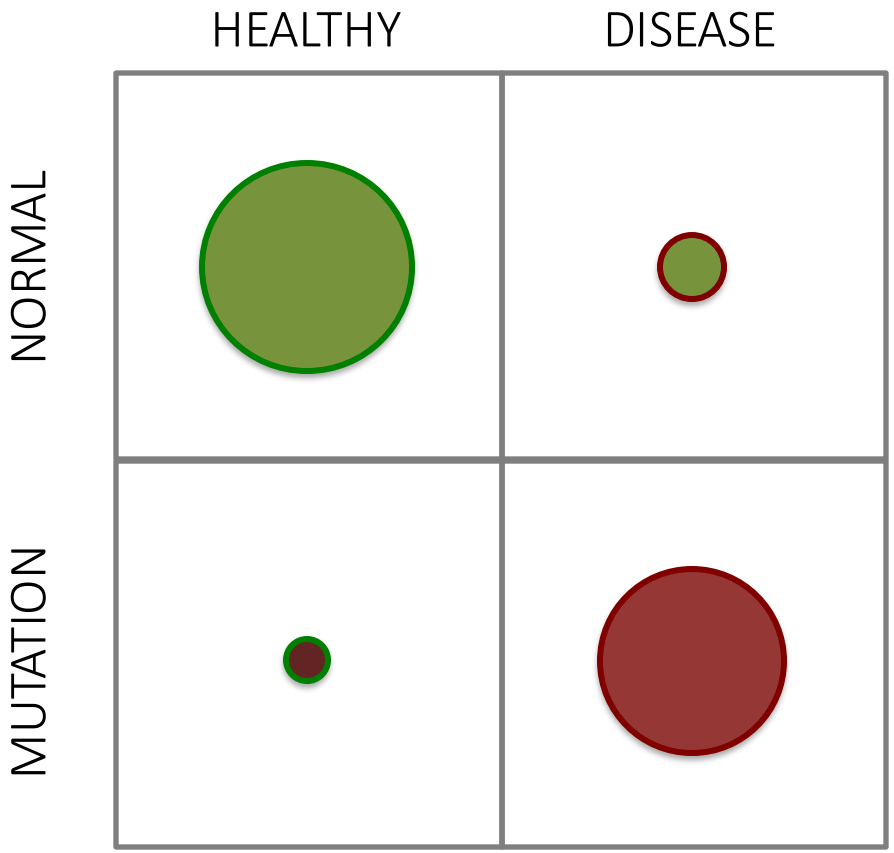
Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Gerstung et al 2015. Nature Communications

INTRODUCTION

17 figures:

- Patterns of Mutation and differential expression
- Genetic mechanisms of differential expression
- Prediction of blood and marrow counts
- Prediction of survival



e.g. CFTRΔF508

deletion of three nucleotides spanning positions 507 and 508 of the CFTR gene on chromosome 7,

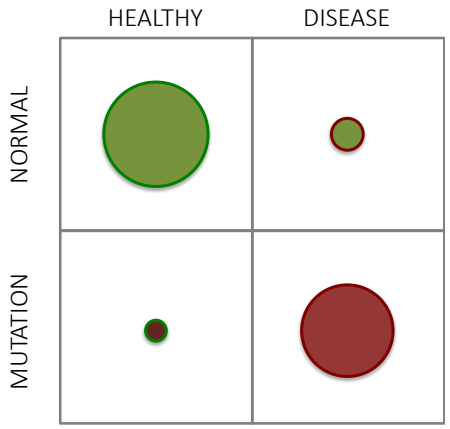
→ loss of the codon for phenylalanine (F).

The CFTRΔF508 mutation produces an abnormal CFTR protein that cannot fold properly and which does not escape the endoplasmic reticulum for further processing.

Having two copies of this mutation (one inherited from each parent) is by far the most common cause of cystic fibrosis (CF), responsible for nearly two-thirds of cases worldwide

- previously, cancer diagnosis would be at a high level, e.g. resolved to the organ level.
- improvement in imaging techniques has improved detection and diagnosis
- at the genetic level, biomarkers have been identified that can identify pre-disposition
- e.g. BreastNextTM gene panel- sequence 17 genes (ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, MRE11A, MUTYH, NBN, NF1, PALB2, PTEN, RAD50, RAD51C, RAD51D, and TP53)

This panel can identify a correlation with pre-disposition to breast cancer, but doesn't address causality



Once we start investigating multiple genes/polygenic disease, the effects can be less obvious and the analysis becomes more complicated.

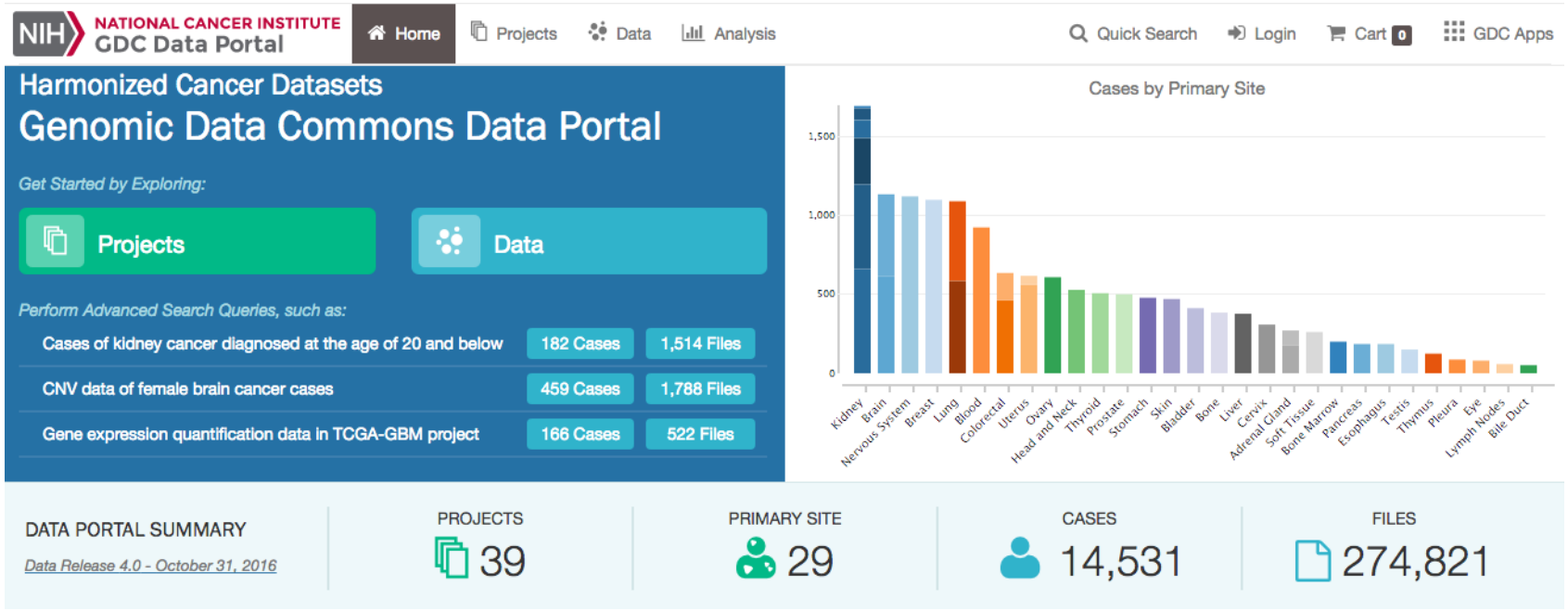
If we add in the effect of a spectrum of phenotypes, then things become even more complex.

Now, we have to identify meaningful associations across both genotype and phenotype spectrums. This is what the Gerstung paper is trying to address.

INTRODUCTION

There are information sources that attempt to link genotype, phenotype and biology.

For example, the GDCD portal



But this doesn’t address prognosis or response to treatment

But there are efforts to do this as well

Mutations → genes → pathways

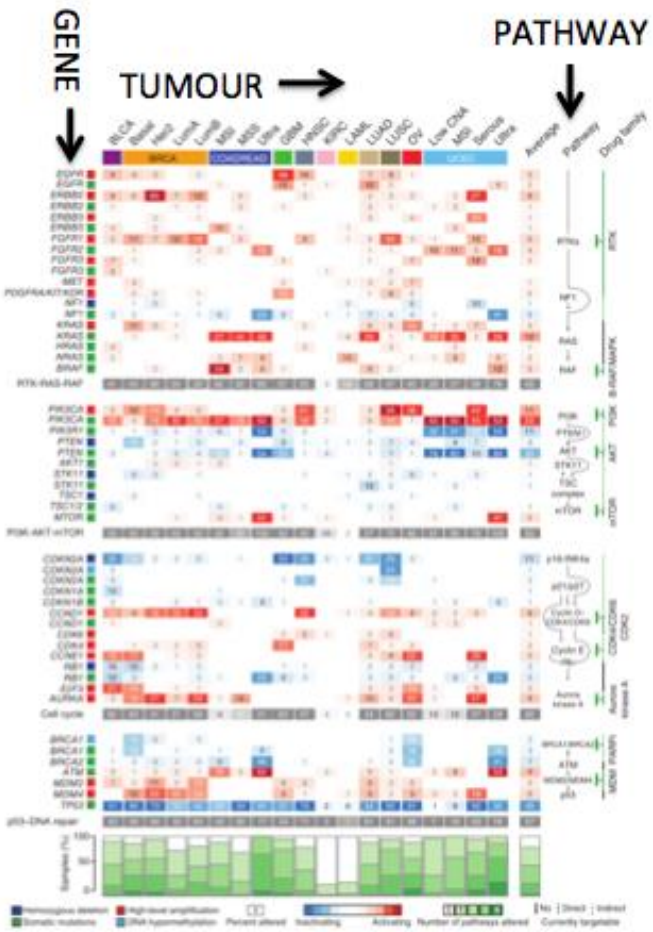


Figure 4 Map of functional and actionable alterations across 12 tumor types. Genes (rows) encoding components of four major oncogenic pathways (RTK-RAS-RAF, PI3K-AKT-mTOR, cell cycle and p53–DNA repair; shown schematically in the pathway column) are affected by selected functional events (percent of samples altered and types of alteration are represented by colored squares) across tissue-specific tumor types (columns). Alterations to at least one of these pathways are observed in almost all samples of almost all tumor types (stacked green bar plots at bottom), except in KIRC and LAML. A sizable fraction of these alterations are directly or indirectly therapeutically actionable given the current availability of anticancer drugs (the column with drug family information shows the targets of specific inhibitors). Tumor types abbreviated as in **Table 1**.

nature genetics
OPEN

Emerging landscape of oncogenic signatures across human cancers

Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz & Chris Sander

DATA

PHENOTYPE

a range of clinical parameters (primarily based on analysis of blood) that are commonly used as markers

✚

Demographics

GENETIC

Sequence 111 genes in 738 MDS patients

⇩

12 key genes strongly associated with MDS

✚

4 cytogenetic abnormalities overrepresented in MDS

✚

Microarray data from 159 MDS patients + 17 controls

The 12 genes were identified in an earlier study

The genetic basis of phenotypic heterogeneity in myelodysplastic syndromes

Azra Raza and Naomi Galili

Abstract | Myelodysplastic syndromes (MDS) are malignant clonal disorders of haematopoietic stem cells and their microenvironment, affecting older individuals (median age ~70 years). Unique features that are associated with MDS — but which are not necessarily present in every patient with MDS — include excessive apoptosis in maturing clonal cells, a pro-inflammatory bone marrow microenvironment, specific chromosomal abnormalities, abnormal ribosomal protein biogenesis, the presence of uniparental disomy, and mutations affecting genes involved in proliferation, methylation and epigenetic modifications. Although emerging insights establish an association between molecular abnormalities and the phenotypic heterogeneity of MDS, their origin and progression remain enigmatic.

Leukemia (2010) 24, 756–764

www.nature.com/leu

ORIGINAL ARTICLE

Deregulated gene expression pathways in myelodysplastic syndrome hematopoietic stem cells

A Pellagatti[1], M Cazzola[2], A Giagounidis[3], J Perry[1], L Malcovati[2], MG Della Porta[2], M Jädersten[4], S Killick[5], A Verma[6], CJ Norbury[7], E Hellström-Lindberg[4], JS Wainscoat[1] and J Boultwood[1]

[1]LRF Molecular Haematology Unit, NDCLS, John Radcliffe Hospital, Oxford, UK; [2]Department of Hematology Oncology, University of Pavia Medical School, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; [3]Medizinische Klinik II, St Johannes Hospital, Duisburg, Germany; [4]Division of Hematology, Department of Medicine, Karolinska Institutet, Stockholm, Sweden; [5]Department of Haematology, Royal Bournemouth Hospital, Bournemouth, UK; [6]Albert Einstein College of Medicine, Bronx, NY, USA and [7]Sir William Dunn School of Pathology, University of Oxford, Oxford, UK

To gain insight into the molecular pathogenesis of the myelodysplastic syndromes (MDS), we performed global gene expression profiling and pathway analysis on the hematopoietic stem cells (HSC) of 183 MDS patients as compared with the HSC of 17 healthy controls. The most significantly deregulated pathways in MDS include interferon signaling, thrombopoietin signaling and the Wnt pathways. Among the most significantly deregulated gene pathways in early MDS are immunodeficiency, apoptosis and chemokine signaling, whereas advanced MDS is characterized by deregulation of DNA damage response and checkpoint pathways. We have identified distinct gene expression profiles and deregulated gene pathways in patients with del(5q), trisomy 8 or −7/del(7q). Patients with trisomy 8 are characterized by deregulation of pathways involved in the immune response, patients with −7/del(7q) by pathways

the World Health Organization.[6,7] Patients with refractory anemia (RA) with or without ringed sideroblasts, according to the French–American–British classification, were subdivided based on the presence or absence of multilineage dysplasia. In addition, patients with RA with excess blasts (RAEB) were subdivided into two categories, RAEB1 and RAEB2, based on the percentage of bone marrow blasts. Patients with >20% blasts were classified as AML. MDS patients with <5% bone marrow blasts and an isolated deletion of the long arm of chromosome 5 [del(5q)] were categorized as 5q– syndrome, and a new category termed as MDS-unclassifiable was created incorporating patients not fulfilling the other criteria. The main prognostic factors of MDS for progression to AML and survival include

Nature Reviews Cancer 12, 849-859 (December 2012) | doi:10.1038/nrc3321

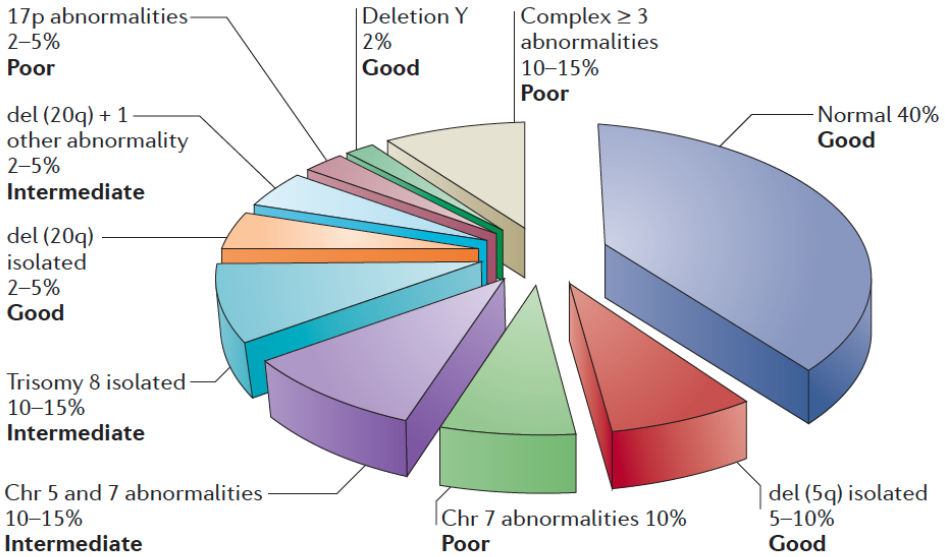


Figure 1 | **Incidence and prognosis of common chromosomal abnormalities in MDS.** The most common chromosomal abnormalities are shown in the figure with the proportions of total myelodysplastic syndromes (MDS) cases represented as wedges. The percentages and International Prognostic Scoring System (IPSS) predictions are included. Chr, chromosome.

Table 2 | **Common gene mutations in MDS**

Genes	Chromosomal location	Frequency in MDS	Prognosis (with refs)	Function of encoded protein
Oncogenes and tumour suppressors				
RUNX1	21q22	15%	Poor[81,82]	Core binding transcription factor important in haematopoiesis
TP53	17p13	5–10%	Poor[82,83]	Tumour suppressor activity regulating downstream target genes involved in cell cycle arrest, apoptosis, senescence, DNA repair and metabolism
NRAS	1p13.2	10%	Poor for lower-risk[82,84]	GTPase that functions as an oncogene when mutated and constitutively active
KRAS	12p12.1	2–5%	Poor[82,84]	GTPase that functions as an oncogene when mutated and constitutively active
ETV6	12p13.2	2–5%	Poor[82]	ETS transcription factor required for haematopoiesis and maintenance of developing vascular network
EVI1	3q26	1–2%	Poor[84]	Transcriptional regulator and oncoprotein that may be involved in haematopoiesis, apoptosis, development, differentiation and proliferation
Methylation of CpG islands				
TET2	4q24	20%	Unclear[63–65]	Methylcytosine dioxygenase that converts methylcytosine to 5-hydroxymethylcytosine; required for myelopoiesis
IDH1 and *IDH2*	2q33.3 and 15q26.1	5–10%	Unknown	Isocitrate dehydrogenase converts isocitrate to α-ketoglutarate; regulates TET2 activity
DNMT3A	2p23.3	5–10%	Poor[58,59]	DNA methyltransferase that functions in *de novo* methylation and is coordinated with histone methylation to repress transcription
Histone modification				
ASXL1	20q11.2	10–15%	Poor[72,73,82]	Histone-binding protein that disrupts chromatin in localized areas to enhance or repress transcription
EZ2H	7q36.1	5%	Poor[75,82,89]	Histone methyltransferase that helps maintain gene repression
Spliceosome				
SF3B1	2q33.1	20% (in MDS) and 65% (in MDS-RS)	Better for lower-risk[89]; no difference for ringed sideroblasts[97–103]	Spliceosome protein component essential for spliceosome complex assembly
SRSF2	17q25.1	Not known	Poor[103]	Spliceosome protein component essential for spliceosome complex assembly
U2AF1	21q22.3	Not known	None[100,103]	Spliceosome protein component essential for spliceosome complex assembly
ZRSR2	Xp22.1	Not known	None[103]	Spliceosome protein component essential for spliceosome complex assembly
Others				
JAK2	9p24.1	50% in RARS-T	Unknown	Non-receptor protein tyrosine kinase with roles in cell cycle, genomic instability, apoptosis and mitotic recombination
CBL	11q23.3	2–5%	Unknown	E3 ubiquitin ligase that participates as a negative regulator of transduction in haematopoietic cells
RPS14	5q33.1	5q⁻ syndrome	Unknown	Ribosomal protein of the 40S subunit

ASXL1, additional sex-combs-like 1; *DNMT3A*, DNA methyltransferase 3A; *ETV6*, ETS variant 6; *EZ2H*, enhancer of zeste homologue 2; IDH, isocitrate dehydrogenase; *JAK2*, Janus kinase 2; MDS, myelodysplastic syndromes; RARS-T, refractory anaemia with ringed sideroblasts and thrombocytosis; *RPS14*, ribosomal protein S14; RS, ringed sideroblasts; *RUNX1*, runt-related transcription factor 1; *SF3B1*, splicing factor 3B subunit 1; SRSF2, serine/arginine-rich splicing factor 2; *TET2*, tet methylcytosine dioxygenase 2; *U2AF1*, U2 small nuclear RNA auxiliary factor 1; *ZRSR2*, zinc finger (CCCH type) RNA-binding motif and serine/arginine-rich 2.

