UMEÅ UNIVERSITY

Cancer subtype identification using cluster analysis on high-dimensional omics data

Linda Vidman

Department of Mathematics and Mathematical Statistics
Umeå 2020

Doctoral Thesis No. 70/20

ISBN: 978-91-7855-172-9
ISSN: 1653-0829

Electronic version available at: http://umu.diva-portal.org/
Printed by: TMG Tabergs AB
Umeå, Sweden 2020

If you tell the truth, you don´t have to remember anything.

–Mark Twain

Table of Contents

List of papers

The thesis is based on the following papers:

I. **Vidman L**, Källberg D, Rydén P. Cluster analysis on high dimensional RNA-seq data with applications to cancer research - An evaluation study. *PLOS ONE*. 2019;14(12):e0219102.

II. Källberg D, **Vidman L**, Rydén P. Comparison of methods for variable selection in clustering of high-dimensional RNA-sequencing data to identify cancer subtypes (manuscript).

III. Thysell E, **Vidman L**, Ylitalo EB, et al. Gene expression profiles define molecular subtypes of prostate cancer bone metastases with different outcomes and morphology traceable back to the primary tumor. *Mol Oncol*. 2019;13(8):1763-1777.

IV. Andersson-Evelönn E, **Vidman L**, Källberg D, Landfors M, Liu X, Ljungberg B, Hultdin M, Degerman S, Rydén P. Combining epigenetic and clinicopathological variables improves prognostic prediction in clear cell Renal Cell Carcinoma (manuscript).

Abstract

Identification and prediction of cancer subtypes are important parts in the development towards personalized medicine. By tailoring treatments, it is possible to decrease unnecessary suffering and reduce costs. Since the introduction of next generation sequencing techniques, the amount of data available for medical research has increased rapidly. The high dimensional omics data produced by various techniques requires statistical methods to transform data into information and knowledge.

All papers in this thesis are related to distinguishing of disease subtypes in patients with cancer using omics data. The high dimension and the complexity of sequencing data from tumor samples makes it necessary to pre—process the data. We carry out comparisons of feature selection methods and clustering methods used for identification of cancer subtypes. In addition, we evaluate the effect that certain characteristics of the data have on the ability to identify cancer subtypes. The results show that no method outperforms the others in all cases and the relative ranking of methods is very dependent on the data. We also show that the benefit of receiving a more homogeneous data by analyzing genders separately can outweigh the possible drawbacks caused by smaller sample sizes. One of the major challenges when dealing with omics data from tumor samples is that the patients are generally a very heterogeneous group. Factors that lead to heterogeneity include age, gender, ethnicity and stage of disease. How big the effect size is for each of these factors might affect the ability to identify the subgroups of interest.

In omics data, the feature space is often large and how many of the features that are informative for the factors of interest will also affect the complexity of the problem. We present a novel clustering approach that can identify different clusters in different subsets of the feature space, which is applied on methylation data to create new potential biomarkers. It is shown that by combining clinical data with methylation data for patients with clear cell renal carcinoma, it is possible to improve the currently used prediction model for disease progression.

Using unsupervised clustering techniques, we identify three molecular subtypes of prostate cancer bone metastases based on gene expression profiles. The robustness of the identified subtypes is confirmed by applying several clustering algorithms with very similar results.

Sammanfattning

Identifiering och prediktion av cancer undergrupper är viktiga delar i utvecklingen mot personalized medicin. Genom att skräddarsy behandling är det möjligt att reducera både onödigt lidande och kostnader. Sedan introduktionen av next generation sequencing tekniken så har mängden data som kan användas för medicinsk forskning ökat snabbt. Det högdimensionella data som produceras av olika tekniker kräver statistiska metoder för att omvandlas till information och kunskap.

Alla artiklar i den här avhandlingen är relaterade till särskiljning av sjukdomsundergrupper hos patienter med cancer genom användning av omikdata. Den höga dimensionen och komplexiteten hos sekvensieringsdata från tumörprover gör det nödvändigt att bearbeta data innan analys. Vi jämför olika variabelselektion- och klustermetoder som används för att identifiera cancerundergrupper. Vi utvärderar även effekten vissa utmärkande drag hos data har på förmågan att identifiera cancerundergrupper. Resultatet visar att ingen metod utklassar de övriga metoderna i alla fallen och att den relativa rankingen av metoderna var väldigt beroende av data. Det visades också att det kunde vara fördelaktigt att analysera könen var för sig, eftersom fördelen med ett mer homogent data kan uppväga nackdelen med en mindre stickprovsstorlek. En av de stora utmaningarna med omikdata från tumörprover är att patientgruppen oftast är väldigt heterogen. Patienterna skiljer sig i allt från ålder och kön, till etnicitet och sjukdomsstadie. Hur stor effektstorleken är för dessa faktorer kan påverka förmågan att identifiera undergrupperna av intresse.

I omikdata är antalet variabler ofta stort, och hur många av dessa som innehåller information kopplat till faktorerna av intresse, påverkar också komplexiteten av problemet. Vi presenterar en ny klustermetod som kan identifiera olika kluster bland olika delar av variablerna och denna används på metyleringsdata för att skapa nya potentiella biomarkörer. Vi visar att det är möjligt att förbättra prediktionsmodellen för sjukdomsprogression hos patienter med njurcellscarinom genom att kombinera kliniskt data med metyleringsdata.

