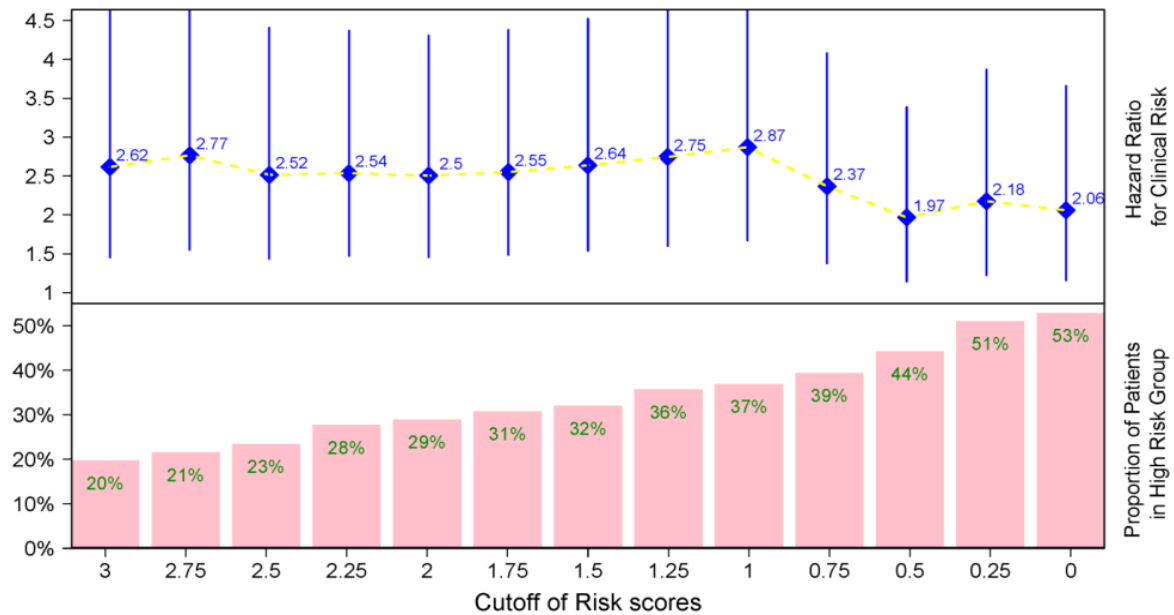


Figure 3.7 Hazard Ratios (HR) with 95% CIs for gene signatures in high- versus low-risk groups

The x-axis indicates the risk scores (derived from 15-year survival probabilities) used as cutoff values. The pink bars (bottom) indicate the proportion of patients in the high-risk group from the first validation dataset; blue diamonds (top) denote HRs adjusted for clinical risk based on 15-year survival probability and the blue vertical lines denote the corresponding 95% CIs.

We then computed risk scores on 41 samples with predicted 1p/19q co-deletion from two datasets that were not included in the training set (**Table 3.1**). There was no significant difference in overall survival between the two resulting groups in this second validation set (log-rank test p-value=0.25, HR = 2.7) which is likely due to the low number of death events (n = 5) (**Figure 3.8A**). Combining both validation sets into a single analysis showed a highly significant difference in survival between high and low risk classes (log-rank test p-value=0.00058, HR = 2.65) (**Figure 3.8B**).

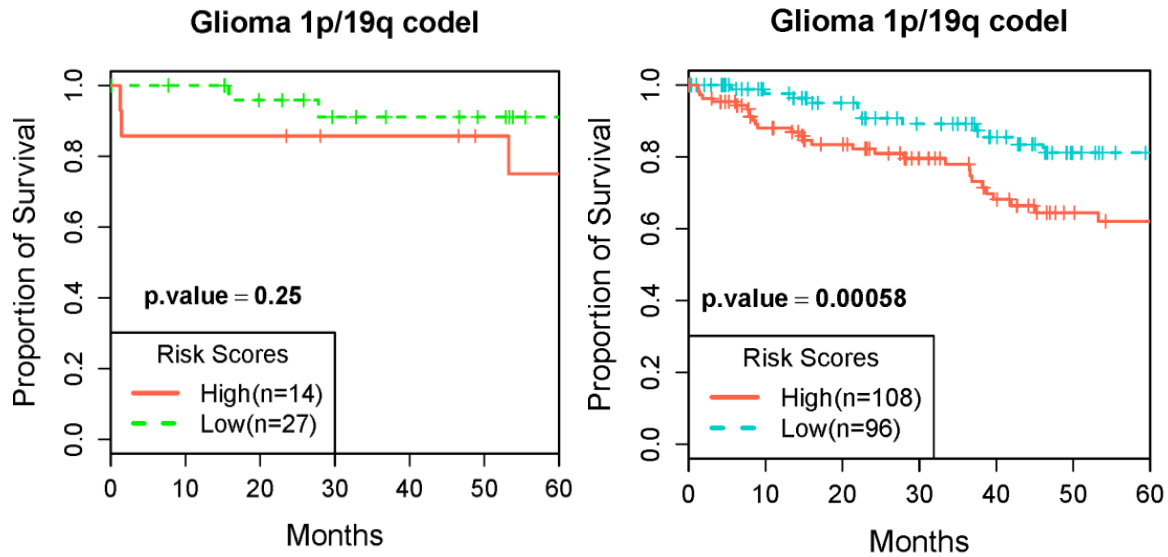


Figure 3.8 Kaplan-Meier survival analysis of co-deletion glioma according to 35-gene signature derived risk scores

(A) Kaplan-Meier survival curves for the glioma patients with 1p/19q co-deletion an independent dataset compiled from publications by Guan et al and Weller et al. predicted according to gene expression in an independent second validation dataset based on the median cut-off of 35-gene signature derived risk scores are shown. Patients in the second validation dataset were divided into high-risk (red line) and low-risk (green line) subgroups according to the risk scores and cutoff based on the median value in the training cohort. ($P = 0.25$, log-rank test; hazard ratio = 3.30).

(B) Kaplan-Meier survival curves for 1p/19q co-deletion glioma patients from combined first and second validation dataset, separated into two groups based risk score. The survival of the high-risk patients (red line) shown significantly worse than that of the low-risk patients (green line; $P = 0.00058$, log-rank test; hazard ratio = 2.65).

In addition, we examined scaled Schoenfeld residuals to verify the proportional hazards assumption (**Figure 3.9**) and analyzed martingale-residuals (**Figure 3.10**) to verify linearity, as well as Variance Inflation Factor (VIF) analysis on 35 genes included in the final multivariable Cox model to assess the potential for multi-collinearity on those 35 signature genes (**Table 3.4**). The results of residual analysis with overall VIF were acceptable without high correlation, with all VIF values being less than ten in the training dataset. These results

suggest that the 35-gene signature is significantly associated with the survival of 1p/19q code patients.

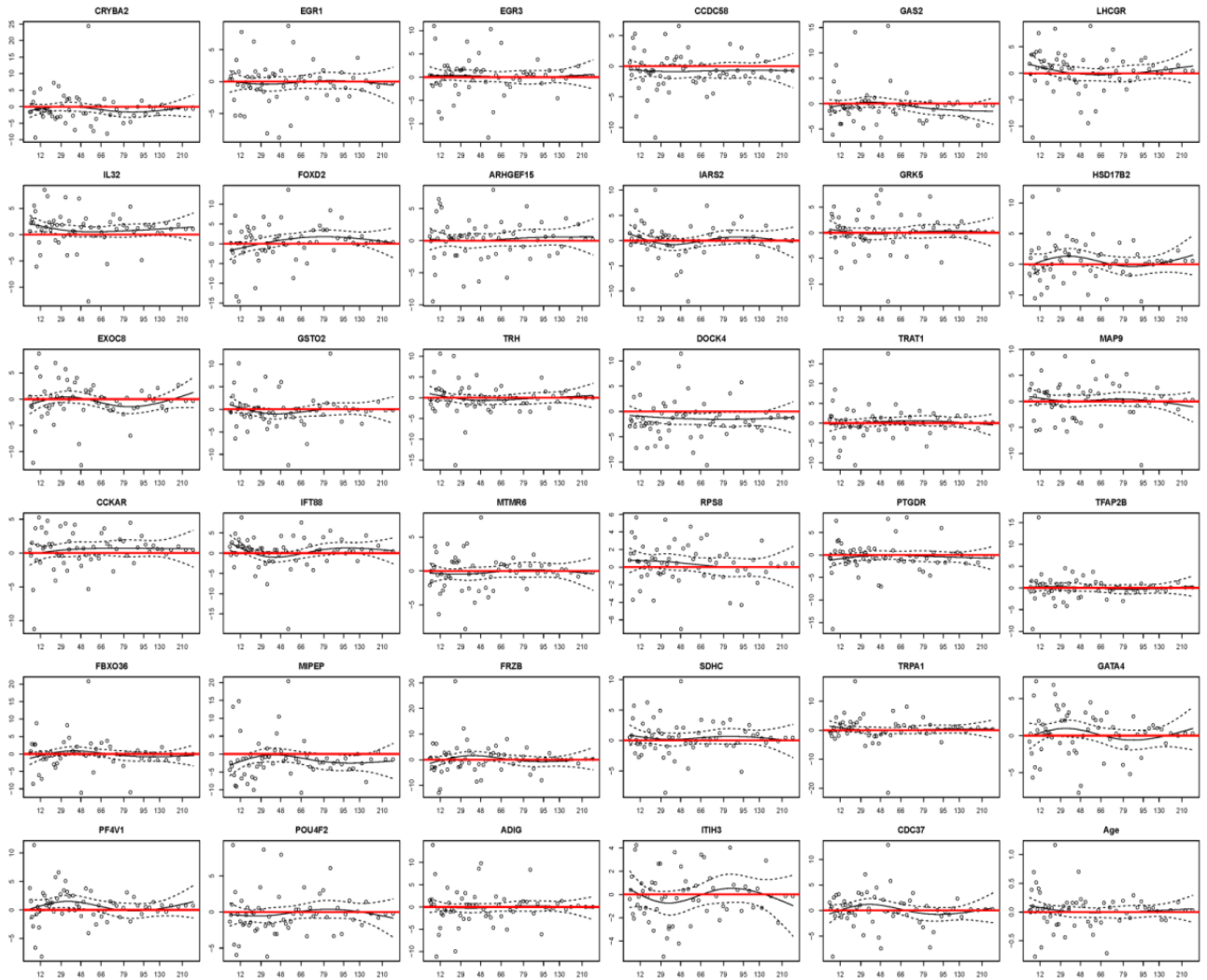


Figure 3.9 Schoenfeld residual plots to evaluate the proportional hazard assumption of signature genes

The X-axis represents the overall survival time for each patient; Y-axis represents scaled Schoenfeld residuals, the black dots represent the Schoenfeld residuals for each predictor in the model. The horizontal red line refers to Schoenfeld value as zero on the individual plots. The solid black curve is a cubic split fit to the data and the dashed black lines represent upper and lower approximate 95% confidence limits.

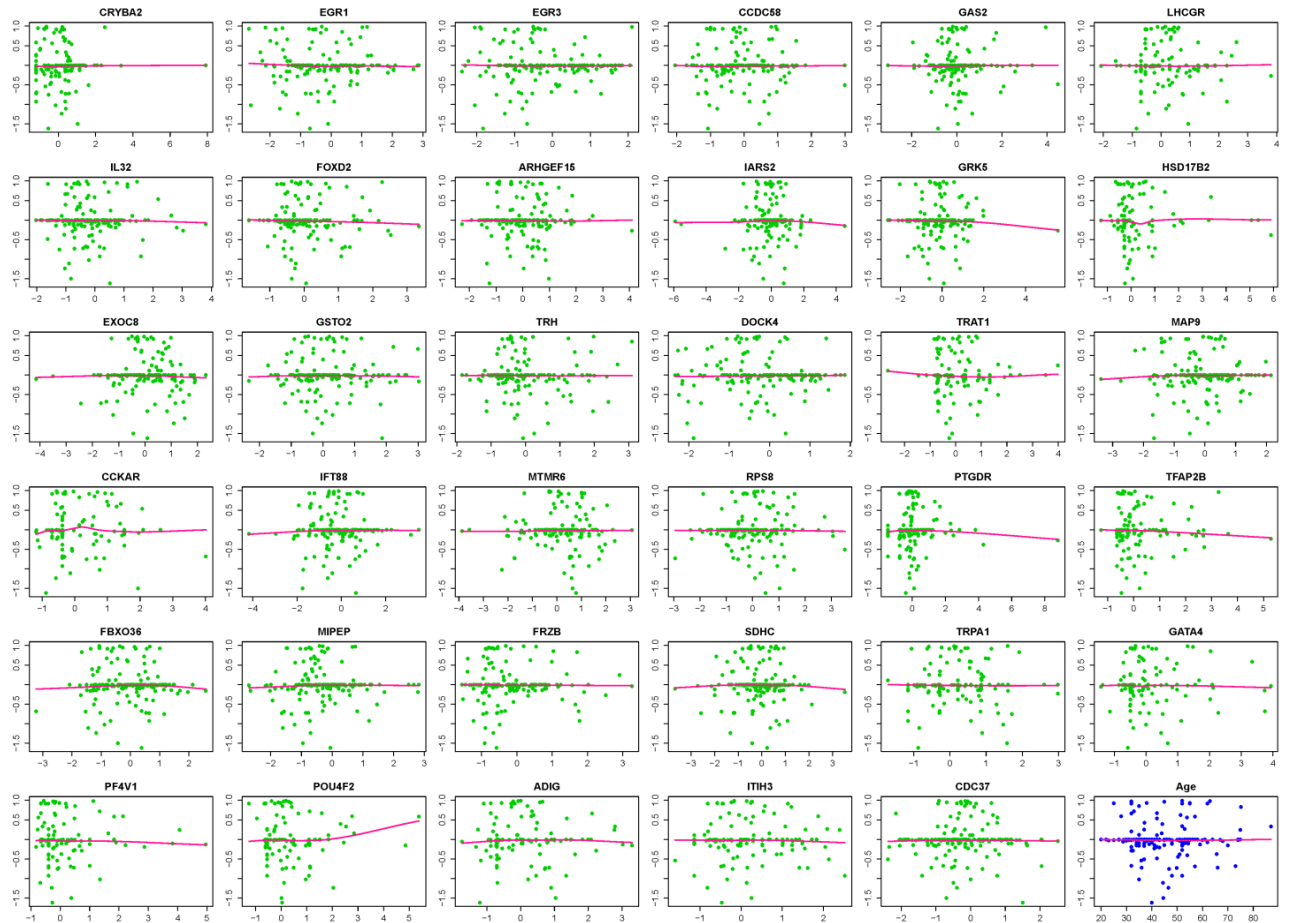


Figure 3.10 Martingale residuals to evaluate the linearity of log hazard on signature genes

The X-axis represents the expression value of each signature gene in all the patients; Y-axis represents martingale residuals, the green dots represent the martingale residuals for each predictor (covariates) in the hazard model. The horizontal red line refers to a nonparametric loess regression line on the individual plots.

When we examine the association between risk scores and overall survival time, we observed a trend for the high risk group towards reduced survival compared to the low risk group (**Figure 3.11**).

Table 3.4 Variance inflation factors (VIF) of 35 gene signature for prediction in the training dataset

Gene Signature	Variance Inflation Factor (VIF)
PF4V1	2.144599
GAS2	2.774342
TRAT1	2.836738
CRYBA2	3.037724
CCKAR	3.254183
IL32	3.264624
EGR1	3.347227
ITIH3	3.353768
CCDC58	3.379878
MTMR6	3.395117
CDC37	3.404695
TRH	3.525869
TFAP2B	3.573467
ARHGEF15	3.621795
ADIG	3.635986
MAP9	3.702387
RPS8	3.713787
GRK5	3.76141
SDHC	3.836232
DOCK4	4.013433
EXOC8	4.109896
FOXD2	4.131283
GSTO2	4.212632
FBXO36	4.245492
EGR3	4.420287
TRPA1	4.536422
POU4F2	4.666415
HSD17B2	5.37498
LHCGR	5.83129
GATA4	5.871038
IFT88	6.036097
PTGDR	6.606242
FRZB	7.335438
IARS2	7.560403
MIPEP	9.128166

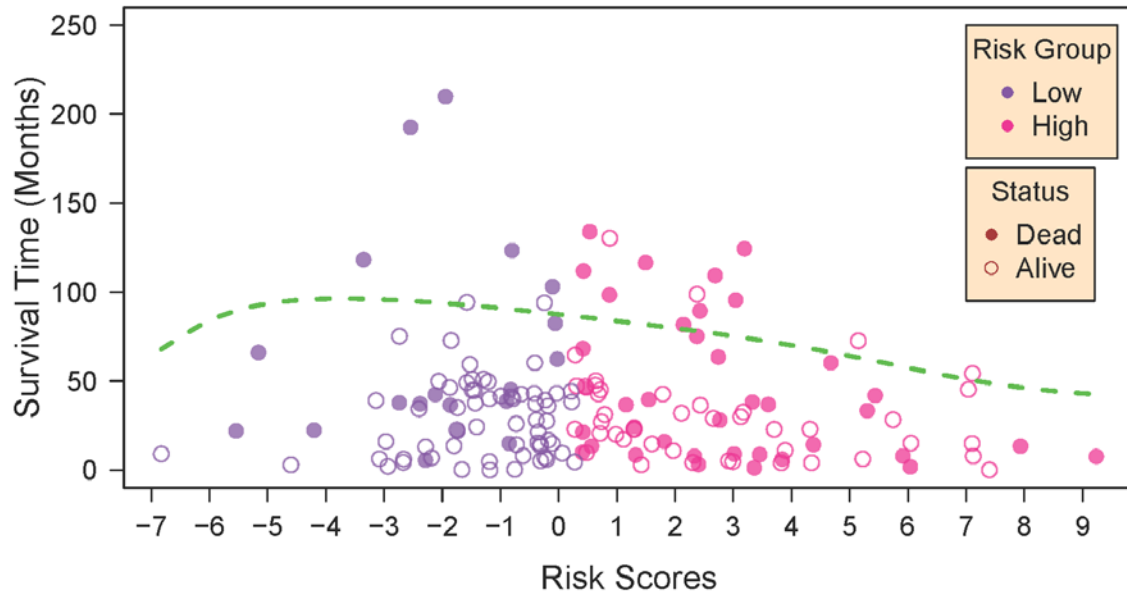


Figure 3.11 Correlation between gene signatures derived risk scores and overall survival time.

The purple dots represent low-risk patients (negative risk scores), and the red dots represent high-risk patients (positive risk scores). The soft dots represent living patients; the solid dots represent dead patients. The green line plots a smooth, nonparametric estimation of the quantile distribution of overall survival (OS) as a function of patients' risk scores.

3.3.4 Multivariable analysis shows prognostic power of 35-gene signature

Using prognostic index (risk score) as a continuous covariate, we determined the predictive accuracy by computing the C-index of the gene signature, age and grade. We also examined the C-index value of the gene signature combined with age and grade. Both analyses were performed in the joint validation set restricted to the patients whose age and grade information is available. The C-index of the gene signature (0.626 ± 0.044) was comparable to that of age (0.640 ± 0.048), or grade (0.640 ± 0.073) alone (**Table 3.5**). The highest C-index was achieved in when combining the three variables (0.663 ± 0.041). These results suggest that risk prediction is most accurate when combining the 35-gene signature with age and tumor grade.

Table 3.5 Performance of Multivariable Analysis in Validation Dataset

Predictor	Gene signature	Age	Grade	Age + Grade	Gene signature + Age + Grade
C-Index \pm SE	0.626 \pm 0.044	0.640 \pm 0.048	0.640 \pm 0.073	0.656 \pm 0.041	0.663 \pm 0.041
C-Index(CI)	0.540, 0.712	0.545, 0.734	0.497, 0.785	0.574, 0.737	0.583, 0.743
HR (95% CI)	1.78 (1.02 - 3.11)	1.71 (0.99 - 2.97)	1.26 (0.74 - 2.15)	2.06 (1.18 - 3.60)	3.23 (1.73 - 6.04)

Abbreviations: C-index, concordance index; HR, hazard ratio; CI, confidence interval

Estimates are based on data from patients in the combined validation dataset (n=191, 1st validation dataset + 2nd validation dataset), with both age and grade information available. The hazard ratios (HR) and their 95% confidence intervals between two groups of patients in the validation dataset were calculated based on their risk scores above and below the median risk score computed from the training dataset.

3.3.5 Functional Annotation of 35 Gene Signatures

We compared gene expression between the high- risk and low- risk groups and found that 32 out of 35 signature genes showed a significant difference (**Figure 3.12, Table 3.3**).

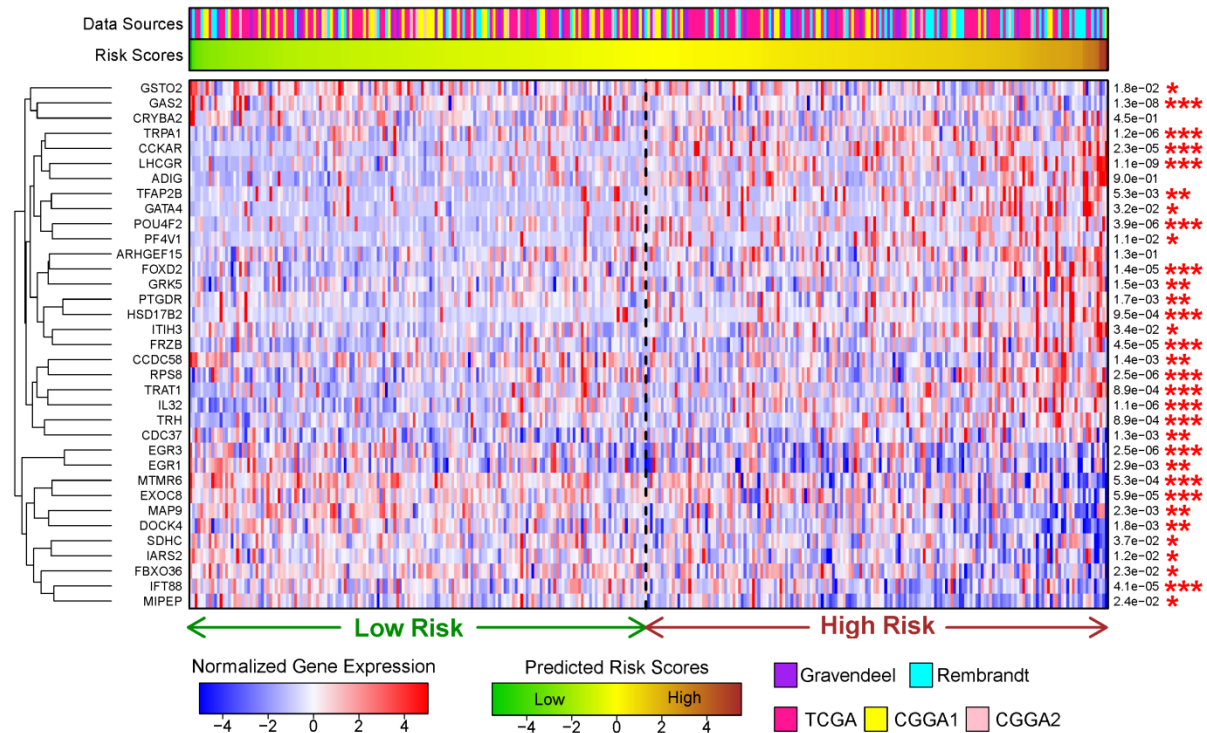


Figure 3.12 mRNA differential expression patterns of 35 signature genes in two risk groups

The normalized level of mRNA expression was plotted, with genes labelled on the left along with their corresponding p-values on the right, asterisks indicating the significance of differential gene expression (t-test) between two groups with predicted high risk (right part) and low risk (left part) (**P < 0.01, *P < 0.05). The dendrogram on the left showed the cluster by genes. The top bar shows each patient's predicted risk scores (green indicates low risk and brown indicates high risk) and its source dataset. (TCGA=red, Gravendeel=purple, Rembrandt=cyan, CGGA1_microarray=yellow, CGGA2_RNAseq=pink).

Gene set variation analysis revealed that the gene ontology (GO) terms which mapped to the genes that were differentially expressed in high- and low-risk groups were associated with acetylation activity, response to copper ions, prostaglandins and inflammation (**Figure 3.13**). The corresponding biological functions of protein acetylation, inflammatory response and copper homeostasis may contribute to these patients' high risk and poor clinical outcomes.

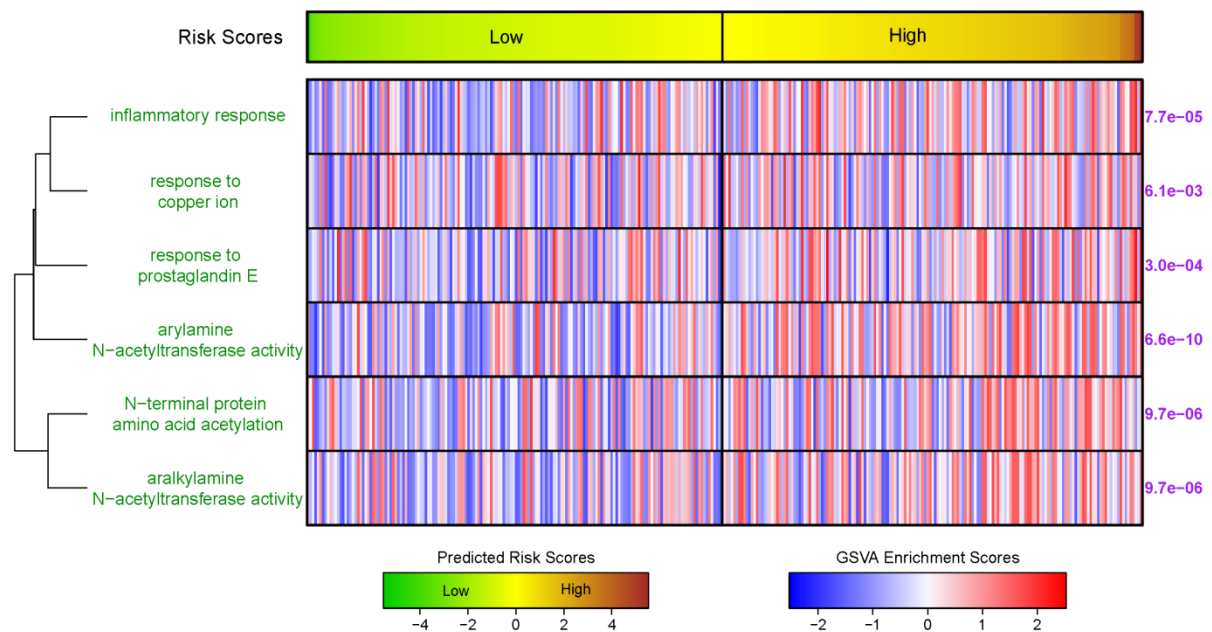


Figure 3.13 Association of risk groups with gene ontology (GO) function

Risk scores for each patient (top bar; in ascending order, from left to right) were derived from a multi-variable Cox model. Gene set variance analysis (GSVA) was used to calculate gene set enrichment scores (bottom). The P-values on the right were obtained using a t-test of enrichment scores from high-risk and low-risk groups for each GO term.

The presence of the inflammation category amongst the differentially activated GO terms suggested differences in the tumor microenvironment between high-risk and low-risk groups. We applied the ESTIMATE algorithm to predict tumor purity using the gene expression profiles (Yoshihara, Shahmoradgoli et al. 2013) and found a significant increase in ESTIMATE scores in the high risk group (**Figure 3.14**), suggesting that a greater presence of inflammatory microenvironment components is associated with progressive tumorigenesis.

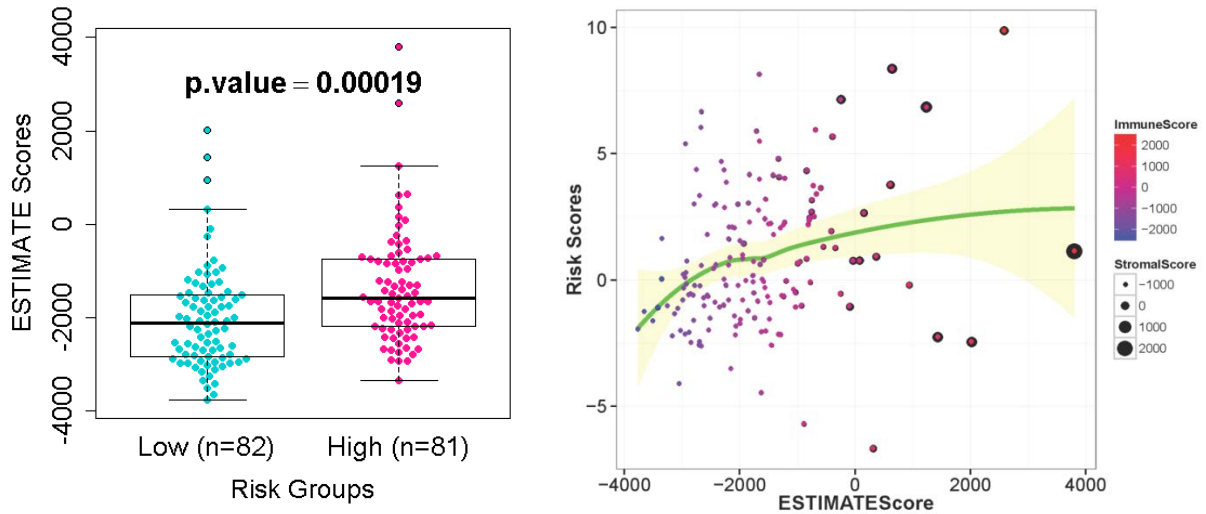


Figure 3.14 Comparison of ESTIMATE scores in high- and low-risk group in the first validation dataset of 1p/19q code1

(A) Scores computed by the ESTIMATE tumor purity algorithm were plotted for the high-risk (red) and low-risk group (blue) from the first validation dataset, with p-values indicating the significance of the difference between ESTIMATE scores (t-test) between two groups (** $P < 0.001$).

(B) The correlation of ESTIMATE scores (X-axis) and risk scores (Y-axis) are plotted. The LOESS curve in green fitted to ESTIMATE scores from risk scores. The confidence intervals of loess regression are in yellow.

3.3.6 Applying the 35-Gene Signature across Glioma

We asked whether the 35-gene signature model is also associated with patient survival in patients with IDH-wildtype or IDH-mutant-non-code1 gliomas. From TCGA, we obtained the gene expression profiles from 223 IDH-mutant gliomas that were wild type for chromosome arms 1p and 19q, as well as the transcriptional profiles from 221 IDH-wildtype gliomas. The analysis was restricted to cases with available outcome data and expression data generated by RNA-sequencing. After computing risk scores for all samples we separated the two datasets into a low risk and high risk group based on median risk score respectively and observed significant differences in overall survival for both glioma categories (**Figure 3.15**).

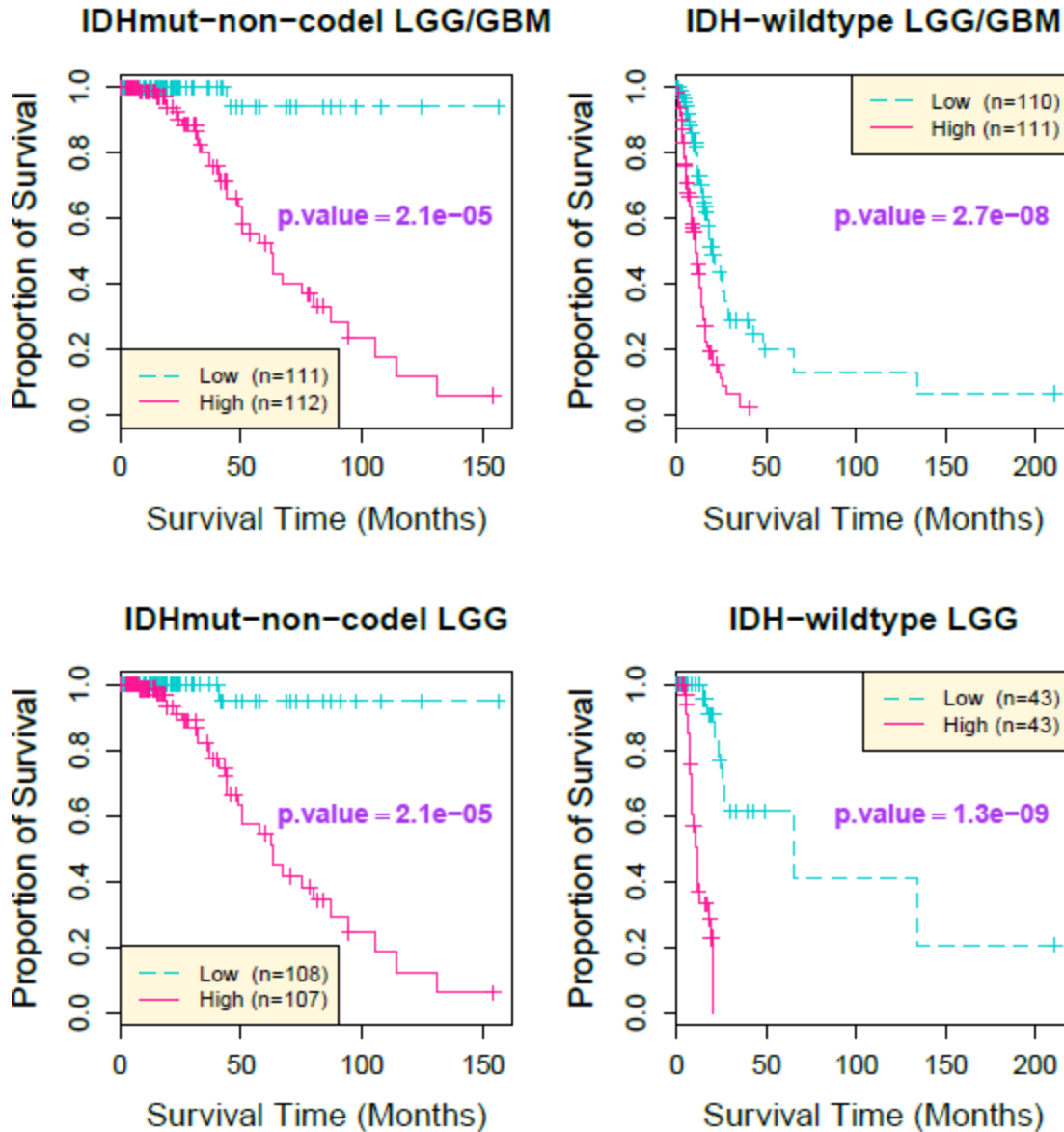


Figure 3.15 Prediction of outcome in non-codel IDH-mutant glioma and IDH-wildtype glioma

Kaplan-Meier cumulative survival curves for TCGA diffuse glioma patients whose tumors carry IDH-mutation but not the 1p/19q co-deletion (left) and with IDH-wildtype tumors (right), classified into two groups based on 35-gene signature derived risk scores. p-value is the result of a log-rank test between the two groups shown in each panel.

To gain insight into the universal relevance of the 35-gene signature across different molecular subtypes of glioma, we also applied the ESTIMATE algorithm to compare tumor

purity between high- and low- risk groups. With IDH-mutant-codel gliomas, we found that ESTIMATE based tumor purity scores were significantly lower in the low-risk group of IDH-wildtype glioma samples, compared to their high-risk counterpart. This was not the case for IDH-wildtype LGG, nor for IDH-mutant-non-codel gliomas regardless of grade (**Figure 3.16 A, B**). The difference in microenvironment presence between high- and low-risk group of IDH-wildtype glioblastoma emphasizes the facilitating role that tumor-associated microglia play in promoting disease progression (Hanahan and Weinberg 2011).

