THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Disentangling patterns of gene expression in high grade serous ovarian cancer

Student Exam 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

Here is the abstract

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 and 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 for HGSOC 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

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 (HRD) – 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.

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 [6] and [7].

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 [8].

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 [8].

Wang *et al* [9] 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* [10] 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 [11]. 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 [12]. 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 attention to batch effects. A final classifier took a consensus of the two methods based on a minimal gene set (order 40 of genes), 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, never-the-less 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 [13]. 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 (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.

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

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 Maringer *et al* [14] 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 GEO (199 samples) which included survival information, and found no significant predictive power [15]. 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 [16]. 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 myloma 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 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 [17]. A more direct molecular link is proposed by Cuello *et al* [18]. 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 been 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. [19]) 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 (mRMA 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 separation 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. See [23] for a full explanation, from which figure 1 is taken.

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 assumed to be so. PCA results in components which are ordered by the proportion of total variance they explain, and are mutually orthogonal.

The contrasting results of the three methods is illustrated in figure 1. *TODO: add explanation of this diagram*

Conventionally, transcriptomics expression arrays are oriented with genes (or transcripts) as 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* these are the $M \times K$ *pattern matrix* and the $K \times N$ *amplitude matrix*. K (or k as we will use) 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 sometimes used. The original matrix to be factorized is variously named X or V.

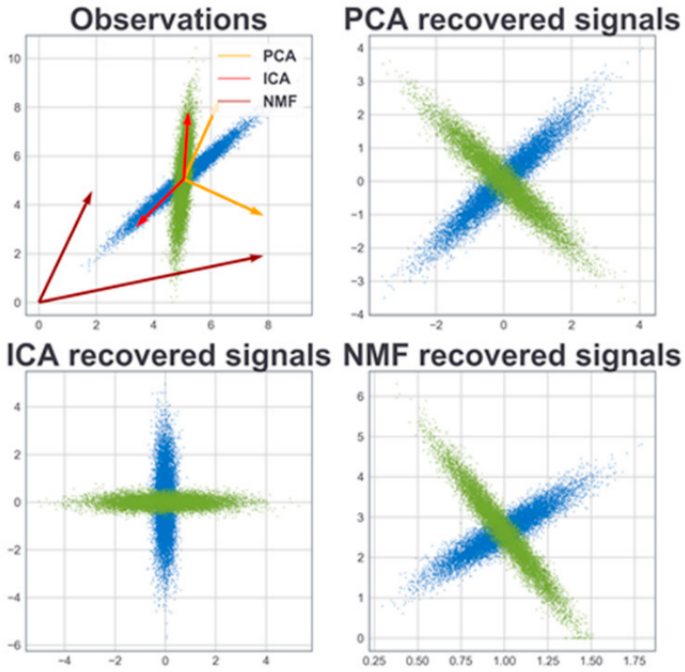


Figure 1: Illustrating the differing results of component extraction by PCA, ICA and NMF. From [23]

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

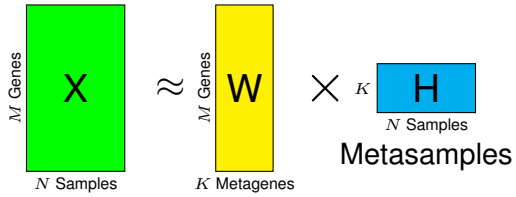
$$X \approx WH \tag{1}$$

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

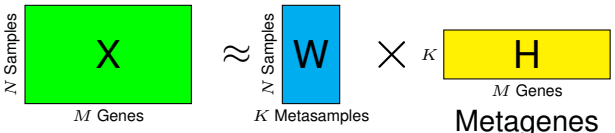
$$X_{M,N} \approx W_{M,K} H_{K,N} \tag{2}$$

as illustrated in figure 2 (a).

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



(a) Optimising independence of Metagenes



(b) Optimising independence of Metasamples

Figure 2: 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.

gene expression analysis with exchanged and transposed meanings of W and H, and this is illustrated in figure 2 (b). In this case it is the properties of the *metasamples* which will be 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” [23], 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” [24]. According to Cantini *et al* references [17], [25], [26] and [27] optimise metagenes, while references [28] and [29] 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 [20].

The problem of determining the optimal k for a given expression matrix is addressed by Kairov *et al* [25], based on optimizing the *stability* of the components over multiple algorithm initializations.

1.4 Research goals

This work is concerned with analysing the patterns gene expression in HGSOC by dimensionality reduction techniques of matrix factorization, and relating these patterns to clinical and genomic features.

The following research questions were set out early in the project. *TODO: tweak these goals!*

1. Which are the influential genes in HGSOC according to gene expression data? Do these confirm published results?
2. Which of PCA, ICA and NMF is best suited to this analysis?
3. For identified genes, what is the underlying biology?
4. Can we discriminate subtypes of HGSOC from expression data?
5. Is there value in incorporating somatic and germline genomic data?

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. 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.

2.1.1 Datasets

Two gene expression datasets were assailable for this work (n refers to number of patients, g refers to number of genes).

1. The Cancer Genome Atlas (TCGA) derived, n=274, g=19,601, with metadata on survival
2. Austrailian Ovarian Cancer Study (AOCS) [3], n=80, g=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, 18, 17].

TODO: *more explanation of the provenance of these datasets*.

Expression data for the AOCS and TCGA datasets was received for this project in a spreadsheet format, having been derived from the RNA-Seq data as described in [5], including normalization by variance stabilizing transformation.

2.1.2 Method outline

An overview of the methodology adopted in this work is shown in figure 3, 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 determined factors by gene enrichment analysis against the Gene Ontology (GO).
5. Computation by least squares optimisation of metasamples associated with 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 relationship 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.

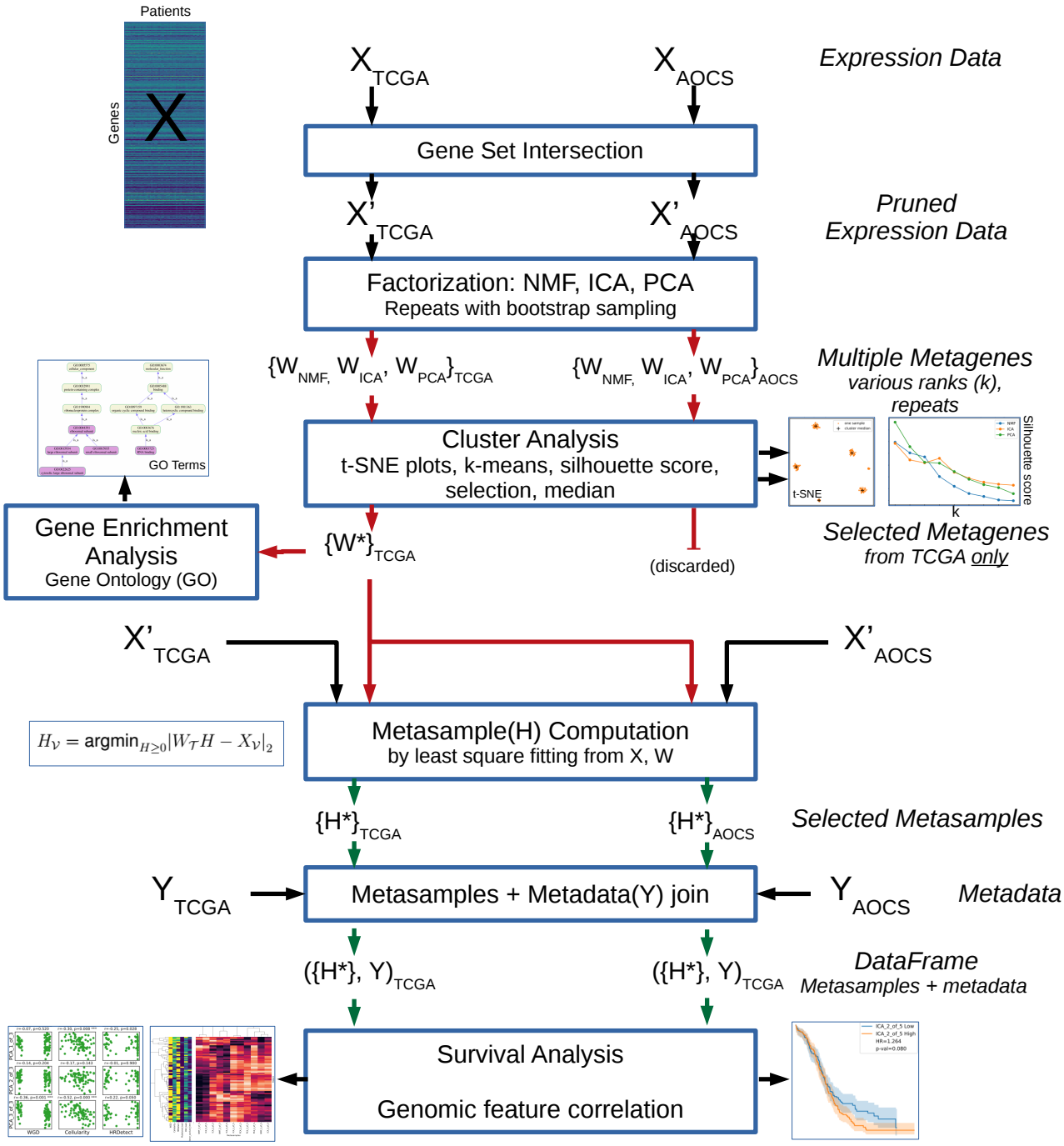


Figure 3: Overview of methodology as information flow. The diagramatic convention here is that at each stage, both datasets – TCGA and AOCS – are processed separately by the given algorithm. The exception is Gene Set Intersection, which involves interaction between the datasets.

2.1.3 Tools

- Python 3.6
- PyCharm Integrated Development Environment
- Numpy for high performance matrix manipulation
- Pandas for data frame handing
- Matplotlib and Seaborn for general plotting
- Scikit-learn for matrix factorization and k-means clustering
- GOATOOLS package for gene enrichment analysis against GO
- Lifelines package for survival analysis.

A full list of software and libraries used, with versions, is given in appendix 6.3.

2.2 Gene set intersection

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. Other parameters of interest are `alpha` (multipluer 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. Additionally, options for the entropy functional 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 stochastic behaviour.

TODO: probably best to state the chosen hyper parameters here

2.4 Metagene selection by cluster coherence

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 overall process is summarised as pseudocode in figure 4.

A complication arises in the computation of ICA and PCA facorizations, in that essentially the same factor can arise as w or $-w$ – i.e. $180°$ rotated vectors. These would appear as separate cluster 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 for $r \geq 10$.

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*. 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 *not* in the same cluster. See Wikipedia: Silouette (clustering). *TODO: make this a reference* 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, those median metagenes to take forward to gene enrichment and 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 (this will be seen in the presented results). 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.

