Identifying the mechanisms driving pancreatic ductal adenocarcinoma stem cell characteristics using single-cell RNA-sequencing

Andrei-Florian Stoica

St. Hilda's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences
Hilary Term 2023

Declaration

I confirm that the work presented in this thesis was performed by me between October 2019 and April 2023 as a DPhil student in the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford. All contributions made by additional parties are acknowledged within the main text of the thesis. This thesis has been written by me. I have not previously submitted this thesis for the award of a degree.

Andrei-Florian Stoica
St. Hilda's College
Hilary Term 2023

Acknowledgements

I would like to thank my supervisors, Dr Siim Pauklin, Prof Udo Oppermann and Prof Adam Cribbs for their guidance and support. I also thank my funding sources, the Engineering and Physical Sciences Research Council (EPSRC) and the Structural Genomics Consortium (SGC). In addition, I would like to express my thanks to Dr Martin Pook for mentoring me as a rotation student in Dr Siim Pauklin's lab, and to Prof Lei Jiang, Wanzi Hong and Yaoxin Liu for generating the single-cell RNA-sequencing patient data that I analysed in this thesis. For their contribution to generating the single-cell RNA-sequencing A13A data, I thank Dr Siim Pauklin, Dr Stefania Militi and Dr Martin Philpott. Furthermore, I am grateful for having benefited from access to two high-performance computing servers provided by Prof Adam Cribbs (Nanopore), and Prof Liye Cheng and Feng Liu. I am also grateful to all the other past and present members of the Pauklin group not mentioned already: Dr. Chao-Hui Chang, Siwei Deng, Egle-Helene Ervin, Dr. Yuliang Feng, Dr. Rhiannon French, Dr. Mai Abdel Mouti, Hannah Pook, Linh Huyen Truong and Shihong Wu, and to other people in Botnar who have helped me with advice and discussions: Prof Afsie Sabokbar, Prof Claire Edwards, Dr. Reshma Nibhani and Dr. James Dunford. Finally, I would like to express my gratitude to my family and friends for their continuous support over the last few years, especially to my amazing wife, Xue Stoica.

List of abbreviations

Abbreviation	Description
2-HG	2-hydroxyglutarate
ABCB1	ATP binding cassette subfamily B member 1
ABCC2	ATP binding cassette subfamily C member 2
ABCG2	ATP binding cassette subfamily G member 2 (Junior blood group)
AC099850.3	PRR11 antisense RNA 1
ACC	Acinar cell carcinoma
ADAMTS6	ADAM metallopeptidase with thrombospondin type 1 motif 6
ADGRE5	Adhesion G protein-coupled receptor E5
ADGRF1	Adhesion G protein-coupled receptor F1
ADM	Acinar-to-ductal metaplasia
ADM	Adrenomedullin
AFL	Atypical flat lesions
AGR2	Anterior gradient 2, protein disulphide isomerase family member
AKAP12	A-kinase anchoring protein 12
AKR1B10	Aldo-keto reductase family 1 member B10
AKR1C1	Aldo-keto reductase family 1 member C1
AKR1C2	Aldo-keto reductase family 1 member C2
AKR1C3	Aldo-keto reductase family 1 member C3
Akt	AKT serine/threonine kinase
ALDH	Aldehyde dehydrogenase
ALDH1	Aldehyde dehydrogenase 1 family member A1
ALDH1A1	Aldehyde dehydrogenase 1 family member A1
AMBP	Alpha-1-microglobulin/bikunin precursor
ANKRD1	Ankyrin repeat domain 1
ANKRD36C	Ankyrin repeat domain 36C
ANKRD37	Ankyrin repeat domain 37
ANLN	Anillin, actin binding protein
ANP32E	Acidic nuclear phosphoprotein 32 family member E
ANXA1	Annexin A1
APC	APC regulator of WNT signaling pathway
ARHGAP11A	Rho GTPase activating protein 11A
ARL4A	ADP ribosylation factor like GTPase 4A
ASCO	The American Society of Clinical Oncology
ASF1B	Anti-silencing function 1B histone chaperone
ASPH	Aspartate beta-hydroxylase
ASPM	Assembly factor for spindle microtubules
ATF3	Activating transcription factor 3
ATM	ATM serine/threonine kinase
AURKA	Aurora kinase A
AURKB	Aurora kinase B
BAG3	BAG cochaperone 3
BBC3	BCL2 binding component 3
Bcl-2	BCL2 apoptosis regulator
BIRC5	Baculoviral IAP repeat containing 5
BORA	BORA aurora kinase A activator
BRCA1	BRCA1 DNA repair associated

BRCA2	BRCA2 DNA repair associated
BRD	Bromodomain
BRD9	Bromodomain containing 9
BSA	Bovine serum albumin
BTG1	BTG anti-proliferation factor 1
BTG2	BTG anti-proliferation factor 2
BUB1	BUB1 mitotic checkpoint serine/threonine kinase
BUB1B	BUB1 mitotic checkpoint serine/threonine kinase B
C18orf54	Chromosome 18 open reading frame 54
C1orf112	FIGNL1 interacting regulator of recombination and mitosis
CA 19-9	Carbohydrate antigen 19-9
CAF	Cancer-associated fibroblast
CALM1	Calmodulin 1
CBR3	Carbonyl reductase 3
CBX3	Chromobox 3
CCDC18	coiled-coil domain containing 18
CCDC77	Coiled-coil domain containing 77
CCNA2	Cyclin A2
CCNB1	Cyclin B1
CCNB2	Cyclin B2
CCRSA	Cell-cycle-related stemness-associated
CD133	Prominin 1
CD24	CD24 molecule
CD44	CD44 molecule (Indian blood group)
CD63	CD63 molecule
CD68	CD68 molecule
CD9	CD9 molecule
CD90	CD90 molecule
CD93	CD93 molecule
CDC20	Cell division cycle 20
CDC25A	Cell division cycle 25A
CDC45	Cell division cycle 45
CDCA2	Cell division cycle associated 2
CDCA3	Cell division cycle associated 3
CDCA5	Cell division cycle associated 5
CDCA8	Cell division cycle associated 8
CDH1	Cadherin 1
CDH2	Cadherin 2
CDK1	Cyclin dependent kinase 1
CDKN1A	Cyclin dependent kinase inhibitor 1A
CDKN2A	Cyclin dependent kinase inhibitor 2A
CDKN3	Cyclin dependent kinase inhibitor 3
CEACAM5	CEA cell adhesion molecule 5
CEACAM6	CEA cell adhesion molecule 6
CENPA	Centromere protein A
CENPC	Centromere protein C
CENPF	Centromere protein F
CENPH	Centromere protein H
CENPI	Centromere protein I
CENPK	Centromere protein K
CENPN	Centromere protein N

CENPU	Centromere protein U
CENPW	Centromere protein W
CEP128	Centrosomal protein 128
CEP55	Centrosomal protein 55
CEP78	Centrosomal protein 78
CGA	Glycoprotein hormones, alpha polypeptide
CHEK1	Checkpoint kinase 1
CHK1	Checkpoint kinase 1
CHORDC1	Cysteine and histidine rich domain containing 1
CIT	Citron rho-interacting serine/threonine kinase
CKAP2	Cytoskeleton associated protein 2
CKAP2L	Cytoskeleton associated protein 2 like
CKS1B	CDC28 protein kinase regulatory subunit 1B
CKS2	CDC28 protein kinase regulatory subunit 2
CLDN	Claudin
CLSPN	Claspin
CLU	Clusterin
c-Met	MET proto-oncogene, receptor tyrosine kinase
c-FLIP	FLICE-like inhibitory protein
CNTRL	Centriolin
COL1A1	Collagen type I alpha 1 chain
COL1A2	Collagen type I alpha 2 chain
COL3A1	Collagen type III alpha 1 chain
COL5A1	Collagen type V alpha 1 chain
COL5A2	Collagen type V alpha 2 chain
CPEB2	Cytoplasmic polyadenylation element binding protein 2
CRAN	The Comprehensive R Archive Network
CSC	Cancer stem cell
CT	Computerized tomography
CXCR4	C-X-C motif chemokine receptor 4
CYP4F3	Cytochrome P450 family 4 subfamily F member 3
DBF4	DBF4 zinc finger
DCLK1	Doublecortin like kinase 1
DCN	Decorin
DDIAS	DNA damage induced apoptosis suppressor
DEPDC1B	DEP domain containing 1B
DIAPH3	Diaphanous related formin 3
DKC1	Dyskerin pseudouridine synthase 1
DLGAP5	DLG associated protein 5
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DNA2	DNA replication helicase/nuclease 2
DNAJB1	DnaJ heat shock protein family (Hsp40) member B1
DNAJB4	DnaJ heat shock protein family (Hsp40) member B4
DNAJB6	DnaJ heat shock protein family (Hsp40) member B6
DSG2	Desmoglein 2
DSP	Desmoplakin
DST	Dystonin
DTL	Denticleless E3 ubiquitin protein ligase homolog
E2F1	E2F transcription factor 1
E2F2	E2F transcription factor 2

EBLN2	Endogenous Bornavirus like nucleoprotein 2
EBNA1BP2	EBNA1 binding protein 2
ECM	Extracellular matrix
ECT2	Epithelial cell transforming 2
EDA	Ectodysplasin A
EDTA	Ethylenediaminetetraacetic acid
EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
ELOVL6	ELOVL fatty acid elongase 6
EMT	Epithelial-mesenchymal transition
EP300	E1A binding protein p300
EPCAM	Epithelial cell adhesion molecule
ER	Endoplasmic reticulum
ERK	Mitogen-activated protein kinase 1
ESA	Epithelial cell adhesion molecule
ESCO2	Establishment of sister chromatid cohesion N-acetyltransferase 2
ESMO	European Society for Medical Oncology
EUS	Endoscopic ultrasound
EXO1	Exonuclease 1
EXT1	Exostosin glycosyltransferase 1
EZH2	Enhancer of zeste 2 polycomb repressive complex 2 subunit
EZR	Ezrin
F12	Ham's F-12 Nutrient Mix
F2RL2	Coagulation factor II thrombin receptor like 2
FAAP24	FA core complex associated protein 24
FAM72B	Family with sequence similarity 72 member B
FAM83A	Family with sequence similarity 83 member A
FAM83D	Family with sequence similarity 83 member D
FAMMM	Familial atypical multiple mole and melanoma
FANCI	FA complementation group I
FBXO43	F-box protein 43
FBXO5	F-box protein 5
FDA	Food and Drug Administration
FGF	Fibroblast growth factor
FGF2	Fibroblast growth factor 2
FLJ10213	Endogenous Bornavirus like nucleoprotein 2
FN1	Fibronectin 1
FOXM1	Forkhead box M1
FRMD4A	FERM domain containing 4A
FST	Follistatin
FTH1	Ferritin heavy chain 1
FTL	Ferritin light chain
G2E3	G2/M-phase specific E3 ubiquitin protein ligase
GADD45A	Growth arrest and DNA damage inducible alpha
GADD45B	Growth arrest and DNA damage inducible beta
GALK1	Galactokinase 1
GCLM	Glutamate-cysteine ligase modifier subunit
GDA	Guanine deaminase
GDF15	Growth differentiation factor 15
GEO	Gene Expression Omnibus
GEPIA	Gene Expression Profiling Integrative Analysis

GINS1	GINS complex subunit 1
GINS2	GINS complex subunit 2
GLIPR1	GLI pathogenesis related 1
GLRX3	Glutaredoxin 3
GNAS	GNAS complex locus
GnP	Gemcitabine and nab-paclitaxel
GO	Gene Ontology
GPAT3	Glycerol-3-phosphate acyltransferase 3
GPRC5A	G protein-coupled receptor class C group 5 member A
GPSM2	G protein signaling modulator 2
GPX3	Glutathione peroxidase 3
GSTM3	Glutathione S-transferase mu 3
H1-2	H1.2 linker histone, cluster member
H1-4	H1.4 linker histone, cluster member
H2A.Z	H2A.Z variant histone 1
H2AC21	H2A clustered histone 21
H2AZ1	H2A.Z variant histone 1
H4C3	H4 clustered histone 3
HAS2	Hyaluronan synthase 2
HASPIN	Histone H3 associated protein kinase
HDAC1	Histone deacetylase 1
HDAC2	Histone deacetylase 2
HDAC2-AS2	HDAC2 and HS3ST5 antisense RNA 2
HDAC4	Histone deacetylase 4
HDAC6	Histone deacetylase 6
HDAC7	Histone deacetylase 7
HDAC8	Histone deacetylase 8
HERPUD1	Homocysteine inducible ER protein with ubiquitin like domain 1
HGD	Homogentisate 1,2-dioxygenase
HIF-1α	Hypoxia inducible factor 1 subunit alpha
HIP1R	Huntingtin interacting protein 1 related
HIPK2	Homeodomain interacting protein kinase 2
HIST1H1B	H1.5 linker histone, cluster member
HIST1H1E	H1.4 linker histone, cluster member
HIST1H2BF	H2B clustered histone 7
HIST1H3B	H3 clustered histone 2
HIST2H2AB	H2A clustered histone 21
HJURP	Holliday junction recognition protein
HMGA2	High mobility group AT-hook 2
HMGB1	High mobility group box 2
HMGB2	High mobility group box 2
HMGB3	High mobility group box 2
HMMR	Hyaluronan mediated motility receptor
HNF1A	HNF1 homeobox A
HNRNPA2B1	Heterogeneous nuclear ribonucleoprotein A2/B1
HSP90AA1	heat shock protein 90 alpha family class A member 1
HSP90AB1	Heat shock protein 90 alpha family class B member 1
HSPA1A	Heat shock protein family A (Hsp70) member 1A
HSPA1B	Heat shock protein family A (Hsp70) member 1B
HSPA8	Heat shock protein family A (Hsp70) member 8
HSPB8	Heat shock protein family B (small) member 8

HSPD1	Heat shock protein family D (Hsp60) member 1
HSPH1	Heat shock protein family H (Hsp110) member 1
IAP	Inhibitor of apoptosis
IDH1	Isocitrate dehydrogenase (NADP(+)) 1
IDH2	Isocitrate dehydrogenase (NADP(+)) 2
IF	Interferon
IFI44L	Interferon induced protein 44 like
IFI6	Interferon alpha inducible protein 6
IL33	Interleukin 33
IL6	Interleukin 6
INHBA	Inhibin subunit beta A
IOPN	Intraductal oncocytic papillary neoplasms
IPMN	Intraductal papillary mucinous neoplasms
ISLR	Immunoglobulin superfamily containing leucine rich repeat
ITGA11	Integrin subunit alpha 11
ITGB3BP	Integrin subunit beta 3 binding protein
ITPN	intraductal tubulopapillary neoplasms
IVT	In vitro transcription
JAK	Janus kinase 1
JAM	Junctional adhesion molecule
JNK	Mitogen-activated protein kinase 8
JUN	Jun proto-oncogene, AP-1 transcription factor subunit
JUNB	JunB proto-oncogene, AP-1 transcription factor subunit
KCNMA1	Potassium calcium-activated channel subfamily M alpha 1
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIAA0101	PCNA clamp associated factor
KIAA0319	KIAA0319
KIF11	Kinesin family member 11
KIF14	Kinesin family member 14
KIF15	Kinesin family member 15
KIF18A	Kinesin family member 18A
KIF18B	Kinesin family member 18B
KIF20A	Kinesin family member 20A
KIF23	Kinesin family member 23
KIF2C	Kinesin family member 2C
KIF4A	Kinesin family member 4A
KIFC1	Kinesin family member C1
KNL1	Kinetochore scaffold 1
KNSTRN	Kinetochore localized astrin (SPAG5) binding protein
KNTC2	NDC80 kinetochore complex component
KPNA2	Karyopherin subunit alpha 2
KRAS	KRAS proto-oncogene, GTPase
KRT18	Keratin 18
KRT19	Keratin 19
KRT8	Keratin 8
L3.6pl	L3.6 pancreas-liver
LAP3	C-X-C motif chemokine receptor 4
LGR5	Leucine rich repeat containing G protein-coupled receptor 5
LIMCH1	LIM and calponin homology domains 1
LINC00536	Long intergenic non-protein coding RNA 536
LIPH	Lipase H

LMNB1	Lamin B1
LOX	Lysyl oxidase
LRRFIP1	LRR binding FLII interacting protein 1
LUM	Lumican
MAD2L1	Mitotic arrest deficient 2 like 1
MAN1A1	Mannosidase alpha class 1A member 1
MAP1B	Microtubule associated protein 1B
MAPK	Mitogen-activated protein kinase
MCM10	Minichromosome maintenance 10 replication initiation factor
MCN	Mucinous cystic neoplasms
MDM2	MDM2 proto-oncogene
MDR1	ATP binding cassette subfamily B member 1
ME1	Malic enzyme 1
MELK	Maternal embryonic leucine zipper kinase
MET	MET proto-oncogene, receptor tyrosine kinase
MFAP5	Microfibril associated protein 5
MIF	Macrophage migration inhibitory factor
MIR4713HG	MIR4713 host gene
MIRN21	MicroRNA 21
MIRN221	MicroRNA 221
MIS18A	MIS18 kinetochore protein A
MK	Midkine
MKI67	Marker of proliferation Ki-67
MLH1	MutL homolog 1
MND1	Meiotic nuclear divisions 1
MRI	Magnetic resonance imaging
mRNAsi	mRNA expression-based stemness index.
MRP1	ATP binding cassette subfamily C member 1
MSH2	MutS homolog 2
MSH6	MutS homolog 6
MSMB	Microseminoprotein beta
MTFR2	Mitochondrial fission regulator 2
MUC5AC	Mucin 5AC, oligomeric mucus/gel-forming
MXD1	MAX dimerization protein 1
MYBL2	MYB proto-oncogene like 2
MYC	MYC proto-oncogene, bHLH transcription factor
Nanog	Nanog homeobox
NCAPG	Non-SMC condensin I complex subunit G
NCAPH	Non-SMC condensin I complex subunit H
NCCN	National Comprehensive Cancer Network
NCL	Nucleolin
NCSLC	non-small-cell lung carcinoma
NDC1	NDC1 transmembrane nucleoporin
NDC80	NDC80 kinetochore complex component
NDE1	NudE neurodevelopment protein 1
NDRG1	N-myc downstream regulated 1
NEGR	Neuronal growth regulator 1
NEIL3	Nei like DNA glycosylase 3
NEK2	NIMA related kinase 2
NES	Nestin
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells

NID2	Nidogen 2
Notch	Notch receptor
NQO1	NAD(P)H quinone dehydrogenase 1
NR0B1	Nuclear receptor subfamily 0 group B member 1
NRF2	Nuclear factor erythroid 2–related factor 2
NTS	Neurotensin
NUF2	NUF2 component of NDC80 kinetochore complex
NUP107	Nucleoporin 107
NUSAP1	Nucleolar and spindle associated protein 1
Oct4	POU class 5 homeobox 1
OCT4	POU class 5 homeobox 1
ODF2	Outer dense fiber of sperm tails 2
OIP5	Opa interacting protein 5
ORC1	Origin recognition complex subunit 1
ORC6	Origin recognition complex subunit 6
OS	Overall survival
OSGIN1	Oxidative stress induced growth inhibitor 1
OSR1	Odd-skipped related transcription factor 1
OTUD1	OTU deubiquitinase 1
p53	Tumor protein p53
PALB2	partner and localizer of BRCA2
PALM2-AKAP2	PALM2 and AKAP2 fusion
PALS2	Protein associated with LIN7 2, MAGUK p55 family member
PanIN	Pancreatic intraepithelial neoplasia
PanIN1	Pancreatic intraepithelial neoplasia 1
PanIN2	Pancreatic intraepithelial neoplasia 2
PanIN3	Pancreatic intraepithelial neoplasia 3
PAPPA	PAPPA
PARP1	Poly(ADP-ribose) polymerase 1
PARPBP	PARP1 binding protein
PBK	PDZ binding kinase
PC	Pancreatic cancer
PCA	Principal component analysis
PCLAF	PCNA clamp associated factor
PCR	Polymerase chain reaction
PCSC	Pancreatic cancer stem cell
PDAC	Pancreatic ductal adenocarcinoma
PDGF	Platelet derived growth factor
PDGFD	Platelet derived growth factor D
PDIA4	Protein disulfide isomerase family A member 4
PECAM1	Platelet and endothelial cell adhesion molecule 1
PFS	Progression-free survival
PGD	Phosphogluconate dehydrogenase
PHLDA2	Pleckstrin homology like domain family A member 2
PI3K	Phosphoinositide 3-kinase
PIF1	PIF1 5'-to-3' DNA helicase
PIMREG	PICALM interacting mitotic regulator
PKIB	CAMP-dependent protein kinase inhibitor beta
PLCB1	Phospholipase C beta 1
PLEKHG3	Pleckstrin homology and RhoGEF domain containing G3
PLK1	Polo like kinase 1

PLK4	Polo like kinase 4
PMAIP1	Phorbol-12-myristate-13-acetate-induced protein 1
PMS2	PMS1 homolog 2, mismatch repair system component
PNPT1	Polyribonucleotide nucleotidyltransferase 1
POSTN	Periostin
POU2F2	POU class 2 homeobox 2
POU5F1	POU class 5 homeobox 1
PP	Pancreatic polypeptide
PPIN	Protein-protein interaction network
PPP1R15A	Protein phosphatase 1 regulatory subunit 15A
PRC1	Protein regulator of cytokinesis 1
PRC2	Polycomb repressive complex 2
PRDX1	Peroxiredoxin 1
PRDX6	Peroxiredoxin 6
PROM1	Prominin 1
PROS	Protein S
PRR11	Proline rich 11
PRSS1	Serine protease 1
PSC	Pancreatic stellate cells
PSRC1	Proline and serine rich coiled-coil 1
PTEN	Phosphatase and tensin homolog
PTF1A	Pancreas associated transcription factor 1a
PTMA	Prothymosin alpha
PTN	Pleiotrophin
PTPRM	Protein tyrosine phosphatase receptor type M
PTTG1	PTTG1 regulator of sister chromatid separation, securin
RACGAP1	Rac GTPase activating protein 1
RAD51	RAD51 recombinase
RAD54L	RAD54 like
RAS	Rat sarcoma
RASAL2	RAS protein activator like 2
RBC	Red blood cells
RCC1	Regulator of chromosome condensation 1
RCCD1	RCC1 domain containing 1
REG4	Regenerating family member 4
REV3	REV3 like, DNA directed polymerase zeta catalytic subunit
RFC4	replication factor C subunit 4
RGS4	Regulator of G protein signaling 4
ROS	Reactive oxygen species
RPS6	Ribosomal protein S6
RRM2	Ribonucleotide reductase regulatory subunit M2
RSAD2	Radical S-adenosyl methionine domain containing 2
RSPO3	R-spondin 3
RUNX1	RUNX family transcription factor 1
S100A11	S100 calcium binding protein A11
S100A6	S100 calcium binding protein A6
S100P	S100 calcium binding protein P
SAPCD2	suppressor APC domain containing 2
scRNA-seq	Single-cell RNA-sequencing
SCT	SCTransform
SDCBP2	Syndecan binding protein 2

SEMA3A	Semaphorin 3A
SEMA6	Semaphorin 6
SGO1	Shugoshin 1
SGO2	Shugoshin 2
SGOL2	Shugoshin 2
SH3RF1	SH3 domain containing ring finger 1
SKA1	Spindle and kinetochore associated complex subunit 1
SKA3	Spindle and kinetochore associated complex subunit 3
SLC40A1	Solute carrier family 40 member 1
SMAD4	SMAD family member 4
SMC4	Structural maintenance of chromosomes 4
SMTN	Smoothelin
Snai1	Snail family transcriptional repressor 1
Snai2	Snail family transcriptional repressor 2
SNORA1	Small nucleolar RNA, H/ACA box 1
SNORA16B	Small nucleolar RNA, H/ACA box 16B
SNORA25	Small nucleolar RNA, H/ACA box 25
SNORD122	Small nucleolar RNA, C/D box 122
SNORD5	Small nucleolar RNA, C/D Box 5
SNORD53	Small nucleolar RNA, C/D box 53
SOX2	SRY-box transcription factor 2
SP	Side population
SPAG5	Sperm associated antigen 5
SPC25	SPC25 component of NDC80 kinetochore complex
SPLCL	Side population genes from the L3.6pl cell line
SPP1	Secreted phosphoprotein 1
SPRY2	Sprouty RTK signaling antagonist 2
SQSTM1	Sequestosome 1
Src	SRC proto-oncogene, non-receptor tyrosine kinase
SSEA-1	Stage-specific embryonic antigen-1
SSEA-4	Stage-specific embryonic antigen-4
STAT	Signal transducer and activator of transcription
STIL	STIL centriolar assembly protein
STK11	Serine/threonine kinase 11
STMN1	Stathmin 1
SUV39H2	SUV39H2 histone lysine methyltransferase
SYNE2	Spectrin repeat containing nuclear envelope protein 2
TAGLN	Transgelin
TALDO1	Transaldolase 1
TCGA	The Cancer Genome Atlas
TDPM	Traditionally-derived PCSC markers
TET	Tet methylcytosine dioxygenase
TFPI2	Tissue factor pathway inhibitor 2
TFPI-2	Tissue factor pathway inhibitor 2
TGFBR1	Transforming growth factor beta receptor 1
TGF-β	Transforming growth factor beta
THY1	THY1 molecule
TIMELESS	Timeless circadian regulator
TK1	Thymidine kinase 1
TM4SF1	Transmembrane 4 L six family member 1
TM4SF19	Transmembrane 4 L six family member 19

TMEM156	Transmembrane protein 156
TMEM92	Transmembrane protein 92
TNFAIP6	TNF alpha induced protein 6
TNFRSF10B	TNF receptor superfamily member 10b
TNFRSF11B	TNF receptor superfamily member 11b
TOP2A	DNA topoisomerase II alpha
TOPK	PDZ binding kinase
TP53	Tumor protein p53
TPX2	TPX2 microtubule nucleation factor
TSPAN1	Tetraspanin 1
TSPAN8	Tetraspanin 8
TTK	TTK protein kinase
TUBA1B	Tubulin alpha 1b
TUBB4A	Tubulin beta 4A class IVa
Twist	Twist family bHLH transcription factor
TXN	Thioredoxin
TXNRD1	Thioredoxin reductase 1
UACA	Uveal autoantigen with coiled-coil domains and ankyrin repeats
UBB	Ubiquitin B
UBC	Ubiquitin C
UBE2C	Ubiquitin conjugating enzyme E2 C
UBE2T	Ubiquitin conjugating enzyme E2 T
UGDH	UDP-glucose 6-dehydrogenase
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
USPSTF	The United States Preventive Services Task Force
VCAN	Versican
VEGFA	Vascular endothelial growth factor A
VEGFC	Vascular endothelial growth factor C
VIM	Vimentin
VSIG1	V-set and immunoglobulin domain containing 1
WGCNA	Weighted gene coexpression network analysis
WNK2	WNK lysine deficient protein kinase 2
WNT5A	Wnt family member 5A
WP	WikiPathways
XRCC2	X-ray repair cross complementing 2
XRCC4	X-ray repair cross complementing 4
Zeb1	Zinc finger E-box binding homeobox 1
Zeb2	Zinc finger E-box binding homeobox 2
ZFAND2A	Zinc finger AN1-type containing 2A
ZWILCH	Zwilch kinetochore protein
ZWINT	ZW10 interacting kinetochore protein

Table of contents

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is among the deadliest human malignancies. Surgery, the only curative treatment, is precluded by the late stage at diagnosis in 80% of cases. Recurrence after surgery is common, and the disease does not respond well to chemotherapy and radiotherapy. Resistance to therapy and recurrence are thought to be driven by pancreatic cancer stem cells (PCSCs), a subset of cells with self-renewal and differentiation capacities. Annihilating these cells is therefore of paramount importance for treating PDAC. Identifying these cells, however, has proven challenging. Currently, there is no gene signature able to identify PCSCs. In this thesis, I employed single-cell RNA-sequencing to integrate multiple approaches towards the identification of PCSCs (experimentally-derived markers, bioinformatics-based gene sets, and computational tools to infer developmental potential from expression data), to uncover the genes and processes characterizing PCSCs, and to assess the effects of I-BRD9, an inhibitor of BRD9, a bromodomain-containing protein involved in chromatin remodelling upon the PCSCs, using two single-cell RNA-sequencing PDAC datasets. The results evidenced cell cycle abnormalities as crucial to cancer stemness in PDAC, with multiple lines of evidence converging towards the identification of clusters whose markers significantly overlapped with cell cycle-related stemness-associated gene sets as PCSCs. Traditionally-derived PCSC markers were found to be largely of low reliability. A transitional cell population distinct from both PCSCs and the bulk of the cells, and one with an advanced stage of differentiation which however regained partial stemness-like characteristics, were identified as highly drug-resistant, suggesting that greater than previously believed PDAC cell heterogeneity, not merely PCSCs, is involved in chemoresistance. I-BRD9 achieves a ~6-fold reduction of PCSCs in one dataset, likely mediated by the demonstrated downregulation of key G2/M DNA replication checkpoint-linked genes such as *TOP2A* and *CDK1*, but the effects are partially reversed by the addition of PDAC drug gemcitabine.

List of figures

List of tables

CHAPTER 1

INTRODUCTION

1 Introduction

1.1 Pancreatic cancer

1.1.1 The origin and development of PDAC

1.1.1.1 The pancreas

The human pancreas is composed of an exocrine portion, involved in the secretion of digestive enzymes and accounting for 85% of the organ mass, and an endocrine portion involved in the secretion of hormones[1]. The functional unit of the exocrine pancreas consists of two elements: an acinus and its associated ductule[1]. Thus, the exocrine component of the pancreas presents two main classes of cells: acinar cells, which form the vast majority of all pancreatic cells and synthetize and secrete digestive enzymes[2], and the associated ductal cells which form the epithelial lining of the tubes that transport the digestive enzymes into the duodenum[3]. Ductal cells make up about 10% of all pancreatic cells and about 4% of the pancreatic volume[3]. Dispersed throughout the exocrine pancreas and amounting only to 1-2% of the pancreatic volume[2], the endocrine islets of Langerhans are divided in several classes, based on the hormones they secrete: glucagons for α cells, insulin for β cells, somatostatin for δ cells and pancreatic polypeptide for PP cells[2].

Notably, acinar cells show a high degree of plasticity that appears to account for the regenerative capabilities of the pancreas, even in the absence of a clearly delineated stem cell compartment in the pancreas[4]. Through acinar-to-ductal metaplasia (ADM), an event that can be triggered by injury, inflammation and stress conditions, acinar cells acquire a more epithelial, ductal-like phenotype, as well as progenitor cell-like characteristics[4].

Another important class of pancreatic cells is composed of the pancreatic stellate cells (PSC), slender star-shaped cells that wrap themselves around acinar and ductal structures, and around

the islets of Langerhans[1]. These cells assist the proper formation of epithelial structures, while their deregulation bears a role in several pathological states of the pancreas, including pancreatic cancer (PC)[1].

1.1.1.2 PDAC: description and delineation

Exocrine pancreatic cancers account for 95% of all pancreatic cancers. The remaining 5% are neuroendocrine pancreatic cancers, which overall show a better prognosis[5].

Pancreatic ductal adenocarcinomas (PDACs) account for more than 90% of all pancreatic malignancies[4]. Because PDAC is by a large margin the most common type of PC, pancreatic cancer is often used as a synonym for PDAC[6]. PDACs of the classical type represent 85% of all exocrine pancreatic cancers[5], while several rare morphological types of PDAC also exist.

Morphological variants of PDAC, of which some share the molecular pathogenesis of the classical PDAC type (adenosquamous carcinomas, undifferentiated carcinoma, undifferentiated carcinomas with osteoclastic giant cells, signet-ring cell carcinoma etc.) but others do not (colloid carcinomas, medullary carcinomas etc.)[6] make up about 10% of exocrine pancreatic cancers[5]. In adenosquamous carcinomas of the pancreas, defined as those pancreatic carcinomas where the squamous component forms 30% or more of the total tumour mass, and in the undifferentiated carcinomas of the pancreas, prognosis might be even grimmer than in classical PDAC[6]. These tumours can both be identified by the presence of large polymorphous tumour cells, such as multinuclear tumour giant cells[6]. Undifferentiated carcinomas of the pancreas are to be distinguished, however, from pancreatic undifferentiated carcinomas with osteoclastic giant cells. The latter group includes carcinomas that present histiocytic giant cells, which are distinguished by the expression of the CD68 marker, and remarkably show a better 5-year survival, in the region of 60%[6]. Another subgroup, colloid carcinomas, shows likewise a significantly higher 5-years survival than the classical PDAC type, situated around 50%[6]. Extremely rare subtypes of

ductal carcinomas of the pancreas include poorly cohesive carcinomas, signet-ring cell carcinomas and medullary carcinomas, among others[5].

Non-ductal exocrine pancreatic cancers also occur. Pseudopapillary tumours account for 3% of all exocrine pancreatic cancers and are more frequent in young people, with a median age at diagnosis of 28 in women and of 35 in men[5]. They are, however, indolent cancers which rarely metastasize and show a 5-year survival rate of 97% even for metastatic disease[5]. Acinar cell carcinomas (ACCs) make up about 2% of the exocrine pancreatic cancers, are typically larger than PDACs, but more amenable to surgery at diagnosis (38% of newly diagnosed ACCs are resectable), which results in better overall prognosis[5]. Pancreatoblastoma is a devastating pediatric cancer with a dismal median overall survival of 5 months for patients not eligible for surgery, and it accounts for 25% of all pancreatic tumours in children, but it is extremely rare in adults[5].

Importantly, despite the ductal phenotype of PDAC, the tumour does not always arise from originally ductal cells as it had been postulated earlier[7], but it can also arise from acinar cells via reprogramming[8] through the aforementioned ADM events that also favour the acquisition of traits resembling those of progenitor cells in the cells undergoing these events[4]. In turn, these traits increase the susceptibility of the affected cells to pro-oncogenic mutations such as activating mutations in *KRAS*[4], the critical driver of PDAC, mutated in almost 100% of PDAC cases[9], a figure which renders PDAC the most *RAS*-addicted of all cancers[9].

1.1.1.3 Incidence and mortality

Pancreatic ductal adenocarcinoma is the most lethal of all common cancers[10]. It is regarded as one of greatest outstanding challenges in oncology[11]. The survival rates of surgically resected PDAC patients have notably risen in recent decades, from 1.5% in 1975 to 17.4% in 2011[12]. However, for patients with unresectable disease the improvements in expected outcomes have been very modest[12]. Median survival after diagnosis is in between 4 and 6 months, and survival

after 5 years may not exceed 5%[12]. PDAC is currently the fourth leading cause of all cancer-related deaths[13], and, due to a combination of increasing incidence, late diagnoses and restricted treatment options[11], is expected to climb to the second position by 2030[14], overtaking breast, prostate and colorectal cancer in the process[15]. The rapid increase in the incidence of PDAC has been noted especially in Europe, North America, Australia and New Zealand, and it has been associated with improvements in socioeconomic conditions, resulting in extended life expectancy and increased prevalence of risk factors such as obesity[11] and metabolic syndrome[16].

The mean age at diagnosis of PDAC is 68 and the male-to-female ratio at diagnostic is 1.6:1[17]. Under the age of 40, a diagnosis of PDAC is rare[10].

1.1.1.4 Risk factors

Lifestyle aspects, such as tobacco smoking, heavy alcohol consumption and low physical activity are regarded as risk factors for developing PDAC[18]. The emergence of the disease is also favored by obesity, chronic pancreatitis, type 2 diabetes[18], metabolic syndrome[19] and non-O ABO blood type[18]. Chronic pancreatitis increases the risk of developing PDAC 13.3-fold, tobacco smoking increases it 1.7-fold to 2.6-fold, diabetes increases it 1.5-fold to 2-fold and obesity increases it 1.1-fold to 1.5-fold[20]. Consuming more than three standard alcoholic drinks per days produce a 1.22-fold to 1.36 increased risk of developing PDAC, and a dose-response relation has been evidenced[21]. High consumption of red meat and fatty diet may likewise increase the risk of developing PDAC[21].

Around 10% of PDAC cases are thought to be familial[22], relating to a number of syndromes such as: familial atypical multiple mole and melanoma (FAMMM) syndrome, hereditary breast and ovarian cancer, Lynch syndrome, Peutz-Jeghers syndrome[22] and hereditary pancreatitis[20]. The Peutz-Jeghers syndrome, characterized by mutations in the *STK11* gene, is thought to increase the risk of developing PDAC by more than two orders of magnitude (132-fold)[20]. A major risk

increase has been observed also for hereditary pancreatitis, associated with mutations in *PRSS1* (53-fold)[20]. FAMMM, through mutations in *CDKN2A*, increases the risk of developing PDAC 13-fold to 39-fold[20]. Familiar breast and ovarian cancer syndrome increases the same risk twofold when due to mutations in *BRCA1* and threefold to ninefold when due to mutations in *BRCA2*[20]. The Lynch syndrome, characterized by a set of mutations in one or several of the *MLH1*, *MSH6*, *MSH2*, *PMS2* and *EPCAM* genes brings a ninefold to elevenfold increase of the risk[20]. The broad spectrum cancer predisposition syndrome known as Li-Fraumeni[23], characterized by *TP53* mutations, triggers a sevenfold increase in the risk of developing PDAC[20]. An increase in the risk of developing PDAC has also been noted for the Fanconi anemia and breast cancer syndrome, linked to *PALB2*, familial adenomatous polyposis syndrome, linked to *APC*, and ataxia-telengiectasia syndrome, linked to *ATM*[20].

However, the listed lifestyle factors and rare familiar syndromes account together only for at most 20% of all PDAC cases[20]. The majority of PDAC cases are, thus, considered sporadic cancers[20].

1.1.1.5 Prevention and early detection

Early detection of PDAC precursors is of crucial importance in reducing PDAC-associated mortality[20]. For example, patients with a pancreatic precursor noninvasive cystic lesion (see **Section 1.1.1.6**) are typically cured after surgery, but once the lesion has gained an invasive character, 5-year mean survival drops by more than 50%[20]. However, only about 17 in every 100,000 of all cysts detected incidentally on a MRI are ductal cancers, which raised the question if the benefits of imaging surveillance do in fact exceed potential risks[20].

Ultrasound, otherwise used as the common preliminary screening tool in patients with abdominal symptoms, is not recommended for usage in the early detection of PDAC, as it generally fails to

detect abnormalities with diameters below 2 cm or to provide an informative visualization of the topography of changes[13].

Population-wide screening is not considered feasible because of the low incidence of PDAC in the general population[24]. Periodic screening is thus recommended only for particular genetic profiles, for instance in germ-line mutations in *PRSS1* that manifest themselves as hereditary pancreatitis, with recurrent episodes of pancreatitis starting in childhood and contribute to a greater than 50-fold increased risk of developing PDAC[20].

The United States Preventive Services Task Force (USPSTF) does not recommend periodic screening for PDAC in patients with sporadic chronic pancreatitis without *PRSS1* mutations[20].

Prevention of the disease centers on avoiding lifestyle choices that increase the risk of developing PDAC[24]. Interestingly, the regular use of statin medications was found to be protective against mortality from all cancers[24].

1.1.1.6 Precursor lesions and tumourigenesis

An exome analysis of PDAC from 2008 revealed an average of 63 genomic alterations per patient genome[25]. These alterations involved 12 core signalling pathways, of which KRAS signalling, G1/S phase cell cycle transition, TGF-β signalling, integrin signalling, cell invasion, homophilic cell interaction, and small guanine triphosphate (GTPase)-dependent signalling were the most frequently involved ones[25]. *KRAS*, *TP53*, *CDKN2A*, and *SMAD4* are four genes frequently mutated in PDAC, and the corresponding mutations appear to emerge sequentially[25] as PDAC develops from its most common microscopic precursor, pancreatic intraepithelial neoplasia (PanIN)[6]. While the initial step of the genesis of PDAC remains to be established, the mutation of KRAS is certainly a key event, being found even in some pancreata with no detectable lesions and thus regarded as normal, and in the early PanIN1[25], with its prevalence increasing through the evolution to PanIN2 and PanIN3[7].

PanIN are phenotypically ductal mucinous-papillary intraepithelial neoplasms, with diameters typically below 0.5 cm[6]. 82% of pancreata with PDAC display PanIN, but they can also be found in 16% of normal pancreata[6]. Unlike *KRAS* mutations, mutations in *TP53*, *CDKN2A* and *SMAD4* are typically only found in high-grade PanIN, and in frequencies much lower than in invasive PDAC[6]. Studies in mouse models have found that *KRAS* mutations alone are enough to induce PanIN[7]. However, these PanIN usually do not evolve to PDAC in the absence of other mutations, and a second genetic hit in form of the loss of one of the tumor suppressors *TP53* and *PTEN* commonly occurs before the progression to PDAC[7]. Furthermore, mutated *KRAS* was shown to induce senescence in healthy cells, which is sometimes countered by the mutation or deletion of tumour suppressor genes[7].

In addition to PanIN, it has been posited that other rare microscopic precursors of PDAC exist, in the form of atypical flat lesions (AFL)[6], small tubular lesions enclosed by reactive stroma detected in subjects with a familiar predisposition for PDAC[6].

In 5% to 10% of cases, PDAC arises instead from macroscopic precursors, also called cystic precursors[20]. These include intraductal papillary mucinous neoplasms (IPMN), mucinous cystic neoplasms (MCN), intraductal tubulopapillary neoplasms (ITPN) and intraductal oncocytic papillary neoplasms (IOPN), the latter having been only recently recognized as separate entities[6]. IPMN also show mutations in *KRAS*, but also in *GNAS* in two thirds of the cases[6]. Much like in the case of PanIN, mutations in *TP53*, *CDKN2A* and *SMAD4* are characteristic of high-grade IPMN[6].

In contrast with microscopic precursors, cystic precursors can be imagined using computerized tomography (CT) and magnetic resonance imaging (MRI) and thus can be kept under longitudinal surveillance[20]. Imaging data has suggested that between 2% to 3% of the general population is likely to harbour asymptomatic pancreatic cysts, and the number is thought to increase more than tenfold in advanced age[20].

1.1.1.7 Diagnosis

Most patients are symptomatic at diagnosis[20]. However, the initial symptoms are nonspecific, resulting in a median delay of more than 2 months between presentation and diagnosis[20].

The most common initial symptoms reported in PDAC are fatigue (86%), weight loss (85%), anorexia (83%), abdominal pain (79%), jaundice (56%) and nausea (51%)[20].

The diagnosis cannot be established solely based on symptoms[20]. Patients displaying jaundice or abdominal pain receive multiple examinations, including a complete blood count and a blood chemistry panel, as well as liver function tests that aim to detect and quantify the extent of possible cholestasis, liver metastasis and hepatitis[20]. Patients with epigastric pain can be assessed for acute pancreatitis using serum lipase tests[20]. Carbohydrate antigen 19-9 (CA 19-9), a sialylated Lewis^a blood group antigen, is the only PC biomarker endorsed by the FDA[26], with specificity and sensitivity in between 70% and 90% in symptomatic patients[20]. However, it is not reliable as a diagnostic tool in asymptomatic patients[20]. Its utility is hindered by the fact that it is often elevated in some other cancers[20], being originally isolated from a colorectal carcinoma cell line[26], but also in some benign ailments of the pancreas[20]. Moreover, about 5% to 10% of the general population does not express Lewis antigens at all[20]. However, the combination of CA 19-9 with other biomarkers such as *MUC5AC* has shown some encouraging preliminary results, and it may serve as the basis of a blood test-based diagnosis approach in the future[20].

After suspicions of a PDAC diagnosis arise, patients are subjected to computerized tomography (CT) as an initial assessment technique[20]. CT has comparable sensitivity and specificity in the detection of PDAC to MRI, but it is generally preferred because of its diminished costs and widespread availability[20], as well as its superior spatial resolution and robustness in the face of respiratory motion artifacts[20]. MRI can be used in patients with contraindications to CT and to resolve ambiguous CT results, in particular liver lesions[20]. In a few select cases, endoscopic

ultrasound (EUS) can be used to assist identifying a tumour mass that has eluded both CT and MRI detection[20].

Most commonly, PDAC presents itself as a solitary lesion[17]. It is situated in the head of the pancreas in 60-70% of cases, and in the body or tail of the pancreas in the other cases[17]. PDACs tend to firm, multinodular and sclerotic tumours, with no distinct margins[17].

1.1.1.8 Staging

It is estimated that fewer than 20% of PDAC patients present resectable disease at diagnosis[14] and that in only 15% of patients, the surgery has the potential to be curative[27]. Resectability is defined in terms of meeting the following three conditions: no tumour-artery interface can exist, any tumour-vein interface must be lower than 180°, and no metastatic disease can exist, where the capture of lymph nodes outside the surgical basin would be considered metastasis[22].

Borderline resectable PDAC forms an important clinical entity on its own[22], and it accounts for an additional 20% of all PDAC cases[28]. In borderline resectable PDAC as defined per the National Comprehensive Cancer Network (NCCN) criteria, celiac artery encasement can exceed 180° only if neither the aorta no the gastroduodenal artery are involved[20]. The superior mesenteric artery can be encased only at angles below 180°, and the tumour can come in contact with the common hepatic artery as long as no extensions into the celiac artery or hepatic artery bifurcations exist[20]. The portal vein and the superior mesenteric vein can be encased. If the angle of encasement is lower than 180°, contour irregularities and thrombosis are also tolerated, as long as the portal vein and the superior mesenteric vein are reconstructible[20].

Thus, in locally advanced cancer, defined likewise by the NCCN criteria, at least one of the following conditions are met: the tumour has encased the celiac artery or the superior mesenteric artery at an angle greater than 180°; the tumour is in contact with the celiac artery

and the aorta, or the inferior vena cava; the portal vein or the superior mesenteric vein is not reconstructible[20].

1.1.1.9 Treatment

1.1.1.9.1 Current standards of care

Metastatic disease represents the most frequent clinical presentation of PDAC[20]. FOLFIRINOX, a chemotherapy regimen combining 5-fluorouracil, leucovorin, oxaliplatin and irinotecan, is presently the standard of care for fit patients with metastatic PDAC[20]. Before FOLFIRINOX, the standard of care for metastatic PDAC was gemcitabine monotherapy[20]. FOLFIRINOX achieved improvements in terms of both median OS: 11.1 months for FOLFIRINOX versus 6.8 months for gemcitabine, and progression-free survival (PFS): 6.4 months for FOLFIRINOX versus 3.3 months for gemcitabine[20].

An alternative to FOLFIRINOX is another aggressive combination chemotherapy regimen, where gemcitabine is administered in conjunction with nab-paclitaxel, with an improved median OS and PFS of 1.8 months when compared to gemcitabine monotherapy[20]. FOLFIRINOX is considered the more challenging but potentially the more effective regimen out of the two, and is generally preferred in patients with better performance status and ability to follow the treatment schedule[20].

The recent FDA approval of PARP inhibitor olaparib in PDAC patient with germline *BRCA1* or *BRCA2* mutations marks the first time a targeted therapy drug was approved for the treatment of PDAC[20]. The approval was granted after a group of patients with metastatic PDAC that was unresponsive to platinum-based chemotherapy recorded a median PFS of 7.4 months after being given olaparib, while the median PFS was only 3.8 months in the group given placebo[20].

Significant challenges in treating PDAC apply also to the case of resectable disease. The long-term results of surgically removing PDAC are currently unsatisfactory, with a median survival of 14-17

months and a 5-year survival variously reported in between 10% and 27%[13]. The surgical goal of a R0 resection, defined as a resection with no tumour cells within 1mm of any surface, is considered of paramount importance for survival[13]. Some presentations of PDAC, such as those where either the portal vein or the superior mesenteric vein is involved, can be still resected, but R0 is rarely reached due to the aggressiveness of the tumour[13].

In general, the effectiveness of neoadjuvant (before surgery) chemotherapy in resectable PDAC is uncertain, and conflicting reports exist[13], but it is recommended in a few specific cases: when prompt surgery is impossible, when the tumours are large or express high amounts of CA 19-9, or when patients experience severe pain[13]. No particular chemotherapeutic drug or combination is regarded as a standard of care for neodjuvant therapy in resectable PDAC[13].

In borderline resectable PDAC, however, neoadjuvant therapy plays a clear role: reducing the tumour size and, ideally, the tumour stage and thus increasing the chances of obtaining a R0 resection[13]. These limited goals are often met; patients with resectable PDAC show similar percentages of R0 resection and similar overall survival and 5-year survival values as patient with borderline resectable PDAC who underwent neoadjuvant therapy[13].

Patients with borderline resectable PDAC are recommended to be placed in clinical trials if possible as per the European Society for Medical Oncology (ESMO) guidelines[13]. Alternatively, they can be given chemotherapy based on either gemcitabine or FOLFIRINOX, followed by chemoradiotherapy and surgery[13].

Notably, neoadjuvant chemotherapy has also been used in locally advanced PDAC, with some positive results[13]. In a recent meta-analysis of 12 studies the FOLFIRINOX regime was shown capable to reduce a locally advanced PDAC to a resectable one in 28% of patients[13]. It has been estimated that about one third of the patients diagnosed with locally advanced PDAC will have the tumour successfully reduced to a resectable one after neoadjuvant therapy[13]. Thus,

neoadjuvant protocols are regarded as suitable for patients with locally advanced PDAC, which can be treated as such and then reassessed for surgery[13].

Following surgery, adjuvant chemotherapy may be applied. Protocols include six months of either gemcitabine monotherapy, or 5-FU with leucovorin, beginning between 8 to 12 weeks after the surgery[13]. The general effectiveness of adjuvant chemotherapy is, however, disputed[13]. The American Society of Clinical Oncology (ASCO) guidelines suggest it as an option in patients who were not given neoadjuvant therapy, had a suboptimal resection (R1, with a positive resection margin), have developed metastases to lymph nodes or have received 4-6 months of adjuvant chemotherapy[13]. The ESMO guidelines advise against any use of adjuvant chemoradiotherapy[13].

1.1.1.9.2 Persistent hurdles

Several factors continue to impede the development of effective therapies for PDAC.

Situated deep in the lower abdomen, behind the stomach and in between the aorta and its abdominal branches, the pancreas is an organ where tumours can grow while being shielded from early detection and encase early the aforementioned blood vessels, rendering surgery unfeasible[20]. This contributes to the frequent advanced stage at detection and to the low resectability rate of newly diagnosed cases[20].

A second obstacle is the aggressive biology of PDAC, most markedly represented by its capacity for early metastasis[20]. More than 50% of newly diagnosed PDAC patients present distant metastases, and most patients than undergo resection show metastases within 4 years of surgery, indicating the presence of micro-metastases in seemingly localized presentations of PDAC[20].

Thirdly, PDAC shows extremely severe physiologic effects, with 80% of patients displaying the wasting symptom of cachexia already at diagnosis, which PDAC-induced exocrine and endocrine disruptions may add upon[20]. Cachexia results in poor treatment tolerance, and cachexic patients

show reduced survival rates after surgery and chemotherapy when compared to their non-cachexic counterparts[20].

The fourth factor addresses resistance to therapy. PDAC shows rapid progression and low rates of complete response even after the administration of the most effective systemic agents currently available coupled with radiotherapy[20]. An increasing amount of evidence suggests that resistance is conferred to a significant extent by pancreatic cancer stem cells (PCSCs), which have also been implicated in metastasis[29]. They will be further discussed in **Section 1.1.3**. In addition, PDAC displays a highly prominent extracellular matrix (ECM) which negatively impacts treatment success, but disrupting it also shows unavoidable adverse effects[30]. Desmoplasia induced by the tumour greatly increases the percentage of connective tissue in the PDAC tumour mass, from about 5% to averages of 60%, but values as high as 90% have also been recorded[31]. Cancer-associated fibroblasts (CAFs) are thought to be the source of the desmoplasmic stroma that develops in PDAC[4]. In turn, CAFs are thought to originate from pancreatic stellate cells that become activated through inflammation or injury, and subsequently deposit ECM in the form of laminins, fibronectins, collagen and hyaluronan[4]. The dense tumour stroma has been implicated in gemcitabine resistance, occurring as a result of the scarce diffusion of the drug to the tumour cells[29].

1.1.1.10 Prognosis

Both GnP (gemcitabine and nab-paclitaxel) and FOLFIRINOX have achieved some limited improvements in overall survival in metastatic PDAC[25]. Marked improvements have been obtained for resected PDAC, with the overall survival rate increasing from 22.1 to 35 months in the last 10 years, arguably as a result of the development of better adjuvant therapies[25]. But the major issues of chemoresistance and frequent recurrence after surgery remain[25]. Metastatic organotropism to liver and lungs is a significant factor that drives the high mortality rate of PDAC[32].

Involvement of lymph nodes in patients with resectable PDAC is an important prognostic factor[13]. The persistence of higher than normal levels of CA 19-9 after surgery is a factor of poor prognosis suggestive of occult metastasis[20]. In general, declining CA 19-9 levels after systemic therapy indicate an improved prognosis of OS[20]; conversely, a return to high levels of CA 19-9 after a nadir in these values can suggest treatment failure: recurrence or metastasis[20].

Furthermore, PDAC cells that have acquired a more mesenchymal phenotype through EMT resist better both traditional chemotherapy and the therapeutic targeting of the cell cycle (e.g., through MAPK inhibitors)[33].

1.1.2 Pancreatic cancer cell heterogeneity

Typically, the PDAC tumour mass encompasses a variety of malignant and stromal cell types, such as ductal cells, acinar cells, endocrine cells, endothelial cells, fibroblasts, stellates, macrophages, T and B cells[34]. Malignant cells in the tumour mass show variation in terms of aspects important to the biology of tumour, such as the progression of epithelial-mesenchymal transition, metabolic alterations and the differentiation stage.

1.1.2.1 Epithelial-mesenchymal transition

Epithelial-mesenchymal transition is the process through which epithelial cells acquire a phenotype resembling that of mesenchymal cells. It is defined by changes in cell morphology, the expression of mesenchymal markers and changes in the migratory behaviour[33]. The process is reversible and occurs in varying degrees in different cancer cells. It is regulated by a number of transcription factors: Snai1, Snai2, Zeb1, Zeb2, and Twist, which are also employed by stem cells during embryonic development and can be frequently found expressed in human cancers, including PDAC[33]. Thus, cells undergoing EMT become more stem-like, with implications upon their proliferative abilities, which are reduced, and upon their migratory abilities, which are increased[33]. Further, the cell death mechanisms of the cells suffer alterations as a result of EMT,

which contributes to resistance to therapy[33]. Several signalling pathways are associated with EMT, including TGF-β, Notch, Wnt/β-catenin, inflammatory JAK/STAT and NF-κb[33]. Mouse models show that EMT occurs very early in the development of PDAC, anticipated by inflammation[33]. Circulating PDAC cells that have undergone EMT, as evidenced by the presence of EMT markers, have stem cell-like characteristics which allow them to initiate new tumours at distant locations[33].

Heterogeneous populations of CAFs infiltrated through PDAC's dense stroma secrete humoural factors that promote EMT[33]. In the absence of CAFs, PDAC cells that have undergone EMT can preserve their associated markers for a short time, but they will subsequently restore their initial epithelial phenotype, which sheds light upon the fact that EMT is, in fact, not irreversible[33].

1.1.2.2 Metabolic changes

Hypoxia is typical for PDAC, an indication of poor prognosis, and a factor that has been evidenced to increase cancer cell proliferation and survival, EMT, invasiveness, metastasis, and resistance to both chemotherapy and radiotherapy[35]. This occurs both through the mediation of HIF-1α and independently of it[35]. An aspect that plays an important causative role in the development of hypoxia is the exceptionally extensive and poorly vascularized desmoplastic stromal reaction of PDAC, almost unparalleled in the other carcinomas[35].

Notably, hypoxia and the associated nutrient deprivation do not result in major cell death in PDAC[35]. Thus, hypovascularized PDAC presents certain early adaptations to the challenging environment[35]. Glycolysis and the amino acid production both increase in response to oxygen deprivation[35]. Amino acid production is realized via protein degradation, protein glycosylation and fatty acid synthesis[35]. Cellular components are also often recycled and scavenged[35]. Collectively, the early adaptations of PDAC to hypoxic settings are known as the metabolic switch[35].

The increase in amino acid demands seems to be a very early phenomenon in tumour development, and the metabolic reprogramming necessary in order to provide cancer cells with branched-chain amino acids was suggested to anticipate a diagnosis of PDAC by 5 years[35].

Further, PDAC, like most cancers, manifests a dependence to the otherwise non-essential amino acid glutamine, which it metabolizes through a non-canonical pathway centered on transaminases, while the *KRAS* and *MYC* oncogenes perform the reprogramming of key enzymes in the glutamine pathway[35].

In addition to development of hypoxia, the acidification of the tumour microenvironment by lactic acid is another metabolic factor that creates favourable conditions for the tumour[35]. It promotes chronic inflammation and suppressed the T-cell mediated adaptive immune response, and thus supports a pro-tumour immunologic remodeling[35]. The production of interleukin-17 and interleukin-23, dependent on lactate, is conducive to an inflammatory tumour microenvironment that attracts pro-tumoural immune cells[35].

A mouse model has shown that mutations that activate *KRAS* induce the development of PDAC and its maintenance through alterations in the regulation of anabolic glucose metabolism[36]. Furthermore, it was demonstrated that subsequently turning off *KRAS* after it had been activated to induce PDAC led to the initial regression of the tumour, but also that a subset of cells would nonetheless manage to survive and trigger tumour relapse[36]. These surviving cells were then transcriptionally characterized and found to have stem cell-like features and to rely more on oxidative phosphorylation than glycolysis, which suggests that the metabolism of CSCs is markedly different from that of non-CSCs in PDAC[36].

In addition, PCSCs also present alternative metabolic pathways, relying on fatty acids and mevalonate for their survival[36].

1.1.2.3 Differentiation

PDAC is, at root, most often a disease of acinar differentiation, in which mature acinar cells get reprogrammed to acquire a ductal phenotype[37]. It has been found that *PTF1A*, the master regulator of acinar differentiation, is a major suppressor of PDAC initiation, and that its downregulation is necessary condition for the *KRAS*-induced acinar-to-ductal metaplasia (ADM) in both mice and humans[37].

ADM can be favoured by the downregulation of signal transduction pathways such as TGF-β/SMAD[37]. The combination of *SMAD4* deficiency and *TGF-β* overexpression is able to trigger pancreatic metaplasia further leading to PanIN development, while also being implicated in the fibrosis of pancreatic stroma and the autocrine activation of PSC[38]. The roles of TGF are carried further by PI3K and MAPK signalling, through both SMAD-dependent and SMAD-independent mechanisms[38].

In mouse models, it was evidenced that although PDAC can alternatively arise from ductal cells, it appears that ductal cells with mutations in *KRAS* alone cannot form PanIN lesions[38]. However, acquiring further aberrations such as the deletion of *PTEN*, does lead to the formation of lesions resembling human IPMNs that, in turn, progress to PDAC[38].

After initiation, PDAC cells eventually de-differentiate through the EMT, gaining a mesenchymal phenotype and stem cell-like traits in the process[39]. EMT and its reverse process play a role in the regulation of invasion, metastasis, and the generation of PCSCs[39].

1.1.3 Pancreatic cancer stem cells

1.1.3.1 Definition and discovery

According to the cancer stem cell (CSC) paradigm, CSCs constitute about 5% of the total number of cancer cells and are essentially at the apex of the tumour hierarchy, serving as the main regulators of tumour progression[32]. The term “cancer stem cells” is used in reference to two

characteristics of these cells, shared with normal stem cells: self-renewal, and the ability to produce differentiated progeny[40].

Cancer stem cells were first identified in 1997, in acute myelogenous leukemia[41], and since then the existence of cancer stem cells was discovered in brain[42], breast[43], colon[44], esophagus[45], liver[46], lung[47], ovarian[48], and prostate[49] cancers. In the case of PDAC, the first report of cancer stem cells dates to 2007[50].

1.1.3.2 Origins of PCSCs and their role in PDAC

It is presently not known how PCSCs originate[32]. Their resemblance with normal stem cells suggests their origins may lie among transformed tissue-specific stem cells or progenitor, bone-marrow derived cells. Another hypothesis is that they originate in dedifferentiated cells present in adult tissue[32] through de novo mutations[51].

Since their discovery, PCSCs have been conclusively shown to be involved in PDAC resistance to chemotherapy, displaying increased prevalence within the tumor after treatment with gemcitabine[52]. In general, CSCs are innately more resistant to both chemotherapy and radiotherapy and present superior invasive and metastatic abilities when compared to differentiated cancer cells[32]. The chemoresistance of CSCs is largely due to enhanced DNA repair abilities and tolerance to DNA damage, high levels of detoxification enzymes, quiescence and epigenetic changes, as well as due to their specific interactions with the tumour environment[32]. Autophagy and the scavenging of reactive oxygen species are also increased in CSC relative to normal cancer cells, with these two mechanisms also contributing to the marked chemoresistance of CSCs[53]. In addition, upregulated multidrug resistance transporters, the frequent dysregulation of anti-apoptotic proteins from the Bcl-2 family, increased aldehyde dehydrogenase activity, the overexpression of c-FLIP and that of IAP family proteins in CSC also contribute to the distinctive resistance of CSCs to anti-cancer therapies[53]. Finally, similarities between CSCs and normal stem cells have also been argued to play a central role in

chemoresistance[54]. Because the preservation of normal stem cell is of essential importance in an organism due to their key role in maintaining the organism's cell pool, mechanisms to prevent apoptosis and senescence have evolved in these cells, and these mechanisms are believed to be hijacked by CSCs[54].

Furthermore, CSCs display the aberrant activation of developmental signaling pathways and exhibit the upregulation of pathways of central importance in metastasis[40].

In addition to drug resistance and the progression of metastasis, PCSCs have been shown to control other essential aspects of the evolution of PDAC: primary tumour growth and disease recurrence[32]. PDAC has been thus described as a prototypical example of a CSC-driven disease[32].

The understanding of CSCs as a state rather than as a hardwired type has been growing in recent years[32]. This is of particular significance in the context of chemotherapy and radiotherapy response, when the cells that make up the bulk of the tumour can get eradicated, but the CSCs more often survive, which in turns leads to the emergence of a CSC niche that favours stemness features in PDAC cells[32]. Thus, differentiated tumour cells can be converted to CSCs via reprogramming to replenish the pool of CSCs[32]. During or after treatment, PCSCs exhibit markedly enhanced DNA repair abilities, drug efflux activity, metabolic reprogramming, quiescence, EMT and autophagy, as well as epigenetic alterations, interactions with the tumour microenvironment and changes in the regulation of developmental pathways, all which contribute to their resilience within the tumour, their resistance against anti-proliferative therapies, and their involvement in disease recurrence[32].

Notably, the phenotypic and functional variability of PCSCs is comparable to the interpatient variability identified in primary pancreatic tissue[32]. This diversity suggests that different events in the natural history of the disease, such as relapse and disease progression, might have different CSC signatures[32].

1.1.3.3 The identification of PCSCs

PCSCs have been asserted to exhibit a diverse array of markers, such as CD24, CD44, CD133, CXCR4, Oct4 and c-Met[32].

CD24 is a glycosylphosphatidylinositol-anchored membrane protein that regulates EMT phenotypes associated with the activation of the Wnt/β-catenin pathway during tumour differentiation[32]. It has been linked with metastasis, high grade tumours and lower OS in PDAC[32]. Surface CD24 is usually considered a marker of presumed PCSC when it is expressed in conjunction with epithelial specific antigen (ESA) and CD24[32], and the molecular signature composed of these three markers (CD44 + CD24 + ESA +) was the earliest identification introduced for potential CSCs in PDAC[32].

CD44 is a transmembrane receptor that binds hyaluronan, considered a bona fide marker of CSCs[32]. It is thought to regulate EMT, tumour plasticity, and the development of therapy resistance and tumour recurrence[32].

CD133 is a glycosylated pentaspan protein that is also recognized as a CSC marker in other cancers[32]. It regulates multiple signalling pathways, including Akt, Bcl-2, Src and Ras, as well as some of their downstream effectors of Ras: ERK, JNK, PI3K and STAT. CD133 also activates the Wnt pathway through its physical association with β-catenin and HDAC6[32], thus promoting EMT, cancer cell migration and metastasis. Hypoxia has been shown to increase the expression of CD133 in PDAC cells, mainly through the mediation of HIF-1α[32]. Furthermore, overexpression of CD133 was found to increase dye efflux and ALDH activity, features regarded as facets of genuine CSCs[32]. Through the metabolic plasticity observed in response to the accumulation of reactive oxygen species (ROS) in CD133+ PDAC cells, CD133 offers enhanced survival and thus contributed to drug resistance[32]. CD133 also contributes to drug efflux[32]. In addition, high expression of CD133 is also correlated with negative prognosis factors such as lymph node metastasis[32]. It has

also been observed that cells presenting both CD133 and CXCR4 can induce primary tumour growth with full differentiation in permissive recipients[32]. However, it must be added that recent studies asserts that CD133 is, in fact, not a useful marker of either tumour stage or disease activity, bearing insignificant associations with both factors[32]. Thus, the possibility that CD133 describes cells that can evolve to CSCs rather than bona fide CSCs exists, and more research is needed to assess the clinical value of this proposed PCSC marker[32].

CXCR4 is a chemokine receptor significantly overexpressed in several cancers, which acts as a moderator of the tumour microenvironment and of tumour-stroma interactions[32]. In part through its crosstalk with key oncogenic pathways such as Akt, ERK and β-catenin, the chemokine axis plays a role in metastasis[32]. PDAC cells presenting both CXCR4 and CD133 have an enhanced ability to generate liver metastases[32], and high expression of CXCR4 in PDAC patients is associated with lower OS, higher chances of developing lymph node metastases, and higher chances of a liver recurrence[32].

Oct4 is the main factor in pluripotency and contributes to the proliferation, migration and invasive character of PCSCs[55]. Pancreatic tumours where *Oct4* was knocked down together with *Nanog* exhibited a reduction in the aforementioned characteristics of the PC stem cells, and sensitized them to gemcitabine[55]. The mechanism that accounts for the observed events is thought to involve the regulation of the expression of *Bcl-2* and *Caspase-3*[55].

c-Met is a receptor tyrosine kinase that can promote invasive behaviour in cancer cells[32]. It interacts with the hepatocyte growth factor (HGF) to stimulate a variety of signalling pathways: PI3K/Akt, JAK/STAT, Ras/mitogen-activated protein kinase (MAPK), Src, and Wnt/β-catenin[32], being thus involved in tumour proliferation, angiogenesis, resistance to apoptosis, tumour-stroma crosstalk, EMT, invasion and metastasis[32].

Other PCSC markers reported in the literature include ALDH[51], DCLK1[32], ESA[51], LGR5[56], Nanog[56], Nestin[57] and SOX2[57].

However, a comprehensive understanding of the role the PCSC markers play in characterizing PCSC subpopulations is lacking, while lines of contradictory or at any rate not easily reconcilable evidence about these markers have occasionally emerged, and there is no global, general molecular signature that separates PCSCs from the other PDAC cells[32].

Computational methods to quantify cancer stemness from transcriptomics data have also been developed in recent years, such as the machine learning-based mRNA expression-based stemness index (mRNAsi)[58]. Methods to computationally quantify cellular pluripotency (e.g. ORIGINS[59], CytoTRACE[60]), and to determine the trajectory of differentiation in single-cell RNA-sequencing data (e.g. Slingshot[61], Monocle[62]) also exist.

1.2 Single-cell RNA-sequencing

The question of identifying cell subpopulations of biological interest can be addressed by using the technology of single-cell RNA-sequencing (scRNA-seq), which is able to delineate cell subpopulations grouped based on their transcriptomic patterns. Single-cell RNA-sequencing has become the state-of-the-art technology for uncovering different cell types and functions within organs and tissues[63]. The first instance of mRNA-sequencing of a single cell dates back to 2009[64]. The process consisted of the following steps: cell lysis, cDNA synthesis, primer removal, poly (A) tailing, second-strand cDNA synthesis, PCR amplification, cDNA shearing, adaptor ligation and library amplification[64]. Since then, several novel scRNA-seq methodologies have been developed, with improvements in key steps, such as single-cell capture, cDNA amplification, library preparation, thus achieving a massive reduction in the costs of the process[64].

Currently, single-cell RNA-sequencing technologies differ in terms of the methods chosen for some of the main steps of the process. Cell capture can be realized using FACS (Smart-seq, Smart-seq2, MATQ-seq, CEL-seq, MARS-seq), microfluidics (Fluidigm C1, Seq-Well) or microdroplets (Drop-seq, 10x Genomics, inDrop-seq, DNBelab C4)[63]. Most technologies, but not CEL-seq, MARS-seq and

inDrop-seq, which use in vitro transcription (IVT), use PCR for the cDNA amplification step[63]. Smart-seq, Smart-seq2, MATQ-seq, Fluidigm C1 and MATQ-seq perform full-length sequencing, whereas the other listed methods perform 3′ or 5′ sequencing, which is cheaper but shows limitations in terms of reporting the mRNA isoform where poly (A) tails are attached[63]. Biases in the amplification step can occur for both PCR and IVT, therefore most methods use unique molecular identifiers (UMIs) to barcode each individual mRNA molecule, exceptions here being Smart-seq, Smart-seq2 and Fluidigm C1, which however cannot be used for larger numbers of cells[63].

Single-cell RNA-sequencing has been used to investigate cellular heterogeneity in cancer, and in particular, CSCs. For instance, a recent single-cell RNA-sequencing study in colorectal cancer cells collected from eight patients, aiming to identify CSCs using 11 traditional markers of colon cancer stem cells, found minimal overlap between the subpopulations characterized by each of the markers, thus demonstrating the heterogeneity of the putative CSCs in colorectal cancer[65], while a scRNA-seq study in glioblastoma showcased some limitations of CD133 as a CSC marker, while uncovering new potential CSC markers for that cancer[66]. Meanwhile, a single-cell RNA-sequencing study in intrahepatic cholangiocarcinoma identified CSCs by quantifying the developmental potential of the cells using CytoTRACE, rather than by using conventional CSC markers[67]. Thus, scRNA-seq technology offers ample possibilities for the identification of cells evidencing cancer stemness, using both marker-based, experimentally-derived, descriptions of these cells, and computational methods for the automated identification of CSCs from the expression data without invoking prior knowledge of putative CSC markers.

Additionally, scRNA-seq has been used, for instance, to identify drug resistant populations in lymphoma and their associated pathways[68], and to evidence distinct drug resistant states in non-small-cell lung carcinoma (NSCLC), after the treatment of NSCLC cells in culture with EGFR inhibitors erlotinib and gefitinib[69].

1.3 The promise of epigenetic therapies in solid tumours

Epigenetic modifications can be understood as those aspects of chromatin biology[70], most importantly covalent modifications to histones or to DNA[71], that impact gene expression without altering the sequence of DNA[70]; it is the interpretation of the genome what gets modified, not the genetic code itself[71]. DNA methylation takes place at cytosine residues and has a silencing effect upon gene expression[71], while hydroxymethylation, which also occurs at cytosine residues, has been linked with the opposite effect[72]. Histone methylations and acetylations, among other epigenetic modifications, can affect several amino acids and have a silencing or an activating effect depending on the affected residue[71]. Epigenetic changes can be inherited[73].

As evidenced in recent years, the majority of cancers display epigenetic deregulations[73], often as a direct result of the alteration of epigenetic machinery[71]. Mutations in genes that regulate epigenetic changes have been found to a contribute to a wide array of tumour-related phenomena: tumour initiation, cell growth, immune evasion, metastasis, heterogeneity and drug resistance[73]. Thus, epigenetic mutations have become attractive targets in modern oncology research[73].

Human cancers can display mutations in gene encoding all the principal classes of epigenetic proteins, such as DNA methylation enzymes, chromatin remodelling complexes, histone modification enzymes, histone mark readers and histone proteins[73]. Epigenetic modifications can also occur in an indirect fashion, through mutations in gene encoding metabolic enzymes that in turn impact the function of epigenetic proteins[73].

Four major phenotypes are known to occur in human cancers due to epigenetic mutations[73]. These are: DNA promoter hypermethylation, genome-wide DNA hypomethylation, abnormal histone modifications (including abnormal binding of histone readers), and abnormal chromatic structures[73]. In turn, these phenotypes arise through several biological mechanisms, and achieve

distinctive effects[73]. DNA promoter hypermethylation can be triggered in two manners: the upregulation of DNA methyltransferases, and loss-of-function mutations in DNA demethylases, and results in the selective suppression of gene expression[73]. Genome-wide hypomethylation appears through loss-of-function mutations in DNA methyltransferases, and results in genome instability[73]. Diminished histone acetylation can occur as an effect of either the upregulation of histone acetyltransferases or loss-of-function mutations in histone acetylases; analogously, abnormal histone methylation is caused by mutations in histone methyltransferases or demethylases[73]. In addition, mutations in the gene encoding histone variant H3.3 can also trigger aberrant histone methylation[73]. Alterations of histone methylation can also take place in an indirect way, through mutations in the genes encoding two metabolic enzymes, IDH1 and IDH2, that result in the production of the 2-HG oncometabolite[73]. In turn, 2-HG inhibits histone demethylases and TET proteins, thus increasing histone methylation. These histone modifications, as well as mutations in genes encoding histone readers (e. g. BRD-containing proteins), can affect multiple facets of cellular activity, such as transcription, splicing, DNA replication, DNA repair, and cell cycle control[73]. Furthermore, abnormal chromatin structures, which occur due to loss-of-function mutations in the switch/sucrose non-fermentable chromatin-remodelling complex or alterations in histone modification and DNA methylation, likewise impact transcription, DNA repair and cell cycle control[73].

About 30 years ago, it was first observed that cancer cells display aberrant methylation, in the form of a combination of genome-wide hypomethylation and promoter hypermethylation[73]. Both aspects are tumourigenic: genome-wide hypomethylation, through genomic instability, as well as through the potential activation of oncogenes and transposable elements, promoter hypermethylation, through the selective suppression of tumour suppressor genes[73]. Abnormal cancer-associated DNA methylation has no known cause, but aging and diet are believed to play a role[73].

Other epigenetic modification promoting the development of cancer include: mutations affecting the nucleosome remodeling complex, mutations in genes encoding histone modifiers, dysregulation of histone readers, mutations in genes encoding histones, and mutations in genes encoding metabolic enzymes[73].

From a mechanistic standpoint, all epigenetic modifications can be divided into two classes, gain and function and loss of function, of which loss of function modifications are significantly more challenging to target[73].

In PDAC, the list of epigenetic modifications that have been documented in the literature includes: altered gene methylation due to aberrant expression of DNA methyltransferases 1, 3a and 3b[74], high expression of *HDAC7*[75], the early downregulation of tumour suppressor *WNK2* through promoter hypermethylation[76], the downregulation of *TFPI-2*, a Kunitz-type serine proteinase inhibitor which counters invasion and metastasis by protecting the matrix from degradation, driven by aberrant methylation[77], the overexpression of histone variant *H2A.Z*, which helps PDAC cells overcome the senescence barrier and thus favouring tumor growth and drug resistance and correlating with poor prognosis[78].

While the exploitation of epigenetic modifications is overall less well established in oncology research than the exploitation of genetic ones, there are several reasons why the area holds significant promise[73]. These include: the high prevalence of epigenetic modifications in cancer; the potential pleiotropic effect of epigenetic targeting, by simultaneously obstructing multiple cancer-associated pathways; the important role played by some epigenetic proteins in immune evasion; the reversibility of epigenetic modifications, which opens new therapeutic avenues, for example through the reactivation of epigenetically silenced tumour suppressor genes[73]. Perhaps most fundamentally, epigenetics builds upon the realization that cancer is a genomic disease, with the implication that epigenetic drugs can target the genome in its entirety[79].

Historically, epigenetic drugs have found applications in the treatment of several hematological malignancies, such as myelodysplatic syndromes and T-cell lymphomas, due to the mutational profiles frequently displayed in these diseases, which often exhibit mutated or otherwise dysregulated genes encoding epigenetic regulators[70]. Epigenetic drugs have shown a comparatively lower efficacy in the treatment of solid tumours[79], although remarkably positive preliminary results were also obtained, as epigenetic drugs were found to promote the chemosensitization or counter chemoresistance in solid tumours, when used as part of combination regimens[80].

Encouragingly, in 2020, tazemetostat became the first epigenetic therapy drug to gain FDA approval for a solid tumour indication: locally advanced or metastatic epithelioid sarcoma[81]. Tazemetostat is a lysine methyltransferase inhibitor that targets EZH2, the enzymatic subunit of the PRC2 complex involved in transcription silencing[81], and a first in class drug, being the first histone methyltransferase inhibitor to gain FDA approval[82]. At present, many other histone methyltransferase inhibitors are in development, and some are being studied in clinical trials[82].

With the FDA approval of tazemetostat, the number of classes of epigenetic drugs currently approved by FDA for oncological indications has increased to four: histone acetylases, DNA methyltransferases, IDH1, EZH2[83]. Overall, the fast-growing number of drugs successfully targeting epigenetic modifications in cancer indicate that epigenetics is an area that holds remarkable prospects in terms of treating both hematological malignancies and solid tumours[84]

Further, epigenetics can offer new avenues for targeting CSCs and thus improving cancer therapies[85]. The deregulation of epigenetics pathways is thought to contribute to tumourigenesis, especially through its promoting of the maintenance and survival of CSCs[85]. Moreover, because epigenetic modifications such as histone modifications and DNA methylation are crucial in the differentiation of normal stem cells to specific lineages, abnormal epigenetic modifications may have the ability to transform normal stem cells into CSCs[85]. The epigenetic regulation of

important signalling pathways involved in embryonic development such as Wnt/β-catenin, Hedgehog and Notch has been found to be disrupted in several cancers, and in turn these abnormalities support the occurrence of pro-tumour events such as the Notch-mediated development of drug resistance, which stimulates the overexpression of drug efflux transporters such as ABCG2, MDR1 and MRP1, which use ATP to transport drugs out of the cell against the concentration gradient[85].

1.4 Research objectives

In this thesis, single-cell RNA-sequencing is used in order to identify PCSCs by combining different descriptors of these cells from the literature with multiple computational tools for the prediction of cellular development potential, to determine the genes and processes distinctively activated in the detected PCSC, and to study the role of epigenetic processes in stemness characteristics in PDAC by chemically inhibiting BRD9, a promising epigenetic target reportedly involved in the genesis of several cancers[86], upon PCSC.

The specific aims of this thesis are:

1. To identify cancer stem cells in single-cell RNA-sequencing data by combining evidence coming from the traditional identification of PCSC using putative markers and from novel bioinformatics-based methods.
2. To determine the genes and processes characterizing these cancer stem cells, thus uncovering new regulators of PCSC and potential targets for their annihilation.
3. To examine the effects of the chemical inhibition of BRD9, an epigenetic regulator recently reported as a driver of tumorigenesis, upon the identified PCSC.

CHAPTER 2

MATERIALS AND METHODS

2 Materials and methods

2.1 Cell culture

PDAC cells from the A13A cell line[87], which originated in a primary tumour, provided by the Christine Iacobuzio-Donahue lab, were grown in Thermo Scientific™ Nunc™ EasYFlask™ Cell Culture Flasks (Thermo Scientific™ 156367) having a growth area of 25 cm^2, in 5 mL of adherent cell culture medium of composition listed in **Table 2.1**. Cells were split at 80-90% confluence by incubating them with 0.05% trypsin-EDTA (ThermoFisher 25300062) for 7-10 minutes at 37°C.

Component	**Supplier and catalog number**	**Dilution factor in solution**
DMEM, high glucose, GlutaMAX™ Supplement, pyruvate	Life Technologies 31966047	1X
Fetal Bovine Serum, heat inactivated	Sigma-Aldrich F9665-500ML	10X
MEM Non-Essential Amino Acids Solution	Thermo Fisher Scientific 11140035	100X
MEM Vitamin Solution (100X) 100 ml	Life Technologies 11120037	100X
Penicillin Streptomycin (10,000 U/mL)	Thermo Scientific 15140122	100X

Table 2.1. Components of the adherent growth condition medium used for the culture of the A13A cells used for scRNA-sequencing.

Cells were then treated for 24 h under the following three treatment conditions:

- Activin A (10 µM)

- Activin A (10 μM) and I-BRD9 (10 μM)
- SB-431542 (10 μM)

This experimental setting evaluated the effects of inhibiting BRD9, an epigenetic target recently reported to play an oncogenic role in several cancers[86], and of modulating Activin signalling, which was reported to drive the self-renewal and tumorigenicity of PCSC[88]. Activin A was used for the activation of Activin signalling, and SB-431542 was used for its inhibition.

For the patient cells, the processing of the tumour sample, the compound treatment and the single-cell RNA-sequencing were performed by Wanzi Hong and Yaoxin Liu, in the laboratory of Lei Jiang, Guangdong Provincial Geriatrics Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences. The PDAC tumour sample was collected from a 50-year-old female patient. The medium used for the growth of the PDAC patient cells consisted of the following components: DMEM/F12 (1X), L-glutamine (100X), Pen-strep (100X), FGF (1000X), EGF from 20 μg/ml stock (1000X), insulin from 10 μg/ml stock (2000X) and B27 (100X). The PDAC biopsy was processed in accord with the protocol below:

1. Before collecting the sample, a collagenase IV aliquot (Collagenase Type IV: Stem Cell Technologies 7909, 2.5 mg/ml in PBS) was defrosted in the water bath and ice blocks were put in the sample collection container. The sample was collected on ice.
2. The sample was processed in a Class 2 Microbiological Safety Cabinet (MSC). Transport media containing tissue was poured into a 10ml Petri dish.
3. The sample was broken into pieces as small as possible using a scalpel.
4. Using a forceps, the tissue was inserted to a 15ml falcon tube containing 10ml of primary growth medium + 2.5 mg/ml collagenase IV + 2 mg/ml dispase II (Sigma-Aldrich 4942078001) + 1 mg/ml trypsin inhibitor (Sigma-Aldrich T6522) + 1 unit/ml DNase I (NEB M0303S). Then, the tissue was incubated at 37°C with shaking speed set at 50rpm for 40 min.

5. The dissociated cells were repeatedly collected at intervals of 20 min to increase cell yield and viability.

6. Cell suspensions were filtered using a 70um cell strainer. Red blood cells (RBC) were removed by RBC lysis buffer (Invitrogen, Cat. no. 1966634) with 1 unit/ml DNase I.

7. The digested tissue was decanted into a 15ml falcon with growth medium and spun at 300xg for 10mins to sediment single cells.

8. Cells were resuspended in primary growth medium, counted and divided into four groups, corresponding to the four treatment conditions.

9. Samples were transferred to low-adherent flat-bottomed plates.

Cells were then treated for 72 h under the following four treatment conditions:

- DMSO
- I-BRD9 (10 µM)
- DMSO + gemcitabine (1µM)
- I-BRD9 (10 µM) + gemcitabine (1µM)

This experimental setting assessed the effects of inhibiting BRD9, in comparison to treatment with standard-of-care drug gemcitabine, a combination of I-BRD9 and gemcitabine and a control.

2.2 Single-cell RNA-sequencing

For the A13A scRNA experiment, single-cell capture and reverse transcription were performed using the Drop-seq method[89], by Siim Pauklin and Stefania Militi. Cells were loaded into a microfluidics cartridge at a concentration of 310 cells per µl. Cell capture, lysis, and reverse transcription were all performed using a Nadia instrument available from Dolomite Bio. Reverse transcription reactions were performed using ChemGene beads also captured in the microfluidic wells. Beads were collected from the device, and subsequently cDNA amplification was

performed. Prior to PCR, the beads were treated with Exonuclease I. Following PCR, purified cDNA was used as an input for Nextera tagmentation reactions. The quality of the cDNA library was assessed using a TapeStation (Agilent Technologies). All cDNA purification steps were performed using Ampure XP beads (Beckman). High quality samples were then sequenced on a NextSeq 500 sequencer using a 75-cycle High Output kit (Illumina).

For the patient scRNA experiment, 10x Genomics technology was used, with the v3 chemistry. Cells were collected, centrifuged and washed twice with 1mL 1X PBS containing 0.04% BSA, then a single-cell suspension in PBS was obtained. Automatic cytometry was used to determine the cell concentration, and the sample volume was calculated based on the optimal cell sampling concentration supplied by 10X official website and the target capture number. Samples were then put on ice for the subsequent preparation of gel bead-in-emulsions and reverse transcription. For library construction, SPRIselectbeads was used to purify the product, followed by adaptor ligation and SPRIselectpurification. Library concentration was determined using Qubit® 3.0 Flurometer (Life Technologies) and Agilent 2100 High Sensitivity DNA Assay Kit (Agilent Technologies) was used to determine the distribution of library product fragments. The qualified library was then sequenced on an Illumina HiSeq platform.

For the A13A dataset, count matrices obtained from the raw single-cell RNA-sequencing, for both spliced and unspliced reads, were generated using kb-python[90], with a velocity index created using Ensembl[91] annotations for human. For the patient dataset, the corresponding count matrices were obtained using Cell Ranger[92] and velocyto[93].

2.3 Analysis of single-cell RNA-sequencing data

2.3.1 General considerations

All scRNA-sequencing data analysis was performed in R[94,95], using the RStudio integrated development environment[96]. The main package used for the analysis of the data was the R toolkit Seurat[97-100].

The visualisation of gene expression patterns was done using the native Seurat functions FeaturePlot and VlnPlot, and the visualisation of categorical descriptors of cells was done using DimPlot. Kernel density gene expression were made using Nebulosa[101].

Generic custom visualisations, including histograms, bar plots and scatter plots, were created using the ggplot2[102] package. Venn-Euler diagrams were made using the eulerr package[103]. The intersections of more than three sets were visualized using the ComplexUpset package[104]. Symmetric tabular data, such as correlation matrices, were visualised with the corrplot package[105], with adjustments for flexible display made using the editGrob function. To make the plots amenable for being easily joined by other plots in the same picture and for being saved with the ggsave function from ggplot, the grid.echo and grid.grab functions from the gridGraphics package[106] were used. Gene-concept network plots illustrating the results of enrichment analysis were drawn using the cnetplot function from the enrichplot package, included in the DOSE suite from Bioconductor[107]. The conversion of plots to ggplot objects employed the as.ggplot function from the ggplotify package[108].

The dplyr package[109] was used to facilitate operations on data frames, for instance the passing of data frame columns rather than column names (strings) as function arguments.

The DOSE package from Bioconductor[107] was used to perform enrichment analysis, on terms from the Gene Ontology database[110], the KEGG database[111] and the WikiPathways database[112]. The

enrichment analysis relied upon the corresponding annotation of the human genome, imported from the org.Hs.eg.db annotation package from Bioconductor[113].

2.3.2 Quality control

Genes expressed in fewer than 10 cells were removed from the data. Next, low-quality cells were identified. First, the identification of doublets was performed using the scDblFinder R package, available from Bioconductor[114], selected for the task because it showed the best performance among eight doublet identification methods in a recent benchmarking study, in terms of both the accuracy of its predictions and its computational efficiency[115]. As the standard scDblFinder pipeline uses random simulated doublets in order to train the model to recognize real ones, thus inducing variability in the results, 100 scDblFinder runs were used to construct an aggregate (consensus) prediction. After the 100 runs resulted in the allocation of 100 singlet or doublet predictions for each cell, a cut-off was determined for the minimum number of doublet predictions at which a cell would be classed as a doublet, as described below:

- All cells in the Seurat object were grouped by the number of times they were predicted to be doublets. Thus, a vector V1 with 100 elements was created to store the number of cells predicted to be doublets in exactly N runs, where N took values from 1 to 100. The cells never predicted to be doublets were not included.
- A reverse cumulative sum was constructed and stored in a vector V2 with 100 elements. Thus, each V2[i] was defined as follows, for $1 \leq i \leq 100$:

$$V2[i] = \sum_{j=i}^{100} V1[j]$$

- The index at which the corresponding element in the vector of reverse cumulative sums most closely approximated the mean number of doublets predicted in the 100 runs as measured by the absolute value of the difference was taken as the desired cut-off.

The rationale for evaluating the optimal choice for the cut-off in this manner was assuring that the number of doublets as per the consensus doublet prediction does not stray far from the numbers for the same predicted in the scDblFinder runs, thus benefitting from scDblFinder's estimation of the optimal proportion of doublets in the dataset.

To evaluate both run-to-run variability and the improvements made by the consensus prediction, Jaccard similarity scores were employed. For two sets A and B, the Jaccard similarity J(A, B) is defined as follows[116]:

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Here, |.| denotes the cardinality of a set (number of elements), while ∩ denotes the set intersection. The Jaccard similarity score takes values between 0, when the two sets share no elements and 1, when the two sets share all elements. Higher scores correspond to higher similarities.

After doublets were identified, the subsequent steps involved removal of outliers for metrics of interest from each experimental condition. Since no fixed filtering cut-offs that can account for the massive variability of single-cell RNA-sequencing readings exists[117], different numerical cut-offs were provided for each experimental condition, based on the visual inspection of the histograms showing the distribution of metrics for each condition.

The metrics employed in quality control were:

- The complexity (novelty) score of each cell, defined as the log10 ratio of its number of features and number of detected unique molecular identifiers (UMIs)[118], short random nucleotide sequences ligated to template molecules, used to count the absolute number of molecules[119].

- Outliers at the low end of the spectrum of novelty scores were identified for each condition and marked for removal.

