Computational Inferences of Mutations Driving Mesenchymal Differentiation in Glioblastoma

James Chen

Submitted in partial fulfillment of the requirements for the Doctor of Philosophy Degree in the Graduate School of Arts and Sciences

Columbia University
2013

ABSTRACT

Computational Inferences of Mutations Driving Mesenchymal Differentiation in Glioblastoma

James Chen

This dissertation reviews the development and implementation of integrative, systems biology methods designed to parse driver mutations from high-throughput array data derived from human patients. The analysis of vast amounts of genomic and genetic data in the context of complex human genetic diseases such as Glioblastoma is a daunting task. Mutations exist by the hundreds, if not thousands, and only an unknown handful will contribute to the disease in a significant way. The goal of this project was to develop novel computational methods to identify candidate mutations from these data that drive the molecular differentiation of glioblastoma into the mesenchymal subtype, the most aggressive, poorest-prognosis tumors associated with glioblastoma.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank the Califano Lab and my mentor, Andrea Califano, for their intellectual and motivational support during my stay in their lab. I would also like to thank the Dept. of Genetics and Development for their ongoing support of my work, to Angela Christiano for introducing me to the bioinformatic wing of Columbia University, and to Michael Shen for his ongoing collaborations with my research and me. I would also like to thank Michael Berens of TGEN, AZ and his group for their welcome collaborative efforts. Finally, many thanks to my family and friends.

CHAPTER 1 – Background of Glioblastoma and Systems Biology Approaches

Glioblastoma (GBM), the most prevalent brain cancer found in humans[30]][35], is an undoubtedly complex genetic disease. Along with other fields of cancer study, recent research is beginning to reveal the extent of physiological variability in tumors affecting different patients. Although GBM tumors share traits common enough to be classified under the same umbrella term (such as basic histological markers), there is significant variability in other traits associated with the patient's prognosis; the aggressiveness of the tumor, its metastasis, and its resistance to therapeutics are not uniform in GBM. At both genomic and genetic levels, GBM tumors are as heterogeneous as their physiology suggests[3][5][32][42]. A GBM tumor does not develop solely with the activation of oncogenic driver[48]; although this is a necessary and essential step, tumors exhibit the concomitant, large-scale genomic alterations that result in GBM tumors differentiating into three molecularly distinct subtypes: the Mesenchymal, Proneural, and Proliferative subtypes. The Mesenchymal subtype earmarks specific physiological behaviors of the tumor that exist independently of oncogenic processes and uniquely affect the prognosis associated with the disease: mesenchymal GBM are the most aggressively growing, poorest prognosis GBM tumors[31][35].

This finding corroborates the notion that GBM, and cancers by extension, are not diseases resulting solely from the activation of oncogenic processes. There are other physiological, developmental, and molecular processes that are regulated independently of oncogenesis, yet contribute significantly to the overall behavior of the disease. It is equally important to the understanding of GBM biology to characterize the genetic regulation of these other processes, and how their co-occurrence with oncogenic mutations can alter the progression of GBM. This is not straightforward when studying a disease characterized by hundreds, if not thousands, of genomic mutations co-occurring in a single patient. Mutations occur in different patterns that can render every single GBM patient unique from another[71].

The goal of this work was to develop a computational framework that integrates several sources of high-throughput data and systems biology approaches to predict the driver mutations that induce Mesenchymal differentiation in Glioblastoma. Molecular perturbations that result in mesenchymal differentiation should correlate with the genomic mutations that are responsible for their aberrant expression, and these perturbations can be used to reverse engineer the genomic-genetic integration.

Glioblastoma and the Mesenchymal Subtype: a Molecular Perspective

Glioblastoma (GBM) is a subcategory of high-grade gliomas, the most common type of brain tumors found in humans. Over 50% of brain tumors in functional brain tissue of human patients are classified into this category[35]. There is a slightly higher incidence in males, and the average age of onset is >50[31]. Brain cancers are pathologically identified as GBM based primarily on the presence of necrotic or necrotizing tissue at the core of the tumor, surrounded by anaplastic (un- or de-differentiated cells) and typically an extensive vasculature[2][23][57].

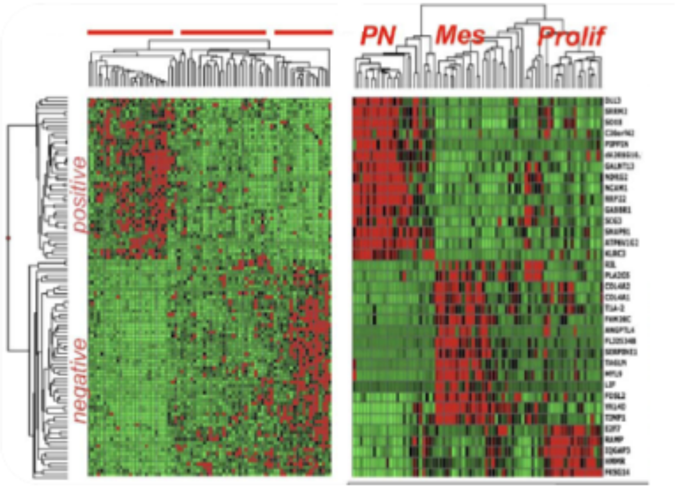
These cancers are virtually incurable and highly aggressive. The average post-diagnosis survival of patients is projected as twelve months with treatment, and typically less than four months without treatment[46]. They are extremely difficult to detect, and are typically only identified later in the cancer progression due to the presentation of secondary symptoms: the tumor develops large enough in size to cause increased intracranial pressure, impairment of neural and cognitive function, and other generalized symptoms such as chronic headaches, nausea, etc. Furthermore, GBM has displayed a robust resistance to standard cancer therapeutics in addition to the naturally dangerous risks of treating diseases in the brain[46] [51] [53]. Transporting drugs across the blood brain barrier has always been a difficult medical procedure, and tumor resections in the brain carry the inherent risk of causing irrevocable secondary damage.

At a genetic level, GBM is a disease characterized by significant *de novo* somatic chromosomal copy changes and rearrangements, and there are subsequently no known hereditary risk factors[22]. There are also no significant associations of brain cancer with specific environmental factors[46]. These tumors typically present with genomic alterations in “classic” oncogenic drivers and tumor suppressors: amplifications of EGFR[63][68][71] or losses of p16[24] are found in over 60% of patients, and typically in combination[67][68]. Despite the significant mutation rate of oncogenic genes, the variable rates of these mutations combined with extensive mutations throughout the genome establish GBM as a highly heterogeneous genetic disease. While a large number of patients may bear mutations at the EGFR or p16 loci, each patient exhibits evidence for chromosomal alterations affecting hundreds to thousands of other genes in unique and unpredictable combinations. This results in a genetically and physiologically diverse cancer, which, in combination with the resistance of GBM to conventional therapeutics, has led the field to begin refining our definitions of GBM in order to improve our understanding of the mechanics and etiology of this complex genetic disease.

In the past, the study of cancer in general focused primarily on the identification of genetic variants and genomic loci that predispose or induce tumor formation, such as EGFR and p16. Genes whose increased expression leads to tumorigenesis are called *oncogenes*, and genes whose loss induces tumorigenesis*, tumor suppressors*. Tumor suppressors and oncogenes are typically genes participating in key developmental pathways, such as the Notch and Wnt signaling cascades[47][67][70], or metabolic and cellular proliferation pathways[79]. Aberrant activation or re-activation of these developmental pathways trigger unchecked, accelerated proliferation that could lead to formation of tumor masses, differentiation or de-differentiation that could result in metastasis, and the co-opting of “normal” biological processes such as angiogenesis to provide nutrients for the growing mass.

These genetic elements are typically insufficient individually to induce the full development of cancer, but subsequent studies identified that a series of accumulated genetic abnormalities could push cells into tumorigenic behavior. Cancer has been shown, along these lines, to be a complex genetic disease requiring the input of multiple genetic processes to fully develop into the unique pathogenic tumors and metastases. In cancers like GBM, this coincides with the diverse genomic mutation architecture observed in patients. However, prior to the development of high-throughput genetic and genomic screening, candidate-based approaches have primarily been used to make sense of how individual mutations and biological processes contribute to the disease.

The availability and cost-effectiveness of gene expression profiling has increased significantly in recent years, allowing for new means of defining cancers such as GBM at the molecular level. Numerous studies have been conducted that indicate that cancers that have been presumed to be a single disease actually segregate into distinct molecular subtypes[32][42][52][59]. While a great deal of research has been focused on understanding the oncogenic drivers and oncogenic and angiogenic properties of



Phillips et al, 2006

Figure legend: Unsupervised hierarchical clustering of genes whose expression correlated with patient prognosis in GBM revealed three distinct, mutually exclusive molecular subtypes. The Mesenchymal, Proneural, and Proliferative subtypes were coined after the expression of marker genes that canonically define mesenchyme, proneural tissue, and cellular proliferation processes, respectively. The mesenchymal expression profile in particular was associated specifically with poorest patient prognosis, and the proneural signature with the best.

GBM, it has also become apparent that the heterogeneity of GBM has ramifications that extend far beyond the “classic” tumorigenic pathways. The umbrella classification of Glioblastoma itself contains an amalgamation of molecularly and physiologically distinct brain tumors that bear extensive, yet unique, profiles of mutated genes, and equally diverse distributions of unique patterns of gene expression. Research on the unique expression patterns of various GBM samples has revealed that brain tumors falling under the umbrella classification of “Glioblastoma” cosegregate into three distinct subtypes, which were coined the *Mesenchymal* (MES), *Proneural* (PN), and *Proliferative* (PRO) subtypes[52]. Each subtype was named and defined based upon the expression of gene markers primarily expressed in mesenchyme (YKL40, FN1), in tissues of neural and proneural origin (OLIG2, BCAN), and upon activation of cellular proliferation and angiogenesis (TOP2A, PCNA), respectively. Phillips *et al.* reported that the expression of these three marker panels is mutually exclusive: MES samples both express MES markers and show suppression of PN and PRO markers. Moreover, GBM subtypes do not show any significant correlation with classical oncogenic mutations.

Further analysis revealed that tumors expressing the MES expression signature are at maximal risk of being highly aggressive, even more so than other GBM tumors, and predictive of the poorest prognosis in the overall patient cohort. It was hypothesized that the expression of MES markers contributes to the alteration of the tumor’s biology, rendering it highly aggressive and resistant to treatment, though the particular mechanics behind the contribution to poor prognosis are unknown. It was proposed that the MES expression pattern is a quantitative molecular predictor of the poorest prognosis GBM that warranted further study. This profile could also be used as a biological readout for poor prognosis tumors in subsequent experiments instead of relying on more noisy, indirect, and qualitative measures such as World Health Organization grade and prognosis.

Studies such as these have allowed the field to model GBM as a molecular disease, rather than a “physiological” one. “GBM tumors” that exhibit behaviors across a general spectrum can now be broken down into a series of discrete expression profiles. Each set of genes in an expression profile could provide insight into the precise genetic pathways that are required to develop a specific subtype of GBM. Even more specifically, it was shown that distinct subtypes of cancers exist beyond what could medically be distinguished by histology and other clinical assays. These molecular subtypes also provided an explanation as to why different patients responded drastically differently to both the disease and treatments: the diseases being treated as the same were not the same. The focus of research for many shifted to understanding how to integrate this molecular information into modeling GBM for both an understanding of its etiology, and potential development of more focused, effective treatments.

Concurrently, The Cancer Genome Atlas (TCGA) had launched a massive initiative to gather and catalogue patient-matched clinical, gene expression, genomic, and DNA sequencing data on tumor samples obtained from patients around the world. The objective was to gather this large-scale data and make it publicly available to researches, providing an enormous sample pool of data that would be daunting and impractical, if not impossible, for any individual researcher to attempt to gather alone[67][68]. We were able to obtain gene expression, gene copy number, and clinical annotation data on 252 GBM tumors from human patients to complement the work of Phillips *et al*. This amalgamation of data provided a unique opportunity and resource with sufficient power to study the molecular and genetic regulation of the differentiation of these three subtypes of GBM with systems biology approaches geared towards understanding the regulation of large-scale molecular phenotypes.

Systems Biology Approaches and MES Master Regulators

In order to place my work into the context of a specific systems biology approach, it is important to understand how the Califano lab as a whole approaches problems such as MES subtype differentiation. We specifically sought to further understand the biology of the MES subtype GBM by attempting to identify the genetic regulators that drive the activation of the MES marker panel. The more traditional bioinformatic approach is to identify key expression patterns in the tumor and isolate developmental or metabolic pathways for further characterization and study[21][23]. When the expression of angiogenic regulators is discovered amidst a panel of differentially expressed genes when compared to “normal” control tissue, research is then geared towards specifically targeting aberrant angiogenesis[66]. While this has produced fruitful research, a fundamental limitation exists in the approach. It is not designed to elucidate novel, important processes or genetic regulators, and only accounts for a relatively arbitrary selection of candidate genes with already-known or implied functions in a cancer such as GBM. Furthermore, fundamental assumptions are made in candidate selection that are not necessarily true. It is assumed that there exists one gene, or relatively few genes, that cause the phenotype. It is assumed that these genes are differentially expressed at a statistically detectable level, and that all other genes that are differentially expressed are either unrelated or downstream consequences of the select few drivers. While any of these can be true, none can be assumed *a priori* when doing genome-wide, array-based studies for the selection of causal contributors to the disease.

Instead, the Califano lab employed a systems biology approach to understanding subtype differentiation in GBM. Rather than selecting candidate genes from a panel of differentially expressed genes by their involvement with biological processes such as oncogenesis and angiogenesis, we used the entire set of differentially expressed genes as a phenotypic readout. This is a systemic approach used to identify the genetic regulators of an entire expression profile, including all of the biological processes that are altered, to account for the tumor’s unique behavior in its entirety. This was accomplished by implementing the ARACNe algorithm in the TCGA GBM dataset, a method of reverse-engineering context-specific transcriptional networks[3][10][11].

The ARACNe algorithm developed by the Califano lab uses Mutual Information, a measure of probabilistic dependency between variables that is capable of measuring non-linear and non-monotonic correlations, and an extension of Data Processing Inequality (DPI), to infer genetic regulatory interactions directly from data with human origins[39][40]. Targets of transcription factors are identified by finding genes whose expression has a high degree of correlation via mutual information with the expression of a transcription factor. The non-transcription factor genes are presumed to be regulatory targets of the transcription factor, based on the hypothesis that transcription factors are more likely to regulate multiple targets than that targets are regulated by large numbers of transcription factors. The DPI is then used to systematically eliminate likely indirect regulatory targets to create transcriptional "regulons," or sets of genes that are specific targets of each transcription factor[39]. Using this methodology we have been able to reverse engineer context-specific regulatory networks with up to 70% validation in contexts such as B-cell lymphoma and GBM[3][11]. This result is essentially a transcriptional map of all genes expressed in the analyzed context, complete with all transcription factors that are expressed in the tissue and the predicted gene targets that they specifically, and directly, regulate.

The transcriptional network can subsequently be interrogated to identify “Master Regulators,” or MRs, which are transcription factors whose regulons are enriched in a molecular gene expression signature that corresponds to a phenotype of interest. In doing so, we essentially identify the fewest number of transcription factors that would be necessary to specifically recapitulate the observed molecular expression phenotype. Interrogating the GBM network for Master Regulators of the mesenchymal GBM gene expression signature identified three transcription factors: CEBPB, CEBPD, and STAT3[11]. Our results indicated that over 70% of the MES gene expression signature panel defined by Phillips *et al.* could be activated by expressing a combination of only these three transcription factors.

The biological importance of these findings was established when we observed that shRNA-mediated co-silencing of CEBPB/D and STAT3 was sufficient to suppress the expression of mesenchymal markers in MES GBM tumor-derived cell lines that were intercranially injected into mice. Tumor cells with the co-silencing of these master regulators did not proliferate and they develop into solid masses, whereas individually silenced and un-silenced cells universally developed into tumors that were fatal to the mice. Furthermore, the protein expression of these three TFs in independent cohorts of human GBM patients stratified with the worst patient prognoses in a similar manner: tumors sections staining double-positive for CEBP and STAT3 proteins associated with significantly poorer prognoses than in patients that were negative or singly-positive for these proteins[11].

These results validated the computational inferences that predicted the CEBPs and STAT3 as master regulators specifically for the activation of the mesenchymal subtype behavior in GBM. It had become possible to computationally model a complex genetic disease as a gene expression signature and to predict the master regulators of that signature, thereby predicting the genetic regulators of the disease or phenotype of interest. The master regulator modules served as a molecular “bottleneck” through which all transcriptional and regulatory processes were integrated to produce a specific molecular effect. Any genetic event or change that directly or indirectly perturbed the behavior of the CEBPs and STAT3 would be predicted to induce mesenchymal transformation in GBM. This provided a novel way of approaching the study of genetic disease etiology: identifying or predicting the genes that regulate the behavior of master regulators allows for more targeted screening methods than using genome-wide approaches. In addition, traditional genomic approaches introduce confounding factors such as passenger mutations, mutations that contribute to unrelated disease behaviors, and they lose valuable power to the correction of large numbers of tested statistical hypotheses.

These results also provide support for the notion that clinically relevant physiological traits of GBM are dictated by more than the classical oncogenic signaling pathways. Whether by an addiction mechanism or otherwise, molecular programs activated in tandem with the classical oncogenic programs can contribute to the progression of disease. Equally important is the implication that diseases such as GBM can potentially be treated through more molecular avenues than the extensively studied oncogenic and angiogenic signaling cascades. Understanding the genetic events that drive the expression of these unique gene panels, and how they relate to the progression of the disease, is becoming as scientifically important as the genetic events that drive the formation of tumors, and evidence is mounting that the plethora of mutations that exist along with oncogenic mutations cannot be ignored for contribution to cancers such as GBM.

Genomic Mutations and eQTLs: a genomic perspective

My thesis began as a natural extension of this work, and to ask the question: **what actually happened in the genome of the patients to induce the activation of these master regulators?** These master regulators have never been specifically associated with classical oncogenic drivers in network or biological analyses. They have never even been extensively studied in the context of GBM, and their role was completely unknown prior to our application of ARACNe to the Phillips classification. In a complex disease characterized by extensive alteration of its genome, we hypothesized that other mutations originally considered as “background” or “passenger” alterations to the more prominent oncogenic drivers must have contributed to the differentiation of the distinct molecular subtypes, and we sought a method to identify and validate them in the context of mesenchymal GBM.

This idea serves as the conceptual basis for the now-ubiquitous genome-wide association studies. A genomic region demarcated by some combination of genetic markers, ranging from balancers to microsatellites to restriction length polymorphisms, is tested for association to some phenotype. This phenotype can be a disease or higher-level trait, or can be the expression of specific proteins and markers. For polygenic and complex traits, this typically results in multiple regions of the genome associating with the trait, typically including several genes; it is hypothesized that some combination of genes in these associated regions provide combinatorial contribution to the overall trait being observed. However, a common issue that stymies GWAS is the significant degree of identifiable mutations or SNPs that exist in any given individual. The sheer number of loci that must be tested for association across a cohort requires extensive correction for statistical testing in order to pare down the genomic data to an interpretable, testable set of candidates. Furthermore, diseases with diverse genetic causes prove difficult to parse because the association signals of each cause are diluted when comparing against a cohort of patients that include other causal genetic variants.

This issue can be addressed by integrating additional information into the analytic framework to shorten the list of candidate loci. The introduction and development of sophisticated gene expression studies has made integrating gene expression with gene mutation / genotype a promising choice. The concept of linking genetics (gene expression, expression patterns) and genomics (chromosomal and mutational data) is not a novel one, and is itself an extension of studying Quantitative Trail Loci. Mapping Quantitative Trait Loci (QTLs) in model organisms was a natural extension of linkage analyses tracing the association of linkage markers and polymorphisms associated with polygenic traits[36]. Genomic regions associated with genetic markers (SNPs or microsatellites, etc) were correlated with gene expression panels obtained while studying “quantitative traits,” or traits that do not have canonical Mendelian inheritance patterns. These traits were typically binary and discrete, and regulated by a minimal number of genes (yellow vs red, on vs off, etc).

An example of a quantitative trait is a person’s height. Human height exists across a continuous spectrum of measurements, and no individual “height” gene exists that is solely responsible for the regulation of how tall a person can/will develop. While there are undoubtedly environmental and developmental considerations to be taken for how tall or short an individual will be, height regardless tends to show distinct patterns of heritability. A QTL analysis applied to this issue would obtain the genotypes of a cohort of human subjects with varied height, and genomic loci whose genotype differences were predictive of height difference were identified in combination, based on statistical and computational methods. Genomic loci that maximized the prediction of height would subsequently be identified as QTLs, loci that contributed to the quantitative trait of height. This methodology was not geared towards identifying causal genomic loci via identification of specific mutations, but instead to identify by linkage and association the genomic loci that appeared to contribute to a quantitative state.

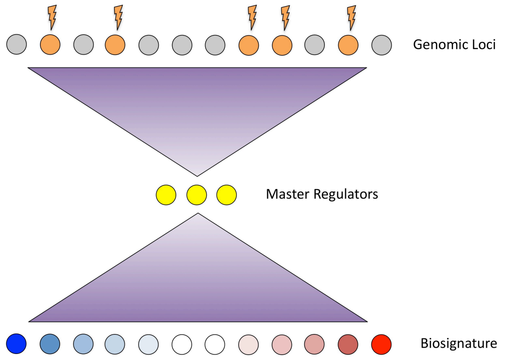
While this methodology was originally designed to analyze “organism-level” traits, or physiological traits that were directly observable, the advent of genetic and molecular profiling and the development of gene expression studies provided a molecular extension for this methodology, much in a manner similar to what was discussed for GBM and cancer previously. It had become possible to look at how genetic markers segregated with the transcriptional behavior of genes- to directly correlate gene expression to genotype, rather than looking at the effects of gene expression with reference to the genotype. A trait could be redefined as a product of the gene expression that drives the trait, which in turn was presumed to be regulated at a genomic level in a manner predicted by the genotype[29].

The development of the “Systems Biology” perspective was readily compatible with these approaches. With the increased understanding of transcriptional regulation in molecular systems, eQTLs could be modeled and identified by how they affect the behavior of major transcriptional regulators in both *cis*- and *trans*-regulatory interactions. Genetic variants that associated directly with the coding regions of major transcription regulators could be linked to the altered expression of the transcription factor itself and / or the behavior of its targets (*cis*-regulation) by adapting more traditional correlative and statistical metrics integrated with transcriptional assays and binding assays. Conversely, *trans*-regulatory elements were defined as genetic variants that, while not falling within the coding region of a regulator or gene directly, nonetheless directly associated with the altered behavior of the gene's expression. These variants could affect the behavior of other genes in a variety of ways. As an example, a variant could fall in a region of another gene that regulates the transcription or activity of the gene being, or alter the methylation state of the genomic region resulting in silencing or activation. In the context of transcriptional regulation, researchers began improving the detection of such genetic variants that altered gene activity by integrating additional information to maximize *a priori* knowledge of the predicted function of the genes being studied through several different methods.

One approach was to include transcription factor binding information. Transcription factors are proteins that regulate the expression of other genes by binding to DNA at a coding region to initiate transcription of mRNAs. Each transcription factor recognizes a distinct subset of genes and is able to activate (or repress) the transcription specifically of these targets. Multiple transcription factors can share the same target, but the overlap is variable. A core binding sequence defines the specificity of these regulatory regimes. This motif must be present in the promoter regions of a gene in order for a transcription factor to recognize, bind, and initiate transcription. Subsequently, any gene that is a target of a given transcription factor will bear a binding motif for that transcription factor in its promoter region. These motifs and their specific affinities can be computationally predicted using a combination of binding assays and thermodynamic models.

The availability of genomic sequencing allowed for the searching of promoter regions for these binding motifs, and to predict the binding activity based on the sequence similarities found in each specific promoter. Integrating this information with gene expression allows for an accurate measure of the activity of a transcription factor as a function of its transcriptional targets, identified by binding motifs. Differential activity across different genotypes could then be used to identify genomic regions that co-segregated with the differential activity of each transcription factor, defining genomic regions, or aQTLs, that associate with the differential activity of specific transcription factors[6] .

Identifying Driver Mutations by Integrating Master Regulators



Algorithms such as ARACNe and MINDy have identified molecular “bottlenecks,” small modules of master regulators and modulators that integrate signals required to activate the expression of gene expression profiles that define and potentially drive cancer subtype differentiation. We hypothesize that the genomic mutations that drive the differentiation, subsequently, must in some way interact with the master regulator bottleneck- a small number of master regulators regulate a large gene panel, and genomic mutations that drive the phenotype must interact with the master regulators.

Our reverse-engineered networks and master regulator analysis allow us to add an additional, pivotal dimension to these approaches: the ability to capitalize on our molecular “bottleneck” to measure the effect of genomic mutations on the transcriptional activity of master regulators that control a specific gene expression set. Identifying the mutations that control a gene expression set, by proxy, implies that the mutations control the phenotype associated with the expression set. Rather than asking, “What mutations associate with the MES phenotype,” I instead ask, “What mutations perturb the molecular behavior of the MES master regulators? What mutations are predicted to induce the expression of the MES phenotype?” These questions can be answered using an algorithm that incorporates transcriptional networks, genomic profiling, and gene expression profiling. I set out to develop this algorithm and perform an analysis on the TCGA dataset to identify candidate mutations that could drive the expression and differentiation of the Mesenchymal subtype, and biologically validate any subsequent results. Candidate mutations were expected to specifically perturb, directly or indirectly, the molecular behavior of the CEBP and STAT3 master regulators.

In my case, GBM was considered the ideal model to develop an algorithm to link genomic mutations to regulatory network perturbations (and subsequently gene expression profiles) due to the availability of patient-matched data through the TCGA. This matched data allowed us to classify GBM samples, reconstruct the GBM transcriptional network and interrogate for subtype master regulators, and finally attempt to integrate genomic information all in the same patient cohort to establish causality between mutations and molecular behaviors within patient samples.

This thesis details the development, implementation, and validation of my genetic genomic analytic framework in the following steps, broken down by chapter:

Chapter 2: Identifying mesenchymal cohorts in the TCGA dataset and selecting optimal parameters for downstream analysis.

Chapter 3: The definition of functional CNVs and parsing them from genomic and genetic data.

Chapter 4: Picking candidate driver mutations of the MES subtype using conditional association metrics

Chapter 5: A manuscript submitted to *Nature* on the identification and biological validation of the candidate MES driver, KLHL9

Chapter 5a: Additional work-in-progress addressing reviewer comments for the manuscript

Chapter 6: A discussion of the algorithm and biological results obtained in KLHL9

CHAPTER 2 – TCGA ARRAY PROCESSING AND SAMPLE CLASSIFICATION

At its core, studying the genomic drivers of differentiation to the MES GBM subtype requires the integration of several types of data: gene expression, genomic status (mutations, copy number, etc), molecular subtype classification, and regulatory networks. The functional genetic-genomics algorithm itself requires only CNV and gene expression data, but all of these types must be available to complete the downstream analyses associated with the entire workflow.

While the original subtypes defined by Phillips *et al.* were defined in a fairly large cohort, data generated by TCGA contained independent patient-matched genomic and gene expression data across a cohort of >230 patients. This was precisely the type and amount of data required for the analyses intended for my thesis work, but before any progress could be made in predicting mesenchymal drivers, I had to ensure that the TCGA cohort was comparable to the Phillips cohort, and that the subtypes could be accurately recapitulated in this independent dataset.

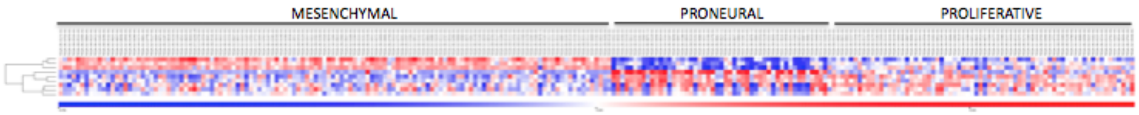
Furthermore, none of these resources are standardized to be directly integrated into a framework that I had proposed, so prior to actually running and validating the analysis, data processing was conducted to ensure that the data made available by the TCGA could be formatted and curated to generate biologically meaningful results. The details of each step are entailed here.

Classifying TCGA GBM Samples

The first task at hand in analyzing the TCGA dataset was the reclassification of patient samples into the MES, PN, and PRO subtypes originally defined by Phillips *et al*. These subtypes exist as user-defined molecular classifications; there is no predefined way to directly recapitulate these classifications solely from the TCGA data. In order to classify the TCGA samples as closely to the original Phillips samples as possible, I opted to build a centroid-based classifier that was trained on the Phillips classification and dataset and use this to separate the subtypes in the TCGA dataset. This approach allows us to more closely match the specific molecular profiles defined by Phillips, instead of attempting a new hierarchical classification in the TCGA set alone using the Phillips marker panels.

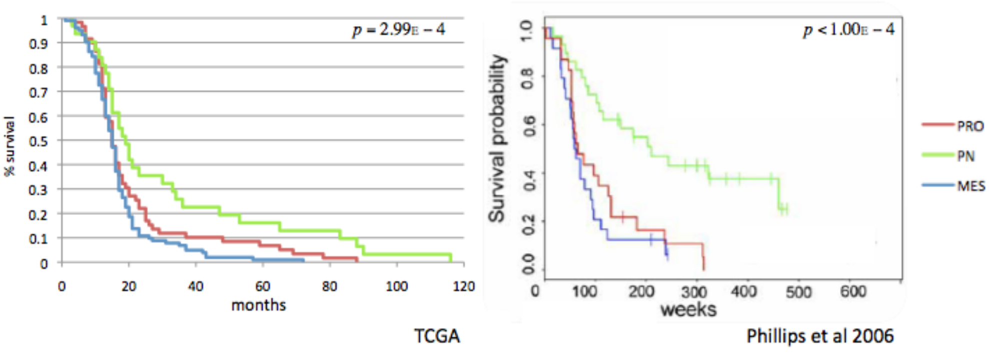
I selected three markers to represent each subtype (nine markers total), which were chosen by two criteria: the genes were not transcription factors and had the highest coefficient of variation between the samples. The markers selected by these criteria were: YKL40, SERPINE1, and TIMP1 for the MES signature; BCAN, OLIG2, and KLRC3 for the PN signature; and HMMR, TOP2A, and PCNA for the PRO signature. A centroid representing each Phillips subtype was created using the nine markers by defining a point in the search space with minimal average Euclidean distance between all samples of a single subtype in the Phillips dataset. These parameters were then used on the incoming TCGA samples for classification. Each TCGA sample was classified as the subtype of the nearest centroid, again measured by Euclidean distance. I selected markers that were not transcription factors to remove the possibility of bias in subsequent analyses that are informed by transcription regulatory networks; training a classifier on a transcription factor will artificially increase the enrichment of its regulon in patient samples when a transcription regulatory network is interrogated. Maximizing the coefficients of variation allowed for the selection of the minimum number of markers while maximizing their information, ensuring that I did not overfit our classification by using excessively large marker panels.

The application of this classifier to the TCGA dataset identified 164 MES samples, 64 PN samples, and 24 PRO samples (A list of samples and their classifications is enclosed in the appendix [APPEND01]). These samples exhibited robust expression of the appropriate panel markers to the exclusion of panel markers of the other two subtypes, and clustering the TCGA samples by these subtypes correctly reproduced three distinct expression clusters that concurred with the Phillips et al panels, as shown below (red indicates increased expression, blue indicates decreased expression).



After classifying the TCGA cohort into subtypes as defined by Phillips et al, hierarchical clustering of the patients according to the original classifying panel, when separate by subtype, reveals robust clustering of lineage-specific markers to the exclusion of markers of the other classes, as originally reported by Phillips et al.

When these samples were then checked for reciprocation of the reported separation of prognosis, I found a strong, statistically significant separation of the MES and PRO subtypes compared to the PN samples, as originally reported by Phillips *et al*. The p-value associated with the separation of Kaplan-Meier curves for the MES and PN samples was 2.99e-4. It should be noted that the TCGA dataset does NOT contain data from patients who survived cancer; all patient samples in the set have a time of death, which accounts for the discrepancies between the KM curves for the PN patients between the Phillips and TCGA datasets.



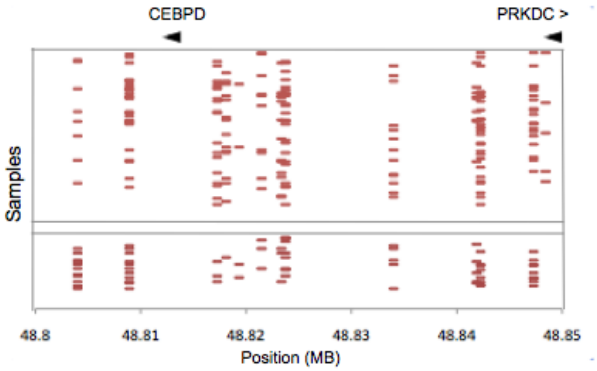
When separated into Phillips subtypes using a nearest-neighbor centroid classifier, TCGA tumor samples reciprocate the stratification of patient prognoses reported by Phillips et al: MES being indicative of the poorest prognosis, PN being indicative of the best. NOTE: the TCGA patient cohort does NOT include patients that survived the cancer; all patients in the set have a time of death, accounting for the discrepancies in the end curves.

Gene set enrichment analysis also confirmed that genes in the regulon of our predicted MES master regulators were significantly and specifically upregulated in TCGA samples classified as MES using this centroid-based classifier, leading us to conclude that I had accurately recapitulated the Phillips classification in the TCGA dataset. I used this classification scheme for the TCGA dataset as the basis for all subsequent analytic work.

Processing CNV and Gene Expression Arrays

While Affymetrix SNP arrays are commonly used as a proxy to infer copy number alterations, our analysis of the data showed several technical issues that reduced their usefulness in searching for functional genomic mutations. Agilent CGH (comparative genomic hybridization) arrays are specifically designed to detect copy number alterations at gene loci, and they bear oligos that hybridize along the coding region of their target genes, as well as probes interspersed throughout non-coding regions. This ensures, at minimum, cover of genic regions in the genome. The Affymetrix SNP array is not designed with coding information in mind, *a priori.* SNP arrays contain panels of SNPs that have been identified as informative to the LD of underlying populations and are scattered throughout the genome. The implication of this, and the first issue, is that there are regions in the genome with less or inadequate coverage to accurately infer smaller-scale CNVs.

As an example, CGH arrays successfully identified a focal amplification of the CEBPD locus on chromosome 8 as highly predictive of MES differentiation, even though the genomic region surrounding the locus

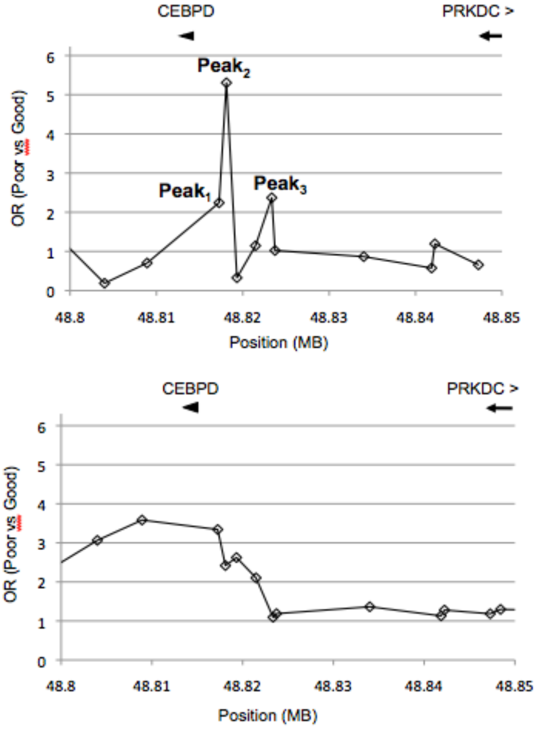


Affymetrix SNP arrays sparsely populate numerous gene-coding regions, preventing the detection of significant associations to molecular phenotypes such as the MES signature without employing integrative metrics. Even using these metrics never generates signals as significantly correlated to both expression and subtype classification as CGH array segmentation data.

was largely devoid of evidence for CNVs. I knew this was likely to be true because we have biologically validated CEBPD as a master regulator of MES differentiation. Furthermore, amplifications of CEBPD as reported by CGH arrays significantly correlated with increased expression of the CEBPD transcript, and increased expression of

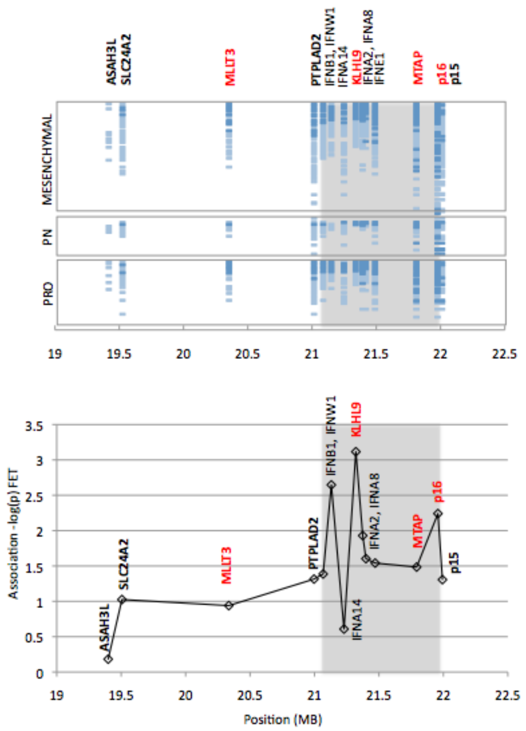
the CEBPD regulon. However, when the same region was analyzed using the SNP arrays, I found that there were no SNPs in the region that fell directly in the coding or promoter regions of CEBPD as shown in the figure provided (red hashes indicate SNPs that called an amplification event).

