Cancer Genome Analysis (CONEXIC)

Akavia et al. Cell, 2010.

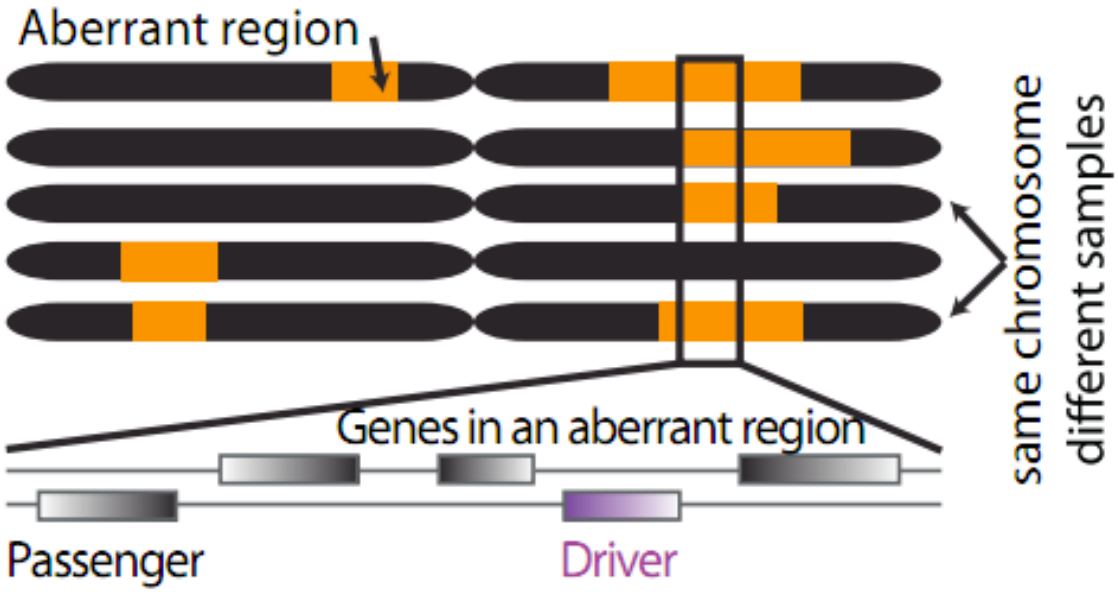
02-715 Advanced Topics in Computational Genomics

Integrated Approach for Discovering Drivers in Cancer

- Previous methods: find frequently occurring mutations
- Copy number variation in tumor samples can involve a large region containing multiple genes
  - Many are passengers: how to distinguish passenger and driver genes in the copy number variation region?
- Integrative analysis of copy number variations and gene expressions

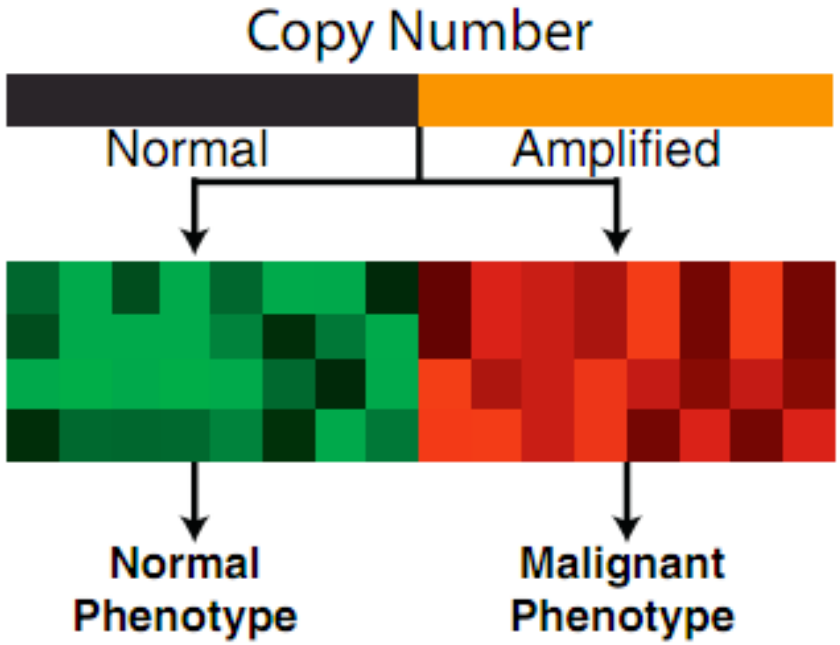
CONEXIC Modeling Assumption I

- A driver mutation should occur in multiple tumors more often than would be expected by chance



