SplicingFactory—splicing diversity analysis for transcriptome data

Benedek Dankó, Péter Szikora, Tamás Pór, Alexa Szeifert, Endre Sebestyén

2021

Students: Zahra Ahmad Nezhad, Samie Baniasadi

January 2022

What is splicingFactory?

They created an R package called SplicingFactory to meet different criteria,
They calculated RNA isoforms and examined them under different conditions.

SplicingFactory R :

The SplicingFactory R package uses transcript-level expression values to analyze splicing diversity based on various statistical measures, like Shannon entropy or the Gini index.

Items that have been tested using packages on a variety of isoforms:

- the effect of RNA-seq quantification tools
- quantification uncertainty
- gene expression levels
- isoform numbers

alternative splicing

- is well-known and described in most eukaryotic organisms
- expands the RNA repertoire of most genes

outline

Materials and methods

dataset processing

Comparison

Implementation

Results

Myelodysplastic syndrome dataset processing

- data from SRA (SRP133442, SRP149374) using SRA-tools
- they quantified transcript-level expression using Kallisto and Salmon and the full GENCODE v34 transcriptome annotation.
- Salmon-SAF was run both in alignment-based using the STAR alignments and mapping-based mode using a decoy-aware transcriptome index.
- they run the tool with 100 bootstraps, the unstranded paired-end library option and the additional -seqBias and -gcBias parameters.

Comparison of Kallisto, Salmon and Salmon-SAF results

- they selected the 17 control samples from the SRP133442 dataset.
- calculated the normalized naive-entropy and Gini-index for all genes.
- they selected all genes with non-NA diversity values and calculated their Spearman correlation.

Performance benchmarks

- using a single core on a server with Intel® Xeon® Gold 4118 2.30 GHz CPU type, a total of 157 GB memory and Ubuntu 18.04.5 LTS operating system.
- they used the Salmon-based quantification for the benchmarks.
- Using a fixed number of 60 669 starting genes, they increased the sample number from 10 to 130 using steps of 10, 20, 40, 60, 80, 100 and 130 for the same calculations.

Comparison to other tools

splicehetero

- ran SpliceHetero with parameters -slb True and -prn 17
- To prepare the input data keeping only those splice junctions that appear in the STAR index sjdbList.fromGTF.out.tab file
- converted the filtered STAR junction files
- ran SpliceHetero on these sorted BED files

SEVA

- ran SEVA in RStudio Server to FALSE
- To prepare the input data we first performed RPM normalization on the STAR splice junction files

Comparison to other tools

whippet

- ran Whippet with default parameters
- To prepare the input data, we converted the raw fastq files to have only a single '+' character on every third line
- using Whippet's whippet-index.jl script, created an index file based on the GENCODE v34 transcriptome annotation gtf file, and ran Whippet's quantification algorithm with the whippet-quant.jl script with the additional -biascorrect parameter

SF3B1 differential diversity and enrichment analysis

analyzed the SRP149374 dataset

compared the MDS samples without any known mutation to MDS samples with known *SF3B1* somatic mutations

calculated the normalized naive-entropy, normalized Laplace-entropy and Gini-index using the Salmon-based quantification

compared the *SF3B1* mutated and *SF3B1* wild-type samples using Wilcoxon-test and performed *P*-value adjustment with the Benjamini–Hochberg method.

performed enrichment analysis using the significant genes (|mean difference| >0.1 and adjusted *P*-value <0.05) separately with either a mean diversity increase or decrease between the *SF3B1* mutant and *SF3B1* wild-type groups

used the bone marrow marker gene sets originating

Implementation

- the package calculates a diversity value for each gene in each sample, using splicing isoform expression values.

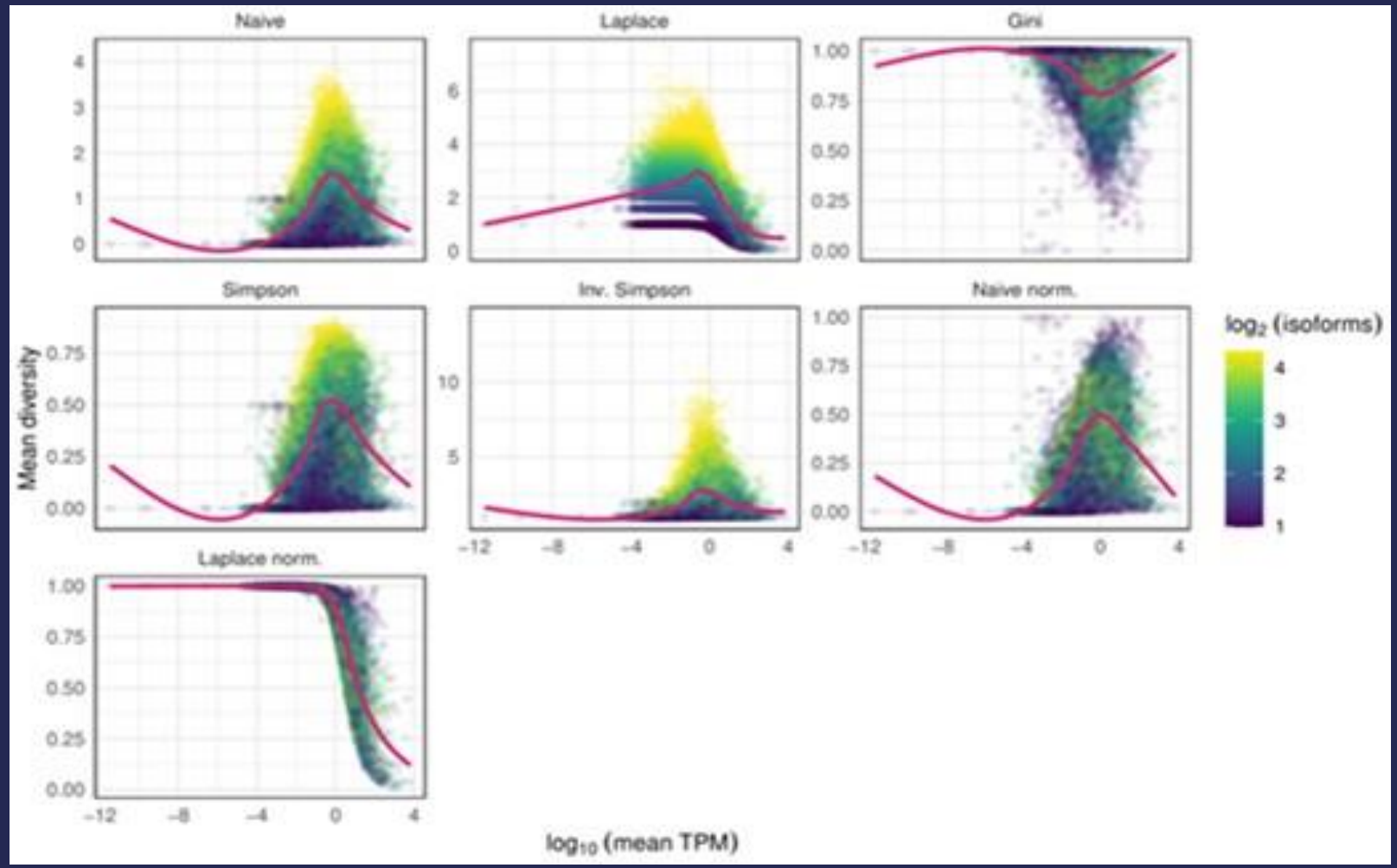
- calculates differential diversity results between conditions.

