Using Multi-omic Cancer Data to Find Ways to Improve the Treatment of Bladder Cancer

Zhangdaihong Liu, Giovanni Mizzi, Iliana Peneva, Esther Wershof

June 17, 2015

Abstract

In recent literature on bladder cancer there is call time and time again for better identification of cancer subtypes. Broadly, bladder cancer can be split into two types: muscle invasive and non-muscle invasive. By successfully identifying sub-phenotypes of bladder cancer using multi-omic data, we hope to be better able to personalise cancer care and choose appropriate treatments. To achieve this goal, we analysed the rich data source available to us via The Cancer Genome Atlas, in particular focussing on Methylation, Gene Expression and Copy Number Variation data. We found evidence supporting current research on using the gene AQP1 as a bladder cancer biomarker. We classified a subgroup of patients with a higher mean age, who had highly methylated genes but better survival prognosis. Further, we identified ways in which multi-omic data could be used to provide more detail than the currently used TNM system. Finally, we performed Multiple Dataset Integration (MDI) to provide a way of combining the different data types.

1 Introduction

In recent years there have been huge advances in the quantity, variety and affordability of multi-omic data. The cost of genome sequencing has decreased from $100m in 2001 to under $1000 as of 2014, outpacing Moore's law of complexity [1]. These remarkable developments mean that in the near future we could reasonably expect these data to be available for all cancer patients marking a major shift in modern medicine to more personalised care. There is a need to use these data effectively to aid in the diagnosis, treatment and prevention of cancer.

Bladder cancer is the most common cancer of the urinary tract and it is the ninth most common cancer worldwide [2]. In Europe and North America, more than 90% of bladder cancers are urothelial carcinoma [3]. These tumours are currently assessed using the Tumour-Node-Metastasis system (TNM), which amongst other measures, gives the tumour a grade (Ta-T4). T2-T4 describe tumours which are muscle invasive and frequently metastasize.

The aims of our project, suggested by our external partner University of Birmingham Cancer Sciences, were as follows:

- Look at specific data types to identify sub-phenotypes associated with survival.
- Compare outcome prediction from various data types. Which types of data give us the best predictions of likely disease course, outcomes, and patient survival?
- Combine multiple data types to improve outcome prediction. Can we get better predictions by combining a range of different types of data?

To achieve these goals, we analysed the urothelial bladder carcinoma dataset available to us via The Pan-Cancer Analysis Project as part of The Cancer Genome Atlas (TCGA) [4]. TGCA is a large online catalogue containing multi-omic data and the Pan-Cancer project analyses 12 tumour types profiled by TCGA [5]. It is hugely beneficial not only to look at the multi-omic datasets individually but to fuse them into one multidimensional dataset and stratify patients in this way. By doing this we can see where the data types agree and can learn from areas where they contradict each other.

2 Methods

2.1 Data

In the urothelial bladder carcinoma dataset from TCGA, there are a variety of different data types available. We chose three data types: copy number variation (CNV), gene expression (GE) and

	CNV	*MUT*	*GE*	*METH*
CNV	300	4	1	4
MUT		300	6	10
GE			300	3
METH				300

Table 1: This shows the overlap between different data types when looking at the 300 most significantly altered genes for each.

methylation (METH).

Copy number variations are alterations of the DNA that result in the cell having an abnormal number of copies of one or more sections of the DNA. They correspond to relatively large regions of the genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosomes. Variations in copy number have been associated to many diseases. Recent evidence also shows that in cancer cells, copy number can be elevated [6]. Elevating the copy number of oncogenes can increase the expression of the protein that they encode so that cancer can be exacerbated. On the other hand, elevating the copy number of suppressor genes can help mitigating cancer [7, 8].

Gene expression measures the amount of mRNA being produced by each gene, participates in protein production to a significant degree and captures gene regulation which gives the cell control over structure and function [9]. Gene expression is a very fundamental level at which the genotype gives rise to phenotype. Analysing gene expression level is a typical and straightforward way of identifying cancer related genes including both inhibitors and drivers.

DNA methylation regulates the normal gene expression and protein function together with other epigenetic mechanisms [10]. It is also involved in mRNA processing and facilitates the chromatin organisation within cells. Aberrant DNA methylation patterns are usually developed during early carcinogenic formations and are associated with epigenetic silencing of genes.

In the complete dataset, information was collected on 153 patients. Of these patients, a subset of 95 contained information on copy number variation, gene expression and methylation.

2.2 Selecting genes

For each of the data type, a Wilcoxon Rank-Sum test was applied to select the 300 most significant genes by comparing the data from tumour tissue and normal control tissue. This allowed for identification of genes that are significantly altered in bladder cancer.

Specially, for gene expression, there are excess amount of zeros in the data that might be caused by the noise. Therefore, we removed all the genes that have more than 30 zeros (out of 95 patients). On the Synapse website [11], from which we obtained the data, the authors also provide a list of significant genes according to mutation data. Using the top 300 significant genes from GE, CNV, METH and Mutation (MUT) respectively, we looked at the overlaps, to see if any genes were significant in multiple data types. The results are given in Table 1 and are discussed in §3.6. We see that there is only a small overlap between any of the data types. Therefore, taking 300 significant genes in each data type would be a reasonably representative set given the computational time expenses.

2.3 Bayesian Hierarchical Clustering

One of the aims of this project is to stratify the patients according to multi-omic data, i.e. to group patients that have similar gene alterations, and look at if this is correlated with their survival rate or other clinical parameters. The algorithm we used to perform this is called Bayesian Hierarchical Clustering (BHC) [12, 13]. Hierarchical clustering is a bottom-up agglomerative algorithm that, starting from a set of n data points that are considered to be grouped in n clusters, gives as output a *dendrogram*, i.e. a binary tree constructed in $n-1$ steps by merging together at each time step the two clusters that are closest in the features space. In many ordinary clustering algorithms (such as hierarchical clustering and k-means clustering), it is necessary for the user to specify certain parameters such as the metric used to determine the distance between points and the number of clusters the dendrogram should be split into. This can lead to bias in the underlying distribution for the data. BHC does not make use of these parameters. Instead it considers the probability that two points (or two clusters) are to be merged. This probability is updated in a Bayesian fashion during the construction of the dendrogram. BHC has thus the advantage that it does not require the number of clusters, or the distance at which to cut the dendrogram for clusters, and not even a notion of distance on the feature space, but instead infers the number of clusters from the data.

The generative model for the BHC algorithm is a Dirichlet process mixture model and the algorithm performs an approximate inference of the model by computing the weights of exponentially many partitions with the constraint that they are consistent with the binary tree structure. Practically, it computes the probability that the data contained in the two clusters that it is trying to merge were in fact generated from the same mixture component. This is done for every pair of clusters and then the merging with the highest probability is performed.

