Methods Supplement

Mathematical Description

Inputs to GISPA

1. Separate, genomewide, normalized data sets for each data type of interest (for transcript level expression or probe level GE data, transcript compatibility files will be needed)
2. Specification of the profile of interest that is based upon the input data types. The determination of each profile as biologically plausible is at the discretion of the user and may be based on a combination of factors, including the following guiding principles: i) promoter hypermethylation is consistent with gene silencing; ii) promoter hypomethylation is consistent with GE; iii) copy gain is consistent with over-expression and iv) copy loss is consistent with under-expression.
3. Specification of a class to characterize against (at least) two other classes.

Statistics

Within-Feature Profile Statistic (WFPS). Within the context of three single-sample classes, X, Y_1 and Y_2, in which X is to be characterized relative to both Y_1 and Y_2 classes, in terms of some single feature profile of increased values for some data type, we define a within-feature (increased value) profile statistic (WFPS) for each $g^{th.}$ gene as follows:

$$WFPS_{g+}(x,y_1,y_2) = \delta_g^{+}(x,y_1,y_2) = f_1(x,y_1,y_2) + f_2\,(x,y_1,y_2) = \left|\frac{(x-y_1)}{(x-y_2)} - 1\right| + \frac{y_1 y_2}{x^2} \qquad \text{(Eq.1)}$$

where $WFPS_{g+}(x,y_1,y_2)\colon \mathbb{R}^{+} \to \mathbb{R}^{+}$ such that for y_1=y_2, $\lim_{x\to\infty} WFPS_{g+}(x,y_1,y_2) = 0$, and thus low values of our WFPS_{g+} are desirable. The first term (ratio of differences of each comparison with x) forces the comparison samples to be similar in feature values to reduce the potential for ambiguities that may arise. The second term penalizes genes whose feature values for the comparison samples are close to the sample of interest, and in turn, rewards genes whose magnitude of difference between the sample of interest and comparison samples is large. Alternatively, if interest lies in characterizing X with respect to a profile defined by decreased feature values, we define a within-feature (decreased value) profile statistic by taking the inverse of the penalty as indicated below, in which case, similar to our WFPS, low values are also desirable:

$$WFPS_{g-}(x,y_1,y_2) = \delta_g^{-}(x,y_1,y_2) = f_1(x,y_1,y_2) + f_2^{-1}\,(x,y_1,y_2)$$

For a general case of M (single-sample) classes, in which the class X is to be characterized against the M-1 classes based on a single feature profile of increased values, we have the following generalized WFPS within a gth. gene:

$$WFPS_{g+}(x,y_1,\dots,y_{M-1}) = \delta_g^{+} = \binom{M-1}{2}^{-1} \sum_{i<j=1}^{M-1} f_1(x,y_i,y_j) + f_2\,(x,y_i,y_j) \qquad \text{(Eq. 2)}$$

Similarly, a generalized form for the WFPS_{g-} is defined by taking the inverse of the second term, as in the three class case. The above statistic is readily able to be applied to many settings, including the defining of gene sets with similar profiles as characteristic of one subtype versus all other subtypes, or time course data to define changes specific to a single time point compared to all other time points.

In the case of cell line or mouse data, one may average technical replicates to form a composite value for each data type within each gene. In the case of patient data, one may average over samples within each class for very few samples, or form an average filter statistic based on all combinations formed among patient samples.

Between-Genes, Within-Feature Profile Statistic (BGWFPS). To obtain a measure of the relative, among genes, within-feature profile, we construct an empirical cumulative distribution function (ecdf) corresponding to the WFPS, and define a between-gene, within-feature profile statistic (BGWFPS) by:

$$BGWFPS(wfps) = F(wfps) = Pr(WFPS \leq wfps) \qquad \text{(Eq. 3)}$$

such that $BGWFPS: \mathbb{R}^+ \rightarrow [0,1]$ for realized values of the WFPS, denoted by wfps. In this case, we assume the existence of a latent construct defined by the genes whose WFPS optimally characterizes a single class from other comparison classes. Thus, conditional on each class sample, we view the genome-wide constructed WFPS's as forming a distribution with error in estimating the true underlying latent construct from which we are able to describe genes selected as characteristic of some class of interest. The BGWFPS, as an ecdf, defines the percentiles corresponding to genome-wide WFPS values, such that for a given gene, for example, BGWFPS(wfps=10)=0.20 is interpreted as 20% of the WFPS's among all genes fall at or below a value of 10. The BGWFPS creates a standardization among the WFPS values such that regardless of feature, the range will be the same between 0 and 1, which is an important consideration for the between feature profile analyses. In this case, similar to the WFPS, and because of the monotonically increasing property of cdf's, low BGWFPS values are also desirable.

Between-Features Profile Statistic (BFPS). We construct a summary statistic of between feature profiles by summing the BGWFPS among the features. For example, if interest is in characterizing a class with the two feature profile of decreased values in feature 1 (F1) and increased values in feature 2 (F2), we form a between-feature profile statistic as below:

$$BF1F2PS_g^{-+} = BGWF1PS_g^{-} + BGWF2PS_g^{+} \qquad \text{(Eq. 4)}$$

Similar to our WFPS and BGWFPS, low values of this statistic are also desirable.

Gene Set Selection. We apply a multiple change point model (cpm) to successive differences in $-log_{10}$ transformed WFPS's in the case of a single feature profile or $-log_{10}$ transformed BFPS in the case of two or more feature-defined profiles, using a variance-based binary segmentation method (1), and implemented by the Bioconductor package, "change point" (2). The application of a multiple cpm results in several gene sets of similar profiles based on our statistics that are ranked according to the level of support for the a priori profile such that the gene set defined by the right most tail of the distribution (defined by change point1) has the most support, and so on, downward with increasing change points. The number of gene sets identified varies according to the empirical distributions for the profile of interest. For implementation, we have allowed up to the maximum number of change points to be specified, where possible, and defined an empty set for the case in which a change point is not able to be identified. For details on the implementation of a multiple change point model that includes diagnostics to guide the selection on the number of change points for a given profile, see our R package, GISPA-NGS (https://github.com/BhaktiDwivedi/GISPA.NGS, available for download.

Prominent Feature. Within each gene, we examine the proportionate contribution of each feature to the BFPS. Since small values of our BFPS are desirable, we require small values to be associated with a larger proportion of this statistic. To achieve this inverse relation, we define the following:

$$p_g^{F1} = \frac{\left(BGWF1PS_g^-\right)^{-1}}{\left(BGWF1PS_g^-\right)^{-1} + \left(BGWF2PS_g^+\right)^{-1}}; p_g^{F2} = 1 - p_g^{F1} \quad \text{(Eq. 5)}$$

Using these proportions, a prominent feature (if it exists) may be defined as that feature which is associated with the maximum proportion. For example, if the summed percentile is 1.10 based on two features, feature 1 percentile=0.90 and feature 2 percentile=0.20, then the percent contribution from feature 1 is 18% and that of feature 2 is 82%, and thus feature 2 is the defined prominent feature driving the BFPS.

Transcript Level Results. When forming a combined data set, a transcript incompatibility may occur that is considered. For example, it may be the case that for a given gene, a CpG site does not correspond to a transcript in the microarray GE array data. In such cases, we filter the combined data to include compatible combinations of features.

