THE UNIVERSITY of EDINBURGH
School of Biological Sciences

Disentangling patterns of gene expression in high grade serous ovarian cancer by matrix factorization

Ian Poole

Student Examination Number: **B156476**

In partial fulfilment of the requirement for the Degree of Master of Science in Systems and Synthetic Biology at the University of Edinburgh, 2019 / 2020

Disssertation Supervisor: Dr. Ailith Ewing

Contents

Abstract

Gene expression analysis – transcriptomics – of a set biopsied tumour tissues can identify genes whose expression is linked to the pathological processes of the tumour. In contrast to supervised contrastive approaches unsupervised techniques work with datasets of mixed and unknown tumour status. Matrix factorization (MF) methods have become a popular means of reducing and "disentangling" the salient components of variation in such datasets, known as metagenes. From these metagenes the most influential genes can be identified, and when subjected to gene enrichment analysis against the Gene Ontology (GO) can provide insights into the underlying biological processes in play.

In this work the biological target is high-grade serous ovarian cancer (HGSOC) – the most common and aggressive type of ovarian cancer. Two gene expression datasets are analysed; the Australian Ovarian Cancer Study dataset (N=80) has patient survival and genomic metadata, while the larger dataset from The Cancer Genome Atlas (N=374) has only survival metadata. This motivates an approach of unsupervised metagenes discovered by MF techniques in the larger dataset then transferring that information to the smaller dataset to test for their power to stratify patient survival and for correlation with genomic features.

Three MF methods are applied and compared: non-negative matrix factorization (NMF), independent component analysis (ICA) and principal component analysis (PCA). While these are all linear factorization methods, they have very different properties. A key problem is deciding on the number of components to extract – the factorization rank. Too high a rank will result in metagenes which are unstable and so fail to generalise. The role of sampling error in the instability seems under appreciated in the literature, which this work addresses by bootstrap resampling then assessing stability through k-means clustering and the Silhouette score. Similarity of metagenes found by the three methods is assessed via the Jaccard index.

An integrative analysis of the GO enrichment results, survival analysis, genomic feature correlations and metagene similarity allows some tentative biological findings to be suggested and corroboration sought in the literature. The main methodological findings are that MF is effective in identifying and removing technical batch effects, but to beware of instability due to sampling error when applying MF to small datasets. The question of which of the three methods – NMF, ICA or PCA – is most effective for these purposes is considered but no clear conclusion can be drawn.

1 Introduction

1.1 Epithelial Ovarian Cancer

Ovarian cancer (OC) is a heterogeneous disease, presenting with a wide range of pathologies and molecular characteristics . The World Health Organisation sets out five subtypes of epithelial ovarian carcinomas (EOC) grouped into two main types[1, 2]:

Type I are low grade with generally good prognosis, with two subtypes:

Low grade serous affecting the epithelial membrane which secretes serous fluid), < 5% of cases

Mucosal affecting membrane rich mucous glands, 2 - 3%

Clear cell relating to the presence of clear cells in histology, 5 - 10%.

Type II are more aggressive (rapidly growing), typically carrying P53 mutations with defects in mechanisms of DNA repair. Three subtypes:

Endometrial affecting the the inner lining of the uterus, 10 % of cases.

High grade serous OC (HGSOC) affects the epithelium of the ovaries and fallopian tubes. It is the most common type of OC, accounting for 70% of cases, and most serious of the OC subtypes. HGSOC is the focus of this dissertation.

HGSOC typically occurs in older patients and is diagnosed at a late stage. Histology of HGSOC is similar to the low grade subtype, but the cancer develops along distinct molecular pathways. Treatment options depend on tumour stage and include surgery (tumour debulking) or platinum-based chemotherapy (e.g. Cisplatin). Acquired resistance to platinum is a major cause of disease recurrance[3] and so motivates research into the genomic events which lead to such resistance.

1.2 Patterns of gene expression in HGSOC

Identifying the patterns – or "signatures" – of gene expression has been an active research area, for the purposes of understanding the carcinogenesis pathway, subtype classification, predicting survival prognosis, predicting treatment response and understanding the causative mechanism of risk factors; there is much cross-over between these areas, however.

1.2.1 Carcinogenesis pathways including homologous repair deficiency (HRD)

BRCA1 inactivation is known to cause chromosomal instability in many cancers. Pradhan *et al* [4] investigated the role of BRCA1 in HGSOC by copy number and expression analysis, finding surprisingly that inactivation has no relationship with gross genomic alteration. They leave open the question whether DNA repair by PARP plays a role. The relationship between BRCA1/2 and DNA repair – specifically homologous repair deficiency – is picked up by Ewing *et al* [5]. They study the complex interplay of single nucleotide variants and large structural variants as they impact BRCA1/2 disruption and thence HRD. This work suggests that when BRCA1/2 loss is detected in HGSOC patients, there may be a clinical role for PARP inhibitors, since this would prevent error-prone non-homologous repair by PARP, and so selectively kill BRCA1/2 disrupted cells.

A gene expression signature of HRD has been reported [6] and shown to be predictive of overall and progression free survival in OC patients. The signature was derived by a supervised approach, from tumour samples known to be with or without HRD. Examining the paper's supplementary material, the signature is based on 116 genes, with +1 /-1 weights. Notable positively weighted genes include ABCF1, ACTL6B and ARID5A; negatively weighted include ZNF780A, WDR36 and TRPM7.

Studying the combined landscape of chromosomal rearrangements, driver mutations (particularly TP53, BRCA1/2), methylation and gene expression profiles has lead to a better understanding of the molecular events involved in the progression HGSOC and the development of chemoresistance [3]. The sequence of events is complex, however, and not easily summarised here. A similar integrative / multiplatform approach is

taken in [7] and [8].

1.2.2 Subtype classification

A relatively early (2003) expression microarray based work demonstrated clustering into lymphocyte related genes (IGH3, IGKC, IGLJ3), extracellular matrix/stromal related (COL11A1, COL3A1, MMP2, SPARC, RBBP1) and six other clusters [9].

Wang *et al* [10] use NMF clustering (see section 1.3.1) to identify five subtypes of HGSOC informative of outcomes: 1:mesenchymal, 2:immunoreactive, 3:proliferative, 4:differentiated and 5:anti-mesenchymal. Subtypes 2 and 5 are found to be associated with better survival.

Disruption of the PI3K-AKT signalling pathway is known to be significant in several cancers, due to its role in regulating apoptosis. Espinosa *et al* [11] used unsupervised clustering techniques to investigate expression of 22 genes of this pathway. They found that HGSOC cases formed two separate clusters; the cluster with high expression of CASP3, XIAP, NFKB1, FAS and GSK3B was linked with better outcomes.

An integrative network approach has identified two distinct subtypes of ovarian cancer, distinguished by differences in expression of genes related to transcription factors which influence angiogenesis [12]. The authors suggest that the regulatory networks highlighted by their analysis can lead to targetting with specific drugs.

A recent pre-publication develops a pipeline for sub-type prediction with the potential to be used clinically [13]. The NanoString gene expression platform is used. Their strategy was to focus on small set of 513 genes (vs $\sim$21,000 protein coding genes) known from the literature to have relevance to subtyping. Two approaches were followed by different teams. “All Array”: used expression array data from 1650 patients across 14 studies evaluating 9 supervised learning algorithms by bootstrap, selecting an AdaBoost -like method. The “TCGA” team used 434 patients from The Cancer Genome Atlas (TCGA[1]) evaluating 5 algorithms by cross-validation, selecting a random forest. “All Array” had the advantage of more data but needed careful attention

[1] I’m always amused by the genetics pun in the naming of this resource!

to batch effects. A final classifier took a consensus of the two methods based on a minimal gene set (around 40), validated by a leave-one-out (patient level) approach. Survival analysis was carried out stratified by predicted subtype.

1.2.3 Treatment response and survival

As we have seen in section 1.1, EOC subtypes are closely linked with survival, nevertheless much research is directed at finding direct links with expression.

The prognostic value of expression signatures is demonstrated by the "Classification of Ovarian Cancer" (CLOVAR) system [14]. Single-sample gene set enrichment analysis (ssGSEA) is used to obtain scores for previously reported expression signatures: differentiated, immunoreative, mesenchymal and proliferative. An outcome prediction model based on these signatures gave good survival stratification,, with likelihood ratio 10.8 between above/below median groups on a validation dataset which improved to 15.7 when augmented with BRCA1/2 mutation status and other post-operative clinical factors.

An explicit supervised learning approach can be taken, as in Berchuck *et al*, where a hybrid decision tree plus linear discriminant model demonstrated survival stratification in both cross-validation and external validation experiments, based on microarray expression data. They found expression of CSFT3, ABCD3, MAL and APMCF1 to be the most influential; in fact the decision tree was almost entirely driven by the first two of these.

Mairinger *et al* screened 770 immune related genes to identify 11 differentially expressed genes associated with response to platinum treatment. They find that expression of HS11B1, DNBT1, CKLF, NUP107, CCL18, LY96, ATG7, SLAMF7, CXCL9 is associated with better survival, while IKBKG and SDHA associate with with poor survival.

The platinum based drug Cisplatin is a key treatment in ovarian cancer, yet tumours often develop resistance, possibly related to the host's immune response. Understanding and predicting such resistance is therefore of clinical importance and is

addressed by Mairinger *et al* [15] through expression analysis using the NanoString platform with a panel of 770 immune related genes. They find the following genes to be significantly related to platinum resistance: KLRC1, TCF7, CD274, HSD11B1, COLEC12, PDGFC, FCF1, BMI1, TNFRSF9, ATG10, EWSR1.

Finding a clear, robust correlation between gene expression patterns and survival is not straight forward. One substantial meta-analysis applied 16 previously published gene sets to two studies from the Gene Expression Omnibus (GEO) (199 samples) which included survival information, and found no significant predictive power [16]. Although, within the same study, the authors did identify predictive genes, in particular SNGC, MAPT, ESR2 and PGR.

Expression of a single gene – CD38, which codes for a transmembrane glycoprotein – has recently been shown to have survival prognostic value on its own [17]. CD38 has a role in the breakdown of nicotinamide adenine dinucleotide (NAD^+), and is also a cell marker of lymphocytes. CD38 is over expressed in myeloma cells, thence its importance as a prognostic marker and a potential drug treatment target. It is worth mentioning that the analysis in this paper made use of the Gene Expression Profiling Interactive Analysis (GEPIA) database, which conveniently supports survival analysis with respect of any given gene, against a range of TCGA hosted gene expression databases.

1.2.4 Association with other risk factors

Obesity has been shown to negatively impact prognosis in ovarian cancer, possibly through difficulties in matching chemotherapy dose with patient BMI [18]. A more direct molecular link is proposed by Cuello *et al* [19]. They used NMF based clustering to demonstrate that expression of genes linked with obesity and lipid metabolism impart poorer progression free survival, independent of known HGSOC driver gene expression. The link is complex however, there being many confounding non-molecular based reasons for a link between high BMI and HGSOC survival, including later diagnosis and compromised operability. Some studies (e.g. [20]) conclude no overall link at all.

1.3 Matrix factorization for dimensionality reduction

Dimensionality reduction concerns representing high dimensional data in far fewer dimensions, with minimal loss of information. This is possible in many cases because dimensions – or equivalently elements of a feature vector – are highly correlated. In the case of gene expression analysis, the dimensionality is of the order 20,000, i.e. the number of transcripts (mRNA transcribed genes) being considered. However, it is known that sets of genes act in concert, their expression thus being correlated, so reducing the true degrees of freedom in the data. The challenge is to uncover these true degrees of freedom, which constitute our "patterns of gene expression".

The three matrix factorization methods considered in this work are described below. All are forms of dimensionality reduction, but have differing criteria they aim to optimise. The number of target dimensions (or components, factors, metagenes – these terms are used somewhat synonymously) is referred to as the rank, k.

1.3.1 Non-negative matrix factorization (NMF)

As the name implies, this is only applicable to matrices with +ve or zero elements, and finds components which are themselves +ve or zero. This property make the resulting factorization more interpretable since the components are strictly additive. NMF has been applied widely in -omics research, originally to gene expression microarrays and latterly to RNA-seq datasets. A particular value of NMF is that it allows for input samples to be directly assigned to one of k clusters. The optimization function on which NMF is based can be tuned to achieve a *sparse* factorization result, that is one in which only a small number of elements in the achieved factorization are non-zero, a property which further simplifies interpretation [21].

1.3.2 Independent component analysis (ICA)

The goal of ICA is to represent the given matrix with components which are statistically independent of each other. It was originally proposed to solve the blind source sep-

aration problems [22]. If the input matrix of samples were drawn from a multivariate Gaussian distribution, then the result would be no different to that of PCA. Where the data is non-Gaussian however, ICA results in components which separate out independent sources of variation.

1.3.3 Principal component analysis (PCA)

PCA is a well known technique for dimensionality reduction, based on eigenvalue decomposition. It is arguably the ideal, efficient solution for data which is multivariate Gaussian distributed – or can be assumed to be so. PCA results in components which are ordered by the proportion of total variance they explain, and are mutually orthogonal.

Conventionally, expression arrays are oriented with genes (or transcripts) in rows, and samples (e.g. patients) in columns. A typical expression array might have in the order of tens of thousands of rows and hundreds of columns. MF methods reduces this large $m \times n$ (genes $\times$ samples) matrix into two smaller matrices. In the general terminology of Stein-O'Brien *et al* [23] these are the $m \times k$ *pattern matrix* and the $k \times n$ *amplitude matrix*. k is typically small of the order 10, and refers to the number of extracted *factors* (or components); it is the rank of the factorization. Since we are focussing on transriptomics, rather than pattern and amplitude we will use the terms *metagenes* and *metasamples* respectively, which are in common useage. The symbol conventions for these matrices varies depending whether NMF, ICA or PCA is being discussed. NMF generally uses W and H respectively while ICA generally uses S and A. PCA tends to be differently formulated, but W and T are often used. The original matrix to be factorized is variously named X or V.

In this work I wish to treat and discuss the three factorization methods in a unified way. Thus, for the remainder of this dissertation the following notational conventions are adopted:

Expression matrix: $X\ (m \times n)$

Metagene matrix: $W\ (m \times k)$

Metasample matrix: $H\ (k \times n)$

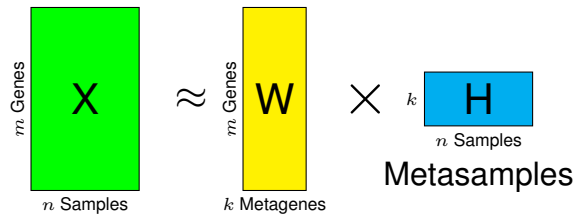
where N is the number of patients (or samples), m is the number of genes and K is the factorization rank, i.e. the number of components or factors. Thus the factorization is written

$$X \approx WH \tag{1}$$

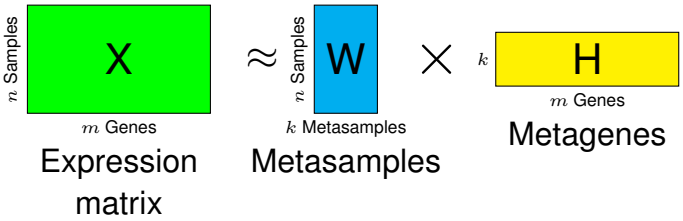
Adding subscripts to indicate the row,column orientation of the matrices:

$$X_{m,n} \approx W_{m,k} H_{k,n} \tag{2}$$

as illustrated in figure 1 (a).



(a) Optimising independence of Metagenes



(b) Optimising independence of Metasamples

Figure 1: Two ways of configuring matrix factorization in the context of gene expression analysis. In configuration (a) NMF and ICA will optimise the *metagenes*, while in (b) the *metasamples* are optimised.

In the case of NMF and ICA, the W and H matrices cannot be trivially exchanged and transposed. This is because the optimisations (sparsity and independence respectively) which define these algorithms are focussed on the W matrix. It is perfectly possible to apply these algorithms to gene expression analysis with exchanged and transposed meanings of W and H, and this is illustrated in figure 1 (b). In this case

it is the properties of the *metasamples* which are optimised. Thus, the two forms are different in substance, not simply in notational convention.

There is substantial confusion and lack of clarity in the way that matrix factorization, particularly ICA, is applied in transcriptomics research. “Surprisingly, both ways of applying ICA to omics data are wide-spread, and sometimes it takes an effort to figure out in which way ICA was applied” [24], and “Different protocols to apply ICA to transcriptomic data exist and currently no single standard approach has been defined. The main difference in the existing approaches consists in what is considered as source signal matrix in the decomposition” [25]. According to [25], references [18, 26, 27, 28], optimise metagenes, while references [29, 30] optimises metasamples.

1.3.4 Determining the optimum number of factors

The number of factors, or rank k, to extract is a key decision in any matrix factorization approach. In this regard, PCA differs from ICA and NMF. In PCA it is reasonable to extract all factors, setting $k = n$, the factors being ranked by the associated eigenvalue which also ranks the proportion of variance explained. However, that approach is not valid for ICA and NMF, since differing sets of factors will be obtained for different k, and for a given choice of k all factors are equally important – they do not rank [23].

The problem of determining the optimal k for a given expression matrix is addressed by Kairov *et al* [26], based on optimizing the *stability* of the components over multiple algorithm initializations.

1.3.5 Comparisons dimensionality reduction methods for gene expression analysis

Recent work by Way *et al* [31] was aimed directly at answering the question of which dimensionality reduction method was best suited to gene expression data. They considered PCA, ICA, NMF and also non-linear neural network methods – directional autoencoders (DAE) and variational autoencoders (VAE). Briefly, these latter methods work by training a neural network to reconstruct an input signal while being forced through a network “bottleneck”. Their conclusion was that no single method was optimal, but

best performance – judged by capturing the most biological pathway-associated features – was achieved by taking the extracted latent dimensions of all the methods, an approach they call the “BioBombe”.

Meng *et al* [29] extensively review dimensionality reduction methods applied to multi-omic datasets. PCA, ICA and NMF are discussed, but the main focus is on less well known algorithms applicable to multiple datasets, including multiple coinertia analysis and consensus PCA. The essential purpose of these methods is to transform diverse datasets – transcriptomics, genomics and metabolomics – into a crommon and manageable representation space.

