Investigation of the tumour microenvironment of prostate cancer

Stefano Mangiola
orcid.org/0000-0001-7474-836X

PhD - Dec 2018

Medicine, Dentistry & Health Sciences
The University of Melbourne

Submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy

Declaration

I Stefano Mangiola declare that:

(i) The present thesis comprises only my original work towards the PhD except where indicated in the preface
(ii) Due acknowledgement has been made in the text to all other material used; and
(iii) The thesis is 34,046 words long as approved by the Research Higher Degrees Committee.

Sincerely,

Acknowledgements

I thank my whole close and extended family and the countless people that has invested any time to support my work. My work was done with the hope of improving directly of indirectly the life of at least one patient suffering from prostate cancer, or other forms of cancer.

Table of contents

CHAPTER 1

Context

One in seven men in Australia is at risk of developing prostate cancer before the age of 75. This disease is a leading cause of male death worldwide, with a mortality rate of 62 men per 100,000. Treatments for prostate cancer exist, including surgery, radiotherapy and androgen deprivation therapy. However, the results achieved by the combination of these therapies can lead to variable outcomes, mainly due to the genetic heterogeneity of tumour cells and/or emergence of resistance. In contrast to the cancer cells, the non-cancerous portion of the tumour microenvironment, such as immune cells, fibroblasts and endothelial cells, is a genetically stable target that has a key role in cancer development. Improving our knowledge of the genetic and molecular interactions existing between cancer cells and other non-cancerous cell populations, both in primary or metastatic prostate cancer will provide new key insights in the biology of the disease and give new treatment opportunity. Here, we review the landscape of molecular interactions between cancerous and non-cancerous cells within the prostate tumoral mass.

The prostate tumour microenvironment

Within the primary prostate tumoural mass, cancer cells modulate the tumour microenvironment (TME)—the union of the extracellular matrix and diverse range of cell types, including immune and stromal cells, such as fibroblasts, endothelial cells, T- and B- cells and macrophages—resulting in more favourable conditions for tumour development. In doing this, cancer cells activate four key hallmarks: (i) cell cycle perturbation and proliferation; (ii) cell mobility, including immune cell infiltration and immune suppression; (iii) epithelial to mesenchymal transition; and (vi) angiogenesis. The cellular interactions within the TME at the molecular level is complex, involving the synergistic action of multiple cell types creating positive cross-cellular stimulatory loops, that ultimately support cancer progression (Fig. 1.1).

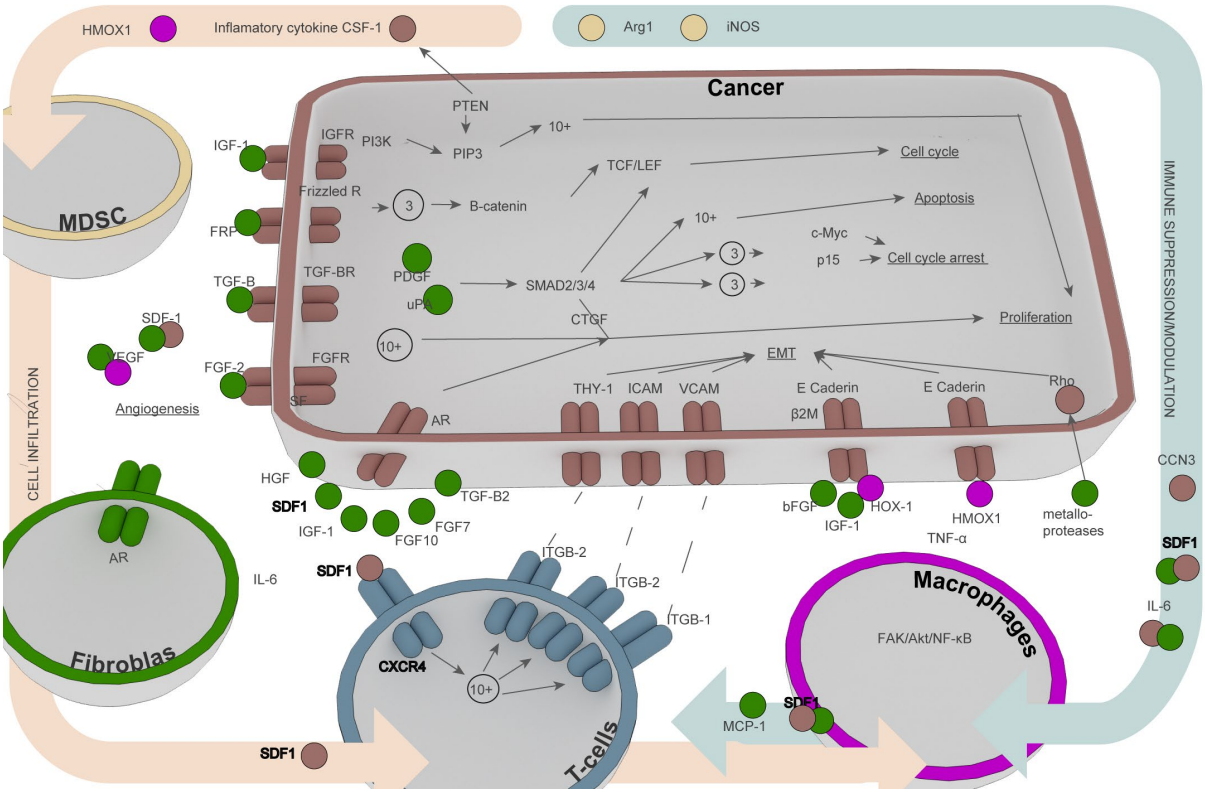


Fig. 1: Landscape of the major intercellular interactions in the prostate primary tumour microenvironment at the molecular level. Circles represent ligands coloured accordingly to the cell of provenance. Arrows represent genetic/physical gene interactions. Dashed lines represent physical receptor interactions.

Cell cycle perturbation and proliferation

The microenvironmental contribution to cancer cell proliferation has principally been linked to the activity of cancer associated fibroblasts[1]. Several proteins secreted by cancer associated fibroblasts promote cell proliferation via nine leading molecular mechanisms (Fig. 1.1), of which TGF-β and androgen are key axes, being part of 3 alternative pathways each. The molecule TGF-β is mainly secreted by cancer associated fibroblasts[2], immune cells[3] and cancer cells, and binds to its receptor on cancer cells acting as enhancer of cell proliferation. The transduction of the TGF-β signalling from the receptor to the nucleus is driven by three redundant axes: (i) SMAD3/4 and CTGF[4–6]; (ii) SMAD3/4, c-Myc and p15 that directly alters the cell cycle arrest mechanism[7]; and (iii) by SMAD3/4 linked to a longer downstream transduction chain (>10 molecules)[8]. Another important pathway of cancer cell proliferation is initiated by the frizzled-related proteins (FRP) secreted by cancer associated fibroblasts, that indirectly enhances the cancer cell cycle via a pathway including the frizzled receptor, B-cathenin and TCF/LEF[7]. Cancer associated fibroblasts are also secretors of insulin growth factor (IGF)-1 that drive two complementary kinase pathways in cancer cells. The two axes include (i) PIP3[9,10], and (ii) FGF-2 that interacts with ERK signalling[11–13]. Androgens, besides having both a direct role in cancer cell progression and evolution [14–16], also stimulate the secretory activity of cancer associated fibroblasts for growth factors such as IGF-1, stromal cell derived factor (SDF)-1, hepatocyte growth factor (HGF), transforming growth factor (TGF)-β2 and fibroblasts growth factor (FGF)-7 and FGF-10[17]. The latter molecule, which is regulated by TGF-β, forms a positive feedback loop increasing androgen receptor in the epithelium[18]. Macrophages, similarly to associated fibroblasts, have a major role in tumour development for prostate cancer (as well as for several cancer types)[9,19–21]. For example, in the primary prostate tumour cancer associated macrophages contribute to cancer cell proliferation through E-cadherin[22] via the secretion of heme oxygenase-1.

Cell mobility, immune cell infiltration and immune suppression

The enhanced infiltration of immune cells at the tumour site, and the alteration of their phenotype (=modulation) in favour of a more anti-inflammatory and wound-healing-like role, result in a favourable environment for the proliferation of cancerous cells. The infiltration of immune cells in the primary prostate TME is promoted above all by a range of molecules secreted by cancer cells, cancer associated fibroblasts and tumour associated macrophages (Fig. 1.1). For example,

MCP-1 secreted by cancer cells and SDF-1 secreted by both cancer cells and cancer associated fibroblasts increase the migration of monocytes into the TME [1]. The protein encoded by CSF-1 and other inflammatory cytokines secreted from cancer cells — with an increased rate if the PTEN gene is mutated [23] — stimulate immune cell infiltration that includes macrophages and myeloid-derived suppressor cells (MDSCs)[24–26]. An analogous role has the molecule Heme oxygenase-1 produced by tumour associated macrophages[27].

Immune cell modulation in the primary prostate TME is mostly carried out by cancer cells and myeloid-derived suppressor cells with the main target being T-cells and macrophages. For example, another important role of CSF-1 is in immune modulation, namely to lower the antigen presentation by macrophages and lower the anti-tumour T cell response [28]. Further modulation of leukocytes is enhanced by molecules Arg-1 and iNOS secreted mainly by macrophages[29,30]. The latter cell type is another that is commonly modulated by cancer cells in prostate cancer as well as many other tumours. The molecules CCN3 (secreted by cancer cells and myeloid-derived suppressor cells), IL-6, secreted by cancer cells[1], and the molecule SDF-1 secreted by cancer cells and cancer associated fibroblasts, act in concert to shift the macrophage phenotype from an inflammatory M1 to an anti-inflammatory M2 polarization, resulting in a growth factor rich environment[1].

Epithelial to mesenchymal transition

Epithelial to mesenchymal transition (EMT) is a hallmark of prostate cancer that is promoted by several microenvironmental processes, dominated by the role of cancer associated fibroblasts, cancer associated macrophages and T-cells (Fig. 1.1). Cancer associated fibroblasts boost epithelial to mesenchymal transition[31] via bFGF and IGF-1[32]; and indirectly via secreted metalloproteases, which stimulate the release of ROS from cancer cells (via Rac1b/cycloxygenase-2; COX-2)[33]. Also, tumour associated macrophages influence directly or indirectly the mobility properties of prostate cancer cells. On one hand they indirectly contribute to the modulation of fibroblasts to cancer associated fibroblasts, which are key to enhance the motility of prostate cancer cells[1]; on the other hand they directly support the epithelial to mesenchymal transition via the repression of E-cadherin with a molecular mechanism that is dependent on the present of Heme oxygenase-1 in the TME[22]. Not surprisingly, within cancer cells themselves, several pathways are strongly associated with the epithelial to mesenchymal transition, including the MYC signalling and stem-cell development pathways, as well as pathways regulated by NOTCH, FGFR and

WNT[34]. Overall, within the prostate primary TME, basal and luminal cells have distinct molecular profiles and functions, being neurogenic-like and stem-like respectively[35].

Angiogenesis

Blood supply, via the formation of new vessels (angiogenesis) within the tumoral mass, is essential for cancer growth[36]. The enhanced angiogenesis activity has been linked principally to cancer associated fibroblasts and macrophages, via the secretion of vascular endothelial growth factor (VEGF)[37,38] (Fig. 1.1). Cancer associated fibroblasts upregulate the production and secretion of VEGF via androgen stimulation, which is increased in the prostate TME[39]; whereas tumour associated macrophages enhance their secretion of VEGF via the AK/Akt/NF-κB pathway triggered by the cancer secreted molecule CCN3[40]. Furthermore, the oxidative metabolism of endothelial cells, controlled mainly by ADRB2 and E4ORF1, affects the adrenergic innervation, which promotes angiogenesis[41].

The metastatic tumour microenvironment

While the molecular interactions between cancer and benign cells within the primary TME could in principle also be present in the metastatic bone TME, the metastatic environment is characterised by specific processes, related to cell migration, tissue penetration and bone formation and resorption.

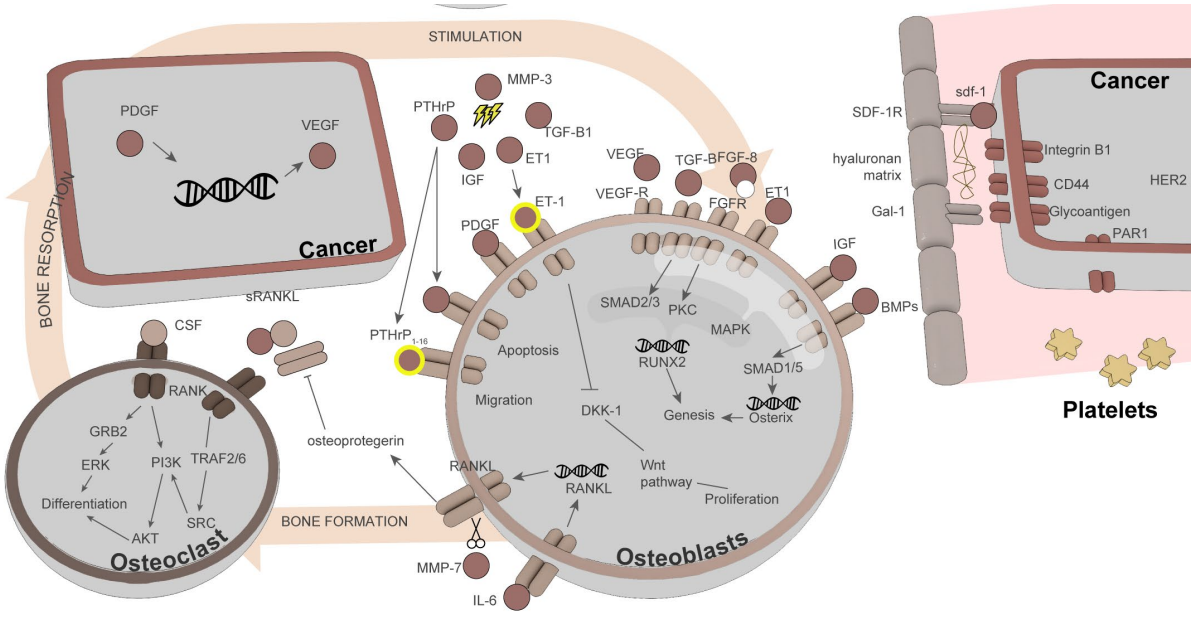


Figure 1.2: Representation of intercellular interactions proper of prostate metastatic tumour microenvironment. Circles represent ligands coloured accordingly to the cell of provenience. Arrows represent genetic/physical gene interactions. On the right-hand side a cancer cell entering from the bone marrow is represented. On the left-hand side is represented the bone marrow tumour microenvironment.

The migration to the bone metastatic site and the initiation of the "vicious cycle"

Metastatic cells that leave the primary tumour site migrate through the bloodstream (and often through lymphatic vessels) to reach distant organs and establish metastases. The bloodstream is a hostile environment for cancer cells. Within this migration phase the survival of migrating cancer cells is facilitated by the interactions with platelets that work as a physical protective barrier against immune cells; this interaction is promoted by the cancer transmembrane protein PAR1[42] (Fig. 1.2). In order to establish a niche in a distant organ, metastatic cells need to adhere and penetrate through the blood vessel wall; they achieve this interacting with endothelial cells via several axes, such as SDF-1/SDF-1R[43], integrin B1[44,45], CD44 to hyaluronan matrix[46], glycoantigen[45] and/or beta-galactosidase-binding-lectin-3 hyaluronan matrix [45]. Within the bone marrow, cancer cells trigger and maintain a positive stimulatory loop or "vicious cycle"[47], during which osteoclasts enhance the production of bone matrix and stimulate by consequence the activity of osteoclasts, which in turn stimulates osteoclasts. This cycle results in an enhanced turnover of the bone matrix. The benefit of such enhanced turnover for tumour development is the increased availability of growth factors, which, after having been injected into the bone matrix by osteoblasts during bone formation, become available when the bone matrix is resorbed by osteoclasts. Such cycles can result in either an overall increase of bone matrix formation (i.e., osteoblastic metastases), as often observed for prostate cancer[48]; or an overall reduction of bone matrix at the metastatic site (i.e., osteoclastic metastasis), as often observed for breast cancer[49].

Bone formation, osteoblasts migration, activation and proliferation.

The enhanced bone formation is a consequence of an increased migration, activation and proliferation of osteoblasts, driven by cancer cells and supported by other cell types. A key molecular signal that enhances osteoblast migration to the metastatic site includes the molecules PDGF[46] and PTHrP1-16. The availability of PTHrP1-16 depends on the presence of extracellular matrix proteases (e.g., MMP-3), which are produced and secreted by cancer cells them self and

whose function is to produce the final cleaved version of PTHrP[50] (Fig. 1.2). Cancer cells also modulate the activation of osteoblasts at the metastatic site via the secretion of VEGF, bone morphogenetic protein (BMP), TGF-β and several growth factors. The secreted molecules TGF-β[51], IGF, ET1 and FGF-8[52] lead to osteoblastogenesis via the phosphorylation of MAPK, and the upregulation of RUNX2. Cancer secreted BMP upregulates key transcription factors for osteoblastogenesis, such as osterix via SMAD1/5 and RUNX2 via MAPK. VEGF and BMP are part of a positive loop where BMP secreted by cancer stimulates the expression of VEGF in cancer cells themselves[53] (Fig. 1.2). Following their activation, the sustained proliferation of osteoblasts within the bone TME is promoted by cancer secreted molecules such as PDGF[46], and ET-1 that being activated by extracellular proteases inhibits the expression of DDK-1 in osteoblasts, allowing the Wnt pathway to enhance the proliferative mechanism[54]. The uncleaved molecule PTHrP can also act directly as an inhibitor of apoptosis for osteoblasts[55–57].

Bone resorption, osteoclast activation and proliferation

Bone matrix resorption is driven by osteoclasts that, in doing so, release a range of growth factors and signalling molecules sequestered in the bone matrix crystals. Such molecules sustain the vicious cycle and subsequently enhance cancer development. The molecule RANK is the key player in bone resorption; it is mainly produced by osteoblasts and prostate cancer cells[58–61] (Fig. 1.2). As a receptor, RANK is attached to the cell membrane; however, it can also be present in the extracellular matrix in its soluble form sRANK. Cancer cells enhance the availability of RANK in the TME via two principal mechanisms, the secretion of the pro-inflammatory cytokine IL-6 that increase RANK expression by osteoblasts[62], and the secretion of MMP-7 that solubilize RANK present on cell membrane[63] increasing the chance of reaching osteoclasts.

Therapeutic opportunities

Similarly, to several other tumour types, primary and metastatic prostate cancer are characterised by a complex microenvironment, where diverse cell types interact and synergise leading to tumour progression. A total of seven hallmarks of the disease progression have been discussed here, from tumour growth to metastasis. Several of these hallmarks represent treatment targets in other cancer types.

For example, immune therapy mainly focuses on targeting inhibitory mechanisms that cancer cells trigger to avoid immune cell anti-cancer activity. Inhibitors of PD-1 and CTLA-4

target cancer T-cells immune evasion and have shown high efficacy in eradicating several cancers types, such as melanoma. Current research is focused on understanding clinical and molecular profiles that are associated with immunotherapy efficacy. For example, prostate cancer does not benefit from such treatment, therefore further investigation is needed to understand whether an alternative mechanism for immune evasion and/or modulation exists. Epithelial to mesenchymal transition is being tackled by several experimental drugs across a diverse range of cancers [64]. The main molecular targets are Calcineurin, TGFBR1, EGFR the MF-kB pathway and SMAD molecules. All such treatments are at most in phase II, and so far, no clinical trial exists for targeting epithelial to mesenchymal transition in prostate cancer. Anti-angiogenic treatments are at an advanced stage, with more than ten compounds approved for clinical use [65]. Most of those compounds target tyrosine kinases, VEGF, VEGF receptors and the mTOR pathway. The cancer types that most benefit from those treatments are colorectal, lung, renal, brain hepatocellular and pancreatic. So far, no anti-angiogenic drugs have been approved for prostate cancer. No compounds tackling tumour cell migration through the circulation to metastatic sites, nor the development of bone metastasis are currently in advanced experimentation.

The lack of approved alternative treatments for prostate cancer tackling the biology of the disease at a systematic level represents a call for further investigation. The emergence of high-throughput cellular and molecular technologies represents an opportunity to gain a better understanding of the unique features of prostate cancer that cause the poor translation of treatments from other cancers.

The study of the tissue microenvironment: an evolving concept

The tissue microenvironment have been extensively studied through *in vitro* and *in vivo* experiments, such as migration assays [66] and xenograft mouse models [67] respectively. More recently, thanks to the development of high-throughput cellular and molecular technologies, it is possible to infer the biology at the molecular level for a wider range of cell types, for a given experiment. For example, fluorescence-activated cell sorting allows the isolation of several selected cell types from a tissue, which can be analysed at the molecular level with nucleotide sequencing or mass spectroscopy. In recent years, advances in machine learning techniques and the increased availability of nucleotide sequencing data for a wide range of cell types, has allowed the computational inference of both the cellular composition of tissues and molecular composition

of specific cell types within a tissue, from whole tissue gene expression data. This approach increases the throughput of cell types that can be studied in a given experiment and allows the integrated analysis of novel and public data sets. Furthermore, such analysis can be performed on data generated from fresh frozen samples, as opposed to FACS which requires processed tissue, and in vitro and *in vivo* experiments which rely on physical models of the tissue. Despite their attractiveness, these computational approaches present major challenges. For example, most computational methods strictly rely on prior information, that can be sparse and/or misrepresent the data.

The mathematics behind the modelling of the tissue microenvironment

The inference of (i) tissue composition, and (ii) transcriptional profiles of distinct cell types within a tissue can be described more precisely as (i) the proportions of each cell type within a tissue sample, and (ii) the true transcription rate of each gene for each cell type (theoretically observable only if each cell type was purified and unchanged by tissue processing). For each gene, the value we observe from the whole tissue is a combination of (i) and (ii) for each cell type. For one gene and one sample, this data structure can be framed mathematically as a weighted average (Eq. 1). That is, the tissue gene expression value y (observed) is equal to the weighted sum of specific gene expression for each cell type a (observed or inferred), weighted by its absolute proportion π within each tissue sample (inferred).

$$(1) \quad \mathrm{y} \approx a_1 * \pi_1 + a_2 * \pi_2 + \ldots + a_a * \pi_P$$

For multiple genes and multiple samples, this data structure can be framed mathematically as a system of linear equations (Eq. 2). That is, the observed matrix of gene transcription values Y (observed) is equal to the matrix of specific gene transcription for each cell type A (observed or inferred) multiplied by the matrix of absolute proportions of each cell types within each tissue sample Π (inferred).

$$(2) \quad \mathrm{Y}\begin{bmatrix} y_{g1,s1} & y_{g1,s2} & \cdots & y_{g1,sS} \\ y_{g2,s1} & y_{g2,s2} & \cdots & y_{g2,sS} \\ \cdots & \cdots & \cdots & \cdots \\ y_{gG,s1} & y_{gG,s2} & \cdots & y_{gG,sS} \end{bmatrix} \approx A\begin{bmatrix} a_{g1,p1} & a_{g1,p2} & \cdots & a_{g1,pP} \\ a_{g2,p1} & a_{g2,p2} & \cdots & a_{g2,pP} \\ \cdots & \cdots & \cdots & \cdots \\ a_{gG,p1} & a_{gG,p2} & \cdots & a_{gG,pP} \end{bmatrix} \times \Pi\begin{bmatrix} m_{p1,s1} & m_{p1,s2} & \cdots & m_{p1,sS} \\ m_{p2,s1} & m_{p2,s2} & \cdots & m_{p2,sS} \\ \cdots & \cdots & \cdots & \cdots \\ m_{pP,s1} & m_{pP,s2} & \cdots & m_{pP,sS} \end{bmatrix}$$

The subscript $g \in \{1,..,G\}$ represents genes, $s \in \{1,..,S\}$ represents samples and $p \in \{1,..,P\}$ represents cell type (populations). The selection of genes in this framework is done based on how well a gene can segregate one or more cell types from the rest; ideally, genes would be exclusively expressed/methylated in one cell type. Such genes are referred to as markers.

Algorithms that infer tissue composition and/or cell type specific transcriptional signatures differ from each other for the amount of prior information and the statistical framework used. In principle, observing Y (Eq. 2) for enough samples and marker genes, would be possible to infer both the matrices A and Π that best explain the data. In practice, considering the high tissue complexity and the data bottleneck, most algorithms use a priori information for A to confidently infer Π or vice versa. Information for A is usually gathered from public resources of gene transcription for pure/purified cell types, while the information on Π can be observed experimentally with flow cytometry or immunohistochemistry of different sections of the tissue sample, or related samples. The inference of the matrix A and/or Π is performed through a wide variety of statistical methods, including: (i) linear regression with least square noise model coupled with QR factorization optimization[68], or ε-insensitive noise model coupled with maximal margin optimization[69]; (ii) quadratic programming optimization[70]; (iii) mixed membership model[71]; or (iv) non-parametric gene enrichment[72].

Table 1.1. List of available algorithms for the inference of tissue composition and/or cell type specific transcriptional signatures. * refers to semi supervised methods.

Software	Input			Output		Method	Main Platform	Year of publication	References
	Prop	Sign	Mark	Prop	Sign				
***De novo* algorithms**									

MMAD*				+	+	Linear regression	MATLAB	2013	[73]
CAM				+	+	Scatter space	R	2016	[74]
mixture_estimation				+	+		R	2010	[75]
DeMix*				+	+	Nelder–Mead (maximum likelihood)	R	2013	[76]
Ab initio algorithms									
MMAD*	+	+		+	+	Linear regression	MATLAB	2013	[73]
DSA			+	+	+	Quadratic regression	R	2013	[70]
PERT		+		+		mixed membership model	Octave	2012	[71]
NNML		+		+		mixed membership model	MATLAB	2012	[71]
ISOLATE		+		+		mixed membership model	MATLAB	2009	[77]
DeconRNASeq	+	+		+*		Quadratic programming (non-negative decomposition)	R	2013	[78]
TEMT	+				+	Mixture model (MCMC)	Python	2013	[79]
ESTIMATE		+		+		Gene enrichment	R	2013	[80]
Cibersort		+		+		maximal margin	R	2015	[69]
Comics		+		+		Linear regression (RNA) + ANOVA (DNA)	R	2016	[81]
DeMix*		+	+	+	+	*Linear regression,* L_1 norm ML	R	2013	[76]
NA						Minimum ratio		2010	[75]
csSAM	+				+	Linear regression	–	2010	[82]
Zuckerman et al.				+	+	Linear regression	–	2013	[83]
MCP-counter		+		+		Linear regression	R	2016	[84]
ssGSEA		+		+		Gene enrichment	R	2016	[85]
xCell		+		+		Gene enrichment	R	2017	[72]
TIminer (pipeline)		+		+		Gene enrichment	Bash	2017	[86]
TIMER		+		+		Linear regression	Web	2016	[87]
EPIC		+		+		Linear regression	R	2017	[88]
quanTIseq (pipeline)		+		+		Linear regression	Bash	2017	[89]
ssKL		+		+		NA	R	2014	[90]
ssFrobenius		+		+*		NA	R	2014	[90]

In the next section we introduce these methods grouping them by being *ab initio* or *de novo*, and subgrouping them by the statistical approach used.

Ab initio methods

Methods that use a priori information can either infer the tissue composition using transcriptomic cell type specific signatures, or vice versa. For complex tissues, where a high number of cell types including both distinct and highly similar cell populations are present, some a priori information is necessary to resolve the tissue mixture. Such an approach carries the risk of integrating information from public sources that is not representative of the query tissue.

Inference of tissue composition

Several statistical methods have been implemented for the inference of tissue composition using a priori information, including linear or quadratic regression with a diverse range of noise models and optimisation techniques, mixed membership model approaches or nonparametric gene enrichment.

Least square error linear regression through QR factorization

A multiple linear regression can be employed to solve the linear equation system Eq. 2. The noise is modelled with the square of the error distances (pairwise for each gene and sample) between the inferred $\hat{Y}$ and the observed Y (Eq. 2). Least square optimization was first applied for identifying cellularity patterns on simulated data[68] and on gene expression data from blood for the systemic lupus erythematosus disease[91]; using the nnls function (from Matlab) or the lsfit function (from R) respectively. The algorithm MMAD[73], predicts Π minimizing the least square error using a non-linear gradient descent algorithm with non-negative constraint. Similarly, the CsSAM algorithm, predicts cell type proportions using least square regression. The method of Zuckerman et al. predicts cell-type proportions (Π) using an external reference signatures just as initialization source. First non-negative matrix factorization is used to estimate an intermediate matrices $\hat{A}$ and $\hat{\Pi}$ initializing $\hat{A}$ from the reference signatures. Second, the most likely set of cell types present in the mix is estimated (given a maximum number of cell types provided) using a method called SKLD[92], and a surrogate transcription signature matrix $\hat{\hat{A}}$ is composed. Third, non-negative least square is used to estimate Π from Y and $\hat{\hat{A}}$. DeMix utilizes a list of marker genes and/or a set of

reference background samples (e.g., benign component within the tumour mass) as a priori information and infers the proportion of the two major components (e.g., cancer and benign). In case marker genes are not defined, this algorithm uses Nelder–Mead procedure to identify an optimum gene set. The inference of two-component proportion is based on a custom optimization algorithm that evaluates the difference between the observed and predicted data.

E-insensitive loss linear regression through Maximal margin optimization

The algorithm Cibersort [69] applies a robust linear regression based on ε-insensitive loss function (i.e., svm R function using a linear kernel). The ε-insensitive loss function allows zero loss within the area surrounding the regression line and a quadratic loss beyond that area. This loss function confers more robustness against noise from non-informative genes. Furthermore, svm implements ridge regression to increase robustness for autocorrelated cell types. Cibersort also provides a curated pre-compiled signature data set including the most informative genes useful to separate 22 distinct immune cell types[69].

Linear regression through quadratic programming

Quadratic programming can be used to solve the linear equation system Eq. 2 with positive constraint. Formally, a quadratic programming solver is defined as an algorithm that finds a fit for a quadratic optimisation function applying linear constraints. The main difference to other linear regression solvers is the possibility to apply a positive constraint to the inferred proportions. Two algorithms make use of quadratic programming for gene expression deconvolution. The algorithm DeconRNAseq is RNA-seq dedicated[78], and uses the quadratic programming algorithm limSolve (from R). The algorithm DSA[70] only uses a list of marker genes for each selected cell type as prior information. DSA first predicts $\hat{\Pi}$ using $\hat{A}$, a surrogate of the matrix A (i.e., a diagonal matrix, including marker genes), it then predicts A from $\hat{\Pi}$, and subsequently Π with the inferred gene signatures. It uses the R quadprog solver [93].

Mixed membership model through Markov Chain Monte Carlo

A mixed membership model can be used for inferring proportion parameters in the system of linear equations Eq. 2. A series of algorithms based on latent Dirichlet allocation [94] (LDA) have been implemented: ISOLATE[77], NNML[71] and PERT[71]. Latent Dirichlet allocation can be represented as a product conditional probabilities. Both Π and A are modelled here as proportions (as arising from a Dirichlet distribution) of each cell type of being present in each sample, and of each gene

expressed in each cell type, respectively. The algorithm ISOLATE implements latent Dirichlet allocation in its simplest form, while the algorithm NNML introduces three further parameters that allows the modelling of an unknown/background source of RNA, beside the signatures provided, which includes all those cell types for which a signature is not available. The algorithm PERT expands on ISOLATE with three further parameters that allow to model the artefactual differences between reference signature and fresh hidden signatures within the mix, caused by the processing needed to obtain such purified signatures (e.g., cell sorting).

Gene enrichment through non-parametric statistics

Some methods do not attempt to solve directly the system of linear equations Eq. 2, but rather use a publicly available non-parametric algorithm (e.g., GSEA[95]) for calculating a score associated to the likelihood of a set of marker genes that define a cell type of being in enriched in the top or bottom rank in the tissue gene expression data set. For example, ESTIMATE uses ssGSEA[96] to calculate enrichment scores for stromal tissue and immune cell infiltration. The tumour purity score is integrated using the ABSOLUTE-based[97] score. All three scores are integrated into a formula identified using Eureqa Formulize[98]. The algorithm xCell[72] uses GSEA enrichment score that is first transformed to a linear scale and then adjusted using a spill over compensation technique trained on *in silico* mixtures, with the goal of reducing autocorrelation biases among similar cell types.

Inference of cell type specific transcriptome

As opposed to the inference of tissue composition, a limited number of methods have been implemented for the *ab initio* inference of cell type specific transcriptomes from whole tissue nucleotide information.

Least square error linear regression

The algorithm MMAD, in its alternative setting, is able to infer cell specific transcriptional profiles if Π is known[73]. The reference profiles are calculated independently for each gene, minimizing the least square error of cell types that have similar expression for the selected gene, using a k-means mixed membership model. The most conservative mixed membership model is estimated using AIC parsimony criterion.

Linear regression through quadratic programming

The algorithm DSA[70] estimates the matrix Π and A given a list of marker genes for each selected cell type. Briefly, DSA first predicts Π using surrogate of the matrix A (i.e., a diagonal matrix, including marker genes); then, pure signatures of cell types (A) are calculated using the predicted Π using the algorithm quadprog (from R)[93].

De novo methods

Several algorithms predict both cell type proportions (Π) and cell type specific signatures (A) for a selected/inferred number of cell types within a tissue without any pre-existing information. The main advantages of such approaches are avoiding the bias generated by using partially representative, publicly available gene marker signatures for the selected cell types; and potentially discover novel cell types in cases where the identity of cell markers is inferred. These methods use a diverse range of statistics of which least square linear regression is the most popular, although a range of statistical tools, such as multidimensional-corner identification and AIC are used in integration.

Least square error linear regression

For the algorithm CAM[74] genes are plotted in a "N" dimensional space depending on their expression values, with N being the number of given mixed samples; for example three samples genes would be plotted in a 3D space forming a gene "cloud". The CAM algorithm uses Minimum description length (MDL) to identify "corners" in the cloud that will include marker genes for the most represented cell-types. Then, Π is estimated solving the system of linear equation, using the predicted marker genes to build a surrogate of A (i.e., a diagonal matrix, including marker genes). The matrix Π is then used to predict A using lsfit function (from R). The algorithm MMAD in its *de novo* setting, attempts to infer both A and Π without a priori information. It uses a maximum likelihood approach to minimize sum of square error between the inferred mixed value for each gene and the observed mixed value. DeMix is designed for a two-ways deconvolution (e.g., cancer versus benign component) and uses a maximum likelihood approach to solve the two components of the linear system Eq. 2. An optimal set of tumour purities for every sample is derived using the Nelder-Mean procedure[99], based on global properties (mean and standard deviation) of the mixed tumour samples; then those proportions are used to deconvolve the mixed expression samples into their pure components according to the maximum likelihood of the combined probability of (i)

observing each gene expression for pre tumour considered global properties of pure tumour gene expression, and (ii) the estimated pure normal gene expression considered global properties of pure normal gene expression.

Minimum ratio

The algorithm Mixture_estimation[75] infers the expression signatures and proportions of a two-components transcriptional mix. The two-component proportions are estimated using a minimum ratio paradigm; defining the ratio for each gene as being the gene expression in the mix divided by the pure gene expression. When the ratios are calculated for a gene across all possible proportion values (between 0 and 1), they form an unimodal increasing curve; the proportion of the mix is found by identifying the point where the second derivative of the curve (where the trade-off between pulling low values up and low values down) is optimal.

Comparative analysis of the methods for transcriptional signature deconvolution

Considering that methods for infer cell type transcriptomic profiles from whole tissue data are not so well established, I will focus on the benchmarking of the deconvolution methods that infer tissue composition from whole tissue transcriptomic data, including *ab initio* and *de novo* implementations. In order to compare the accuracy and robustness of publicly available methods, the inferences of tissue composition were performed using a selection of published and novel cell type specific signature training data sets. This permits the assessment of the robustness of each method to diverse, a priori information. Those training data sets were: (i) the LM22 training transcriptomic signatures[69]; (ii) a recompiled LM22 (called LM22 redo) from the source signatures; (iii) an enriched LM22 with stromal signatures from epithelial, endothelial and fibroblast cells; (vi) a novel RNA sequencing data set that includes 28 cell types, and compiled integrating the BLUEPRINT[100], FANTOM[101] and ENCODE[102] databases; (v) a novel microarray based data set including 28 cell types compiled integrating the GSE86362[84] and LM22 data sets. In total, we tested each combination of algorithm and training signature across three validation data sets: a TCGA RNA sequencing based data set (named TCGA), a pure cell type RNA sequencing data set (named Pure), and a peripheral blood mononuclear cell (PBMC) microarray based data set[69] (named PBMC).

The TCGA data set is composed by 2352 samples from 19 major epithelial cancer types. The goal of the TCGA data set is to observe the agreement on tumour purity evaluation with a

selection of DNA based measures across cancer types[103], including: (i) ABSOLUTE score [97]; (ii) CPE[104], a consensus measurement of purity estimation ; and (iii) LUMP[105], based on methylation of immune-specific CpG sites. For the TCGA validation data set only the novel RNA sequencing and microarray-based training data sets were used as included epithelial cell transcriptional signatures. Figure 1.3 shows similar overall performances across training datasets and a diverse range of performances over cancer types for almost all algorithms. The combination of the algorithm Cibersort with RNA based transcriptional signature shows significantly better performances compared to the other combinations. As expected, the gain in performance of Cibersort declined when paired with the microarray-based signature. Comics and Decon are the second and third best performing algorithms on this validation data set.

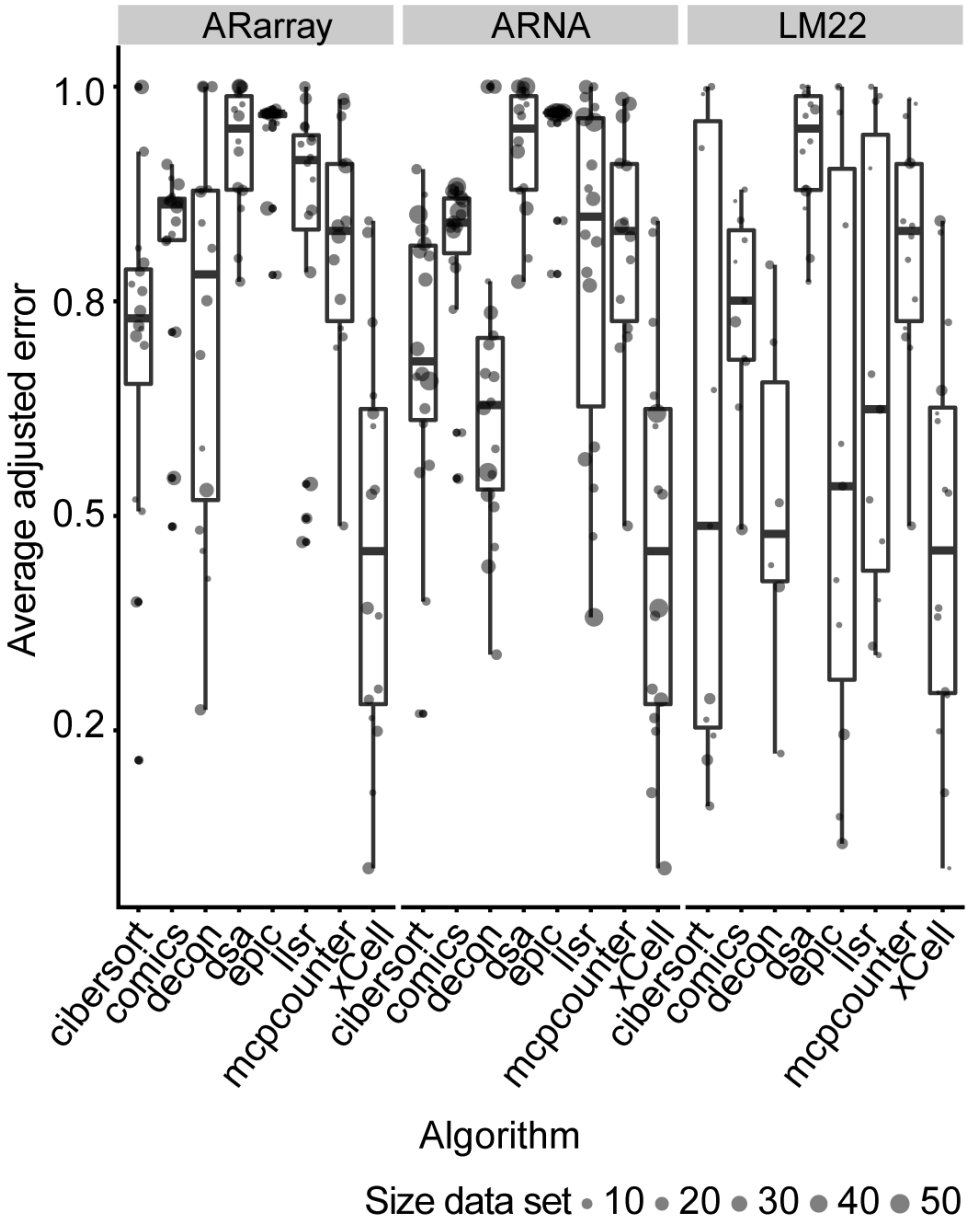


Fig. 1.3 Comparison of the adjusted inference error of various deconvolution algorithms for the dataset TCGA across selected training data sets. The size of the dots represents the number of samples present for each cancer type.

The Pure data set is composed by 190 samples from 20 pure cell types. The goal of the Pure cohort is the assessment of the accuracy in classifying correctly pure cell types. Figure 1.4 shows a highly diverse performances across algorithms, training data sets and cell types. Overall all algorithms show poor accuracy, except for xCell. The algorithm decon showed the second best

performance. Interestingly, considering that the validation data set is RNA-seq based, the training data set LM22 (based on microarray) gave higher performances across algorithms compared with the other RNA sequencing and microarray-based data sets.

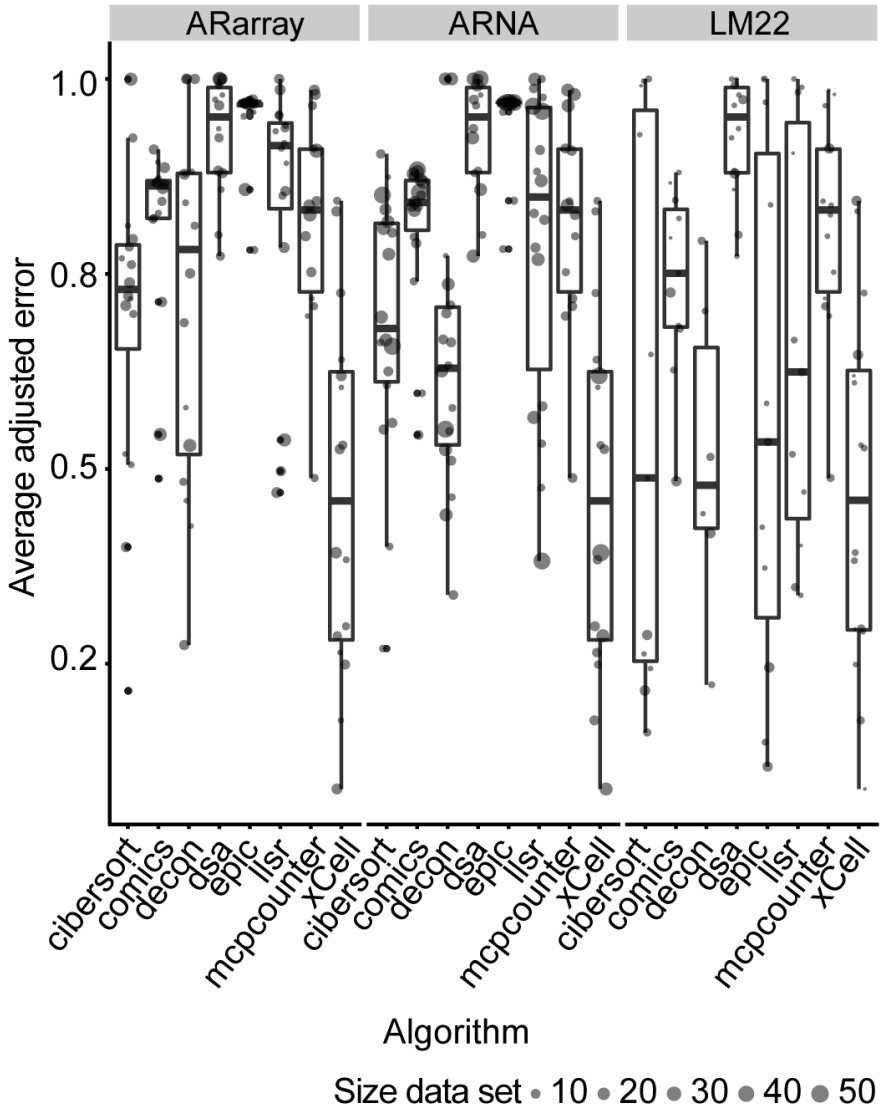


Fig. 1.4 Comparison of the adjusted inference error of various deconvolution algorithms for the dataset Pure across selected training data sets. The size of the dots represents the number of samples present for each cancer type.

The PBMC data set is composed by 20 samples for which 9 blood cell types were experimentally measured. The goal of the PBMC training data set is to evaluate the accuracy in

inferring tissue composition in complex tissues as well as to perform an independent assessment of this publicly available data set previously analysed[69]. For this validation data set all available training signatures were used for evaluation. Figure 1.5 shows an overall good accuracy across algorithms and training data sets. As expected, the combination of LM22 and Cibersort shows the highest accuracy amongst all. The data set LM22 gave the highest overall accuracy on this validation set. The accuracy advantage of Cibersort decreases with any other training data set used.

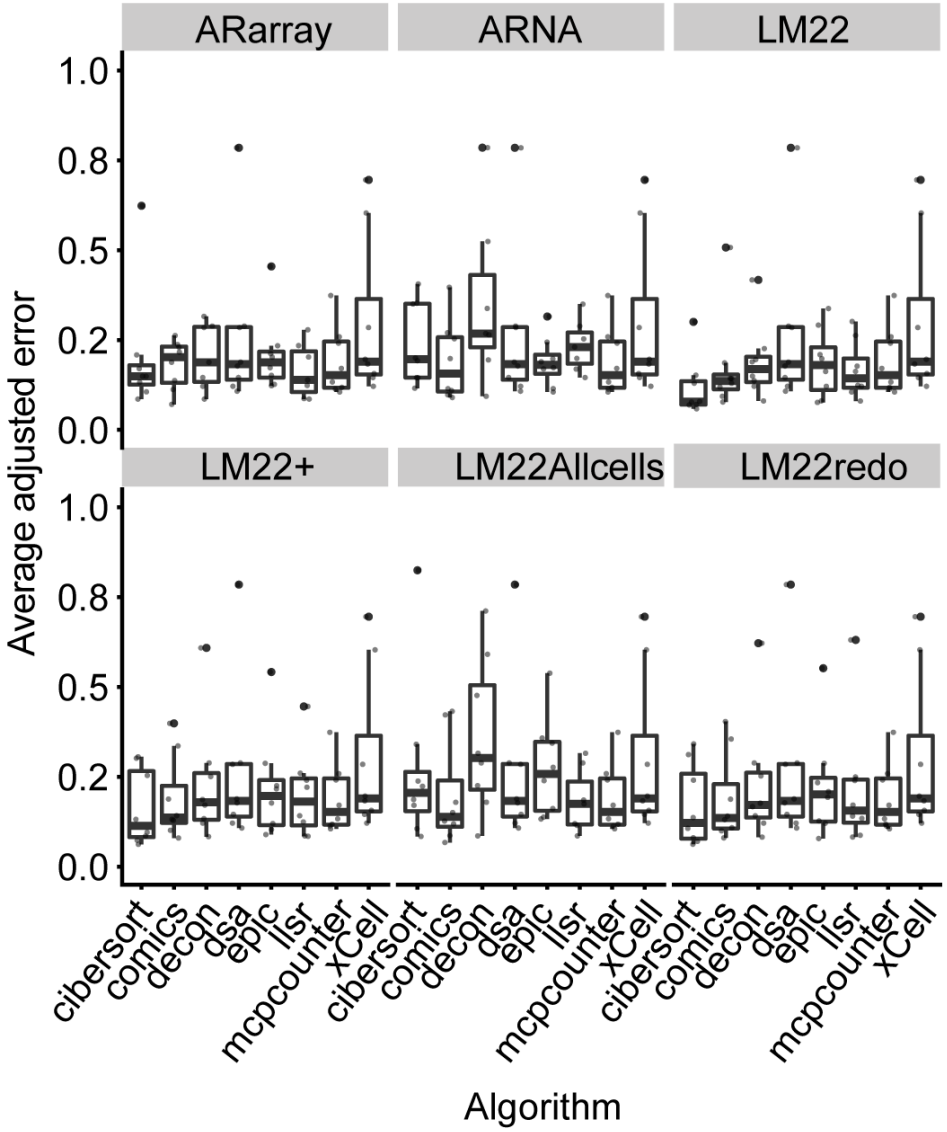


Fig. 1.5 Comparison of the adjusted inference error of various deconvolution algorithms for the dataset PBMC across selected training data sets. The size of the dots represents the number of samples present for each cancer type.

Figure 1.6 shows the overall variation in performances of all algorithms across training and validation data sets. Overall, Cibersort shower lower adjusted mean error rate (Eq. 3) compared to the other algorithms. However, for each validation data set we observed variability in the rank of most accurate algorithms. For example, xCell outperforms all algorithms for the classification task using the Pure validation data set, while Cibersort outperforms the other algorithms for the PBMC and TCGA validation data sets. Regardless of the absolute performances, the algorithms comics and mcpcounter showed the highest consistency in performances across training data sets.