Gene Level Results. The variable number of units (e.g., probes, transcripts, CpG sites) representing a single gene may create an issue such that genes defined by several units are likely to be selected simply based on their larger number of units representation. For this reason, we implement a 'carry one forward' approach that selects the unit associated with the smallest WFPS for a single-feature profile to carry forward in subsequent analysis, thus producing gene level results. In the case of a multi-feature profile, we select the combination of units (e.g., variant and CpG site) that is associated with the smallest BFPS to carry forward for a single gene, and the BGWFPS is updated based on the gene level units selected in subsequent analysis. While this is a simple approach, another is to apply GISPA (Fig. 1) to all units defined for a gene, such as all CpG sites within a gene, and identify the set of CpG sites based on change point 1 results to 'move forward', and update GISPA results using these selected sites to represent a single gene. The same may be done for all possible combinations of say, CpG methylation sites, CN segments and GE data, within each gene.

Gene Set Significance. We assess the significance of gene sets by comparing our data results through randomly assigning genes to sets (i.e. 'row randomization') based on the following algorithm.

1. Randomly assign genes to each set, the size of which is based on our data results, and apply the GISPA method to this randomized dataset to obtain a mean between-feature profile statistic (BFPS; Eq. 4) for each gene set.
2. Repeat step 1 for 1,000 gene randomizations and calculate a mean BFPS for each gene set.
3. Estimate a monte-carlo p-value by the proportion of times that the mean BFPS based on the randomized data is greater than the mean BFPS based on our data results for each gene set.

Once statistically significant gene sets are defined based on the above algorithm, the significance of individual genes in characterizing a class within a set may be further examined by randomly assigning classes within each gene. This permutation approach has been previously applied for use with single sample classes and shown to produce consistent results in such cases (3,4).

SISPA Profile Score. We demonstrate the calculation of a SISPA profile score based on p genes with the profile of increased GE (feature 1) with decreased methylation (feature 2) from n samples. Let z_{gj}^1 denote feature 1 GE values and z_{gj}^2, feature 2 methylation beta values for a $g^{th.}$ gene and $j^{th.}$ sample, such that $\mathbf{z}_g^1 = \left(z_{g1}^1, \dots, z_{gn}^1\right)$ and $z_g^2 = \left(z_{g1}^2, \dots, z_{gn}^2\right)$ each have mean zero and variance one across all n samples within a gene. A profile score for feature 1 within a $j^{th.}$ sample is defined by:

$$z_j^1 = \frac{1}{\sqrt{p}} \sum_{g=1}^{p} z_{gj}^1 \quad \text{(Eq.6)}$$

such that the variance of z_{+j}^{1} is one within each sample. A profile score is similarly defined for feature 2 within a jth. sample and denoted by z_j^2 . A composite, between features 1 and 2 profile score in the context of increased GE with decreased methylation score within a jth. sample is defined by $z_j^{+\,-} = (z_j^1 - z_j^2)$. A multiple change point model is then applied to this (-log10 transformed) composite score to define samples with and without profile activity.

Methods Supplement

Data Collection, Preprocessing, Normalization, Transformations

The microarray datasets have been deposited in NCBI's Gene Expression Omnibus (GSE68258 for SNP-CN; GSE68259 for methylation). The RNA-Seq data has been deposited in NCBI's Sequence Read Archive under Bioproject ID SRP057322.

Methylation. CpG island methylation data was generated based on the Illumina Infinium Human Methylation 450K Beadchip and run in duplicate for each set of three cell lines. Intensities from both runs, among all three cell lines, were processed for CpG site and sample quality checks, and normalized using the bioconductor packages, 'lumi' (5) and 'methylumi' (6). Data preprocessing included, color balance adjustment, sample quality assessment based on CpG intensity, background level correction, and data normalization used simple scaling normalization. A mean beta value was obtained between the two runs within each cell line. Prior to analysis, CpG sites with a detection p-value greater than 0.0001 in at least 50% of the samples were removed.

SNP-CN. Copy number (CN) data was generated based on the Illumina Omni1 Quad SNP-CN Array on each cell line and processed using in Illumina Genome Studio. Segmentation was performed on the CN data using Partek software (v6.6, Partek Inc.) in which a segment is defined according to the following criteria: 1) neighboring regions have significant ($p < 0.001$) mean intensity differences; 2) breakpoints are chosen to provide optimal statistical significance and 3) detected regions contain at least 10 SNPs. HapMaP samples were used for assessment of 'normal' CN. CN segments were further defined as either focal or large-scale, according to whether the segment's chromosome arm fraction was less than or greater than or equal to 2%, respectively. The threshold of 2% corresponded to approximately the 95th percentile among chromosome arm fractions within each cell line. The grouping of CN segments in this way allows for differential variability in relation to segment size, and has been similarly applied to CN data (7). A segment was categorized according to its mean CN as follows: homozygous (CN < 0.7), heterozygous (0.7 <= CN <= 1.7), normal (1.7 < CN <= 2.3) or amplified (> 2.3). Similar thresholds have been applied to MM CN data (8). Prior to analysis, segments were removed if the call (e.g., gain, loss, no change) was not specific to any one of the three cell lines. No segments were identified with a copy gain or loss in two cell lines that was not present in the third. An inferred gene CN change is based on the segment in which the gene was included, either in total or part.

RNA-Seq. Illumina TruSeq RNA protocols were used to generate RNA-Seq libraries that were sequenced using the Illumina HiSeq2000 to obtain 100bp paired-end reads. After performing sample QC on raw FASTQ files, determining any necessary trimming that may be required using FASTQC and applying some additional QC, files were aligned to Human Reference Genome hg19 (GRCh37) using Bowtie2 (9) and Tophat2 (10). DESeq (11), Varscan2 (7) , and ANNOVAR (12) was used to obtain RNA-Seq derived GE and variant data, and RefSeq annotation of coding variants, respectively. Variants located in exon, splicing site, upstream/downstream, or 5'/3 UTR regions that were associated with a non-synonymous change were included. Variants present in dbSNP137 were removed unless the same variant was also present in COSMIC67. Variant calls were made based on a minimum read depth of eight with at least two supporting reads at a position. Variant calls were used in the analysis by defining the proportion of reads supporting a variant among the total number of reads. The GE analysis was done using DESeq normalization at gene-level with Cufflinks/Cuffdiff (13).

SUPPLEMENTAL TABLES

MM_expcnv.xlsx

Table 1. GISPA selected (change point1) genes that satisfy the (two-feature) profile of decreased copy number with decreased RNA-Seq GE, specific to KMS11, MM1s and RPMI8266 cell lines.

MM_expvar.xlsx

Table 2. GISPA selected (change point 1) genes that satisfy the (two-feature) profile of RNA-Seq coding variants with increased RNA-Seq gene expression, specific to KMS11, MM1s and RPMI8266 cell lines.

MM_metvar.xlsx

Table 3. GISPA selected (change point 1) genes that satisfy the (two-feature) profile of RNA-Seq coding variants with increased CpG island methylation, specific to KMS11, MM1s and RPMI8266 cell lines.