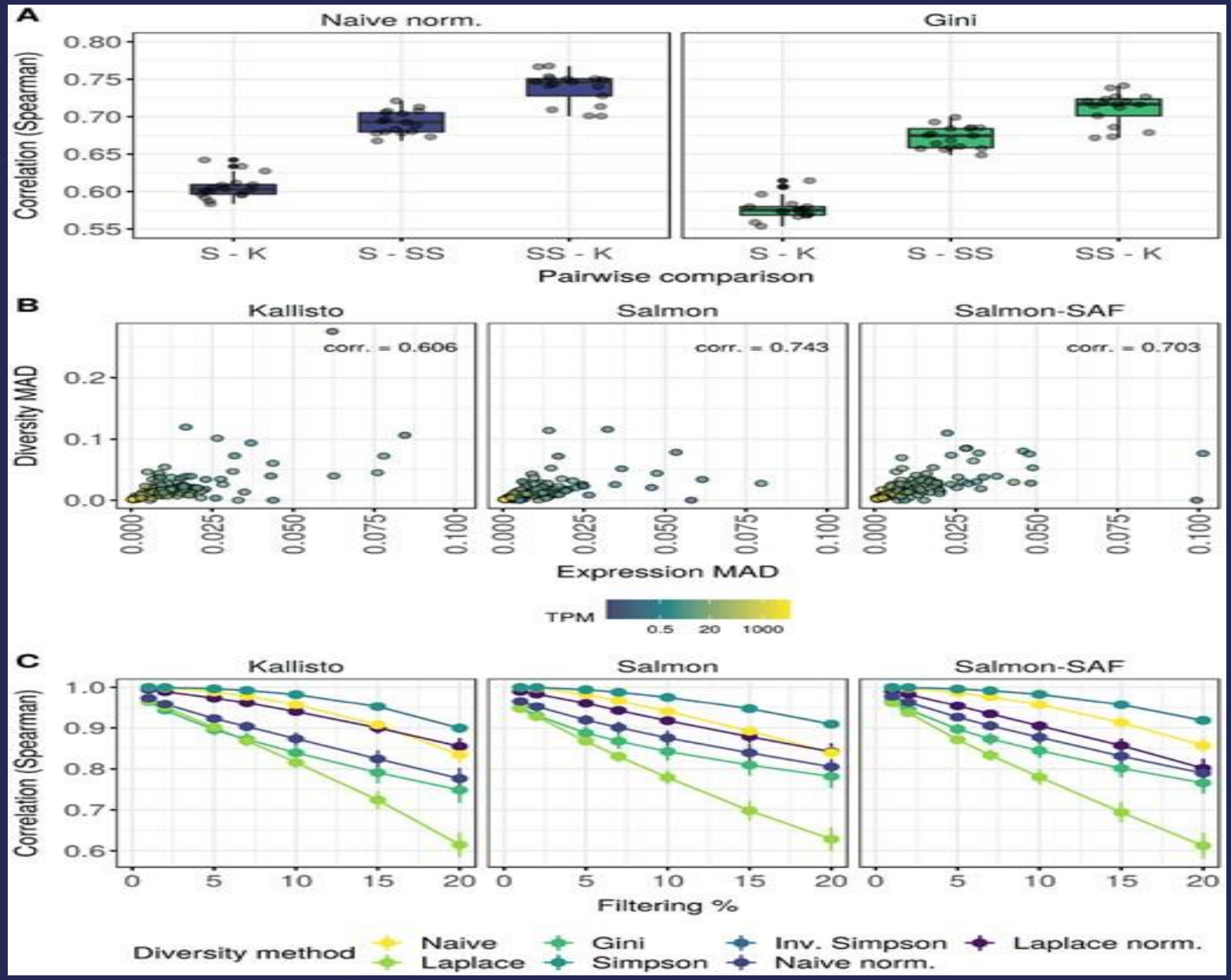
Association of diversity metrics

The Gini-index shows an opposing pattern compared to all others.