To see this, let $\mathcal{D}$ be the set of all data points, and $\mathcal{D}_i \subset \mathcal{D}$ the set of data points at the leaves of tree T_i. Then, in considering the merging of the trees T_i and T_j in a new tree T_k, with associated dataset $\mathcal{D}_i \cup \mathcal{D}_j = \mathcal{D}_k$, the algorithm will consider two hypotheses: hypothesis $\mathcal{H}_1^k$ is that all the elements of $\mathcal{D}_k$ were actually generated from the same probabilistic model $p(\mathbf{x}|\theta)$ with unknown parameters θ. If $p(\theta|\beta)$ is the prior over the parameters of the model, the probability of the clustering $\mathcal{D}_k$ under hypothesis $\mathcal{H}_1^k$ is given by

$$p(\mathcal{D}_k|\mathcal{H}_1^k) = \int p(\mathcal{D}_k|\theta)p(\theta|\beta)\, d\theta. \tag{1}$$

The alternative hypothesis $\mathcal{H}_2^k$ is that data in $\mathcal{D}_i$ and $\mathcal{D}_j$ should not be merged because it is unlikely that they were generated from a single probabilistic model. By calling $\pi_k \overset{\text{def}}{=} p(\mathcal{H}_1^k)$ the prior that all points in $\mathcal{D}_k$ belong to one cluster, the marginal probability of the data in tree T_k is given by:

$$\begin{aligned} p(\mathcal{D}_k|T_k) = \pi_k p(\mathcal{D}_k|\mathcal{H}_1^k) + \\ + (1-\pi_k)p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j), \end{aligned} \tag{2}$$

which is defined recursively. The probability of data $\mathcal{D}_k$ under hypothesis $\mathcal{H}_2^k$ is then:

$$p(\mathcal{D}_k|\mathcal{H}_2^k) = p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j). \tag{3}$$

Then using Bayes rule it is possible to find the posterior probability for the hypothesis of merging the two clusters,

$$p(\mathcal{H}_1^k|\mathcal{D}_k) = \frac{\pi_k p(\mathcal{D}_k|\mathcal{H}_1^k)}{\pi_k p(\mathcal{D}_k|\mathcal{H}_1^k) + (1-\pi_k)p(\mathcal{D}_k|\mathcal{H}_2^k)} \tag{4}$$

which is used to decide which two trees to merge and which merges in the final hierarchy are justified. If the largest $p(\mathcal{H}_1^k|\mathcal{D}_k) > 0.5$, patients or genes will be clustered into the same cluster. Otherwise, merging still proceeds with the largest $p(\mathcal{H}_1^k|\mathcal{D}_k)$ value but puts the targets into different clusters.

2.4 Heatmaps

We plotted heatmaps of the 300 most significant genes for each data type by using R packages *BHC* and *gplots* [13, 14]. They are shown in Figures 1-3. BHC was applied both to patients (columns) and genes (rows), and displayed in the heatmaps. Using a heatmap is a very efficient way to display correlation between group of features and groups of data points. It also helps to achieve the aim of patient stratification.

2.5 Survival Analysis

We used Kaplan-Meier curves to visualize the survival relationships between the different clusters, and a log-rank test to compare the survival distributions for two or more clusters. With this approach, we aimed to identify clusters of patients that have significantly different survival outcome prognosis, and to determine in the analysis stage the factors that have driven these differences.

We performed survival analysis on the three data types using the *survival* package in R [15]. We fitted Cox proportional hazards models to investigate the relationships between the most relevant clinical variables such as tumour stage: extent of local invasion (T), spread to local lymph nodes (N), metastatic spread (M), gender, and age at diagnosis, and the survival of the different clusters of patients. The model assumes that the hazard of death is proportional to the exponential of a linear predictor formed of the covariates, and it deals appropriately with right-censored data where patients leave before the end of the medical study.

2.6 Gene Enrichment

Gene set enrichment analysis was applied to the clusters of genes from each data type using the databases Gene Ontology (Biological Process, Molecular Function and Cellular Component)

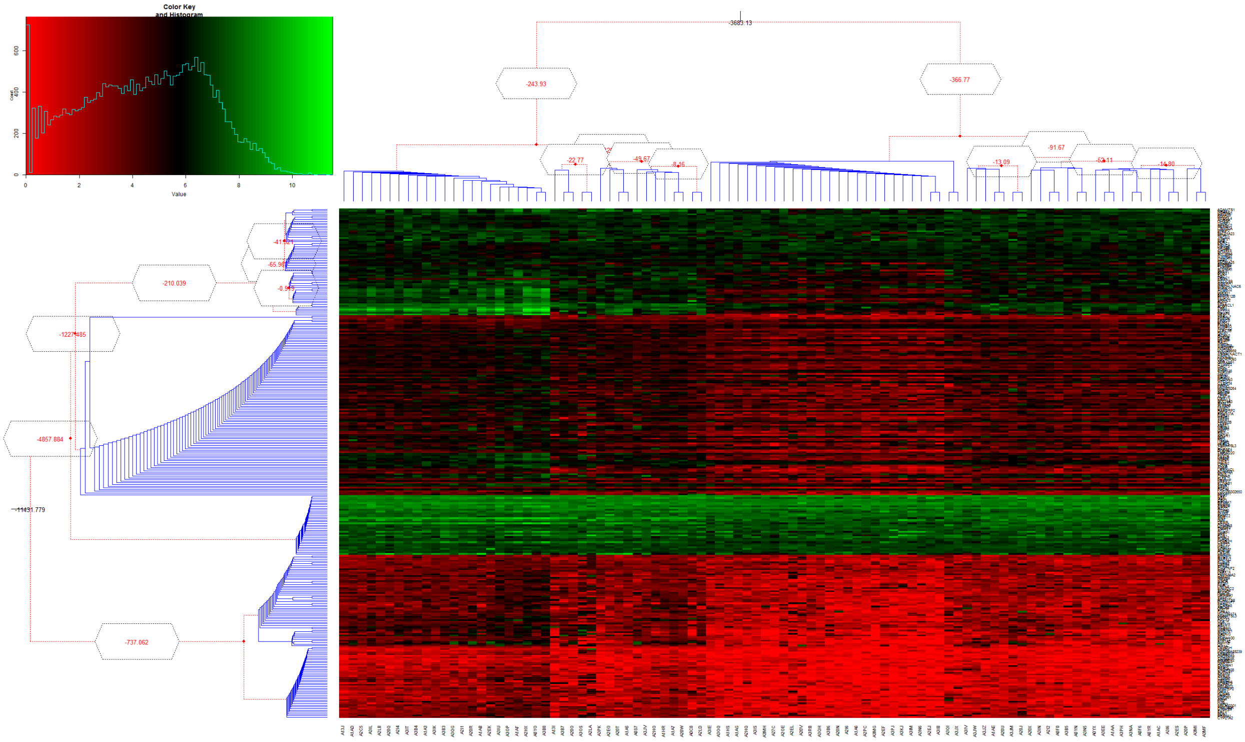

Figure 1: Heatmap for the GE dataset. Genes are shown in rows and patients are shown in columns. Gene cluster 3 has a significantly higher expression level compared with the other clusters, and there are mixed levels of gene expression in the top gene clusters (clusters 5-9) across different patients clusters.