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
         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
      for each factorizer:
         plot score vs k
```

Figure 4: 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 components 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 to consider the intersection of the gene sets which each metagene highlights. In the gene enrichment analysis which follows, candidate genes will be identified for each meata-

gene. 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 heatmap.

2.6 Gene enrichment analysis

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 HGSOC. In order to gain insights into what biological processes are involved, gene enrichment analysis against the Gene Ontology (GO) was carried out for each of the 11 metagenes individually.

The essence of gene enrichment analysis 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 [30].

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 gene enrichment analysis run.

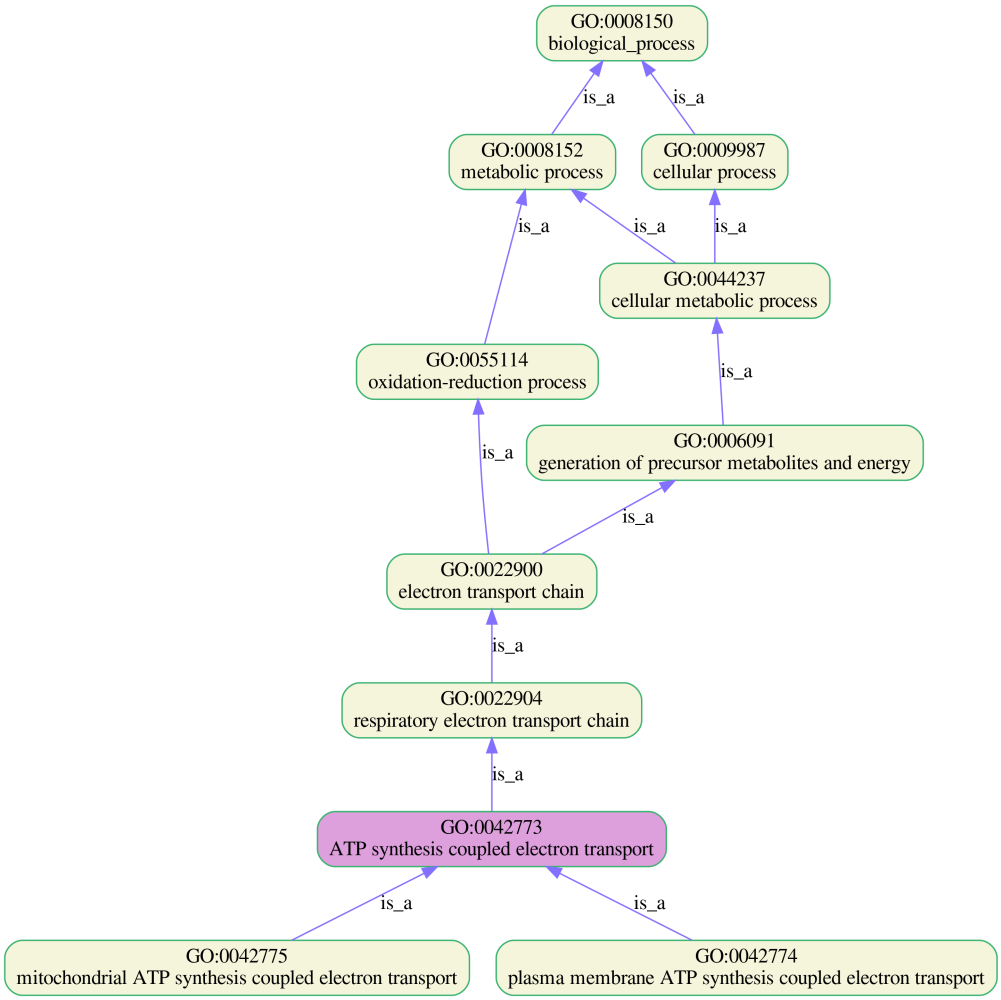


Figure 5: Example of a small section of the Gene Ontology (GO), focussed GO:0042773 (in purple), showing parents and children of the term. Generated by GOATOOLS.

The result of this analysis, per metagene, is a list of enriched (or depleted, however in our case 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 5

2.7 Transfer of learned metagenes to novel a dataset

It is fundamental to our approach that *metagenes* determined on the basis of one cohort of data can be used to generate *metasamples* for a different cohort. 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 $(g_{\mathcal{T}}, n_{\mathcal{T}})$, with $g_{\mathcal{T}}$ the number genes and $n_{\mathcal{T}}$ the number of patients. $W_{\mathcal{T}}$ is the metagene matrix of shape $(g_{\mathcal{T}}, k)$ and H_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 $(g_{\mathcal{V}}, n_{\mathcal{V}})$. 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 $g_{\mathcal{T}} = g_{\mathcal{V}} = g$, in which case $W_{\mathcal{T}}$ and $W_{\mathcal{V}}$ have the same shape of (g, 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 Python library function `scipy.optimize.nnls` is used.

For ICA and PCA we use ordinary least square regression, formulated a 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 we wish to evaluate the efficacy of the three factorization methods – NMF, ICA and PCA – we in fact take forward three different H matrices, each having ranks as for the associated metagenes.

2.8 Reconciling computed metasamples and metadata

To carry out patient level analysis – survival analysis, heatmaps and boxplot – 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 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 (rows), and with the metadata with respect to patient identifiers in the columns. 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→TCGA (OS), 2) TCGA→AOCS (OS) and 3) TCGA→AOCS (PFS). *TODO: describe these three versions of analysis in words*.

Analysis was performed using the Python `Lifelines` package [31]. Metasample values were

binariased 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 plot 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) pre-determined per-patient high-level genomic features were available as follows:

WGD : Whole genome doubling – a marker of genome instability. This is 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 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 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. Unit tests are based on tiny expression matrices, just 100 genes by 10 patients, thus the whole test suite of over 60 tests executes in around 40 seconds and can be run frequently. These tests are particularly valuable when re-factoring code. 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 was used for plots relating to NMF, ICA and PCA respectively. Metagenes are consistently referred to by, for example, "NMF-2-of-3" – meaning the 2nd component of the rank $k = 3$ NMF factorization.

3 Results

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

From figure 6 it can be seen that with sampling error excluded, 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 7 makes the same comparison for the lager 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 PCA and NMF can be found in the appendix, figures

For this reason it is appropriate 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.

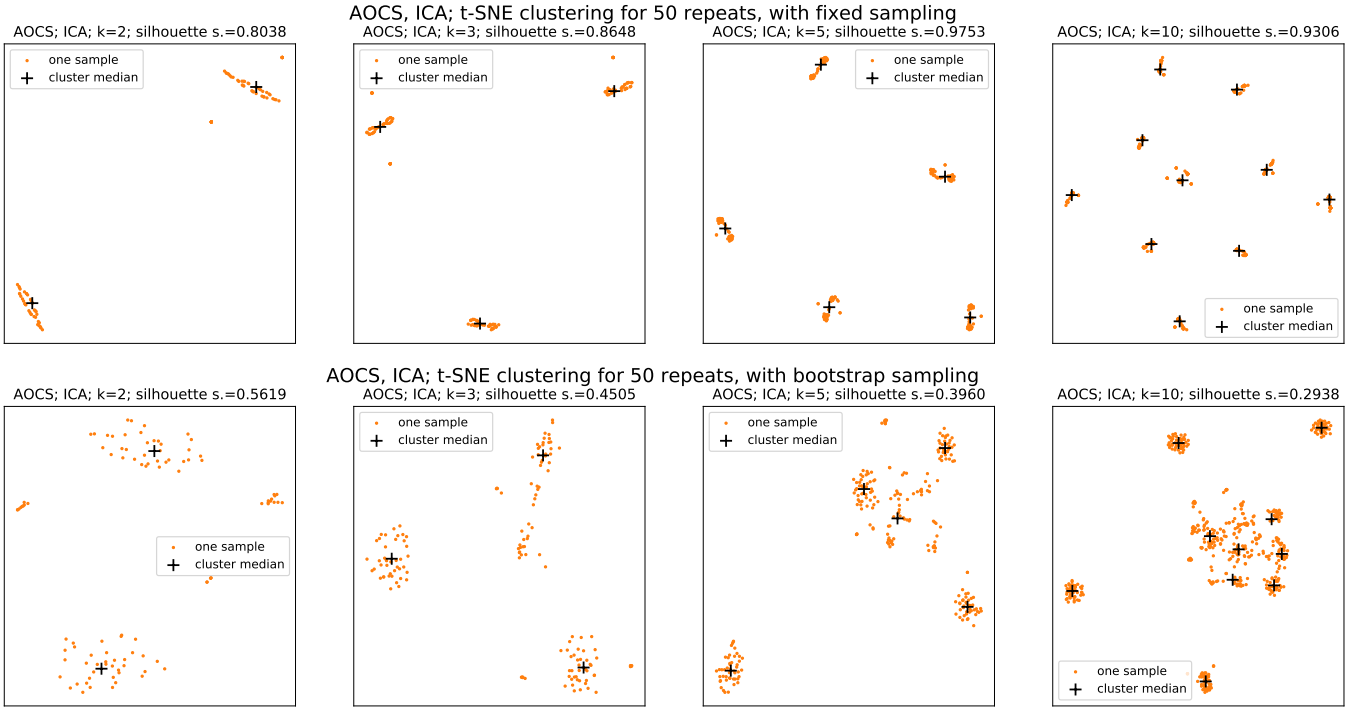


Figure 6: Clustering of metagenes from ICA facorizations on the N=80 AOCS dataset, comparing *fixed* sampling (top row) with *bootstrap* sampling (bottom row) for a selection of factorization ranks.

In choosing the factorization rank for each method, both the t-SNE plots and silhouette scrore graphs in figure 8 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

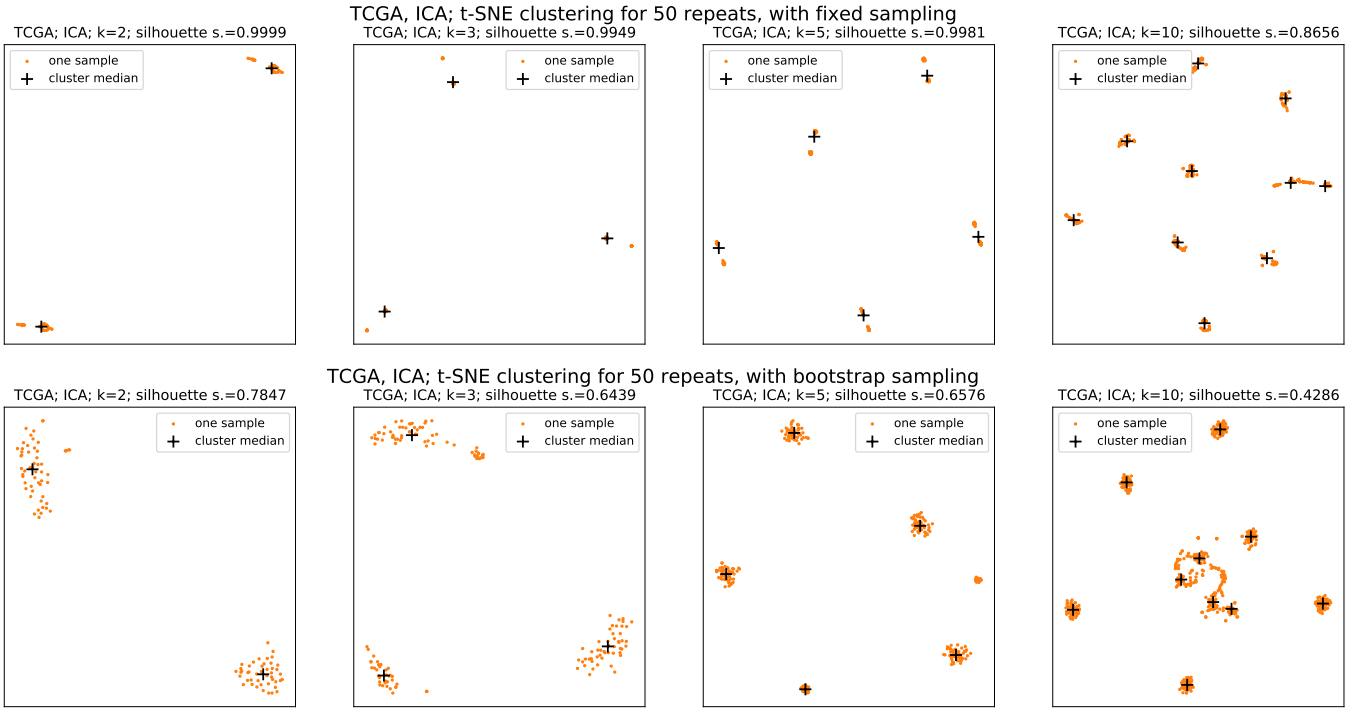


Figure 7: Clustering of metagenes from ICA facorizations on the N=374 TCGA dataset, again comparing fixed and bootstrap sampling.

of silhouette scores (figure 8, bottom), ICA appears to have a sweet spot at $k = 5$. For PCA, $k = 3$ or 4 both seem reasonable. The following choices were made: $K_{\text{NMF}} = 3, K_{\text{ICA}} = 5, K_{\text{PCA}} = 3$. These choices predicate all the results which follow.

There is a curious artefact visible in the NMF factorization clusterings of figure 8, 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. *TODO: hypothesis?*

3.2 Gene sets identified by metagenes of different methods intersect in some cases

The Jaccard heatmap of figure 9 shows that some pairs of metagenes identify overlapping sets of genes, in particular the pairs (ICA-1-of-5, PCA-2-of-3: J=0.61) and (ICA-5-of-5, PCA-3-of-3: J=0.49). As expected however, there is very little similarity between components *within* a factorization method.

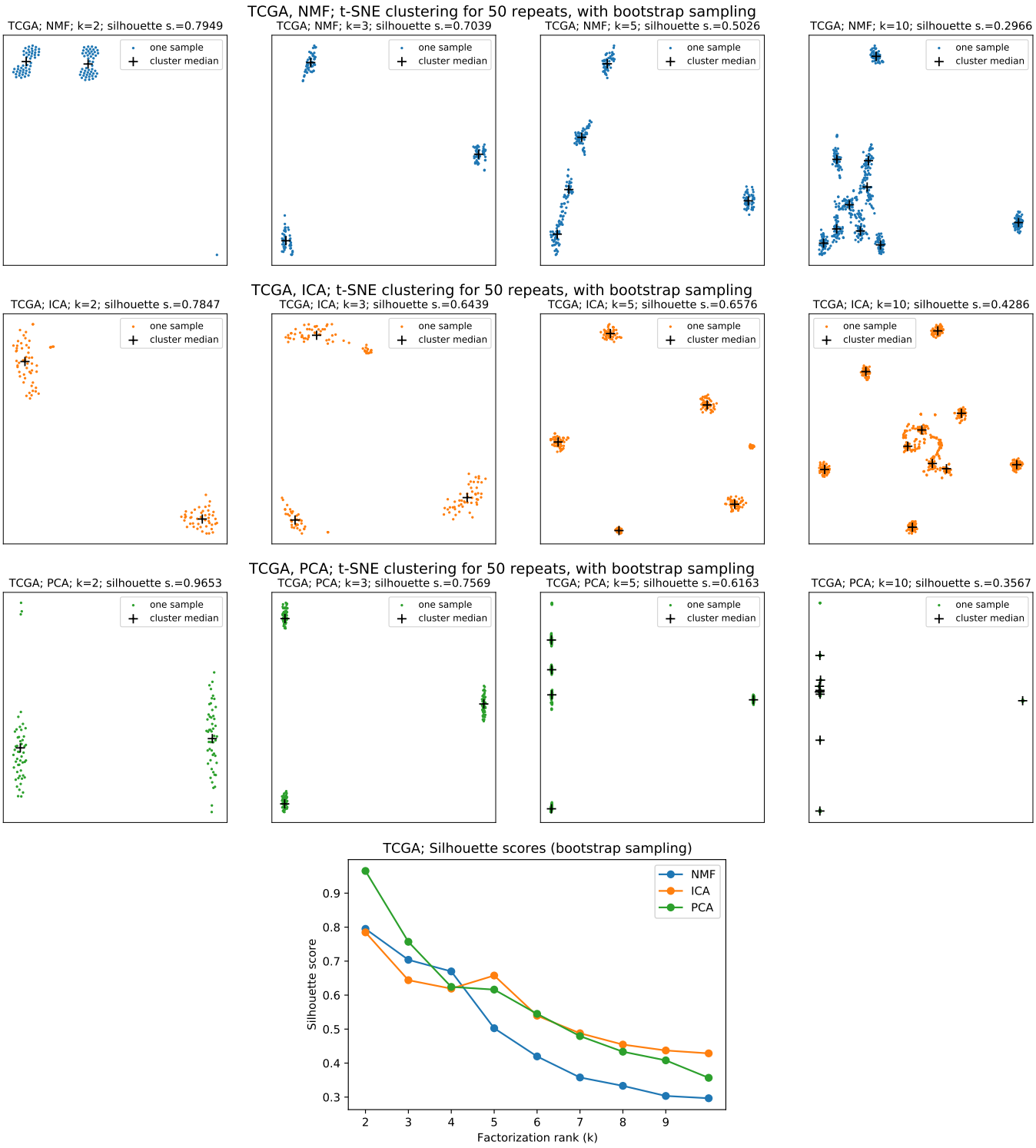


Figure 8: Metagene clustering for all three methods applied to the N=374 TCGA dataset with bootstrap sampling, over a range of factorization ranks. The silhouette scores are also plotted (bottom). It is on the basis of this figure that the factorization ranks for each method were selected.

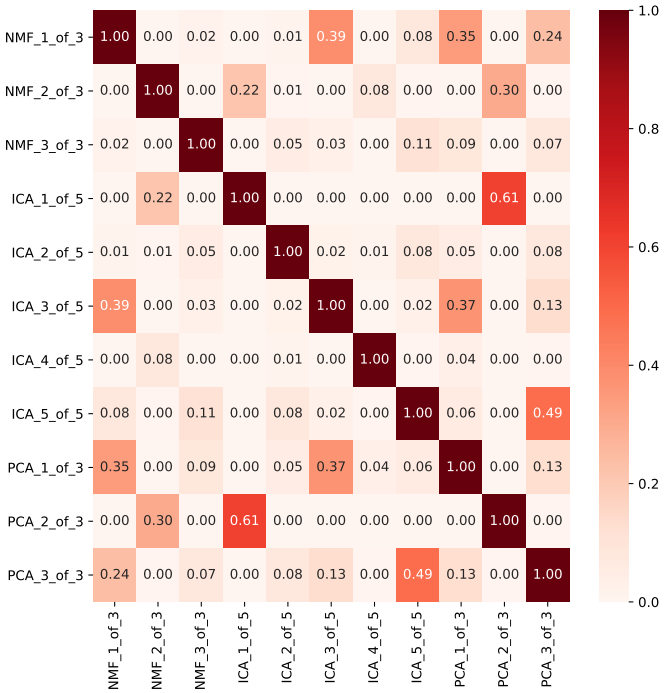


Figure 9: Heatmap of Jaccard similarities between the candidate genes identified by the 11 components.

3.3 Metagenes highlight genes which are enriched for particular biological processes

Gene enrichment analysis 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; this table can be found in the appendix, table 6.2. To take full advantage of the hierarchical nature of the GO, results are presented as lineage maps in figures 10 to 13. 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-of-5 – 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-of-3, ICA-1-of-5, ICA-2-of-5 and PCA-2-of-3 yielded no significant enriched terms (for FDR ≤ 0.01).

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