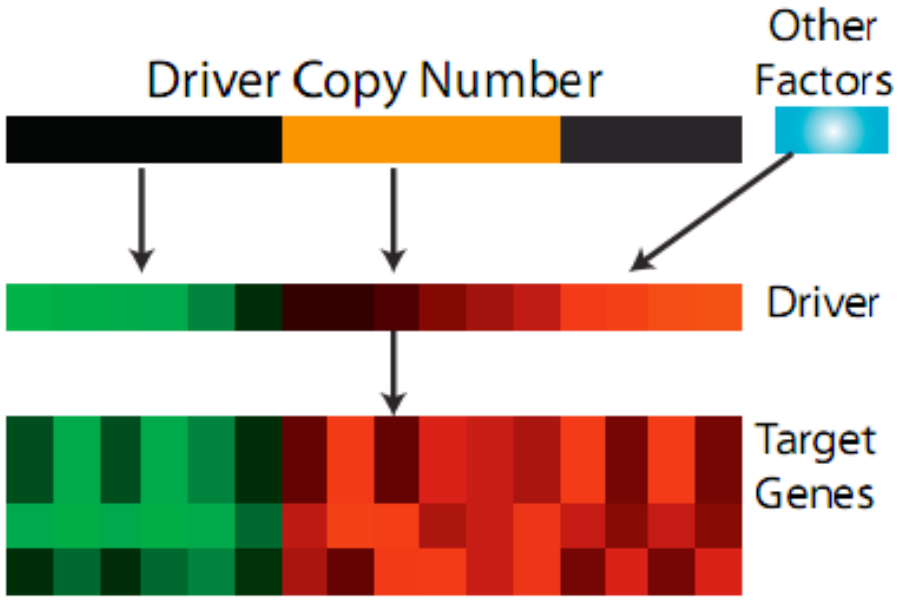
CONEXIC Modeling Assumption II

- A driver mutation may be associated (correlated) with the expression of a group of genes that form a module



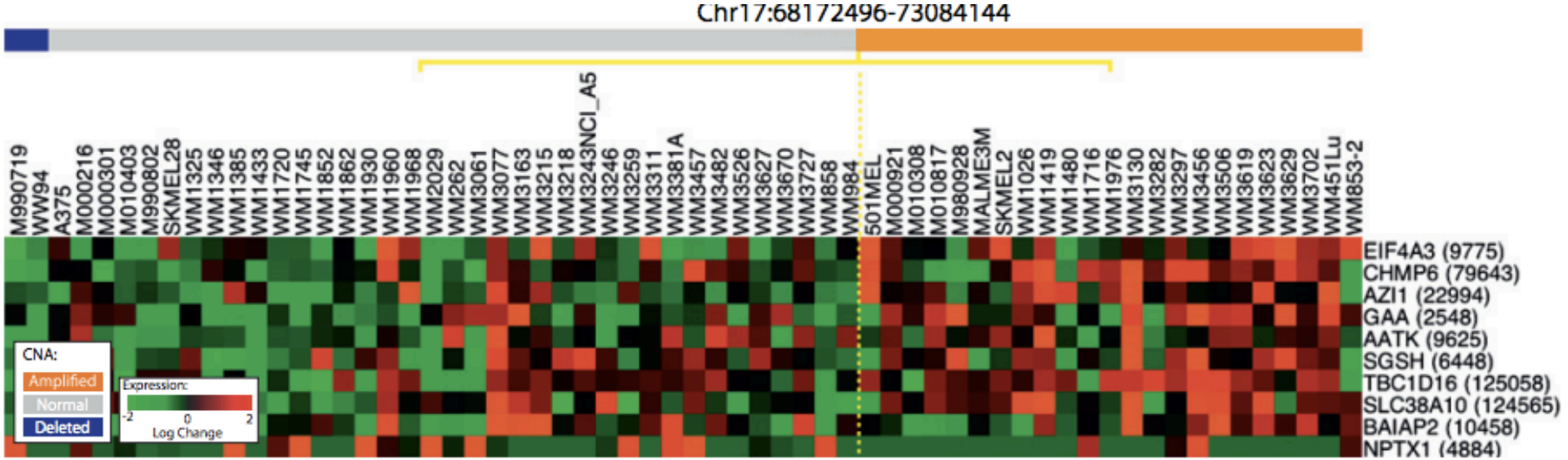
CONEXIC Modeling Assumption III

- A driver mutation may be associated (correlated) with the expression of a group of genes that form a module



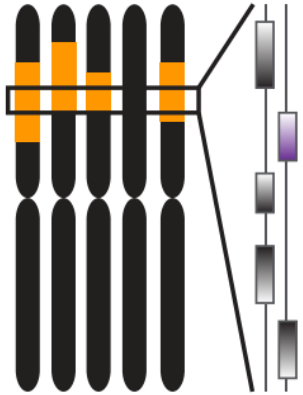
Copy Number Variation and Gene Expression

- Even among the individuals with amplification in copy numbers, the expression levels for those genes can differ.



