



# Development and validation of a novel cross-cancer immunogenomics analysis model & web-based software

Dr Reza Rafiee, 26/06/2018



# Development and validation of a novel cross-cancer immunogenomics analysis model & web-based software

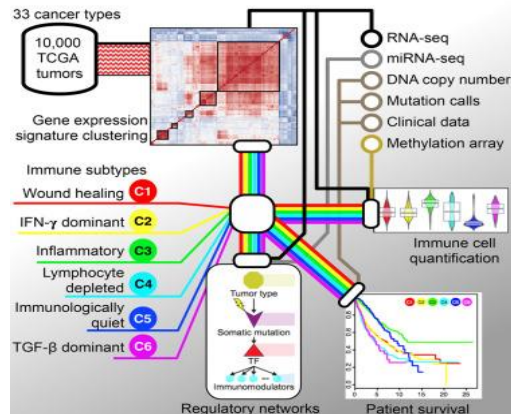
- Develop an analytically validated, gene expression analysis model and software applicable to multiple platforms, including RNA-seq and cDNA microarray technology which can
  - Score the immune gene expression signatures of patient FFPE tumour samples (or a single sample)
  - Prospectively classify patient FFPE tumour samples into one of the six consensus immune molecular subtypes reported by [PanCan Atlas](#).
- Apply the software to clinically annotated gene expression data from multiple sites including prostate, breast, ovarian, lung, melanoma and colon cancer provided by Almac Diagnostics.
- Assess the association between each immune molecular subtype and patient outcomes following conventional and immune-based therapies.

Completed   In progress   Not started

- VM server
  - Specifying minimum hardware requirements for the project, assessing external prices from AWS and DigitalOcean company (8 cores, 16GB RAM, 256GB system disk)
  - Specifying software and libraries/packages required for the project
  - Install software, libraries and packages
- PanCan Atlas dataset preparation
  - Downloading, extracting signature genes (based on *The Immune Landscape of Cancer*), extracting corresponding RNA-seq data matched with signature genes, extracting the overlapped genes across signatures (in two dataset)
  - Harmonising variables (sample Ids and gene Ids) in different datasets including retrieving alternative gene ids from genecard.org, assessing the distribution of missing data
  - Single sample gene set enrichment analysis (ssGSEA)
  - Training & validation dataset preparation and clustering analysis
- Assess machine learning API tools and libraries for our platform such as TensorFlow
- DDRD implementation (supervision and implementation)
  - Addressing missing with multiple imputation, exception handling modules (Amanda)
  - Identifying the DDRD threshold in an empirical manner (Amanda)
  - GUI designing (Myself)
  - Assessing confidence interval for each DDRD score (Myself)

# Dataset, resources and methods

- **30** Cancer subtypes (non-hematologic)
- Total number of samples including immune subgroups (after removing samples without subgroups): **9126**
- Number of genes in 5 immune signatures: **446** (5 overlapped)
- Number of missing genes in RNA-seq dataset: **1**
- Input data: Raw or normalised gene expression data from RNA-seq and cDNA microarray platforms
  - Normalisation:
    - Microarray: Log2 transformed of read count
    - RNA-seq: FPKM (normalised based on only the immune signature genes)
- Immune gene expression signatures scoring (ssGSEA & WGCNA)
- An ensemble classification model (including SVM classifier) or Deep learning model
- Web-based software development running on a VM server hosted in Queen's university Belfast:  
<http://smg.qub.ac.uk:3838>