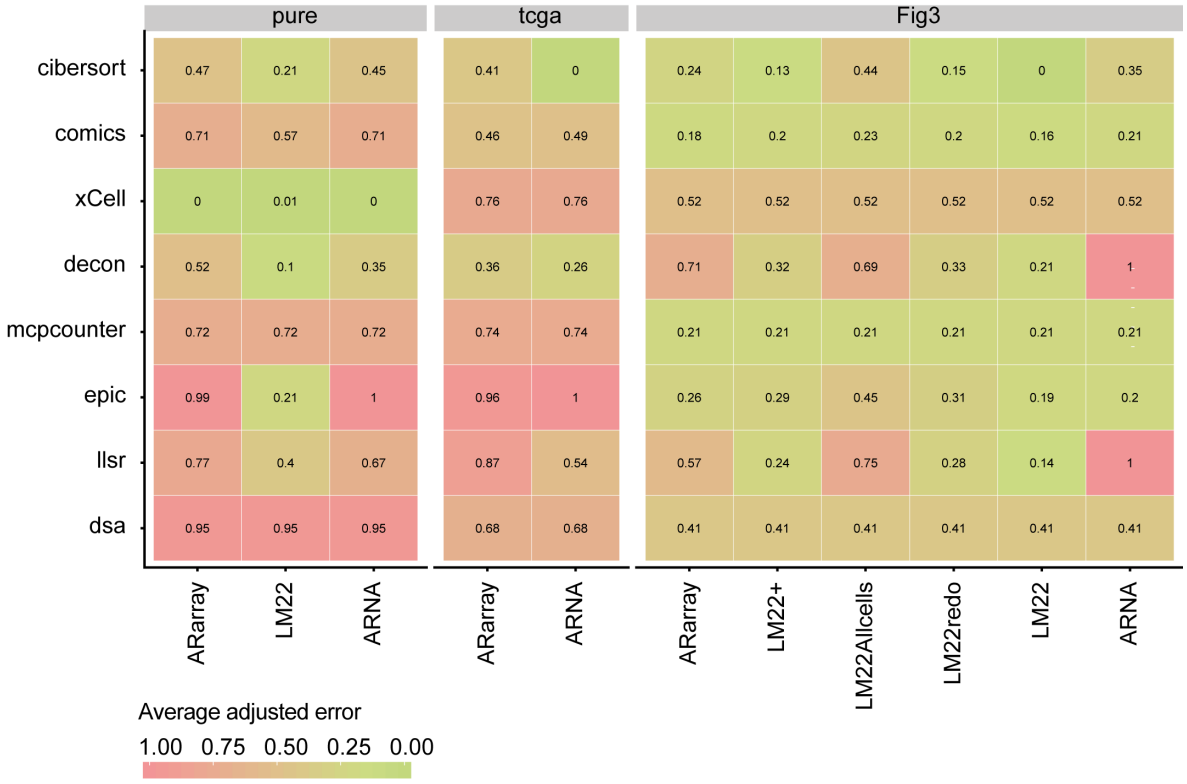


Fig. 1.6 Heatmap representing the overall performances of all algorithms across the three-validation data set, employing a diverse range of training gene expression signatures. The performance is expressed in term of adjusted absolute error (Eq. 3)

Methods

Compilation of the novel training data set signatures

For the RNA sequencing based signature, cell type specific transcriptomes were collected from three databases: BLUEPRINT[100], FANTOM5[101], and ENCODE[102]. All cell type identifiers were

harmonized to the following 28 cell types: epithelial, endothelial, fibroblast, adipocyte, eosinophil, neutrophil, monocyte, macrophage M0, macrophage M1, macrophage M2, dendritic resting, dendritic activated, mast cell activated, mast cell resting, b cell naive, b cell memory, t CD4 naive, t CD8, t helper 1, t helper 2, t helper follicular, t gamma delta, t reg, t memory central, t memory effector, NK activated, NK resting and plasma cell. All data sets were integrated based on gene symbols. Considering the large amount of samples, the integration was performed in a sparse manner, in order to avoid gene loss. That is, if a gene symbol was missing from a sample, its value was marked as missing data (i.e., NA). After harmonization, the data set was normalized using the TMM method[106]. In order to not discard genes in the normalization phase, the calculation of the normalization factor has first been performed on non-parse, filtered genes (> 0.5 count per million in at least 2/3 of the samples), and then applied to the whole data set. Then, the missing values for each gene were estimated using the closest cell type category according to the hierarchy of cell differentiation (Fig. S1.1). In order to remove unwanted variation between data sources, the algorithm RUV4[107] was employed using the highest level descendants of the hierarchy cut at level 2 as covariate of interest, and using a selection of 600 housekeeping genes[107] as negative controls. The number of unwanted covariates (i.e., parameter K in RUV4 algorithm) was arbitrarily chosen for parsimony as the double of the number of the integrated databases (e.g., k = 6 for RNA sequencing), in the absence of a more meaningful criteria. For the microarray based signature, cell type specific transcriptomes were collected from two data sets: LM22[69] and GSE86362[84]. An analogous methodology was used to integrate those data sets, except for the normalization step, which was based on quantile normalization.

Gene marker selection

The set of marker genes for each cell type was identified performing a global differential gene expression analysis across all other cell types, using edgeR[108] on the RNA sequencing based training data set. For each query cell type genes were shortlisted if having false discovery rate < 0.05 and fold change > 0, based on the edgeR top table; then, genes were ranked by fold change and the top 200 were shortlisted; then, genes were ranked based on the variance of the rest of cell types and the top 100 genes were shortlisted. From those genes, the ones that showed bimodality (i.e., that don't show heterogeneity within) for any cell type were discarded.

Calculation of the performance score

In order to create a performance score applicable to all validation data-sets, we created an error metrics based on the absolute distance between ground truth and prediction, adjusted proportionally to the abundance of a cell type. The rationale for not using R^2 and p-value from the linear regression of predicted proportions against ground truth is: (i) that this score only applies when the ground truth proportion is a continuous value, and does not apply to classification problems with binary proportion values (i.e., either 0 or 1); and (ii) because a two-values score is not suitable for creating a unique rank. The score is calculated with the following formula

$$(3) \quad \text{adjusted absolute error} = ||\pi_{predicted}\pi_{real}|| * (1 + ||logit(\pi_{real})||)$$

Where π is a cell type proportion in a sample, and where || || represents the L1 regularization. The first component on the left side of Eq. 3 represents the absolute error, while the second component represents a normalization constant. The normalization constant is needed because, if we assume that a proportion variable is Beta distributed (Fig. S1.2) the degrees of freedom are the largest at the point 0.5 and the lowest at the plateaus 0 and 1. The trend of change of degrees of freedom can be modelled with a logit function (Fig. S1.2). With such adjusted rate, the error for cell types with proportion close to 0 and 1 will obtain more weight proportionally to the absolute of the inverse of the logit function. The summation of 1 to the logit transformation (Eq. 3) have the function of producing a unitary baseline (multiplication by 1) for proportions of 0.5.

Conclusions

The problem of transcriptional deconvolution has been approached for 17 years (from 2001[68]). During this time, a diverse range of statistical paradigms has been adopted with success, and the improvement of cell-type specific transcriptomic signatures has allowed to infer the composition of complex tissues[69]. Overall, there is no a single algorithm that outperform others in all scenarios. Cibersort, xCell and Comics have overall better performances compared with the other tested algorithms; while Cibersort has superior performance on the PBMC validation data set and good performances on the other data sets, xCell have superior performances on the Pure data set.

Although premature, in recent years deconvolution methods based on methylation profiles have emerged[109]. In principle, the use of methylation cell-type-specific signatures offers more stability compared to transcriptional signatures as are less sensitive to extemporal biological changes and provide long-term gene regulatory profile for diverse cell types. For lack of direct comparative measures, such methodologies have been omitted from this work.

In seeking associations between tissue composition and biological/clinical properties/outcomes, molecular profiles deconvolution across a sample cohort represents the first step; a statistical model then needs to be applied to the inferred cell-type proportions to infer an association between the abundance of a cell type and biological/clinical properties/outcomes. Compared to differential transcription analyses, working with tissue composition has a further challenge: while the inference of gene counts carries low uncertainty (considering the specificity of nucleotide gene sequences and the sequenced read length of modern sequencing techniques), the inference of tissue composition carries a much bigger uncertainty. Therefore, an integrated algorithm that performs tissue deconvolution and regression and/or hypothesis testing is necessary, to be able to transfer the uncertainty from the first inference phase to the second.

Supplementary figures

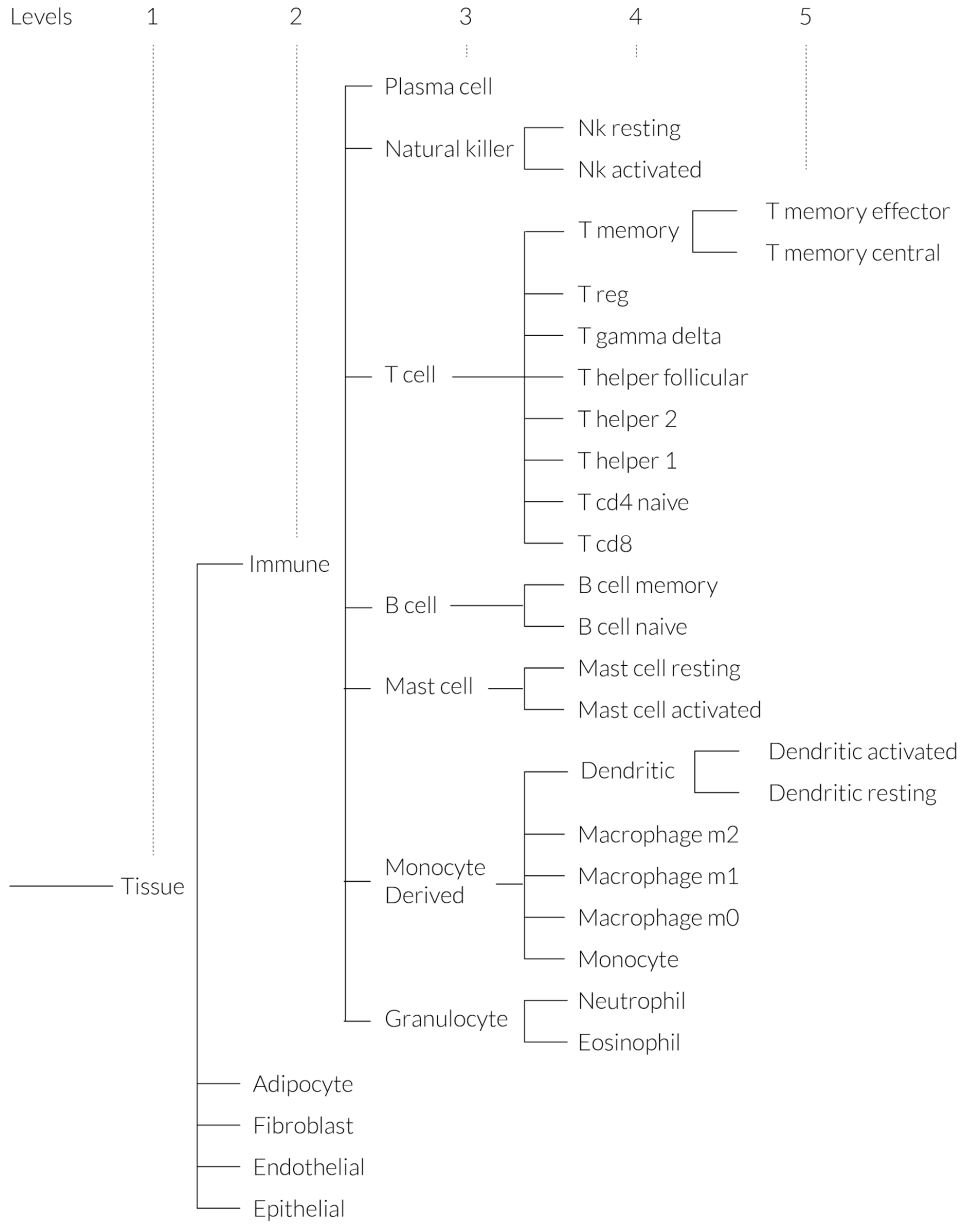


Fig. S1.1 Graph representation of the cell type hierarchical structure.

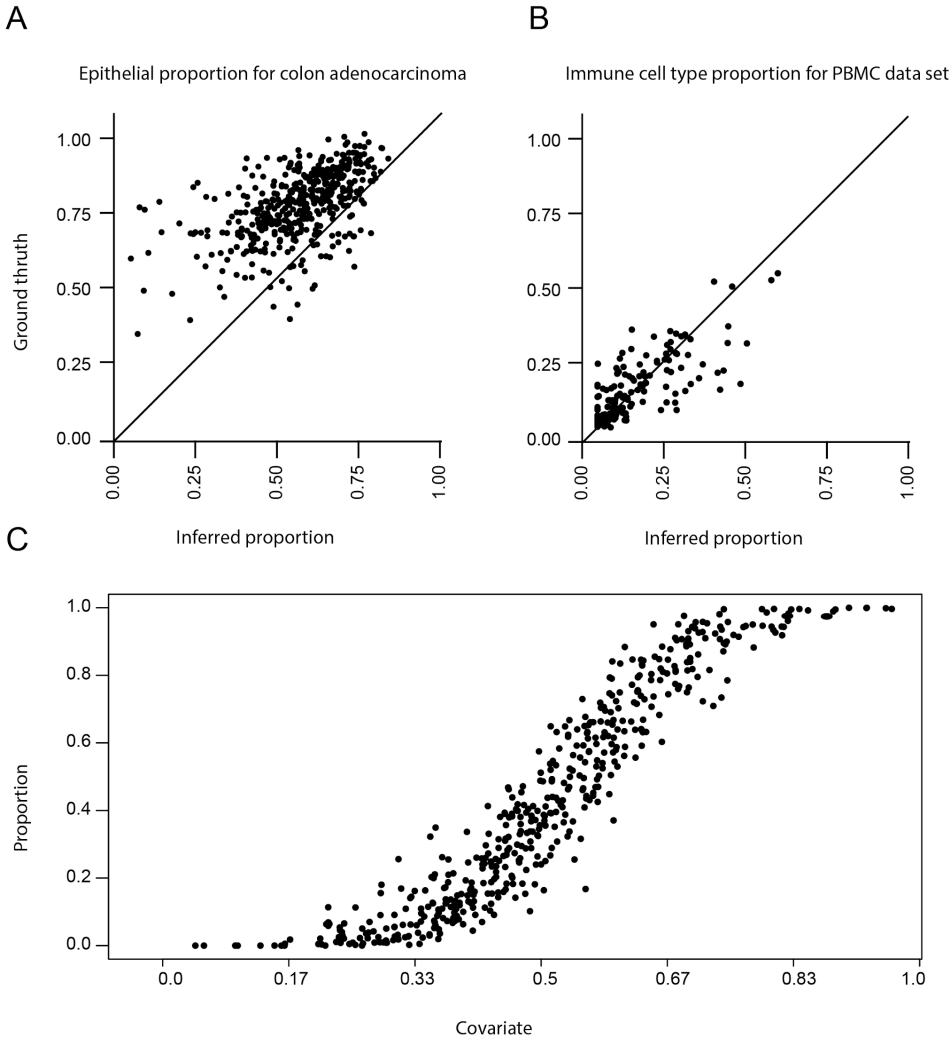


Fig. S1.2 A — Inference of epithelial proportion in TCGA colon adenocarcinoma samples, showing the compression of the error near 1, for highly pure cancer samples (i.e., where the proportion of epithelial cell is near 1). **B** — Inference of various blood cell type proportion in the PBMC validation data set, showing the compression of the error near 0, for rare cell types. **C** — Example of beta distribution showing the dimensionality compression near 0 and 1.

Summary

The tumour microenvironment is a complex evolving system; and although an old concept, it is being intensively investigated across diverse disciplines such as cellular, molecular and computational biology. In this thesis, I investigated multiple aspects of the prostate tumour microenvironment using high throughput experimental (i.e., Fluorescence-activated cell sorting; FACS) and computational techniques (Bayesian inference). I focused both on periprostatic adipose tissue and primary prostate cancer tissue.

- In chapter two, I investigated the diagnostic potential of periprostatic fat, using transcriptomic abundance data to detect a "field-effect" of the cancer-immuno activity. Here, I applied machine learning techniques to perform a 2-step feature selection and cross validation, with the goal to classify prostate cancer risk category from periprostatic fat gene transcription abundance.
- In chapter three, I investigated transcriptomic changes of the periprostatic adipose tissue in association with supercastration therapy. Here, I performed differential transcription analysis of treated against untreated patients with the goal to identify tissue metabolic changes at the molecular level that could negatively affect treatment outcome (e.g., obesity and/or inflammation).
- In chapter four, I investigated transcriptional changes along increasing risk score (i.e. CAPRA risk score), for several enriched cell types from prostate cancer tissue (i.e., Epithelial, fibroblasts, T- and myeloid cells). Here, I developed and applied an innovative differential transcription approach that was able to perform differential transcription analyses on (pseudo-)continuous covariates (i.e., CAPRA risk score), with the goal to gain knowledge about the synergic contribution of different cell types to the development of hallmarks of prostate cancer (e.g., angiogenesis, immune modulation and tissue remodelling).
- In chapter five, I developed a novel regression method targeted to proportional data. Here, I developed a Bayesian inference model with the goal of inferring both intrinsic and extrinsic proportional changes along a covariate of interest. Although stand-alone, this inference model was integrated with the method for differential tissue composition analyses developed and used in chapter six.
- In chapter six, I developed and used a novel method for differential tissue composition analysis that is able to integratively infer (i) the tissue composition from whole tissue transcriptional

data (and prior information about the transcriptional profiles of pure cell types); and (ii) infer the association between cell type abundance and a covariate of interest (e.g., risk score or tumour relapse). Here, I applied such method to the TCGA database with the goal of defining a landscape of associations between cell type abundance and cancer relapse, identifying key cell types for a diverse range of cancers.

CHAPTER 2

Context

Evidence suggests that altered adipose tissue homeostasis may be an important contributor to the development and/or progression of prostate cancer. In this study, we investigated the adipose transcriptional profiles of low- and high-risk disease to determine both prognostic potential and possible biological drivers of aggressive disease. RNA was extracted from periprostatic adipose tissue from patients categorised as having prostate cancer with either a low or high risk of progression based on tumour characteristics at prostatectomy and profiled by RNA sequencing. The expression of selected genes was then quantified by qRT-PCR in a cross-validation cohort. In the first phase, a total of 677 differentially transcribed genes were identified, from which a subset of 14 genes was shortlisted. In the second phase, a 3 gene (IGHA1, OLFM4, RERGL) signature was refined and evaluated using recursive feature selection and cross-validation, obtaining a promising discriminatory utility (area under curve 0.72) at predicting the presence of high-risk disease. Genes implicated in immune and/or inflammatory responses predominated. Periprostatic adipose tissue from patients with high-risk prostate cancer has a distinct transcriptional signature that may be useful for detecting its occult presence. Differential expression appears to be driven by a local immune/inflammatory reaction to more advanced tumours, than any specific adipose tissue-specific tumour-promoting mechanism. This signature is transferable into a clinically usable PCR-based assay, which in a cross-validation cohort shows diagnostic potential.

Periprostatic fat tissue transcriptome reveals a signature diagnostic for high-risk prostate cancer

Stefano Mangiola[1,2,3,*], Ryan Stuchbery[1,*], Geoff Macintyre[4,5,6], Michael J Clarkson[1], Justin S Peters[7], Anthony J Costello[1,2,7], Christopher M Hovens[1,2,7] and Niall M Corcoran[1,2,7,8]

1. Australian Prostate Cancer Research Centre Epworth,Richmond, Victoria, Australia
2. Department of Surgery,The University of Melbourne, Parkville, Victoria, Australia
3. Division of Bioinformatics,Walter and Eliza Hall Institute, Parkville, Victoria, Australia
4. Centre for Neural Engineering,Department of Computing and Information Systems, The University of Melbourne, Parkville, Victoria, Australia
5. Cancer Research UK Cambridge Institute,University of Cambridge, Cambridge, UK
6. Diagnostic Genomics,NICTA, Victoria Research Laboratory, The University of Melbourne, Parkville, Victoria, Australia
7. Department of Urology,Royal Melbourne Hospital, Parkville, Victoria, Australia
8. Department of Urology,Frankston Hospital, Frankston, Victoria, Australia

DOI: doi.org/10.1530/ERC-18-0058

Introduction

Prostate cancer is the most widely diagnosed cancer among men in developed countries and the second most diagnosed cancer in men worldwide[110]. The widespread use of serum PSA measurement as an ad hoc community-based screening test has meant that the majority of patients are diagnosed very early in the natural history of the disease, in the absence of any significant cancer-related symptoms[111]. It is clear both from prospective cohort studies as well as the control arms of randomised intervention trials, that many of these cancers will not pose a threat to patient life or even wellbeing within their lifetime[112–114]. Counterbalancing this is the recognition that numerically, many diagnosed cancers are biologically aggressive, and prostate cancer is the primary cause of death in over 160,000 men worldwide each year[110]. However, current methods fail to reliably discriminate indolent tumours from those with metastatic potential[115,116]. This uncertainty leads many clinicians and/or patients to choose to complete radical therapy, even when the chance of benefit is small or absent. In addition, although the majority of patients recover well, a minority will develop debilitating post-operative morbidity, which is both time consuming and

costly to manage, and significantly impacts quality of life[117]. At the same time, misidentification of biologically aggressive disease is equally troublesome, with the possibility of patients with potentially lethal tumours being inappropriately observed.

As the majority of tumours are neither palpable nor visible sonographically, a key limitation of the current diagnostic strategy is the reliance on prostate sampling, with transrectal ultrasound used to obtain needle biopsy cores from pre-specified areas within the gland[118]. This inevitably leads to a 'sampling error', in which clinically significant tumours may be missed resulting in a false-negative biopsy, or tumour grade, consistently the most important predictor of disease natural history, under- or over-estimated[115]. Indeed a recently reported large cohort study suggests that standard transrectal biopsy fails to diagnose 27–45% of clinically significant cancers[119].

Increasing evidence over the last decade indicates that altered adipose tissue homeostasis may be an important contributor to the development and/or progression of a number of solid organ tumours, including prostate cancer[120]. This was initially suggested by observations linking obesity to the development of aggressive prostate cancer[121–125], and findings that obesity-induced biochemical changes in periprostatic adipose tissue can directly affect tumour growth[126]. As prostate tumours progress locally, they frequently invade the periprostatic adipose compartment, thereby gaining immediate access to adipose tissue and locally produced adipokines that have been associated with tumour cell growth, invasion and metastases[127–130]. Similarly, tumour-derived factors may simultaneously influence the metabolic profile of periprostatic adipose tissue, perhaps culturing a niche within which the cancer can progress to a more advanced stage[131]. Although the evidence so far is the strongest in support of a role for periprostatic fat in promoting prostate cancer progression in obese patients, the association between changes in local adipose tissue metabolism and prostate cancer aggressiveness in non-obese patients is less clear.

Given the potential role of the local fat depot in prostate cancer progression, as well as the ongoing need for new strategies to improve risk stratification at the time of diagnosis, we investigated the possibility that alterations in periprostatic adipose tissue may be associated with disease risk. We performed RNA sequencing on periprostatic adipose tissue obtained at the time of prostatectomy, to determine if there existed a transcriptional signature that would allow differentiation between groups of patients at high or low risk of progression. We found significant alterations in expression affecting 677 genes and identified a distinct signature that we confirmed

by qPCR in an expanded cohort. Analyses of the genes involved suggest that differential expression is due to reactive changes within the tissue rather than local adipose tissue acting as a distinct driver of tumour progression.

Materials and methods

Ethics statement

The collection and use of tissue for this study had Epworth Healthcare institutional review board approval and patients provided written informed consent (HREC approval number 34506).

Study cohort selection

Patients with localised prostate cancer from whom adipose tissue was obtained for research purposes at the time of prostatectomy were identified from a prospectively collected adipose tissue bank[132]. Prior to ligation of the dorsal venous complex and prostate pedicles, the anterior prostate was defatted and the specimen was removed immediately, placed in a sterile container and transferred on ice for long-term storage in the vapour phase of liquid nitrogen. Selected patients had enough quantities of periprostatic adipose tissue collected at the time of surgery, had no prior local or systemic prostate cancer therapies, could be categorised into low- or high-risk cohorts based on their pathological stage, prostatectomy Gleason score and tumour volume and were free from significant systemic medical conditions. CAPRA-S scores were calculated for all patients as described[133]. Patients were grouped into a discovery phase (n = 20) and then a second, cross-validation (n = 58) phase, in a balanced fashion between low and high- risk.

Gene expression screen

A total of 50–100 µg of adipose tissue were separated from fresh frozen samples stored at ~−180°C. RNA was isolated using the Qiagen RNeasy Lipid Tissue Mini Kit and eluted in 35 µL nuclease-free water. 0.5–1 µg of total RNA was used as the input for cDNA library synthesis using TruSeq RNA Sample Prep Kit v2 (Illumina), and libraries were constructed according to manufacturer's instructions. Samples were sequenced on a HiSeq 2500 (Illumina) using 101 bp paired-end chemistry, aiming for 50 million mapped paired-end reads per sample.

Data pre-processing and differential expression analysis

A schematic summarising the overall workflow is provided in Supplementary Fig. 2.1. The RNA sequencing (RNA-seq) quality for each sample was checked using the fastqc algorithm[134]. Reads were trimmed for Illumina adapters and low-quality fragments using the trimmomatic algorithm, and short reads filtered out from the pools according to default settings[135]. The remaining reads were aligned to the reference genome hg19 using the STAR aligner with default settings[136]. The gene abundance for each sample was quantified in terms of nucleotide reads per gene (read-count) using featureCounts[137]. Low abundant genes were filtered from the analysis if not present in at least 0.5 parts per million in two-thirds of the samples in each disease group (i.e., low- and high-risk). The gene read counts were normalised based on their log-medians for each sample. The normalised gene abundances were adjusted for unknown variation using RUVseq (using default settings)[138]. The number of unwanted covariates (i.e., hidden batches) to account for was chosen through an iterative process. RUVseq was run for an increasing number of covariates (from 1 to 15); for each run, the covariate matrix calculated was added to the design matrix (i.e., low-/high-risk labels) and our samples tested for differential transcription using the edgeR package[108]. Based on the ranking of selected putative positive gene controls identified in previous studies[120] (IL6, TNF, LEP, NFKB1, CD68) as well as the P-value distribution of each run, 14 covariates were chosen accordingly(Supplementary Figs 2.2 and 2.3). The differential transcription analysis was performed on the resulting data set, utilising the identified 14 covariate parameters in the linear model of edgeR. Pathway analyses were performed on the differentially transcribed genes (false discovery rate, FDR < 0.05) using two algorithms in parallel, SPIA and GSEA[96,139]. A potential transcriptional signature of 14 genes was selected for further analysis prioritised on low FDR, high fold-change and transcript abundance.

Classification using quantitative qRT-PCR

Selected genes were analysed in an extended cross-validation cohort of 58 patients (28 low-risk, 30 high-risk) through quantitative real-time (qRT-) PCR, using 1 µL of cDNA, 0.5–1 µL qRT-PCR primers (see below), 5 µL of TaqMan Fast Advanced Master Mix (Applied Biosystems) and made up to 10 µL volume per well with UltraPure distilled water (Gibco). Primers to 14 genes from the initial exploratory cohort including IGHA1 (Hs00733892_m1), SAA2 (Hs01667582_m1), MYH11 (Hs00224610_m1), RERGL (Hs00922947_m1), SOCS3 (Hs02330328_s1), PLA2G2A (Hs00179898_m1), SLC2A1 (Hs00892681_m1), COL6A6

(Hs01029204_m1), GPR34 (Hs00271105_s1), CLDN1 (Hs00221623_m1), PCDH10 (Hs00252974_s1), SELE (Hs00174057_m1), OLFM4 (Hs00197437_m1) and DES (Hs00157258_m1) were pre-designed and commercially available from Applied Biosystems. Samples were run on a 384-well plate using a Viia7 qRT-PCR machine (Applied Biosystems) under the following conditions: UNG incubation at 50°C for 2 min; polymerase activation at 95°C for 20 s; followed by 40 cycles of denature at 95°C for 1 s; anneal/extend at 60°C for 20 s. Expression levels of target genes were normalised to the geometric mean of GAPDH (Hs00266705_g1, Applied Biosystems), TBP (Hs00427621_m1, Applied Biosystems) and POLR2A(Hs00427621_m1, Applied Biosystems) using the formula $2^{-\Delta C(T)}$.

Machine learning algorithms including support vector machine[140] (with the following settings; SVM-Type: C-classification; SVM-Kernel: radial; cost: 1; gamma: 0.3), random forest[141] (with the following settings; Number of trees: 500; number of variables tried at each split: 1) and generalised linear model[142] (with the following settings; degrees of freedom: 49; residuals: 46) were employed to classify patients as low or high risk both in the training and cross-validation phases. Genes were prioritised based on recursive feature selection using the rfe function from Caret R package[143] on a randomly subsampled training portion of the data set (~70%). The classification performances of the resulting gene rank were cross-validated against the remaining data set (~30%) across 13 (gene N-1) iterations, first including only the two top genes, with increments of one gene per iteration. This gene feature selection/cross-validation procedure was iterated 30 times with different random, balanced subsampling of training and cross-validation fractions and the mean area under curve (AUC; expressing the prediction performances of a binary classifier) was calculated for each combination. The signature size N was chosen according to the performance trend observed using from 2 to 14 classifiers. The most recurrent N genes across cross-validation were then selected and a final round of cross-validation was performed using such genes. The patients in this cohort who were also present in the RNA-seq analysis were excluded from any cross-validation set and limited to the training set for the qRT-PCR classifier. The Cibersort tool[69] was used to test for epithelial cell infiltration within the profiled periprostatic adipose tissue.

Analysis of TCGA data

Read counts and sample annotations of the TCGA prostate adenocarcinoma RNA-seq dataset were taken from the website portal.gdc.cancer.gov[144]. Data were filtered, normalised and tested for global differential expression accounting for unwanted variation as described earlier for our RNA-seq data set. The 14 genes analysed with qRT-PCR were employed to classify low-/high-grade patients in a cross-validation fashion analogously to the RNA-seq classification procedure described earlier. Furthermore, the potential infiltration of cancerous cell from the prostate into the periprostatic fat was tested with Cibersort[69] using an ad hoc signature based on LM22, with fibroblast and endothelial gene expression signatures were taken from ENCODE, BLUEPRINT and FANTOM5 data sets.

Data and computational algorithms

The raw data of sequence reads can be retrieved at ega-archive.org with the code EGAS00001002446. The informatics code used for the analyses in this work can be retrieved at github.com/stemangiola/Fat-classification-RNAseq-2017.

Results

Patient characteristics

For the initial screen, we selected 20 patients with high- or low-risk disease respectively, as determined by the prostatectomy Gleason score, pathological stage and total tumour volume (Supplementary Table 2.1). Patients in the low-risk group had median CAPRA-S score of 1 (range 0–4) with an estimated 91.0% 5-year progression-free survival, compared to a median of 7 (range 3–9) in the high-risk group and an estimated 5-year disease-free survival of only 26.9%[145]. Similarly, tumours in the high-risk patients were significantly larger than those with low-risk disease (mean 8.3 cc vs 0.8 cc, $P = 0.003$ Students t-test). The groups were however well matched for body mass index (BMI) (low-risk mean 28.0 vs high-risk mean 26.2, $P = 0.56$ Students t-test). Further patients with similar characteristics were later selected for validation studies (Table 2.1).

Gene expression of adipose tissue in prostate cancer patients

Samples were sequenced to an average depth of 67 million reads. After data filtering and normalisation, the distribution of gene read counts followed an expected log-normal distribution.

Preliminary clustering revealed two outlying samples (one each in the low and high-risk group), which were removed as previously recommended[146,147]. A multi-dimensional scaling (MDS) analysis[148] of gene read counts in the low- and high-risk cohorts demonstrated a noticeable separation for periprostatic adipose tissue (Figs 2.1 and 2.2A), which increased significantly after reduction of unwanted variation (RUV) that eliminated sample processing batch effects (Supplementary Fig. 2.4A). No significant clustering was noted based on BMI or statin use, indicating that the effect of tumour risk on transcription was greater than either of these two covariates (Supplementary Figs 2.5 and 2.6). On differential expression analysis a total of 677 differentially transcribed genes (FDR < 0.05) were identified between low- and high-risk disease in periprostatic tissue (of which the top 25 genes ranked by FDR P-values are listed in Table 2.2; full list provided as a supplementary excel file). Overall, the range of fold changes was low (from −3 to 2, Fig. 2.2B). Interestingly, the majority of the top ranked genes have roles in inflammation and immune response, including IGHA1, SAA1, SAA2, SELE, LYZ, CXCL2 and ITGAD as well SOCS3 (upregulated), which is known to be upregulated by increased levels of the inflammatory cytokines IL6 and IL10, suggesting that differences in immune activation are important to the differing severity of prostate cancers. When differentially transcribed genes were ranked according to their log fold-change (logFC), 5 of the top 20 genes overrepresented in high-risk disease encode various types of immunoglobulins. Further differences in gene expression were identified in genes involved in the transport of calcium or dependent upon calcium for their action; SCGN, GRIN2A, PLA2G2A, genes responsible for forming or maintaining the extracellular matrix; CLDN1, ITIH3, NPNT and genes encoding muscle proteins; MYH11, MUSTN1, DES, MYOZ1. Looking specifically at adipokines that have previously been implicated in driving obesity-related cancer progression, a significant increase in expression in high-risk disease was noted for IL6 (logFC 0.42, FDR = 0.04) and CCL2 (logFC 0.57, FDR = 6.4E-06), although the fold-change was quite small (Supplementary Table 2.2). No significant difference in expression was identified for TNF, LEP (leptin), ADIPOQ (adiponectin), IL10 or IL8.

Table 2.1. Clinical characteristics of study cohort

	Low-risk	**High-risk**
n	28	30

Age (yrs.)	Median	60	66
	Range	44-74	49-80
PSA (ng/dl)	Median	5.8	7.3
	Range	0.7-41.0	2.7-81.0
	<10	22	20
	Oct-20	5	5
	>20	1	5
Pathological Stage	pT2a	8	-
	pT2c	18	1
	pT3a	2	20
	pT3b	-	9
Gleason Sum (ISUP Group)	6 (1)	20	-
	3+4 (2)	8	6
	4+3 (3)	-	11
	8 (4)	-	2
	9-10 (5)	-	11
Tumour Volume (cm^3)	Mean	0.63	6.9
	Range	0.1-3.3	1.2-32.1
Recurrence	No	28	13
	Yes	0	17
Follow up (months)	Mean	27	29
	Range	2-59	3-50

Pathway analysis of the list of differentially transcribed genes using both SPIA and GSEA algorithms (Table 3) identified 18 differentially regulated pathways (FDR < 0.05), with an overlap

of 7 pathways. Most differences observed in functional pathway regulation demonstrate cancer-related alterations to immune response and inflammation. A gene signature to be refined and cross-validated with qRT-PCR was selected based on FDR and logFC, including IGHA1, SAA2, MYH11, DES, RERGL, SOCS3, PLA2G2A, SLC2A1, COL6A6, GPR34, CLDN1, PCDH10, SELE and OLFM4.

qRT-PCR refinement of the gene signature

qRT-PCR was used to interrogate the 14 selected genes across a larger cohort of 58 patients (28 low-risk, 30 high-risk). Of these genes, IGHA1, MYH11, RERGL, SOCS3, PLA2G2A, CLDN1 and OLFM4 were confirmed to be significantly differentially transcribed across the two groups (P-value <0.05, one-sided Student's t-test) and GPR34, PCDH10 and SELE were found to have P-values <0.1 (Supplementary Fig. 2.7). Overall, qRT-PCR fold-change between low- and high-risk tumours (calculated on delta-TC-value) showed a positive linear correlation with the RNA-seq transcription fold-change (calculated on mean read counts), with a slope of 1.5 and a P-value (linear model; lm R package)[149] of 0.06 and an R 2 of 0.26 (Fig. 2.3A). Of the four genes that did not validate, two (SLC2A1 and GPR34) had a logFC <0.7. The second phase feature selection using qRT-PCR gene abundance values led to the decision to set the signature size to 3 (Fig. 2.3B) as the performances of the three classifiers peaked at this value. The best performance expressed in mean AUC was 0.73 (S.D. = 0.14; Fig. 2.3B). Among the 3 gene signatures across iterations, the genes IGHA1, OLFM4 and RERGL occurred most often and further cross-validation using only these genes led to a mean AUC of 0.72 (S.D. = 0.14; Fig. 2.3C), with an out-of-bag estimate of error rate ranging from 31 to 39% across iterations.

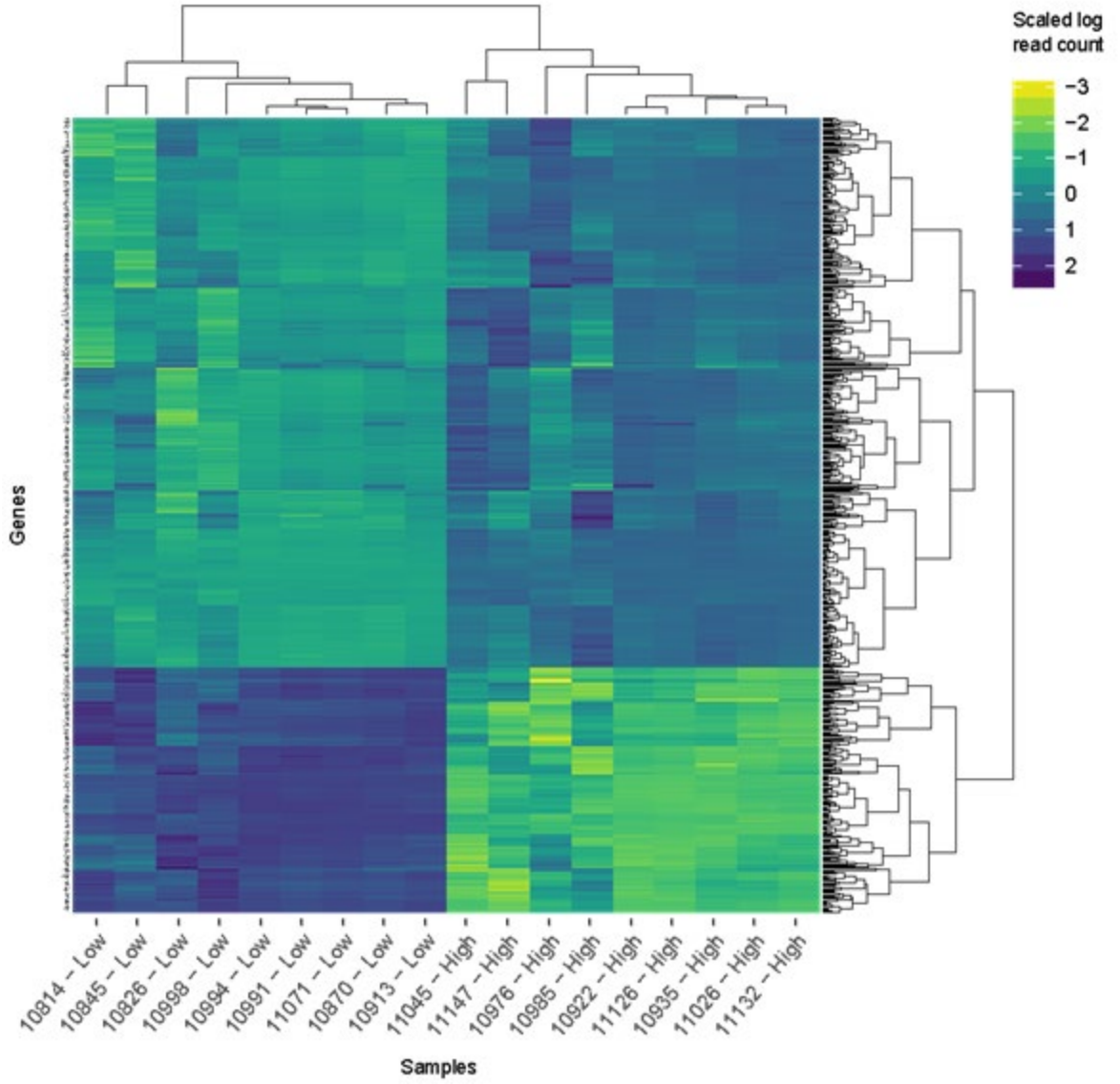


Figure 2.1 Heatmap of the differentially expressed genes in periprostatic adipose tissue between patients with low- and high-risk prostate cancer.

Specificity of the gene signature to fat

To confirm that the signature was specific to fat and did not just represent sample contamination by invasive prostate cancer, several approaches were taken. Firstly, we examined global differential expression in the TCGA prostate cancer data set and found little overlap in differentially transcribed genes compared to the periprostatic fat cohort with just two of the genes in the 14-gene signature reaching FDR < 0.05 (Supplementary Table 2.3), neither of which was a

component of the final 3-gene assay. Secondly, we applied the 14-gene signature developed for adipose tissue to the TCGA dataset, which showed an overall negligible ability to classify high- and low-risk cancer cancers (Fig. 2.3C and Supplementary Fig. 2.4B), with a mean AUC of 0.61 compared with 1.0 using the adipose tissue RNA-seq dataset, and 0.72 using the less complex qRT-PCR for our 3-gene signature (Fig. 2.3C), indicating that the signature is specific to adipose tissue. In addition, analysis of epithelial cell infiltration of the periprostatic fat using Cibersort indicated that epithelial cells were rare in the adipose tissue in both groups, and not significantly different between low- or high-risk prostate cancer patients (P = 0.84, Supplementary Fig. 2.8). Several immune cell subtypes, however, were differed significantly between groups, including eosinophils (P = 0.009) and T memory effector cells (P = 9.8E-8). We also did not detect any expression of prostate epithelial cell-specific transcripts such as KLK3, KLK2 and PCA3 in any sample.

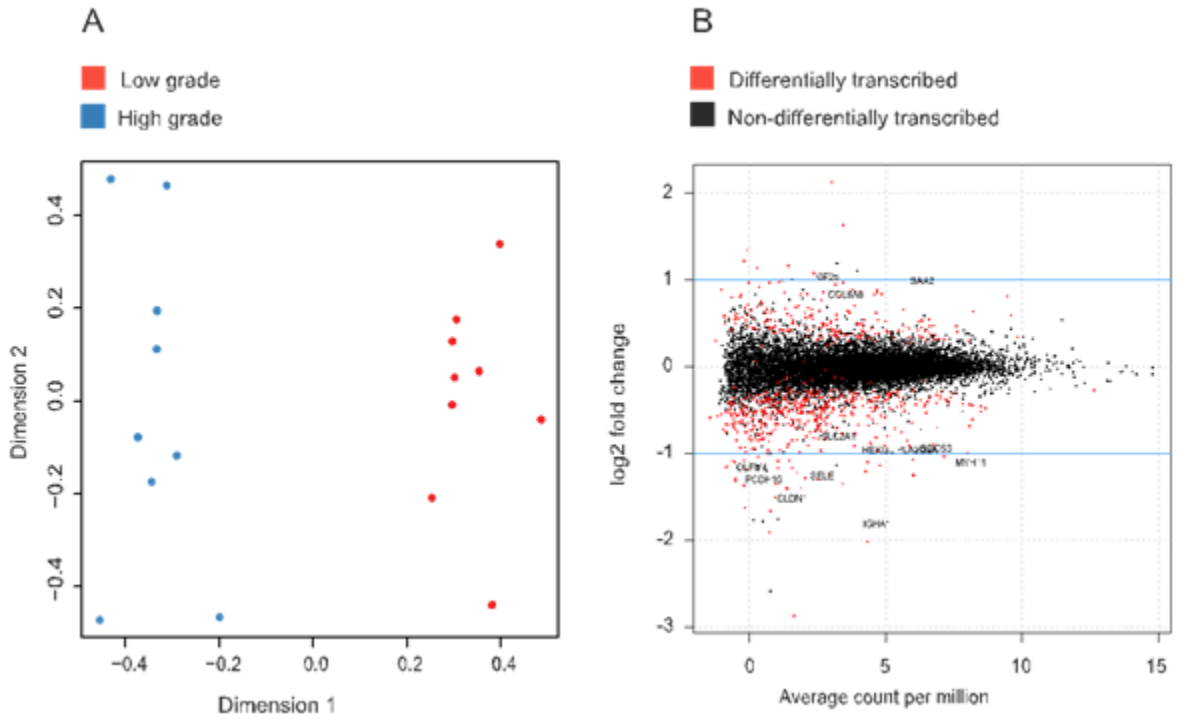


Figure 2.2. (A) Multi-dimensional-scaling (MDS) plot of RNA-seq samples after filtering, normalisation and removal of unwanted variation. (B) Smear plot of genes showing the relation between log fold changes (logFC) and gene abundance (count per million, CPM). The genes selected for qRT-PCR Dimension 1 analysis are labelled in the plot.

Table 2.2. Top differentially transcribed genes; ordered based on increasing P-value.

Gene symbol	logFC	logCPM	LR	p-value	FDR
IGHA1	1.7	4.6	138.9	4.50E-32	7.00E-28
SAA2	-1.1	6.3	111.9	3.60E-26	2.80E-22
SYCE1	-1.2	1.5	98.7	2.90E-23	1.50E-19
SAA1	-0.9	9.5	96.1	1.00E-22	4.00E-19
CLDN1	1.4	1.5	91.9	8.70E-22	2.70E-18
SCGN	1.3	1.1	87.3	9.30E-21	2.40E-17
MYH11	1	8.1	82.1	1.20E-19	2.70E-16
SELE	1.1	2.7	76.2	2.40E-18	4.70E-15
MUSTN1	0.8	4	69.1	9.20E-17	1.50E-13
LYZ	-0.8	4.7	68.9	1.00E-16	1.50E-13
GRIN2A	1.2	0.2	64.8	8.00E-16	1.10E-12
DES	1.1	4.4	64.5	9.40E-16	1.10E-12
MYOZ1	1.1	0.5	64.4	9.80E-16	1.10E-12
SHOX2	0.9	2.4	64.2	1.00E-15	1.20E-12
ITIH3	1.2	1.8	63	2.00E-15	2.10E-12
SUSD5	1	2.8	60.4	7.40E-15	7.10E-12

KRT19	1.7	1.1	59.9	9.90E-15	8.90E-12
CXCL2	0.8	3.9	59.8	1.00E-14	8.90E-12
NPNT	0.8	4.1	57.7	2.90E-14	2.30E-11
SOCS3	0.8	6.8	57.2	3.80E-14	2.90E-11
TBX5	-1.3	2.1	56.7	4.90E-14	3.60E-11
LOC284454	0.7	4.9	56.5	5.30E-14	3.80E-11
ITGAD	-1.5	0.1	56	6.80E-14	4.60E-11
PLA2G2A	0.8	6.2	55.2	1.00E-13	6.90E-11
RERGL	0.8	4.7	54.6	1.40E-13	8.90E-11

Table 2.3. Pathways enriched in periprostatic adipose tissue derived from patients with low-risk compared to high-risk prostate cancer as determined by two independent algorithms, SPIA and GSEA.