SNPs falling nearest the CEBPD locus generated erratic and statistically weak associations to the poor prognosis MES sample subtype (see peaks 1, 2, and 3). Integrating over the region using a sliding window did improve the association of the region spanning CEBPD to the poor prognosis subtype, but ultimately did not perform as well as segmentation data produced by CGH. I also observed that no individual SNP or integration at the CEBPD genomic locus was able to produce as significant a correlation with



Segmentation mapping of SNPs in the CEBPD region of chromosome 8 reveals that no SNPs in the array existed within the coding frame of the CEBPD gene. Because of this lack of coverage, no direct call on the CNV status was available, and standard GISTIC measures called the region diploid. CGH arrays, conversely, identified a focal amplification of the locus.

CEBPD transcript as simply using CGH segmentation mapping data: rMAX|SNP = 0.148 compared to rMAX|CGH = 0.258 using a simple Spearman correlation.



Segmentation mapping of CGH probes to genes on chr9.p show significant evidence for independent deletion frequencies of genes across the TCGA sample cohort. GISTIC-processed SNP arrays assign this entire region as deleted in all of the included patients. Blue hashes denote significant evidence of deletion (Dark blue: homozygous deletion, Light blue: heterozygous deletion). Greyed area indicates the minimum region marked as deleted when applying the most lenient GISTIC thresholds.

Additionally, a comparative analysis was performed between using Agilent CGH segmentation files and GISTIC-processed Affymetrix SNP arrays. GISTIC is a data pre-processing algorithm designed to construct “minimum common regions” of genomic alterations by integrating signals obtained from adjacent SNPs, similar to a sliding window integration algorithm. The first issue that occurred was the a significant loss of resolution in minimum deleted/amplified regions due to a combination of the sparseness of SNPs, their non-uniform distribution in the genome, and GISTIC’s tendency to favor the joining of two mutated fragments over keeping separate segments. Shown here is an image of the segmentation mapping of the oft-deleted chromosome arm 9.p. This genomic locus bears the most common oncogenic suppressor that is deleted in GBM: p16. This chromosome arm typically suffers significant deletions and tumors are frequently found with this entire arm missing. However, CGH segmentation mapping of the area across a sampling of TCGA patients reveals significant evidence of irregular, partial, and nonequivalent alterations specifically of the p16 region, including significant probe-based evidence of differential copy number counts among genes in the region (while p16 is almost uniformly lost, the probability of losing other genes in a patient decreases as the genomic distance from p16 increases). GISTIC mapping and SNP arrays call the regions equivalent; all are called as completely deleted for the entire region, which obscured a candidate identified by both the algorithm and MINDy: KLHL9.

For these reasons, I opted to primarily use data obtained on Agilent CGH arrays for the CNV portion of the integrative analysis. Furthermore, I chose not to employ least common region mapping methods and post-processing such as GISTIC. These methodologies, while useful for inferring and mapping large-scale genomic alterations, tend to artificially bias against smaller, focal genomic changes since the integrations across focal regions will dilute out a true, small signal of change with large amounts of true signal of diploid status. This was also observed in, but not limited to, the focal amplification of CEBPD: individual segmentation mapping of the locus reveals significant evidence for an amplification when considering the probes that hybridize directly to the CEBPD coding region, but no other gene loci in the area show any evidence for alterations in an overwhelming majority of the patients assayed.

These analyses produced several results. I successfully established the ideal platform and formatting for the genomic CNV data. The CGH arrays provided by Agilent platforms proved to have better coverage of coding regions throughout the genome, and could most accurately reconstruct locus-specific CNVs without the reliance of sliding window integration algorithms such as GISTIC, which tend to compromise resolution of the CNV topography in highly-mutated areas and bias against the identification of focal genomic changes. Additionally, I have shown that the Phillips GBM subtype signatures are robust and extendable to independent datasets. In independent datasets, I was able to reproduce the subtype classifications using a centroid-based classifier that was trained on the Phillips cohort, and this classification successfully and robustly reproduced the association of poor patient prognosis to the MES subtype in the TCGA dataset. With this data in hand, I moved forward in using the TCGA dataset to identify genomic mutations that drive MES differentiation in GBM.

CHAPTER 3 – DIGGIn, Part 1: Developing an algorithm to predict Drivers Inferred from Genetic-Genomic Interactions

Independently of the TCGA subtype classifications and CGH array standardizations, I developed an approach to infer causal mutations in GBM using the TCGA patient set. The approach required an algorithm capable of integrating genomic and gene expression data to assign what we coined “Functional” CNVs, which are then integrated with ARACNe and master regulator analysis results to produce candidate drivers of a molecular phenotype of interest. The general workflow is presented below, where the functional genetic-genomic analysis occurs naïvely and independently to molecular classifications and ARACNe/master regulator analysis and then used in tandem with these data to arrive at a consensus for candidates. Since the ARACNe, master regulator, and classification analyses are done independently and methodologies already exist to characterize the master regulators and downstream components, I focused on development of the DIGGIn algorithm: a pipeline to identify “functional” CNVs with the intent of making mesenchymal **D**rivers **I**nference from **G**enetic-**G**enomic **In**teractions.

As a first step, I sought to develop an approach to pare down the likely candidate mutations. CGH array data covers the entire genome, including >20,000 known genes and associated regulatory regions; we hypothesized that very few of these genes would actually contribute meaningfully to the differentiation of the mesenchymal subtype, and testing for all loci would significantly reduce statistical power simply by including a majority of these irrelevant loci. Since GBM is a disease primarily defined by chromosomal alterations (compared to point mutations), I sought to assign functionality to a genomic alteration by correlating it with a change in the transcriptional expression of genes that fell within the altered region in a fashion similar to a QTL. I hypothesized that a genomic alteration at a gene locus that functionally alters the transcriptional behavior of that gene should result in a correlation between the genomic event and transcriptional expression across the patient samples. This assumption also allows us to firmly attribute a causal direction to the perturbations and subsequent transcriptional cascade. An example is provided in the figure below: amplifications of the EGFR locus are biologically validated oncogenic drivers that are observed in a significant fraction of GBM patients. Concomitantly, patients

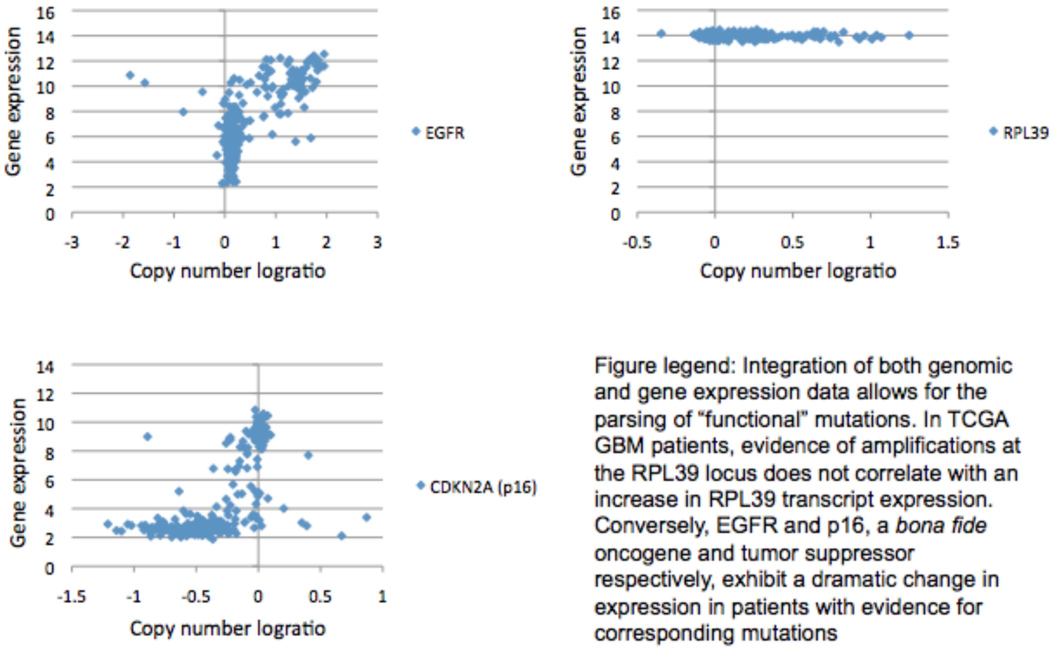


Figure legend: Integration of both genomic and gene expression data allows for the parsing of "functional" mutations. In TCGA GBM patients, evidence of amplifications at the RPL39 locus does not correlate with an increase in RPL39 transcript expression. Conversely, EGFR and p16, a *bona fide* oncogene and tumor suppressor respectively, exhibit a dramatic change in expression in patients with evidence for corresponding mutations that bear evidence for EGFR amplifications in tumors also show greater transcriptional expression levels of EGFR. The mutation at this locus was the type of mutation I wished to uniquely identify.

A natural extension of this methodology is to then infer the indirect molecular perturbations associated with a genomic mutation. After establishing that a genomic copy number change at a gene locus correlates with a change in transcription of that gene, the expression of all other genes expressed in the tissue can be tested for association with the genomic mutation. Effectively, this approach generates a transcriptional dysregulation network associated with each genomic mutation in the same vein as an ARACNe transcriptional network. However, I do not apply the Data Processing Inequality because the direct effect is already known (the transcription of the gene bearing the alteration), and indirect effects are the most informative in establishing how a genomic mutation propagates through the molecular network and affects the expression of specific gene sets. Through this workflow, I aimed to be able to infer candidate driver mutations from genomic mutations that can be predicted to alter the molecular behavior of a GBM tumor.

Algorithm Workflow

DIGGIn is designed to import two matrix files: one corresponding to gene expression data across a cohort of patients, and another corresponding to the genomic CNV segmentation data of the same cohort. The program is coded to accept matrices that have gene / coding loci as matrix rows, and individual samples as the columns. Capitalizing on the computational simplicity of Mutual Information (discussed below) the program has been designed to dynamically match data between these two matrices using multi-dimensional associative arrays. Data is dynamically called from both matrices by their given sample and gene IDs, meaning that neither the row orders nor the column orders in the two matrices matters, so long as the labels used are the same between the two matrices.

From these two files, DIGGIn follows several steps in order to predict functional genomic alterations. The details of each step are discussed in the subsequent sections.

DIGGIn Phase I: Genetic-genomics analysis

1. Find optimized kernel for Mutual Information estimation
2. Build Mutual Information Null Distribution and compute p-value function
3. Identify functional CNVs
4. Link differential expression of other genes to functional CNVs (regulons)

While the core of the genetic-genomic analysis identifies the predicted functional CNVs that should be considered for further analysis, the assigning of particular f-CNVs as candidate drivers of any particular phenotype must be done in additional analytic steps.

DIGGIn Phase 2: Identifying candidate drivers

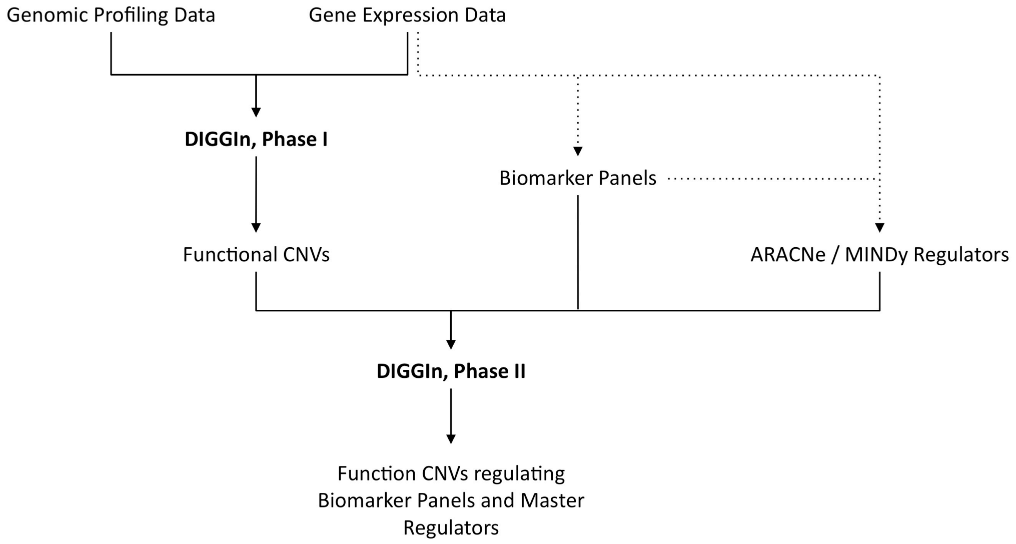
5. Identify f-CNV regulons enriched in a gene panel of interest

6. Identify increased activities in ARACNe-predicted master regulators

7. Eliminate passenger mutations and identify maximum effect size mutations

The integration with ARACNe and MINDy predictions, as well as statistical tests for association to molecular, clinical, and phenotypic outputs occurs at a post-processing level, which is discussed in chapter 4. A full schematic of the DIGGIn architecture is provided below, detailing the inputs and outputs to each of the two phases of DIGGIn, and how they integrate with other systems biological methods such as ARACNe and MINDy. Solid lines indicate DIGGIn-specific flow that was developed specifically in this thesis. Broken lines indicate established methodologies such as ARACNe/MINDy and generalized identification of biomarker panels utilizing hierarchical clustering and classification metrics.

DIGGIn Architecture



The DIGGIn algorithm can be broken into two general phases. Phase I integrates genomic profiling and gene expression data to identify a set of CNVs that exist in genes with a concomitant change in gene expression. The expression of all other genes in the genome are then compared to these “functional” CNVs to assign indirect dysregulation. Phase II of DIGGIn interrogates these functional CNVs for statistical enrichment of gene marker panels and/or enriched activation of master regulators and modulators provided by ARACNe/MINDy. Phase II also implements several conditional association metrics to identify combinations of mutations for the maximum effect size across the available sample space. DIGGIn-specific flow is outlined with solid arrows. Previously available methodologies and how they relate to the DIGGIn framework are outlined in broken arrows.

Mutual Information

I opted to measure the dependence of differential gene expression on genomic mutation using the information theoretic metric, Mutual Information. This is the central metric used by the ARACNe algorithm and it has been shown in several models to detect biologically-validated transcriptional interactions between transcription factors and their targets to a degree that is missed by more traditional statistical tests. Mutual information is a probabilistic measure capable of detecting non-monotonic correlations between continuous variables, and has several advantages over more traditional statistical methods. It does not require arbitrary discretization of data, it has low computational complexity, and does not require *a priori* information or inferences on the distribution or topology of the data used. The first point was a particularly important consideration given the use of CGH arrays for detecting copy number variants. Traditional copy number analysis involves assigning a statistical threshold to reject the null hypothesis, locus count = 2. Using these types of methods in an integrative analysis would require a statistically dependent discretization of the CGH data, followed by another statistically dependent measure of co-information, introducing multiple additional hypothesis corrections and results that are highly dependent on the original thresholds set. Instead, mutual information allowed us to consider the entire range of CGH array values as a vector of continuous random variables and tie them to patient-matched expression vectors.

Mutual information between a matched pair of continuous random variables, x and y, is defined as:

$$MI[x;y] = \sum_{x \in X} \sum_{y \in Y} p(x,y) * \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

wherein p(x) and p(y) are the marginal probability functions for random variables in sets X and Y, and p(x,y) is the observed joint probability density function of the matched variables.

The mutual information function measures statistical depdence as the ratio of the observed probability of two variables co-occurring, p(x,y) to the expected probability given statistical independence, p(x)*p(y), weighted by the expected probability of p(x,y). If X and Y were conditionally independent (the outcomes of X never affect Y and vice versa) then the predicted probability of the two variables co-occurring is equal to the product of each event occurring separately, p(x,y) = p(x)*p(y). In this instance, the function simplifies to log(1) = 0 information, and it can be claimed that they are mutually independent or that no *information* exists between the variables (again, x does not affect y; therefore knowing x tells nothing about y). If the events are not mutually independent, p(x,y) > p(x)*p(y), and the function produces a non-zero value, providing a quantitative value of the information, or correlation, contained between the two variables.

For the purposes of inferring interactions between gene expression or between genomic-genetic pairings in a rank-sorted dataset, we measure the probability density functions for the log ratio using a kernel bandwidth estimator that weights neighboring datapoints based on their distance from each tested point (discussed below). However, the expected probability of any individual pairing of variables (or each individual point), p(x,y), is equal to 1/*M*, where *M* is the number of samples in the probability space. Thus, the definition of mutual information in our context can be reduced to:

$$MI[x;y] = \frac{1}{M}\sum_{x\in X}\sum_{y\in Y}\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

wherein the remaining term corresponding to the observed p(x,y) is inferred by a probability density function created from the sets X and Y. In the case of genetic-genomics, the X vector would represent a vector of continuous variables corresponding to the copy number status of a tested genomic locus, indexed by sample. Conversely, the Y vector would represent patient-matched expression values of genes being tested for co-information with the mutation status at locus in vector X.

Mutual information across continuous random variables can be efficiently estimated using a Gaussian kernel estimator to construct non-Normal probability density functions for variable sets X and Y, defined:

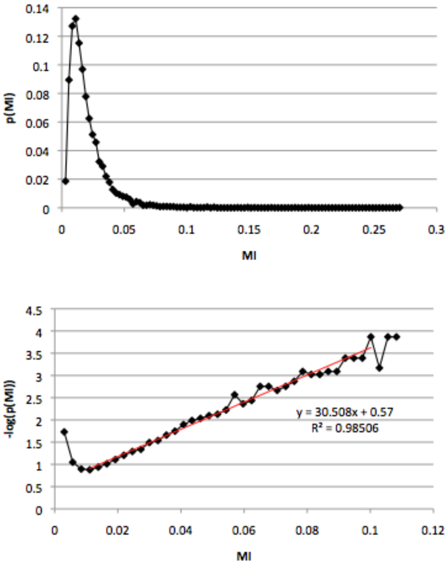
$$p(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Fitting a Gaussian kernel with an optimized bandwidth estimator, *h*, as a window for measuring variance allows for efficient estimation of MI between variables, resulting in the fully developed function for MI estimation:

$$MI[x;y] = \frac{1}{M}\sum_{x\in X}\sum_{y\in Y}\log\left(\frac{\frac{1}{2\pi h^2}\sum_{i=1}^{M} e^{-\left(\frac{(x-x_i)^2+(y-y_i)^2}{2h^2}\right)}}{\frac{1}{\sqrt{2\pi h^2}}\sum_{i=1}^{M} e^{-\left(\frac{(x-x_i)^2}{2h^2}\right)} * \frac{1}{\sqrt{2\pi h^2}}\sum_{i=1}^{M} e^{-\left(\frac{(y-y_i)^2}{2h^2}\right)}}\right)$$

While such an estimator is asymptotically unbiased in infinite or large datasets, in finite datasets a bias does exist and depends on the kernel width used. Since kernel selection is largely heuristic and empirical, this can lead to MI estimates whose accuracy varies, based on kernel selection. However, the performance of MI in this context is not directly dependent upon the fidelity of the MI estimate to the true MI value, but instead depends on the accuracy of MI *rank* estimates. Statistical significance is established by testing $MI_{xy} \geq MI_0$, where MI_0 is defined as the mutual information threshold obtained from the null distribution at a given significance[8][9]. In this context, the bias attributable to kernel selection is minimized, so long as a kernel that is optimally fitted for the dataset being analyzed is held constant across all comparisons and the modeled null distribution. The end results are MI values whose ranks are directly comparable to those produced in the null distribution to ascertain a p-value.

As reported by Margolin *et al*., selecting a variety of kernel bandwidth values did not significantly alter the results of the analysis for a majority of the mutated loci, as long as the same kernel estimator was used in both the construction of the null distribution and the analytics, allowing *h* optimization to be a largely heuristic process. The most significant changes in results involved loci whose mutational frequency was low (<5% of patients tested exhibited evidence for a locus alteration) and the gene expression vector had a low coefficient of variation. The ramifications and solutions to this are discussed below. Therefore, rather than optimizing a kernel for every M^2 number of comparisons (an extremely resource-intensive process), I instead optimize a single kernel width by selecting for the kernel value that minimizes the MI between random vectors. I take a Monte Carlo model approach and compute over 10^5 iterations the MI from randomly paired CNV and gene expression vectors under different kernel widths starting from *h = 0.9* to *h=0.01* and select the maximum value *h* wherein the 90th percentile of the recorded pairs falls below MI=0.05. This kernel width is then used for all



The null distribution for estimating MI p-values is made by measuring the MI between randomized CNV and gene expression vectors over 10e5 iterations and measuring the frequency of each MI value. The exponential decay of the function in the right tail allows a linear fit to log-transformed data for the extrapolation of arbitrarily small p-values subsequent analysis and the null distribution generated is then used for the calculation and extrapolation of p-values, and for all subsequent analytic measurements of MI in the dataset. This approach ensures the selection of a maximum value *h* such that the (MI | *h*) in the null distribution allows for maximum dynamic range in detecting non-random co-information between tested CNV and gene expression vectors.

The null distribution corresponding to the optimal kernel, *h,* is used to estimate MI to arbitrarily small p-values. 10^5 randomized iterations allows for the construction of a probability density function from which I can assign a p-value, *p*, to any recorded MI value. This distribution is asymptotically distributed on the right tail, which allows us to log-transform the empirically determined p-values and fit a linear function to the data. From this fit, I can estimate arbitrarily low p-values without having to run large numbers of iterations in the MI null distribution. For the work with TCGA GBM, the critical MI value for rejecting H_0 was set as the MI_0 with an FDR < 0.1. based on 10^5 iterations.

Results and Analysis

Following the standardization of the TCGA data, I performed the analysis to infer functional genetic mutations. Of the ~9,000 genes expressed in the patient samples and full ~20,000 gene loci tested for genomic alterations, only 1489 genes had alterations that shared significant mutual information with their gene expression (a list of these genes is included in its entirety in [APPEND02]).

These alterations were subsequently flagged as functional CNV genes (f-CNVGs), genes whose genomic loci showed evidence of an alteration that correlated with a change in transcription levels of the same gene. Any f-CNVG could subsequently be considered a candidate causal driver mutation, since the genomic event could be definitively linked to a molecular perturbation. Testing the performance quality of DIGGIn included two basic tests: reciprocation of known ARACNe transcriptional regulons when those TFs were mutated, and the successful identification of *bona fide* oncogenic drivers previously reported in GBM.

Successful recapitulation of Transcriptional Regulons

Among the 1489 identified f-CNVGs were several transcription factors including the validated mesenchymal master regulator, CEBPD. I compared the panel of genes that showed statistically significant MI with the mutated transcription factors against the ARACNe-predicted targets of each transcription factor in GBM and the mesenchymal signature as a whole. As expected, there was statistically significant overlap between the differential expression panel predicted by the f-CNVG analysis and the predicted transcriptional regulon: the overlap of ARACNe-predicted targets of CEBPD and the genetic-genomic predicted targets of CEBPD was highly enriched ($p<6.58e-17$). The enrichment of mesenchymal signature genes associated with the deletion of KLHL9 was also highly significant ($p<2.00e-9$). DIGGIn was able to properly associate the dysregulation of known transcriptional targets to the corresponding mutated transcription factors.

Successful identification of gold standard Oncogenic Drivers

As an additional quality control metric, I tested whether DIGGIn was capable of identifying common mutations that have already been reported in the literature. DIGGin was successfully able to identify several classical oncogenic drivers as f-CNVGs. Of the 18 classical oncogenic drivers reported[21] as GBM oncogenic drivers, the algorithm positively identified 14, including loci such as EGFR, CDK4, MYCN, p16, PTEN, RB1, and NF1, for a statistically significant enrichment of true-positive mutations ($p<1.93e-10$). The remaining four were either too rare in the TCGA population tested to obtain statistical significance (all candidates missed with significant mRNA calls were present in <10 TCGA samples, <5% of the set) or were disregarded as potential drivers because the transcript was not found in the tissues, implying that changes in expression were not possible and were therefore non-perturbing. This latter point was one of the initial goals in designing the algorithm; the omission of artificial candidates via the integration of additional biological information drastically improved the computational power in parsing the data into meaningful mutations. These findings showed that the algorithm was able to attribute changes in gene expression levels of oncogenic drivers to the mutations that occurred at their genomic loci (a list of these genes is provided in [APPEND03]). Based on these results, I concluded that any candidates successfully identified by the algorithm could be expected to be a legitimately functional CNV to be included in subsequent analysis.

Detection of rare, yet functional mutations via the MINDy algorithm

One caveat of the use of mutual information is the relatively large data requirement to obtain usable MI estimates. It was established that roughly 100 independent samples was the absolute minimum required for ARACNe to have sufficient statistical power to produce robust transcriptional networks[1][8]. While the TCGA dataset included over 200 patients, there was an added constraining limitation on the power: a non-uniform distribution of mutations across the patients. Certain loci, such as EGFR and p16, were altered in over 60% of the patients and showed a very dynamic range of gene expression, providing ample information across “affected” and “unaffected” patients to generate co-information with gene expression. In contrast, copy number changes that were more rare or whose gene expression had lower coefficients of variation provided less information. Though I did not bin or discretize the data, using continuous variables when a rarer mutation is being tested results in highly-clustered points in a joint probability density function, diluting the ability to detect information over the null distribution without a specifically-optimized kernel. Although it may appear conceptually ideal to favor mutations with the highest effect size when considering candidate driver mutations, there is always the possibility of a very important, yet rare, contributor to the etiology of a highly heterogeneous disease.

To address this issue, we also implemented the MINDy (Modulator Inference by Network Dynamics) algorithm[2][22][23] to detect and identify *a priori* candidate modulators of the MES master regulators. Implementation of the genetic-genomic algorithm on the TCGA GBM dataset showed that, while the algorithm was highly successful at parsing functional mutations when the mutations were present in relatively large portions of the population, it had difficulty detecting functional CNVs when they existed in fewer than ~5% of patients overall (<10 patients) if the variance of gene expression or CNV reads was relatively high. The MINDy algorithm provided a complementary approach wherein potential modulators of the MES master regulators were predicted *a priori* from the gene expression profiles. These candidates were then cross-referenced with CNV data to ascertain whether any patients carried mutations in gene loci of the MINDy-predicted modulators. This approach does not require any minimum number of affected patients, but simultaneously does not provide statistical evidence that can be used to predict effect size. However, including the MINDy modulators significantly expanded our coverage of reported oncogenic drivers.

For comparison, DIGGIn was modified to use statistics to measure correlation between CNVs and gene expression. Both the Mann-Whitney U-test and Student's T-test were used to measure a change in gene expression between samples that were binned into "diploid," "amplified," or "deleted" groups. This methodology requires assigning two statistical thresholds: one to assign significant evidence for copy number change, and another to ascertain significant differential expression between samples with vs. without copy number change.

While these approaches were able to reasonably detect functional CNVs at a population threshold lower than MI, these approaches did *not* outperform a simple cross-comparison of CNVs at gene loci and MINDy-predicted modulator loci, due largely to the dependency on multiple statistical thresholds with corrections for multiple hypothesis testing. Related to this, the cross-comparison method is a much more intuitive approach, requiring significantly less pre-processing and fewer variable thresholds. The results of the parametric analyses could significantly change based on the thresholds used for differential expression, copy number alteration, and minimum effect size considered. Since MINDy successfully identified nearly every rare locus that was detected in the traditional methods, I opted to use MINDy.

In summary, the DIGGIn algorithm was designed to implement mutual information to identify what I refer to as f-CNVs. f-CNVs are genomic copy number variations occurring at gene loci that present with a concomitant alteration in the transcriptional expression of the genes contained in that CNV. This allows for a powerful filtering step to remove a significant amount of probes (and therefore hypotheses) to be tested when identifying candidate mutations for further study. However, additional processing is required to assign specific f-

CNVs as driver mutations for specific molecular expression panels like the MES gene expression signature, which will be detailed in Chapter 4.

CHAPTER 4 – DIGGIn, Part 2: MES-specific candidate drivers are identified from f-CNV results based on enrichment analysis and conditional association

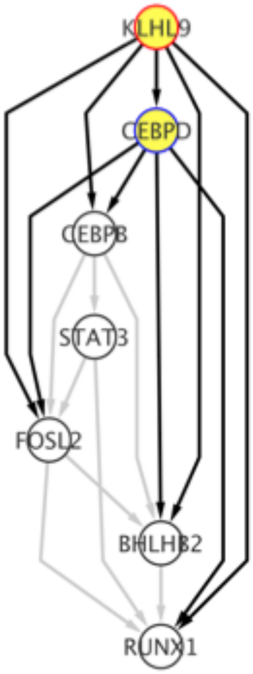
Interrogation for MES Gene Expression and Master Regulator Activity

The initial steps of DIGGIn are designed to identify functional CNVs. These genomic mutations are the mutations that can directly be associated with a molecular perturbation of the genes expressed in a tumor. In order to do so, I adapted the mutual information function derived for the ARACNe algorithm and applied it to the integration of genomic and gene expression array data. Phase I of DIGGIn defines genomic loci that have a traceable link between a mutation of a gene and differential expression of that gene. It also defines all the genes whose expression is affected indirectly by this gene locus. These significant functional CNVs, or f-CNVs are defined as a subset of candidate driver loci, isolated from a field of hundreds of thousands of candidate loci.

However, CNVs do not fall exactly within single genes at a time. They occupy variably large swaths of genomic regions and can manipulate any combination of genes expressed in the tumor, which can regulate any number of independent biological processes. In order to specifically identify candidate drivers of mesenchymal differentiation, I needed to create an algorithm with additional metrics to assign likely function to this specific molecular profile. Phase II of DIGGIn is a suite of analyses integrating conditional associations, and modeled off of classical genetic testing in order to rank and select candidate f-CNVGs that are maximally likely to drive the phenotype of interest.

I ascertained whether any of these 1500 f-CNVGs could specifically regulate the genes in the mesenchymal gene expression panel by conducting an enrichment analysis on the f-CNVG's "regulon," comprised of all other genes whose mRNA expression shared significant information with the genomic perturbation. f-CNVG loci were flagged as potential MES differentiation drivers if their regulons were statistically enriched in genes in the MES gene expression panel defined by Phillips *et al*. This was subsequently cross-referenced with the results of the MINDy algorithm as applied to the TCGA GBM cohort to obtain consensus candidates (genes identified by both methods) and potential candidates (genes identified by at least one method). This enrichment analysis was conducted using the standard GSEA protocol with an FDR < 0.1. The genes that carried statistically significant enrichment of differentially expressed MES marker genes are included in [APPEND03]. Of the ~1500 f-CNVG loci successfully identified, only 41 were significantly enriched specifically in genes that define the MES gene expression signature, and identified by both genetic-genomic approaches and MINDy as a potential driver, including amplifications in the previously identified master regulator, CEBPD. Two genomic loci presented high enrichment in both the activation of the MGES and increased activity of *all* predicted MES master regulators: amplifications of CEBPD and deletions of KLHL9.

The above figure displays two types analyses that DIGGIn is capable of conducting on defined f-CNVs. In the event that an ARACNe network is not available, a more straightforward approach can be conducted by testing each f-CNV for increased activity of a biomarker panel of interest as a whole, pitted against the total size of the f-CNV's dysregulatory hub. This is represented in the figure by the blue squares. Alternatively, ARACNe master regulators can be integrated into DIGGIn and we can instead measure master regulator *activity*. We define activity of a transcription factor in this context as the collective change in expression of its ARACNe-predicted targets. By measuring the coordinated change of a TF's targets, we have a proxy for the TF protein's activity (more active TFs will produce more target transcripts). Using this measure, DIGGIn can infer which f-CNVs are predicted to contribute to the increased activity of specific master regulators, such as the mesenchymal master regulators indicated as yellow squares. This analysis is conducted independently of the standard statistical enrichment, but generally produces better results when f-CNV hubs become large.



Of the 41 genomic CNVs that had significant MI with MES master regulators, the two loci with the most significant enrichment of both the MES signature genes and, subsequently, the activity of the ARACNe-predicted master regulators (yellow squares) were amplifications of CEBPD and deletions of KLHL9. Master regulator activity was measured as a function of the differential expression of the MR's ARACNe-predicted targets. While simple enrichment by Fisher's Exact test of MGES genes was significant for these two loci, testing for master regulator activity provided more power in identifying loci.

Selecting MES Driver Candidates

The functional CNVs at CEBPD and KLHL9 were significantly enriched in MES genes and MES master regulator activity, and were subsequently considered potential driver mutations that could induce mesenchymal differentiation in GBM tumors among the 41 candidates. However, there are still two issues that had to be addressed before positively stating that KLHL9 and CEBPD were mutated master regulators/modulators of MES transformation in GBM:

1) There are an unknown number of causal driver mutations amidst hundreds, if not thousands, of mutated loci.

2) CNVs tend to affect large genomic regions; whole chromosomal deletions are affected in GBM, which means that any mutated locus that drives the differentiation a subtype will usually come with several other neighboring loci that are also associated with the subtype, but are biologically irrelevant.

To address the first issue, I implemented a recursive greedy search algorithm described below:

```
$X_0 = \arg\max(m \bullet s)$

$while(p_{FET}(X_i) < \alpha)\{$
    $m_i = \{m \setminus X_{i-1}\}$
    $s_i = \{s \setminus X_{i-1}\}$
    $X_i = \arg\max(m_i \bullet s_i)$
$\}$
```

The TCGA GBM cohort was classified into MES, PN, and PRO subtypes according to the classifier trained from the Phillips work. A vector set, *s*, was defined containing each sample classified as MES or PN. A complementary vector set, *m*, was defined containing the mutation state of each candidate genomic locus across all patients in *s*. A candidate driver f-CNV, X_0, was selected by finding the maximum the dot product between the candidate f-CNV's vectors *m* and *s.* X_0 was essentially tested for association specifically to the MES subtype compared to the PN subtype and a p-value could be assigned by a Fisher's Exact Test (FET). X_0 is subsequently the first candidate driver identified and accounts for some portion of samples in the vectors *m* and *s*.

Subsequently, the vectors *m* and *s* are updated to contain only the set of samples that existed in the original vectors that are NOT also contained in the set of samples that bear the mutation X_0. Using the new sets m_1 *and* s_1, the next most significant locus associated to the mesenchymal subtype is identified, X_1.

This analysis is recursively repeated until statistical significance of association can no longer be achieved with the remaining samples in each vector.

The use of the *argmax* function with the Fisher's Exact Test in a greedy search indirectly pushes the algorithm towards selecting candidate f-CNVs that offer the maximum effect size detectable by statistical methods, which manifests in the selection of two types of candidates. The value to this approach is that it is capable of detecting significant associations resulting of both high-frequency mutations (more prevalent mutations will provide more statistical power, reflecting higher effect size) and rarer but highly specific mutations (mutations may be more rare but only occur in specific sets of samples, producing high associations) as long as the distribution of mutations are approximately equal.

However, due to the unequal distributions of mutations throughout samples, there is a significant risk of the analysis favoring very common mutations (>50% of the samples) to the exclusion of very rare ones (<5% of samples) simply due to the difference in power associated with limited sample sets; stronger p-values are more readily obtainable when the mutation is more prevalent, such that a significant driver mutation that occurs in only 5% of samples may be overshadowed and placed in a superset with a more common mutation simply because the minimum p-value obtainable in this context is an order of magnitude higher than other candidates. This can lead to false positive where multiple, independent true driver mutations are placed in the superset of another, noncausal mutation simply because this mutation is so common that it co-occurs with multiple drivers.

