BMC Bioinformatics

moGSA: integrative single sample gene-set analysis of multiple omics data

--Manuscript Draft--

Manuscript Number:	BINF-D-15-00655R1
Full Title:	moGSA: integrative single sample gene-set analysis of multiple omics data
Article Type:	Methodology article
Abstract:	Background: The increasing availability of multi-omics datasets has created an opportunity to understand how different biological pathways and molecules interact to cause disease. However, there is a lack of analysis methods that can integrate and interpret multiple experimental and molecular data types measured over the same set of samples. Result: To address this challenge, we introduce moGSA, a multivariate single sample gene-set analysis method. It uses multivariate latent variable decomposition to discover correlated global variance structure across datasets and calculates an integrated gene set enrichment score using the most informative features in each data type. Integrating multiple diverse sources of data, reduces the impact of missing or unreliable information in any single data type, and may increase the power to discover subtle changes in gene-sets. We show that integrative analysis with moGSA outperforms existing single sample GSA methods on simulated data. We apply moGSA to two studies with real data. First we discover similarities and differences in mRNA, protein and phosphorylation profiles of induced pluripotent and embryonic stem cell lines. Secondly we report that three molecular subtypes are robustly discovered when copy number variation and mRNA profiling data of 308 bladder cancers from The Cancer Genome Atlas are integrated using moGSA. Our method provides positive or negative gene-set scores (with p-values) of each gene set in each sample. We demonstrate how to assess the influence of each data type or gene to a moGSA gene set score. With moGSA, there is no requirement to filter data to the intersect of features. All molecular features on all platforms may be included in the analysis. Conclusion: moGSA provides a powerful yet simple tool to perform integrated simple sample gene-set analysis. Its latent variable approach is fundamentally different to existing single sample GSA approaches. It is an attractive approach for data integration and is particularly suited to integrated cluster or molecular subtype discovery. It is available in the Bioconductor R package "mogsa".
Response to Reviewers:	See uploaded file response2reviewers_final.docx in which we provide a point by point response to all concerns of the two reviewers. Our comments are in red. This document and the cover letter are enclosed with the manuscript file.

Technische Universität München

Technische Universität München . Lehrstuhl für Proteomik und Bioanalytik
Emil Erlenmeyer Forum 5 . 85354 Freising. Germany

Wissenschaftszentrum
Weihenstephan für Ernährung,
Landnutzung und Umwelt

Lehrstuhl für Proteomik und
Bioanalytik

Dr. Amin Moghaddas Gholami

Emil Erlenmeyer Forum 5
85354 Freising
Germany

Tel +49.8161.71.2065
Fax +49.8161.71.5931

amin@tum.de
www.wzw.tum.de/proteomics

Date: 2016.01.25

Editorial Office
BMC Bioinformatics

Freising, 25 January 2016

Revised manuscript submission
Manuscript ID: BINF-D-15-00655

Dear BMC Bioinformatics Editorial Office,

Today, we are submitting our revised manuscript to of our revised manuscript, entitled

moGSA: a multivariate approach for integrative gene-set analysis of multiple omics data

We thank the reviewers for their valuable suggestions. We have carefully considered all the points raised and provide a detailed point to point response. We have performed a number of new experiments, new data analysis, modified main figures and include several new supplementary figures. Moreover the text of the manuscript is considerably restructured and edited for clarity.

Collectively, the revision process has made the manuscript much stronger and we hope the reviewers now find it acceptable for publication in BMC bioinformatics.

On behalf of all authors and with best wishes,

(Amin Moghaddas Gholami)

(Aedín Culhane)

We thank the reviewers for their time, insight and valuable suggestions. We have carefully considered all of their concerns and include our point by point response below.

To address reviewer concerns, we have considerably revised the manuscript. We have performed several new experiments and have added Figure 4B and additional supplementary figures. In addition, the manuscript has been extensively edited to correct grammatical errors and improve the clarity of the manuscript.

We are grateful to the reviewers, as we feel their critic has made the manuscript substantially stronger. We hope the reviewers now find it acceptable for publication in BMC bioinformatics.

#Reviewer 1:

moGSA: a multivariate approach for integrative gene-set analysis of multiple omics data

The authors propose a gene-set analysis when integrating multiple 'omics data sets using Multiple Factor Analysis (MFA) coupled with the inclusion of gene-set annotation matrices which are incidence matrices indicating whether an 'omic feature belong to a given gene-set in a particular 'omic data set. Dimension reduction is achieved by calculating gene-set scores as linear combinations of factors weighted by the gene-set space weight. Clustering analysis on the observation is performed using consensus clustering. The approach is compared to existing single 'omics gene-set analysis approaches on a simulation study, and further applied to two multi 'omics case studies.

*General comments

The methodology is interesting and has promising potential in the near future, as it aligns with current research in the field. The manuscript however needs some serious polishing, editing and English proofread - I cannot list extensively the sentences that would need to be fixed as the review would be too lengthy. I also have specific comments in order to improve the clarity of the manuscript, but also further validate the approach.

We are pleased that the reviewer finds the method interesting. The manuscript has been extensively edited and polished by a native English speaker. Moreover additional analyses were performed to validate the moGSA approach.

*Specific comments for revision

*Major

1.Results section moGSA outperforms single sample GSA methods: this section needs to be extended as I do not think it addresses all methodological aspects.

a. The description of the simulation study is repetitive with he Methods section. I would have liked to see the effect of the number of features in each gene-set since the method proposes a gene-set normalisation to account for various number of genes in a gene-set (p19 l4).

We have edited the manuscript to reduce repetition, but since the methods are at the end of the manuscript, we included a brief description of the simulated data within the result section.

The effect of the number of differentially expressed features in gene-sets was explored. We examine the effect of having 5, 10 and 25 differentially expressed genes in each gene-set of length 50. In each simulation we randomly select the differentially expressed subset.

All simulations were performed on gene-sets of length 50. Since the gene-set length was the same in all simulated analyses, normalization would not affect the score. The normalized Gene Set Score is the mean expression of genes in a gene-set. We use the mean, rather than the sum of genes because the range of the latter would be dependent on the number of features in a gene set (length), would hinder comparison of GSS between gene sets. We added two new supplementary figures and S22, which show non-normalized GSS of gene-sets in figures 3 and 5 which were computed using the gene expression sum rather than the gene expression mean.

b. In addition, I would suggest to report the first singular value used in A to weight each data set (p16, l16). Would this information be useful to evaluate the degree of information contained in each data set related to the other (as mentioned in the discussion p14 l22.

Thank you for this suggestion, this in now indicated in the manuscript.

For the stem cell data, we now state on line 193-196

"The three datasets contributed similar variance in the integrated analysis, as indicated by weighting of each dataset in MFA. The first eigenvalue (square of singular values) of each PCA were 0.24, 0.26 and 0.26 for the transcriptome, proteome and PhosoProteome dataset respectively."

We also include a scree plot showing the eigen values of the first 3 PCs in Supplementary Figure 4.

For the BLCA data:

The first eigenvalue (square of singular value) of a PCA of BLCA mRNA and CNV data are `0.0004` and `0.0003` respectively. A scree plot of the first 10 eigenvalues of PCA of each dataset are shown in supplementary figure S7.

c. In the methods section, I would have liked to see more details about the other competitor approaches, GSVA and ssGSEA. For those I would suggest to add another comparison where the data sets are weighted by their first singular value to be fairer.

Whilst we compare moGSA to GSVA and ssGSEA, these cannot be considered true "competitor approaches" to moGSA as there were designed for analysis of a single data matrix. moGSA performs an integrated gene set analysis over many datasets.

We now include a new analysis in which repeated all analyses that use simulated data, with datasets that were weighted by their first eigenvalue. This result is shown in Supplementary Figure S3. The

result are consistent with our previous results (Figure 2) and confirm that moGSA also outperforms ssGSEA and GSVA.

In the manuscript we now state (line 172-175): "Finally, since MFA weights input matrices by their first singular value before moGSA, we examined the effect of data set weighting on the other methods, but found moGSA still outperformed ssGSEA and GSVA when data matrices of the triplet were weighted before concatenation (Figure S3)."

d. Finally, this section does not address the effect of the choice of the # of PC to choose in MFA. How much information would be missed if the user selected a smaller number of PCs?

Yes, we thank the reviewer for this important point. Reducing the number of PCs may affect results. If too few PCs are selected, the user may miss information. However adding additional PCs may not add more information (new gene-sets) once the first few PCs with highest eigenvalues are included. The optimal number of PCs will depend on the variance of the datasets.

We study this effect using simulated data (Figure 2C and 2D).

In addition, we now include an experiment in which we vary the number of PCs from 1 to 12 to show the effect of component number on the results of the BLCA data (Figure 4B, Supplementary Figure S9). The result of this experiment is described on Lines 250-263:

"In a typical analysis, we use a scree plot to select the number of components. The scree plot indicated that five components should capture sufficient variance for input to moGSA. We confirmed that this was the optimal number of components as input to moGSA, in the following experiment. We performed moGSA on the BLCA mRNA gene expression and CNV data (n=308) with a range of components ranged from 1 to 12. For each gene-set in the GSS matrix, gene-sets were ranked by the number of tumors in which they were significantly regulated (either positive or negative GSS, $p<0.05$), such that gene-sets that were significant in most tumors had highest rank. The distribution of the number of tumors in which gene-sets were significant at $p<0.05$, $p<0.01$, and $p<0.001$ is shown Figure S7. No gene-set was significant in all 308 tumors and most gene-sets were insignificant in all tumors (Figure S7). For $p<0.05$, we examined the 10, 20, 40, 100, 200, 500 and 1000 highest ranked gene-sets and examined the stability of gene-set ranking when additional components were included (Figure S8). Increasing the number of components (from 1 to 5) increased the stability of gene set lists, however there was little additional gain after 5 components (Figure S9). Among the top 100 ranked genesets, few new gene-sets were identified after 5 component (Figure 4B)"

In addition, we now include a paragraph in the discussion section (lines 402-417):

"The number of components is an important input parameter to consider when applying moGSA to gene-set analysis or cluster discovery. Similar to PCA, the optimal number of MFA components may be assessed by examining the variance associated with the each component. The first component will

capture most variance and the variance associated with subsequent component decreases monotonically. Scree plots (Figure 2C, 4A) may be used to visualize if there is an elbow point in the eigenvalues, allowing one to select the components before the elbow point. Alternatively one may select the number of components that capture a certain proportion of variance (50%, 70% etc). In addition, one may include components that are of biological interest. For example, in the iPS ES example, there is a clear biological meaning in the third component (ES vs iPS cell line). In analysis of the BLCA data, we examined a range of components (1-12), and show that there is little gain of information once a minimum number of components with high variance are included (Figure 4B). In addition, the variance of retained components should not be dominated by one or few of the datasets. To facilitate the biological interpretation of components, the GSS could be decomposed with regard to components. In the BLCA example, the second and forth component are largely contributed by CNV, whereas mRNA is more important in defining the third and fifth components. Including five components ensured that both datasets contributed relatively similar variance to the global variance."

2. Results section, TCGA data. Some repeat that are already mentioned in the Methods section.

We have removed repeated sentences in the methods and results section.

a. 'it is hard to analyze the gene-set score of individual patient' Would boxplots help w.r.t to clusters?

We agree that the boxplot is a better visualization when the number of sample is large. Therefore, we include a new supplementary figure that shows the gene sets scores w.r.t cluster. See figure S18.

b. Some references are missing

We have reviewed the manuscript and now include these missing citations.

c. Justify the cutoff 200.308 to select the gene sets. Give pointers on how a user would be able to choose that cutoff.

moGSA is a single sample GSA methods, which provides a GSS and p value for each gene-set in each patients. Whilst we could include all gene-sets that are significant (at p<0.05, p<0.01, p<0.001) in any patients, in reality a gene-set that is significant in only 1 patient is unlikely to be biologically informative.

In this case, we wished to identify the few gene-sets that were significantly regulated (p<0.05) is most patients, and the choice of 200 /308 was somewhat arbitrary. The cluster analysis indicated that three subtypes existed in the BLCA data, each of which had between 57 and 148 patients. So a threshold of 200, likely included gene-sets that were up or down in 2 or more subtypes.

We now include a brief discussion and mentioned alternative criteria that could be used to subset significant gene-sets.

Line 294-299:

"To further characterize BLCA, we focused on gene-sets that were differentially regulated in most patients. There were 73 gene-sets that were significantly regulated (positive or negative GSS, p value<0.05) in 200 or more of the 308 patients (Table S3 and Figure S17). Alternatively a lower cutoff would include more gene-sets that are regulated fewer tumors, fewer gene-set could be selected using a lower p-value (p<0.01, 0.001) or a supervised analysis could be used to select GSS that most discriminate groups of tumors. "

d. p11: what does the n notation stand for (used in the Methods section for the number of observations)

We removed "n", which may be confused with the number of observations.

e. l12 p12: how is significance of a GSS assessed? (which test?)

The statistical inference is based on the central limited theorem, which was previously included in the method section 'statistical inferential aspect'. To better highlight this important aspect of the method, we have renamed this section to "Evaluation of significance of gene-set scores" and placed it directly after the section that explains the calculation of gene set score. (line 504)

f. Figure 4A: the percentage of variance would be more appropriate

We agree and have updated figure 4A to shown the percentage of variance.

g. Figure 4E: most correlations are pretty low (0.4). I assume the authors used cor.test in R which assesses whether a correlation is significantly different from 0. Another type of test uses the degrees of freedom and might be more appropriate here. In S5 I would certainly not state that a correlation of 0.35 is 'highly correlated'.

We have rephrased "highly correlated" to "significantly correlated". Cor.test was significant at p value < 10-16 indicating an association between the measured TF mRNA expression and GSS based on the TF targets. We also confirmed the significance of the correlation, using random re-sampling (We used 10000 randomly sampling and confirmed that p<0.0001). To the best of our knowledge, the cor.test function in R, does use the degree of freedom in calculating the p-value.

Although the correlation coefficient is not high (0.4), a correlation coefficient of 0.4 to 0.5 is comparable to and within the expected range of mRNA to protein correlation studies.

There are multiple reasons why a correlation of 0.4 might be considered a high correlation for this type of data. 1) GSS is calculated based on the mean gene expression of the TF targets, essentially measuring gene activity. We would only expect a subset of genes with binding sites to be activated at any one genes 2) TFT prediction is often inaccurate. We included experimental valid and computational predicted transcription factor binding sites, many of the latter are likely to be inaccurate 3) mRNA counts are not a direct measure of TF activity. Given many unavoidable limitations in gene expression data analysis, we believe a correlation coefficient of 0.4 is biologically meaningful.

h. I would have liked to see some PCA scatter plots on the samples, maybe with colors based on the C1, C2 and C3 to understand the data better.

We now include a PCA on each individual dataset and the results in shown in figure S7, where the patients are colored according to their subtype.

3. Methods section, data simulation: I struggled to understand that section.

a. It was not clear to me how many features belong to each gene set.

b. There are few repetitions (n= 30 mentioned l5 and l9 p21, the fact that 100 sets of triplet l4 p21 and l13 p22), statement 'ith row and j'th column' is not clear given that the matrix dimensions are not given

c. I was not clear how many observations belonged to each cluster - perhaps a diagram would help to clarify this.

We regret this was not clear. We have considerably edited this section for clarity. In addition, we include Figure S1 to illustrate the data simulation.

50 features belonged to each gene set.

We are removed repetitions

There were 30 observations, 6 clusters and 5 observations per cluster.

4. Methods section, moGSA algorithm: this section would need serious rewriting as some sentences do not make sense

These sections has been edited for clarity

a. Could the method handle genes belonging to several gene sets? If not, that is a limitation to mention in the discussion

Yes a gene can belong to multiple gene-sets. This is not a limitation in moGSA.

We now state (line 107-110): “Rows of the gene-set annotation matrix contain the features and each column is an independent annotation vector for a gene-set. A feature may belong to multiple gene-sets simultaneously, that is a row sum may exceed 1.”

b. More effort must be made for the notations (value p l20 p16 should be introduced here, not l3 p17), wrong use of '-' instead of '…', e.g. in 'X_1 - X_k' should be 'X_1, …, X_k'. It would help to mention the dimensions of some matrices, such as P, Q, W_k. F_k l6 p23 is not introduced before.

We updated the notation and now state the dimension of matrices

for example "**X** is transpose so that **P** is a $n \times r$ matrix, **Q** is a $p \times r$ matrix, **Δ** is an $r \times r$ square matrix," (line 460); "where **F** has the same dimension as **P**" (line 462); "The overall gene-set space **W** ($m \times r$ matrix)" (ling 479) …

c. In GSVD, I assume the number of latent variables is the rank of X? that would be worth mentioning. Some pointers as of how many PCs to choose in the final model would be welcomed to the reader.

In SVD, the number of latent variable is the rank of X. In GSVD, the number of latent variable cannot exceed the rank of X. We now include this in the manuscript, on line 461 "the maximum number of r is the rank of **X**." In point 1d (above) we include pointers on selecting the number of PCs.

d. Step 3: is there a key reference for that step or is that a novel contribution? If so, it would need to be better justified, especially equation (11- 13)

This is a novel contribution. We have edited the description for clarity so that the calculation of overall gene set score are first introduced, and the decomposition is simply the same calculation of subsets of the original grand matrix. This description is more straightforward.

e. 'normalized gene-sets score': is calculated after the matrix reconstruction. Wouldn't it be better to include it in the W_k calculation instead?

Thanks for this suggestion. This was a helpful suggestion. We modified our algorithm to implement this.

In our previous implementation, the gene set annotation matrix is a binary matrix where rows are features and columns are gene sets. Each columns is a gene set annotation vector where 1 indicates the association between feature and gene set, and 0 otherwise.

In the updated version, we divide each gene annotation vector by its sum of the vector, so that the sum of the vector equals 1. This will lead to the normalized W_k and normalized gene sets score. Please refer to page 19, line 7.

f. 'gene influential score'. What is the motivation for calculating the standard deviation? Why not use a correlation instead? Presumably that score would be limited in a small sample size setting (e.g. the cell line data set?)

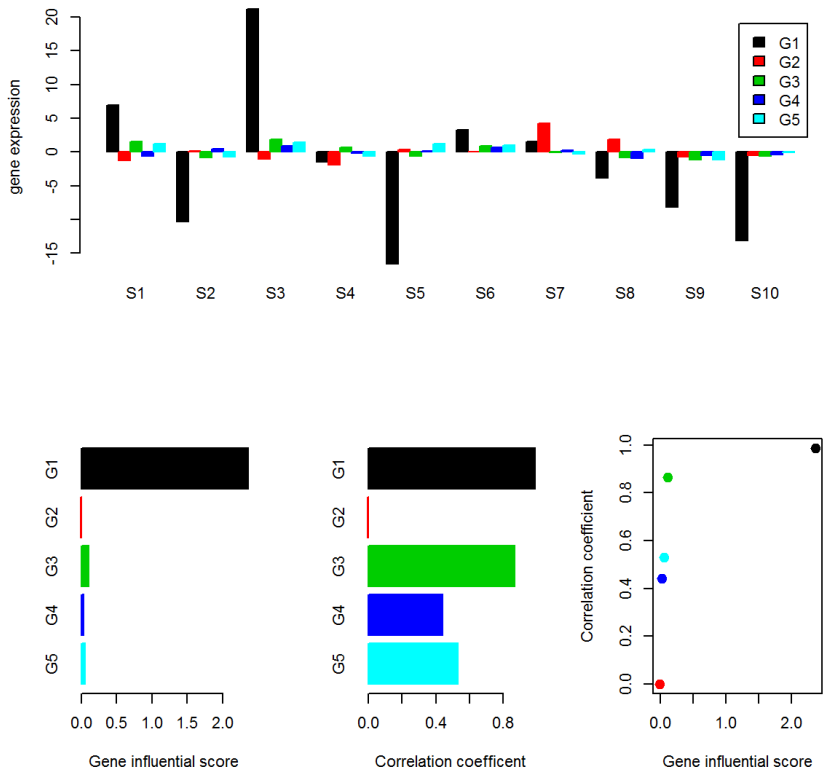
Thank you for your suggestion, we have added correlation as a GIS measurement to the mogsa Bioconductor package, and this will be available in the next release of the package.

However whilst the correlation and the leave-one-out procedure may be similar in many cases, the correlation ignores the actual range of expression. This is a problem in some cases, as illustrated below.

In the following toy example (see figure below), we examine 5 genes (G1 to G5) which belong to the same gene set in 10 observations (S1 to S10). Expression values for the 5 genes in each of the 10 samples are shown in barplot (top panel).

We can see that G1 (black) gene has the largest range and variance, whereas other genes (G2 to G5) have much smaller ranges, and the final gene set scores (GSS) are dominate by G1. The gene influential score (GIS; based on leave-one-out procedure) clearly tells that G1 has the biggest influence (bottom left panel), but the correlation coefficients suggest that G3 is almost as important as G1 (bottom middle panel). However the range of G3 is very low. In this situation, there would be a "poor" correlation between correlation base method and the leave-one-out procedure based methods (bottom right panel).

In 'omics data, we have thousands of values, that are detected in low numbers close to the instrument detection range. That is why we tend to use moderated t-statistics, rather than a classical t-tests. In future we may implement a more statistically rigorous approach, but for now, we find the current implementation of GIS useful.



g. 'statistical inferential aspect'. I did not understand what were the implications of that section, please expand as why it would be useful.

This section is renamed evaluation of significance of gene-set scores and we have modified the text such that it is clearer.

h. Sections 'Clustering latent variable(s)' and 'Prediction strength.' should probably be moved after moGSA if that is part of the moGSA framework.

We now make it clear that these are not part of the moGSA framework, however we find it intuitive to include these at this position in the manuscript as it highlights options that users may wish to consider.

Single sample GSA results is a matrix of gene-sets by observations. This is easily parsed for small datasets. But when there are many observations, clustering is informative. Numerous clustering approaches exist. In the article we employ two different approaches, but we have not sought to determine an optimal clustering approach. This is an interesting question and would be the subject of future work.

In addition, users may their own preferences and we do not want to restrict users, therefore, we did not clustering within the moGSA framework.

All code was written in R and is available.

5. Methods section, TCGA data. That section report non-specific filtering, while the abstract mention 'no requirement to pre-filter the features in a study'. Explain. There is no library size/ TMM or any specific RNA seq normalisation applied to the mRNA data?

Many integrative methods require that features from different datasets are mapped to a common identifier, and exclude features that do not map. This reduces the number of features available for study and is likely to reduce the power to detect differentially regulated gene-sets.

But this is not the case in our method. This is what we mean by 'no requirement to pre-filter". To avoid potential confusion, we removed this text in the updated version.

We did a minimal amount of filtering to exclude low quality features; those with low expression values or low variance. These are unlikely to provide useful information. This filtering was not required to run the method.

We downloaded and used pre-normalized level 3 TCGA data. RNAseq data had been normalized using RNASeqV2, which uses the MapSplice for the alignment and RSEM for the normalization and quantitation. The only processing that we performed was a median centering of columns (tumors). We now state this more clearly in the text.

6. Methods section, clustering latent variable. Explain the choice of the value 80% of patients used in resampling

We used default parameters. 80% is the default number in consensus clustering. Whilst we report the results with the default settings, we also vary this parameter and found that this clustering is robust with respect to changes in 50% to 90% of this parameter. We include an additional Figure which shows that clustering is stable with between 50% to 90% resampling (Figure S11).

In the manuscript, on line 272, we state "Stability analysis showed there was no effect when different resampling proportions (50%, 60%, 70%, 80% and 90%) were used in the inner and outer loop of consensus clustering (Figure S11)"

7. Methods section, prediction strength. Explain the choice of the value 9-nearest neighbours

We updated the text and include figure S12 and S13 to justify the choice of 9 nearest neighbors. We tested K = 1,3,5,7,9,11,13,15 and 17 and found it had minimal influence on the results.

On Line 277-279 we state: “Data were divided into training and test, and a KNN classifier was used to iteratively predict the class of each patient. Though no good choice of *K* existed (Figure S12), this had minimal influence on the final result, which clearly supported three subtypes (Figure S13).”

8. Methods section, Processing of the iPS ES 4-plex data. The total number of samples is not clear from that data set, nor the number of features left after filtering. Be consistent with the TCGA data set in terms of notations ('log10(value+1)' and '(normalized count + 1) were logarithm transformed (base 10').)

We regret our lack of clarity and have now updated the corresponding text.

Line 589: “. In this study, we used the 4-plex data, which consists of 17347 genes, 7952 proteins and 10499 sites of phosphorylation in four cell lines. “

Line 596: “ After filtering, 10,961, 5817 and 7912 features were retained in the transcriptomic, proteomic and phosphor-proteomic datasets”

Notation terms are now uniform. For example, we state “all the data were logarithm transformed (base 10).”

9. Methods section, Subtype calling…' proper references missing. This section should appear after the TCGA preprocessing section. Similarly 'Exam of somatic mutations of patients': I assume this is part of the TCGA bladder data set but it is not clear from the title of that section.