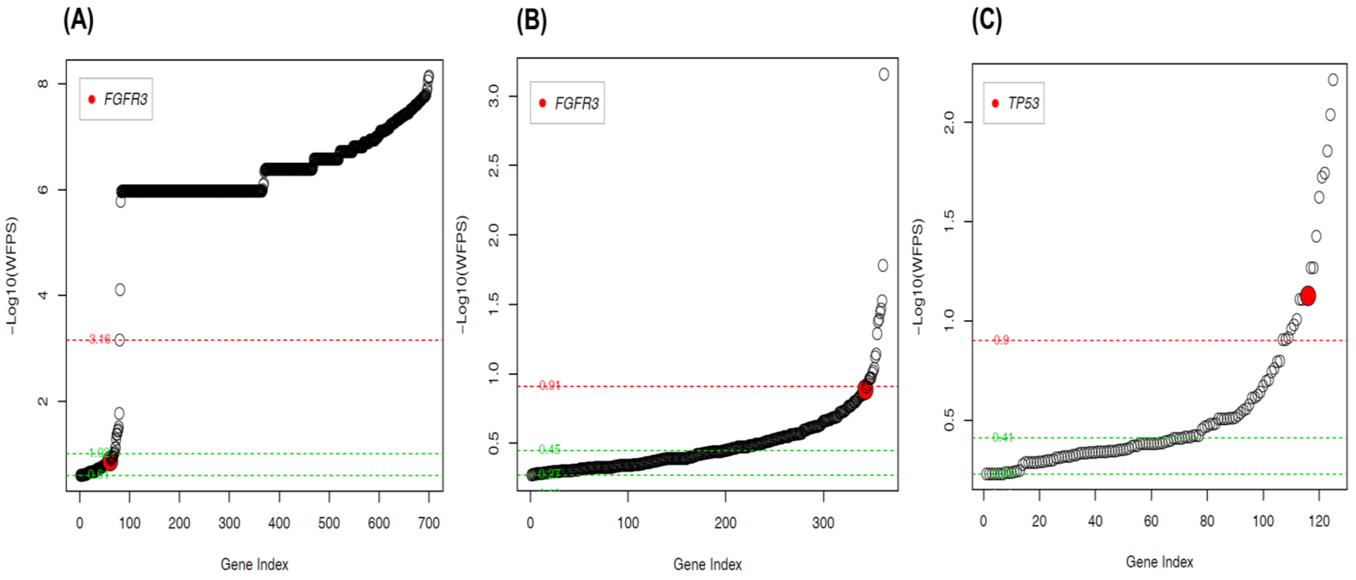
MM_expmetcnv.xlsx

Table 4. GISPA selected (change point 1) genes that satisfy the (three-feature) profiles combining RNA-Seq gene expression, CpG island methylation, and copy number combination of changes (increased vs. decreased), specific to KMS11, MM1s and RPMI8266 cell lines.

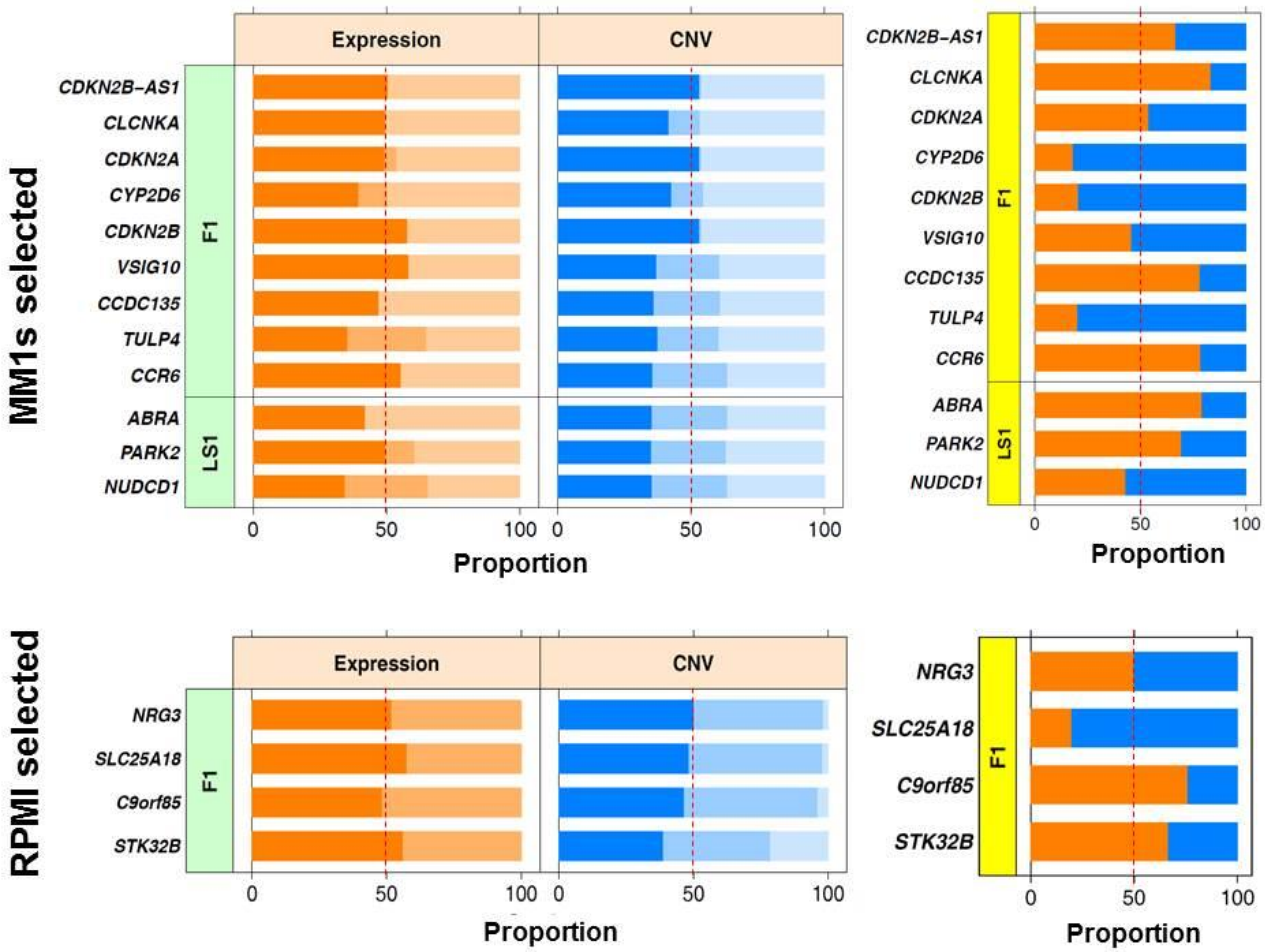
MM_mirnamerger.xls
x

Table 5. miRNAs predicted or known to target the majority of GISPA selected (change point 1) gene set results that satisfy the profile of decreased RNA-Seq gene expression with increased CpG island methylation and decreased copy number, specific to KMS11, MM1s and RPMI8226 cell lines based on mirnamerger (http://mirnamerger.org/).