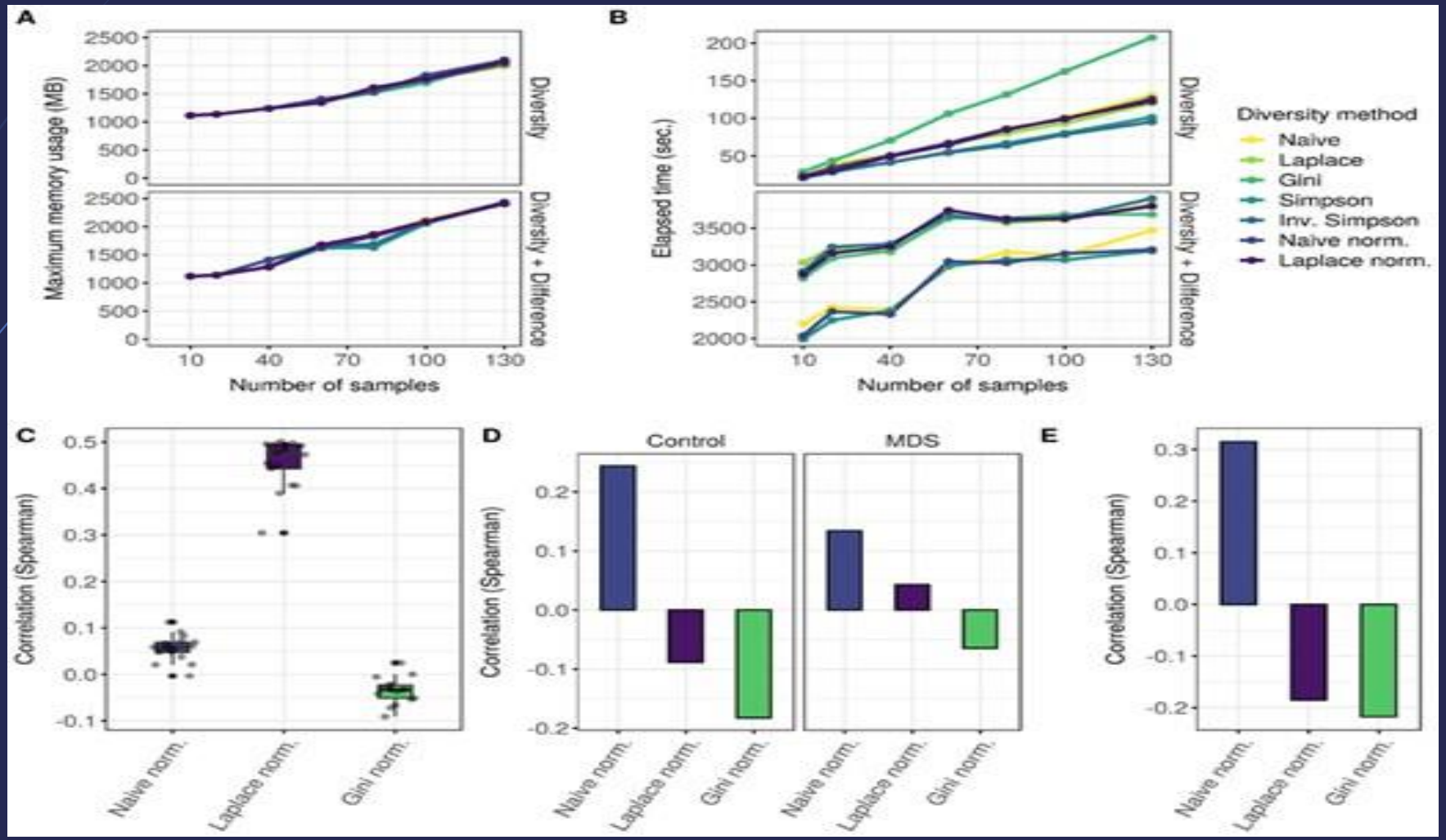
The nonnormalized naive- and Laplace-entropy, besides the Simpson-index and the inverse Simpson-index shows a clear association with isoform number.