CONEXIC Overview

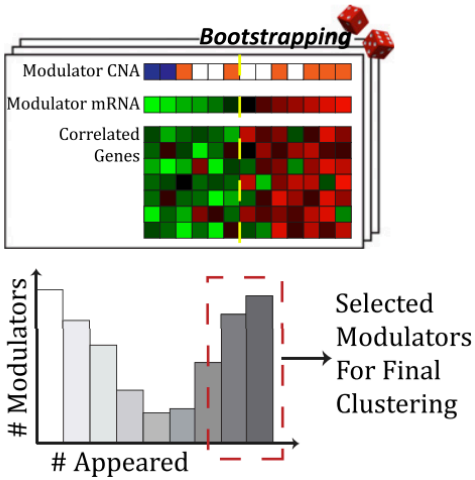
1. GISTIC



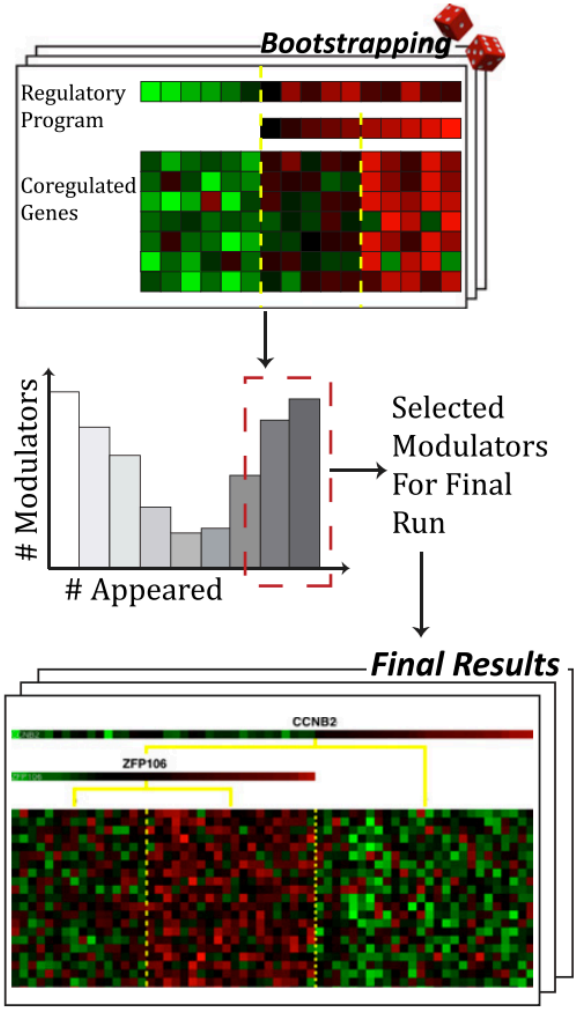
Amplified Genes:
1. CCND1
2. MITF
3.....

Deleted Genes:
1. CDKN2A
2. KLF6
3.....

2. Single Modulator



3. Network Learning



CONEXIC Overview

- Extends module networks to handle cancer copy number variation and gene expression data to find driver mutation
- Assumes a driver mutation affects “gene modules” rather than individual genes
- The gene expression as a proxy to distinguish between driver and passenger mutations in the large region of copy number variations

CONEXIC: Selecting Candidate Drivers

- Apply GISTIC to find frequently occurring regions of copy number variations

- Run CONEXIC with only the genes in the selected regions as candidate driver genes

GISTIC

- Detect frequently occurring CNV regions in cancer samples

- Typically CNV regions are large
  - Often involve the whole chromosome arms
  - Within the amplified region, there are small regions with peaks that often contain driver genes
  - How can we identify these relatively small regions?

GISTIC

- For each locus, compute G-scores

$$G^{AMP}(m) = \sum_i CN(m,i) \times I(CN(m,i) > \Theta^{AMP})$$

  - $CN(m,i)$: copy number measurement for sample i, marker m
  - Average copy number scores for marker m

