Supplementary Information

Further information regarding patient selection and controls

The control samples were obtained from healthy members of staff at The Royal London Hospital during the same time collection of samples from the cases was being undertaken. It was not possible to obtain blood counts on these controls. The age range of these control individuals was 22-63 years and there was a female to male ratio of 2:1.

For the characterised patient groups, all cases underwent a chromosomal breakage test and only those that were found to have a mutation in either *FANCA* or *FANCG* gave a positive result. The disease-causing mutations in these cases were identified by screening with an in-house custom capture targeted exon panel containing 111 disease genes (Supplemental Table 2). Sanger sequencing confirmed the mutation. Characterisation of the DKC1 and SDS patients was initially based on the clinical presentation. Targeted Sanger sequencing of *DKC1* or *SBDS* respectively, identified the underlying genetic cause.

The uncharacterised/unsolved cases were selected for inclusion in the study based on their initial clinical presentation. All cases were under the clinical care of the same haematologist. All were negative for the chromosomal breakage test ruling out FA. The grouping of these cases is based on their clinical presentation and related blood counts. While it is accepted that this may not be the most reliable way of grouping these samples, it is logical that for uncharacterised bone marrow failure cases matching by blood counts in the first instance makes the most sense with the caveat that the underlying genetic cause may be different[1]. All cases were also screened for potential disease-causing variants by using the in-house candidate gene custom capture panel containing 111 genes that are associated with known bone marrow failure. No potentially pathogenic variants were detected by this panel. Further work needs to be undertaken to elucidate the underlying genetic variant(s) in these cases.

RNA sequencing is highly reproducible between batches

RNA-sequencing analysis was performed on a group of three patients with genetically characterised FA and three unrelated healthy controls. This was performed in duplicate to assess the reproducibility of the experiment (analysis: FANC and FANC.1). Differential expression analysis of the FANC data set showed there were 875 dysregulated genes with a False discovery rate (FDR) <0.05 and of these 710 were up regulated and 165 were down regulated. For the FANC.1 experiment the numbers of dysregulated genes were 804, 615 and 189, respectively. This is also seen in the volcano plots showing a skewing to up regulation in these experiments (Supplemental Figure 1A and 1B). A plot of the $\log_2$ fold change of FANC against FANC.1 showed very similar levels of dysregulation, with a correlation of $R^2 = 0.94$ (Supplemental Figure 1C). To validate the data, we selected a small subset of genes for analysis by qPCR. After normalising the expression to three control genes, we confirmed the relative levels in the patients were as predicted by the RNAseq results. Fold change detected by qPCR from total RNA correlated very well with the fold change recorded in mRNA by RNAseq for the significantly dysregulated genes (Supplemental Figure 1D). These data demonstrate that batch has little effect on the reproducibility of the data. Where possible sample libraries were prepared and run in the same batch thus further minimising variability.

Additional reference

1. Blombery P, Fox L, Ryland GL, Thompson ER, Lickiss J, McBean M, et al. Utility of clinical comprehensive genomic characterization for diagnostic categorization in patients presenting with hypocellular bone marrow failure syndromes. Haematologica. 2021;106(1):64-73.

Supplementary Tables:

Supplemental Table 1 – Blood counts for the patients enrolled in the study.

Supplemental Table 2 – List of genes included in the in-house bone marrow failure targeted gene panel.

Supplemental Table 3 – DESeq2 analysis data for all patient groups (FANC, SBDS, DKC1, TRI and SNGL (Excel spread sheet with multiple tabs)

Supplemental Table 4 – GSEA analysis of all comparisons detailing all the positively and negatively enriched gene sets with a cut off FWDR q-val <0.05. (Excel spread sheet)

Supplemental Table 5 – Shared significantly up regulated genes as identified by 3-way and 5-way Venn diagrams.

Supplemental Table 6 – Shared and gene specific dysregulated gene lists for the intersection of up and down regulated genes. (Excel spread sheet)

Supplemental Table 7 – Panther analysis of shared and gene-specific dysregulated genes (Excel spread sheet with multiple tabs)

Supplementary Figures:

Supplemental Figure 1 - Reproducible dysregulation of gene expression in Fanconi anemia demonstrating minimal batch variability.