We have edited the manuscript considerably to correct for this. This section now appears after the “Preprocessing of Bladder Cancer TCGA data” section. Also, there is no “subtype calling” we include the description of published subtypes at the start of the results section so that users are aware of the background to the analysis. The references were included and updated.

10. Availability of the method. The bioconductor vignette would be to be updated - this is really important to ensure the method can be used by others.

The method is free, open and publically available in Bioconductor.

https://www.bioconductor.org/packages/release/bioc/html/mogsa.html

We include two vignettes with the package and will continue to update the package and vignettes so they can be used by others. New and modified functions (such as adding correlation measurement of gene influential score) are in the development version and it will be automatically propagated to the release version as part of the next release of Bioconductor (April 2016).

*Minor

1. 'Overview of the moGSA algorithm' Step 3 in Fig 1 is not mentioned. The paragraph should also state that all details are in the Methods section.

We now include more details of the methods in the results section. We describe the calculations performed in each step, for example, page 6, line 12 “This is done by multiplying the components for features and the gene set annotation matrix.”. line 15: “moGSA generates a gene-set score (GSS) matrix through multiplying the components for observation and components for gene-sets. As a result, a gene-set score is calculated for each”. Secondly, we explicitly say “step 1/2/3” in our description so that readers can associate the description to the figure.

#Reviewer 2:

The article tackles the issue of analyzing GO information associated with multi-omic data. The method is relevant to analyze have multiple omic data obtained on a common set of individuals and for which all variables in all datasets are described by a common set of GO terms. The approach uses MFA to represent the multi-omic dataset in a small dimensional subset and projects the GO terms as additional variables. Then, scores for the different variables and individual are derived. The method is simple, the article well written and illustrated with convincing experiments on simulated and real datasets. However, the article suffers from imprecise statements and questionable choices related to GO use. I divide my review into three parts: main comments and minor comments which is mainly a list of typos.

* Main comments

- The first main comment is related to the use of GO terms. First, GO terms are usually organized into a tree: how do you select which level of the tree you are using? does the number of GO terms used in the analysis is important? which choices do you advice so as which GO terms and how many GO terms should be included in the analysis? Can you include redundant informations like two GO terms which are children/ancestor?

Gene sets were downloaded from MSigDB. Users typically apply GSEA to all MSigDB genesets and therefore we implemented our methods so that it is consistent and comparable with the widely used methods (eg GSEA, GSVA).

The MSigDB database do filter GO Terms, and we used their filtered version of GO. Specifically they exclude broad and narrow terms and those with redundancy. According to the description of the database: "GO gene sets for very broad categories, such as Biological Process, have been omitted from MSigDB. GO gene sets with fewer than 10 genes have also been omitted. Gene sets with the same members have been resolved based on the GO tree structure: if a parent term has only one child term and their gene sets have the same members, the child gene set is omitted; if the gene sets of sibling terms have the same members, the sibling gene sets are omitted."
(http://software.broadinstitute.org/gsea/msigdb)

We expanded the description of the genesets in methods section entitled “Sources of Gene-set annotation”, so it now reads:

Gene-sets from the Molecular Signature Database MSigDB (version 4.0) [34] were used in this analysis. The following MSigDB categories were included; MSigDB C2 curated pathways, C3 motiff pathways which included the transcription factor target (TFT) target gene-set and C5 gene ontology (GO) gene-sets which included biological process (BP), cellular component (CC) and molecular function(MF) GO terms. Among GO gene-sets, there were 825, 233 and 396 genesets in the BP, CC and MF categories respectively. There were 617 TFT genesets. The pathway databases, Biocarta, KEGG and Reactome had 217, 186 and 674 gene-sets respectively. We excluded gene ontology terms that have more than 500 genes and less than 5 genes mapped to datasets For example, in the BLCA analysis, gene-sets (1,454 in total) were filtered to exclude those with less than 5 genes in a list of the concatenated features of CNV and mRNA data resulting in 1,125 retained gene-sets.

In particular, page 8, you write "The 228 GO terms are generally classified into 19 categories" and I do not understand if you are analysing the GO terms themselves or the categories in Figure 3A. If you analyse the categories, why not directly apply the method to them? Please, provide more details about all this.

Sorry for the imprecise description. Here we cluster the GO terms using gene overlap. Gene overlap was calculated using Hamming distance using gene-set annotation matrix. We calculated the pairwise distance between the columns of the gene-set annotation matrix. Therefore we only considered the genes that were present in the dataset.

Line 198-201 “There were 228 GO terms (out of 825) that had significant up or down-regulated gene-set scores (GSSs) in at least one cell line (BH corrected p value < 0.01). There was gene overlap among many GO terms and hierarchical clustering analysis (Hamming distance and complete linkage) was used to group the 288 GO terms into 21 broad categories (Table S1)”

Also, GO terms correspond to incomplete information: when a GO term is related to a gene, it is an information but when a GO term is not related to a gene, you do not know if it is an information or a missing information (this gene is related to this term but has never been referenced as such). How does your method handle this problem and how is it sensitive to missing information?

If a gene is not annotated to a gene set, then we assume there is no association. We used MSigDB for GO terms, and thus we approach is limited to the quality of that data.

If there is no association, we cannot distinguish if this is because the data is unknown or there really is no relation. Therefore, our method, similar to other gene-set analysis approaches is bias toward to the pathways or gene-sets that are well documented. This is a problem common to most all gene set methods.

This is a limitation of the gene sets data rather than moGSA itself. When a “complete” gene set database is created, then the of moGSA results would be unbiased.

- Second, I do not really understand the motivation behind gene influence score: why using cross validation and not the correlation (squared cosine) between the projection of gene j and GO term i in the projection space? Your method is computationally expensive and does not seem to be the most natural in the situation where you have all your quantity represented in the same (low-dimensional) feature space... Also, as Y is a linear combination of the observation, I think that at least you could derived a GCV criterion which would be faster?

This is a good point and was raised by both reviewers. We answer this in depth above. Please refer to the response to reviewer 1, the correlation based method would ignore the actual range of genes. This is true for the correlation in the reconstructed expression space or projection space. So we used the leave-one-out procedure.

This method is not computationally costly and takes less than a few minutes to compute the TCGA analysis on a standard laptop computer.

- Third, I have a real issue with your experiment on cancer data. You start your analysis with a clustering of the observations which is based on a PCA. Then you are performing a MFA on the same (?) dataset to analyze them together with GO terms. First, I would have liked to know which data you are precisely using to cluster your observations (the same that are further analyzed?). Then, you are performing a clustering based on the first k components: why k components? why using a PCA as preprocessing before the clustering? And most importantly, why not using the first k components of the MFA? It would be more relevant to make explicit that the groups and GO term analysis are both obtained from the same projection of the data in a small dimensional subspace. I am not sure whether it can or not biased the conclusions of your analysis but you should discuss this point.

We regret that the description of our approach lacked clarity. We have edited the manuscript considerably and we hope that the method is now easier to follow.

We did not use the PCA for clustering. We used the components (integrated data) defined by MFA for the clustering. PCA is not a preprocessing before clustering.

We regret and recognize that our previous description was misleading because we used “PC” to refer to latent variables generated by MFA. The have changed “PC” to “components” throughout the manuscript to avoid this confusion.

- I think that section moGSA step 3 must be rewritten so as to explain that the Y corresponds to the gene set projection of the space reconstructed by the first k components. As you have written it, it seems to have multiple steps although it is just

$Y = G^T X[d]$

in which X[d] is the reconstruction of X on the first d components. Equivalently, it is equal to WF^T and I do not really see the point of so many details about the different sums. I would better like the other approach: explain what is Y from a global formula and that it can be decomposed into partial scores which can be of interest in some situations. This makes more sense in light of the use of the GO terms as the projection of additional variables.

Thanks for this suggestion, we modified the method section accordingly. Please see text starting from page 20, line 7.

* Minor comments

- page 7: you write (twice) that gene-set were "selected" as differentially expressed. I think that you mean that they were "simulated". This is confusing because, before reading the Method section, I have thought that you were describing some pre-processing of a simulated dataset.

We changed the “selected” to “simulated”.

- page 8: I do not think that the term "proportionally" strictly apply to related concatenation (which is not a numeric quantity) and signal-to-noise ratio.

We changed “proportionally” to “accordingly”

- page 9: "encouraged by the results" is probably not something to write in a scientific paper...

We have removed this

- page 16: "each individual matrix contributes almost equal variance to" is not grammatically correct I think

Line 467: One dataset may contribute disproportionately to the integrated analysis in MVA approaches when all datasets have equal weight. To correct for this, MFA weights datasets by dividing each by their first eigenvalue.

- page 21, line 9: n=30 is useless (written 5 lines above)

This is removed and this section is considerably revised.

- page 21, line 13: I do not see why you are insisting on the random selection of the DE in the gene-sets because these are simulated data: you could have taken the first ones, that would have changed nothing. Maybe you wanted to point out that there was an overlap between DE in different gene-sets? Please clarify.

Randomly selected DE genes in a gene-set will results in overlapped DE genes, we clarified this point on line 453 “Random selection of DEGs means that the DEGs in different datasets may overlap.”

We did this for two reasons: (i) this setting is more realistic because in reality, the DE genes could be overlapped between datasets. (ii) in one set of simulation, we simulate up to 25 genes as differentially expressed in a gene set (there are 50 genes in total), so we cannot avoid overlapping DE genes.

- page 21, line 15: I know that statistics are probably a drug, but are you sure that you were using an "addictive model"? ;)

Thank you for your humor ;-) It made us all smile. We have changed addictive to additive

- page 22, between lines 8 and 9: the notation i \in DEG_j is confusing because DEG are related to clusters and not to observation (j). The same holds for \beta_j: you write \beta_j ~ N(\mu=0, \gamma=s) but explains that \beta_j is equal for two different j. The notations in this part must be clarified.

DE gene sets is related to clusters, that is, DE gene sets are the same among observations within the same cluster. However, DE genes (DEGs) are different among observations because they are randomly selected from these DE gene sets. To clarify this point, please see page 22, line 6 "Within each cluster, the same set of DE gene-sets were randomly selected. For a DE gene-set, a number of genes were randomly simulated as DE genes (DEG), denoted as DEG_j".

We do not use beta_j in the revised manuscript, we changed the notation to bete_l and stated that " For observations belongs to the same cluster l, the same \beta_l was applied ". (page 22, line 13.)

- The section Gene-set annotation must come before or after the data description but not in-between two data description. The same holds for Other GSA methods, Clustering latent variable, Prediction strength

We moved this section and others, and considerably revised the manuscript for clarity.

- page 24, line 6: what does "by one subtype model" mean? Do you want to refer to one clustering obtained for a given number of clusters?

This is right, we change this sentence to “In prediction strength method, all samples were assigned a “true” subtype label according to the clustering obtained from a given number of clusters” (line 626).

- page 24, line 12: "and the most subtypes": what does that mean?

We changed this sentence to “Therefore, the model with the greatest number of subtypes and prediction strength > 0.8 can be considered “optimal”. (Line 635)

moGSA: integrative single sample gene-set analysis of multiple omics data

Chen Meng[1], Bernhard Kuster[1,2], Bjoern Peters[3], Aedín C Culhane[4,5]* and Amin Moghaddas Gholami[1,6]*

[1] Chair of Proteomics and Bioanalytics, Technische Universität München, Freising, Germany

[2] Center for Integrated Protein Science Munich, Freising, Germany

[3] La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA

[4] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA.

[5] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA

[6] Current address: La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA

* Correspondence: aedin@jimmy.harvard.edu; agholami@lji.org

Abstract

Background: The increasing availability of multi-omics datasets has created an opportunity to understand how different biological pathways and molecules interact to cause disease. However, there is a lack of analysis methods that can integrate and interpret multiple experimental and molecular data types measured over the same set of samples.

Result: To address this challenge, we introduce moGSA, a multivariate single sample gene-set analysis method. It uses multivariate latent variable decomposition to discover correlated global variance structure across datasets and calculates an integrated gene set enrichment score using the most informative features in each data type. Integrating multiple diverse sources of data, reduces the impact of missing or unreliable information in any single data type, and may increase the power to discover subtle changes in gene-sets. We show that integrative analysis with moGSA outperforms existing single sample GSA methods on simulated data. We apply moGSA to two studies with real data. First we discover similarities and differences in mRNA, protein and phosphorylation profiles of induced pluripotent and embryonic stem cell lines. Secondly we report that three molecular subtypes are robustly discovered when copy number variation and mRNA profiling data of 308 bladder cancers from The Cancer Genome Atlas are integrated using moGSA. Our method provides positive or negative gene-set scores (with p-values) of each gene set in each sample. We demonstrate how to assess the influence of each data type or gene to a moGSA gene set score. With moGSA, there is no requirement to filter data to the intersect of features. All molecular features on all platforms may be included in the analysis.

Conclusion: moGSA provides a powerful yet simple tool to perform integrated simple sample gene-set analysis. Its latent variable approach is fundamentally different to existing single sample GSA approaches. It is an attractive approach for data integration and is particularly suited to integrated cluster or molecular subtype discovery. It is available in the Bioconductor R package “mogsa”.

Keywords

Gene-set analysis, Multivariate analysis, Data integration, Omics, Bladder cancer, molecular subtype stratification

Introduction

Technological innovations have enabled the acquisition of unprecedented amounts of multi-scale molecular, genotype and phenotype information. Advances in high-throughput sequencing allow quantification of global DNA variation and RNA expression in tissue or blood samples [1, 2]. Mass spectrometry (MS)-based proteomics has undergone rapid progress in recent years, and systematic MS analyses can now identify and quantify the majority of proteins expressed in a human cell line [3]. More and more studies report comprehensive molecular profiling using multiple different experimental approaches on the same set of biological samples. These data can potentially yield insights into the molecular machinery of biological systems. However, integrating, interpreting and generating biological hypothesis from such complex datasets is a considerable challenge.

Our groups and others have described multivariate analysis (MVA) approaches that uncover latent correlated structure within and between omics datasets [4-7]. MVA use extensions of principal component analysis (PCA) to project data onto a lower dimensional space so that trends or relationships between multiple datasets, observations (cases) and features (e.g. genes) can be identified. MVA methods identify global correlated patterns among observations, and therefore do not require pre-filtering of gene identifiers in each dataset to a common intersecting subset of features (genes/proteins). All features whether they have annotation or not can be included in the analysis. This is particularly important when analyzing experimental platforms that include novel genes, or use identifiers that are difficult to be map. A further attractive feature of latent variable approaches is that supplementary data

such as gene-set information (e.g. Gene Ontology annotations) can be projected onto the MVA to aid interpretation [5, 6, 8].

Gene-set analysis (GSA) is widely used in the analysis of genome scale data and is often the first step in the biological interpretation of lists of genes or proteins that are differentially expressed between phenotypically distinct groups [9]. These methods use external biological information to reduce thousands of genes or proteins into short lists of functional related gene-sets (e.g. cellular pathways, subcellular localization, transcription factors or miRNA targets), thus facilitating hypothesis generation. The simplest GSA based methods rely on over-representation analysis and only require a list of genes as input. Hypergeometric tests or Fisher's exact test are often used to identify statistically significant overlap between a shortlist of genes or proteins and a database of gene-sets [10]. Gene-set enrichment analysis (GSEA) and significance analysis of function and expression (SAFE) not only require a list of genes, but also take advantage of quantitative information in omics data [11, 12]. More recently, pathway topology approaches also consider the network structure of biological pathways in over-representation analysis [13]. However, these methods are supervised tests that require predefined groups of samples using known experimental, clinical, phenotypic or conditional data (e.g. tumor vs. normal cases).

Modern omics studies frequently explore a panel of experimental conditions or tissue samples with multiple phenotypes, for example The Cancer Genome Atlas (TCGA), ENCyclopedia of DNA Elements (ENCODE) projects [14] and other studies [15]. Such studies frequently wish to discover new molecular subtypes and thus traditional GSA methods which require known subsets have limited application in such cases. To address this issue, several unsupervised, single sample GSA (ssGSA) methods have been developed [16-19]. These methods do not require prior availability of phenotypic or clinical data. One of the most popular approaches is single-sample GSEA (ssGSEA) that ranks genes according to the empirical cumulative distribution function and calculates a single sample-wise gene-set score by comparing the

scores of genes that are inside and outside a gene-set [18]. Another related method described recently, gene-set variation analysis (GSVA), also calculates sample-wise gene set enrichment as a function of the genes that are inside and outside a gene set. GSVA uses a similar Kolmogorov-Smirnov-like rank statistic to assess the enrichment score, but genes are ranked using a kernel estimation of a cumulative density function [16]. Each of these unsupervised single-sample GSA methods are designed for the analysis of a single dataset. To the best of our knowledge no GSA method exists which integrates and calculates a single sample GSA score on multiple datasets simultaneously.

Here, we present a novel unsupervised single-sample gene-set analysis that calculates an integrated enrichment score using all of the information in multiple 'omics datasets. We call this approach multiple omics GSA (moGSA). We show that moGSA has higher sensitivity and specificity to detect gene-sets compared to single dataset GSA and demonstrate that moGSA outperforms existing unsupervised GSA methods when applied to simulated data. We apply moGSA to both small and large scale data from multiple omics studies.

Results

moGSA integrates and discovers gene-sets that are enriched in features in two or more omics data matrices obtained on the same set of observations (Figure 1). Omics studies generate multiple data matrices such as RNA sequencing counts of gene expression, measurements of proteins, metabolites, lipids, DNA copy number variations and several other biological molecules that can be mapped to gene-sets. In each, the number of features frequently exceeds the number of observations (rows and columns of the matrix, respectively). In this paper, we refer to genes or other biological molecules as features for simplicity.

Figure 1 describes the three steps of the algorithm. Input quantitative or qualitative data matrices must have matched observations but may have different and unmatched features. The number of features

may exceed the number of observations. In order to map features to gene-sets, moGSA requires an incidence matrix of gene to gene-set membership associations for each data matrix and in each "gene-set annotation matrix", a value of 1 indicates that a feature (e.g. gene) is a member of a gene-set. Rows of the gene-set annotation matrix contain the features and each column is an independent annotation vector for a gene-set. A feature may belong to multiple gene-sets simultaneously, that is a row sum may exceed 1.

In the first step, several (k) input data matrices are integrated using multiple factor analysis (MFA) [20]. MFA is a multiple table extension of principal component analysis (PCA) that is well suited to integrating multiple omics data since it reduces high dimensional omics data to a relative small number of components that capture the most prominent correlated structure among different datasets [20]. To prevent datasets with more features or different scales dominating a MFA, each dataset is weighted by dividing it by the first eigenvalue of a decomposition of each individual dataset. MFA generates matrices of latent variables (components) in observation (P) and feature (Q) space. The number of components typically equals the number of observations minus one. We retain and examine the first few components as these represent most of the variance in the data. Approaches for choosing the number of components are discussed later. In the next step (step 2) each gene-set annotation matrix ($G_{1..k}$) is projected as additional information onto the gene-set space ($Q_{1..k}$) generating a score for each gene-set in the same projected space ($W_{1..k}$). In the final step (step 3), moGSA multiplies the latent variables of the observations (P) and latent variables of gene-sets ($W_{1..k}$) to generate a matrix (Y) with a gene-set score (GSS) for each gene-set in each observation (Y).

A gene-set with a high GSS value has features that explain a large proportion of the global correlated information among data matrices. These features could be from any or all data matrices, and may be non-overlapping, for example a GSS of a gene set with features A-H, could be driven by high levels of gene expression in genes A,B,C, and increased protein levels in proteins C,D,E and amplifications in copy

number in gene H. The GSS matrix (Y) may be decomposed with respect to each dataset (X) or latent variable space (P,Q) so that the contribution of each individual dataset or component to the overall score can be evaluated (see Methods).

moGSA outperforms existing single sample GSA methods

Methods to perform integrated ssGSA on multiple 'omics datasets are not yet described. Therefore, we compared the performance of moGSA to ssGSA methods that were developed for analysis of one dataset. One-table ssGSA methods were generally optimized for analysis of gene expression data and include the widely used GSVA and ssGSEA and naïve matrix multiplication (NMM) [16, 18].

Figure 2 shows the performance of each method applied to 100 simulated datasets, each run simulated a study of 30 observations with three omics datasets that measured 1,000 features each (Figure S1; see Methods section). Each features was a member of one of the 20 gene-sets. Each gene-set had 50 genes. The observations grouped into 6 clusters and each clusters had 5 differentially expressed (DE) gene-sets when compared to the other observations. Within DE gene-sets, 5, 10 and 25 out of 50 genes were randomly simulated to be DE genes (DEG). The triplets were analyzed by moGSA directly, however matrices were concatenated for NMM, GSVA and ssGSEA as these methods can only accept one matrix as input.

We anticipated that moGSA might be especially powerful at identifying altered gene-sets in heterogeneous or noisy data. That is because moGSA, uses only the top few most informative latent variables, thus omitting the signal of many features with little variance, which are potentially noise. Therefore we explored the power of the methods to detect DE gene-sets when there was a strong or weak gene expression signal. First we simulated increasing DEG signal to noise by changing the mean gene expression of DEGs in the cluster and secondly we altered the number of DE genes in a DE gene-set (5, 10 and 25 genes). As expected, the performance of all methods was better when signal-to-noise ratio

or the number of DE genes in DE gene-sets increased (Figure 2A and 2B). moGSA consistently outperformed the other methods and the difference were even more apparent when the signal-to-noise ratio was low or when there were few DE genes (5 or 10 of 50 genes) (Figure 2B).

Next we compared the performance of each method using data with a simple or complex phenotype. In data with a simple phenotype a few components should easily capture most of the variance in the data. However in data with a complex phenotype for example a heterogeneous tumor dataset, with mixed histology, grade and response to treatment, there are many signals and many latent variables may be required to capture even half of the variance. Specificity and sensitivity of the methods detecting the DE gene-sets (measured as the area under the receiver operating characteristic curve; AUC) were evaluated. In the simulated data, observations grouped into six clusters, each with highly correlated genes and these six clusters could be captured by the first five components. Therefore we simulated data such that the first 5 components captured 50%, 30% or only 25% of the total variance (Figure 2C). Again, moGSA outperformed the other methods and was relatively robust to changes in the variance retained (Figure 2D). The performance (AUC) of all methods decreased when greater variance was retained, which can be explained by higher intra-cluster correlation that leads to a lower signal-to-noise ratio (see methods).

Given the many fundamental differences between moGSA and the other ssGSA methods, we repeated the simulations adjusting for technical aspects of the moGSA approach that might give it an "unfair edge", but these did little to improve the performance of the others methods. Since, GSVA and ssGSEA were designed for analysis of single datasets, we compared the performance of GSVA and ssGSEA on a single datasets of the triplet compared to the concatenated triplet. Concatenating multiple data matrices neither improved nor decreased the performance compared to analysis of single datasets, most likely because the signal-to-noise ratio increased accordingly with concatenation (Figure S2). In addition, since MFA weights input matrices by their first singular value before moGSA, we examined the

effect of data set weighting on the other methods, but found moGSA still outperformed ssGSEA and GSVA when data matrices of the triplet were weighted before concatenation (Figure S3).

Application of moGSA to stem cell mRNA and proteomics data

We applied moGSA to study a dataset consisting of mRNA, protein and phospho-protein profiling of four cell lines – two embryonic stem cell lines (ESC; H1 and H9), one induced pluripotent cell line (iPSC; DF19.7) and a fibroblast cell line (newborn foreskin fibroblast; NFF). Induced pluripotent stem cells (iPSC) are adult cells that have been reprogrammed to be more like embryonic stem cells (ESC) and have great potential in the field of regenerative medicine. These cells express ESC markers and can differentiate into different cell types [21]. Induced pluripotent cells are often derived from NFF cells. The data was downloaded from [21].

After filtering low abundant features, there were 10,961; 5,817; and 7,912 unique mRNAs, proteins and phosphorylation sites features respectively (see Methods). Principal component analysis (PCA) of each individual dataset is shown in Figure S4. The strongest signal (first PCs) in all three datasets was the difference between NFF cells and the stem cell lines, and this difference was particularly apparent in the proteomics datasets. The second and third components represented subtle differences between iPSC and ESC lines, thus we retained the top 3 components when we applied MFA to transform all of the data onto the same space and scale. The three datasets contributed similar variance in the integrated analysis, as indicated by weighting of each dataset in MFA. The first eigenvalues (square of singular values) of each PCA were 0.24, 0.26 and 0.26 for the transcriptome, proteome and PhosoProteome dataset respectively. MFA recapitulated the PCA of the individual datasets. Most of the variance was captured in that the first component and it discriminated between NFF and other cell lines. The variance of the molecular differences between the ESC cells (captured on the second component) was greater than the difference between ESC and iPSC cell lines (component 3) (Figure S5).

moGSA was used to annotate the features with gene ontology (GO) biological processes. There were 228 GO terms (out of 825) that had significant up or down-regulated gene-set scores (GSSs) in at least one cell line (BH corrected p value < 0.01). There was gene overlap among many GO terms and hierarchical clustering analysis (Hamming distance and complete linkage) was used to group the 288 GO terms into 21 broad categories (Table S1). Gene-set scores of representative GO terms from each category are shown in Figure 3A. Biological processes associated with more differentiated cell types were associated with the NFF cells and included up-regulation of vesicle-mediated transport, immune related responses and cell adhesion. In contrast cell proliferation GO terms such DNA replication, and cell cycle processes had significantly higher GGS in the highly proliferative stem cell lines. These results confirm previous findings [21].