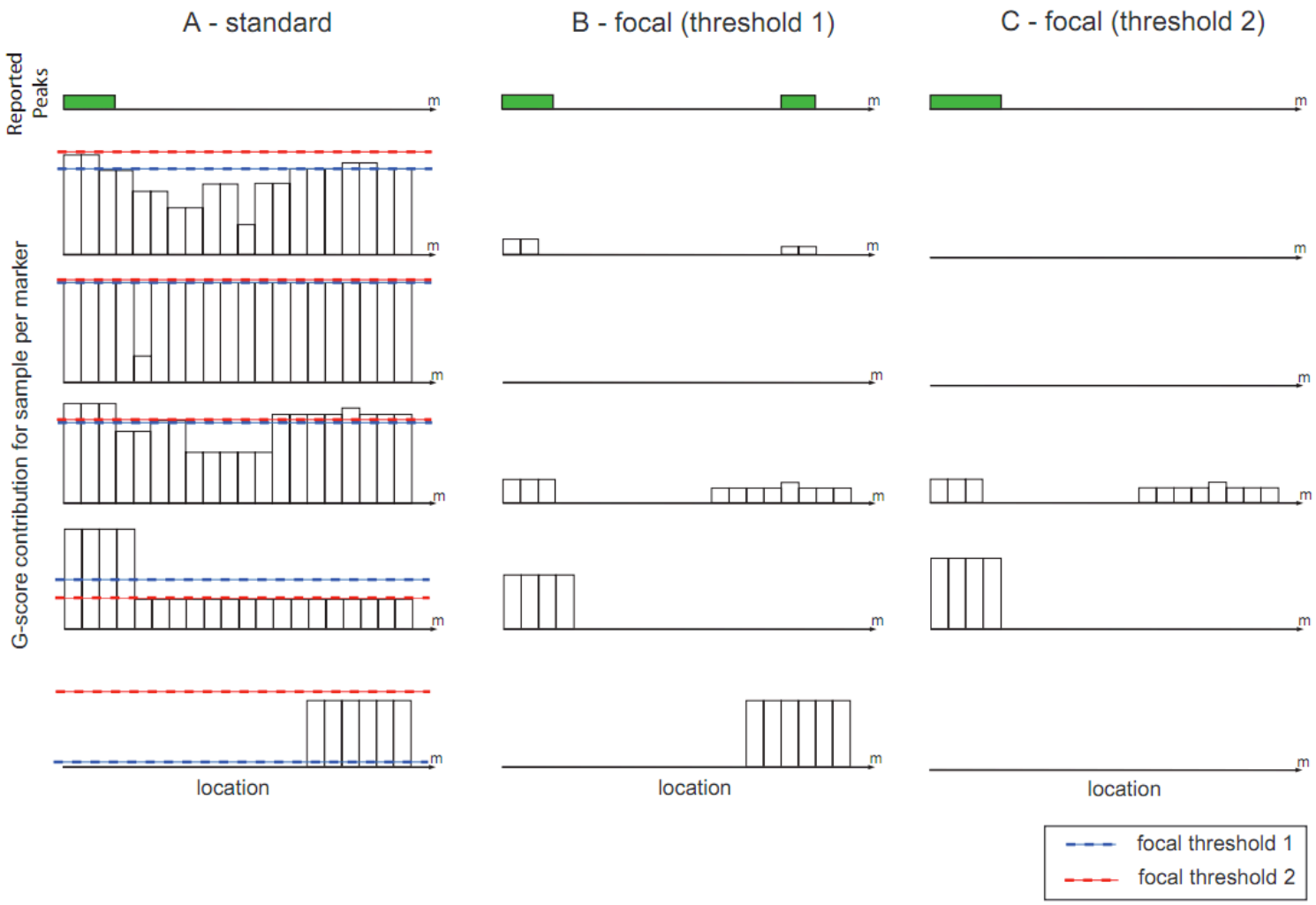
- Peel-off strategy: Find the peak within the large region of amplification and set the consecutive markers to zero before looking for the next peak in order to avoid the overlap
  - Detects only the highest peak in the large region

GISTIC

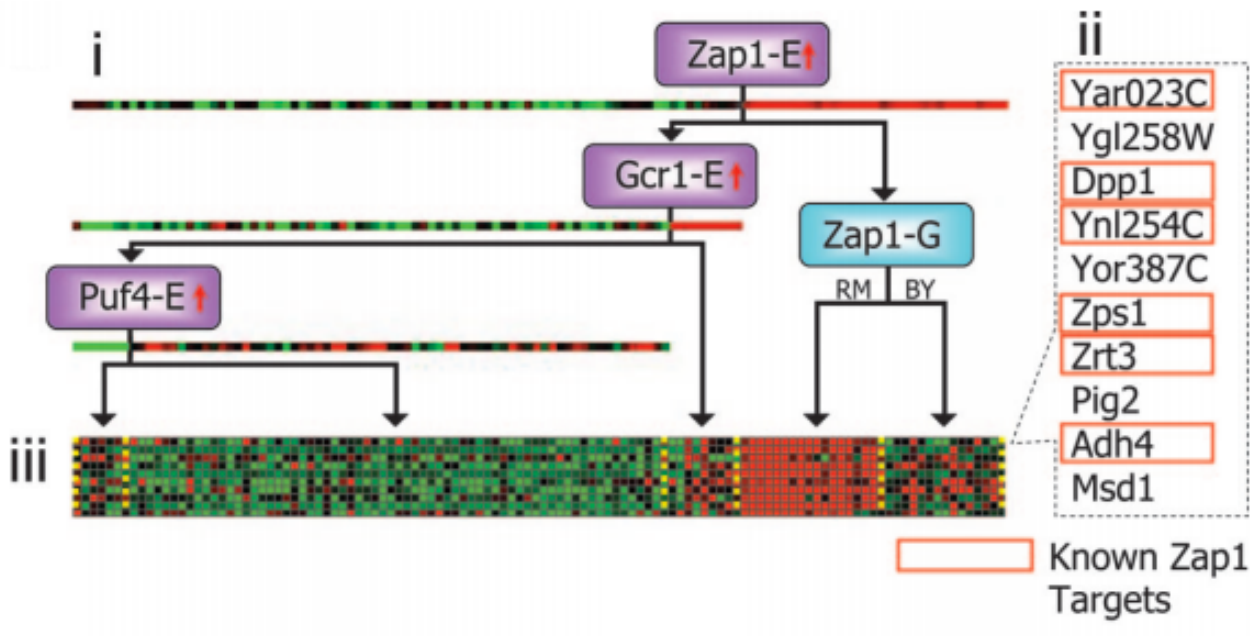
- How to distinguish a single large peak and multiple small peaks

- Extension of GISTIC for detecting focal copy number variations
  - During peel off, apply the threshold estimated from other regions of DNA
    - Different thresholds for different samples
    - Usually the highest broad aberration genome-wide as the threshold
    - Results are sensitive to the threshold value

GISTIC


A - standard
B - focal (threshold 1)
C - focal (threshold 2)
Reported Peaks
G-score contribution for sample per marker
m
location
focal threshold 1
focal threshold 2