Vi identifierar tre undergrupper av benmetastaser från prostatacancer baserat på genuttrycksprofiler genom att använda oövervakade klustringstekniker. Vi visar att grupperna är robusta genom att applicera flera olika klustringstekniker som alla gav liknande resultat.

Acknowledgements

I would like to start by thanking my supervisor Patrik Rydén for his support during these years. His devotion for research has been truly inspiring. I would like to thank Anna Ivarsson for all the work she has put down to keep the department at order, her angry notes in the kitchen have never failed to make me smile. I would like to extend my thanks to the staff at the service office for cleaning up our mess and making sure the coffee machines are filled with beans. Even though I have not, after more than ten years at the university, learned to appreciate the bitter taste of coffee, it has not passed me by, the importance the beverage has on morale of many of my colleagues. Coffee-lovers or not, I would like to thank my colleagues at the department of Mathematics and Mathematic Statistics for all interesting conversations and for motivating me to finish my PhD.

I also feel sincere gratitude towards all my co-authors; without them, this had not been possible. To David, Jun and Therese for reading and helping me improve this thesis. To the members in Umeå fallskärmsklubb for believing I can fly and for filling my life with joy, even at the worst times.

Finally, I would like to give my warmest thanks to Mikael for listening to all my complaints and for cheering me up when feeling down.

Linda Vidman
Umeå, January 2020

1. Introduction

Disease – a condition deviating from the normal, with a negative effect on function or structure of an organism. Diseases of different kinds have plagued the inhabitants of the earth since the beginning of time. Archeological discoveries of prehistoric people show signs of disease in form of gross external features, and documentation of disease can be found as far back as in the 17th century BC. It was not until about the fourth century BC that scientists, influenced by the Greek physician Hippocrates, started to believe that disease was not a punishment from the Gods, but rather caused by earthly influences. Since then, researchers all over the world have made lasting contributions to the field of pathology [1].

The causes of disease are many. Pathogenic microbial agents such as viruses and bacteria can cause a variety of infectious diseases [2]. Diseases can also be caused by epigenetic changes or genetic defects and may or may not be hereditary [3,4]. One example of such a disease is cancer, which is the second leading cause of death worldwide [5]. The suffering and costs connected to cancer related diseases are huge, which makes cancer research a high priority target. Cancer is actually a general term for a group of diseases that involves cells that grow and divide in an uncontrolled manner. The connection to genetic damage was made over 100 years ago by Theodore Boveri who published a paper suggesting that cancer tumors originate from a single cell with chromosomal damage and that inheritance could play a role in the risk of cancer development [6]. It would however take until 2003 before the Human Genome Project completed the task of determining the sequence of nucleotide base pairs that constitutes the human DNA [7]. It took additional three years before the first report of cancer genome sequencing appeared [8]. The first methods used for sequencing were both slow and costly. The introduction of next generation sequencing, which enabled researchers to sequence data at much higher speed and at lower costs than before, opened a new era in genomic and medical research. Researchers were provided with opportunities of investigating the role played by genomic variants in health and disease of humans. As the amount of data generated by sequencing methods continue to increase exponentially, the technical and ethical challenges arise with it. The huge amount of data also sets higher demands on researchers to transform data into information and knowledge, which requires development of advanced statistical methods to handle the high dimension and complexity of the data.

When studying diseases using genetic or epigenetic data, the process typically involves at least three steps. A pre-processing procedure with the aim to remove technical noise is usually the first step. After that, a feature selection step is often necessary to remove redundant and uninformative features and therefore reduce the dimension. Thereafter statistical methods can be used to test hypotheses,

discover novel subtypes and predict diagnosis etc. Each of these steps requires the researcher to make choices of which method to use. It is crucial to evaluate the effect of these choices and to investigate how different characteristics of the data influence the performance of the methods.

Finding relevant underlying subgroups in high-dimensional genetic and epigenetic data where the patients often have diverse backgrounds and the features are affected by many factors of which only a few are known, is an arduous task. Detection of novel disease subtypes requires unsupervised methods and different clustering algorithms have been frequently used for this purpose [9,10].

Classification methods can be used for assessing risk of disease progression or survival time for patients, but this requires training data where the outcome is known. The classification result can then be used to determine e.g. follow-up or treatment strategies.

In this thesis, gene expression data from cancer patients are used to evaluate feature selection/extraction methods as well as clustering methods where the aim is to identify cancer subtypes. Properties of the data, such as imbalance of underlying subgroups and the total number of observations are studied to determine which impact they have on ability to identify new subgroups. A novel clustering method that can identify clusters in different parts of the feature space is proposed. The approach is applied to methylation data from patients with kidney cancer to create potential biomarkers. Several clustering techniques are applied to gene expression data from bone metastases from patients with prostate cancer to detect novel subtypes. In addition, we test if it is possible to improve the currently used risk-classifier for disease progression in patients with clear cell renal cell carcinoma, by combining clinical data with methylation data.