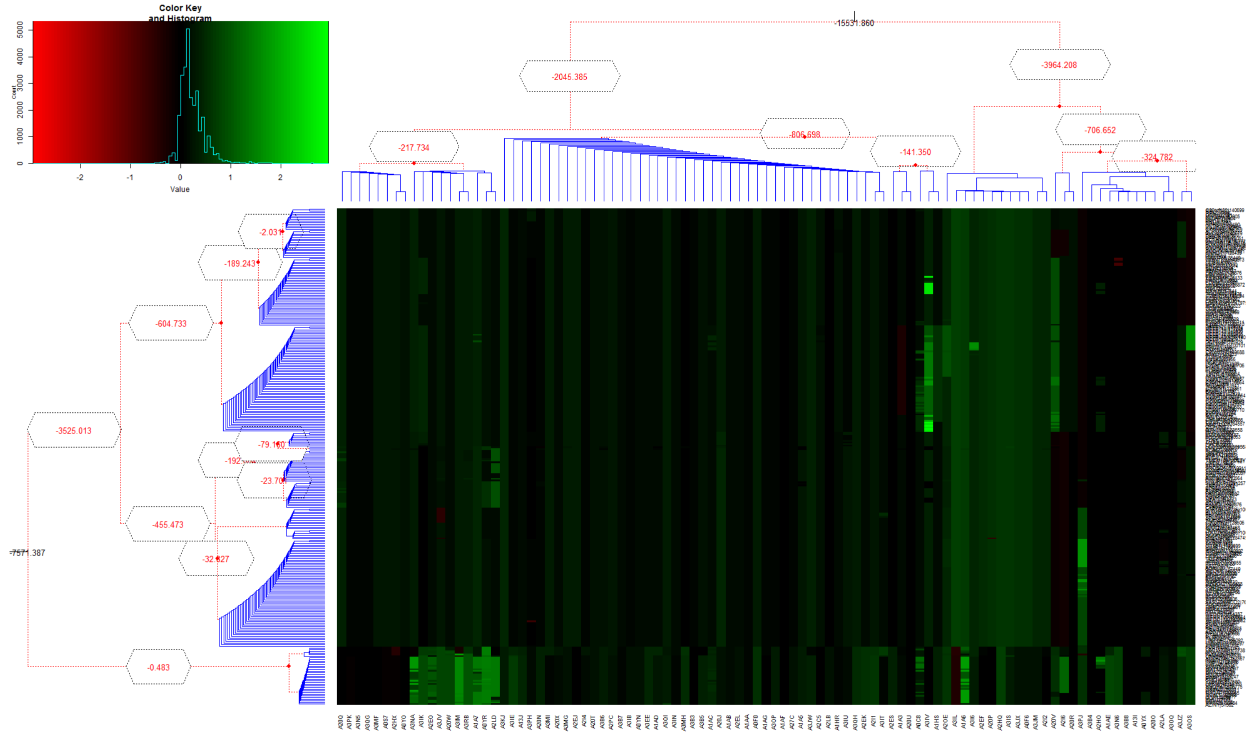

Figure 2: Heatmap for the CNV dataset. Genes are shown in rows and patients are shown in columns. Gene clusters 1 and 2 at the bottom have a concentrated high level variation for patient cluster 2.

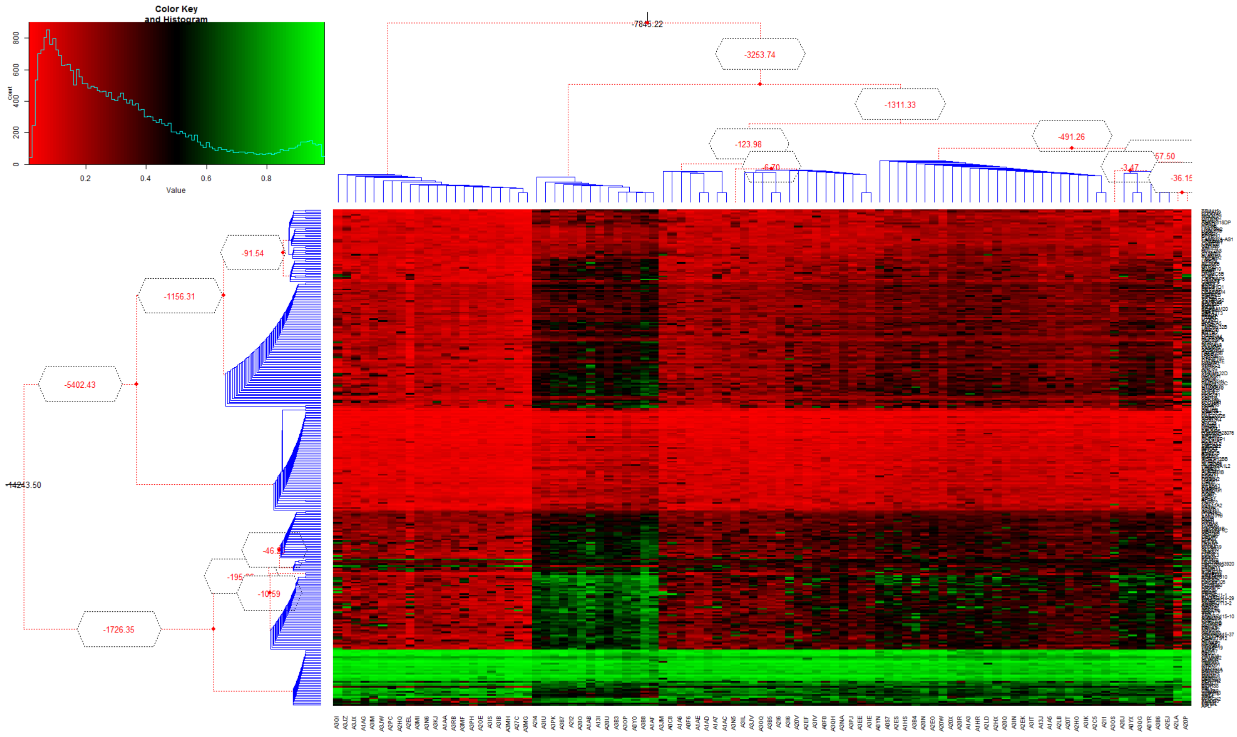

Figure 3: Heatmap for the Methylation dataset. The patients are on the horizontal axis and they have been clustered into six clusters, with a few outliers. The genes are on the vertical axis. Patients in cluster 2 appear to have significantly hypermethylated genes. There is one gene cluster that is consistently hypermethylated (along the bottom) and another that is unmethylated across all groups of patients.

and KEGG pathways. The three categories of Gene Ontology provide information on the biological objective to which certain genes contribute, the biochemical activity of these genes and the place in the cell where they are active. The KEGG pathways are widely used to highlight pathways that are over-represented in the specific gene cluster compared to the set of all genes.