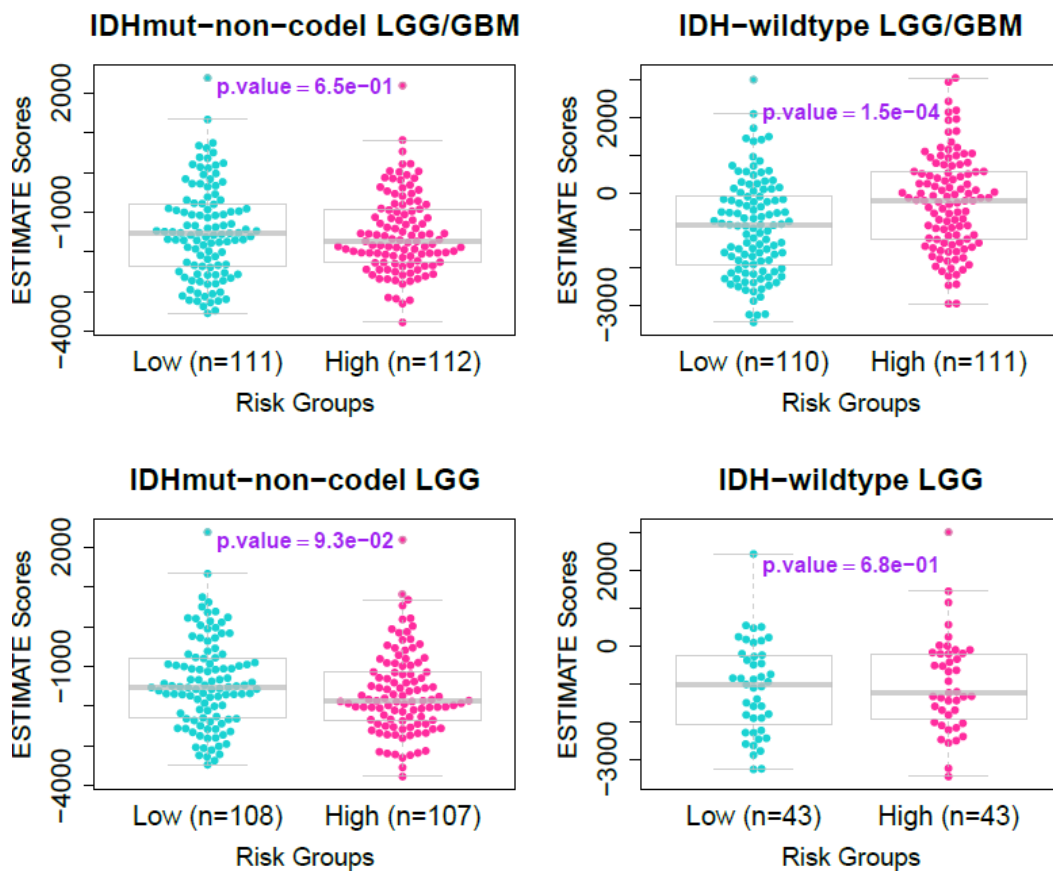
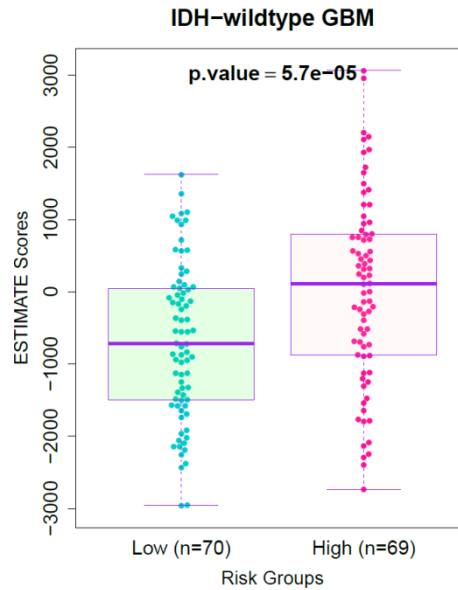


Figure 3.16 Comparison of ESTIMATE scores in high- and low-risk group of IDH-mutant-non-codel and IDH-wildtype glioma

(A) Scores computed by the ESTIMATE tumor purity algorithm were plotted for the high-risk (red) and low-risk (blue) groups from IDH-mutant-non-codel (left) and the IDH-wildtype (right) TCGA cohorts, with p-values reflecting the significance of the difference between ESTIMATE scores (t-test) between two groups.



(B) Scores computed by the ESTIMATE tumor purity algorithm were plotted for the high-risk (red) and low-risk group (blue) from the IDH-wildtype-GBM, with p-values reflecting the significance of difference between ESTIMATE scores (t-test) between two groups.

3.4 Discussion

High-throughput gene profiling and sequencing have yielded new insights on the molecular aberrations underlying glioma progression.(Verhaak, Hoadley et al. 2010, Ceccarelli, Barthel et al. 2016). As our perspective on the optimal clinical and molecular marker based classification of adult diffuse glioma harboring 1p/19q co-deletion progresses, biomarkers for risk based classification may provide additional value. Our systematic analysis identified a 35-gene signature, which classified 1p/19q codel glioma patients according to their overall survival. Remarkably, the median survival of the group of patients classified as high risk was 75 months, as opposed to 45 months for non 1p/19q codel patients, confirming that 1p/19q codel patients have a favorable overall survival and suggesting a relatively homogeneous disease subtype. Our gene signature was derived from multiple data sources, representing patients from a mixture of diffuse glioma grades and histologies. We validated the prognostic performance in independent validation datasets and showed the signature to

have added predictive signal when combined with known prognostic markers such as age and grade. Due to the unavailability of some of the files needed, normalization of training and validation datasets could not be performed entirely separately. Patients in our cohort were treated using a variety of different modalities and treatment annotation was lacking for a substantial portion of the dataset. With the recent introduction of a potential standard of care for low grade glioma (Buckner, Shaw et al. 2016), it is important to repeat and validate the gene signature on a coherently treated patient data set, while considering additional prognostic factors such as tumor size, location, and extent of resection. In order to pursue validation studies, risk scores can be computed using the gene signature and regression coefficients derived from the training dataset.

The univariate Cox model alone is insufficient for feature selection through estimation of survival as clinical endpoints when solving regression problems with high dimensional data. To prevent overfitting, the ridge regression Cox model demonstrates the best performance in tested datasets (Bovelstad, Nygard et al. 2007). Therefore, we applied the univariate Cox model to select genes related to OS time and regularized regression coefficients calculated by an elastic net regression Cox model, which combined the algorithm of ridge and lasso regression, to improve the predictive performance in the independent datasets. Owing to the relatively small number of events (64 deaths /170 patients in training dataset), we applied a 3-fold cross validation for Cox PH regression and selected the signature genes with optimized λ based on penalty regularization. While none of the individual genes showed an exceptionally high coefficient in our Cox model, multiple genes cumulatively exhibited an effect on survival prediction. Finally, we used multivariable Cox regression to adjust the selected genes for the clinical factors age and grade and to generate our prognostic index. As treatment information was unavailable for a considerable portion of our cohort (43%), we did not consider treatment variables in our statistical modeling.

Pathway analysis suggested that N-terminal acetyltransferases (NATs), protein acetylation, response to copper ions, prostaglandins and inflammation may be involved in 1p/19q glioma progression. Therapeutic agents including tamoxifen and cisplatin have been reported to demonstrate their anti-cancer effects through NAT inhibitory activity (Lu, Lin et al. 2001, Lee, Lu et al. 2004, Ragunathan, Dairou et al. 2008, Kalvik and Arnesen 2013), suggesting targeting of NATs as a potential therapeutic strategy in high risk 1p/19q co-deleted cases. In addition, copper depletion may act as an effective anti-angiogenesis strategy (Goodman, Brewer et al. 2004), and prostaglandins play an important role in cell adhesion, migration, and invasion during cancer development (Menter and Dubois 2012). Accordingly, the genes involved in protein acetylation and response to inflammation and copper are highly expressed in high-risk glioma patients (Di Cerbo and Schneider 2013). These data indicate that alterations in the expression levels of these signature genes might exert significant roles in glioma progression by promoting growth and conveying cell survival advantages. In addition to our pathway analyses, we noted that the stromal and immune related signals quantitated via ESTIMATE scores were significantly increased in the high-risk group relative to the low-risk group of 1p/19q codeleted glioma as well as in IDH-wildtype glioblastoma. This observation implies an association between the survival risk predicted by our gene signature and the infiltration by tumor-associated normal cells, which play a crucial role in microenvironment regulation during tumor progression (Quail and Joyce 2013). Four genes (*ITIH3*, *TRAT1*, *FRZB*, *IL-32*) from 35 genes signature overlap with the stromal and immune gene signatures used to define ESTIMATE scores, further suggesting the tumor microenvironment as a potential risk factor for subsets of glioma patients.

Collectively, our findings highlighted signature genes that might be involved in critical tumor progression and fundamental biological functions in gliomas with the 1p/19q co-deletion. The lack of treatment standardization amongst our patient cohort means that further research is needed to determine whether this or other gene signatures are able to

serve as treatment biomarkers. Ideally, clinical decisions would be based on a predictive model integrating clinical variables, tumor phenotypic and molecular factors. While further and prospective validation is needed, the gene signature approach may provide a starting point to better understand prognostic risk factors in 1p/19q co-deletion glioma. The results described here provide a first report investigating the heterogeneity of the relatively novel entity of 1p/19q codeletion glioma.

CHAPTER 4

Prediction of emerging fusion transcripts with oncogenic potential in The Cancer Genome Atlas pan- cancers

4.1 Introduction

Fusion transcripts are chimeric genes derived from partial DNA fragments of two previously independent gene partners. The prevalence of gene fusions varies widely between cancer types; gene fusions have been previously reported in 90% of all lymphomas, over half of leukemia, and one-third of soft-tissue tumors (Parker and Zhang 2013). Overall, recurrent fusion events occur at low frequencies; for example, *KIF5B-RET* fusion presents in 1%–2% of lung adenocarcinomas (Qian, Chai et al. 2014, Huang, Schneeberger et al. 2016).

Fusion transcripts are a result of mechanisms such as genomic rearrangements, including chromosome translocations (Guarnerio, Bezzi et al. 2016) and interstitial deletions (Hermans, van Marion et al. 2006, Meyer, Brieger et al. 2009); transcriptional read-through of neighboring transcription units (Nacu, Yuan et al. 2011, Varley, Gertz et al. 2014); and trans- and cis-splicing of adjacent genes (Zhang, Gong et al. 2012, Velusamy, Palanisamy et al. 2013, Jividen and Li 2014, Qin, Song et al. 2015). Many fusion transcripts are associated with oncogenic potential, which is correlated with the chimera proportion of the transcripts, and thus fusion transcripts play a critical role as driver mutations in a wide spectrum of cancer types (Bos, Gardizi et al. 2013, Watson, Takahashi et al. 2013, Zhou, Yang et al. 2013). Most gene fusions exert their oncogenic impact by regulating a fusion protein with oncogenic activity (e.g., by triggering constitutive activation of tyrosine kinase, on which cancer cells become dependent through “oncogene addiction”), by disrupting a function (e.g., by truncating the coding sequence of tumor suppressor genes), or by

deregulating one of the partner genes (e.g., by gain or loss of ubiquitination sites that modulate the stability of the original proteins).

Recent advances in bioinformatics have elucidated many aspects of oncogenic gene fusions, from the causative genomic features of fusion events to the structural and regulatory properties of fusion proteins. Gain in the understanding of both common and rare gene fusions have shown promising impact on clinical applications such as identifying molecular subtypes of cancers, stratifying patients according to fusion occurrence, monitoring residual disease after treatment, and predicting relapse (Eguchi, Faria et al. 2014, Wyatt, Mo et al. 2014). Fusion transcripts also have served as efficacious therapeutic targets in both solid and hematologic malignancies (Parker and Zhang 2013); for instance, the *ALK* inhibitor crizotinib emerged as an effective therapy for a subset of non–small cell lung cancer patients harboring *EML4-ALK* fusions (Sasaki, Rodig et al. 2010).

Fusion genes have been recognized for over 30 years as critical drivers of cancer progression and effective therapeutic targets. However, the complexity of the cancer transcriptome, the high dynamic range of gene expression, and the unavoidable presence of sequencing errors confound computational fusion detection, and the gene fusions identified by different studies and in different cancer types show few distinct trends. Fusion rank algorithms could help biologists and clinicians prioritize putative fusion lists and predict the biologic functions of these fusions on the basis of their genomic features and protein structures. Using The Cancer Genome Atlas (TCGA) RNA sequence profiling, we have comprehensively characterized fusion transcripts across 33 tumor types and provided an integrative platform for the identification and annotation of both novel and known fusion genes that could serve as diagnostic tools and molecular targets for therapeutic interventions.

4.2 Methods

4.2.1 Data resources

TCGA DNA and RNA sequencing data were downloaded from the Cancer Genomics Hub (CGHub, <https://cghub.ucsc.edu>). Copy number segmentation data and gene expression data were downloaded from Firehose (<https://gdac.broadinstitute.org/>).

Somatic mutation data were downloaded from UCSC Xena repository.

([https://xenabrowser.net/datapages/?cohort=TCGA%20Pan-Cancer%20\(PANCAN\)\)](https://xenabrowser.net/datapages/?cohort=TCGA%20Pan-Cancer%20(PANCAN))) All data used in this study were summarized in **Table S4.1**. A panel of normal samples (n=689-22) were assembled as controls to filter out potential germline fusion events, after excluding 22 normal samples that resembled tumor samples on the basis of unsupervised hierarchical clustering. The clustering was done within each cancer type using expression of all genes and ward's method.

4.2.2 Identification of fusion transcripts

We applied PRADA (Torres-Garcia, Zheng et al. 2014) to all RNAseq samples for data preprocessing and fusion calling. In brief, RNA sequencing reads were aligned to a composite reference consisting of both genome (hg19) and transcriptome (Ensembl 64), followed by a remapping step that aligns transcriptome coordinates to the reference genome (Berger, Levin et al. 2010). GATK best practices were implemented in the pipeline, including marking duplication and base quality recalibration. More information about PRADA can be found at <http://bioinformatics.mdanderson.org/main/PRADA:Overview>.

PRADA detects fusion transcripts based on discordant read pairs (reads mapping to different protein-coding genes) and junction spanning reads (reads mapping to the exon–exon junctions). We required at least two discordant read pairs and one junction spanning read to call a fusion candidate. Multiple steps were applied to remove possible artifacts, including gene pairs with high sequence similarity (Blastn E-value > 0.001), low transcriptional allelic fraction (TAF, minimum 0.01 for both partner genes), and high partner

gene variety (PGV, maximum 10). TAF was calculated as the ratio of fusion supporting junction spanning reads to the total number of junction spanning reads. PGV was defined as the number of chromosomal arms where partner genes were localized for a fusion gene per cancer type. We excluded all fusions where any partner gene had a PGV more than 10. To test the stringency of this cutoff we ran 100,000 permutations and observed a less than 0.001% chance by random to see such promiscuity, suggesting fusions involving these genes were highly unlikely to be *bona fide*. Finally we removed all fusions found in normal samples.

4.2.3 Validation of fusion transcripts through integrating structure variants and copy number changes

For cases where both copy number profiles and gene fusions were available, we aligned fusion points with copy number breakpoints allowing an extension of 100 Kb to the expected direction for both partner genes. We used Speedseq to detect structural variants (SVs) (Chiang, Laver et al. 2015) from whole genome sequencing (WGS) data. We filtered SVs requiring more than 3 supporting reads > 3, at least one split read and one discordant read pair. For fold-back inversions (BND on the same chromosome) we required more than 9 supporting reads. We removed SVs with breakpoints falling in low-complexity regions (eg. repeat region DNA), or stacking across different tumor types. We further removed SVs with high sequence similarity of the 100bp window flanking the breakpoints. Germline events were filtered out by comparing with matched normal samples.

We scanned the intersection between the edge of the confidence interval from the supported structure variants including large fragment duplication, deletion, insertion and inversion and truncated intron region flanking to the junction upon fusion events. First we assigned two partner genes into three categories based on their relative location to the adjacent break point of structure variants separately, (A,B,C,D definition only consider

individual fusion junction that mapped to any break point of SV calls, here fusion Gene-A and Gene-B are calculated/ scanned separately) where category A: a break point of structure variants fall into the exact intron region following the transcriptional orientation of adjacent junction point of the fusion; category B: a break point of structure variants fall into the maximum edge exon region of an entire gene depending on the transcriptional direction side expanded with 100Kb wiggle room; Next we assigned each fusion pair into four categories based on the fidelity supported by structure variants. High fidelity was defined when both partner genes are assigned with category A, middle fidelity was assigned for those fusion genes with one partner gene assigned with category A and another partner gene assigned as category B, low fidelity was defined when both fusion partner genes are assigned as category B; For those fusion pairs with only one partner gene supported by structure variants, we assigned as one-sided.

4.2.4 Pathway and fusion centrality analysis

Pathway enrichment of fusion genes was tested against Gene Ontology (GO) based on hypergeometric distribution. Fusion transcript centrality score was calculated based on domain-based fusion model, to predict the oncogenic driver in which partner genes act as hubs in a cancer pathway network. (Wu, Kannan et al. 2013) We set the threshold value to 0.37 for fusion degree centrality score, the fusion transcripts assigned with centrality score > 0.37 were considered as prioritized cancer fusion drivers.

4.2.5 Exon expression analysis

We calculated the reads per kilobase per million mapped for exon expression based on Ensembl 64 annotation using realigned files generated from PRADA, and then we performed Z-score transformation for the expression of each exon across all the samples in

each cancer type. Welch's t-test was used to compare the expressions of exons before and after the junction point of each partner gene.

4.3 Results

4.3.1 Distribution of fusion transcripts in different cancer types

We analyzed mRNA sequencing data of 9968 tumors and 665 normal samples from 33 cancer types available in TCGA (**Table S4.1**) and identified 56,198 candidate fusion events. After removing 35,263 potential artifacts with stringent filters and 862 fusions detected in normal samples, we obtained 20,112 somatic fusions for further analyses (**Figure 4.1A**). On average sarcoma and uterine carcinosarcoma harbored the most fusions whereas kidney cancers (papillary carcinoma, chromophobe carcinoma, clear cell and renal cell carcinoma), uveal melanoma, and thymoma harbored the least fusions, indicating the lower tail of the spectrum.

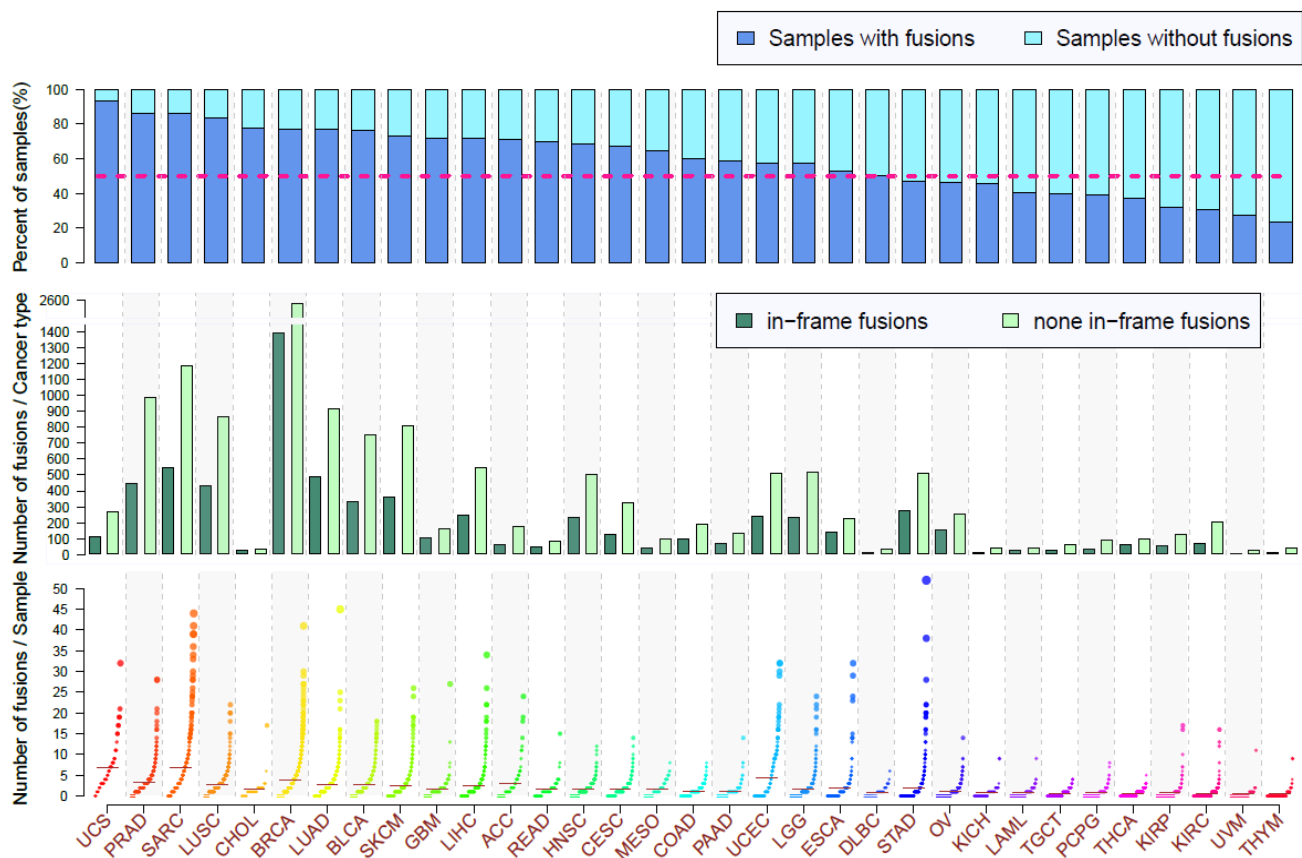


Figure 4.1 Spectrum of filtered fusion transcripts across 33 cancer types

The top panel shows the percentage of samples where at least one fusion transcript was detected per cancer type (dark blue). The middle panel shows total number of fusion transcripts detected in each cancer type. The bottom panel shows the number of fusions detected in each tumor sample. The brown horizontal marker indicates the mean number of fusions in each cancer type. The cancer types are sorted according to the percentage of samples detected with fusion transcripts.

Breaking into chromosome arms, sarcoma presented a strong enrichment of fusions on 12q (Figure 4.2), a pattern reminiscent of focal chromothripsis in a subset of this malignancy (Nord Karolin et al., 2013, Human Molecular Genetics).

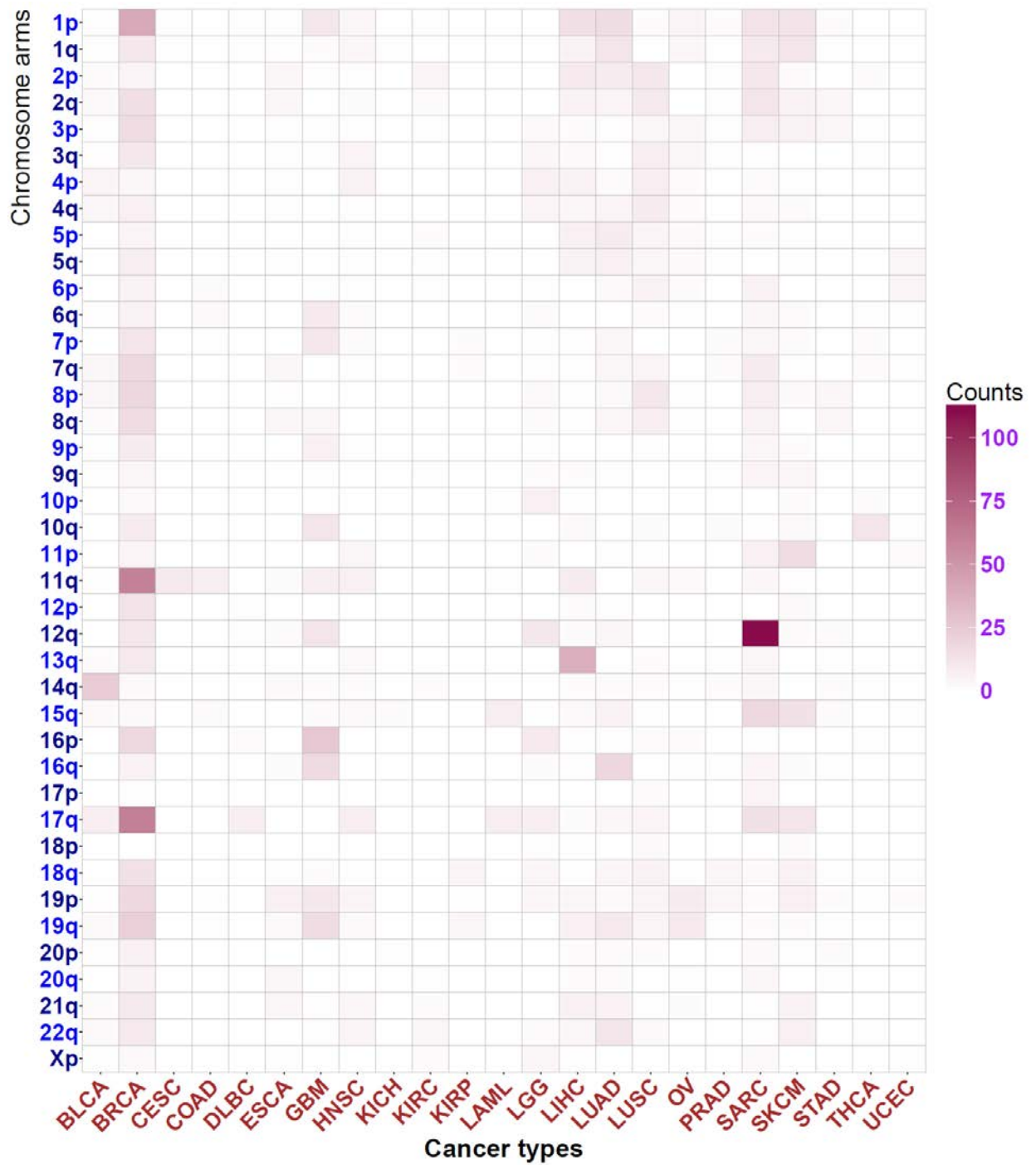


Figure 4.2 Counts of fusion transcripts supported by structure variants at different chromosome arms in 23 cancer types

Dark red on the heat-map represents the counts of fusion transcripts.

By integrating copy number data we found 32% of fusions had both junctions close to a DNA breakpoint and ~58.5% of fusion transcripts had at least one partner gene in amplified or deleted regions. We next separated fusions into short (<1 Mb), intermediate (1-10 Mb) and long (>10 Mb) intra-chromosomal events and inter-chromosomal events per genomic localization of the two partner genes. We observed disparate association patterns between these categories and copy number profile. For instance, almost 50% of fusions were short intra-chromosomal events in high grade ovarian serous carcinoma (OV), and they were predominantly associated with copy number changes. In contrast, 70% of fusions detected in thyroid carcinoma (THCA) were copy number neutral inter-chromosomal events. Acute myeloid leukemia (AML), which also exhibited a quiet genome, displayed a similar percentage (65%) of copy number neutral inter-chromosomal fusions (**Figure 4.3**). A rare cancer, pheochromocytoma and paraganglioma (PCPG), demonstrated predominantly copy number associated inter-chromosomal fusions which might reflect frequent large segmental copy number alterations in this disease (Flynn, Benn et al. 2015). Other outliers included cervical cancer, which showed minimal copy number related fusions.

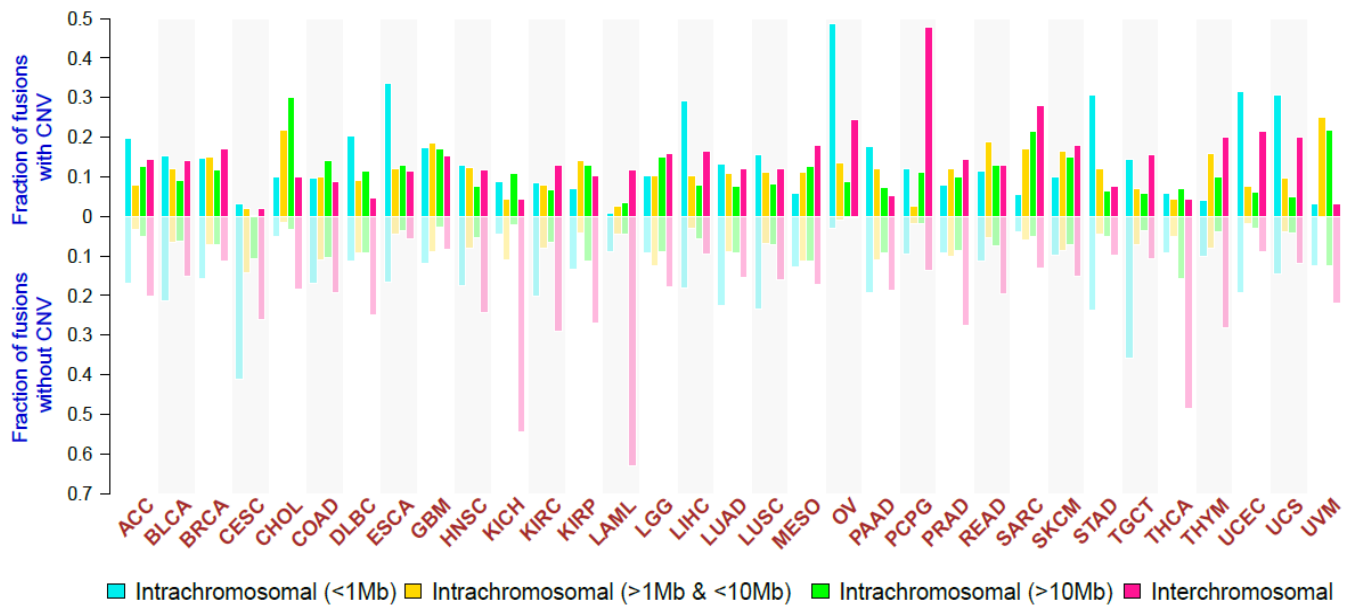


Figure 4.3 Distribution of genomic features among fusion transcripts across 33 cancer types

The bar plots represent the fractions of different categories of fusions according to the distance between two partner genes and presence of DNA copy number variations within a 100-Kb window up-stream and down-stream of the junction point. CNV: copy number variation

4.3.2 Gene fusion annotation and tight association with genomic structure alterations

The finding of frequent copy number neutral interchromosomal fusions prompted us to investigate to what extent balanced translocations may drive fusion formation. Using speedseq, we realigned whole genome sequencing data of 971 pairs of tumor and matched germline samples, and detected 272,638 somatic structural variations (SVs). Speedseq categorized SVs into deletion (6.6%), duplication (9.2%), inversion (65.6%) and translocation (18.5%) events depending on the read alignment patterns. No SVs were found in 70 tumors, of which 35 were acute myeloid leukemia or thyroid carcinoma. Similar with gene fusion, SVs also showed a continuous spectrum across human cancers (**Figure 4.4**). Of the 272,638 SVs, 151,564 had both ends falling into genic regions. In comparison, we detected a total of 2,588 fusions in these WGS cases, suggesting WGS was more sensitive

in calling SVs, or the majority of the SVs were not expressed. Filtering parameters of both DNA and RNA SV analyses had an impact on the observed discrepancy but were unlikely the major factors.

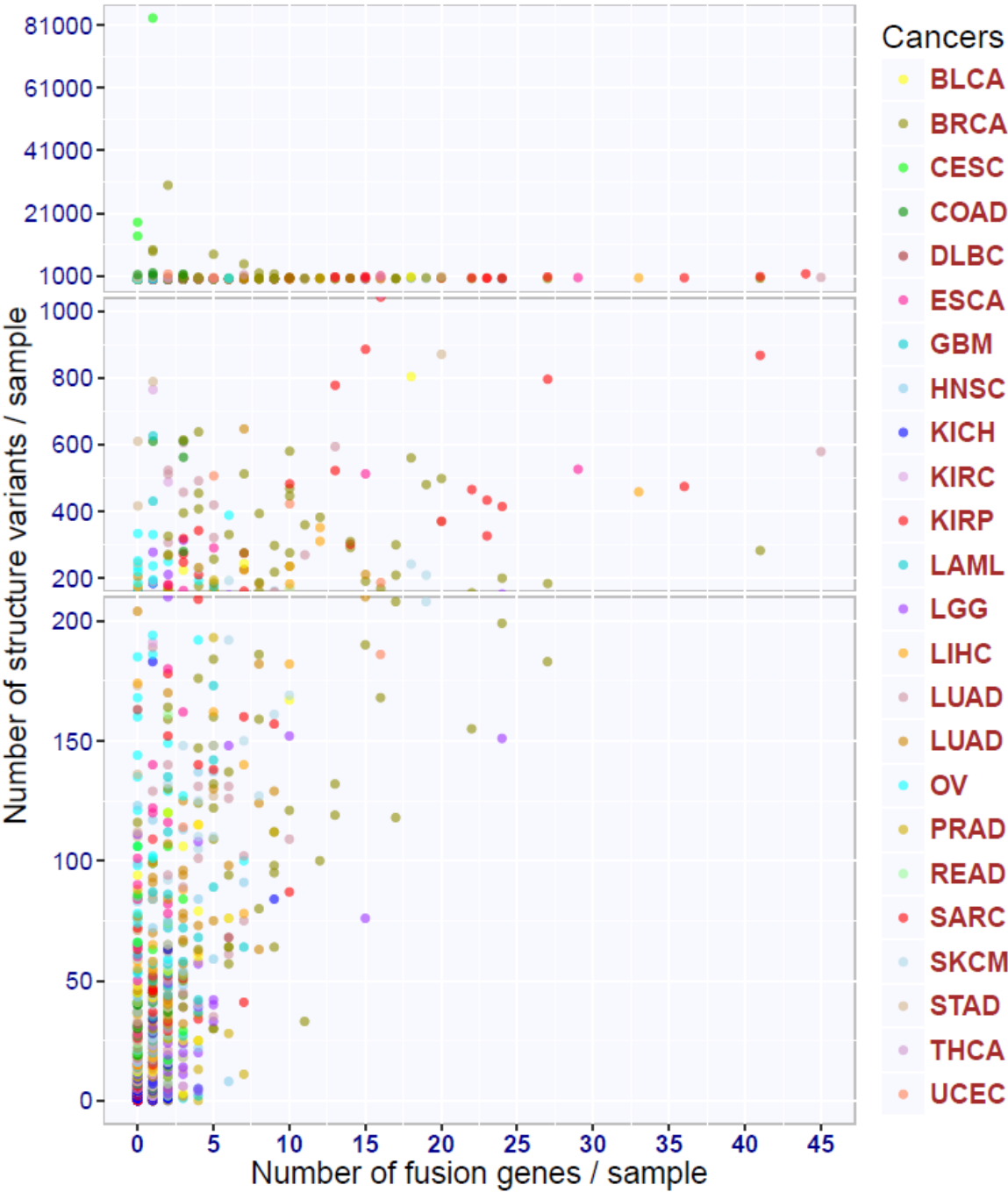


Figure 4.4 Counts of fusion transcripts and structure variants in each tumor sample
Each dot denotes one sample, and each color represents a cancer type.

We next examined the DNA structure variants for mechanistic evidence of the observed fusion events. Overall we found supporting evidence for 717 fusions (28%) from all four categories of structure variants (**Figure 4.5**).

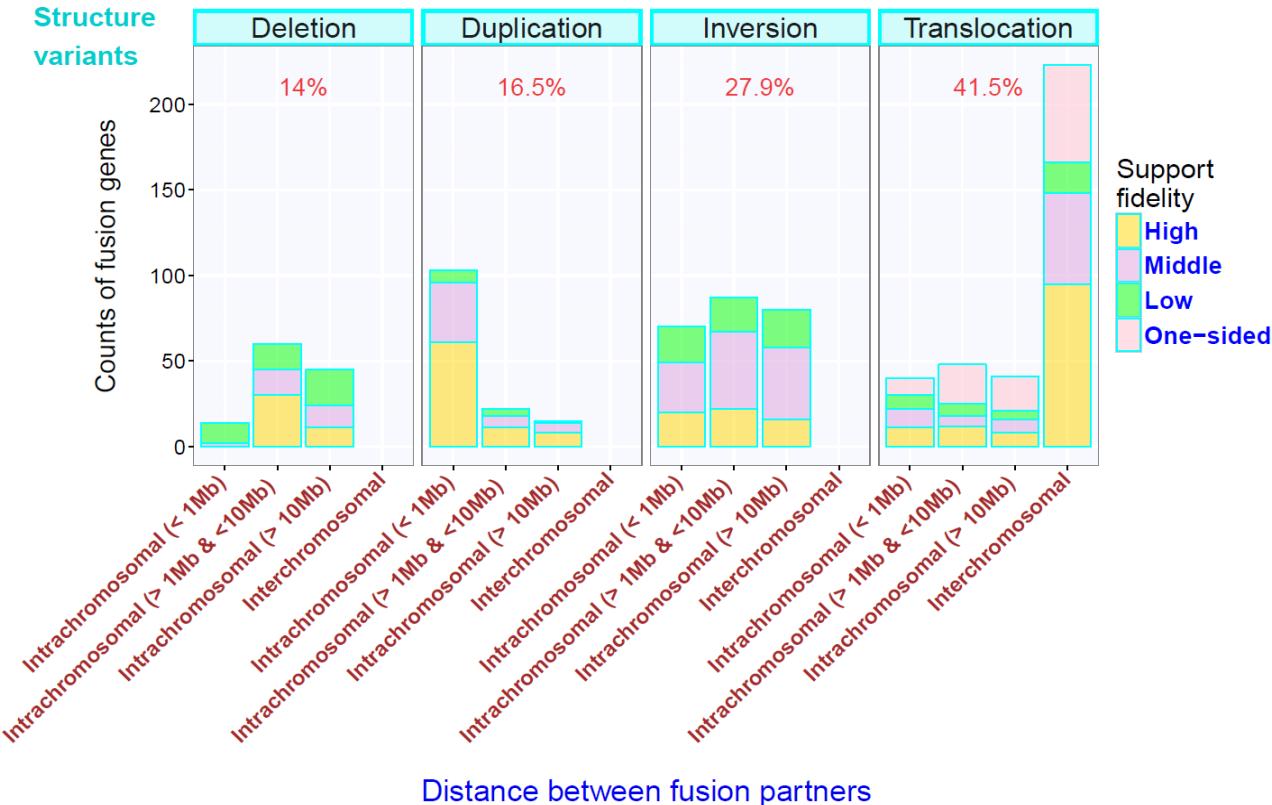


Figure 4.5 Distribution of fusion transcripts associated with structure variants

The number of fusion transcripts supported by different types of structure variants at different confidence levels. Each group was divided according to the distance of the chromosome location of two junction points from the fusion transcript (x-axis). Confidence level in stacked bar plot: high, breakpoints of structure variants in the intron region of adjacent fusion junction; middle, breakpoints of structure variants in the entire exon region of adjacent fusion gene; low, breakpoints of structure variants in the 100-Kb spanning region to the exon of adjacent fusion (y-axis).

Interestingly duplication explained the largest proportion of short intra-chromosomal fusions, and translocation accounted for all inter-chromosomal fusions. This pattern was independent of levels in supporting evidences. It should be noted that short intra-

chromosomal fusion was likely underestimated in the deletion SV group because our tool discards discordant read pairs in short distance to exclude possible read through events. A few examples are illustrated in **Figure 4.6**. Given the abundance of fusions supported by translocation, we will examine the translations in more detail.