- The number of genes per cell.
- The number of UMIs per cell.
    - For the number of genes and the number of UMIs per cells, outliers at both ends were removed. However, the identification of outliers at the higher end of these metrics was done quite conservatively, in order to stay close to the number of doublets predicted by scDblFinder.
- The percentages of mitochondrial genes among the genes detected in each cell. Mitochondrial genes have been evidenced to be distinctly upregulated in broken cells due to the loss of cytoplasmic content[120].
- The percentage of ribosomal genes among the genes detected in each cell. Generally abundant and showing high cell-to-cell variability, ribosomal genes tend to be frequently assessed as variable features[121]. However, this variability is thought be driven, in part, by technical reasons, as the expression of ribosomal genes is generally not restricted to only a few clusters[121]. Removing cells with an unusually high percentage of ribosomal genes may thus bring improvements in the feature selection step, favouring the selection of a more biologically informative set of variable features.
    - Both percentages were calculated using the PercentageFeatureSet function from Seurat, and only outliers at the higher end were removed. The patterns used to identify mitochondrial and ribosomal genes, respectively, are listed in **Table 2.2**
    - Marking for removal cells with a high percentage of ribosomal genes was done conservatively, as a harsh filtering would have carried the risk of inadvertently removing valid biological signal.

Gene type	**Pattern**

Mitochondrial	“MT^-“
Ribosomal	"^RP[SL][[:digit:]]|^RPLP[[:digit:]]|^RPSA"

Table 2.2. The patterns used to identify mitochondrial and ribosomal genes.

- Shannon diversity. Cells with low Shannon diversity are likely to be low-quality[122], therefore outliers at the low end were removed. For each cell *x*, Shannon diversity index, denoted by H(x), is defined as follows[123]:

$$H(x) = -\sum_{i=1}^{N} p_{x,i} \ln p_{x,i}$$

  Here, $p_{x,i}$ represents the proportional abundance of gene *i* reads in cell x, and *i* takes values from 1 to the total number of genes recorded in the dataset, denoted by *N*. Shannon diversity was implemented using the vegan package for R[124].

- Simpson diversity. Also calculated using vegan and used for cell filtering in the same fashion. Simpson diversity, denoted by $\lambda(x)$ for each cell *x*, is defined as follows:

$$\lambda(x) = 1 - \sum_{i=1}^{N} p_{x,i}^2$$

  The notations $p_{x,i}$, i and N have the same meaning as before.

Of note, Shannon and Simpson diversity both combine the two main aspects of diversity, richness and evenness[125], defined in this context as the number of genes expressed in a cell and the relative abundance of gene expression profiles in a cell, respectively[126]. An advantage of Shannon diversity over Simpson diversity is that it is more sensitive to rare genes. Conversely, Simpson diversity is more sensitive to moderately abundant genes[127].

Following the removal of low-quality cells, the Seurat subsets corresponding to the experimental condition were merged again, and the step of filtering out genes expressed in fewer than 10 cells was repeated.

2.3.3 Regression of undesired sources of variation and integration

After quality control, regression of undesired sources of variation was performed, in order to obtain clusters that accurately represented the biological variation of interest within the cells. The Seurat object was again split by experimental condition. Cell cycle predictions were made on the subsets using Seurat's CellCycleScoring function, preceded by an initial normalization step performed with Seurat's NormalizeData function, necessary in order to provide the adequate type of input for CellCycleScoring.

Next, the SCTransform[128] normalization method, version 2[129], was employed, regressing out the number of UMI counts, the cell cycle phase and the percentages of mitochondrial and ribosomal genes per cell in the process. SCTransform was chosen for the normalization and scaling of the data and the selection of highly variable features in lieu of Seurat's native NormalizeData, FindVariableFeatures and ScaleData functions because it represents optimally the statistical properties of scRNA data and provides substantial improvements in terms of the detection of variable genes and differential expression analysis[129], having been used to identify biological substructures in scRNA data that had previously eluded detection with the standard Seurat workflow[128]. To assist the variable feature selection step, the return.only.var.genes parameter was set to FALSE and the min_cells parameter was set 0 in the SCTransform runs. Thus, all the genes in the Seurat object were scaled for each of the subsets corresponding to the experimental conditions. The union of the 3000 variable features detected for each of the Seurat subsets was then taken and used as the anchor.features parameter in the PrepSCTIntegration function. In addition, the variable features of the Seurat object obtained when the subsets were merged again at the conclusion of the SCTransform runs were set to the same union. This method of

variable feature selection was implemented because the default Seurat option for this setting, namely the SelectIntegrationFeatures function, may miss small subpopulations of cells existing in only one of the Seurat subsets corresponding to the experimental conditions, if these small subpopulations are characterized by genes that are not variable in the other subsets as well.

Then, PrepSCTFindMarkers was run on the Seurat object, making the merged Seurat object amenable to differential expression analysis. Finally, rare genes (expressed in < 10 cells) were removed again from counts slot of the SCT assay, created for the Seurat object as a result of the SCTransform run.

With the scaling and normalization steps now completed, a PCA was run and the object was integrated using Harmony[130], selected because it was recommended as the first method of choice for single-cell RNA-sequencing data integration in a recent benchmarking study that compared the results of 14 available methods[131]. The group.by.vars argument was set to "orig.ident" in the RunHarmony function, and the assay.use argument was set to "SCT".

2.3.4 Dimensionality reduction and clustering

Next, a UMAP dimensionality reduction was performed. The reduction parameter in RunUMAP was set to "harmony", and the cut-off for dimension selection was set at the point where the amount of variance explained by the 50 dimensions changed by less than 0.1% between consecutive dimensions of the Harmony reduction.

For the patient dataset, an additional step was necessary at this stage. Because the surgical sample did not contain solely tumour cells but stromal cells were also present in the mix, these non-tumour cells had to be filtered out. This was implemented by identifying regions of the plot (clusters) characterized by the expression of established ductal markers (*KRT19*, *KRT8*, *KRT18*) or stromal markers (*ISLR*, *MFAP5*, *LUM*), and removing the clusters characterized by the latter from the data. Then, genes expressed in fewer than 10 cells were again removed, and the whole

dataset was subjected again to the same normalization and integration procedure detailed upon in the previous section, following by dimensionality reduction with UMAP.

After the UMAP reduction was generated, clustering was performed. First, nearest neighbours in terms of gene expression were calculated using the FindNeighbors function from Seurat, with the reduction parameter set to "umap". The determination of the clusters was performed with the FindCluster function, using the default Louvain algorithm[132].

To establish the most biologically informative clustering, different cluster resolutions, followed by the identifications of cluster markers, were explored iteratively. In order to obtain the desired cluster boundaries that satisfactorily followed the expression of observed markers, clustering was performed at high resolutions, resulting in large numbers of clusters that were subsequently manually merged in accord with marker expression. Cluster merging was performed using the RenameIdents function from Seurat, followed by the renumbering of the clusters in the Seurat metadata. Because the mergers abrogated the increasing order in which cluster numbers were displayed on plots, the order was restored by dropping the levels of the seurat_clusters column from the Seurat metadata and adding new, sorted levels.

2.3.5 Identification of clusters associated with stemness

The identification of clusters of putative CSC character was performed using multiple descriptors of cancer stem cells available in the literature, combined with predictions assigned using ORIGINS, a novel R package designed to identify stem and progenitor cells and to quantify pluripotency using protein-protein interaction networks[59], trajectory analysis using Slingshot, a tool developed to identify cell lineages[61], and gene set knowledge discovery using Enrichr[133].

Firstly, a list of commonly proposed PCSC markers was selected from five recent reviews on the subject: *C-Met (MET)*[55,134-136], *CD24*[29,55,134-136], *CD44*[29,55,134-136], *CD93*[134], *CD133 (PROM1)*[29,55,134-136],

DCLK1[55,134], *EPCAM (ESA)* [29,55,134-136], *NANOG*[134], *LGR5*[134,136], *ABCG2*[29,135], *ALDH1 (ALDH1A1)*[29,135], *CXCR4 (LAP3)*[29,55,135] , *NES*[135], *ABCB1 (MDR1)*[55], *OCT4 (POU5F1)*[29,55] and *CD90 (THY1)*[136] and extended with a list of other genes experimentally identified as PCSC markers and reported as such in individual publications: *KIT*[137], *SOX2*[57], *TSPAN8*[138], *REG4*[139], *CD9*[140], *EZH2*[141], *SSEA-1*[88], *SSEA-4*[88], *AGR2*[142], *GLRX3*[143] and *HNF1A*[144] reaching a total of 27 genes.

Secondly, 13 sets of genes computationally predicted to be directly linked with cancer stemness, or predicted to be linked with poor prognosis, carcinogenesis and progression, traits associated with cancer stemness, were collected from published studies. These lists included: 1 cancer stem cell signature, 1 gene set associated with characteristics of CSC cells in PDAC, 9 gene sets associated with the characteristics of CSC in 9 different cancers (gastric cancer, lung squamous cell carcinoma, ovarian cancer, breast cancer, glioma, endometrial cancer, bladder cancer, colon cancer, liver cancer), 1 list of genes indicative of poor prognosis in PDAC, and 1 list of markers of carcinogenesis and progression in PDAC.

A description of the composition of these gene lists and the way they were obtained is provided below:

- A cancer stem cell signature of 14 genes: *TTK*, *CDC20*, *TOPK (PBK), KNTC2 (NDC80), KIF4A*, *MELK*, *PRC1*, *KIF20A*, *ECT2*, *DTL*, *KIF2C*, *GPSM2*, *OIP5* and *KIAA0101 (PCLAF),* obtained by taking an initial set of 84 genes found to be frequently and highly upregulated in cancer cells and to play important roles in the survival and proliferation of cancer cells, while being minimally or not at expressed in healthy tissue, and restricting it to those genes among them that were found overexpressed in *in silico* studies in human embryonic stem cells, induced pluripotent stem cells, and CSC-like breast cancer cells[145].

- A set of 36 genes associated with characteristics of PDAC stem cells: *NEK2*, *PBK*, *NCAPH*, *CENPA*, *TPX2*, *PLK1*, *CDC20*, *KIF4A*, *MKI67*, *HJURP*, *CKS2*, *CCNA2*, *KIF11*, *ZWINT*, *DTL*, *UBE2C*, *CDCA5*, *GINS1*, *CDKN3*, *PTTG1*, *RAD51*, *CCNB2*, *CDK1*, *GINS2*, *KIFC1*, *SKA3*, *NUF2*,

CEP55, *BUB1*, *KIF18B*, *CDC45*, *BIRC5*, *ASF1B*, *AURKA*, *E2F1* and *UBE2T*, obtained by mRNA stemness index (mRNAsi) calculations, a metric develop recently to quantify stemness associated with oncogenic dedifferentiation using machine learning[58] , followed by weighted gene coexpression network analysis (WGCNA)[146] based on 169 normal samples and 142 tumour samples[147].

- A set of 16 markers associated with characteristics of gastric cancer stem cells: *BUB1*, *BUB1B*, *NCAPH*, *KIF14*, *RACGAP1*, *RAD54L*, *TPX2*, *KIF15*, *KIF18B*, *CENPF*, *TTK*, *KIF4A*, *SGOL2 (SGO2)*, *PLK4*, *XRCC2*, and *C1orf112*, obtained through mRNAsi calculations followed by WGCNA, based on 30 normal samples and 343 gastric cancer samples[148].

- A set of 10 markers associated with characteristics of lung squamous cell carcinoma stem cells: *BUB1B*, *CDC25A*, *CDCA5*, *CENPA*, *DKC1*, *NCAPH*, *RAD51*, *SKA3*, *SPAG5*, and *TIMELESS*, obtained likewise through mRNAsi calculations and WGCNA, using data from The Cancer Genome Atlas (TCGA)[149].

- A set of 7 markers associated with characteristics of ovarian cancer stem cells: *BUB1*, *CDC20*, *CCNB2*, *DLGAP5*, *KIF4A*, *NEK2* and *NUSAP1*, identified through WGCNA and GeneMania analysis and with their expression in ovarian cancer stem cells confirmed experimentally through quantitative PCR[150].

- A set of 32 markers associated with characteristics of breast cancer stem cells: *TPX2*, *HJURP*, *CDCA8*, *PLK1*, *KIFC1*, *CENPA*, CCNB2, *KIF2C*, *EXO1*, *TTK*, *KIF4A*, *CDC25A*, *MELK*, *NDC80*, *NCAPG*, *CEP55*, *NCAPH*, *RAD54L*, KIF20A, *KIF18B*, *ORC1*, *CDC45*, *KIF23*, *CDC20*, *BUB1*, *AURKB*, *SKA1*, *FOXM1*, *SGO1*, *DLGAP5*, *CDCA3*, and *BUB1B*, found by employing mRNAsi and WGCNA on data from the TGCA, Oncomine, Gene Expression Omnibus (GEO), Gene Expression Profiling Integrative Analysis (GEPIA) databases[151].

- A set of 51 markers associated with characteristics of glioma stem cells: *GINS2*, *DBF4*, *OIP5*, *PBK*, *CCNB2*, *C1orf112*, *CDCA8*, *MELK*, *DEPDC1B*, *DDIAS*, *RCC1*, *DLGAP5*, *NUF2*, *AC099850.3*, *FAM72B*, *RAD51*, *ORC1*, *CDK1*, *SGO1*, *CDCA2*, *KIF2C*, *SKA3*, *PRR11*, *BUB1*, *TTK*, *ESCO2*, *FBXO5*, *NCAPG*, *NDC80*, *SGO2*, *KIF4A*, *TPX2*, *NCAPH*, *CKAP2L*, *HASPIN*, *CENPI*, *SMC4*, *CKAP2*, *KIF15*, *TOP2A*, *ZWINT*, *KIFC1*, *LMNB1*, *E2F2*, *KNL1*, *KIF14*, *ASPM*, *C18orf54*, *MCM10*, *KIF11* and *CENPF*, found using epigenetically regulated mRNAsi followed by WGCNA, based on glioma data from TGCA[152].

- A set of 19 markers associated with characteristics of endometrial cancer stem cells: *ORC6*, *C1orf112*, *RAD54L*, *SGO2*, *BUB1*, *PLK4*, *KIF18B*, *BUB1B*, *TTK*, *NCAPG*, *XRCC2*, *CENPF*, *KIF15*, *RACGAP1*, *ARHGAP11A*, *TPX2*, *KIF14*, *KIF4A* and *NCAPH,* found using mRNAsi and WGCNA, based on data from TGCA and RNA-seq of 552 endometrial cancer samples obtained from the University of California Santa Cruz Genome Browser[153].

- A set of 13 markers associated with characteristics of bladder cancer stem cells: *AURKA*, *BUB1B*, *CDCA5*, *CDCA8*, *KIF11*, *KIF18B*, *KIF2C*, *KIFC1*, *KPNA2*, *NCAPG*, *NEK2*, *NUSAP1*, and *RACGAP1*, found using mRNAsi and WGCNA, based on data from TGCA that has also been validated using the Oncomine and GEO databases[154].

- A set of 27 markers associated with characteristics of colon cancer stem cells: *CHEK1 (CHK1)*, *BUB1*, *KIF18A*, *TTK*, *PLK4*, *NUP107*, *SPC25*, *DNA2*, *DDIAS*, *MCM10*, *RFC4*, *NCAPG*, *BUB1B*, *SUV39H2*, *NCAPH*, *KIF23*, *CDK1*, *MELK*, *DEPDC1B*, *NEIL3*, *MTFR2*, *PNPT1*, *ORC6*, *CCNA2*, *MAD2L1*, *CENPA* and *XRCC2*, using WGCNA and mRNAsi on RNA sequencing data from TCGA[155].

- A set of 15 markers associated with characteristics of liver cancer stem cells: *KIF4A*, *TTK*, *CCNB1*, *CDC20*, *NCAPG*, *CCNB2*, *CDC45*, *UBE2C*, *CENPA*, *AURKB*, *RRM2*, *CDCA8*, *BIRC5*, *TPX2*, and *KIF2C*, found using mRNAsi, WGCNA and the maximal clique centrality method using data from TGCA[156].

- A set of 20 poor prognosis genes, identified as hub genes through WGCNA on a dataset of 25 PDAC samples and 7 healthy pancreatic samples: *ANLN*, *ZWINT*, *CEP55*, *TOP2A*, *UBE2C*, *ZWILCH*, *CDK1*, *STIL*, *KIAA0101 (PCLAF)*, *GINS1*, *CENPF*, *PRC1*, *RRM2*, *ASPM*, *FANCI*, *KIF20A*, *CENPU*, *NUSAP1*, *CENPK* and *TK1*[157]:

- A set of 14 markers of PDAC carcinogenesis and progression, identified as hub genes from three microarray PDAC datasets analysed with Cytoscape: *KIF4A*, *RRM2*, *FAM83D*, *ASPM*, *NCAPG*, *TPX2*, *NUSAP1*, *RACGAP1*, *MKI67*, *KIF20A*, *CENPF*, *UBE2C*, *BUB1*, and *KIF11*[158].

Together, 118 genes appeared in at least one of these 13 gene lists. The interaction network of the proteins encoded by these genes, as obtained using the STRING database[159], was highly connected, showing a mean of 78.19 interactions per protein. The number of interactions with other CCRSA proteins reported for each CCRSA protein is displayed in **Figure 2.1**:

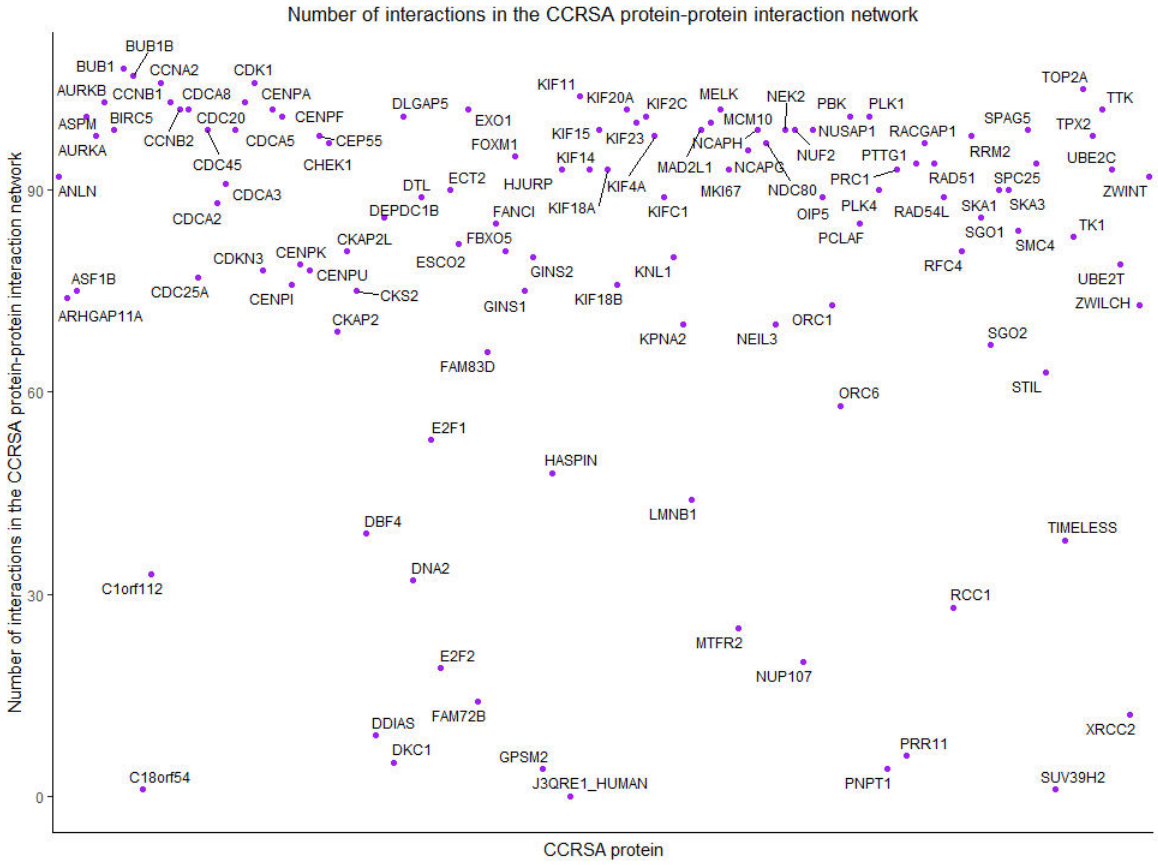


Figure 2.1. Protein-protein interaction network of the proteins encoded by the CCRSA genes.

The gene lists enumerated above showed a considerable amount of overlap, indicating that similar biological processes recurrently emerge as predicted regulators of CSC characteristics in a variety of human cancers, including PDAC, and that the same mechanisms are closely connected to poor prognosis and progression in PDAC. **Table 2.3** showcases the genes which made more than one appearance in the 13 gene sets:

Number of occurrences among the 13 CCRSA gene sets	
Gene(s)	**Number of occurrences**
KIF4A	9
BUB1	8
TTK, NCAPH, TPX2 and *NCAPG*	7
BUB1B	6
CDC20, KIF2C, CENPA, CCNB2, CDK1, KIF18B and *CENPF*	5
MELK, KIF20A, KIF11, UBE2C, KIFC1, RACGAP1, NUSAP1 and *CDCA8*	4
PBK, NDC80, NEK2, PLK1, ZWINT, CDCA5, RAD51, SKA3, CEP55, CDC45, KIF14, RAD54L, KIF15, SGO2, PLK4, XRCC2, C1orf112, DLGAP5, ASPM and *RRM2*	3
PRC1, DTL, OIP5, PCLAF, MKI67, HJURP, CCNA2, GINS1, GINS2, NUF2, BIRC5, AURKA, CDC25A, ORC1, KIF23, AURKB, SGO1, DEPDC1B, DDIAS, TOP2A, MCM10, ORC6, and *MAD2L1*	2

Table 2.3. The 65 genes that appeared more than once in the 13 CCRSA gene sets.

Out of the 78 pairs selected from the 13 sets, 35 (44.9%) shared at least 5 genes, while only four pairs of gene sets shared no genes. Every set had one set with which it shared at least 5 genes, as illustrated in **Figure 2.2**:

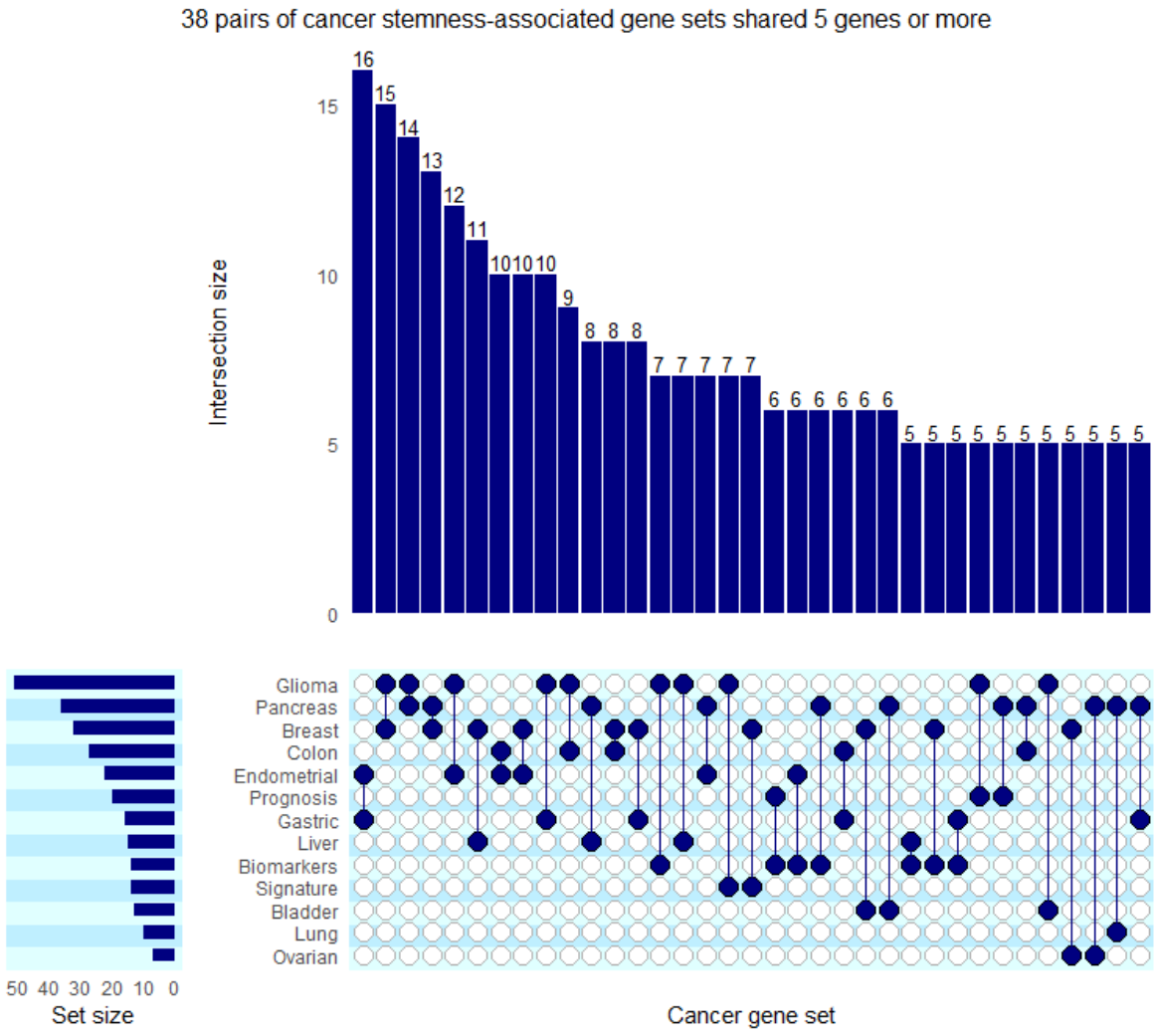


Figure 2.2. Intersections of CCRSA gene sets containing at least 5 genes

Importantly, for a number of these genes, the association with the formation or maintenance of CSCs in various cancers has received additional experimental support – while not being necessarily reported explicitly as "CSC markers". Selected examples include:

- *GPSM2*[160], *BIRC5*[160], *AURKB*[161] and *ASPM*[162] in PDAC;
- *BUB1*[163], *TPX2*[164], *RAD51*[165], *GINS2*[166], *MKI67*[167], *FOXM1*[168], *AURKA*[168] and *MCM10*[169] in breast cancer;

- *NCAPH*[170] and *SPC25*[171] in non-small cell lung cancer;
- *PRC1*[172], *SKA3*[173], *NEK2*[174], *KIF15*[175], *GINS1*[176], *UBE2T*[177], *FAM83D*[178] and *ASF1B*[179] in hepatocellular carcinoma;
- *CDK1*[180] in bladder cancer;
- *CDC20*[181], *CENPA*[182], *MTFR2*[183], *TTK*[183], *MELK*[184], *PLK1*[185] and *PBK*[186] in glioma;
- *KIFC1*[187], *KIF11*[187], *ECT2*[188], *E2F1*[189] and *PRR11*[190] in gastric cancer;
- *RAD54L* in head and neck squamous cell carcinoma[191]
- *CDC25A* in colorectal cancer[192]
- *CENPK* in cervical cancer[193].

Of particular note, a recent gene expression study based on 22 surgical tumour specimens from non-small-cell lung cancer patients revealed the greater than 8-fold overexpression of six genes (*PTTG1*, *TOP2A*, *CEP55*, *BIRC5*, *TK1*, and *ASPM*), all also listed among the 118 predicted cancer stemness-associated genes, in the putative CSC population ($ALDH^{high}$) when compared to the $ALDH^{low}$ population[194]. In the same study, *CDK1*, *PRC1*, *ASF1B*, *CDKN3*, *DLGAP5*, *FOXM1*, *KIAA0101 (PCLAF)*, *CDCA3*, *KIF20A*, *NUSAP1*, *CENPM*, *PLK1*, *CDCA8*, *DTL* and *CENPF* were all found to show a greater than 4-fold overexpression in the $ALDH^{high}$ non-small-cell lung cancer cell population[194]. Furthermore, *AURKA* has been reported to regulate the pluripotency of embryonic stem cells[195], while *PLK4* controls the self-renewal of pluripotent stem cells[196] and *CDK1* is important to the maintenance of pluripotency in human pluripotent stem cells[197]. All in all, the findings cited above support the idea that the 118 genes of predicted cancer stemness associations do indeed significantly correspond to stemness-related traits, and that they thus can be used, in conjunction with other lines of evidence, for the identification of cancer stem cells.

Of note, the most significantly enriched GO term corresponding to the 118 genes, as obtained using StringDB[110], was “Cell cycle” (GO:0007049) and the most significantly enriched Reactome[198] pathway was also “Cell cycle” (HSA-1640170). Together, the two “Cell cycle” terms covered 105

out of the 118 genes. An additional 4 genes (*ARHGAP11A*, *DEPDC1B*, *C18orf54* and *FAM72B*) were reported as components of the human mitotic cell cycle machinery[199], and cell cycle associations have also been evidenced for all the remaining genes: *ASF1B*[200], *CKAP2L*[201], *MTFR2*[202], *C1orf112*[203], *AC099850.3*[204], and the DNA repair genes *DTL*[205], *UBE2T*[206], *KPNA2*[207] and *NEIL3*[208]. Therefore, these genes will be further referred to in this thesis as the **Cell-cycle-related stemness-associated (CCRSA) genes**. Given the functional commonalities of the 118 genes, a 14th set was constructed as the union of all the 13 sets. Thus, both the 13 sets and their union were used to identify CSCs in the A13A and patient PDAC scRNA-seq data.

Thirdly, a list of 39 genes overexpressed in a side population (SP) of PDAC cells generated in the L3.6pl cell line, which showcased important attributes of CSCs even in the absence of displaying previously postulated PCSC markers, was collected from its associated publication[209]. In cancer, cells that successfully realize the efflux of the Hoechst 33342 dye are referred to as SP cells, and they present characteristics of CSCs, in particular tumour initiation abilities, chemoresistance, and the presence of stemness-related genes[210]. The attributes of the SP cells identified in the L3.6pl cell line in the aforementioned study were: self-renewal, ability to differentiate into non-SP cells, high capacity for tumorigenesis and metastasis after orthotopic injections in nude mice[209].

The 39 genes identified as differentially overexpressed in the L3.6pl SP were: *AKR1B10*, *ABCG2*, *EID3*, *MIRN221*, *F2RL2*, *GDF15*, *NR0B1*, *AGPAT9*, *MIRN21*, *MAP1B*, *CYP4F3*, *HIST1H1B*, *LRRFIP1*, *SNORD122*, *KIAA0319*, *HIST2H2AB*, *HIST1H1E*, *TXNRD1*, *TM4SF19*, *NID2*, *NTS*, *WNT5A*, *SNORA1*, *MAN1A1*, *S100P*, *SNORA16B*, *AKR1C3*, *GDA*, *HIST1H3B*, *LIMCH1*, *SNORA25*, *SNORD53*, *CPEB2*, *LIPH*, *FLJ10213*, *SNORD5*, *RSPO3*, *TMEM156* and *HIST1H2BF*[209].

Thus, the 27 experimentally-derived proposed PCSC markers, the 118 computationally-predicted genes associated with cancer stemness or associated traits, and the 39 SP markers in L3.6pl PDAC cells were employed as distinct lines of evidence for identification of PCSC in the single-cell RNA

data. No overlap existed between the 118 computationally-predicted genes and either the 27 proposed PCSC markers or the 39 SP markers, while the overlap between the 27 proposed CSC markers and the 39 SP markers was limited to one marker, *ABCG2*.

The lack of overlap between the TDPM genes and the CCRSA genes is likely to be mostly due to the fact the former largely correspond to cell surface markers, aimed to **identify** PCSC, but not necessarily asserted to have a causal role in their formation or maintenance. On the other hand, the CCRSA genes are asserted to be **drivers** of stemness, corresponding to proteins localized in the nucleoplasm, found to be involved in the cell cycle machinery, rather than to surface markers.

The three gene sets will be referred further with the acronyms listed in **Table 2.4**, which also details the principles that were employed in the analysis of the variation in their expression:

Gene set	**Acronym**	**Usage in differential expression analysis**	**Rationale for the usage choice**
27 traditionally-derived PCSC marker genes	TDPM genes	Individually (as single genes)	They are curated from multiple different studies where they were typically reported as single genes. They may not correspond to the same type of CSCs.
118 cell-cycle related stemness-associated genes	CCRSA genes	The 13 sets are used both as individual groups of genes and collectively, by taking their union to be the 14 set.	Each set comes from one study. Because the sets have significant overlaps with each other, they are also considered collectively as a union.

39 side population genes from the L3.6pl cell line	SPLCL genes	Collectively	They all come from the same study.

Table 2.4. The usage notes of the three sets of genes: acronyms, the way they were employed in differential analysis and the rationale for it.

Next, ORIGINS activity scores were used to identify stemness-associated clusters in two ways:

- The statistical significance of the differential representation of each cluster among top-ranking activity cells was calculated. No fixed cut-off was set for the selection of top-ranking activity cells. Instead, the cut-off was allowed to vary between ~0.2% and ~16% of the cells, following estimates about the expected proportion of CSC in PDAC[211], and the corresponding p-values were generated using the methods described in **Section 2.3.8, Setting C.II**.
- Pairwise Wilcoxon tests were used to evaluate differences in the medians of the distributions of the activity scores between all the possible ordered clusters pairs, using the wilcox.test function, with the alternative parameter set to "greater", thus evaluating whether the median activity score was significantly higher between each pair of clusters. The results were subjected to a Benjamini-Yekutieli[212] correction for multiple testing (as detailed in **Section 2.3.6**) available as the BY function from the sgof[213] package, with the significance threshold set at 0.05, and a ranking of the clusters was constructed based on the **difference** between the number of times each cluster was found to have a significantly **higher** median activity score than other clusters and the number of times the same cluster was found to have a significantly **lower** median activity score than other clusters.

The step of calculating ORIGINS activity scores by using the default activity function was found to be computationally expensive. Because the calculations involved one cell at a time, with a min-

max normalization step at the end of the calculations, a modified version of the activity function was constructed, involving parallelization using the parLapply, makeCluster and clusterExport functions from the parallel package and the running of the new function on a server with 60 available cores. In order to minimize memory usage, the Seurat data matrix was not passed directly to the modified activity function, but divided based on columns into 1000 slices, on which activity calculations were computed individually. The results were concatenated at the end of the run and subjected to a min-max normalization, arriving at the same scores as provided by the standard activity function available from ORIGINS, but with a major decrease in execution time.

Next, Slingshot trajectory analysis was used to identify lineages of differentiation along which the clusters were ordered, with the clusters linked to stemness being at the beginning of the lineages reported by Slingshot.

Finally, gene set enrichment was performed using the *CellMarker_Augmented_2021* and *PanglaoDB_Augmented_2021* databases from Enrichr, in order to identify clusters with markers that overlapped with the markers of cell types linked to stemness and cancer stemness in these two databases. To this end, the rba_enrichr function from the rbioapi[214] package was used.

A summary of the six lines of evidence used to identify CSCs in this dataset is available in **Table 2.5**:

PCSC identification method	**Details**
27 experimentally-derived putative markers of PCSC	Curated from 5 reviews and 7 individual publications, in which they were reported as PCSC markers, as supported by experimental evidence.
118 genes computationally predicted to be linked with cancer stemness or with related	Identified mainly through mRNAsi and WGCNA methods, as associated with stemness in

traits (poor prognosis, rapid progression), all associated with the cell cycle.	different cancers: PDAC (1 gene set) and other cancers (10 gene sets), or as associated with prognosis in PDAC (1 gene set) or as biomarkers of progression in PDAC (1 gene set). In the present thesis, these sets were analysed both individually as distinct gene sets and collectively as the union of all the genes reported by the aforementioned studies, amounting to a total of 118 genes. Some are supported by recent experimental evidence in a number of cancer types.
39 markers of a side population (SP) reported in L3.6pl PDAC cells	Selected from a publication which experimentally identified markers of a PDAC cell population that showed some stemness characteristics even in the absence of traditional PCSC markers, amounting to a total of 39 genes.
ORIGINS predictions	No previous knowledge of CSC markers; predictions are automated based on protein-protein interaction networks.
Slingshot pseudotime determination	No previous knowledge of CSC markers; predictions are automated based on gene expression data
Enrichr gene set enrichment analysis	Used to determine the clusters for which associations with stem-like cells are stronger

	than associations with other cell types, based on marker overlap assessments with Enrichr databases.

Table 2.5. An overview of the six lines of evidence used to characterize potential PCSC in the scRNA-sequencing data.

2.3.6 Differential expression analysis

The generation of the list of all differentially upregulated genes in each subset of cells selected from the scRNA data, grouped by cluster, condition or any other criteria, was performed using Seurat's native FindMarkers function, with the only.pos parameter set to TRUE. For most applications, including all the marker lists used in gene overlap assessments, the min.pct and logfc.threshold parameters were lowered to 0 from the default values of 0.1 and 0.25. The min.pct parameter refers to the minimum fraction of cells – in either of two compared groups – in which the gene has to be detected in order to be factored into the differential expression analysis, while the logfc.threshold refers to the minimum average $\log_2$ fold-change between the two groups that a gene included in the analysis must show. These parameters were set to 0 in order to avoid missing any signal, in exchange for an acceptable decrease in speed. In addition, the densify parameter was set to TRUE, obtaining a noted decrease in execution time in return for an acceptable increase in memory usage.

In all cases, marker lists originating as outputs of the FindMarkers function were filtered by a statistical significance threshold. To this end, the Bonferroni-adjusted p-values were selected, rather than the unadjusted ones, as the assessment of the differential expression of a gene between two groups is an implicit instance of multiple testing - one out of a number of possible comparisons equalling the number of genes in the dataset. While each comparison has individually a low probability of delivering a false positive, a larger number of comparisons drives an increasing frequency of false positives rendering the unadjusted p-values inappropriate[215]. The

Bonferroni correction is the simplest and the most conservative (least likely to result in false positives) of the methods that have been developed to address this issue[216], and it was preferred for the selection of markers step because of its computational efficiency. But methods with more power (ability to avoid false negatives) exist, including analytic approaches such as the Benjamini-Yekutieli technique mentioned previously.

Because marker lists were always generated for several groupings of interest rather than for one (for instance, for all the clusters rather than just for one), the Bonferroni-corrected p-values (*p* in the formula below) were further subjected to an additional multiple testing correction, in which they were again adjusted with Bonferroni for the testing of multiple groupings, as follows:

$$p_{final} = \min(p|G|, 1)$$

Here, |G| denotes the cardinality of the grouping of interest, such as the number of clusters or the number of condition selections.

Finally, the genes whose doubly-adjusted p-values did not exceed the conventional threshold of 0.05 were regarded as markers of statistical significance and retained in the corresponding marker list, while the other genes returned by FindMarkers were filtered out from the list.

All parameters in FindMarkers, other than the ones listed above, were kept at default values. Thus, the comparison of the differential expression between the two groups of interest was performed using the default option provided by Seurat, the Wilcoxon rank-sum test[217].

2.3.7 The evaluation of the overlaps of sets of genes

A frequent question occurring in scRNA data analysis is whether the overlap of two subsets *A* and *B* of genes of interest selected from the same set of genes (for example, genes overexpressed in one of the cluster and genes underexpressed after a treatment) reaches statistical significance. This setting is illustrated in **Figure 2.3**. The illustrated gene sets A and B can originate as markers

of the groupings available in the scRNA dataset (e.g., markers of different clusters or of different experimental conditions), or from a source external to the scRNA dataset (for instance, gene sets selected from the literature) after a filtering step removing genes not detected in the gene set N corresponding to the single-cell RNA experiment has been performed.

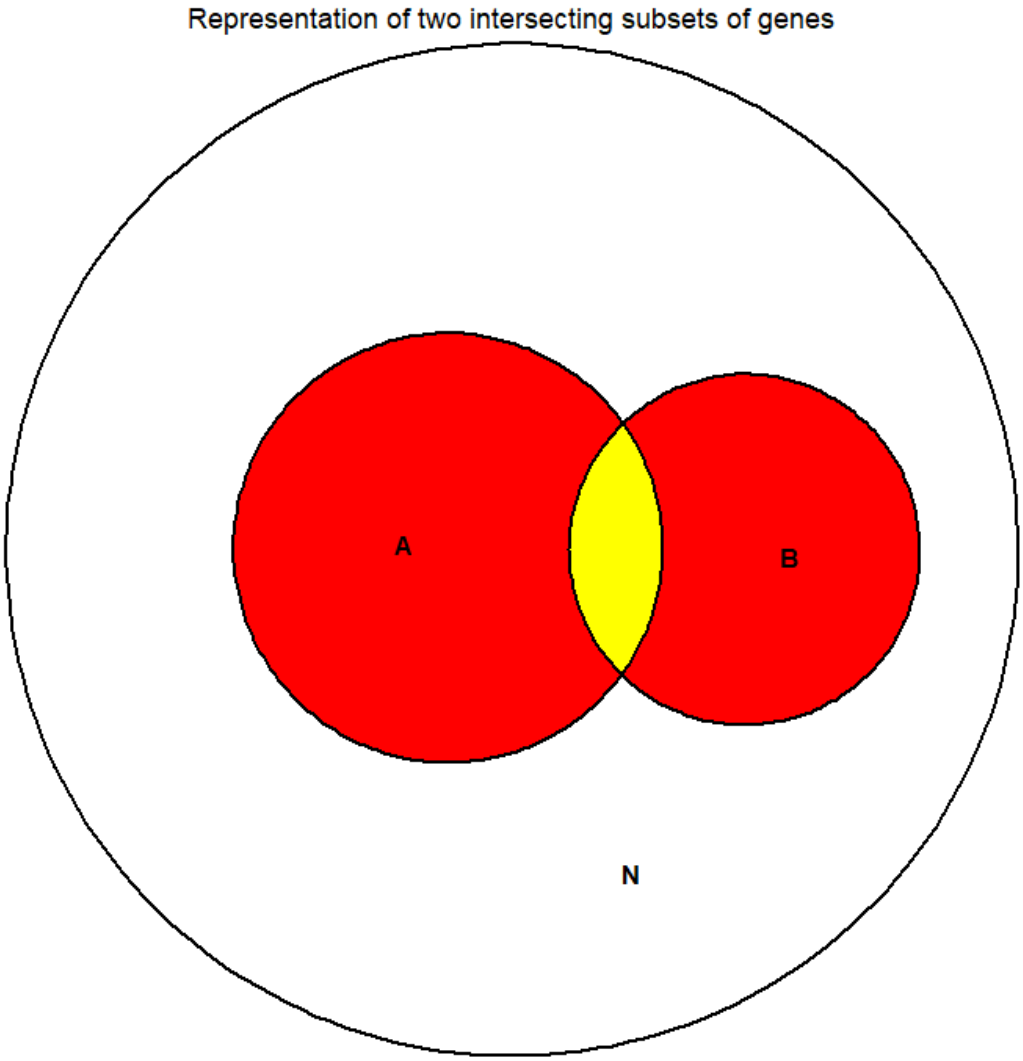


Figure 2.3. The overlap of two subsets belonging to the same set of genes.

To address this problem, a hypergeometric p-value indicating the probability that the overlap of the two subsets could have been of equal or larger size under the null hypothesis was computed using the phyper function with parameters k – 1, a, n – a, b, lower.tail = FALSE, where these arguments signify:

- *k*: The number of genes in the intersection of the two subsets of genes.
- *a*: The cardinality of subset *A (|A| = a).*
- *n*: The total number of genes in the set of genes.
- *a*: The cardinality of subset *B (|B| = b).*
- lower.tail: A Boolean variable that can be adjusted to select either the lower or the upper tail of the probability distribution of the variable of interest, as needed. For this problem, the upper tail of the distribution was the one that was needed.

With the above parameters, phyper calculates the probability that the two subsets of fixed sizes *a* and *b* intersect in at least k points.

The suitability of the usage of a hypergeometric p-value to evaluate the significance of gene overlaps follows from the following observations:

- The probability that two subsets *A* and *B* of a of fixed cardinalities *a* and *b* selected from a set of fixed size *n* share exactly *k* elements can be calculated as below:
    - The number of sets with k elements from n possible elements, known as "*n* choose *k*" and denoted as $\binom{n}{k}$ can be found according to the formula:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

      Here, n! represents the factorial, n! = 1 · 2 · 3 · … · n.
    - The probability that two subsets of fixed sizes *a* and *b* intersect in exactly k points is the ratio of the number of ways to form 2 subsets of fixed sizes *a* and *b* of the original set with n elements that meet this condition (sharing exactly k elements) and the number of ways to form two subsets of fixed sizes *a* and *b* with no other restrictions.

- The numerator of the ratio described above can be calculated as described below:

$$\binom{n}{a}\binom{a}{k}\binom{n-a}{b-k}$$

The first factor describes the number of ways to choose the first subset, of cardinality *a*, selected from a set of *n* elements. The second factor describes the number of ways to choose *k* elements from the *a* elements of the first set that will constitute the intersection with the second set. The third factor describes the number of ways to choose the remaining *b* – *k* elements of the second set, as *k* of them were already chosen as part of the intersection. These *b* – *k* elements must not be in the first set – otherwise the intersection of the two sets would exceed the desired value *k*. Therefore, only *n* – *a* elements of the original *n* are available for this choice.

- The denominator of the aforementioned ratio is:

$$\binom{n}{a}\binom{n}{b}$$

This is simply the number of ways to form two sets of cardinalities *a* and *b* from a set of size *n*.

- With the cancellation of the first factor, the probability that was sought is:

$$\frac{\binom{a}{k}\binom{n-a}{b-k}}{\binom{n}{b}}$$

Using probability and set notation:

$$P(|A \cap B| = k) = \frac{\binom{|A|}{k}\binom{n-|A|}{|B|-k}}{\binom{n}{|B|}}$$

- This is the formula of the probability mass function of a hypergeometric distribution of a population of size *n*, with $|A|$ successes and *k* obtained successes among $|B|$ trials[218].
- Thus, the probability that the two sets of fixed sizes intersect in at least *k* points equals:

$$P(|A \cap B| \geq k) = \sum_{i=k}^{\min(|A|,|B|)} \frac{\binom{|A|}{i}\binom{N-|A|}{|B|-i}}{\binom{N}{|B|}}$$

This sums all the probabilities that the number of shared elements between the two sets is *k*, *k + 1* etc., all the way up to the minimum of *a* and *b* and equals the p-value computed with phyper as indicated above.

This p-value was then used to calculate the probability that the intersection of two sets of differentially expressed genes (e.g. one upregulated in a cluster of choice, one downregulated after a treatment) could have reached the number of detected shared genes or a greater one merely by chance. The significance threshold was set at 0.05.

Because the genes identified by FindMarkers as differentially expressed in the clusters, treatment conditions or any other groupings of interest showed different strengths of the effect (average $\log_2$ fold-change), and the variety of the possible choices for average $\log_2$ fold-change filters can impacts the results of the significance assessment of the overlap between two sets of genes, the following approach was employed for evaluating the overlap of two sets of genes provided by FindMarkers:

- First, all the $\log_2$ fold-change values recorded for the genes in the two marker lists originating as FindMarker outputs were iteratively used as filters for **both** marker lists, whenever a filtering was possible, that is, when the threshold was not high enough to render one of the marker lists empty.

- A hypergeometric p-value was calculated for each overlap of the filtered lists.
- The resulting list of p-values, each corresponding to a different choice of average $\log_2$ fold-change filters, was corrected for multiple testing using the Benjamini-Yekutieli method with an alpha value (significance threshold) of 0.05.
- From the list of Benjamini-Yekutieli-adjusted p-values, the **median** was selected to assess the statistical significance of the overlap in question.

This approach was preferred upon other threshold-free hypergeometric gene expression comparison algorithms existing in the literature, such as the rank–rank hypergeometric overlap[219] because of the reduced computational complexity: O(M + N) for the number of assessed overlaps rather than O(MN) in the Big O notation, where M and N denote the lengths of the compared marker lists, providing thus ample improvements in speed while making acceptable concessions in terms of the quantity of information provided. Comparisons that can be reasonably argued to be less illuminating for the assessment of the significance of the overlap (e.g., between all markers in one list and only the top markers of the other list) were thus skipped by the algorithm.

The same procedure was used to judge the overlap of pairs of gene sets among which one was obtained with FindMarkers, while the other was selected from the literature and included average $\log_2$ fold-change values. Literature gene sets for which no such values were available for the genes were also compared with FindMarkers outputs in this analysis. In this case, the procedure was similar but simpler, as only the gene set originating from FindMarkers was iteratively filtered based on its own average $\log_2$ fold-change, while the set with no available $\log_2$ fold-change values was used in its entirety at each step.

All gene sets selected from the literature were first filtered to remove genes not found in the scRNA dataset, thus becoming subsets of the set of genes expressed in the dataset.

In addition, the overlap of three subsets of genes from the same set was assessed for statistical significance – for instance, genes overexpressed in a cluster, underexpressed after a treatment,

and selected from a gene list from the literature. This setting is illustrated in **Figure 2.4**. As before, the gene sets A, B and C can originate as markers of the groupings available in the scRNA dataset (e.g., markers of different clusters or of different experimental conditions), or from a source external to the scRNA dataset (for instance, gene sets selected from the literature) after a filtering step removing genes not detected in the gene set N corresponding to the scRNA experiment has been performed.

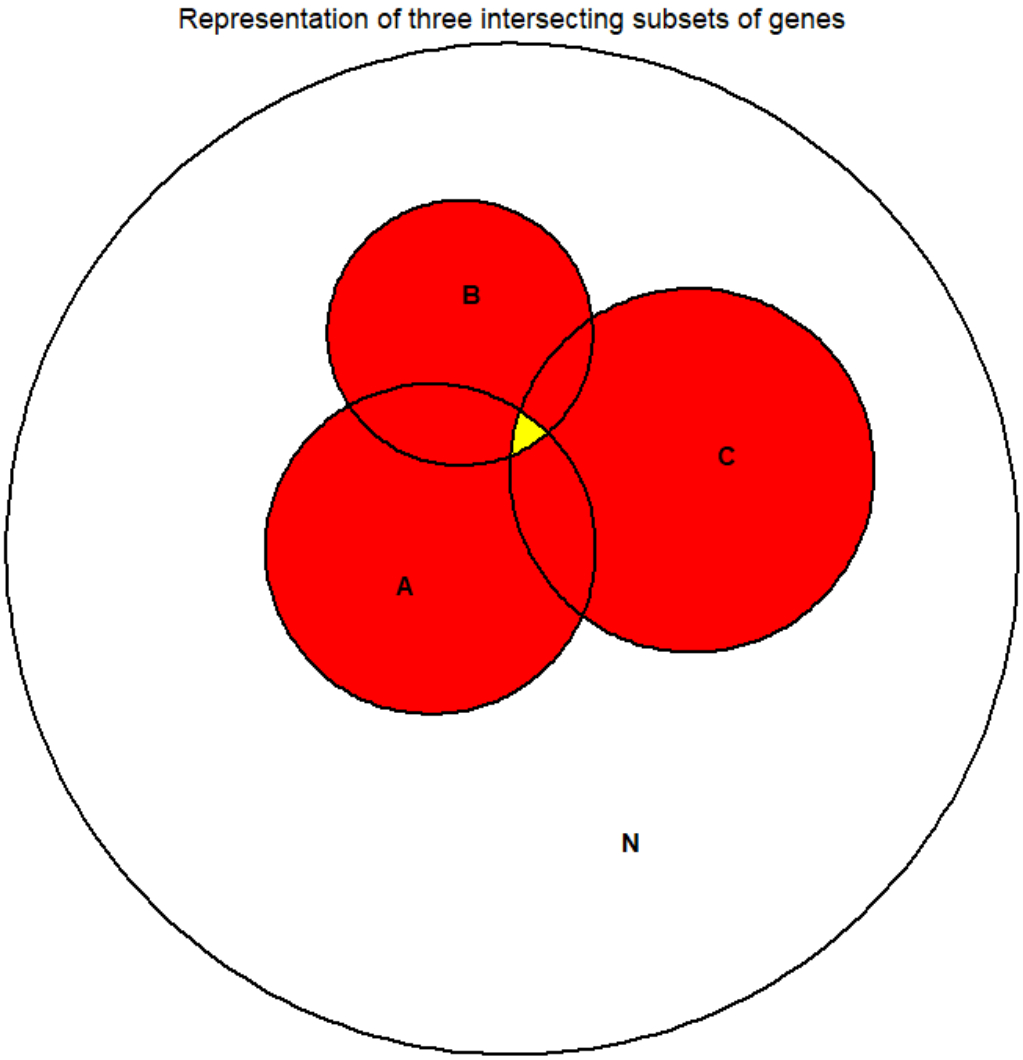


Figure 2.4. The overlap of three subsets belonging to the same set of genes.

The p-value used to evaluate statistical significance in this setting was calculated as follows:

- First, the probability that two subsets *A*, *B* and C of fixed cardinalities *a, b* and c selected from a set N of fixed size *n* share exactly *k* elements was computed as described below:
    - Without losing generality, we will assume a ≤ b ≤ c.
    - The probability that three subsets of fixed sizes *a*, *b* and *c* intersect in exactly k points is the ratio of the number of ways to form 3 subsets of fixed sizes *a*, *b* and *c* of the original set with n elements that meet this condition (sharing exactly k elements) and the number of ways to form three subsets of fixed sizes *a*, *b* and *c* with no other restrictions.
    - The numerator of the ratio described above can be calculated using this expression:

$$\binom{n}{a}\sum_{i}\binom{a}{i}\binom{n-a}{b-i}\binom{i}{k}\binom{n-i}{c-k}$$

      First, the set *A* is chosen by picking *a* elements out of the existing *n*. Then, a sum is taken over all the possible values of $|A\cap B|$, denoted by *i* here, where choosing $A\cap B$ amounts to choosing *i* elements out of *a*, finalizing the choice of *B* (i. e. choosing $B \setminus A$) amounts to choosing $b - i$ elements out of $n - a$ available ones, and *C* is constructed in a similar fashion with *B*, by first choosing *k* values shared with $A\cap B$ and $c - k$ elements outside of it.

      The above expression operates under the convention that $\binom{n}{k}$ is *0* when $k < 0$ or $k > n$. The bounds for *i* will be specified later in this section.

    - The denominator of the ratio has this expression:

$$\binom{n}{a}\binom{n}{b}\binom{n}{c}$$

      This is the number of ways to form three subsets of cardinalities *a*, *b* and *c* from a set with *n* elements.

- After the cancelling out of the first factor, the probability equals:

$$\frac{\sum_i \binom{a}{i}\binom{n-a}{b-i}\binom{i}{k}\binom{n-i}{c-k}}{\binom{n}{b}\binom{n}{c}}$$

- This can be rewritten as below:

$$\sum_i \left(\frac{\binom{a}{i}\binom{n-a}{b-i}}{\binom{n}{b}}\right)\left(\frac{\binom{i}{k}\binom{n-i}{c-k}}{\binom{n}{c}}\right)$$

- We observe that the first factor corresponds to the probability mass function of a random variable following the hypergeometric distribution[218] with the parameters:
  - *n*: population size
  - *a*: number of successes in the population
  - *b*: number of draws
  - *i*: number of observed successes
- Similarly, the second factor corresponds to the probability mass function of a random variable following the hypergeometric distribution with the parameters:
  - *n*: population size
  - *i*: number of successes in the population
  - *c*: number of draws
  - *k*: number of observed successes
- Thus, both quantities were computed efficiently using the dhyper function with parameters *i*, a, *n* – *a* and *b* for the first bracketed factor and *k*, *i*, *n* – *i* and *c* for the second bracketed factor.

- For the practical implementation of the formula, bounds for *i* (the cardinality of the intersection of A and B) had to be explicitly set.
- These were obtained from the relations demanded by the existence conditions of each factor listed above, namely:

$$0 \leq i \leq a$$

$$0 \leq b - i \leq n - a$$

$$0 \leq k \leq i$$

$$0 \leq c - k \leq n - i$$

- Thus:

$$\max(k, a + b - n) \leq i \leq \min(a, n + k - c)$$

- The expression has now the form listed below, where *f(x, y, z, t)* denotes a random variable with a hypergeometric distribution with *x* observed successes, a population size of *y*, *z* successes in the population and *t* draws.

$$\sum_{\mathrm{i} = \max(k,\ a+b-n)}^{\min(a,\ n+k-c)} f(i, n, a, b)\, f(k, n, i, c)$$

- Thus, the probability that the three subsets A, B and C of the same set N, with |A| ≤ |B| ≤ |C|, intersect in exactly three points can be expressed as:

$$P(|A \cap B \cap C| = k) = \sum_{i=\max(k,|A|+|B|-|N|}^{\min(|A|,|N|+k-|C|)} f(i, |N|, |A|, |B|) f(k, |N|, i, |C|)$$

- Consequently, the expression used to compute the p-value, signifying the probability that results as extreme as those observed (the intersection of the three sets being of cardinality *k*), equals:

$$P(|A \cap B \cap C| \geq k)$$

$$= \sum_{i = \max(k,|A|+|B|+|C|-|N|)}^{|A|} \sum_{j = \max(k,|A|+|B|-|N|)}^{\min(|A|,|N|+k-|C|} f(j,|N|,|A|,|B|)\, f(k,|N|,j,|C|)$$

- To verify the correctness of the calculation and the implementation, the following identity was verified through the iterative testing of numerical examples:

$$\sum_{i = 0}^{|A|} \sum_{j = \max(k,|A|+|B|-|N|)}^{\min(|A|,|N|+k-|C|} f(j,|N|,|A|,|B|)\, f(k,|N|,j,|C|) = 1$$

  Because all the probabilities were non-negative, additive under countable unions (by the definition given above, the probability that three sets intersect in exactly *k* elements, where *k* is iterated over a countable set *X*, equals the sum of the probabilities for each *k*) and summed to 1, the provided expression does amount to a probability measure, as desired.

The cases of overlap assessment involved in this scRNA analysis typically involved the comparison of multiple groups (e.g., different clusters and conditions). This induced another multiple testing scenario, driven by the groups, for which a correction was thus required. The Benjamini-Yekutieli procedure was selected for the task.

When multiple sets of Seurat-based groupings were evaluated against each other (e.g. experimental conditions and clusters), all the p-values were aggregated into an array and subsequently adjusted with the Benjamini-Yekutieli correction. Literature gene sets were not compared against each other for overlap, therefore each literature gene set received its own Benjamini-Yekutieli correction for the assessment of overlap with cluster or condition markers.

A detailed overview of the approaches that were used to determine whether the overlap of two or three subsets of genes selected from the same set is statistically significant in all the settings where the question occurred in this study is available below.

- **Setting G.I**.: Two subsets of the same set of genes. For the first subset, an ordering is available via sorting by average $\log_2$ fold-change. The second set is unranked.
    - **Example:**
        - One list of markers determined using FindMarkers and one set among the CCRSA gene sets (see **Section 2.3.5**).
    - **Procedure:**
        - All $\log_2$ fold changes available in the FindMarkers output, except the maximum, are used iteratively to filter the corresponding marker list – at each step, only genes with $\log_2$ fold changes values above the filter are taken into consideration for the significance assessment.
        - For each filter, a hypergeometric p-value assessing the statistical significance of the overlap of the filtered gene sets with the gene set from the literature is calculated.
        - Thus, a number of p-values equal to the number of distinct $\log_2$ fold changes recorded for the ranked gene set, excluding the maximum, is provided for the overlap. The list of p-values is subjected to the Benjamini-Yekutieli correction, after which the **median** of the p-values is taken to represent the p-value of the overlap of the two gene sets.
    - **Further adjustment for multiple testing:**
        - Because of the implicit multiple testing scenario entailed by the usage of the FindMarkers function from Seurat, the p-values obtained from **all** the groupings of interests (clusters, treatment conditions etc.) are

aggregated into an array and corrected once again with Benjamini-Yekutieli.

- Thus, the **doubly-corrected** p-values are the **final** p-values that assess the overlaps of interest.

- **Setting G.II**: Two subsets of the same set of genes. Average $\log_2$ fold-change values are provided for both.
  - **Examples**:
    - Two lists of markers determined using FindMarkers, using different groupings, for instance, one by experimental condition (orig.ident), one by cluster (seurat_clusters).
    - One list of markers determined using FindMarkers and the list of side population markers (see **Section 2.3.5**), for which average $\log_2$ fold-change values are provided in the additional information section of the associated publication[209].
  - **Procedure**:
    - The union of all the average $\log_2$ fold changes available for both gene sets is taken. Values higher or equal to either of the maxima of the two average $\log_2$ fold changes sets are excluded from it.
    - Both marker lists are iteratively filtered using **all** the average $\log_2$ fold changes existing in the aforementioned union. The same average $\log_2$ fold change filter is applied to **both** gene sets.
    - The p-value of the overlap is then computed as in **Setting G.I.**
  - **Further adjustment for multiple testing:**
    - As in **Setting G.I**.
  - **Note**:

        - If the setting involved two lists of FindMarkers outputs, the length of the final array of p-values will equal the product of the cardinalities of the two groupings (for example, the number of clusters multiplied by the number of conditions).
        - If the setting involved one list of FindMarker outputs and the list of side population markers, the length of the final array of p-values will equal the cardinality of the grouping (for example, the number of clusters).

- **Setting G.III:** Three subsets of the same set of genes. For the first two subsets, an ordering is available via sorting by average $\log_2$ fold-change. The third set is unranked.
    - **Example**:
        - Two lists of markers determined using FindMarkers with different groupings, and one set among the CCRSA gene sets (**Section 2.3.5**).
    - **Procedure**:
        - Average $\log_2$ fold changes are obtained as in **Setting G.II** and iteratively applied to both sortable sets.
        - A p-value assessing the significance of the overlap when each filter is applied is calculated with the methods for three-way overlap significance assessments listed earlier in this section and these p-values are used to obtain the p-value of the overlap of the three sets as in **Setting G.I**.
    - **Further adjustment for multiple testing:**
        - As in **Setting G.II**.
    - **Note**:
        - The length of the final array of p-values will equal the product of the cardinality of the two groupings.
- **Setting G.IV:** Three subsets of the same set of genes. All three have orderings imposed by average $\log_2$ fold-change values.

- **Example**:
  - Two lists of markers determined using FindMarkers with different groupings and the side population markers (**Section 2.3.6**).
- **Procedure**:
  - The union of all the average $\log_2$ fold changes available for the three gene sets is taken. Values higher or equal to any of the maxima of the three average $\log_2$ fold changes sets are excluded from it.
  - The p-value is computed as in **Setting G.III**.
- **Further adjustment for multiple testing:**
  - As in **Setting G.I**.
- **Note:**
  - As in **Setting G.III**.

In addition to sets of genes, single genes – for instance, TDPM genes - were assessed for their differential expression in cells coming from different groupings such as clusters, conditions, or combinations thereof, as detailed below:

- **Setting SG:** A single gene tested in a single type of FindMarkers groups.
  - **Example:** A TDPM gene tested for expression in different clusters or experimental conditions.
  - **Procedure:** The doubly Bonferroni-corrected marker lists corresponding to the groups of interest are assessed for the differential expression of the gene.
  - **Further adjustment for multiple testing:** None. All multiple testing correction steps were already performed during marker selection.

Additionally, the problem of determining the significance of the overlap of two subsets of genes selected from different sets also surfaced in this analysis – namely, as the assessment of the

overlap between the markers of groupings in the A13A dataset and those of groupings in the patient dataset. This setting is illustrated in **Figure 2.5**:

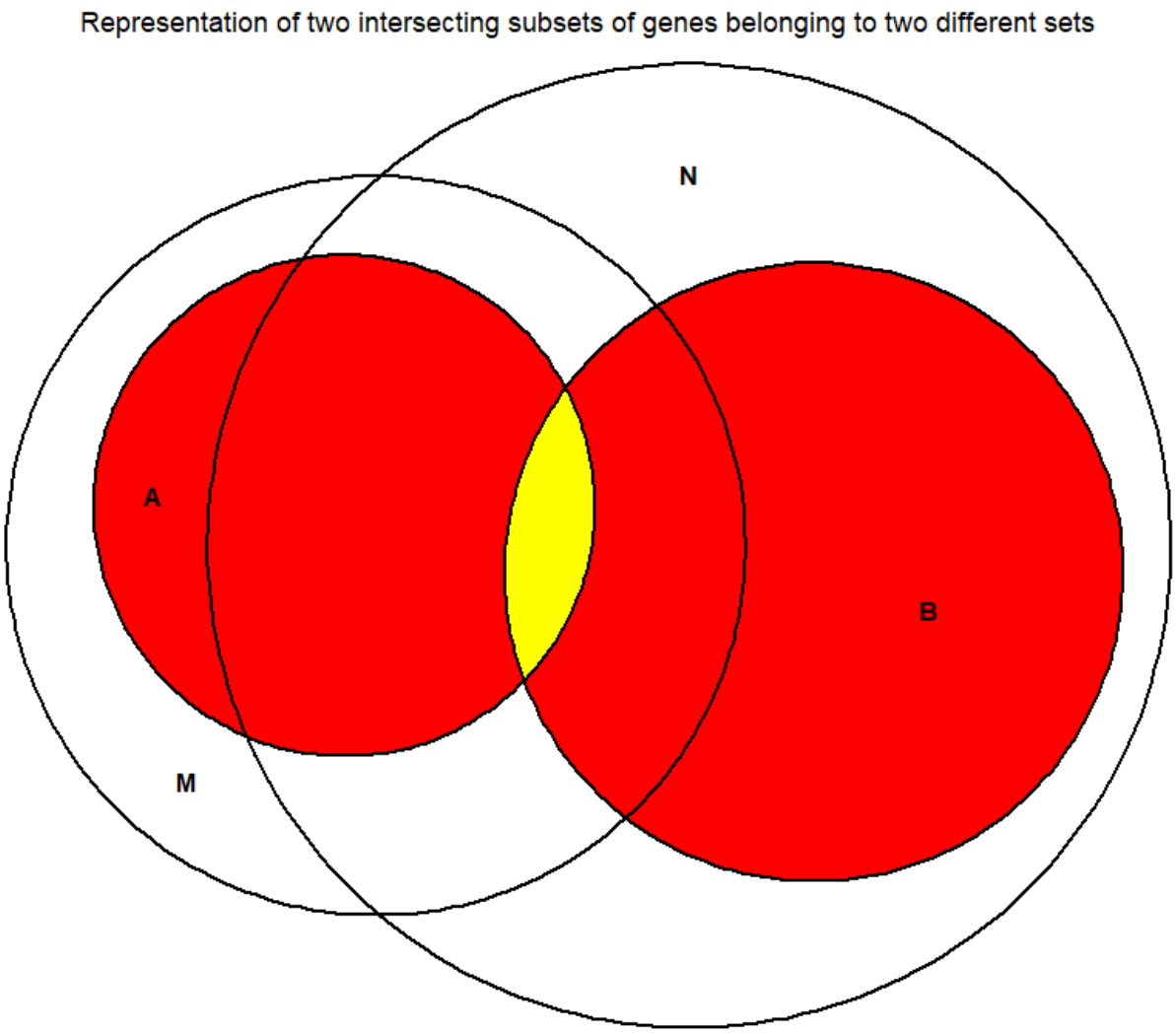


Figure 2.5. The overlap of two subsets taken from different sets of genes.

To assist the presentation of the calculation of the p-value that will be used to assess statistical significance for this scenario, we employ the following notations for the quantities listed below – all known in this setting:

- a = |A|
- b = |B|

- m = |M|
- n = |N|
- c = |A ⋂ N|
- d = |B ⋂ M|
- p = |M ⋂ N|
- k = |A ⋂ B|

Now we proceed with the determination of the formula for the p-value:

- The probability that two sets, *A* and *B* of fixed cardinalities and of fixed cardinalities of their respective intersections with sets M and N intersect in exactly *k* elements is the ratio of the number of ways to choose two sets A and B meeting the conditions listed above and the number of ways to choose two sets A and B meeting all the conditions related to the cardinality of their intersection with sets M and N, but with no restrictions about the cardinality of their intersection with each other.
- The numerator of the fraction described above can be computed as follows:

$$\binom{|M \cap N|}{|A \cap N|}\binom{|M \backslash N|}{|A \backslash N|}\binom{|A \cap N|}{|A \cap B|}\binom{|N \cap M \backslash A|}{|B \cap M \backslash A|}\binom{|N \backslash M|}{|B \backslash M|}$$

  First, the set *A* is selected while heeding the intersection requirements with the sets M and N; the choice of A is thus divided into two parts, picking the elements of A that are shared with N, out of $|M \cap N|$ available elements, then picking the elements of A that are not found in N, out of $|M/N|$ available elements. For the choice of B, first the elements shared with A are chosen out of $|A \cap N|$ available elements, then the elements shared with M but not with A, that is, the $B \cap M \backslash A$ subset, out of $|N \cap M \backslash A|$ available elements, then finally the elements not in M are chosen, out of |N\M| available elements.

- The denominator of the fraction equals:

$$\binom{|M \cap N|}{|A \cap N|}\binom{|M \backslash N|}{|A \backslash N|}\binom{|N \cap M|}{|B \cap M|}\binom{|N \backslash M|}{|B \backslash M|}$$

This is simply the number of ways to choose the sets *A* and *B* within the restrictions imposed by their respective intersections with *M* and *N*, but with no conditions about their own intersection.

- After cancelling out the factors shared between the numerator and the denominator, the probability has the following expression:

$$\frac{\binom{|A \cap N|}{|A \cap B|}\binom{|N \cap M \backslash A|}{|B \cap M \backslash A|}}{\binom{|M \cap N|}{|B \cap M|}}$$

  - This can be rewritten as:

$$\frac{\binom{|A \cap N|}{|A \cap B|}\binom{|M \cap N| - |A \cap N|}{|M \cap B| - |A \cap B|}}{\binom{|M \cap N|}{|B \cap M|}}$$

- Thus, we have arrived at an expression involving only known quantities. Per the earlier definitions, the probability equals:

$$\frac{\binom{c}{k}\binom{p-c}{d-k}}{\binom{p}{d}}$$

- Therefore, the sought probability is:

$$P(|A \cap B| = k) = \frac{\binom{|A \cap N|}{k}\binom{|M \cap N| - |A \cap N|}{|M \cap B| - k}}{\binom{|M \cap N|}{|B \cap M|}}$$

- Correspondingly, the p-value has the expression listed below, where the upper bound for the sum was determined using the observation that the maximum number of elements than *A* and *B* can share in this setting cannot be greater than the number of elements shared by A and *N* (because *B* is included in *N*), and symmetrically it also cannot be greater than the number of elements shared by *B* and *M*, with equality attainable when *A* = *N* or *B* = *M*.

$$P(|A \cap B| \geq k|$$

$$= \binom{|M \cap N|}{|B \cap M|}^{-1} \sum_{i=k}^{\min(|A \cap N|,\ |B \cap M|)} \binom{|A \cap N|}{i} \binom{|M \cap N| - |A \cap N|}{|M \cap B| - i}$$

- As previously, the correctness of the implementation was assessed by assuring that the following identity held through numerical examples:

$$\binom{|M \cap N|}{|B \cap M|}^{-1} \sum_{i=0}^{\min(|A \cap N|,\ |B \cap M|)} \binom{|A \cap N|}{i} \binom{|M \cap N| - |A \cap N|}{|M \cap B| - i} = 1$$

Thus, it was confirmed that the expression does amount to a probability measure, as desired.

Unlike the formulas for p-values developed earlier, this formula is not computable using functions that have a readily available implementation in R. Therefore, an implementation was written for it, as described below:

- Because direct calculation of the factorials involved in the expression listed above was prohibitive in terms of variable storage as the resulting numbers were far greater than the maximum value R allocates for integers (2^{32} - 1), the following approach to compute the expression was pursued:

- Factorials were not calculated explicitly, but instead their prime factors and their corresponding exponents were stored as vectors:
  - First, a list of all the prime numbers less than equal the number of genes found in **both** the A13A and the patient datasets was generated using the primes package from CRAN.
    - This choice was driven by the limiting factor in the p-value formula listed above, $|M \cap N|$. Because no factorials higher than $|M \cap N|!$ were involved in the formula, $|M \cap N|$ served as a natural bound for the list of primes that was generated.
  - Afterwards, the prime factor decomposition of each factorial was performed using Legendre's formula for each prime p[220]:

$$v_p(n!) = \sum_{i=1}^{\infty} \left\lfloor \frac{n}{p^i} \right\rfloor$$

    Here, $v_p(n!)$ denotes the exponent of p in n!, while the $\lfloor x \rfloor$ operator denotes the floor function – the largest integer less than *x*.
- Then, the list of exponents in each n choose k term was calculated by vector addition and a function F implementing the formula listed above: F(n) – F(k) –F(n – k). Using the list of exponents thus obtained and the ordered list of primes below N, the prime factors of each n choose k factor was then computed.
- The products of n choose k factors were likewise calculated, as arrays of prime factors and their exponents.
- The probability terms were then summed up to obtain the desired p-value

Because this procedure was used for FindMarkers outputs coming from different datasets, and thus having an order imposed by the available average $\log_2$ fold changes, the calculation applied the exact strategy as in **Setting G.II** above, with a single modification: for each filtering choice of

the two marker sets, the p-value used to evaluate the overlap followed the formula above corresponding to subsets selected from two different sets, rather than the hypergeometric p-value used in **Setting G.II**. For clarity, this scenario will be referred to as **Setting TSG**.

The calculation of the p-value in this scenario was computationally expensive. Therefore, the function used to calculate the p-value of the overlap of markers of clusters from the two scRNA-sequencing datasets was subjected to parallelization by employing the parLapply, makeCluster, and clusterExport function from the parallel package and run on 60 cores.

Finally, to visualize significant overlaps, alluvial plots were made using ggalluvial[221]. Connecting curves between the categories involved in the overlaps were distinguished by their thickness, with thicker connecting curves corresponding to lower p-values. The same visualization approach was used to represent the significant overlaps of enriched GO terms (**Section 2.3.9**).

2.3.8 The evaluation of the overlaps of sets of cells

The statistical significance of the overlap of subsets *A* and *B* composed of cells from the same set was assessed following the same principles as those described in the previous section, using a hypergeometric p-value computed via phyper.

Two types of problems involved assessments of the statistical significance of cell sets overlaps: the differential representation of cells from one Seurat grouping in another Seurat grouping (e.g. from an experimental condition in a cluster), and the representation of ranked cells (e.g. using ORIGINS activity, see **Section 2.3.5**) among Seurat clusters or conditions. The methodology for both cases is detailed upon below:

- **Setting C.I:** Two subsets of the same set of cells, both unordered.
    - **Example:**
        - The list of cells from one cluster and the list of cells from one dataset corresponding to one treatment condition.

- **Procedure**:
  - A one-time hypergeometric p-value is calculated using the methods described in **Section 2.3.7**, with subsets of cells in lieu of subsets of genes. For the overlap of one pair of two lists, no Benjamini-Yekutieli correction is performed as there is no multiple testing involved, both cell sets being unordered and thus not subjected to repeated filtering choices that would result in different values of overlap.
- **Adjustment for multiple testing:**
  - The comparison of multiple groups (different cluster – condition pairs), however, does require a correction for multiple testing, and Benjamini-Yekutieli was selected again for the task. As before, the results are aggregated into an array equaling the product of the cardinalities of the two groupings (e.g. conditions and clusters), but this time the p-values are **singly-corrected**.

- **Setting C.II:** Two subsets of the same set of cells, both having a comparable ranking available.
  - **Example:**
    - The list of cells from each cluster, ranked by their ORIGINS activity score.
  - **Procedure**:
    - The ranking values of interest – in the case of activity scores, the scores between the top ~0.2% and ~16% of all the cells, as detailed in **Section 2.3.7**, are sorted and iteratively used as filters for both cells sets.
    - Instead of filtering out values too high to be shared by all the groupings of interest, as in the analysis of the overlap of genes with available average $\log_2$ fold-change values (**Setting G.II**), p-values of the overlaps

corresponding to filters where no cells exist in some of the groupings are simply assigned a value of 1.

- A hypergeometric p-value of overlap is calculated in the other cases, for each filtering.
- The list of p-values for all the filters corresponding each overlap is subjected to the Benjamini-Yekutieli correction and the median is taken to represent the p-value of the overlap.

- **Further adjustment for multiple testing:**
  - The multiple testing induced by the groupings results in a Benjamini-Yekutieli adjustment of the p-values, as earlier.

2.3.9 The evaluation of the overlaps of enriched GO terms

The list of GO terms enriched in each grouping of interest was generated using the methods described in **Section 2.3.1**. For each grouping, the full list of doubly Bonferroni-corrected grouping markers was used as input.

The p.adjust column of each of the DOSE enrichResult object, storing the corrected p-value corresponding to the significance assessment of each GO term enrichment, was subjected to a Bonferroni adjustment for the testing of multiple groupings, akin to the procedure used for the identification of grouping markers in **Section 2.3.6**. GO terms with doubly adjusted p-values higher than 0.05 were subsequently filtered out.

In order to compute the overlap of GO terms enriched in different groupings from the same scRNA-sequencing dataset by the methods developed previously, the set of all GO terms appearing in each scRNA dataset was identified by performing enrichment analysis on the whole dataset, with no corrections or filtering. Thus, we have obtained a situation analogous with **Section 2.3.7, Setting G.II**, and thus suitable for the same overlap assessment method, through the following bijections:

- The GO terms enriched in each grouping ↔ The genes differentially expressed in each grouping;
- All the GO terms associated with any gene found in the dataset ↔ All the genes found in the data;
- The order imposed by the doubly-adjusted p-values in each enrichment result ↔ The order imposed by the average $\log_2$ fold changes in each FindMarkers result.

Therefore, the overlap assessment was performed in the same manner as in the case of the aforementioned setting, while adjusting for the fact that the ordering used in **Setting G.II** (based on average $\log_2$ fold changes) was **decreasing**, while the one used in this setting (based on p-values) is **increasing**, with the appropriate modifications (changes of signs and of extrema) applied to the functions used for the generation of filters and applying them to the marker lists. Thus, the p-values are used iteratively to filter out values **above** them rather than **below**, while the most restrictive filter now involved the **maximum of the minima** of the p-values of the two GO term lists, rather than the minimum of the maxima of average $\log_2$ fold changes of the two markers lists.

The setting described in this section (two ordered sets of GO terms originating as the results of enrichment analysis on the same dataset) will be referred as **Setting EA.II**.

The overlap of GO terms enriched in different groupings from the different scRNA-seq datasets (the A13A and the patient one), thus having different sets of GO terms found in the data, was also performed. The procedure followed the same lines as those described for **Setting EA.1**, with the sole difference lying in the formula used for the calculation of the p-value of the overlap between two filtered sets of GO terms coming from the two datasets, namely the formula found in the final part of **Section 2.3.7**, in lieu of a hypergeometric p-value. Thus, this setting is an analogue of **Setting TSG** in light of the correspondences described above, again with the

appropriate reversal of signs and of extrema detailed upon above. It will be referred further as **Setting TSEA.**

Finally, an analysis of the enrichment of epigenetic GO terms was performed. Epigenetic GO terms were identified from the list of all GO terms associated with the genes of the two datasets using the following pattern:

"histone|chromatin|epigenetic|protein-DNA|packaging|nucleosome|DNA methylation|imprinting|silencing|ncRNA-mediated|conformation|geometric|miRNA|inactivation of X chromosome".

Their differential representation among the enriched GO terms enriched for different grouping marker lists was evaluated following the methods of **Setting G.I**, again with the aforementioned correspondences and reversal of the signs and extrema in the corresponding functions. This setting will be referred further as **Setting EA.I**.

2.3.10 Trajectory analysis

The tradeSeq package[222] was used to perform trajectory-based differential analysis using the Slinsghot trajectory prediction results. The evaluateK function was used to calculate the number of knots to be used as the nknots parameter in the fitGAM function from tradeSeq, which fit a generalized additive model to the Slingshot pseudotime values, using also their respective cell weights. Parallelization of these functions was implemented using the bparam function from BiocParallel[223]. Early differences between lineages were assessed with the earlyDESeq function, and the plotSmoothers and plotGeneCount functions were employed for visualization of the evolution of gene expression within and between lineages.

Next, a weighted average of the pseudotime values obtained for the lineages found by Slingshot was implemented, in order to obtain one unified pseudotime scoring for all the cells. Then, the identification of genes whose expression variation is linked to the variation in the values of the

two numerical measures assessing stemness used in this analysis (ORIGINS activity and Slingshot aggregate pseudotime) was performed using the following approach:

- For both metrics of stemness, **median** and **mean** scores were calculated over each gene, over all the cells in which the gene was detected. The number of cells in which each gene was detected was also stored for each gene in the same data frame.
- The values obtained for the median and mean were sorted, respectively, based on the order of decreasing inferred stemness scores (that is, decreasingly for ORIGINS activity scores, increasingly for Slingshot pseudotime values).
- Then, the genes varying in accord with the stemness metric were identified, using combined information coming from the sorted median and mean scores, respectively, and the number of cells in which each gene was detected. An initial cutoff for the number of cells in which a gene needed to be detected to be taken into consideration was set based on visual inspection of the data frame constructed as described above. Afterwards, **genes detected in a greater number of cells than the preceding selected gene were selected**, with the gene list being traversed in the order imposed by the sorted median and mean scores.
- Thus, four gene sets were obtained, linked with the variation of the median and mean of the two numerical stemness predictors. These gene sets will be referred further as **characteristic gene sets** of their corresponding central tendency measure and stemness predictor.

The approach developed here constructs the upper boundary of the convex hull - the smallest set that contains a set of points while also comprising the whole line segment that connects any two points belonging to it - of the 2D figure with the number of cells expressing each gene on the X axis, and the aggregate stemness scores for each gene over the cells in which the gene is

expressed on the Y axis, minus points below the initial cut-off for the number of cells, omitted because genes expressed in very few cells may link with extreme stemness scores just by chance.

The genes thus identified are genes whose variation has the same directionality as the variation of stemness scores. Conversely, selecting genes in the order imposed by sorting of the central tendency measure of stemness in the direction of increasing stemness in the same fashion described above finds the lower boundary of the convex hull of the figure described above. For better visualization, the coordinates were flipped in the figure displaying the identification of characteristic gene sets. While other methods to assess the dependence of the variation of the stemness metrices (for instance, based upon correlation methods or obtained by training a generalized additive model) can also provide an additional piece of information, namely a numerical value corresponding to the strength of the link between the stemness metric and the expression of the gene of interest, the approach developed here has the advantage of not requiring an arbitrary cut-off above which genes are deemed to be significantly linked to the stemness measure.

In order to assess whether the three assays linked to developmental potential depend upon similar sets of genes and biological processes, the characteristic gene sets were then assessed for gene overlaps, and their biological associations were identified using StringDB and enrichment analysis.

Finally, RNA velocity analysis was performed using velocyto.R, using the spliced and unspliced count matrices obtained as described in **Section 2.2**.

2.3.11 Analysis of cell-cell communication

The analysis of cell-cell communication used two R packages: SingleCellSignalR[224] and CellChat[225].

With SingleCellSignalR, cell-cell interactions were discovered using the cell_signaling function and visualised using the mv_interactions function.

With CellChat, cell-cell interactions were found using the identifyOverExpressedInteractions function, having previously assigned the variable features of the Seurat object as the variable features of the CellChat object (the “features” slot of the var.features assay) in lieu of using variable features found using CellChat’s native feature selection method. Communication patterns were identified using the identifyCommunicationPatterns function, with the number of expected patterns having been chosen through the visual inspection of silhouette and cophenetic plots done using selectk.

CHAPTER 3

CHARACTERIZATION OF CANCER STEM CELLS IN SINGLE-CELL RNA-SEQUENCING DATA IN A13A PDAC CELLS

3 Characterization of cancer stem cells in single-cell RNA-sequencing data in A13A PDAC cells

3.1 Introduction

This chapter describes the results of a single-cell RNA-sequencing experiment performed in PDAC cells from the A13A cell line under three different treatment conditions: Activin A, Activin A and I-BRD9, and SB-431542. The treatment duration was 24 h for each condition. The identification and characterization of subpopulations (clusters) in the datasets was performed, with a focus on discerning cell groups ranking high in cancer stemness and on examining the trajectory of differentiation. Subsequently, an assessment of the differential effects of the experimental conditions upon these subpopulations was performed, aiming to assess the effects of inhibiting BRD9, an epigenetic regulator recently suggested to drive tumorigenesis[86] and pluripotency through chromatin remodelling[226], upon cancer stemness. In addition, Activin signalling was evaluated as a possible regulator of cancer stemness, as reported previously in the literature[88],

Section 3.2.1 covers quality control. Normalization, integration, dimensionality reduction and clustering are treated in **Section 3.2.2**. In **Section 3.2.3**, clusters associated with stemness are identified, and a functional characterization of all the clusters is performed in **Section 3.2.4**.

Section 3.2.5 includes trajectory analysis and RNA velocity analysis. Cell-cell communication is studied in **Section 3.2.6**, and the epigenetics processes enriched for the markers of clusters are identified in **Section 3.2.7**.

Beginning with **Section 3.2.8,** the differential effects of the treatment conditions are assessed. In **Section 3.2.8**, the effects of the treatment conditions are evaluated at the global (pseudobulk) level. The intracluster effects of the treatment conditions are examined in **Section 3.2.9**. In order to assess the specificity of the effects of the treatment conditions upon stemness, overlap assessments between the genes and processes characterizing the clusters and the experimental

conditions, respectively, are performed in **Section 3.2.10**. The extents to which stemness-linked gene sets account for the gene overlaps identified in **Section 3.2.10** is determined in **Section 3.2.11**. Finally, a discussion of the results of this chapter is provided in **Section 3.3**.

3.2 Results

3.2.1 Quality control

Genes expressed in very few cells (< 10) were removed, in order to avoid inducing potential bias upon subsequent analysis. Afterwards, the eight steps listed in **Section 2.3.2** were performed for the identification and elimination of low-quality cells: removal of doublets identified using scDblFinder and removal of cells with a very low complexity (novelty) score, a very high percentage of mitochondrial or ribosomal genes, a very low or very high percentage of number of counts or UMIs per cells, or a very low Shannon or Simpson diversity.

100 scDblFinder runs predicted an average of 1102.32 doublets, most closely approximated by a doublet prediction significance cutoff of 47 runs, resulting in 1104 identified doublets (**Figure 3.1**).

The Jaccard similarity scores reported between all the pairs of scDblFinder runs ranged between 0.54 and 0.74, while Jaccard scores (see **Section 2.3.2)** taken with the consensus prediction ranged between 0.66 and 0.81.

The mean of the Jaccard scores between the consensus predictions and each of the scDblFinder runs was 0.76, above the mean of the maxima of all the Jaccard scores between any two pairs of runs (0.72).

The results of the calculations of Jaccard similarity scores between scDblFinder runs and the consensus prediction are illustrated in **Figure 3.2**. Purple up-triangles represent Jaccard scores taken with the consensus prediction, while the maxima, means and minima of the Jaccard similarities between each prediction and the other 99 are displayed with navy squares, medium blue circles and light blue triangles, respectively. Dashed lines represent group averages.

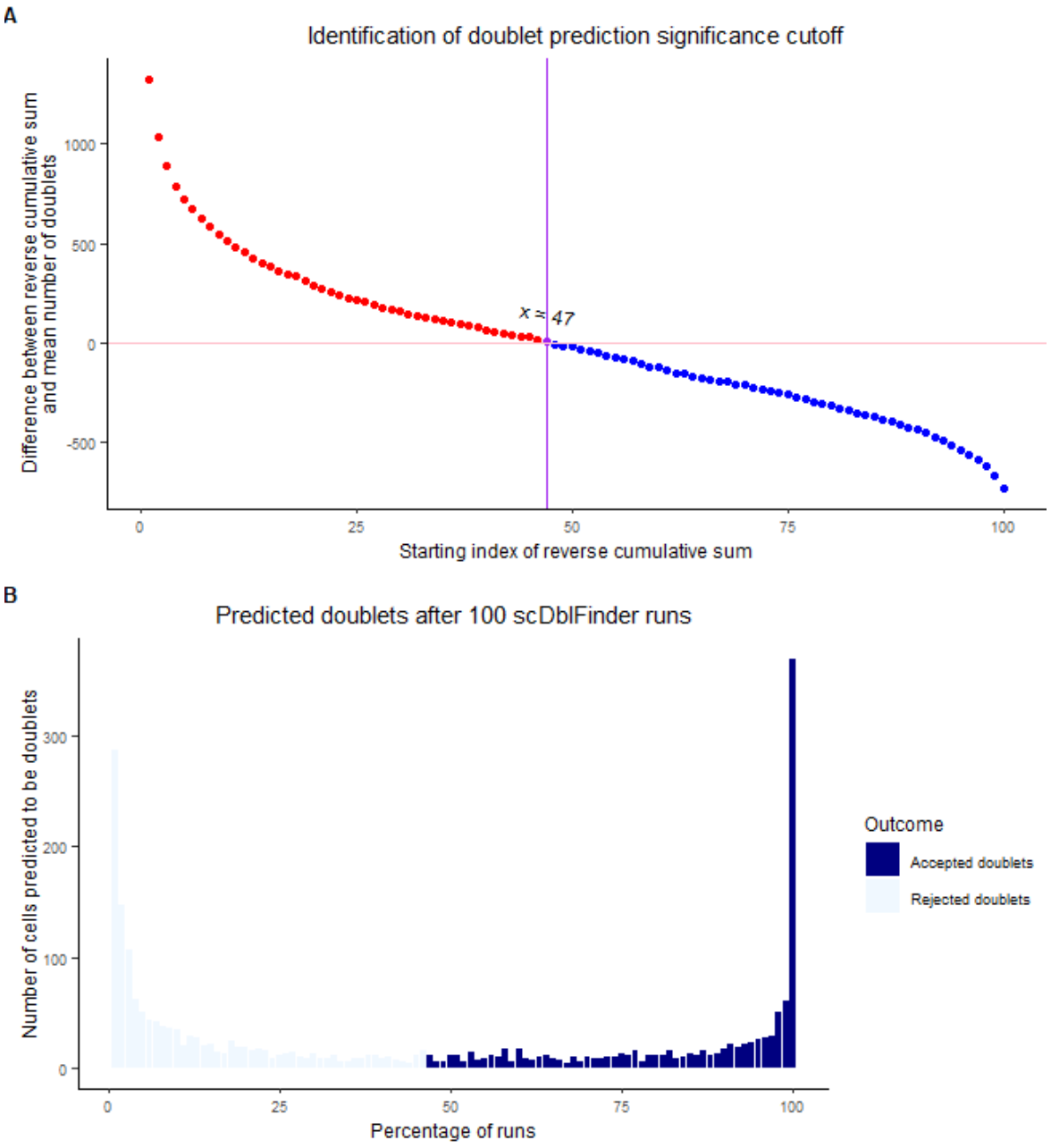


Figure 3.1. **A)** Identification of the doublet prediction cutoff. **B)** The distribution of accepted and rejected doublets.