Name	**ID**	**FDR**	**Status**	**Algorithm**
Cytokine-cytokine receptor interaction	4060	3.76-04	Inhibited	SPIA
Malaria	5144	2.30E-03	Activated	SPIA
Graft versus host disease	5332	3.80E-03	Inhibited	SPIA

Circadian rhythm	4710	3.80E-03	Inhibited	SPIA
cAMP signalling pathway	4024	8.90E-03	Activated	SPIA
Autoimmune thyroid disease	5320	1.90E-02	Inhibited	SPIA
Allograft rejection	5330	1.90E-02	Inhibited	SPIA
Leishmaniasis	5140	1.90E-02	Activated	SPIA
Type I diabetes mellitus	4940	2.10E-02	Inhibited	SPIA
Intestinal immune network for IgA production	4672	2.80E-02	Activated	SPIA
Pathways in cancer	5200	3.30E-02	Activated	SPIA
Systemic lupus erythematosus	5322	4.00E-02	Activated	SPIA
Antigen processing and presentation	M16004	≈0	NA	GSEA
Allograft rejection	M18615	≈0	NA	GSEA
Type I diabetes mellitus	M12617	≈0	NA	GSEA
Graft versus host disease	M13519	2.00E-04	NA	GSEA

Autoimmune thyroid disease	M13103	6.00E-04	NA	GSEA
Asthma	M13950	1.20E-03	NA	GSEA
Leishmania infection	M3126	2.80E-03	NA	GSEA
Intestinal immune network for IGA production	M615	8.20E-03	NA	GSEA
Viral myocarditis	M12294	9.00E-03	NA	GSEA
Metabolism of xenobiotics by cytochrome p450	M16794	1.30E-02	NA	GSEA
Systemic lupus erythematosus	M4741	2.10E-02	NA	GSEA

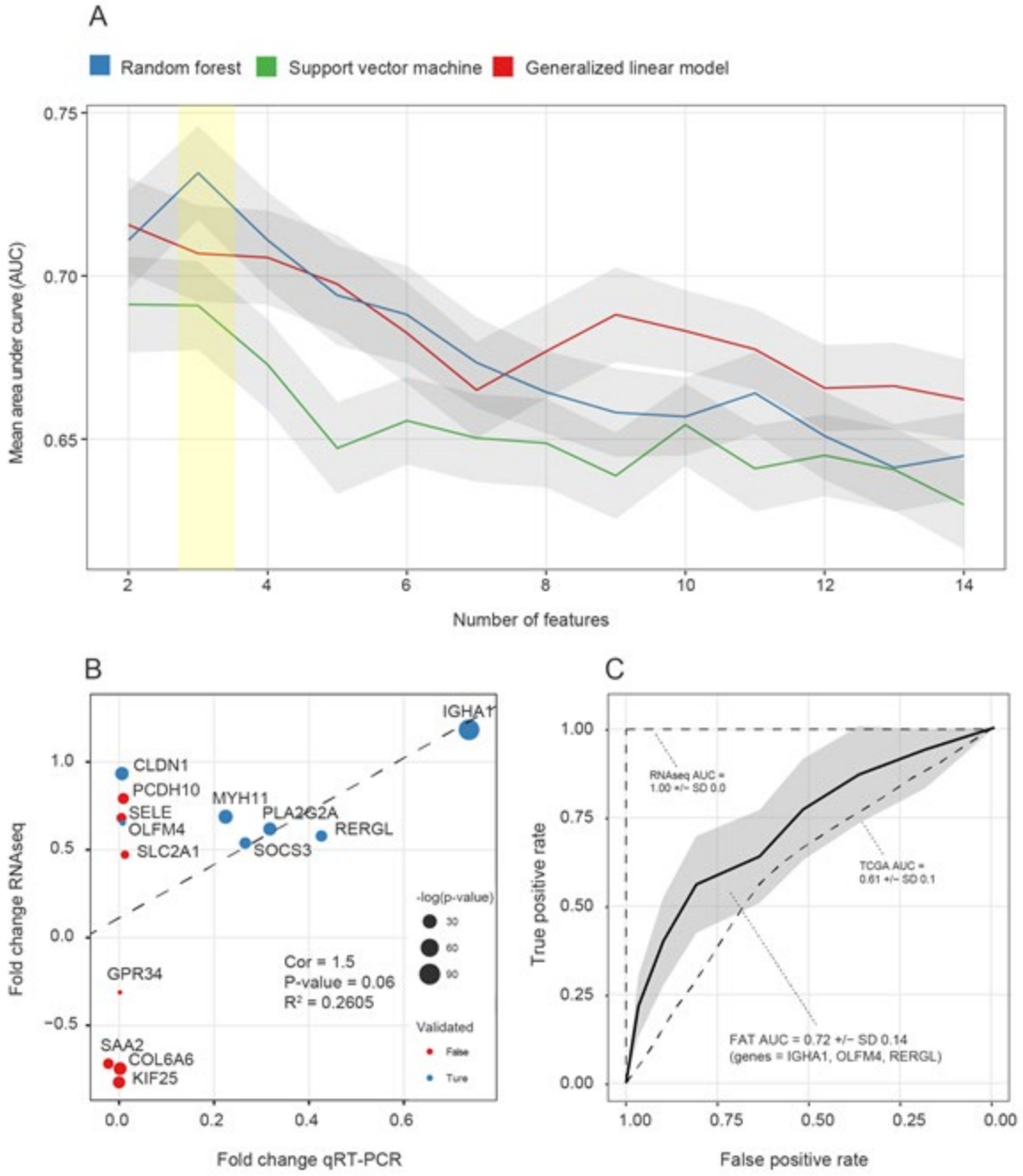


Figure 2.3. (A) The correlation plot of the fold-change for the selected genes between qRT-PCR and RNA sequencing. (B) Area under curve (AUC), of the ROC curve representing the classification performances of three different classifiers on qRT-PCR for a range of gene signature sizes. (C) The performances of the classifier created with the selected gene signature (n = 3 genes) on the qRT-PCR cohort, represented as a ROC curve. For comparison, the dashed lines represent

the performances for the same genes on our RNA-seq (Supplementary Fig. 2.4A) and the TCGA prostate adenocarcinoma data sets.

Discussion

Due to the sampling error associated with the standard prostate biopsy technique, there is always the inherent risk that potentially lethal tumours may be missed, even after multiple biopsies[150]. Although a number of advances have been made to reduce these errors, such as the use of transperineal mapping biopsies[119] as well as MRI-guided needle placement[151], patient misclassification remains a common clinical problem, and there are significant trade-offs in terms of increased costs and potential side effects. Contemporaneously with advances in biopsy technique, two transcriptional signatures have been developed, which are used clinically to predict the presence of more advanced pathological features or an increased risk of recurrence[133,152], both of which can reduce misclassification error particularly in those patients diagnosed with low-risk disease. However, both these tests depend upon the presence of adequate tumour tissue in the diagnostic core to obtain usable RNA, which may be limited, particularly in patients with low volume disease, and the risk scores obtained can vary with the individual tissue core selected for input from any given cancer due to inherent genomic heterogeneity within primary prostate tumours[153]. A potential alternative strategy involves the identification of a 'field-change' specifically within benign tissue that is associated with adverse pathological features or clinical outcome. Certainly, a number of studies have shown that gene expression in benign prostate tissue of cancer-bearing organs differ significantly from that of benign tissue obtained from cancer-free glands[154–156], and the altered expression in 39 genes in tumour-adjacent normal-appearing tissue has recently been associated with the risk of recurrence post treatment[157].

Given the epidemiological and experimental associations between periprostatic adipose tissue homeostasis and prostate cancer aggression, we believed that analysis of periprostatic fat could potentially yield a useful signal that could help detect the presence of high-risk tumours. We therefore performed a genome-wide analysis of gene expression in periprostatic adipose tissue obtained from patients with low- and high-risk disease, and despite the low dynamic range of differential expression, have identified a transcriptional signature that could distinguish between the two groups with high accuracy. The overall lack of significance in correlation between RNA-seq fold-change and qRT-PCR fold-change (Fig. 2.3A; P-value = 0.06) was expected, considering

the increase of sample size and the change of technology for the cross-validation cohort. On the other hand, 7 genes in particular were shown to be consistent between the two cohorts, providing more confidence on the applicability of our signature into a clinical setting. We have then translated and refined this signature into a 3-gene qRT-PCR-based assay, which demonstrates promising discriminatory ability in an expanded cross-validation cohort, confirming the presence of a cancer-related 'field effect' in periprostatic adipose tissue that may be clinically exploitable. Indeed, the use of a fat-based expression assay has many advantages over a tumour tissue-based test, most importantly in that it does not rely on the actual detection of tumour in the diagnostic biopsy, and the anterior fat pad is amenable to biopsy, particularly with transperineal needle placement. There is also the potential that there is less heterogeneity in the signal across the fat depot, although this has not been tested.

The role of adipose tissue in promoting cancer initiation and development is best understood in relation to obesity, which has been the subject of intense research interest over the last two decades. It is now well recognised that obesity induces a low-grade chronic inflammatory state within adipose depots, with infiltration by cells of both the innate and specific immune system. This results in the elaboration of various cytokines such as IL6, MCP-1 and TNF-α, which given the prostate gland is surrounded in fatty tissue, may promote prostate cancer progression in a paracrine manner[158]. Certainly conditioned media derived from periprostatic adipose tissue from obese patients significantly increased the proliferation of PC-3 cells to a significantly greater extent than similar media from lean counterparts[126]. However, multiple studies have demonstrated that similarly conditioned media derived from non-obese patients can promote the proliferation and invasion of human prostate tumour cells[130,159,160]. Indeed, prostate cancer has been demonstrated to induce many of these changes in the surrounding fat, establishing a reciprocal loop that promotes tumour progression[131,160]. Besides changes in cytokine expression, adipose tissue can secrete a number of fat-specific adipokines, including leptin and adiponectin, which can affect tumour cell growth and have previously been found to be altered in the anterior fat pad of prostate cancer patients[161]. Additionally, changes in the expression of proteins involved in adipocyte lipolysis and/or lipogenesis have been implicated in promoting prostate cancer growth, by increasing local supply of lipids required for intratumoural energy production[162].

Given these observations, a priori we anticipated that any distinguishing expression signature would largely comprise of these recognised changes, particularly given the differences

in clinical aggressiveness between the two cohorts used for the screen. However, we could only identify significant changes in two potentially relevant cytokines (IL6 and CLL2), and the differential expression was low with a logFC between cohorts being < 0.6. This is perhaps not surprising, given that the cohorts were well matched for BMI and suggests previously described changes may be obesity state specific, and not contributing significantly to the progression of high-grade high-stage disease in non-obese individuals. The optimal 3-gene signature that did emerge comprising IGHA1(immunoglobulin heavy chain constant alpha 1), OLFM4 (olfactomedin 4) and RERGL (RAS-related and oestrogen-regulated growth inhibitor-like protein). IGHA1 encodes the first part of the constant region the IgA1 isoform of the immunoglobulin IgA, a secretory antibody that is produced predominantly at mucosal surfaces where it binds to and prevents pathogen entry. The IgA1 isoform however is predominantly found in tissue and serum, and functions to opsonise foreign antigens to initiate phagocytosis and antibody-mediated cytotoxicity. Interestingly, prostate tumour-induced immune tolerance has previously been linked to the selective recruitment of IgA expressing B-cells, that suppress induction of a specific immune response through the expression of PD-L1 and IL10[163]. OLFM4 encodes an extracellular matrix protein, which is involved in cell adhesion and has anti-apoptotic effects. Counter-intuitively, intratumoural loss of expression of olfactomedin 4 has been linked to prostate cancer development and progression, reportedly through activation of the hedgehog signalling pathway[164]. In contrast, our data indicate that OLFM4 is overexpressed in periprostatic adipose tissue associated with high-risk disease, although the source of the transcript is unclear, as the expression level in normal fat is very low[165]. One potential source is through tissue infiltration by a specific subset of neutrophils, which express the transcript and are defined by it[166]. The function of RERGL is unknown, but it shares significant sequence homology with the RAS-superfamily member RERG, which encodes a tumour-suppressing GTPase and is predicted to share many of its functions[167]. Loss of expression of RERGL has been identified in colorectal cancer, where it is associated with poorer overall survival[167]. Although RERGL transcript is ubiquitously expressed, including in adipose tissue, preliminary immunohistochemical studies suggest that adipocytes do not elaborate the protein, although both fibroblasts and lymphocytes stain strongly[165]. Certainly, tissue decomposition analysis using Cibersort supports alteration in the immune cell composition of adipose tissue as a potential source of the discriminatory signal, with significant changes in the proportion of several immune cells types. However, given that immune cells make up a very small proportion of the

tissue, and the performance of these types of algorithms in complex tissues is untested, further work is required to validate these finding.

Despite the use of a cross-validation rather than an independent validation cohort this pilot study shows a diagnostic potential for periprostatic tissue. Although the AUC of 0.72 is modest, it is considerable considering the size of the discovery cohort in this study, and it is of the same predictive range of other approved polygenic tests such as Mi-Prostate Score (AUC = 0.77), SelectMDx (AUC = 0.76) and ExoDx (AUC = 0.71) for high-risk disease[168–170]. This suggests that an extended discovery cohort (n > 200) may result in a more robust signature with greater classification accuracy when translated into a PCR-based assay. Although we have not formally tested the adequacy of periprostatic tissue sampling, anecdotal experience (NMC) indicates that it may be biopsied inadvertently when obtaining anterior cores during transperineal prostate biopsy without adverse effects, and certainly yields sufficient RNA for qRT-PCR (up to 100 ng of RNA from a single biopsy core). Feasibility of this approach will however require formal testing in a prospective study.

In summary, we have identified a transcriptional signature in periprostatic fat that can distinguish patients with clinically localised prostate cancer at low or high risk of progression and have successfully translated it into a 3-gene qRT-PCR-based assay. The basis of this signature appears to be related more to a local immune and/or inflammatory reaction to the presence of high-risk tumour rather than a specific adipose tissue-based tumour-promoting mechanism as previously described, although the latter may be more obvious in obese and severely obese patients. Significant developmental work is required to assess utility in more marginal cases as well its specificity in the presence of benign prostatic conditions such as benign prostatic hyperplasia and prostatitis, before it can translated into a clinically usable test.

Supplementary data

This is linked to the online version of the paper at https://doi-org.ezp.lib.unimelb.edu.au/10.1530/ERC-18-0058.

CHAPTER 3

Context

Prostate cancer is a leading cause of morbidity and cancer-related death worldwide. Androgen deprivation therapy (ADT) is the cornerstone of management for advanced disease. The use of these therapies is associated with multiple side effects, including metabolic syndrome and truncal obesity. At the same time, obesity has been associated with both prostate cancer development and disease progression, linked to its effects on chronic inflammation at a tissue level. The connection between androgen deprivation therapy, obesity, inflammation, and prostate cancer progression is well-established in clinical settings; however, an understanding of the changes in adipose tissue at the molecular level induced by castration therapies is missing. Here we investigated the transcriptional changes in periprostatic fat tissue induced by profound androgen deprivation therapy in a group of patients with high-risk tumours compared to a matching untreated cohort. We find that the deprivation of androgen is associated with a pro-inflammatory and obesity-like adipose tissue microenvironment. This study suggests that the beneficial effect of therapies based on androgen deprivation may be partially counteracted by metabolic and inflammatory side effects in the adipose tissue surrounding the prostate.

Androgen deprivation therapy promotes an inflammatory and obesity-like microenvironment in periprostatic fat

Introduction

For over 80 years, androgen deprivation by surgical or medical castration has been the cornerstone of treatment for advanced prostate cancer[171]. As new cytotoxic and androgen receptor targeted therapies have been developed, demonstrating survival benefit in combination with androgen deprivation in a number of clinical settings, the duration a patient can expect to be in a castrated state prior to death has been extended significantly[172]. Given that androgen signalling is important for homeostasis in a number of different organ systems, it is not surprising that both short and long term use is associated with a number of deleterious effects (reviewed in Rhee et al., 2015[173]).

Forefront of these is the association of androgen deprivation with metabolic syndromes such as diabetes mellitus[174] and obesity[175], as androgens play a key role in the regulation of intermediate metabolism and tissue composition[176]. Increased fat tissue mass (known in conjunction with loss of muscle mass as sarcopenic obesity) is one of the main metabolic side effects of androgen deprivation therapy (ADT)[177], even for short-term treatment[178–180]. At the molecular level, lack of androgen related hormones leads to changes in tissue lipid composition and decreased insulin sensitivity[174]. For example, gonadotropin-releasing hormone agonists have been shown to alter tissue lipid profiles with cholesterol levels, triglycerides, and high-density lipoproteins shown to increase up to 10.6%, 25%, and 8–20% respectively[179,180].

The promotion of an obese-like phenotype by androgen deprivation is highly clinically relevant, as obesity (expressed as body mass index; BMI) is itself associated with the development of prostate cancer, post-prostatectomy biochemical failure, and risk of death from prostate cancer. Although the link between elevated body mass index and increased risk of prostate cancer is still controversial[181,182], several studies have found a positive association between body mass index and cancer grade and/or stage at the time of radical prostatectomy[183–185]. Two recent studies identified an association between body mass index and biochemical failure rates following radical prostatectomy, based on a large scale, multi-ethnic cohort[186,187]. The relationship between body mass index and prostate cancer-specific mortality is also widely supported[122–124,188].

Although the connection between androgen deprivation therapy, obesity, and prostate cancer progression is well-established in clinical settings, a molecular understanding of the changes in adipose tissue associated with castrating therapies is still missing, in part due to a paucity of appropriate clinical specimens. This is especially important for periprostatic adipose tissue due to its proximity to the cancer site and its potential to influence prostate hormonal and immune homeostasis[189]. Here for the first time, based on a unique cohort of patients with six months profound androgen suppression and receptor blockade, we performed an integrative study of the molecular and cellular changes in periprostatic fat associated with androgen deprivation. In this study we show that androgen deprivation therapy is associated with a pro-inflammatory and obesity-like adipose tissue microenvironment.

Materials and Methods

Ethics statement.

The collection and use of tissue for this study had Epworth Healthcare institutional review board approval and patients provided written informed consent (HREC approval number 34506).

Study cohort selection

Androgen deprivation therapy treated patients (n=11) were recruited from an open label neoadjuvant phase II study in which patients with high-risk disease received a 'supercastration' regimen consisting of degarelix 240/80 mg subcutaneously every four weeks; abiraterone acetate 500 mg orally daily titrating upwards every two weeks by 250 mg to a final dose of 1000 mg daily; bicalutamide 50 mg orally daily; and prednisolone 5 mg orally twice daily for a total of 6 months (Australian New Zealand Clinical Trials Registry 12612000772842). Untreated patients with similar pre-treatment characteristics were obtained from a prospective prostatectomy biorepository[132,189]. Prior to ligation of the dorsal venous complex and prostate pedicles, the anterior prostate was defatted and the specimen was removed immediately, placed in a sterile container and transferred on ice for long-term storage in the vapour phase of liquid nitrogen.

Gene expression screen

A total of 50–100 µg of adipose tissue was separated from fresh frozen samples stored at −160°C. RNA was isolated using the Qiagen RNeasy Lipid Tissue Mini Kit and eluted in 35 µL nuclease-

free water. 0.5–1 µg of total RNA was used as the input for cDNA library synthesis using TruSeq RNA Sample Prep Kit v2 (Illumina), and libraries were constructed according to manufacturer's instructions. Samples were sequenced on a HiSeq 2500 (Illumina) using 101 base paired-end chemistry, aiming for 50 million mapped paired-end reads per sample.

Data pre-processing and mapping

The RNA sequencing quality for each sample was controlled using the FastQC algorithm[134]. Reads were trimmed for Illumina adapters and low-quality fragments using the Trimmomatic algorithm, and short reads filtered out from the pools according to default settings [135]. The remaining reads were aligned to the reference genome (hg19) with the STAR aligner using default settings[136]. The gene abundance for each sample was quantified in terms of reads per gene (read-count) using featureCounts[137]. Low abundance genes were filtered from the analysis, if not present in at least 0.5 parts per million in two-thirds of the samples in each treatment group (i.e., treated and naïve).

Differential expression and gene set enrichment analyses

Considering the sparse batch distribution, the gene abundances were adjusted for unknown variation using RUVseq with one unwanted covariate (using default settings)[138]. The resulting covariate matrix for the unwanted covariate was appended to the design matrix (i.e., treated vs. naïve, plus the intercept term); then, all samples were tested for differential transcription using the edgeR package[108], considering differentially transcribed genes with a false discovery rate < 0.05. Ensemble pathway analyses were performed using the algorithm EGSEA[190]. In order to test for the enrichment of an obesity molecular phenotype among the differentially transcribed genes, an ad hoc signature data set (supplementary file ijo2014210x1[191]) was queried using the algorithm GSEA[96].

Differential tissue composition analyses

The associations between (i) the abundance of stromal and immune cell types within the tissue and (ii) the treatment status (i.e., treated or naïve) was inferred using two distinct approaches. Both approaches included a two-step inference, where the cellular composition of each sample is inferred first (i.e., the proportion of several cell types within the tissue sample), and an association analysis is performed integrating such inference with the treatment status. The first approach applied the algorithm Cibersort[69] for the inference of tissue composition, in combination with

DirichletReg[192] for the regression of the proportional estimates produced by Cibersort[192]. Considering that Cibersort was designed mainly for microarray data, and only for PBMC cell types, a custom probabilistic Bayesian model was also implemented (based on the Markov chain Monte Carlo probabilistic framework Stan[193]), which natively models RNA sequencing data and performs association analysis in an integrative manner preserving uncertainty information between the two steps. This probabilistic model can be described by a joint probability density formula and a series of sampling statements (see Supplementary Material), the full methodology and validation is described in Chapter 6.

qRT-PCR validation

In order to validate the methodology used for the inference of differential transcription, qRT-PCR was used for an independent observation of gene transcript abundance. A total of nine differentially transcribed genes were selected for validation with qRT-PCR, based on false discovery rate (< 0.05), log fold change (> 2) and on the absence of clear outliers. The qRT-PCR validation was performed using 1 μL of cDNA, 0.5 μL qRT-PCR primers (see below), 5 μL of TaqMan Fast Advanced Master Mix (Applied Biosystems) and 3.5 μL of UltraPure distilled water (Gibco). The primers, including ART3 (Hs00922621_m1), CSDC2 (Hs00411093_m1), DIO2 (Hs05050546_s1), FCGR1B (Hs00174081_m1), LYZ (Hs00426232_m1), OR51E2 (Hs00258239_s1), SLC16A12 (Hs01584854_m1), SUSD5 (Hs01394532_m1), and TRIM67 (Hs01595609_m1), were pre-designed and commercially available from Applied Biosystems. Samples were run on a 384-well plate using a Viia7 qRT-PCR machine (Applied Biosystems) under the following conditions: UNG incubation at 50°C for 2 min; polymerase activation at 95°C for 20 s; followed by 40 cycles of denature at 95°C for 1 s; anneal/extend at 60°C for 20 s. Expression levels of target genes were normalized to the geometric mean of GAPDH (Hs00266705_g1, Applied Biosystems), TBP (Hs00427621_m1, Applied Biosystems) and POLR2A (Hs00427621_m1, Applied Biosystems) using the formula $2^{-\Delta C(T)}$. One-sided Student's t-test was used for hypothesis testing; then, Bonferroni multiple-test correction was applied to the produced p-values.

Patient characteristics

The treated and naïve groups comprised 11 and 10 patients respectively; their clinical and pathological characteristics are shown in Table 3.1. Given that pre-operative risk assessment is frequently inaccurate[115,116], being biased towards underestimation of tumour grade and stage, patients in the high risk cohort were selected based on the stage, grade, and volume of tumour in the prostatectomy specimen. All patients in the treated cohort had high risk disease at the time of initial assessment, although the ultimate response to androgen deprivation was highly variable.

Differentially transcribed genes represent three main functional groups

The RNA sequencing libraries had an average of 55 million reads across the 21 samples. All samples had a Phred quality score exceeding 28 following filtering and trimming[134]. As expected, the distribution of the multi-dimensional scaling (MDS) analysis[148] including both treated and naïve groups showed the improvement in clustering obtained through the removal of unwanted variation (RUVseq; Fig. 3.1A and 3.1B). However, the overall magnitude of differences between the two groups was low (i.e., log fold difference < 3; Fig. 3.1B and 3.1C). No significant difference was found between the two treatment categories for body mass index or CAPRA-S risk score distributions (adjusted p-value = 1.0 and 1.0 respectively; Fig. S3.1).

A total of 70 genes were identified as differentially transcribed (false discovery rate < 0.05; Table S3.1), characterized by a median fold change of 3.23. Of these, 49 genes were characterized by a fold change greater than 2. Among the differentially transcribed genes with fold change greater than 2, three recurring biological processes (from grouping analogous gene ontology annotations; GO[194]; Table S3.2) were identified: hormonal and fat homeostasis (n = 8), inflammation (n = 8) and neural plasticity (n = 4). Several genes involved in cholesterol metabolism were found to be upregulated from the hormonal homeostasis gene set. One such gene encodes for cytochrome P450, family 1, member A1 (CYP1A1), which catalyses several reactions involved in the synthesis of cholesterol, steroids and other lipids, as well as drug metabolism[195]. Another upregulated gene, fatty acid desaturase 2 (FADS2), is a known modulator of lipid composition in skin[196]. Within in the treated cohort, several genes were decreased in abundance such as iodothyronine deiodinase 2 (DIO2), which is associated with the biosynthesis of thyroid hormone[197]; and cyclin A1 (CCNA1), which is involved in spermatogenesis[198]. For inflammation,

upregulated genes were enriched over downregulated genes (n = 7 vs. 1 respectively). The transcriptional changes with larger magnitude involved two paralog genes (i.e., IGKV1D−39 and IGKV1−39) encoding for "v" region of the variable domain of immunoglobulin light chains, mainly secreted by B lymphocytes and participating in antigen recognition[199]. The only downregulated gene within the inflammation category was WAP four-disulphide core domain 1 (WFDC1), which is linked to negative regulation of the inflammatory response[200].

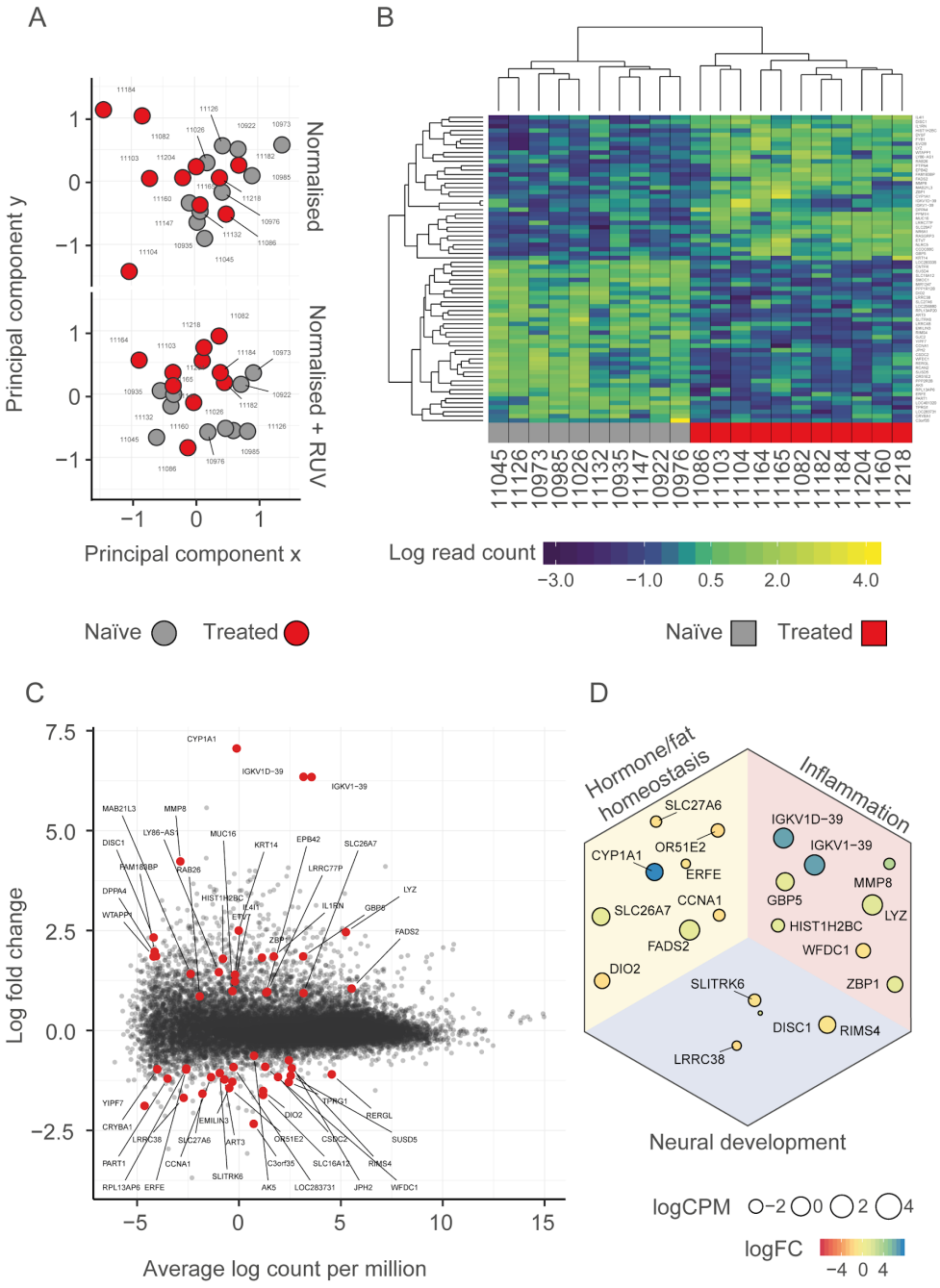


Figure 3.1: A — Heatmap of the top differentially regulated genes, with unsupervised hierarchical clustering for samples and genes. B — Multidimensional scaling (MDS) plot of the treated and naïve cohorts before and after removal of the unwanted variation (K = 1). C — Smear plot indicating the differentially transcribed genes in red. D — recurrent functional groups within the differentially transcribed genes.

For neural development, the transcript abundance of most genes was decreased in treated patients, including several genes regulating synapse formation such as regulating synaptic membrane exocytosis 4 (RIMS4). Among nine differentially transcribed genes, a total of seven validated with qRT-PCR, after correcting for multiple hypothesis testing (i.e., adjusted p-value <0.05; Fig. S3.2).

Table 3.1. Clinical characteristics of study cohort. PSA = prostate specific antigen; BMI = body mass index.

		Naïve	**Treated**
Age (yrs)	Median	66	65
	Range	49-72	63-72
PSA (ng/dl)	Median	7.5	14.4
	Range	2.7-27	4.4-95
	<10	7	5
	10-20	2	2
	>20	1	4
Clinical Stage	cT1	3	2
	cT2	7	4
	cT3	0	5
Biopsy Grade	ISUP2	2	1
	ISUP3	3	0
	ISUP4	2	3

	ISUP5	3	7
Pathological Stage	pT0	0	1
	pT2	0	3
	pT3	10	8
Prostatectomy Grade	ND	0	1
	ISUP1	0	2
	ISUP2	0	1
	ISUP3	3	1
	ISUP4	1	2
	ISUP5	6	5
Tumour Volume	Median	7.1	1
	Range	0.7-17.8	0-9.3
BMI (kg/m2)	Mean	26.9	28.2
	SD	2.9	4

Table 3.2. EGSEA results

GeneSet	**Direction**	**p.value**	**p.adj**
Hallmark Signatures			
Hallmark allograft rejection	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
Hallmark kras signalling up	Up	$< 1.0x10^{-16}$	$1.0x10^{-06}$
Hallmark inflammatory response	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
Hallmark IL6 jak stat3 signalling	Up	$8.0x10^{-06}$	$5.0x10^{-05}$
Hallmark interferon gamma response	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
Gene ontology			

GO regulation of innate immune response	Up	$2.0x10^{-06}$	$3.8x10^{-05}$
GO innate immune response	Up	$< 1.0x10^{-16}$	$9.0x10^{-06}$
GO positive regulation of defense response	Up	$4.0x10^{-06}$	$8.4x10^{-05}$
GO positive regulation of immune response	Up	$< 1.0x10^{-16}$	$9.0x10^{-06}$
GO immune system process	Up	$4.9x10^{-05}$	$7.1\ x10^{-4}$
KEGG			
hsa04612 Antigen processing and presentation	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
hsa05152 Tuberculosis	Up	$1.7x10^{-05}$	$1.6\ x10^{-4}$
hsa05164 Influenza A	Up	$2.2x10^{-05}$	$2.0\ x10^{-4}$
hsa05332 Graft-versus-host disease	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
hsa05140 Leishmaniasis	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
Immune signatures			
GSE7509 Genes down-regulated in immature dendritic cells	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
GSE2706 Genes down-regulated in comparison of unstimulated DC	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
GSE19888 Genes up-regulated in HMC-1 (mast leukaemia) cells	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
GSE34156 Genes down-regulated in monocytes	Up	$< 1.0x10^{-16}$	$< 1.0x10^{-16}$
GSE37416 Genes up-regulated in activated neutrophils	Up	$7.0x10^{-06}$	$9.7x10^{-05}$

Enriched inflammatory signature

Overall, the gene enrichment analysis performed by EGSEA showed a pro-inflammatory signature for all query data sets (e.g., Hallmarks, Gene Ontology, KEGG, and Immune Signatures; Table 2; Supplementary file 1)[190]. The pathways within the immune signature data set included IL6/JAK/STAT3 signalling, interferon gamma response, positive regulation of immune response,

and antigen processing and presentation. Specifically for the Immune Signature dataset, transcriptional changes pointed to the differentiation of immature immune cell types (i.e., immature dendritic cells and monocytes), as well as neutrophil and mast cell activation.

Consistent with the gene enrichment analyses, the differential tissue composition analysis based on our Bayesian inference model showed a positive association between overall immune cell abundance and treatment status (Fig. 3.2A). In the two approaches employed for differential tissue composition analysis, monocyte derived cells dominated the immune population within adipose tissue across the treated and naïve cohorts. Signatures of macrophages, monocytes and granulocytes were enriched by our model within the immune cell population in treated patients compared to naïve. This inference was partially consistent with that of the Cibersort-DirichletReg approach (i.e., for monocytes and macrophages; Fig. 3.2B). The latter approach uniquely identified an association involving CD4 memory resting, NK cells resting, and Mast cells resting. Although a significant enrichment of CD8+ T-cells in treated patients was not observed using our statistical model and the Cibersort-DirichletReg approach, a positive association appears to exist when observing the distributions of the estimated cell type proportions (Fig. S3.3). As expected, considering the absence of a robust adipocyte transcriptomic signature within the model, the fibroblast cell type appears to have captured the adipocyte transcriptomic profile (Fig S3.3). The differences observed in the average estimated proportions for immune cell types between Cibersort and our statistical method are in part due to the inclusion of non-immune cells (e.g., fibroblasts, endothelial and epithelial) in our model, while Cibersort models selectively estimate immune cells as composing the totality of the tissue.

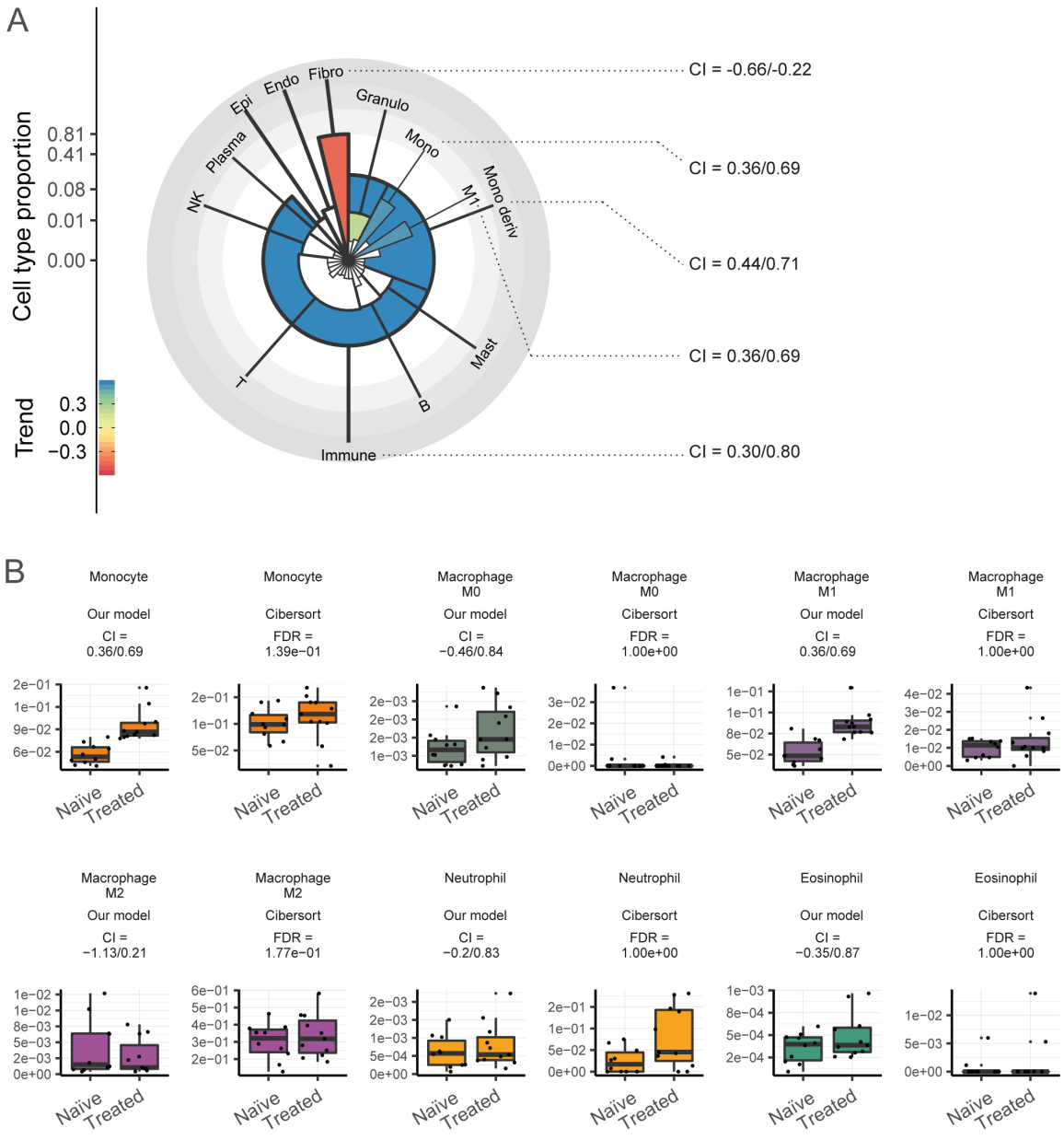


Figure 3.2: Differential tissue composition analysis. A — Polar plot representing the overall cell type abundance (i.e., radius dimension) and the significant associations with androgen deprivation therapy (i.e., white for non-significant associations). Cell types are labelled if more abundant than 1%. CI = 95% credible interval of the association. B — Boxplots of the inferred cell type proportions by our Bayesian probabilistic model and Cibersort, for the cell types that correspond

or are part of significantly differentially abundant cell type categories (e.g., the differentially abundant category "granulocytes" include eosinophils and neutrophils) between the two treatment categories (i.e., treated and naïve) according to our model. FDR = false discovery rate linked to an association being different non null.

Enriched obesity signature

The analysis of a previously published obesity transcriptional signature for adipose tissue[191] revealed a positive association with androgen deprivation treatment independent of body mass index (false discovery rate of $8.4x10^{-3}$; Fig. 3.3). Within the ten top ranked genes present in the obesity signature, the majority were linked to inflammation (Table S3.2), including Fc fragment of IgG binding protein (FCGBP), lysozyme (LYZ), chemokine ligand motif 10 (CXCL10), myeloid cell nuclear differentiation antigen (MNDA), toll like receptor 8 (TLR8), and a member of the STAT family (STAT1), which is activated by various ligands including interferon-alpha, interferon-gamma (IFNɣ), epidermal growth factor (EGF), platelet derived growth factor (PDGF) and interleukin 6 (IL6). The third top ranked gene (included in the obesity signature) is a key regulator of hormonal homeostasis (DHRS9), which is able to convert 3-alpha-tetrahydroprogesterone to dihydroxyprogesterone and 3-alpha-androstanediol to dihydroxyprogesterone in the cytoplasm[201]; also, it is a marker for regulatory macrophages[202]. Regulatory genes for calcium homeostasis were also present, including S100 calcium-binding protein A1 (S100A1) and stanniocalcin 2 (STC2), which regulate renal and intestinal calcium and phosphate transport[203].

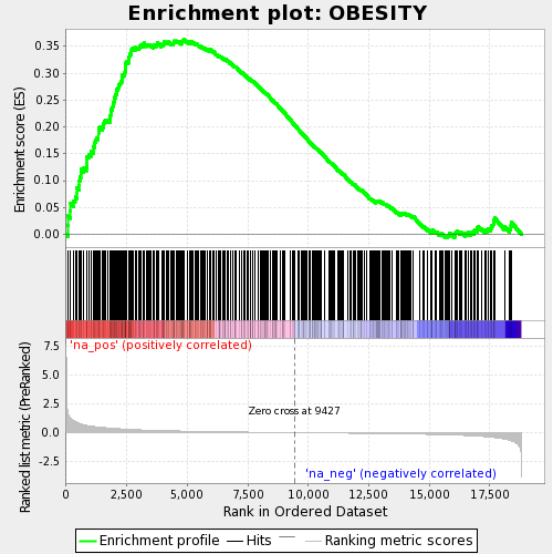


Figure 3.3: GSEA enrichment plot showing the significant enrichment of the obesity signature among the most differentially transcribed genes.

Conclusions

Obesity induces a persistent inflammatory and hormone rich tissue microenvironment that contributes to high risk disease[204,205]. Androgen deprivation therapy is a known cause of increased fat body mass[183–185], yet the cellular and molecular processes that are altered in association with androgen deprivation therapy, especially in the periprostatic adipose tissue microenvironment, have not been completely resolved. In this study we showed that androgen deprivation therapy based on a 6-month combination treatment of degarelix, bicalutamide, and abiraterone is associated with a pro-inflammatory adipose tissue microenvironment, as well as with altered obesity-related gene transcription linked with cholesterol and hormonal homeostasis (Fig. 3.1, 3.2 and 3.3).

Overall, the periprostatic adipose tissues of the treated and naïve cohorts were transcriptionally similar. This may indicate that some differences observed in tissue profiles reside within tissue infiltrating cells (e.g., immune cells). With gene integration (e.g., differential tissue composition and gene enrichment analyses), it was possible to extract global properties about differences between the two cohorts at the molecular level, despite a small number of single genes being significantly differentially transcribed. For example, both differential tissue composition and gene enrichment analyses pointed to an enrichment of infiltrating immune cell types within the tissue. Monocytes and macrophages had the greatest presence within the periprostatic adipose

tissue, compared with other immune cells. The abundance of these immune cell types was positively associated with androgen deprivation, suggesting their infiltration of the tissue, which is consistent with in vivo studies[206]. Macrophages have been shown to interact with adipose tissue in a paracrine manner, where TNF-α secretion from macrophages interferes with adipocyte insulin signalling and induces fatty acid lipolysis, which commences a vicious inflammatory cycle and contributes to insulin resistance[207]. Furthermore, an elevated blood monocyte count is an independent prognostic predictor for poor prostate cancer outcome in cancer-specific and overall survival studies[208,209].

This study suggests that the beneficial effect of androgen deprivation therapy may be partially counteracted by metabolic and inflammatory side effects in the adipose tissue encompassing the prostate. This may be particular pertinent when the primary tumour is in situ, as tumour response within the prostate appears less profound compared to that observed for metastatic disease[210,211]. Further studies will need to investigate the immune infiltration profile associated with androgen deprivation, as well as the potential impact of anti-inflammatory therapies on local tumour response.

Online methods and raw data

The code used to conduct the analyses is available at github.com/stemangiola/ADT_fat. The sequenced reads raw files are available at ega-archive.org with the identifier EGAS00001003286.

Supplementary material

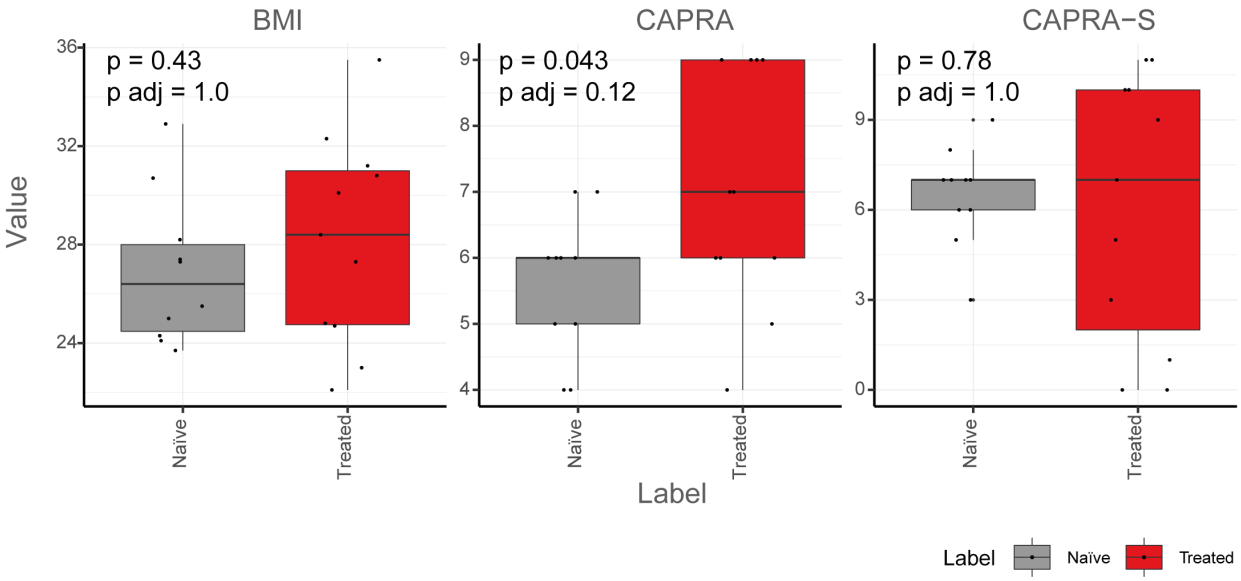


Figure S3.1. Boxplot of the body mass index (BMI), CAPRA and CAPRA-S risk scores for the two treatment categories (i.e., treated and naïve). Hypothesis tests were performed with t-test and the adjustment for multiple test correction was performed with the Bonferroni technique.

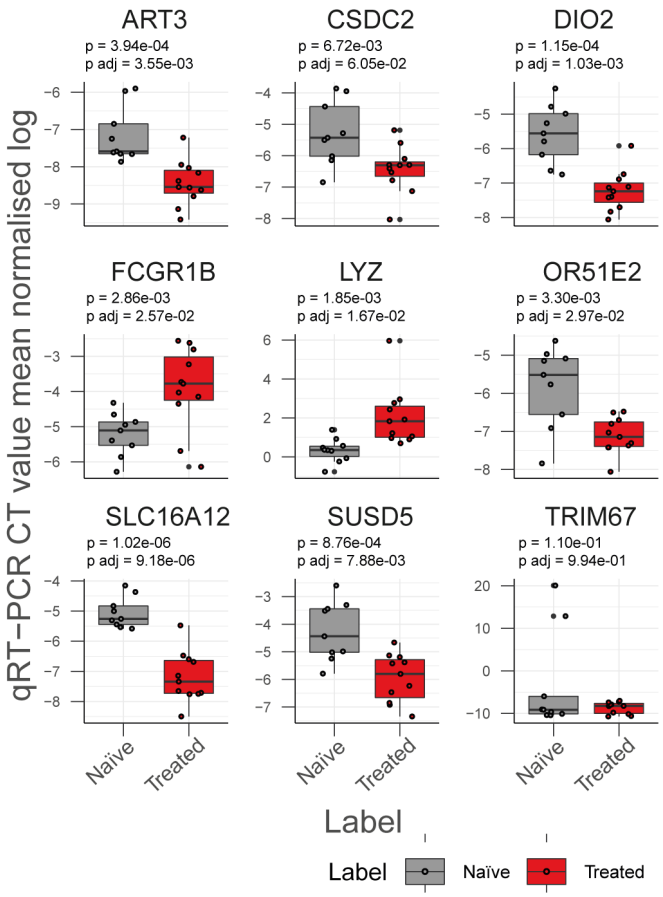


Figure S3.2 qRT-PCR validation. A total of seven out of nine probed genes did validated accordingly to the differential transcription analyses performed on RNA sequencing data. Hypothesis tests were performed with one side t-test and the adjustment for multiple test correction was performed with the Bonferroni technique.

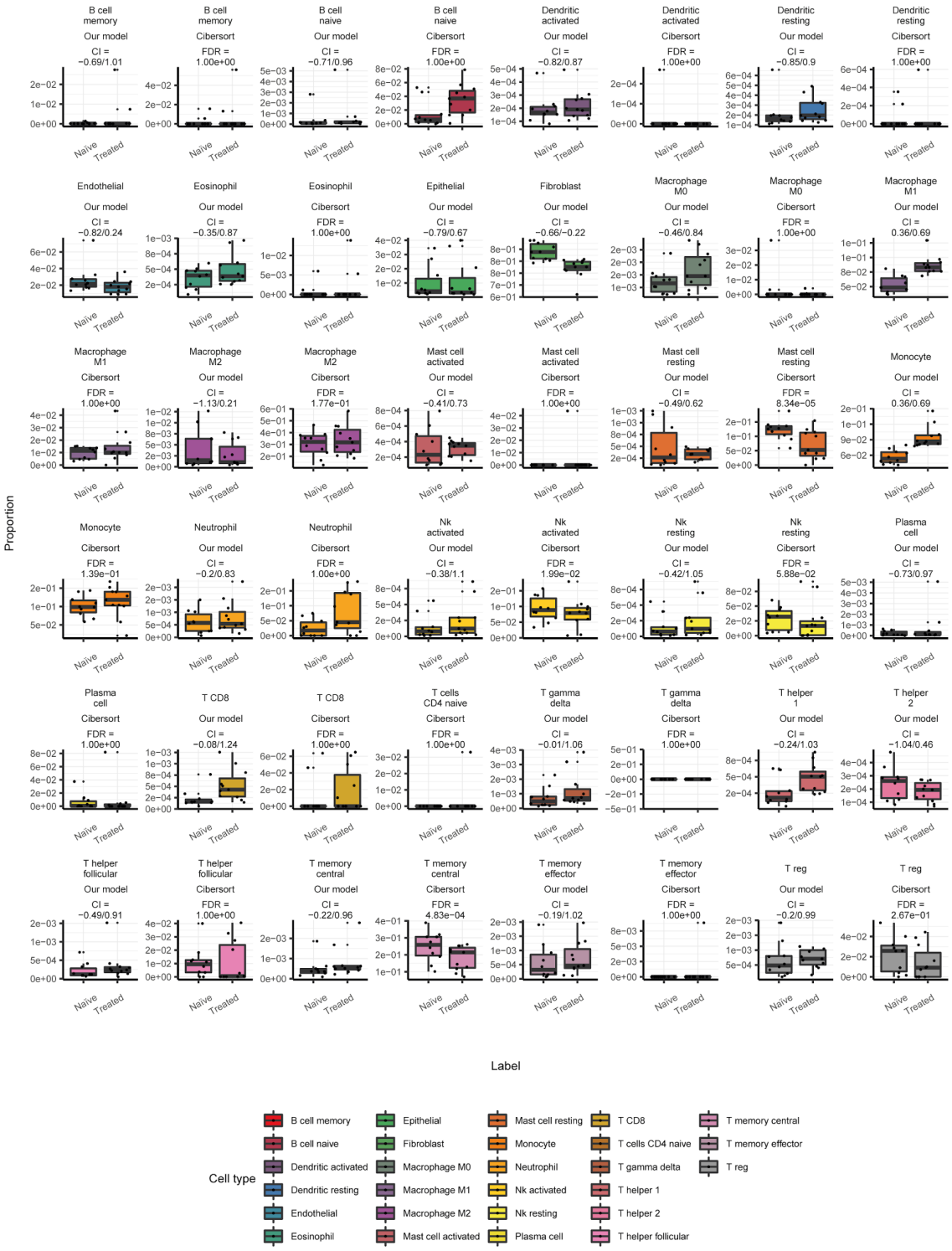


Figure S3.3. Boxplots of the inferred cell type proportions by our Bayesian probabilistic model and Cibersort.