Supplemental Figure 2 – Comparison of overall gene expression as determined by DESeq2 analysis as a measure of Fold change and False discovery rate (FDR)

Supplemental Figure 3 – 2D PCA plots showing cases v controls for FANC, DKC1 and SBDS.

Supplemental Table 1 - Blood counts on patients enrolled in this study.

ID	Group	Hb (M: 13-17; F: 12-16 g/dl)	WCC (4-11 x10^9/l	Platelets (150-410 x10^9/l)	Neutro-phils (2-7 X10^9/l)	Lympho-cytes (1-3 X10^9/l	Mono-cytes (0.2-1 X10^9/l)	Eosino-phils (0-0.5 X10^9/l)	CD3+ T cells (918-2023 X10^6/l)	CD4+ T cells (455-1320 X10^6/l	CD8+ T cells (140-906 X10^6/l)	CD19+ B cells (42-461 X10^6/l)	CD56+ NK cells (90-600 X10^6/l)
4457	FANC	14.4	**2.9**	207	**1.7**	**0.5**	0.6	0.2	**679**	**217**	467	**1**	160
4462	FANC	12.6	**2.4**	177	**1.4**	**0.9**	**0.1**	0	970	**358**	615	**1**	**33**
4501	FANC	13.5	**2.7**	161	**0.9**	1.3	0.3	0.2	1115	549	532	**1**	257
4705	DKC1	13.6	6.2	189	3.3	1.9	0.8	0.1	1323	627	520	315	294
4740	DKC1	14.6	7.5	**84**	2.3	**4.3**	0.7	0.1	**4150**	**1527**	**2549**	74	279
4702	DKC1	15.1	9.3	279	6.2	2.0	0.8	0.2	1430	704	704	243	406
4551	SBDS	14.2	**2.2**	**103**	**0.8**	1.1	0.4	0	**832**	**446**	386	**35**	164
4575	SBDS	14.6	**2.6**	**141**	**1.1**	1.3	0.2	0	1078	737	328	63	**85**
4642	SBDS	14.0	**2.1**	**101**	**0.7**	1.1	0.3	0	**836**	**429**	371	179	196
4504	TRI	**8.3**	**2.6**	**109**	**1.6**	**0.6**	0.2	0.2	**381**	**231**	153	89	**74**
4545	TRI	**7.5**	**2.0**	**48**	**0.6**	1.2	0.2	0.1	1091	572	532	**39**	**68**
4592	TRI	**7.9**	**2.1**	**9**	**0.6**	1.3	0.2	0	**-**	**-**	**-**	**-**	**-**
4520	SNGL	**10.7**	6.7	410	4.2	1.7	0.5	0.3	1998	725	417	260	273
4633	SNGL	14.2	**3.5**	197	**1.9**	1.2	0.3	0	1187	847	333	154	**71**
4680	SNGL	13.1	**2.3**	289	**1.4**	**0.6**	0.2	0	**410**	**319**	**108**	132	**79**

Normal ranges are given in brackets. M – males, F- females. Abnormal values are in bold (underlined bold is higher than normal). For 4740 the high T-lymphocyte count gives a low B cell percentage/ratio. – data not available