160 Immune  
signatures based  
on opinions of  
immuno-  
oncologist experts

Including 4691 genes

Source	SetName	Source	SetName	Source	SetName	Source	SetName
Senbabaoğlu	Angiogenesis	Wolf	CD8_PCA_16704732	c7atoms	C7_Atom_17	Gbersort	T.cells.gamma.delta
Senbabaoğlu	APM1	Wolf	GRANS_PCA_16704732	c7atoms	C7_Atom_18	Gbersort	NK.cells.resting
Senbabaoğlu	APM2	Wolf	LYMPHS_PCA_16704732	c7atoms	C7_Atom_19	Gbersort	NK.cells.activated
Wolf	ICS5_score	Wolf	T_cell_PCA_16704732	c7atoms	C7_Atom_20	Gbersort	Monocytes
Wolf	Llexpression_score	Wolf	TGFB_PCA_17349583	c7atoms	C7_Atom_21	Gbersort	Macrophages.M0
Wolf	Chemokine12_score	Wolf	Rotterdam_BRneg_PCA_15721472	c7atoms	C7_Atom_22	Gbersort	Macrophages.M1
Wolf	NHL_5gene_score	Wolf	HER2_Immune_PCA_18006808	c7atoms	C7_Atom_23	Gbersort	Macrophages.M2
Wolf	CD68	Wolf	IR7_score	c7atoms	C7_Atom_24	Gbersort	Dendritic.cells.resting
Wolf	CD8A	Wolf	Buck14_score	c7atoms	C7_Atom_25	Gbersort	Dendritic.cells.activated
Wolf	PD1_data	Wolf	TAMsurr_score	c7atoms	C7_Atom_26	Gbersort	Mast.cells.resting
Wolf	PDL1_data	Wolf	Immune_NSCLC_score	c7atoms	C7_Atom_27	Gbersort	Mast.cells.activated
Wolf	PD1_PDL1_score	Wolf	Module3_IFN_score	c7atoms	C7_Atom_28		
Wolf	CTLA4_data	Wolf	Module4_TcellBcell_score	c7atoms	C7_Atom_29		
Wolf	Bcell_mg_IQJ	Wolf	Module5_TcellBcell_score	c7atoms	C7_Atom_30		
Wolf	Bcell_receptors_score	Wolf	Module11_Prolif_score	c7atoms	C7_Atom_31		
Wolf	STAT1_score	Wolf	GP11_Immune_IFN	c7atoms	C7_Atom_32		
Wolf	CSF1_response	Wolf	GP2_ImmuneTcellBcell_score	Bindea	aDC		
Wolf	TcClassII_score	Wolf	CD8_CD68_ratio	Bindea	B cells		
Wolf	IL12_score_21050467	Wolf	TAMsurr_TcClassII_ratio	Bindea	CD8 T cells		
Wolf	IL4_score_21050467	Wolf	CHANG_CORE_SERUM_RESPONSE_UP	Bindea	Cytotoxic cells		
Wolf	IL2_score_21050467	Wolf	CSR_Activated_15701700	Bindea	DC		
Wolf	IL13_score_21050467	Wolf	CD103pos_CD103neg_ratio_25446897	Bindea	Eosinophils		
Wolf	IFNG_score_21050467	Attractors	LYM	Bindea	iDC		
Wolf	TGFB_score_21050467	Attractors	IFIT3	Bindea	Lymph vessels		
Wolf	TREM1_data	Attractors	G_GIMAP4	Bindea	Macrophages		
Wolf	DAP12_data	Attractors	G_HLA.DPA1	Bindea	Mast cells		
Wolf	Tcell_receptors_score	Attractors	G_SLAMF6	Bindea	Neutrophils		
Wolf	IL8_21978456	Attractors	G_LILRB4	Bindea	NK CD56bright cells		
Wolf	IFN_21978456	Attractors	G_SIGLEC9	Bindea	NK CD56dim cells		
Wolf	MHCI_21978456	Attractors	G_CYTH4	Bindea	NK cells		
Wolf	MHC2_21978456	Attractors	G_CD3E	Bindea	pDC		
Wolf	Bcell_21978456	ICR	ICR_SCORE	Bindea	T cells		
Wolf	Tcell_21978456	ICR	ICR_INHIB_SCORE	Bindea	T helper cells		
Wolf	CD103pos_mean_25446897	ICR	ICR_ACT_SCORE	Bindea	Tcm cells		
Wolf	CD103neg_mean_25446897	c7atoms	C7_Atom_1	Bindea	Tem cells		
Wolf	IgG_19272155	c7atoms	C7_Atom_2	Bindea	Tfh cells		
Wolf	Interferon_19272155	c7atoms	C7_Atom_3	Bindea	Tgd cells		
Wolf	LQK_19272155	c7atoms	C7_Atom_4	Bindea	Th1 cells		
Wolf	MHCI_19272155	c7atoms	C7_Atom_5	Bindea	Th17 cells		
Wolf	MHCII_19272155	c7atoms	C7_Atom_6	Bindea	Th2 cells		
Wolf	STAT1_19272155	c7atoms	C7_Atom_7	Bindea	Treg cells		
Wolf	Troester_WoundSig_19887484	c7atoms	C7_Atom_8	Gbersort	B.cells.naive		
Wolf	MDACC.FNA.1_20805453	c7atoms	C7_Atom_9	Gbersort	B.cells.memory		
Wolf	IGG_Cluster_21214954	c7atoms	C7_Atom_10	Gbersort	Plasma.cells		
Wolf	Minterferon_Cluster_21214954	c7atoms	C7_Atom_11	Gbersort	T.cells.CD8		
Wolf	Immune_cell_Cluster_21214954	c7atoms	C7_Atom_12	Gbersort	T.cells.CD4.naive		
Wolf	MCD3_CD8_21214954	c7atoms	C7_Atom_13	Gbersort	T.cells.CD4.memory.resting		
Wolf	Interferon_Cluster_21214954	c7atoms	C7_Atom_14	Gbersort	T.cells.CD4.memory.activated		
Wolf	B_cell_PCA_16704732	c7atoms	C7_Atom_15	Gbersort	T.cells.follicular.helper		
		c7atoms	C7_Atom_16	Gbersort	T.cells.regulatory..Tregs.		