Table S3.1. Attached as file

Table S3.2. Attached as file

Supplementary file 3.1. Attached as file

Probabilistic Bayesian inference model

$$(1)\ P(\sigma)\,P(\phi)\,P(\delta)\prod_{r=1}^{R}\prod_{p=1}^{P}P(\alpha_{r,p}|\delta)\prod_{p=1}^{P}\prod_{s=1}^{S}P(\pi_{p,s}|X_s,\alpha_p,\phi)\prod_{s=1}^{S}\prod_{g=1}^{G}P(Y_{s,g}|x_g,\pi_s,\sigma_s,\theta^*)$$

$$(1)\ P(Y|x,\pi,\sigma,\theta^*)\sim lognormal\,(x*\pi,\sigma)$$

$$(2)\ P(\pi|X,\alpha,\phi)\sim Dirichlet(\widehat{\widehat{Y}})$$

$$(3)\ P(\alpha|\delta)\sim Dirichlet\left(\left[\delta_1,...,\delta_k\right]\right)$$

$$(4)\ P(\sigma)\sim normal\,(0,0.1)$$

$$(5)\ P(\phi)\sim normal\,(1,...)$$

$$(6)\ P(\delta)\sim cauchy\,(1,2)$$

$$(7)\ \widehat{\widehat{Y}}=softmax(\widehat{Y})*\sigma$$

$$(8)\ \widehat{Y}=X\cdot\alpha$$

$$(9)\ \mathrm{softmax(z\ )}_j=\frac{e^{z_j}}{\sum_{k=1}^{K}e^{z_k}}\,for j=1,...,K$$

$$(10)\ \ \mathrm{logit}(p)=\log\left(\frac{p}{1p}\right)=\log(p)\log(1p)=\ \log\left(\frac{1}{p}1\right)$$

The parameter α represents the rates of change of each cell type category along the biological conditions. The parameter π represents the matrix of proportions for each cell type category and sample. The parameters σ, φ and δ define the noise model. The point estimate and credible intervals for both cell type proportions and trends of change are calculated from the posterior distribution.

CHAPTER 4

Context

Treatments for prostate cancer include surgery, radiotherapy and androgen deprivation therapy. The results achieved by the combination of these therapies can lead to variable and often temporary results, mainly due to cancer cells genetic plasticity and heterogeneity, leading to the emergence of drug resistance. In contrast, the non-cancerous portion of the tumour microenvironment, including immune cells and fibroblasts, is a genetically stable target that has a key role in cancer development. Improving our knowledge of the genetic and molecular interactions between cancer cells and other non-cancerous cell populations, both in primary or metastatic prostate cancer will provide new key insights in the biology of the disease and give new treatment opportunities. Here, analysing the transcriptional profiles of enriched cell types from the tumour microenvironment of primary prostate cancer, we provide an extensive landscape of intercellular transcriptional synergies between cancerous and benign cells, revealing the activation of key hallmarks of prostate cancer progression.

The interplay among cell types in the prostate tumour microenvironment contributes to the activation of key hallmarks

Introduction

Prostate cancer is a leading cause of male death worldwide. Treatments for prostate cancer include surgery, radiotherapy and androgen deprivation therapy. Due mainly to the high levels of heterogeneity and the fast adaptability nature of prostate cancer cells, such therapies lead to variable and often temporary results. For example, although androgen deprivation therapy confers biochemical effects for more than 90% of patients [212] and clinical effects for ~70% of patients [213], such effects can be temporary leading to recurrence within two-years in more than 50% of patients [214,215]. In contrast, the non-cancerous portion of the tumour microenvironment, including immune and stromal cells, is genetically more stable compared with cancer cells and plays a key role in cancer development. For example, the benign component is important for the tumour development[216,217], metastasis[218] as well as in drug resistance[219] and has potential as diagnostic tool[220]. Therefore, improving our knowledge on the genetic and molecular interactions existing between cancer cells and other non-cancerous cell populations, both in primary or metastatic prostate cancer will provide new key insights in the biology of the disease and give new treatment opportunities.

The cellular components of the tumoral mass and the extracellular matrix are referred as tumour microenvironment (TME). The tumour microenvironment has been extensively studied through *in vitro* and *in vivo* experiments, such as migration assays[66] and xenograft mouse models respectively[67]. These functional approaches are extremely valuable, as they can provide evidence on the role of specific cell types and/or genes in the development of the disease. However, due to the high financial and logistic burden associated with such approaches, they are limited in the throughput of targets that can be analysed in each given experiment. More recently, high-throughput cellular and molecular technologies, such as fluorescence-activated cell sorting (FACS) and low-input nucleotide sequencing, allow exploratory studies of a wider range of cell types and their gene expression patterns within single experiments. That is, fluorescence-activated

cell sorting allows the simultaneous isolation of several selected cell types from a tissue, which can be analysed at the transcriptome level with RNA sequencing.

For prostate cancer, several studies have adopted the combination of fluorescence-activated cell sorting and RNA sequencing. These studies mainly focused on the process of epithelial to mesenchymal transition (MET) [34,35]. Epithelial cell populations and/or stromal cells were enriched from prostate cancer tissue cores (n = 20 and n = 6 respectively) and differential transcriptional analysis was performed. These studies demonstrated the power of cell type enrichment coupled with high-throughput molecular analyses; however, an integrative investigation of the transcriptional changes happening among epithelial, immune and stromal cells is still pending. In this study, adopting (i) cell-specific enrichment; (ii) RNA sequencing; and (iii) a novel statistical model for continuous differential transcription analyses; we performed a hypothesis-generating study focusing on the emergence of molecular signatures for hallmarks of prostate cancer, resolving the contribution of cancer, immune and stromal cells.

Methods

Tissue sampling and processing

Following prostatectomy (n=13 patients), a tissue biopsy was collected from the prostate tumour site using a four millimetre punch, conditionally to histopathological verification[132,189]. If not otherwise specified, all procedures were carried out at 4°C. Tissue blocks were washed in Phosphate-buffered saline (PBS) solution for 2 minutes and minced for 2 minutes with a scalpel. Homogenised tissue was added to a solution (total volume of 7 ml) composed by of 1 mg/ml collagenase IV, 0.02 mg/ml DNase 1, 0.2 mg/ml dispase. The solution was serially digested at 37°C at 180 rpm, through three digestion steps of duration 5, 20 and 20 minutes duration, with the final 3 minutes dedicated to sedimentation at 0 rpm. After each digestion step, the supernatant was aspirated and filtered through a 70 μm strainer into a pre-chilled tube, diluting the solution with 15 ml of 2% bovine serum PBS to quench the enzymatic reaction. The resulting cumulative solution was then centrifuged at 1500 rpm for five minutes, with the supernatant collected and the cell pellet resuspended into 1 ml 2% PBS-serum prior to labelling (Fig. S4.1).

Antibody labelling, flow cytometry and cell storage

In order to identify and enrich for epithelial, fibroblast, T-cell, and myeloid cell populations, the antibodies EpCAM-PE, CD31-APC; CD90-PerCP; CD45-APC-Cy7, CD3-BV711; CD16-PB were used respectively, at a dilution factor of 1/40 with a labelling time of 30 minutes. To remove unbound antibody, the labelled cell solution was diluted with 5 ml of 2% calf serum PBS and centrifuged at 1500 rpm for five minutes. The supernatant was removed, the cell pellet was resuspended in 2 ml of 2% PBS-serum and the PE fluorescence conjugated viability dye was added at a dilution of 1/40 from the stock. Due to the heterogeneity of the prostate tissue, the cell sorting strategy utilised a robust three stage design: (i) A series of shared physical parameter plots, based on cell size and morphology to remove cell debris and cell doublets or clusters; (ii) the fluorochrome selection for cell specific surface marker immunophenotype ; and (iii) a second cell-type specific side-scatter gate based on expected cell morphology. The four enriched cell populations collected in 1.5 ml tubes and stored in dry ice immediately after collection, before centrifugation and permanent storage at -80°C.

RNA extraction, library preparation and RNA sequencing

RNA extraction was performed in two batches (comprising 6 and 7 patients, for a total of 24 and 28 samples respectively) on consecutive days. The two patient batches included a balanced distribution of Gleason score and days passed from tissue processing (in order to eliminate time-dependent methodological biases). The RNA extraction was performed using the miRNeasy Micro Kit (Qiagen; Cat # 217084), according to manufacturer's protocol. Briefly, cell pellets were lysed with QIAzol lysis reagent, treated with chloroform and centrifugation carried out to separate the aqueous phase. Total RNA was precipitated from aqueous phase using absolute ethanol, filtered through the MinElute spin column and treated with DNase I to remove genomic DNA. The RNA bound columns were washed with the buffers RWT and RPE before eluting the total RNA with 14µl of RNase-free water. RNA estimation was carried out using Tapestation (Agilent).

Whole transcriptome analysis on low input total RNA samples (up to 10ng) was carried out using SMART-Seq v4 Ultra Low Input RNA Kit (Clontech), according to manufacturer's protocol. The first-strand cDNA synthesis utilised 3’ SMART-Seq CDS Primer II A and the SMART-Seq v4 Oligonucleotide and the cDNA amplification was carried out on Thermocycler using PCR Primer II A and PCR conditions: 95°C 1min, 12 cycles of 98°C 10sec, 65°C 30sec and

68°C 3min; 72°C 10min and 4°C for ever. The PCR-amplified cDNA was purified using AMPure XP beads and processed with the Nextera XT DNA Library Preparation Kits (Illumina, Cat. # FC-131-1024 and FC- 131-1096) as per the protocol provided by the manufacturer.

The whole transcriptome analysis on input total RNA samples (10 – 100ng) was carried out using Truseq RNA Sample Preparation Kit v2. The poly-A containing mRNA was purified using oligo-dT bound magnetic beads followed by fragmentation. The first strand cDNA synthesis utilised random primers and second strand cDNA synthesis was carried out using DNA Polymerase I. The cDNA fragments then underwent end repair process, the addition of a single 'A' base, and ligation of the RNA adapters. The adaptor ligated cDNA samples were bead-purified and enriched with PCR (15 cycles) to generate the final RNAseq library.

The SMART-Seq v4 RNA and Truseq RNA libraries were sequenced on an Illumina Nextseq 500 to generate 15-20 million 75 bp paired-end reads for each sample. The batch effect due to sequencing runs was minimised by pooling all 52 libraries and carrying out three sequential runs on a Nextseq500 sequencer.

Sequencing data quality control, mapping and gene counting

The RNA sequencing quality for each sample was checked using the Fastqc [134]. Reads were trimmed for custom Nextera Illumina adapters, low quality fragments and short reads were filtered out from the pools according to default settings. All remaining reads were aligned to the reference genome Hg38 using the STAR aligner with default settings[136] (Fig. S4.1). The quality control on the alignment was performed with RNA-SeQC[221]. For each sample, the gene transcription abundance was quantified in terms of nucleotide reads per gene (read-count) using FeatureCounts[137] with the following settings: isPairedEnd = T, requireBothEndsMapped = T, checkFragLength = F, useMetaFeatures = T. All sequenced reads that did not align to the reference human genome were aligned against bacterial and viral reference genomes using kraken[222] with default settings.

Statistical inference

Changes of transcriptional levels along the post-prostatectomy CAPRA-S risk score[223] were inferred using regression of the raw gene counts along a pseudo-continuous covariate, independently for each cell type (i.e., epithelial, fibroblast, T-cell and monocyte-derived cells). The CAPRA-S risk score is a combination of: (i) concentration of blood prostate serum antigen

(PSA); (ii) presence of surgical margin (SM); (iii) Gleason score; (iv) presence of seminal vesicle invasion (SVI); (v) the extent of extracapsular extension (ECE); and (vi) the lymph node involvement. The RNA extraction batch was used as further covariate. Due to the absence of publicly available model for non-linear monotonic regression along a continuous covariate, a new non-linear model was implemented. This model is based on the simplified Richard's curve[224] (Eq.1), but re-parameterised to improve numerical stability (Eq. 2).

$$(1)\ GL(X, \alpha, \beta, \kappa) = \frac{k}{1 + e^{-(\alpha + X\beta)}}$$

$$(2)\ GLA(X, y_\circ, \beta, \eta) = \frac{y_\circ \left(1 + e^{\eta\dot{\beta}}\right)}{1 + e^{\eta\dot{\beta} - X\beta}}$$

Where $y_\circ$ represents the intercept on the y axes, η represents the point of inflection on the x-axis, β represents the matrix of coefficients (i.e., slope coefficients, without the intercept term), $\dot{\beta}$ represents the coefficient of interest (i.e., main slope), and k the upper plateau of the generalised sigmoid function.

Bayesian inference was used to infer the values of all parameters of the model. The probabilistic framework Stan[193] was used to encode the joint probability function of the model, partitioning the transcriptomic data set in blocks of 5000 genes to decrease the analysis run-time. This Bayes model is based on a negative binomial distribution (parameterised as mean and dispersion) of the raw gene counts. In order to account for diverse sequencing depths across samples and a possible asymmetry of transcriptional changes (i.e., the overall transcriptional output along the covariate on interest is not zero) a normalisation parameter has been added to the negative binomial expected value. This parameter is identified by the regularised horseshoe[225] prior over the covariate of interest. The role of this prior is to impose a sparsity assumption on the gene-wise transcriptional changes; that is, most genes are not differentially transcribed. The precision of the negative binomial distribution is conditional to the expected value[146] following a generalised sigmoid function (Eq. 1). The overall distribution of the gene intercepts follows a gamma probability function. The statistical model is defined by the following joint probability density.

$$(3)\ P(\gamma)\,P(\delta)\,P(\sigma)\,P(\eta)\,P(\xi)\,P\left(\dot{\beta}|\xi\right)\prod_{r=2}^{R}P(\beta_r|\sigma)\prod_{g=1}^{G}P\left(y_{\circ g}|\gamma',\gamma''\right)\prod_{g=1}^{G}\prod_{s=1}^{S}F$$

$$(4)\ Y = NB_{\text{ reparam 2 log }}\left(\hat{Y},\delta,\omega\right)$$

$$(5)\ \hat{Y} = GLA\left(X,y_{\circ},\beta,\eta\right)$$

$$(6)\ \omega = GL\left(\hat{Y},\alpha,\kappa\right)$$

$$(7)\ \beta_1 \sim regHorseshoe\left(...\right)$$

$$(8)\ \beta_{2..R} \sim normal\left(0,\sigma\right)$$

$$(9)\ y_{\circ} \sim gamma\left(\gamma' * \gamma''^{-1},\gamma''^{-1}\right)$$

$$(10)\ \alpha_1,\kappa,\sigma,\delta \sim normal\left(0,1\right)$$

$$(11)\ \alpha_0 \sim normal\left(0,10\right)$$

$$(12)\ \eta,\gamma \sim normal\left(0,5\right)$$

Where Y represents raw gene counts, $\hat{Y}$ represents the expected values of gene counts and X represents the design matrix (with no intercept term and with scaled covariates). The regression function also includes β which represents the gene-wise matrix of factors (i.e., slopes excluding the intercept term), y^{o} and η which represent the gene-wise y-intercept and the inflection point of the generalised reparametrized sigmoid function (Eq. 2), while γ represent the hyperparameters of y^{o}. Other parameters of the negative binomial function are δ, which represents the normalisation factors; and ω, which represents overdispersion. The regularising prior (for imposing the sparsity assumption) over the covariate of interest β^{dot} (first column of β) is defined by the hyperparameter list ξ[225], while σ represents the standard deviation of the factor not of interest. The hyperprior of the overdispersion parameter ω is defined by α representing the intercept and slope, and κ representing the upper plateau of the generalised sigmoid function (Eq. 1).

Gene annotation

Each gene (g) was considered well fitted by the model if at most 3 samples had counts outside the 95th percentile of the generated quantities (according to posterior predictive checks standards[226,227]). Among the well fitted genes, those for which the 0.95 credible interval of the

posterior distribution of the factor of interest $\dot{\beta}_g$ did not include the value 0 were labelled as differentially transcribed. In order to interpret the inflection points over the CAPRA-S covariate in a biologically meaningful way: the inflection point was adjusted to the log-scale, and the covariate was converted to the natural scale.

$$(13)\ log\{\ (GLA\,(...)) = \frac{log\,(\ \text{upper plateau of}\ GLA\,(...))}{2}$$

$$(14)\ log\,(y^{\circ}) + log\left(1 + e^{\eta\dot{\beta}}\right) log\left(1 + e^{\eta\dot{\beta}\dot{X}\dot{\beta}}\right) = \frac{log\,(y^{\circ}) + log\left(1 + e^{\eta\dot{\beta}}\right)}{2}$$

$$(15)\ \dot{X} = \frac{\dot{\beta}\eta log\left(e^{\frac{y^{\circ}}{2}}\sqrt{e^{y^{\circ}\eta} + 11}\right)}{y^{\circ}}$$

This point was calculated (in log space) as the value of the x-axis $\dot{X}$ at half distance between zero and the upper plateau of the generalised reparametrized sigmoid function (Eq. 2). Genes were functionally annotated with gene ontology categories[194] using BiomaRt[228]. Furthermore, genes were functionally annotated with the protein atlas database[165] for identifying those that interface with the extracellular environment, encoding for membrane and secreted proteins.

Results and discussion

Quality control

After library preparation, the amplified cDNA sequences showed the expected nominal solution concentration and distribution in fragment length, except for monocyte-derived samples, which concentration was lower compared to other cell types, and the proportion of larger fragments (> 600bp) was higher, indicating a less efficient enzymatic process (Fig. S4.2). After filtering and trimming, the sequenced reads of all samples showed a mean Phred quality score above 28. The sequencing output ranged from 70 million (for sample 2C; Fig. 4.1) to 1 million reads (for sample 4C; Fig. 4.1). The proportion of reads uniquely mapped to the Hg38 reference genome was > 80% for most samples (n = 41), with an exonic rate in the range of 50 to 70%. Overall, myeloid samples were characterised by a lower sequencing output, and mapping coverage compared to the other three cell types.

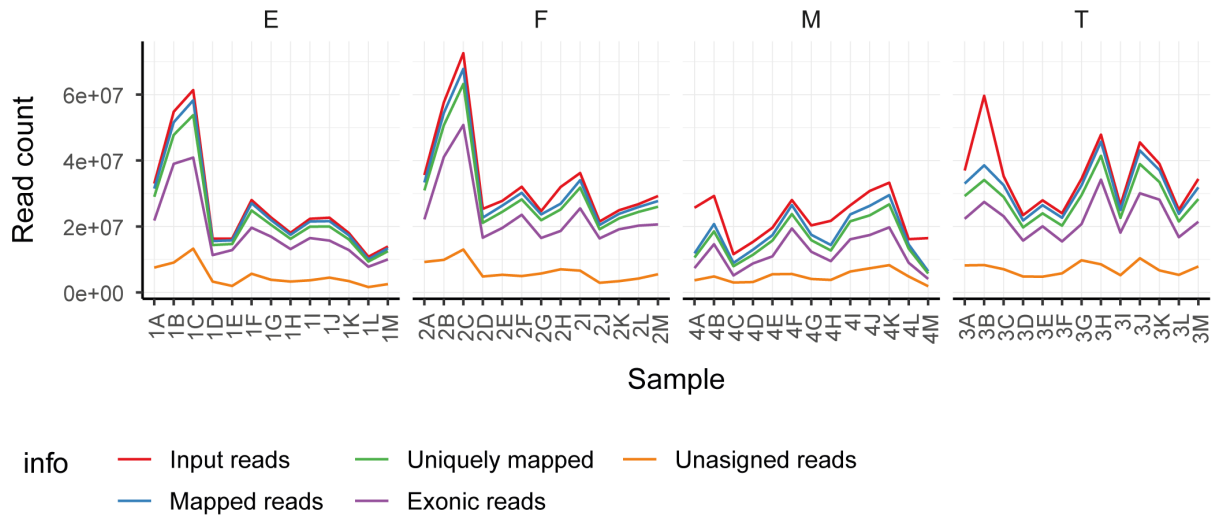


Figure 4.1: Descriptive statistics of sequencing output and mapping for each sample (n = 52) grouped by cell type.

For myeloid samples, a positive association can be observed between CAPRA-S risk score and (i) number of mapped reads, as well as (ii) number of mapped reads to exons (Fig. 4.2). Such association is partly due to an abundant amount of bacterial and viral genomic content for surrounding benign and low grade cancers (Fig. 4.3).

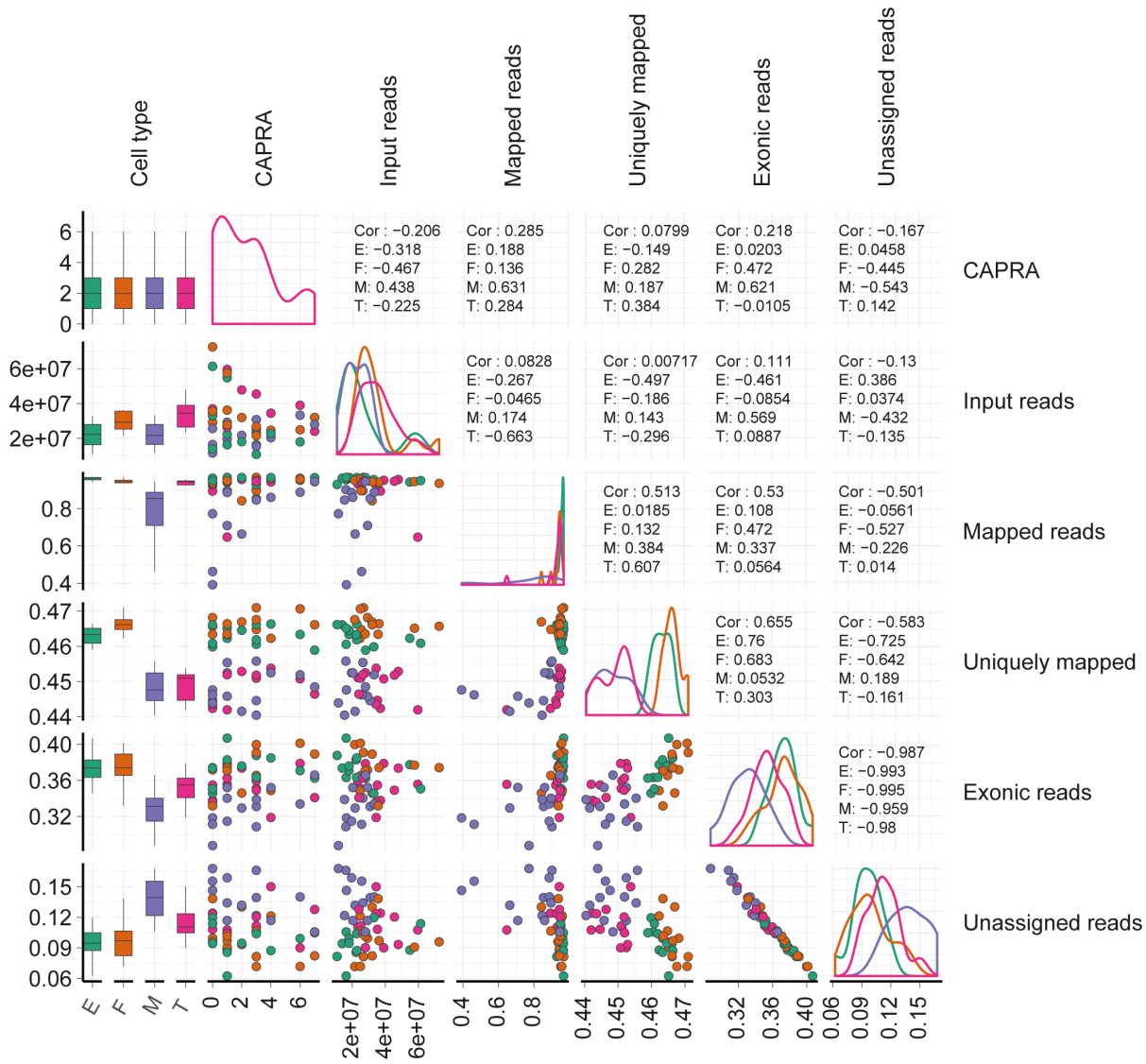


Figure 4.2: Pair plot showing the relations among sequencing and mapping statistics, stratified by cell type and CAPRA-S risk score.

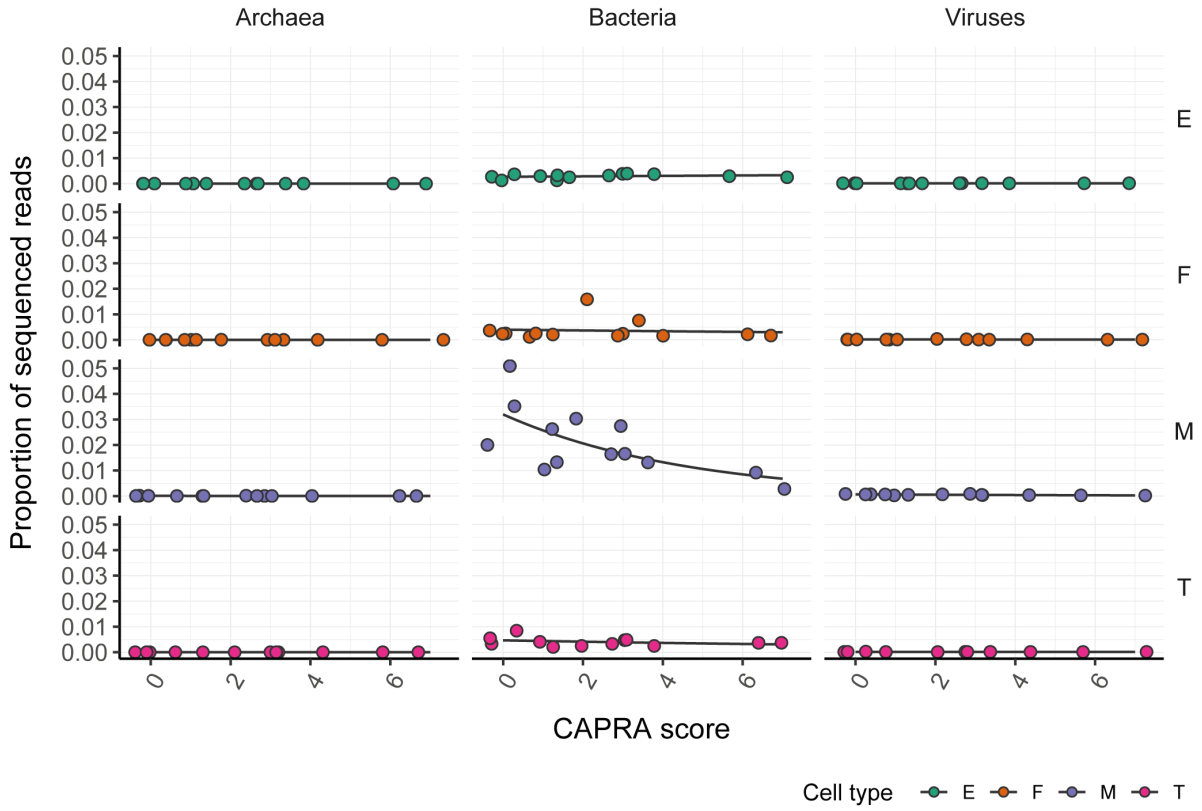


Figure 4.3: Proportion of bacterial and viral sequences in each sample, and its association with CAPRA-S risk score, cohort across cell types.

Differential transcription analyses

On average across the four cell types, 40% of genes were quarantined as having 0 sequenced reads in more than half of samples. As expected, the analysis of the two principal components of the transcriptional profiles of our cohort grouped by cell type shows a clear association with CAPRA-S risk score (Fig. 4.4), especially strong for epithelial and fibroblast cells. Following statistical inference, an average of 10% of genes were quarantined following the posterior distribution check (Table 4.1), with such proportion associated with the overall sequencing coverage and mapping rate of each cell type.

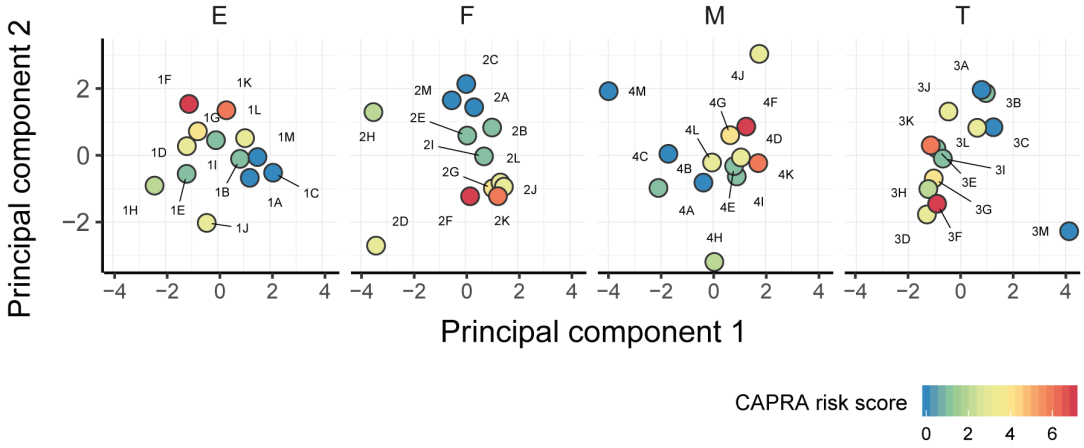


Figure 4.4: Multidimensional scaling (MDS) plot of all samples, grouped by cell type. The colour coding is linked to the CAPRA-S risk score

A total of 1,626 Genes were overall identified as differentially transcribed across the four cell types (Table 4.1). Surprisingly, epithelial cells showed the smallest association with the CAPRA-S risk score in terms of number of differentially transcribed genes; myeloid cells conversely showed the largest association (Table 4.1). The inferred trends of the top differentially transcribed genes for each cell type are showed in Fig. S4.3.

Table 4.1: Summary statistics for the differential transcription analysis. CI = confidence interval; DT differentially transcribed. "Of which" refers to the gene selection relative to the category adjacent on the left.

Cell type	Tot genes	Quarant. sparse	Quarant. CI	DT (up/down)	Of which interface	Of which cancer genes (consistent)	Of which PC genes (consistent)
E	21.618	5.408	189	171 (139/32)	80	48 (73%)	29 (80%)
F	21.510	7141	651	267 (156/111)	97	24 (68%)	11 (64%)
T	21.716	8807	540	288 (195/93)	83	35 (70%)	19 (79%)
M	22.507	13836	2695	900 (827/73)	261	37 (84%)	14 (71%)

As expected, the distribution of adjusted inflection points (i.e., point at which the trend has the steepest increase or decrease) are roughly unimodal and centred within the CAPRA-S risk score interval 0-7 (Fig. 4.5), with the exception of the myeloid cell population showing a bimodal distribution for the negatively associated genes.

The differential transcription analysis based on the algorithm edgeR and the patient stratification pivot CAPRA-S risk score of 2 failed to identify any differentially transcribed genes, due to limited sample size and multiple testing correction.

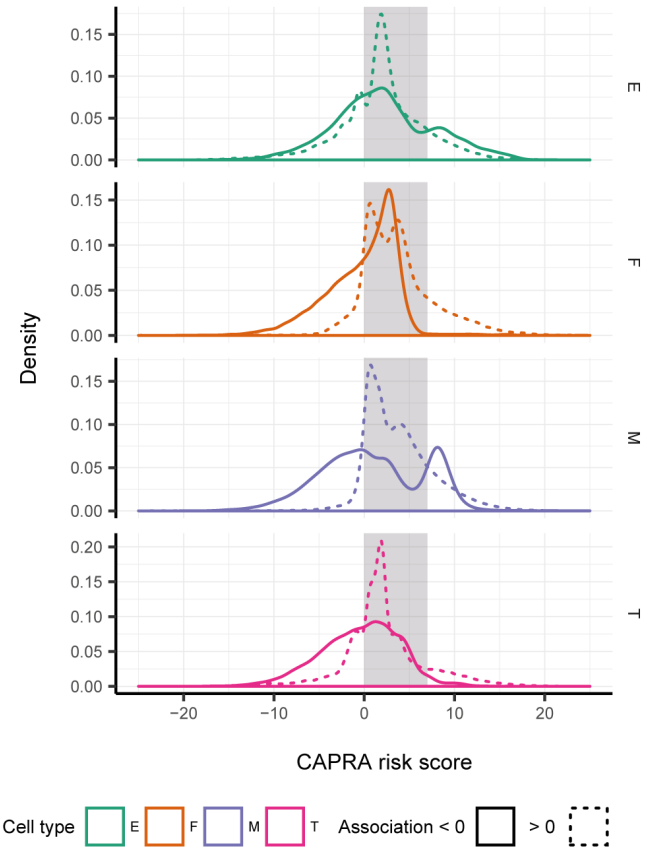


Figure 4.5: Distribution of the adjusted inflection points along the CAPRA-S risk score, across cell types. The grey shade represents the range. of CAPRA-S risk score in the 13-patient cohort. Inflection points outside that range represents exponential-like trends that did not approach a plateau within the CAPRA-S risk score range.

Gene annotation

On average across the four cell types, 35% of genes identified encode for proteins that interface to the extracellular space (i.e., secreted or transmembrane; Table 4.1). An average of 33% of these genes have been previously identified as cancer related; and of those, an average of 51% have been previously described as prostate cancer related genes. For all cell types, most cancer genes have a direction of change consistent with the literature (20 vs. 7 for epithelial; 14 vs. 7 for fibroblasts; 18 vs. 8 for T-cells; 28 vs. 8 for myeloid cells).

Of all differentially transcribed genes that encode for proteins that interface with the extracellular space (transmembrane or secretory proteins), a total of six recurring hallmarks were identified (from grouping analogous gene ontology annotations; GO[194]): (i) epithelial/cancer cell growth; (ii) angiogenesis; (iii) cancer cell migration (as integration of tissue remodelling and epithelial to mesenchymal transition); (iv) osteogenesis; (v) hormone/fat homeostasis; and (vi) macrophage-fibroblast interplay.

The pro- and anti-inflammatory balance evolves during tumour progression

The balance between inflammatory and anti-inflammatory signals evolves along the disease progression (i.e., CAPRA-S risk score range 0-7). The balance between pro- and anti-inflammatory signals evolves along the disease progression. While the pro-inflammatory transcriptional signature is prominent in the initial stage of the cancer progression (i.e., CAPRA-S risk score 0-2), it appears to decrease in the more advanced stages. On the contrary, the anti-inflammatory signature significantly expands in the late stage of the disease (i.e., CAPRA score > 2; p-value 0.015). Overall, the majority of the inflammatory signal it targeted toward the recruitment of monocytes/macrophages[28,229,230,231,232,233,234,235,236] (labelled with an asterisk in Fig. 4.6).

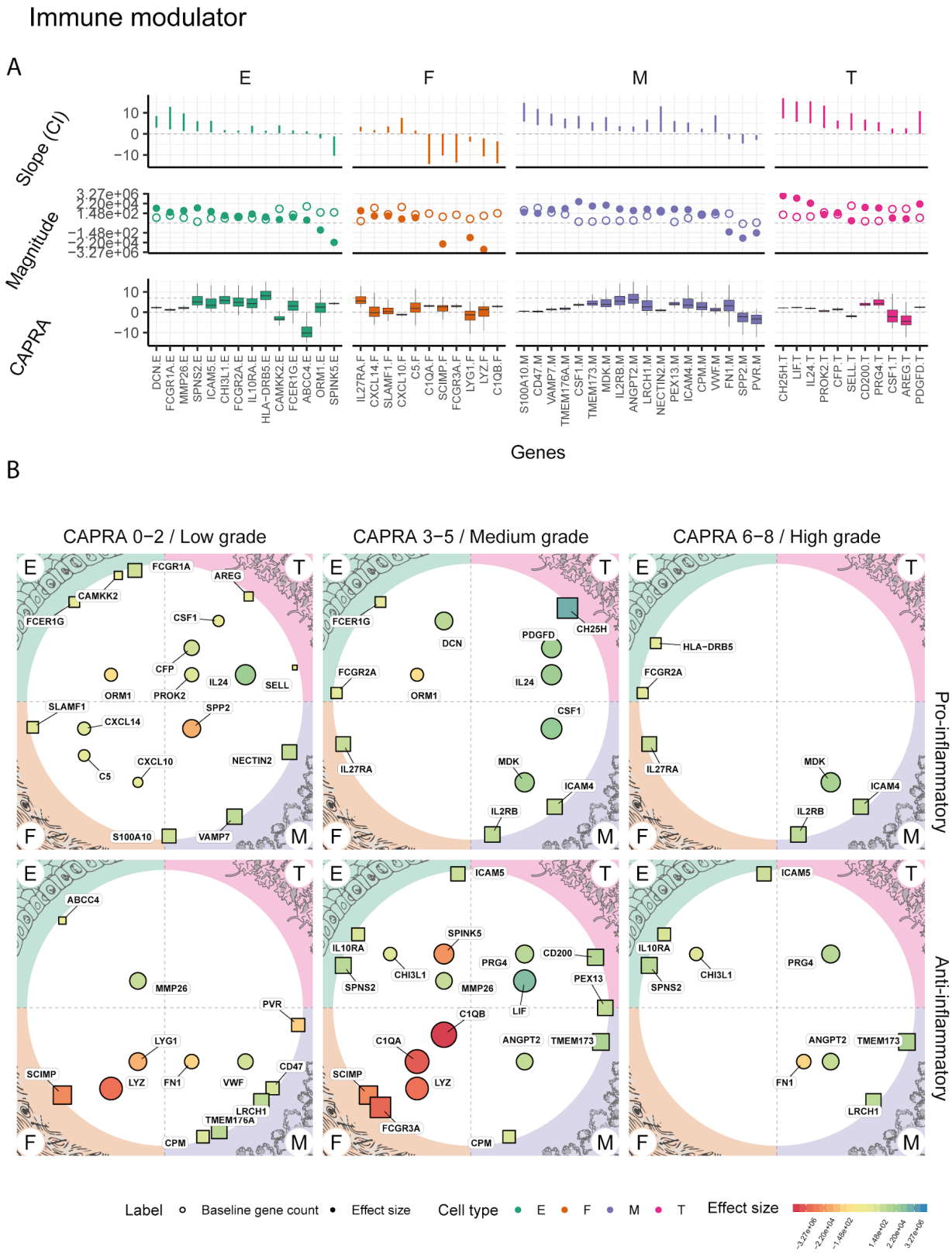


Figure 4.6: A - regression descriptive statistics. Slope (CI) represent the credible interval of the

angular coefficient; Magnitude represents the fold change (full dot) and the baseline transcription abundance (intercept, empty circle); CAPRA represents the value of the inflection point estimate. B -Map of the secretory (represented as circles) and transmembrane (represented as squares) protein coding genes, that are differentially transcribed across the four cell types (i.e., epithelial, fibroblasts, myeloid cells and T-cells), that contribute to the immune modulation. Genes are grouped according the range value of the inflection point (at what stage of the disease a transcriptional change happens; CAPRA-S risk score 0-2, 3-5 and 6-8), and according to their role (Pro- or anti-inflammatory). The colour coding and the size are linked with the fold change of the expected transcription value (between the baseline CAPRA-S risk score = 0, and CAPRA-S risk score = 7) on a logarithmic scale.

In a more advanced stage of the disease (i.e., CAPRA-S risk score 3-5), the sustained inflammatory signature is mainly maintained by the myeloid population with contribution of the T-cell and epithelial cell populations, still having monocyte/macrophages as main target[237,238,239]. The transcriptional alterations (labelled with an asterisk in Fig. 4.6) for monocyte and macrophage recruitment includes the genes IL2RB[240–242], ICAM4[243,244], DCN[245,246] and MDK[247,248] in myeloid cells and CSF1[28], PDGFD[249] in T-cells. In the most advanced stage of the disease (i.e., CAPRA-S risk score 6-8), several myeloid chemotactic genes maintain their upregulation. Interestingly, epithelial cells upregulates the gene HLA-DRB5 which is HLA-DR is an MHC class II cell surface receptor normally encoded by the human leukocyte for antigen presentation[250,251]. It is tempting to speculate that together with the upregulation of FC receptors, epithelial cells might promote an immune mimicry activity.

The epithelial pro-migratory phenotype is promoted by three complementary hallmarks

A synergy among all four cell types promotes an epithelial pro-migratory signature across three routes: direct modulation of epithelial-to-mesenchymal transition (EMT), a pro-fibrotic stimulus and a matrix degradation activity. A transcriptomic signature[252] for epithelial-to-mesenchymal transition phenotype is enriched among the epithelial differentially transcribed genes (adjusted p-value 0.022; Fig. S4.2). Modulatory genes of epithelial-to-mesenchymal transition are synergically activated along the disease progression by epithelial, myeloid and T-cell population[253,254,255].

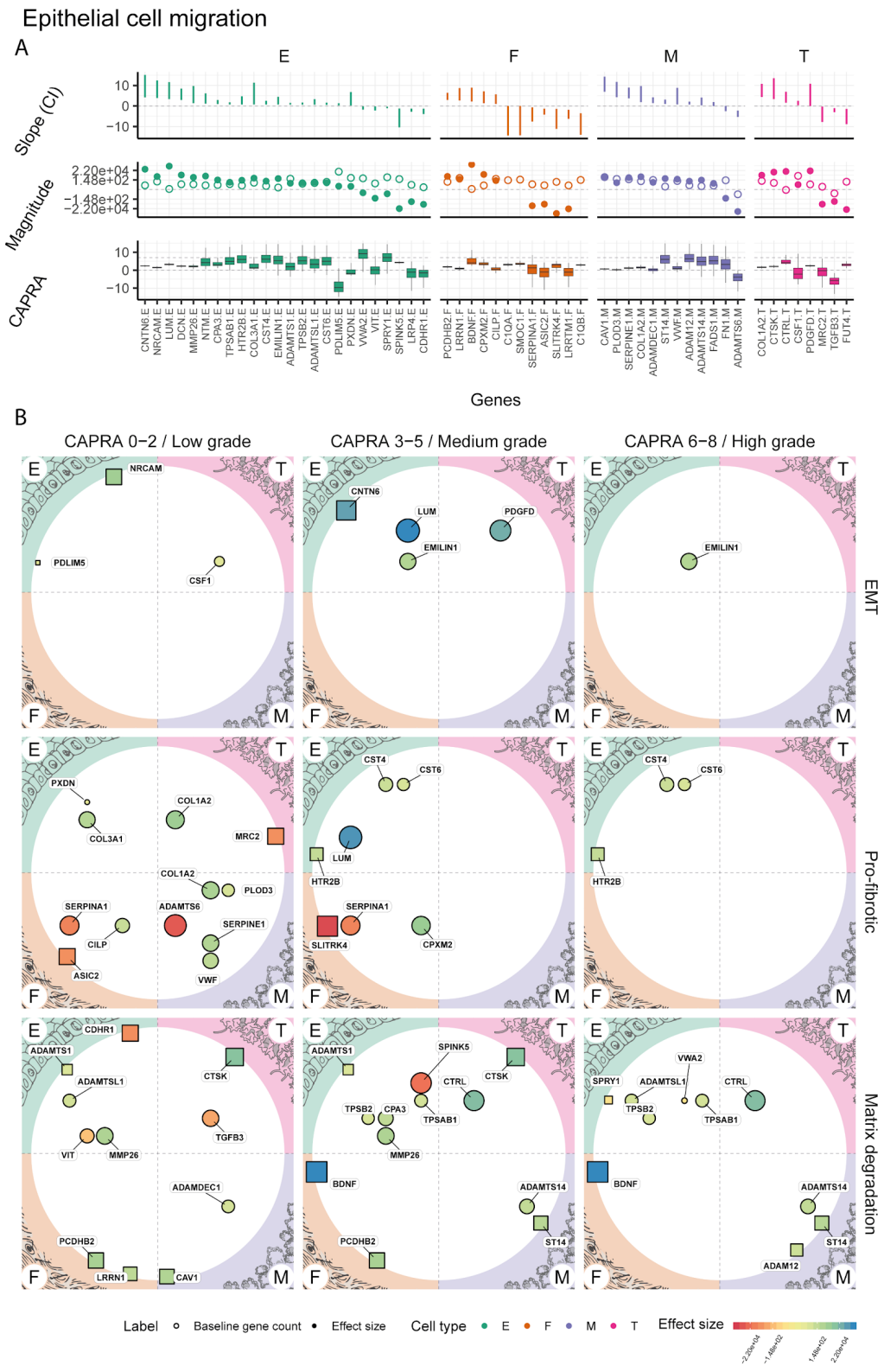


Figure 4.7: A - regression descriptive statistics. Slope (CI) represent the credible interval of the

angular coefficient; Magnitude represents the fold change (full dot) and the baseline transcription abundance (intercept, empty circle); CAPRA represents the value of the inflection point estimate. B - Map of the secretory (represented as circles) and transmembrane (represented as squares) protein coding genes, that are differentially transcribed across the four cell types (i.e., epithelial, fibroblasts, myeloid cells and T-cells), that contribute to the epithelial cell migration. Genes are grouped according the range value of the inflection point (at what stage of the disease a transcriptional change happens; CAPRA-S risk score 0-2, 3-5 and 6-8), and according to their role (directly promoting epithelial-to-mesenchymal transition, pro-fibrotic or promoting matrix remodelling). The colour coding and the size are linked with the fold change of the expected transcription value (between the baseline CAPRA-S risk score = 0, and CAPRA-S risk score = 7) on a logarithmic scale.

The balance between a fibrotic and matrix remodelling/degradation signatures evolves during tumour progression. The induction of fibrosis appears to be an early event in cancer progression; on the contrary, the anti-fibrotic signalling is altered by all four cell types consistently along the CAPRA-S risk score. In the initial stage of disease progression (i.e., CAPRA-S risk score 0-2), several differentially transcribed genes encoding protein that interface the extracellular environment are connected with collagen production, including COL3A1 in epithelial and COL1A2 in T-cell and myeloid cell populations; and PLOD3 which is upregulated in myeloid cells and is an essential gene for collagen catabolism[256].

Both the upregulated genes SERPINE1[257,258] from myeloid cell population and PXDN[259–261] from epithelial encode for fibrotic factors downstream of TGFβTGFB1; similarly, the upregulated gene VWF[262,263] from myeloid cells and CILP[264,265] from fibroblasts contribute to a fibrotic environment. In more advanced stages of the disease (i.e., CAPRA-S risk score >2) both the cysteine proteases inhibitors CST4[266] and CST6[267] are upregulated by epithelial cells. Such cell type and fibroblasts also upregulate the a pro-fibrotic factors HTR2B[268] and CPXM2[269] respectively.

The matrix remodelling and degradation is mainly driven by the upregulation of proteases, such as MMP26[270], ADAMTS1[271] and ADAMTSL1[272]. in the first stages of the disease (i.e., CAPRA-S risk score 0-2), and TPSB2[273], TPSAB1 and CPA3[274] in more advanced stages. Furthermore, epithelial cells upregulate an inhibitor of a family of serine proteinases involved in extracellular matrix remodelling, SPINK5[275–277]. Myeloid cells upregulate ADAMDEC1 and

ADAMTS6 during the initial stage of the disease, and ADAM12 and ADAMTS14 and ST14 in later stages[278,279]. T-cell in advanced disease upregulate the potent trypsin protease CTRL. Fibroblasts promote tissue remodelling by increasing mobility with the upregulation of several genes linked with neural plasticity and connectivity, including the genes PCDHB2, LRRN1, BDNF, SLITRK4 and ASIC2[280–282].

The synergy among cell populations to promote angiogenesis evolves during disease progression

The angiogenesis signalling appears to be sustained along the whole disease development. In the initial stage of cancer development (i.e., CAPRA-S risk score 0-2) the pro-angiogenic signal is not enriched compared with the anti-angiogenic (or repression of pro-angiogenic) signalling (adjusted p-value 1), on the contrary of later stages of the disease (i.e., CAPRA-S risk score >2; adjusted p-value 0.03; Fig. 4.8). Along the disease progression, a gene alteration signature of platelet recruitment, that promote angiogenesis and endothelial cell migration, is expressed in synergy by the immune cell types. T-cells upregulate PRG4[283] and PDGFD[284], while myeloid cells upregulate CHPT1[285] which favours the secretion of platelet activation factor, and VWF, a potent coagulation factor[286–288]. Interestingly, fibroblasts seem to negatively modulate the myeloid role in platelet activation by upregulating ADAMTS13 (Willebrand factor-cleaving protease), which cleaves extracellular VWF, promoting at the same time angiogenesis[289–291].