To hedge against this occurrence, each subgroup of TCGA patients associated to a significant candidate f-CNV was redefined as the entire MES cohort, and the recursive analysis was repeated to ensure that there were no viable subcategories of sample sets as defined by individual candidate drivers.

As a result of this analysis, we were surprised to find that ~50% of the MES samples could be accounted for by bearing a mutation in only one of two loci: amplifications of CEBPD (one of the biologically validated master regulators of the MGES), and deletions of KLHL9, a previously unreported gene in tight linkage disequilibrium with one of the most prevalent mutations in GBM, deletions of p16.

To parse out false positives resulting from being in linkage disequilibrium with a true driver mutation, I devised a computational algorithm based on the following hypothesis: *Of all mutations that are statistically associated with each other (co-occurring), no mutation can be more significantly associated with the MES phenotype than a true driver MES differentiation.*

This can expressed mathematically by comparing the association of two hypothetical associated genes, $gene_x$ and $gene_y$, to the MES subtype compared to the PN subtype. A mutation (*mut*) at gene$_x$ can be correctly identified as a candidate driver if, when all other co-mutated genes *y* in superset *Y* are wildtype (*WT*), it fits the criteria:

$$\underset{y \in Y}{p_{FET}}(MES; gene_x{}^{mut} \mid gene_y{}^{WT}) < p_{FET}(MES; gene_y{}^{mut} \mid gene_x{}^{WT})$$

This indicates that any co-information between gene$_y$ and the MES subtype is actually an artifact of information passed to it from gene$_x$ by statistical association. Conversely, any gene for which this does not hold true can be identified as a passenger mutation since its association to the mesenchymal subtype is conditionally dependent upon the mutation at another locus. If the mutation frequencies are such that there are no samples that are mutated at gene$_x$ and wildtype at gene$_y$, the complementary association can be taken:

$$\underset{y \in Y}{p_{FET}}(PN; gene_x{}^{WT} \mid gene_y{}^{mut}) < p_{FET}(PN; gene_y{}^{WT} \mid gene_x{}^{mut})$$

I constructed associative networks of co-occurring mutations by performing a pair-wise Fisher’s Exact test matching all ~1500 f-CNV genes with each other. From this map, I extracted all genes that were statistically associated with amplifications of CEBPD and deletions of KLHL9. Any of these loci could be a true driver of the MES subtype instead of KLHL9 or CEBPD, due to their high co-occurrence with these genes. Based on our stated hypothesis, I tested the conditional association of each locus to the MES classification across TCGA samples, given that another co-mutated gene was *not* mutated. If the conditioned locus is the true driver, the association of the tested locus should be abrogated (no condition, or gene in this case, can be more associated to the effect than the cause).

This analysis concurred that the most likely driver genes were CEBPD and KLHL9, as conditioning on the absence of these two mutations completely removed the association of any other genomic loci that were co-mutated with them. Conversely, the associations of KLHL9 and CEBPD to the MES subtype were the most robust to conditioning on the absence of other mutations.

In order to identify and biologically validate mesenchymal drivers in this GBM patient cohort, a series of experiments were designed to assign both statistical association of candidate drivers to the MES subtype and associated poor prognosis, and biological validation of the molecular function of these mutations alongside a possible mechanism of action.

CHAPTER 5 – Deletion of the KLHL9 E3 Ligase Complex Adaptor Protein Induces Mesenchymal Signature in High-Grade Glioma

DIGGIn was designed to identify candidate genomic mutations that drive the activation of distinct molecular expression panels. This is accomplished using a two-step analytic process where functional genomic mutations are identified, then subsequently interrogated for statistical enrichment of the gene panels of interest (in this case, mesenchymal differentiation). The results of the DIGGIn analysis on the TCGA cohort identified KLHL9 as a previously unreported candidate driver of mesenchymal differentiation in GBM. I followed up this candidate driver with a series of statistical analyses on independent human cohorts and biological experimentation in cell lines. The full process of implementation of DIGGIn and subsequent analysis are detailed in this manuscript.

This manuscript is currently undergoing revisions for resubmission to *Nature* and contains the implementation of DIGGIn as described in this thesis. It additionally details the biological methods and experiments used to validate the identified candidate, KLHL9, and the subsequent identification of the mechanism by which it interacts with the validated mesenchymal master regulators, CEBPB and CEBPD. All figures and figure legends for this manuscript have been attached to appendices 5 and 6 ([APPEND05] and [APPEND06]). Additional experiments conducted to address reviewer comments are discussed in the following Chapter 5a.

Deletion of *KLHL9* E3 Ligase Complex Adapter Protein Induces Mesenchymal Signature in High-Grade Glioma

James C. Chen[1,4], Mariano J. Alvarez[1], Gabrielle E. Rieckhof[1,2], Francesco Niola[8], Archana Iyer[1], Kristin L. Diefes[10], Kenneth Aldape[10], and Andrea Califano[1,2,3,5,8,9]
[1] Columbia Initiative in Systems Biology,
[2] Center for Computational Biology and Bioinformatics,
[3] Department of Biomedical Informatics,
[4] Department of Genetics and Development,
[5] Department of Biochemistry and Molecular Biophysics,
[6] Department of Neurology,
[7] Department of Pediatrics,
[8] Institute for Cancer Genetics, Columbia University,
[9] Herbert Irving Comprehensive Cancer Center, Columbia University, New York, New York, 10032 USA.
[10] Department of Pathology, M.D. Anderson Cancer Center, Houston, Texas 77030, USA.

Correspondence should be addressed to: Andrea Califano, califano@c2b2.columbia.edu; Antonio Iavarone, ai2102@columbia.edu; Anna Lasorella, al2179@columbia.edu.

The most aggressive subtype of human high-grade glioma (HGG) is characterized by expression of mesenchymal genes (the "mesenchymal signature"), a phenotype driven by aberrant activation of master regulators *C/EBPβ*, *C/EBPδ* and *STAT3*[1]. Yet, the specific genetic alterations contributing to this tumor subtype remain largely unknown, despite availability of large-scale mutational and gene copy number alteration datasets[2]. We hypothesized that the master regulators of the mesenchymal subtype represent a natural bottleneck, responsible for canalizing aberrant upstream signals from multiple genetic alterations. Confirming this, unbiased genome-wide regulatory-network analysis of these genes and of their upstream regulators identified focal amplifications of *C/EBPδ* and frequent deletions of *KLHL9*, an adaptor of Cullin3-based E3 ligases, in poorest prognosis mesenchymal HGG tumors. Loss of *KLHL9* in tumors leads to *C/EBPβ* and *C/EBPδ* protein accumulation, while re-expression of the gene triggered ubiquitin-mediated, proteasome-dependent destruction of these transcription factors, abrogated the expression of mesenchymal genes, and promoted cell cycle arrest. An independent HGG patient cohort confirmed *KLHL9* deletion in 70% of poor prognosis cases. Taken together, these data elucidate a previously unidentified *KLHL9* deletion as the most frequent alteration promoting aggressive mesenchymal subtype of HGG and provide a novel, regulatory-network based paradigm for the elucidation of driver mutations in cancer.

High-Grade Glioma (HGG), including astrocytoma grade III and IV, are the most common human brain tumors and are virtually incurable, with an average survival of 12 months post diagnosis[3]. Gene expression profiling of HGG samples from large patient cohorts, using a gene subset that optimally correlates with disease prognosis, has revealed three subtype-specific signatures, including mesenchymal (MGES), proliferative (PGES), and pro-neural (PNGES) markers respectively[4]. Among these, the MGES was found to be associated with the worst prognosis, as further confirmed by analysis of additional Glioma datasets[1], including TCGA[5] and Rembrandt[6]. More recently, co-segregation analysis with large-scale gene copy number (GCN) alteration and mutational data[5] revealed an alternative signature stratification, including Proneural, Neural, Classic, and Mesenchymal[7], with the Mesenchymal expression signature again associated with the worst disease prognosis. This allows us to use the Mesenchymal gene expression signature is a quantitative molecular proxy for poor prognosis GBM.

Analysis of the mesenchymal signature, using a regulatory network inferred *de novo* by the ARACNe algorithm, elucidated three Master Regulator (MR) genes – the transcription factors (TFs) *C/EBPβ, C/EBPδ,* and *STAT3* – as synergistic drivers of the mesenchymal gene expression signature of high-grade glioma[1]_ENREF_3. The *C/EBPβ* and *C/EBPδ* subunits form both homo- and heterodimers that regulate the same targets. Co-ectopic expression of *C/EBPβ* and *STAT3*, but not of either gene in isolation, was sufficient to reprogram neural stem cells along an aberrant mesenchymal lineage. Conversely, co-silencing of the two genes abrogated the mesenchymal phenotype in short-term cultures of human HGG-initiating cells and established glioma cell lines, both *in vitro* and *in vivo*. Finally, stratification of an independent cohort using *C/EBPβ* and phospho-*STAT3* antibodies revealed that functional co-activation of these TFs was associated with the worst prognosis in ~70% of the patients.

Among the relatively large panel of genetic alterations reported by the TCGA Consortium[5], only *NF1* mutations were statistically associated with the MGES, accounting however for < 25% of the worst prognosis samples, thus leaving the majority of genetic variance associated with worst prognosis unaccounted for. Despite the growing availability of data and the identification of clinically relevant molecular subtypes within HGG, the genetic alterations contributing to this subtype remain virtually unknown and none has been mechanistically elucidated.

Here we introduce and experimentally validate the “bottleneck hypothesis,” i.e. the concept that master regulators of a tumor subtype implement functional bottlenecks that canalize and integrate aberrant upstream signals from a spectrum of driver genetic alterations, thus constituting a central dependency of the phenotype (oncogene or non-oncogene addiction[8]). Specifically, if the co-activation of *C/EBP* and *STAT3* is both necessary and sufficient to establish an MGES subtype, then relevant genetic alterations must be harbored either by the master regulators themselves or by their upstream functional regulators. Integration of copy number alteration data and regulatory network analysis upstream of these MR genes confirmed this hypothesis, leading to the identification and functional validation of the focal amplification of the *C/EBPδ* locus and deletion of the *KLHL9* locus, which account for the majority of mesenchymal and worst prognosis samples in glioblastoma (GBM). Additionally, this analysis suggests a possible functional role for *NF1* as an upstream post-translational regulator of *STAT3*. Indeed, its inactivation by deletion and mutation co-segregates with *C/EBPδ* amplifications, consistent with the established mechanism of synergistic mesenchymal signature activation in GBM, which requires co-activation of *C/EBP* and *STAT3* transcriptional activity (Supplemental Figure 9).

RESULTS

Identifying Functional Copy Number Variations

To reduce the number of candidate copy number variations (CNV) that may play a functional role in dysregulating MRs of the MGES signature, thereby regulating a major molecular program associated with poor prognosis in GBM, we defined functional CNV Genes (f-CNVGs) as genes within CNV loci whose copy number was informative of their expression level by either mutual information analysis or t-test statistics, see methods section. The analysis was performed using a set of 229 TCGA samples for which both gene expression and copy number profiles were available from Affymetrix (HU-133) and Agilent (CGH) respectively, with no subtype selection. Only genes passing these criteria were subsequently considered as candidate MGES-relevant genetic alterations.

The analysis identified 1486 f-CNVGs at a p-value $p < 0.05$ (Bonferroni corrected); see Figure 1b. f-CNVGs that frequently co-occur in the same samples were grouped into 34 clusters, based on sample co-segregation analysis, see methods. As expected, cluster membership was mostly determined by genomic proximity, since f-CNVGs at distant loci were relatively independent of each other, except for cases where large fragment of a chromosome were recurrently deleted or amplified. For instance, f-CNVGs on chromosome 9, which is frequently deleted in GBM, were clustered together. However, they were statistically independent of f-CNVGs on chromosome 12 (Figure 1a).

Based on this metric, the vast majority of CNVs did not appear to affect expression of their corresponding genes (see for instance Supplemental Figure 2a). We thus tested whether this filtering may be too conservative by checking for the successful identification of established oncogenic drivers as f-CNVGs. The vast majority of established GBM genetic alterations were preserved by the analysis, including 14/18 gene copy number alterations previously identified as classical GBM tumorigenesis drivers[5], such as EGFR, CDK4, PDGFRA, MDM2, MDM4, MET, AKT3, MYCN, PIK3CA, CDKN2A, CDKN2C, RB1, PTEN, and NF1, see Supplemental Figure 10. The remaining four genes were not identified as f-CNVGs due either to insufficient analytical power, because the corresponding CNV frequencies were too low for statistical association with the corresponding gene expression, or because there was no evidence of differential expression due to the CNV. For example, the CDKN2B locus was omitted as a candidate f-CNVG, despite a high frequency of deletion and linkage to CDKN2A because CDKN2B expression was not detected in these GBM samples. Previous studies that lack this selection step would consider this locus as a candidate gene based on the genomic array data, despite the fact that the expression array data precludes it from functionally altering the molecular behavior of these tumors. Supplemental Figure 2b, for instance, shows the strong association between the CNV harboring the EGFR gene and its mRNA expression. Additionally, among previously identified MGES MR genes (*C/EBPβ/δ*, *STAT3*, *FOSL2*, *BHLHB2*, and *RUNX1*), only *C/EBPδ* was identified as an f-CNVG by this analysis, based on coordinated amplification and overexpression in ~22% of the samples; see Figure 1. This suggests that aberrant functional activation of *C/EBP* and *STAT3* most frequently arise from upstream genetic or epigenetic events rather than from direct amplification events.

Identification of f-CNVGs as candidate modulators of MGES

To identify f-CNVGs that may drive the aberrant activity of MGES MR genes, we applied two complementary approaches to the TCGA data. First the MINDy algorithm[9] was used as a genetic-genetic approach to identify all upstream candidate functional modulators of the transcriptional activity of *C/EBPβ*, *C/EBPδ*, and *STAT3*. Inferred modulators that were also detected as f-CNVGs or that harbored mutations reported in [5] were then selected for further analysis. Succinctly, MINDy tests whether the expression of a candidate modulator *M* may affect the strength of the regulatory relationship between a TF and its targets t_i. This is accomplished by computing the conditional mutual information $I = [TF; t_i \mid M]$ between TF and target, given the modulator availability. MINDy was successful in the identification of both known and novel modulators of MYC in human B cells, which were experimentally validated[9], and in the analysis of interactions between all signaling proteins and TFs also in human B cells[10].

Concurrently, we used a genetic-genomic approach, inspired by [11], to find f-CNVGs whose presence in specific samples would be associated with master regulator activity, measured as a function of MES marker expression. This was accomplished by computing the mutual information between each f-CNVG and the of the established MGES MR genes, including *C/EBPβ*, *C/EBPδ*, *STAT3*, *FOSL2*, *BHLHB2*, and *RUNX1*. MR activity in each sample was inferred from the expression of their ARACNe-inferred transcriptional targets[1], see methods section. The combination of these two analysis identified 184 of the original 1486 f-CNVGs as candidate modulators of MGES MR genes by either analysis, and 41 of those 184 were identified by both methods (see Supplementary Table 8). These 1486 f-CNVGs were then clustered into co-mutated groups via simple pairwise statistical association methods, revealing that the majority of these mutations fell into linked regions on various chromosomes (Figure 1a).

Functional amplifications of C/EBPδ and deletions of KLHL9 associate with MGES subtype

The 184 f-CNVGs emerging from the previous analysis were then tested for actual association with the MGES molecular subtype. The subtype classification was established using a signature-based sample classification, as described in [1] _ENREF_3, see methods section. Specifically, samples were classified as either MGES or non-MGES, and the association of samples bearing mutations at each f-CNVG to each of the subtypes was assayed. Consistent with MR analysis, all but two of these f-CNVGs displayed statistically significant enrichment of the expression MGES marker genes.

These loci were further tested to identify the loci with the greatest effect size associated with the mesenchymal tumors. A recursive analysis was performed to determine the smallest subset of the f-CNVGs that was both maximally associated with the MES subtype, and accounted for the maximum number of MES tumor samples; see the methods section, Testing for CNV association by recursion, for details.

Two f-CNVGs emerged as having both high effect size and statistically significance to the MGES subtype by genomic and molecular genetic association. These include the single-gene focal amplifications of the *C/EBPδ* locus on chromosome 8, in 31 of 144 (22%) MGES samples vs. 1 of 51 non-MGES ($p \leq$ 2.1E-5), and the deletion of the *KLHL9* locus on chromosome 9, in 55 of 144 (38%) MGES samples vs. 3 of 51 non-MGES (p ≤ 8.14e-7). Of the 144 MGES samples, 17 (12%) presented genetic alterations at both loci (synopsis in

Figure 1b). Overall, this resulted in highly differential distributions, with significant independent *p*-values for $C/EBP\delta^{amp}/KLHL9^{\mathrm{WT}}$ ($p \leq$ 9.9E-3) and $KLHL9^{-}$ $/C/EBP\delta^{\mathrm{WT}}$ (*p*). Overall, 69 of 144 MGES samples (48%) presented at least one of the two genetic alterations vs. only 4 of 51 non-MGES samples ($p \leq$ 1.15e-9). This result was based on a highly conservative threshold, normally used for genome-wide CNV inference, suggesting that these alterations may be even more frequent.

One additional test was implemented to ensure that *KLHL9* and *C/EBPδ* were the most likely drivers among the mutations that they co-occur with in patients. In particular, *KLHL9* is located in a frequently deleted chromosomal region that includes *CDKN2A* (p16), one of the most frequently deleted tumor suppressors in GBM. It is thus legitimate to ask whether the identification of *KLHL9* as an association with the MGES subtype may be an artifact due to its proximity to *CDKN2A*. To address this issue, all mutations that statistically tended to co-occur with these two mutations in patients (obtained from Figure 1a, shown in Figure 2a) were tested for association to the MES subtype given the absence of another co-occurring gene. This procedure is explained in greater detail in the Methods section: Testing for candidate f-CNVGs among co-mutated clusters. Figures 2b and 2c synopse the results of the analysis. This analysis revealed that, of all the tested f-CNVGs, only those for *C/EBPδ* and *KLHL9* could abrogate the MGES association of every other cluster f-CNVGs, while still showing significant conditional association across virtually all tests. Indeed, no other f-CNVG was statistically significant after conditioning the analysis on these two alterations. These results suggested that the two genetic alterations were the most likely causal ones among those considered in the analysis. We conclude from this that the mutations of *KLHL9* and *C/EBPδ* are in fact the most likely drivers of the MES phenotype based on a consensus of genetic, genomic, and associative analysis.

This analysis was also conducted on genes selected by chromosomal proximity to the *C/EBPδ* and *KLHL9* loci. Supplementary Figure 8 shows that statistical association of the *KLHL9* deletion to the MGES subtype is substantially increased when only *CDKN2A* deleted samples are considered (Supplementary Figure 8b, blue line), while the statistical association of *CDKN2A* deletions with the MGES subtype is completely abrogated when conditioned on the absence of *KLHL9* deletions (Supplementary Figure 8b, red line). This suggests that *KLHL9* deletions rather than *CDKN2A* deletions account for the association of this genomic region with the MGES.

Interestingly, deletions and mutations of the *NF1* gene were also associated with MGES samples. However, these events tended to co-occur with *C/EBPδ* amplification and were not statistically significant following conditional association analysis. Since *NF1* was inferred by MINDy as a *STAT3* modulator and since activation of both *C/EBP* and *STAT3* is necessary for reprogramming along the mesenchymal lineage, this suggests that these two events may cooperate, i.e.

that MGES samples harboring the *C/EBPδ* amplification may also harbor *STAT3* activating alterations, including *NF1* deletions and mutations[7].

Using a stringent threshold for their identification, *C/EBPδ*+ and *KLHL9*-alterations account for 48% of the MGES samples in the TCGA dataset, with deletions/mutations at the *NF1* locus covering an additional 8% independently of *C/EBPδ* amplifications or *KLHL9* deletions[7], suggesting that these may constitute the most common alterations associated with the MGES subtype of GBM, especially since many mutated samples may be missed duet to the conservative threshold selection to minimize false positives.

Alterations of C/EBPδ and KLHL9 predict poor prognosis independently of molecular classification

Since alterations of *C/EBPδ and KLHL9* were both associated to and predicted to regulate the mesenchymal signature, the molecular predictor of poor prognosis GBM,_we tested whether or not these mutations are sufficient to predict poor prognosis. We obtained an independent set of genomic DNA from 63 primary GBM tumor samples provided by Ken Aldape and assayed them for deletions of KLHL9. We used Kaplan-Meier statistics on the original TCGA dataset to test whether alterations in *C/EBPδ* and *KLHL9* were also good predictors of poor patient prognosis, independently of prior molecular classification, as originally observed. The Kaplan-Meier survival curve of samples with mutations in either of these loci differed significantly from the survival curve of “good” prognosis GBM patients (Figure 3a, $p \leq 3.46e-4$). Additionally it was statistically significantly distinct from a cohort of all samples that are diploid at these two loci, regardless of molecular- or prognosis-based classification (Figure 3a, $p \leq 0.0319$).

The *C/EBPδ* gene codes for one of the master regulators that induce direct activation of MGES genes[1]. Thus the mechanistic relevance of its amplification in mesenchymal samples is obvious. Conversely, the mechanism by which deletion of *KLHL9* drives expression of mesenchymal genes in glioma is unknown. Thus, we set out to determine the functional significance of *KLHL9* deletions for mesenchymal transformation of HGG.

KLHL9 deletions are enriched in an independent cohort of poor prognosis patients and predict elevated CEBP protein levels

We asked whether *KLHL9* deletions might be frequently found in an independent cohort of poor-prognosis glioma samples, compared to good prognosis ones. We analyzed the status of the *KLHL9* genomic locus by quantitative genomic PCR (qgPCR) from a set of 63 formalin-fixed, paraffin-embedded (FFPE) primary glioma samples collected at the MD Anderson Cancer Center from two separate cohorts. These included 10 poor-prognosis (<35 weeks survival) and 9 good-prognosis (>130 weeks survival) samples. The other primary samples were samples obtained from the TCGA that were not part of our original analytical set, also classified into good and poor prognosis. qgPCR analysis revealed a significantly higher frequency of *KLHL9* homozygous deletions in poor-prognosis samples (21/40) compared to good prognosis samples (4/23) (Figure 3a, b), resulting in a very significant p-value (p < 5.6e-3 by FET), see Figure 3b. This suggests that *KLHL9* may be frequently deleted (>50%) in poor prognosis samples, above the frequency determined by a stringent cutoff in TCGA CNV data analysis (38%). Genomic DNA sequencing of the samples lacking deletion of *KLHL9* failed to reveal the presence of mutations in the coding sequence of *KLHL9*.

Concurrently, we performed IHC assays to check the protein levels of the master regulators CEBPβ and C/EBPδ. We observed that, as shown before, MES GBM tumors are characterized by unique expression of these two proteins. We were subsequently able to show that KLHL9 deletions strongly predict the presence of mesenchymal levels of CEBPβ and C/EBPδ (odds ratio 12.25, p=0.0283) on a cohort of 20 primary samples tested, shown in Figure 3d.

Re-expression of KLHL9 in KLHL9$^{-/-}$; CDKN2A$^{-/-}$ human glioma depletes C/EBPβ and C/EBPδ proteins.

To assess whether *KLHL9* deletions activate the function of the previously validated MRs of the MGES subtype (*C/EBPβ*, *C/EBPδ*, and *STAT3*), we asked whether restoring the expression of *KLHL9* in cells carrying homozygous deletion of the endogenous *KLHL9* gene may affect expression of their mRNAs and/or proteins. From the genomic analysis of *KLHL9*, *CDKN2A*, *C/EBPδ* and *EGFR* genes in eight human glioma cell lines we found that the SF210 cell line harbors homozygous deletion of both *KLHL9* and *CDKN2A* whereas all the glioma cell lines were diploid at the *C/EBPδ* locus (Supplemental Figure 4). Thus, the SF210 line provides an ideal cellular system to investigate the functional consequences of *KLHL9* restoration in *KLHL9* and *CDKN2A* double-deleted glioma.

We used a lentiviral vector to transduce SF210 with a doxycycline inducible full-length *KLHL9* gene*.* We selected two SF210 stable clones showing (DOX)-induced expression of *KLHL9* mRNA 48 hours after induction this effect was sustained for at least 96 hours (Figures 4a). Consistently, *KLHL9* protein levels were stably detected by western blot, up to 96 hours post induction (Figure 4c). An inducible GFP clone was also validated and used as a control in all subsequent experiments.

RNAseq experiments on these cells revealed that 48 hours of restored KLHL9 expression coincided with a significant shift in expression of CEBP-predicted transcriptional targets (Figure 4b) compared to GFP mock transfected cells. Furthermore, the mesenchymal marker genes predicted to be regulated by either CEBP shifted to suppressed expression, despite no significant changes in the expression levels of either *C/EBPβ* and *C/EBPδ* (inset 4b), including genes such as YKL40 and FN1.

While their mRNA levels remained unchanged, *KLHL9* expression coincided with markedly decreased protein levels of master regulators *C/EBPβ* and *C/EBPδ* but not of *STAT3* (Figure 4c). Additionally, ectopic *KLHL9* expression triggered similar down-regulation of the positive control AuroraB protein, which is known to be destabilized and degraded by a *KLHL9*-containing, cullin3-based E3 ubiquitin ligase complex[12,13]. No equivalent protein changes were detected when the GFP control SF210 cells were treated with DOX. Taken together, these results indicate that re-expression of *KLHL9* induces the suppression of MGES marker genes via the loss of the two MGES master regulators *C/EBPβ* and *C/EBPδ* at the protein level.

Furthermore, the suppression of *C/EBPβ/δ* protein levels was observed in a *CDKN2A* null background, thus confirming that *CDKN2A* deletion in isolation is not sufficient to maintain high protein expression of the master regulators. This suggests that deletion of *KLHL9* is sufficient to activate the two previously validated master regulators of the MGES, thus significantly contributing to the induction of mesenchymal transformation in GBM, independently of *CDKN2A* expression.

KLHL9 promotes poly-ubiquitylation and proteasomal-mediated degradation of C/EBPβ and C/EBPδ

Given that *KLHL9* is an adaptor of cullin 3-based E3 ligases[13], and the observation that *C/EBPβ/δ* proteins decrease without change in their mRNA levels following *KLHL9* expression, we tested whether ectopic expression of *KLHL9* in glioma cells may trigger ubiquitylation-dependent, proteasome-mediated degradation of *C/EBP* TFs. To test this hypothesis, we measured the half-life of *C/EBPβ* and *C/EBPδ* proteins in the presence or absence of *KLHL9* in SF210 cells treated with the proteasome inhibitor MG-132 versus controls. The proteins' half-life was measured while protein synthesis had been abrogated by the translational inhibitor cyclohexamide (CHX). These experiments revealed that the half-life of *C/EBPβ/δ* was markedly reduced from >4 hours in control SF210 cells, lacking *KLHL9* (GFP-expressing clones in the presence of DOX and *KLHL9*-inducible clones in the absence of DOX), to ~1-2 hours in the SF210 cells in which *KLHL9* had been restored by treatment with DOX. Inhibition of the proteasome by MG-132 restored accumulation of the *C/EBP* proteins in the presence of *KLHL9* (Figure 5a). The results indicate that re-expression of *KLHL9* in glioma cells triggers proteasome-mediated degradation of the *C/EBP* TFs. Furthermore, an interaction was detected between the CEBP proteins and the KLHL9 protein, as assayed by a co-immunoprecipitation using KLHL9 to pull down the CEBPs (Figure 5b).

To test whether proteasome-mediated degradation of *C/EBPβ/δ* proteins by *KLHL9* was also ubiquitylation-dependent, we prepared cell lysates in the presence of MG-132, with and without *KLHL9* expression, and tested for ubiquitylation of immunoprecipitated *C/EBPβ* and *C/EBPδ* by western blot. Following expression of *KLHL9* and proteasomal inhibition, poly-ubiquitylated *C/EBPβ* and *C/EBPδ* were significantly increased in comparison to uninduced controls, see Figure 5c, thus confirming that *KLHL9* promotes both poly-ubiquitylation and proteasome-dependent degradation of *C/EBP* TFs.

Rescuing with a mutant KLHL9 protein suppresses ubiquitylation of CEBPs

Finally we cloned a KLHL9 protein bearing a 70 aa deletion in the N-terminal end of the protein corresponding to the cullin-interacting BTB domain of the protein (Figure 6a). This domain is responsible for bringing the ligase/target complex to the cullin scaffold, which also brings in an E2 adaptor bearing ubiquitin, mediating the transfer of the ubiquitin to the target. Upon exogenous rescue with this mutant construct, we successfully abrogated both the detection of ubiquitylated CEBP species upon immunoprecipitation (Figure 6b) to levels that match a GFP-transfected, KLHL9 null molecular behavior, as well as the suppression of CEBP proteins 48 hours post expression (Figure 6c).

KLHL9 expression suppresses the proliferation of glioma cells

Expression of *C/EBP* TFs and presentation of a mesenchymal phenotype are hallmarks of aggressiveness in HGG. We thus assayed the effects of *KLHL9* expression on cellular growth over 96 hours in the DOX-dependent, *KLHL9*-expressing SF210 clones.

Immunofluorescence microscopy, following *KLHL9* induction, revealed the emergence of large, extensively spread cells with enlarged nuclei that failed to incorporate EdU (red signal), compared to uninduced controls (Figure 7a), suggesting that *KLHL9* expression may suppress mesenchymal glioma cell proliferation. These large cells appeared only upon induction of *KLHL9* and accounted for 38% of the cell population. Less than 5% of these large cells had any detectable EdU signal, compared to 70% incorporation frequency observed in the GFP control cells. To further quantify this effect, we measured BrdU incorporation via flow cytometry. Cells expressing *KLHL9* for 48 hours (Figure 7b, red series) showed a significant reduction in BrdU incorporation relative to uninduced controls (Figure 7b, black series) following a 24-hour BrdU pulse, based on integrations of the area under the BrdU-positive and -negative peaks. To corroborate this observation, we also measured cell growth by normalized cell counts of DOX-induced clones versus DOX induced GFP clones and uninduced controls over a 96-hour timecourse. DOX treatment of GFP controls did not significantly alter the growth of the cells, whereas expression of *KLHL9* correlated with a significant decrease in cell growth that was detectable at 72 hours post-induction, and was maintained through at least 96 hours (Figure 7c).

DISCUSSION

Analysis of large CNV and mutation datasets is providing an extraordinary window over the genetic events that underlie tumorigenesis and tumor progression. Unfortunately, the number of genetic alterations that are statistically associated with most solid tumors tends to be very high, due also to recurrent large-scale genomic rearrangements. As a result, an increasing challenge of cancer research is to be able to separate driver mutations from passenger ones.

Equally importantly, extensive knowledge on established oncogenes and tumor suppressors usually hampers elucidation of genes located in their chromosomal proximity as viable candidate driver mutations. For instance, even though we showed that *KLHL9* is frequently mutated in the MGES subtype of glioblastoma, its proximity to *CDKN2A* prevented it (as well as many other genes in that region) from being previously considered as independent contributions to the subtype etiology.

In contrast, regulatory network based analysis established KLHL9 as an ideal candidate for functional validation, independent of its proximity to CDKN2A, because of its computationally inferred role as a strong modulator of MGES MR activity. Not only could we elucidate the specific mechanism by which *KLHL9* modulates turnover of *C/EBP* TFs, by poly-ubiquitylation dependent proteasomal degradation, but analysis of an independent cohort of poor versus good prognosis GBM patients revealed this gene as even more frequently deleted than originally suspected from TCGA data analysis (>70% versus 38%). This suggests that current thresholds for mutational analysis may be over-conservative, likely to minimize false positives detection in genome-wide studies, and that more realistic threshold could be used if the number of candidate genes could be reduced via regulatory network based approaches, as shown in this study.

Recently, integration of regulatory network based approaches with GWAS data has been successful in identifying a handful of phenotype-relevant genetic alterations[14]. However, this analysis has always proceeded in a genome wide fashion, thus requiring highly conservative thresholds for evaluation of statistical significance. In this manuscript, we presented and validated the “bottleneck hypothesis,” i.e., that some cancers are characterized by functional bottlenecks, implemented by master regulator TFs, which integrate aberrant signals originating from a spectrum of genetic and epigenetic alterations in their upstream regulators. Under such assumption, analysis of genetic alterations in the master regulators and in their upstream regulators can elucidate key genetic alterations that would have otherwise been missed. Interestingly, master regulator bottlenecks may fail to harbor genetic alterations, making their identification difficult by conventional mutational analysis approaches. For instance, using gene candidate approaches, we have previously elucidated Nf-κB as a master integrator of aberrant events in its upstream pathways within the ABC subtype of Diffuse Large B Cell Lymphoma even though it is never itself mutated[15]. Similarly, of the three MGES master regulators in GBM, only *C/EBPδ* was significantly amplified. Thus, despite their critical functional role, none of these key genes could have been identified by traditional mutational or copy number variation analysis.

A first corollary of the bottleneck hypothesis is that regulatory network based analysis of genes that are upstream functional regulators or modulators of master regulators of a tumor subtype may be much more effective in providing candidates driver mutations than unconstrained GWAS. Importantly, as shown, the collapse in the number of candidate mutations made possible by regulatory network analysis allows efficient use conditional association methods to separate driver from passenger mutations. This approach would be completely implausible if it had to be performed on all mutations that are statistically associated with a phenotype of interest.

A second corollary of the bottleneck hypothesis, especially when combined with the results of [1], is that individual genetic events, such as *C/EBPδ* amplifications or *KLHL9* deletions, may be too rare or unlikely to provide appropriate targets for pharmacological intervention. Conversely, by integrating an entire spectrum of aberrant signals from upstream genetic alterations, functional bottlenecks implemented by master regulators may constitute more universal biomarkers and pharmacological targets (i.e., universal oncogene or non-oncogene addition points of the cancer subtype) because of their ability to integrate the effect of many low-frequency mutations.

A final corollary is that the bottleneck hypothesis may help identify key genetic alterations that are either not focal or are harbored by genes located in close chromosomal proximity to well-established oncogenes and tumor suppressors. In the past, the approach has been to simply ignore such genes to reduce false positives. Yet, there is no functional reason why genes within large, frequently deleted or amplified regions or in close proximity to established oncogenes should be less likely to be drivers of the phenotype. Indeed, regulatory network based analysis was successful in identifying the role of *KLHL9* deletions, which are both non-focal and in close proximity to CDKN2A, one of the most frequently deleted genes in GBM.

KLHL9 deletions in HGG result in mesenchymal transformation because of aberrant stabilization of the master regulator C\EBP TFs. At least two other genes coding for E3 ubiquitin ligases undergo loss-of-function genetic alterations in HGG. The first gene codes for Fbw7, an F-box protein of the SCF complex that is mutated in several forms of human cancer including HGG[16]. Fbw7 mutations stabilize the oncoprotein substrates cyclin E, Myc and Notch[17]. The second gene coding for an E3 ligase, which can be deleted in HGG, is Huwe1, a Hect-domain ubiquitin ligase that normally triggers initiation of differentiation and loss of self-renewal in the developing brain by targeting the N-Myc oncoprotein for ubiquitin-mediated degradation by the proteasome[18]. Our findings indicate that loss-of-function events targeting E3 ubiquitin ligases in human cancer not only promote aberrant stabilization of classical oncoproteins thus contributing to cancer development but they can also trigger accumulation of key TFs responsible for specific tumor signatures and aggressive phenotypes.

Clearly, the ability to identify both cancer bottlenecks and their candidate upstream functional regulators depends critically on the availability of accurate and comprehensive repertoires of cell-context specific molecular interactions (interactomes). While the assembly of integrated transcriptional, post-

transcriptional, and post-translational interactomes is still in its infancy, the genome-wide integration of experimental and computational approaches appears to be providing increasingly descriptive and biologically relevant models, suggesting that network based biology may be an increasingly valuable tool in our repertoire of approaches to elucidate the mechanism of key physiological and disease related processes.

Keywords: [Insert MeSH Keywords 4-8 in Results section]

ACKNOWLEDGEMENTS

This work was supported by NCI Grand Opportunities TDDN Network 1RC2CA148308-01 (A.C.), In Silico Research Centre of Excellence NCI-caBIG 29XS192 (A.C.), National Centers for Biomedical Computing NIH Roadmap Initiative U54CA121852 (A.C.), National Institute of Neurological Disorders and Stroke (NINDS) grant R01NS061776 (A.I.). F.N. is supported by a fellowship from the Italian Ministry of Welfare/Provincia di Benevento.