The question of reproducibility across datasets of matrix factorization methods is considered by Cantini *et al*. [25]. They experiment with several MF methods applied to a total of 26 cancer transcriptomic datasets (including ovarian). They conclude the best approach to be ICA, stabilized by the use of Reciprocally Best Hit graphs; in essence requiring reciprocal identification of metagenes between at least two datasets.

1.4 Research gaps and questions

Considering the foregoing review of the literature, there has clearly been much work by large teams of researchers directed at understanding the gene expression patterns in HGSOC, as one would expect given its clinical importance. Dimensionality reduction is a huge and active field – the bread-and-butter of machine learning research. As we have seen, conclusions both on the underlying biology and the optimal methodology are far from consistent, however. Some approaches – such as the BioBombe reviewed above – are complex with results likely to be difficult to reproduce.

Robustness of research results is of course fundamental. MF methods offer a means of *unsupervised* learning on transcriptomic datasets to identify key genes active in a disease process. Would two teams working with identical data and using similar MF methods obtain similar metagenes and so arrive at the same biological conclusions? The *stability* of MF methods is well considered by many authors, although this has typically been focussed on stability of the underlying algorithm – NMF and ICA are

stochastic in nature, sensitive to initial conditions. Much less considered is stability in the presence of *sampling error* which inevitably exists particularly in small datasets.

These and wider considerations suggest the following research questions.

1. What underlying biological processes are at play in HGSOC as seen through the patterns of gene expression? Do these confirm published results?
2. How do the patterns of gene expression uncovered relate to genome level features, such as those marking genome instability?
3. Do MF methods yield factors which are predictive of patient survival?
4. How should MF methods be applied to achieve robust results on datasets of limited size?
5. Does MF detect and remove technical batch effects?
6. Which MF method is best suited to transcripomics analysis?

2 Methodology

2.1 Outline

Our overall approach is to use *unsupervised* machine learning methods to represent gene expression data in a small number of features, then to study whether these features correlate with clinical and biological variables. Unsupervised methods have the advantage of being applicable to large datasets for which metadata (e.g. clinical information) is not available. Results can then be transferred to smaller datasets which do have metadata. This is appropriate in our case, since of the two datasets available (described below), one (TCGA) has relatively large n but little available metadata, while the other (AOCS) is much smaller but has more useful metadata of genomic features. This motivates an approach of unsupervised learning on the TCGA dataset which is then transferred to the AOCS dataset for evaluation.

2.1.1 Datasets

Two gene expression datasets were made available for this work (N refers to number of patients, m refers to number of protein coding genes).

1. The Cancer Genome Atlas (TCGA) derived, N=374 patiens, m=19,601 genes, with metadata on survival
2. Australian Ovarian Cancer Study (AOCS) [3], N=80, m=19,730, with metadata on survival, cellularity and additional genomic features. Several of the studies reviewed earlier also use data from the Australian Ovarian Cancer Study (AOCS): [3, 5, 19, 18].

Expression data for the AOCS and TCGA datasets was received for this project in spreadsheet format, having been derived from the RNA-Seq data with normalization by variance stabilizing transformation applied. The RNA-Seq data is publicaly available, downloadable from the Pan-cancer Analysis of Whole Genomes (PCAWG) data portal for AOCS, and from the Genomic Data Commonas (GDC) data portal for TCGA. The provenance and processing of these datasets is described in [5].

2.1.2 Methods

An overview of the methodology adopted in this work is shown in figure 2, consisting of:

1. Identification of a consistent gene set between the TCGA and AOCS datasets.
2. Unsupervised metagene extraction by matrix factorization methods : NMF, ICA and PCA.
3. Metagene selection for robustness by k-means clustering of bootstrap resampled factorizations evaluated by silhouette score, based on the TCGA dataset.
4. Investigation of biological significance of extracted metagenes by gene enrichment analysis (GEA) against the Gene Ontology (GO).

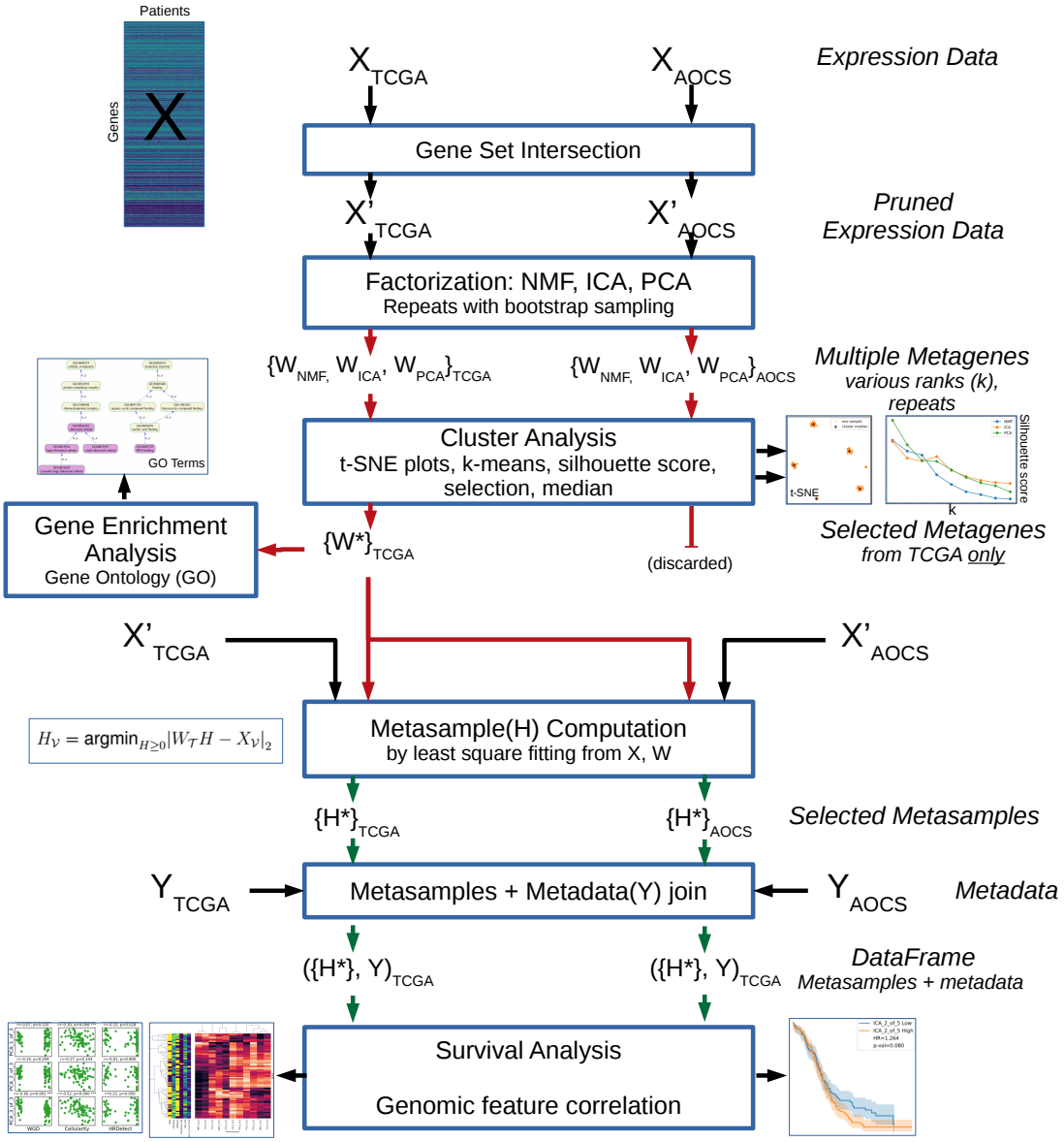


Figure 2: Overview of methodology as information flow. The diagramatic convention here is that at each stage, both datasets – TCGA and AOCS – are processed separately by the given algorithm. The exception is Gene Set Intersection, which involves interaction between the datasets. (The investigation of batch effects on the combined dataset is not shown in this diagram to avoid over complexity).

5. Computation by least squares optimisation of metasamples associated with each TCGA derived metagene against TCGA and AOCS datasets.

6. Alignment of available metadata with computed metasamples, for TCGA and

AOCS.

7. Survival analysis on TCGA and AOCS to investigate relationships between metasamples and patient survival.
8. Scatter plots and heat maps to investigate the relationships between metasamples and genomic features available for the AOCS dataset.
9. Investigation of batch effects by repeating the above analysis pipeline on the *combined* AOCS / TCGA dataset (N=80+273=353).

These steps are detailed in the following main sections.

2.1.3 Tools

- DELL XPS 13, 16GB RAM, running Linux Mint version v19.
- Python v3.6.8
- PyCharm Integrated Development Environment, v2020.1.3
- Numpy for high performance matrix manipulation, v1.18.15
- Pandas for data frame handing, v1.0.5
- Matplotlib and Seaborn for general plotting, v3.2.2 and v0.10.1
- SciPy for least-square solutions and statistical tests, v1.5.0
- Scikit-learn for matrix factorization and k-means clustering, v0.23.1
- GOATOOLS package for GEA against GO, v1.0.6
- Lifelines package for survival analysis, v0.24.16.
- Github for source control.

2.2 Syncrhonisation of gene sets accross AOCS and TCGA

In order to allow factorizations found in the TCGA dataset to be applied to the AOCS dataset it is necessary (or at least convenient) to synchronize the set of genes over which the expression matrices are defined. As provided to this project, the TCGA and AOCS datasets cover 19,610 and 19,730 protein coding genes respectively, 19,566 of which are common to both (according to ENSG encodings). Thus, both datasets were pruned to the 19,566 intersection set and ordered consistently.

2.3 Matrix factorization computation

One of the aims of this work is to compare the efficacy of three methods of dimensionality reduction – NMF, ICA and PCA – as explained in the introduction. Some of the key algorithm hyper parameters of each method were investigated and tuned with respect to reconstruction accuracy, specifically the root-mean square (RMS) difference between X and WH. The following parameters were explored in each case:

NMF: Parameters `max_iter` (algorithm iterations) and `tol` (tolerance of convergence) were optimised for good accuracy and acceptable execution time. `max_iter` = 5000 and `tol` = 0.01 was used. Other parameters of interest are `alpha` (multiplier for regulation term) and `l1_ratio` (multiplier for L1 regularization, when `alpha` > 0). L1 regularization favours elements being precisely zero, whereas L2 regularization will encourage them to be small. These parameters have not been explored in this project, and the default `alpha = 0` (no regularization) was used.

ICA: Parameters `max_iter` and `tol` were investigated as for NMF above. `max_iter` = 5000 and `tol` = 10^{-5} was used. Additionally, options for the entropy function which forms the basis of the optimiziation were investigated.

PCA: This is in principle a deterministic algorithm based on eigenvector decomposition. However, `sklearn.decomposition.PCA` uses a more efficient 'randomized' algorithm when the given matrix is larger than 500 in both dimensions. Thus in our use case PCA is seen to have (slightly) stochastic behaviour.

2.4 Metagene extraction by cluster coherence under bootstrap resampling

Deciding on the number of components (factors, metagenes) to extract – that is the factorization rank – is key. Taking more components results in more accurate representation of the observed expression matrix and provides more avenues to explore the underlying biology. However, it is important that the metagenes are *stable*, that is that

they have reliable meaning when transfered to other datasets. There are two sources of variation or instability to consider.

Firstly, NMF and ICA are inherently stochastic algorithms, sensitive to their starting state, so repeated runs give different results.

A second and more fundamental source of variation of relevance to all three factorization methods is *sampling error*. Our factorizations are based on a small ($N = 80$ or $N = 374$) sample of patients drawn from the population of HGSOC patients; our particular datasets are just two examples of many different "draws" which could have been made from that population.

Bootstrap sampling (also know as *Monte Carlo* simulation) is a common method of empirically propagating the consequence of sampling error when the distribution or processing operations are difficult to model mathematically. This is implemented by performing factorizations multiple times, at each iterations choosing N samples from the N available *with replacement*. For this work 50 repeats were performed, being a compromise between achieving an adequate simulation without overly burdensome computation time. The process is described below and summarised as pseudocode in figure 3.

A complication arises in the computation of ICA and PCA factorizations, in that essentially the same factor can arise as w or $-w$ – i.e. $180°$ rotated vectors. These would appear as separate clusters in repeated sampling, and confound attempts to collect and aggregate. The solution adopted here is to normalise each factor by requiring that the most extreme element – i.e. having the greatest absolute value – is positive, the whole factor being negated if this is not the case. This is arguably over simplistic, but seems to be effective in practice.

Each of the three factorizer methods was evaluated for rank k between 2 and 10. In each case, 50 iterations of factorization are performed on bootstrap samples, generating $50k$ instances of 19,566 dimensioned metagenes. Two dimensional t-SNE plots were generated for visualization purposes. If sampling and algorithm initialisation error are modest then we expect to see k tight clusters of points. To avoid lengthy computation in the t-SNE clustering, the metagenes were first reduced to $r = 20$ dimensions by

PCA; brief experiments showed that this reduction had negligible effect on the t-SNE visualisation providing $r \geq 10$.

In order to obtain median estimates of the k metagenes from the $50k$ which were generated, k-means clustering was performed in the PCA reduced ($r = 20$ dimension) space, delivering k sets of points. These were referenced back to the original 19,566 dimensioned metagenes and the per-dimension median calculated. These k median metagenes were saved to a file named for the specific factorizer and rank k.

The k-means clustering further allowed for a quantitative assessment of cluster coherence for the particular choice of factorizer and rank, via the *silhouette score* (see [32]). In brief, this is a measure of how close (by Euclidean distance) each point in a cluster is to other points in the cluster versus points in *other* clusters. The score is in the range -1 to +1, with 0 implying random scatter and 1 implying all points of a cluster perfectly overlay.

The t-SNE plots and silhouette scores were assessed visually to decide, for each of the three factorizers, the highest rank with good cluster coherence. The median points of those clusters then formed the metagenes to take forward to gene enrichment, survival and genomic feature correlation analysis. Note that the t-SNE plots are used only for visualisation; they are not involved in the computation of median metagenes or silhouette score.

Initial investigation on the N=80 AOCS dataset showed very poor cluster coherence, even for k=2. Thus, it was decided to compute and select metagenes only from the N=374 TCGA dataset. The selection rationale is set out in the results section, but it is convenient to state here that the following ranks were selected: $K_{\text{NMF}} = 3, K_{\text{ICA}} = 5, K_{\text{PCA}} = 3$. Thus, there were $3 + 5 + 3 = 11$ metagenes taken forward for follow-on analysis.

A point to be aware of is that because PCA metagenes are treated the same as NMF and ICA metagenes, being extracted by the same clustering approach, the *ordering* of PCA components is not necessarily maintained. Thus, component 'PCA-1' is not necessarily the first PCA component in the usual sense – i.e. the component capturing greatest variance. This could be fixed but would require special case code.

Metagene Stability Assessment

```
for factorizer in NMF, ICA, PCA:
   for k in 2..10:
      Ws = []
      for j in 1..50:
         X = bootstrap sample N from N with replacement
         W, H = factorizer(X, rank=k, seed=j)
         append W to Ws

      # Ws is a list of 50*k metagenes, each of length 19,566
      # Pragmatically reduce metagenes to 20 dimensions by PCA

      Ws_reduced = PCA(Ws, rank=20)
      plot t-SNE(Ws_reduced)
      clustering = k-means clustering(Ws_reduced, n_clusters=k)
      score[factorizer, k] = silhouette score(clustering)
      median_metagenes = calculate medians for k clusters of metagenes
      save median_metagenes to file by factorizer and k
   plot score vs for factorizer k
```

Figure 3: Pseudocode for generating median metagenes and assessing their stability to random algorithm initialization and sampling error.

2.5 Determining metagene similarity by Jaccard index

It is pertinent to ask "how similar are the 11 metagenes which emerge from the foregoing factorization and cluster analysis?". We expect those coming from a single factorization method to be distinct by construction. But perhaps ICA and PCA identified similar components. Similarity with NMF derived components is possible but less likely due to the positivity constraint. A possible approach to measuring similarity might be to treat the components as vectors and determine the angle between them, via the scalar or "dot" product. But in such a high dimension space (19,566) any pair of vectors will be very close to $90°$.

An alternative is to consider the intersection of the gene sets which each metagene highlights. In the GEA which follows, candidate genes will be identified for each meatagene. A standard measure of set similarity is the Jaccard index, (or similarity), defined on a pair of sets as:

$$\text{Jaccard}(A, B) = \frac{A \cap B}{A \cup B}.$$

The measure lies between 0 (no similarity) and 1 (identical), and is symmetric. Jaccard similarity was thus calculated for all pairs of the 11 genes sets and visualised as a square heatmap.

2.6 Gene enrichment analysis (GEA)

The metagenes extracted by the above described factorization and clustering process provide valuable information into which genes vary in expression in the study samples; thus in our case, which genes are influential in the dominant expression signal in HGSOC. In order to gain insights into what biological processes are involved, GEA against the Gene Ontology (GO) was carried out for each of the 11 metagenes individually.

The essence of GEA is to compare a candidate set of genes with many functional gene sets, asking the question "are there significantly more (or less, i.e. depletion) intersecting genes than would be expected for the same number of genes being drawn at random (without replacement) from the total population of genes in the study.

For each metagene, the candidate gene set was determined as those genes having weights outwith three standard deviations of the mean. This set was analysed using the Python GOATOOLS package [33].

The gene ontology was downloaded from `http://purl.obolibrary.org/obo/go/go-basic.obo`. (`Purl.org` is a resource for managing permanent URLs; `obo` refers to the Open Biological and Biomedical Ontology (OBO)). Annotations linking human genes to GO concepts was downloaded from the GO website, specifically `http://geneontology.org/gene-associations/goa_human.gaf.gaf`. The gene population was defined as the common 19,566 protein coding genes (see section 2.2) against which the metagenes were computed.

Uncorrected p-value threshold was set to 0.01. Multiple hypothesis significance

testing used the false discovery rate (FDR) method of Bengamini and Hochberg, the FDR threshold being set to 0.01. This is more stringent than the 0.05 value which is commonly used, and was chosen since multiple metagenes are being analysed, implying multiple hypothesis testing over and above that which is accounted for by FDR filtering within a single GEA run. (Arguably, $0.05/11 \approx 0.005$ should have been used).