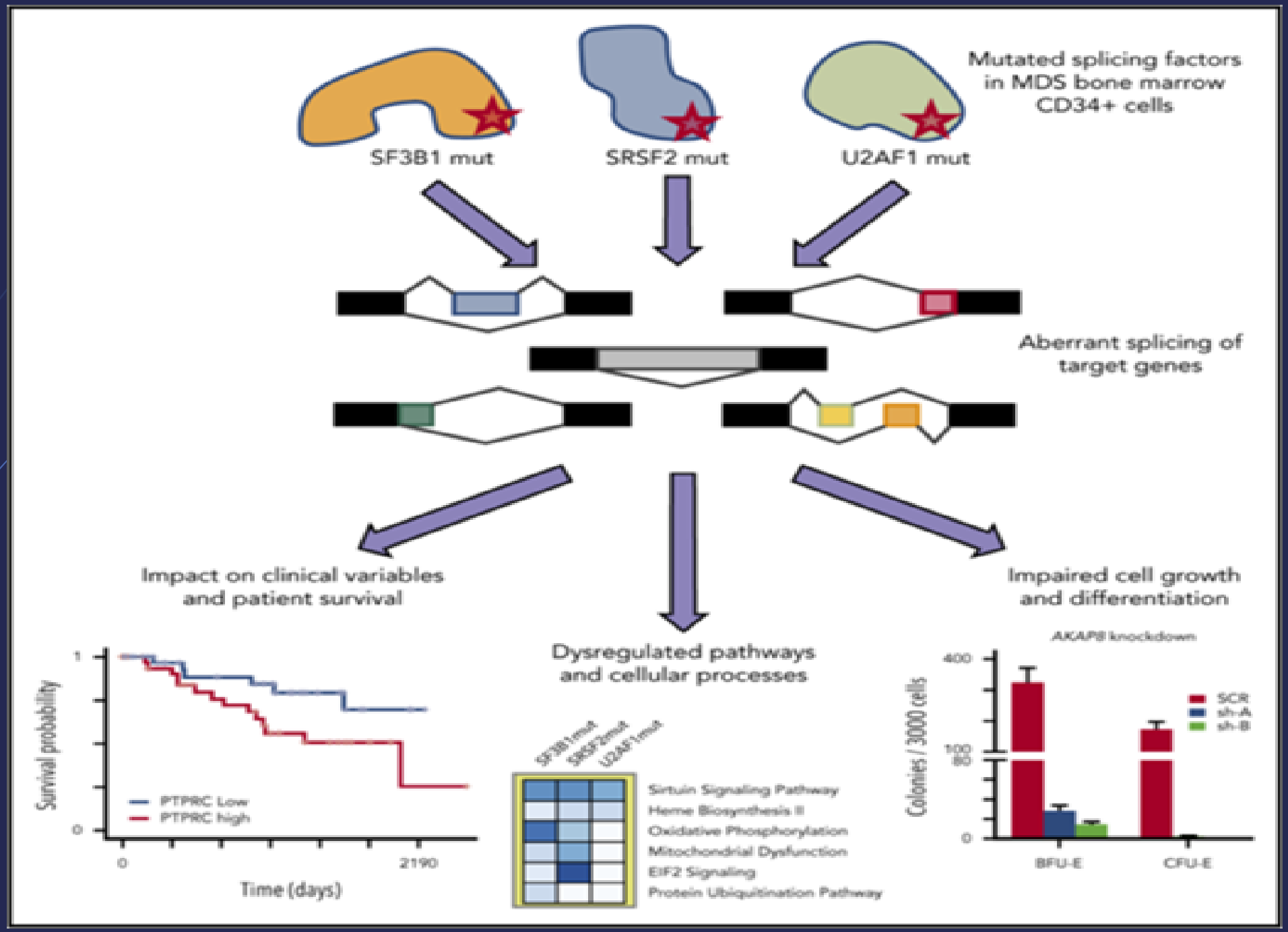
the strongly biased pattern for the normalized Laplace-entropy is again the result of the +1 pseudocount, as all genes with only zero or very low isoform expression levels are given an expression of ~1, leading to the maximum possible entropy of 1.

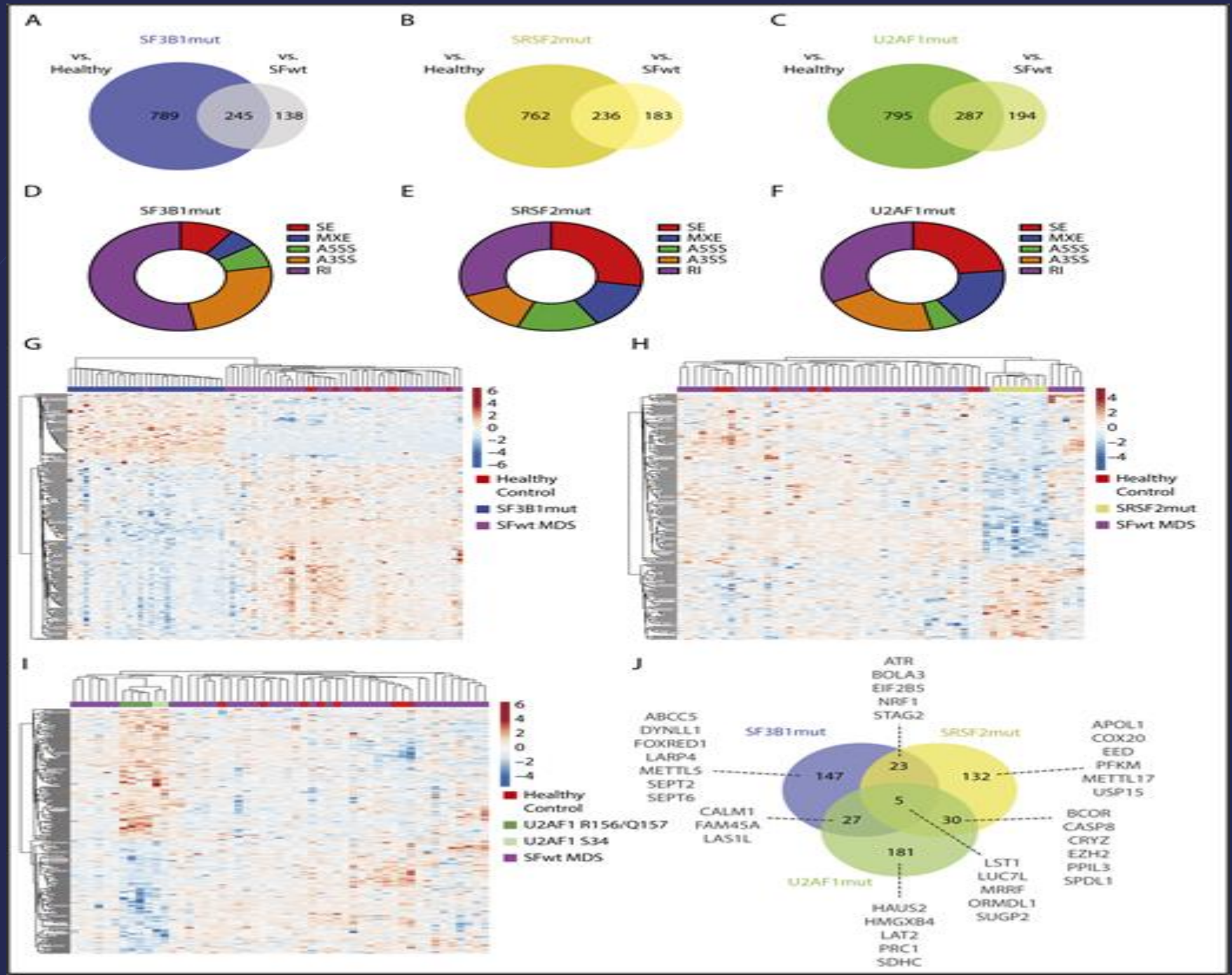

Naive
Laplace
Gini
Simpson
Inv. Simpson
Naive norm.
Laplace norm.
Mean diversity
log10 (mean TPM)
log2 (isoforms)