# 83 Immune signatures from 4 studies (n=2665 genes)

Which are known to be associated with immune activity in tumour tissue

Source	SetName	# of genes	Source	SetName	# of genes	Source	SetName	# of genes	Source	SetName	# of genes
Şenbabaoğlu	Angiogenesis		Wolf	IL13_score_21050467		Wolf	IGG_Cluster_21214954		Wolf	GP2_ImmuneTcellBcell_score	
Şenbabaoğlu	APM1		Wolf	IFNG_score_21050467		Wolf	Minterferon_Cluster_21214954		Wolf	CD8_CD68_ratio	
Şenbabaoğlu	APM2		Wolf	TGFB_score_21050467	80	Wolf	Immune_cell_Cluster_21214954		Wolf	TAMsurr_TcClassII_ratio	
Wolf	ICS5_score		Wolf	TREM1_data		Wolf	MCD3_CD8_21214954		Wolf	CHANG CORE SERUM RESPONSE UP	212
Wolf	L1expression_score	18	Wolf	DAP12_data		Wolf	Interferon_Cluster_21214954		Wolf	CSR_Activated_15701700	
Wolf	Chemokine12_score		Wolf	Tcell_receptors_score		Wolf	B_cell_PCA_16704732		Wolf	CD103pos_CD103neg_ratio_25446897	
Wolf	NHI_5gene_score		Wolf	IL8_21978456		Wolf	CD8_PCA_16704732		Attractors	LYM	
Wolf	CD68		Wolf	IFN_21978456		Wolf	GRANS_PCA_16704732		Attractors	IFIT3	
Wolf	CD8A		Wolf	MHC1_21978456		Wolf	LYMPHS_PCA_16704732		Attractors	G_GIMAP4	
Wolf	PD1_data		Wolf	MHC2_21978456		Wolf	T_cell_PCA_16704732		Attractors	G_HLA.DPA1	
Wolf	PDL1_data		Wolf	Bcell_21978456		Wolf	TGFB_PCA_17349583		Attractors	G_SLAMF6	
Wolf	PD1_PDL1_score		Wolf	Tcell_21978456		Wolf	Rotterdam_ERneg_PCA_15721472		Attractors	G_LILRB4	
Wolf	CTLA4_data		Wolf	CD103pos_mean_25446897		Wolf	HER2_Immune_PCA_18006808		Attractors	G_SIGLEC9	
Wolf	Bcell_mg_IGJ		Wolf	CD103neg_mean_25446897		Wolf	IR7_score		Attractors	G_CYTH4	
Wolf	Bcell_receptors_score		Wolf	IgG_19272155		Wolf	Buck14_score		Attractors	G_CD3E	
Wolf	STAT1_score		Wolf	Interferon_19272155		Wolf	TAMsurr_score		ICR	ICR_SCORE	
Wolf	CSF1_response	112	Wolf	LCK_19272155		Wolf	Immune_NSCLC_score		ICR	ICR_INHIB_SCORE	
Wolf	TcClassII_score		Wolf	MHC.I_19272155		Wolf	Module3_IFN_score	24	ICR	ICR_ACT_SCORE	
Wolf	IL12_score_21050467		Wolf	MHC.II_19272155		Wolf	Module4_TcellBcell_score				
Wolf	IL4_score_21050467		Wolf	STAT1_19272155		Wolf	Module5_TcellBcell_score				
Wolf	IL2_score_21050467		Wolf	Troester_WoundSig_19887484		Wolf	Module11_Prolif_score				
			Wolf	MDACC.FNA.1_20805453		Wolf	GP11_Immune_IFN				