AUTHOR CONTRIBUTIONS

Andrea Califano (AC) conceived the overall computational and experimental framework for the elucidation and experimental validation of genetic alterations upstream of master regulator modules and wrote the manuscript. James C. Chen (JCC) and AC conceived and designed the genetical-genomics analysis, the recursive association methods, and the classification methods; JCC implemented all the computational algorithms necessary for the analysis of the data; JCC also designed and performed all biological experiments with the training and guidance of Mariano J. Alvarez (MJA) and under the supervision of AC, performed data analysis, and wrote the Results, Methods, Figures, and Supplemental Figures sections of the manuscript, providing editing throughout. MJA additionally retrieved and performed quality control and normalization of the Affymetrix gene expression data, contributed to all relevant discussions, and provided editing for the manuscript. Gabrielle Rieckhof (GR) performed additional cell growth experiments and provided editing for the manuscript. Archana Iyer (AIyer) helped perform a gQPCR experiment and assisted in preparing and performing the FACS. Francesco Niola (FN) provided cell pellets for all huGBM cell lines used in the project, as well as the SF210 cell line used for KLHL9 cloning, under the supervision of AI and AL. Kristin L. Diefes (KLD) isolated the DNA from the MD Anderson human GBM cohort, Kenneth Aldape (KA) identified the samples in the MD Anderson cohort and supervised the work of KLD.

METHODS

TCGA data processing

Somatic copy number variation (Agilent) and gene expression data for 229 TCGA tumor samples were downloaded from the TCGA Data Portal. Clinical data for a subset of these patients was also acquired from the Data Portal. The gene expression arrays were GCRMA-normalized, and the copy number variation log ratios were extracted from the CNV array data.

Agilent CGH arrays were used instead of the also available Affymetrix SNP array data because probe coverage of key loci was sparse in the latter; For instance, the *C/EBPδ* locus had no probes within the coding region of the gene, and associations that were detected in the CGH arrays were not detectable in the Affymetrix data without using more sophisticated, sliding-window integration methods across probes in the region (Supplemental Figure 5). Additionally, overall stronger CNV – gene-expression dependencies were detected using the Agilent CGH array data. For instance, Affymetrix probes proximal to the *C/EBPδ* locus showed no correlation with *C/EBPδ* expression, without sliding window integration, and were overall less correlated than those reported by CGH arrays (Supplemental Figure 6).

Inference of Functional CNV Genes (f-CNVG)

f-CNVGs were identified by integrating gene expression and copy number variation using a statistical test based on the Mutual Information (MI),

$I_F = I[CNV_i; mRNA_i]$ To allow identification of low-frequency genetic alterations, genes lacking significant MI between their CNV and their expression were also tested for differential expression conditional on copy number changes, using a by T-test for all alterations with >1 prevalence and a Z-test for single-occurrence mutations.

The dependency between CNV at a locus *x* and a gene *y* at the same locus was measured based on the pair-wise mutual information between the vector of gene expression values of *y* across all samples*,* and the vector of CNV log ratios of *x* across the same samples: $MI[x;y]=\Sigma_x\Sigma_y\log(p(xy) / p(x)p(y))$, using a Gaussian kernel estimator. Values of *MI[x;y]* that were statistically significant at a p-value $p < 0.05$, Bonferroni corrected for the total number of tested pairs, were used to identify candidate f-CNVGs. The statistical threshold for *MI* significance was determined from a null distribution built by computing the *MI* between 10,000 randomly paired CNV and gene expression vectors (Supplemental Figure 3).

This approach offers two advantages. First, it can detect statistical dependencies originating from non-linear relationships between the two vectors that may be missed by other measures of statistical independence, such as Pearson correlation. It also removes the need for three independent statistical significance tests per gene: one to detect significant gene differential expression, one to detect a significant CNV, and one to assess significant correlation between the two. This allows for increased statistical power (Supplemental Figure 2c).

f-CNVG clustering

All f-CNVGs identified by the previous test were then clustered based on same-sample co-segregation. This was done using Fisher's Exact Test on the number of overlapping samples presenting the alteration, at a $p < 0.05$ statistical significance threshold, Bonferroni corrected for the number of multiple tests. All f-CNVG pairs that showed significant correlation are connected by an edge in Figure 1b. Clusters of co-mutated f-CNVGs were identified by higher association scores between genes in the cluster than between those genes and genes outside of the cluster; genes that had a much higher probability of being co-mutated clustered together when using the association p-value as a metric.

For each inferred cluster of co-segregating f-CNVGs, we computed the mutual information between the corresponding f-CNVGs and the activity of each of the mesenchymal master regulators, *C/EBPβ/δ*, *STAT3*, *FOSL2*, *BHLHB2*, and *RUNX1* originally identified as MRs of the MGES signature. The mutual information was computed and tested as discussed for the CNV – gene-expression case, using testing for statistical enrichment of each MR's targets, as identified from the ARACNe-inferred transcriptional networks[19,20] rather than gene expression (Supplemental Figure 2a).

Network-Based Association Study: Testing for f-CNVG association by recursion

Following classification of the TCGA GBM tumors into poor- and good-prognosis, the candidate mutations identified previously enriched in differential expression of the MES genes were tested for association to the MGES and poor-prognosis phenotype recursively. Across all available samples, each f-CNVG was tested for association to the MES subtype. The f-CNVG with the highest association across all comparisons was identified, and all patient samples bearing that f-CNVG were removed from the dataset. This association analysis was then repeated to identify the next highest-association f-CNVG until no additional significant associations could be identified.

Network-Based Association Study: Testing for candidate f-CNVGs among co-mutated clusters

Once a candidate f-CNVG for the MES subtype was identified, it was subjected to an additional analysis to account for the possibility that its association is an artifact of another mutation that co-occurs with it in patients. In order to test this an analysis was designed under the following hypothesis: among all co-mutated genes, only the true, causal mutation will remain associated to the molecular subtype across various genetic backgrounds. Therefore, all of the mutations that were found to statistically co-occur with a candidate driver (obtained from the association map in Figure 1A) were conditionally tested for association to the MES subtype, given that another gene in the co-mutated cluster was not mutated. This pair-wise analysis was performed for every pairing of genes in the co-mutated cluster, searching for a mutation that, when removed from the background, caused the association of all other genes in the cluster to become insignificant.

Classification of TCGA GBM tumors

TCGA tumor samples were reclassified into poor-prognosis and good-prognosis phenotypes based on the activity of master regulators originally reported as drivers of the most aggressive subtypes of GBM: the C/EBP*β/δ*, *STAT3*, *FOSL2*, *BHLHB2*, and *RUNX1* (Supplemental Figure 1a). Molecular classification by the activity of these genes via a centroid-based, nearest-neighbor classifier produced two groups separable by prognosis at a statistically significant level (Supplemental Figure 1b), and served as the basis for subsequent associative analyses. Clustering TCGA tumor samples by prognosis and testing for differential activity of both MGES master regulators and signature genes recapitulates the finding that these genes are accurate predictors of poor prognosis. See figure 1b).

Genomic KLHL9 copy number characterization in an independent HGG cohort

Genomic DNA was extracted from ten poor-prognosis (post-diagnosis survival <35 weeks) and nine better-prognosis (>135 weeks) paraffin-embedded HGG obtained from the MD Anderson Cancer Center and tested for copy number changes of relevant genes by quantitative genomic qPCR. The copy-number status of the *KLHL9* gene was analyzed by quantitative amplification of two 200-bp amplicons, at the 5' and 3' ends of the *KLHL9* coding sequence respectively, according to the methods discussed in the qRT-PCR Methods section. The entire coding sequence of the *KLHL9* locus was also sequenced to scan for possible mutations from samples that showed successful amplification.

Plasmid constructs

Bacterial cultures were grown on agar plates with appropriate selection at 28C. Transformations into DH5α cells (Invitrogen) were performed using the recommended protocol.

The coding region of the *KLHL9* locus was amplified from genomic DNA obtained from 293T cells using the following primers: *KLHL9*-BsshII-F (5'-GGCAGCGCGCatgaaagtgcccttggtaacg-3') and *KLHL9*-XhoI-R (5'-GCGCTCGAGctaagaatgatctgaaggtgctga-3') with the AccuPrime TAQ system (Invitrogen). This PCR product was digested with BssHII and XhoI and ligated into the pEN_TTmcs inducible expression vector with the Rapid DNA Ligation Kit (Roche) according to the kit's protocol.

After sequencing for mutations, the *KLHL9* locus insert was introduced to the lentiviral packaging vector pSLIK via Gateway cloning (Invitrogen). A GFP-pEN_TTmcs was also cloned to a pSLIK vector and included as a negative control for all subsequent cell culture work.

Cell lines/Cell culture

SF210 and 239T-FT cells were grown in DMEM +10%Fetal Bovine Serum (Gibco,BRL), incubated at 37C with 5% CO2.

Stable, inducible *KLHL9* and GFP-SF210 cells were generated by transfecting the appropriate pSLIK vectors and supplementing plasmids into 239T-FT cells with JetPEI Transfection Reagent (Polypus Transfection). 24 hours post-transfection, the virus-bearing medium was aspirated off, vacuum-filtered, and placed over pre-confluent SF210 cells. After 48 hours of infection, SF210 cells were placed under G418 selection at 1 mg/ml for 7 days.

KLHL9-infected SF210 cells were then cloned via dilution limit to obtain monoclonal cell populations. GFP-control transfected cells were left as a polyclonal population. Cells were then checked for *KLHL9* or GFP expression by induction with 2ug/ml doxycycline (Sigma) for 24 hours. GFP production in GFP controls was verified by fluorescent microscopy, and *KLHL9* expression was verified by qRT-PCR at 24 hours, and Western Blotting at 72 hours. Isolated clones were maintained with 200ug/ml G418 while growing for subsequent experiments.

qRT-PCR

Total RNA was prepared from cells using the Cells-to-cDNA kit (Ambion), and reverse-transcribed to cDNA via first-strand cDNA synthesis using the qScript cDNA kit (Quanta Biosciences) according to manufacturer protocols. Real-time PCR was performed using SYBR Green PCR Master Mix (Applied Biosystems). DNA samples were run in biological triplicates and technical duplicates. Comparative fold changes were computed using the δδCT method normalized to internal controls of GAPDH expression.

RNAseq experiments

Total RNA from six samples (3 each of KLHL9-rescued and mock-rescued) were prepared via TriZOL precipitation and purified using Qiagen RNeasy columns. Samples were tested for integrity via Bioanalyzer and submitted to the Columbia Sequencing center.

Differential analysis was performed using a t-test. These p-values were used as the ranking for a subsequent gene set enrichment analysis (GSEA) to ascertain whether specific biomarker sets were differentially expressed when KLHL9 was rescued or not.

Western Blotting

Cell lysates were prepared from SF210-*KLHL9* and SF210-GFP cells after 72 hours of either doxycycline treatment or control medium by lysing them in RIPA buffer (Sigma Aldrich) with Complete MINI EDTA-free protease inhibitors (Roche). Lysates were quantified using the BCA Protein Assay (Pierce) following manufacturer protocol.

The antibodies were used at 1:500 (*KLHL9*), 1:10,000 (B-actin), 1:1000 (*C/EBPβ*, *C/EBPδ*, *STAT3*, Ubiquitin), 1:10,000 (*AURKB*, goat-anti-mouse), and 1:20,000 (goat-anti-rabbit). Blocking and antibody incubations were done in SuperBlock T20 TBS Blocking Buffer (ThermoScientific). All antibodies were obtained from Santa Cruz.

Protein half-life time courses

KLHL9-4 (SF210 cells transfected with *KLHL9*) and GFP control (SF210 transfected with GFP) cells were grown to pre-confluence in 10cm plates. Plates were then split into 6-well plates in DMEM-10%FBS with 2ug/ml doxycycline and left for 24 hours. 30 minutes prior to the start of cyclohexamide treatment, one *KLHL9*-4 series was treated with 10uM MG-132 while the others were treated with DMSO (the MG-132 solution used was dissolved in DMSO). After 30 minutes, all cells were treated with a DMEM-10%FBS cocktail of doxycycline (2ug/ml) and cyclohexamide (20uM), and additionally with MG-132 (10uM) where appropriate.

At the end of the time course, cells were washed with ice-cold PBS and scraped from the plates, and lysed in RIPA buffer with protease inhibitors. Lysates were quantified using the BCA Protein Assay Kit (Thermo Scientific), and separated by SDS-PAGE and immunoblotted as previously described.

Densiometric analysis was done using the ImageJ software suite.

Immunoprecipitation

Whole lysates were prepared in either RIPA buffer (ubiquitin IP experiments) or Cell Lysis Buffer from Cell Signaling (Co-IP) from *KLHL9*-induced and -uninduced clones subjected to 24 hours of doxycycline (2ug/ml) treatment. IPs for ubiquitylated protein species were additionally treated with the proteasome inhibitor MG-132 for 4 hours after doxycycline treatment (10uM). *C/EBPβ/δ* proteins were immunoprecipitated using antibodies from Santa Cruz and the DynaBeads G Immunoprecipitation kit (Invitrogen) following manufacturer protocols. Eluted *C/EBPβ/δ* proteins were separated by SDS-PAGE and transferred to nitrocellulose membranes according to standard Western blotting protocols and probed for ubiquitin, and *C/EBPβ/δ* (Santa Cruz).

Generating KLHL9 mutants

The deletion of the BTB domain in KLHL9 was generated using PCR fusion. Primers were designed for the 5’ and 3’ so to generate a full length KLHL9. These primers were then paired with 50-bp primers that were designed to be homologous to the 25 base pairs immediately before and after the BTB domain. After amplifying two separate DNA products corresponding to the KLHL9 fragments before and after the BTB domain, bearing 25-bp homologous ends, these two products were combined into a third PCR amplification reaction with the original full-length product primers. This final PCR reaction fuses the two fragments together, producing in frame KLHL9 DNA without the BTB domain.

This construct was then cloned into the expression vector pLCPX for transfection, alongside clones containing the wild type KLHL9 and the empty pLCPX parent vector for controls.

Cell growth time courses

For an initial time point, and each desired time point, a 6-well plate was prepared by seeding ~500 SF210-*KLHL9* or SF210-GFP cells into the wells. Induced wells were plated with 100ul of DMEM 10% FBS and a final concentration of 2ug/ml doxycycline, while the remaining three received 100ul DMEM 10% FBS. Cell counts for seeding were determined using a Countess automated cell counter (Invitrogen). Cells were seeded in biological triplicates.

At each time point, cell growth per well was quantified by counting the cells on the plate using the Countess cell counter. Growth curves were built by normalizing cell counts at each time point to the counts obtained from the appropriate initial time point.

BrdU Flow Cytometry

SF210; *KLHL9*-4 clones were grown for 96 hours with or without induction via doxycycline according to the methods already provided. After 96 hours, BrdU (Calbiochem) was introduced to the cells at a 1:2000 concentration for 24 hours as instructed by manufacturer protocols.

The following day the cells were fixed in BD Cytofix/Cytoperm reagent (BD Biosciences) and stained with fluorescent anti-BrdU (BD biosciences) according to the manufacturer’s protocols. These cells were then analyzed via flow cytometry (20,000 cells per treatment recorded) and analyzed with the FlowJo software suite. The upper/lower limit for left/right peak intervals was defined using the negative control distribution; all events exceeding the 99th percentile BrdU measure in the negative control distribution were considered BrdU-positive, and the remainder BrdU negative (demarcated by the dotted line). The integrals for these peaks were computed and displayed as [left peak : right peak] percentage ratios.

EdU Immunofluorescent Microscopy

Visualization of EdU incorporation was performed using the Click-It EdU HCS Assay Kit (Alexa 647) from Invitrogen. Cells were seeded at 500, 1000, 2000, and 4000 cells per well in a 96-well plate, in biological triplicates for each treatment (induced and uninduced). Cells were left to grow for 72 hours. After 72 hours, all subsequently exposed to EdU at a 1:2000 concentration for 24 hours.

Cells were then fixed in 4% paraformaldehyde, and both staining and visualization were then carried out according to manufacturer protocols.

Raw grayscale images generated by the microscopy analysis were then colored and composites were created using the ImageJ software suite. All image processing was conducted simultaneously on paired experiment-controls.

References

1 Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318-325, doi:nature08712 [pii]
10.1038/nature08712 (2010).

2 Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110, doi:S1535-6108(09)00432-2 [pii]
10.1016/j.ccr.2009.12.020.

3 Ohgaki, H. & Kleihues, P. Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. *J Neuropathol Exp Neurol* **64**, 479-489 (2005).

4 Phillips, H. S. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157-173, doi:S1535-6108(06)00056-0 [pii]
10.1016/j.ccr.2006.02.019 (2006).

5 TCGA-Consortium. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:nature07385 [pii]
10.1038/nature07385 (2008).

6 Sun, L. *et al.* Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* **9**, 287-300, doi:S1535-6108(06)00084-5 [pii]
10.1016/j.ccr.2006.03.003 (2006).

7 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713, doi:10.1038/nature09270 (2010).

8 Weinstein, I. B. Cancer. Addiction to oncogenes--the Achilles heal of cancer. *Science* **297**, 63-64, doi:10.1126/science.1073096
297/5578/63 [pii] (2002).

9 Wang, K. *et al.* Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat Biotechnol* **27**, 829-839, doi:nbt.1563 [pii]
10.1038/nbt.1563 (2009).

10 Wang, K. *et al.* Dissecting the interface between signaling and transcriptional regulation in human B cells. *Pac Symp Biocomput*, 264-275 (2009).

11 Yang, X. *et al.* Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat Genet*, doi:ng.325 [pii]
10.1038/ng.325 (2009).

12 Maerki, S. *et al.* The Cul3-KLHL21 E3 ubiquitin ligase targets aurora B to midzone microtubules in anaphase and is required for cytokinesis. *J Cell Biol* **187**, 791-800, doi:10.1083/jcb.200906117 (2009).

13 Sumara, I. *et al.* A Cul3-based E3 ligase removes Aurora B from mitotic chromosomes, regulating mitotic progression and completion of cytokinesis in human cells. *Dev Cell* **12**, 887-900, doi:10.1016/j.devcel.2007.03.019 (2007).

14 Califano, A., Butte, A., Friend, S. H., Ideker, T. & Schadt, E. E. Integrative Network-based Association Studies: Leveraging cell regulatory models in the

post-GWAS era. *Nature Genetics, in press* **currently in Nature Preceedings (http://precedings.nature.com/documents/5732/version/1)** (2011).

15 Compagno, M. *et al.* Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature* **459**, 717-721, doi:nature07968 [pii]

10.1038/nature07968 (2009).

16 Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:10.1038/nature07385 (2008).

17 Nakayama, K. I. & Nakayama, K. Ubiquitin ligases: cell-cycle control and cancer. *Nat Rev Cancer* **6**, 369-381, doi:10.1038/nrc1881 (2006).

18 Zhao, X. *et al.* The N-Myc-DLL3 cascade is suppressed by the ubiquitin ligase Huwe1 to inhibit proliferation and promote neurogenesis in the developing brain. *Dev Cell* **17**, 210-221, doi:S1534-5807(09)00293-7 [pii]

10.1016/j.devcel.2009.07.009 (2009).

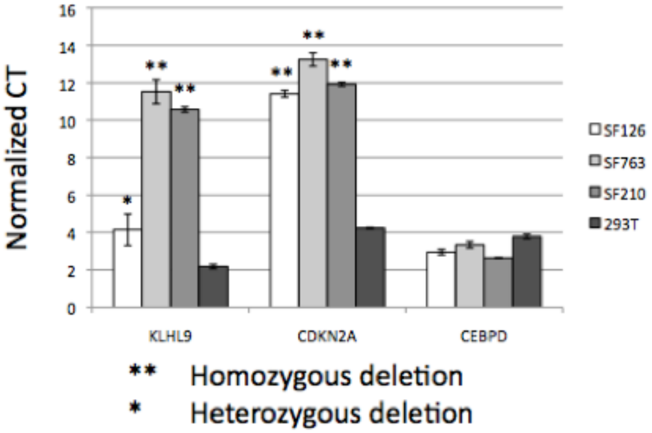
19 Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**, 382-390, doi:ng1532 [pii]

10.1038/ng1532 (2005).

20 Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7 Suppl 1**, S7 (2006).

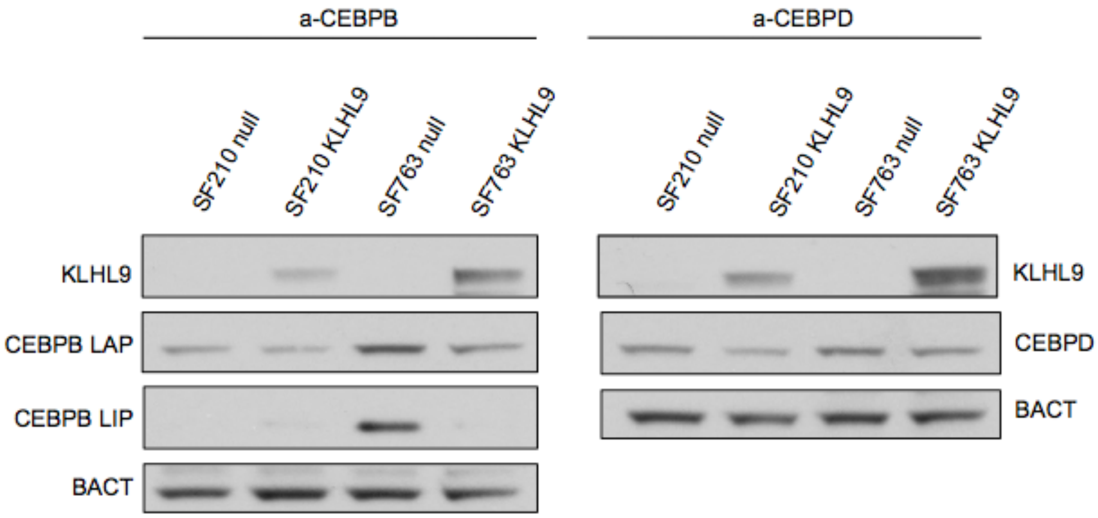
CHAPTER 5a – Supplemental Information to Manuscript / Work in Progress

During the review process for our manuscript, I continued to work on improving the biological validations of the KLHL9 master regulator. I received a fresh selection of huGBM cell lines from the University of California, San Francisco after they performed quality control experiments to ensure that the aliquots matched data generated when they were first isolated. Of the panel of five aliquots was one vial of low-passage SF210, and an additional cell line that was validated as bearing a homozygous co-deletion of KLHL9 and p16: SF763. Verification was conducted using the same genomic qPCR methods described previously, with results shown in [Figure 5a.1]. These validated aliquots were selected to allay issues that may arise from unknown maintenance and accumulated mutations that may have occurred in the original aliquots we received and worked with. The KLHL9 gene was additionally cloned into the lentiviral expression vector pLOC and validated by sequencing to provide an independent vector with a more active constitutive promoter for validation.

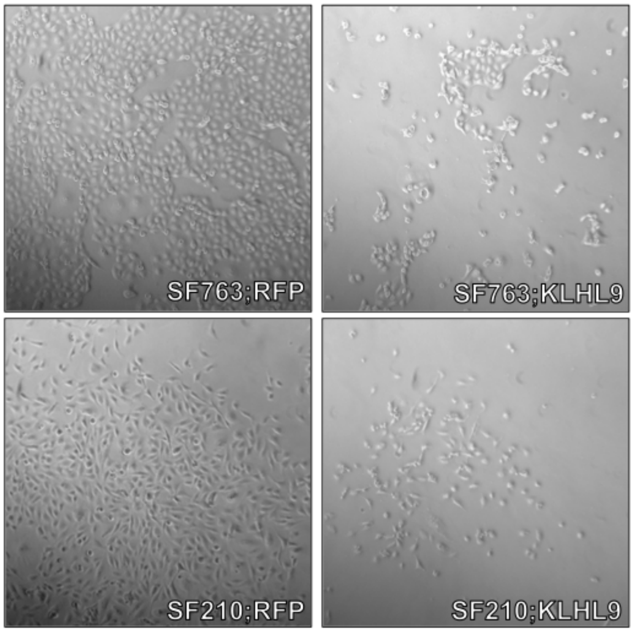


Two validated cell lines obtained from UCSF were verified as bearing homozygous deletions for KLHL9 and CDKN2a (p16): SF210 and SF763

These cells were subsequently amplified and used for repeat validation experiments in the rescue of KLHL9. Both cell lines reciprocated the reported increased protein turnover of the CEBPB and CEBPD proteins, detectable as early as 48-hours post transfection. Transfection of KLHL9 into both SF210 and SF763 also revealed a more robust growth phenotype than those we were able to obtain from the original SF210 cell line. Transient transfection of the lenti-KLHL9 construct for 24 hours resulted in almost complete cell growth arrest and extensive cell death in both SF210 and SF763 48 hours post-transfection.



Exogenous expression of KLHL9 in two new aliquots of human-derived GBM cell lines (the original SF210 and a newly verified line SF763) provided by UCSF results in the loss of CEBPB and CEBPD proteins



Transient transfection of a constitutive KLHL9 for 48 hours in both cell lines, SF210 and SF763, resulted in a marked decrease in cellular prolifleration and increased number of dead or dying cells

Additionally, our collaborator Michael Berens provided access to a series of huGBM primary tumors that were passaged as xenografts in mice. These tumors were grown in a much more *in vivo* biological context than immortalized cell lines maintained in petri dishes without feeder cells, and provided an ideal model to address reviewer requests of *in vivo* experiments. Through qPCR of both cDNA transcripts and genomic DNA, we successfully identified two primary tumor grafts that exhibited evidence for a homozygous deletion of KLHL9, designated GBM64 and HF2354.

Lentiviral infection of our KLHL9 construct and subsequent selection via Blasticidin in these contexts mirrored our observations in cell lines SF210 and SF763. Michael Berens's group reported complete arrest following Blasticidin selection of both tumors when KLHL9 was introduced, but not when the RFP control vector was stably integrated. This work is currently ongoing and no data has been made available at the writing of this thesis.

CHAPTER 6 – Discussion of Results both Computational and Biological, and DIGGIn functionality

The technological advancements in generating human-derived, genome-wide panels of quantitative data have provided an extraordinary window into the complex genetic events that underlie the development and progression of cancers such as GBM. The ability to generate these comprehensive datasets across hundreds of patients is rapidly becoming a feasible, efficient means of acquiring data for analysis of human diseases, which immediately attaches translational relevance to experimental findings. However, the large, systemic genomic arrangements associated with cancers such GBM result in thousands of detectable genomic mutations or rearrangements, any combination of which could contribute to multiple biological and metabolic processes in a highly complex disease. An increasing challenge in the field of cancer research *vis à vis* the increase in high-throughput data acquisition is the development of methods to extract meaningful information from this sea of data. Furthermore, the already extensive knowledge of oncogenes and tumor suppressors has the unfortunate corollary of biasing research towards these well-known, well-characterized processes. While it has been reliably shown that multiple cancers share the same oncogenic pathways, this *a priori* selection bias inherently selects against the discovery of mutations that drive physiological behaviors independent of oncogenesis, which may nonetheless be vital to the understanding of the tumor.

Instead, by creating and implementing a regulatory network approach we were able to identify two master regulators of the mesenchymal differentiation of GBM, despite the fact that neither of these loci has ever been implicated in GBM before. Even further, our approaches were able to detect KLHL9 as a contributor specifically to mesenchymal differentiation despite the fact that it is highly linked to the most common oncogenic mutation in GBM: deletions of p16. We were able to accomplish this by integrating multiple genomic datasets and dynamically interrogating them directly for genes that were predicted to regulate the unique gene panel identifying mesenchymal GBM.

DIGGIn is capable of highly accurate detection of functional genomic alterations. We successfully detected 14/18 bona fide oncogenes and tumor suppressors from a list of ~20,000 genomic loci at a statistically significant enrichment of p<1.93e-10. The number of genomic loci successfully identified as f-CNVs was only ~1500, or about 7.5% of the available loci. This result was surprising in the context of a disease with changes in whole chromosome arms, but coincides with the hypothesis that very few mutations would meaningfully contribute to the behavior of any given biological context. A relatively few number of genes are expressed in any tissue at a given time [ref], and genomic mutations affecting loci whose transcripts are not expressed should not be considered relevant to the disease. This was one of the primary purposes of devising DIGGIn: the elimination of gene loci from consideration based on biological evidence that they would be unrelated to the disease. This circumvents a prime limitation of genome-wide statistical studies: multiple hypothesis testing. Conversely, biological relevance can be directly assigned to any genomic locus that passes the DIGGIn's criteria.

DIGGIn, ARACNe, and MINDy: fully parsing regulatory interactions

When placed into the greater context of the systems biology methodologies developed in the Califano lab, DIGGIn occupies a complementary, but unique, niche in the full analytic framework. ARACNe and MINDy are algorithms designed to reverse-engineer a complete, comprehensive molecular network of transcriptional (ARACNe) and post-translational (MINDy) interactions. From these networks, master regulators and modulators of master regulators can be inferred for a phenotype of interest, and this enriched set of genes can be subsequently interrogated for mutations. This approach adds an additional filter to circumvent the statistical power limitations that have traditionally stymied GWAS and other genome-wide analytics. Whereas these other methods must correct for multiple hypothesis testing on every gene expressed in the genome, or every gene tested for a mutation, ARACNe/MINDy-informed analysis is only concerned with a relatively small subset of genes that are computationally inferred to directly affect the phenotype being studied.

DIGGIn, on the other hand, is designed specifically to identify genomic mutations and assign to them molecular perturbations, which may or may not be placed into the context of an interaction network. Rather than reconstructing a genome-wide interaction network and identifying master regulators to search for mutations, DIGGIn identifies mutations and checks to see if they are candidate master regulators based on their perturbations. As a corollary, transcriptional and activity dependencies on candidate master regulators/drivers are made directly by using genomic reads rather than transcriptional reads of each candidate regulator. These intuitively minor distinctions have two major ramifications for the analysis that render DIGGIn unique from ARACNe/MINDy.

First, DIGGIn is designed to identify regulators in the context of genomic mutations present in the samples, not molecular regulators of a reconstructed, “general” context network. This means that the DIGGIn algorithm is primarily designed to identify genomic mutations that modulate the expression of target biomarker panels – the primary concern is to identify mutations that induce a phenotype measured by a biomaker panel, not to define a comprehensive master regulator list. The mutations will be mutations of master regulators or modulators, but again, defining a comprehensive set master regulators is not the goal of DIGGIn. DIGGIn can subsequently be informed by ARACNe and MINDy network analysis to verify the results, as has been implemented in this thesis. As a corollary, although DIGGIn may be unable to comprehensively identify master regulators, any mutations in master regulators will be immediately identifiable as significant regulators, even if they were undetectable by ARACNe/MINDy and have no *a priori* information that would imply an important role in the phenotype being studied.

Secondly, as a direct corollary of the first difference, the use of genomic data in one dimension of the mutual information analysis drastically alters the underlying map of the probability surface used in estimating mutual information. Gold standard and validated genomic loci that bear CNVs show excellent correlation with gene expression. This shows that genomic CNVs are excellent predictors of differential expression, and in these cases will produce similar estimates of mutual information regardless of whether genomic or gene expression data is used as long as the kernels are properly selected. **The difference in performance of DIGGIn in these contexts is not an increase or decrease of the estimates of MI for a given gene-gene pair, but rather in the *ranking* of the MI estimate.** As defined in the DIGGIn chapters, the null distribution of mutual information is defined by randomized pairing of genomic and gene expression vectors. This is done to generate a simulated set in which genomic mutations and genetic expression are entirely independent. When using exclusively gene expression data, there is a significant range of background noise generated by artificial correlations or indirect effects of expression. It is very likely that a significant set of genes that have no common regulator will be correlated with each other. This generates a null distribution with a relatively significant level of background noise, and can lead to complications with the application of DPI, or in the detection of modulating interactions that are real, but perhaps not distinguishable statistically from background noise.

Conversely, the variance in genomic reads when samples are not mutated is extremely low. Any potentially real interaction between a genomic locus and gene expression will be immediately and significantly ranked above any genomic loci that do not bear mutations at all, simply because the coefficient of variance is significantly greater when mutations in the patient cohort exist. This immediately removes any possibility of genomic variance clouding real signals. The primary concern of DIGGIn then becomes parsing true signals out of LD blocks, which is detailed extensively in DIGGIn, part II.

Combined, these two differences allow DIGGIn to detect biologically relevant regulators and modulators from genomic data. The gene KLHL9 was not identified as a master regulator for GBM mesenchymal induction by the ARACNe algorithm, nor did was MINDy immediately able to identify it as a post-translational modulator. Initial analysis with ARACNe yielded no information on KLHL9 because KLHL9 is not a transcription factor, and subsequent analysis on a subset of signal transduction molecules including KLHL9 failed to identify it as a significant master regulator. MINDy, on the other hand, was able to identify KLHL9 as a modulator locus only after we explicitly searched for KLHL9 as a candidate modulator. Even then, the locus would not have come up as a significant modulator of the CEBP master regulators in a blind, genomic analysis based, again, on enrichment ranks or p-value. The locus, while statistically significant, would not have appeared as a significant candidate modulator among the hundreds of modulators that are identified by MINDy. Yet, KLHL9 is clearly an important modulator of the activities of CEBPB and CEBPD, as evidenced by my biological experiments with the gene. The deletion is significantly enriched in differential expression of the MES marker panel as a whole, and the activities of the master regulators CEBPB and D are significantly increased when KLHL9 is deleted (as inferred by the enrichment of each of the master regulator's predicted regulons in the KLHL9 genomic-genetic hub).

DIGGIn is able to detect these loci explicitly because its regulatory inferences are drawn from genomic-genetic data compared to genetic-genetic data.

Conversely, DIGGIn's primary weakness is the strength of ARACNe and MINDy. DIGGIn cannot detect master regulators or modulators if their respective genomic loci do not bear mutations. It is strictly designed to identify the driver mutations that both affect master regulators and modulators and exist in the samples being studied. What this means is that DIGGIn cannot detect or predict interactions between genes contributing to the development of a phenotype that do not bear mutations. ARACNe and MINDy interactomes provide interaction maps of hundreds of genes. These interactions can be detected across a swath of samples regardless of whether or not the genes involved bear mutations. These interactions can subsequently be used to explore the full extent and breadth of potential interactions for the development of treatments, predicting key mutations, and general abstractions of how a given disease is regulated. DIGGIn is designed to **identify the functional mutations that exist in patients**, ARACNe/MINDy is designed to identify all genes that could affect the patients **if they were to be mutated.**

The end result is that DIGGIn provides a valuable complementary perspective to ARACNe/MINDy. The use of genomic information as a proxy for one dimension of transcriptional information can be used to identify both mutations that directly exist in patients that affect a molecular phenotype, and master regulators or modulators that are not detectable by ARACNe and MINDy due to technical limitations specific to using transcription-only inferences.

Modularity of DIGGIn and Application to Other Models

As an analytic algorithm, the genetic-genomic analysis was implemented specifically with modularity in mind. GBM was selected as a prototype model for the development of these approaches primarily because of the availability of patient-matched datasets made available by the TCGA. However, these methods are applicable to the study of any genetic disease in which stable molecular gene expression profiles and accurate regulatory networks can be generated. This approach is directly applicable to any biological context in which the following criteria are met: The traits being studied are primarily caused by genetic contributions, genomic and gene expression arrays are obtainable, and the biological context in question exists in a relatively stable, homogeneous molecular state.

The algorithm is also set up in a framework that allows for the inclusion of additional metrics to define copy-neutral genomic alterations and assign to them functional molecular changes. Although they have not been implemented, analytic modules can be added to the framework to account for genomic and epigenetic events such as methylation and point mutations – any genomic event that can be detected by experimental methods on a genome-wide scale could theoretically be integrated into this analysis. For GBM, these extra metrics were deemed unnecessary to the scope of this thesis work because GBM is characterized primarily by copy number alterations, and because deep sequencing and methylation data was not available in format or in quantifies to allow for useful analysis.

The identification of f-CNVs in these contexts are possible with as little as 80-100 samples, although this presents the bare minimum required to achieve the needed statistical power. If these results are to be integrated with regulatory networks generated by methods such as ARACNe and MINDy, additional samples will be required to ensure accurate reconstruction of these regulatory networks, as outlined in [Margolin *et al*.]. However, the genetic-genomic analyses can be supplemented with any post-processing methods to add additional biological context to the results.