NMF 1-of-3 : this is related to the cellular structures and processes of the extra-cellular matrix (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 [32]. Thus, this metagene may simply reflect heterogeneity of tissues in the biopsied sample, or may have some deeper disease related significance.

NMF 2-of-3 : 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 ribosomal 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 [33].

NMF 3-of-3, ICA 1-of-5 and ICA 2-of-5 : no significant biological enrichment found.

ICA 3-of-5 : this component relates to processes of multicellular / extracellular organisation, and the ECM. It is therefore similar to NMF 1-of-3 above.

ICA 4-of-5 : this component is seen to relate to the ribosomal subunit (as NMF 2-of-3),

additionally membrane proteins in the respiratory complex, mitochondrial and the NADH dehydrogenase complex. Processes of organonitrogen biosynthesis are also highlighted.

ICA 5-of-5 : processes relating to regulation of immune response are enriched in this component, featuring genes from the major histocompatibility complex 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.

PCA 1-of-3 : there is mention of ECM related terms, as for ICA 3-of-5 and NMF 1-of-3. 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 [34]. We have already seen (section 1.2.2, [11]) there is evidence for distinct subtypes of OC distinguished by angiogenesis related genes, and that these subtypes have been found to inform clinical outcome [11]. Thus, this component might contain useful prognostic value.

PCA 2-of-3 : no significant biological enrichment found.

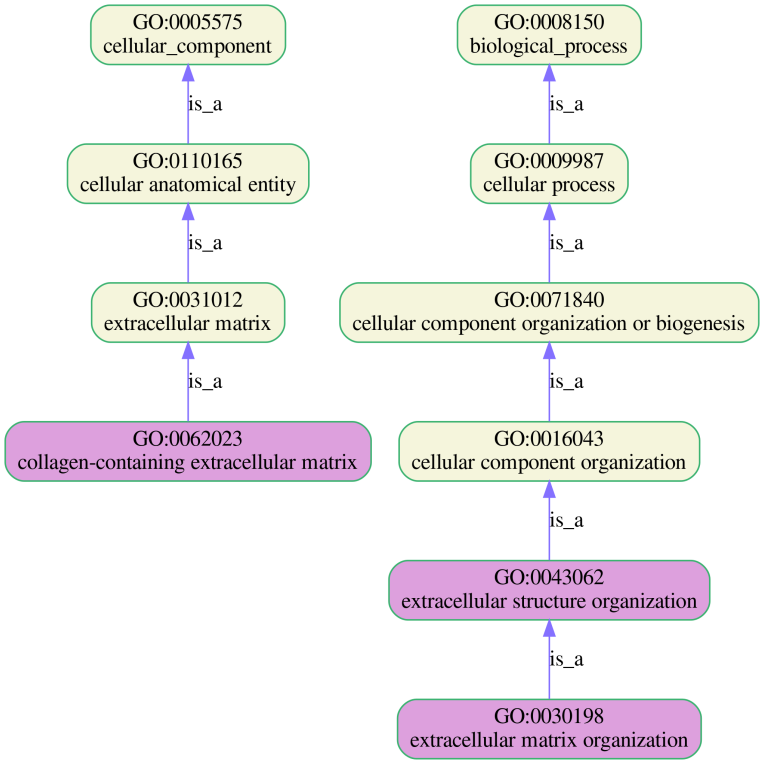
PCA 3-of-3 : this component has some similarity with ICA 5-of-5, in that it refers to regulation of immune processes (HLA- genes) and ER membrane. However, chemotaxis (cell movement) is also highlighted.

We have noted above that both NMF 1-of-3 and ICA 3-of-5 may be related to tissue type heterogeneity. We have also seen from metasample heatmap analysis that these two metagenes are associated with closely correlated metasamples (in AOCS), and also to have a moderately high Jaccard similarity of 0.39. Further, we observed that these components correlate negatively with cellularity. This is as one would expect: low cellularity implies a mixture of tumour and non-tumour cell types, resulting in variation of ECM composition.

Interestingly, the above summary of the gene enrichment results tentatively suggests component PCA 1-of-3 is most likely to have clinical prognostic value, based on a cursory look at the literature. This perhaps supports an earlier observation that, while not conclusive, component PCA-1-of-3 comes closest to demonstrating correlation with patient survival (section 2.9, figure 16).

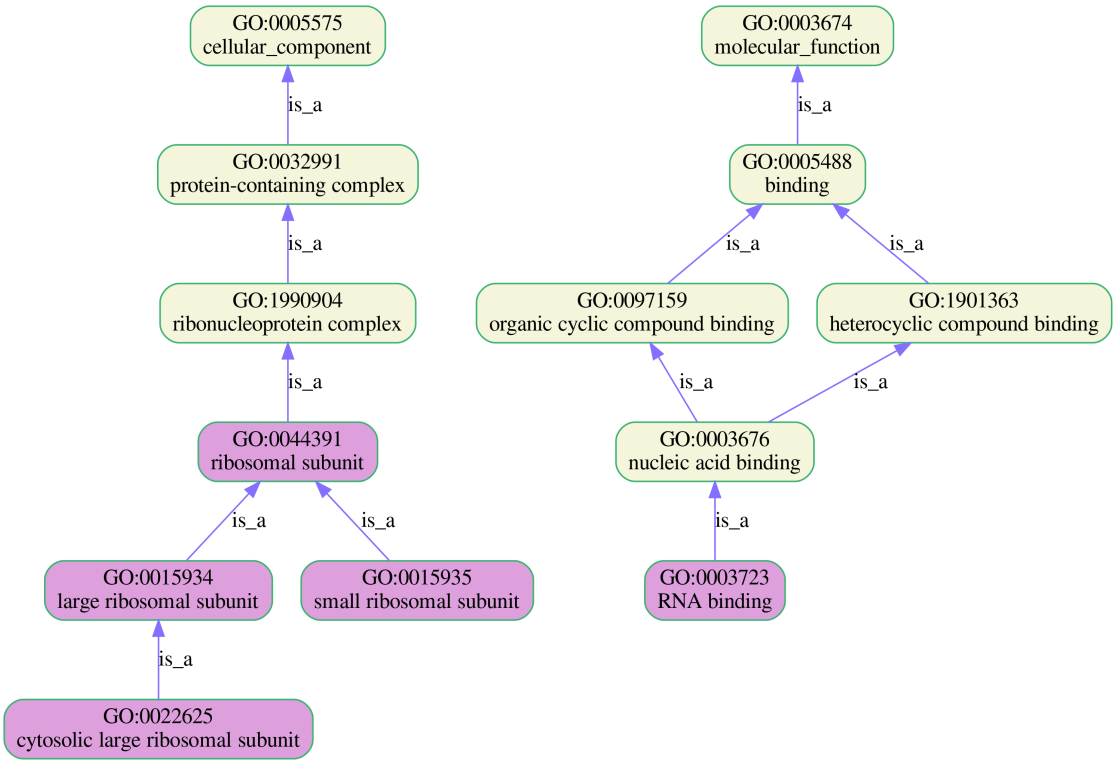
How should we interpret those four components which demonstrate no significant enrichment with biological meaning? It is notable that all three methods have at least one such component.

Among these four, PCA 1-of-3 and ICA 1-of-5 have high (the highest) Jaccard similarity of 0.61. For all other pairs among these four Jaccard similarity is negligible. Perhaps these two components are picking up on similar technical variation between the samples?



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

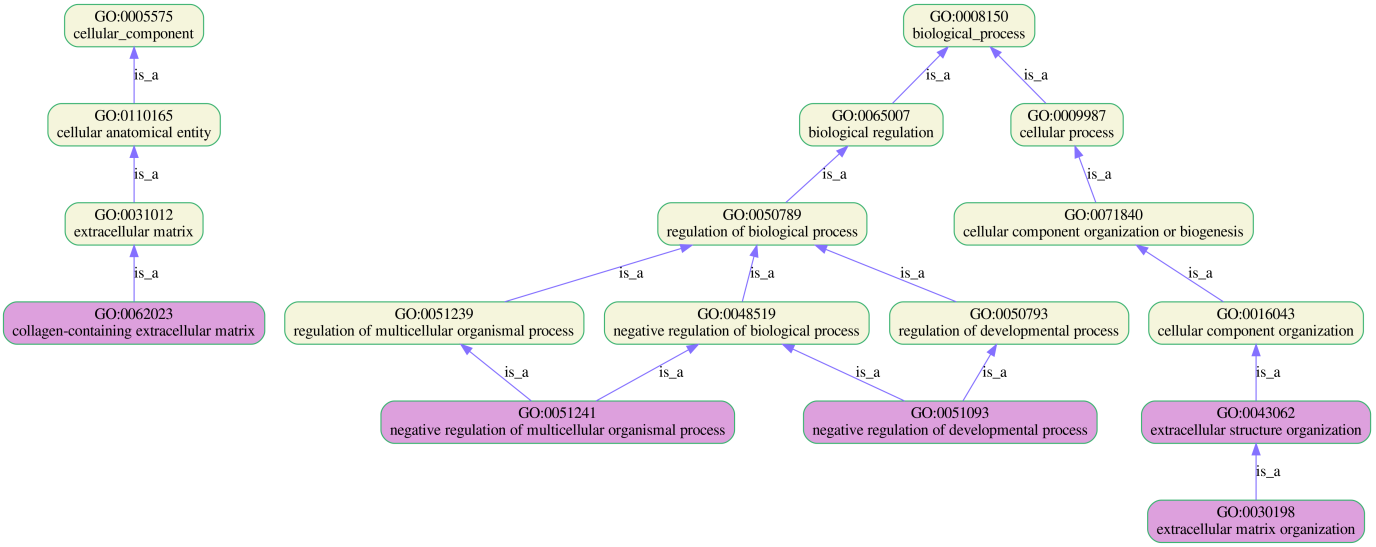
Component NMF-1-of-3



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

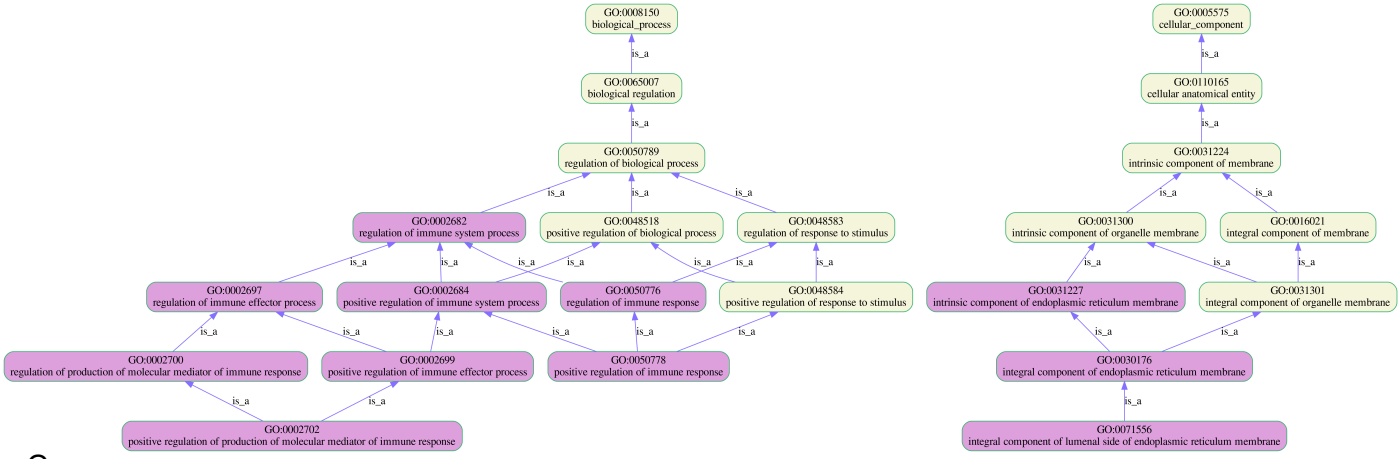
Component NMF 2-of-3