Pre-cursor for CONEXIC (Lee et al., PNAS 2006)



CONEXIC: Single Modulator

- Step 1: 347 candidate drivers after applying GISTIC
- Step 2: Run K-means clustering on gene expression levels of candidate drivers to determine the expression threshold between normal and amplified/deleted samples
- Step 3: Determine target gene modules influenced by each candidate driver
  - Split the target gene expressions with respect to the threshold in Step 2
  - Assess the quality of split

CONEXIC Network-Learning Algorithm

- The single modulators as initialization

- Iterate between the two steps until fewer than 10% of the target genes have been re-assigned to a different module
  - Step 1: Learning the regulation program for each module
    - Construct a regression tree by splitting samples according to the drivers
    - Continue splitting until regression model fits the influence of modulator on the modules well at the leaf
  - Step 2: Re-assign each gene into the module that best models it behavior

LitVAn (Literature Vector Analysis)

- Literature-based analysis tool for inference of gene module functionality
  - enrichment analysis for gene modules
- NCBI database that associates each gene with manually curated papers (70,000 papers)
- Bag-of-words assumption

LitVAn

- TF*IDF score: score to words which are overrepresented in a subset of documents relative to the full corpus
  - Inverse Document Frequency (IDF):
    - a score based on the portion of documents each term appears in, with high scores for low coverage
    - Computed once for the whole corpus
  - Term Frequency (TF):
    - a direct count for the number of times the term appears in the subset of documents
    - Computed for each module

- For each set of genes (a module), LitVAn counts the term frequency in papers associated with these genes and compare this count to the null distribution, using a TF*IDF score

Dataset

- Melanoma, gene expression and copy number from 101 samples

- 64 modulators and 7869 target genes found by CONEXIC

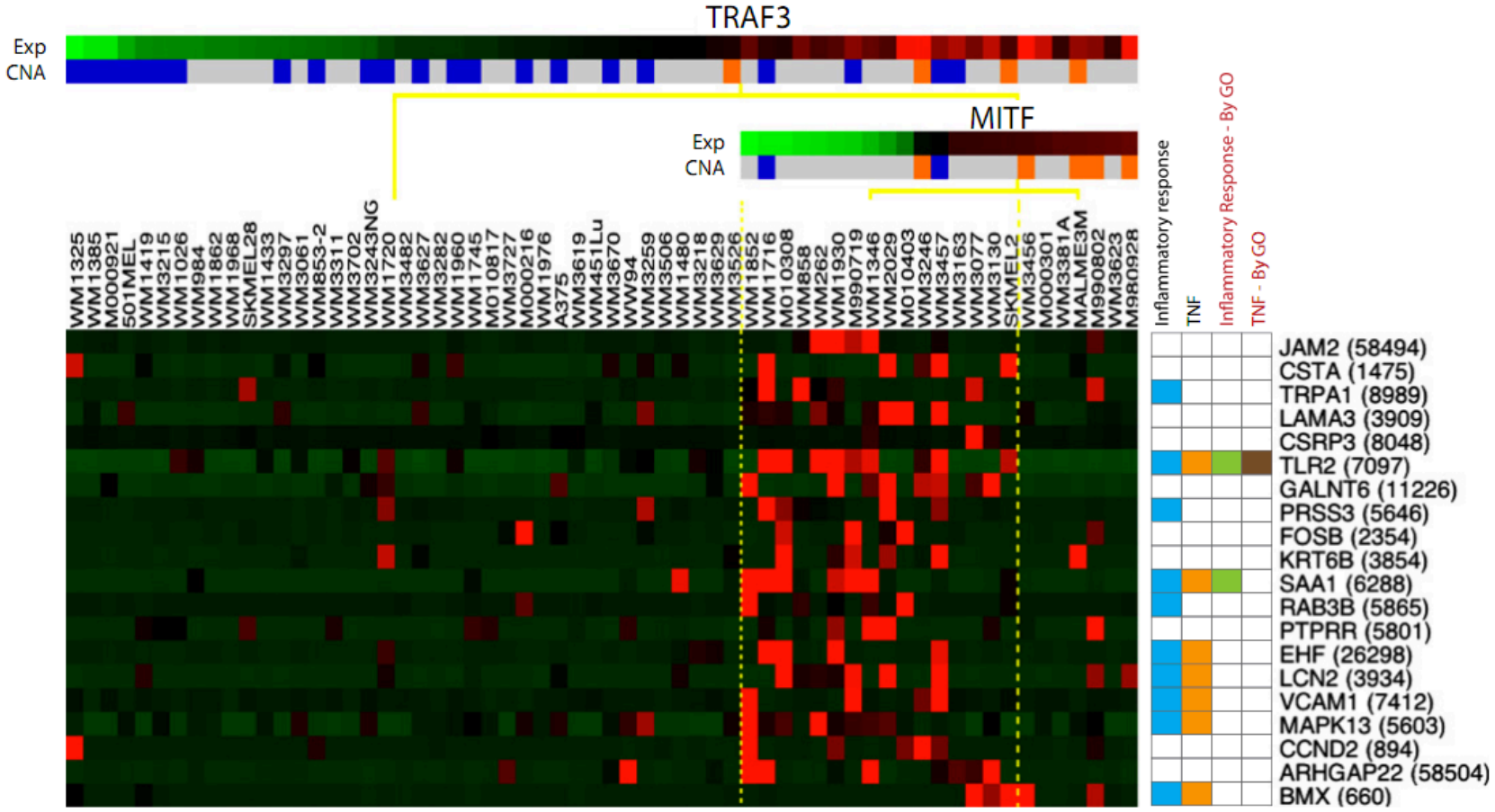
Highest Modulators Identified by CONEXIC

