Cell

An Integrated Approach to Uncover Drivers of Cancer

Uri David Akavia,[1,2,5] Oren Litvin,[1,2,5] Jessica Kim,[3,4] Felix Sanchez-Garcia,[1] Dylan Kotliar,[1] Helen C. Causton,[1] Panisa Pochanard,[3,4] Eyal Mozes,[1] Levi A. Garraway,[3,4] and Dana Pe'er[1,2,*]

[1]Department of Biological Sciences, Columbia University, 1212 Amsterdam Avenue, New York, NY 10027, USA
[2]Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, New York, NY 10032, USA
[3]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA
[4]Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA
[5]These authors contributed equally to this work
*Correspondence: dpeer@biology.columbia.edu
DOI 10.1016/j.cell.2010.11.013

SUMMARY

Systematic characterization of cancer genomes has revealed a staggering number of diverse aberrations that differ among individuals, such that the functional importance and physiological impact of most tumor genetic alterations remain poorly defined. We developed a computational framework that integrates chromosomal copy number and gene expression data for detecting aberrations that promote cancer progression. We demonstrate the utility of this framework using a melanoma data set. Our analysis correctly identified known drivers of melanoma and predicted multiple tumor dependencies. Two dependencies, *TBC1D16* and *RAB27A*, confirmed empirically, suggest that abnormal regulation of protein trafficking contributes to proliferation in melanoma. Together, these results demonstrate the ability of integrative Bayesian approaches to identify candidate drivers with biological, and possibly therapeutic, importance in cancer.

INTRODUCTION

Large-scale initiatives to map chromosomal aberrations, mutations, and gene expression have revealed a highly complex assortment of genetic and transcriptional changes within individual tumors. For example, copy number aberrations (CNAs) occur frequently in cancer due to genomic instability. Genomic data have been collected for thousands of tumors at high resolution using array comparative genomic hybridization (aCGH) (Pinkel et al., 1998), high-density single-nucleotide polymorphism (SNP) microarrays (Beroukhim et al., 2010; Lin et al., 2008), and massively parallel sequencing (Pleasance et al., 2010). Although multiple new genes have been implicated in cancer through sequencing and CNA analysis (Garraway et al., 2005), these studies have also revealed enormous diversity in genomic aberrations in tumors among individuals. Each tumor is unique and typically harbors a large number of genetic lesions, of which only a few drive proliferation and metastasis. Thus, identifying driver mutations (genetic changes that promote cancer progression) and distinguishing them from passengers (those with no selective advantage) has emerged as a major challenge in the genomic characterization of cancer.

The most widely used approaches are based on the frequency that an aberration occurs: if a mutation provides a fitness advantage in a given tumor type, its persistence will be favored, and it is likely to be found in multiple tumors. For example, GISTIC identifies regions of the genome that are aberrant more often than would be expected by chance and has been used to analyze a number of cancers (Beroukhim et al., 2007, 2009; Lin et al., 2008). However, there are limitations to analytical approaches based on CNA data alone: CNA regions are typically large and contain many genes, most of which are passengers that are indistinguishable in copy number from the drivers. CNA data have statistical power to detect only the most frequently recurring drivers above the large number of unrelated chromosomal aberrations that are typical in cancer. Finally, these approaches rarely elucidate the functional importance or physiological impact of the genetic alteration on the tumor. These limitations highlight the need for new approaches that can integrate additional data to identify drivers of cancer. Gene expression is readily available for many tumors, but how best to combine it with information on CNA is not obvious.

We postulate that driver mutations coincide with a ''genomic footprint'' in the form of a gene expression signature. We developed an algorithm that integrates chromosomal copy number and gene expression data to find these signatures and identify likely driver genes located in regions that are amplified or deleted in tumors. Each potential driver gene is altered in some, but not all, tumors and, when altered, is considered likely to play a contributing role in tumorigenesis. Unique to our approach, each driver is associated with a gene module, which is assumed to be altered by the driver. We sometimes gain insight into the likely role of a candidate driver based on the annotation of the genes in the associated module. We demonstrate the utility of our method using a data set (Lin et al., 2008) that includes paired measurements of gene expression and copy number from 62 melanoma samples. Our analysis correctly identified known drivers of melanoma and connected them to many of their

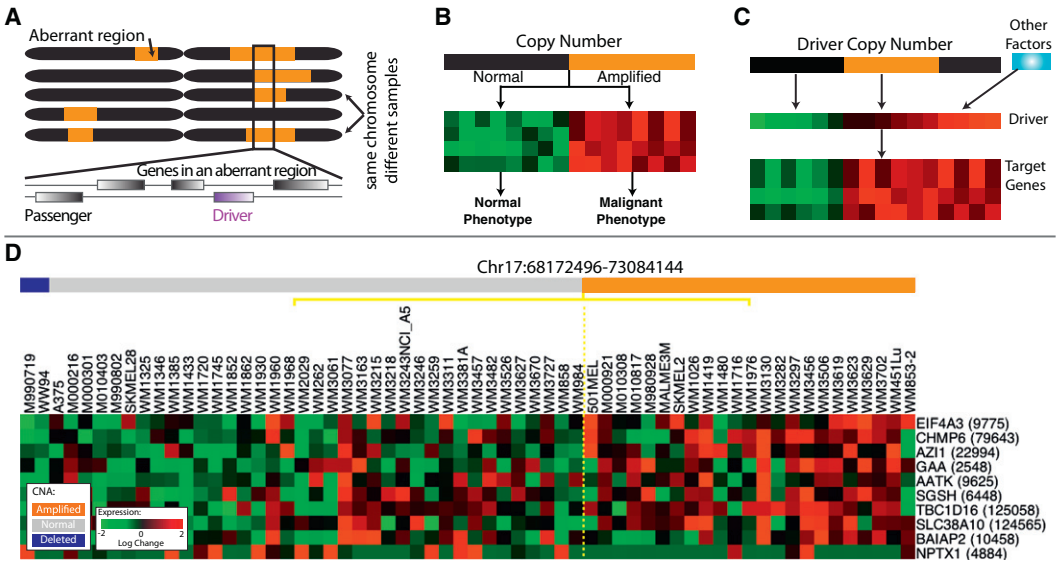


Figure 1. Modeling Assumptions

For all heat maps, each row represents a gene and each column represents a tumor sample.

(A) The same chromosome in different tumors; orange represents amplified regions. The box shows regions amplified in multiple tumors.

(B) An idealized signature in which the target genes are upregulated (red) when the DNA encoding the driver is amplified (orange).

(C) A driver may be overexpressed due to amplification of the DNA encoding it or due to the action of other factors. The target genes correlate with driver gene expression (middle row), rather than driver copy number (top row).

(D) Data representing amplified region on chromosome 17. Heat maps of expression for 10 of 24 genes that passed initial expression filtering (Extended Experimental Procedures).

Samples are ordered according to amplification status of the region (orange, amplified; blue, deleted). These genes are identical in their amplification status, and though gene expression is correlated with amplification status to some degree, the expression of each gene is unique. It is these differences that facilitate the identification of the driver. See also Extended Experimental Procedures, Figure S1, and Table S1.

targets and biological functions. In addition, it predicted novel melanoma tumor dependencies, two of which, *TBC1D16* and *RAB27A*, were confirmed experimentally. Both of these genes are involved in the regulation of vesicular trafficking, which highlights this process as important for proliferation in melanoma.

RESULTS

The Genomic Signature of a Driver

We define a ''driver mutation'' to be a genetic alteration that provides the tumor cell with a growth advantage during carcinogenesis or tumor progression (Stratton et al., 2009). We reasoned that driver mutations might leave a genomic ''footprint'' that can assist in distinguishing between driver and passenger mutations based on the following assumptions: (1) a driver mutation should occur in multiple tumors more often than would be expected by chance (Figure 1A); (2) a driver mutation may be associated (correlated) with the expression of a group of genes that form a ''module'' (Figure 1B); (3) copy number aberrations often influence the expression of genes in the module via changes in expression of the driver (Figure 1C).

Driver mutations are frequently associated with the abnormal regulation of processes such as proliferation, differentiation, motility, and invasion. Given that many cancer phenotypes are reflected in coordinated differences in the expression of multiple genes (a module) (Golub et al., 1999; Segal et al., 2004), a driver mutation might be associated with a characteristic gene expression signature or other phenotypic output representing a group of genes whose expression is *modulated* by the driver. In addition, CNAs do not typically alter the coding sequence of the driver and so are expected to influence cellular phenotype via changes in the driver's expression. In consequence, changes in expression of the driver are important, so approaches that measure association between the expression of a candidate driver (as opposed to its copy number) and that of the genes in the corresponding module are likely to promote the identification of drivers.

Gene expression is particularly useful for identifying candidate drivers within large amplified or deleted regions of a chromosome: whereas genes located in a region of genomic copy gain/loss are indistinguishable in copy number, expression permits the ranking of genes based on how well they correspond with the phenotype (Figure 1D). CNA data aids in determining the direction of influence, which cannot be derived based on correlation in gene expression alone (Figure 3A). This permits an unbiased approach for identifying candidate drivers from any functional family, beyond transcription factors or signaling proteins.

A Bayesian Network-Based Algorithm to Identify Driver Genes

We developed a computational algorithm, copy number and expression in cancer (CONEXIC), that integrates matched copy

Gene Symbol	Pathway	Band	Genes in Region	Validation p-value
MITF	Melanoma	3p14.2-p14.1	1	$<10^{-6}$
TBC1D16	Vesicular Trafficking	17q25.3	24	$<10^{-6}$
ZFP106	Insulin/Ras	15q15.1	7	$<10^{-6}$
DIXDC1	Wnt/JNK/PI3K	11q23.1	17	0.0001
OIP5	Cell Cycle	15q15.1	13	$<10^{-6}$
TTBK2		15q15.2	7	0.0383
TRAF3	NFkappaB/JNK	14q32.32	19	0.0121
RAB27A	Vesicular Trafficking	15q15-q21.1	33	$<10^{-6}$
C12orf35		12p11.21	45	$<10^{-6}$
WBP2		17q25	92	0.0275
MOCS3		20q13.13	16	$<10^{-6}$
NDUFB2		7q34	10	$<10^{-6}$
ST6GALNAC2		17q25.1	92	$<10^{-6}$
GRB2	EGFR/Ras	17q24-q25	92	0.1373
ECM1		1q21	55	0.0083
KCNG1		20q13	16	0.202
DPM1		20q13.13	16	0.097
PFKP	Metabolism	10p15.3-p15.2	3	0.0801
KLF6	Cell cycle, c-JUN (JNK)	10p15	3	$<10^{-6}$
TIMM8B	Mitochondria	11q23.1-q23.2	17	0.7622
PI4KB		1q21	55	0.0003
PSMB4		1q21	55	0.0005
VPS72		1q21	55	$<10^{-6}$
TARS2		1q21.3	55	0.0001
MNS1		15q21.3	33	0.0908
TDRD3	RNA processing	13q21.2	203	$<10^{-6}$
CCNB2	Cell Cycle	15q22.2	33	$<10^{-6}$
EIF5	Cell Cycle	14q32.32	19	0.1096
RAB7A	Vesicular Trafficking	3q21.3	16	$<10^{-6}$
PIK3CB	PI3K signaling	3q22.3	15	$<10^{-6}$

Figure 2. The Highest-Scoring Modulators Identified by CONEXIC

Gene names are color coded based on the role of the gene in cancer. Ten genes have been previously identified as oncogenes or tumor suppressors (blue); of these, two in melanoma (brown). Column 3 represents chromosomal location, orange represents amplification, and blue represents deletion. These genes were identified within regions containing multiple genes, and the number of genes in each aberrant region is listed in column 4. Column 5 lists the p value for modulator validation in independent data (for a full list, see Table S2 and Figure S3C). p values are shown for the Johansson data set unless the modulator was missing from this data set, and then p value from the Hoek data set is shown. See also Extended Experimental Procedures, Table S2, and Figure S3.

number (amplifications and deletions) and gene expression data from tumor samples to identify driver mutations and the processes that they influence. CONEXIC is inspired by Module Networks (Segal et al., 2003) but has been augmented by a number of critical modifications that make it suitable for identifying drivers (see Extended Experimental Procedures available online). CONEXIC uses a score-guided search to identify the combination of modulators that best explains the behavior of a gene expression module across tumor samples and searches for those with the highest score within the amplified or deleted regions (Extended Experimental Procedures and Figure S1).

The resulting output is a ranked list of high-scoring modulators that both correlate with differences in gene expression modules across samples and are located in amplified or deleted regions in a significant number of these samples. The fact that the modulators are amplified or deleted indicates that they are likely to control the expression of the genes in the corresponding modules (see Figure 3). Because the modulators are amplified or deleted in a significant number of tumors, it is reasonable to assume that expression of the modulator (altered by copy number) contributes a fitness advantage to the tumor. Therefore, the modulators likely include genes whose alteration provides a fitness advantage to the tumor.

Identifying Candidate Driver Genes in Melanoma

We applied the CONEXIC algorithm to paired gene expression and CNA data from 62 cultured (long- and short-term) melanomas (Lin et al., 2008). A list of candidate drivers was generated using copy number data available for 101 melanoma samples by applying a modified version (Sanchez-Garcia et al., 2010) of GISTIC (Beroukhim et al., 2007) (see Table S1). Next, we integrated copy number and gene expression data (available for 62 tumors) to identify the most likely drivers (Extended Experimental Procedures). Statistical power is gained by integrating all data and by combining statistical tests on thousands of genes to support the selected modulators. This resulted in the identification of 64 modulators that explain the behavior of 7869 genes. We consider the top 30 scoring modulators, presented in Figure 2, as likely drivers (see Table S2 for the complete list).

Many Modulators Are Involved in Pathways Related to Melanoma

The top 30 modulators (likely drivers) include 10 known oncogenes and tumor suppressors (Figure 2). In many cases, CONEXIC chose the cancer-related gene out of a large aberrant region containing many genes. For example, *DIXDC1*, a gene known to be involved in the induction of colon cancer (Wang et al., 2009b), was selected among 17 genes in an aberrant region (Figure S2). *CCNB2*, a cell-cycle regulator, was selected from a large amplified region containing 33 genes. The modulators span diverse functional classes, including signal transducers (*TRAF3*), transcription factors (*KLF6*), translation factors (*EIF5*), and genes involved in vesicular trafficking (*RAB27A*).

Performing a comprehensive literature search for all genes is tedious and time consuming, so we developed an automated procedure, literature vector analysis (LitVAn), which searches for overrepresented terms in papers associated with genes in a gene set. LitVAn uses a manually curated database (NCBI Gene) to connect genes with terms from the complete text of more than 70,000 published scientific articles (Extended Experimental Procedures). LitVAN found a number of overrepresented terms (Figure S3E) among the top 30 modulators, including "PI3K" and "MAPK," which are known to be activated in melanoma; "cyclin," representing proliferation, which is common in

all cancers; and "RAB." Rabs regulate vesicular trafficking, a process not previously implicated in melanoma.

The Association between a Modulator and the Genes in a Module

Beyond generating a list of likely drivers (modulators), the CONEXIC output includes groups of genes that are associated with each modulator (modules). We tested how reproducible the modulators and their associated modules are using gene expression data from two other melanoma cohorts with 45 (Hoek et al., 2006) and 63 (Johansson et al., 2007) samples (see Extended Experimental Procedures and Figure S3). We found that 51 of 64 (80%) of the selected modulators are conserved across data sets in a statistically significant manner. Modules (statistically associated genes) are likely enriched with genes whose expression is biologically affected by the modulator (Figure 3). In consequence, the processes and pathways represented by genes in a module can help us to gain insight into how an aberration in the modulator might alter the cellular physiology and contribute to the malignant phenotype.

Annotation of data-derived sets of genes is typically carried out based on gene set enrichment using Gene Ontology (GO) annotation. Although this approach is useful, there are modules for which GO annotation does not capture the known biology. For example, the "TNF module" is enriched with the GO terms "developmental process" and "cell differentiation" (q value = 0.0014 and 0.004, respectively). We used LitVAn to carry out a systematic literature search and found 11 of 20 genes in the module related to the TNF pathway, inflammation, or both (Figure 3C and Table S3), although only two of these genes were annotated for these processes in GO. *TRAF3*, the modulator chosen by CONEXIC, is known to regulate the NF-κB pathway (Vallabhapurapu et al., 2008), a major downstream target of TNF. Although *TRAF3* has not been previously implicated in melanoma, the importance of the NF-κB pathway in melanoma is well supported (Chin et al., 2006).

A Known Driver, MITF, Is Correctly Associated with Target Genes

CONEXIC identified microphthalmia-associated transcription factor (*MITF*) as the highest-scoring modulator. MITF is a master regulator of melanocyte development, function, and survival (Levy et al., 2006; Steingrímsson et al., 2004), and the overexpression of *MITF* is known to have an adverse effect on patient survival (Garraway et al., 2005).

To test the association between modulator and module, we obtained an experimentally derived list of MITF targets (Hoek et al., 2008b) and asked whether the modules identified by CONEXIC associate MITF with its known targets. The *MITF*-associated modules contained 45 of 80 previously identified targets (p value $< 1.5 \times 10^{-45}$) supporting a match between the transcription factor (TF) and its known targets. However, a few targets (*TBC1D16*, *ZFP106*, and *RAB27A*) are both associated with *MITF* and are themselves modulators of additional modules. CONEXIC limits each gene to a single module, so association with an MITF target would preclude association with MITF. If we permit indirect association to *MITF* through the modules of these additional modulators, CONEXIC correctly identifies 76 of the 80 targets identified by Hoek et al. (p value $< 1.5 \times 10^{-78}$). Similar target sets are not available for any other modulator, precluding a more rigorous evaluation of our other predictions.

MITF Expression Correlates with Targets Better Than Copy Number

Expression of *MITF* correlates with the expression of its targets better than *MITF* copy number, though both correlations are statistically significant (p value of 0.0001 versus 0.04; Figures 4A and 4B). This relationship is unidirectional: *MITF* is significantly overexpressed when its DNA is amplified (p value 0.0004), but overexpressed *MITF* does not always correspond with *MITF* amplification. We find that *MITF* is less correlated with its copy number (rank 294th) than most other genes in aberrant regions (see Table S1C), and more than half of the tumors that overexpress *MITF* do not have a CNA that spans the *MITF* gene. Comparison of *MITF* target expression between samples with and without *MITF* amplification did not show an effect of DNA amplification on expression of the targets (Extended Experimental Procedures).

MITF Correctly Annotated with Its Known Role in Melanoma

We used GO gene set enrichment to identify the biological processes and pathways represented in each module associated with *MITF*. The module containing the genes most significantly upregulated by *MITF* (Figure 4B and Figure S4A) is significantly enriched for the terms "melanosome" and "pigment granule" (q value = $4.86e^{-6}$ for each). It includes targets involved in proliferation such as *CDK2*, consistent with the observation that MITF can promote proliferation via lineage-specific regulation of *CDK2* (Du et al., 2004). The module containing genes most strongly inhibited by *MITF* (Figure 4B and Figure S4B) has a metastatic signature strongly associated with invasion, angiogenesis, the extracellular matrix, and NF-κB signaling. These modules and their annotation suggest that *MITF* serves as a developmental switch between two types of melanoma, in which high *MITF* expression promotes proliferation and low *MITF* expression promotes invasion. Thus, our automated, computationally derived findings dissect a complex response and accurately recapitulate the known literature, including the experimental characterization of MITF (Hoek et al., 2008a).

LitVAN annotated additional modulators with their known role (e.g., *CCNB2* with cell cycle and mitosis; data not shown). The detailed match between the CONEXIC output and empirically derived knowledge of the role of known modulators in melanoma provides confidence in CONEXIC's predictions for modulators that are not well characterized.