Total = 446 genes

Of 160 signatures, 77 signatures did not affect the identified signature clusters and therefore have been removed from the final analysis.

# Dataset preparation and harmonisation

	IFN- $\gamma$ response	TGB- $\beta$ response	Activation of <b>Macrophage</b> /monocytes	Overall <b>Lymphocyte</b> infiltration	Wound healing	Total # of genes	Note
	24	80	112	18	212	446	5 overlapped genes between immune groups
	24	80	111	15	210	440	<b>IGLC1/IGLC</b> is missing* (a protein coding gene, no RNA- seq data)
		ITGB2	ITGB2 CCL5 CD8A IL7R MSN	CCL5 CD8A	IL7R MSN		Overlapped genes  In RNA-seq data (FPKM), 9 genes had alternative names (next slide)

- Harmonising datasets including gene ids for the two datasets (FPKM RNA-seq and signature/subgroup datasets)

\* Initially we had 10 missing genes ids ==> cross-checking all the alternative names of the missing gene ids (>30) across 20k genes

# Dataset preparation and harmonisation – gene ids

#	Alternative/Aliases gene ids (used in RNA-seq dataset)	Gene ids (used in signature dataset)	Note
1	CD247	CD3Z	Lymphocyte (signature group)
2	CD8B	CD8B1	Lymphocyte
3	VCAN	CSPG2	TGF- $\beta$
4	CTSL1	CTSL	Macrophage
5	CELF2	CUGBP2	Macrophage
6	FPR3	FPRL2	Macrophage
7	HDC	IGHG3	Lymphocyte
8	CYTIP	PSCDBP	Macrophage
9	C13orf1	SPRYD7	Wound healing



# Dataset preparation and harmonisation – cancer subtypes

	# of RNA-seq samples (440 genes)	# of samples (with subgroups)	Note
	10167	9126*	1041 samples without any immune subgroup (Could be used for testing by the final classifier)

## All 30 non-hematologic cancer types

	Cancer Subtype	Number of Samples
1	ACC	78
2	BLCA	397
3	BRCA	1083
4	CESC	300
5	CHOL	35
6	COAD	441
7	ESCA	173
8	GBM	154
9	HNSC	514
10	KICH	65
11	KIRC	515
12	KIRP	279
13	LGG	514
14	LIHC	362