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	$<10^{-6}$

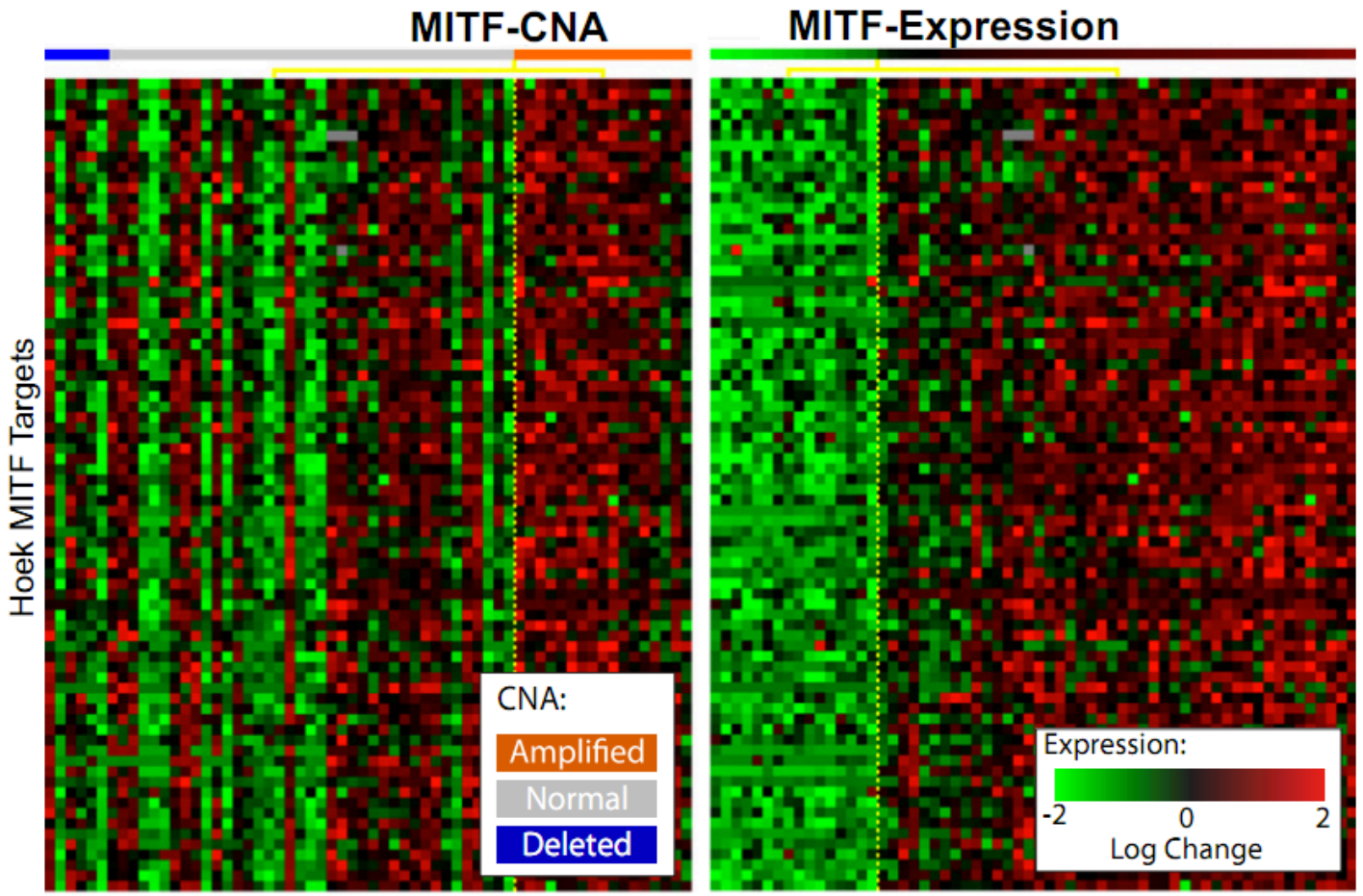
Establishing Directionality

- Copy number variation can be used to