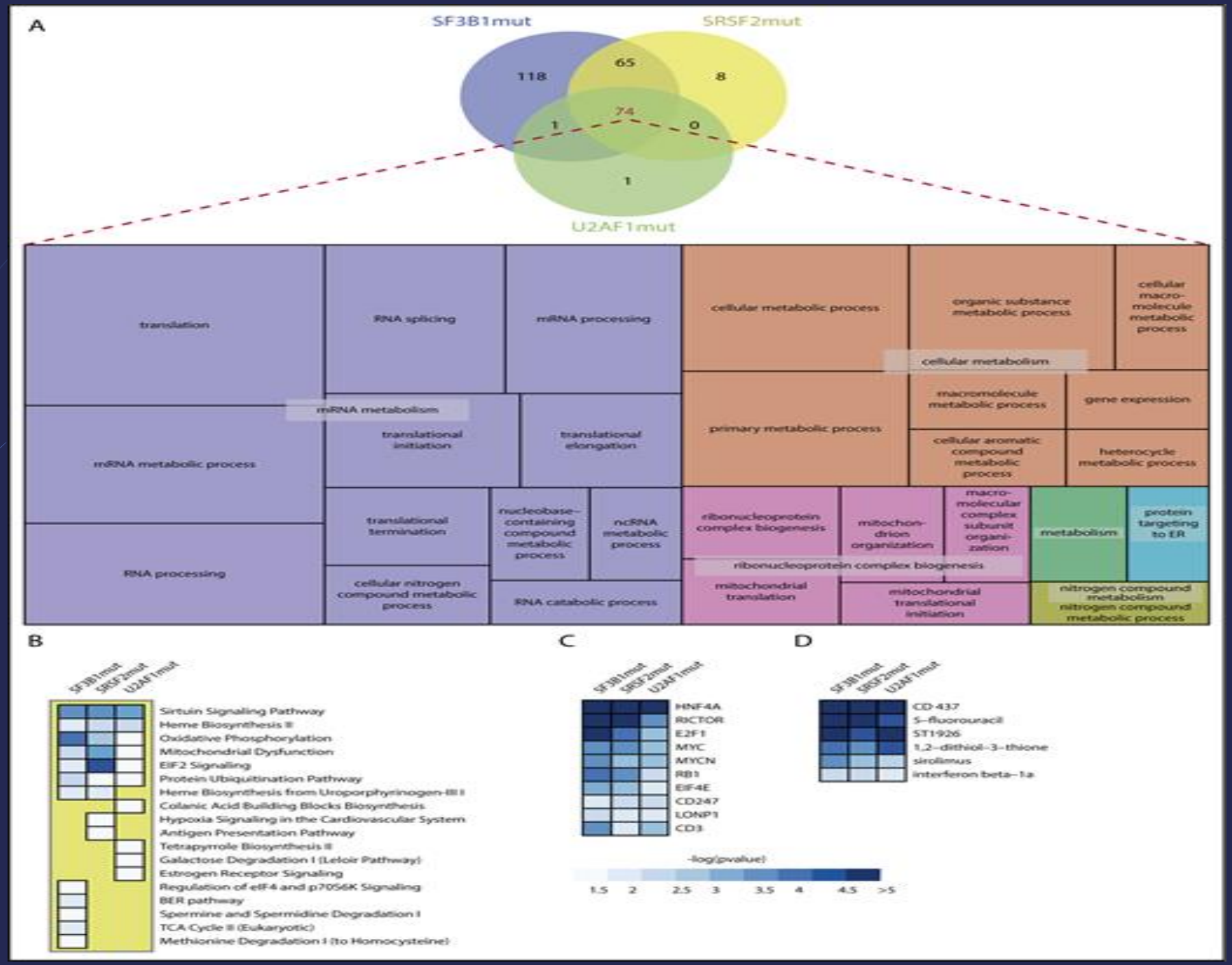

A
Naive norm.
Gini
Correlation (Spearman)
0.80
0.75
0.70
0.65
0.60
0.55
S - K
S - SS
SS - K
Pairwise comparison
B
Kallisto
Salmon
Salmon-SAF
corr. = 0.606
corr. = 0.743
corr. = 0.703
Diversity MAD
0.2
0.1
0.0
0.000
0.025
0.050
0.075
0.100
Expression MAD
TPM
0.5
20
1000
C
Kallisto
Salmon
Salmon-SAF
Correlation (Spearman)
1.0
0.9
0.8
0.7
0.6
0
5
10
15
20
Filtering %
Diversity method
Naive
Laplace
Gini
Simpson
Inv. Simpson
Naive norm.
Laplace norm.