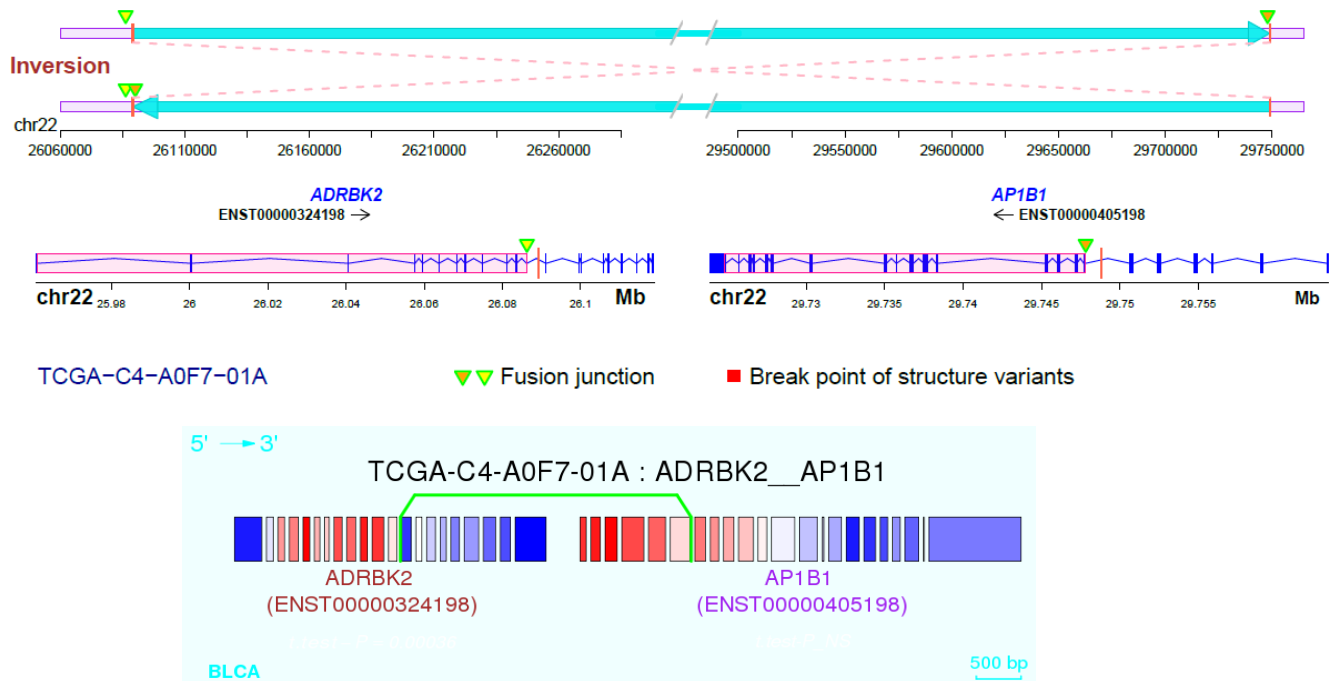


Figure 4.6A Fusion transcript derived from inversion in BLCA

A somatic *ADRBK2-AP1B1* fusion resulting from a balanced inversion (chr22: 26,089,092-29,748,825) in BLCA (TCGA-C4-A0F7) was measured by both whole-genome sequencing and RNA sequencing. The inversion is shown on the top (middle region in gray block truncated for visualization purposes). The partner genes *ADRBK2* and *AP1B1* formed the fusion transcript *ADRBK2-AP1B1* shown on the bottom. The red vertical lines represent breakpoints of structure variants detected by SpeedSeq.

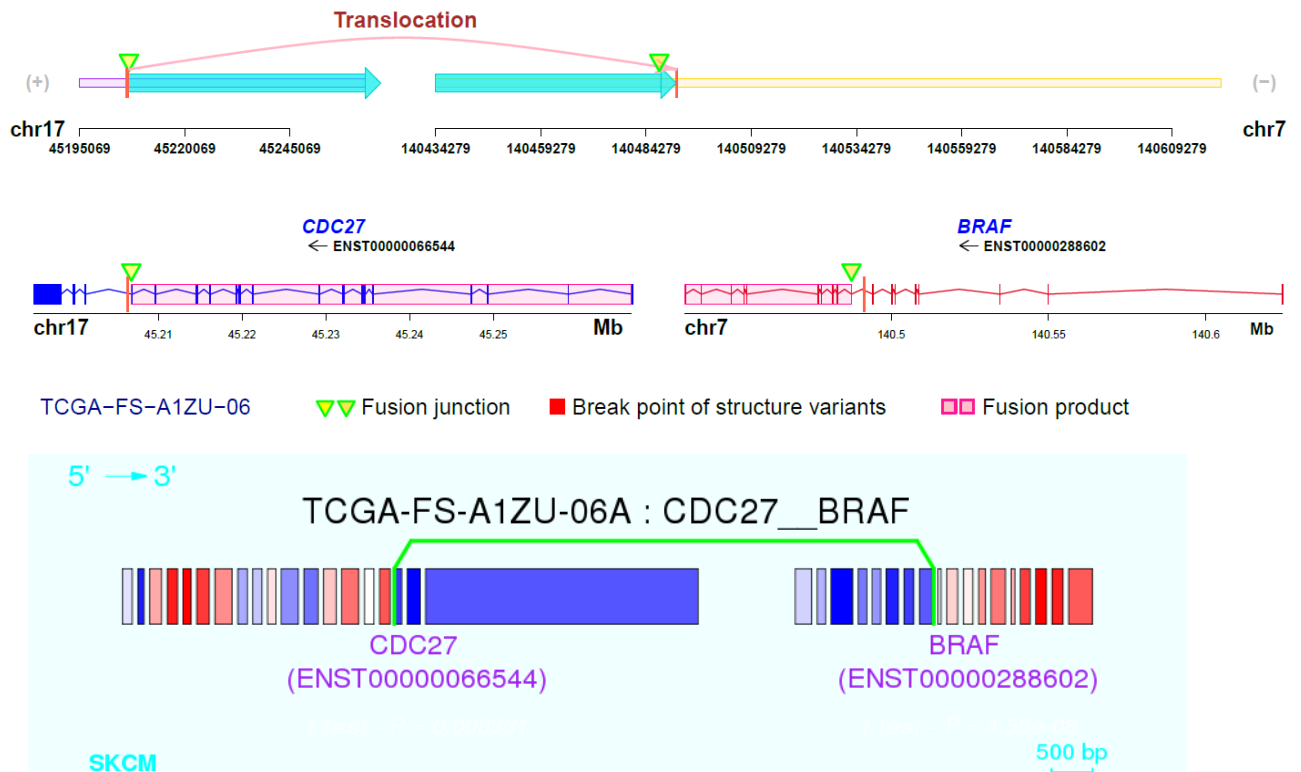


Figure 4.6B Fusion transcript derived from translocation in SKCM

A somatic *CDC27-BRAF* fusion resulting from a balanced translocation (chr17:45,206,337– chr7:140,491,457) in SKCM (TCGA-FS-A1ZU) was measured by both whole-genome sequencing and RNA sequencing. The inter-chromosomal translocation is shown on the top (middle region in gray block truncated for visualization purposes). The partner genes *CDC27* and *BRAF* formed the fusion transcript *CDC27-BRAF* shown on the bottom. The red vertical lines represent breakpoints of structure variants detected by SpeedSeq.

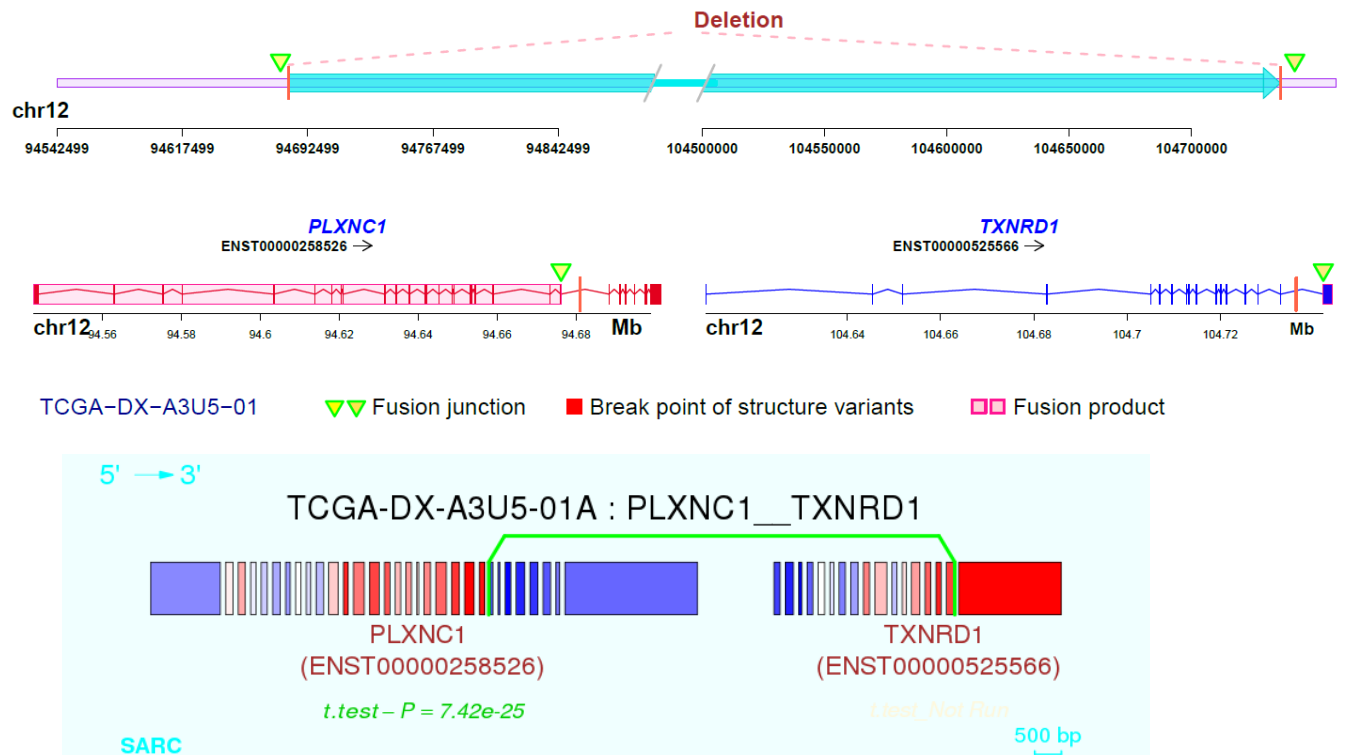


Figure 4.6C Fusion transcript derived from large fragment deletion in SARC

A somatic *PLXNC1-TXNRD1* fusion resulting from a large fragment deletion (chr12: 94,681,044-104,736,300) in SARC (TCGA-DX-A3U5) was measured by both whole-genome sequencing and RNA sequencing. The deletion is shown on the top (middle region in gray block truncated for visualization purposes). The partner genes *PLXNC1* and *TXNRD1* formed the fusion transcript *PLXNC1-TXNRD1* shown on the bottom. The red vertical lines represent breakpoints of structure variants detected by SpeedSeq.

4.3.3 Hotspot fusion transcripts are associated with genomic instability

Among all the fusion transcripts detected across all cancer types, the most frequent fusion transcripts, containing kinase partner gene *FRS2* (n=66 in 12 cancer types), were located at hotspots of chromosome 12q breakpoints, where most condensed fusion transcripts presented in sarcomas (n=39 for *FRS2* fusions in sarcomas). The genomic hotspots for fusions overlapped with regions that were frequently amplified in sarcomas and were accompanied by complex genomic rearrangement with intra-chromosome inversion. Only genomic events detected in samples overlapping with fusion-positive samples are shown in

a circos plot (**Figure 4.7**), suggesting that these fusion events are derived from genomic rearrangement. The genomic hotspots for fusions are observed on chromosomes 12q15; these regions are coincident with frequent focal gains, chromosomal translocations, and inversions. Increased levels of local genomic instability were linked to concomitant higher frequency of complex genomic rearrangement and consequently to fusion formation where both fusion points were adjacent to genomic breakpoints detected with multiple segments between the two breakpoints. This finding suggests that unbalanced genomic rearrangements occur on high-density specific DNA double-strand breaks and rejoin into fusion transcripts.

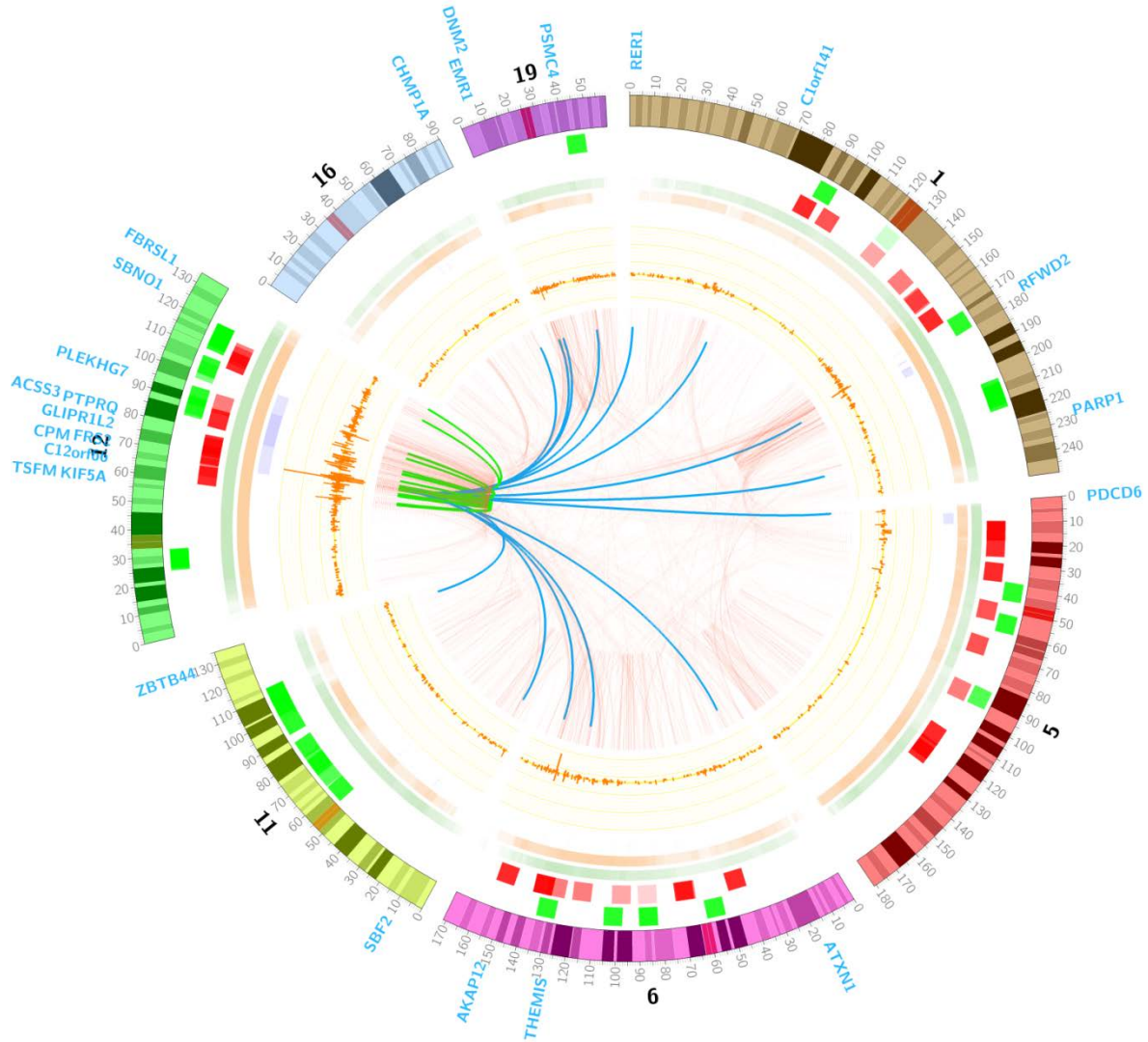


Figure 4.7 Association between hotspots of recurrent fusion transcripts with chromosome rearrangement and copy number alterations in SARC

A circo plot represents all the curated fusions identified in sarcomas from TCGA dataset. Fusion transcripts formed by *FRS2* kinase in SARC are represented by thick arcs ($n=39$) and labeled in blue outside the ideogram. The green arcs represent intra-chromosomal fusions; the blue arcs represent inter-chromosomal fusions. The light red arcs show genomic translocations in SARC. The middle ring in orange shows the frequency of fusion genes in a density histogram at the given genomic location. The next ring shows structure variants of inversion (purple), duplication (tomato), and deletion (olive-green) detected in SARC, and the outermost ring shows the copy number alterations of amplification (red) and deletion (green) in sarcomas. The frequencies of structure variants and copy number alterations are presented by the color density. Only samples detected with *FRS2* fusions are calculated for the frequency of structure variants and copy number alterations.

4.3.4 Prioritizing functional fusions

We observed 476 recurrent fusions across 33 cancer types, including those found in one cancer type or multiple cancer types. The former included *TMPRSS2-ERG* in prostate cancer, *PML-RARA* and *RUNX1-RUNX1T1* in leukemia, *PTPRK-PSPO3* in colorectal cancer, *FGFR2-BICC1* in cholangiocarcinoma, *EML4-ALK* in lung cancer and *FGFR3-TACC3* in glioblastoma, cervical cancer, head and neck cancer (**Figure 4.8**). We also identified less characterized recurrent fusions *UBTF-MAML3* in pheochromocytoma and paraganglioma.

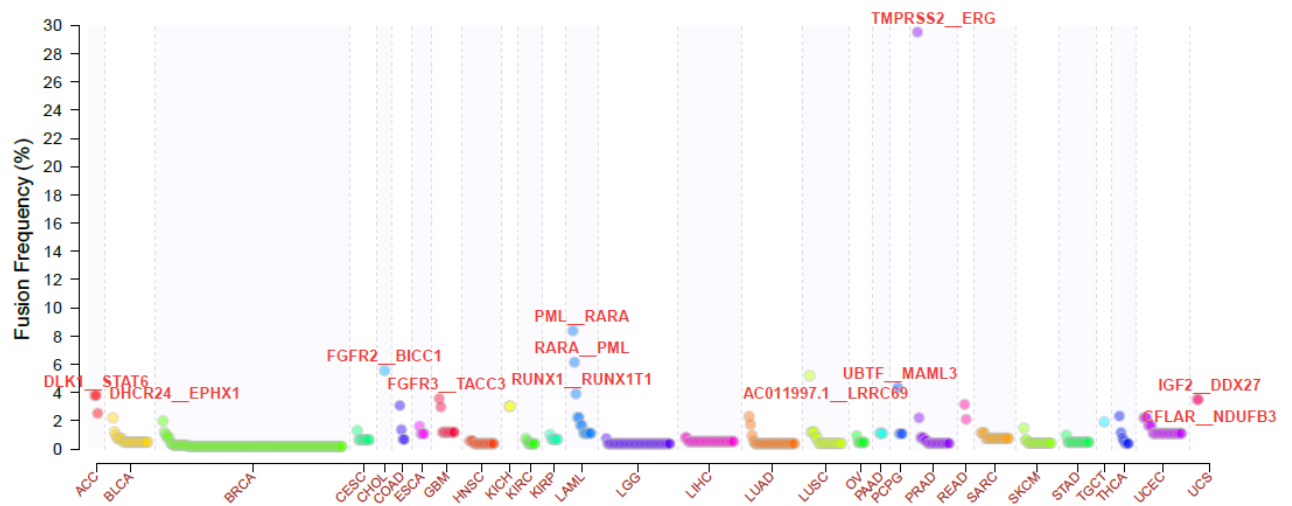


Figure 4.8 Frequency of recurrent fusion transcripts across 33 cancer types

The most recurrent ($n \geq 2$) fusion transcripts are shown, and the most prevalent fusion transcripts are labeled in red. The fusion frequency was defined as the number of samples detected with the number of recurrent fusions divided by the total number of samples analyzed in each cancer type.

Gene wise, *FGFR3* and *TACC3* represented the most frequent genes involved in recurrent fusions, followed by *BRAF*, *RARA*, *RET* and *PML* (**Figure 4.9**). Recently Pan FGFR inhibitors have been under test in phase I-II clinical trials (Shaw, Hsu et al. 2013), which may lead to targeted inhibition of tumors with FGFR fusions.

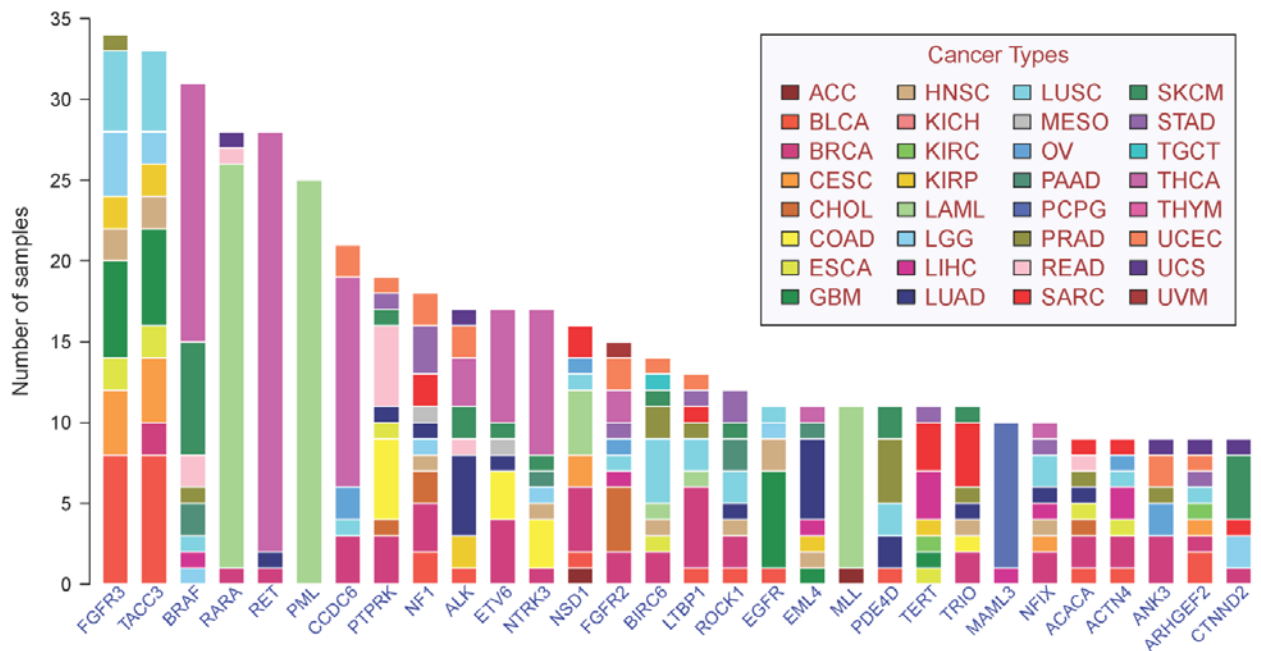


Figure 4.9 Most frequent partner genes in recurrent fusion transcripts across 33 cancer types

The distributions of partner genes that form the fusion transcripts are shown. The y-axis represents the number of patients found to have partner genes that form fusion transcripts in each cancer.

We applied two approaches to systematically nominate possible functional fusions, and we employed a network based approach to assign a centrality score to every fusion (Wu, Kannan et al. 2013). The underlying hypothesis is derived from the empirical observation that at least one of the partner genes of most known cancer fusions are hub genes in a gene network; thus the fusion centrality score (a multi-gene-based centrality metric) may reflect the functional significance of a fusion. We found thyroid cancer and acute myeloid leukemia distinguished from other cancers for high centrality scores, and both cancer types were known for fusion genes (TCGA, unpublished data). The third lead cancer type was cholangiocarcinoma, in which frequent FGFR fusions manifested a distinct molecular subtype (TCGA, unpublished data). Using an arbitrary cutoff (centrality score > 0.37), thyroid, acute myeloid leukemia and cholangiocarcinoma were the top three cancer types with 75%, 68%, and 67% of fusions being predicted to have functional roles in these malignancies

(**Figure 4.10**). Using the same cutoff, we were able to find 6,175 in-frame fusion transcripts. In addition to established driver fusions such as BCR-ABL (0.396), MLL-MLLT4 (0.395), PML-RARA (0.39), ETV6-NTRK3 (0.39), SND1-BRAF (0.389), CCDC6-RET (0.387), and NCOA4-RET (0.388), we also identified novel recurrent fusions with high centrality scores, such as *AFF1-PTPN13* (0.372) in bladder cancer, *BUB1B-EIF2AK4* (0.376) in breast cancer, *UBTF-MAML3* (0.373) in pheochromocytoma and paraganglioma, *HIPK2-PARP12* (0.384), *SRGAP3-LHFPL4* (0.370) in uterine corpus endometrial carcinoma.

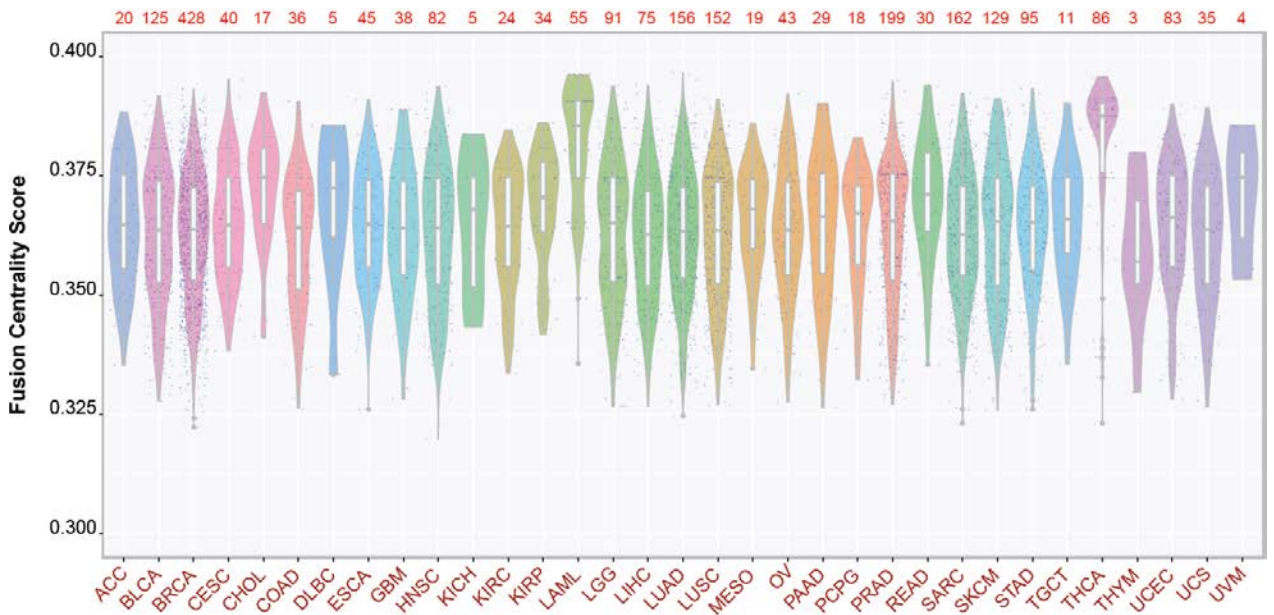


Figure 4.10A Centrality scores of fusion partner genes across 33 cancer types

The centrality distributions of all the cancer fusion partner genes after the filter procedure (6,175 in-frame fusion transcripts including 6,528 partner genes) in each cancer type are shown in a violin plot. The gray horizontal bars in the boxplot represent the mean centrality scores in each cancer type. The numbers of potential fusion drivers (centrality scores of >0.37) in each cancer type are labeled in red on the top.

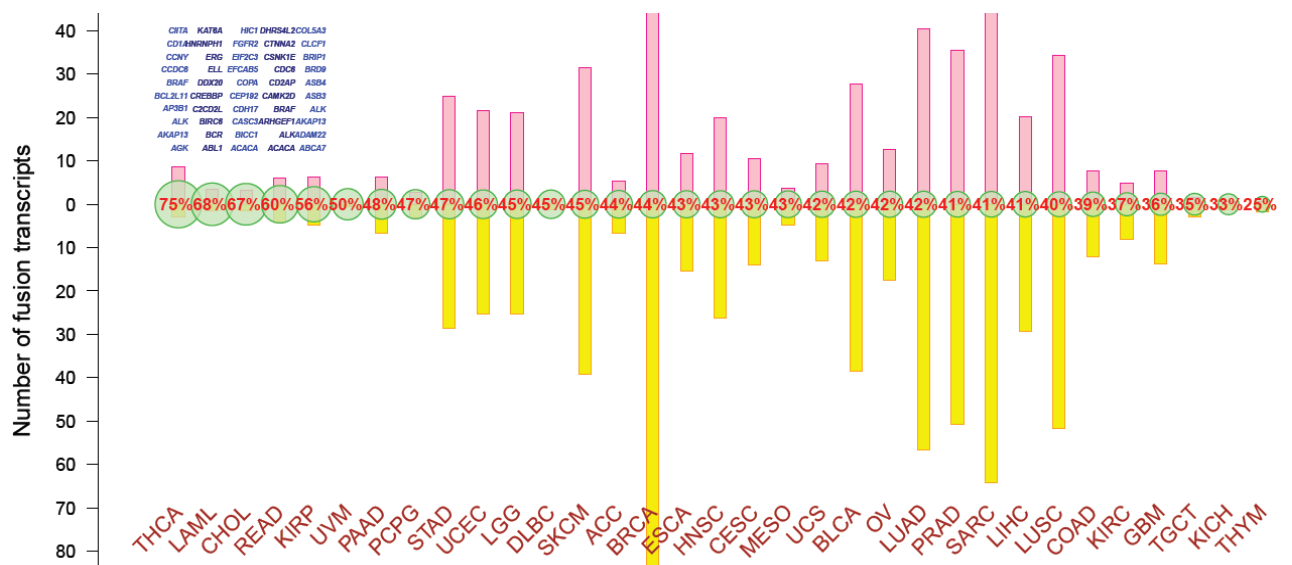


Figure 4.10B Proportion of driver fusions among all fusion transcripts across 33 cancer types

The numbers and distributions of partner genes that form the fusion transcripts are shown. The y-axis represents the number of patients found to have partner genes that form fusion transcripts in each cancer type.

We also accessed the predicted fusion transcripts derived from oncogenes and tumor suppressor genes. Based on the cosmic database of cancer gene census and cancer genome landscapes (Vogelstein, Papadopoulos et al. 2013), we found 412 in-frame fusion transcripts involved in tumor suppressor genes (TSGs) and 481 in-frame fusion genes comprised of oncogenes (OGs) in either partner genes. (**Figure 4.11**) The percentage of in-frame fusion transcripts involved in OGs is highest in THCA (29.9%) and LAML(19.2%), which is consistent with the proportion of driver genes in each cancer type (**Figure 4.10B**), implying fusion transcripts might play a dominant role in tumor progression in a subset of THCA and LAML, which could be applied for molecular classification and prognosis markers.

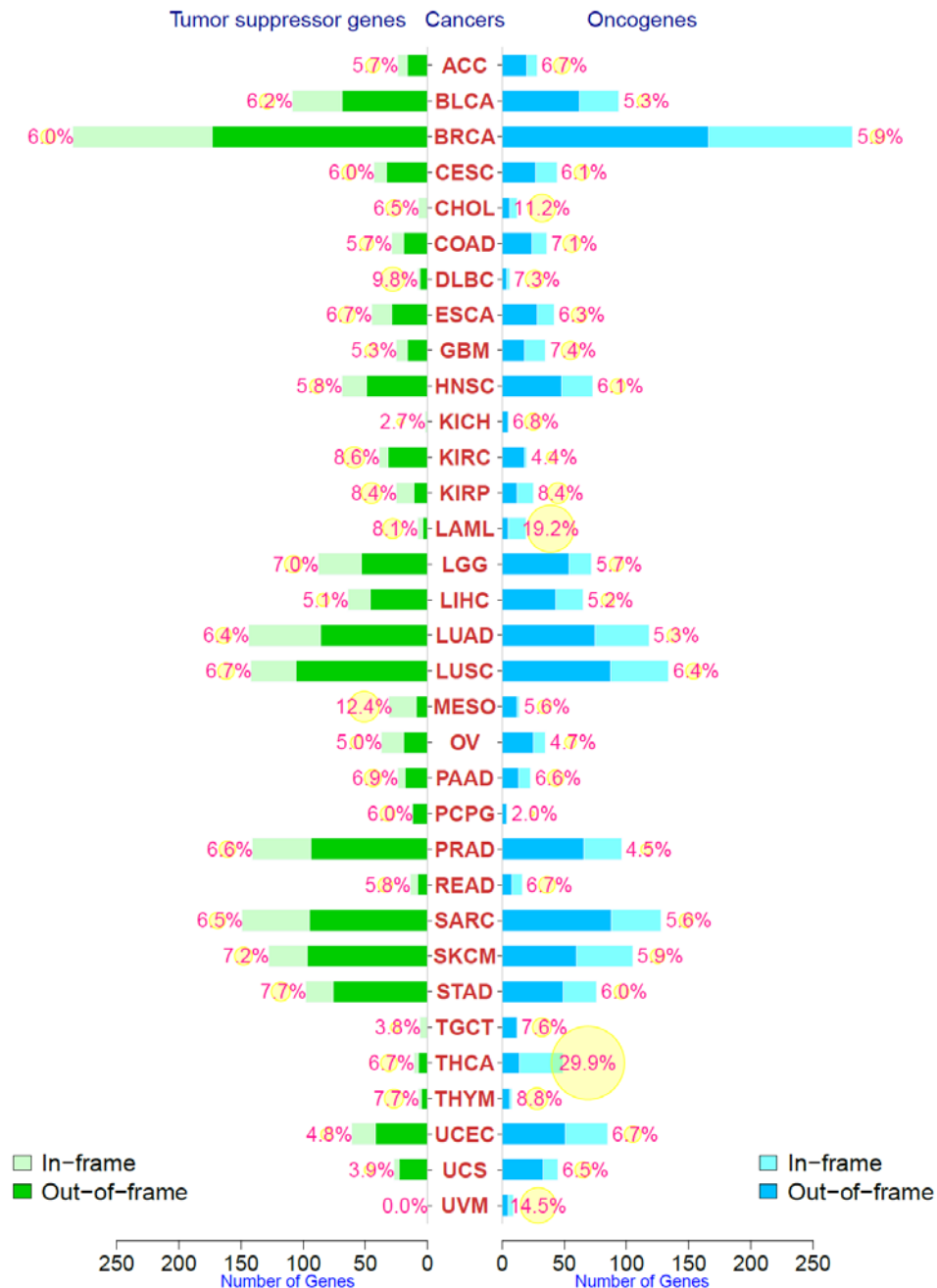


Figure 4.11 Number of fusion transcripts composed of tumor suppressor genes and oncogenes across 33 cancer types

The counts of fusion transcripts comprised of tumor suppressor genes or oncogenes in either gene partner are represented in the bar plot. The percentage of tumor suppressor gene/oncogene–formed fusions dominated by total fusion transcripts in each cancer type are denoted in yellow circles, with the circle size proportional to the ratio of oncogenes or tumor suppressor genes to all fusion transcripts in each cancer type. The in-frame and out-of-frame fusion transcripts were counted separately.

In addition, we sought to examine the effects of fusion transcripts on gene expression level of their partner genes. Using distance based outlier clustering; we have detected 1,723 fusion transcripts with significantly altered expression compared with that of remaining tumor samples in each cancer type, indicating the potential regulatory effects on the transcriptional level upon fusion events. Applying Kernel density functions, we found 2,835 fusion transcripts (75% quantile as critical value for outlier detection) displaying deviated expression level of their partner genes compared with their expression in those samples without corresponding fusion transcripts detected. For example, we observed the altered gene expression of partner genes in resulted fusion transcripts positive samples versus those samples without specific fusion transcripts detected. For example, in tumor samples detected with fusion gene *C10orf68-CCDC7* from multiple types of cancers including BLCA, BRCA, COAD, KIRC, LUAD, PRAD, UCEC, both partner genes *C10orf68* and *CCDC7* exhibited increased expression compared with their expression levels in samples without *C10orf68-CCDC7* fusion events(**Figure 4.12A**), while accompanied with overexpression of exons flanking to fusion junction points in several samples (**Figure 4.12B**), indicating the potential of those fusion events associated with overexpression and dysregulation of their partner genes.