Biological Relevance of Computational Findings

Using this integrative genetic-genomic approach, we were able to positively identify two candidate master regulators of the mesenchymal subtype without any *a priori* information of the molecular classification of the tumors. Of the two, CEBPD had already been validated as a mesenchymal master regulator, and we subsequently identified KLHL9 as a post-translational regulator of the CEBP master regulators of MES transformation. We were able to identify a post-translational regulator of MES transformation in tight linkage disequilibrium with a common oncogene using only genomic and transcriptional data. In addition, we were able to positively identify almost the entirety of field-accepted *bona fide* tumor suppressors and oncogenes as f-CNVs; the only loci that were missed were ones that either were so rare that statistical power could not be reached (and indeed, these loci would not have been found by traditional methods in this dataset had they not already been established as oncogenes), or they were actively disregarded as functional because, although they may have been shown to induce oncogenesis in general, there was no evidence that these genes are expressed or functional in the context of GBM.

These results demonstrate that the genetic-genomic approach is capable of the statistically enriched identification of biologically relevant genomic loci from a pool of tens of thousands of candidate mutated genomic loci, and thousands of genes expressed in GBM. Additionally, DIGGIn is capable of detecting the presence of multiple independent driver mutations that contribute to the etiology of a disease. This stands in contrast to traditional genomic approaches, which are extensively limited in their detection power due to extensive hypothesis correcting, and due to the lack of dynamic partitioning. The accuracy of raw ANOVA or other statistic methods in a biological context deteriorates as the number of underlying genetic causes in a cohort increases, due to the presence of other causal genes polluting the association signal of any individual locus being tested. Whereas traditional statistical approaches are actively stymied by heterogeneity, DIGGIn was tailor-made with such biological contexts in mind and is actively designed to both circumvent and capitalize on the heterogeneity of the patient cohort.

Implications of Biological Results (KLHL9) in GBM

In the context of GBM specifically, this work provides significant evidence for the value of studying GBM, and cancers in general, not only in the context of oncogenesis and tumor progression, but also in a context that elucidates metabolic and physiological nuances that render tumors unique from patient to patient. This work and an increasing amount of gene expression studies in cancer show that the umbrella categories of cancer that have been traditionally assigned to tumors do not capture the diversity of the disease, even in relatively specific contexts such as glioblastoma, or astrocytoma[14][16][21]. Tumors under most of these classifications are still separable into distinct molecular subtypes based on their gene expression profiles, and it is possible to use these methodologies to elucidate molecular programs that associate with behaviors unique to specific subtypes. These unique molecular programs are functionally distinct, but occur in tandem with the molecular changes associated with tumor development and progression, the classical field of cancer study. Subsequent application of regulatory networks allowed us to further identify that the entire signature could be regulated via a relatively small number of master regulators and showed that manipulation of these other molecular programs independent of classical oncogenic pathways could be sufficient to inhibit tumor growth.

This work adds a complementary approach to the identification of master regulators by identifying the actual mutations that occur in tumor samples to alter the behavior of the master regulators, rather than inferring them from gene expression data. Based on the observation that large molecular panels can be manipulated by a small number of regulators, we were able to devise a computational framework to identify genomic events that showed evidence for perturbing the master regulators, instead of searching on a candidate-by-candidate basis for involvement in classical cancer pathways or by a genome-wide statistical study. This work allowed us to specifically target and identify the regulators of a specific molecular behavior in GBM with clinically relevant effects. The combination of network analysis and genetic-genomics allows for an approach that capitalizes on the information density of these large scale datasets without being hampered by statistical threshold limitations, and allows a focused approach to studying genome-wide molecular behavior using a small number of regulatory hubs.

The identification of KLHL9 as a completely novel post-translational regulator of mesenchymal differentiation in GBM came directly as a computational prediction from exclusively human-derived data, which we subsequently validated with biological experimentation. Our integrative methods were able to identify this locus as a candidate master regulator despite it being deleted in “only” ~30% of mesenchymal tumors in our TCGA cohort, and despite it being in close proximity to the oncogene, p16. DIGGIn in its current implementation is not capable of detecting focal (promoter deletions), copy neutral (point mutations / frameshifts), or epigenetic (methylation) changes. We hypothesize that it is very likely that the remaining mesenchymal samples bear these undetected mutations in KLHL9, the master regulators themselves, or other upstream components that regulate the master regulators. Additionally, most genomic loci proximal to classical oncogenes are actively disregarded as contributory to tumor etiology in any way because they are assumed to be an artifact of association to those oncogenes. We were, instead, able to provide evidence that the high frequency of Mesenchymal tumors, which consist of over 50% of tumors obtained from the TCGA, is due to the high likelihood of obtaining losses of chromosome 9 that would span both p16 and KLHL9. This mutation would be sufficient to induce both tumorigenesis and mesenchymal transformation, and the development of mesenchymal GBM comes as a result of the simultaneous activation of at least two distinct molecular programs: oncogenesis and mesenchymal transformation.

As the rapid advances in biotechnology continue to produce vast amounts of genetic and genomic data at decreasing cost, a primary concern and field of research is the development of computational methods to meaningfully process this data in a biological context. Systems biological approaches have provided a novel perspective on the modeling of complex genetic traits and diseases, capitalizing on the availability of genomic data. The ability to infer regulatory networks has allowed us to integrate years of genetic research into a framework to understanding how large, modular molecular programs are regulated in cell-specific contexts.

Concurrently, it is becoming increasingly apparent that diseases as complex as cancer should not and cannot be addressed simply as a function of oncogenic behavior. Individual tumors can acquire multiple mutations in addition to oncogenic drivers that nonetheless can drastically alter the physiology of the tumor with very real clinical ramifications. These other mutations and their effects cannot be simply dismissed in the interest of studying oncogenesis or angiogenesis. The difference between mesenchymal and proneural GBM is a significantly shorter prognosis - mesenchymal patients in the TCGA cohort do not survive beyond 36 months post-diagnosis, while even proneural patients who succumb to the disease can still survive beyond 60-80 months (surviving patients are predominantly patients with proneural tumors and were not included in the TCGA cohort). The characterization of how these mutations affect the behavior of the tumor via their individual molecular programs not only broadens our understanding of the disease, but also opens up new avenues of research for treatment. Conventional therapeutics that prove ineffective for certain cancers could potentially be replaced by more targeted agents that select against physiological behaviors unique to the cancer subtype, and increasing diagnostic panels to include cancer subtyping can grant valuable insight into improving the diagnosing of the disease.

The goal of this work was to bridge the gap between these two rising issues and elucidate the genetic architecture of mesenchymal tumors in Glioblastoma multiforme. We have successfully created the computational methodology, DIGGIn, to predict driver mutations for individual molecular programs directly from human data. We were able to parse out and identify KLHL9, a novel, highly prominent post-translational regulator of mesenchymal differentiation in GBM. We were able to identify deletions of this gene as a functional genomic perturbation without any *a priori* information as to its relevance to GBM, and were subsequently able to predict its functionality in subtype differentiation out of hundreds of thousands of candidate loci, and to biologically identify its mechanism of action experimentally. Furthermore, the analytic rational and software architecture are readily applicable to any biological context that is appropriately addressed with systems biological methods.

References

1 Ash, RB. *Information Theory*. Dover Books (1990)

2 Bao, S *et al*. Stem cell-like glioma cells promote tumor angiogenesis through vascular endothelial growth gactor. *Cancer Res* **66** (2006)

3 Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**, 382-390, doi:ng1532 [pii]
10.1038/ng1532 (2005).

4 Beier, D *et al*. CD133+ and CD133- gliobloastoma-derived cancer stem cells show differential growth characteristics. *Cancer Res* **67** (2007)

5 Beroukhim, R. *et al*. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci.* **104** (2006)

6 Boorsma, A *et al*. Inferring condition-specific moldulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. *PLoS ONE* **3** (2008)

7 Brem, RB. *et al*. Genetic interactions between polymporphisms that affect gene expression in yeast. *Nature* **436** (2005)

8 Bromberg, JF. *et al* Stat3 as an oncogene. *Cell* **98** (1999)

9 Butte, AJ. and Kohane, IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput.* **5** (2002).

10 Califano, A., Butte, A., Friend, S. H., Ideker, T. & Schadt, E. E. Integrative Network-based Association Studies: Leveraging cell regulatory models in the post-GWAS era. *Nature Genetics, in press* **currently in Nature Preceedings (http://precedings.nature.com/documents/5732/version/1)** (2011).

11 Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318-325, doi:nature08712 [pii]
10.1038/nature08712 (2010).

Chickering, DM. *Learning from Data: Artificial Intelligence and Statistics*. Springer-Verlag, New York. (2000)

12 Compagno, M. *et al.* Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature* **459**, 717-721, doi:nature07968 [pii]
10.1038/nature07968 (2009).

13 Cooper, GF. Herskovits, E. *A Bayesian method for the induction of probabilistic networks from data*. Mach.Learn. 9:309-347 (1992)

14 David, J. and Mackay, C. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press (2003)

15 Demuth, T. and Berens ME. Molecular mechanisms of glioma cell migration and invasion. *J.Nerooncol.* **70** (2004)

16 Doss, S. *et al.* Cis-acting expression quantitative trait loci mapping in mice. *Genome Res.* **15**. (2005)

17 Ein-Dor, L. *et al.* Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21** (2005)

18 Emilson, V. *et al*. Genetics of gene expression and its effect on disease. *Nature* **452** (2009)

19 Erkstrand, AJ. *et al*. Amplified and rearranged epidermal growth factor recpetor genes in human glioblastomas reveal deletions of sequences encoding portions of the N- and/or C-terminal tails. *Proc Natl Acad Sci* **89** (1992)

20 Friedman, N. *et al*. Using Bayesian networks to analyze expression data. *J.Comp.Bio.* **7.** (2000)

21 Freiji, WA. *et al*. Gene expression profiling of gliomas strongly predicts survival. *Cancer Res* **64.** (2004)

22 Furnari, FB *et al.* Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev.* **21** (2007)

23 Godard, S *et al*. Classification of human astrocytic gliomas on the basis of gene expression: A correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. *Cancer Res*. **63** (2003)

24 Hartman, C. *et al*. The rate of homozygous CDKN2A/p16 deletions in glioma cell lines and primary tumors. *Int Jrnl Oncology* **15** (1999).

25 Hecker, M. Lambeck, S. Toepfer, S. *et al*. Gene regulatory network inference: Data integration in dynamic models - A review. *Biosystems* (2009)

26 Heinberger, A. Hlatky, R. Suki, D et al. Prognostic Effect of Epidermal Growth Factor Receptor and EGFRvIII in Glioblasotma Multiforme Patients*. Clinical Cancer Research* **11** (2005)

27 Huttenhower, C *et al.* Nearest neighbor networks: clustering expression data based on gene neighborhoods. *BMC Bioinformatics* **8** (2007)

28 Jaynes, ET. *Information Theory and Statistical Mechanics.* Phys.Rev **106** (1957)

29 Kendziorski, CM. *et al.* Statistical methods for expression quantitative trailt loci (eQTL) mapping. *Biometrics* **62**. (2005)

30 Kitange, GJ. Templeton, KL. and Jenkins, RB. Recent advances in the molecular genetics of primary gliomas. *Curr Opn Oncology* **15** (2003)

31 Krex, D. *et al*. Long term survival with Glioblastoma multiforme. *Brain* **130** (2007)

32 Liang, Y. *et al*. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc. Natl. Acad. Sci*. **102** **(**2005**)**

33 Lee, E. and Bussemaker, H. Identifying genetic determinants of transcription factor activity. *Mol.Sys.Bio.* **6**:**412**. (2010)

34 Li, M and Vitanyi, P. *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York (1996)

35 Louis, DN *et al*. The 2007 WHO classification of tumors of the central nervous system. IARC Press, Lyon, France (2007)

36 Lynch, M and Walsh, B. *Genetics and Analysis of Quantitative Traits*. Sinauer. (1998)

37 Maerki, S. *et al.* The Cul3-KLHL21 E3 ubiquitin ligase targets aurora B to midzone microtubules in anaphase and is required for cytokinesis. *J Cell Biol* **187**, 791-800, doi:10.1083/jcb.200906117 (2009).

38 Mani, K. Lefebvre, C. Wang, K. *et al.* A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Molecular Systems Biology* **169** (2008)

39 Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7 Suppl 1**, S7 (2006).

40 Margolin, A. A. *et al.* Reverse engineering cellular networks. *Nature Protocols **106*** (2006).

41 May, R. Member, S. Dandy, G. *et al*. Critical Values of a Kernel Density-based Mutual Information Estimator. *Intl Joint Conference on Neural Networks* **1** (2006)

42 Mischel, PS. Shai, R. *et al.* Identification of Molecular Subtypes of Glioblastoma by Gene Expression Profiling. *Oncogenomics* **22** (2003)

43 Mitchell, T. *Machine Learning*. McGraw Hill (1997)

44 Morely, M *et al*. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430** (2004)

45 Nakayama, K. I. & Nakayama, K. Ubiquitin ligases: cell-cycle control and cancer. *Nat Rev Cancer* **6**, 369-381, doi:10.1038/nrc1881 (2006).

46 Ohgaki, H. & Kleihues, P. Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. *J Neuropathol Exp Neurol* **64**, 479-489 (2005).

47 Ohgaki, H. Kleihues, P. Genetic alterations and signaling pathways in the evolution of gliomas. *Cancer Science* **100** (2009)

48 Deyell, R. and Attiyeh, E. Advances in the understanding of constitutional and somatic genomic alterations in neuroblastoma. *Cancer Genetics* **204:3** (2011)

49 Nutt, CL. et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res* **63**. (2003)

50 Nadeau, S. *et al*. A transcriptional role for CEBPB in the neuronal response to axonal injury. *Mol Cell Neurosci* **29** (2005)

51 Pelloski, CE *et al*. YKL-40 expression is associated with poorer prognosis response to radiation and shorter overall survival in glioma. *Clin Cancer Res* **11** (2005).

52 Phillips, H. S. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157-173, doi:S1535-6108(06)00056-0 [pii]
10.1016/j.ccr.2006.02.019 (2006).

53 Prados, MD. and Levin, V. Biology and treatment of malignant gliomas. *Seminars Oncol* **27** (2000)

54 Quniata, FJ. *et al* Systems biology approaches for the study of multiple sclerosis. *J Cell Mol Med* **12:4** (2008)

55 Rockman, M. Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature* **456** (2008)

56 Ronald, J and Akey, JM. The evolution of gene expression QTL in *Saccharomyces cerevisiae*. *PLoS ONE* **2** (2007)

57 Sakariassen, PO. *et al*. Angiogenesis-independent tumor growth mediated by stem-like cancer cells. *Proc. Natl. Acad. Sci*. **103** (2006)

58 Schmidt, MC. *et al*. Impact of genotype and morphology on the prognosis of glioblastoma. *J Neuropath Exp Neuro* **61** (2002)

59 Shai, R. *et al*. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene* **22** (2003)

60 Sieberts, SK. *et al*. Moving towards a system genetics view of disease. *Mamm.Genome* **18** (2007)

61 Simmons, ML *et al*. Analysis of complex relationshipts between age, p53, epidermal growth factor receptor, and survival in glioblastoma multiforme. *Cancer Res* **61** (2001)

62 Showalter, TN. *et al*. Multifocal Glioblastoma Multiforme: Prognostic Factors and Patterns of Progression*. Intl Jrnl Radiation OncogologyBiologyPhysics* **69** (2007)

63 Smith, JS. *et al*. PTEN mutation, EGFR amplification, and outcome in patients with anaplastic astrocytoma and glioblastoma multiforme. *JNCI* **16** (2001)

64 Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102** (2005)

65 Sumara, I. *et al.* A Cul3-based E3 ligase removes Aurora B from mitotic chromosomes, regulating mitotic progression and completion of cytokinesis in human cells. *Dev Cell* **12**, 887-900, doi:10.1016/j.devcel.2007.03.019 (2007).

66 Sun, L. *et al.* Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* **9**, 287-300, doi:S1535-6108(06)00084-5 [pii]
10.1016/j.ccr.2006.03.003 (2006).

67 TCGA-Consortium. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:nature07385 [pii]
10.1038/nature07385 (2008).

68 TCGA-Consortium. Integrated genomic analysis identifies clinically relevant subtypes of glioblasotma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **1:17** (2010)

69 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713, doi:10.1038/nature09270 (2010).

70 von Deimling, A. *et al*. Molecular pathways in the formation of gliomas. *Glia* **15** (1995).

71 Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110, doi:S1535-6108(09)00432-2 [pii]
10.1016/j.ccr.2009.12.020.

72 Wang, H. *et al*. Analysis of the activation status of Akt, NFkB, and STAT3 i n human diffuse gliomas. *Lab. Invest.* **84** (2004)

73 Wang, K. *et al.* Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat Biotechnol* **27**, 829-839, doi:nbt.1563 [pii]
10.1038/nbt.1563 (2009).

74 Wang, K. *et al.* Dissecting the interface between signaling and transcriptional regulation in human B cells. *Pac Symp Biocomput*, 264-275 (2009).

75 Watanabe, K. *et al*. Overexpression of the EGF receptor and p53 mutations are mutually exlusive in Primary and Secondary Glioblastomas. *Brain Pathology* **5** (2008).

76 Weinstein, I. B. Cancer. Addiction to oncogenes--the Achilles heal of cancer. *Science* **297**, 63-64, doi:10.1126/science.1073096
297/5578/63 [pii] (2002).

77 Weiss, WA *et al*. Genetic determinants of malignancy in a mouse model f or oligodendogliomas. *Cancer Res.* **63** (2003)

78 Witten, I. and Frank, E. *Data Mining: Practical Learning Tools and Techniques*. Morgan Kaufman, Amsterdam. (2005)

79 Yang, X. *et al.* Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat Genet*, doi:ng.325 [pii]
10.1038/ng.325 (2009).

80 Yao, Y. Y. Information-theoretic measures for knowledge discovery and data mining, in *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, Karmeshu (ed.), Springer, pp. 115–136. (2003)

81 Zhao, X. *et al.* The N-Myc-DLL3 cascade is suppressed by the ubiquitin ligase Huwe1 to inhibit proliferation and promote neurogenesis in the developing brain. *Dev Cell* **17**, 210-221, doi:S1534-5807(09)00293-7 [pii]

82 Zhu, J *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat.Genetic.* **40** (2008)

83 Zhu, Y. *et al*. Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma. *Cancer Cell* **8**. (2005)

APPENDICES

APPEND01 – Classification of TCGA samples

TCGA-02-0001-01	MES
TCGA-02-0002-01	MES
TCGA-02-0003-01	PN
TCGA-02-0004-01	MES
TCGA-02-0006-01	MES
TCGA-02-0007-01	PRO
TCGA-02-0009-01	MES
TCGA-02-0010-01	PN
TCGA-02-0011-01	PN
TCGA-02-0014-01	PN
TCGA-02-0015-01	MES
TCGA-02-0016-01	PRO
TCGA-02-0021-01	PRO
TCGA-02-0023-01	MES
TCGA-02-0024-01	PN
TCGA-02-0025-01	MES
TCGA-02-0026-01	PN
TCGA-02-0027-01	MES
TCGA-02-0028-01	PN
TCGA-02-0033-01	MES
TCGA-02-0034-01	MES
TCGA-02-0037-01	MES
TCGA-02-0038-01	MES
TCGA-02-0039-01	MES
TCGA-02-0043-01	MES
TCGA-02-0046-01	PN
TCGA-02-0047-01	PN
TCGA-02-0048-01	PN
TCGA-02-0051-01	MES
TCGA-02-0052-01	MES
TCGA-02-0054-01	MES
TCGA-02-0055-01	MES
TCGA-02-0057-01	MES
TCGA-02-0058-01	PN
TCGA-02-0059-01	MES
TCGA-02-0060-01	PN
TCGA-02-0064-01	MES
TCGA-02-0068-01	MES
TCGA-02-0070-01	MES
TCGA-02-0071-01	MES
TCGA-02-0074-01	PN
TCGA-02-0075-01	MES
TCGA-02-0079-01	MES
TCGA-02-0080-01	PN
TCGA-02-0083-01	MES
TCGA-02-0084-01	PN
TCGA-02-0085-01	MES
TCGA-02-0086-01	MES
TCGA-02-0087-01	PN
TCGA-02-0089-01	MES
TCGA-02-0099-01	MES
TCGA-02-0102-01	MES
TCGA-02-0104-01	PN
TCGA-02-0106-01	MES
TCGA-02-0107-01	MES
TCGA-02-0111-01	MES
TCGA-02-0113-01	MES
TCGA-02-0114-01	PN
TCGA-02-0115-01	MES
TCGA-02-0116-01	MES
TCGA-02-0117-01	MES
TCGA-02-0258-01	PN
TCGA-02-0260-01	MES
TCGA-02-0266-01	MES
TCGA-02-0269-01	MES
TCGA-02-0271-01	MES
TCGA-02-0281-01	PN
TCGA-02-0285-01	MES
TCGA-02-0289-01	PRO
TCGA-02-0290-01	MES
TCGA-02-0317-01	MES
TCGA-02-0321-01	MES
TCGA-02-0324-01	MES
TCGA-02-0325-01	PRO
TCGA-02-0326-01	MES
TCGA-02-0330-01	PRO
TCGA-02-0332-01	PN
TCGA-02-0333-01	MES
TCGA-02-0337-01	MES
TCGA-02-0338-01	PN
TCGA-02-0339-01	PN
TCGA-02-0422-01	MES
TCGA-02-0430-01	MES
TCGA-02-0432-01	PN

TCGA-02-0439-01	PRO
TCGA-02-0440-01	PN
TCGA-02-0446-01	PRO
TCGA-02-0451-01	MES
TCGA-02-0456-01	MES
TCGA-06-0122-01	MES
TCGA-06-0124-01	MES
TCGA-06-0125-01	MES
TCGA-06-0126-01	PN
TCGA-06-0127-01	MES
TCGA-06-0128-01	PN
TCGA-06-0129-01	PN
TCGA-06-0130-01	MES
TCGA-06-0132-01	MES
TCGA-06-0133-01	PRO
TCGA-06-0137-01	MES
TCGA-06-0137-01	MES
TCGA-06-0137-01	MES
TCGA-06-0137-01	MES
TCGA-06-0138-01	PRO
TCGA-06-0139-01	MES
TCGA-06-0140-01	MES
TCGA-06-0141-01	MES
TCGA-06-0142-01	PN
TCGA-06-0143-01	MES
TCGA-06-0145-01	MES
TCGA-06-0145-01	MES
TCGA-06-0145-01	MES
TCGA-06-0145-01	MES
TCGA-06-0146-01	PN
TCGA-06-0147-01	MES
TCGA-06-0148-01	MES
TCGA-06-0148-01	MES
TCGA-06-0148-01	MES
TCGA-06-0148-01	MES
TCGA-06-0149-01	MES
TCGA-06-0152-01	MES
TCGA-06-0154-01	MES
TCGA-06-0154-01	MES
TCGA-06-0156-01	MES
TCGA-06-0156-01	PN
TCGA-06-0156-01	PN
TCGA-06-0157-01	PN
TCGA-06-0158-01	MES
TCGA-06-0160-01	PN
TCGA-06-0162-01	PRO
TCGA-06-0164-01	MES
TCGA-06-0166-01	MES
TCGA-06-0167-01	PN
TCGA-06-0168-01	MES
TCGA-06-0169-01	MES
TCGA-06-0171-01	PN
TCGA-06-0173-01	MES
TCGA-06-0174-01	PN
TCGA-06-0175-01	MES
TCGA-06-0176-01	MES
TCGA-06-0177-01	MES
TCGA-06-0178-01	PN
TCGA-06-0179-01	MES
TCGA-06-0182-01	PRO
TCGA-06-0184-01	MES
TCGA-06-0185-01	MES
TCGA-06-0187-01	MES
TCGA-06-0188-01	PN
TCGA-06-0189-01	MES
TCGA-06-0190-01	MES
TCGA-06-0192-01	MES
TCGA-06-0194-01	MES
TCGA-06-0195-01	PN
TCGA-06-0197-01	MES
TCGA-06-0201-01	MES
TCGA-06-0206-01	MES
TCGA-06-0208-01	PN
TCGA-06-0210-01	MES
TCGA-06-0211-01	MES
TCGA-06-0211-01	PRO
TCGA-06-0213-01	MES
TCGA-06-0214-01	PRO
TCGA-06-0216-01	PN
TCGA-06-0216-01	PRO
TCGA-06-0221-01	PN
TCGA-06-0237-01	PN
TCGA-06-0238-01	PN
TCGA-06-0240-01	PN
TCGA-06-0241-01	PN
TCGA-06-0394-01	MES
TCGA-06-0397-01	MES
TCGA-06-0402-01	MES
TCGA-06-0409-01	MES
TCGA-06-0410-01	PN
TCGA-06-0412-01	MES
TCGA-06-0413-01	PN

TCGA-06-0414-01	PN
TCGA-06-0644-01	MES
TCGA-06-0645-01	MES
TCGA-06-0646-01	PRO
TCGA-06-0648-01	PN
TCGA-06-0649-01	MES
TCGA-06-0673-11	PN
TCGA-06-0676-11	PN
TCGA-06-0678-11	MES
TCGA-06-0680-11	PN
TCGA-06-0681-11	PN
TCGA-06-0686-01	PN
TCGA-06-0743-01	PRO
TCGA-06-0744-01	PRO
TCGA-06-0745-01	PN
TCGA-06-0747-01	MES
TCGA-06-0749-01	MES
TCGA-06-0750-01	MES
TCGA-07-0249-20	MES
TCGA-08-0244-01	MES
TCGA-08-0246-01	MES
TCGA-08-0344-01	PN
TCGA-08-0345-01	MES
TCGA-08-0346-01	MES
TCGA-08-0347-01	PRO
TCGA-08-0348-01	PRO
TCGA-08-0349-01	PRO
TCGA-08-0350-01	PN
TCGA-08-0351-01	PN
TCGA-08-0352-01	MES
TCGA-08-0353-01	PN
TCGA-08-0354-01	MES
TCGA-08-0355-01	PRO
TCGA-08-0356-01	MES
TCGA-08-0357-01	MES
TCGA-08-0358-01	PN
TCGA-08-0359-01	MES
TCGA-08-0360-01	MES
TCGA-08-0373-01	MES
TCGA-08-0375-01	PRO
TCGA-08-0380-01	PRO
TCGA-08-0385-01	PN
TCGA-08-0386-01	MES
TCGA-08-0389-01	PN
TCGA-08-0390-01	MES
TCGA-08-0392-01	MES
TCGA-08-0509-01	MES
TCGA-08-0510-01	MES
TCGA-08-0511-01	MES
TCGA-08-0512-01	MES
TCGA-08-0514-01	MES
TCGA-08-0516-01	MES
TCGA-08-0517-01	PN
TCGA-08-0518-01	MES
TCGA-08-0520-01	MES
TCGA-08-0521-01	MES
TCGA-08-0522-01	MES
TCGA-08-0524-01	PN
TCGA-08-0525-01	MES
TCGA-08-0529-01	MES
TCGA-08-0531-01	MES
TCGA-08-0623-11	PN
TCGA-08-0626-11	PN
TCGA-08-0627-11	PN
TCGA-12-0616-01	PN
TCGA-12-0618-01	PN
TCGA-12-0619-01	MES
TCGA-12-0620-01	MES
TCGA-12-0653-01	MES
TCGA-12-0654-01	MES
TCGA-12-0656-01	MES
TCGA-12-0657-01	MES
TCGA-12-0688-01	MES
TCGA-12-0692-01	MES
TCGA-12-0703-01	MES
TCGA-12-0707-01	MES
TCGA-12-0772-01	MES
TCGA-12-0773-01	MES
TCGA-12-0775-01	MES
TCGA-12-0776-01	MES
TCGA-12-0778-01	MES
TCGA-12-0780-01	MES
TCGA-15-0742-01	MES

APPEND02 – List of GBM f-CNVs

AAAS
AARS
ABCC4
ABCD3
ABCD4
ABCF2
ABHD11
ABHD14A
ABHD4
ABHD5
ABI1
ABLIM3
ACN9
ACOT7
ACSL1
ACSL5
ACTL6A
ACTN4
ACTR1A
ACTR5
ACVR2A
ADCK2
ADCY6
ADIPOR1
ADNP
ADORA2A
ADRA2A
ADRM1
ADSL
ADSS
AES
AGPAT4
AHCYL1
AHNAK
AHSA1
AKAP6
AKAP8
AKR1C2
AKR7A2
AKT1
ALDH18A1
ALDH1A3
ALDH1B1
ALG12
ALMS1
ALPK3
AMD1
ANAPC10
ANAPC13
ANKMY2
ANKRD11
ANKRD17
ANP32B
ANXA11
ANXA7
AOX1
AP1B1
AP1M2
AP2A2
AP2S1
AP3D1
APITD1
APOBEC3C
APOE
APOL3
APP
APTX
ARCN1
ARF3
ARF5
ARFGAP3
ARFIP2
ARHGAP22
ARHGEF10L
ARHGEF12
ARID4A
ARIH1
ARL1
ARL3
ARMC1
ARPC1A
ASB6
ASF1A
ATF5
ATG3
ATG5
ATP10D
ATP12A
ATP2B4
ATP5C1
ATP5F1
ATP5S
ATP6V0A1
ATP6V0D1
ATP6V1D
ATP6V1E1
ATP6V1F
ATRN
ATRNL1
ATXN10
ATXN3
AUTS2
AVEN
AVIL
AVPI1
B4GALNT1
B4GALT1
B4GALT5
BAG3
BAG5
BAHD1
BAZ1A
BAZ1B
BAZ2A
BCAR3
BCAS2
BCAT2
BCKDHA
BCL2L2
BCL6
BCL7B
BCMO1
BCR
BDH1
BECN1
BIRC2
BLNK
BNC2
BPGM
BRP44
BRP44L
BSN

BST2
BTBD1
BTN2A2
BXDC5
BYSL
C10orf119
C10orf72
C10orf76
C11orf48
C11orf60
C11orf63
C11orf68
C12orf4
C12orf41
C13orf1
C13orf18
C13orf23
C14orf1
C14orf101
C14orf135
C14orf147
C14orf156
C14orf166
C15orf44
C16orf61
C17orf48
C17orf75
C19orf22
C1D
C1GALT1
C1orf174
C1orf25
C1QTNF3
C1R
C20orf11
C20orf24
C20orf29
C20orf4
C21orf2
C22orf9
C3orf18
C5
C5AR1
C5orf15
C5orf22
C6orf64
C6orf66
C7orf10
C7orf26
C8orf33
C9orf46
C9orf82
CACNA1D
CACNA2D3
CACNB2
CACNG3
CALU
CAMK1D
CAMK2G
CAMK2N1
CAMTA1
CAND1
CAPZA2
CARHSP1
CASD1
CBARA1
CBR4
CC2D1A
CCDC101
CCDC106
CCDC25
CCDC56
CCDC6
CCDC69
CCDC94
CCND3
CCNG2
CCT2
CCT6A
CCT7
CD164
CD2
CD24
CD2BP2
CD63
CD93
CDC14B
CDC16
CDC2L5
CDC2L6
CDC37
CDC40
CDC42
CDH10
CDH6
CDK3
CDK4
CDK5
CDK5RAP2
CDK9
CDKN2A
CDKN2C
CDS2
CEBPD
CEP135
CERK
CFI
CGRRF1
CH25H
CHAC1
CHCHD3
CHD8
CHIC2
CHMP4A
CHMP5
CHST12
CHST3
CIB1
CIRBP
CIZ1
CLCF1
CLDN10
CLEC11A
CLMN
CLN5
CLOCK
CLPTM1
CLU
CNOT3
CNOT4
CNOT7
CNR1
COG2
COG5
COL13A1
COL2A1
COL5A1
COMMD4

COMMD9
COMT
COPA
COPE
COPS2
COPS6
COQ7
COQ9
COX15
COX4NB
COX6B1
CPEB3
CPM
CPNE1
CRB1
CREB3
CREB3L2
CRELD2
CRISPLD2
CRKL
CROT
CRTAM
CRYBB2
CRYL1
CSDA
CSDE1
CSF2RB
CSNK1A1
CSNK1G2
CSPG5
CSTF1
CTDSP2
CTNNB1
CTNNBIP1
CTNNBL1
CUGBP2
CUL1
CUL2
CUL4A
CUL5
CUTC
CWF19L1
CXCL12
CYCS
CYorf15B
CYP27B1
CYP51A1
DAG1
DBI
DCLRE1C
DCTN2
DCTN3
DCTN5
DCTN6
DCUN1D4
DDIT3
DDIT4
DDO
DDX17
DDX21
DDX27
DDX31
DDX39
DDX3X
DDX41
DDX50
DDX56
DDX58
DEAF1
DENND2A
DENND2D
DEPDC6
DERA
DFFA
DGCR2
DGKA
DGKI
DHDDS
DHRS7
DHX32
DHX38
DIO2
DKK1
DLC1
DLG1
DLGAP1
DLGAP4
DMPK
DNAJA1
DNAJB5
DNAJB6
DNAJC12
DNAJC17
DNAJC3
DNAL4
DNTTIP2
DOCK1
DPAGT1
DPEP2
DPY19L4
DRAP1
DRG1
DTX3
DUS4L
DUSP12
DUSP6
DYNLT1
E2F3
EBNA1BP2
ECD
ECHDC1
ECHDC3
ECHS1
EDNRB
EED
EFCAB2
EFNB2
EGFR
EHD4
EIF1AX
EIF1AY
EIF2AK1
EIF2AK3
EIF4EBP1
EIF4EBP2
EIF4G2
EIF5
ELAVL2
ELMO2
ELN
EMP3
ENG
ENTPD5
ENTPD6
EPHB3
EPOR
EPS15
EPS15L1

EPS8L2
ERCC1
ERCC2
ERCC5
ESRRA
ETNK1
ETNK2
EWSR1
EXOC1
EXOC3
EXOC7
EXOSC7
EXOSC8
EXPH5
EYA2
EZH2
F13A1
F3
FAF1
FAM105A
FAM35A
FAM45A
FAM46A
FAM53B
FAM5C
FAM65A
FAM69A
FAM82B
FANCC
FANCG
FARP1
FBN1
FBXO28
FBXO34
FBXO38
FBXO7
FCER2
FDFT1
FDX1
FECH
FEM1B
FER1L3
FHOD1
FIBP
FIP1L1
FIS1
FKBP14
FKBP1A
FKBP3
FKBP5
FLJ10357
FLJ20323
FMOD
FNDC3A
FNTA
FNTB
FOXJ2
FPGT
FRAT1
FRMD4A
FUCA1
FXR1
FXYD1
FYCO1
G3BP2
GALC
GALK2
GALNAC4S-6ST
GARNL1
GARS
GAS7
GAS8
GATAD1
GATAD2A
GBAS
GCA
GCC1
GCH1
GCLM
GDI2
GFAP
GFOD2
GFPT1
GFRA1
GFRA2
GGA1
GHITM
GINS1
GLI1
GLTSCR1
GLUD1
GMEB1
GMFB
GMPR2
GNA11
GNA12
GNAI1
GNAI2
GNAI3
GNB5
GNG7
GNL3
GNPTAB
GOLGA1
GOLGA2
GPHN
GPR6
GPR65
GPSM2
GPX4
GRHPR
GSN
GSTO1
GSTT1
GSTZ1
GTF2B
GTF2F2
GTF2H5
GTF3C1
GTF3C2
GTF3C4
GTF3C5
GTPBP1
GTPBP2
GTPBP4
GUF1
GUSB
GYS1
GZMB
H1F0
H2AFV
HABP4
HADHB
HBEGF
HBS1L
HBXIP
HDAC2
HDAC9

HDHD1A
HERC1
HIC2
HIF1A
HIP1
HISPPD2A
HIST1H1C
HIVEP2
HK1
HLA-DQA1
HLCS
HMG20B
HMGCL
HMGN3
HOXA10
HOXC10
HOXC13
HOXC4
HOXC8
HP1BP3
HPR
HPRT1
HPS6
HRAS
HSBP1
HSD17B12
HSF2
HSP90AB1
HSPB1
HSPH1
HUS1
ICMT
IDI1
IFI35
IFI6
IFIH1
IFIT2
IFIT3
IFRD1
IFT74
IL15RA
IL6
IL6ST
IMPDH2
INHBE
INPP5A
INPP5E
INPP5F
INTS5
INTS6
INVS
IQCE
IRGQ
ISLR
ISOC2
ITCH
ITGA8
ITGB1
ITIH2
ITM2B
ITPA
IVD
JARID1A
JMJD2C
JRK
KARS
KBTBD2
KCNA3
KCNH2
KCNMA1
KCNMB2
KCNMB4
KCTD12
KCTD15
KDELR2
KIAA0247
KIAA0284
KIAA0355
KIAA0391
KIAA0415
KIAA0495
KIAA0562
KIAA0564
KIAA0649
KIAA0892
KIAA1128
KIAA1279
KIAA1539
KIAA1598
KIAA1704
KIAA1797
KIF5A
KIF5B
KIN
KIT
KLF11
KLF9
KLHDC2
KLHDC4
KLHL12
KLHL20
KLHL24
KLHL9
KLRK1
KPNA3
KRT18
KTN1
LAMA5
LANCL2
LAPTM4A
LARP5
LCMT1
LDB3
LEMD3
LEPROTL1
LETM1
LGALS3
LGR4
LHFP
LIAS
LILRB4
LIMK2
LIN7C
LIPA
LMO2
LMO4
LOC90379
LPIN2
LRIG2
LRP3
LRP5
LRRC15
LRRC8D
LSM3
LSM7
LSM8
LUC7L2
LY6H