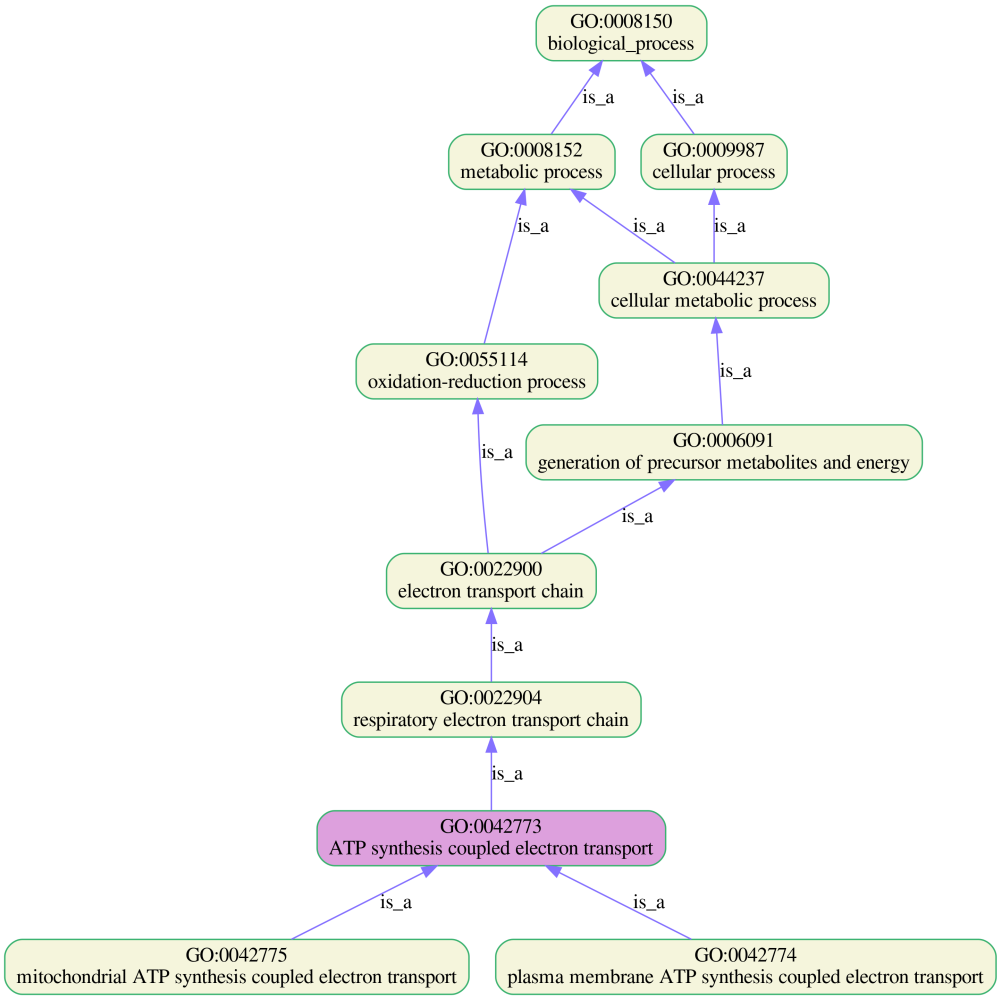


Figure 4: Example of a small section of the Gene Ontology (GO), focussed GO:0042773 (in purple), showing parents and children of the term. Generated by GOATOOLS.

The result of this analysis, per metagene, is a list of enriched (or depleted, however no depleted terms were found) GO terms, each with an associated list of involved genes and a FDR significance level. This list can be inspected directly for insights, but that misses the point of the GO, which is to organise the terms hierarchically by 'is a' relationships. Thus, the enriched terms were rendered graphically to show them in the

context of their parent terms; an example is shown in figure 4

2.7 Transfer of learned metagenes to a novel dataset

It is fundamental to our approach that *metagenes* determined on the basis of one (large but metadata poor) dataset can be used to generate *metasamples* for a different (small but metadata rich) dataset. In the exposition below we refer to these as the *training* and *validation* cohorts respectively. Rank k matrix factorization on the training dataset $\mathcal{T}$ (e.g. TCGA) results in:

$$X_{\mathcal{T}} \approx W_{\mathcal{T}} H_{\mathcal{T}} \tag{3}$$

where $X_{\mathcal{T}}$ is the expression matrix of shape $(m_{\mathcal{T}}, n_{\mathcal{T}})$, with $m_{\mathcal{T}}$ the number genes and $n_{\mathcal{T}}$ the number of patients. $W_{\mathcal{T}}$ is the metagene matrix of shape $(m_{\mathcal{T}}, k)$ and $H_{\mathcal{T}}$ is the metasamples matrix of shape $(k, n_{\mathcal{T}})$.

We wish to apply the factorization learned on $\mathcal{T}$ to a novel dataset $\mathcal{V}$ (specifically the AOCS dataset) of shape $(m_{\mathcal{V}}, n_{\mathcal{V}})$. Importantly, $n_{\mathcal{V}} = 1$ reflects the application of these methods to a single patient in a clinical setting. To apply the learned factorization we need to find $H_{\mathcal{V}}$ as in the factorization

$$X_{\mathcal{V}} \approx W_{\mathcal{V}} H_{\mathcal{V}} \tag{4}$$

In both the experimental and clinical situation we are given $X_{\mathcal{V}}$ but not $W_{\mathcal{V}}$ and require to find $H_{\mathcal{V}}$. We *cannot* simply perform the matrix factorization on dataset $\mathcal{V}$ since $n_{\mathcal{V}}$ may be small, or even a single patient. However, if patients in datasets $\mathcal{T}$ and $\mathcal{V}$ are drawn from the same population (ovarian cancer patients), then $X_{\mathcal{T}}$ and $X_{\mathcal{V}}$ can be expected to have similar distributions w.r.t to their columns, and thus $W_{\mathcal{T}}$ and $W_{\mathcal{V}}$ can be expected to be equivalent within sampling error. This only makes sense if the two expression matrices $X_{\mathcal{T}}$ and $X_{\mathcal{V}}$ are defined over the *same set of genes*, so that $m_{\mathcal{T}} = m_{\mathcal{V}} = m$, in which case $W_{\mathcal{T}}$ and $W_{\mathcal{V}}$ have the same shape of (m, k).

We thus need to solve for $H_{\mathcal{V}}$ in

$$X_{\mathcal{V}} \approx W_{\mathcal{T}} H_{\mathcal{V}} \tag{5}$$

This can be solved by the method of least squares. In the case that the original factorization was by NMF, we use non-negative least square regression (NNLS):

$$H_{\mathcal{V}} = \text{argmin}_{H \geq 0} |W_{\mathcal{T}} H - X_{\mathcal{V}}|_2 \tag{6}$$

where $|\cdot|_2$ indicates Euclidean distance or L2-norm. The function `scipy.optimize.nnls` is used. For ICA and PCA we use ordinary least square regression, formulated as above but dropping the $H \geq 0$ constraint, using `scipy.linalg.lstsq`.

The end result of the analysis is the $H_{\mathcal{V}}$ matrix of shape $(k, n_{\mathcal{V}})$, thus delivering for each dataset in our $\mathcal{V}$ a feature vector – or metasample – of length k.

Since in this work the efficacy of three factorization methods are being studied (NMF, ICA and PCA), three different H matrices are taken forward, each having ranks as for the associated metagenes (3, 5 and 3 respectively).

2.8 Reconciling computed metasamples and metadata

To carry out patient level analysis – survival analysis, heatmaps and genome feature correlation – as described in the following sections, it is necessary to compile tables (Pandas DataFrames) which bring together per-patient metasamples relating to each of the selected metagenes, with associated metadata – such as cellularity and survival information. The metagenes always derive from the TCGA dataset (because of its larger size as explained in section 2.4), but we wish to apply these metagenes as metasamples to the expression data and associated metadata of either the TCGA or AOCS datasets. The mathematics of transferring metagenes across datasets has been described above. Care is required to ensure that expression matrices are aligned correctly with metagenes with respect to their genes, and that metadata is aligned correctly with metasamples with respect to patient identifiers. Note that the same transferring

approach is taken when metasamples are required for TCGA, even though we could in that case obtain the H matrix directly from the factorization.

2.9 Survival analysis

Survival analysis was performed to investigate whether the metasamples derived from the selected metagenes correlate with patient survival. Overall survival (OS) data is available for both the TCGA and AOCS datsets. Additionally, progression free survival (PFS) data is available for AOCS. As described above, metagenes were obtained by factorization on the TCGA dataset only. Application to TCGA (for OS) thus represents an in-sample test, while application to AOCS (for OS and PFS) is a more exacting out-of-sample test. There are thus three analyses to consider: 1) TCGA$\rightarrow$TCGA (OS) meaning TCGA derived metagenes are tested for OS on the TCGA dataset, 2) TCGA$\rightarrow$AOCS (OS) meaning TCGA derived metagenes are tested for OS on the AOCS dataset, and 3) TCGA$\rightarrow$AOCS (PFS)... as 2) but for PFS.

Analysis was performed using the Python `Lifelines` package [34]. Metasample values were binarized to 0, 1 by thresholding at the median value. Kaplan-Meier plots were made in respect of derived metasample for each of the three analyses – making 11×3 plots each with two survival curves with 95% confidence intervals. A hazard ratio (HR) was calculated for each analysis by fitting Cox's proportional hazards model with p-value relating to the hypothesis that HR is significantly different to 1.0.

2.10 Investigating correlation between metasamples and genomic features

For the AOCS dataset (only) per-patient high-level genomic features were available as follows:

WGD : Whole genome doubling – a marker of genome instability, a binary feature.

Cellularity : proportion of cells belonging to the tumour (as opposed to surrounding normal tissue).

HR Detect : A predictor of homologous repair deficiency based on established mutational features [5].

Mutational Load : A measure of the total number of mutations present in the tumour genome.

CNV Load : Copy-number variation, i.e. deviation from the normal diploid cell compliment.

SV Load : Structural variation, a measure of the degree of chromosomal rearrangements such as translocations and inversions.

Scatter plots were generated between each metasample and each genomic features – a grid of $11 \times 6 = 66$ plots in all. These were based on the reconciled metasamples and metadata as described in section 2.8. Pearson's correlation coefficients (r) and associated p-values (i.e. the probability that r does not differ from zero) were calculated. However, since WGD is a binary feature, the more appropriate Point-Biserial correlation (see [35]) was used in that case. A p-value significance threshold of 0.01 was chosen, although with 66 hypotheses being tested, this does risk false discovery.

As an alternative visualisation of the relationship between metasamples and genomic features, and to see the relationship between metasamples themselves, a clustered heatmaps was generated using the Seaborn `clustermap()` function based on the same underlying data as above.

2.11 Investigating batch effects by analysis of a combined dataset

It has been claimed (e.g. [23, 36]) that matrix factorization methods are an effective means of identifying and removing batch effects. To verify this, the TCGA and AOCS datasets (after gene set intersection), were horizontally concatenated producing an $N = 80 + 374 = 454$ patient dataset with a known batch artefact. All of the previously described analyses were repeated for this combined dataset (however, survival analysis is not shown). Ideally, the factorization rank selected for each method should be reconsidered, since the larger dataset could justify extracting more components.

For simplicity this was not done however, the previously selected $k = 3, 5, 3$ for NMF, ICA and PCA respectively being maintained. Note however that we should not expect the extracted metagenes and derived metasamples to be equivalent. Genomic feature correlation w.r.t these newly extracted components was performed using the AOCS datasets. These results were inspected for evidence of some metasamples aligning with technical batch effects while others align with biological processes.

2.12 Codebase and plotting conventions

The described methodology was implemented in Python. All code is available in a github repository: `https://github.com/ipoole/HgsocTromics`. Good software engineering practice has been followed, with an object-oriented design, frequent commits and unit testing. A base class is used to present a uniform interface for the three factorization methods. Intermediate results are cached to file with lazy computation for efficient working without imposing evaluation order dependency. Unit tests are based on tiny expression matrices, just 100 genes by 10 patients, thus the whole test suite of over 65 tests executes in around 45 seconds. The fully automated (except for rank selection) analysis pipeline takes approximately three hours from scratch, the majority of that time being taken for the repeated factorizations over the range of considered ranks. The total codebase is approximately 2,900 lines of Python.

All plots are generated in vector graphics pdf format to ensure smooth scaling to any resolution. A consistent colour scheme of blue, orange and green is used for plots relating to NMF, ICA and PCA respectively. In figures metagenes / metasamples are consistently referred to by, for example, “NMF-2-of-3” – meaning the 2nd component of the rank $k = 3$ NMF factorization. In text this is shortened to “NMF-2” where the rank is understood.

3 Results

3.1 Consideration of sampling error is crucial to finding robust metagene signals

From figure 5 it can be seen that with sampling error excluded (top), clusters appear reasonably coherent, but when sampling error is modelled by bootstrap sampling then the factorizations become much less stable. This demonstrates that *sampling* error is far greater than errors due to algorithm initialisation when the dataset is small (N=80), making metagene extraction unreliable. Figure 6 makes the same comparison for the larger N=374 TCGA dataset, in which it can be seen that the impact of sampling error is not so severe. For brevity only results for ICA are shown here; similar result for NMF and PCA can be found in the appendix, figures 21 and 22.

For this reason it is was decided to perform all metagene extraction on the larger TCGA dataset, in the expectation that the obtained metagenes will be more robust and likely to better generalise to other datasets.

In choosing the factorization rank for each method, both the t-SNE plots and silhouette scrore graphs in figure 7 were carefully considered. Firstly, $k > 2$ is desirable to have sufficient information to work with. NMF cluster coherence seems to deteriorate after $k = 3$. Looking at the graph of silhouette scores (figure 7, bottom), ICA appears to have a sweet spot at $k = 5$. For PCA, $k = 3$ or 4 both seem reasonable. As already noted earlier, the following choices were made: $K_{\text{NMF}} = 3, K_{\text{ICA}} = 5, K_{\text{PCA}} = 3$.

There is a curious artefact visible in the NMF factorization clusterings of figure 7, top row. For $k = 2, 3$ and 5 several of the clusters show a bi-modal character. This is not observed in the fixed sampling case (not shown). The artefact is difficult to explain. It cannot be the $180°$ rotation issue discussed earlier, since this does not apply to the all +ve components; I have no explanation.

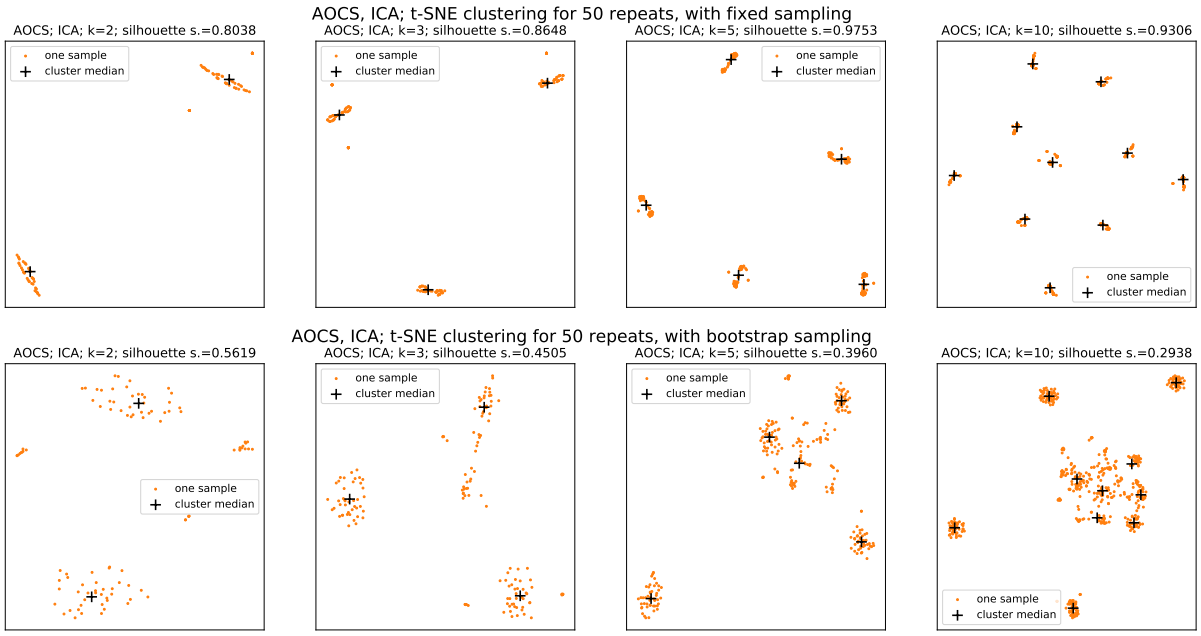


Figure 5: Clustering of metagenes from ICA facorizations on the N=80 AOCS dataset, comparing *fixed* sampling (top row) with *bootstrap* sampling (bottom row) for a selection of factorization ranks.

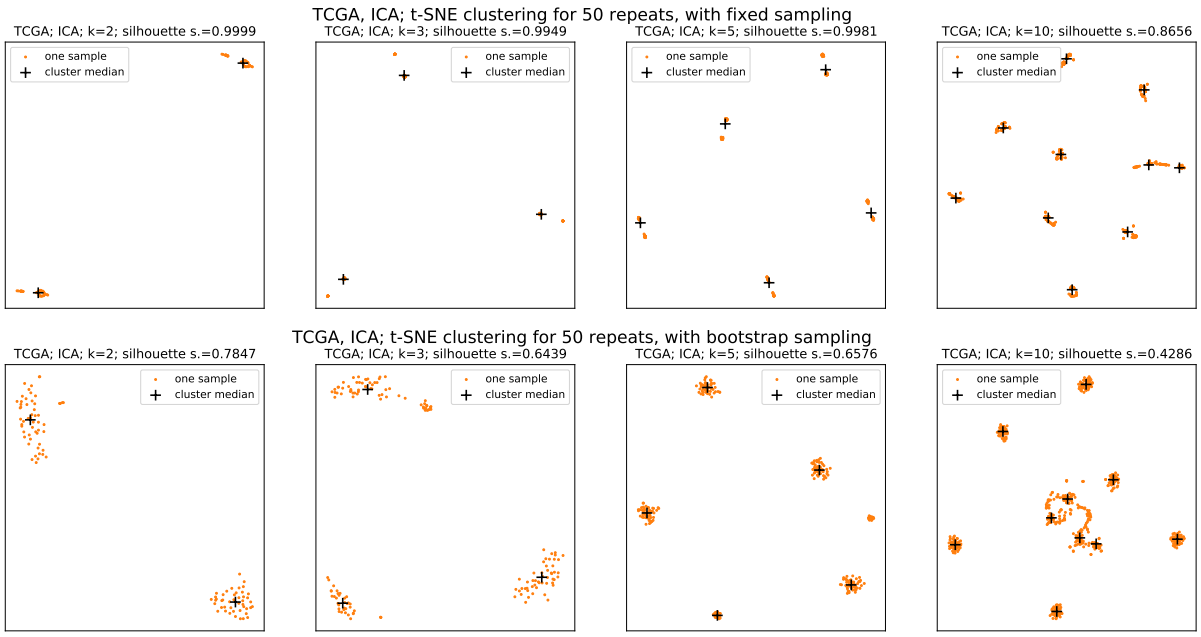


Figure 6: Clustering of metagenes from ICA facorizations on the N=374 TCGA dataset, again comparing fixed and bootstrap sampling.