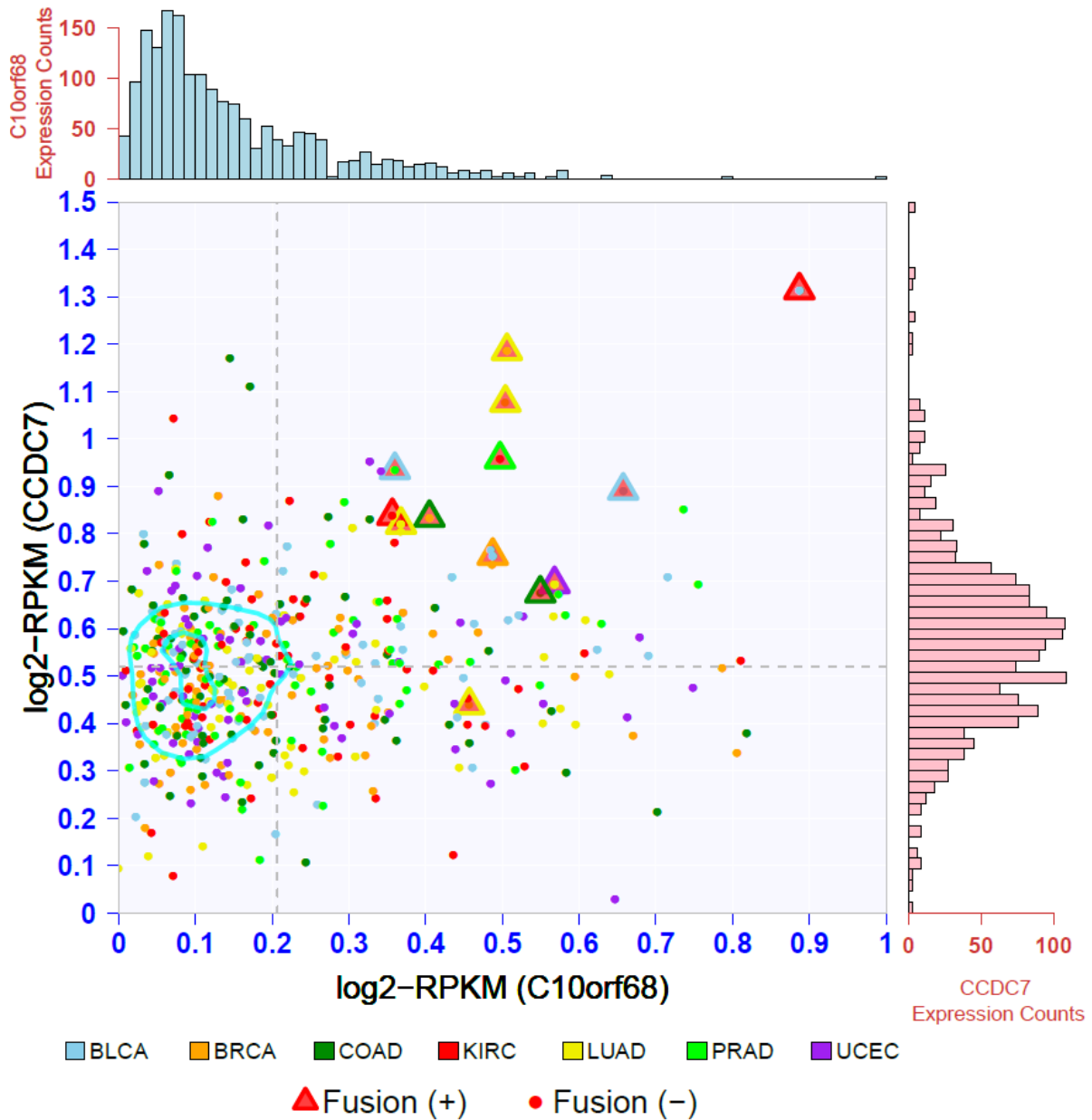


Figure 4.12A Increased transcription levels in recurrent fusion transcripts of *C10orf68*-*CCDC7* in multiple cancer types

The expression levels (log2RPKM) of the *C10orf68*-*CCDC7* fusion transcript in multiple cancer types (number of fusion transcripts in each cancer type: BLCA: n=3, BRCA: n=11, COAD: n=11, KIRC: n=4, LUAD: n=12, PRAD: n=2, UCEC: n=4) are shown. The dots show the expression of the partner genes *C10orf68* and *CCDC7* in each sample, the triangle shows the expression levels of *C10orf68* and *CCDC7* in the samples with *C10orf68*-*CCDC7* fusion transcripts, and the cyan smoothing lines show the kernel density estimation of transcript expression levels of partner genes with *C10orf68* and *CCDC7* in all samples with *C10orf68*-*CCDC7* fusion (n=45).

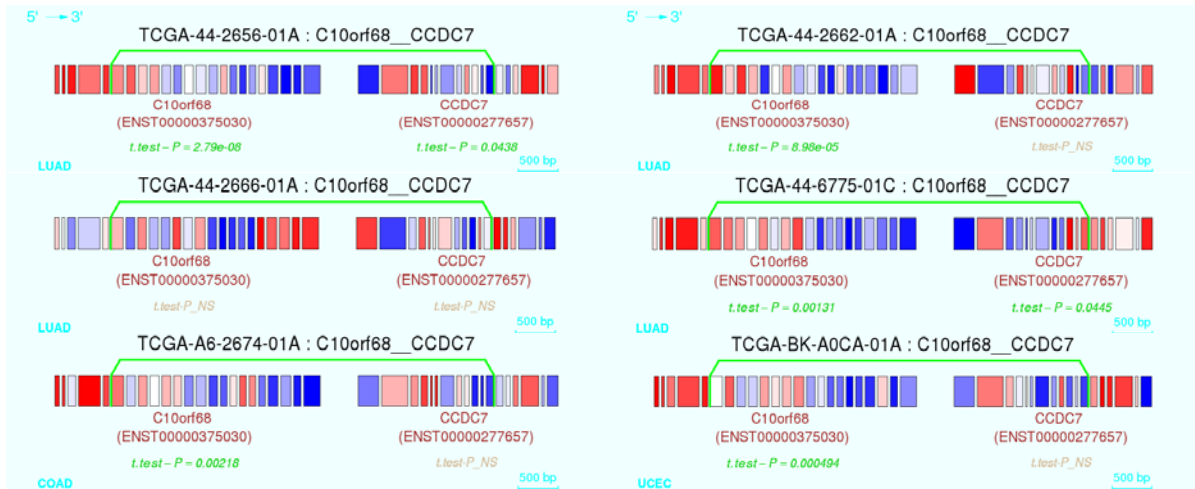


Figure 4.12B Z-normalized exon expression of C10orf68-CCDC7 in multiple cancer types

Blue and red represent relatively low and high level of expression based on the reads per kilobase per million mapped (RPKM) value of each exon.

4.3.5 Recurrent fusion transcripts in normal tissues

To explore the germline events of chimeric fusions, we have detected the fusion transcripts in 23 normal samples from the same patients that match to tumor samples applying the same fusion call method followed by filtering. We have identified 779 fusion transcripts in 742 normal tissue samples. Of these, 95% are fusion transcripts whose junction of either partner genes mapping to the region with germline copy number alterations based on the database of genomic variants, indicating that most fusion transcripts detected in normal are originated from the breakpoints located between the functional elements of two genes. We found a few recurrent fusion transcripts in different cancer types, such as *CHST5-TMEM231* in READ and *CRHR1-KIAA1267* in BLCA, HNSC, KIRC, PRAD, THCA, and these highly recurrent fusions also frequently occur in multiple cancer types, therefore are considered as germline events. Since these fusions are also implicated in previous GWAS analysis (Leslie, O'Donnell et al. 2014), indicating the germline events occur at both genomic and transcriptomic level (**Figure 4.13**).

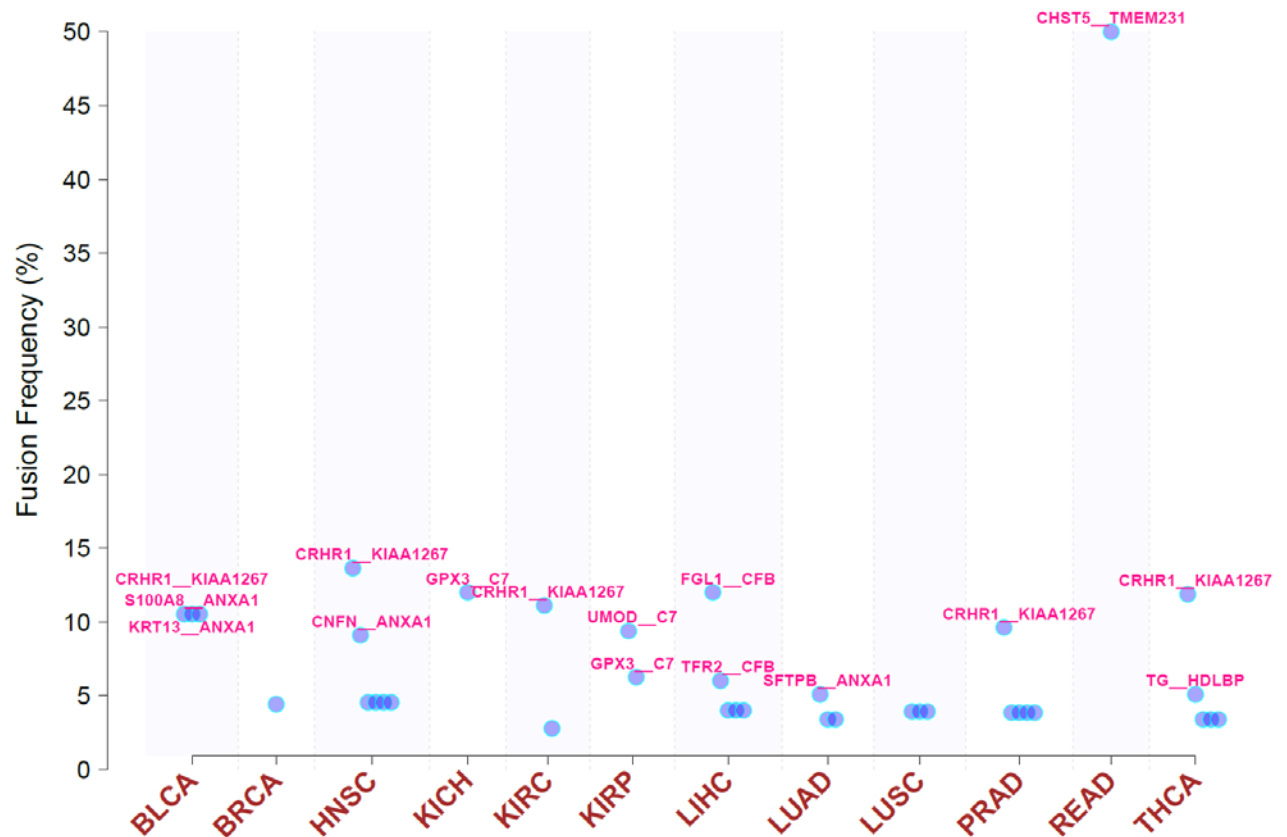


Figure 4.13 Frequency of recurrent fusion transcripts in normal tissue across 12 tissue types

The most recurrent ($n \geq 2$) fusion transcripts are shown, and the most prevalent fusion transcripts are labeled in red. The fusion frequency was defined as the number of samples with each recurrent fusion divided by the total number of normal samples analyzed in each tissue type.

4.3.6 Recurrent fusion transcripts are mutually exclusive with somatic mutations

We selected the samples with both fusion and mutation data sets available to compare the mutation frequency between the samples with or without fusion transcripts constitute the specific gene. The emerging patterns of mutual exclusivity between gene fusions and somatic mutations in their partner genes are observed across multiple cancer types. For example, *TYK2* kinase was detected to be fused with 42 different partner genes in 17 cancer types, based on the presence of *TYK2* constitute fusion, we have divided the samples from different cancer types into two groups, notably the overall mutation frequency

is significantly lower in fusion positive group than that in fusion negative group, (Welch's t-test, p-value =0.0055), and there are few somatic mutations occurring in those samples with *TYK2* containing fusions (**Figure 4.14A**), similar phenomena in *EIF2AK2* comprised fusion pairs (Welch's t-test, p-value =0.0071), (**Figure 4.14B**), this finding has set the framework to unravel the genomic basis of fusion transcripts. The common trend between recurrent, in-frame fusion transcripts and reduced number of somatic mutations in the same patients is seen in many cancer types, implying the properties of fusion events occurring in those patients drive cancer growth and progression instead of somatic mutations.

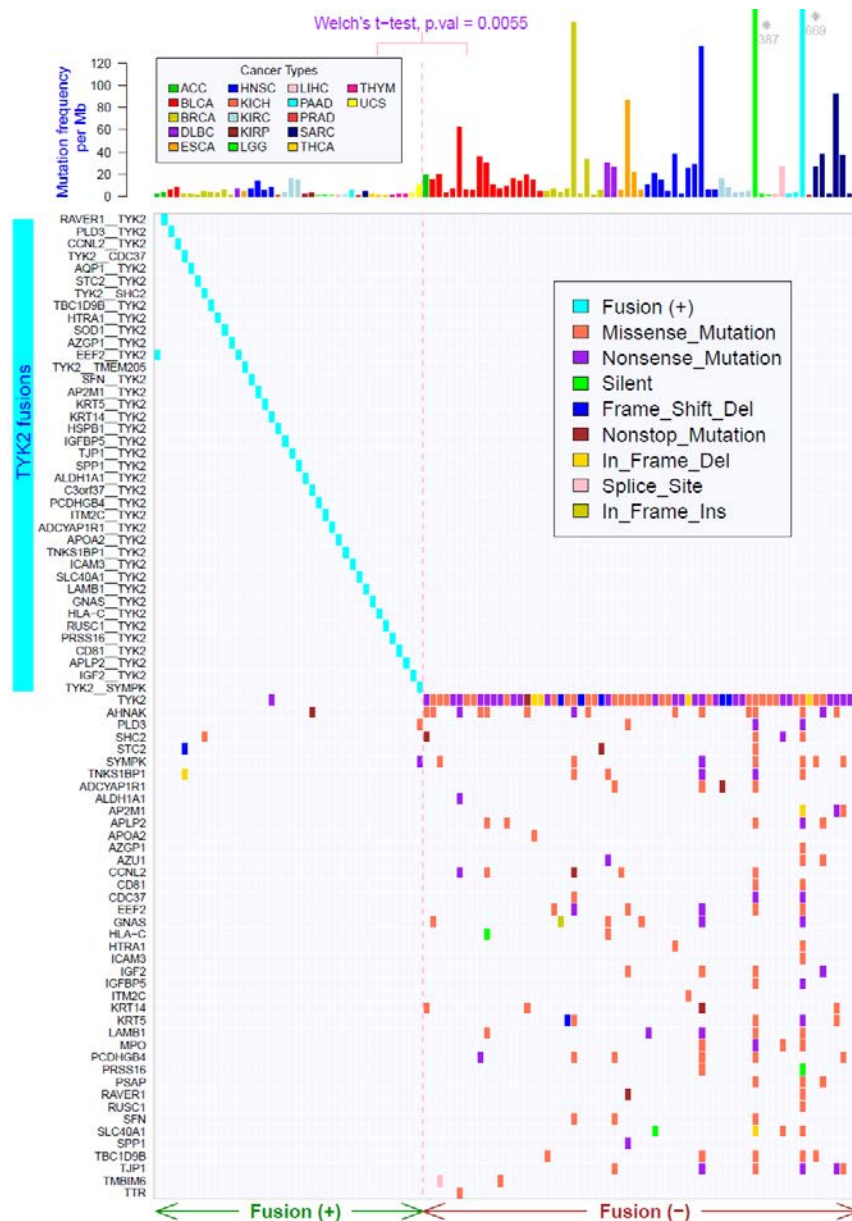


Figure 4.14A Recurrent *TYK2*-containing fusion transcripts are mutually exclusive with somatic mutation events in multiple cancer types

The *TYK2*-containing fusion transcripts in multiple cancer types are shown. (ACC: n=2, BLCA: n=2, BRCA: n=8, DLBC: n=1, ESCA: n=1, HNSC: n=2, KICH: n=1, KIRC: n=5, KIRP: n=2, LAML: n=4, LGG: n=3, PAAD: n=3, PRAD: n=1, SARC: n=1, THCA: n=2, THYM: n=3, UCS: n=2.) The top panel represents the mutation frequencies of the somatic mutations; Welch's t-test was performed to compare the mutation frequency between samples with or without fusion transcript-containing *TYK2* kinase. The bottom panel illustrates the fusion and somatic mutations detected with *TYK2* and its partner fusion genes in the samples with and without *TYK2* fusion transcripts detected.

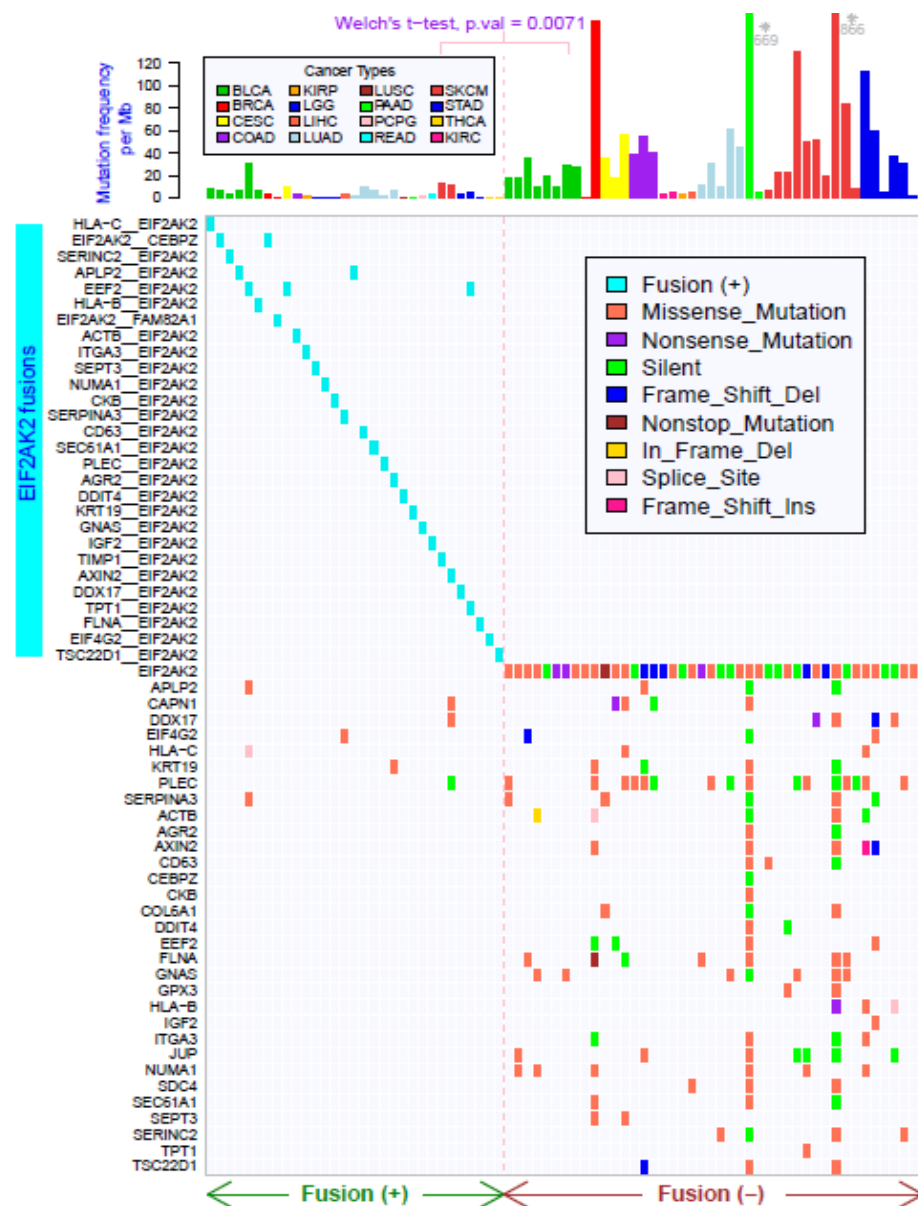


Figure 4.14B Recurrent *EIF2AK2*-containing fusion transcripts are mutually exclusive with somatic mutation events in multiple cancer types

The *EIF2AK2*-containing fusion transcripts in multiple cancer types are shown. (BLCA: n=6, BRCA: n=2, CESC: n=1, COAD: n=1, KIRC: n=3, KIRP: n=2, LGG: n=3, LIHC: n=1, LUAD: n=5, LUSC: n=4, PAAD: n=1, PCPG: n=1, READ: n=1, SKCM: n=2, STAD: n=4, THCA: n=2.) The top panel indicates the mutation frequencies of the somatic mutations; Welch's t-test was performed to compare the frequency between the samples with or without fusion transcript-containing *EIF2AK2* kinase. The bottom panel illustrates the fusion and somatic mutations detected with *EIF2AK2* and its partner fusion genes in the samples with or without *EIF2AK2* fusion transcripts detected.

4.3.7 Perturbed pathways associated with fusion formation

The fusion transcripts largely triggered pathway perturbations through which the functional domains in partner genes were truncated by fusion formation or recombined with regulatory domains donated by their partner genes. We explored the enriched gene ontology terms and reactome pathways of all fusion transcripts in each cancer type. We found that the perturbed pathways were enriched for kinase activity and small GTPase signaling pathways and for transferase and transmembrane receptors in most cancer types (**Figure 4.15**).

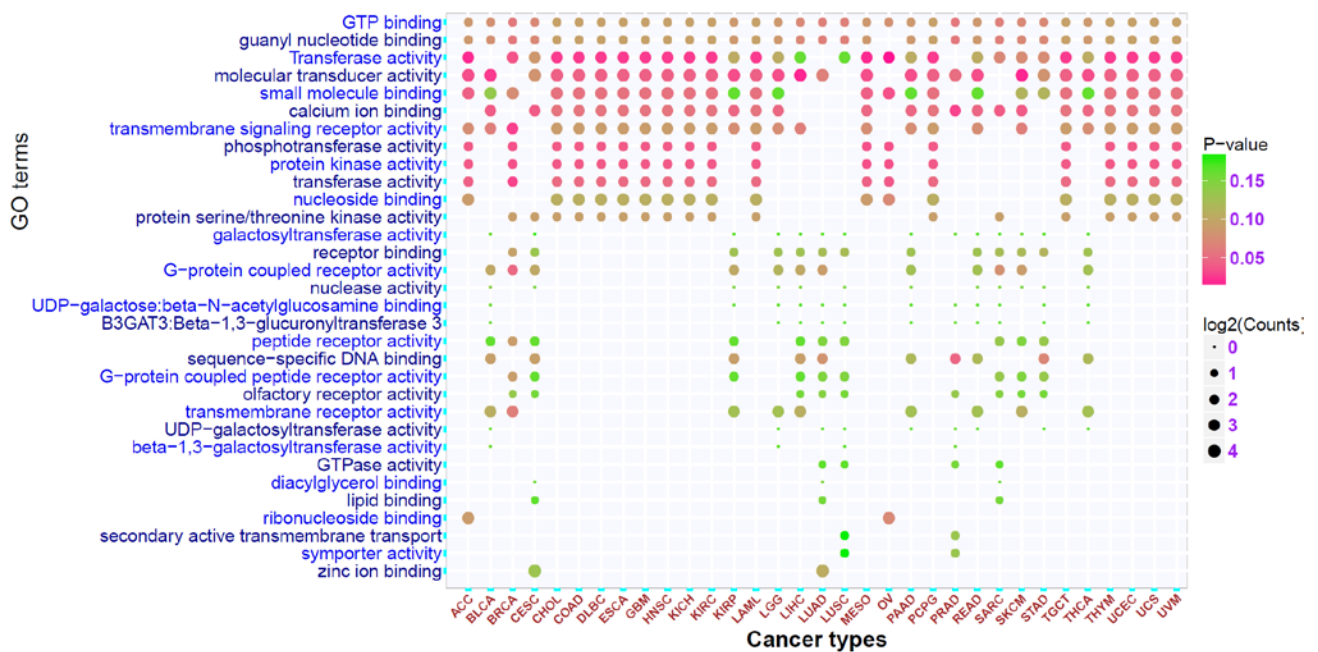


Figure 4.15 Most represented ontological categories in gene ontology analysis of fusion transcripts in 33 cancer types

The perturbed gene ontology terms enriched at fusion transcripts in each cancer type are shown in dots; the colors of the dots represent the adjusted p value of gene ontology terms from partner genes of fusion transcripts, and the size of the dots represents the number of genes annotated in each gene ontology term.

VEGF-VEGF receptor pathways and axon guidance and PDGF/NGF pathways, as well as chromatin remodeling pathways, were predominantly perturbed in urothelial bladder carcinoma, breast cancer, lung cancer, and head and neck cancer, suggesting that similar

pathways are perturbed in association with tumor genesis in these cancer types. In sarcomas, only a few pathways were disrupted, including SUMOylation, TGF-beta receptor, and ECM interaction pathways, despite the high frequency of fusion formations in sarcomas, indicating the heterogeneous effects of fusion events in different cancer types (**Figure 4.16**).

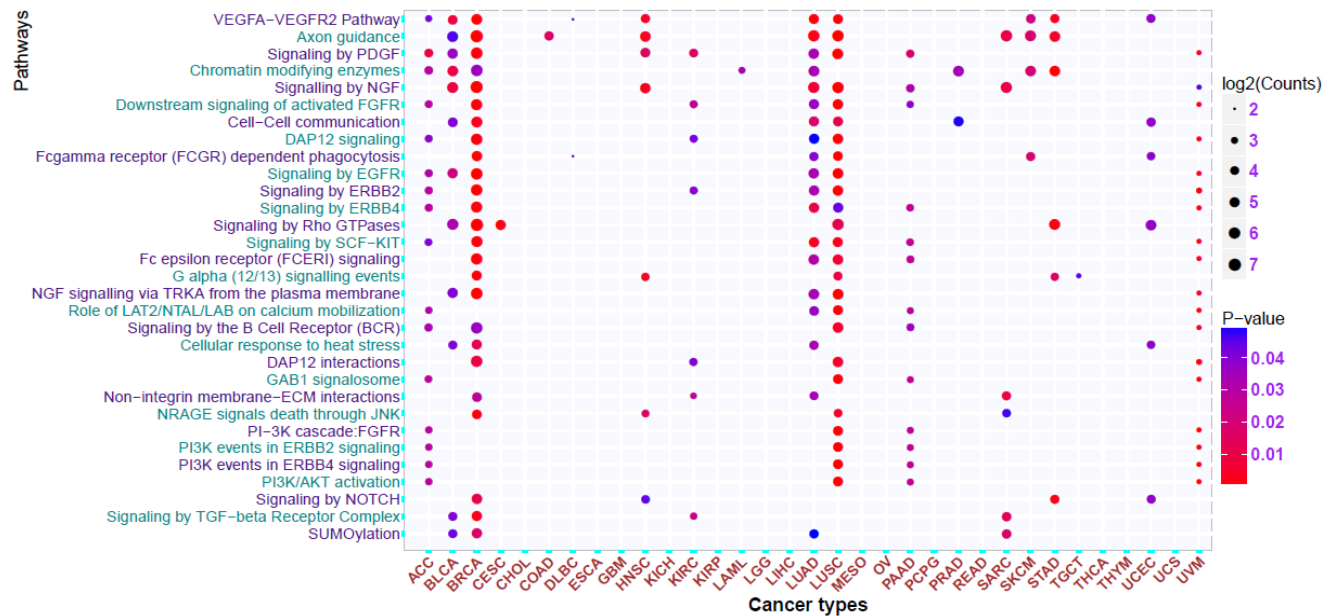


Figure 4.16 Perturbed pathways from fusion transcripts across 33 cancer types

The reactome pathways of fusion transcripts enriched in each cancer type are shown in dots; the colors of the dots represent the adjusted p-value of the enriched pathway from partner genes of fusion transcripts, and the size of the dots represents the number of genes annotated in each pathway.

4.3.8 Cancer type-specific fusion gene networks and hubs

Some fusion transcripts lead to biologic perturbations in which partner genes recombine with a large number of alternative partner genes, and a mutual fusion partner may link such genes, resulting in the clusters of interrelated fusion transcripts. We built the occurrence of degree in each cancer type to determine empirical hub genes. In cancer types with 100 to 1000 fusion transcripts detected, we observed several common hub genes shared by multiple cancer types, such as *STAT6* and *EEF2* (**Figure 4.17**).

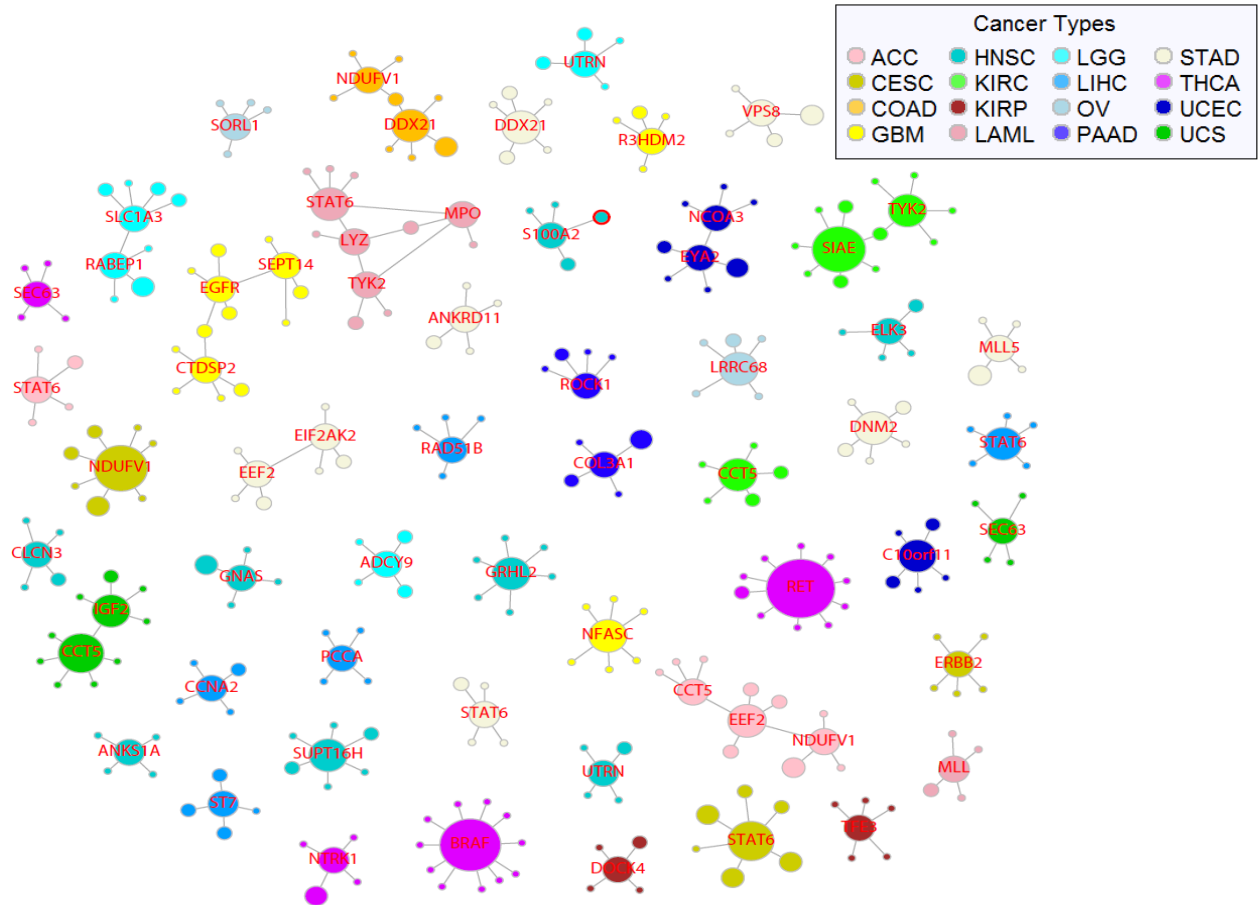


Figure 4.17 Network of gene fusions in multiple cancers

All the unique fusion pairs that include more than three partner genes in 16 cancer types are plotted. The nodes (circles) indicate genes, and edges (gray lines) indicate the occurrence of a fusion transcript between genes. Node size indicates the number of partner genes with which the node/gene fuses. The hub genes that are interconnected with more than three partner genes are labeled in red.

In the cancer types with a high frequency of fusion transcripts, we observed an intertwined network of fusion transcripts, with *FRS2* as central hub gene predominantly fused with multiple partner genes in sarcoma and BRCA. *FRS2* is a tyrosine kinase adaptor protein in FGFR pathway, inducing downstream activation of Ras-MAPK pathway (Luo, Kim et al. 2015), suggesting its oncogenic potential upon fusion events. We also found that *RARA* and *BCAS3* were hub genes in BRCA; *BCAS3* (breast cancer anti-estrogen resistance gene 3)

Among 717 fusion transcripts comprising 1,259 unique partner genes detected in liver cancer, only five genes (*ST7*, *STAT6*, *CCNA2*, *PCCA*, and *RAD51B*) recombined with at least four partners, while there were few interconnected networks, and 1,151 (91.4%) of the partner genes were not part of any network (**Figure 4.19**).

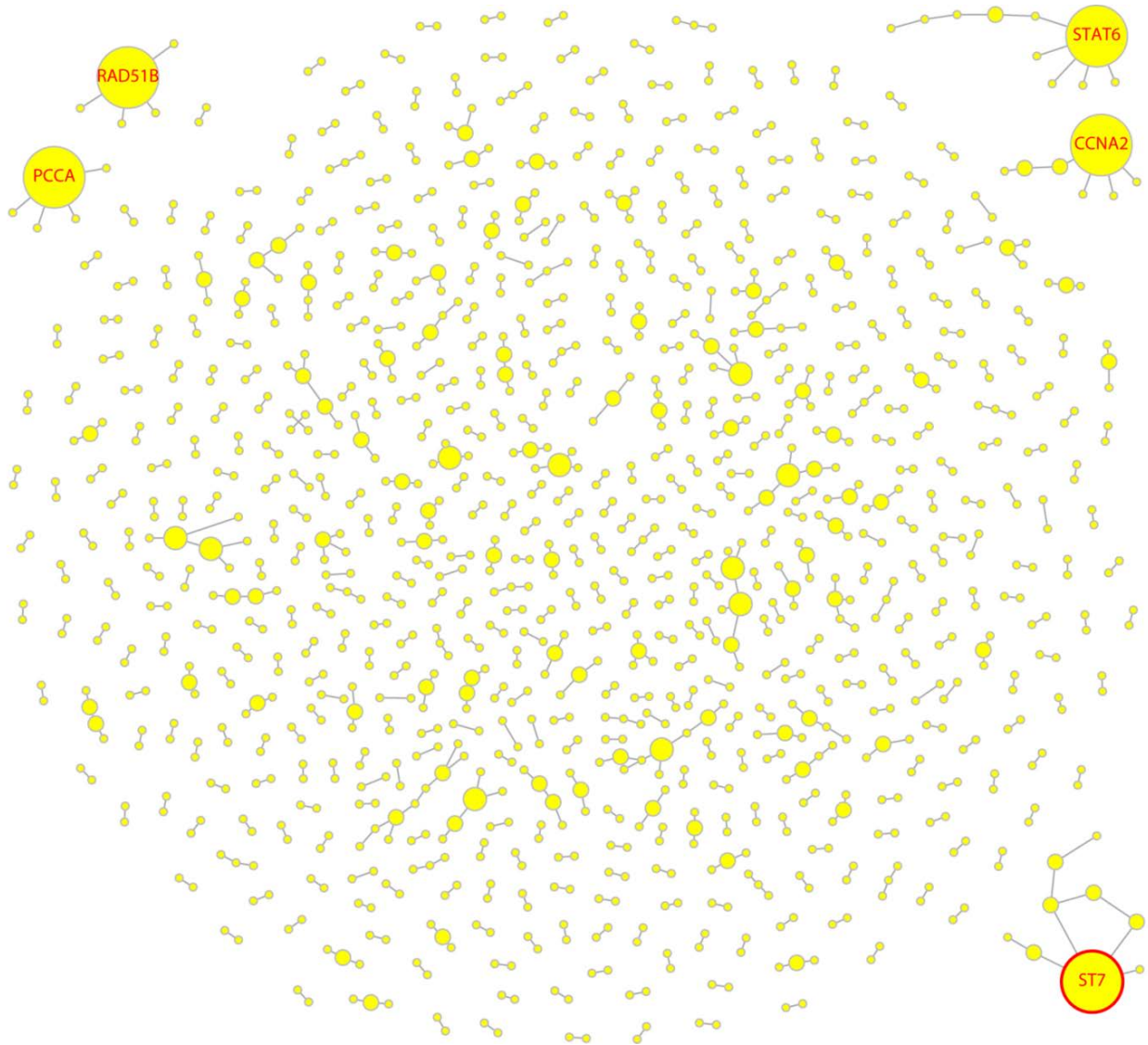


Figure 4.19 Network of gene fusions in liver cancer (LIHC)

All the unique fusion pairs including all of their partner genes in LIHC are plotted. The nodes (yellow circles) indicate genes, and edges (gray lines) indicate the occurrence of a fusion transcript between genes. Node size indicates the number of partner genes with which the node/gene fuses. The hub genes that are interconnected with more than four partner genes are labeled in red.

A similar pattern was found in stomach adenocarcinoma, with only a few genes, such as *MLL5*, *DNM2*, and *DDX21*, fused with multiple partners (**Figure 4.20**). Fusion-triggered network alteration may also occur via node loss or changes of regulatory functions in the context of loss or a shift in protein functional domains.