Performance benchmarks




Mutated splicing factors in MDS bone marrow CD34+ cells
SF3B1 mut
SRSF2 mut
U2AF1 mut
Aberrant splicing of target genes
Impact on clinical variables and patient survival
Dysregulated pathways and cellular processes
Impaired cell growth and differentiation
Survival probability
1
0
PTPRC Low
PTPRC high
0
2190
Time (days)
Sirtuin Signaling Pathway
Heme Biosynthesis II
Oxidative Phosphorylation
Mitochondrial Dysfunction
EIF2 Signaling
Protein Ubiquitination Pathway
AKAP8l knockdown
Colonies / 3000 cells
400
100
80
0
SCR
sh-A
sh-B
BFU-E
CFU-E


A
SF3B1mut
vs. Healthy
vs. SFwt
789
245
138
B
SRSF2mut
vs. Healthy
vs. SFwt
762
236
183
C
U2AF1mut
vs. Healthy
vs. SFwt
795
287
194
D
SF3B1mut
E
SRSF2mut
F
U2AF1mut
SE
MXE
A5SS
A3SS
RI
G
Healthy Control
SF3B1mut
SFwt MDS
H
Healthy Control
SRSF2mut
SFwt MDS
I
Healthy Control
U2AF1 R156/Q157
U2AF1 S34
SFwt MDS
J
ATR
BOLA3
EIF2BS
NRF1
STAG2
ABCC5
DYNLL1
FOXRED1
LARP4
METTL5
SEPT2
SEPT6
SF3B1mut
SRSF2mut
APOL1
COX20
EED
PFKM
METTL17
USP15
147
23
132
5
CALM1
FAM45A
LAS1L
27
30
BCOR
CASP8
CRYZ
EZH2
PPIL3
SPDL1
181
LST1
LUC7L
MRRF
ORMDL1
SUGP2
U2AF1mut
HAUS2
HMGXB4
LAT2
PRC1
SDHC


A
SF3B1mut
SRSF2mut
U2AF1mut
118
65
8
74
1
0
1
translation
RNA splicing
mRNA processing
mRNA metabolism
mRNA metabolic process
translational initiation
translational elongation
RNA processing
translational termination
nucleobase-containing compound metabolic process
ncRNA metabolic process
cellular nitrogen compound metabolic process
RNA catabolic process
cellular metabolic process
organic substance metabolic process
cellular macro-molecule metabolic process
cellular metabolism
primary metabolic process
macromolecule metabolic process
gene expression
cellular aromatic compound metabolic process
heterocycle metabolic process
ribonucleoprotein complex biogenesis
mitochon-drion organization
macro-molecular complex subunit organi-zation
metabolism
protein targeting to ER
ribonucleoprotein complex biogenesis
mitochondrial translation
mitochondrial translational initiation
nitrogen compound metabolism
nitrogen compound metabolic process
B
SF3B1mut
SRSF2mut
U2AF1mut
Sirtuin Signaling Pathway
Heme Biosynthesis II
Oxidative Phosphorylation
Mitochondrial Dysfunction
EIF2 Signaling
Protein Ubiquitination Pathway
Heme Biosynthesis from Uroporphyrinogen-III I
Colanic Acid Building Blocks Biosynthesis
Hypoxia Signaling in the Cardiovascular System
Antigen Presentation Pathway
Tetrapyrrole Biosynthesis II
Galactose Degradation I (Leloir Pathway)
Estrogen Receptor Signaling
Regulation of eIF4 and p70S6K Signaling
BER pathway
Spermine and Spermidine Degradation I
TCA Cycle II (Eukaryotic)
Methionine Degradation I (to Homocysteine)
C
SF3B1mut
SRSF2mut
U2AF1mut
HNF4A
RICTOR
E2F1
MYC
MYCN
RB1
EIF4E
CD247
LONP1
CD3
-log(pvalue)
1.5
2
2.5
3
3.5
4
4.5
>5
D
SF3B1mut
SRSF2mut
U2AF1mut
CD437
5-fluorouracil
ST1926
1,2-dithiol-3-thione
sirolimus
interferon beta-1a

List of aberrant splicing events, identified in *SF3B1*mut MDS, significantly correlated with a clinical variable

Event ID	Gene	Event type	Chr	Strand	Start position	End position	Spearman correlation variable	cor_estimate	*P*	adj.*P*
6795	PARVG	A3SS	22	+	44582456	44583758	BM blast %	0.58	4.38E-07	.003031
5420	RPRD1A	RI	18	–	33605560	33607038	BM blast %	0.58	6.21E-07	.004292
3146	DOM3Z	RI	6	–	31938382	31938924	ANC	−0.57	1.00E-06	.006916
12280	CXXC1	SE	18	–	47811694	47811721	ANC	−0.56	1.28E-06	.00888

3305	**AP1 G2**	**RI**	**14**	**–**	**24031170**	**24031624**	**ANC**	**−0.56**	**1.68E-06**	**.011579**
2362	SNRPN	A3SS	15	+	25219434	25219603	ANC	−0.55	3.14E-06	.021689
5460	TCEA2	A3SS	20	+	62703210	62703294	ANC	0.54	3.24E-06	.022393
6627	NICN1	A3SS	3	–	49462381	49462579	Plt	−0.54	3.80E-06	.026281
6153	ABCC5	A3SS	3	–	183703091	183703243	ANC	−0.54	4.23E-06	.029229
518	ERCC3	RI	2	–	128046912	128047400	ANC	−0.54	4.71E-06	.032551
2359	SNRPN	A3SS	15	+	25219457	25219603	ANC	−0.54	5.02E-06	.03468