Identification of *TBC1D16* as a Tumor Dependency in Melanoma

The second highest-scoring modulator identified by CONEXIC is *TBC1D16*, a Rab GTPase-activating protein of unknown biological function. Rabs are small monomeric GTPases involved in membrane transport and trafficking. TBC1D16 is well conserved, and although its targets are not known, a close paralog, TBC1D15, regulates RAB7A (also selected as a modulator; Figure 2) (Itoh et al., 2006). We used a module associated with *TBC1D16* to infer its potential role in melanoma (Figure 5A)

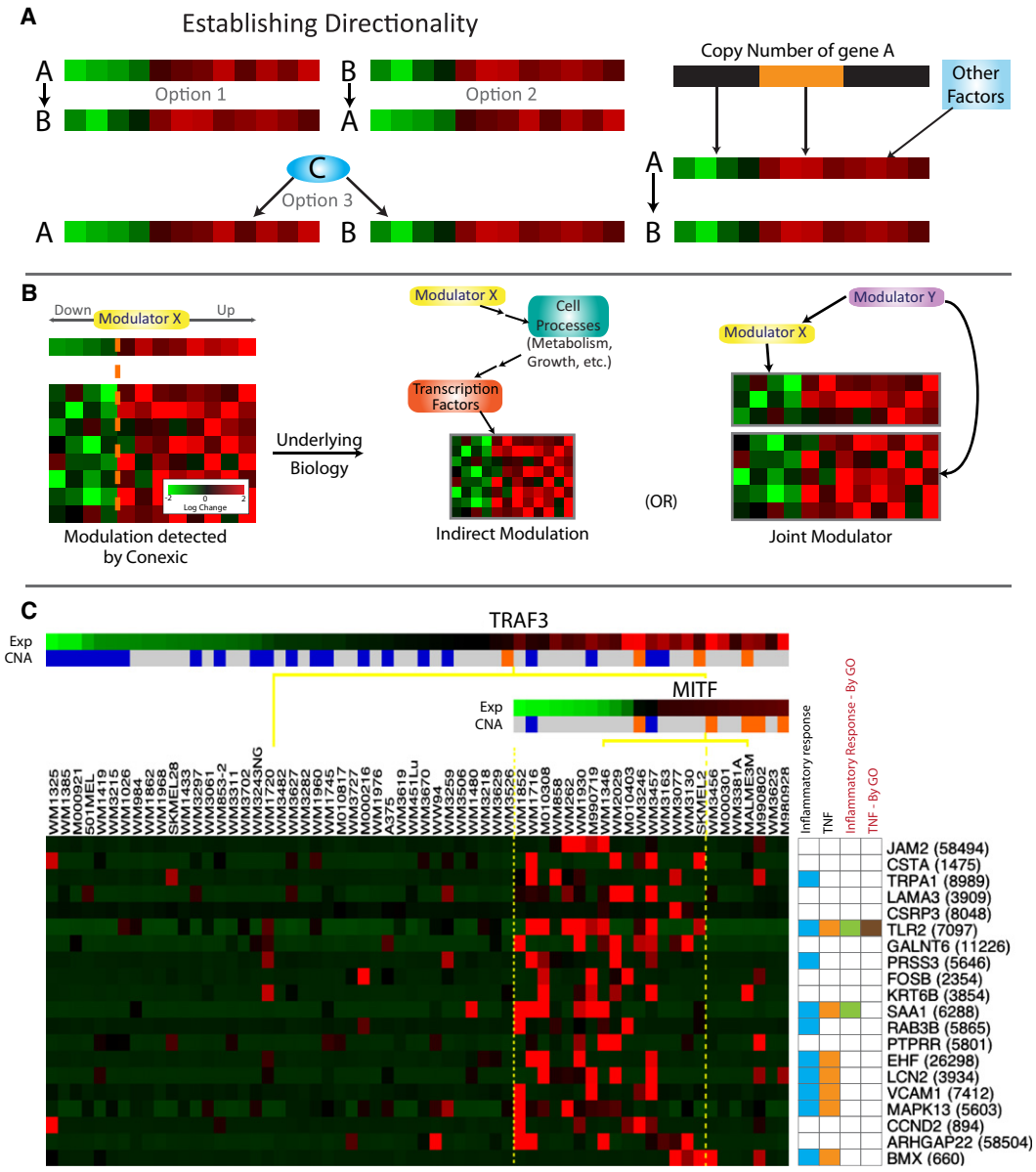


Figure 3. Associating Modulators to Genes

(A) Three scenarios could explain a correlation between a candidate driver (gene A) and its target (gene B): A could influence B, B could influence A, or both could be regulated by a common third mechanism (Pearl, 2000). The availability of both gene expression and chromosomal copy number data allows us to establish the likely direction of influence. If the expression of gene A is correlated with its DNA copy number and the copy number is altered in a large number of tumors, it is likely that the copy number alteration results in a change in expression of A in these tumors. So the model in which A influences the expression of B and other correlated genes is the most likely. In this way, examination of both copy number and gene expression in a single integrated computational framework facilitates identification of candidate drivers.

(B) Modulator influence on a module can go beyond direct transcriptional cascades involving transcription factors or signaling proteins and their targets. Genetic alteration of any gene (e.g., a metabolic enzyme) can alter cell physiology, which is sensed by the cell and subsequently leads to a transcriptional response through a cascade of indirect influences and mechanisms. Whereas modules are typically enriched for genes influenced by the modulator, they also contain genes that are coexpressed with the modulator ("joint modulator"). Both types are helpful for annotating the module and determining the functional role of the modulator.

(C) The TNF module. The modulators include *TRAF3* and *MITF*, wherein high *TRAF3* and low *MITF* are required for upregulation of the genes in the module. The annotation for each gene is represented in a color-coded matrix. Blue and orange squares represent literature-based annotation (see Table S3); green and brown are from GO. LitVAN associated the genes in this module with TNF and the inflammatory response.

See also Figure S2 and Table S3.

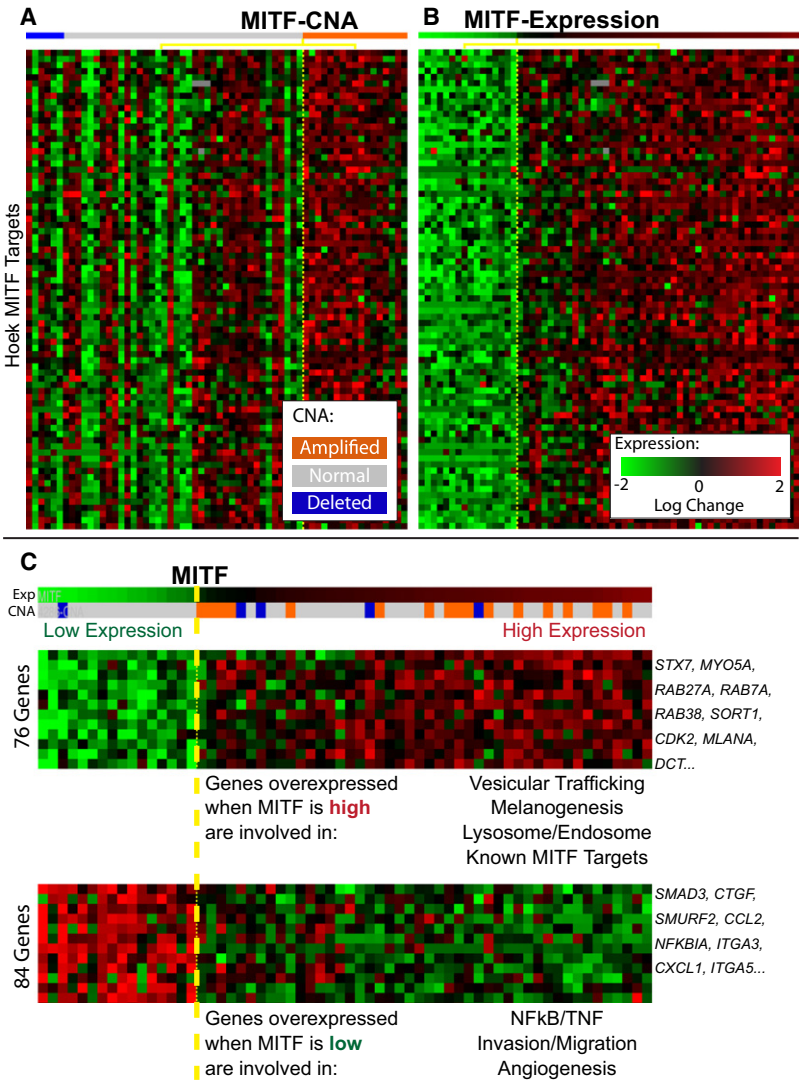


Figure 4. *MITF* Expression Correlates with Expression of the Genes in the Associated Module

(A) Each row represents the gene expression of 1 of 78 MITF targets identified by Hoek (Hoek et al., 2008b); the tumor samples are split into two groups based on the copy number of *MITF* (Welch t test p value = 0.04).

(B) The rows represent the same genes, in the same order as in (A), but here, the tumor samples are split into a group of samples that express *MITF* at high (n = 46) or low levels (n = 16) (Welch t test p value = 0.0001).

(C) Two modules associated with *MITF*, showing a selected subset of genes. LitVAN annotation for the genes in each module is shown below the heat map. The complete modules with all genes are available in Figure S4.

and discovered that diverse biological processes are represented by genes in the module and that more than half are annotated for processes such as melanogenesis, vesicular trafficking, and survival/proliferation (Table S4A). This suggests that *TBC1D16* plays a role in cell survival and proliferation.

TBC1D16 is an uncharacterized gene located in an amplified region that contains 23 other genes, including *CBX4*, which is known to play a role in cancer (Satijn et al., 1997). Expression of *TBC1D16* is not highly correlated with *TBC1D16* copy number compared to other genes in the region (ranked 7th out of 24) or to all candidate drivers (252th out of 428). Nevertheless, *TBC1D16* is the top-scoring gene in the region and the second highest-scoring modulator, so it was selected for experimental verification.

The module exhibits a dose-response relationship between *TBC1D16* expression and the expression of genes in the module such that higher expression of *TBC1D16* is correlated with higher expression of genes in the module (correlation coefficient 0.76).

We carried out western blotting and RT-PCR on some of the short-term cultures (STCs) used to generate the Lin data set and asked whether the *TBC1D16* transcript correlates with protein levels. The results confirmed that the expression of *TBC1D16* corresponds well with the amount of the 45 kD isoform of TBC1D16 (data not shown). These results suggest that knockdown of *TBC1D16* expression in tumors that have high levels of *TBC1D16* will lead to a reduction in proliferation.

TBC1D16 Is Required for Proliferation

To test whether *TBC1D16* is required for proliferation of melanoma cultures, we carried out a knockdown experiment. We selected two STCs with high levels of *TBC1D16*, WM1960 (16-fold higher expression than WM1346, DNA not amplified) and WM1976 (34-fold higher expression, amplified DNA) and control STCs, WM262 and WM1346 that express *TBC1D16* at a lower level. We used two shRNAs to knock down *TBC1D16* expression in each of the four STCs and measured growth over 8 days (Extended Experimental Procedures). RT-PCR was used to confirm that the reduction in the amount of the *TBC1D16* transcript was similar for all of the STCs (Figure S5). Knockdown of *TBC1D16* expression reduced cell growth in WM1960 and WM1976 to 16% and 40%, respectively, relative to controls infected with GFP shRNA in the same STCs (Figures 5B–5D). This result is specific for cultures with high levels of *TBC1D16*, as the controls, WM262 and WM1346, grow at similar rates to cultures infected with shGFP (75%–90%). As predicted, growth inhibition at day 8 is proportional to the amount of the *TBC1D16* transcript and is independent of *TBC1D16* copy number (Figures 5C and 5D). Taken together, these results support CONEXIC's prediction that *TBC1D16* is required for proliferation in melanomas that overexpress the gene.

RAB27A Identified and Experimentally Confirmed as a Tumor Dependency

The *TBC1D16* module contains a second modulator, *RAB27A*, also known to be involved in vesicular trafficking (Figure 5A).

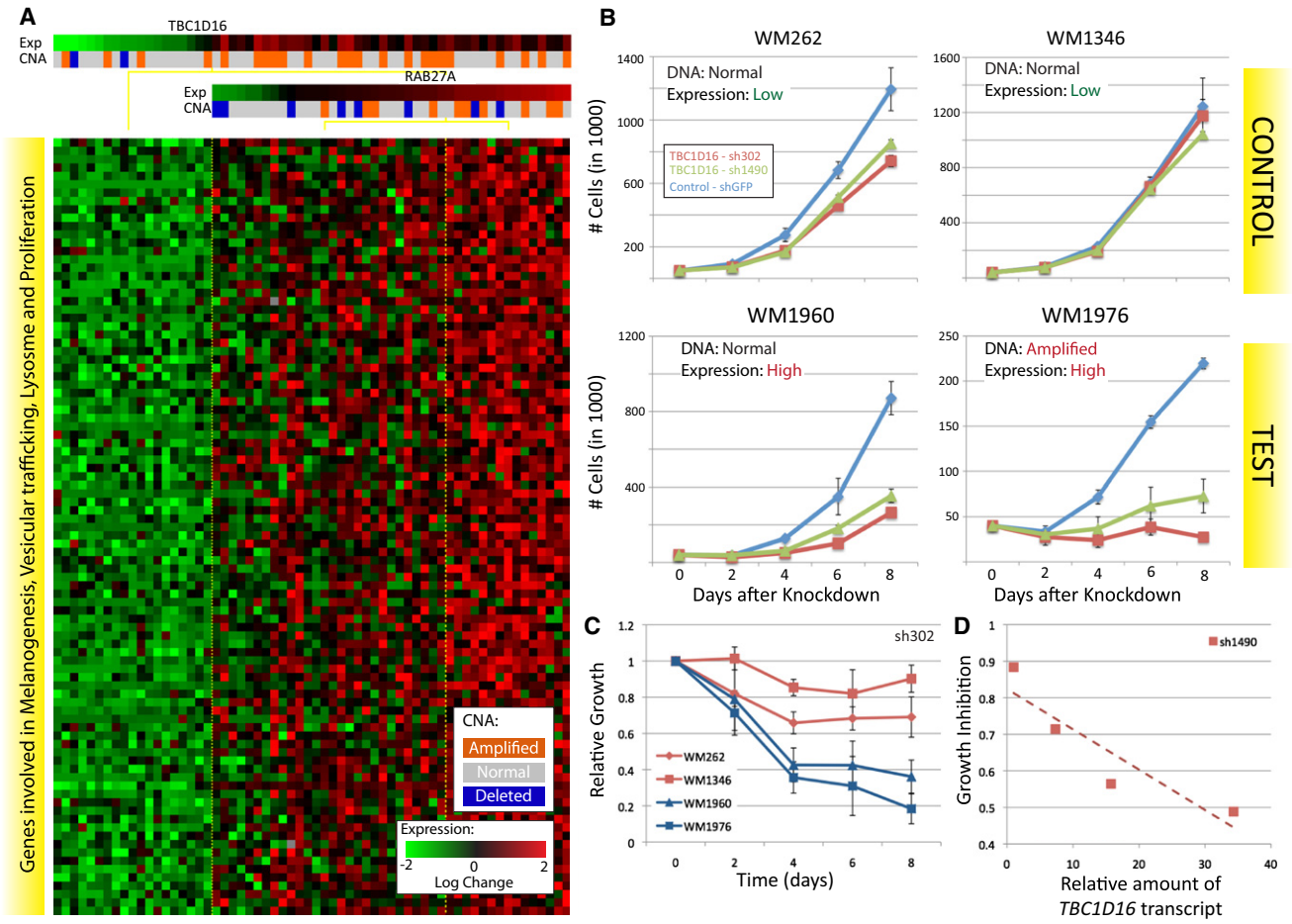


Figure 5. *TBC1D16* Is Necessary for Melanoma Growth

(A) A module associated with *TBC1D16* and *RAB27A*. The genes in the module are involved in melanogenesis, survival/proliferation, lysosome, and protein trafficking (see Table S4A for details).

(B) Representative growth curves for each of the four STCs infected with *TBC1D16* shRNA. Each curve represents three technical replicates. RT-PCR was used to confirm that the reduction in the amount of the *TBC1D16* transcript was similar for all of the STCs (Figure S5).

(C) Change in growth over time, relative to the number of cells plated, averaged over all replicates (Extended Experimental Procedures). Mean over three biological replicates × three technical replicates for each STC. See Figure S5 and Table S4B for additional replicates and hairpins.

(D) Growth inhibition at 8 days is directly proportional to the amount of the *TBC1D16* transcript and is independent of the *TBC1D16* copy number.

RAB27A functions with RAB7A to control melanosome transport and secretion. RAB7A localizes to early melanosomes, whereas RAB27A is found in mature melanosomes (Jordens et al., 2006). CONEXIC selected both *RAB27A* and *RAB7A* as modulators.

RAB27A is in an amplified region that did not pass the standard GISTIC q value threshold for significance, and expression of the gene is not highly correlated with *RAB27A* copy number compared to other candidate drivers (323th out of 428). Nevertheless, CONEXIC identified it as the top-scoring modulator out of the 33 genes in this region and ranked it 8th out of 64 modulators, and it was therefore selected for empirical assessment.

To test the prediction that *RAB27A* is important for proliferation in tumors with high levels of *RAB27A*, we tested the effect of shRNA knockdown of the *RAB27A* transcript on proliferation. We chose two STCs in which the gene is highly expressed WM1385 (28-fold higher expression compared with A375, DNA amplified) and WM1960 (38-fold higher expression, DNA not amplified) and two controls that express *RAB27A* at a lower level (A375 and WM1930). Western blots show that expression of RAB27A correlates with expression of the cognate gene in these cultures (data not shown).

Knockdown of *RAB27A* expression using shRNA was similar for all cultures (Figure S6) but only reduced cell growth significantly in the STCs that overexpress *RAB27A* (18% or 35% in WM1385 or WM1960 relative to the same cultures infected with GFP shRNA). *RAB27A* shRNA had less impact (growth rates of 65%–80%) in the control STCs that have low RAB27A (Figures 6A and 6B). Growth inhibition at 6 days is correlated with the amount of the *RAB27A* transcript and is independent of *RAB27A* copy number (Figures 6B and 6C). Taken together, these results support CONEXIC's prediction that *RAB27A* is a tumor dependency in melanomas that overexpress *RAB27A*.

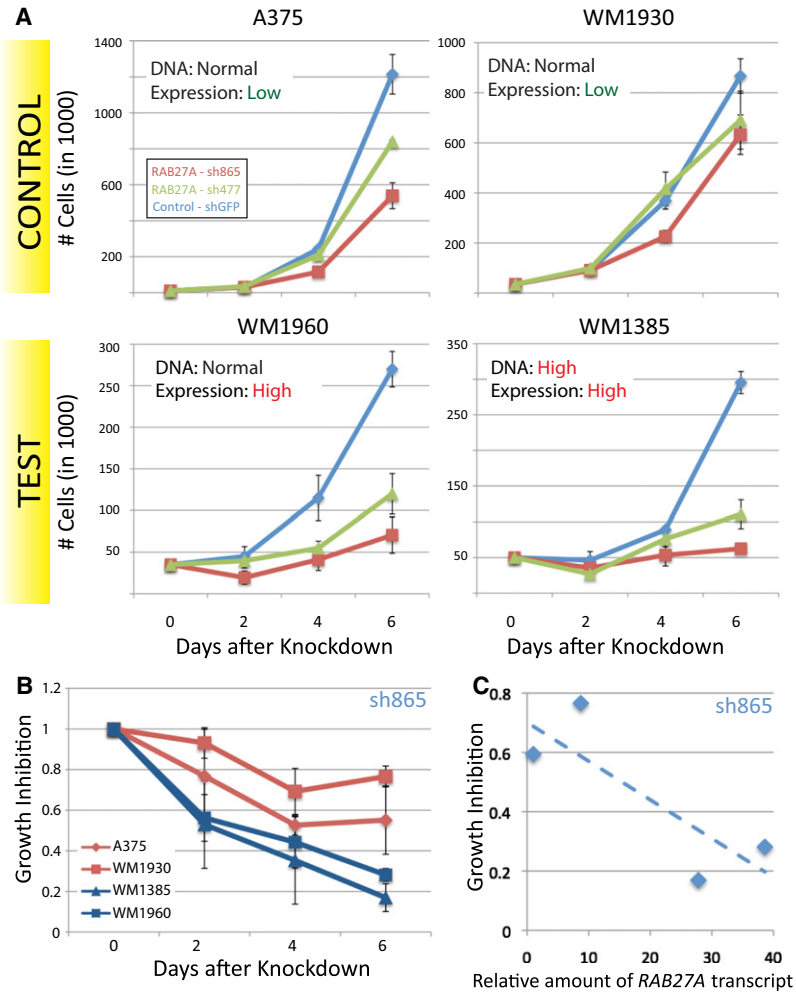


Figure 6. *RAB27A* Is Necessary for Melanoma Growth

(A) Representative growth curves for each of the four STCs infected with *RAB27A* shRNA. Each curve represents three technical replicates. RT-PCR was used to confirm that the reduction in the amount of the *RAB27A* transcript was similar in all of the STCs (Figure S6).

(B) Change in growth over time, relative to the number of cells plated, averaged over all replicates. Knockdown of *RAB27A* expression in cells that express this gene at high levels reduces proliferation. Data averaged over three biological replicates × three technical replicates for each STC. See Figure S6 and Table S5 for all data.

(C) Growth inhibition at 6 days is dependent on the amount of the *RAB27A* transcript and is independent of *RAB27A* copy number.

RAB27A Affects the Expression of Genes in Associated Modules

To test whether *RAB27A* affects the expression of genes in associated modules, as predicted by CONEXIC, we carried out microarray profiling after knockdown of *RAB27A* in the test STCs (WM1385 and WM1960). We compared the expression profile after *RAB27A* knockdown to a control profile generated by infecting the same STC with GFP shRNA. We used *g*ene *s*et *e*nrichment *a*nalysis (GSEA) (Subramanian et al., 2005) to test whether each of the three modules associated with *RAB27A* are enriched with genes that are differentially expressed (DEG) after knockdown (see Extended Experimental Procedures). We found that all three *RAB27A*-associated modules are significantly enriched for genes affected by *RAB27A* (p values $< 10^{-5}$ for all three modules; see Figure 7C) and that these modules responded in the direction predicted by CONEXIC.

These results support our computational prediction that the expression of *RAB27A* affects the expression of the genes in the associated modules. We note that RAB27A functions as a vesicular trafficking protein, suggesting that it influences gene expression through an unknown and likely indirect mechanism.

We used LitVAN to identify the biological processes and pathways represented among the DEGs. Cell cycle-related terms are significant among the downregulated genes, which might be expected given the reduced growth after *RAB27A* knockdown. In addition, we found that genes annotated for the ERK pathway are upregulated (including *MYC*, *FOSL1*, and *DUSP6*). We used GSEA to measure enrichment of an experimentally derived set of genes that respond to MEK inhibition in melanoma (Pratilas et al., 2009). The resulting p value $< 4.7 \times 10^{-5}$ suggests that ERK signaling is altered after *RAB27A* knockdown in these STCs.

TBC1D16 Influences the Expression of Genes in Associated Modules

We carried out microarray profiling after knockdown of *TBC1D16* to evaluate whether expression of *TBC1D16* affects the expression of genes in the four modules associated with it. We used two shRNAs to knock down *TBC1D16* in the test STCs (WM1960 and WM1976) and compared the gene expression to controls infected with GFP shRNA (in the same STCs). GSEA analysis established that all four modules are significantly enriched for genes affected by differences in *TBC1D16* expression (p values $< 10^{-5}$, 0.0002, 0.008, and 0.009, respectively; see Figure 7). Two modules responded to *TBC1D16* knockdown in the direction predicted by CONEXIC. In addition, GSEA analysis ranked genes in the *TBC1D16* module (Module 25) highest out of 177 (based on the GSEA p value), demonstrating that the genes in this module are the most highly differentially expressed genes in the data set.