2. Omics data

Deoxyribonucleic acid or DNA consists of four nucleobases (T, C, G and A) and carries the genetic instructions in all known organisms. The double strained helix shaped structure was described in a scientific paper 1953 written by James Watson and Francis Crick for which they later received the Nobel Prize together with Maurice Wilkins [11]. The process of determining the order of the nucleobases is called DNA sequencing. Since the first DNA sequences were obtained in the 1970s, the methods have developed in a rapid speed. The next generation DNA sequencing (NGS) is the name of a collection of sequencing methods, including Illumina and Roche 454 sequencing, that allows us to sequence DNA and RNA at a much higher speed and considerably lower price than before. NGS can be used to sequence whole genomes or specific areas. The knowledge obtained by DNA sequences has played a major role in several areas, including medical diagnosis of diseases.

A DNA molecule is divided into functional units called genes. The information in the genes becomes useful when it is transcribed into RNA and later translated to a protein. This flow of information is called the central dogma of molecular biology [12], see Figure 1.

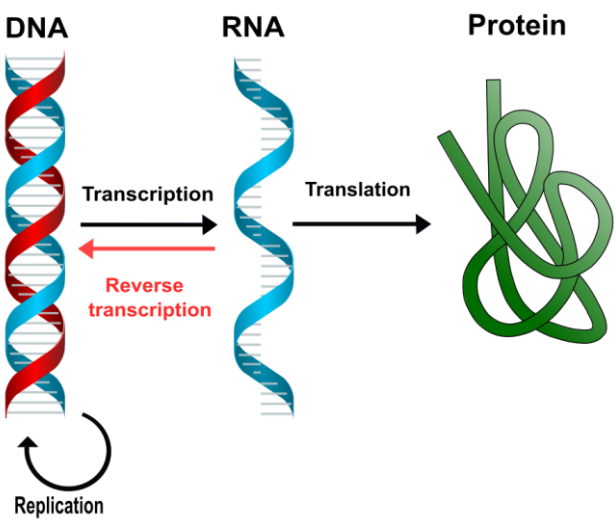

Figure 1. The central dogma of molecular biology.

Proteins play a critical role in the body and do most of the work in the cells. They are required for structure, function and the regulation of the different tissues and organs in our body. Gene regulation or regulation of gene expression is the process used by cells to control the production of certain gene products (proteins or RNA).

2.1. Methylation

Aside from the genetic code that is stored in the DNA, there are other mechanisms involved in the regulation of gene products. Epigenetics is the study of heritable changes in phenotype that does not involve changes in the genetic code. One example of an epigenetic mechanism is the folding of the DNA. The DNA is tightly folded into units called chromosomes. How tight the DNA is folded affects the transcription to RNA and therefore also the gene expression [13]. Another well studied mechanism is DNA methylation, which is the process where a methyl group is added to the DNA molecule. A methyl group consists of four atoms, three hydrogen atoms that are bonded to one carbon atom. Both cytosine (C) and adenine (A) can be methylated, but in mammals, methyl groups are almost exclusively added to cytosines (C) at CpG sites, see Figure 2. A CpG site is a section of the DNA, where a cytosine (C) nucleobase is followed by a guanine (G) nucleobase. In human DNA, about 80% of the CpG sites are methylated [14].

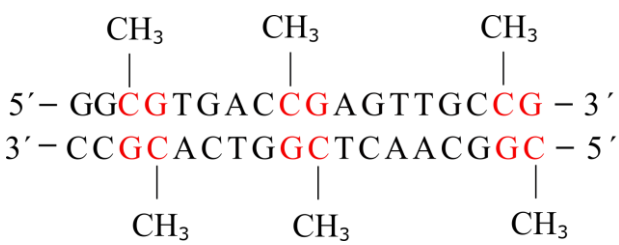

Figure 2. DNA methylation is a mechanism occurring when a methyl group (CH_3) is added to the DNA molecule. The majority of DNA methylation occurs on cytosine (C).

Differences in methylation pattern have been connected to several diseases, including obesity, Rett syndrome and cancer [15]. In paper IV, we use the degree of methylation measured at 450 000 sites in the DNA to build new variables for prediction of risk for cancer progression in patients with clear cell renal cell carcinoma.

2.2. Gene expression

Measuring of gene expression is an important part of pathology and the data can be used for various purposes. A common aim is to identify genes that are

differentially expressed between two treatment groups or between disease subtypes. Those genes can in turn be used to identify disease subtypes.

Measuring of gene expression can be done using different techniques, including RNA sequencing and microarrays, which are described below.

Microarrays

DNA microarrays are slides that contain thousands of tiny spots. Each spot contains known single stranded DNA sequences from a specific gene and the length of the sequence can vary between different platforms. The spots are attached to the surface on the slide on a defined position, corresponding to a gene. How the spot is attached to the solid surface also varies between methods. Messenger RNA (mRNA) is RNA that codes for proteins. After extracting mRNA from the samples, the mRNA is reverse-transcribed into complementary DNA (cDNA). In the hybridization process, cDNA from the samples that have been dyed with fluorescent dye are applied on the microarray chip and bind to the DNA in the spots. In a two-channel microarray there are two colours, one for the sample of interest and another for a control sample, which can be for example a normal sample. The chip is then washed to remove any unbound labelled DNA strands and placed in a laser scanner that activates the fluorescence dye. The intensity of the color is proportional to the amount of cDNA that bounded to the probes and is used as a relative measure of gene expression.