STAG2?

Is it possible to deconvolute the gene expression data into contributions from each genetic and cytogenetic mutation?

PDID	PD6175a
GEOID	GSM1420393
File	GSM1420393_MDS009.CEL
Described in Papaemmanuil et al, Blood 2013	yes
Described in Pellagatti et al, Leukemia 2010	yes
Type	MDS
Gender	1
Age	76
WHO_category	RA
Survival_days	2575
Status	0
AML_progression_days	NA
AML_status	NA
Karyotype	46, XX, t(1;3)(p33:p14), del(5)(q14;q34)[21] / 46, XX [4]
Cytogenetic_risk	1
IPSS	int-1
WPSS	NA
Transfusion_dep	NA
Serum_ferritin	NA
PB_cytopenia	NA
Haemoglobin	9,4
Absoulte_neutrophile_count	1,2
Platelet_count	331
BM_blasts_pct	NA
ME_ratio	NA
Ring_sideroblasts_pct	NA

SF3B1	0	NF1	0
TET2	0	WT1	0
SRSF2	0	IRF1	0
ASXL1	0	RAD21	0
DNMT3A	0	ATRX	0
RUNX1	0	CDKN2A	0
U2AF1	0	ETV6	0
TP53	0	KDM6A	1
EZH2	0	CEBPA	0
IDH2	0	FLT3	0
STAG2	0	GNAS	0
ZRSR2	0	PTEN	0
CBL	0	SH2B3	0
BCOR	0	BRAF	0
NRAS	0	CTNNA1	0
JAK2	0	rearr.3q	1
CUX1	0	del(5q)	1
IDH1	0	-7/del(7q)	0
KRAS	0	tri(8)	0
PHF6	0	del(11q)	0
EP300	0	del(12p)	0
GATA2	0	abn.17	0
NPM1	0	tri(19)	0
MLL2	0	del(20q)	0
PTPN11	0	del(Y)	0
CREBBP	0	other	0
KIT	0	complex	0
MPL	0		

date		1,1627
pb_cytopenia	NA	
hb		9,4
anc_log		0,183154543
plt_log		5,802118375
bm_blasts_logit	NA	
ring_sideroblasts_logit	NA	
ipss		2
age_imp		76

TABLE 1 (SUMMARY) & SUPPLEMENTARY TABLE 1

ANALYSIS OVERVIEW

myelodysplastic syndromes (MDS)

- myelodysplastic syndromes (MDS) (bone marrow disorders) represent a heterogeneous group of chronic blood cancers.
- Dysplasia (abnormal cell morphology)
- characterized by ineffective haematopoiesis resulting in peripheral cytopenias (reduction in the number of mature blood cells), and patients typically have a hypercellular bone marrow
- About 30-40% of patients evolve to acute myeloid leukaemia (AML) over months to years after diagnosis.

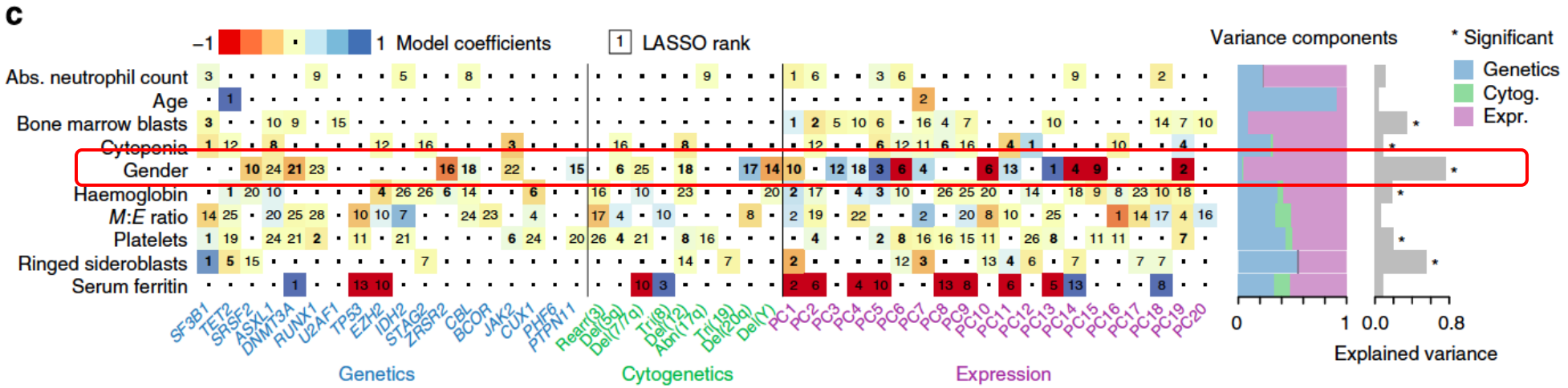
WHAT ARE THEY TRYING TO DO?

Seek to model association between phenotype and genotype in myelodysplastic syndromes (MDS).

- For genotype they use a set of 12 genes that are commonly mutated in MDS and 4 cytogenetic abnormalities that are overrepresented in MDS
- For phenotype, they use a range of clinical data (primarily based on analysis of blood)
- They then generate a series of statistical models to evaluate the relationship between mutations and general levels of gene expression genetic mechanisms impacted by mutations
- They try to associate these genetic signals with phenotype
- Finally, they attempt to predict outcome for a patient

OVERVIEW

- Challenging as trying to combine different types of patient data.
- primarily interested in isolating and defining the aggregate effects of driver mutations on the transcriptome.
- However, this will be confounded by factors such as age, sex and genetic background.
- It raises a lot of challenges regarding how to test for significant associations between the different data types.



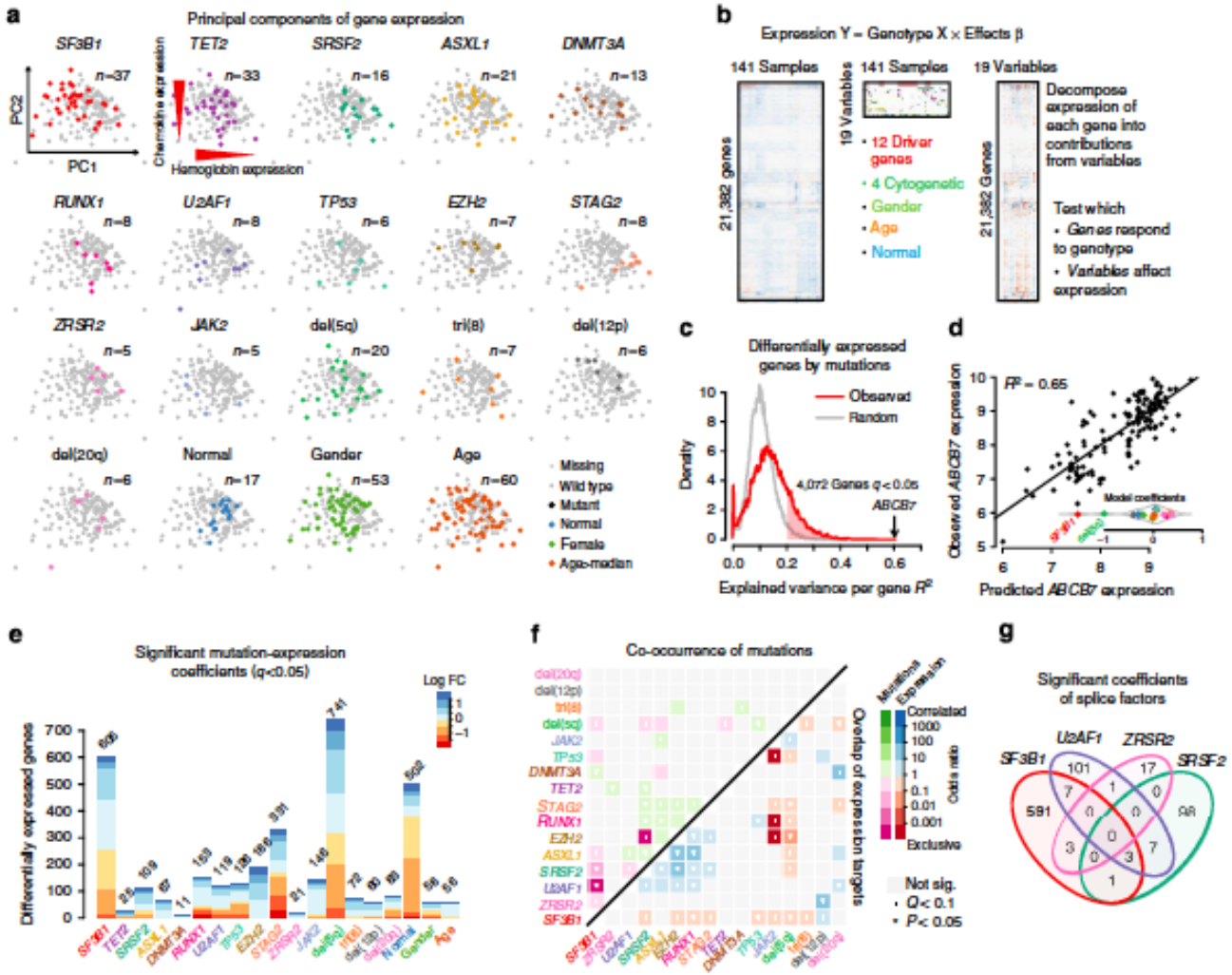
myelodysplastic syndromes (MDS)

- MDS is well studied, and commonly mutated genes, chromosomal aberrations and gene expression changes have been reported
- The most commonly mutated genes in MDS are regulators of RNA splicing and epigenetic modifiers, but signal transduction pathways and transcription factors are also frequent targets.
- MDS is also characterized by cytogenetic aberrations such as deletions on the long arms on chromosomes 5, 7 and 20, as well as more complex karyotypes.
- This information is used in the Gerstung study to try and determine which mutations and chromosomal aberrations should be prioritized

ANALYSIS DETAILS

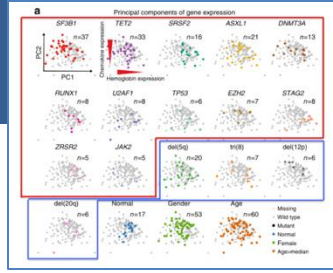
FIGURE 1

This figure contains the results from a range of different analyses



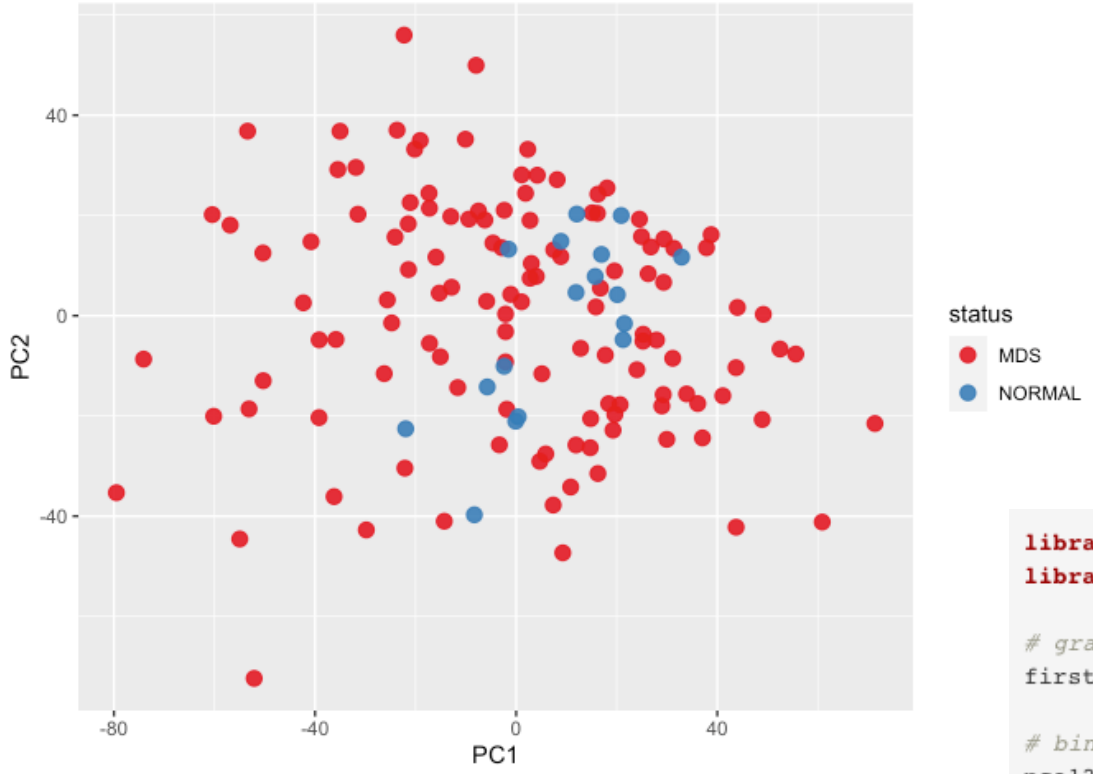
PCA ANALYSIS

PCA



Good first way to explore the data

The basic PCA plot is completely uninformative.



No apparent separation between `normal` and `MDS` patients

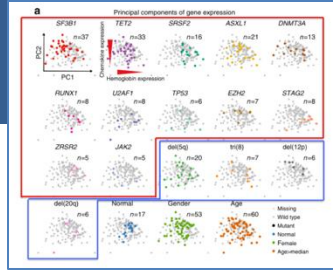
So, now we look at individual genetic events. Is a specific mutation associated with a major chang

```
library(ggplot2)
library(RColorBrewer)

# grab the PCA data
first2PCcomponents<-pca$x[,1:2]

# bind the design data to the PCA data so we are able to distinguish the patients
pca12WDesign<-transform(merge(first2PCcomponents, design,  by="row.names"), row.names=Row.names)
pca12WDesign$status=factor(pca12WDesign$Normal, levels=c(0,1), labels=c("MDS", "NORMAL"))
# plot
ggplot(pca12WDesign, aes(x=PC1, y=PC2, color=status)) +
  geom_point(size=3, alpha = 0.9) +
  scale_color_brewer(palette="Set1")
```