We performed the analysis using the R package *HTSanalyzeR* [16]. It performs hypergeometric test for overlap between a gene cluster and the gene set defined in Gene Ontology, and gene set enrichment analysis for concordant trends in the cluster. The results are then mapped to a network and enriched subnetworks are identified.

In an enrichment map, the node marker size signifies the number of genes in this category/pathway and the thickness of the edges denotes the Jaccard similarity coeffiecient between two categories.

A p-value cut-off of 0.05 was used for the hypergeometric test. Enrichment maps were then plotted, using the 20 most significant GO and KEGG terms. p-values were adjusted using the Benjamini-Hochberg correction [17]. The pictures showing the results can be found in §S2 of the supplementary material.

2.7 Multiple dataset Integration

We fused the CNV and GE datasets using Multiple dataset Integration (MDI) [18, 19]. MDI is an unsupervised algorithm that allows for the identification of genes which cluster together in multiple datasets by giving an interpretation of the degree of similarity between data types. Crucially it uses a Gibbs Sampler to sample parameters ϕ_{kl} which measure the strength of association between data types k and l. In this way, MDI may be thought of as a correlated clustering method.

The probability density for each data type is modelled using an N-component Dirichlet-multinomial mixture model

$$p(x) = \sum_{c=1}^{N} \pi_c f(x \mid \theta_c).$$

The π_c's are mixture proportions, f is a parametric density and θ_c denotes the vector of parameters associated with the c^{th} component. We may choose f to appropriately model the kind of data we have. As our data are continuous, a Gaussian distribution is suitable. Given observations $x_1, \ldots, x_n$, we perform Bayesian inference to find posteriors for the parameters of this model. For details on the priors used, see [18]. It is useful to note, however, that $\pi_1, \ldots, \pi_n \sim Dir(\frac{\alpha}{N}, \ldots, \frac{\alpha}{N})$ where α is a parameter which may also be inferred.

To combine K data types, we take a mixture model for each, with each data type k allowed different concentration parameter α_k. The MDI model is as follows:

$$p(c_{i1}, c_{i2}, \ldots c_{iK} \mid \phi) \propto \prod_{k=1}^{K} \pi_{c_{ik}k} \prod_{k=1}^{K-1} \prod_{l=k+1}^{K} (1 + \phi_{kl}\mathbb{I}(c_{ik} = c_{il})), \quad (5)$$

where c_{ik} is the cluster of observation i in data type k and ϕ_{kl} is the dependency parameter showing the strength of association between data types k and l. If $\phi_{kl} = 0$ for every k and l, it would be equivalent to looking at the K mixture models separately i.e. having complete independence between the data types.

Parameter posterior distributions are obtained via a Gibbs Sampler and full details of the priors used can be found in the supplementary material of [18].

Whilst it is important to consider data types individually, MDI will provide more insight into the underlying structure of the sample and therefore into the true clustering structure of bladder cancer.

3 Results

We introduce the following notation: PC:= patient cluster, GC:= gene cluster. In each of the following subsections, we took one data type and clustered patients and genes according to that data. Kaplan-Meier curves are plotted for the patient clusters that have size bigger than 5. We looked at patient clusters in pairs and generated an unadjusted p-value for the null hypothesis that both clusters were drawn from the same distribution. A summary of the performed hypothesis tests is presented in Table 2.

3.1 Gene Expression

PCs 5 (containing 6 patients) and 7 (containing 26 patients) had significantly different survival outcomes with a p-value of 0.0159 and PCs 7 and 10 (containing 6 patients) had significantly different survival outcomes with a p-value of 0.00224. This last p-value is particularly small and of interest. In both cases, patients in PC7 had a better prognosis than those in clusters 5 and 10. Referring to the Gene Expression heatmap (Figure 1) we see that patients in PC7 stand out as having the lowest levels of gene expression across GCs 1, 2, 4 and 5 compared to the other PCs, and that this difference is most notable in GC5.

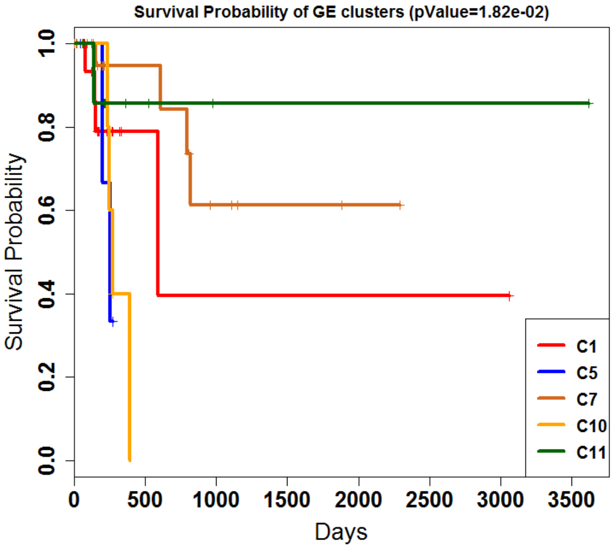

Figure 4: Kaplan-Meier plot of survival for the 5 largest patient clusters for the GE dataset

The enrichment analyses of GCs 1, 2, 4 and 5 revealed several pathways that are overrepresented in bladder cancer and have effect on the disease outcome. For example, changes in genes from GCs 1, 2 and 4 involved in focal adhesion, which is responsible for the cell attachment to the extracellular matrix, influence the cancer metastasis, growth and invasion [20]. Endocytosis, which regulates the accessibility of various receptors and ligands and is also the main regulator of the cell fate determination, appears to be dysregulated in GCs 1 and 4. This has a negative effect on the activation of major pathways such as the NOTCH pathway [21]. A more detailed analysis of the enriched pathways in GCs 1, 2 and 4 is provided in the §3.5.

Since the difference in the survival outcome of the PCs is the most notable in GC5, we looked at the functions of the genes in the cluster. Two of the genes, MYH11 and SPARCL1, have been identified as biomarkers for bladder cancer and are used in diagnostic method, which has had

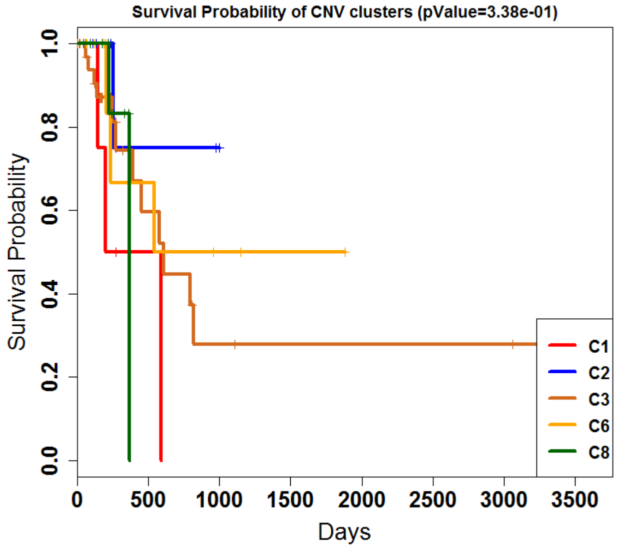

Figure 5: Kaplan-Meier plot of survival for the 5 largest patient clusters for the CNV dataset