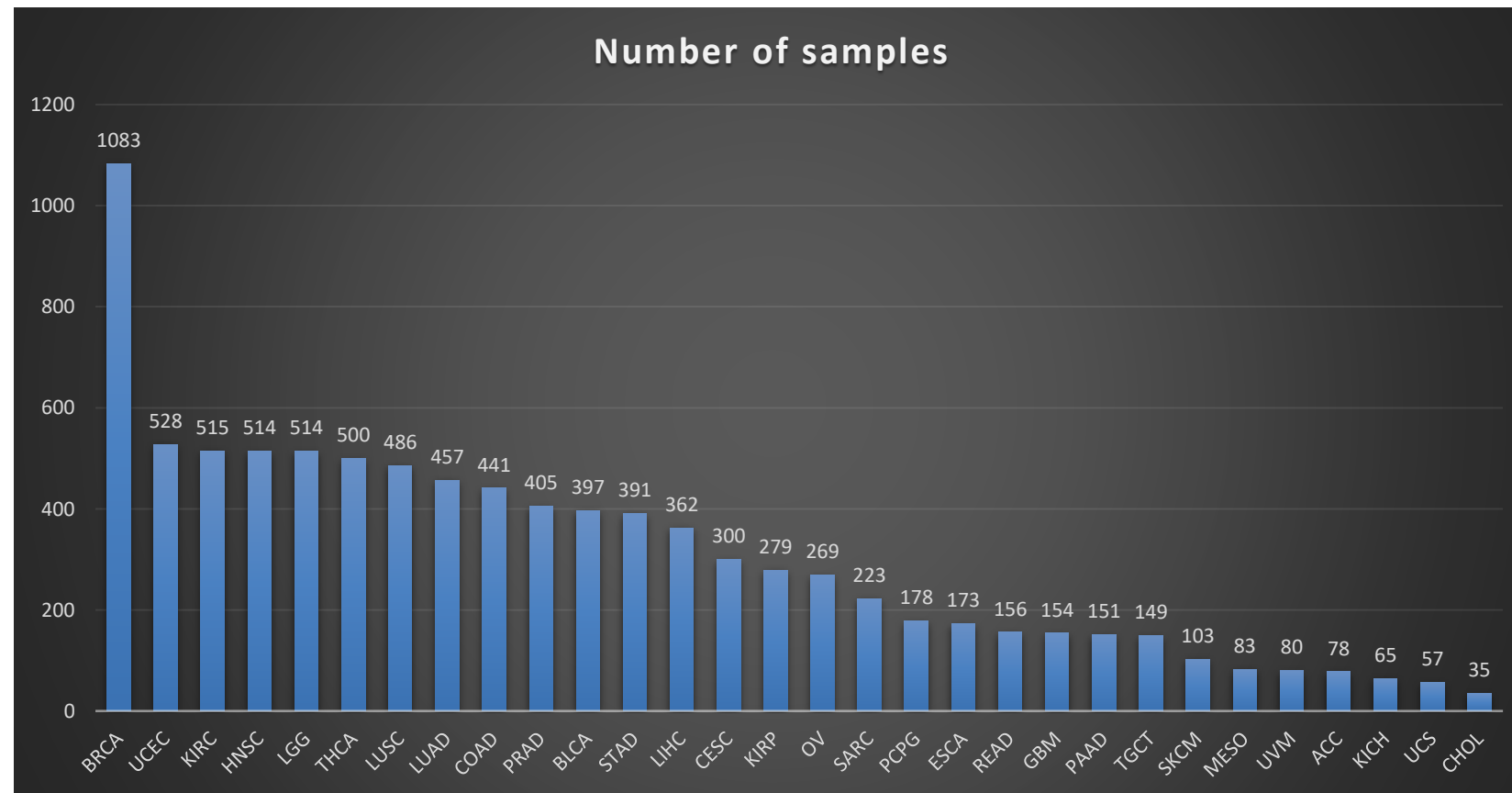
	Cancer Subtype	Number of Samples
15	LUAD	457
16	LUSC	486
17	MESO	83
18	OV	269
19	PAAD	151
20	PCPG	178
21	PRAD	405
22	READ	156
23	SARC	223
24	SKCM	103
25	STAD	391
26	TGCT	149
27	THCA	500
28	UCEC	528
29	UCS	57
30	UVM	80

\* Matched with the number of samples in the paper prior to model-based clustering