PCA

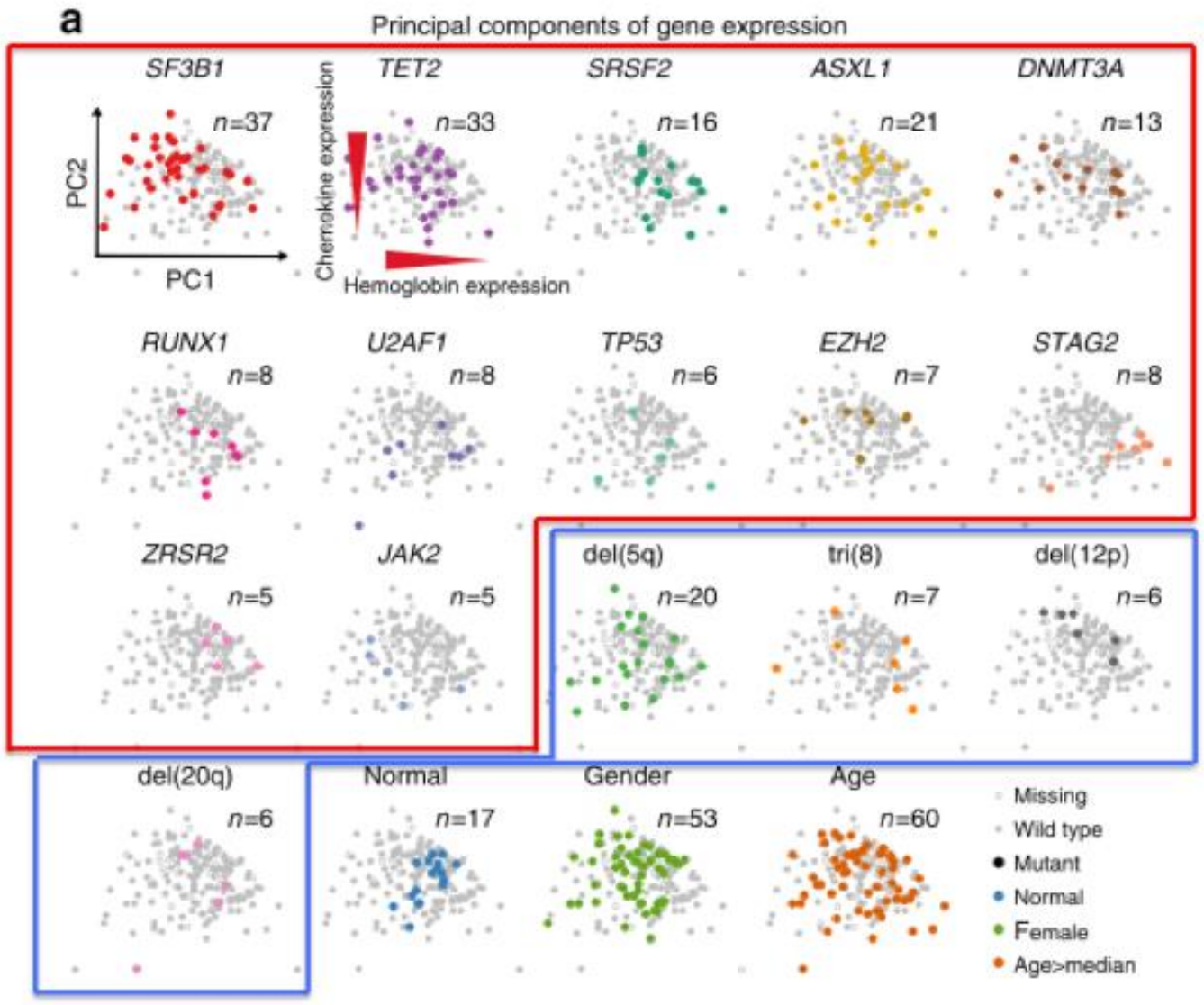


So, look at the components associated with each gene

i.e., superimpose the mutation information on the PCA plots

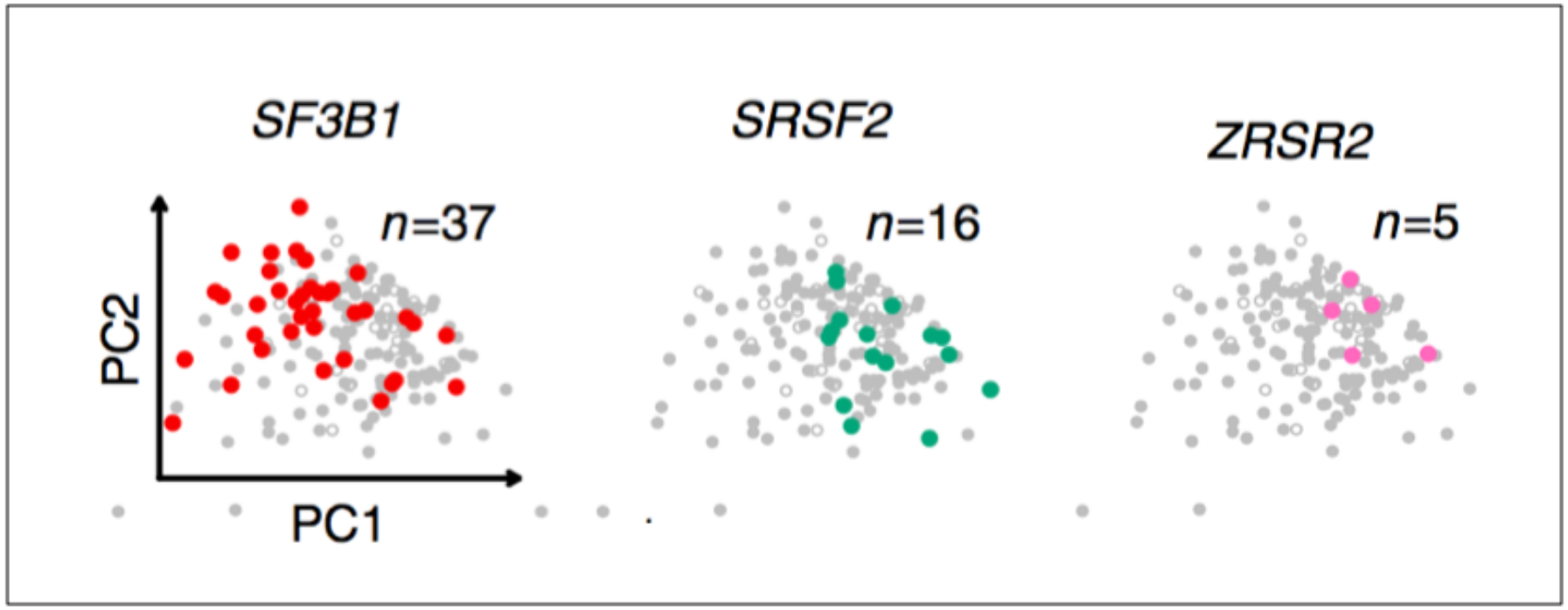
Are there specific mutations that are associated with major disruptions to the overall gene expression patterns ?

FIGURE 1a


a
Principal components of gene expression
SF3B1
n=37
TET2
n=33
SRSF2
n=16
ASXL1
n=21
DNMT3A
n=13
PC2
PC1
Chemokine expression
Hemoglobin expression
RUNX1
n=8
U2AF1
n=8
TP53
n=6
EZH2
n=7
STAG2
n=8
ZRSR2
n=5
JAK2
n=5
del(5q)
n=20
tri(8)
n=7
del(12p)
n=6
del(20q)
n=6
Normal
n=17
Gender
n=53
Age
n=60
Missing
Wild type
Mutant
Normal
Female
Age>median

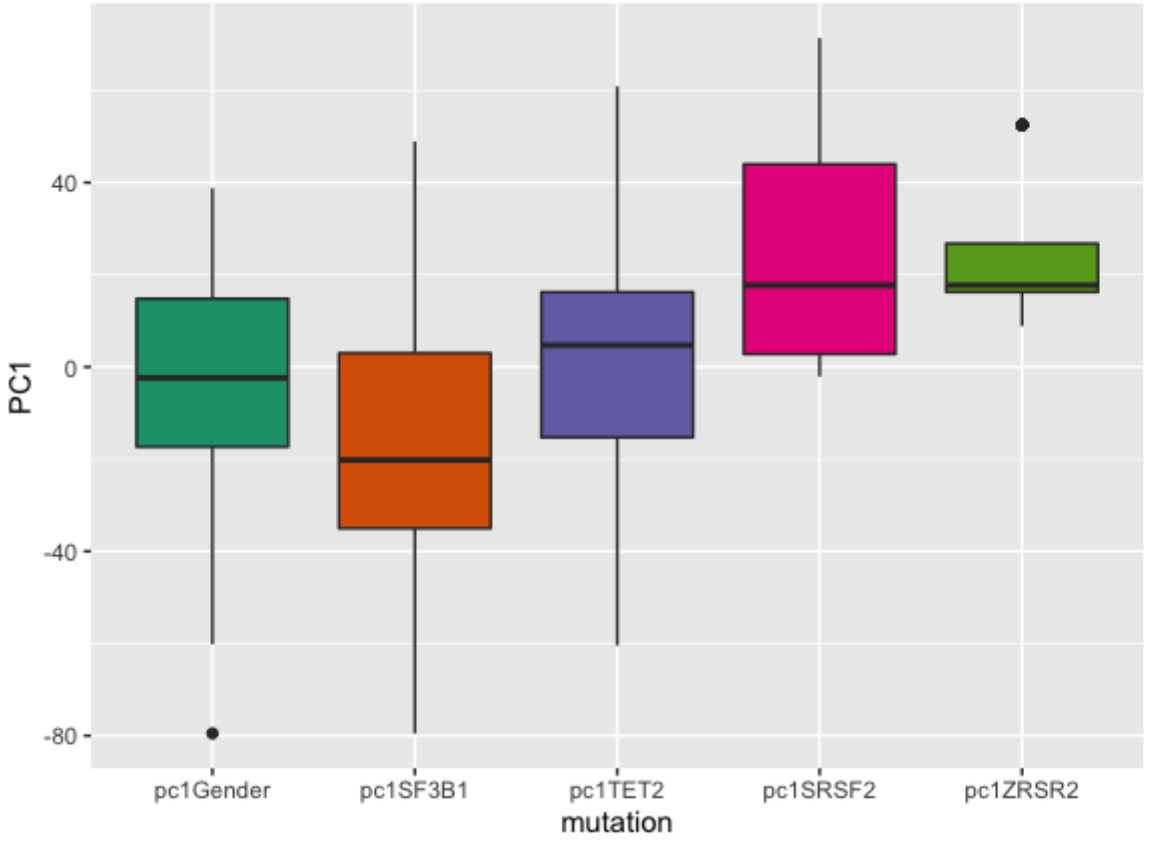
INTRODUCTION

"patients with mutations in the RNA splicing factor SF3B1 tend to have low scores on the first principal component, whereas patients with mutations in two other splicing factors, SRSF2 and ZRSR2, have high scores."



INTRODUCTION

"patients with mutations in the RNA splicing factor SF3B1 tend to have low scores on the first principal component, whereas patients with mutations in two other splicing factors, SRSF2 and ZRSR2, have high scores."




```
pc1Gender<-pca12WDesign[which(pca12WDesign$Gender==1), "PC1"]
pc1SF3B1<-pca12WDesign[which(pca12WDesign$SF3B1==1), "PC1"]
pc1TET2<-pca12WDesign[which(pca12WDesign$TET2==1), "PC1"]
pc1SRSF2<-pca12WDesign[which(pca12WDesign$SRSF2==1), "PC1"]
pc1ZRSR2<-pca12WDesign[which(pca12WDesign$ZRSR2==1), "PC1"]

pc1x3<-cbind(pc1Gender,pc1SF3B1, pc1TET2, pc1SRSF2, pc1ZRSR2 )
```

```
library(reshape2)
meltedPc1x3=melt(pc1x3)
colnames(meltedPc1x3)<-c("ID", "mutation", "PC1")
ggplot(data=meltedPc1x3, aes(x=mutation, y=PC1)) +
  geom_boxplot(aes(fill=mutation)) + scale_fill_brewer(palette="Dark2")
```

We don’t know what this means yet, only that the different mutations produce different effects on gene expression

DATA MODELING

A linear model to deconvolute gene expression and mutations. To explore predictors of gene expression in a multivariate framework, we developed a linear modelling approach that measures the association of expression levels on a gene-by-gene basis with a number of potential predictors, including driver mutations and nuisance variables (Fig. 1b). The normal samples were included to identify changes common to all MDS samples. Somatically acquired mutations and cytogenetic lesions were encoded as being present/absent. The model assumes that each mutation is associated with a certain set of expression changes and that the expression pattern in cases with a complex genotype comprising multiple alterations is the sum of the changes induced by each mutation. We chose a linear model due to its interpretability and established statistical methods, enabling us to test which transcripts are deregulated in the presence of specific alterations, after correcting for other confounding variables such as the nuisance factors and coexisting driver mutations. The additivity assumption ignores potential interactions between genetic lesions that may arise from the cellular signalling circuitry. However, inferring these interactions systematically from the data would require many more cases, as the number of gene:gene interaction pairs is quadratic and yields combinatorially many terms for higher order interactions. Hence, this assumption appears necessary for statistically robust inference in a data set of this size.

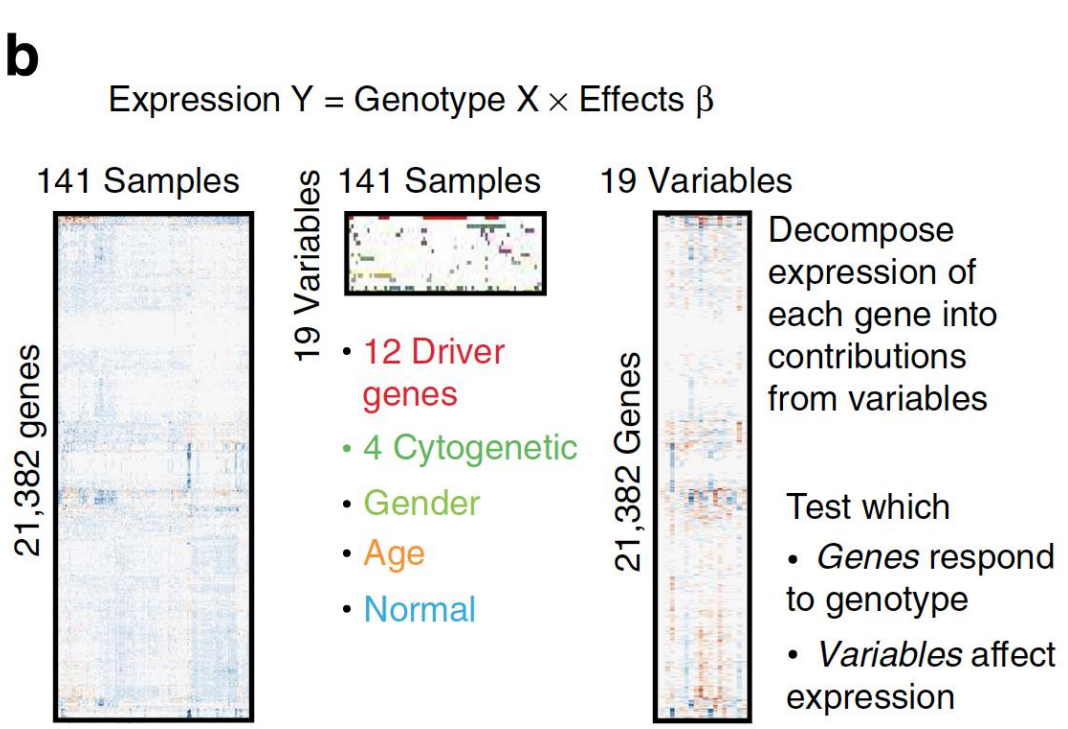
VG Live VGTV VG+ Sport TV-guide Tips oss Kjøp VG+

Bane Nor har sendt ut dette bildet, som viser deler av ødeleggelsene etter nattens brann. Foto: Bane Nor

Stengt for togtrafikk ved Nationaltheatret - utsetter oppdatering

Mandag er det fortsatt stengt for togtrafikk ved Nationaltheatret i Oslo etter en kabelbrann.

MODELING THE DATA



observed expression for gene k in patient k

$$Y_{ik} = \sum_{j=1} X_{ij}\beta_{ij} + \beta_{0i}$$

where:

X_{ij} is the mutation matrix for patient k and mutation j

$X_{ij} = 1$: patient i has an oncogenic mutation j

$X_{ij} = 0$ no mutation

for gender $X_{ij} = 1 =$ female; $X_{ij} = 0 =$ male

for age X_{ij} takes integer values.

β_{jk} is the the expression change in gene k induced by the presence of mutation j.

β_{0j} denotes the baseline expression level of gene j

glmPrediction <- geneExprlm$coefficients %*% t(design)

Usage

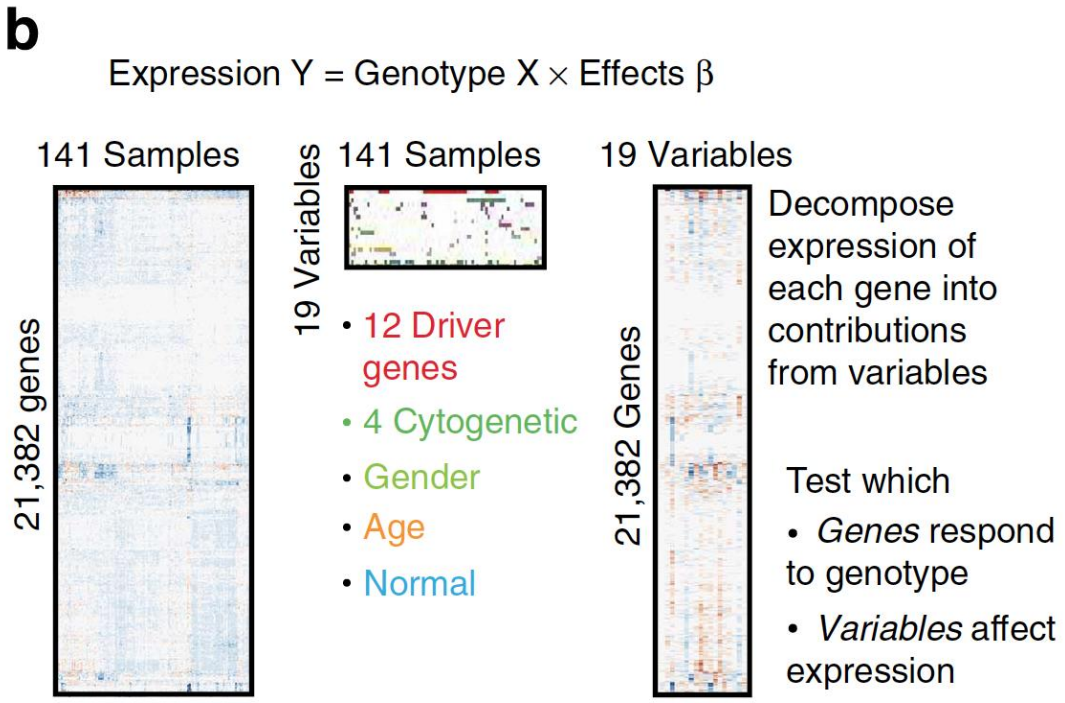
```
lmFit(object, design=NULL, ndups=1, spacing=1, block=NULL, correlation, weights=NULL, method="ls", ...)
```

Arguments

object	A matrix-like data object containing log-ratios or log-expression values for a series of arrays, with rows corresponding to genes and columns to samples. Any type of data object that can be processed by `getEAWP` is acceptable.
design	the design matrix of the microarray experiment, with rows corresponding to arrays and columns to coefficients to be estimated. Defaults to the unit vector meaning that the arrays are treated as replicates.

The advantage with using **lmFit** is you can simply provide gene expression data and a design.
Trying to use the in-built function would require more effort.
Interactions can be handled but they must be specified in separate commands in **Limma**

MODELING THE DATA



glmPrediction <- geneExprlm$coefficients %*% t(design)

Observed gene expression

Perturbation to expression from mutations

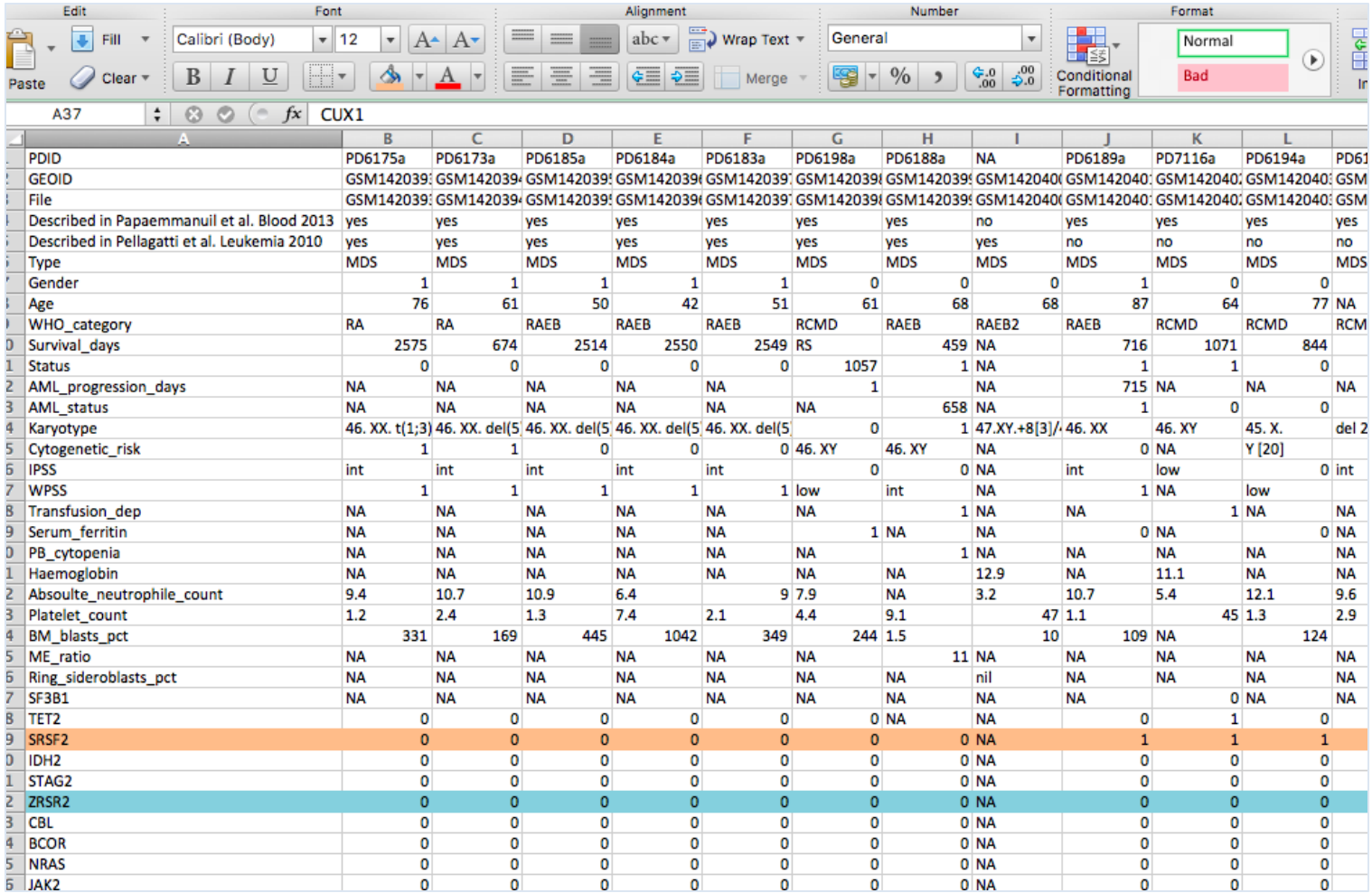
Baseline expression of gene k

$$Y_{ik} = \sum_{j=1} X_{ij}\beta_{jk} + \beta_{0k} + \varepsilon$$

Patient i
Gene k
Mutation j

SUPPLEMENTARY DATA

The gene mutation information is contained in the supplementary file SuppTableS1GEO.txt