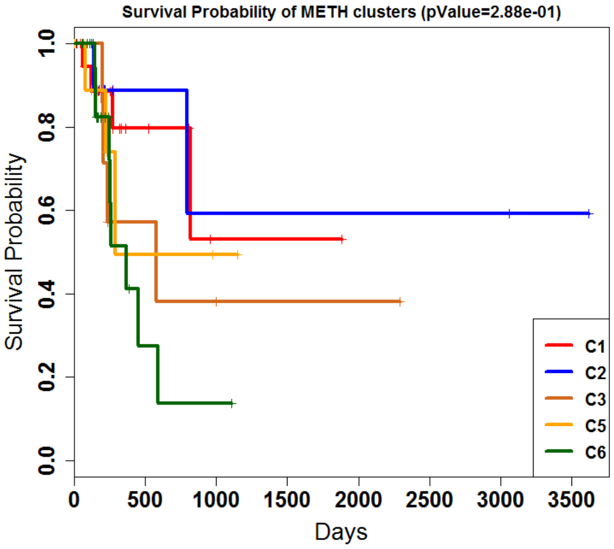

Figure 6: Kaplan-Meier plot of survival for the 5 largest patient clusters for the METH dataset

100% accuracy in clinical trials [22].

In addition, the overexpression of AQP1, which facilitates the transport of water across cell membranes and is involved in the control of the cell cycle, has been linked to tumorigenesis [23]. It is over-expressed in multiple tumours, such as breast cancer, renal cell carcinoma and adenoid cystic carcinoma, and Yoshida *et al.* [24] have shown in their study that patients with positive expression of AQP1 have much worse outcome prognosis than those with negative expression. Moreover, it has been suggested that AQP1 can be used as a potential biomarker for early diagnosis of tumours [25] and there is ongoing research into the possible applications of AQP inhibitor to prevent the development and progression of tumours [26].

3.2 Copy Number Variation

Comparing PCs 1 (containing 6 patients) and 2 (containing 10 patients) in terms of survival outcomes gave the smallest p-value (0.09), with patients in PC1 having a worse prognosis. We examine the different properties of PCs 1 and 2 despite the large p-value, but for future research it would be worth looking at a larger sample size to see if there is any significant difference between the two clusters. Referring to the Copy Number Variation heatmap (Figure 2), we see that in PC2 patients, in GC1 and GC2 there are more copy number gains than in PC1, where there are no gains amongst these genes, only deletions. Therefore, there may be a link between copy number loss of these genes in GC1 and 2 and a poorer survival outcome.

We performed enrichment analysis on the GCs 1 and 2 to investigate the different survival outcomes on molecular level. The MAPK signalling pathway, which is involved in cell differentiation and proliferation, and apoptosis[1] [27], is dysregulated in GC1. Moreover, major metabolic pathways are involved in serious bladder cancer pathophysiology in GC1, and this has a detrimental effect on the homeostasis. The enrichment map analysis suggests that the p53 signalling pathway, which has many links to cancer, has been altered and can result in several responses, including cell cycle arrest, the inhibition of angiogenesis[2] and DNA repair [28].

3.3 Methylation

PCs 1 (containing 21 patients) and 6 (containing 24 patients) have significantly different survival outcome, with a p-value of 0.048. The patients of PC6 have significantly worse prognosis and referring to the Methylation heatmap (Figure 3), we see that they have consistently higher DNA methylation levels compared with PC1, and cancer cells tend to be more methylated in general. Curiously, PC2 (containing 11 patients) stands out as a distinct band of hypermethylation, but the patients in this cluster have the best survival outcome prognosis. There are a number of pos-

[1] **apoptosis**: programmed cell death

[2] **angiogenesis**: the formation of new blood vessels, often as a result of a tumour

sible explanations to explore.

Firstly, as Witte *et al.* [29] suggest in their review low methylation groups of patients are associated with more frequent mutations in the TP53 gene, which is one of the most commonly altered genes in cancer, further work is needed to establish the characteristics of patients who would fall into PC6.

Secondly, computation of the mean age of each patient cluster reveals that the patients in PC2 are older than in the other two clusters, with a mean age of 70.91, as compared with a mean age of 66.33 for PC1 and of 67.63 for PC6. Since bladder cancer is an age-driven disease, the correlation between age and DNA methylation levels should also be taken into consideration.

3.4 Where has TNM failed patients?

For the past 30 years, treatment of bladder cancer has largely been determined by TNM staging. It gives a grade for the extent of the primary tumour (T), how much the cancer has spread to lymph nodes (N) and whether it has metastasized (M). The cancer is then treated according to where it lies on the TNM scale. It is hoped that analysis of multi-omic data can improve on TNM staging by providing more subtypes of bladder cancer or by identifying areas where TNM is incorrectly determining appropriate treatment for patients. We ranked patients based on TNM rating and identified those whose vital status was unexpected given their TNM rating. It is important to note that for some patients, we did not have follow-up data over a long duration. In future work, it will be useful to have this information to do more accurate survival analysis. Four patients were identified who had not survived but had tumours at stage T2. Three of these patients belonged to PC3 for copy number variation. PC3 is large and more work needs to be done to establish the characteristics of this cluster.

At the other end of the spectrum, seven patients were identified with tumours at stage T4 who had not died. Three of these patients belonged to PC1 for gene expression and three patients belonged to PC8 for copy number variation. This last finding is particularly interesting as this cluster is fairly small (containing ten patients) and nearly a third of patients in this cluster have done better than expected. Furthermore, patients belonging to PC8 for copy number variation showed lower levels of copy number variation, across all genes except for those in GC1. This could imply that checking patient CNV levels could be a useful addition to TNM classification. If patients are found to have a high tumour stage but have low levels of copy number variation, they may have a better outcome than other patients with T4 tumours. This information could be used to make better informed choices on appropriate treatments. Further exploration of this theory needs to be done looking at larger samples of patients and other survival data such as tumour recurrence and treatment given.

3.5 Gene Enrichment

A better understanding of the molecular pathophysiology of bladder cancer and the pathways involved in muscle invasion and metastasis could lead to a targeted therapy. By addressing specific dysregulated pathways connected to a more progressive disease, we will be better able to choose appropriate treatments based on the patient genotype. We performed gene enrichment analysis on the GCs that may contribute to the different patient survival to determine the frequently aberrant pathways.