Figure 10: Lineage maps of enriched Gene Ongology (GO) terms for components NMF-1-of-3 and NMF-2-of-3. In these diagrams, enriched terms are coloured purple, while there ancestors in the ontology are yellow. (NMF-3-of-3 produced no significant enrichment results)



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

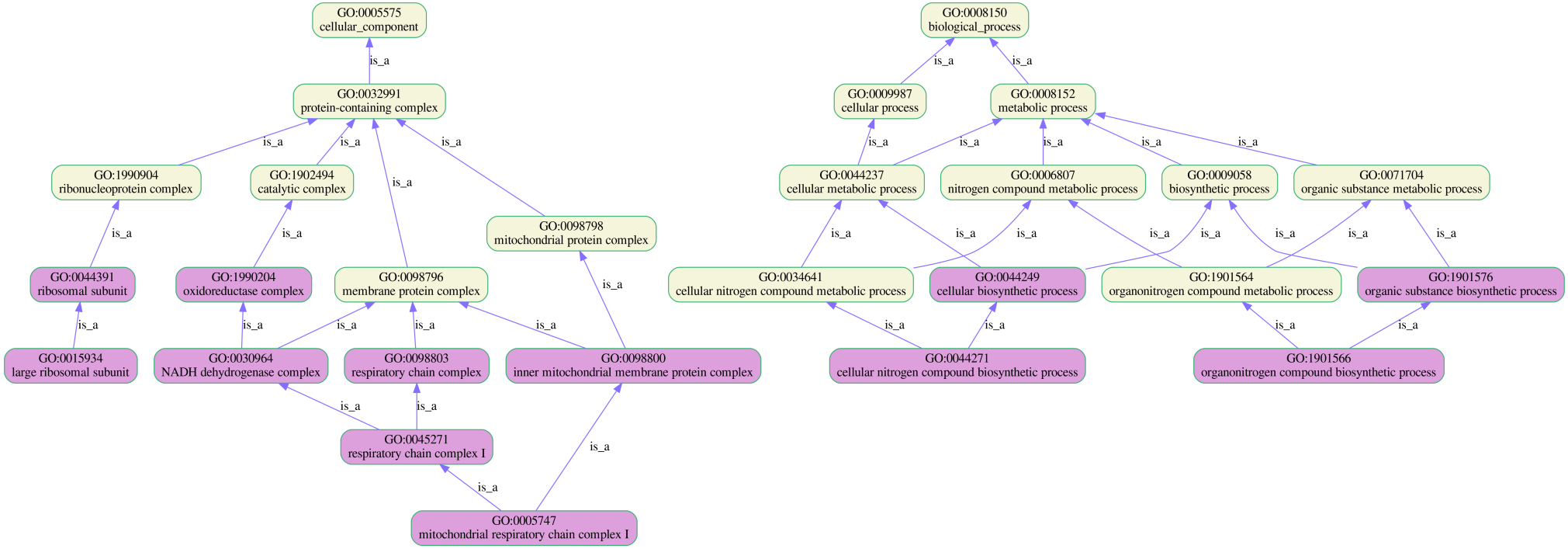
Component ICA-3-of-5



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

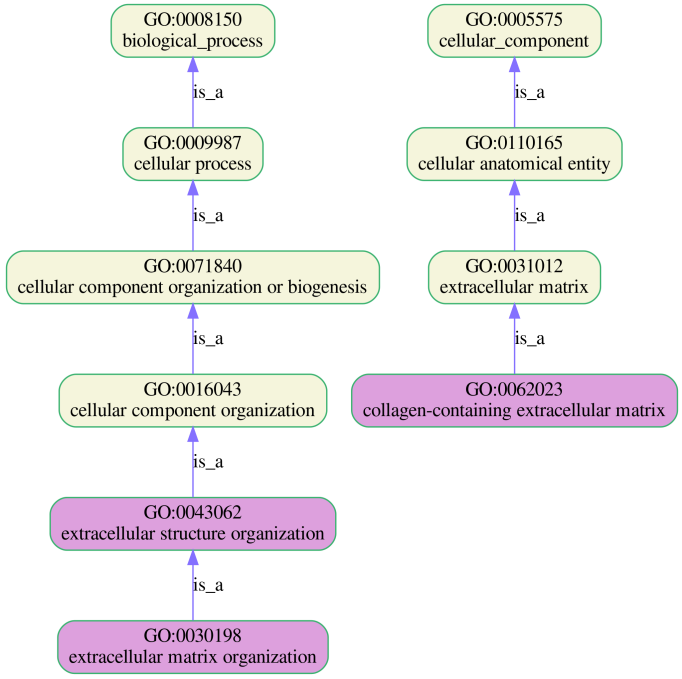
Component ICA-5-of-5

Figure 11: Lineage maps of enriched Gene Ongology (GO) terms for components ICA-3-of-5 and ICA-5-of-5



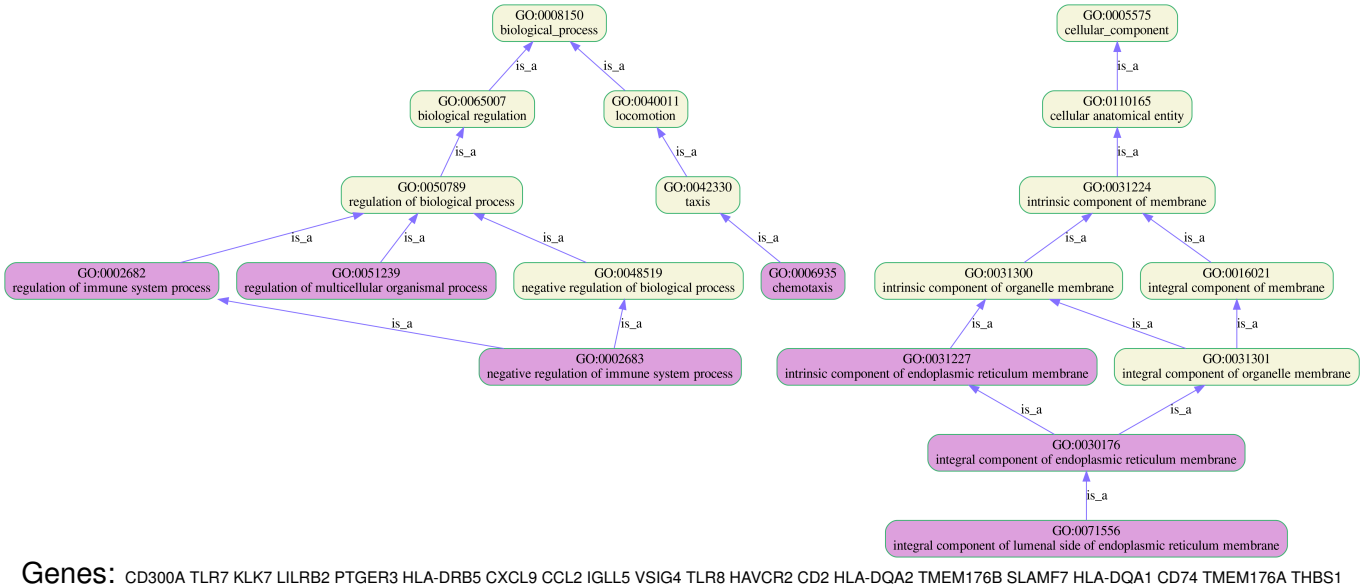
Genes: NDUFA7 RPL36 RPS21 TMA7 ATP5PF MRPL12 RPL32 RPL37 ATP5MC3 ATP5MPL RPL29 NDUFB1 MRPL54 COX7B RPL14 RPL9 RPL18A ATP5MD NDUFA2 MRPL41 RPS27 COX6C NDUFS6 POLR2I RPL28 MRPL52 ATP5MG NDUFB9 NDUFB11 ATP5F1D ATP5PD RPL39 NDUFA1 COX7C NDUFB8 RPS18 NDUFA4 RPS25 RPS15 COX5A COX6A1 NDUFB6 SMG6 RPL31 UQCRB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

Figure 12: Lineage maps of enriched Gene Ongology (GO) terms for component ICA-4-of-5. For this component it was necessary to additionally filter the enriched terms – see text.



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

Component PCA-1-of-3



Genes: CD300A TLR7 KLK7 LILRB2 PTGER3 HLA-DRB5 CXCL9 CCL2 IGLL5 VSIG4 TLR8 HAVCR2 CD2 HLA-DQA2 TMEM176B SLAMF7 HLA-DQA1 CD74 TMEM176A THBS1 CXCR6 ITGAX CXCL1 HLA-DQB2 SASH3 HLA-B HLA-DOA THBS2 VTCN1 PAEP HLA-DPB1 HLA-DRB1 CCR1 PAX2 B2M HLA-DPA1 HLA-DQB1 FCGR1A C3 TYROBP CCL11 HLA-DRA FCGR2B SLAMF8 VSIR

Component PCA-3-of-3

Figure 13: Lineage maps of enriched Gene Ongology (GO) terms for components 1 and 3 from the rank=5 PCA factorization. (Component PCA-2-of-3 produced no significant enrichment results)

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 TCGA dataset in figure 14 and for the AOCS dataset in figure 15. Plots for progression free survival (PFS) on AOCS can be found in the appendix, figure 21. 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-of-3, 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.

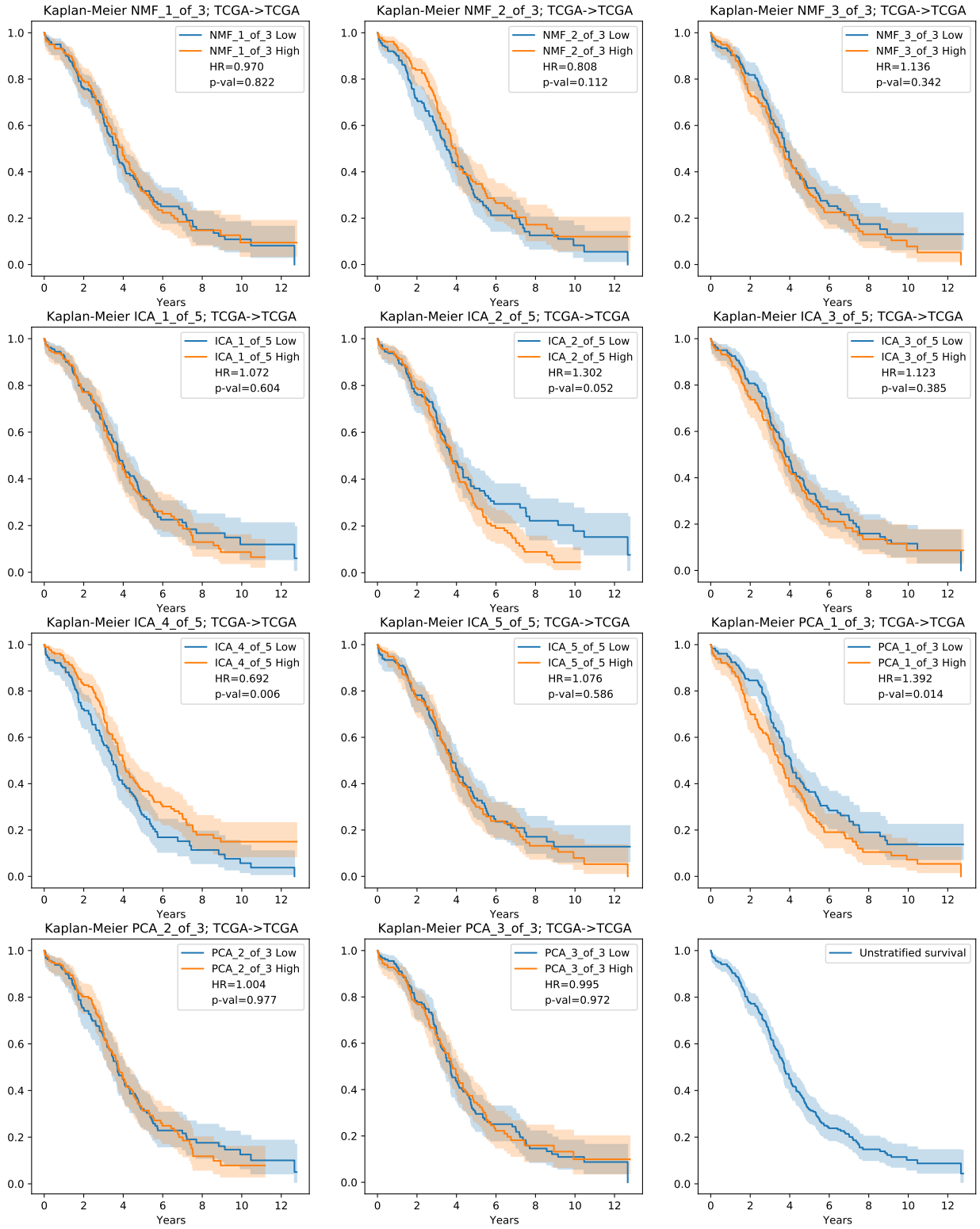


Figure 14: Kaplan-Meier plots for each metasample component, stratified at the median value, for TCGA $\rightarrow$ TCGA for overall survival (OS) case. Hazard ratio and p-value is shown for each case. The final plot (bottom right) is unstratified overall survival

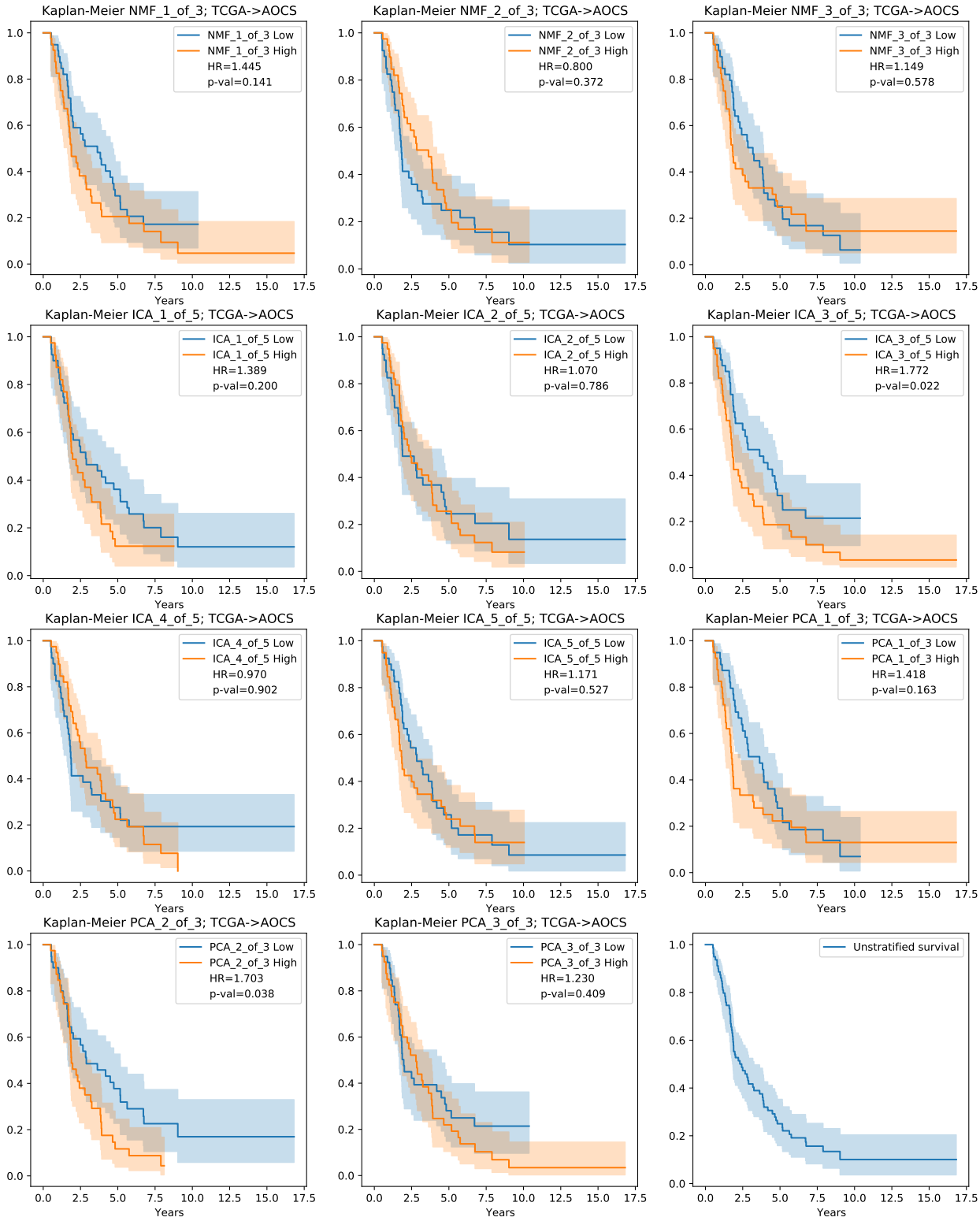


Figure 15: Kaplan-Meier plots for TCGA → TCGA for overall survival (OS).

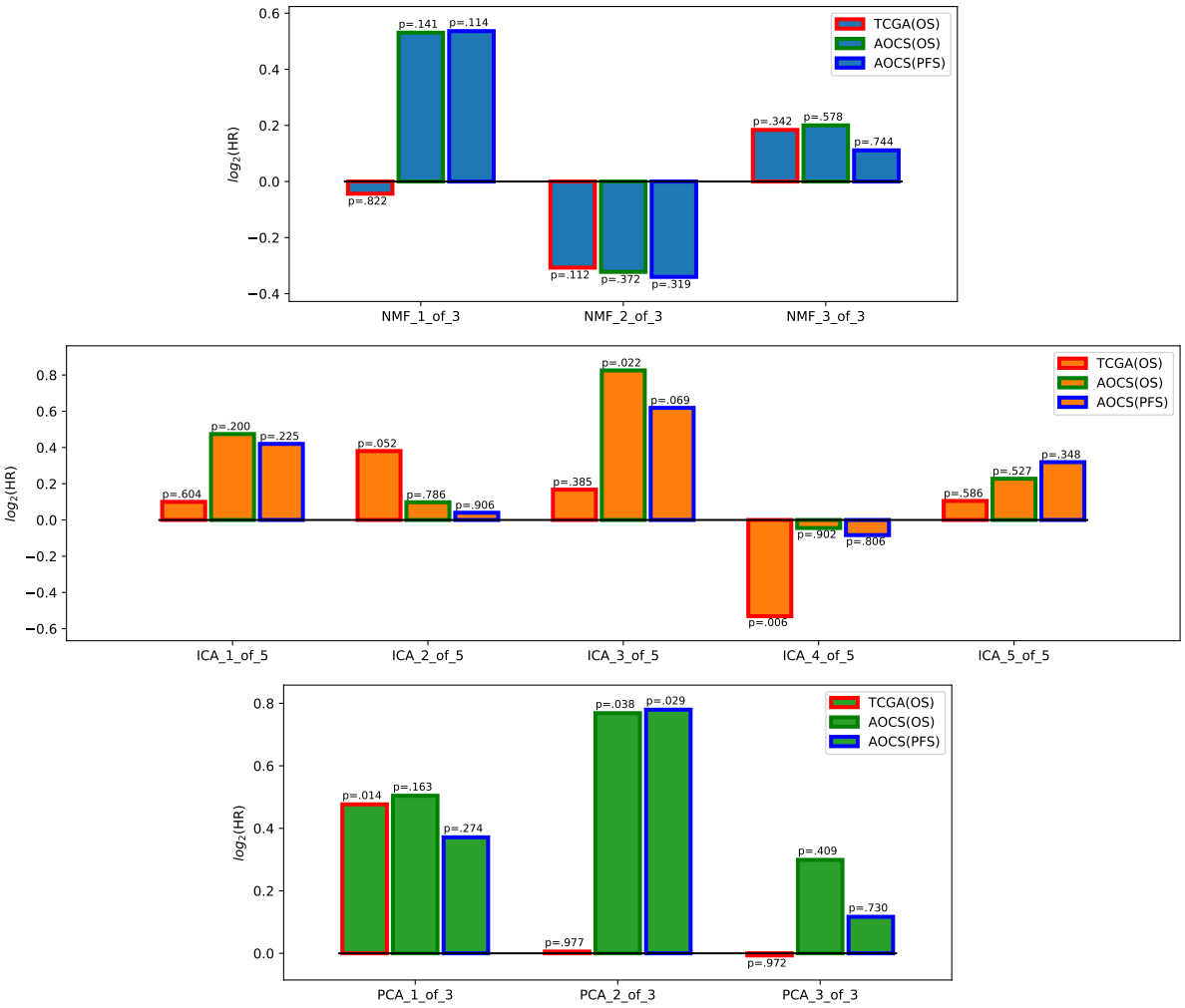


Figure 16: Visual summary of survival analysis as applied to TCGA(OS), AOCS(OS) and AOCS(PFS). Plots are divided by factorization method. Bar heights show $\log_2(\text{HR})$ with p-value also shown.

3.5 Metasample correlation with genomic features

From the grid of scatter plots shown in figure 17 it can be seen there are some significant correlations. The sense (sign) of the correlation is arbitrary – particularly for ICA and PCA – recall normalisation w.r.t $180°$ was applied.

The significant correlations are summarised in table 1. WGD and Cellularity show correlation with at least one component from all three methods and indeed three components correlating with WGD also correlate with Cellularity and in the same sense. This might suggest simply that WGD and Cellularity are themselves correlated, but this is not so: $r = -0.14, p = 0.201$ for Point-Biserial correlation (data not shown). HR Detect correlates with two NMF and one ICA component, not intersecting with those above, suggesting distinct biological processes are involved.