# Sample distribution across cancers (n=9126)

**BRCA:** Breast Invasive Carcinoma  
**UCEC:** Uterine Corpus Endometrial Carcinoma  
**KIRC:** Kidney Renal Clear Cell Carcinoma  
**HNSC:** Head and Neck Squamous Cell Carcinoma  
**LGG:** Low Grade Glioma  
**THCA:** Thyroid Carcinoma  
**LUSC:** Lung Squamous Cell Carcinoma  
**LUAD:** Lung Adenocarcinoma  
**COAD:** Colon Adenocarcinoma  
**PRAD:** Prostate Adenocarcinoma  
**BLCA:** Bladder Urothelial Carcinoma  
**STAD:** Stomach Adenocarcinoma  
**LIHC:** Liver Hepatocellular Carcinoma  
**CESC:** Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma  
**KIRP:** Kidney Renal Papillary Cell Carcinoma  
**OV:** Ovarian Cancer  
**SARC:** Sarcoma  
**PCPG:** Pheochromocytoma and Paraganglioma  
**ESCA:** Esophageal (Oesophageal) Carcinoma  
**READ:** Rectum Adenocarcinoma  
**GBM:** Glioblastoma Multiforme  
**PAAD:** Pancreatic Adenocarcinoma  
**TGCT:** Testicular Germ Cell Tumours  
**SKCM:** Skin Cutaneous Melanoma  
**MESO:** Mesothelioma  
**UVM:** Uveal Melanoma  
**ACC:** Adrenocortical carcinoma  
**KICH:** Kidney Chromophobe  
**UCS:** Uterine Carcinosarcoma  
**CHOL:** Cholangiocarcinoma

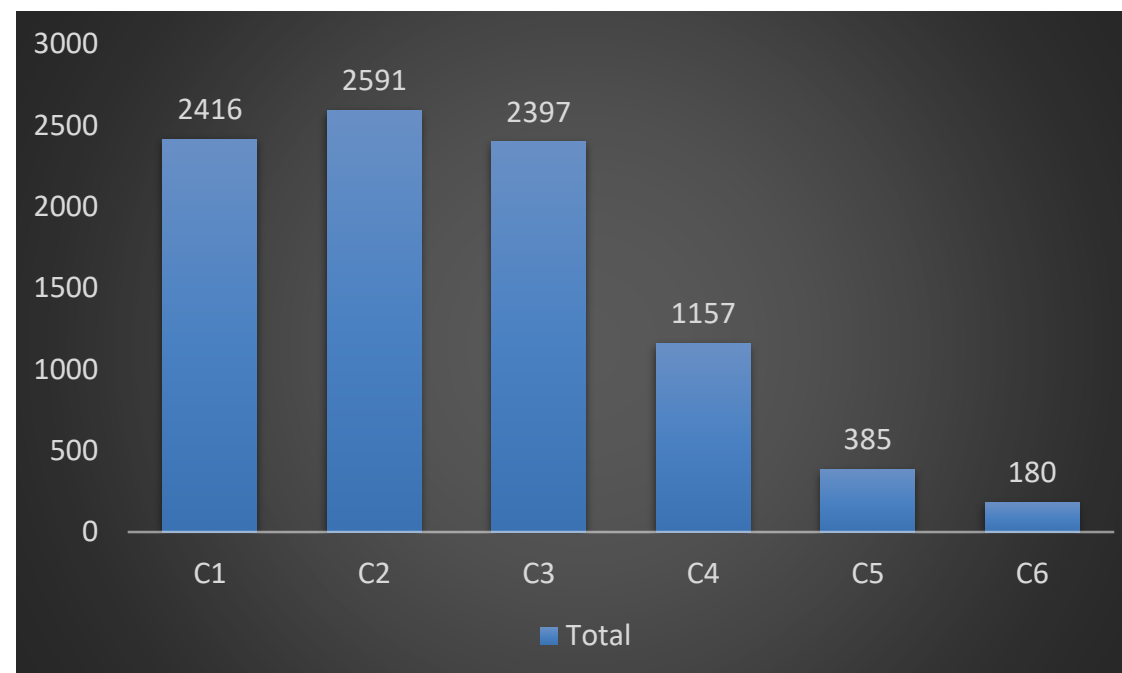


## Samples from below cancers (hematologic) without any immune subtypes

**DLBC:** Diffuse Large B-cell Lymphoma  
**LAML:** Acute Myeloid Leukemia  
**THYM:** Thymoma

# Sample distribution across immune subtype (n=9126)

Immune subtype	Number of samples in each immune subtype
C1	2416
C2	2591
C3	2397
C4	1157
C5	385
C6	180