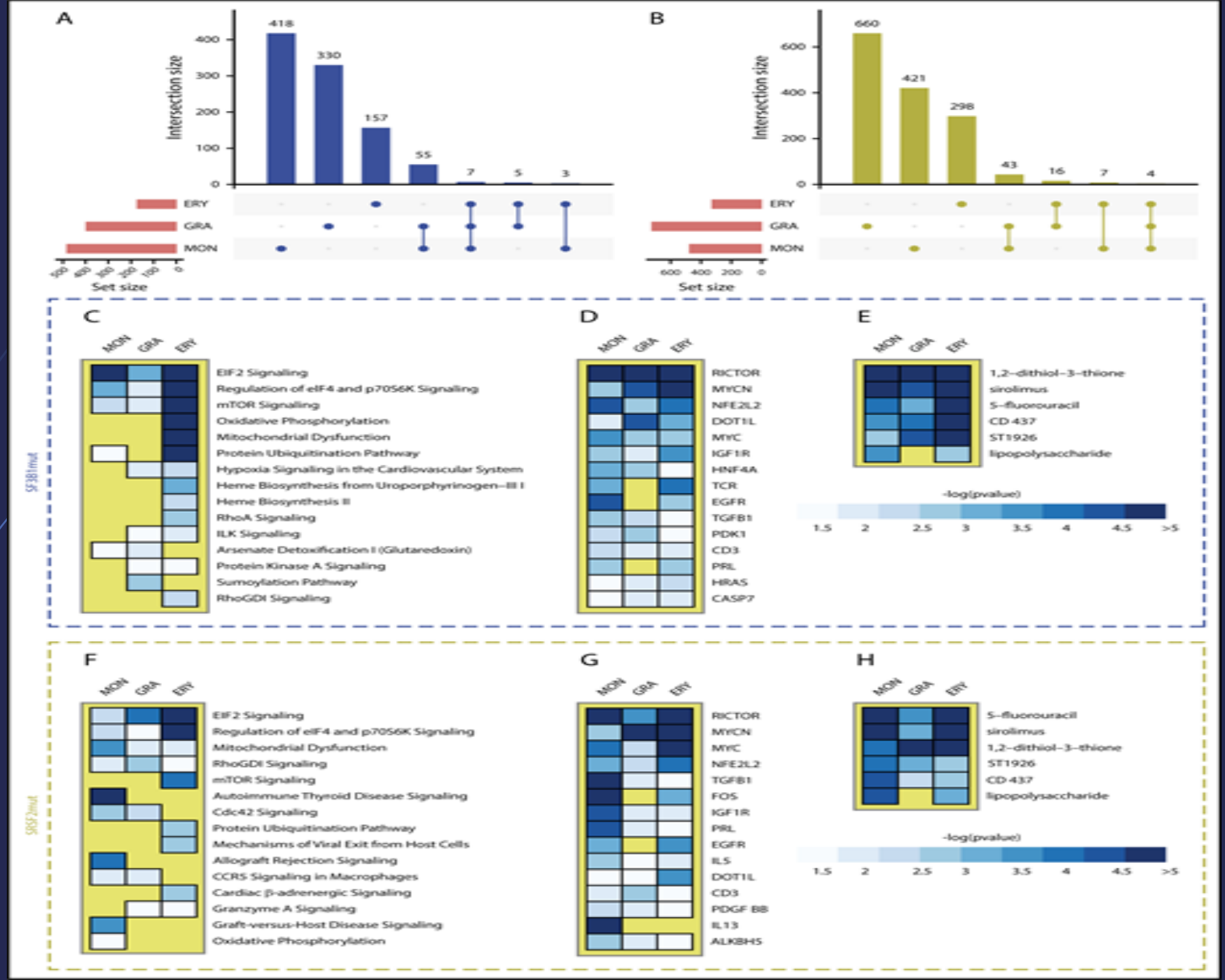
7563	**PPOX**	**A3SS**	**1**	**+**	**161137128**	**161137276**	**ANC**	**0.53**	**5.99E-06**	**.041411**
4975	GPR108	A3SS	19	–	6730997	6731122	ANC	−0.53	6.43E-06	.044443
5816	PSTPIP1	A3SS	15	+	77328142	77328276	ANC	−0.53	6.49E-06	.044832
4728	NICN1	RI	3	–	49462381	49462871	Plt	−0.53	6.75E-06	.046655

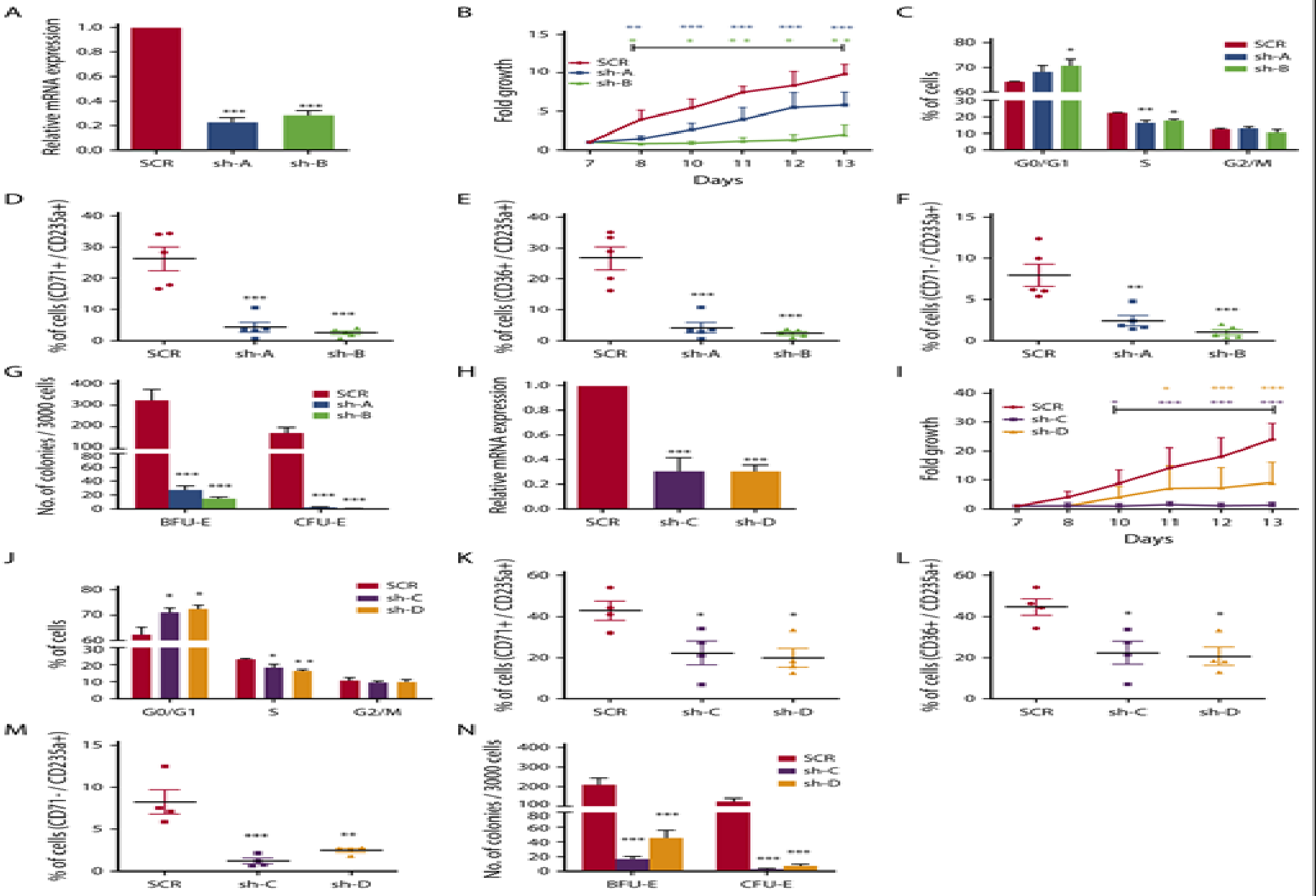
Genes with isoforms that significantly predicted survival in MDS in multivariate models