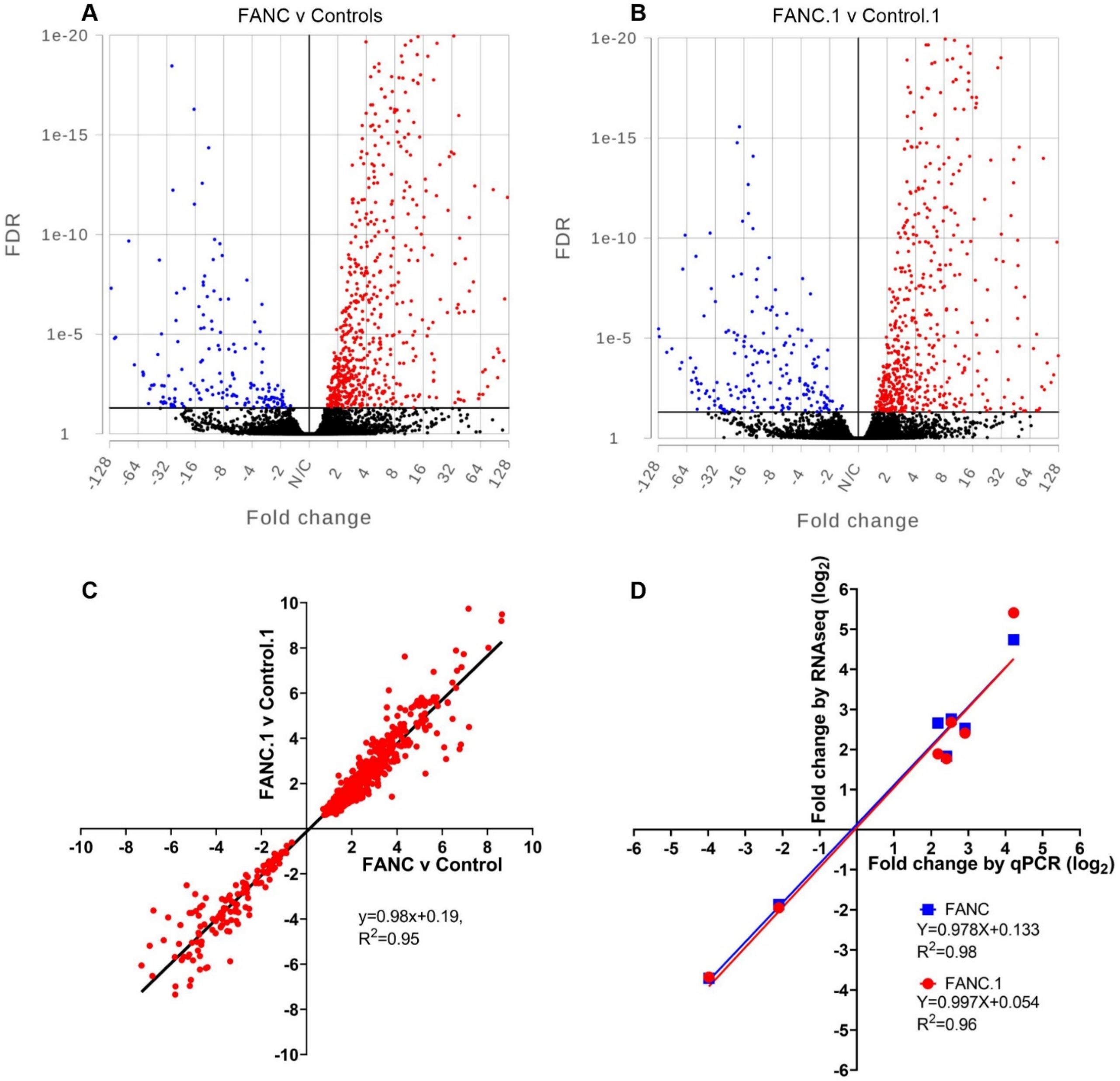
Supplemental Table 2 – List of genes included on the in-house candidate gene screening panel.