The GCs 1, 2 and 4 from GE (Figures S2-S4 in the supplementary material) appear to be enriched for pathways, known to be related to bladder cancer, for example, Wnt and mTOR signalling pathways. The secreted signalling factors of the Wnt protein family have been found to regulate many cellular processes, including cell fate decisions and cell proliferation, and so aberrant Wnt signalling is associated with tumorigenesis [30]. The mTOR signalling pathway is invloved in many major cellular processes, including survival, metabolism and autophagy[3] [31]. The defective mTOR signalling pathway in GC4 affects the PI3K pathway, which is an important regulator of the cell cycle. The chemokine signalling pathway, which is altered in GCs 2 and 4, is known to regulate growth, survival and migration [32]. Thus it has a vital role in metastasis. In addition, the tight junctions govern the permeability of endothelial and epithelial cells, and the dismantling of their structure can lead to the uncontrolled growth and invasion of cancer cells [33]. Other enriched pathways in these clusters

[3] **autophagy**: controlled digestion of damaged organelles within a cell

	Clusters	*Unadjusted p-value*	*Cluster size*	*Median Survival Time (days)*
GE	PC5	0.0159	6	200
	PC7*		26	>2000
GE	PC7*	0.00224	26	>2000
	PC10		6	272
CNV	PC1	0.09	6	396
	PC2*		10	>1000
METH	PC1*	0.048	21	>1500
	PC6		24	370
MDI	PC1*	0.062	34	393
	PC6		47	370

*Table 2: Summary of the pairwise hypothesis tests for each of the data types and MDI. The PC with better prognosis is denoted with *.*

include: protein digestion and absorption, calcium signalling pathway, vascular smooth muscle contraction, cell adhesion molecules and purine metabolism.

The analyses of GCs 1 (Figure S5 in the supplementary material) and 2 from CNV reveal that the clusters are enriched for pathways, known to be related to bladder cancer, for example: p53 signalling pathway, neurotrophin signalling pathway and purine metabolism. The dysregulation of the p53 signalling pathway has been implicated in cancer migration, invasion, survival and growth, and it may lead to cell cycle arrest and apoptosis [28]. The defective neurotrophin signalling pathway affects the PI3K signalling pathway too. This has an adverse effect on the cell survival, growth, motility and angiogenesis [34]. Purine metabolism, which is required for the production and recycling of adenine and guanine, is also involved in serious bladder cancer pathophysiology, and this has a detrimental effect on the DNA repair process [35]. Other enriched pathways include: cell cycle, pyrimidine metabolism, mRNA surveillance pathway and RNA transport.

We integrated the genes from GCs 2, 3, 4 and 5 (Figure S6 in the supplementary material) from METH into one cluster to perform enrichment analysis and to investigate further the differenent survival outcomes of PCs 1 and 6. This was mainly done because GCs 3 and 4 have insufficient number of genes to obtain an enrichment map on their own. The analysis implies that processes associated with cell proliferation and cell survival are defective. For example, the MAPK and NOTCH signalling pathways are enriched. The deviation from the strict control of the MAPK signalling pathway impairs the cell growth, proliferation and differentiation processes [27], whereas the defective NOTCH signalling pathway, which regulates the cell fate determination during the development stage [36], affects the Wnt, hypoxia and TGF-beta pathways [37] too.

The analysis also shows that some adhesion processes are enriched. For example, the cell adhesion molecules, which contribute to a variety of functions including cell growth, differentiation, motility and inflammation, participate in tumour invasiveness and metastasis and thus play a pivotal role in the development and progression of the cancer [38]. In addition, the activation of the JAK/STAT signalling pathway is known to have influence on proliferation, migration, and apoptosis in cancer [39]. Other processes that are enriched in the integrated gene cluster include carbohydrate digestion and absorption, calcium signalling pathway and protein digestion and absorption.

3.6 Overlapping genes

We also looked at the functions of the genes that appear in two or more of the data types. Most of the genes are known to be related to different types of cancer: for example, FAM84B is associated with breast, colorectal, ovarian and prostate cancer and adenocarcinoma [40], whereas MYBL2 and SNRPB are associated with breast and ovarian cancer [41, 42]. MYBL2 is important for the cell cycle progression and has be identified as a potential tumour suppres-

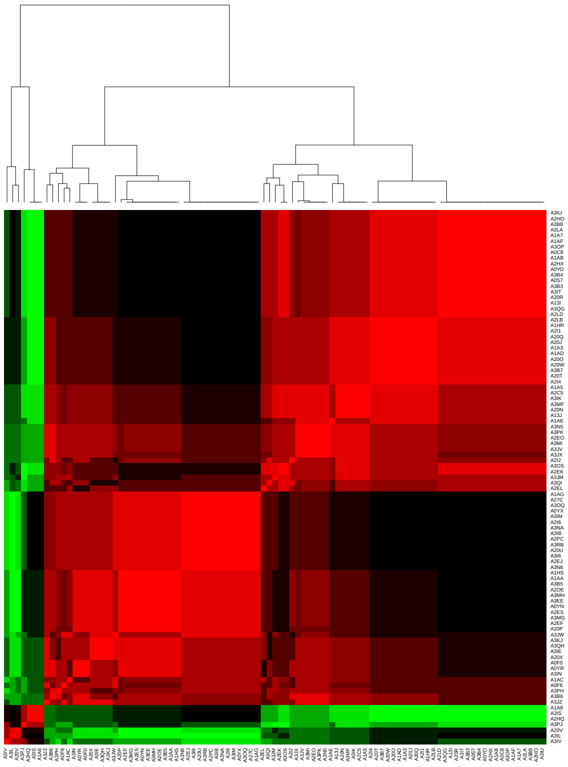

Figure 7: MDI averaged clustering over patients

sor for acute myeloid leukemia [43]. In addition, CREB3L3, which is involved in sequence-specific DNA binding and transcription factor activity, is frequently mutated in patients with hepatocellular carcinoma, CAMKK2, which appears both in MUT and METH, and is associated with prostate cancer [44], influences various processes, such as gene transcription, cell survival and apoptosis. TNXB, which mediates interactions between cells and extracellular matrix and plays an important role in the growth of epithelial tumours, has been found to be overexpressed in bladder cancer [45].

3.7 Multiple dataset Integration

Preliminary runs of the MDI algorithm on each single dataset showed that in the methylation data, patients are put in a single cluster. The reason for this is as yet unclear to us. Since this affects the algorithm, we decided to run MDI using only the GE and CNV datasets. The algorithm was run 6 times, each time generating a MCMC with $n = 20000$ steps and using the last 100 time steps to perform the following analysis. From the output of each run, a posterior similarity matrix between the patients was produced, and an average of these matrices was calculated.

A standard hierarchical clustering was performed on this matrix, and the result is shown in Figure 7. From the picture and from the dendrogram it is possible to see that the patients are clearly divided into 4 clusters. Two of these clusters are very big, and made of $\simeq 40$ patients each, while the two remaining clusters are made of only $\simeq 5$ patients.