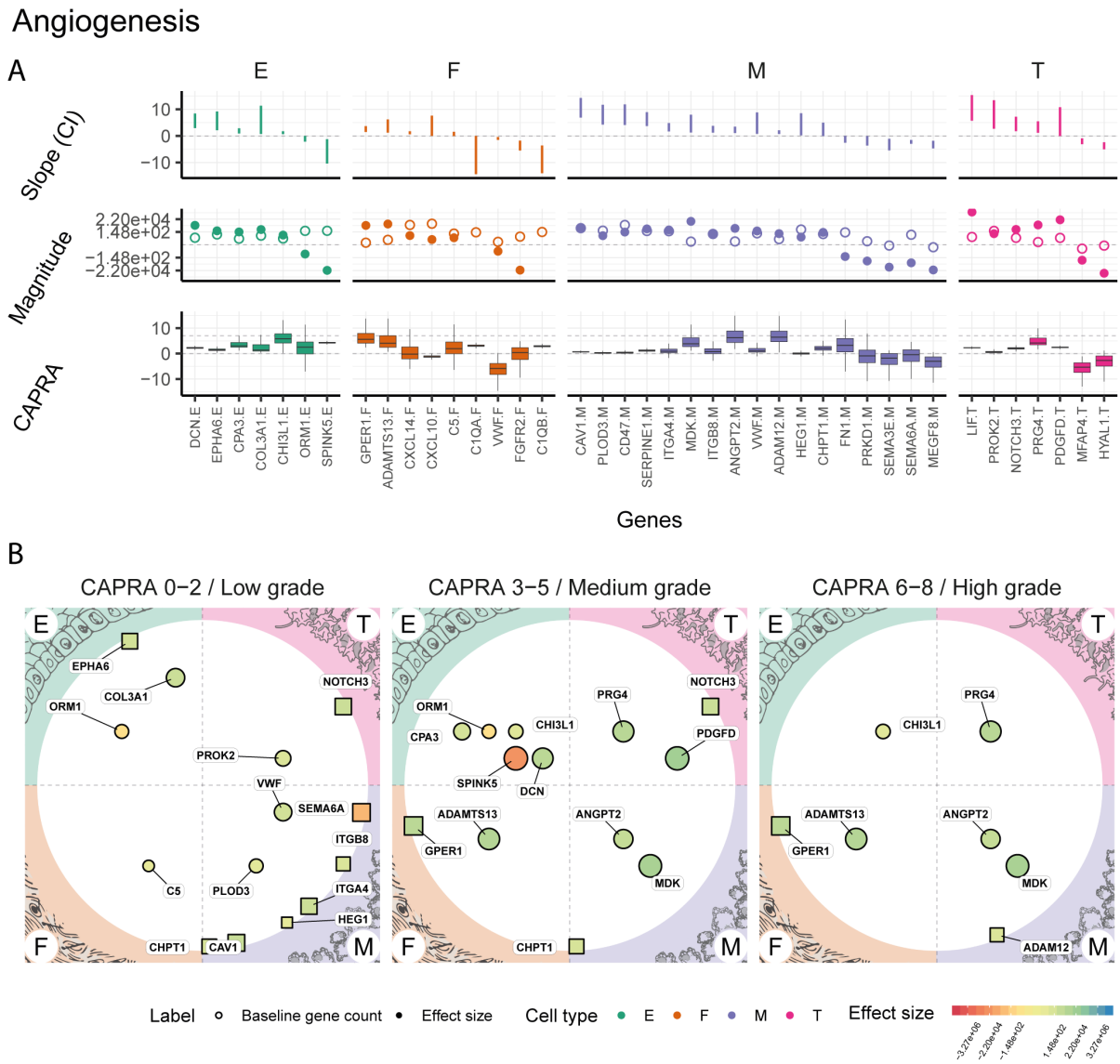


Figure 4.8: A - regression descriptive statistics. Slope (CI) represent the credible interval of the angular coefficient; Magnitude represents the fold change (full dot) and the baseline transcription abundance (intercept, empty circle); CAPRA represents the value of the inflection point estimate. B - Map of the secretory (represented as circles) and transmembrane (represented as squares) protein coding genes, that are differentially transcribed across the four cell types (i.e., epithelial, fibroblasts, myeloid cells and T-cells), that contribute to angiogenesis. Genes are grouped according the range value of the inflection point (at what stage of the disease a transcriptional change happens; CAPRA-S risk score 0-2, 3-5 and 6-8). The colour coding and the size are linked with the fold change of the expected transcription value (between the baseline CAPRA-S risk score = 0, and CAPRA-S risk score = 7) on a logarithmic scale.

At a later stage (i.e., CAPRA-S risk score >2), the epithelial cell population regulate a total of five secreted protein genes including the further downregulation of ORM1, the downregulation of a key inhibitor of serine proteases involved in angiogenesis SPINK5[275], and the upregulation of three pro-angiogenic genes including CHI3L1[292], DNC[246] and CPA3[293,294], a proteinase normally secreted by mast cells that favours micro-vessel formation. Myeloid cells upregulate the receptor ADAM12 known in cancer biology to indirectly promote angiogenic phenotypes[295], and upregulate the secreted ANGPT2 that is a known pro angiogenic factor that favours metastases and an immune target in clinics[296].

Hormonal homeostasis

Along the disease progression, our data suggest a synergy in hormonal molecules/cholesterol production and secretion, and hormonal sensing that is exerted by all four cell types. The most recurring metabolite that is linked with differentially transcribed genes (encoding for secreted of transmembrane proteins) across the four cell types is cholesterol (labelled with an asterisk in Fig. 4.9). Overall however, the set of interface genes (i.e., secreted and transmembrane) for cholesterol production and secretion was not enriched compared with those for cholesterol metabolism.

In an initial stage of the disease (i.e., CAPRA-S risk score 0-2), several differentially transcribed genes encode for transmembrane proteins with a role in increasing the extracellular cholesterol concentration, including CYP51A1 which silences cholesterol degradation enzymes[297], STARD3NL linked to cholesterol transport[298,299], and SC5D that drives cholesterol and steroid synthesis at the cell membrane[300], which gene is upregulated by both T-cells in early stages and by myeloid cells in more advanced stages (i.e., CAPRA-S risk score >2). At late stages of the disease, the epithelial secretory CES3 promotes the production of free cholesterol from cholesteryl esters for steroid hormone production[301,302]. The myeloid transmembrane protein gene CH25H have a key role in converting cholesterol into its hydrated form[303].

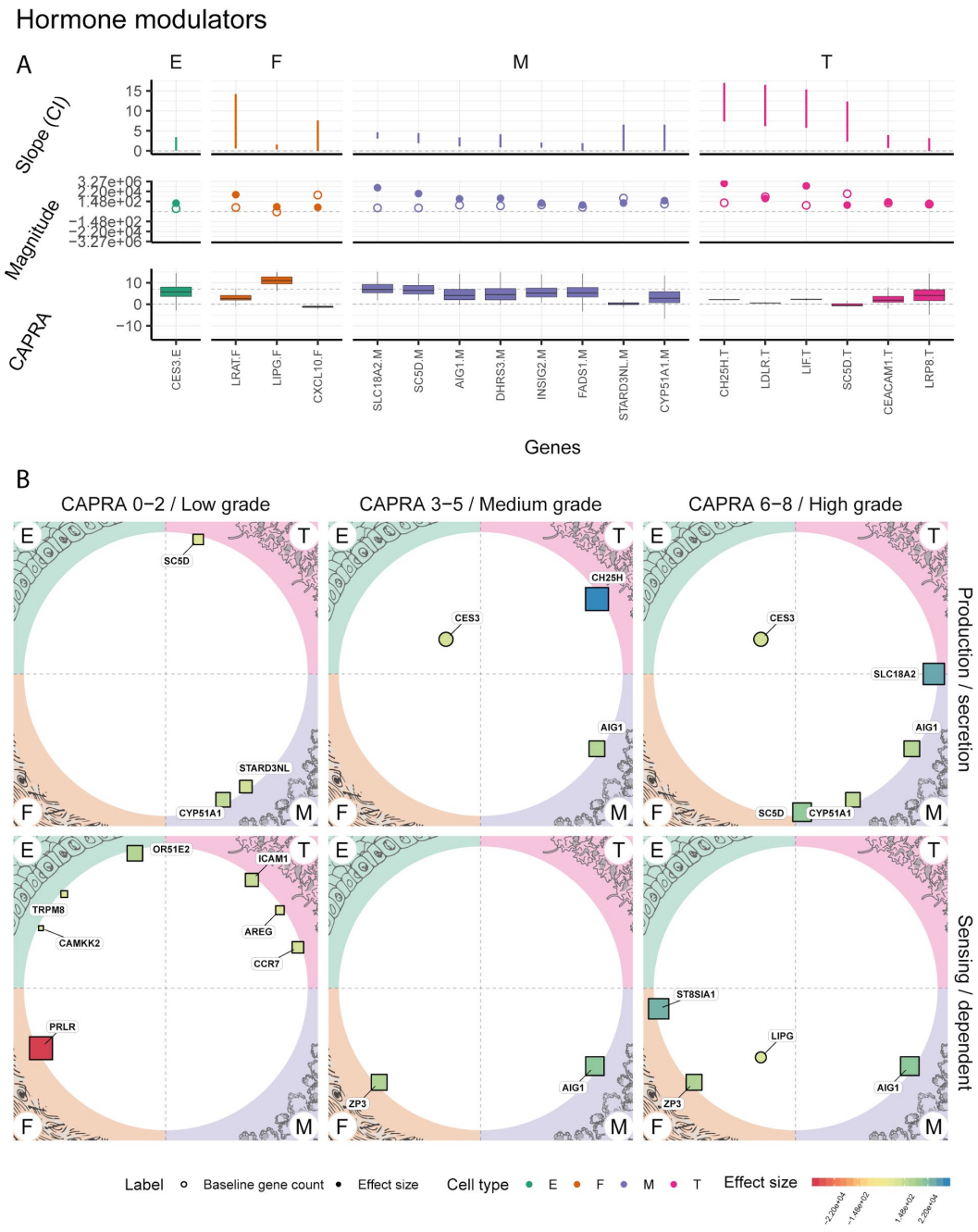


Figure 4.9: A - regression descriptive statistics. Slope (CI) represent the credible interval of the angular coefficient; Magnitude represents the fold change (full dot) and the baseline transcription abundance (intercept, empty circle); CAPRA represents the value of the inflection point estimate. B -Map of the secretory (represented as circles) and transmembrane (represented as squares) protein coding genes, that are differentially transcribed across the four cell types (i.e., epithelial, fibroblasts, myeloid cells and T-cells), that contribute to hormonal homeostasis. Genes are grouped

according the range value of the inflection point (at what stage of the disease a transcriptional change happens; CAPRA-S risk score 0-2, 3-5 and 6-8), and according to their role (production/secretion or sensing/signal transduction of hormone related molecules). The colour coding and the size are linked with the fold change of the expected transcription value (between the baseline CAPRA-S risk score = 0, and CAPRA-S risk score = 7) on a logarithmic scale. A related signature of hormone sensing and signal transduction is present among the differentially transcribed genes[304],[305],[306,307].

Conclusions

The tumour microenvironment has been shown to be important for cancer progression [216,217] and treatment resistance[219], as well as being a reliable diagnostic marker in cases of metastatic disease. The inference of differential transcription on enriched key cell types (i.e., epithelial, fibroblasts, myeloid and T-cells), with a statistical model that is able to infer the stage of the disease at which a transcriptional change takes place, allowed to map the contribution of different cancerous and benign cell types to the development of hallmarks of prostate cancer through the disease progression.

Both the library preparation and sequencing efficacy were heterogeneous across the four enriched cell types. In particular, myeloid cells have both a lower sequencing output and lower mapping rate. The latter aspect (and possibly the first) appears to be caused by the abundance of bacterial sequence contaminants. This was to be expected as the enriched myeloid cell population includes several phagocytic immune cells types. The ability of macrophages and other phagocytic immune cell populations to probe bacterial and viral nucleotidic material could be of use in future studies. We observed a decrease in bacterial content for high grade cancer, this may be due to the permanent inflammation within the tissue, and whether such sterile inflammatory environment is of clinical importance remains to be established.

Intercellular heterogeneity is presented in the analysis of principal components, where T-cells cluster poorly accordingly to CAPRA-S risk score, compared to the other three cell types (Fig. 4.4). The diversity and complexity of the cell phenotypes within the enriched T-cells might play a role in such poorly correlated sample distribution. A possible approach to resolve intra cell-type diversity is to integrate sample-wise cellular composition as a confounding covariate in the regression model. Such composition can be inferred from the whole tissue RNA abundances using

algorithms such as ARMET (included in the present thesis) and Cibersort [69]. However, the relatively low sample size of the present study (i.e., n = 13) does not allow for such integration.

The differential transcription analysis was performed with a novel approach. The use of a model able to detect changes in transcript abundance along a (pseudo-)continuous covariate (i.e., CAPRA-S risk score) allows: (i) to avoid the arbitrary choice of thresholds for converting the risk score into binary risk categories (i.e., low- or high-risk); and (ii) to provide the confidence information about the stage of the disease at which a gene-wise change in transcription most likely happen. For example in case of a discrete change of transcription, our statistical model would provide high confidence about the inflection point around a particular coordinate of the CAPRA-S risk score, while in case of a smooth transcriptional change along the CAPRA-S risk score the confidence would be spread along the CAPRA-S risk score. Although it is important to not regard the lack of confidence as information (e.g., continuous change along the covariate of interest), the previous example become more relevant as the confidence of the transcriptional change increases (e.g., when the posterior distribution of the slope parameter do not include the zero value; i.e., for differentially transcribed genes).

The common parameterisation of the generalised sigmoid function has shown to be numerically unstable for transcriptional data mainly for the frequent lack of the upper plateau (parameter *k* in Eq. X) in differential transcription trends (e.g., for exponential changes along the covariate of interest). In this study, we proposed a new parameterisation of the generalised sigmoid function that allows robust detection of monotonic changes for transcriptional data.

Overall, we observed a smaller number of changes in transcriptional abundance for more advanced stages of the disease. Such bias could be caused by the smallest representation of high-grade samples compared to lower-grade, thus providing less information to base the inference on. Across the four cell types, the myeloid cell population provided the longest list of differentially transcribed genes, compared with the other cell types (900 compared with an average of 242). On one hand, such difference could be just due to the overall lower coverage of the myeloid cell samples that can cause higher noise and more false positives. On the other hand, such difference in number of differentially transcribed genes could be simply due to the richness in biological response of myeloid cells that include key modulators of the immune response (e.g., myeloid, macrophage, and dendritic cells).

Our study showed that the majority of differentially transcribed genes encoding for proteins that interface the extracellular environment (e.g., transmembrane and secreted) target the chemotaxis and modulation of macrophages, which are known to drive a wound healing environment[30] . We were able to identify two axes of interaction between monocytes and macrophages with other cell types including epithelial cells and fibroblasts, creating positive chemotactic and modulatory loops. Furthermore, our results show that monocyte macrophages may be involved in the hormonal homeostasis of the extracellular matrix. While cholesterol is known to be an inflammatory messenger with a key role in monocyte and macrophage activity and is a precursor of membrane lipids needed for phagocytosis and cytotoxic activity. However, a side effect of such metabolism could be enrichment of free cholesterol available in the extracellular matrix, that could be used by cancer cells for driving the synthesis of testosterone. In previous studies, the abundance of monocytes in circulation have been shown to be prognostic for poor prognosis[208].

Interestingly, the epithelial cell population altered the transcription of transmembrane protein genes with an inflammatory role is expressed by immune cells. If confirmed, such gene regulation could allow an immunoglobulin poor environment in the vicinity of such cell, still allowing a chronic inflammation within the prostate tissue. Alternatively, such activity could also promote immune escape[308]. The tissue remodelling and fibrosis represented an important recurrent hallmark in our data. Although TGFβ was not among the DE genes for epithelial cells, several downstream and related genes were differentially regulated. The only cell population where we could observe an upregulation of TGFβ3 was myeloid, although did not reach significance. On the contrary T-cells significantly downregulate TGFβ2 in advanced stages of the disease (i.e., CAPRA-S risk score 3-5). Regarding angiogenesis, an important axe seems to be platelet recruitment and activation. This axe is sustained by myeloid, epithelial and T-cells, while fibroblasts appear regulate an opposite signature of anticoagulation. Overall, this study proves the utility of being able to map transcriptional alterations to specific cell populations as well as to cancer developmental stages. The present study gives an extensive landscape on the possible synergies existing among cancer, immune and stromal cells, and represents an useful hypothesis generating resource for future studies.

Supplementary material

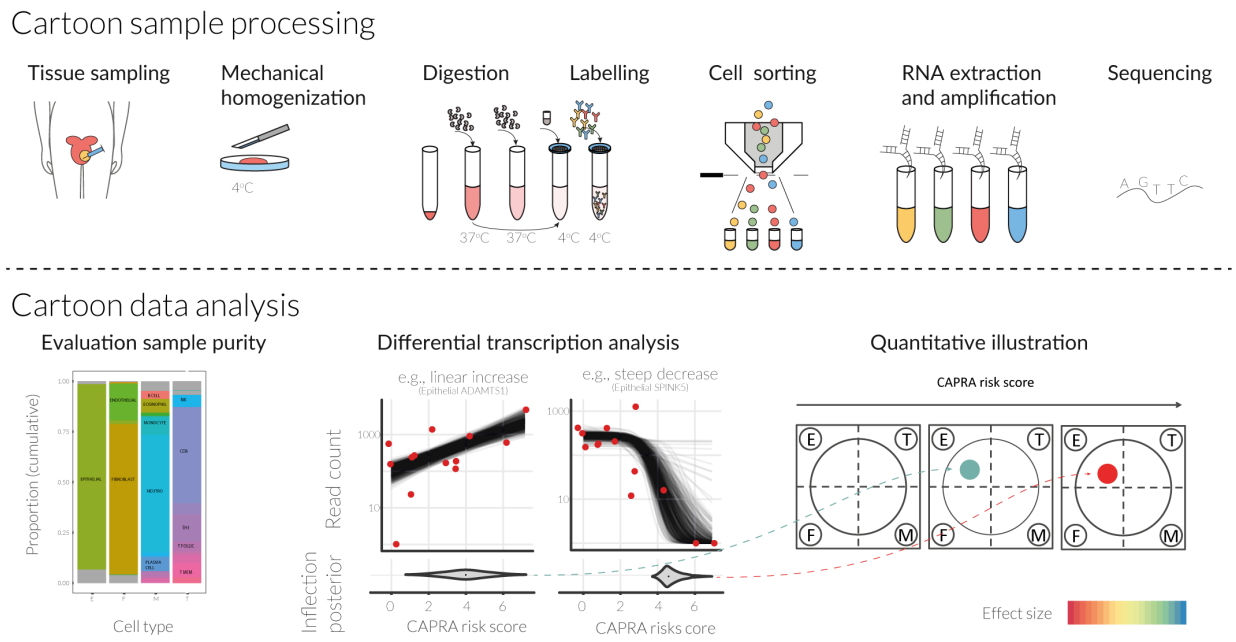


Fig. S4.1. Diagram of the tissue processing and data analysis

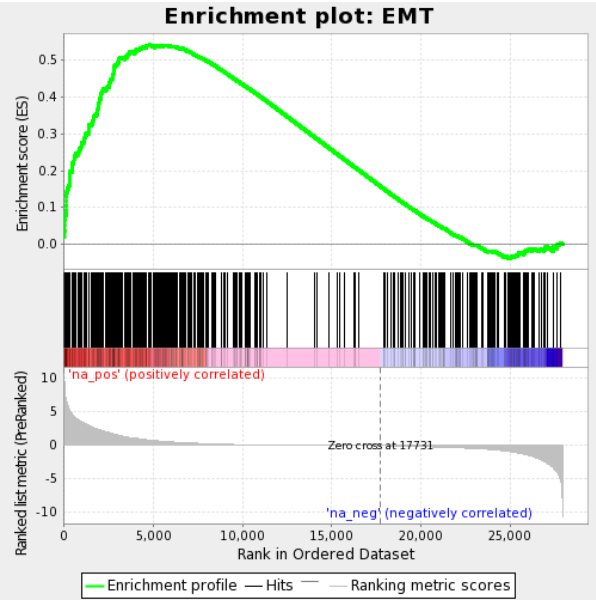


Fig S4.2. GSEA analysis of an epithelial to mesenchymal transition signature[252] for the epithelial samples, along the CAPRA risk score.


ANKRD7
AVPR1A
CNN1
CNTN6
DACT3
DCN
DDIAS
HEYL
LOC102723704
LRRC10B
LUM
MARC1
NCCRP1
NRCAM
OR51E2
PATE2
RASD1
SIM2
SPAG17
TARP
ALG10
ARHGAP31
BMT2
C1QB
CP
CYP1B1-AS1
ELOVL6
FBXW9
GGN
HBA2
HBB
ICAM4
IRF8
NELL1
PCDHB2
RPS6KA6
TMEM178A
ZNF45
ZNF500
ZSCAN20
ATP1B3
CNIH1
FAM209B
FKBP4
GOSR1
GPD1L
JADE1
KRT10
LRRC47
PEA15
PMF1
PMP22
RBM42
RBM7
RPS4X
SEC61G
SEM1
TGIF1
TXNDC9
VPS18
ARL4A
CCNF
CD80
CDC42EP2
CENPL
CH25H
DLEC1
DNAJC18
FOSL1
IL24
LDLR
LIF
LINC01353
LOC101928370
LOC728975
NRARP
PODXL
RAVER2
RNASE4
WEE1
CAPRA_TOTAL
y

Fig. S4.3. Trend plots of the top differentially transcribed genes, The red dots represent data, the error bars represent the posterior confidence interval of each data point. The grey lines represent all the possible trends according to the model.

CHAPTER 5

Context

Proportional data is often relevant in statistical inference, either because data is directly acquired as such or because acquired count data underlies a multinomial process, which is driven by the proportions of its components. For example, the analysis of tissue composition via the direct observation of the different cell types within, represents a multinomial observation driven by cell type proportions; similarly, the inference of tissue composition through molecular make-up of the tissue models the components of a tissue as proportions. Observing/inferring the changes in the composition of a mixture along a factor of interest can be highly informative. For example, observing/inferring the changes in tissue composition along cancer grade can improve the knowledge about the role of specific cell types in cancer progression. Identifying driving changes in a mixture is a challenging task, due to the apparent inverse correlation emerging among components solely as result of the sum-to-one compression, which is characteristic of all proportional data. For example, two or more components of a system that remain unchanged in count size along a factor of interest in the real space, can appear as having non-zero changing rates in proportional space due to the change of a third component. Publicly available algorithms fail to model such aspects, and thus to correctly identify drivers of change. Here, we present a probabilistic generative Bayes model that can identify drivers of change from a simplex space under the parsimony assumption that two or more component of the system remain unchanged in the real space.

Inference of extrinsic changes in simplex space under parsimony assumption

Introduction

Proportional data is often relevant, either because data is directly acquired as such or because acquired count data underlies a multinomial process that is driven by the proportions of its components. For example, the analysis of tissue composition via the direct observation of the different cell types within represents a multinomial observation driven by cell type proportions; similarly, the inference of tissue composition through molecular make-up of the tissue, models the components of a tissue as proportions. A vector of proportions that sums to one is referred to as a simplex. More formally, a simplex is a vector of reals of size K with K-1 degrees of freedom (Eq. 1).

$$(1)\quad \mathrm{S}^d = \{(x_1, \ \ldots \ , x_d) : x_i > 0 \, (i = 1, \ \ldots \ , d) \, , x_1 + \ \ldots \ + x_d < 1\}$$

Geometrically, a simplex of size K can be represented as a point in a K-1-dimensional polytope. For example, a simplex of K=3 correspond to 2-dimensional triangle surface (Fig. 5.1).

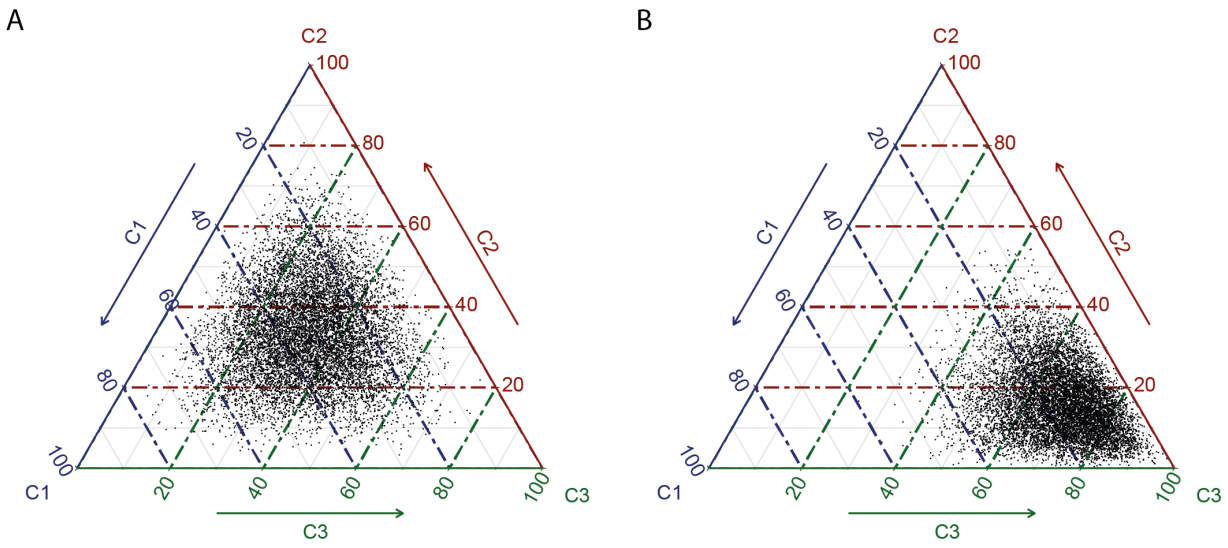


Figure 5.1: Geometric visualization of a three components (K = 3) simplex sample space in two dimension, with a degree of uncertainty associated with each observation. Each point in the plot represent a simplex (e.g., [0.2, 0.4, 0.4]). On the left, all three components are equally represented, on the right the C3 component is over-represented.

Inferring changes in the composition of a mixture (i.e., the trends of change of the proportion of its components) along a factor of interest, can inform about the role of each component within. For example, the change in cellular composition of the tumoral mass along cancer grade can inform about the importance of any cell types for cancer progression. Performing such inference presents several challenges. For example, the simplex space is characterised by non-symmetric and heteroscedastic noise with extremes at the plateaus 0 and 1. Also, the independence assumption of the simplex components is complex and has multiple definitions[309]. Furthermore, a linear increase in real space appears as curved in the simplex space [192], therefore a suitable regression function must be chosen. Another point of complexity is the two-ways interpretation of the simplex space[309]. The extrinsic interpretation considers the existence of a real system where changes regarding the absolute quantity of its components (e.g., a cell population increase of 1 million units in a tissue area/organ; Fig. 5.2A) are inferred. Of such a system, only a small proportion can be usually observed at any given time; the observations are therefore the result of a multinomial process, that is driven by the proportions of its components. Therefore, even if a system could be numerically described on an unbounded positive natural scale, the proportional observations must be described in a simplex space. On the contrary, the intrinsic interpretation of the simplex space considers such space as the only one existing (Fig. 5.2B). The

analysis of either the extrinsic or intrinsic interpretations can inform on different aspects of a system.

As practical analogy, we can define a five-components (e.g., cell types) system of which only a small portion can be observed at any given time. Through time, the count size of two components increases (solid red and irregular orange lines; Fig. 5.2A); the count size of one component decreases (dashed blue line in Fig. 5.2A); while two components remain stable in size (dash-dotted purple and dotted green lines in Fig. 5.2A). The relative proportions of the five groups are observed through time (Fig. 5.2B). The visualisation of the real space (unobserved; extrinsic interpretation of the simplex) and the proportional space (observed; intrinsic interpretation of the simplex) give different perspectives on which component is changing. In the real space two different groups appear stable, maintaining a unchanged size (dash-dotted purple and dotted green lines in Fig. 5.2A); while in the proportional space, only one different component maintains a stable relative proportion (irregular line; Fig. 5.2B). In this example, the extrinsic interpretation can provide information on the growing forces in the system and can be helpful for example if the goal is to target the component(s) that are "driving" the change, in order to re-establish the equilibrium. The intrinsic interpretation can provide information on the mere proportional relationships among the components in a system and can be helpful for example if the ratio of two or more components is important for the stability of the system. In an alternative scenario, where the system can be completely observed but is characterised by a lower and upper boundary (e.g., voting system), the extrinsic interpretation can still give information on the growing forces in a system. That is, even if a quantity reaches a plateau toward its upper or lower bound, the force applied to that boundary can live in an unbounded space (-Inf, +Inf).

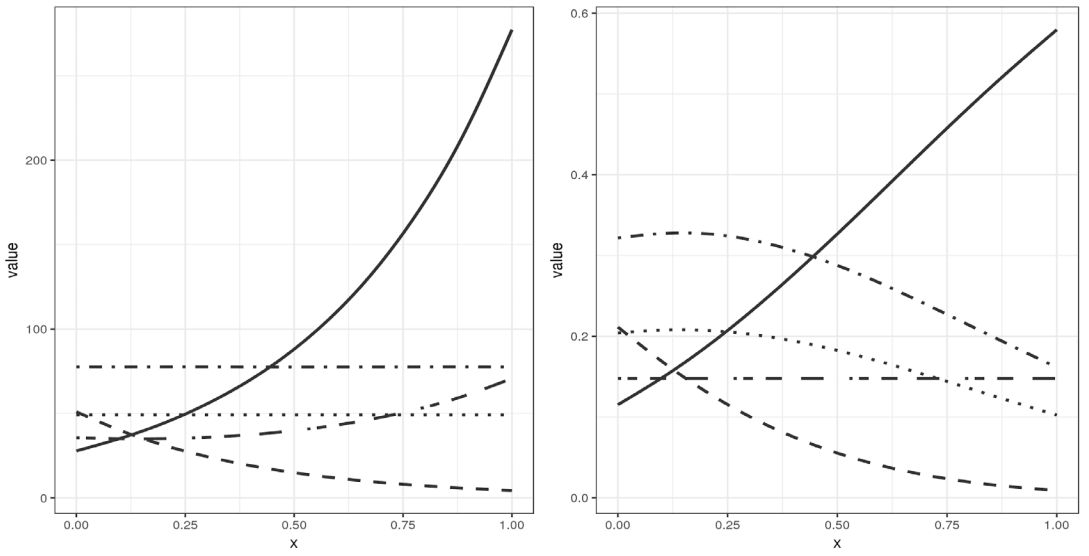


Figure 5.2: A - Extrinsic interpretation of the real unconstrained space (unobserved). B - Intrinsic representation of A in terms of probability in simplex space (observed).

This problem has been referred to as simplex, Dirichlet or softmax regression. Generally, two main approaches are adopted to model changes in this "awkward" [309] sample space. First, the simplex data can be transformed to a linear space using methods such as isometric, centred, or additive log-ratio, enabling the direct use of familiar statistical tools such as multiple linear regression. This approach although attractive presents some drawbacks, including the more laborious interpretation of the numerical generating process and the loss in accuracy in cases of extreme non-symmetricity and/or heteroscedasticity. Second, the simplex can be modelled in its native sample space. The probability density distributions of choice are either beta[310,311], Dirichlet[192], or simplex[312]. For regression, the use of such distributions is coupled usually with a sigmoid function, or with other monotone functions defined within the interval zero to one. A statistics that works on untransformed data allows a more direct construction of generative Bayesian models, which have been shown to work well in regimes of low or missing data [313]. Moreover, such models can be integrated naturally in larger hierarchical models. Considering the characteristics of the statistical model described in this article, we will focus on the methods that work in a native simplex space, from now on.

Beta distribution and Beta regression

The Beta distribution is described by Eq. 2

$$(2)\quad 1_{\frac{}{\mathrm{B}(\alpha,\beta)x^{\alpha-1}(1-x)^{\beta-1}}}$$

where the shape and the location are defined by combinations of the two parameters α and β (> 0). The beta regression underlies a generalized linear model where the link function is an inverse logit and the noise around such function follows a beta probability density function. Ferrari and Cribari-Neto proposed a beta-regression model and implemented the statistical estimation in the R language as the package betareg[310,311]. The maximum likelihood (ML)[314] is used for the estimation of the parameters. Also, an R package gamlss [315–317] provides semi-parametric beta regression type models for proportional data. The advantage of the beta regression is that the beta distribution follows the non-symmetricity and heteroscedasticity characteristics of the simplex sample space, allowing flexibility to model both centre and variance.

Dirichlet distribution and Dirichlet regression

The Dirichlet probability distribution is defined by Eq. 3

$$(3)\quad f\left(x_1,\ldots,x_K;\alpha_1,\ldots,\alpha_K\right)=\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K}x_i^{\alpha_i 1}$$

where K is the order of the distribution (i.e., the number of components of which proportions sum to 1), α (>0) is the parameter vector of length K[318], and , and B(α) being the multivariate Beta function. The modality (unimodal or bimodal) of the K components is defined by the sign of the logarithm of α: if positive the distribution is unimodal or bimodal otherwise (with the modes being at the values zero and one). The overall precision around the modes is dependent on the overall magnitude of the logarithm of α, where values distant from zero concentrate the mass around the mode. For example, for log(α)>0 (i.e., aα > 1) the precision equals to ∑log(α). The location of the modes of the K components is defined within the interval 0-1 and depends on the relative ratios of the components of the α parameter. For example, the modes of the two Dirichlet distributions defined by A = [1,2,3], and B = [10, 20, 30] are the same (approximately equal to 0.17, 0.3, and 0.5) with B characterized by lower variances around the modes.

A publicly available regression algorithm based on the Dirichlet probability distribution is DirichletReg[192], which implements such distribution under two possible parametrizations: common and alternative. Under the common parametrization, the α parameter is modelled by log-link; while under the alternative parametrization, the α parameter is modelled analogously to betareg[310,311] with mean and variance, allowing for heteroskedasticity. The maximum likelihood is used for the estimation of the parameters[314]. The Dirichlet distribution of K components has a relatively small degrees of freedom compared to K Beta independent distributions, with the variance defined by just one; this aspect can suite low-data regimes. A disadvantage of Dirichlet regression is due to the strong independence foundation of its distribution, which can produce only convex sampling (Fig. 1); while for some data the negative correlation among component car result in concave distributions[309].

Simplex distribution

The Simplex distribution is derived from the generalised inverse Gaussian distribution, and is defined by Eq. 4

$$(4)\ p\left(y;\alpha_1,\alpha_2,\mu,\sigma^2\right) = c(\alpha_1,\alpha_2,\mu,\sigma^2)y_1^{a_1-1}\left(1-y_1\right)^{\alpha_2-1} e^{-\frac{1}{2\sigma^2}d_{\alpha_1,\alpha_2}(y;\mu)},$$
$$y \in (0,1)$$

with mu and σ being the centre and variance; and α1 and α2 being shape parameters [319]. The Simplex distribution limits to a Beta distribution if α1 and α2 >0 and σ2 → inf [320], and limits to a standard normal distribution if σ2 → 0[320]. This distribution can be unimodal or bimodal depending on the σ parameter (point of inversion 4/sqrt(3))[312].

Such probability distribution has been implemented in a regression model in R called simplexreg[312]. This algorithm is based on a generalized linear model with a simplex distributed noise, using a diverse range of regression functions including logit, probit, cloglog, neglog or any other monotonic, differentiable function that map the space (0,1) → (-inf, +inf). The iteratively reweighted least squares algorithm is used for the maximum likelihood estimation of the parameters. The advantages of the simplex distribution are its adaptability to overdispersed noise models; however it comes at the cost of being defined by four parameters for each component, which leaves less residual degrees of freedom for inference compared with other methods.

Independently from of which model they adopt, publicly available algorithms do not attempt to infer the extrinsic rates of change but limit their inference to the intrinsic ones. Here, we present a generative Bayesian model cable to infers both extrinsic and intrinsic changes from data simplex space, using a Dirichlet and a Beta generative models respectively.

Methods

The probabilistic model

A Bayesian probabilistic model was used to infer the trends of change of the components of a simplex. In order to infer the trends of change for both the intrinsic and extrinsic interpretations of the simplex space, a Dirichlet framework and a Beta framework were used respectively. The two models can be represented by the probability densities Eq. 5 and Eq. 15

Dirichlet framework

$$(5)\ \mathrm{P}(\sigma)\,P(\phi)\prod_{r=1}^{R}P(\alpha_r|\sigma)\prod_{n=1}^{N}P(Y_n|X_n,\alpha,\phi)$$

$$(6)\ P(Y|X,\alpha,\sigma)\sim Dirichlet(\widehat{\widehat{Y}})$$

$$(7)\ \widehat{\widehat{Y}}=softmax(\widehat{Y})*\sigma$$

$$(8)\ \widehat{Y}=X\cdot\alpha$$

$$(9)\ \sum_{p=1}^{P}\alpha_{p,1}=0$$

$$(10)\ P(\sigma)\sim cauchy\,(0,2.5)$$

$$(11)\ P(\phi)\sim cauchy\,(0,2.5)$$

$$(12)\ P(\alpha_{1..P,1})\sim normal\,(0,5)$$

$$(13)\ \mathrm{P}(\alpha_{1..P,2..R})\sim\ \mathrm{reg.horseshoe}\ \ (1,1,1,4,4)$$

$$(14)\ softmax(\mathbf{z})_j=\frac{e^{z_j}}{\sum_{k=1}^{K}e^{z_k}}forj=1,...,K$$

Beta framework

$$(15)\ \mathrm{P}(\phi_2)\prod_{r=1}^{R} P(\alpha_{2r})\prod_{n=1}^{N} P\left(Y_n|X_n,\alpha_2,\phi_2\right)$$

$$(16)\ P\left(Y|X,\alpha_2,\sigma_2\right)\sim Beta(\widehat{\widehat{Y_2}}*\phi_2,(1\widehat{Y_2})*\phi_2))$$

$$(17)\ \widehat{\widehat{Y_2}}=inv.logit(\widehat{Y_2})$$

$$(18)\ \widehat{Y_2}=X\cdot\alpha_2$$

$$(19)\ P\left((\alpha_2)_{1..P,1}\right)\sim beta\left(2,(11/P)*2P\right)$$

$$(20)\ P\left((\alpha_2)_{1..P,2..R}\right)\sim normal\left(0,1\right)$$

For the two sets of formulae, Y represents the input simplex of dimension K (number of components), N (number of observations) and X represents the design matrix of dimensions N, R (number of covariates including the intercept term).

For the Dirichlet framework, α represents the matrix of the factors of interest of size N, R; φ is the precision parameter for the Dirichlet prior distribution for Y; and σ represents the vector parameter for the regularised horseshoe[225] prior distribution for α. Specifically, Eq. 5 represents the joint probability distribution that is used to identify the most likely values of the parameters; Eq. 6 represents the sampling statement for the likelihood of the data from a Dirichlet distribution; Eq. 7 represents the calculation of the hyperparameter of the Dirichlet probability distribution, parameterized as expected values and precision[193]; Eq. 8 is calculation of the expected values for the linear system; Eq. 9 represents the sum-to-zero constraint of the intercept values, which allows the degrees of freedom on the P components to fit those of the expected value of the Dirichlet hyperparameter (i.e., P-1); Eq. 10, 11, 12 and 13 represent the prior distributions attributed to σ, φ and α respectively; and Eq. 15 represents the softmax transformation, which is the exponential equivalent of the zero-to-one normalization.

For the Beta framework, α_2 is analogous to α; and φ_2 represents the precision parameter for the Beta distribution for Y. Specifically, Eq. 16 represents the joint probability distribution that is used to identify the most likely values of the parameters; Eq. 17 represents the sampling statement for the likelihood of the data from a Beta distribution, parameterized as expected value and precision; Eq. 18 represents the inverse logit transformation of the expected value of the linear system; Eq. 19 is analogous to Eq. 8; and Eq. 20 and 21 represent the prior distributions attributed to φ_2 and α_2.

Such models can be also represented as two oriented, plated graphs (Dirichlet - Fig. 3, Beta - Fig. S1), where the bold nodes represent the parameters, the light nodes represent data, where the edges between node represent the conditionality between the probabilities distributions and/or data, and the plates represent the iterations across the dimension of each node.

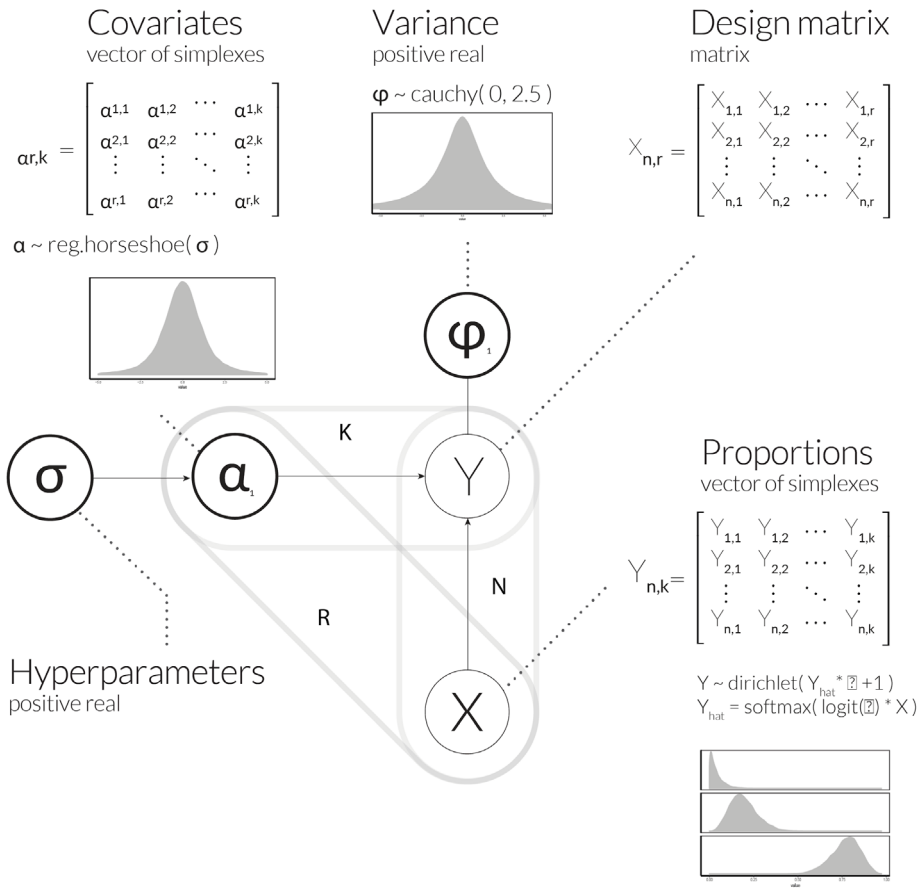


Figure 5.3: Graphical representation of the Dirichlet generative regression model, used to infer parameters for the extrinsic prospective.

Benchmark

A two-way benchmark for the intrinsic and extrinsic interpretation of the simplex space was designed with simulated data, with the aim of testing the accuracy in detection of non-zero changes. For both benchmarks, the numerical process that generated the data is based on the union of K negative binomials draws simulating the counts of K components of a system (hidden to the algorithms). At every point along the covariate of interest, the expected value of the K generating

negative binomials varies following an exp-linear regression function (i.e., log-link). For every point along the covariate of interest, the counts are then mapped to a simplex using a softmax transformation.

For each of the two benchmarks, the data set were simulated with four different flavours representing different selections of tuning parameters: (i) the variance around the regression function; (ii) the magnitude of the independent variables α (i.e., slopes); (iii) the number of observations; and (iv) the number of components of the simplex (i.e., K). For each simulation, 30 replicates were produced, for a total of 4800 simulations.

The extrinsic and intrinsic benchmarks were created with two different strategies to set the changing versus non changing components. For the extrinsic benchmark, half of the components do not change in size (e.g., dash-dotted and dotted lines in Fig. 5.2A) while half of the components increase or decrease in size (e.g., solid, irregular and dashed blue in Fig. 5.2A). For the intrinsic benchmark, half of the components change in the real space in a way to retain a non-changing proportion in the simplex space (e.g., irregular line; Fig. 5.2B). For both benchmarks, each component was labelled 1 if changing or 0 if not. The performance of the algorithms was measured in terms of area under curve (AUC) based on this accuracy of the binary classification.

Results and discussion

Benchmark

The intrinsic benchmark was designed to test the accuracy in detecting significant changes in the relative abundance of each component. For this test, a flat line in simplex space represents a zero change. The algorithms betaReg and simplexReg and our model performed similarly across the four data set flavours (Fig. 5.4; increasing precision, slope, number of samples and number of components) in terms of area under the curve (AUC). Our model performed slightly better compared with the other algorithms for the three flavours: increasing precision, slope and number of samples. Overall, the algorithm dirichletReg performed poorly for all data set flavours, compared with the other algorithms.

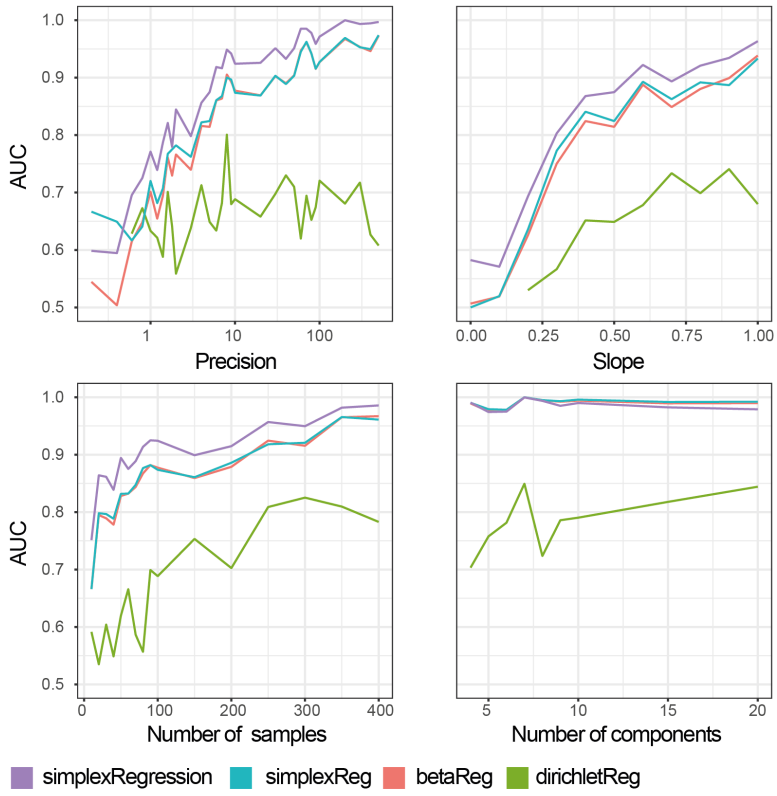


Figure 5.4: Area under curve (AUC) of four different methods when varying 4 parameters of simulation data.

The extrinsic benchmark was designed to test the accuracy in detecting changes in the real space under the parsimony assumption. In this scenario, null changes in the real space for two or more components may appear as distinct non-zero trends in the simplex space. For this extrinsic test, the true positive rate of betaReg, simplexReg and our model is high and comparable, with our statistical model performing slightly poorlier than the other two. Overall, the algorithm dirichletReg performed poorly compared with the other algorithms in terms of true positive rate. Fig. 5.5 shows the unique ability of our statistical model of classifying extrinsic changes from the simplex space in the four simulation flavours, with false positive rate < 1%. Counterintuitively, the false positive rate increases for betaReg and simplexReg as the input data becomes less noisy; that is, when the (i) variance around the regression function decreases; (ii) value of the independent variables α (i.e., slopes) increases; and (iii) number of observations increase. Such a result is a consequence of the independent evaluation of the K components of the simplex, rather than a

generative, integrative modelling. Interestingly, the number of components of the system did not affect the performances for simplexRegression, betaReg or simplexReg; and rather appears to be positively correlated with the performances of dirichletReg. The algorithm DirichletReg showed overall lower false positive rate for the extrinsic interpretation compared with betaReg and simplexReg; however, this is likely to be a side effect of the overall poor performance of such an algorithm in classifying true positives.