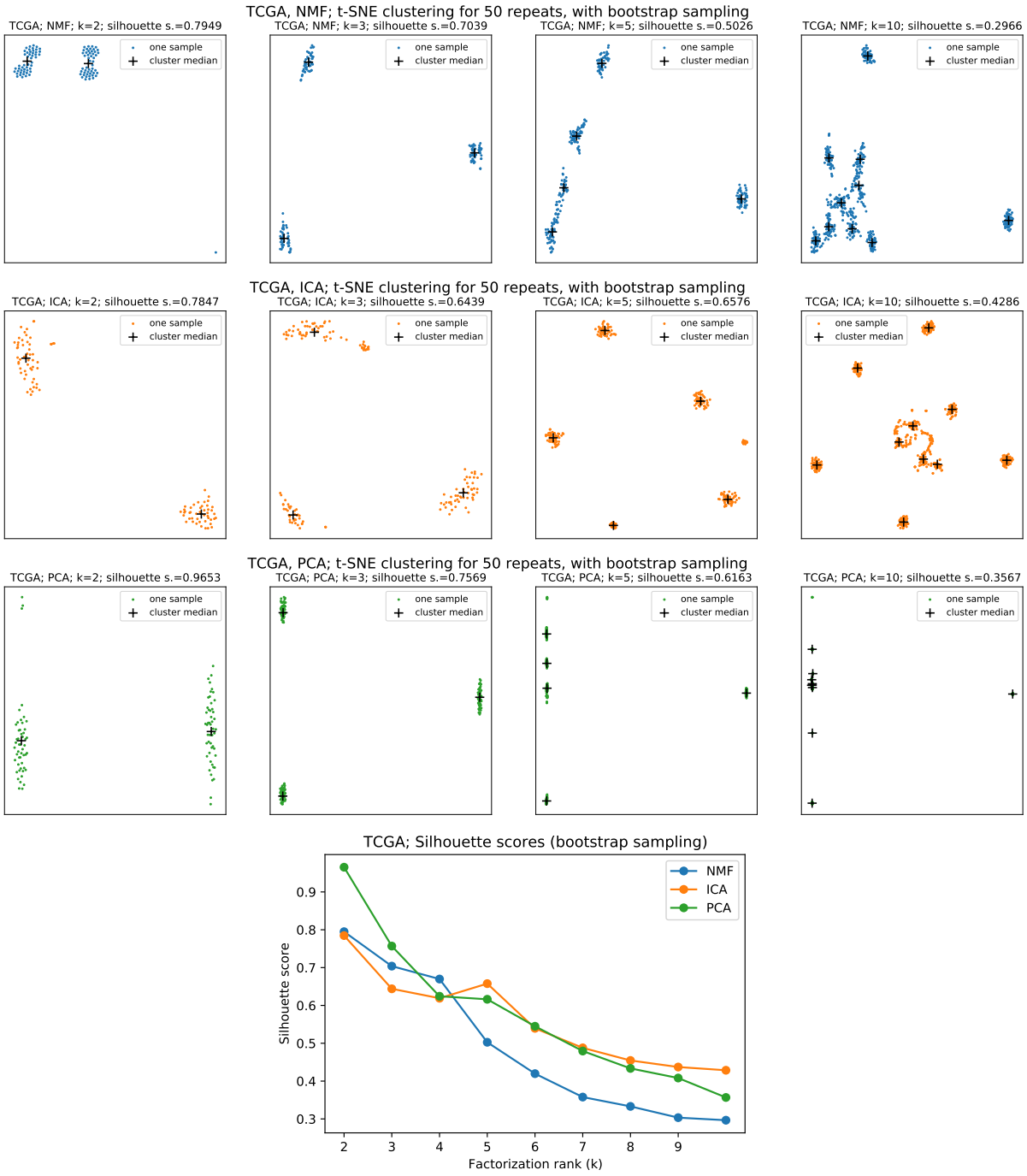


Figure 7: Metagene clustering for all three methods applied to the N=374 TCGA dataset with bootstrap sampling, over a range of factorization ranks. The silhouette scores are also plotted (bottom). It is on the basis of this figure that the factorization ranks for each method were selected.

3.2 Jaccard similarity shows metagenes identify intersecting gene sets in some cases

The Jaccard heatmap of figure 8 shows that some pairs of metagenes identify overlapping sets of genes, in particular the pairs (ICA-1, PCA-2: J=0.61) and (ICA-5, PCA-3: J=0.49). As expected however, there is very little similarity between components *within* a factorization method.

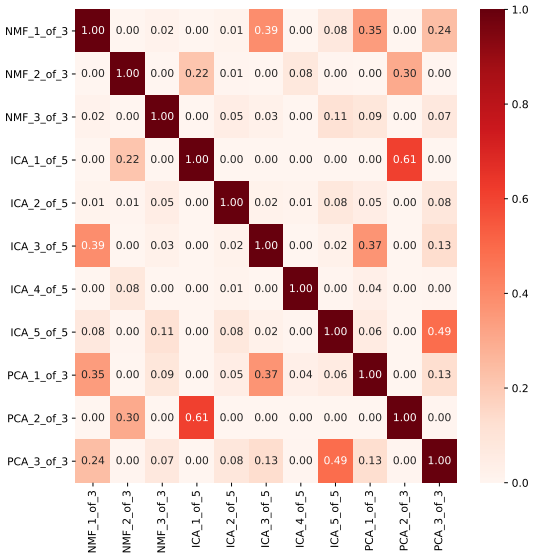


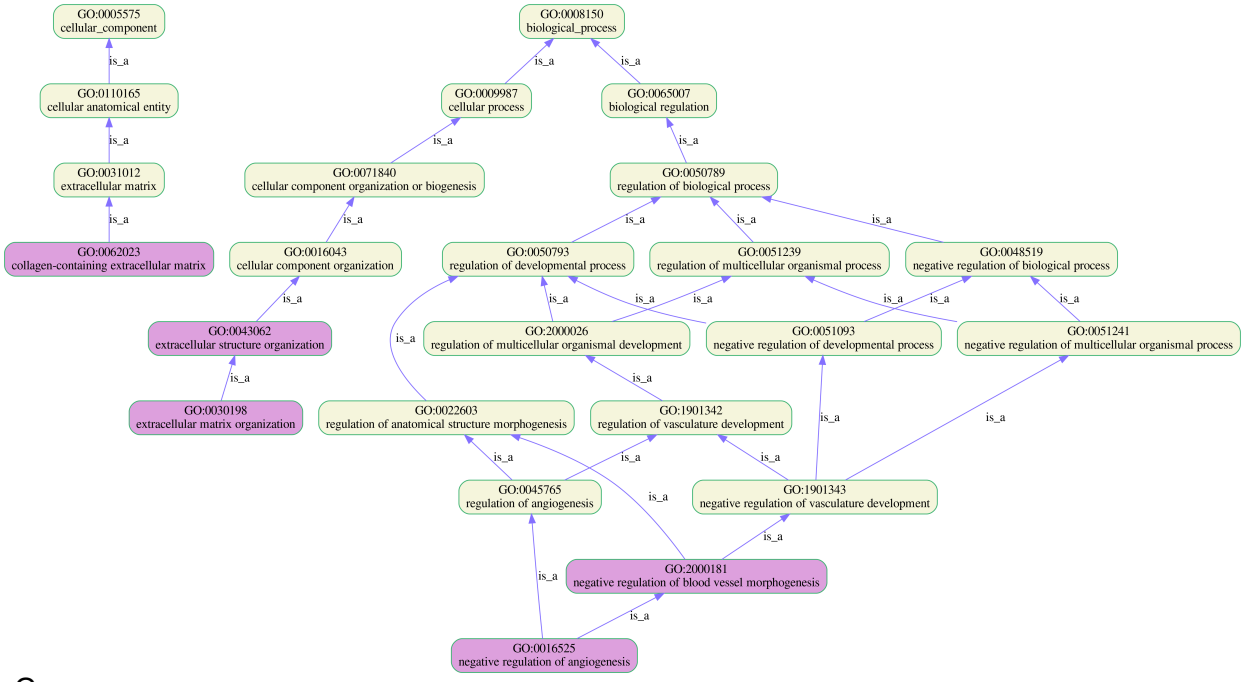
Figure 8: Heatmap of Jaccard similarities between the candidate genes identified by the 11 components.

3.3 Metagenes highlight genes which are enriched for biological processes

GEA against the GO results in, for each of the 11 metagenes, a table of GO terms with a list of the candidate genes which have an annotated association to that term. See appendix 6.5 to access the full table. To take full advantage of the hierarchical nature of the GO, results are presented as lineage maps in figures 9 to 14. (The large graphic for ICA-4 is in the appendix, figure 23). The contributing candidate genes for all significantly enriched terms are shown beneath each figure. High-level terms at depth less than 3 were removed, since these are generic (e.g. “regulation of biological

process") and so uninteresting. One component – ICA-4 – was problematic in that 49 GO terms (with depth ≥ 3) were identified as significant. For this component it was necessary to limit the graphic to the 12 terms having the largest number of associated candidate genes.

The four components NMF-3, ICA-1, ICA-2 and PCA-2 yielded no significant enriched terms (for FDR ≤ 0.01).



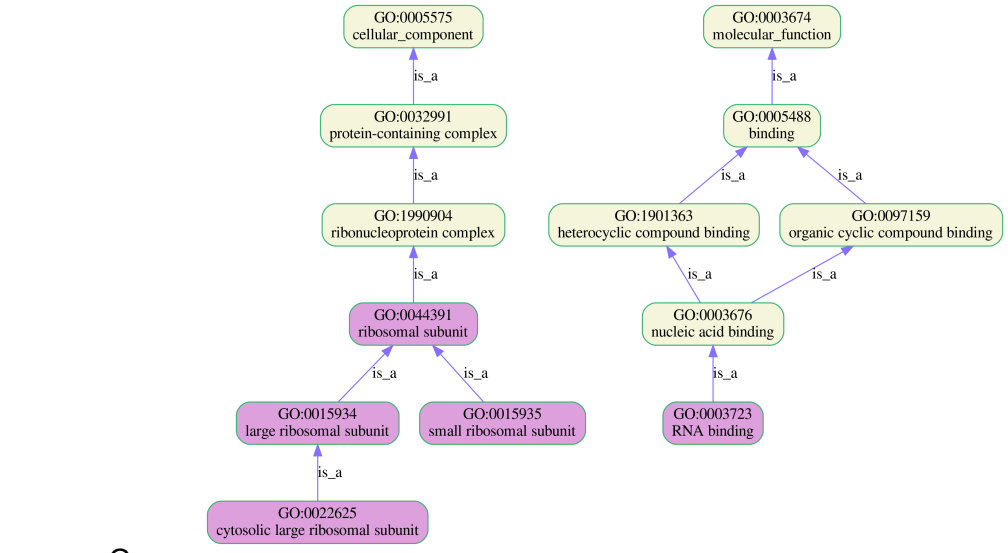
Genes: COL16A1 ECM1 ISM1 EFEMP1 ADAMTSL1 COL3A1 ADAMTS2 FGL2 PCOLCE COMP MMP19 CCDC80 THBS1 NID2 MMP9 THBS2 HTRA1 COL6A2 LAMA4 TNC

Component NMF-1

Figure 9: Lineage maps of enriched Gene Ongology (GO) terms for components NMF-1 In these diagrams, enriched terms are coloured purple, while there ancestors in the ontology are yellow.

Considering the biological significance of each of the 11 extracted metagenes, with reference to the GO enrichment results:

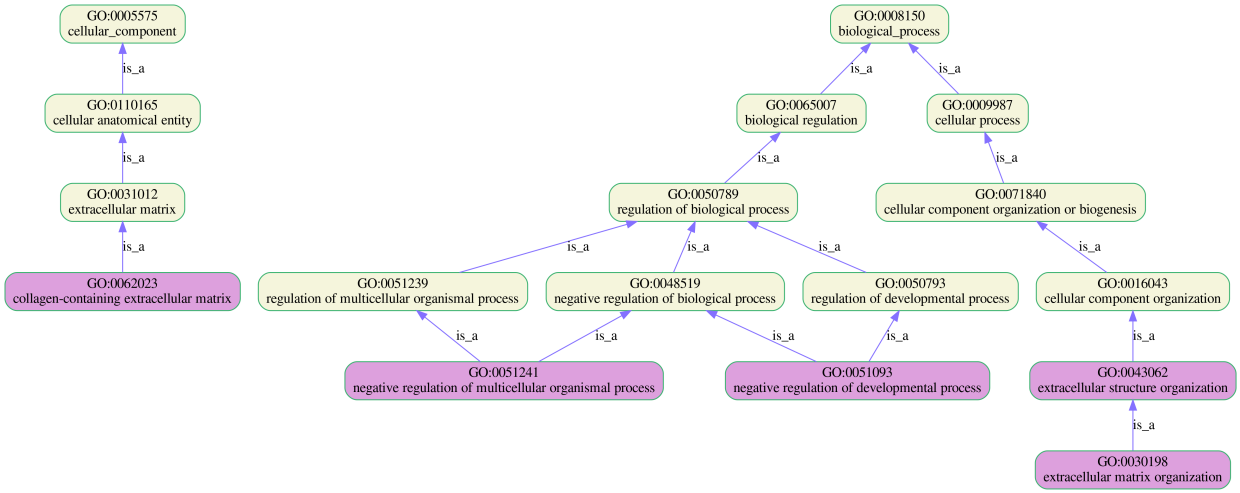
NMF-1 : there is mention of extra-cellular matrix (ECM) related terms, as in ICA-3 and NMF-1 below. However, there are also terms relating to the regulation of angiogenesis. It is known that tumours have a need for increased blood supply, and that expression level of factors promoting angigenesis are associated with aggresiveness of tumour growth [37]. We have already seen (section 1.2.2, [12])



Genes: RPL28 RPL32 RPLP0 RPL13A RPS14 RPS4X RPS18 RPS27 EEF1A1 RPL37 RPS20 RPL18A RPL8 RPL15 RPL13 RPS11

Component NMF 2

Figure 10: Lineage maps of enriched GO terms for components NMF-2. (NMF-3 produced no significant enrichment results)



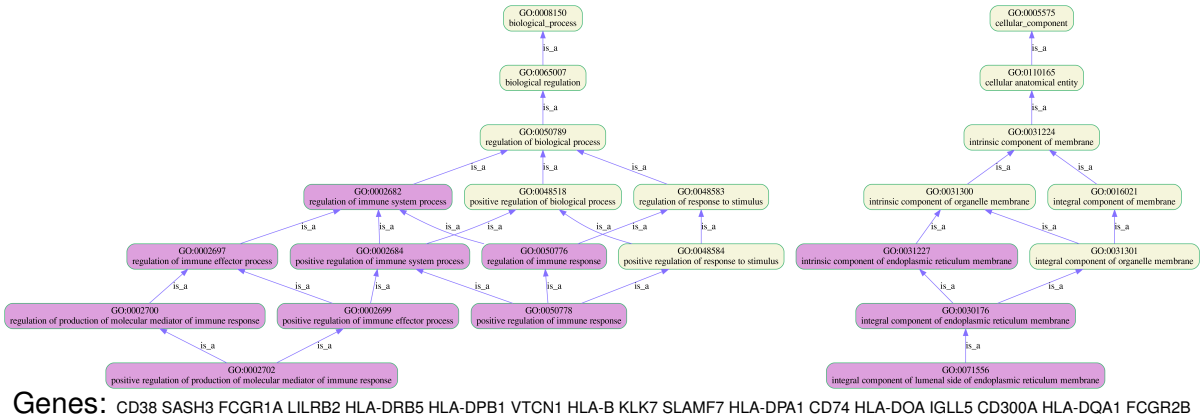
Genes: COL16A1 ECM1 EFEMP1 COL3A1 ADAMTS2 ADAMTS16 PCOLCE MMP19 THBS1 PTGER3 COL6A6 COL26A1 PRICKLE1 OGN RFLNA THBS2 ADAMTS7 COL6A2 ENPP1 LAMA4 ISM1 ADAMTSL1 MATN3 CCDC80 COMP RFLNB NID2 TMEM176A MMP9 TMEM176B ADAMTS5 DPT HTRA1 TNC

Component ICA-3

Figure 11: Lineage maps of enriched GO terms for components ICA-3

there is evidence for distinct subtypes of OC distinguished by angiogenesis related genes, and that these subtypes have been found to inform clinical outcome [12]. Thus, this component might contain useful prognostic value.