RNA-seq

Unlike microarray experiments where the DNA sequence must be known in advance, RNA-sequencing does not necessarily require any prior sequence knowledge. The exact procedure varies between platforms, but the general steps are similar. Most instruments use DNA for the sequencing and the RNA is therefore converted into a cDNA library. When the RNA sequences are extracted from the sample, the molecules of interest are isolated (mRNA in this case). The fragments are reverse-transcribed into cDNA and fragmented in shorter pieces. The fragmentation can in some cases be performed before the conversion to cDNA. Sequencing adaptors are attached to both ends of the cDNA fragments before a size selection is performed. The cDNA library is then (often) amplified before it is sequenced on a NGS-platform. How long fragments the platforms can read varies between the instruments, but commonly around 50-400 bases long. The third generation sequencing platforms can handle much longer reads [16]. Longer fragments are easier to align to the genome and therefore preferred in the downstream analysis. The sequencing step includes addition of nucleotides that are coloured in different fluorescent colours, one for each of the four bases. The sequences are then determined by reading the colour of each incorporated

nucleotide. The reads from the fragments are stored in FASTQ files, which besides from the sequence contain a per-base quality measure. The raw reads are then aligned to the reference genome to determine where on the genome each read belongs. The human reference genome is continually updated as novel techniques are developed and new discoveries are made [17]. The number of reads that have been aligned to a specific gene is then proportional to the gene expression. The read count is affected by several factors which must be taken into account in the downstream analysis. The length of the gene affects how many reads that are aligned to it, where a longer gene will have more reads. The gene length does not matter so much when the same feature is compared in two or more samples, but could be an issue if two features are compared within one sample. Another factor that affects the read count is the sequencing depth, which is the number of sequenced reads for a sample. Different samples can have different sequencing depth and appropriate normalization should therefore be applied to make gene expression levels comparable between samples.

There are several conceptual differences between microarray and RNA-seq techniques. RNA-seq is better for detecting low expressed genes and the background noise is lower, but the technique is more expensive than microarrays. Papers I and II are based on RNA sequencing data from human cancer tumors, while paper III utilizes microarray data.

3. Pre-processing

Both gene expression data and methylation data require pre-processing before performing analyses to remove e.g. batch effects and technical variation, which can come from both the library preparation and from the sequencing itself. The normalization process aims to remove noise and batch effects and to compensate for e.g. difference in library size. One critical point is that the procedures often assume that the distribution of gene expression values will be the same for all samples, which is not always the case.

3.1. Normalization

Which type of normalization to use on RNA-seq data depends on the aim of the study. Comparisons within a sample requires normalization that accounts for factors like gene length and GC-content, while comparisons between samples requires normalization methods that compensate for differences in library size. All included studies focus on between-sample comparisons. In paper I and II, we used publically available gene expression data. The data sets were generated using RNA-sequencing and quantified using the tool RSEM, which applies the Expectation-Maximization algorithm [18]. There exist other tools for quantification of gene expression from RNA-seq data, e.g. HTSeq and Cufflinks, but we focused on data where RSEM was used for estimating expression levels [19,20]. In paper I, the raw counts were divided by the 75:th percentile for each patient after removing zeroes, followed by multiplication by 1000 and a log-transform. In paper II, the raw counts were instead normalized and transformed using a variance stabilizing transform offered in the R-package Deseq2 [21]. In paper III, we combined gene expression data from two different platforms. The mean-value for each platform was subtracted to remove batch effects. Before centering by the mean, the arrays were quantile normalized. In paper IV methylation data from two different bead types were normalized using the BMIQ method, which is a model based intra-array normalization strategy that is specifically constructed to correct for probe design bias in Illumina Infinium 450k DNA methylation data [22].

4. Feature reduction

When analyzing data where the feature dimension is considerably larger than the number of observations, it is a common procedure to reduce the number variables. We reduce the number of features both by filtering non-informative features and by transforming features into a lower space.

4.1. Filtering

Filtering of low expressed genes is common in analysis of gene expression data. It is especially important to filter genes expressed at low levels when the goal is to identify differentially expressed genes (DEG). Sha et al. [23] showed that the sensitivity for DEG detection increased after filtering up to 20% of low expressed genes. In paper I, the genes with expression values below the 15^{th} percentile in more than 75% of the sample, were removed prior to the analysis. In paper II, we applied a slightly different approach. Genes were given a score based on how many of the samples that had expression values lower than the 25^{th} gene percentile. Based on the score, 25 % of the lowest expressed genes were filtered out. In paper IV we analyzed β-values, which is the estimated level of methylation (ranging from 0 to 1), for 450 000 sites. These were filtered down to approximately 169 0000 by removing e.g. probes located at the X and Y-chromosomes and probes with very low signals. To enable use of our method on data generated from other platforms, we also excluded probes without representation on the Illumina EPIC methylation array.

4.2. Feature selection and feature extraction

After initial filtering of non-informative features, it can be necessary to reduce the feature space further. This can be done either by transforming the features into a space of lower dimension by using methods such as principal component analysis and partial least squares regression, or by ranking features according to how informative they are and selecting only the highest ranked features. The procedure can look very different depending on whether the aim is to perform supervised classification or unsupervised learning. For supervised problems, the known labels or outcomes can be used in order to select informative features, whereas clustering problems requires methods based on other characteristics of the data. Examples of such characteristics are high variation over samples or the presence of two or more “peaks” in the data. A bimodal or multimodal distribution suggests underlying groups in the data, see e.g. a density plot of gene expression value of one gene in the brain cancer data set used in paper I and II (Figure 1). Two distinct peaks in the distribution will make the gene a good candidate for disease biomarker. In paper I and II, we evaluate different feature reduction techniques used for cluster analysis of RNA-seq data.