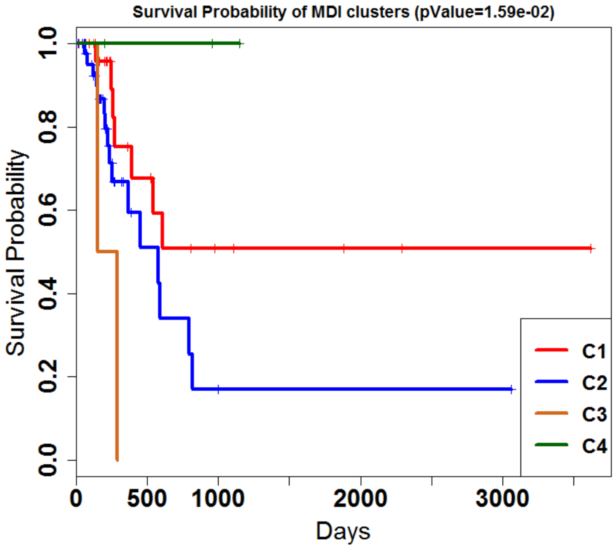

Figure 8: Kaplan-Meier plot of survival for the MDI patient clusters

The comparison of PC1 (containing 34 patients) and PC2 (containing 47 patients) in terms of survival outcome gave a p-value of 0.062, with the patients in PC1 having a better prognosis (Figure 8). Although the difference between the PCs is not statistically significant, it is definitely an improvement with respect to the single data type ones. Considering the TNM staging, we observed that in most of the patients in PC2, the tumour had already spread to numerous distant and regional lymph nodes. On molecular level, the patients in PC2 have higher methylation levels (Figure 9) and more overexpressed genes (Figure S8) than the patients in PC1.

To check if this new clustering of the patients is meaningful, the dendrogram resulting from Figure 7 was imposed on the three different data types. For CNV the result is not much different from Figure 2, although some new structure appears for the 2 big clusters.

Imposing the new dendrogram on METH and GE data results in a significant loss of the previous structure, as shown in Figure 9 for METH and Figure S8 for GE, but a different kind of structure seems to emerge.

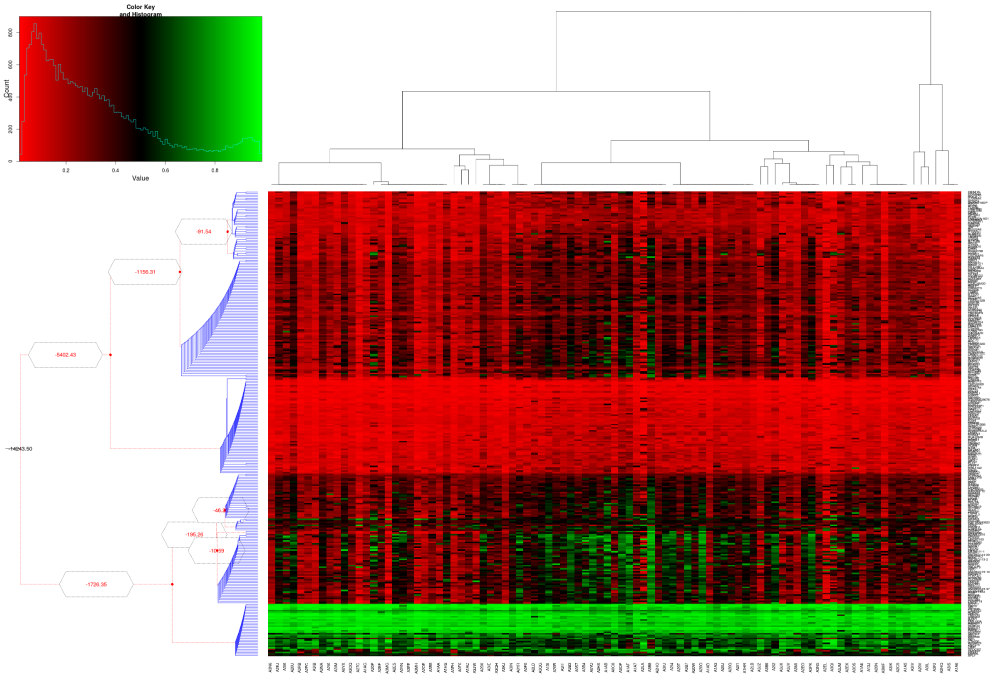

Figure 9: Heatmap for the METH dataset with patients dendrogram obtained from MDI

4 Conclusions

We have presented analysis of multi-omic data from TCGA and identified potential subphenotypes of bladder cancer based on survival outcome. These may be more patient specific than the current TNM system used for diagnosis. We looked at the 300 most significantly altered genes for each data type, clustered using BHC and then performed survival analysis and gene enrichment. After looking at the data types individually we also examined how they could be fused using MDI. Our main findings can be summarised as follows:

- The analysis of the gene expression data revealed a link between gene expression levels and patient survival outcome: lower levels of certain clusters of genes are associated with better outcome prognosis. As patients with overexpressed AQP1 gene have considerably worse disease outcome, it will be worth researching the possible application of AQP1 as a bladder cancer biomarker.
- Using methylation data, we identified a group of patients with consistently hypermethylated genes but also with the best survival outcome. The patients in this cluster had a higher mean age than the rest of the cohort, which implies that the patient's age is an important factor in predicting disease outcome and should be taken into account when working with DNA methylation data.
- We also found patients whose survival outcome was not expected given their TNM staging. Our analysis of the CNV dataset suggests that patient CNV levels may provide some explanation about why these cases occurred and that they could be a useful addition to the TNM classification.
- The fusion of CNV and GE using MDI improved the disease outcome prediction. It clustered the patients into fewer groups than BHC did with the individual data types, and a distctinctive genetic profile emerges even in the data that was not used in the dataset integration.

As with all work of this nature and particularly in our case, our analysis would have benefitted from a larger sample size, enabling us to have larger clusters. It will also be an important next step to use data containing other clinical information such as tumour recurrence and progression. Data fusion is still in its early development and as we have seen, it is still difficult to understand the underlying clustering structure using data types which do not show strong levels of agreement. It will be interesting to see how this field develops and how we may combine different data types in the future.

Acknowledgments

This research was funded by EPSRC and MRC and carried in collaboration with University Hospital Birmingham. Our gratitude goes to our supervisors Dr Rich Savage (University of Warwick, Systems Biology) and Professor Jean-Baptiste Cazier (University of Birmingham, School of Cancer Sciences) for their guidance and motivation. Thanks also goes to Professor Nick James (University of Warwick, Warwick Medical School) for his insight into some of the biological aspects of the project.

References

[1] E. C. Hayden. The '$ 1, 000 genome'. *Nature*, 507:294–295, 2014.

[2] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, 127:2893–2917, 2010.

[3] M. A. Knowles and C. D. Hurst. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nature Reviews Cancer*, 15:25–41, 2015.

[4] The Cancer Genome Atlas. `http://cancergenome.nih.gov/`. Accessed: 2015-03-25.

[5] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, and B. A. Ozenberger. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013.

[6] F. Cappuzzo, F. R. Hirsch, E. Rossi, S. Bartolini, G. L. Ceresoli, L. Bemis, J. Haney, S. Witta, K. Danenberg, and I. Domenichini. Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non–small-cell lung cancer. *Journal of the National Cancer Institute*, 97:643–655, 2005.

[7] M. R. Atkinson, M. A. Savageau, J. T. Myers, and A. J. Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in escherichia coli. *Cell*, 113: 597–607, 2003.