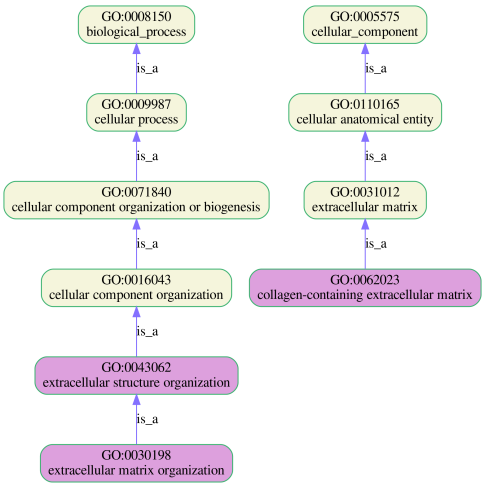
NMF-2 : here we see enrichment of genes relating to the ribosomal subunit and the processes of RNA binding, implying perhaps a link with assembly of the riboso-



Genes: CD38 SASH3 FCGR1A LILRB2 HLA-DRB5 HLA-DPB1 VTCN1 HLA-B KLK7 SLAMF7 HLA-DPA1 CD74 HLA-DOA IGLL5 CD300A HLA-DQA1 FCGR2B HLA-DQB2 SLAMF8 HLA-DRA HLA-DMB HLA-DQB1 MFAP4 DOCK8 HLA-DQA2 C3 RSAD2 HAVCR2

Component ICA-5

Figure 12: Lineage maps of enriched GO terms for components ICA-5



Genes: COL16A1 COL26A1 OGN ADAMTSL1 COL3A1 ADAMTS2 ADAMTS16 PCOLCE CCDC80 COMP COL6A2 ADAMTS9 ADAMTS7 NID2 TNC COL14A1

Component PCA-1

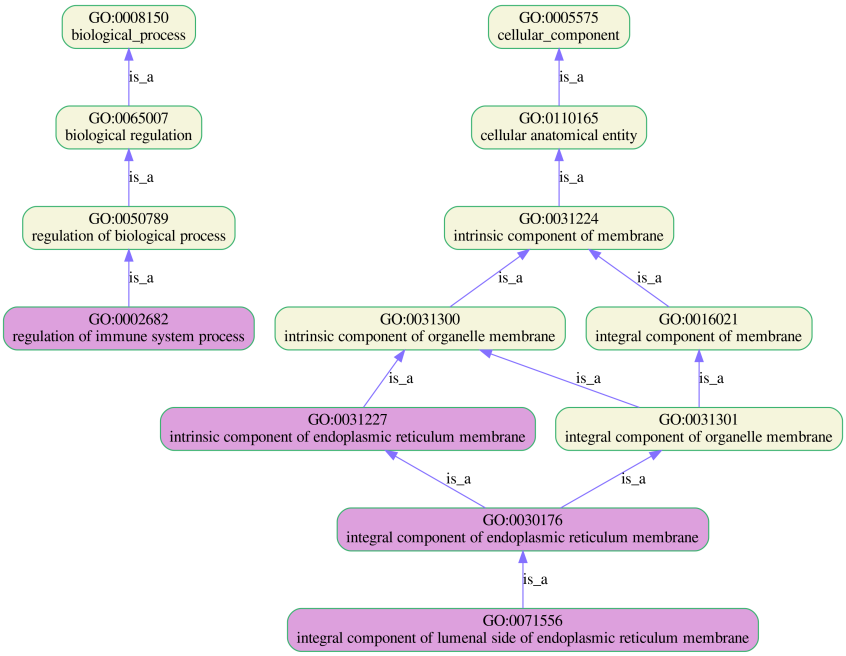
Figure 13: Lineage maps of enriched GO terms for component PCA-1

mal RNA-protein complex. Ribosomes are known to have a role in carcinogenesis, by dysregulation the RNA $\rightarrow$ protein translation, or mutations in ribosomal subunits impacting on cellular metabolism [38].

NMF-3, ICA-1 and ICA-2 : no significant biological enrichment found.

ICA-3 : this component relates to processes of multicellular / extracellular organisation, and the ECM. It is therefore similar to NMF-1 above.

ICA-4 : many processes are highlighted by the component. These include processes relating to the ribosomal subunit (as NMF-2), additionally membrane pro-



Genes: SASH3 FCGR1A LILRB2 HLA-DRB5 HLA-DPB1 VTCN1 HLA-B KLK7 SLAMF7 HLA-DPA1 TMEM176A HLA-DOA CD74 IGLL5 CD300A FCGR2B HLA-DQA1 HLA-DQB2 SLAMF8 TMEM176B HLA-DRA HLA-DQB1 VSIR HLA-DQA2 C3 HAVCR2

Component PCA-3

Figure 14: Lineage maps of enriched GO terms for component PCA-3. (Component PCA-2 produced no significant enrichment results)

teins in the respiratory complex, mitochondrial and the NADH dehydrogenase complex. Alsp processes of organonitrogen biosynthesis are also highlighted, purine nucleotide biosynthesis, RNA binding, proton membrane transport.

ICA-5 : processes relating to regulation of immune response are enriched in this component, featuring genes from the major histocompatibility complex (MHC) group, in particular the HLA- genes, allowing the immune system to recognise self from non-self. GO terms relating to the endoplasmic reticulum (ER) are also highlighted. It may be that this component is mainly sensitive to the immunohistochemical signature of patients, and therefore not of clinical interest [2].

PCA-1 : this is related to the cellular structures and processes of the ECM, that is the proteins such as collagens which mediate the three-dimensional organisation of cells in a tissues. ECM molecular composition will vary substantially between tissue types, but is also known to play a part in many disease processes [39]. Thus, this metagene may simply reflect heterogeneity of tissues in the biopsied

[2] This turns out to be an incorrect conclusion: note presence of gene CD38 and see the Discussion!

sample, or may have some deeper disease related significance.

PCA-2 : no significant biological enrichment found.

PCA-3 : this component has some similarity with ICA-5, in that it refers to regulation of immune processes (HLA- genes) and ER membrane. However, chemotaxis (cell movement) is also highlighted.

3.4 Association of unsupervised gene expression patterns with patient survival is inconclusive

Kaplan-Meier overall survival (OS) plots stratified by each metasample component are shown for the for the AOCS dataset in figure 15 and for the TCGA dataset in figure 24 (appendix). Plots for progression free survival (PFS) on AOCS are also in the appendix, figure 25. All plots show 95% confidence intervals and hazard ratio (HR) with associated p-value.

All of these results are summarised in figure 16, which brings together the three sets of results – TCGA (OS), AOCS (OS) and AOCS (PFS) – with respect to the 11 metasample components. $\log_2$ HR is used in order that the *sense* of the survival impact can be readily appreciated. If a metasample component has a robust correlation with survival, then we expect the p-value for all three sets of results to show significance *and* for the effects to have the same sense – be in the same direction. None of the 11 components pass this test.

The component with the largest observed effect on survival is PCA-1, showing a reasonably consistent hazard ratio of around 1.3 ($2^{0.4}$) across the three experiments. This has $p < 0.05$ in the larger TCGA dataset, but not in the smaller AOCS dataset.

3.5 Metasamples correlate with some genomic features

From the grid of scatter plots shown in figure 17 it can be seen there are some significant correlations at play. These are summarised and analysed in the next section.

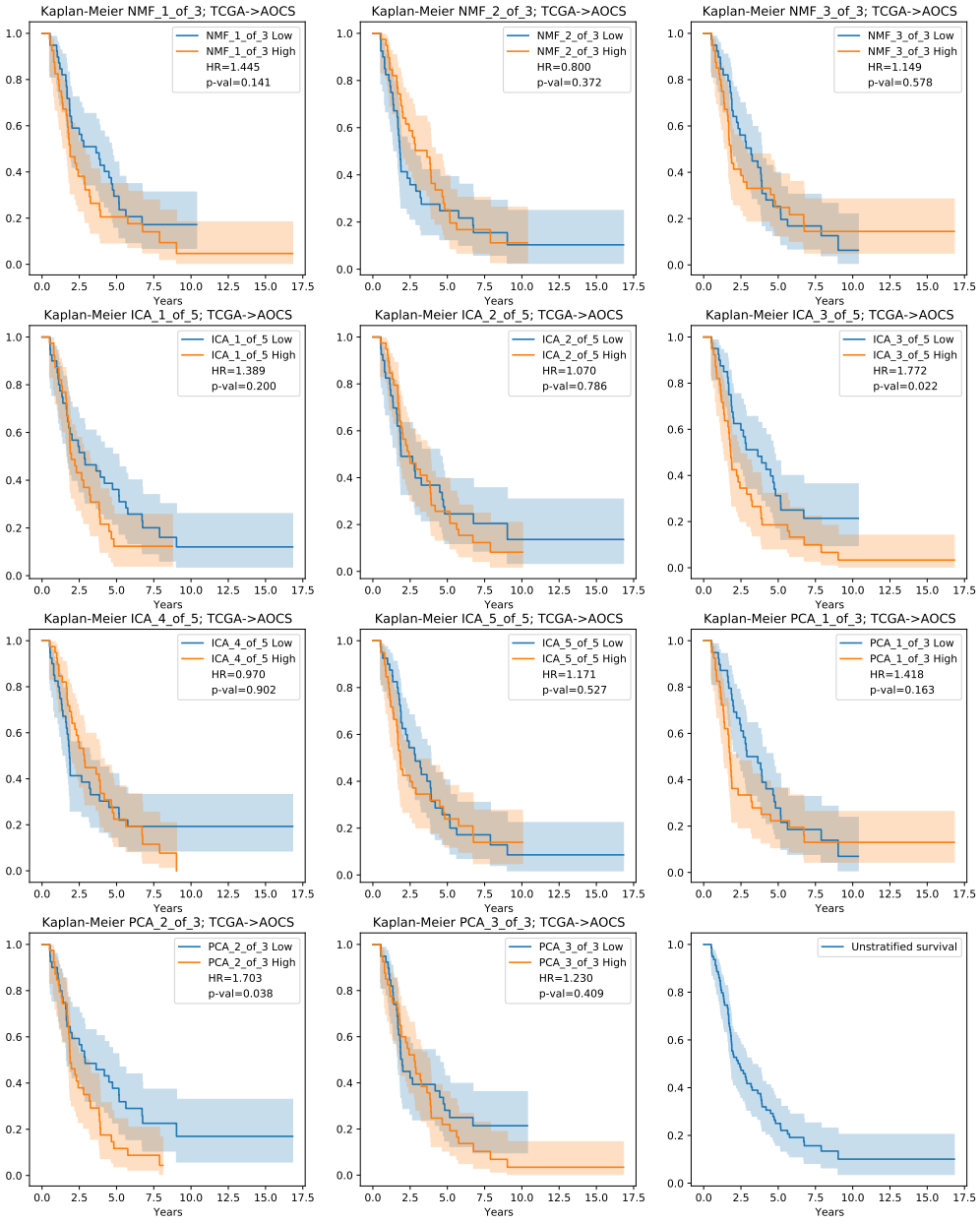


Figure 15: Kaplan-Meier plots for TCGA → AOCS for overall survival (OS).

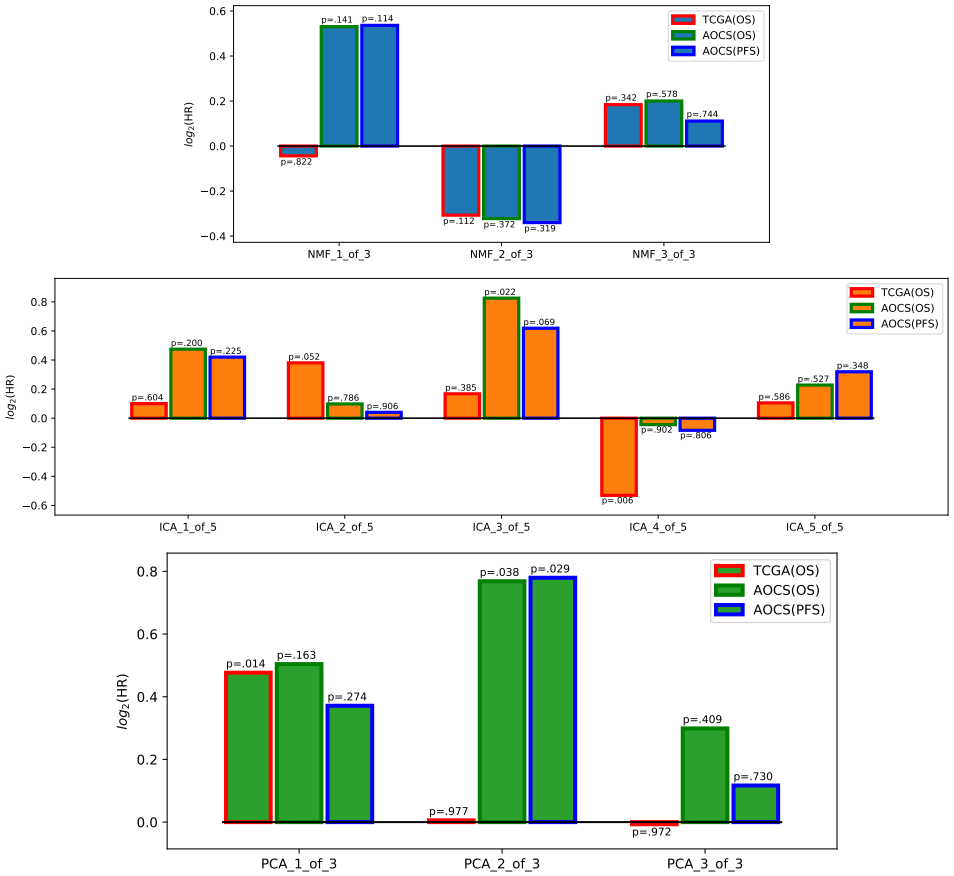


Figure 16: Visual summary of survival analysis as applied to TCGA(OS), AOCS(OS) and AOCS(PFS). Plots are divided by factorization method. Bar heights show $\log_2(\text{HR})$ with p-value also shown.

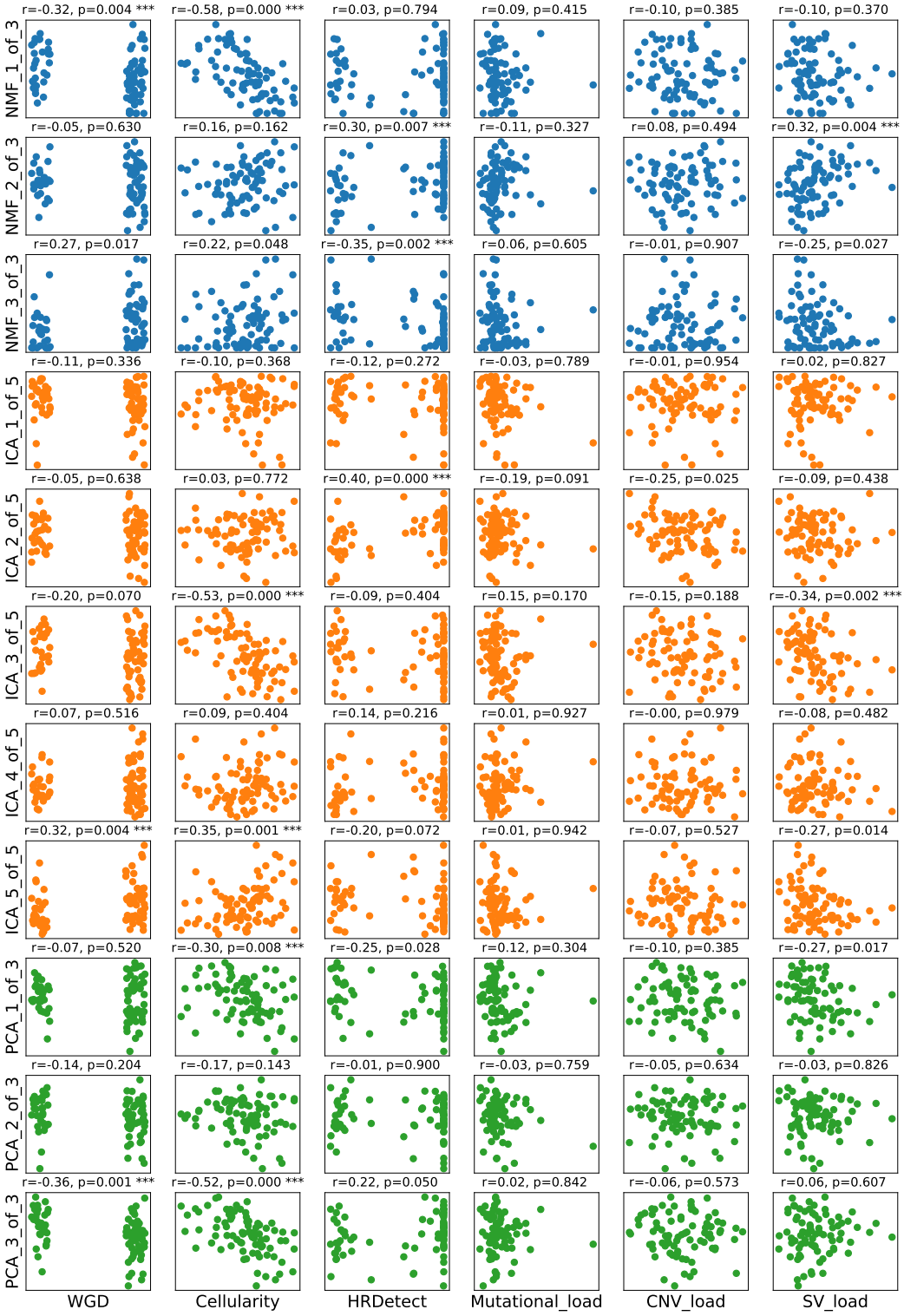


Figure 17: Grid of scatter plots to visualise the correlation between metasamples and genomic features. In the case of WGD which has binary values (0, 1), jitter is applied for visual effect only. Above each plot is shown a correlation coefficient r and associated p-value, highlighted with '***' where $p < 0.01$. In the case of WGD , the Point-Biserial correlation is used; for all others Pearson's r is used.

3.6 Integrative analysis of results

Table 1 summarises the key results of all the forgoing analysis.