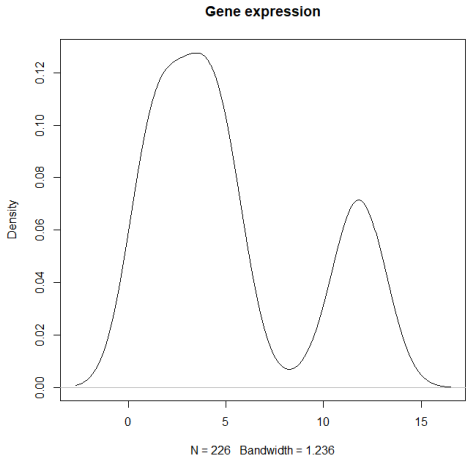

Figure 3. Density plot of gene expression of one gene in 226 patients with lower grade glioma. Two distinct peaks are visible, suggesting the presence of underlying groups among the samples.

5. Clustering

Clustering is an unsupervised learning process, where similar objects are grouped together to form clusters. The definition of a cluster is quite loose and different clustering algorithms utilize different techniques to search for an optimal solution, and since there exists no uniform definition on how to measure similarity of a group of objects, the clustering results can differ substantially. Some clustering techniques require that the user specify how many clusters the data should be divided into, while others give a hierarchy of the objects as output.

Cluster analysis is commonly used within medical research to identify disease subgroups [9,24]. In paper I, we compared different clustering approaches on RNA-seq data from tumor samples, where the disease subgroups were known. We also studied how the data characteristics and the choice of pre-processing method affected how well the clustering techniques were able to identify the subtypes. In paper III, we studied the robustness of the identified subgroups by applying several clustering methods.

Clustering of high-dimensional data is challenging due to several reasons. High dimensional data tend to be sparse and causes all observations to appear equidistant from each other, meaning that the ratio between the nearest and farthest points approaches 1. This is especially true for some distance metrics [25]. A high dimensional feature space also causes different clusters to form in different subspaces of the data. These effects are commonly referred to as “curse of dimensionality”, an expression first used by Richard Bellman [26].

When the feature space is large, relevant features can become masked behind irrelevant features [27]. In paper IV, we developed a novel clustering method called Directed Cluster Analysis (DCA) that captures clusters defined in different subspaces. The method consists of two steps. First clustering was performed on each variable (methylation site) separately, which divides the objects (samples) into two groups. So each variable is a 0/1 vector. Next, the variables are divided into groups. Variables with similar 0/1 profiles will cluster together, see Figure 4. Both clustering steps are made using k-means clustering. The idea behind this approach is that variables that are affected by the same factors will give similar partitions of the patients. We made consensus variables of the clusters by calculating the mean methylation (β-value) for each sample of the all variables included in a cluster. However, as an alternative you could calculate a majority vote for each cluster, which will label each sample as either 0 or 1. This would yield several cluster outputs that could be compared to known partitions from factors such as gender to reduce the list of partitions possible related to disease subtype. This approach is closely related to the concept of biclustering, which is a

data mining technique that identifies a subset of rows with similar patterns across a subset of columns [28]. Unlike biclustering techniques, we cluster in two steps and require similar pattern across all columns (patients) and not just in a subset. Some techniques allow for overlap of the features, whereas our technique allows features to be included in only one cluster.

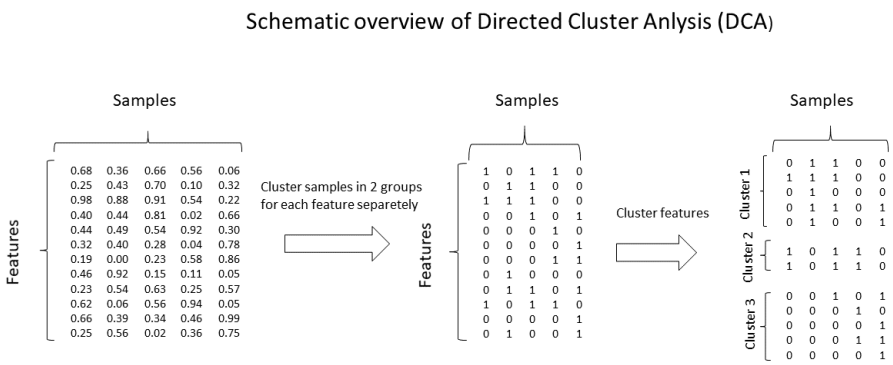

Figure 4. Schematic overview of Directed Cluster Analysis.

6. Classification

Classification is the process of determining which class a new observation belongs to by training a model on data with known class belonging. A challenge when working with high-dimensional data is that classification methods tend to perform poorly when handling data with few observations and huge amount of variables, wherefore some kind of feature reduction technique often is required. Generally, the classification error tends to decrease in the training data when adding more variables to the model, but the model will eventually suffer from overfitting.