The function of TBC1D16 is unknown, but it is predicted to be involved in vesicular trafficking. In our knockdown analysis, LitVAN annotated the upregulated genes with terms related to vesicular trafficking. These include *RAB3C*, *RAB7A*, *CHMP1B*, *RAB18*, *SNX16*, *COPB1*, and *CAV1* (see Table S6A). However, it is not clear how *TBC1D16* affects gene expression or how changes in expression affect vesicular trafficking.

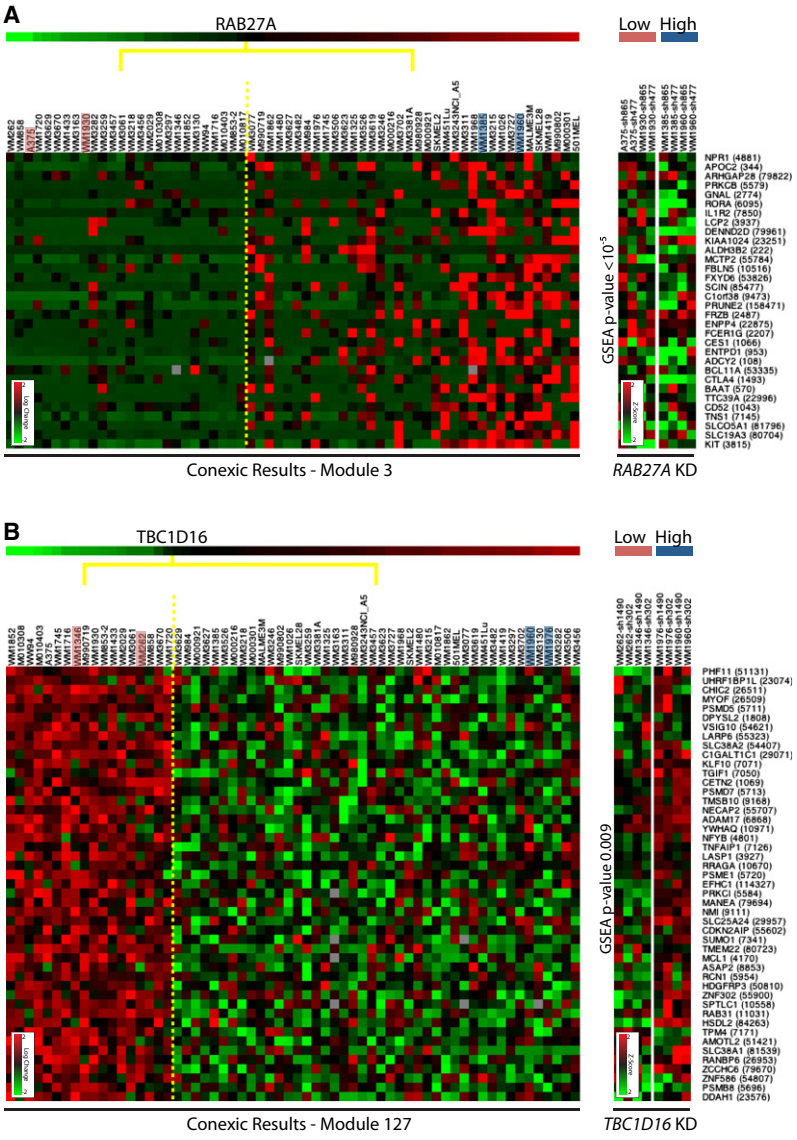


Figure 7. Results of Knockdown Microarrays for *RAB27A* and *TBC1D16*

(A) To the left is one of the modules associated with *RAB27A*, and to the right are data generated following knockdown (KD) of *RAB27A* for the same genes in the STCs indicated (pink and blue). The expression of genes in the module goes down relative to shGFP, as predicted. KD expression heat map shows Z scores (see Extended Experimental Procedures) showing that these are some of the most differentially expressed genes (DEGs) in the genome.

(B) To the left is one of the modules associated with *TBC1D16*, and to the right are data generated following KD of *TBC1D16* in the STCs indicated. The expression of genes in the module goes up relative to shGFP, as predicted. The test STCs (blue) and control STCs (pink) respond differently, demonstrating the importance of context (*TBC1D16* overexpression status) in determining the response.

(C) GSEA p value and ranking (relative to 177 CONEXIC modules) for *RAB27A*- and *TBC1D16*-associated modules (see Figure S7 for data). GSEA was calculated using the median of four profiles (two cell lines × two hairpins) on the test STCs. Significant p values indicate that knockdown of *RAB27A* and *TBC1D16* each affects the subset of genes predicted by CONEXIC (note that 10^{-5} is the smallest p value possible given that 100,000 permutations are used). The color of the module name represents the predicted direction of response to knockdown (red and green represent up- and downregulated, respectively). The arrow represents the observed response to knockdown. The direction of response was correctly predicted for two of four *TBC1D16* modules and for all *RAB27A* modules.

See also Figure S7 and Table S6.

C

RAB27A Moduled Modules

Module	GSEA p-value	Rank
Module 3	$<10^{-5}$	3 ↓
Module 31	$<10^{-5}$	2 ↑
Module 75	$<10^{-5}$	7 ↓

TBC1D16 Moduled Modules

Module	GSEA p-value	Rank
Module 25	$<10^{-5}$	1 ↓
Module 75	0.008	21 ↓
Module 147	2×10^{-5}	5 ↓
Module 127	0.009	22 ↑

DISCUSSION

We have demonstrated that combining tumor gene expression and copy number data into a single framework increases our ability to identify likely drivers in cancer and the processes affected by them. Gene expression allows us to distinguish between multiple genes in an amplified or deleted region (many of which are indistinguishable based on copy number) and to identify those that are likely to be drivers. The combination of data types allows us to identify regions that would be overlooked using methods based on DNA copy number alone.

Expression of a Driver, Not Its Copy Number, Drives Phenotype

The novelty of our method and the key to its success is our modeling paradigm: the expression of a driver should correspond with the expression of genes in an associated module. Examination of MITF and its targets supports our assumptions. Expression of *MITF* best correlates with the expression of its targets, but *MITF* overexpression does not always correspond with *MITF* amplification. A change in DNA copy number is only one of many ways that gene expression can be altered. For example, *MITF* expression can be upregulated via signaling from the Ras/Raf (oncogenic BRAF occurs frequently in melanoma) (Wellbrock et al., 2008) and Frizzled/Wnt pathways (Chin et al., 2006).

Most methods for identifying drivers within aberrant regions focus on genes whose expression is well correlated with the copy number of the cognate DNA (Lin et al., 2008; Turner

et al., 2010). The expression of many of the predicted drivers that we identify is poorly correlated with their copy number, relative to other genes in the region and to all other candidate drivers *MITF* (294th), *TBC1D16* (252th) and *RAB27A* (323th) (see Table S1C). We believe that the discrepancies between CNA and expression arise because there are multiple ways to up- or downregulate a gene. For example, *TBC1D16* and *RAB27A* were both identified as transcriptional targets of MITF (Chiaverini et al., 2008; Hoek et al., 2008b) and are therefore upregulated when *MITF* is overexpressed. Moreover, we postulate that many drivers are less correlated with their copy number than passengers due to selective pressure; if there is a fitness advantage to up- or downregulate expression, the tumor will find a mechanism to do so.

TBC1D16 and *RAB27A* Are Required for Proliferation

We tested two drivers predicted by CONEXIC with knockdown experiments and showed that tumors that express either *TBC1D16* or *RAB27A* at high levels are dependent on the corresponding gene for growth. Our results demonstrate that these dependencies are determined by expression of the gene (in both cases), rather than DNA amplification status, further supporting the assumptions underlying our approach. Thus, we not only identify tumor dependencies, but also the tumors in which these genes are crucial for proliferation. Identifying dependencies that are critical for tumor survival is needed for drug-targeted therapies; for example, FLT3 inhibitors in AML, which have had successful phase II trials (Fischer et al., 2010). Our approach is unbiased with respect to protein function and does not incorporate prior knowledge, thus enabling the identification of dependencies in genes involved with vesicular trafficking. *TBC1D16* and *RAB27A* validate the ability of our approach to correctly identify tumor dependencies and the genes that they affect.

Association between Modulator and Module

A key feature of our approach is that CONEXIC goes beyond identifying drivers. By associating candidate drivers with gene modules and annotating them using information from the literature, CONEXIC provides insight into the physiological roles of drivers and associated genes. We used LitVAn to find biological processes and pathways overrepresented in each module and to associate drivers with functions, accurately identifying targets of *MITF* and annotating the functions of known drivers (*MITF*, *CCBN2*, and *TRAF3*).

The results of microarray profiling following knockdown further support the association between modulator and module and confirm our ability to identify genes affected by *TBC1D16* and *RAB27A*. We successfully connected genes involved in vesicular trafficking to their effects on gene expression, likely through a cascade of indirect influences. In addition to profiling the STCs that highly express each of these genes (test STCs), we also profiled two lower-expressing STCs (control STCs), in which the effect of knockdown is less detrimental to growth. For *TBC1D16*, there is substantial overlap in the DEGs in the test STCs (p value $< 6.6 \times 10^{-22}$), but not in the DEGs between control and test STCs (p value > 0.76). This reflects the complexity of the transformed state and demonstrates that genetic context has a fundamental impact on the effect of a perturbation.

Genes Involved in Trafficking Are Important in Melanoma

Of the top 30 drivers selected by CONEXIC, three genes (*TBC1D16*, *RAB27A*, and *RAB7A*) are known to be involved in vesicular trafficking (Itoh et al., 2006; Jordens et al., 2006). All of these genes are amplified (DNA) and highly expressed (RNA) in multiple melanomas. There is increasing evidence that genes controlling trafficking play a role in melanoma. Germline variation in *Golgi phosphoprotein 3* (*GOLPH3*), a gene involved in vesicular trafficking, is associated with multiple cancers (Scott et al., 2009). Our data identify two novel dependencies that are encoded in somatic CNAs, demonstrate the dependency of melanoma on *TBC1D16* and *RAB27A* expression for proliferation, and highlight the potential role of vesicular trafficking in this malignancy.

The role of vesicular trafficking in melanoma has yet to be characterized. Vesicular trafficking regulates many receptor tyrosine kinases (RTKs) both spatially and temporally and thus determines both the duration and intensity of signaling (Ying et al., 2010). For example, RAB7A is involved in the regulation of ERK signaling (Taub et al., 2007), and ERK is known to play an important role in melanoma (Chin et al., 2006). Tight control of ERK expression could potentially be important in melanocytes because of its influence on MITF: ERK is required for the activation of MITF, but high levels of ERK lead to MITF degradation (Wellbrock et al., 2008). It is possible that recurrent aberrations in vesicular trafficking genes might involve control of ERK signaling intensity. This is further supported by the upregulation of an ERK signature (Pratilas et al., 2009) following *RAB27A* knockdown in our data (p value $< 4.7 \times 10^{-5}$).

CONEXIC and Other Approaches

CONEXIC differs from other methods in a number of ways. First, it uses the gene expression of a candidate driver, rather than its copy number, as a proxy to report on the status of the gene, e.g., two tumors that overexpress a driver are treated equivalently even if there is amplification in the DNA of only one of them. Second, it associates a candidate driver with a module of genes whose expression corresponds with that of the predicted driver, which was critical for identification of *TBC1D16* as a modulator. Third, combining copy number and gene expression provides greater sensitivity for identifying significantly aberrant regions that would not be selected based on DNA alone; this was critical for the identification of *RAB27A*.

Methods based on copy number data are limited to detecting large regions containing multiple genes, such that the driver cannot be readily identified among them. Recent efforts have focused on integrating additional sources of information into the analysis. Some methods use prior information, such as the role of a gene in other cancers (Beroukhim et al., 2010). Others, like CONEXIC, integrate gene expression data (Adler et al., 2006), but the results of these methods fall short of CONEXIC's. We systematically compared CONEXIC to other methods using the same data and found that they did not identify *MITF* or any other known driver in melanoma (see Extended Experimental Procedures).

Statistical dependencies in gene expression have been used to connect a regulator to its target (Friedman et al., 2000; Lee et al., 2006; Segal et al., 2003) and for uncovering important regulators in cancer (Adler et al., 2006; Carro et al., 2010; Wang et al., 2009a). These approaches typically only detect transcription factors and signaling molecules and do not connect the altered regulatory networks to upstream genetic aberrations. Incorporating information on amplification or deletion status allows us to consider any functional class of genes and thus permits detection of vesicular trafficking genes that would not be identified using other methods. It also allows us to relate the malignant phenotype to genetic aberrations from which it is likely to have originated.

We tuned our method toward reducing the selection of modulators that are not drivers. To gain this specificity, we do not detect all genes and pathways that drive tumors. First, some drivers in amplified and deleted regions do not pass the stringent statistical tests employed in our method. Second, CONEXIC only identifies candidate drivers that are encoded in amplified or deleted regions. In consequence, it would not detect drivers of melanoma such as *BRAF* and *NRAS* that are typically associated with point mutations. Third, CONEXIC detects drivers based on the assumptions delineated above; though these hold for many drivers, it is likely that they are not appropriate for all drivers.

To meet the challenge of finding all driving alterations in cancer, a number of complementary approaches are needed. Experimental approaches such as screening using pooled short hairpin RNAs (shRNAs) (Bric et al., 2009; Zender et al., 2008) are likely to detect a set of drivers different from those detected by CONEXIC. These screens are dependent on the genetic background and are limited to drivers that influence processes that can be readily measured, such as proliferation, whereas CONEXIC scans all of the genetic data together and can potentially identify drivers of any function across different genetic backgrounds. In the future, we envision that CONEXIC will be used to guide in vivo screening initiatives and to assist in the choice of regions, functional assays, and genetic backgrounds probed.

Beyond Melanoma

The challenge of finding candidate drivers is considerable: tumors are heterogeneous, the data are noisy and highly correlated, and there are a large number of possible combinations of drivers and genes in modules. Our approach is successful because it couples simple modeling assumptions with powerful computational search techniques and rigorous statistical evaluation of the results at each step.

Both the principles underlying CONEXIC and the software can be applied to any tumor cohort containing matched data for copy number aberrations and gene expression. The principle of associating any type of mutation (e.g., epigenetic alterations and coding sequence) with gene expression signatures or other phenotypic outputs that differ among samples will be of increasing importance as sequence and epigenetic data accumulate. Not only does this help to distinguish between driver and passenger mutations, but the genes in the associated module can also provide insight into the role of the driver. This approach can be used to identify the genetic aberrations responsible for tumorigenesis and to find those that relate to any other measurable phenotype, such as the resistance of tumors to drugs. We anticipate that our approach will make an important contribution toward a basic mechanistic understanding of cancer and in revealing associations of clinical significance. Cancer is a heterogeneous disease in which we are only just beginning to appreciate the importance of genetic background and the myriad ways in which the cellular machinery can be redirected toward the transformed state. Methods that begin to dissect this complexity move us another step closer to a world where personalized therapies are routine.

EXPERIMENTAL PROCEDURES

Statistical Methods

A detailed description of the statistical methods and computational algorithms used can be found in the Extended Experimental Procedures. The CONEXIC and LitVAN algorithms were developed for this research, and the software is available at http://www.c2b2.columbia.edu/danapeerlab/html/software.html.

Experimental Methods

Cells were grown using standard culture conditions, and knockdown was carried out by infection with lentivirus using RNAi sequences designed by the RNAi Consortium. shRNA lentivirus were prepared according to TRC protocols (http://www.broadinstitute.org/rnai/trc), with minor modifications. Cell proliferation assays, RT-PCR, microarrays, and immunoblotting were carried out using standard techniques. Primer sequences and detailed methods can be found in the Extended Experimental Procedures.

ACCESSION NUMBERS

All primary data are available at the Gene Expression Omnibus (GSE23884).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, eight figures, and six tables and can be found with this article online at doi:10.1016/j.cell.2010.11.013.

ACKNOWLEDGMENTS

The authors will like to thank Nir Hacohen, Antonio Iavarone, Daphne Koller, Liz Miller, Itsik Pe'er, Suzanne Pfeffer, Neal Rosen, and Olga Troyanskaya for valuable comments. This research was supported by the National Institutes of Health Roadmap Initiative, NIH Director's New Innovator Award Program through grant number 1-DP2-OD002414-01, and National Centers for Biomedical Computing Grant 1U54CA121852-01A1. D.P. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and Packard Fellowship for Science and Engineering.

Received: May 13, 2010
Revised: August 31, 2010
Accepted: October 22, 2010
Published online: December 2, 2010

REFERENCES

Adler, A.S., Lin, M., Horlings, H., Nuyten, D.S., van de Vijver, M.J., and Chang, H.Y. (2006). Genetic regulators of large-scale transcriptional signatures in cancer. Nat. Genet. *38*, 421–430.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc. Natl. Acad. Sci. USA *104*, 20007–20012.

Beroukhim, R., Brunet, J.-P., Di Napoli, A., Mertz, K.D., Seeley, A., Pires, M.M., Linhart, D., Worrell, R.A., Moch, H., Rubin, M.A., et al. (2009). Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. Cancer Res. *69*, 4674–4681.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Bric, A., Miething, C., Bialucha, C.U., Scuoppo, C., Zender, L., Krasnitz, A., Xuan, Z., Zuber, J., Wigler, M., Hicks, J., et al. (2009). Functional identification of tumor-suppressor genes through an in vivo RNA interference screen in a mouse lymphoma model. Cancer Cell *16*, 324–335.

Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., et al. (2010). The transcriptional network for mesenchymal transformation of brain tumours. Nature *463*, 318–325.

Chiaverini, C., Beuret, L., Flori, E., Busca, R., Abbe, P., Bille, K., Bahadoran, P., Ortonne, J.-P., Bertolotto, C., and Ballotti, R. (2008). Microphthalmia-associated transcription factor regulates RAB27A gene expression and controls melanosome transport. J. Biol. Chem. *283*, 12635–12642.

Chin, L., Garraway, L.A., and Fisher, D.E. (2006). Malignant melanoma: genetics and therapeutics in the genomic era. Genes Dev. *20*, 2149–2182.

Du, J., Widlund, H.R., Horstmann, M.A., Ramaswamy, S., Ross, K., Huber, W.E., Nishimura, E.K., Golub, T.R., and Fisher, D.E. (2004). Critical role of CDK2 for melanoma growth linked to its melanocyte-specific transcriptional regulation by MITF. Cancer Cell *6*, 565–576.

Fischer, T., Stone, R.M., Deangelo, D.J., Galinsky, I., Estey, E., Lanza, C., Fox, E., Ehninger, G., Feldman, E.J., Schiller, G.J., et al. (2010). Phase IIB trial of oral Midostaurin (PKC412), the FMS-like tyrosine kinase 3 receptor (FLT3) and multi-targeted kinase inhibitor, in patients with acute myeloid leukemia and high-risk myelodysplastic syndrome with either wild-type or mutated FLT3. J. Clin. Oncol. *28*, 4339–4345.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601–620.

Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J., et al. (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature *436*, 117–122.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531–537.

Hoek, K.S., Eichhoff, O.M., Schlegel, N.C., Döbbeling, U., Kobert, N., Schaerer, L., Hemmi, S., and Dummer, R. (2008a). In vivo switching of human melanoma cells between proliferative and invasive states. Cancer Res. *68*, 650–656.

Hoek, K.S., Schlegel, N.C., Eichhoff, O.M., Widmer, D.S., Praetorius, C., Einarsson, S.O., Valgeirsdottir, S., Bergsteinsdottir, K., Schepsky, A., Dummer, R., and Steingrimsson, E. (2008b). Novel MITF targets identified using a two-step DNA microarray strategy. Pigment Cell Melanoma Res *21*, 665–676.

Hoek, K.S., Schlegel, N.C., Brafford, P., Sucker, A., Ugurel, S., Kumar, R., Weber, B.L., Nathanson, K.L., Phillips, D.J., Herlyn, M., et al. (2006). Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. Pigment Cell Res. *19*, 290–302.

Itoh, T., Satoh, M., Kanno, E., and Fukuda, M. (2006). Screening for target Rabs of TBC (Tre-2/Bub2/Cdc16) domain-containing proteins based on their Rab-binding activity. Genes Cells *11*, 1023–1037.

Johansson, P., Pavey, S., and Hayward, N. (2007). Confirmation of a BRAF mutation-associated gene expression signature in melanoma. Pigment Cell Res. *20*, 216–221.

Jordens, I., Westbroek, W., Marsman, M., Rocha, N., Mommaas, M., Huizing, M., Lambert, J., Naeyaert, J.M., and Neefjes, J. (2006). Rab7 and Rab27a control two motor protein activities involved in melanosomal transport. Pigment Cell Res. *19*, 412–423.

Lee, S.-I., Pe'er, D., Dudley, A.M., Church, G.M., and Koller, D. (2006). Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc. Natl. Acad. Sci. USA *103*, 14062–14067.

Levy, C., Khaled, M., and Fisher, D.E. (2006). MITF: master regulator of melanocyte development and melanoma oncogene. Trends Mol. Med. *12*, 406–414.

Lin, W.M., Baker, A.C., Beroukhim, R., Winckler, W., Feng, W., Marmion, J.M., Laine, E., Greulich, H., Tseng, H., Gates, C., et al. (2008). Modeling genomic diversity and tumor dependency in malignant melanoma. Cancer Res. *68*, 664–673.

Pearl, J. (2000). Causality: models, reasoning, and inference (Cambridge, U.K.; New York: Cambridge University Press).

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. *20*, 207–211.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.-L., Ordóñez, G.R., Bignell, G.R., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Pratilas, C.A., Taylor, B.S., Ye, Q., Viale, A., Sander, C., Solit, D.B., and Rosen, N. (2009). (V600E)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. Proc. Natl. Acad. Sci. USA *106*, 4519–4524.