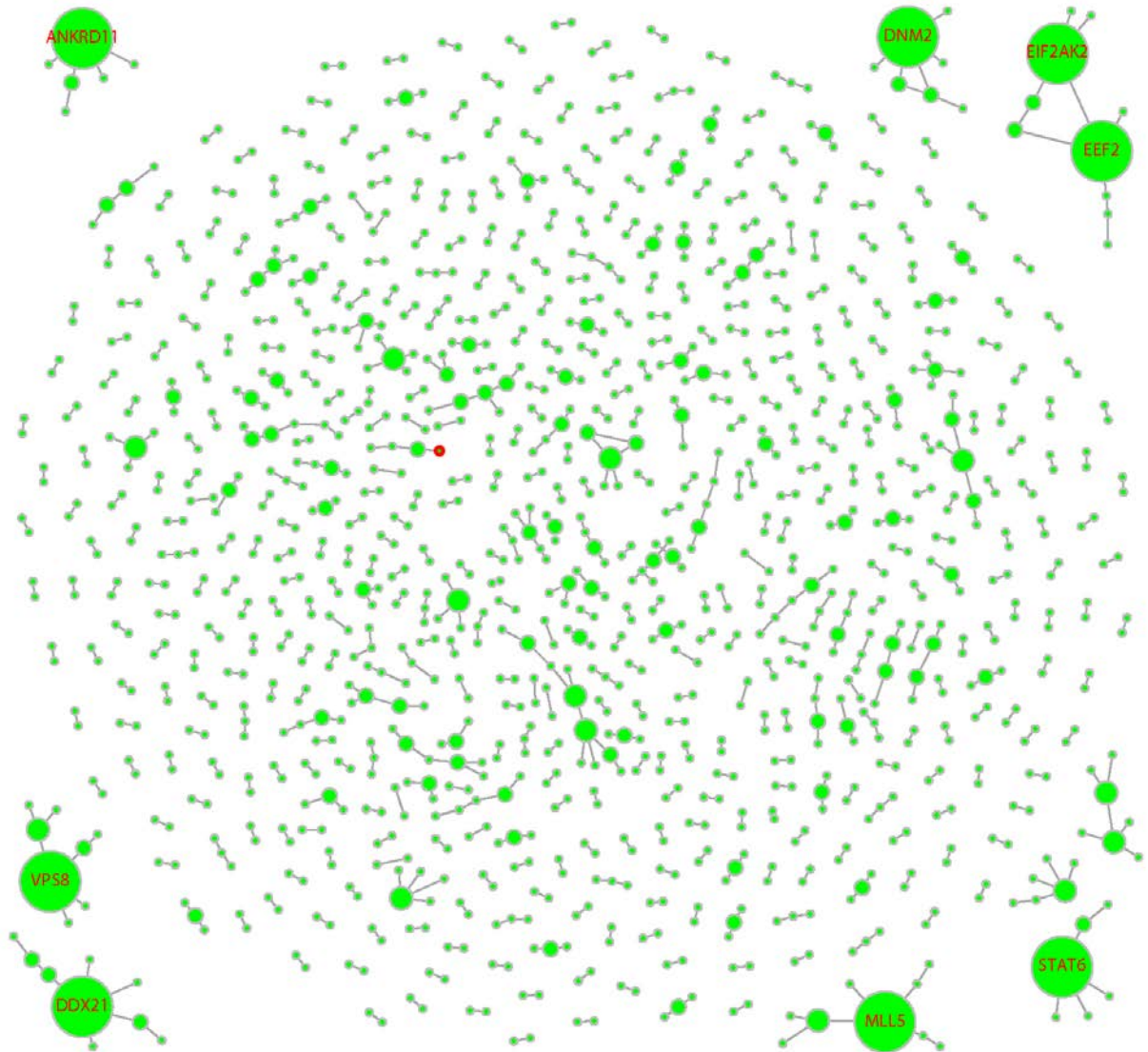


Figure 4.20 Network of gene fusions in stomach adenocarcinoma (STAD)

All the unique fusion pairs including all their partner genes in STAD are plotted. The nodes (green circles) indicate genes, and edges (gray lines) indicate the occurrence of a fusion transcript between genes. Node size indicates the number of partner genes with which the node/gene fuses. The hub genes that are interconnected with more than four partner genes are labeled in red.

4.3.9 Altered tyrosine kinase domains in predicted fusion transcripts may deregulate oncogene functionality

Tyrosine kinase fusion genes consist of a subset of therapeutic relevant oncogenes associated with both circulating and solid tumors (Medves and Demoulin 2012). In some of these kinase fusions, the kinase domains in the fusion oncogenes fuse with partner genes harboring regulatory domains and consequently loss of response to ligand and become constitutively activated through ligand-independent dimerization and/or oligomerization induced by their fusion partners. In other cases, the kinase domains are fused with their partner genes, leading to intracellular delocalization of the kinase domains, and the properties of crossing the cell membrane is needed for therapeutic antibodies against the original kinase targets. Using as references the Swiss-Prot Protein database (<http://www.uniprot.org/docs/pkinfam>), the Human Protein Reference Database (<http://www.hprd.org/>), and the Pfam database (<http://pfam.xfam.org/>), we identified 1,969 fusion transcripts harboring a protein kinase, including 752 in-frame protein kinase fusions. Most in-frame kinase fusions belonged to the Ser/Thr protein kinase family, followed by the tyrosine kinase family. We identified 152 in-frame fusion transcripts harboring a kinase in either partner gene whose kinase domains remained intact upon fusion events, suggesting potential oncogenic functions under modulation by their partner genes. Interestingly, we identified *PDGFRA-USP8* in sarcomas (**Figure 4.21**), in which tyrosine kinases from platelet-derived growth factor receptor (*PDGFRA*) flanked adjacent to the coiled-coil domain and Rho-related GTP-binding domain (RHOD) domain from *USP8*. Charged residues in the coiled-coil domain may function as a hinge that allows the N-terminal kinase domains to interact with C-terminal RHOD, a sulfurtransferase involved in cyanide detoxification, and coiled-coil interactions may mediate constitutive multimerization, mitochondrial binding, and kinase activity of its adjacent protein kinase. Notably, the expression level in exons of the *PDGFRA-USP8* fusion transcript was significantly higher than that in their parental

exons, indicating that the novel fusion *PDGFRA-USP8* may trigger structural and functional alterations in *PDGFRA*, whose oncogenic properties have critical implications in cancer (Velghe, Van Cauwenberghe et al. 2014).

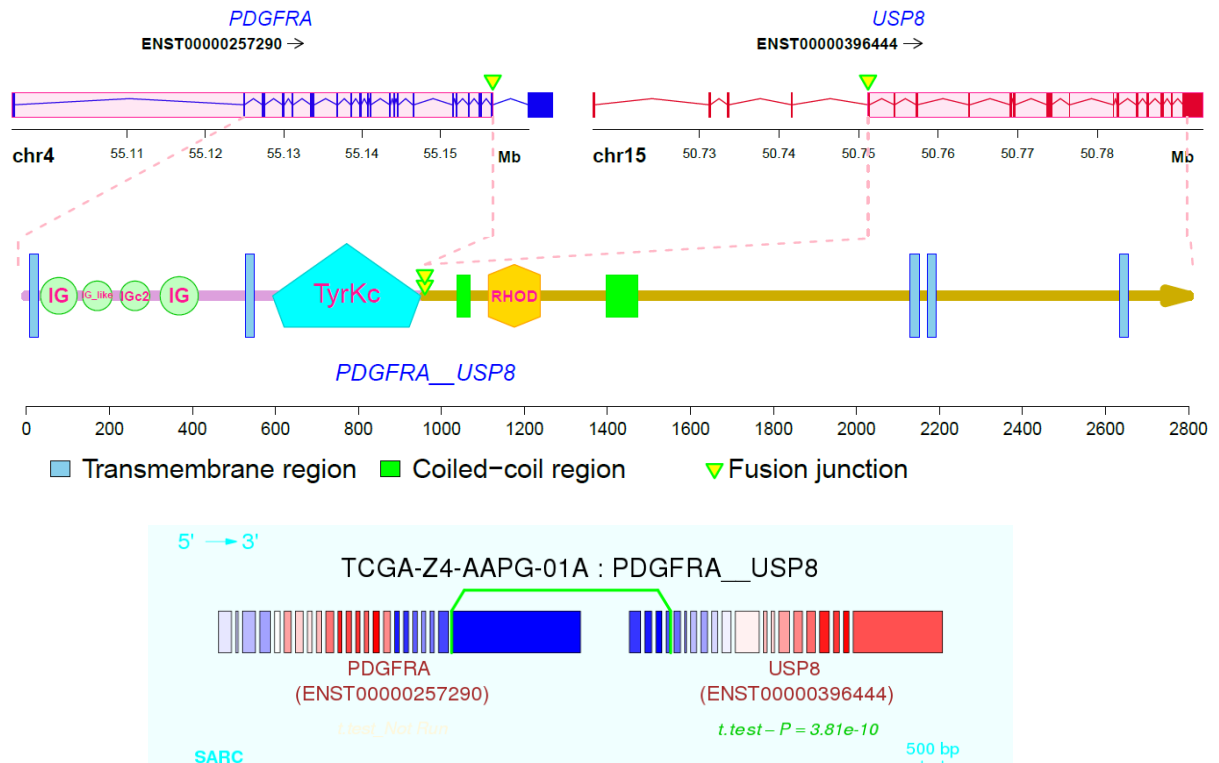


Figure 4.21 Novel fusion transcript *PDGFRA-USP8* harboring kinase domains in sarcomas

The top panel showed the partner genes *PDGFRA* and *USP8* are illustrated in the top panel. The red frame represents the remaining fragment of two transcripts that recombined into the *PDGFRA-USP8* fusion protein. The protein structure domains are illustrated in the bottom panel, with the tyrosine kinase domain from *PDGFRA* flanked closely to the coiled-coil domain from *USP8*. The bottom panel showed Z-normalized expression for each exon of *PDGFRA* and *USP8* in sarcomas. Red and blue represent relatively high exon expression and low exon expression, respectively.

Another intriguing kinase fusion, composed of *protein kinase C beta* (*PRKCB*), which is reported to promote mammary tumorigenesis in the tumor microenvironment (Wallace, Pitarresi et al. 2014), was fused with different gene partners in multiple types of cancer including PRAD, LUSC, LUAD, LGG, and GBM. Thus, this fusion was predicted to truncate

at the N-terminal part of *PRKCB* harboring an auto-inhibitory domain and consequently to cause constitutive activation of *PRKCB*. The exon expressions of *PRKCB*-containing fusions displayed a similar pattern, with higher expression levels in fusion regions than in parental exons (**Figure 4.22**). Taken together, these findings reveal a set of fusion genes with kinase domains conferring critical roles in tumor progression in multiple types of cancers.

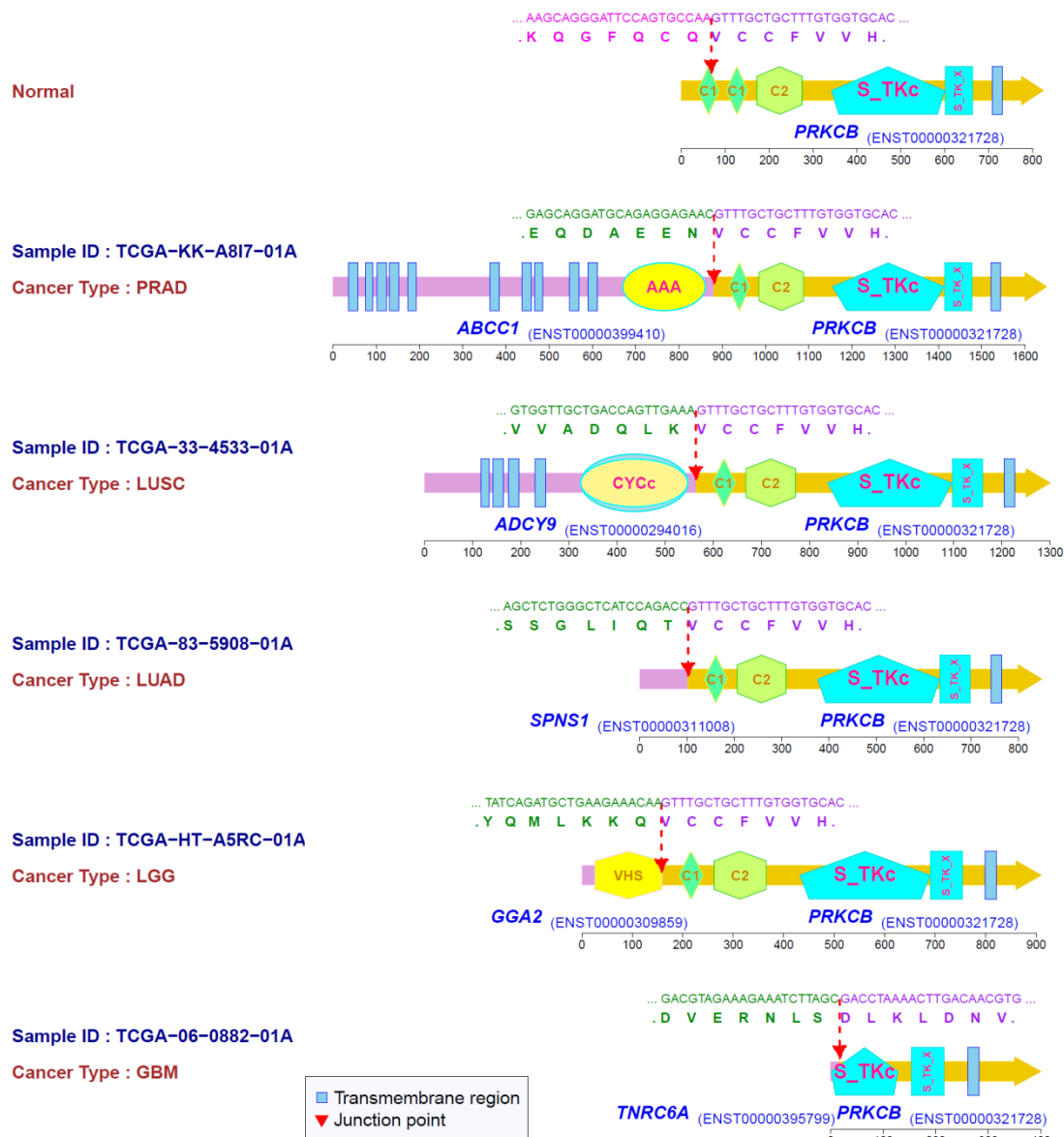


Figure 4.22 Novel fusion transcript involved kinase PRKCB in different cancer types

The protein structure of *PRKCB* and its partner genes in five cancer types are illustrated in the top panel. The TCGA sample ID, cancer type, and coding protein domains in fusion genes, as well as nucleic acid and amino-acid sequences, are annotated for each predicted *PRKCB* kinase fusion protein. The 5' partner genes are denoted in the purple segment, and the 3' partner gene is denoted in the gold segment. In addition, the fusion breakpoints are denoted by red arrows. Serine/threonine protein kinase (S_TKc) domains and kinase extension domains are in cyan; the transmembrane domains (TM) are in blue rectangles, C1 and C2 conserved regulatory domains are in green polygons, and other domains are in yellow. Other SMART Domain annotation: AAA, ATPases; CYCc, guanylate cyclases that catalyze the formation of cGMP from GTP; VHS, membrane targeting/cargo recognition; C1/2, protein kinase C conserved region domains.

4.3.10 Loss of epigenetic modification domains in predicted fusion transcripts may deregulate activity of tumor suppressor genes

Chromatin remodeling proteins function as gatekeepers and constitute a major determinant for transcriptional regulation, triggering a wide repertoire of biologic functions. Loss or gain of chromatin remodeling domains confers a unique ability to reprogram the cancer genome to alter oncogenic phenotypes (Muntean and Hess 2009) (Nair and Kumar 2012). Applying the epigenetic modifier database dbEM (<http://crdd.osdd.net/raghava/dbem/>), we found that 2,688 (13.4%) fusion transcripts constituted epigenetic modifiers in one of the partner genes. There were 26 in-frame fusion transcripts involving chromatin modification domains with potential loss of functions in tumor suppressor genes. A typical novel fusion transcript, *MANF-SETD2*, could lead to loss of the SET domain in *SETD2* in BRCA, and several lines of study have shown that SETD2 acts as an epigenetic modifier with tumor suppressor properties (Li, Duns et al. 2016). In addition, SETD2 plays important roles in maintaining genome integrity, and its loss of function could foster branched evolution through impaired DNA repair and replication stress in renal cancer (Kanu, Gronroos et al. 2015). Notably, the significantly increased expression of SETD2 in exons flanking the resulting fusion junction point was observed in several samples, demonstrating the potential of *MANF-SETD2* as a potential oncogenic driver (**Figure 4.23**). Eight in-frame fusion transcripts involved chromatin modification domains with oncogenic properties, and the function and stability of these onocogenic domains are potentially modulated by the domains donated from its partner genes.

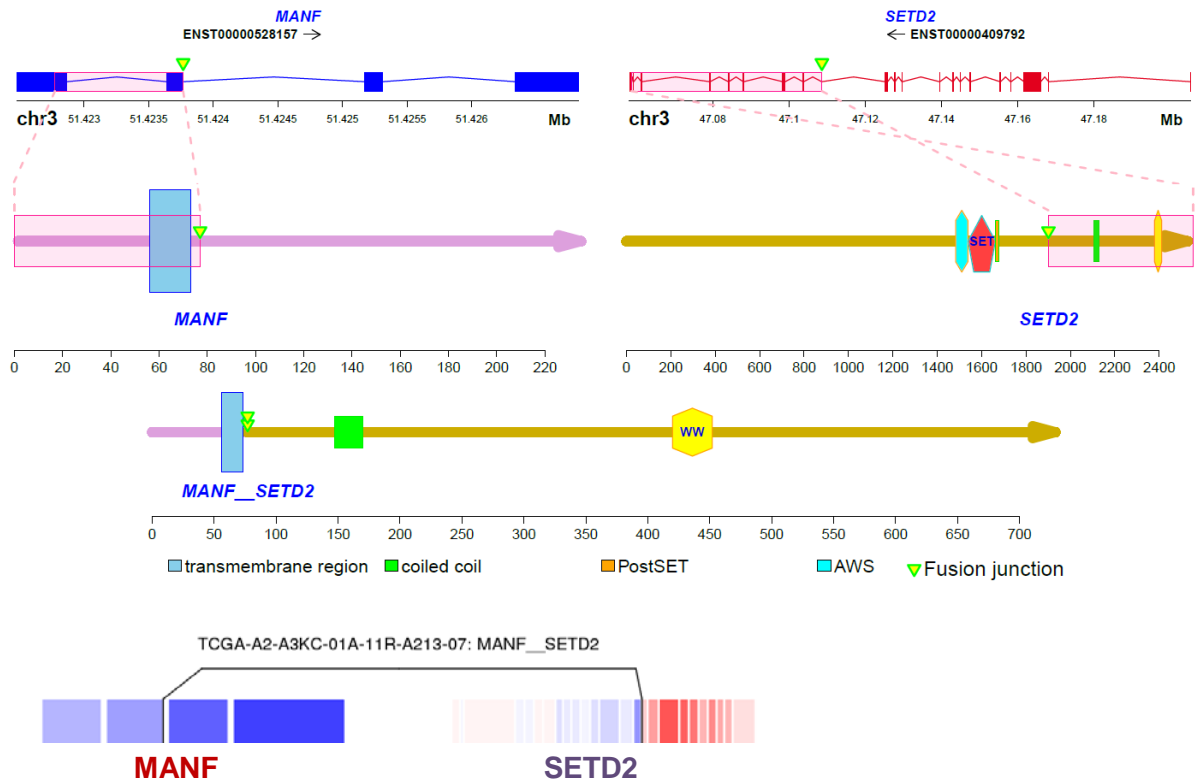


Figure 4.23 Novel fusion transcript affects chromatin remodeling domain (SET) in BRCA

The top panel showed both transcriptomic (upper) and protein (bottom) structures of parent genes and the fusion gene *MANF-SETDD2* are shown. The SET domain of *SETD2* is depleted in the predicted fusion sequence. Ensembl transcript ID and genomic and protein coordinates are indicated. The bottom panel showed the exon expression of *MANF-SETD2* in BRCA.

Another example of chromatin modifier involved fusion is tumor suppressor gene *ARID1B*, which codes an A/T-rich interaction domain as a subunit of SWI/SNF complex, as a 5' partner gene. Upon fusing with multiple partner genes in BRCA, BLCA, LUAD, UCEC, LGG, GBM, and SKCM, *ARID1B* lost the ARID domain upon the fusion event, accompanied by reduced expression of *ARID1B* at both the exon and the transcript level. Its loss of function is implicated in multiple types of cancers (Aso, Uozaki et al. 2015), indicating that *ARID1B* fusion confers bona-fide tumor suppressor properties (**Figure 4.24**).

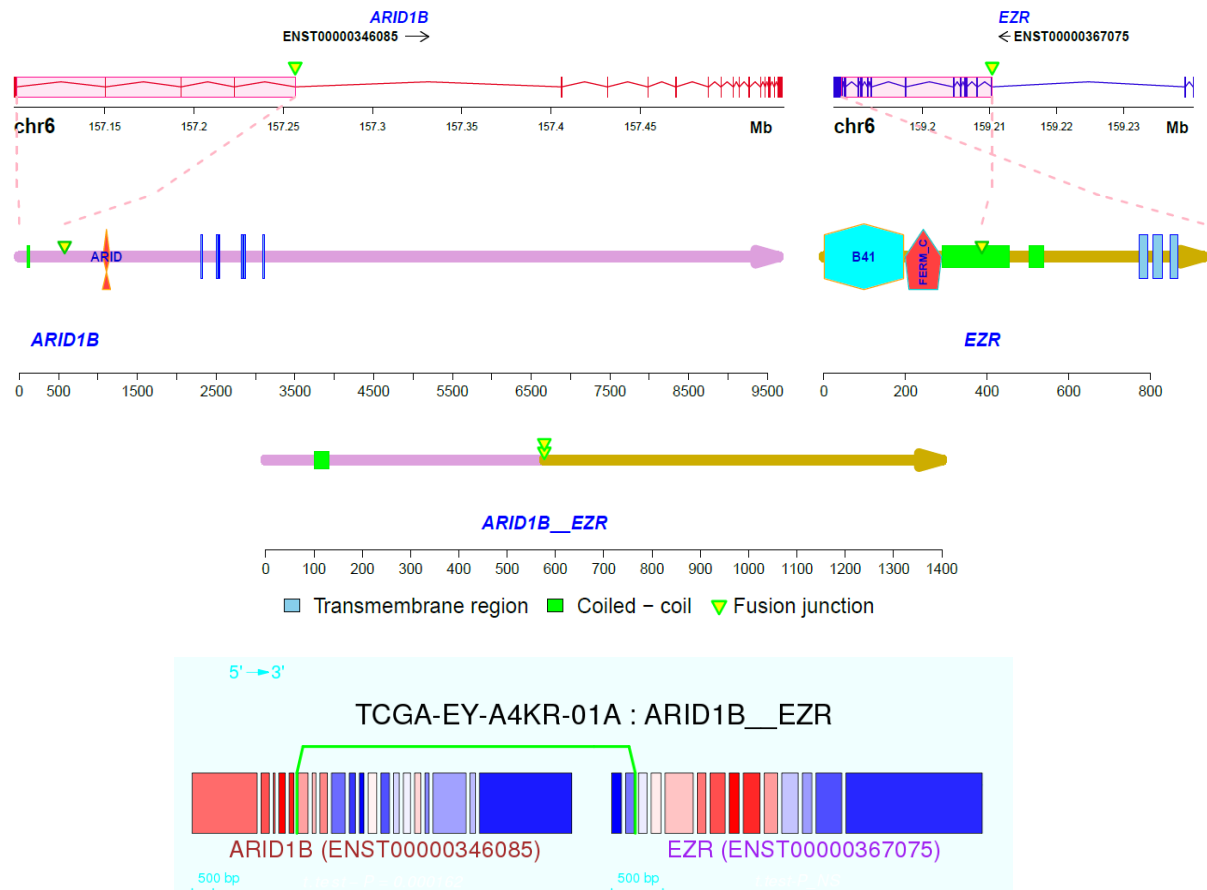


Figure 4.24 Novel fusion transcripts *ARID1B-EZR* in Uterine Corpus Endometrial Carcinoma (UCEC)

The top panel showed both transcriptomic (upper) and protein (bottom) structures of parent genes and the fusion gene *ARID1B-EZR* are shown. The ARID domain of *ARID1B* is depleted in the predicted fusion sequence. Ensembl transcript ID and genomic and protein coordinates are indicated. The bottom panel showed the exon expression of *ARID1B-EZR* in Uterine Corpus Endometrial Carcinoma (UCEC).

4.3.11 Gain or loss of post-translational modification sites in fusion may deregulate oncogenes and tumor suppressor genes

Post-translational modification sites play important roles in regulatory switches in protein functions and pathways. Frequent mutations in ubiquitination sites are implicated as novel driver mechanisms involved in carcinogenesis, including angiogenesis, and post-translational modification-specific mutations are associated with decreased patient survival

(Narayan, Bader et al. 2016). We explored the loss or gain of ubiquitination sites in fusion transcripts composed of oncogenes or tumor suppressor genes, which could upregulate oncogene activity or downregulate tumor suppressor gene activity owing to the role of ubiquitination sites in mediating protein stability and degradation (Mani and Gelmann 2005, Kirkin and Dikic 2011). Taking advantage of ubiquitination site predictor CKSAAP_UbSite (http://protein.cau.edu.cn/cksaap_ubsite) (Chen, Zhou et al. 2014, Chen, Zhou et al. 2015), we found that four fusion proteins constitute oncogene *NET1* loss ubiquitination sites in PRAD, TGCT, STAD (5' partner gene), and BLCA (3' partner gene). As a typical example, the specific pattern of segment retention in *ATP11B-NET1* fusion proteins leads to ubiquitination site loss at the N-terminal of *neuroepithelial cell transforming 1* (*NET1*), which may confer increased stability of *NET1*, a RhoA guanine exchange factor, to exert its oncogenic function of promoting motility and metastasis (Bennett, Sadlier et al. 2011) (**Figure 4.25**).

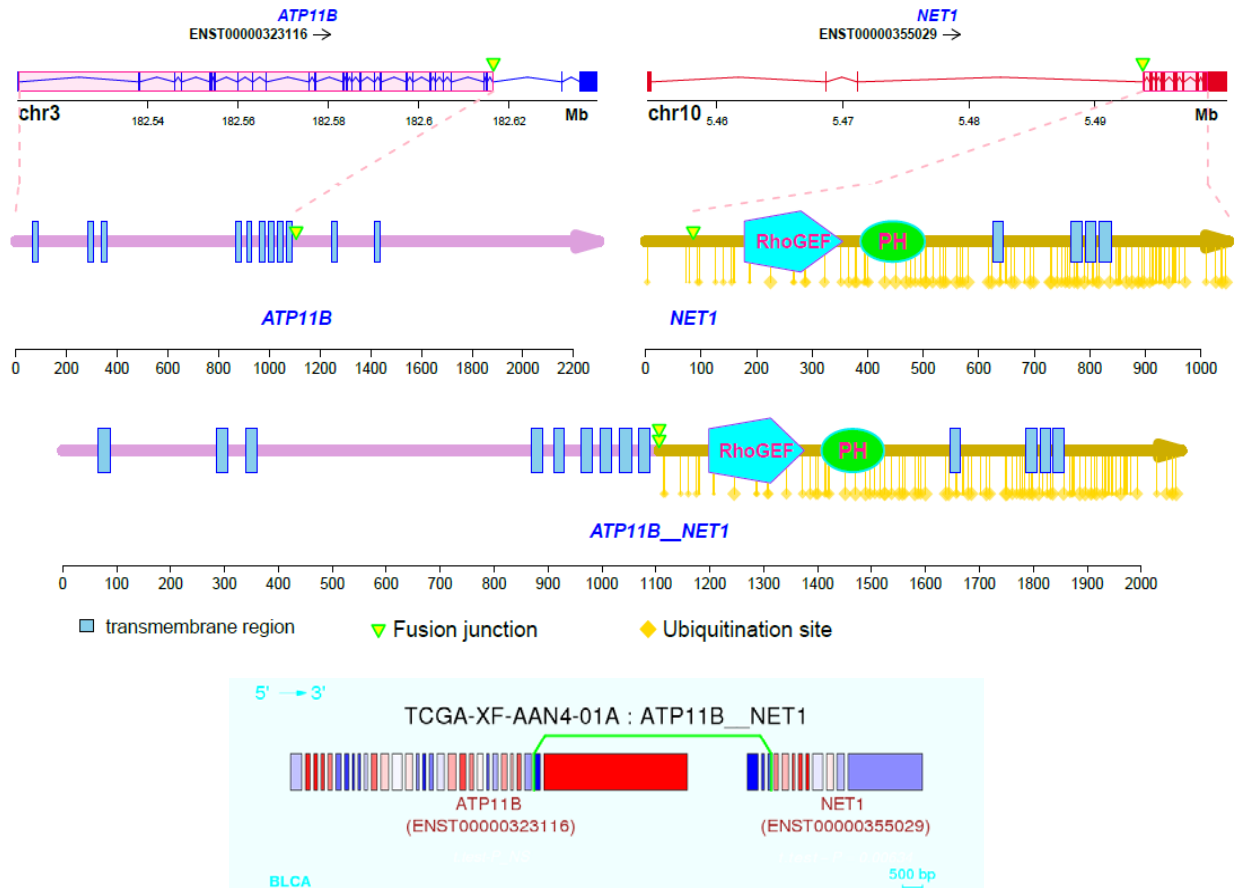


Figure 4.25 Fusion-induced ubiquitination binding site losses in oncogene

The top panel showed both transcriptomic (upper) and protein (bottom) structures of parent genes and the fusion gene *ATP11B-*NET1** are shown. *NET1* showed loss of four ubiquitination sites at its N-terminal upon fusion with *ATP11B* in BLCA. The bottom panel showed the exon expression of *ATP11B-*NET1** in BLCA.

Conversely, we found that four fusion transcripts formed by tumor suppressor genes in sarcoma, SKCM, BRCA, and ESCA gain ubiquitination sites upon fusion events. A typical example of ubiquitination site gain adjacent to a tumor suppressor gene is *HKR1-*CEACAM7**, which may trigger the amalgamation of a ubiquitinated segment from *HKR1* with a short portion of the *CEACAM7* tumor suppressor domain, leading to fusion-mediated loss of immunoglobulin-like C2-type domain function in *CEACAM7*, which mediates numerous cellular functions, including proliferation, differentiation, tumor suppression, immunity, and

infection through self-multimerization or multimerization with other family members (Bonsor, Beckett et al. 2015, Chang, Huang et al. 2016) (**Figure 4.26**).

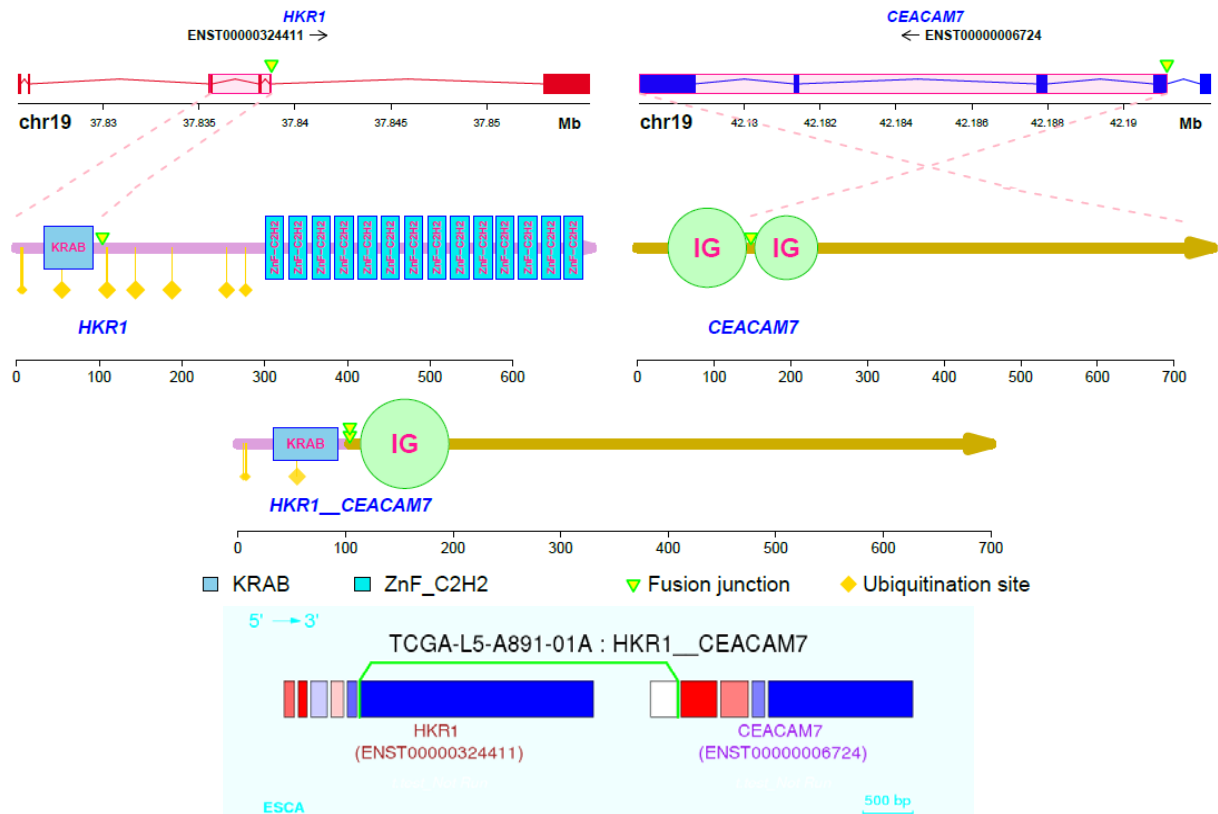


Figure 4.26 Fusion-induced ubiquitination binding sites gained in tumor suppressor genes from fusion *HKR1-CEACAM7* in esophageal carcinoma (ESCA)

The top panel showed both transcriptomic (upper) and protein (bottom) structures of parent genes and the fusion gene *HKR1-CEACAM7* are shown. *CEACAM7* gained four ubiquitination sites at its N-terminal upon fusion with *HKR1* in ESCA. The bottom panel showed the exon expression of *HKR1-CEACAM7* in ESCA.

4.4 Discussions

Our study has prospectively provided an integrated gene fusion database with 20,905 fusion transcripts detected from 9,964 paired-end RNA sequencing tumor samples. We comprehensively characterized the genomic features of fusion transcripts across 33 tumor types, and we found that certain fusion transcripts were associated with somatic mutations, complex genomic rearrangements, and copy number variation.

Notably, we observed significant mutual exclusion between fusion transcripts and somatic mutations per sample in most tumor types. Although the detailed mechanisms remain to be explored, this mutual exclusion suggests a compensatory process of genetic alterations acquired during tumorigenesis. The vast majority of fusion events were accompanied by increased expression in the posterior exons flanking the breakpoint detected in fusion transcripts, suggesting that fusion events affect transcriptional level. Importantly, there was substantial overlap between fusion transcripts and the breakpoints of complex genomic rearrangements, suggesting that these fusion events are caused by chromosomal rearrangement in a complicated fashion.

Since genomic instability can produce passenger fusions and since fusion preferentially affects highly central, widely varying interaction-prone elements, the observed density of interaction-mediating features in parent proteins is in accordance with their centrality in interaction networks (Wu, Kannan et al. 2013). Many parent genes serve as essential genes, which dovetails with the concept of “edge” manner perturbations in cancer, (i.e., in-frame mutations or causal non-synonymous single-nucleotide polymorphisms disrupt specific interactions or edges of proteins rather than the entire node), and fusion events may perturb the functionality of or rewire specific interactions of only a portion of essential proteins. Therefore we implemented centrality scores to predict network disruptions that may be involved in oncogenesis triggered by fusion events. Expectedly, the centrality scores for predicted driver fusions aligned with a high proportion of fusion transcripts harboring oncogenes in THCA, LAML, and CHOL, suggesting the fidelity of centrality scores predicting most novel fusions that exhibited validated oncogenic relevance. Using cut-off values based on centrality scores to select predicted driver fusions from the passenger events, we not only observed the most prevalent fusion transcripts but also revealed new cancer types harboring known fusions, such as *RUNX1-RUNX1T1*, *PML-RARA* and *MLL-MLLT10* in LAML; *FGFR2-BICC1* in CHOL; *FGFR3-TACC3* in BLCA and GBM; *CCDC6-RET* in THCA;

and *PTPRK-RSPO3* in READ and COAD. These fusions were also detected in other types of cancers with low frequency and supported by previous reports (Brandimarte, Pierini et al. 2013, Suzuki, Makinoshima et al. 2013, Williams, Hurst et al. 2013, Storm, Durinck et al. 2016) (Arai, Totoki et al. 2014). The novel fusion *UBTF-MAML3* frequently occurred in pheochromocytoma and paraganglioma; the nucleolar transcription factor *UBTF* fused with *MAML3*, the transcriptional coactivator, might confer altered transcript activity and could be used as biomarker for patient classification.

Notably, our explorations uncovered a set of a set of fusion events that cause tumorigenesis through constitutive or modulated activation of kinase proteins. One such protein is the receptor tyrosine kinase *PDGFRB*, which serves as the receptor for platelet-derived growth factors and plays important roles in wound healing, atherosclerosis, fibrosis, and regulation of tissue oncotic pressure. RHOD domain in *USP8*, which is involved in endosome dynamics and coordinates membrane transport with reorganization of actin cytoskeleton and focal adhesion dissolution, is also involved in internalization and trafficking of activated tyrosine kinase receptor *PDGFRB* (Moroncini, Paolini et al. 2010, Nehru, Voytyuk et al. 2013). When fused to the membrane-spanning region of *PDGFRB* through the hinge provided by coiled-coil domain, RHOD could lead to dimerization of *PDGFRB* and activation of its kinase activity. It is plausible that the constitutive signaling from *PDGFRB-USP8* drives cell proliferation. Since *PDGFRB* could be inhibited by imatinib, *PDGFRB-USP8* represents a potential therapeutic target. We also found that protein kinase C beta (*PRKCB*) fused with several partner genes in different cancer types, including *ABCC1* in PRAD, *ADCY9* in LUSC, *SPNS1* in LUAD, *GGA2* in LGG, and *TNRC6A* in GBM. Depletion of the N-terminal inhibitory domain of *PRKCB* could cause constitutive activation of *PRKCB*, which is known to mediate angiogenesis, immunity, fibroblast function, and adipogenesis (Wallace, Pitarresi et al. 2014). Therefore, targeted therapy using enzastaurin against *PRKCB* in patients harboring *PRKCB* fusions could be of benefit as precision treatment.