	A	B	C	D	E	F	G	H	I	J	K	L	
	PDID	PD6175a	PD6173a	PD6185a	PD6184a	PD6183a	PD6198a	PD6188a	NA	PD6189a	PD7116a	PD6194a	PD61
	GEOID	GSM142039	GSM142039	GSM142039	GSM142039	GSM142039	GSM142039	GSM142039	GSM142040	GSM142040	GSM142040	GSM142040	GSM
	File	GSM142039	GSM142039	GSM142039	GSM142039	GSM142039	GSM142039	GSM142039	GSM142040	GSM142040	GSM142040	GSM142040	GSM
	Described in Papaemmanuil et al. Blood 2013	yes	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	yes
	Described in Pellagatti et al. Leukemia 2010	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no
	Type	MDS	MDS	MDS	MDS	MDS	MDS	MDS	MDS	MDS	MDS	MDS	MDS
	Gender	1	1	1	1	1	0	0	0	1	0	0	
	Age	76	61	50	42	51	61	68	68	87	64	77	NA
	WHO_category	RA	RA	RAEB	RAEB	RAEB	RCMD	RAEB	RAEB2	RAEB	RCMD	RCMD	RCM
	Survival_days	2575	674	2514	2550	2549	RS	459	NA	716	1071	844	
	Status	0	0	0	0	0	1057	1	NA	1	1	0	
	AML_progression_days	NA	NA	NA	NA	NA	1		NA	715	NA	NA	NA
	AML_status	NA	NA	NA	NA	NA	NA	658	NA	1	0	0	
	Karyotype	46. XX. t(1;3)	46. XX. del(5	46. XX. del(5	46. XX. del(5	46. XX. del(5	0	1	47.XY.+8[3]/	46. XX	46. XY	45. X.	del 2
	Cytogenetic_risk	1	1	0	0	0	46. XY	46. XY	NA	0	NA	Y [20]	
	IPSS	int	int	int	int	int	0	0	NA	int	low	0	int
	WPSS	1	1	1	1	1	low	int	NA	1	NA	low	
	Transfusion_dep	NA	NA	NA	NA	NA	NA	1	NA	NA	1	NA	NA
	Serum_ferritin	NA	NA	NA	NA	NA	1	NA	NA	0	NA	0	NA
	PB_cytopenia	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA
	Haemoglobin	NA	NA	NA	NA	NA	NA	NA	12.9	NA	11.1	NA	NA
	Absoulte_neutrophile_count	9.4	10.7	10.9	6.4	9	7.9	NA	3.2	10.7	5.4	12.1	9.6
	Platelet_count	1.2	2.4	1.3	7.4	2.1	4.4	9.1	47	1.1	45	1.3	2.9
	BM_blasts_pct	331	169	445	1042	349	244	1.5	10	109	NA	124	
	ME_ratio	NA	NA	NA	NA	NA	NA	11	NA	NA	NA	NA	NA
	Ring_sideroblasts_pct	NA	NA	NA	NA	NA	NA	NA	nil	NA	NA	NA	NA
	SF3B1	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	NA	NA
	TET2	0	0	0	0	0	0	NA	NA	0	1	0	
	SRSF2	0	0	0	0	0	0	0	NA	1	1	1	
	IDH2	0	0	0	0	0	0	0	NA	0	0	0	
	STAG2	0	0	0	0	0	0	0	NA	0	0	0	
	ZRSR2	0	0	0	0	0	0	0	NA	0	0	0	
	CBL	0	0	0	0	0	0	0	NA	0	0	0	
	BCOR	0	0	0	0	0	0	0	NA	0	0	0	
	NRAS	0	0	0	0	0	0	0	NA	0	0	0	
	JAK2	0	0	0	0	0	0	0	NA	0	0	0	

SUPPLEMENTARY DATA

Absolute_neutrophile_count		9	7.9	NA	3.2	10.7	5.4
Platelet_count	2.1		4.4	9.1	47	1.1	
BM_blasts_pct		349	244	1.5	10	109	NA
ME_ratio	NA		NA	11	NA	NA	NA
Ring_sideroblasts_pct	NA		NA	NA	nil	NA	NA
SF3B1	NA		NA	NA	NA	NA	
TET2		0	0	NA	NA	0	
SRSF2		0	0	0	NA	1	
IDH2		0	0	0	NA	0	
STAG2		0	0	0	NA	0	
ZRSR2		0	0	0	NA	0	
CBL		0	0	0	NA	0	
BCOR		0	0	0	NA	0	
NRAS		0	0	0	NA	0	

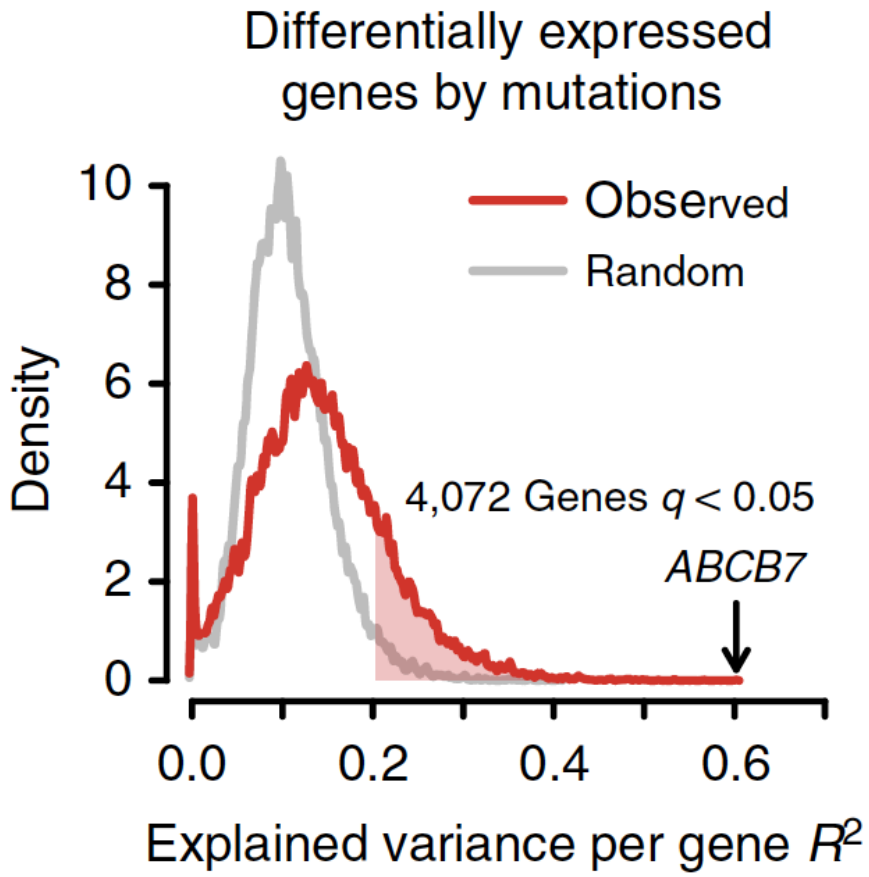
for example, if patient 1 is *female*, has mutations in *SF3B1* and *del*(5*q*), then, for gene *ABCB7*

$$Y_{ABCB7,1} = \beta_{SF3B1,ABCB7} + \beta_{del(5q),ABCB7} + \beta_{female,ABCB7} + \beta_{ABCB7,1}$$

Can we associate gene expression changes with specific mutations?

FIGURE 1C

FIGURE 1C: EXPLANED VARIANCE / GENE



This is why they call the classifyTestsF function in the code

THERE WERE 4072 GENES THAT HAD CHANGES THAT COULD BE ASSOCIATED WITH A MUTATION

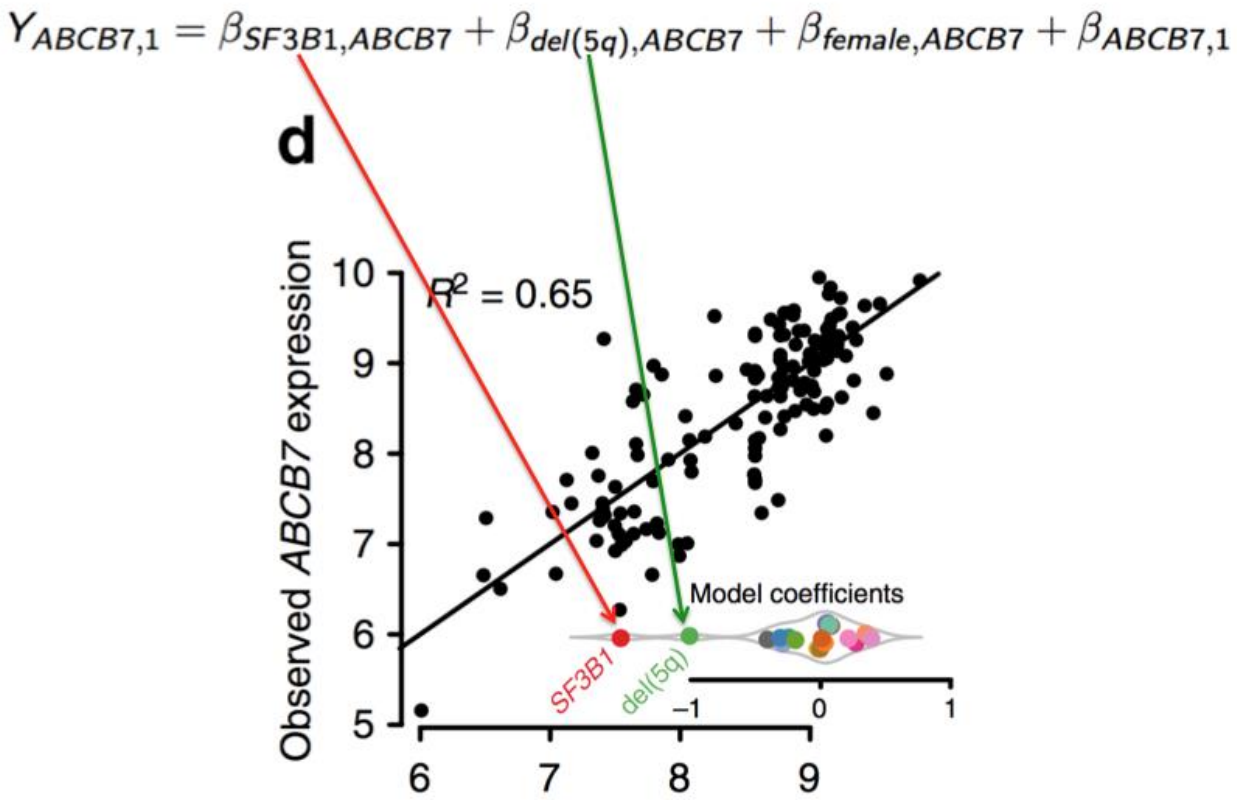
THE MINIMUM VARIANCE THAT COULD BE EXPLAINED WAS 20%. THE MOST VARIANCE WAS ASSOCIATED WITH ABCB7 (ATP Binding Cassette Subfamily B Member 7)

https://www.genecards.org/cgi-bin/carddisp.pl?gene=ABCB7

Can we use this information to predict how mutations will affect expressions of a single gene?

FIGURE 1D

FIGURE 1D: PREDICTED ABCB7 EXPRESSION



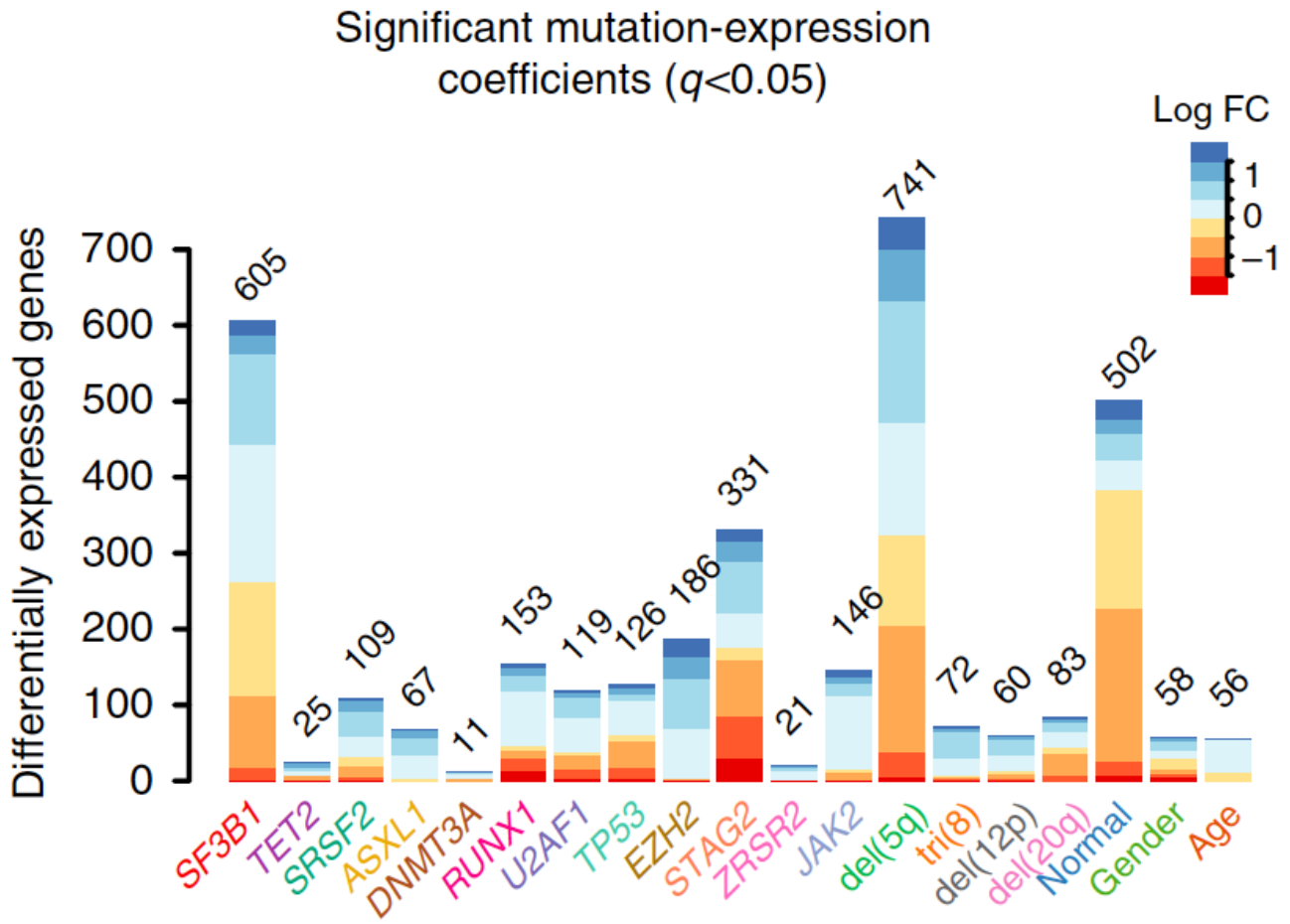
IN MOST CASES THE MODEL COEFFICIENTS WERE SMALL FOR ABCB7, BUT LARGE VALUES WERE PREDICTED FOR SF3B1 & del(5q).

THESE MODIFICATIONS ARE KNOWN TO PRODUCE STRONG DOWN-REGULATION OF ABCB7

What is the cumulative effect of a single mutation on system wide gene expression?

FIGURE 1E

FIG 1E: DIFFERENTIALLY EXPRESSED GENES BY GENE



SF3B1 & del(5q).have comparable numbers of differentially expressed genes
But the del(5q) is deletion of a region, compared to a point mutation for SF3B1

What the occurence between co-mutations and the impact on system wide gene expression?