LZTFL1
M6PRBP1
MAD1L1
MAD2L1BP
MAGI2
MAN1A1
MAN1A2
MANSC1
MAP3K7IP2
MAP4
MAP7
MAPK1
MAPK14
MAPK6
MAPKAP1
MAPKBP1
MAPRE1
MARS
MAT2B
MAX
MBD6
MBIP
MBTPS1
MCF2
MDH2
MDM1
MDM2
ME1
ME3
MED4
MED6
MEN1
MEOX2
MET
METTL3
METTL4
MFGE8
MFN1
MFN2
MGAT1
MGAT3
MGC2752
MICAL1
MINPP1
MIZF
MKL1
MKRN1
MLC1
MLL
MLLT10
MLLT3
MN1
MOXD1
MPDZ
MPHOSPH6
MPI
MPP5
MRPL17
MRPL18
MRPL19
MRPL24
MRPL39
MRPL4
MRPL40
MRPS17
MRPS2
MRPS22
MRPS31
MSRB2
MTAP
MTERF
MTHFD1
MTIF2
MTMR3
MTRF1
MUM1
MYH9
MYL6
MYO9B
MYRIP
NADK
NAGA
NANOS1
NARS2
NCBP2
NCL
NCOA1
NCOA6
NDEL1
NDFIP1
NDUFA2
NDUFA3
NDUFA6
NDUFA8
NDUFAB1
NDUFAF1
NDUFB5
NDUFB6
NDUFB7
NDUFB8
NDUFS6
NEDD8
NEK1
NEK3
NELL1
NF2
NFASC
NFE2L2
NFKBIA
NFX1
NIPA2
NIPSNAP1
NLGN4X
NMT2
NMU
NNMT
NOL6
NOL7
NOLC1
NOSIP
NPC2
NPEPPS
NPM1
NPM3
NPR2
NPTN
NPTX2
NR2F6
NRBF2
NRD1
NRP1
NSFL1C
NT5C2
NTRK3
NUAK2
NUDCD3
NUDT1
NUDT15

NUFIP1
NUP107
NUP133
NUP205
NUP214
NUP37
NUP43
NUP50
NUP85
NXT1
OAZ1
OBFC1
OGDH
OGFOD1
OIP5
OLFML1
OPA1
OPRS1
OPTN
ORC3L
ORC5L
OS9
OSBPL10
OSBPL2
OSBPL9
OSTM1
OVGP1
OXA1L
OXSR1
PAICS
PAK2
PAK4
PAK6
PANK2
PANK4
PAOX
PAPD1
PAPOLA
PARD3
PARK7
PARVA
PARVB
PAXIP1
PCBD1
PCCA
PCDH21
PCDH7
PCDH9
PCID2
PCMT1
PCMTD2
PCNA
PDAP1
PDCD11
PDCD4
PDGFRA
PDLIM7
PDSS2
PEF1
PELI2
PER3
PES1
PEX1
PEX3
PEX5
PEX7
PFKP
PFTK1
PGAP1
PGLS
PGM3
PGPEP1
PHACTR2
PHB2
PHF10
PHF11
PHIP
PHKG1
PHLDA2
PHTF2
PHYH
PIGB
PIGK
PIGN
PIGO
PIGV
PIK3C2B
PIN4
PINK1
PLAA
PLAGL1
PLAUR
PLEKHA1
PLEKHA4
PLEKHA6
PLEKHF1
PLK4
PLXNB1
PMM1
PMPCB
PMVK
PNKP
PNMT
PODXL2
POFUT1
POFUT2
POLDIP3
POLH
POLR1E
POLR2B
POLR2F
POSTN
PPAT
PPFIBP2
PPM1A
PPM1F
PPME1
PPP1R13L
PPP1R15A
PPP1R8
PPP2R5C
PPP2R5E
PPP3CA
PPP3CB
PPP3CC
PPP6C
PQLC1
PRC1
PRDX3
PREB
PRG3
PRIM1
PRKAB2
PRKACB
PRKCDBP
PRKCQ
PRKG1
PRKRIP1

PRKY
PRMT1
PRMT5
PROSC
PROX1
PSAP
PSD
PSEN1
PSMA4
PSMB4
PSMC6
PSMD1
PSMD13
PSMD14
PSMD4
PSMD5
PSME1
PTBP2
PTCD1
PTDSS2
PTGES3
PTMS
PTOV1
PTP4A1
PTPN21
PTPRA
PTPRD
PTPRK
PUS7L
PVRL2
PXMP4
PYGB
PYGL
QRICH1
RAB11A
RAB11FIP2
RAB14
RAB27A
RAB5B
RABAC1
RABGGTB
RABIF
RABL4
RAC1
RAD23B
RAE1
RAF1
RAI14
RALA
RALGDS
RALGPS1
RAP2A
RARRES2
RARRES3
RARS
RASL11B
RASSF2
RB1
RBBP5
RBBP9
RBM16
RBM28
RBM5
RBP4
RBX1
RCBTB1
RCBTB2
RCP9
RDH11
RER1
RERE
REXO4
RFC2
RFK
RGS10
RGS16
RGS17
RGS6
RHOA
RHOBTB1
RHOC
RIC8A
RIN1
RINT1
RIPK5
RNF11
RNF111
RNF128
RNF141
RNF31
RNF6
RNF7
RNF8
RNH1
RPL11
RPL22
RPL28
RPP30
RPS21
RPS23
RPS24
RPS25
RPS27L
RPS9
RRAGA
RRAGD
RSU1
RTF1
RUVBL2
RWDD1
RWDD3
RXRA
SACM1L
SACS
SAE1
SAFB2
SAMM50
SAP18
SAPS2
SAPS3
SAR1A
SARS
SASH1
SC5DL
SCAMP2
SCAMP3
SCAMP4
SCARA3
SCFD1
SCRG1
SCUBE2
SCYL3
SDCCAG1
SDCCAG8
SDF2L1
SEC24C
SEC61B
SEC61G

SEC63
SEL1L
SEMA6D
SENP2
SERINC1
SETD2
SETX
SEZ6L
SF3B3
SF3B5
SF4
SFRS14
SFRS2IP
SGCB
SGPL1
SGPP1
SGTA
SH3GLB2
SH3PXD2A
SHARPIN
SHB
SHOC2
SIAH1
SIGLEC7
SIN3B
SIP1
SIPA1L1
SKI
SLC10A2
SLC16A1
SLC1A1
SLC1A4
SLC24A1
SLC25A17
SLC25A20
SLC25A22
SLC25A28
SLC26A10
SLC29A3
SLC2A5
SLC30A9
SLC33A1
SLC35A1
SLC35E3
SLC38A1
SLC4A2
SLC6A3
SLC7A8
SLC7A9
SLC9A1
SLK
SMAP1
SMARCA2
SMARCD3
SMO
SMU1
SMURF1
SNAP23
SNAP29
SNAPC2
SNAPC3
SNAPC4
SNCG
SNRPD3
SNRPF
SNTB1
SNW1
SNX13
SNX3
SNX5
SNX6
SOCS6
SOD1
SORCS3
SOS2
SOX13
SPAG6
SPATA5L1
SPCS2
SPG20
SPG7
SPHK2
SRGAP2
SRP54
SRP72
SRPK1
SRPK2
SRPR
SRPRB
SS18L1
SS18L2
SSBP1
SSNA1
SSR1
SSX2IP
ST13
ST3GAL4
ST6GALNAC4
ST7
STAM
STAT3
STIM1
STK16
STK24
STK32B
STOML2
STS
STX16
SUCLA2
SUPT16H
SUPV3L1
SURF1
SUZ12
SYF2
SYNJ2
SYT13
TACC3
TADA3L
TAF10
TAF2
TANK
TASP1
TAX1BP1
TBC1D2
TBCC
TBL1XR1
TBL2
TBP
TBPL1
TCEA1
TCEB3
TCF20
TCF4
TDRD3
TDRD7
TEAD1
TEK
TERF2

TEX10
TFAM
TFB1M
TFDP1
TFR2
TFRC
THBS1
THNSL1
THOC5
THPO
THY1
TIAM1
TIMM23
TIMM44
TIMM9
TINF2
TJP1
TKT
TM2D1
TM9SF1
TM9SF2
TMED10
TMED2
TMED3
TMEM106B
TMEM115
TMEM135
TMEM147
TMEM30A
TMEM39A
TMEM5
TMEM62
TMEM8
TMEM80
TMEM87A
TMEM9B
TNKS2
TNPO3
TOMM22
TOPORS
TOR1A
TP53
TP53BP1
TPD52L2
TPP2
TPST2
TRA2A
TRIM22
TRIM24
TRIM8
TRIP13
TRPM2
TRPM4
TRRAP
TSEN34
TSFM
TSG101
TSPAN13
TSPAN31
TSPYL4
TTC26
TTF1
TTK
TTLL12
TUBB2B
TUBG2
TUBGCP2
TUBGCP5
TXNL4A
TXNRD2
TYRO3
TYRP1
UBAP1
UBAP2
UBE2A
UBE2D1
UBE2D4
UBE2H
UBE2L3
UBE2L6
UBE2Q1
UBE3C
UBIAD1
UBL3
UBTD1
UCHL3
UFD1L
UFM1
UGCGL2
UNC50
UPF3A
UPP1
UQCR
USF2
USP10
USP13
USP14
USP2
USP4
USP46
USP9X
USPL1
UTP14C
UTX
VAMP3
VAPA
VCP
VGF
VISA
VLDLR
VPS13C
VPS13D
VPS26A
VPS37B
VPS37C
VPS39
VPS41
VPS4A
VRK3
WAC
WAPAL
WASL
WBP4
WBSCR22
WDR18
WDR3
WDR32
WDR37
WDR42A
WDR48
WDR7
WDR77
WDR8
WEE1
WIPI2
WTAP
XPOT
XRCC5

YARS2
YEATS2
YEATS4
YES1
YKT6
YME1L1
YTHDC1
YTHDF1
ZBED4
ZBTB1
ZBTB24
ZC3H7B
ZCCHC14
ZCWPW1
ZDHHC4
ZDHHC7
ZFHX4
ZFP106
ZFP30
ZFX
ZFYVE26
ZHX3
ZMYM5
ZNF131
ZNF132
ZNF134
ZNF14
ZNF180
ZNF212
ZNF214
ZNF223
ZNF226
ZNF227
ZNF23
ZNF238
ZNF248
ZNF250
ZNF264
ZNF277
ZNF281
ZNF282
ZNF3
ZNF324
ZNF337
ZNF394
ZNF419
ZNF43
ZNF44
ZNF468
ZNF473
ZNF544
ZNF551
ZNF576
ZNF586
ZNF587
ZNF671
ZNF74
ZNF8
ZNF83
ZNF85

APPEND03 – Candidate MES fCNVs

KLHL9
ABI1
TSFM
GHITM
TOPORS
CEBPD
CLOCK
ST7
STAM
MDM4
BLNK
DDIT3
TFR2
AKT1
MARS
PFKP
CDK4
CAMK2N1
DDX56
GTPBP4
IRGQ
MLC1
MET
ECD
CTDSP2
CCDC6
IFT74
DTX3
PDGFRA
RSU1
NPTX2
NFKBIA
SOX13
CUTC
CAPZA2
KIF5B
CARHSP1
FRMD4A
TSPAN31
NUP107
NRBF2
CYorf15B
SAR1A
GLI1
DCLRE1C
ECHDC3
CLEC11A
ITGB1
TEK
ATP5C1
RPL28
LANCL2
MAX
ITGA8
ARL3
SEC61G
KIAA1797
RBBP5
GSTT1
VGF
NMT2
ZNF134
MLLT3
ZCWPW1
USP9X
MDH2
USP2
FAM53B
LY6H
ETNK2
CDKN2A
KIN
ELAVL2
HLA-DQA1
METTL1
KCNH2
PAPD1
AVIL
SPAG6
CBARA1
BCAT2
DUS4L
CUL2
PIN4
PRC1
MTAP
IL15RA
PRKCQ
HP1BP3
GFAP
KCTD15
TAX1BP1
PIK3C2B
SERINC1
CDK5
EIF4G2
THNSL1
PRKG1
PHYH
FIP1L1
NRP1
PRKY
SNX13
SLC25A22
CHIC2
WDR37
OPTN
PDAP1
ZNF586
CYP27B1
PDLIM7
EIF1AX
ITIH2
C7orf26
VPS26A
STS
PPP3CB
IL6
UPP1
MINPP1
PHKG1
KIF5A
CCT6A
MEOX2
PLEKHA6
DDX21
CCDC106
GLUD1
DDX3X
TMEM8
AP3D1
OBFC1

RPS24
MBD6
AKR1C2
EIF1AY
HOXA10
RPP30
WIPI2
MRPS17
DIO2
MDM2
CAMK2G
C9orf82
B4GALNT1
SCRG1
DDX50
ANXA7
NDUFA3
CH25H
EGFR
DCTN2
RHOBTB1
CPM
UBE2D1
PANK4
GBAS
ANXA11
MSRB2
OS9
KIT
KIAA1128
ZNF132
GNA12
UBAP1
GNAI1
PSPH
ZNF14
SLC35E3
HDAC9
SMARCD3
EMP3
SLC26A10
INHBE
KDELR2
C10orf72
CUGBP2
ZNF671
SNAPC2
GAS7
SLC25A28
GDI2

APPEND04 - SCRIPTS

SCRIPTS: EUCLIDEAN DISTANCE SUBTYPE CLASSIFIER

```
#!usr/bin/perl -w

$count=0;

open(IN,"/Users/SupernerdMkII/Desktop/EUC classifier data/classifiers3.tab") or die;
while(<IN>){
        chomp $_;

        @array=split/\t/,$_;

        $vector{$array[0]}=$_;
}
close IN;

@hubs=keys(%vector);

$count=0;

open(IN,"list3.tab") or die;
#open(IN,"/Users/SupernerdMkII/Desktop/whole tumor expression
sets/rembrant_data.exp") or die;
while(<IN>){
        chomp $_;
        @array=split/\t/,$_;
        @distances=();

        if($count==0){$count++;next;}

#       print $array[0]."\n";

        foreach $hub(@hubs){
                chomp $hub;
                @hub=split/\t/,$vector{$hub};
                $vector="vector";
                $distance{$hub}=EUCDIST(\@array,$vector{$hub});
                push(@distances, $distance{$hub});

                print "$hub\t$distance{$hub}\t";
        }

#       print "$array[0]\t@distances\n";

        @distances = sort {$a<=>$b} @distances;
        $min=$distances[0];

        foreach $hub(@hubs){
                if($distance{$hub}==$min){$classy=$hub;}
        }

        print "$array[0]\t$classy\n";
}
```

```
sub EUCDIST{
        my($arrayref,$coord)=@_;

        my @array1=@$arrayref;
        my @array2=split/\t/,$coord;

        my $eucd=0;

        foreach $i(1..scalar(@array1)-1){
                $eucd+=($array1[$i]-$array2[$i])**2
        }

        $eucd=sqrt($eucd);
        return $eucd;
}
```

SCRIPTS: CO-MUTATION NETWORK BUILDER

```
#!usr/bin/perl -w

$thresh=0.168; #threshold of CNV reads to call amp or del

open(IN,"<list1.tab") or die;
while(<IN>){
        chomp $_;
        $candidates{$_}=1;    #defined as hash to extract exact CNV vectors from later
arrays
        $defined{$_}=0;                 #defined to mark genes whose comparisons have
been done to avoid redundant operations
}
close IN;

@candidatekeys=keys(%candidates);        #defined for the actual pairwise checking

print "Candidates read.\n";

open(IN, "<wholeCNVsgenesonly.tab") or die; #cnv matrix file
while(<IN>){
        chomp $_;
        @array=split/\t/,$_;

        if(defined($candidates{$array[0]})){
                $candidates{$array[0]}=$_;
        }
}
close IN;

print "CNV vectors read.\n";

open(OUT, ">pairwise CNV results2.tab") or die;

foreach $gene(@candidatekeys){
        print "$gene START\n";

        foreach $gene2(@candidatekeys){

                ## do not compare a gene against itself, and do not repeat comparisons
that have already been done
                next if($gene eq $gene2);
                next if($defined{$gene2}==1);

                ## define/reset counters
                $Aamp=0;
                $Bamp=0;
                $Camp=0;
                $Damp=0;

                $Adel=0;
                $Bdel=0;
                $Cdel=0;
                $Ddel=0;

                @array1=split/\t/,$candidates{$gene};
                @array2=split/\t/,$candidates{$gene2};

                foreach $i(1..scalar(@array1)-1){
                        #check AMP
                        if($array1[$i]<=-$thresh and $array2[$i]<=-$thresh){ $Aamp++; }
                        elsif($array1[$i]<=-$thresh and $array2[$i]>-$thresh){ $Bamp++; }
                        elsif($array1[$i]>-$thresh and $array2[$i]<=-$thresh){ $Camp++; }
                        elsif($array1[$i]>-$thresh and $array2[$i]>-$thresh){ $Damp++; }

                        #check DEL
                        if($array1[$i]>=$thresh and $array2[$i]>=$thresh){ $Adel++; }
                        elsif($array1[$i]>=$thresh and $array2[$i]<$thresh){ $Bdel++; }
                        elsif($array1[$i]<$thresh and $array2[$i]>=$thresh){ $Cdel++; }
                        elsif($array1[$i]<$thresh and $array2[$i]<$thresh){ $Ddel++; }

                }

                $pamp=PValue($Aamp,$Bamp,$Camp,$Damp);
                $pdel=PValue($Adel,$Bdel,$Cdel,$Ddel);

#               if($gene eq "ECHDC3"){          print
"$gene2\t$Adel\t$Bdel\t$Cdel\t$Ddel\n";         }

                print OUT "$gene\t$gene2\tAMP:\t$pamp\tDEL:\t$pdel\n";
        }

        $defined{$gene}=1;
}


##############
#LNFACTORIAL
###############
sub LnFactorial{
        my $n=shift;
        $lnn=0;

        while($n>=1){
                $lnn+=log($n);
                $n--;
        }

        return $lnn;
}

##############
```

```
#Probability of one table
###############
sub ProbTable{
        my ($a , $b , $c, $d) = @_;
        my $n = $a + $b + $c + $d;
        my $LnNumerator     = LnFactorial($a+$b)+
                LnFactorial($c+$d)+
                LnFactorial($a+$c)+
                LnFactorial($b+$d);

        my $LnDenominator   = LnFactorial($a) +
                LnFactorial($b) +
                LnFactorial($c) +
                LnFactorial($d) +
                LnFactorial($n);

  my $LnP = $LnNumerator - $LnDenominator;
  return exp($LnP);
}

##############
#p-value calculator
###############
sub PValue{
        my ($a, $b, $c, $d) = @_;

        my $n = $a + $b + $c + $d;

        my $p = 0;

        my $min;

        $p+=ProbTable($a,$b,$c,$d);

#       if($a*$d >= $b*$c){
                $min = ($c < $b) ? $c : $b;
                $i=0;

                while($i<=$min){
                        $a++;
                        $b--;
                        $c--;
                        $d++;

                        $p+=ProbTable($a,$b,$c,$d);
                        $i++;
                }
                if($p>1){$p=1;}

                return $p;
}
```

SCRIPTS: GENETIC-GENOMIC ALGORITHM

```
#!/usr/bin/perl -w

($exp,$cnv,$out,$m,$b, $FDR)=@ARGV; #input filepaths
chomp $exp;	#gene expression matrix
chomp $cnv; #CNV matrix
chomp $out; #base name of results file (statistics such as the kernel used will be
appended to this)
chomp $m;	#slope of linear fit of function -log(p) null distribution
chomp $b;	#intercept of linear fit of function -log(p) null distribution
chomp $FDR; #desired FDR threshold

#	parameters for NBL p-value estimation: derived from linear fit of -log(pnull)
#	$m = 12.22
#	$b = 0.4545

#	new NBL set
#	$m = 20.654
#	$b = 0.3035

#$kernel=0.12297;

$kernel=0.852;

$first=0;

###################################################################
######################## ACQUIRE DATA ##########################
###################################################################

## NOTE ##
## This script is hardcoded to accept tab-delimited files with the first row
## and column corresponding to the gene names and patient IDs, respectively.
## These IDs must match in formatting across the exp and cnv files, but do NOT
## have to be matched in order.

## This stage of the script indexes all of the information contained in both the
## CNV and expression files using a multi-dimensional hash. From here in, all data
## for the analysis can be dynamically called from the hashes stored in memory,
allowing
## for maximum computational efficiency

## input a file with <geneID> <expression vector>
open(IN, "<$exp") or die;
while(<IN>){
	chomp $_;
	@array=split/\t/,$_;

	if($first==0){
		foreach $i(1..scalar(@array)-1){
```

```
            $expressionnameposition{$array[$i]}=$i-1; #first entry will not be in
the vectors
        }
        $first++;
    }

    else{
        $string=join("\t",@array[1..scalar(@array)-1]);
        $geneexpressionvector{$array[0]}=$string;
    }

}
close IN;

$first=0;

print "Expression array finished.\n";

## input a CNV file <genename> <CNVvector>
open(IN,"<$cnv") or die;
while(<IN>){
    chomp $_;
    @array=split/\t/,$_;

    if($first==0){
        foreach $i(1..scalar(@array)-1){
            $cnvnameposition{$array[$i]}=$i-1; #first entry will not be in the
vectors
        }

        $first++;
    }

    else{
        $string=join("\t",@array[1..scalar(@array)-1]);
        $geneCNVvector{$array[0]}=$string;
        push(@genes,$array[0]);

    }
}
close IN;

print "CNV array finished.\n";


################################################
############ BEGIN      IDENTIFICATION #############
################################################


################################ define fCNVGs
```

```
## This stage of the script is the active analysis of identifying functional f-CNVs
## Each gene locus's cnv and expression vectors are retrieved and the MI between
these
## is computed. If the MI value passes the specified FDR, it is flagged as an f-CNV,
## but no second-degree analysis is conducted yet. This module will output a list of
## f-CNVs. The user can then allow the script to proceed or break this output list into
## smaller lists and parallel process to minimize the time needed to complete the
analysis


@pvalues=();

print "***Defining fCNVGs***\n";

foreach $gene(@genes){
        chomp $gene;

        next if(defined($geneexpressionvector{$gene})==0);

        @CNVvector=split/\t/,$geneCNVvector{$gene};
        @EXPvector=split/\t/,$geneexpressionvector{$gene};

        $MI=MI(\@CNVvector,\@EXPvector,\%cnvnameposition,\%expressionnameposit
ion,$kernel);

        $p= 10**-($m*$MI+$b);

        push(@pvalues,$p);

        $fCNVtest{$gene}=$p;

        print "$gene\t$MI\t$p\n";       #output to screen or dumpfile so users can actively
track progress
}

$q=QTHRESH(\@pvalues,$FDR);

@keys=keys(%fCNVtest);

foreach $gene(@keys){
        if($fCNVtest{$gene}<=$q){
                $fCNVG{$gene}=$fCNVtest{$gene}."\t";
        }
}

@FCNVGs=keys(%fCNVG);           #now contains all CNVs passing FDR specified for
f-CNVG
@genes=keys(%geneexpressionvector);

open (OUT, ">$out"."_k_".$kernel."_FDR_".$FDR."_uncharacterized.tab") or die;
```

```
foreach $gene(@FCNVGs){   print OUT "$gene\t$fCNVtest{$gene}\n";      }
close OUT;

print "***".scalar(@FCNVGs)." fCNVGs defined. Characterizing.***\n";


################################ characterize fCNVGs

## This stage of the script takes the results from the previous analysis and recurses,
## measuring the MI between the CNV vector of the fCNV locus and the expression of
## every gene in the genome.
## This results in the linking of genes whose expression shows significant correlation by
MI
## to the mutational state of the fCNV, and is therefore potentially regulated by the fCNV


open(OUT, ">$out"."_k_".$kernel."_FDR_".$FDR.".tab") or die;

select((select(OUT),$|=1)[0]); #flush writing to OUT so that log can be checked

foreach $gene(@FCNVGs){
        chomp $gene;

        $string="$gene\t$fCNVG{$gene}";
        @pvalues=();
        @CNVvector=split/\t/,$geneCNVvector{$gene};

        foreach $gene2(@genes){
                chomp $gene2;
                next if($gene eq $gene2);

                @EXPvector=split/\t/,$geneexpressionvector{$gene2};


        $MI=MI(\@CNVvector,\@EXPvector,\%cnvnameposition,\%expressionnameposit
ion,$kernel);

                $p= 10**-($m*$MI+$b);

                push(@pvalues,$p);

                $target{$gene2}=$p;
        }

        $q=QTHRESH(\@pvalues,$FDR);

        @keys=keys(%target);

        foreach $temp(@keys){
                if($target{$temp}<=$q){
                        $string=$string.$temp."\t".$target{$temp}."\t";
                }
        }

        print OUT "$string\n";
}

close OUT;

################################
######## FUNCTIONS #############
################################

sub MI{
        my($CNVref,$expref,$CNVnames,$expnames,$kernel2)=@_;
        my @CNV=@$CNVref;
        my @exp=@$expref;

        my @samplekey=keys(%{$expnames});

        my $xo;
        my $yo;

        my $top;
        my $bottom;

        my $MI=0;
        my $topsum=0;
        my $bottomsum1=0;
        my $bottomsum2=0;
        my $bottomtotal=0;

        my $M=0;

        foreach $sample(@samplekey){
                next if(defined(${$CNVnames}{$sample})==0);
                $xo=$CNV[${$CNVnames}{$sample}];
                $yo=$exp[${$expnames}{$sample}];
                $topsum=0;
                $bottomsum1=0;
                $bottomsum2=0;

                $M=0;

                foreach $sample2(@samplekey){
                        next if(defined(${$CNVnames}{$sample2})==0);
                        next if($sample2 eq $sample);
                        $M++;


        $top=JOINT($xo,$CNV[${$CNVnames}{$sample2}],$yo,$exp[${$expnames}{$sa
mple2}],$kernel2);
```

```
            $bottom1=MARG($xo, $CNV[${$CNVnames}{$sample2}] ,
$kernel2 );
            $bottom2=MARG($yo, $exp[${$expnames}{$sample2}] , $kernel2
);


            $topsum+=$top;
            $bottomsum1+=$bottom1;
            $bottomsum2+=$bottom2;

        }

        $topsum*=(1/$M)*((1/(2*3.14159*$kernel2**2)));
        $bottomsum1*=(1/$M)*((1/(sqrt(2*3.14159*$kernel2))));
        $bottomsum2*=(1/$M)*((1/(sqrt(2*3.14159*$kernel2))));

        $bottomtotal=$bottomsum1*$bottomsum2;

        # The following condition was added as a failsafe in the event that a zero
value
        # is somehow obtained from the bottom marginal functions (resulting in
division by zero)
        if($bottomtotal==0){    $bottomtotal=0.0000001;      }

        #calculate MI here
        $MI+=(log($topsum/$bottomtotal)/log(10));

    }

    $MI*=(1/$M);

    return $MI;
}


#join probability density function of 2 variables using Gaussian kernel
sub JOINT{
    my($xo, $xi, $yo, $yi, $h)=@_;

    my $joint = exp(-((($xo-$xi)**2+($yo-$yi)**2)/(2*$h**2)));

    return $joint;
}

#marginal probability density function of a variable using Gaussian kernel
sub MARG{
    my($xo, $xi, $h)=@_;

    my $marg= exp(-((($xo-$xi)**2)/(2*$h**2)));

    return $marg;
```

```
}

#compute qscore for FDR signficance
sub QTHRESH{
	my($arrayname,$threshold)=@_;

	my $q=0;
	my @scores = sort {$a<=>$b} @$arrayname;
	my $m=scalar(@scores);

	foreach $i(0..scalar(@scores)-1){
		my $k=$i+1;

		if($scores[$i]<=(($k/$m)*$threshold)){		$q=$scores[$i];		}
	}

	return $q;
}
```

SCRIPTS: NULLDISTRIBUTION / KERNEL OPTIMIZER

```
#!usr/bin/perl -w

($exp,$cnv,$out)=@ARGV; #input filepaths
chomp $exp;
chomp $cnv;
chomp $out;

#$kernel=0.0164;
$kernel=0.852;

$first=0;

######################################################################
####
######################## ACQUIRE DATA
####################################
######################################################################
####

open(IN, "<$exp") or die;
while(<IN>){
	chomp $_;
	@array=split/\t/,$_;

	if($first==0){
		foreach $i(1..scalar(@array)-1){
			$expressionnameposition{$array[$i]}=$i-1; #first two entries will
not be in the vectors
		}
		$first++;
	}

	else{
		$string=join("\t",@array[1..scalar(@array)-1]);
		$geneexpressionvector{$array[0]}=$string;
		push(@genes,$array[0]);
	}

}
close IN;

$first=0;

print "Expression array finished.\n";

open(IN,"<$cnv") or die;
while(<IN>){
	chomp $_;
	@array=split/\t/,$_;
```

```
        if($first==0){
                foreach $i(1..scalar(@array)-1){
                        $CNVnameposition{$array[$i]}=$i-1; #first entry will not be in the
vectors
                }

                $first++;
        }

        else{
                $string=join("\t",@array[1..scalar(@array)-1]);
                $geneCNVvector{$array[0]}=$string;
        }
}
close IN;

print "CNV array finished.\n";

#################################################################
#################### NULLDISTRIBUTION ####################
#################################################################

print "Computing Null Distribution.\n";

$i=0;

open(OUT,">$out") or die;

while($i<=10000){
        $int=int(rand(scalar(@genes)-1));
        next if(defined($geneCNVvector{$genes[$int]})==0);
        @CNVs=split/\t/,$geneCNVvector{$genes[$int]};


        $int=int(rand(scalar(@genes)-1));
        @expression=split/\t/,$geneexpressionvector{$genes[$int]};

        $information=MI(\@CNVs, \@expression, \%expressionnameposition,
\%CNVnameposition, $kernel);


        print $information."\n";
        print OUT $information."\n";

        $i++;
}

close OUT;

####################################################
```

```
################ SUBROUTINES ################
################################################

sub MI{
        my($CNVref,$expref,$CNVnames,$expnames,$kernel2)=@_;
        my @CNV=@$CNVref;
        my @exp=@$expref;

        my @samplekey=keys(%{$expnames});

        my $xo;
        my $yo;

        my $top;
        my $bottom;

        my $MI=0;
        my $topsum=0;
        my $bottomsum1=0;
        my $bottomsum2=0;
        my $bottomtotal=0;

        my $M=0;

        foreach $sample(@samplekey){
                next if(defined(${$CNVnames}{$sample})==0);
                $xo=$CNV[${$CNVnames}{$sample}];
                $yo=$exp[${$expnames}{$sample}];
                $topsum=0;
                $bottomsum1=0;
                $bottomsum2=0;

                $M=0;

                foreach $sample2(@samplekey){
                        next if(defined(${$CNVnames}{$sample2})==0);
                        next if($sample2 eq $sample);
                        $M++;

                        $top=JOINT($xo , $CNV[${$CNVnames}{$sample2}] , $yo ,
$exp[${$expnames}{$sample2}] , $kernel2);
                        $bottom1=MARG($xo, $CNV[${$CNVnames}{$sample2}] ,
$kernel2 );
                        $bottom2=MARG($yo, $exp[${$expnames}{$sample2}] , $kernel2
);


                        $topsum+=$top;
                        $bottomsum1+=$bottom1;
                        $bottomsum2+=$bottom2;

                }
        $topsum*=(1/$M)*((1/(2*3.14159*$kernel2**2)));
        $bottomsum1*=(1/$M)*((1/(sqrt(2*3.14159*$kernel2))));
        $bottomsum2*=(1/$M)*((1/(sqrt(2*3.14159*$kernel2))));

        if($topsum ==0) {       $topsum=0.0000001; }

        $bottomtotal=$bottomsum1*$bottomsum2;

        if($bottomtotal==0){    $bottomtotal=0.0000001;       }

        #calculate MI here
        $MI+=(log($topsum/$bottomtotal)/log(10));

    }

    $MI*=(1/$M);

    return $MI;
}


#join probability density function of 2 variables using Gaussian kernel
sub JOINT{
    my($xo, $xi, $yo, $yi, $h)=@_;

    my $joint = exp(-((($xo-$xi)**2+($yo-$yi)**2)/(2*$h**2)));

    return $joint;
}

#marginal probability density function of a variable using Gaussian kernel
sub MARG{
    my($xo, $xi, $h)=@_;

    my $marg= exp(-((($xo-$xi)**2)/(2*$h**2)));

    return $marg;

}
```

SCRIPTS: GENETIC-GENOMICS BY U-TEST (REPLACED w/ MI)

```
#!usr/bin/perl -w

$initial=0;
$CNVthreshold=0.168;

open(IN, "<wholeCNVsgenesonly.tab") or die;
while(<IN>){
        chomp $_;
        @array=split/\t/,$_;

        #hash tcga sample labels (sample name as keys, position as variable)
        if($initial==0){
                foreach $i(1..scalar(@array)-1){
                        $CNVaddress{$array[$i]}=$i;
                }
                $initial++;
        }


        #hash CNV vectors (gene name as key, entire string as variable)
        else{
                $CNVvector{$array[0]}=$_;
        }
}
close IN;

#CNVaddress hash = addresses
#CNVvector hash = values

#reset initializer
$initial=0;

open(IN, "<gene_in_network_expression.exp") or die;
while(<IN>){
        chomp $_;
        @array=split/\t/,$_;

        #hash tcga sample labels again (sample positions as keys, labels as variable
        if($initial==0){
                foreach $i(2..scalar(@array)-1){
                        $EXPaddress{$i}=$array[$i];
                }
                $initial++;
        }

        else{
                @expression2=();

                next if(exists($CNVvector{$array[1]})==0);
                #for each gene, push expression into an array and hash the expression
(key) to the sample (address)
                foreach $i(2..scalar(@array)-1){
                        push(@expression2, $array[$i]);
                        $EXPvector{$array[$i]}=$i;
                }

                #call and array CNV vector of same gene
                @CNVrefarray=split/\t/,$CNVvector{$array[1]};

                for ($i=scalar(@expression2)-1; $i>=0; $i--){

        if(defined($CNVaddress{$EXPaddress{$EXPvector{$expression2[$i]}}})==0){
                    splice(@expression2,$i,1);
                        }
                }

                #reorder expression array ascending
                @expression=sort {$a <=> $b} @expression2;

                #define 3 variables: amp del norm, initialize to zero
                $amp=0;
                $del=0;

                $ampcount=0;
                $delcount=0;
                $normcount=0;

                @amparray=();
                @delarray=();

                foreach $i(0..scalar(@expression)-1){

        $CNVvalue=$CNVrefarray[$CNVaddress{$EXPaddress{$EXPvector{$expressio
n[$i]}}}];

                        if($CNVvalue >= $CNVthreshold){
                    push(@delarray,$expression[$i]);
                        }

                        elsif($CNVvalue <= -$CNVthreshold){
                    push(@amparray,$expression[$i]);
                        }

                        else{
                    push(@delarray,$expression[$i]);
                    push(@amparray,$expression[$i]);
                    $normcount++;
                        }
                }

                foreach $i(0..scalar(@amparray)-1){

        $CNVvalue=$CNVrefarray[$CNVaddress{$EXPaddress{$EXPvector{$expressio
n[$i]}}}];

                        if($CNVvalue <= -$CNVthreshold){$amp+=($i+1); $ampcount++;}
                }

                foreach $i(0..scalar(@delarray)-1){

        $CNVvalue=$CNVrefarray[$CNVaddress{$EXPaddress{$EXPvector{$expressio
n[$i]}}}];

                        if($CNVvalue >= $CNVthreshold){$del+=($i+1); $delcount++;}
                }

                $UscoreAMP=$amp - ($ampcount * ($ampcount+1))/2 ;
                $UscoreDEL=$del - ($delcount * ($delcount+1))/2 ;

                $Uampmax=$ampcount*$normcount;
                $Udelmax=$delcount*$normcount;

                if($UscoreAMP < $Uampmax/2){ $UscoreAMP=$Uampmax-
$UscoreAMP; }

                $zAMP=NullDist($ampcount,$normcount, $UscoreAMP);

                print "$array[1]\t$ampcount\t$zAMP\n";
        }
}
close IN;




sub NullDist{
        my($n1,$n2,$Ufound)=@_;
        my $total=$n1+$n2;
        my @Uray=();
        my @n1=();

        #print "$n1\t$n2\t$U\n";
        my $trials=5000;
        #$count=0;

        foreach $i(1..$trials){
                my $temp=$total;
                my @n2=(0..$total);
```

```
            #print "entering first loop\n";
            while(scalar(@n1) <= $n1){
                my $int=int(rand($temp-1))+1;

                push(@n1,$n2[$int]);
                splice(@n2,$int,1);
                $temp--;
            }

            my $sum=0;

            foreach $entry(@n1){
                $sum+=$entry;
            }

            $U=$sum-(scalar(@n1) * (scalar(@n1)+1)/2);

            if($U < $n1 * $n2 - $U){ $U = $n1 * $n2 - $U;}

            push(@Uray,$U);
            @n1=();
        }

        my $average=0;

        foreach $entry(@Uray){
            $average+=$entry;
        }

        $average=$average/scalar(@Uray);

        my $stdev=0;

        foreach $entry(@Uray){
            $stdev+=($entry-$average)**2;
        }

        $stdev=sqrt($stdev/(scalar(@Uray)-1));

        #print "$n1\t$n2\t$average\t$stdev\n";

        my $z=($Ufound-$average)/$stdev;
        @Uray=();
        return $z;
    }
```

APPEND05 – Manuscript Figures and Figure Legends

Figure1.