In integrative analysis of multiple omics data, it is important to evaluate the relative contribution (either concordant or discrepant) of each dataset to the overall GSS. Data-wise decomposition of the GSSs (see Methods) are shown in Figure 3B. The three data sets have concordant contributions to most of the GO terms, including vesicle mediate transport, cell matrix adhesion, cell cycle processes in NFF line; chromosome organization and biogenesis in H9 and NFF cell lines.

However, in other GO classes, we also observed differences in the contribution of mRNA, proteins and phosphor-protein data to the GSS. Chromosome organization and biogenesis had significant positive GSS in the stem cells and significant negative GSS in the NFF cells, and was driven by differences in the phosphorylation data. Another case where the mRNA and protein data were incongruent was the GO class “glycoprotein metabolic process”. It had GSS scores of 9.7 ($p<0.001$), -8.6 ($p<0.01$), -5.3 ($p<0.01$) and 0 ($p>0.05$) in NFF, iPSC, H9 and H1 cells respectively. Up-regulation in NFF mainly reflects up-regulation on the protein level. However, down-regulation in iPSC DF19.7 cells is due to low expression of related mRNAs. The GO term wound healing has previously been shown to be differentially upregulated in fibroblast NFF cells compared to ESC [21]. Consistently, we also found wound healing was

upregulated in NFF compared to ESC; the GSS for wound healing were 14.2 ($p<0.01$), -5.4 ($p<0.01$), -5.2 ($p<0.01$) and -3.6 ($p<0.001$) for NFF, iPSC, H9 and H1 cells respectively (Table S1). Down-regulation of wound healing in H9 cell line was dominated by mRNA data, and the two proteomics datasets contributed little to the negative GSS. In contrast to previous studies [21], we did not observe significant differences in wound healing between iPSC and ESC. This difference could be because moGSA is more sensitive (than single data GSA) in detecting gene-sets that have subtle but consistent changes in multiple datasets. More importantly, the contribution of individual gene-set could be evaluated by the decomposition of GSS with respect to datasets

Application of moGSA to TCGA Bladder cancer data analysis

Since moGSA performs unsupervised integrative single sample GSA, it is particularly useful approach for cluster discovery in multi 'omics data. Therefore we applied moGSA to extract an integrative subtype model of BLCA from copy number variation (CNV) and mRNA data of 308 muscle invasive urothelial bladder cancer (BLCA) patients (obtained as part of the TCGA project).

BLCA is a molecularly heterogeneous cancer with between 2 and 5 molecular subtypes (reviewed by [38]). Briefly, Sjödahl et al. first defined five major subtypes termed urobasal A (UroA), UroB, genomically unstable (GU), squamous cell carcinoma-like (SCCL) and 'infiltrated' [22]. The TCGA study defined four expression clusters (I–IV) [23] . The two subtype model consists of basal-like and luminal subtypes [24] which was extended by Choi et al. who defined a 'p53-like' luminal subtype apart from basal-like and luminal subtypes [25].

Data were downloaded from the TCGA website and after filtering out features with low variance (see Methods), CNV and RNA-seq mRNA expression data contained 12,447 and 14,710 genes respectively, in which 7,644 genes were common to both datasets (Figure S4). Filtering of features is not required by moGSA but we filter low quality features as they are unlikely to contribute to the analysis. PCA of each

individual dataset is shown in Figure S7. From scree plots of the first 10 eigenvalues, an elbow in each plot appears between 4-6 components suggesting this number of components are needed to capture most of the variance (Figure S7), which we anticipated given the known molecular heterogeneity in these data. The first eigenvalue (square of singular value) of the PCA of BLCA mRNA and CNV data are `0.0004` and `0.0003` respectively. We applied a preliminary MFA on the data and Figure 4A shows the eigenvalues of the resulting components. The top five components captured a quarter of the total variance and were not dominated by either CNV or mRNA (CNV 50.6%, mRNA 49.4%). Also, these five components were not correlated with batches (TCGA batch ID), plates, shipping date or tissue source sites.

In a typical analysis, we use a scree plot to select the number of components. The scree plot indicated that five components should capture sufficient variance for input to moGSA. We confirmed that this was the optimal number of components as input to moGSA, in the following experiment. We performed moGSA on the BLCA mRNA gene expression and CNV data (n=308) with a range of components ranged from 1 to 12. For each gene-set in the GSS matrix, gene-sets were ranked by the number of tumors in which they were significantly regulated (either positive or negative GSS, $p<0.05$), such that gene-sets that were significant in most tumors had highest rank. The distribution of the number of tumors in which gene-sets were significant at $p<0.05$, $p<0.01$, and $p<0.001$ is shown Figure S7. No gene-set was significant in all 308 tumors and most gene-sets were insignificant in all tumors (Figure S7). For $p<0.05$, we examined the 10, 20, 40, 100, 200, 500 and 1000 highest ranked gene-sets and examined the stability of gene-set ranking when additional components were included (Figure S8). Increasing the number of components (from 1 to 5) increased the stability of gene set lists, however there was little additional gain after five components (Figure S9). Among the top 100 ranked gene-sets, few new gene-sets were identified after five components (Figure 4B).

Therefore we used moGSA to perform single sample GSA analysis with 1,125 gene-sets on an MFA of the mRNA and CNV BLCA data in which five components were retained. The number of significant gene-sets per patient ($p<0.05$) ranged from 183 to 595 and these contained both gene-sets with positive and negative GSS. To identify the number of BLCA molecular subtypes, we performed consensus clustering on the five components, which resulted in a three-subtype model (Figure 4B and Figure S10-13). We performed several experiments, to confirm that three subtypes was optimal particularly since between 2 and 5 subtypes have been previously reported in BLCA [23]. Whilst consensus clustering analysis indicated high confidence in either two or three subtypes (Figure S10B-D), silhouette analysis (Figure S10E) suggested three subtypes. Stability analysis showed there was no effect when different resampling proportions (50%, 60%, 70%, 80% and 90%) were used in the inner and outer loop of consensus clustering (Figure S11). A recent report highlighted limitations in consensus clustering [26], and therefore in parallel, we also used the "prediction strength" algorithm, to discover the number of stable subtypes that can be predicted from the data [27] (see Methods). Data were divided into training and test, and a KNN classifier was used to iteratively predict the class of each patient. Though no good choice of *K* existed (Figure S12), this had minimal influence on the final result, which clearly supported three subtypes (Figure S13). Therefore using two independent approaches, we determined that the data (5 components of the integrated analysis) supported three BLCA molecular subtypes.

The three BLCA subtypes identified in our integrative analysis overlapped with the BLCA subtypes identified in previous studies (Table S2, Figure S14). Our integrative BLCA subtypes consisted of two larger subtypes C1, C2 containing 148 and 103 patients respectively, and a smaller group C3 with 57 patients. The smaller subtype, C3, was the most robust (Figure S10E, S11). The integrative subtype C1 harbored a high number of patients in the type III and IV of the TCGA subtypes, the infiltrated and SCCL subtypes of the Sjödahl study [22] and the basal-like subtype identified by Damrauer (BH corrected p-value < 0.05, Table S2) [24]. Subtypes C2 and C3 were more similar to the Damrauer luminal subtype.

But, the C3 subtype contained more low grade tumors and showed a strong overlap with the UroA subtype of the Sjödahl study and type I of the TCGA subtype model. Subtype C2 tumors overlapped with the genomically unstable subtype defined by Sjödahl (Table S2). Accordingly, we observed higher mutation rate in the C2 patients (Figure S15). In single sample gene-set analysis with moGSA, C1 patients had more significant GSS ($p<0.05$) than C2 or C3 (Figure S16).

To further characterize BLCA, we focused on gene-sets that were differentially regulated in most patients. There were 73 gene-sets that were significantly regulated (positive or negative GSS, p value<0.05) in 200 or more of the 308 patients (Table S3 and Figure S17). Alternatively a lower cutoff would include more gene-sets that are regulated fewer tumors, fewer gene-set could be selected using a lower p-value ($p<0.01$, 0.001) or a supervised analysis could be used to select GSS that most discriminate groups of tumors. Cluster analysis of the GSS matrix (73 selected gene-sets x 308 tumors) revealed 3 clusters of gene-sets. A large cluster of 51 gene-sets had positive GSS scores in C1 but negative scores in C2 or C3. Two smaller clusters of gene-sets of 16 and 6 gene-sets had positive GSS scores in C2 and C3 respectively (Figure S17).

The large C1 gene-sets cluster was dominated by 31 gene-sets with terms associated with “immune response” which had significant strongly positive GSS in the C1 basal-like/SCC-like BLCA subtypes. Associations between immune regulation and the basal-like cluster have been previously reported [22]. The remaining 20 gene-sets in the C1 cluster of gene-sets included terms associated with “extracellular”, function, cell morphogenesis, migration and muscle cell development, “apoptosis” (2 gene- sets), and “G protein coupled receptor” (6 gene-sets) (Figure S17, S18) and EMT related gene sets (Figure S19), which recent reports that the Basal-like subtype tend to have more muscle-invasive and metastatic disease at presentation [22]. The remaining gene-sets could broadly be defined by biological processes of “cell cycle” (9 gene-sets) and “DNA repair and chromosome related” (7 gene-sets) which had high GSS in C2 (and some C1) and “mitochondrion” (4 gene-sets) in C3. A heatmap of the GSSs of representative gene-

set of each category is shown in Figure 4C and S17. We found that most of these gene-sets have been associated with subtype of bladder cancer. Increased cell-cycle and DNA repair GSS were associated with the "genomically unstable" luminal C2 cluster [28] (Figure S14, S16). The mitochondrial component has been described in bladder cancer and other cancers previously [28, 29], our study particularly associated this function with C3 low-grade papillary-like subtype in BLCA. However other gene-sets may be associated with C3 that were excluded when GSS were filtered to those that were broadly significant in 200 or more patients.

The GSSs clearly distinguished the three BLCA molecular subtypes. The most significant gene-sets, "immune response" and "immune system process" have significant positive or negative GSS in 270 and 265 of 308 patients respectively (Table S3). The median GSS for the gene-set "immune system process" was 0.82, -0.75, -0.61 in C1, C2 and C3 respectively (Figure S17, S18) indicating that immune related processes have high gene expression or CNV in the C1 subtype and much lower in C2 and C3. Next, we determined the importance of individual genes in each gene-set by calculating a gene influential score (GIS) using a leave-one-out procedure (see methods). The maximum GIS value for a gene in a gene-set is 1, which indicates that gene contributes a high proportion of variance to the overall variance of the GSSs. A GIS close to 1 often suggests a high correlation between the gene expression value and GSS. Gene influential score of the gene-set immune system process in BLCA suggested that the top ranked genes included *ITGB2*, *SPI1*, *DOCK2*, *LILRB2* and *LAT2*. Other highly ranked genes included drug target genes such as *CD4*, *IL6*, the interferon induced proteins *IFITM2* and *IFITM3* and the G protein coupled receptors *GPR183* and *CMKLR1* (Table S4). Top positive influencers in "regulation of apoptosis" were also related to the immune response, such as *STK17A*, *ANXA5* and *BCL2A1*, *STAT1*, Serpin B, *TGFB* and *ANXA1* (Table S4). Moreover, several epithelial to mesenchymal transition (EMT) related gene-sets, such as "collagen" (including *COL6A3*, *COL1A1*, *COL5A1* and *COL3A1*), "extracellular matrix proteins" (e.g. glycoproteins *SRGN* and *FBN1*) and mesenchymal gene-sets were elevated in C1 (Figure S19; Table S4).

The C3 subtype tumors had higher GSSs in mitochondrial related gene-set and lower expression of genes related to cell cycle process and DNA replication. GIS analysis suggested that two families of genes, NADH dehydrogenases (NDUFs) and mitochondrial ribosomal proteins (*ABCC1*/*MRP*) influenced the mitochondrial proteins (Table S4).

To identify transcription factors (TF) that may regulate gene expression in the three tumor subtypes, we used transcriptional factor target (TFT) gene-sets to annotate the tumors. Similar to the selection of GO terms, we focused on TFT gene-sets with more than 200 significant GSSs across 308 patients (Table S2). The GSSs of the E2F family target gene-set were significantly different in most of the tumors and are particularly low for the C3 tumors. The rest of the four identified TFs were highly elevated in the C1 subtype. Among them, we identified an *MADS* (*MCM1*, Agamous, Deficiens, and *SRF*) box superfamily member, *SRF* and several TFs associated with transactivation of cytokine and chemokine genes, including *NFkB1*, *ETS1* and *IRF1* (Figure 4D). The genes exhibiting the largest GIS in the *IRF1* and *NFkB1* target gene-sets include *ACTN1*, *CXorf21*, *ICAM1*, *MSN*, *TNFSF13B*, *IL12RB1* and *CDK6* (Table S5). Further, we examined the correlations between GSSs and the mRNA expression. All five TFs showed that the TF mRNA and GSSs are significantly correlated (Figure 4E, Figure S20). The boxplot of GSS with respect to subtypes in Figure 4C and D are shown in Figure S18,S21.

In order to identify the contribution of each dataset, we decomposed the GSSs with respect to the datasets or components. Figure 5A shows the means of data-wise decomposed GSSs in each subtype for "cell cycle process", where we found that mRNA expression strongly influenced the GSS, particularly the low GSS of the C3 subtype patients. The gene influential score (GIS) analysis supports this finding as the top 30 most influential genes are all based on mRNA expression (Figure 5B), including *RACGAP1*, *DLGAP5*, *FBXO5*, *AURKA*, *KERA* (*CNA2*) and *CDKN3* (Figure 5C). By contrast, both CNV and mRNA data influenced the gene-set "G protein coupled receptor activity" (Figure 5D) and the GIS analysis shows that the most influential genes include those from both mRNA and CNV data (Figure 5E). However, the CNV and mRNA

expression patterns in the C3 subtype shows a clear difference for this gene-set (Figure 5F). Top gene influencers of "G protein couple receptor activity" included CNV of *GRM6*, *NMUR2*, *PDGFRB* and adrenergic receptors, the gene expression of *ADGRL4* (*ELTD1*), *CMKLR1* and *PDGFRB* (Figure 5F). In addition, the data-wise decomposition of GSS identified several GSSs that were only contributed by the mRNA data, including the immune system process, DNA replication and mitochondrion gene-set (Figure S21).

Discussion

In this paper, we introduced a new multivariate single sample gene-set analysis approach, moGSA that enables discovery of biological pathways with correlated profiles across multiple complex datasets. moGSA uses multivariate latent variable analysis to explore correlated global variance structure across datasets and then extracts the set of gene-sets or pathways with highest variance and most strongly associated with this correlated structure across observations. By combining multiple data types, we can compensate for missing or unreliable information in any single data type so we may find gene-sets that cannot be detected by single omics data analysis alone [4].

moGSA uses the maximum variance of the concordant structure across of datasets to calculate the gene-set scores for each observation. This is fundamentally different from other gene-set enrichment analysis methods which use a 'within observation summarization' such as the mean or median of gene expression of genes in a gene-set. It has several characteristics that make it attractive for data integration. First moGSA uses MFA, a multi-table extension of PCA to reduce the complexity of the original data by transforming high dimensional data to a small number of components (latent variables). The components with highest eigenvalues (largest variance) capture the most prominent structure among the different datasets. Excluding components with low variance may strengthen the signal-to-noise ratio of data, as it reduces low variant, noise or artifact variance [30, 31]. In moGSA, the entire set

of features from each platform is decomposed onto a lower dimension space. The linear combination of feature loadings is used in the calculation of the gene-set scores. Features that contribute low variance contribute little to the score and thus the dimension reduction within moGSA comes with an intrinsic filtering of noise. The advantages of intrinsic variance filtering of features can be clearly seen when we applied moGSA to simulated data. moGSA outperformed ssGSA approaches including ssGSEA and GSVA which do not include a noise-filtering component. Second, data integration of features is achieved at the gene-sets level rather than scoring individual features. This greatly facilitates the biological interpretation among multiple integrated datasets. There is no requirement to pre-filter features in a study or map features from different datasets to a set of common genes. Therefore, moGSA can be used to compare technological platforms that have different or missing features.

There is great potential for applying multi-table unsupervised GSA approaches for discovery of new subtypes and pathways in integrated data analysis of complex diseases such as cancer. In this study, we applied moGSA in combination with clustering analysis. Dimension reduction approaches such as moGSA and MFA are well suited to cluster discovery data because these approaches consider the global variance in the data and as such are complementary to hierarchical or k-means clustering approaches which focus on the pair-wise distance between observations [31-33].

The number of components is an important input parameter to consider when applying moGSA to gene-set analysis or cluster discovery. Similar to PCA, the optimal number of MFA components may be assessed by examining the variance associated with the each component. The first component will capture most variance and the variance associated with subsequent component decreases monotonically. Scree plots (Figure 2C, 4A) may be used to visualize if there is an elbow point in the eigenvalues, allowing one to select the components before the elbow point. Alternatively one may select the number of components that capture a certain proportion of variance (50%, 70%, etc). In addition, one may include components that are of biological interest. For example, in the iPS ES example,

there is a clear biological meaning in the third component (ES vs iPS cell line). In analysis of the BLCA data, we examined a range of components (1-12), and show that there is little gain of information once a minimum number of components with high variance are included (Figure 4B). In addition, the variance of retained components should not be dominated by one or a few datasets. To facilitate biological interpretation of components, the GSS could be decomposed with regard to components. In the BLCA example, the second and forth component are largely contributed by CNV, whereas mRNA is more important in defining the third and fifth components. Including five components ensured that both datasets contributed relatively similar variance to the global variance.

An issue might arise with latent variables analysis if components with the large variance capture information unrelated to biological variance [30], such as technical artifacts or batch effects. In practice this is rare in MFA, because it focuses on components that capture global correlation among all datasets. Often batch effects are specific to a platform and thus a component that captures information that is entirely uncorrelated to the global structure will be omitted from the set of highly variant integrated components. However it is still wise to perform careful batch effect control, especially in the large scale omics studies. A more detailed description of batch effect detection is described in [34].

Another consideration when applying moGSA, is that it is most efficient in detecting gene-sets that have broad correlation patterns among data types. It may fail to discover gene-sets with few genes, particularly if they had low variances on the selected components.

Methods

moGSA algorithm

Input data and gene-set annotation matrix

The inputs to moGSA are pairs of multiple matrices ($\mathbf{X}_K$, $\mathbf{G}_K$). $\mathbf{X}_K$ is a set of matrices, denoted $\mathbf{X}_1$, ..., $\mathbf{X}_k$, … $\mathbf{X}_K$, where K is the total number of quantitative matrices. Matrix $\mathbf{X}_k$ is a $p_k \times n$ matrix of quantitative omic data, which contains p_k rows of features (e.g. genes) measured over the same n observations. Each of the matrices $\mathbf{X}_1$, …, $\mathbf{X}_K$ has a corresponding gene-set annotation matrix, $\mathbf{G}_1$, ..., $\mathbf{G}_k$, …, $\mathbf{G}_K$. The gene-set annotation matrix $\mathbf{G}_k$ is a $p_k \times m$ binary incidence matrix of gene to gene-set membership associations, where m is the number of gene-sets. The element $g_{k[i,j]}$ in $\mathbf{G_k}$ has the value 1 if the ith feature is a member of the gene-set j and 0 otherwise. $\mathbf{G}_k$ is constructed using predefined gene-set information such as the Gene Ontology [35, 36] GeneSigDb [37] or MSigDB [38]

moGSA step 1 multivariate integration

The first step of the moGSA involves data integration with a multiple table multivariate analysis method. In this study, we use MFA because of its simplicity and computational efficiency. MFA can be viewed as a generalization of principal component analysis (PCA) for a multi-table problem [20]. We briefly describe MFA using the nomenclature of Abdi et al. 2013 [20].

When integrating multiple data matrices, one must decide if all datasets should have equal weight, or if some data are “more important”, for example those with higher quality, fewer features, higher variance, etc. Simple tensor decomposition approaches, or PCA on a concatenated matrix, give every dataset equal weight and results are often dominated by the matrix (or matrices) with the large variance or most features. To correct for this, MFA weights datasets by dividing each by their first eigenvalue. The weight of each matrix is expressed as

$$\alpha_k = \frac{1}{\lambda_{k,1}^2} \quad (1)$$

Where $\lambda_{k,1}^2$ is the first singular value of data matrix $\mathbf{X}_k$. For convenience, the weights of matrices are stored in a diagonal matrix **A**, whose diagonal elements are

$$\mathrm{diag}\{\mathbf{A}\} = [\alpha_1 \mathbf{1}_1^{\mathrm{T}}, \cdots, \alpha_k \mathbf{1}_k^{\mathrm{T}}, \cdots, \alpha_K \mathbf{1}_K^{\mathrm{T}}] \quad (2)$$

The transpose of a matrix is denoted by superscript $^{\mathrm{T}}$. $\mathbf{1}_k^{\mathrm{T}}$ is a vector of 1 in the length of p_k. As a result, **A** is a $p \times p$ diagonal matrix, the diagonal elements of **A** representing the weight of features in $\mathbf{X}_1$, ..., $\mathbf{X}_k$. Similarly, the weight of each observation is an $n \times n$ diagonal matrix, **M**. In the present study, we use $m_{ii}=1/n$, namely, all observations are equally weighted.

We then transpose and concatenate all $\mathbf{X}_k$ to a complete pxn matrix ($p = \sum_k p_k$):

$$\mathbf{X} = [\mathbf{X}_1^{\mathrm{T}} | \ldots | \mathbf{X}_k^{\mathrm{T}} | \ldots | \mathbf{X}_K^{\mathrm{T}}]^{\mathrm{T}} \quad (3)$$

After deriving the matrix weights, observation weights and the concatenated matrix, MFA is reduced to an analysis of the triplet (**X**, **A**, **M**). The solution of the problem is given by generalized singular value decomposition (GSVD):

$$\mathbf{X}^{\mathrm{T}} = \mathbf{P}\Delta\mathbf{Q}^{\mathrm{T}} \text{ with the constraint that } \mathbf{P}^{\mathrm{T}}\mathbf{M}\mathbf{P} = \mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} = \mathbf{I} \quad (4)$$

X is transpose so that **P** is a $n \times r$ matrix, **Q** is a $p \times r$ matrix, **Δ** is an $r \times r$ square matrix, the maximum number of r is the rank of **X**. The components of MFA, **F**, are given by

$$\mathbf{F} = \mathbf{P}\Delta \quad (5)$$

where **F** has the same dimension as **P**. In the PCA framework, the matrix **P** contains the PCs or latent variables. We also call it *sample space* in this paper. The column vectors in **P** may be plotted on a two dimensional space to visualize the contribution of each observation to the variance captured by each PC. The matrix **Q** is the loading matrix or *gene space*. Because **X** is a concatenation of multiple matrices, the gene space matrices $\mathbf{Q}_1$, ..., $\mathbf{Q}_k$ may also be concatenated or partitioned in the same manner, namely,

$$\mathbf{Q} = [\mathbf{Q}_1^{\mathrm{T}} \mid \cdots \mid \mathbf{Q}_k^{\mathrm{T}} \mid \cdots \mid \mathbf{Q}_K^{\mathbf{T}}]^{\mathbf{T}} \quad (6)$$

moGSA step 2 project gene-set annotation matrix as supplementary data

Different gene-sets have different candidate genes, therefore, in order to facilitate the comparison of gene-set score across gene-sets, we normalized the gene-set annotation matrix so that the sum of each column in **G** equals 1, that is,

$$\hat{g}_{[i,j]} = \frac{\hat{g}_{[i,j]}}{\sum_i \hat{g}_{[i,j]}} \quad (7)$$

where $\hat{g}_{[i,j]}$ is the elements on the *i*th row and *j*th column in the normalized gene-set annotation matrix $\hat{\mathbf{G}}$. The gene-set score calculated using un-normalized gene-set annotation matrix for gene-sets in Figure 3 and 4 are shown in Figure S22 and S23.