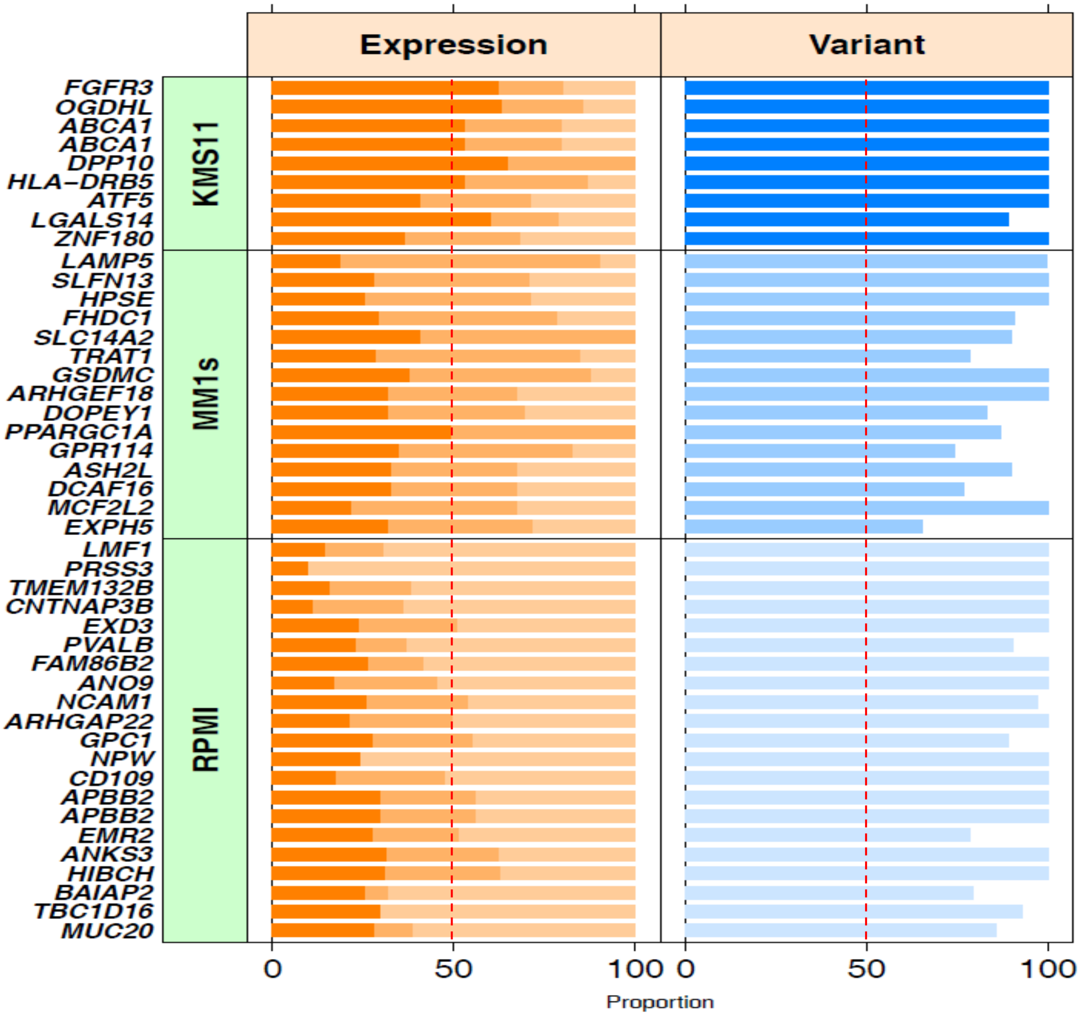
SUPPLEMENTAL FIGURES



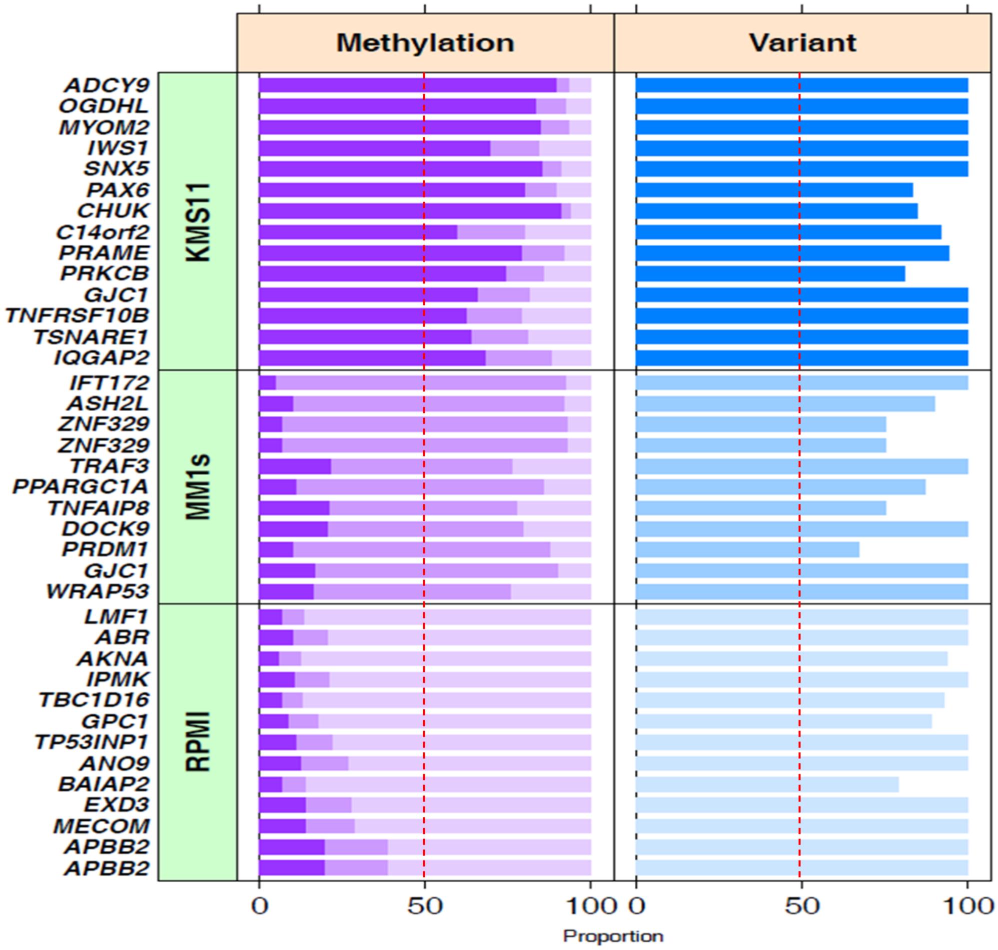
Supplementary Fig.1. GISPA identifies known molecular features characteristic of multiple myeloma cell lines. Plots show transformed (-log10) between-feature profile statistic (BFPS; see Fig. 1) on the y-axis and gene index on the x-axis. The first three change points (cpt) are plotted as horizontal lines with cpt1 in red and cpt2 and 3 in green. A) KMS11 high GE shows *FGFR3* (red dot) falling above cpt3. The ‘extreme cases’ of genes expressed in KMS11 but with zero GE values for both RPMI and MM1s cell lines are defined by the first two change points. B) KMS11 increased GE excluding genes with zero GE values for both MM1s and RPMI: *FGFR3* falls above in cpt2. C) KMS11 decreased copy number: *TP53* is selected in cpt1 as part of a deletion specific to KMS11.