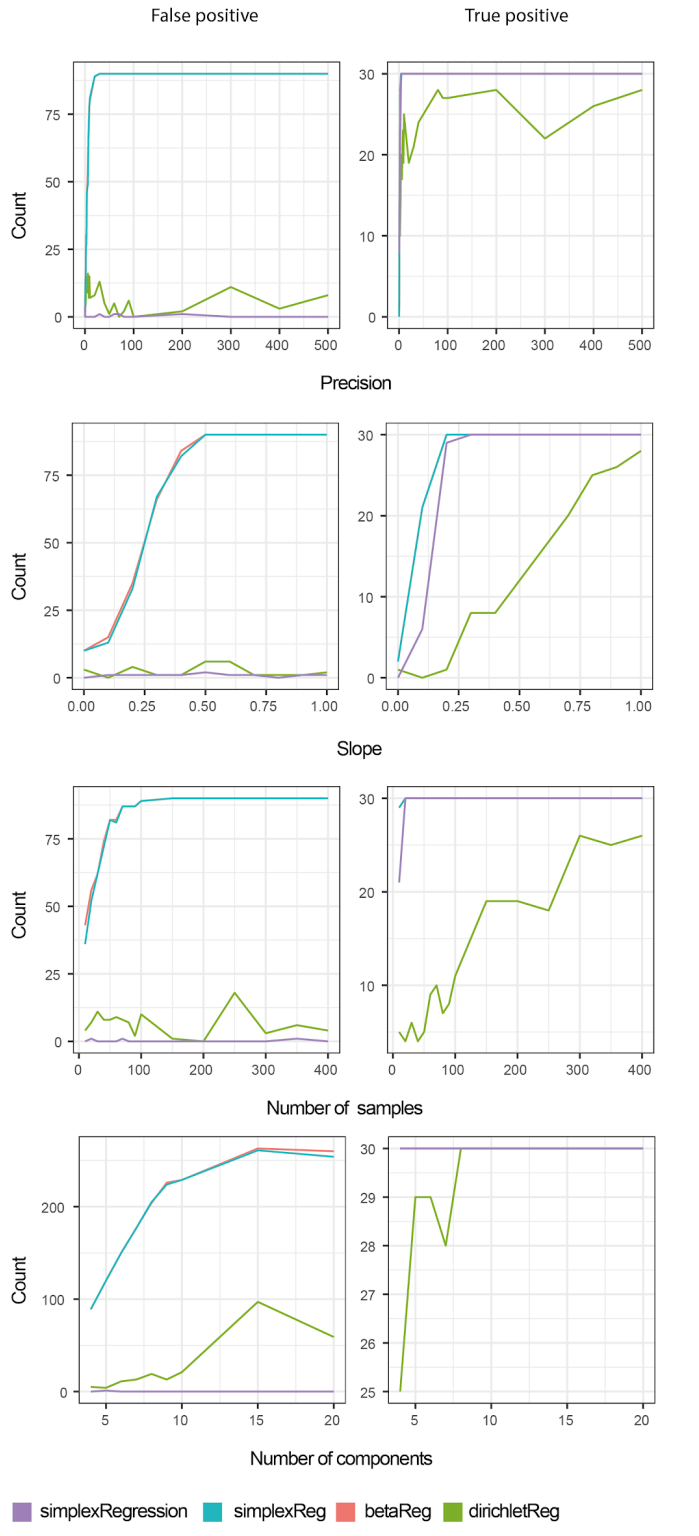


Figure 5.5: False positive and true positive counts of four different methods when varying 4 parameters of simulation data.

Probabilistic model implementation

Our Dirichlet-beta Bayes model has been encoded in the probabilistic language Stan[193,321], which permits the declaration of conditional probabilities and samples from the resulting joint distribution with the efficient non-U-turns-sampler (NUTS). The encoding of the probabilistic model in the Stan language is as follows.

```
functions{

 vector reg_horseshoe(
                                                    vector zb,
                                                    real aux1_global ,
                                                    real aux2_global,
                                                    vector aux1_local ,
                                                    vector aux2_local ,
                                                    real caux,
                                                    real scale_global ,
                                                    real slab_scale
                                                    ) {
  int K = rows(zb);

  // Horseshoe variables
                        real tau ; // global shrinkage parameter
                        vector [ K] lambda ; // local shrinkage parameter
                        vector [ K] lambda_tilde ; // ' truncated ' local shrinkage parameter
                        real c; // slab scale

                        // Horseshoe calculation
                        lambda = aux1_local .* sqrt ( aux2_local );
                        tau = aux1_global * sqrt ( aux2_global ) * scale_global * 1 ;
                        c = slab_scale * sqrt ( caux );
                        lambda_tilde = sqrt ( c ^2 * square ( lambda ) ./ (c ^2 + tau ^2* square ( lambda )) );
                        return zb .* lambda_tilde * tau ;
 }
}
data{
            int K;                  // Number of groups
            int N;                  // Number of observations
            int<lower=1> R;             // Number of covariates
            matrix[N,R] X;              // Design matrix
            simplex[K] beta[N];                 // Proportions

            // Horseshoe
            real < lower =0 > par_ratio ; // proportion of 0s
            real < lower =1 > nu_global ; // degrees of freedom for the half -t prior
            real < lower =1 > nu_local ; // degrees of freedom for the half - t priors
            real < lower =0 > slab_scale ; // slab scale for the regularized horseshoe
            real < lower =0 > slab_df; // slab degrees of freedom for the regularized

}
transformed data{
            real < lower =0 > scale_global = par_ratio / sqrt(1.0 * N); // scale for the half -t prior for tau
}
parameters {
            real<lower=0> phi_raw;                      // Unique variance of the dirichlet distribution
            matrix[R,K] extrinsic_raw;                  // Covariates
```

```
		// Horseshoe
		real < lower =0 > aux1_global ;
		real < lower =0 > aux2_global ;
		vector < lower =0 >[ K] aux1_local ;
		vector < lower =0 >[ K] aux2_local ;
		real < lower =0 > caux ;

		real<lower=0> phi2;
 matrix[R,K] intrinsic;
		real<lower=0> bg_sd2;		// Variance of the prior distribution to alpha

}
transformed parameters{
		matrix[N,K] beta_hat;
		vector[K] beta_hat_hat[N];
		matrix[N, K] beta_hat2;
		matrix[R,K] extrinsic;		// Covariates
		real<lower=0> phi = inv(sqrt(phi_raw));		// Unique variance of the dirichlet distribution

		// Building matrix factors of interest
		extrinsic[1] = extrinsic_raw[1];
		extrinsic[2] = to_row_vector(
			reg_horseshoe(
				to_vector(extrinsic_raw[2]),
				aux1_global ,
				aux2_global,
				aux1_local ,
				aux2_local ,
				caux,
				scale_global,
				slab_scale
			)
		);

		beta_hat = X * extrinsic;
		for(n in 1:N) beta_hat_hat[n] = softmax( to_vector(beta_hat[n]) ) * phi ;

		beta_hat2 = inv_logit(X * intrinsic);
}
model {

		// Priors
		phi_raw ~ normal(0,1);
		phi2 ~ cauchy(0,2.5);
		bg_sd2 ~ cauchy(0,2.5);

		// Linear model
		for(n in 1:N) beta[n] ~ dirichlet(beta_hat_hat[n]);

		extrinsic_raw[1] ~ normal(0,5);
		sum(extrinsic_raw[1]) ~ normal(0,0.01 * K) ;
		extrinsic_raw[2] ~ normal (0 , 1);
		if(R > 2) for(r in 3:R) extrinsic_raw[r] ~ normal(0,1);

		// Horseshoe
		aux1_local ~ normal (0 , 1);
		aux2_local ~ inv_gamma (0.5* nu_local , 0.5* nu_local );
		aux1_global ~ normal (0 , 1);
		aux2_global ~ inv_gamma (0.5* nu_global , 0.5* nu_global );
		caux ~ inv_gamma (0.5* slab_df , 0.5* slab_df );
```

```
        // Linear model for beta regression
        for(n in 1:N) beta[n] ~ beta(beta_hat2[n] * phi2, (1 - beta_hat2[n]) * phi2);

        // Prior distribution on the background cluster
        intrinsic[1] ~ normal(1.0/K, bg_sd2);
        intrinsic[2] ~ normal(0, 2);
        if(R > 2) for(r in 3:R) intrinsic[r] ~ normal(0,1);

}
generated quantities{
        vector[K] beta_gen[N];
        for(n in 1:N) beta_gen[n] = dirichlet_rng(beta_hat_hat[n]);
}
```

Interface

Our model has been implemented as R package (simplexRegression). Our function accepts three inputs; (i) a proportion matrix oriented with column-wise components and row-wise observations; (ii) a design matrix (including the intercept); and (iii) the label of the covariate to perform the hypothesis test on. With the command "simplexRegression(proportion_matrix, design_matrix, cov_to_test)", two different hypothesis testing are shown, one for the extrinsic scenario (cov - extrinsic; where independent variables α of interest are inferred for the generating numerical process), and for the intrinsic scenario (cov - intrinsic; where independent variables α of interest are inferred for the mere simplex space).

```
-------------------------------------------------------------------------------------------
Beta-Coefficients for variable no. 1:
                Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)       0.27       0          -8.71        1.03e-05 ***
cov - extrinsic    0.94       0          -22.43       0 ***
cov - intrinsic    1.14       0          -20.4        0 ***
-------------------------------------------------------------------------------------------
Beta-Coefficients for variable no. 2:
                Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)       0.26       0          -4.74        0.31
cov - extrinsic    -0.82       0          17.04        1.33e-10 ***
cov - intrinsic     -1.23       0          16.9         0 ***
-------------------------------------------------------------------------------------------
Beta-Coefficients for variable no. 3:
                Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)       0.27       0          -8.09        4.44e-05 ***
cov - extrinsic    -0.09       0          1.87         0.39
```

```
        cov - intrinsic     -0.44      0         7.18        7.03e-13 ***
        ---------------------------------------------------------------------------------------------
        Beta-Coefficients for variable no. 4:
                    Estimate   Std. Error   z value     Pr(>|z|)
        (Intercept)      0.2       0          5.37       0 ***
        cov - extrinsic    0.08      0         -1.59      0.39
        cov - intrinsic     -0.2      0         2.8         5.18e-03 **
---------------------------------------------------------------------------------------------
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Number of Observations: 100
Link: linear-softmax
```

Plots

A plot is produced for each inference run, which summarizes the inference for both the intrinsic and extrinsic interpretations. The densities represent the posterior estimate distributions of the slope parameters. The densities that have their mass distant from zero represent components with positive/negative change. Fig. 6 shows the posterior distributions for a simulated system with three stationary components (in linear space) and one increasing component along the covariate of interest. This example shows how significative differences between intrinsic and extrinsic interpretations of the simplex space can exist. In the intrinsic interpretation, all four components are inferred to have non-zero changes as they appear in the simplex space. On the contrary, in the extrinsic interpretation only one component is inferred having non-zero changes (i.e., red density curve; as it was simulated in the real space). For this interpretation, the increase in count size of such component in real space drives all the changes apparent in the simplex space.

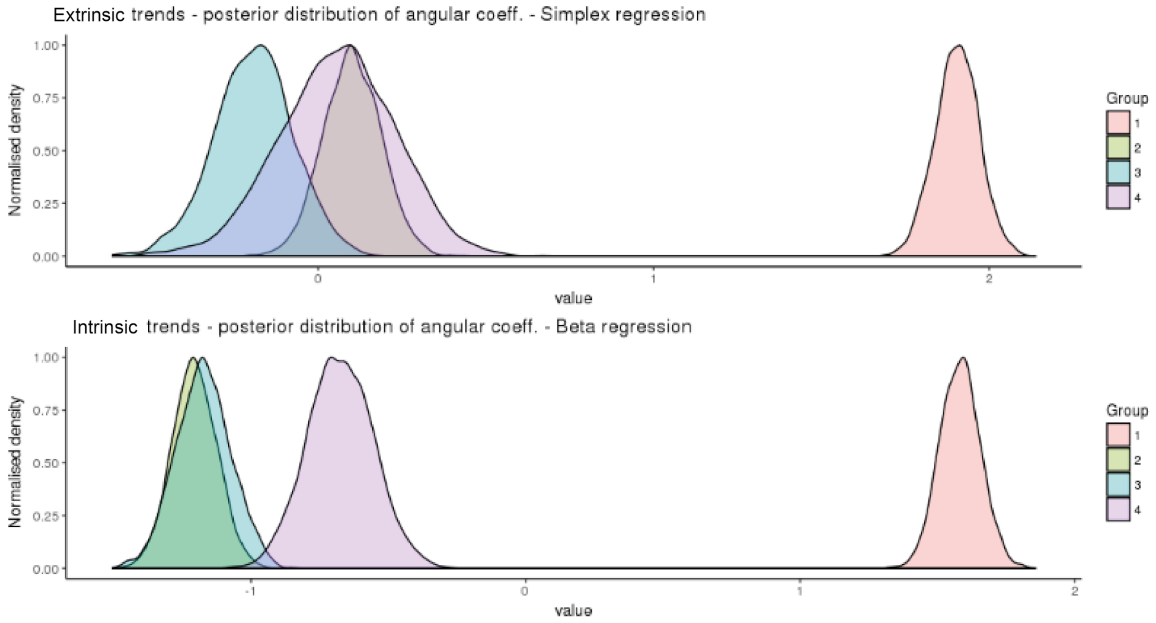


Figure 5.6: Example result plot, for a system with four components of which only one has a non-zero, positive rate of change in the real space. On the top is shown the inference on the extrinsic interpretation of the simplex space. On the bottom is shown the intrinsic interpretation. The density curves represent the posterior distribution of each of the four slope parameters.

Conclusions

Every experiment that collects data as proportions, and many experiments that collect count data of a small portion of a system, can be modelled in simplex space. Being able to correctly model changes happening in simplex space is therefore important. Publicly available algorithms focus on the intrinsic interpretations of the simplex space, analysing trends as they visually appear within the zero to one range, with the interpretation of a flat line as zero change. However, investigating which of the K components of a system is likely to drive the changes that are observed in the simplex space is often relevant. Such question has validity only if a parsimony assumption is made, being that more than one component is not changing in the physical system in analysis. For example, if we consider a tissue composed by K main cell types and want to observe the evolution of tissue composition during cancer progression, the assumption to be made is that two or more cell types do not change in number within the tissue, and thus that the tumour development is associated with a limited number of drivers. The ability of our model to identify such drivers is due to the integrated and generative modelling of all components (for the

Dirichlet part of the model). This approach is possible through the modelling of the intercept term of the K-components linear system with a variable having K-1 degrees of freedom, which introduces an interdependence among the components, accounting for the inverse relations among them characteristic of the simplex space. Our statistical model can accurately inform about both extrinsic and intrinsic interpretations of the simplex space, using a Dirichlet and beta models respectively. In order to build our generative model, the statistical framework Stan was employed. Although our model was implemented in a R package, it constitutes a module that can also be integrated in more extensive models, where the regression in simplex space may be just a component of a complex workspace. For example, where the simplex input is itself a parameter inferred upstream, and/or the coefficients of change (i.e., slopes) are used downstream. This modularity increases the values of the design and testing of generative probability Bayes models.

CHAPTER 6

Context

The properties of the tissue microenvironment are key for many pathological conditions such as cancer. For example, for several types of cancer, the tissue microenvironment shows altered properties that actively promote tumour development and/or decrease the treatment response. Furthermore, some molecular markers of the tissue microenvironment have been selected as a valuable diagnostic tool. Both *in vitro* and *in vivo* techniques have been used for the study of the biology of tissue microenvironment at the molecular level and its association to cancer progression. Such techniques must necessarily focus on specific genes/cell type targets due to the financial and logistic burden. Computational methods allow large scale exploratory analyses across many genes, cell types and samples. Recently, improved algorithms and a wider availability of high throughput molecular data have allowed large scale epidemiological studies inferring the cellular composition of tumour microenvironments using both novel and publicly available data sets. Inferring changes in tissue composition, that we name here differential tissue composition (DTC) analysis is essential to readily connect such inference with biology. Differential tissue composition analyses implies the inference of both tissue composition and its rate of change along a factor of interest. So far, only the first stage of the problem, but not the second, has been tackled. Here, we present a hierarchical Bayesian model that, besides improving on the "state of the art" in the inference of tissue composition, permits the inference and testing of the significance of changes in tissue composition across biological conditions.

Differential tissue composition analyses from whole tissue transcriptional levels

Introduction

The properties of the tissue microenvironment are key for many pathological conditions such as cancer [19,40,46,160], involving both immune (e.g., t-cells, b-cells, monocytes, granulocytes) and stromal (e.g., fibroblasts, endothelial, smooth muscle, nervous) cell compartments. For example, increased proportion of infiltrating immune cells within the tumour microenvironment (TME) affects tumour growth in colorectal [322,323] and breast cancer [324,325]; similarly, stromal cells are key for cancer development, such as cancer associated fibroblasts in colorectal [326,327], in prostate [1] and breast cancer [328], endothelial in breast cancer [329,330] and adipose in prostate cancer [160]. Improving our knowledge on the biology of the tumour microenvironment will lead to improved diagnosis and a better understanding of drug resistance [331–333]. Given their key role and genetic stability, tumour associated stromal and immune cells represent a valuable target for treating cancer [324,331,334,335].

For cancer, the properties of specific cell types within the tumoral mass have been investigated through *in vitro* and *in vivo* experiment, such as migration and proliferation assays, and drug response or xenograft mouse models [67]. Despite their effectiveness, such approaches must focus on specific cell types and on a limited number of tissue samples, due to the financial and logistic burden associated to the experimental procedures. The computational inference of tissue composition using tissue transcriptional levels is a high-throughput alternative that enables large scale epidemiological investigations, using novel and publicly available data. However, such inference must link with biological/clinical traits to gain meaning. The inference of changes in tissue composition along biological/clinical traits, that we name here "differential tissue composition" (DTC) analysis is key to readily provide clinical applications. The analysis of differential tissue composition can be presented as two-staged: (i) the inference of tissue composition for each sample; and (ii) the execution of a trend analysis through the integration of sample-wise tissue composition across biological conditions.

More precisely, the inference of tissue composition (i) refers to the estimation of the proportions of each cell type within. For the special case of one gene and one sample (Eq. 1), the tissue gene transcription level *y* (observed) is equal to the weighted sum of specific gene transcriptional level for each cell type *a* (observed or unobserved), weighted by its absolute proportion π within such tissue sample (unobserved). For the generalised case of many genes and many samples (Eq. 2), the observed matrix of gene transcription levels Y (observed) is equal to the matrix of specific gene transcriptional levels for each cell type A (observed or unobserved) multiplied by the matrix of absolute proportions of each cell types within each tissue sample M (unobserved).

$$(1)\quad \mathrm{y} = \mathrm{a}_1 * \pi_1 + a_2 * \pi_2 + \ \ldots\ + a_a * \pi_P$$

$$(2)\quad Y \begin{bmatrix} y_{g1,s1} & y_{g1,s2} & \cdots & y_{g1,sS} \\ y_{g2,s1} & y_{g2,s2} & \cdots & y_{g2,sS} \\ \cdots & \cdots & \cdots & \cdots \\ y_{gG,s1} & y_{gG,s2} & \cdots & y_{gG,sS} \end{bmatrix} = A \begin{bmatrix} a_{g1,p1} & a_{g1,p2} & \cdots & a_{g1,pP} \\ a_{g2,p1} & a_{g2,p2} & \cdots & a_{g2,pP} \\ \cdots & \cdots & \cdots & \cdots \\ a_{gG,p1} & a_{gG,p2} & \cdots & a_{gG,pP} \end{bmatrix} \times \Pi \begin{bmatrix} m_{p1,s1} & m_{p1,s2} & \cdots & m_{p1,sS} \\ m_{p2,s1} & m_{p2,s2} & \cdots & m_{p2,sS} \\ \cdots & \cdots & \cdots & \cdots \\ m_{pP,s1} & m_{pP,s2} & \cdots & m_{pP,sS} \end{bmatrix}$$

where g {1..G} represents genes, s {1..S} represents samples and p {1..P} represents cell type populations. In order to solve this system of linear equations and infer the value of the matrix Π, several approaches have been employed so far. Four main approaches are: (a) regression; (b) generative probabilistic mixture model; (c) minimum ratio paradigm; and (d) gene set enrichment. The approach (a) via regression commonly treats each sample separately, reducing the problem to a series of multiple regressions. This approach was first applied by Venet et al.[68]; since then, has been tackled with a diverse range of optimization methods such linear regression[336], quadratic programming[78] and support vector linear regression[337]. The approach (b) via generative probabilistic mixture is based on the latent Dirichlet allocation (LDA) framework[94], and treats gene transcriptional signatures for each cell types as words in topics, and cell types in samples as topics in documents. The optimal solution for such hierarchical probabilistic model is identified via Markov chain Monte Carlo. The approach (c) via minimum ratio paradigma[75] treats each tissue sample separately, modelling them as a mixture of two components A and B with proportions p and 1-p. A unimodal decreasing curve is drawn ordering the gene-wise ratios $p_{g;\ \{1..G\}}$ obtained dividing the tissue gene transcriptional levels with the component A. The point where the second derivate of such curve equals zero indicates the best estimate for p. The approach (d) via gene enrichment analysis[72] treats each sample separately; it infers their composition using a gene

enrichment scores calculated using a publicly available non parametric method[95], that is transformed to a linear scale before further compensation.

More precisely, the analysis of differential tissue composition (ii) refers to the regression of the proportions of cell types (i.e., matrix Π) along a factor of interest (e.g., cancer grade). Such analysis provides confidence on the changes in cell type proportions associated with a factor of interest. For such analysis, two main approaches can be adopted: (a) multiple linear regression on a linearized homoscedastic data space; and (b) multiple generalised linear regression on the native data space. For the approach (a), the proportional data can be transformed to a linear space using methods such as isometric, centred, or additive log-ratio[309]. The resulting transformed data space is assumed to have normally distributed noise and trends of change that follow a linear function. This approach, although attractive, presents some drawbacks, including the more laborious interpretation of the results in face of the numerical generating process, and the loss in accuracy in cases if non-symmetric of heteroscedastic noise persists. For the approach (b), the sample space is assumed to have a either Beta[310,311], Dirichlet[192], or simplex[312] distributed noise and having trends of change that follow a either sigmoid function, or other monotone functions defined within the interval 0, 1[312]. In principle, a statistics that works on untransformed data allows a more direct construction of generative probabilistic models, which have been shown to work well in regimes of little or sparse data[313], and it can be integrated in larger hierarchical models.

Several algorithms have approached the deconvolution stage (i) of the problem, providing wide evidence for the accuracy of such computational approach for microarray data[69–71,73,75–78,80,81,83,338], and one piece of evidence for RNA sequencing data[72]. However, so far no publicly available algorithms attempted to integrate the regression stage (ii). Without such integration, the link to interpret the results in a biologically meaningful context, with statistical robustness, is missing. A serial inference of stages (i) and (ii) using standalone algorithms (rather that integrated in the same probabilistic model) would also not be an optimal approach, as the information about the uncertainty of the inference around (i) would be forgotten for inference of (ii), resulting in a significance test skewed toward false positives[339,340]. Here, we present ARMET-tc: a hierarchical generative Bayesian model that, beside improving on the "state of the art" in the inference of tissue composition (i), allows for the first time to infer and test the significance of changes in tissue composition across biological conditions (ii).

Hierarchical structure of the data

Analogously to cell lineage differentiation[341], the data and the inference problem were structured in a hierarchical fashion. Such hierarchy that can be represented as a directed tree graph, where nodes represent cell type categories and branches represent the parental relation between cell types categories (Fig. 6.1). The term "cell type category" defines a cell phenotype that is characterised by common traits (e.g., immune cell, t cell or epithelial cell), and it is represented by a node in the hierarchical structure. The hierarchical structure was organized in 3 levels, where level 1 represents main cell types (e.g., epithelial and immune cells) and level 3 defines specific cell phenotypes (e.g., activated dendritic cells, M2 polarised macrophages and regulatory t cells). Each cell type category is characterised by a list of marker genes, defined as genes that are preferentially transcribed in such category compared with others. Given the recursive nature of data manipulation happening within a tree, in order to facilitate the understanding of this section we introduce here few key recurring concepts. An ancestor of a node is the node of a lower level connected with it. The direct ancestor of a node is the ancestor directly connected with it (yellow dot in Fig. 6.1). The descendants of a node are all the nodes of a higher level that relate to it (green dots in Fig. 6.1). The direct descendants of node are those nodes of higher levels that are directly connected with it (red dots in Fig. 6.1). The direct peers of a node are all the other nodes with the same direct ancestor (purple dots in Fig. 6.1). The recursive peers of a node are all direct peers and the peers of indirect ancestors (blue dots in Fig. 6.1). If not specified, the terms ancestors, descendant or peer refer to their generic definition.

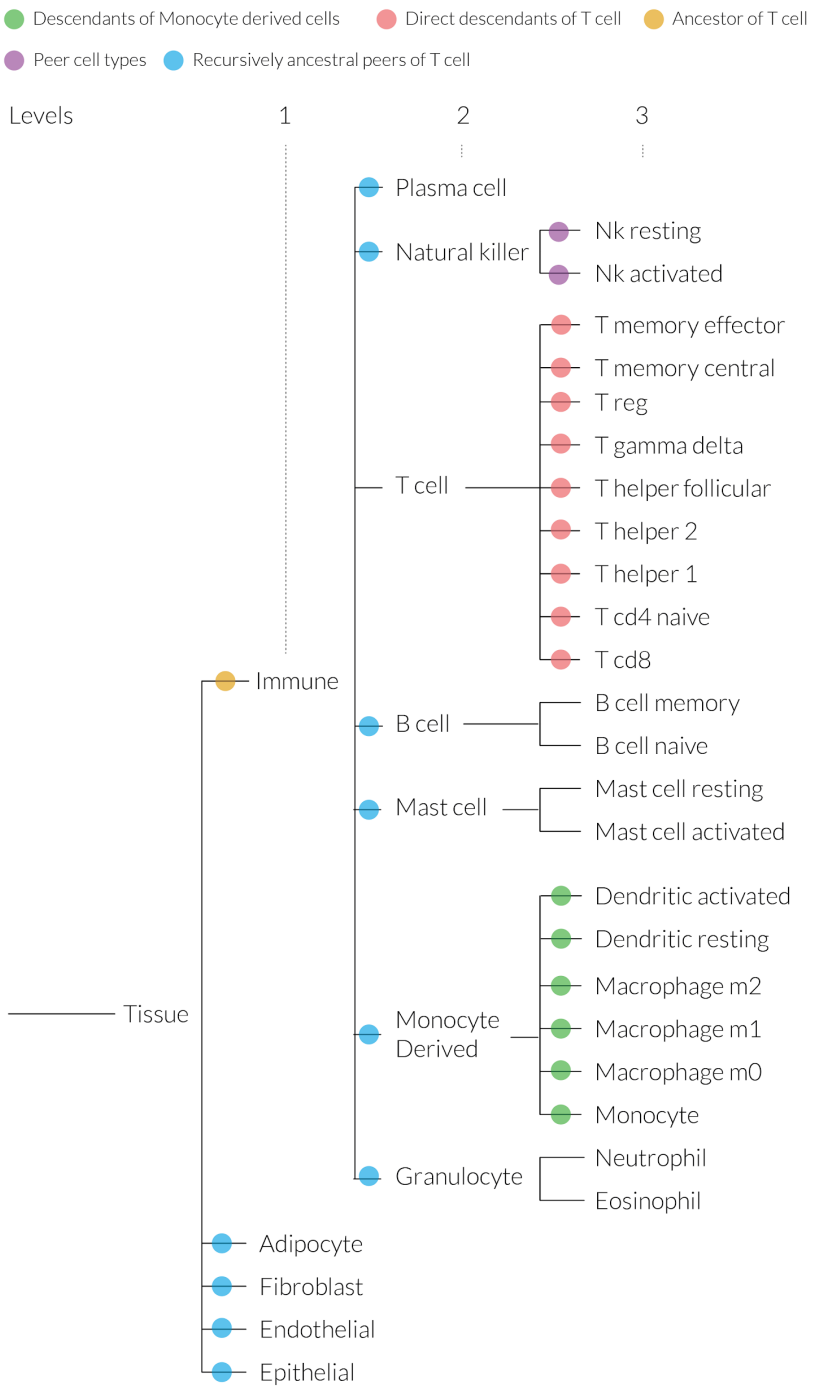


Figure 6.1: Graph representation of the cell type hierarchical structure used by ARMET-tc. Colours represent relations among elements of the hierarchy that are mentioned through the article. Elements of the hierarchy are referred as cell type categories or nodes within this article. The green nodes are descendants of the monocyte derived cell type category. The red nodes are direct descendants of the T cell type category. The yellow node is ancestor of the T cell type category.

The purple nodes are peers to each other. The blue nodes are recursive peers of the T cell type category.

Transcriptional signatures of cell type categories

The gene transcriptional levels for each cell type category were inferred from publicly available data sets. In order to accommodate both RNA sequencing and microarray input queries, two separate signature data sets were developed, named A-RNAseq and A-microarray respectively. For A-RNAseq, a total of 621 pure cell type samples were selected from the databases ENCODE[102], FANTOM5[101] and BLUEPRINT[100]; for A-microarray, a total of 603 pure cell type samples were selected from the GSE86362 and LM22[69] data sources, and integrated with A-RNAseq for the missing cell type categories. The algorithm trimmed mean of M-values (TMM)[106] was used to normalise A-RNAseq, while quantile normalization [342,343] was used to normalise A-microarray. In order to remove unwanted variation between data sources, the algorithm RUV4[107] was employed using level-2 cell type categories (Fig. 6.1) as covariate of interest, and using a selection of 600 housekeeping genes[107] as negative controls. The number of unwanted covariates (i.e., parameter K in RUV4 algorithm) was chosen with a parsimony criterion as the double of the number of the integrated databases (e.g., k = 6 for RNA sequencing). In order to normalize genes that had sparse information across any cell type, the missing gene transcription levels were estimated recursively from the closest recursive peers for which such information was available.

Gene markers selection

The set of marker genes for each cell type category was identified performing a differential gene transcription analysis including its recursive peers, using edgeR[108] on the RNA sequencing data of the cell type specific gene transcription signatures.

Structural design of the differential tissue composition analysis

The differential tissue composition analysis is performed recursively for each cell type category (i.e., node of the hierarchy; Fig. 6.1) having descendants, starting from level 1 of the hierarchy. Such analysis for each cell type category aims to estimate the trends of change of proportions of the direct descendants of such category. For example, the analysis for t cell tracks the change in proportions of the t cell subtypes (red dots in Fig. 6.1). Each local analysis is performed in a context-aware fashion integrating the contribution of recursive peers inferred for lower layers of

the cell type hierarchy. In order to achieve this, the linear equations (Eq. 1) is not applied directly, but rather adjusted (Eq. 3). That is, for each gene, its transcription level y in the tissue is made from the contribution of the cell type categories in analysis y_{fg}, and the contribution of the background y_{bg} cell type categories (i.e., recursive peers; Eq. 3).

$$(3)\ y = y_{fg} + y_{bg}$$

$$(4)\ y_{fg} = \hat{y} * \pi_{category}$$

$$(5)\ \hat{y} = a_1 * \hat{\pi}_1 + a_2 * \hat{\pi}_2 + ... + a_A * \hat{\pi}_P$$

$$(6)\ y_{bg} = a_{bg1} * \pi_{bg1} + a_{bg2} * \pi_{bg2} + ... + a_{bgB} * \pi_{bgP}$$

The probabilistic model

The differential tissue composition analysis is performed using a hierarchical probabilistic model where a first stage (i) infers the tissue composition for each sample, and where the second stage (ii) infers trends in tissue compositions across biological conditions. The first stage implements a multiple linear model on the logarithmic scale. The bimodality of RNA sequencing data is modelled using a zero-inflated log-normal noise model. The local composition $p_1...p_P$ of cell types for each sample is represented by a simplex (i.e., a vector defined in (0,1) which components sum to 1). The second stage implements both simplex regression and a beta regression for inferring and testing extrinsic and intrinsic rates of change[309] respectively. For integrity of the generative probabilistic model, while the simplex regression was integrated to the hierarchical model as generative process, the beta regression was implemented as a separate model that accepts as input the posterior distribution of the proportion matrix, so to not lose the information about the uncertainty around the inferred proportions. In total, the probabilistic model has 5 parameters, and can be described by a joint probability density formula (Eq. 7) or a graphical model (Fig. 6.2).

$$(7)\ P(\sigma)\,P(\phi)\,P(\delta)\prod_{r=1}^{R}\prod_{p=1}^{P}P(\alpha_{r,p}|\delta)\prod_{p=1}^{P}\prod_{s=1}^{S}P(\pi_{p,s}|X_s,\alpha_p,\phi)\prod_{s=1}^{S}\prod_{g=1}^{G}P(Y_{s,g}|x_g,\pi_s,\sigma_s,\theta^*)$$

$$(8)\ P\left(Y|x,\pi,\sigma,\theta^{*}\right)\sim lognormal\left(x*\pi,\sigma\right)$$

$$(9)\ P\left(\pi|X,\alpha,\phi\right)\sim Dirichlet(\widehat{\widehat{Y}})$$

$$(10)\ P\left(\alpha|\delta\right)\sim Dirichlet\left(\left[\delta_{1},...,\delta_{k}\right]\right)$$

$$(11)\ P\left(\sigma\right)\sim normal\left(0,0.1\right)$$

$$(12)\ P\left(\phi\right)\sim normal\left(1,...\right)$$

$$(13)\ P\left(\delta\right)\sim cauchy\left(1,2\right)$$

$$(14)\ \widehat{\widehat{Y}}=softmax(\widehat{Y})*\sigma$$

$$(15)\ \widehat{Y}=X\cdot\alpha$$

$$(16)\ \mathrm{softmax(z\ )}_{j}=\frac{e^{z_{j}}}{\sum_{k=1}^{K}e^{z_{k}}}forj=1,...,K$$

$$(17)\ \ \mathrm{logit}(p)=\log\left(\frac{p}{1p}\right)=\log(p)\log(1p)=\ \log\left(\frac{1}{p}1\right)$$

The parameter α represents the rates of change of each cell type category along the biological conditions. The parameter π represents the matrix of proportions for each cell type category and sample. The parameters σ, φ and δ define the noise model. The probabilistic model for beta regression was previously described elsewhere[344]. The point estimate and credible intervals for both cell type proportions and trends of change are calculated from the posterior distribution.

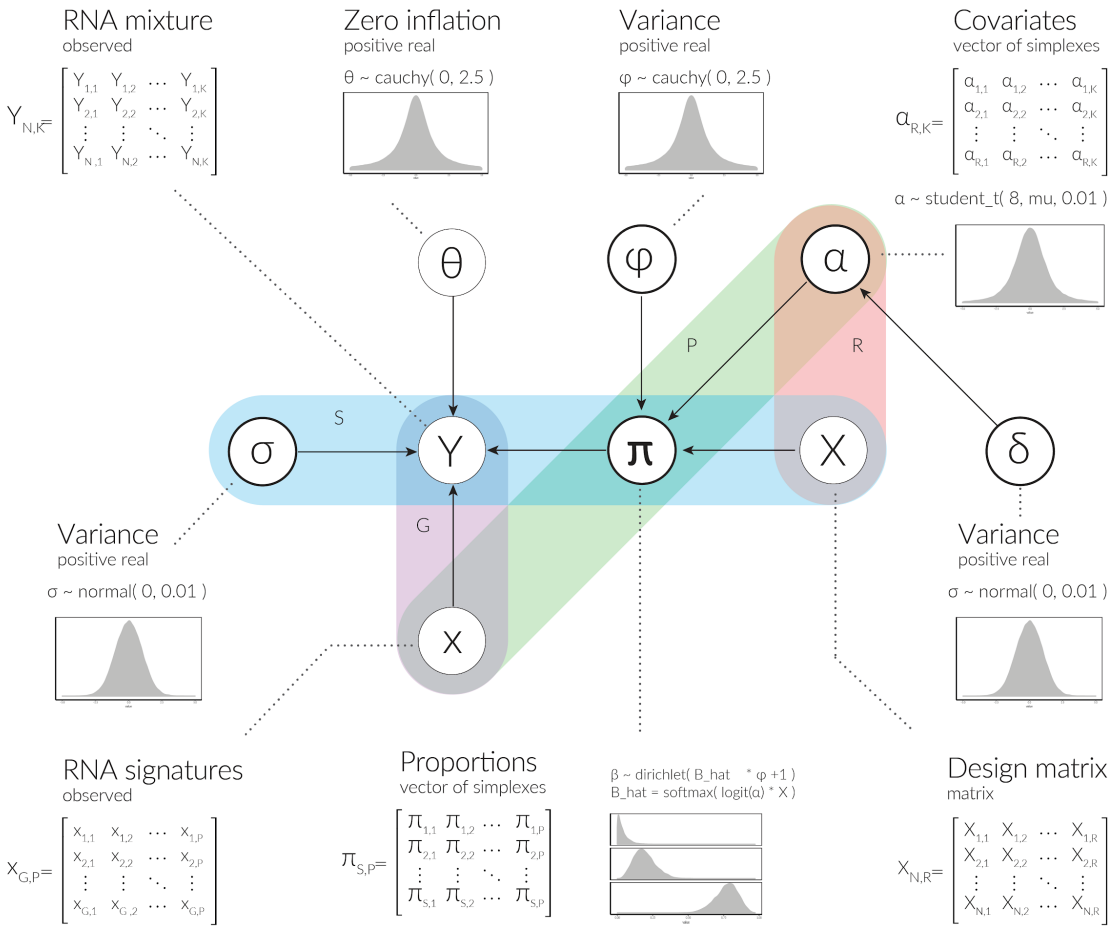


Figure 6.2: Graphical plate model representing the joint distribution as directed graph, where bold nodes represent parameters, these nodes represent the data, edges represent the conditionalities between nodes, and plates define the shared dimensionality of the nodes. For example, the parameter π shares the P (cell types) dimension with α, and the S (samples) dimension with the observed tissue gene expression

Implementation

This probabilistic model was implemented in R[345] using the Bayesian framework Stan [193,321]. Stan permits the definition of a probabilistic model with a declarative symbolic language and sample from it with the non-U-turns-sampler (NUTS)[346], based on the Hamiltonian monte carlo algorithm[347]. This probabilistic model is available as R package ARMET (with *install_github("stemangiola/ARMET", args = "--preclean", build_vignettes = FALSE*).

Regression benchmarks

A first simulation benchmark was designed to test the accuracy of the inference of changes in tissue composition across a continuous covariate, for cell types belonging to level 1, 2 and 3 (Fig. 6.1). The matrix of P cell type proportions across S tissue samples (i.e., matrix $\Pi_{S,P}$ in Eq. 2), was simulated as previously described[344], with positive trend of change for a cell type and null trend for the others. For each sample, the transcriptional signature for P cell types were sampled from A-RNAseq and mixed according to the simulated proportions. In order to evaluate the performances in a range of possible scenarios, in silico tissue samples were simulated with varying S and rates of change.

Comparative benchmark

A second benchmark based on validated data was designed to compare the performances of ARMET-tc to other publicly available methods in inferring tissue composition (i.e., stage (i) of the differential tissue composition analysis) from three validation data sets. First, a validation data set (named Pure) composed of a selection of purified cell types which gene transcription levels were detected with RNA sequencing. This dataset enabled the measurement of the specificity of the inference. Second, a validation data set (named TCGA) composed by cancer TCGA samples for which cancer purity has been estimated as consensus of an array of publicly available methods based on genomic and transcriptomic data[104]. This validation data set enabled the measurement of the accuracy of cancer purity estimation. Third, a validation data set (named PBMC) proposed by Newman [69] et al. based on gene transcription of samples of peripheral blood mononuclear cells, detected with microarray, for which cellular composition has been observed with flow cytometry. This validation data set enabled the measurement of the accuracy against a complex tissue with experimentally detected tissue composition. For each validation data set, an array of training cell type specific signature data sets was used to infer tissue composition, with the goal of testing the robustness of the methods to training data with variable size and origin. Such training transcriptional signatures data sets were: A-RNAseq; A-microarray; LM22; LM22-redo being LM22 reconstituted from the original raw data; LM22-Becht being the integration of LM22 with a second data set of microarray based immune cell signatures (i.e., GSE86362); and LM22+ being an integration of LM22-Becht with stromal transcriptomic signatures. Given the absence of the

whole transcriptome for LM22, and the different gene selection for ATMET-tc compared to Cibersort, LM22 could not be used in combination with ARMET-tc.

In order to achieve a more direct comparison across validation data sets, an adjusted inference error measure was adopted (Eq. 18)[344]. This error measures the difference between the inferred and the observed proportions of a specific cell type, adjusted by the absolute value plus one of the logit transformed value of the observed proportion[344]. Such adjustment permits the reduction in the bias of estimating rare or dominating cell types, compared to cell type with ~50% abundance[344]; normally, such bias leads to a relatively low or high error rate respectively.

$$(18)\quad \mathrm{Err}_{adjusted} = |P_{observed} - P_{inferred}| * \overline{|logit(P_{observed})| + 1}$$

Inference of associations between tissue composition and cancer relapse

In order to show the utility of ARMET-tc, we inferred the association between tissue composition and cancer relapse, for several primary tumour types within the TCGA database. The cancer type included were acute myeloid leukemia, breast invasive ductal carcinoma, breast invasive lobular carcinoma, cervical squamous cell carcinoma, colon adenocarcinoma, head and neck squamous cell carcinoma, hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, pancreatic adenocarcinoma, papillary thyroid cancer, prostate adenocarcinoma, renal clear cell carcinoma, serous ovarian cancer, stomach adenocarcinoma, uterine endometrioid carcinoma. RNA sequencing gene counts were taken from GDC portal[348]; the sample information were taken from Cbioportal[349]. A stratification strategy for disease free survival time was adopted, as a proportional hazard model (e.g., Cox-regression[350]) is currently not part of our statistical model. For each cancer type, samples were grouped according to whether the patient relapsed before time T1 or had not relapsed after time T2. Time T1 and T2 were selected for each cancer as the 0.2 and the 0.8 quantile of the disease-free survival time. The binary relapse variable was used as main covariate of interest and patient age was used as confounding factor.

Results and discussion

Regression benchmark

The first benchmark shows the performance of ARMET-tc in inferring trends of change (i.e., slope) across cell type levels (i.e., 1, 2, and 3; Fig. 6.1), with varying: (i) the effect size (i.e.,

proportional to the slope term in the linear model); and (ii) sample size (Fig. 6.3). As expected, overall the true positive rate increased with the sample size and effect size. If we consider 0.8 the minimum true positive rate acceptable, for cell types belonging to level 1 (Fig. 6.1) such positive rate can be achieved for a cell type doubling in size with ~40 samples; for level 2, the same rate can be achieved with ~100 samples. For level 3, 0.8 true positive rate is achievable with a threefold change in proportion for a cell type category across 100 samples.

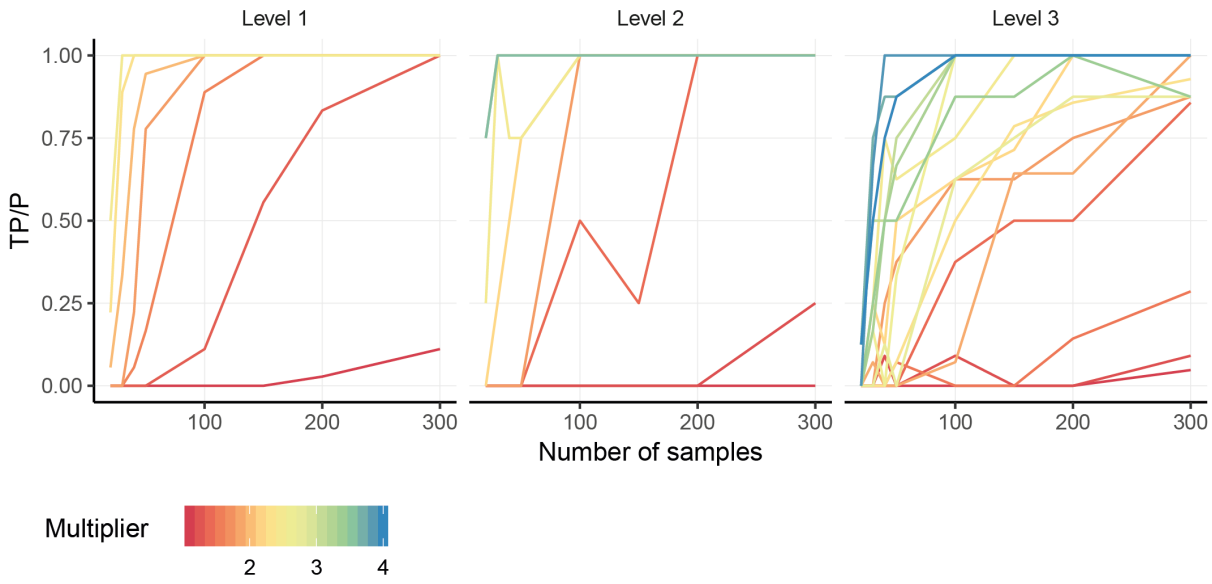


Figure 6.3: A - trend lines showing the true positive rate of ARMET-tc in classifying trends of change across varying degrees of change and sample size, for three different level of cell differentiation. The three dots represent the specific runs that are visualised in section B. B - Specific inference runs showing in black all the non-changing populations, in blue the ground truth of the changing population and in red the inference of cell type proportions in every sample with the inferred trend of change. The error bars represent the 0.95 credible interval.

Comparative benchmark

The second benchmark (Fig. 6.4) shows the overall higher accuracy of ARMET-tc in estimating tissue composition from tissue gene expression data, compared to other two reference publicly available methods: an implementation of linear equation system solver lsfit[336] and Cibersort[69]. In particular, for RNA sequencing based validation data, ARMET-tc was able to provide higher accuracy across all training data sets, while for the validation data set PBMC ARMET-tc shows a

higher robustness outperforming other methods across all training data sets except for LM22, where Cibersort was able to achieve better performances.

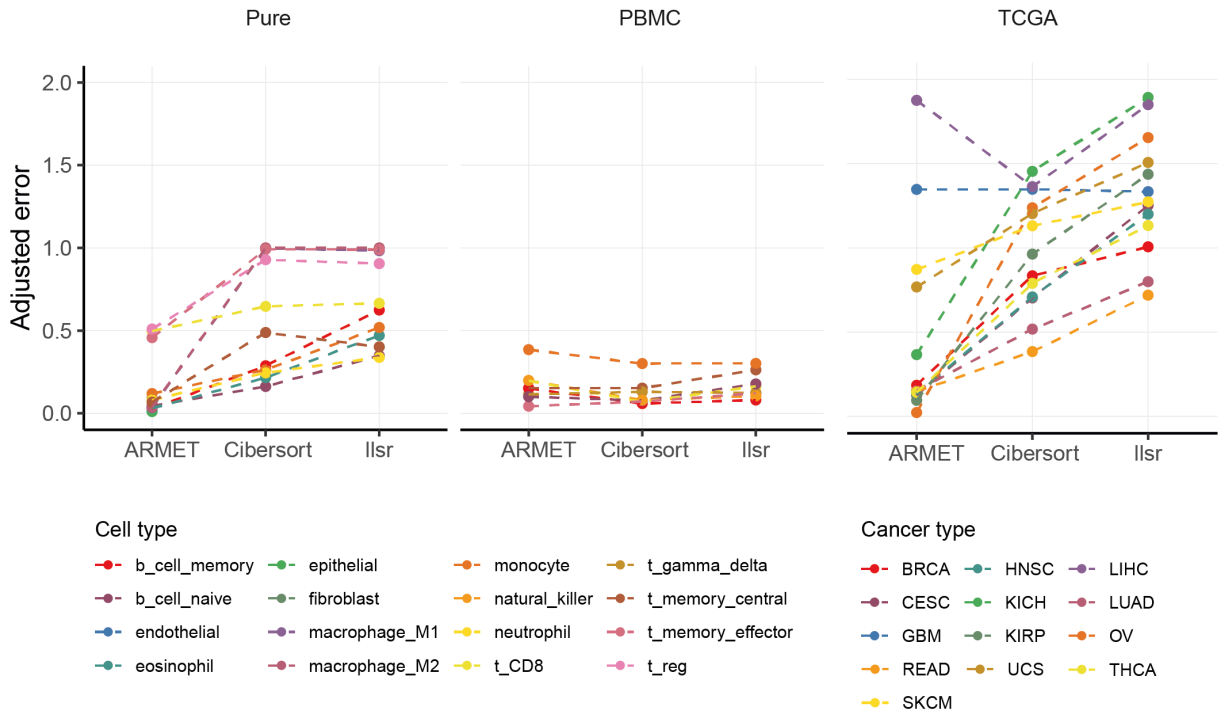


Figure 6.4: Line plot of the performances of three methods: ARMET-tc, Cibersort and Lsfit, on three validation data sets: Pure, composed by a selection of purified cell types which gene expression was detected with RNA sequencing; TCGA, composed by cancer TCGA samples for which cancer purity have been estimated as consensus of an array of publicly available methods based on genomic and transcriptomic data; and PBMC, gene expression samples derived from peripheral blood mononuclear cells, produced with microarray, for which cellular composition has been observed with flow cytometry. Each validation data set have been tested using an array of training signature data: A-RNAseq, A-microarray, LM22, LM22-redo being LM22 reconstituted from the original raw data; LM22-Becht being the integration of LM22 with another data set of microarray based immune cell signatures; LM22+ being an integration of LM22-Becht with stromal transcriptomic signatures.

Landscape of associations between cell types and cancer relapse

Differential tissue composition analyses were performed on an extensive array of cancer types, to infer the association between abundance of cell types with disease relapse. In order to visualise both overall abundance of cell types and significant changes of cell type abundance novel graphics were produced (Fig. 5.5A).

For prostate cancer, ARMET-tc was able to identify monocyte derived cells, and in particular macrophages as positively associated with cancer relapse. The infiltration of such cell types has been previously linked with proliferative inflammatory atrophy lesions, chronic prostatic inflammation and cancer[351]. Prostate cancer-specific and overall survival studies identified an elevated monocyte count as an independent prognostic factor for poor outcome[208,209]. Furthermore, TAM infiltration in prostate needle biopsy specimens is a useful predictive factor for PSA failure or progression of PCa after hormonal therapy[209].

For acute myeloid leukemia, several associations between blood composition and tumour relapse were identified. The abundance of eosinophil was inferred to be negatively associated with tumour relapse, in agreements with previous findings[352], focused on the relationship between eosinophilia and inversion 16, with remission and resistance to chemotherapy[353]. Similarly, the presence of γδ t cells were inferred to be negatively associated with relapse. The anti-cancer role of such t cell phenotype is extensively supported [354] (although pro-tumour activity has also be observed). Specifically for myeloid leukemia, a fourfold increase of γδ t cells has been observed in patients with very early morphological or molecular relapse[355]; such t cells were also able to kill leukemic target cells in vitro. On the contrary, the monocyte population was identified as positively associated with relapse, as supported by previous studies[352], which identified a predictive power of the reduction of blood monocyte counts with leukemia-free survival after the first HDC/IL-2 treatment cycle. Furthermore, plasma cells were inferred to be highly negatively associated with relapse, suggesting their beneficial role in disease progression.