Next, we project the annotation matrix as supplementary data [35] to generate the *gene-set space* matrix $\mathbf{W}_k$ (*m*×*r*), which is calculated as a product of the normalized gene annotation matrix and loading matrix.

$$\mathbf{W} = \hat{\mathbf{G}}^{\mathrm{T}}\mathbf{Q} \text{ where } \hat{\mathbf{G}} = [\hat{\mathbf{G}}_1^{\mathrm{T}} \mid \cdots \mid \hat{\mathbf{G}}_k^{\mathrm{T}} \mid \cdots \mid \hat{\mathbf{G}}_K^{\mathrm{T}}]^{\mathrm{T}} \quad (8)$$

$\hat{\mathbf{G}}$ is the grand annotation matrix with dimension *p*×*m*. The overall gene-set space **W** (*m*×*r* matrix) could also be expressed as the sum of individual $\hat{G}_k$ and $\mathbf{Q}_k$, that is,

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{W}_k \text{ where } \mathbf{W}_k = \hat{\mathbf{G}}_k^{\mathrm{T}}\mathbf{Q}_k \quad (9)$$

moGSA step 3 reconstruction of gene-set-observation matrix

The main output of MOGSA is a *gene-set score (GSS)* matrix, denoted by **Y**, whose rows are *m* gene-sets and columns are *n* observations. It is calculated as

$$\mathbf{Y} = \hat{\mathbf{G}}^{\mathrm{T}}\mathbf{Q}^{[R]}\boldsymbol{\Delta}^{[R]}\mathbf{P}^{[R]^{\mathrm{T}}} = \mathbf{W}^{[R]}\mathbf{F}^{[R]^{\mathrm{T}}} = \hat{\mathbf{G}}^{\mathrm{T}}\mathbf{X}^{[R]} \quad (10)$$

where $\mathbf{Q}^{[R]}$ and $\mathbf{P}^{[R]}$ are the gene space and observation space within top *R* components. $\boldsymbol{\Delta}^{[R]}$ is the diagonal matrix containing top *R* singular values. As a result, $\mathbf{X}^{[R]}$ is the reconstruction of **X** using top *R* components. In practice, it is interesting to evaluate the contribution of a dataset or a component to the overall gene-set score. Therefore, we decompose gene-set scores with respect to data sets and components. The GSS matrix for dataset $\mathbf{X}_k$ and component *r* is calculated as

$$\mathbf{Y}_k^r = \mathbf{W}_k^r \mathbf{F}_k^{r^{\mathrm{T}}} \quad (11)$$

we use superscript *r* to indicate the *r*th component and the subscript *k* to indicate the *k*th matrix ($\mathbf{X}_k$). Similarly, $\mathbf{W}_k^r$ denotes the *r*th dimension of gene-set space of matrix $\mathbf{X}_k$, $\mathbf{F}_k^r$ is the *r*th component of the sample space. The outer product of the two vectors results in a GSS matrix for a specific components and dataset. Consequently, the overall gene-set score for component r (i.e. component-wise decomposed gene-set scores) is the sum of the gene-set score matrix of the components across all datasets, that is,

$$\mathbf{Y}^r = \sum_k \mathbf{Y}_k^r = \sum_{k=1}^{K} \mathbf{W}_k^r \mathbf{F}_k^{r^{\mathrm{T}}} \quad (12)$$

Similarly, the overall gene-set score matrix by a single dataset (i.e. data-wise decomposed gene-set scores) is the sum of the matrices by all the components retained.

$$\mathbf{Y}_k = \sum_t \mathbf{Y}_k^r = \sum_{r=1}^{R} \mathbf{W}_k^r \mathbf{F}_k^{r\mathrm{T}} \quad (13)$$

Therefore, the contribution of an individual dataset and/or component may be calculated. Finally, the complete gene-set score matrix is given by

$$\mathbf{Y} = \sum_t \mathbf{Y}^r = \sum_k \mathbf{Y}_k = \sum_{k=1}^{K} \sum_{r=1}^{R} \mathbf{W}_k^r \mathbf{F}_k^{r\mathrm{T}} \quad (14)$$

which is the sum of all contributions by individual components and dataset. In practice, only the components with greatest variances (highest eigenvalues) should be retained in the analysis. If all components are retained, the result would be similar or exactly the same as naïve matrix multiplication (NMM; see later).

Evaluation of the significance of gene-set scores (calculating p-values)

The expression (7) and (10) say that, for each observation, a gene-set score could be viewed as the mean of gene expression (in the reconstructed expression values $\mathbf{X}^{[R]}$) of genes in a particular gene-set.

If the candidate genes in a gene-set are randomly drawn from all features in $\mathbf{X}^{[R]}$ (null hypothesis), the distribution of the means of selected genes is given by central limited theorem (CLT),

$$\bar{x} \sim N(\mu, \sigma_{\bar{x}}) \text{ with } \sigma_{\bar{x}} = c\frac{\sigma}{\sqrt{h}} \quad (15)$$

Where μ is the mean of a column (observation) in $\mathbf{X}^{[R]}$, $\sigma_{\bar{x}}$ is the sampling standard deviation of means, σ is the standard deviation of the column in $\mathbf{X}^{[R]}$, h is the number of candidate genes mapped to $\mathbf{X}$ in a gene-set and $c = \sqrt{(p-h)/(p-h)}$ is the finite population correction factor (p is the number of features in $\mathbf{X}$). It is used since each gene was only selected once in one gene-set.

Gene influential score

Gene-sets are composed of genes, and therefore we calculate the contribution of each feature to the GSS, as it is interesting from a biological point of view to identify "driver" genes in a gene-set. In moGSA, feature contribution, denoted by gene influential score (GIS), is calculated via a leave-one-out procedure. The GSS of gene-set *i*, $\mathbf{Y}_{[i]}$, for all the observations are

$$\mathbf{Y}_{[i]} = \hat{\mathbf{G}}_{[i]}{}^{\mathrm{T}}\mathbf{X}^{[R]} \tag{16}$$

where $\hat{\mathbf{G}}_{[i]}$ is the gene-set annotation vector for gene-set *i*. Correspondingly, the gene-set score for *i*th gene-set excluding gene *g* is

$$\mathbf{Y}_{[i]}^{-g} = \hat{\mathbf{G}}_{[i]}^{-g\,\mathrm{T}}\mathbf{X}^{[R]} \tag{17}$$

Where $\hat{\mathbf{G}}_{[i]}^{-g}$ is the gene-set annotation vector for gene-set *i* but without gene *g*. The influence of the gene *g* is measured by

$$E_{[i]}^{g} = -\log_2 \frac{sd(\mathbf{Y}_{[i]}^{-g})}{sd(\mathbf{Y}_{[i]})} \tag{18}$$

where $sd(\cdot)$ stands for the function of calculating standard deviation. For convenience, the feature influential score then is rescaled, such that the gene with maximum influence always equals 1. Therefore, a positive $E_{[i]}^{g}$ suggests that gene *g* tends to have a positive correlation with gene-set score of gene-set *I*, whereas a gene with a negative value tends to have a negative correlation.

Data simulation

We simulated 100 multiple 'omics data projects. Each simulated dataset was a triplet (*K*=3) containing three data matrices (Figure S1), each matrix had the dimension 1000×30, representing 30 matched observations (*n*=30) and 1,000 features (p_k=1,000). Each of dataset of features had an annotation matrix,

which assigned each feature to one of 20 non-overlapping "gene-sets". The binary annotation matrix had dimensions 1,000 features × 20 gene-sets. Each gene-set contained 50 genes.

The 30 observations were defined by 6 equal sized clusters with 5 samples per cluster.

In each observation, 5 out of 20 gene-sets were simulated as differentially expressed (DE). With same a cluster, the same set of DE gene-sets were randomly selected as we assume that differentially expressed (DE) gene-sets define the difference between clusters and observations. For a DE gene-set, a number of genes were randomly simulated as DE genes (DEG), denoted as DEG_j. Random selection of DEGs means that the DEGs in different datasets may overlap. In different simulations (Figure 2) we varied the number of DEGs per gene-set (eg 5, 10 and 25 out of 50) or mean signal:noise.

We used the following linear additive model adapted from [16], the expression or abundance of gene on *i*th row and *j*th column is simulated as

$$y_{ij} = \alpha_i + \beta_l + \gamma_{ij} + \varepsilon_{ij} \quad (19)$$

where with *i = 1, …, n* is gene specific effect. $\beta_l \sim \mathrm{N}(\mu = 0, \sigma = s)$ is the cluster effect. For observations belongs to the same cluster *l*, the same β_l was applied. The cluster effect factor (categorical variable) is introduced following the hypothesis that observations from the same clusters are driven by some common pathways or “gene-sets” and ensures that observations from the same cluster have a higher within than between cluster correlation. The six correlated clusters in the simulated data are captured by first five components. We adjust the variance of each cluster, so that different variance would be captured by the top five components. The cluster effect $\beta_l \sim \mathrm{N}(\mu = 0, \sigma = s)$ is sampled from a distribution with a mean of 0 and standard deviation *s.* The standard deviation (*s*) adjusts the correlation between observations in the same cluster, and thus each cluster can have different variance. In this study, we set s = 0.3, 0.5 and 1.0, which lead to 25%, 30% and 50% of total variance are captured by the

top 5 components. $\varepsilon \sim \mathrm{N}(\mu = 0, \sigma = 1)$ is the noise factor. γ_{ij} is a factor, if a gene is differentially expressed (DE):

$$\gamma_{ij} \begin{cases} \sim \mathrm{N}(\mu = m, \sigma = 1) & if \quad i \in DEG_j \\ = 0 & otherwise \end{cases} \tag{20}$$

Apart from the retained variance, two other parameters are tuned in the simulation study. First is the number of DEGs in a DE gene-set (5, 10 and 25 out of 50 DEGs). The second parameter is different signal-to-noise ratio, which is tuned through modifying *m* in (20). The candidate *m* are 0.3, 0.5 and 0.8 standing for low, medium and high signal-to-noise ratio. In total, 100 projects of triplet datasets were generated. The three matrix triplets were analyzed by moGSA. NMM, GSVA and ssGSEA, only accept one matrix as input; therefore the three simulated matrices in one triplet set were concatenated. The performance was assessed by the area under the ROC curve (AUC).

Data

Downloading and Processing of Bladder Cancer TCGA data

Normalized mRNA gene expression, copy number variation (CNV), microRNA (miRNA) expression data and clinical information of BLCA were downloaded from TCGA (Date: 26/09/2014) using TCGA assembler [39]. The processed mRNA gene expression had been obtained on the Illumina HiSeq platform and the MapSplice and RSEM algorithm had been used for the short read alignment and quantification (Referred as RNASeqV2 in TCGA) [40, 41]. The gene level CNV was estimated by the mean of copy number of genomic region of a gene (retrieved by TCGA assembler directly). Patients that were present in both gene expression and the CNV data were included in the analysis (n=308).

Before applying moGSA, minimal non-specific filtering of low variance genes was performed on both datasets. RNA sequencing data (normalized count + 1) were logarithm transformed (base 10). Genes were filtered to retain those with a total row sum greater than 300 and median absolute deviation (MAD)

greater than 0.1, which retained 14,692 unique genes (out of 20,531 genes). Then, RNA-seq gene expression data were median centered. For the CNV data, genes with standard deviation greater than the median were retained.

Genome instability in TCGA BLCA tumors

GISTIC2.0 [42] data for copy number gains/deletion in 24,776 unique genes were downloaded from TCGA firehouse (http://gdac.broadinstitute.org/; download date 2015-03-09). The GISTIC encodes homozygous deletion, heterozygous deletion, low-level gain and high-level amplification as -2, -1, 1 and 2 respectively. The four types of events were counted for each of the patients. The total number of events were calculated by sum all four types of events

Downloading and Processing of the iPS ES 4-plex data

The transcriptomic (RNA-sequencing), proteomic and phoshphoproteomics data were downloaded from Stem Cell-Omic Repository (Table S1, S2 and S5 from http://scor.chem.wisc.edu/data.php) [21]. In this study, we used the 4-plex data, which consists of 17347 genes, 7952 proteins and 10499 sites of phosphorylation in four cell lines. For the transcriptomics data, the expression levels of genes were represented by RPKM values. Three replicates were available and we used the mean RPKM value of the three replicates. Genes with duplicated symbols and low expression (summed RPKM < 12) were removed. The iTRAQ quantification of protein and phosphorylation sites were performed by TagQuant [43], as describe in [21]. The protein and sites of phosphorylation with low intensity (summed intensity <20) were removed. In the proteomics data, proteins that are not mapped to an official symbol were removed. Finally, all the data were logarithm transformed (base 10). After filtering, 10,961, 5817 and 7912 features were retained in the transcriptomic, proteomic and phosphor-proteomic datasets. A few missing values still present and replaced with zero. The enrichment analysis was done on the gene symbol levels, the specific phosphorylation sites were not considered.

Sources of Gene-set annotation

Gene-sets from the Molecular Signature Database MSigDB (version 4.0) [38] were used in this analysis. The following MSigDB categories were included; MSigDB C2 curated pathways, C3 motiff pathways which included the transcription factor target (TFT) target gene-set and C5 gene ontology (GO) gene-sets which included biological process (BP), cellular component (CC) and molecular function (MF) GO terms. Among GO gene-sets, there were 825, 233 and 396 gene-sets in the BP, CC and MF categories respectively. There were 617 TFT gene-sets. The pathway databases, Biocarta, KEGG and Reactome had 217, 186 and 674 gene-sets respectively. We excluded gene ontology terms that have more than 500 genes and less than 5 genes mapped to datasets. For example, in the BLCA analysis, gene-sets (1,454 in total) were filtered to exclude those with less than 5 genes in a list of the concatenated features of CNV and mRNA data resulting in 1,125 retained gene-sets.

Other GSA methods (including NMM)

Single gene-set method, including GSVA and ssGSEA methods were implemented using the R/Bioconductor package GSVA [16]. Default settings were used for these methods. Naïve gene-set score Y_{naive} was calculated through matrix multiplication (NMM).

$$\mathbf{Y}_{naive} = \hat{\mathbf{G}}^{\mathrm{T}}\mathbf{X} \tag{21}$$

Therefore, the result of NMM is exactly the same as moGSA if all of the axes are retained.

Clustering latent variable

Consensus clustering was used [44, 45] to cluster the top five latent variables with Pearson correlation distance and Ward linkage for the inner loop clustering. Eighty percent of patients were used in the re-sampling step of clustering. In addition, different percentage of patients in the resampling was evaluated. The results suggested the subtype model is robust with regard to different percentage of

samples resampling (Figure S20). Average agglomeration clustering was used in the final linkage (linkage for consensus matrix) [44].

Prediction strength to determine the optimal number of subtypes

We used the "prediction strength" algorithm to assess the number of subtypes that can be predicted from the data [22]. In prediction strength method, all samples were assigned a "true" subtype label according to the clustering obtained from a given number of clusters. Then, the patients were then divided into "training" and "testing" sets. KNN classifier was used to classify the patients in testing set. Cross-validation suggested that there is no obvious good choice of *K* (Figure S21), but the number of *K* does not have a big influence on the result (figure S22). We finally selected to use 9 nearest neighbors (the middle of evaluated numbers). For each test, the agreement in assignment between predicted and true labels were computed. The prediction strength is defined by the lowest proportion among all the subtypes. It indicates the similarity between the true and predicted labels and ranges from 0 to 1, where a value > 0.8 suggests a robust subtype classification [22]. Therefore, the model with the greatest number of subtypes and prediction strength > 0.8 can be considered "optimal". In this study, we performed 100 random separations of training and testing sets and the prediction strength of each randomization was calculated.

List of Abbreviations

ANOVA – analysis of variance

AUC – area under the ROC curve

BLCA – bladder cancer

BP – biological process

CC – cellular component

CCA – canonical correlation analysis

CIA – co-inertia analysis

CLT – central limited theorem

DE – differentially expressed

DEGS – differentially expressed gene-set

EMT – Epithelial to mesenchymal transition

GIS – gene influential score

GO – gene ontology

GS – gene-set

GSA – gene-set analysis

GSEA – gene-set enrichment analysis

GSS – gene-set score

MAD - median absolute deviation

MCIA – multiple co-inertia analysis

MF – molecular function

MFA – multiple factorial analysis

MVA – multivariate analysis

NMM – naïve matrix multiplication

PCA – principal component analysis

ROC - Receiver operating characteristic

SVD – singular value decomposition

TCGA – the cancer genome atlas

TF – transcriptional factor

TFT – transcriptional factor target

Competing interests

The authors declare no conflict of interest

Authors' contribution

AC conceived the study with CM and AMG. AC, CM and AMG developed the concept and experimental design and wrote the manuscript. CM wrote the R code and conducted the experiments. AMG and AC supervised the project. BK and BP had intellectual contribution to both the experimental design and drafting the manuscript.

Description of additional data files

SupplementaryFigures.pdf – 23 supplementary figures

Table_S1.xlsx - Table S1 - the gene-set score (GSS) matrix of Gene ontology (GO) for iPS ES 4-plex data.

Table_S2.xlsx - Table S2: The Chi square test of association between integrative subtypes and previously published subtypes.

Table_S3.xlsx - Table S3 - the gene-set score (GSS) matrix of Gene ontology (GO) and transcriptional factor target (TFT) gene-set with more than 200 significant GSSs for BLCA data.

Table_S4.xlsx - Table S4 - the gene influential score (GIS) for selected gene-sets (from Gene Ontology). The document contains GIS analysis for 9 gene-sets.

Table_S5.xlsx - Table S5 - the gene influential score (GIS) for selected transcriptional factor gene-sets. The document contains GIS analysis for 2 gene-sets.

Acknowledgements

We wish to thank Prof. Joaquim Bellmunt for the insightful discussions about bladder cancer molecular subtypes and treatment. We also thank Dr. Hannes Hanne and Dominic Helm for reading the manuscript and giving the valuable suggestions. Funding for this work was provided by DFCI BCB Research Scientist Developmental Funds, National Cancer Institute at the National Institutes of Health [grant numbers 2P50 CA101942-11, 1U19 AI111224-01, 1U19 AI109755-01] and Department of Defense BCRP [award number W81XWH-15-1-0013]. Views and opinions of, and endorsements by the author(s) do not reflect those of the US Army or the Department of Defense

Reference

1. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. *Nature reviews Genetics* 2011, **12**(2):87-98.
2. Metzker ML: **Sequencing technologies - the next generation**. *Nature reviews Genetics* 2010, **11**(1):31-46.
3. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al*: **Mass-spectrometry-based draft of the human proteome**. *Nature* 2014, **509**(7502):582-587.
4. Meng C, Kuster B, Culhane AC, Gholami AM: **A multivariate approach to the integration of multi-omics datasets**. *BMC bioinformatics* 2014, **15**:162.
5. de Tayrac M, Le S, Aubry M, Mosser J, Husson F: **Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach**. *BMC genomics* 2009, **10**:32.
6. Fagan A, Culhane AC, Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007, **7**(13):2162-2171.
7. Le Cao KA, Martin PG, Robert-Granie C, Besse P: **Sparse canonical methods for biological data integration: application to a cross-platform study**. *BMC bioinformatics* 2009, **10**:34.
8. Busold CH, Winter S, Hauser N, Bauer A, Dippon J, Hoheisel JD, Fellenberg K: **Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data**. *Bioinformatics* 2005, **21**(10):2424-2429.
9. Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS computational biology* 2012, **8**(2):e1002375.
10. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic acids research* 2009, **37**(1):1-13.
11. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach**. *Bioinformatics* 2005, **21**(9):1943-1949.
12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.
13. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R: **A novel signaling pathway impact analysis**. *Bioinformatics* 2009, **25**(1):75-82.
14. Consortium EP: **The ENCODE (ENCyclopedia Of DNA Elements) Project**. *Science* 2004, **306**(5696):636-640.
15. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegner J: **Data integration in the era of omics: current and future challenges**. *BMC systems biology* 2014, **8 Suppl 2**:I1.
16. Hanzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data**. *BMC bioinformatics* 2013, **14**:7.
17. Tomfohr J, Lu J, Kepler TB: **Pathway level analysis of gene expression using singular value decomposition**. *BMC bioinformatics* 2005, **6**:225.
18. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C *et al*: **Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1**. *Nature* 2009, **462**(7269):108-112.
19. Lee E, Chuang HY, Kim JW, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification**. *PLoS computational biology* 2008, **4**(11):e1000217.

20. Abdi H, Williams LJ, Valentin D: **Multiple factor analysis: principal component analysis for multitable and multiblock data sets**. *Wiley Interdisciplinary Reviews: Computational Statistics* 2013, **5**(2):31.
21. Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, Bailey DJ, Swaney DL, Tervo MA, Bolin JM, Ruotti V *et al*: **Proteomic and phosphoproteomic comparison of human ES and iPS cells**. *Nature methods* 2011, **8**(10):821-827.
22. Sjodahl G, Lauss M, Lovgren K, Chebil G, Gudjonsson S, Veerla S, Patschan O, Aine M, Ferno M, Ringner M *et al*: **A molecular taxonomy for urothelial carcinoma**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2012, **18**(12):3377-3386.
23. Knowles MA, Hurst CD: **Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity**. *Nature reviews Cancer* 2015, **15**(1):25-41.
24. Damrauer JS, Hoadley KA, Chism DD, Fan C, Tiganelli CJ, Wobker SE, Yeh JJ, Milowsky MI, Iyer G, Parker JS *et al*: **Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(8):3110-3115.
25. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, Roth B, Cheng T, Tran M, Lee IL *et al*: **Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy**. *Cancer cell* 2014, **25**(2):152-165.
26. Senbabaoglu Y, Michailidis G, Li JZ: **Critical limitations of consensus clustering in class discovery**. *Sci Rep* 2014, **4**:6207.
27. Tibshirani R, Walther G: **Cluster Validation by Prediction Strength**. *Journal of Computational and Graphical Statistics* 2005, **14**(3):18.
28. Lindgren D, Frigyesi A, Gudjonsson S, Sjodahl G, Hallden C, Chebil G, Veerla S, Ryden T, Mansson W, Liedberg F *et al*: **Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome**. *Cancer research* 2010, **70**(9):3463-3472.
29. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Perez C, Lopez-Bigas N, Kamoun A, Neuzillet Y, Gestraud P *et al*: **Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes**. *Cell reports* 2014, **9**(4):1235-1245.
30. Chang W-C: **On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions**. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1983, **32**(3):9.
31. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(18):10101-10106.
32. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P: **'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns**. *Genome biology* 2000, **1**(2):RESEARCH0003.
33. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV: **Fundamental patterns underlying gene expression profiles: simplicity from complexity**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(15):8409-8414.
34. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data**. *Nature reviews Genetics* 2010, **11**(10):733-739.
35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

36. Gene Ontology C: **Gene Ontology Consortium: going forward**. *Nucleic acids research* 2015, **43**(Database issue):D1049-1056.
37. Culhane AC, Schroder MS, Sultana R, Picard SC, Martinelli EN, Kelly C, Haibe-Kains B, Kapushesky M, St Pierre AA, Flahive W *et al*: **GeneSigDB: a manually curated database and resource for analysis of gene expression signatures**. *Nucleic acids research* 2012, **40**(Database issue):D1060-1066.
38. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.
39. Zhu Y, Qiu P, Ji Y: **TCGA-assembler: open-source software for retrieving and processing TCGA data**. *Nature methods* 2014, **11**(6):599-600.
40. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**(4):493-500.
41. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM *et al*: **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery**. *Nucleic acids research* 2010, **38**(18):e178.
42. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers**. *Genome biology* 2011, **12**(4):R41.
43. Wenger CD, Phanstiel DH, Lee MV, Bailey DJ, Coon JJ: **COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA**. *Proteomics* 2011, **11**(6):1064-1074.
44. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data**. *Machine Learning* 2003, **52**(1-2):28.
45. Wilkerson MD, Hayes DN: **ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking**. *Bioinformatics* 2010, **26**(12):1572-1573.

Figure legends

Figure 1 - Schematic view of the moGSA algorithm. The algorithm requires pairs of matrices as input; multiple omics data matrices and corresponding gene-set (GS) annotation matrices. In step 1, the multiple matrices are analyzed with a multivariate analysis (MVA) method resulting in an observation space and gene space. Next, the gene-set annotation matrices are projected on the same space, and the resulting matrix contains the gene-set space. The last step is to reconstruct gene-set-observation through multiplying the observation and gene-set spaces.

Figure 2 – Comparison of moGSA with NMM, GSVA and ssGSEA. The performance of methods was accessed by their ability to identify differentially expressed gene-sets over 100 simulations in every condition (as indicated by the area under the ROC curve; AUC). (A) Comparison of GSA methods using data with different signal-to-noise ratios. (B) Comparison of data with different number of differentially expressed (DE) genes in each of the DE gene-set. From left to right, 5, 10 and 25 of total 50 genes are differentially expressed in each of the three simulated data matrices if a gene-set is defined as DE gene-sets. (C) Scree plots show representative eigenvalues in each of the conditions in (D). (D) AUCs with different proportion of variance are capture by top 5 components. From left to right, 25%, 30% and 50% of total variance are captured. The darker bars represent the top 5 components.

Figure 3 – integrative gene-set analysis of iPS ES 4-plex data. (A) A heatmap shows the gene-set score (GSS) for significantly regulated gene-sets in the cell lines, the white colored blocks/cells indicates the change of gene-sets are non-significant. (B) Data-wise decomposition of the GSS for some of the gene-sets. The contribution of each of the data is represent by a bar. The Y-axis is the data-wise decomposed gene-set score.