Metagene	S'val	WGD	Cell.	HRD	Mut.L.	CN.L.	SV.L.	Enriched GO terms	Similar to
NMF-1	.	Y-	Y-	.	.	.	.	ECM, Reg. angiogenesis	ICA-3, PCA-1
NMF-2	.	.	.	Y+	.	.	Y+	Ribo. subunit, RNA-binding	.
NMF-3	.	.	.	Y-	.	.	.	.	.
ICA-1	.	.	.	.	.	.	.	.	PCA-2
ICA-2	.	.	.	.	.	.	.	.	.
ICA-3	.	.	Y-	.	.	.	Y-	ECM, multicellular/extracellular organisation	NMF-1, PCA-1
ICA-4	.	.	.	.	.	.	.	Ribo. subunit, Mitochondria, NADH dehydrogenase, purine nucleotide synthesis, RNA binding, Proton membrane transport	.
ICA-5	.	Y+	Y+	.	.	.	.	ER, Immune response, MHC	PCA-3
PCA-1	?+	.	Y-	.	.	.	.	Extra-cellular matrix / organisation,	NMF-1, ICA-3
PCA-2	.	.	.	.	.	.	.	.	ICA-1
PCA-3	.	Y-	Y-	.	.	.	.	ER, Immune response, MHC, chemotaxis	ICA-5

Table 1: For each metagenes a 'Y' indicates that a significant correlation was observed with: Survival, WGD, HRD, Mutational load, CNV load and SV load; '?' indicates a questionable link; +ve / -ve correlation is indicated. The next column gives a terse summary of the GEA results. The final column lists the metagenes which are apparently similar – having a Jaccard similarity of ≥ 0.35.

From this summary table we can make some interesting observations:

ECM relationship to Cellularity? NMF-1, ICA-3 and PCA-1 each have an ECM association and we have noted this could be related to tissue type heterogeneity. These three components (one from each of the methods) are also observed to be correlated with Cellularity. The sense (+/-) of the correlation is meaningful for NMF (not so for ICA or PCA) and we see that ECM and thence tissue heterogeneity is negatively correlated with cellularity. This seems reasonable, in that a low Cellularity implies that a greater proportion of surrounding non-tumour tissue is included in the sample, so increasing heterogeneity. That these three components turn out to be similar by the Jaccard index is as would be expected given their functional similarity.

No consistent link with survival. The only component showing even questionable correlation with patient survival is PCA-1. This is associated with ECM terms,

which may simply be due to Cellularity. Component ICA-4 highlights the greatest number of biological processes – indeed the GO lineage map 23 needed to be pruned to fit on a page. And yet this component shows no correlation with any disease related factors, neither survival nor genomic features. It may be that this component is simply picking up on the normal metabolic activity of the cells.

Is HRD linked with ribosomal processing? HRD is correlated with NMF-2 and NMF-3. The first of these is associated with the ribonucleoprotein complex while the second shows no biological enrichment.

Technical batch effects? There are four components (NMF-3, ICA-2, ICA-4 and PCA-2) showing no enrichment of GO terms. These components also show almost no correlation with disease related factors (the exception being NMF-3 linking with HRD). With that exception aside, we could conclude that these components have little biological meaning, and are instead picking up technical effects. I do not know how the TCGA dataset was assembled; perhaps there are multiple batches within the set?

WGD and Cellularity have curious common influences. The three components NMF-1, ICA-5 and PCA-3 (one from each method) correlate with both WGD and Cellularity, and in the same sense. This might suggest simply that WGD and Cellularity are themselves positively correlated, but this is not so: $r = -0.14, p = 0.201$ for Point-Biserial correlation (data not shown). Neither is there any close similarity between the three components by Jaccard, although ICA-5 and PCA-3 are both linked to immune response and MHC. Is there some curious non-linear interaction?

Mutational load CNV load are not captured by these components. This is somewhat surprising, perhaps the link would emerge with more extracted components.

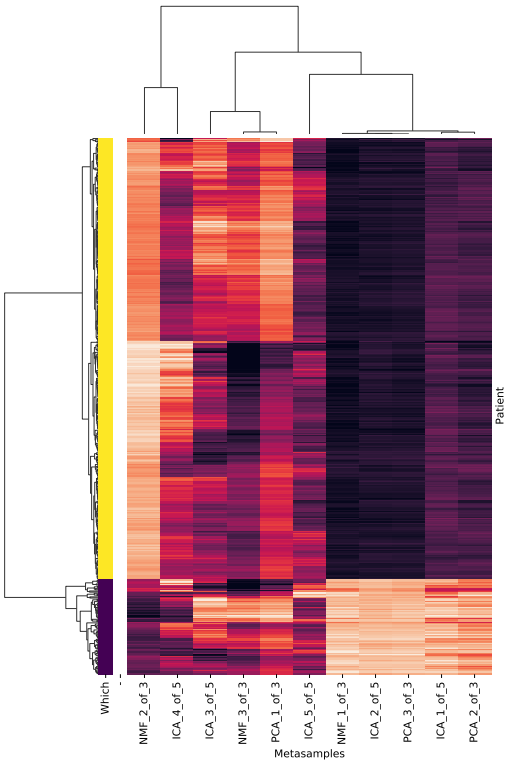


Figure 18: Heatmap of the metasamples matrix for the combined AOCS + TCGA dataset. The 'Which' field indicates the source dataset; purple: AOCS, yellow:TCGA.

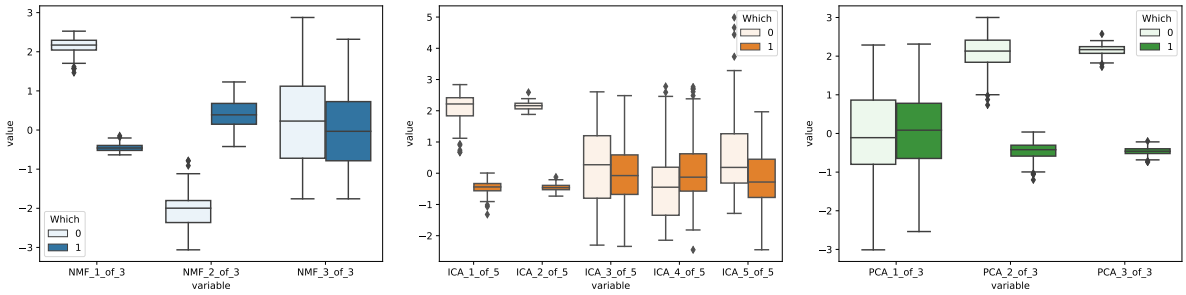


Figure 19: Boxplots of metasamples from the combined AOCS + TCGA datasets, showing relationship between metasamples dataset origin; Which=0 or 1 for AOCS or TCGA respectively.

3.7 Matrix factorization discovers and effectively removes batch effects

Here are presented results of applying the three MF methods to the *combined* AOCS and TCGA datasets. From the clustered heatmap of figure 18 it is clear that the five components clustered to the left are highly sensitive to (100% predictive of) the batch,

while in the other components batch is hardly discernable. This is confirmed by the boxplots in figure 19. Within the larger TCGA dataset there is clear clustering into two groups with a roughly 5:6 ratio split. Does this perhaps indicate that the TCGA set is collected from two separate sources? Is the difference technical or are there substantially different cohorts of patients?

Table 2 summarises observations on this combined dataset. It can be seen that, with a few exceptions, a component is sensitive *either* to Batch *or* to some biological process, but seldom both. This is a useful property.

Metagene	Batch	WGD	Cell.	HRD	Mut.L.	CN.L.	SV.L.	Enrichment?
NMF-1	Y	.	.	.	.	.	.	Y
NMF-2	Y	Y	Y	.	.		.	Y
NMF-3	.	Y	Y	.	.	.	.	Y
ICA-1	Y	.	.	.	.	.	.	.
ICA-2	Y	.	.	.	.	.	.	.
ICA-3	.	.	.	.	.		.	Y
ICA-4	.	Y	Y	.	.	.	.	Y
ICA-5	.	.	.	.	.	.	.	Y
PCA-1	.	Y	Y	.	.	.	.	Y
PCA-2	Y	.	.	.	.	.	.	.
PCA-3	Y	.	.	.	.	.	.	.

Table 2: In a similar format as table 1, this summarises results of in respect of the *combined AOCS and TCGA* datasets. The key point here is that a known batch effect has been introduced, reflected in the 'Batch' column. The 'Enrichment' column simply indicates whether or not biological enrichment was found (the details have not been analysed). Note that since these metagenes have been re-extracted from different data, they are distinct from those analysed in the previous section

4 Discussion

Addressing each research questions in turn:

What underlying biological processes are at play in HGSOC as seen through the patterns of gene expression?

Biological processes identified by GEA have been set out in section 2.6. Processes relating to the ECM and extracellular organisation may reflect cell type heterogeneity – perhaps within the tumour its self, or due to sampling of adjacent normal tissue. Processes of negative regulation of angiogenesis could reflect tumour growth and therefore be of prognostic value.

These observations are based on GEA on the GO only. Possibly more insights could have been gained by GEA against KEGG[3] Pathways. (This was begun but had to be dropped due to shortage of time).

How do the patterns of gene expression uncovered relate to genome level features, such as those marking genome instability?

The integrative analysis set out in section 3.6 suggests that Cellularity is correlated to the processes of the the ECM, reflecting heterogeneity of tissue type. Of course, Cellularity is not really a biological characteristic – it is a technical artefact of the tumour biopsy process, and so we might conclude that signatures relating to ECM might also be considered as technical artefacts. We found that whole genome doubling (WGD) and Cellularity correlate together with three components even though they are not themselves directly correlated – something which warrants further investigation.

Weak evidence of a link between homologous repair deficiency (HRD) and RNA-binding was found. Being the site of protein synthesis, ribosomes will be associated with most of what goes on in a the cell. One paper suggests a link : “a comprehensive

[3] Kyoto Encyclopedia of Genes and Genomes

review of the literature reveals a role for conventional DNA repair proteins in ribosome biogenesis, and conversely, ribosome biogenesis proteins in DNA repair" [40]. Seek and ye shall find!

A note of caution: these results are based on testing $6 \times 11 = 66$ hypotheses at a significance level of < 0.01, so there will likely be at least one false discovery. A significance level of 0.001 would have been more appropriate.

Expression signature correlations with mutational load and CNV load have not been detected.

Do MF methods yield factors which are predictive of patient survival?

We found no convincing evidence of this, for an individual factor (metasample). Some components were predictive in one or other of the two datasets, but not both, which is unconvincing. The factor for which GEA indicated regulation of angiogenesis (NMF-1), and thence might have prognostic value, performed poorly in survival analysis. So overall a negative result.

Is there a signal for survival to be found in these data? We have noted that according to [17], the single gene CD38 is highly predictive of survival in EOC. Does this hold for the TCGA and AOCS datasets we are working with? This is easy to check, running survival analysis stratified by a hand-crafted metagene, setting the CD38 element to 1, all others to 0. This was applied as in section 2.9 to TCGA (OS), AOCS (OS) and AOCS (PFS) – see figure 20. The effects are in a consistent direction across the three tests, with high CD38 expression predicting favourable survival. This is highly significant in the TCGA dataset and $p < 0.05$ in the smaller AOCS. So there is a signal, but it did not emerge unbidden from matrix factorization. (CD38 is in fact highlighted by component ICA-5, see figure 12, but must be diluted by other genes).

Several other specific genes have been noted in section 1.2; it would be interesting to test these also.

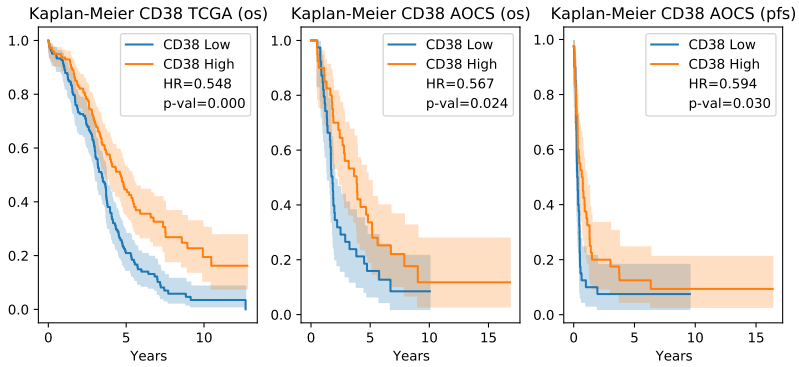


Figure 20: Survival analysis plots stratified by above/below median expression of CD38 only.

How should MF methods be applied to achieve robust results on datasets of limited size?

One valuable lesson from this work is the importance of accounting for sampling error when considering the stability of factorizations and thence the selection of factorization rank. Sevaral authors comment on the variability of NMF and ICA factorization due to algorithm initialisation [26, 24, 31], but most neglect sampling error which we have shown to be far more significant. Cantini *et al* [41] mention bootstrapping, in reference to the BIODICA package (for ICA factorization), but it is unclear whether this is applied to the input data to model sampling error. In the initial stages of this project, experiments were conducted with BIODICA on the AOCS dataset. The package recommend ranks of 14, 19 or 30. Yet we have seen that such high ranks on the small N=80 dataset leads to highly unstable factorizations when bootstrap modelling of sampling error is included. Suspicion of this result lead to me roll my own implementation based on bootstrap resampling as described, resulting in much lower ranks (fewer metagenes) being selected than would otherwise be the case if only algorithmic randomness were considered.

Does MF detect and remove technical batch effects?

Yes. We have been able to confirm the observations of [42, 36, 29] that detection and removal of technical batch effects naturally emerge from MF methods. That is, the technical and biological variation largely (not completely) separate into different com-

ponents. This is a useful property, particularly as the growing availability of transcriptomics datasets from resources such as TCGA and GEO allows for the simultaneous analysis of dozens of distinct datasets with likely different technical provenance, for which hand crafted normalisation would be infeasible.

Which MF method is best suited to transcripomics analysis?

The original technical focus of this dissertation was ICA and NMF since these are know to be able to disentangle variation in data with complex structure, not multivariate normally distributed. PCA was added as a baseline dimensionality reduction method, just for comparison. What would we have been missed if only PCA had been used? Reviewing the integrative analysis of results in table 1 we see that all three methods picked up on WGD and cellularity. PCA included the two components having the largest observed effect on overall survival. NMF and ICA both showed sensitivity to SV load. NMF further added sensitivity to HRD and identified genes relating to angiogenesis with possible pathological significance.

So it's a mixed picture. Of course, had more components been selected from each method then likely more would have been found, but at the expense of robustness and thence false discovery, as we have discussed. It seems we must concur with the conclusion of Way *et al*[31] that no one method excels, each brings individual value. At least these three methods are simple and linear (unlike the VAE for example), and so can be used together easily.

4.1 Further work

This work has thrown up several possible lines of further study:

1. **Algorithm hyper-parameters**. NMF and ICA algorithms have a number of hyper-parameters, only a few of which were explored in this work. For example, NMF has parameters to encourage sparsity through L1 regularization. This should result in many (most?) metagene elements reducing to zero, and might offer a

more robust means of selecting candidate genes to feed into GEA – replacing the current arbitrary 3 SD from the mean rule.

2. **Selecting components from several factorization ranks**. In the current work, a single rank was selected for each method – 3, 5 an 3 for NMF, ICA and PCA respectively. Unlike PCA, components extracted by NMF and ICA do not 'nest' with rank; that is adding rank in general yields a new set of components. Thus one might, for NMF and ICA, perform factorizations at $k = 2, 3, 4, 5$, say, yielding $2 \times 15 = 30$ potential components in total. Jaccard similarity could be used to identify that subset of components which had the least overlap in detected genes. In this way, a larger number of components could be obtained without use of high factorization ranks which we have seen to be unstable. In fact this is similar to the approach of [31] mentioned in section 1.3.5. It will be important to take a principled approach to the setting of the FDR theshold when applying GEA, to account for the increased number of hypothesis being tested.

3. **Cross-dataset factorization stability**. We went to considerable effort – through bootstrap sampling and cluster analysis – to select factorization ranks which we hoped would be robust and generalise well to other datasets. A way to confirm this robustness would be to perform the factorization / clustering pipeline on *both* datasets separately, yielding two sets of metagenes. In a perfect world, these metagenes would pair up identically (in-so-far as the two datasets had identical technical characteristics, drawn from identical population of patients). The Jaccard similarity heatmap could be used to verify this, and reject components which showed low cross-dataset similarity. This proposal chimes to some extent with the use of Reciprocally Best Hit graphs in [25] reviewed in section 1.3.5.

4. **Systematic comparison with published gene expression patterns**. In analysing the meaning of each metagene above, some tentative links were made with the research literature, by informal searching. A more systematic and ideally automated approach is required. The Geo Profiles database at NCBI, or the Expression Atlas at EMBL-BI might be possible starting points.

5 Conclusions

This work has investigated the application of unsupervised matrix factorization (MF) methods – specifically non-negative matrix factorization (NMF), independent component analysis (ICA) and principal component analysis (PCA) – to gene expression analysis in the context of high grade serous ovarian cancer (HGSOC). These methods aim to find metagenes which capture biologically or technically significant variation in the gene expression signal. Metagenes were extracted from a N=374 patient dataset. Key findings are:

- Biological processes relating to extracellular matrix (ECM), ribosomal subunits, RNA binding, angiogenesis and immune response have been identified and may have significance to HGSOC.
- Metagenes have been found with possible correlation to genome level features of whole chomosome doubling, homologous repair deficiency and structural variant load. However, these findings have marginal statistical significance given the small size (N=80) of the available dataset.
- A metagene predictive of patient survival did *not* emerge from these unsupervised methods.
- However, expression of a single gene, CD38, identified from the literature, was confirmed to have substantial impact on patient survival.
- When applying MF methods, particularly to small datasets, it is important to consider the impact of sampling error when deciding the number of metagenes to extract. Bootstrap resampling and cluster analysis is an effective approach.
- MF methods have been found effective in separating technical variation (batch effect) from biological variation.
- NMF, ICA and PCA each find mostly different metagenes. No clear evidence has been found to support one over the others.

Acknowledgements

I am most grateful to Dr. Ailith Ewing for proposing this project and making available the datasets. Ailith provided excellent, insightful supervision throughout and gave valuable feedback on an early draft of this dissertation. I'd like to thank my employer, Canon Medical Research Europe Ltd., for allowing me the flexibility and study leave to undertake this MSc. Thank you to my family for excusing me from cooking and washing up duties during the more stressful phases of MSc study and dissertation writing!

References

[1] M. Kossaï, A. Leary, J. Y. Scoazec, and C. Genestie, "Ovarian Cancer: A Heterogeneous Disease," *Pathobiology*, vol. 85, no. 1-2, pp. 41–49, 2018.

[2] M. A. Lisio, L. Fu, A. Goyeneche, Z. H. Gao, and C. Telleria, "High-grade serous ovarian cancer: Basic sciences, clinical and therapeutic standpoints," *International Journal of Molecular Sciences*, vol. 20, no. 4, 2019.