ACD	EFL1	GATA1	PALB2	RPS10	STN1
ADA2	ELANE	GATA2	PARN	RPS17	TAZ
ANKRD26	ERBB3	GFI1	POLA1	RPS19	TERC
BRCA1	ERCC4	GRHL2	POT1	RPS24	TERT
BRCA2	ERCC6L2	HAX1	RAD51	RPS26	THPO
BRIP1	ETV6	HOXA11	RAD51C	RPS27	TINF2
C15orf41	FANCA	JAGN1	RBM8A	RPS28	TP53
CDAN1	FANCB	KIF23	RECQL4	RPS29	TYMS
CEBPA	FANCC	KLF1	RFWD3	RPS7	UBE2T
CSF3R	FANCD2	LIG4	RMRP	RTEL1	USB1
CTC1	FANCE	LPIN2	RPL11	RUNX1	VPS45
CXCR4	FANCF	MAD2L2	RPL15	SAMD9	WAS
CYCS	FANCG	MECOM	RPL18	SAMD9L	WRAP53
DCLRE1B	FANCI	MPL	RPL26	SBDS	XRCC2
DDX41	FANCL	MYSM1	RPL27	SEC23B	ZCCHC8
DKC1	FANCM	NAF1	RPL31	SHQ1	
DNAJC21	FYB1	NHP2	RPL35A	SLX4	
DNAJC3	G6PC	NOP10	RPL5	SP1	
DUT	G6PC3	NPM1	RPL9	SRP54	

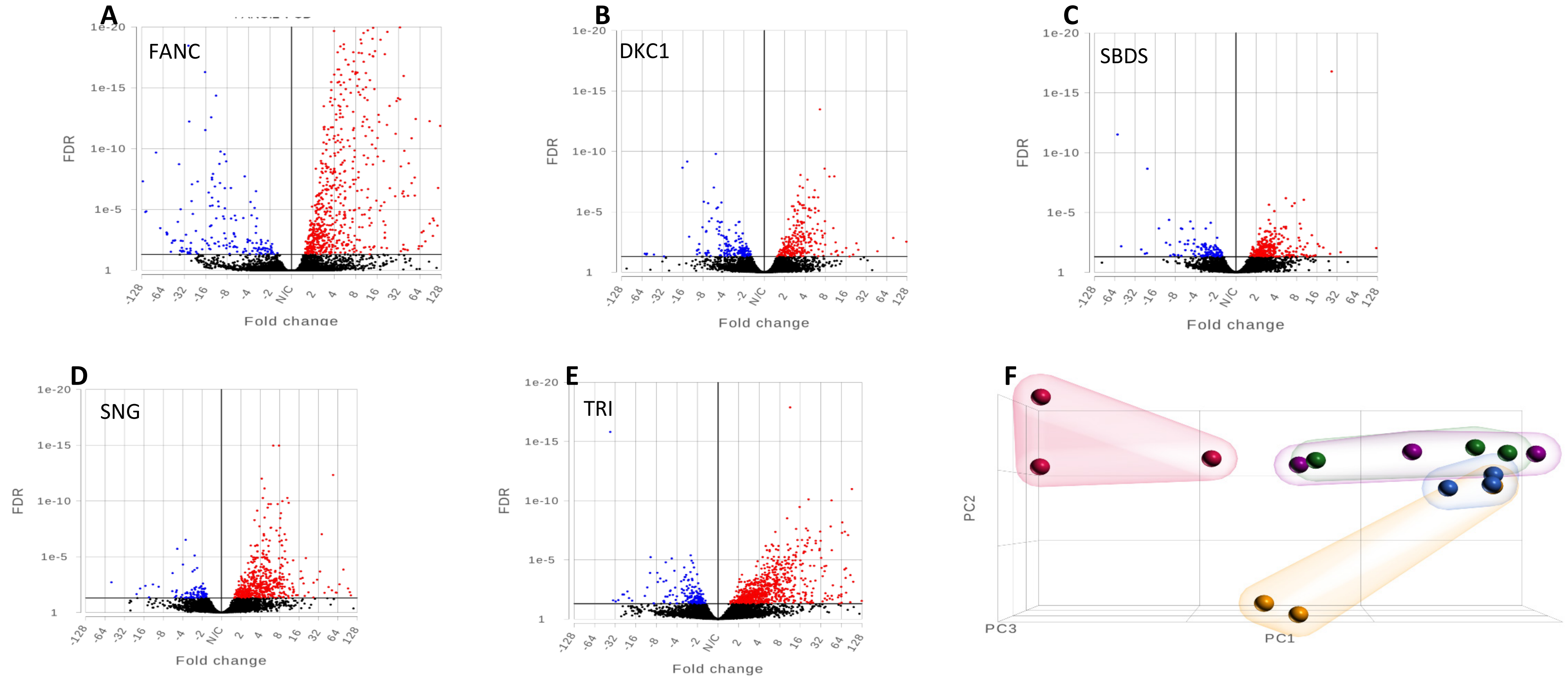
Supplemental Table 5 – Shared significantly up regulated genes as identified by 3-way and 5-way Venn diagrams.