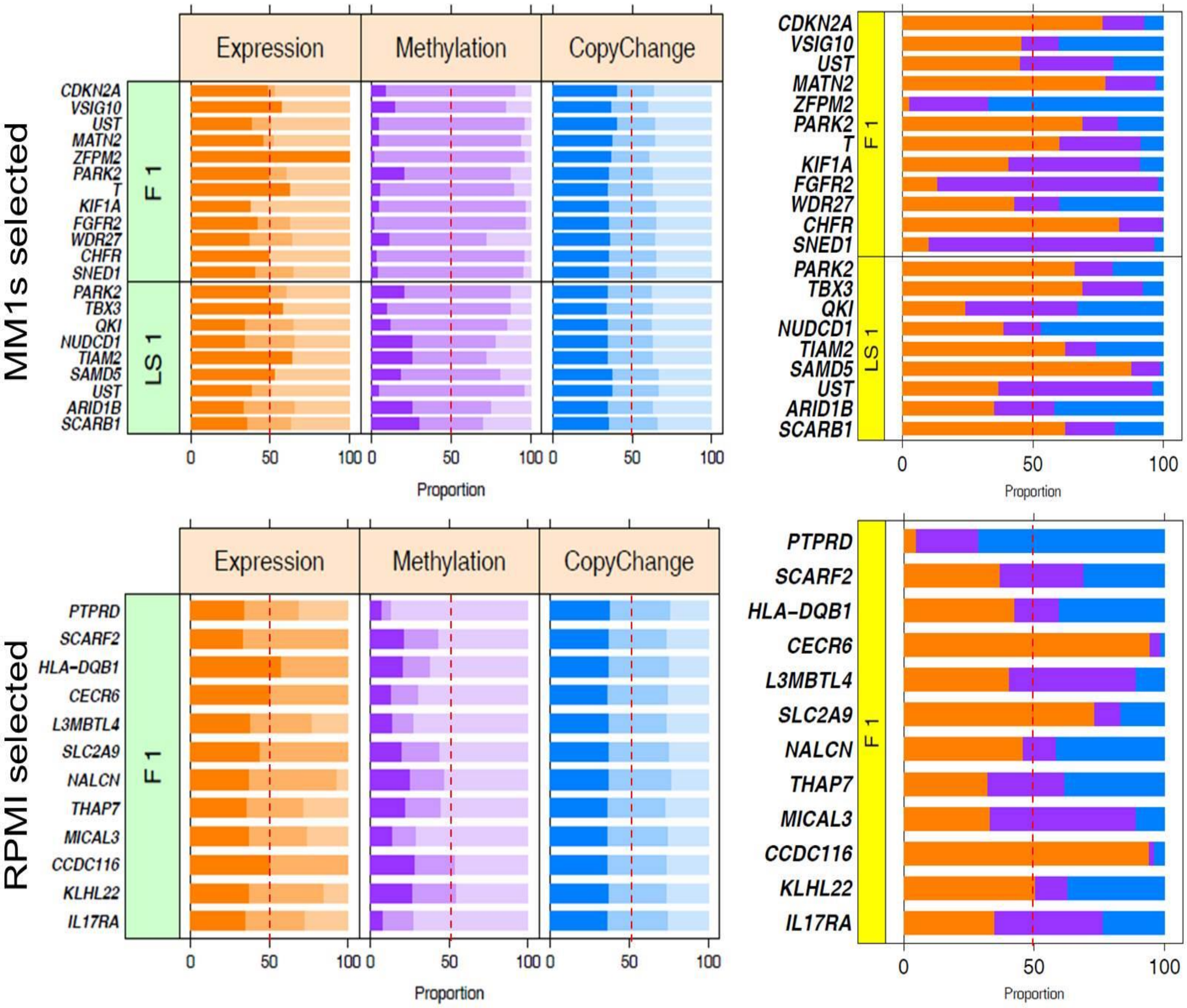
Supplementary Fig. 2. Two feature GISPA identifies MM1s and RPMI8226 selected genes with decreased gene expression (GE) and decreased copy number (CN). Change point 1 (cpt1) gene set results that satisfy the profile of decreased GE with decreased CN by CN segment as focal (F) or large=scale (LS) according to whether the segment's chromosome arm fraction was less than or greater than or equal to 2%, respectively (F1=focal in cpt1l; LS1=Large-Scale in cpt1) are displayed by cell line. Genes are sorted from the smallest to largest between-feature profile statistic. Left: *Between-cell line differences.* Within each data type: GE (in orange) and CN change (in blue), a stacked bar denoting the percent contribution to the total change from each cell line is displayed along a color gradient from darkest (KMS11) to medium (MM1s) to lightest (RPMI) shades. Among all genes selected, the percent contribution to total changes in each feature is the smallest for both GE and CN for the selected cell line. Right: *Between-feature Differences.* The percent contribution from each feature to the profile is displayed as a stacked bar. Among the MM1s selected genes, *CYP2D6, CDKN2B, VSIG10, TULP4*, and *NUDCD1* show CN changes as the prominent feature driving the profile, with the remaining showing GE. Among the RPMI selected genes, *SLC25A18* shows CN, while *NRG3* shows both CN and GE as prominent features, and the remaining show GE.



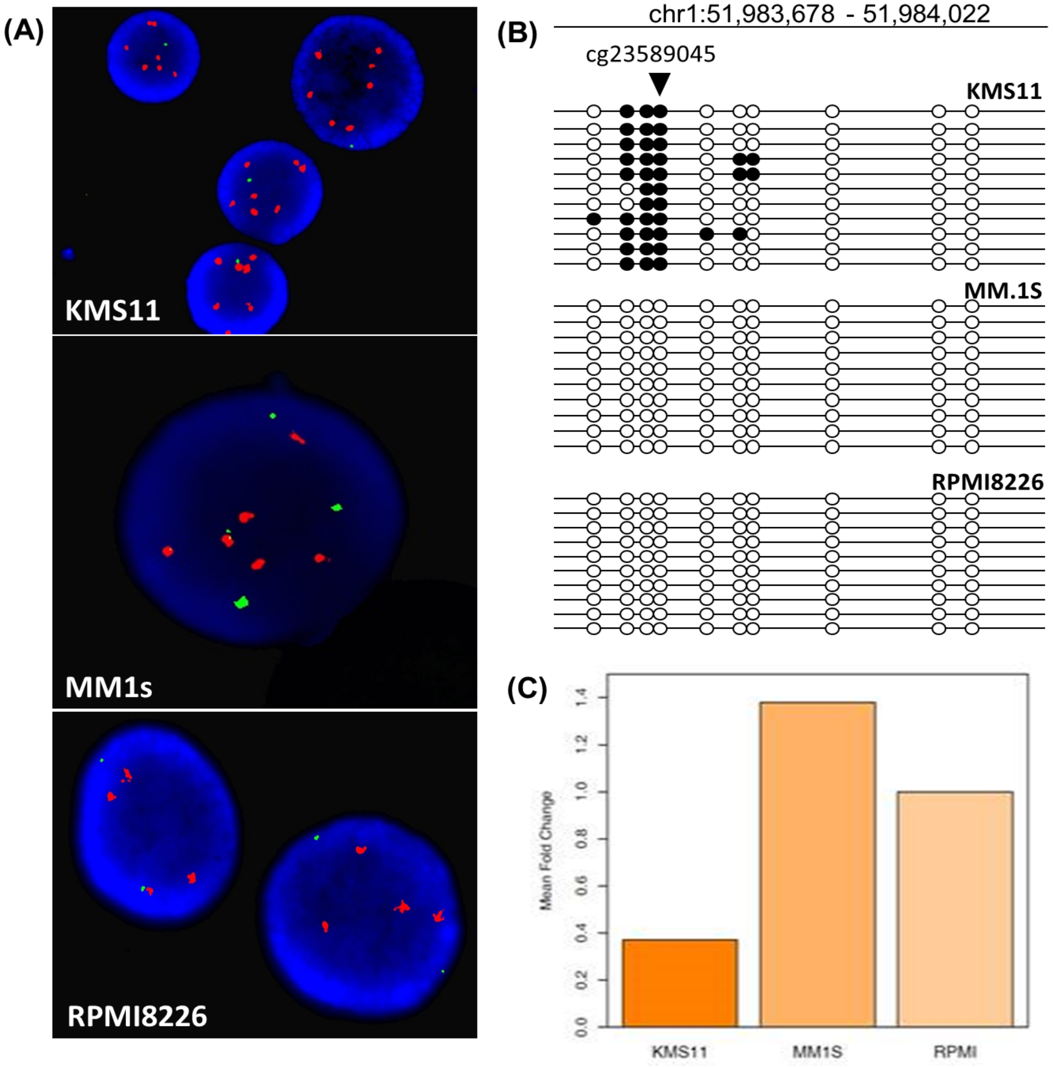
Supplementary Fig. 3. Two-feature GISPA identifies KMS11, MM1s and RPMI8226 selected coding variants with increased gene expression (GE). Change point 1 (cpt1) gene set results that satisfy the profile of coding variants with increased GE are displayed by cell line. Genes are sorted from the smallest to largest between-feature profile statistic. Within each data type: GE (in orange) and variant proportions (in blue), a stacked bar denoting the percent contribution from each cell line is displayed along a color gradient from darkest (KMS11) to medium (MM1s) to lightest (RPMI) shades. Among the KMS11 expressed variants, *FGFR3* is selected as the topmost gene showing a missense Y373C mutation (COSMIC ID, COSM718) with all reads supporting the variant allele (see Supplemental Table 2), and increased *FGFR3* GE specific to the KMS11 cell line, results that are well established data features of KMS11 (14). Among all other genes identified as having expressed variants, some have shown to be associated with cancer. *MYEOV* over-expression in MM patients is indicative of poorer prognosis (15). *MYEOV* has also been shown to be overexpressed in MM cell lines with t(11;14) (16) and colon cancer (17), and amplified in breast cancer (18). *CAMKK2* has been shown as overexpressed in near-tetraploid mantel cell lymphoma (19).