For colon adenocarcinoma, the cell type that was most negatively associated with tumour relapse was γδ t cells; which anti-cancer activity have been demonstrated for this cancer type, mediated by granule exocytosis and dependent on isoprenoid production by tumour cells[356]. However, the pro-anti-cancer activity balance of γδ t cells in colon adenocarcinoma is not fully established[354]. The cell type that was most positively associated with tumour relapse was macrophages; surprisingly the cell phenotype leading this change was M1. Such a phenotype of

macrophages has been mainly linked to anti-cancer activity[357]; however, M1 macrophages have been shown to damage endothelial monolayers in vitro[358], and that such damage can lead to increased tumour-cell adhesion to the vasculature[359]. Furthermore, the roles of M1 macrophages can change during the progression of the disease: going from an early stage elimination of tumour cells with the activation of an adaptive immunity response, to a late polarisation to pro-tumorigenic M2 macrophages after such response is ablated by cancer[358].

For hepatocellular carcinoma, T-cells had a negative association to relapse similarly to other cancers. Such association have been observed (together with other immunoscores) for both $CD3_+$ and $CD8_+$ linked to rate of recurrence as well as disease free survival[360]. On the contrary, macrophages had an overall increase in association with relapse, showing a switch toward M1 phenotype. Such counterintuitive result could be potentially caused by an erroneous classification of M2 macrophages as M1. However, some aspects of the mechanisms responsible for the regulation and maintenance of M1 and M2 polarization imbalance are still unclear[361].

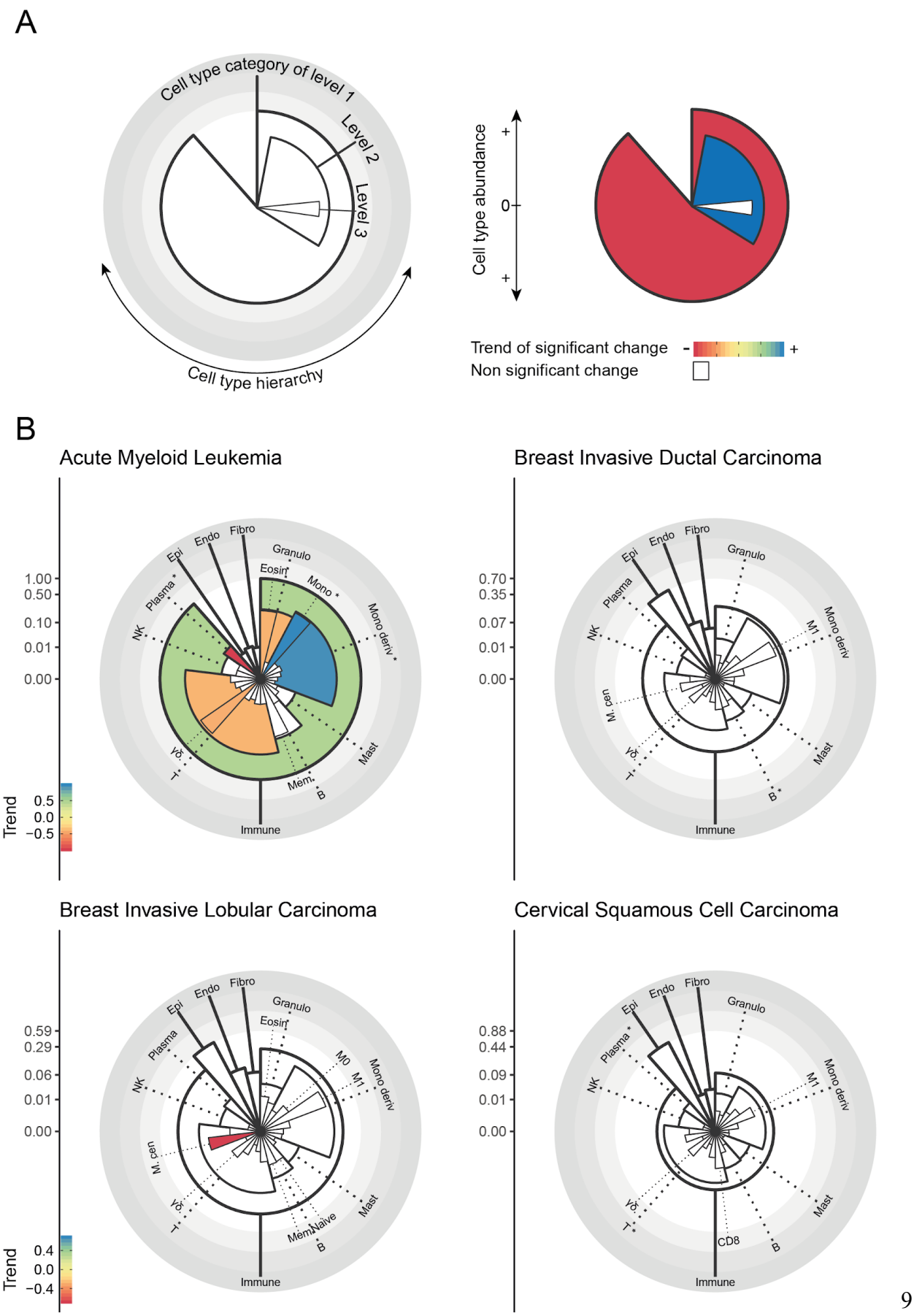
For lung squamous carcinoma, T-cell abundance had similar negative association with cancer relapse to most of other cancers. Previous studies[362] identified the association of the abundance of $CD3_+$ T-cells with improved overall survival, as well as the association of $CD4_+$ t cells (in the stromal extra tumour compartment) and of $CD8_+$ T-cells (in the tumour compartment) with increased overall and disease specific survival. The density of infiltrating t cells have been identified as powerful prognostic factor, more than standard pathological criteria[363]. On the contrary of most other cancers, the macrophage cell population (specifically M1 phenotype) was negatively associated with relapse. This result agrees with previous evidence[364], although a general consensus on the anti- or pro-tumour potential of M1 macrophages is missing[365]. Although rare in the tissue, both eosinophils and neutrophils seem to be negatively associated with cancer relapse.

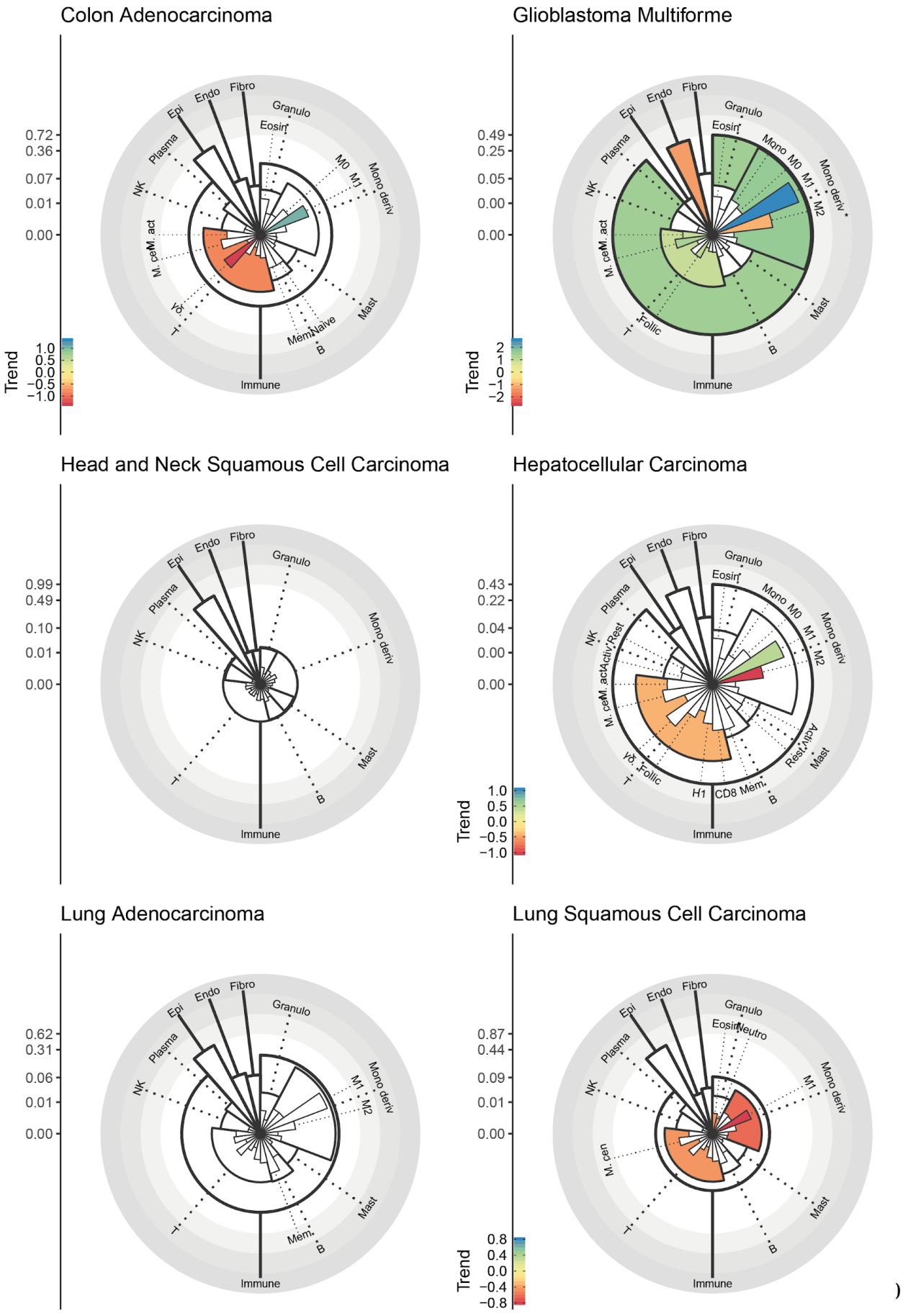
For pancreatic adenocarcinoma, a global negative association with relapse was inferred for the presence of immune cells. The main contributions came from t cells (including memory and γδ), mast cells and macrophages. For T-cells, monocyte derived cells and mast cell the direction of the association with relapse is not consistent with previous studies[366–368]. However, the concurrent decrease of all immune cell type, and the absence of evidence for any driver role from ARMET-tc inference is an indication that such overall decrease might be apparent in simplex space because of the increase of the epithelial component (estimate non-significant association = 0.57).

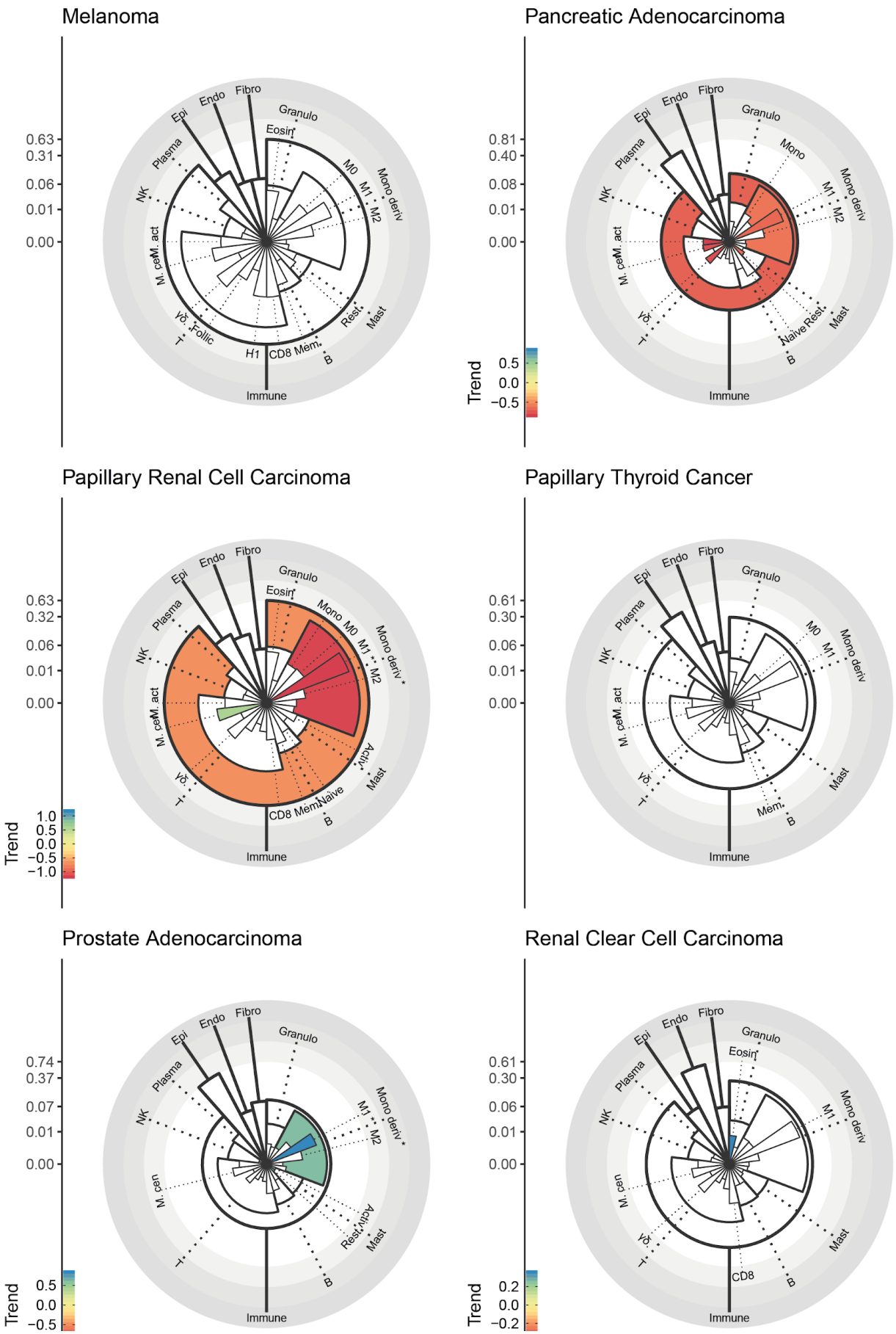
For serous ovarian cancer, the associations with relapse were similar to prostate cancer; with the abundance of the macrophage population, specifically the M1 phenotype, being positively correlated to worst outcome. Although several studies[369,370] support the specific role of tumour associated macrophages in cancer progression, it has been shown that in late stages of the disease cancer cells might become resilient to the toxic activity of M1 macrophages and modulate them to M2 phenotype, making their presence a risk factor. It is possible however that the challenges in distinguishing M1 from M2 macrophages may lead to an inaccurate inference of the two phenotypes.

For stomach adenocarcinoma, the association pattern was similar to pancreatic adenocarcinoma, involving overall immune cell abundance as a factor of positive outcome. The cell type category involved are t cells, granulocytes and monocyte/macrophages. For stomach adenocarcinoma, the ratio of neutrophil to lymphocytes is a prognostic indicator of overall and disease-free survival, with a low ratio being associated with better outcome. This shift in ratio is supported by ARMET-tc inference. Although both neutrophil and T-cell abundance is negatively associated with relapse, the expected neutrophil to lymphocytes ratio is respected due to the abundance of T-cells who drive the change. T-cells, and specifically the memory phenotype, are associated with better survival[371]. Also, the decreased expression of granulocyte-macrophage colony-stimulating factor has been associated with adverse clinical outcome[372].

For uterine endometrioid carcinoma, ARMET-tc inferred the overall immune cell infiltration as negatively associated with relapse, mainly involving T-cells with the phenotype memory and γδ.


A
Cell type category of level 1
Level 2
Level 3
Cell type hierarchy
Cell type abundance
+
0
+
Trend of significant change
-
+
Non significant change
B
Acute Myeloid Leukemia
Breast Invasive Ductal Carcinoma
Breast Invasive Lobular Carcinoma
Cervical Squamous Cell Carcinoma
Trend
Immune
Epi
Endo
Fibro
Granulo
Eosin
Mono
Mono deriv
Mast
B
Mem
Naive
M0
M1
M. cen
NK
Plasma
T
γδ
CD8


Colon Adenocarcinoma
Glioblastoma Multiforme
Head and Neck Squamous Cell Carcinoma
Hepatocellular Carcinoma
Lung Adenocarcinoma
Lung Squamous Cell Carcinoma
Trend
Immune
Granulo
Eosin
Neutro
Mono M0
Mono M1
Mono M2
Mono deriv
Mast
B
Mem
Naive
Follic
γδ
M. ceM. act
M. cen
Plasma
NK
Epi
Endo
Fibro
Rest
Activ
H1
CD8 Mem
)


Melanoma
0.63
0.31
0.06
0.01
0.00
Fibro
Endo
Epi
Granulo
Eosin
Plasma
M0
M1
M2
Mono deriv
NK
M. cen
M. act
Rest
Mast
γδ
Follic
H1
CD8 Mem
B
Immune
Pancreatic Adenocarcinoma
0.81
0.40
0.08
0.01
0.00
Mono
Naive Rest
Trend
0.5
0.0
−0.5
Papillary Renal Cell Carcinoma
0.63
0.32
0.06
0.01
0.00
Activ
Naive
1.0
0.5
0.0
−0.5
−1.0
Papillary Thyroid Cancer
0.61
0.30
0.06
0.01
0.00
Mem
Prostate Adenocarcinoma
0.74
0.37
0.07
0.01
0.00
M. cen
Rest
Activ
0.5
0.0
−0.5
Renal Clear Cell Carcinoma
0.61
0.30
0.06
0.01
0.00
CD8
0.2
0.0
−0.2

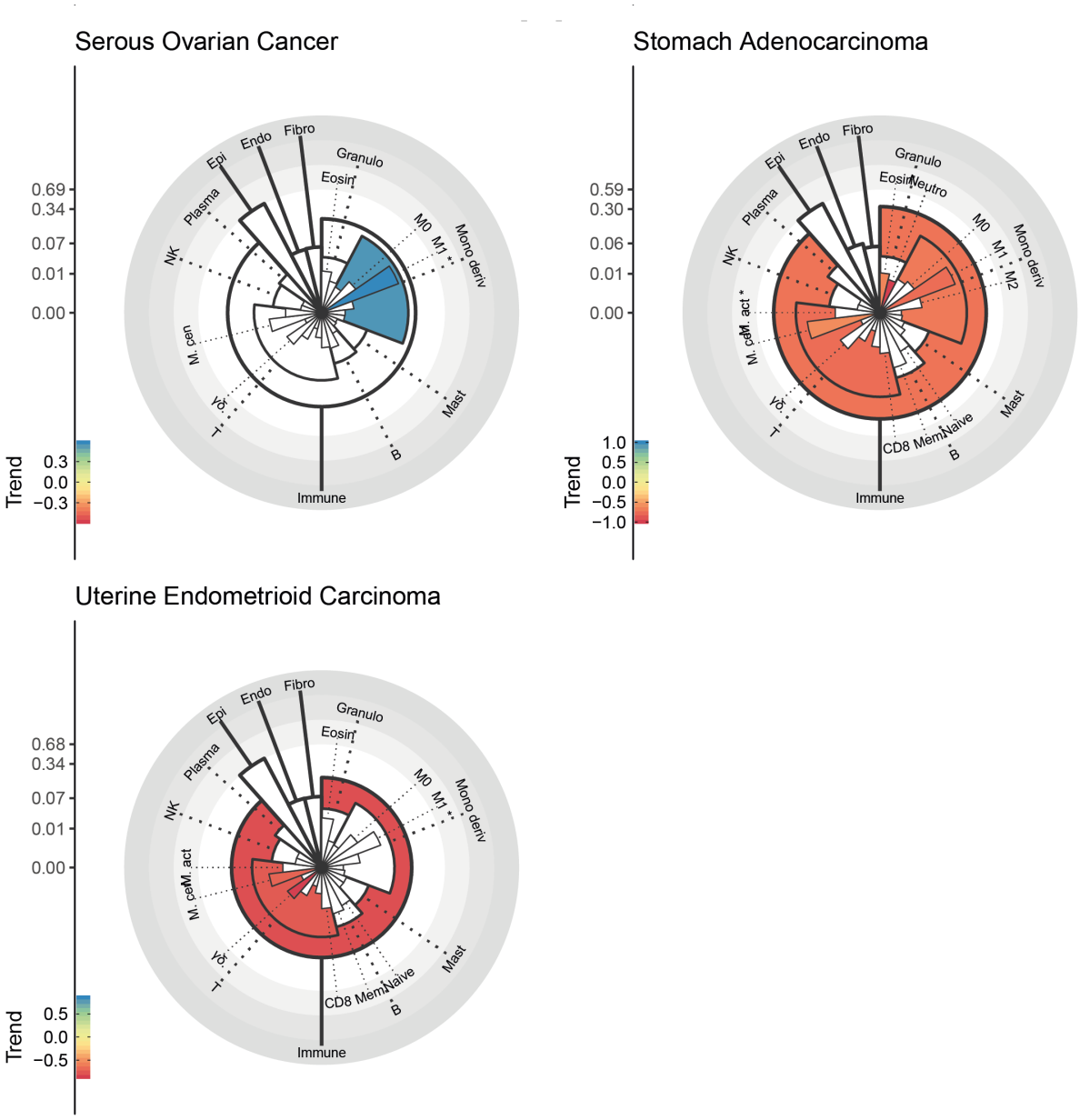


Figure 6.5: A - Schematics of the information content of the differential tissue composition plots. The radial axis (i.e., width of each slice) represent the cell type hierarchy (e.g., the slice representing the immune cell include the slice representing t-cells, which includes the slice representing memory t-cells); the y axis (i.e., the depth of each slice) represents the abundance of each cell type category; and the colour coding represent the degree of significant change of cell type proportion in association with cancer relapse (i.e., blue if positive or red if negative). B - Plots representing the cell abundance and the changes in cell type proportion associated with cancer relapse.

Conclusions

Here we presented ARMET-tc, a novel tool that allows for this first-time differential tissue composition analyses. This tool outperforms publicly available alternatives in deconvolving whole tissue RNA sequencing data, and exclusively performs an integrated analysis of the trends of change of tissue composition along covariates of interest preserving the uncertainty information though the whole inference process.

For the inference of tissue composition (i.e., gene transcription deconvolution), our method uses an innovative hierarchical approach that increase the information/noise ratio and transcriptional signatures autocorrelation. That is, rather than approaching the inference of tissue composition in an unstructured manner, ARMET-tc performs such task recursively across the cell differentiation hierarchy. The principle is that identifiable transcriptional signature that segregate diverse cell type categories exist at multiple level of such hierarchy: for example, a robust signature for immune cells that distinguish them from epithelial cells exist, as well as a robust signatures that allow to distinguish distinct immune cell types within the immune cell population. The isolation of such two inferences allows to better target the use of marker genes, avoiding the introduction of markers that even though are specific for a node of the hierarchy (e.g., phenotypes of b cells) are useless for another (e.g., phenotypes of epithelial cells). For example, a gene that segregates activated from central memory t cells is not informative on the segregation of epithelial from fibroblast. Thanks to a higher information/noise ratio and lower cell type transcriptional signatures autocorrelations, it is possible to model RNA sequencing data on a logarithmic scale. Such modelling strategy allows to avoid the marked heteroskedasticity of RNA sequencing data leading to better resolution for rare cell populations and changes therein. Furthermore, the use of a simplex for modelling the proportion of cell types within a tissue avoids the anomaly of negative proportion present in most of algorithms based on pure linear regression.

It should be noted that in its current form, ARMET (as well as the other alternative methods presented in this thesis) is based on a linear additivity assumption (Eq. 2; gene transcription across samples for each cell type is equal), however this is needed to keep the parameter space smaller than the data space; this is an acceptable assumption that becomes irrelevant when N genes are used as the random errors will cancer out for $N>>1$. Furthermore, ARMET (as well as the other alternative methods presented in this thesis) ignores a key bias, being the relation between cell size and absolute mRNA content. For example, macrophages have larger internal volume than T cells,

and have a larger mRNA content. Due to this, ARMET infers a biased measure of cell type proportion, that will be adjusted for cell type average size in future releases.

For the inference of changes in tissue composition, a novel simplex regression approach was designed and implemented [344]; which allows the detection of relative proportional changes (intrinsic perspective) as well as the detection of driver changes in absolute abundance within the tissue (extrinsic perspective). As shown for pancreatic adenocarcinoma, this aspect is particularly important as overall changes in relative proportions (e.g., decrease of immune cells) could be an artifactual effect driven by the change (i.e., increase of epithelial tumour cells) of another single cell type [344]. The Bayesian inference model employed by ARMET-tc allows to preserve the information about the uncertainty of the estimation of the tissue composition (step i) across the inference of its changes along a covariate of interest. This ability is fundamental in the context of transcriptional deconvolution as the uncertainty associated with step (i) is substantial. A key unique quality of our method is the report of the direct probabilistic uncertainty of each estimation to the user, including both tissue composition (deconvolution) and trends of change. The landscape of associations between tissue composition and cancer relapse for wide types of tumour types was revealed with higher consistency compared to previous studies and demonstrated the utility of ARMET-tc in a clinical context. Future directions are the improvement of the noise model for RNA sequencing data; the addition of microarray specific noise models; and the improvement of gene selection for cell-type-specific signatures, for example improving the internal separation of macrophage and T-cell phenotypes.

CHAPTER 7

Conclusions

In this thesis, I have demonstrated the importance of investigating the prostate tumour microenvironment at the molecular level, integrating experimental and computational tools. An improved understanding of prostate tumour microenvironment will open novel routes both for diagnosis and treatment. The diagnostic potential of molecular and cellular features of tumour microenvironment have been previously shown (for example[220]). Considering this potential as well the challenges that multifocality poses in the stratification of prostate cancer patients based on their tissue pathology, the identification of changes at systemic level that could have diagnostic power is an attractive route.

In Chapter 2, I showed the value in probing the adipose tissue surrounding the prostate at the molecular level. Being both near the cancer location and metabolically active, adipose tissue has the potential to be both affected by and/or affect cancer development. In this first part of the thesis, I have provided evidence that changes at the transcriptomic level happen in association with cancer grade. Such findings give hope for overcoming the multifocality limitation of prostate cancer in biopsies-based diagnosis.

In chapter 3, I provided further evidence for the molecular alterations that surrounding benign adipose tissue acquires in consequence of androgen deprivation therapy. Interestingly, the identified changes involve mostly inflammation and likely enrichment of inflammatory cells within the tissue. The relevance of the investigation of adipose tissue is increased by its accessibility, compared with prostate tissue.

The therapeutic potential of targeting benign elements of the tumour microenvironment has been proven, especially with immune checkpoint inhibitors for a wide range of cancers. Immune checkpoint inhibitors (i.e., PD-1 and PD-L1 inhibitor) have improved life expectancy for several cancer types, including melanoma and lung cancer. However, the benefits of these novel therapies are not translated to prostate cancer so far. Currently, there is an urgent need for further investigation of the prostate tumour microenvironment, with the goal of discovering new immune escape mechanism of prostate cancer, or alternative non-immune routes of adaptation and development. Chapters 4 have shown the value of probing prostate tissue at the site of primary prostate cancer. In this chapter, I have provided evidence that the molecular profiles of cancerous

and non-cancerous cells types are associated with the development of the disease. This investigation was aided by both an innovative experimental design approach and a novel statistical inference model for differential transcription analyses. The differential transcription analysis approach enabled inference of associations between gene abundances and a continuous risk score, rather than a binary label as it is commonly recommended (e.g., low- vs. high grade). This analysis, applied to enriched epithelial, fibroblasts, T- and myeloid cell types from cancerous prostate tissue, allowed the creation of landscape plots were transcriptomic changes were mapped to both the cellular source and CAPRA risk score, which is a surrogate of disease development if we assume that prostate cancer is characterised by a progressive nature. This approach favoured a much deeper hypothesis generation compared to classic differential transcriptional analysis approaches. The current approach can be viewed as in a middle ground between bulk tissue and single cell transcriptomic analysis. Enriching for key cell types before bar-coding allows for much more meaningful interpretations of the results, still allowing for a medium size patient cohort. A limitation to be considered is the number of cell types that can be enriched for each experiment with fluorescence-activated cell sorting (FACS) technology. Although serial collection of more than four cell types is possible for a given sample, the starting total number of cells in the sample might not allow that. A higher dimensional analysis on a minimal representative selection of patients will be possible when the cost linked to single cell sequencing technology will significantly drop.

In cases when the physical enrichment/isolation of single cell types is not possible, statistical models can be used to extract information about the tissue at the cellular and molecular levels, from bulk tissue molecular profiles (e.g., transcriptomic). For example, inferring the cellular composition of a selection of tumour samples, representative of clinical stages/conditions can inform about the importance of single cell types for cancer development and risk status.

Chapter 5 and 6 provide evidence that the abundance of some non-cancerous cell types are associated with the development of cancer. Two stand-alone statistical models have been developed in chapters 5 and 6 respectively. While the inference of tissue composition from bulk tissue transcriptomic data has been approached in the past, not much attention has been paid to connecting changes of cell type abundance to biological/clinical conditions. Chapter 5 was focused on this latter point, with the development of an improved regression model for proportional data. This inference model is able to estimate both the observable (i.e., intrinsic) and driver (i.e.,

extrinsic)[309] associations from proportional data. Besides being a stand-alone, this regression model was integrated with the deconvolution model implemented in chapter 6. With the integration of the deconvolution and regression statistical models, it is possible to perform robust differential tissue composition analyses from bulk tissue transcriptomic data. I compared the performance of the deconvolution model with representative publicly available algorithms. I also examined the performance on simulated data with increasing noise content. Finally, I applied the model for inferring associations between the abundance of specific cell types and cancer relapse across several cancer types in a prognostic setting. Using TCGA data, I produced a landscape of cell type abundance profiles, across a wide range of cancer types; and identified prognostic cellular signatures. The integrative nature of the two phases of the differential tissue composition analysis (i.e., deconvolution and regression) is important considering the uncertainty linked with the deconvolution phase. For example, the use of a deconvolution algorithm and a separate regression algorithm is possible, passing the point estimate of the first to the second model. However, since the deconvolution phase carries a high level of uncertainty, the regression phase will be biased toward false positive identification.

Future work

Beside contributing to improve the biological and statistical knowledge related to prostate tumour microenvironment, this thesis established several lines of works that can be developed further. Considering the promising results of chapter 2, an extended discovery cohort ($n \simeq 100$) of periprostatic adipose tissue may reveal an improved gene signature. This signature may be useful as a complementary tool for the stratification low- and high-grade tumours and reduce the chances of incorrect stratification due to high prostate cancer multifocality. Although the choice of a two-stage feature selection procedure allowed a cost-efficient operation, it also represents a limitation of the present study. An expanded discovery cohort would allow more robust feature selection, which would produce a more robust gene list to validate on an independent patient cohort, with an orthogonal technology (e.g., qRT-PCR). Furthermore, the addition of cellular tissue composition analysis with flow-cytometry technologies such as CyTOF or FACS would enrich the feature set that could be informative of the field effect of the tumour.

Future work should also be focused on investigating several hypotheses that chapter 4 has generated. For example, it appears that monocytes and macrophages play a major role in prostate

cancer. A coupled cell-type-enrichment and differential transcription analysis can be applied specifically to the diverse phenotype of monocyte-derived cells, in order to investigate further the role of each cell type category in prostate cancer progression. The role of immune cell in the metabolism of hormone-related molecules may be further investigated experimentally with metabolomics and proteomics approaches. The enrichment of hormone-related molecules in the extracellular matrix may be a side effect of inflammation, which may represent a potential therapeutic target. Furthermore, future work will be aimed to using the cell-type-specific transcriptional signatures identified in chapter 4 for patient risk stratification. Employing transcriptome deconvolution, the presence of the cell-type-specific gene signatures identified experimentally could be sought within publicly available data of bulk tissue transcriptomes of prostate cancer.

Methods developed in chapters 4, 5 and 6 significantly improve on the state of the art, however future work should aim to further improve the noise models. For example, the differential transcription inference model developed in chapter 4 could use a better noise model of RNA sequencing data. In the present thesis, a negative binomial distribution has been used to model the uncertainty of gene count observations; this distribution is convenient because it models discrete counts and the overdispersed noise typical of RNA sequencing data. Although the multinomial distribution better represents the numerical generating process that goes from sequencing to gene counts, it does not allow for the necessarily overdispersion. The Dirichlet-multinomial distribution improves on this but does not overcome it completely. A hierarchical normal-multinomial or student-t-multinomial may also be a valid alternative.

An improved noise model could be directly beneficial for the differential tissue composition analysis algorithm. Currently, the prior information about the gene-wise abundance of cell-type-specific signatures is used as point estimate. In reality, such abundance carries a variability that could be modelled as described above, allowing a better inference of the tissue composition. This hierarchical approach would integrate (i) the uncertainty linked to the transcriptional signatures of single cell types, (ii) the uncertainty linked to the inference of tissue composition, and (iii) the uncertainty linked with the inference of associations between cell type abundance and a biological/clinical factor of interest. Furthermore, the selection of the genes within the cell-type-specific transcriptional signatures can be improved. This selection would require a better cross-validation approach based on common differential transcription analysis

tools, or a fully probabilistic approach based on Bayesian inference. Given the hierarchical structure of the deconvolution model presented in this thesis, more cell types will be added to the model, allowing a deeper analysis of a given sample. The regression model for proportional data has also improvement margins. For example, although the Dirichlet distribution is suitable for modelling trends in proportional data, it does not account for the overdispersed noise proper of some data sources. For avoiding issues with outliers leading to false positive associations (i.e., apparent non-null trends of change), a hierarchical normal-multinomial or student-t-multinomial model could be introduced.

Final remarks

Both the genetic alteration of prostate cancer cells and the interaction of cancer cells with surrounding benign cells (e.g., immune system and stroma) play an essential role in prostate cancer development. To better diagnose and cure this disease, we need to take an integrative approach that tries to include both aspects in the investigative effort. Currently, the amount of resources invested in high-throughput genetic studies of prostate cancer (e.g., pan-cancer analysis of whole genomes (PCAWG); Pan Prostate Cancer Consortium; and other cancers) is skewed toward the study of genetic alteration (i.e., mutations, rearrangements, copy number variations) in cancer cells. This skewed approach underlies a belief that the cancer evolution can be understood in isolation, without considering the environment (i.e., patient specific physiology) that surrounds and interact with cancer cells. The present thesis represents an attempt to enrich the genomic and computational biology research with a microenvironmental focus, looking toward the more ambitious goal of integrating both aspects (i.e., tumour and non-tumour cell biology) of cancer development.

References

1. Comito, G. *et al.* Cancer-associated fibroblasts and M2-polarized macrophages synergize during prostate carcinoma progression. *Oncogene* **33,** 2423–2431 (2014).
2. Yu, Y. *et al.* Cancer-associated fibroblasts induce epithelial–mesenchymal transition of breast cancer cells through paracrine TGF-β signalling. *Br. J. Cancer* **110,** 724 (2013).
3. Yang, L., Pang, Y. & Moses, H. L. TGF-beta and immune cells: an important regulatory axis in the tumor microenvironment and progression. *Trends Immunol.* **31,** 220–227 (2010).
4. Lin, X. *et al.* PPM1A Functions as a Smad Phosphatase to Terminate TGFβ Signaling. *Cell* **166,** 1597 (2016).
5. Barron, D. A. & Rowley, D. R. The reactive stroma microenvironment and prostate cancer progression. *Endocr. Relat. Cancer* **19,** R187–204 (2012).
6. Tuxhorn, J. A., McAlhany, S. J., Dang, T. D., Ayala, G. E. & Rowley, D. R. Stromal cells promote angiogenesis and growth of human prostate tumors in a differential reactive stroma (DRS) xenograft model. *Cancer Res.* **62,** 3298–3307 (2002).
7. Joesting, M. S. *et al.* Identification of SFRP1 as a candidate mediator of stromal-to-epithelial signaling in prostate cancer. *Cancer Res.* **65,** 10423–10430 (2005).
8. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
9. Chung, L. W. K., Baseman, A., Assikis, V. & Zhau, H. E. Molecular insights into prostate cancer progression: the missing link of tumor microenvironment. *J. Urol.* **173,** 10–20 (2005).
10. Siddle, K. *et al.* Specificity in ligand binding and intracellular signalling by insulin and insulin-like growth factor receptors. *Biochem. Soc. Trans.* **29,** 513–525 (2001).
11. Bhowmick, N. A. *et al.* TGF-beta signaling in fibroblasts modulates the oncogenic potential of

adjacent epithelia. *Science* **303,** 848–851 (2004).

12. Ao, M. *et al.* Cross-talk between paracrine-acting cytokine and chemokine pathways promotes malignancy in benign human prostatic epithelium. *Cancer Res.* **67,** 4244–4253 (2007).
13. Orimo, A. & Weinberg, R. A. Stromal fibroblasts in cancer: a novel tumor-promoting cell type. *Cell Cycle* **5,** 1597–1601 (2006).
14. Lonergan, P. E. & Tindall, D. J. Androgen receptor signaling in prostate cancer development and progression. *J. Carcinog.* **10,** 20 (2011).
15. Tan, M. H. E., Li, J., Xu, H. E., Melcher, K. & Yong, E.-L. Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta Pharmacol. Sin.* **36,** 3–23 (2015).
16. Zhou, Y., Bolton, E. C. & Jones, J. O. Androgens and androgen receptor signaling in prostate tumorigenesis. *J. Mol. Endocrinol.* **54,** R15–29 (2015).
17. Yu, S. *et al.* Androgen receptor in human prostate cancer-associated fibroblasts promotes prostate cancer epithelial cell growth and invasion. *Med. Oncol.* **30,** 674 (2013).
18. Memarzadeh, S. *et al.* Role of autonomous androgen receptor signaling in prostate cancer initiation is dichotomous and depends on the oncogenic signal. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 7962–7967 (2011).
19. Turley, S. J., Cremasco, V. & Astarita, J. L. Immunological hallmarks of stromal cells in the tumour microenvironment. *Nat. Rev. Immunol.* **15,** 669–682 (2015).
20. Joyce, J. A. & Pollard, J. W. Microenvironmental regulation of metastasis. *Nat. Rev. Cancer* **9,** 239–252 (2009).
21. Psaila, B. & Lyden, D. The metastatic niche: adapting the foreign soil. *Nat. Rev. Cancer* **9,** 285–293 (2009).
22. Nemeth, Z. *et al.* Heme oxygenase-1 in macrophages controls prostate cancer progression. *Oncotarget* **6,** 33675–33688 (2015).
23. Bryant, G., Wang, L. & Mulholland, D. J. Overcoming Oncogenic Mediated Tumor Immunity in Prostate Cancer. *Int. J. Mol. Sci.* **18,** (2017).

24. Holmgaard, R. B., Zamarin, D., Lesokhin, A., Merghoub, T. & Wolchok, J. D. Targeting myeloid-derived suppressor cells with colony stimulating factor-1 receptor blockade can reverse immune resistance to immunotherapy in indoleamine 2,3-dioxygenase-expressing tumors. *EBioMedicine* **6,** 50–58 (2016).
25. Caescu, C. I. *et al.* Colony stimulating factor-1 receptor signaling networks inhibit mouse macrophage inflammatory responses by induction of microRNA-21. *Blood* **125,** e1–13 (2015).
26. Dai, X.-M. *et al.* Targeted disruption of the mouse colony-stimulating factor 1 receptor gene results in osteopetrosis, mononuclear phagocyte deficiency, increased primitive progenitor cell frequencies, and reproductive defects. *Blood* **99,** 111–120 (2002).
27. Trikha, P. & Carson, W. E., 3rd. Signaling pathways involved in MDSC regulation. *Biochim. Biophys. Acta* **1846,** 55–65 (2014).
28. Zhu, Y. *et al.* CSF1/CSF1R blockade reprograms tumor-infiltrating macrophages and improves response to T-cell checkpoint immunotherapy in pancreatic cancer models. *Cancer Res.* **74,** 5057–5069 (2014).
29. Garcia, A. J. *et al.* Pten null prostate epithelium promotes localized myeloid-derived suppressor cell expansion and immune suppression during tumor initiation and progression. *Mol. Cell. Biol.* **34,** 2017–2028 (2014).
30. Koh, T. J. & DiPietro, L. A. Inflammation and wound healing: the role of the macrophage. *Expert Rev. Mol. Med.* **13,** e23 (2011).
31. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119,** 1420–1428 (2009).
32. Brawer, M. K., Deering, R. E., Brown, M., Preston, S. D. & Bigler, S. A. Predictors of pathologic stage in prostatic carcinoma. The role of neovascularity. *Cancer* **73,** 678–687 (1994).
33. Giannoni, E., Bianchini, F., Calorini, L. & Chiarugi, P. Cancer associated fibroblasts exploit reactive oxygen species through a proinflammatory signature leading to epithelial mesenchymal transition and stemness. *Antioxid. Redox Signal.* **14,** 2361–2371 (2011).

34. Smith, B. A. *et al.* A basal stem cell signature identifies aggressive prostate cancer phenotypes. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E6544–52 (2015).

35. Zhang, D. *et al.* Stem cell and neurogenic gene-expression profiles link prostate basal cells to aggressive prostate cancer. *Nat. Commun.* **7,** 10798 (2016).

36. Nishida, N., Yano, H., Nishida, T., Kamura, T. & Kojiro, M. Angiogenesis in cancer. *Vasc. Health Risk Manag.* **2,** 213–219 (2006).

37. Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nat. Med.* **19,** 1423–1437 (2013).

38. Riabov, V. *et al.* Role of tumor associated macrophages in tumor angiogenesis and lymphangiogenesis. *Front. Physiol.* **5,** 75 (2014).

39. Levine, A. C. *et al.* Androgens induce the expression of vascular endothelial growth factor in human fetal prostatic fibroblasts. *Endocrinology* **139,** 4672–4678 (1998).

40. Chen, P.-C. *et al.* Prostate cancer-derived CCN3 induces M2 macrophage infiltration and contributes to angiogenesis in prostate cancer microenvironment. *Oncotarget* **5,** 1595–1608 (2014).

41. Zahalka, A. H. *et al.* Adrenergic nerves activate an angio-metabolic switch in prostate cancer. *Science* **358,** 321–326 (2017).

42. Chay, C. H. *et al.* A functional thrombin receptor (PAR1) is expressed on bone-derived prostate cancer cell lines. *Urology* **60,** 760–765 (2002).

43. Sun, Y.-X. *et al.* Skeletal localization and neutralization of the SDF-1(CXCL12)/CXCR4 axis blocks prostate cancer metastasis and growth in osseous sites in vivo. *J. Bone Miner. Res.* **20,** 318–329 (2005).

44. Edlund, M. *et al.* Integrin expression and usage by prostate cancer cell lines on laminin substrata. *Cell Growth Differ.* **12,** 99–107 (2001).

45. Glinskii, O. V. *et al.* Mechanical entrapment is insufficient and intercellular adhesion is essential for metastatic cell arrest in distant organs. *Neoplasia* **7,** 522–527 (2005).

46. Morrissey, C. & Vessella, R. L. The role of tumor microenvironment in prostate cancer bone

metastasis. *J. Cell. Biochem.* **101,** 873–886 (2007).

47. Gartrell, B. A. & Saad, F. Managing bone metastases and reducing skeletal related events in prostate cancer. *Nat. Rev. Clin. Oncol.* **11,** 335–345 (2014).
48. Dai, J., Hensel, J., Wang, N., Kruithof-de Julio, M. & Shiozawa, Y. Mouse models for studying prostate cancer bone metastasis. *Bonekey Rep* **5,** 777 (2016).
49. Woolf, D. K., Padhani, A. R. & Makris, A. Assessing response to treatment of bone metastases from breast cancer: what should be the standard of care? *Ann. Oncol.* **26,** 1048–1057 (2015).
50. Frieling, J. S., Pamen, L. A., Cook, L. M., Yang, S. & Lynch, C. C. Abstract 5061: Roles for matrix metalloproteinase-3 (MMP-3) in the prostate tumor-bone microenvironment. *Cancer Res.* **73,** 5061–5061 (2013).
51. Hienert, G., Kirchheimer, J. C., Christ, G., Pflüger, H. & Binder, B. R. Plasma urokinase-type plasminogen activator correlates to bone scintigraphy in prostatic carcinoma. *Eur. Urol.* **15,** 256–258 (1988).
52. Logothetis, C. J. & Lin, S.-H. Osteoblasts in prostate cancer metastasis to bone. *Nat. Rev. Cancer* **5,** 21–28 (2005).
53. Mayr-Wohlfart, U. *et al.* Vascular endothelial growth factor stimulates chemotactic migration of primary human osteoblasts. *Bone* **30,** 472–477 (2002).
54. Ide, H. *et al.* Growth regulation of human prostate cancer cells by bone morphogenetic protein-2. *Cancer Res.* **57,** 5022–5027 (1997).
55. Cornish, J. *et al.* Stimulation of osteoblast proliferation by C-terminal fragments of parathyroid hormone-related protein. *J. Bone Miner. Res.* **14,** 915–922 (1999).
56. Chen, H.-L. *et al.* Parathyroid hormone and parathyroid hormone-related protein exert both pro- and anti-apoptotic effects in mesenchymal cells. *J. Biol. Chem.* **277,** 19374–19381 (2002).
57. Karaplis, A. C. & Vautour, L. Parathyroid hormone-related peptide and the parathyroid hormone/parathyroid hormone-related peptide receptor in skeletal development. *Curr. Opin. Nephrol. Hypertens.* **6,** 308–313 (1997).

58. Brown, J. M. *et al.* Osteoprotegerin and rank ligand expression in prostate cancer. *Urology* **57,** 611–616 (2001).

59. Brown, J. M. *et al.* Serum osteoprotegerin levels are increased in patients with advanced prostate cancer. *Clin. Cancer Res.* **7,** 2977–2983 (2001).

60. Boyle, W. J., Simonet, W. S. & Lacey, D. L. Osteoclast differentiation and activation. *Nature* **423,** 337–342 (2003).

61. Josson, S., Matsuoka, Y., Chung, L. W. K., Zhau, H. E. & Wang, R. Tumor-stroma co-evolution in prostate cancer progression and metastasis. *Semin. Cell Dev. Biol.* **21,** 26–32 (2010).

62. Zheng, Y. *et al.* Targeting IL-6 and RANKL signaling inhibits prostate cancer growth in bone. *Clin. Exp. Metastasis* **31,** 921–933 (2014).

63. Lynch, C. C. *et al.* MMP-7 promotes prostate cancer-induced osteolysis via the solubilization of RANKL. *Cancer Cell* **7,** 485–496 (2005).

64. Voon, D. C., Huang, R. Y., Jackson, R. A. & Thiery, J. P. The EMT spectrum and therapeutic opportunities. *Mol. Oncol.* **11,** 878–891 (2017).

65. Abdalla, A. M. E. *et al.* Current Challenges of Cancer Anti-angiogenic Therapy and the Promise of Nanotherapeutics. *Theranostics* **8,** 533–548 (2018).

66. Salvatore, V. *et al.* The tumor microenvironment promotes cancer progression and cell migration. *Oncotarget* **8,** 9608–9616 (2017).

67. Katt, M. E., Placone, A. L., Wong, A. D., Xu, Z. S. & Searson, P. C. In Vitro Tumor Models: Advantages, Disadvantages, Variables, and Selecting the Right Platform. *Front Bioeng Biotechnol* **4,** 12 (2016).

68. Venet, D., Pecasse, F., Maenhaut, C. & Bersini, H. Separation of samples into their constituents using gene expression data. *Bioinformatics* **17 Suppl 1,** S279–87 (2001).

69. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).

70. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. L. & Liu, Z. Digital sorting of complex tissues for

cell type-specific gene expression profiles. *BMC Bioinformatics* **14,** 89 (2013).

71. Qiao, W. *et al.* PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLoS Comput. Biol.* **8,** e1002838 (2012).
72. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18,** 220 (2017).
73. Liebner, D. A., Huang, K. & Parvin, J. D. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics* **30,** 682–689 (2014).
74. Wang, N. *et al.* Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **6,** 18909 (2016).
75. Clarke, J., Seo, P. & Clarke, B. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics* **26,** 1043–1049 (2010).
76. Ahn, J. *et al.* DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics* **29,** 1865–1871 (2013).
77. Quon, G. & Morris, Q. ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics* **25,** 2882–2889 (2009).
78. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29,** 1083–1085 (2013).
79. Li, Y. & Xie, X. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics* **14 Suppl 5,** S11 (2013).
80. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).
81. Altboum, Z. *et al.* Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.* **10,** 720 (2014).
82. Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat.*

Methods **7,** 287–289 (2010).

83. Zuckerman, N. S., Noam, Y., Goldsmith, A. J. & Lee, P. P. A self-directed method for cell-type identification and separation of gene expression microarrays. *PLoS Comput. Biol.* **9,** e1003189 (2013).
84. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17,** 218 (2016).
85. Şenbabaoğlu, Y. *et al.* Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* **17,** 231 (2016).
86. Tappeiner, E. *et al.* TIminer: NGS data mining pipeline for cancer immunology and immunotherapy. *Bioinformatics* **33,** 3140–3141 (2017).
87. Li, B. *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17,** 174 (2016).
88. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6,** (2017).
89. Finotello, F. *et al.* quanTIseq: quantifying immune contexture of human tumors. *bioRxiv* 223180 (2017). doi:10.1101/223180
90. Qi, L. *et al.* Deconvolution of the gene expression profiles of valuable banked blood specimens for studying the prognostic values of altered peripheral immune cell proportions in cancer patients. *PLoS One* **9,** e100934 (2014).
91. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* **4,** e6098 (2009).
92. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **22,** 79–86 (1951).
93. Turlach, B. A. & Weingessel, A. quadprog: Functions to solve quadratic programming problems. *R package version* 1–4 (2007).

94. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **3,** 993–1022 (2003).

95. GSEA (gene set enrichment analysis). in *SpringerReference* (Springer-Verlag, 2011).

96. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

97. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30,** 413–421 (2012).

98. Eureqa Formulize. Available at: http://nutonian.wikidot.com/. (Accessed: 9th May 2018)

99. Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **7,** 308–313 (1965).

100. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167,** 1897 (2016).

101. Abugessaisa, I. *et al.* FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database* **2016,** (2016).

102. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).

103. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6,** 8971 (2015).

104. Zheng, X., Zhang, N., Wu, H.-J. & Wu, H. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.* **18,** 17 (2017).

105. Benelli, M., Romagnoli, D. & Demichelis, F. Tumor purity quantification by clonal DNA methylation signatures. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty011

106. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).

107. Gagnon-Bartsch, J. A. Removing Unwanted Variation from Microarray Data with Negative

Controls. (UC Berkeley, 2012).

108. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

109. Titus, A. J., Gallimore, R. M., Salas, L. A. & Christensen, B. C. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum. Mol. Genet.* **26,** R216–R224 (2017).

110. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J. Clin.* **65,** 87–108 (2015).

111. O'Brien, B. A., Cohen, R. J., Ryan, A., Sengupta, S. & Mills, J. A new preoperative nomogram to predict minimal prostate cancer: accuracy and error rates compared to other tools to select patients for active surveillance. *J. Urol.* **186,** 1811–1817 (2011).

112. Hamdy, F. C. *et al.* 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. *N. Engl. J. Med.* **375,** 1415–1424 (2016).

113. Lu-Yao, G. L. *et al.* Outcomes of localized prostate cancer following conservative management. *JAMA* **302,** 1202–1209 (2009).

114. Parker, C., Muston, D., Melia, J., Moss, S. & Dearnaley, D. A model of the natural history of screen-detected prostate cancer, and the effect of radical treatment on overall survival. *Br. J. Cancer* **94,** 1361–1368 (2006).

115. Corcoran, N. M. *et al.* Upgrade in Gleason score between prostate biopsies and pathology following radical prostatectomy significantly impacts upon the risk of biochemical recurrence. *BJU Int.* **108,** E202–10 (2011).

116. Corcoran, N. M. *et al.* Underestimation of Gleason score at prostate biopsy reflects sampling error in lower volume tumours. *BJU Int.* **109,** 660–664 (2012).

117. Michaelson, M. D. *et al.* Management of complications of prostate cancer treatment. *CA Cancer J. Clin.* **58,** 196–213 (2008).

118. Gore, J. L. *et al.* Optimal combinations of systematic sextant and laterally directed biopsies for the detection of prostate cancer. *J. Urol.* **165,** 1554–1559 (2001).

119. Ahmed, H. U. *et al.* Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate

cancer (PROMIS): a paired validating confirmatory study. *Lancet* **389,** 815–822 (2017).

120. Prieto-Hontoria, P. L. *et al.* Role of obesity-associated dysfunctional adipose tissue in cancer: a molecular nutrition approach. *Biochim. Biophys. Acta* **1807,** 664–678 (2011).

121. Møller, H. *et al.* Prostate cancer incidence, clinical stage and survival in relation to obesity: a prospective cohort study in Denmark. *Int. J. Cancer* **136,** 1940–1947 (2015).

122. Snowdon, D. A., Phillips, R. L. & Choi, W. Diet, obesity, and risk of fatal prostate cancer. *Am. J. Epidemiol.* **120,** 244–250 (1984).

123. Andersson, S. O. *et al.* Body size and prostate cancer: a 20-year follow-up study among 135006 Swedish construction workers. *J. Natl. Cancer Inst.* **89,** 385–389 (1997).

124. Calle, E. E., Rodriguez, C., Walker-Thurmond, K. & Thun, M. J. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N. Engl. J. Med.* **348,** 1625–1638 (2003).

125. Freedland, S. J., Bañez, L. L., Sun, L. L., Fitzsimons, N. J. & Moul, J. W. Obese men have higher-grade and larger tumors: an analysis of the duke prostate center database. *Prostate Cancer Prostatic Dis.* **12,** 259–263 (2009).

126. Venkatasubramanian, P. N. *et al.* Periprostatic adipose tissue from obese prostate cancer patients promotes tumor and endothelial cell proliferation: a functional and MR imaging pilot study. *Prostate* **74,** 326–335 (2014).

127. Cheng, L. *et al.* Correlation of margin status and extraprostatic extension with progression of prostate carcinoma. *Cancer* **86,** 1775–1782 (1999).

128. Stenman, U.-H. Re: Periprostatic Adipose Tissue as a Modulator of Prostate Cancer Aggressiveness. *Eur. Urol.* **57,** 541–542 (2010).

129. van Roermund, J. G. H. *et al.* Periprostatic fat correlates with tumour aggressiveness in prostate cancer patients. *BJU Int.* **107,** 1775–1779 (2011).

130. Ribeiro, R. *et al.* Human periprostatic adipose tissue promotes prostate cancer aggressiveness in vitro. *J. Exp. Clin. Cancer Res.* **31,** 32 (2012).

131. Ribeiro, R. J. T. *et al.* Tumor cell-educated periprostatic adipose tissue acquires an aggressive cancer-promoting secretory profile. *Cell. Physiol. Biochem.* **29,** 233–240 (2012).

132. Kerger, M. *et al.* Microscopic assessment of fresh prostate tumour specimens yields significantly increased rates of correctly annotated samples for downstream analysis. *Pathology* **44,** 204–208 (2012).

133. Klein, E. A. *et al.* A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. *Eur. Urol.* **66,** 550–560 (2014).

134. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. (2010).

135. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

136. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

137. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

138. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32,** 896–902 (2014).

139. Tarca, A. L. *et al.* A novel signaling pathway impact analysis. *Bioinformatics* **25,** 75–82 (2009).

140. Bennett, K. P. & Demiriz, A. Semi-Supervised Support Vector Machines. in *Advances in Neural Information Processing Systems 11* (eds. Kearns, M. J., Solla, S. A. & Cohn, D. A.) 368–374 (MIT Press, 1999).

141. Svetnik, V. *et al.* Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43,** 1947–1958 (2003).

142. Dobson, A. J. & Barnett, A. *An Introduction to Generalized Linear Models, Third Edition*. (Taylor & Francis, 2008).

143. Kuhn, M. & Others. Caret package. *J. Stat. Softw.* **28,** 1–26 (2008).

144. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer.

Cell **163,** 1011–1025 (2015).

145. Cooperberg, M. R., Hilton, J. F. & Carroll, P. R. The CAPRA-S score: a straightforward tool for improved prediction of outcomes after radical prostatectomy. *Cancer* **117,** 5039–5046 (2011).

146. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).

147. Liu, R. *et al.* Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res.* **43,** e97 (2015).

148. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).

149. Wilkinson, G. N. & Rogers, C. E. Symbolic Description of Factorial Models for Analysis of Variance. *J. R. Stat. Soc. Ser. C Appl. Stat.* **22,** 392–399 (1973).

150. Nafie, S., Pal, R. P., Dormer, J. P. & Khan, M. A. Transperineal template prostate biopsies in men with raised PSA despite two previous sets of negative TRUS-guided prostate biopsies. *World J. Urol.* **32,** 971–975 (2014).

151. Pokorny, M. R. *et al.* Prospective study of diagnostic accuracy comparing prostate cancer detection by transrectal ultrasound-guided biopsy versus magnetic resonance (MR) imaging with subsequent MR-guided biopsy in men without previous prostate biopsies. *Eur. Urol.* **66,** 22–29 (2014).

152. Cooperberg, M. R. *et al.* Validation of a cell-cycle progression gene panel to improve risk stratification in a contemporary prostatectomy cohort. *J. Clin. Oncol.* **31,** 1428–1434 (2013).

153. Wei, L. *et al.* Intratumoral and Intertumoral Genomic Heterogeneity of Multifocal Localized Prostate Cancer Impacts Molecular Classifications and Genomic Prognosticators. *Eur. Urol.* **71,** 183–192 (2017).

154. Schlomm, T. *et al.* Molecular cancer phenotype in normal prostate tissue. *Eur. Urol.* **55,** 885–890 (2009).

155. Risk, M. C. *et al.* Differential gene expression in benign prostate epithelium of men with and without prostate cancer: evidence for a prostate cancer field effect. *Clin. Cancer Res.* **16,** 5414–5423 (2010).

156. Kosari, F. *et al.* Shared gene expression alterations in prostate cancer and histologically benign prostate from patients with prostate cancer. *Am. J. Pathol.* **181,** 34–42 (2012).

157. Magi-Galluzzi, C. *et al.* Gene expression in normal-appearing tissue adjacent to prostate cancers are predictive of clinical outcome: evidence for a biologically meaningful field effect. *Oncotarget* **7,** 33855–33865 (2016).

158. Taylor, R. A., Lo, J., Ascui, N. & Watt, M. J. Linking obesogenic dysregulation to prostate cancer progression. *Endocr Connect* **4,** R68–80 (2015).

159. Sacca, P. A. *et al.* Human periprostatic adipose tissue: its influence on prostate cancer cells. *Cell. Physiol. Biochem.* **30,** 113–122 (2012).

160. Laurent, V. *et al.* Periprostatic adipocytes act as a driving force for prostate cancer progression in obesity. *Nat. Commun.* **7,** 10230 (2016).

161. Zhang, Q., Sun, L.-J., Yang, Z.-G., Zhang, G.-M. & Huo, R.-C. Influence of adipocytokines in periprostatic adipose tissue on prostate cancer aggressiveness. *Cytokine* **85,** 148–156 (2016).

162. Guaita-Esteruelas, S., Gumà, J., Masana, L. & Borràs, J. The peritumoural adipose tissue microenvironment and cancer. The roles of fatty acid binding protein 4 and fatty acid binding protein 5. *Mol. Cell. Endocrinol.* **462,** 107–118 (2018).

163. Shalapour, S. *et al.* Immunosuppressive plasma cells impede T-cell-dependent immunogenic chemotherapy. *Nature* **521,** 94–98 (2015).

164. Li, H. *et al.* Olfactomedin 4 deficiency promotes prostate neoplastic progression and is associated with upregulation of the hedgehog-signaling pathway. *Sci. Rep.* **5,** 16974 (2015).

165. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).

166. Clemmensen, S. N. *et al.* Olfactomedin 4 defines a subset of human neutrophils. *J. Leukoc. Biol.* **91,** 495–500 (2012).

167. Liu, H. Y. & Zhang, C. J. Identification of differentially expressed genes and their upstream regulators in colorectal cancer. *Cancer Gene Ther.* **24,** 244–250 (2017).

168. McKiernan, J. *et al.* A Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer at Initial Biopsy. *JAMA Oncol* **2,** 882–889 (2016).

169. Tomlins, S. A. *et al.* Urine TMPRSS2:ERG Plus PCA3 for Individualized Prostate Cancer Risk Assessment. *Eur. Urol.* **70,** 45–53 (2016).

170. Van Neste, L. *et al.* Detection of High-grade Prostate Cancer Using a Urinary Molecular Biomarker-Based Risk Score. *Eur. Urol.* **70,** 740–748 (2016).

171. Huggins, C. & Hodges, C. V. Studies on prostatic cancer. *Cancer Res.* (1941).

172. James, N. D. *et al.* Survival with Newly Diagnosed Metastatic Prostate Cancer in the 'Docetaxel Era': Data from 917 Patients in the Control Arm of the STAMPEDE Trial (MRC PR08, CRUK/06/019). *Eur. Urol.* **67,** 1028–1038 (2015).

173. Rhee, H. *et al.* Adverse effects of androgen-deprivation therapy in prostate cancer and their management. *BJU Int.* **115 Suppl 5,** 3–13 (2015).

174. Faris, J. E. & Smith, M. R. Metabolic sequelae associated with androgen deprivation therapy for prostate cancer. *Curr. Opin. Endocrinol. Diabetes Obes.* **17,** 240–246 (2010).

175. Braunstein, L. Z., Chen, M.-H., Loffredo, M., Kantoff, P. W. & D'Amico, A. V. Obesity and the Odds of Weight Gain following Androgen Deprivation Therapy for Prostate Cancer. *Prostate Cancer* **2014,** 230812 (2014).

176. Comitato, R., Saba, A., Turrini, A., Arganini, C. & Virgili, F. Sex hormones and macronutrient metabolism. *Crit. Rev. Food Sci. Nutr.* **55,** 227–241 (2015).

177. van Londen, G. J., Levy, M. E., Perera, S., Nelson, J. B. & Greenspan, S. L. Body composition changes during androgen deprivation therapy for prostate cancer: a 2-year prospective study. *Crit. Rev. Oncol. Hematol.* **68,** 172–177 (2008).

178. Smith, J. C. *et al.* The effects of induced hypogonadism on arterial stiffness, body composition, and metabolic parameters in males with prostate cancer. *J. Clin. Endocrinol. Metab.* **86,** 4261–4267 (2001).

179. Dockery, F., Bulpitt, C. J., Agarwal, S., Donaldson, M. & Rajkumar, C. Testosterone suppression in

men with prostate cancer leads to an increase in arterial stiffness and hyperinsulinaemia. *Clin. Sci.* **104,** 195–201 (2003).

180. Smith, M. R., Lee, H. & Nathan, D. M. Insulin sensitivity during combined androgen blockade for prostate cancer. *J. Clin. Endocrinol. Metab.* **91,** 1305–1308 (2006).
181. Freedland, S. J. & Aronson, W. J. Examining the relationship between obesity and prostate cancer. *Rev. Urol.* **6,** 73–81 (2004).
182. Chow, K. *et al.* Obesity suppresses tumor attributable PSA, affecting risk categorization. *Endocr. Relat. Cancer* **25,** 561–568 (2018).
183. Amling, C. L. *et al.* Relationship between obesity and race in predicting adverse pathologic variables in patients undergoing radical prostatectomy. *Urology* **58,** 723–728 (2001).
184. Rohrmann, S., Roberts, W. W., Walsh, P. C. & Platz, E. A. Family history of prostate cancer and obesity in relation to high-grade disease and extraprostatic extension in young men with prostate cancer. *Prostate* **55,** 140–146 (2003).
185. Mydlo, J. H., Tieng, N. L., Volpe, M. A., Chaiken, R. & Kral, J. G. A pilot study analyzing PSA, serum testosterone, lipid profile, body mass index and race in a small sample of patients with and without carcinoma of the prostate. *Prostate Cancer Prostatic Dis.* **4,** 101–105 (2001).
186. Freedland, S. J. *et al.* Impact of obesity on biochemical control after radical prostatectomy for clinically localized prostate cancer: a report by the Shared Equal Access Regional Cancer Hospital database study group. *J. Clin. Oncol.* **22,** 446–453 (2004).
187. Amling, C. L. *et al.* Pathologic variables and recurrence rates as related to obesity and race in men with prostate cancer undergoing radical prostatectomy. *J. Clin. Oncol.* **22,** 439–445 (2004).
188. Rodriguez, C. *et al.* Body mass index, height, and prostate cancer mortality in two large cohorts of adult men in the United States. *Cancer Epidemiol. Biomarkers Prev.* **10,** 345–353 (2001).
189. Mangiola, S. *et al.* Periprostatic fat tissue transcriptome reveals a signature diagnostic for high-risk prostate cancer. *Endocr. Relat. Cancer* **25,** 569–581 (2018).
190. Alhamdoosh, M., Ng, M. & Ritchie, M. E. EGSEA: Ensemble of Gene Set Enrichment Analyses. *R*

package version **1,** (2017).

191. Das, S. K., Ma, L. & Sharma, N. K. Adipose tissue gene expression and metabolic health of obese adults. *Int. J. Obes.* **39,** 869–873 (2015).
192. Maier, M. J. DirichletReg: Dirichlet regression for compositional data in R. (2014).
193. Carpenter, B. *et al.* Stan: A probabilistic programming language. *J. Stat. Softw.* **20,** 1–37 (2016).
194. Botstein, D. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).
195. Nebert, D. W. & Russell, D. W. Clinical importance of the cytochromes P450. *Lancet* **360,** 1155–1162 (2002).
196. Sampath, H. & Ntambi, J. M. The role of fatty acid desaturases in epidermal metabolism. *Dermatoendocrinol.* **3,** 62–64 (2011).
197. Bianco, A. C. & Kim, B. W. Deiodinases: implications of the local control of thyroid hormone action. *J. Clin. Invest.* **116,** 2571–2579 (2006).
198. Panigrahi, S. K., Manterola, M. & Wolgemuth, D. J. Meiotic failure in cyclin A1-deficient mouse spermatocytes triggers apoptosis through intrinsic and extrinsic signaling pathways and 14-3-3 proteins. *PLoS One* **12,** e0173926 (2017).
199. Lefranc, M.-P. Immunoglobulin and T Cell Receptor Genes: IMGT® and the Birth and Rise of Immunoinformatics. *Front. Immunol.* **5,** (2014).
200. Ressler, S. J. *et al.* WFDC1 is a key modulator of inflammatory and wound repair responses. *Am. J. Pathol.* **184,** 2951–2964 (2014).
201. Jones, R. J., Dickerson, S., Bhende, P. M., Delecluse, H.-J. & Kenney, S. C. Epstein-Barr virus lytic infection induces retinoic acid-responsive genes through induction of a retinol-metabolizing enzyme, DHRS9. *J. Biol. Chem.* **282,** 8317–8324 (2007).
202. Riquelme, P. *et al.* DHRS9 Is a Stable Marker of Human Regulatory Macrophages. *Transplantation* **101,** 2731–2738 (2017).
203. Xiang, J. *et al.* Regulation of Intestinal Epithelial Calcium Transport Proteins by Stanniocalcin-1 in Caco2 Cells. *Int. J. Mol. Sci.* **17,** (2016).

204. Baumgartner, R. N. Body composition in healthy aging. *Ann. N. Y. Acad. Sci.* **904,** 437–448 (2000).

205. Stenholm, S. *et al.* Sarcopenic obesity: definition, cause and consequences. *Curr. Opin. Clin. Nutr. Metab. Care* **11,** 693–700 (2008).

206. Oh, D. Y., Morinaga, H., Talukdar, S., Bae, E. J. & Olefsky, J. M. Increased macrophage migration into adipose tissue in obese mice. *Diabetes* **61,** 346–354 (2012).

207. Suganami, T., Nishida, J. & Ogawa, Y. A paracrine loop between adipocytes and macrophages aggravates inflammatory changes: role of free fatty acids and tumor necrosis factor alpha. *Arterioscler. Thromb. Vasc. Biol.* **25,** 2062–2068 (2005).

208. Wang, Y.-Q. *et al.* Peripheral monocyte count: an independent diagnostic and prognostic biomarker for prostate cancer - a large Chinese cohort study. *Asian J. Androl.* **19,** 579–585 (2017).

209. Nonomura, N. *et al.* Infiltration of tumour-associated macrophages in prostate biopsy specimens is predictive of disease progression after hormonal therapy for prostate cancer. *BJU Int.* **107,** 1918–1922 (2011).

210. Montgomery, B. *et al.* Neoadjuvant Enzalutamide Prior to Prostatectomy. *Clin. Cancer Res.* **23,** 2169–2176 (2017).

211. Tombal, B. *et al.* Enzalutamide monotherapy in hormone-naive prostate cancer: primary analysis of an open-label, single-arm, phase 2 study. *Lancet Oncol.* **15,** 592–600 (2014).

212. Palmberg, C., Koivisto, P., Visakorpi, T. & Tammela, T. L. PSA decline is an independent prognostic marker in hormonally treated prostate cancer. *Eur. Urol.* **36,** 191–196 (1999).

213. Gittes, R. F. Carcinoma of the prostate. *N. Engl. J. Med.* **324,** 236–245 (1991).

214. Crawford, E. D. *et al.* A controlled trial of leuprolide with and without flutamide in prostatic carcinoma. *N. Engl. J. Med.* **321,** 419–424 (1989).

215. Denis, L. J. *et al.* Goserelin acetate and flutamide versus bilateral orchiectomy: a phase III EORTC trial (30853). EORTC GU Group and EORTC Data Center. *Urology* **42,** 119–29; discussion 129–30 (1993).

216. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune

landscape in human cancer. *Immunity* **39,** 782–795 (2013).

217. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–196 (2016).

218. Pearce, O. M. T. *et al.* Deconstruction of a Metastatic Tumor Microenvironment Reveals a Common Matrix Response in Human Cancers. *Cancer Discov.* **8,** 304–319 (2018).

219. Ostman, A. The tumor microenvironment controls drug sensitivity. *Nat. Med.* **18,** 1332–1334 (2012).

220. Mlecnik, B. *et al.* The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis. *Sci. Transl. Med.* **8,** 327ra26 (2016).

221. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28,** 2184–2185 (2012).

222. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).

223. Cooperberg, M. R., Hilton, J. F. & Carroll, P. R. The CAPRA-S score: a straightforward tool for improved prediction of outcomes after radical prostatectomy. *Cancer* (2011).

224. Richards, F. J. A Flexible Growth Function for Empirical Use. *J. Exp. Bot.* **10,** 290–301 (1959).

225. Piironen, J. & Vehtari, A. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11,** 5018–5051 (2017).

226. Gelman, A. *et al. Bayesian Data Analysis, Third Edition*. (CRC Press, 2013).

227. Gabry, J. Graphical posterior predictive checks using the bayesplot package. (2018). Available at: https://cran.r-project.org/web/packages/bayesplot/vignettes/graphical-ppcs.html. (Accessed: 13th July 2018)

228. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–3440 (2005).

229. Kumar, V. *et al.* Cancer-Associated Fibroblasts Neutralize the Anti-tumor Effect of CSF1 Receptor Blockade by Inducing PMN-MDSC Infiltration of Tumors. *Cancer Cell* **32,** 654–668.e5 (2017).

230. Liu, M. *et al.* CXCL10/IP-10 in infectious diseases pathogenesis and potential therapeutic

implications. *Cytokine Growth Factor Rev.* **22,** 121–130 (2011).

231. Lu, J., Chatterjee, M., Schmid, H., Beck, S. & Gawaz, M. CXCL14 as an emerging immune and inflammatory modulator. *J. Inflamm.* **13,** 1 (2016).
232. Latour, S. *et al.* Regulation of SLAM-mediated signal transduction by SAP, the X-linked lymphoproliferative gene product. *Nat. Immunol.* **2,** 681–690 (2001).
233. Wang, G. *et al.* Migration of myeloid cells during inflammation is differentially regulated by the cell surface receptors Slamf1 and Slamf8. *PLoS One* **10,** e0121968 (2015).
234. Consequences of the crosstalk between monocytes/macrophages and natural killer cells. Front. (2012).
235. Rovis, T. L., Brlic, P. K., Kaynan, N. & Lisnic, V. J. Inflammatory monocytes and NK cells play a crucial role in DNAM-1–dependent control of cytomegalovirus infection. *Journal of* (2016).
236. O'Connell, P. A., Surette, A. P., Liwski, R. S., Svenningsson, P. & Waisman, D. M. S100A10 regulates plasminogen-dependent macrophage invasion. *Blood* **116,** 1136–1146 (2010).
237. Wu, X. *et al.* Angiopoietin-2 as a Biomarker and Target for Immune Checkpoint Therapy. *Cancer Immunol Res* **5,** 17–28 (2017).
238. Vijayan, V. *et al.* A New Immunomodulatory Role for Peroxisomes in Macrophages Activated by the TLR4 Ligand Lipopolysaccharide. *J. Immunol.* **198,** 2414–2425 (2017).
239. Cohen, N. *et al.* Fibroblasts drive an immunosuppressive and growth-promoting microenvironment in breast cancer via secretion of Chitinase 3-like 1. *Oncogene* **36,** 4457–4468 (2017).
240. Jounaidi, Y., Cotten, J. F., Miller, K. W. & Forman, S. A. Tethering IL2 to Its Receptor IL2Rβ Enhances Antitumor Activity and Expansion of Natural Killer NK92 Cells. *Cancer Res.* **77,** 5938–5951 (2017).
241. ImmuNet. Available at: http://immunet.princeton.edu/genes/detail/homo-sapien/IL2RB/. (Accessed: 26th August 2018)
242. Espinoza-Delgado, I. *et al.* Interleukin-2 and human monocyte activation. *J. Leukoc. Biol.* **57,** 13–19 (1995).

243. Kim, M. *et al.* Novel natural killer cell-mediated cancer immunotherapeutic activity of anisomycin against hepatocellular carcinoma cells. *Sci. Rep.* **8,** 10668 (2018).

244. Ihanus, E., Uotila, L. M., Toivanen, A., Varis, M. & Gahmberg, C. G. Red-cell ICAM-4 is a ligand for the monocyte/macrophage integrin CD11c/CD18: characterization of the binding sites on ICAM-4. *Blood* **109,** 802–810 (2007).

245. Järvinen, T. A. H. & Ruoslahti, E. Target-seeking antifibrotic compound enhances wound healing and suppresses scar formation in mice. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21671–21676 (2010).

246. Zhang, W. *et al.* Decorin is a pivotal effector in the extracellular matrix and tumour microenvironment. *Oncotarget* **9,** 5480–5491 (2018).

247. Narita, H., Chen, S., Komori, K. & Kadomatsu, K. Midkine is expressed by infiltrating macrophages in in-stent restenosis in hypercholesterolemic rabbits. *J. Vasc. Surg.* **47,** 1322–1329 (2008).

248. Fan, N. *et al.* Midkine, a potential link between obesity and insulin resistance. *PLoS One* **9,** e88299 (2014).

249. Yang, P. *et al.* Role of PDGF-D and PDGFR-β in neuroinflammation in experimental ICH mice model. *Exp. Neurol.* **283,** 157–164 (2016).

250. GeneCards Human Gene Database. HLA-DRB5 Gene - GeneCards | DRB5 Protein | DRB5 Antibody. Available at: https://www.genecards.org/cgi-bin/carddisp.pl?gene=HLA-DRB5. (Accessed: 26th August 2018)

251. Genetics Home Reference. HLA-DRB5 gene. *Genetics Home Reference* Available at: https://ghr.nlm.nih.gov/gene/HLA-DRB5. (Accessed: 26th August 2018)

252. Kolijn, K. *et al.* Epithelial-Mesenchymal Transition in Human Prostate Cancer Demonstrates Enhanced Immune Evasion Marked by IDO1 Expression. *Cancer Res.* **78,** 4671–4679 (2018).

253. Kai, K. *et al.* CSF-1/CSF-1R axis is associated with epithelial/mesenchymal hybrid phenotype in epithelial-like inflammatory breast cancer. *Sci. Rep.* **8,** 9427 (2018).

254. Moen, I. *et al.* Hyperoxic treatment induces mesenchymal-to-epithelial transition in a rat adenocarcinoma model. *PLoS One* **4,** e6381 (2009).

255. Pivetta, E., Colombatti, A. & Spessotto, P. A Rare Bird among Major Extracellular Matrix Proteins: EMILIN1 and the Tumor Suppressor Function. *Journal of Carcinogenesis & Mutagenesis* **0,** 1–11 (2013).

256. Tsai, C.-K. *et al.* Overexpression of PLOD3 promotes tumor progression and poor prognosis in gliomas. *Oncotarget* **9,** 15705–15720 (2018).

257. Higgins, P. J. Expression of the p53 target SERPINE1 (PAI-1) gene is required for human tumor cell migration upon plastic conversion to a stem cell-like phenotype in response to TGF- β1+EGF. (2016). doi:10.4172/1948-5956.S1.02

258. Sonnylal, S. *et al.* Connective tissue growth factor causes EMT-like cell fate changes in vivo and in vitro. *J. Cell Sci.* **126,** 2164–2175 (2013).

259. Tauber, S. *et al.* Transcriptome analysis of human cancer reveals a functional role of heme oxygenase-1 in tumor cell adhesion. *Mol. Cancer* **9,** 200 (2010).

260. Péterfi, Z. *et al.* Peroxidasin is secreted and incorporated into the extracellular matrix of myofibroblasts and fibrotic kidney. *Am. J. Pathol.* **175,** 725–735 (2009).

261. Sitole, B. N. & Mavri-Damelin, D. Peroxidasin is regulated by the epithelial-mesenchymal transition master transcription factor Snai1. *Gene* **646,** 195–202 (2018).

262. Terraube, V., Marx, I. & Denis, C. V. Role of von Willebrand factor in tumor metastasis. *Thromb. Res.* **120 Suppl 2,** S64–70 (2007).

263. Joshi, N. *et al.* Von Willebrand factor deficiency reduces liver fibrosis in mice. *Toxicol. Appl. Pharmacol.* **328,** 54–59 (2017).

264. van Nieuwenhoven, F. A. *et al.* Cartilage intermediate layer protein 1 (CILP1): A novel mediator of cardiac extracellular matrix remodelling. *Sci. Rep.* **7,** 16042 (2017).

265. Zhang, C.-L. *et al.* Cartilage intermediate layer protein-1 alleviates pressure overload-induced cardiac fibrosis via interfering TGF-β1 signaling. *J. Mol. Cell. Cardiol.* **116,** 135–144 (2018).

266. Xie, L. *et al.* Cystatin C increases in cardiac injury: a role in extracellular matrix protein modulation. *Cardiovasc. Res.* **87,** 628–635 (2010).

267. Hitomi, K., Yamagiwa, Y., Ikura, K., Yamanishi, K. & Maki, M. Characterization of human recombinant transglutaminase 1 purified from baculovirus-infected insect cells. *Biosci. Biotechnol. Biochem.* **64,** 2128–2137 (2000).

268. Königshoff, M. *et al.* Increased expression of 5-hydroxytryptamine2A/B receptors in idiopathic pulmonary fibrosis: a rationale for therapeutic intervention. *Thorax* **65,** 949–955 (2010).

269. Jara, P. *et al.* Matrix metalloproteinase (MMP)-19-deficient fibroblasts display a profibrotic phenotype. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **308,** L511–22 (2015).

270. Bonnans, C., Chou, J. & Werb, Z. Remodelling the extracellular matrix in development and disease. *Nat. Rev. Mol. Cell Biol.* **15,** 786–801 (2014).

271. de Arao Tan, I., Ricciardelli, C. & Russell, D. L. The metalloproteinase ADAMTS1: A comprehensive review of its role in tumorigenic and metastatic pathways. *Int. J. Cancer* **133,** 2263–2276 (2013).

272. Kim, Y.-K. *Understanding Depression: Volume 1. Biomedical and Neurobiological Background*. (Springer, 2018).

273. Singhania, A. *et al.* Altered Epithelial Gene Expression in Peripheral Airways of Severe Asthma. *PLoS One* **12,** e0168680 (2017).

274. de Souza Junior, D. A., Santana, A. C., da Silva, E. Z. M., Oliver, C. & Jamur, M. C. The Role of Mast Cell Specific Chymases and Tryptases in Tumor Angiogenesis. *Biomed Res. Int.* **2015,** 142359 (2015).

275. Jayakumar, A. *et al.* Consequences of C-terminal domains and N-terminal signal peptide deletions on LEKTI secretion, stability, and subcellular distribution. *Arch. Biochem. Biophys.* **435,** 89–102 (2005).

276. Kouzaki, H. *et al.* Endogenous Protease Inhibitors in Airway Epithelial Cells Contribute to Eosinophilic Chronic Rhinosinusitis. *Am. J. Respir. Crit. Care Med.* **195,** 737–747 (2017).

277. Azouz, N. P. *et al.* The antiprotease SPINK7 serves as an inhibitory checkpoint for esophageal epithelial inflammatory responses. *Sci. Transl. Med.* **10,** (2018).

278. El Khoury, L. *et al.* Polymorphic variation within the ADAMTS2, ADAMTS14, ADAMTS5, ADAM12 and TIMP2 genes and the risk of Achilles tendon pathology: a genetic association study. *J. Sci. Med. Sport* **16,** 493–498 (2013).

279. Saneyasu, T., Akhtar, R. & Sakai, T. Molecular Cues Guiding Matrix Stiffness in Liver Fibrosis. *Biomed Res. Int.* **2016,** 2646212 (2016).

280. Tossell, K. *et al.* Lrrn1 is required for formation of the midbrain-hindbrain boundary and organiser through regulation of affinity differences between midbrain and hindbrain cells in chick. *Dev. Biol.* **352,** 341–352 (2011).

281. Warnecke, A. *et al.* Stable release of BDNF from the fibroblast cell line NIH3T3 grown on silicone elastomers enhances survival of spiral ganglion cells in vitro and in vivo. *Hear. Res.* **289,** 86–97 (2012).

282. Dudás, J. *et al.* Fibroblasts produce brain-derived neurotrophic factor and induce mesenchymal transition of oral tumor cells. *Oral Oncol.* **47,** 98–103 (2011).

283. Emanuele, N. *et al.* Effect of Recombinant Lubricin on Human Blood Coagulation Parameters and Platelet Aggregation. *The FASEB Journal* (2016).

284. Uutela, M. *et al.* PDGF-D induces macrophage recruitment, increased interstitial pressure, and blood vessel maturation during angiogenesis. *Blood* **104,** 3198–3204 (2004).

285. Henneberry, A. L., Wistow, G. & McMaster, C. R. Cloning, genomic organization, and characterization of a human cholinephosphotransferase. *J. Biol. Chem.* **275,** 29808–29815 (2000).

286. Starke, R. D. *et al.* Endothelial von Willebrand factor regulates angiogenesis. *Blood* **117,** 1071–1080 (2011).

287. Schmid, M. C. & Varner, J. A. Myeloid cell trafficking and tumor angiogenesis. *Cancer Lett.* **250,** 1–8 (2007).

288. Randi, A. M., Laffan, M. A. & Starke, R. D. Von Willebrand factor, angiodysplasia and angiogenesis. *Mediterr. J. Hematol. Infect. Dis.* **5,** e2013060 (2013).

289. Tsai, H.-M. ADAMTS13 and microvascular thrombosis. *Expert Rev. Cardiovasc. Ther.* **4,** 813–825

(2006).

290. Xu, H. *et al.* ADAMTS13 controls vascular remodeling by modifying VWF reactivity during stroke recovery. *Blood* **130,** 11–22 (2017).

291. Muia, J. *et al.* Allosteric activation of ADAMTS13 by von Willebrand factor. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 18584–18589 (2014).

292. Libreros, S., Garcia-Areas, R. & Iragavarapu-Charyulu, V. CHI3L1 plays a role in cancer through enhanced production of pro-inflammatory/pro-tumorigenic and angiogenic factors. *Immunol. Res.* **57,** 99–105 (2013).

293. Lundequist, A., Tchougounova, E., Abrink, M. & Pejler, G. Cooperation between mast cell carboxypeptidase A and the chymase mouse mast cell protease 4 in the formation and degradation of angiotensin II. *J. Biol. Chem.* **279,** 32339–32344 (2004).

294. de Souza Junior, D. A., Borges, A. C., Santana, A. C., Oliver, C. & Jamur, M. C. Mast Cell Proteases 6 and 7 Stimulate Angiogenesis by Inducing Endothelial Cells to Release Angiogenic Factors. *PLoS One* **10,** e0144081 (2015).

295. Roy, R., Dagher, A., Butterfield, C. & Moses, M. A. ADAM12 Is a Novel Regulator of Tumor Angiogenesis via STAT3 Signaling. *Mol. Cancer Res.* **15,** 1608–1622 (2017).

296. Lewis, C. E., Harney, A. S. & Pollard, J. W. The Multifaceted Role of Perivascular Macrophages in Tumors. *Cancer Cell* **30,** 18–25 (2016).

297. Wang, Y. *et al.* Regulation of Cholesterologenesis by the Oxysterol Receptor, LXRα. *J. Biol. Chem.* **283,** 26332–26339 (2008).

298. McDonough, C. W. *et al.* Atenolol induced HDL-C change in the pharmacogenomic evaluation of antihypertensive responses (PEAR) study. *PLoS One* **8,** e76984 (2013).

299. STARD3 StAR related lipid transfer domain containing 3 [Homo sapiens (human)] - Gene - NCBI. Available at: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=10948. (Accessed: 27th August 2018)

300. Krakowiak, P. A. *et al.* Lathosterolosis: an inborn error of human and murine cholesterol synthesis

due to lathosterol 5-desaturase deficiency. *Hum. Mol. Genet.* **12,** 1631–1641 (2003).

301. Wei, E. *et al.* Loss of TGH/Ces3 in mice decreases blood lipids, improves glucose tolerance, and increases energy expenditure. *Cell Metab.* **11,** 183–193 (2010).
302. Ross, M. K., Streit, T. M. & Herring, K. L. Carboxylesterases: Dual roles in lipid and pesticide metabolism. *J. Pestic. Sci.* **35,** 257–264 (2010).
303. Vigne, S. *et al.* IL-27-Induced Type 1 Regulatory T-Cells Produce Oxysterols that Constrain IL-10 Production. *Front. Immunol.* **8,** 1184 (2017).
304. Racioppi, L. CaMKK2: a novel target for shaping the androgen-regulated tumor ecosystem. *Trends Mol. Med.* **19,** 83–88 (2013).
305. Asuthkar, S., Velpula, K. K., Elustondo, P. A., Demirkhanyan, L. & Zakharian, E. TRPM8 channel as a novel molecular target in androgen-regulated prostate cancer cells. *Oncotarget* **6,** 17221–17236 (2015).
306. Aupperlee, M. D. *et al.* Epidermal growth factor receptor (EGFR) signaling is a key mediator of hormone-induced leukocyte infiltration in the pubertal female mammary gland. *Endocrinology* **155,** 2301–2313 (2014).
307. Kariagina, A., Xie, J., Leipprandt, J. R. & Haslam, S. Z. Amphiregulin mediates estrogen, progesterone, and EGFR signaling in the normal rat mammary gland and in hormone-dependent rat mammary cancers. *Horm. Cancer* **1,** 229–244 (2010).
308. Lodoen, M. B. & Lanier, L. L. Viral modulation of NK cell immunity. *Nat. Rev. Microbiol.* **3,** 59–69 (2005).
309. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Series B Stat. Methodol.* **44,** 139–177 (1982).
310. Ferrari, S. & Cribari-Neto, F. Beta Regression for Modelling Rates and Proportions. *J. Appl. Stat.* **31,** 799–815 (2004).
311. Cribari-Neto, F. & Zeileis, A. Beta Regression in R. 22 (2009).
312. Zhang, P., Qiu, Z. & Shi, C. simplexreg: An R Package for Regression Analysis of Proportional Data

Using the Simplex Distribution. *Journal of Statistical Software, Articles* **71,** 1–21 (2016).

313. Ng, A. Y. & Jordan, M. I. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. in *Advances in Neural Information Processing Systems 14* (eds. Dietterich, T. G., Becker, S. & Ghahramani, Z.) 841–848 (MIT Press, 2002).
314. Eliason, S. R. *Maximum Likelihood Estimation: Logic and Practice*. (SAGE, 1993).
315. Stasinopoulos, D. M. & Rigby, R. A. Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* (2007).
316. Rigby, R. A. & Mikis Stasinopoulos, D. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Stat. Modelling* **6,** 209–229 (2006).
317. Stasinopoulos, D. M., Rigby, R. A. & Akantziliotou, C. Instructions on how to use the GAMLSS package in R. *Accompanying documentation in the current GAMLSS help files,(see also http://www. gamlss. org/)* (2006).
318. Connor, R. J. & Mosimann, J. E. Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *J. Am. Stat. Assoc.* **64,** 194–206 (1969).
319. Barndorff-Nielsen, O. E. & Jørgensen, B. Some parametric models on the simplex. *J. Multivar. Anal.* **39,** 106–116 (1991).
320. Jorgensen, B. *The Theory of Dispersion Models*. (CRC Press, 1997).
321. Gabry, J. & Goodrich, B. rstanarm: Bayesian applied regression modeling via Stan. *R package version* **2,** (2016).
322. Meraviglia, S. *et al.* Distinctive features of tumor-infiltrating γδ T lymphocytes in human colorectal cancer. *Oncoimmunology* **6,** e1347742 (2017).
323. Trella, E. *et al.* The interplay between neutrophils and CD8+ T cells improves survival in human colorectal cancer. *Clin. Cancer Res.* clincanres–2047 (2017).
324. Bense, R. D. *et al.* Relevance of Tumor-Infiltrating Immune Cell Composition and Functionality for Disease Outcome in Breast Cancer. *J. Natl. Cancer Inst.* **109,** (2017).
325. Araujo, J. *et al.* CCL5 expression and tumor infiltrating immune cells in triple negative breast

cancer. *J. Clin. Orthod.* **35,** 11553–11553 (2017).

326. Hawinkels, L. J. A. C. *et al.* Interaction with colon cancer cells hyperactivates TGF-β signaling in cancer-associated fibroblasts. *Oncogene* **33,** 97–107 (2014).

327. Tommelein, J. *et al.* Cancer-associated fibroblasts connect metastasis-promoting communication in colorectal cancer. *Front. Oncol.* **5,** 63 (2015).

328. Busch, S. *et al.* TGF-beta receptor type-2 expression in cancer-associated fibroblasts regulates breast cancer cell growth and survival and is a prognostic marker in pre-menopausal breast cancer. *Oncogene* **34,** 27–38 (2015).

329. Slebe, F. *et al.* FoxA and LIPG endothelial lipase control the uptake of extracellular lipids for breast cancer growth. *Nat. Commun.* **7,** 11199 (2016).

330. Wang, C.-A., Harrell, J. C., Iwanaga, R., Jedlicka, P. & Ford, H. L. Vascular endothelial growth factor C promotes breast cancer progression via a novel antioxidant mechanism that involves regulation of superoxide dismutase 3. *Breast Cancer Res.* **16,** 462 (2014).

331. Nagarsheth, N., Wicha, M. S. & Zou, W. Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy. *Nat. Rev. Immunol.* **17,** 559–572 (2017).

332. Zhao, Y. & Adjei, A. A. Targeting Angiogenesis in Cancer Therapy: Moving Beyond Vascular Endothelial Growth Factor. *Oncologist* **20,** 660–673 (2015).

333. Bovy, N. *et al.* Endothelial exosomes contribute to the antitumor response during breast cancer neoadjuvant chemotherapy via microRNA transfer. *Oncotarget* **6,** 10253–10266 (2015).

334. Mantovani, A., Marchesi, F., Malesci, A., Laghi, L. & Allavena, P. Tumour-associated macrophages as treatment targets in oncology. *Nat. Rev. Clin. Oncol.* (2017). doi:10.1038/nrclinonc.2016.217

335. Dougan, M. & Dranoff, G. Immune therapy for cancer. *Annu. Rev. Immunol.* **27,** 83–117 (2009).

336. Becker, R. A., Chambers, J. M. & Wilks, A. R. The new S language. *Pacific Grove, Ca.: Wadsworth & Brooks, 1988* (1988).

337. Bennett, K. P. & Campbell, C. Support Vector Machines: Hype or Hallelujah? *SIGKDD Explor. Newsl.* **2,** 1–13 (2000).

338. Chen, L. *et al.* CAM-CM: a signal deconvolution tool for in vivo dynamic contrast-enhanced imaging of complex tissues. *Bioinformatics* **27,** 2607–2609 (2011).

339. Cressie, N., Calder, C. A., Clark, J. S., Ver Hoef, J. M. & Wikle, C. K. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecol. Appl.* **19,** 553–570 (2009).

340. de Valpine, P. Shared challenges and common ground for Bayesian and classical analysis of hierarchical statistical models. *Ecol. Appl.* **19,** 584–588 (2009).

341. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157,** 714–725 (2014).

342. Bolstad, B. Probe level quantile normalization of high density oligonucleotide array data. *Unpublished manuscript* (2001).

343. Boldstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* **19,** 185–193 (2003).

344. Inference of extrinsic rates of change from simplex space using a Dirichlet regression model. *Google Docs* Available at: https://docs.google.com/document/d/1VpuljFJOs4VR-__ZvJZlw0X6cTOCVmDPOTr6D0okr-w/edit. (Accessed: 23rd February 2018)

345. Team, R. C. & Others. R: A language and environment for statistical computing. (2013).

346. Hoffman, M. D. & Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv [stat.CO]* (2011).

347. Neal, R. M. & Others. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* **2,** (2011).

348. GDC. Available at: http://portal.gdc.cancer.gov. (Accessed: 20th July 2018)

349. cBioPortal for Cancer Genomics. Available at: http://cbioportal.org/public-portal. (Accessed: 20th July 2018)

350. Cox, D. R. Regression models and life-tables (with discussions). *Jr Stat Soc B* **34,** 187–220 (1972).

351. Lanciotti, M. *et al.* The role of M1 and M2 macrophages in prostate cancer in relation to extracapsular tumor extension and biochemical recurrence after radical prostatectomy. *Biomed Res. Int.* **2014,** 486798 (2014).

352. Rydström, A. *et al.* Dynamics of myeloid cell populations during relapse-preventive immunotherapy in acute myeloid leukemia. *J. Leukoc. Biol.* **102,** 467–474 (2017).

353. Pulsoni, A. *et al.* M4 acute myeloid leukemia: the role of eosinophilia and cytogenetics in treatment response and survival. The GIMEMA experience. *Haematologica* **93,** 1025–1032 (2008).

354. Lafont, V. *et al.* Plasticity of γδ T Cells: Impact on the Anti-Tumor Response. *Front. Immunol.* **5,** 622 (2014).

355. Aswald, J. M. *et al.* Flow cytometric assessment of autologous gammadelta T cells in patients with acute myeloid leukemia: potential effector cells for immunotherapy? *Cytometry B Clin. Cytom.* **70,** 379–390 (2006).

356. Todaro, M. *et al.* Efficient killing of human colon cancer stem cells by gammadelta T lymphocytes. *J. Immunol.* **182,** 7287–7296 (2009).

357. Pan, X.-Q. The mechanism of the anticancer function of M1 macrophages and their use in the clinic. *Chin. J. Cancer* **31,** 557–563 (2012).

358. Erreni, M., Mantovani, A. & Allavena, P. Tumor-associated Macrophages (TAM) and Inflammation in Colorectal Cancer. *Cancer Microenviron.* **4,** 141–154 (2011).

359. Gül, N. *et al.* Macrophages mediate colon carcinoma cell adhesion in the rat liver after exposure to lipopolysaccharide. *Oncoimmunology* **1,** 1517–1526 (2012).

360. Gabrielson, A. *et al.* Intratumoral CD3 and CD8 T-cell Densities Associated with Relapse-Free Survival in HCC. *Cancer Immunol Res* **4,** 419–430 (2016).

361. Zhang, Y.-L. *et al.* SPON2 Promotes M1-like Macrophage Recruitment and Inhibits Hepatocellular Carcinoma Metastasis by Distinct Integrin-Rho GTPase-Hippo Pathways. *Cancer Res.* **78,** 2305–2317 (2018).

362. Soo, R. A. *et al.* Prognostic significance of immune cells in non-small cell lung cancer: meta-

analysis. *Oncotarget* **9,** 24801–24820 (2018).

363. Remark, R. *et al.* The non-small cell lung cancer immune contexture. A major determinant of tumor characteristics and patient outcome. *Am. J. Respir. Crit. Care Med.* **191,** 377–390 (2015).

364. Jackute, J. *et al.* Distribution of M1 and M2 macrophages in tumor islets and stroma in relation to prognosis of non-small cell lung cancer. *BMC Immunol.* **19,** 3 (2018).

365. Hedbrant, A., Wijkander, J., Seidal, T., Delbro, D. & Erlandsson, A. Macrophages of M1 phenotype have properties that influence lung cancer cell progression. *Tumour Biol.* **36,** 8715–8725 (2015).

366. Karakhanova, S. *et al.* Prognostic and predictive value of immunological parameters for chemoradioimmunotherapy in patients with pancreatic adenocarcinoma. *Br. J. Cancer* **112,** 1027–1036 (2015).

367. Daley, D. & Miller, G. The role of γδ T cells in pancreatic cancer: what could this mean for the clinic? *Expert Rev. Gastroenterol. Hepatol.* **11,** 609–610 (2017).

368. Guo, X. *et al.* Mast Cell Tryptase Contributes to Pancreatic Cancer Growth through Promoting Angiogenesis via Activation of Angiopoietin-1. *Int. J. Mol. Sci.* **17,** (2016).

369. Yuan, X. *et al.* Prognostic significance of tumor-associated macrophages in ovarian cancer: A meta-analysis. *Gynecol. Oncol.* **147,** 181–187 (2017).

370. Wang, X., Zhao, X., Wang, K., Wu, L. & Duan, T. Interaction of monocytes/macrophages with ovarian cancer cells promotes angiogenesis in vitro. *Cancer Sci.* **104,** 516–523 (2013).

371. Kang, B. W., Kim, J. G., Lee, I. H., Bae, H. I. & Seo, A. N. Clinical significance of tumor-infiltrating lymphocytes for gastric cancer in the era of immunology. *World J. Gastrointest. Oncol.* **9,** 293–299 (2017).

372. Liu, H. *et al.* Decreased expression of granulocyte-macrophage colony-stimulating factor is associated with adverse clinical outcome in patients with gastric cancer undergoing gastrectomy. *Oncol. Lett.* **14,** 4701–4707 (2017).