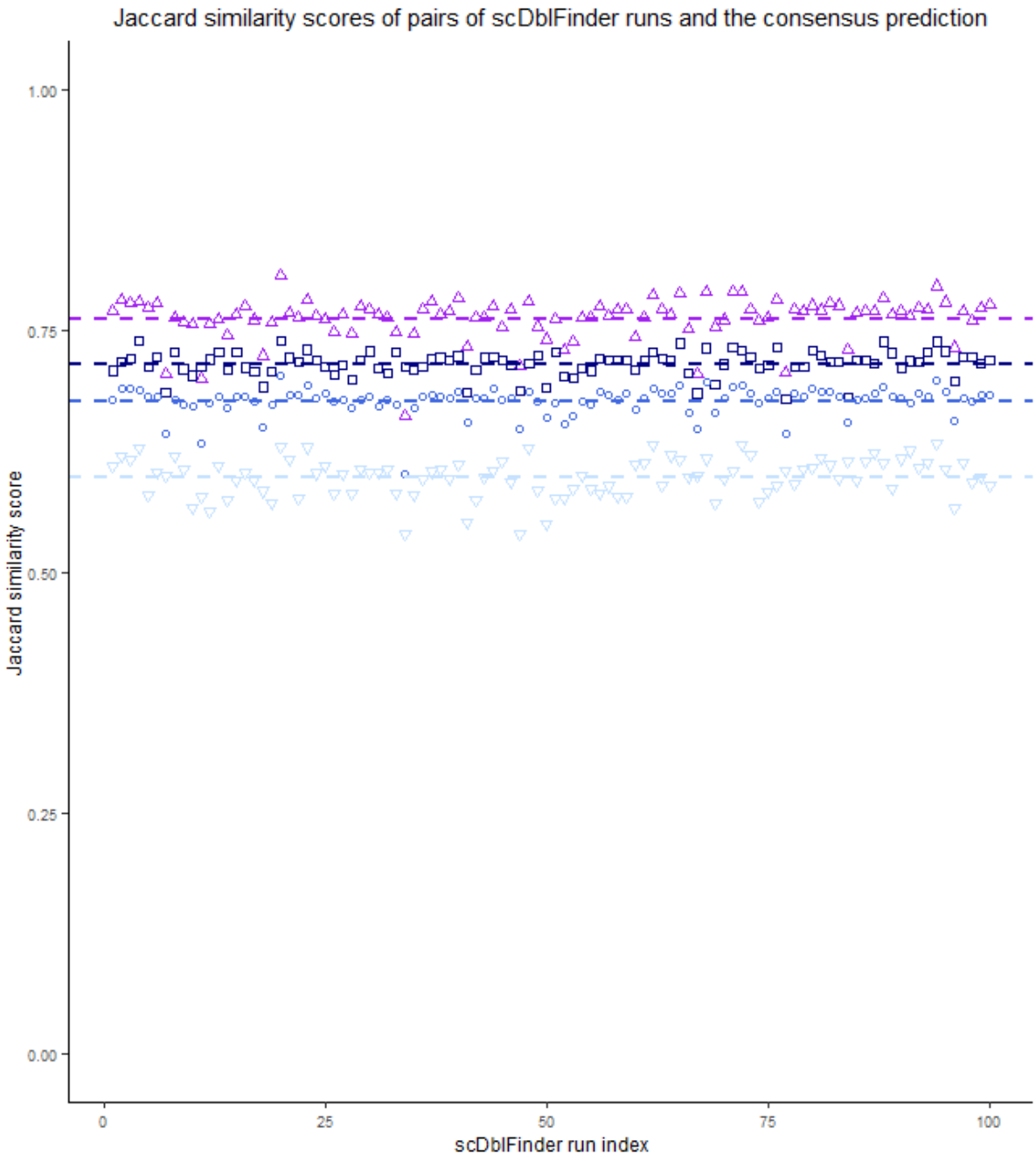


Figure 3.2. Jaccard similarity scores between each prediction and the consensus prediction, and between each prediction and the maxima, means and minima of the other 99.

Next, the cells were filtered based on the other criteria listed in **Section 2.3.2**. **Figure 3.3**, **Figure 3.4** and **Figure 3.5** illustrate the distributions of the filtering variables across experimental conditions. On each plot, the section between the two vertical lines represents cell retained at each filtering step.

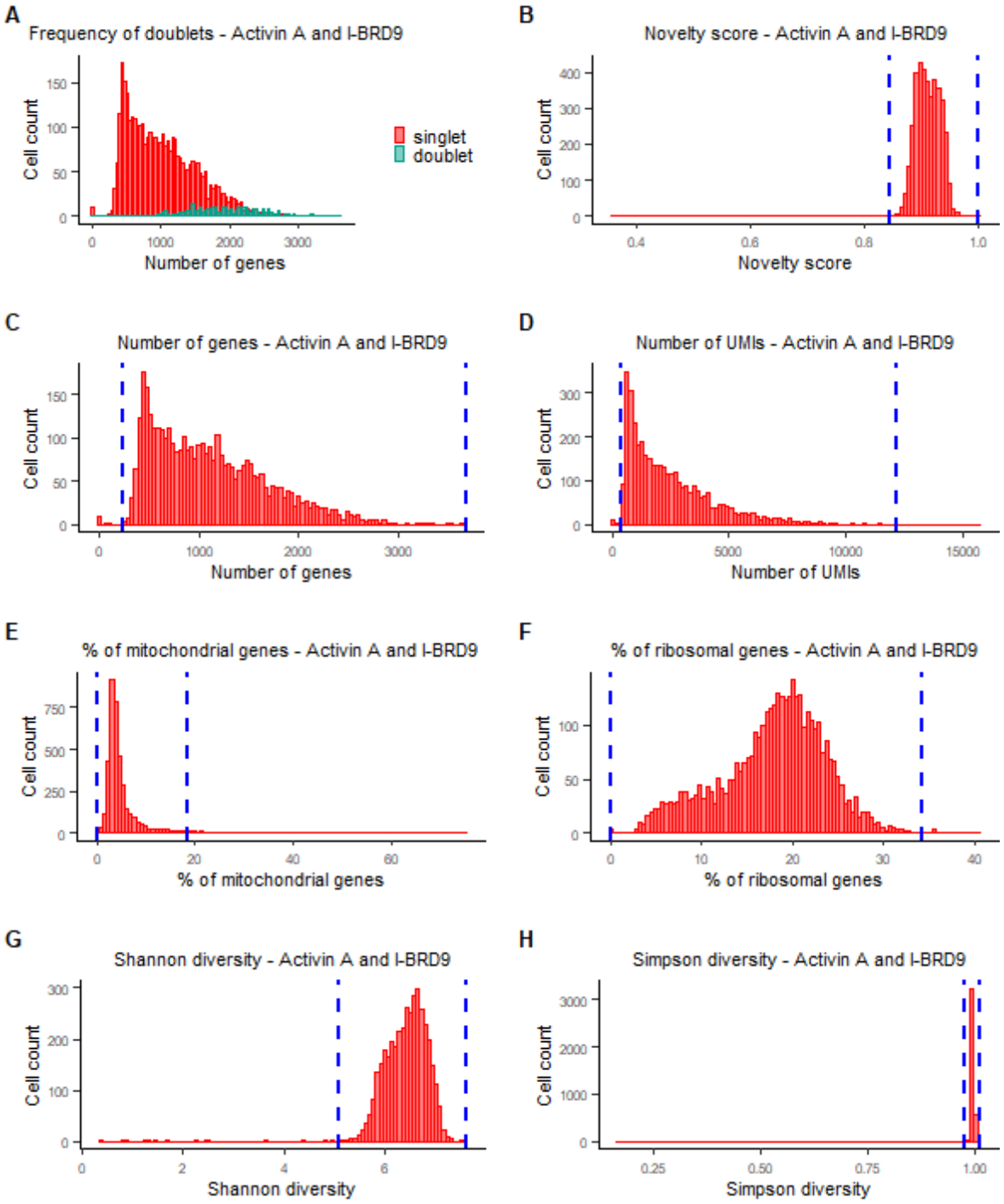


Figure 3.3. Quality control selection criteria for cells in the Activin A and I-BRD9 condition: **A)** Singlet status as predicted by scDblFinder; **B)** Novelty score (log10 (nFeature_RNA) / log10 (nCount_RNA)) above 0.85; **C)** Number of detected genes (nFeature_RNA) between 250 and 3670; **D)** Number of UMIs (nCount_RNA) between 400 and 12200; **E)** Percentage of mitochondrial genes below 18.6%; **F)** Percentage of ribosomal genes below 34.3%; **G)** Shannon diversity above 5.1; **H)** Simpson diversity above 0.98.

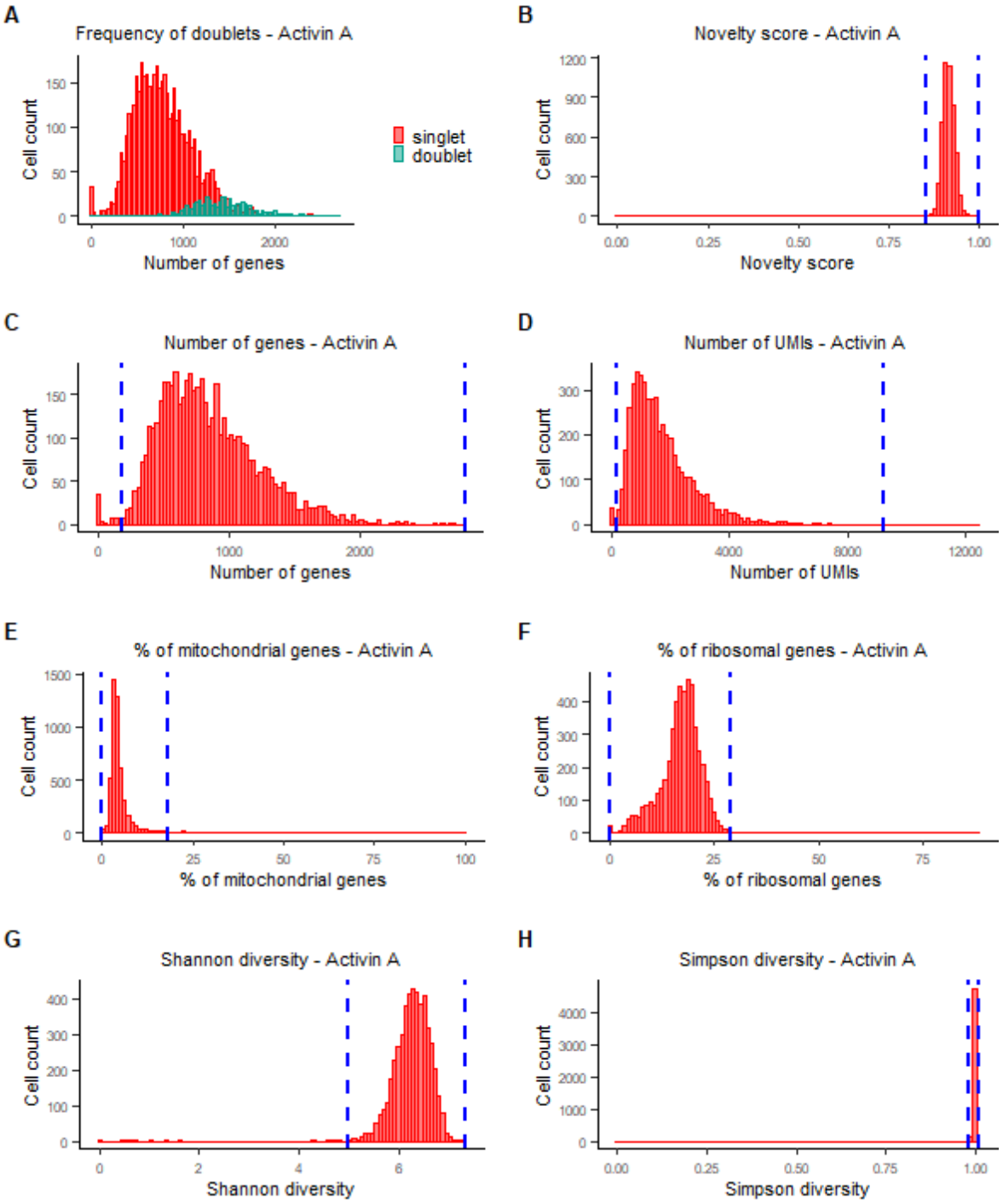


Figure 3.4. Quality control selection criteria for cells in the Activin condition: **A)** Singlet status as predicted by scDblFinder; **B)** Novelty score (log10 (nFeature_RNA) / log10 (nCount_RNA)) above 0.86; **C)** Number of detected genes (nFeature_RNA) between 190 and 2800; **D)** Number of UMIs (nCount_RNA) between 220 and 9200; **E)** Percentage of mitochondrial genes below 18%; **F)** Percentage of ribosomal genes below 29%; **G)** Shannon diversity above 4.98; **H)** Simpson diversity above 0.98.

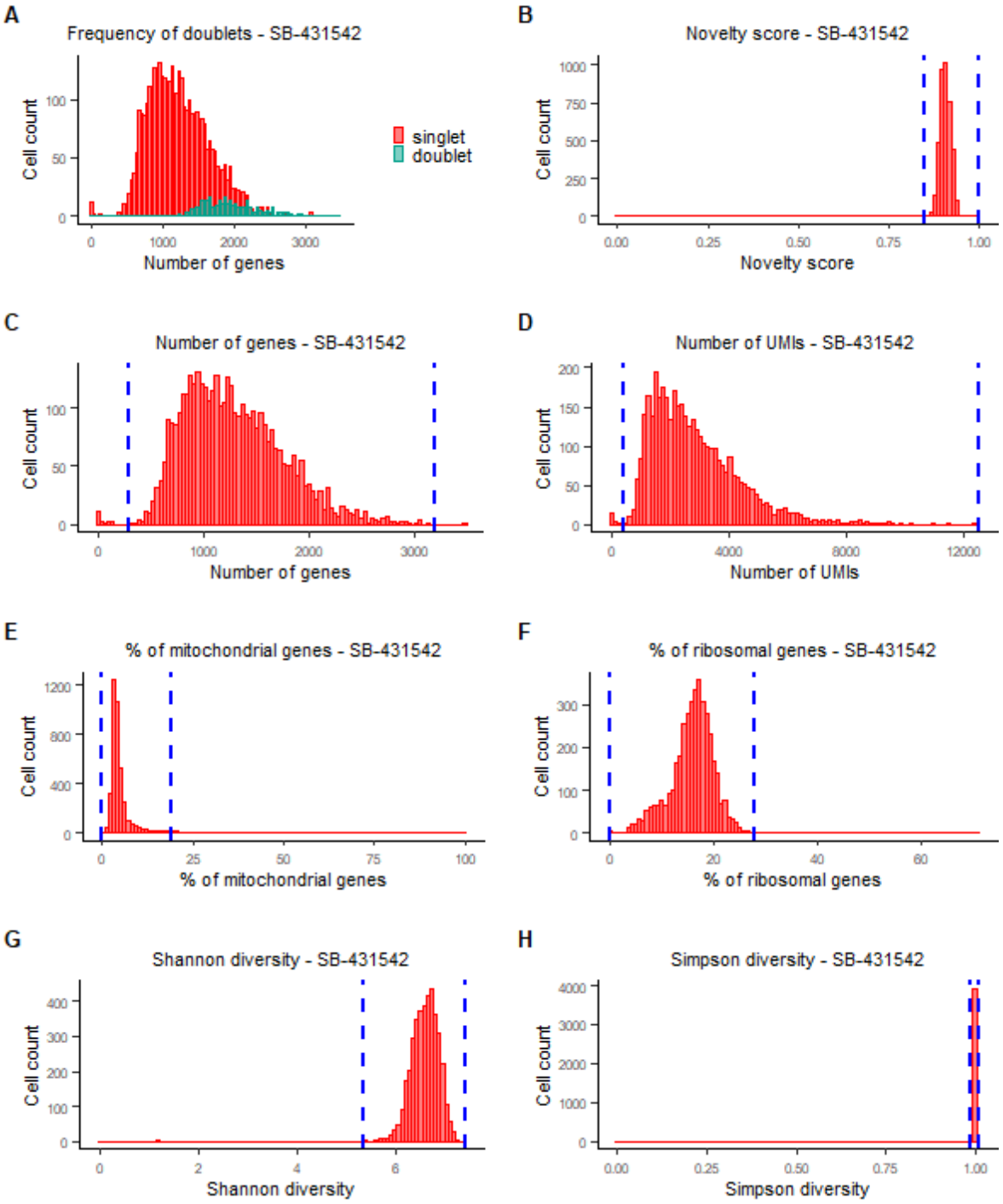


Figure 3.5. Quality control selection criteria for cells in the SB-431542 condition: **A)** Singlet status as predicted by scDblFinder; **B)** Novelty score (log10 (nFeature_RNA) / log10 (nCount_RNA)) above 0.85; **C)** Number of detected genes (nFeature_RNA) between 300 and 3200; **D)** Number of UMIs (nCount_RNA) between 430 and 12500; **E)** Percentage of mitochondrial genes below 18.8%; **F)** Percentage of ribosomal genes below 27.8%; **G)** Shannon diversity above 5.35; **H)** Simpson diversity above 0.99.

The cut-offs employed for each of the criteria are summarized in **Table 3.1**:

Cell filtering criterion/Experimental condition	Activin A and I-BRD9	Activin A	SB-431542
Novelty score	Above 0.85	Above 0.86	Above 0.85
Number of genes	Between 250 and 3670	Between 190 and 2800	Between 300 and 3200
Number of UMIs	Between 400 and 12200	Between 220 and 9200	Between 430 and 12500
Percentage of mitochondrial genes	Below 18.6	Below 18	Below 18.8
Percentage of ribosomal genes	Below 34.3	Below 29	Below 27.8
Shannon diversity	Above 5.1	Above 4.98	Above 5.35
Simpson diversity	Above 0.98	Above 0.98	Above 0.99

Table 3.1. A summary of the quality control selection criteria used for cells in all three experimental conditions (Activin A and I-BRD9, Activin A, SB-431542).

The numbers of cells that were removed at each step of the filtering are illustrated in **Figure 3.6**. Doublet filtering removed a total of 1104 cells and filtering by the percentage of mitochondrial genes removed 230 cells. The other cell filtering steps removed very few cells.

Once all the filtering was complete, the merged Seurat object contained 3400 cells from the Activin A and I-BRD9 condition, 4290 cells from the Activin condition and 3580 cells from the SB-431542 condition.

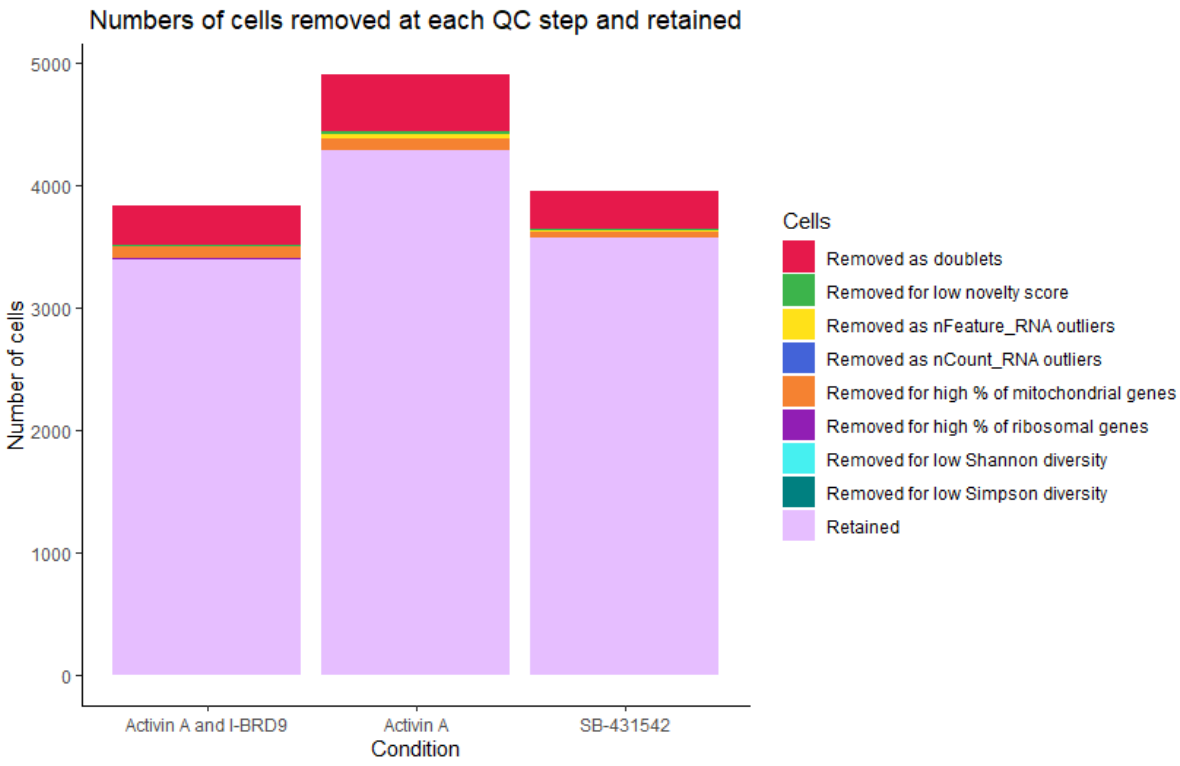


Figure 3.6. The number of cells removed at each step of the filtering and retained at the end.

After the filtering of low-quality cells, 196 genes became expressed in fewer than 10 cells and were consequently removed from the data.

The next section will discuss normalization and scaling using SCTransform, dimensionality reduction, integration and clustering.

3.2.2 Clustering

Next, clustering was performed, in order to group the cells into subpopulations of transcriptionally similar cells.

The Seurat object was split by experimental condition, and SCTransform was applied individually on the Seurat objects obtained after splitting, regressing the cell cycle phase, the percentage of mitochondrial genes and the percentage of ribosomal genes in the process, then the objects were again merged.

With the gene counts adjusted by SCTransform, 117 genes now appearing in fewer than 10 cells were removed from the data. 10964 genes were thus retained for further analysis. Then, the Seurat object was integrated with Harmony on the SCT assay, thus allowing the cells to be grouped by their functional type rather than by the experimental condition, and an UMAP reduction was performed using the first 17 Harmony-generated dimensions. The result of the integration and UMAP are displayed in **Figure 3.7**:

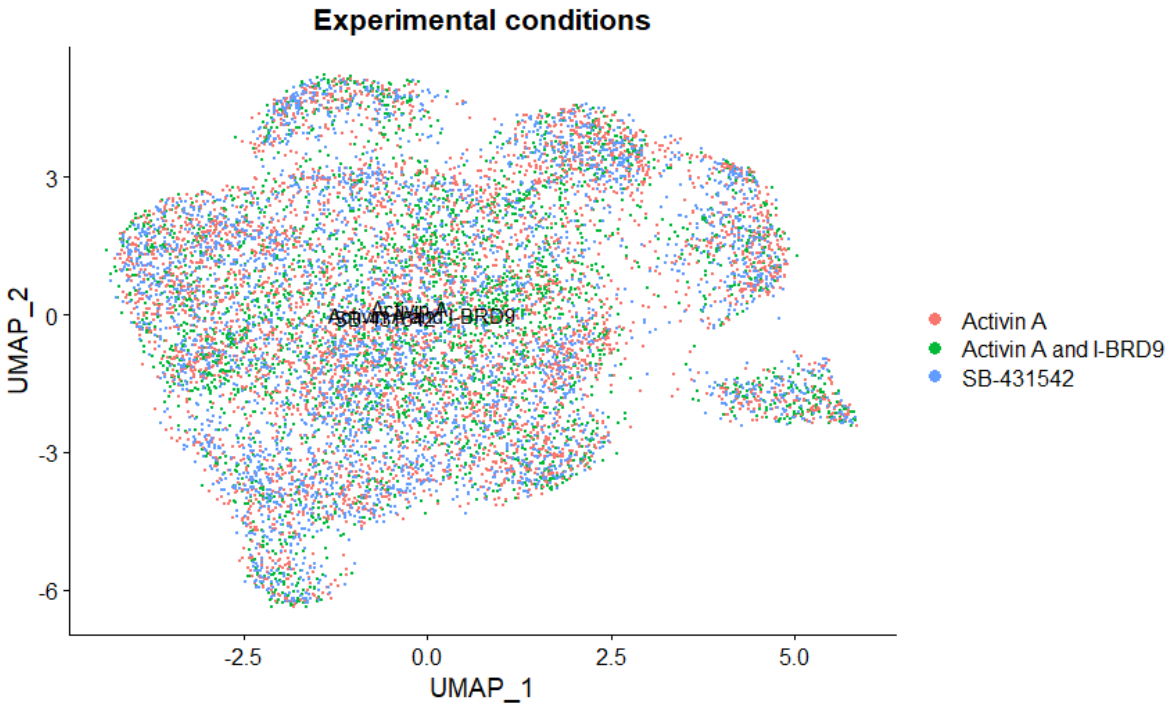


Figure 3.7. UMAP plot showing cells grouped by the experimental condition of origin, after regression and integration.

Initial clustering was then performed, with the goal of obtaining clusters as clearly delineated by distinctive markers as possible. Different resolutions were picked iteratively, and neighbouring clusters were merged based on the same principle of increasing the distinctiveness of cluster markers, by improving upon measures such as the p-value and the logarithm of the fold-change. A resolution of 2.5 was found to be optimal, resulting in 54 initial clusters (**Figure 3.8**). We mention here that other methods to assist in the determination of clusters also exist, such as clustree, which visualises clustering trees in order to study the relationship between clusters at different resolutions[227].

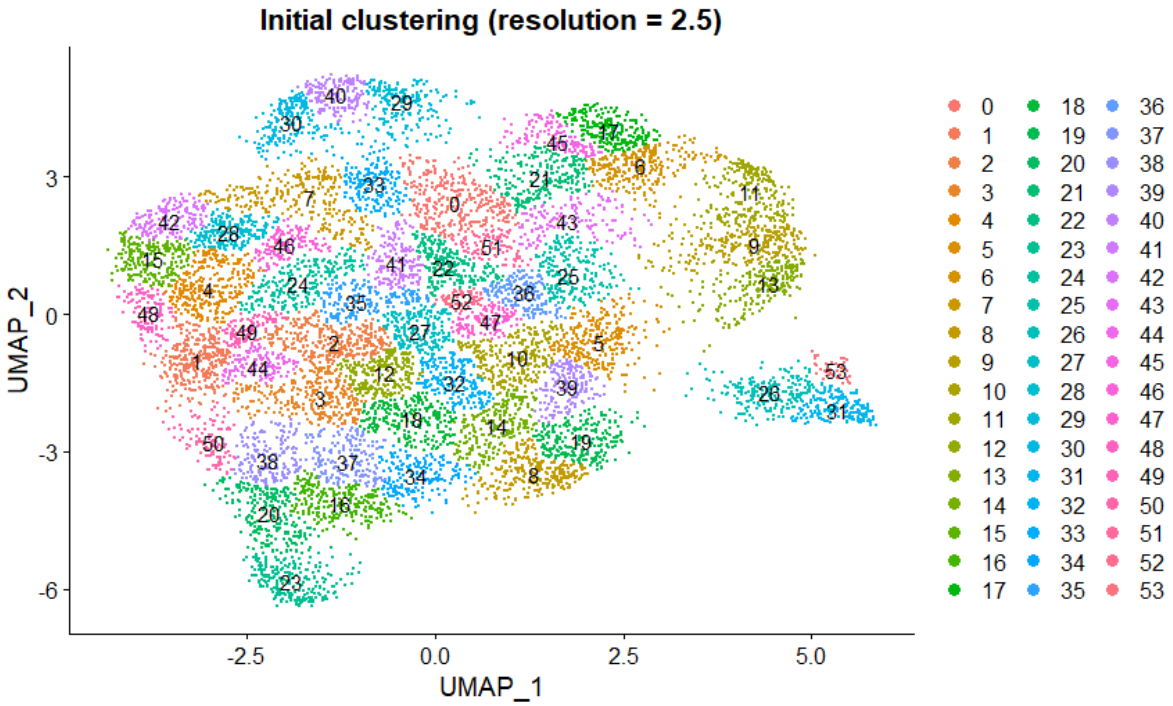


Figure 3.8. The initial configuration of clusters, prior to the mergers based on the distribution of marker genes, was obtained with a resolution set at 2.5 and contained 57 clusters.

After cluster mergers based on marker expression, 11 final clusters were obtained (**Figure 3.9**).

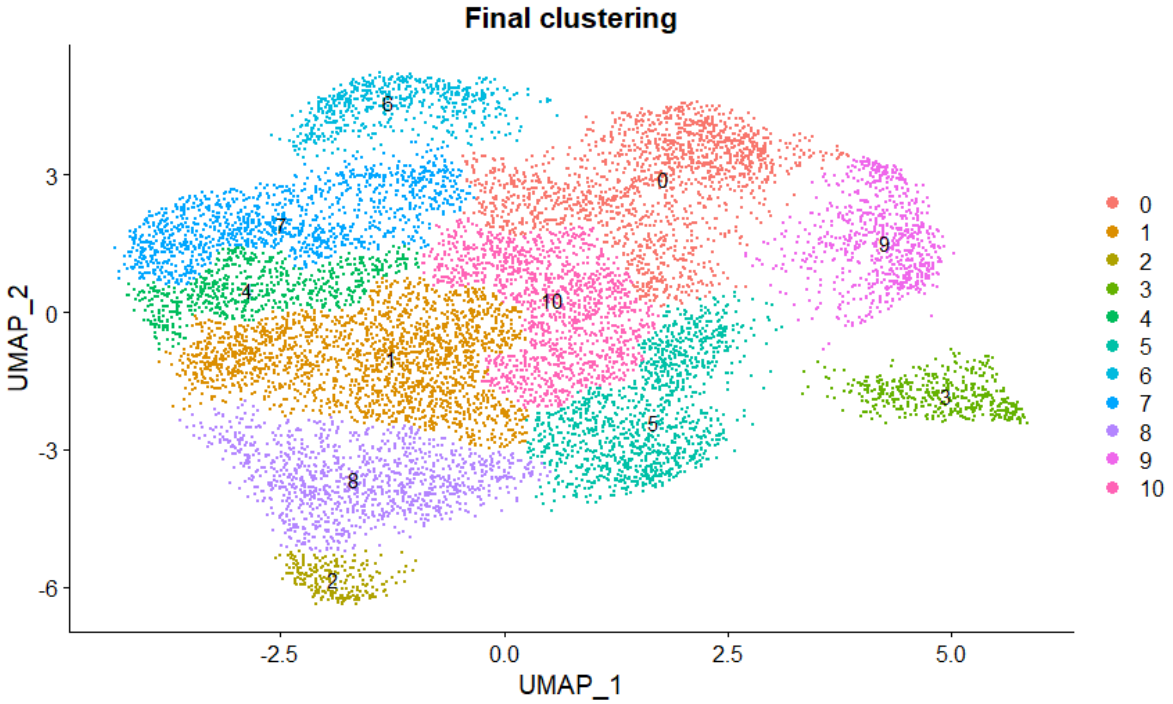


Figure 3.9. The final configuration of clusters.

The details of the construction of the final clusters are provided in **Table 3.2**:

Final cluster	**Original clusters**	**Markers**
0	0, 43, 21, 45, 6, 17 and 25	*AKAP12*; *NDRG1*
1	1, 44, 49, 2, 3, 12, 35, 27 and 18	*PTTG1*; *CENPF*
2	23	*AKR1B10*; *CEACAM6*
3	26, 53 and 31	*IFI6*; *MSMB*
4	4, 48 and 24	*H4C3*; *NCL*
5	5, 19, 14, 8 and 39	*HERPUD1*; *PDIA4*
6	30, 40 and 29	*AURKA*; *DLGAP5*
7	7, 15, 42, 28, 46 and 33	*H4C3*; *H1-4*
8	50, 38, 37, 20, 16 and 34	*AKR1C1*; *AKR1C2*
9	9, 11 and 13	*CGA*; *FST*
10	10, 52, 22, 47, 41, 36, 51 and 32	*DSP*; *DST*

Table 3.2. 11 final clusters were constructed from the initial 57 clusters via merging and renumbering, based on the expression of shared markers that displayed a localized expression in distinct regions of the UMAP plot.

On the following pages, the expression of 12 genes important to the delineation of clusters, namely *AKAP12*, *CEACAM6* and *AKR1C1* (**Figure 3.10**), *PTTG1*, *SYNE2* and *DST* (**Figure 3.11**); *IFI6*, *H4C3* and *NDC80* (**Figure 3.12**), *PDIA4*, *AURKA* and *CGA* (**Figure 3.13**), is illustrated through feature and violin plots.

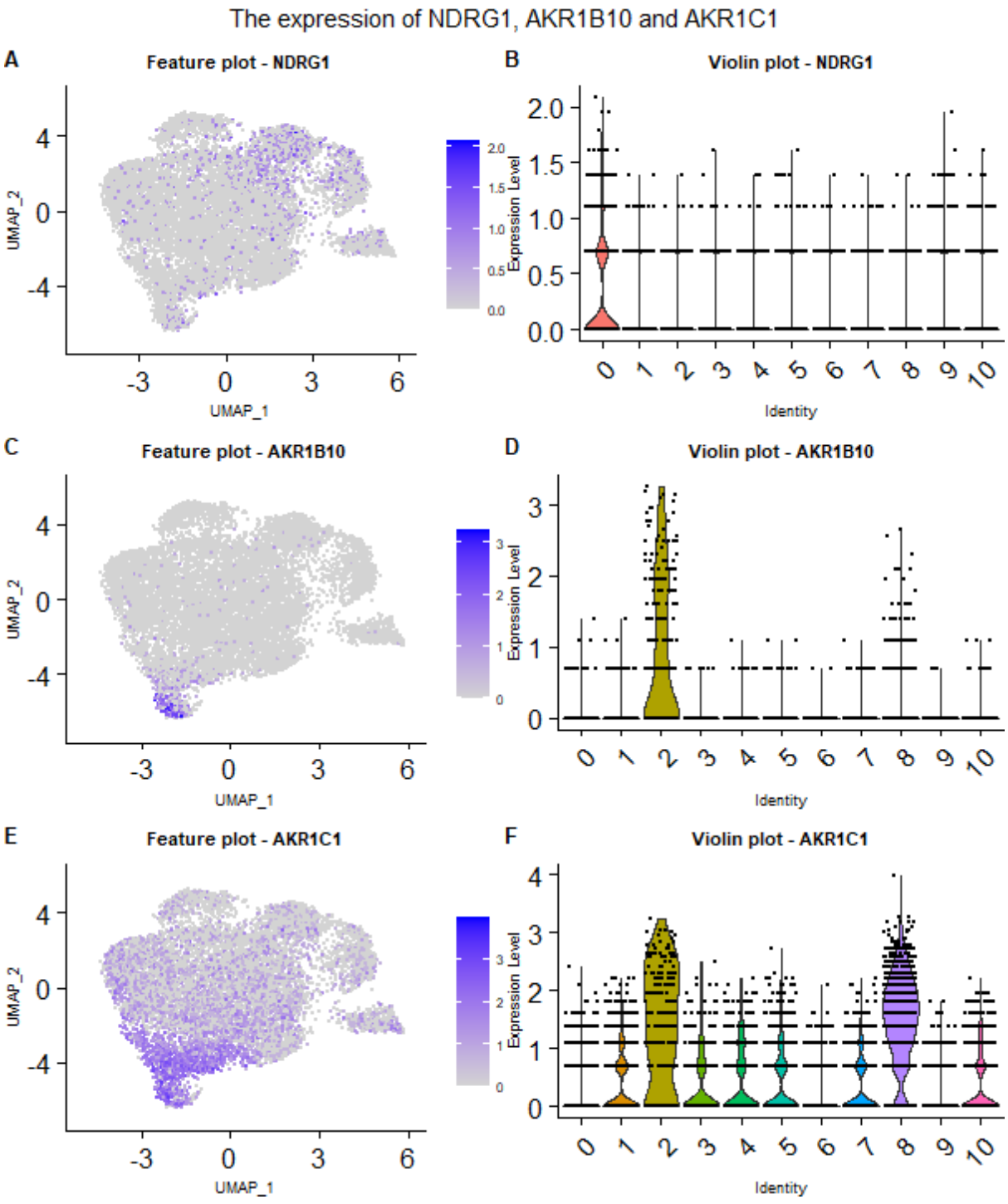


Figure 3.10. The contour of *NDRG1* delineates Cluster 0 **(A, B)**. The expression pattern of *AKR1B10* establishes Cluster 2 **(C, D)**. Cluster 8 shows an overexpression of markers also overexpressed in Cluster 2, such as *AKR1C1* **(E, F)**, but not *AKR1B10*.

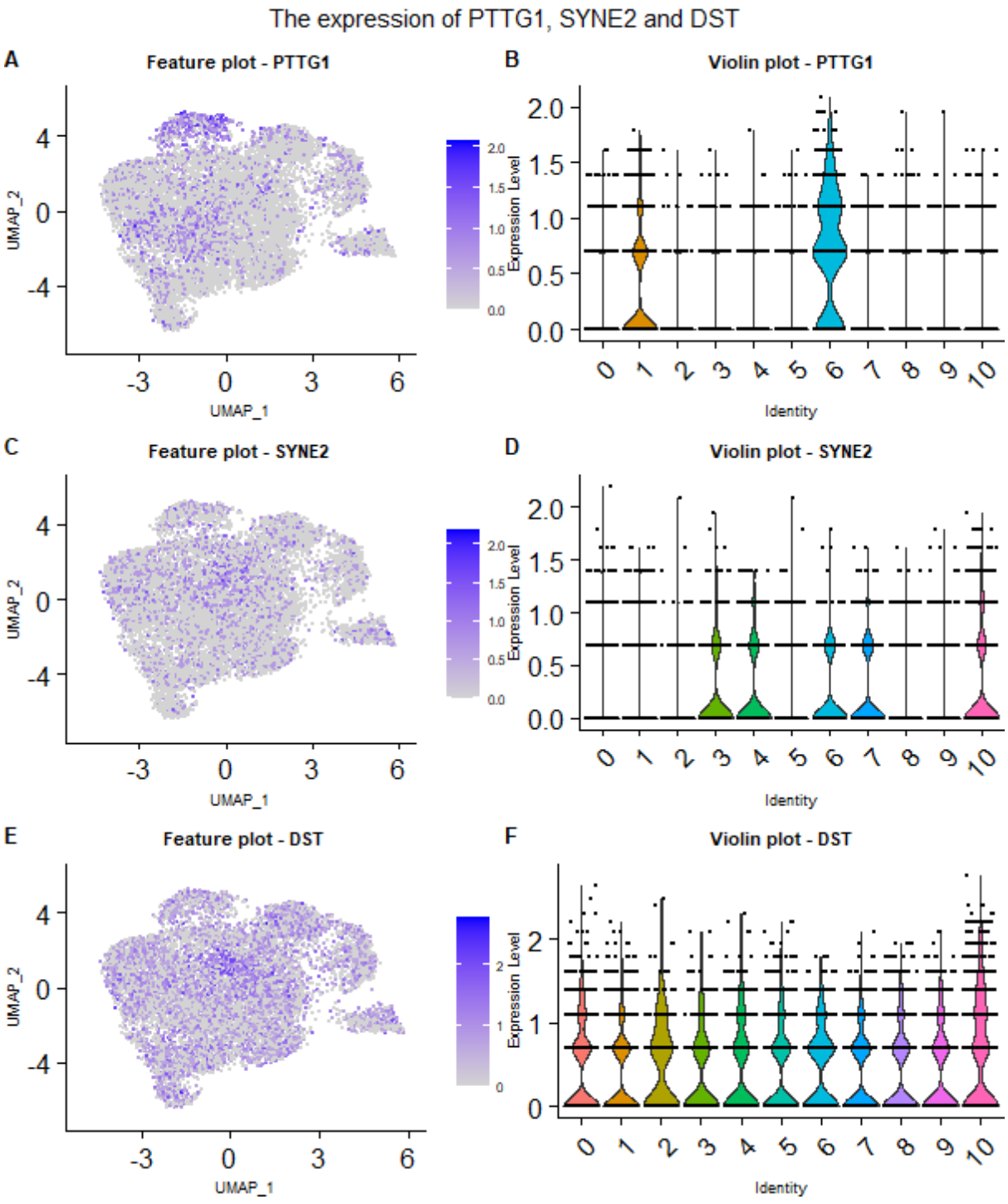


Figure 3.11. Outside of its main cluster of expression (Cluster 6), *PTTG1* was also overexpressed in the region of the plot labelled as Cluster 1 **(A, B)**. The markers of Cluster 10 were less distinctive. These included *SYNE2* **(C, D)** and *DST* **(E, F)**.

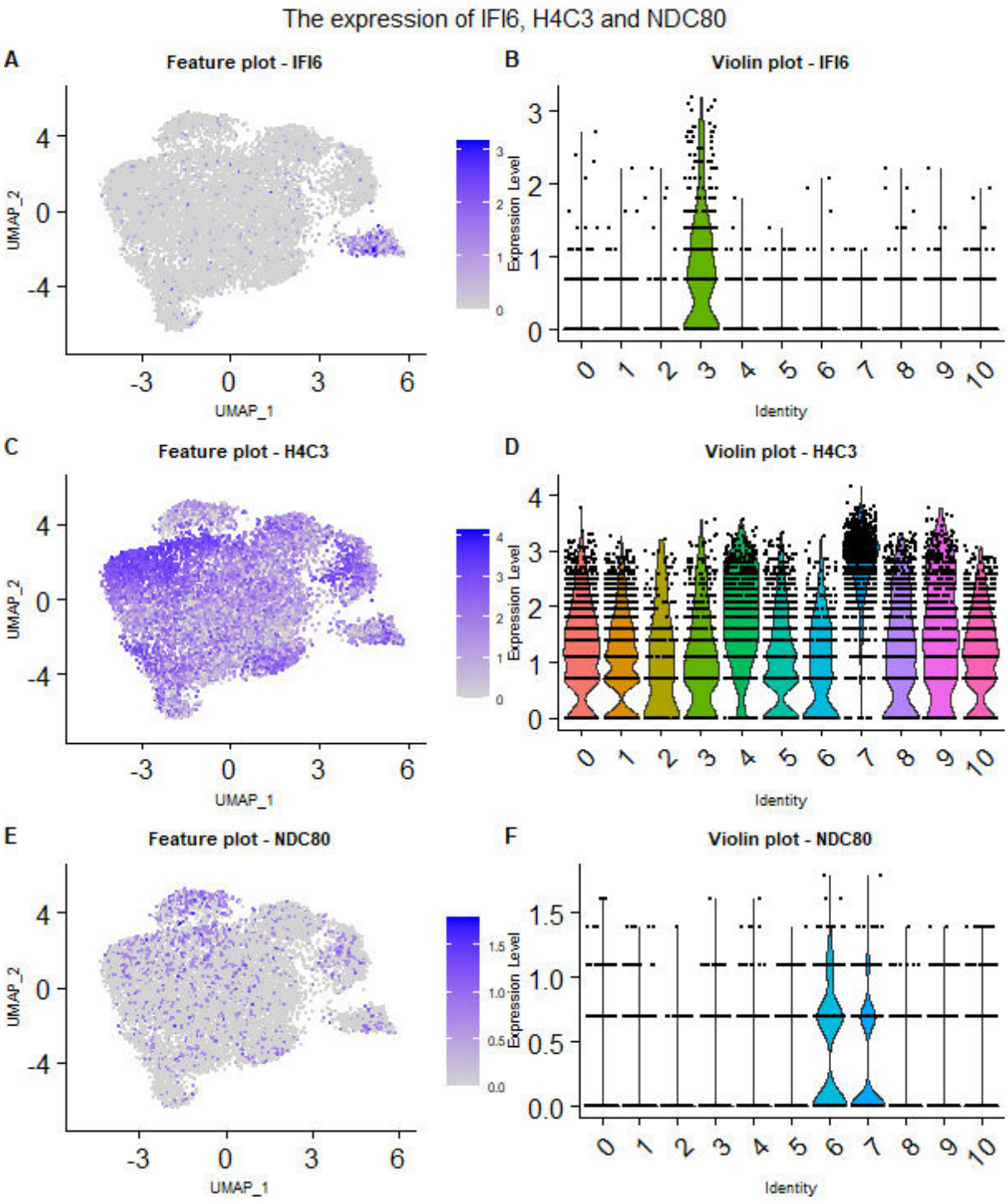


Figure 3.12. The expression of *IFI6* marked Cluster 3 **(A, B)**. *H1-4* was distinctly overexpressed in Cluster 7 **(C, D)**. *NDC80* was overexpressed in Cluster 7, but not in Cluster 4 **(E, F)**.

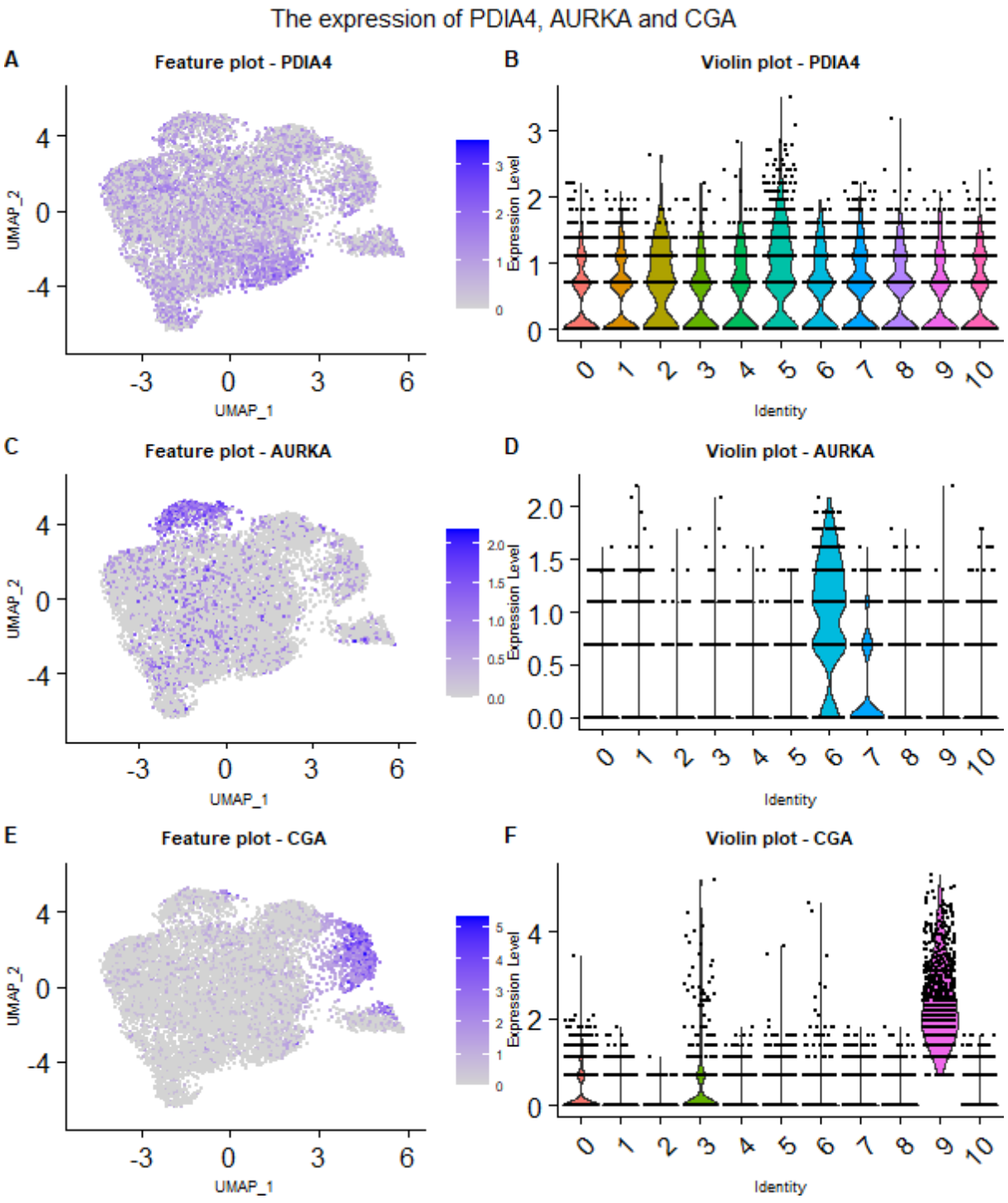


Figure 3.13. *PDIA4* was overexpressed in Cluster 5 **(A, B)**. *AURKA* showed an overexpression in Cluster 6 **(C, D)**. *CGA* was markedly overexpressed in Cluster 6 **(E, F)**.

With the clusters now established, the next section will focus on identifying the clusters associated with cancer stemness.

3.2.3 Identification of clusters associated with cancer stemness

The identification of the clusters associated with cancer stemness followed the six lines of evidence introduced in **Section 2.3.5**: the 27 TDPM genes, the 14 CCRSA gene sets, the 39 SPLCL genes, ORIGINS activity predictions, Slingshot pseudotime inference, and Enrichr gene enrichment analysis using the *CellMarker_Augmented_2021* and *PanglaoDB_Augmented_2021* databases.

15 markers among **the TDPM genes** were found expressed in the A13A scRNA dataset: *MET*, *CD24*, *PROM1*, *EPCAM*, *ALDH1A1*, *LAP3*, *KIT*, *SOX2*, *TSPAN8*, *REG4*, *CD9*, *EZH2*, *AGR2*, *GLRX3* and *HNF4A*. The Spearman correlations of their respective gene expressions (**Figure 3.14**) were small, with their maximum being 0.21, registered for *REG4* and *TSPAN8*.

7 out of the 15 markers were found to be differentially overexpressed in at least one cluster, of which *REG4*, *TSPAN8* and *ALDH1A1* showed an important localization in **Cluster 2** (adjusted p-values: 0, 3.71e-145 and 3.35e-67), compatible with the idea that they may indeed mark CSCs in this dataset. The expression of the other TDPM genes did not align closely to any region of the plot. *AGR2* was also overexpressed in **Cluster 2** (adjusted p-value: 4.16e-46), but it had a high expression all over the dataset.

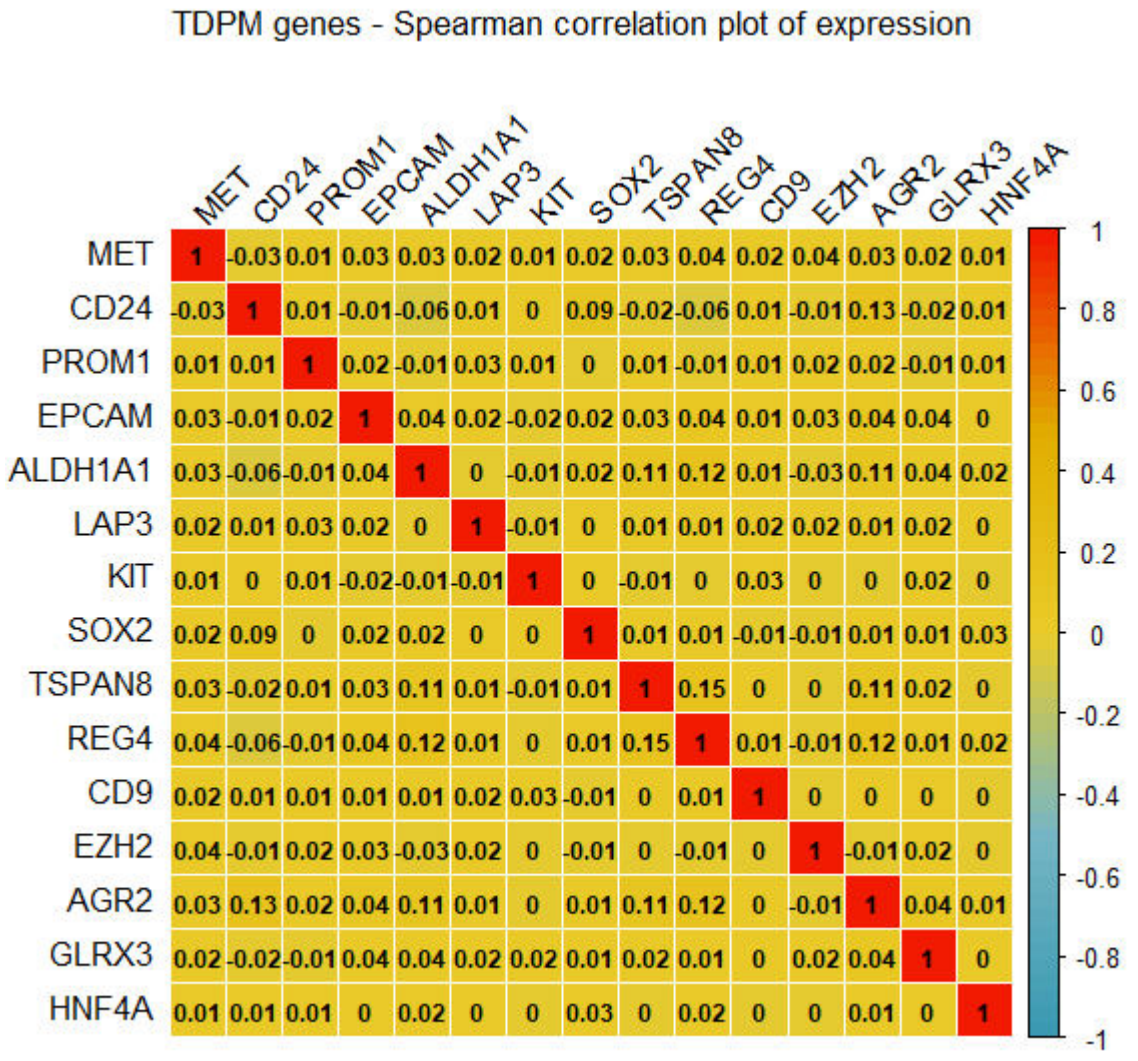


Figure 3.14. Spearman correlation plot of the expression of the 15 TDPM genes that were found in the dataset.

Next, 109 **CCRSA genes** were found expressed in the dataset. **Cluster 6** and **Cluster 7** registered significant marker overlaps with 13 and 10 out of the 14 CCRSA gene sets, respectively. No other cluster recorded significant marker overlaps. In each case, a lower p-value of overlap was obtained for **Cluster 6** than for **Cluster 7** (e.g., p-values of overlap for the union CCRSA gene set: 1.06e-64 for **Cluster 6**, 8.76e-26 for **Cluster 7**).

Cluster 6 and **Cluster 7** showed a similar number of CCRSA genes among their markers (68 and 63, respectively), but the overexpression of CCRSA genes was generally much stronger in the

Cluster 6, with 42 of the CCRSA markers of **Cluster 6** showing a $\log_2$ fold-change above 0.25, compared to just 4 CCRSA markers of **Cluster 7**: *NDC80*, *CENPU*, *TOP2A* and *MKI67*. The distribution of CCRSA genes among cluster markers, both with and without the above-mentioned threshold, is visualised in **Figure 3.15**. For each cluster, the shared CCRSA markers are defined as the CCRSA markers identified as markers of at least one other cluster, while the exclusive CCRSA markers are the CCRSA markers that were not identified as markers of any other cluster.

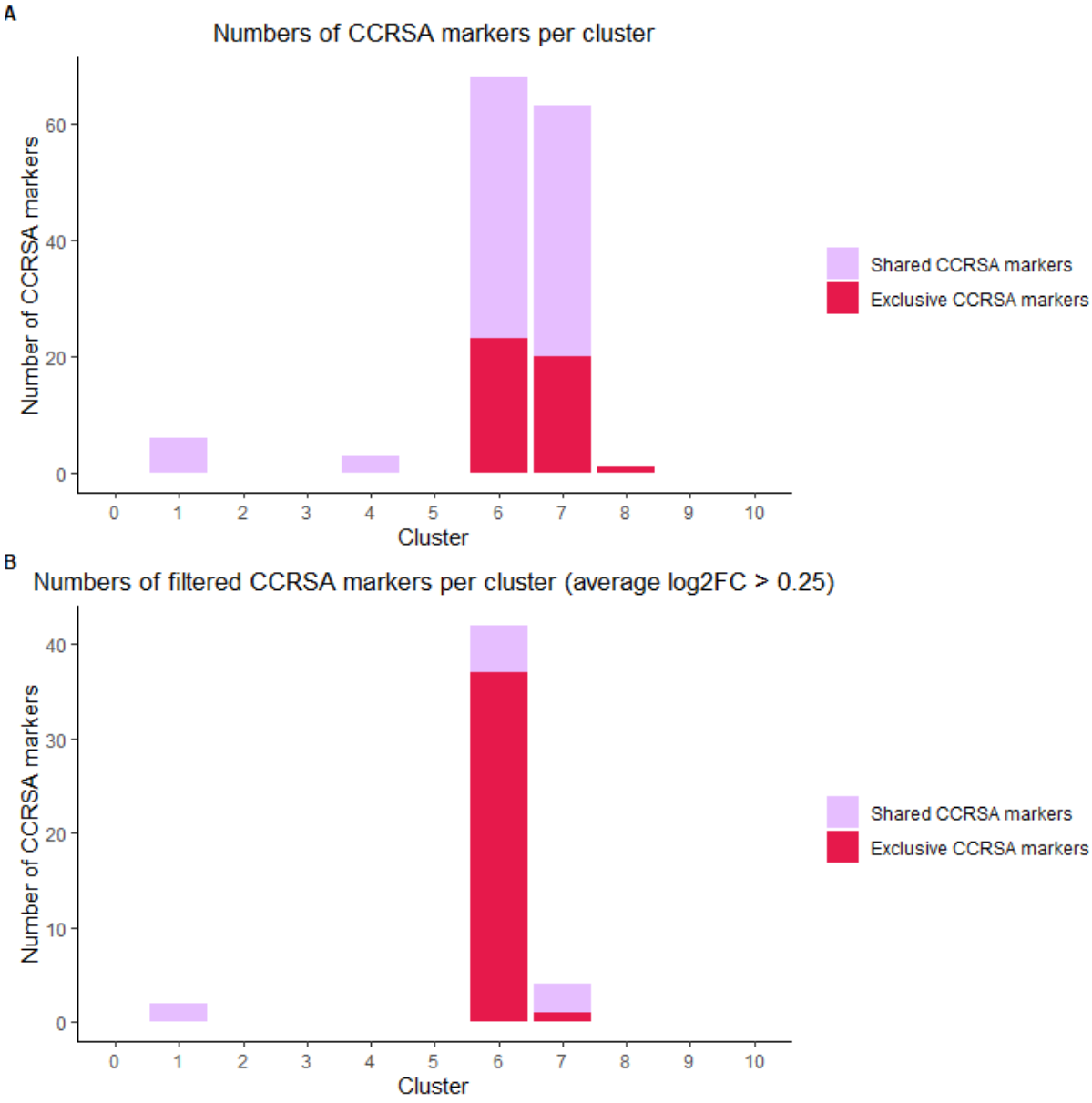


Figure 3.15. Cluster 6 and Cluster 7 both register more than 60 CCRSA genes among their markers **(A)**. But Cluster 6 accounts for the vast majority of CCRSA genes found among its markers when a 0.25 log2 fold-change threshold was applied **(B)**.

Cluster 2 (adjusted p-value: 2.58e-04) and **Cluster 8** (adjusted p-value: 2.58e-04) recorded significant marker overlaps with **the SPLCL genes**. 20 SPLCL genes were found in the data: *AKR1B10*, *EP300*, *GDF15*, *GPAT3*, *MAP1B*, *CYP4F3*, *LRRFIP1*, *KIAA0319*, *H2AC21*, *H1-4*, *TXNRD1*, *NTS*, *WNT5A*, *MAN1A1*, *S100P*, *AKR1C3*, *GDA*, *LIMCH1*, *LIPH* and *TMEM156*, of which 8 were overexpressed in **Cluster 2** (*AKR1B10*, *S100P*, *TXNRD1*, *GDF15*, *AKR1C3*, *TMEM156*, *GPAT3* and *LIPH*) and 5 were overexpressed in **Cluster 8** (*AKR1C3*, *TXNRD1*, *AKR1B10*, *CYP4F3* and *GDF15*).

The two **ORIGINS activity**-based assessments, namely the pairwise Wilcoxon comparisons of the activity scores between clusters (**Figure 3.16**) and the detection of cluster overrepresentation among high activity cells (**Table 3.3**), identified strong stemness associations for **Cluster 2** and **Cluster 6**.

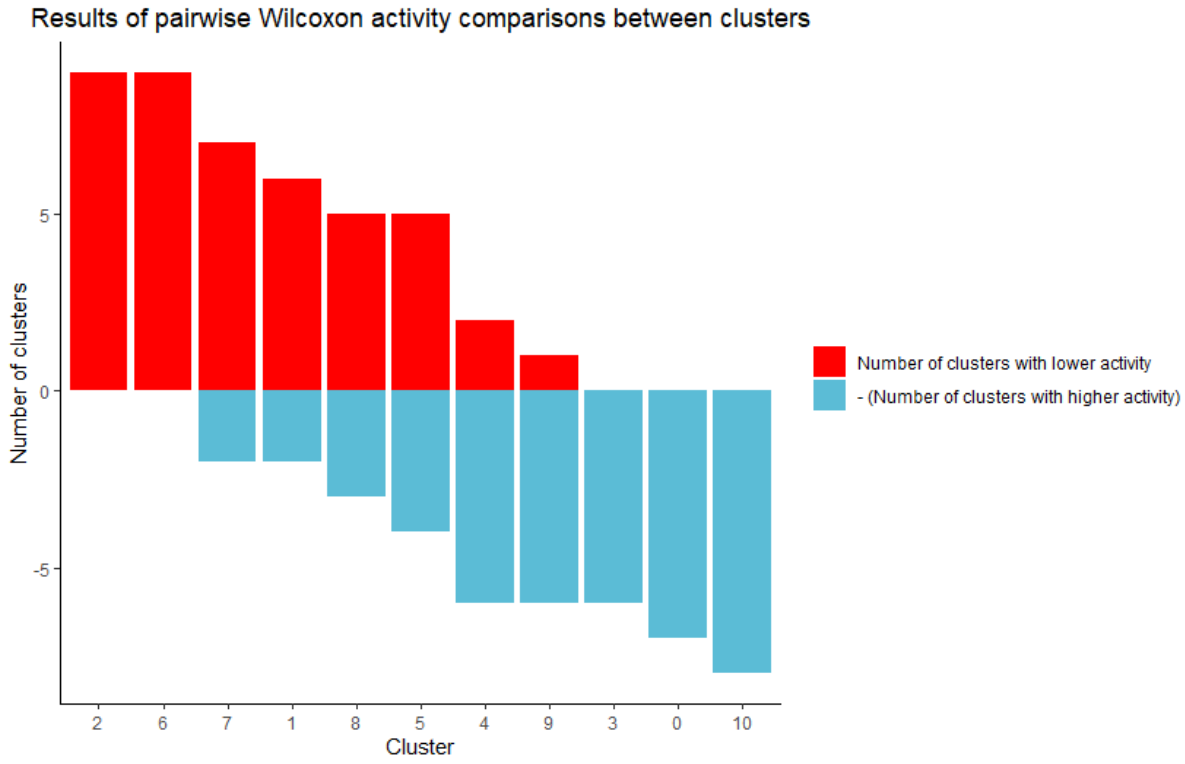


Figure 3.16. Cluster 2 and Cluster 7 showed significantly higher activity scores than the other clusters.

Cluster	p-value of the overrepresentation of the cells among the high ORIGINS activity cells

6	3.01e-11
2	4.24e-11
1	1.25e-06

Table 3.3. The clusters evidencing an overrepresentation of cells with top activity scores.

Slingshot trajectory predictions identified three differentiation lineages (**Figure 3.17**), all originating in **Cluster 6**, and having **Cluster 7** as the next cluster in the trajectory of differentiation.

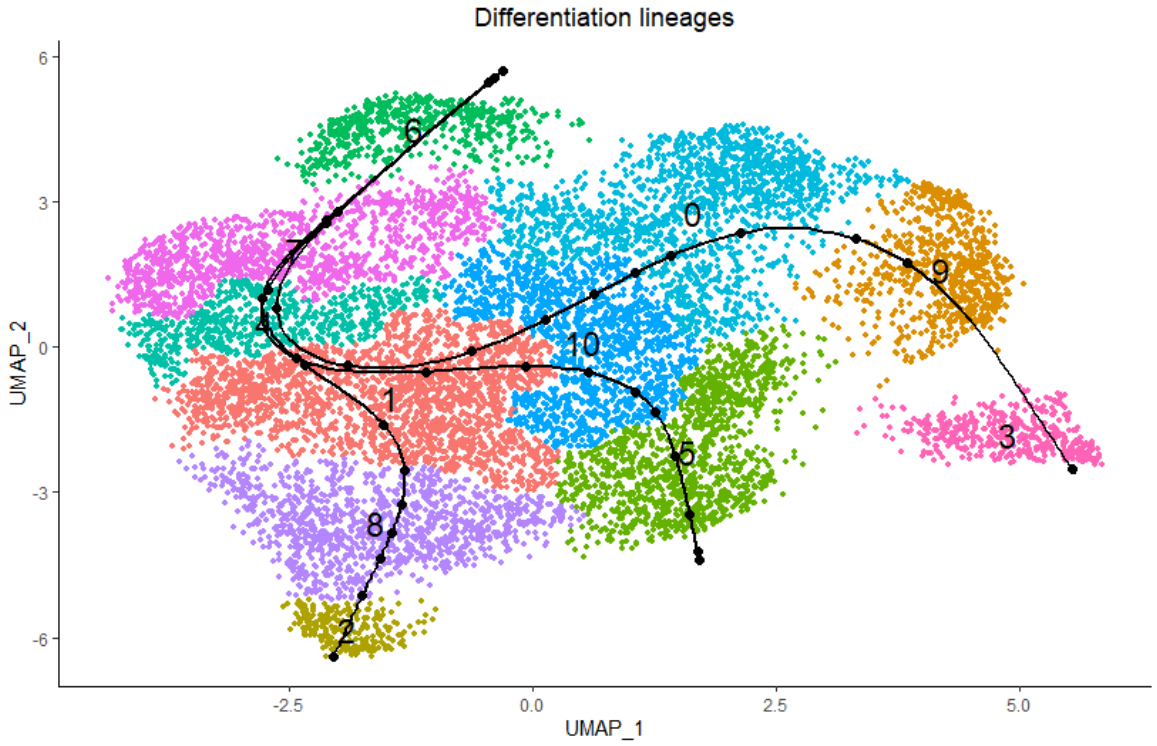


Figure 3.17. Differentiation lineages identified using Slingshot.

The first of the three lineages is illustrated in **Figure 3.18**. After starting in **Cluster 6** and passing through **Cluster 7**, the trajectory of differentiation enters **Cluster 4**, **Cluster 1**, **Cluster 10**, **Cluster 0**, **Cluster 9** and **Cluster 3**, in that order.

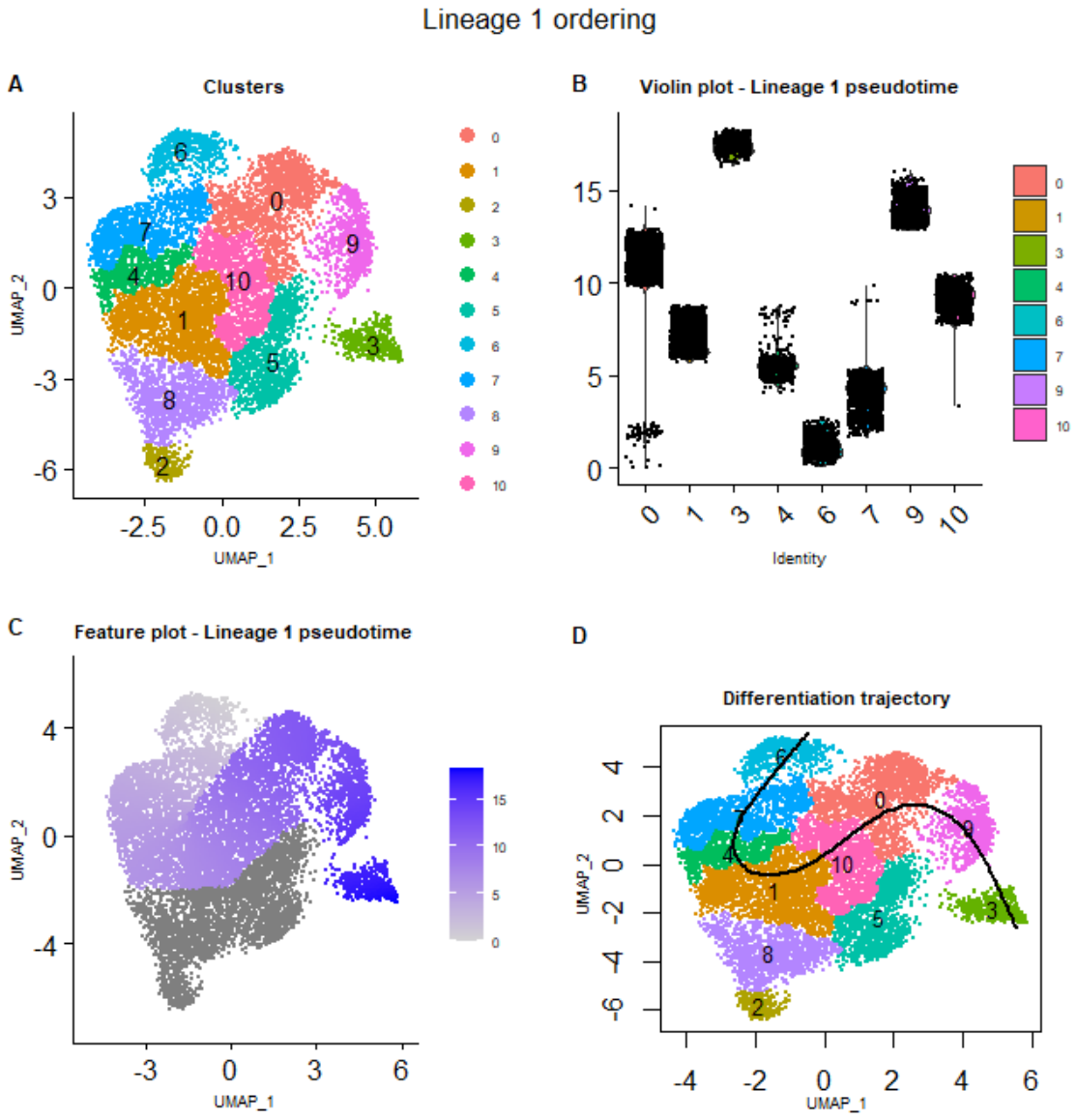


Figure 3.18. Lineage 1 ordering: **A)** Clusters; **B)** Pseudotime violin plot; **C)** Pseudotime feature plot; **D)** Differentiation trajectory.

Next, the second lineage reaches, in order, **Cluster 6**, **Cluster 7**, **Cluster 4**, **Cluster 1**, **Cluster 10** and **Cluster 5** (**Figure 3.19**). Although the first lineage also crosses through the first five clusters listed above, the bifurcation between the first two lineages occurs, as showcased earlier in **Figure 3.17**, already in **Cluster 7**.

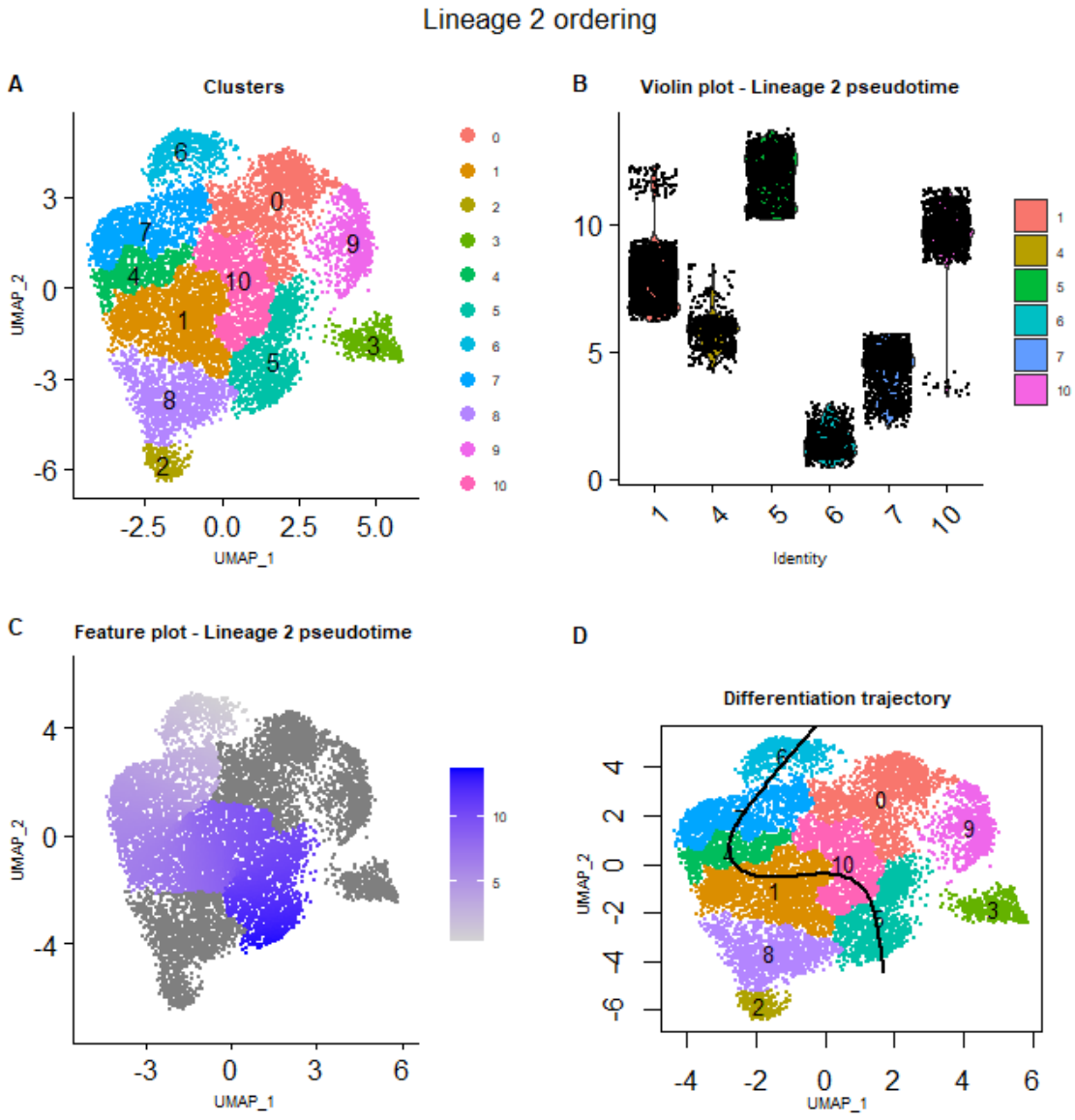


Figure 3.19. Lineage 2 ordering: **A)** Clusters; **B)** Pseudotime violin plot; **C)** Pseudotime feature plot; **D)** Differentiation trajectory.

The third lineage reaches **Cluster 8** and **Cluster 2**, after passing through **Cluster 6**, **Cluster 7**, **Cluster 4** and **Cluster 1** (**Figure 3.20**).

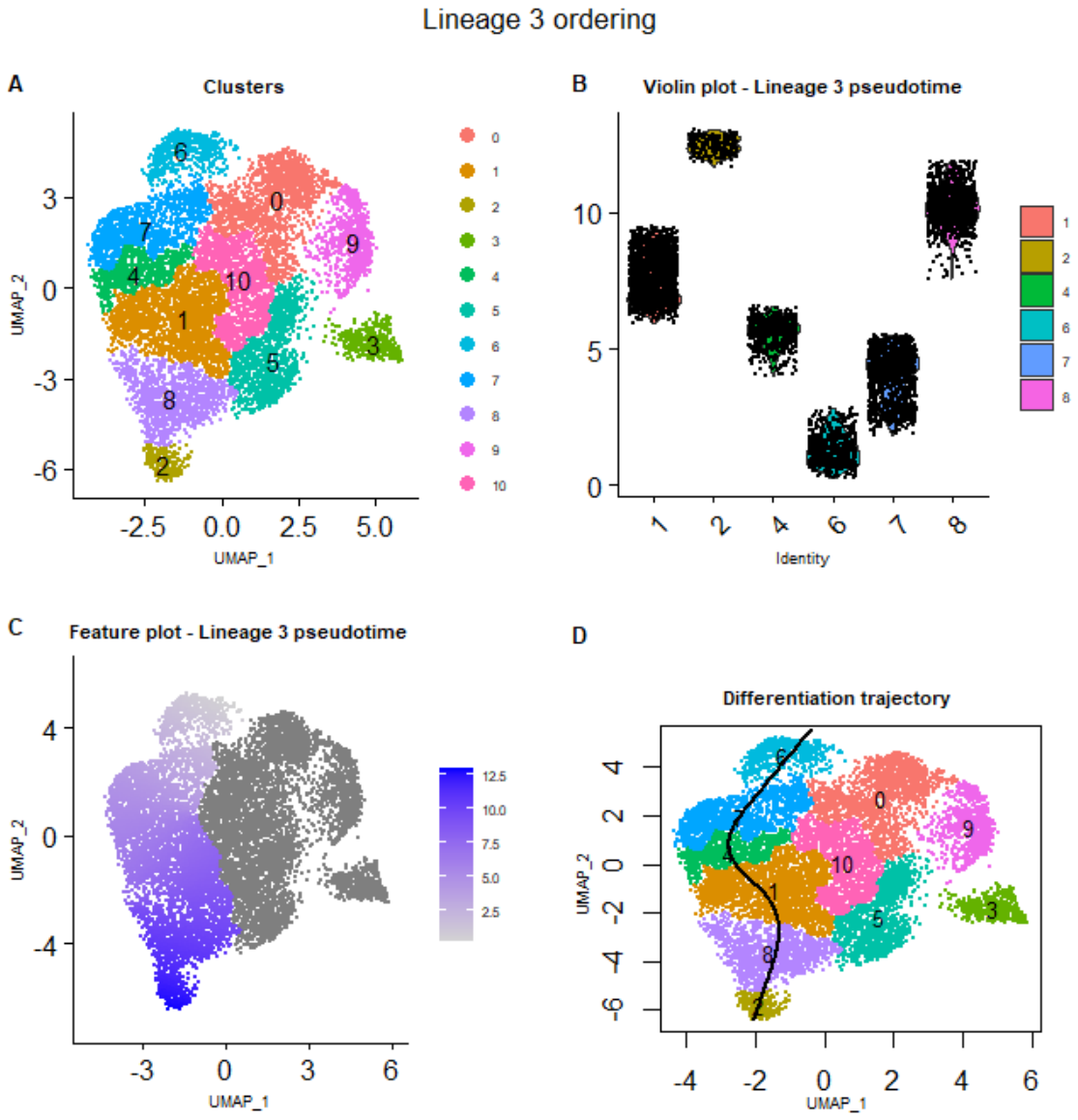


Figure 3.20. Lineage 3 ordering: **A)** Clusters; **B)** Pseudotime violin plot; **C)** Pseudotime feature plot; **D)** Differentiation trajectory.

Remarkably, despite the notable indications of stemness that surfaced earlier for **Cluster 2**, Slingshot posits that this cluster is located at the terminus, rather than at the beginning, of a differentiation trajectory. Therefore, the findings provide evidence against a strictly hierarchical model of stemness, with a "CSC" population at the apex of the differentiation trajectory, gradually differentiating into phenotypes with ever less pronounced developmental potential. Rather, they support the view that a progenitor population can differentiate away from the

original progenitor phenotype into a new population which may later *regain* significant stemness abilities.

Finally, **Cluster 6** and **Cluster 7** obtained the most significant enrichments for stem cell gene sets from the *CellMarker_Augmented_2021* Enrichr database among all clusters. The top seven enriched *CellMarker_Augmented_2021* terms for **Cluster 6** and **Cluster 7** were linked to stem and progenitor cells. **Cluster 2** showcased four *CellMarker_Augmented_2021* terms directly related to stem and progenitor cells among its top enriched five terms, with the other one being labelled "Side-population cell:Undefined". Meanwhile, the *PanglaoDB_Augmented_2021* enrichment analysis evidenced stemness associations for **Cluster 1**, **Cluster 4**, **Cluster 6** and **Cluster 7**. The "Pluripotent Stem Cells" termed ranking the first among all *PanglaoDB_Augmented_2021* terms for all the four clusters. Conspicuously, **Cluster 2** shows no enrichment for this term.

To conclude, strong evidence of stemness has emerged for **Cluster 6**, and to a lesser extent, **Cluster 7**. Noted evidence of stemness has also surfaced for **Cluster 2**, but one lacking some traditional attributes of cancer stemness (being located at the top of the differentiation hierarchy). Crucially, evidence from the Slingshot trajectory analysis places **Cluster 6** as the progenitor of the cells, with **Cluster 7** as the second cluster in all the three lineages that were identified, and **Cluster 2** situated at the end of one of the three lineages. In the light of the findings provided by the other lines of evidence, **Cluster 2** can be interpreted as a population in which a regain of some stemness characteristics occurs.

In the next section, the clusters will be functionally characterized.

3.2.4 Functional characterization of the clusters

Based on the results from the previous section, **Cluster 6** was named **Top stemness** and **Cluster 7** was named **High stemness**. **Figure 3.21** shows the top 5 enriched GO terms for **Cluster 6** and

Cluster 7. These terms are related to the cell cycle, an expected finding given the significant representation of CCRSA genes among the markers of these clusters.

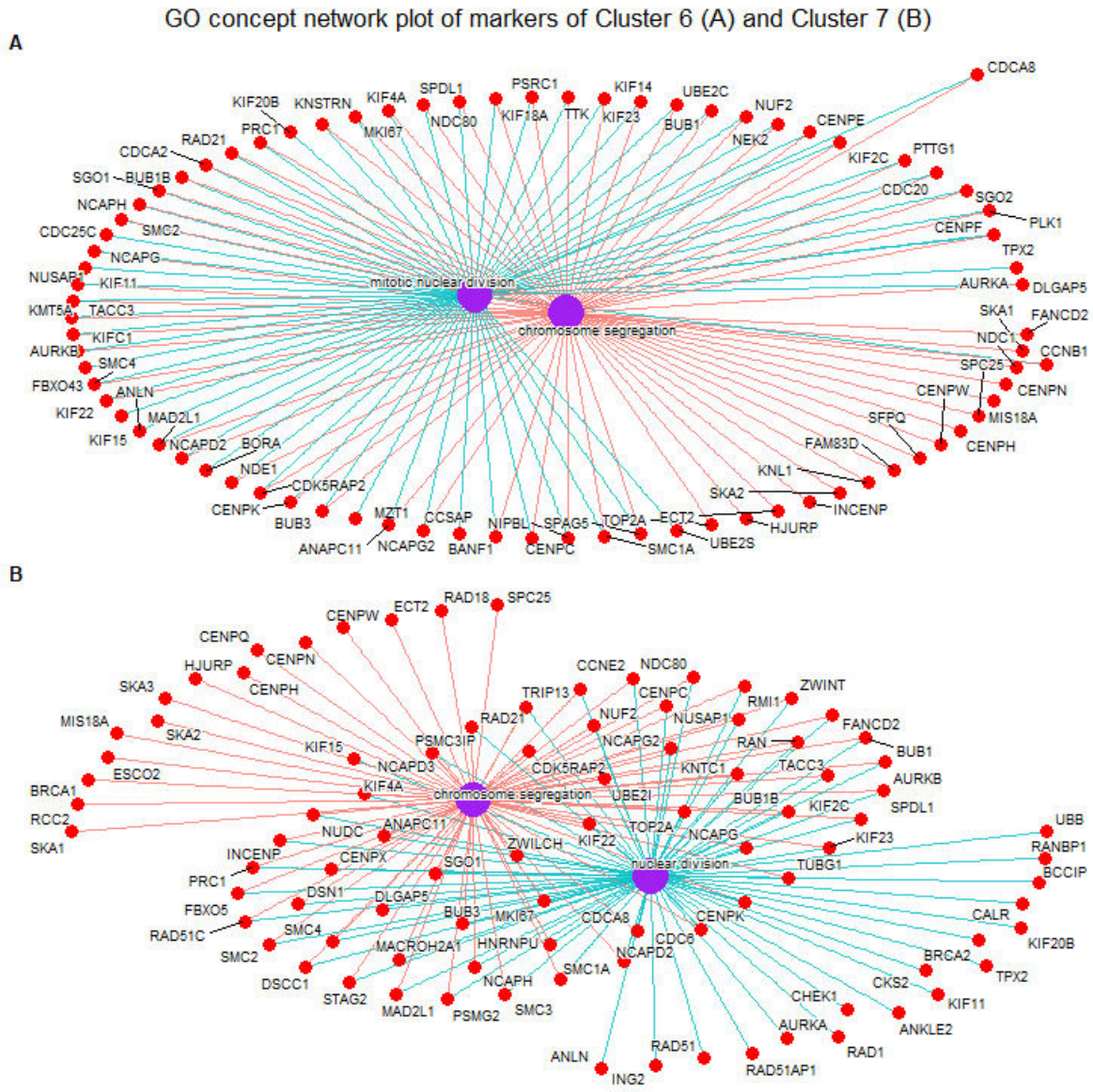


Figure 3.21. Concept network plot of the top 2 enriched GO terms and their associated genes for: **A)** Cluster 6; **B)** Cluster 7.

Next, because of the overlap between its markers and the SPLCL genes, **Cluster 2** was named **SP**, in accord with its side population-like character. Its markers had enriched GO terms related to detoxification. **Cluster 8**, also reaching statistical significance in terms of marker overlap with the SPLCL genes, also showed commonalities with **Cluster 1** in terms of both markers and their

enriched GO terms. As an intermediary between **SP** and **Cluster 1**, **Cluster 8** was named **Transitional SP**, while **Cluster 1**, evidencing distinctly enriched processes related to mRNA and rRNA processing, was named **RNA processing**.

Meanwhile, **Cluster 4** was named **Telomere maintenance**, **Cluster 0** was named **Hypoxia-like** and **Cluster 9** was named **Glycolysis**, in accordance with the GO terms enriched for their markers. **Cluster 3** and **Cluster 5** bore associations with interferon (IF) response and endoplasmic reticulum (ER) stress, respectively. Therefore, they were named **IF response** and **ER stress**. Lacking any significantly enriched GO terms for its markers, **Cluster 10** was also the only cluster for which terms from the *Jensen_DISEASES* Enrichr database were enriched, covering a total of eight cancers. **Cluster 10** thus showcases gene modules overexpressed in multiple cancers, but not linked to cancer stemness. It was therefore named **Bulk cells**.

The functional characterizations of the clusters are listed in **Table 3.4** and illustrated in **Figure 3.22**. These will be used in the subsequent part of this analysis instead of cluster numbers.

Cluster number	**Functional characterization**
0	Hypoxia-like
1	RNA processing
2	SP
3	IF response
4	Telomere maintenance
5	ER stress
6	Top stemness
7	High stemness
8	Transitional SP
9	Glycolysis

10	Bulk cells

Table 3.4. The functional characterization of the clusters.