Imbalance data across 5 classes

Penalising c5 and c6, imbalanced data/classifier



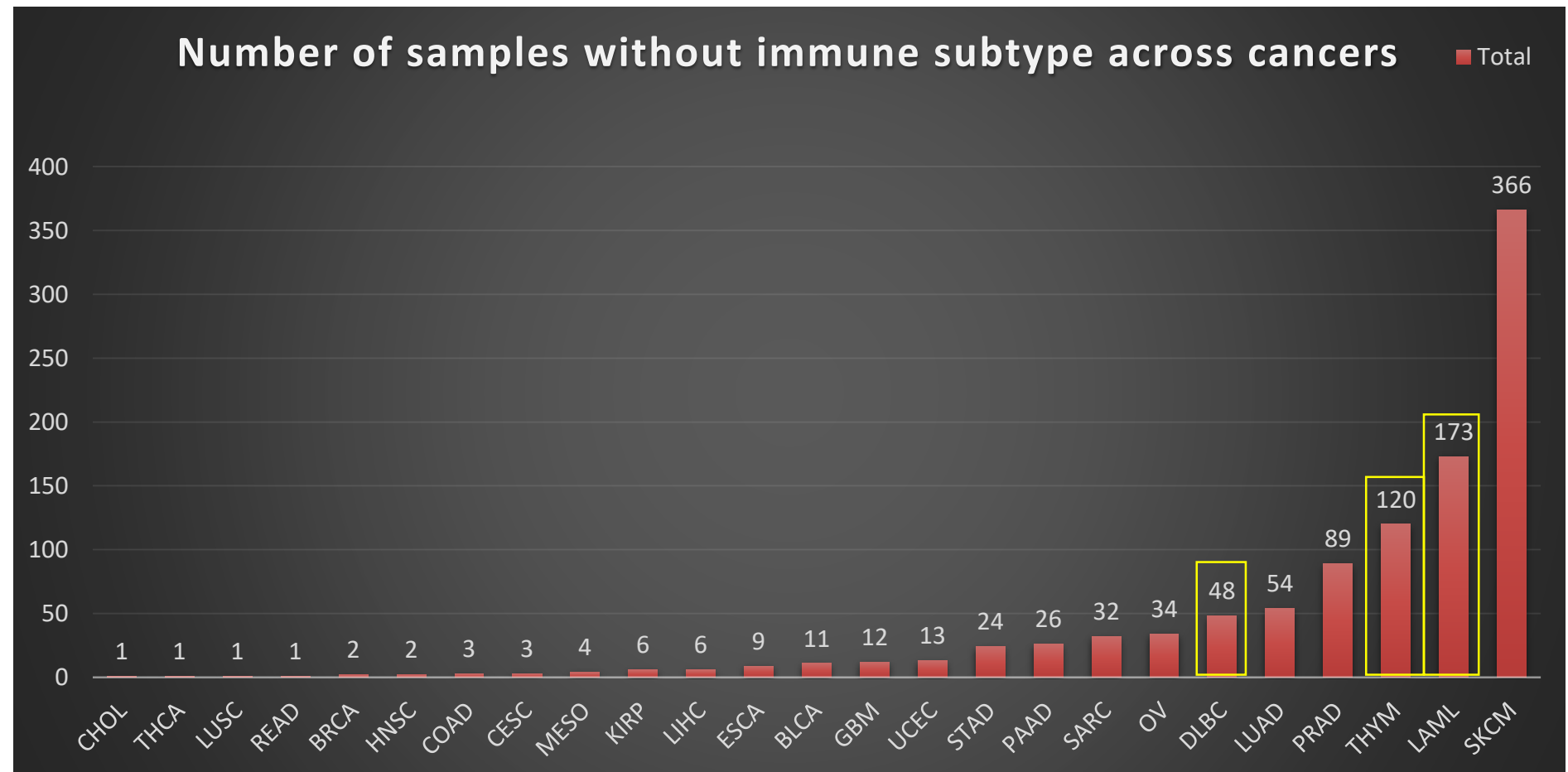
# Distribution of samples without immune subtype (n=1041)

No any immune subtype for these cancers:

**DLBC:** Diffuse Large B-cell Lymphoma

**LAML:** Acute Myeloid Leukemia

**THYM:** Thymoma




# Assessing missing data across genes [9126×440]

All the analyses in the paper are based on non-missing values (genes/samples with missing data were excluded )