Importantly, we discovered a dozen fusion transcripts involving loss of epigenetic reader domains. For instance, *MANF-SETD2* was identified in breast cancer, with the SET domain completely depleted upon fusion events, and with loss of *SETD2*, a histone-lysine N-methyltransferase that plays an important role in tumor progression and chemotherapy resistance through attenuation of DNA damage response in multiple cancers (Huang, McPherson et al. 2015, Kanu, Gronroos et al. 2015, Parker, Rose-Zerilli et al. 2016). As another example, *ARID1B*, a subunit of chromatin remodeling complex SWI/SNF, which function as tumor suppressor in human cancer (Aso, Uozaki et al. 2015), was identified with loss of ARID domains upon fusion to different gene partners. Therefore, both *SETD2* and *ARID1B* fusions could serve as attractive biomarkers to predict clinical outcomes and stratify patients based on the presence of the involved fusions.

In addition to the impact of fusion events on parental protein structure, post-translational and co-translational modification sites on retained sequences of fusion proteins could regulate protein stability (e.g., by ubiquitination) or subcellular localization (e.g., N-myristoylation) as well. We identified loss of ubiquitination sites at the N-terminal of oncogene NET1, a RhoA guanine nucleotide exchange factor that contributes to cancer cell motility and invasion (Carr, Zuo et al. 2013). The loss of ubiquitination sites confer the potential to promote NET1 activity by increasing its stability, as ubiquitin acts as a versatile cellular signal that governs various of biologic processes including transcription, DNA repair, endocytosis, protein degradation, autophagy, immunity and inflammation (Husnjak and Dikic 2012). Fusion events altering ubiquitin binding sites adjacent to functional domains of parental proteins could trigger fundamental effects on their oncogenic properties.

To our knowledge, the present study has produced the first repository of fusion genes identified in all major cancer types by unified criteria from RNA-Seq datasets (<http://www.tumorfusions.org>). Our study reflected the distinct genomic features of fusion composition in different cancer types and unraveled a series of common fusion hub genes

and fusion transcripts associated with potential tumorigenesis shared in multiple cancer types, including some less well-characterized fusion transcripts. Our findings not only build the foundation for clinical utility of various onco-fusion proteins but also facilitate thorough dive into the integrative molecular plateau driving cancers. Collectively, our findings could be instrumental to developing new prognostic and therapeutic strategies targeting onco-fusion proteins that have escaped from normal regulatory pathways in various tumor types.

CHAPTER 5

Conclusions and future perspectives

5.1 Summary

In my dissertation, I have explored omics data and their clinical relevance in order to pave the way to precision medicine. Toward this goal, we performed studies through three aspects: 1) We characterized how transcriptomic complexity was correlated with increased intra-tumoral heterogeneity and the contents of the tumor microenvironment, decoded cellular components contributing to the classification of GBM, and elucidated the clinical relevance towards therapeutic intervention in each subtype of the patients harboring distinct genomic patterns of both intrinsic tumor and infiltrated tumor microenvironments; 2) We identified gene signatures associated with inter-tumor heterogeneity and used this potential to stratify patients based on prognostic index in LGG; 3) We characterized fusion transcripts and their associated genomic features across different cancer types and identified some novel fusion events with oncogenic potential as diagnostic tools and/or therapeutic targets.

5.2 Significance, pitfalls and perspective explorations

5.2.1 Classification of glioma integrating transcriptomic profiling from both intrinsic tumor and infiltrated microenvironment

Numerous studies have been carried out to elucidate the mechanism for radiation resistance of GBM. Some studies reported glioma cells undergoing Epithelial-Mesenchymal Transition (EMT) involving in GBM recurrence (Kubelt, Hattermann et al. 2015), while other studies reported only a subset of proneural GBM patient-derived glioma sphere cultures (GSCs) differentiated to MES state mediated by $\text{TNF-}\alpha/\text{NF-}\kappa\text{B}$, accompanied with CD44 subpopulations showing enrichment for radiation resistant phenotypes (Bhat, Balasubramanian et al. 2013). These controversial findings led to our hypothesis that GBM

subtypes may change during tumor evolution. Is there an intrinsic transition from proneural to mesenchymal upon GBM recurrence, or do proneural & mesenchymal GSC co-exist within individual tumors and radiation-resistant Mes-like cells preferentially survive in recurrent cases and emerge as dominant populations? On the other hand, pathological examination revealed the neurosis exhibit higher content in mesenchymal subtype, yet no significant difference of overall survival between mesenchymal subtype and non-mesenchymal subtypes, indicating other cellular components intertwined with mesenchymal signaling and phenotypic subtype, and differential expression of signature genes might reflect the molecular fluctuation signaling from both tumor and associated stromal or immune components.

To test this hypothesis, I dissected GBM intrinsic transcription phenotypes and their association with different cellular components of tumor immune environment, using RNA sequencing data derived from glioma patients and their derived glioma sphere cultures (GSC). I performed computational modeling on RNA sequencing data and integrated somatic mutation data to deconvolute the complex nature of stromal-tumor microenvironments and their association with established glioma subtypes. I also performed comparisons of molecular subtypes between matching primary and recurrent gliomas, and elucidated treatment-induced phenotypic tumor evolution. I found proneural to mesenchymal transitions occur only in a subset of GBM patients, while intra-tumoral heterogeneity acts as a predominant factor associated with subtype transition upon recurrence. Moreover decoding the components that infiltrate tumor microenvironment evolving different glioblastoma transcriptomic subtypes provided the rationale for more effective immunotherapy trials.

This study provided a more precise strategy to determine transcriptional subtype of GBM, revealed the tight association between intrinsic tumor transcriptome and dynamics of tumor-associated immuno-environment, and unraveled the essential contributions of

microenvironment to clinical outcomes. The alterations in cellular components of the immune system during GBM progression, such as reduced invading monocytes but a subtype dependent increase in M2 macrophages/microglia cells upon disease recurrence, implied that activated M2-macrophages may be potential targets for immunotherapy. Such therapies might include inhibition of macrophage recruitment, suppression of TAM survival, and blockage of M2-like tumor-promoting activity. Our work also sheds new light on how immune diversity in each patient contributes to inter-tumoral heterogeneity, which could explain the different responses to the same treatment regimen, indicating that therapeutic regimens should be determined through monitoring the state of both angiogenic and immune parameters in patients at different disease stages. In addition, this study has disputed the conventional wisdom that epithelial-to-mesenchymal transition (EMT) inducing signals drive glioblastoma progression in all subtypes and revealed that signaling from molecular classification is derived from complex mixed cellular traits rather than a single homogenous cell origin. Taken together this study has built significant foundation towards precision medicine through comprehensive characterization of transcriptional and cellular landscape of IDH wild type GBM during tumor evolution modulated by different treatments.

Regarding future perspectives, given the significant contribution of tumor microenvironment to glioma-genesis and intra/inter tumor heterogeneity which is associated with clinical outcomes, I propose to develop novel immune/ stromal gene signatures to predict treatment response in GBM. In light of immune classification of GBM subtypes, I hypothesize that tumor subtypes are determined by two axes, inflammation and adaptive immunity, as shown in **Figure 5.1**. The tumor immunological environment contributes to clinical outcome along these two major axes, which could be applied to stratify tumor molecular subgroups. The adaptive anti-tumor immune response is associated with favorable outcomes, whereas pro-tumor inflammation leads to poor outcome. Therefore the

development of multiple markers regulated in different axes simultaneously is necessary to interpret the immune contexture of GBM and accurately predict prognosis and response to radio/chemo therapies. Characterization of interactions and crosstalk pathways between the immune system and bulk tumor will also facilitate targeting immunotherapeutic treatments. Developing computational modeling of single cell sequencing on specifically sorted cell types (based on cell surface markers) to discover signaling networks between cancer cells, stromal and immune cells involved in disease evolution and developing machine-learning and network analysis through single-cell transcriptomics profiling to decipher clonal evolution and mechanisms mediating microglia/macrophage polarization to the immunosuppressive state in GBM.

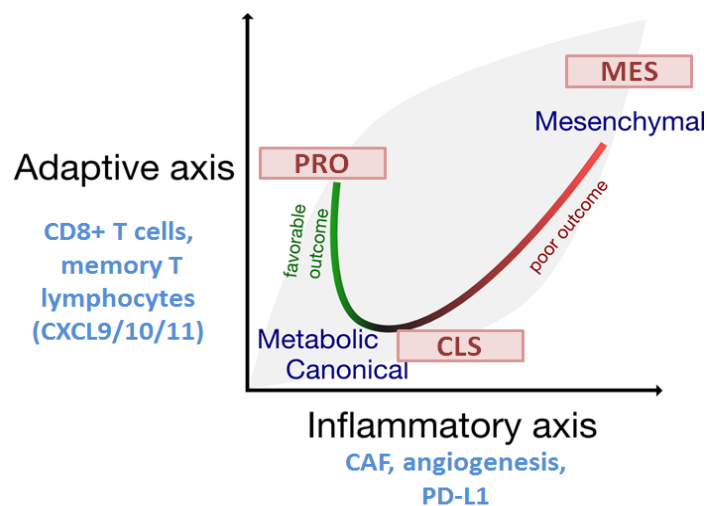


Figure 5.1 Clinically relevant transcriptomic subtypes determined by both inflammatory and adaptive immune components

Transcriptomic predicted subgroups that correspond to GBM tumor based subtypes (proneural, classical, mesenchymal) are labeled in red rectangles. The smooth regression line and gray polygon represent the hypothetical trend of polarization of immune cells in tumor microenvironment.

Taking advantage of single-cell technology, we could seek insights into unique subsets of immune cells that predict clinical outcomes, and sort out rare cell populations (i.e.

antigen-specific T cells) that confer protection or trigger the pathology of disease. We would then build system-level models to decipher the spatiotemporal and functional dynamics of intercellular interactions at single-cell resolution, so that we could evaluate the responses to antiangiogenic therapy in combination with immunotherapy. Based on the hypothesis that multipotent adipose-derived stromal cells (ASCs) could alter the tumor microenvironment in ways that facilitate the transition to more malignant status, I propose to qualify the ratio of stromal fibroblast to adipocytes, in order to predict the patients' outcome based on their immune / stromal cell composition. As another arm of the project, we could quantify the intra-tumor heterogeneity based on single cell profiling of functionally distinct cell types in tumor associated microenvironment infiltration (TAM).

5.2.2 Identify gene signatures associated with clinical outcomes in low grade glioma

Given 1p/19q codeletion cohort only account for a small proportion of all intracranial gliomas in adults and those patients usually exhibit prolonged survival with censored data (**Figure 5.2**), I have searched all over the public datasets to retrieve seven individual datasets of glioma with 1p/19q codeletion where both survival and gene expression data are available. Since the distributions of overall survival and death rates vary largely from one dataset to another, and are further confounded with measurement using different platforms including microarray and RNA-seq, different treatment options from different eras, combinations of two or three independent datasets failed to produce sufficient training data for penalty based feature selection or validation by standard Cox models.

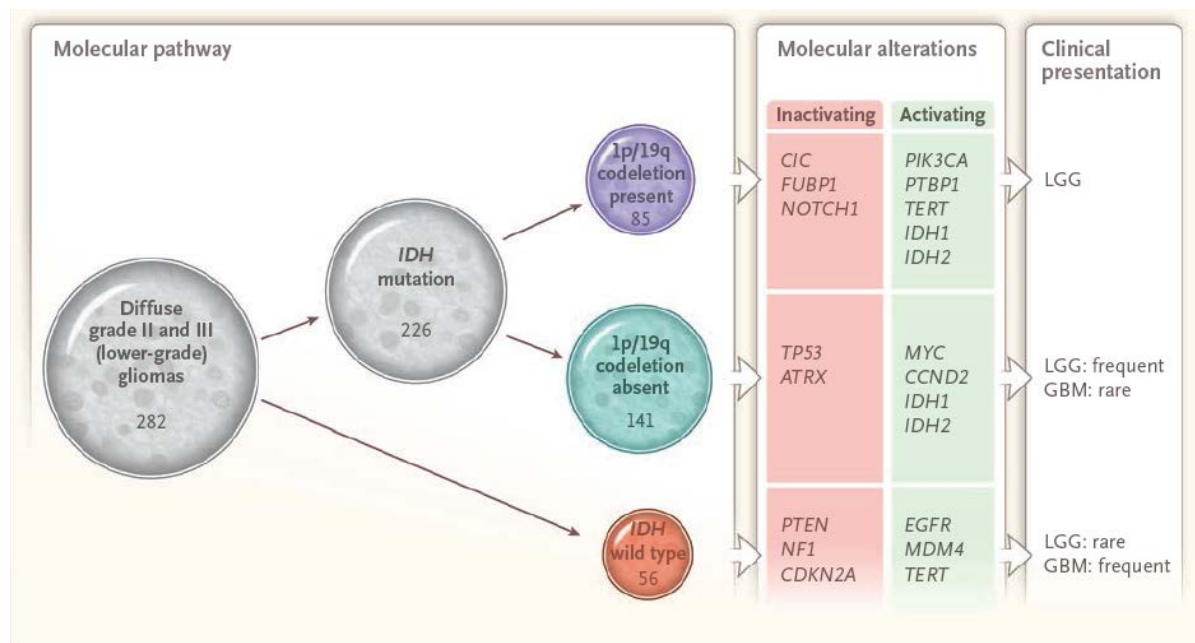


Figure 5.2 Subtype-specific molecular alterations and distinct clinical presentations in Lower-Grade Gliomas (LGG)

The numbers of patients in each molecular subtype are shown. The molecular alterations and associated disease grade are shown on the right.

(Cited from Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas

The New England Journal of Medicine June 25, 2015 vol. 372 no. 26)

In order to accommodate the heterogeneous nature of the datasets and the limited number of events in some datasets, I have combined five major datasets through scaled normalization to reduce the batch effects, then I split the combined dataset evenly through random sampling so that the distributions of survival, treatment options, grade, age are similar for training and validation datasets respectively. First I used the training dataset (n=164) to perform feature selection applying the Elastic Net (3 fold cross-validation using glmnet), a regularized Cox model that is a robust hybrid of lasso and ridge regressions), then I made predictions for the validation dataset (n=170). The resulting C-index derived from gene signatures is similar to that from age and grade, with the hazard ratio (HR) for each gene predictor ranging between 0.5-1.5. I have assessed the linear correlation of each predictor and overall survival and found the effect of alteration of each gene expression on

overall survival is little, (**Figure 5.3**) indicating the individual predictors are weak, while the C-index is generated by the cumulative power of the entire gene signature (n=35).

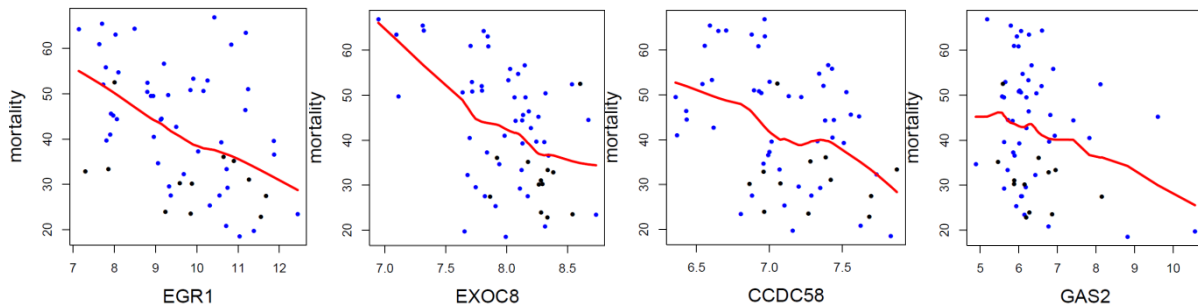


Figure 5.3 Partial plots for gene predictors from codel glioma

The values on the y-axis represent the expected number of deaths for a given predictor, after adjusting for all other predictors. The values on the x-axis represent the relative gene expression values of each predictor. The red solid lines represent LOESS curves fitting gene expression versus death rate.

In an attempt to develop an optimized gene signature, I have also explored different methods for feature selection through modeling the association of gene expression with overall survival, including L1 based feature selection (LASSO), Support Vector Machine (SVM), ensemble method (Survival Ensembles), Random Survival Forest (randomForestSRC) where ensemble estimates are derived for the cumulative hazard function, and Model-Based Boosting methods (Likelihood-based boosting for Cox models: CoxBoost). In ensemble algorithms, bagging process construct several instances of black-box estimators based on random sampling from training set and then aggregates each individual prediction into a final prediction. The goal is to make an ensemble by introducing randomization into its itinerary construction process and reduce the variance of decision tree as base estimator. Amongst these exploratory methods, LASSO outperforms the other methods through simultaneous estimation and variable selection, boosting exhibits intermediate performance, whereas Ensemble models generate poor predictions. The plausible explanation in terms of overfitting reduction is, bagging methods perform the best

on complex but strong models (fully developed decision trees), whereas boosting methods usually fit better with weak models. Of note, there is high overlap of top significant genes selected by the different methods, suggesting the selected gene signature is robust to certain extent. In addition, I combined two independent datasets to validate the predictive power of the gene signature. Though not significant due to extremely small sample sizes, there is a distinct trend between high and low risk groups based on their prognostic index.

Interestingly my investigation revealed that some of these signature genes are associated with inflammatory response revealed by corresponding GO terms mapped to the signature genes. Moreover, I calculated the estimate score for each sample to infer their content of infiltrated immune signal and stroma components and found estimated scores in the high risk group are significantly higher than those in the low risk group. The risk scores deduced for each patient are positively correlated with their estimate scores, plus some selected signature genes are overlap with immune and stroma signature applied in ESTIMATE, suggesting inflammation and immune signal might contribute to increased risk in these patients. This is the first study tackling the source of heterogeneity in codel glioma patients.

In terms of future directions, I will apply backward selection to reduce the number of signature genes to further reduce overfitting and the complexity of the model. I will fit a Cox model with current gene signatures, then dropped the least significant genes at some critical level, repeating the process by successively re-fitting reduced models and applying the same critical level until all remaining signature genes are statistically significant. I will also apply the adaptive lasso to fine-tune the gene signature selection process, where adaptive weights are used for penalizing different coefficients in the L1 penalty. Finally, I will look for newly generated datasets for independent validations of the potential predictions for application in clinical settings in the future.

5.2.3 Characterize distinct genomic patterns of fusion transcripts and identify novel fusion events conferring oncogenic potential in pan-cancers

By investigating fusion transcripts of 33 different tumor types using TCGA RNAseq data, I have observed a large variance of fusion occurrence rates across fusion genes with distinct genomic features and cancer types. By integrating with multi-dimensional TCGA data, including WGS, WES and SNP array data, we have annotated a subset of fusion transcripts supported by corresponding structure variants in a genomically coordinated manner. My study revealed common and diverse fusion spectrums with hotspot fusion events across different tumor types. We found a peak enrichment of fusion transcripts at 12q in sarcomas, which is coincident with structure variants and copy number alterations. The majority of fusion transcripts are associated with translocation at their adjacent regions followed by deletion, duplication and inversion.

I have also explored the functional predictions on the fusion transcripts involving kinase activity, chromatin remodeling and post-translational modifications, demonstrating the potential functional relevance of transcripts in diverse biological and malignant contexts and nominating a set of novel fusion transcripts such as *PDGFRB-USP8*, *MANF-SETD2*, *GGA2-PRKCB*, etc with predicted functional associations for further study. This study will improve our understanding of fusion transcripts and their correlated genomic features, the oncogenic potential and perturbed pathways triggered by fusion transcripts, which will facilitate translational efforts in therapeutic target discovery and diagnostic tool development. The fusion database and annotation generated from this analysis are uploaded for public access. TCGA Fusion Gene Data Portal (www.tumorfusions.org).

Since there is a distinct correlation between recurrent gene fusions and specific cancer types, we could develop the fusion gene detection for common tumor subtype screening, which will provide a roadmap for precision therapies. Some of the novel fusion transcripts identified in this study act as important drivers of malignancy, so they could serve

as potentially diagnostic markers in the clinical setting. These include recurrent fusions in *PTPRK-RSPO3*, *TFG-GPR128* in READ, *UBTF-MAML3* in PCPG, etc. In addition, this study provided integrative annotation of gene fusions associated with both balanced and unbalanced chromosomal rearrangements, as well as their association with copy number alterations, which lay the foundation for unraveling the complex nature of genomic structural aberrations. Collectively the fusion database detected from the world's largest sample size and tumor types in uniform platform revealed fusion transcripts with clinical potential that have not been reported as yet.

Regarding future directions, I will further explore the genomic features of hotspot fusion transcripts and elucidate the plausible elements that make those loci prone to fusion formation. I will also identify the fusion transcripts that trigger transcriptional alterations of their target genes by trans/cis-regulation. I will scan the fusion transcripts with gain or loss of sumoylation loci that are analogous to ubiquitination. I will decode the fusion genes that mediate rewiring of protein-protein interaction (PPI) networks in different cancer types, to address whether and how fusion transcripts disrupt the interactive networks, alter the regulatory sites and rewire the signaling that associated with tumor genesis. Collectively I will generate a comprehensive landscape revealing the molecular principles and heterogeneous patterns of fusion events across different types of cancers and their clinical association, with a subset of fusion transcripts as the classifier to stratify the patients.

Insight into molecular heterogeneity, environmental risk factors, tumor plasticity and crosstalk between tumor and their associated microenvironments have laid the foundation towards potential novel strategies for early detection and efficacy therapy, while further basic and translational investigation to address heterogeneous nature and dynamics of cancer evolution are required. This will lead my study to dive into a promising continuation of efforts from the current thesis and demonstrate the contribution of my Ph.D. work to a series of biological and therapeutic perspectives towards precision medicine.

Supplementary Tables

Table S4.1 Number of TCGA samples in 33 cancer types

Cancer	Tumor	Normal
ACC	79	0
BLCA	414	19
BRCA	1119	113
CESC	306	3
CHOL	36	9
COAD	309	0
DLBC	48	0
ESCA	185	13
GBM	170	5
HNSC	522	44
KICH	66	25
KIRC	541	72
KIRP	291	32
LAML	179	0
LGG	534	0
LIHC	374	50
LUAD	541	59
LUSC	502	51
MESO	87	0
OV	428	0
PAAD	179	4
PCPG	184	3
PRAD	502	52
READ	95	10
SARC	263	2
SKCM	472	1
STAD	414	37
TGCT	156	0
THCA	513	59
THYM	120	2
UCEC	185	24
UCS	57	0
UVM	80	0
sum	9951	689

Table S4.2 Number of fusions filtered by each step in 33 cancer types

Cancer	Raw	E.value > 0.001	PGV > 10	TAF > 0.01	No detection in normal
ACC	448	394	335	247	240
BLCA	2386	1971	1930	1159	1122
BRCA	8460	6306	5413	4170	4060
CESC	1242	900	790	445	426
CHOL	103	83	83	64	61
COAD	809	472	461	340	307
DLBC	107	61	61	45	44
ESCA	691	439	439	374	367
GBM	684	360	360	295	289
HNSC	5156	3959	1520	770	730
KICH	190	137	125	57	47
KIRC	1045	602	447	291	271
KIRP	715	528	371	206	185
LAML	279	174	174	127	112
LGG	2277	1909	1367	851	785
LIHC	2764	2154	1257	761	753
LUAD	3551	2712	1962	1481	1451
LUSC	3584	2531	2008	1426	1356
MESO	445	208	197	141	136
OV	790	675	675	419	416
PAAD	989	482	378	201	192
PCPG	532	473	225	125	119
PRAD	4096	3361	2026	1441	1412
READ	330	207	207	148	132
SARC	3946	2906	2372	1764	1751
SKCM	3844	3273	1869	1195	1170
STAD	1313	960	946	784	768
TGCT	311	155	144	101	82
THCA	3094	2726	481	206	143
THYM	169	96	96	58	55
UCEC	1191	1004	981	787	780
UCS	579	497	478	389	385
UVM	78	57	57	37	33

References

- Arai, Y., Y. Totoki, F. Hosoda, T. Shiota, N. Hama, H. Nakamura, H. Ojima, K. Furuta, K. Shimada, T. Okusaka, T. Kosuge and T. Shibata (2014). "Fibroblast growth factor receptor 2 tyrosine kinase fusions define a unique molecular subtype of cholangiocarcinoma." *Hepatology* **59**(4): 1427-1434.
- Aso, T., H. Uozaki, S. Morita, A. Kumagai and M. Watanabe (2015). "Loss of ARID1A, ARID1B, and ARID2 Expression During Progression of Gastric Cancer." *Anticancer Res* **35**(12): 6819-6827.
- Bennett, G., D. Sadlier, P. P. Doran, P. Macmathuna and D. W. Murray (2011). "A functional and transcriptomic analysis of NET1 bioactivity in gastric cancer." *BMC Cancer* **11**: 50.
- Bonsor, D. A., D. Beckett and E. J. Sundberg (2015). "Structure of the N-terminal dimerization domain of CEACAM7." *Acta Crystallogr F Struct Biol Commun* **71**(Pt 9): 1169-1175.
- Bos, M., M. Gardizi, H. U. Schildhaus, R. Buettner and J. Wolf (2013). "Activated RET and ROS: two new driver mutations in lung adenocarcinoma." *Transl Lung Cancer Res* **2**(2): 112-121.
- Brandimarte, L., V. Pierini, D. Di Giacomo, C. Borga, F. Nozza, P. Gorello, M. Giordan, G. Cazzaniga, G. Te Kronnie, R. La Starza and C. Mecucci (2013). "New MLLT10 gene recombinations in pediatric T-acute lymphoblastic leukemia." *Blood* **121**(25): 5064-5067.
- Capelletti, M., M. E. Dodge, D. Ercan, P. S. Hammerman, S. I. Park, J. Kim, H. Sasaki, D. M. Jablons, D. Lipson, L. Young, P. J. Stephens, V. A. Miller, N. I. Lindeman, K. J. Munir, W. G. Richards and P. A. Janne (2014). "Identification of recurrent FGFR3-TACC3 fusion oncogenes from lung adenocarcinoma." *Clin Cancer Res* **20**(24): 6551-6558.
- Carr, H. S., Y. Zuo, W. Oh and J. A. Frost (2013). "Regulation of focal adhesion kinase activation, breast cancer cell motility, and amoeboid invasion by the RhoA guanine nucleotide exchange factor Net1." *Mol Cell Biol* **33**(14): 2773-2786.
- Chang, J., L. Huang, Q. Cao and F. Liu (2016). "Identification of colorectal cancer-restricted microRNAs and their target genes based on high-throughput sequencing data." *Onco Targets Ther* **9**: 1787-1794.
- Chen, T., T. Zhou, B. He, H. Yu, X. Guo, X. Song and J. Sha (2014). "mUbiSiDa: a comprehensive database for protein ubiquitination sites in mammals." *PLoS One* **9**(1): e85744.
- Chen, Z., Y. Zhou, Z. Zhang and J. Song (2015). "Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features." *Brief Bioinform* **16**(4): 640-657.
- Chiang, C., R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose, E. P. Garrison, G. T. Marth, A. R. Quinlan and I. M. Hall (2015). "SpeedSeq: ultra-fast personal genome analysis and interpretation." *Nat Methods* **12**(10): 966-968.

Eguchi, F. C., E. F. Faria, C. Scapulatempo Neto, A. Longatto-Filho, C. Zanardo-Oliveira, S. R. Taboga and S. G. Campos (2014). "The role of TMPRSS2:ERG in molecular stratification of PCa and its association with tumor aggressiveness: a study in Brazilian patients." Sci Rep **4**: 5640.

Flynn, A., D. Benn, R. Clifton-Bligh, B. Robinson, A. H. Trainer, P. James, A. Hogg, K. Waldeck, J. George, J. Li, S. B. Fox, A. J. Gill, G. McArthur, R. J. Hicks and R. W. Tothill (2015). "The genomic landscape of pheochromocytoma." J Pathol **236**(1): 78-89.

Guarnerio, J., M. Bezzi, J. C. Jeong, S. V. Paffenholz, K. Berry, M. M. Naldini, F. Lo-Coco, Y. Tay, A. H. Beck and P. P. Pandolfi (2016). "Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations." Cell **165**(2): 289-302.

Guo, J., L. Canaff, C. V. Rajadurai, N. Fils-Aime, J. Tian, M. Dai, J. Korah, M. Villatoro, M. Park, S. Ali and J. J. Lebrun (2014). "Breast cancer anti-estrogen resistance 3 inhibits transforming growth factor beta/Smad signaling and associates with favorable breast cancer disease outcomes." Breast Cancer Res **16**(6): 476.

Hermans, K. G., R. van Marion, H. van Dekken, G. Jenster, W. M. van Weerden and J. Trapman (2006). "TMPRSS2:ERG fusion by translocation or interstitial deletion is highly relevant in androgen-dependent prostate cancer, but is bypassed in late-stage androgen receptor-negative prostate cancer." Cancer Res **66**(22): 10658-10663.

Huang, K. K., J. R. McPherson, S. T. Tay, K. Das, I. B. Tan, C. C. Ng, N. Y. Chia, S. L. Zhang, S. S. Myint, L. Hu, V. Rajasegaran, D. Huang, J. L. Loh, A. Gan, A. N. Sairi, X. X. Sam, L. T. Dominguez, M. Lee, K. C. Soo, L. L. Ooi, H. S. Ong, A. Chung, P. K. Chow, W. K. Wong, S. Selvarajan, C. K. Ong, K. H. Lim, T. Nandi, S. Rozen, B. T. Teh, R. Quek and P. Tan (2015). "SETD2 histone modifier loss in aggressive GI stromal tumours." Gut.

Huang, Q., V. E. Schneeberger, N. Luetkeke, C. Jin, R. Afzal, M. M. Budzevich, R. J. Mankanji, G. V. Martinez, T. Shen, L. Zhao, K. M. Fung, E. B. Haura, D. Coppola and J. Wu (2016). "Preclinical Modeling of KIF5B-RET Fusion Lung Adenocarcinoma." Mol Cancer Ther **15**(10): 2521-2529.

Husnjak, K. and I. Dikic (2012). "Ubiquitin-binding proteins: decoders of ubiquitin-mediated cellular functions." Annu Rev Biochem **81**: 291-322.

Jividen, K. and H. Li (2014). "Chimeric RNAs generated by intergenic splicing in normal and cancer cells." Genes Chromosomes Cancer **53**(12): 963-971.

Kanu, N., E. Gronroos, P. Martinez, R. A. Burrell, X. Yi Goh, J. Bartkova, A. Maya-Mendoza, M. Mistrik, A. J. Rowan, H. Patel, A. Rabinowitz, P. East, G. Wilson, C. R. Santos, N. McGranahan, S. Gulati, M. Gerlinger, N. J. Birkbak, T. Joshi, L. B. Alexandrov, M. R. Stratton, T. Powles, N. Matthews, P. A. Bates, A. Stewart, Z. Szallasi, J. Larkin, J. Bartek and C. Swanton (2015). "SETD2 loss-of-function promotes renal cancer branched evolution through replication stress and impaired DNA repair." Oncogene **34**(46): 5699-5708.

Kirkin, V. and I. Dikic (2011). "Ubiquitin networks in cancer." Curr Opin Genet Dev **21**(1): 21-28.

Leslie, R., C. J. O'Donnell and A. D. Johnson (2014). "GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database." Bioinformatics **30**(12): i185-194.

Li, J., G. Duns, H. Westers, R. Sijmons, A. van den Berg and K. Kok (2016). "SETD2: an epigenetic modifier with tumor suppressor functionality." Oncotarget.

Luo, L. Y., E. Kim, H. W. Cheung, B. A. Weir, G. P. Dunn, R. R. Shen and W. C. Hahn (2015). "The Tyrosine Kinase Adaptor Protein FRS2 Is Oncogenic and Amplified in High-Grade Serous Ovarian Cancer." Mol Cancer Res **13**(3): 502-509.

Mani, A. and E. P. Gelmann (2005). "The ubiquitin-proteasome pathway and its role in cancer." J Clin Oncol **23**(21): 4776-4789.

Medves, S. and J. B. Demoulin (2012). "Tyrosine kinase gene fusions in cancer: translating mechanisms into targeted therapies." J Cell Mol Med **16**(2): 237-248.

Meyer, C., A. Brieger, G. Plotz, N. Weber, S. Passmann, T. Dingermann, S. Zeuzem, J. Trojan and R. Marschalek (2009). "An interstitial deletion at 3p21.3 results in the genetic fusion of MLH1 and ITGA9 in a Lynch syndrome family." Clin Cancer Res **15**(3): 762-769.

Moroncini, G., C. Paolini, A. Grieco, G. Nacci, M. Cuccioloni, M. Mozzicafreddo, C. Tonnini, S. Svegliati, M. Angeletti, E. Avvedimento, A. Funaro and A. Gabrielli (2010). "PDGF receptor as therapeutic and diagnostic target in systemic sclerosis." Clinical and Experimental Rheumatology **28**(5): S71-S72.

Muntean, A. G. and J. L. Hess (2009). "Epigenetic dysregulation in cancer." Am J Pathol **175**(4): 1353-1361.

Nacu, S., W. Yuan, Z. Kan, D. Bhatt, C. S. Rivers, J. Stinson, B. A. Peters, Z. Modrusan, K. Jung, S. Seshagiri and T. D. Wu (2011). "Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples." BMC Med Genomics **4**: 11.

Nair, S. S. and R. Kumar (2012). "Chromatin remodeling in cancer: a gateway to regulate gene transcription." Mol Oncol **6**(6): 611-619.

Narayan, S., G. D. Bader and J. Reimand (2016). "Frequent mutations in acetylation and ubiquitination sites suggest novel driver mechanisms of cancer." Genome Med **8**(1): 55.

Nehru, V., O. Voytyuk, J. Lennartsson and P. Aspenstrom (2013). "RhoD binds the Rab5 effector Rabankyrin-5 and has a role in trafficking of the platelet-derived growth factor receptor." Traffic **14**(12): 1242-1254.

Parker, B. C. and W. Zhang (2013). "Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment." Chin J Cancer **32**(11): 594-603.

Parker, H., M. J. Rose-Zerilli, M. Larrayoz, R. Clifford, J. Edelmann, S. Blakemore, J. Gibson, J. Wang, V. Ljungstrom, T. K. Wojdacz, T. Chaplin, A. Roghanian, Z. Davis, A. Parker, E. Tausch, S. Ntoufa, S.

Ramos, P. Robbe, R. Alsolami, A. J. Steele, G. Packham, A. E. Rodriguez-Vicente, L. Brown, F. McNicholl, F. Forconi, A. Pettitt, P. Hillmen, M. Dyer, M. S. Cragg, C. Chelala, C. C. Oakes, R. Rosenquist, K. Stamatopoulos, S. Stilgenbauer, S. Knight, A. Schuh, D. G. Oscier and J. C. Strefford (2016). "Genomic disruption of the histone methyltransferase SETD2 in chronic lymphocytic leukaemia." Leukemia.

Qian, Y., S. Chai, Z. Liang, Y. Wang, Y. Zhou, X. Xu, C. Zhang, M. Zhang, J. Si, F. Huang, Z. Huang, W. Hong and K. Wang (2014). "KIF5B-RET fusion kinase promotes cell growth by multilevel activation of STAT3 in lung cancer." Mol Cancer **13**: 176.

Qin, F., Z. Song, M. Babiceanu, Y. Song, L. Facemire, R. Singh, M. Adli and H. Li (2015). "Discovery of CTCF-sensitive Cis-spliced fusion RNAs between adjacent genes in human prostate cells." PLoS Genet **11**(2): e1005001.

Sasaki, T., S. J. Rodig, L. R. Chirieac and P. A. Janne (2010). "The biology and treatment of EML4-ALK non-small cell lung cancer." Eur J Cancer **46**(10): 1773-1780.

Storm, E. E., S. Durinck, F. de Sousa e Melo, J. Tremayne, N. Kljavin, C. Tan, X. Ye, C. Chiu, T. Pham, J. A. Hongo, T. Bainbridge, R. Firestein, E. Blackwood, C. Metcalfe, E. W. Stawiski, R. L. Yauch, Y. Wu and F. J. de Sauvage (2016). "Targeting PTPRK-RSPO3 colon tumours promotes differentiation and loss of stem-cell function." Nature **529**(7584): 97-100.

Suzuki, M., H. Makinoshima, S. Matsumoto, A. Suzuki, S. Mimaki, K. Matsushima, K. Yoh, K. Goto, Y. Suzuki, G. Ishii, A. Ochiai, K. Tsuta, T. Shibata, T. Kohno, H. Esumi and K. Tsuchihara (2013). "Identification of a lung adenocarcinoma cell line with CCDC6-RET fusion gene and the effect of RET inhibitors in vitro and in vivo." Cancer Sci **104**(7): 896-903.

Teles Alves, I., T. Hartjes, E. McClellan, S. Hiltemann, R. Bottcher, N. Dits, M. R. Temanni, B. Janssen, W. van Workum, P. van der Spek, A. Stubbs, A. de Klein, B. Eussen, J. Trapman and G. Jenster (2015). "Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer." Oncogene **34**(5): 568-577.

Torres-Garcia, W., S. Zheng, A. Sivachenko, R. Vegesna, Q. Wang, R. Yao, M. F. Berger, J. N. Weinstein, G. Getz and R. G. Verhaak (2014). "PRADA: pipeline for RNA sequencing data analysis." Bioinformatics **30**(15): 2224-2226.

Varley, K. E., J. Gertz, B. S. Roberts, N. S. Davis, K. M. Bowling, M. K. Kirby, A. S. Nesmith, P. G. Oliver, W. E. Grizzle, A. Forero, D. J. Buchsbaum, A. F. LoBuglio and R. M. Myers (2014). "Recurrent read-through fusion transcripts in breast cancer." Breast Cancer Res Treat **146**(2): 287-297.