Figure 4 – Data integration with moGSA and integrative subtype defined by latent variables. (A) Bar plot showing the eigenvalues of components defined by MFA. The top 5 components were selected in the

analysis. (B) Effect of including additional component (1-12) on the identification of new genesets among the top 100 genesets (C) Prediction strength was used to evaluate the robustness of classification into two to eight subtypes. The boxplot shows the prediction strength of 100 randomizations. Two and Three are relative robust subtype models (prediction strength > 0.8). (D) Gene ontology (GO) and transcriptional target (TFT) gene-sets annotation of tumors. Heatmap showing the GSSs for selected gene-sets. The gene-sets "immune-related, apoptosis, G protein receptor, collagen, extracellular region and cell migration" are strong in the C1 (basal-like) subtype, whereas the mitochondrial related gene-sets are over represented in the C3 (luminal A-like) subtype of tumors. (E) The most significant transcriptional factor (TF) target gene-sets. The gene-set scores suggest that 4 out of the 5 TFs are hyperactive in the C1 subtype, except E2F family is active in the C2 subtype of cancer. The white spaces in (A) and (B) denote non-significant GSSs. (F) The scatter plots display the correlation between gene-set scores and the mRNA level of selected TFs. The expression of selected TFs is significantly correlated with their gene-set scores (also see Figure S16).

Figure 5 – CNV and mRNA data contribute unequally to defining subtype and gene-set scores. (A) Data-wise decomposition of gene-set scores for "cell cycle process". The bar plot shows the normalized mean of data-wise decomposed GSSs in each subtype (the black vertical line on the bars show the 95% confidence interval of the mean). (B) The bar plot shows the gene influential scores (GISs) of genes in the "cell cycle process" gene-sets. The expression of the top 30 most influential genes in the gene-set are shown in (C). (D-F) Same as (A-C) for "G protein couple receptor activity". Gene names in (F) with asterisks indicate genes from CNV data.

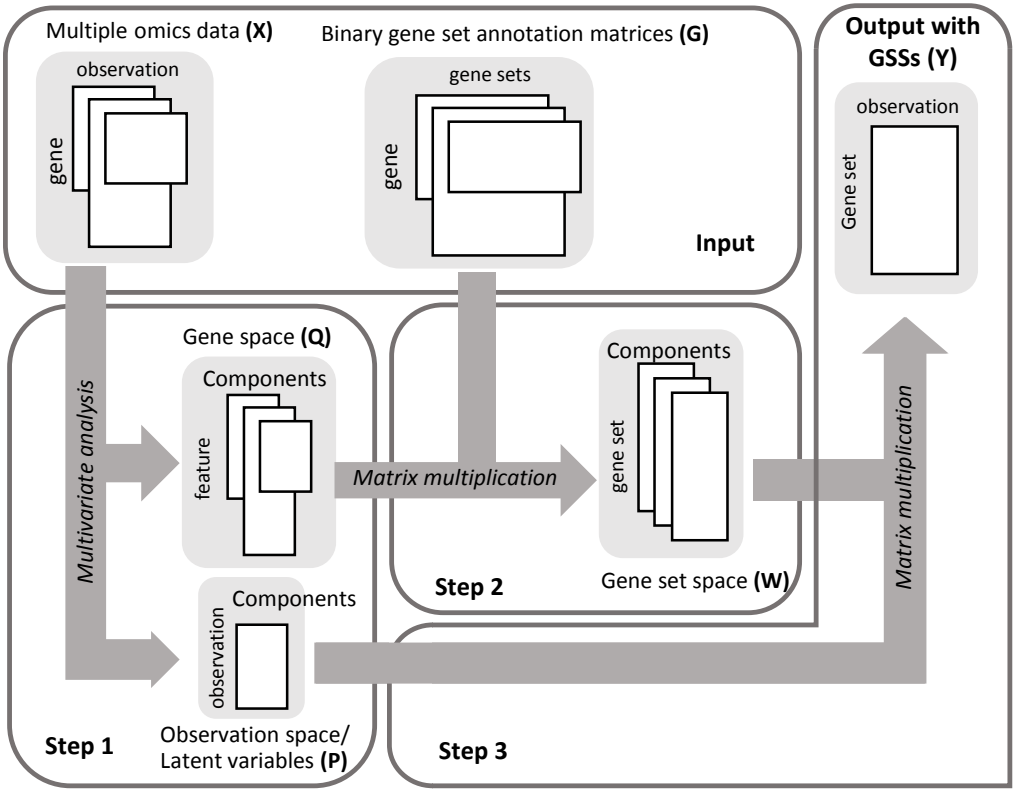


Figure 1

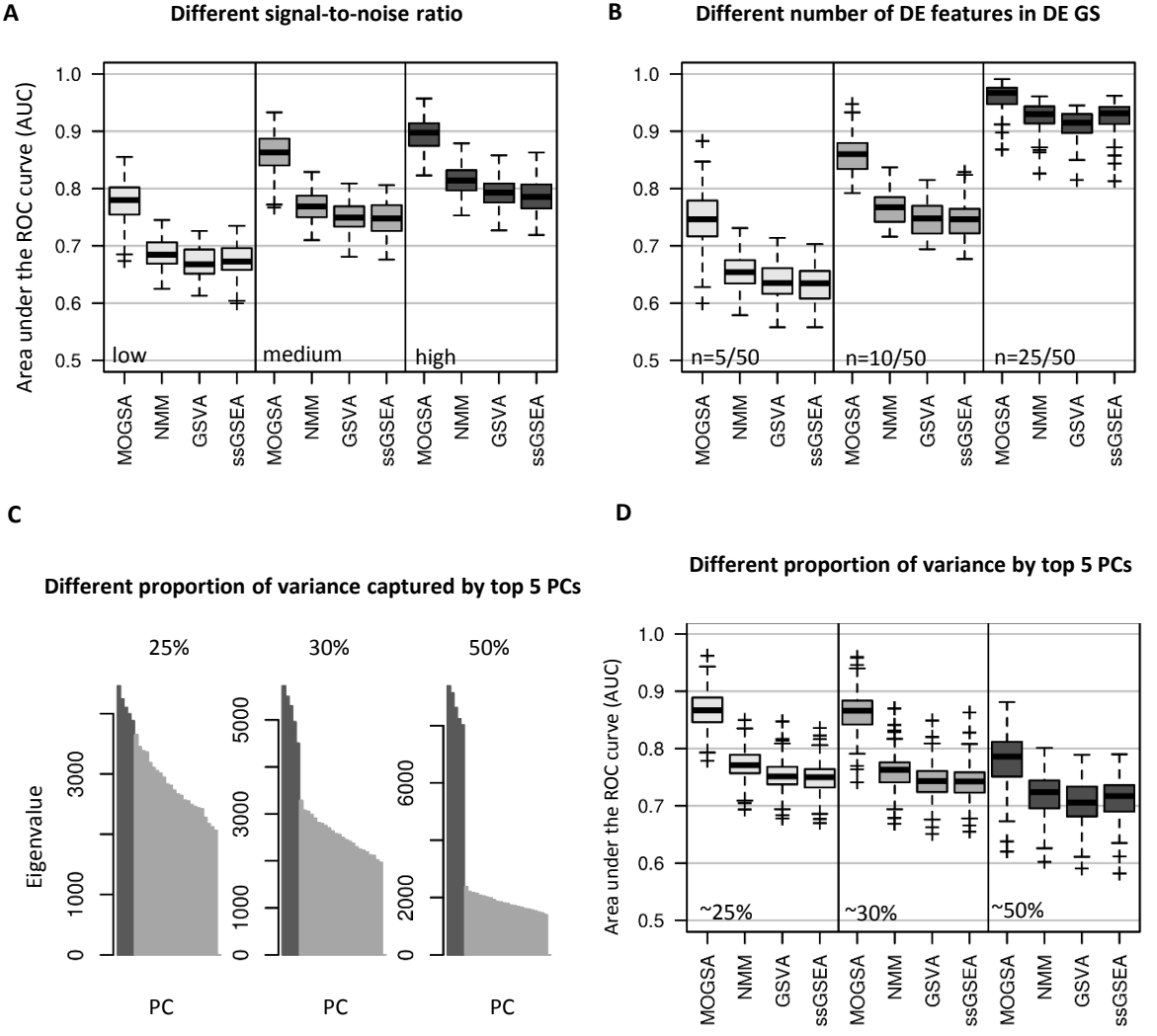

A
Different signal-to-noise ratio
Area under the ROC curve (AUC)
low
medium
high
MOGSA
NMM
GSVA
ssGSEA
B
Different number of DE features in DE GS
n=5/50
n=10/50
n=25/50
C
Different proportion of variance captured by top 5 PCs
25%
30%
50%
Eigenvalue
PC
D
Different proportion of variance by top 5 PCs
~25%
~30%
~50%

Figure 2

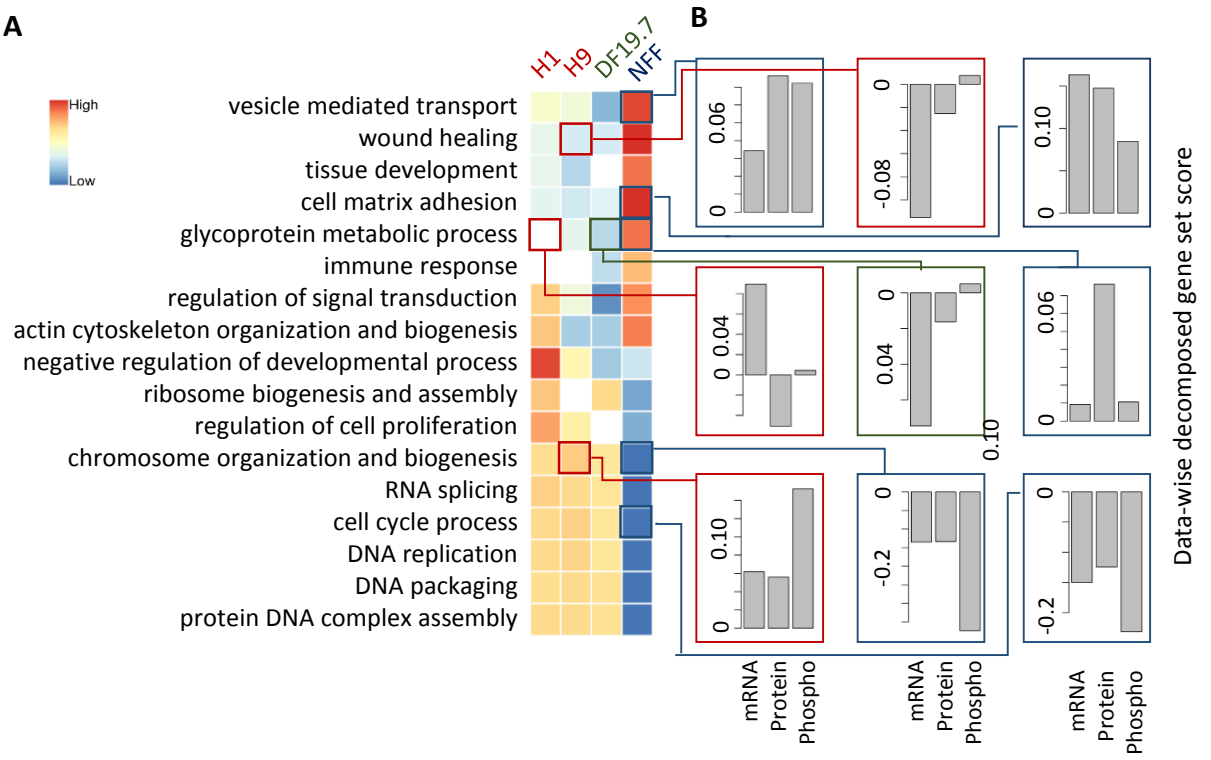

A
B
High
Low
H1
H9
DF19.7
NFF
vesicle mediated transport
wound healing
tissue development
cell matrix adhesion
glycoprotein metabolic process
immune response
regulation of signal transduction
actin cytoskeleton organization and biogenesis
negative regulation of developmental process
ribosome biogenesis and assembly
regulation of cell proliferation
chromosome organization and biogenesis
RNA splicing
cell cycle process
DNA replication
DNA packaging
protein DNA complex assembly
mRNA
Protein
Phospho
Data-wise decomposed gene set score

Figure 3

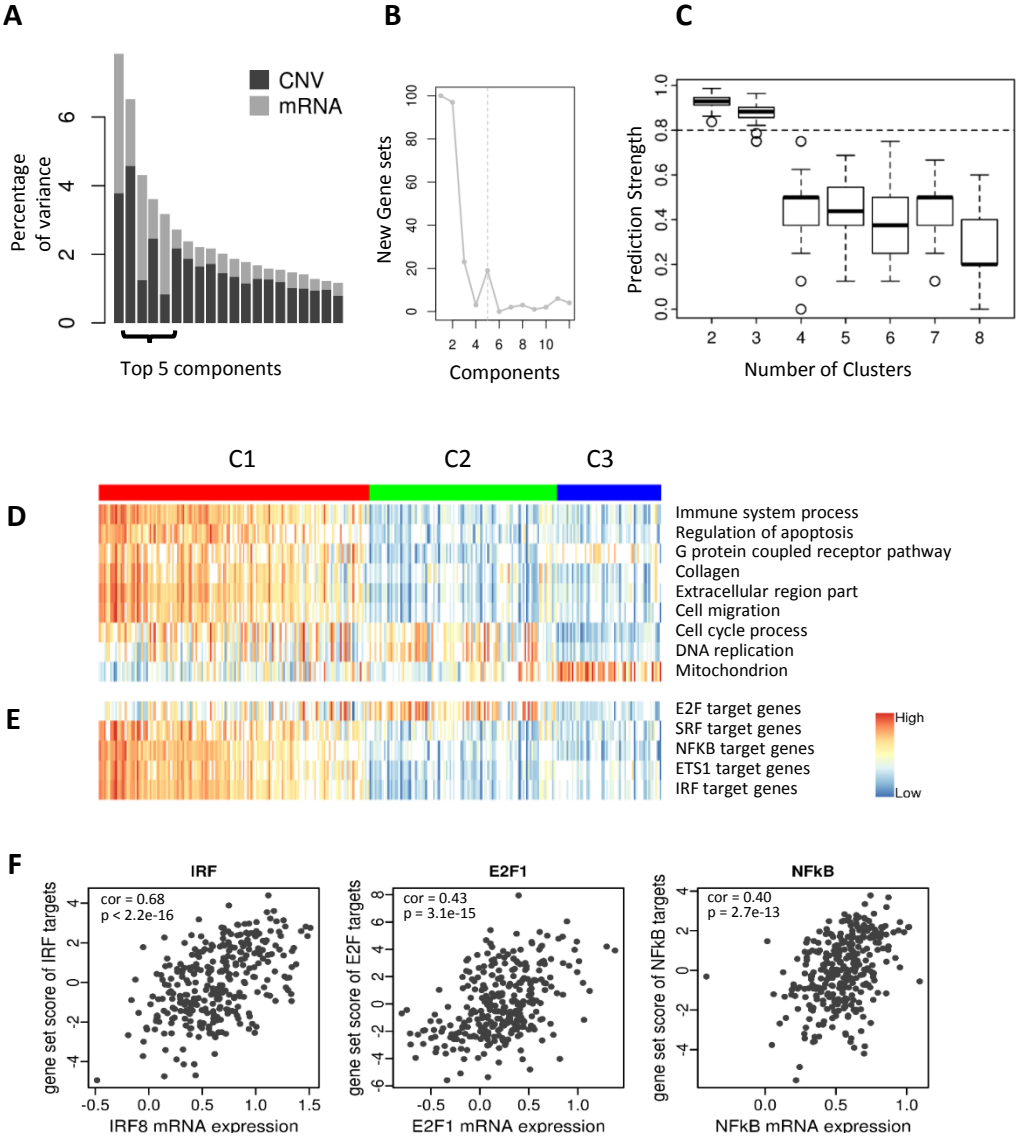

A
CNV
mRNA
Percentage of variance
Top 5 components
B
New Gene sets
Components
C
Prediction Strength
Number of Clusters
C1
C2
C3
D
Immune system process
Regulation of apoptosis
G protein coupled receptor pathway
Collagen
Extracellular region part
Cell migration
Cell cycle process
DNA replication
Mitochondrion
E
E2F target genes
SRF target genes
NFKB target genes
ETS1 target genes
IRF target genes
High
Low
F
IRF
cor = 0.68
p < 2.2e-16
gene set score of IRF targets
IRF8 mRNA expression
E2F1
cor = 0.43
p = 3.1e-15
gene set score of E2F targets
E2F1 mRNA expression
NFkB
cor = 0.40
p = 2.7e-13
gene set score of NFkB targets
NFkB mRNA expression

Figure 4

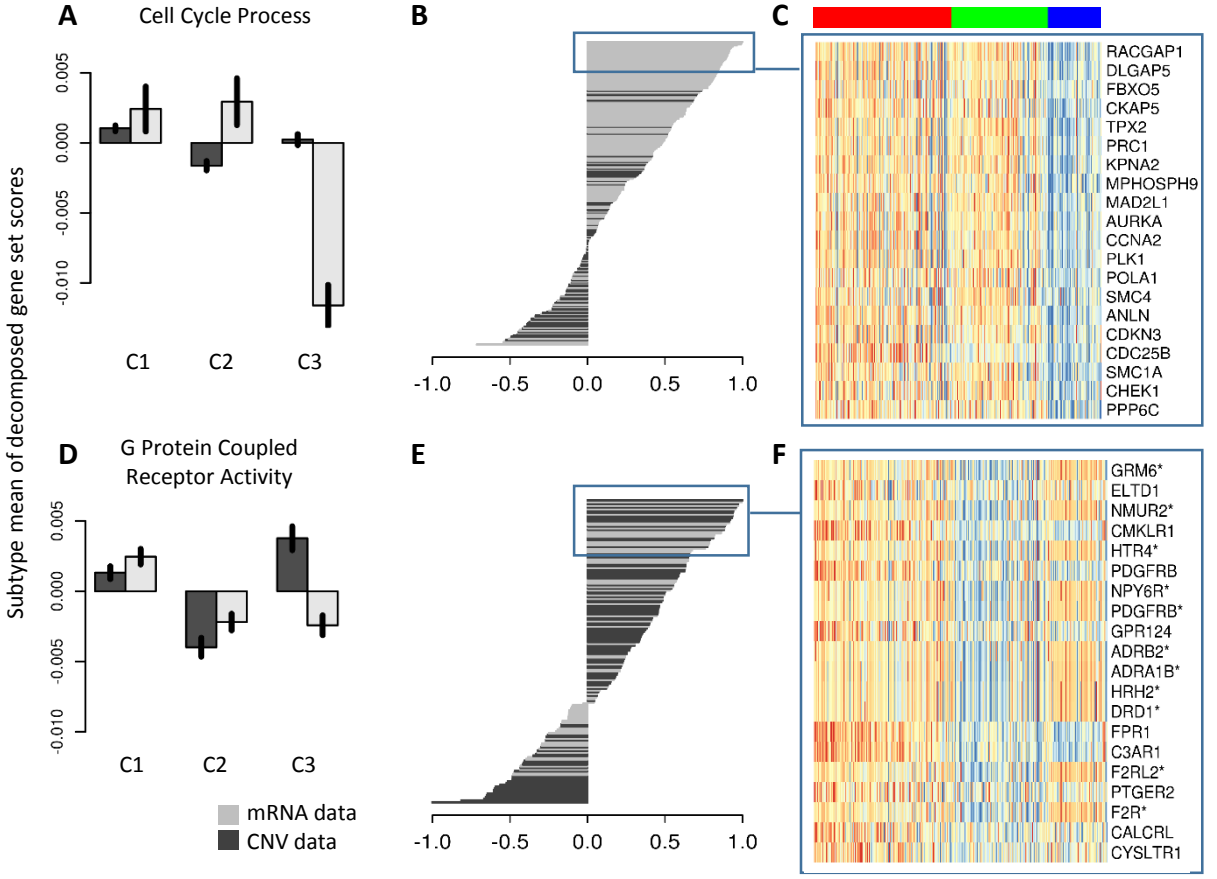

A
Cell Cycle Process
Subtype mean of decomposed gene set scores
0.005
0.000
-0.005
-0.010
C1
C2
C3
B
-1.0
-0.5
0.0
0.5
1.0
C
RACGAP1
DLGAP5
FBXO5
CKAP5
TPX2
PRC1
KPNA2
MPHOSPH9
MAD2L1
AURKA
CCNA2
PLK1
POLA1
SMC4
ANLN
CDKN3
CDC25B
SMC1A
CHEK1
PPP6C
D
G Protein Coupled
Receptor Activity
mRNA data
CNV data
E
F
GRM6*
ELTD1
NMUR2*
CMKLR1
HTR4*
PDGFRB
NPY6R*
PDGFRB*
GPR124
ADRB2*
ADRA1B*
HRH2*
DRD1*
FPR1
C3AR1
F2RL2*
PTGER2
F2R*
CALCRL
CYSLTR1

Figure 5

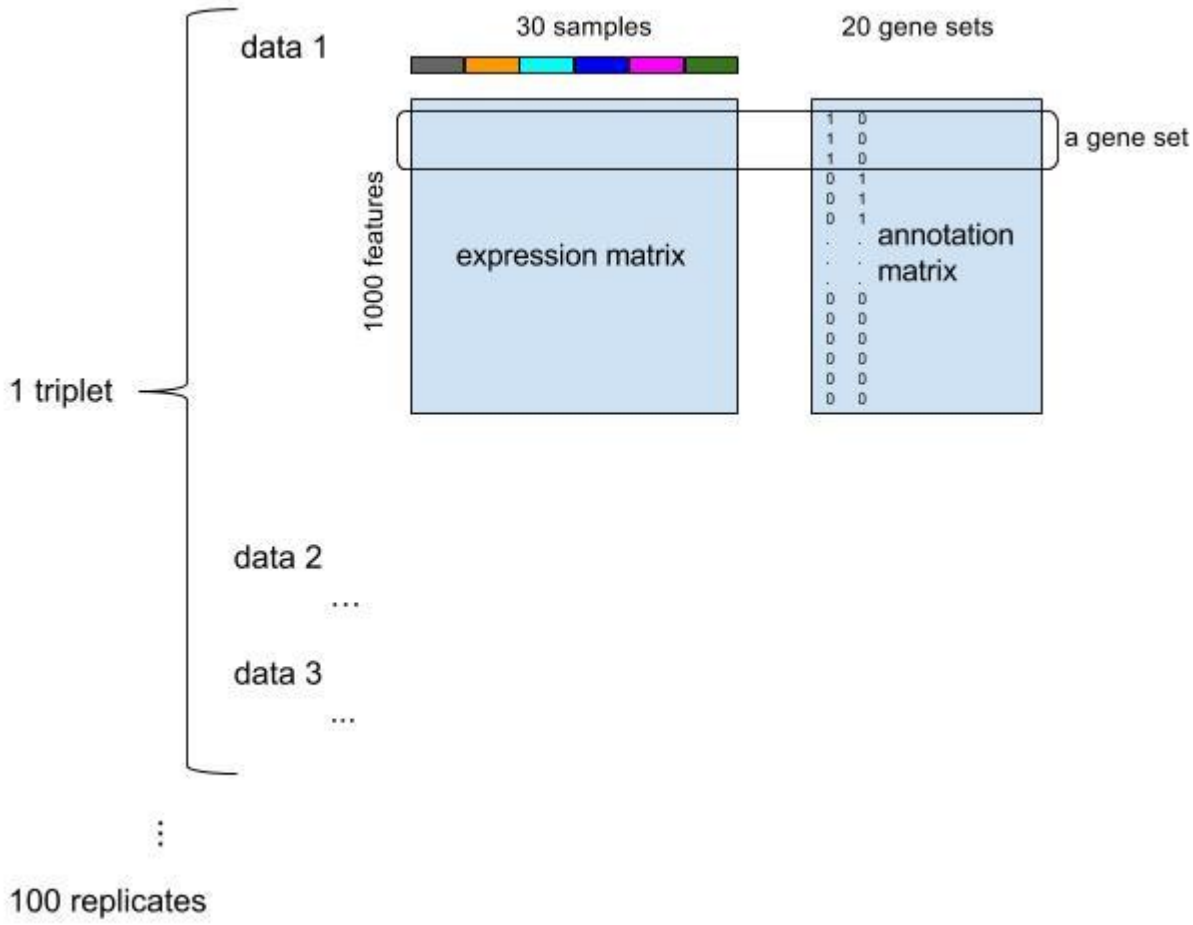


Figure S1 – A diagram showing the simulation data. One dataset contained of a matrix triplet (data 1, data 2, data 3). Each contained 1,000 features and 30 observations. The 30 observations were divided into six clusters. The 1,000 features had an annotation matrix which assigned features to 20 gene-sets, each gene-set had 50 genes. 100 triplets were simulated in this analysis.