(a) 1 Network map showing co-mutated f-CNVG loci as a function of the statistical associations of genes harboring amplifications (blue nodes) or deletions (red nodes). Edges denote a statistically significant association between connected f-CNVGs ascertained by Fisher Exact's Test (FET) ($p < 0.05$ Bonferroni corrected). The connected nodes' clustering distance is also based on strength of association i.e. juxtaposed nodes are more significantly associated to each other than distantly connected nodes. Chromosome location for the larger clusters, and the location of the *C/EBPδ* and *KLHL9* nodes are provided. (b) 483 loci were identified as bearing functional CNV genes (f-CNVG). Presented is a statistical summary of associations of the f-CNVGs passing these criteria to the poor-prognosis subtype versus the good-prognosis subtype, including "classical" oncogenesis f-CNVGs. Marks indicate amplifications (+) deletions (-) and diploid (WT) for each gene.

Figure2.

(a) The f-CNVGs statistically co-occurring with amplifications of *C/EBPδ* or deletions of *KLHL9*, and associated with the poor prognosis phenotype across all TCGA samples were retrieved from the f-CNVG association network from Figure 1a (amplifications as blue nodes, deletions as red nodes, edges denoting significant association). Each f-CNVG in the cluster for *C/EBPδ* (b) or *KLHL9* (c) was then conditionally tested in pair-wise fashion for association to the poor prognosis subtype; color grading corresponds to the -log(*p*) of the association (white cells indicate $p > 0.1$) of the tested locus (rows) to the poor prognosis phenotype upon conditioning for the absence of the indicated, conditioned locus (columns). I.e., conditioning on *KLHL9* abrogates the association of all other f-CNVGs in its cluster to the poor-prognosis phenotype (white column), while only one locus can abrogate association of *KLHL9* (red row). Only *KLHL9* and *C/EBPδ* (indicated in bold and boxed in) remove all associations across their respective co-mutated clusters when conditioned for, yet remained robust to conditioning on the other genes in the cluster. The average -log(*p*) value associated with each conditioned locus across all tested loci is provided in the last row of each heatmap.

Figure3.

(a) Genomic q-PCR analysis of an independent cohort of 63 patients reveals a high enrichment of KLHL9 deletions in patients with poor prognosis. y-axis is reported in CT values with the cutoff for statistically significant evidence of a deletion presented as a red line; all CT values above the red line indicate evidence of deletion. CT values are reported as mean ±SEM (b) Statistical analysis of the results presented in Figure 3b shows a highly significant association of samples bearing a KLHL9 deletion to the poor prognosis cohorts. (c) Kaplan-Meier curves of patients based on their genotype at the KLHL9 and CEBPD loci. *x*-axis represents post-diagnostic survival in months, *y*-axis the percentage of patients surviving at measured time points. The presence of *C/EBPδ* amplifications and *KLHL9* deletions in patient samples (red line) is sufficient to separate poor prognosis from good prognosis patients in GBM TCGA samples, as defined by expression of PN signature markers (black line), and even compared to pooled Good and Poor prognosis samples lacking those mutations (blue line) at a statistically significant level (p_1 = red vs. blue, p_2 = red vs black, alpha=0.05). Distribution of prognoses of individual patients with tumors bearing the *C/EBPδ* amplification (*C/EBPδ+*) and *KLHL9* deletions (*KLHL9-*) are indicated as hashes in the labeled boxes below the graph. Hashes note the time of death after diagnosis of each patient bearing the corresponding mutation across all samples. (d) IHC probing for CEBPB and CEBPD proteins in these primary tissue samples reveals a strong correlation between elevated CEBP expression in GBM tumors and the MES subtype, as well as a significant association to KLHL9 deletions.

Figure4.

Effect of *KHLH9* expression on *C/EBPβ/δ* and *YKL40* mRNA and protein levels 96 hours post-induction. (a) *KHLH9* mRNA expression levels by qPCR in the two inducible *KHLH9* clones and GFP control cells. (b) RNAseq analysis of these cells after 48 hours of induction reveals significant dysregulation in ARACNE-predicted transcriptional targets of *C/EBPβ*, *C/EBPδ*, and a significant reprogramming away from the canonical MES subtype. Benchmark mesenchymal markers are indicated on the barcodes for reference. No significant change in *C/EBPβ* or *C/EBPδ* expression levels were observed (c) KHLH9, *C/EBPβ*, *C/EBPδ*, and *STAT3* protein levels 72h after Dox-mediated induction of *KHLH9* expression. B-actin was using as housekeeping control gene.

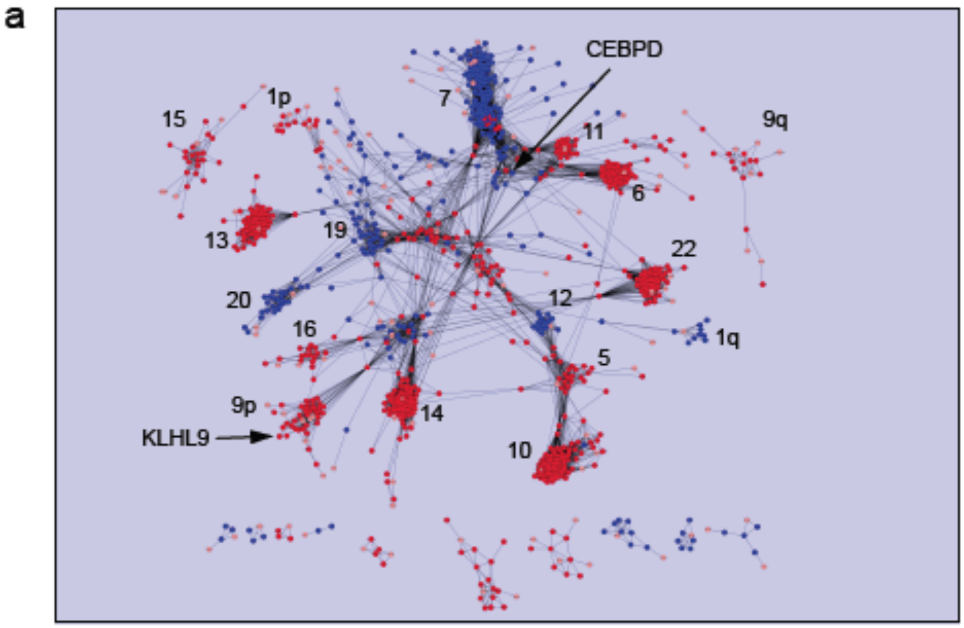
Figure5.

Protein half-life time course for *C/EBPβ/δ* conditioned on *KLHL9* expression. Abbreviations: DOX = doxycycline, MG132 = proteasome inhibitor. (a) A 4-hr exponential time courses conducted for protein half-life in *KLHL9*-induced , *KLHL9*-induced-proteasome-inhibited, and GFP-induced SF210 cells with cyclohexamide treatment. Cells expressing *KLHL9* in the presence of cyclohexamide showed a protein half-life of ~1hr for *C/EBPβ/δ* proteins that was not observed in GFP controls (>2hr half life). Addition of MG132 to *KLHL9*-expressing cells restored the half-life of the CEBP proteins to those observed in the *KLHL9*-null GFP controls. (b) Co-immunoprecipitation of KLHL9 shows an interaction between KLHL9 and CEBP proteins (c) Immunoprecipitation of *C/EBPβ* and *C/EBPδ* proteins and probing for ubiquitylation reveals increased concentrations of poly-ubiquitylated CEBP proteins only when *KLHL9* is expressed; conversely, precipitating ubiquitylated species and probing for CEBPs corroborates this observation. These IPs (in Figure 5c) were performed under the presence of MG132.

Figure6.
(a) A basic representation of the KLHL9 protein showing key functional domains, and the structure of the mutant KLHL9-ΔBTB. (b) Ubiquitin IPs of the three constructs following 24 hours of expression and 4 hours of MG-132 treatment shows suppression of ubiquitylated CEBP species when a mutant, inactive form of KLHL9 is used for rescue. (c) Western blotting of whole cell lysates after 48 hours of exogenous expression of either KLHL9, KLHL9-ΔBTB, or NT.

Figure7.
(a) Induction of *KLHL9* is followed by the appearance of large, circular cells with large nuclei, visualized by blue nuclear stain. These cells also do not incorporate EdU (red) when exposed to it over 24 hours, unlike the normal, fibroblast-like counterparts. Cells that do not incorporate EdU (and are therefore considered non-proliferative) have nuclei that appear blue in the composite image; arrows demarcate the nuclei of these cells. (b) Flow cytometry of BrdU incorporation by *KLHL9*-induced (red) and uninduced cells (black) after 24 hours of BrdU exposure is presented as histograms internally normalized to the highest peak. A BrdU negative control is provided (gray). Color-coded integrations for the area under the defined peaks are provided as left-peak : right-peak percentage ratios. (c) Cell viability measured as a function of ATP activity in *KLHL9*-expressing vs *KLHL9*-nonexpressing cells and GFP controls. Closed datapoints represent normalized cell proliferation in DOX-induced samples and open datapoints represent DOX-negative cells. Data is presented as the mean ± SEM.

Figure 1. Genetical Genomics identifies functional amplifications of CEBPD and deletions of KLHL9 as candidate drivers of poor-prognosis GBM



b

Functional Mutation	Odds Ratio (Poor vs Good)	pFET (Poor v Good)	# Poor Prognosis
CEBPD^{amp}; KLHL9^{-}	**undef***	**1.940-3**	**17**
CEBPD^{amp}; KLHL9^{WT}	**9.50**	**9.900e-3**	**14**
KLHL9^{-} ; CEBPD^{WT}	**5.47**	**2.120e-5**	**38**
EGFR^{amp} ; KLHL9^{WT},CEBPD^{WT}	1.35	0.2267	**Total: 69/144, 48%**
CDKN2A^{-} ; KLHL9^{WT},CEBPD^{WT}	1.03	0.5243	
CDK4^{amp} ; KLHL9^{WT},CEBPD^{WT}	0.631	0.9096	
PDGFRA^{amp} ; KLHL9^{WT},CEBPD^{WT}	0.282	0.9980	

*no good prognosis samples bear this genotype

Figure 2. f-CNVGs CEBPD and KLHL9 account for the association to poor prognosis of all co-occuring mutations

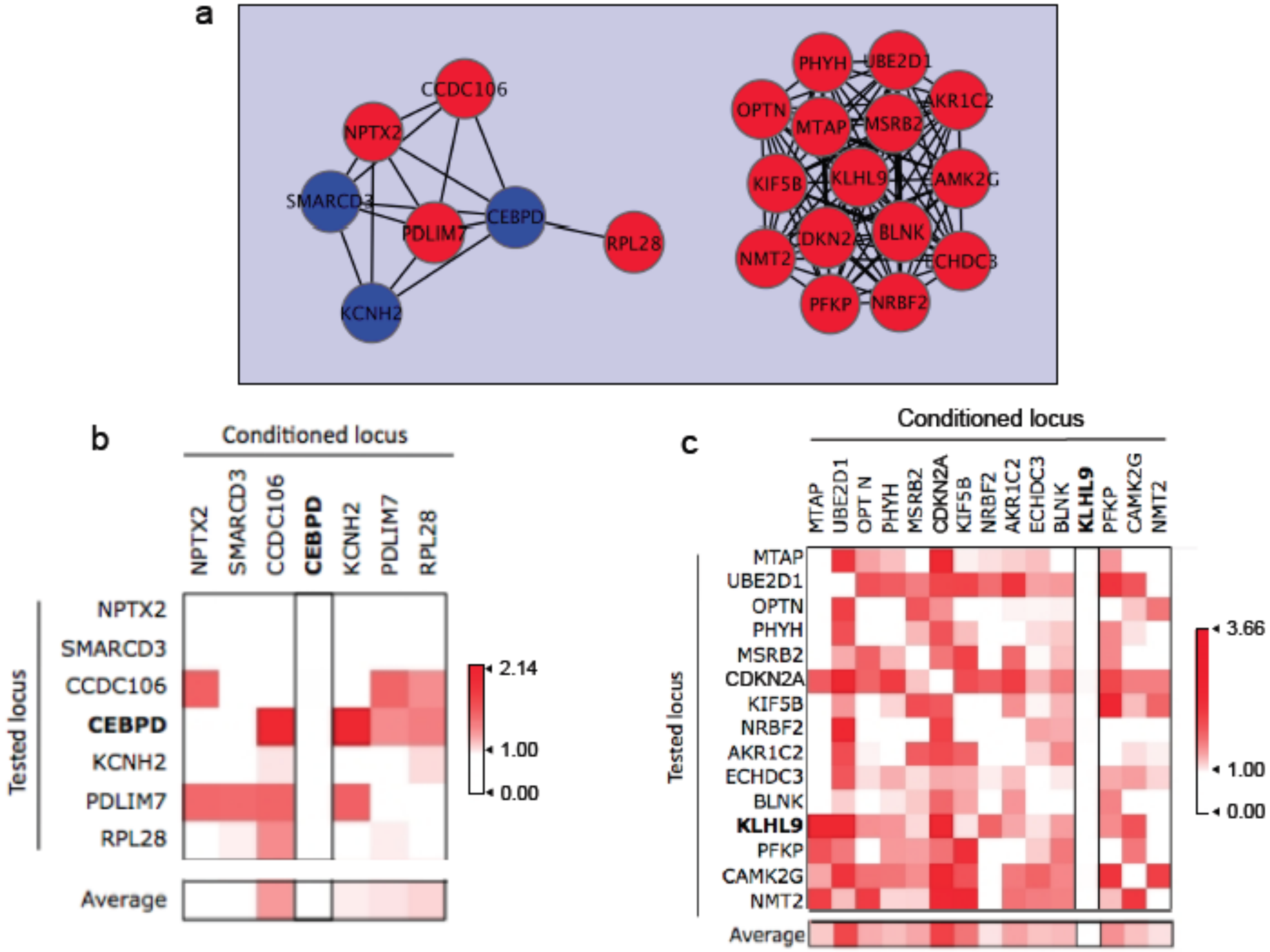
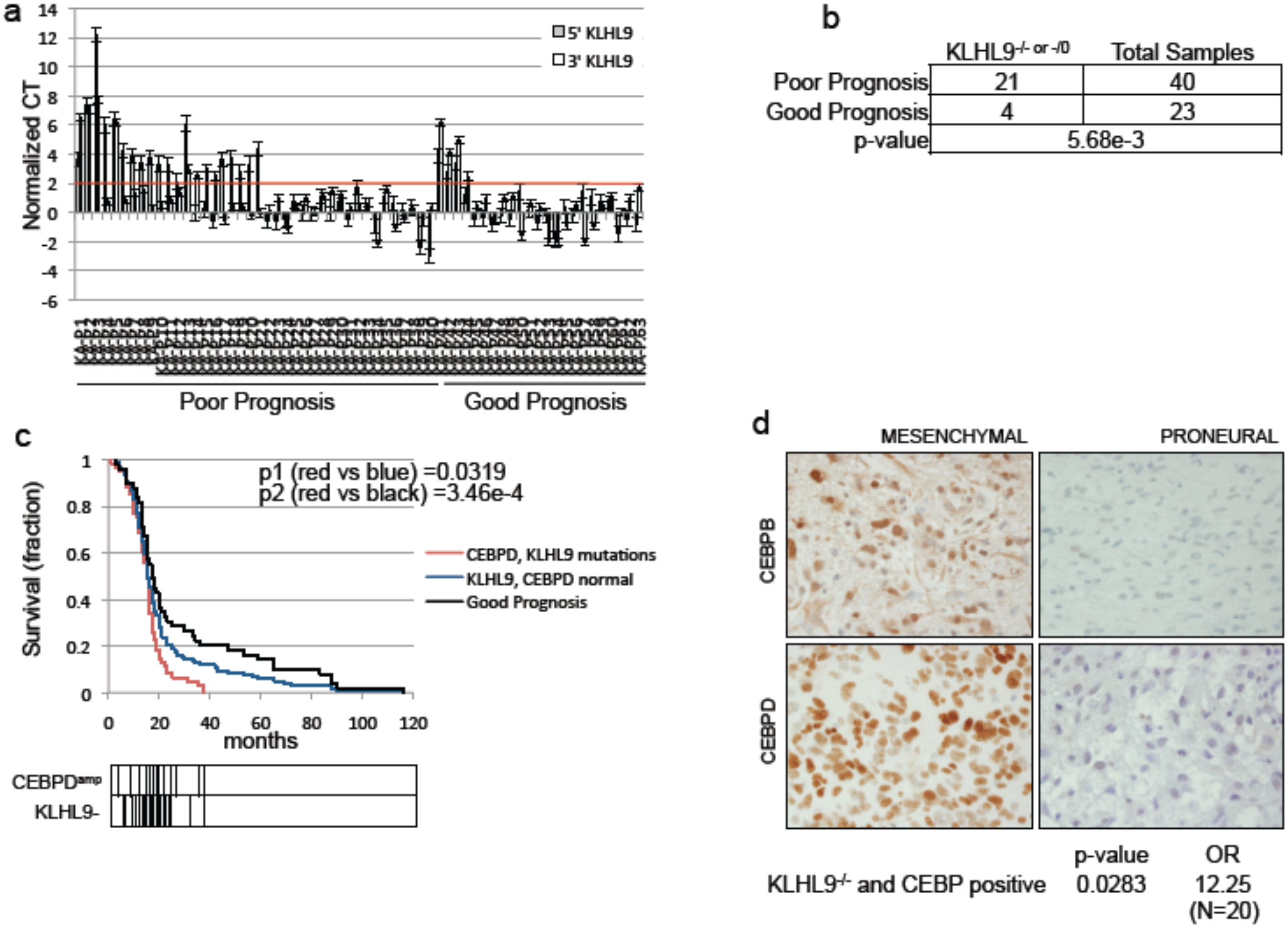


Figure 3. CNVs in CEBPD and KLHL9 predict poor prognosis and elevated CEBP protein levels independently of molecular classification in independent patient cohorts



	KLHL9$^{-/- \text{ or } -/0}$	Total Samples
Poor Prognosis	21	40
Good Prognosis	4	23
p-value	5.68e-3	

	p-value	OR
KLHL9$^{-/-}$ and CEBP positive	0.0283	12.25 (N=20)

Figure 4. Exogenous KLHL9 affects expression of predicted CEBP targets and suppresses CEBPB/D proteins

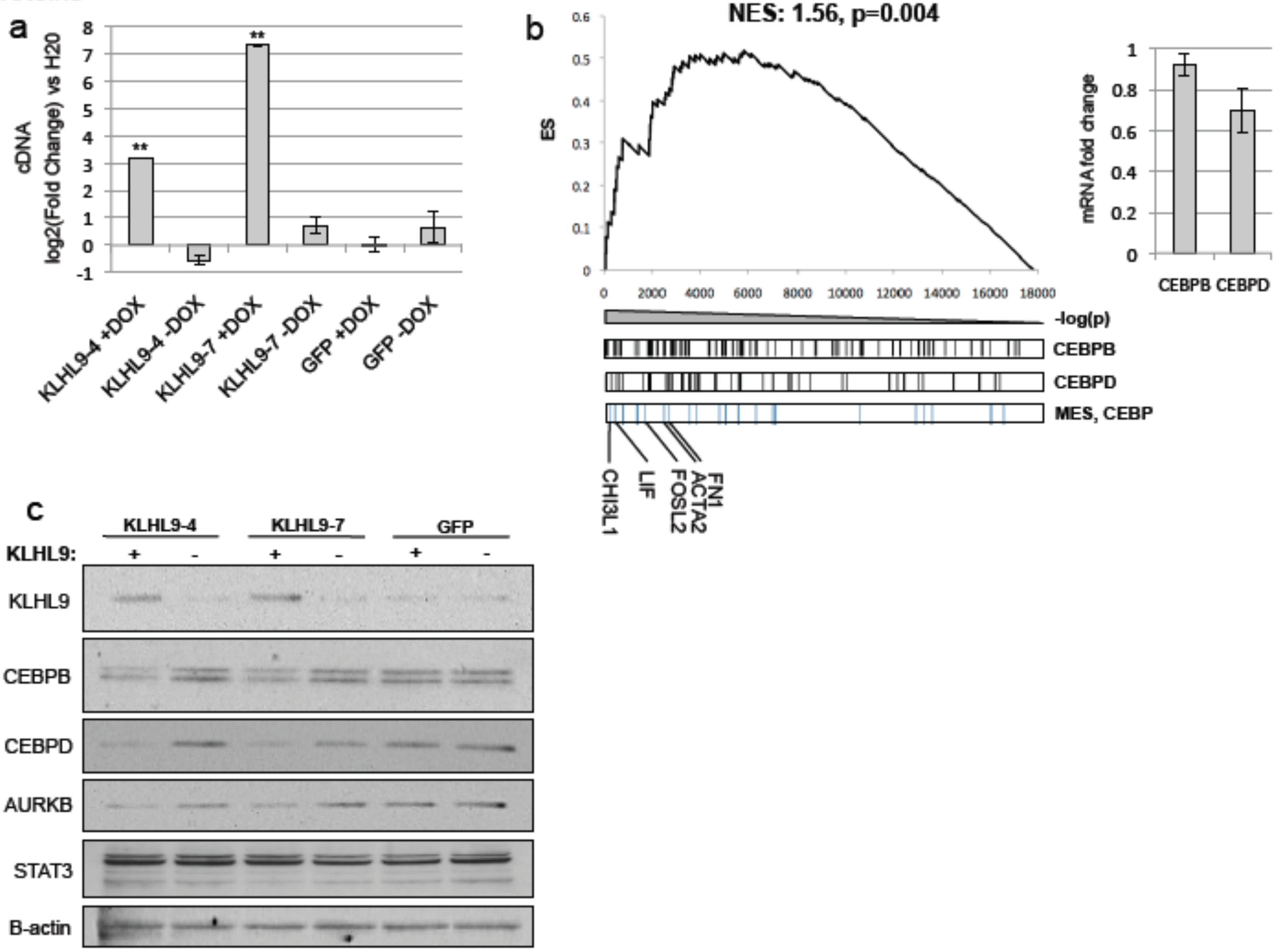
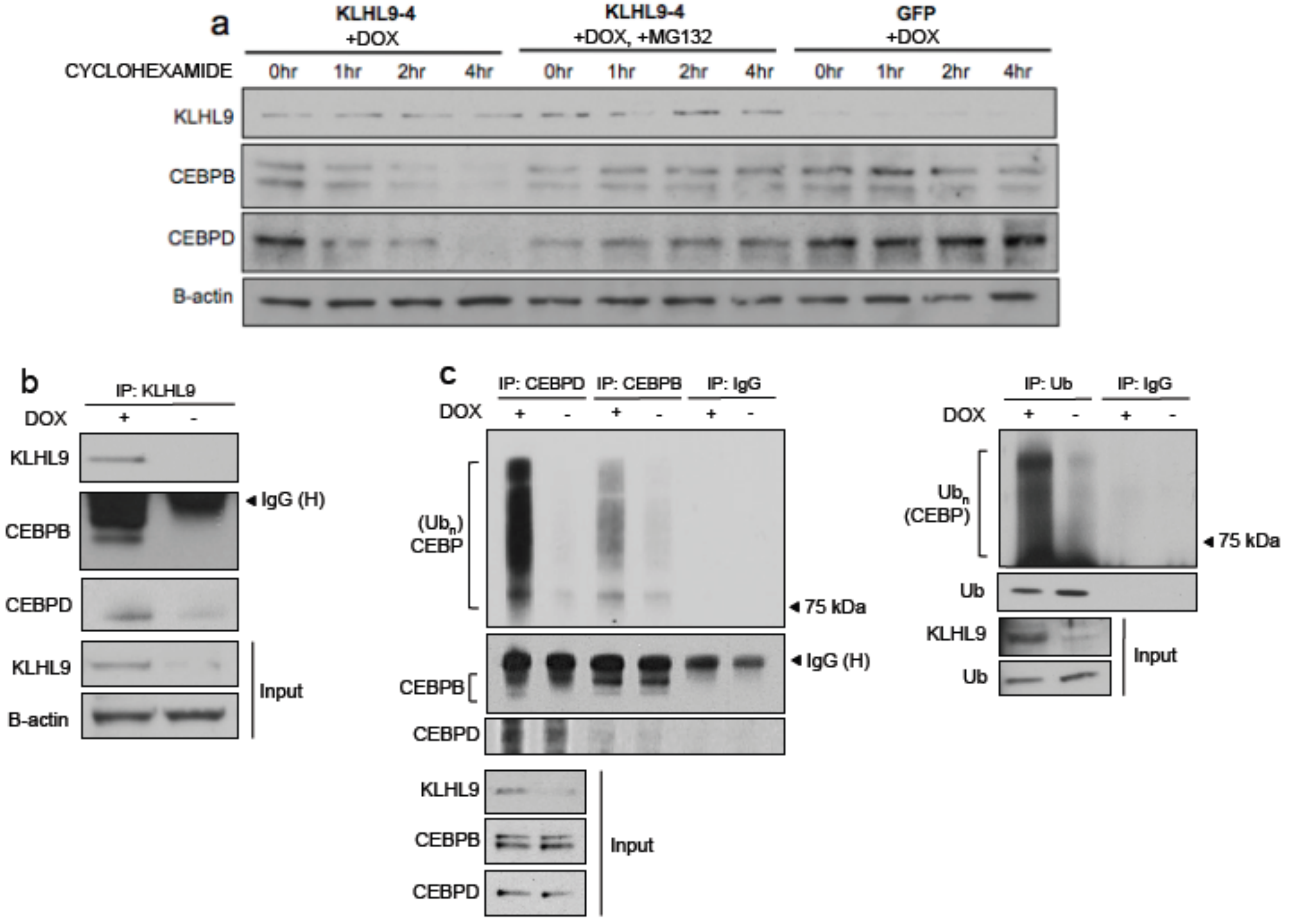


Figure 5. KLHL9 expression induces proteasome-mediated degradation of CEBPB/D via ubiquitylation



*Ub IPs Performed in presence of MG132

Figure 6. Deletion of the KLHL9 BTB domain suppresses CEBP ubiquitylation and degradation

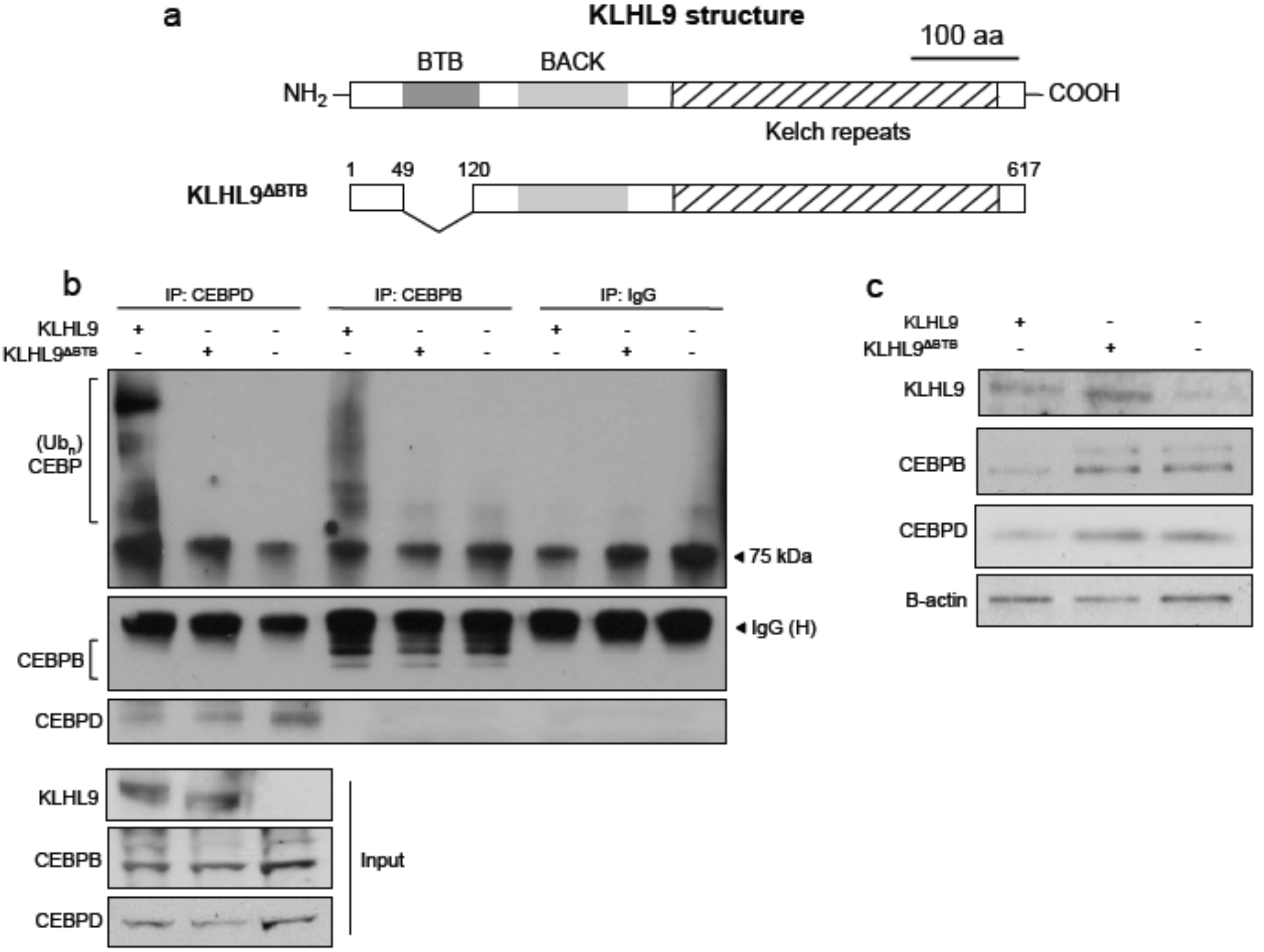
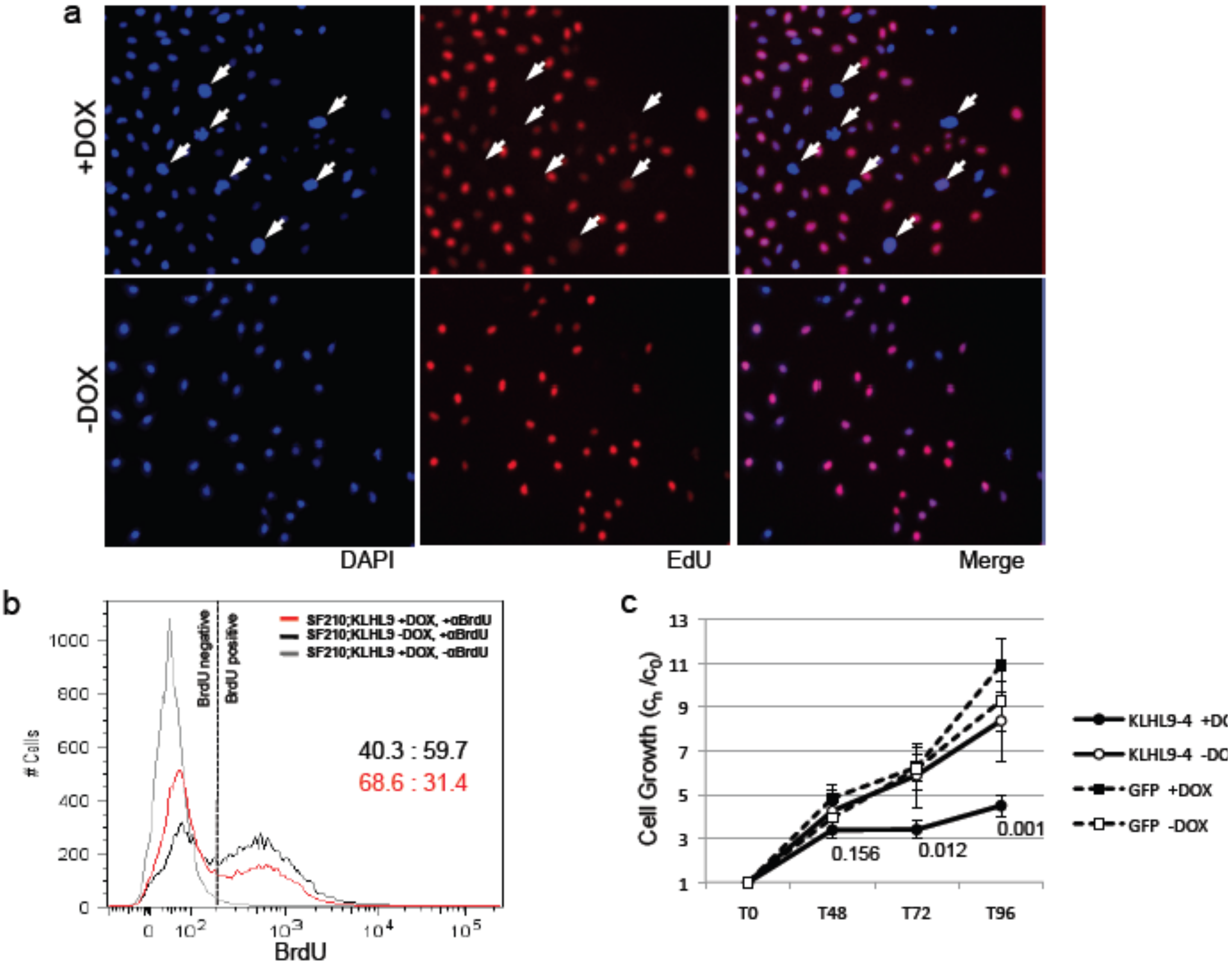


Figure 7. KLHL9 expression induces a cell morphology phenotype and arrested cellular proliferation



APPEND06 – Manuscript Supplemental Figures and Figure Legends

SuppFigure1.

(1a) Five hierarchically self-regulating transcription factors (*C/EBPβ/δ*, *STAT3*, *FOSL2*, *BHLHB2*, and *RUNX1*) were identified as the master regulators of aggressive, poor prognosis GBM by the ARACNE algorithm. (1b) TCGA samples classified into "good" and "poor" prognosis using the activity of these five transcription factors in TCGA data shows a statistically significant separation in survival curves (p<0.05).

SuppFigure2.

Functional copy number variation genes (f-CNVGs) are defined as copy number variations in gene loci where differential expression of the gene is detected in correlation with the observed CNV. (2a) Nonfunctional CNVs will show no differential expression of host genes between amplified (red) and diploid (blue) samples. (2b) Functional CNVs show a measurable differential expression between amplified (red) and normal (samples). (2c) Measuring statistical dependencies between CNVs and gene expression via traditional statistical tests yields little statistical power, producing only 51 functional CNVs (U-test). In constrast, information theoretic approaches (Mutual Information) more than double the amount (124 total) of detectable dependencies at the tested significance threshold, and include all those detectable by traditional statistical methods.

SuppFigure3.