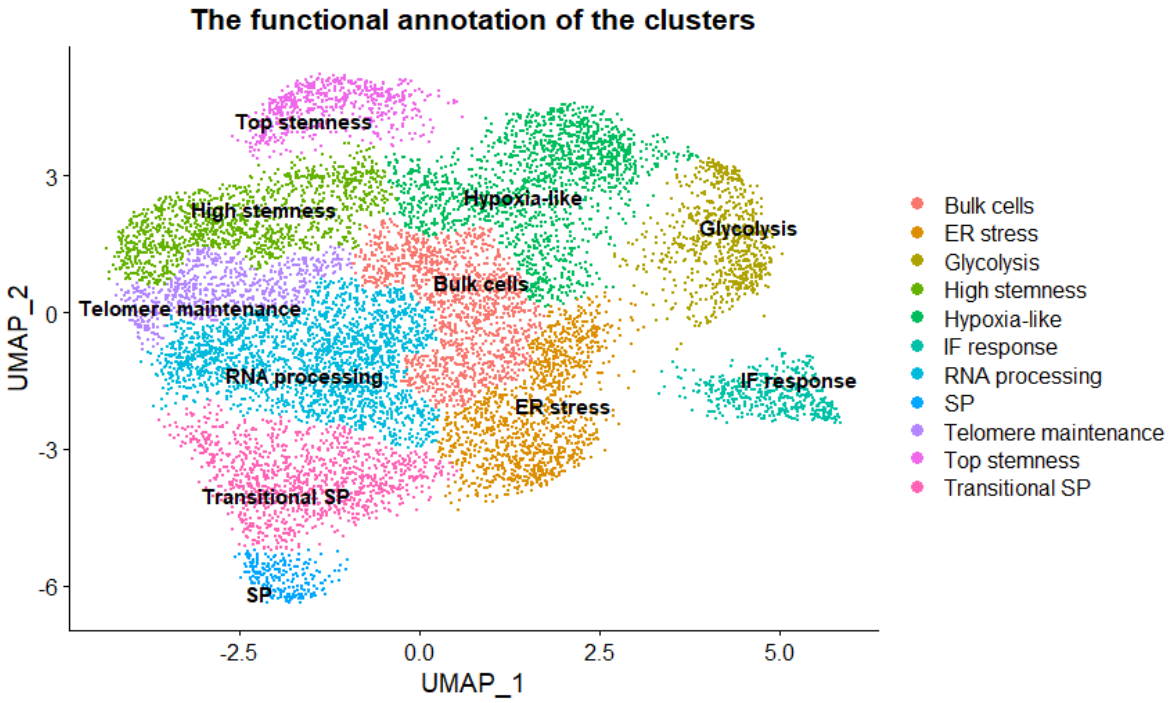


Figure 3.22. The annotation of the clusters

With the clusters now functionally characterized, a detailed analysis of the differentiation trajectory will be provided in the next section.

3.2.5 Analysis of the differentiation trajectory

Using the evaluateK function, the optimal number of knots for the tradeSeq generalised additive model used to fit to the Slingshot trajectories in order to identify changes in gene expression along the lineages was found to be 13. Afterwards, the generalised additive model was generated using this parameter.

Among the three lineages discovered using Slingshot and illustrated in the previous section, Lineage 3 was of considerable importance, as it drove the emergence of the **SP** population, for

which links with stemness have been evidenced despite its advanced position on the pseudotime axis. Therefore, this section will include a thorough characterization of this lineage.

The separation of Lineage 3 from Lineage 1 and Lineage 2 was found to correspond to the region of the plot between the tradeSeq knot 4 (pseudotime: 6.69) and knot 6 (pseudotime: 9.06)

By setting filters for the outputs of the earlyDETest function from tradeSeq (p-value > 0.05; median fold-change > 0.1; Wald test score > 20), 41 genes were found to drive the separation of lineage 3: *FTL*, *TALDO1*, *AGR2*, *OSGIN1*, *AKR1C1*, *TXN*, *S100A6*, *GCLM*, *AKR1C3*, *PRDX1*, *AKR1C2*, *ANKRD36C*, *S100A11*, *ABCC2*, *TXNRD1*, *EBNA1BP2*, *TSPAN1*, *PKIB*, *TMEM92*, *GPAT3*, *SQSTM1*, *PRDX6*, *PGD*, *CLU*, *ME1*, *UACA*, *ELOVL6*, *TM4SF1*, *UGDH*, *KRT8*, *HGD*, *FTH1*, *AMBP*, *ALDH1A1*, *H1-2*, *GSTM3*, *AKR1B1*, *PHLDA2*, *CD63*, *CBR3* and *SDCBP2*, of which 32 genes corresponded to a connected PPIN in the STRING database.

The 41 identified early drivers of the separation of lineage 3 revealed significantly enriched GO terms related to drug resistance, while WP enrichment analysis found the activation of NRF2 pathway, playing roles in cancer stemness and chemoresistance[228], as the main enriched term for the same gene set.

The expression of *FTL*, *AKR1C1*, *AKR1C2*, *NQO1*, *TXNRD1*, *TALDO1* along pseudotime, all important early drivers of the separation of Lineage 3, is displayed in **Figure 3.23**:

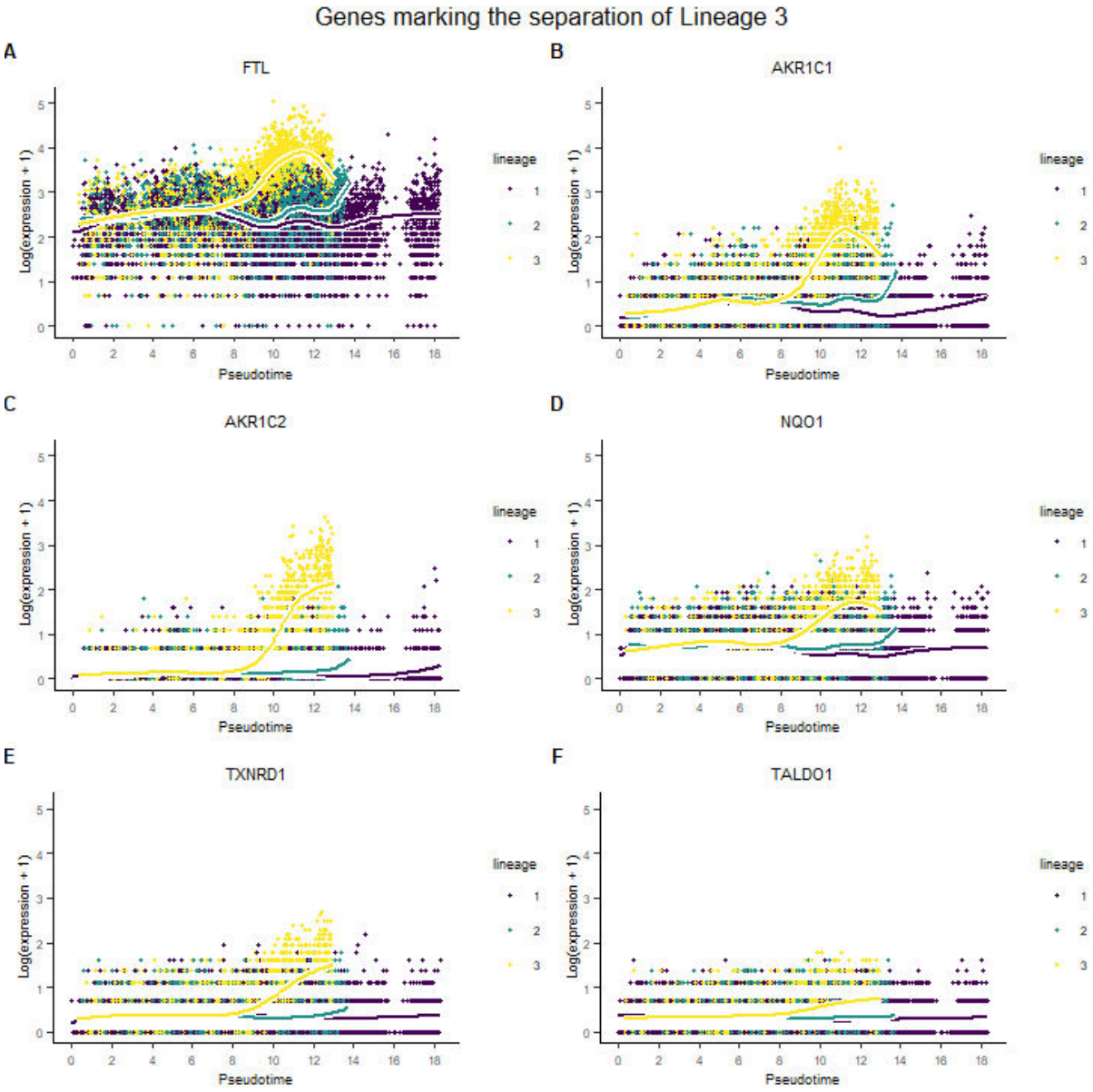


Figure 3.23. The expression of: **A)** *FTL*; **B)** *AKR1C1*; **C)** *AKR1C2*; **D)** *NQO1*; **E)** *TXNRD1* and **F)** *TALDO1* along pseudotime.

The middle course of Lineage 3 is marked by an increase in the expression of several genes including *AKR1B10*, *ASPH*, *PALS2*, *EZR* and *TNFRSF11B*, and by the resurgence of *GDF15* – a **Top stemness** cluster marker and also one of the SPLCL genes. Pseudotime plots of these genes, with distinctive bursts in expression between knot 6 (pseudotime: 9.06) and knot 8 (pseudotime: 10.37) are provided in **Figure 3.24**.

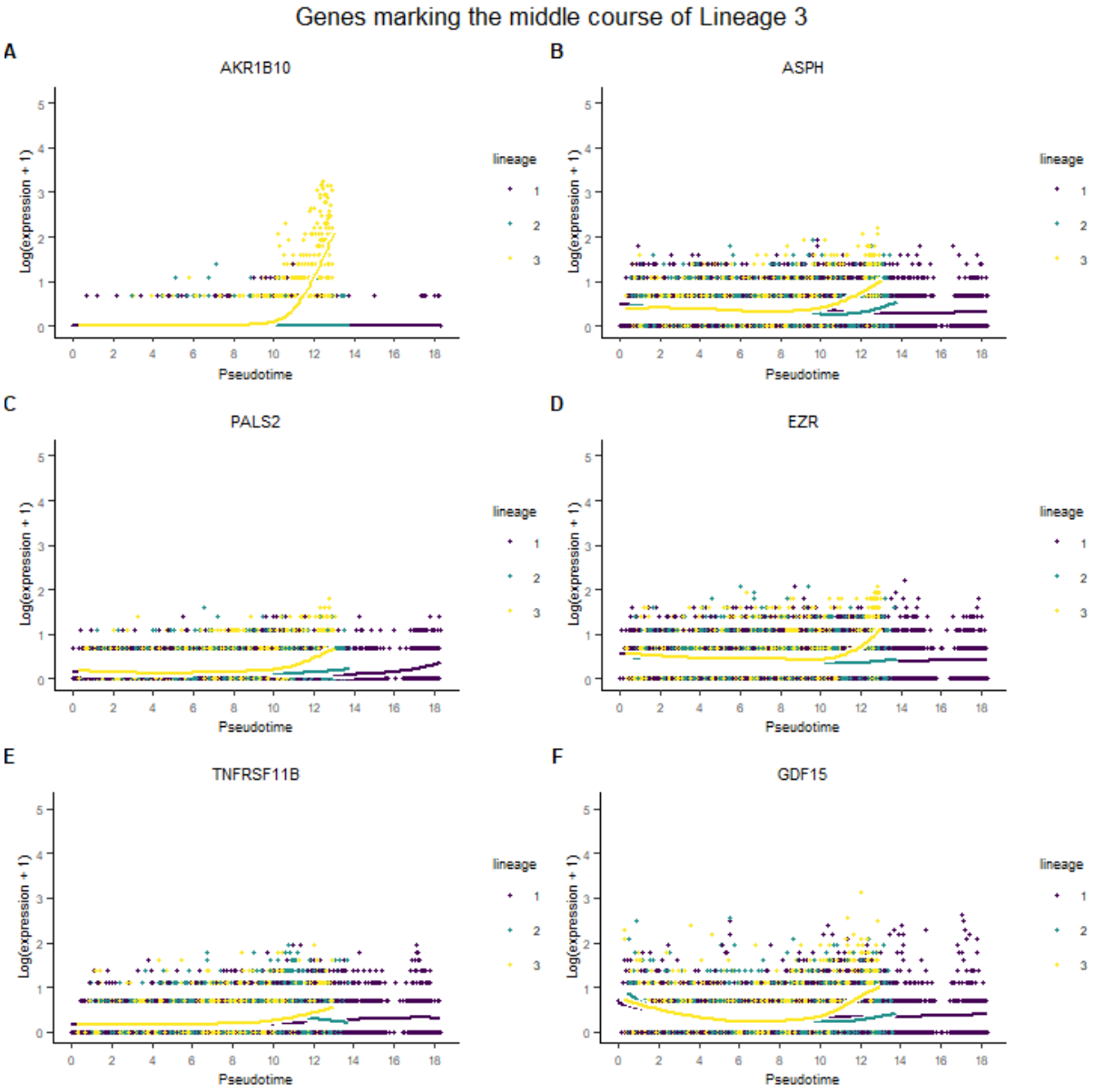


Figure 3.24. The expression of: **A)** *AKR1B10*; **B)** *ASPH*; **C)** *PALS2*; **D)** *EZR*; **E)** *TNFRSF11B* and **F)** *GDF15* along pseudotime.

Later in the evolution of this lineage, a few genes only minimally expressed earlier significantly increased their expression selectively in Lineage 3, between knot 9 (pseudotime: 10.89) and knot 11 (pseudotime: 13.03). The increases in expression were early for *CEACAM6* and *REG4* and late for *ANKRD1* and *GPRC5A* (**Figure 3.25**). 171 markers of the **SP** cluster significantly increased their expression between knots 9 and 11 (p-value > 0.05; median fold-change > 0.1; Wald test score > 20). WP enrichment analysis over these genes revealed an activation of VEGFA signalling.

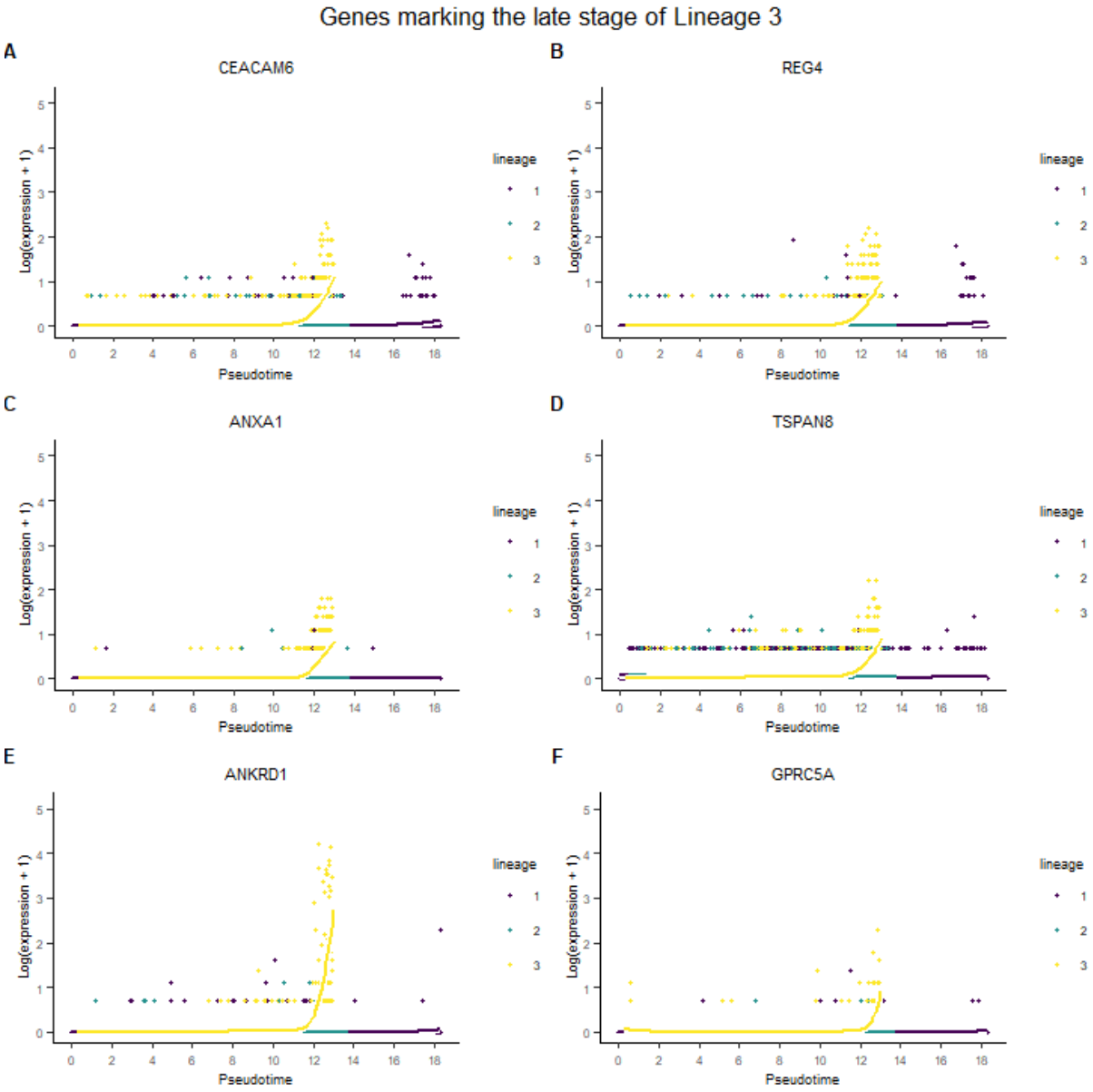


Figure 3.25. The expression of: **A)** *CEACAM6*; **B)** *REG4*; **C)** *ANXA1*; **D)** *TSPAN8*; **E)** *ANKRD1* and **F)** *GPRC5A* along pseudotime.

After knot 11, the number of novel **SP** markers emerging as markers of Lineage 3 within the same constraints as mentioned above drops substantially; only *ADGRF1* surfaces as such between knots 11 and 12, and no new markers characterize the region between knots 12 and 13. Together with the decline in the expression of some of the early drivers of lineage differentiation as illustrated by the concave shape seen in the pseudotime plots of *FTL*, *AKR1C1* and *NQO1* (**Figure 3.23)**, this fact indicates a decline in **SP**-like stemness at the very late stages of the pseudotime.

The combined Slingshot pseudotime had a maximum of 18.32, and an absolute Pearson correlation of 0.16 with ORIGINS activity. No genes were expressed solely at the apex of the developmental potential hierarchy – highest activity or lowest pseudotime (**Figure 3.26**).

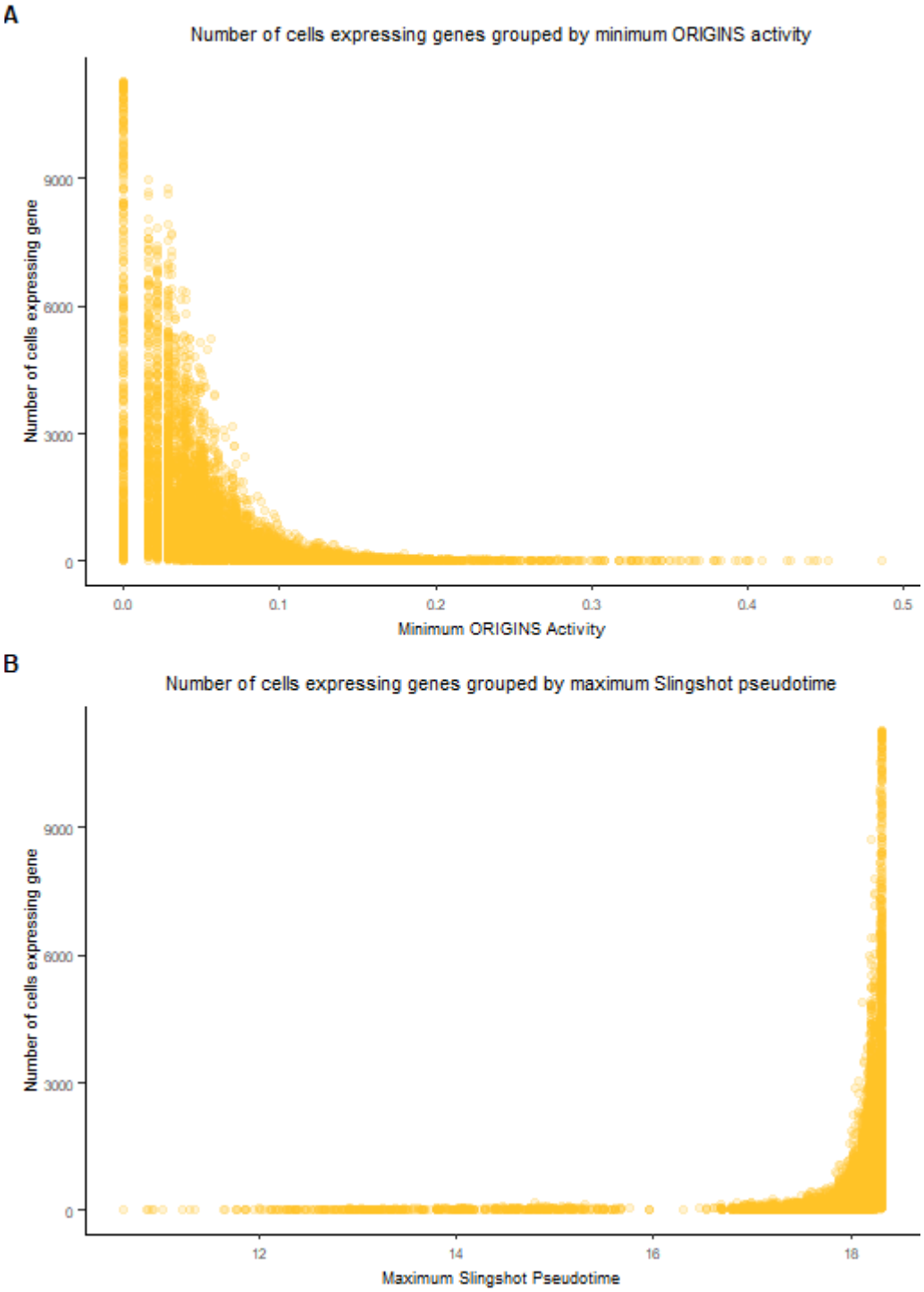


Figure 3.26. The distribution of the number of cells expressing each gene, grouped by: the ORIGINS activity minima **(A)**, Slingshot pseudotime maxima **(B)** of each gene.

Characteristic gene sets, defined in **Section 2.3.10**, were determined for activity medians and means per gene (**Figure 3.27**), and Slingshot medians and means per gene, having 46, 39, 29 and 25 genes, respectively (**Figure 3.28**).

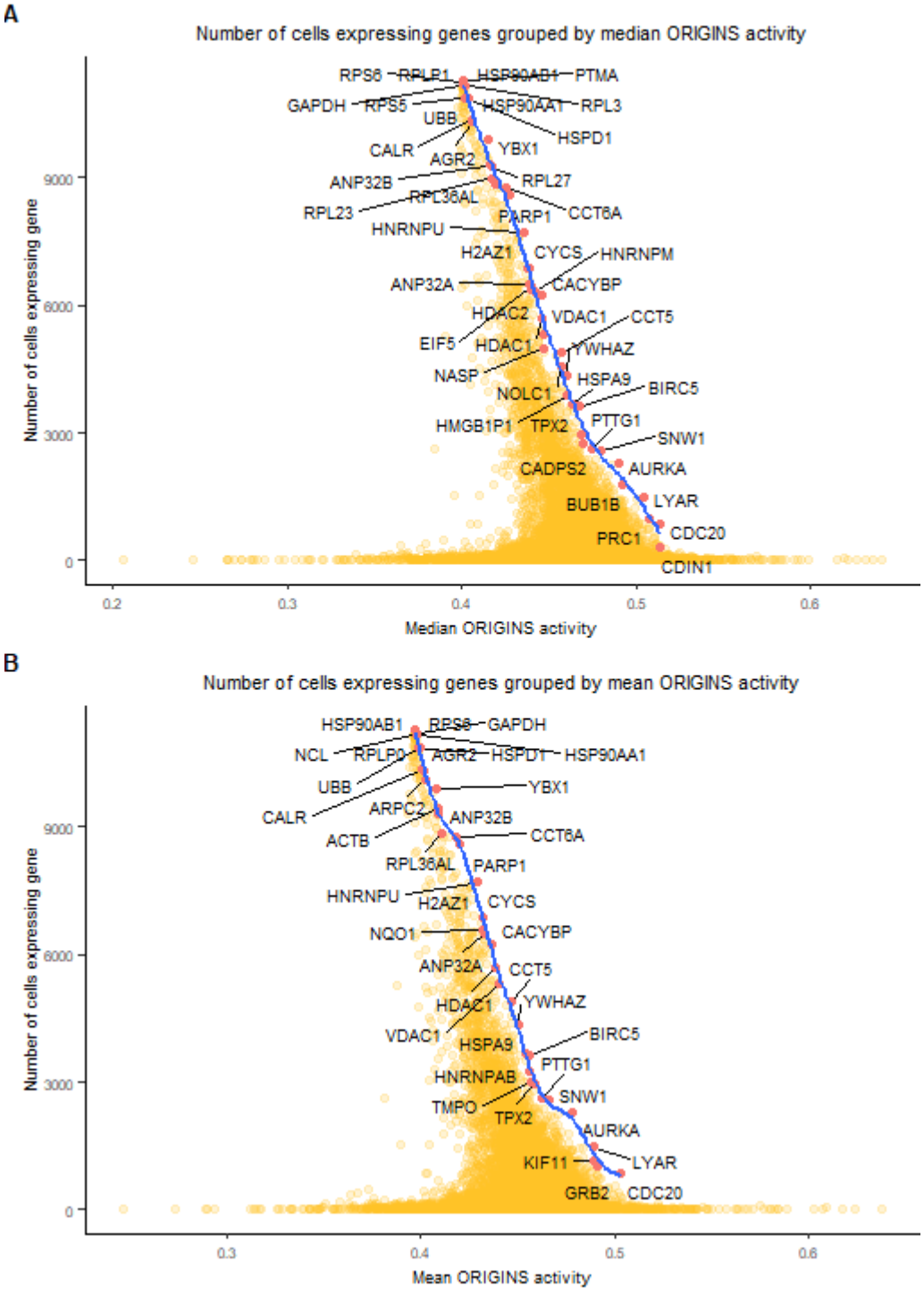


Figure 3.27. The distribution of the number of cells expressing each gene, grouped by: ORIGINS activity medians **(A)**; ORIGINS activity means **(B)**. The characteristic genes are also displayed on the plot.

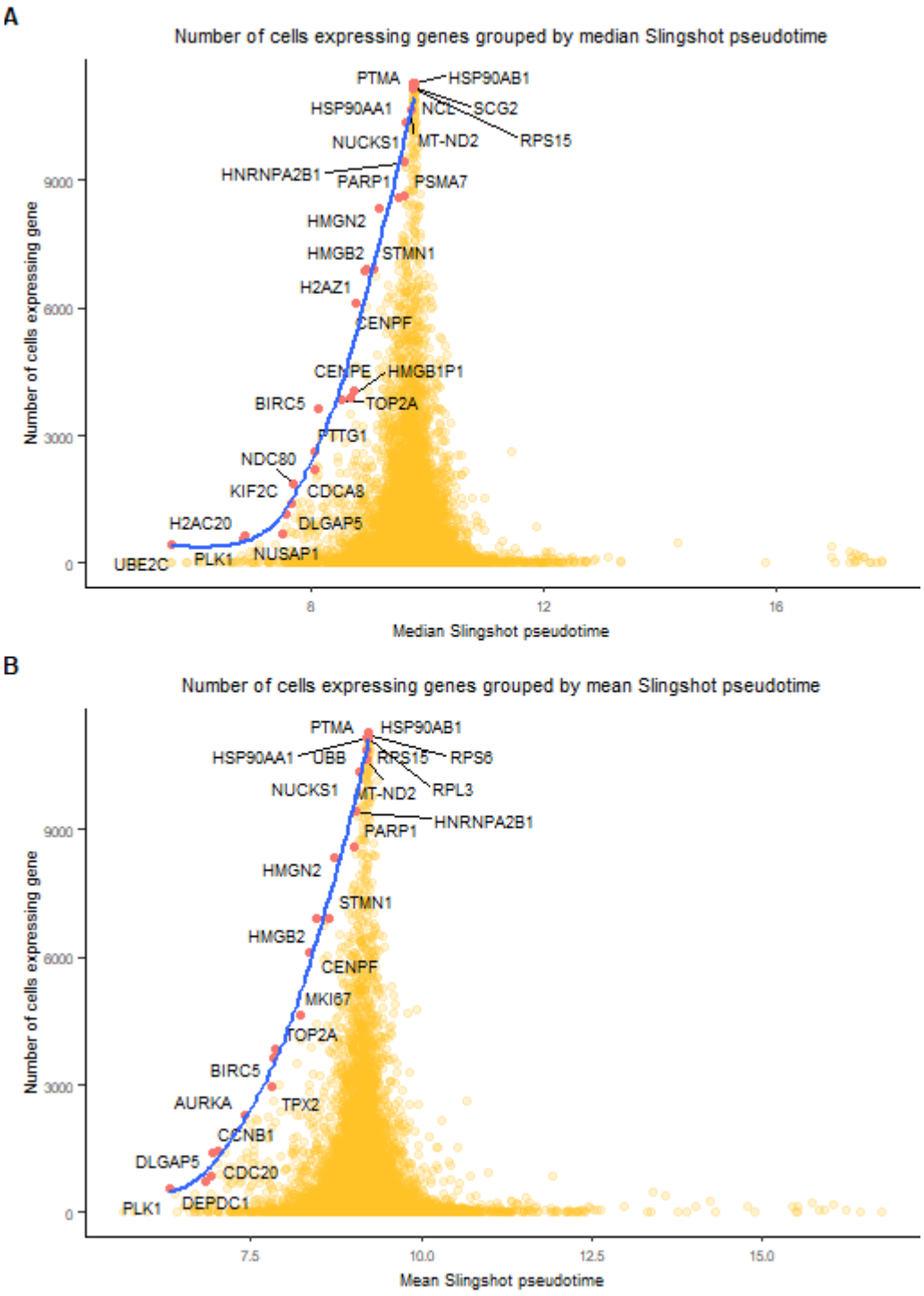


Figure 3.28. The distribution of the number of cells expressing each gene, grouped by: Slingshot pseudotime medians **(A)**; Slingshot pseudotime means **(B)**. The characteristic genes are also displayed on the plot.

12 genes appeared in at least three of the four characteristic sets (**Table 3.5**):

Genes	**Number of occurrences in the characteristic**

	gene sets
PARP1, HSP90AB1, HSP90AA1 and *BIRC5*	4
UBB, TPX2, RPS6, PTTG1, PTMA, H2AZ1, CDC20 and *AURKA*	3

Table 3.5. The 12 genes appearing at least thrice in the four characteristic gene sets.

All pairs of characteristic gene sets overlapped to a statistically significant extent, with p-values of overlaps ranging from 5.32e-68 (activity medians set and activity means set) to 6.02e-12 (Slingshot pseudotime medians set and activity means set). The cardinalities of all intersections of two, three and all four characteristic gene sets are illustrated in **Figure 3.29**.

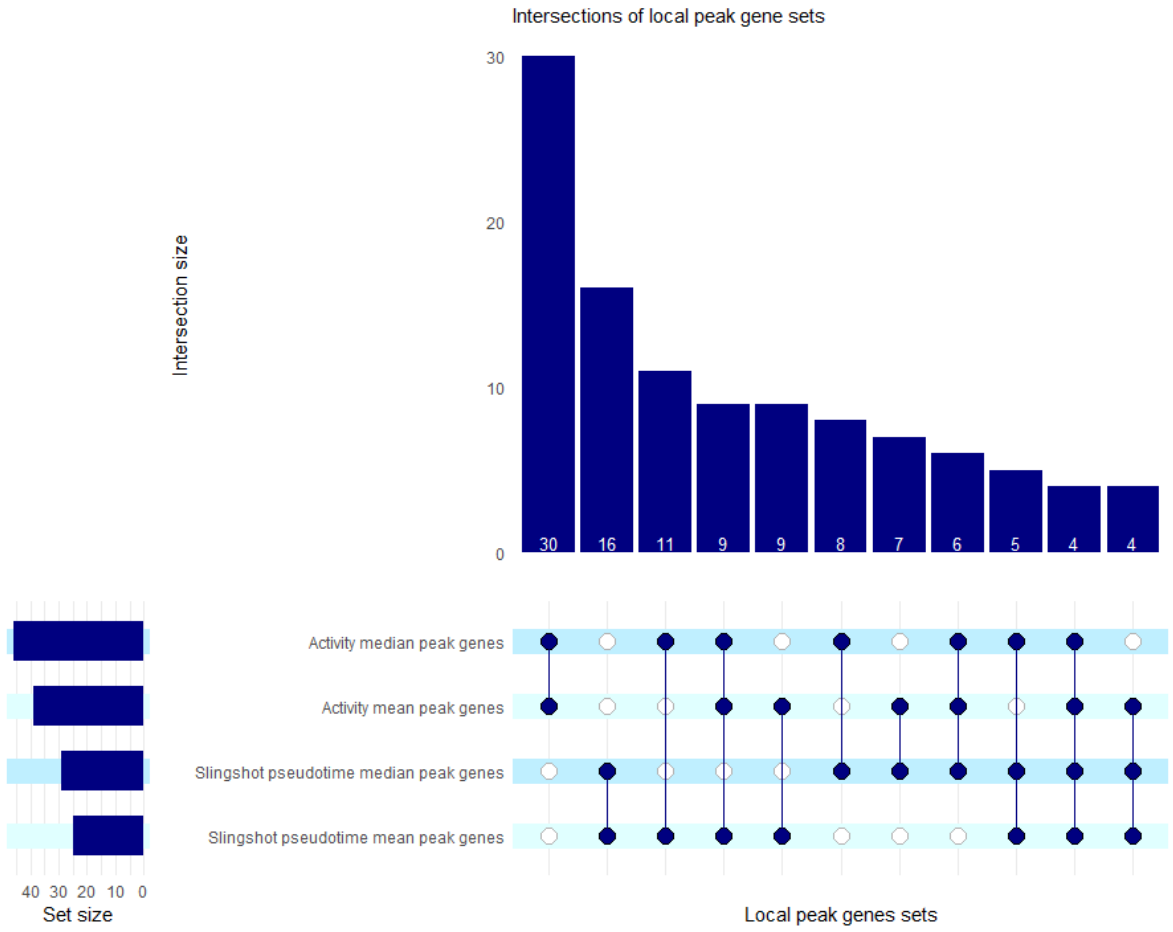


Figure 3.29. The intersections of the characteristic gene sets.

Thus, although ORIGINS activity and Slingshot pseudotime scores showed a limited correlation, similar sets of genes were found to lie at the root of the variation of developmental potential within PDAC cells for both assessments. Furthermore, all the four characteristic gene sets showed significant overlaps with the union CCRSA gene set, with p-values ranging from 1.39e-14 (Slingshot pseudotime medians set) to 2.1e-06 (Activity means set).

Among the 12 genes shared by at least three characteristic gene sets, 5 were CCRSA genes, and the others were *HSP90AA1*, *HSP90AB1*, *PARP1*, *H2AZ1*, *PTMA*, *RPS6* and *UBB*. These 7 genes had patterns of expression resembling those of ORIGINS activity, paralleling some of the extra-**Top stemness** activity local peaks seen in the **High stemness** and **RNA processing** clusters (**Figure 3.30**).

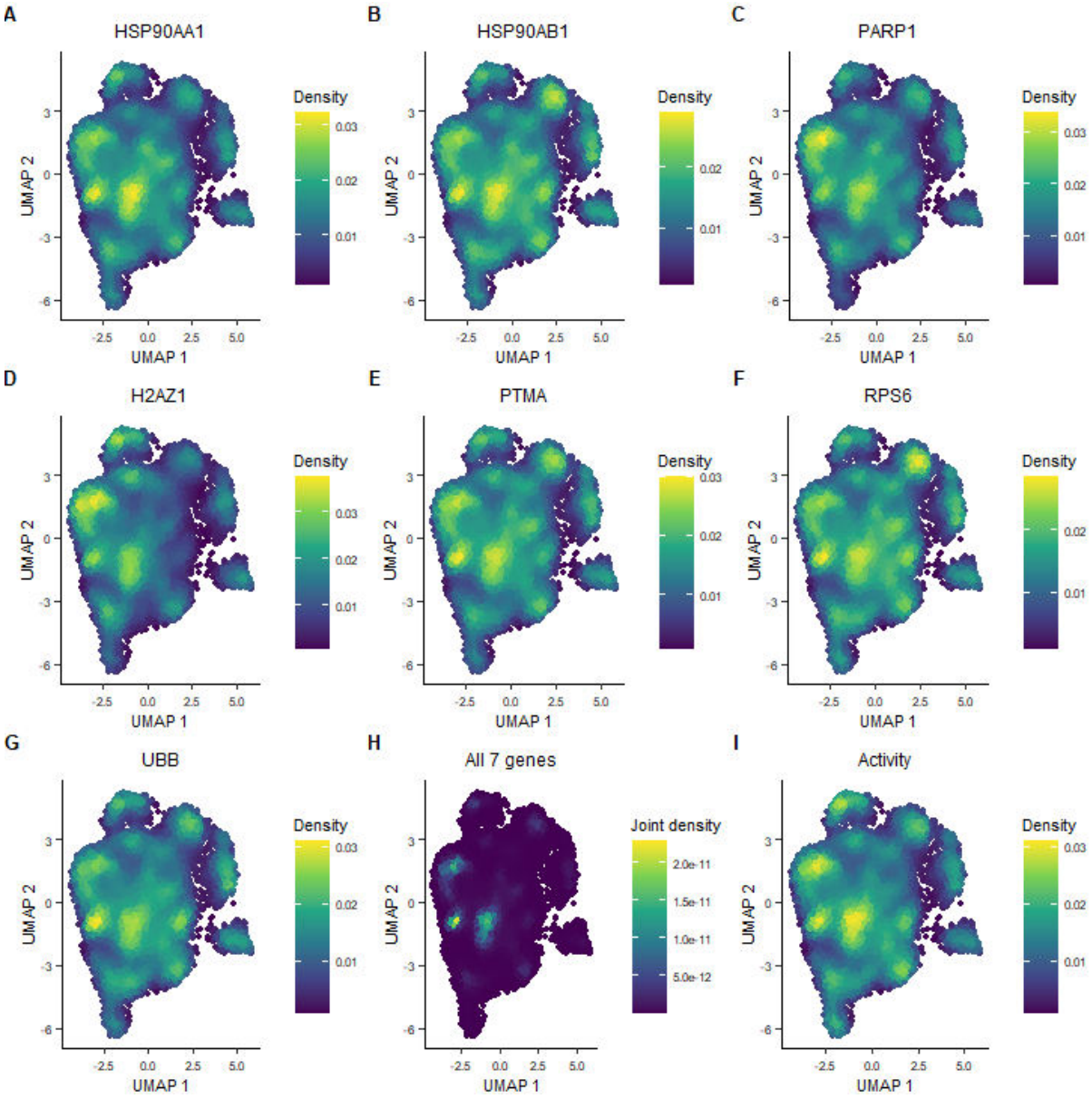


Figure 3.30. Nebulosa plots of the 7 genes found in the intersection of at least three characteristic sets but not among the CCRSA genes: **A)** HSP90AA1; **B)** HSP90AB1; **C)** PARP1; **D)** H2AZ1; **E)** PTMA; **F)** RPS6; **G)** UBB1, of **H)** their joint density, and of **I)** ORIGINS activity.

Finally, RNA velocity assessment using velocyto.R (**Figure 3.31**) suggests that movement from the **Transitional SP** to the **SP** cluster occurs through the left part of the contact region between the two clusters, followed by a displacement towards the centre of the **SP** cluster, corresponding to the local overexpression of genes such as *ANKRD1*, *VSIG1* and *GPRC5A* (**Figure 3.32**), markers of the late evolution of Lineage 3. In addition, cells from the **Top stemness** cluster are moving away from the **Hypoxia-like** cluster, while the **IF response** cluster appears to be disintegrating, with a

part of the cells strongly pulling further away and the other part returning to the centre of the plot. The rightmost regions of the **RNA processing** cluster, of moderate stemness character, are differentiating towards **Bulk cells**.

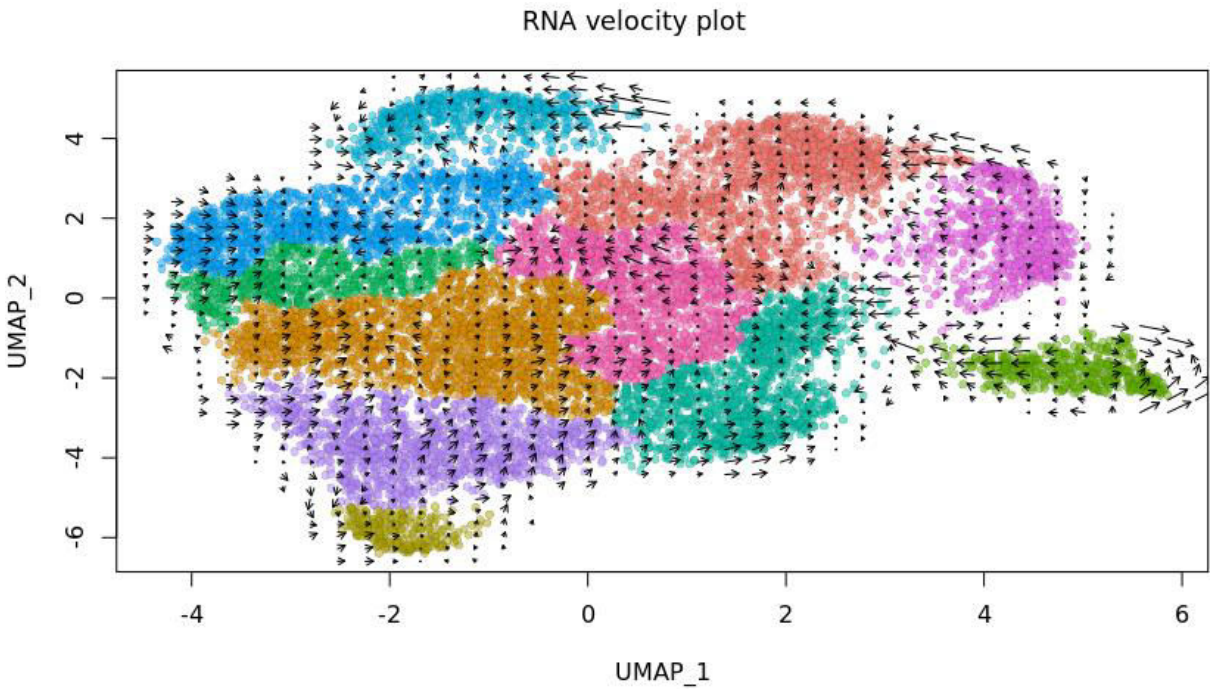


Figure 3.31. RNA velocity plot of the cells in the dataset.

To conclude this section, the analysis of the differentiation trajectory shed light upon the three lineages present in the dataset, identified the early genes responsible for the emergence and further evolution of the drug resistant **SP** subpopulation, found sets of genes marking the change in developmental potential shared between ORIGINS activity and Slingshot pseudotime, and determined the trends in the direction of nascent RNA.

The next section will provide a comparative look at the inter-cluster and intra-cluster cell-cell communication characterizing the clusters, using SingleCellSignalR and CellChat.

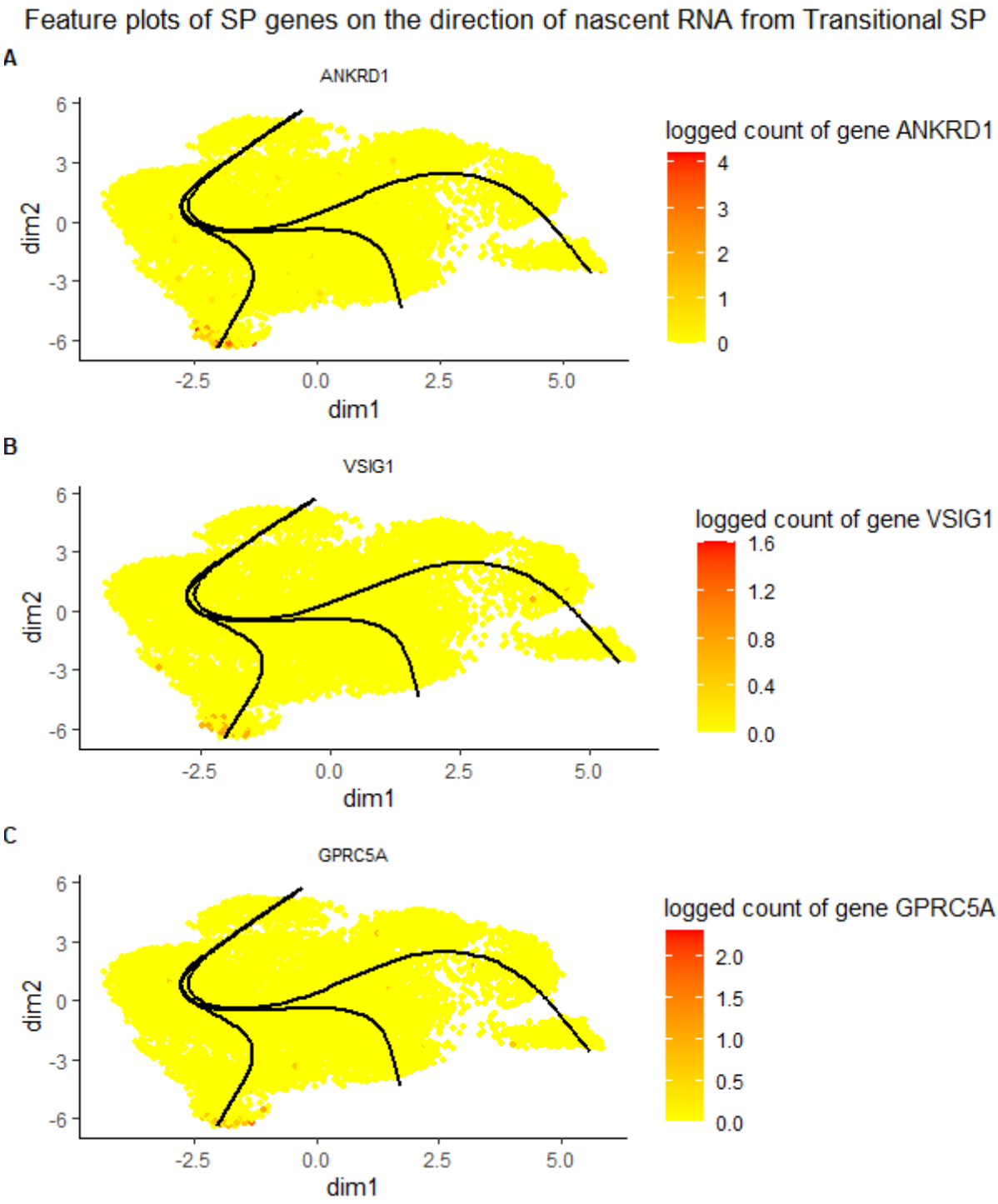


Figure 3.32. TradeSeq feature plots of **A)** ANKRD1; **B)** VSIG1 and **C)** GPRC5A.

3.2.6 Analysis of cell-cell communication

SingleCellSignalR analysis revealed that **SP** lacks cell-cell interactions involving TGFBR1, shared by all the other clusters. On the other hand, **SP** also displays its own specific cell-cell interactions

involving ANXA1, CEACAM5 and CEACAM6. The most variable 30 cell-cell interactions found by SingleCellSignalR are displayed in **Figure 3.33**:

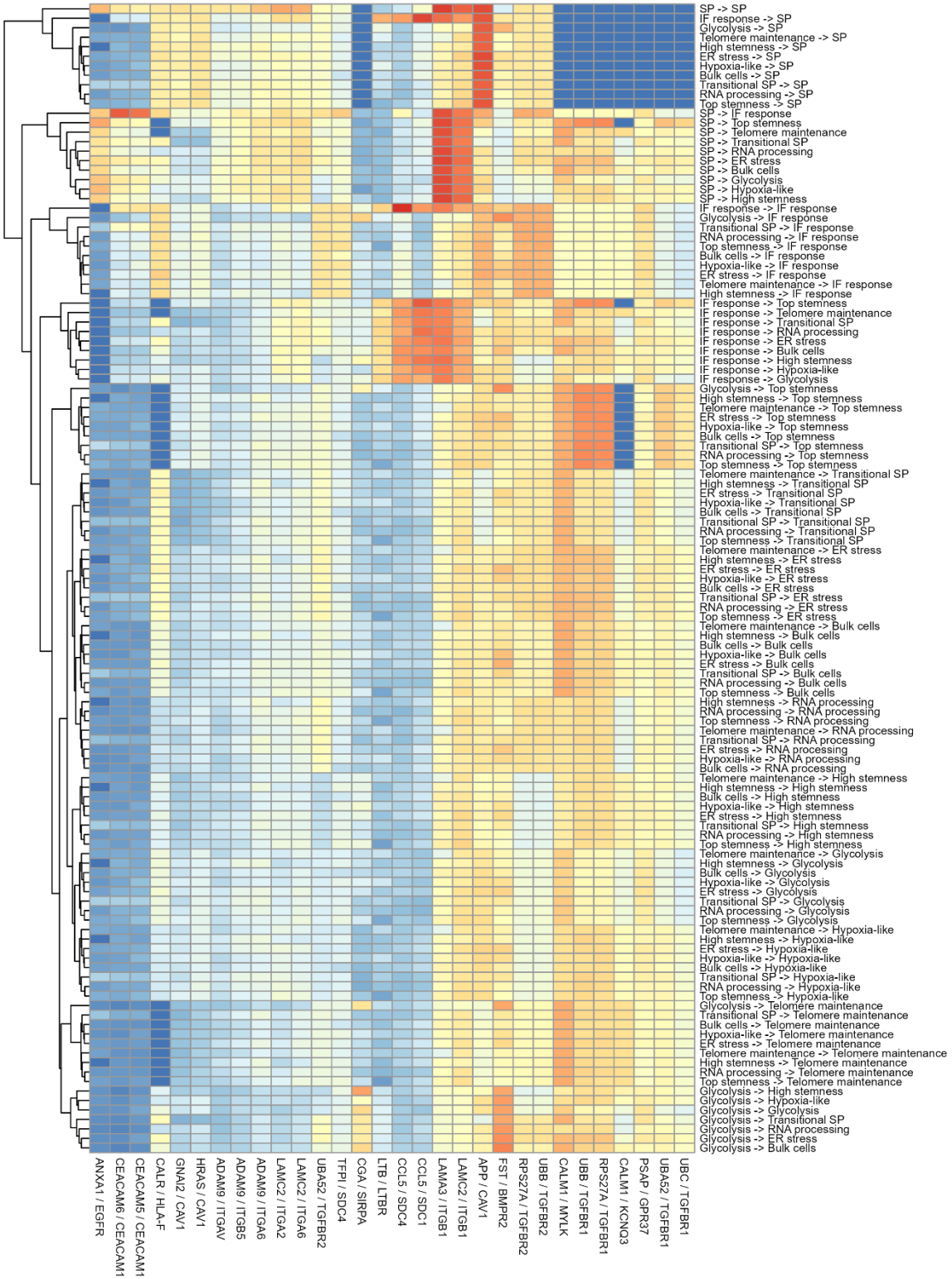


Figure 3.33. Top 30 variable cell-cell interactions found by SingleCellSignalR.

CellChat analysis found 16 signalling pathways significantly activated between pairs of clusters or within clusters. MK and laminin signalling were active between most pairs of clusters, while CADM signalling involved the **Top stemness** cluster to an important extent (**Figure 3.34**):

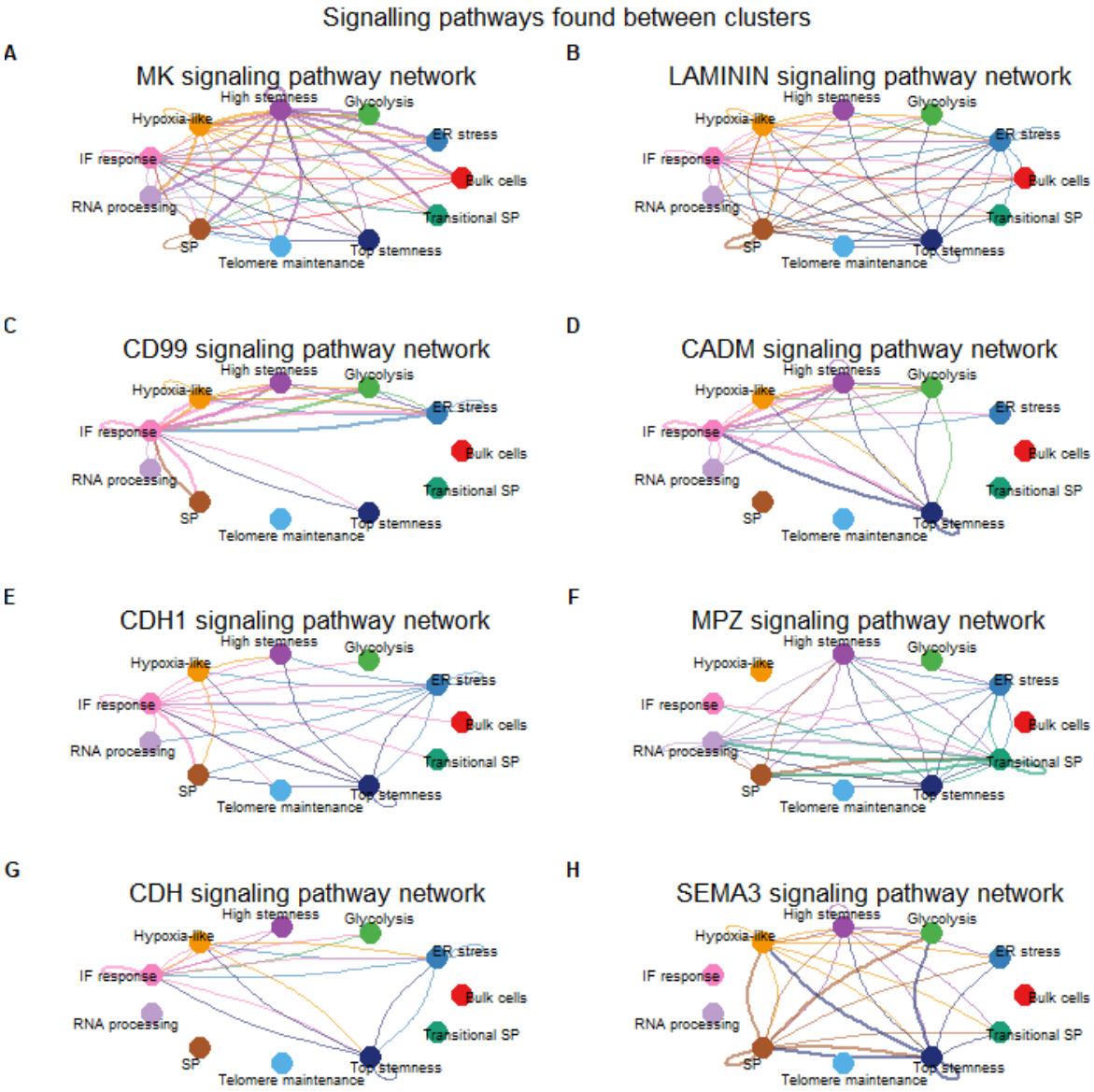


Figure 3.34. The representation of **A)** MK; **B)** laminin; **C)** CD99; **D)** CADM; **E)** CDH1; **F)** MPZ; **G)** CDH and **H)** SEMA3 signalling within clusters and between pairs of clusters.

In addition, the **SP** cluster was involved in ADGRE5 signalling, the **IF response** and **Glycolysis** clusters were linked by tenascin signalling, and NEGR signalling was found to be activated within the **Top stemness** cluster (**Figure 3.35**).

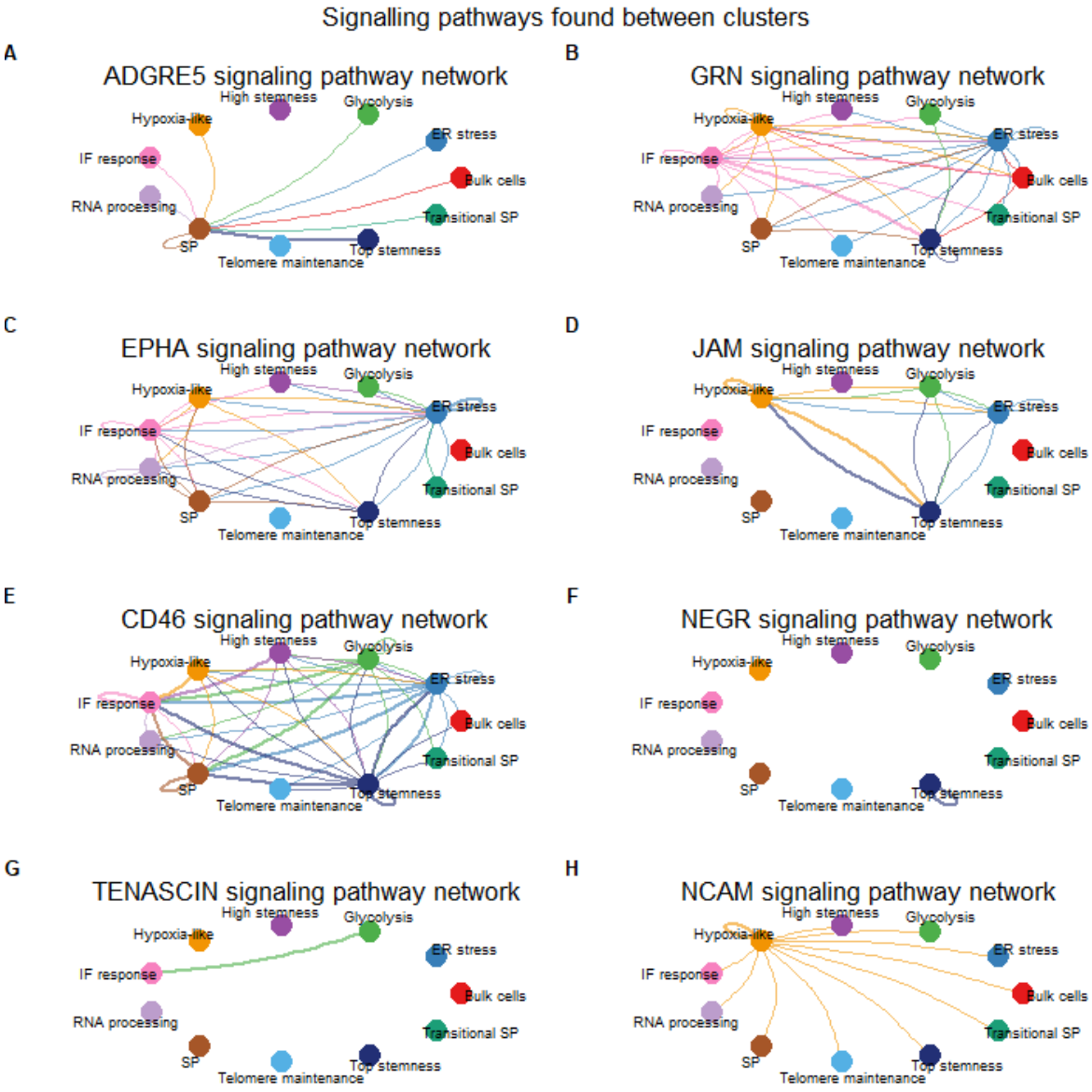


Figure 3.35. The representation of **A)** ADGRE5; **B)** GRN; **C)** EPHA; **D)** JAM; **E)** CD46; **F)** NEGR; **G)** TENASCIN, and **H)** NCAM signalling within clusters and between pairs of clusters.

Five patterns of communications of secreting cells were identified, grouping **Bulk cells** with **IF response**, **High stemness** and **Hypoxia-like** with **Telomere maintenance**, **ER stress** with **Glycolysis**, **RNA processing** with **SP** and **Transitional SP**, while **Top stemness** had its own pattern (**Figure 3.36**). Regarding communications of target cells, three patterns were found, grouping **Glycolysis** and **Hypoxia-like** with **Top stemness**, having **IF response** alone with a second pattern, and a third pattern covering all the other clusters (**Figure 3.37**).

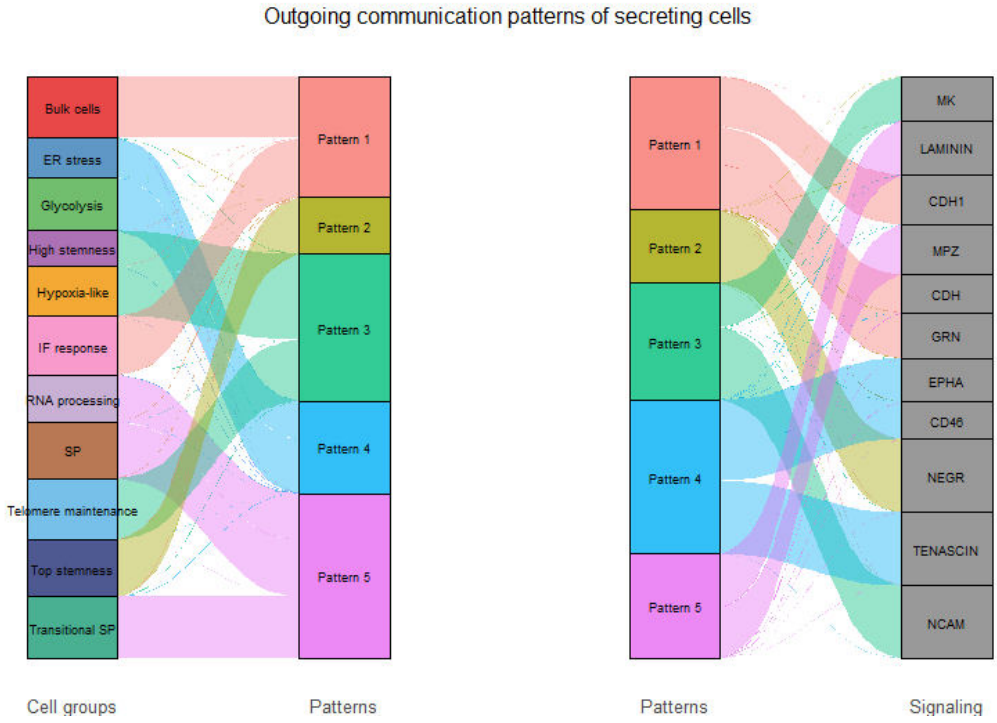


Figure 3.36. The clusters grouped by their outgoing communication patterns of secreting cells, and the pathways corresponding to each pattern.

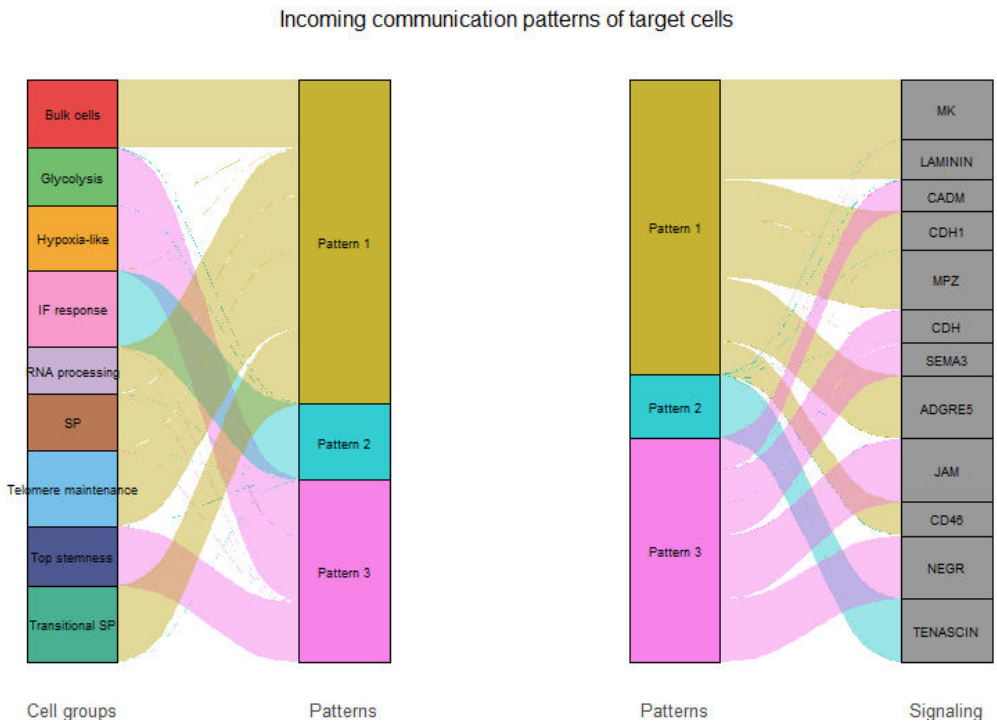


Figure 3.37. The clusters grouped by their incoming communication patterns of target cells, and the pathways corresponding to each pattern.

In **Figure 3.38**, both outgoing and incoming interactions are grouped per clusters and visualised through a dot plot:

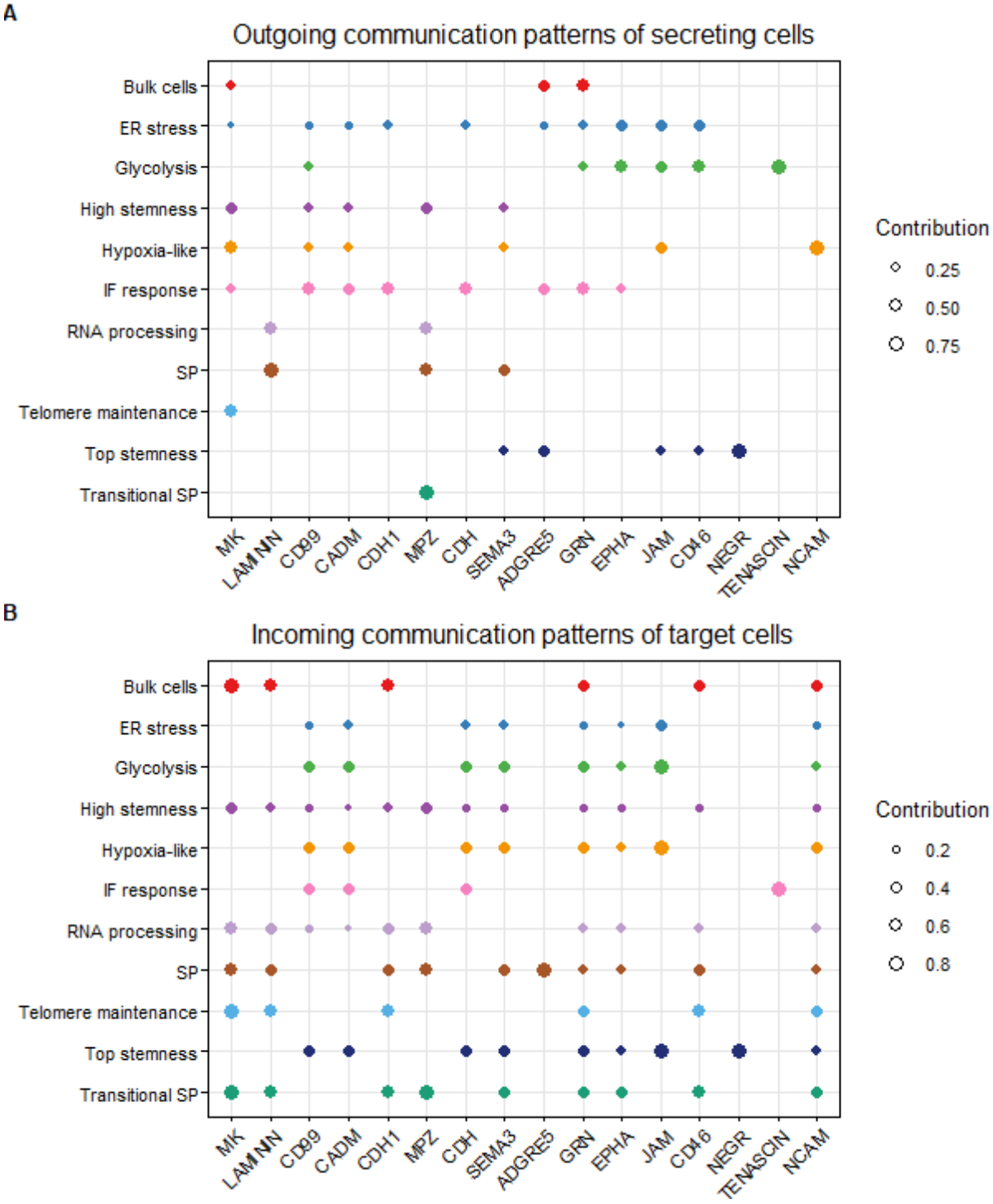


Figure 3.38. Dot plot displaying the activation of **A)** outgoing and **B)** incoming signalling pathways for each cluster.

Overall, the analysis establishes commonalities in cell signalling between the **Transitional SP** cluster and both **RNA processing** and **SP** clusters – its predecessor and successor in the differentiation trajectory, respectively, in terms of both outcoming and ingoing interactions, but also marked discrepancies between the outgoing interactions characterizing the **Top stemness** and **High stemness** clusters.

The next section will discuss the epigenetics processes enriched for the markers of the clusters in order to determine whether epigenetic processes are linked to cancer stemness.

3.2.7 The epigenetic mechanisms characterizing the clusters

120 GO terms linked to epigenetics mechanisms were found in association with the genes detected in the data. 19 were enriched for the cluster markers, of which all were found for the **High stemness** cluster, 5 were found for the **RNA processing** cluster and 4 were found in the **Top stemness** cluster.

Epigenetic pathways were not found to be overrepresented among all enriched GO terms to a statistically significant level in any of the clusters. However, histone genes *H1-4* and *H4C3* are the strongest markers of the **High stemness** cluster, and changes in the expression of epigenetics-related genes, as represented most prominently by these histone genes, characterize the differentiation of **Top stemness** cells into **High stemness** ones. Other epigenetic mechanisms, primarily related to nucleosome organization and chromatin remodeling, are also enriched for the **High stemness** cluster (**Figure 3.39**).

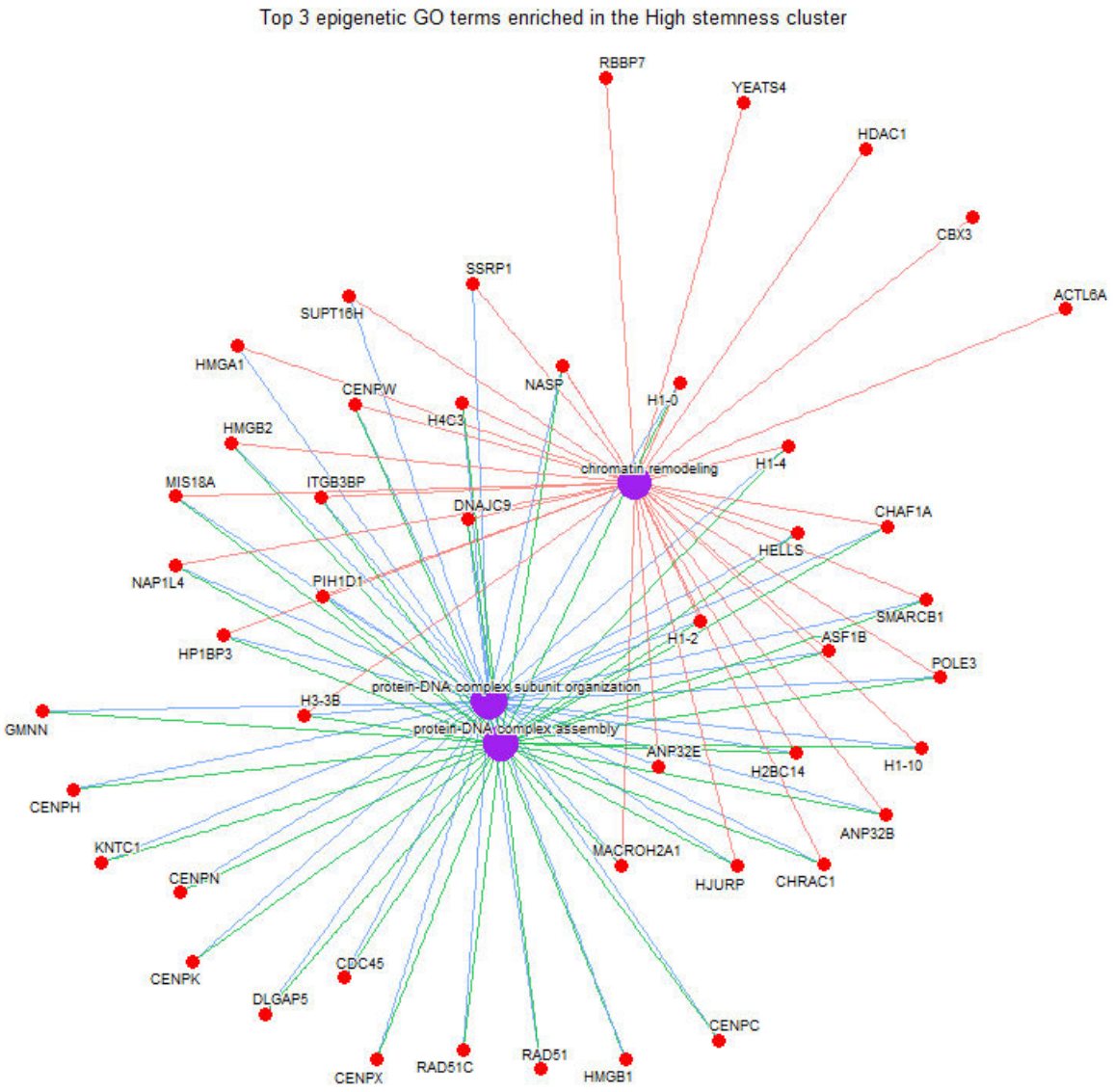


Figure 3.39. Epigenetic GO terms enriched in the High stemness cluster.

The following sections of this chapter will look at the effects of the experimental conditions, starting with assessment of the variation of stemness associations due to the experimental conditions seen in the dataset at the pseudo-bulk level.

3.2.8 The global (pseudo-bulk) comparison of stemness between experimental conditions

As individual genes fitting the traditional "CSC marker" idea did not surface in this dataset, only the TDPM genes sharing an overexpression in a cluster related to stemness, namely *AGR2*, *ALDH1A1*, *REG4* and *TSPAN8* (all overexpressed in **SP**) were still considered for further analysis, as a combination. Their joint density is displayed in **Figure 3.40**:

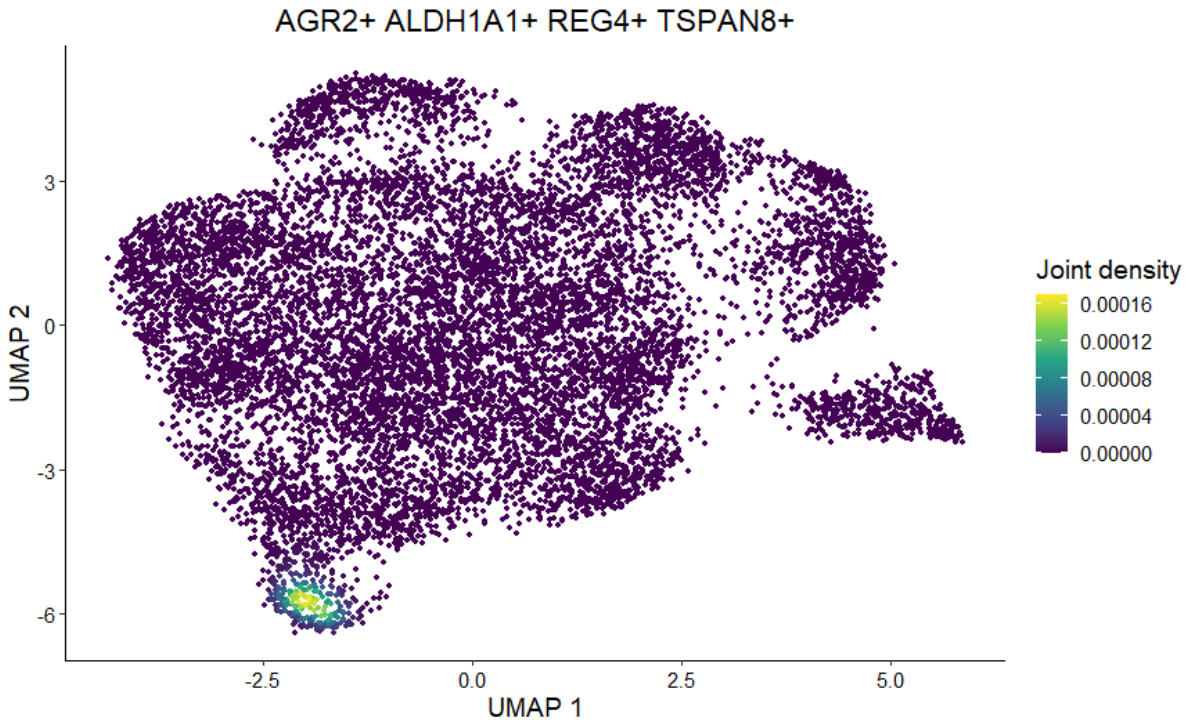


Figure 3.40. Nebulosa plot of the joint density of *AGR2*, *ALDH1A1*, *REG4* and *TSPAN8*.

39 cells in the dataset expressed the *AGR2+ ALDH1A1+ REG4+ TSPAN8+* signature, 11 of them in the Activin A condition, 3 of them in the Activin A and I-BRD9 condition and the remaining 25 in the SB-431542 condition. Among cells expressing this signature, cells from the Activin A and I-BRD9 condition were thus significantly underrepresented (p-value: 4e-03), but significance was not retained when only Activin A and Activin A and I-BRD9 cells were included in the comparison. Thus, no evidence of a significant reduction in the number of cells expressing the *AGR2+*

ALDH1A1+ REG4+ TSPAN8+ signature in Activin A and I-BRD9-treated cells relative to the number of Activin A-treated cells is found.

In order to assess the relation between the treatments and the expression of stemness-linked genes, all the 12 possible selections of one or two versus one or two experimental conditions were assessed for the overrepresentation of stemness-associated genes in this section. The notation **[X] vs. [Y]** will be used throughout the analysis of the effects linked with condition selections, and it signifies that markers of the condition selection X (ident.1 in FindMarkers) were taken relative to condition selection Y (ident.2 in FindMarkers).

For the overlaps between the CCRSA gene sets and the markers of condition selections, the results of significance are displayed in **Table 3.6**.

Cancer gene set	**Condition selections with significant marker overlap**	
	Condition selection	**p-value**
Pancreas	[SB-431542] vs. [Activin A, Activin A and I-BRD9]	9.02e-08
	[SB-431542] vs. [Activin A and I-BRD9]	1.32e-06
	[SB-431542] vs. [Activin A]	1.49e-06
Breast	[SB-431542] vs. [Activin A, Activin A and I-BRD9]	1.86e-03
	[SB-431542] vs. [Activin A and I-BRD9]	1.94e-02
Glioma	[SB-431542] vs. [Activin A, Activin A and I-BRD9]	2.35e-06
	[SB-431542] vs. [Activin A]	2.46e-05
	[SB-431542] vs. [Activin A and I-BRD9]	2.46e-05
Endometrial	[SB-431542] vs. [Activin A, Activin A and I-BRD9]	8.18e-03
	[SB-431542] vs. [Activin A]	1.2e-02
Colon	[SB-431542] vs. [Activin A, Activin A and I-BRD9]	1.01e-05
	[SB-431542] vs. [Activin A]	5.29e-03

	[SB-431542] vs. [Activin A and I-BRD9]	1.26e-02
Union	[SB-431542] vs. [Activin A, Activin A and I-BRD9]	3.99e-18
	[SB-431542] vs. [Activin A and I-BRD9]	2.34e-13
	[SB-431542] vs. [Activin A]	9.38e-13
	[Activin A, SB-431542] vs. [Activin A and I-BRD9]	2.13e-03

Table 3.6. Condition selections whose markers significantly overlap with the CCRSA gene sets.

Thus, no significant underexpression of CCRSA genes in the Activin and I-BRD9 condition relative to the Activin one was recorded, while SB-431542 was associated with increases in stemness associations, rather than decreases.

For the SPLCL genes, statistical significance was not reached for any of the 12 ordered pairs of experimental conditions.

The analysis of the enrichment of treatment conditions among high ORIGINS activity revealed that SB-431542-treated cells were overrepresented among the high activity score cells (p-value: 3.99e-28) with Activin A and I-BRD9 and Activin corresponding to underrepresentations (p-values: 1.74e-08 and 1.2e-06). When considering only Activin and Activin and I-BRD9 cells, no differential representation reached statistical significance, however.

The Wilcoxon rank sum test showed a higher median of activity scores for the SB-431542 condition than for Activin A (p-value: 4.02e-208) and Activin A and I-BRD9 (p-value: 6.67e-166).

To conclude, the SB-431542 condition is associated with higher stemness as per the pseudo-bulk assessment. The next part will analyse the intra-cluster effects of the treatment conditions.

3.2.9 The intra-cluster effects of the treatment conditions upon stemness

This section provides a look at the effects upon cancer stemness in response to treatment in each cluster. The representation of experimental conditions among the cells from each cluster will be analysed, in order to detect any significant treatment-associated reduction or increase of the prevalence of the cells from stemness-linked clusters in response to treatment, then intracluster effects upon cancer stemness will be assessed using the TDPM, the CCRSA and the SPLCL genes, and ORIGINS activity.

All statistically significant overrepresentations of cells from any of the experimental conditions in any of the clusters are listed in **Table 3.7**.

Condition	**Cluster**	**p-value**	**Percentage in cluster**	**Percentage among all cells**
Activin A and I-BRD9	Bulk cells	9.8e-17	40.81	30.17
SB-431542	Transitional SP	1.72e-16	43.47	31.77
SB-431542	High stemness	4.54e-03	36.54	31.77
Activin A and I-BRD9	RNA processing	9.18e-03	33.35	30.17

Table 3.7. The statistically significant overrepresentations of cells from any of the experimental conditions in any of the clusters.

All statistically significant underrepresentations of cells from any of the experimental conditions in any of the clusters are listed in **Table 3.8**:

Condition	**Cluster**	**p-value**	**Percentage in cluster**	**Percentage among all cells**

Activin A and I-BRD9	Transitional SP	1.5e-11	20.89	30.17
SB-431542	Bulk cells	1.27e-10	23.62	31.77
Activin A and I-BRD9	High stemness	1.11e-09	22.17	30.17
Activin A and I-BRD9	Glycolysis	3.41e-03	24.25	30.17

Table 3.8. The statistically significant underrepresentations of cells from any of the experimental conditions in any of the clusters.

When comparing only the Activin and Activin and I-BRD9 conditions, the listed conditions and clusters display an overexpression (**Table 3.9**). An underrepresentation of cells from the Activin and I-BRD9 condition surfaced in the **High stemness** cluster (p-value: 8.12e-07)

Condition	**Cluster**	**p-value**	**Percentage in cluster among compared cells**	**Percentage among all compared cells**
Activin A and I-BRD9	Bulk cells	1.12e-08	53.42	44.21
Activin A	High stemness	8.12e-07	65.07	55.79
Activin A	Transitional SP	1.71e-03	63.05	55.79
Activin A	Glycolysis	3.71e-03	63.44	55.79
Activin A and I-BRD9	RNA processing	1.53e-02	47.9	44.21

Table 3.9. The differential overexpression of the cells from the Activin and I-BRD9 and Activin conditions within clusters.

Among the 39 cells displaying the **TDPM genes**-based *AGR2+ ALDH1A1+ REG+ TSPAN8+* signature (see **Section 3.2.6**), 33 were found in the **SP** cluster: 10 in the Activin A condition, 3 in the Activin A and I-BRD9 and 20 in the SB-431542 condition, with the underrepresentation of Activin A and I-

BRD9 cells reaching statistical significance (p-value: 0.01), which was, however, not retained when only the Activin A and Activin A and I-BRD9 were compared.

The intra-cluster assessment of changes in stemness due to the treatment conditions involving **the CCRSA genes** found results of significance only for the union CCRSA gene set, in all cases involving markers of the SB-431542 condition (**Table 3.10**):

Cluster	**Condition selections with significant marker overlap with the union CCRSA genes**	
	Condition selections	**p-values**
Hypoxia-like	[SB-431542] vs. [Activin A, Activin A and I-BRD9]	0.01
RNA processing	[SB-431542] vs. [Activin A and I-BRD9] [SB-431542] vs. [Activin A, Activin A and I-BRD9]	0.02 0.03

Table 3.10. Condition selections with significant marker overlap with the union CCRSA genes within clusters.

For the **SPLCL genes**, no statistically significant intra-cluster overlaps were recorded.

Next, the intracluster Wilcoxon pairwise activity comparisons of **ORIGINS activity scores** performed for the treatment conditions found a link between the SB-431542 and higher activity scores within each cluster, most significantly within the **Hypoxia-like** cluster for [SB-431542] vs. [Activin A] (adjusted p-value: 3.32e-52; Cohen's d: 0.97). For the **Top stemness** cluster, the variation in activity scores due to the experimental condition was smaller, with the most significant difference (adjusted p-value: 3.13e-04; Cohen's d: 0.45) obtained for [SB-431542] vs. [Activin A and I-BRD9]. Of note, the [Activin A] vs. [Activin A and I-BRD9] grouping did reach significance in the **High stemness**, **Top stemness** and **SP** clusters. The complete results are illustrated in **Figure 3.41**:

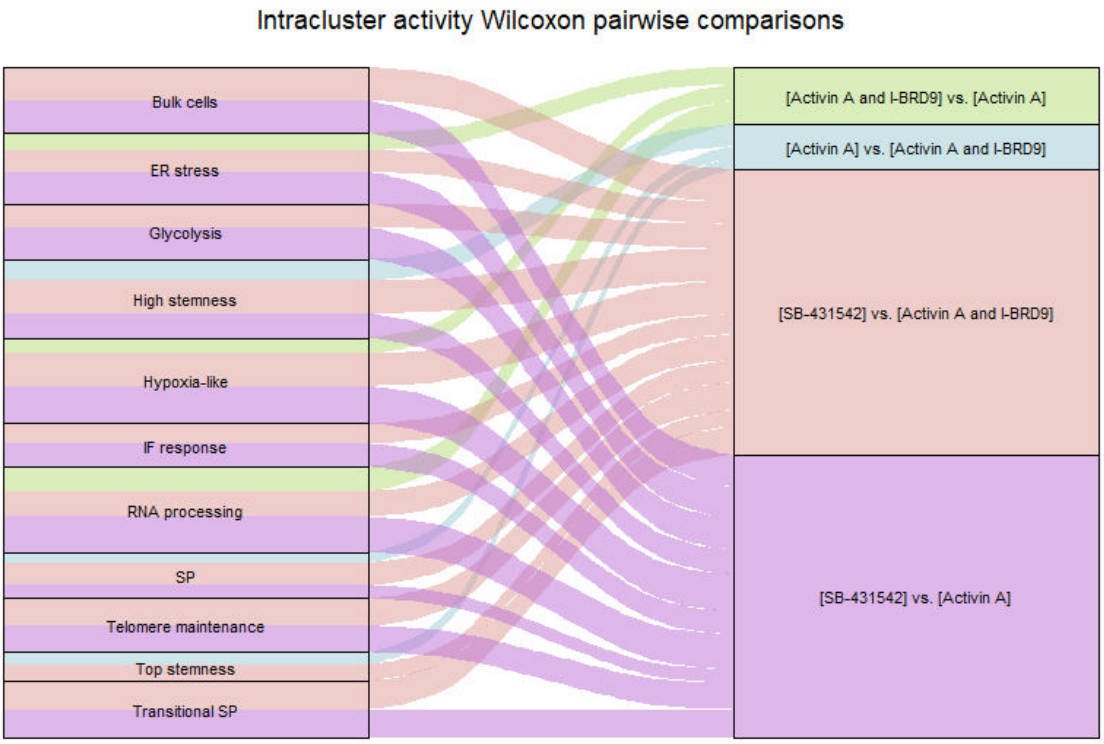


Figure 3.41. The results of Wilcoxon pairwise activity comparisons between clusters. Thicker connecting lines correspond to lower p-values.

In conclusion, cells from the Activin A and I-BRD9 are underrepresented relative to the cells from the Activin A condition in the **High stemness** cluster, and their ORIGINS activity scores within the **Top stemness**, **High stemness** and **SP** clusters are reduced to a statistically significant, if limited, extent. The SB-431542 condition shows again the most frequent associations with stemness.

The next section will examine the commonalities between the genes and the GO terms associated with the clusters and the condition selections.

3.2.10 The overlaps of markers of clusters and markers of condition selections

In this section, the effects of the treatment conditions upon the functional types of cells found in the data are uncovered by evaluating the significance of the overlaps of markers of clusters and

those of condition selections. The same assessment is then performed for the enriched GO terms of these markers.