Sanchez-Garcia, F., Akavia, U.D., Mozes, E., and Pe'er, D. (2010). JISTIC: identification of significant targets in cancer. BMC Bioinformatics *11*, 189.

Satijn, D.P., Olson, D.J., van der Vlag, J., Hamer, K.M., Lambrechts, C., Masselink, H., Gunster, M.J., Sewalt, R.G., van Driel, R., and Otte, A.P. (1997). Interference with the expression of a novel human polycomb protein, hPc2, results in cellular transformation and apoptosis. Mol. Cell. Biol. *17*, 6076–6086.

Scott, K.L., Kabbarah, O., Liang, M.-C., Ivanova, E., Anagnostou, V., Wu, J., Dhakal, S., Wu, M., Chen, S., Feinberg, T., et al. (2009). GOLPH3 modulates mTOR signalling and rapamycin sensitivity in cancer. Nature *459*, 1085–1090.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nat. Genet. *36*, 1090–1098.

Steingrímsson, E., Copeland, N.G., and Jenkins, N.A. (2004). Melanocytes and the microphthalmia transcription factor network. Annu. Rev. Genet. *38*, 365–411.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Taub, N., Teis, D., Ebner, H.L., Hess, M.W., and Huber, L.A. (2007). Late endosomal traffic of the epidermal growth factor receptor ensures spatial and temporal fidelity of mitogen-activated protein kinase signaling. Mol. Biol. Cell *18*, 4698–4710.

Turner, N., Lambros, M.B., Horlings, H.M., Pearson, A., Sharpe, R., Natrajan, R., Geyer, F.C., van Kouwenhove, M., Kreike, B., Mackay, A., et al. (2010). Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. Oncogene *29*, 2013–2023.

Vallabhapurapu, S., Matsuzawa, A., Zhang, W., Tseng, P.-H., Keats, J.J., Wang, H., Vignali, D.A.A., Bergsagel, P.L., and Karin, M. (2008). Nonredundant and complementary functions of TRAF2 and TRAF3 in a ubiquitination cascade that activates NIK-dependent alternative NF-kappaB signaling. Nat. Immunol. *9*, 1364–1370.

Wang, K., Saito, M., Bisikirska, B.C., Alvarez, M.J., Lim, W.K., Rajbhandari, P., Shen, Q., Nemenman, I., Basso, K., Margolin, A.A., et al. (2009a). Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat. Biotechnol. *27*, 829–839.

Wang, L., Cao, X.X., Chen, Q., Zhu, T.F., Zhu, H.G., and Zheng, L. (2009b). DIXDC1 targets p21 and cyclin D1 via PI3K pathway activation to promote colon cancer cell proliferation. Cancer Sci. *100*, 1801–1808.

Wellbrock, C., Rana, S., Paterson, H., Pickersgill, H., Brummelkamp, T., and Marais, R. (2008). Oncogenic BRAF regulates melanoma proliferation through the lineage specific factor MITF. PLoS ONE *3*, e2734.

Ying, H., Zheng, H., Scott, K., Wiedemeyer, R., Yan, H., Lim, C., Huang, J., Dhakal, S., Ivanova, E., Xiao, Y., et al. (2010). Mig-6 controls EGFR trafficking and suppresses gliomagenesis. Proc. Natl. Acad. Sci. USA *107*, 6912–6917.

Zender, L., Xue, W., Zuber, J., Semighini, C.P., Krasnitz, A., Ma, B., Zender, P., Kubicka, S., Luk, J.M., Schirmacher, P., et al. (2008). An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. Cell *135*, 852–864.

EXTENDED EXPERIMENTAL PROCEDURES

Data and Processing

We used copy number data for 101 melanoma samples generated by Lin et al. (Lin et al., 2008b). The SNP locations were translated from HG17 to HG18 using the UCSC liftOver application (Kent et al., 2002). Gene expression was available for 62 of these samples using HT-HGU133A Affymetrix chip, which measures the expression of 12725 genes (Lin et al., 2008b). We removed probe sets whose standard deviation was smaller than 0.25 on a log2 scale, resulting in 12,101 probe sets measuring 8243 unique genes. We merged probe sets for genes if these agreed and removed inconsistent genes, resulting in a final set of 7981 genes. Expression values were normalized to mean of zero and a standard deviation of one for each gene.

Statistical Tests

Validation of Modules Using Other Melanoma Data Sets

We evaluated the reproducibility of the regulation programs and their associated modules (see Regulation Programs below) using gene expression data from two other melanoma cohorts with 45 (Hoek et al., 2006) and 63 (Johansson et al., 2007) samples downloaded from the GEO website (GSE4843 & GSE7127, respectively). We used CONEXIC to infer modules and their regulation programs using the data of Lin et al. (Lin et al., 2008b) (training data) and tested these in two additional datasets (test data): (Hoek et al., 2006) and (Johansson et al., 2007). We removed data from one sample (SKMEL28) that was included in the Lin and Hoek datasets. Each dataset was processed separately as described above. We used the models (regulation programs and the module gene assignments) learned from the Lin dataset to evaluate the likelihood that the model generated the test data. The regulation program (modulators, their order, logic and thresholds) remains exactly the same as defined by the training data. Permutation testing was used to evaluate how well the regulation programs predict the behavior of the genes in the modules. For each modulator, the samples in its sub-tree (where this modulator is root) were permutated 1,000,000 times and were scored for each permutation. This constructs a null distribution representing Normal Gamma scores when there is no dependency between modulator and module genes. The score for the non-permuted data is compared to this distribution, generating a p-value. When a modulator was associated with more than one module the highest scoring was selected. These permutation p-values are listed in Figure 2, as well as Figure S3 and Table S2. Figure 2 and Figure S3 show the p-values from the Johansson dataset, unless the modulator was missing in this dataset, and then the p-value from the Hoek dataset is shown. Table S2 shows the p-values for both datasets.

Analysis of MITF Targets

We used a list of MITF targets as defined by Hoek et al. (Hoek et al., 2008), of which 80 genes were expressed in the Lin et al. (Lin et al., 2008b) data set. Association between MITF copy number and the expression of the target genes was tested using permutations on Welch's t test statistic (Welch, 1938), which evaluates the difference in both mean and in variance. Samples were split into two groups: 17 samples with amplification were compared to 47 samples with deleted or normal *MITF* DNA copy number (see Figure 4A). To test association between *MITF* expression and the expression of the target genes, the CONEXIC split point was used, splitting the data into 46 samples with high levels of *MITF* and 16 samples with lower levels of *MITF* expression (see Figure 4B). The p-value of *MITF* copy number was 0.04, compared to a p-value of 0.0001 for *MITF* expression.

To determine if *MITF* amplification has an additional effect besides modifying *MITF* expression we identified samples in which *MITF* expression was similar in tumors with and without DNA amplification. Two sets of samples were defined: 18 samples that have high *MITF* expression and 11 samples, with medium levels of *MITF* expression. Comparison of *MITF* target expression between samples with and without *MITF* amplification in these two groups did not show an effect of DNA amplification on the target expression. Using Welch t test and 1000 permutations, no gene passed FDR significance (t test q-value of 0.92 and 0.56, respectively).

The *MITF* target set defined by Hoek et al. (Hoek et al., 2008) includes both direct and indirect targets positively regulated by *MITF*. To evaluate how well CONEXIC associates *MITF* with this target set, we test both direct and indirect modulation. We combined all the genes contained in modules for which *MITF* is the top positive modulator yielding a set of 307 genes, 45 of which are Hoek identified *MITF* targets (p-value $< 1.5 \times 10^{-45}$). Among these 307 are genes selected as modulators (*TBC1D16*, *ZFP106*, and *RAB27A*). Modules associated with *TBC1D16* contain 22 genes from the *MITF* target set, and since each gene is only associated with one module, these are not in the *MITF* modules. If we include indirect modulation (e.g., *MITF* positively modulates gene Y which positively modulates gene X), then 76/80 targets are associated with *MITF*.

Gene Set Enrichment

We used Genatomy for hypergeometric gene-set enrichment. The tool is available for download from http://www.c2b2.columbia.edu/danapeerlab/html/software.html.

Genatomy computes Hyper-geometric enrichment based p-values combined with False Discovery Rate to account for multiple testing. Gene Ontology (GO) sets were obtained from GeneOntology.org. Genatomy also enables the interactive visualization of all modules and regulation programs inferred by CONEXIC.

Analysis of Growth Data

Technical replicates were averaged and standard deviation was calculated (see Figure 5B, Figure 6A, Table S4B, and Table S5). For each hairpin, the fold change of growth was calculated; number of cells at each day was compared to the number of cells plated. The

effect of test hairpins was assessed by comparing the fold change of the test hairpins to the control hairpin (shGFP) of the corresponding day. Effects from biological replicates were then averaged (see Figure 5C, Figure 6B, Figure S5, Figure S6, Table S4B, and Table S5).

Processing of KD Microarrays

Knockdown experiments were hybridized on Affymetrix Human Gene 1.0 ST arrays. Gene expression data was processed using RMA. For each STC, we computed a *baseline profile* by averaging data from two biological replicate arrays, generated following infection with shGFP. We computed a *KD profile* representing the effect of knockdown with each hairpin in each STC using the log ratio of the data generated following knockdown with the shRNA in the STC, with the baseline profile for the same STC. This gives us two KD profiles for each STC, one for each of the two hairpins. The KD profiles were used for subsequent analysis (e.g., GSEA below).

For displaying the heat maps in Figure 7 and Figure S7 the KD profiles are represented as Z-scores: each KD profile was treated as a distribution and the fold change for each gene was transformed to number of standard deviations away from the mean. This shows that the changes in expression of genes in these modules are among the largest, genome-wide.

Analysis of KD Microarray Data

We used Gene Set Enrichment Analysis (Subramanian et al., 2005) to test whether the genes associated with *RAB27A* and *TBC1D16* are differentially expressed after knockdown. As a ranking function for GSEA analysis, we did not use each microarray as a rank, but generated only one signature for each gene knocked down. We used the median of 4 experiments (both hairpins, both cell lines) in all *test* samples, to generate rankings, one for *RAB27A* and one for *TBC1D16*. This was to control for the off-target effects of shRNA (use of 2 hairpins) and differences between genetic contexts (use of 2 cell-lines), and gives us an answer to the question "which genes change following knockdown of the modulator." We used the standalone version of GSEA 2.06, available from http://www.broadinstitute.org/gsea/ and ran it with the default parameters and 100,000 permutations.

We selected the modules significantly associated with each modulator: 4 for *TBC1D16* and 3 for *RAB27A.* GSEA was run on all 177 modules (size > 15 genes) to obtain the relative rank of each. All of the modules associated with *TBC1D16* or *RAB27A* had a significant p-value for the corresponding KD signature. For *RAB27A*, all modules were significant with a p-value $< 10^{-5}$ (note that for 100,000 permutations, this is the lowest possible p-value), responding in the direction predicted by CONEXIC. For *TBC1D16*, all modules were significant, although 2/4 modules responded in a direction opposite to our CONEXIC prediction. The highest scoring module (Normal Gamma score) also ranked first in GSEA.

Compared to the p-values in Figure 7 the *RAB27A* KD signature applied to *TBC1D16* modules gave poor results for all modules (module 25: p-value > 0.09, module 147: p-value > 0.16 and module 127: p-value > 0.06). Applying the *TBC1D16* KD signature to the *RAB27A* modules also gave an inferior result (module 3: p-value > 0.046 and module 31: p-value > 0.005) showing that the match between KD signatures and associated modules is specific for each of *RAB27A* and *TBC1D16*.

We used the *TBC1D16* and *RAB27A* KD profiles to identify sets of DEGs for each modulator, using a Z-score of 2 to define a threshold for up and downregulation. For *RAB27A*, 432 genes were upregulated and 412 genes were downregulated (Table S6B, S6C). LitVAN identified a number of terms including cell-cycle, MAPK/ERK, TNF/NFkB, immune system, extra-cellular matrix, invasiveness and metastasis. Following this analysis, we took a set of 52 genes whose expression changes significantly and consistently following MEK inhibition across multiple melanoma cell lines (Pratilas et al., 2009). We ran GSEA on these genes and found that their expression is substantially different following *RAB27A* KD (p-value $< 4.7 \times 10^{-5}$). For *TBC1D16*, 489 genes were upregulated and 304 genes were downregulated (Tables S6D and S6E). The terms identified by LitVAN include trafficking, ER, Golgi, RAB and immune system.

The hyper-geometric p-value was used to evaluate the significance of the overlap in DEGs between the two test STCs for each modulator. For *TBC1D16* the upregulated genes overlap with p-value $< 6.6 \times 10^{-22}$ and the downregulated genes overlap with p-value $< 1.6 \times 10^{-6}$. For *RAB27A* the upregulated genes overlap with p-value $< 1.5 \times 10^{-47}$ and the downregulated genes overlap with p-value $< 1.2 \times 10^{-21}$. The same analysis was carried out, comparing between the test and the control STC data. For *TBC1D16*, the upregulated genes overlap with p-value > 0.65 and the downregulated genes overlap with p-value $< 6 \times 10^{-9}$. These results suggest that knockdown of *TBC1D16* influences the control and test STCs in substantially different ways. The test and control STCs are more similar for *RAB27A.*

Experimental Methods

Cell Culture Conditions

Melanoma short-term cultures (STCs) derived from metastatic foci (Lin et al., 2008b) were cultured in RPMI medium (MediaTech) supplemented with 10% fetal bovine serum (Gemini). A375 melanoma cell line and 293T virus packaging cells were cultured in DMEM medium (MediaTech) with 10% fetal bovine serum. All cells were maintained in 100 units/ml penicillin, 100 μg/ml streptomycin at 37°C under 5% CO_2.

shRNA

shRNA knockdown sequences for TBC1D16 and RAB27A were those designed by the RNAi Consortium (TRC): shRAB27A_865 (TRCN0000005296), CGGATCAGTTAAGTGAAGAAA; shRAB27A_477 (TRCN0000005298), CAGGAGAGGTTTCGTAGCTTA; shTBC1D16_302 (TRCN0000061889), CCTGTGCTTGTACATGGAGAA; and shTBC1D16_1490 (TRCN0000061891), GCGAAAGGAGTACTCTGAGAT.

The negative control was: shGFP, GCAAGCTGACCCTGAAGTTCA.

The shRNA lentiviruses were produced according to TRC protocols (http://www.broadinstitute.org/rnai/trc). Briefly, 2 × 10^6 293T cells were seeded in 100 mm plates, and at 24 hr were co-transfected with 3 μg of pLKO.1-shRNA, 2.7 μg of Δ8.9, and 0.3 μg of VSV-G vectors using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. At 48h post-transfection, viral supernatant was collected, passed through a 0.45 μm filter (Nalgene), and stored in small aliquots in −80°C until use. To reduce nonspecific viral cytotoxicity, each viral supernatant was titrated: Target cells were seeded in 6-well plates (1 × 10^5 – 2 × 10^5/ well), and at 24h post-seeding, were infected with dilutions (1:5, 1:10, 1:20, 1:40, 1:100, 1:200) of each lentivirus preparation. Infections were carried out in duplicate, in 2 ml medium/ well containing 6 μg/ml polybrene for 6h. 48h post-infection, puromycin was added to one set of infected cells, leaving the remaining set unselected. Cells were washed with PBS, trypsinized, and collected for counting using a Vi-Cell XR Cell Viability Analyzer (Beckman Coulter) after 3-4 days of selection. For each virus, the ratio of cells surviving in media with puromycin versus media without puromycin was determined for each dilution. Titers that yielded approximately 50% survival were used for subsequent infections.

To perform knockdown experiments, target cells (1.2 × 10^6 – 3 × 10^6) were plated the day before infection to obtain 30%–40% confluence at the time of infection. Cells were then incubated with virus at the dilution established above, the virus-containing media was removed and fresh media was added. The next day, puromycin was added to select for infected cells (WM1960 and WM1385 at 2 μg/ml; WM1346, WM1976, WM1930, WM262 and A375 at 1 μg/ml). After selection, cells were plated for proliferation assays or used for RT-PCR or immunoblotting.

Cell Proliferation Assays

Cultures with stable expression of each shRNA construct were seeded in triplicate (technical replicates) in 12-well plates at 1x10^4 - 5x10^4 cells/ well in 1 ml of medium. Cells were washed with PBS, trypsinized and counted using a Vi-Cell XR Cell Viability Analyzer (Beckman Coulter) or Coulter particle counter (Beckman Coulter) at the times indicated.

Quantitative Reverse Transcription-PCR Analysis

Total RNA was harvested using the RNeasy Mini Kit (QIAGEN), and cDNA prepared with the SuperScript III First-Strand Synthesis Supermix Kit (Invitrogen) according to the manufacturers' recommendations. All real-time PCRs were performed in triplicate using PCR SYBR Green Master Mix (Applied Biosystems) on an ABI 7300 (Applied Biosystems). The data were normalized to *TBP*. Gene-specific primer sequences follow:

TBP: forward, 5′- CCACTCACAGACTCTCACAAC-3′, reverse, 5′- CTGCGGTACAATCCCAGAACT-3′; RAB27A: forward, 5′- GAAACTGGATAAGCCAGCTACAG-3′, reverse, 5′- ATATTTCTCTGCGAGTGCTATGG-3′; TBC1D16_114: forward, 5′- CTACT CCAAGAACAATGTCTGCG-3′, reverse, 5′- GCCTCTGGATGCGAGAGTTG-3′; *TBC1D16*_1211: forward, 5′- CGCCCCCGATAAGA CATGC-3′, reverse, 5′- CCTTCCGCAGCTTGTACTC-3′; *TBC1D16*_1752: forward, 5′- GATGAGTCAGACACCTTC-3′, reverse, 5′- GGTACAGCAGTTGTTTCT-3′.

Microarray Analysis

Total RNA was harvested using the RNeasy Mini Kit (QIAGEN). The Affymetrix GeneChip Human Gene 1.0 ST Arrays were used for gene expression profiling, performed by the Dana Farber Cancer Institute facility according to the manufacturer's protocols.

Immunoblot Analysis

Western blotting was carried out using standard methods on cell lysates, normalized for total protein. Primary antibodies used were RAB27A (Santa Cruz, 1:500), TBC1D16 (Novus Biologicals, 1:500), β-actin (Sigma-Aldrich, 1:12,000) or α-tubulin (Cell Signaling Technology, 1:1000). Secondary antibodies were anti-rabbit or anti-mouse IgG, HRP-linked; Cell Signaling Technology, 1:1000 dilution). The signal was detected using SuperSignal West Pico or West Dura Chemiluminscent Substrate ECL reagent (Thermo Scientific).

Copy Number and Expression In Cancer

*Co*py *n*umber and *e*xpression *in ca*ncer (CONEXIC) is a data-driven algorithm that takes matched copy number and gene expression data from tumors as input and combines these to identify driving aberrations and associate these with the genes they modulate. CONEXIC is based on a Bayesian scoring function that evaluates each candidate driver, or 'modulator'. The score measures how well a modulator (or combination of modulators) predicts the behavior of a gene expression module across tumor samples. CONEXIC identifies the most likely drivers by searching for the highest scoring modulators in a stepwise manner (Figure S1) that seeks to improve the Bayesian score.

The resulting output is a set of predicted modulators that map within an amplified or deleted region. Each is associated with a module, a set of genes whose expression is likely altered by changes in expression of the modulator. In some cases, a module is associated with multiple modulators that form a regulation program (Figure S3A).

The tool is available from http://www.c2b2.columbia.edu/danapeerlab/html/software.html.

Regulation Programs

We adopted the concept of the regulation program from Segal et al. (Segal et al., 2003). Formally, the regulation program is a conditional probability distribution for the module gene expression, conditioned on the gene expression of the modulators. A regulation program of a module *M* specifies a set of contexts and the expected expression values for each context. A context is determined by the expression of a small set of modulators that influence *M*'s expression. This set of contexts is organized as

a regression tree composed of two basic building blocks: decision nodes and leaf nodes. Each decision node corresponds to a modulator and a query on its value (for example, ''is MITF $\geq$ threshold''). Each decision node has two child nodes: the right child node is chosen when the answer to the corresponding query is true; the left node is chosen when it is false. For each sample, one begins at the root node and continues down the tree in a path according to the answers to the queries in that particular sample until a leaf node is reached. Each leaf encodes a probability distribution representing how the module's genes are expected to behave in that sample. The expression of genes in *M* in each context is modeled as a normal distribution; this distribution is encoded using a mean and variance stored at the corresponding leaf.

Regression trees are particularly well suited for modeling driver mutations in cancer because (i) these can capture combinatorial and condition specific relations that frequently occur in cancer (e.g., both overexpression of EGFR and deletion of P16 are required) (ii) These can capture changes in both the mean and the variance of the module gene expression.

The CONEXIC Algorithm

The CONEXIC learning algorithm consists of three key steps:

1. Selection of candidate driver genes (Figure S1A).
2. Single Modulator step that creates an initial association between candidate drivers and gene modules (Figure S1B).
3. An iterative Network Learning step to improve on the initial model (Figure S1C).

CONEXIC searches for a model that can explain the variation in the gene expression across samples as a function of a small number of modulators. The search is driven by the optimization of a Bayesian scoring function. The search begins with an initial starting point and then proceeds in making stepwise changes that improve the score at each step.

Selection of Candidate Drivers

Motivation

First, we want to identify regions of the DNA that are recurrently amplified/deleted in tumors and consider genes within or neighboring to those regions as candidate drivers. We expect that many of the driver mutations will be contained in this candidate list.

Details

We applied the (GISTIC) algorithm (Beroukhim et al., 2007), using the JISTIC implementation available from http://www.c2b2.columbia.edu/danapeerlab/html/software.html (Sanchez-Garcia et al., 2010), to all 101 samples. To increase sensitivity, we used a q-value threshold of 0.3, compared to 0.25 previously used with this data (Lin et al., 2008b). Genes that overlap a significant aberrant region are chosen as candidate driver genes. To capture aberrations in regulatory regions, for each aberrant region, the closest non-overlapping genes on each side are also chosen as candidate drivers, if their distance from the edge of the region is less than 100Kb.