FIGURE 1F

Figure 1F: INTERACTIONS BETWEEN GENES

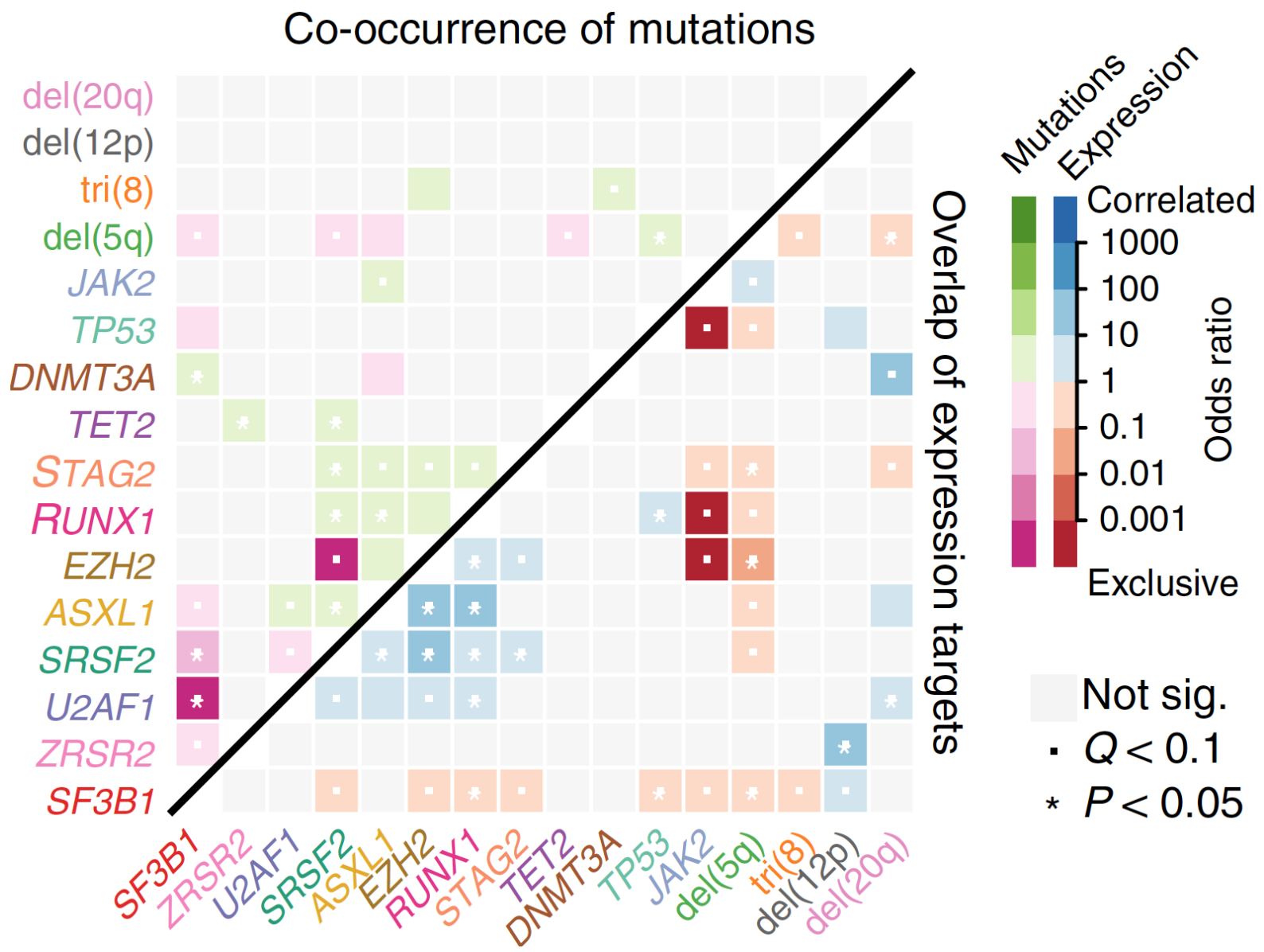


Figure 1F: INTERACTIONS BETWEEN GENES

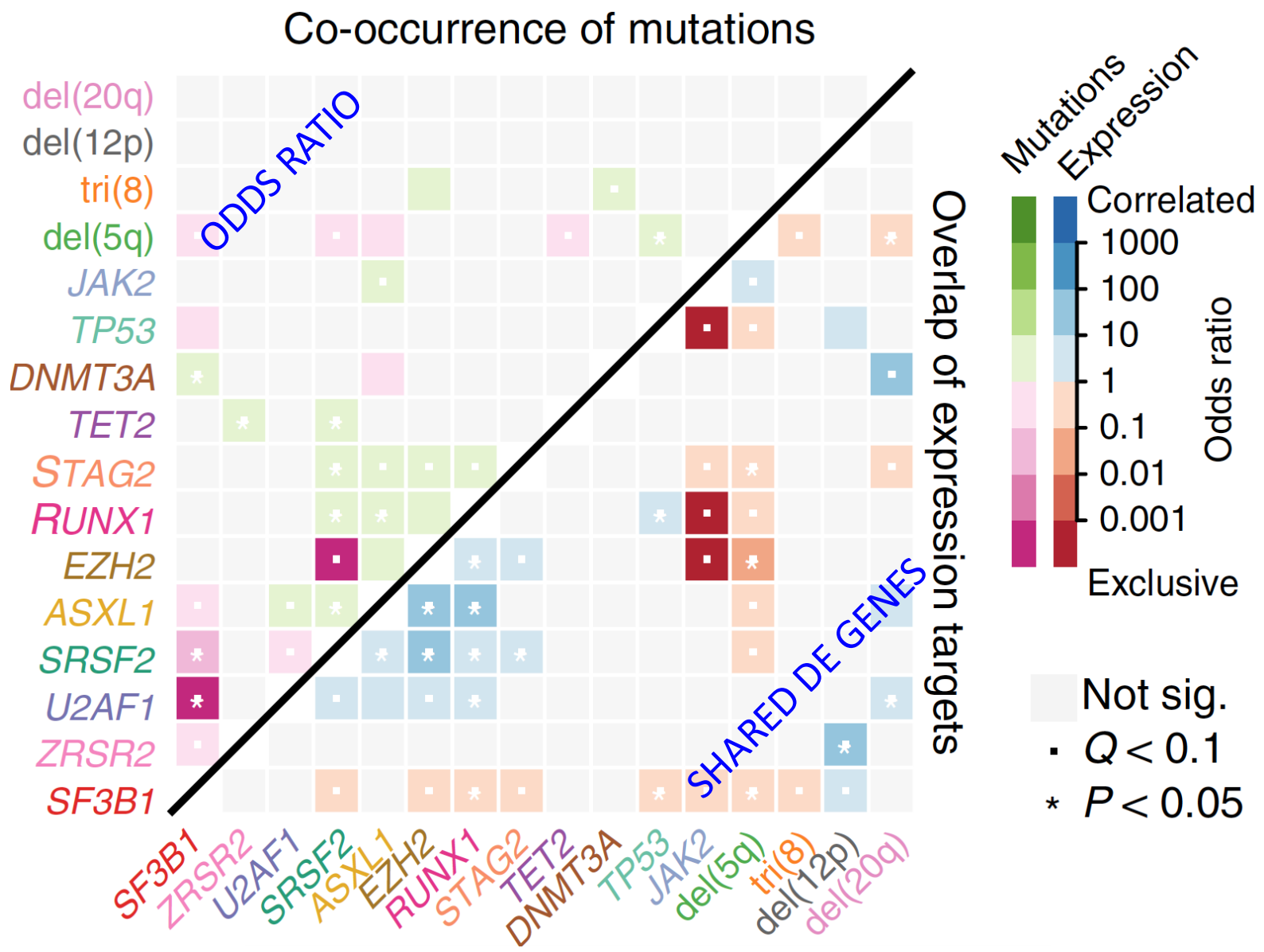
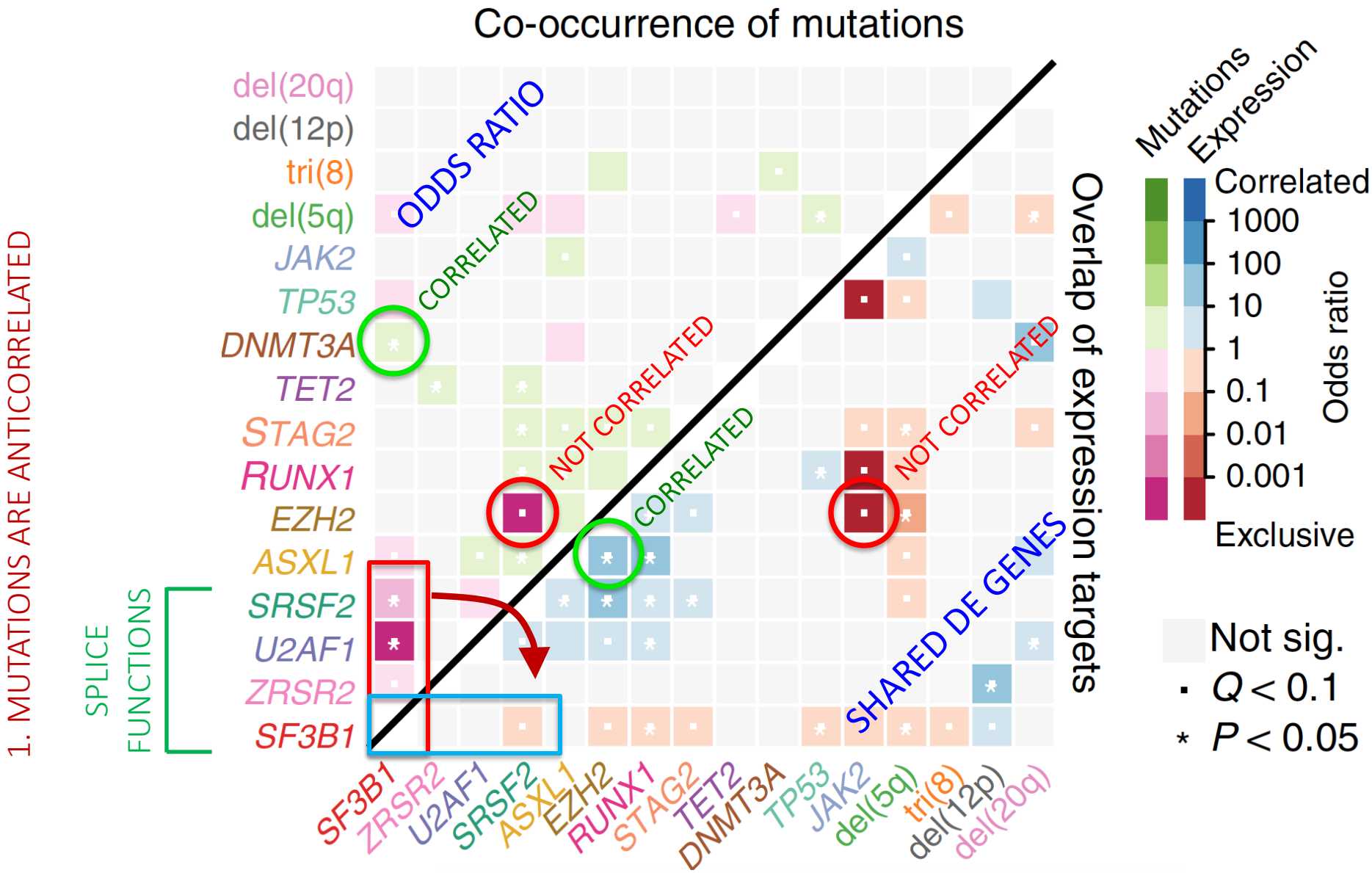


Figure 1F: INTERACTIONS BETWEEN GENES



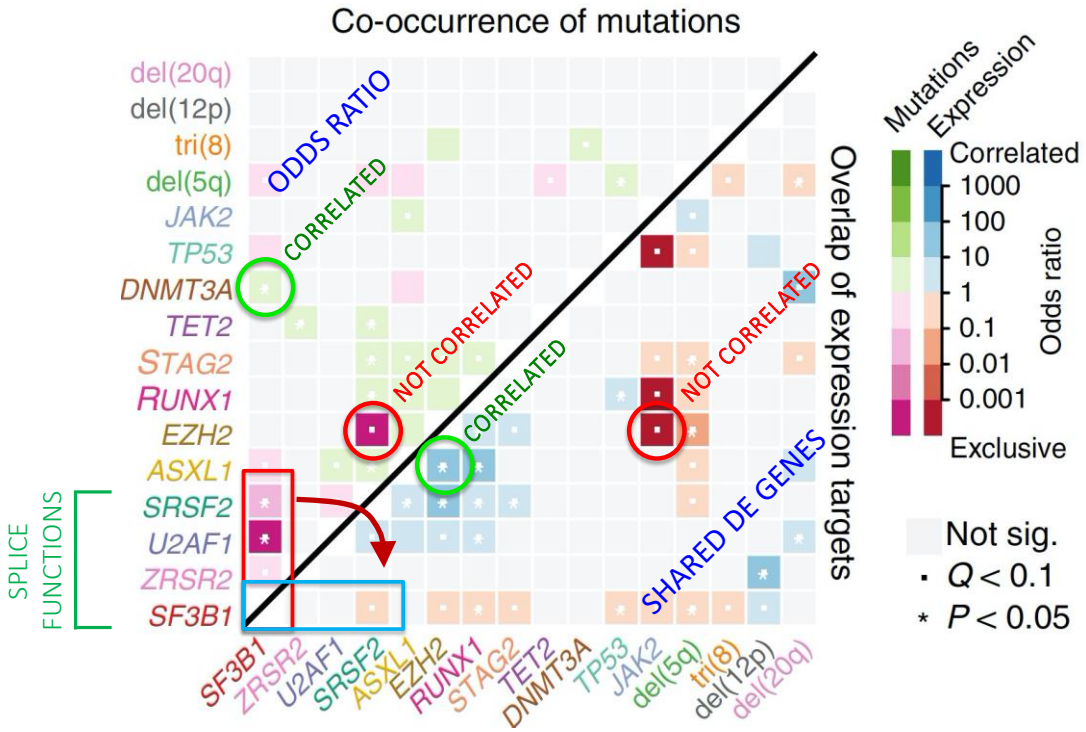
1. MUTATIONS ARE ANTICORRELATED

SPLICE FUNCTIONS

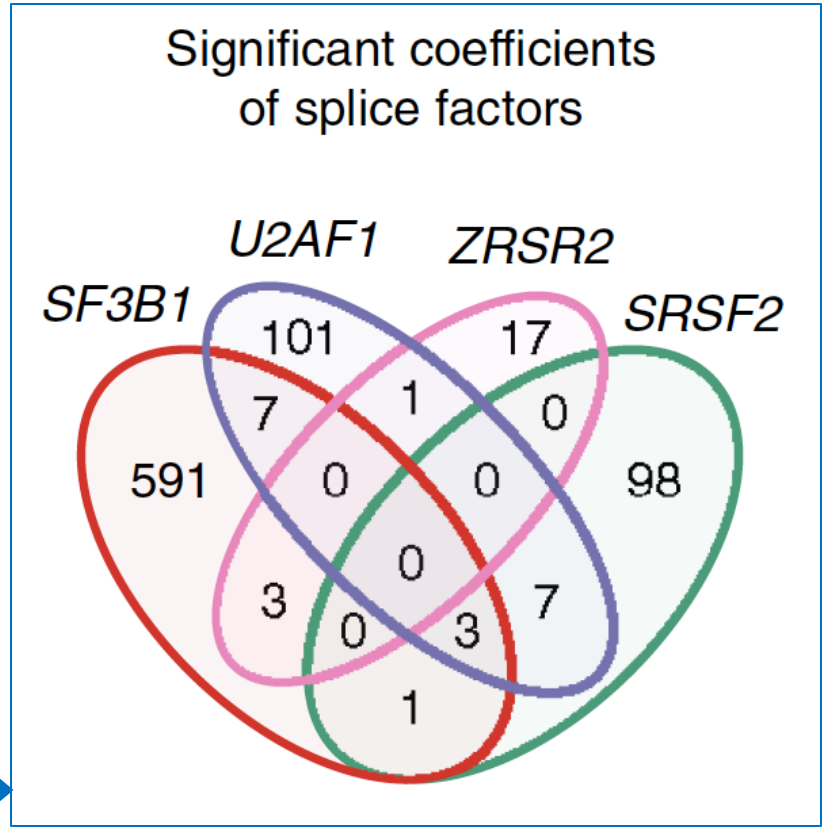
2. GENE EXPRESSION IS NOT CORRELATED

Figure 1F: INTERACTIONS BETWEEN GENES

1. MUTATIONS ARE ANTICORRELATED



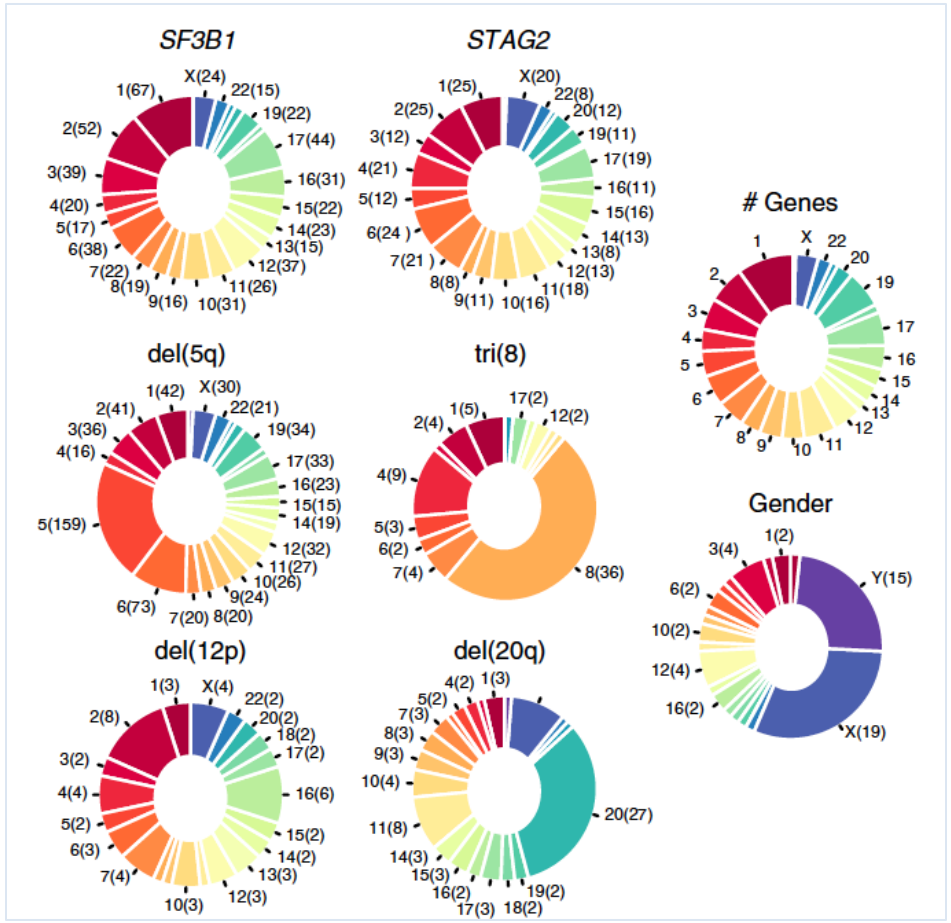
2. GENE EXPRESSION IS NOT CORRELATED



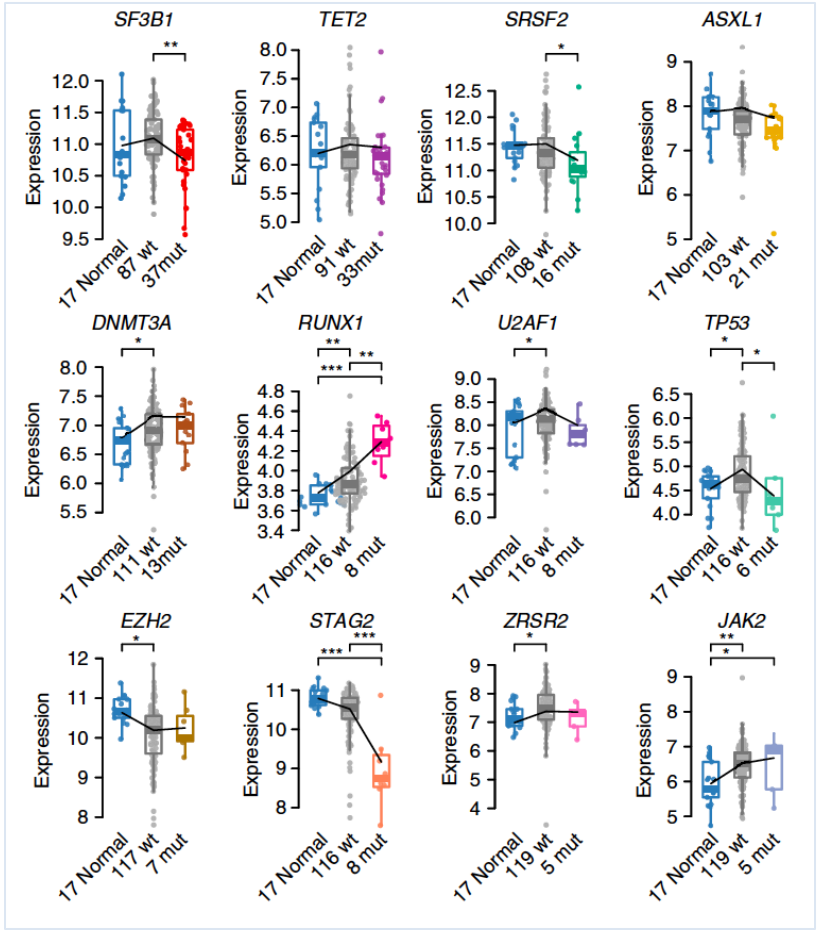
Further characterisation of the system wide impact of single mutations