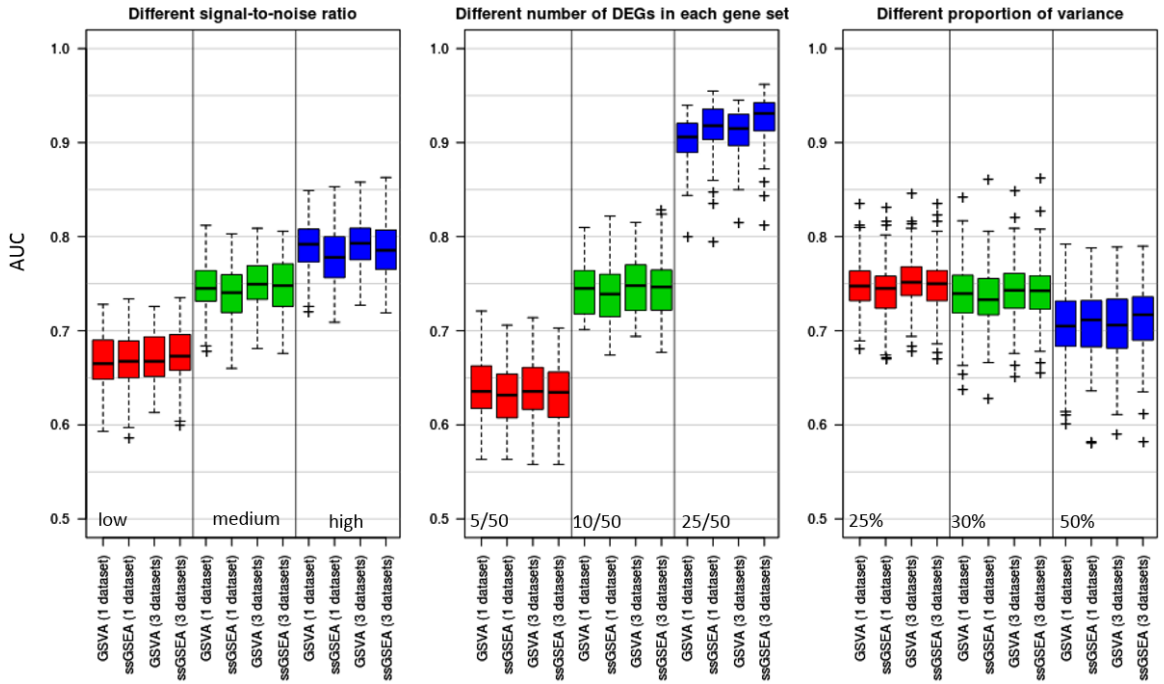


Figure S2 – Simple concatenation of multiple data sets did not improve the performance of GSVA and ssGSEA. Results show area under the curve (AUC) performance of GSVA and ssGSEA analysis of a single dataset (referred as 1 data set) and concatenated data sets (referred as 3 data sets) . Methods, data and evaluation are the same those in Figure 2.

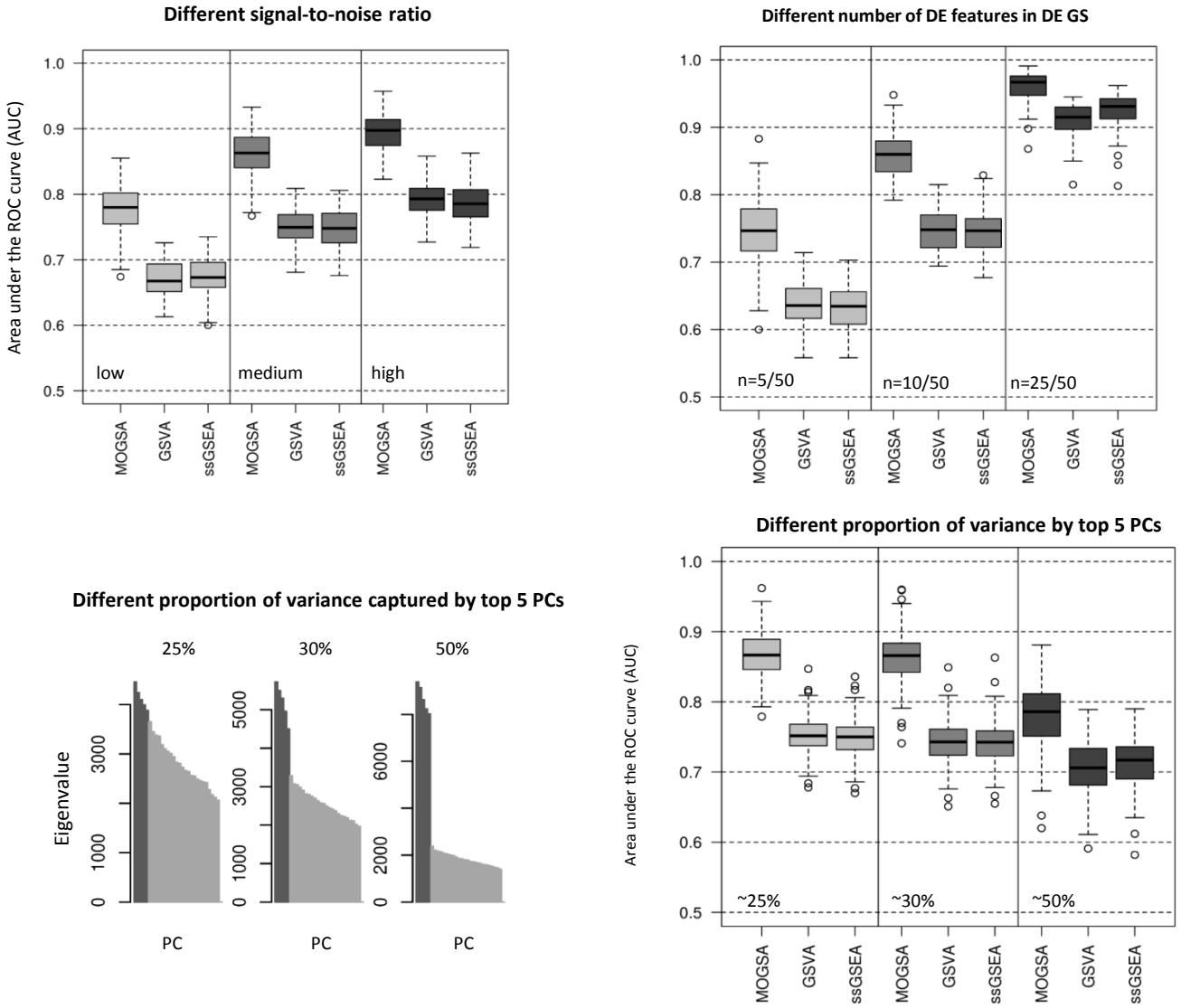
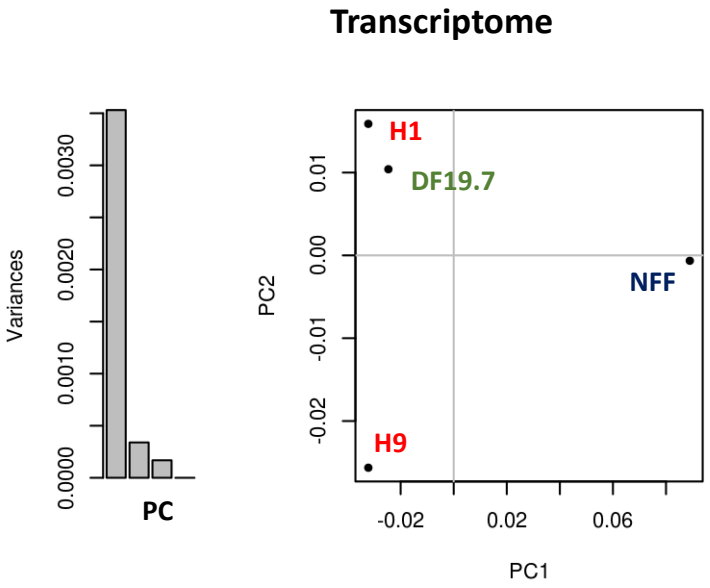
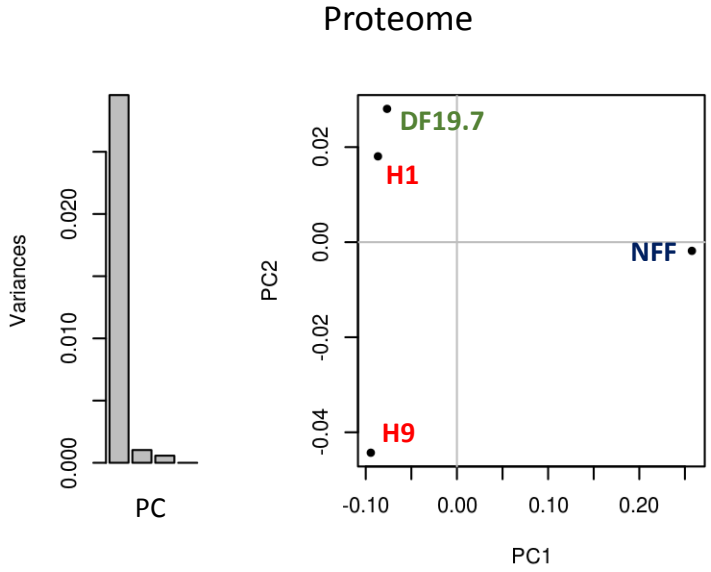


Figure S3 –moGSA outperforms GSVA and ssGSEA using weighted matrices. Because moGSA weights input matrices by their first singular value, we weighted the matrices in a triplet by their first singular value before concatenation. Methods were performed as described in Figure 2.



The iPS ES dataset contains:

- 1 fibroblast cell line (newborn foreskin fibroblast; NFF).
- 1 induced pluripotent cell line (iPSC; DF19.7)
- 2 embryonic stem cell lines (ESC; H1 and H9)



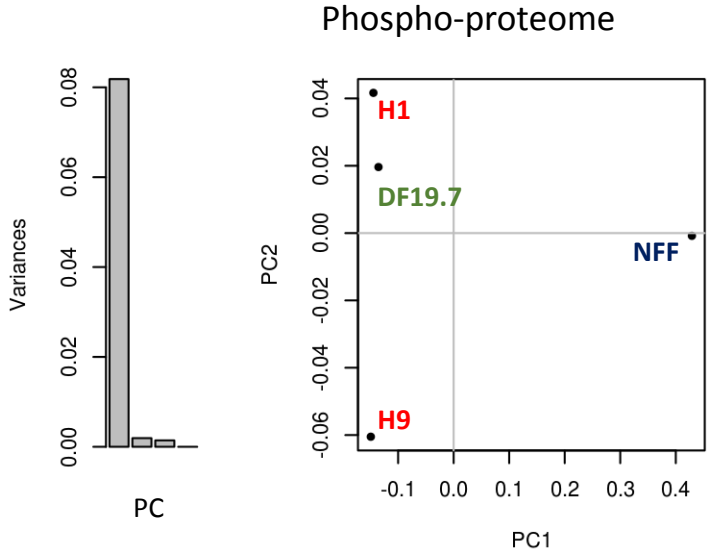


Figure S4 – PCA of data in iPS ES dataset. Principal component analysis of the 3 datasets in the iPS ES triplet. Most of the variance was captured by the first component which captures the difference between the fibroblast foreskin cells and the other samples. The second component captured information about molecular that distinguished the induce pluripotent and embryonic cells.

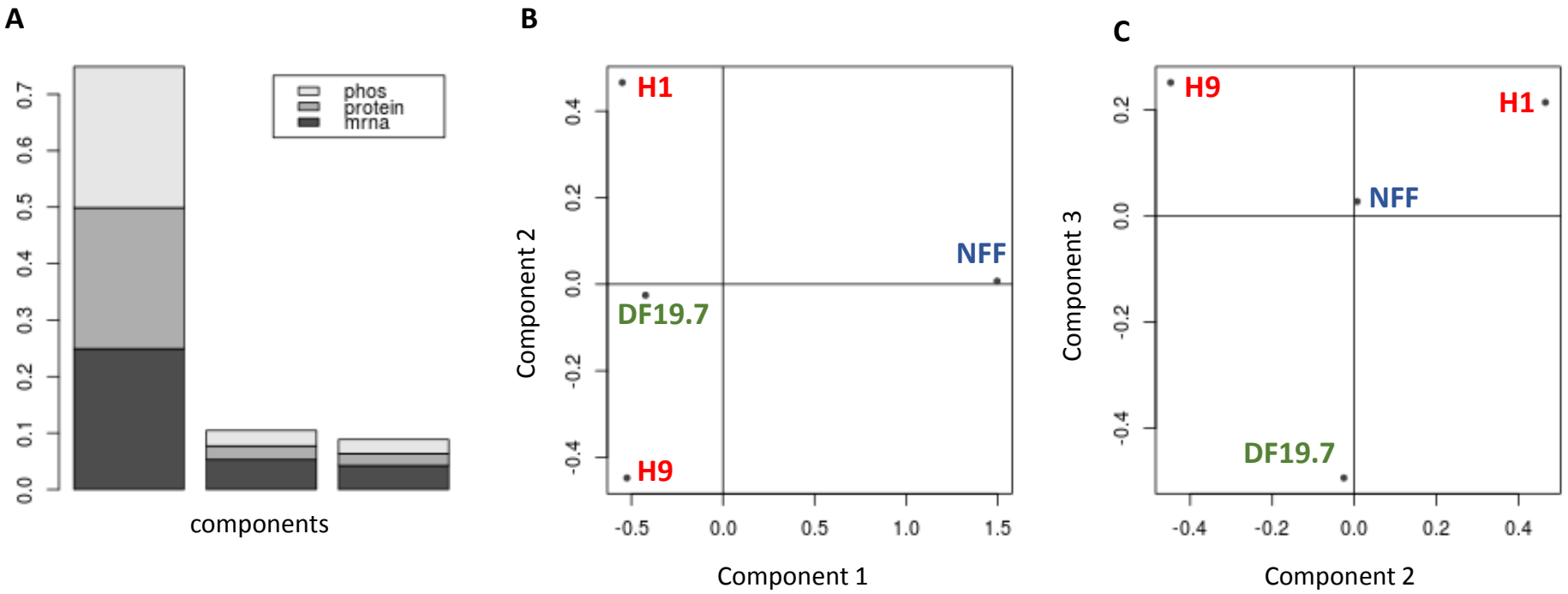


Figure S5 – moGSA of the iPS ES 4-plex data. (A) A scree plot of the eigenvalues of the MFA. Grayscale shades represent the contribution of each individual dataset and show each dataset contributes equally to the variance. Similar to PCA of the individual datasets (Figure S4) the first component captures most of the variance in the data. By plotting the first components of MFA (B) it can be seen that this first component captures the difference between NFF and pluripotent cell lines and the (C) third component represents the difference between **iPSC** and **ESC** lines. The three datasets contributed similar variance in the integrated analysis, as indicated by weighting of each dataset in MFA. The first eigenvalue (square of singular values) of each PCA were 0.24, 0.26 and 0.26 for the transcriptome, proteome and PhosoProteome dataset respectively.

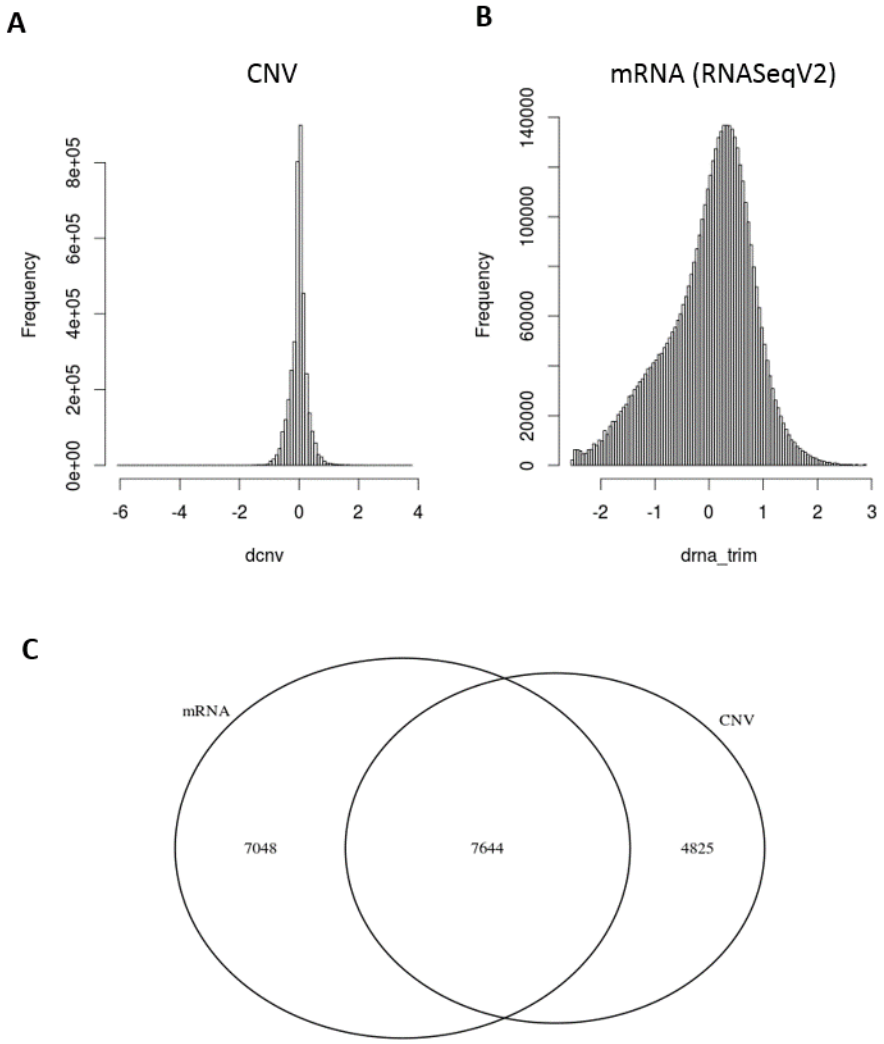


Figure S6 – Features in Copy number variation (CNV) and mRNA data of 308 TCGA muscle invasive urothelial bladder cancer (BLCA) patients. After filtering features with low variance (see Methods), CNV and RNA-seq data contained 12,447 and 14,710 genes respectively, in which 7,644 genes were common to both datasets. The distribution of the (A) CNV and (B) mRNA data and (C) overlap of common features is shown using a venn diagram

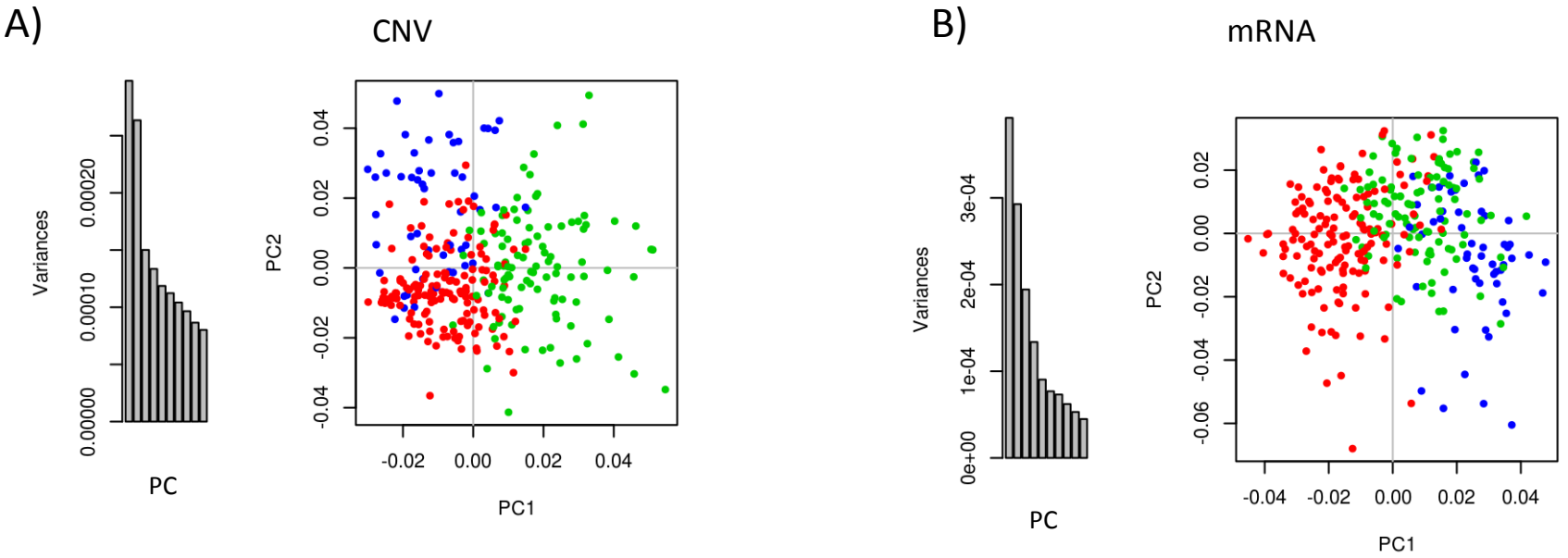


Figure S7 – Results PCA of (A) CNV and (B) mRNA RNA-seq gene expression of BLCA tumors (n=308).

Each panel shows a scree plot of the variance captured by the first 10 components and a plot of the first two components (PC1, PC2). Tumors are colored by molecular subtype; C1 (red), C2 (green),C3 (blue). The first two components of the CNV decomposition distinguishes these 3 subtypes. . The first eigenvalue (square of singular value) of a PCA of BLCA mRNA and CNV data were 0.0004 and 0.0003 respectively. This was used to weight each data in MFA and indicates both datasets contributed similar variance to the integrated analysis.

A

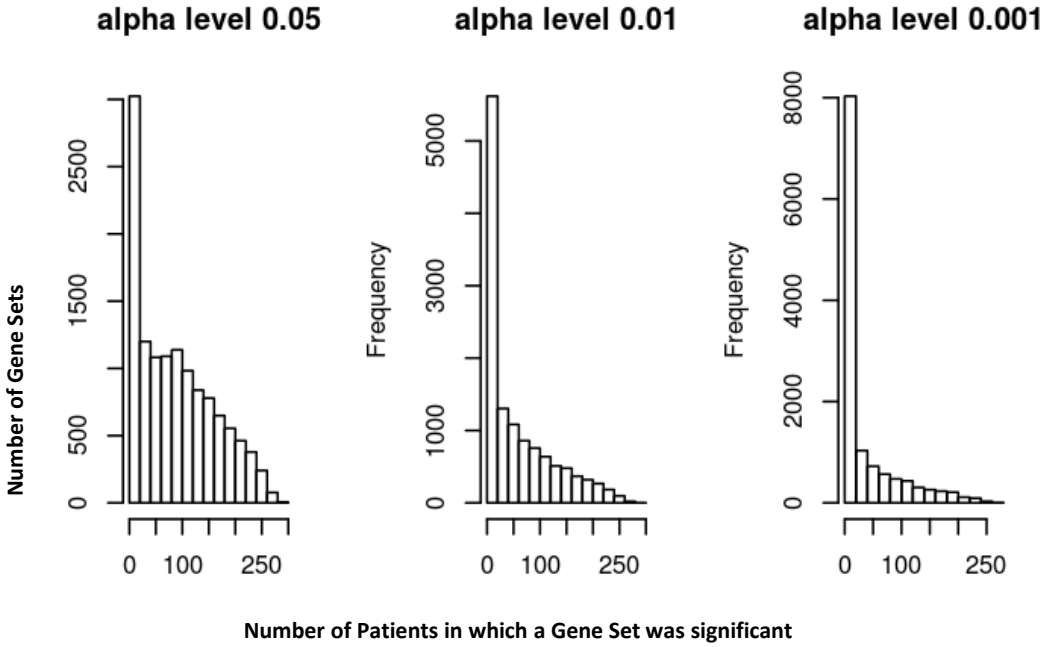


Figure S8 – Number of significant gene sets reported by moGSA in BLCA

moGSA is a single sample GSA approach that reported significant genesets in each BLCA tumor (A) shows the distribution of significant gene set at $p < 0.05$, $p < 0.01$ and $p < 0.001$. moGSA was performed with 5 components. Most genesets were insignificant across all or most patients and no gene set achieved a sum of 308 (significant in all patients) . Among the genesets that were significant in at least 1 patient, the median number of patients in which a geneset was significant was 93, 61 and 46 for $p < 0.05$, $p < 0.01$ and $p < 0.001$ respectively.