Result

This resulted in 27 amplified regions containing 513 peak genes and 23 deleted regions with 384 peak genes (Table S1). In subsequent steps, we aim to identify which of these 897 candidates are likely drivers.

Expression Filtering

We now integrate copy number and gene expression data, which is available for 62 tumors. As an initial filter, we require candidate drivers to be differentially expressed across the different tumor samples. This removes genes that are expressed at a constant level across all tumors and not influenced by their copy number. Additionally, this removes genes that are not expressed, and are therefore unlikely to be drivers. Our final set of candidate drivers included 428 genes in significantly amplified or deleted regions, whose gene expression varied with standard deviation greater than 0.25 (Table S1C).

Single Modulator Step

Motivation

The Single Modulator step constructs an initial model, which focuses the subsequent search on variation in gene expression that can be explained by drivers encoded in CNAs (as opposed to variation due to other types of aberrations such as coding mutations). The Single Modulator step (Figure S1B), establishes an initial pairing between candidate drivers and gene expression modules by associating each target gene with the single driver gene that fits it best (based on corresponding gene expression profiles). As a result, each gene is clustered into a module consisting of those genes for which the same driver gene was found to be the best fit and the module is associated with that candidate driver.

Details

To aid the identification of a good starting model, the Single Modulator step considers a smaller search space and more conservative set of candidate drivers, only those whose gene expression is significantly altered by either their amplification or deletion status. Candidates are filtered using a Welch t test (p-value < 0.05), comparing amplified versus normal or deleted versus normal - 347 candidate drivers pass this test. Amplification or deletion status of a gene in a specific sample is determined using the average copy-number value for all SNP markers inside the gene; if the gene contains no SNP markers we take the copy-number value for the single closest SNP marker. Using the same thresholds as in (Lin et al., 2008b), if the

average value of the SNP markers around a gene is above 0.3, the gene is marked as amplified, if the average value is below −0.3,it is marked as deleted.

For each candidate driver gene, we use the gene expression values of the amplified/deleted samples to guide the choice of threshold, and consider the gene expression of the amplified/deleted samples to represent appropriate high/low expression levels. We use k-means clustering, using k = 2 and the normal and amplified/deleted samples as the two initial clusters to fit two normal distributions. The boundary between the two clusters is the selected expression threshold level for this driver gene. The expression of each target gene is split into two sets: those in the tumor samples in which the driver's expression is below the threshold, and those in the tumor samples in which the driver's expression is above the threshold. The Normal Gamma scoring function is used to compute the quality of this split, thus measuring a target gene's fit with a candidate driver.

After the score is computed for all pair-wise combinations of candidate drivers and target genes, each gene is assigned to the single highest scoring candidate driver. Permutation testing is used to verify the statistical significance of association between driver and gene. Driver gene expression is randomly permuted 10,000 times, conserving the composition of values, but rendering the order random and independent of the target gene. Each of these permutations is scored, creating a null distribution to compare with the unpermuted order; thus providing a p-value for the association between gene and candidate driver. If this p-value < 0.001, we associate between gene and candidate driver, declaring the candidate driver a modulator of its associated gene. Care must be taken to avoid spurious associations due to the dense correlation structure of genes encoded in the same aberrant region, as it is easy to obtain associations between all candidate drivers in a region with the same target gene. Thus, it is important to find only the single best association between a gene and its modulator (see Figure S2).

We established a number of additional criteria to ensure the robustness of our results; these criteria were guided by results on randomly permuted data. First, we require that each modulator be chosen by at least 20 genes. If some of the modules have less than 20 members, we break up the smallest module and reassign its genes to the next best scoring modulator, repeating until all modules have at least 20 members. Second, we apply non-parametric bootstrapping (Efron, 1979) and repeat this procedure 100 times, generating variations on the dataset using random re-sampling with replacement. This ensures that the association is not an artifact of the specific set of samples and is robust across different subsets of the original data. We select candidate drivers that were selected in at least 90% of the runs.

We then make one final run of the Single Modulator step, using this filtered set of 130 candidate driver genes. The selected set of modulators all reside in significantly aberrant regions and their gene expression best corresponds with at least 20 genes. The statistical significance of this association is ensured both by permutation testing and non-parametric bootstrap.

Result

Single Modulator step identified 78 modulators that explain the behavior of 4018 genes. Each of the 78 modulators is associated with a module containing at least 20 genes. These will be refined in subsequent steps.

Network Learning step

Motivation

The Network Learning step (Figure S1C) uses the modules generated by the Single Modulator step as a starting point and uses an iterative approach to improve the score of the modules and their regulation programs. The Network Learning step is based on the Module Networks algorithm (Lee et al., 2006; Segal et al., 2003) with a few critical improvements, designed to remove spurious association, described below.

Details

The algorithm iteratively alternates between two tasks: (i) learning the regulation program for each module; (ii) and re-assigning each gene into the module that best models its behavior. The score improves at each iteration and these terminate if fewer than 10% of the target genes have been re-assigned to a different module during the gene re-assignment step.

Given a set of modules, we learn a regulation program for each module. We recursively learn the regulation program by choosing, at each point, the candidate driver that best splits the gene expression of the module genes into two distinct behaviors. All candidate drivers and potential split values are evaluated and the driver-split combination that achieves the highest improvement in score is selected. Only if the score improvement is greater than a pre-defined penalty, the split is selected. Unlike Single Modulator, all 428 candidate drivers are considered and each candidate driver is not limited to a single split threshold; rather the optimal threshold is chosen to maximize the score. While multiple thresholds are possible across the different regulation programs, the score includes a penalty on the number of different split values, thus limiting the number selected. The tree is recursively grown from the root to its leaves. At each new split, the driver gene that provides the best improvement in score is chosen; permutation testing (as in the Single Modulator step) is used to ensure the statistical significance of the split; and the outlier-removal test (described below) is used to ensure the modulator was not selected due to outliers in the data. For each new split, linear influence of the modulator on each side of the split is tested (as described below), and if linear influence is found further sub-splits on this side of the split are forbidden. The process terminates when no query that improves the score and passes these two tests can be found, allowing for up to a total of five splits in each regulation program.

In addition to permutation testing, the regulation program learning process includes two additional statistical tests, designed to remove spurious splits. (i) To ensure that modulators were not selected due to outliers in the data, splits are subjected to

a outlier-removal test, as follows: for each candidate split, we remove the highest 4% of expression values in the side that has higher mean expression, and re-calculate the score improvement with the remaining 96% of the data. We reject the split if the score improvement is less than 0.6 times the score improvement with the entire data. We perform a similar test removing the lowest 4% of expression values in the side that has lower mean expression. The resulting splits are robust to outliers in the data.

(ii) Some of the modulators have a linear influence on the target gene expression and their values are enough to explain the variation in the expression of the target genes without additional splits. In most cases, such influence is only on one side of the split, e.g., target gene expression of one leaf is linearly correlated with the modulator expression, while the expression of the other leaf is not. Our goal is to remove additional splits when the modulator is correlated across all samples and retain splits in cases the correlation is one sided. Correlation alone is not sufficiently sensitive to distinguish between a strong correlation that is limited to one side of the split, versus a weaker correlation across all samples. To make this distinction, for each new split, and for each side of that split, we calculate the Pearson correlation and regression slope between the expression values of the modulator and of the members of the module. First, we ask whether the modulator is correlated with the module, and require a Pearson correlation coefficient > 0.6 for at least one of the leaves. Next, we evaluate whether the slope on both sides of the split is similar, requiring that the slope in the leaf with higher correlation and the slope of the combined data (on both leaves) be within a ratio of 0.7 to1.4. If both criteria hold, we forbid any further sub-splits on the correlated side.

Given the inferred regulation programs, we determine the module whose associated regulation program best predicts each gene's behavior. Specifically, we iterate over all genes, one at a time, and move each gene into the module that provides the highest improvement in the score. This step is guaranteed to improve the score, or leave it the same (if the gene is not moved). We repeat this reassignment process for all genes three times, at every iteration.

Similarly to the Single Modulator step we boost robustness using non-parametric bootstrap. The iterative learning algorithm is run 100 times. We then filter the set of candidate driver genes, leaving only genes that appeared in at least one regulation program in at least 40% of the runs. 64 modulators pass this threshold and continue to the next step. We then make one final run of the Network Learning algorithm, using this filtered set of candidate driver genes.

Results

This resulted in the identification of 64 modulators that explain the behavior of 7869 genes. We compared the models at the beginning versus at the end of the Network Learning step and found the end model superior by a number of measures:

1. The final model can explain the behavior of 7869 genes, relative to only 4018 at the end of Single Modulator (starting model).
2. The test log-likelihood is significantly higher for the final model, relative to the initial model, in a leave-one-out cross validation (see Figure S8D and robustness section below).
3. *TBC1D16* is a good example of the need for the more aggressive search performed by network learning. It was the second-highest scoring modulator at the end of network learning and empirically validated. Due to its more limited search, Single Modulator did not select *TBC1D16* at all.

Model Refinement

The candidate drivers used for the regulation programs include only genes residing in CNA regions, which are expected to explain only part of the global changes in gene expression. Observed changes in gene expression can also result from additional factors, such as somatic mutations that are not included in our data. While the algorithm attempts to assign all genes to modules, some genes expression is not influenced by any CNA based driver. Therefore, it is important to remove genes that can't be explained by any regulation program, or more formally, no program significantly improves the gene's likelihood. For each gene, we calculate the difference between the likelihood of its data using its assigned regulation program and the likelihood of its data without any regulation program. For each module, we calculate the distribution of these delta likelihoods; those genes for which the delta likelihood is two standard deviations or more below the mean are removed from the module. These genes can't be explained by any regulation program and will not be members of any module. A final iteration of learning the regulation program is executed after these genes are removed.

Score Function

We use a Bayesian scoring approach that maximizes the overall joint probability of both the data and of the model structure. Let D represent the data and S represent the structure of the network, then the scoring function is expressed as $\log P(D,S) = \log P(D|S) + \log P(S)$. Where the first term is the likelihood of the data for a given model (in the Bayesian approach we integrate over all possible model parameters) and the second term is the prior on the structure for which we use a penalty score on model complexity.

Following the Module Networks approach (Segal et al., 2003) we use Normal Gamma distribution for our likelihood function, see Segal et al. (Segal et al., 2005) for full details. Normal Gamma gives a higher score to data with lower variance and hence finds splits that create two different contexts that represent two distinct behaviors (normal distributions). The Normal Gamma score is described below:

$$NormalGamma(Leaf, \lambda, \alpha):$$

$$N = Size(Leaf)$$

$$\beta = Max\left(1, \frac{\lambda^*(\alpha - 2)}{\lambda + 1}\right)$$

$$\beta^+ = \beta + \frac{Var(Leaf)^*N}{2} + N^*\lambda^* \frac{\overline{Leaf^2}}{2^*(N+\lambda)}$$

$$\alpha^+ = \alpha + \frac{N}{2}$$

$$Score = -N^*\ln\left(\sqrt{2\pi}\right) + \frac{\ln\left(\frac{\lambda}{\lambda+N}\right)}{2} + \ln\left(\Gamma\left(\alpha^+\right)\right) - \ln(\Gamma(\alpha)) + \alpha^*\ln(\beta) - \alpha^{+*}\ln\left(\beta^+\right)$$

Leaf is a vector of gene expression values contained in the leaf and α,λ and β are parameters. A split is scored by comparing the score of the split data to the score without the split, along with a penalty for the split.

$$NormalGamma(Left\_Leaf) + NormalGamma(Right\_Leaf) > = NormalGamma(Entire\_data) + Penalty$$

Our penalty function is comprised of two parts. Following Module Networks (Segal et al., 2003), we use a complexity prior that penalized the number of leaves in each regulation program, using the exponential distribution over total number of leaves. Denoting the regulation program as T and L as number of leaves, $\log P(T) = -\beta L$. Following genetic Module Networks (Geronemo (Lee et al., 2006)), in addition to a penalty specific to each regulation program, we have a network wide penalty function that penalizes the total number of modulators. The prior takes the form of a power-law distribution on the number of modulators. This prior encourages the algorithm to select a sparse number of modulators, which is particularly important in this application, whose main purpose is to identify a small set of potential drivers. Full details are available in (Lee et al., 2006).

The scoring function has 5 parameters, α and λ for the Normal Gamma distribution and β, x and y for the complexity prior. These were selected using 10-fold cross validation and the parameters used were $\alpha = 2$, $\lambda = 1$, $\beta = 20$, x = 15 and y = 0.

Parameter Selection and Robustness

Selection of Candidate Drivers

Selection of candidate drivers requires determining a q-value threshold for GISTIC, the higher the threshold, the more candidate regions and genes will be selected, 0.25 is typically used as a threshold for determining the final list of significant regions (Beroukhim et al., 2009; Beroukhim et al., 2007; Lin et al., 2008b; Walter et al., 2009). Within CONEXIC, GISTIC is used to only generate a pool of candidate genes for further selection, so we used the more permissive threshold of 0.3. It is likely that there are additional drivers even beyond a threshold of 0.3, but too many candidate modulators burden CONEXIC both computationally and statistically. Therefore, we selected a threshold of 0.3 and correctly identified *CCNB2* and *RAB27A* as drivers in region below the 0.25 threshold, demonstrating increased sensitivity.

Single Modulator Step

The Single Modulator requires a confidence threshold for non-parametric bootstrap. We selected 90, meaning that we only selected modulators chosen in more than 90% of the bootstrap runs. Before removing modules containing fewer than 20 genes the median single modulator run included 295 modulators. After removing small modules, a median of 202 modulators still remained. Following bootstrap with a threshold of 90% only 78 remained.

Why did we choose 90? In a histogram representing the number of modulators at each confidence threshold (Figure S8A) we observe that below 90 the distribution of modulators at each confidence level flattens and becomes uniform. It is important to note that this threshold does not define a filter, but rather only a starting point for Network Learning, which reconsiders all 428 candidate drivers. Indeed, 10 modulators that are not selected at this stage are included in the final model, including *TBC1D16* and *ZPF106*.

CONEXIC achieves similar results across a broad range of thresholds and the final results bear significant similarity, even in a comparison between using 80 versus 95 as a threshold. Using 80 as a threshold results in 60 modulators and using 95 as a threshold

results in 57 modulators, the overlap between them is 45 modulators. The final model using 80 as a threshold is much closer to our final model, with complete overlap in all the modulators discussed in this manuscript.

Network Learning Step

While the model resulting from Network Learning is higher scoring than the model resulting from Single Modulator, the former has more parameters and hence over-fitting is a concern. To evaluate whether the learned models can be generalized to unseen data, we compared the models derived from Single Modulator versus Network Learning using leave-one-out cross validation. For each sample (tumor), the two models were learned using the 61 other samples. We calculated the test log-likelihood of the held out 'test' sample for each of the models. In essence, we are testing how well the model "predicts" a new sample. The likelihood of the held out sample was consistently better in the Network Learning model, for almost all of the 62 samples (Figure S8D).

Network learning reconsiders all candidate drivers, so bootstrap is needed for this step, for the same reason it is needed in Single Modulator. Here, a clear threshold cannot readily be determined from a histogram representing the number of modulators at each confidence threshold (Figure S8B). Instead, we used a similar histogram generated from the same data, only randomly permuted to determine a confidence threshold that is beyond spurious associations. Observing the histogram in Figure S8C we see there are no spurious associations above 35, so we selected 40 as our threshold.

A concern in this domain is its dense correlation structure; candidate drivers residing in the same region are all correlated to the region's copy number and among themselves. Bootstrapping has a pivotal role in dealing with the spurious correlations that can arise: if a candidate driver corresponds to a module only through its correlation to its own copy number, other candidates in the region are equally likely to be selected and will not hold pass bootstrapping. Only when the candidate driver provides a substantially better score than its neighboring candidates in the region does the association hold. A threshold of 40 is more than enough to serve this role, as most regions have at least 5 genes, typically dozens.

This threshold largely determines the final output and changes in it directly add or remove modulators from the final model. The result is a relatively short, ranked list of putative drivers, however discretion must still be applied. The ranking is informative; the top 10 are more reliable than those ranked 50 to 60. Additional filters can be applied to evaluate this list, such as whether it replicates in other datasets or whether it associates with a module annotated for cancer related processes. Our goal was generating a list of modulators that have a signal above spurious noise using a less stringent threshold, leaving the final evaluation to the informed user.

LitVAn

LitVAn (Literature Vector Analysis) is an automatic literature-based analysis tool for inference of gene module functionality. The basic principle is similar to other gene set enrichment methods, identifying over-represented terms associated with a subset of genes.

We use the NCBI database, which associates each gene with manually curated papers. Our corpus contains around 70,000 full-text papers. The algorithm is based on TF*IDF score, which gives a higher score to words which are overrepresented in a subset of documents relative to the full corpus. Inverse Document Frequency (IDF), gives each "term" (a word) a score based on the portion of documents it appears in, with high scores for low coverage. Term Frequency (TF), is calculated for a subset of documents rather than the entire set, and is a direct count for the number of times the term appears in the subset. For each set of genes (a module), we count the term frequency in papers associated with these genes and compare this count to the null distribution, using a TF*IDF score (Salton and Buckley, 1988).

The TF*IDF score takes the "bag of words" approach, ignoring the order of which the words appear and their location in the text (headers, legends, etc.). The documents are first processed by a semantic stemming algorithm (Van Rijsbergen et al., 1980), which converts words to their most basic form in order to treat different forms of the same word equally. The IDF score is calculated once for the entire compendium and stored, and the TF score is calculated for each module separately, using all the papers linked to genes in the module. To avoid biasing the module score with terms related to only one gene that has many papers associated with it, we use a "Leave-One-Out" score. With this approach, for each term we identify the gene that contributes the most to it, and remove its contribution from the TF score of that term. Although TF is generally defined as a linear score, our tests show clear advantage of using a log2 scale.

To evaluate the significance of the TF*IDF score, we generate an expected score based on random modules. Sets of genes in different sizes, varying from 5 to several hundred genes, were randomly selected and their best LitVAn result was used to determine the expected score (calculated using a linear regression between the top TF*IDF score and the number of papers associated with the random module). Based on the randomized scores, the 95% confidence intervals of the linear regression were used to determine the threshold of significance for a given number of papers.

The output of LitVAn is a ranked list of terms with an indication of their significance level, as well as a map linking the genes, significant terms and papers that contribute to the score. An online version of LitVAn is available at http://litvan.bio.columbia.edu.

Comparison to Other Methods

Comparison to Module Networks

CONEXIC uses an algorithm similar to Module Networks (Segal et al., 2003) as its key statistical engine. Module Networks is based on two principles: (i) influences and interactions between proteins often generate statistical dependencies in gene expression and (ii) testing dependencies on entire modules of genes enables statistical discovery that is undetectable when considering each gene in isolation. Module Networks was designed to infer regulatory models of transcription factors and their upstream signaling proteins.

Module Networks handles the ambiguity between correlation and influence using prior knowledge: Taking a precompiled list of transcription factors and signaling proteins it assumes: 'If a protein that has a known role in transcriptional regulation and is correlated (or anti-correlated) with the expression of genes in a module, it is likely to regulate the genes in that module.' Module Networks was applied to a yeast dataset with 173 samples, to build a regulatory model over 2355 genes (Segal et al., 2003).

CONEXIC has a fundamentally different goal: the identification of drivers of cancer. In this application, the primary role of the module is to provide support for a gene as a driver, the 'network interpretation' is secondary. We apply CONEXIC to a human cancer dataset with 62 samples, to build a model over 7869 genes involving influences that go beyond direct transcriptional cascades. CONEXIC is based on a different set of assumptions. It assumes that perturbations originate in the DNA and this provides the direction of influence (Figure 3A). The set of candidate modulators includes all genes contained within frequently amplified and deleted regions of any functional class and thus extends beyond transcriptional regulation. It aims to capture modulation of expression in response to altered cell physiology (Figure 3B). Aberrations in the DNA lead to perturbations which provide a rich source of variation that can be used to help uncover molecular influences in the cell. Numerous adaptations were made to the Module Networks algorithm to make it more suitable for this application.

We provided the Module Networks algorithm, as originally implemented (Segal et al., 2003) with the same candidate driver set (GISTIC output) used by CONEXIC. The resulting model includes 347 modulators, which are a large fraction of the 428 candidates that pass expression filtering and is of limited use as a selective method for identifying drivers. Additionally, only one of the empirically confirmed *MITF* targets (as defined by Hoek (Hoek et al., 2008) was associated with *MITF*, in contrast to the 45 identified by CONEXIC. The Single Modulator step was used to derive modules for initialization of Module Networks, as opposed to clustering initialization. Starting from the Single Modulator (with bootstrap) defined modules, the final output included 109 modulators, and *MITF* was associated with 46 of its experimentally derived targets (both numbers are median values across 100 runs). We conclude that the Single Modulator initialization method is critical for CONEXIC's success in associating a modulator to genes it influences. It focuses the learning on changes in expression altered by CNA, as opposed to those altered via other mechanisms (e.g., coding mutations in the ERK pathway).

The output from running Module Networks with bootstrap, using a candidate set defined by GISTIC and initialized using Single Modulator, yields results similar to those of CONEXIC. Each of the additional refinements: permutation testing, outlier removal, removal of linear splits and gene removal provide smaller improvements to the final results. For example, the removal of linear splits only removes 21 out of 489 splits in the model. Nevertheless, each of these steps improves the model, as assessed by leave-one-out likelihood tests.