HUGO	Name	Relevant biological function*	Key process
ANXA1	Annexin A1	inflammatory process	inflammation
ATP5MG	ATP synthase subunit g, mitochondrial	ATP synthesis	mitochondrion
COX7C	Cytochrome c oxidase subunit 7C, mitochondrial	mitochondrial, oxidase	mitochondrion
HP	Haptoglobin	haemoglobin binding	acute-phase response
HSP90AA1	Heat shock protein HSP 90-alpha	molecular chaperone	stress response
MYL6	Myosin light polypeptide 6	actin family cytoskeletal protein, calmodulin	muscle protein
RBX1	E3 ubiquitin-protein ligase RBX1	ubiquitin-protein ligase	ubiquitination
RETN	Resistin	promotes chemotaxis in myeloid cells	neutrophil
RPL7	60S ribosomal protein L7	ribosomal protein	translation
RPL9	60S ribosomal protein L9	ribosomal protein	translation
RPL11	60S ribosomal protein L11	ribosomal protein	translation
RPL21	60S ribosomal protein L21	ribosomal protein	translation
RPL23	60S ribosomal protein L23	ribosomal protein	translation
RPL26	60S ribosomal protein L26	ribosomal protein	translation
RPL31	60S ribosomal protein L31	ribosomal protein	translation
RPL41	60S ribosomal protein L41	ribosomal protein	translation
RPS3A	40S ribosomal protein S3a	ribosomal protein, cysteine protease	translation
RPL21P16	ribosomal protein L21 pseudogene 16	n/a	n/a
S100A8	Protein S100-A8	calmodulin, signalling molecule	signalling
SELENOK	Selenoprotein K	T-cell proliferation	other
SERF2	Small EDRK-rich factor 2	amyloid protein aggregation	other
TMA7	Translation machinery-associated protein 7	cytoplasmic translation	translation
TOMM7	Mitochondrial import receptor subunit TOM7 homolog	protein targeting to mitochondrion	mitochondrion
TPT1	Translationally controlled tumour protein	non-motor microtubule binding protein	other
TXN	Thioredoxin	mitochondrial reactive oxygen species homeostasis	mitochondrion
UQCR11	Cytochrome b-c1 complex subunit 10	part of the mitochondrial respiratory chain	mitochondrion