determine causal relationship

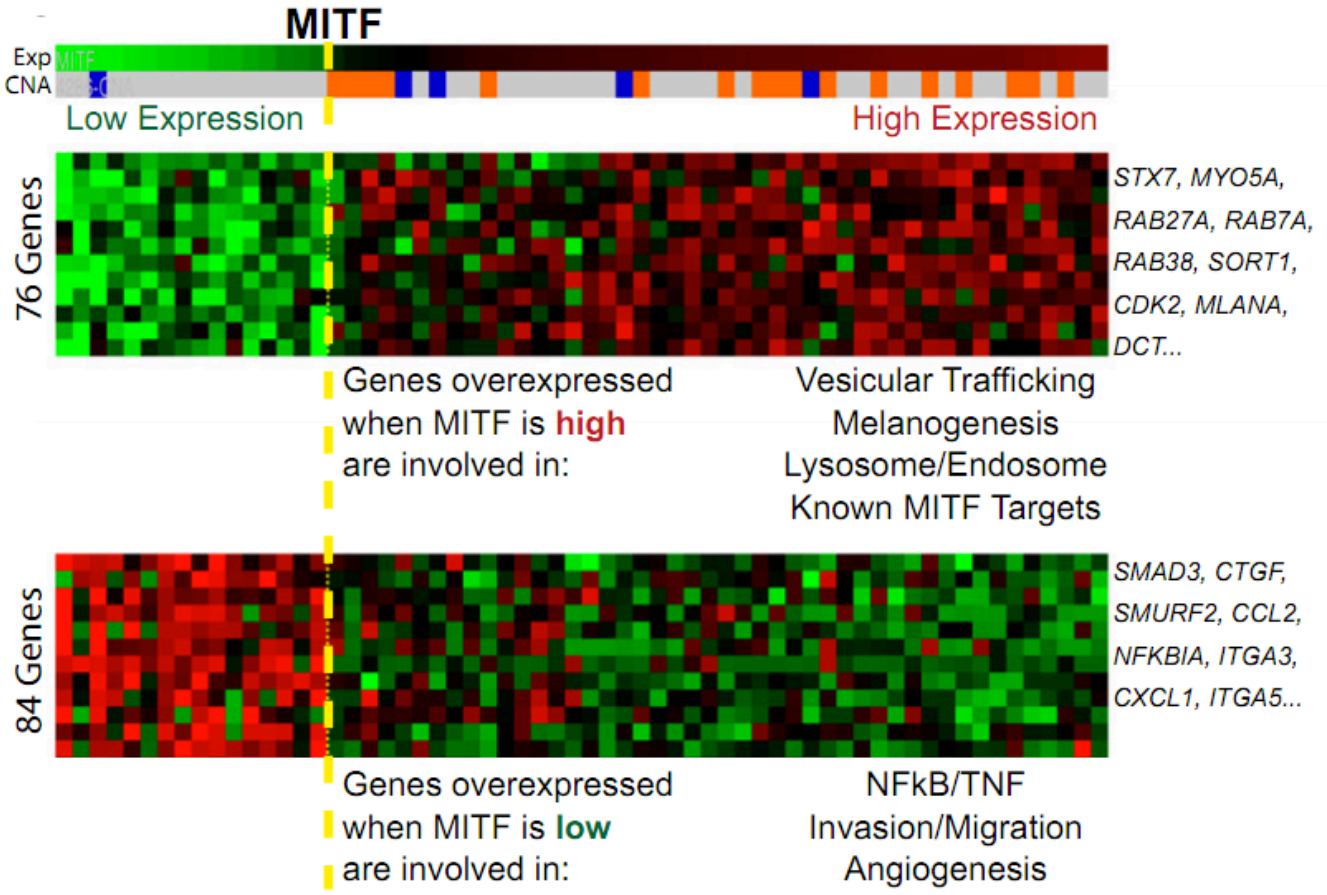
Multiple Modulators



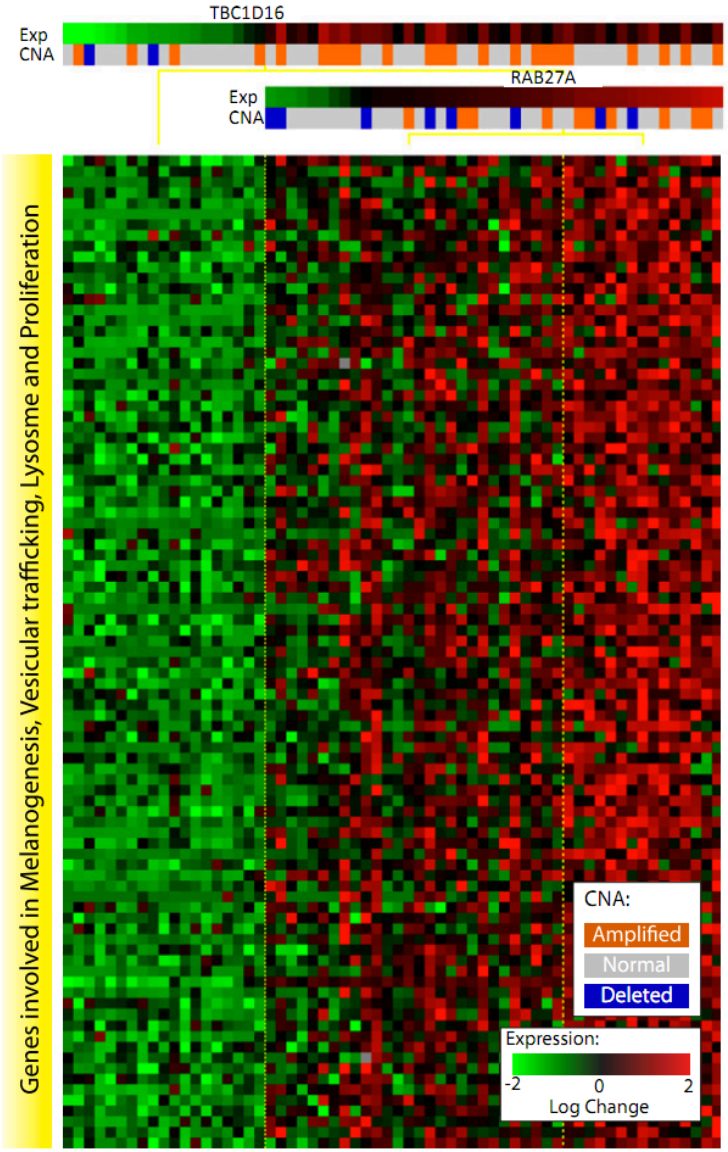
MITF Gene Expression/Copy Number Variation