The most significant overlap between cluster markers and selection markers was registered for the **High stemness** cluster and the [SB-431542] vs. [Activin A, Activin A and I-BRD9] selection (adjusted p-value: 1.98e-154; Jaccard score: 0.22). For the **Top stemness** cluster, the strongest overlap was found for the [SB-431542] vs. [Activin A] selection (adjusted p-value: 1.6e-24; Jaccard score: 0.08). For the [Activin A] vs. [Activin A and I-BRD9] selection, a significant overlap is registered with the markers of the **High stemness** cluster. The top 20 significant overlaps are showcased in **Figure 3.42**:

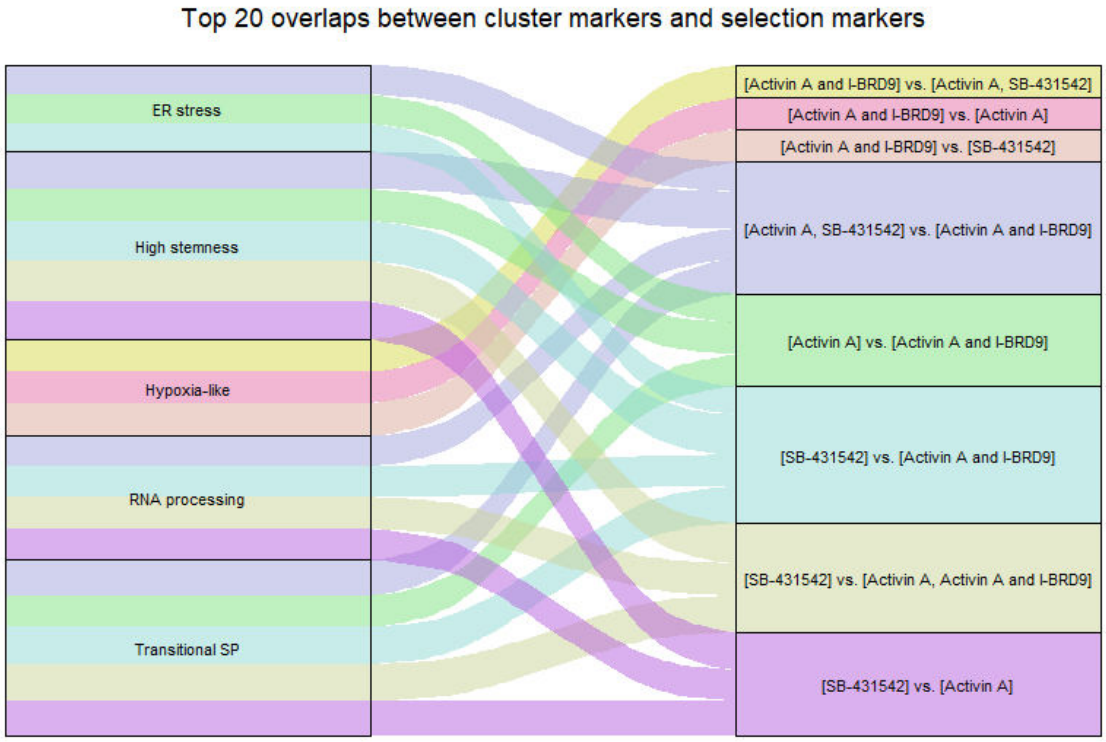


Figure 3.42. The top 20 overlaps recorded between cluster markers and selection markers.

Sizeable overlaps between the GO terms enriched for the **High stemness** and those enriched for the markers of SB-431542 condition likewise emerged, together with overlaps for the same condition and the markers of the **Top stemness**, **RNA processing** and **Transitional SP** clusters. No

indications of differential associations between the GO terms enriched for the markers of the Activin A and Activin A and I-BRD9 conditions and the markers of any of the stemness clusters surfaced. The most significant overlap of enriched GO terms was recorded for the **High stemness** cluster and the [SB-431542] vs. [Activin A, Activin A and I-BRD9] condition (adjusted p-value: 7.33e-137; Jaccard score: 0.47). For the **Top stemness** cluster, the most significant overlap was registered for the same condition selection (adjusted p-value: 7.33e-137; Jaccard score: 0.22). The top 20 overlaps are displayed in **Figure 3.43**:

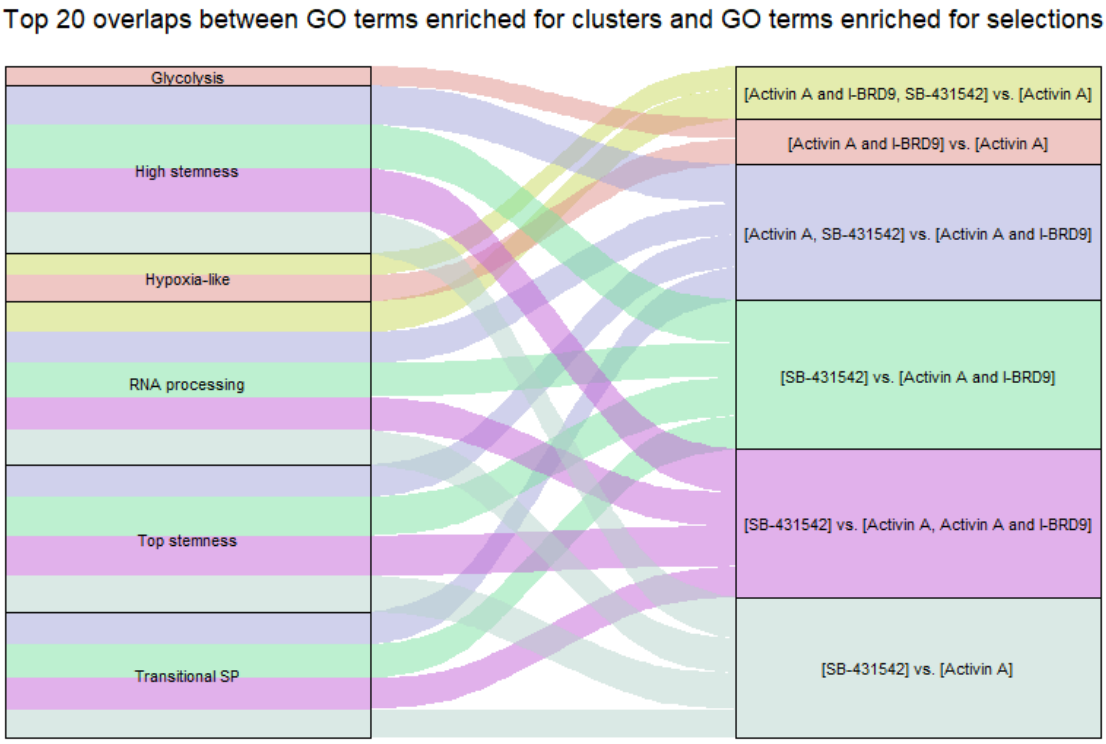


Figure 3.43. The top 20 overlaps recorded between GO terms enriched for cluster markers and selection markers.

In conclusion, this section further supports the links between increased stemness and the SB-431542, condition, while little variation in stemness can be reliably adjudged between the other two experimental conditions. In the next section, three-way overlaps in which the selection markers and cluster markers are assessed against the CCRSA genes, followed by SPLCL genes, will be analysed.

3.2.11 The overlaps of markers of clusters, markers of condition selections and stemness-linked gene sets

This section assesses the impact of the stemness-linked gene sets upon the overlaps between the markers of clusters and the markers of condition selections, in order to evaluate whether the effects seen in stemness-linked clusters can be attributed to a downregulation of stemness-linked genes.

All the significant three-way overlaps between cluster markers, selection markers and CCRSA sets involved the SB-431542 condition, and one of the **Top stemness**, **High stemness** and **Telomere maintenance** clusters (**Figure 3.44**, **Figure 3.45** and **Figure 3.46**).

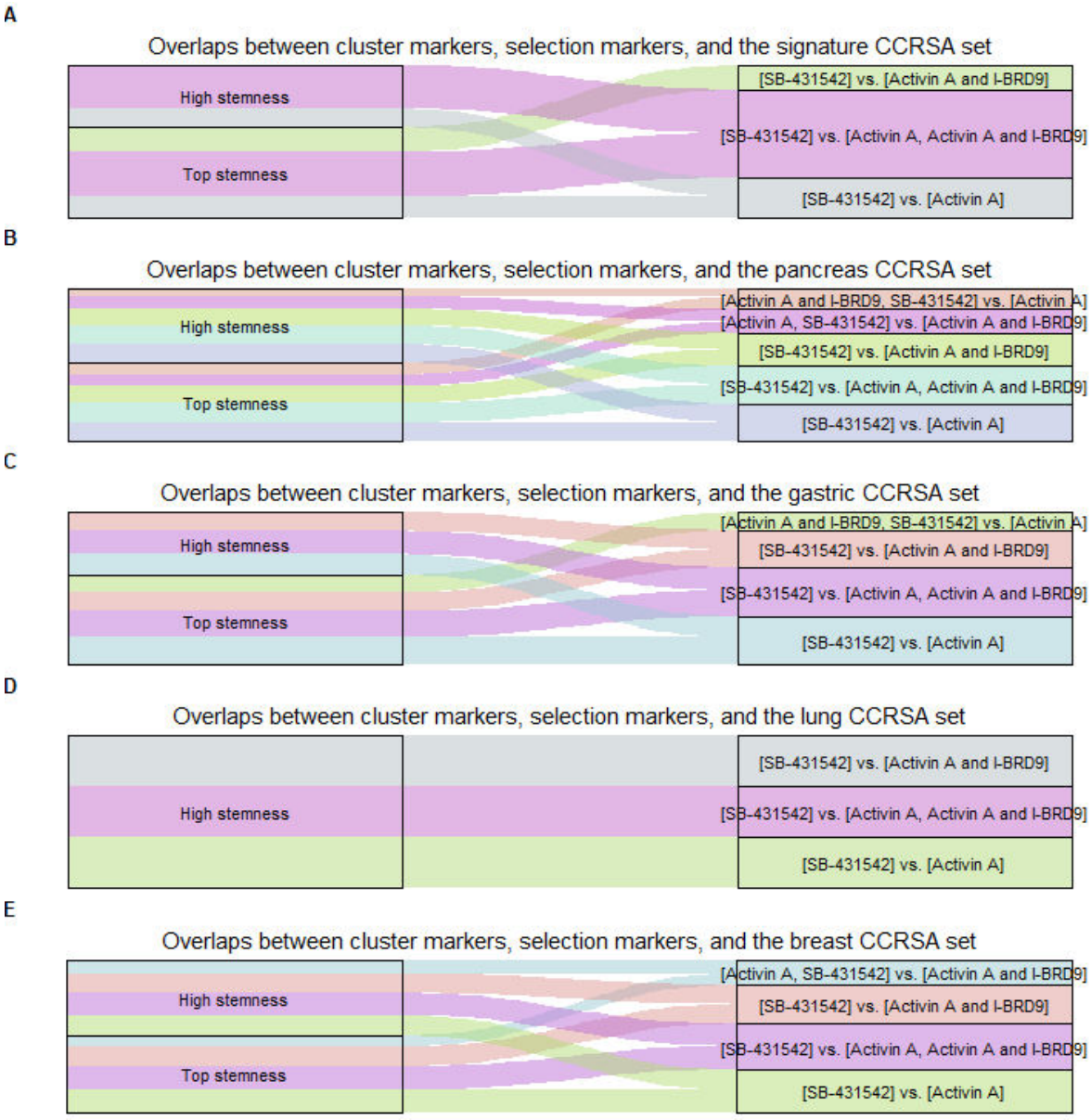
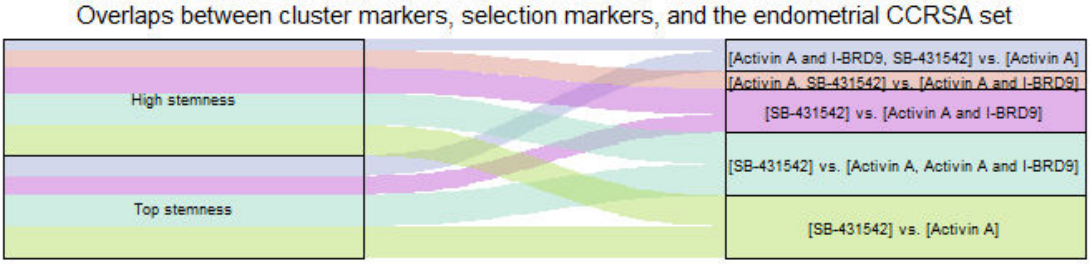


Figure 3.44. Significant three-way overlaps between cluster markers, selection markers, and the signature **(A)**, pancreas **(B)**, gastric **(C)**, lung **(D)**, breast **(E)** CCRSA gene sets. Thicker lines connecting cluster and condition selections correspond to lower p-values.

A

B



C

D

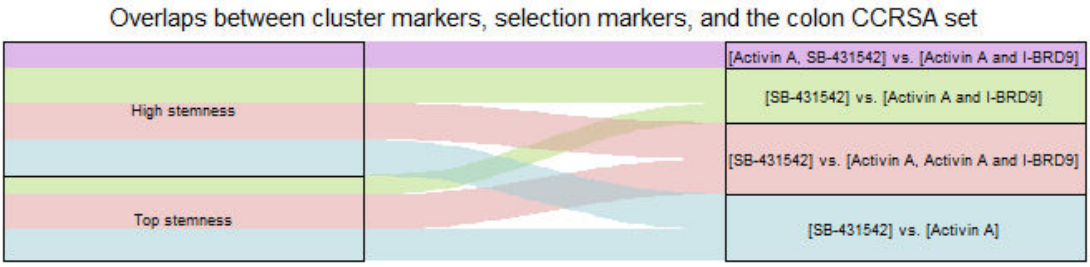


Figure 3.45. Significant three-way overlaps between cluster markers, selection markers, and the glioma **(A)**, endometrial **(B)**, bladder **(C)** and colon **(D)** CCRSA gene sets. Thicker lines connecting clusters and condition selections correspond to lower p-values.

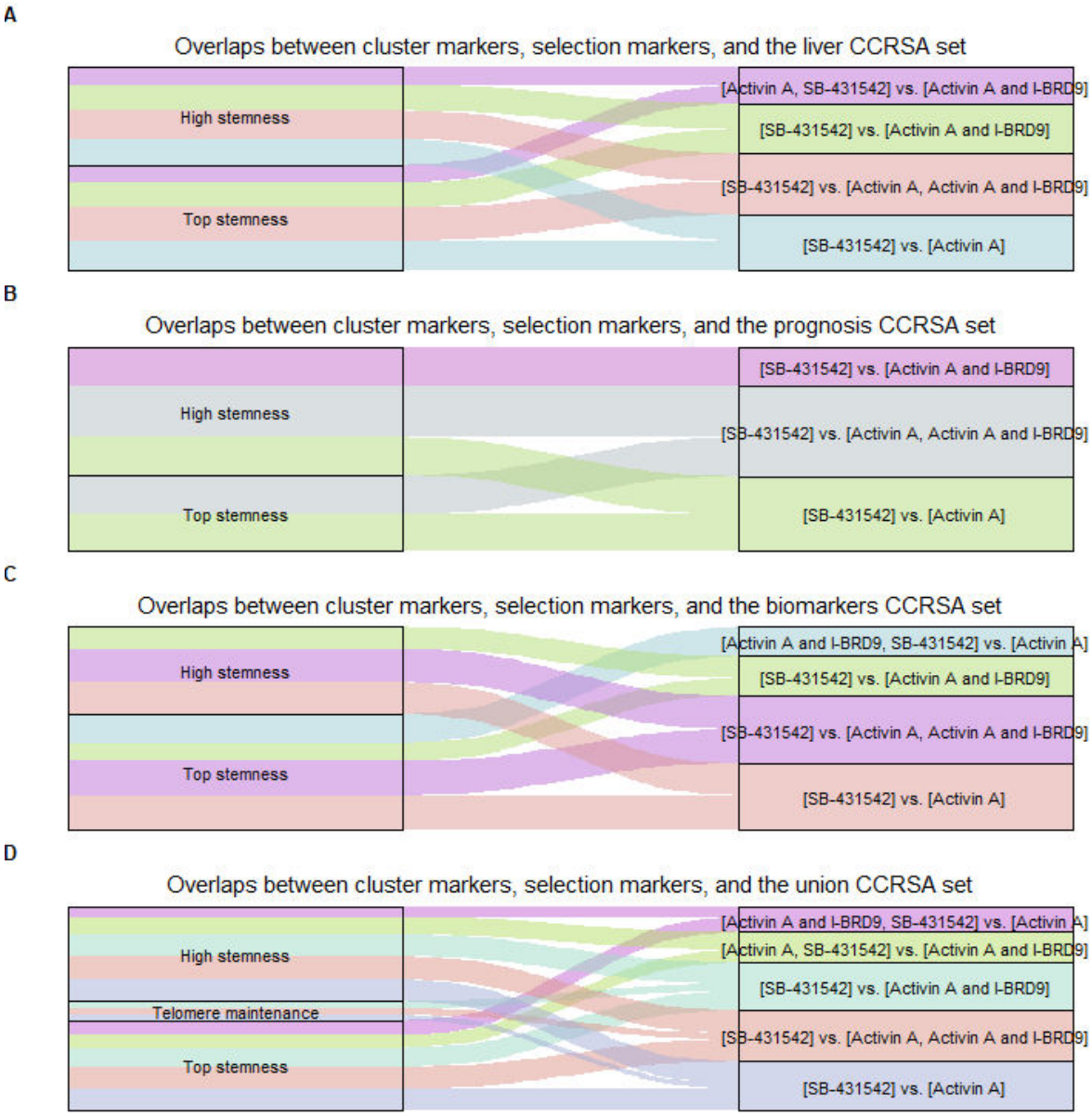


Figure 3.46. Significant three-way overlaps between cluster markers, selection markers, and the liver **(A)**, prognosis **(B)**, biomarkers **(C)** and union **(D)** CCRSA gene sets. Thicker lines connecting clusters and condition selections correspond to lower p-values.

For **the SPLCL genes**, the three-way overlap assessment resulted in significance for only in the **SP**,

Transitional SP and **High hypoxia** clusters (**Figure 3.47**):

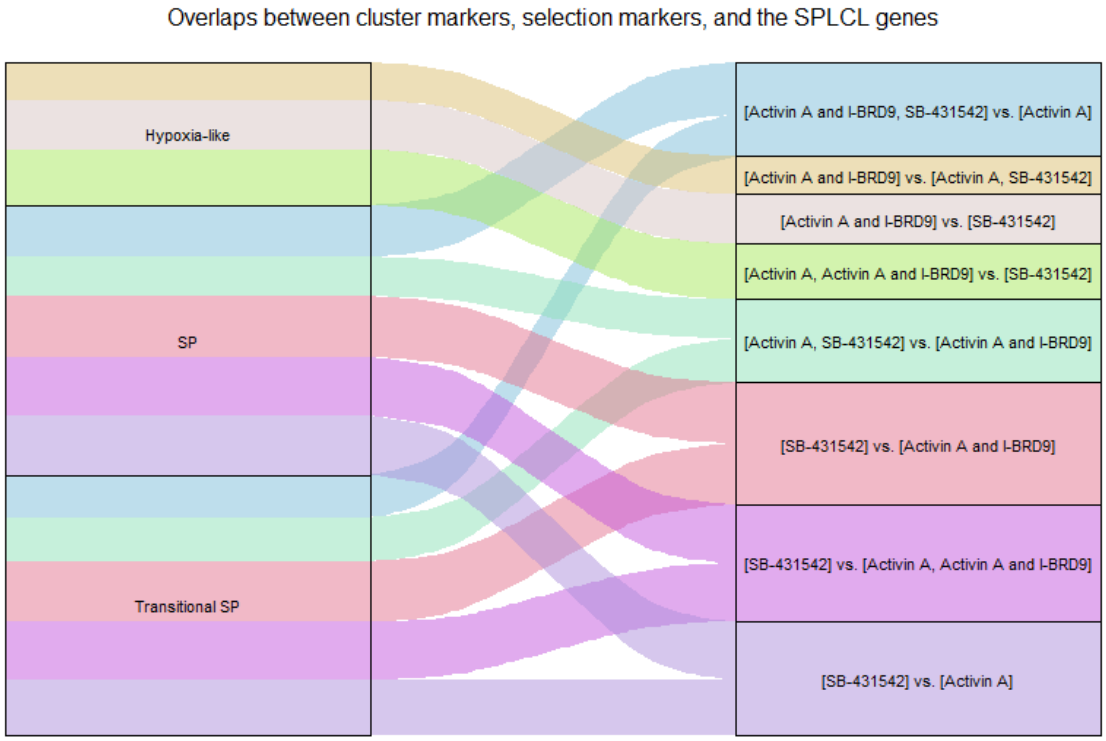


Figure 3.47. The statistically significant overlaps between the markers of clusters, the markers of condition selections, and the SPLCL genes.

SB-431542 markers accounted for the overlaps seen with the **SP** and **Transitional SP** cluster markers, but the opposite effect was seen in the **High hypoxia** cluster.

Overall, the results showed that the previously noted overlaps between the markers of stemness-related clusters and the SB-431542 condition are significantly driven by stemness-linked gene sets, evidencing further the links between the SB-431542 condition and stemness in this dataset.

The next section will provide a discussion of the findings obtained in this chapter.

3.3 Discussion

A **Top stemness** population with markers strongly overlapping with the CCRSA gene sets, having a high development potential as per ORIGINS activity, placed at the origin (highest stemness) of all

three lineages detected with Slingshot, and having significant overlaps with markers of cancer stemness in various cancers as per two Enrichr databases, is identified in the dataset.

Of note, the **Top stemness** population is *not* characterized also by the overexpression of single-gene traditional markers – which prove unable to identify this population. In general, the 28 traditionally derived putative PCSC markers correlate poorly with each other in terms of expression, and identify only minimally overlapping cell populations, an exception being the *ALDH1A1+ REG4+ TSPAN8+ AGR2+* marker signature, which links well with the **SP** population, characterized by its marker overlap with the L3.6 side population genes. This population also scored high in terms of ORIGINS activity. However, as per the Slingshot trajectory inference assessment, **SP** is not a bona fide "CSC" population, being associated with a terminal cluster in one of the lineages, rather than an initial one.

An interpretation of the findings is that cells can regain some stem-like properties at an advanced stage in their differentiation from CSC. For instance, *AKR1B10*, a strong marker of **SP** cells in the dataset, was found overexpressed in multiple cancer types and precancerous lesions[229]. Interestingly, *CEACAM6*, another strong marker of the **SP** cells but not one of the SPLCL genes, and in addition a promoter of platinum resistance[230], was recently evidenced as a marker of quiescent residual cancer cells driving chemoresistance and tumour recurrence[231]. Thus, it appears bona fide CSC are not the only cancer cell subpopulations that showcase tumour initiation properties and distinctively enhanced drug resistance abilities, but rather the findings support the idea that these traits can be developed at later stages of the differentiation trajectory. Therefore, the disagreement in the literature about different sets of CSC markers being reported in the literature may be due to the existence of multiple classes of cell populations that show key features associated with stemness (most prominently, tumorigenesis and drug resistance).

Thus, targeting putative CSC identified using few or single-gene signatures is unlikely to eliminate CSC. Proposed CSC identified in such a manner might not, in fact, lie in the apex of the differentiation trajectory, but represent cellular entities more closely corresponding to the **SP** cluster.

At the same time, it does not appear that fully eliminating cellular types resembling the **Top stemness** cluster described in this analysis will be curative in and of itself, because other cell populations might display enhanced drug resistance, while still retaining enough tumourigenesis potential that would make them able to regenerate the whole tumour.

When understood as a gradient, the variation in stemness between the cells in this dataset was found to be highly influenced by cell cycle related-genes, markedly overrepresented among the genes whose expression varies concomitantly with ORIGINS activity and Slingshot pseudotime scores.

Contrary to previous findings on the role of TGF-β/Activin/NODAL signaling in PCSC[88], its inhibition via SB-431542 was associated with a limited, but consistent over the lines of evidence that were analysed, increase in cancer stemness relative to the Activin A treatment conditions. This is explainable through the lens of the fact that the CSC-like cells identified in this dataset are proliferative rather than quiescent, hence Activin A treatment, which inhibits DNA replication, does not result in increases in stemness. In addition, the inconsistency can also be explained through the identification of CSC in the works cited above, performed using putative CSC markers for which no evidence of links with stemness surfaced in this dataset.

The 24 h Activin A and I-BRD9 treatment did not result in sizeable decreases in stemness, relative to the Activin A treatment.

The results also show that a multitude of epigenetic mechanisms (e.g. related to DNA packaging, chromatin remodelling, nucleosome assembly) are enriched in the **High stemness** cluster, the

second cluster in the differentiation trajectory in all the three Slingshot lineages. A few epigenetic processes are also enriched for the **Top stemness** population. For the **SP** cells, a population matching several descriptions of cancer stemness, but without being at the origin of the differentiation trajectory in this dataset, the findings do not relay a vulnerability to epigenetic inhibitors.

Of note, among the strongest markers of the **Top stemness** cluster, *AURKA* has been implicated in promoting EMT by regulating histone modifications through Wnt/β-catenin and PI3K/Akt signaling in gastric cancer[232], and the same role was noted for *TOP2A* in cervical cancer, again through PI3K/Akt signaling[233], and for *ASPM* in non-small cell lung cancer, through Wnt/β-catenin signaling[234]. In addition, *AURKA* induces expression of typically hypoxia-induced genes in normoxic conditions (pseudohypoxia) in breast cancer, where it also drives early metastasis[235].

The next chapter evaluates the effects of I-BRD9 for a longer treatment duration (72 h), with standard-of-care PDAC drug gemcitabine also included in the experimental conditions – a compound for which CSC-driven chemoresistance is a noted hurdle seriously impeding effective therapy.

CHAPTER 4

CHARACTERIZATION OF CANCER STEM CELLS IN SINGLE-CELL RNA-SEQUENCING DATA IN PATIENT PDAC CELLS

4 Characterization of cancer stem cells in single-cell RNA-sequencing data in patient PDAC cells

4.1 Introduction

This chapter describes the results of a single-cell RNA-sequencing experiment performed in PDAC cells collected from a 50-year-old female patient, under four different treatment conditions: DMSO, gemcitabine, I-BRD9, I-BRD9 and gemcitabine. The treatment duration was 72 h for each condition. The identification and characterization of subpopulations (clusters) in the datasets was performed, with a focus on discerning cell groups ranking high in cancer stemness and on examining the trajectory of differentiation, followed by an assessment of the differential effects of the experimental conditions upon these subpopulations. The choice of experimental conditions aimed to assess the effects of I-BRD9 upon CSCs for a longer treatment time than the one employed previously (24 h), and to study the effects of gemcitabine upon identified CSCs, in the light of the previously documented resistance of CSC to this drug[236], a possible explanation of therapy failure.

Section 4.2.1 covers quality control. Non-tumour cells are identified and removed in **Section 4.2.2**. **Section 4.2.3** covers normalization, integration, dimensionality reduction and clustering. Clusters associated with stemness are identified in **Section 4.2.4**. A functional characterization of all the clusters is performed in **Section 4.2.5**. In **Section 4.2.6**, trajectory analysis and RNA velocity analysis are performed, followed by the inference of cell-cell communication in **Section 4.2.7**. The epigenetics processes enriched for the markers of clusters are identified in **Section 4.2.8**.

Starting with **Section 4.2.9,** the differential effects of the treatment conditions are evaluated. In **Section 4.2.9**, the effects of the treatment conditions are assessed at the global (pseudobulk) level. **Section 4.2.10** covers the intracluster effects of the treatment conditions. In order to determine whether the treatment conditions specifically affect stemness-linked clusters, overlap

assessments between the genes and processes characterizing the clusters and the experimental conditions, respectively, are performed in **Section 4.2.11**. The extents to which stemness-linked gene sets account for the gene overlaps identified previously is determined in **Section 4.2.12**, and a discussion of the results obtained in this chapter is provided in **Section 4.3**.

4.2 Results

4.2.1 Quality control

Genes expressed in very few cells (< 10) were removed. Next, doublets were identified using scDblFinder and then removed, followed by the removal of low-quality cells, having a very low complexity (novelty) score, a very high percentage of mitochondrial or ribosomal genes, a very low or very high percentage of number of counts or UMIs per cells, or a very low Shannon or Simpson diversity.

100 scDblFinder runs predicted an average of 2429.06 doublets, best approximated by a doublet prediction significance cutoff of 44 runs, resulting in 2435 predicted doublets (**Figure 4.1**).

The Jaccard similarity scores (defined in **Section 2.3.2**) registered between all the pairs of scDblFinder runs ranged between 0.68 and 0.79, while the Jaccard scores taken with the consensus prediction ranged between 0.75 and 0.83.

The mean of the Jaccard scores between the consensus predictions and each of the scDblFinder runs was 0.81, above the mean of the maxima of all the Jaccard scores between any two pairs of runs (0.77).

The results of the calculations of Jaccard similarity scores between scDblFinder runs and the consensus prediction are illustrated in **Figure 4.2**. Purple up-triangles represent Jaccard scores taken with the consensus prediction, while the maxima, means and minima of the Jaccard

similarities between each prediction and the other 99 are displayed with navy squares, medium blue circles and light blue triangles, respectively. Dashed lines represent group averages.

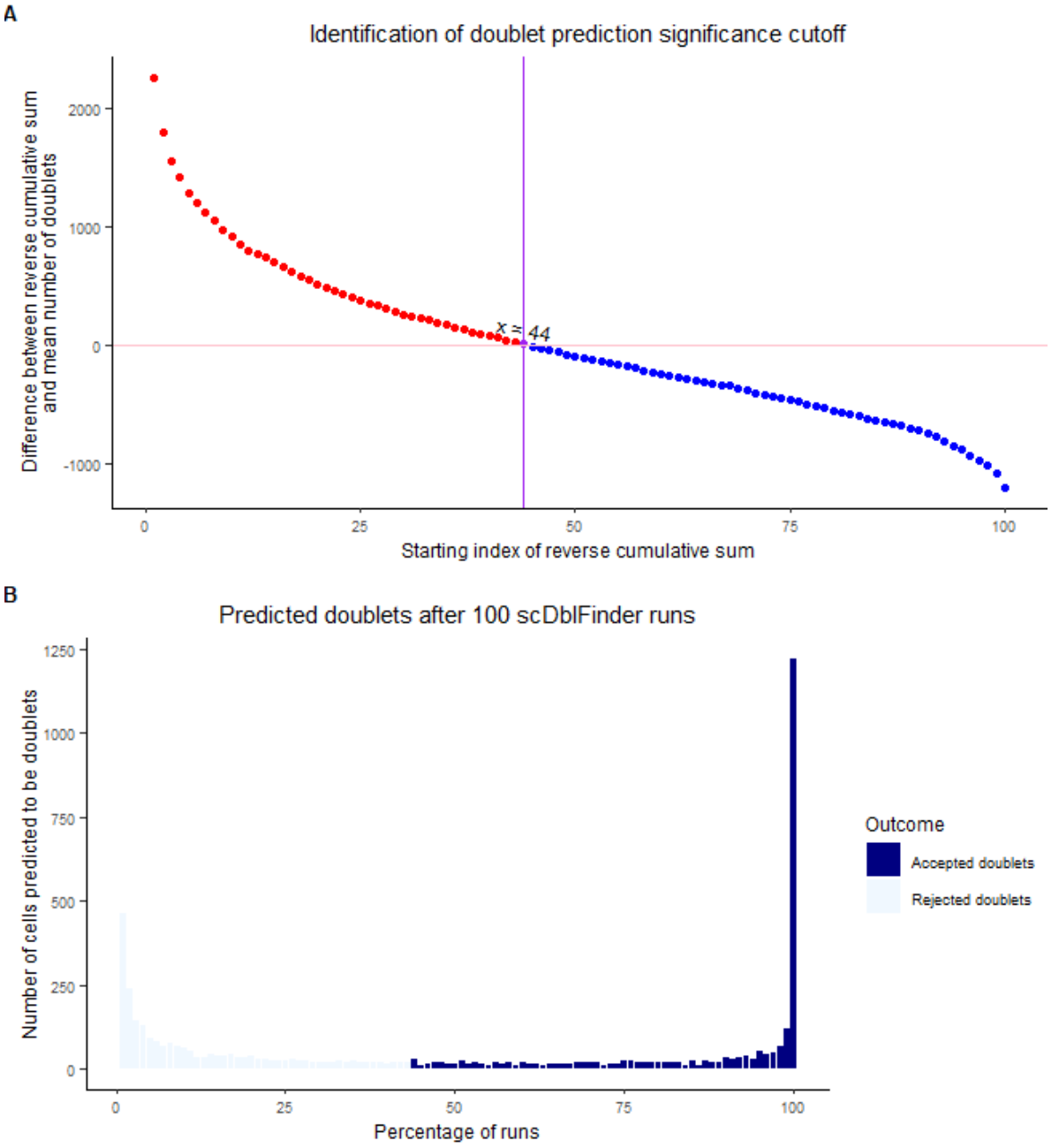


Figure 4.1. **A)** Identification of the doublet prediction cutoff. **B)** The distribution of accepted and rejected doublets.

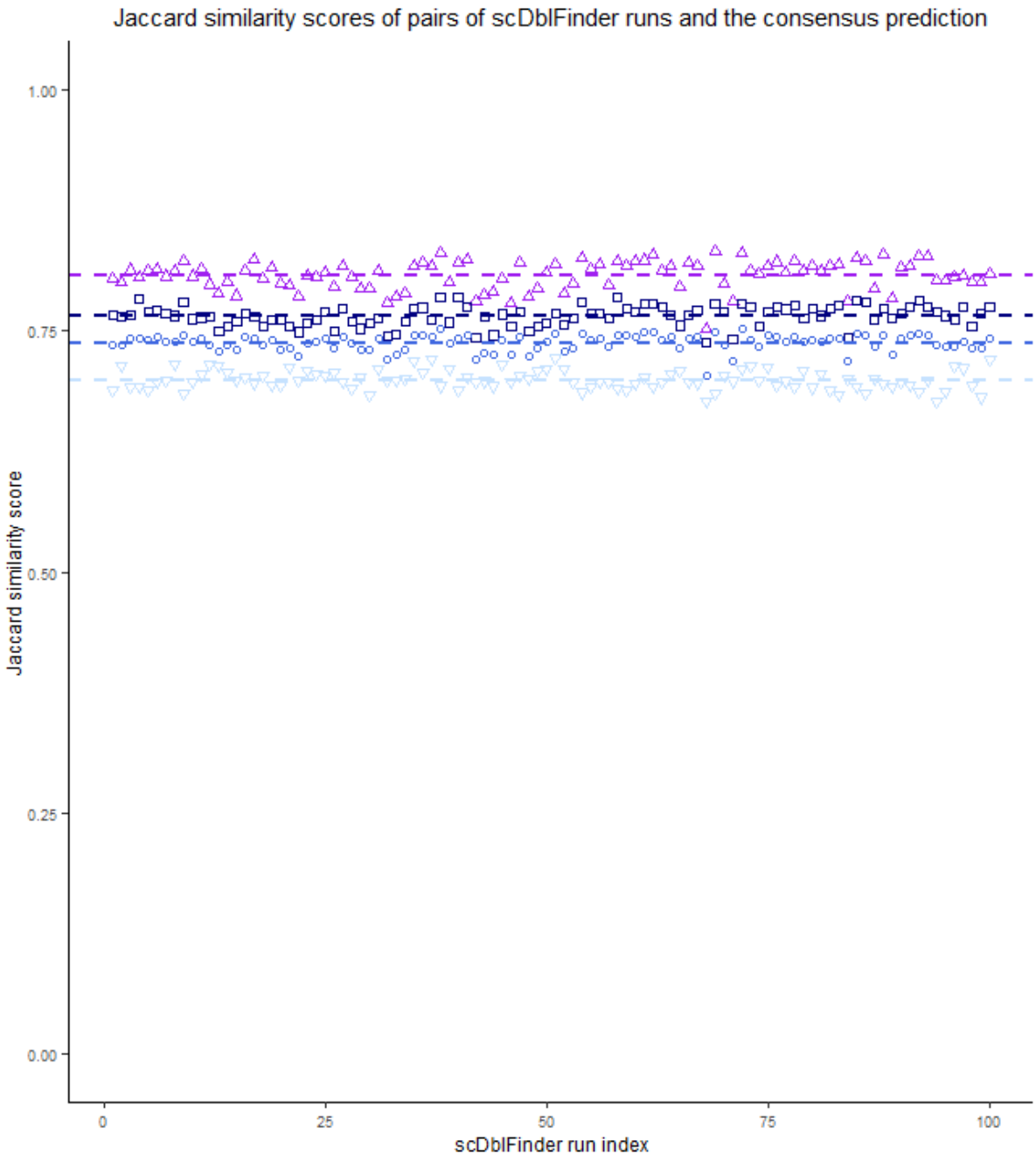


Figure 4.2. Jaccard similarity scores between each prediction and the consensus prediction, and between each prediction and the maxima, means and minima of the other 99.

Next, the cells were filtered based on the other criteria listed in **Section 2.3.2**. The histograms in **Figure 4.3**, **Figure 4.4**, **Figure 4.5** and **Figure 4.6** illustrate the distributions of the filtering variables across the experimental conditions. On each plot, the section between the two blue dotted vertical lines represents the cells retained in the dataset at each filtering step.

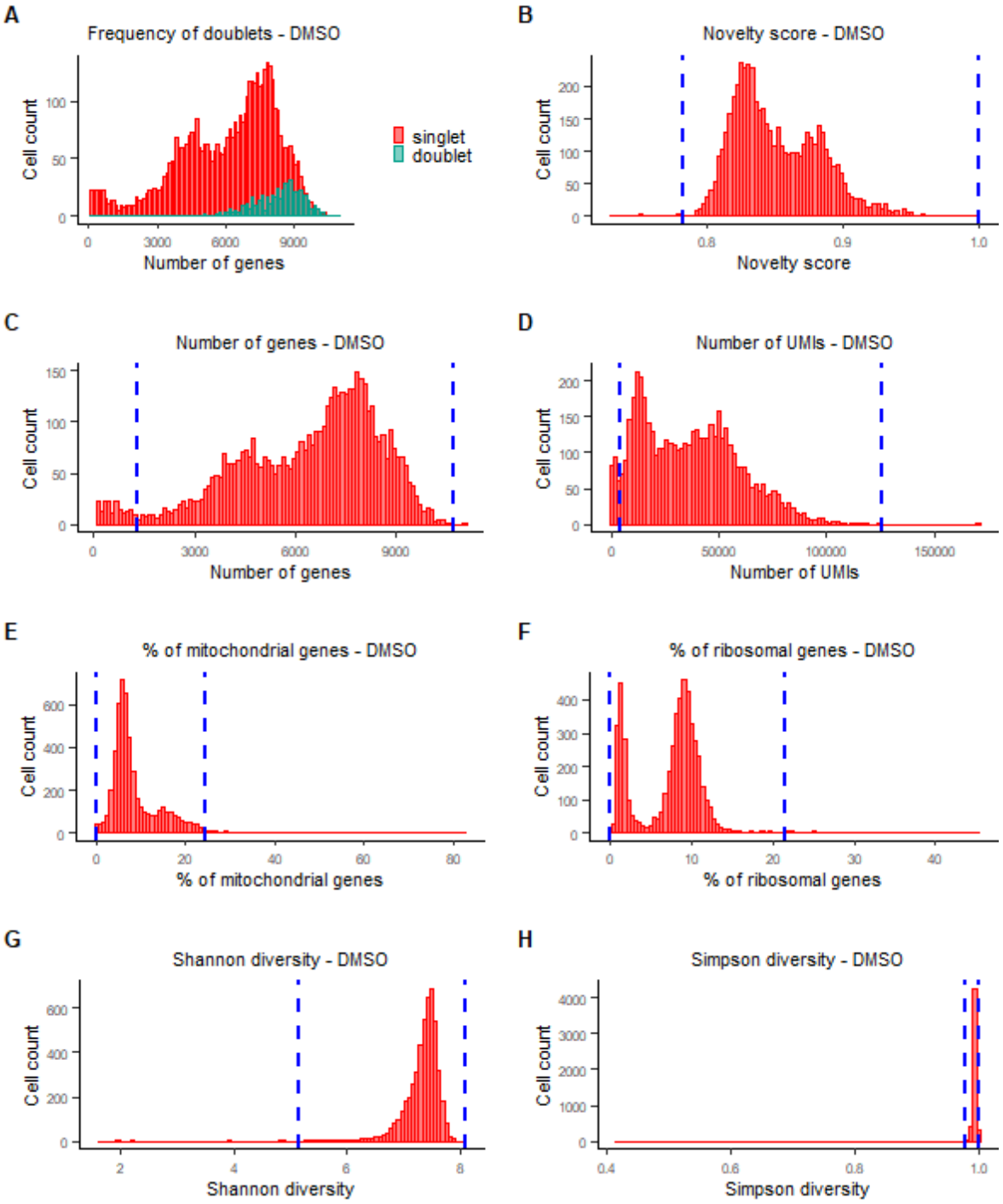


Figure 4.3. Quality control selection criteria for cells in the DMSO condition: **A)** Singlet status as predicted by scDblFinder; **B)** Novelty score (log10 (nFeature_RNA) / log10 (nCount_RNA)) above 0.78; **C)** Number of detected genes (nFeature_RNA) between 1300 and 10700; **D)** Number of UMIs (nCount_RNA) between 4000 and 126000; **E)** Percentage of mitochondrial genes below 24.6%; **F)** Percentage of ribosomal genes below 21.5%; **G)** Shannon diversity above 5.15; **H)** Simpson diversity above 0.98.

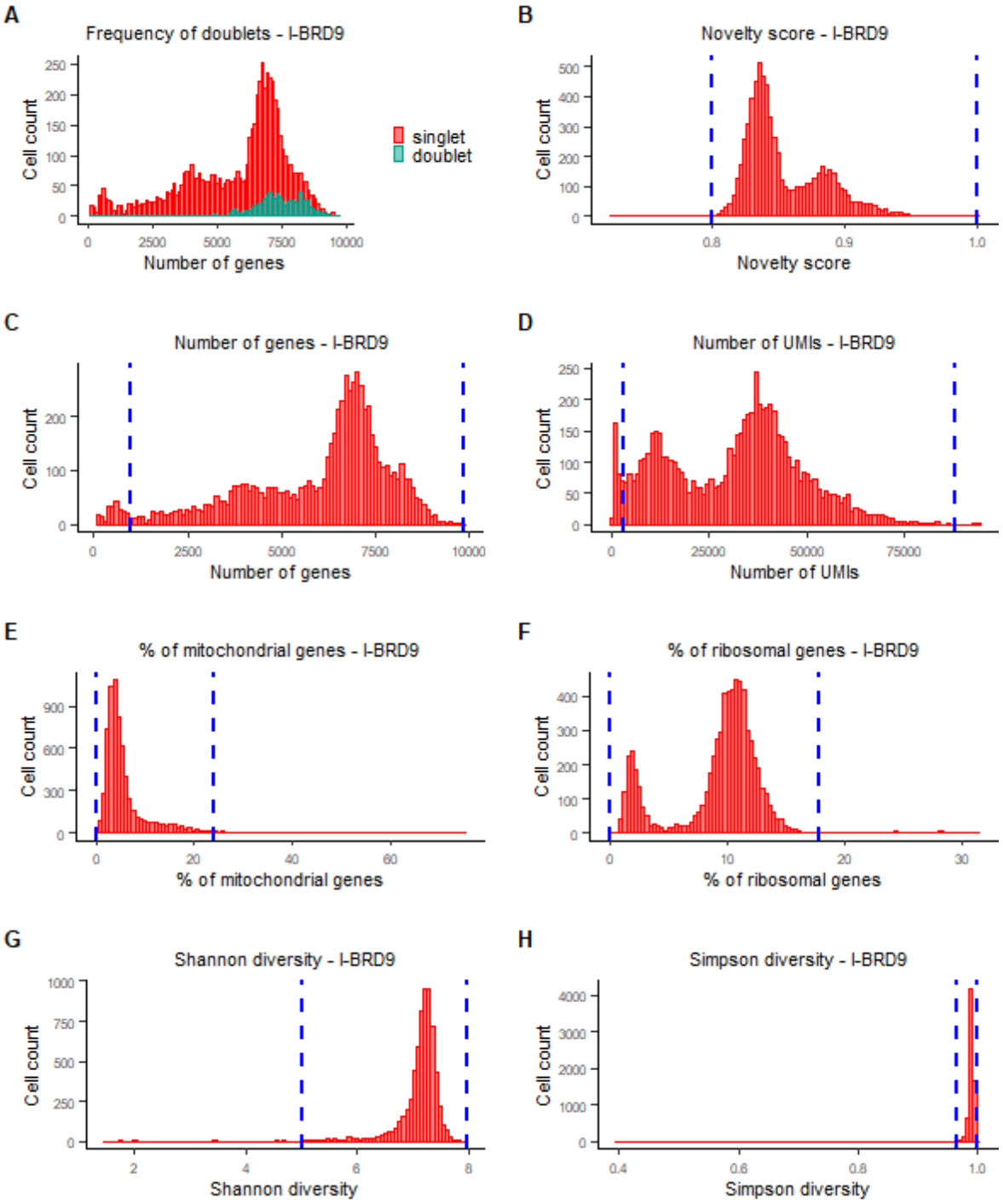


Figure 4.4. Quality control selection criteria for cells in the I-BRD9 condition: **A)** Singlet status as predicted by scDblFinder; **B)** Novelty score (log10 (nFeature_RNA) / log10 (nCount_RNA)) above 0.8; **C)** Number of detected genes (nFeature_RNA) between 1000 and 9900; **D)** Number of UMIs (nCount_RNA) between 3400 and 88000; **E)** Percentage of mitochondrial genes below 24%; **F)** Percentage of ribosomal genes below 17.8%; **G)** Shannon diversity above 5; **H)** Simpson diversity above 0.96.

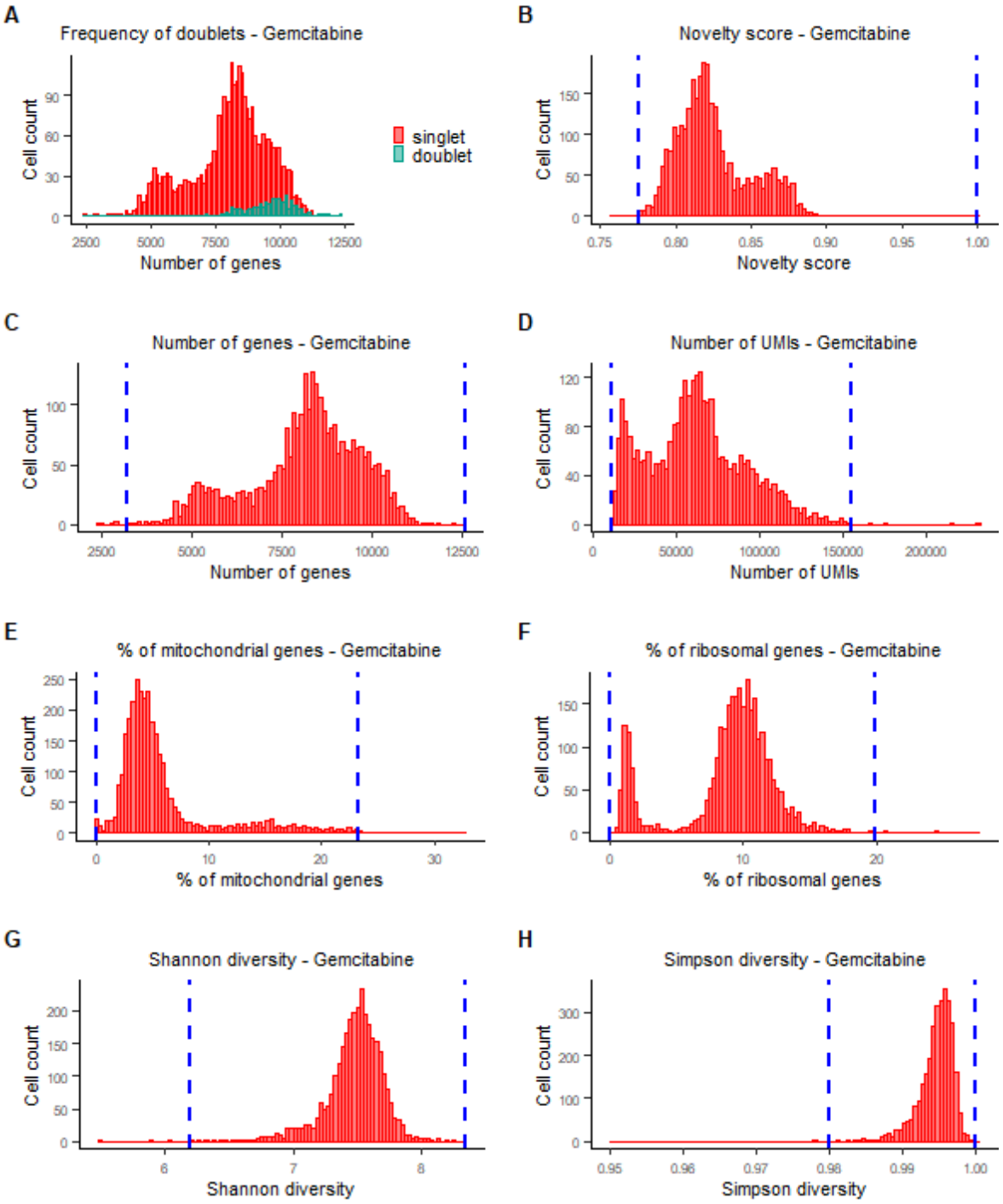


Figure 4.5. Quality control selection criteria for cells in the Gemcitabine condition: **A)** Singlet status as predicted by scDblFinder; **B)** Novelty score (log10 (nFeature_RNA) / log10 (nCount_RNA)) above 0.78; **C)** Number of detected genes (nFeature_RNA) between 3200 and 12600; **D)** Number of UMIs (nCount_RNA) between 11000 and 155000; **E)** Percentage of mitochondrial genes below 23.2%; **F)** Percentage of ribosomal genes below 19.8%; **G)** Shannon diversity above 6.2; **H)** Simpson diversity above 0.98.

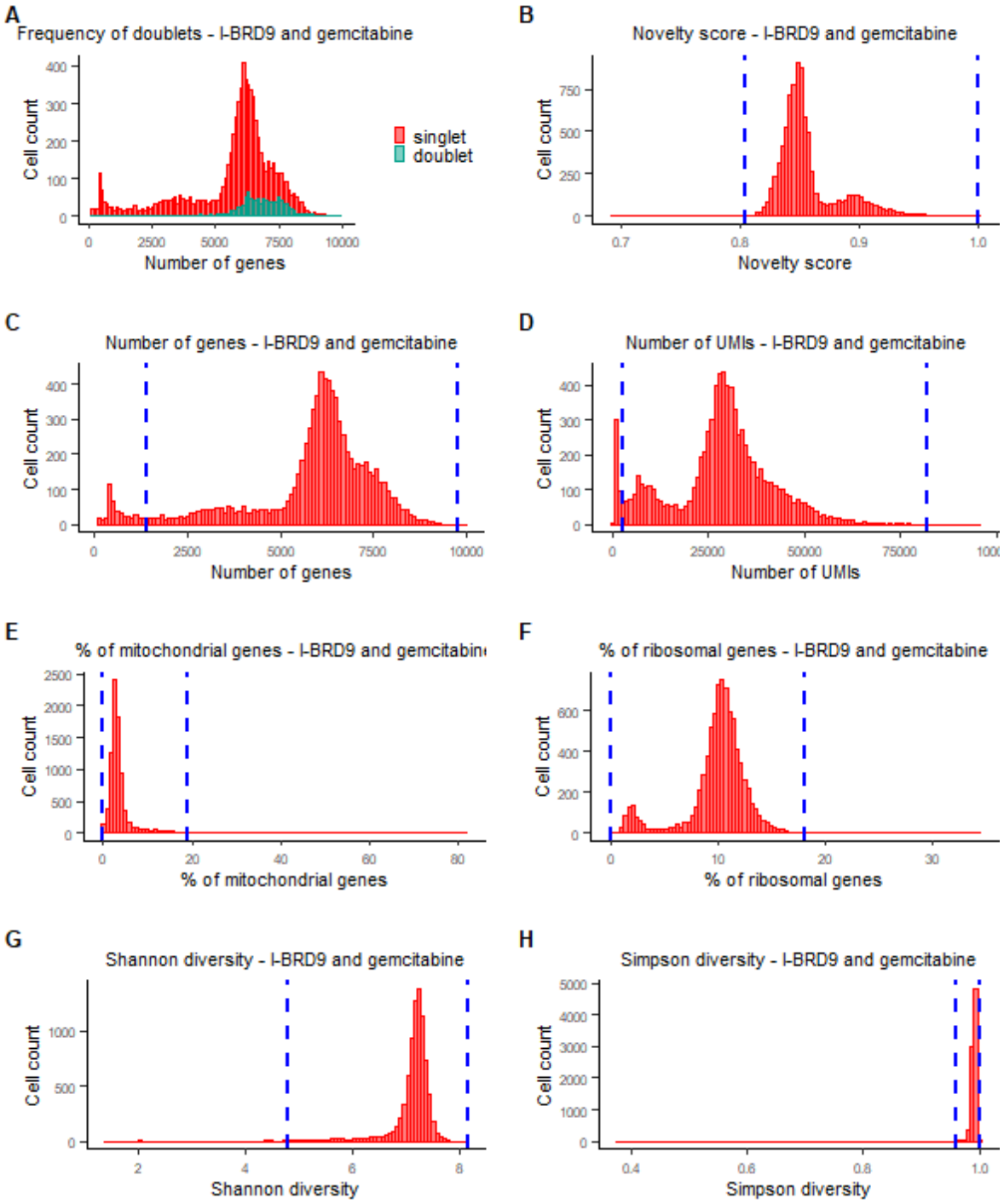


Figure 4.6. Quality control selection criteria for cells in the I-BRD9 and Gemcitabine condition: **A)** Singlet status as predicted by scDblFinder; **B)** Novelty score (log10 (nFeature_RNA) / log10 (nCount_RNA)) above 0.8; **C)** Number of detected genes (nFeature_RNA) between 1380 and 9800; **D)** Number of UMIs (nCount_RNA) between 3000 and 82000; **E)** Percentage of mitochondrial genes below 19%; **F)** Percentage of ribosomal genes below 18%; **G)** Shannon diversity above 4.8; **H)** Simpson diversity above 0.96.

The cut-offs employed for each of the criteria are summarized in **Table 4.1**:

Cell retention criterion/Experimental condition	DMSO	I-BRD9	Gemcitabine	I-BRD9 and gemcitabine
Novelty score	Above 0.78	Above 0.8	Above 0.78	Above 0.8
Number of genes	Between 1300 and 10700	Between 1000 and 9900	Between 3200 and 12600	Between 1380 and 9800
Number of UMIs	Between 4000 and 126000	Between 3400 and 88000	Between 11000 and 155000	Between 3000 and 82000
Percentage of mitochondrial genes	Below 24.6	Below 24	Below 23.2	Below 19
Percentage of ribosomal genes	Below 21.5	Below 17.8	Below 19.8	Below 18
Shannon diversity	Above 5.15	Above 5	Above 6.2	Above 4.8
Simpson diversity	Above 0.98	Above 0.96	Above 0.98	Above 0.96

Table 4.1. A summary of the quality control selection criteria used for cells in all four experimental conditions (DMSO, I-BRD9, Gemcitabine, I-BRD9 and gemcitabine).

The numbers of cells that were removed at each step of the filtering and retained in the Seurat object are illustrated in **Figure 4.7**. After the filtering of low-quality cells, the merged Seurat object contained 4314 DMSO cells, 2813 Gemcitabine cells, 5607 I-BRD9 cells, and 6857 I-BRD9 and gemcitabine cells. 382 genes that became expressed in fewer than 10 cells were removed.

The next section will discuss the removal of non-tumour cells.

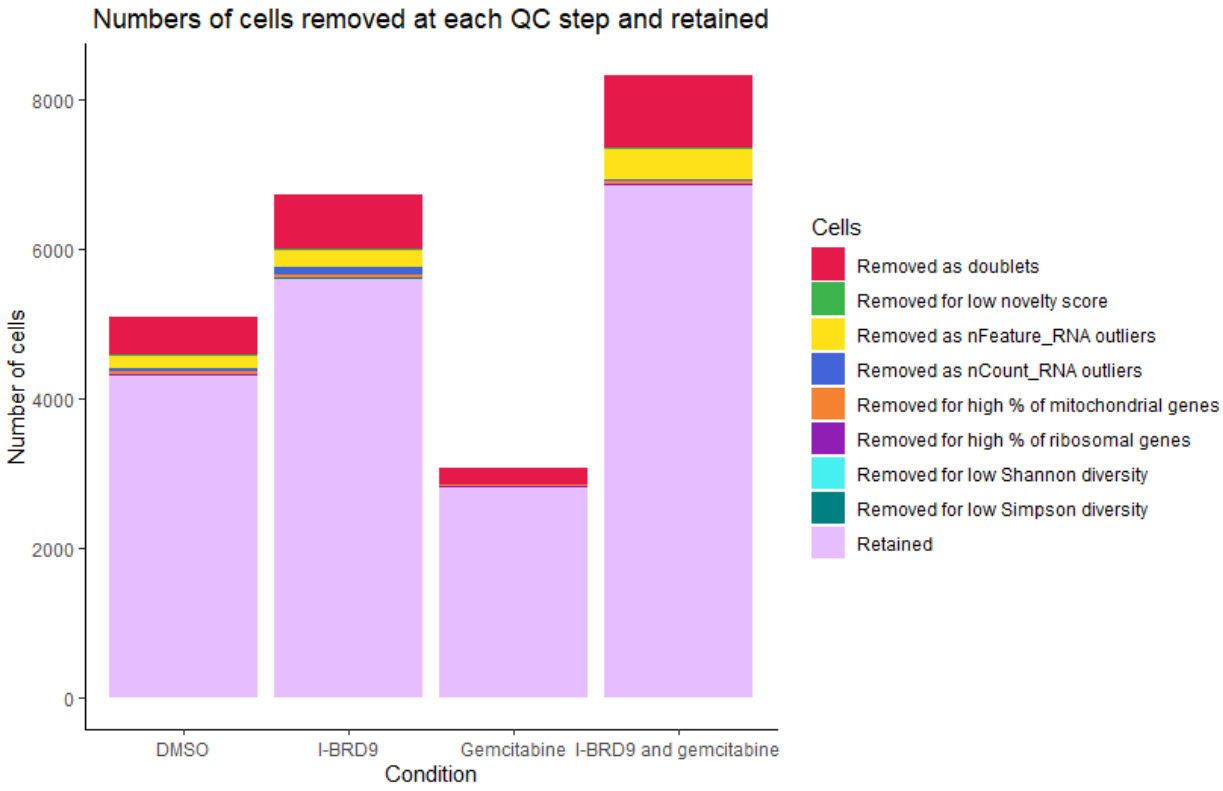


Figure 4.7. The numbers of cells removed at each step of the filtering and retained at the end.

4.2.2 Removal of non-tumour cells

In order to study the effects of the treatments upon PDAC cells specifically, the identification of non-tumour cells was performed, and these cells were subsequently removed from the data.

The Seurat object was split by experimental condition. Then, SCTransform was applied individually on the Seurat objects obtained after splitting, regressing the percentages of mitochondrial genes and ribosomal genes in the process, and the objects corresponding to different experimental conditions were merged again. With the gene counts now adjusted by SCTransform, 53 genes appeared in fewer than 10 cells and were therefore removed.

Next, PCA was run, the object was integrated with Harmony, thus allowing the cells to be grouped by their functional type rather than by the experimental condition, and an UMAP reduction was performed using the first 15 Harmony-generated dimensions as input (**Figure 4.8**).

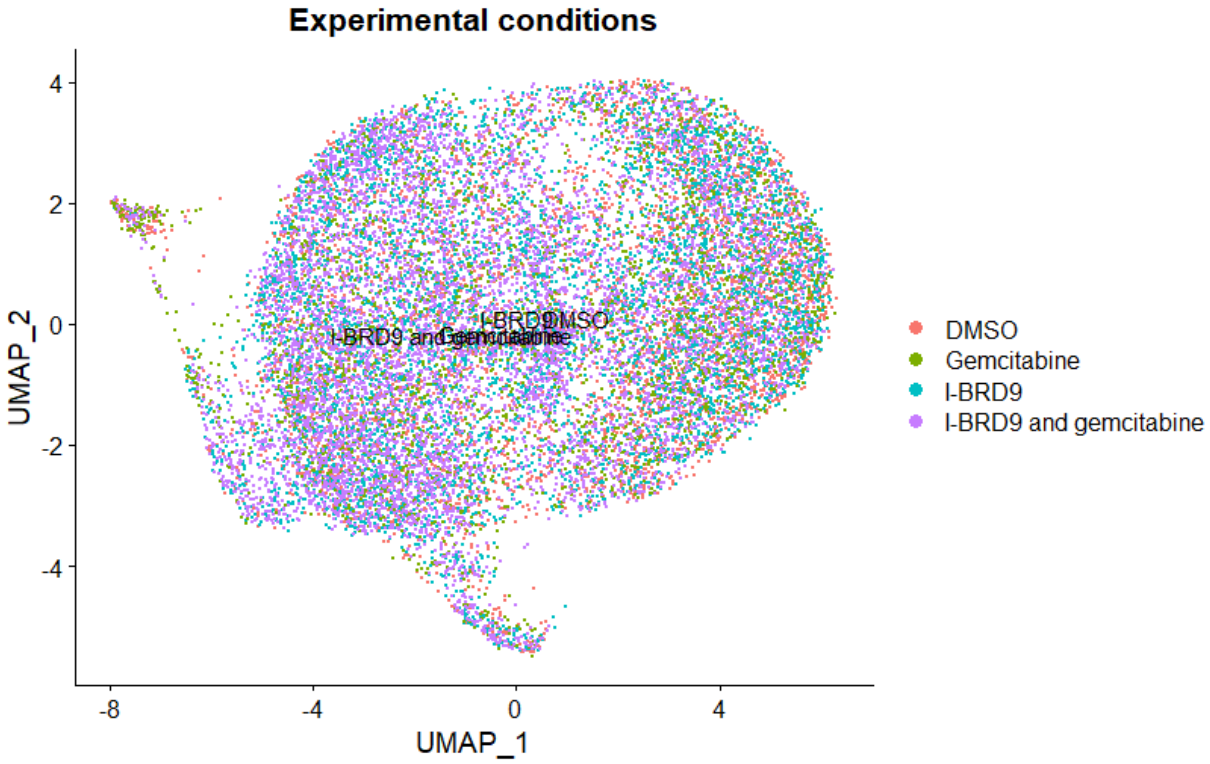


Figure 4.8. Experimental condition in the integrated Seurat object, prior to the identification of cell types.

Tumour cells were then separated from non-tumour cells. The cells were clustered at a low resolution of 0.03 which placed the weakly connected regions of the plot into different clusters (**Figure 4.9**).

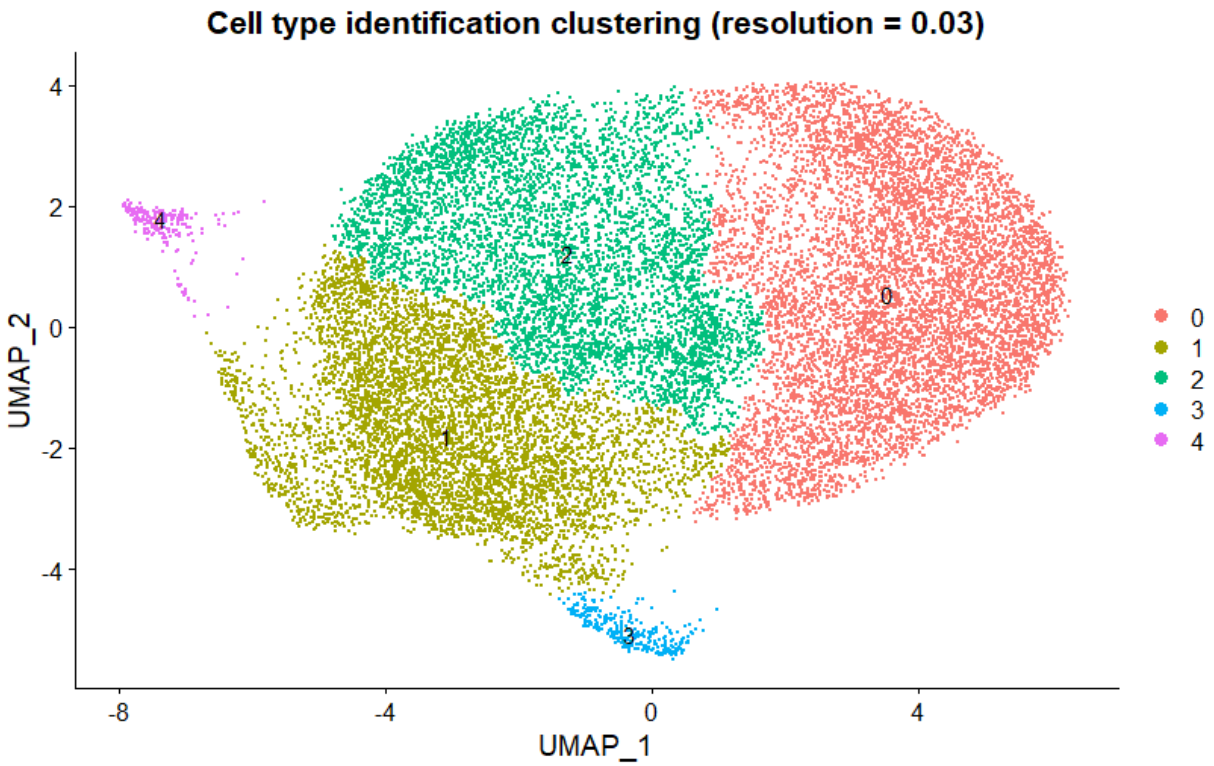


Figure 4.9. The clusters used to determine cell types.

KRT19, a reliable marker of ductal cells, for which a stromal presence detected in the context of PDAC has been attributed entirely to contamination with epithelial cells[237], although subtypes of fibroblasts able to express *KRT19* have also been noted to exist in association with hepatic cholangiocarcinoma[238], was strongly expressed over most of the UMAP plot, but in reduced amounts in **Cluster 3**. A similar situation was registered for other keratins such as *KRT8* and *KRT18*, and also for *DSG2*, listed among four putative serum markers that could be used to distinguish PDAC from benign cysts[239], thus suggesting that the bulk of the cells in this dataset consists of tumour ductal cells. In **Figure 4.10**, the expression of *KRT19*, *KRT8* and *DSG2* is illustrated.

However, this pattern of expression was found to occur together with the expression of multiple genes previously reported to characterize stromal cells, especially of the activated stellate variety: *COL3A1*, *COL5A2*, *FN1*, *VCAN*, *COL1A2*, *COL5A1*, *ITGA11*, *CDH11* and *COL1A1*. This

seeming inconsistency admits an explanation by positing a predominantly ductal character for the cells, as inferred from the presence of *KRT19*, seldom expressed in bona fide stromal cells, but coupled with a very marked shift to the mesenchymal end of the EMT in the vast majority of the cells, an explanation that has been previously invoked to explain the occurrence of "fibroblastoid" phenotypes in cancer cells, characterized by the ample expression of genes of high expression in the mesenchymal stroma[240]. This interpretation is supported by the abundance of *CDH2* and *VIM* in the dataset, progressively upregulated during EMT[241], and the almost complete absence of *CDH1*, downregulated in EMT[241] (**Figure 4.11**).

Meanwhile, **Cluster 3** – lower in *KRT19* expression – shows several markers of fibroblasts of only minimal expression anywhere else in the UMAP plot (**Figure 4.12**). An example is *ISLR*, reported to be expressed in fibroblasts and mesenchymal stromal cells, but not in epithelial, endothelial or smooth muscle cells[242] . Thus, it was assessed as a mixed fibroblast cluster that may also contain elements transcriptionally akin to *KRT19*-expressing fibroblast population mentioned above.

Consequently, most of the cells in the dataset were identified as ductal tumour cells shifted to the mesenchymal end of EMT, while one small cluster was identified as consisting of fibroblasts (**Figure 4.13**) and therefore removed from the data. The preponderance of tumour cells relative to the non-tumour cells is explained by the low-adherent cell culture conditions, favoring the survival of tumour cells due to their resistance to anoikis.

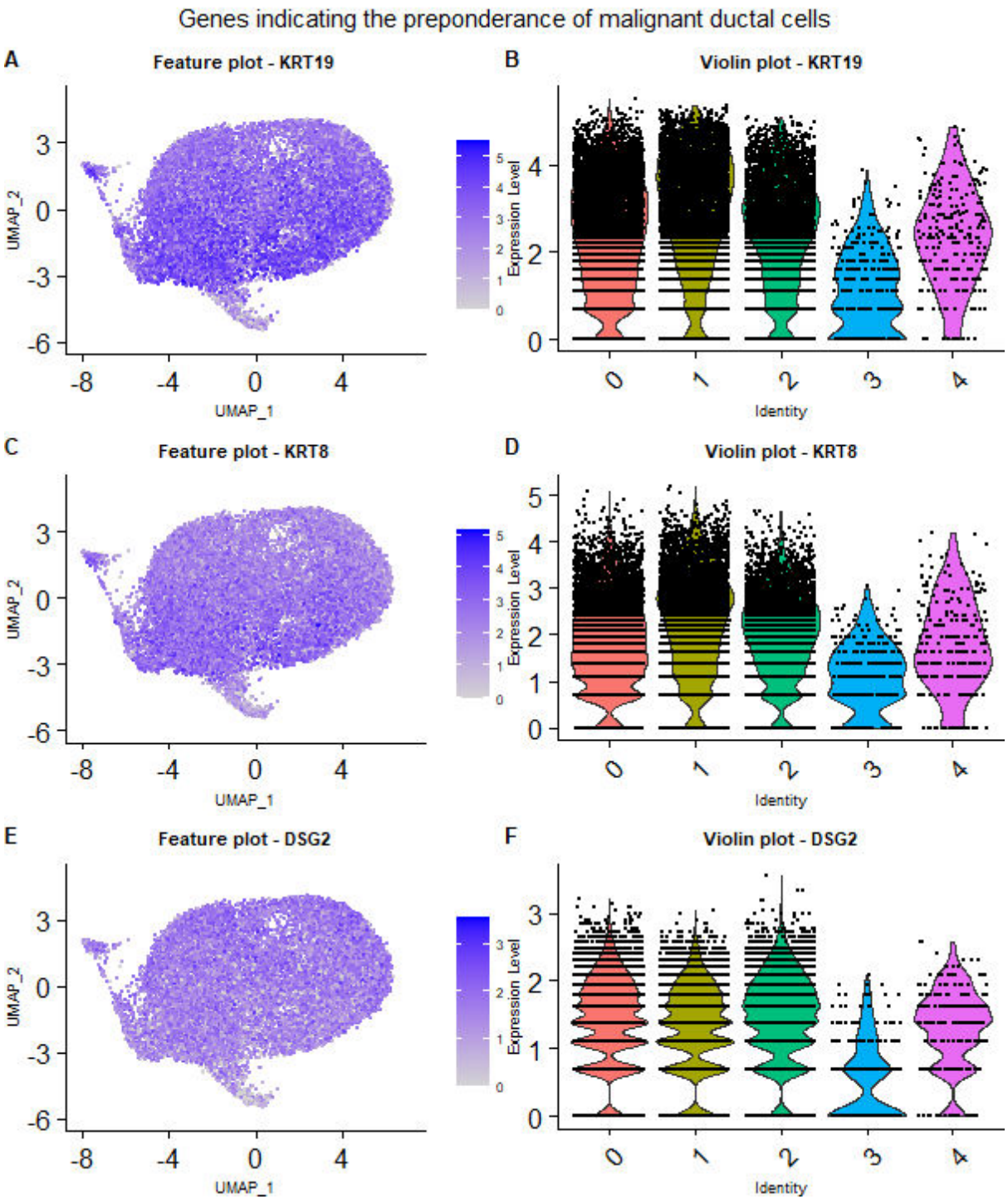


Figure 4.10. *KRT19* and *KRT8*, markers of ductal cells, and *DSG2*, a putative marker of PDAC, show ample expression in the dataset, but more restricted in Cluster 3.

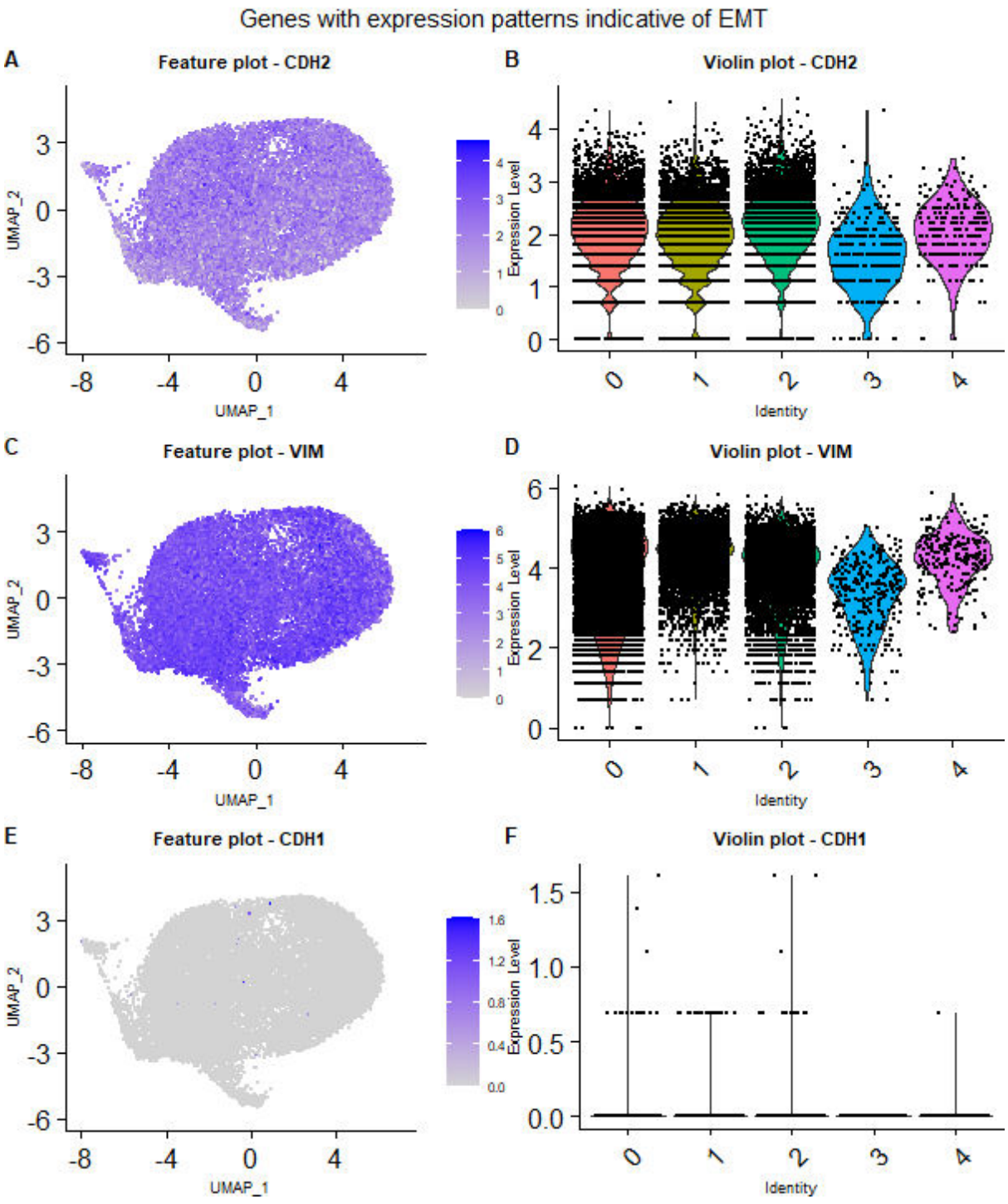


Figure 4.11. The very high expression of *CDH2* and *VIM* contrasts with the nearly absent expression of *CDH1*, marking the epithelial-mesenchymal transition.

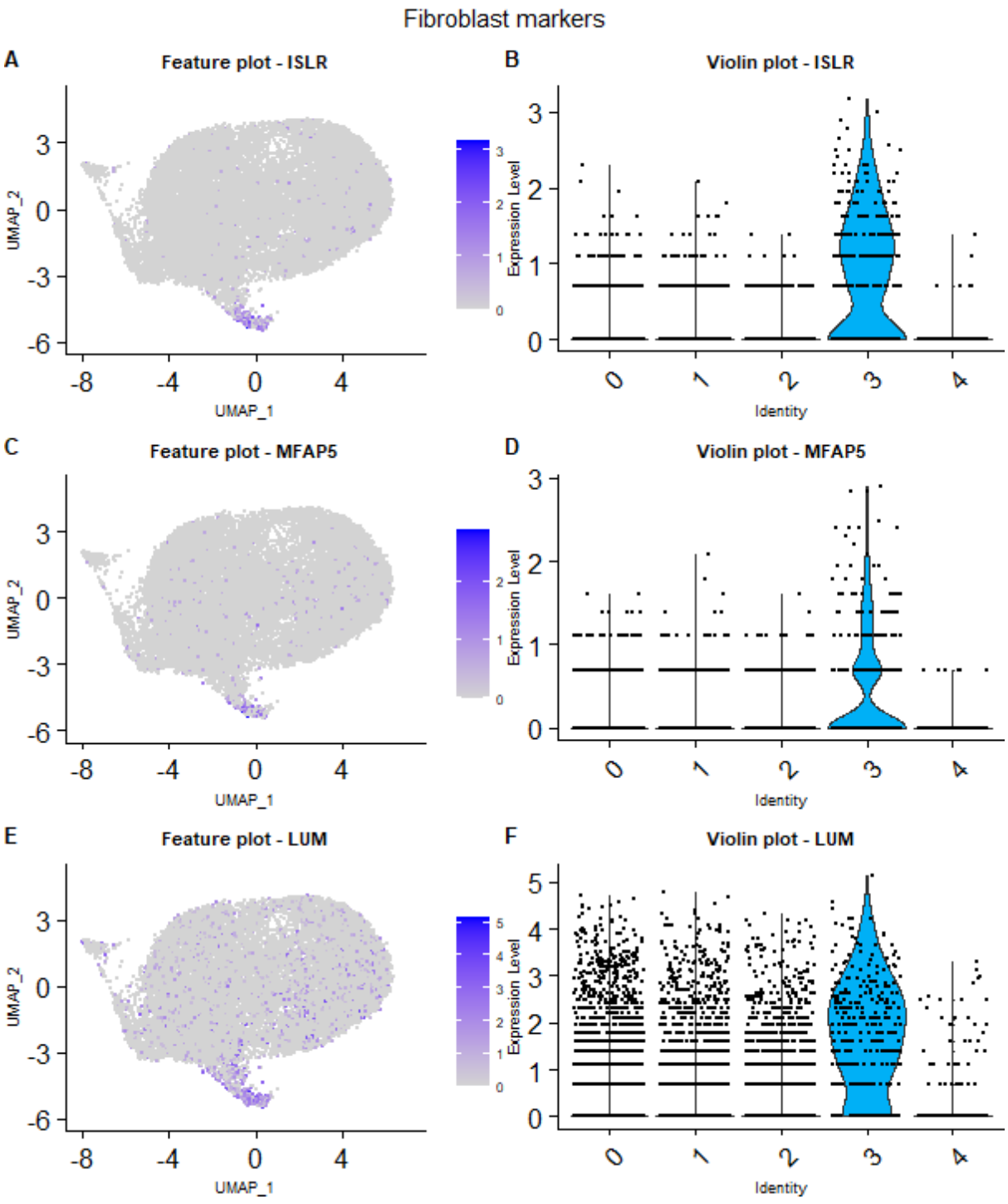


Figure 4.12. Fibroblast markers such as *ISLR*, *MFAP5* and *LUM* are distinctively overexpressed in Cluster 3.

After the removal of fibroblasts, 28 genes expressed in < 10 cells were removed. 21931 genes and 17509 cells were retained, and the workflow between quality control and dimensionality reduction was repeated, followed by a UMAP on the first 22 Harmony dimensions (**Figure 4.14**).

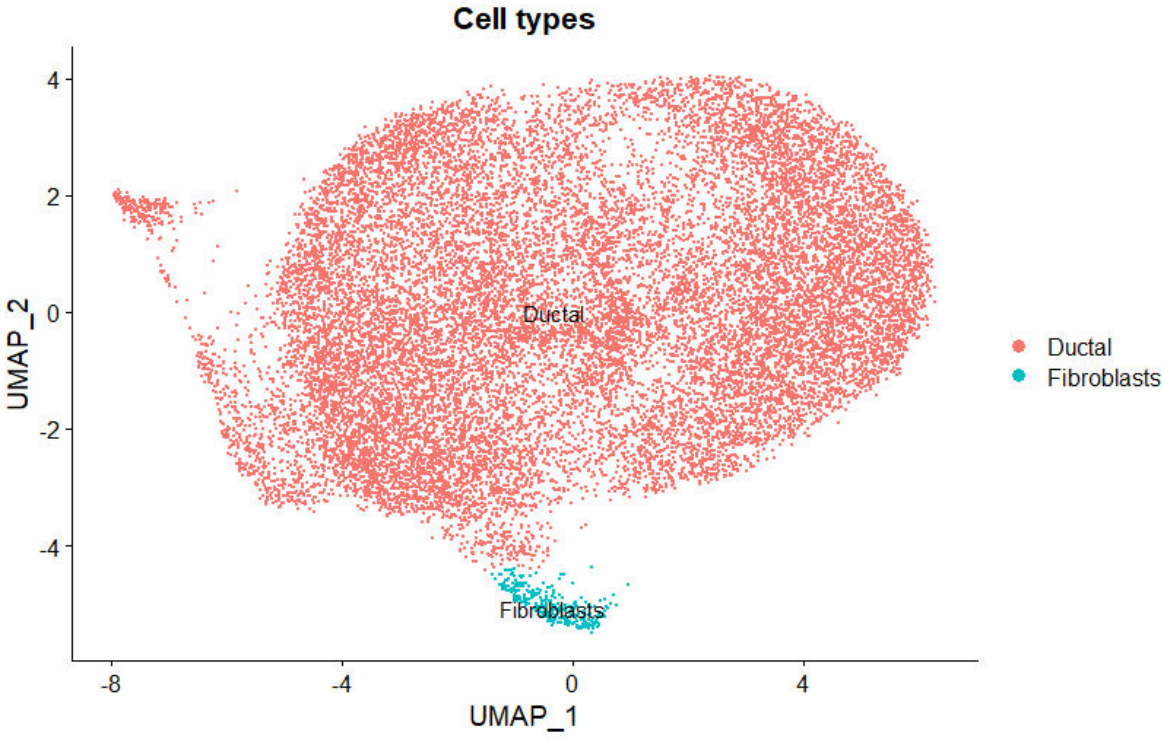


Figure 4.13. Cell type identification.

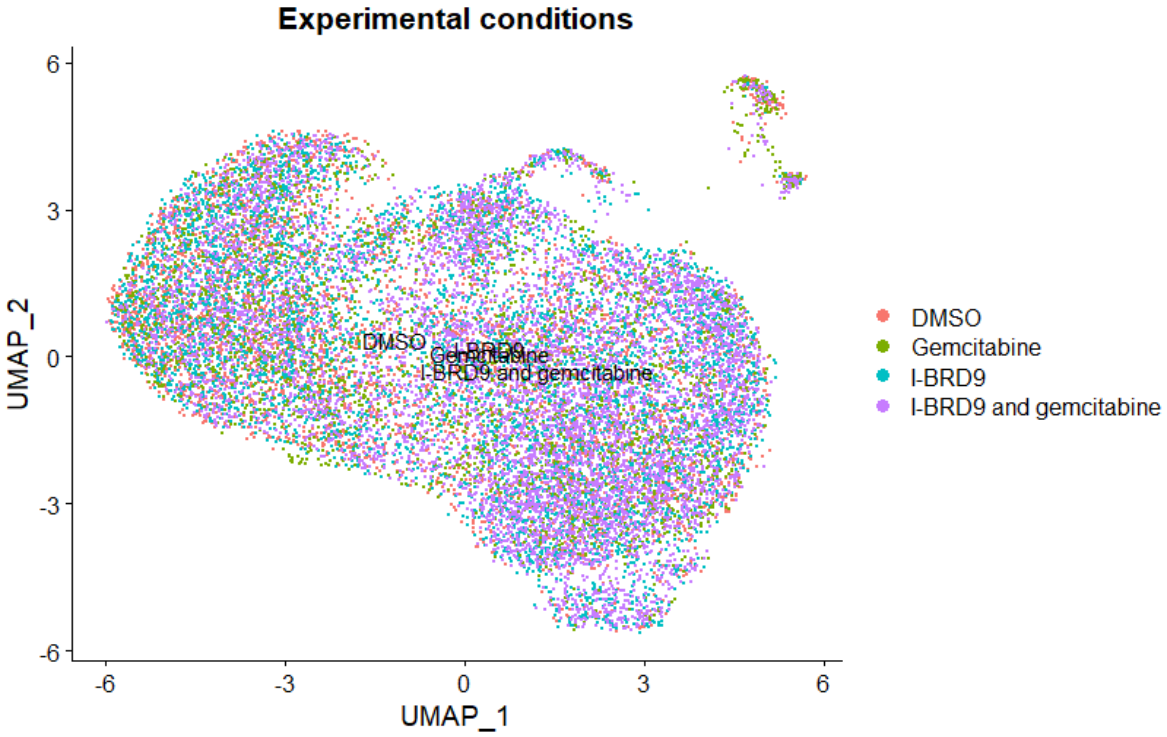


Figure 4.14. UMAP plot showing cells grouped by the experimental condition.

The next section will describe the clustering.

4.2.3 Clustering

Next, clustering was performed, in order to identify subpopulations of cells similar in terms of their gene expression.

As described in **Section 2.3.4**, clustering was initially performed at a high resolution, obtaining a much larger of number of clusters than the final one. Then, the clusters were merged based on marker expression, in order to obtain the most biologically relevant clustering. Iteratively, a resolution value of 2.5 was found to be suitable for the task.

Thus, 67 clusters were found at a resolution of 2.5 (**Figure 4.15**). Subsequently, they were merged into 9 clusters (**Figure 4.16**).

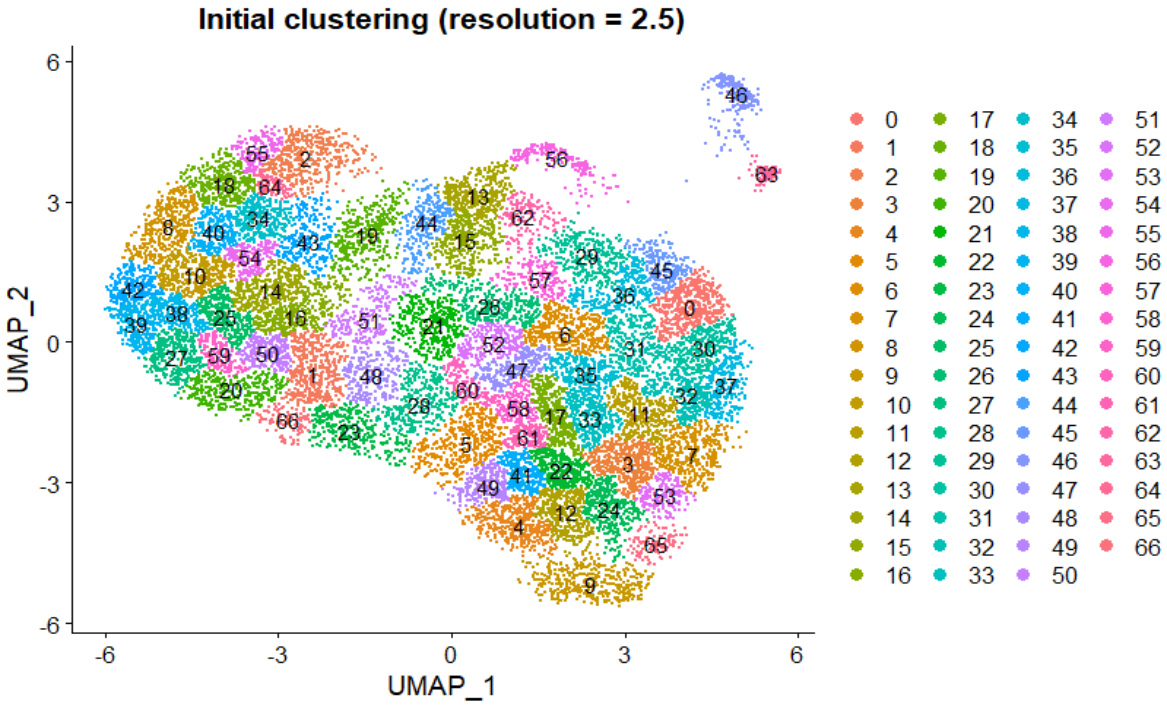


Figure 4.15. The initial configuration of clusters.