In paper I, we used classification as a kind of reference on how strong the genetic signal connected to disease subtype was in the datasets. The method we used is called random forest, which is a classification algorithm consisting of several decision trees. The algorithm uses a version of bagging where a subset of both samples and features are used in each split and therefore prevent the trees from becoming too correlated [29].

In paper IV, we used logistic regression to classify patients into high or low risk groups for disease progression. Logistic regression falls within the category of linear classifiers and models a binary dependent variable or response by a linear combination of the independent variables or predictors, which can be both continuous and binary. By fitting data to a logit function, it predicts the probability of an event occurring (in our case disease progression within five years). The logistic regression model is described by:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

where p is the probability of an event i.e. $P(Y = 1)$, Y is the response variable. $\beta_0, \ldots, \beta_k$ are the model parameters and $x_1, \ldots, x_k$ are the predictors.

The left hand side of the equation gives the log odds of an event, which can be converted to get the probability p:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

The posterior probability (the probability of an event given a set of predictors) is then used to classify patients. The standard cutoff is at 0.5. By altering the cutoff, we controlled the true positive rate to 85%, which allowed for comparison to the Mayo scoring system, which is currently in use in Sweden for prediction of

outcome and treatment stratification for patients with clear cell renal cell carcinoma.

In paper III, we first applied cluster analysis to identify subtypes of bone metastases in patients with prostate cancer. Using supervised analysis, we then identified the top 20 differentially expressed genes, which were included in a prediction model. Prediction was made using orthogonal projections to latent structures discriminant analysis (OPLS-DA), which is a latent variable method that creates new components as linear combinations of the original variables and use these as prediction variables [30]. OPLS-DA can model both variation connected to subtype information as well as uncorrelated (orthogonal) variation, i.e. within class variation. It uses the class information to decompose the data matrix and remove variation which is not correlated to cancer subtype.

7. Evaluation

Evaluation of clustering methods is a difficult task, both since the true partitions are seldom known in advance and since there exist several ways of determining similarity between partitions. In papers I,II and IV, we assumed that the true partitions were known. In papers I and II, we evaluated the performance of the clustering methods by comparing the clustering result to the known partition using adjusted Rand index (ARI). ARI is a modified version of Rand index, that adjusts for agreement by chance [31]. The value 1 indicates that the compared partitions are identical, and the value 0 indicates that the agreement is as poor as expected by chance.

We consider the known disease subgroups as the gold standard partition. However, we know that there exists several other partitions in the data as well. One such known partition is the gender of the patients. Different groupings of the samples may exist in different feature subspaces. One way to evaluate the clustering result would be to define the gold standard as the groups defined by all know factors, which then will be several small groups of patients. But this will make the comparisons more difficult since a patient misclassified with respect to gender will be judged equally as a misclassification w.r.t disease subtype.

How well a clustering method is able to identify groups defined by a factor of interest depends on both the clustering algorithm itself, but maybe even more on how many other factors that differentiate the objects, and how strongly the variables are affected by the factors. A low value of for example adjusted Rand index does not necessarily imply that the method has low performance, it can indicate the presence of a factor with stronger signal than the one defining the gold standard partition.

There is a substantial difference between unsupervised clustering and supervised classification. Since supervised classification trains the model to differentiate on the factor of interest, it is not as sensitive to the presence of irrelevant features as clustering techniques are. In paper I, we used supervised classification as a positive control when evaluating clustering techniques.

In paper IV, we evaluated the classification performance by comparing the sensitivity and specificity between different models. The sensitivity is the true positive rate, which in this case means the proportion of patients with disease progression within five years that were classified as high risk. The specificity is the true negative rate, which corresponds to the proportion of patients without disease progression within five years that were classified as low risk. One could choose to look only at the percentage of correctly classified patients, which would

make it easier to compare models with each other. That could however be misleading in the case of very skew subgroups, since the models can classify all patients to belong to one group and that would yield quite high performance. If the consequences of misclassifying patients in one of the groups are more severe than misclassifying patients in the other group, it is beneficial to use sensitivity and specificity rather than only percentage of correctly classified patients. It does however make it more complicated to compare different models. In our case, we chose the cutoff for the posterior probability to fixate the sensitivity. In that way, we could compare the models using only the specificity.

In paper III, the classification model was validated on external data for which the disease subgroups were unknown. Since the subgroups were unknown, the validation did not give a measure of performance, but the relative distribution of the predicted subtypes could be compared to that obtained in our data.

8. Results

In paper I, we showed that choice of clustering algorithm had an effect on the ability to identify cancer subtypes using RNA-seq data, but no method completely outperformed the others in all data sets. The relative distribution of the cancer subtypes affected the clustering performance, where very skewed distributions often gave lower accuracy. A limited negative effect was observed for reduction of the sample size. The presence of other partitions in the data can lower the ability of finding partitions related to disease subtypes. By analyzing the genders separately, we observed that the gain of analyzing a more homogeneous data could outweigh the negative effect of a smaller sample size.

The results of comparisons between feature selection methods in paper II, showed that the choice of selection procedure can have major effect on the cluster analysis. Relatively low overlap of the 1000 selected genes between the best performing methods was observed. This suggests that it might be beneficial to combine feature selection methods. The performance of different feature selection methods were compared to a case where no feature selection was applied. The results showed that the average performance of the 13 selection methods was lower than the negative control (no selection) in three of the data sets. This highlights some of the dangers with unsupervised feature selection, where if unlucky, the selection will reduce the performance.