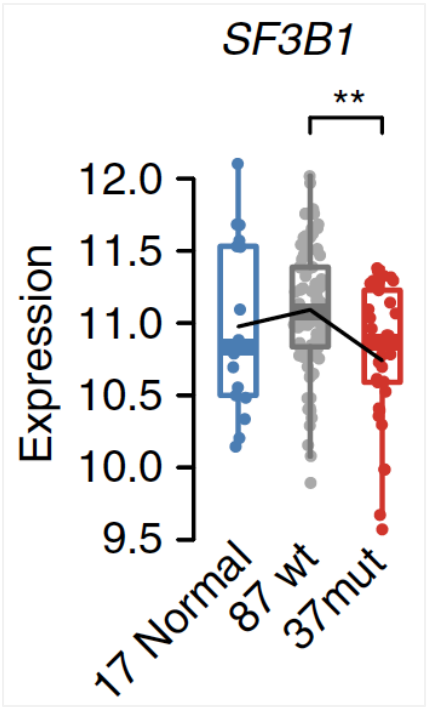
FIGURE 2

FIGURE 2A/B



DISTRIBUTION OF AFFECTED GENES BY CHROMOSOME (ONLY DELETIONS SHOW ANY PATTERN)

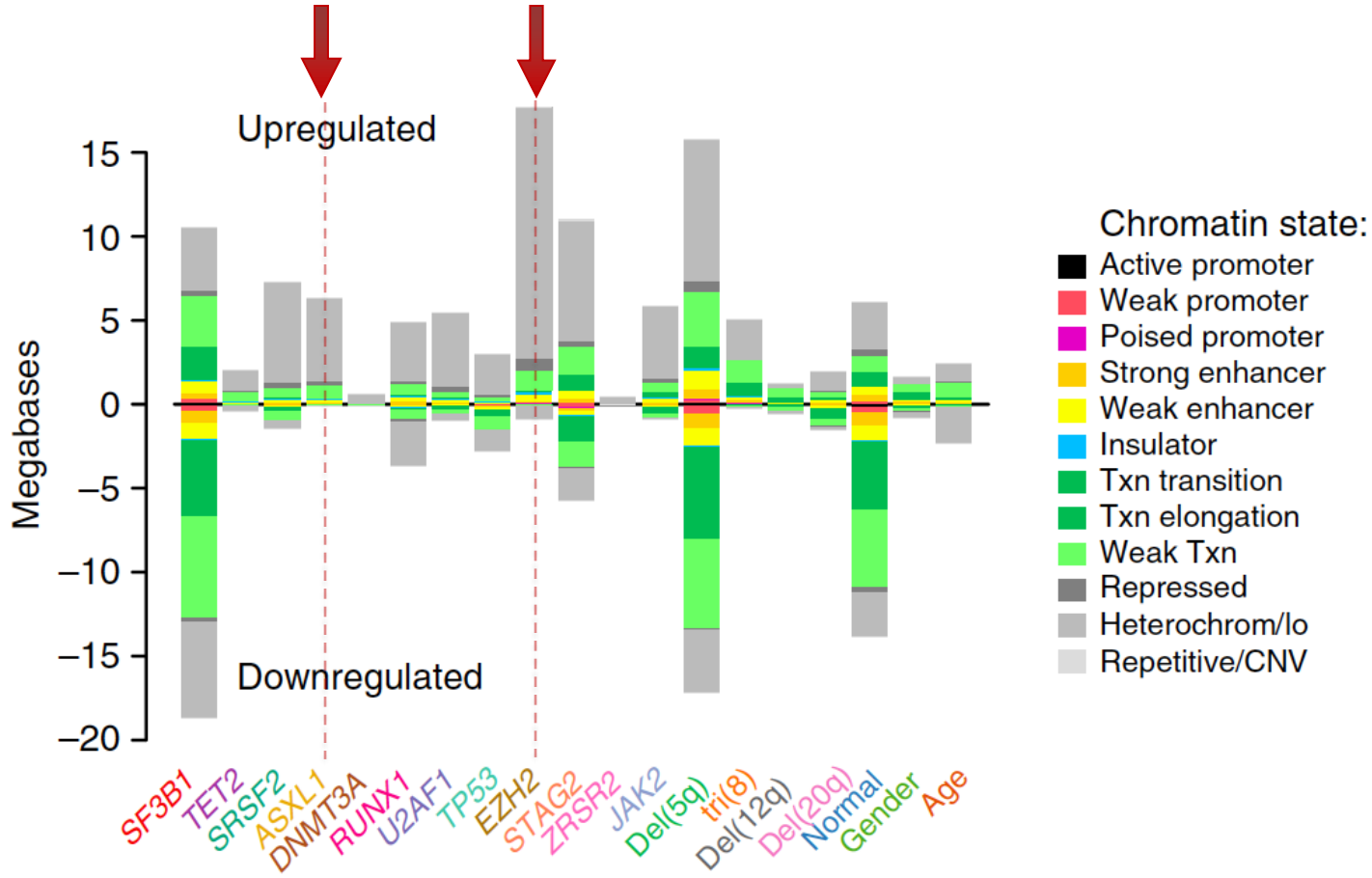




VARIATION IN AFFECTED GENES BETWEEN **NORMAL/WT/MUTATION** FOR KEY GENES

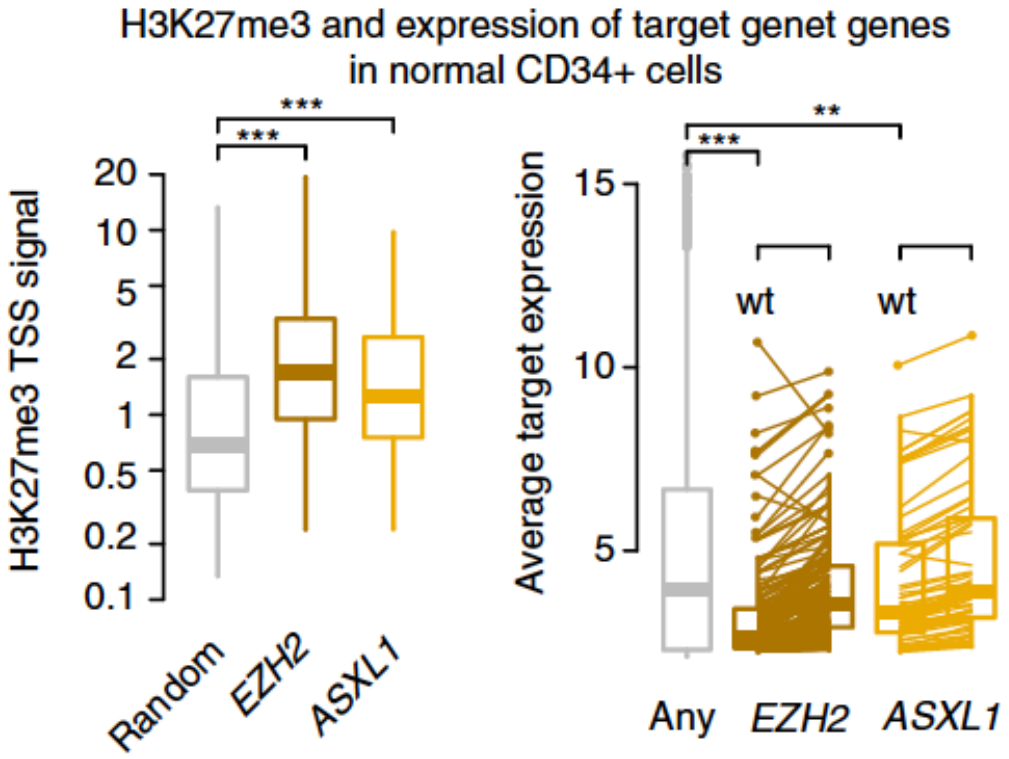
FIGURE 2C

MEGABASES, NOT GENE NUMBER



Heterochromatin states are highly disrupted in samples with driver mutations in the specified genes.
Usually, heterochromatin is densely packed and not accessible to transcription factors. But the mutations seem to disrupt this, particularly in ASXL1, and EZH2, both of which play roles in chromatin silencing.

FIGURE 2D



Left: H3K27me3 ChIP-seq enrichment data at transcription start sites (TSSs) of 1000 randomly selected genes that are predicted to be differentially expressed in EZH2 and ASXL1 mutants (there are many more H3K27me3 modifications at these TSSs when there is a mutation in the EZH2 or ASXL1 gene)
Right: These changes lead to significant expression changes (generally upregulation, consistent with Figure 2C)

Making predictive models for clinical characteristics

FIGURE 3

FIGURE 3

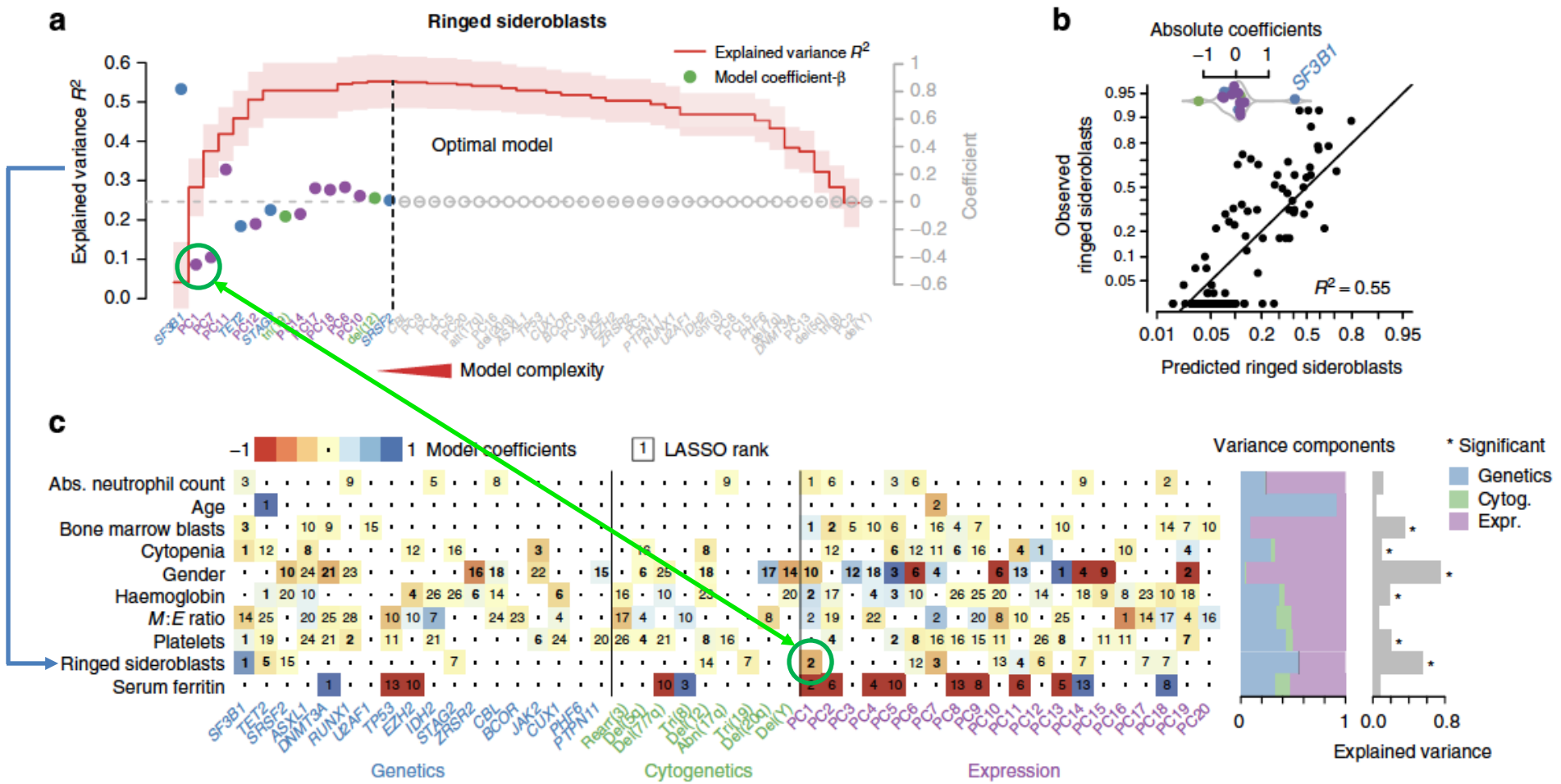


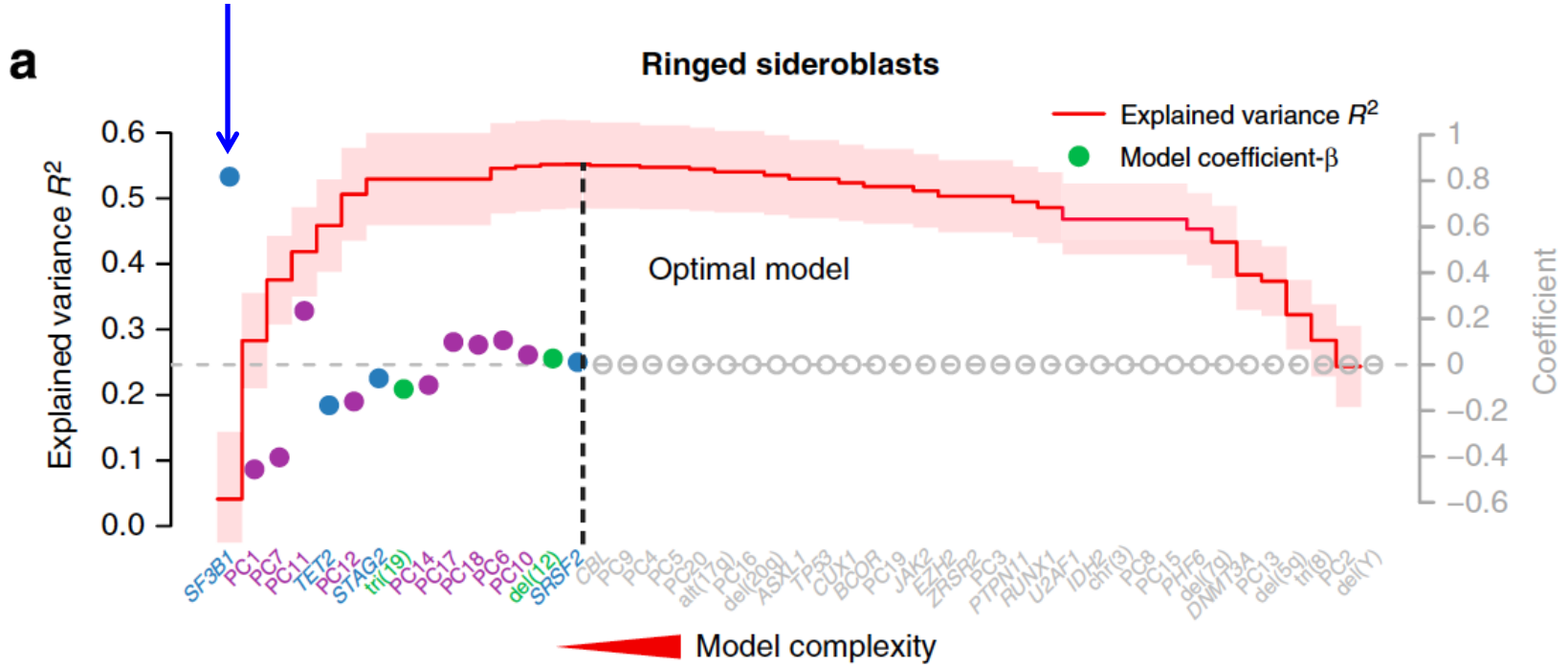
Figure 3 | Prediction of blood and bone marrow counts. (a) Variance explained by selected driver genes, cytogenetic lesions and first 20 transcriptome principal components (red line±1 s.d.; fivefold cross validation) ordered by their occurrence in a LASSO penalized model. The optimal model maximizes the explained variance R^2. The right axis indicates the effect of each standardized covariate in the optimal model.

(b) Scatter plot of predicted and observed amounts of ringed sideroblasts on a double logit axis. The inset shows the model coefficients indicating the magnitude of each fold change of driver alterations or a unit fold change in the expression components.

(c) Heatmap of optimal model coefficients for eight blood and bone marrow counts plus gender and age. LASSO-selected coefficients are coloured. The numbers on each tile denote the order in which variables are included indicating their relative importance. Bold fonts are used for highly significant coefficients in which the explained variance is one s.d. below the maximum. The right bar plot shows the estimated distribution of variance explained by genetic, cytogenetic and transcriptomic variables. Stars (*) denote models where R^2 is greater than zero by a margin of more than one s.d.