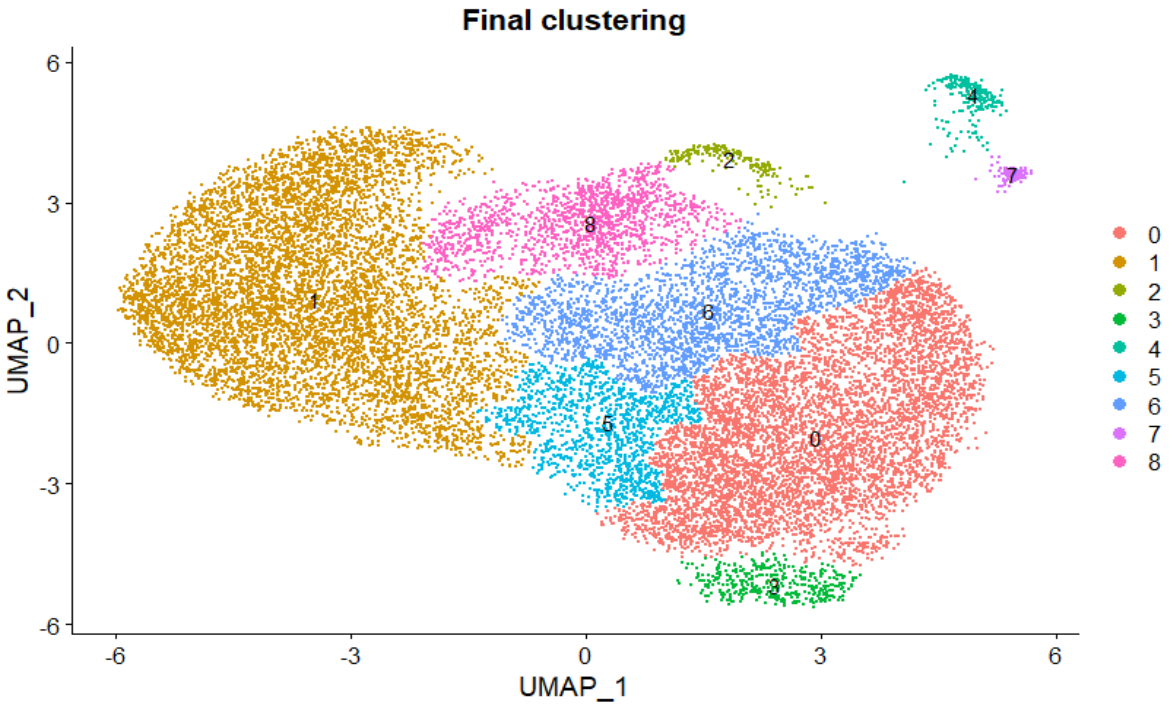


Figure 4.16. The final configuration of clusters.

The cluster mergers and the corresponding markers are listed in **Table 4.2**:

Final cluster	**Original clusters**	**Markers**
0	0, 30, 32, 11, 33, 31, 17, 22, 61, 41, 12, 4, 35, 37, 7, 53, 24, 3 and 65	*RGS4*; *SPRY2*
1	1, 66, 20, 50, 59, 27, 25, 39, 38, 16, 14, 54, 10, 42, 43, 34, 64, 2, 55, 18, 40, 8, 23, 51 and 48	*ANKRD37*; *ADM*
2	56	*ATF3*; *PMAIP1*
3	9	*TUBB4A*; *GPX3*
4	46	*CLSPN*; *CENPF*
5	5, 58, 60, 28 and 49	*OSR1*; *GALK1*

6	6, 57, 29, 36, 26, 52, 47, 21 and 45	*PDGFD*; *SLC40A1*
7	63	*BTG2*; *GDF15*
8	13, 62, 15, 44 and 19	*IFI44L*; *RSAD2*

Table 4.2. 9 final clusters were constructed from the initial 67 clusters via merging, based on the expression of shared markers that displayed a localized expression in distinct regions of the UMAP plot.

On the next few pages, the expression of 9 genes important to the delineation of clusters, namely *RGS*, *ANKRD37* and *PMAIP1* (**Figure 4.17**), *TUBB4A*, *CLSPN* and *OSR1* (**Figure 4.18**), *PDGFD*, *GDF15* and *IFI44L* (**Figure 4.19**) is illustrated through feature and violin plots.

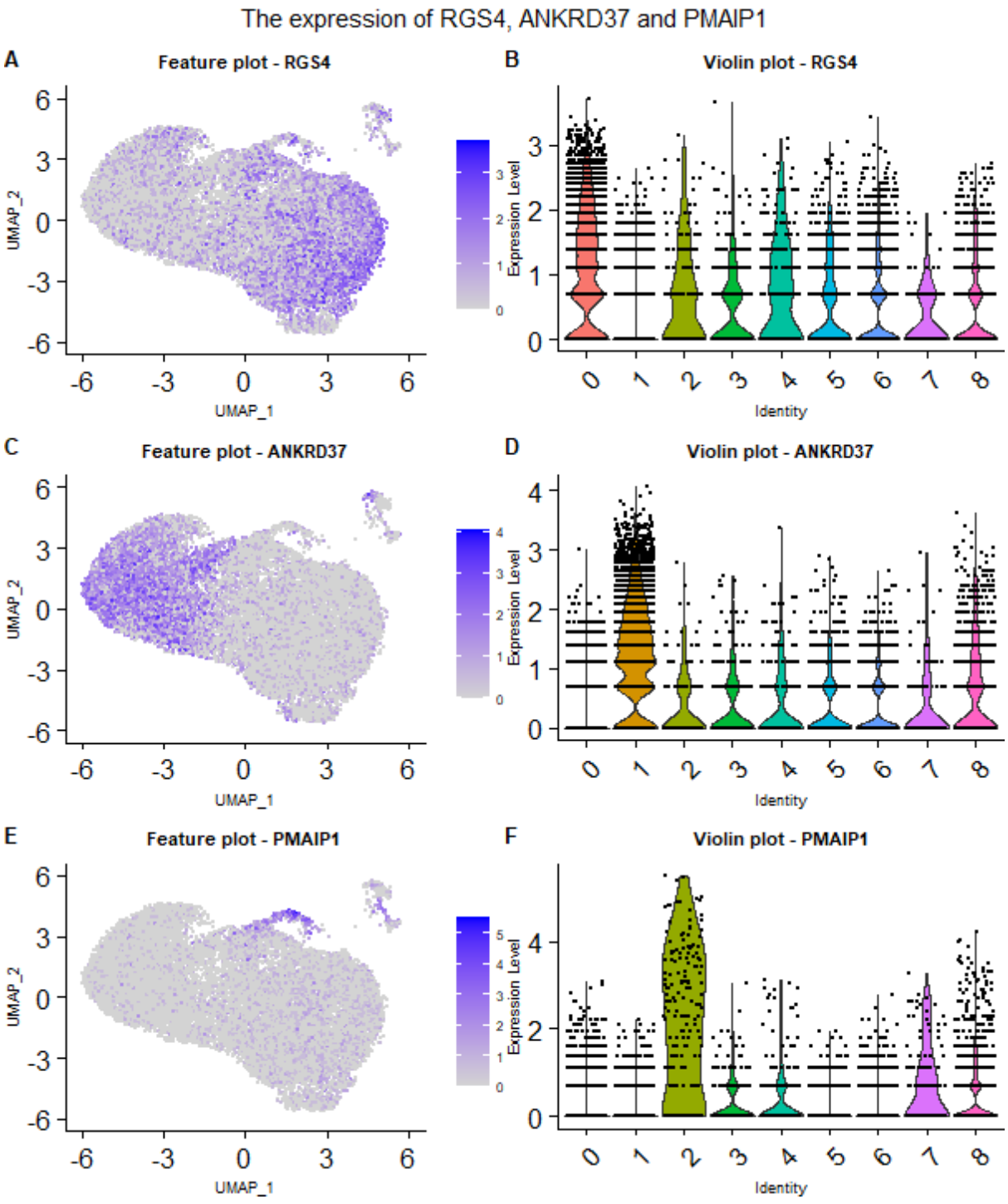


Figure 4.17. The contour of *RGS4* delineates Cluster 0 (**A, B**). The expression pattern of *ANKRD37* establishes Cluster 1 (**C, D**). *PMAIP* is overexpressed in the region of the plot defined as Cluster 2 (**E, F**).

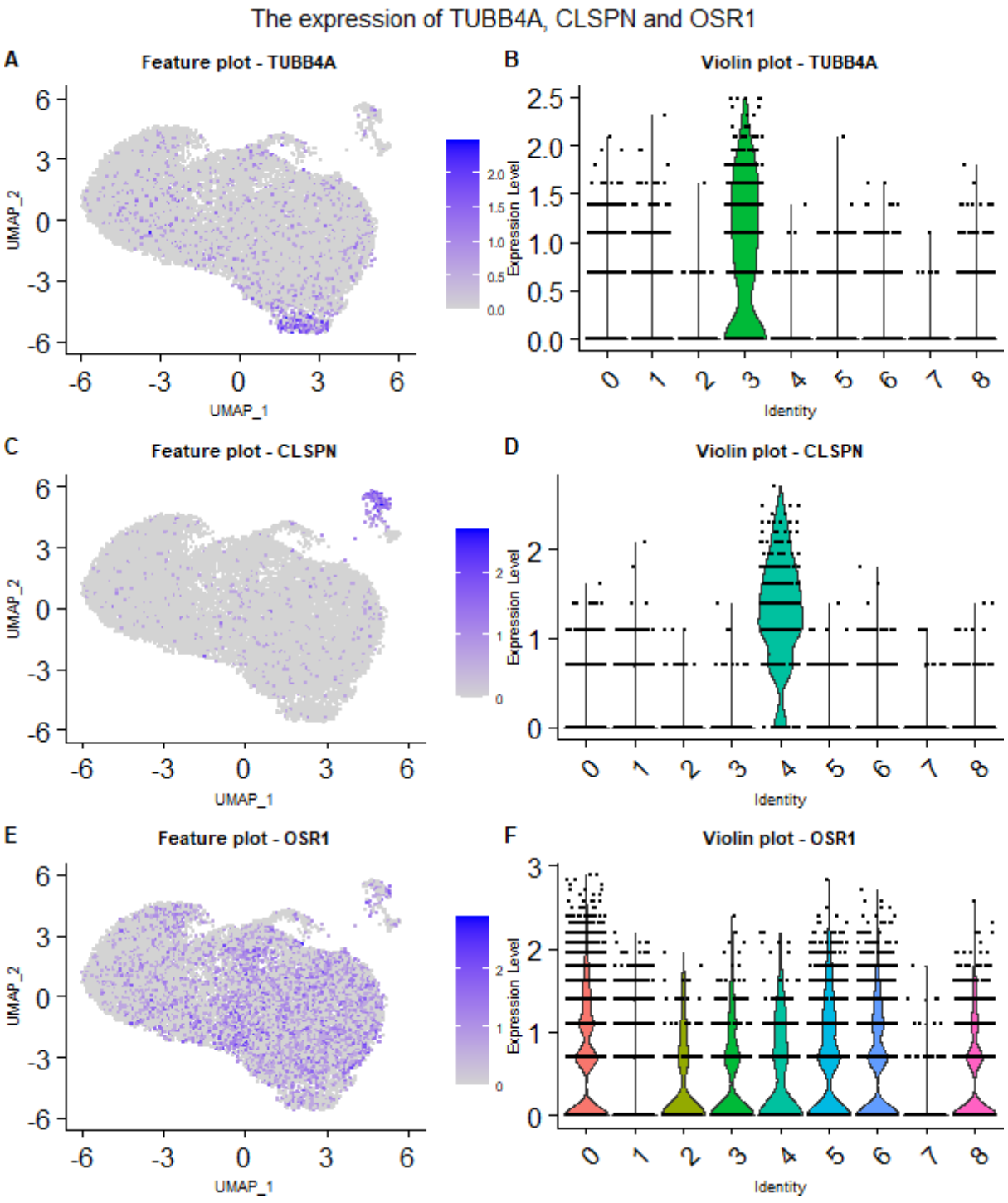


Figure 4.18. The high expression of *TUBB4A* marks Cluster 3 (**A, B**). *CLSPN* is very strongly expressed in Cluster 4 (**C, D**). Cluster 5 is the least distinctive among all clusters in terms of specific markers. *OSR1* show a higher expression in this cluster (**E, F**).

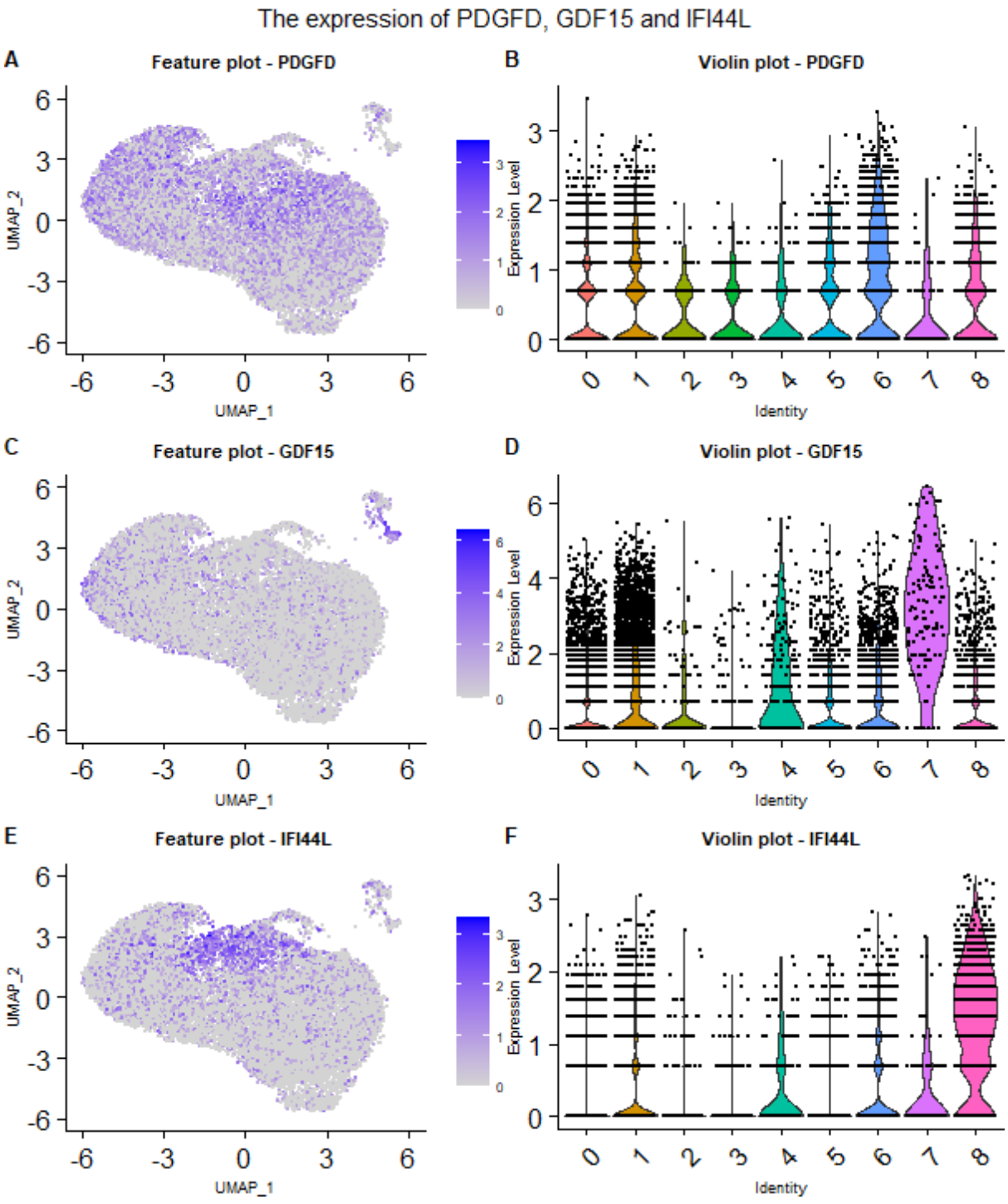


Figure 4.19. Less distinctive than most clusters, Cluster 6 is delineated by the overexpression of *PDGFD* (**A, B**). *GDF15* is markedly overexpressed in Cluster 7 (**C, D**). The contour determined by the high expression of *IFI44L* demarcates Cluster 8 (**E, F**).

The next section will identify the clusters associated with cancer stemness.

4.2.4 Identification of clusters associated with cancer stemness

The identification of the clusters associated with cancer stemness followed six lines of evidence (see **Section 2.3.5**) the 27 TDPM genes, the 14 CCRSA gene sets, the 39 SPLCL genes, ORIGINS activity predictions, Slingshot pseudotime inference, and Enrichr gene enrichment analysis using the *CellMarker_Augmented_2021* and *PanglaoDB_Augmented_2021* databases (see **Section 2.3.5**).

17 markers among the TDPM genes were found to expressed in the dataset: *MET*, *CD24*, *CD44*, *DCLK1*, *EPCAM*, *LGR5*, *ABCG2*, *LAP3*, *NES*, *ABCB1*, *POU5F1*, *CD34*, *TSPAN8*, *CD9*, *EZH2*, *SSEA-1* (*FUT4*) and *GLRX3*. The Spearman correlation of their gene expressions is displayed in **Figure 4.20**:

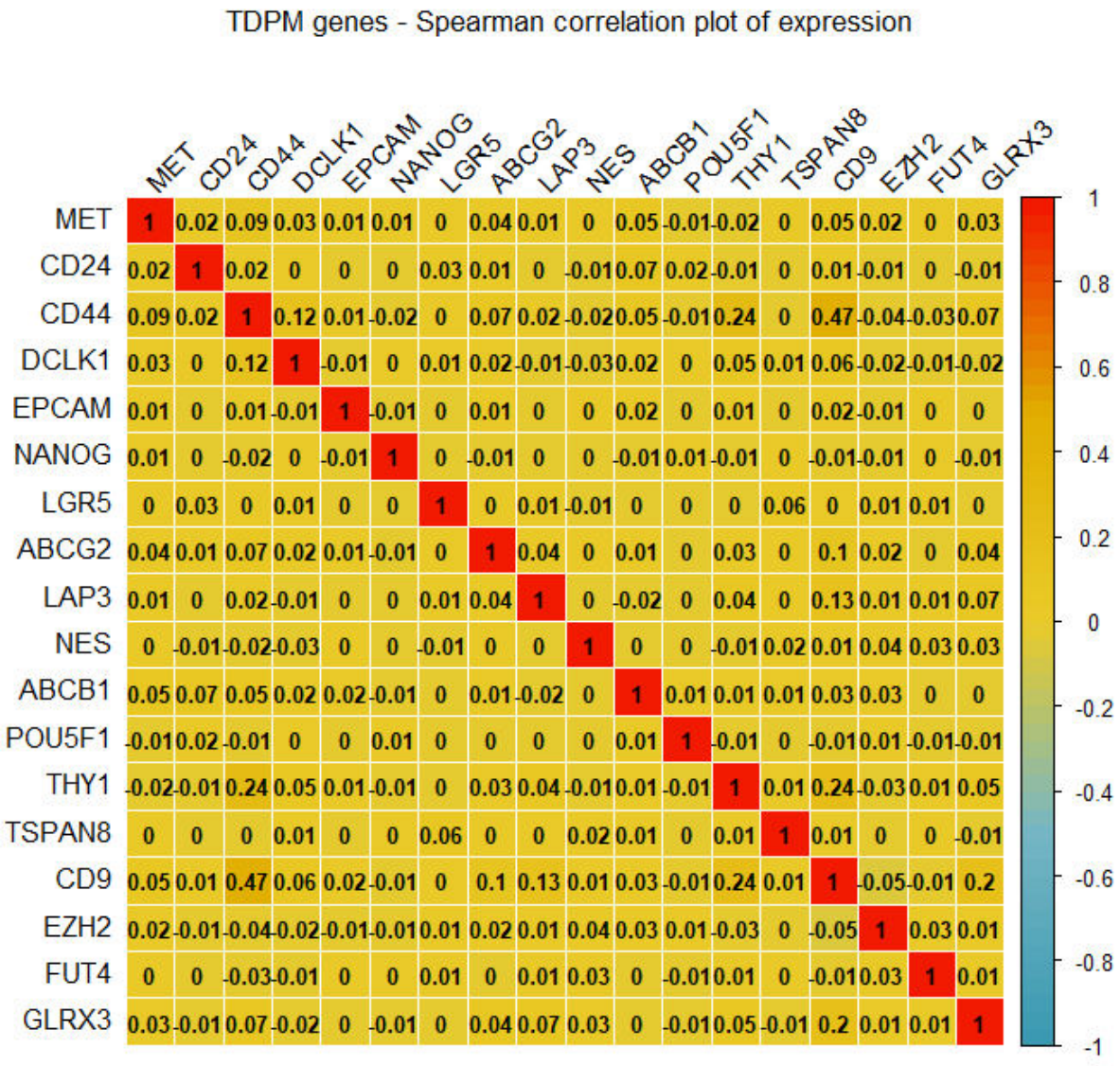


Figure 4.20. Spearman correlation plot of the expression of the 17 TDPM genes that were found in the dataset.

A singular instance of a high correlation can be observed, namely between *CD9* and *CD44* (0.47), while the correlation between *THY1* and *CD44* was also notable (0.24). Otherwise, the TDPM genes did not correlate well with each other.

11 out of the 18 markers were found to be differentially overexpressed in at least one cluster. However, only *EZH2* linked distinctly to a particular region of the UMAP plot, namely **Cluster 4** (adjusted p-values: 2.58e-176). For all the other TDPM genes, the expression was not localized

and thus not suggestive of marking CSC. In **Cluster 4**, *NES*, *THY1*, *ABCG2* and *ABCB1* also reached statistical significance (adjusted p-values: 2.21e-21, 1.42e-16, 1.8e-03 and 4.27e-03).

Next, the markers of **Cluster 4** significantly overlapped with each of the 14 CCRSA gene sets, with p-values between 5.32e-117 (the union CCRSA gene set) and 5.38e-06 (the ovarian CCRSA gene set), while no other cluster registered significant overlaps with any of these sets. Among the 117 CCRSA genes detected in the dataset, 114 were found to be significantly overexpressed in **Cluster 4**, 86 of which exclusively, while **Cluster 7** had the second greatest number of overexpressed CCRSA genes, 16, but none were expressed exclusively in this cluster (**Figure 4.21**).

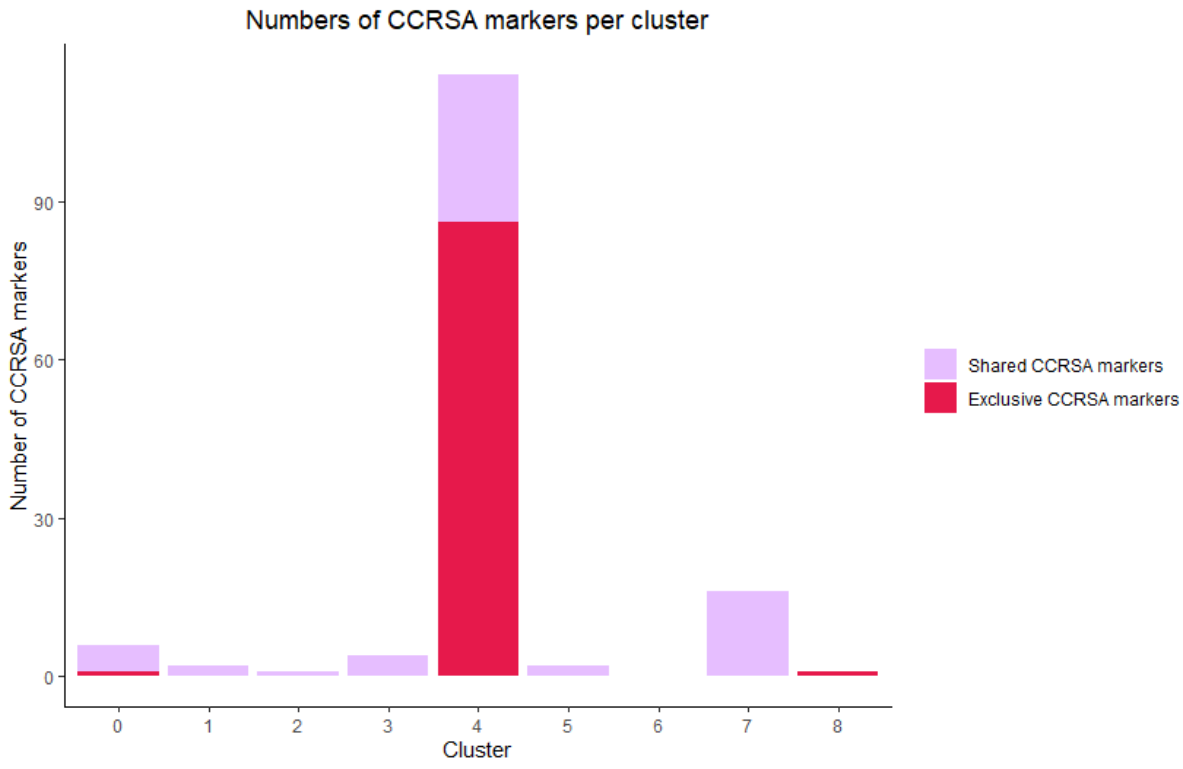


Figure 4.21. Nearly all the CCRSA genes found in the dataset are overexpressed in Cluster 4, while only a few of them are expressed in the other clusters.

No cluster registered a significant overlap with **SPLCL genes**, indicating the absence of significantly side population-like clusters in this dataset.

Both assessments involved **ORIGINS activity**, pairwise Wilcoxon comparisons (**Figure 4.22**), and the detection of overrepresentation of high activity cells (**Table 4.3**) showed stemness links for **Cluster 4**:

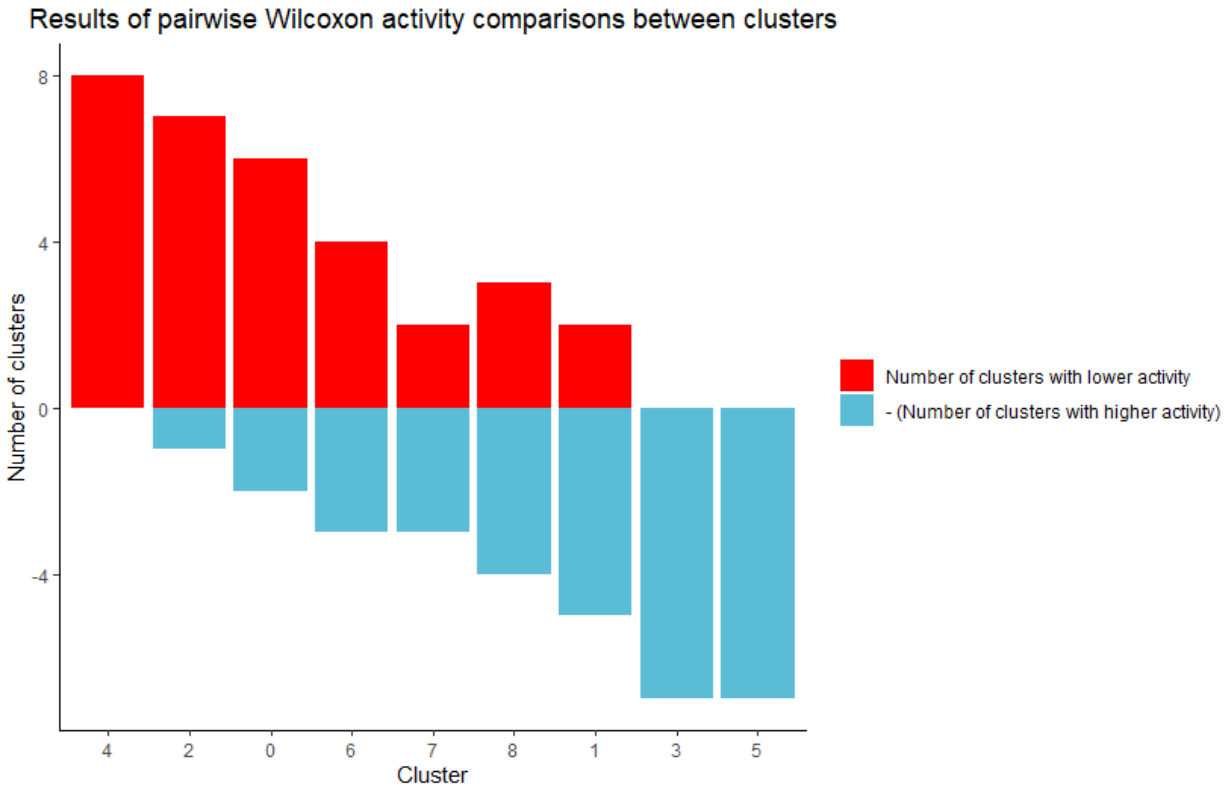


Figure 4.22. Cluster 4 showcased significantly higher activity scores than all the other clusters.

Cluster	p-value of the overrepresentation of the cells among the high ORIGINS activity cells
4	1.43e-40
0	2.38e-22
2	4.5e-09
6	1.41e-02

Table 4.3. The clusters evidencing an overrepresentation of cells with top activity scores.

Next, Slingshot identified three differentiation lineages (**Figure 4.23**), all of which started in **Cluster 4** and had **Cluster 7** as the next cluster in the trajectory.

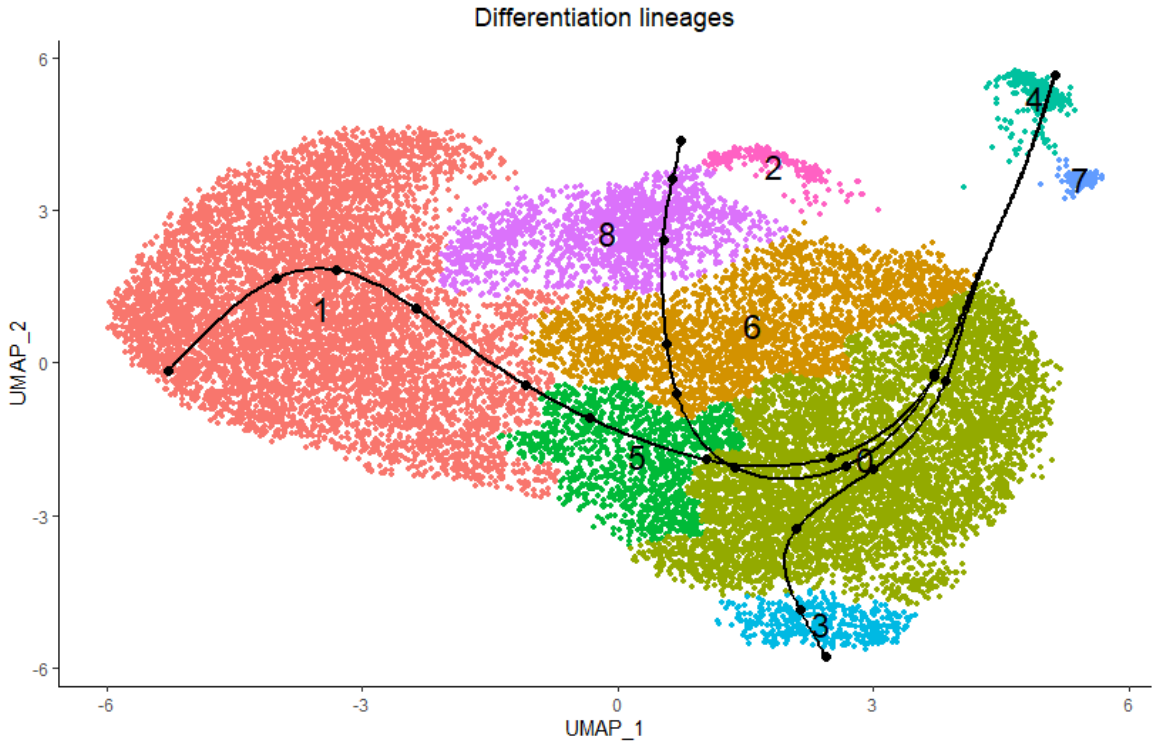


Figure 4.23. Differentiation lineages identified using Slingshot.

The first of the three lineages is illustrated in **Figure 4.24**. After **Cluster 4** and **Cluster 7**, the trajectory of differentiation enters, in order: **Cluster 0**, **Cluster 5**, **Cluster 6** and **Cluster 8**, **Cluster 1**:

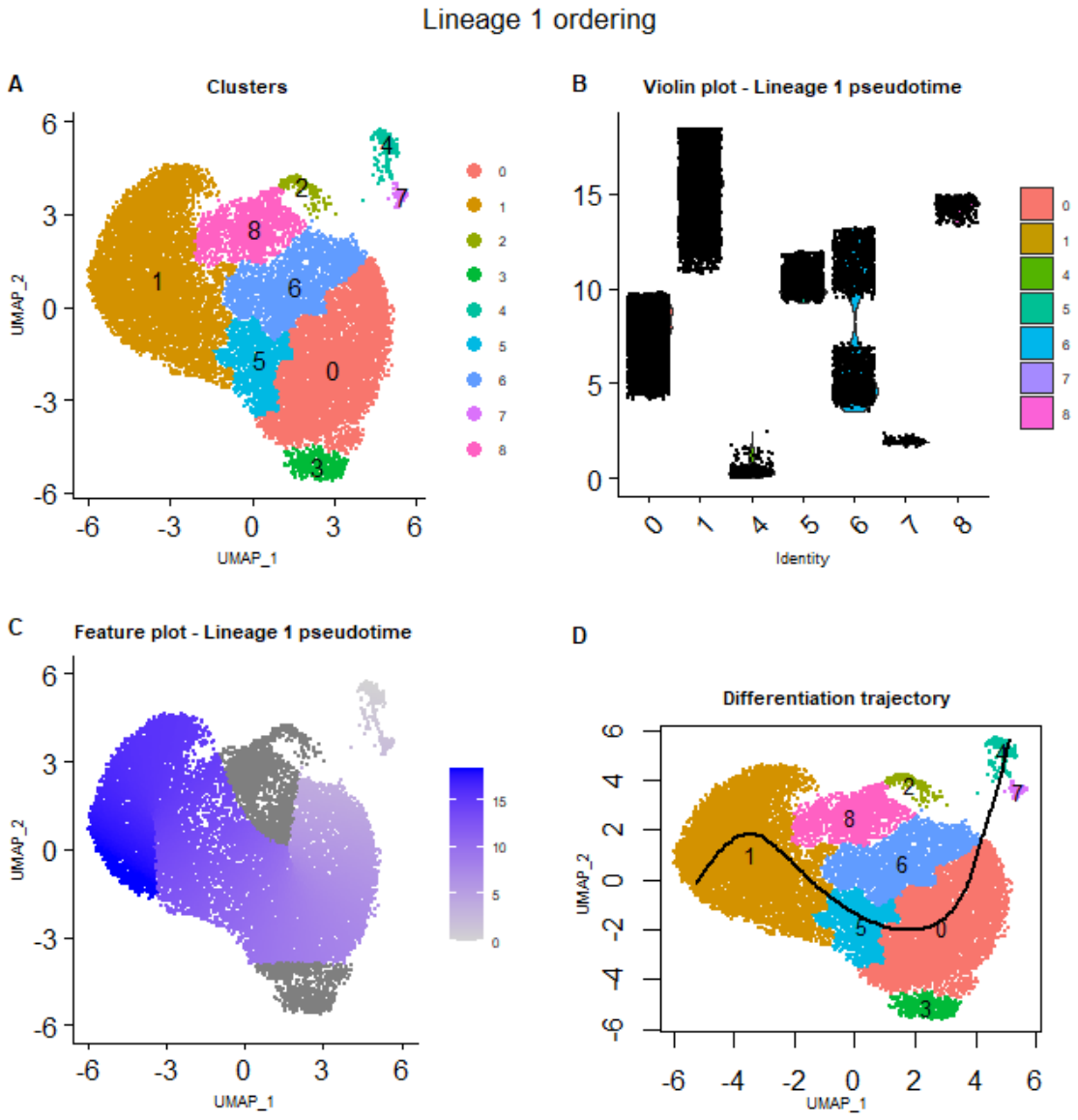


Figure 4.24. Lineage 1 ordering: **A)** Clusters; **B)** Pseudotime violin plot; **C)** Pseudotime feature plot; **D)** Differentiation trajectory.

The second of the three lineages passed through the same clusters as the first lineage, apart from the terminal cluster. The second lineage ends in **Cluster 2** rather than in **Cluster 1**, as displayed in **Figure 4.25**:

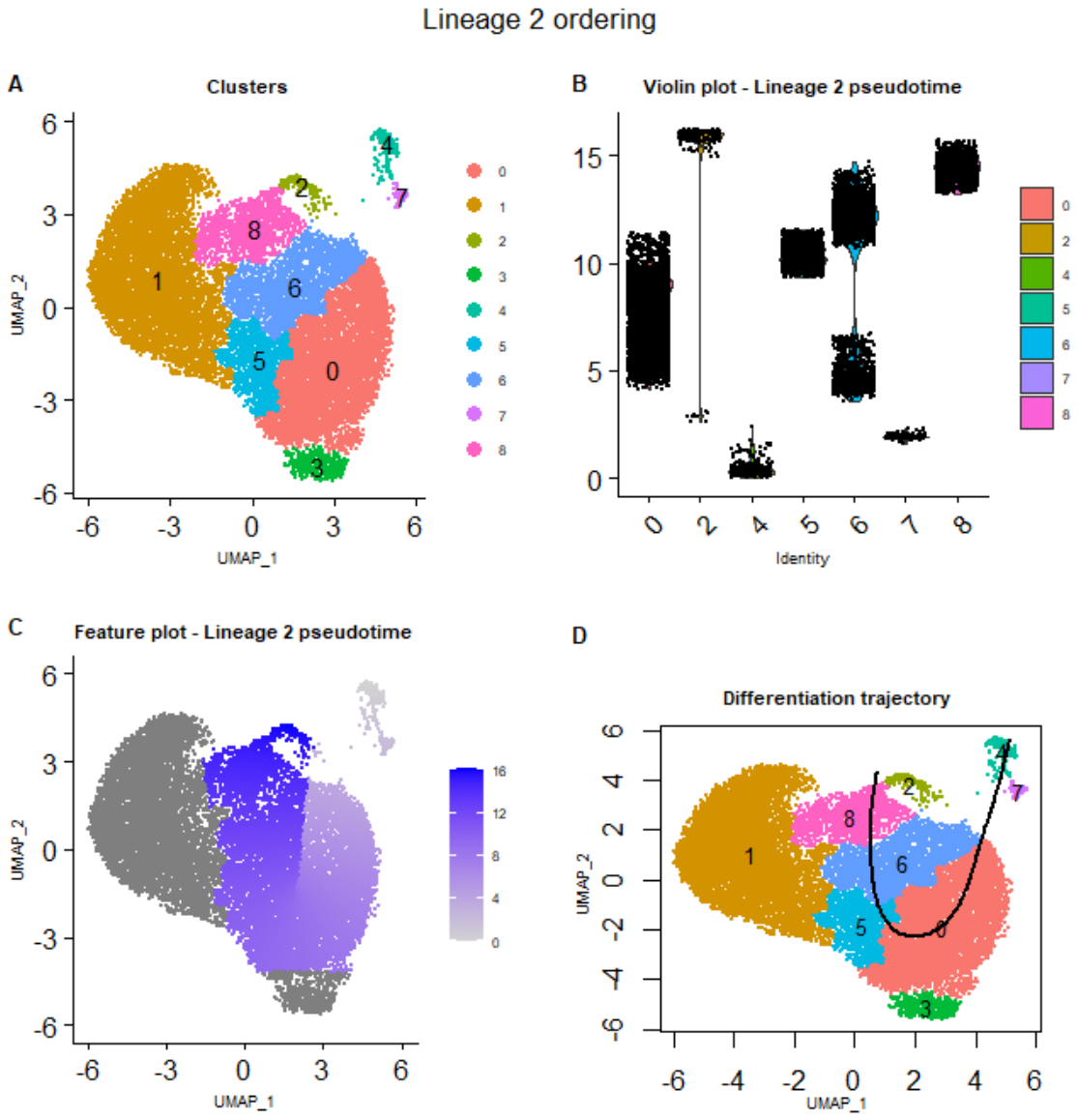


Figure 4.25. Lineage 2 ordering: **A)** Clusters; **B)** Pseudotime violin plot; **C)** Pseudotime feature plot; **D)** Differentiation trajectory.

The third lineage also passes first through **Cluster 4**, **Cluster 7** and **Cluster 0**, but unlike the previous two lineages, it never enters **Cluster 5**, instead reaching **Cluster 3**, where it ends (**Figure 4.26**):

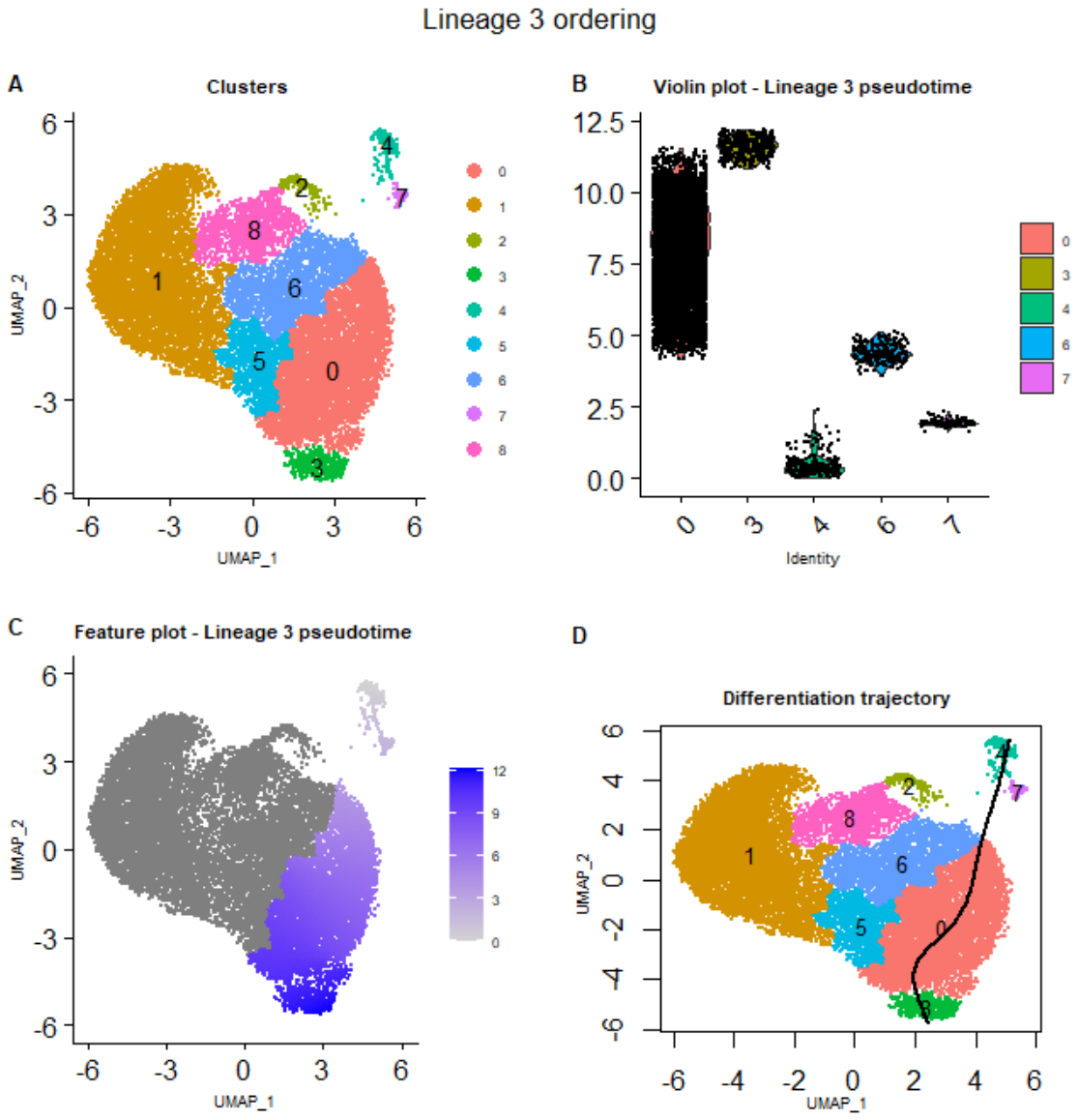


Figure 4.26. Lineage 3 ordering: **A)** Clusters; **B)** Pseudotime violin plot; **C)** Pseudotime feature plot; **D)** Differentiation trajectory.

Finally, for the *CellMarker_Augmented_2021* Enrichr database, the top six enriched terms for **Cluster 4** and **Cluster 7** were related to stem and progenitor cells. **Cluster 4** evidenced the lowest adjusted p-values for the top six terms, ranging between 3.67e-84 ("Oxyntic Stem cell:Stomach") and 3.15e-82 ("Vascular Progenitor cell:Blood"). The *PanglaoDB_Augmented_2021* analysis had "Pluripotent stem cells" as the top enriched term for **Cluster 4** (adjusted p-value: 1.72e-26), while the top enriched term was not related to stem or progenitor cells for any of the other clusters.

In conclusion, five out of the six lines of evidence suggest **Cluster 4** has the highest association with stemness in this dataset, while **SPLCL genes**-linked stemness was not found to characterize any of the clusters.

The next section will provide a functional characterization of the clusters.

4.2.5 Functional characterization of the clusters

Based on the results from the previous section, **Cluster 4** was named **Top stemness**. The weakly CCRSA genes-linked **Cluster 7** was found to have enriched terms related to p53 signal transduction. Therefore, **Cluster 7** was named **p53 signalling**. The top two enriched GO terms for **Cluster 4** and **Cluster 7** and their associated genes are displayed in **Figure 4.27**.

Otherwise, **Cluster 0** was found to be involved in actin filament organization and endosomal transport. It was named **Actin organization**. **Cluster 1** showed enriched GO terms related to response to hypoxia and oxygen levels, hence it was named **Hypoxia-like**. **Cluster 2** was found to regulate response to chemical stress, apoptosis and oxidative stress, and it was therefore named **Stress response**. **Cluster 3** showcased enriched terms related to ubiquitin-mediated protein catabolism, hence it was named **Protein degradation**. **Cluster 5** evidenced an enrichment of oxidative phosphorylation, cellular respiration and ATP metabolism. It was named **Cellular respiration**. **Cluster 6** was named **Wnt signalling**, in accord with its enrichment of GO terms linked to this pathway. **Cluster 8** presented an activation of interferon-gamma and type I interferon response, and was named **IF response**.

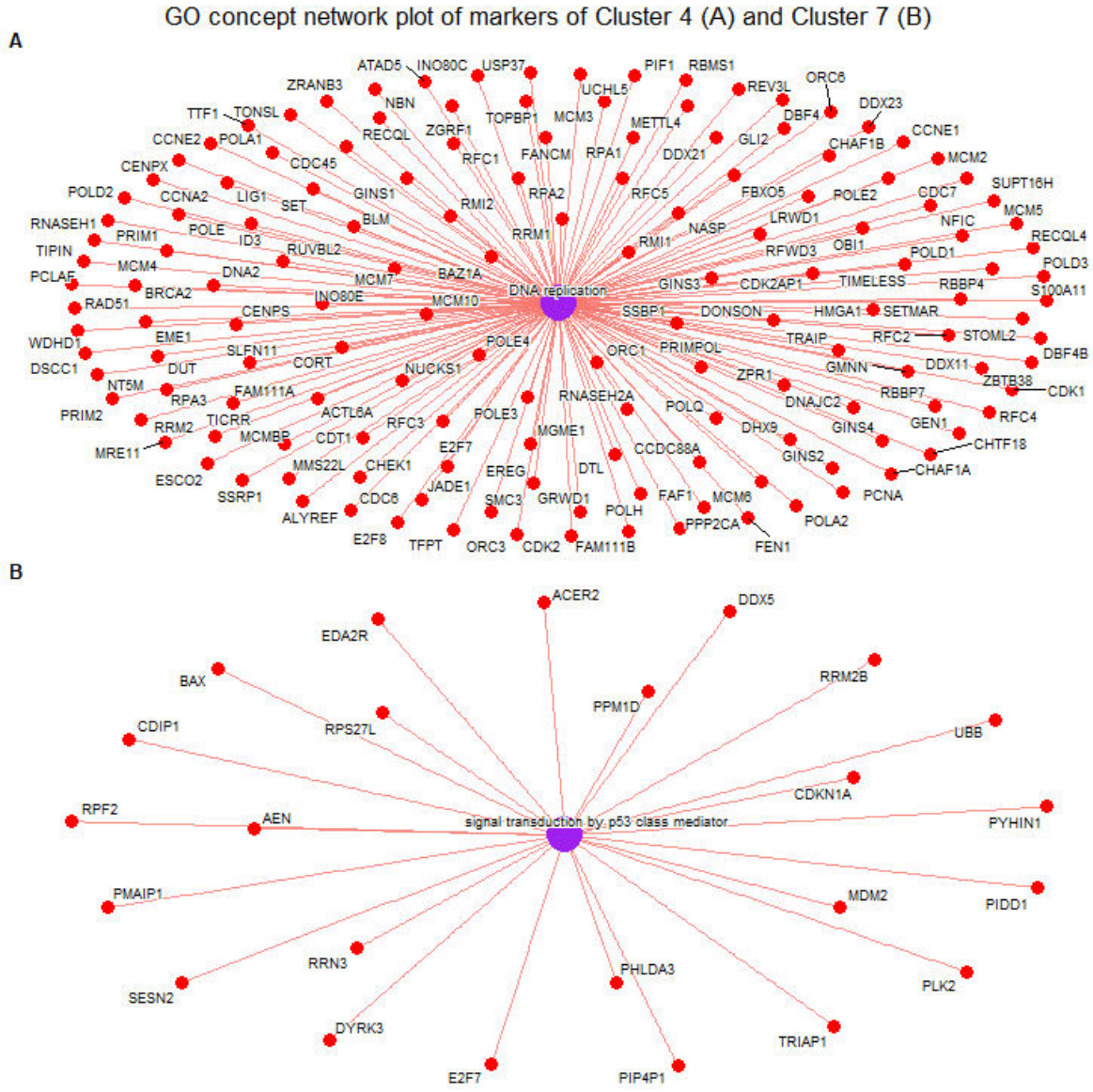


Figure 4.27. Concept network plot of the top 3 enriched GO terms and their associated genes for: **A)** Cluster 4; **B)** Cluster 7.

The functional characterizations of the clusters are listed in **Table 4.4** and illustrated in **Figure 4.28**. These will be used in the subsequent part of this analysis instead of cluster numbers.

Cluster number	**Functional characterization**
0	Actin organization
1	Hypoxia-like
2	Stress response

3	Protein degradation
4	Top stemness
5	Cellular respiration
6	Wnt signalling
7	p53 signalling
8	IF response

Table 4.4. The functional characterization of the clusters.

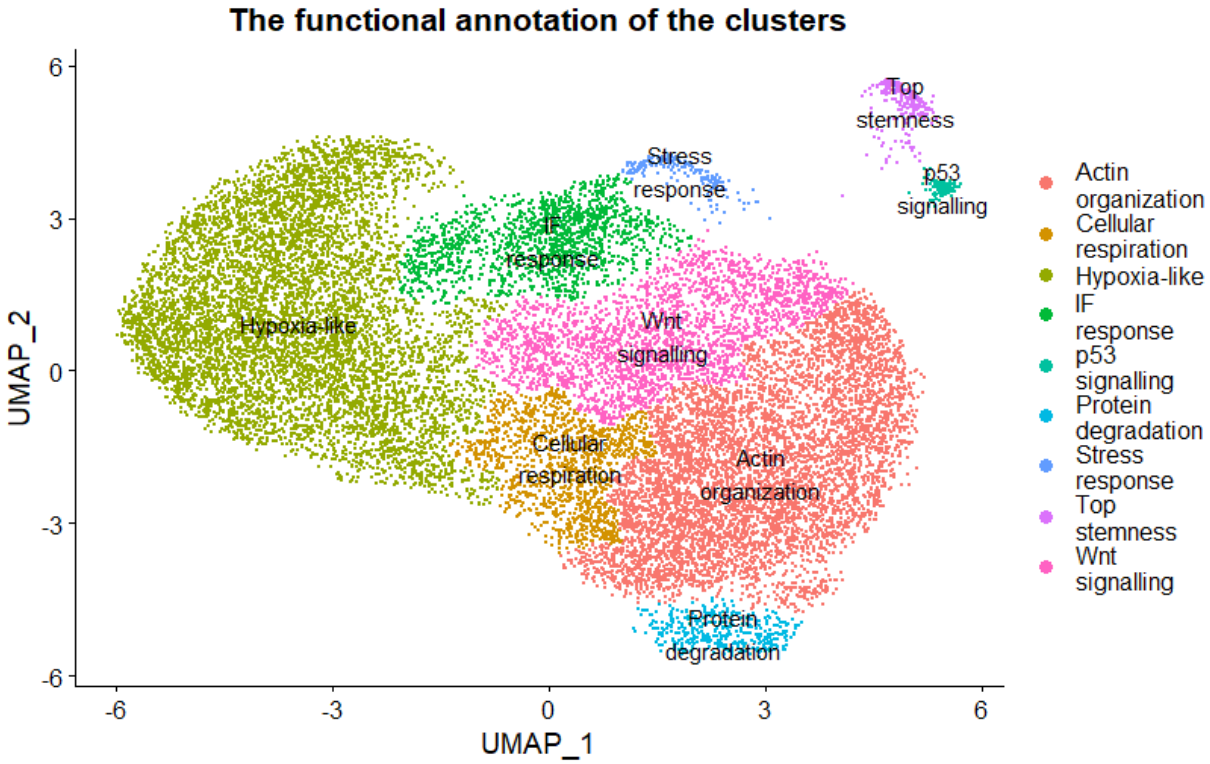


Figure 4.28. The annotation of the clusters.

The next section will analyse the Slingshot lineages and perform an RNA velocity analysis.

4.2.6 Analysis of the differentiation trajectory

Using the evaluateK function, the optimal number of knots for the tradeSeq generalised additive model used to fit to the Slingshot trajectories was found to be 10, and the generalised additive model was generated using this parameter. As using all the 23025 genes in the dataset would have been prohibitive in terms of computational resources, only the most informative genes were included for the fitting of the generalised additive model, consisting of the union of cluster markers (11192 genes) and the variable features of the Seurat object (4386 genes), 11639 genes in total.

As Lineage 2 ended in a cluster close to the **Top stemness** one on the UMAP plot, namely **Stress response**, its late evolution was examined. Lineage 2 showed only minimal differential gene expression after knot 9, so the region between knot 8 (pseudotime: 15.47; pseudotime percentile: 77.78%) and knot 9 (pseudotime: 16.21; pseudotime percentile: 88.89%) was assessed for upregulated genes. 31 genes strongly upregulated (p-value < 0.05; median fold-change > 0.8; Wald test score > 160) were found: *PPP1R15A*, *HSPH1*, *HSPA1A*, *HSP90AA1*, *JUN*, *HSPD1*, *DNAJB1*, *HSPA1B*, *BAG3*, *GADD45A*, *DNAJB4*, *UBC*, *EXT1*, *DNAJB6*, *OTUD1*, *TNFRSF10B*, *HSP90AB1*, *MXD1*, *ARL4A*, *GADD45B*, *ZFAND2A*, *BBC3*, *SEMA3A*, *CHORDC1*, *JUNB*, *HIP1R*, *MIR4713HG*, *HSPA8*, *HSPB8*, *REV3* and *BTG1*, with significantly enriched GO terms linked to heat response and response to unfolded protein. Thus, the tradeSeq assessment finds no evidence of dedifferentiation of the **Stress response** cluster to the **Top stemness** one, as the markers and processes associated with the late stage of this lineage are not linked to the **Top stemness** cluster.

The Slingshot pseudotime combined over the lineages had values between between 0 and 18.4, and an absolute Pearson correlation of 0.21 with the ORIGINS activity score. No genes were found expressed exclusively at the apex of the developmental potential hierarchy (highest activity, lowest Slingshot pseudotime), as illustrated in **Figure 4.29**.

Next, characteristic gene sets (**Section 2.3.10**), were determined. The set of activity medians and median per gene had 79 and 82 characteristic genes each (**Figure 4.30**). The set of Slingshot medians and means per gene had 60 and 54 characteristic genes, respectively (**Figure 4.31**).

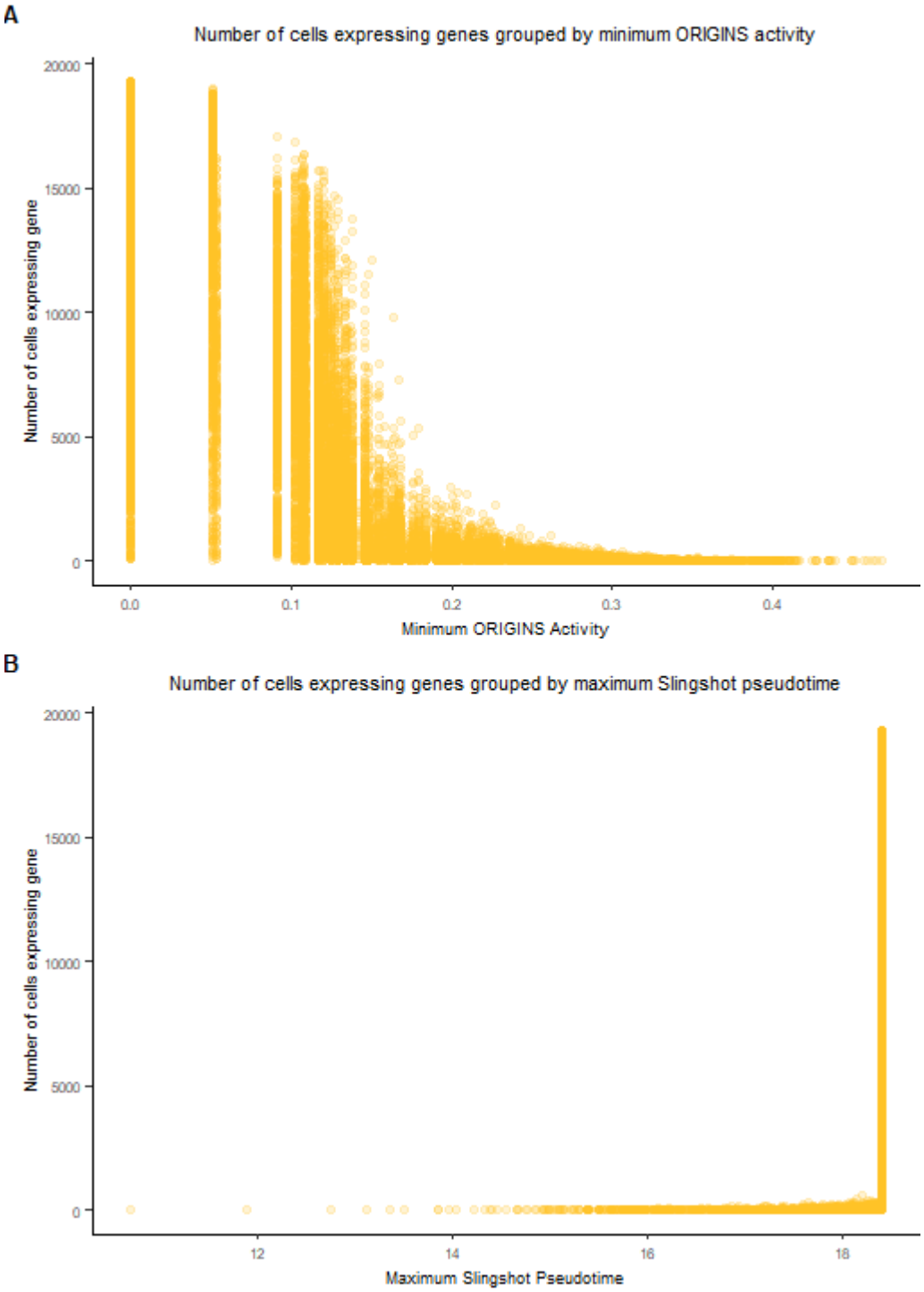


Figure 4.29. The distribution of the number of cells expressing each gene, grouped by: (**A**) the ORIGINS activity minima; (**B**) Slingshot pseudotime maxima of each gene.

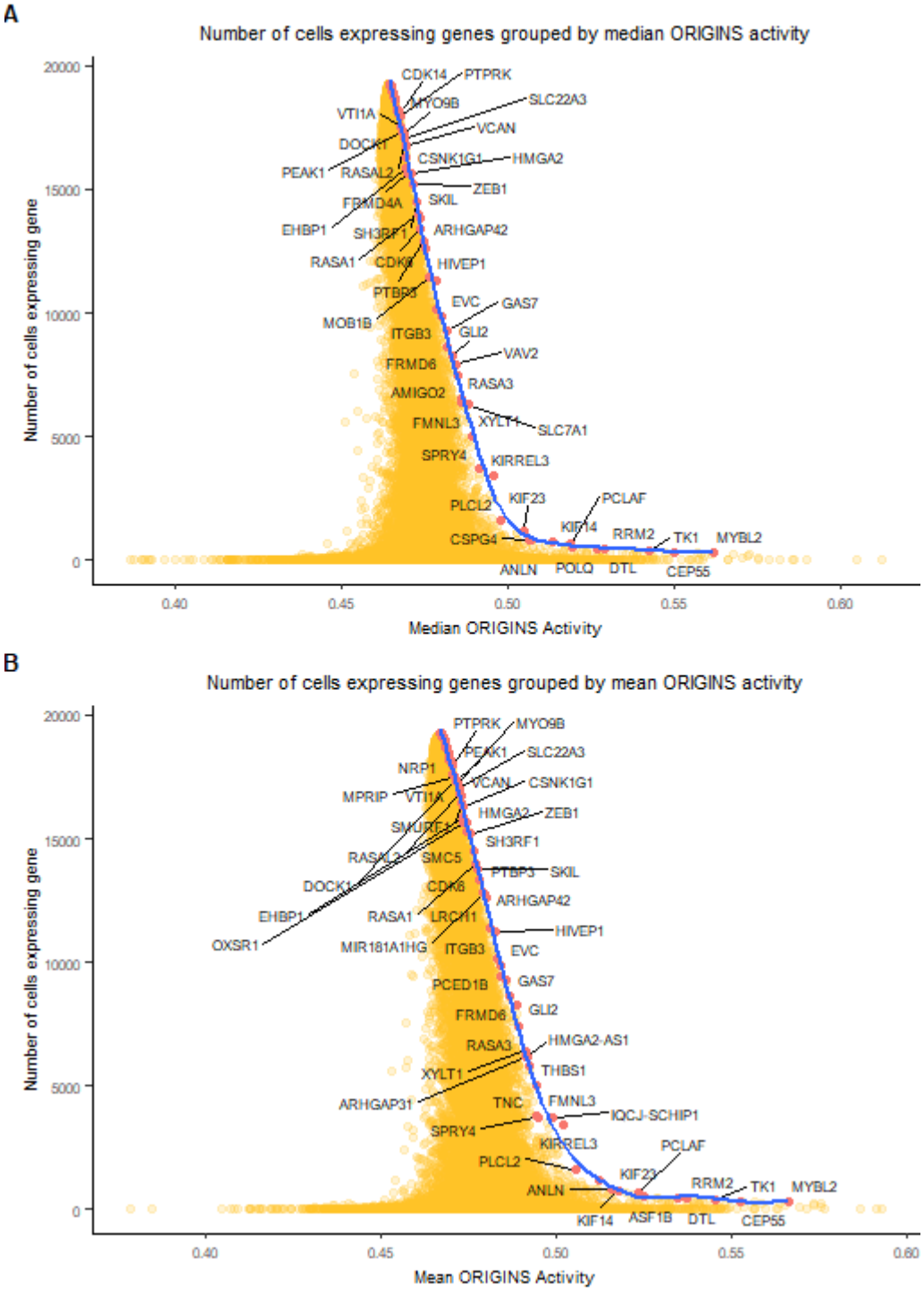


Figure 4.30. The distribution of the number of cells expressing each gene, grouped by: ORIGINS activity medians (**A**); ORIGINS activity means (**B**). The characteristic genes are also displayed on the plot.

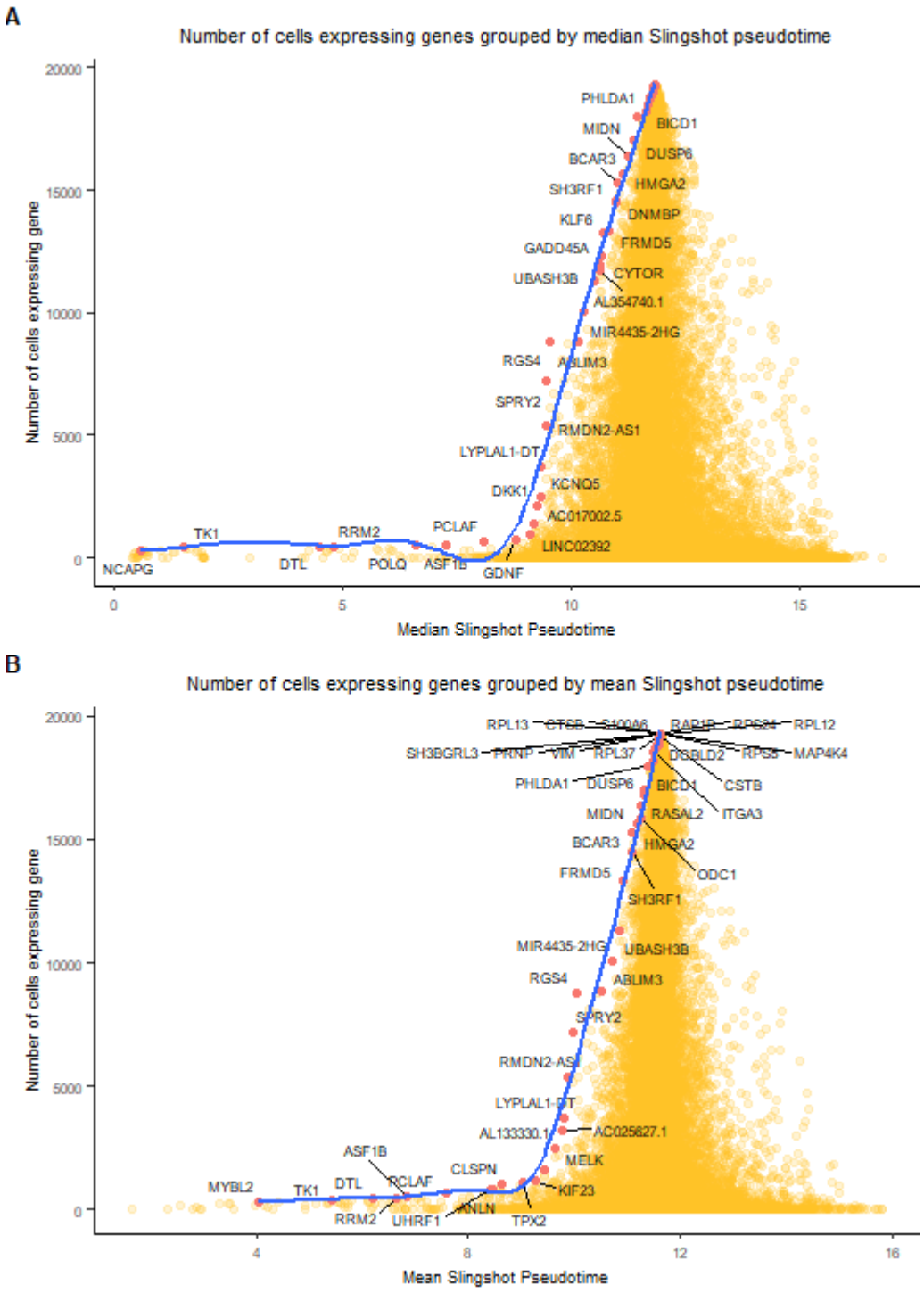


Figure 4.31. The distribution of the number of cells expressing each gene, grouped by: Slingshot pseudotime medians (**A**); Slingshot pseudotime means (**B**). The characteristic genes are also displayed on the plot.

14 genes appeared in at least three of the four characteristic gene sets (**Table 4.5**):

Genes	**Number of occurrences in the characteristic gene sets**
TK1, SH3RF1, S100A6, RRM2, PCLAF, HMGA2 and *DTL*	4
RASAL2, PALM2-AKAP2, MYBL2, KIF23, HIPK2, ASF1B and *ANLN*	3

Table 4.5. The 14 genes appearing at least thrice in the four characteristic gene sets, and their number of occurrences.

All the four pairs of characteristic gene sets overlapped to a statistically significant extent (**Table 4.6**), with p-values ranging from 2.1e-131 (the Activity medians set and Activity means set) to 4.02e-11 (Slingshot pseudotime medians set and Activity means set), indicating that similar genes are linked with the variation in both assays of developmental potential. The cardinalities of all intersections of two, three and all four characteristic gene sets are illustrated in **Figure 4.32**. Furthermore, all four characteristic gene sets showed significant overlap with the union CCRSA gene set, with p-values ranging from 6.05e-11 (for the Slingshot pseudotime means set) to 6.06e-07 (for the Slingshot pseudotime medians set).

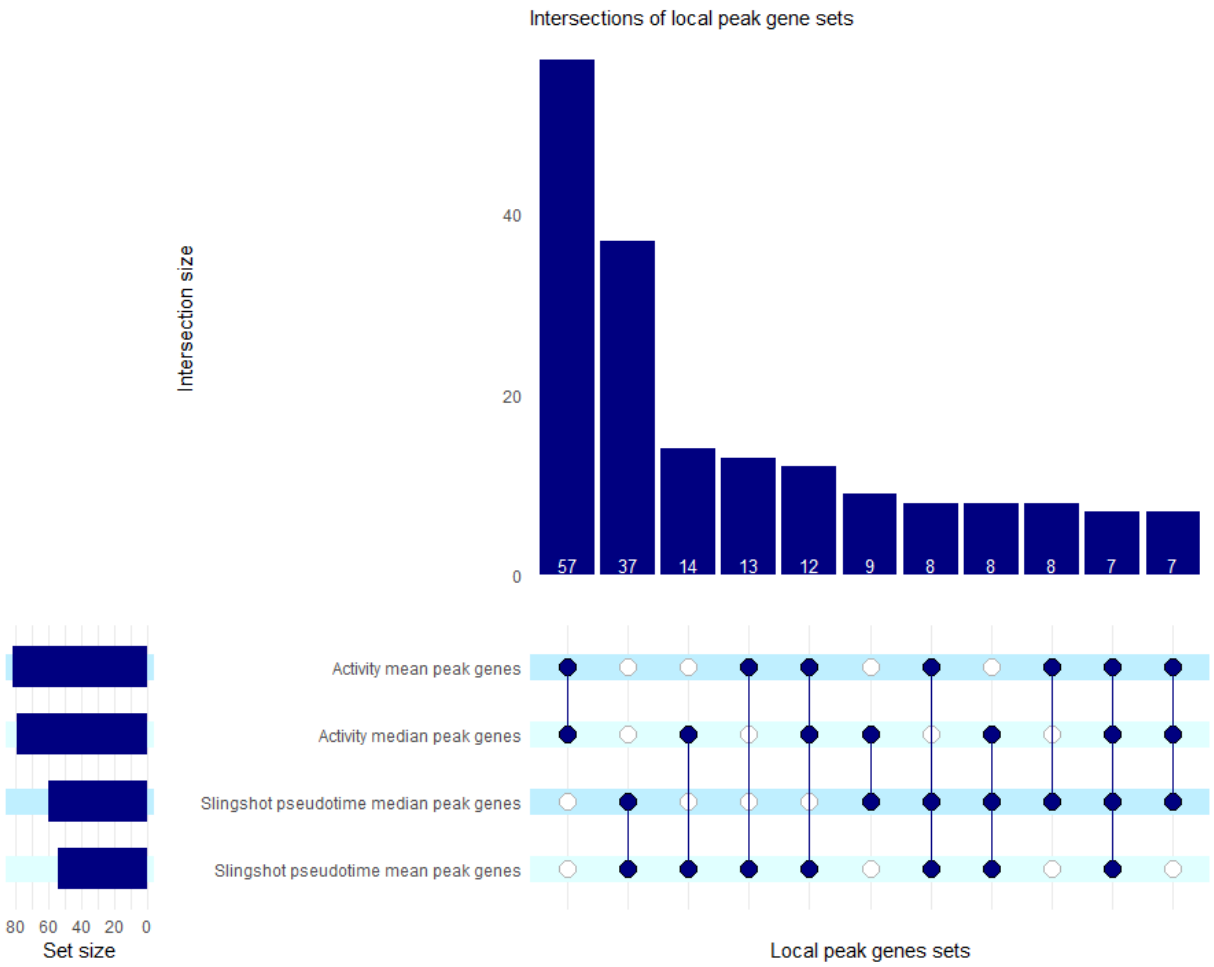


Figure 4.32. The intersections of the characteristic gene sets.

7 out of the 14 genes found in at least three characteristic genes sets were non-CCRSA genes: *HMGA2*, *S100A6*, *SH3RF1*, *HIPK2*, *MYBL2*, *PALM2-AKAP2* and *RASAL2*. *MYBL2* was a marker of the **Top stemness** cluster, and peaks in the expression of other six genes corresponded to local peaks in ORIGINS activity scores, while the joint density of the seven genes marked a region of the **Top stemness** cluster (**Figure 4.33**)

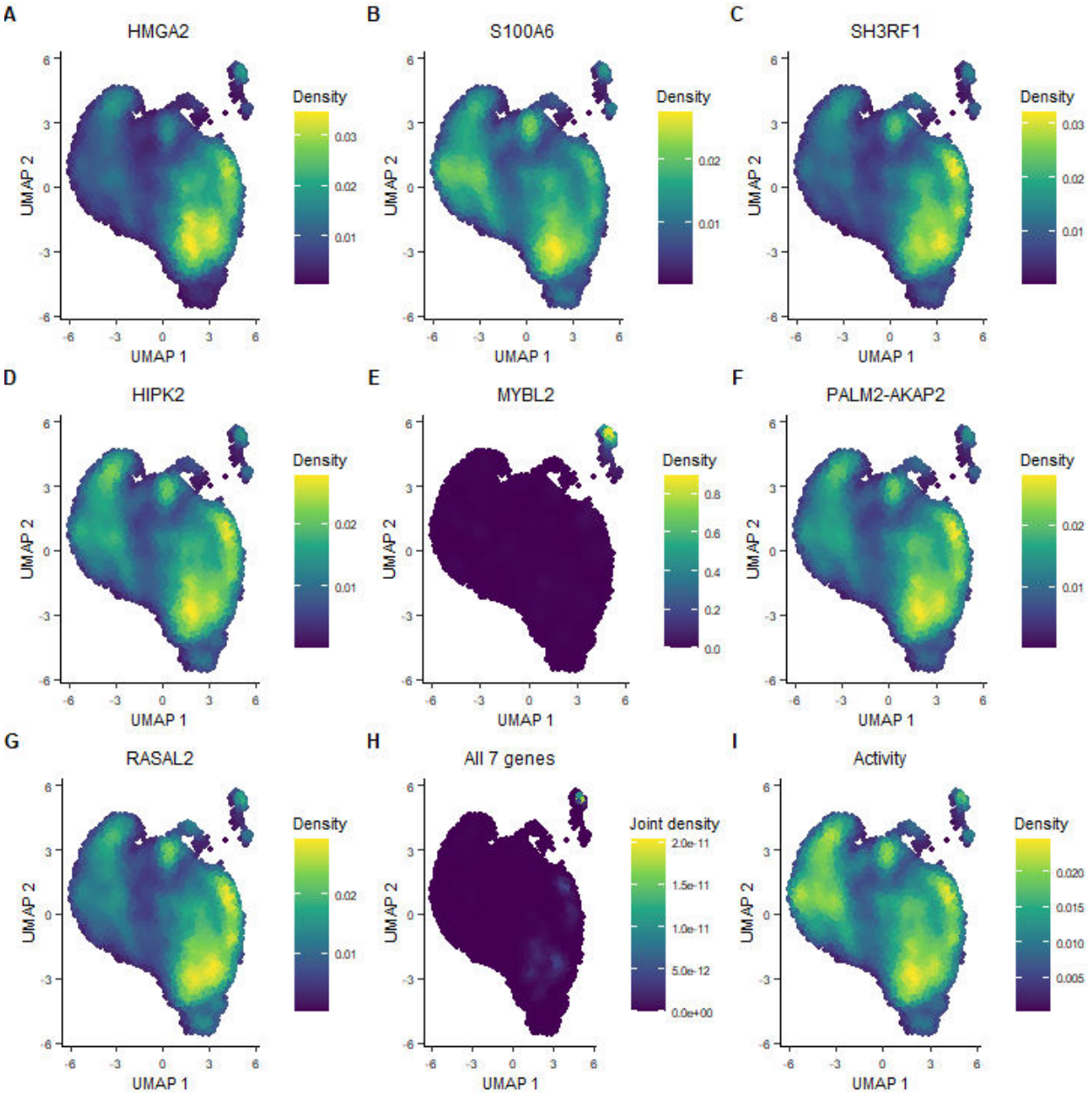


Figure 4.33. Nebulosa plots of 7 genes found in the intersection of at least three characteristic sets but not among the CCRSA genes: **A)** HMGA2; **B)** S100A6; **C)** SH3RF1; **D)** HIPK3; **E)** MYBL2; **F)** PALM2-AKAP2; **G)** RASAL2, of **H)** their joint density, and of **I)** ORIGINS activity.

Finally, the RNA velocity assessment (**Figure 4.34**) found that trends in the evolution of nascent RNA from **Stress response** cluster towards the bulk of the cells. Meanwhile, movement from the **p53 signalling** cluster to the **Top stemness** cluster appears possible, indicating that the former cluster retains an ability of reverting to a CSC phenotype. The **Top stemness** cluster displayed both peripheral cells in the cluster being attracted to its centre, but also possible differentiation

through an intermediary stage not otherwise evidence in this dataset, towards the left side of the cluster. This suggests that the differentiation of CSCs can occur through multiple trajectories, creating distinct intermediary populations in the process.

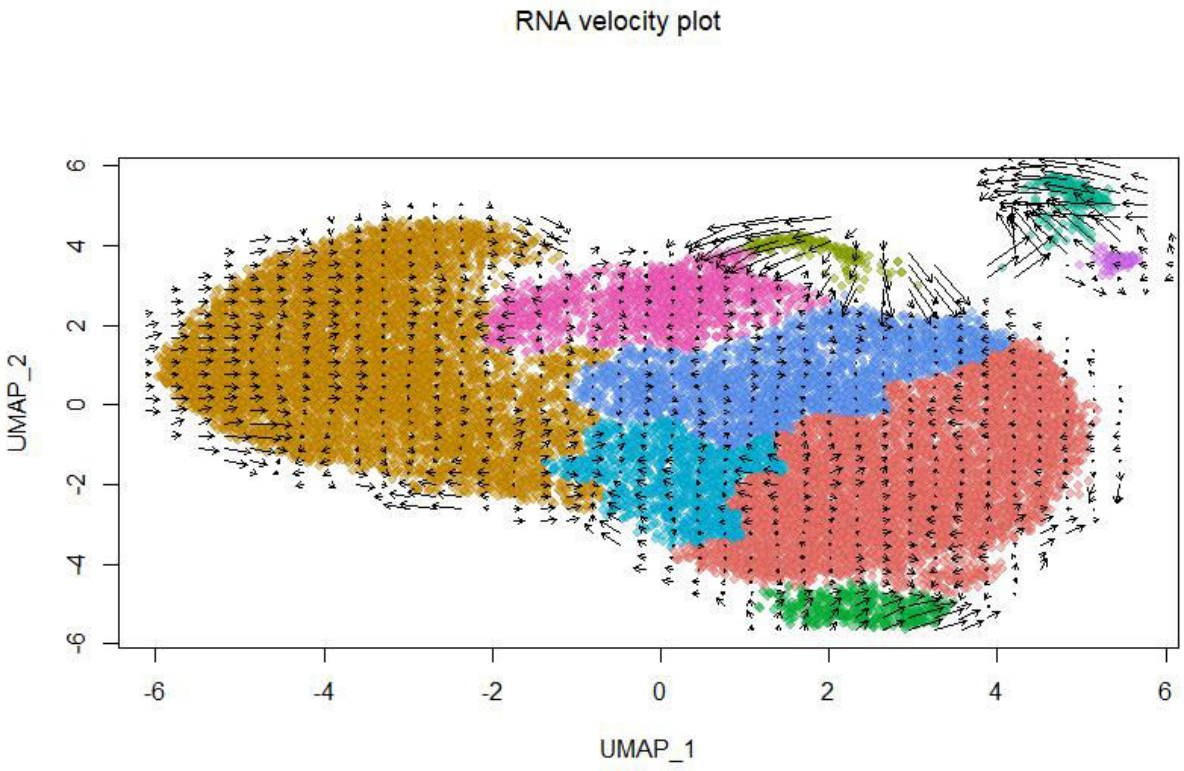


Figure 4.34. RNA velocity plot of the cells in the dataset.

To conclude, a capacity of dedifferentiation towards a **Top stemness** phenotype was found for the **p53 signalling** cluster but not for the **Stress response** one, while sets of genes marking the change in developmental potential were found to significantly overlap between ORIGINS activity and Slingshot pseudotime. The next section will discuss cell-cell communication.

4.2.7 Analysis of cell-cell communication

SingleCellSignalR analysis revealed CALM1/HMMR interactions involving the **Top stemness** cluster and multiple distinctive interactions for the **Stress response** cluster (**Figure 4.35**).

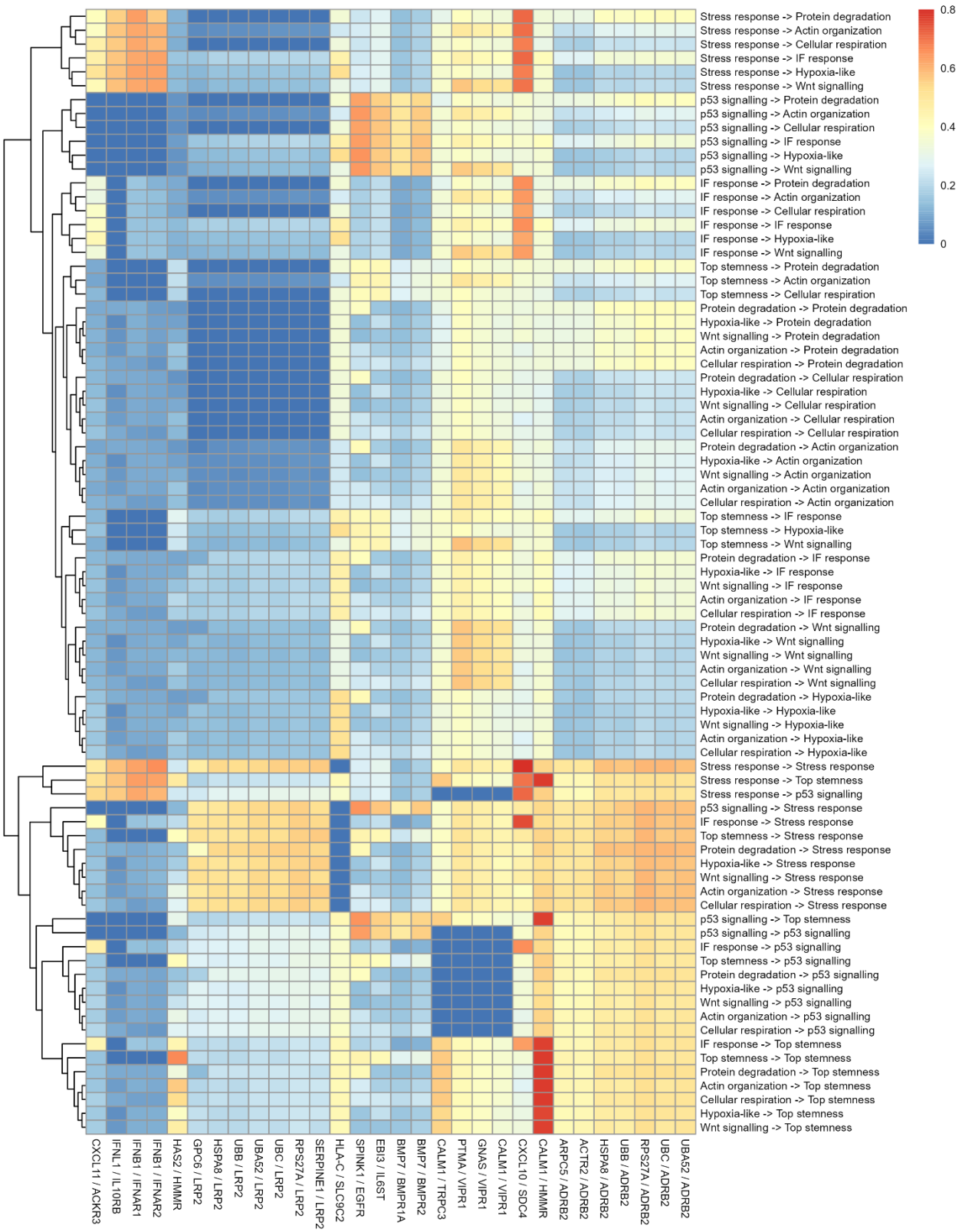


Figure 4.35. Top 30 variable cell-cell interactions found by SingleCellSignalR.

CellChat analysis found 58 signalling pathways significantly activated between pairs of clusters or within clusters. Among these, MK, EGF, MIF, THY1, PTN, EDA, CLDN and PECAM1 signaling were distinctly activated for the **Top stemness** cluster (**Figure 4.36**).

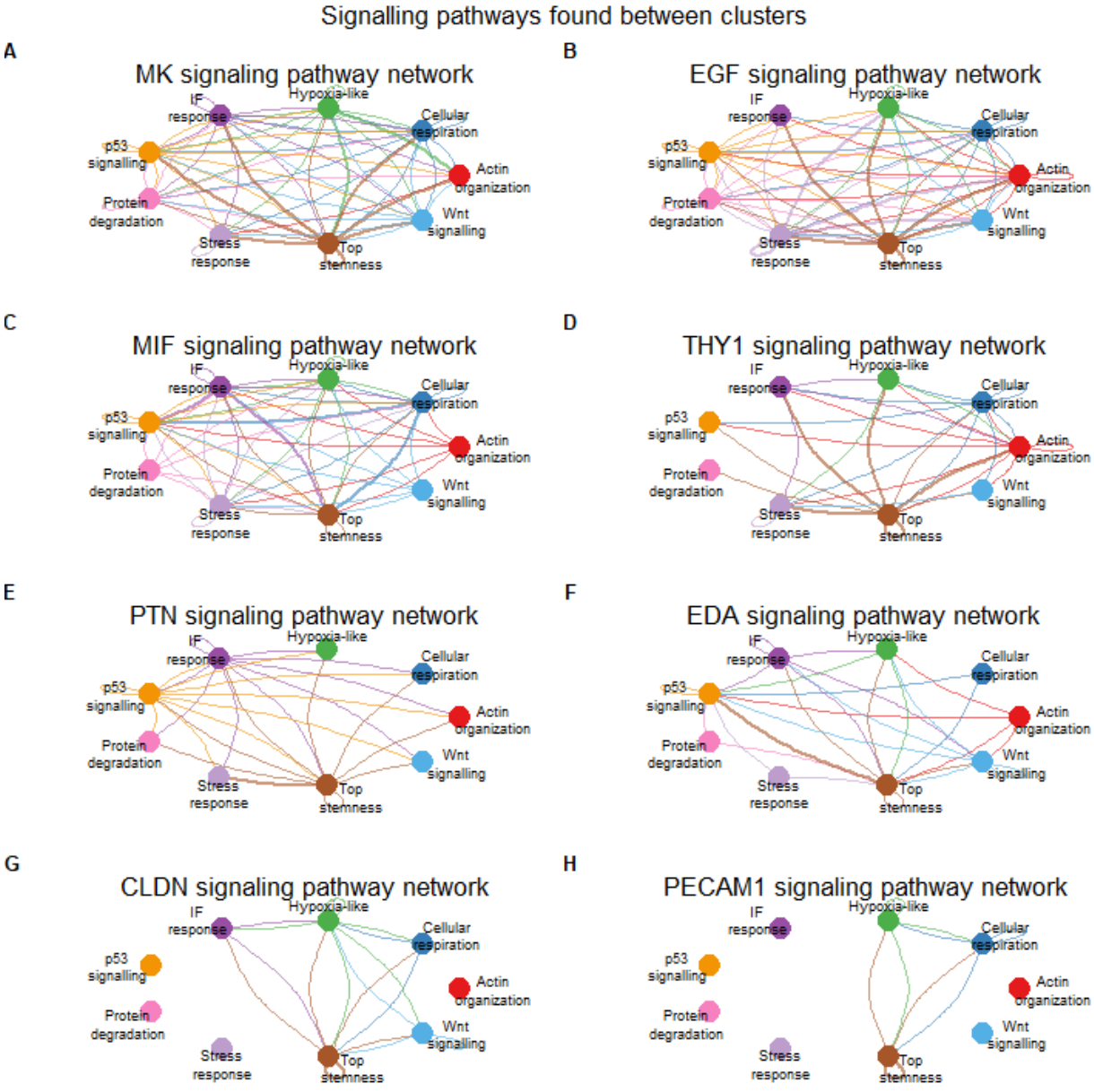


Figure 4.36. Signalling pathways distinctly activated for the Top stemness cluster: **A)** MK; **B)** EGF; **C)** MIF; **D)** THY1; **E)** PTN; **F)** EDA; **G)** CLDN; **H)** PECAM.