[3] A. M. Patch, E. L. Christie, D. Etemadmoghadam, D. W. Garsed, J. George, and E. Al., "Whole-genome characterization of chemoresistant ovarian cancer," *Nature*, vol. 521, pp. 489–494, may 2015.

[4] M. Pradhan, B. Risberg, C. G. Tropé, M. van de Rijn, C. B. Gilks, and C. H. Lee, "Gross genomic alterations and gene expression profiles of high-grade serous carcinoma of the ovary with and without BRCA1 inactivation," *BMC Cancer*, vol. 10, no. 1, pp. 1–8, 2010.

[5] A. Ewing, A. Meynert, M. Churchman, G. R. Grimes, R. L. Hollis, C. S. Herrington, T. Rye, C. Bartos, I. Croy, M. Ferguson, T. Mcgoldrick, N. Mcphail, N. Siddiqui, and S. Dowson, "Structural variants at the BRCA1/2 loci are a common source of homologous repair deficiency in high grade serous ovarian carcinoma," pp. 1–37, 2020.

[6] J. Lu, D. Wu, C. Li, M. Zhou, and D. Hao, "Correlation between gene expression and mutator phenotype predicts homologous recombination deficiency and outcome in ovarian cancer," *Journal of Molecular Medicine*, vol. 92, no. 11, pp. 1159–1168, 2014.

[7] Z. He, J. Zhang, X. Yuan, Z. Liu, B. Liu, S. Tuo, and Y. Liu, “Network based stratification of major cancers by integrating somatic mutation and gene expression data,” *PLoS ONE*, vol. 12, no. 5, pp. 1–12, 2017.

[8] K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, and X. et al Huang, H; Zhou, “Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin,” *Cell*, vol. 158, no. 4, pp. 929–944, 2014.

[9] M. Schaner and Others, “Gene Expression Patterns in Ovarian Carcinomas Marci,” *Molecular Biology of the Cell*, vol. 14, no. December, pp. 5069 –5081, 2003.

[10] C. Wang, S. M. Armasu, K. R. Kalli, M. J. Maurer, E. P. Heinzen, G. L. Keeney, W. A. Cliby, A. L. Oberg, S. H. Kaufmann, and E. L. Goode, “Pooled clustering of high-grade serous ovarian cancer gene expression leads to novel consensus subtypes associated with survival and surgical outcomes,” *Clinical Cancer Research*, vol. 23, no. 15, pp. 4077–4085, 2017.

[11] I. Espinosa, L. Catasus, B. Canet, E. D’Angelo, J. Mũoz, and J. Prat, “Gene expression analysis identifies two groups of ovarian high-grade serous carcinomas with different prognosis,” *Modern Pathology*, vol. 24, no. 6, pp. 846–854, 2011.

[12] K. Glass, J. Quackenbush, D. Spentzos, B. Haibe-Kains, and G. C. Yuan, “A network model for angiogenesis in ovarian cancer,” *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–17, 2015.

[13] A. Talhouk, J. George, C. Wang, T. Budden, T. Z. Tan, S. Derek, S. Kommoss, H. S. Leong, S. Chen, and M. P. Intermaggio, “Development and validation of the gene-expression Predictor of high-grade-serous Ovarian carcinoma molecular subTYPE (PrOTYPE),” 2020.

[14] R. G. Verhaak, P. Tamayo, J. Y. Yang, D. Hubbard, H. Zhang, and E. al., “Prognostically relevant gene signatures of high-grade serous ovarian carcinoma,” *Journal of Clinical Investigation*, vol. 123, pp. 517–525, jan 2013.

[15] F. Mairinger, A. Bankfalvi, K. W. Schmid, E. Mairinger, P. Mach, R. F. Walter, S. Borchert, S. Kasimir-Bauer, R. Kimmig, and P. Buderath, “Digital immune-related gene expression signatures in high-grade serous ovarian carcinoma: Developing prediction models for platinum response,” *Cancer Management and Research*, vol. 11, pp. 9571–9583, 2019.

[16] T. Fekete, E. Rásõ, I. Pete, B. Tegze, I. Liko, G. Munkácsy, N. Sipos, J. Rigõ, and B. Györffy, “Meta-analysis of gene expression profiles associated with histological classification and

survival in 829 ovarian cancer samples," *International Journal of Cancer*, vol. 131, no. 1, pp. 95–105, 2012.

[17] Y. Zhu, Z. Zhang, Z. Jiang, Y. Liu, and J. Zhou, "CD38 Predicts Favorable Prognosis by Enhancing Immune Infiltration and Antitumor Immunity in the Epithelial Ovarian Cancer Microenvironment," *Frontiers in Genetics*, vol. 11, no. April, pp. 1–13, 2020.

[18] G. Au-Yeung, P. M. Webb, A. Defazio, S. Fereday, M. Bressel, and L. Mileshkin, "Impact of obesity on chemotherapy dosing for women with advanced stage serous ovarian cancer in the Australian Ovarian Cancer Study (AOCS)," *Gynecologic Oncology*, vol. 133, no. 1, pp. 16–22, 2014.

[19] M. A. Cuello, S. Kato, and F. Liberona, "The impact on high-grade serous ovarian cancer of obesity and lipid metabolism-related gene expression patterns: the underestimated driving force affecting prognosis," *Journal of Cellular and Molecular Medicine*, vol. 22, no. 3, pp. 1805–1815, 2018.

[20] K. E. Hew, A. Bakhru, E. Harrison, M. O. Turan, R. MacDonald, D. D. Im, and N. B. Rosenshein, "The Effect of Obesity on the Time to Recurrence in Ovarian Cancer: A Retrospective Study," *Clinical Ovarian and Other Gynecologic Cancer*, vol. 6, no. 1-2, pp. 31–35, 2013.

[21] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 12, pp. 4164–4169, 2004.

[22] P. Comon, "Independent component analysis, A new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[23] G. L. Stein-O'Brien, R. Arora, A. C. Culhane, A. V. Favorov, L. X. Garmire, C. S. Greene, L. A. Goff, Y. Li, A. Ngom, M. F. Ochs, Y. Xu, and E. J. Fertig, "Enter the Matrix: Factorization Uncovers Knowledge from Omics," oct 2018.

[24] N. Sompairac, E. Barillot, P. V. Nazarov, U. Czerwinska, L. Cantini, A. Biton, A. Molkenov, Z. Zhumadilov, F. Radvanyi, A. Gorban, U. Kairov, and A. Zinovyev, "Independent component analysis for unraveling the complexity of cancer omics datasets," *International Journal of Molecular Sciences*, vol. 20, no. 18, 2019.

[25] L. Cantini, U. Kairov, A. De Reyniès, E. Barillot, F. Radvanyi, A. Zinovyev, and I. Birol, "Assessing reproducibility of matrix factorization methods in independent transcriptomes," *Bioinformatics*, vol. 35, no. 21, pp. 4307–4313, 2019.

[26] U. Kairov, L. Cantini, A. Greco, A. Molkenov, U. Czerwinska, E. Barillot, and A. Zinovyev, "Determining the optimal number of independent components for reproducible transcriptomic data analysis," *BMC Genomics*, vol. 18, no. 1, pp. 1–13, 2017.

[27] W. Kong, C. R. Vanderburg, H. Gunshin, J. T. Rogers, and X. Huang, "A review of independent component analysis application to microarray gene expression data.," *BioTechniques*, vol. 45, pp. 501–20, nov 2008.

[28] S. I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology*, vol. 4, no. 11, 2003.

[29] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane, "Dimension reduction techniques for the integrative analysis of multi-omics data," *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 628–641, 2016.

[30] E. Barillot, L. Calzone, and P. Hupe, "Review of Computational Systems Biology of Cancer," *BioMedical Engineering OnLine*, vol. 12, no. 1, p. 76, 2013.

[31] G. Way and M. Zietz, *Sequential compression of gene expression across dimensionalities and methods reveals no single best method or dimensionality*. 2019.

[32] Wikipedia, "Silouette clustering."

[33] D. V. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. W. Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, W. Dampier, C. Dessimoz, P. Flick, and H. Tang, "GOATOOLS: A Python library for Gene Ontology analyses," *Scientific Reports*, vol. 8, no. 1, pp. 1–17, 2018.

[34] C. Davidson-Pilon, J. Kalderstam, N. Jacobson, Sean-reed, B. Kuhn, P. Zivich, M. Williamson, AbdealiJK, D. Datta, A. Fiore-Gartland, A. Parij, D. WIlson, Gabriel, L. Moneda, K. Stark, A. Moncada-Torres, H. Gadgil, Jona, K. Singaravelan, L. Besson, M. S. Peña, S. Anton, A. Klintberg, J. Noorbakhsh, M. Begun, R. Kumar, S. Hussey, D. Golland, Jlim13, and A. Flaxman, "CamDavidsonPilon/lifelines: v0.24.16," jul 2020.

[35] Wikipedia, "Point-biserial correlation coefficient."

[36] E. Renard, S. Branders, and P. A. Absil, “Independent component analysis to remove batch effects from merged microarray datasets,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9838 LNCS, pp. 281–292, 2016.

[37] N. Nishida, H. Yano, T. Nishida, T. Kamura, and M. Kojiro, “Angiogenesis in cancer,” *Vascular Health and Risk Management*, vol. 2, no. 3, pp. 213–219, 2006.

[38] S. O. Sulima, I. J. F. Hofman, K. De Keersmaecker, and J. D. Dinman, “How Ribosomes Translate Cancer.,” *Cancer discovery*, vol. 7, no. 10, pp. 1069–1087, 2017.

[39] A. D. Theocharis, D. Manou, and N. K. Karamanos, “The extracellular matrix as a multitasking player in disease,” *FEBS Journal*, vol. 286, no. 15, pp. 2830–2869, 2019.

[40] L. M. Ogawa, S. J. Baserga, N. Haven, and N. Haven, “Crosstalk between the nucleolus and the DNA damage response,” vol. 13, no. 3, pp. 443–455, 2018.

[41] L. Cantini, U. Kairov, A. De Reyniès, E. Barillot, F. Radvanyi, A. Zinovyev, and I. Birol, “SUPPLEMENTARY INFORMATION: Assessing reproducibility of matrix factorization methods in independent transcriptomes,” *Bioinformatics*, vol. 35, no. 21, pp. 4307–4313, 2019.

[42] W. Zhou and R. B. Altman, “Data-driven human transcriptomic modules determined by independent component analysis,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–25, 2018.

6 Appendices

6.1 Metagene clustering – additional figures

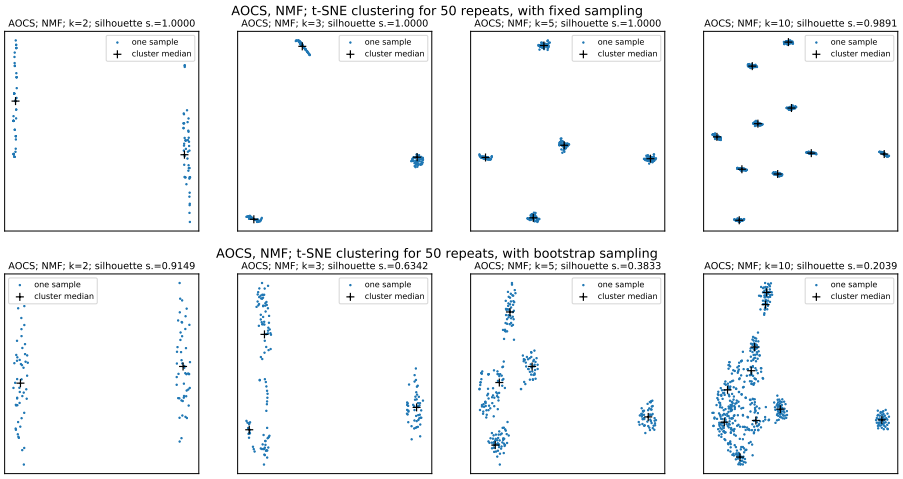


Figure 21: Clustering of metagenes from NMF facorizations on the AOCS dataset, comparing fixed and bootstrap sampling

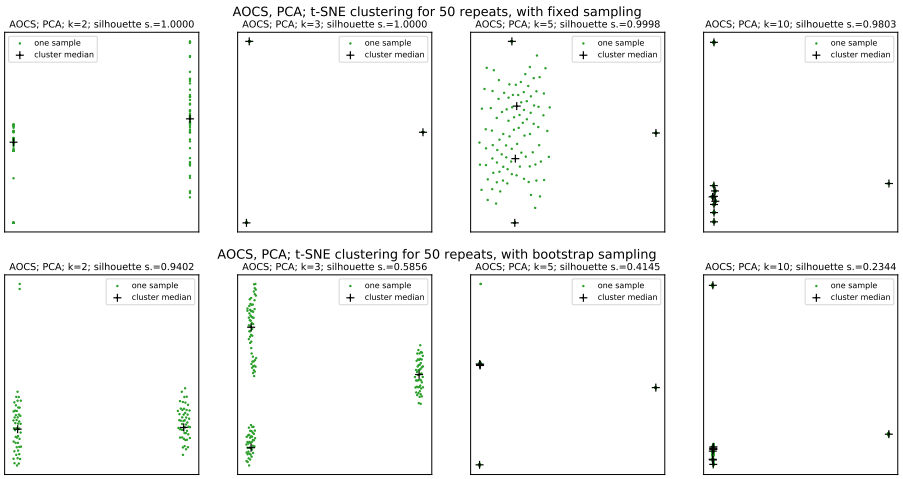
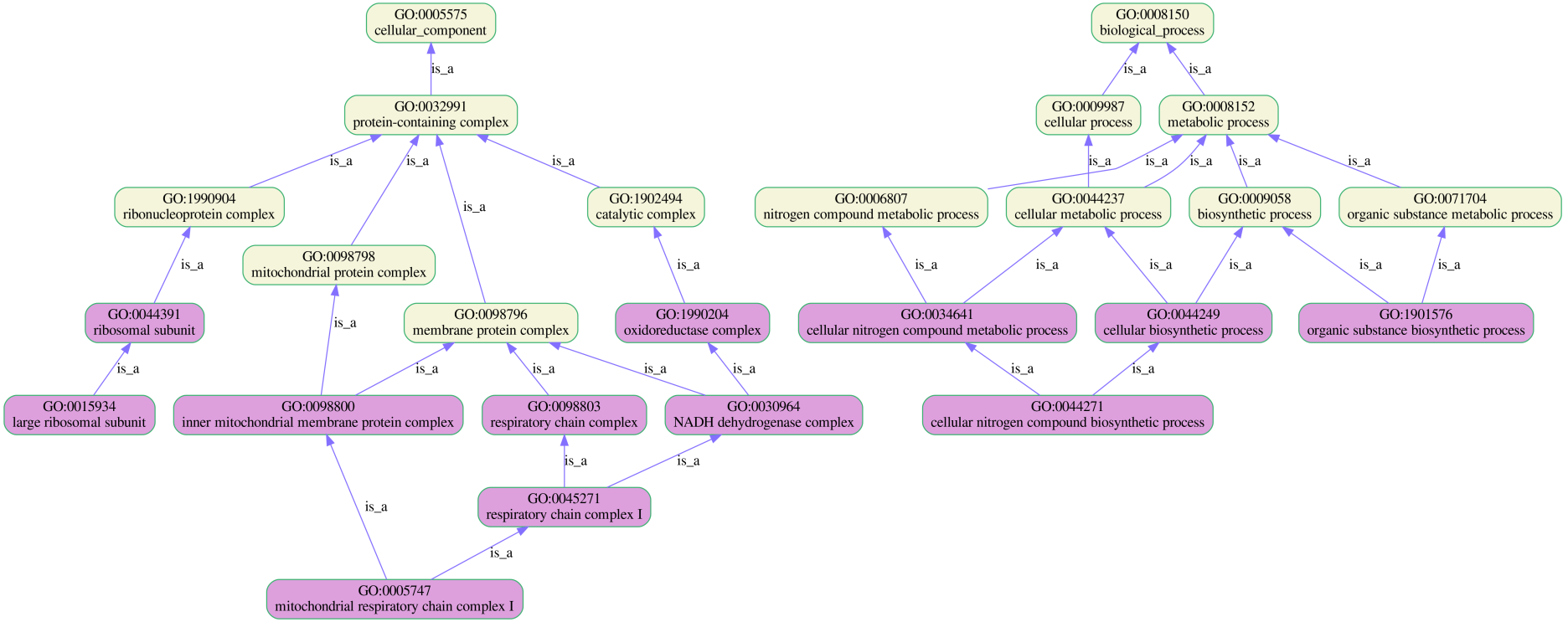


Figure 22: Clustering of metagenes from PCA facorizations on the AOCS dataset, comparing fixed and bootstrap sampling

6.2 GEA – additional plot



Genes: COX5B SNRPG RPL28 ROMO1 RPL26L1 SMDT1 NDUFA11 RPL36 RPS16 COX8A NDUFA7 ATP5PF NOP10 NDUFB1 ATP5MD NDUFC1 COA6 NDUFA3 NHP2 NDUFB8 NDUFC2 RPS5 RPL37 NDUFB10 COX5A RPS25 APRT ATP5ME NDUFA1 RPL35A RPS21 RPL14 ATP5MC2 COX7B COX6A1 NDUFB6 RPL39 RPL9 NDUFB9 ATP5MG UQCRH ATP5F1D MRPL41 ATP5MPL SNRPF RPS15 RPL22L1 ATP5PD POLR2I RPL29 RPL18A RPL8 NDUFA2 RPSA MRPL21 MRPL51 RPL18 SMG6 RPS13 UQCRB COX7C MRPL57 RPS18 RPL27 NDUFB2 NDUFB11 ATP5MC3 COX6C RPL13 RPL19 RPS11 RPL26 NDUFB7 MRPL36 NDUFA4 NDUFS6 RPS19 NDUFB4 TMA7 MRPL11 MRPL54 RPS23 RPL32 NDUFA13 DNPH1 RPL10A RPL35 RPS14 MRPL12 RPL31 COX4I1 UQCR10 RPL36A LAGE3 MRPL52 RPS27 RPS20 RPL30 RPL23 RPS9

Figure 23: Lineage maps of enriched GO terms for component ICA-4. For this component it was necessary to additionally filter the enriched terms – see text.