The null distribution for measuring mutual information is empirically-determined by randomly pairing 10,000 CNV and gene expression vectors and measuring the mutual information between them, representing the mutual information obtained under the null hypothesis of random pairing (no correlation). A regression function is then fit to this distribution that is used to estimate the p-values of mutual information measured in the analysis that is no longer bound by sampling size.

SuppFigure4.

Eight huGBM-derived cell lines assayed for deletions in the *KLHL9* or *CDKN2A* CDS locus. y-axis represents normalized CTs; positive CTs indicate less genomic DNA present, and each CT represents a fold change of 2 relative to the control GAPDH levels, set at 0 CT. CT values are reported as mean ±SEM. Red line indicates threshold for statistically-significant evidence for genomic deletions.

SuppFigure5.

(5a) Probe mapping of the Affymetrix SNP arrays reveals lack of coverage of the *C/EBPδ* gene locus (red hashes denote probes across different samples) and sparse coverage of the genomic region compared to CGH arrays. (5b) Probe-wise association mapping of the chromosome 8 locus using Affy SNP arrays does not reveal significant association of the gene locus with poor prognosis as reported by the CGH arrays. (5c) Using a sliding window integration, significant association of the *C/EBPδ* locus is detected using Affymetrix arrays, though not as significantly as reported by CGH arrays. X-axes represent chromosome position in megabases, y-axis depicts odds ratios of association.

SuppFigure6.

Probes available in Affymetrix SNP arrays do not recapitulate the CNV;gene expression correlations observed when integrating the CGH arrays with gene expression profiles. The three maximally associated probes in the region, labeled as peaks -1,0, and +1, were each individually tested for correlation to the mRNA expression of *C/EBPδ*; those probes were unable to predict mRNA levels with any significance. Integration of probes using a sliding window reveals a correlation with *C/EBPδ* mRNA, but still with less information than the probes provided by the CGH arrays.

SuppFigure7.

Segmentation maps pre-GISTIC processing of the CEBPD and KLHL9 genomic loci were used to deconvolve the mutational topography for genetical-genomics analysis. GISTIC processing removes the CEBPD locus entirely as a false signal and renders deletions at the KLHL9 and CDKN2A loci as mutually inclusive and equivalent across all samples bearing deletions at CDKN2A.

SuppFigure8.

Association scores presented as $–\log_{10}(p)$ on the y-axes. X-axes plot the location of each locus by megabase along the chromosomes. Gene names in red indicate genes harboring functional CNVs. Bolded gene names indicate genes expressed in the TCGA GBM tumors, and plain text indicate genes that are not expressed in TCGA GBM. Gene locations are indicated as diamonds. (7a) Probe-wise association mapping across the locus on chromosome 8 harboring *C/EBPd* reveals a focal amplification that associates with the poor-prognosis subtype (red line). When samples bearing *C/EBPδ* amplifications are removed, the association across the region is also removed (blue line). Gene locations are indicated (diamonds). (7b) Probe-wise association mapping across the locus on chromosome 9 harboring *KLHL9* reveals a deletion that associates with the MES subtype versus PN/PRO (red line) spanning ~21MB-22MB, including *KLHL9* and *CDKN2A*. When samples bearing *C/EBPd* amplifications are removed, the association across the region is also removed (blue line). (7c) At the *KLHL9* locus, removing all samples carrying a deletion of the *CDKN2A* locus but diploid at the *KLHL9* locus enhances the association of the *KLHL9* locus to the poor-prognosis phenotype (blue line) compared to the complete set (red line).

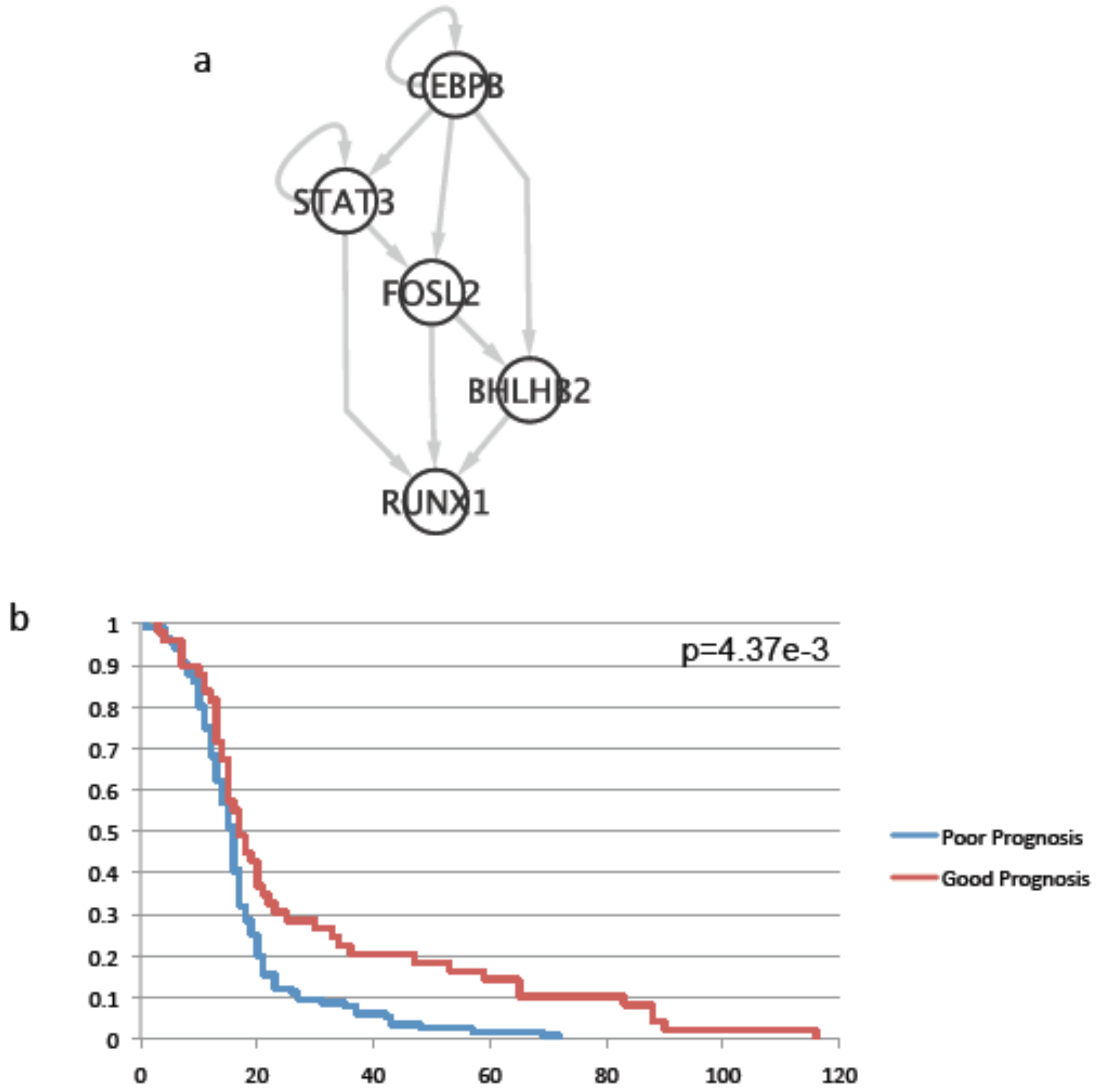
SuppFigure9.

Transcriptional cross-talk experiments using siRNAs reveals regulatory interactions between the MES master regulators CEBPD and STAT3. STAT3 silencing induces a reduction in transcription of CEBPD, placing CEBPD and CEBPB as the most downstream regulators of the three

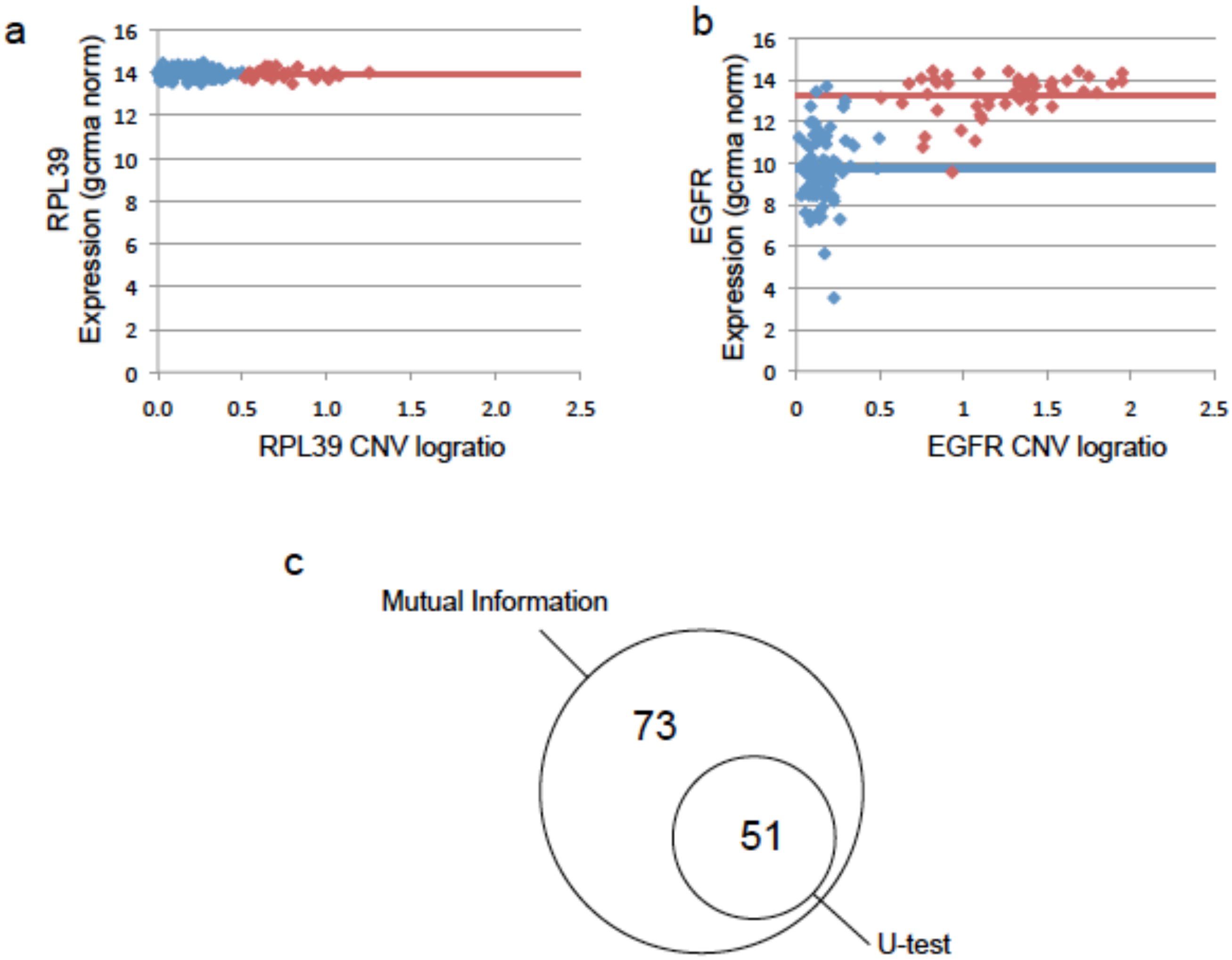
SuppFigure10.

Gene lists of genes bearing CNVs functionally correlating with MGES behavior (f-CNVGs) are listed by their methods of identification: genetical genomics and MINDy modulator analysis. In addition, a list of commonly-used “classical” CNV markers of GBM oncogenesis are listed, and the genes whose CNVs were identified as f-CNVs by our genetical genomic analysis are represented in **bold**. All classical CNVs that were not identified appeared in fewer than 10 samples in the total TCGA set (<5% of samples).

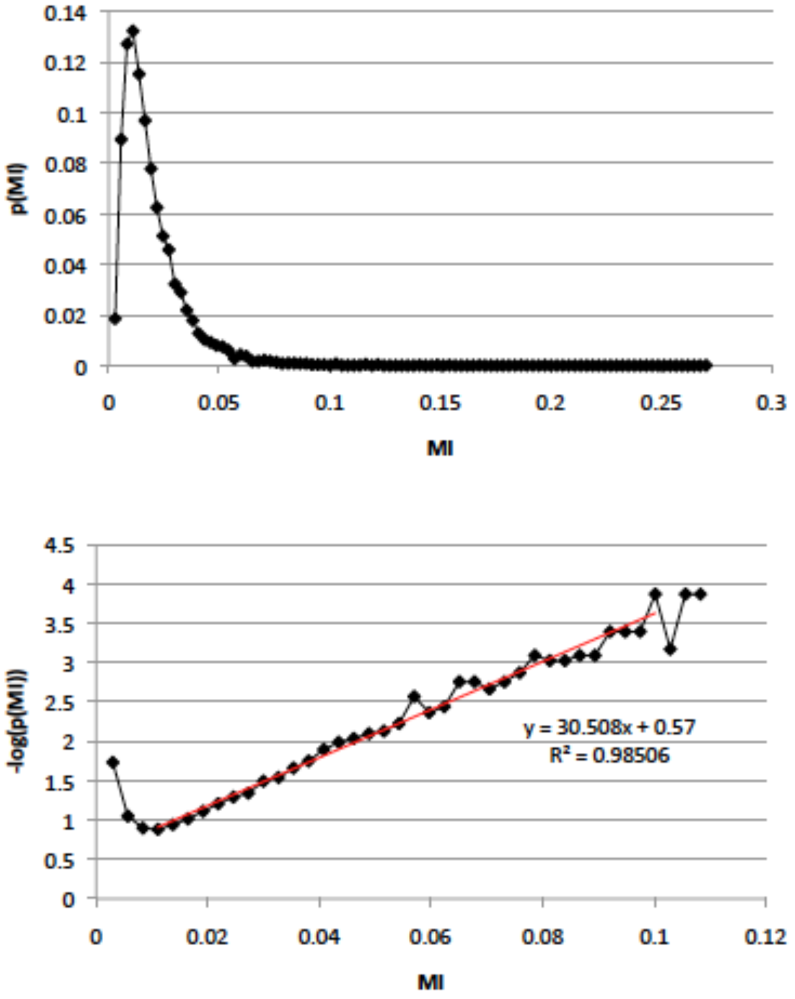
Supplemental Figure 1. Defining the poor prognosis subtype by MGES expression in the TCGA dataset



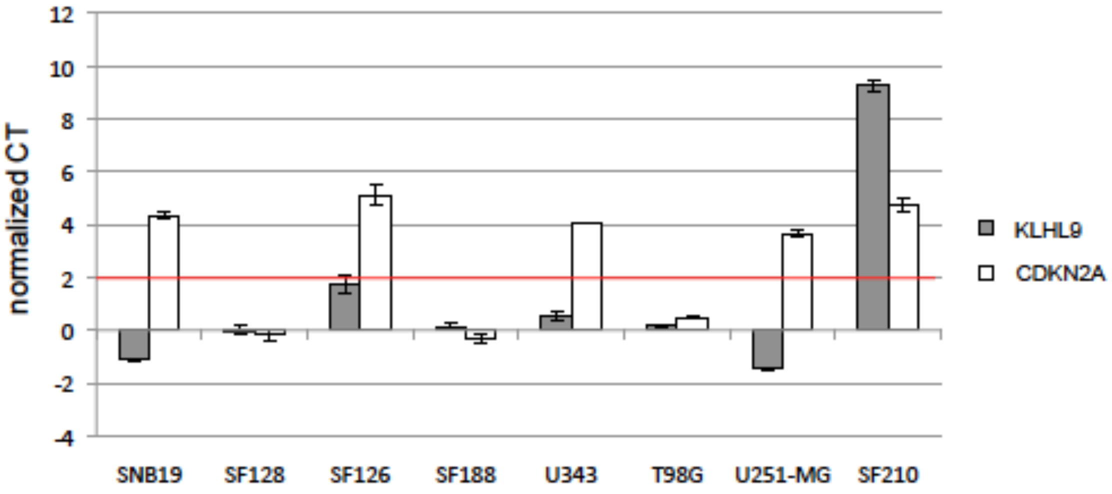
Supplemental Figure 2. f-CNVGs are defined based on an integration of CNV and gene expression data using mutual information



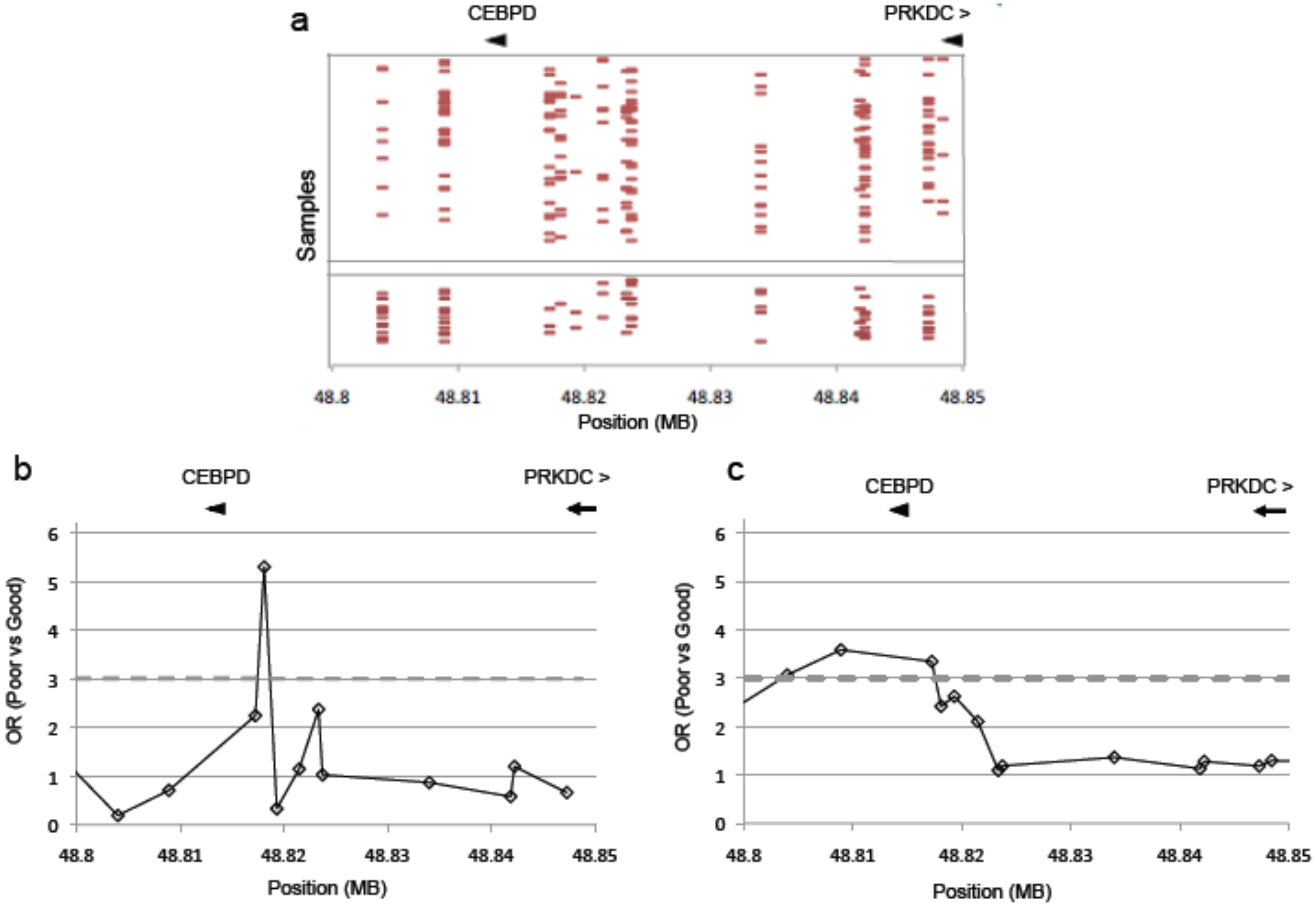
Supplemental Figure 3. The null distribution of MI[CNV;mRNA] in the TCGA dataset, estimating p-values



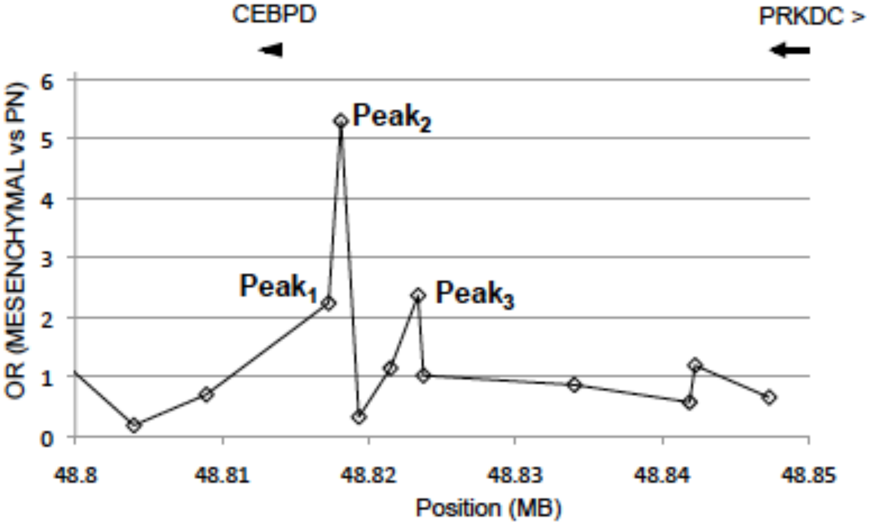
Supplemental Figure 4. The huGBM-derived SF-210 cell line bears a homozygous deletion of KLHL9 and of p16



Supplemental Figure 5. Affymetrix SNP Array probe coverage is sparse at key genes and requires integrative methods to detect association to the MGES

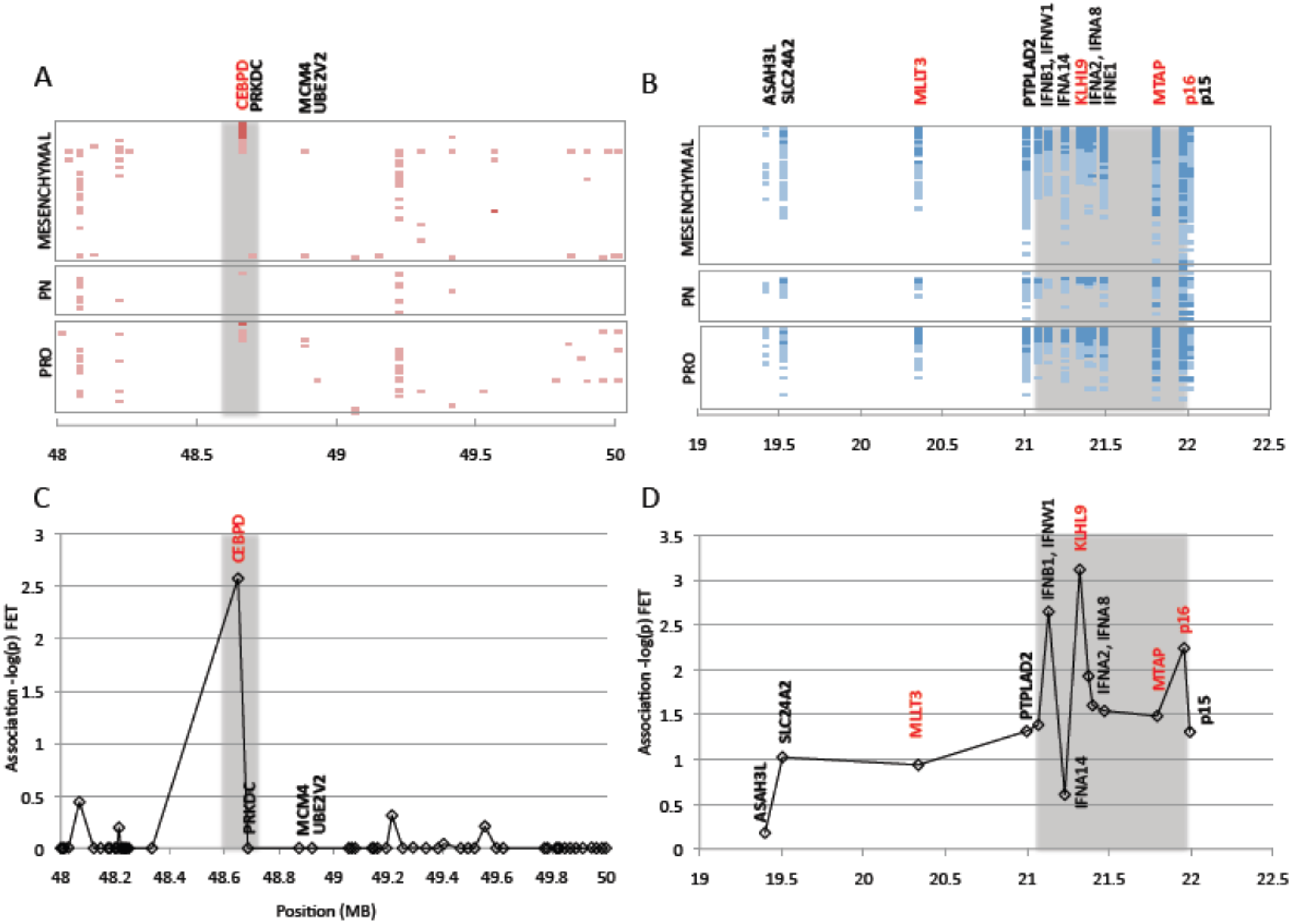


Supplemental Figure 6. Agilent array data provides better correlation with expression data than individual segments from Affymetrix SNP array data

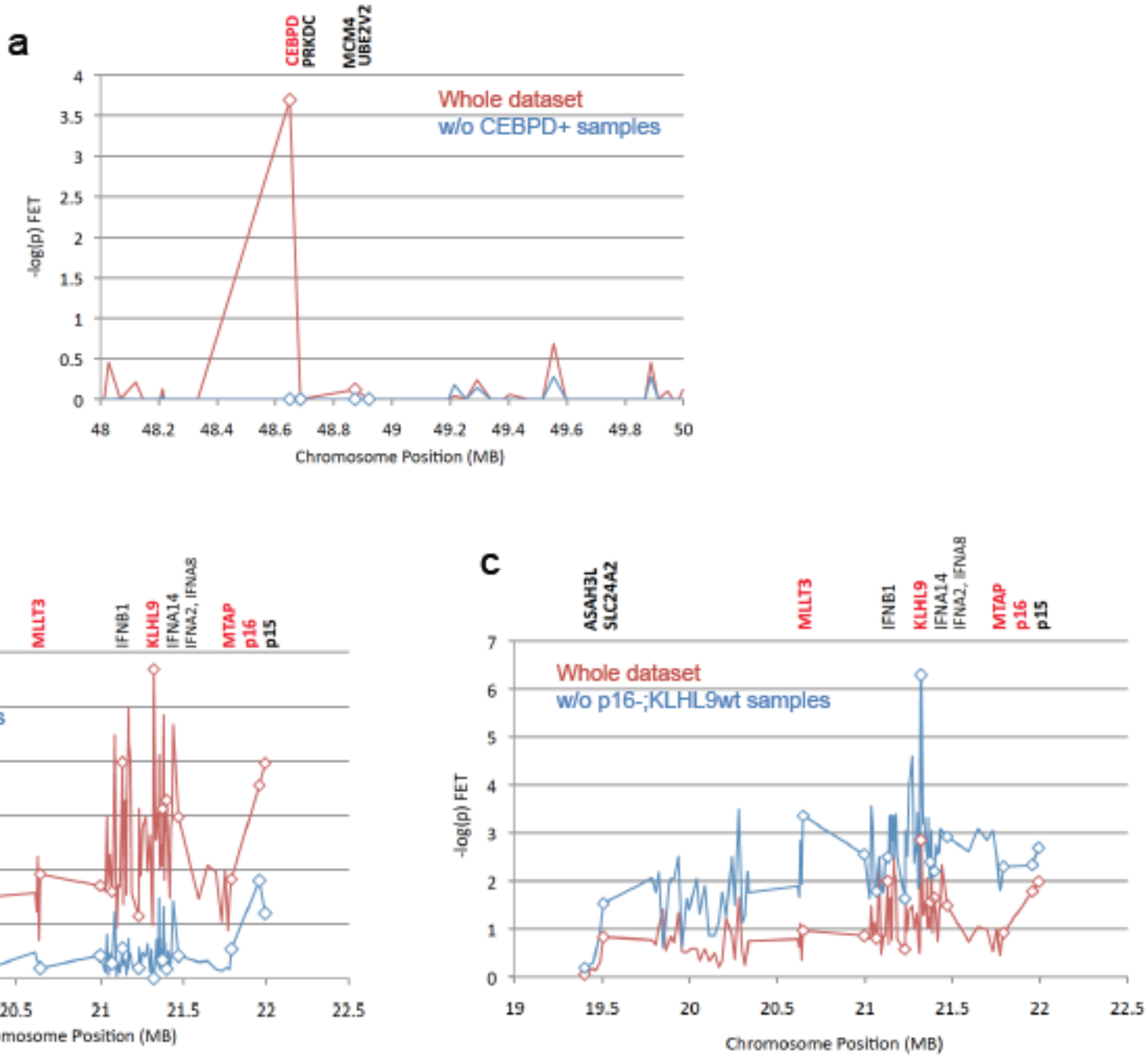


CEBPD CNV Probe	Correlation with CEBPD Expression
Affy: Peak$_1$	0.0904
Affy: Peak$_2$	0.0931
Affy: Peak$_3$	-0.0172
Affy: Integration	0.148
Agilent: CEBPD locus	0.258

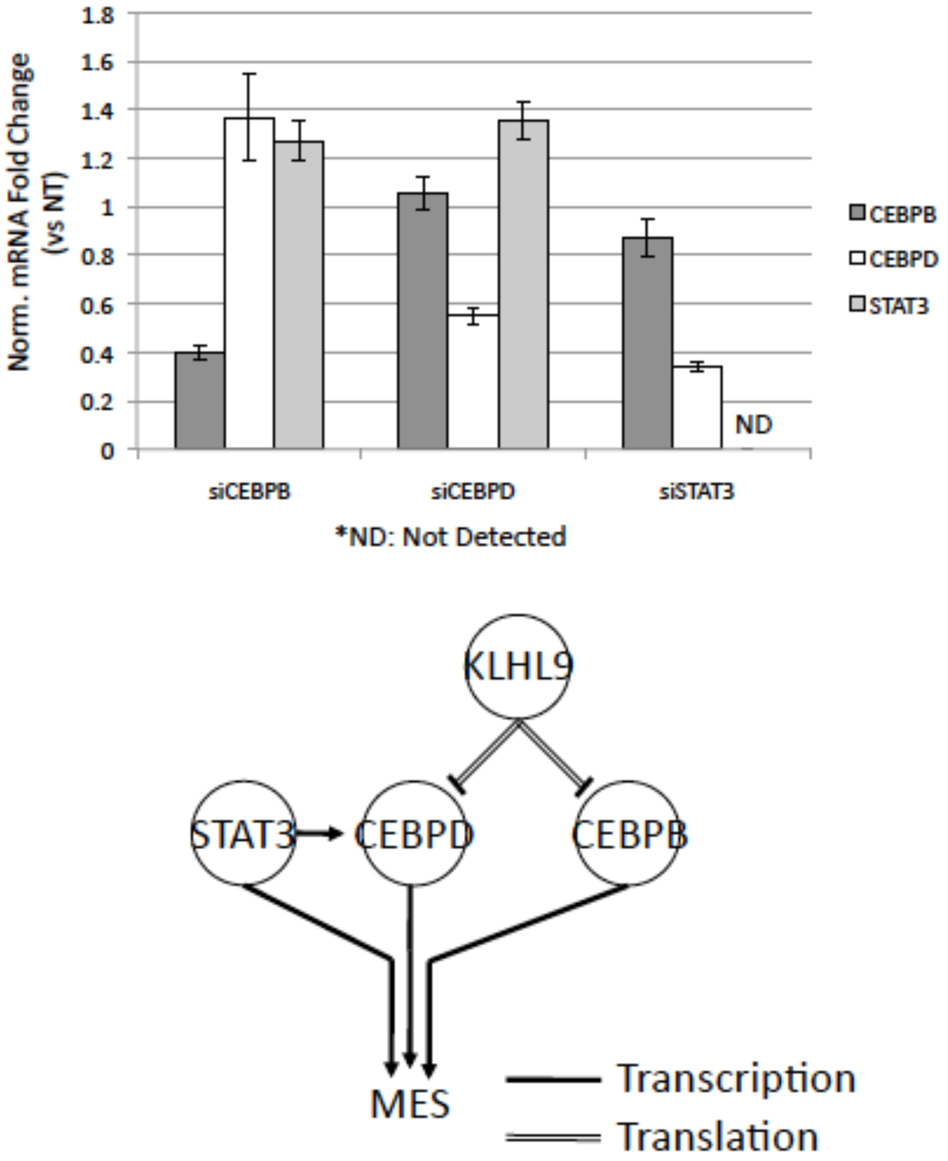
Supplemental Figure 7. Segmentation mapping reveals focal CNVs associated with the MES phenotype



Supplemental Figure 8. Amplification of CEBPD and deletions of KLHL9 account for the association of their chromosomal region to the MGES



Supplemental Figure 9. siRNA-mediated crosstalk experiments reveal CEBPB proteins are the most downstream MES master regulators



Supplemental Figure 10. Genes identified as f-CNVGs include a majority of previously-defined oncogenesis CNVs

MGES fCNVGs identified by BOTH Genetical Genomics and MINDy

KLHL9
CEBPD
BLNK
PFKP
IRGQ
CCDC6
NPTX2
KIF5B
NRBF2
ECHDC3
RPL28
SEC61G
NMT2
MDH2
CDKN2A
KCNH2
BCAT2
MTAP
KCTD15
CDK5
PHYH
SNX13
OPTN
PDLIM7
VPS26A
UPP1
CCT6A
CCDC106
AP3D1
AKR1C2
WIPI2
CAMK2G
DDX50
EGFR
UBE2D1
MSRB2
ZNF132
PSPH
SMARCD3
KDELR2
SNAPC2

MGES fCNVGs identified by EITHER Genetical Genomics or MINDy

ABI1
CLOCK
DDIT3
CDK4
MLC1
IFT74
NFKBIA
CARHSP1
CYorf15B
CLEC11A
LANCL2
KIAA1797
ZNF134
USP2
KIN
PAPD1
DUS4L
IL15RA
TAX1BP1
EIF4G2
FIP1L1
SLC25A22
PDAP1
EIF1AX
STS
MINPP1
MEOX2
GLUD1
OBFC1
EIF1AY
MRPS17
C9orf82
ANXA7
DCTN2
PANK4
OS9
GNA12
ZNF14
EMP3
C10orf72
GAS7
TSFM
ST7
TFR2
CAMK2N1
MET
DTX3
SOX13
FRMD4A
SAR1A
ITGB1
MAX
RBBP5
MLLT3
FAM53B
ELAVL2
AVIL
CUL2
PRKCQ
PIK3C2B
THNSL1
NRP1
CHIC2
ZNF586
ITIH2
PPP3CB
PHKG1
PLEKHA6
DDX3X
RPS24
HOXA10
DIO2
B4GALNT1
NDUFA3
RHOBTB1
GBAS
KIT
UBAP1
SLC35E3
SLC26A10
CUGBP2
SLC25A28
GHITM
STAM
AKT1
DDX56
ECD
PDGFRA
CUTC
TSPAN31
GLI1
TEK
ITGA8
GSTT1
ZCWPW1
LY6H
HLA-DQA1
SPAG6
PIN4
HP1BP3
SERINC1
PRKG1
PRKY
WDR37
CYP27B1
C7orf26
IL6
KIF5A
DDX21
TMEM8
MBD6
RPP30
MDM2
SCRG1
CH25H
CPM
ANXA11
KIAA1128
GNAI1
HDAC9
INHBE
ZNF671
GDI2
TOPORS
MDM4
MARS
GTPBP4
CTDSP2
RSU1
CAPZA2
NUP107
DCLRE1C
ATP5C1
ARL3
VGF
USP9X
ETNK2
METTL1
CBARA1
PRC1
GFAP

“Classical” oncogenesis CNVs

EGFR	Amplified loci
CDK4	
PDGFRA	
MDM2	
MDM4	
MET	
AKT3	
MYCN	
CCND2	
PIK3CA	
CDK6	
CDKN2A	Deleted loci
CDKN2B	
CDKN2C	
RB1	
PTEN	
PARK2	
NF1	

*bold indicates identification as fCNVG