Other pathways with a differential pattern of activation are desmosome, PDGF and PROS signalling (strongest for the **Wnt signalling** cluster), PTPRM signalling (shared between the **IF**

response, **Hypoxia-like**, **Actin organization** and **Wnt signalling** clusters), SPP1 signalling (strongest for the **Actin organization** cluster), SEMA6 and JAM signalling (for the **Hypoxia-like** cluster) and IFN-I signalling (strongest for the **Stress response** cluster), all illustrated in **Figure 4.38**.

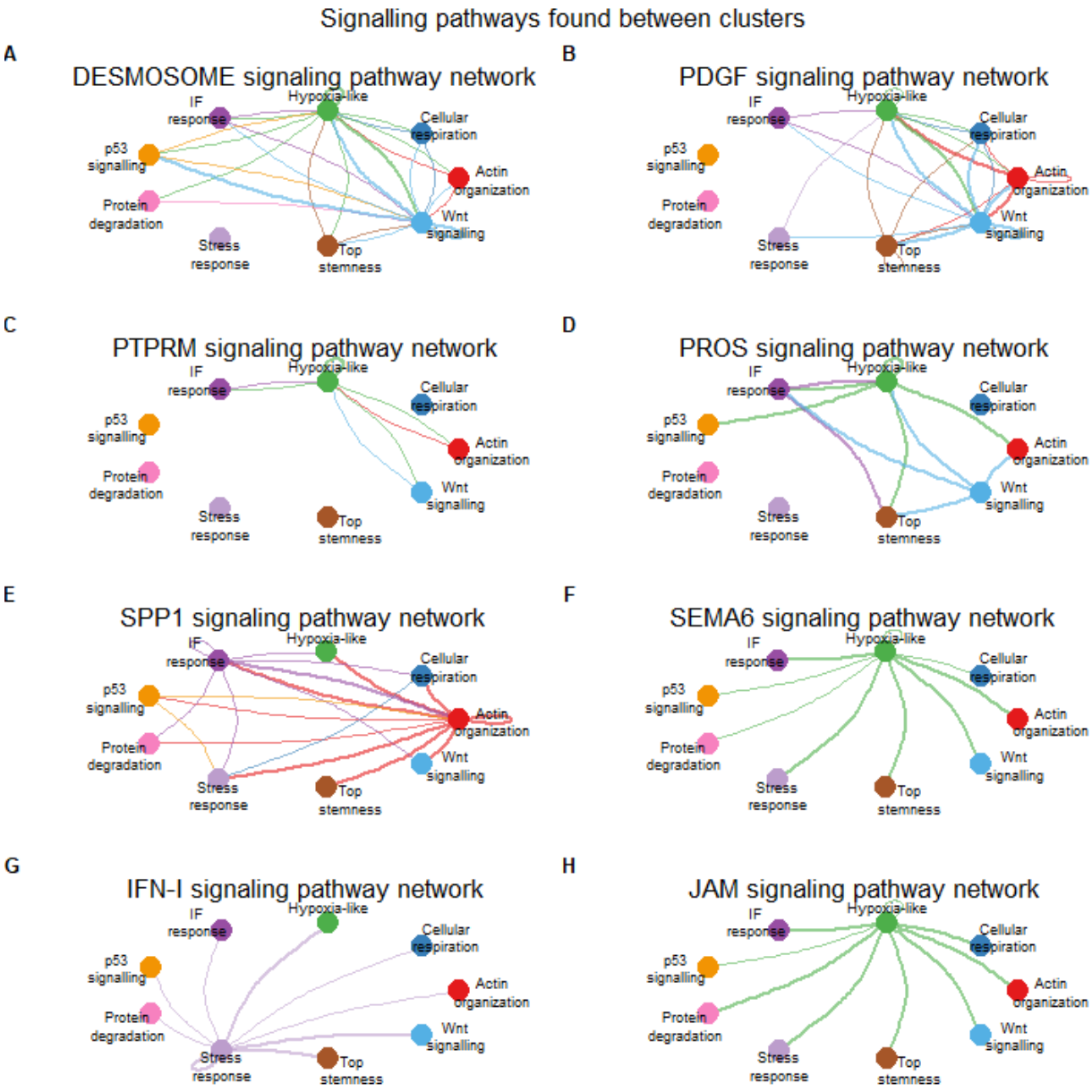


Figure 4.37. Signalling pathways distinctly activated for clusters other than Top stemness: **A)** Desmosome; **B)** PDGF; **C)** PTPRM; **D)** PROS; **E)** SPP1; **F)** SEMA6; **G)** IFN-I; **H)** JAM.

In addition, four patterns of communications of secreting cells (**Figure 4.38**) and six patterns of communications of target cells (**Figure 4.39**) were identified. In both cases, the pattern of **Top stemness** was not shared with any other cluster.

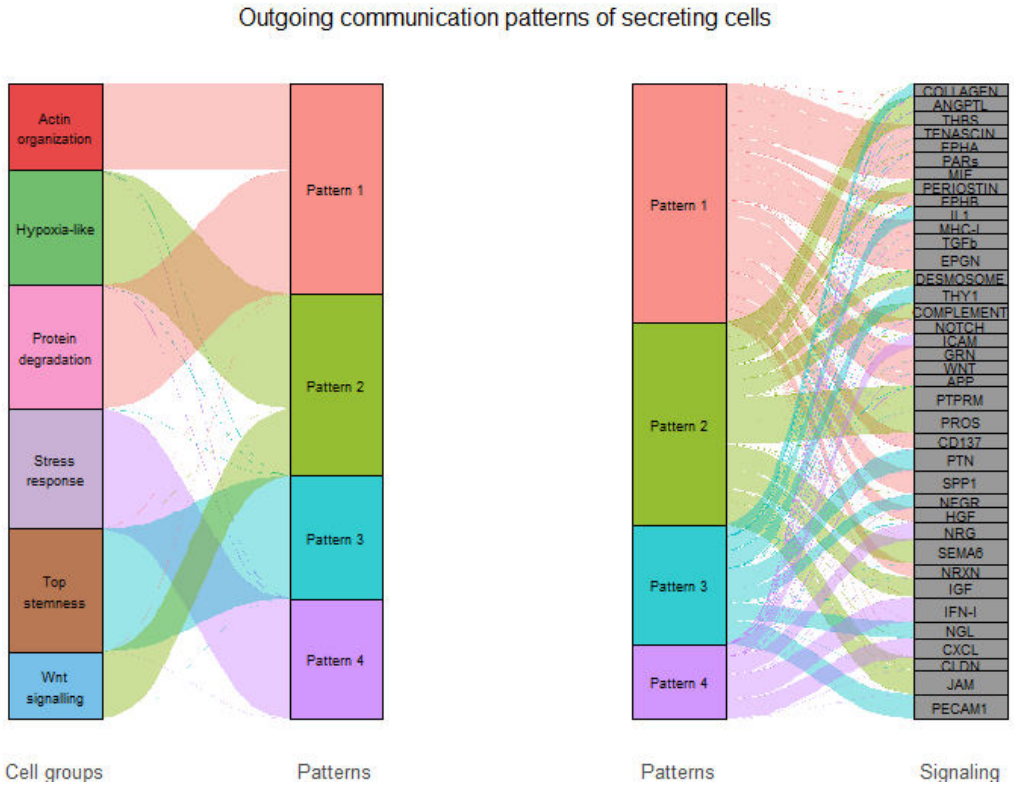


Figure 4.38. The clusters grouped by their outgoing communication patterns of secreting cells, and the pathways corresponding to each pattern.

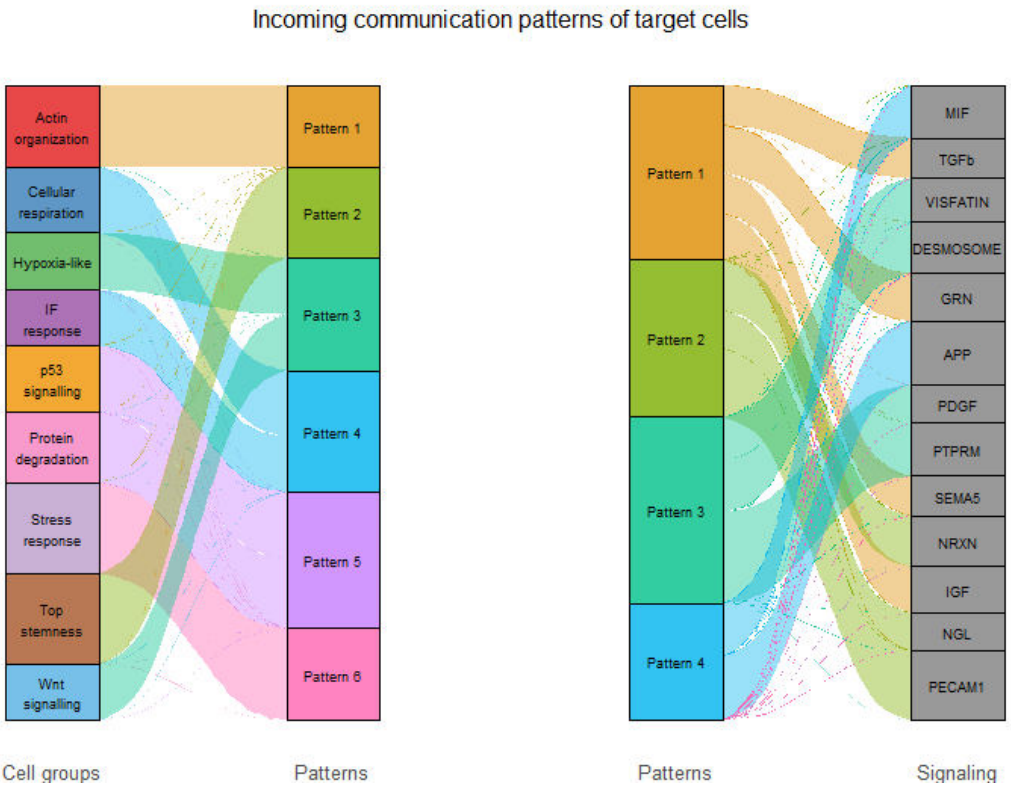


Figure 4.39. The clusters grouped by their incoming communication patterns of secreting cells, and the pathways corresponding to each pattern.

In **Figure 4.41**, both outgoing and incoming interactions are displayed for each cluster.

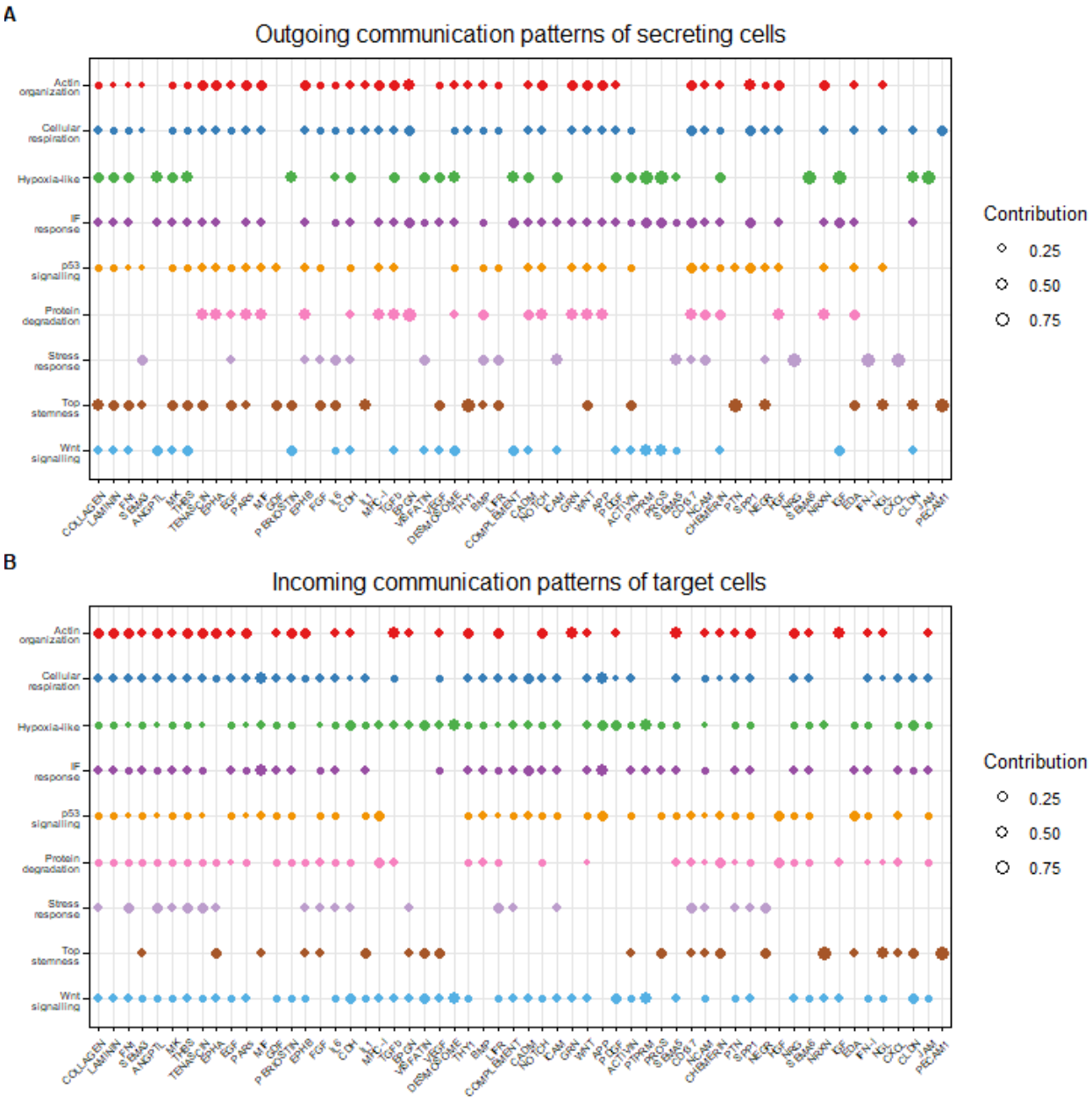


Figure 4.40. Dot plot displaying the activation of **A)** outgoing and **B)** incoming signalling pathways for each cluster.

Overall, the analysis has determined commonalities in cell signalling between the **Hypoxia-like** and **Wnt signalling** clusters, while **Top stemness** shows a distinctive pattern of signalling. The next section will discuss the epigenetics processes enriched for the markers of the clusters, in order to assess their importance to stemness-linked clusters in this dataset.

4.2.8 The epigenetic mechanisms characterizing the clusters

120 GO terms linked to epigenetics mechanisms were found in association with the genes detected in the data. Among these GO terms, 13 were found enriched for the cluster markers, distributed among clusters as follows: 7 in **Top stemness** (**Figure 4.41**), 6 in **Wnt signalling**, 3 in **Stress response**, and 1 in **Hypoxia-like**. No cluster evidenced a statistically significant overrepresentation of epigenetics-related terms among its enriched GO terms.

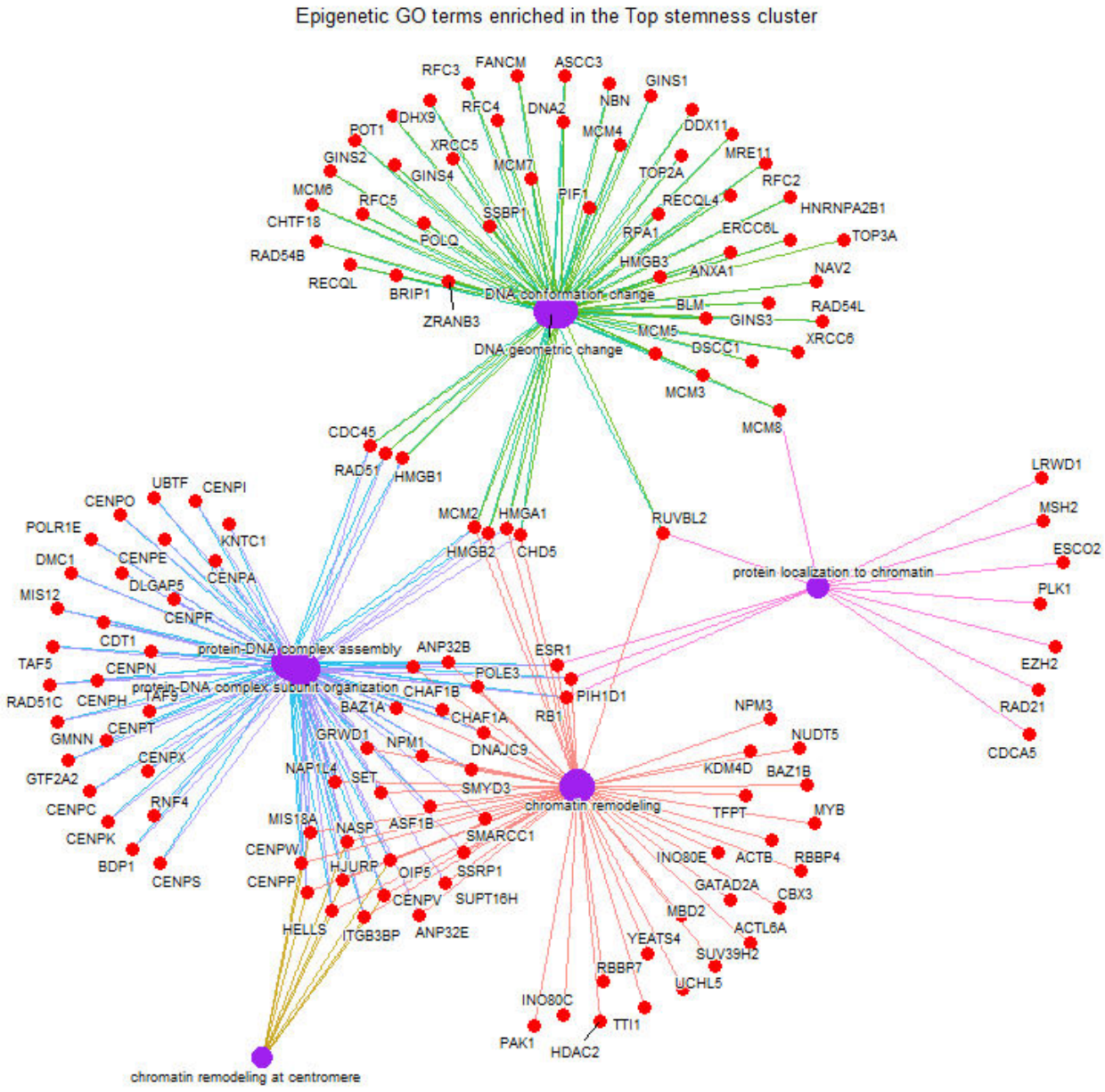


Figure 4.41. Epigenetic GO terms enriched in the Top stemness cluster.

In conclusion, only a few epigenetics processes are enriched for the clusters, with the highest number of them found in the **Top stemness** cluster, represented mostly by changes in the geometry and conformation of DNA and in the assembly and organization of protein-DNA complexes. The remainder of this chapter will look at the effects of the treatments, starting with assessment of the changes upon cancer stemness at the pseudo-bulk level.

4.2.9 The global (pseudo-bulk) comparison of stemness between experimental conditions

The five TDPM genes overexpressed in the **Top stemness** cluster (*EZH2*, *NES*, *THY1*, *ABCG2* and *ABCB1*) were co-expressed in only 1 cell. When only the stronger markers *EZH2*, *NES* and *THY1* were assessed, I-BRD9 and gemcitabine and I-BRD9 cells were underrepresented (adjusted p-values: 5.36e-05 and 4.57e-02). The *EZH2+NES+THY1+* signature was detected in 35 DMSO cells, 38 Gemcitabine cells, 21 I-BRD9 cells and 18 I-BRD9 and gemcitabine cells. **Figure 4.42** displays a kernel density plot corresponding to this signature:

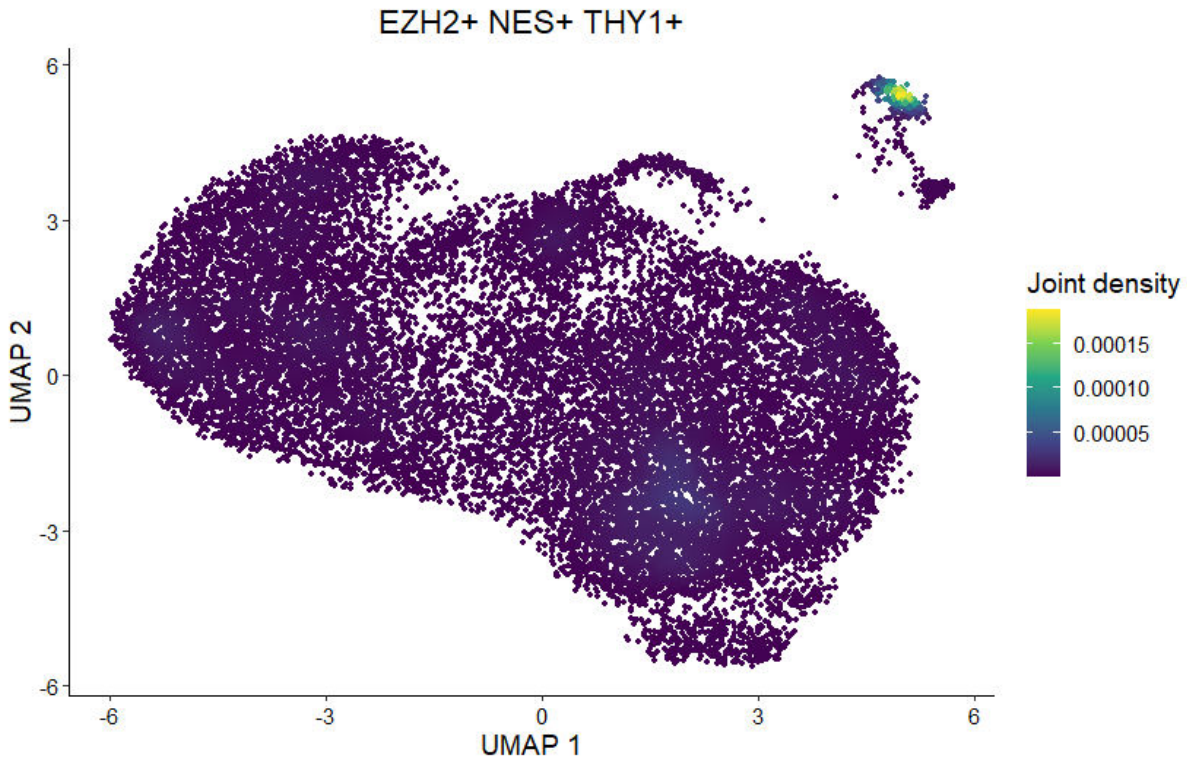


Figure 4.42. Kernel density plot of the EZH2+NES+THY1+ signature.

In order to investigate the link between the changes in the expression of stemness-linked genes in response to the treatments, overlaps between the CCRSA gene sets and the markers of experimental conditions were assessed. The condition selections considered were all the 4 x 3 = 12 selections of markers of one condition relative to another one and the selection of I-BRD9 conditions relative to the others and its reverse, to a total of 14 condition selections. As before, the notation **[X] vs. [Y]** means that the markers of condition(s) X were taken relative to condition(s) Y. Significant overlaps were identified for only two gene sets, prognosis and union (**Table 4.6**).

Cancer gene set	**Condition selections with significant marker overlap**	
	Condition selection	**p-value**
Prognosis	[Gemcitabine] vs. [I-BRD9] [Gemcitabine] vs. [DMSO] [Gemcitabine] vs. [I-BRD9 and gemcitabine] [DMSO, Gemcitabine] vs. [I-BRD9, I-BRD9 and gemcitabine]	1.03e-04 1.03e-04 3.41e-04 8.52e-03
Union	[Gemcitabine] vs. [I-BRD9]	1.5e-02

Table 4.6. Condition selections whose markers significantly overlap with the CCRSA gene sets.

For the SPLCL genes, statistical significance was not reached for any of the 14 ordered pairs of experimental conditions.

The Wilcoxon rank sum test showed significantly higher **ORIGINS activity** scores for the Gemcitabine condition, and significantly lower for the I-BRD9 conditions (**Table 4.7**). Gemcitabine cells were overrepresented among top ORIGINS activity cells, and I-BRD9 (adjusted p-value 1.9e-10) and I-BRD9 and gemcitabine (adjusted p-value 1.15e-02) cells were underrepresented.

Condition pair	p-value (for the alternative hypothesis: greater)
[Gemcitabine] vs. [I-BRD9]	1.9e-45
[DMSO] vs. [I-BRD9]	2.93e-45
[Gemcitabine] vs. [I-BRD9 and gemcitabine]	2.75e-40
[DMSO] vs. [I-BRD9 and gemcitabine]	2.21e-38
[Gemcitabine] vs. [DMSO]	1.87e-02

Table 4.7. Pairs of conditions where the alternative hypothesis ("greater") reached significance for the Wilcoxon test.

In conclusion, this section provides some evidence for a reduction in stemness associated with both I-BRD9 conditions, while the Gemcitabine condition is associated with higher stemness. The effect is most strongly evidenced by the Wilcoxon pairwise assessment of ORIGINS activity scores, but also surfaces in the results coming from the analysis of the representation of the experimental conditions among cells with high ORIGINS activity scores, and among cells displaying the *EZH+NES+* signature.

The next section will assess the effect of treatment conditions upon stemness as evidenced within clusters.

4.2.10 The intra-cluster effects of the treatment conditions upon stemness

The representation of experimental conditions among the cells from each cluster will be analysed, and the TDPM, the CCRSA and the SPLCL genes, followed by ORIGINS activity, will be employed to assess intracluster changes in cancer stemness due to the treatment conditions.

All statistically significant overrepresentations of cells from any of the experimental conditions in any of the clusters are listed in **Table 4.8**.

Condition	Cluster	p-value	Percentage in cluster	Percentage among all cells
DMSO	Hypoxia-like	3.47e-80	29.55	21.96
I-BRD9 and gemcitabine	Actin organization	5.95e-49	42.74	35.02
Gemcitabine	Top stemness	1.71e-30	46.03	14.37
I-BRD9 and gemcitabine	Protein degradation	7.57e-10	51.59	35.02
I-BRD9 and gemcitabine	Wnt signalling	1.55e-06	39.78	35.02
Gemcitabine	p53 signalling	2.16e-06	33.33	14.37
I-BRD9	Hypoxia-like	2.85e-03	30.21	28.65
I-BRD9 and gemcitabine	IF response	3.77e-03	39.32	35.02
Gemcitabine	Hypoxia-like	3.46e-02	15.33	14.37

Table 4.8. The statistically significant overrepresentations of cells from any of the experimental conditions in any of the clusters.

All statistically significant underrepresentations of cells from any of the experimental conditions in any of the clusters are listed in **Table 4.9**.

Condition	Cluster	p-value	Percentage in cluster	Percentage among all cells

I-BRD9 and gemcitabine	Hypoxia-like	6.86e-113	24.9	35.02
DMSO	Actin organization	5.68e-40	16.08	21.96
I-BRD9	Top stemness	1.63e-17	5.86	28.65
DMSO	Protein degradation	1.35e-09	9.26	21.6
DMSO	IF response	6.38e-05	17.28	21.96
I-BRD9 and gemcitabine	Top stemness	6.38e-05	21.34	35.02
Gemcitabine	Protein degradation	7.93e-05	6.88	14.37
Gemcitabine	Actin organization	8.71e-04	12.9	14.37
Gemcitabine	Wnt signalling	2.32e-03	12.04	14.37

Table 4.9. The statistically significant underrepresentations of cells from any of the experimental conditions in any of the clusters.

Thus, a ~4.9-fold underrepresentation of I-BRD9 cells in the **Top stemness** cluster relative to their representation among all cells was detected. For the I-BRD9 and gemcitabine condition, the same figure is ~1.6, indicating a marked but not complete reversal of the effects of I-BRD9 in terms of reducing the prevalence of CSC-liked cells due to the addition of gemcitabine. Meanwhile, Gemcitabine cells are overrepresented in the **Top stemness** cluster by a factor of ~3.2. With the presence of DMSO cells in the same cluster at ~26.78% (~1.2-fold greater than their presence among all cells), this translates into a ~5.9-fold underrepresentation of I-BRD9 cells, and a ~1.9-fold underrepresentation of I-BRD9 and gemcitabine cells among the cells in the **Top stemness** cluster relative to the DMSO control.

Gemcitabine, meanwhile, is overrepresented in the **Top stemness** cluster relative to its representation among all Gemcitabine cells by a factor of ~3.2. Considering the ~1.2-fold overrepresentation of DMSO cells in the same cluster relative to all DMSO cells, this translates to a ~2.7-fold overrepresentation of Gemcitabine cells in the **Top stemness** cluster relative to the DMSO control.

Figure 4.43 displays the cells from each experimental condition plotted separately, showcasing a visible underrepresentation of I-BRD9 cells in the **Top stemness** cluster:

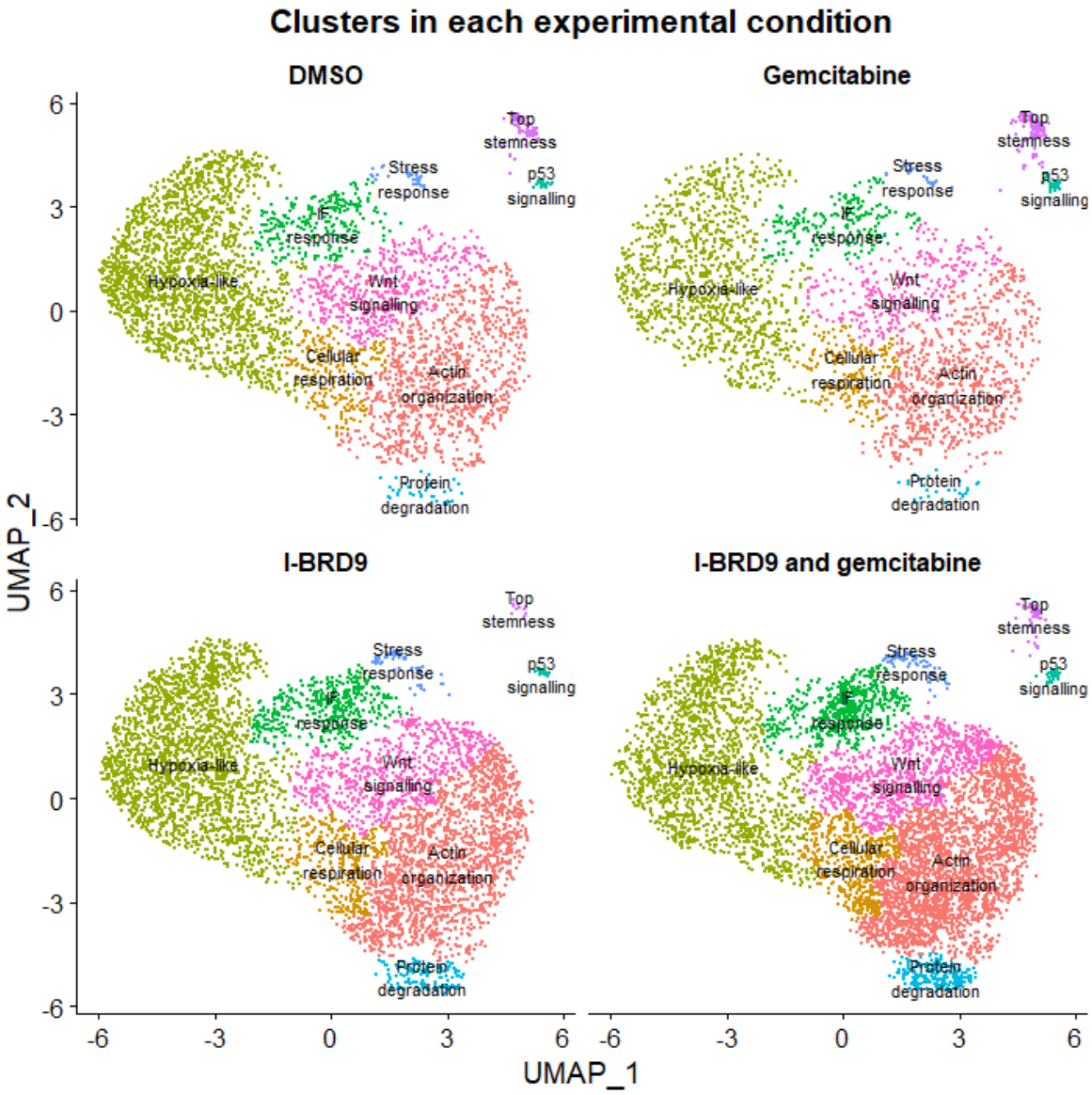


Figure 4.43. Clusters in each experimental condition.

With regards to the *EZH2+NES+* signature, significant underrepresentations of I-BRD9 and I-BRD9 and gemcitabine were detected in the **Hypoxia-like** cluster: 20 DMSO cells, 15 gemcitabine cells, 5 I-BRD9 cells (adjusted p-value: 2.57e-02), 1 I-BRD9 and gemcitabine cell (adjusted p-value: 9.38e-04).

The intra-cluster assessment of changes in stemness due to the treatment conditions returned no significant overlaps for either **the CCRSA gene sets** or the **SPLCL genes**.

To further elucidate the mechanisms behind the underrepresentation of I-BRD9 cells in the **Top stemness**, the unfiltered markers (that is, the outputs from FindMarkers used without applying any p-value cut-off) of the [DMSO] vs. [I-BRD9] and [DMSO] vs. [I-BRD9 and gemcitabine] selections were assessed for intersections with the CCRSA genes. The rationale for this approach was that the small size of the **Top stemness** cluster would render statistical significance for the intracluster markers of condition selections difficult to reach using the Wilcoxon test of FindMarkers, unless the effect in question is very strong. But because the CCRSA genes are closely related functionally, the **collective trend** that might emerge from the distribution of CCRSA genes as markers of condition selections is unlikely to be spurious, even when significance was not reached for individual CCRSA markers given the constraint imposed by the small cluster size.

112 out of the 117 CCRSA genes appeared as unfiltered markers of the [DMSO] vs [I-BRD9] selection within the **Top stemness** cluster, with only 4 CCRSA genes appearing as markers of the opposite selection. One CCRSA gene, *E2F2*, did not appear in either group, due to being found exclusively in gemcitabine-treated cells with within the **Top stemness** cluster.

In **Figure 4.44**, the 112 genes are displayed. *HJURP*, *TOP2A*, *PBK*, *CENPF* and *CDK1* appear as outliers, with average $\log_2$ fold changes recorded for the [DMSO] vs. [I-BRD9] selection above 1.2.

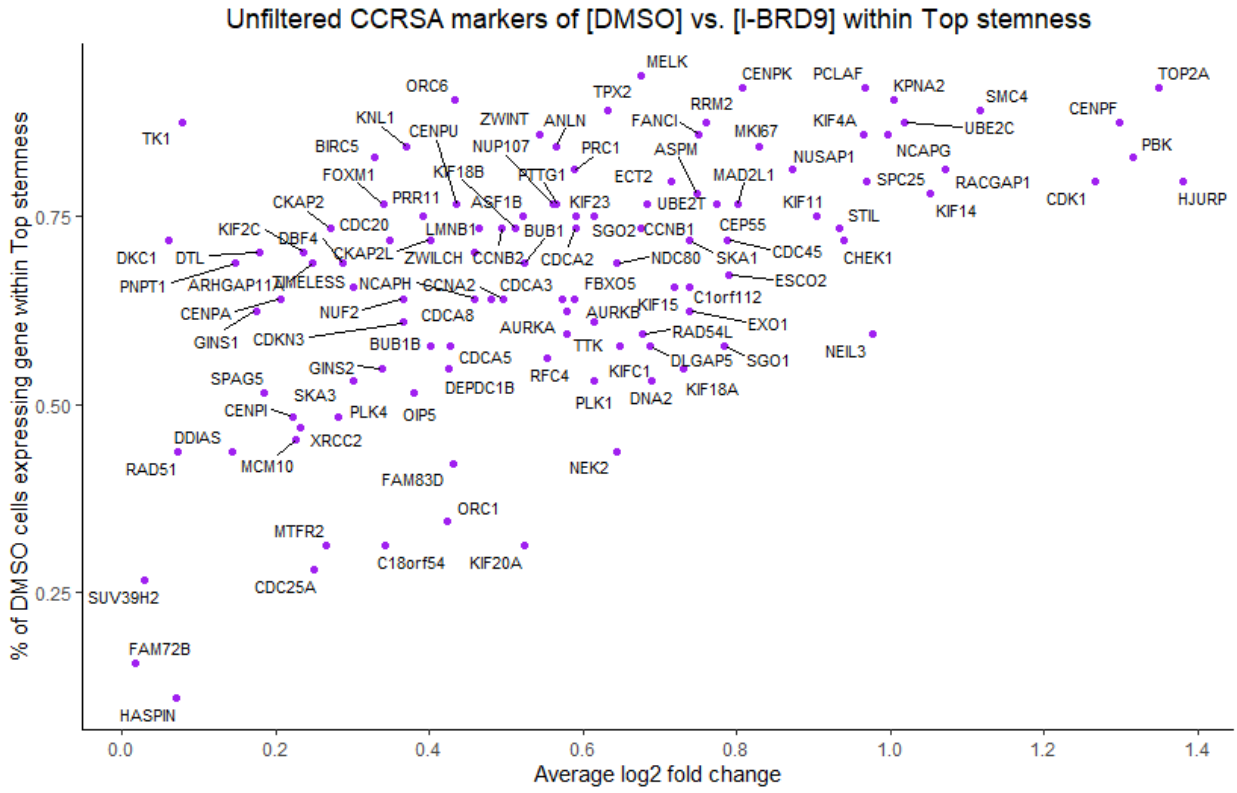


Figure 4.44. Unfiltered CCRSA markers of [DMSO] vs. [I-BRD9] within the Top stemness cluster.

For the [DMSO] vs [I-BRD9 and gemcitabine] selection, a reduction of the effect is seen, paralleling the findings regarding the distribution of I-BRD9 and gemcitabine cells within the **Top stemness** cluster. When no p-value filters where applied, 89 CCRSA genes mark this condition selection within the **Top stemness** cluster compared to 28 CCRSA genes marking the opposite condition selection. No CCRSA genes that marked neither selection existed.

In **Figure 4.45**, the 89 genes are displayed. *HJURP*, *TOP2A*, *CCNB1*, *CENPF*, *UBE2C* and *KPNA2* are visual outliers on the plot, with average $\log_2$ fold changes recorded for the [DMSO] vs. [I-BRD9 and gemcitabine] selection above 1.2.

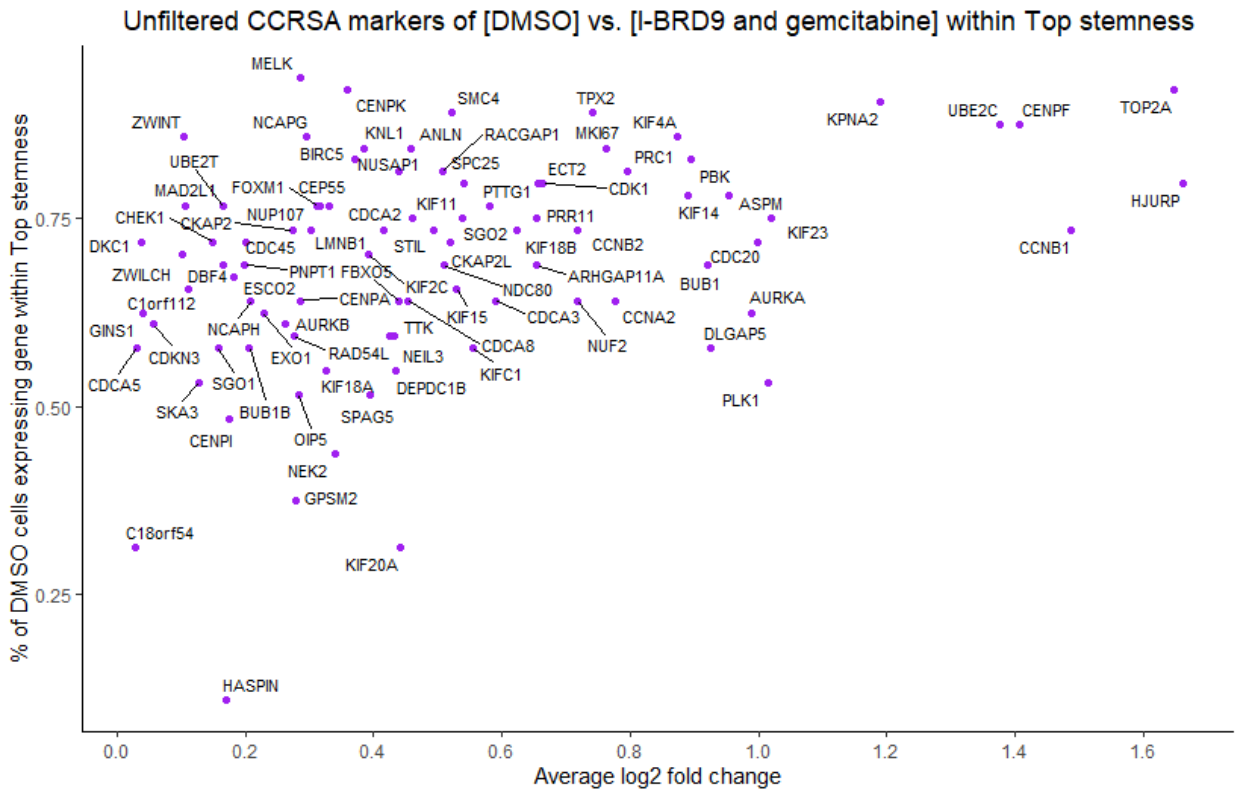


Figure 4.45. Unfiltered CCRSA markers of [DMSO] vs. [I-BRD9 and gemcitabine] within the Top stemness cluster.

The trend of downregulation of the CCRSA genes in the I-BRD9 conditions disappeared entirely in the **p53 signalling** cluster: 32 CCRSA genes appeared as unfiltered markers of [DMSO] vs. [I-BRD9], while 45 CCRSA genes did so for the opposite selection. This finding matches with the lack of significant underrepresentation of I-BRD9 cells in this cluster relative to the DMSO cells. With regards to the remaining 40 CCRSA genes, 23 were found expressed only in Gemcitabine cells (the number of Gemcitabine and I-BRD9 and gemcitabine cells in this cluster were very close, 41 and 40), 12 were found expressed only in Gemcitabine and I-BRD9 and gemcitabine cells, 1 was found expressed only in I-BRD9 and gemcitabine cells, and 4 were not found expressed at all within this cluster.

Furthermore, the **p53 signalling** cluster displayed a salient upregulation of unfiltered CCRSA markers of the [Gemcitabine] vs. [DMSO] selection, paralleling the ~3.3-fold overrepresentation of Gemcitabine cells relative to DMSO ones in this cluster. 95 such markers were found, with

PTTG1, *CKS2* and *PCLAF* being clear outliers, with an average $\log_2$ fold change above 1.5 (**Figure 4.46**). 14 CCRSA markers characterized the reverse selection, 4 CCRSA genes were not found in this cluster, and 4 genes were exclusively found in one or both I-BRD9 conditions.

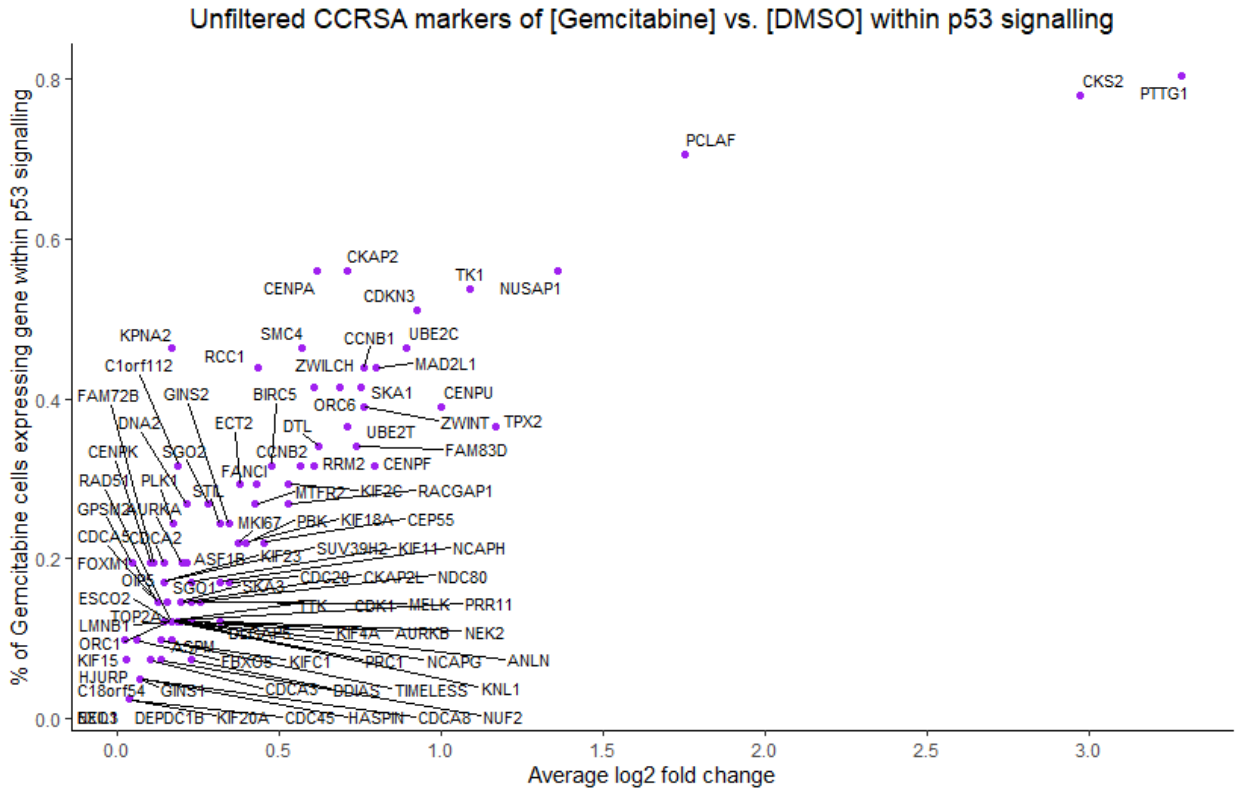


Figure 4.46. Unfiltered CCRSA markers of [Gemcitabine] vs. [DMSO] within the p53 signalling cluster.

Remarkably, the gemcitabine condition showed highly distinctive markers relative to the other three conditions even within the very small **p53 signalling** cluster, of which 563 were retained even with a statistical significance threshold of 0.05 for doubly Bonferroni-corrected p-value. The top five GO terms enriched for these 563 markers were intrinsic apoptotic signaling pathway, regulation of autophagy, translational initiation, intrinsic apoptotic signaling pathway by p53 class mediator and intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator (adjusted p-values: 1.21e-08, 4.92e-08, 4.39e-07, 7.59e-07 and 1.15e-06).

Next, the Wilcoxon pairwise activity comparisons of **ORIGINS activity scores** were performed in each cluster, in order to find treatment conditions associated with lower ORIGINS activity within

the clusters. It revealed significantly lower activity scores for I-BRD9 and gemcitabine cells relative to Gemcitabine cells (within six clusters) and DMSO (within five clusters). The I-BRD9 condition also registered significantly lower activity scores than DMSO in three clusters, and the same result was obtained relative to the Gemcitabine condition. In the **Top stemness** cluster, however, only the comparison between Gemcitabine and DMSO cells reached statistical significance, with the latter group corresponding to significantly lower activity scores. Of note, no intracluster pairwise comparison of activity scores registered an instance in which an I-BRD9 condition had significantly higher activity scores than one without I-BRD9. Overall, the strongest four effects were seen in the **Hypoxia-like** cluster (adjusted p-values: 1.12e-62, 7.19e-49, 5.44e-44 and 2.41e-36; Cohen's d scores: 0.46, 0.46, 0.59 and 0.55), all involving a non-I-BRD9 condition showing significantly higher activity scores that an I-BRD9 one. For the **Top stemness** cluster, only [Gemcitabine] vs. [DMSO] reached significance (adjusted p-value: 2.39e-02; Cohen's d score: 0.42). The complete results are illustrated in **Figure 4.47**.

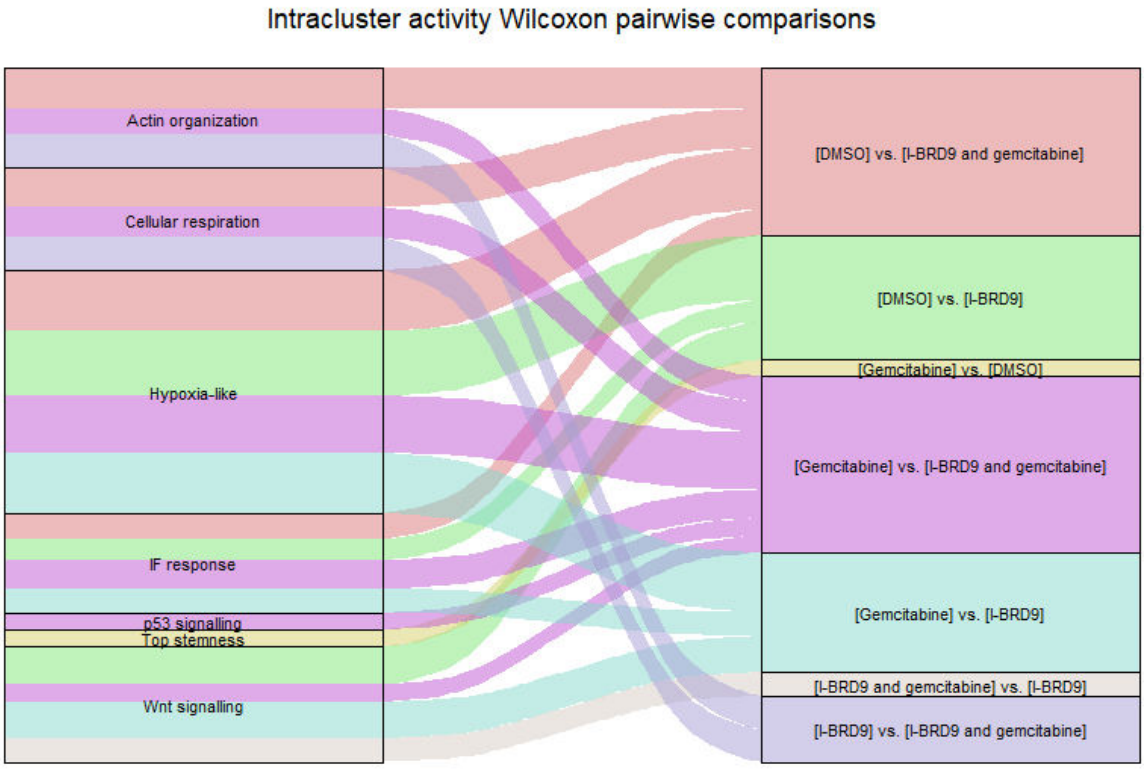


Figure 4.47. The results of Wilcoxon pairwise activity comparisons between clusters. Thicker connecting lines correspond to lower p-values.

In summary, the intracluster assessments evaluated a remarkable, nearly ~6-fold underrepresentation of the I-BRD9-treated cells in the **Top stemness** cluster relative to the DMSO control. These effects are, however, partly reversed when I-BRD9 is used in combination with gemcitabine, with I-BRD9 and gemcitabine cells showing a less marked, slightly under ~2-fold underrepresentation relative to DMSO-treated cells. Gemcitabine, meanwhile, is associated with a ~2.7 overrepresentation in the **Top stemness** cluster relative to the DMSO control. Otherwise, the TDPM-based *EZH2+NES+THY1+* signature was associated with a lower expression for the I-BRD9 condition, and in the **Hypoxia-like** cluster, the findings reached statistical significance. The assessments involving CCRSA gene sets or SPLCL genes returned no significant overlaps. However, nearly all CCRSA genes are downregulated in the **Top stemness** cluster; while statistical significance is rarely reached individually by CCRSA genes given the very small size of the cluster,

the collective trend displayed by these genes makes a clear case for a decrease in stemness due to I-BRD9 treatment in this cluster. Meanwhile, the **p53 signalling** cluster does not display decreases in stemness due to I-BRD9, but strong increases due to gemcitabine. Finally, the Wilcoxon pairwise assessment of ORIGINS activity scores with cluster revealed consistently lower activity scores for the I-BRD9 conditions, most prominently in the **Hypoxia-like** cluster. However, for the **Top stemness** cluster, only the comparison of Gemcitabine and DMSO cells returned results of significance.

The next section will evaluate the overlaps between cluster markers and condition selection markers, as well as the overlaps between the enriched GO terms associated with these markers.

4.2.11 The overlaps of markers of clusters and markers of condition selections

In this section, the effects of the treatment conditions upon the functional types of cells found in the data are uncovered by evaluating the significance of the overlaps of markers of clusters and those of condition selections. The same assessment is then performed for the enriched GO terms of these markers.

First, the overlap between the sets of genes upregulated in each cluster and the sets of genes upregulated in each of the condition selections was evaluated for statistical significance. The top 20 registered overlaps are visualized in **Figure 4.48**:

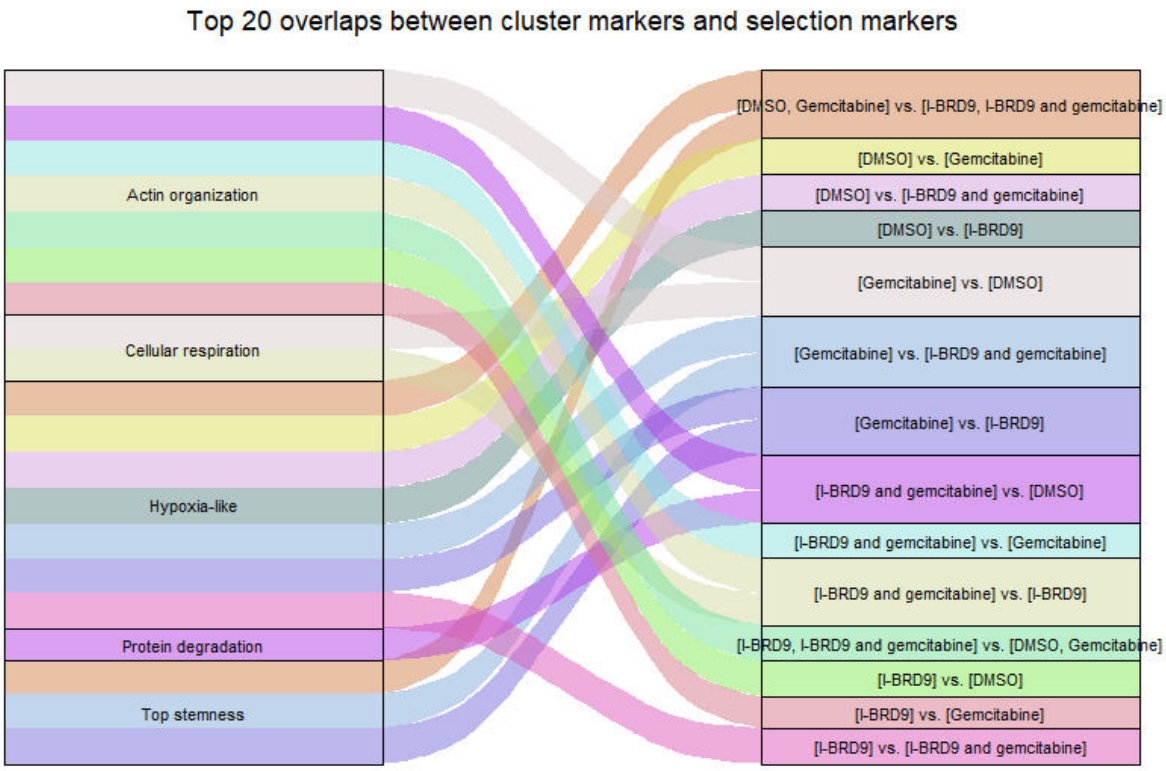


Figure 4.48. The top 20 overlaps recorded between cluster markers and selection markers.

Thus, the **Top stemness** cluster made three appearances in the list, and in each case the markers of the Gemcitabine condition (together with markers of DMSO condition, in one case) were taken relative to one or both I-BRD9 conditions, namely:

- [Gemcitabine] vs. [I-BRD9] (adjusted p-value: 4.03e-317; Jaccard score: 0.23; ranking: 11).
- [Gemcitabine] vs. [I-BRD9 and gemcitabine] (adjusted p-value: 1.5e-247; Jaccard score: 0.21; ranking: 14).
- [DMSO, Gemcitabine] vs. [I-BRD9, I-BRD9 and gemcitabine] (adjusted p-value: 5.51e-192; Jaccard score: 0.18; ranking: 17).

Significance was also reached for the [DMSO] vs. [I-BRD9] selection (adjusted p-value: 9.6e-132; Jaccard score: 0.15) and the **Top stemness** cluster. [I-BRD9] vs. [DMSO], and the other three reverses of the selections listed above, registered no significant overlaps with the markers of the same cluster.

5155 genes appeared as markers of the [DMSO] vs. [I-BRD9] selection, of which 195 had an average $\log_2$ fold-change was above 1. For these 195 genes, the enrichment of the "stem cell proliferation" GO term did reach significance (adjusted p-value: 7.78e-03), driven by five genes: *WNT5A*, *FGF2*, *VEGFC*, *RUNX1* and *HMGA2*.

Remarkably, 97 of these 195 strong markers of the [DMSO] vs. [I-BRD9] selection were also found among the 2915 markers of the **Top stemness** cluster. In turn, out of these 97 markers, 8 were among the 193 strong markers (average $\log_2$ fold-change > 1) of the **Top stemness** cluster: *LOX*, *TFPI2*, *FRMD4A*, *TAGLN*, *GLIPR1*, *GDF15*, *KCNMA1* and *HAS2*.

7 out of the 13 strongest markers of the [DMSO] vs. [I-BRD9] selection were also markers of the **Top stemness** cluster (*LINC00536*, *DCN*, *IL33*, *INHBA*, *ADAMTS6*, *IL6* and *TFPI2*), although only TFPI2 was a strong marker (average $\log_2$ fold-change > 1). The remaining 6 markers were *POSTN*, *TNFAIP6*, *COL3A1*, *COL1A2*, *PLCB1* and *PAPPA*.

Next, the overlaps between the GO terms enriched for the markers of these clusters and condition selections were analysed. The top 20 statistically significant overlaps are illustrated in **Figure 4.49**:

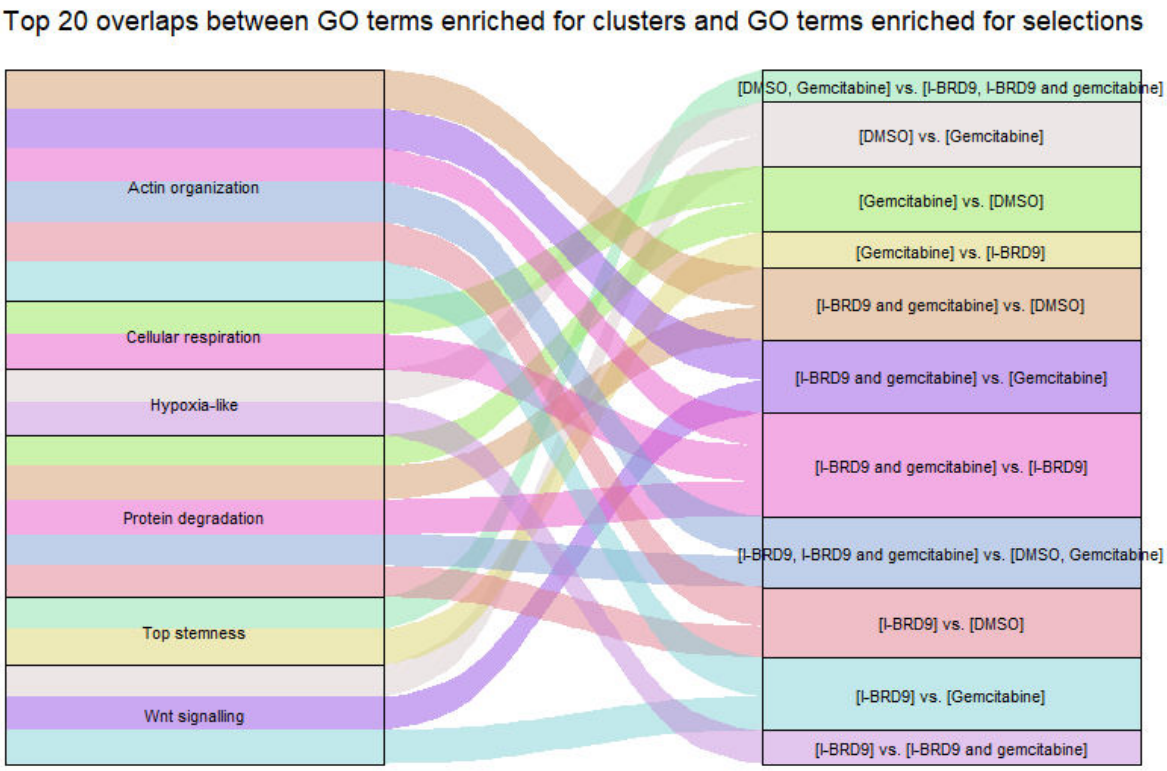


Figure 4.49. The top 20 overlaps recorded between GO terms enriched for cluster markers and selection markers.

Two appearances were obtained for the **Top stemness** cluster:

- [Gemcitabine] vs. [I-BRD9] (adjusted p-value: 3.29e-75; Jaccard score: 0.26; ranking: 6).
- [DMSO, Gemcitabine] vs. [I-BRD9, I-BRD9 and gemcitabine] (adjusted p-value: 2.43e-47; Jaccard score: 0.18; ranking: 16).

The [DMSO] vs. [I-BRD9] selection did not reach statistical significance for the **Top stemness** cluster, but the [Gemcitabine] vs. [I-BRD9 and gemcitabine] one did (adjusted p-value: 2.01e-37; Jaccard score: 0.15). All the reverse selections of the four selections indicated above reached no statistical significance for the **Top stemness** cluster.

In summary, this section has evidenced the fact that significantly overlapping markers and enriched GO terms exist between the **Top stemness** cluster and the selections taking gemcitabine-treated cells relative to I-BRD9-treated ones. The results were among the strongest overlaps identified between any pairs of clusters and condition selections. For markers, the

findings also showed a corresponding overlap with markers of the DMSO condition relative to the I-BRD9 one, indicating a distinctive loss of markers associated with **Top stemness** cluster when under I-BRD9 treatment, but this result was not paralleled when enriched GO terms rather than markers were involved in the assessment. No indication of significant overlaps in the opposite direction, that is, between **Top stemness** markers and markers of the I-BRD9 conditions surfaced. Therefore, the results do provide a measure of evidence towards a reduction in stemness-linked processes due to I-BRD9 treatment, which is consistent with the other findings reported earlier in this chapter.

4.2.12 The overlaps of markers of clusters, markers of condition selections and stemness-linked gene sets

This section assesses the impact of the stemness-linked gene sets upon the overlaps between the markers of clusters and the markers of condition selections, in order to evaluate whether the effects seen in stemness-linked clusters can be attributed to a downregulation of stemness-linked genes.

Two significant overlaps were registered for the pancreas and the glioma CCRSA gene sets each, with the **Top stemness** cluster, and the [Gemcitabine] vs. [I-BRD9 and gemcitabine] and [Gemcitabine] vs. [I-BRD9] selections. For both CCRSA gene sets, both selections showed the same adjusted p-value: 1.53e-04 and 3.31e-02, respectively.

A larger number of overlaps were recorded for the prognosis and the union CCRSA gene sets, involving two clusters **Top stemness**, and **p53 signalling** cluster in both cases, with the [Gemcitabine] vs. [I-BRD9] and [Gemcitabine] vs. [I-BRD9 and gemcitabine] corresponding to the most significant overlaps, as follows:

- Prognosis CCRSA gene set:
    - With **Top stemness** and [Gemcitabine] vs. [I-BRD9] (adjusted p-value: 1.87e-11)

- With **Top stemness** and [Gemcitabine] vs. [I-BRD9 and gemcitabine] (adjusted p-value: 6.44e-11).

- Union CCRSA gene set:
  - With **Top stemness** and [Gemcitabine] vs. [I-BRD9] (adjusted p-value: 1.55e-14)
  - With **Top stemness** and [Gemcitabine] vs. [I-BRD9 and gemcitabine] (adjusted p-value: 3.5e-14).

All three-way overlaps obtained for the prognosis and union CCRSA gene sets are illustrated in **Figure 4.50**:

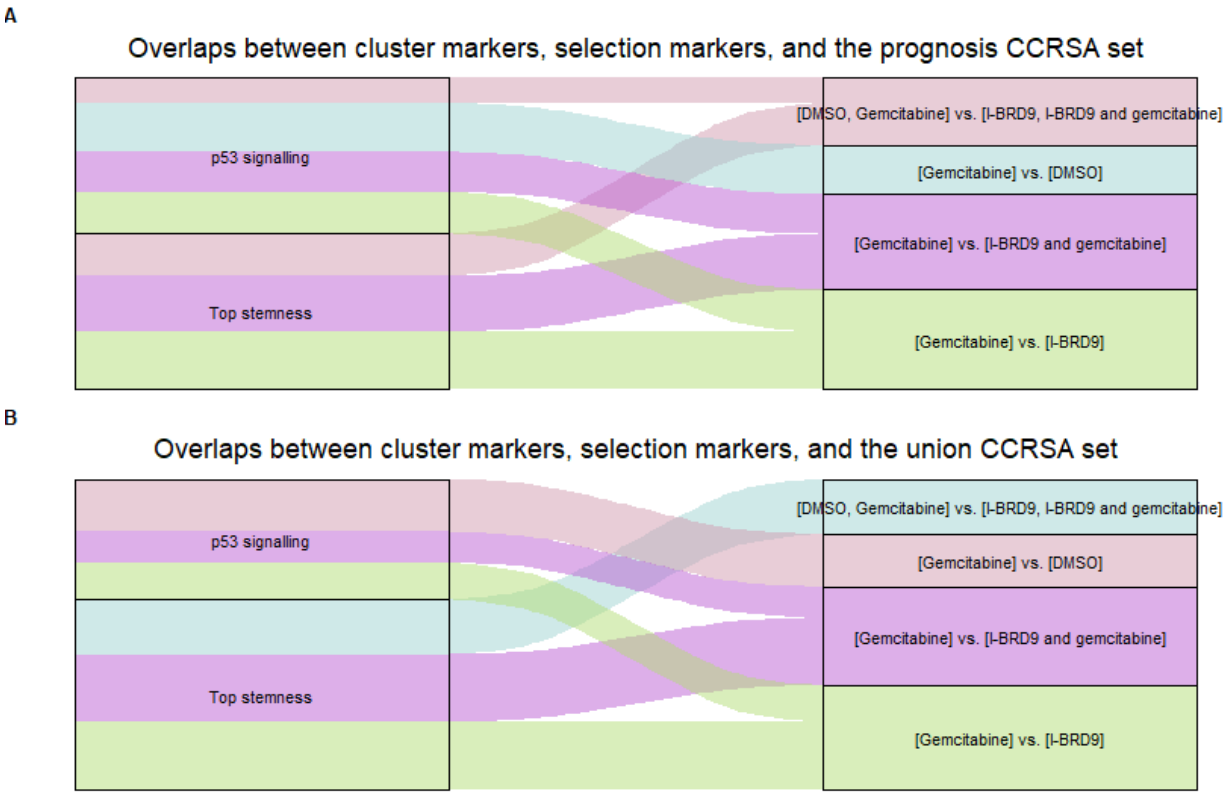


Figure 4.50. Significant three-way overlaps between cluster markers, selection markers, and the prognosis CCRSA gene (**A**) and the union CCRSA gene set (**B**). Thicker lines connecting cluster and condition selections correspond to lower p-values.

No significant three-way overlaps were registered for the other CCRSA gene sets or for the SPLCL genes.

In conclusion, this section has established the significant direct involvements of the CCRSA genes in some of the overlaps between markers of the **Top stemness** clusters and the markers overexpressed in non-I-BRD9 conditions relative to I-BRD9 conditions noted earlier.

The next section will provide a discussion of the findings obtained in this chapter.

4.3 Discussion

Five out of the six lines of evidence employed in this analysis converged towards the identification of as small **Top stemness** population in this dataset, also overexpressing *EZH2*, a histone methyltransferase reported as a putative CSC marker in PDAC and proposed as a druggable target for a subgroup of medulloblastoma CSC[243] and in uveal melanoma[244].

Remarkably, EZH2 has been identified as a regulator of *CCNB2*, *MKI67*, *KIF20A*, *CCNA2*, *CDC20*, *PCLAF* and *TOP2A*[245], all of them being both CCRSA genes and some of the strongest markers of **Top stemness** (adjusted p-values: 0), rendering it a promising epigenetic target in PDAC.

The **Top stemness** population displays gemcitabine resistance, with Gemcitabine cells showcasing an overrepresentation of a factor of ~3.2 in the **Top stemness** cluster. Of note, many of the strongest markers of the **Top stemness** cluster have been recently associated in the literature with chemoresistance in cancer, including *CLSPN*[246], *CDCA8*[247], *TOP2A*[248], *RRM2*[249], *TYMS*[250], *UBE2C*[251], *CDK1*[252], further highlighting the importance of these genes and their associated mechanisms in frustrating therapeutic strategies. Other very strong markers of **Top stemness**, such as *CDC20*[253] and *NEK2*[254], have been noted to promote the progression of hepatocellular carcinoma through the regulation of EMT, and the same has been found for *SPC25* in oral squamous cell carcinoma[255].

With regards to the effects of I-BRD9, a ~4.9-fold underrepresentation of I-BRD9 cells in the **Top stemness** cluster relative to their representation among all cells was detected, contrasting with the lack of a comparable efficacy of the I-BRD9 treatment reported in the previous chapter, likely

driven by the difference in the treatment duration (24 h for the A13A scRNA-seq experiment, 72 h for the patient one), and potentially also by biological differences between the two samples of PDAC cells.

However, combining I-BRD9 with gemcitabine substantially, but not entirely, reversed the effects of I-BRD9 with regards to reducing the prevalence of **Top stemness** cells. For the I-BRD9 and gemcitabine cells, only a ~1.6-fold underrepresentation in the **Top stemness** cluster was detected. Thus, I-BRD9 and gemcitabine do not synergize towards eliminating CSCs. Rather, the effect registered is simply a combination of the two effects seen for the two compounds: towards reducing the prevalence of CSC for I-BRD9, towards increasing it for gemcitabine.

In addition to reducing the prevalence of **Top stemness** cells, I-BRD9 was also associated with a reduction in the expression of CCRSA genes, without generally reaching statistical significance at the level of individual CCRSA genes but evidencing a clear collective trend.

The **p53 signalling** cluster evidenced lower stemness but salient chemoresistance, with I-BRD9 showing no associations with reduced numbers of **p53 signalling** cells, whereas Gemcitabine was associated with a more than 3-fold increase of **p53 signalling** cells relative to the DMSO control, as well as with the presence of 23 CCRSA genes not otherwise found in cells from the other three experimental conditions within this cluster. Autophagy was associated with the gemcitabine cells from this cluster, a phenomenon previously note to arise as a side effect of cancer therapy and to assist cancer cell survival[256]. Furthermore, autophagy induces pluripotency, a mechanism upon which CSCs depend[256], raising the possibility that **p53 signalling** retains sufficient stemness to be capable of regenerating the entire tumour, including the **Top stemness** population, in the event that the original **Top stemness** population is annihilated by chemotherapy, with the RNA velocity analysis suggesting that such **p53 signalling** cells can indeed dedifferentiate to **Top stemness** cells. Strong markers of the **p53 signalling** cluster with noted recent associations with chemoresistance in cancer include *GDF15*[257], *GADD45A*[258], *MDM2*[259] and *CDKN1A*[260].

The effects of BRD9 upon cancer stemness likely involve G2/M DNA replication checkpoint-linked genes as key targets, such as *TOP2A* and *CDK1*, whose downregulation after BRD9 inhibition was recently demonstrated[261], and also *UBE2C*, *CCNB1*, *PBK* and *CENPF*, all genes downregulated in the I-BRD9 condition relative to the DMSO control in this dataset. Remarkably, among six pathways related to cell proliferation, the G2/M checkpoint pathway alone was found to predict survival in PDAC[262].

Histone deacetylases have been conjectured as mediators of the TOP2A downregulation after the inhibition of BRD9[261], an idea that exploits the fact that HDAC1 and HDAC2 interact directly with *TOP2A*[263]. Both *HDAC1* and *HDAC2*, as well as *HDAC2-AS2*, *HDAC4* and *HDAC8*, were found to be downregulated in the I-BRD9 condition relative to the DMSO one (adjusted p-values: 5.62e-07, 2.08e-51, 2.7e-15, 2.36e-15 and 5.03e-118). In addition, the effects of I-BRD9 upon *TOP2A* could be driven by its downregulation of *EZH2* (adjusted p-value: 3.97e-19). Of note, BRD9 was previously found highly enriched at some of the enhancer regions of EZH2[264]. BRD9 has also been noted to facilitate the interaction of *RAD54* and *RAD51*, essential for homologous recombination-mediated DNA repair[265]. The “DNA recombination”, “recombinational repair” and “double-strand break repair via homologous recombination” GO terms were strongly enriched in the **Top stemness** cluster (adjusted p-values: 3.41e-28, 6.98e-24, 1.45e-23).

The next chapter will provide an aggregate analysis of the subpopulations discovered in the two scRNA-seq datasets and identify shared genes and processes governing stemness in the two datasets, and perform a comparative analysis of the changes in gene expression due to I-BRD9 treatment between the two datasets.

CHAPTER 5

A COMPARATIVE ASSESSMENT OF THE SINGLE-CELL RNA-SEQUENCING RESULTS

5 A comparative assessment of the single-cell RNA-sequencing results

5.1 Introduction

This chapter provides a comparative analysis of the findings emerging from the two single-cell RNA-sequencing experiments discussed previously.

Section 5.2.1 discusses overlaps between cluster markers from the two datasets, and between their associated enriched GO terms. A look at the stemness markers conserved between the two datasets is provided in **Section 5.2.2**. Epigenetic associations characterizing conserved stemness markers are identified in **Section 5.2.3**, and the overlap of condition selection markers and GO terms between the two datasets are discussed in **Section 5.2.4**. Finally, **Section 5.3** comprises a discussion.

5.2 Results

5.2.1 Overlaps of markers and GO terms enriched for clusters from the two scRNA-seq datasets

A comparison of the overlap of cluster markers between the two datasets is displayed in **Figure 5.1**. Generally, clusters with a similar or identical functional annotation in both datasets recorded overlaps reaching statistical significance.

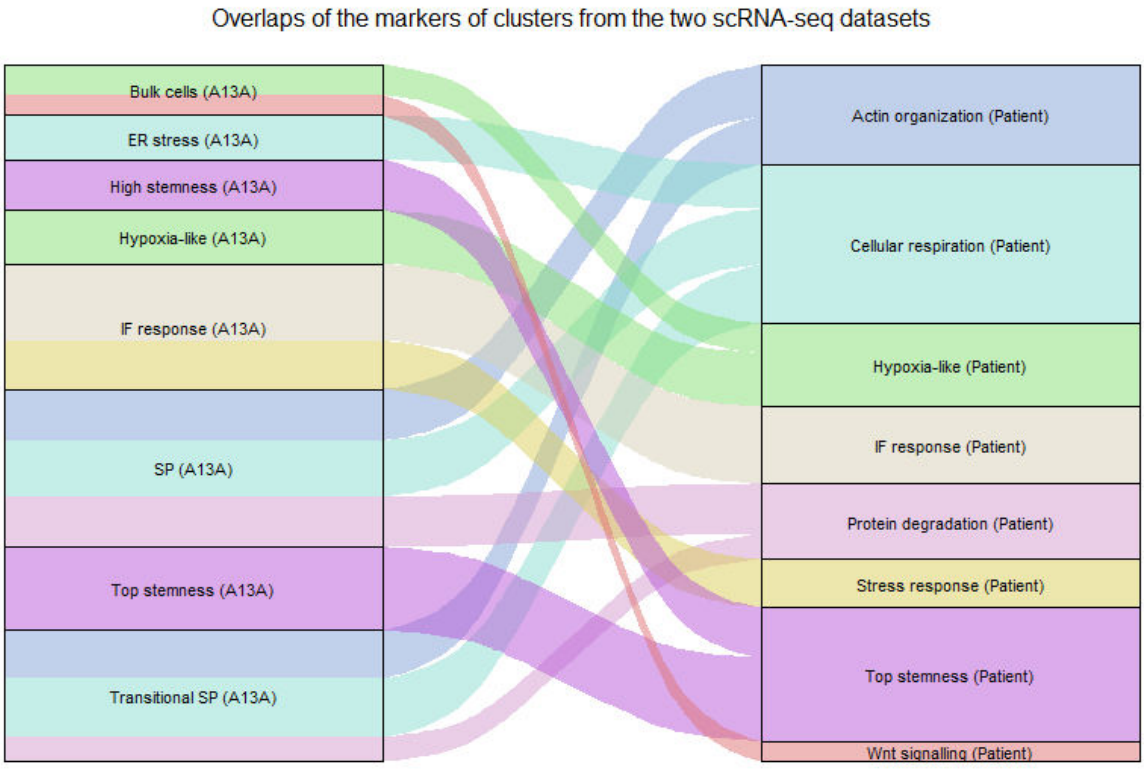


Figure 5.1. The significant overlaps of cluster markers between the two datasets. Thicker connecting lines correspond to lower p-values.

The top overlap was recorded between the **Top stemness** clusters in both datasets (adjusted p-value: 6.12e-50), followed by the one between the **IF response** clusters in both datasets (adjusted p-value: 4.09e-35). The **Hypoxia-like** clusters also reached a significant marker overlap (adjusted p-value: 9.84e-10, 6th overall).

Of note, the **SP** and **Transitional SP** clusters from the A13A dataset significantly overlapped with the same three clusters from the patient dataset (**Actin organization**, **Cellular respiration**, **Protein degradation**).

While the **Top stemness** cluster from the A13A dataset did not record any significant marker overlaps with any cluster except its functionally analogous cluster from the patient dataset, the latter also significantly overlapped with the **High stemness** cluster from the A13A dataset (adjusted p-value: 9.8e-10, ranking: 6th).

Next, overlaps of GO terms enriched for the markers of clusters from the two datasets were assessed for statistical significance. The results are illustrated in **Figure 5.2**:

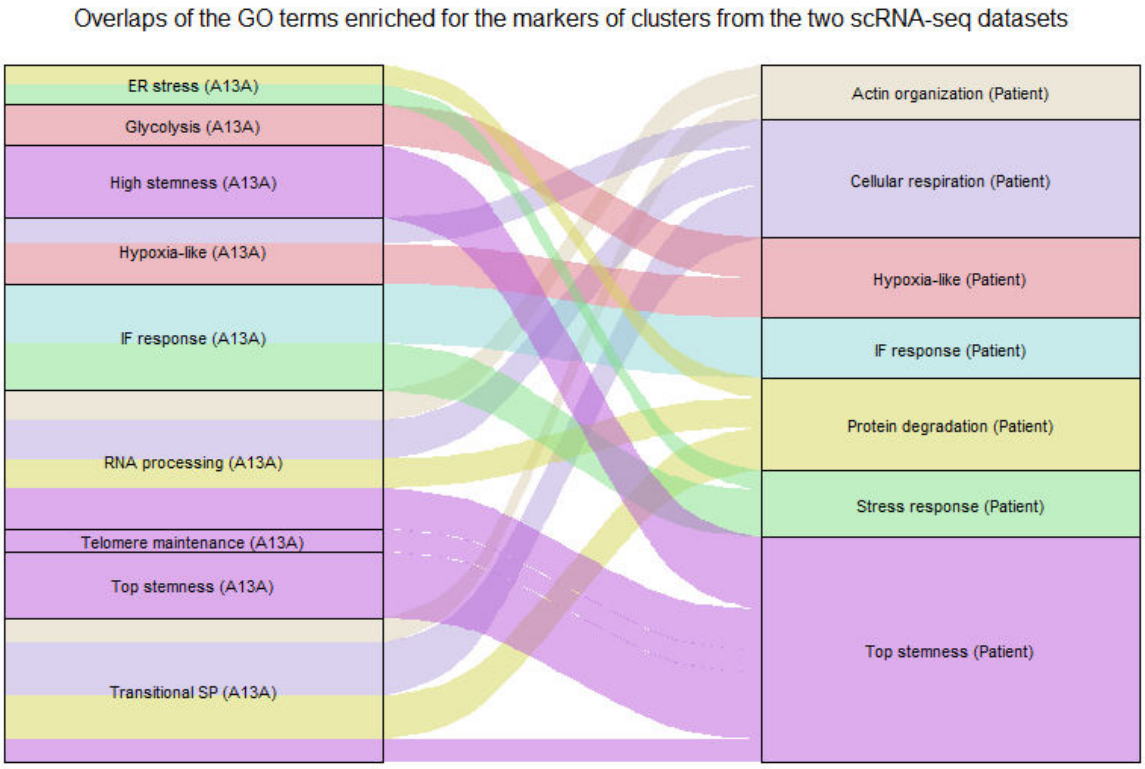


Figure 5.2. The significant overlaps of enriched GO terms for cluster markers between the two datasets. Thicker connecting lines correspond to lower p-values.

In terms of enriched GO terms, the strongest overlap again involved the **Top stemness** cluster from the patient dataset, but with the **High stemness** cluster from the A13A dataset (adjusted p-value: 1.2e-183), while the overlap with the **Top stemness** cluster from the A13A dataset was the second strongest (adjusted p-value: 1.2e-183). The top three GO terms significantly enriched in both the **Top stemness** cluster from the Patient dataset and the **High stemness** cluster from the A13A dataset were: chromosome segregation, nuclear division and organelle fission (adjusted p-values for the patient **Top stemness** cluster: 1.37e-55, 5.67e-44 and 2.37e-42; rankings for the patient **Top stemness** cluster: 1st, 2nd and 3rd; adjusted p-values for the patient **Top stemness**

cluster: 1.37e-55, 5.67e-44 and 2.37e-42; adjusted p-values for the A13A **High stemness** cluster: 2.77e-45, 6.91e-41 and 7.1e-40; rankings for the A13A **High stemness** cluster: 3rd, 4th and 5th).

The **IF response** and **Hypoxia-like** clusters from both datasets again registered significant overlaps (adjusted p-values: 1.12e-64 and 2.44e-12; rankings: 3rd and 9th).

As in the case of the markers, the **Top stemness** cluster from A13A dataset overlapped only with the **Top stemness** cluster from the patient dataset. However, the asymmetry was more pronounced for the overlaps of the latter cluster, which overlapped additionally not only with the **High stemness** cluster as indicated in the previous paragraph, but also with the **RNA processing**, **Transitional SP** and **Telomere maintenance** clusters (adjusted p-values: 1.43e-13, 8.44e-04 and 1.35e-03; rankings: 8th, 15th and 16th).

The absence of overlaps with the **SP** cluster in terms of the enriched GO terms is consistent with the fact that no SPLCL genes-related clusters were identified in the patient dataset. Of note, the marker overlaps with the **SP** cluster mentioned earlier in this chapter included 4 SPLCL genes for **Actin organization** (*TXNRD1*, *TMEM156*, *GPAT3* and *LIPH*), 5 SPLCL genes for **Protein degradation** (*AKR1B10*, *TXNRD1*, *AKR1C3*, *GPAT3* and *LIPH*), and only 1 SPLCL gene for **Cellular respiration** (*AKR1C3*).

Finally, we note that the **p53 signalling** cluster from the patient dataset was the only cluster not to record significant overlaps with any other clusters, in terms of either markers or their associated enriched GO terms.

The next section will identify the markers of stemness clusters occurring in both datasets.

5.2.2 Markers of stemness clusters shared by both datasets

The **Top stemness** cluster from the A13A dataset had markedly fewer markers (336) than the **Top stemness** cluster from the patient dataset, with 208 of them being also found among the 2915 markers of the latter cluster.

The top five enriched GO terms corresponding to these 208 shared markers were: mitotic nuclear division (adjusted p-value: 7.47e-63), chromosome segregation (adjusted p-value: 1.35e-62), nuclear division (adjusted p-value: 2.27e-62), organelle fission (adjusted p-value: 1.11e-60, and nuclear chromosome segregation (adjusted p-value: 7.07e-52).

The 128 markers of the **Top stemness** cluster from the A13A dataset but not of the functionally related cluster from the patient dataset showcased only four enriched GO terms: mRNA processing, positive regulation of protein localization to plasma membrane, positive regulation of protein localization to cell periphery and regulation of cell cycle phase transition, all with an adjusted p-value of 0.04.

Meanwhile, the 2707 markers of the opposite set difference showed enrichment for GO terms related to DNA replication and repair, having the following top five enriched GO terms: DNA replication (adjusted p-value: 2.38e-55), DNA-templated DNA replication (adjusted p-value 9.21e-46), double-strand break repair (adjusted p-value: 3.53e-30), DNA recombination (adjusted p-value: 6.27e-26) and regulation of DNA metabolic process (adjusted p-value: 2.66e-25). The significant enrichment of GO terms related to DNA repair and replication for the patient **Top stemness** raises the question of possible BRCAness in the patient sample – a feature of homologous recombination deficiency resembling a BRCA1 mutation, noted to correlate with enhanced DNA repair and higher expression of MKI67 in breast cancer, regardless of whether the BRCAness phenotype was due to BRCA mutations or not[266]. This increase in DNA repair is due to

the fact that multiple compensatory mechanisms, such as RAD51 overexpression, p53 mutation and 53BP1 loss, can emerge in cancer in response to a BRCA1 mutation[267].

The **High stemness** cluster from the A13A dataset had 427 markers, of which 336 were also found among the markers of the **Top stemness** cluster from the patient dataset.

The shared markers of the **Top stemness** clusters in both datasets are illustrated in **Figure 5.3**. Visual outliers on the plot are *DIAPH3*, *H2AZ1*, *STMN1*, *MKI67*, *TUBA1B*, *PTTG1*, *BIRC5*, *TPX2*, *HMGB2*, *AURKA*, *TOP2A* and *CENPF*.

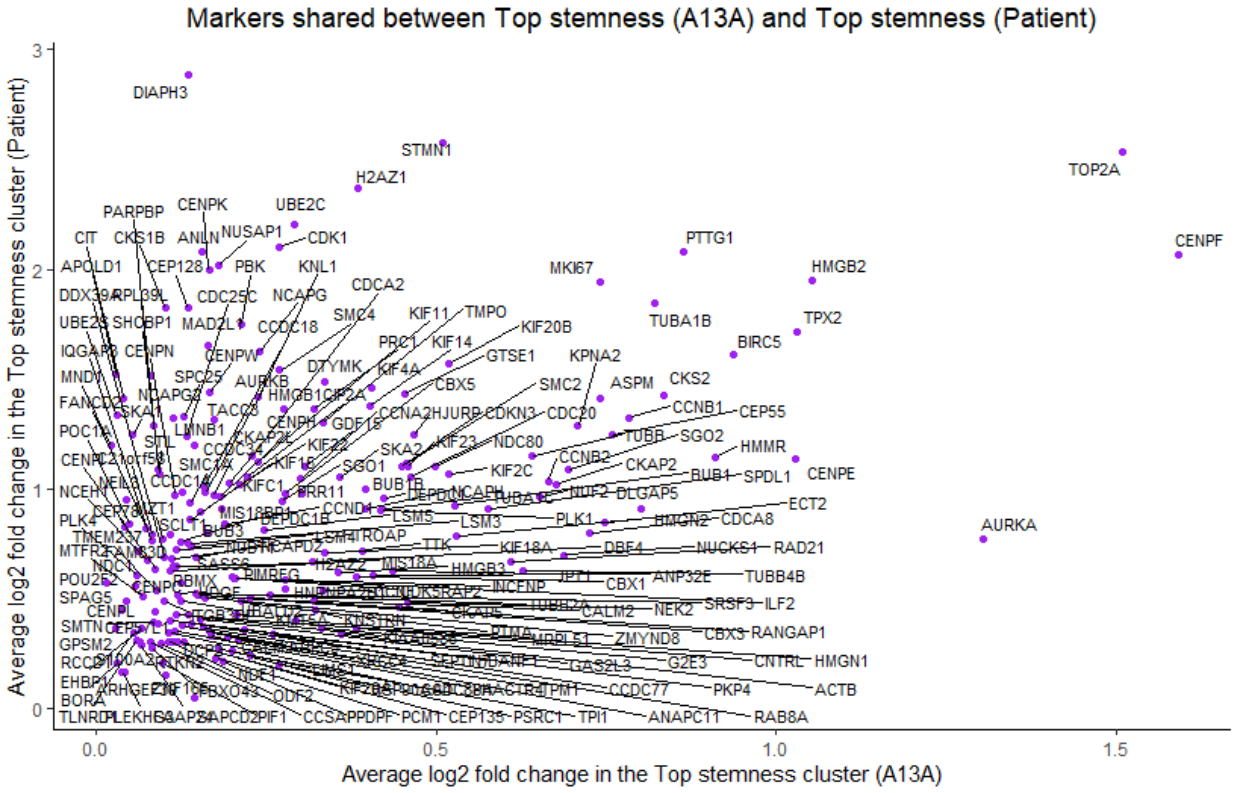


Figure 5.3. Markers shared by the Top stemness clusters in both datasets.