In paper III, we identified three molecular subtypes of bone metastasis from patients with prostate cancer using cluster analysis. We showed that the results were robust, by applying five different clustering algorithms that all generated very similar partitions of the patients, see Figure 5. Using the 20 most differentially expressed genes with respect to the defined subtypes, we constructed a classification model. The model was applied to an external data set and the classification result showed frequencies of the subtypes comparable to those observed in our data.

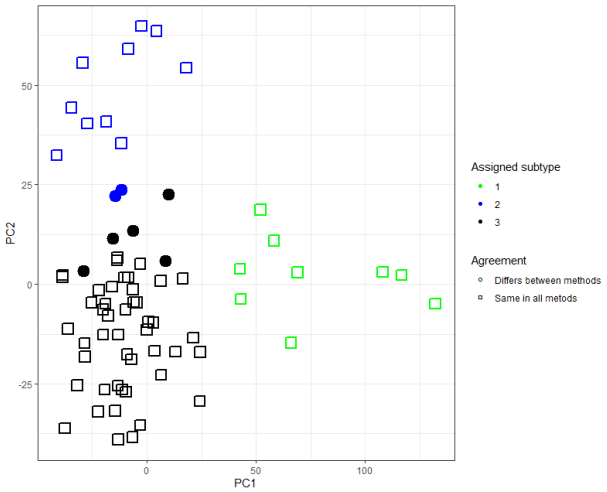

Figure 5. Agreement of five clustering partitions of bone metastasis samples from patients with prostate cancer. The clustering was performed on the first two principal components.

In paper IV, we presented a novel clustering method that can detect clusters in different parts of the feature space. We used it to construct possible biomarkers based on methylation data from patients with kidney cancer (clear cell renal carcinoma). Today, a model utilizing only clinical data is used to classify patients into risk categories with different follow up-strategies. Our results showed that methylation data could be used together with clinical data to improve risk classifications.

9. Discussion

Talking about performance of feature selection methods or clustering methods can be confusing. A low similarity to the gold standard does not necessarily imply that the method worked poorly. It can indicate that the method identified some other partition in the data. When considering data from human samples, there will exist several factors that can be used to distinguish the samples and many of them are not known to us.

In the unsupervised case, we cannot use class information or follow-up data to select informative features. We can search for features with high variation among samples or features with clear peaks in their distribution, but we have no way to tell if these attributes are related to disease subtype. The example in Figure 3, show the density of gene expression for one gene measured in tumor samples from patients with brain cancer. Although the distribution has two quite distinct peaks, the gene holds very little information related to our defined gold standard partition but it is probably very informative for some other partition of the data.

For three of the data sets in paper II, the average clustering performance (based on genes selected by the 13 different feature selection methods) were lower than the performance using clustering based on all genes. This suggest the presence of strong genetic signals unrelated to our defined gold standard.

There are many factors that affect the ability to identify cancer subtypes. Aside from clustering methods, we have analyzed some pre-processing choices and data characteristics. However, there is a need for further investigations to see how feature selection methods are affected by e.g. imbalance of subtypes and determining how different normalization procedures affect feature selection methods. We observed that the overlap of selected genes between high performing selection methods were low, and it would be of interest to investigate further if it is possible to combine different gene selection methods to improve clustering results.

10. Summary of papers

10.1. Paper I

In paper I, entitled *Cluster analysis on high dimensional RNA-seq data with applications to cancer research – An evaluation study*, we studied changes in the ability to identify cancer subtypes based on different choices of pre-processing and clustering algorithms. The performance was evaluated on publicly available gene expression data from four cancer types. The true subtypes were known in advanced and used for evaluation, whereas the analyses were unsupervised. Both the choice of feature reduction/selection method and clustering algorithm had an effect on the ability of identifying cancer subtypes, but with big differences in ranking of the methods between the data sets, it was hard to draw any general conclusions. The results showed that the benefit of obtaining a more homogeneous data by dividing the samples by gender was greater than the disadvantage caused by smaller sample sizes. The study was performed on a variety of different sample sizes and using different distributions of the underlying subtypes.

10.2. Paper II

Paper II, *Comparison of methods for variable selection in clustering of high-dimensional RNA-sequencing data to identify cancer subtypes,* compares 13 feature selection methods by applying them to four human cancer RNA-seq data sets before performing cluster analysis to identify cancer subtypes. The performance was evaluated by comparing to the case were 1) no selection was performed, 2) to a supervised approach and 3) to a random selection. The characteristics of the top ranked genes were studied and the overlap of selected genes was compared between the different methods. The study showed that the dip-test and the bimodality index performed among the best, whereas two methods based on co-expression between genes performed poorly. However, the performance was dependent on data set and the distribution of cancer subtypes. Low overlap of top ranked methods suggests that it might be beneficial to combine two or more gene selection methods.

10.3. Paper III

Robust clustering identified subtypes in patients with prostate cancer in the third paper, *Gene expression profiles define molecular subtypes of prostate cancer bone metastasis with different outcome and morphology traceable back to the primary tumor*. The most common place for prostate cancer to spread is to the bone. At this advanced stage there is no cure, so the treatment is palliative. The aim in this paper was to identify variability in bone metastasis that could be of

importance for therapy. Clustering of gene expression profiles from bone metastases, revealed three subtypes with differences in outcome. Several clustering techniques were applied to the data, with very similar results, confirming the robustness of the identified subgroups. A classification model for the identified subtypes was built and later tested on an external data set, which resulted in subtype frequencies comparable to the training set.