MITF Modules

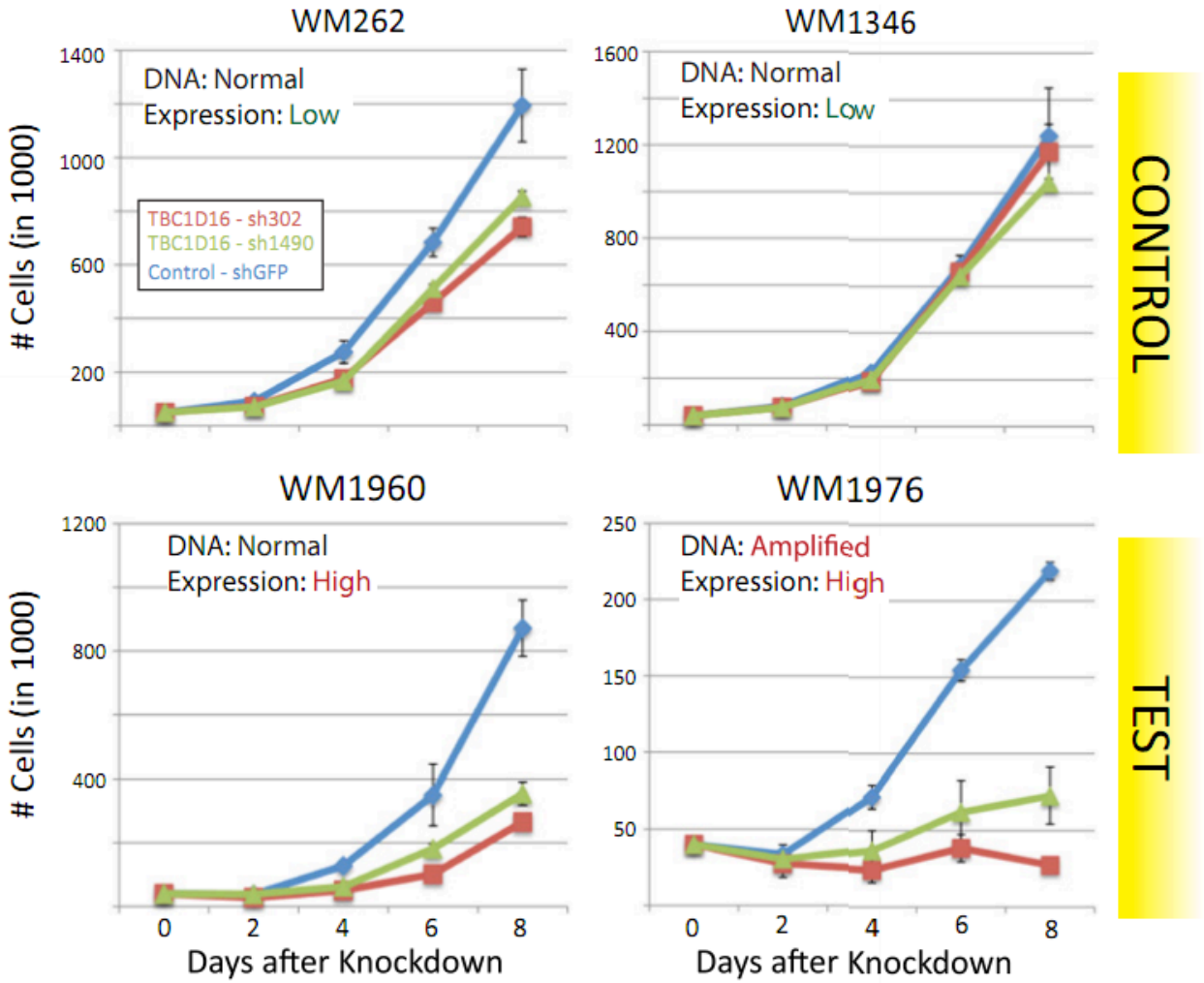


TBC1D16 for Melanoma Growth



- High correlation between TBC1D16 expression and target expressions
- Low correlation between expression and TBC1D16 copy number variation

TBC1D16 for Melanoma Growth: Experimental Validation



Summary

- Copy number variation in tumors can involve a large region that contains many genes

- CONEXIC integrates gene expression and copy number variation data collected from tumors to identify driver genes