Gene	Associated splice factor mutation or mutations	Function/pathway	*P* (multivariate survival)	
CAP1	SRSF2	Focal adhesion & extracellular exosomes	.0044	
PTPRC	U2AF1 (S34)	Focal adhesion & extracellular exosomes	.0093	
IFI44	SRSF2, U2AF1 (S34)	Interferon, extracellular exosomes	.012	
IFI44L	U2AF1 (S34)	Interferon, extracellular exosomes	.0086	

CD46	**U2AF1 (S34)**	**Focal adhesion & extracellular exosomes**	**.039**
CRTC2	SF3B1	Extracellular exosomes	.035
FCGR2A	U2AF1 (S34)	Extracellular exosomes	.016
PPOX	SF3B1, U2AF1 (R156/Q157)	Heme biosynthesis	.031
AHSA2	SF3B1	HSP90 ATPase	.029
DHP5	SF3B1	Translation elongation factor 2 modification	.026
MECR	U2AF1 (S34 & R156/Q157)	Mitochondrial reductase	.022

NASP	U2AF1 (S34)	HSP90 binding	.014
PFDN5	U2AF1 (R156/Q157)	Prefoldin subunit	.042
PABPC4	U2AF1 (R156/Q157)	NMD mRNA decay	.036


A
Intersection size
418
330
157
55
7
5
3
ERY
GRA
MON
Set size
B
Intersection size
660
421
298
43
16
7
4
ERY
GRA
MON
Set size
C
MON
GRA
ERY
SF3B1mut
EIF2 Signaling
Regulation of eIF4 and p70S6K Signaling
mTOR Signaling
Oxidative Phosphorylation
Mitochondrial Dysfunction
Protein Ubiquitination Pathway
Hypoxia Signaling in the Cardiovascular System
Heme Biosynthesis from Uroporphyrinogen-III I
Heme Biosynthesis II
RhoA Signaling
ILK Signaling
Arsenate Detoxification I (Glutaredoxin)
Protein Kinase A Signaling
Sumoylation Pathway
RhoGDI Signaling
D
MON
GRA
ERY
RICTOR
MYCN
NFE2L2
DOT1L
MYC
IGF1R
HNF4A
TCR
EGFR
TGFB1
PDK1
CD3
PRL
HRAS
CASP7
E
MON
GRA
ERY
1,2-dithiol-3-thione
sirolimus
5-fluorouracil
CD437
ST1926
lipopolysaccharide
-log(pvalue)
1.5
2
2.5
3
3.5
4
4.5
>5
F
MON
GRA
ERY
SRSF2mut
EIF2 Signaling
Regulation of eIF4 and p70S6K Signaling
Mitochondrial Dysfunction
RhoGDI Signaling
mTOR Signaling
Autoimmune Thyroid Disease Signaling
Cdc42 Signaling
Protein Ubiquitination Pathway
Mechanisms of Viral Exit from Host Cells
Allograft Rejection Signaling
CCR5 Signaling in Macrophages
Cardiac β-adrenergic Signaling
Granzyme A Signaling
Graft-versus-Host Disease Signaling
Oxidative Phosphorylation
G
MON
GRA
ERY
RICTOR
MYCN
MYC
NFE2L2
TGFB1
FOS
IGF1R
PRL
EGFR
IL5
DOT1L
CD3
PDGF BB
IL13
ALKBH5
H
MON
GRA
ERY
5-fluorouracil
sirolimus
1,2-dithiol-3-thione
ST1926
CD437
lipopolysaccharide
-log(pvalue)
1.5
2
2.5
3
3.5
4
4.5
>5


A
Relative mRNA expression
SCR
sh-A
sh-B
B
Fold growth
Days
C
% of cells
G0/G1
S
G2/M
D
% of cells (CD71+ / CD235a+)
E
% of cells (CD36+ / CD235a+)
F
% of cells (CD71+ / CD235a+)
G
No. of colonies / 3000 cells
BFU-E
CFU-E
H
Relative mRNA expression
sh-C
sh-D
I
Fold growth
J
% of cells
K
% of cells (CD71+ / CD235a+)
L
% of cells (CD36+ / CD235a+)
M
% of cells (CD71+ / CD235a+)
N
No. of colonies / 3000 cells

Thank you