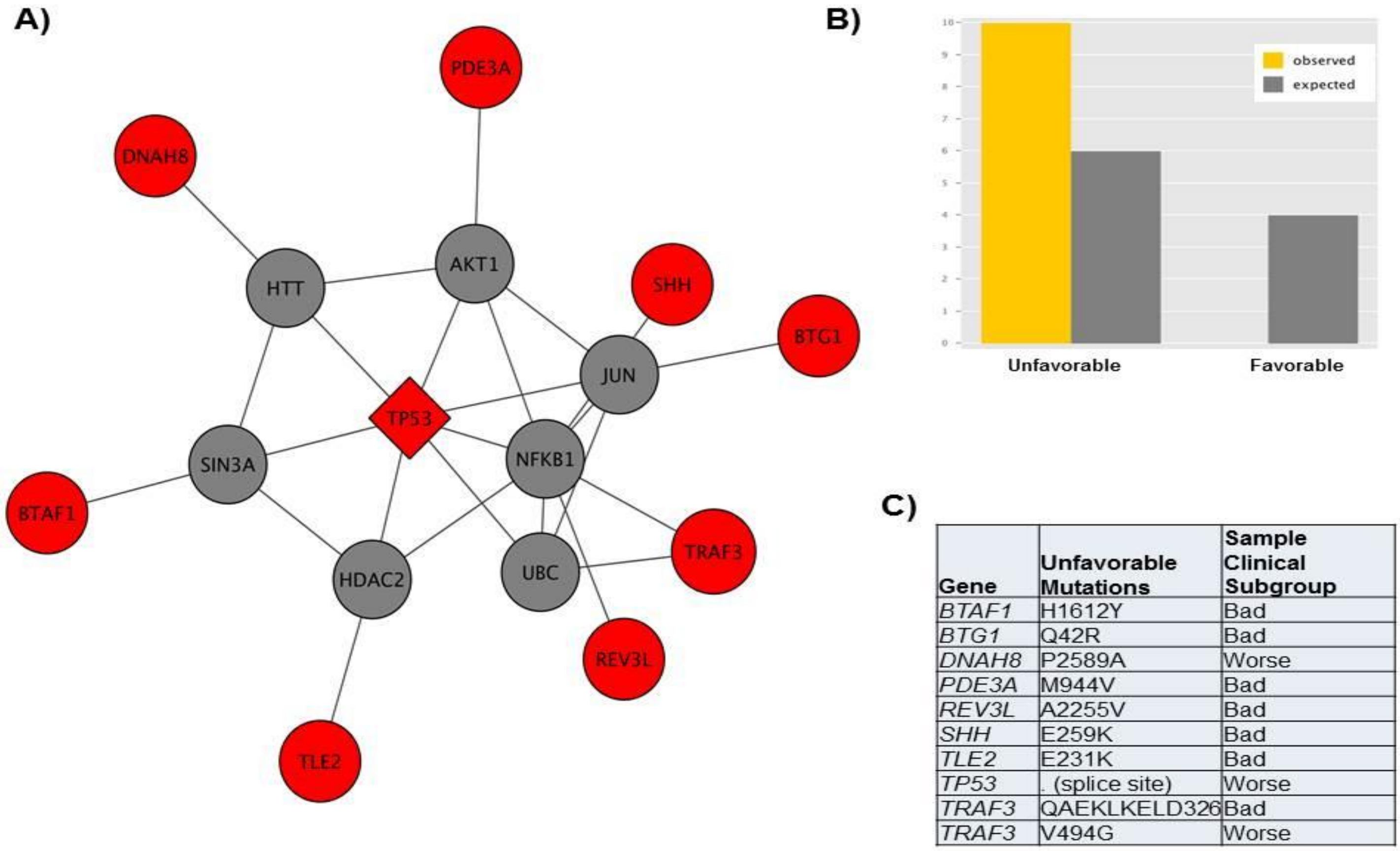
Supplementary Fig. 4. Two-feature GISPA identifies KMS11, MM1s and RPMI8226 selected coding variants with increased CpG island methylation. Change point 1 (cpt1) gene set results that satisfy the profile of coding variants with increased CpG methylation are displayed by cell line. Genes are sorted from the smallest to largest between-feature profile statistic. Within each data type: CpG methylation (in orange) and variant proportions (in blue), a stacked bar denoting the percent total contribution from each cell line is displayed along a color gradient from darkest (KMS11) to medium (MM1s) to lightest (RPMI) shades. Among the KMS11 selected genes, *PAX6* has been shown to be expressed in pancreatic cancer (20), and hypermethylated in breast cancer (21). *TNFRS10* has been shown to be over-expressed in MM cell lines that over-express *TP53* (22) and hypermethylated in neuroblastoma (23). *OGDHL* is hypermethylated in breast, cervix, lung, esophageal, pancreatic, and colon cancers (24-27). *PRKCB* and *GJC1* were associated with hypermethylation in breast and colorectal cancer, respectively (28,29). *IQGAP2* has been associated with under-expression in colorectal cancer (30) and is silenced via methylation in gastric cancer; which is also associated with poor prognosis (31).



Supplementary Fig. 5. Three Feature GISPA identifies KMS11 selected genes with decreased gene expression (GE), increased CpG methylation and decreased copy number (CN). Change point 1 (cpt1) gene set results that satisfy the profile of decreased GE with increased CpG methylation and decreased (heterozygous) CN by CN segment as focal (F) or large=scale (LS) according to whether the segment's chromosome arm fraction was less than or greater than or equal to 2%, respectively (F1=focal in cpt1l; LS1=Large-Scale in cpt1) are displayed by cell line. Genes are sorted from the smallest to largest between-feature profile statistic. Left: *Between-cell line differences.* Within each data type: GE (in orange), CpG methylation beta values (in purple), and CN change (in blue), a stacked bar denoting the percent contribution from each cell line is displayed along a color gradient from darkest (KMS11) to medium (MM1s) to lightest (RPMI) shades. Among all genes selected, as expected, the percent contribution to total changes in each feature is the smallest for GE, largest for CpG methylation and smallest for CN change within each cell line. Right: *Between-feature Differences.* The percent contribution from each feature to the profile is displayed as a stacked bar. Among the MM1s selected genes, the top gene, *CDKN2A* shows GE as the prominent feature driving the profile, with a combination of prominent feature changes from all other genes. Among the RPMI selected genes, the top gene, *PTPRD* shows CN change as the prominent feature, while the remaining genes show a combination of feature changes, some genes with all three features as prominent (e.g., *SCARF2, IL17RA*).