Integration of CNA with Gene Expression

Methods based on copy number information alone, e.g., GISTIC (Beroukhim et al., 2009; Beroukhim et al., 2007; Lin et al., 2008b; Walter et al., 2009) are typically limited to detecting large regions containing multiple genes, such that the driver cannot be readily identified among them. Applied to this melanoma dataset, GISTIC (with q-value of 0.25) finds multiple regions containing 588 genes (Lin et al., 2008b), the drivers and passengers indistinguishable. While GISTIC is a valuable method to filter a genome of ~23,000 genes and derive a set of hundreds of candidate drivers, additional data types are required to narrow this list down further.

A number of different approaches integrate CNA and gene expression by identifying genes with significant correlations between DNA copy number and gene expression. Lin et al. (Lin et al., 2008b) applied the approach to this data and predicted *KLF6* and *CUL2* as putative drivers. Recently (Huh et al., 2010), *KLF6* was validated as a driver in melanoma. *MITF* CNA is poorly correlated with its gene expression and hence not identified with this approach.

SLAMS (Adler et al., 2006) bears some conceptual similarity to CONEXIC, but there are critical differences. SLAMS requires an initial signature that is used to divide the samples into classes and runs SAM (Tusher et al., 2001) to find the copy number that best separates the classes. The algorithm then finds a gene (or multiple genes) in the selected region, where the expression of the gene is a good predictor of the expression signature. In contrast to SLAMS, CONEXIC does not require a pre-defined expression signature, but identifies one or more signatures de novo. To test SLAMS on the melanoma dataset, we used the *MITF* targets identified by Hoek (Hoek et al., 2008) as a signature. SLAMS identified the copy number of 1444 genes as significant, ranking *MITF* as 75th. In contrast, CONEXIC correctly identified *MITF* as the top ranked gene, associated *MITF* with its targets de-novo and predicted many additional drivers.

Witten et al. (Witten et al., 2009) described a method based on applying penalized canonical correlation analysis (CCA) to the cross product matrix of gene expression and CNA data, identifying the regions and correlated genes in a single step. We applied the method to the melanoma dataset using the same steps and parameters as those used in the original paper. This method identified 7980 genes as significant, including almost all the genes influenced by CNA, but did not distinguish the drivers among them.

Methods that Integrate CNAs with Other Data Types

In addition to expression, other data types have been used with CNA to help identify drivers. GRAIL (Raychaudhuri et al., 2009) prioritizes genes within GISTIC regions based on prior knowledge and known gene annotation. GRAIL identified *MCL1* (using 3000 samples across multiple cancers), but failed to find *MITF* or *KLF6* (Beroukhim et al., 2010).

Another approach, NetBox (Cerami et al., 2010), uses a curated human protein-protein interaction database as an additional source of information. It constructs protein-protein interaction networks from genes within recurrently aberrant regions and defines drivers as hubs in these networks. Applied to the Lin melanoma dataset NetBox did not find any significant networks (lowest p-value 0.15). Even considering networks of low significance, NetBox did not identify *MITF*, *KLF6* or any other known melanoma oncogene/

tumor suppressor. Both GRAIL and NetBox are strongly based on prior knowledge and annotations. While they present a powerful approach for identifying oncogenes in new contexts (e.g., *MCL1* which has not yet been verified in melanoma), they only predict drivers among well annotated genes. The advantage of assaying copy number, gene expression, sequencing and other technologies genome-wide is that the data are comprehensive and unbiased. To fully exploit this data we need methods that go beyond the realm of the well annotated.

SUPPLEMENTAL REFERENCES

Adler, A.S., Lin, M., Horlings, H., Nuyten, D.S., van de Vijver, M.J., and Chang, H.Y. (2006). Genetic regulators of large-scale transcriptional signatures in cancer. Nat. Genet. *38*, 421–430.

Annunziata, C.M., Davis, R.E., Demchenko, Y., Bellamy, W., Gabrea, A., Zhan, F., Lenz, G., Hanamura, I., Wright, G., Xiao, W., et al. (2007). Frequent engagement of the classical and alternative NF-kappaB pathways by diverse genetic abnormalities in multiple myeloma. Cancer Cell *12*, 115–130.

Appel, S., Bringmann, A., Grünebach, F., Weck, M.M., Bauer, J., and Brossart, P. (2006). Epithelial-specific transcription factor ESE-3 is involved in the development of monocyte-derived DCs. Blood *107*, 3265–3270.

Baumann, H., and Gauldie, J. (1994). The acute phase response. Immunol. Today *15*, 74–80.

Bellanger, S., de Gramont, A., and Sobczak-Thépot, J. (2007). Cyclin B2 suppresses mitotic failure and DNA re-replication in human somatic cells knocked down for both cyclins B1 and B2. Oncogene *26*, 7175–7184.

Beroukhim, R., Brunet, J.-P., Di Napoli, A., Mertz, K.D., Seeley, A., Pires, M.M., Linhart, D., Worrell, R.A., Moch, H., Rubin, M.A., et al. (2009). Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. Cancer Res. *69*, 4674–4681.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc. Natl. Acad. Sci. USA *104*, 20007–20012.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Brauweiler, A., Lorick, K.L., Lee, J.P., Tsai, Y.C., Chan, D., Weissman, A.M., Drabkin, H.A., and Gemmill, R.M. (2007). RING-dependent tumor suppression and G2/M arrest induced by the TRC8 hereditary kidney cancer gene. Oncogene *26*, 2263–2271.

Cerami, E., Demir, E., Schultz, N., Taylor, B.S., and Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. PLoS ONE *5*, e8918.

Chan, Y.R., Liu, J.S., Pociask, D.A., Zheng, M., Mietzner, T.A., Berger, T., Mak, T.W., Clifton, M.C., Strong, R.K., Ray, P., and Kolls, J.K. (2009). Lipocalin 2 is required for pulmonary host defense against Klebsiella infection. J. Immunol. *182*, 4947–4956.

Chiang, P.W., Oiso, N., Gautam, R., Suzuki, T., Swank, R.T., and Spritz, R.A. (2003). The Hermansky-Pudlak syndrome 1 (HPS1) and HPS4 proteins are components of two complexes, BLOC-3 and BLOC-4, involved in the biogenesis of lysosome-related organelles. J. Biol. Chem. *278*, 20332–20337.

Chiaverini, C., Beuret, L., Flori, E., Busca, R., Abbe, P., Bille, K., Bahadoran, P., Ortonne, J.-P., Bertolotto, C., and Ballotti, R. (2008). Microphthalmia-associated transcription factor regulates RAB27A gene expression and controls melanosome transport. J. Biol. Chem. *283*, 12635–12642.

Di Fulvio, M., Henkels, K.M., and Gomez-Cambronero, J. (2007). Short-hairpin RNA-mediated stable silencing of Grb2 impairs cell growth and DNA synthesis. Biochem. Biophys. Res. Commun. *357*, 737–742.

Dynek, J.N., Chan, S.M., Liu, J., Zha, J., Fairbrother, W.J., and Vucic, D. (2008). Microphthalmia-associated transcription factor is a critical transcriptional regulator of melanoma inhibitor of apoptosis in melanomas. Cancer Res. *68*, 3124–3132.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. *7*, 1–26.

Endo, R., Saito, T., Asada, A., Kawahara, H., Ohshima, T., and Hisanaga, S. (2009). Commitment of 1-methyl-4-phenylpyrinidinium ion-induced neuronal cell death by proteasome-mediated degradation of p35 cyclin-dependent kinase 5 activator. J. Biol. Chem. *284*, 26029–26039.

Fargnoli, M.C., Fargnoli, M.C., Pike, K., Pfeiffer, R.M., Tsang, S., Rozenblum, E., Munroe, D.J., Golubeva, Y., Calista, D., Seidenari, S., et al. (2008). MC1R variants increase risk of melanomas harboring BRAF mutations. J. Invest. Dermatol. *128*, 2485–2490.

Fischer, J., Weide, T., and Barnekow, A. (2005). The MICAL proteins and rab1: a possible link to the cytoskeleton? Biochem. Biophys. Res. Commun. *328*, 415–423.

Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J., et al. (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature *436*, 117–122.

Gissen, P., Johnson, C.A., Gentle, D., Hurst, L.D., Doherty, A.J., O'Kane, C.J., Kelly, D.A., and Maher, E.R. (2005). Comparative evolutionary analysis of VPS33 homologues: genetic and functional insights. Hum. Mol. Genet. *14*, 1261–1270.

Goedert, M., Cuenda, A., Craxton, M., Jakes, R., and Cohen, P. (1997). Activation of the novel stress-activated protein kinase SAPK4 by cytokines and cellular stresses is mediated by SKK3 (MKK6); comparison of its substrate specificity with that of other SAP kinases. EMBO J. *16*, 3563–3571.

Graves, A.R., Curran, P.K., Smith, C.L., and Mindell, J.A. (2008). The Cl^-/H^+ antiporter ClC-7 is the primary chloride permeation pathway in lysosomes. Nature *453*, 788–792.

Grottke, C., Mantwill, K., Dietel, M., Schadendorf, D., and Lage, H. (2000). Identification of differentially expressed genes in human melanoma cells with acquired resistance to various antineoplastic drugs. Int. J. Cancer *88*, 535–546.

Hamza, I., Faisst, A., Prohaska, J., Chen, J., Gruss, P., and Gitlin, J.D. (2001). The metallochaperone Atox1 plays a critical role in perinatal copper homeostasis. Proc. Natl. Acad. Sci. USA *98*, 6848–6852.

Hamza, I., Prohaska, J., and Gitlin, J.D. (2003). Essential role for Atox1 in the copper-mediated intracellular trafficking of the Menkes ATPase. Proc. Natl. Acad. Sci. USA *100*, 1215–1220.

Hanein, S., Martin, E., Boukhris, A., Byrne, P., Goizet, C., Hamri, A., Benomar, A., Lossos, A., Denora, P., Fernandez, J., et al. (2008). Identification of the SPG15 gene, encoding spastizin, as a frequent cause of complicated autosomal-recessive spastic paraplegia, including Kjellin syndrome. Am. J. Hum. Genet. *82*, 992–1002.

Hoek, K.S., Schlegel, N.C., Brafford, P., Sucker, A., Ugurel, S., Kumar, R., Weber, B.L., Nathanson, K.L., Phillips, D.J., Herlyn, M., et al. (2006). Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. Pigment Cell Res. *19*, 290–302.

Hoek, K.S., Schlegel, N.C., Eichhoff, O.M., Widmer, D.S., Praetorius, C., Einarsson, S.O., Valgeirsdottir, S., Bergsteinsdottir, K., Schepsky, A., Dummer, R., and Steingrimsson, E. (2008). Novel MITF targets identified using a two-step DNA microarray strategy. Pigment Cell Melanoma Res *21*, 665–676.

Homma, M.K., Wada, I., Suzuki, T., Yamaki, J., Krebs, E.G., and Homma, Y. (2005). CK2 phosphorylation of eukaryotic translation initiation factor 5 potentiates cell cycle progression. Proc. Natl. Acad. Sci. USA *102*, 15688–15693.

Houng, A., Polgar, J., and Reed, G.L. (2003). Munc18-syntaxin complexes and exocytosis in human platelets. J. Biol. Chem. *278*, 19627–19633.

Høyer-Hansen, M., Bastholm, L., Szyniarowski, P., Campanella, M., Szabadkai, G., Farkas, T., Bianchi, K., Fehrenbacher, N., Elling, F., Rizzuto, R., et al. (2007). Control of macroautophagy by calcium, calmodulin-dependent kinase kinase-beta, and Bcl-2. Mol. Cell *25*, 193–205.

Huh, S.J., Chen, Y.-L., Friedman, S.L., Liao, J., Huang, H.-J.S., Cavenee, W.K., and Robertson, G.P. (2010). KLF6 Gene and early melanoma development in a collagen I-rich extracellular environment. J. Natl. Cancer Inst. *102*, 1131–1147.

Hunziker, W., and Peters, P.J. (1998). Rab17 localizes to recycling endosomes and regulates receptor-mediated transcytosis in epithelial cells. J. Biol. Chem. *273*, 15734–15741.

Itoh, S., Kim, H.W., Nakagawa, O., Ozumi, K., Lessner, S.M., Aoki, H., Akram, K., McKinney, R.D., Ushio-Fukai, M., and Fukai, T. (2008). Novel role of antioxidant-1 (Atox1) as a copper-dependent transcription factor involved in cell proliferation. J. Biol. Chem. *283*, 9157–9167.

Jing, X.T., Wu, H.T., Wu, Y., Ma, X., Liu, S.H., Wu, Y.R., Ding, X.F., Peng, X.Z., Qiang, B.Q., Yuan, J.G., et al (2008). DIXDC1 Promotes Retinoic Acid-Induced Neuronal Differentiation and Inhibits Gliogenesis in (Cells. Cell Mol. Neurobiol.), p. 19.

Johansson, P., Pavey, S., and Hayward, N. (2007). Confirmation of a BRAF mutation-associated gene expression signature in melanoma. Pigment Cell Res. *20*, 216–221.

Jörgensen, P.M., Gräslund, S., Betz, R., Ståhl, S., Larsson, C., and Höög, C. (2001). Characterisation of the human APC1, the largest subunit of the anaphase-promoting complex. Gene *262*, 51–59.

Keats, J.J., Fonseca, R., Chesi, M., Schop, R., Baker, A., Chng, W.-J., Van Wier, S., Tiedemann, R., Shi, C.-X., Sebag, M., et al. (2007). Promiscuous mutations activate the noncanonical NF-kappaB pathway in multiple myeloma. Cancer Cell *12*, 131–144.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kizil, C., Otto, G.W., Geisler, R., Nüsslein-Volhard, C., and Antos, C.L. (2009). Simplet controls cell proliferation and gene transcription during zebrafish caudal fin regeneration. Dev. Biol. *325*, 329–340.

Knebel, A., Haydon, C.E., Morrice, N., and Cohen, P. (2002). Stress-induced regulation of eukaryotic elongation factor 2 kinase by SB 203580-sensitive and -insensitive pathways. Biochem. J. *367*, 525–532.

Knecht, W., Cottrell, G.S., Amadesi, S., Mohlin, J., Skåregärde, A., Gedda, K., Peterson, A., Chapman, K., Hollenberg, M.D., Vergnolle, N., and Bunnett, N.W. (2007). Trypsin IV or mesotrypsin and p23 cleave protease-activated receptors 1 and 2 to induce inflammation and hyperalgesia. J. Biol. Chem. *282*, 26089–26100.

Lee, S.-I., Pe'er, D., Dudley, A.M., Church, G.M., and Koller, D. (2006). Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc. Natl. Acad. Sci. USA *103*, 14062–14067.

Levy, C., Khaled, M., and Fisher, D.E. (2006). MITF: master regulator of melanocyte development and melanoma oncogene. Trends Mol. Med. *12*, 406–414.

Li, Y., Li, L., Brown, T.J., and Heldin, P. (2007). Silencing of hyaluronan synthase 2 suppresses the malignant phenotype of invasive breast cancer cells. Int. J. Cancer *120*, 2557–2567.

Lin, S., Wang, J., Ye, Z., Ip, N.Y., and Lin, S.C. (2008a). CDK5 activator p35 downregulates E-cadherin precursor independently of CDK5. FEBS Lett. *582*, 1197–1202.

Lin, W.M., Baker, A.C., Beroukhim, R., Winckler, W., Feng, W., Marmion, J.M., Laine, E., Greulich, H., Tseng, H., Gates, C., et al. (2008b). Modeling genomic diversity and tumor dependency in malignant melanoma. Cancer Res. *68*, 664–673.

Ludwig, R.J., Zollner, T.M., Santoso, S., Hardt, K., Gille, J., Baatz, H., Johann, P.S., Pfeffer, J., Radeke, H.H., Schön, M.P., et al. (2005). Junctional adhesion molecules (JAM)-B and -C contribute to leukocyte extravasation to the skin and mediate cutaneous inflammation. J. Invest. Dermatol. *125*, 969–976.

McGill, G.G., Horstmann, M., Widlund, H.R., Du, J., Motyckova, G., Nishimura, E.K., Lin, Y.-L., Ramaswamy, S., Avery, W., Ding, H.-F., et al. (2002). Bcl2 regulation by the melanocyte master regulator Mitf modulates lineage survival and melanoma cell viability. Cell *109*, 707–718.

Miccoli, L., Frouin, I., Novac, O., Di Paola, D., Harper, F., Zannis-Hadjopoulos, M., Maga, G., Biard, D.S., and Angulo, J.F. (2005). The human stress-activated protein kin17 belongs to the multiprotein DNA replication complex and associates in vivo with mammalian replication origins. Mol. Cell. Biol. *25*, 3814–3830.

Nan, H., Kraft, P., Hunter, D.J., and Han, J. (2009). Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians. Int. J. Cancer *125*, 909–917.

Narla, G., Heath, K.E., Reeves, H.L., Li, D., Giono, L.E., Kimmelman, A.C., Glucksman, M.J., Narla, J., Eng, F.J., Chan, A.M., et al. (2001). KLF6, a candidate tumor suppressor gene mutated in prostate cancer. Science *294*, 2563–2566.

Neubrand, V.E., Will, R.D., Möbius, W., Poustka, A., Wiemann, S., Schu, P., Dotti, C.G., Pepperkok, R., and Simpson, J.C. (2005). Gamma-BAR, a novel AP-1-interacting protein involved in post-Golgi trafficking. EMBO J. *24*, 1122–1133.

Nielsen, M.S., Madsen, P., Christensen, E.I., Nykjaer, A., Gliemann, J., Kasper, D., Pohlmann, R., and Petersen, C.M. (2001). The sortilin cytoplasmic tail conveys Golgi-endosome transport and binds the VHS domain of the GGA2 sorting protein. EMBO J. *20*, 2180–2190.

Nomiyama, T., Nakamachi, T., Gizard, F., Heywood, E.B., Jones, K.L., Ohkura, N., Kawamori, R., Conneely, O.M., and Bruemmer, D. (2006). The NR4A orphan nuclear receptor NOR1 is induced by platelet-derived growth factor and mediates vascular smooth muscle cell proliferation. J. Biol. Chem. *281*, 33467–33476.

Nomura, T., Huang, W.-C., Seo, S., Zhau, H.E., Mimata, H., and Chung, L.W.K. (2007). Targeting beta2-microglobulin mediated signaling as a novel therapeutic approach for human renal cell carcinoma. J. Urol. *178*, 292–300.

Osborn, L., Hession, C., Tizard, R., Vassallo, C., Luhowskyj, S., Chi-Rosso, G., and Lobb, R. (1989). Direct expression cloning of vascular cell adhesion molecule 1, a cytokine-induced endothelial protein that binds to lymphocytes. Cell *59*, 1203–1211.

Palmer, C.D., Mutch, B.E., Workman, S., McDaid, J.P., Horwood, N.J., and Foxwell, B.M. (2008). Bmx tyrosine kinase regulates TLR4-induced IL-6 production in human macrophages independently of p38 MAPK and NFkappB activity. Blood *111*, 1781–1788.

Pankiv, S., Alemu, E.A., Brech, A., Bruun, J.A., Lamark, T., Overvatn, A., Bjørkøy, G., and Johansen, T. (2010). FYCO1 is a Rab7 effector that binds to LC3 and PI3P to mediate microtubule plus end-directed vesicle transport. J. Cell Biol. *188*, 253–269.

Piaggi, S., Raggi, C., Corti, A., Pitzalis, E., Maschera, M.C., Saviozzi, M., Pompella, A., and Casini, A. (2010). Glutathione transferase omega 1-1 (GSTO1-1) plays an anti-apoptotic role in cell resistance to cisplatin toxicity. Carcinogenesis *31*, 804–811.

Pratilas, C.A., Taylor, B.S., Ye, Q., Viale, A., Sander, C., Solit, D.B., and Rosen, N. (2009). (V600E)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. Proc. Natl. Acad. Sci. USA *106*, 4519–4524.

Rae, J.M., Johnson, M.D., Cordero, K.E., Scheys, J.O., Larios, J.M., Gottardis, M.M., Pienta, K.J., and Lippman, M.E. (2006). GREB1 is a novel androgen-regulated gene required for prostate cancer growth. Prostate *66*, 886–894.

Rae, J.M., Johnson, M.D., Scheys, J.O., Cordero, K.E., Larios, J.M., and Lippman, M.E. (2005). GREB 1 is a critical regulator of hormone dependent breast cancer growth. Breast Cancer Res. Treat. *92*, 141–149.

Rangel, J., Nosrati, M., Leong, S.P., Haqq, C., Miller, J.R., III, Sagebiel, R.W., and Kashani-Sabet, M. (2008). Novel role for RGS1 in melanoma progression. Am. J. Surg. Pathol. *32*, 1207–1212.

Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C.Y., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D., and Daly, M.J.; International Schizophrenia Consortium. (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. *5*, e1000534.

Roma, C., Ferrante, P., Guardiola, O., Ballabio, A., and Zollo, M. (2007). New mutations identified in the ocular albinism type 1 gene. Gene *402*, 20–27.

Rosales, J.L., Ernst, J.D., Hallows, J., and Lee, K.Y. (2004). GTP-dependent secretion from neutrophils is regulated by Cdk5. J. Biol. Chem. *279*, 53932–53936.

Rosengren Pielberg, G., Golovko, A., Sundström, E., Curik, I., Lennartsson, J., Seltenhammer, M.H., Druml, T., Binns, M., Fitzsimmons, C., Lindgren, G., et al. (2008). A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. Nat. Genet. *40*, 1004–1009.

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. *24*, 513–523.

Sanchez-Garcia, F., Akavia, U.D., Mozes, E., and Pe'er, D. (2010). JISTIC: identification of significant targets in cancer. BMC Bioinformatics *11*, 189.

Schallreuter, K.U., Rübsam, K., Chavan, B., Zothner, C., Gillbro, J.M., Spencer, J.D., and Wood, J.M. (2006). Functioning methionine sulfoxide reductases A and B are present in human epidermal melanocytes in the cytosol and in the nucleus. Biochem. Biophys. Res. Commun. *342*, 145–152.