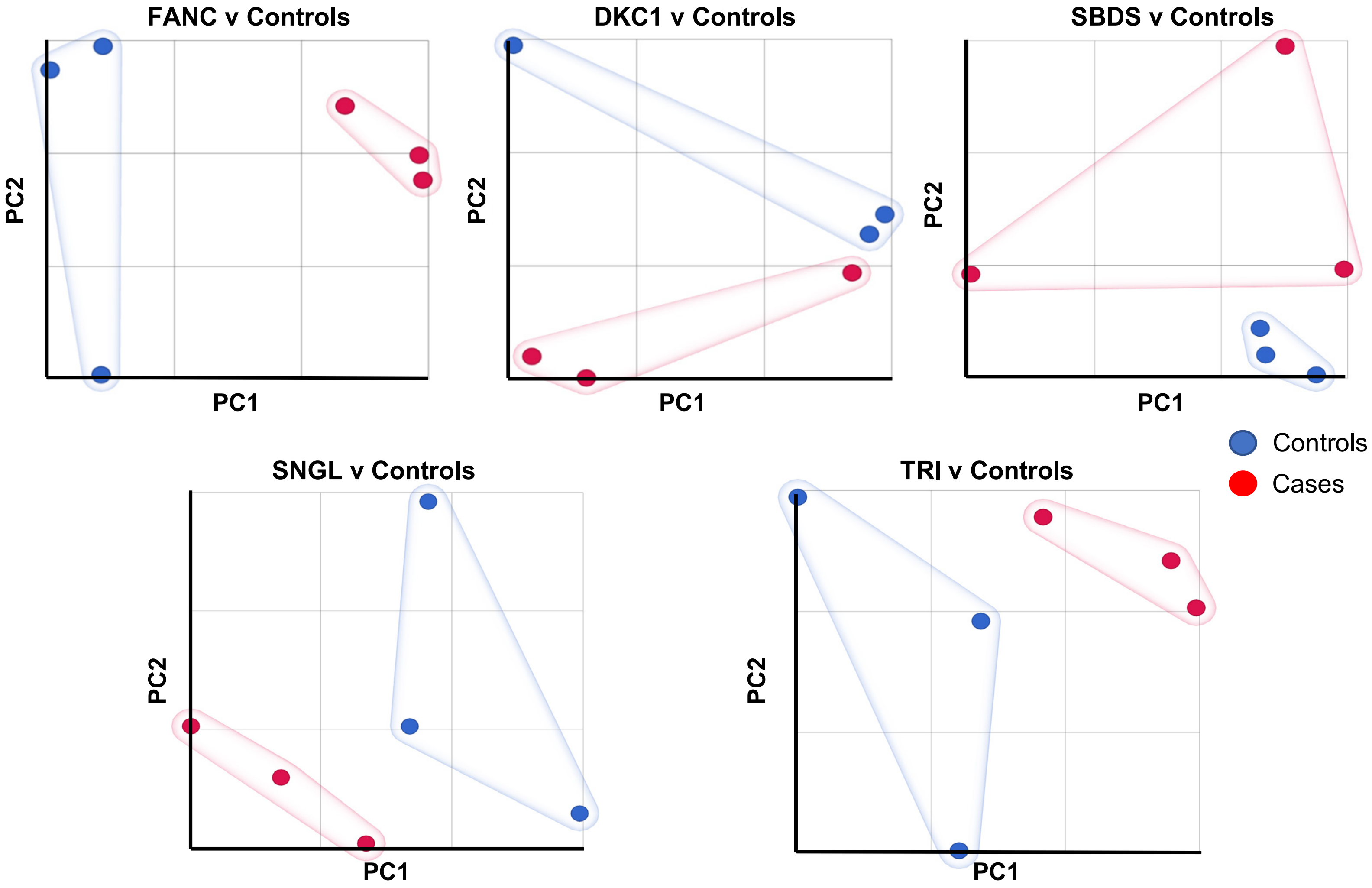
Listing relevant biological function as described by Uniprot* and the key process. Genes highlighted in red are only present in the intersection of upregulated genes from FANC DKC1 and SBDS.



Supplemental Figure 1 – Reproducible dysregulation of gene expression in Fanconi anemia demonstrating minimal batch variability. A and B Volcano plots showing a similar pattern of dysregulation between the duplicate experiments (FDR<0.05, red -upregulated genes, blue-down regulated genes). **C.** Scatter plot of the fold changes observed for the shared dysregulated genes shows good reproducibility. **D.** Scatter plot of $\log_2$ fold change determined by qPCR compared with RNA-seq data from two FANC experiments.



Supplemental Figure 2 – Comparison of overall gene expression as determined by DESeq2 analysis as a measure of Fold change and False discovery rate (FDR) A-E- Volcano plots of the patient groups included in this study showing differential gene expression when compared with healthy controls. For clarity scales are set the same so some points may be omitted from the plot. Each point represents a gene. red - up regulated genes, blue – down regulated genes (all at FDR<0.05) black – genes that fail to meet the FDR<0.05 threshold. F – PCA plot showing all the patient used in this study (blue-FANC cases, red-DKC1 cases, yellow-SBDS cases, green-TRI cases, purple – SNGL cases).



Supplemental Figure 3 – 2D PCA plots showing cases v controls for all groups of cases in this study. In each case the red circles represent the cases and the blue circles the controls