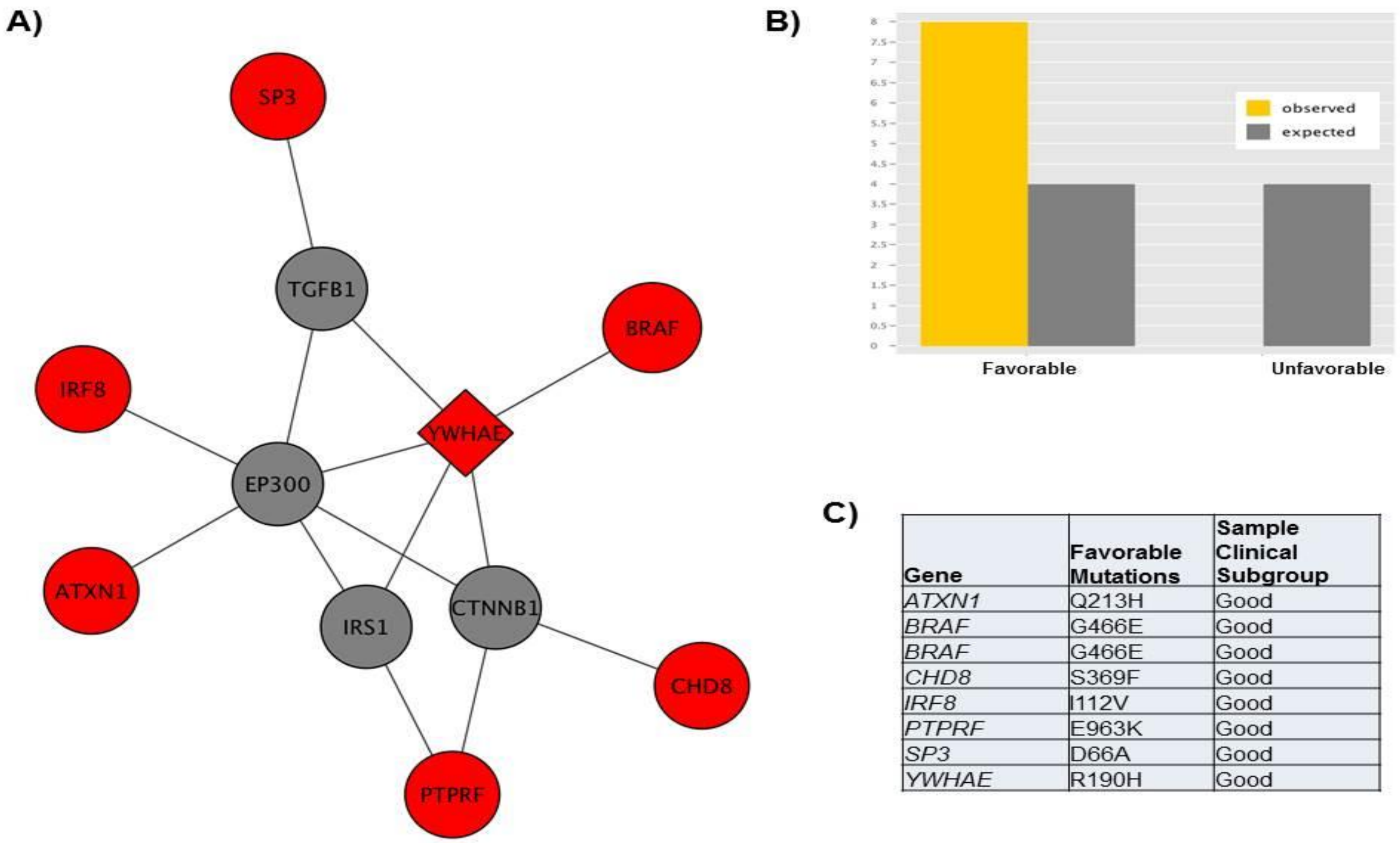


Supplementary Fig. 6. Experimental validation that *EPS15* locus allelic loss and hypermethylation is specific to KMS11. A) FISH results for 1p32.3 (band containing *EPS15)* based on CDKN2C, located upstream of *EPS15*. 1p/1q FISH was performed according to manufacturer's instructions using CDKN2C (1p32.3) probes (shown in green) and CSK1B (1q21.2) probes (shown in orange) (Cytocell LTD, Cambridge, UK, Cat.# LPH 039-A) (32,33) B) DNA methylation of the GISPA selected EPS15 locus was determined by bisulfite sequencing (34). Primers were designed to avoid CpG sites and to recognize the bisulfite-modified sequence and span a 345bp amplicon on (Hg19) chr1:51,983,678 - 51,984,022. Shown are 8-10 independently cloned alleles from the indicated cell line. Filled circle, methylated CpGs; open circles, unmethlyated CpGs. Arrow indicates the CpG site interrogated by the Illumina 450K platform (cg23589045). C) qPCR cDNA results of mean fold change versus control based on triplicate data (SD=0.02 for KMS11, SD=0.13 for MM1s, SD=0 for RPMI8226).



Gene	Unfavorable Mutations	Sample Clinical Subgroup
BTAF1	H1612Y	Bad
BTG1	Q42R	Bad
DNAH8	P2589A	Worse
PDE3A	M944V	Bad
REV3L	A2255V	Bad
SHH	E259K	Bad
TLE2	E231K	Bad
TP53	. (splice site)	Worse
TRAF3	QAEKLKELD326	Bad
TRAF3	V494G	Worse

Supplementary Fig. 7. Somatic mutations with skewed allelic gene expression (GE) subnetwork (module) enriched in unfavorable prognostic class. (A) Detected subnetwork (p=0.10) enriched in 10 newly diagnosed MM patients with unfavorable prognosis based on 1,000 permuted networks applied to the GISPA-defined, cpt1 mutations with increased GE from the 29 patients subgroup in the coMMpass trial. (B) Fisher's exact test of association between prognostic status (unfavorable vs. favorable) and mutation in the subnetwork. (C) Gene mutations for the subnetwork identified as enriched in the 10 patients. Genes are color-coded (red=mutated gene; grey=non-mutated gene). Genes with a diamond shape are seed genes used as part of the input network.



Gene	Favorable Mutations	Sample Clinical Subgroup
ATXN1	Q213H	Good
BRAF	G466E	Good
BRAF	G466E	Good
CHD8	S369F	Good
IRF8	I112V	Good
PTPRF	E963K	Good
SP3	D66A	Good
YWHAE	R190H	Good

Supplementary Fig. 8. Somatic mutations with skewed allelic gene expression (GE) subnetwork (module) enriched in the favorable prognostic group. (A) Detected subnetwork (p=0.05) enriched in 8 newly diagnosed MM patients with unfavorable prognosis based on 1,000 permuted networks applied to the GISPA-defined, cpt1 mutations from the 29 patients subgroup in the coMMpass trial. (B) Fisher’s exact test of association between prognostic status (unfavorable vs. favorable) and mutation in the subnetwork. (C) Gene mutations for the subnetwork identified as enriched in the 8 patients. Genes are color-coded (red=mutated gene; grey=non-mutated gene). Genes with a diamond shape are seed genes used as part of the input network.