Scott, G.A., McClelland, L.A., Fricke, A.F., and Fender, A. (2009). Plexin C1, a receptor for semaphorin 7a, inactivates cofilin and is a potential tumor suppressor for melanoma progression. J. Invest. Dermatol. *129*, 954–963.

Segal, E., Pe'er, D., Regev, A., Koller, D., and Friedman, N. (2005). Learning Module Networks. JMLR *6*, 557–588.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Seye, C.I., Yu, N., Jain, R., Kong, Q., Minor, T., Newton, J., Erb, L., González, F.A., and Weisman, G.A. (2003). The P2Y2 nucleotide receptor mediates UTP-induced vascular cell adhesion molecule-1 expression in coronary artery endothelial cells. J. Biol. Chem. *278*, 24960–24965.

Shen, F., Hu, Z., Goswami, J., and Gaffen, S.L. (2006). Identification of common transcriptional regulatory elements in interleukin-17 target genes. J. Biol. Chem. *281*, 24138–24148.

Shiba, T., Takatsu, H., Nogi, T., Matsugaki, N., Kawasaki, M., Igarashi, N., Suzuki, M., Kato, R., Earnest, T., Nakayama, K., and Wakatsuki, S. (2002). Structural basis for recognition of acidic-cluster dileucine sequence by GGA1. Nature *415*, 937–941.

Smith, A.N., Lovering, R.C., Futai, M., Takeda, J., Brown, D., and Karet, F.E. (2003). Revised nomenclature for mammalian vacuolar-type H^+ -ATPase subunit genes. Mol. Cell *12*, 801–803.

Steingrímsson, E., Copeland, N.G., and Jenkins, N.A. (2004). Melanocytes and the microphthalmia transcription factor network. Annu. Rev. Genet. *38*, 365–411.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Sugiura, T., Noguchi, Y., Sakurai, K., and Hattori, C. (2008). Protein phosphatase 1H, overexpressed in colon adenocarcinoma, is associated with CSE1L. Cancer Biol. Ther. *7*, 285–292.

Suzuki, T., Li, W., Zhang, Q., Karim, A., Novak, E.K., Sviderskaya, E.V., Hill, S.P., Bennett, D.C., Levin, A.V., Nieuwenhuis, H.K., et al. (2002). Hermansky-Pudlak syndrome is caused by mutations in HPS4, the human homolog of the mouse light-ear gene. Nat. Genet. *30*, 321–324.

Thaker, N.G., Zhang, F., McDonald, P.R., Shun, T.Y., Lewen, M.D., Pollack, I.F., and Lazo, J.S. (2009). Identification of survival genes in human glioblastoma cells by small interfering RNA screening. Mol. Pharmacol. *76*, 1246–1255.

Trevisani, M., Siemens, J., Materazzi, S., Bautista, D.M., Nassini, R., Campi, B., Imamachi, N., Andrè, E., Patacchini, R., Cottrell, G.S., et al. (2007). 4-Hydroxynonenal, an endogenous aldehyde, causes pain and neurogenic inflammation through activation of the irritant receptor TRPA1. Proc. Natl. Acad. Sci. USA *104*, 13519–13524.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA *98*, 5116–5121.

Van Rijsbergen, C.J., Robertson, S.E., and Porter, M.F. (1980). New models in probabilistic information retrieval (London: British Library Research and Development Department).

Vetrini, F., Auricchio, A., Du, J., Angeletti, B., Fisher, D.E., Ballabio, A., and Marigo, V. (2004). The microphthalmia transcription factor (Mitf) controls expression of the ocular albinism type 1 gene: link between melanin synthesis and melanosome biogenesis. Mol. Cell. Biol. *24*, 6550–6559.

Waghray, A., Keppler, D., Sloane, B.F., Schuger, L., and Chen, Y.Q. (2002). Analysis of a truncated form of cathepsin H in human prostate tumor cells. J. Biol. Chem. *277*, 11533–11538.

Walter, M.J., Payton, J.E., Ries, R.E., Shannon, W.D., Deshmukh, H., Zhao, Y., Baty, J., Heath, S., Westervelt, P., Watson, M.A., et al. (2009). Acquired copy number alterations in adult acute myeloid leukemia genomes. Proc. Natl. Acad. Sci. USA *106*, 12950–12955.

Wang, J.-S., Wang, F.-B., Zhang, Q.-G., Shen, Z.-Z., and Shao, Z.-M. (2008). Enhanced expression of Rab27A gene by breast cancer cells promoting invasiveness and the metastasis potential by secretion of insulin-like growth factor-II. Mol. Cancer Res. *6*, 372–382.

Wang, L., Cao, X.X., Chen, Q., Zhu, T.F., Zhu, H.G., and Zheng, L. (2009). DIXDC1 targets p21 and cyclin D1 via PI3K pathway activation to promote colon cancer cell proliferation. Cancer Sci. *100*, 1801–1808.

Wee, S., Wiederschain, D., Maira, S.-M., Loo, A., Miller, C., deBeaumont, R., Stegmeier, F., Yao, Y.-M., and Lengauer, C. (2008). PTEN-deficient cancers depend on PIK3CB. Proc. Natl. Acad. Sci. USA *105*, 13057–13062.

Weidle, U.H., Evtimova, V., Alberti, S., Guerra, E., Fersis, N., and Kaul, S. (2009). Cell growth stimulation by CRASH, an asparaginase-like protein overexpressed in human tumors and metastatic breast cancers. Anticancer Res. *29*, 951–963.

Welch, B.L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika *29*, 350–362.

Witten, D.M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics *10*, 515–534.

Yang, R.Z., Lee, M.J., Hu, H., Pollin, T.I., Ryan, A.S., Nicklas, B.J., Snitker, S., Horenstein, R.B., Hull, K., Goldberg, N.H., et al. (2006). Acute-phase serum amyloid A: an inflammatory adipokine and potential link between obesity and its metabolic complications. PLoS Med. *3*, e287.

Yoshida, H., Jono, H., Kai, H., and Li, J.D. (2005). The tumor suppressor cylindromatosis (CYLD) acts as a negative regulator for toll-like receptor 2 signaling via negative cross-talk with TRAF6 AND TRAF7. J. Biol. Chem. *280*, 41111–41121.

Zhang, R., Xu, Y., Ekman, N., Wu, Z., Wu, J., Alitalo, K., and Min, W. (2003). Etk/Bmx transactivates vascular endothelial growth factor 2 and recruits phosphatidylinositol 3-kinase to mediate the tumor necrosis factor-induced angiogenic pathway. J. Biol. Chem. *278*, 51267–51276.

Zhu, S., Wurdak, H., Wang, Y., Galkin, A., Tao, H., Li, J., Lyssiotis, C.A., Yan, F., Tu, B.P., Miraglia, L., et al. (2009). A genomic screen identifies TYRO3 as a MITF regulator in melanoma. Proc. Natl. Acad. Sci. USA *106*, 17025–17030.

Zou, L., Zhou, J., Zhang, J., Li, J., Liu, N., Chai, L., Li, N., Liu, T., Li, L., Xie, Z., et al. (2009). The GTPase Rab3b/3c-positive recycling vesicles are involved in cross-presentation in dendritic cells. Proc. Natl. Acad. Sci. USA *106*, 15801–15806.

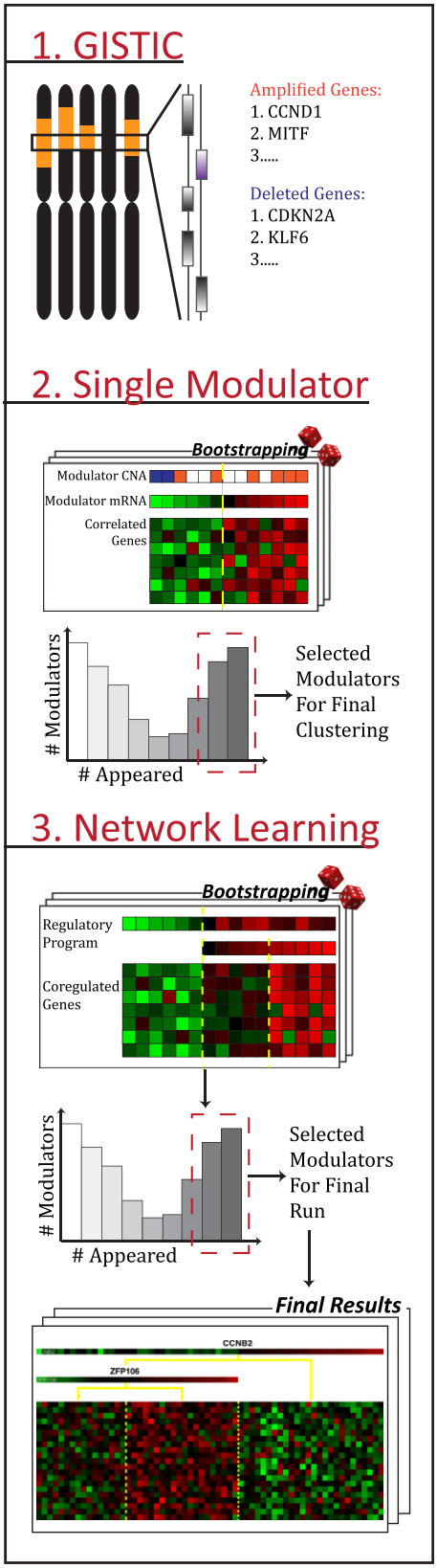
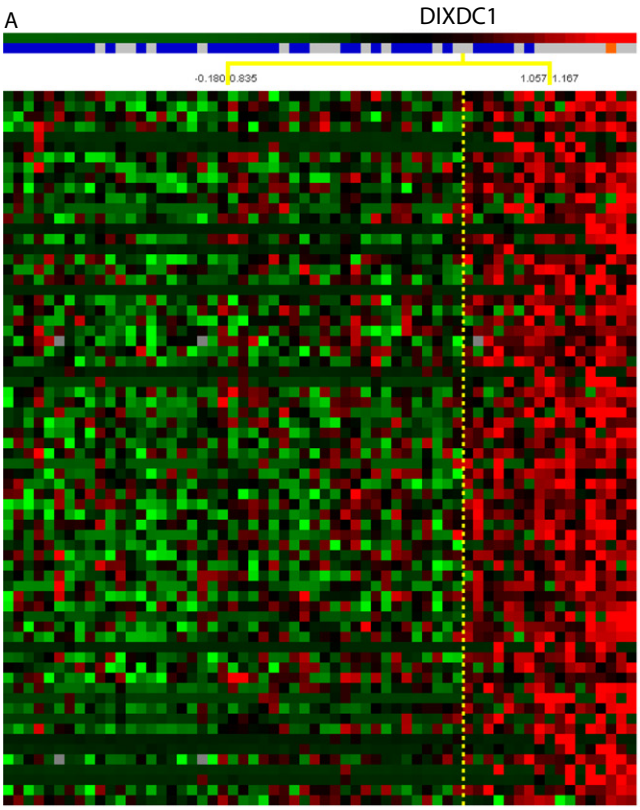

1. GISTIC
Amplified Genes:
1. CCND1
2. MITF
3.....
Deleted Genes:
1. CDKN2A
2. KLF6
3.....
2. Single Modulator
Bootstrapping
Modulator CNA
Modulator mRNA
Correlated Genes
Modulators
Appeared
Selected Modulators For Final Clustering
3. Network Learning
Bootstrapping
Regulatory Program
Coregulated Genes
Modulators
Appeared
Selected Modulators For Final Run
Final Results
CCNB2
ZFP106

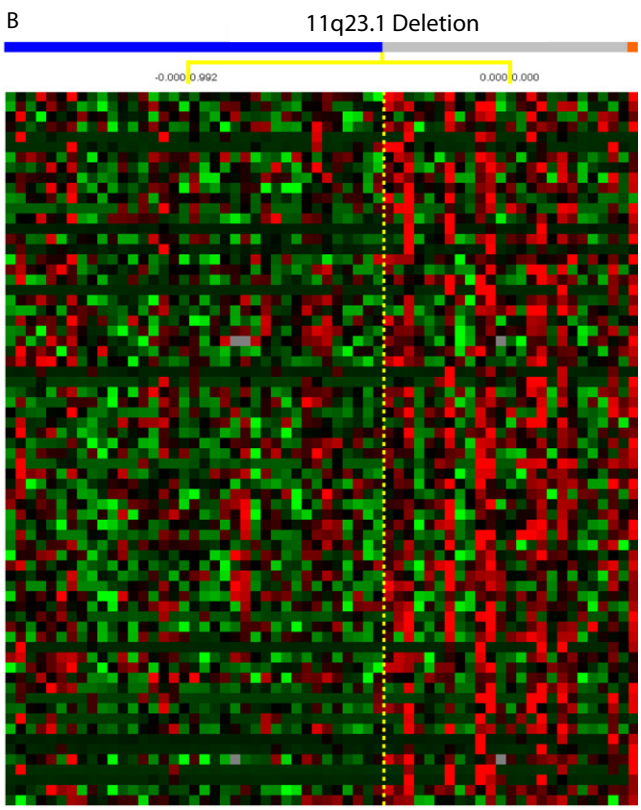
Figure S1. Overview of the CONEXIC Learning Algorithm, Related to Figure 1

(1) Selection of candidate driver genes (modulators). The same chromosome is represented in different tumors and orange represents amplified regions. The box shows a region amplified in multiple tumors, considered significantly amplified. Multiple genes reside in this region, represented as shaded boxes along two strands of DNA. Only the purple box represents a driver gene, whereas the gray boxes represent passenger genes and these are indistinguishable based on copy number alone. All genes located in significantly amplified or deleted regions are selected as candidate driver genes. (2) Single Modulator step. Modules of genes are each associated with the best possible candidate driver, based on gene expression of the gene and the candidate driver. The heat-map represents a good association where both copy number of the modulator influences the expression of the modulator and the expression of the modulator corresponds to the expression of the genes in the module. Random re-sampling with replacement is used to generate perturbations on the initial dataset and this step is repeated across datasets (bootstrapping). The histogram represents the number of runs (datasets) in which each selected modulator appears, the final run is performed with candidate drivers chosen for 90% of the runs (dotted red box). (3) Network Learning step. Using the set of Single Modulator of modules as a starting point, the algorithm refines the selected modulators and modules, now allowing for more than one modulator associated with each module. Bootstrapping is used similarly to the Single Modulator step, with the distinction being that modulators must be selected in 40% of the runs.

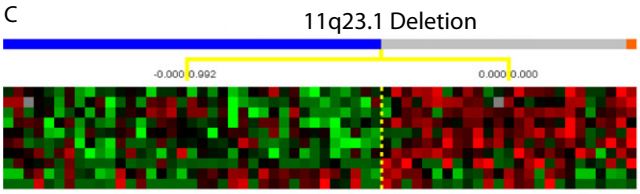
A



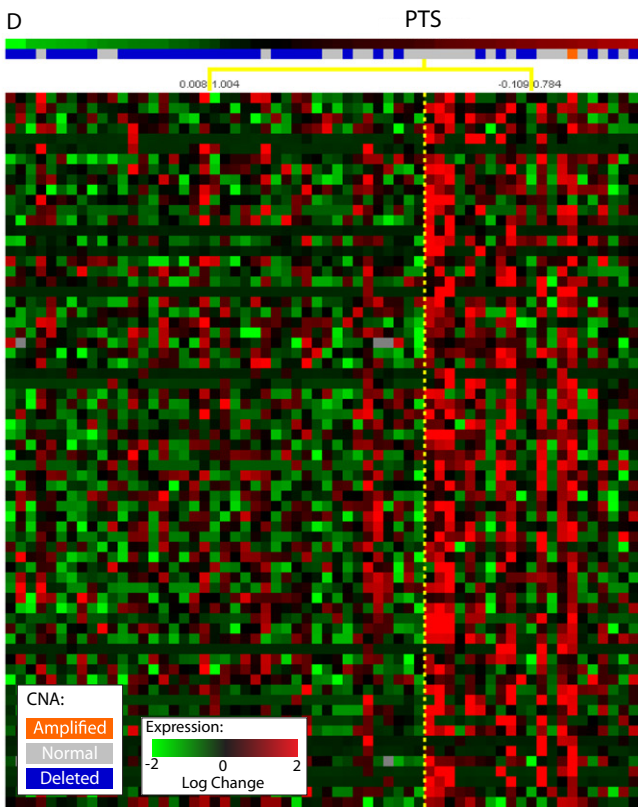
B



C



D



E

Regulator	NormalGamma	P-value
11q23.1 Deletion	-5,951.6	0.2888
PTS	-5,769.4	0.0013
DIXDC1	-5,307.4	0

Figure S2. Pinpointing DIXDC1, Related to Figure 3

The figure demonstrates the dense correlation structure of the data and clarifies why we associate each gene only with a single module.

(A) A *DIXDC1* associated module, *DIXDC1* expression correlates with expression of the genes in the associated module.

(B) The rows represent the same genes, in the same order as in A, but here the tumor samples are split according to copy number (deletion) of the region containing *DIXDC1*. The split still partitions the samples into two distinctly behaving groups, but not as well as *DIXDC1* expression.

(C) Data for deleted region on chromosome 11, representing expression for 10/17 genes that passed initial expression filtering. Samples are ordered according to deletion status of the region and genes are correlated with the region.

(D) The rows represent the same genes, in the same order as in A, but here the tumor samples are split according to the expression of *PTS*. The split in *PTS* expression is an improvement on the split defined by deletion status and looks compelling on its own. The split becomes inferior only in comparison to *DIXDC1* (panel A).

(E) The scores and p-values of each of the splits in panels A, B and D. Both the splits for *DIXDC1* and *PTS* are better scoring than the split on DNA alone and both are significant under permutation testing. If genes were allowed more than one modulator, most genes in the module would associate with *DIXDC1*, *PTS* and other genes in the chromosome 11 region. Selecting only one gene identifies *DIXDC1*, the highest scoring split and a gene known to be involved in WNT signaling, cell cycle and cancer (Jing et al., 2008; Wang et al., 2009).

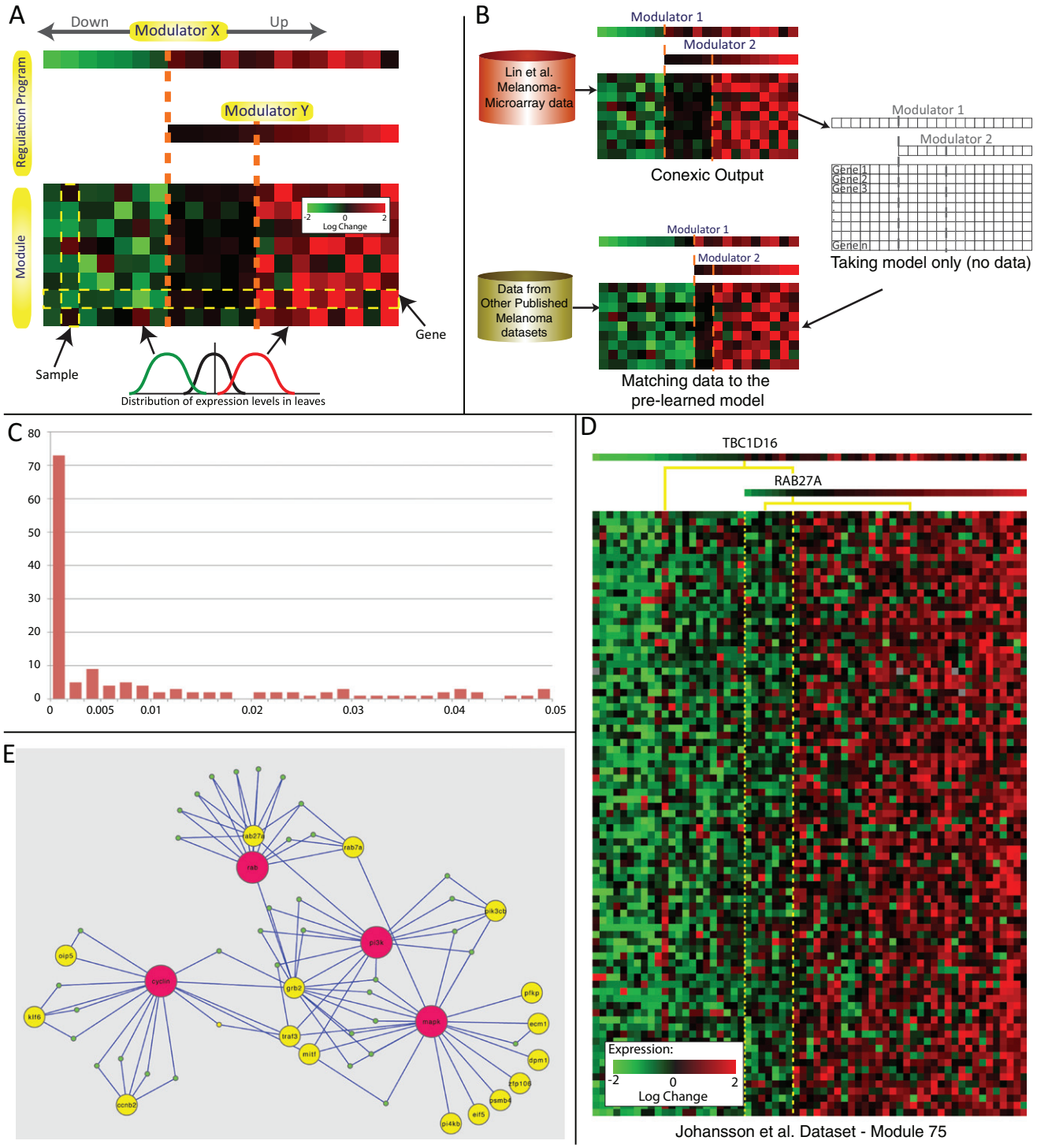


Figure S3. The Network Structure Is Verified in Other Data Sets, Related to Figure 2

(A) A module describes the expression of a set of genes, based on the conditional probability distribution ("regulation program"), represented as a regression tree (Segal et al., 2003). The expression profiles of all genes in the module are depicted, where the rows are genes, and the columns are tumors. Queries proceed from the root downwards, if the modulators expression is above the threshold, the queries proceed to the right and otherwise to the left of the split (dotted orange line). On the right side are groups of samples that follow the same branches down the tree, tumors that express modulators X and Y at a high level. Each leaf of the tree represents a Normal distribution.