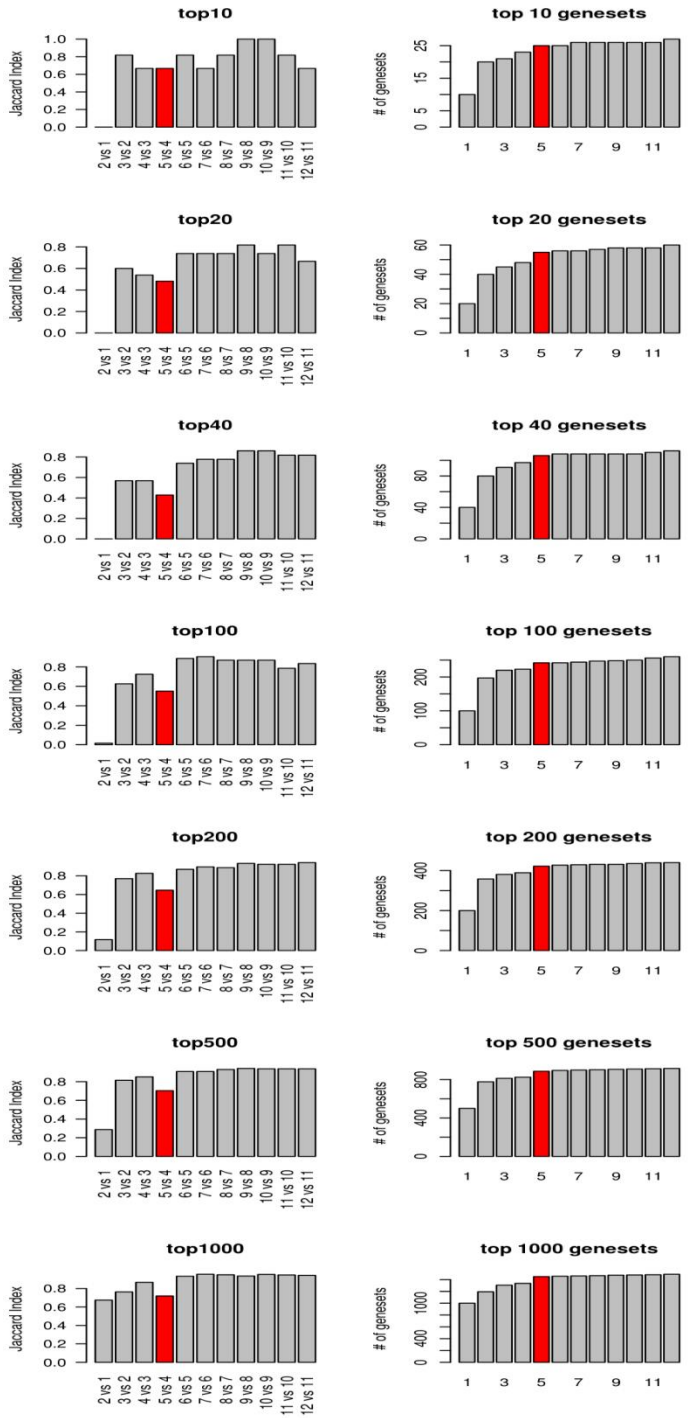


Figure S9 Effect of number of component (from 1 to 12) on moGSA of BLCA tumors (n=308) integrating mRNA and CNV data.

Genesets were filtered to those that significant ($p<0.05$) in each patient and the sum of patients in which each geneset was significant was calculated. Genesets were ranked (high to low) by the number of patients in which it was significant.

Then the top N (N=10, 20, 40, 100, 200) gene sets were selected and the Jaccard similarity coefficient was used to compare the overlap in highly ranked genesets when the number of components was between 1 and 12. A Jaccard Index (JI) of 1 would indicate that the sets are identical, and an index of 0 indicates no overlap. The left panels the similarity in the top genesets between pairs of analyses in which X or X+1 component were used between where X is between 1 and 11. Across a range of components (2 to 12), the top 10 most highly ranked genesets have high overlap, the JI varies between 0.6 and 1.0, reflecting an intersection of size 7 to 10. When a larger number of genesets are examined, increasing the number components is associated with a higher and more stable JI, however no additional gain in JI is achieved after 5 components.

The right panel shows the union of genesets identified when additional components are examined .In figure F, with one component we extract 10 genesets, when we add a second components , we extract 20 genesets and these have little overlap (consistent with panel A). However only a few new genesets were identified by adding a further component (3 components) as 3 components identified 22 genesets (????). In general additional component identify further genesets but there is little gain in genesets after five components.

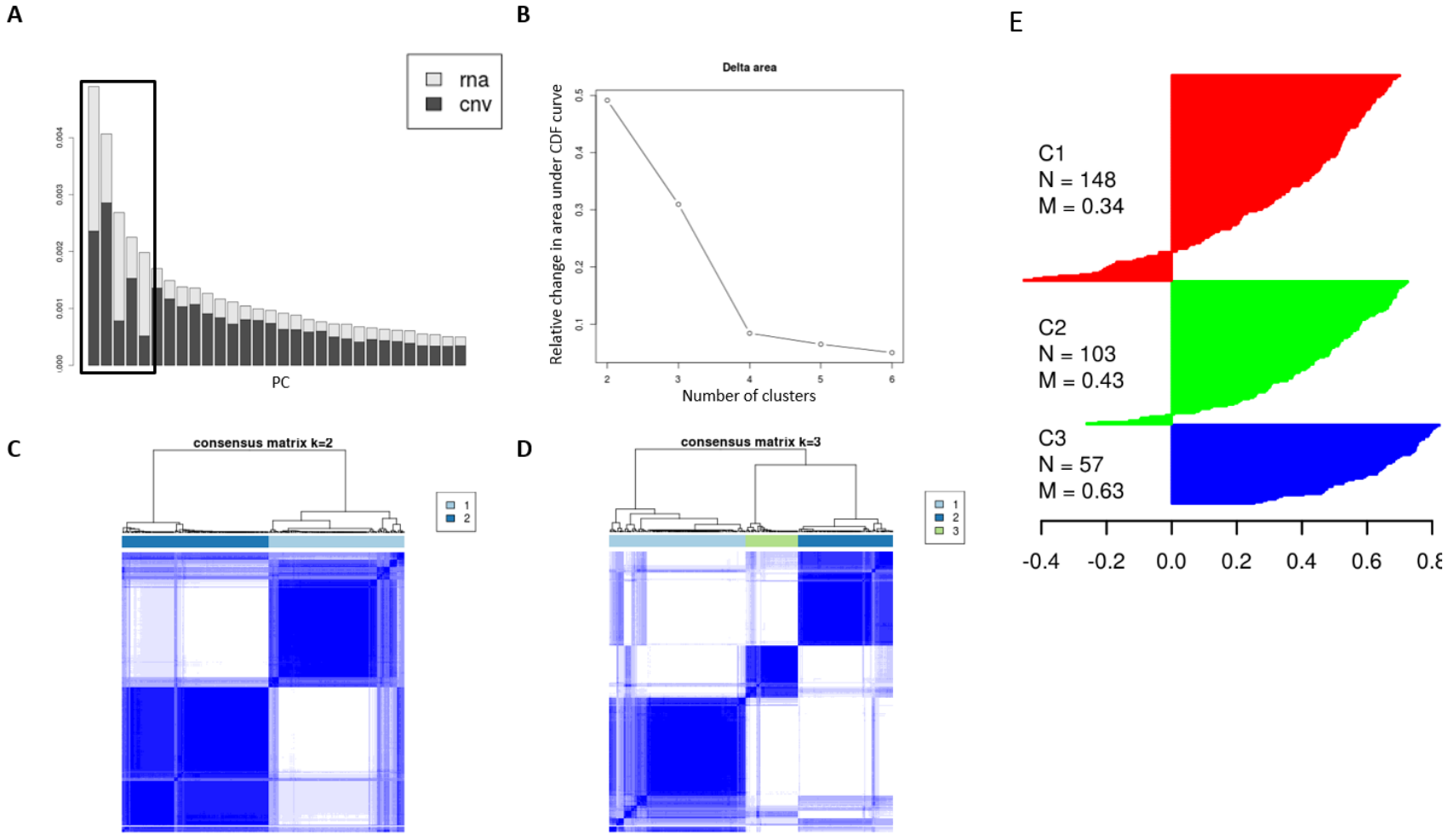


Figure S10 –Clustering of MFA latent variables identify three BLCA subtypes. MFA of mRNA and CNV of BLCA patients was performed . (A) shows the eigenvalues of each of the latent variables and top five PCs are marked. Five latent variables were used in consensus clustering and (B) the relative change area under the CDF curve (y-axis) over different pre-defined number of clusters (x-axis), which is used to determine the number of clusters. For both 2 and 3 clusters, the relative change in area under the CDF cure is high, indicating that either that the BLCA tumors may contain 2 or 3 subtypes . However hierarchical of the consensus matrix for (C) 2 or (D) 3 subtypes together stability analysis (Figure S10), predictive strength analysis (shown in Figure 4b) and (E) silhouette analysis the data supported 3 clusters. C3 was highly robust but there was a number of unstable patients in C1 and C2

Figure S11 - Stability of BLCA molecular subtypes.

Clustering results using different proportion of samples in the resampling. We used model defined by 80% samples resampling (top bar). But the clustering result is similar in terms of different proportion of resampling samples Compared to 80%, a few C1 samples did cluster with C2 or C3 with 50%,60% or 70%, 80% respectively

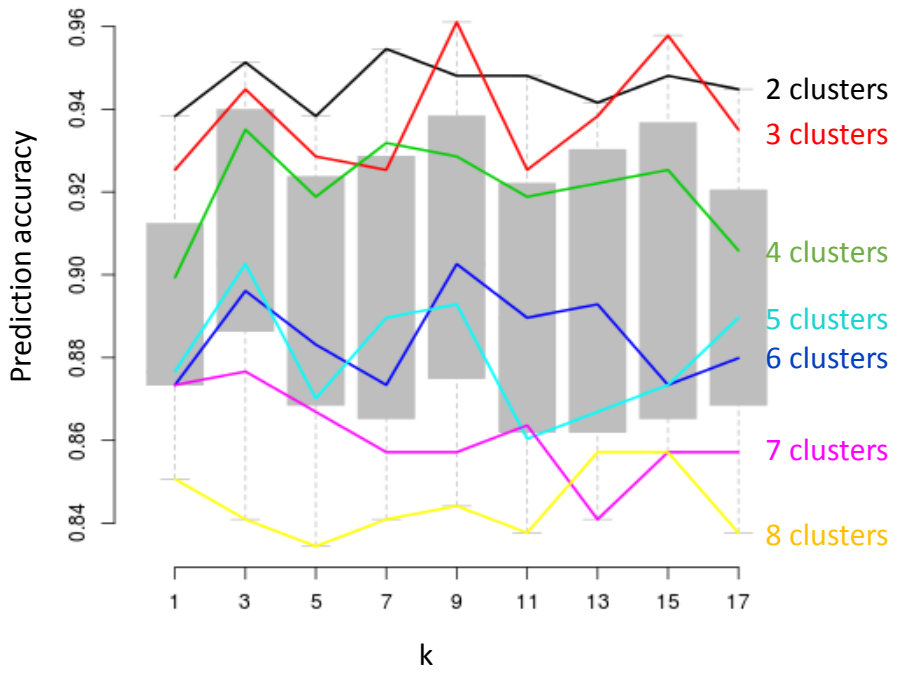
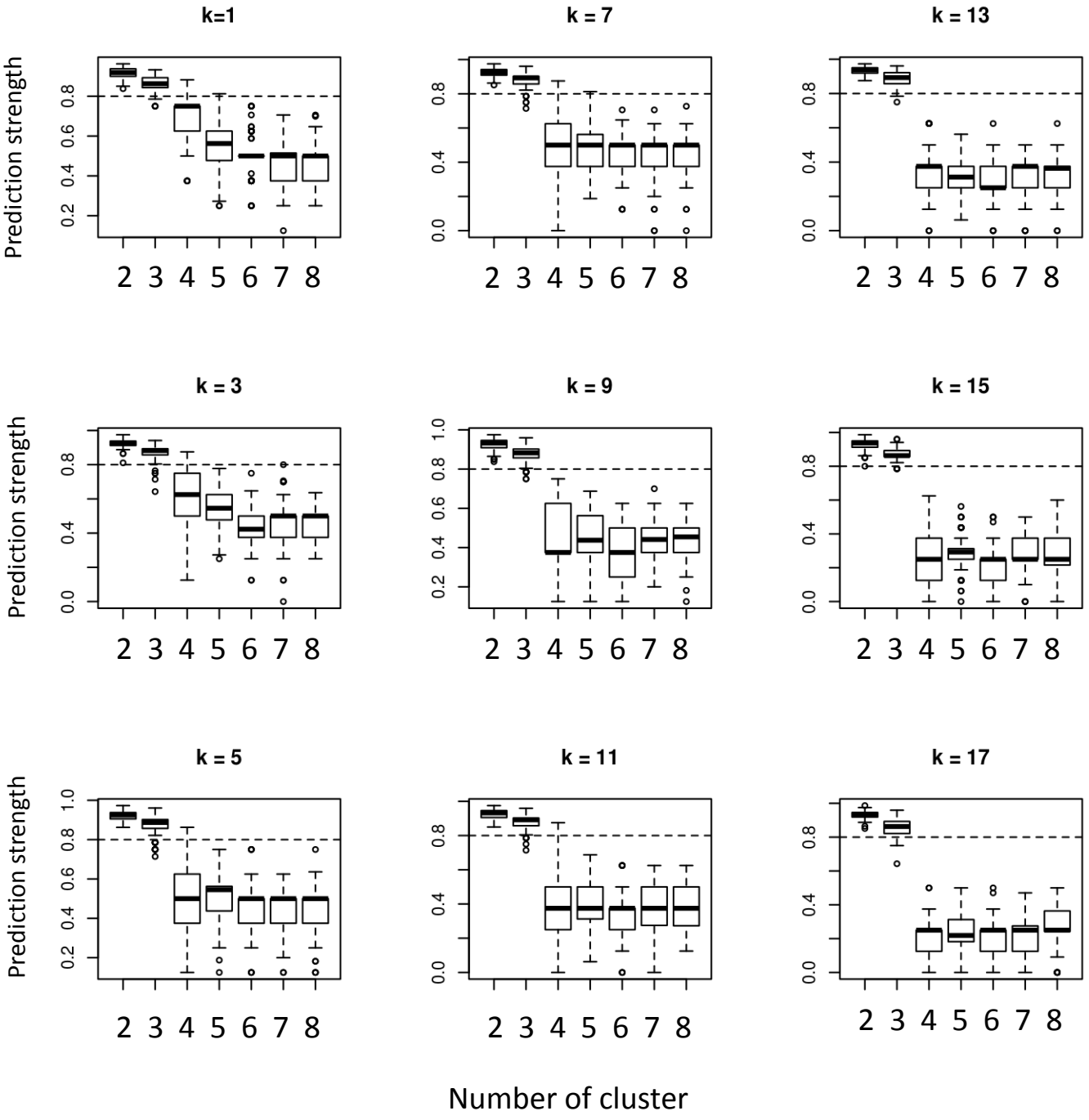


Figure S12 – Determining the number of clusters in the BLCA data (KNN/Prediction Strength)

Cross-validation were used to optimize the optimal number of K in the KNN classifier. We evaluated odd numbers K from 1 to 17. The performance of classifier were measured with prediction accuracy (y-axis). There is not a K clearly better than the others.

Figure S13 – Prediction strength using different K in KNN classifier.

All K suggest that three subtype is the robust number of subtype in the integrated BLCA datasets.



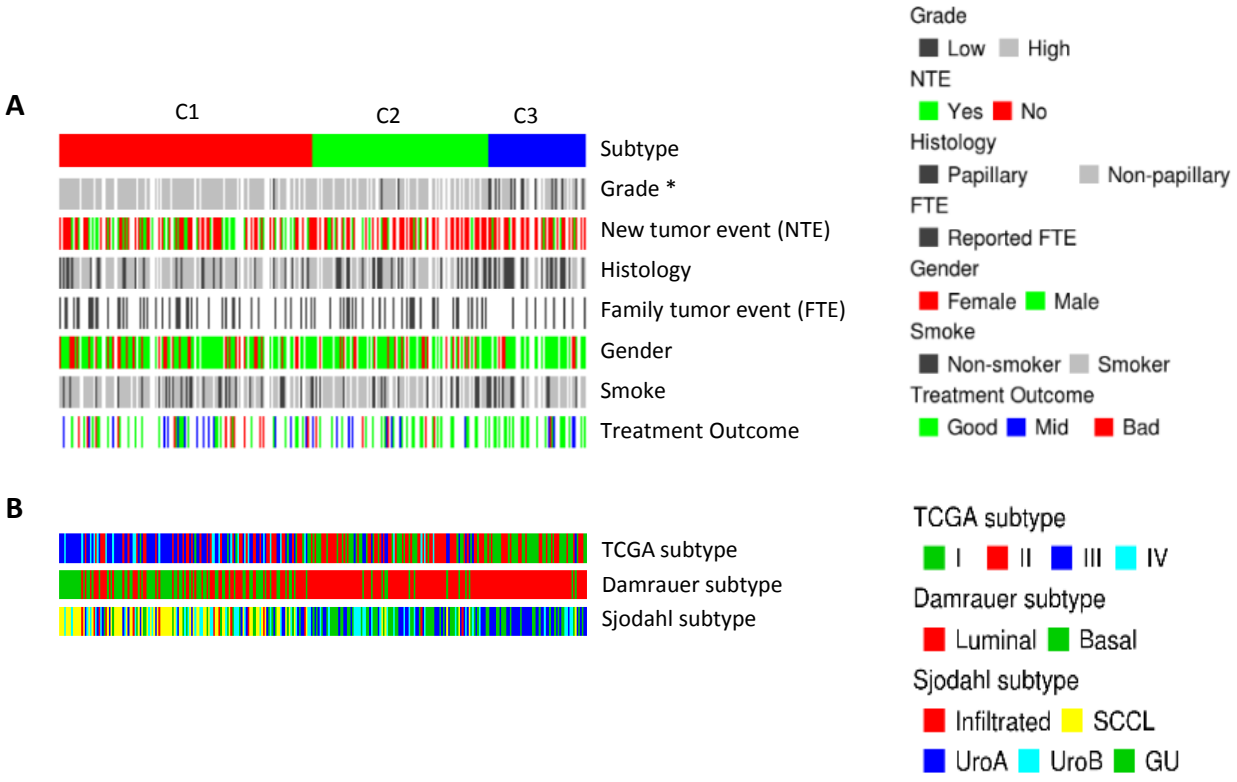


Figure S14 – Characteristics of the BLCA molecular subtypes. (A) Enrichment of clinical/phenotype factors including smoking gender, new tumor events, etc ib subtypes was studies. Grade was significantly correlated with the subtypes ($\chi 2$ test, FDR BH corrected p value < 0.01). (C) There was strong concordance between theC1,C2,C3 subtypes and molecular subtypes previously reported by the TCGA, Damrauer et al. and Sjodhl et al. C1 was enriched with III and IV for TCGA subtype, Basal subtype in Damrauer subtype and the SCCL and Infiltrated subtypes in Sjodahl subtype. C2 and C3 is comparable to the luminal subtype in Damrauer subtype model. C3 also enriched with UroA subtype in Sjodahl subtype and type I in TCGA subtype model.

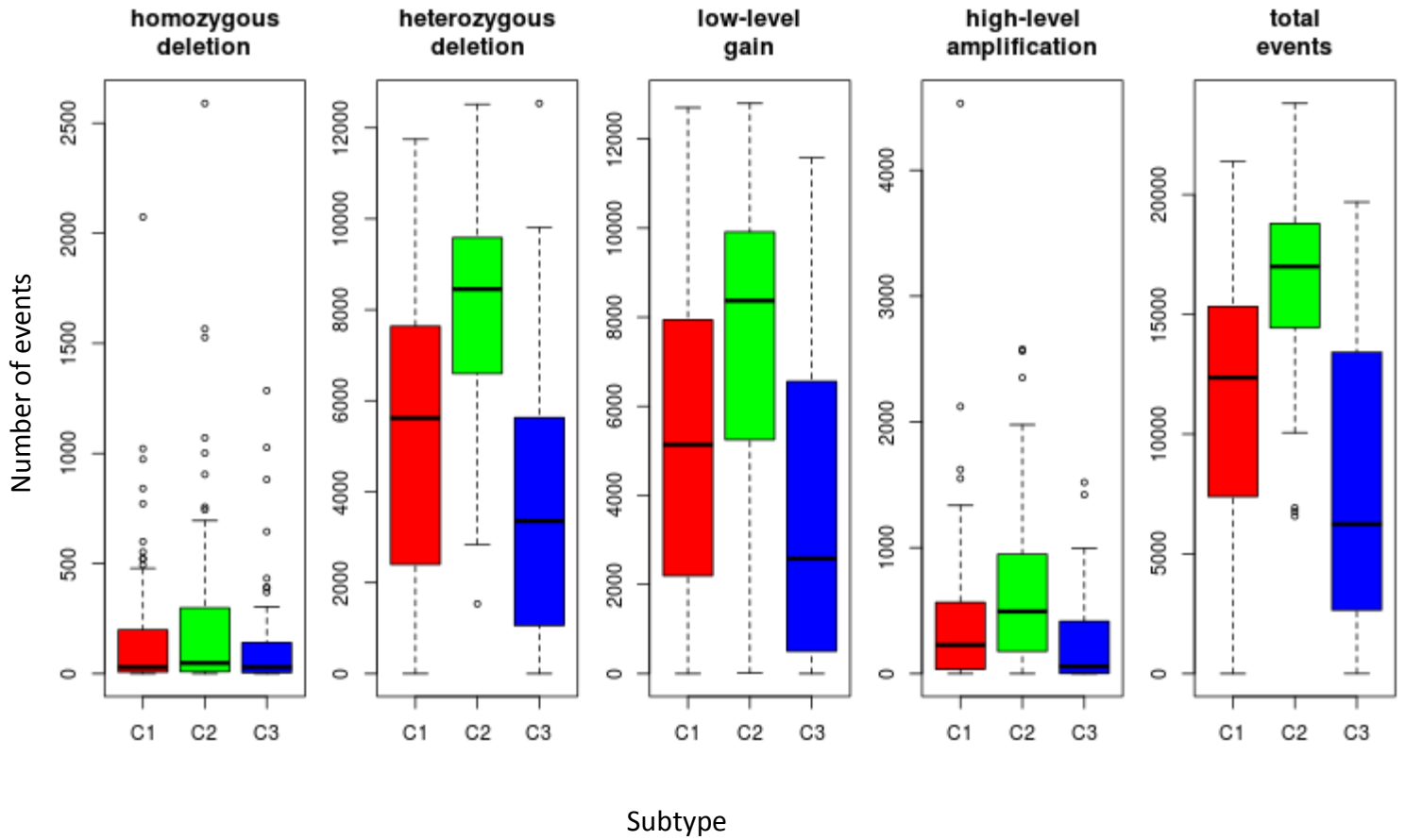


Figure S15– BLCA subtype C2 tumors have more instability and higher numbers of mutation events

Plots show the numbers of homozygous or heterozygous deletions, low and high level gains in addition to total CNV events in the genome of BLCA patients (n=308).