The shared markers of the **High stemness** cluster in the A13A dataset and of the **Top stemness** cluster in the patient dataset are illustrated in **Figure 5.4**. Visual outliers on the plot are *RRM2*, *DIAPH3*, *TOP2A*, *STMN1*, *H2AZ1*, *TUBA1B*, *HMGB2* and *H1-4*.

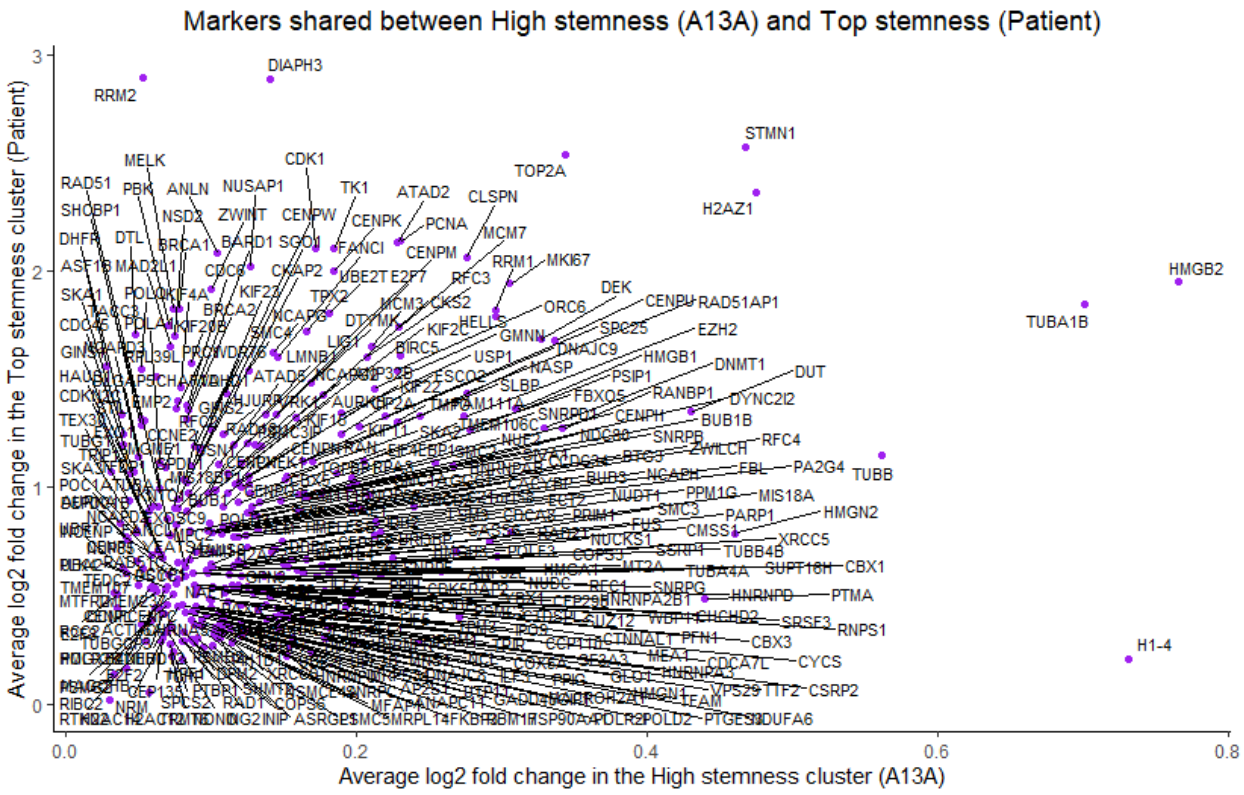


Figure 5.4. Markers shared by the High stemness cluster in the A13A dataset and the Top stemness cluster in the patient dataset.

Finally, markers shared by the **High stemness** cluster from the A13A dataset and by both **Top stemness** clusters are displayed in **Figure 5.5**, with the average $\log_2$ fold change in the Top stemness clusters on the coordinate axes. In total, 103 such markers were identified.

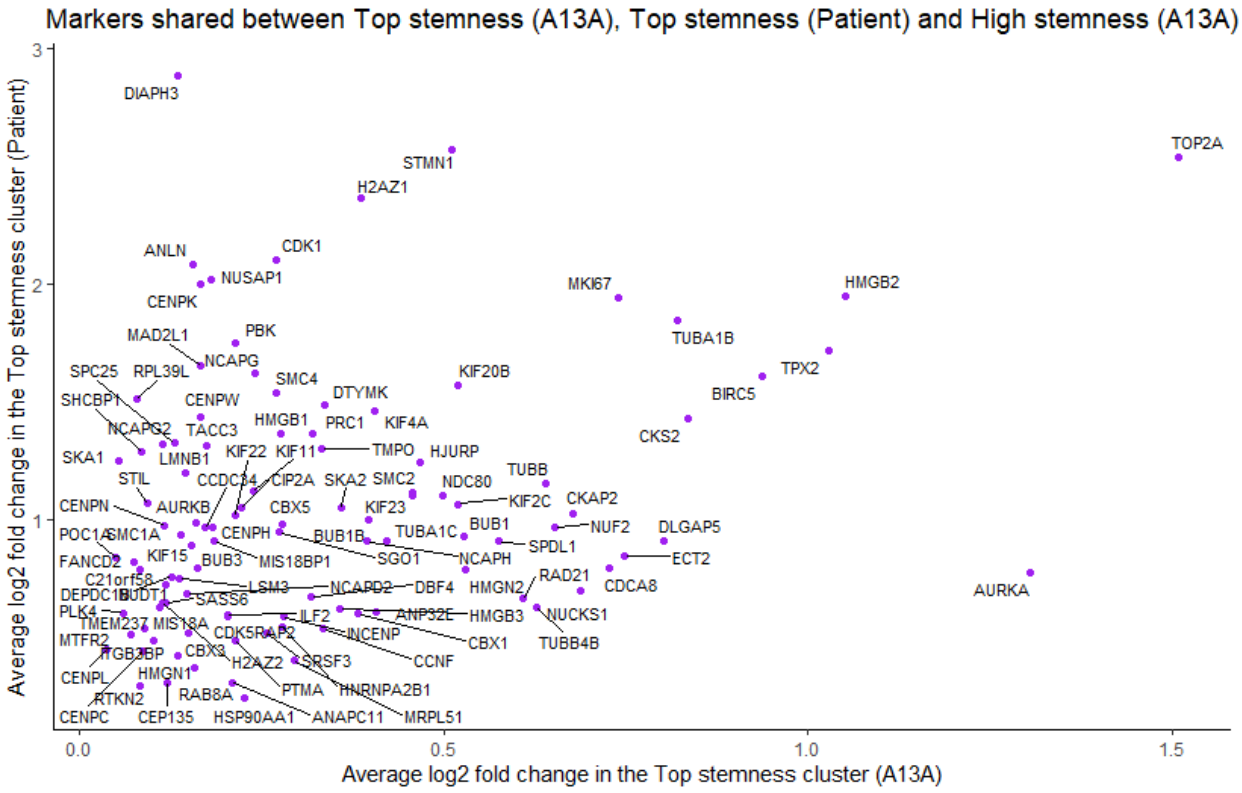


Figure 5.5. Markers shared by the High stemness cluster in the A13A dataset and the Top stemness clusters in both datasets.

In conclusion, 103 genes distinctively highly expressed in both **Top stemness** clusters and in the **High stemness** cluster are found, with CCRSA genes *TOP2A* and non-CCRSA genes *HMGB2*, *TUBA1B*, *STMN1* and *H2AZ1* showing a particularly strong overexpression in all three clusters. The next section will identify the epigenetic mechanisms with a shared activation in the stemness clusters.

5.2.3 The epigenetic mechanisms characterizing the shared markers of stemness-linked clusters

To discover novel regulators of stemness linked with genetics mechanisms, the CCRSA genes were subtracted from the 103 markers shared by the **Top stemness** clusters and the A13A **High stemness** clusters identified in the previous section, and then enrichment analysis was performed on the remaining 63 genes. 13 of the 63 genes were found to be involved in epigenetic

mechanisms, namely *HMGB2*, *ANP32E*, *CENPW*, *HMGB1*, *CENPH*, *MIS18A*, *CENPN*, *CENPC*, *ITGB3BP*, *CBX3*, *HMGB3* and *HNRNPA2B1*. The top five GO terms related to epigenetics and enriched for the 63 genes were: protein-DNA complex subunit organization (adjusted p-value: 3.01e-06), protein-DNA complex assembly (adjusted p-value: 9.74e-06), histone exchange (adjusted p-value: 2.37e-05), chromatin remodeling at centromere (adjusted p-value: 1.48e-04), and nucleosome organization (adjusted p-value: 2.38e-03). All enriched GO terms for the set of 63 non-CCRSA shared markers of the aforementioned three clusters are visualized in **Figure 5.6**.

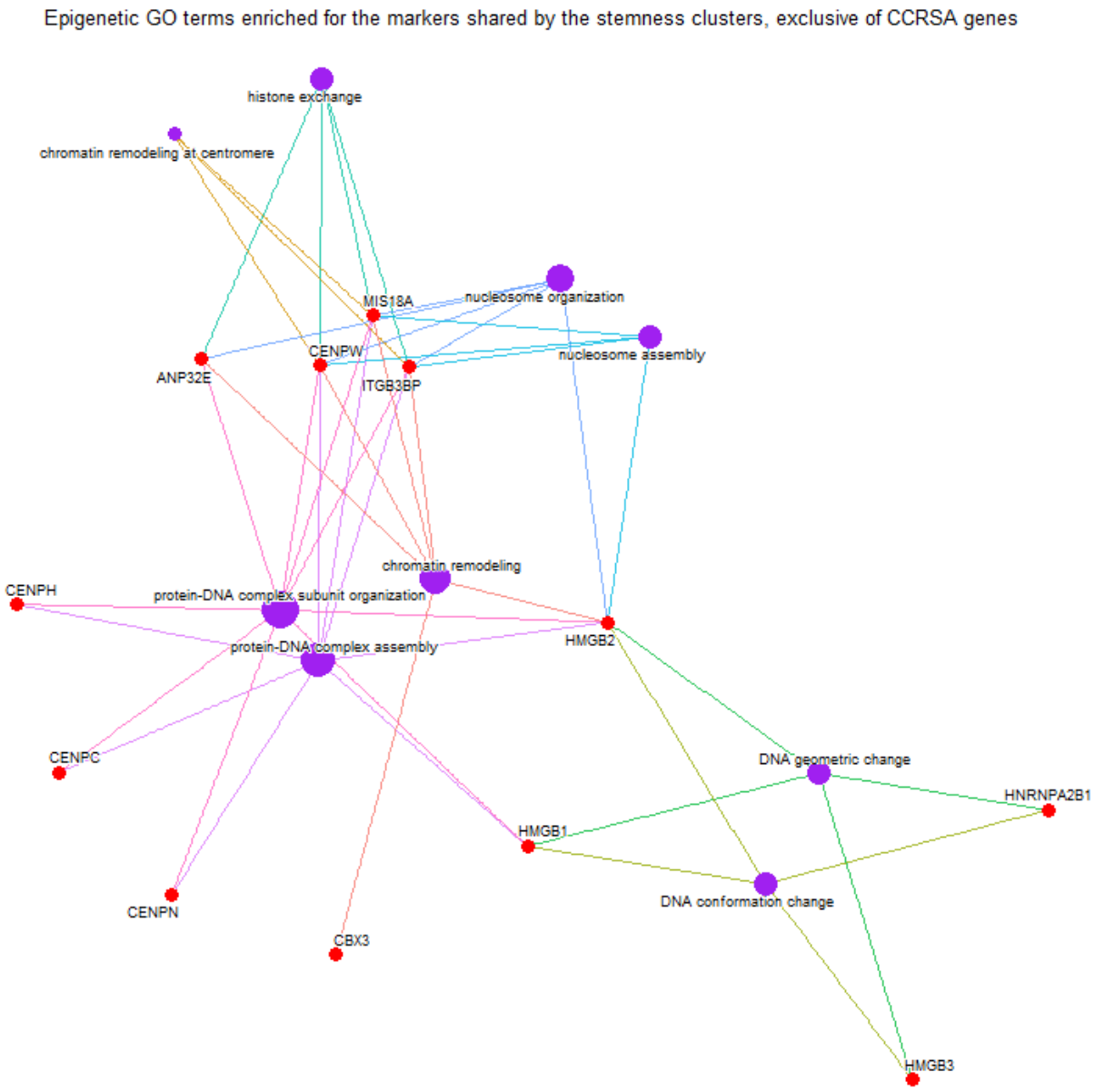


Figure 5.6. GO terms enriched for the 63 non-CCRSA shared markers of the Top stemness clusters, and of the High stemness cluster from the A13A dataset.

Finally, an identification of the genes who appeared exclusively as markers of both **Top stemness** clusters, but not of any other cluster in either dataset, was performed. 44 such genes were found, of which 16 were CCRSA genes (*CCNB2*, *ASPM*, *PLK1*, *CDC20*, *KIF14*, *TTK*, *KIF18A*, *CCNA2*, *KIF20A*, *CKAP2L*, *KNL1*, *KIFC1*, *GPSM2*, *CENPI*, *SPAG5* and *NEIL3*), while the remaining 28 genes were: *HMMR*, *DEPDC1*, *PSRC1*, *KNSTRN*, *G2E3*, *CNTRL*, *PIMREG*, *PIF1*, *FBXO43*, *SAPCD2*, *CKS1B*,

XRCC4, *CEP128*, *PARPBP*, *BORA*, *CCDC77*, *NDE1*, *CCDC18*, *SMTN*, *ODF2*, *FAAP24*, *NDC1*, *CIT*, *PLEKHG3*, *RCCD1*, *MND1*, *CEP78* and *POU2F2*. However, enrichment analysis identified no significantly enriched epigenetic terms for these 28 non-CCRSA genes, or for the 28 non-CCRSA genes taken together with the 16 CCRSA genes.

In conclusion, 13 genes related to epigenetic GO terms are identified as markers of the clusters linked to stemness in both datasets to a statistically significant extent, namely the two **Top stemness** clusters and the A13A **High stemness** cluster. However, an analysis of the specific markers shared by both **Top stemness** clusters, but by no other cluster in either dataset, revealed no connection with epigenetic mechanisms.

The next section will provide a comparative look of the effects of the experimental conditions in both datasets.

5.2.4 Overlaps of markers and GO terms enriched for condition selections from the two scRNA-seq datasets

The overlaps of condition selection markers between the two datasets were both less pronounced and less interrelated (e.g I-BRD9 or non-I-BRD9 conditions from both datasets, respectively, did not show distinctively strong overlaps), suggesting that indeed the shorter duration of treatment might explain the lack of noted effects upon stemness seen in the A13A dataset to an important extent. 43 such overlaps were recorded, of which the strongest 20 are displayed in **Figure 5.7**. The top overlap was registered for the [Activin A and I-BRD9] vs. [Activin A] selection from the A13A dataset and the [Gemcitabine] vs. [DMSO] one from the patient dataset (adjusted p-value: 1.33e-18).

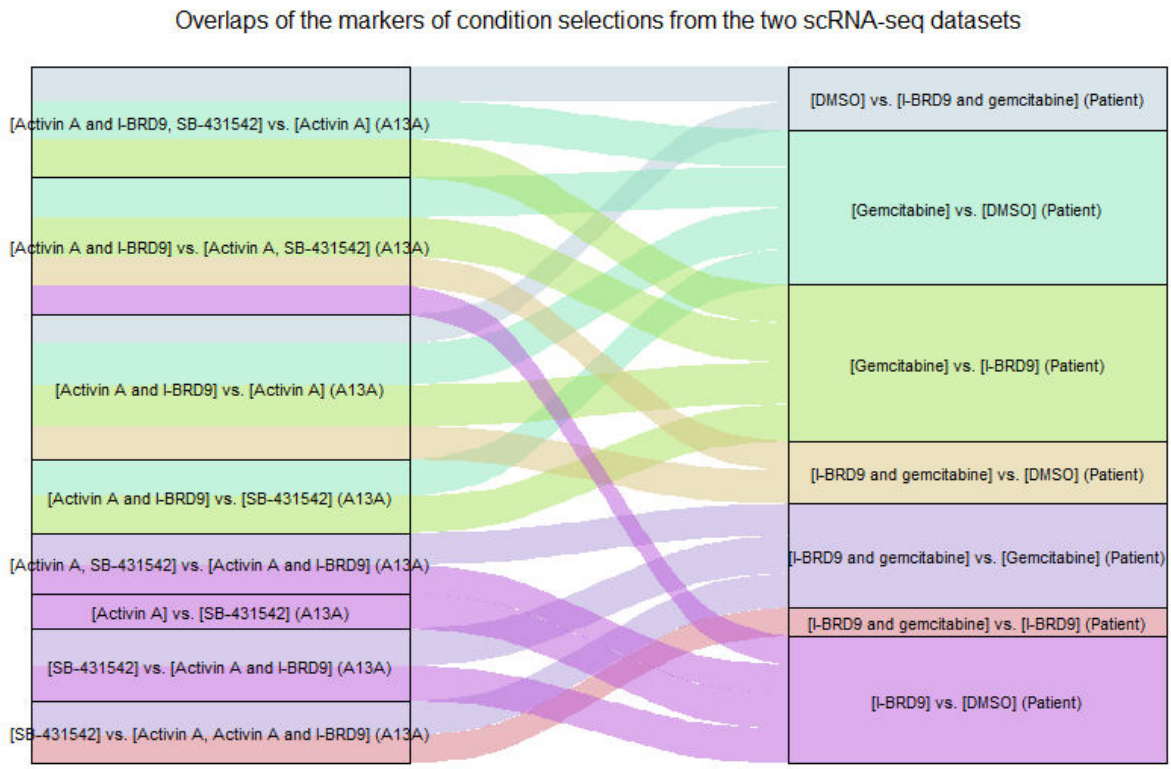


Figure 5.7. The top 20 overlaps of condition selection markers between the two datasets. Thicker connecting lines correspond to lower p-values.

With regards to the overlaps of GO terms enriched for the markers of condition selections between the two datasets, 69 significant overlaps were recorded, of which the strongest 20 are displayed in **Figure 5.8**. The top overlap was recorded for the [Activin A, SB-431542] vs. [Activin A and I-BRD9] condition from the A13A dataset and the [I-BRD9 and gemcitabine] vs. [I-BRD9] condition from the patient dataset (adjusted p-value: 8.47e-115).

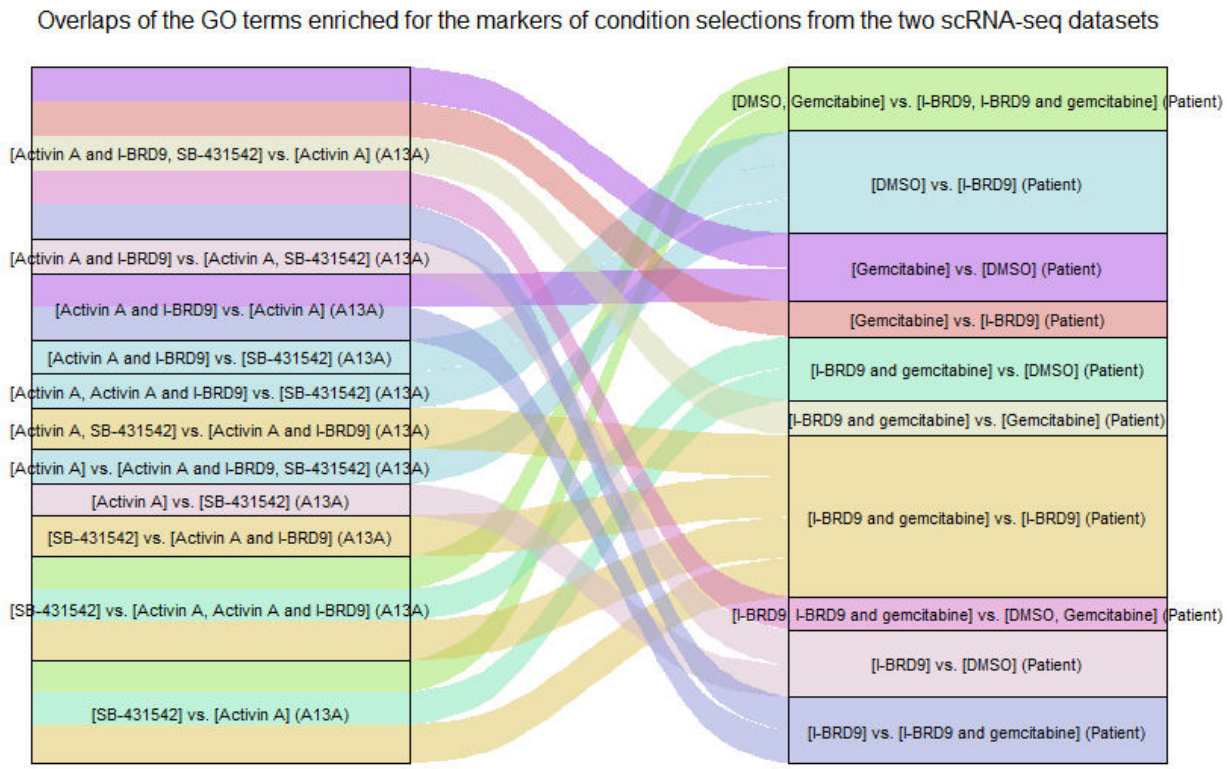


Figure 5.8. The top 20 overlaps of GO terms enriched for the condition selection markers between the two datasets. Thicker connecting lines correspond to lower p-values.

Of note, among the 51 CCRSA markers of the [DMSO] vs. [I-BRD9] selection in the patient dataset, a sizeable number matching with the reduction in stemness seen in this dataset, only 4 were shared with the [Activin A] vs. [Activin A and I-BRD9] selection from the A13A dataset: *CHEK1*, *ECT2*, *AURKA* and *NDC80*. An additional 9 CCRSA genes surfaced as markers of the [Activin A] vs. [Activin A and I-BRD9] selection, but not of the [DMSO] vs. [I-BRD9] selection: *NCAPH*, *GINS2*, *DKC1*, *CDCA8*, *FBXO5*, *MCM10*, *ORC6*, *RFC4* and *CENPU*.

In conclusion, there are important differences between the sets of markers downregulated after I-BRD9 treatments in the two datasets, suggestive of the fact that the duration of treatment in the A13A experiment was not optimal. In addition, the fact that despite a generally much lower downregulation of CCRSA genes after I-BRD9 treatment in the patient dataset rather than in the A13A one, nine CCRSA genes were found downregulated after I-BRD9 treatment only in the A13A

dataset, suggests that biological differences between the A13A and the patient sample also account for the observed discrepancy.

5.3 Discussion

Clusters with a similar functional annotation in both datasets overlapped substantially, in terms of both markers and their associated enriched GO terms. Notably, while the **Top stemness** cluster from the A13A dataset overlapped only with the **Top stemness** cluster from the patient dataset with regards to both markers and GO terms, the latter cluster overlapped with four additional A13A clusters in terms of one or both compared items: **High stemness**, **RNA processing**, **Transitional SP** and **Telomere maintenance**. Furthermore, the **Top stemness** cluster from the patient dataset overlapped stronger with the **High stemness** cluster from the A13A dataset rather than with the **Top stemness** one. All in all, this suggests that the A13A cells displayed stronger stemness characteristics, with the lack of an effect seen for I-BRD9 in this dataset being, most probably a combination of the shorter treatment duration and higher stemness characterizing the A13A dataset. Stronger overall drug resistance, besides bona fide stemness, as evidenced by the **SP** population and the **Transitional SP** one in the A13A dataset, contrasting with the **Protein degradation** population in the patient dataset, which shows only a sub-threshold overlap with SPLCL markers, is probably a third aspect that contributes to the discrepancy.

Epigenetic mechanisms are enriched for shared markers of **High stemness** and of the two **Top stemness** clusters, and 13 epigenetic regulators – distinct from the CCRSA genes – are found overexpressed in all these three clusters. Several of these genes have been previously associated with pluripotency or cancer stemness in the literature e.g. *HMGB2* was noted as a regulator of pluripotency through interactions with *OCT4*[268], *ANP32E* was noted to promote the proliferation and self-renewal of pancreatic cancer stem cells[269], *CENPW* was recently demonstrated to induce the proliferation of CSC-like phenotypes in kidney renal clear cell carcinoma[270], and *HMGB1*,

induced by hypoxia, was shown to drive the self-renewal and tumorigenicity of glioma cancer stem cells[271].

For the 28 marker genes overexpressed exclusively in the **Top stemness** clusters, no significant epigenetic associations were identified, however. Notably, among these genes, *POU2F2* was found to drive the formation of CSCs in hepatocellular carcinoma[272].

CHAPTER 6

SUMMARY AND CONCLUSIONS

6 Summary and conclusions

6.1 Summary of research aims

Pancreatic ductal adenocarcinoma (PDAC) is the most lethal of all common cancers[10], being the fourth leading cause of all cancer-related deaths[13] and projected to climb to the second position by 2030[14]. It remains one of the greatest outstanding challenges in oncology[11].

No curative treatment for PDAC currently exists, except surgery. However, 80% of newly diagnosed PDAC cases are not amenable to surgery, and have a grim prognosis: median survival after a diagnosis of unresectable PDAC is under 6 months, and survival after 5 years is around 5%[12].

A key factor obstructing the development of effective therapies for PDAC is the presence of cancer stem cells (CSCs)[29], cells capable of both differentiation and self-renewal[40], proposed as the main regulators of tumour progression[32], and involved in chemoresistance, evidencing an increased prevalence within the tumour after treatment with gemcitabine[52]. In general, CSCs display enhanced chemoresistance, radioresistance and metastatic abilities when compared to differentiated cancer cells[32]. CSC are traditionally identified using a number of putative CSC markers. However, there is currently no general signature able to identify CSC, and the role of these in cancer stemness has not been satisfactorily elucidated yet, while contradictory reports regarding the suitability of CSC markers exist in the literature[32]. Recently, computational approaches to identify CSC have emerged, including a machine learning-based mRNA stemness index (mRNAsi)[58], methods to computationally quantify cellular pluripotency[59,60] and trajectory inference methods[61]. Reconciling the different perspectives provided by these approaches is therefore an essential step towards the reliable identification of CSC.

For the task of identifying cell subpopulations of interest, such as CSC, single-cell RNA-sequencing (scRNA-seq) is a powerful technology, able to delineate clusters of transcriptionally related cells.

Single-cell RNA-sequencing has become the state-of-the-art technology for uncovering different cell types and functions within organs and tissues[63], and has been used to investigate cellular heterogeneity in cancer[65].

Following the identification of CSC, differential expression analysis on the scRNA-seq data, followed by gene set enrichment analysis, can be used to identify the genes and molecular mechanisms characterizing CSC.

With regards to the question of eliminating CSC, a promising avenue that has emerged in recent years consists of epigenetic therapies[85], as the deregulation of epigenetics pathways is thought to promote the maintenance and survival of CSCs[85].

This thesis aimed to reliably identifying pancreatic cancer stem cells (PCSC) in scRNA-seq data by using multiple lines of evidence, of determining their distinctive genes and enriched processes, and to test the effects of inhibiting BRD9, an epigenetic regulator reportedly involved in tumorigenesis, upon PCSC. **Figure 6.1** summarizes the objectives, the datasets and the methods employed:

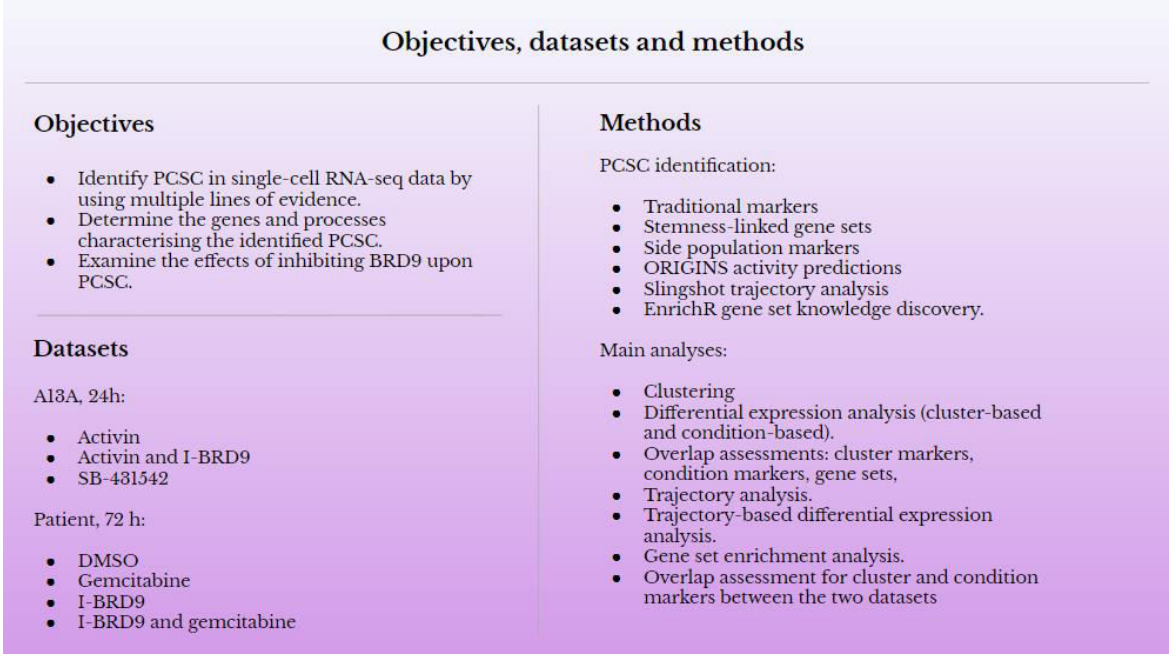


Figure 6.1. Research objectives, datasets and methods

In the following three sections, the main results linked to each of the three aims will be reported, followed by conclusions in the last section.

6.2 Aim I: Identifying PCSCs

The identification of PCSCs has traditionally been performed using putative PCSC markers derived experimentally, but often lacking a clear mechanism of action by which they would be associated with cancer stemness. Gene signatures consisting of markers derived using this paradigm typically involve a very small number of genes, or even single genes. However, the results presented in this thesis do not support this approach towards the identification of PCSC. Traditionally derived PCSC markers were found to describe generally non-overlapping populations of no associations with stemness uncovered using the other lines of evidence, and thus were generally unable to identify PCSC. An exception was the promising epigenetic target *EZH2*, linking to a significant extent with both the **High stemness** population in the A13A dataset and the **Top stemness** population in the patient dataset.

Most importantly, no genes expressed solely in cells at the higher end were detected at all, that is, gene fitting a traditional "PCSC marker" idea. This further suggests the unsuitability of identifying PCSC through descriptors built on a single gene, or on only a few genes.

Contrastingly, the CCRSA gene sets, largely derived computationally by used mRNA stemness index-based approaches, were able to make an identification of PCSC that was consistently supported by the other lines of evidence: ORIGINS activity predictions, Slingshot trajectory inference, and enrichment using stemness-linked Enrichr databases.

With regards to the side population-related SPLCL genes, they were shown able to identify a population with some stemness characteristics, also marked by the strong expression of a few traditionally-derived PCSC markers (*REG4*, *TSPAN8*, *ALDH1A1*) and by high ORIGINS activity scores, but one that was shown not to lie, in fact, at the origin of the trajectory of differentiation of the cells.

Thus, the population identified using these genes corresponds only to some aspects of the CSC concept – it represents a group of cells that regains some developmental potential in its evolution.

Clinically, the results suggest that a move away from the usage of proposed "PCSC markers" towards the functionally related CCRSA genes is necessary for the correct identification and annihilation of these highly tumourigenic cells.

Furthermore, the results do not support a binary division between CSC and non-CSC but assert the existence of intermediary populations of clinical relevance, a description fitting both the **SP** population found in the A13A dataset, on one hand, and the **High stemness** cluster from the A13A dataset as well as **p53 signalling** cluster from the patient dataset, of which the former represents a regain in stemness at a later stage of evolution, while the latter represents a partial retention of developmental potential, more pronounced for the **High stemness** cluster than for the **p53 signalling** one.

6.3 Aim II: Establishing genes and processes characteristic to stemness in PDAC

Stemness was identified to be crucially linked with **cell cycle abnormalities**, affecting processes such as chromosome segregation, nuclear division, cell cycle phase transition, spindle assembly and kinetochore organization.

Despite the salient overlap between the **Top stemness** clusters in the two datasets, in terms of both markers and enriched GO terms, notable differences between the two did exist, in that processes related to DNA replication and DNA repair were very strongly enriched for the **Top stemness** cluster in the Patient dataset, but not in the A13A one. Contrastingly, in the A13A dataset it was the **High stemness** cluster, of lower stemness character, where the two processes were strongly enriched, although not to the extent seen in the **Top stemness** cluster from the patient dataset.

Thus, in contrast with the idea that CSC resist chemotherapy through their quiescence, *proliferative* and *highly proliferative* CSC are identified in the A13A and the patient dataset, respectively, with the latter also showing outstanding DNA repair capabilities.

With regards to drug resistance, however, it appears that the stemness and the drug resistance peaks did not coincide among the identified subpopulations, with both datasets displaying exceptionally drug resistant subpopulation that were not bona fide CSC, albeit still retaining (**p53 signalling** from the Patient dataset) or having regained (**SP** from the A13A dataset) considerable amounts of developmental potential. Eliminating CSCs may, therefore, not be sufficient to eliminate the tumour, as non-CSC highly drug-resistant populations, may not only survive the concurrent therapeutic targeting of typical bulk cells and of CSC due to being neither, but may also regain enough developmental potential to regenerate the tumour; in particular, dedifferentiation from the **p53 signalling** cluster in the patient dataset to the **Top stemness** one appears possible. Among the two drug resistant clusters, **SP** was transcriptionally of little relation to the CCRSA-linked **Top stemness** and **High stemness** populations from both datasets, while **p53 signalling** did show a markedly higher representation of CCRSA genes than any cluster in the patient dataset except **Top stemness**, but one that still fell below the significance threshold.

Clinically, DNA repair enzymes appears to be the most promising targets for combination therapy towards the annihilation of the highly proliferative CSC, while the annihilation of highly drug resistant but non-CSC cells emerges as a necessary additional goal of therapy. In particular, the non-CCRSA-linked but highly drug resistant **SP** cluster from the A13A dataset would be unlikely to respond to the targeting of regulators of CCRSA genes, but rather needs an inhibition of its own characteristic enriched mechanisms, such as detoxification.

As a final note, while CCRSA-linked alterations strongly emerged as central to stemness, it is possible that **SP**-like populations, that is, drug-resistant PDAC cells that regain a substantial fraction of the developmental potential of PCSC during their evolution, might additionally exist, and be guided by

different mechanisms of actions. Similarly, other populations analogous to **High stemness** from the A13A dataset and **p53 signalling** from the patient dataset (that is, retaining an overexpression of numerous CCRSA markers) might also exist, as hinted from the fact that the RNA velocity analysis in the patient dataset suggested that differentiation in directions other than that of **p53 signalling**, although that cluster was found to be the second in all Slingshot lineages, may be possible. This observation suggests that scRNA-seq may prove extremely useful in precision medicine, being able to identify the specific cells with full of partial stemness characteristics in a PDAC patient and their molecular signatures and inform the choice of different therapeutic courses of action.

6.4 Aim III: Evaluating the effects of I-BRD9 in PCSCs

The fact that identified CSCs did not evidence any genes with the expression pattern expected for bona fide CSC markers (that is, essentially absent in non-CSCs), but rather an ample set of distinctly overexpressed cell cycle-related genes, under the control of several regulators, suggest that epigenetic therapies, given their potential of targeting the genome in its entirety, might be a promising avenue for the treatment of PDAC. EZH2, a modulator of gene silencing, might serve as a potentially useful target in this regard, as a regulator of multiple crucial CCRSA genes.

With regards to the effects of I-BRD9, the lower duration of 24 h in the A13A experiment was not sufficient for a decrease in stemness to occur for the A13A dataset, possibly also due to genetic differences between the A13A and the patient cells, but considerable effects were seen for the 72 h treatment implemented for the patient dataset. The most notable of these effects was the ~6-fold underrepresentation of I-BRD9-treated cells relative to DMSO-treated ones in the **Top stemness** cluster. This underrepresentation was partly reversed in the I-BRD9 and gemcitabine condition, however, with I-BRD9 and gemcitabine having opposite effect upon cancer stemness and the cumulative effect simply being the additive result of these opposite trends. This suggests that co-

administering I-BRD9 with gemcitabine would not be optimal as a treatment regimen, but rather I-BRD9 could perhaps be combined with another cytotoxic agent, or perhaps it could be beneficial to administer I-BRD9 and gemcitabine sequentially, with a delay, rather than concurrently.

I-BRD9 was shown to downregulate important CCRSA genes *TOP2A*, *CDK1*, *UBE2C*, *CCNB1*, *PBK* and *CENPF* in the patient dataset, and these effects appear likely to be mediated by the downregulation of *EZH2*, also identified to be downregulated after the administering of I-BRD9 in the patient dataset.

6.5 Conclusion

By analyzing single-cell RNA-sequencing data in PDAC, this thesis has evidenced that genes and processes linked to the CCRSA gene sets, previously predicted as stemness regulators in a variety of cancers by using mRNAsi-based approaches, lie at the crux of stemness in PDAC. Contrastingly, putative PCSC markers were largely unable to identify PCSC, the exception being the epigenetic regulator EZH2. Correspondingly, the administration of Activin A, previously reported to stimulate proposed PCSC as identified using traditional markers, does not stimulate CSC in the A13A dataset, while on the contrary, the chemical inhibition of Activin signalling had a small but consistent effect in this direction.

Therefore, cell cycle abnormalities are the central element governing PDAC stemness. Promising molecular targets for the elimination of PCSC that were identified in this thesis include TOP2A and CENPF, strongly overexpressed in both **Top stemness** clusters.

Highly drug resistant populations, exceeding CSC in this regard without being themselves bona fide CSC, were identified, namely **SP** in the A13A dataset and **p53 signalling** in the patient dataset. While functionally different otherwise, the two drug resistant clusters shared an overexpression of *GDF15*, the only SPLCL gene overexpressed in both the **SP** and the **Top stemness** clusters in the A13A dataset. In the **p53 signalling** cluster, *GDF15* is the strongest marker by average $\log_2$ fold-change,

thus suggesting it might be a suitable target for the elimination of highly drug resistant PDAC cell populations.

I-BRD9 achieved a ~6-fold reduction of **Top stemness** cells in the patient dataset relative to the DMSO control, likely as a result of the downregulation of *TOP2A*. This downregulation might be mediated by another epigenetic regulator, *EZH2*, an indicator of both **High stemness** cells in the A13A dataset and **Top stemness** cells in the Patient dataset. Synergy with gemcitabine was not noted, with the aforementioned effect significantly diminishing its magnitude when I-BRD9 was administered together with gemcitabine. Otherwise, gemcitabine drove increases in cancer stemness, and was associated with the emergence of cells with enhanced autophagy capacities in the **p53 signalling** cluster. Sizeable effects due to I-BRD9 in the A13A dataset, where the duration of treatment was three times shorter (24 h), are not noted. The discrepancy in the effect of I-BRD9 in the two dataset probably being a combined effect of both the different duration of treatment and biological heterogeneity, the genetical differences between the two samples. The enhanced DNA repair and DNA replication in the patient **Top stemness** cluster, suggestive of possible BRCA-ness in the patient sample, might have played a role in rendering the patient sample more vulnerable to I-BRD9 treatment.

In conclusion, this thesis integrates distinct lines of evidence to arrive at an identification of CSC in scRNA-seq PDAC data, uncovers genes and processes distinctly activates in CSC, detects highly drug resistant populations of partial stemness characteristics, and finds novel potential therapeutic targets: *TOP2A* and *CENPF* for addressing cancer stemness, *GDF15* for addressing chemoresistance. I-BRD9 is shown to achieve a remarkable reduction in the number of CSC in the patient dataset, possibly mediated in part by another epigenetic regulator, *EZH2*, but synergy with gemcitabine is not observed.

7 References

1. Pandol, S.J. The exocrine pancreas. in *Colloquium series on integrated systems physiology: from molecule to function*, Vol. 3 1-64 (Morgan & Claypool Life Sciences, 2011).
2. Carrière, C. & Korc, M. Regulatory Signaling in Pancreatic Organogenesis: Implications for Aberrant Signaling in Pancreatic Cancer. in *Handbook of Cell Signaling* 2611-2620 (Elsevier, 2010).
3. Grapin-Botton, A. Ductal cells of the pancreas. *The international journal of biochemistry & cell biology* **37**, 504-510 (2005).
4. Orth, M.*, et al.* Pancreatic ductal adenocarcinoma: Biological hallmarks, current status, and future perspectives of combined modality treatment approaches. *Radiation Oncology* **14**, 1-20 (2019).
5. Niger, M.*, et al.* One size does not fit all for pancreatic cancers: A review on rare histologies and therapeutic approaches. *World Journal of Gastrointestinal Oncology* **12**, 833 (2020).
6. Haeberle, L. & Esposito, I. Pathology of pancreatic cancer. *Translational gastroenterology and hepatology* **4**(2019).
7. Singh, K.*, et al.* Kras mutation rate precisely orchestrates ductal derived pancreatic intraepithelial neoplasia and pancreatic cancer. *Laboratory Investigation* **101**, 177-192 (2021).
8. Lee, A.Y.*, et al.* Cell of origin affects tumour development and phenotype in pancreatic ductal adenocarcinoma. *Gut* **68**, 487-498 (2019).
9. Waters, A.M. & Der, C.J. KRAS: the critical driver and therapeutic target for pancreatic cancer. *Cold Spring Harbor perspectives in medicine* **8**, a031435 (2018).
10. Ryan, D.P., Hong, T.S. & Bardeesy, N. Pancreatic adenocarcinoma. *New England Journal of Medicine* **371**, 1039-1049 (2014).
11. Lambert, A.*, et al.* An update on treatment options for pancreatic adenocarcinoma. *Therapeutic advances in medical oncology* **11**, 1758835919875568 (2019).
12. Bengtsson, A., Andersson, R. & Ansari, D. The actual 5-year survivors of pancreatic ductal adenocarcinoma based on real-world data. *Scientific reports* **10**, 1-9 (2020).
13. Winter, K.*, et al.* Diagnostic and therapeutic recommendations in pancreatic ductal adenocarcinoma. Recommendations of the Working Group of the Polish Pancreatic Club. *Przeglad Gastroenterologiczny* **14**, 1 (2019).
14. Bekkali, N.L. & Oppong, K.W. Pancreatic ductal adenocarcinoma epidemiology and risk assessment: Could we prevent? Possibility for an early diagnosis. *Endoscopic ultrasound* **6**, S58 (2017).
15. Stoica, A.-F., Chang, C.-H. & Pauklin, S. Molecular Therapeutics of Pancreatic Ductal Adenocarcinoma: Targeted Pathways and the Role of Cancer Stem Cells. *Trends in Pharmacological Sciences* **41**, 977-993 (2020).
16. Huang, J.*, et al.* Worldwide burden of, risk factors for, and trends in pancreatic cancer. *Gastroenterology* **160**, 744-754 (2021).
17. Schawkat, K., Manning, M.A., Glickman, J.N. & Mortele, K.J. Pancreatic ductal adenocarcinoma and its variants: pearls and perils. *Radiographics* **40**, 1219-1239 (2020).
18. Li, S.-S.*, et al.* ABO blood type, smoking status, other risk factors and prognosis of pancreatic ductal adenocarcinoma. *Medicine* **99**(2020).
19. Park, J.-H.*, et al.* Changes in metabolic syndrome status are associated with altered risk of pancreatic cancer: A nationwide cohort study. *Gastroenterology* **162**, 509-520. e507 (2022).
20. Grossberg, A.J.*, et al.* Multidisciplinary standards of care and recent progress in pancreatic ductal adenocarcinoma. *CA: A Cancer Journal for Clinicians* **70**, 375-403 (2020).

21. Mario, C*., et al.* Epidemiology and risk factors of pancreatic cancer. *Acta Bio Medica: Atenei Parmensis* **89**, 141 (2018).
22. Sohal, D.P., Willingham, F.F., Falconi, M., Raphael, K.L. & Crippa, S. Pancreatic adenocarcinoma: improving prevention and survivorship. *American Society of Clinical Oncology Educational Book* **37**, 301-310 (2017).
23. Kratz, C.P*., et al.* Analysis of the Li-Fraumeni Spectrum Based on an International Germline TP53 Variant Data Set: An International Agency for Research on Cancer TP53 Database Analysis. *JAMA oncology* **7**, 1800-1805 (2021).
24. Yeo, T.P. Demographics, epidemiology, and inheritance of pancreatic ductal adenocarcinoma. in *Seminars in oncology*, Vol. 42 8-18 (Elsevier, 2015).
25. Hayashi, H., Higashi, T., Miyata, T., Yamashita, Y.i. & Baba, H. Recent advances in precision medicine for pancreatic ductal adenocarcinoma. *Annals of Gastroenterological Surgery* **5**, 457-466 (2021).
26. Wu, E., Zhou, S., Bhat, K. & Ma, Q. CA 19-9 and pancreatic cancer. *Clinical advances in hematology & oncology: H&O* **11**, 53 (2013).
27. Ansari, D., Gustafsson, A. & Andersson, R. Update on the management of pancreatic cancer: surgery is not enough. *World journal of gastroenterology: WJG* **21**, 3157 (2015).
28. Javed, A.A*., et al.* Outcome of patients with borderline resectable pancreatic cancer in the contemporary era of neoadjuvant chemotherapy. *Journal of Gastrointestinal Surgery* **23**, 112-121 (2019).
29. Di Carlo, C., Brandi, J. & Cecconi, D. Pancreatic cancer stem cells: perspectives on potential therapeutic approaches of pancreatic ductal adenocarcinoma. *World journal of stem cells* **10**, 172 (2018).
30. Tian, C*., et al.* Proteomic analyses of ECM during pancreatic ductal adenocarcinoma progression reveal different contributions by tumor and stromal cells. *Proceedings of the National Academy of Sciences* **116**, 19609-19618 (2019).
31. Liot, S*., et al.* Stroma involvement in pancreatic ductal adenocarcinoma: An overview focusing on extracellular matrix proteins. *Frontiers in Immunology* **12**, 709 (2021).
32. Patil, K., Khan, F.B., Akhtar, S., Ahmad, A. & Uddin, S. The plasticity of pancreatic cancer stem cells: Implications in therapeutic resistance. *Cancer and Metastasis Reviews* **40**, 691-720 (2021).
33. Bulle, A. & Lim, K.-H. Beyond just a tight fortress: contribution of stroma to epithelial-mesenchymal transition in pancreatic cancer. *Signal Transduction and Targeted Therapy* **5**, 1-12 (2020).
34. Peng, J*., et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research* **29**, 725-738 (2019).
35. Cohen, R*., et al.* Targeting cancer cell metabolism in pancreatic adenocarcinoma. *Oncotarget* **6**, 16832 (2015).
36. Lanfranca, M.P*., et al.* Metabolism and epigenetics of pancreatic cancer stem cells. in *Seminars in cancer biology*, Vol. 57 19-26 (Elsevier, 2019).
37. Krah, N.M*., et al.* The acinar differentiation determinant PTF1A inhibits initiation of pancreatic ductal adenocarcinoma. *Elife* **4**, e07125 (2015).
38. Parte, S., Nimmakayala, R.K., Batra, S.K. & Ponnusamy, M.P. Acinar to ductal cell trans-differentiation: A prelude to dysplasia and pancreatic ductal adenocarcinoma. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 188669 (2021).
39. Schmidtlein, P.M*., et al.* A comparative endocrine trans-differentiation approach to pancreatic ductal adenocarcinoma cells with different EMT phenotypes identifies quasi-mesenchymal tumor cells as those with highest plasticity. *Cancers* **13**, 4663 (2021).
40. Simeone, D.M. Pancreatic cancer stem cells: implications for the treatment of pancreatic cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **14**, 5646-5648 (2008).

41. Bonnet, D. & Dick, J.E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* **3**, 730-737 (1997).
42. Piccirillo, S.G., Binda, E., Fiocco, R., Vescovi, A.L. & Shah, K. Brain cancer stem cells. *Journal of molecular medicine* **87**, 1087-1095 (2009).
43. Velasco-Velázquez, M.A., Homsi, N., De La Fuente, M. & Pestell, R.G. Breast cancer stem cells. *The international journal of biochemistry & cell biology* **44**, 573-577 (2012).
44. Ricci-Vitiani, L., Fabrizi, E., Palio, E. & De Maria, R. Colon cancer stem cells. *Journal of Molecular Medicine* **87**, 1097-1104 (2009).
45. Tomizawa, Y., Wu, T.T. & Wang, K.K. Epithelial mesenchymal transition and cancer stem cells in esophageal adenocarcinoma originating from Barrett's esophagus. *Oncology letters* **3**, 1059-1063 (2012).
46. Sell, S. & Leffert, H.L. Liver cancer stem cells. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **26**, 2800 (2008).
47. Eramo, A., Haas, T. & De Maria, R. Lung cancer stem cells: tools and targets to fight lung cancer. *Oncogene* **29**, 4625-4635 (2010).
48. Bapat, S.A. Human ovarian cancer stem cells. *Reproduction (Cambridge, England)* **140**, 33-41 (2010).
49. Fioriti, D.*, et al.* Cancer stem cells in prostate adenocarcinoma: a target for new anticancer strategies. *J Cell Physiol* **216**, 571-575 (2008).
50. Li, C.*, et al.* Identification of pancreatic cancer stem cells. *Cancer Res* **67**, 1030-1037 (2007).
51. Fitzgerald, T.L. & McCubrey, J.A. Pancreatic cancer stem cells: association with cell surface markers, prognosis, resistance, metastasis and treatment. *Advances in biological regulation* **56**, 45-50 (2014).
52. Mueller, M.T.*, et al.* Combined targeted treatment to eliminate tumorigenic cancer stem cells in human pancreatic cancer. *Gastroenterology* **137**, 1102-1113 (2009).
53. Safa, A.R. Resistance to drugs and cell death in cancer stem cells (CSCs). *Journal of translational science* **6**(2020).
54. Prieto-Vila, M., Takahashi, R.-u., Usuba, W., Kohama, I. & Ochiya, T. Drug resistance driven by cancer stem cells and their niche. *International journal of molecular sciences* **18**, 2574 (2017).
55. Gzil, A.*, et al.* Markers of pancreatic cancer stem cells and their clinical and therapeutic implications. *Molecular biology reports* **46**, 6629-6645 (2019).
56. Amsterdam, A., Raanan, C., Schreiber, L., Polin, N. & Givol, D. LGR5 and Nanog identify stem cell signature of pancreas beta cells which initiate pancreatic cancer. *Biochemical and biophysical research communications* **433**, 157-162 (2013).
57. Herreros-Villanueva, M.*, et al.* SOX2 promotes dedifferentiation and imparts stem cell-like features to pancreatic cancer cells. *Oncogenesis* **2**, e61-e61 (2013).
58. Malta, T.M.*, et al.* Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* **173**, 338-354. e315 (2018).
59. Senra, D., Guisoni, N. & Diambra, L. ORIGINS: a protein network-based approach to quantify cell pluripotency from scRNA-seq data. *bioRxiv* (2022).
60. Gulati, G.S.*, et al.* Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405-411 (2020).
61. Street, K.*, et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19**, 1-16 (2018).
62. Cao, J.*, et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
63. Jovic, D.*, et al.* Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine* **12**, e694 (2022).

64. Tang, F.*, et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**, 377-382 (2009).
65. Lin, K. & Shen, J.P. Elucidating cancer stem cells heterogeneity in colorectal cancer by single-cell RNA sequencing. *Cancer Research* **82**, 6066-6066 (2022).
66. He, Y.*, et al.* Quantitative Evaluation of Stem-like Markers of Human Glioblastoma Using Single-Cell RNA Sequencing Datasets. *Cancers* **15**, 1557 (2023).
67. Bian, J.*, et al.* Characterization of immunogenicity of malignant cells with stemness in intrahepatic cholangiocarcinoma by single-cell RNA sequencing. *Stem Cells International* **2022**(2022).
68. Wang, L., Mo, S., Li, X., He, Y. & Yang, J. Single-cell RNA-seq reveals the immune escape and drug resistance mechanisms of mantle cell lymphoma. *Cancer biology & medicine* **17**, 726 (2020).
69. Aissa, A.F.*, et al.* Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature communications* **12**, 1628 (2021).
70. Gallipoli, P. & Huntly, B.J. Novel epigenetic therapies in hematological malignancies: current status and beyond. in *Seminars in Cancer Biology*, Vol. 51 198-210 (Elsevier, 2018).
71. Mohammad, H.P., Barbash, O. & Creasy, C.L. Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer. *Nature medicine* **25**, 403-418 (2019).
72. Lee, M.K., Brown, M.S., Wilkins, O.M., Pattabiraman, D.R. & Christensen, B.C. Distinct cytosine modification profiles define epithelial-to-mesenchymal cell-state transitions. *Epigenomics* **14**, 519-535 (2022).
73. Pfister, S.X. & Ashworth, A. Marked for death: targeting epigenetic changes in cancer. *Nature reviews Drug discovery* **16**, 241-263 (2017).
74. Gao, J.*, et al.* Aberrant DNA methyltransferase expression in pancreatic ductal adenocarcinoma development and progression. *Journal of Experimental & Clinical Cancer Research* **32**, 1-10 (2013).
75. Ouaissi, M.*, et al.* High histone deacetylase 7 (HDAC7) expression is significantly associated with adenocarcinomas of the pancreas. *Annals of Surgical Oncology* **15**, 2318-2328 (2008).
76. Dutruel, C.*, et al.* Early epigenetic downregulation of WNK2 kinase during pancreatic ductal adenocarcinoma development. *Oncogene* **33**, 3401-3410 (2014).
77. Omura, N. & Goggins, M. Epigenetics and epigenetic alterations in pancreatic cancer. *International journal of clinical and experimental pathology* **2**, 310 (2009).
78. Ávila-López, P.*, et al.* H2A. Z overexpression suppresses senescence and chemosensitivity in pancreatic ductal adenocarcinoma. *Oncogene* **40**, 2065-2080 (2021).
79. Jones, P.A. At the tipping point for epigenetic therapies in cancer. *The Journal of clinical investigation* **124**, 14-16 (2014).
80. Ronnekleiv-Kelly, S.M., Sharma, A. & Ahuja, N. Epigenetic therapy and chemosensitization in solid malignancy. *Cancer Treatment Reviews* **55**, 200-208 (2017).
81. Rothbart, S.B. & Baylin, S.B. Epigenetic therapy for epithelioid sarcoma. *Cell* **181**, 211 (2020).
82. Rugo, H.S.*, et al.* The promise for histone methyltransferase inhibitors for epigenetic therapy in clinical oncology: a narrative review. *Advances in therapy* **37**, 3059-3082 (2020).
83. Lee, J.E. & Kim, M.-Y. Cancer epigenetics: Past, present and future. in *Seminars in Cancer Biology* (Elsevier, 2021).
84. Zheng, Y.-C. & Feng, S.-Q. Epigenetic modifications as therapeutic targets. *Current Drug Targets* **21**, 1046-1046 (2020).
85. Toh, T.B., Lim, J.J. & Chow, E.K.-H. Epigenetics in cancer stem cells. *Molecular cancer* **16**, 29 (2017).

86. Zhu, X., Liao, Y. & Tang, L. Targeting BRD9 for cancer treatment: a new strategy. *OncoTargets and therapy* **13**, 13191 (2020).
87. Fu, B.*, et al.* Evaluation of GATA-4 and GATA-5 methylation profiles in human pancreatic cancers indicate promoter methylation patterns distinct from other human tumor types. *Cancer biology & therapy* **6**, 1546-1552 (2007).
88. Lonardo, E.*, et al.* Nodal/Activin signaling drives self-renewal and tumorigenicity of pancreatic cancer stem cells and provides a target for combined drug therapy. *Cell stem cell* **9**, 433-446 (2011).
89. Macosko, E.Z.*, et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).
90. Melsted, P.*, et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature biotechnology* **39**, 813-818 (2021).
91. Cunningham, F.*, et al.* Ensembl 2022. *Nucleic acids research* **50**, D988-D995 (2022).
92. Zheng, G.X.*, et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
93. La Manno, G.*, et al.* RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
94. Team, R.C. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. (2022).
95. Team, R.C. R: A language and environment for statistical computing. (2013).
96. RStudio, T. RStudio: integrated development for R. *Rstudio Team, PBC, Boston, MA URL http://www. rstudio. com* (2020).
97. Hao, Y.*, et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587. e3529 (2021).
98. Stuart, T.*, et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).
99. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420 (2018).
100. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495-502 (2015).
101. Alquicira-Hernandez, J. & Powell, J.E. Nebulosa recovers single-cell gene expression signals by kernel density estimation. *Bioinformatics* **37**, 2485-2487 (2021).
102. Wickham, H., Chang, W. & Wickham, M.H. Package 'ggplot2'. *Create elegant data visualisations using the grammar of graphics. Version* **2**, 1-189 (2016).
103. Larsson, J.*, et al.* Package 'eulerr'. (2021).
104. Krassowski, M. ComplexUpset: Create complex UpSet plots using ggplot2 components. *R package version 0.5* **18**(2020).
105. Wei, T.*, et al.* Package 'corrplot'. *Statistician* **56**, e24 (2017).
106. Murrell, P. The gridGraphics Package. *R Journal* **7**(2015).
107. Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**, 608-609 (2015).
108. Yu, G. ggplotify: Convert Plot to 'grob'or 'ggplot'Object. R package version 0.0. 5. (2020).
109. Wickham, H., Francois, R., Henry, L. & Müller, K. dplyr. in *useR! Conference* (2014).
110. Consortium, G.O. The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research* **47**, D330-D338 (2019).
111. Kanehisa, M. The KEGG database. in *Novartis foundation symposium* 91-100 (Wiley Online Library, 2002).
112. Slenter, D.N.*, et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research* **46**, D661-D667 (2018).
113. Carlson, M., Falcon, S., Pages, H. & Li, N. org. Hs. eg. db: Genome wide annotation for Human. *R package version* **3**, 3 (2019).

114. Germain, P.-L., Lun, A., Macnair, W. & Robinson, M.D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research* **10**, 979 (2022).
115. Xi, N.M. & Li, J.J. Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. *STAR protocols* **2**, 100699 (2021).
116. Jaccard, P. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* **44**, 223-270 (1908).
117. Grandi, F*., et al.* popsicleR: AR Package for Pre-processing and Quality Control Analysis of Single Cell RNA-seq Data. *Journal of Molecular Biology*, 167560 (2022).
118. Surumbayeva, A*., et al.* Preparation of mouse pancreatic tumor for single-cell RNA sequencing and analysis of the data. *STAR protocols* **2**, 100989 (2021).
119. Staadig, A., Hedman, J. & Tillmar, A. Applying Unique Molecular Indices with an Extensive All-in-One Forensic SNP Panel for Improved Genotype Accuracy and Sensitivity. *Genes* **14**, 818 (2023).
120. Ilicic, T*., et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome biology* **17**, 1-15 (2016).
121. M Ascensión, A., Ibáñez-Solé, O., Inza, I., Izeta, A. & Araúzo-Bravo, M.J. Triku: a feature selection method based on nearest neighbors for single-cell data. *GigaScience* **11**, giac017 (2022).
122. Lin, W*., et al.* Single-cell transcriptome analysis of tumor and stromal compartments of pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome medicine* **12**, 1-14 (2020).
123. Shannon, C.E. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* **5**, 3-55 (2001).
124. Oksanen, J*., et al.* Package 'vegan'. *Community ecology package, version* **2**, 1-295 (2013).
125. Wang, R*., et al.* Influence of species richness, evenness, and composition on optical diversity: A simulation study. *Remote Sensing of Environment* **211**, 218-228 (2018).
126. Seweryn, M.T., Pietrzak, M. & Ma, Q. Application of information theoretical approaches to assess diversity and similarity in single-cell transcriptomics. *Computational and structural biotechnology journal* **18**, 1830-1837 (2020).
127. Sherwin, W.B., Jabot, F., Rush, R. & Rossetto, M. Measurement of biological information with applications from genes to landscapes. *Molecular Ecology* **15**, 2857-2869 (2006).
128. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome biology* **20**, 1-15 (2019).
129. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome biology* **23**, 1-20 (2022).
130. Korsunsky, I*., et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**, 1289-1296 (2019).
131. Tran, H.T.N*., et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology* **21**, 1-32 (2020).
132. Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).
133. Xie, Z*., et al.* Gene set knowledge discovery with enrichr. *Current protocols* **1**, e90 (2021).
134. Xia, P. & Liu, D.-H. Cancer stem cell markers for liver cancer and pancreatic cancer. *Stem cell research*, 102701 (2022).
135. Ishiwata, T*., et al.* Pancreatic cancer stem cells: features and detection methods. *Pathology & Oncology Research* **24**, 797-805 (2018).
136. Valle, S., Martin-Hijano, L., Alcalá, S., Alonso-Nocelo, M. & Sainz Jr, B. The ever-evolving concept of the cancer stem cell in pancreatic cancer. *Cancers* **10**, 33 (2018).

137. Amsterdam, A.*, et al.* Modulation of c-kit expression in pancreatic adenocarcinoma: a novel stem cell marker responsible for the progression of the disease. *Acta histochemica* **116**, 197-203 (2014).
138. Wang, H., Rana, S., Giese, N., Büchler, M.W. & Zöller, M. Tspan8, CD44v6 and alpha6beta4 are biomarkers of migrating pancreatic cancer-initiating cells. *International journal of cancer* **133**, 416-426 (2013).
139. Bishnupuri, K.S., Sainathan, S.K., Ciorba, M.A., Houchen, C.W. & Dieckgraefe, B.K. Reg4 interacts with CD44 to regulate proliferation and stemness of colorectal and pancreatic cancer cells. *Molecular Cancer Research* **20**, 387-399 (2022).
140. Wang, V.M.-Y.*, et al.* CD9 identifies pancreatic cancer stem cells and modulates glutamine metabolism to fuel tumour growth. *Nature cell biology* **21**, 1425-1435 (2019).
141. van Vlerken, L.E.*, et al.* EZH2 is required for breast and pancreatic cancer stem cell maintenance and can be used as a functional cancer stem cell reporter. *Stem cells translational medicine* **2**, 43-52 (2013).
142. Dumartin, L.*, et al.* ER stress protein AGR2 precedes and is involved in the regulation of pancreatic cancer initiation. *Oncogene* **36**, 3094-3103 (2017).
143. Jo, J.H.*, et al.* GLRX3, a novel cancer stem cell-related secretory biomarker of pancreatic ductal adenocarcinoma. *BMC cancer* **21**, 1-19 (2021).
144. Abel, E.V.*, et al.* HNF1A is a novel oncogene that regulates human pancreatic cancer stem cell properties. *Elife* **7**, e33947 (2018).
145. Park, J.-H., Chung, S., Matsuo, Y. & Nakamura, Y. Development of small molecular compounds targeting cancer stem cells. *MedChemComm* **8**, 73-80 (2017).
146. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 1-13 (2008).
147. Huang, X.-Y.*, et al.* A New Stemness-Related Prognostic Model for Predicting the Prognosis in Pancreatic Ductal Adenocarcinoma. *BioMed Research International* **2021**(2021).
148. Huang, C., Hu, C.-G., Ning, Z.-K., Huang, J. & Zhu, Z.-M. Identification of key genes controlling cancer stem cell characteristics in gastric cancer. *World journal of gastrointestinal surgery* **12**, 442 (2020).
149. Liao, Y., Xiao, H., Cheng, M. & Fan, X. Bioinformatics analysis reveals biomarkers with cancer stem cell characteristics in lung squamous cell carcinoma. *Frontiers in genetics* **11**, 427 (2020).
150. Luo, M., Zeng, H., Ma, X.-Y. & Ma, X.-L. Identification of hub genes for ovarian cancer stem cell properties with weighted gene co-expression network analysis. *Sichuan da xue xue bao. Yi xue ban= Journal of Sichuan University. Medical Science Edition* **52**, 248-258 (2021).
151. Pei, J., Wang, Y. & Li, Y. Identification of key genes controlling breast cancer stem cell characteristics via stemness indices analysis. *Journal of translational medicine* **18**, 1-15 (2020).
152. Xia, P., Li, Q., Wu, G. & Huang, Y. Identification of glioma cancer stem cell characteristics based on weighted gene prognosis module co-expression network analysis of transcriptome data stemness indices. *Journal of Molecular Neuroscience* **70**, 1512-1520 (2020).
153. Liu, Y.*, et al.* Comprehensive analysis of the control of cancer stem cell characteristics in endometrial cancer by network analysis. *Computational and mathematical methods in medicine* **2021**(2021).
154. Pan, S., Zhan, Y., Chen, X., Wu, B. & Liu, B. Identification of biomarkers for controlling cancer stem cell characteristics in bladder cancer by network analysis of transcriptome data stemness indices. *Frontiers in oncology*, 613 (2019).

155. Zhang, J., Yin, H. & Tong, Z. Weighted Gene Coexpression Network Analysis of Key Genes Controlling Cancer Stem Cell Characteristics by Mrnasi in Colon Adenocarcinoma. (2021).
156. Zhao, Z*., et al.* Identification of Biomarkers Associated with Hepatocellular Carcinoma Stem Cell Characteristics Based on Co-Expression Network Analysis of Transcriptome Data and Stemness Index. *Critical Reviews™ in Eukaryotic Gene Expression* **32**(2022).
157. Zhou, Z*., et al.* Overexpression of topoisomerase 2-alpha confers a poor prognosis in pancreatic adenocarcinoma identified by co-expression analysis. *Digestive diseases and sciences* **62**, 2790-2800 (2017).
158. Zhang, M*., et al.* Screening and identification of key biomarkers in pancreatic cancer: evidence from bioinformatic analysis. *Journal of Computational Biology* **27**, 1079-1091 (2020).
159. Szklarczyk, D*., et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research* **49**, D605-D612 (2021).
160. Dang, S.-C*., et al.* G-protein-signaling modulator 2 expression and role in a CD133+ pancreatic cancer stem cell subset. *OncoTargets and therapy* **12**, 785 (2019).
161. Bian, Y*., et al.* Target Deconvolution of a Multikinase Inhibitor with Antimetastatic Properties Identifies TAOK3 as a Key Contributor to a Cancer Stem Cell–Like PhenotypeTAOK3 Is a Novel Target in Pancreatic Cancer Stem Cells. *Molecular cancer therapeutics* **18**, 2097-2110 (2019).
162. Wang, W.Y*., et al.* A gene expression signature of epithelial tubulogenesis and a role for ASPM in pancreatic tumor progression. *Gastroenterology* **145**, 1110-1120 (2013).
163. Han, J.Y., Han, Y.K., Park, G.-Y., Kim, S.D. & Geun Lee, C. Bub1 is required for maintaining cancer stem cells in breast cancer cell lines. *Scientific reports* **5**, 1-10 (2015).
164. Huang, C., Han, Z. & Wu, D. Effects of TPX2 gene on radiotherapy sensitization in breast cancer stem cells. *Oncology letters* **14**, 1531-1535 (2017).
165. Turdo, A*., et al.* Effective targeting of breast cancer stem cells by combined inhibition of Sam68 and Rad51. *Oncogene* **41**, 2196-2209 (2022).
166. Peng, L*., et al.* GINS2 regulates matrix metallopeptidase 9 expression and cancer stem cell property in human triple negative Breast cancer. *Biomedicine & Pharmacotherapy* **84**, 1568-1574 (2016).
167. Cidado, J*., et al.* Ki-67 is required for maintenance of cancer stem cells but not cell proliferation. *Oncotarget* **7**, 6281 (2016).
168. Yang, N*., et al.* FOXM1 recruits nuclear Aurora kinase A to participate in a positive feedback loop essential for the self-renewal of breast cancer stem cells. *Oncogene* **36**, 3428-3440 (2017).
169. Murayama, T*., et al.* MCM10 compensates for Myc-induced DNA replication stress in breast cancer stem-like cells. *Cancer science* **112**, 1209-1224 (2021).
170. Xiong, Q*., et al.* miR-133b targets NCAPH to promote β-catenin degradation and reduce cancer stem cell maintenance in non-small cell lung cancer. *Signal transduction and targeted therapy* **6**, 1-3 (2021).
171. Chen, J., Chen, H., Yang, H. & Dai, H. SPC25 upregulation increases cancer stem cell properties in non-small cell lung adenocarcinoma cells and independently predicts poor survival. *Biomedicine & Pharmacotherapy* **100**, 233-239 (2018).
172. Liao, S*., et al.* PRC1 and RACGAP1 are Diagnostic Biomarkers of Early HCC and PRC1 Drives Self-Renewal of Liver Cancer Stem Cells. *Frontiers in Cell and Developmental Biology* **10**, 864051 (2022).
173. Bai, S*., et al.* Spindle and kinetochore-associated complex subunit 3 (SKA3) promotes stem cell-like properties of hepatocellular carcinoma cells through activating Notch signaling pathway. *Annals of Translational Medicine* **9**(2021).

174. Lin, S.*, et al.* NEK2 regulates stem-like properties and predicts poor prognosis in hepatocellular carcinoma. *Oncology reports* **36**, 853-862 (2016).
175. Li, Q.*, et al.* Kinesin family member 15 promotes cancer stem cell phenotype and malignancy via reactive oxygen species imbalance in hepatocellular carcinoma. *Cancer letters* **482**, 112-125 (2020).
176. Li, S.*, et al.* GINS1 induced sorafenib resistance by promoting cancer stem properties in human hepatocellular cancer cells. *Frontiers in cell and developmental biology*, 2057 (2021).
177. Ho, N.P.-y. PE-152: UBE2T: A Molecular Regulator for Cancer Stemness in Hepatocellular Carcinoma. *춘· 추계 학술대회(KASL)* **2018**, 204-205 (2018).
178. Lin, B.*, et al.* FAM83D associates with high tumor recurrence after liver transplantation involving expansion of CD44+ carcinoma stem cells. *Oncotarget* **7**, 77495 (2016).
179. Fang, X.*, et al.* ASF1B potentiates stem cell traits and tumor progression in hepatocellular carcinoma via histone H3. 3-dependent transcriptional reprogramming. *Cancer Research* **82**, 6065-6065 (2022).
180. Heo, J.*, et al.* Phosphorylation of TFCP2L1 by CDK1 is required for stem cell pluripotency and bladder carcinogenesis. *EMBO molecular medicine* **12**, e10880 (2020).
181. Xie, Q.*, et al.* CDC20 maintains tumor initiating cells. *Oncotarget* **6**, 13241 (2015).
182. Behnan, J., Grieg, Z., Joel, M., Ramsness, I. & Stangeland, B. Gene knockdown of CENPA reduces sphere forming ability and stemness of glioblastoma initiating cells. *Neuroepigenetics* **7**, 6-18 (2016).
183. Wang, J.*, et al.* Targeting dual specificity protein kinase TTK attenuates tumorigenesis of glioblastoma. *Oncotarget* **9**, 3081 (2018).
184. Joshi, K.*, et al.* MELK-dependent FOXM1 phosphorylation is essential for proliferation of glioma stem cells. *Stem cells* **31**, 1051-1063 (2013).
185. Li, X.*, et al.* Dual inhibition of Src and PLK1 regulate stemness and induce apoptosis through Notch1-SOX2 signaling in EGFRvIII positive glioma stem cells (GSCs). *Experimental Cell Research* **396**, 112261 (2020).
186. Joel, M.*, et al.* Targeting PBK/TOPK decreases growth and survival of glioma initiating cells in vitro and attenuates tumor growth in vivo. *Molecular cancer* **14**, 1-15 (2015).
187. Oue, N., Sentani, K., Sakamoto, N., Uraoka, N. & Yasui, W. Molecular carcinogenesis of gastric cancer: Lauren classification, mucin phenotype expression, and cancer stem cells. *International Journal of Clinical Oncology* **24**, 771-778 (2019).
188. Gong, D.-Y.*, et al.* Upregulation of ECT2 is associated with transcriptional program of cancer stem cells and predicts poor clinical outcome in gastric cancer. *Oncology letters* **20**, 1-1 (2020).
189. Enjoji, S.*, et al.* Stemness Is Enhanced in Gastric Cancer by a SET/PP2A/E2F1 AxisSET/PP2A/E2F1 Axis Maintains Gastric Cancer Cell Stemness. *Molecular cancer research* **16**, 554-563 (2018).
190. Hu, H.*, et al.* Proline-rich protein 11 regulates self-renewal and tumorigenicity of gastric cancer stem cells. *Cellular Physiology and Biochemistry* **47**, 1721-1728 (2018).
191. Nathansen, J.*, et al.* Oct4 confers stemness and radioresistance to head and neck squamous cell carcinoma by regulating the homologous recombination factors PSMC3IP and RAD54L. *Oncogene* **40**, 4214-4228 (2021).
192. Manoochehri, H., Asadi, S., Tanzadehpanah, H., Sheykhhasan, M. & Ghorbani, M. CDC25A is strongly associated with colorectal cancer stem cells and poor clinical outcome of patients. *Gene Reports* **25**, 101415 (2021).
193. Lin, X.*, et al.* CENPK Interacts with SOX6 to Promote Cervical Cancer Stemness and Chemoresistance Through Wnt and P53 Signaling. (2021).

194. Masciale, V.*, et al.* New perspectives in different gene expression profiles for early and locally advanced non-small cell lung cancer stem cells. *Frontiers in Oncology* **11**, 613198 (2021).
195. Lee, D.-F.*, et al.* Regulation of embryonic and induced pluripotency by aurora kinase-p53 signaling. *Cell stem cell* **11**, 179-194 (2012).
196. Renzova, T.*, et al.* Inactivation of PLK4-STIL module prevents self-renewal and triggers p53-dependent differentiation in human pluripotent stem cells. *Stem cell reports* **11**, 959-972 (2018).
197. Neganova, I.*, et al.* CDK1 plays an important role in the maintenance of pluripotency and genomic stability in human pluripotent stem cells. *Cell death & disease* **5**, e1508-e1508 (2014).
198. Jassal, B.*, et al.* The reactome pathway knowledgebase. *Nucleic acids research* **48**, D498-D503 (2020).
199. Giotti, B.*, et al.* Assembly of a parts list of the human mitotic cell cycle machinery. *Journal of molecular cell biology* **11**, 703-718 (2019).
200. Chen, C.*, et al.* ASF1b is a novel prognostic predictor associated with cell cycle signaling pathway in gastric cancer. *Journal of Cancer* **13**, 1985 (2022).
201. Monteverde, T.*, et al.* CKAP2L Promotes Non–Small Cell Lung Cancer Progression through Regulation of Transcription Elongation. *Cancer research* **81**, 1719-1731 (2021).
202. Li, H.*, et al.* Association of High Expression of Mitochondrial Fission Regulator 2 with Poor Survival of Patients with Esophageal Squamous Cell Carcinoma. *Journal of Cancer Prevention* **26**, 250 (2021).
203. Edogbanya, J.*, et al.* Evolution, structure and emerging roles of C1ORF112 in DNA replication, DNA damage responses, and cancer. *Cellular and Molecular Life Sciences* **78**, 4365-4376 (2021).
204. Wu, F., Wei, H., Liu, G. & Zhang, Y. Bioinformatics profiling of five immune-related lncRNAs for a prognostic model of hepatocellular carcinoma. *Frontiers in oncology* **11**, 1869 (2021).
205. Chen, Y.-C.*, et al.* Targeting DTL induces cell cycle arrest and senescence and suppresses cell growth and colony formation through TPX2 inhibition in human hepatocellular carcinoma cells. *OncoTargets and therapy* **11**, 1601 (2018).
206. Wang, X.*, et al.* UBE2T contributes to the prognosis of esophageal squamous cell carcinoma. *Pathology & Oncology Research* **27**, 632531 (2021).
207. Gao, C.L., Wang, G.W., Yang, G.Q., Yang, H. & Zhuang, L. Karyopherin subunit-α 2 expression accelerates cell cycle progression by upregulating CCNB2 and CDK1 in hepatocellular carcinoma. *Oncology letters* **15**, 2815-2820 (2018).
208. Reis, A. & Hermanson, O. The DNA glycosylases OGG1 and NEIL3 influence differentiation potential, proliferation, and senescence-associated signs in neural stem cells. *Biochemical and biophysical research communications* **423**, 621-626 (2012).
209. Niess, H.*, et al.* Side population cells of pancreatic cancer show characteristics of cancer stem cells responsible for resistance and metastasis. *Targeted oncology* **10**, 215-227 (2015).
210. Wu, C. & Alman, B.A. Side population cells in human cancers. *Cancer letters* **268**, 1-9 (2008).
211. Zhan, H.-x., Xu, J.-w., Wu, D., Zhang, T.-p. & Hu, S.-y. Pancreatic cancer stem cells: new insight into a stubborn disease. *Cancer letters* **357**, 429-437 (2015).
212. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188 (2001).
213. Castro-Conde, I. & de Uña-Álvarez, J. sgof: An R Package for Multiple Testing Problems. *R Journal* **6**(2014).

214. Rezwani, M., Pourfathollah, A.A. & Noorbakhsh, F. rbioapi: user-friendly R interface to biologic web services' API. *Bioinformatics* **38**, 2952-2953 (2022).
215. Noble, W.S. How does multiple testing correction work? *Nature biotechnology* **27**, 1135-1137 (2009).
216. Weisstein, E.W. Bonferroni correction. *https://mathworld. wolfram. com/* (2004).
217. Wilcoxon, F. Individual comparisons by ranking methods. in *Breakthroughs in statistics* 196-202 (Springer, 1992).
218. Johnson, N.L., Kotz, S. & Kemp, A.W. *Univariate discrete distributions*, (John Wiley & Sons, 2005).
219. Plaisier, S.B., Taschereau, R., Wong, J.A. & Graeber, T.G. Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic acids research* **38**, e169-e169 (2010).
220. Legendre, A.M. *Théorie des nombres*, (Firmin Didot Frères, 1830).
221. Brunson, J.C. Ggalluvial: layered grammar for alluvial plots. *Journal of Open Source Software* **5**, 2017 (2020).
222. Van den Berge, K.*, et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications* **11**, 1-13 (2020).
223. Morgan, M., Obenchain, V., Lang, M., Thompson, R. & Turaga, N. BiocParallel: Bioconductor facilities for parallel evaluation. *R package version* **1**(2014).
224. Cabello-Aguilar, S.*, et al.* SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Research* **48**, e55-e55 (2020).
225. Jin, S.*, et al.* Inference and analysis of cell-cell communication using CellChat. *Nature communications* **12**, 1-20 (2021).
226. Gatchalian, J.*, et al.* A non-canonical BRD9-containing BAF chromatin remodeling complex regulates naive pluripotency in mouse embryonic stem cells. *Nature communications* **9**, 5139 (2018).
227. Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* **7**, giy083 (2018).
228. Kahroba, H., Shirmohamadi, M., Hejazi, M.S. & Samadi, N. The Role of Nrf2 signaling in cancer stem cells: From stemness and self-renewal to tumorigenesis and chemoresistance. *Life sciences* **239**, 116986 (2019).
229. Zhao, Y.*, et al.* Targeting cancer stem cells and their niche: perspectives for future therapeutic targets and strategies. in *Seminars in cancer biology*, Vol. 53 139-155 (Elsevier, 2018).
230. Du, H.*, et al.* CEACAM6 promotes cisplatin resistance in lung adenocarcinoma and is regulated by microRNA-146a and microRNA-26a. *Thoracic Cancer* **11**, 2473-2482 (2020).
231. Du Manoir, S.*, et al.* In high-grade ovarian carcinoma, platinum-sensitive tumor recurrence and acquired-resistance derive from quiescent residual cancer cells that overexpress CRYAB, CEACAM6, and SOX2. *The Journal of Pathology* **257**, 367-378 (2022).
232. Liu, X.*, et al.* AURKA induces EMT by regulating histone modification through Wnt/β-catenin and PI3K/Akt signaling pathway in gastric cancer. *Oncotarget* **7**, 33152 (2016).
233. Wang, B.*, et al.* TOP2A Promotes Cell Migration, Invasion and Epithelial–Mesenchymal Transition in Cervical Cancer via Activating the PI3K/AKT Signaling. *Cancer Management and Research* **12**, 3807 (2020).
234. Xia, C.*, et al.* Abnormal spindle-like microcephaly-associated protein enhances cell invasion through Wnt/β-catenin-dependent regulation of epithelial-mesenchymal transition in non-small cell lung cancer cells. *Journal of Thoracic Disease* **13**, 2460 (2021).
235. Whately, K.M.*, et al.* Nuclear Aurora-A kinase-induced hypoxia signaling drives early dissemination and metastasis in breast cancer: implications for detection of metastatic tumors. *Oncogene* **40**, 5651-5664 (2021).

236. Hu, G.*, et al.* Intrinsic gemcitabine resistance in a novel pancreatic cancer cell line is associated with cancer stem cell-like phenotype. *International journal of oncology* **40**, 798-806 (2012).
237. Robin, F.*, et al.* Molecular profiling of stroma highlights stratifin as a novel biomarker of poor prognosis in pancreatic ductal adenocarcinoma. *British journal of cancer* **123**, 72-80 (2020).
238. Zhang, M.*, et al.* Single-cell transcriptomic architecture and intercellular crosstalk of human intrahepatic cholangiocarcinoma. *Journal of hepatology* **73**, 1118-1130 (2020).
239. Kosanam, H.*, et al.* Laminin, gamma 2 (LAMC2): a promising new putative pancreatic cancer biomarker identified by proteomic analysis of pancreatic adenocarcinoma tissues. *Molecular & Cellular Proteomics* **12**, 2820-2832 (2013).
240. Schuetz, C.S.*, et al.* Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. *Cancer research* **66**, 5278-5286 (2006).
241. Dai, S.*, et al.* HNRNPA2B1 regulates the epithelial–mesenchymal transition in pancreatic cancer cells through the ERK/snail signalling pathway. *Cancer cell international* **17**, 1-13 (2017).
242. Takahashi, M.*, et al.* Roles of the mesenchymal stromal/stem cell marker Meflin/Islr in cancer fibrosis. *Frontiers in Cell and Developmental Biology*, 2687 (2021).
243. Miele, E.*, et al.* The histone methyltransferase EZH2 as a druggable target in SHH medulloblastoma cancer stem cells. *Oncotarget* **8**, 68557 (2017).
244. Jin, B.*, et al.* Verification of EZH2 as a druggable target in metastatic uveal melanoma. *Molecular cancer* **19**, 1-15 (2020).
245. Zhu, J.*, et al.* Coexpression analysis of the EZH2 gene using the cancer genome atlas and oncomine databases identifies coexpressed genes involved in biological networks in breast cancer, glioblastoma, and prostate cancer. *Medical Science Monitor* **26**(2020).
246. Yamada, S.*, et al.* Cisplatin resistance driver claspin is a target for immunotherapy in urothelial carcinoma. *Cancer Immunology, Immunotherapy*, 1-9 (2023).
247. Qi, G.*, et al.* CDCA8, targeted by MYBL2, promotes malignant progression and olaparib insensitivity in ovarian cancer. *American journal of cancer research* **11**, 389 (2021).
248. Zheng, X.*, et al.* TOP2A is a potential biomarker and promotes platinum drug resistance in ovarian cancer: a bioinformatics and experimental analysis. (2022).
249. Zhan, Y.*, et al.* Inhibiting RRM2 to enhance the anticancer activity of chemotherapy. *Biomedicine & Pharmacotherapy* **133**, 110996 (2021).
250. Sato, Y., Tomita, M., Soga, T., Ochiai, A. & Makinoshima, H. Upregulation of thymidylate synthase induces pemetrexed resistance in malignant pleural mesothelioma. *Frontiers in pharmacology* **12**, 718675 (2021).
251. Xiong, Y.*, et al.* UBE2C functions as a potential oncogene by enhancing cell proliferation, migration, invasion, and drug resistance in hepatocellular carcinoma cells. *Bioscience reports* **39**(2019).
252. Zhu, Y.*, et al.* Inhibition of CDK1 reverses the resistance of 5-Fu in colorectal cancer. *Cancer management and research* **12**, 11271 (2020).
253. Yang, G.*, et al.* CDC20 promotes the progression of hepatocellular carcinoma by regulating epithelial-mesenchymal transition. *Molecular medicine reports* **24**, 1-10 (2021).
254. Zhang, Y.*, et al.* NEK2 promotes hepatocellular carcinoma migration and invasion through modulation of the epithelial-mesenchymal transition. *Oncology reports* **39**, 1023-1033 (2018).
255. Zhang, Q., Zeng, Z., Xie, W. & Zeng, Z. Highly Expressed SPC25 Promotes the Stemness, Proliferation and EMT of Oral Squamous Cell Carcinoma Cells via Modulating the TGF-β Signaling Pathway. *Journal of Hard Tissue Biology* **31**, 195-204 (2022).

256. Smith, A.G. & Macleod, K.F. Autophagy, cancer stem cells and drug resistance. *The Journal of pathology* **247**, 708-718 (2019).
257. Du, Y.*, et al.* GDF15 negatively regulates chemosensitivity via TGFBR2-AKT pathway-dependent metabolism in esophageal squamous cell carcinoma. *Frontiers of Medicine*, 1-13 (2022).
258. Liu, J.*, et al.* Down-regulation of GADD45A enhances chemosensitivity in melanoma. *Scientific reports* **8**, 1-11 (2018).
259. Xing, Y.*, et al.* TNFAIP8 promotes the proliferation and cisplatin chemoresistance of non-small cell lung cancer through MDM2/p53 pathway. *Cell communication and signaling* **16**, 1-15 (2018).
260. Hu, K.*, et al.* The novel roles of virus infection-associated gene CDKN1A in chemoresistance and immune infiltration of glioblastoma. *Aging (Albany NY)* **13**, 6662 (2021).
261. Yang, Q.*, et al.* The Functional Role and Regulatory Mechanism of Bromodomain-Containing Protein 9 in Human Uterine Leiomyosarcoma. *Cells* **11**, 2160 (2022).
262. Oshi, M.*, et al.* G2M checkpoint pathway alone is associated with drug response and survival among cell proliferation-related pathways in pancreatic cancer. *American journal of cancer research* **11**, 3070 (2021).
263. Tsai, S.-C.*, et al.* Histone deacetylase interacts directly with DNA topoisomerase II. *Nature genetics* **26**, 349-353 (2000).
264. Alpsoy, A.*, et al.* BRD9 is a critical regulator of androgen receptor signaling and prostate cancer progression. *Cancer research* **81**, 820-833 (2021).
265. Zhou, Q.*, et al.* The bromodomain containing protein BRD-9 orchestrates RAD51–RAD54 complex formation and regulates homologous recombination-mediated repair. *Nature communications* **11**, 2639 (2020).
266. Oshi, M.*, et al.* Development of a novel BRCAness score that predicts response to PARP inhibitors. *Biomarker Research* **10**, 1-11 (2022).
267. Willers, H., Pfäffle, H.N. & Zou, L. Targeting homologous recombination repair in cancer. in *DNA Repair in Cancer Therapy* 119-160 (Elsevier, 2012).
268. Campbell, P.A. & Rudnicki, M.A. Oct4 interaction with Hmgb2 regulates Akt signaling and pluripotency. *Stem Cells* **31**, 1107-1120 (2013).
269. Ma, Y.-S.*, et al.* Long non-coding RNA NORAD promotes pancreatic cancer stem cell proliferation and self-renewal by blocking microRNA-202-5p-mediated ANP32E inhibition. *Journal of Translational Medicine* **19**, 1-14 (2021).
270. Deng, S., Han, T., Huang, Q., Lu, J. & Yu, Z. CENPW as a biological indicator: predicting prognosis and guiding treatment in a patient with Kidney Renal Clear Cell Carcinoma. (2023).
271. Ye, C.*, et al.* Hypoxia-induced HMGB1 promotes glioma stem cells self-renewal and tumorigenicity via RAGE. *Iscience* **25**, 104872 (2022).
272. Yuan, C.*, et al.* POU2F2-IL-31 Autoregulatory Circuit Converts Hepatocytes into the Origin Cells of Hepatocellular Carcinoma. *Advanced Science* **8**, 2004683 (2021).