The heatmap dendrogram (figure 18) shows close correlation between metasamples pairs (NMF-1-of-3, ICA-5-of-5), (ICA-1-of-5, PCA-2-of-3) and (NMF-1-of-3, ICA-3-of-5).

Genomic feature	Significant correlations
WGD	NMF-1-of-3, ICA-5-of-5, PCA-3-of-3
Cellularity	NMF-1-of-3, ICA-3-of-5, ICA-5-of-5, PCA-3-of-3
HR Detect	NMF-2-of-3, NMF-3-of-3, ICA-2-of-5
Mutational load	None
CNV load	None
SV load	NMF-2-of-3

Table 1: Summary of metasample components showing a significant correlation with each genomic feature.

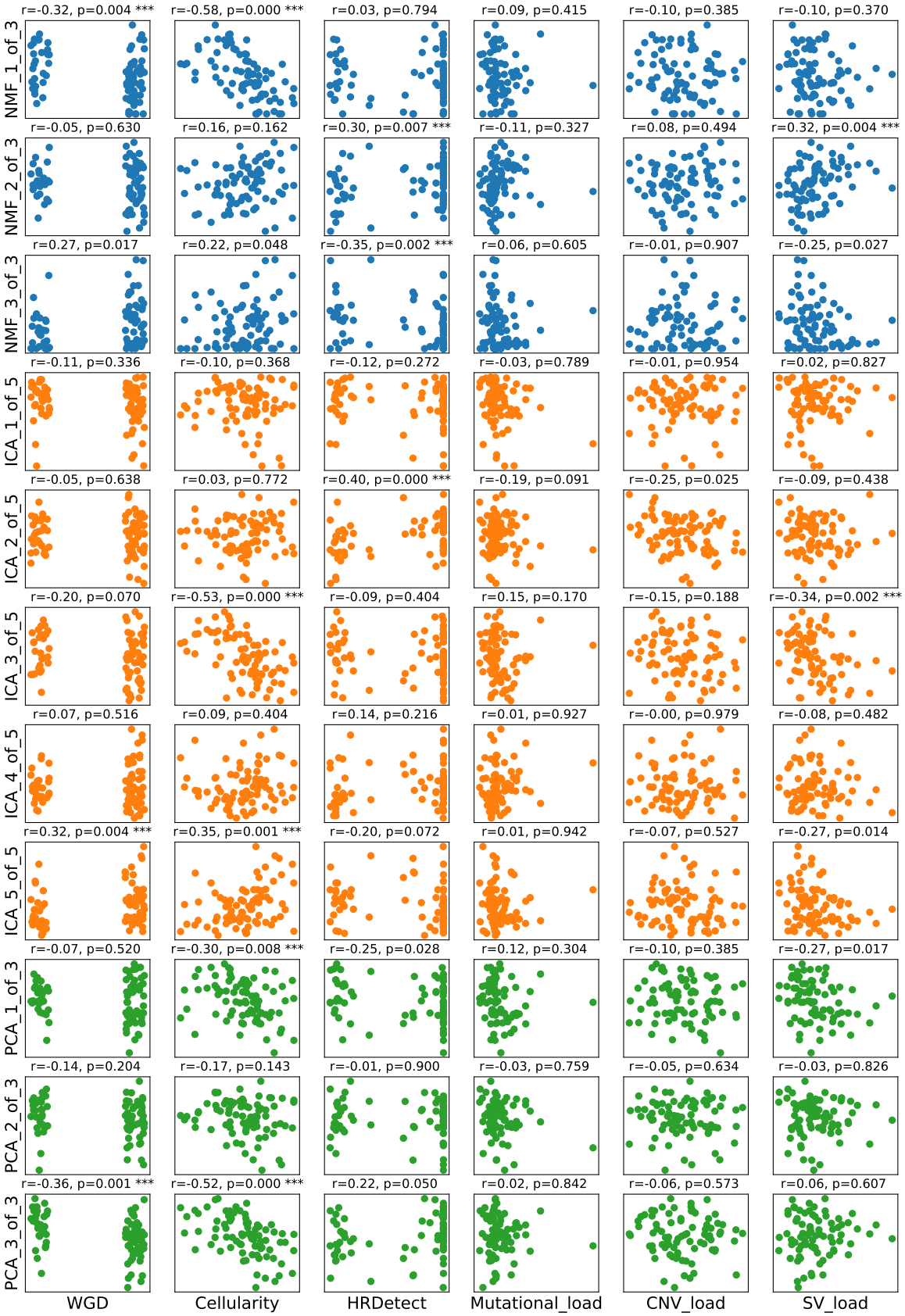


Figure 17: Grid of scatter plots to visualise the correlation between metasamples and genomic features. In the case of WGD which has binary values (0, 1), jitter is applied for visual effect only. Above each plot is shown a correlation coefficient r and associated p-value, highlighted with '***' where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

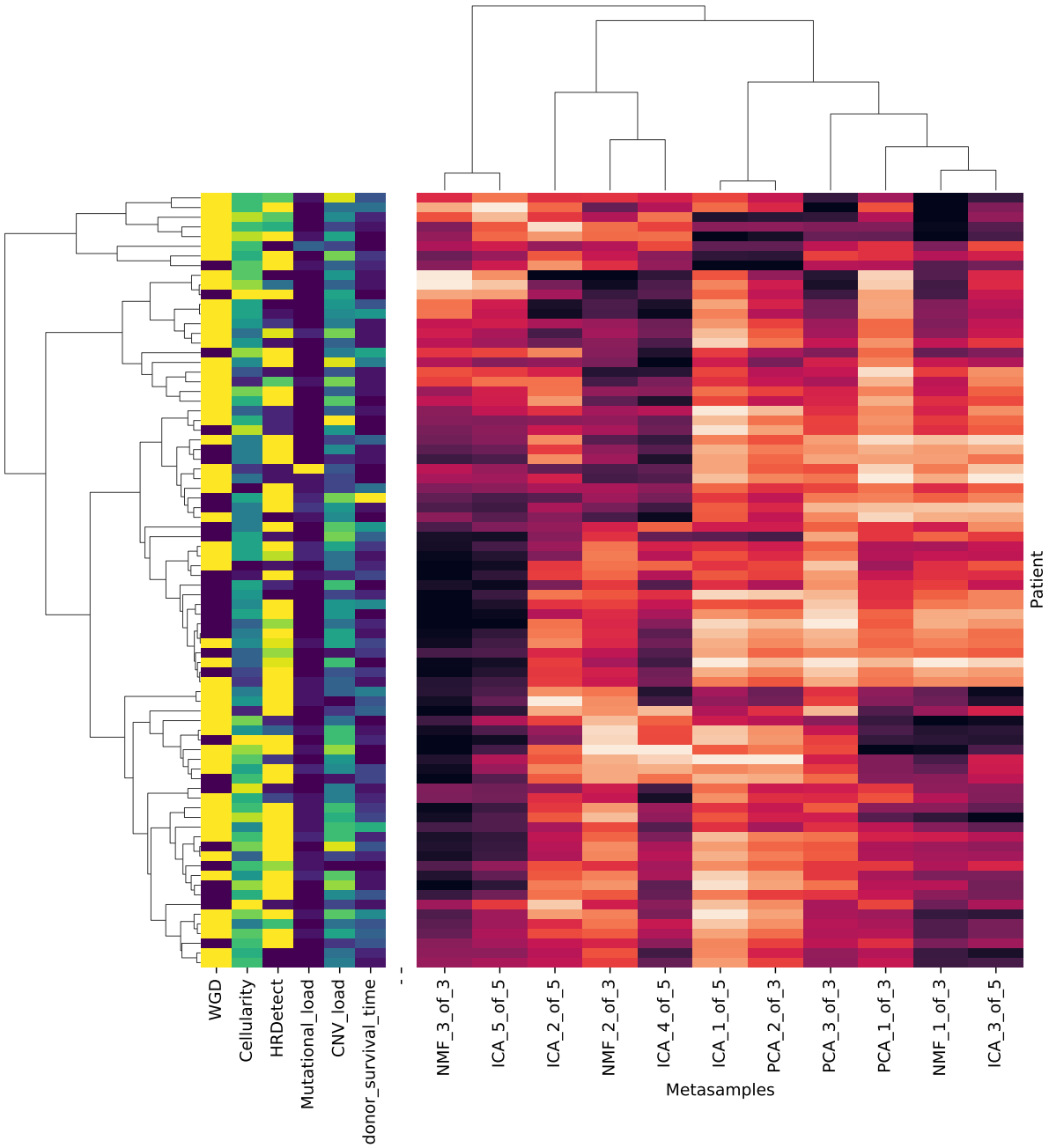


Figure 18: Heatmap of the metasamples matrix for the AOCS dataset, computed from metagenes factorized from the TCGA dataset. Several patient metadata variables are shown in columns to the left.

4 Discussion

One valuable lesson from this work is the importance of accounting for sampling error, by bootstrapping, 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 [25, 23, 35], but most neglect sampling error which we have shown to be far more significant. Cantini *et al* [36] 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 the current work, BIODICA was used, and was found to 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.

These considerations lead to the selection of much lower ranks (fewer metagenes) than would otherwise be the case.

TODO: much more discussion!

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 gene enrichment analysis – replacing the current arbitrary 3 SD from the mean rule.

2. **Discovering batch effects**. It has been claimed (e.g. [20]) that matrix factorization methods are an effective means of identifying and removing batch effects. This could be explored by horizontally concatenating the TCGA and AOCS datasets (after gene set intersection), producing an $n = 80 + 274 = 354$ dataset with a substantial batch artefact. Factorizations of this combined dataset should be more robust because of the larger n. Does each of NMF, ICA and PCA naturally produce a component which correlates with batch?

3. **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.

4. **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.

5. **Contrasting gene expression patterns between cancers**. The current work set out to find patters of gene expression in HGSOC. Yet it is hard to say the highlighted patterns are specific to HGSOC, or generic to all cancers. One way of teasing out cancer specific patterns might be to take a similar approach to the previous item, but instead of looking for metagenes which are consistent between datasets, find those which are distinct. That said, there will likely be many more sophisticated approached in the literature, given the clinical importance of the topic.

6. **Systematic comparison with published gene expression patterns**. In analysing the meaning of each metagene above, some tentative links were made with the research literature. But this was anecdotal and frankly not particularly scientific; searching on PubMed for GO terms associated with HGSOC and looking at one or two hits! 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

TODO