Velghe, A. I., S. Van Cauwenberghe, A. A. Polyansky, D. Chand, C. P. Montano-Almendras, S. Charni, B. Hallberg, A. Essaghir and J. B. Demoulin (2014). "PDGFRA alterations in cancer: characterization of a gain-of-function V536E transmembrane mutant as well as loss-of-function and passenger mutations." Oncogene **33**(20): 2568-2576.

Velusamy, T., N. Palanisamy, S. Kalyana-Sundaram, A. A. Sahasrabuddhe, C. A. Maher, D. R. Robinson, D. W. Bahler, T. T. Cornell, T. E. Wilson, M. S. Lim, A. M. Chinnaiyan and K. S. Elenitoba-Johnson

(2013). "Recurrent reciprocal RNA chimera involving YPEL5 and PPP1CB in chronic lymphocytic leukemia." Proc Natl Acad Sci U S A **110**(8): 3035-3040.

Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, Jr. and K. W. Kinzler (2013). "Cancer genome landscapes." Science **339**(6127): 1546-1558.

Wallace, J. A., J. R. Pitarresi, N. Sharma, M. Palettas, M. C. Cuitino, S. T. Sizemore, L. Yu, A. Sanderlin, T. J. Rosol, K. D. Mehta, G. M. Sizemore and M. C. Ostrowski (2014). "Protein kinase C Beta in the tumor microenvironment promotes mammary tumorigenesis." Front Oncol **4**: 87.

Watson, I. R., K. Takahashi, P. A. Futreal and L. Chin (2013). "Emerging patterns of somatic mutations in cancer." Nat Rev Genet **14**(10): 703-718.

Williams, S. V., C. D. Hurst and M. A. Knowles (2013). "Oncogenic FGFR3 gene fusions in bladder cancer." Hum Mol Genet **22**(4): 795-803.

Wu, C. C., K. Kannan, S. Lin, L. Yen and A. Milosavljevic (2013). "Identification of cancer fusion drivers using network fusion centrality." Bioinformatics **29**(9): 1174-1181.

Wyatt, A. W., F. Mo, K. Wang, B. McConeghy, S. Brahmabhatt, L. Jong, D. M. Mitchell, R. L. Johnston, A. Haegert, E. Li, J. Liew, J. Yeung, R. Shrestha, A. V. Lapuk, A. McPherson, R. Shukin, R. H. Bell, S. Anderson, J. Bishop, A. Hurtado-Coll, H. Xiao, A. M. Chinnaiyan, R. Mehra, D. Lin, Y. Wang, L. Fazli, M. E. Gleave, S. V. Volik and C. C. Collins (2014). "Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer." Genome Biol **15**(8): 426.

Yoshihara, K., Q. Wang, W. Torres-Garcia, S. Zheng, R. Vegesna, H. Kim and R. G. Verhaak (2015). "The landscape and therapeutic relevance of cancer-associated transcript fusions." Oncogene **34**(37): 4845-4854.

Zhang, Y., M. Gong, H. Yuan, H. G. Park, H. F. Frierson and H. Li (2012). "Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation." Cancer Discov **2**(7): 598-607.

Zhou, J. X., H. Yang, Q. Deng, X. Gu, P. He, Y. Lin, M. Zhao, J. Jiang, H. Chen, Y. Lin, W. Yin, L. Mo and J. He (2013). "Oncogenic driver mutations in patients with non-small-cell lung cancer at various clinical stages." Ann Oncol **24**(5): 1319-1325.

Aine, M., P. Eriksson, F. Liedberg, M. Hoglund and G. Sjodahl (2015). "On Molecular Classification of Bladder Cancer: Out of One, Many." Eur Urol **68**(6): 921-923.

Alentorn, A., C. Dehais, F. Ducray, C. Carpentier, K. Mokhtari, D. Figarella-Branger, O. Chinot, E. Cohen-Moyal, C. Ramirez, H. Loiseau, S. Elouahdani-Hamdi, P. Beauchesne, O. Langlois, C. Desenclos, J. S. Guillo, P. Dam-Hieu, F. Ghiringhelli, P. Colin, J. Godard, F. Parker, F. Dhermain, A. F. Carpentier, J. S. Frenel, P. Menei, L. Bauchet, T. Faillot, M. Fesneau, D. Fontaine, M. J. Motuo-Fotso, E. Vauleon, C. Gaultier, C. Le Guerinel, E. M. Gueye, G. Noel, N. Desse, X. Durando, E. Barrascout, M. Wager, D. Ricard, I. Carpiuc, J. Y. Delattre, A. Idbaih and P. Network (2015). "Allelic loss of 9p21.3 is a prognostic factor in 1p/19q codeleted anaplastic gliomas." Neurology **85**(15): 1325-1331.

Arai, Y., Y. Totoki, F. Hosoda, T. Shiota, N. Hama, H. Nakamura, H. Ojima, K. Furuta, K. Shimada, T. Okusaka, T. Kosuge and T. Shibata (2014). "Fibroblast growth factor receptor 2 tyrosine kinase fusions define a unique molecular subtype of cholangiocarcinoma." Hepatology **59**(4): 1427-1434.

Aran, D., M. Sirota and A. J. Butte (2015). "Systematic pan-cancer analysis of tumour purity." Nat Commun **6**: 8971.

Aso, T., H. Uozaki, S. Morita, A. Kumagai and M. Watanabe (2015). "Loss of ARID1A, ARID1B, and ARID2 Expression During Progression of Gastric Cancer." Anticancer Res **35**(12): 6819-6827.

Bao, S., Q. Wu, R. E. McLendon, Y. Hao, Q. Shi, A. B. Hjelmeland, M. W. Dewhirst, D. D. Bigner and J. N. Rich (2006). "Glioma stem cells promote radioresistance by preferential activation of the DNA damage response." Nature **444**(7120): 756-760.

Bao, Z. S., H. M. Chen, M. Y. Yang, C. B. Zhang, K. Yu, W. L. Ye, B. Q. Hu, W. Yan, W. Zhang, J. Akers, V. Ramakrishnan, J. Li, B. Carter, Y. W. Liu, H. M. Hu, Z. Wang, M. Y. Li, K. Yao, X. G. Qiu, C. S. Kang, Y. P. You, X. L. Fan, W. S. Song, R. Q. Li, X. D. Su, C. C. Chen and T. Jiang (2014). "RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas." Genome Res **24**(11): 1765-1773.

Barbie, D. A., P. Tamayo, J. S. Boehm, S. Y. Kim, S. E. Moody, I. F. Dunn, A. C. Schinzel, P. Sandy, E. Meylan, C. Scholl, S. Frohling, E. M. Chan, M. L. Sos, K. Michel, C. Mermel, S. J. Silver, B. A. Weir, J. H. Reiling, Q. Sheng, P. B. Gupta, R. C. Wadlow, H. Le, S. Hoersch, B. S. Wittner, S. Ramaswamy, D. M. Livingston, D. M. Sabatini, M. Meyerson, R. K. Thomas, E. S. Lander, J. P. Mesirov, D. E. Root, D. G. Gilliland, T. Jacks and W. C. Hahn (2009). "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1." Nature **462**(7269): 108-112.

Barbieri, C. E., F. Demichelis and M. A. Rubin (2012). "Molecular genetics of prostate cancer: emerging appreciation of genetic complexity." Histopathology **60**(1): 187-198.

Baysan, M., S. Bozdog, M. C. Cam, S. Kotliarova, S. Ahn, J. Walling, J. K. Killian, H. Stevenson, P. Meltzer and H. A. Fine (2012). "G-cimp status prediction of glioblastoma samples using mRNA expression data." PLoS One **7**(11): e47839.

Becht, E., N. A. Giraldo, B. Beuselinck, S. Job, L. Marisa, Y. Vano, S. Oudard, J. Zucman-Rossi, P. Laurent-Puig, C. Sautes-Fridman, A. de Reynies and W. H. Fridman (2015). "Prognostic and theranostic impact of molecular subtypes and immune classifications in renal cell cancer (RCC) and colorectal cancer (CRC)." Oncoimmunology **4**(12): e1049804.

Bengtsson, H., A. Ray, P. Spellman and T. P. Speed (2009). "A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods." Bioinformatics **25**(7): 861-867.

Bennett, G., D. Sadlier, P. P. Doran, P. Macmathuna and D. W. Murray (2011). "A functional and transcriptomic analysis of NET1 bioactivity in gastric cancer." BMC Cancer **11**: 50.

Berger, M. F., J. Z. Levin, K. Vijayendran, A. Sivachenko, X. Adiconis, J. Maguire, L. A. Johnson, J. Robinson, R. G. Verhaak, C. Sougnez, R. C. Onofrio, L. Ziaugra, K. Cibulskis, E. Laine, J. Barretina, W. Winckler, D. E. Fisher, G. Getz, M. Meyerson, D. B. Jaffe, S. B. Gabriel, E. S. Lander, R. Dummer, A. Gnirke, C. Nusbaum and L. A. Garraway (2010). "Integrative analysis of the melanoma transcriptome." Genome Res **20**(4): 413-427.

Bhat, K. P., V. Balasubramaniyan, B. Vaillant, R. Ezhilarasan, K. Hummelink, F. Hollingsworth, K. Wani, L. Heathcock, J. D. James, L. D. Goodman, S. Conroy, L. Long, N. Lelic, S. Wang, J. Gumin, D. Raj, Y. Kodama, A. Raghunathan, A. Olar, K. Joshi, C. E. Pelloso, A. Heimberger, S. H. Kim, D. P. Cahill, G. Rao, W. F. Den Dunnen, H. W. Boddeke, H. S. Phillips, I. Nakano, F. F. Lang, H. Colman, E. P. Sulman and K. Aldape (2013). "Mesenchymal differentiation mediated by NF-kappaB promotes radiation resistance in glioblastoma." Cancer Cell **24**(3): 331-346.

Blank, C. U., J. B. Haanen, A. Ribas and T. N. Schumacher (2016). "CANCER IMMUNOLOGY. The "cancer immunogram". " Science **352**(6286): 658-660.

Boerno, S. T., C. Grimm, H. Lehrach and M. R. Schweiger (2010). "Next-generation sequencing technologies for DNA methylation analyses in cancer genomics." Epigenomics **2**(2): 199-207.

Bonsor, D. A., D. Beckett and E. J. Sundberg (2015). "Structure of the N-terminal dimerization domain of CEACAM7." Acta Crystallogr F Struct Biol Commun **71**(Pt 9): 1169-1175.

Boots-Sprenger, S. H., A. Sijben, J. Rijntjes, B. B. Tops, A. J. Idema, A. L. Rivera, F. E. Bleeker, A. M. Gijtenbeek, K. Diefes, L. Heathcock, K. D. Aldape, J. W. Jeuken and P. Wesseling (2013). "Significance of complete 1p/19q co-deletion, IDH1 mutation and MGMT promoter methylation in gliomas: use with caution." Mod Pathol **26**(7): 922-929.

Bos, M., M. Gardizi, H. U. Schildhaus, R. Buettner and J. Wolf (2013). "Activated RET and ROS: two new driver mutations in lung adenocarcinoma." Transl Lung Cancer Res **2**(2): 112-121.

Bovelstad, H. M., S. Nygard, H. L. Storvold, M. Aldrin, O. Borgan, A. Frigessi and O. C. Lingjaerde (2007). "Predicting survival from microarray data--a comparative study." Bioinformatics **23**(16): 2080-2087.

Brachtel, E. F., T. N. Operana, P. S. Sullivan, S. E. Kerr, K. A. Cherkis, B. E. Schroeder, S. M. Dry and C. A. Schnabel (2016). "Molecular classification of cancer with the 92-gene assay in cytology and limited tissue samples." Oncotarget **7**(19): 27220-27231.

Brandimarte, L., V. Pierini, D. Di Giacomo, C. Borgia, F. Nozza, P. Gorello, M. Giordan, G. Cazzaniga, G. Te Kronnie, R. La Starza and C. Mecucci (2013). "New MLLT10 gene recombinations in pediatric T-acute lymphoblastic leukemia." Blood **121**(25): 5064-5067.

Brennan, C. W., R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, R. Beroukhi, B. Bernard, C. J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, S. A. Shukla, G. Ciriello, W. K. Yung, W. Zhang, C. Sougnez, T. Mikkelsen, K. Aldape, D. D. Bigner, E. G. Van Meir, M. Prados, A. Sloan, K. L. Black, J. Eschbacher, G. Finocchiaro, W. Friedman, D. W. Andrews, A. Guha, M. Iacocca, B. P. O'Neill, G. Foltz,

J. Myers, D. J. Weisenberger, R. Penny, R. Kucherlapati, C. M. Perou, D. N. Hayes, R. Gibbs, M. Marra, G. B. Mills, E. Lander, P. Spellman, R. Wilson, C. Sander, J. Weinstein, M. Meyerson, S. Gabriel, P. W. Laird, D. Haussler, G. Getz, L. Chin and T. R. Network (2013). "The somatic genomic landscape of glioblastoma." *Cell* **155**(2): 462-477.

Buckner, J. C., D. Gesme, Jr., J. R. O'Fallon, J. E. Hammack, S. Stafford, P. D. Brown, R. Hawkins, B. W. Scheithauer, B. J. Erickson, R. Levitt, E. G. Shaw and R. Jenkins (2003). "Phase II trial of procarbazine, lomustine, and vincristine as initial therapy for patients with low-grade oligodendroglioma or oligoastrocytoma: efficacy and associations with chromosomal abnormalities." *J Clin Oncol* **21**(2): 251-255.

Buckner, J. C., E. G. Shaw, S. L. Pugh, A. Chakravarti, M. R. Gilbert, G. R. Barger, S. Coons, P. Ricci, D. Bullard, P. D. Brown, K. Stelzer, D. Brachman, J. H. Suh, C. J. Schultz, J. P. Bahary, B. J. Fisher, H. Kim, A. D. Murtha, E. H. Bell, M. Won, M. P. Mehta and W. J. Curran, Jr. (2016). "Radiation plus Procarbazine, CCNU, and Vincristine in Low-Grade Glioma." *N Engl J Med* **374**(14): 1344-1355.

Cairncross, G., M. Wang, E. Shaw, R. Jenkins, D. Brachman, J. Buckner, K. Fink, L. Souhami, N. Laperriere, W. Curran and M. Mehta (2013). "Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402." *J Clin Oncol* **31**(3): 337-343.

Cancer Genome Atlas Research, N. (2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." *Nature* **455**(7216): 1061-1068.

Cancer Genome Atlas Research, N., D. J. Brat, R. G. Verhaak, K. D. Aldape, W. K. Yung, S. R. Salama, L. A. Cooper, E. Rheinbay, C. R. Miller, M. Vitucci, O. Morozova, A. G. Robertson, H. Nounshmehr, P. W. Laird, A. D. Cherniack, R. Akbani, J. T. Huse, G. Ciriello, L. M. Poisson, J. S. Barnholtz-Sloan, M. S. Berger, C. Brennan, R. R. Colen, H. Colman, A. E. Flanders, C. Giannini, M. Grifford, A. Iavarone, R. Jain, I. Joseph, J. Kim, K. Kasaian, T. Mikkelsen, B. A. Murray, B. P. O'Neill, L. Pachter, D. W. Parsons, C. Sougnez, E. P. Sulman, S. R. Vandenberg, E. G. Van Meir, A. von Deimling, H. Zhang, D. Crain, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, T. Shelton, M. Sherman, P. Yena, A. Black, J. Bowen, K. Dicostanzo, J. Gastier-Foster, K. M. Leraas, T. M. Lichtenberg, C. R. Pierson, N. C. Ramirez, C. Taylor, S. Weaver, L. Wise, E. Zmuda, T. Davidsen, J. A. Demchok, G. Eley, M. L. Ferguson, C. M. Hutter, K. R. Mills Shaw, B. A. Ozenberger, M. Sheth, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, B. Ayala, J. Baboud, S. Chudamani, M. A. Jensen, J. Liu, T. Pihl, R. Raman, Y. Wan, Y. Wu, A. Ally, J. T. Auman, M. Balasundaram, S. Balu, S. B. Baylin, R. Beroukhi, M. S. Bootwalla, R. Bowlby, C. A. Bristow, D. Brooks, Y. Butterfield, R. Carlsen, S. Carter, L. Chin, A. Chu, E. Chuah, K. Cibulskis, A. Clarke, S. G. Coetzee, N. Dhalla, T. Fennell, S. Fisher, S. Gabriel, G. Getz, R. Gibbs, R. Guin, A. Hadjipanayis, D. N. Hayes, T. Hinoue, K. Hoadley, R. A. Holt, A. P. Hoyle, S. R. Jefferys, S. Jones, C. D. Jones, R. Kucherlapati, P. H. Lai, E. Lander, S. Lee, L. Lichtenstein, Y. Ma, D. T. Maglinte, H. S. Mahadeshwar, M. A. Marra, M. Mayo, S. Meng, M. L. Meyerson, P. A. Mieczkowski, R. A. Moore, L. E. Mose, A. J. Mungall, A. Pantazi, M. Parfenov, P. J. Park, J. S. Parker, C. M. Perou, A. Protopopov, X. Ren, J. Roach, T. S. Sabedot, J. Schein, S. E. Schumacher, J. G. Seidman, S. Seth, H. Shen, J. V. Simons, P. Sipahimalani, M. G. Soloway, X. Song, H. Sun, B. Tabak, A. Tam, D. Tan, J. Tang, N. Thiessen, T. Triche, Jr., D. J. Van Den Berg, U. Veluvolu, S. Waring, D. J. Weisenberger, M. D. Wilkerson, T. Wong, J. Wu, L. Xi, A. W. Xu, L. Yang, T. I. Zack, J. Zhang, B. A. Aksoy, H. Arachchi, C. Benz, B. Bernard, D. Carlin, J. Cho, D. DiCara, S. Frazer, G. N. Fuller, J. Gao, N. Gehlenborg, D. Haussler, D. I. Heiman, L. Iype, A. Jacobsen, Z. Ju, S. Katzman, H. Kim, T. Knijnenburg, R. B. Kreisberg, M. S. Lawrence, W. Lee, K. Leinonen, P. Lin, S. Ling, W. Liu, Y. Liu, Y. Liu, Y. Lu, G. Mills, S. Ng, M. S. Noble, E. Paull, A. Rao, S.

Reynolds, G. Saksena, Z. Sanborn, C. Sander, N. Schultz, Y. Senbabaoglu, R. Shen, I. Shmulevich, R. Sinha, J. Stuart, S. O. Sumer, Y. Sun, N. Tasman, B. S. Taylor, D. Voet, N. Weinhold, J. N. Weinstein, D. Yang, K. Yoshihara, S. Zheng, W. Zhang, L. Zou, T. Abel, S. Sadeghi, M. L. Cohen, J. Eschbacher, E. M. Hattab, A. Raghunathan, M. J. Schniederjan, D. Aziz, G. Barnett, W. Barrett, D. D. Bigner, L. Boice, C. Brewer, C. Calatuzzolo, B. Campos, C. G. Carlotti, Jr., T. A. Chan, L. Cuppini, E. Curley, S. Cuzzubbo, K. Devine, F. DiMeco, R. Duell, J. B. Elder, A. Fehrenbach, G. Finocchiaro, W. Friedman, J. Fulop, J. Gardner, B. Hermes, C. Herold-Mende, C. Jungk, A. Kendler, N. L. Lehman, E. Lipp, O. Liu, R. Mandt, M. McGraw, R. McLendon, C. McPherson, L. Neder, P. Nguyen, A. Noss, R. Nunziata, Q. T. Ostrom, C. Palmer, A. Perin, B. Pollo, A. Potapov, O. Potapova, W. K. Rathmell, D. Rotin, L. Scarpacci, C. Schilero, K. Senecal, K. Shimmel, V. Shurkhay, S. Sifri, R. Singh, A. E. Sloan, K. Smolenski, S. M. Staugaitis, R. Steele, L. Thorne, D. P. Tirapelli, A. Unterberg, M. Vallurupalli, Y. Wang, R. Warnick, F. Williams, Y. Wolinsky, S. Bell, M. Rosenberg, C. Stewart, F. Huang, J. L. Grimsby, A. J. Radenbaugh and J. Zhang (2015). "Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas." N Engl J Med **372**(26): 2481-2498.

Candido, J. and T. Hagemann (2013). "Cancer-related inflammation." J Clin Immunol **33 Suppl 1**: S79-84.

Carr, H. S., Y. Zuo, W. Oh and J. A. Frost (2013). "Regulation of focal adhesion kinase activation, breast cancer cell motility, and amoeboid invasion by the RhoA guanine nucleotide exchange factor Net1." Mol Cell Biol **33**(14): 2773-2786.

Carter, S. L., K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, R. Beroukhim, D. Pellman, D. A. Levine, E. S. Lander, M. Meyerson and G. Getz (2012). "Absolute quantification of somatic DNA alterations in human cancer." Nat Biotechnol **30**(5): 413-421.

Ceccarelli, M., F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta, S. Anjum, J. Wang, G. Manyam, P. Zoppoli, S. Ling, A. A. Rao, M. Grifford, A. D. Cherniack, H. Zhang, L. Poisson, C. G. Carlotti, Jr., D. P. Tirapelli, A. Rao, T. Mikkelsen, C. C. Lau, W. K. Yung, R. Rabadan, J. Huse, D. J. Brat, N. L. Lehman, J. S. Barnholtz-Sloan, S. Zheng, K. Hess, G. Rao, M. Meyerson, R. Beroukhim, L. Cooper, R. Akbani, M. Wrensch, D. Haussler, K. D. Aldape, P. W. Laird, D. H. Gutmann, T. R. Network, H. Nounshmehr, A. Iavarone and R. G. Verhaak (2016). "Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma." Cell **164**(3): 550-563.

Chang, J., L. Huang, Q. Cao and F. Liu (2016). "Identification of colorectal cancer-restricted microRNAs and their target genes based on high-throughput sequencing data." Onco Targets Ther **9**: 1787-1794.

Chen, F., X. Zhuang, L. Lin, P. Yu, Y. Wang, Y. Shi, G. Hu and Y. Sun (2015). "New horizons in tumor microenvironment biology: challenges and opportunities." BMC Med **13**: 45.

Chen, T., X. Y. Xu and P. H. Zhou (2016). "Emerging molecular classifications and therapeutic implications for gastric cancer." Chin J Cancer **35**: 49.

Chen, T., T. Zhou, B. He, H. Yu, X. Guo, X. Song and J. Sha (2014). "mUbiSiDa: a comprehensive database for protein ubiquitination sites in mammals." PLoS One **9**(1): e85744.

Chen, Y. and S. H. Tseng (2014). "Targeting tropomyosin-receptor kinase fused gene in cancer." Anticancer Res **34**(4): 1595-1600.

Chen, Z., Y. Zhou, Z. Zhang and J. Song (2015). "Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features." Brief Bioinform **16**(4): 640-657.

Chiang, C., R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose, E. P. Garrison, G. T. Marth, A. R. Quinlan and I. M. Hall (2015). "SpeedSeq: ultra-fast personal genome analysis and interpretation." Nat Methods **12**(10): 966-968.

Choe, C., Y. S. Shin, C. Kim, S. J. Choi, J. Lee, S. Y. Kim, Y. B. Cho and J. Kim (2015). "Crosstalk with cancer-associated fibroblasts induces resistance of non-small cell lung cancer cells to epidermal growth factor receptor tyrosine kinase inhibition." Onco Targets Ther **8**: 3665-3678.

Cortesi, E., M. Palleschi, V. Magri and G. Naso (2015). "The promise of liquid biopsy in cancer: a clinical perspective." Chin J Cancer Res **27**(5): 488-490.

Dai, L., P. Koutrakis, B. A. Coull, D. Sparrow, P. S. Vokonas and J. D. Schwartz (2016). "Use of the Adaptive LASSO Method to Identify PM2.5 Components Associated with Blood Pressure in Elderly Men: The Veterans Affairs Normative Aging Study." Environ Health Perspect **124**(1): 120-125.

Dai, M., P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson and F. Meng (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucleic Acids Res **33**(20): e175.

Deng, L., H. Liang, B. Burnette, M. Beckett, T. Darga, R. R. Weichselbaum and Y. X. Fu (2014). "Irradiation and anti-PD-L1 treatment synergistically promote antitumor immunity in mice." J Clin Invest **124**(2): 687-695.

Di Cerbo, V. and R. Schneider (2013). "Cancers with wrong HATs: the impact of acetylation." Brief Funct Genomics **12**(3): 231-243.

Doucette, T., G. Rao, A. Rao, L. Shen, K. Aldape, J. Wei, K. Dziurzynski, M. Gilbert and A. B. Heimberger (2013). "Immune heterogeneity of glioblastoma subtypes: extrapolation from the cancer genome atlas." Cancer Immunol Res **1**(2): 112-122.

Dubbink, H. J., P. N. Atmodimedjo, J. M. Kros, P. J. French, M. Sanson, A. Idbaih, P. Wesseling, R. Enting, W. Spliet, C. Tijssen, W. N. Dinjens, T. Gorlia and M. J. van den Bent (2015). "Molecular classification of anaplastic oligodendroglioma using next-generation sequencing: a report of the prospective randomized EORTC Brain Tumor Group 26951 phase III trial." Neuro Oncol.

Dunn, G. P., M. L. Rinne, J. Wykosky, G. Genovese, S. N. Quayle, I. F. Dunn, P. K. Agarwalla, M. G. Chheda, B. Campos, A. Wang, C. Brennan, K. L. Ligon, F. Furnari, W. K. Cavenee, R. A. Depinho, L.

Chin and W. C. Hahn (2012). "Emerging insights into the molecular and cellular basis of glioblastoma." Genes Dev **26**(8): 756-784.

Eckel-Passow, J. E., D. H. Lachance, A. M. Molinaro, K. M. Walsh, P. A. Decker, H. Sicotte, M. Pekmezci, T. Rice, M. L. Kosel, I. V. Smirnov, G. Sarkar, A. A. Caron, T. M. Kollmeyer, C. E. Praska, A. R. Chada, C. Halder, H. M. Hansen, L. S. McCoy, P. M. Bracci, R. Marshall, S. Zheng, G. F. Reis, A. R. Pico, B. P. O'Neill, J. C. Buckner, C. Giannini, J. T. Huse, A. Perry, T. Tihan, M. S. Berger, S. M. Chang, M. D. Prados, J. Wiemels, J. K. Wiencke, M. R. Wrensch and R. B. Jenkins (2015). "Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors." N Engl J Med **372**(26): 2499-2508.

Eder, K. and B. Kalman (2014). "Molecular heterogeneity of glioblastoma and its clinical relevance." Pathol Oncol Res **20**(4): 777-787.

Eguchi, F. C., E. F. Faria, C. Scapulatempo Neto, A. Longatto-Filho, C. Zanardo-Oliveira, S. R. Taboga and S. G. Campos (2014). "The role of TMPRSS2:ERG in molecular stratification of PCa and its association with tumor aggressiveness: a study in Brazilian patients." Sci Rep **4**: 5640.

Elkhattouti, A., M. Hassan and C. R. Gomez (2015). "Stromal Fibroblast in Age-Related Cancer: Role in Tumorigenesis and Potential as Novel Therapeutic Target." Front Oncol **5**: 158.

Engler, J. R., A. E. Robinson, I. Smirnov, J. G. Hodgson, M. S. Berger, N. Gupta, C. D. James, A. Molinaro and J. J. Phillips (2012). "Increased microglia/macrophage gene expression in a subset of adult and pediatric astrocytomas." PLoS One **7**(8): e43339.

Figarella-Branger, D., K. Mokhtari, C. Dehais, A. Jouvet, E. Uro-Coste, C. Colin, C. Carpentier, F. Forest, C. A. Maurage, J. M. Vignaud, M. Polivka, E. Lechapt-Zalcman, S. Eimer, G. Viennet, I. Quintin-Roue, M. H. Aubriot-Lorton, M. D. Diebold, D. Loussouarn, C. Lacroix, V. Rigau, A. Laquerriere, F. Vandenbos, S. Michalak, H. Sevestre, M. Pech, F. Labrousse, C. Christov, J. L. Kemeny, M. P. Chenard, D. Chiforeanu, F. Ducray, A. Idhah and P. Network (2014). "Mitotic index, microvascular proliferation, and necrosis define 3 groups of 1p/19q codeleted anaplastic oligodendrogliomas associated with different genomic alterations." Neuro Oncol **16**(9): 1244-1254.

Flynn, A., D. Benn, R. Clifton-Bligh, B. Robinson, A. H. Trainer, P. James, A. Hogg, K. Waldeck, J. George, J. Li, S. B. Fox, A. J. Gill, G. McArthur, R. J. Hicks and R. W. Tothill (2015). "The genomic landscape of pheochromocytoma." J Pathol **236**(1): 78-89.

Francis, P., H. M. Namlos, C. Muller, P. Eden, J. Fernebro, J. M. Berner, B. Bjerkehagen, M. Akerman, P. O. Bendahl, A. Isinger, A. Rydholm, O. Myklebost and M. Nilbert (2007). "Diagnostic and prognostic gene expression signatures in 177 soft tissue sarcomas: hypoxia-induced transcription profile signifies metastatic potential." BMC Genomics **8**: 73.

Freije, W. A., F. E. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, L. M. Liau, P. S. Mischel and S. F. Nelson (2004). "Gene expression profiling of gliomas strongly predicts survival." Cancer Res **64**(18): 6503-6510.

Frenkel-Morgenstern, M., V. Lacroix, I. Ezkurdia, Y. Levin, A. Gabashvili, J. Prilusky, A. Del Pozo, M. Tress, R. Johnson, R. Guigo and A. Valencia (2012). "Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts." Genome Res **22**(7): 1231-1242.

Fridman, W. H., F. Pages, C. Sautès-Fridman and J. Galon (2012). "The immune contexture in human tumours: impact on clinical outcome." Nat Rev Cancer **12**(4): 298-306.

Friedman, J., T. Hastie and R. Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." J Stat Softw **33**(1): 1-22.

Gabrusiewicz, K., B. Rodriguez, J. Wei, Y. Hashimoto, L. M. Healy, S. N. Maiti, G. Thomas, S. Zhou, Q. Wang, A. Elakkad, B. D. Liebelt, N. K. Yaghi, R. Ezhilarasan, N. Huang, J. S. Weinberg, S. S. Prabhu, G. Rao, R. Sawaya, L. A. Langford, J. M. Bruner, G. N. Fuller, A. Bar-Or, W. Li, R. R. Colen, M. A. Curran, K. P. Bhat, J. P. Antel, L. J. Cooper, E. P. Sulman and A. B. Heimberger (2016). "Glioblastoma-infiltrated innate immune cells resemble M0 macrophage phenotype." JCI Insight **1**(2).

Gagan, J. and E. M. Van Allen (2015). "Next-generation sequencing to guide cancer therapy." Genome Med **7**(1): 80.

Galli, R., E. Binda, U. Orfanelli, B. Cipelletti, A. Gritti, S. De Vitis, R. Fiocco, C. Foroni, F. Dimeco and A. Vescovi (2004). "Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma." Cancer Res **64**(19): 7011-7021.

Gill, B. J., D. J. Pisapia, H. R. Malone, H. Goldstein, L. Lei, A. Sonabend, J. Yun, J. Samanamud, J. S. Sims, M. Banu, A. Dovas, A. F. Teich, S. A. Sheth, G. M. McKhann, M. B. Sisti, J. N. Bruce, P. A. Sims and P. Canoll (2014). "MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma." Proc Natl Acad Sci U S A **111**(34): 12550-12555.

Goodman, V. L., G. J. Brewer and S. D. Merajver (2004). "Copper deficiency as an anti-cancer strategy." Endocr Relat Cancer **11**(2): 255-263.

Gravendeel, L. A., M. C. Kouwenhoven, O. Gevaert, J. J. de Rooi, A. P. Stubbs, J. E. Duijm, A. Daemen, F. E. Bleeker, L. B. Bralten, N. K. Kloosterhof, B. De Moor, P. H. Eilers, P. J. van der Spek, J. M. Kros, P. A. Sillevius Smitt, M. J. van den Bent and P. J. French (2009). "Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology." Cancer Res **69**(23): 9065-9072.

Greco, F. A. (2013). "The impact of molecular testing on treatment of cancer of unknown primary origin." Oncology (Williston Park) **27**(8): 815-817.

Guan, X., J. Vengoechea, S. Zheng, A. E. Sloan, Y. Chen, D. J. Brat, B. P. O'Neill, J. de Groot, S. Yust-Katz, W. K. Yung, M. L. Cohen, K. D. Aldape, S. Rosenfeld, R. G. Verhaak and J. S. Barnholtz-Sloan (2014). "Molecular subtypes of glioblastoma are relevant to lower grade glioma." PLoS One **9**(3): e91216.

Guarnerio, J., M. Bezzi, J. C. Jeong, S. V. Paffenholz, K. Berry, M. M. Naldini, F. Lo-Coco, Y. Tay, A. H. Beck and P. P. Pandolfi (2016). "Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations." Cell **165**(2): 289-302.

Guo, J., L. Canaff, C. V. Rajadurai, N. Fils-Aime, J. Tian, M. Dai, J. Korah, M. Villatoro, M. Park, S. Ali and J. J. Lebrun (2014). "Breast cancer anti-estrogen resistance 3 inhibits transforming growth factor beta/Smad signaling and associates with favorable breast cancer disease outcomes." Breast Cancer Res **16**(6): 476.

Hambardzumyan, D., D. H. Gutmann and H. Kettenmann (2015). "The role of microglia and macrophages in glioma maintenance and progression." Nat Neurosci **19**(1): 20-27.

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Hanzelmann, S., R. Castelo and J. Guinney (2013). "GSVA: gene set variation analysis for microarray and RNA-seq data." BMC Bioinformatics **14**: 7.

Helin, K. and D. Dhanak (2013). "Chromatin proteins and modifications as drug targets." Nature **502**(7472): 480-488.

Hermans, K. G., R. van Marion, H. van Dekken, G. Jenster, W. M. van Weerden and J. Trapman (2006). "TMPRSS2:ERG fusion by translocation or interstitial deletion is highly relevant in androgen-dependent prostate cancer, but is bypassed in late-stage androgen receptor-negative prostate cancer." Cancer Res **66**(22): 10658-10663.

Hoadley, K. A., C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, J. Zhang, C. Kandoth, R. Akbani, H. Shen, L. Omberg, A. Chu, A. A. Margolin, L. J. Van't Veer, N. Lopez-Bigas, P. W. Laird, B. J. Raphael, L. Ding, A. G. Robertson, L. A. Byers, G. B. Mills, J. N. Weinstein, C. Van Waes, Z. Chen, E. A. Collisson, N. Cancer Genome Atlas Research, C. C. Benz, C. M. Perou and J. M. Stuart (2014). "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin." Cell **158**(4): 929-944.

Huang, K. K., J. R. McPherson, S. T. Tay, K. Das, I. B. Tan, C. C. Ng, N. Y. Chia, S. L. Zhang, S. S. Myint, L. Hu, V. Rajasegaran, D. Huang, J. L. Loh, A. Gan, A. N. Sairi, X. X. Sam, L. T. Dominguez, M. Lee, K. C. Soo, L. L. Ooi, H. S. Ong, A. Chung, P. K. Chow, W. K. Wong, S. Selvarajan, C. K. Ong, K. H. Lim, T. Nandi, S. Rozen, B. T. Teh, R. Quek and P. Tan (2015). "SETD2 histone modifier loss in aggressive GI stromal tumours." Gut.

Huang, Q., V. E. Schneeberger, N. Luetkeke, C. Jin, R. Afzal, M. M. Budzevich, R. J. Mankanji, G. V. Martinez, T. Shen, L. Zhao, K. M. Fung, E. B. Haura, D. Coppola and J. Wu (2016). "Preclinical Modeling of KIF5B-RET Fusion Lung Adenocarcinoma." Mol Cancer Ther **15**(10): 2521-2529.

Huang, Y. T., T. Hsu, K. T. Kelsey and C. L. Lin (2015). "Integrative analysis of micro-RNA, gene expression, and survival of glioblastoma multiforme." Genet Epidemiol **39**(2): 134-143.

Hughey, J. J. and A. J. Butte (2015). "Robust meta-analysis of gene expression using the elastic net." Nucleic Acids Res **43**(12): e79.

Hunter, C., R. Smith, D. P. Cahill, P. Stephens, C. Stevens, J. Teague, C. Greenman, S. Edkins, G. Bignell, H. Davies, S. O'Meara, A. Parker, T. Avis, S. Barthorpe, L. Brackenbury, G. Buck, A. Butler, J.

Clements, J. Cole, E. Dicks, S. Forbes, M. Gorton, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, V. Kosmidou, R. Laman, R. Lugg, A. Menzies, J. Perry, R. Petty, K. Raine, D. Richardson, R. Shepherd, A. Small, H. Solomon, C. Tofts, J. Varian, S. West, S. Widaa, A. Yates, D. F. Easton, G. Riggins, J. E. Roy, K. K. Levine, W. Mueller, T. T. Batchelor, D. N. Louis, M. R. Stratton, P. A. Futreal and R. Wooster (2006). "A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy." Cancer Res **66**(8): 3987-3991.

Huse, J. T., H. S. Phillips and C. W. Brennan (2011). "Molecular subclassification of diffuse gliomas: seeing order in the chaos." Glia **59**(8): 1190-1199.

Husnjak, K. and I. Dikic (2012). "Ubiquitin-binding proteins: decoders of ubiquitin-mediated cellular functions." Annu Rev Biochem **81**: 291-322.

Isella, C., A. Terrasi, S. E. Bellomo, C. Petti, G. Galatola, A. Muratore, A. Mellano, R. Senetta, A. Cassenti, C. Sonetto, G. Inghirami, L. Trusolino, Z. Fekete, M. De Ridder, P. Cassoni, G. Storme, A. Bertotti and E. Medico (2015). "Stromal contribution to the colorectal cancer transcriptome." Nat Genet **47**(4): 312-319.