Gene	# of missing	Signature group	Note
RGS8	1396	Wound healing	Total:  2536 out of [9126×440] (0.06%)
PLG	833	Wound healing	
MMP3	307	TGB-β	



## Developed GUI for the DDRD assay project

http://127.0.0.1:6535 [Open in Browser](#) 

# DNA Damage Response Deficiency (DDRD) Assay

**Gene expression platform:**

- ☒ qPCR, Normalised
- ☐ Microarray, Normalised
- ☐ NanoString, Normalised
- ☐ RNA-Seq, Normalised

**Cancer subtype:**

- ☒ Breast
- ☐ Ovarian
- ☐ Oesophageal
- ☐ Colorectal

**Gene expression CSV file upload:**

[Browse...](#) No file selected

DDRD Assay will score gene expression data

Download test data to try the Assay for scoring: [Test CSV file](#)

DDRD Score Table

Classification Plot

Informative Genes Table

Sample QC

Heatmap

Gene expression

About


Help


Unclassifiable samples are those for which a confident subgroup call could not be made

[Download table as .csv](#)

---

Almac's DDRD assay has the potential to be used as a companion diagnostic for DNA damaging therapy across a range of disease areas.

 **QUEEN'S  
UNIVERSITY  
BELFAST**

 **ALMAC**



# DDRD assay main page GUI

http://127.0.0.1:7371 | Open in Browser | Publish

## DNA Damage Response Deficiency (DDRD) Assay

Gene expression platform:

- ☒ qPCR, Normalised
- ☐ Microarray, Normalised
- ☐ NanoString, Normalised
- ☐ RNA-Seq, Normalised

Cancer subtype:

- ☒ Breast
- ☐ Ovarian
- ☐ Oesophageal
- ☐ Colorectal

Gene expression CSV file upload:

Browse... No file selected

DDRD Assay will score gene expression data



Download test data to try the Assay for scoring: [Test CSV file](#)

DDRD Score Table | Classification Plot | Sample QC | Heatmap | About | Help

Unclassifiable samples are those for which a confident subgroup call could not be made

Download table as .csv

Almac's DDRD assay has the potential to be used as a companion diagnostic for DNA damaging therapy across a range of disease areas.

Partnering to Advance Human Health

© 2018 Stratified Medicine Group, CCRCB, Queen's University Belfast

Timescale and deliverables		Year1				Year2		
		Q1*	Q2	Q3	Q4	Q1	Q2	Contributors
<b>WP1: Classification model development</b>								
<b>WP1.1: Supervised learning model development &amp; deployment</b>								
Immunogenomics classification model development	Resource preparation: normalised gene expression*							Andrena
	Resource preparation: 5 signatures & 6 immune subgroups							Reza
	Training and test datasets preparation							Reza & Andrena
	Supervised learning model selection							Reza
	Model test and validation (methods)							Reza
	Model validation by independent cohorts							Reza
Web-based package development and deployment	Application/Software adaptation							Reza
	GUI web-based design, development & test (localhost)							Reza
	Deployment, system/user-accepting test & bugs check (VM)							Reza
	Documentation: software manuals (soft version) & handover							Reza
	Post-development support and ongoing maintenance							Reza & Andrena
<b>WP1.2: Reproducible immune signature generation, DDRD signature generation and classification</b>								
Immune signature generation	Application code development 1 (ssGSEA & WGCNA)							Andrena & Reza
	Application code development 2 (ssGSEA & WGCNA)							Andrena & Reza
	Integration, test and adaptation							Andrena & Reza
DDRD signature generation and classification	Application code development3 (DDRD assay classification)							Reza, Andrena & Amanda
	GUI web-based design & integration with existing apps							Reza & Amanda
	Deployment, system/user-accepting test & bugs check (VM)							Reza, Andrena & Amanda
	Documentation: software manuals (soft version) & handover							Reza, Andrena & Amanda
	Conference/Journal preparation & presentation							
<b>WP2: Cross-cancer clinical subgroup application</b>								
Applying the classification software to clinically annotated gene expression data (provided by Almac Diagnostics)	Assessing the association between each immune molecular subtype and patient outcomes across prostate, breast, ovarian, lung, melanoma and colon cancer Predication of drug responses (?)							