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, S. Fereday, K. Nones, P. Cowin, K. Alsop, P. J. Bailey, K. S. Kassahn, F. Newell, M. C. Quinn, S. Kazakoff, K. Quek, C. Wilhelm-Benartzi, E. Curry, H. S. Leong, A. Hamilton, L. Mileshkin, G. Au-Yeung, C. Kennedy, J. Hung, Y. E. Chiew, P. Harnett, M. Friedlander, M. Quinn, J. Pyman, S. Cordner, P. O'Brien, J. Leditschke, G. Young, K. Strachan, P. Waring, W. Azar, C. Mitchell, N. Traficante, J. Hendley, H. Thorne, M. Shackleton, D. K. Miller, G. M. Arnau, R. W. Tothill, T. P. Holloway, T. Semple, I. Harliwong, C. Nourse, E. Nourbakhsh, S. Manning, S. Idrisoglu, T. J. Bruxner, A. N. Christ, B. Poudel, O. Holmes, M. Anderson, C. Leonard, A. Lonie, N. Hall, S. Wood, D. F. Taylor, Q. Xu, J. Lynn Fink, N. Waddell, R. Drapkin, E. Stronach, H. Gabra, R. Brown, A. Jewell, S. H. Nagaraj, E. Markham, P. J. Wilson, J. Ellul, O. McNally, M. A. Doyle, R. Vedururu, C. Stewart, E. Lengyel, J. V. Pearson, N. Waddell, A. Defazio, S. M. Grimmond, and D. D. Bowtell, "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] 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.

[7] K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, J. Zhang, C. Kandoth, R. Akbani, H. Shen,

L. Omberg, A. Chu, A. A. Margolin, L. J. Van't Veer, N. Lopez-Bigas, P. W. Laird, B. J. Raphael, L. Ding, A. G. Robertson, L. A. Byers, G. B. Mills, J. N. Weinstein, C. Van Waes, Z. Chen, E. A. Collisson, C. C. Benz, C. M. Perou, J. M. Stuart, R. Abbott, S. Abbott, B. A. Aksoy, K. Aldape, A. Ally, S. Amin, D. Anastassiou, J. T. Auman, K. A. Baggerly, M. Balasundaram, S. Balu, S. B. Baylin, S. C. Benz, B. P. Berman, B. Bernard, A. S. Bhatt, I. Birol, A. D. Black, T. Bodenheimer, M. S. Bootwalla, J. Bowen, R. Bressler, C. A. Bristow, A. N. Brooks, B. Broom, E. Buda, R. Burton, Y. S. Butterfield, D. Carlin, S. L. Carter, T. D. Casasent, K. Chang, S. Chanock, L. Chin, D. Y. Cho, J. Cho, E. Chuah, H. J. E. Chun, K. Cibulskis, G. Ciriello, J. Cleland, M. Cline, B. Craft, C. J. Creighton, L. Danilova, T. Davidsen, C. Davis, N. D. Dees, K. Delehaunty, J. A. Demchok, N. Dhalla, D. DiCara, H. Dinh, J. R. Dobson, D. Dodda, H. V. Doddapaneni, L. Donehower, D. J. Dooling, G. Dresdner, J. Drummond, A. Eakin, M. Edgerton, J. M. Eldred, G. Eley, K. Ellrott, C. Fan, S. Fei, I. Felau, S. Frazer, S. S. Freeman, J. Frick, C. C. Fronick, L. L. Fulton, R. Fulton, S. B. Gabriel, J. Gao, J. M. Gastier-Foster, N. Gehlenborg, M. George, G. Getz, R. Gibbs, M. Goldman, A. Gonzalez-Perez, B. Gross, R. Guin, P. Gunaratne, A. Hadjipanayis, M. P. Hamilton, S. R. Hamilton, L. Han, Y. Han, H. A. Harper, P. Haseley, D. Haussler, D. N. Hayes, D. I. Heiman, E. Helman, C. Helsel, S. M. Herbrich, J. G. Herman, T. Hinoue, C. Hirst, M. Hirst, R. A. Holt, A. P. Hoyle, L. Iype, A. Jacobsen, S. R. Jeffreys, M. A. Jensen, C. D. Jones, S. J. Jones, Z. Ju, J. Jung, A. Kahles, A. Kahn, J. Kalicki-Veizer, D. Kalra, K. L. Kanchi, D. W. Kane, H. Kim, J. Kim, T. Knijnenburg, D. C. Koboldt, C. Kovar, R. Kramer, R. Kreisberg, R. Kucherlapati, M. Ladanyi, E. S. Lander, D. E. Larson, M. S. Lawrence, D. Lee, E. Lee, S. Lee, W. Lee, K. V. Lehmann, K. Leinonen, K. M. Leraas, S. Lerner, D. A. Levine, L. Lewis, T. J. Ley, H. I. Li, J. Li, W. Li, H. Liang, T. M. Lichtenberg, J. Lin, L. Lin, P. Lin, W. Liu, Y. Liu, Y. Liu, P. L. Lorenzi, C. Lu, Y. Lu, L. J. Luquette, S. Ma, V. J. Magrini, H. S. Mahadeshwar, E. R. Mardis, M. A. Marra, M. Mayo, C. McAllister, S. E. McGuire, J. F. McMichael, J. Melott, S. Meng, M. Meyerson, P. A. Mieczkowski, C. A. Miller, M. L. Miller, M. Miller, R. A. Moore, M. Morgan, D. Morton, L. E. Mose, A. J. Mungall, D. Muzny, L. Nguyen, M. S. Noble, H. Noushmehr, M. O'Laughlin, A. I. Ojesina, T. H. O. Yang, B. Ozenberger, A. Pantazi, M. Parfenov, P. J. Park, J. S. Parker, E. Paull, C. S. Pedamallu, T. Pihl, C. Pohl, D. Pot, A. Protopopov, T. Przytycka, A. Radenbaugh, N. C. Ramirez, R. Ramirez, G. Rätsch, J. Reid, X. Ren, B. Reva, S. M. Reynolds, S. K. Rhie, J. Roach, H. Rovira, M. Ryan, G. Saksena, S. Salama, C. Sander, N. Santoso, J. E. Schein, H. Schmidt, N. Schultz, S. E. Schumacher, J. Seidman, Y. Senbabaoglu, S. Seth, S. Sharpe, R. Shen, M. Sheth, Y. Shi, I. Shmulevich, G. O. Silva, J. V. Simons, R. Sinha, P. Sipahimalani, S. M. Smith, H. J. Sofia, A. Sokolov, M. G. Soloway, X. Song, C. Sougnez, P. Spellman, L. Staudt, C. Stewart, P. Stojanov, X. Su, S. O. Sumer, Y. Sun, T. Swatloski,

B. Tabak, A. Tam, D. Tan, J. Tang, R. Tarnuzzer, B. S. Taylor, N. Thiessen, V. Thorsson, T. Triche, D. J. Van Den Berg, F. Vandin, R. J. Varhol, C. J. Vaske, U. Veluvolu, R. Verhaak, D. Voet, J. Walker, J. W. Wallis, P. Waltman, Y. Wan, M. Wang, W. Wang, Z. Wang, S. Waring, N. Weinhold, D. J. Weisenberger, M. C. Wendl, D. Wheeler, M. D. Wilkerson, R. K. Wilson, L. Wise, A. Wong, C. J. Wu, C. C. Wu, H. T. Wu, J. Wu, T. Wylie, L. Xi, R. Xi, Z. Xia, A. W. Xu, D. Yang, L. Yang, L. Yang, Y. Yang, J. Yao, R. Yao, K. Ye, K. Yoshihara, Y. Yuan, A. K. Yung, T. Zack, D. Zeng, J. C. Zenklusen, H. Zhang, J. Zhang, N. Zhang, Q. Zhang, W. Zhang, W. Zhao, S. Zheng, J. Zhu, E. Zmuda, and L. Zou, "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, vol. 158, no. 4, pp. 929–944, 2014.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] R. G. Verhaak, P. Tamayo, J. Y. Yang, D. Hubbard, H. Zhang, C. J. Creighton, S. Fereday, M. Lawrence, S. L. Carter, C. H. Mermel, A. D. Kostic, D. Etemadmoghadam, G. Saksena, K. Cibulskis, S. Duraisamy, K. Levanon, C. Sougnez, A. Tsherniak, S. Gomez, R. Onofrio, S. Gabriel, L. Chin, N. Zhang, P. T. Spellman, Y. Zhang, R. Akbani, K. A. Hoadley, A. Kahn, M. Köbel, D. Huntsman, R. A. Soslow, A. Defazio, M. J. Birrer, J. W. Gray, J. N. Weinstein, D. D. Bowtell, R. Drapkin, J. P. Mesirov, G. Getz, D. A. Levine, and M. Meyerson, "Prognostically relevant gene signatures of high-grade serous ovarian carcinoma," *Journal of Clinical Investigation*, vol. 123, pp. 517–525, jan 2013.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[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] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

6 Appendices

6.1 Additional figures and plots

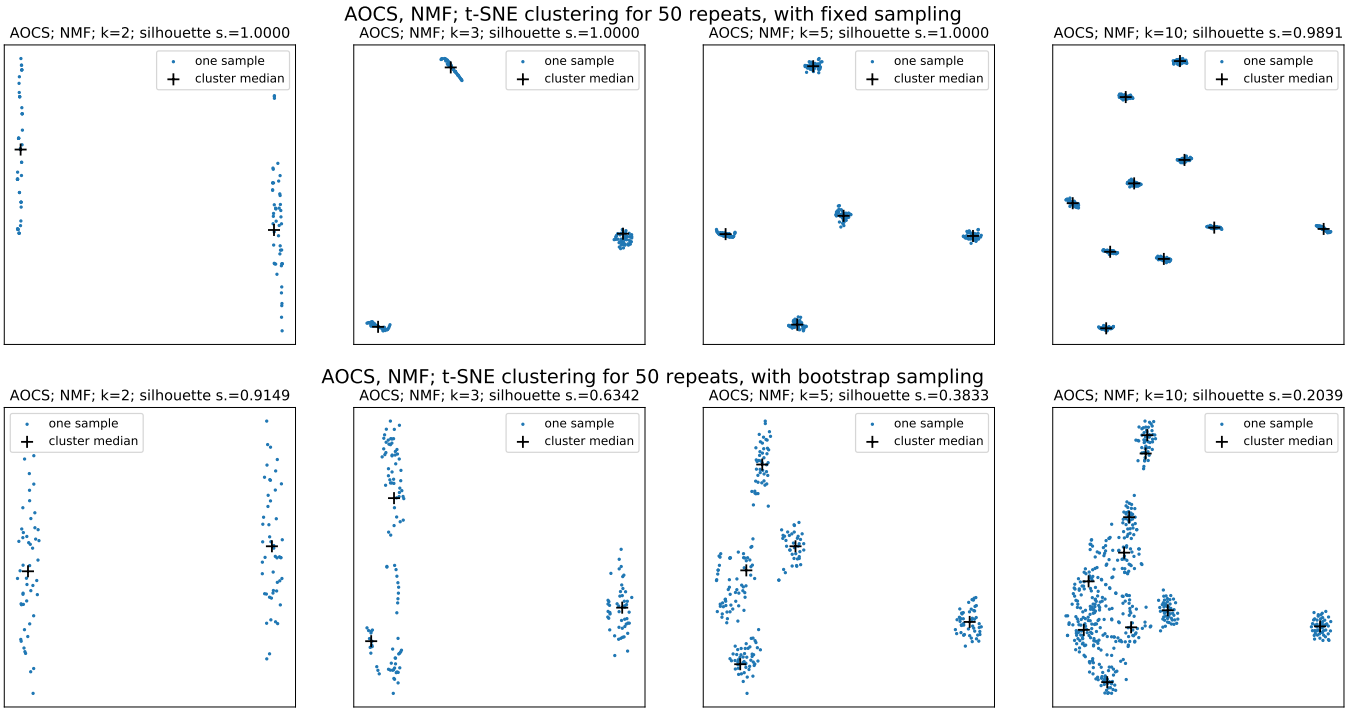


Figure 19: Clustering of metagenes from NMF facorizations on the AOCS dataset, comparing fixed and bootstrap sampling

TODO - I’ll likely move the AOCS survivial plots here, and perhaps some of the GO lineages.

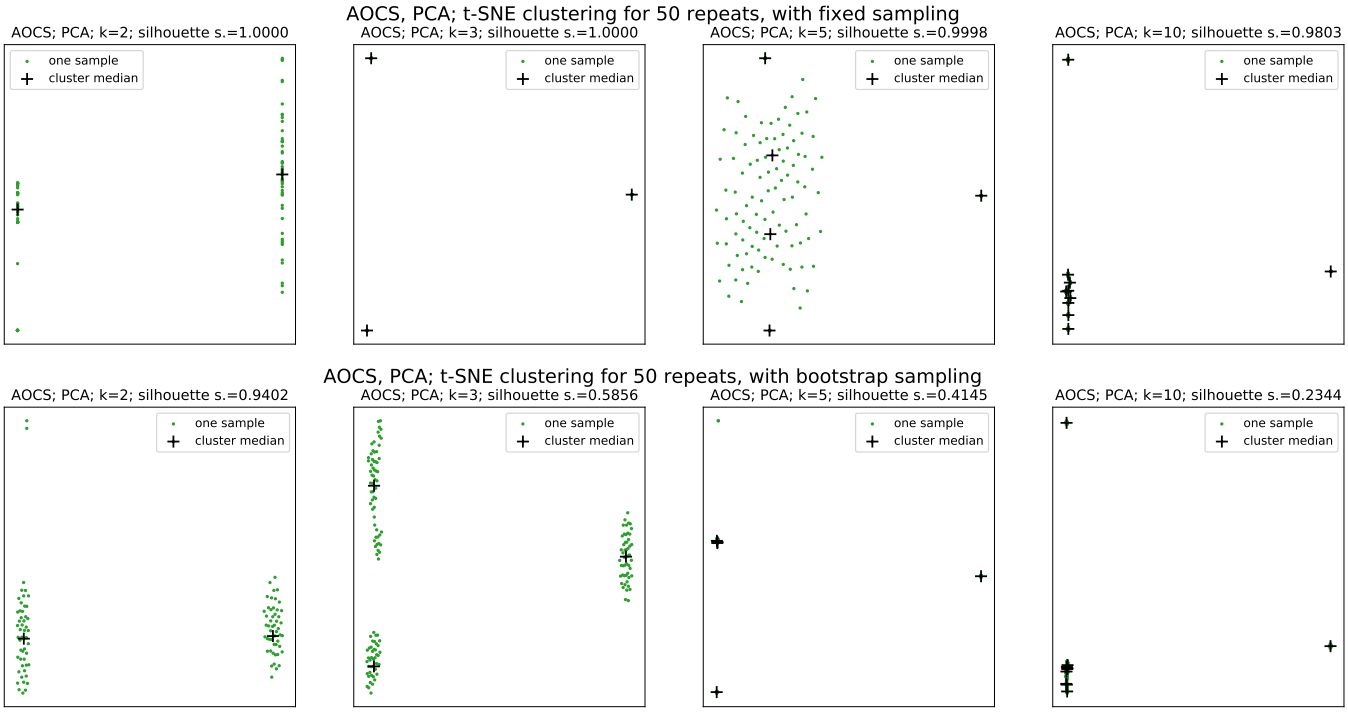


Figure 20: Clustering of metagenes from PCA facorizations on the AOCS dataset, comparing fixed and bootstrap sampling

6.2 Gene enrichment raw results

TODO

6.3 Software libraries and versions

TODO

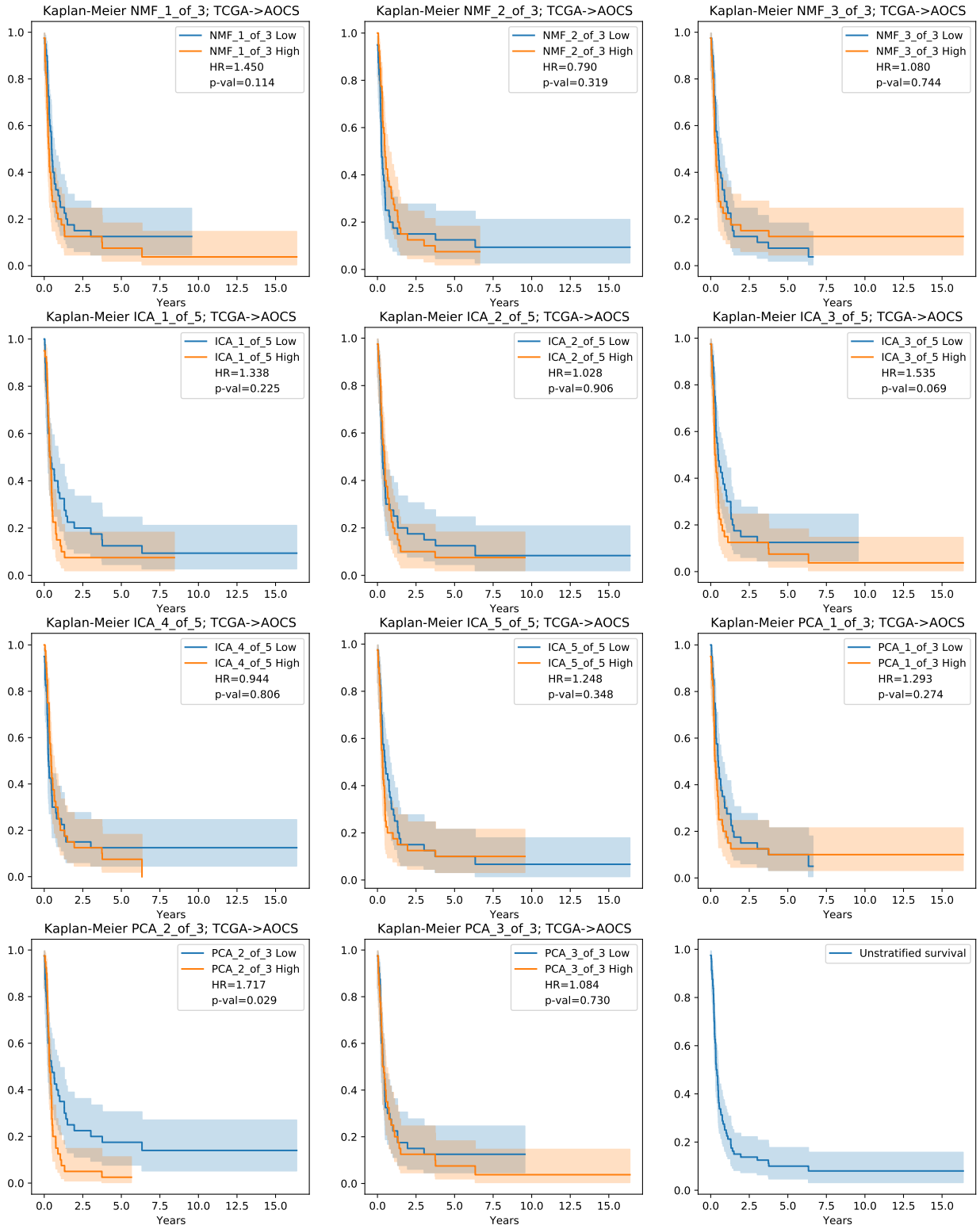


Figure 21: Kaplan-Meier plots for TCGA → TCGA for progression free survival (PFS).

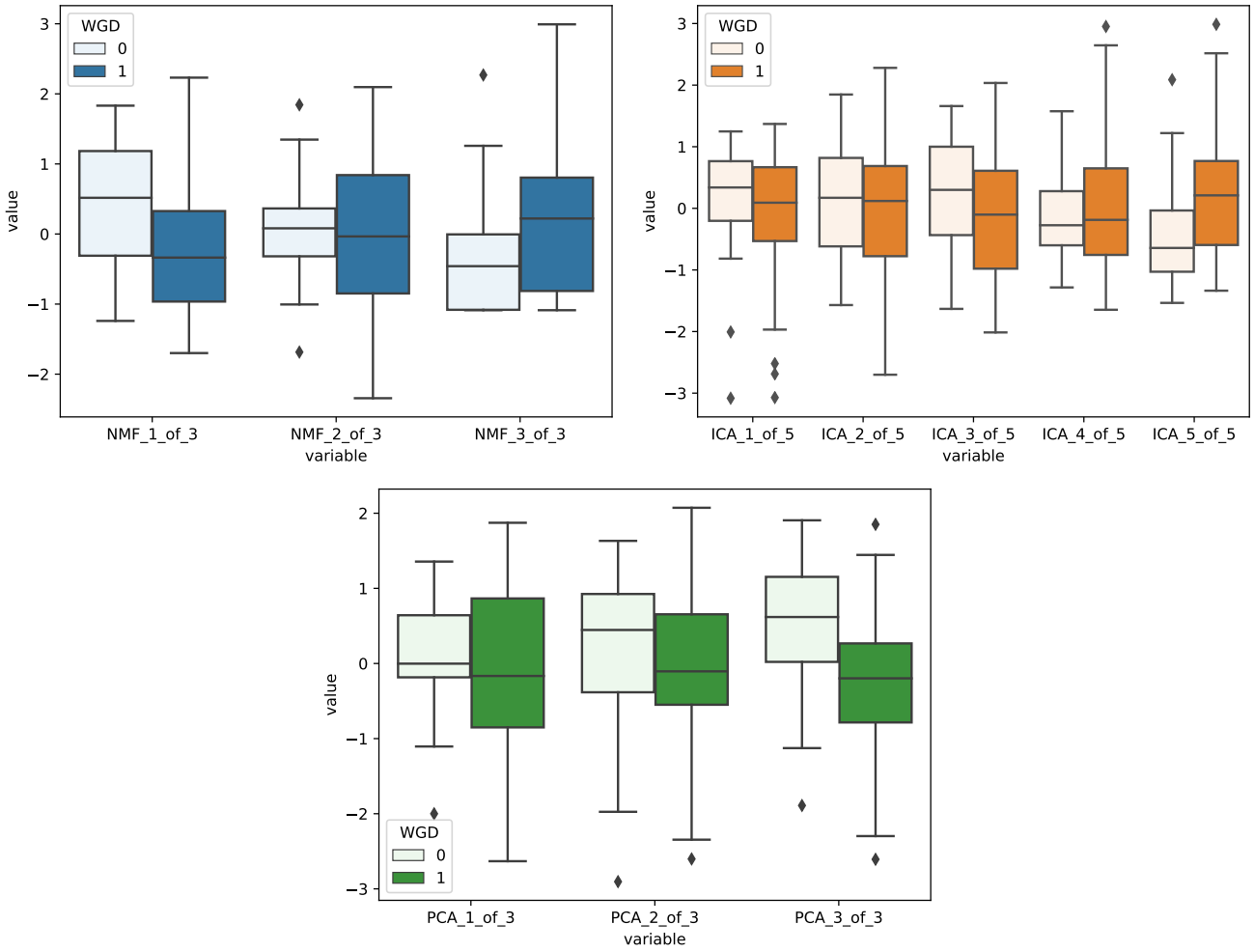


Figure 22: Boxplots showing relationship between metasamples (derived from TCGA metagenes, applied to AOCS data) and whole genome doubling (WGD) status. Metasamples from NMF, ICA and PCA are shown separately. To allow convenient visualisation, each metasample is normalised to zero mean unit variance, hence between metasample variation is removed.