FIGURE 3A

MOST SIGNIFICANT CONTRIBUTION COMES FROM SF3B1



GLM WITH A LASSO MODEL (USING PRESENCE OF MUTATION + PCA OF GENE EXPRESSION)

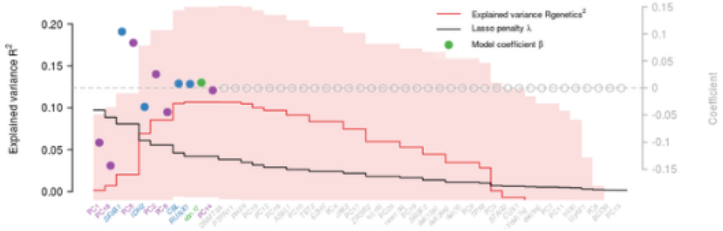
ADD FIRST COMPONENT THAT IS MOST CORRELATED WITH BLOOD COUNTS, THEN ADD SECOND…

FIGURE 3A - SUPPLEMENTAL

THE OTHER MODELS WEREN'T SO EFFECTIVE

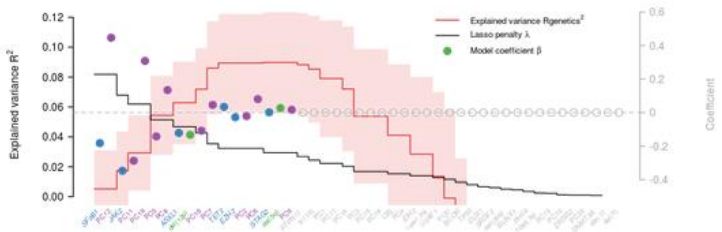
SF3B1

ABSOLUTE NEUTROPHIL COUNT

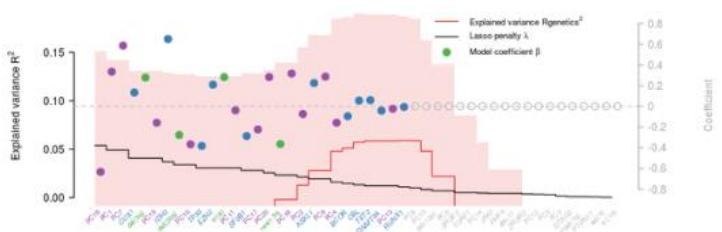


SF3B1

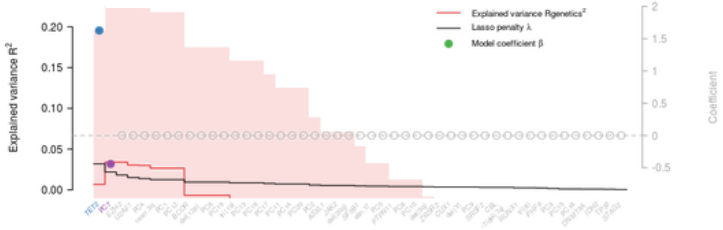
CYTOPENIA



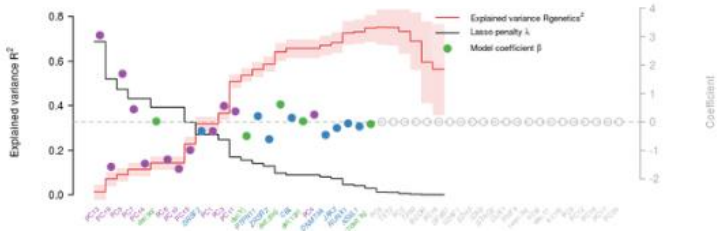
M:E RATIO



AGE

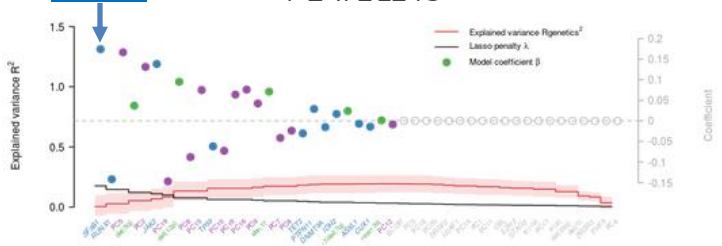


GENDER



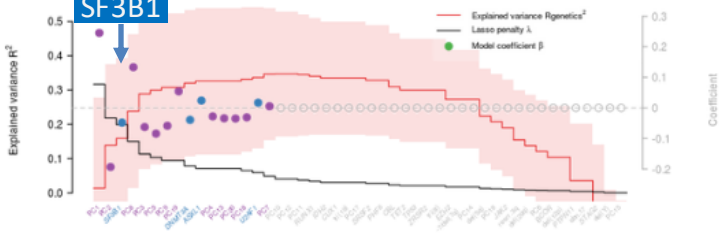
SF3B1

PLATELETS

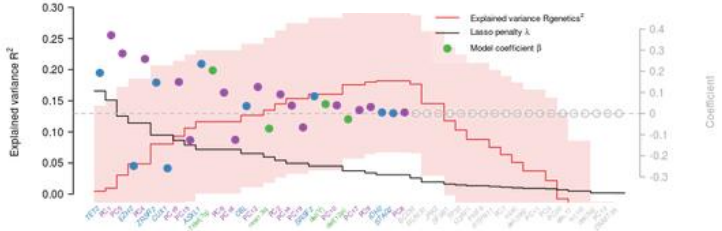


BONEMARROW BLASTS

SF3B1



HAEMOGLOBIN



SF3B1

RINGED SIDEROBLASTS

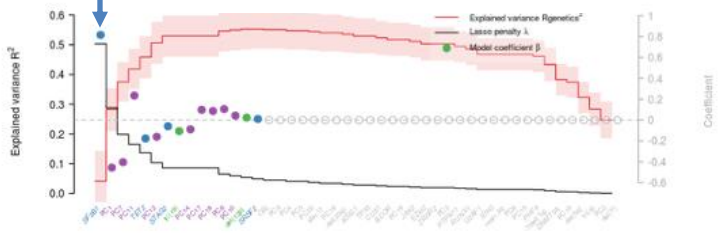
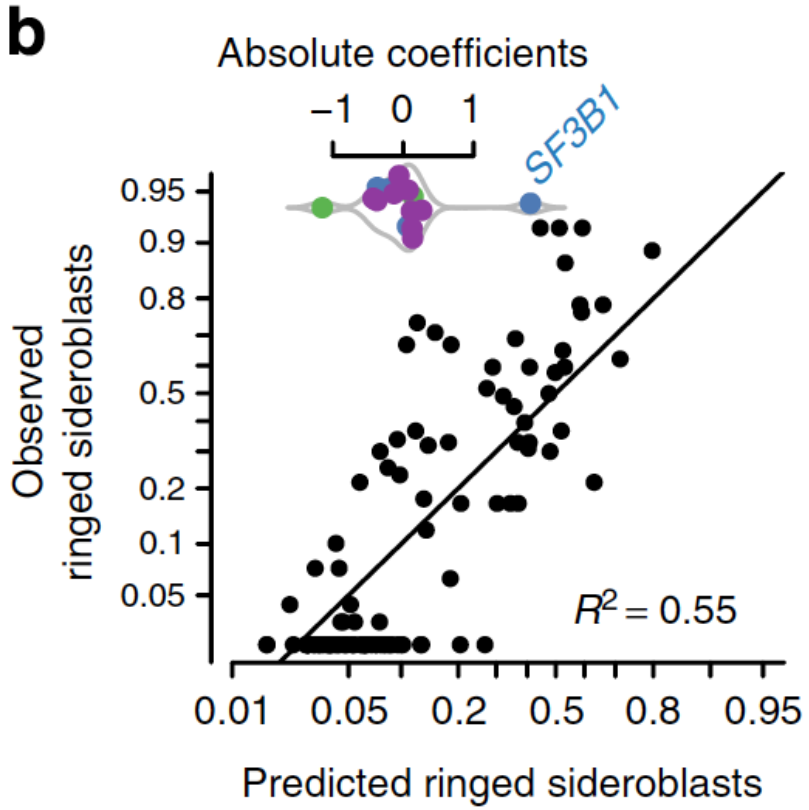
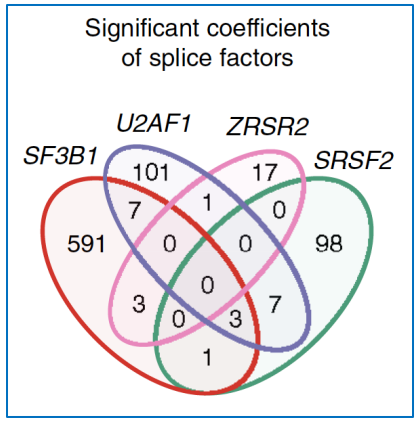
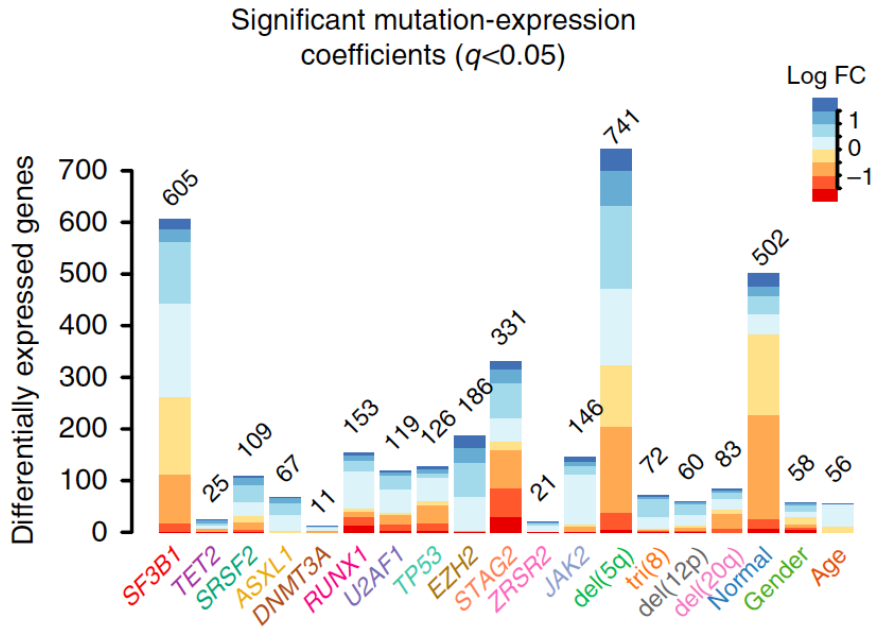
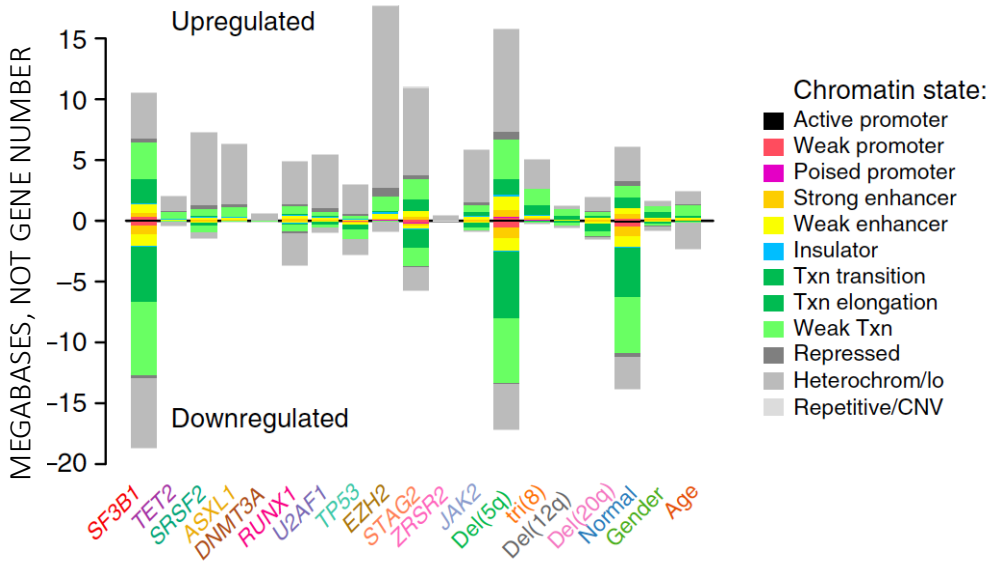


FIGURE 3B



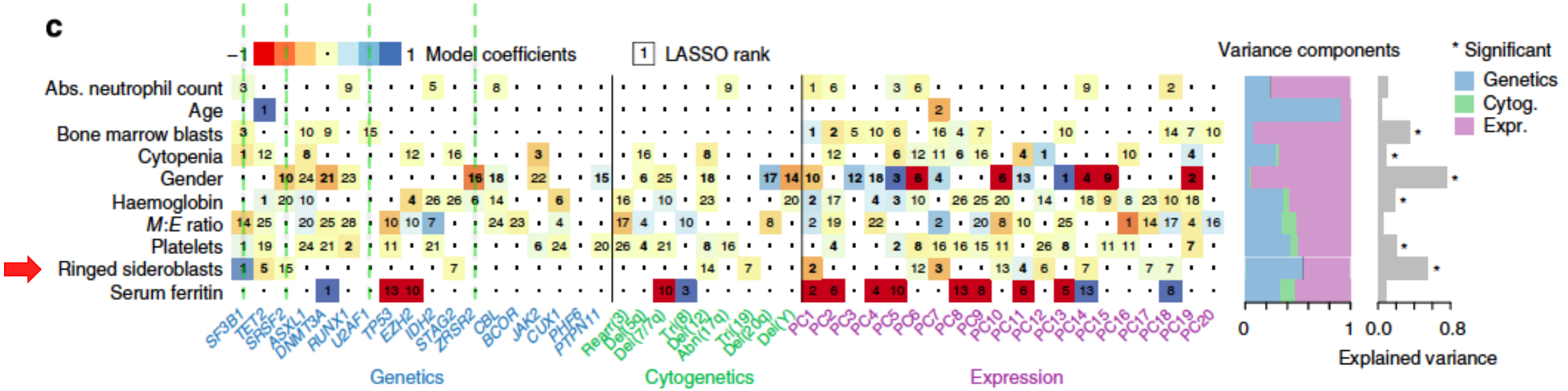


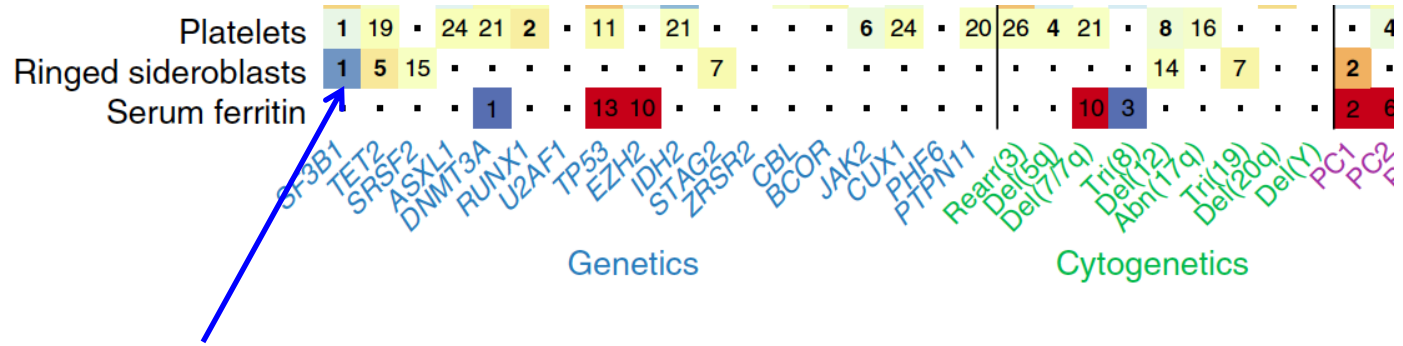




THIS SHOWS MODEL PERFORMANCE

FIGURE 3C





SF3B1 HAS FIRST LASSO RANK FOR RINGED SIDEROBLAST MODEL

RESULTS VARY WILDLY AMONGST MODELS

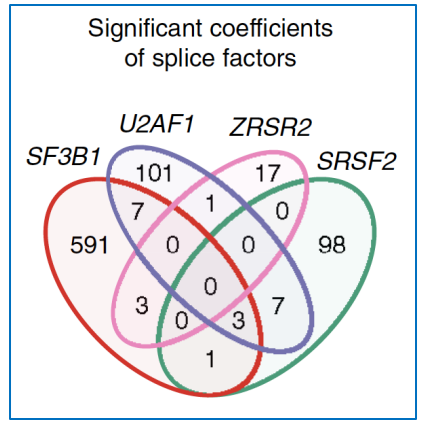
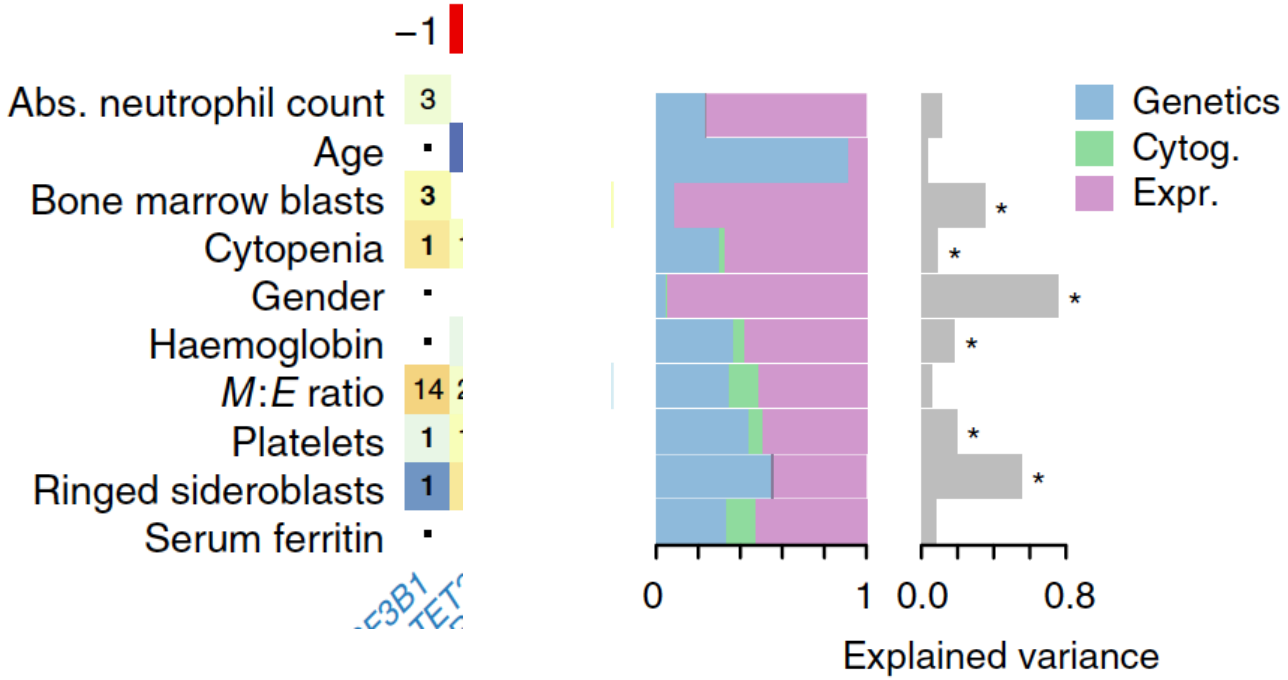