10.4. Paper IV

In Paper IV, with the title: *Combining epigenetic and clinicopathological variables improves prognostic prediction in clear cell Renal Cell Carcinoma*, we combined methylation data with clinical data to improve prognostic prediction of patients with kidney cancer. The current clinical practice in Sweden to predict risk of progression for patients with clear cell renal cell carcinoma is based on clinical variables such as tumor diameter and histologic grade. The idea in this paper was to create new variables based on methylation profiles from tumor samples that could complement the clinical variables in the classification of disease progression. The variables were constructed by a novel two step clustering method, Directed Cluster Analysis. First, for each methylation site the samples were clustered into two groups, resulting in 0/1 profiles. At the next step, the profiles were clustered into groups and the mean methylation value taken over all methylation sites included in each group constituted the new variables. We treated the variables as potential biomarkers and used them in a classification model. Using our constructed variables resulted in approximately the same classification accuracy as using only the clinical variables. By combining our variables with previously identified biomarkers and clinical data we were able to build a classifier that was slightly better than the one that is in clinical use today.

The idea behind the Directed Cluster Analysis is that sites that are affected by the same factor, e.g. gender, should have similar 0/1 profiles and hence cluster together. Calculating the majority vote for the samples in each cluster will result in different clusters for different subsets of the feature space. One can then compare the sample partitions with known factors to disregard partitions that are not of interest.

References

1. van den Tweel JG, Taylor CR. A brief history of pathology: Preface to a forthcoming series that highlights milestones in the evolution of pathology as a discipline. *Virchows Archiv : an international journal of pathology*. 2010;457(1):3-10.
2. Janeway CA, Jr, Travers P, Walport M, Shlomchik MJ. *Immunobiology: The Imune System in Health and Disease*. Vol 5th edition. New York: Garland science; 2001.
3. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. *Journal of molecular biology*. 2013;425(21):3919-3936.
4. Tycko B, Ashkenas J. Epigenetics and its role in disease. *The Journal of clinical investigation*. 2000;105(3):245-246.
5. Global Burden of Disease Cancer C, Fitzmaurice C, Allen C, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA oncology*. 2017;3(4):524-548.
6. Hansford S, Huntsman DG. Boveri at 100: Theodor Boveri and genetic predisposition to cancer. *The Journal of Pathology*. 2014;234(2):142-145.
7. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931-945.
8. Sjöblom T, Jones S, Wood LD, et al. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science (New York, N.Y.)*. 2006;314(5797):268.
9. Lapointe J, Li C, Higgins JP, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101(3):811-816.
10. Bertucci F, Finetti P, Rougemont J, et al. Gene Expression Profiling Identifies Molecular Subtypes of Inflammatory Breast Cancer. *Cancer Research*. 2005;65(6):2170-2178.
11. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*. 1953/04/01 1953;171(4356):737-738.
12. Crick F. Central Dogma of Molecular Biology. *Nature*. 1970/08/01 1970;227(5258):561-563.
13. Gibcus JH, Dekker J. The context of gene expression regulation. *F1000 Biol Rep*. 2012;4:8-8.
14. Lavie L, Kitova M, Maldener E, Meese E, Mayer J. CpG Methylation Directly Regulates Transcriptional Activity of the Human Endogenous Retrovirus Family HERV-K(HML-2). *Journal of Virology*. 2005;79(2):876-883.

15. Jin Z, Liu Y. DNA methylation in human diseases. *Genes Dis*. 2018;5(1):1-8.
16. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect*. 2018;24(4):335-341.
17. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*. 2017/03/01/ 2017;109(2):83-90.
18. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*. August 04 2011;12(1):323.
19. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511-515.
20. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*. 2015;31(2):166-169.
21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014/12/05 2014;15(12):550.
22. Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics (Oxford, England)*. 2013;29(2):189-196.
23. Sha Y, Phan JH, Wang MD. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. *Conf Proc IEEE Eng Med Biol Soc*. 2015 2015;2015:6461-6464.
24. Ren Z, Wang W, Li J. Identifying molecular subtypes in human colon cancer using gene expression and DNA methylation microarray data. *International Journal of Oncology*. 2016;48(2):690-702.
25. Aggarwal CC, Hinneburg A, Keim DA. On the Surprising Behavior of Distance Metrics in High Dimensional Space2001; Berlin, Heidelberg.
26. Bellman R. *Adaptive Control Processes: A Guided Tour*: Princeton University Press; 1961.
27. Ronan T, Qi Z, Naegle KM. Avoiding common pitfalls when clustering biological data. *Science Signaling*. 2016;9(432):re6.
28. Padilha VA, Campello RJGB. A systematic comparative evaluation of biclustering techniques. *BMC bioinformatics*. 2017/01/23 2017;18(1):55.
29. Breiman L. Random Forests. *Mach. Learn*. 2001;45(1):5-32.
30. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*. 2002;16(3):119-128.
31. Hubert L, Arabie P. Comparing partitions. *Journal of Classification*. 1985/12/01 1985;2(1):193-218.