Supplement References

1. Edwards, A.W.F. and Cavalli-Sforza, L.L. (1965) A Method for Cluster Analysis. *Biometrics*, **21**, 14.
2. Killick, R. and Eckley, I.A. (2014) Changepoint: An R package for Changepoint Analysis. *Journal of Statistical Software*, **58**, 19.
3. Kowalski, J. and Powell, J. (2004) Inference for Stochastic Linear Hypotheses: Application to High Dimensional Data. *Biometrika*, **91**, 16.
4. Kowalski, J., Drake, C., Schwartz, R.H. and Powell, J. (2004) Non-parametric, hypothesis-based analysis of microarrays for comparison of several phenotypes. *Bioinformatics (Oxford, England)*, **20**, 364-373.
5. Du, P., Kibbe, W.A. and Lin, S.M. (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics (Oxford, England)*, **24**, 1547-1548.
6. Davis, S., Du, P., Bilke, S., Triche, T. and Bootwalla, M. (2013) methylumi: Handle Illumina methylation data. *R package version 8.0*.
7. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L. and Wilson, R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**, 568-576.
8. Walker, B.A., Leone, P.E., Chiecchio, L., Dickens, N.J., Jenner, M.W., Boyd, K.D., Johnson, D.C., Gonzalez, D., Dagrada, G.P., Protheroe, R.K.M. *et al.* (2010) A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood*, **116**, E56-E65.
9. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.
10. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, **14**, R36.
11. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology*, **11**, R106.
12. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, **38**, e164.
13. Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics (Oxford, England)*, **27**, 2325-2329.
14. Chesi, M., Nardini, E., Brents, L.A., Schrock, E., Ried, T., Kuehl, W.M. and Bergsagel, P.L. (1997) Frequent translocation t(4;14)(p16.3;q32.3) in multiple myeloma is associated with increased expression and activating mutations of fibroblast growth factor receptor 3. *Nature genetics*, **16**, 260-264.
15. Moreaux, J., Hose, D., Bonnefond, A., Reme, T., Robert, N., Goldschmidt, H. and Klein, B. (2010) MYEOV is a prognostic factor in multiple myeloma. *Experimental hematology*, **38**, 1189-1198.e1183.
16. Janssen, J.W., Vaandrager, J.W., Heuser, T., Jauch, A., Kluin, P.M., Geelen, E., Bergsagel, P.L., Kuehl, W.M., Drexler, H.G., Otsuki, T. *et al.* (2000) Concurrent activation of a novel putative transforming gene, myeov, and cyclin D1 in a subset of multiple myeloma cell lines with t(11;14)(q13;q32). *Blood*, **95**, 2691-2698.
17. Moss, A.C., Lawlor, G., Murray, D., Tighe, D., Madden, S.F., Mulligan, A.-M., Keane, C.O., Brady, H.R., Doran, P.P. and MacMathuna, P. (2006) ETV4 and Myeov knockdown impairs colon cancer cell line proliferation and invasion. *Biochemical and Biophysical Research Communications*, **345**, 216-221.
18. Janssen, J.W., Cuny, M., Orsetti, B., Rodriguez, C., Valles, H., Bartram, C.R., Schuuring, E. and Theillet, C. (2002) MYEOV: a candidate gene for DNA amplification events occurring centromeric to CCND1 in breast cancer. *International journal of cancer. Journal international du cancer*, **102**, 608-614.
19. Neben, K., Ott, G., Schweizer, S., Kalla, J., Tews, B., Katzenberger, T., Hahn, M., Rosenwald, A., Ho, A.D., Müller-Hermelink, H.K. *et al.* (2007) Expression of centrosome-associated gene products is linked to tetraploidization in mantle cell lymphoma. *International Journal of Cancer*, **120**, 1669-1677.

20. Mascarenhas, J.B., Young, K.P., Littlejohn, E.L., Yoo, B.K., Salgia, R. and Lang, D. (2009) PAX6 Is Expressed in Pancreatic Cancer and Actively Participates in Cancer Progression through Activation of the MET Tyrosine Kinase Receptor Gene. *Journal of Biological Chemistry*, **284**, 27524-27532.
21. Conway, K., Edmiston, S., May, R., Kuan, P., Chu, H., Bryant, C., Tse, C.-K., Swift-Scanlan, T., Geradts, J., Troester, M. *et al.* (2014) DNA methylation profiling in the Carolina Breast Cancer Study defines cancer subclasses differing in clinicopathologic characteristics and survival. *Breast Cancer Research*, **16**, 450.
22. Xiong, W., Wu, X., Starnes, S., Johnson, S.K., Haessler, J., Wang, S., Chen, L., Barlogie, B., Shaughnessy, J.D., Jr. and Zhan, F. (2008) An analysis of the clinical and biologic significance of TP53 loss and the identification of potential novel transcriptional targets of TP53 in multiple myeloma. *Blood*, **112**, 4235-4246.
23. van Noesel, M.M., van Bezouw, S., Voûte, P.A., Herman, J.G., Pieters, R. and Versteeg, R. (2003) Clustering of hypermethylated genes in neuroblastoma. *Genes, Chromosomes and Cancer*, **38**, 226-233.
24. Sen, T., Sen, N., Noordhuis, M.G., Ravi, R., Wu, T.-C., Ha, P.K., Sidransky, D. and Hoque, M.O. (2012) OGDHL is a modifier of AKT-dependent signaling and NF-κB function. *PloS one*, **7**, e48770.
25. Ostrow, K.L., Park, H.L., Hoque, M.O., Kim, M.S., Liu, J., Argani, P., Westra, W., Criekinge, W.V. and Sidransky, D. (2009) Pharmacologic Unmasking of Epigenetically Silenced Genes in Breast Cancer. *Clinical Cancer Research*, **15**, 1184-1191.
26. Hoque, M.O., Kim, M.S., Ostrow, K.L., Liu, J., Wisman, G.B.A., Park, H.L., Poeta, M.L., Jeronimo, C., Henrique, R., Lendvai, Á. *et al.* (2008) Genome-Wide Promoter Analysis Uncovers Portions of the Cancer Methylome. *Cancer research*, **68**, 2661-2670.
27. Guerrero-Preston, R., Hadar, T.A.L., Ostrow, K.L., Soudry, E., Echenique, M., Ili-Gangas, C., PÉRez, G., Perez, J., Brebi-Mieville, P., Deschamps, J. *et al.* (2014) Differential promoter methylation of kinesin family member 1a in plasma is associated with breast cancer and DNA repair capacity. *Oncology Reports*, **32**, 505-512.
28. Roessler, J., Ammerpohl, O., Gutwein, J., Steinemann, D., Schlegelberger, B., Weyer, V., Sariyar, M., Geffers, R., Arnold, N., Schmutzler, R. *et al.* (2014) The CpG island methylator phenotype in breast cancer is associated with the lobular subtype. *Epigenomics*, 1-13.
29. Sirnes, S., Honne, H., Ahmed, D., Danielsen, S.A., Rognum, T.O., Meling, G.I., Leithe, E., Rivedal, E., Lothe, R.A. and Lind, G.E. (2011) DNA methylation analyses of the connexin gene family reveal silencing of GJC1 (Connexin45) by promoter hypermethylation in colorectal cancer. *Epigenetics*, **6**, 602-609.
30. Gröschl, B., Bettstetter, M., Widmann, T., Hofstädter, F. and Dietmaier, W. (2014) Abstract 5004: Iqgap2 is downregulated in colorectal cancer (crc) and involved in cellular migration. *Cancer research*, **74**, 5004.
31. Jin, S.-H., Akiyama, Y., Fukamachi, H., Yanagihara, K., Akashi, T. and Yuasa, Y. (2008) IQGAP2 inactivation through aberrant promoter methylation and promotion of invasion in gastric cancer cells. *International Journal of Cancer*, **122**, 1040-1046.
32. Sawyer, J.R. (2011) The prognostic significance of cytogenetics and molecular profiling in multiple myeloma. *Cancer genetics*, **204**, 3-12.
33. Ross, F.M., Avet-Loiseau, H., Ameye, G., Gutierrez, N.C., Liebisch, P., O'Connor, S., Dalva, K., Fabris, S., Testi, A.M., Jarosova, M. *et al.* (2012) Report from the European Myeloma Network on interphase FISH in multiple myeloma and related disorders. *Haematologica*, **97**, 1272-1277.
34. Levine, J.J., Stimson-Crider, K.M. and Vertino, P.M. (2003) Effects of methylation on expression of TMS1/ASC in human breast cancer cells. *Oncogene*, **22**, 3475-3488.