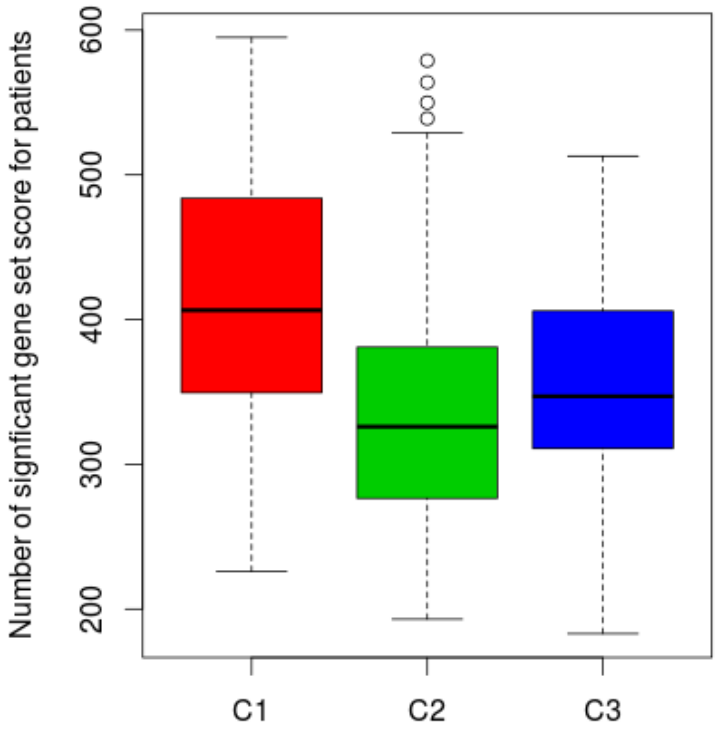


Figure S16:. The number of gene-sets with significant GSS at p<0.05 in each molecular subtypes of BLCA tumors (include positive and negative GSS)

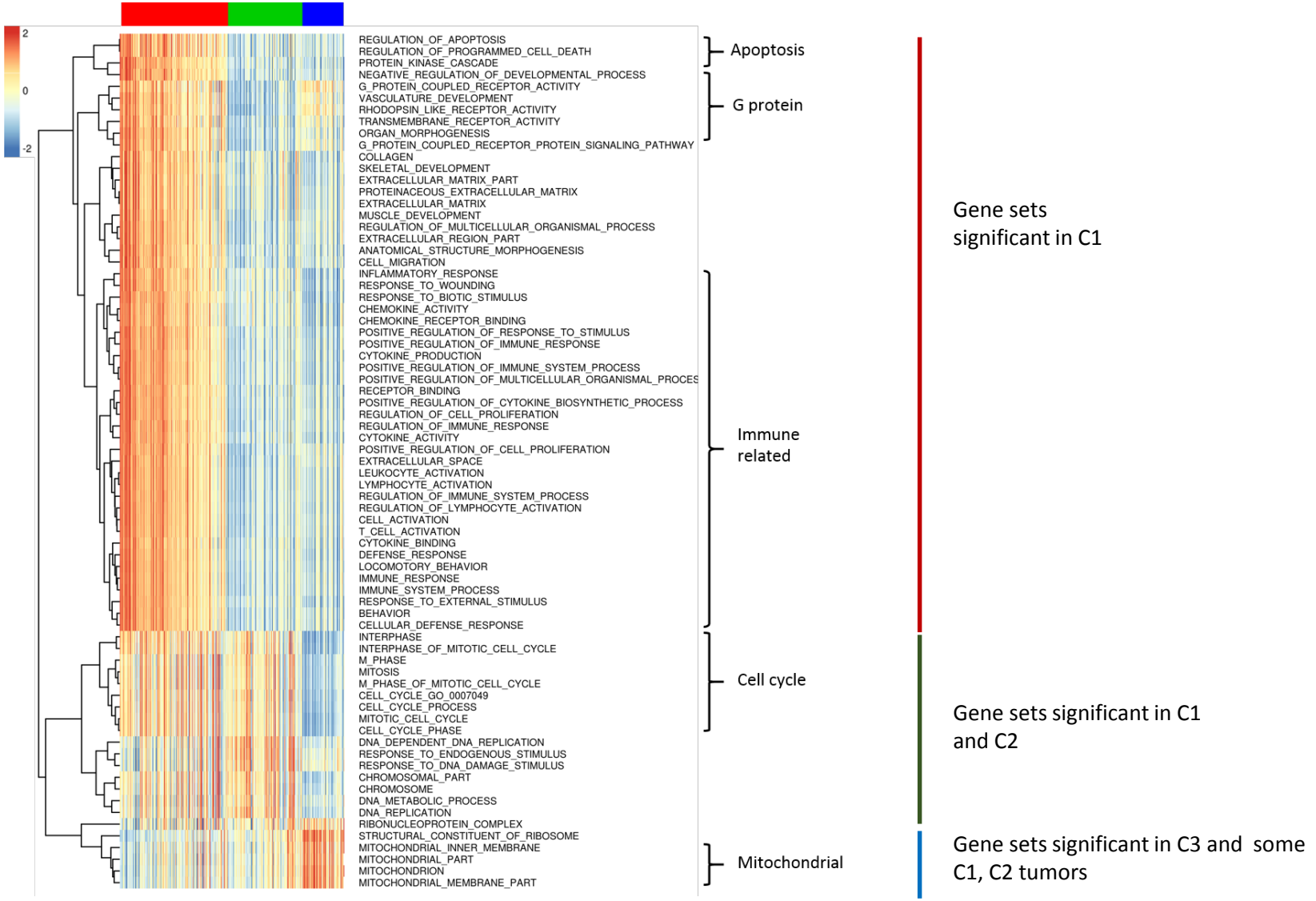


Figure S17 – Genesets with Gene Set Scores (GSS) that were significant (p<0.05) in many BLCA patients (n >=200 patients). The rows (gene sets) of the heatmaps are clustered so that the gene sets with similar GSS scores across patients are grouped. Columns are ordered BLCA tumor molecular subtype (C1,C2,C3). Genesets formed three broad clusters (those significant in C1, C1 and C2 or C3 and other tumors). Significant genesets in C1 were associated with apoptosis, G protein coupled proteins, extracellular function, muscle development and Immune reponse. Gene sets significant in both C1 and C2 were mostly associated with the cell cycle, DNA repair and replication. Gene sets significant in C3 patients were associated with the mitochondria.

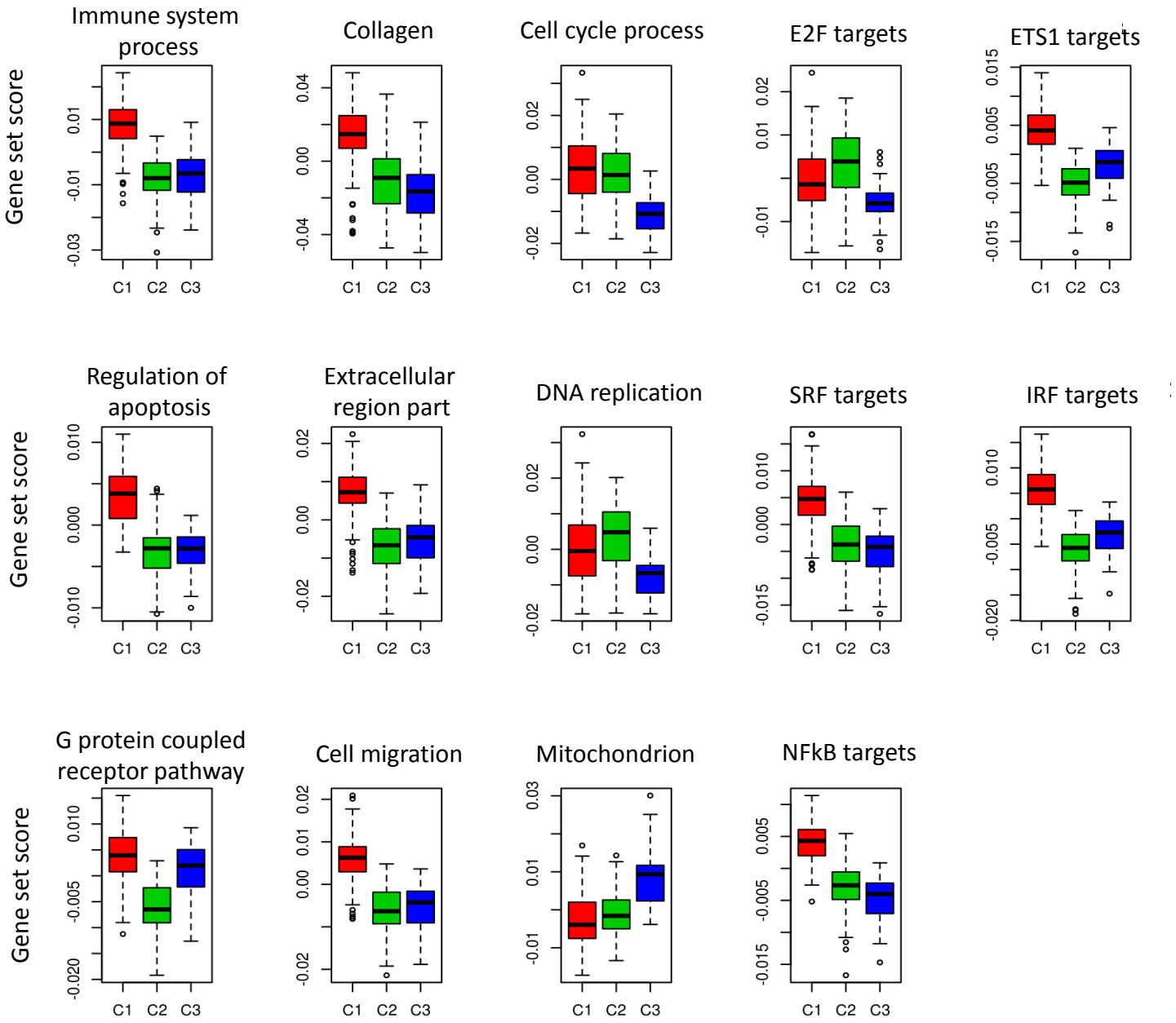
Figure S18 –

The distribution of gene set scores in different BLCA molecular subtypes.
Boxplot of a subset of genesets (from Figure S12) Gene sets are also shown in Figure 4C and D.

Immune processes, Regulation of apoptosis, and cytosketal genesets were upregulated in C1.

C2 was characterized by downregulation of ETS1 and IRF targets, G-protein coupled receptors pathways and increased DNA related pathway (possibly associated with increased genome instability)

C3 had lower expression of cell cycle and DNA replication genes compared to C1 and C2



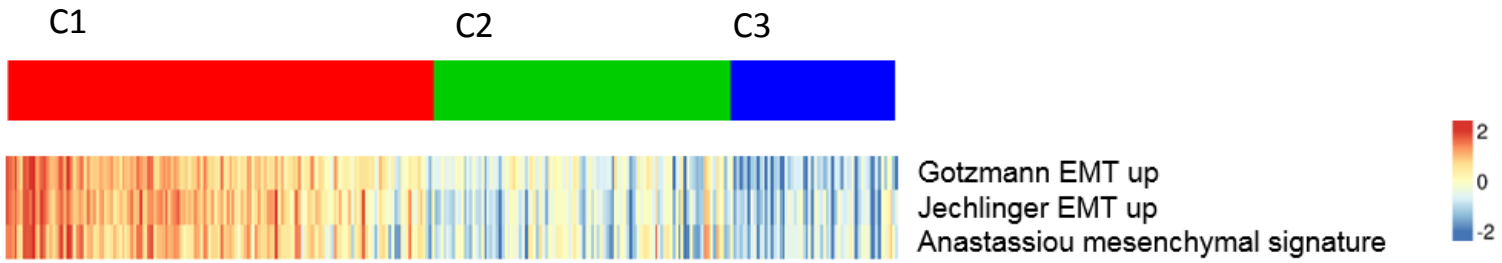


Figure S19 –EMT related gene sets are highly activated in the C1 subtype.

Heatmap displays GSS for three mesenchymal related gene sets

* The three genes set are derived from MSigDB C2 curated signatures. The original names as annotated in the MSigDB are:
“GOTZMANN_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP”,
“JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP” and
“ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE”.

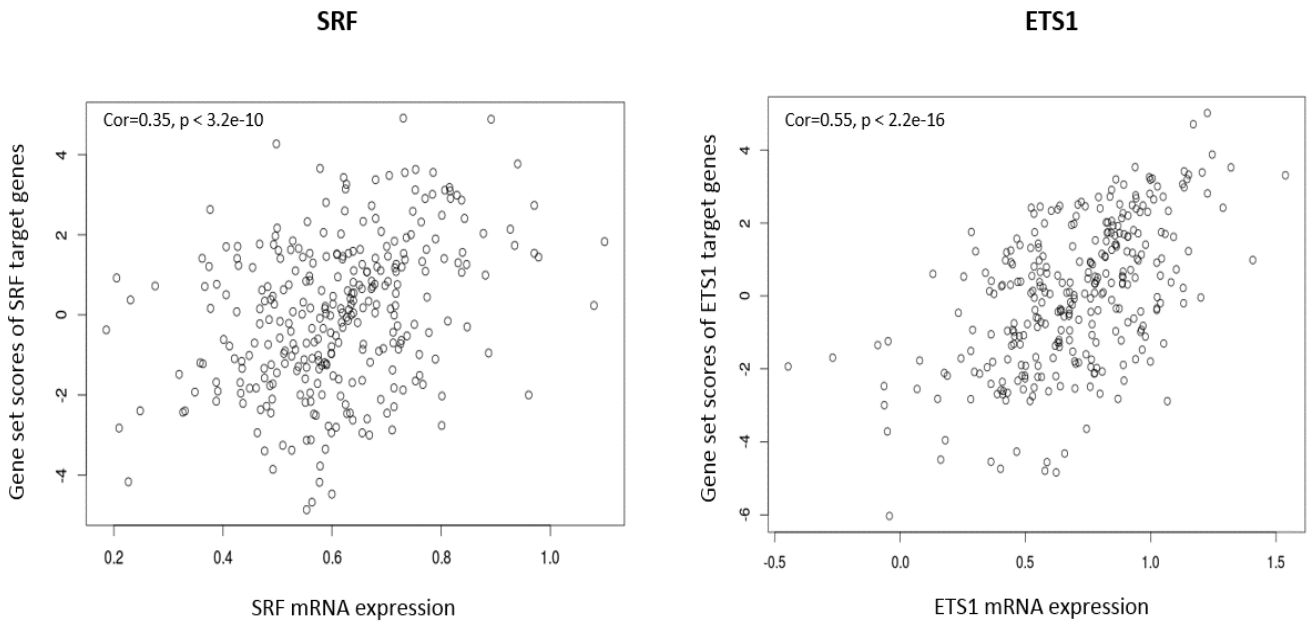


Figure S20 – Gene sets scores of transcription factor (TF) target genes were highly correlated with the mRNA expression of their transcript factors in tumors. Scatter plots show gene set score and mRNA expression levels of transcription factors (A) SRF and (B) ETS1 in the 308 BLCA tumors.

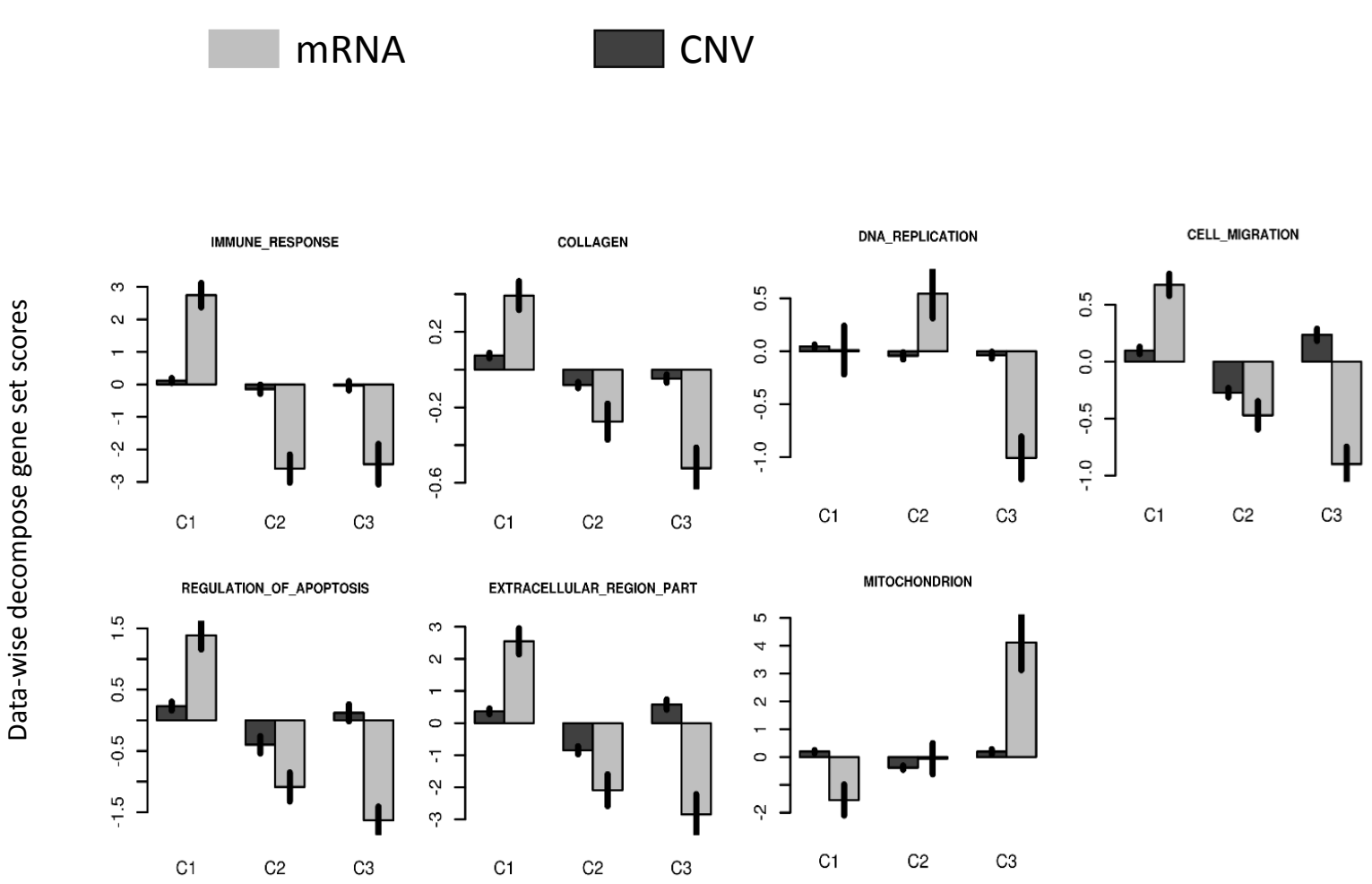


Figure S21 – moGSA gene set scores integrate data from multiple data sources

Barplots of decomposed gene set scores for selected gene sets (shown in Figure 5). The mean of decomposed GSSs for mRNA (light grey) and CNV (dark grey) is shown for each molecular subtype. Black segments on the bars represent 95% confidence interval of the mean. Y-axis is decomposed gene set score.

Figure S22 –Non-normalized gene set scores of genesets that were significant in iPS ES cells (shown in Figure 3).

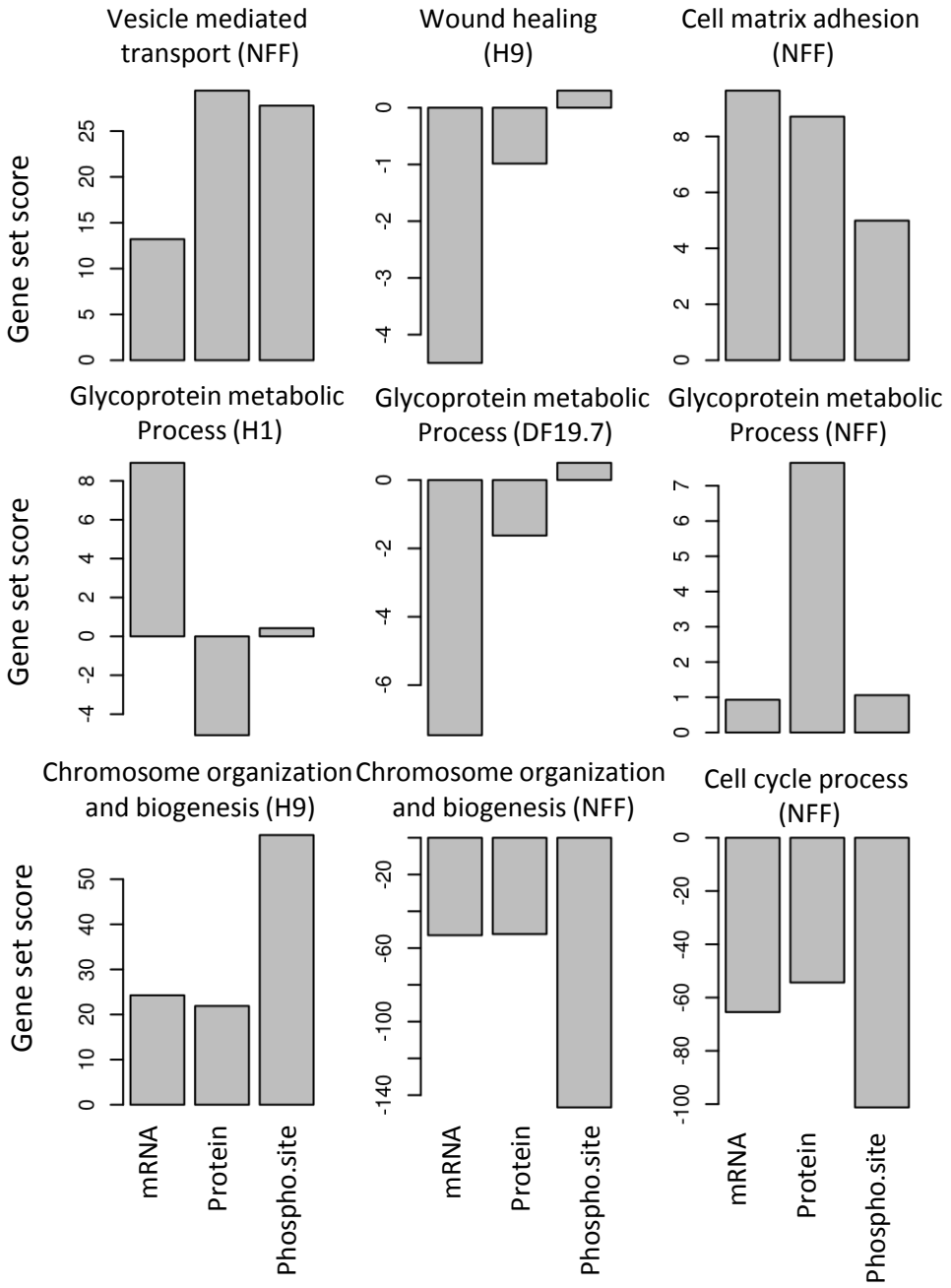
Normalized geneset scores are reported throughout this article, but here we shown gene set scores that have not been normalized by gene set length.

A raw gene set score is the sum of the contributions of all genes in that gene set. As a result gene sets with more genes tend to have higher scores and GSS are comparable.

This plot shows that the non-normalized GSS range (y-axis) varies between gene sets and is generally associated with the number of genes in a gene set. The number of matched features per geneset is:

vesicle mediated transport: 385
wound healing: 39
cell matrix adhesion: 59
glycoprotein metabolic process: 100
chromosome organization and biogenesis: 392
cell cycle process: 437

Therefore moGSA normalizes the gene set scores by dividing by the length of the geneset.



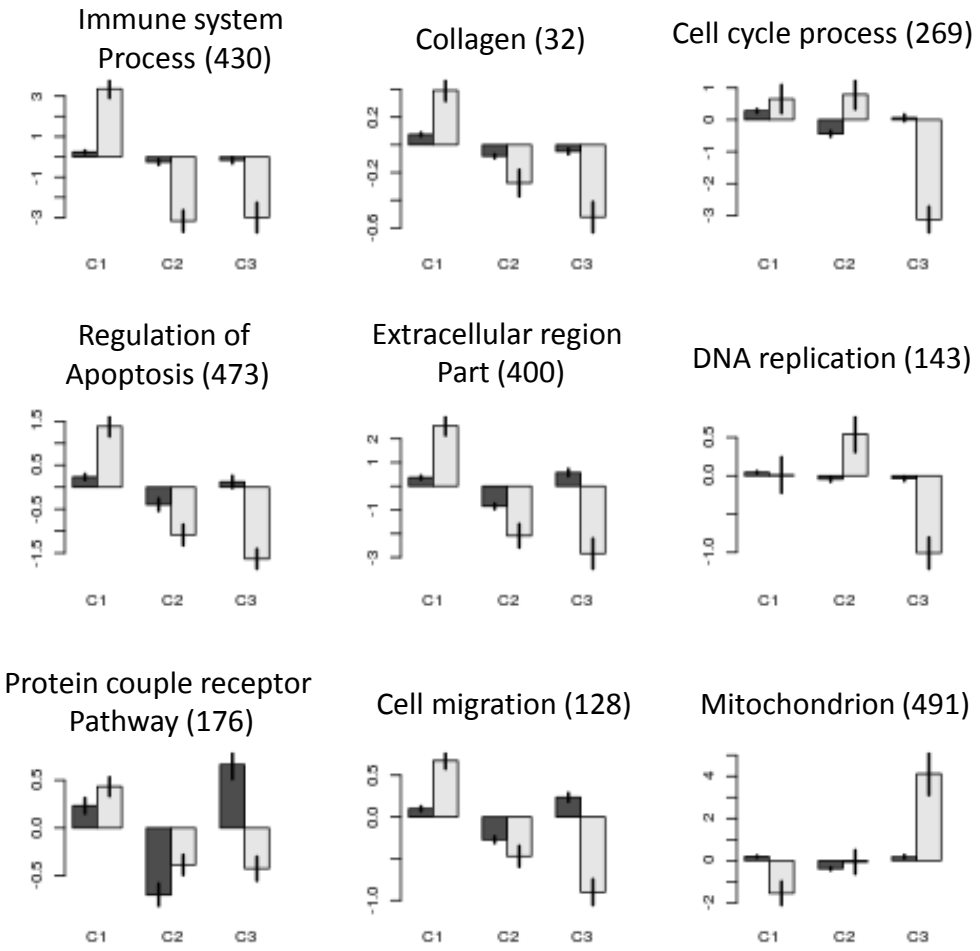


Figure S23 –Non-normalized gene set scores of genesets that were significant in BLCA tumors (shown in Figure 4,5 and S18).

The raw GSS is sum of the contributions of genes in a gene set and there fore the scale of non-normalized gene set scores (y-axis) are different. Gene sets with more genes will have higher scores and GSS will not be comparable within a study. Therefore moGSA normalizes raw GSS by gene set length. Normalized GSS are reported throughout this article. Further description of the GSS in the legend of Figure S21) and in the methods. Plots are labelled with the gene set name and the number features (genes) in each gene set which is shown in parenthesis.

Table S1

Table S2

Click here to access/download

Supplementary Material

Table_S2_BLCASubtypeComparisonContingencyTable.xlsx

Table S3

Table S4

Table S5