6.3 Survival analysis – additional figures

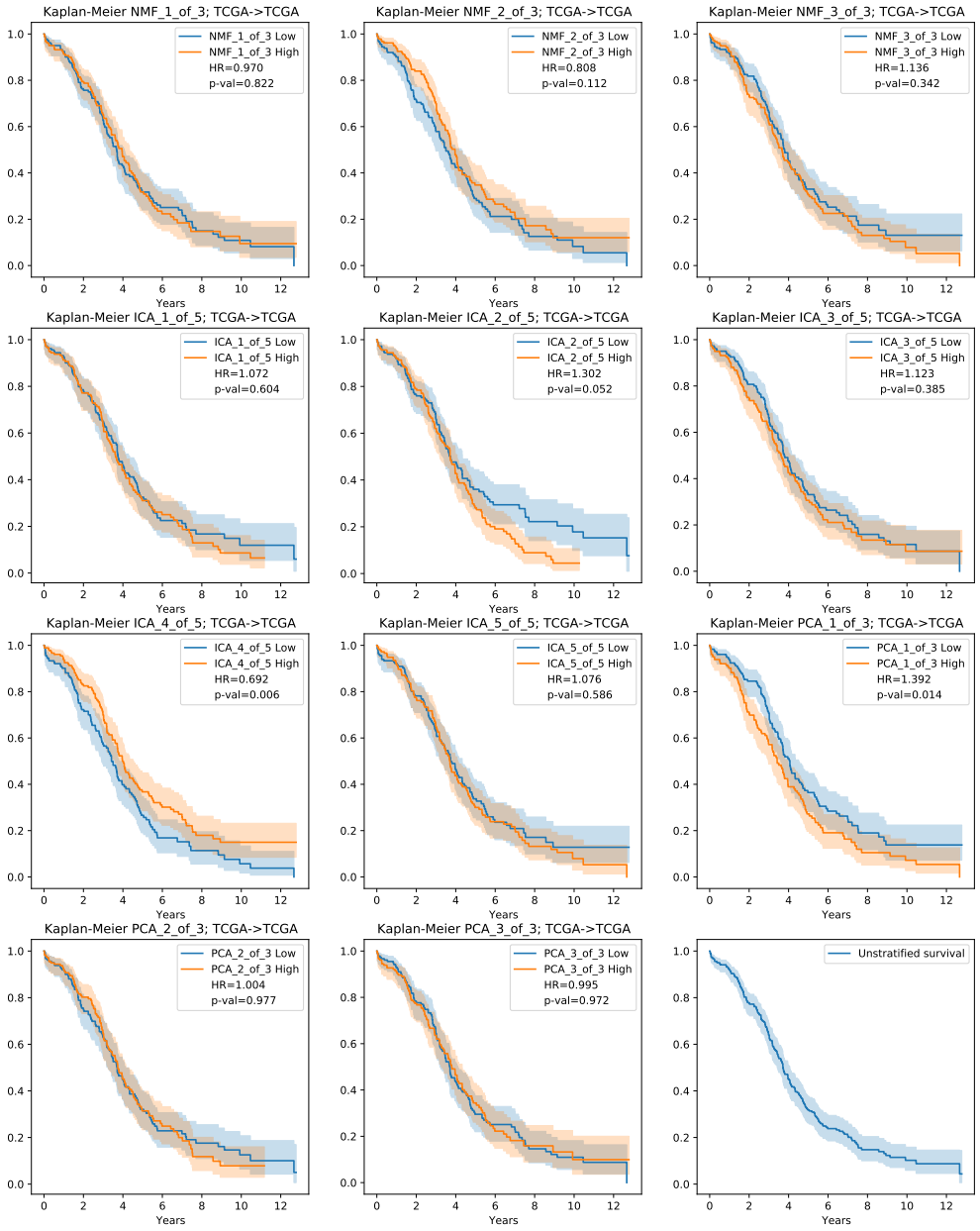


Figure 24: Kaplan-Meier plots for each metasample component, stratified at the median value, for TCGA $\rightarrow$ TCGA for overall survival (OS) case. Hazard ratio and p-value is shown for each case. The final plot (bottom right) is unstratified overall survival

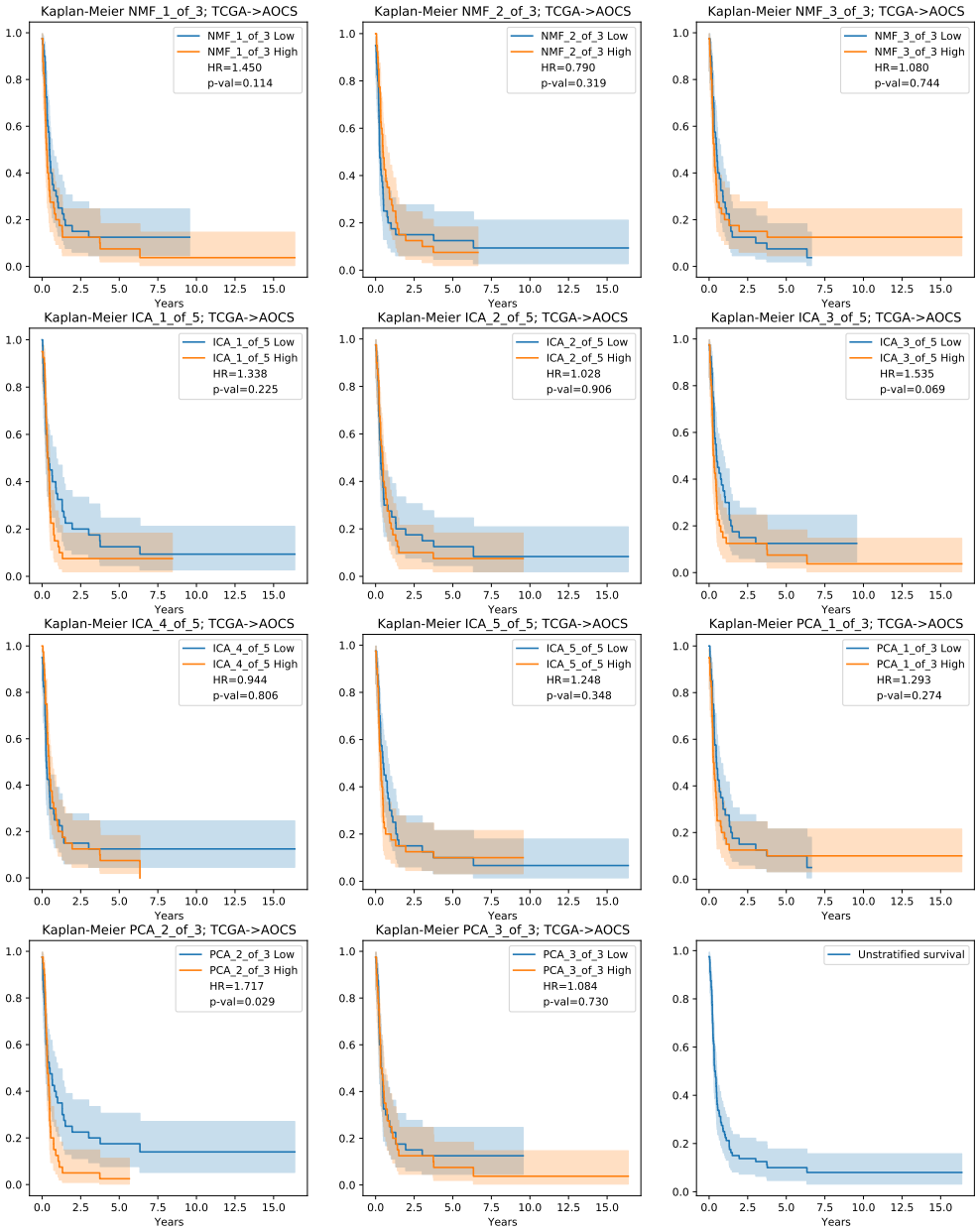


Figure 25: Kaplan-Meier plots for TCGA $\rightarrow$ TCGA for progression free survival (PFS).

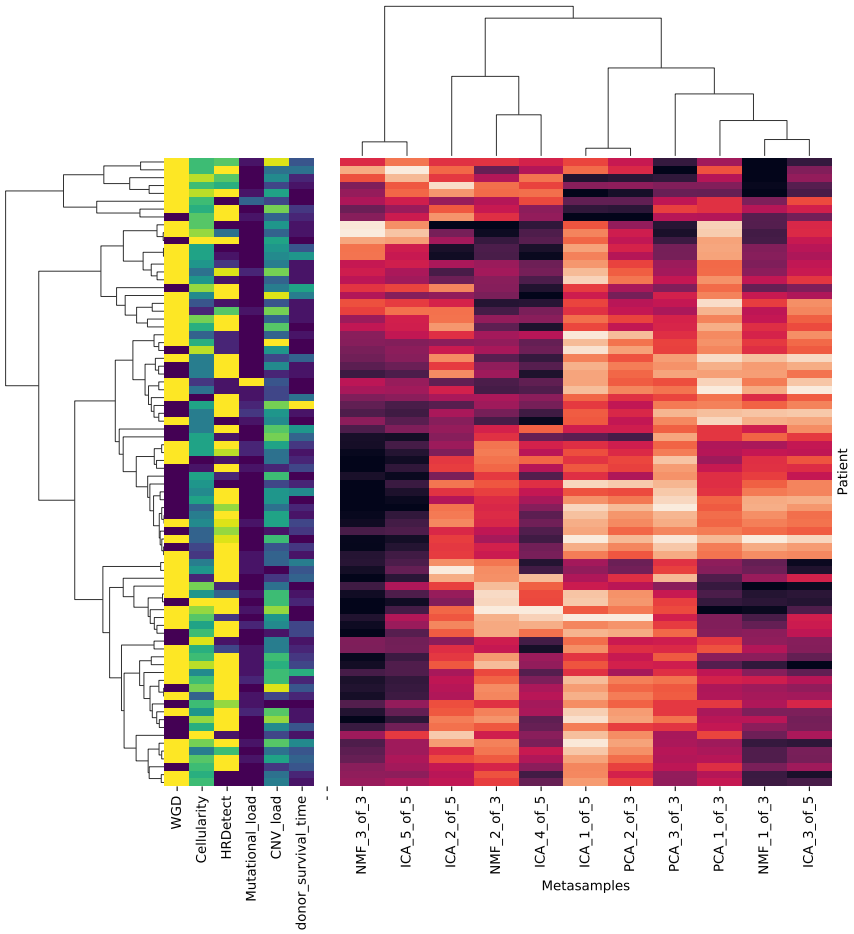


Figure 26: Heatmap of the metasamples matrix for the AOCS dataset, computed from metagenes factorized from the TCGA dataset. Several patient metadata variables are shown in columns to the left.

6.4 Batch effect investigation – Additional figures

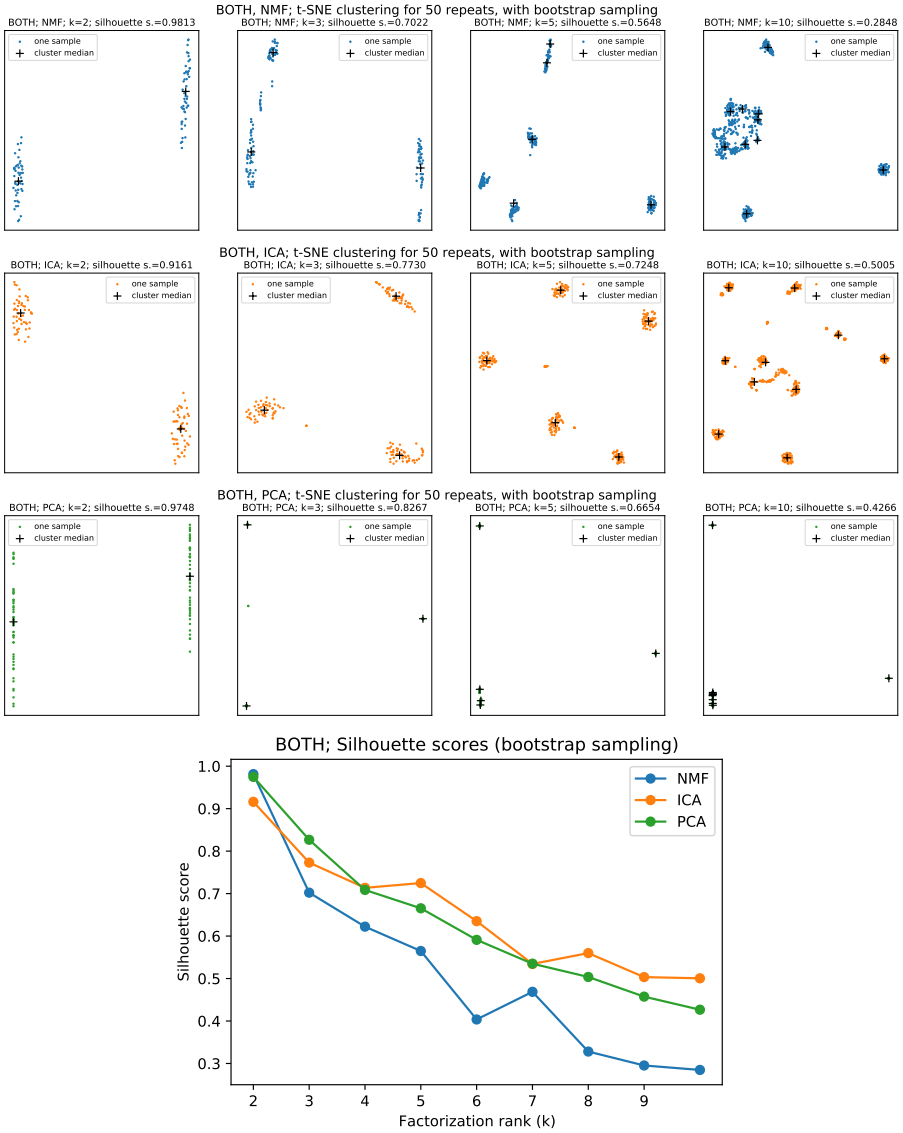


Figure 27: Metagene clustering for all three methods applied to the combined N=80+374 AOCS + TCGA dataset with bootstrap sampling, over a range of factorization ranks. The silhouette scores are also plotted (bottom).

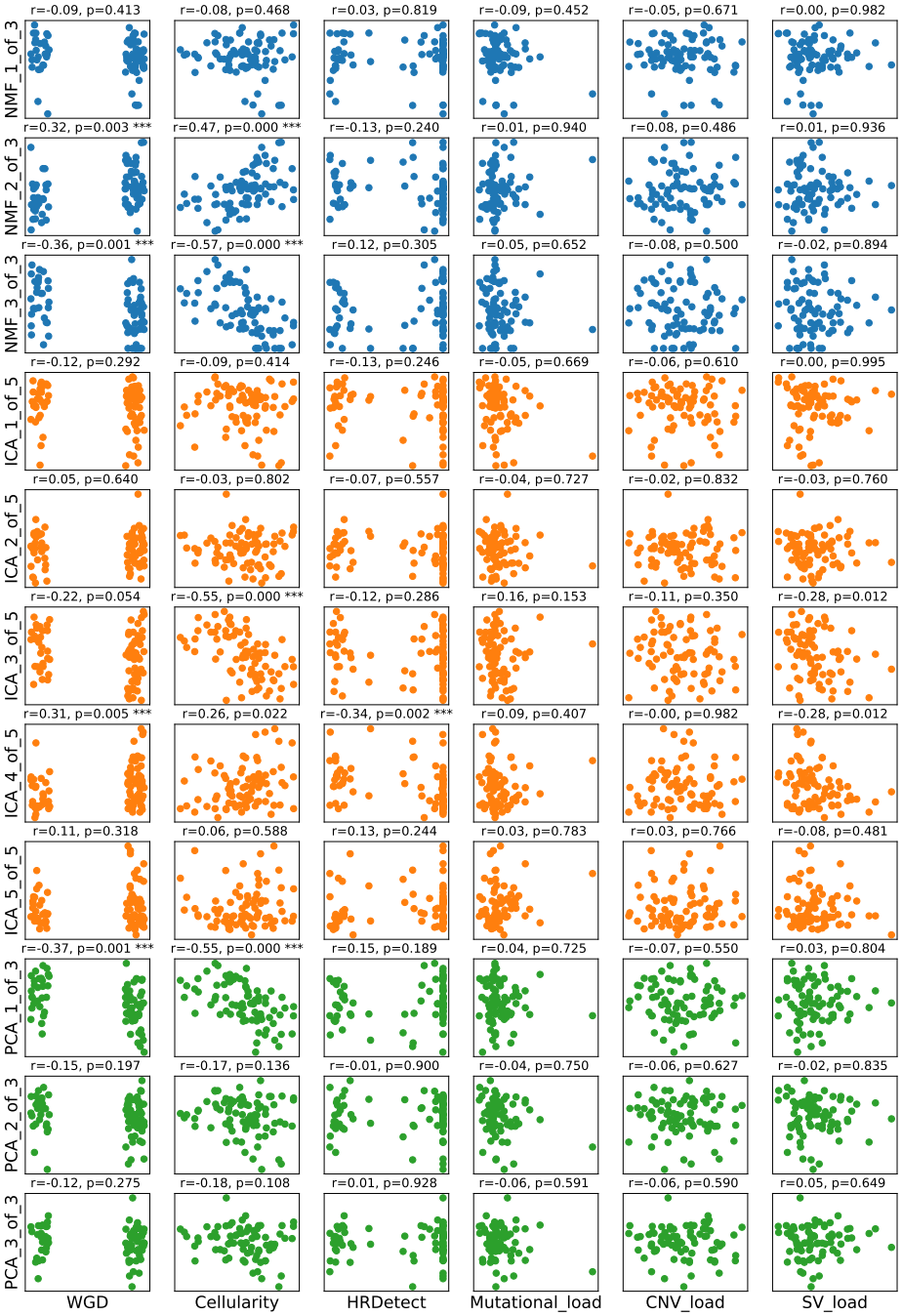


Figure 28: Grid of scatter plots to visualise the correlation between metasamples derived from the metagenes of the combined AOCS + TCGA datasets, and genomic features of the AOCS.

6.5 Access to gene enrichment analysis table

The table of GEA results is too large to conveniently include here. If required it can be obtained from my github repository at https://github.com/ipoole/HgsocTromics/

blob/master/Results/combined_gea_fdr.tsv.