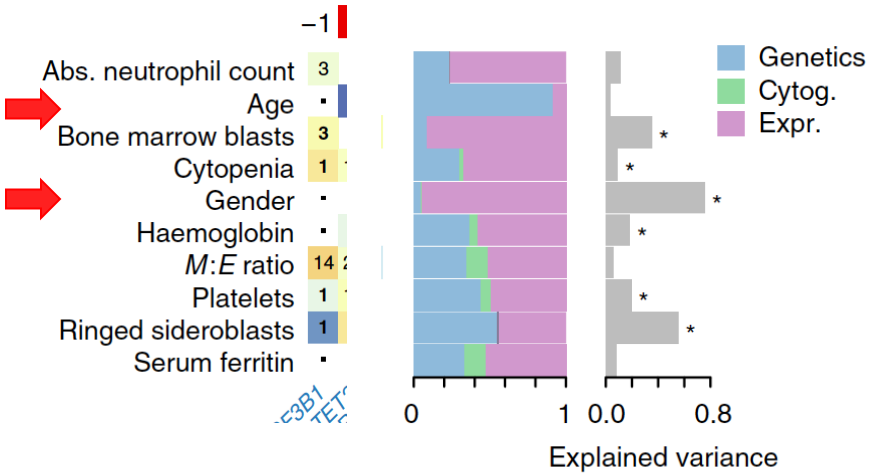
FIGURE 3C

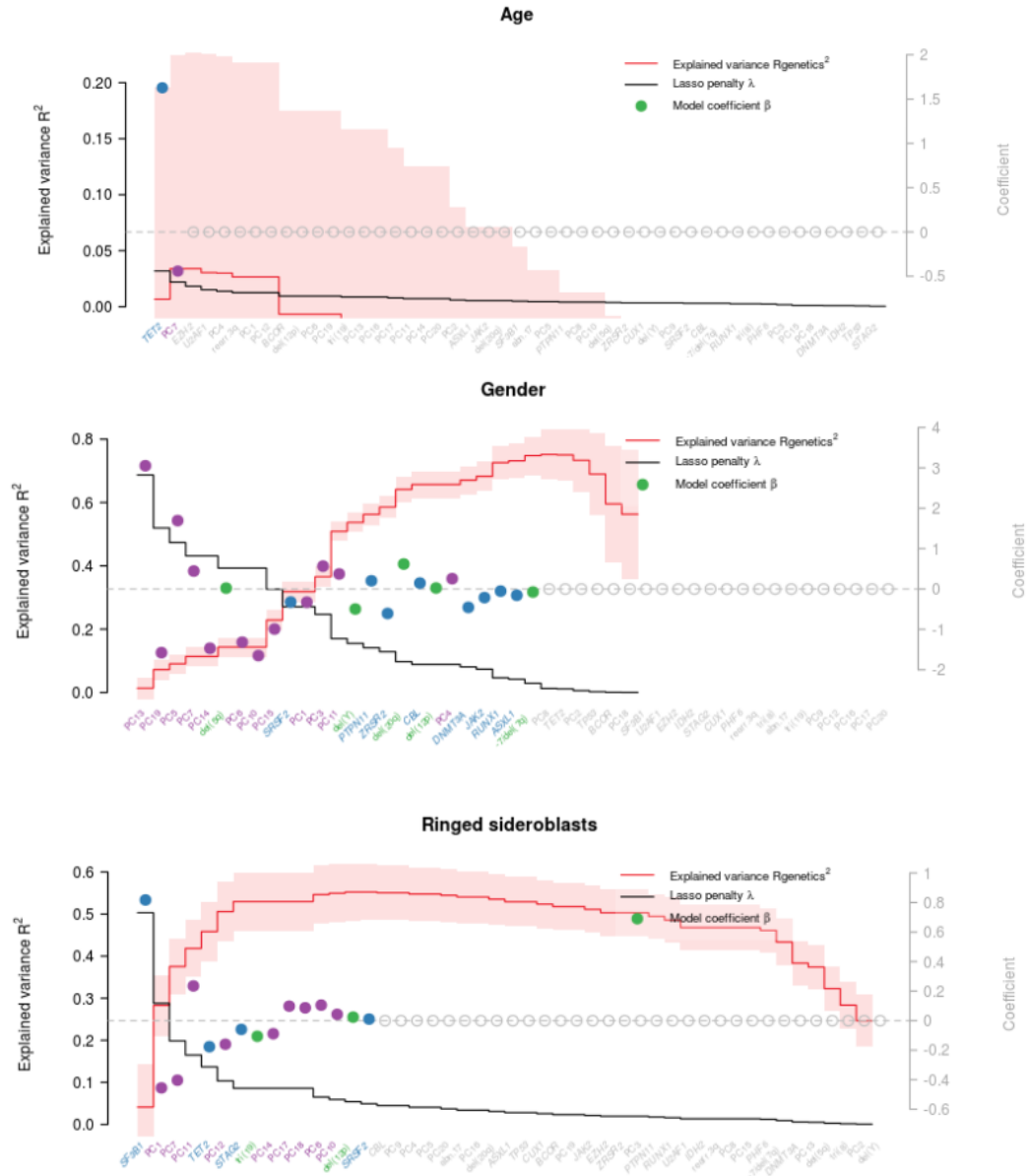
THIS SHOWS THE COMBINATION FROM THE THREE DATA TYPES



CONTRIBUTIONS FROM THE THREE CLASSES VARY AMONGST MODELS

FIGURE 3C


−1
Abs. neutrophil count
Age
Bone marrow blasts
Cytopenia
Gender
Haemoglobin
M:E ratio
Platelets
Ringed sideroblasts
Serum ferritin
Genetics
Cytog.
Expr.
0
1
0.0
0.8
Explained variance


Age
Gender
Ringed sideroblasts
Explained variance Rgenetics²
Lasso penalty λ
Model coefficient β
Explained variance R²
Coefficient

Applying these findings in a practical context – patient prognosis

FIGURE 4

FIGURE 4: SURVIVAL AND PROGNOSIS

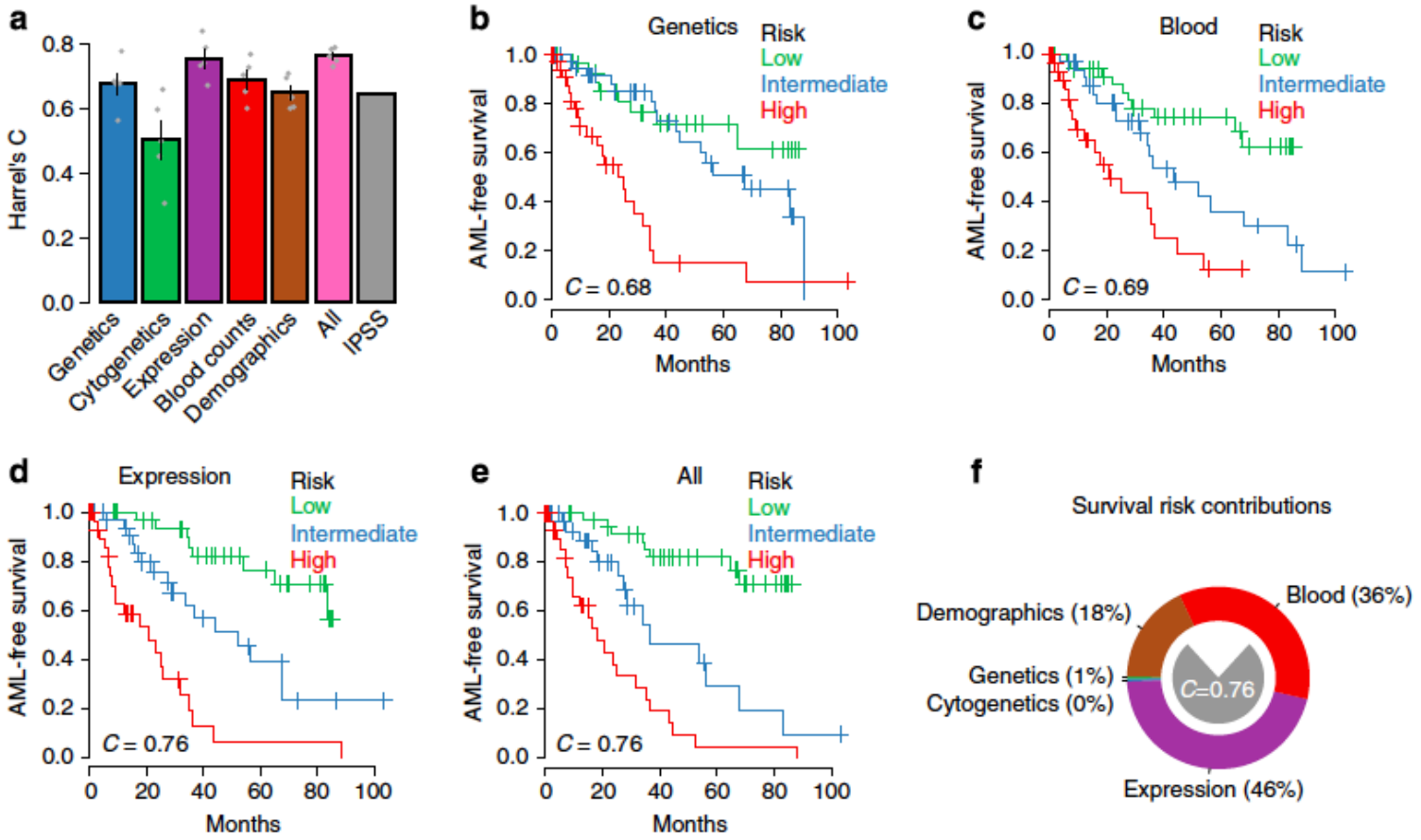


FIGURE 4: SURVIVAL AND PROGNOSIS

all this is very interesting, but can this integrated dataset be used in a practical setting to deliver a patient progress?

Currently, the International Prognostic Scoring System is used

multivariate Cox proportional hazards model

For patient i, the hazard function is given by

$$\lambda_i(t) = \lambda_0(t) \exp\left(- \sum_{j=1}^{p} X_{ij}\beta_j \right)$$

Where $X_{ij}(t)$, is the covariation of hazard X_j at time t and β_I is the fitted coefficient.

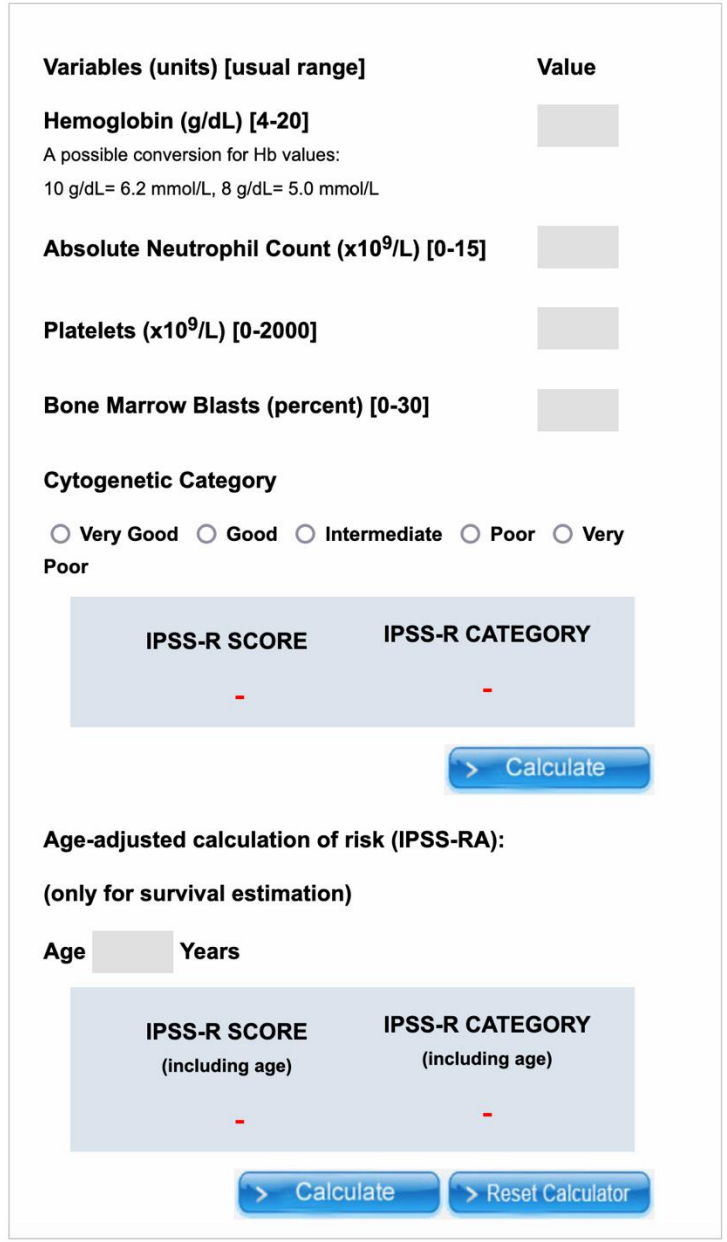
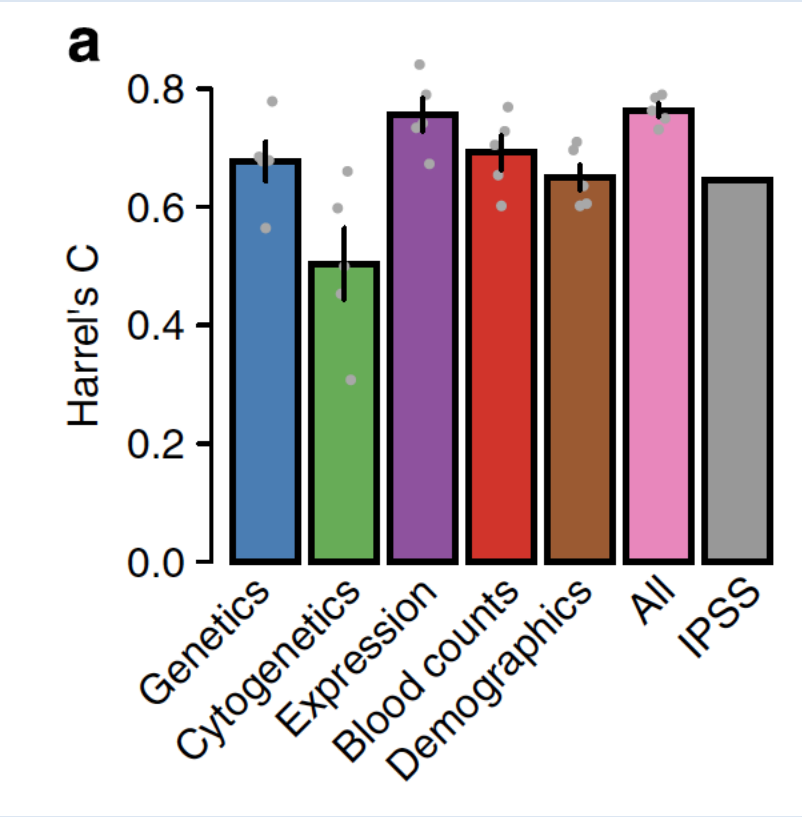


FIGURE 4: SURVIVAL AND PROGNOSIS

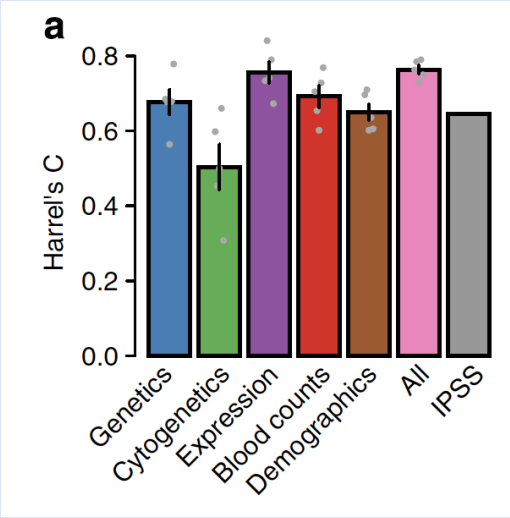
$$Harrel's\ C\ Statistic = \frac{NO\ OF\ CONCORDANT\ PAIRS}{NO\ OF\ CONCORDANT\ PAIRS + NO\ OF\ DISCORDANT\ PAIRS}$$



If the hazard risk model is good, patients who had shorter survival times should have higher risk scores.

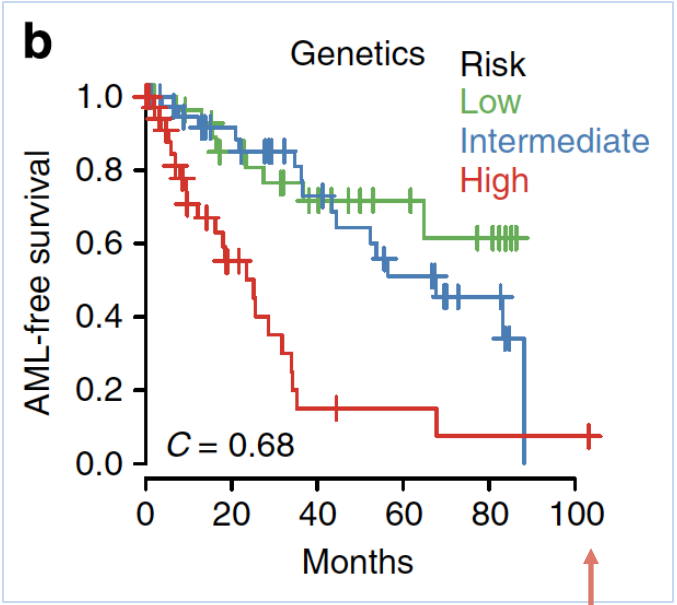
Best model performance is from Expression data (PCA components) or putting everything together

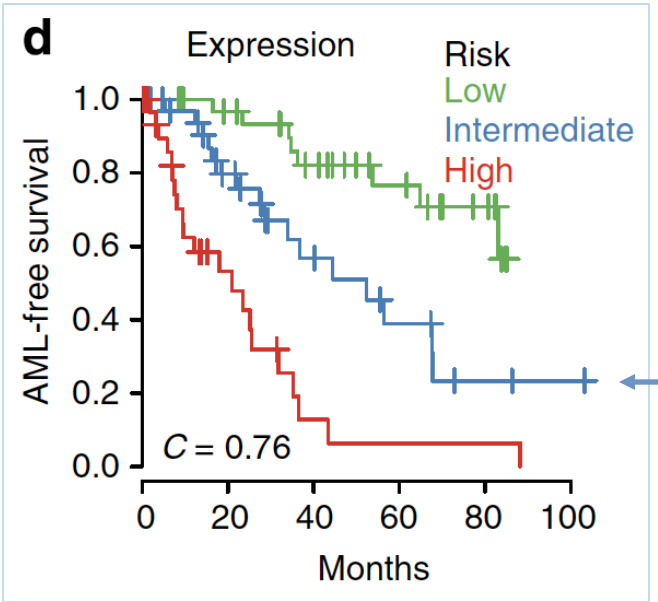
FIGURE 4: SURVIVAL AND PROGNOSIS



Kapler-Meier curves are also presented to show more details of each model performance

The + marks right censored data points (where the patient left the study before their outcome was determined)





The genetics based model has poorer predictive performance. Patients predicted to be high risk are surviving out to 100+ months

The expression based model has better predictive performance. This patient is now classified as intermediate risk

CONCLUDING COMMENTS

The paper demonstrates it is possible to integrate multiple datatypes to gain insight into the molecular processes associated with a condition.

It also highlights the interconnectivity of the datatypes and demonstrate the prognostic potential of the approach

But, there are challenges with ensuring the reproducibility of the analysis and data handling (integration, standardization and sharing)