(B) Evaluation of a network structure: CONEXIC learns all modules and their regulation programs based on data from Lin et al. (Lin et al., 2008b). The modulators,

their regulation programs and associated modules are kept constant. Two additional melanoma datasets were used to test how well the regulation program predicts the expression of the genes in the module for each tumor.

(C) A histogram of p-values (based on permutations) representing how well each modulator predicts gene expression in the test data for each pair of modulator and module selected by CONEXIC. The p-value for more than 50 modulators is close to zero. P-values are shown from the Johansson dataset, unless the modulator was missing in this dataset, and then the p-value from the Hoek dataset is shown.

(D) A module inferred from the Lin data and applied to the Johansson data, the module as it appears with the Lin data is presented in Figure 5A. The gene expression of the genes in the module matches that of the modulators.

(E) LitVAn was used to analyze literature trends in the 30 selected modulators appearing in Figure 2, the top four most significant terms: 'MAPK', 'PI3K', 'cyclin' and 'RAB'. The figure represents the graphic output of LitVAn, where significant terms (red circles) are associated (graph edge) with multiple genes in the query (yellow dots). An edge with a green dot represents a publication that significantly associates between the gene and the term in itself (typically through repeated use of the term throughout the paper) and the interactive version includes a link to the PubMed abstract. Only the top 5 most significant papers for each term-gene combination are represented (if any).

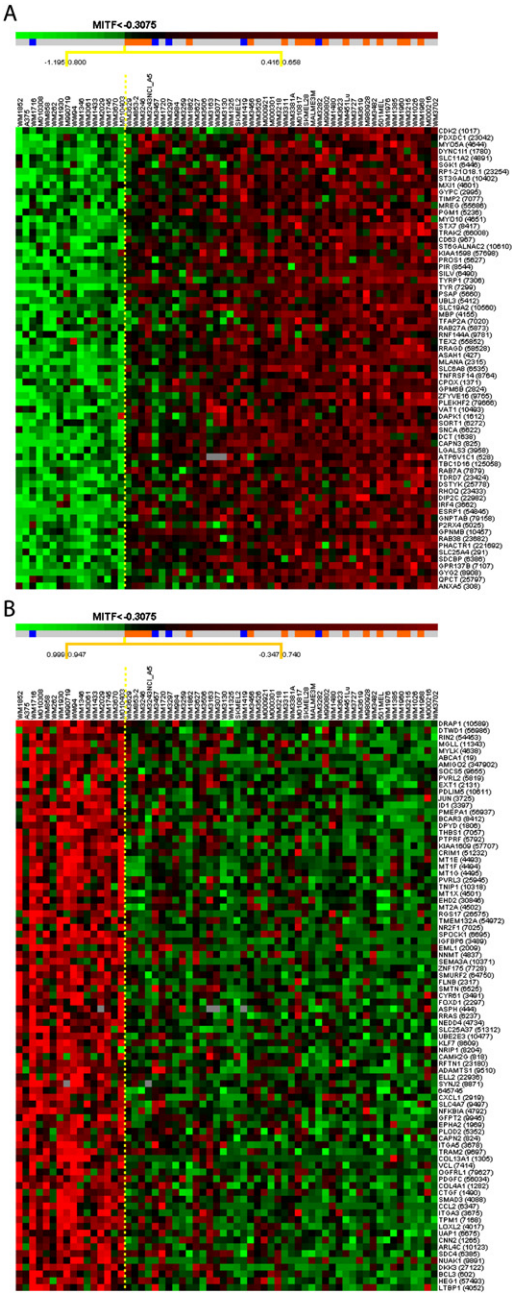


Figure S4. MITF-Associated Modules, Related to Figure 4

(A) *MITF* upregulated module. This presents the full module summarized in the upper part of Figure 4D. This module contains genes that were associated with *MITF* and correlate with its expression. The genes are listed with gene symbol and Entrez Gene ID and they are enriched for Vesicular Trafficking and Melanogenesis.

(B) *MITF* downregulated Module. This presents the full module summarized in the lower part of Figure 4D. This module contains genes that were associated with *MITF* and anti-correlate with its expression. The genes are listed with gene symbol and Entrez Gene ID and they are enriched for NFkB/TNF and Invasion/Migration.

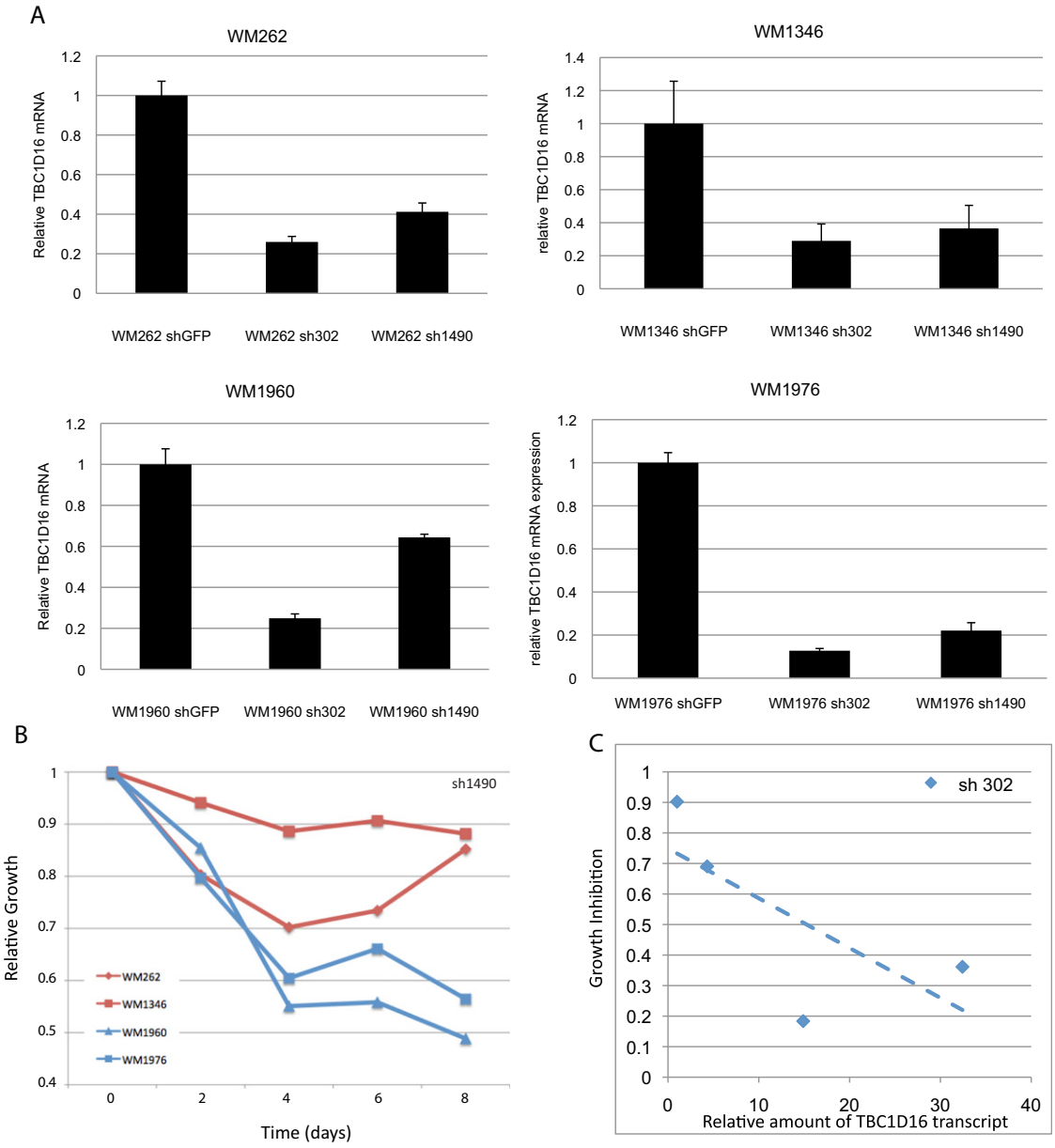


Figure S5. *TBC1D16* mRNA Knockdown and Growth Effects, Related to Figure 5

(A) RT-PCR levels of *TBC1D16* mRNA after shRNA knockdown for each of the 4 STCs tested. The reduction in the amount of the *TBC1D16* transcript was similar in all of the STCs, where each bar represents data from 3 biological replicates (see Table S4B).

(B) Representative growth curves for each of the 4 STCs tested with *TBC1D16* knockdown with the hairpin sh1490, each curve represents 3 technical replicates.

(C) Growth inhibition at 8 days is directly proportional to the amount of the *TBC1D16* transcript and is independent of *TBC1D16* copy number. Data averaged on 3 biological replicates X 3 technical replicates for each STC.

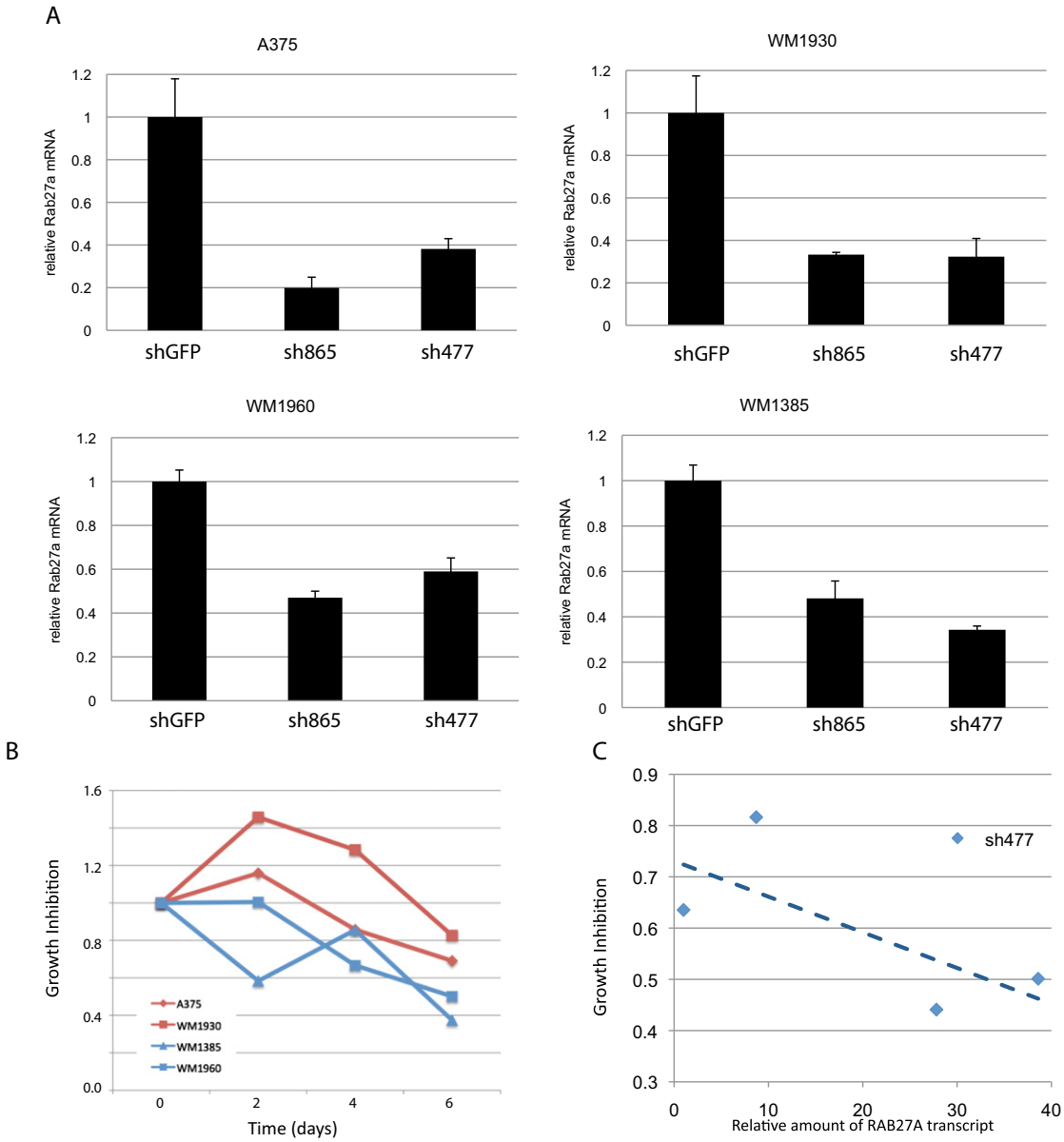


Figure S6. *RAB27A* mRNA Knockdown and Growth Effects, Related to Figure 6

(A) RT-PCR levels of *RAB27A* mRNA after shRNA knockdown for each of the 4 STCs tested. The reduction in the amount of the *RAB27A* transcript was similar in all of the STCs, where each bar represents data from 3 biological replicates (see Table S5).

(B) Representative growth curves for each of the 4 STCs tested with *RAB27A* knockdown with sh477, each curve represents 3 technical replicates.

(C) Growth inhibition at 6 days is dependent on the amount of the *RAB27A* transcript and is independent of *RAB27A* copy number. Data averaged over all replicates for each STC (methods).

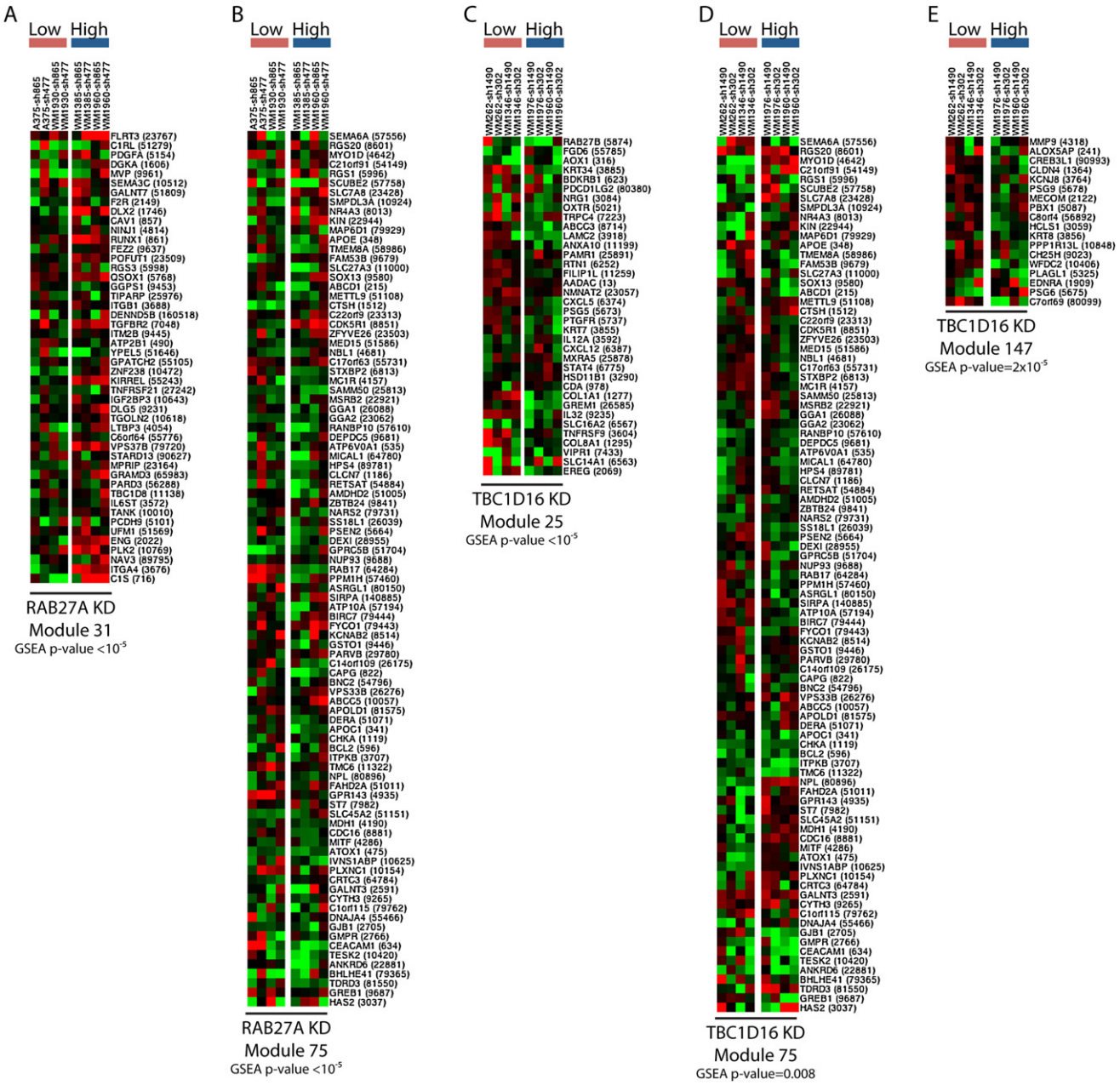


Figure S7. Microarray Results after Knockdown of *RAB27A* or *TBC1D16*, Related to Figure 7

(A) Data generated following knockdown (KD) of *RAB27A* is presented for module 31 (*RAB27A* associated). Heat-map represents Z-scores.

(B) Data generated following knockdown (KD) of *RAB27A* is presented for module 75 (*RAB27A* associated).

(C) Data generated following knockdown (KD) of *TBC1D16* is presented for module 25 (*TBC1D16* associated).

(D) Data generated following knockdown (KD) of *TBC1D16* is presented for module 75 (*TBC1D16* associated).

(E) Data generated following knockdown (KD) of *TBC1D16* is presented for module 147 (*TBC1D16* associated). Z-score data for all 5 modules demonstrates that the genes in each module are responding to perturbation in the predicted modulator. GSEA p-values were calculated using the median of 4 test samples and appear below each module.

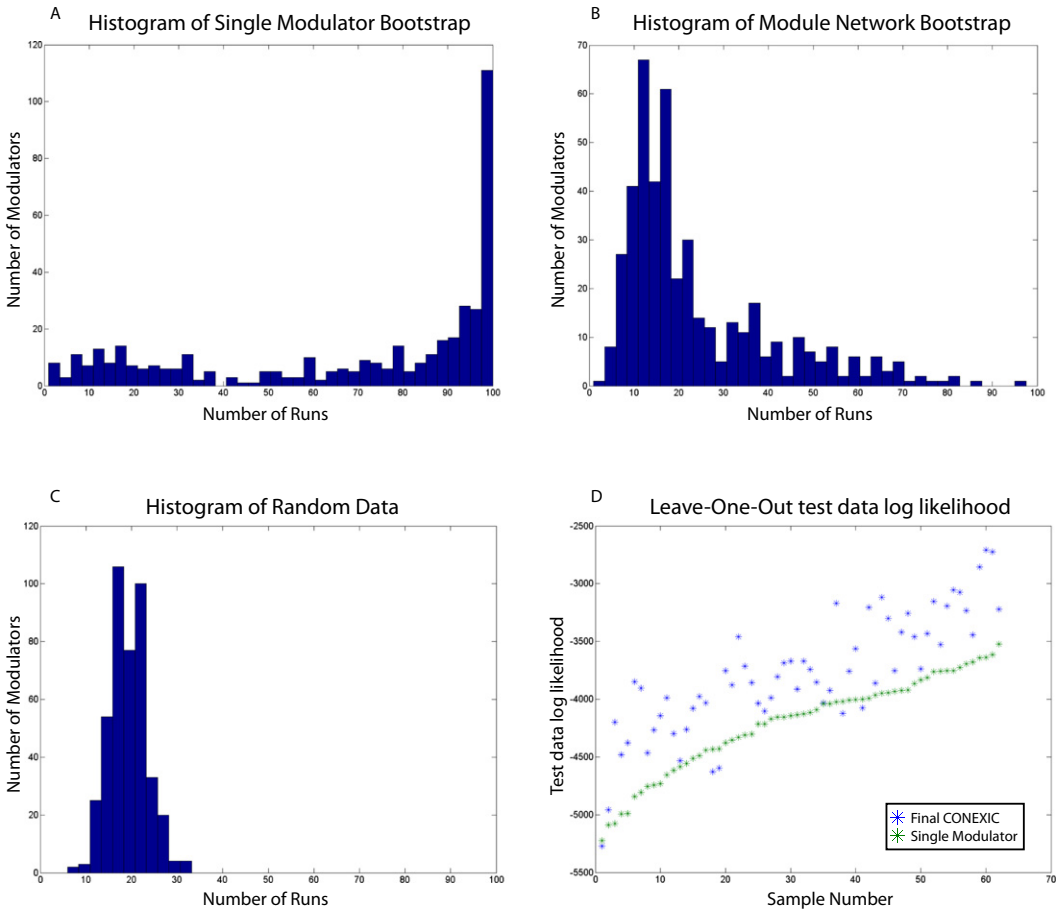


Figure S8. Robustness Analysis, Related to Experimental Procedures

(A) Histogram representing confidence values for bootstrap of Single Modulator.

(B) Histogram representing confidence values for bootstrap of Network Learning.

(C) Histogram representing confidence values for bootstrap of Network Learning, where the data has been randomly permuted. These values represent spurious modulators and none passed a confidence threshold of 35.

(D) Test likelihood for each sample in leave-one-out cross-validation. Each column is matched for the same sample, where green and blue stars represent test likelihood for single modulator and test likelihood for final CONEXIC model, respectively. The full model significantly outperforms Single Modulator on test data.