[8] G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, and R. Misra. Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39:1256–1260, 2007.

[9] S. L. Pereira, A. S. Rodrigues, M. I. Sousa, M. Correia, T. Perestrelo, and J. Ramalho-Santos. From gametogenesis and stem cells to cancer: common metabolic themes. *Human reproduction update*, 20:924–943, 2014.

[10] S. M. Gasser and E. Li. *Epignenetics and Disease*. Springer Basel, 2011.

[11] Sage Synapse: Contribute to the Cure. `https://www.synapse.org/#!Synapse:syn1461149/wiki/`. Accessed: 2015-03-25.

[12] K. A. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, 2005.

[13] R. S. Savage, K. Heller, Y. Xu, Z. Ghahramani, W. M. Truman, G. Murray, K. Denby, and D. L. Wild. R/BHC: fast bayesian hierarchical clustering for microarray data. *BMC Bioinformatics*, 10, 2009.

[14] G. Warnes and original R port by Thomas Lumley. Package 'gplots'. 2009.

[15] T. Therneau and T. Lumley. survival: Survival analysis, including penalized likelihood, 2011. *R package version*, pages 2–36, 2011.

[16] X. Wang, C. Terfve, JC. Rose, and F. Markowetz. Htsanalyzer: an r/bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics*, 27:879–880, 2011.

[17] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.

[18] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28:3290–3297, 2012.

[19] R. S. Savage, Z. Ghahramani, J. E. Griffin, P. Kirk, and D. L Wild. Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. *arXiv:1304.3577*, 2013.

[20] G. W. McLean, N. O. Carragher, E. Avizienyte, J. Evans, V. G. Brunton, and M. C. Frame. The role of focal-adhesion kinase in cancer- a new therapeutic opportunity. *Nature Reviews Cancer*, 5:505–515, 2005.

[21] P. P. Di Fiore. Endocytosis, signaling and cancer, much more than meets the eye. *Molecular Oncology*, 3:273–279, 2009.

[22] L. Dong, A. J. Bard, W. G. Richards, M. D. Nitz, D. Theodorscu, R. Bueno, and G. J. Gordon. A gene expression ratio-based diagnostic test for bladder cancer. *Advances and Applications in Bioinformatics and Chemistry*, 2:17–22, 2009.

[23] M. O. Hoque, J.-C. Soria, J. Woo, T. Lee, J. Lee, B. Trink, and C. Moon. Aquaporin 1 is overexpressed in lung cancer and stimulates NIH-3T3 cell proliferation and anchorage-independent growth. *American Journal of Pathology*, 168:3973–3976, 2006.

[24] T. Yoshida, S. Hojo, and S. Sekine. Expression of aquaporin-1 is a poor prognostic factor for stage ii and iii colon cancer. *Molecular Clinical Oncology*, 1:953–958, 2013.

[25] J. J. Morrissey, J. Mobley, and R. S. Figenshau. Urine aquaporin 1 and perilipin 2 differentiate renal carcinomas from other imaged renal masses and bladder and prostate cancer. *Mayo Clinical Procedure*, 90:35–42, 2015.

[26] C. Stigliano, S. Aryal, and M. D. de Tullio. siRNA-chitosan complexes in poly (lactic-co-glycolic acid) nanoparticles for the silencing of aquaporin-1 in cancer cells. *Molecular Pharmacology*, 10:3186–3194, 2013.

[27] A. S. Dhillon, S. Hagan, O. Rath, and W. Kolch. MAP kinase signalling pathways in cancer. *Oncogene*, 26:3279–3290, 2007.

[28] A. J. Levine, J. Bargonetti, G. L. Bond, J. Hoh, and K. Onel. *The p53 tumour suppressor pathway and cancer.* Springer US, 2005.

[29] T. Witte, C. Plass, and C. Gerhauser. Pan-cancer patterns of DNA methylation. *Genome Medicine*, 66:138–140, 2014.

[30] L. R. Howe and A. M. C. Brown. Wnt signaling and breast cancer. *Cancer Biology and Therapy*, 3:36–41, 2004.

[31] M. Laplante and D. M. Sabatini. mTOR signaling in growth control and disease. *Cell*, 149:274–293, 2012.

[32] S. L. Hembruff and N. Cheng. Chemokine signaling in cancer: Implications on the tumor microenvironment and therapeutic targeting. *Cancer Therapy*, 7:254, 2009.

[33] T. A. Martin and W. G. Jiang. Loss of tight junction barrier function and its role in cancer metastasis. *BBA Biomembranes*, 1788: 872–891, 2009.

[34] R. M. Kostrzewa. *Handbook of Neurotoxicity.* Springer, New York, 2014.

[35] A. Ertel, A. Verghese, S. W. Byers, M. Ochs, and A. Tozeren. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Molecular Cancer*, 5, 2006.

[36] M. Baron. An overview of the notch signalling pathway. *Seminars in Cell and Developmental Biology*, 14:113–119, 2003.

[37] E. R Andersson, R. Sandberg, and U. Lendahl. Notch signalling: simplicity in design, versaility in function. *Development*, 138:3593–3612, 2011.

[38] T. Okegawa, R. C. Pong, Y. Li, and J. T. Hsieh. The role of cell adhesion molecule in cancer progression and its applications in cancer therapy. *Acta Biochimica Polonica*, 51:445–457, 2004.

[39] J. S. Rawlings, K. M. Rosler, and D. A. Harrison. The JAK/STAT signaling pathway. *Journal of Cell Science*, 117:1281–1283, 2004.

[40] M. Ghoussaini, H. Song, T. Koessler, K. E. Driver, K. A. Pooley, and S. J. Ramus. Multiple loci with different cancer specificities within the 8q24 gene desert. *Journal of National Cancer Institute*, 100:962–966, 2008.

[41] M. M. Tanner, S. Grenman, A. Koul, O. Johannsson, and P. Meltzer. Frequent amplification of chromosomal region 20q12-q13 in ovarian cancer. *Clinical Cancer Research*, 6: 1833–1839, 2000.

[42] X. Wang, V. Shane Pankratz, Z. Freder-

icksen, R. Tarrell, M. Karaus, and A. M. Dunning. Common variants associated with breast cancer in genome-wide association studies are modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers. *Human Molecular Genetics*, 6:2886–2897, 2000.

[43] S. Heinrichs, L. F. Conover, C. E. Bueso-Ramos, O. Kilpivaara, K. Stevenson, and D. Neuberg. MYBL2 is a sub-haploinsufficient tumor suppressor gene in myeloid malignancy. *eLIFE*, 2, 2013.

[44] L. Racioppi and A. R. Means. Calcium/calmodulin-dependent protein kinase kinase 2: Roles in signaling and pathophysiology. *The Journal of Biological Chemistry*, 287:31658–31665, 2012.

[45] H. Niu, H. Jiang, B. Cheng, L. Xinhui, L. Shao, Q. Dong, S. Liu, and X. Wang. Stromal proteome expression profile and muscle-invasive bladder cancer. *Cancer Cell International*, 12, 2012.