Iuliano, A., A. Occhipinti, C. Angelini, I. De Feis and P. Lio (2016). "Cancer Markers Selection Using Network-Based Cox Regression: A Methodological and Computational Practice." Front Physiol **7**: 208.

Jiang, B., J. Mason, A. Jewett, M. L. Liu, W. Chen, J. Qian, Y. Ding, S. Ding, M. Ni, X. Zhang and Y. G. Man (2013). "Tumor-infiltrating immune cells: triggers for tumor capsule disruption and tumor progression?" Int J Med Sci **10**(5): 475-497.

Jividen, K. and H. Li (2014). "Chimeric RNAs generated by intergenic splicing in normal and cancer cells." Genes Chromosomes Cancer **53**(12): 963-971.

Johnson, W. E., C. Li and A. Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." Biostatistics **8**(1): 118-127.

Joo, K. M., J. Kim, J. Jin, M. Kim, H. J. Seol, J. Muradov, H. Yang, Y. L. Choi, W. Y. Park, D. S. Kong, J. I. Lee, Y. H. Ko, H. G. Woo, J. Lee, S. Kim and D. H. Nam (2013). "Patient-specific orthotopic glioblastoma xenograft models recapitulate the histopathology and biology of human glioblastomas in situ." Cell Rep **3**(1): 260-273.

Kalvik, T. V. and T. Arnesen (2013). "Protein N-terminal acetyltransferases in cancer." Oncogene **32**(3): 269-276.

Kannan, K., C. Coarfa, P. W. Chao, L. Luo, Y. Wang, A. E. Brinegar, S. M. Hawkins, A. Milosavljevic, M. M. Matzuk and L. Yen (2015). "Recurrent BCAM-AKT2 fusion gene leads to a constitutively activated AKT2 fusion kinase in high-grade serous ovarian carcinoma." Proc Natl Acad Sci U S A **112**(11): E1272-1277.

Kanu, N., E. Gronroos, P. Martinez, R. A. Burrell, X. Yi Goh, J. Bartkova, A. Maya-Mendoza, M. Mistrik, A. J. Rowan, H. Patel, A. Rabinowitz, P. East, G. Wilson, C. R. Santos, N. McGranahan, S. Gulati, M. Gerlinger, N. J. Birkbak, T. Joshi, L. B. Alexandrov, M. R. Stratton, T. Powles, N. Matthews, P. A. Bates,

A. Stewart, Z. Szallasi, J. Larkin, J. Bartek and C. Swanton (2015). "SETD2 loss-of-function promotes renal cancer branched evolution through replication stress and impaired DNA repair." Oncogene **34**(46): 5699-5708.

Katsios, C., D. E. Ziogas, T. Liakakos, O. Zoras and D. H. Roukos (2012). "Translating cancer genomes sequencing revolution into surgical oncology practice." J Surg Res **173**(2): 365-369.

Kim, H. and R. G. Verhaak (2015). "Transcriptional mimicry by tumor-associated stroma." Nat Genet **47**(4): 307-309.

Kim, H., S. Zheng, S. S. Amini, S. M. Virk, T. Mikkelsen, D. J. Brat, J. Grimsby, C. Sougnez, F. Muller, J. Hu, A. E. Sloan, M. L. Cohen, E. G. Van Meir, L. Scarpance, P. W. Laird, J. N. Weinstein, E. Lander, S. Gabriel, G. Getz, M. Meyerson, L. Chin, J. Barnholtz-Sloan and R. G. W. Verhaak (2014). "Whole genome and multi-sector sequencing of primary and post-treatment glioblastoma illustrates patterns of tumor evolution." Submitted.

Kim, H., S. Zheng, S. S. Amini, S. M. Virk, T. Mikkelsen, D. J. Brat, J. Grimsby, C. Sougnez, F. Muller, J. Hu, A. E. Sloan, M. L. Cohen, E. G. Van Meir, L. Scarpance, P. W. Laird, J. N. Weinstein, E. S. Lander, S. Gabriel, G. Getz, M. Meyerson, L. Chin, J. S. Barnholtz-Sloan and R. G. Verhaak (2015). "Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution." Genome Res **25**(3): 316-327.

Kirkin, V. and I. Dikic (2011). "Ubiquitin networks in cancer." Curr Opin Genet Dev **21**(1): 21-28.

Kocarnik, J. M., S. Shiovitz and A. I. Phipps (2015). "Molecular phenotypes of colorectal cancer and potential clinical applications." Gastroenterol Rep (Oxf) **3**(4): 269-276.

Kohno, T., H. Ichikawa, Y. Totoki, K. Yasuda, M. Hiramoto, T. Nammo, H. Sakamoto, K. Tsuta, K. Furuta, Y. Shimada, R. Iwakawa, H. Ogiwara, T. Oike, M. Enari, A. J. Schetter, H. Okayama, A. Haugen, V. Skaug, S. Chiku, I. Yamanaka, Y. Arai, S. Watanabe, I. Sekine, S. Ogawa, C. C. Harris, H. Tsuda, T. Yoshida, J. Yokota and T. Shibata (2012). "KIF5B-RET fusions in lung adenocarcinoma." Nat Med **18**(3): 375-377.

Kreutzberg, G. W. (1996). "Microglia: a sensor for pathological events in the CNS." Trends Neurosci **19**(8): 312-318.

Kubelt, C., K. Hattermann, S. Sebens, H. M. Mehdorn and J. Held-Feindt (2015). "Epithelial-to-mesenchymal transition in paired human primary and recurrent glioblastomas." Int J Oncol **46**(6): 2515-2525.

Kwon, S. M., S. H. Kang, C. K. Park, S. Jung, E. S. Park, J. S. Lee, S. H. Kim and H. G. Woo (2015). "Recurrent Glioblastomas Reveal Molecular Subtypes Associated with Mechanistic Implications of Drug-Resistance." PLoS One **10**(10): e0140528.

Ladanyi, A. (2013). "[Prognostic value of tumor-infiltrating immune cells in melanoma]." Magy Onkol **57**(2): 85-95.

Lee, J. H., H. F. Lu, D. Y. Wang, D. R. Chen, C. C. Su, Y. S. Chen, J. H. Yang and J. G. Chung (2004). "Effects of tamoxifen on DNA adduct formation and arylamines N-acetyltransferase activity in human breast cancer cells." Res Commun Mol Pathol Pharmacol **115-116**: 217-233.

Leslie, R., C. J. O'Donnell and A. D. Johnson (2014). "GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database." Bioinformatics **30**(12): i185-194.

Leyten, G. H., D. Hessels, S. A. Jannink, F. P. Smit, H. de Jong, E. B. Cornel, T. M. de Reijke, H. Vergunst, P. Kil, B. C. Knipscheer, I. M. van Oort, P. F. Mulders, C. A. Hulsbergen-van de Kaa and J. A. Schalken (2014). "Prospective multicentre evaluation of PCA3 and TMPRSS2-ERG gene fusions as diagnostic and prognostic urinary biomarkers for prostate cancer." Eur Urol **65**(3): 534-542.

Li, H. and Y. Luan (2005). "Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data." Bioinformatics **21**(10): 2403-2409.

Li, J., G. Duns, H. Westers, R. Sijmons, A. van den Berg and K. Kok (2016). "SETD2: an epigenetic modifier with tumor suppressor functionality." Oncotarget.

Lipinski, K. A., L. J. Barber, M. N. Davies, M. Ashenden, A. Sottoriva and M. Gerlinger (2016). "Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine." Trends Cancer **2**(1): 49-63.

Liu, J., A. Masurekar, S. Johnson, S. Chakraborty, J. Griffiths, D. Smith, S. Alexander, C. Dempsey, C. Parker, S. Harrison, Y. Li, C. Miller, Y. Di, Z. Ghosh, S. Krishnan and V. Saha (2015). "Stromal cell-mediated mitochondrial redox adaptation regulates drug resistance in childhood acute lymphoblastic leukemia." Oncotarget **6**(40): 43048-43064.

Liu, L. Z., F. X. Wu and W. J. Zhang (2014). "A group LASSO-based method for robustly inferring gene regulatory networks from multiple time-course datasets." BMC Syst Biol **8 Suppl 3**: S1.

Louis, D. N., A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues and D. W. Ellison (2016). "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary." Acta Neuropathol **131**(6): 803-820.

Lu, K. H., K. L. Lin, T. C. Hsia, C. F. Hung, M. C. Chou, Y. M. Hsiao and J. G. Chung (2001). "Tamoxifen inhibits arylamine N-acetyltransferase activity and DNA-2-aminofluorene adduct in human leukemia HL-60 cells." Res Commun Mol Pathol Pharmacol **109**(5-6): 319-331.

Luo, L. Y., E. Kim, H. W. Cheung, B. A. Weir, G. P. Dunn, R. R. Shen and W. C. Hahn (2015). "The Tyrosine Kinase Adaptor Protein FRS2 Is Oncogenic and Amplified in High-Grade Serous Ovarian Cancer." Mol Cancer Res **13**(3): 502-509.

Madhavan, S., J. C. Zenklusen, Y. Kotliarov, H. Sahni, H. A. Fine and K. Buetow (2009). "Rembrandt: helping personalized medicine become a reality through integrative translational research." Mol Cancer Res **7**(2): 157-167.

Mani, A. and E. P. Gelmann (2005). "The ubiquitin-proteasome pathway and its role in cancer." J Clin Oncol **23**(21): 4776-4789.

Markman, J. L. and S. L. Shiao (2015). "Impact of the immune system and immunotherapy in colorectal cancer." J Gastrointest Oncol **6**(2): 208-223.

Martinez-Ledesma, E., J. F. de Groot and R. G. Verhaak (2015). "Seek and destroy: relating cancer drivers to therapies." Cancer Cell **27**(3): 319-321.

Martinez, E., K. Yoshihara, H. Kim, G. M. Mills, V. Trevino and R. G. Verhaak (2015). "Comparison of gene expression patterns across 12 tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects." Oncogene **34**(21): 2732-2740.

Mazor, T., A. Pankov, B. E. Johnson, C. Hong, E. G. Hamilton, R. J. Bell, I. V. Smirnov, G. F. Reis, J. J. Phillips, M. J. Barnes, A. Idbaih, A. Alentorn, J. J. Kloezean, M. L. Lamfers, A. W. Bollen, B. S. Taylor, A. M. Molinaro, A. B. Olshen, S. M. Chang, J. S. Song and J. F. Costello (2015). "DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors." Cancer Cell **28**(3): 307-317.

McGowan, M. L., R. A. Settersten, Jr., E. T. Juengst and J. R. Fishman (2014). "Integrating genomics into clinical oncology: ethical and social challenges from proponents of personalized medicine." Urol Oncol **32**(2): 187-192.

Medves, S. and J. B. Demoulin (2012). "Tyrosine kinase gene fusions in cancer: translating mechanisms into targeted therapies." J Cell Mol Med **16**(2): 237-248.

Meldrum, C., M. A. Doyle and R. W. Tothill (2011). "Next-generation sequencing for cancer diagnostics: a practical perspective." Clin Biochem Rev **32**(4): 177-195.

Meng, Y., M. A. Beckett, H. Liang, H. J. Mauceri, N. van Rooijen, K. S. Cohen and R. R. Weichselbaum (2010). "Blockade of tumor necrosis factor alpha signaling in tumor-associated macrophages as a radiosensitizing strategy." Cancer Res **70**(4): 1534-1543.

Menter, D. G. and R. N. Dubois (2012). "Prostaglandins in cancer cell adhesion, migration, and invasion." Int J Cell Biol **2012**: 723419.

Meyer, C., A. Brieger, G. Plotz, N. Weber, S. Passmann, T. Dinger mann, S. Zeuzem, J. Trojan and R. Marschalek (2009). "An interstitial deletion at 3p21.3 results in the genetic fusion of MLH1 and ITGA9 in a Lynch syndrome family." Clin Cancer Res **15**(3): 762-769.

Moroncini, G., C. Paolini, A. Grieco, G. Nacci, M. Cuccioloni, M. Mozzicafreddo, C. Tonnini, S. Svegliati, M. Angeletti, E. Avvedimento, A. Funaro and A. Gabrielli (2010). "PDGF receptor as therapeutic and diagnostic target in systemic sclerosis." Clinical and Experimental Rheumatology **28**(5): S71-S72.

Muntean, A. G. and J. L. Hess (2009). "Epigenetic dysregulation in cancer." Am J Pathol **175**(4): 1353-1361.

Nacu, S., W. Yuan, Z. Kan, D. Bhatt, C. S. Rivers, J. Stinson, B. A. Peters, Z. Modrusan, K. Jung, S. Seshagiri and T. D. Wu (2011). "Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples." BMC Med Genomics **4**: 11.

Nair, S. S. and R. Kumar (2012). "Chromatin remodeling in cancer: a gateway to regulate gene transcription." Mol Oncol **6**(6): 611-619.

Nakasone, E. S., H. A. Askautrud, T. Kees, J. H. Park, V. Plaks, A. J. Ewald, M. Fein, M. G. Rasch, Y. X. Tan, J. Qiu, J. Park, P. Sinha, M. J. Bissell, E. Frengen, Z. Werb and M. Egeblad (2012). "Imaging tumor-stroma interactions during chemotherapy reveals contributions of the microenvironment to resistance." Cancer Cell **21**(4): 488-503.

Narayan, S., G. D. Bader and J. Reimand (2016). "Frequent mutations in acetylation and ubiquitination sites suggest novel driver mechanisms of cancer." Genome Med **8**(1): 55.

Nehru, V., O. Voytyuk, J. Lennartsson and P. Aspenstrom (2013). "RhoD binds the Rab5 effector Rabankyrin-5 and has a role in trafficking of the platelet-derived growth factor receptor." Traffic **14**(12): 1242-1254.

Newman, A. M., C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn and A. A. Alizadeh (2015). "Robust enumeration of cell subsets from tissue expression profiles." Nat Methods **12**(5): 453-457.

Noushmehr, H., D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloso, E. P. Sulman, K. P. Bhat, R. G. Verhaak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, K. Aldape and N. Cancer Genome Atlas Research (2010). "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma." Cancer Cell **17**(5): 510-522.

Olar, A. and K. D. Aldape (2014). "Using the molecular classification of glioblastoma to inform personalized treatment." J Pathol **232**(2): 165-177.

Ortiz de Mendibil, I., J. L. Vizmanos and F. J. Novo (2009). "Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer." PLoS One **4**(3): e4805.

Ozawa, T., M. Riester, Y. Cheng, J. T. Huse, M. Squatrito, K. Helmy, N. Charles, F. Michor and E. C. Holland (2014). "Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like precursor glioma." Cancer Cell **26**(2): 288-300.

Palucka, K. and J. Banchereau (2012). "Cancer immunotherapy via dendritic cells." Nat Rev Cancer **12**(4): 265-277.

Parker, B. C. and W. Zhang (2013). "Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment." Chin J Cancer **32**(11): 594-603.

Parker, H., M. J. Rose-Zerilli, M. Larrayoz, R. Clifford, J. Edelmann, S. Blakemore, J. Gibson, J. Wang, V. Ljungstrom, T. K. Wojdacz, T. Chaplin, A. Roghanian, Z. Davis, A. Parker, E. Tausch, S. Ntoufa, S. Ramos, P. Robbe, R. Alsolami, A. J. Steele, G. Packham, A. E. Rodriguez-Vicente, L. Brown, F. McNicholl, F. Forconi, A. Pettitt, P. Hillmen, M. Dyer, M. S. Cragg, C. Chelala, C. C. Oakes, R. Rosenquist, K. Stamatopoulos, S. Stilgenbauer, S. Knight, A. Schuh, D. G. Oscier and J. C. Strefford (2016). "Genomic disruption of the histone methyltransferase SETD2 in chronic lymphocytic leukaemia." Leukemia.

Patel, A. P., I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suva, A. Regev and B. E. Bernstein (2014). "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma." Science **344**(6190): 1396-1401.

Phillips, H. S., S. Kharbanda, R. Chen, W. F. Forrest, R. H. Soriano, T. D. Wu, A. Misra, J. M. Nigro, H. Colman, L. Soroceanu, P. M. Williams, Z. Modrusan, B. G. Feuerstein and K. Aldape (2006). "Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis." Cancer Cell **9**(3): 157-173.

Prat, A., E. Pineda, B. Adamo, P. Galvan, A. Fernandez, L. Gaba, M. Diez, M. Viladot, A. Arance and M. Munoz (2015). "Clinical implications of the intrinsic molecular subtypes of breast cancer." Breast **24 Suppl 2**: S26-35.

Prins, R. M., H. Soto, V. Konkankit, S. K. Odesa, A. Eskin, W. H. Yong, S. F. Nelson and L. M. Liau (2011). "Gene expression profile correlates with T-cell infiltration and relative survival in glioblastoma patients vaccinated with dendritic cell immunotherapy." Clin Cancer Res **17**(6): 1603-1615.

Pyonteck, S. M., L. Akkari, A. J. Schuhmacher, R. L. Bowman, L. Sevenich, D. F. Quail, O. C. Olson, M. L. Quick, J. T. Huse, V. Teijeiro, M. Setty, C. S. Leslie, Y. Oei, A. Pedraza, J. Zhang, C. W. Brennan, J. C. Sutton, E. C. Holland, D. Daniel and J. A. Joyce (2013). "CSF-1R inhibition alters macrophage polarization and blocks glioma progression." Nat Med **19**(10): 1264-1272.

Qian, Y., S. Chai, Z. Liang, Y. Wang, Y. Zhou, X. Xu, C. Zhang, M. Zhang, J. Si, F. Huang, Z. Huang, W. Hong and K. Wang (2014). "KIF5B-RET fusion kinase promotes cell growth by multilevel activation of STAT3 in lung cancer." Mol Cancer **13**: 176.

Qin, F., Z. Song, M. Babiceanu, Y. Song, L. Facemire, R. Singh, M. Adli and H. Li (2015). "Discovery of CTCF-sensitive Cis-spliced fusion RNAs between adjacent genes in human prostate cells." PLoS Genet **11**(2): e1005001.

Quail, D. F. and J. A. Joyce (2013). "Microenvironmental regulation of tumor progression and metastasis." Nat Med **19**(11): 1423-1437.

Raeisi Shahraki, H., S. Pourahmad and S. M. Ayatollahi (2016). "Identifying the Prognosis Factors in Death after Liver Transplantation via Adaptive LASSO in Iran." J Environ Public Health **2016**: 7620157.

Raggi, C., P. Invernizzi and J. B. Andersen (2015). "Impact of microenvironment and stem-like plasticity in cholangiocarcinoma: molecular networks and biological concepts." J Hepatol **62**(1): 198-207.

Ragunathan, N., J. Dairou, B. Pluvineau, M. Martins, E. Petit, N. Janel, J. M. Dupret and F. Rodrigues-Lima (2008). "Identification of the xenobiotic-metabolizing enzyme arylamine N-acetyltransferase 1 as a new target of cisplatin in breast cancer cells: molecular and cellular mechanisms of inhibition." Mol Pharmacol **73**(6): 1761-1768.

Razavi, S. M., K. E. Lee, B. E. Jin, P. S. Aujla, S. Gholamin and G. Li (2016). "Immune Evasion Strategies of Glioblastoma." Front Surg **3**: 11.

Remark, R., M. Alifano, I. Cremer, A. Lupo, M. C. Dieu-Nosjean, M. Riquet, L. Crozet, H. Ouakrim, J. Goc, A. Cazes, J. F. Flejou, L. Gibault, V. Verkarre, J. F. Regnard, O. N. Pages, S. Oudard, B. Mlecnik, C. Sautes-Fridman, W. H. Fridman and D. Damotte (2013). "Characteristics and clinical impacts of the immune environments in colorectal and renal cell carcinoma lung metastases: influence of tumor origin." Clin Cancer Res **19**(15): 4079-4091.

Ries, C. H., M. A. Cannarile, S. Hoves, J. Benz, K. Wartha, V. Runza, F. Rey-Giraud, L. P. Pradel, F. Feuerhake, I. Klamann, T. Jones, U. Jucknischke, S. Scheiblich, K. Kaluza, I. H. Gorr, A. Walz, K. Abiraj, P. A. Cassier, A. Sica, C. Gomez-Roca, K. E. de Visser, A. Italiano, C. Le Tourneau, J. P. Delord, H. Levitsky, J. Y. Blay and D. Ruttinger (2014). "Targeting tumor-associated macrophages with anti-CSF-1R antibody reveals a strategy for cancer therapy." Cancer Cell **25**(6): 846-859.

Rizvi, N. A., M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmir, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher and T. A. Chan (2015). "Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer." Science **348**(6230): 124-128.

Roukos, D. H. (2010). "Novel clinico-genome network modeling for revolutionizing genotype-phenotype-based personalized cancer care." Expert Rev Mol Diagn **10**(1): 33-48.

Roukos, D. H. (2011). "Cancer genome explosion and systems biology: impact on surgical oncology?" Ann Surg Oncol **18**(1): 12-15.

Roychowdhury, S. and A. M. Chinnaiyan (2016). "Translating cancer genomes and transcriptomes for precision oncology." CA Cancer J Clin **66**(1): 75-88.

Ruffell, B. and L. M. Coussens (2015). "Macrophages and therapeutic resistance in cancer." Cancer Cell **27**(4): 462-472.

Samantarrai, D., S. Dash, B. Chhetri and B. Mallick (2013). "Genomic and epigenomic cross-talks in the regulatory landscape of miRNAs in breast cancer." Mol Cancer Res **11**(4): 315-328.

Sandberg, R. and O. Larsson (2007). "Improved precision and accuracy for microarrays using updated probe set definitions." BMC Bioinformatics **8**: 48.

Sasaki, T., S. J. Rodig, L. R. Chirieac and P. A. Janne (2010). "The biology and treatment of EML4-ALK non-small cell lung cancer." Eur J Cancer **46**(10): 1773-1780.

Schroder, M. S., A. C. Culhane, J. Quackenbush and B. Haibe-Kains (2011). "survcomp: an R/Bioconductor package for performance assessment and comparison of survival models." Bioinformatics **27**(22): 3206-3208.

Schumacher, T. N. and R. D. Schreiber (2015). "Neoantigens in cancer immunotherapy." Science **348**(6230): 69-74.

Schwaederle, M., M. Zhao, J. J. Lee, A. M. Eggermont, R. L. Schilsky, J. Mendelsohn, V. Lazar and R. Kurzrock (2015). "Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials." J Clin Oncol **33**(32): 3817-3825.

Shaw, A. T., P. P. Hsu, M. M. Awad and J. A. Engelman (2013). "Tyrosine kinase gene rearrangements in epithelial malignancies." Nat Rev Cancer **13**(11): 772-787.

Shugay, M., I. Ortiz de Mendibil, J. L. Vizmanos and F. J. Novo (2012). "Genomic hallmarks of genes involved in chromosomal translocations in hematological cancer." PLoS Comput Biol **8**(12): e1002797.

Siegal, T. (2015). "Clinical impact of molecular biomarkers in gliomas." J Clin Neurosci **22**(3): 437-444.

Storm, E. E., S. Durinck, F. de Sousa e Melo, J. Tremayne, N. Kljavin, C. Tan, X. Ye, C. Chiu, T. Pham, J. A. Hongo, T. Bainbridge, R. Firestein, E. Blackwood, C. Metcalfe, E. W. Stawiski, R. L. Yauch, Y. Wu and F. J. de Sauvage (2016). "Targeting PTPRK-RSPO3 colon tumours promotes differentiation and loss of stem-cell function." Nature **529**(7584): 97-100.

Stransky, N., E. Cerami, S. Schalm, J. L. Kim and C. Lengauer (2014). "The landscape of kinase fusions in cancer." Nat Commun **5**: 4846.

Sturm, D., S. Bender, D. T. Jones, P. Lichter, J. Grill, O. Becher, C. Hawkins, J. Majewski, C. Jones, J. F. Costello, A. Iavarone, K. Aldape, C. W. Brennan, N. Jabado and S. M. Pfister (2014). "Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge." Nat Rev Cancer **14**(2): 92-107.

Sturm, D., H. Witt, V. Hovestadt, D. A. Khuong-Quang, D. T. Jones, C. Konermann, E. Pfaff, M. Tonjes, M. Sill, S. Bender, M. Kool, M. Zapatka, N. Becker, M. Zucknick, T. Hielscher, X. Y. Liu, A. M. Fontebasso, M. Ryzhova, S. Albrecht, K. Jacob, M. Wolter, M. Ebinger, M. U. Schuhmann, T. van Meter, M. C. Fruhwald, H. Hauch, A. Pekrun, B. Radlwimmer, T. Niehues, G. von Komorowski, M. Durken, A. E. Kulozik, J. Madden, A. Donson, N. K. Foreman, R. Drissi, M. Fouladi, W. Scheurlen, A. von Deimling, C. Monoranu, W. Roggendorf, C. Herold-Mende, A. Unterberg, C. M. Kramm, J. Felsberg, C. Hartmann, B. Wiestler, W. Wick, T. Milde, O. Witt, A. M. Lindroth, J. Schwartzentruber, D. Faury, A. Fleming, M. Zakrzewska, P. P. Liberski, K. Zakrzewski, P. Hauser, M. Garami, A. Klekner, L. Bognar, S. Morrissy, F. Cavalli, M. D. Taylor, P. van Sluis, J. Koster, R. Versteeg, R. Volckmann, T. Mikkelsen, K. Aldape, G. Reifenberger, V. P. Collins, J. Majewski, A. Korshunov, P. Lichter, C. Plass, N. Jabado and S. M. Pfister (2012). "Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma." Cancer Cell **22**(4): 425-437.

Sun, Y. (2016). "Tumor microenvironment and cancer therapy resistance." Cancer Lett **380**(1): 205-215.

Suzuki, M., H. Makinoshima, S. Matsumoto, A. Suzuki, S. Mimaki, K. Matsushima, K. Yoh, K. Goto, Y. Suzuki, G. Ishii, A. Ochiai, K. Tsuta, T. Shibata, T. Kohno, H. Esumi and K. Tsuchihara (2013). "Identification of a lung adenocarcinoma cell line with CCDC6-RET fusion gene and the effect of RET inhibitors in vitro and in vivo." Cancer Sci **104**(7): 896-903.

Torres-Garcia, W., S. Zheng, A. Sivachenko, R. Vegesna, Q. Wang, R. Yao, M. F. Berger, J. N. Weinstein, G. Getz and R. G. Verhaak (2014). "PRADA: pipeline for RNA sequencing data analysis." Bioinformatics **30**(15): 2224-2226.

Van Allen, E. M., D. Miao, B. Schilling, S. A. Shukla, C. Blank, L. Zimmer, A. Sucker, U. Hillen, M. H. Geukes Foppen, S. M. Goldinger, J. Utikal, J. C. Hassel, B. Weide, K. C. Kaehler, C. Loquai, P. Mohr, R. Gutzmer, R. Dummer, S. Gabriel, C. J. Wu, D. Schadendorf and L. A. Garraway (2015). "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma." Science **350**(6257): 207-211.

van den Bent, M. J., A. A. Brandes, M. J. Taphoorn, J. M. Kros, M. C. Kouwenhoven, J. Y. Delattre, H. J. Bernsen, M. Frenay, C. C. Tijssen, W. Grisold, L. Sipos, R. H. Enting, P. J. French, W. N. Dinjens, C. J. Vecht, A. Allgeier, D. Lacombe, T. Gorlia and K. Hoang-Xuan (2013). "Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951." J Clin Oncol **31**(3): 344-350.

Varley, K. E., J. Gertz, B. S. Roberts, N. S. Davis, K. M. Bowling, M. K. Kirby, A. S. Nesmith, P. G. Oliver, W. E. Grizzle, A. Forero, D. J. Buchsbaum, A. F. LoBuglio and R. M. Myers (2014). "Recurrent read-through fusion transcripts in breast cancer." Breast Cancer Res Treat **146**(2): 287-297.

Velghe, A. I., S. Van Cauwenberghe, A. A. Polyansky, D. Chand, C. P. Montano-Almendras, S. Charni, B. Hallberg, A. Essaghir and J. B. Demoulin (2014). "PDGFRA alterations in cancer: characterization of a gain-of-function V536E transmembrane mutant as well as loss-of-function and passenger mutations." Oncogene **33**(20): 2568-2576.

Velusamy, T., N. Palanisamy, S. Kalyana-Sundaram, A. A. Sahasrabudhe, C. A. Maher, D. R. Robinson, D. W. Bahler, T. T. Cornell, T. E. Wilson, M. S. Lim, A. M. Chinnaiyan and K. S. Elenitoba-Johnson (2013). "Recurrent reciprocal RNA chimera involving YPEL5 and PPP1CB in chronic lymphocytic leukemia." Proc Natl Acad Sci U S A **110**(8): 3035-3040.

Verhaak, C. M., A. M. Lintsen, A. W. Evers and D. D. Braat (2010). "Who is at risk of emotional problems and how do you know? Screening of women going for IVF treatment." Hum Reprod **25**(5): 1234-1240.

Verhaak, R. G., K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. Hayes and N. Cancer Genome Atlas Research (2010). "Integrated genomic analysis identifies

clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1." Cancer Cell **17**(1): 98-110.

Verhaak, R. G., P. Tamayo, J. Y. Yang, D. Hubbard, H. Zhang, C. J. Creighton, S. Fereday, M. Lawrence, S. L. Carter, C. H. Mermel, A. D. Kostic, D. Etemadmoghadam, G. Saksena, K. Cibulskis, S. Duraisamy, K. Levanon, C. Sougnez, A. Tsherniak, S. Gomez, R. Onofrio, S. Gabriel, L. Chin, N. Zhang, P. T. Spellman, Y. Zhang, R. Akbani, K. A. Hoadley, A. Kahn, M. Kobel, D. Huntsman, R. A. Soslow, A. Defazio, M. J. Birrer, J. W. Gray, J. N. Weinstein, D. D. Bowtell, R. Drapkin, J. P. Mesirov, G. Getz, D. A. Levine and M. Meyerson (2013). "Prognostically relevant gene signatures of high-grade serous ovarian carcinoma." J Clin Invest **123**(1): 517-525.

Vitucci, M., D. N. Hayes and C. R. Miller (2011). "Gene expression profiling of gliomas: merging genomic and histopathological classification for personalised therapy." Br J Cancer **104**(4): 545-553.

Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, Jr. and K. W. Kinzler (2013). "Cancer genome landscapes." Science **339**(6127): 1546-1558.

Wallace, J. A., J. R. Pitarresi, N. Sharma, M. Palettas, M. C. Cuitino, S. T. Sizemore, L. Yu, A. Sanderlin, T. J. Rosol, K. D. Mehta, G. M. Sizemore and M. C. Ostrowski (2014). "Protein kinase C Beta in the tumor microenvironment promotes mammary tumorigenesis." Front Oncol **4**: 87.

Wang, J. and F. Xue (2015). "[Group Lasso Penalized Classifier for Diagnosis of Diseases with Categorical Data]." Sheng Wu Yi Xue Gong Cheng Xue Za Zhi **32**(5): 965-969.

Watson, I. R., K. Takahashi, P. A. Futreal and L. Chin (2013). "Emerging patterns of somatic mutations in cancer." Nat Rev Genet **14**(10): 703-718.

Weller, M., M. van den Bent, K. Hopkins, J. C. Tonn, R. Stupp, A. Falini, E. Cohen-Jonathan-Moyal, D. Frappaz, R. Henriksson, C. Balana, O. Chinot, Z. Ram, G. Reifenberger, R. Soffietti, W. Wick and G. European Association for Neuro-Oncology Task Force on Malignant (2014). "EANO guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma." Lancet Oncol **15**(9): e395-403.

Weller, M., R. G. Weber, E. Willscher, V. Riehmer, B. Hentschel, M. Kreuz, J. Felsberg, U. Beyer, H. Löffler-Wirth, K. Kaulich, J. P. Steinbach, C. Hartmann, D. Gramatzki, J. Schramm, M. Westphal, G. Schackert, M. Simon, T. Martens, J. Bostrom, C. Hagel, M. Sabel, D. Krex, J. C. Tonn, W. Wick, S. Noell, U. Schlegel, B. Radlwimmer, T. Pietsch, M. Loeffler, A. von Deimling, H. Binder and G. Reifenberger (2015). "Molecular classification of diffuse cerebral WHO grade II/III gliomas using genome- and transcriptome-wide profiling improves stratification of prognostically distinct patient groups." Acta Neuropathol **129**(5): 679-693.

Williams, S. V., C. D. Hurst and M. A. Knowles (2013). "Oncogenic FGFR3 gene fusions in bladder cancer." Hum Mol Genet **22**(4): 795-803.

Wu, C. C., K. Kannan, S. Lin, L. Yen and A. Milosavljevic (2013). "Identification of cancer fusion drivers using network fusion centrality." Bioinformatics **29**(9): 1174-1181.

Wyatt, A. W., F. Mo, K. Wang, B. McConeghy, S. Brahmbhatt, L. Jong, D. M. Mitchell, R. L. Johnston, A. Haegert, E. Li, J. Liew, J. Yeung, R. Shrestha, A. V. Lapuk, A. McPherson, R. Shukin, R. H. Bell, S. Anderson, J. Bishop, A. Hurtado-Coll, H. Xiao, A. M. Chinnaiyan, R. Mehra, D. Lin, Y. Wang, L. Fazli, M. E. Gleave, S. V. Volik and C. C. Collins (2014). "Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer." Genome Biol **15**(8): 426.

Yan, J., L. Y. Kong, J. Hu, K. Gabrusiewicz, D. Dibra, X. Xia, A. B. Heimberger and S. Li (2015). "FGL2 as a Multimodality Regulator of Tumor-Mediated Immune Suppression and Therapeutic Target in Gliomas." J Natl Cancer Inst **107**(8).

Yan, W., W. Zhang, G. You, J. Zhang, L. Han, Z. Bao, Y. Wang, Y. Liu, C. Jiang, C. Kang, Y. You and T. Jiang (2012). "Molecular classification of gliomas based on whole genome gene expression: a systematic report of 225 samples from the Chinese Glioma Cooperative Group." Neuro Oncol **14**(12): 1432-1440.

Ye, X. Z., S. L. Xu, Y. H. Xin, S. C. Yu, Y. F. Ping, L. Chen, H. L. Xiao, B. Wang, L. Yi, Q. L. Wang, X. F. Jiang, L. Yang, P. Zhang, C. Qian, Y. H. Cui, X. Zhang and X. W. Bian (2012). "Tumor-associated microglia/macrophages enhance the invasion of glioma stem-like cells via TGF-beta1 signaling pathway." J Immunol **189**(1): 444-453.

Yoshihara, K., M. Shahmoradgoli, E. Martinez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Trevino, H. Shen, P. W. Laird, D. A. Levine, S. L. Carter, G. Getz, K. Stemke-Hale, G. B. Mills and R. G. Verhaak (2013). "Inferring tumour purity and stromal and immune cell admixture from expression data." Nat Commun **4**: 2612.

Yoshihara, K., Q. Wang, W. Torres-Garcia, S. Zheng, R. Vegesna, H. Kim and R. G. Verhaak (2015). "The landscape and therapeutic relevance of cancer-associated transcript fusions." Oncogene **34**(37): 4845-4854.

Zadeh, G., O. H. Khan, M. Vogelbaum and D. Schiff (2015). "Much debated controversies of diffuse low-grade gliomas." Neuro Oncol **17**(3): 323-326.

Zeng, J., A. P. See, J. Phallen, C. M. Jackson, Z. Belcaid, J. Ruzevick, N. Durham, C. Meyer, T. J. Harris, E. Albesiano, G. Pradilla, E. Ford, J. Wong, H. J. Hammers, D. Mathios, B. Tyler, H. Brem, P. T. Tran, D. Pardoll, C. G. Drake and M. Lim (2013). "Anti-PD-1 blockade and stereotactic radiation produce long-term survival in mice with intracranial gliomas." Int J Radiat Oncol Biol Phys **86**(2): 343-349.

Zhang, Y., M. Gong, H. Yuan, H. G. Park, H. F. Frierson and H. Li (2012). "Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation." Cancer Discov **2**(7): 598-607.

Zheng, S., M. G. Chheda and R. G. Verhaak (2012). "Studying a complex tumor: potential and pitfalls." Cancer J **18**(1): 107-114.

Zhou, J. X., H. Yang, Q. Deng, X. Gu, P. He, Y. Lin, M. Zhao, J. Jiang, H. Chen, Y. Lin, W. Yin, L. Mo and J. He (2013). "Oncogenic driver mutations in patients with non-small-cell lung cancer at various clinical stages." Ann Oncol **24**(5): 1319-1325.

VITA

Xin Hu, the daughter of Shirong Hu and Kui Ye, was born on January 24th, 1976 in Qingdao, Shandong province, China. She graduated from Qingdao No.2 high school in China in June 1994. After that, she entered Shanghai University, majored in Biochemistry Engineering, and earned her bachelor degree in June 1998. She continued to pursue a master's degree of Microbiology in Shanghai Normal University and obtained her master's degree in June 2002. Then she joined Chinese National Human Genomic Center at Shanghai as a research associate after her internship in July 2002. Following January, 2005, she traveled to Houston, Texas to pursue her research internship in MD Anderson Cancer Center. Then she transferred to Baylor College of Medicine, working as a research assistant until August, 2009, when she was enrolled at The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences (GSBS), pursuing her Ph.D. degree in the field of epigeomics and genomics, mentored by Dr. Jean-Pierre Issa. Eventually, she conducted her thesis research under the supervision of Dr. Roel Verhaak in the department of Genomic Medicine and the department of Bioinformatics and Computational Biology, MD Anderson Cancer Center and received her Ph.D. degree in Bioinformatics, Biostatistics and Systems Biology in spring, 2017.

Permanent address:

Unit 702, 81 Dengzhou Road, Shibei District,
Qingdao, Shandong, China, 266012